EMPIRICAL PROCESSES
BASED ON REGRESSION RESIDUALS: THEORY AND APPLICATIONS

by

Gemai Chen

M.Sc. Simon Fraser University 1987

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in the Department of Mathematics and Statistics

of

Simon Fraser University

SIMON FRASER UNIVERSITY

August, 1991

APPROVAL

Name: Gemai Chen

Degree: Doctor of Philosophy

Title of thesis: **EMPIRICAL PROCESSES BASED ON REGRESSION RESIDUALS THEORY AND APPLICATIONS**

Examining Committee: Dr. Alistair H. Lachlan
Chairman

Dr. Richard A. Lockhart, Senior Supervisor

Dr. Michael A. Stephens

Dr. Rick Routledge

Dr. Charmaine Dean

Dr. Reg Kulperger, External Examiner

Date Approved: August 19, 1991

PARTIAL COPYRIGHT LICENSE

I hereby grant to Simon Fraser University the right to lend my thesis, project or extended essay (the title of which is shown below) to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users. I further agree that permission for multiple copying of this work for scholarly purposes may be granted by me or the Dean of Graduate Studies. It is understood that copying or publication of this work for financial gain shall not be allowed without my written permission.

Title of Thesis/Project/Extended Essay

Empirical Processes Based on Regression Residuals: Theory and Applications

Author: _____
(signature)

GEMAI CHEN
(name)

Aug 19, 1991
(date)

Abstract

The problem of testing goodness-of-fit when fitting linear regression models is approached in this thesis through a careful study of the weak convergence properties of the empirical processes based on regression residuals. For normal theory linear regression models, the effect of estimating regression parameters and the effect of estimating the standard deviation of the error distribution are shown to be separable, and are each identified. Also identified is the effect of the Box-Cox transformation on estimation of regression parameters and error standard deviation. The weak convergence properties established here cover two different situations: (1) the number of regression parameters is fixed finite; (2) the number of regression parameters increases as sample size increases.

When the error distribution is not normal, a group of contiguous alternatives are studied in detail, and weak convergence properties of residual empirical processes under these contiguous alternatives are obtained.

Applications of the above mentioned weak convergence properties are sought in three areas: (a) testing overall goodness-of-fit when fitting linear regression models; (b) testing overall goodness-of-fit for Box-Cox transformations; and (c) testing composite goodness-of-fit hypotheses for continuous distributions.

Proposals are made to extend the basic ideas of EDF tests to the areas of generalized linear models (GLIM) and transform-both-sides (TBS) models.

Acknowledgements

Opportunities are believed to be everywhere and of various kinds, no matter whether you are optimistic or pessimistic, yet the chance of having a particular opportunity is hardly known, even to a great statistician. Therefore, I feel especially fortunate to have met Dr. Stephens in China. With his help I came to North America, and throughout the time I was at SFU, he influenced me in his unique way. To Dr. Stephens I give my sincerest thanks.

Being a student, I can not think of anything nicer than being able to knock at the doors of my professors and have thought-provoking talks with them. Dr. Routledge, Professor Villegas and Dr. Weldon, among other faculties, were always kind and available to open their doors. To all of them, I want to say thank you, and I wish you the best.

I strongly believe that by luck or fate alone I cannot explain how lucky and fortunate I am to have had Dr. Lockhart as my supervisor. Not only because he is one of the most friendly and knowledgeable professors in the department, but his openness, his insights into hard problems and his care and help which go well beyond academic matters stimulated and are still stimulating me as well. My sincerest thanks also go to you, Dr. Lockhart.

Finally, I want to thank Sylvia, Maggie, Diane and other department staff. It was their kindness and help that made my stay at SFU comfortable. A special thank is to the Department of Mathematics and Statistics which supported me financially.

Dedication

To the memory of
my mother and my father.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

It is exciting experience studying the works concerning empirical distribution functions and the related empirical processes.

Let $y_1, y_2, \ldots, y_n$ be n real-valued observations from an experiment. For any real value y, let the proportion of the y_i's that are less than or equal to y be denoted by $F_n(y)$. For a fixed y, $F_n(y)$ is a function of the sample values y_i, thus it is random; for a fixed sample $\{y_1, y_2, \ldots, y_n\}$, $F_n(y)$ is a real-valued step function of y, also called a *sample path*, taking values between zero and one. Such a (random) function as $F_n(y)$, which also enjoys all the properties of an ordinary distribution function, is called an *empirical distribution function*.

1.1 Why Empirical Distribution Function?

One of the most basic statistical problems is to learn about the behaviour of a population as a whole through taking sample observations on certain representative states of the population. It is not difficult for one to see that some principles are needed to guide the way in which the population is defined properly for a planned study, and the way in which data are collected and analyzed to answer questions proposed in the planned study. Among these considerations, one sooner or later needs to think about the following fundamental questions:

How can one know that one can come to understand population behaviour by just gathering and studying sample observations of the population? If one can, how well can one know the population?

Although the world is full of uncertainties and mysteries, within the realm of statistics the answers to the above questions are certain and clear. In its simplest yet perhaps the most informative form the answers read: *Yes, sample observations do reveal population properties, and with the increase of sample information, the understanding of the population characteristics is continuously improved, and when sample information piles up to infinity, one is bound to know the population for sure.*

In a more formal way to express the above answers, suppose the population distribution function is $F(y)$, from which a random sample $\{y_1, y_2, \ldots, y_n\}$ is taken. Then the famous Glivenko-Cantelli theorem says: almost surely,

$$\sup_{-\infty<y<+\infty} |F_n(y) - F(y)| \longrightarrow 0$$

as $n \to \infty$. That is, the empirical distribution function $F_n(y)$ serves the basic role of exploring population properties through sample observations. Loève (1955) called the above theorem "the fundamental theorem of statistics" and Pitman (1972, p79) called the above theorem "the existence theorem for statistics as a branch of applied mathematics."[1] It is noted here that the sample information $F_n(y)$ and the population $F(y)$ are compared over the whole range directly.

1.2 Empirical Processes

Almost sure properties are indeed good properties if estimation problems are of concern. However, distribution properties are often needed in order to test statistical hypotheses. For the above $F_n(y)$ and $F(y)$, that $|F_n(y) - F(y)|$ converges almost surely to zero uniformly in y implies that for every y the limiting distribution of $|F_n(y) - F(y)|$ is degenerate, i.e.,

[1] Another fundamental theorem is the central limit theorem of probability theory.

taking the value zero with probability one. This is a result that is of little use in telling the distributional difference between $F_n(y)$, which is known, and $F(y)$, which is usually unknown.

In fact, one does not have to work with the limiting distribution if one can find out the finite sample distribution directly. In the present context, it is possible to do so for $|F_n(y) - F(y)|$ or its supremum over y provided $\{y_1, y_2, \ldots, y_n\}$ is a random sample from the population $F(y)$ and $F(y)$ does not contain unknown parameters which need to be estimated. In general, however, finding exact finite sample distributions involving empirical distribution functions can be very difficult, if not impossible.

Truth is often told in ideal situations as can be seen from the above arguments. But to work towards seeing the truth, fine balance between ideal situations and practical situations is necessarily required. Luckily, another fundamental theorem exists in the study of empirical distribution functions which points a way out. Let

$$Y_n(y) = \sqrt{n}\,\{F_n(y) - F(y)\}$$

which is called the *empirical process* of the sample $\{y_1, y_2, \ldots, y_n\}$. For random sampling, this theorem says: as $n \to 0$, $Y_n(y)$ converges weakly[2] to a Gaussian process[3], whose distribution is non-degenerate and calculable. This is true even when some unknown parameters of $F(y)$ have to be estimated.

The balance that this thesis is to take can be described as below: After randomly sampling a hypothesized distribution with known form but unknown parameters, some parameters of interest are estimated, say, through doing regression, and quantities called *residuals* are produced and used to construct *estimated* or *residual empirical distribution function* $\hat{F}_n(y)$ and further *estimated* or *residual empirical process* $\hat{Y}_n(y)$[4]. The limiting distribution of the residual empirical process $\hat{Y}_n(y)$ is then found and used to check the

[2]See section 2.1 for the definition of weak convergence of stochastic processes.
[3]See section 2.1 for the definition of a Gaussian process.
[4]See section 2.1 for detailed definitions.

adequacy of the hypothesized distribution in regard to describing the sample observations.

In other words, residuals from doing regression will take the place of the random sample in the ideal situations mentioned above, and results similar to those obtainable in the ideal situations using random samples will be obtained using the residuals. This approach for finding distributions is well supported by the rich literature in the general theory of function spaces and has proved successful. As a result, it is termed *the stochastic process approach.*

1.3 A Review of Literature

To clarify some of the points made in the above two sections further, and to put this thesis into perspective, a review of the literature on which this thesis is based is given below. This review is not meant to be comprehensive. Instead, this review begins by introducing the historical background in some detail, then a highly selected and highly condensed summary of those papers which either influenced the research done in this thesis directly, or pointed to different approaches and directions in the area of empirical distribution functions and empirical processes, is presented. For a good one-volume source, one can consult Shorack and Wellner (1986).

The pioneer work was J. Doob's 1949 paper, *Heuristic approach to the Kolmogorov-Smirnov theorems.* There are two important themes studied in this paper: the first theme challenged the later theoretical statisticians; the second theme is called *the invariance principle* today. Together, these two themes laid down the corner-stone of the stochastic process approach, mentioned in section 1.2.

Since if a random variable Y is distributed according to a continuous distribution $F(y)$, then $U = F(Y)$ is distributed as a uniform random variable $U(0,1)$, one can consider the basic case of a random sample $\{u_1, u_2, \ldots, u_n\}$ from $U(0,1)$. In this case let $U_n(t)$ be the proportion of the u_i's that are less than or equal to t, $0 \leq t \leq 1$, and let $B_n(t) = \sqrt{n}\{U_n(t) - t\}$, $0 \leq t \leq 1$.

For any fixed t, $nU_n(t)$ is a binomial (n,p) random variable, where $p=t$, and $B_n(t)$ is $nU_n(t)$ standardized, therefore, as $n \to \infty$, $B_n(t)$ converges in distribution to the normal distribution $N(0, t-t^2)$. Similarly, for any pair (t_1,t_2), $(B_n(t_1), B_n(t_2))$ converges in distribution to the bivariate normal distribution with zero mean and covariance $\min(t_1,t_2)-t_1t_2$. In fact, for any finite k and any $t_1,t_2,\ldots,t_k$, the k-vector $(B_n(t_1), B_n(t_2),\ldots,B_n(t_k))$ converges in distribution to $(B(t_1), B(t_2),\ldots,B(t_k))$, a k-variate normal distribution with zero mean and covariance Σ whose $(i,j)^{th}$ element is $\min(t_i,t_j)-t_it_j$. This type of convergence is called *convergence in finite distributions* and is denoted $B_n \to_{f.d.} B$, as $n \to \infty$. For the present case, the process $B(t)$ is the so-called *tied-down Brownian motion* or *Brownian bridge* characterized by:

(B1) B has continuous sample paths between the two fixed points $t=0$ and $t=1$,

(B2) For any finite k and any $t_1,t_2,\ldots,t_k$, $(B_n(t_1),\ldots,B_n(t_k))$ has a multivariate normal distribution with zero mean and a covariance matrix whose $(i,j)^{th}$ element is

$$\rho(t_i,t_j) = \mathrm{Cov}(B(t_i), B(t_j)) = \min(t_i,t_j) - t_it_j.$$

To test whether a random sample really come from the uniform distribution, many statistics have been suggested. Among those already available in the late 40's, the Kolmogorov-Smirnov (two-sided) statistic is defined by

$$D_n = \sup_{0\le t\le 1} |U_n(t)-t|.$$

The finite sample distribution of D_n was difficult and unknown[5] when Doob wrote his paper, but it was known that for any $b > 0$,

$$P(\sup_{0\le t\le 1} |B(t)| \ge b) = 2\sum_{k=1}^{\infty} e^{-2k^2b}.$$

Since $B_n(t) = \sqrt{n}\{U_n(t)-t\} \to_{f.d.} B(t)$, Doob argued (he wrote x_n and x instead of B_n and B):

[5]This is known now, see Durbin, J. (1973), section 2.4.

> We shall assume, until a contradiction frustrates our devotion to heuristic reasoning, that in *calculating asymptotic $x_n(t)$ process distributions when $n \to \infty$ we may simply replace the $x_n(t)$ process by the $x(t)$ process.* It is clear that this cannot be done in all possible situations, but let the reader who has never used this sort of reasoning exhibit the first counter example.

Two things are implied by Doob's heuristic argument. The first thing is the relationship between convergence in finite distributions and what nowadays is called *weak convergence*[6]. Doob was quite right to say "...this cannot be done in all possible situations, ..." because convergence in finite distributions does not imply weak convergence in general. This problem was subtle and theoretical statisticians and probabilists worked hard for years to develop a useful theory of weak convergence. See Billingsley (1968).

The second thing implied is the invariance property which roughly says that the limiting distributions of the statistics calculated from B_n in a continuous fashion should be the same as that calculated from B, the weak limit of B_n. The invariance property allows one to design different statistics to catch different features of the problem in such a way in which one does not need to worry about finding the limiting distributions of the statistics proposed, as long as the basic weak convergence from B_n to B has been established.

As with many landmark works, Doob's conjecture has been proved right (Donsker, (1952)). More importantly, his conjecture directed people to develop useful theories and to find other interesting applications.

Given the above historical background, the following is an annotated list of references given in chronological order.

1. Kac, M. and Siegert, A.J.F. (1947). *An explicit representation of a stationary Gaussian process.* A Gaussian process with a positive definite and square-integrable covariance function (or kernel) can be written as a weighted infinite sum of independent

[6]See section 2.1 for definition.

chi-square random variables, each of which is on one degree of freedom. The weights are all positive, and are eigenvalues to a Fredholm integral equation of the first kind in which the covariance function is the kernel. This result is the theoretical basis of some numerical calculations.

2. Donsker, M.D. (1952). *Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems.* The material here is a version of the invariance principle discussed above.

3. Anderson, T.W. and Darling, D.A. (1952). *Asymptotic theory of certain 'goodness-of-fit' criteria based on stochastic processes.* A detailed account of the stochastic approach to finding goodness-of-fit test statistics and the corresponding (limiting) distributions is given in this paper.

4. Darling, D.A. (1955). *The Cramér-von Mises test in the parametric case.* Centered at the concept of convergence in finite distributions, a fairly general formulation of constructing Cramér-type goodness-of-fit tests is discussed.

5. Kac, M., Kieffer, J. and Wolfowitz, J. (1955). *On tests of normality and other tests of goodness-of-fit based on distance method.* The effects of estimating mean and variance on the limiting distribution of the test statistics based on integral-type distance are found explicitly. Many later works can be viewed as extensions of this basic result.

6. Imhof, J.P. (1961). *Computing the distribution of quadratic forms in normal variables.* A general working method is presented here to calculate probabilities related to quadratic forms in normal random variables. This is needed to carry through the stochastic approach to perform goodness-of-fit tests.

7. Sukhatme, S. (1972). *Fredholm determinant of a positive kernel of a special type and its applications.* Ways of finding eigenvalues to Fredholm integral equations of the first

kind with a special type of kernels are investigated. The results are useful to perform goodness-of-fit tests when some parameters are estimated.

8. Durbin, J. (1973a). *Weak convergence of the sample distribution function when parameters are estimated.* A long-waited serious treatment of the weak convergence of some basic empirical processes is given here for the independent and identically distributed case.

9. Stephens, M.A. (1974). *EDF statistics for goodness-of-fit and some comparisons.* Practically usable EDF tests of goodness-of-fit and their empirical powers are discussed.

10. Rao, J.S. and Sethuraman, J. (1975). *Weak convergence of empirical distribution functions of random variables subject to perturbations and scale factors.* A method for rigorously proving weak convergence of empirical processes is demonstrated.

11. Stephens, M.A. (1976). *Asymptotic results for goodness-of-fit statistics with unknown parameters.* Some fine relationships are given between eigenvalues and the means and variances of some EDF statistics for testing goodness-of-fit of normal and exponential distributions with unknown parameters.

12. Neuhaus, G. (1976). *Weak convergence under contiguous alternatives of the empirical processes when parameters are estimated: The D_k approach.* Weak convergence of multi-dimensional empirical processes is studied.

13. Mukantseva, L.A. (1977). *Testing normality in one-dimensional and multi-dimensional linear regression.* Empirical processes defined using residuals from straight line regressions with normal errors are investigated in detail. Extension to general linear regression situation is indicated.

14. Pierce, D.A. and Kopecky, K.J. (1979). *Testing goodness-of-fit for the distribution of errors in regression models.* It is shown here that the covariance function of the

empirical processes defined using residuals from fitting linear regression models with normal errors and with a constant in the model matrix has the information structure of a location and scale family.

15. Loynes, R.M. (1980). *The empirical distribution function of residuals from generalized regression.* Weak convergence of residual empirical processes is considered for independent but non-identically distributed error distributions.

16. Khmaladze, E.V. (1981). *Martingale approach in the theory of goodness-of-fit tests.* Weak convergence of empirical processes is studied through martingale theory.

17. Koul, H.L. (1984). *Tests of goodness-of-fit in linear regression.* A fairly general formulation for studying ordinary and weighted empirical processes and rank processes with general errors is presented here.

18. Shorack, G.R. (1984). *Empirical and rank processes of observations and residuals.* A unified approach to investigate empirical and rank processes with or without nuisance parameters is provided.

19. Pierce, D.A. (1985). *Testing normality in autoregressive models.* This paper tries to extend the results that hold for residual empirical processes from normal theory linear regression to stationary and invertible autoregressive models.

20. D'Agostino, and Stephens, M.A. (1986). *Goodness-of-fit techniques.* A comprehensive treatment of commonly used goodness-of-fit methods is given here for applications.

21. Meester, S.G. and Lockhart, R.A. (1988). *Testing for normal errors in designs with many blocks.* This paper studies EDF tests of goodness-of-fit in analysis of variance models when the number of parameters increases linearly with sample size.

22. Koul, H.L. and Levental, S. (1989). *Weak convergence of the residual empirical process in explosive autoregression.* This paper tries to show that the results valid for residual

empirical processes from normal theory linear regression can be extended to explosive autoregression.

23. Hjort, N.L. (1990). *Goodness-of-fit tests in models for life history data based on cumulative hazard rates.* This paper represents the current status of a new approach on goodness-of-fit problems in life testing and survival analysis. The theory of counting-process is used and cumulative hazard functions are studied directly.

24. Terry, M. *et al.* (1990). *Martingale-based residuals for survival models.* Martingale, score and deviance residuals are reviewed for graphical methods of diagnostics for survival models. Analytic methods based on these residuals would be of potential applications and are worth investigating.

The plan for this thesis is the following. Chapter 2 presents a careful study of the weak convergence properties of the empirical processes based on regression residuals. For normal theory linear regression models, the effect of estimating regression parameters and the effect of estimating standard deviation of the error distribution are shown to be separable, and are each identified. Also identified is the effect of the Box-Cox transformation on estimation of regression parameters and error standard deviation. The weak convergence properties established here cover two different situations: (1) the number of regression parameters is fixed finite; (2) the number of regression parameters is increasing as sample size increases.

When the error distribution is not normal, a group of contiguous alternatives are studied in detail, and weak convergence properties of residual empirical processes under these contiguous alternatives are obtained. Chapter 3 contains this material.

Chapters 4 to 6 are direct applications of the previous two chapters. The problem of testing overall goodness-of-fit when fitting linear regression models is studied in Chapter 4; the problem of testing overall goodness-of-fit for Box-Cox transformations is investigated in Chapter 5; and the problem of testing composite goodness-of-fit hypotheses for continuous distributions is examined in Chapter 6.

Chapter 7 proposes extensions and further developments of the basic ideas of EDF tests into the areas of generalized linear models (GLIM) and transform-both-sides (TBS) models.

Chapter 2

Residual Empirical Processes: Linear Regression

Empirical processes constructed by using residuals from ordinary normal theory linear regression are studied in this chapter. To this end, the concept of weak convergence in metric spaces is introduced in section 2.1, and some useful properties of least squares estimators of regression parameters are given in section 2.2. The next three sections are devoted to three important situations.

Section 2.3 presents two theorems which handle the effect of estimating the standard deviation of the error distribution. Theorem 2.3.1 identifies the effect of estimating the standard deviation σ of the error distribution; Theorem 2.3.2 separates the effect of estimating σ and the effect of estimating regression parameters θ and shows that any weak convergence result for the empirical processes of residuals when σ is known implies a corresponding result when σ must be estimated. These two theorems allow one to concentrate, without losing any rigour, on the effect of estimating regression parameters θ.

Section 2.4 studies the effect of estimating regression parameters θ when the dimension of θ is fixed finite. This effect is identified in Theorem 2.4.1; and a rigorous treatment of

the related weak convergence problems is given in Theorem 2.4.2. The main results for estimating both σ and θ are summarized in Theorem 2.4.3. It is noted that Theorem 2.4.2 is a special but interesting application of Loynes (1980) and Rao and Sethuraman (1975). Some comments are given at the beginning of Section 2.4.

Section 2.5 studies empirical processes related to analysis of variance models. By paying close attention to the simple structure of the model matrix, a less well-known weak convergence result, which was obtained for one-way layout by Meester and Lockhart (1988), is shown to hold for more general analysis of variance models, including balanced two-way layout (without interactions and with interactions, respectively), randomized complete block designs, and two-factor nested models.

2.1 Weak Convergence in Metric Spaces

Throughout this thesis, let $(\Omega, \mathcal{A}, P)$ be the underlying probability space. Let $(S, \mathcal{E}, d)$ be a metric space, where S is a set, d is a metric on S and $\mathcal{E}$ is the Borel σ-field generated from all d-open subsets of S. Any map M: $\Omega \to S$ is said to be a *random element* taking values in S if $M^{-1}(E) \in \mathcal{A}$ for every $E \in \mathcal{E}$. In short, M is called a random element in S.

2.1.1 Definition of weak convergence in metric spaces

A functional on $(S, \mathcal{E}, d)$ is a mapping g from S to the real line R. R is always endowed with Euclidean metric d_e, the absolute value, and $\mathcal{B}$ is the Borel σ-field on the real line. Then g is $\mathcal{E}$-measurable if $g^{-1}(B) \in \mathcal{E}$ for every $B \in \mathcal{B}$; g is bounded if there exists an $L > 0$ such that $|g(s)| \leq L$ for all $s \in S$; g is d-continuous if $s_n, s \in S$ and $d(s_n, s) \to 0$ imply $|g(s_n) - g(s)| \to 0$, as $n \to \infty$.

Given a sequence of random elements Y_n, $n \geq 0$, in S, one can construct a sequence of probability measures P_n on $(S, \mathcal{E})$ through defining for each n and any $E \in \mathcal{E}$

$$P_n(E) = P(Y_n^{-1}(E)).$$

Here P_n is called the *probability measure induced by* Y_n.

Definition 1 (Weak convergence) *A sequence of random elements* Y_n, $n \geq 1$, *in* S *is said to* converge weakly *to a random element* Y_0 *in* S *if for every bounded, d-continuous,* $\mathcal{E}$*-measurable functional* g *on* S,

$$\int_S g\, dP_n = E\{g(Y_n)\} \longrightarrow E\{g(Y_0)\} = \int_S g\, dP_0$$

as $n \to \infty$. *This type of convergence is denoted* $Y_n \Rightarrow Y_0$ *or* $P_n \Rightarrow P_0$, *as* $n \to \infty$, *and* Y_0 *is also called a weak limit of* Y_n.

For properties and criteria of weak convergence in general metric spaces, one can consult Billingsley (1968), Bergström (1982) and Pollard (1984). In particular, when $(S, \mathcal{E}, d) = (R, \mathcal{B}, d_e)$, where d_e is the absolute value metric on the real line R, the concept of weak convergence coincides with convergence in distribution of real random variables; the latter is denoted, say, $Y_n \to_d Y_0$ for random variables Y_n and Y_0.

Closely related to the concept of weak convergence are the concepts of relative compactness and tightness.

Definition 2 (Relative compactness, Tightness) *A family* Π *of probability measures on* S *is said to be relatively compact, if each sequence of members of* Π *contains a sub-sequence which converges weakly to a probability measure on* S; Π *is said to be tight if for every* $\epsilon > 0$, *there exists a compact subset* K_ϵ *of* S *such that* $\pi(K_\epsilon) > 1 - \epsilon$ *for all* $\pi \in \Pi$.

These two concepts are related by the following result due to Prohorov (1956), which makes it easier to apply the theory of weak convergence:

- For general metric spaces, if Π is tight, then Π is relatively compact;
- For separable metric spaces, if Π is relatively compact, then Π is tight.

2.1.2 Convergence in probability and product spaces

Many concepts about univariate and multivariate random variables can be generalized to the context of random elements in general metric spaces.

For a sequence of random elements Y_n in S and an element $s \in S$, if for each $\epsilon > 0$,

$$P\{ d(Y_n, s) \geq \epsilon \} \longrightarrow 0$$

as $n \to \infty$, then Y_n is said to *converge in probability* to s, denoted by $Y_n \to_p s$. Here, s can be regarded as a constant-valued random element.

Let X, X_n be random elements in metric space S_1, let Y, Y_n be random elements in metric space S_2. If both S_1 and S_2 are separable, then the product space $S^* = S_1 \times S_2$ is separable and (X, Y) and (X_n, Y_n) are random elements in S^*. The following list of facts will be needed later, and their proofs can be found in Billingsley (1968).

Fact 1 $Y_n \to_p s$ if and only if $Y_n \Rightarrow s$.

Fact 2 If $S = S_1 = S_2$ is separable, then $Y_n \Rightarrow Y$ and $d(Y_n, X_n) \to_p 0$ imply $X_n \Rightarrow Y$.

Fact 3 If S_1 and S_2 are separable, then $X_n \Rightarrow X$ and $Y_n \to_p s$ imply $(X_n, Y_n) \Rightarrow (X, s)$, where $s \in S_2$.

Fact 4 If S_1 and S_2 are general metric spaces and $S^* = S_1 \times S_2$, then probability measures on S^* are tight if and only if the two sets of marginal probability measures are tight on S_1 and S_2, respectively.

2.1.3 Space $C[0,1]$ and space $D[0,1]$

For applications of weak convergence theory to the study of empirical distribution functions and empirical processes, the general metric space $(S, \mathcal{E}, d)$ needs to assume a special form.

Define $D[0,1]$ to be the collection of all real-valued functions $x\colon [0,1] \to R$ that are right-continuous and have left-hand limits. For example, the sample paths of the empirical

distribution function $U_n(t)$ of a sample from $U(0,1)$ belong to $D[0,1]$ for every n. $D[0,1]$ will be the basic space used in this thesis.

Also, let $C[0,1]$ be the collection of all real continuous functions defined on [0,1]. Clearly, $C[0,1] \subset D[0,1]$. When endowed with the uniform metric d_u defined for $x, y \in C[0,1]$ by

$$d_u(x,y) = \sup_{0 \le t \le 1} |x(t) - y(t)|,$$

$(C[0,1], d_u)$ becomes a complete separable metric space. Let $\mathcal{C}$ denote the Borel σ-field generated by all the d_u-open subsets of $C[0,1]$.

To introduce a suitable metric on $D[0,1]$, subtleties arise. The problem is essentially the problem of measurability when one wants to work with some functionals of empirical processes, such as Kolmogorov-Smirnov statistics. The situation can be worse. For example, if the uniform metric d_u is adopted for $D[0,1]$ and $\mathcal{E}$ is the Borel σ-field generated by all the d_u-open subsets of $D[0,1]$, then even the empirical process from a $U(0,1)$ sample is not a random element in $D[0,1]$ if one is willing to accept the axiom of choice of set theory (Pollard, (1984)), therefore, there is no way to study any distributional property. With the uniform metric d_u, the problem is that $\mathcal{E}$ contains too many sets.

There are different ways to overcome the difficulty associated with the measurability problem. The way presented below comes from Billingsley (1968). Other approaches can be found in Pollard (1984), Shorack and Wellner (1986) and references therein.

Define, for $x, y \in D[0,1]$,

$$d_s(x,y) = \inf_{\lambda \in \Lambda} \max\left\{\sup_{0 \le t \le 1} |x(t) - y(\lambda(t))|, \sup_{0 \le t \le 1} |t - \lambda(t)|\right\},$$

where Λ consists of all strictly increasing continuous functions from [0,1] onto [0,1]. That d_s is a metric is a standard result now, due to Skorohod (1956). Thus d_s is called the *Skorohod metric.* The Borel σ-field generated by all the d_s-open subsets of $D[0,1]$ is denoted by $\mathcal{D}$.

Equipped with d_s, $D[0,1]$ becomes a complete separable metric space, too[1]. Completeness of $(D[0,1], d_s)$ is not important if the existence of a limiting process is known, but

[1] $(D[0,1], d_u)$ is complete but not separable.

separability is vital to ensure measurability for many interesting applications.

For a sequence of random elements Y_n, $n \geq 0$, in $D[0,1]$, the following sufficient conditions for weak convergence and tightness in $D[0,1]$ are very useful.

- **Weak Convergence**: If $Y_n \rightarrow_{f.d.} Y_0$ (Y_n converges in finite distributions to Y_0) and Y_n, $n \geq 1$, is tight, then $Y_n \Rightarrow Y_0$;

- **Tightness**: For $\delta > 0$, define

$$W_{Y_n}(\delta) = \sup_{|s-t|<\delta} |Y_n(s) - Y_n(t)|.$$

If (i) for any $\eta > 0$, there exists an $a > 0$ such that for all $n \geq 1$,

$$P\{|Y_n(0)| > a\} \leq \eta,$$

and (ii) for any $\eta > 0$ and any $\gamma > 0$, there exist $\delta \in (0,1)$ and $n_0 = n_0(\delta, \eta, \gamma)$ such that for all $n \geq n_0$,

$$P\{W_{Y_n}(\delta) \geq \gamma\} \leq \eta,$$

then Y_n, $n \geq 1$, is tight. See Billingsley (1968).

The above sufficient conditions also apply to random elements in $C[0,1]$. Moreover, between space $(C[0,1], d_u)$ and space $(D[0,1], d_s)$, there is the relationship that for any x, y in $D[0,1]$, $d_u(x,y) \geq d_s(x,y)$ and $\mathcal{C} = C[0,1] \cap \mathcal{D}$, that is, on $C[0,1]$ d_u and d_s determine the same topology. Note also that weak convergence depends only on the topology on S, that is, if two different metrics lead to the same topology on S, the notion of weak convergence on S will be the same regardless of which metric is actually used.

2.1.4 Gaussian process and the stochastic process approach

The limiting processes which will be derived in this thesis are known as *Gaussian processes* in general. Formally, a Gaussian process with index set T is a collection of real-valued random variables $\{Y_t \mid t \in T\}$ such that for every $k \geq 1$ and all indices $t_1, t_2, \ldots, t_k$ in

T, $(Y_{t_1}, \ldots, Y_{t_k})$ is a k-variate normal random vector. Knowing the mean and covariance structure is equivalent to knowing the distributional properties if the process involved is Gaussian.

The Brownian bridge is a continuous sample path Gaussian process indexed by $T = [0,1]$ with zero mean and covariance function

$$\rho_0(s,t) = \min(s,t) - st, \qquad s,t, \in [0,1].$$

Now one can see the starting point of the stochastic process approach more clearly. For a sample $\{u_i\}_{i=1}^n$ from $U(0,1)$, that is, $u_i : \Omega \to [0,1]$ is a random variable, $i = 1,2,\ldots,n$, the empirical distribution function is, strictly speaking, a function of ω and t, i.e., $U_n(\omega,t)$, where $\omega \in \Omega$ and $t \in [0,1]$. One can view this as a stochastic process indexed by $T = [0,1]$ if the "evolution" of the process is of interest. On the other hand, for goodness-of-fit type problems, the behaviour of $U_n(\omega,t)$ over the entire range of t needs to be considered at the same time. When this is the case, for each ω it is better to think of $U_n(\omega,t)$ as a point in $D[0,1]$. This point of view is especially useful when one analyzes functionals defined on $D[0,1]$ involving $U_n(\omega,t)$, because the analysis can then be done in two stages: the analysis of $U_n(\omega,t)$ and the analysis of the functionals. If one can prove weak convergence of $U_n(\omega,t)$, one is ready to analyze any continuous functionals of $U_n(\omega,t)$ in the same manner.

2.2 Some Useful Results for Least Squares Estimators in Linear Regression

Linear regression, assuming a normal error distribution, is one of the most widely used statistical techniques. For this and because of this, methods for fitting models, estimating parameters, checking model adequacies and predicting future responses have been studied extensively.

ASSUMPTIONS	VIOLATIONS
• Linearity and additive errors $y_i = x_i^t\theta + \varepsilon_i$, where $\theta = (\theta_1, \ldots, \theta_p)^t$ is the unknown regression parameter.	Non-linear dependence on θ and/or non-additive errors.
• Independence of errors ε_i are mutually independent.	Correlated errors.
• Constant error variance $V(\varepsilon_i) \equiv \sigma^2$.	Heterogeneous variances.
• Normal error distribution each ε_i is normally distributed.	Skewed and/or long/short tailed error distribution.
• Error-free covariates x_i^t are measured without error.	Random covariates.
• Uniqueness of $\hat{\theta}$ rank$(X) = p$, full rank, $X = (X_1, \ldots, X_p)$.	Collinearity among X_j's, where X_j's are the columns of X.

Table 2.1: Assumptions of normal theory linear regression and possible violations.

2.2.1 Assumptions of linear regression

As with many formalized statistical techniques, linear regression is based on a number of assumptions, each of which has its evil competitor. In the following, these assumptions are listed in such a way that allows one to see the kinds of violations that may occur.

Let y_i be the i^{th} response associated with covariates $x_i^t = (x_{i1}, \ldots, x_{ip})$ (a row vector, here t denotes transpose) and error ε_i. Let $X = (X_1, X_2, \ldots, X_p)$ be the model matrix, where X_j is the j^{th} column of X. Table 2.1 lists six assumptions for a standard linear regression model and some possible violations.

2.2.2 Properties of least squares estimators in linear regression

Suppose the six assumptions in Table 2.1 are met. Using matrix notations, a linear regression model can be written as

$$Y = X\theta + \sigma\varepsilon, \tag{2.2.1}$$

where Y is an $n \times 1$ vector of responses; X is an $n \times p$ matrix of covariates with known values and $\text{rank}(X) = p$; θ is a $p \times 1$ vector of unknown regression parameters; σ is the unknown standard deviation of the error distribution; and $\varepsilon \sim N_n(0, I)$, where I is the $n \times n$ identity matrix.

By minimizing the residual sum of squares $SSR(\theta) = (Y - X\theta)^t(Y - X\theta)$ with respect to θ, one obtains the least squares estimator $\hat{\theta}$ of θ, predicted values $\hat{Y}$, ordinary residuals R and an estimator $\hat{\sigma}^2$ of σ^2. Some useful properties of these estimators and a number of other quantities are listed below for future reference.

1. **Properties of $\hat{\theta}$:**

   1.1 $\hat{\theta} = (X^tX)^{-1}X^tY$.

   1.2 $\text{E}(\hat{\theta}) = \theta$.

   1.3 $\hat{\theta}$ is BLUE, i.e., among the class of all linear unbiased estimators of θ, $\hat{\theta}$ has the smallest variance.

   1.4 $\text{V}(\hat{\theta}) = \sigma^2(X^tX)^{-1}$.

   1.5 $\hat{\theta} - \theta \sim N_p(0, \sigma^2(X^tX)^{-1})$.

2. **Properties of $\hat{Y}$:**

   2.1 $\hat{Y} = X\hat{\theta} = HY$, where $H = X(X^tX)^{-1}X^t$ is called the hat matrix or the projection matrix.

   2.2 $\text{E}(\hat{Y}) = X\theta$.

   2.3 $\text{V}(\hat{Y}) = \sigma^2 H$.

2.4 $\hat{Y} \sim N_n(X\theta, \sigma^2 H)$.

3. **Properties of** R:

3.1 $R = Y - \hat{Y} = (I - H)Y = \sigma(I - H)\varepsilon$.

3.2 $\mathrm{E}(R) = 0$.

3.3 $\mathrm{V}(R) = \sigma^2(I - H)$.

3.4 $R \sim N_n(0, \sigma^2(I - H))$.

3.5 $R^t R/\sigma^2 \sim \chi^2_{n-p}$, where χ^2_{n-p} is a χ^2 random variable on $n-p$ degrees of freedom.

4. **Properties of** $\hat{\sigma}^2$:

4.1 $\hat{\sigma}^2 = R^t R/(n-p) = Y^t(I - H)Y/(n-p) = \sigma^2 \varepsilon^t(I - H)\varepsilon/(n-p)$.

4.2 $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$.

4.3 $\hat{\sigma} \to_p \sigma$, as $\nu = (n-p) \to \infty$.

5. **Properties of** H **and** $I - H$:

5.1 H and $I - H$ are symmetric and idempotent, in particular, $h_{ii} = \sum_{j=1}^n h_{ij}^2$, for $i = 1, 2, \cdots, n$.

5.2 $\mathrm{trace}(H) = p$, $\mathrm{trace}(I - H) = n - p$.

5.3 $0 \le h_{ii} \le 1$ for all i, $-0.5 \le h_{ij} \le 0.5$ for all $i \ne j$.

6. **Independence Lemma**: Let $\hat{\theta}$ and R be defined as above and define the *standardized residuals* r by

$$r = \frac{R}{\hat{\sigma}} = \frac{Y - \hat{Y}}{\hat{\sigma}}.$$

Then the triple $(\hat{\theta}, \hat{\sigma}, r)$ has three mutually independent components.

Proof. It suffices to show that r is independent of $(\hat{\theta}, \hat{\sigma})$, as the other cases are standard results which can be found in Seber (1977) and Chatterijee and Hadi (1988).

It is tricky to prove (Arnold (1981), page 64), but it is true, that $(\hat{\theta}, \hat{\sigma})$ is jointly complete and sufficient for (θ, σ). Note that

$$r = \frac{R}{\hat{\sigma}} = \frac{\sigma(I-H)\varepsilon}{\sigma\sqrt{\varepsilon^t(I-H)\varepsilon/(n-p)}}$$

has a distribution free of σ. By Basu's theorem, r is independent of $(\hat{\theta}, \hat{\sigma})$. □

2.3 Residual Empirical Process: Two Theorems for Handling the Effect of Estimating Standard Deviation σ

When responses $Y = (y_1, y_2, \ldots, y_n)^t$ satisfy the linear regression model defined in section 2.2.2, namely,

$$Y = X\theta + \sigma\varepsilon, \qquad \varepsilon \sim N_n(0, I), \tag{2.3.1}$$

or in its component form

$$y_i = x_i^t\theta + \sigma\varepsilon_i, \quad \varepsilon_i \text{ are independent } N(0,1), \tag{2.3.2}$$

both θ and σ need to be estimated in most real situations. It turns out that the contribution from estimating θ is different from that of estimating σ, and these two sources of contributions can be clearly separated—a miracle that comes from the normality assumption as will be seen. This section is devoted to the study of the effect of estimating σ.

2.3.1 Definitions of residual empirical processes $Y_n(t)$ and $\tilde{Y}_n(t)$

The notations introduced in section 2.2.2 are also used in this section. In particular, $R = Y - \hat{Y}$, and $r = R/\hat{\sigma}$, where $\hat{\sigma}^2 = R^tR/(n-p)$. The component form of R is

$$R_i = y_i - \hat{y}_i = y_i - x_i^t\hat{\theta},$$

while the component form of r is

$$r_i = \frac{y_i - x_i^t\hat{\theta}}{\hat{\sigma}}.$$

Furthermore, define

$$I[a \le b] = \begin{cases} 1 & \text{if } a \le b \\ 0 & \text{if } a > b. \end{cases}$$

Let $\Phi(x) = \int_{-\infty}^{x} \phi(y)\,dy$, where $\phi(y) = (2\pi)^{-1/2}\exp\{-\frac{1}{2}y^2\}$, $-\infty < y < +\infty$. For $t \in [0,1]$, let $J_1(t) = \phi(\Phi^{-1}(t))$, $J_2(t) = J_1(t)\Phi^{-1}(t)$.

Definition 3 (Residual empirical process) *In linear regression (2.3.1), let $e_i = \Phi(R_i)$ and $\tilde{e}_i = \Phi(r_i)$, $i = 1, 2, \ldots, n$. Then for $t \in [0,1]$ define*

$$Y_n(t) = n^{-1/2}\sum_{i=1}^{n}\{I[e_i \le t] - t\}, \qquad \textit{using } e_i\textit{'s}, \tag{2.3.3}$$

$$\tilde{Y}_n(t) = n^{-1/2}\sum_{i=1}^{n}\{I[\tilde{e}_i \le t] - t\}, \qquad \textit{using } \tilde{e}_i\textit{'s}, \tag{2.3.4}$$

as empirical processes based on residuals, or simply, residual empirical processes, for the sample $y_1, y_2, \ldots, y_n$.

2.3.2 The first decomposition of residual empirical process: $\tilde{Y}_n(t) = (Y_n \circ H_n)(t) + Z_n(t)$

Observe that

$$\begin{aligned} \forall\, i \qquad & \tilde{e}_i \le t \\ \text{iff} \quad & y_i - x_i^t\hat{\theta} \le \hat{\sigma}\Phi^{-1}(t) \\ \text{iff} \quad & e_i \le \Phi\left(\hat{\sigma}\Phi^{-1}(t)\right). \end{aligned}$$

Denoting $H_n(t) = \Phi(\hat{\sigma}\Phi^{-1}(t))$ and $Z_n(t) = \sqrt{n}(H_n(t) - t)$, $t \in [0,1]$, one arrives at the following decomposition of $\tilde{Y}_n(t)$:

$$\begin{aligned} \tilde{Y}_n(t) &= (Y_n \circ H_n)(t) + Z_n(t) \\ &= n^{-1/2}\sum_{i=1}^{n}\{I[e_i \le H_n(t)] - H_n(t)\} + Z_n(t). \end{aligned} \tag{2.3.5}$$

It can be seen that in the first component of the above decomposition, $Y_n(t)$ appears, which comes from estimating θ alone, but there is a random change of time involved; in

the second component, the only part which is random is $\hat{\sigma}$. This is exactly the result of estimating σ. The idea here is to find the limiting distributions of Y_n and Z_n first, then through random change of time, to obtain the limiting distribution of $\tilde{Y}_n$.

2.3.3 Study of $Z_n(t)$: $Z_n(t)$ converges weakly to a Gaussian Process

Taylor expansion method will be employed to find the weak limit of Z_n. To simplify notation, some notational conventions are introduced here and used throughout this thesis.

When a function f satisfies the conditions of Taylor expansion about a point a, one has

$$f(t) = f(a) + \sum_{k=1}^{n} \frac{1}{k!} f^{(k)}(a)(t-a)^k + \frac{1}{(n+1)!} f^{(n+1)}(\eta)(t-a)^{n+1},$$

where η is a number lying between t and a. In general, η changes when t and a change, but when the exact location of η is not important, as will be seen in many cases in the rest of this thesis, it is agreed to write at the same time, say,

$$f(t) = f(a) + f'(\eta)(t-a),$$

$$f(t) = f(a) + f'(a)(t-a) + \frac{1}{2} f''(\eta)(t-a)^2,$$

etc, using the same η; when dependence on t and a is important, it is agreed to write, say,

$$f(t) = f(a) + f'(a+\eta t), \text{ where } |\eta| \leq 1.$$

Lemma 2.3.1 *Let $\nu = n-p$ and $\tilde{Z}_n(t) = 2^{-1/2} J_2(t)(2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu)$, $t \in [0,1]$. Then*

(1) $(2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu) \to_d N(0,1)$ *as* $\nu \to \infty$,

(2) $\{\tilde{Z}_n(t)\}_{n=1}^{\infty}$ *is tight in both* $C[0,1]$ *and* $D[0,1]$.

Proof. Without loss of generality, assume $\sigma = 1$. If $X \sim \chi^2_\nu$, the characteristic function of X is

$$\varphi_X(t) = (1-2it)^{-\frac{\nu}{2}}.$$

Let $Y = (X - \nu)/\sqrt{2\nu}$, then the characteristic function of Y is

$$\varphi_Y(t) = e^{-it\sqrt{\nu/2}} \left(1 - it\sqrt{2/\nu}\right)^{-\nu/2}.$$

As $\nu \to \infty$, one can show that $\varphi_Y(t) \to e^{-t^2/2}$. By Lévy-Cramér theorem, $Y \to_d N(0,1)$. Since $\nu\hat{\sigma}^2 \sim \chi^2_\nu$, therefore, $(2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu) \to_d N(0,1)$.

According to section 2.1.3, the following two conditions are sufficient for $\tilde{Z}_n$ to be tight: (i) $\forall\ \eta > 0$, there exists an $a > 0$ such that for all $n \geq 1$,

$$P\left\{|\tilde{Z}_n(0)| > a\right\} \leq \eta,$$

(ii) $\forall\ \gamma > 0$, $\forall\ \eta > 0$, there exist $\delta \in (0,1)$ and $n_0 = n_0(\delta, \gamma, \eta)$ such that for all $n \geq n_0$,

$$P\left\{W_{\tilde{Z}_n}(\delta) \geq \gamma\right\} \leq \eta,$$

where $W_{\tilde{Z}_n}(\delta) = \sup_{|s-t|<\delta} |\tilde{Z}_n(s) - \tilde{Z}_n(t)|$.

Now $\tilde{Z}_n(0) = 0$ because $J_2(0) = 0$, so (i) is satisfied. To prove (ii), notice that

$$\begin{aligned} W_{\tilde{Z}_n}(\delta) &= \sup_{|s-t|<\delta} |\tilde{Z}_n(s) - \tilde{Z}_n(t)| \\ &= 2^{-1/2} |(2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu)| \sup_{|s-t|<\delta} |J_2(s) - J_2(t)|. \end{aligned}$$

Because $J_2(t)$ is continuous on [0,1] (Shorack and Wellner (1986), page 181), for any $\gamma > 0$ and $\eta > 0$, one can choose $\delta \in (0,1)$ such that $\sup_{|s-t|<\delta} |J_2(s) - J_2(t)| = b$, where b has the property that $2\{1 - \Phi(\sqrt{2}\gamma/b)\} \leq \eta/2$. Let $a_\nu = (2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu)$, since $a_\nu \to_d N(0,1)$, one can choose n_0 such that for all $n \geq n_0$

$$P\left\{|(2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu)| \geq \frac{\sqrt{2}\gamma}{b}\right\} = P\{|a_\nu| \geq \frac{\sqrt{2}\gamma}{b}\} \leq \eta.$$

Therefore,

$$\begin{aligned} P\left\{W_{\tilde{Z}_n}(\delta) \geq \gamma\right\} &= P\left\{2^{-1/2}|(2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu)| \sup_{|s-t|<\delta} |J_2(s) - J_2(t)| \geq \gamma\right\} \\ &= P\left\{|(2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu)| \geq \sqrt{2}\gamma/b\right\} \\ &\leq \eta. \quad \square \end{aligned}$$

Lemma 2.3.2 *For any fixed* $\sigma > 0$,

(1) $\phi(\sigma\Phi^{-1}(t))\Phi^{-1}(t)$ *and* $\phi'(\sigma\Phi^{-1}(t))[\Phi^{-1}(t)]^2$ *are continuous functions of* t *on* $[0,1]$,

(2) *Let* $A(\eta,t) = \phi(\eta)\Phi^{-1}(t)$, *and* $B(\eta,t) = \phi'(\eta)[\Phi^{-1}(t)]^2$, *where* η *is random and is lying between* $\hat{\sigma}\Phi^{-1}(t)$ *and* $\Phi^{-1}(t)$. *If* $x_\nu \to_p 0$, *then* $x_\nu A(\eta,t) = x_\nu B(\eta,t) = o_p(1)$.

Proof. To prove $\phi(\sigma\Phi^{-1}(t))\Phi^{-1}(t)$ is continuous on [0,1], it suffices to prove that the right hand limit at 0 and the left hand limit at 1 exist. Let $x = \Phi^{-1}(t)$, then

$$\lim_{t\to 0^+} \phi(\sigma\Phi^{-1}(t))\Phi^{-1}(t) = \lim_{x\to-\infty} \phi(\sigma x)x = 0.$$

Similarly,

$$\lim_{t\to 1^-} \phi(\sigma\Phi^{-1}(t))\Phi^{-1}(t) = \lim_{x\to+\infty} \phi(\sigma x)x = 0.$$

Similar arguments show that $\phi(\sigma\Phi^{-1}(t))[\Phi^{-1}(t)]^2$ is continuous on [0,1].

Since $\hat{\sigma} \to_p \sigma$, for any $\delta, \tau > 0$, there exists a ν_0 such that for $\nu > \nu_0$,

$$\begin{aligned} & P\{d_s(x_\nu A(\eta,t),0) > \delta\} \\ \le\ & P\{\sup_{0\le t\le 1} |x_\nu A(\eta,t)| > \delta\} \\ \le\ & P\{\sup_{0\le t\le 1} |x_\nu A(\eta,t)| > \delta,\ |\hat{\sigma}-\sigma| \le \delta\} + \tau/2. \end{aligned}$$

Without loss of generality, suppose $\sigma - \delta > 0$. Then from (1), there is a constant $M(\sigma) > 0$ such that

$$\begin{aligned} & P\{\sup_{0\le t\le 1} |x_\nu A(\eta,t)| > \delta,\ |\hat{\sigma}-\sigma| \le \delta\} \\ \le\ & P\{|x_\nu| > \delta/M(\sigma)\}. \end{aligned}$$

Because $x_\nu \to_p 0$, one can find an $m \ge \nu_0$ such that for all $\nu \ge m$, $P\{|x_\nu| > \delta/M(\sigma)\} < \tau/2$. Then for $\nu \ge m$,

$$P\{\sup_{0\le t\le 1} |x_\nu A(\eta,t)| > \delta\} < \tau,$$

that is, for any $\delta > 0$, $P\{\sup_{0\le t\le 1} |x_\nu A(\eta,t)| > \delta\} \to 0$, as $\nu = n - p \to \infty$. This proves $x_\nu A(\eta,t) = o_p(1)$. Similarly, one can show that $x_\nu B(\eta,t) = o_p(1)$. □

Theorem 2.3.1 (The effect of estimating σ) *In linear regression model (2.3.1), let $Z_n(t) = \sqrt{n}(H_n(t) - t)$, where $H_n(t) = \Phi(\hat{\sigma}\Phi^{-1}(t))$. If $d = \lim_{n\to\infty} n/\nu$ exists, where $\nu = n - p$ and $1 \leq d < \infty$, then $Z_n(t) \Rightarrow Z(t)$, a Gaussian process with zero mean and covariance function*

$$\rho_2(s,t) = Cov(Z(s), Z(t)) = \frac{d}{2} J_2(s) J_2(t), \; s, t \in [0,1]. \tag{2.3.6}$$

Proof. Without loss of generality, suppose that the true value of σ is 1.

Step 1. $Z_n(t) = \sqrt{d}\tilde{Z}_n(t) + o_p(1)$, where $\tilde{Z}_n(t) = 2^{-1/2} J_2(t)(2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu)$.

This expansion comes from expanding $H_n(t)$ about $\Phi^{-1}(t)$, which gives

$$\begin{aligned} H_n(t) - t &= \Phi(\hat{\sigma}\Phi^{-1}(t)) - t \\ &= \Phi^{-1}(t)\phi(\Phi^{-1}(t))(\hat{\sigma} - 1) + \tfrac{1}{2}\phi'(\eta)(\Phi^{-1}(t))^2(\hat{\sigma} - 1)^2, \end{aligned}$$

where η lies between $\Phi^{-1}(t)$ and $\hat{\sigma}\Phi^{-1}(t)$. Note that

$$\sqrt{\nu}(\hat{\sigma}^2 - 1) = \sqrt{2}\,(2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu),$$

$$\sqrt{\nu}(\hat{\sigma} - 1)^2 = (\hat{\sigma} - 1)\frac{\sqrt{2}}{\hat{\sigma} + 1}(2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu).$$

Because $\hat{\sigma} - 1 \to_p 0$ as $\nu \to \infty$, $(2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu) \to_d N(0,1)$ from (2) of Lemma 2.3.1, by Slutsky's theorem, $\sqrt{\nu}(\hat{\sigma} - 1)^2 = o_p(1)$, hence, by Lemma 2.3.2,

$$\sqrt{\nu}\phi'(\eta)(\Phi^{-1}(t))^2(\hat{\sigma} - 1)^2 = o_p(1).$$

Thus

$$\begin{aligned} Z_n(t) &= \sqrt{n}(H_n(t) - t) \\ &= \sqrt{n/\nu}\sqrt{\nu}(H_n(t) - t) \\ &= \sqrt{n/\nu}\sqrt{\nu}J_2(t)(\hat{\sigma} - 1) + \sqrt{n/\nu}\,o_p(1). \end{aligned}$$

Rewriting $\sqrt{\nu}(\hat{\sigma} - 1)$ as $\sqrt{2}(\hat{\sigma} + 1)^{-1}(2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu)$ and using the fact that $\hat{\sigma} \to_p 1$ and $J_2(t)$ is continuous on [0,1], one has, by Slutsky' theorem again,

$$J_2(t)\left(\frac{1}{\sqrt{2}} - \frac{\sqrt{2}}{\hat{\sigma} + 1}\right)(2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu) = o_p(1),$$

therefore,

$$\begin{aligned} Z_n(t) &= \sqrt{n/\nu}\,2^{-1/2}J_2(t)(2\nu)^{-1/2}(\nu\hat{\sigma}^2-\nu)+\sqrt{n/\nu}\,o_p(1) \\ &= \sqrt{n/\nu}\,\tilde{Z}_n(t)+\sqrt{n/\nu}\,o_p(1) \\ &= \sqrt{d}\tilde{Z}_n(t)+o_p(1). \end{aligned} \tag{2.3.7}$$

Step 2. $\tilde{Z}_n(t) \to_{f.d.}$ normal distributions. For any $k \geq 1$ and any $\{t_i\}_{i=1}^k$ such that $0 \leq t_1 \leq t_2 \leq, \ldots, t_k \leq 1$, consider the random k-vector $W_n = (\tilde{Z}_n(t_1), \ldots, \tilde{Z}_n(t_k))^t$. Clearly, $E(W_n) = 0$, because $E((2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu)) = 0$. Let $a_\nu = (2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu)$, define $\Gamma^n = (\Gamma^n_{jl}) = V(W_n) = E(W_n^t W_n)$, where

$$\Gamma^n_{jl} = E(\tilde{Z}_n(t_j)\tilde{Z}_n(t_l)),\ j, l = 1, 2, \ldots, k.$$

Since $E(a_\nu) = 0$ and $V(a_\nu) = 1$ for any ν, $\Gamma^n_{jl} = \frac{1}{2}J_2(t_j)J_2(t_l)$ is independent of n.

Let $< u, v >= \sum_{i=1}^k u_i v_i$ denote the usual inner product in Euclidean space R^k. Then for $u \in R^k$,

$$\begin{aligned} -\frac{1}{2} < \Gamma^n u, u > &= -\frac{1}{2}\left(\frac{1}{2}\sum_{l=1}^k u_l \sum_{j=1}^k u_j J_2(t_j)J_2(t_l)\right) \\ &= -\frac{1}{4}\sum_{j=1}^k\sum_{l=1}^k u_j u_l J_2(t_j)J_2(t_l). \end{aligned} \tag{2.3.8}$$

On the other hand, the characteristic function of W_n is

$$\begin{aligned} \varphi_{W_n}(u) &= E\{\exp(i < W_n, u >)\} \\ &= E\left\{\exp\left(i\sum_{j=1}^k 2^{-1/2}u_j J_2(t_j)\,a_\nu\right)\right\} \\ &\to \exp\left\{-\frac{1}{2}\left(\sum_{j=1}^k 2^{-1/2}u_j J_2(t_j)\right)^2\right\} \\ &= \exp\left\{-\frac{1}{4}\sum_{j=1}^k\sum_{l=1}^k u_j u_l J_2(t_j)J_2(t_l)\right\}. \end{aligned} \tag{2.3.9}$$

Comparing (2.3.8) with (2.3.9) gives

$$W_n \to_{f.d.} N_k(0, \Gamma^n) \text{ as } \nu \to \infty.$$

Step 3. From Lemma 2.3.1 (2) and step 2, $\tilde{Z}_n(t)$ is tight and converges to normal laws in finite distributions, thus $\tilde{Z}_n(t)$ converges weakly to some Gaussian process, say, $Z(t)$. The above calculations show that $Z(t)$ has zero mean and covariance function

$$\text{Cov}(Z(s), Z(t)) = \frac{1}{2} J_2(s) J_2(t).$$

Since $Z_n(t) = \sqrt{d}\, \tilde{Z}_n(t) + o_p(1)$, where $1 \le d = \lim_{n\to\infty} n/\nu < \infty$, by fact 2 of section 2.1.2, $Z_n(t) \Rightarrow \sqrt{d} Z(t)$, which is a Gaussian process with zero mean and covariance function

$$\rho_2(s,t) = \frac{d}{2} J_2(s) J_2(t). \quad \square$$

2.3.4 Random change of time

Let $D_0[0,1] = \{x : x \in D[0,1], \text{ x is nondecreasing, } 0 \le x(t) \le 1\}$. Let $\mathcal{D}_0 = D_0[0,1] \cap \mathcal{D}$. For $x \in D[0,1]$ and $g \in D_0[0,1]$, define $Q : D[0,1] \times D_0[0,1] \to D[0,1]$ by

$$Q(x,g)(t) = (x \circ g)(t) = x(g(t)),\ t \in [0,1].$$

Then Q is measurable, see Billingsley (1968), page 232.

For random elements X_n, X in $D[0,1]$ and random elements G_n, G in $D_0[0,1]$, $X_n \circ G_n$ is a random element in $D[0,1]$, resulted from subjecting X_n to the random "time" change represented by G_n. From Billingsley (1968), pages 144-145, if

$$(X_n, G_n) \Rightarrow (X, G),$$

$$P\{X \in C[0,1]\} = P\{G \in C[0,1]\} = 1,$$

then $X_n \circ G_n \Rightarrow X \circ G$ in $D[0,1]$.

2.3.5 Separating the effect of estimating standard deviation σ and regression parameter θ

The residual empirical process $\tilde{Y}_n(t)$, constructed essentially from the standardized residuals after estimating regression parameter θ and standard deviation σ, was decomposed into

$$\tilde{Y}_n(t) = (Y_n \circ H_n)(t) + Z_n(t)$$

in section 2.3.2 and the weak limit of the second term $Z_n(t)$ of this decomposition has been found in section 2.3.3.

There are two more things to be noticed here. Firstly, $H_n(t)$ is a random element in $D_0[0,1]$; secondly, $Z_n(t)$ is independent of $\tilde{Y}_n(t)$ for any t, because Z_n depends on $\hat{\sigma}$, $\tilde{Y}_n$ depends on r_i, and $\hat{\sigma}$ is independent of r_i according to the independence lemma of section 2.2.2.

Moreover, $H_n(t) \to_p T$, where T is the identity map on [0,1]. To see this, expand $H_n(t)$ about $\Phi^{-1}(t)$ again, one has

$$\begin{aligned} H_n(t) - t &= \Phi(\hat{\sigma}\Phi^{-1}(t)) - t \\ &= \Phi(\Phi^{-1}(t)) + \Phi^{-1}(t)(\hat{\sigma} - 1)\phi(\eta) - t \\ &= \Phi^{-1}(t)\phi(\eta)(\hat{\sigma} - 1), \end{aligned}$$

where η lies between $\Phi^{-1}(t)$ and $\hat{\sigma}\Phi^{-1}(t)$. Since $(\hat{\sigma} - 1) \to_p 0$, as $\nu = n - p \to \infty$, by Lemma 2.3.2, $H_n(t) - t = o_p(1)$.

It becomes clear now that if Y_n converges to some Gaussian process Y, one has a good chance of finding the weak limit of $\tilde{Y}_n$. The following theorem formalizes this idea.

Theorem 2.3.2 (Random change of time) *In linear regression model (2.3.1), let $Y_n(t)$ and $\tilde{Y}_n(t)$ be given as in Definition 3. If $Y_n \Rightarrow Y$, where Y is a Gaussian process with zero mean and continuous covariance function $\rho_m(s,t)$, and if $1 \leq d = \lim_{n\to\infty} n/(n-p) < \infty$, then $\tilde{Y}_n \Rightarrow \tilde{Y}$, where $\tilde{Y}$ is also a Gaussian process with zero mean and covariance*

function

$$\rho(s,t) = \rho_m(s,t) - \rho_2(s,t), \tag{2.3.10}$$

where $\rho_2(s,t)$ *is given in Theorem 2.3.1 as* $(d/2)J_2(s)J_2(t)$.

Proof. Since $H_n \to T$ and by assumption $Y_n \Rightarrow Y$, one has[2] by fact 3 of section 2.2.2 and random change of time that $Y_n \circ H_n \Rightarrow Y \circ T = Y$. That $\tilde{Y}_n$ is tight comes from the fact that both $Y_n \circ H_n$ and Z_n are tight and $\tilde{Y}_n = Y_n \circ H_n + Z_n$. That $\tilde{Y}_n$ converges in finite distributions to a Gaussian process, say, $\tilde{Y}$, is the consequence of $Y_n \circ H_n \to_{f.d.} Y$ and $Z_n \to_{f.d.} Z$ and Z_n is independent of $\tilde{Y}_n$. Together, one has $\tilde{Y}_n \Rightarrow \tilde{Y}$, $\tilde{Y}$ is independent of Z and $\tilde{Y} = Y + Z$. Finally, for all $s,t \in [0,1]$,

$$E(\tilde{Y}) = E(Y) + E(Z) = 0,$$

$$\begin{aligned} \mathrm{Cov}\{(Y(s),Y(t)\} &= \mathrm{Cov}\{(\tilde{Y} - Z)(s), (\tilde{Y} - Z)(t)\} \\ &= \mathrm{Cov}\{\tilde{Y}(s), \tilde{Y}(t)\} + \mathrm{Cov}\{Z(s), Z(t)\}, \end{aligned}$$

by the independence of $\tilde{Y}$ and Z. The above equation can be rewritten into $\rho(s,t) = \rho_m(s,t) - \rho_2(s,t)$ as desired. □

Remark 1. If let $\hat{Y}_n(t) = Y_n(t) + Z_n(t)$, one also has $\hat{Y}_n \Rightarrow \tilde{Y}$, where $\tilde{Y}$ is the weak limit of $\tilde{Y}_n$. So Theorem 2.3.2 has justified a partial random change of time, that is, change the time t in Y_n only.

Remark 2. Theorem 2.3.2 reduces the problem of finding the weak limit of $\tilde{Y}_n$ into the problem of finding the weak limit of Y_n. This reduction greatly simplifies the study of $\tilde{Y}_n$.

Remark 3. The restriction on the dimensionality of the regression parameter θ is mild in Theorems 2.3.1 and 2.3.2. In particular, the case where p depends on n and increases to infinity with n is allowed as long as $\lim_{n\to} n/(n-p) < \infty$.

[2]Gaussian processes with continuous covariance functions have continuous paths with probability one. See Gikhman and Skorohod (1965).

2.4 Residual Empirical Process: Linear Regression With Fixed Number of Regression Parameters

With Theorem 2.3.2 proven, the limiting distribution of $Y_n(t)$ is what is needed to complete the search for limiting distribution of $\tilde{Y}_n(t)$. Recall that

$$Y_n(t) = n^{-1/2} \sum_{i=1}^{n} \{I[e_i \leq t] - t\},$$

where $e_i = \Phi(y_i - x_i^t\hat{\theta}), i = 1, 2, \ldots, n$, $\hat{\theta}$ is the least squares estimator of θ in linear regression (2.3.2), namely,

$$y_i = x_i^t\theta + \sigma\varepsilon_i, \quad \varepsilon_i\text{'s are independent } N(0,1), \tag{2.4.1}$$

where x_i^t is the i^{th} row of the model matrix $X = (x_1, x_2, \ldots, x_p)$, and x_j is the j^{th} column of X. Assume in this section that the dimension of regression parameter θ is p and p is fixed finite. Assume also that the true value of σ is 1.

Some comments are due before proceeding to present the theorems of this section. Among the authors who made contributions under the title of this section, Mukantseva (1977) studied in detail the case of straight line regressions and indicated extensions to multiple regressions; Pierce and Kopecky (1979) concentrated on showing that the limiting Gaussian process of the residual empirical process has the covariance structure that appears in the study of a location and scale problem; Loynes (1980) formulated the problem for independent, but not identically distributed random variables and obtained some general results; Koul (1984) allowed the error distribution in linear regression to be different from the normal and considered the weak convergence of weighted residual empirical processes; Shorack (1984) presented a unified approach to investigate empirical and rank processes with or without nuisance parameters. Although there are no totally new results in the theorems to be discussed below (except that Theorem 2.4.2 and part of Theorem 2.4.1 are not seen in the literature in the form given in this thesis), compared to the above cited sources, the approach taken here is more direct, the treatment of weak convergence problems is more

careful, and conditions of the theorems are somehow simpler. Of course, this is possible largely because the errors are supposed to be normally distributed.

2.4.1 The second decomposition of residual empirical process:

$$Y_n(t) = Y_{1n}(t) + Y_{2n}(t) + Y_{3n}(t)$$

Observe that

$$\begin{aligned} \forall\, i \qquad e_i &\le t \\ \Leftrightarrow \quad y_i - x_i^t\hat{\theta} &\le \Phi^{-1}(t) \\ \Leftrightarrow \quad y_i - x_i^t\theta &\le \Phi^{-1}(t) + x_i^t(\hat{\theta}-\theta) \\ \Leftrightarrow \quad u_i = \Phi(\varepsilon_i) &\le \Phi\left(\Phi^{-1}(t) + x_i^t(\hat{\theta}-\theta)\right), \end{aligned}$$

where u_i are independent $U(0,1)$ random variables. It is then possible to decompose $Y_n(t)$ into three parts as

$$\begin{aligned} Y_n(t) &= n^{-1/2}\textstyle\sum_{i=1}^n \{I[e_i \le t] - t\} \\ &= n^{-1/2}\textstyle\sum_{i=1}^n \{I[u_i \le \Phi(\Phi^{-1}(t) + x_i^t(\hat{\theta}-\theta))] - \Phi(\Phi^{-1}(t) + x_i^t(\hat{\theta}-\theta)) \\ &\quad -I[u_i \le t] + t\} \\ &+ n^{-1/2}\textstyle\sum_{i=1}^n \left\{\Phi(\Phi^{-1}(t) + x_i^t(\hat{\theta}-\theta)) - t\right\} \\ &+ n^{-1/2}\textstyle\sum_{i=1}^n \{I[u_i \le t] - t\} \\ &\equiv Y_{1n}(t) + Y_{2n}(t) + Y_{3n}(t), \text{ say.} \end{aligned}$$

In this decomposition, $Y_{3n}(t)$ converges weakly to a Brownian bridge as mentioned in chapter 1, namely, a Gaussian process with zero mean and covariance function $\rho_0(s,t) = \min(s,t) - st$. Notice that $Y_{2n}(t)$ is the result of randomly perturbing the time t by $x_i^t(\hat{\theta}-\theta)$, and $Y_{1n}(t)$ is the result of randomly perturbing the empirical process $n^{-1/2}\sum_{i=1}^n \{I[u_i \le t] - t\}$ of a random sample from $U(0,1)$. The plan of this section is to show that (1) $Y_{2n}(t)$ converges weakly to a Gaussian process, representing the effect of estimating θ; (2) $Y_{1n}(t) = o_p(1)$.

2.4.2 Study of $Y_{2n}(t)$: $Y_{2n}(t)$ converges weakly to a Gaussian process

Since $Y_{2n}(t) = n^{-1/2} \sum_{i=1}^{n} \left\{ \Phi(\Phi^{-1}(t) + x_i^t(\hat{\theta} - \theta)) - t \right\}$, expanding $\Phi(\Phi^{-1}(t) + x_i^t(\hat{\theta} - \theta))$ about $\Phi^{-1}(t)$ gives

$$\begin{aligned} Y_{2n}(t) &= n^{-1/2} \sum_{i=1}^{n} \left\{ \phi(\Phi^{-1}(t)) x_i^t(\hat{\theta} - \theta) + \tfrac{1}{2}\phi'(\eta_i) \left(x_i^t(\hat{\theta} - \theta) \right)^2 \right\} \\ &= n^{-1/2}\, \phi(\Phi^{-1}(t)) \sum_{i=1}^{n} x_i^t(\hat{\theta} - \theta) + n^{-1/2} \sum_{i=1}^{n} \tfrac{1}{2}\phi'(\eta_i) \left(x_i^t(\hat{\theta} - \theta) \right)^2, \end{aligned}$$

where η_i lies between $\Phi^{-1}(t)$ and $\Phi^{-1}(t) + x_i^t(\hat{\theta} - \theta)$,

Theorem 2.4.1 (The effect of estimating θ, p is fixed) *In linear regression model (2.4.1), let $H = (h_{ij}) = X(X^tX)^{-1}X^t$ be the projection matrix, let $m_j = \sum_{i=1}^{n} h_{ij}$, $j = 1, 2, \ldots, n$. Then*

(1) $Y_{2n}(t) = \tilde{Y}_{2n}(t) + o_p(1)$, where $\tilde{Y}_{2n}(t) = n^{-1/2}\, \phi(\Phi^{-1}(t)) \sum_{j=1}^{n} m_j \varepsilon_j$, and

(2) $\tilde{Y}_{2n}(t) \Rightarrow \tilde{Y}_2(t)$, provided $b = \lim_{n\to\infty} (\sum_{j=1}^{n} m_j^2)/n$ exists finite, where $\tilde{Y}_2$ is a Gaussian process with zero mean and covariance function

$$\rho_1(s,t) = Cov(\tilde{Y}_2(s), \tilde{Y}_2(t)) = b\, \phi(\Phi^{-1}(s))\, \phi(\Phi^{-1}(t)) = b J_1(s) J_1(t). \tag{2.4.2}$$

Proof. (1) Because $x_i^t(\hat{\theta} - \theta) = x_i^t(X^tX)^{-1}X^t\varepsilon$, $i = 1, \ldots, n$, where $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^t$ is a random sample from $N(0,1)$, the following relationships are observed according to the definitions of h_{ij} and m_j.

$$\begin{aligned} \sum_{i=1}^{n} x_i^t(\hat{\theta} - \theta) &= \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij}\varepsilon_j, \\ \sum_{j=1}^{n} \left(\sum_{i=1}^{n} h_{ij} \right) \varepsilon_j &= \sum_{j=1}^{n} m_j \varepsilon_j, \\ \sum_{i=1}^{n} \left(x_i^t(\hat{\theta} - \theta) \right)^2 &= \varepsilon^t H \varepsilon. \end{aligned}$$

Moreover, let $\sup_{x\in R}|\phi'(x)| = k < \infty$, then for any $\delta > 0$,

$$\begin{aligned}
& P\left\{d_s(n^{-1/2}\textstyle\sum_{i=1}^n \phi'(\eta_i)\left(x_i^t(\hat\theta-\theta)\right)^2, 0) > \delta\right\} \\
\le\ & P\left\{\sup_{0\le t\le 1}\left|n^{-1/2}\textstyle\sum_{i=1}^n \phi'(\eta_i)\left(x_i^t(\hat\theta-\theta)\right)^2\right| > \delta\right\} \\
\le\ & P\left\{\textstyle\sum_{i=1}^n \left(x_i^t(\hat\theta-\theta)\right)^2 > \sqrt{n}\delta k^{-1}\right\} \\
=\ & P\left\{\varepsilon^t H\varepsilon > \sqrt{n}\delta k^{-1}\right\} \\
\le\ & E(\varepsilon^t H\varepsilon)(\sqrt{n}\delta k^{-1})^{-1} \\
=\ & (kp)/(\delta\sqrt{n}) \\
\to\ & 0, \text{ as } n\to\infty.
\end{aligned}$$

Therefore, $Y_{2n}(t) = \tilde{Y}_{2n}(t) + o_p(1)$.

(2) First, $\tilde{Y}_{2n}(t)$ is tight in $C[0,1]$ and $D[0,1]$. This follows readily by checking the two conditions (i) and (ii) listed in section 2.1.3. That (i) is true comes from the fact that $J_1(0) = 0$, where $J_1(t) = \phi(\Phi^{-1}(t))$. Since for $\delta > 0$,

$$\begin{aligned}
W_{\tilde{Y}_{2n}}(\delta) &= \sup_{|s-t|<\delta}|\tilde{Y}_{2n}(s) - \tilde{Y}_{2n}(t)| \\
&= |n^{-1/2}\textstyle\sum_{j=1}^n m_j\varepsilon_j|\ \sup_{|s-t|<\delta}|J_1(s) - J_1(t)|,
\end{aligned}$$

then, $\forall\ \gamma > 0, \forall\ \eta > 0$,

$$\begin{aligned}
& P\left\{W_{\tilde{Y}_{2n}}(\delta) \ge \gamma\right\} \\
=\ & P\left\{|n^{-1/2}\textstyle\sum_{j=1}^n m_j\varepsilon_j| \ge \gamma\,(\sup_{|s-t|<\delta}|J_1(s) - J_1(t)|)^{-1}\right\} \\
=\ & 2\left\{1 - \Phi\left[\gamma\left(\sqrt{n^{-1}\textstyle\sum_{j=1}^n m_j^2}\ \sup_{|s-t|<\delta}|J_1(s) - J_1(t)|\right)^{-1}\right]\right\},
\end{aligned}$$

as $n^{-1/2}\sum_{j=1}^n m_j\varepsilon_j \sim N(0, n^{-1}\sum_{j=1}^n m_j^2)$. Under the assumption that $n^{-1}\sum_{j=1}^n m_j^2 \to b$, where $b < \infty$, there exists an $l > 0$ such that $n^{-1}\sum_{j=1}^n m_j^2 \le l^2$ for all $n \ge 1$. Furthermore, $J_1(t)$ is continuous on [0,1], therefore, is uniformly continuous on [0,1], so there is a $\delta:\ 0 < \delta < 1$ such that $\sup_{|s-t|<\delta}|J_1(s) - J_1(t)| \le h$, where h has the property that

$$1 - \Phi[\gamma/(lh)] \le \eta/2.$$

For this choice of δ,

$$P\left\{W_{\tilde{Y}_{2n}}(\delta) \geq \gamma\right\} \leq 2\{1 - \Phi(\gamma/(lh))\} \leq \eta$$

for all $n \geq 1$, proving the tightness of $\tilde{Y}_{2n}(t)$.

Second, $\tilde{Y}_{2n}(t) \rightarrow_{f.d.}$ normal distributions. To show this, for any k finite and any $\{t_i\}_{i=1}^k$ such that $0 \leq t_1 \leq t_2 \leq \ldots \leq t_k \leq 1$, consider for $n \geq 1$

$$X_n = (\tilde{Y}_{2n}(t_1), \tilde{Y}_{2n}(t_2), \ldots, \tilde{Y}_{2n}(t_k)).$$

Define for $j = 1, 2, \ldots, n$

$$W_j = (J_1(t_1)m_j\varepsilon_j, J_1(t_2)m_j\varepsilon_j, \ldots, J_1(t_k)m_j\varepsilon_j),$$

then $\{W_j\}_{j=1}^n$ are independent normal vectors and $X_n = n^{-1/2}\sum_{j=1}^n W_j$. Let $\Gamma^j = V(W_j)$, the covariance matrix of W_j, it is clear that $\Gamma^j_{ml} = m_j^2 J_1(t_m)J_1(t_l)$, $m, l = 1, 2, \ldots, k$. The characteristic function of $W_j/\sqrt{n}$ is

$$\begin{aligned}\varphi_{W_j/\sqrt{n}}(u) &= E\{\exp(i < u, W_j/\sqrt{n} >\} \\ &= \exp\left\{-\tfrac{1}{2n}u^t\Gamma^j u\right\},\end{aligned}$$

therefore, the characteristic function of X_n is

$$\begin{aligned}\varphi_{X_n}(u) &= E\{\exp(i < u, X_n >)\} \\ &= E\{\exp(i < u, (W_1 + W_2 + \ldots + W_n)/\sqrt{n} >)\} \\ &= \textstyle\prod_{j=1}^n E\{\exp(i < u, W_j/\sqrt{n} >\} \\ &= \textstyle\prod_{j=1}^n \exp\left\{-\frac{1}{2n}u^t\Gamma^j u\right\} \\ &= \exp\left\{-\tfrac{1}{2}u^t\left(\tfrac{1}{n}\textstyle\sum_{j=1}^n \Gamma^j\right)u\right\} \\ &\rightarrow \exp\{-\tfrac{1}{2}u^t\Gamma u\} \sim N_k(0, \Gamma),\end{aligned}$$

as $n \rightarrow \infty$, where $\Gamma_{ml} = J_1(t_m)J_1(t_l)b$, $m, l = 1, 2, \ldots, k$, proving that $\tilde{Y}_{2n}(t) \rightarrow_{f.d.}$ normal distributions.

Since $\tilde{Y}_{2n}(t)$ is tight and converges in finite distributions to normal distributions, $\tilde{Y}_{2n}(t)$ converges weakly to a Gaussian process, say, $\tilde{Y}_2(t)$. The above calculations show that $\tilde{Y}_2(t)$ has zero mean and covariance function

$$\rho_1(s,t) = \mathrm{Cov}(\tilde{Y}_2(s), \tilde{Y}_2(t)) = b\,\phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t)) = bJ_1(s)J_1(t). \quad \square$$

Corollary 2.4.1 *In Theorem 2.4.1, if the space spanned from the columns of model matrix X, denoted span(X), contains a constant column of 1's for all $n \geq n_0$, where n_0 is a fixed positive integer, then $\tilde{Y}_{2n}(t)$ converges weakly to a Gaussian process with zero mean and covariance function $\rho_1(s,t) = \phi(\Phi^{-1}(s))\phi(\Phi^{-1}(t))$; if for all $n \geq n_0$ the orthogonal complement of span(X) contains a column of 1's, then $\tilde{Y}_{2n}(t) = o_p(1)$.*

Proof. If *span*(X) contains a column of 1's, say 1_n, for n greater than some $n_0 > 0$, then $H\,1_n = 1_n$, where $H = X(X^tX)^{-1}X^t$ is the projection matrix. This implies that for $j = 1, 2, \ldots, n$, $m_j = 1$, therefore, $b = 1$; if $1_n \in span(X)^{\perp}$, then $m_j = 0$, hence, $\tilde{Y}_{2n}(t) = o_p(1)$. $\square$

2.4.3 Study of $Y_{1n}(t)$: $Y_{1n}(t) = o_p(1)$

Recall that $Y_{1n}(t)$ is given by

$$n^{-1/2}\sum_{i=1}^{n}\left\{I\left[u_i \leq \Phi\left(\Phi^{-1}(t) + x_i^t(\hat{\theta}-\theta)\right)\right] - \Phi\left(\Phi^{-1}(t) + x_i^t(\hat{\theta}-\theta)\right) - I[u_i \leq t] + t\right\}.$$

Since $\Phi(\Phi^{-1}(t)) = t$, if all $x_i^t(\hat{\theta}-\theta)$ behave nicely, then $Y_{1n}(t) = o_p(1)$ will be expected. It turns out that this is indeed the case. However, a rigorous proof of this result requires some restrictions on the model matrix $X = (x_{ij})_{n\times p}$.

Assumption A $\sqrt{n}(\hat{\theta}-\theta) = O_p(1)$, where $O_p(1)$ means bounded in probability.

Assumption B There exists an $M < \infty$ such that for all n, $n^{-1}\sum_{i=1}^{n}\max_{1\leq j\leq p}|x_{ij}| \leq M$.

Assumption C There exists an $M_1 < \infty$ such that for all $n \geq 1$ and $i = 1, 2, \ldots, n$, $n^{-1/2}\max_{1\leq j\leq p}|x_{ij}| \leq M_1$.

Theorem 2.4.2 (Random perturbation of empirical process, p is fixed) *In linear regression model (2.4.1), if assumptions A, B, and C hold, then* $Y_{1n}(t) = o_p(1)$.

The proof of Theorem 2.4.2 is pretty long, thus it is desirable to establish a few important steps as lemmas. The idea of the proof was outlined in Loynes (1980) for more general case than normal theory linear regression, and Loynes (1980) was based on Rao and Sethuraman (1975). The proof to be given here constitutes an application of the above two papers with some refinements.

Lemma 2.4.1 *In linear regression (2.4.1), let* $B_i(t,\xi,\theta) = \Phi\left(\Phi^{-1}(t) + x_i^t(\xi-\theta)\right)$, $i = 1,2,\ldots,n$. *For* $\gamma > 0$ *and* $L > 0$, *define* $Q_n(t,\gamma,L) = n^{-1/2}\sum_{i=1}^n q_i(t,\gamma,L)$, *where*

$$q_i(t,\gamma,L) = \sup_{|\xi_1-\xi_2|\le n^{-1/2}\gamma,\,|\xi_1-\theta|\le n^{-1/2}L,\,|\xi_2-\theta|\le n^{-1/2}L} |B_i(t,\xi_1,\theta) - B_i(t,\xi_2,\theta)|,$$

and $|\cdot|$ *denotes the maximum norm of* R^p. *If Assumption B holds, then for any* $L > 0$ *fixed,* $Q_n(t,\gamma,L) \to 0$ *uniformly in* n *and* t, *as* $\gamma \to 0$.

Proof. For ξ_1, ξ_2 such that $|\xi_1 - \xi_2| \le n^{-1/2}\gamma$, and $|\xi_i - \theta| \le n^{-1/2}L$, $i = 1,2$,

$$\begin{aligned}
&B_i(t,\xi_1,\theta) - B_i(t,\xi_2,\theta)\\
&= \Phi\left(\Phi^{-1}(t) + x_i^t(\xi_1-\theta)\right) - \Phi\left(\Phi^{-1}(t) + x_i^t(\xi_2-\theta)\right)\\
&= \phi(\eta)\, x_i^t\,(\xi_1-\xi_2)\\
&\le \sup_{x\in R}\phi(x)\, n^{-1/2}\gamma\, p \max_{1\le j\le p}|x_{ij}|.
\end{aligned}$$

So $Q_n(t,\gamma,L) \le \sup_{x\in R}\phi(x)\, pM\,\gamma \to 0$ uniformly in n and t as $\gamma \to 0$. □

Lemma 2.4.2 *Let* $R(u) = \phi(u+a)/\phi(u)$, *where* ϕ *is the density of the standard normal distribution, a is a real constant,* $u \in R$. *Then* $R(u)$ *is increasing if* $a < 0$; $R(u)$ *is decreasing if* $a > 0$.

Proof. Let $u_1, u_2 \in R$, such that $u_1 < u_2$, then

$$\begin{aligned} R(u_2) - R(u_1) &= \phi(u_2 + a)/\phi(u_2) - \phi(u_1 + a)/\phi(u_1) \\ &= \exp\left\{-\tfrac{1}{2}[(u_2 + a)^2 - u_2^2]\right\} - \exp\left\{-\tfrac{1}{2}[(u_1 + a)^2 - u_1^2]\right\} \\ &= \exp(-\tfrac{1}{2}a^2 - au_1)\,\{\exp(-a(u_2 - u_1)) - 1\} \\ &< 0 \text{ if } a > 0, \quad \text{or} \\ &> 0 \text{ if } a < 0. \quad \square \end{aligned}$$

Lemma 2.4.3 *In linear regression (2.4.1), let $a_i = x_i^t(\xi - \theta)$, $i = 1, 2, \ldots, n$. For $\lambda > 0$ and $L > 0$, define $E_n(\lambda, L) = \sup_{0 \le t \le 1} n^{-1/2} \sum_{i=1}^n r_i(\lambda, L)$, where*

$$r_i(\lambda, L) = \sup_{|\xi - \theta| \le n^{-1/2} L} |\Phi(\Phi^{-1}(t + \lambda/\sqrt{n}) + a_i) - \Phi(\Phi^{-1}(t) + a_i)|.$$

If Assumptions B and C hold, then for any fixed $L > 0$, $E_n(\lambda, L) \to 0$ uniformly in n, as $\lambda \to 0$.

Proof. Let $G_i(t) = \Phi(\Phi^{-1}(t + \lambda/\sqrt{n}) + a_i) - \Phi(\Phi^{-1}(t) + a_i)$, $t \in [0, 1 - \lambda/\sqrt{n}]$. Then

$$\frac{dG_i(t)}{dt} = \frac{\phi(\Phi^{-1}(t + \lambda/\sqrt{n}) + a_i)}{\phi(\Phi^{-1}(t + \lambda/\sqrt{n}))} - \frac{\phi(\Phi^{-1}(t) + a_i)}{\phi(\Phi^{-1}(t))}.$$

By Lemma 2.4.2, $G_i(t)$ is increasing if $a_i < 0$; $G_i(t)$ is decreasing if $a_i > 0$.

If $G_i(t)$ is decreasing in t, $t \in [0, 1 - \lambda/\sqrt{n}]$, then $G_i(t) \le G_i(0) = \Phi(\Phi^{-1}(\lambda/\sqrt{n}) + a_i)$. For ξ and θ such that $|\xi - \theta| \le L/\sqrt{n}$ and by Assumption C,

$$|a_i| = |x_i^t(\xi - \theta)| \le (Lp/\sqrt{n}) \max_{1 \le j \le p} |x_{ij}| \le LpM_1 < \infty.$$

Let $\lambda_0 > 0$ be such that $\Phi^{-1}(\lambda_0) = -LpM_1$. For $\lambda < \lambda_0$, $\Phi^{-1}(\lambda/\sqrt{n}) + LpM_1 < 0$. Thus for these arguments, $\phi(\cdot)$ is increasing, therefore, for some η such that $|\eta| < 1$,

$$\begin{aligned} \sup_{|\xi - \theta| \le L/\sqrt{n}} |G_i(t)| &\le \sup_{|\xi - \theta| \le L/\sqrt{n}} |\Phi(\Phi^{-1}(\lambda/\sqrt{n}) + a_i)| \\ &= \sup_{|\xi - \theta| \le L/\sqrt{n}} |\Phi(\Phi^{-1}(\lambda/\sqrt{n})) + \phi(\Phi^{-1}(\lambda/\sqrt{n}) + \eta\, a_i) a_i|, \\ &\le \lambda/\sqrt{n} + \phi(\Phi^{-1}(\lambda/\sqrt{n}) + LpM_1)\,(Lp/\sqrt{n}) \max_{1 \le j \le p} |x_{ij}| \\ &\le \lambda/\sqrt{n} + \phi(\Phi^{-1}(\lambda) + LpM_1)\,(Lp/\sqrt{n}) \max_{1 \le j \le p} |x_{ij}|. \end{aligned}$$

If $G_i(t)$ is increasing in t, $t \in [0, 1-\lambda/\sqrt{n}]$, then

$$G_i(t) \leq G_i(1-\lambda/\sqrt{n}) = 1 - \Phi(\Phi^{-1}(1-\lambda/\sqrt{n}) + a_i).$$

It follows that for some η such that $|\eta| < 1$,

$$\begin{aligned}
&\sup_{|\xi-\theta|\leq L/\sqrt{n}} |G_i(t)| \\
&\leq \sup_{|\xi-\theta|\leq L/\sqrt{n}} |1 - \Phi(\Phi^{-1}(1-\lambda/\sqrt{n}) + a_i)| \\
&= \sup_{|\xi-\theta|\leq L/\sqrt{n}} |1 - \Phi(\Phi^{-1}(1-\lambda/\sqrt{n})) - \phi(\Phi^{-1}(1-\lambda/\sqrt{n}) + \eta a_i)a_i|, \\
&\leq \sup_{|\xi-\theta|\leq L/\sqrt{n}} |\lambda/\sqrt{n} + \phi(\Phi^{-1}(1-\lambda/\sqrt{n}) + \eta)a_i|.
\end{aligned}$$

Now, for $\lambda < \lambda_0$, $1-\lambda/\sqrt{n} \geq 1-\lambda > 1-\lambda_0$, and $\Phi^{-1}(1-\lambda/\sqrt{n}) - LpM_1 > 0$ by symmetry of $\phi(\cdot)$. So for these arguments, $\phi(\cdot)$ is decreasing. Therefore, the above last inequality can be bounded by

$$\begin{aligned}
&\lambda/\sqrt{n} + \phi(\Phi^{-1}(1-\lambda) - LpM_1)(Lp/\sqrt{n}) \max_{1\leq j\leq p} |x_{ij}| \\
&= \lambda/\sqrt{n} + \phi(\Phi^{-1}(\lambda) + LpM_1)(Lp/\sqrt{n}) \max_{1\leq j\leq p} |x_{ij}|.
\end{aligned}$$

Therefore, regardless of each $G_i(t)$ being increasing or decreasing,

$$\begin{aligned}
E_n(\lambda, L) &= \sup_{0\leq t\leq 1} n^{-1/2} \sum_{i=1}^n \sup_{|\xi-\theta|\leq L/\sqrt{n}} |G_i(t)| \\
&\leq n^{-1/2} \sum_{i=1}^n \{\lambda/\sqrt{n} + \phi(\Phi^{-1}(\lambda) + LpM_1)(Lp/\sqrt{n}) \max_{1\leq j\leq p} |x_{ij}|\} \\
&= \lambda + Lp\,(n^{-1} \sum_{i=1}^n \max_{1\leq j\leq p} |x_{ij}|)\ \phi(\Phi^{-1}(\lambda) + LpM_1) \\
&\leq \lambda + LpM\phi(\Phi^{-1}(\lambda) + LpM_1) \\
&\to 0 \text{ uniformly in } n \text{ as } \lambda \to 0. \quad \square
\end{aligned}$$

Proof of Theorem 2.4.2. The goal is to show that $\forall\ \zeta > 0$, $\forall\ \delta > 0$, $\exists\ n_0 = n_0(\zeta, \delta)$ such that for all $n \geq n_0$

$$P\left\{\sup_{0\leq t\leq 1} |Y_{1n}(t)| > \zeta\right\} < \delta.$$

For the given $\delta > 0$, $\sqrt{n}(\hat{\theta} - \theta) = O_p(1)$ implies that there exists an $L = L(\delta) > 0$ such that for all $n \geq 1$,

$$P\left\{|\hat{\theta} - \theta| > L/\sqrt{n}\right\} < \delta/2,$$

where $|\hat{\theta} - \theta| = \max_{1 \leq i \leq p} |\hat{\theta}_i - \theta_i|$, $p = dim(\theta)$ is fixed finite. Hence, for the given $\zeta > 0$,

$$\begin{aligned}
& P\left\{\sup_{0 \leq t \leq 1} |Y_{1n}(t)| > \zeta\right\} \\
&= P\left\{\sup_{0 \leq t \leq 1} |Y_{1n}(t)| > \zeta,\ |\hat{\theta} - \theta| > L/\sqrt{n}\right\} \\
&+ P\left\{\sup_{0 \leq t \leq 1} |Y_{1n}(t)| > \zeta,\ |\hat{\theta} - \theta| \leq L/\sqrt{n}\right\} \\
&= p_1 + p_2, \text{ say} \\
&< \delta/2 + p_2.
\end{aligned}$$

It then remains to show that $p_2 = P\left\{\sup_{0 \leq t \leq 1} |Y_{1n}(t)| > \zeta,\ |\hat{\theta} - \theta| \leq L/\sqrt{n}\right\} < \delta/2$, for n greater than some n_0. To this end, define

$$R_n(t, \xi, \theta) = n^{-1/2} \sum_{i=1}^{n} \{I[u_i \leq B_i(t, \xi, \theta)] - B_i(t, \xi, \theta) - I[u_i \leq t] + t\},$$

where $B_i(t, \xi, \theta) = \Phi(\Phi^{-1}(t) + x_i^t(\xi - \theta))$. Then $Y_{1n}(t) = R_n(t, \hat{\theta}, \theta)$ and

$$p_2 \leq P\left\{\sup_{0 \leq t \leq 1} \sup_{|\xi - \theta| \leq L/\sqrt{n}} |R_n(t, \xi, \theta)| > \zeta\right\}.$$

The following shows that the last term above is less than $\delta/2$ for n greater than some $n_0 \geq 1$. In fact, detailed steps are given for

$$P\left\{\sup_{0 \leq t \leq 1} \sup_{|\xi - \theta| \leq L/\sqrt{n}} R_n(t, \xi, \theta) > \zeta\right\};$$

the steps for

$$P\left\{\sup_{0 \leq t \leq 1} \sup_{|\xi - \theta| \leq L/\sqrt{n}} R_n(t, \xi, \theta) < -\zeta\right\}$$

can be filled in similarly.

Step 1: Discretization. Divide the cube $C(\theta)$ centered at θ (the true parameter value) and of side length $2L/\sqrt{n}$ into $\{[2L/\gamma] + 1\}^p$ closed subcubes of side length $\gamma/\sqrt{n}$, where γ

is to be chosen later, $[x]$ denotes the greatest integer less than or equal to x. Label these subcubes arbitrarily.

Let the k^{th} subcube be C_k. Define $\xi_{ik}^1(t)$, $\xi_{ik}^2(t)$ to be the values of $\xi \in C_k$ such that $B_i(t,\xi,\theta)$ takes its maximum and minimum values, respectively, and define

$$q_i(t,\gamma) = \sup_{|\xi_1-\xi_2|\le\gamma/\sqrt{n},\ |\xi_1-\theta|\le L/\sqrt{n},\ |\xi_2-\theta|\le L/\sqrt{n}} |B_i(t,\xi_1,\theta) - B_i(t,\xi_2,\theta)|$$

as in Lemma 2.4.1, then

$$B_i(t,\xi_{ik}^1(t),\theta) - B_i(t,\xi_{ik}^2(t),\theta) \le q_i(t,\gamma).$$

Hence, let $B_{ik}(t) = B_i(t,\xi_{ik}^1(t),\theta)$, then

$$\begin{aligned} \sup_{\xi\in C_k} R_n(t,\xi,\theta) &\le n^{-1/2}\textstyle\sum_{i=1}^n \{I[u_i \le B_{ik}(t)] - B_{ik}(t) - I[u_i \le t] + t\} \\ &+ n^{-1/2}\textstyle\sum_{i=1}^n q_i(t,\gamma). \end{aligned}$$

By Lemma 2.4.1, $n^{-1/2}\sum_{i=1}^n q_i(t,\gamma)$ can be made small uniformly in n, in t and in k by choosing γ to be small.

Next, divide $[0,1]$ into $[\sqrt{n}/\lambda]+1$ subintervals of length $\lambda/\sqrt{n}$ (except possibly the last subinterval) by points $t_s = s(\lambda/\sqrt{n})$, $s = 0,1,2,\ldots,[\sqrt{n}/\lambda],\sqrt{n}/\lambda$. Let $I_s = [t_s, t_{s+1}]$.

Since $B_i(t,\xi,\theta)$ is an increasing function of t, so is $B_{ik}(t) = B_i(t,\xi_{ik}^1(t),\theta)$. Therefore, for any $t \in I_s$,

$$\begin{aligned} & n^{-1/2}\textstyle\sum_{i=1}^n \{I[u_i \le B_{ik}(t)] - B_{ik}(t) - I[u_i \le t] + t\} \\ \le\ & n^{-1/2}\textstyle\sum_{i=1}^n \{I[u_i \le B_{ik}(t_{s+1})] - B_{ik}(t_{s+1}) - I[u_i \le t_s] + t_s\} \\ & + n^{-1/2}\textstyle\sum_{i=1}^n \{t_{s+1} - t_s + B_{ik}(t_{s+1}) - B_{ik}(t_s)\}. \end{aligned}$$

But

$$\begin{aligned} & n^{-1/2}\sum_{i=1}^{n}\{t_{s+1}-t_s+B_{ik}(t_{s+1})-B_{ik}(t_s)\} \\ = \;& \lambda+n^{-1/2}\sum_{i=1}^{n}\{B_{ik}(t_{s+1})-B_{ik}(t_s)\} \\ = \;& \lambda+n^{-1/2}\sum_{i=1}^{n}\{B_{ik}(t_{s+1})-B_i(t_{s+1},\xi_{ik}^1(t_s),\theta)\} \\ & +n^{-1/2}\sum_{i=1}^{n}\{B_i(t_{s+1},\xi_{ik}^1(t_s),\theta)-B_{ik}(t_s)\} \\ \leq \;& \lambda+n^{-1/2}\sum_{i=1}^{n}q_i(t_{s+1},\gamma)+n^{-1/2}\sum_{i=1}^{n}r_i(\lambda,t_s), \end{aligned}$$

where $q_i(t,\gamma)$ is defined in Lemma 2.4.1, and $r_i(\lambda,t)$ is defined in Lemma 2.4.3 as

$$r_i(\gamma,t)=\sup_{|\xi-\theta|\leq L/\sqrt{n}}\{|B_i(t+\lambda/\sqrt{n},\xi,\theta)-B_i(t,\xi,\theta)|\}.$$

Thus, by Lemma 2.4.1 and Lemma 2.4.3, $n^{-1/2}\sum_{i=1}^{n}\{t_{s+1}-t_s+B_{ik}(t_{s+1})-B_{ik}(t_s)\}$ can be made arbitrarily small, uniformly in n, in k and in s, by choosing γ and λ small.

Together, it has been shown that

$$\begin{aligned} & \sup_{t\in I_s}\sup_{\xi\in C_k}R_n(t,\xi,\theta) \\ \leq \;& n^{-1/2}\sum_{i=1}^{n}\{I[u_i\leq B_{ik}(t_{s+1})]-B_{ik}(t_{s+1})-I[u_i\leq t_s]+t_s\} \qquad (2.4.3) \\ & +o(\gamma,\lambda), \end{aligned}$$

where $o(\gamma,\lambda)\to 0$ uniformly in n, k and s as $\gamma\to 0$ and $\lambda\to 0$. For the given $\zeta>0$, choose γ and λ small such that $o(\gamma,\lambda)<\zeta/2$. This finishes the discretization process.

Step 2: Exponential Bound. In this step, both γ and λ are fixed at the values chosen at the end of last step. The idea of this step is to represent the summation in (2.4.3) in terms of Bernoulli random variables and obtain a bound from this representation. To this

end, define, for fixed k and s,

$$p_{iks} = |B_{ik}(t_{s+1}) - t_s|,\ i = 1, 2, \ldots, n,$$

$$sgn(iks) = \begin{cases} +1\ if\ B_{ik}(t_{s+1}) - t_s \geq 0, \\ -1\ if\ B_{ik}(t_{s+1}) - t_s < 0, \end{cases}$$

$$w_{iks} = (X_{iks} - p_{iks}) sgn(iks),\ i = 1, 2, \ldots, n,$$

where $\{X_{iks}\}_{i=1}^n$ are independent Binomial$(1, p_{iks})$ variables. Then,

$$\begin{aligned} & n^{-1/2} \textstyle\sum_{i=1}^n \{I[u_i \leq B_{ik}(t_{s+1})] - B_{ik}(t_{s+1}) - I[u_i \leq t_s] + t_s\} \\ & = n^{-1/2} \textstyle\sum_{i=1}^n \{I[u_i \leq B_{ik}(t_{s+1})] - I[u_i \leq t_s] - (B_{ik}(t_{s+1}) - t_s)\} \\ & = n^{-1/2} \textstyle\sum_{i=1}^n \{X_{iks} - p_{iks}\}\, sgn(iks) \\ & = n^{-1/2} \textstyle\sum_{i=1}^n w_{iks}. \end{aligned}$$

Let $T_+ = \{i :\ sgn(iks) = +1\}$, $T_- = \{i :\ sgn(iks) = -1\}$. There is

$$n^{-1/2} \sum_{i=1}^{n} w_{iks} = n^{-1/2} \left\{ \sum_{i \in T_+} (X_{iks} - p_{iks}) - \sum_{i \in T_-} (X_{iks} - p_{iks}) \right\}.$$

Therefore,

$$\begin{aligned} & P\left\{ \sup_{t \in I_s} \sup_{\xi \in C_k} R_n(t, \xi, \theta) > \zeta \right\} \\ \leq\ & P\left\{ n^{-1/2} \sum_{i=1}^{n} w_{iks} > \zeta/2 \right\} \\ \leq\ & P\left\{ \sum_{i \in T_+} (X_{iks} - p_{iks}) > \sqrt{n}\zeta/4 \right\} \\ +\ & P\left\{ -\sum_{i \in T_-} (X_{iks} - p_{iks}) > \sqrt{n}\zeta/4 \right\} \\ =\ & p_{21} + p_{22},\ \text{say.} \end{aligned} \tag{2.4.4}$$

Note that

$$\begin{aligned} \forall\, t>0 \qquad & \textstyle\sum_{i\in T_+}(X_{iks}-p_{iks}) \;>\; \sqrt{n}\zeta/4 \\ \Leftrightarrow \qquad & \exp\left\{t\,\textstyle\sum_{i\in T_+}(X_{iks}-p_{iks})\right\} \;>\; \exp\{t\sqrt{n}\zeta/4\}. \end{aligned}$$

Thus,

$$\begin{aligned} p_{21} &= P\left\{\textstyle\sum_{i\in T_+}(X_{iks}-p_{iks}) \;>\; \sqrt{n}\zeta/4\right\} \\ &= P\left\{\exp\{t\,\textstyle\sum_{i\in T_+}(X_{iks}-p_{iks})\} > \exp\{t\sqrt{n}\zeta/4\}\right\} \\ &\le \exp\{-t\sqrt{n}\zeta/4\}\textstyle\prod_{i\in T_+}\; E\left\{\exp\left(t\,(X_{iks}-p_{iks})\right)\right\} \\ &= a(n,k,s,t,\zeta), \text{ for all } t>0, \text{ say,} \end{aligned}$$

by Markov's inequality and independence of $\{X_{iks}-p_{iks}\}$. Furthermore, because $X_{iks}-p_{iks}$ is a centered Bernoulli random variable whose moment generating function is readily available, there is

$$\begin{aligned} E\left\{\exp\left(t\,(X_{iks}-p_{iks})\right)\right\} &= \{1+p_{iks}(e^t-1)\}\exp(-tp_{iks}) \\ &\ge \exp\left\{tE(X_{iks}-p_{iks})\right\} \\ &= 1, \text{ by Jenson's inequality.} \end{aligned}$$

Hence, taking logarithm of $a(n,k,s,t,\zeta)$ and enlarging $\prod_{i\in T_+}$ to $\prod_{i=1}^n$ gives, for all $t>0$,

$$\begin{aligned} & n^{-1}\log\, a(n,k,s,t,\zeta) \\ = \;& n^{-1}\log\left\{\exp\{-t\sqrt{n}\zeta/4\}\prod_{i\in T_+}\; E\left\{\exp\left(t\,(X_{iks}-p_{iks})\right)\right\}\right\} \\ \le \;& n^{-1}\log\left\{\exp\{-t\sqrt{n}\zeta/4\}\prod_{i=1}^{n}\; E\left\{\exp\left(t\,(X_{iks}-p_{iks})\right)\right\}\right\} \\ = \;& n^{-1}\left\{-t\sqrt{n}\zeta/4+\sum_{i=1}^{n}\left[-tp_{iks}+\log\left(1+p_{iks}(e^t-1)\right)\right]\right\} \qquad (2.4.5) \\ = \;& -n^{-1/2}\left\{t\zeta/4+tn^{-1/2}\sum_{i=1}^{n}p_{iks}-n^{-1/2}\sum_{i=1}^{n}\log[1+p_{iks}(e^t-1)]\right\}. \end{aligned}$$

Notice that

$$\begin{aligned}\textstyle\sum_{i=1}^n p_{iks} &= \textstyle\sum_{i=1}^n |B_{ik}(t_{s+1}) - t_s| \\ &\leq \textstyle\sum_{i=1}^n |B_i(t_{s+1}, \xi_{ik}^1(t_{s+1}), \theta) - B_i(t_{s+1}, \theta, \theta)| + \sum_{i=1}^n |t_{s+1} - t_s| \\ &\leq \textstyle\sum_{i=1}^n B_i(t_{s+1}, \xi_{ik}^1(t_{s+1}), \theta) - B_i(t_{s+1}, \theta, \theta) + \sqrt{n}\lambda.\end{aligned}$$

By the definition of $q_i(t, \gamma)$, namely,

$$q_i(t,\gamma) = \sup_{|\xi_1 - \xi_2| \leq \gamma/\sqrt{n},\ |\xi_1 - \theta| \leq L/\sqrt{n},\ |\xi_2 - \theta| \leq L/\sqrt{n}} |B_i(t, \xi_1, \theta) - B_i(t, \xi_2, \theta)|,$$

for fixed $t = t_{s+1}$, θ and L, as long as $|\xi_{ik}^1(t_{s+1}) - \theta| \leq \gamma/\sqrt{n}$, there is

$$|B_i(t_{s+1}, \xi_{ik}^1(t_{s+1}), \theta) - B_i(t_{s+1}, \theta, \theta)| \leq q_i(t_{s+1}, \gamma).$$

However, the maximum norm distance $|\xi_{ik}^1(t_{s+1}) - \theta|$ may be larger than $\gamma/\sqrt{n}$. The worst possible case is when $\xi_{ik}^1(t_{s+1})$ is at one of the corners of $C(\theta)$. For this case, the (Euclidean) distance from $\xi_{ik}^1(t_{s+1})$ to θ is at most

$$(\sqrt{p}/2)\{[2L/\gamma] + 1\}(\gamma/\sqrt{n}),$$

Now, let $L(\gamma) = \{[\sqrt{p}/2] + 1\}\{[2L/\gamma] + 1\}$, and join $\xi_{ik}^1(t_{s+1})$ and θ by a line. Put points along this line a (Euclidean) distance $\gamma/\sqrt{n}$ apart. It is clear that at most $L(\gamma)\gamma/\sqrt{n}$ points are needed. With this observation, there is the inequality

$$|B_i(t_{s+1}, \xi_{ik}^1(t_{s+1}), \theta) - B_i(t_{s+1}, \theta, \theta)| \leq L(\gamma) q_i(t_{s+1}, \gamma).$$

Therefore,

$$\begin{aligned}\textstyle\sum_{i=1}^n p_{iks} &= \textstyle\sum_{i=1}^n |B_{ik}(t_{s+1}) - t_s| \\ &\leq L(\gamma) \textstyle\sum_{i=1}^n q_i(t_{s+1}, \gamma) + \sqrt{n}\lambda \\ &= \left\{ L(\gamma) n^{-1/2} \textstyle\sum_{i=1}^n q_i(t_{s+1}, \gamma) + \lambda \right\} \sqrt{n}.\end{aligned}$$

But

$$\sup_n |n^{-1/2} \sum_{i=1}^n q_i(t_{s+1}, \gamma)| \leq \sup_x \phi(x) p M \gamma < \infty$$

by Assumption B. Let $K(\gamma) = \sup_x \phi(x) p M \gamma$, it follows that

$$n^{-1/2} \sum_{i=1}^{n} p_{iks} \leq L(\gamma) K(\gamma) + \lambda.$$

From $\log(1+x) \leq x,\ x \geq 0$, one has for $t > 0$

$$\begin{aligned} & n^{-1/2} \textstyle\sum_{i=1}^{n} \log\{1 + p_{iks}(e^t - 1)\} \\ \leq\ & n^{-1/2} \textstyle\sum_{i=1}^{n} p_{iks}(e^t - 1) \\ \leq\ & (e^t - 1)\{L(\gamma)K(\gamma) + \lambda\}. \end{aligned}$$

Combining above results gives

$$n^{-1} \log a(n,k,s,t,\zeta) \leq -n^{-1/2}\{t\zeta/4 + a(t + 1 - e^t)\},\ t > 0,$$

where $a = L(\gamma)K(\gamma) + \lambda > 1$.

For $t \geq 0$, let $h(t) = t\zeta/4 + a(t + 1 - e^t)$. Then $h(t)$ is continuous and $h(0) = 0$. Since $dh(t)/dt > 0$ if $0 < t < \log\{1 + \zeta/(4a)\}$, there is a $t_0 > 0$ such that $c = h(t_0) > 0$ and with this positive constant c

$$n^{-1} \log a(n,k,s,t,\zeta) \leq -n^{-1/2} c,$$

or

$$a(n,k,s,t,\zeta) \leq e^{-\sqrt{n}c},$$

or

$$p_{21} \leq e^{-\sqrt{n}c}, \qquad c > 0.$$

Similar arguments will lead to

$$p_{22} \leq e^{-\sqrt{n}c}, \qquad c > 0.$$

Together, it has been shown that for any s and any k,

$$\begin{aligned} & P\left\{\sup_{t \in I_s} \sup_{\xi \in C_k} R_n(t,\xi,\theta) > \zeta\right\} \\ \leq\ & p_{21} + p_{22} \\ \leq\ & 2e^{-\sqrt{n}c}, \qquad c > 0. \end{aligned}$$

Notice that $2e^{-\sqrt{n}c}$ is independent of k and s. This finishes the step of obtaining an exponential bound.

Step 3: $Y_{1n}(t) = o_p(1)$. Now consider all I_s $(i = 1, 2, \ldots, [\sqrt{n}/\lambda], \sqrt{n}/\lambda)$ and all C_k $(k = 1, 2, \ldots, ([2L/\gamma] + 1)^p)$ at the same time. By Bonferroni inequality,

$$\begin{aligned} p_2 &= P\left\{\sup_{0\le t\le 1}\sup_{|\xi-\theta|\le L/\sqrt{n}} R_n(t,\xi,\theta) > \zeta\right\} \\ &\le P\left\{\max_{t\in I_s, 1\le s\le\sqrt{n}/\lambda}\max_{\xi\in C_k, 1\le k\le([2L/\gamma]+1)^p} R_n(t,\xi,\theta) > \zeta\right\} \\ &\le \{[\sqrt{n}/\lambda]+1\}\{([2L/\gamma]+1)^p\}\, 2e^{-\sqrt{n}c} \\ &= A(\gamma,\lambda,L,p,n)e^{-\sqrt{n}c}, \text{ say.} \end{aligned}$$

Because for the chosen γ, λ, L and the fixed p, $A(\gamma,\lambda,L,p,n)$ grows polynomially, while $e^{-\sqrt{n}c}$ decays exponentially, as $n \to \infty$, therefore, there exists an $n_0 = n_0(\gamma,\lambda)$ (more precisely there exists an $n_0 = n_0(\gamma,\lambda,L(\delta),p) \ge 1$) such that for all $n \ge n_0$, (see remarks before **Step 1**),

$$p_2 = P\left\{\sup_{0\le t\le 1}\sup_{|\xi-\theta|\le L/\sqrt{n}} |R_n(t,\xi,\theta)| > \zeta\right\} < \delta/2.$$

Thus,

$$P\left\{\sup_{0\le t\le 1}|Y_{1n}(t)| > \zeta\right\} = p_1 + p_2 < \delta/2 + \delta/2 = \delta,\ \forall\ n \ge n_0,$$

proving $Y_{1n}(t) = o_p(1)$. □

2.4.4 The main result

Theorem 2.4.3 (Weak convergence of residual process: p is fixed) *In linear regression (2.4.1), estimate regression parameter θ and standard deviation σ by the method of least squares. Let $\tilde{e}_i = \Phi\{(y_i - x_i^t\hat{\theta})/\hat{\sigma}\}$ and define for $t \in [0,1]$*

$$\tilde{Y}_n(t) = n^{-1/2}\sum_{i=1}^{n}\{I[\tilde{e}_i \le t] - t\}. \tag{2.4.6}$$

Denote $m_j = \sum_{i=1}^n h_{ij}$, $j = 1, 2, \ldots, n$. *where* $H = (h_{ij}) = X(X^tX)^{-1}X^t$. *If* p *is fixed finite,* $0 \le b = \lim_{n\to\infty} n^{-1}\sum_{j=1}^n m_j^2 < \infty$, *and Assumptions A, B, and C hold, then*

(1) $\tilde{Y}_n(t) = n^{-1/2}\sum_{i=1}^n \{I[u_i \le t] - t\} + J_1(t)n^{-1/2}\sum_{j=1}^n m_j\varepsilon_j$
$+ J_2(t)2^{-1/2}(2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu) + o_p(1)$, *and*

(2) $\tilde{Y}_n(t)$ *converges weakly to a Gaussian process* $\tilde{Y}(t)$ *with zero mean and covariance function*

$$\tilde{\rho}(s,t) = \min(s,t) - st - bJ_1(s)J_1(t) - \frac{1}{2}J_2(s)J_2(t), \tag{2.4.7}$$

where $J_1(t) = \phi(\Phi^{-1}(t))$, $J_2(t) = \Phi^{-1}(t)\phi(\Phi^{-1}(t))$, $t \in [0,1]$, $\nu = n - p$, and u_i are iid $U(0,1)$ variables.

Proof. The desired conclusion comes from applying Theorem 2.3.2, Theorem 2.3.1, Theorem 2.4.1 and Theorem 2.4.2. □

Corollary 2.4.2 (Asymptotic expansions of $\tilde{Y}_n(t)$) *In linear regression (2.4.1), suppose that the conditions of Theorem 2.4.3 hold. If the model matrix* X *contains a column of* 1*'s, then the following asymptotic expansion holds,*

$$\begin{aligned}\tilde{Y}_n(t) &= n^{-1/2}\sum_{i=1}^n \{I[u_i \le t] - t\} + n^{-1/2}J_1(t)\sum_{i=1}^n \varepsilon_i \\ &\quad + 2^{-1}J_2(t)n^{-1/2}\sum_{i=1}^n(\varepsilon_i^2 - 1) + o_p(1).\end{aligned} \tag{2.4.8}$$

Proof. From (1) of Theorem 2.4.3, rewrite the last part of the expansion there as

$$\begin{aligned}&J_2(t)2^{-1/2}(2\nu)^{-1/2}(\nu\hat{\sigma}^2 - \nu) \\ =\ &J_2(t)2^{-1/2}(2\nu)^{-1/2}\{(\textstyle\sum_{i=1}^n \varepsilon_i^2) - \nu\} - J_2(t)2^{-1/2}(2\nu)^{-1/2}\varepsilon^tH\varepsilon.\end{aligned}$$

Let $\sup_{0\le t\le 1} J_2(t) = k < \infty$, then for any $\delta > 0$,

$$\begin{aligned}&P\left\{\sup_{0\le t\le 1}|J_2(t)2^{-1/2}(2\nu)^{-1/2}\varepsilon^tH\varepsilon| \ge \delta\right\} \\ \le\ &P\left\{\varepsilon^tH\varepsilon \ge [(2\delta\sqrt{\nu})/k]\right\} \\ \le\ &[k/(2\delta\sqrt{\nu})]p \\ =\ &[k/(2\delta)](p/\sqrt{\nu}) \\ \to\ &0.\end{aligned}$$

This shows that the expansion in (2.4.8) holds. □

Remark 4. A drawback of requiring $\lim_{n\to\infty} n^{-1}\sum_{j=1}^{n} m_j^2$ to exist is that the condition is not given in terms of the model matrix X directly. It is however possible to do so. For example, in straight line regression through the origin, the model is

$$y_i = x_i\theta + \sigma\varepsilon_i,\ \varepsilon_i \text{ are independent } N(0,1).$$

If $a = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} x_i$ exists and if $0 < h = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} x_i^2$ exists finite, then $n^{-1/2}\sum_{i=1}^{n}\left\{I[\Phi\{(y_i - x_i\hat{\theta})/\hat{\sigma}\} \le t] - t\right\}$ converges weakly to a Gaussian process with zero mean and covariance function

$$\rho(s,t) = \min(s,t) - st - (a^2/h)J_1(s)J_1(t) - \frac{1}{2}J_2(s)J_2(t).$$

To see this, note that $m_j = \{(\sum_{i=1}^{n} x_i)/(\sum_{i=1}^{n} x_i^2\}x_j$ in this special case, and $n^{-1}\sum_{i=1}^{n} m_j^2 = (n^{-1}\sum_{i=1}^{n} x_i)^2/(n^{-1}\sum_{i=1}^{n} x_i^2) \to a^2/b$, as $n \to \infty$.

Remark 5. It is also possible to restrict general model matrices directly. For instance, in linear regression (2.4.1), suppose $\sigma = 1$ is known and Assumptions B and C hold. Suppose further that

1. $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} x_{ij} = b_j,\ j = 1,2,\ldots,p.$

2. $\lim_{n\to\infty} n^{-1}(X^tX) = \Sigma$, Σ is positive definite.

Then $n^{-1/2}\sum_{i=1}^{n}\left\{I[\Phi(y_i - x_i^t\hat{\theta}) \le t] - t\right\}$ converges weakly to a Gaussian process with zero mean and covariance function

$$\rho(s,t) = \min(s,t) - st - \Psi^t(s)\Sigma^{-1}\Psi(t),$$

where $\Psi(t) = J_1(t)B$, $B = (b_1, b_2, \ldots, b_p)^t$ is a p by 1 vector. This is essentially the formulation used by Darling (1955), Durbin (1973a), Loynes (1980) and many other authors.

Remark 6. When the model matrix X contains a column of 1's and when the conditions of Theorem 2.4.3 hold, the conclusion of Theorem 2.4.3 is identical to that when a random $N(\mu,\sigma^2)$ sample is drawn and both μ and σ need to be estimated.

Remark 7. Although the existance of $H = (h_{ij}) = X(X^tX)^{-1}X^t$ is required in several theorems of this section, it is possible to replace $(X^tX)^{-1}$ by any kind of generalized inverse $(X^tX)^-$ to reach the same conclusions.

2.5 Analysis of Variance Models: The Number of Regression Parameters is Changing With the Sample Size

In deriving the results of the previous two sections, two boundedness conditions are imposed on the model matrix X. However, these two conditions are naturally satisfied by analysis of variance models. In this section, advantage is taken of the simple structure of the model matrix X for analysis of Variance models, and a less well-known asymptotic covariance function structure is established for the empirical processes in these models.

2.5.1 One-way layout

In the case of one-way layout, Meester and Lockhart (1988) obtained the following weak convergence result. Consider the one-way layout model

$$y_{ij} = \mu + \tau_j + \varepsilon_{ij}, \tag{2.5.1}$$

where y_{ij} $(i = 1, 2, \ldots, K)$ denote the K observations under treatment j $(j = 1, 2, \ldots, Q)$, μ is grand mean, τ_j denotes the effect due to treatment j, $\sum_{j=1}^{Q} \tau_j = 0$, and ε_{ij} are iid $N(0, \sigma^2)$. If (ordinary) residuals $y_{ij} - \bar{y}_{.j}$ are (completely) standardized, that is,

$$e_{ij} = \frac{y_{ij} - \bar{y}_{.j}}{w_a \hat{\sigma}} = \frac{\varepsilon_{ij} - \bar{\varepsilon}_{.j}}{w_a \hat{\sigma}}, \tag{2.5.2}$$

where $\bar{y}_{.j} = K^{-1} \sum_{i=1}^{K} y_{ij}$, $\bar{\varepsilon}_{.j} = K^{-1} \sum_{i=1}^{K} \varepsilon_{ij}$, $w_a = \sqrt{1 - 1/K}$, and $\hat{\sigma}$ is the square root of the mean squares due to error, then the empirical process

$$(KQ)^{-1/2} \sum_{i=1}^{K} \sum_{j=1}^{Q} \{I[\Phi(e_{ij}) \leq t] - t\} \tag{2.5.3}$$

converges weakly, as $Q \to \infty$, to a Gaussian process with zero mean and covariance function

$$\begin{aligned}\rho(s,t;K) &= \min(s,t) - st \\ &\quad +(K-1)\left[\Phi_2\left(\Phi^{-1}(s),\Phi^{-1}(t);\frac{1}{K-1}\right) - st\right] \\ &\quad -\frac{K}{2(K-1)}J_2(s)J_2(t), \end{aligned} \tag{2.5.4}$$

where $\Phi_2(x,y;\rho)$ is the bivariate cumulative distribution function for a normal distribution with zero mean vector, unit variances, and correlation coefficient ρ, and $J_2(t)$ is defined before as $\Phi^{-1}(t)\phi(\Phi^{-1}(t))$.

As pointed out by Meester and Lockhart (1988), when K also increases to infinity, there is the limit

$$\rho(s,t;\infty) = \min(s,t) - st - J_1(s)J_1(t) - \frac{1}{2}J_2(s)J_2(t),$$

which is essentially the covariance structure studied in section 2.3 and section 2.4. Thus, paying close attention to model matrix does lead to finer results. Meester and Lockhart (1988) also conjectured that the same result holds for three other designs (balanced two-way layout without interactions, balanced incomplete block and Latin squares).

In the following subsection, it is to be shown that the empirical processes associated with a large class of analysis of variance models also converge weakly to a Gaussian process with zero mean and covariance function of the same type as given in (2.5.4).

2.5.2 Two-way layout

The model for a balanced two-way layout (without replications, without interactions) is given by

$$y_{ij} = \mu + \rho_i + \tau_j + \sigma\varepsilon_{ij}, \tag{2.5.5}$$

where μ is grand mean,

ρ_i are row effects, subject to $\sum_{i=1}^{K}\rho_i = 0$,

τ_j are column effects, subject to $\sum_{j=1}^{Q} \tau_j = 0$,

ε_{ij} are iid $N(0,1)$,

σ is a positive constant.

To simplify presentation, it is assumed that $\sigma = 1$ is known whenever arguments do not involve σ. This, according to an analogue of Theorem 2.3.2, does not weaken the conclusion to be drawn.

Denote

$$\begin{aligned}
\bar{\varepsilon}_{..} &= (KQ)^{-1} \textstyle\sum_{i=1}^{K} \sum_{j=1}^{Q} \varepsilon_{ij}, \\
\bar{\varepsilon}_{i.} &= Q^{-1} \textstyle\sum_{j=1}^{Q} \varepsilon_{ij}, \\
\bar{\varepsilon}_{.j} &= K^{-1} \textstyle\sum_{i=1}^{K} \varepsilon_{ij}, \\
w_a &= \left(1 - \tfrac{1}{K}\right)^{1/2}, \\
w_b &= \left(1 - \tfrac{1}{K} - \tfrac{1}{Q} + \tfrac{1}{KQ}\right)^{1/2}, \\
n &= KQ,
\end{aligned}$$

where w_a and w_b are the standard deviations of the residuals $y_{ij} - \hat{y}_{ij}$ for one-way and two-way layout, respectively. For one-way layout, the (ordinary) residuals are $\varepsilon_{ij} - \bar{\varepsilon}_{.j}$; for two-way layout (2.5.5), the (ordinary) residuals are

$$R_{ij}^{*} = y_{ij} - \hat{y}_{ij} = \varepsilon_{ij} - \bar{\varepsilon}_{i.} - \bar{\varepsilon}_{.j} + \bar{\varepsilon}_{..}. \tag{2.5.6}$$

Now let

$$\begin{aligned}
e_{ij}^{*} &= R_{ij}^{*}/w_b, \\
e_{ij}^{**} &= R_{ij}^{*}/w_a, \\
e_{ij} &= (\varepsilon_{ij} - \bar{\varepsilon}_{.j})/w_a,
\end{aligned}$$

then the empirical process for two-way layout (2.5.5) is defined for $t \in [0,1]$ by

$$G_n^{*}(t) = (KQ)^{-1/2} \sum_{i=1}^{K} \sum_{j=1}^{Q} \{I[\Phi(e_{ij}^{*}) \le t] - t\}. \tag{2.5.7}$$

Similarly, define

$$G_n^{**}(t) = (KQ)^{-1/2} \sum_{i=1}^{K} \sum_{j=1}^{Q} \{I[\Phi(e_{ij}^{**}) \le t] - t\}, \tag{2.5.8}$$

$$G_n(t) = (KQ)^{-1/2} \sum_{i=1}^{K} \sum_{j=1}^{Q} \{I[\Phi(e_{ij}) \le t] - t\}. \tag{2.5.9}$$

Since $w_b - w_a = O(1/Q)$ as $Q \to \infty$, it follows from the change of time technique that $G_n^*(t)$ and $G_n^{**}(t)$ have the same weak limit, if such a limit exists. Therefore, it is equivalent to study $G_n^{**}(t)$ instead.

Using the same technique used in section 2.4, the process $G_n^{**}(t)$ can be rewritten as

$$G_n^{**}(t) = G_{1n}(t) + G_{2n}(t) + G_{3n}(t),$$

where

$$\begin{aligned}
G_{1n}(t) &= (KQ)^{-1/2} \textstyle\sum_{i=1}^{K} \sum_{j=1}^{Q} \{I[\Phi(e_{ij}) \le \Phi(\Phi^{-1}(t) + (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})/w_a)] \\
&\quad -\Phi(\Phi^{-1}(t) + (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})/w_a) - I[\Phi(e_{ij}) \le t] + t\}, \\
G_{2n}(t) &= (KQ)^{-1/2} \textstyle\sum_{i=1}^{K} \sum_{j=1}^{Q} \{\Phi(\Phi^{-1}(t) + (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})/w_a) - t\}, \\
G_{3n}(t) &= (KQ)^{-1/2} \textstyle\sum_{i=1}^{K} \sum_{j=1}^{Q} \{I[\Phi(e_{ij}) \le t] - t\}.
\end{aligned}$$

Clearly, $G_{3n}(t)$ is just the empirical process for one-way layout (2.5.1) when $\sigma = 1$ is known, so as $Q \to \infty$, the weak limit of $G_{3n}(t)$ is already known. Because $\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot}$ are estimates for the row effects ρ_i only, $G_{2n}(t) = o_p(1)$ follows from Corollary 2.4.1. It is then plain to see that if $G_{1n}(t) = o_p(1)$, the process $G_n^*(t)$ will have the same weak limit that $G_{3n}(t)$ has.

Before proceeding to prove $G_{1n}(t) = o_p(1)$, an inequality about the moment generating function of a discrete random variable needs to be established.

Lemma 2.5.1 *For any positive integer n and for any random variable X taking values in $\{0, 1, 2, \ldots, n\}$ with probabilities $p_i = P\{X = i\}$, $i = 1, 2, \ldots, n$, there is always the relationship*

$$E\left\{e^{tX}\right\} \le \frac{\mu_n}{n} e^{nt} + 1 - \frac{\mu_n}{n}, \tag{2.5.10}$$

where $t \geq 0$ *and* $\mu_n = E(X) = 0p_0 + 1p_1 + 2p_2 + \ldots + np_n$ *is the mean of* X.

Proof. For $n = 1$, $\mu_1 = 0p_0 + 1p_1 = p_1$, then

$$E\left\{e^{tX}\right\} = p_0 + p_1 e^t = \frac{\mu_1}{1} e^t + 1 - \frac{\mu_1}{1},$$

hence assertion (2.5.10) holds for $n = 1$.

For $n = 2$, $\mu_2 = p_1 + 2p_2$, thus

$$\begin{aligned} E\left\{e^{tX}\right\} &= p_0 + p_1 e^t + p_2 e^{2t} \\ &= (\tfrac{1}{2}p_1 + p_2)e^{2t} + (1 - \tfrac{1}{2}p_1 - p_2) - \tfrac{1}{2}p_1 e^{2t} - (1 - \tfrac{1}{2}p_1 - p_2) \\ &\quad + p_0 + p_1 e^t \\ &= \tfrac{\mu_2}{2} e^{2t} + 1 - \tfrac{\mu_2}{2} - p_1 \left\{e^t[\tfrac{1}{2}e^t - 1] + \tfrac{1}{2}\right\}. \end{aligned}$$

Since $p_1\{e^t[\frac{1}{2}e^t - 1] + \frac{1}{2}\} \geq 0$, for $t \geq 0$, the assertion holds for $n = 2$.

Now suppose the assertion holds for $n = m$, that is,

$$p_0 + p_1 e^t + p_2 e^{2t} + \ldots + p_m e^{mt} \leq \frac{\mu_m}{m} e^{mt} + 1 - \frac{\mu_m}{m},$$

where $\mu_m = 0p_0 + 1p_1 + 2p_2 + \ldots + mp_m$. Let

$$S_m = p_0 + p_1 + p_2 + \ldots + p_m = 1 - p_{m+1},$$

$$\mu = 0\frac{p_0}{S_m} + 1\frac{p_1}{S_m} + \ldots + m\frac{p_m}{S_m},$$

then, with a slight abuse of the notation for μ_m,

$$\begin{aligned} E\left\{e^{tX}\right\} &= p_0 + p_1 e^t + \ldots + p_m e^{mt} + p_{m+1} e^{(m+1)t} \\ &= S_m \left(\tfrac{p_0}{S_m} + \tfrac{p_1}{S_m} e^t + \ldots + \tfrac{p_m}{S_m} e^{mt}\right) + p_{m+1} e^{(m+1)t} \\ &\leq S_m \left(\tfrac{\mu}{m} e^{mt} + 1 - \tfrac{\mu}{m}\right) + p_{m+1} e^{(m+1)t} \\ &= \tfrac{\mu_m}{m} e^{mt} + 1 - p_{m+1} - \tfrac{\mu_m}{m} + p_{m+1} e^{(m+1)t}. \end{aligned}$$

On the other hand,

$$\begin{aligned} & \tfrac{\mu_{m+1}}{m+1}e^{(m+1)t} + 1 - \tfrac{\mu_{m+1}}{m+1} \\ = \; & \tfrac{1}{m+1}\mu_m e^{mt}e^t + 1 - p_{m+1} - \tfrac{1}{m+1}\mu_m + p_{m+1}e^{(m+1)t}, \end{aligned}$$

hence the difference between the above expression and $E\{e^{tX}\}$ is less than or equal to

$$\mu_m \left\{ e^{mt}\left[\frac{1}{m+1}e^t - \frac{1}{m}\right] - \left(\frac{1}{m+1} - \frac{1}{m}\right)\right\} \geq 0$$

for $t \geq 0$, therefore, the assertion holds for $n = m + 1$. By the method of induction, the proof is completed. □

It is now ready to state the main result.

Theorem 2.5.1 *In the two-way layout (2.5.5), define, for* $i = 1, 2, \ldots, K$ *and* $j = 1, 2, \ldots, Q$,

$$e_{ij} = \{\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot j} + \bar{\varepsilon}_{\cdot\cdot}\}/(w_b\hat{\sigma}),$$

where $w_b = \sqrt{1 - 1/K - 1/Q + 1/(KQ)}$, *and* $\hat{\sigma}$ *is the square root of the mean squares due to error, then as* $Q \to \infty$, *the following empirical process*

$$Y_n(t) = (KQ)^{-1/2}\sum_{i=1}^{K}\sum_{j=1}^{Q}\{I[\Phi(e_{ij}) \leq t] - t\} \tag{2.5.11}$$

converges weakly to a Gaussian process with zero mean and covariance function given in (2.5.4).

Proof. Because the degrees of freedom for error is $df = (K-1)(Q-1)$ and $n/df = KQ/df \to K/(K-1)$, as $Q \to \infty$, the last term in (2.5.4) follows from an application of Theorem 2.3.1 and 2.3.2. As discussed before, it is then sufficient to show that $G_{1n}(t) = o_p(1)$ is true, where $G_{1n}(t)$ is defined after (2.5.9).

Since the terms $\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot}$ appearing in $G_{1n}(t)$ concern only the row effects ρ_i ($i = 1, 2, \ldots, K$, and ρ_i corresponds to θ_i of section 2.4.3), where K is fixed finite, and since

$w_a = \sqrt{1 - 1/K}$ is a constant, all arguments from the beginning of section 2.4.3 to equation (2.4.3) hold. In particular, equation (2.4.3) now becomes ($n = KQ$)

$$\begin{aligned} & \sup_{t\in I_s} \sup_{\xi\in C_k} R_n(t,\xi,\theta) \\ \leq\ & n^{-1/2} \sum_{i=1}^{K} \sum_{j=1}^{Q} \{I[u_{ij} \leq B_{ijk}(t_{s+1})] - B_{ijk}(t_{s+1}) - I[u_{ij} \leq t_s] + t_s\} \qquad (2.5.12) \\ & + o(\gamma,\lambda). \end{aligned}$$

However it should be remarked that the u_{ij} here are $U(0,1)$ random variables but are not independent; only the groups of u_{ij} indexed by j are independent. As in section 2.4.3, for the given $\zeta > 0$, choose γ and λ small such that $o(\gamma,\lambda) < \zeta/2$.

In order to obtain an appropriate exponential bound, the arguments in section 2.4.3 need to be modified because of the lack of independence among the u_{ij}. Specifically, for fixed k and s, define

$$\begin{aligned} & X_{ijks} = I[u_{ij} \leq B_{ijk}(t_{s+1})] - I[u_{ij} \leq t_s], \\ & p_{ijks} = |B_{ijk}(t_{s+1}) - t_s|, \\ & T_{+j} = \{i:\ B_{ijk}(t_{s+1}) - t_s \geq 0\},\ j = 1,2,\ldots,Q, \\ & T_{-j} = \{i:\ B_{ijk}(t_{s+1}) - t_s < 0\},\ j = 1,2,\ldots,Q, \\ & X^{+}_{jks} = \textstyle\sum_{i\in T_{+j}} X_{ijks}, \qquad p^{+}_{jks} = \sum_{i\in T_{+j}} p_{ijks}, \\ & X^{-}_{jks} = \textstyle\sum_{i\in T_{-j}} X_{ijks}, \qquad p^{-}_{jks} = \sum_{i\in T_{-j}} p_{ijks}. \end{aligned}$$

Then

$$\begin{aligned} & n^{-1/2} \textstyle\sum_{i=1}^{K} \sum_{j=1}^{Q} \{I[u_{ij} \leq B_{ijk}(t_{s+1})] - B_{ijk}(t_{s+1}) - I[u_{ij} \leq t_s] + t_s\} \\ =\ & n^{-1/2} \textstyle\sum_{j=1}^{Q} \{(X^{+}_{jks} - p^{+}_{jks}) - (X^{-}_{jks} - p^{-}_{jks})\}, \end{aligned}$$

therefore, equation (2.4.4) becomes

$$\begin{aligned} & P\left\{\sup_{t\in I_s} \sup_{\xi\in C_k} R_n(t,\xi,\theta) > \zeta\right\} \\ \leq\ & P\left\{\sum_{j=1}^{Q} (X^{+}_{jks} - p^{+}_{jks}) > \sqrt{n}\zeta/4\right\} \end{aligned}$$

$$
\begin{aligned}
&+ \quad P\left\{-\sum_{j=1}^{Q}(X^{-}_{jks} - p^{-}_{jks}) > \sqrt{n}\zeta/4\right\} \\
&= \quad p_{21} + p_{22}, \text{ say.}
\end{aligned} \tag{2.5.13}
$$

Now by applying Lemma 2.5.1 to the X^{+}_{jks}, which are independent with means p^{+}_{jks}, by Jenson's inequality, and by the fact that $\log(1+x) \leq x$ when $x \geq 0$, equation (2.4.5) becomes, for $t > 0$,

$$
\begin{aligned}
& n^{-1}\log a(n,k,s,t,\zeta) \\
= \quad & n^{-1}\log\left\{\exp\{-t\sqrt{n}\zeta/4\}\prod_{j=1}^{Q}E\left\{\exp(t(X^{+}_{jks} - p^{+}_{jks}))\right\}\right\} \\
\leq \quad & n^{-1}\left\{-t\sqrt{n}\zeta/4 + \sum_{j=1}^{Q}\left[-tp^{+}_{jks} + \log(1 + p^{+}_{jks}(e^{Kt}-1)/K)\right]\right\} \\
\leq \quad & -n^{-1/2}\left\{t\zeta/4 + n^{-1/2}\sum_{i=1}^{K}\sum_{j=1}^{Q}\left\{tp_{ijks} - p_{ijks}(e^{Kt}-1)/K\right\}\right\}.
\end{aligned} \tag{2.5.14}
$$

Because the presence of constant K does not affect the arguments after equation (2.4.5), the proof for $G_{1n}(t) = o_p(1)$ can be completed as in section 2.4.3. □

Corollary 2.5.1 *If there are S replications for each combination of row level and column level in model (2.5.5), that is,*

$$y_{ijk} = \mu + \rho_i + \tau_j + \sigma\varepsilon_{ijk}, \tag{2.5.15}$$

where $k = 1, 2, \ldots, S$ denotes replications, and define

$$e_{ijk} = \{\varepsilon_{ijk} - \bar{\varepsilon}_{i..} - \bar{\varepsilon}_{.j.} + \bar{\varepsilon}_{...}\}/(w_c\hat{\sigma}),$$

where

$$w_c = \left(1 - \frac{1}{KS} - \frac{1}{QS} + \frac{1}{KQS}\right)^{1/2},$$

and the bar over ε indicates averages over the subscripts represented by dots, then the following empirical process

$$Y_n(t) = (KQS)^{-1/2}\sum_{i=1}^{K}\sum_{j=1}^{Q}\sum_{k=1}^{S}\{I[\Phi(e_{ijk}) \leq t] - t\} \tag{2.5.16}$$

converges weakly, as $Q \to \infty$, to a Gaussian process with zero mean and covariance function given in (2.5.4) but with K there replaced by KS.

Proof. The only important change caused by the presence of replications is the replacement of K by KS in (2.5.4); other changes are only notational. □

When interactions are present, namely,

$$y_{ijk} = \mu + \rho_i + \tau_j + (\rho\tau)_{ij} + \sigma\varepsilon_{ijk}, \tag{2.5.17}$$

the corresponding residuals are $\varepsilon_{ijk} - \bar{\varepsilon}_{ij.}$, where $\bar{\varepsilon}_{ij.} = S^{-1}\sum_{k=1}^{S}\varepsilon_{ijk}$. This has essentially the same structure as in one-way layout (2.5.1), and if define

$$e_{ijk} = \{\varepsilon_{ijk} - \bar{\varepsilon}_{ij.}\}/(w_d\hat{\sigma}),$$

where $w_d = \sqrt{1 - 1/S}$, then the corresponding empirical process

$$Y_n(t) = (KQS)^{-1/2}\sum_{i=1}^{K}\sum_{j=1}^{Q}\sum_{k=1}^{S}\{I[\Phi(e_{ijk}) \le t] - t\} \tag{2.5.18}$$

converges weakly, as $Q \to \infty$, to a Gaussian process with zero mean and covariance function given in (2.5.4) but with K there replaced by S.

It is worth pointing out that in the above development, it is the structure of the (ordinary) residuals that determines the forms of the weights w_a to w_d and eventually determines the weak limits of the associated empirical processes. Here are two more examples.

Randomized Complete Block Design. This is a trivial example, as the residuals for this design are exactly the same as those given in (2.5.6) for model (2.5.5).

Two-factor Nested Design. Let y_{ijk} denote the k^{th} observation when factor A is at level i and factor B is nested within factor A and is at level j. The model expression is

$$y_{ijk} = \mu + \rho_i + \tau_{j(i)} + \varepsilon_{k(ij)}, \tag{2.5.19}$$

where μ is a constant,

ρ_i are constants subject to $\sum_{i=1}^{K}\rho_i = 0$,

$\tau_{j(i)}$ are constants subject to $\sum_{j=1}^{Q} \tau_{j(i)} = 0$, $i = 1, 2, \ldots, K$,

$\varepsilon_{k(ij)}$ are iid $N(0, \sigma^2)$.

The residuals for this design are $\varepsilon_{k(ij)} - \bar{\varepsilon}_{.(ij)}$, which are essentially of the same form for residuals of model (2.5.17), therefore, the empirical process of model (2.5.19) has the same weak limit as that of model (2.5.17).

2.5.3 Comments

For analysis of variance models, it is advantageous to standardize the (ordinary) residuals completely (using both $\hat{\sigma}$ and the diagonal elements of $I - H = I - X(X^tX)^{-1}X^t$. This is so because only the σ known case needs to be considered (see Theorems 2.3.1 and 2.3.2), and in this case the standardized residuals are (dependent) standard normal variables, which are easy to handle, and at the same time no complication arises when applying the change of time technique, since $1 - h_{ii}$ is constant.

Although a rigorous weak convergence result has not been established for the Latin square design and balanced incomplete block design, for all the models mentioned in this section, asymptotic critical points for EDF statistics W^2, U^2 and A^2 are available in Meester and Lockhart (1988).

Chapter 3

Contiguous Alternatives

Chapter 2 is centered at the study of residual empirical processes from fitting linear regression model (2.3.1), namely,

$$Y = X\theta + \sigma\varepsilon, \qquad \varepsilon \sim N_n(0, I).$$

In this study, the normality assumption of the error distribution plays a key role in deriving asymptotic distributions of the residual empirical processes, when the method of least squares is used to estimate the unknown regression parameters θ and σ. However, either the normality assumption is not met in some applications of linear regression techniques, or for some theoretical studies, such as the study of powers of a test, the error is deliberately assumed to have a non-normal distribution. It is then necessary to know: What will happen to the limit distributions of the residual empirical processes when the error in the above linear regression model takes on various non-normal distributions? This chapter is devoted to this question.

The approach based on the idea of contiguity is to be taken below; see Rao (1987), Hall and Loynes (1977). Section 3.1 introduces the concept of contiguity and some related results; section 3.2 studies six contiguous (to i.i.d. normal) alternatives, including general

normal distributions, Student t distributions, χ-distributions, Gamma distributions, Log-normal distributions, and Inverse Gaussian distributions; section 3.3 contains some remarks. The tools used in this chapter are well known; the conclusions in section 3.2 seem to be new.

3.1 Contiguity and Some Related Results

3.1.1 Contiguity

For $n = 1, 2, \ldots$, let $(\Omega_n, \mathcal{A}_n)$ be a sequence of measurable spaces, let $\{P_n\}$ and $\{Q_n\}$ be two sequences of probability measures on $(\Omega_n, \mathcal{A}_n)$.

Definition 4 (Contiguity of Q_n to P_n) *The sequence $\{Q_n\}$ is said to be contiguous to the sequence $\{P_n\}$, denoted by $Q_n \ll P_n$, if for any $A_n \in \mathcal{A}_n$,*

$$\lim_{n\to\infty} P_n(A_n) = 0 \text{ implies } \lim_{n\to\infty} Q_n(A_n) = 0.$$

Recall that a measure μ is said to be absolutely continuous to another measure ν if for any A, $\nu(A) = 0$ implies $\mu(A) = 0$. Thus, contiguity is a kind of asymptotic absolute continuity.

In this chapter, the underlying measurable space is fixed, namely, $(\Omega, \mathcal{A})$. However, to emphasize the dependence on sample size n, a subscript n appears in many quantities. For instance, the basic linear regression model is now written as

$$Y_{in} = x_{in}^t \theta + \sigma \varepsilon_{in}. \tag{3.1.1}$$

In the following discussion, the probability measure P_n is specified as the induced probability measure of the joint density of n independent random variables $Y_{1n}, Y_{2n}, \ldots, Y_{nn}$, where Y_{in} has density g_{in} (with respect to Lebesgue measure, say). This is denoted by P_n: $Y_{in} \sim_{indep} g_{in}$. Similarly, one can write Q_n: $Y_{in} \sim_{indep} f_{in}$. Since it is the distribution of ε_{in} that one has interest in, in this chapter, it is agreed to write P_n: $\varepsilon_{in} \sim_{indep} g_{in}$ and Q_n: $\varepsilon_{in} \sim_{indep} f_{in}$.

3.1.2 Le Cam's lemmas in perspective

Definition 5 (Log-likelihood Ratio) *If* P_n: $\varepsilon_{in} \sim_{indep} g_{in}$ *and* Q_n: $\varepsilon_{in} \sim_{indep} f_{in}$, *the likelihood ratio is defined by*

$$\Lambda_n = \begin{cases} \prod_{i=1}^n \frac{f_{in}(\varepsilon_{in})}{g_{in}(\varepsilon_{in})} & \text{if } g_{in}(\varepsilon_{in}) > 0 \text{ for all } i, \\ 1 & \text{if some } g_{in}(\varepsilon_{in}) = \text{some } f_{in}(\varepsilon_{in}) = 0, \\ \infty & \text{if all } f_{in}(\varepsilon_{in}) > 0, \text{ some } g_{in}(\varepsilon_{in}) = 0, \end{cases}$$

and the log-likelihood ratio is defined by $L_n = \log \Lambda_n$.

Note that under P_n, the probability of $\{\Lambda_n = \infty\}$ is zero. If asymptotic normality of $L_n = \log \Lambda_n$ is established, which will be done for the cases to be considered, then $P_n(\Lambda_n = 0) = o(1)$. Thus, without loss of generality, one can effectively write

$$L_n = \log \Lambda_n = \sum_{i=1}^n \log \frac{f_{in}(\varepsilon_{in})}{g_{in}(\varepsilon_{in})}. \tag{3.1.2}$$

As is well known, the log-likelihood ratio has many (asymptotic) optimal properties. So, if one can build up a (strong) relationship between one's statistic at hand and the log-likelihood ratio, one may get some good properties about one's statistic. This intuitive idea is formalized in the following version of Le Cam's third lemma: (Hall and Loynes (1977))

Let T_n be a statistic defined on $(\Omega, \mathcal{A})$, taking values in a metric space S. Let $M = S \times [-\infty, +\infty]$ be the image space of (T_n, L_n). If

(a) (T_n, L_n) converges weakly under P_n to a probability measure P, and

(b) Q_n is contiguous to P_n,

then (T_n, L_n) converges weakly to a probability measure Q under Q_n and $dQ(u, v) = \exp(v) dP(u, v)$.

Let $X_n(t)$ denote the residual empirical process of model (3.1.1) when θ and σ are estimated by the method of least squares. (To avoid ugly or repetitive notation, $X_n(t)$ is used. See section 2.5 for detail.) Consider now $T_n = X_n(t)$ and $S = D[0, 1]$. In order to be

able to compute probabilities of the limiting distributions of $X_n(t)$ under both P_n and Q_n, only Gaussian processes (for $X_n(t)$) and normal laws (for L_n) are considered as candidates for limits. With such a restriction, conditions under which the above mentioned (a) and (b) are satisfied become simple. For justifying contiguity in (b), one has (Rao (1987), section 1.11)

Le Cam's First Lemma: If $L_n = \log \Lambda_n \Rightarrow N(-\delta^2/2, \delta^2)$ for some $\delta > 0$ under P_n, then $Q_n \ll P_n$.

To deal with (a), consider tightness first. Note that $X_n(t)$ is tight on $D[0,1]$ under P_n: $\varepsilon_{in} \sim_{indep} N(0,1)$ and under conditions stated in Theorem 2.4.3; if $L_n = \log \Lambda_n \Rightarrow N(-\delta^2/2, \delta^2)$, then L_n is also tight on $[-\infty, +\infty]$. As a result, $(X_n(t), L_n)$ is tight on $D[0,1] \times [-\infty, +\infty]$ under P_n: $\varepsilon_{in} \sim_{indep} N(0,1)$ by Fact 4 of section 2.1. So the problem of tightness of $(X_n(t), L_n)$ is solved.

In order to satisfy (a), one needs to show further that $(X_n(t), L_n)$ converges to normal laws in finite distributions. This requirement depends heavily on the forms of the alternatives Q_n, and for this reason will be studied case by case in section 3.2. When this requirement is satisfied under P_n, one can use Le Cam's third lemma (the usual version) and Cramér-Wold device to find limiting distributions of $X_n(t)$ under Q_n. See Rao ((1987), section 1.11).

Le Cam's Third Lemma: For random variable S_n, if under P_n

$$\begin{pmatrix} S_n \\ L_n \end{pmatrix} \Rightarrow N_2 \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right]$$

and $\mu_2 = -\sigma_2^2/2$, then $S_n \Rightarrow N(\mu_1 + \sigma_{12}, \sigma_1^2)$ under Q_n.

3.1.3 A lemma for computations

In the course of looking for contiguous alternatives Q_n to P_n: $\varepsilon_{in} \sim_{indep} N(0,1)$, several integrals appear frequently. The following lemma lists the results of these integrals so that the work of the next section will become relatively easier and clearer.

Notations:

(a) $t \in [0,1]$, $x = \Phi^{-1}(t)$, $J_1(t) = \phi(\Phi^{-1}(t))$, $J_2(t) = \Phi^{-1}(t)J_1(t)$, $J_3(t) = \Phi^{-1}(t)J_2(t)$,

(b) $I[w_{in} \le t] = I[\varepsilon_{in} \le x]$, $w_{in} \sim_{iid} U(0,1)$, $\varepsilon_{in} = \Phi^{-1}(w_{in})$,

(c) $L_n = \hat{L}_n + o_p(1)$, here $\hat{L}_n$ is the leading part of L_n,

(d) $X_n(t) = \hat{X}_n(t) + o_p(1)$, where $\hat{X}_n(t)$ is the asymptotic expansion of $X_n(t)$ from Corollary 2.4.2 of Theorem 2.4.3, namely,

$$\begin{aligned}\hat{X}_n(t) &= n^{-1/2}\textstyle\sum_{i=1}^n \{I[w_{in} \le t]\} + J_1(t)n^{-1/2}\sum_{i=1}^n \varepsilon_{in} \\ &\quad + 2^{-1}J_2(t)n^{-1/2}\textstyle\sum_{i=1}^n(\varepsilon_{in}^2 - 1),\end{aligned}$$

(e) Under P_n, $X_n(t) \Rightarrow X(t)$ with zero mean and covariance function $\rho(s,t)$ given in equation (2.4.7) with $b = d = 1$; under Q_n, $X_n(t) \Rightarrow X^Q(t)$ with mean $A(t)$ and covariance function $\rho^Q(s,t)$.

Lemma 3.1.1 (Computation Lemma) *Under P_n: $\varepsilon_{in} \sim_{indep} N(0,1)$,*

(1) *Let $I_k = E\{\varepsilon_{in}^k I[\varepsilon_{in} \le x] \mid P_n\}$, $k = 1,2,\ldots$, then*

$$\begin{aligned}
I_0 &= \Phi(x) &&= t, \\
I_1 &= -\phi(x) &&= -J_1(t), \\
I_2 &= -x\phi(x) + \Phi(x) &&= -J_2(t) + t, \\
I_3 &= -x^2\phi(x) - 2\phi(x) &&= -J_3(t) - 2J_1(t), \\
I_k &= -x^{k-1}\phi(x) + (k-1)I_{k-2}, && k = 2,3,\ldots
\end{aligned}$$

(2)

$$\begin{aligned}
E\{[n^{-1/2}\textstyle\sum_{i=1}^n I[w_{in} \le t]][n^{-1/2}\sum_{j=1}^n \varepsilon_{jn}]\} &= -J_1(t), \\
E\{[n^{-1/2}\textstyle\sum_{i=1}^n I[w_{in} \le t]][n^{-1/2}\sum_{j=1}^n \varepsilon_{jn}^2]\} &= -J_2(t), \\
E\{[n^{-1/2}\textstyle\sum_{i=1}^n I[w_{in} \le t]][n^{-1/2}\sum_{j=1}^n \varepsilon_{jn}^3]\} &= -J_3(t) - 2J_1(t).
\end{aligned}$$

(3)

$$\begin{aligned} E\{[n^{-1/2}\textstyle\sum_{i=1}^n \varepsilon_{in}][n^{-1/2}\sum_{j=1}^n \varepsilon_{jn}]\} &= 1, \\ E\{[n^{-1/2}\textstyle\sum_{i=1}^n \varepsilon_{in}][n^{-1/2}\sum_{j=1}^n \varepsilon_{jn}^2]\} &= 0, \\ E\{[n^{-1/2}\textstyle\sum_{i=1}^n \varepsilon_{in}][n^{-1/2}\sum_{j=1}^n \varepsilon_{jn}^3]\} &= 3. \end{aligned}$$

(4)

$$\begin{aligned} E\{[n^{-1/2}\textstyle\sum_{i=1}^n (\varepsilon_{in}^2 - 1)][n^{-1/2}\sum_{j=1}^n \varepsilon_{jn}]\} &= 0, \\ E\{[n^{-1/2}\textstyle\sum_{i=1}^n (\varepsilon_{in}^2 - 1)][n^{-1/2}\sum_{j=1}^n \varepsilon_{jn}^2]\} &= 2, \\ E\{[n^{-1/2}\textstyle\sum_{i=1}^n (\varepsilon_{in}^2 - 1)][n^{-1/2}\sum_{j=1}^n \varepsilon_{jn}^3]\} &= 0. \end{aligned}$$

Proof. Straightforward computations. □

3.2 Contiguous Alternatives

According to the preparation of section 3.1, in order to show that the alternatives Q_n: $\varepsilon_{in} \sim_{indep} f_{in}$ are contiguous to the standard P_n: $\varepsilon_{in} \sim_{indep} N(0,1)$, and in order to find the limiting distributions of $(X_n(t), L_n)$ under Q_n, one needs to

(i) Show $L_n \Rightarrow N(-\delta^2/2, \delta^2)$ for some $\delta > 0$ under P_n;

(ii) Prove $(X_n(t), L_n) \Rightarrow (X(t), L)$ under P_n;

(iii) Find the mean and the covariance function of $(X(t), L)$;

(iv) Apply Le Cam's third lemma.

Notice that one can equivalently work with expansions $\hat{X}_n(t)$ and $\hat{L}_n$, which will make many computations a lot easier. Notice further that under the present circumstance, once (i) is shown to hold, (ii) will follow immediately as $\hat{X}_n(t)$ and $\hat{L}_n$ will both be sum of independent random variables, hence, convergence in finite distributions to normal laws is a natural consequence; tightness, on the other hand, is also inherited as discussed in section

3.1. Therefore, (ii) will be taken for granted in the following subsections. It is however important to remind the reader that conditions of Corollary 2.4.2 are assumed to hold for the remainder of this chapter.

3.2.1 Not iid normal alternatives

Let P_n: $\varepsilon_{in} \sim_{indep} N(0,1)$, let Q_n: $\varepsilon_{in} \sim_{indep} N(a_{in}, b_{in}^2)$. Then

$$\begin{aligned} \log \frac{f_{in}(\varepsilon_{in})}{g_{in}(\varepsilon_{in})} &= \log \frac{\phi[(\varepsilon_{in}-a_{in})/b_{in}]}{\phi(\varepsilon_{in})} \\ &= 2^{-1}(1-1/b_{in}^2)\varepsilon_{in}^2 + a_{in}\varepsilon_{in}/b_{in}^2 - 2^{-1}a_{in}^2/b_{in}^2. \end{aligned}$$

If $a_{in} = 1/\sqrt{n} + o(n)$, and $1/b_{in}^2 = n/(\sqrt{n}-1)^2 + o(n)$, where $o(n)$ may depend on i, one has

$$L_n = -1/2 + n^{-1/2}\sum_{i=1}^{n}(\varepsilon_{in} - \varepsilon_{in}^2) + o_p(1).$$

Since $E\{\varepsilon_{in} - \varepsilon_{in}^2 | P_n\} = -1$, $V\{\varepsilon_{in} - \varepsilon_{in}^2 | P_n\} = 3$, therefore,

$$L_n = \log \Lambda_n \Rightarrow N(-3/2, 3),$$

satisfying "$\mu = -\delta^2/2$".

For any fixed $t \in [0,1]$, under P_n, one has

$$\begin{pmatrix} \hat{X}_n(t) \\ \hat{L}_n \end{pmatrix} \Rightarrow N_2\left[\begin{pmatrix} 0 \\ -3/2 \end{pmatrix}, \begin{pmatrix} \rho(t,t) & B(t) \\ B(t) & 3 \end{pmatrix}\right],$$

where $B(t) = \lim_{n\to\infty} \mathrm{Cov}\{(\hat{X}_n(t), \hat{L}_n)|P_n\}$. By the computation lemma of section 3.1,

$$\begin{aligned} \mathrm{Cov}\{(\hat{X}_n(t), \hat{L}_n)|P_n\} &= E\{\hat{X}_n(t)\hat{L}_n\} \\ &= -J_1(t) + J_1(t) + 0 + J_2(t) + 0 - J_2(t) \\ &= 0. \end{aligned}$$

That is, $X_n(t)$ has the same weak limit $X(t)$ under P_n: $\varepsilon_{in} \sim_{indep} N(0,1)$ and under Q_n: $\varepsilon_{in} \sim_{indep} N(a_{in}, b_{in}^2)$, where $a_{in} = 1/\sqrt{n} + o(n)$, and $1/b_{in}^2 = n/(\sqrt{n}-1)^2 + o(n)$.

3.2.2 Contiguous Student's t alternatives

Let P_n: $\varepsilon_{in} \sim_{indep} N(0,1)$, let Q_n: $\varepsilon_{in} \sim_{indep} t_n(x)$, where $t_n(x)$ denotes the density of Student's t distribution with n degrees of freedom. Since $t_n(x) \to \phi(x)$ uniformly in x as $n \to \infty$, where $\phi(x)$ is the density of a $N(0,1)$ random variable, direct computations show that the same conclusion as that of the previous subsection holds.

3.2.3 Contiguous χ alternatives

Because $X \sim \chi^2_\nu$ implies $\sqrt{2X} - \sqrt{2\nu} \Rightarrow N(0,1)$ as $\nu \to \infty$, consider

$$P_n : \varepsilon_{in} \sim_{indep} N(0,1), \quad Q_n : \varepsilon_{in} \sim_{indep} \sqrt{2\chi^2_\nu} - \sqrt{2\nu}.$$

In this case,

$$\begin{aligned}
\log \tfrac{f_{in}(\varepsilon_{in})}{g_{in}(\varepsilon_{in})} &= \log\{2^{1-\nu}\Gamma^{-1}(\nu/2)(\varepsilon_{in}+\sqrt{2\nu})^{\nu-1}\exp[-(\varepsilon_{in}+\sqrt{2\nu})^2/4]\} \\
&\quad - \log\{(2\pi)^{-1/2}\exp(-\varepsilon_{in}^2/2)\} \\
&= \log\sqrt{2\pi} - (\nu-1)\log 2 - \log\Gamma(\nu/2) \\
&\quad + (\nu-1)\log(\varepsilon_{in}+\sqrt{2\nu}) - (\varepsilon_{in}-\sqrt{2\nu})^2/4 + \varepsilon_{in}^2/2 \\
&= \log\sqrt{2\pi} - (\nu-1)\log 2 - \log\Gamma(\nu/2) + (\nu-1)\log\sqrt{2\nu} - \nu/2 \\
&\quad - \varepsilon_{in}/\sqrt{2\nu} + \varepsilon_{in}^2/(4\nu) + [(\nu-1)/(6\nu)]\varepsilon_{in}^3/\sqrt{2\nu} \\
&\quad - [(\nu-1)/(16\nu)]\varepsilon_{in}^4/\nu + o_p(\varepsilon_{in}^4/\nu).
\end{aligned}$$

If $\nu = 2n$, $\Gamma(\nu/2) = \Gamma(n) = (n-1)!$. By Stirling's formula

$$n! = e^{a_n} n^{n+1/2} e^{-n}\sqrt{2\pi},$$

where $1/(12n+1) < a_n < 1/(12n)$, one has

$$\begin{aligned}
& n[\log\sqrt{2\pi} - (\nu-1)\log 2 - \log\Gamma(\nu/2) + (\nu-1)\log\sqrt{2\nu} - \nu/2] \\
= \;& n\{[(2n-1)/2]\log n - [(2n-1)/2]\log(n-1) - 1 - a_{n-1}\} \\
= \;& n\{[(2n-1)/2]\log[1+1/(n-1)] - 1 - a_{n-1}\} \\
= \;& [n(2n-1)/(2n-2)] - n - na_{n-1} + o(1) \\
\to \;& -1/12.
\end{aligned}$$

Also with $\nu = 2n$, the log-likelihood ratio L_n can be written as

$$L_n = n^{-1/2}\sum_{i=1}^{n}(\varepsilon_{in}^3/12 - \varepsilon_{in}/2) + n^{-1}\sum_{i=1}^{n}(\varepsilon_{in}^2/8 - \varepsilon_{in}^4/32) - 1/12 + o_p(1).$$

But $n^{-1}\sum_{i=1}^{n}(\varepsilon_{in}^2/8 - \varepsilon_{in}^4/32) \to_p 1/8 - 3/32 = 1/32$, so

$$L_n = -15/288 + n^{-1/2}\sum_{i=1}^{n}(\varepsilon_{in}^3/12 - \varepsilon_{in}/2) + o_p(1),$$

where $-15/288 = -1/12 + 1/32$. Since $E\{\varepsilon_{in}^3/12 - \varepsilon_{in}/2|P_n\} = 0$, $V\{\varepsilon_{in}^3/12 - \varepsilon_{in}/2|P_n\} = 15/144$, hence

$$L_n \Rightarrow N(-15/288, 15/144),$$

that is, "$\mu = -\delta^2/2$" holds for $\nu = 2n$.

Next,

$$\begin{aligned}\mathrm{Cov}(\hat{X}_n(t), \hat{L}_n) &= E\{\hat{X}_n(t)[n^{-1/2}\textstyle\sum_{i=1}^{n}(\varepsilon_{in}^3/12 - \varepsilon_{in}/2)]\} \\ &= \{-J_3(t) - 2J_1(t)\}/12 + 3J_1(t)/12 + J_1(t)/2 - J_1(t)/2 \\ &= J_1(t)/12 - J_3(t)/12,\end{aligned}$$

that is, under Q_n: $\varepsilon_{in} \sim_{indep} \sqrt{2\chi_{2n}^2} - \sqrt{4n}$, $X_n(t) \Rightarrow X^Q(t)$, where $X^Q(t)$ is a Gaussian process with mean $A(t) = J_1(t)/12 - J_3(t)/12$ and covariance function $\rho^Q(s,t) = \rho(s,t)$.

3.2.4 Contiguous Gamma alternatives

Suppose X has a Gamma distribution $Gamma(\alpha, \beta)$ with density

$$\beta^{-\alpha}\Gamma^{-1}(\alpha)x^{\alpha-1}\exp(-x/\beta),$$

where $0 < x$, $0 < \alpha$ and $0 < \beta$. Let

$$P_n : \varepsilon_{in} \sim_{indep} N(0,1), \quad Q_n : \varepsilon_{in} \sim_{indep} Gamma(\alpha, \beta) - \alpha\beta.$$

It follows that

$$\begin{aligned}\log \tfrac{f_{in}(\varepsilon_{in})}{g_{in}(\varepsilon_{in})} &= -\alpha\log\beta - \log\Gamma(\alpha) + (\alpha-1)\log(\varepsilon_{in}+\alpha\beta)\\ &\quad - \beta^{-1}(\varepsilon_{in}+\alpha\beta) + \log\sqrt{2\pi} + \varepsilon_{in}^2/2\\ &= \log\sqrt{2\pi} - \alpha\log\beta + (\alpha-1)\log(\alpha\beta) - \alpha - \log\Gamma(\alpha)\\ &\quad - (\alpha\beta)^{-1}\varepsilon_{in} + 2^{-1}[1-(\alpha-1)/(\alpha^2\beta^2)]\varepsilon_{in}^2 + 3^{-1}[(\alpha-1)/(\alpha^3\beta^3)]\varepsilon_{in}^3\\ &\quad - 4^{-1}[(\alpha-1)/(\alpha^4\beta^4)]\varepsilon_{in}^4 + o_p(\varepsilon_{in}^4/(\alpha^4\beta^4)).\end{aligned}$$

If $\alpha = n$, $\beta = 1/\sqrt{n}$, then

$$\begin{aligned}& n\{\log\sqrt{2\pi} - \alpha\log\beta + (\alpha-1)\log(\alpha\beta) - \alpha - \log\Gamma(\alpha)\}\\ =\ & n\{(n-1/2)\log[1+1/(n-1)] - 1 - a_{n-1}\} \qquad (na_{n-1}\to 1/12)\\ =\ & n(n-1/2)/(n-1) - n - na_{n-1} + o(1)\\ \to\ & -1/12.\end{aligned}$$

Also, for $\alpha = n$ and $\beta = 1/\sqrt{n}$,

$$2^{-1}[1-(\alpha-1)/(\alpha^2\beta^2)]\sum_{i=1}^{n}\varepsilon_{in}^2 \to_p 1/2,$$

$$-4^{-1}[(\alpha-1)/(\alpha^4\beta^4)]\sum_{i=1}^{n}\varepsilon_{in}^4 \to_p -3/4,$$

hence

$$L_n = -1/3 + n^{-1/2}\sum_{i=1}^{n}(\varepsilon_{in}^3/3 - \varepsilon_{in}) + o_p(1),$$

where $-1/3 = -1/12 + 1/2 - 3/4$. Because $E\{\varepsilon_{in}^3/3 - \varepsilon_{in}|P_n\} = 0$, and $V\{\varepsilon_{in}^3/3 - \varepsilon_{in}|P_n\} = 15/9 - 6/3 + 1 = 2/3$, it has been shown that

$$L_n \Rightarrow N(-1/3, 2/3),$$

and "$\mu = -\delta^2/2$" is true. For covariance, one has

$$\begin{aligned}\mathrm{Cov}(\hat{X}_n(t), L_n) &= E\{\hat{X}_n(t)[n^{-1/2}\textstyle\sum_{i=1}^{n}(\varepsilon_{in}^3/3 - \varepsilon_{in})]\}\\ &= \{-J_3(t) - 2J_1(t)\}/3 + J_1(t) + J_1(t) - J_1(t)\\ &= J_1(t)/3 - J_3(t)/3,\end{aligned}$$

that is, under Q_n: $\varepsilon_{in} \sim_{indep} Gamma(n, 1/\sqrt{n}) - \sqrt{n}$, $X_n(t) \Rightarrow X^Q(t)$, where $X^Q(t)$ is a Gaussian process with mean $A(t) = J_1(t)/3 - J_3(t)/3$ and covariance function $\rho^Q(s,t) = \rho(s,t)$.

3.2.5 Contiguous Lognormal alternatives

When X is distributed as Lognormal $L(\alpha, \beta)$ with density

$$(x\beta\sqrt{2\pi})^{-1} \exp\{-2^{-1}[(\log x - \alpha)/\beta]^2\},\ 0 < x,\ 0 < \beta,\ \alpha \in R,$$

$E(X) = \exp(\alpha + \beta^2/2)$, and $V(X) = \{\exp(\beta^2) - 1\}\exp(2\alpha + \beta^2)$. Let $b = E(X)$, then

$$Y = X - b \sim [(y+b)\beta\sqrt{2\pi}]^{-1} \exp\{-2^{-1}[(\log(y+b) - \alpha)/\beta]^2\},\ -b < y.$$

Consider now

$$P_n :\ \varepsilon_{in} \sim_{indep} N(0,1), \quad Q_{in} :\ \varepsilon_{in} \sim_{indep} L(\alpha,\beta) - b.$$

Note that

$$\begin{aligned} \log(\varepsilon_{in} + b) &= \alpha + \beta^2/2 + \log(1 + \varepsilon_{in}/b) \\ &= \alpha + \beta^2/2 + \varepsilon_{in}/b - 2^{-1}\varepsilon_{in}^2/b^2 + 3^{-1}\varepsilon_{in}^3/b^3 - 4^{-1}\varepsilon_{in}^4/b^4 + o_p(\varepsilon_{in}^4/b^4). \end{aligned}$$

Thus,

$$\begin{aligned} & -(2\beta^2)^{-1}[\log(\varepsilon_{in} + b) - \alpha]^2 \\ = \ & -(2\beta^2)^{-1}[\beta^2/2 + \varepsilon_{in}/b - 2^{-1}\varepsilon_{in}^2/b^2 + 3^{-1}\varepsilon_{in}^3/b^3 - 4^{-1}\varepsilon_{in}^4/b^4 + o_p(\varepsilon_{in}^4/b^4)]^2 \\ = \ & -(2\beta^2)^{-1}[\beta^4/4 + \beta^2\varepsilon_{in}/b + (1 - \beta^2/2)\varepsilon_{in}^2/b^2 + (\beta^2/3 - 1)\varepsilon_{in}^3/b^3 \\ & + (11/12 - \beta^2/4)\varepsilon_{in}^4/b^4 + o_p(\varepsilon_{in}^4/b^4)], \end{aligned}$$

therefore,

$$\begin{aligned} \log \tfrac{f_{in}(\varepsilon_{in})}{g_{in}(\varepsilon_{in}} &= -\beta^2/8 - 2^{-1}\varepsilon_{in}/b - (2\beta^2)^{-1}(1 - \beta^2/2)\varepsilon_{in}^2/b^2 - (2\beta^2)^{-1}(\beta^2/3 - 1)\varepsilon_{in}^3/b^3 \\ & - (2\beta^2)^{-1}(11/12 - \beta^2/4)\varepsilon_{in}^4/b^4 - (2\beta^2)^{-1}o_p(\varepsilon_{in}^4/b^4) \\ & - \alpha - \beta^2/2 - \varepsilon_{in}/b + 2^{-1}\varepsilon_{in}^2/b^2 - 3^{-1}\varepsilon_{in}^3/b^3 + 4^{-1}\varepsilon_{in}^4/b^4 - o_p(\varepsilon_{in}^4/b^4) \\ & - \log\beta + \varepsilon_{in}^2/2, \end{aligned}$$

and

$$\begin{aligned} L_n &= n\{-\alpha - \log\beta - 5\beta^2/8\} - (3/2)\textstyle\sum_{i=1}^n \varepsilon_{in}/b \\ &\quad + [3/4 - (2\beta^2)^{-1} + b^2/2]\textstyle\sum_{i=1}^n \varepsilon_{in}^2/b^2 - [1/2 - (2\beta^2)^{-1}]\textstyle\sum_{i=1}^n \varepsilon_{in}^3/b^3 \\ &\quad + [3/8 - 11(24\beta^2)^{-1}]\textstyle\sum_{i=1}^n \varepsilon_{in}^4/b^4 - \textstyle\sum_{i=1}^n [1 + (2\beta^2)^{-1}]o_p(\varepsilon_{in}^4/b^4). \end{aligned}$$

If $\alpha = \log n^{1/2}$, $\beta = n^{-1/2}$, then $b = \sqrt{n}\exp\{-1/(2n)\}$ and

$$n\{-\alpha - \log\beta - 5\beta^2/8\} = n(\log n^{1/2} - \log n^{1/2} - 5/(8n)) = -5/8,$$

$$[3/4 - (2\beta^2)^{-1} + b^2/2]\sum_{i=1}^n \varepsilon_{in}^2/b^2 \to_p 3/4 + 1/2 = 5/4,$$

$$[3/8 - 11(24\beta^2)^{-1}]\sum_{i=1}^n \varepsilon_{in}^4/b^4 \to_p 0 - 33/24 = -33/24,$$

hence

$$L_n = -3/4 + n^{-1/2}\sum_{i=1}^n \{\varepsilon_{in}^3/2 - 3\varepsilon_{in}/2\} + o_p(1),$$

where $-3/4 = -5/8 - 33/24 + 5/4$. Since $E\{\varepsilon_{in}^3/2 - 3\varepsilon_{in}/2|P_n\} = 0$, $V\{\varepsilon_{in}^3/2 - 3\varepsilon_{in}/2|P_n\} = 4^{-1}(15 - 18 + 9) = 3/2$, one has

$$L_n \Rightarrow N(-3/4, 3/2),$$

meeting the requirement that "$\mu = -\delta^2/2$". Moreover,

$$\begin{aligned} \mathrm{Cov}(\hat{X}_n(t), \hat{L}_n) &= E\{\hat{X}_n(t)[n^{-1/2}\textstyle\sum_{i=1}^n (\varepsilon_{in}^3/2 - 3\varepsilon_{in}/2)]\} \\ &= \{-J_3(t) - 2J_1(t)\}/2 + 3J_1(t)/2 + 3J_1(t)/2 - 3J_1(t)/2 \\ &= J_1(t)/2 - J_3(t)/2, \end{aligned}$$

proving that under Q_n: $\varepsilon_{in} \sim_{indep} L(\log n^{1/2}, n^{-1/2}) - \sqrt{n}\exp\{-1/(2n)\}$, $X_n(t) \Rightarrow X^Q(t)$, where $X^Q(t)$ is a Gaussian process with mean $A(t) = J_1(t)/2 - J_3(t)/2$ and covariance function $\rho^Q(s,t) = \rho(s,t)$.

3.2.6 Contiguous inverse Gaussian alternatives

The standard inverse Gaussian distribution has density

$$h(x,\alpha) = (2\pi)^{-1/2}[1+(\alpha x)/3]^{-3/2}\exp\{-(x^2/2)[1+(\alpha x)/3]^{-1}\},$$

where $0 < \alpha$, $-3/\alpha < x$. Note that as $\alpha \to 0$, $h(x,\alpha) \to \phi(x)$, where $\phi(x)$ is the density of a standard normal random variable. Let

$$P_n:\ \varepsilon_{in} \sim_{indep} N(0,1), \quad Q_n:\ \varepsilon_{in} \sim_{indep} h(\varepsilon_{in},\alpha).$$

Then

$$\begin{aligned}\log\tfrac{f_{in}(\varepsilon_{in})}{g_{in}(\varepsilon_{in})} &= (-3/2)\log[1+(\alpha\varepsilon_{in})/3] + (\varepsilon_{in}^2/2)\{1-[1+(\alpha\varepsilon_{in})/3]^{-1}\}\\ &= -\alpha\varepsilon_{in}/2 + 12^{-1}\alpha^2\varepsilon_{in}^2 + 6^{-1}\alpha\varepsilon_{in}^3 - 18^{-1}\alpha^2\varepsilon_{in}^4 + o_p(\alpha^2\varepsilon_{in}^2).\end{aligned}$$

Consider the case $\alpha = n^{-1/2}$, one has

$$L_n = n^{-1/2}\sum_{i=1}^{n}(\varepsilon_{in}^3/6 - \varepsilon_{in}/2) + n^{-1}\sum_{i=1}^{n}(\varepsilon_{in}^2/12 - \varepsilon_{in}^4/18) + o_p(1).$$

Since under P_n, $n^{-1}\sum_{i=1}^{n}(\varepsilon_{in}^2/12 - \varepsilon_{in}^4/18) \to_p -1/12$, $V(\varepsilon_{in}^3/6 - \varepsilon_{in}/2) = 1/6$, thus

$$L_n \Rightarrow N(-1/12, 1/6),$$

satisfying "$\mu = -\delta^2/2$" again. As

$$\begin{aligned}\mathrm{Cov}(\hat{X}_n(t), \hat{L}_n) &= E\{\hat{X}_n(t)[n^{-1/2}\textstyle\sum_{i=1}^{n}(\varepsilon_{in}^3/6 - \varepsilon_{in}/2)]\}\\ &= J_1(t)/2 - J_1(t)/2 - J_3(t)/6 - J_1(t)/3 + J_1(t)/2\\ &= J_1(t)/6 - J_3(t)/6,\end{aligned}$$

it has been proved that under Q_n: $\varepsilon_{in} \sim_{indep} h(\varepsilon_{in}, n^{-1/2})$, $X_n(t) \Rightarrow X^Q(t)$, where $X^Q(t)$ is a Gaussian process with mean $A(t) = J_1(t)/6 - J_3(t)/6$ and covariance function $\rho^Q(s,t) = \rho(s,t)$.

3.3 Comments

It seems to be magic that only for some clever choices of alternatives that $X_n(t)$ converges weakly. One may think that out of all possible alternatives under which $X_n(t)$ converges weakly, contiguous alternatives are just a (small) portion. However, the truth is: when alternatives are concerned, $X_n(t)$ converges weakly under alternatives Q_n implies that both Q_n is contiguous to P_n and $X_n(t)$ converges weakly under P_n. See Hall and Loynes (1977) for more details.

Chapter 4

EDF Statistics and Overall Test of Fit

Any statistical model is based on some assumptions. The linear regression model (2.1.1) is, roughly speaking, based on six assumptions listed in section 2.1.1, namely, linearity (in parameter) and additivity (in errors), independence of the errors, homoscedasticity of the errors, normality of the errors, error-free covariates, and full-rank model matrix. To check the adequacy of these assumptions, many diagnostic statistics have been invented. See Belsley, Kuh and Welsch (1980), Cook and Weisberg (1982).

This chapter considers using EDF statistics to check goodness-of-fit when fitting linear models. Section 4.1 introduces two families of EDF statistics. Section 4.2 lists some key facts about EDF statistics and explains how a tail probability is computed in this thesis. Section 4.3 studies the possibility of using EDF statistics to assess the overall fit of a linear model to a given data set. A point discussed seriously here is that in linear regression analysis, tests based on EDF statistics are not quite equivalent to tests of normality, as is usually thought to be so. Section 4.4 summarizes the chapter.

4.1 An Introduction to EDF Statistics

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a continuous distribution $F(x;\theta)$, where $\theta \in \Theta \subset R^p$ is either known or needs to be estimated. For a fixed θ, applying the probability integral transformation to the X-sample gives a U-sample $U_1, U_2, \ldots, U_n$, where $U_i = F(X_i;\theta), i = 1, 2, \ldots, n$. When θ is the true parameter for the X-sample, $U_1, U_2, \ldots, U_n$ will be an independent and identically distributed sample from the uniform $U(0,1)$.

Now let $F_n(x)$ denote the empirical distribution function (EDF) of the X-sample, that is,

$$F_n(x) = n^{-1} \sum_{i=1}^{n} I[x_i \leq x],$$

where $I[a \leq b] = 1$ if $a \leq b$, and $I[a \leq b] = 0$ if $a > b$. **Any statistic that measures the discrepancy between F_n and F will be called an EDF statistic.** There are mainly two classes of EDF statistics.

4.1.1 The supremum EDF statistics

Let $\hat{\theta}$ be an estimator of θ, the supremum EDF statistics are based on the maximum difference between F_n and F: the *Kolmogorov-Smirnov statistics* are

$$D^+ = \sup_{-\infty<x<+\infty} F_n(x) - F(x;\hat{\theta}), \tag{4.1.1}$$

$$D^- = \sup_{-\infty<x<+\infty} F(x;\hat{\theta}) - F_n(x), \tag{4.1.2}$$

$$D = \sup_{-\infty<x<+\infty} |F_n(x) - F(x;\hat{\theta})| = \max(D^+, D^-), \tag{4.1.3}$$

and the *Kuiper statistic*, which is designed for data on a circle, is

$$V = D^+ + D^-. \tag{4.1.4}$$

If θ is completely specified as θ_0, then θ_0 is used to replace $\hat{\theta}$ in above expressions.

4.1.2 The integral EDF statistics

The integral EDF statistics, also known as quadratic statistics, are based on the weighted and integrated squared discrepancies between F_n and F, namely,

$$Q = n \int_{-\infty}^{+\infty} \{F_n(x) - F(x;\hat{\theta})\}^2 \psi(x) dF(x;\hat{\theta}),$$

where $\psi(x) \geq 0$ is a suitable weight function. As special cases, the *Cramér-von Mises statistic* is obtained when $\psi(x) = 1$,

$$W^2 = n \int_{-\infty}^{+\infty} \{F_n(x) - F(x;\hat{\theta})\}^2 dF(x;\hat{\theta}), \tag{4.1.5}$$

and the *Anderson-Darling statistic* is obtained when $\psi(x) = \{F(x;\hat{\theta})(1 - F(x;\hat{\theta}))\}^{-1}$,

$$A^2 = n \int_{-\infty}^{+\infty} \{F_n(x) - F(x;\hat{\theta})\}^2 \{F(x;\hat{\theta})(1 - F(x;\hat{\theta}))\}^{-1} dF(x;\hat{\theta}). \tag{4.1.6}$$

A variation, which has been devised for data on a circle, is the *Watson statistic* given by

$$U^2 = n \int_{-\infty}^{+\infty} \left\{ F_n(x) - F(x;\hat{\theta}) - \int_{-\infty}^{+\infty} [F_n(x) - F(x;\hat{\theta})] dF(x;\hat{\theta}) \right\}^2 dF(x;\hat{\theta}). \tag{4.1.7}$$

Again, if θ is completely specified as θ_0, then θ_0 is used to replace $\hat{\theta}$ in above expressions.

4.1.3 Computation formulas for EDF statistics

For a given (observed) X-sample $x_1, x_2, \ldots, x_n$, let $u_i = F(x_i;\hat{\theta})$, $i = 1, 2, \ldots, n$. Without loss of generality, suppose x_i's and u_i's have been arranged into ascending order. Then the above EDF statistics can be easily computed as (Stephens (1986))

$$D^+ = \max_{1 \leq i \leq n} \left(\frac{i}{n} - u_i \right), \tag{4.1.8}$$

$$D^- = \max_{1 \leq i \leq n} \left(u_i - \frac{i-1}{n} \right), \tag{4.1.9}$$

$$D = \max(D^+, D^-), \tag{4.1.10}$$

$$V = D^+ + D^-, \tag{4.1.11}$$

$$W^2 = \sum_{i=1}^{n} \left\{ u_i - \frac{2i-1}{2n} \right\}^2 + \frac{1}{12n}, \tag{4.1.12}$$

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \{(2i-1) \ln u_i + (2n+1-2i) \ln(1-u_i)\}, \tag{4.1.13}$$

$$U^2 = W^2 - n \left\{ 0.5 - n^{-1} \sum_{i=1}^{n} u_i \right\}^2. \tag{4.1.14}$$

In the following discussion, D^+, D^-, and V are not included. For general properties and relative merits of the EDF statistics introduced here, see Stephens (1986).

4.1.4 Stephens' procedure for testing normality

To test H_0^N: $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$, where μ and σ^2 are both unkown, proceed as below:

(a) Compute $w_i = (x_i - \bar{x})/s$, where $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ and $s^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$,

(b) Compute $u_i = \Phi(w_i)$, where $\Phi(w)$ is the cumulative distribution function of a standard normal random variable with zero mean and unit variance,

(c) Calculate D, V, W^2, U^2 and A^2 according to (4.1.10), and (4.1.12) to (4.1.14),

(d) Modify D into $D^* = D(\sqrt{n} - 0.01 + 0.85/\sqrt{n})$, W^2 into $W^* = W^2(1 + 0.5/n)$, U^2 into $U^* = U^2(1+0.5/n)$, and A^2 into $A^* = A^2(1+0.75/n+2.25/n^2)$, where n is the sample size, and reject H_0^N at significance level α if the modified statistics exceed the upper tail significance points given in Table 4.7, D'Agostino and Stephens (1986), page 123. A portion of this table is given below as Table 4.1 for ease of reference.

4.2 Some Key Facts About EDF Statistics

Intuitively speaking, large values of EDF statistics indicate that the data at hand do not come from the distribution used to calculate the EDF statistics. There is then one important

Modified Statistics	Significance Level α					
	0.50	0.25	0.15	0.10	0.05	0.01
$D^* = D(\sqrt{n} - 0.01 + 0.85/\sqrt{n})$	0.601	0.708	0.775	0.819	0.895	1.035
$W^* = W^2(1 + 0.5/n)$	0.051	0.074	0.091	0.104	0.126	0.179
$U^* = U^2(1 + 0.5/n)$	0.048	0.070	0.085	0.096	0.117	0.164
$A^* = A^2(1 + 0.75/n + 2.25/n^2)$	0.341	0.470	0.561	0.631	0.752	1.035

Table 4.1: Upper tail significance points for EDF tests of normality when both mean μ and variance σ^2 are estimated. The tests reject normality if the modified EDF statistics exceed the table entries.

thing that is not discussed in the previous section, that is, how to compute tail probabilities of the EDF statistics. This section will list some key theoretical results about EDF statistics together with numerical methods for computing tail probabilities. Attention will be given only to integral type EDF statistics.

4.2.1 Orthogonal representation of stochastic processes

Let $\{X_n(t) : 0 \leq t \leq 1\}$ be a sequence of generic stochastic process. Let $\{X(t) : 0 \leq t \leq 1\}$ be the weak limit of $X_n(t)$ and denote the covariance function of $X(t)$ by $k(s,t)$, that is,

$$k(s,t) = \mathrm{Cov}(X(s), X(t)),\ 0 \leq s,\ t \leq 1,$$

which is symmetric in s and t. If there are real number λ and real function $f(t)$ such that

$$\int_0^1 f(s)k(s,t)ds = \lambda f(t), \tag{4.2.1}$$

then λ is called an eigenvalue of $k(s,t)$, and $f(t)$ is called an eigenfunction of $k(s,t)$ associated with λ.

Now let $\mathcal{L}^2 = \mathcal{L}^2[0,1] = \{g : [0,1] \to R \mid \int_0^1 g^2(t)dt < \infty\}$ and denote the usual inner product on $\mathcal{L}^2$ by $< f, g >= \int_0^1 f(t)g(t)dt$. With this inner product, $(\mathcal{L}^2, < \cdot, * >)$ becomes

a Hilbert space. For notational convenience, it is agreed to write $Z \sim (a, b)$ to mean that the random variable Z has mean a and variance b.

The basic idea here is to represent $X(t)$ in terms of orthogonal components. This was essentially done in Kac and Siegert (1947):

Given a non-null zero mean process $\{X(t): 0 \leq t \leq 1\}$. If

(1) The paths of $X(t)$ are a.s. in a subspace L of $\mathcal{L}^2$ and $k(s,t)$ is continuous,

(2) $\int_0^1 k(t,t)dt < \infty$,

(3) The eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ of $k(s,t)$ satisfy $\lambda_1 \geq \lambda_2 \geq \ldots > 0$,

(4) The associated orthonormal eigenfunctions $\{f_i(t)\}_{i=1}^{\infty}$ form a complete set for the subspace L,

then

1. $k(s,t) = \sum_{i=1}^{\infty} \lambda_i f_i(s) f_i(t)$, $0 \leq s,\ t \leq 1$, where the convergence on the right side is both absolute and uniform,
2. $\sum_{i=1}^{\infty} \lambda_i < \infty$,
3. $Z_i =< X, f_i > \sim (0, \lambda_i)$ and are uncorrelated,
4. $X(t) \overset{a.s.}{=} \sum_{i=1}^{\infty} \sqrt{\lambda_i} f_i(t) Z_i^*$, where $Z_i^* = Z_i/\sqrt{\lambda_i} \sim (0,1)$ and are uncorrelated,
5. $T = \int_0^1 X^2(t)dt \overset{a.s.}{=} \sum_{i=1}^{\infty} Z_i^2 \overset{a.s.}{=} \sum_{i=1}^{\infty} \lambda_i (Z_i^*)^2$.

4.2.2 χ^2 representation of integral EDF statistics

For Gaussian processes satisfying the conditions (1) to (4) listed above, the uncorrelated Z_i^*'s are actually independent and identically distributed $N(0,1)$ random variables, leading to a χ^2 representation of $T = \int_0^1 X^2(t)dt$, which is a generic integral EDF statistic.

In fact, more can be said about this χ^2 representation. Suppose that $Y(t)$ is Gaussian and satisfies conditions (1) to (4) except that the mean of $Y(t)$ is not zero. For $s, t \in [0,1]$, let

$$\begin{aligned} A(t) &= E(Y(t)), \\ l(s,t) &= \mathrm{Cov}(Y(s), Y(t)), \\ X(t) &= Y(t) - A(t), \\ k(s,t) &= \mathrm{Cov}(X(s), X(t)). \end{aligned}$$

Since $l(s,t) \equiv k(s,t)$, the eigenvalues and eigenfunctions of $k(s,t)$ are the same as those of $l(s,t)$. Let

$$A(t) = \sum_{i=1}^{\infty} a_i f_i(t),$$

where $a_i =< A, f_i >$, $i = 1, 2, \ldots$. Because $X(t)$ has zero mean, from

$$X(t) \stackrel{a.s.}{=} \sum_{i=1}^{\infty} \sqrt{\lambda_i} f_i(t) Z_i^*, \; Z_i^* \sim_{iid} N(0,1),$$

one has

$$\begin{aligned} Y(t) &= X(t) + A(t) \\ &= \textstyle\sum_{i=1}^{\infty} (a_i + \sqrt{\lambda_i} Z_i^*) f_i(t) \\ &= \textstyle\sum_{i=1}^{\infty} (a_i + Z_i) f_i(t) \\ &= \textstyle\sum_{i=1}^{\infty} \sqrt{\lambda_i} f_i(t) Z_i^{**}, \end{aligned}$$

where $Z_i^{**} = (a_i + Z_i)/\sqrt{\lambda_i} \sim_{indep} N(a_i/\sqrt{\lambda_i}, 1)$, therefore,

$$\begin{aligned} T &= \int_0^1 Y^2(t) dt \\ &= \sum_{i=1}^{\infty} \lambda_i (Z_i^{**})^2 \\ &= \sum_{i=1}^{\infty} \lambda_i \chi^2_{i,\delta_i^2}, \end{aligned} \tag{4.2.2}$$

where χ^2_{i,δ_i^2} are independent non-central χ^2 random variables on 1 degree of freedom and with non-centrality $\delta_i^2 = a_i^2/\lambda_i$. See Shorack and Wellner (1986).

4.2.3 Calculation of percentage points

As in previous section, let

$$T = \int_o^1 Y^2(t)dt$$

be a generic integral EDF statistic. To find $P\{T > x\}$, $x \geq 0$, three (numerical) methods will be described here. Suppose the representation of T given in (4.2.2) holds.

Smirnov's Method The characteristic function of T is known as

$$\varphi_T(t) = \prod_{j=1}^{\infty}(1 - 2i\lambda_j t)^{-1/2} \exp\left\{i\sum_{j=1}^{\infty}\left(\frac{\delta_j^2\lambda_j t}{1 - 2i\lambda_j t}\right)\right\}. \tag{4.2.3}$$

Let

$$D(u) = \prod_{j=1}^{\infty}(1 - \lambda_j u)^{-1/2} \exp\left\{\frac{1}{2}\sum_{j=1}^{\infty}\left(\frac{\delta_j^2\lambda_j u}{1 - \lambda_j u}\right)\right\}. \tag{4.2.4}$$

Then, $\varphi_T(t) = D(2it)$, and

$$P\{T > x\} = \frac{1}{\pi}\sum_{k=1}^{\infty}(-1)^{k+1}J_k(x), \tag{4.2.5}$$

where

$$J_k(x) = \int_{\frac{1}{2\lambda_{2k-1}}}^{\frac{1}{2\lambda_{2k}}} \frac{1}{u}e^{-ux}|D(2u)|du. \tag{4.2.6}$$

See Schilling (1983) and the references therein.

Pearson's Method Let κ_m denote the m^{th} cumulant of T, let μ_m denote the m^{th} central moment of T, and let μ denote the mean of T. If representation (4.2.2) is true, one has (Anderson and Darling (1952))

$$\kappa_m = 2^{m-1}(m-1)!\sum_{j=1}^{\infty}\lambda_j^m(1 + m\delta_j^2),\ m = 1, 2, \ldots \tag{4.2.7}$$

In particular,

$$\begin{aligned}
\kappa_1 &= \textstyle\sum_{j=1}^{\infty}\lambda_j(1 + \delta_j^2),\\
\kappa_2 &= 2\textstyle\sum_{j=1}^{\infty}\lambda_j^2(1 + 2\delta_j^2),\\
\kappa_3 &= 8\textstyle\sum_{j=1}^{\infty}\lambda_j^3(1 + 3\delta_j^2),\\
\kappa_4 &= 48\textstyle\sum_{j=1}^{\infty}\lambda_j^4(1 + 4\delta_j^2).
\end{aligned}$$

Equivalently, one has

$$\begin{aligned}\mu &= \kappa_1,\\ \mu_2 &= \kappa_2,\\ \mu_3 &= \kappa_3,\\ \mu_4 &= \kappa_4 + 3\kappa_2^2.\end{aligned}$$

Pearson's method uses the first four moments to build a density $f(x)$ which can be used to approximate $P\{T > x\}$. To this end, define (Kendall and Stuard (1977), Vol. 1, chapter 6)

$$b_0 = -\mu_2(4\mu_2\mu_4 - 3\mu_3^2)/A, \tag{4.2.8}$$

$$b_1 = -\mu_3(\mu_4 + 3\mu_2^2)/A, \tag{4.2.9}$$

$$b_2 = -(2\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3)/A, \tag{4.2.10}$$

where $A = 10\mu_2\mu_4 - 18\mu_2^3 - 12\mu_3^2$. Let $\Delta = (b_1^2 - 4b_0b_2)^{1/2}$, let $k = b_1^2/(4b_0b_2)$. Then $k > 1$, which is the case in this thesis, implies that $f(x)$ belongs to Pearson type VI curves. A computationally simple form for $f(x)$ is obtained in terms of the Beta distribution as below. Define

$$p = (1 - b_2)/b_2, \tag{4.2.11}$$

$$q = \{1 + (b_1 + 2b_1b_2)/\Delta\}/(2b_2), \tag{4.2.12}$$

$$y = -2\Delta/\{2b_2(x + \mu) - \Delta + b_1\}, \tag{4.2.13}$$

then,

$$P\{T > x\} = \frac{1}{B(p,q)}\int_0^y u^{p-1}(1-u)^{q-1}du, \tag{4.2.14}$$

where $B(p,q) = \int_0^1 u^{p-1}(1-u)^{q-1}du$.

Imhof's Method Assume again that representation (4.2.2) holds. Imhof (1961) found that

$$P\{T > x\} = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{sin(\theta(u))}{u\rho(u)}du, \tag{4.2.15}$$

where

$$\theta(u) = -\frac{1}{2}xu + \frac{1}{2}\sum_{j=1}^{\infty}\left\{tan^{-1}(\lambda_j u) + \delta_j^2\lambda_j u(1+\lambda_j^2u^2)^{-1}\right\}, \quad (4.2.16)$$

$$\rho(u) = \prod_{j=1}^{\infty}(1+\lambda_j^2u^2)^{1/4}\exp\left\{\frac{1}{2}\sum_{j=1}^{\infty}\frac{(\delta_j\lambda_j u)^2}{1+\lambda_j^2u^2}\right\}, \quad (4.2.17)$$

$$\lim_{u\to 0}\frac{sin(\theta(u))}{u\rho(u)} = -\frac{1}{2}x + \frac{1}{2}\sum_{j=1}^{\infty}\lambda_j(1+\delta_j^2). \quad (4.2.18)$$

It is remarked that there are many other methods of computing (tail) probabilities associated with linear (possibly infinite) combinations of χ^2 random variables. See Kotz, Johnson and Boyd (1967) and Davies (1973). Of the three methods described here, Smirnov's method is found inefficient for the EDF statistics studied in this thesis. For example, for the Cramér-von Mises statistic W^2, the eigenvalues are no greater than $1/(j\pi)^2$, $j = 1, 2, \ldots$ (Sukhatme (1972)). Therefore, the intervals $[1/(2\lambda_{2k-1}), 1/(2\lambda_{2k})]$, $k = 1, 2, \ldots$, become wider and wider. Pearson's method is found to work fairly well, in particular, it is very fast, provided an incomplete Beta routine is available. Imhof's method seems to be one of the most general methods used in the present context. When a good numerical routine for improper integrals is available (as for example, the routines in the IMSL Library), Imhof's method works efficiently and accurately. The following gives implementation details for Imhof's method applied to find probabilities under the null hypothesis, where only central χ^2 random variables are involved. The approach used here tries to correct the mean only, although it is not very difficult to correct both mean and variance.

Let $k(s,t)$, $0 \le s,\ t \le 1$ be the covariance function for an integral EDF statistic T. Discretize the integral equation (see equation (4.2.1))

$$\int_0^1 f(s)k(s,t)ds = \lambda f(t)$$

into

$$\frac{1}{m}\sum_{j=1}^{m} f_j k\left(\frac{i-0.5}{m}, \frac{j-0.5}{m}\right) = \lambda f_i,\ (i = 1, \ldots, m)$$

for a large integer m ($m = 150$ is used in this thesis) and solve for eigenvalues $\hat{\lambda}_i$ ($i = 1, \ldots, m$). The distribution of $\sum_{i=1}^{\infty} \lambda_i \chi_i^2$ can be approximated by $\sum_{i=1}^{m} \hat{\lambda}_i \chi_i^2 + \tau \chi_{m+1}^2$, where χ_{m+1}^2 is a chi-square random variable on 1 degree of freedom and is independent of the χ_i^2 ($i = 1, \ldots, m$), and τ is found by making

$$\int_0^1 k(t,t)dt = \sum_{i=1}^{\infty} \lambda_i = (\sum_{i=1}^{m} \hat{\lambda}_i) + \tau$$

true. For example, when testing normality using W^2, where both mean and variance are estimated from the data, one has $\sum_{i=1}^{\infty} \lambda_i = 0.0492385$ and $\sum_{i=1}^{m} \hat{\lambda}_i = 0.0492413$, so $\tau = -2.8\text{e-}06$. Equations (4.2.15) to (4.2.18) can then be used to approximate $P\{W^2 > x\}$.

4.3 EDF Tests of Overall Fit for Linear Regression

4.3.1 Overall test of fit

Consider the linear regression model

$$Y = X\theta + \sigma\varepsilon = 1_n\alpha + V\beta + \sigma\varepsilon, \quad \varepsilon \sim N_n(0, I_n), \tag{4.3.1}$$

where 1_n denotes an n by 1 column of 1's, and I_n is the n by n identity matrix. The usual overall test of fit is the F-test of H_0: $\beta = 0$ vs H_1: $\beta \neq 0$. When H_0 is true, model (4.3.1) simply says that $Y = (Y_1, Y_2, \ldots, Y_n)^t$ is a random sample from $N(\alpha, \sigma^2)$, thus tests of normality with unknown mean and unknown variance based on EDF statistics (see section 4.1.4) can be used to test H_0, too. But when H_0 is rejected, the (usual) implication of the F-test is quite different from the (usual) implication by EDF statistics tests.

For the F-test, when H_0 is rejected, the usual implication or interpretation is, among other things, that it is worth including regressors other than the grand mean. One does not question the validity of assuming the errors to be i.i.d. $N(0, \sigma^2)$.

For EDF tests, when H_0 is rejected, the usual implication or interpretation is, among other things, that one may not be able to assume that the errors are i.i.d. $N(0, \sigma^2)$. Although this implication includes the case where the errors are independent normal with

equal variances but unequal means ($\beta \neq 0$), in general, one thinks about more general departures such as unequal variances, non-normality, or dependent errors.

In fact, most regression diagnostic statistics focus on a specific problem area. When a problem is identified, the conclusion usually has local implication pointing towards either a violation of one basic assumption, or influential behaviour of a few points. In particular, EDF statistics are usually employed to test for normality of the errors.

In this section, however, the use of EDF statistics for testing overall fit is explored. The phrase overall fit here means that all the key assumptions of normal theory linear regression are assessed in a comprehensive manner. There is nothing new in carrying out the usual EDF statistics tests—the same EDF statistics are used according to the same tables available, but the tests are no longer regarded solely as tests for normality. The reason for this change of point of view is simple: for the weak convergence results of chapter 2 to hold (which are the basis of all EDF tests in the context of normal theory linear regression), normality is one of the many assumptions, therefore, when EDF statistics tests reject H_0, the cause for the rejection could be any or several of the possibilities listed in section 2.2, Table 2.1. In fact, Theorem 2.4.3 allows a more general hypothesis to be tested by EDF tests.

To test H_0^C: the assumptions of the linear regression model (4.3.1) are all satisfied, proceed as below:

(a) Obtain least squares estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\sigma}$,

(b) Compute $u_i = \Phi\{(y_i - \hat{\alpha} - x_i^t\hat{\beta})/\hat{\sigma}\}$, where $\Phi(w)$ is the distribution function of a $N(0,1)$ random variable,

(c) Calculate D, V, W^2, U^2 and A^2 according to (4.1.10), and (4.1.12) to (4.1.14),

(d) Modify D into $D^* = D(\sqrt{n} - 0.01 + 0.85/\sqrt{n})$, W^2 into $W^* = W^2(1 + 0.5/n)$, U^2 into $U^* = U^2(1+0.5/n)$, and A^2 into $A^* = A^2(1+0.75/n+2.25/n^2)$, where n is the sample size, and reject H_0^C at significance level α if the modified statistics exceed the upper

tail significance points given in Table 4.7, D'Agostino and Stephens (1986), page 123. A portion of this table is given in Table 4.1 of section 4.1 for ease of reference.

4.3.2 Examples

One question not touched so far is: At what level should the EDF statistics tests be performed to test H_0^C? Notice that each assumption for model (4.3.1) as listed in Table 2.1 is a part of H_0^C, so to test H_0^C by performing one EDF statistics test, the significance level should be set large. Experience from doing EDF statistics tests suggests that a P-value of 0.25 or larger indicates that H_0^C is reasonable. This is of course a rule of thumb and exceptions are not difficult to find, as will be shown later in this section.

Before presenting numerical examples, some commonly used regression model selection criteria are listed here for notational reference.

- $R^2 = 1 - \frac{SS(Residual)}{SS(Total)} = 1 - \frac{(Y-\hat{Y})^t(Y-\hat{Y})}{(Y-\bar{Y})^t(Y-\bar{Y})}$,
- $s^2 = \frac{(Y-\hat{Y})^t(Y-\hat{Y})}{n-p}$, where $p = dim(\theta)$,
- PRESS= $\sum_{i=1}^{n}(y_i - \hat{y}_{i,-i})^2$, where $\hat{y}_{i,-i}$ is the least squares fitted value when the i^{th} observation y_i is excluded from the estimation process,
- $R^2_{Pred} = 1 - \frac{PRESS}{(Y-\bar{Y})^t(Y-\bar{Y})}$,
- $\sum_{i=1}^{n}|e_{i,-i}| = \sum_{i=1}^{n}|y_i - \hat{y}_{i,-i}|$,
- $C_p = p + \frac{(s^2-\hat{\sigma}^2)(n-p)}{\hat{\sigma}^2}$, where $\hat{\sigma}^2$ is (ideally) an independent estimate of σ^2 and s^2 is the residual mean square for the candidate model.

Example 4.1. Sales Data[1]. The relationship between sales of asphalt roofing shingles for a particular year and factors that are known to influence sales is studied. Four factors under investigation are: promotional accounts (X_1), number of active accounts (X_2), number

[1]The data are taken from Neter and Wasserman (1974), page 391.

District	X_1	X_2	X_3	X_4	Y
1	5.5	31	10	8	79.3
2	2.5	55	8	6	200.1
3	8.0	67	12	9	163.2
4	3.0	50	7	16	200.1
5	3.0	38	8	15	146.0
6	2.9	71	12	17	177.7
7	8.0	30	12	8	30.9
8	9.0	56	5	10	291.9
9	4.0	42	8	4	160.0
10	6.5	73	5	16	339.4
11	5.5	60	11	7	159.6
12	5.0	44	12	12	86.3
13	6.0	50	6	6	237.5
14	5.0	39	10	4	107.2
15	3.5	55	10	4	155.0

Table 4.2: Sales data for Example 4.1.

of competing brands (X_3) and district potential for the sales districts (X_4). Fifteen districts were included in this study and it is of interest to predict sales based on a regression equation. Table 4.2 contains the data.

Myers (1986) studied three models with an emphasis on making good predictions. Some pertinent statistics are as below. A grand mean is included in each model.

Model	R^2	R^2_{Pred}	s^2	PRESS	D	W^2	U^2	A^2
(1) : x_2, x_3	0.994	0.9913	44.552	782.190	0.8534	0.1697	0.1694	0.8329
(2) : x_1, x_2, x_3	0.997	0.9928	24.796	643.358	0.5151	0.0371	0.0371	0.2795
(3) : x_1, x_2, x_3, x_4	0.992	0.9917	26.207	741.756	0.5812	0.0526	0.0517	0.3551

Model (x_1, x_2, x_3) seems to be the best according to the criteria for regression model selection used here. This choice is well supported by examining the EDF statistics. All of

the EDF statistics give a P-value larger than 0.5. This is also seen from Figure 4.1.

Example 4.2. Fitness Data[2]. One objective measure of aerobic fitness is the maximum oxygen consumption in volume per unit body weight per unit time, denoted by Y. An experiment with 31 participants was carried out, and for each participant the following six factors were measured: X_1 is age in years; X_2 is weight in kilograms; X_3 is time to run $1\frac{1}{2}$ miles; X_4 is resting pulse rate; X_5 is pulse rate at the end of running; X_6 is maximum pulse rate during running. The data are in Table 4.3.

Myers (1986) used the data to illustrate the all possible regressions technique in model selection. With PRESS, $\sum_{i=1}^{n}|e_{i,-i}|$, s, C_p and R^2 as selection criteria, sixty-four models are fitted, of which eight models are maintained for further comparison because of their being simultaneously superior in terms of the chosen selection criteria. The eight models are listed below according to PRESS together with the related statistics.

Model	PRESS	$\sum_{i=1}^{n}\|e_{i,-i}\|$	s	C_p	R^2	D	A^2
(a) : x_1, x_2, x_3, x_5, x_6	181.633	54.034	2.275	5.106	0.848	0.7432	0.6956
(b) : x_1, x_3, x_5, x_6	188.599	59.966	2.312	4.880	0.837	0.5670	0.1974
(c) : $x_1, x_2, x_3, x_4, x_5, x_6$	192.788	56.314	2.322	7.000	0.849	0.7484	0.7432
(d) : x_1, x_3, x_4, x_5, x_6	202.402	62.655	2.357	6.846	0.837	0.5495	0.2335
(e) : x_1, x_3, x_5	205.125	61.233	2.441	6.960	0.811	0.8603	0.5038
(f) : x_1, x_2, x_3, x_5	212.272	60.162	2.452	8.104	0.816	0.7598	0.6407
(g) : x_3, x_5, x_6	212.862	64.000	2.448	7.135	0.810	0.5361	0.2905
(h) : x_2, x_3, x_5, x_6	213.158	62.764	2.456	8.206	0.816	0.6249	0.3044

It can be seen that model (a) is very competitive in terms of the model selection criteria. However, the EDF statistics certainly favour model (b). To see why this is the case, normal probability plots for models (a) to (h) are presented in Figure 4.2. It is clear from Figure 4.2 that model (b) is the best among the eight models in terms of normality, and quite likely

[2]The data are taken from SAS User's Guide: Statistics, 1982 Edition, page 106.

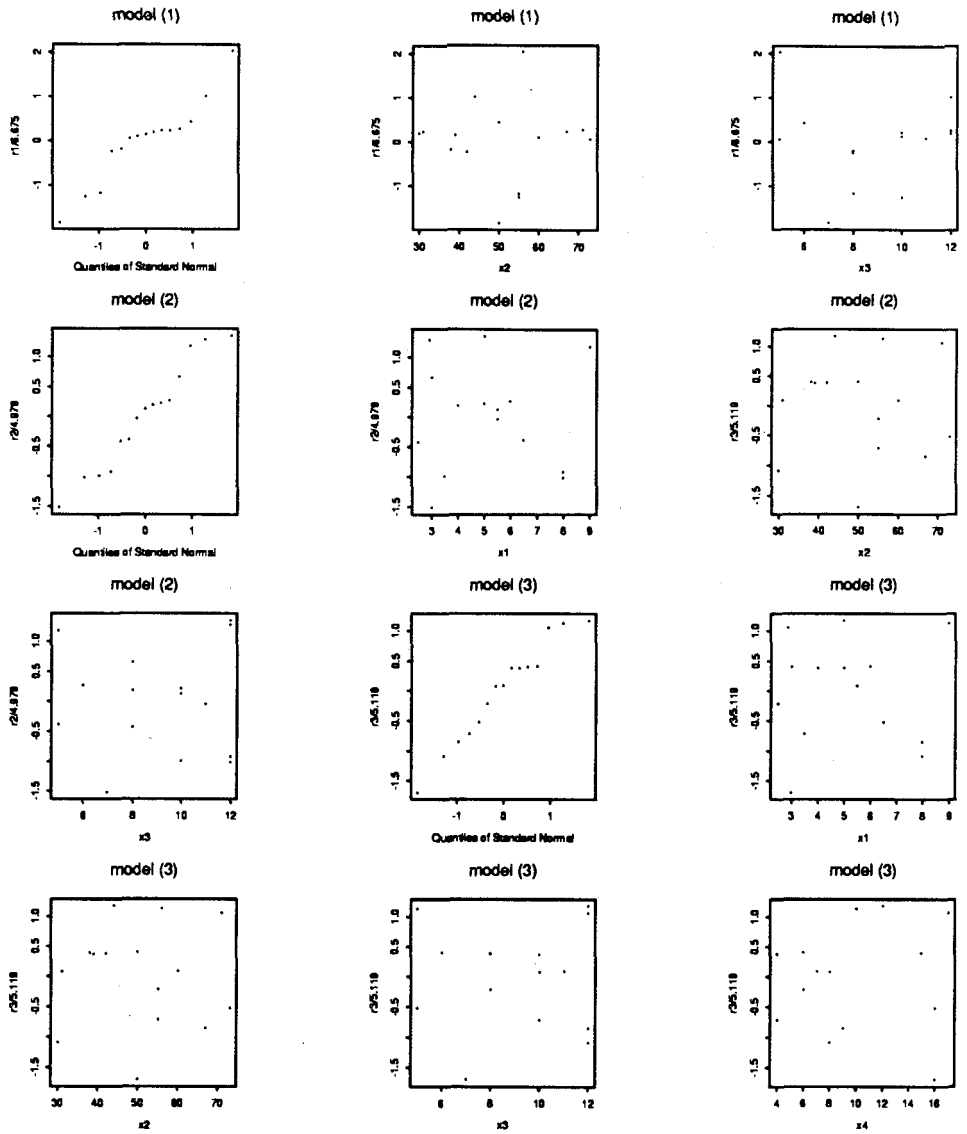


Figure 4.1: Q-Q plots and plots of residuals against regressors for models (1) to (3) in Example 4.1. For model (i), the ordinary residuals are denoted by ri and ri/(residual standard error) is plotted against the regressors. ri/(residual standard error) is also used in the Q-Q plots.

individual	X_1	X_2	X_3	X_4	X_5	X_6	Y
1	44	89.47	11.37	62	178	182	44.609
2	40	75.07	10.07	62	185	185	45.313
3	44	85.84	8.65	45	156	168	54.297
4	42	68.15	8.17	40	166	172	59.571
5	38	89.02	9.22	55	178	180	49.874
6	47	77.45	11.63	58	176	176	44.811
7	40	75.98	11.95	70	176	180	45.681
8	43	81.19	10.85	64	162	170	49.091
9	44	81.42	13.08	63	174	176	39.442
10	38	81.87	8.63	48	170	186	60.055
11	44	73.03	10.13	45	168	168	50.541
12	45	87.66	14.03	56	186	192	37.388
13	45	66.45	11.12	51	176	176	44.754
14	47	79.15	10.60	47	162	164	47.273
15	54	83.12	10.33	50	166	170	51.855
16	49	81.42	8.95	44	180	185	49.156
17	51	69.63	10.95	57	168	172	40.836
18	51	77.91	10.00	48	162	168	46.672
19	48	91.36	10.25	48	162	164	46.774
20	49	73.37	10.08	76	168	168	50.388
21	57	73.37	12.63	58	174	176	39.407
22	54	79.38	11.17	62	156	165	46.080
23	52	76.32	9.63	48	164	166	45.441
24	50	70.87	8.92	48	146	155	54.625
25	51	67.25	11.08	48	172	172	45.118
26	54	91.63	12.88	44	168	172	39.203
27	51	73.71	10.47	59	186	188	45.790
28	57	59.08	9.93	49	148	155	50.545
29	49	76.32	9.40	56	186	188	48.673
30	48	61.24	11.50	52	170	176	47.920
31	52	82.78	10.50	53	170	172	47.467

Table 4.3: Fitness data for Example 4.2.

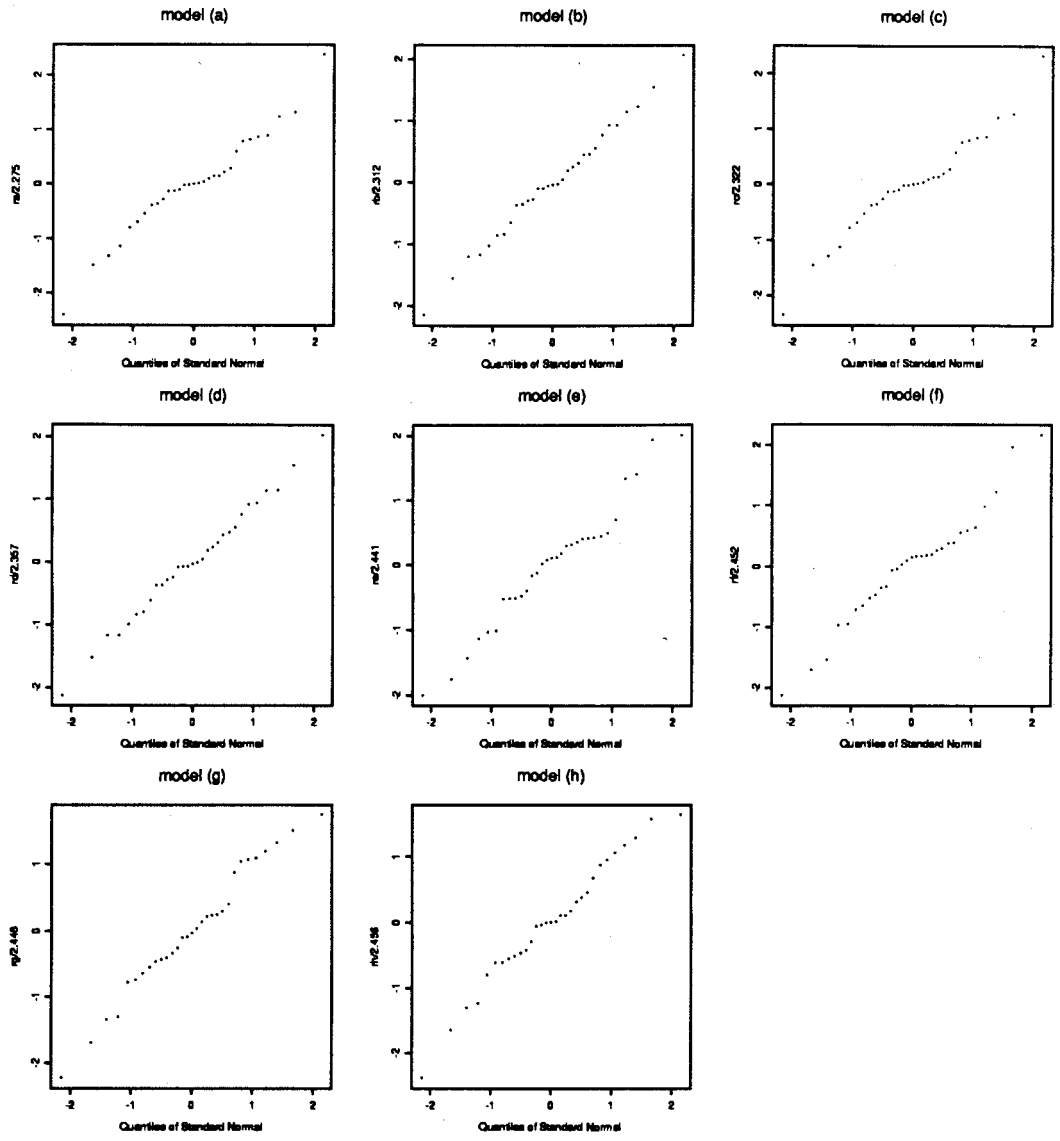


Figure 4.2: Q-Q plots for models (a) to (h) in Example 4.2. For model (a), the ordinary residuals are denoted by ra, and ra/(residual standard error) is used in the Q-Q plot. A similar explanation applies to the rest of the models.

Church	Perimeter X	Area Y	Church	Perimeter X	Area Y
St. Albans	3.48	38.83	Byland	3.14	34.27
Durham	3.69	43.92	Roche	2.04	17.61
Blyth	1.43	9.14	Carmel	1.77	13.37
Binham	2.05	16.66	Bengeo	0.59	2.04
Gloucester	3.05	36.16	Copford	0.69	2.22
Norwich	4.19	38.66	Kempley	0.50	1.46
Leominster	2.43	17.74	Birkin	0.69	1.92
Southwell	2.40	19.46	Hales	0.63	1.86
Chertsey	2.72	23.00	Moccas	0.58	1.69
Hereford	2.99	29.75	Perterchurch	0.86	3.31
Canterbury	4.78	51.19	Little Tey	0.41	1.13
Lindesfarne	1.33	6.60	Melbourne	1.23	6.74
Tintern	1.67	9.04			

Table 4.4: Church data for Example 4.3.

in terms of other model assumptions. It should be noticed here that the EDF statistics also support models (d), (g) and (h). Nevertheless, model (d) has one more regressor (x_4) than model (b) and still does worse than model (b), and models (g) and (h) both suffer from possibly asymmetrical error distributions. Plots of residuals against regressors (not shown here) also indicate that model (b) is the best.

Example 4.3. Church Data[3] . Weisberg (1985) analyzed a data set of 25 post-Conquest Romanesque churches in Britain. Table 4.4 lists the perimeter in hundreds of meters (X) and area in hundreds of square meters for the 25 churches (Y).

[3]The data are taken from Weisberg (1985).

Four models are fitted to the data with the following summary statistics.

Model	D	W^2	U^2	A^2	R^2
$(a): \ y = \alpha + \beta x$	0.4366	0.0232	0.2264	0.1896	0.9629
$(b): \ \sqrt{y} = \alpha + \beta x$	0.8504	0.0986	0.955	0.6776	0.9676
$(c): \ \log y = \alpha + \beta \log x$	0.8097	0.0807	0.0770	0.4641	0.9897
$(d): \ \sqrt{y} = \alpha + \beta\sqrt{x}$	0.4864	0.0277	0.0272	0.1865	0.9804

At first glance, models (a) and (d) are very good fits in terms of EDF statistics tests, because all the EDF statistics give a P-value greater than 0.50. However, plots in Figure 4.3 show that model (c) is superior to the other three models in terms of residual plot and R^2. Notice that among the four models, only model (b) and model (c) are based on the correct units, and model (c) is certainly better than model (b).

The lesson drawn from this example is that like all other criteria used in regression model building, EDF statistics tests alone cannot always point out the overall best model. The following hypothetical example illustrates this point further.

Example 4.4 Anscombe's Data[4]. Anscombe (1973) created four sets of data to demonstrate the usefulness of plots in statistical analysis. Denoted by $\{X_i, Y_i\}$, $i = 1, 2, 3, 4$, the data are in Table 4.5.

Scatter plots of the data shown in Figure 4.4 clearly indicate that it is adequate to fit a straight line model to data set (1) only, because data set (2) shows a curved relationship between X_2 and Y_2; data set (3) most likely has an outlier point; and data set (4) depends crucially on that point staying away from the rest of the data points. However, when fitted to a straight line model $y = \alpha + \beta x$, the four data sets give almost the same estimates: $\hat{\alpha} = 3.0$, $\hat{\beta} = 0.50$, $s = 1.237$, and $R^2 = 0.667$.

[4]The data are taken from Anscombe (1973) .

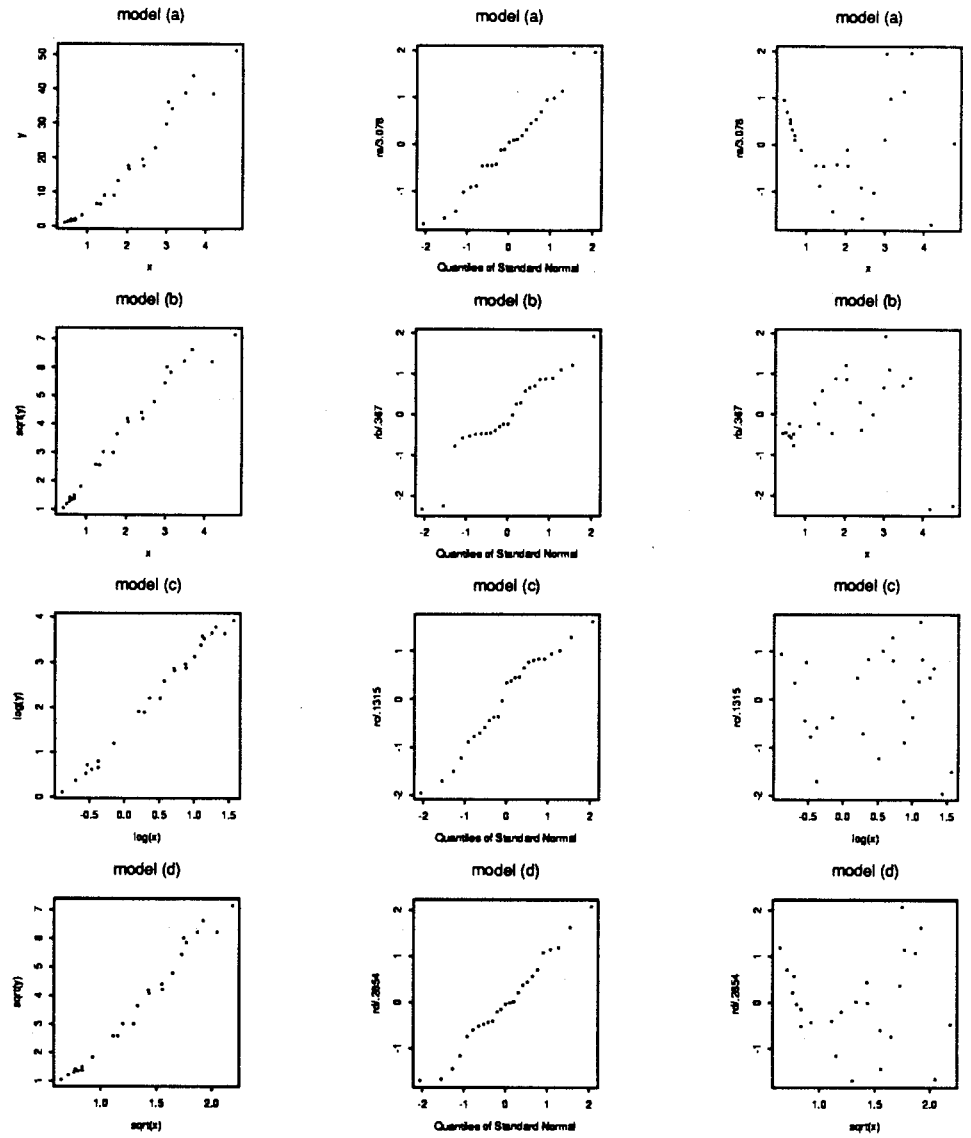


Figure 4.3: Scatter plots, Q-Q plots and plots of residuals against regressors for models (a) to (d) in Example 4.3. For model (a), the ordinary residuals are denoted by ra and ra/(residual standard error) is used in the plots. A similar explanation applies to the other models.

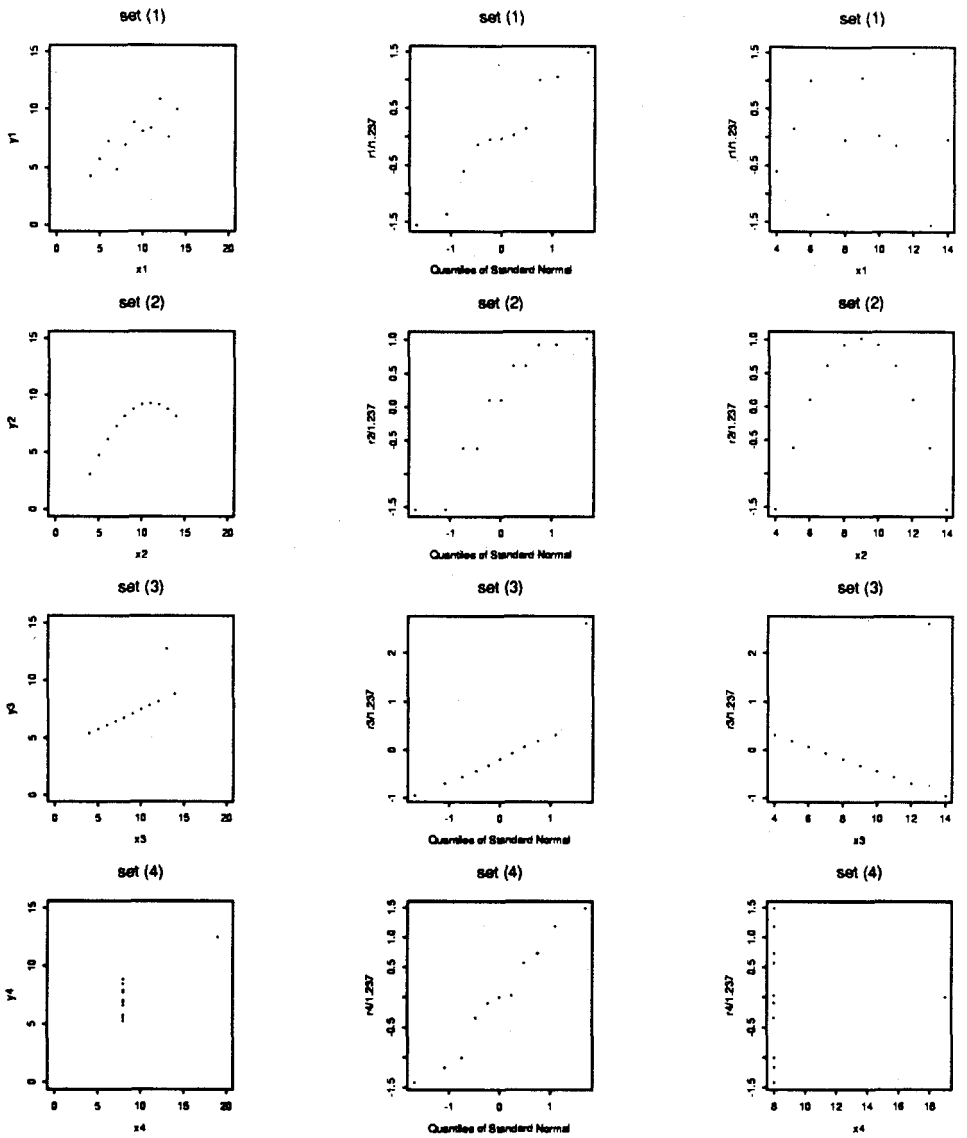


Figure 4.4: Scatter plots, Q-Q plots and plots of residuals against regressors for data sets (1) to (4) in Example 4.4. For data set (i), the ordinary residuals are denoted by ri and ri/(residual standard error) is used in the plots.

X_1, X_2, X_3	Y_1	Y_2	Y_3	X_4	Y_4
10.0	8.04	9.14	7.46	8	6.58
8.0	6.95	8.14	6.77	8	5.76
13.0	7.58	8.74	12.74	8	7.71
9.0	8.81	8.77	7.11	8	8.84
11.0	8.33	9.26	7.81	8	8.47
14.0	9.96	8.10	8.84	8	7.04
6.0	7.24	6.13	6.08	8	5.25
4.0	4.26	3.10	5.39	19	12.50
12.0	10.84	9.13	8.15	8	5.56
7.0	4.82	7.26	6.42	8	7.91
5.0	5.68	4.74	5.73	8	6.89

Table 4.5: Anscobe's data for Example 4.4.

EDF statistics tests are applied to the four data sets. The results are as below:

Data	D	W^2	U^2	A^2
set (1)	0.6136	0.0640	0.0639	0.3643
set (2)	0.6591	0.0776	0.0739	0.5370
set (3)	1.0180	0.1922	0.1716	1.1900
set (4)	0.4410	0.0286	0.0286	0.2057

It can be seen that based on the EDF statistics tests alone, one can make correct decisions for data sets (1) to (3), but one will fail to find any thing wrong with data set (4), and quite possibly will be attracted to the fit for data set (4). This is of course an awkward hypothetical situation for EDF (or other) technique to fail, but the lesson is that one needs to use plotting and other formal techniques together with EDF statistics technique to find a good model. See Figure 4.4.

4.4 Conclusions

In summary, EDF statistics tests can provide evidence that other techniques cannot in regression model building. As overall tests of fit, EDF tests are fairly easy to pass, so the tests should be performed at a significance level of 0.25 or larger. When EDF statistics indicate something is wrong, the problem is usually serious. It seems also true that EDF tests rarely miss a promising model among a group of competitive models.

Chapter 5

EDF Tests for Box-Cox Transformations

When some of the standard assumptions for linear regression models are in doubt, the Box-Cox transformation procedure is usually called in, among a group of procedures available, as a remedy in data analysis. In this procedure, the response variable is subjected to a suitable power transformation so that the standard normal-theory linear regression models can be fitted to the transformed values.

In this chapter, some distribution theory is developed for integral EDF statistics, including the Anderson-Darling statistic A^2 and the Cramér-von Mises statistic W^2, when these statistics are used to test for goodness-of-fit after applying the Box-Cox transformation procedure to fit a linear model. Section 5.1 provides the necessary background for a rigorous treatment of the problem, including a correction to a defect of the usual Box-Cox model; section 5.2 discusses the issue of parameter estimation and describes the procedure of the EDF tests of fit (including a table for carrying out the tests for finite samples); section 5.3 presents the theory behind the test procedure; section 5.4 uses real data to illustrate the procedure described in section 5.2; section 5.5 draws conclusions and discusses some issues

about the Box-Cox transformation procedure; and section 5.6 contains the proof and some technical details for Theorem 5.3.1 and Theorem 5.3.2.

5.1 Introduction

Let $Y_1, \ldots, Y_n$ be positive independent random variables generating responses. For a real number λ, define the modified Box-Cox transformation family by

$$Y_i(\lambda) = \begin{cases} (Y_i^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, \\ \log Y_i & \text{if } \lambda = 0. \end{cases} \tag{5.1.1}$$

The Box-Cox transformation procedure attempts to find a suitable λ and uses it to transform the original responses Y_i into $Y_i(\lambda)$ according to (5.1.1) so that the following linear model is approximately applicable,

$$Y(\lambda) \approx X\beta + \sigma\varepsilon, \tag{5.1.2}$$

where $X = (x_{ij})$ is a known $n \times p$ matrix of constants, $\beta = (\beta_1, \ldots, \beta_p)^t$ are unknown regression parameters (a column vector), σ is an unknown positive constant, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t$ are independent and identically distributed standard normal random variables, and $Y(\lambda) = (Y_1(\lambda), \ldots, Y_n(\lambda))^t$; superscript t denotes transpose.

Now let $\mu_i = x_i^t\beta$ $(i = 1, \ldots, n)$, where x_i^t is the i^{th} row of X. If model (5.1.2) were true exactly, $Y_i(\lambda) \sim N(\mu_i, \sigma^2)$ would follow and this implies that the density of Y_i would be, for $\lambda > 0$,

$$g(y; \lambda, \mu_i, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left[-\{(y^\lambda - 1)/\lambda - \mu_i\}^2/(2\sigma^2)\right] y^{\lambda-1} 1[y > 0], \tag{5.1.3}$$

where $1[A] = 1$ if A is true, and $1[A] = 0$ if A is false. However,

$$\begin{aligned} & \int_0^{+\infty} g(y; \lambda, \mu_i, \sigma^2) dy \\ = & \int_{-1/\lambda}^{+\infty} (2\pi\sigma^2)^{-1/2} \exp\{-(u - \mu_i)^2/(2\sigma^2)\} du \\ = & \int_{-\delta_i}^{+\infty} \phi(v) dv \\ = & \Phi(\delta_i), \end{aligned}$$

where $\delta_i = (\mu_i + 1/\lambda)/\sigma$, $\phi(v)$ and $\Phi(v)$ are the density and distribution function of a standard normal random variable, respectively. Since $\delta_i < +\infty$, $\Phi(\delta_i) < 1$, therefore (5.1.3) is not a proper density and in turn $Y_i(\lambda) \sim N(\mu_i, \sigma^2)$ cannot be true. In this case the left tail of the normal distribution is cut off. If $\lambda < 0$, the right tail will be cut off.

For the purpose of subsequent analysis, a proper density and a proper distribution function are needed, so model (5.1.2) is modified as below: Let $Y_1, \ldots, Y_n$ be positive independent random variables and suppose that there exist $\lambda \in R$, $\beta \in R^p$ and $\sigma > 0$ such that Y_i has the density

$$f(y_i; \lambda, \mu_i, \sigma^2) = \begin{cases} \sigma^{-1}\phi\{(y_i(\lambda) - \mu_i)/\sigma\}\, y_i^{\lambda-1} 1[y_i > 0]/\Phi(\delta_i), & \text{if } \lambda > 0, \\ \sigma^{-1}\phi\{(\log y_i - \mu_i)/\sigma\}\, y_i^{-1}, & \text{if } \lambda = 0, \\ \sigma^{-1}\phi\{(y_i(\lambda) - \mu_i)/\sigma\}\, y_i^{\lambda-1} 1[y_i > 0]/\Phi(-\delta_i), & \text{if } \lambda < 0, \end{cases} \tag{5.1.4}$$

where $\mu_i = x_i^t\beta$, $\delta_i = (\mu_i + 1/\lambda)/\sigma$ $(i = 1, \ldots, n)$. In this formula, $y_i(\lambda)$ is y_i transformed as in (5.1.1).

There are two changes from model (5.1.2) to model (5.1.4). First, model (5.1.2) does not have proper densities while model (5.1.4) does. This is only a technical change. Secondly, model (5.1.2) cannot be the basis for estimating λ by itself, while model (5.1.4) treats λ as a genuine parameter with the same status as β and σ. In other words, model (5.1.4) is taken to model the Y_i. This point of view will be assumed in the remainder of this chapter. See section 5.4 for a discussion.

Notice that in model (5.1.4) $E(Y_i) \neq \mu_i$, therefore model (5.1.4) is a non-linear model in β.

5.2 Parameter Estimation and EDF Tests of Fit

5.2.1 Parameter estimation

The method of maximum likelihood is used to estimate λ, β and ν $(= \sigma^2)$ simultaneously. Denote by L the log-likelihood function of a random sample $Y_1, \ldots, Y_n$ based on model

(5.1.4); then except for a constant,

$$L = \begin{cases} -(n/2)\log\nu - (2\nu)^{-1}\sum_{i=1}^{n}\{(y_i^\lambda - 1)/\lambda - \mu_i\}^2 \\ +(\lambda-1)\sum_{i=1}^{n}\log y_i - \sum_{i=1}^{n}\log\Phi(\delta_i), & \text{if } \lambda > 0, \\ -(n/2)\log\nu - (2\nu)^{-1}\sum_{i=1}^{n}\{\log y_i - \mu_i\}^2 \\ -\sum_{i=1}^{n}\log y_i, & \text{if } \lambda = 0, \\ -(n/2)\log\nu - (2\nu)^{-1}\sum_{i=1}^{n}\{(y_i^\lambda - 1)/\lambda - \mu_i\}^2 \\ +(\lambda-1)\sum_{i=1}^{n}\log y_i - \sum_{i=1}^{n}\log\Phi(-\delta_i), & \text{if } \lambda < 0. \end{cases} \tag{5.2.1}$$

Because $Y_i(\lambda)$ defined by (5.1.1) is differentiable with respect to λ, L is differentiable with respect to λ, β and ν. Thus, for $\lambda > 0$ the likelihood equations are, for $k = 1, 2, \ldots, p$,

$$\frac{\partial L}{\partial \beta_k} = \nu^{-1}\sum_{i=1}^{n}[(y_i^\lambda - 1)/\lambda - \mu_i]x_{ik} - \sum_{i=1}^{n}[\phi(\delta_i)/\Phi(\delta_i)][x_{ik}/\sqrt{\nu}] = 0, \tag{5.2.2}$$

$$\frac{\partial L}{\partial \nu} = -\frac{n}{2\nu} + \frac{1}{2\nu^2}\sum_{i=1}^{n}[(y_i^\lambda - 1)/\lambda - \mu_i]^2 + \sum_{i=1}^{n}[\phi(\delta_i)/\Phi(\delta_i)]\frac{\delta_i}{2\nu} = 0, \tag{5.2.3}$$

$$\begin{aligned}\frac{\partial L}{\partial \lambda} &= -(\lambda^2\nu)^{-1}\sum_{i=1}^{n}[(y_i^\lambda - 1)/\lambda - \mu_i](\lambda y_i^\lambda \log y_i - y^\lambda + 1) \\ &\quad + \sum_{i=1}^{n}[\phi(\delta_i)/\Phi(\delta_i)][1/(\lambda\nu)] + \sum_{i=1}^{n}\log y_i = 0.\end{aligned} \tag{5.2.4}$$

Similarly, for $\lambda < 0$ the likelihood equations are, for $k = 1, 2, \ldots, p$,

$$\frac{\partial L}{\partial \beta_k} = \nu^{-1}\sum_{i=1}^{n}[(y_i^\lambda - 1)/\lambda - \mu_i]x_{ik} + \sum_{i=1}^{n}[\phi(-\delta_i)/\Phi(-\delta_i)][x_{ik}/\sqrt{\nu}] = 0, \tag{5.2.5}$$

$$\frac{\partial L}{\partial \nu} = -\frac{n}{2\nu} + \frac{1}{2\nu^2}\sum_{i=1}^{n}[(y_i^\lambda - 1)/\lambda - \mu_i]^2 - \sum_{i=1}^{n}[\phi(-\delta_i)/\Phi(-\delta_i)]\frac{\delta_i}{2\nu} = 0, \tag{5.2.6}$$

$$\begin{aligned}\frac{\partial L}{\partial \lambda} &= -(\lambda^2\nu)^{-1}\sum_{i=1}^{n}[(y_i^\lambda - 1)/\lambda - \mu_i](\lambda y_i^\lambda \log y_i - y^\lambda + 1) \\ &\quad - \sum_{i=1}^{n}[\phi(-\delta_i)/\Phi(-\delta_i)][1/(\lambda\nu)] + \sum_{i=1}^{n}\log y_i = 0.\end{aligned} \tag{5.2.7}$$

At $\lambda = 0$ the above partial derivatives reduce to the following, respectively,

$$(2\nu)^{-1}\sum_{i=1}^{n}(\log y_i - \mu_i)x_{ik},\ k = 1, \ldots, p, \tag{5.2.8}$$

$$-(n/2) + (2\nu\sqrt{\nu})^{-1}\sum_{i=1}^{n}(\log y_i - \mu_i)^2, \tag{5.2.9}$$

$$-(2\nu)^{-1}\sum_{i=1}^{n}(\log y_i - \mu_i)\log^2 y_i + \sum_{i=1}^{n}\log y_i. \tag{5.2.10}$$

It does not seem possible to find closed-form maximum likelihood estimators for β, ν and λ from the likelihood function directly, or from the likelihood equations. Therefore, iterative numerical methods are necessary.

Now denote by l_{BC} the log-likelihood function of model (5.1.2) discussed by Box and Cox (1964). Then, except for a constant,

$$l_{BC} = -(n/2)\log\nu - (2\nu)^{-1}\sum_{i=1}^{n}\{y_i(\lambda) - \mu_i\}^2 + (\lambda - 1)\sum_{i=1}^{n}\log y_i \tag{5.2.11}$$

in terms of the original random variables. The Box-Cox transformation procedure proceeds to maximize l_{BC} over β and ν with λ kept fixed; the result is a function of λ alone that is given by

$$l_{BC}(\lambda) = -(n/2)\log\tilde{\nu}_{BC}(\lambda) - (n/2) + (\lambda - 1)\sum_{i=1}^{n}\log y_i, \tag{5.2.12}$$

where $n\tilde{\nu}_{BC}(\lambda) = Y(\lambda)^t(I - X(X^tX)^{-1}X^t)Y(\lambda)$ is the residual sum of squares from regressing $Y(\lambda)$ on X. The final value $\tilde{\lambda}$ of λ for subsequent analysis is determined by maximizing $l_{BC}(\lambda)$ over λ, and the estimates for β and ν are given by

$$\tilde{\beta} = (X^tX)^{-1}X^tY(\tilde{\lambda}), \tag{5.2.13}$$

$$\tilde{\nu} = \frac{1}{n}Y(\tilde{\lambda})(I - X(X^tX)^{-1}X^t)Y(\tilde{\lambda}). \tag{5.2.14}$$

Note that the residual sum of squares is usually divided by $n - p$ to obtain an estimate of ν, where p is the number of regression parameters. Examples given in section 5.4 will follow this convention.

It can be seen by comparing (5.2.11) to (5.2.1) that $l_{BC} \approx L$ if $-\sum_{i=1}^{n}\log\Phi(\delta_i) \approx 0$ (or $-\sum_{i=1}^{n}\log\Phi(-\delta_i) \approx 0$). This happens if (1) λ is close to zero, or (2) μ_i's (or $-\mu_i$'s) are large, or (3) ν is small. In practical terms, for $\lambda > 0$, let $\delta_n^+ = \min_{1\le i\le n}\{\delta_i\}$; if $\delta_n^+ > \Phi^{-1}(e^{-c/n})$

for a positive constant c, then

$$-\sum_{i=1}^{n} \log \Phi(\delta_i) \leq -n \log \Phi(\delta_n^+) < c.$$

For instance, for $c = 0.01$ and $n = 50$, $\Phi^{-1}(e^{-0.01/50}) = 3.54$, that is, if the minimum of δ_i has a magnitude 3.54, which is about three and a half standard deviations according to the standard normal distribution, then truncating the left tail cuts off less than 0.01 from the log-likelihood. Similarly, for $\lambda < 0$, let $\delta_n^- = \min_{1 \leq i \leq n}\{-\delta_i\}$; if $\delta_n^- > \Phi^{-1}(e^{-c/n})$ for a positive constant c, then $-\sum_{i=1}^{n} \log \Phi(-\delta_i) < c$. Simple calculations of the above type can serve to assess transformation potentials.

Very often in practice, the term $-\sum_{i=1}^{n} \log \Phi(\delta_i)$ (or $-\sum_{i=1}^{n} \log \Phi(-\delta_i)$) is practically zero so that L and l_{BC} will lead to practically the same parameter estimates for a given set of data.

5.2.2 EDF tests of fit

To test for goodness-of-fit when fitting model (5.1.4) to data, the integral type of EDF statistics can be employed. For the present case, let $\theta = (\lambda, \beta^t, \nu)^t$ and let $\hat{\theta} = (\hat{\lambda}, \hat{\beta}^t, \hat{\nu})^t$ be the maximum likelihood estimate for θ. The cumulative distribution function for Y_i in model (5.1.4) is given by

$$F_i(y_i; \theta) = \begin{cases} \{\Phi(\delta_i(y_i)) + \Phi(\delta_i) - 1\}/\Phi(\delta_i), & \text{if } \lambda > 0, \\ \Phi\{(\log y_i - \mu_i)/\sqrt{\nu}\}, & \text{if } \lambda = 0, \\ \Phi(\delta_i(y_i))/\Phi(-\delta_i), & \text{if } \lambda < 0, \end{cases} \tag{5.2.15}$$

where $\mu_i = x_i^t \beta$, $\delta_i = (\mu_i + 1/\lambda)/\sqrt{\nu}$, $\delta_i(y_i) = (y_i(\lambda) - \mu_i)/\sqrt{\nu}$. Now for each i, let $u_i = F_i(y_i; \hat{\theta})$ and let the empirical distribution function of the u_i's be

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1[u_i \leq t], \ (0 \leq t \leq 1). \tag{5.2.16}$$

The integral EDF statistics are based on the weighted and integrated discrepancies between $\hat{F}_n(t)$ and $F(t) \equiv t$ $(0 \leq t \leq 1)$, as described in section 4.1.2. In particular, for a given data

set $y_1, \ldots, y_n$ (in ascending order), W^2 and A^2 can be computed according to

$$W^2 = \sum_{i=1}^{n} \left\{ u_i - \frac{2i-1}{2n} \right\}^2 + \frac{1}{12n}, \tag{5.2.17}$$

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \{(2i-1) \log u_i + (2n+1-i) \log(1-u_i)\}. \tag{5.2.18}$$

See Durbin (1973), Stephens (1986) and section 4.1.3.

To perform a goodness-of-fit test of H_0^{BC}: model (5.1.4) fits the data, the following steps can be used:

(a) Find $\hat{\lambda}$, $\hat{\beta}$ and $\hat{\nu}$ as outlined above,

(b) Compute $u_i = F_i(y_i; \hat{\theta})$ according to (5.2.15) with the true parameters replaced by their estimates, or approximately, $u_i = \Phi(\delta_i(y_i))$, where $\delta_i(y_i) = (y_i(\hat{\lambda}) - \hat{\mu}_i)/\hat{\sigma}$, $\hat{\sigma} = \sqrt{\hat{\nu}}$,

(c) Calculate W^2 and A^2 according to (5.2.17) and (5.2.18), respectively,

(d) Modify W^2 into $W^{**} = W^2(1 + 2.5/n)$, modify A^2 into $A^{**} = A^2(1 + 4.2/n - 43/n^2)$, where n is the sample size, and reject H_0^{BC} at significance level α if the modified statistics exceed the upper α-percentiles given in Table 5.1.

The entries in Table 5.1 are the upper percentiles of the asymptotic distributions of A^2 and W^2, respectively, as $n \to \infty$, and for the case where $\lambda = 0$. The modified statistics A^{**} and W^{**} shown in Table 1 are obtained through simulations for finite sample sizes and through smoothing of the simulated finite sample percentiles and the asymptotic percentiles. See Stephens (1974) and Linnet (1988).

5.3 Theory of the Tests

In order to obtain and use the asymptotic distributions of A^2 and W^2, a key step is to show that the (estimated) empirical process

$$\hat{Y}_n(t) = \sqrt{n}(\hat{F}_n(t) - t) \tag{5.3.1}$$

Modified Statistics	Upper percentiles								
					α				
A^{**}	0.50	0.40	0.30	0.20	0.15	0.10	0.05	0.025	0.01
	0.2871	0.3211	0.3624	0.4184	0.4572	0.5112	0.6029	0.6945	0.8158
					α				
W^{**}	0.50	0.40	0.30	0.20	0.15	0.10	0.05	0.025	0.01
	0.0428	0.0488	0.0562	0.0663	0.0736	0.0836	0.1007	0.1179	0.1406

Table 5.1: Upper percentiles of the asymptotic distributions of A^2 and W^2 for testing Box-Cox transformations when the true λ value is zero. The statistics A^2 and W^2 are modified into $A^{**} = A^2(1 + 4.2/n - 43/n^2)$ and $W^{**} = W^2(1 + 2.5/n)$, respectively, where n is the sample size, and the modification is aimed at taking care of finite sample situations. See Stephens (1974).

converges weakly to a Gaussian process with zero mean and a manageable covariance function. Some sufficient conditions for such a desired result are presented in the following theorem.

Theorem 5.3.1 *In model (5.1.4), suppose that*

(A) $X = (1_n\ U)$ *is such that* $1_n^t U = 0$, *where* 1_n *is an* $n \times 1$ *vector of 1's,*

(B) $E = \lim_{n\to\infty} n^{-1} X^t \mu^2$ *and* $b = \lim_{n\to\infty} n^{-1} 1_n^t \mu^4$ *exist for any* $\beta \in \Omega$, *where* Ω *is an open convex subset of* R^p, $\mu = X\beta$, μ^k *is an* $n \times 1$ *vector with its* i^{th} *component equal to* $(x_i^t\beta)^k$, $k = 2, 4$,

(C) $\Delta = \lim_{n\to\infty} n^{-1} X^t X$ *exists and is positive definite,*

(D) *there are constants* M_1 *and* M_2 *such that for any* n *and any* $i = 1, \ldots, n$,

$$\frac{1}{n} \sum_{i=1}^{n} \max_{1\le j\le p} |x_{ij}| \le M_1,$$

$$\frac{1}{\sqrt{n}} \max_{1 \le j \le p} |x_{ij}| \le M_2,$$

then

(1) *when* $\lambda = 0$, *the maximum likelihood estimate* $\hat{\theta}$ *is asymptotically normal, that is,* $\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, \Gamma)$, *where*

$$\Gamma = \begin{pmatrix} A & B \\ B^t & C \end{pmatrix}^{-1},$$

with

$$A = (4\nu)^{-1}(7\nu^2 + 10\nu\beta^t\Delta\beta + b),$$

$$B = (-D/2 - E^t/(2\nu),\ -\beta_1/\nu),$$

$$C = \begin{pmatrix} \Delta/\nu & 0 \\ 0^t & 1/(2\nu^2) \end{pmatrix},$$

where $D^t = (1, 0, \ldots, 0)^t$ *is a* $p \times 1$ *vector with its first component equal to 1 and all the other components equal to 0;*

(2) *when* $\lambda = 0$, *the (estimated) empirical process* $\hat{Y}_n(t) = \sqrt{n}(\hat{F}_n(t) - t)$ *converges weakly to a Gaussian process* $Y(t)$ *with zero mean and covariance function*

$$\rho(s,t) = \min(s,t) - st - \frac{1}{1!}J_1(s)J_1(t) - \frac{1}{2!}J_2(s)J_2(t) - \frac{1}{3!}J_3^*(s)J_3^*(t), \qquad (5.3.2)$$

where $J_1(t) = \phi(\Phi^{-1}(t))$, $J_2(t) = \Phi^{-1}(t)J_1(t)$ *are as before, and* $J_3^*(t) = [(\Phi^{-1}(t))^2 - 1]J_1(t)$, $s, t \in [0,1]$.

Remark. It is interesting to notice that the last three terms in equation (5.3.2) correspond to the first three (weighted) Hermite polynomials.

Theorem 5.3.2 *Under similar conditions to those of the above theorem, and for general* λ, σ *and* β,

(1) *the (estimated) empirical process* $\hat{Y}_n(t) = \sqrt{n}(\hat{F}_n(t) - t)$ *converges weakly to a Gaussian process* $Y_G(t)$ *with zero mean and covariance function*

$$\rho_G(s,t) = \min(s,t) - st - \Psi_G^t(s)\Gamma_G\Psi_G(t), \tag{5.3.3}$$

where Γ_G *is a* $(p+2)\times(p+2)$ *matrix and is supposed to be positive definite, and* $\Psi_G(t)$ *is a* $(p+2)\times 1$ *column vector function of* t*, where* $p = dim(\beta)$*,* $t \in [0,1]$*;*

(2) *the covariance function* $\rho_G(s,t)$ *of (5.3.3) converges (pointwise) to the covariance function* $\rho(s,t)$ *of (5.3.2) as (i)* $\lambda \to 0$*, or (ii)* $\sigma \to 0$*, or (iii)* $\beta_1 \to +\infty$*.*

The proof of Theorem 5.3.1 and part (2) of Theorem 5.3.2 is given in section 5.6; part (1) of Theorem 5.3.2 is discussed in section 5.5; and the expressions for $\Gamma_G(s,t)$ and $\Psi_G(t)$ are given in section 5.6.

5.4 Examples

Three examples are given below to illustrate the use of Table 5.1. They illustrate three typical situations where λ is close to zero, positive, and negative.

Example 5.1. Textile Data. Table 4 of Box and Cox (1964) is reproduced below as Table 5.2, which is the result of a single replicate of a 3^3 factorial experiment. The response y is the cycles to failures of worsted yarn, and the three explanatory variables assume three different levels each; see Box and Cox (1964) for details.

Three main effect linear models are fitted to the data. The first model uses y directly and its goodness-of-fit is judged by the EDF tests of section 4.3.1. (Note: In this case, the modified EDF statistics are $A^* = A^2(1 + 0.75/n + 2.25/n^2)$ and $W^* = W^2(1 + 0.5/n)$, and Table 4.1 is used.) The second model transforms y according to (5.1.1) and its goodness-of-fit is assessed using Table 5.1 provided in section 5.2. The third model uses a "nice" value to transform y and its goodness-of-fit is also judged by Table 5.1. Parameter estimates are

Factor levels				Factor levels			
x_1	x_2	x_3	cycles to failures, y	x_1	x_2	x_3	cycles to failures, y
-1	-1	-1	674	0	0	1	438
-1	-1	0	370	0	1	-1	442
-1	-1	1	292	0	1	0	332
-1	0	-1	338	0	1	1	220
-1	0	0	266	1	-1	-1	3636
-1	0	1	210	1	-1	0	3184
-1	1	-1	170	1	-1	1	2000
-1	1	0	118	1	0	-1	1568
-1	1	1	90	1	0	0	1070
0	-1	-1	1414	1	0	1	566
0	-1	0	1198	1	1	-1	1140
0	-1	1	634	1	1	0	884
0	0	-1	1022	1	1	1	360
0	0	0	620				

Table 5.2: Textile data reproduced from Table 4 of Box and Cox (1964), Example 5.1.

Model	Parameter Estimates		Modified EDF (P-value)	
	$\hat{\nu}$	$\hat{\lambda}$	A^*	W^*
y:	488.2	—	1.394 (<0.01)	0.241 (<0.01)
$y(\lambda)$:	0.126	−0.059	0.3697 (>0.20)	0.0541 (>0.30)
$\log y$:	0.186	0	0.2719 (>0.50)	0.0353 (>0.50)
$y(\lambda)$:	Minimum of $-\delta_i = 82.577$		$-\sum_{i=1}^{27} \log \Phi(-\delta_i) = 0$	

Table 5.3: EDF tests of goodness-of-fit for three main effect linear models, textile data, Example 5.1.

obtained by directly maximizing the log-likelihood function and by applying the Box-Cox transformation procedure separately; the results are practically the same. See Table 5.3.

As can be seen from Table 5.3, the transformed models are better fits to the data.

Example 5.2. Tree Data The data in Table 5.4 is taken from *Minitab Student Handbook* (Ryan, Joiner and Ryan (1976), page 278). The heights (x_1), the diameters (x_2) at 4.5 ft above ground level and the volumes (y) were measured for a sample of 31 black cherry trees in the Allegheny National Forest, Pennsylvania. The data were collected to provide a basis for determining an easy way of estimating the volume of a tree based on its height and diameter. Again, three linear models are fitted to the data, using y, $y(\lambda)$ and $y(\frac{1}{3})$, where $\frac{1}{3}$ is chosen according to dimension consideration of volume vs length. Parameter estimates are obtained by directly maximizing the log-likelihood function and by applying the Box-Cox transformation procedure separately; the estimates are virtually the same; Table 5.5 contains the results. In this example, the need for transformation is suggested by dimension consideration and the Box-Cox estimate $\hat{\lambda} = 0.307$ agrees with this consideration. All the three models pass the EDF tests, especially, the untransformed model is the "best" in the sense of EDF tests. However, a close look at residual plots (Figure 5.1) shows that the transformed models are better than the untransformed one. It seems in this case that normality was sacrificed a little bit to obtain overall better fits in doing the Box-Cox transformation.

Example 5.3. Biological Data. Table 1 of Box and Cox (1964) is reproduced as Table 5.6 below. The entries are the survival times (unit is 10 hours) of animals in a 3×4 completely randomized factorial experiment. The factors are Poison with three levels and Treatment with four levels.

Three main effect models are fitted to the data as in the above two examples. Table 5.7 clearly shows that a power transformation improves model fit a great deal.

x_1	x_2	y	x_1	x_2	y	x_1	x_2	y
8.3	70	10.3	11.4	76	21.0	14.5	74	36.3
8.6	65	10.3	11.4	76	21.4	16.0	72	38.3
8.8	63	10.2	11.7	69	21.3	16.3	77	42.6
10.5	72	16.4	12.0	75	19.1	17.3	81	55.4
10.7	81	18.8	12.9	74	22.2	17.5	82	55.7
10.8	83	19.7	12.9	85	33.8	17.9	80	58.3
11.0	66	15.6	13.3	86	27.4	18.0	80	51.5
11.0	75	18.2	13.7	71	25.7	18.0	80	51.0
11.1	80	22.6	13.8	64	24.9	20.6	87	77.0
11.2	75	19.9	14.0	78	34.5			
11.3	79	24.2	14.2	80	31.7			

Table 5.4: Tree data for Example 5.2.

Model	Parameter Estimates		Modified EDF (P-value)	
	$\hat{\sigma}$	$\hat{\lambda}$	A^*	W^*
1 y:	3.882	—	0.2548 (>0.50)	0.0366 (> 0.50)
2 $y(\lambda)$:	0.227	0.307	0.3192 (>0.40)	0.0487 (>0.40)
3 $y(\frac{1}{3})$:	0.249	$\frac{1}{3}$	0.2983 (>0.40)	0.0440 (>0.40)
$y(\lambda)$:	Minimum of $\delta_i = 29.22$		$-\sum_{i=1}^{25} \log \Phi(\delta_i) \approx 0$	

Table 5.5: EDF tests of goodness-of-fit for three straight line models, tree data, Example 5.2.

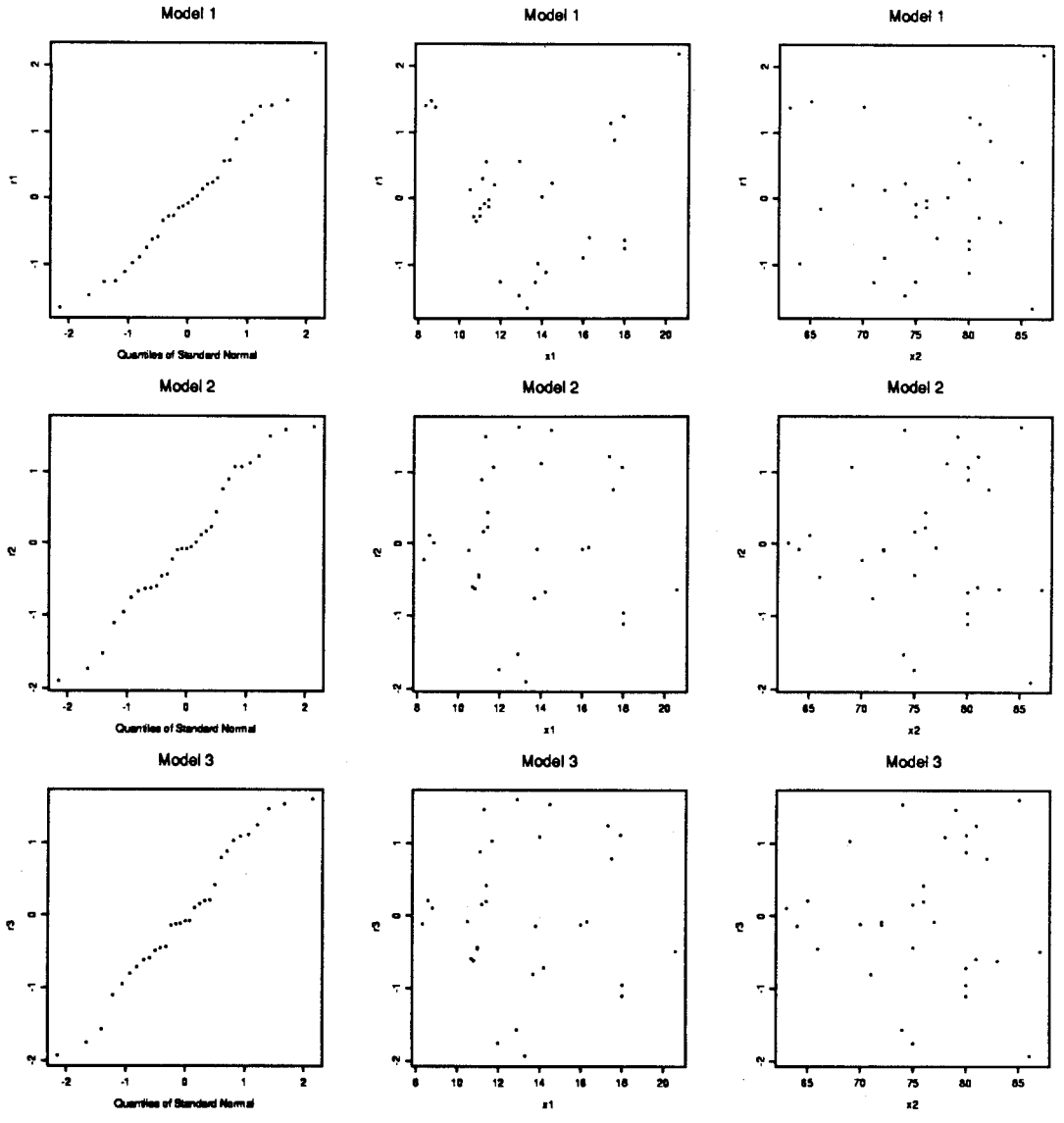


Figure 5.1: Q-Q plots and plots of residuals against regressors for models 1 to 3, tree data, Example 5.2. ri denotes the standardized residuals for model i, i=1, 2, 3.

	Treatment			
Poison	A	B	C	D
I	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
II	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
III	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

Table 5.6: Biological data reproduced from Table 1 of Box and Cox (1964), Example 5.3.

Model	Parameter Estimates		Modified EDF (P-value)	
	$\hat{\sigma}$	$\hat{\lambda}$	A^*	W^*
y:	0.1582	—	1.0550 (<0.01)	0.1589 (<0.025)
$y(\lambda)$:	0.3916	−0.75	0.2110 (>0.50)	0.0296 (>0.50)
$y(-1)$:	0.4931	−1.00	0.3056 (>0.40)	0.0407 (>0.50)
$y(\lambda)$:	Minimum of $-\delta_i = 3.667$		$-\sum_{i=1}^{48} \log \Phi(-\delta_i) = 0.0005$	

Table 5.7: EDF tests of goodness-of-fit for three main effect linear models, biological data, Example 5.3.

5.5 Conclusions and Conjectures

Linnet (1988) empirically studied the use of the Anderson-Darling statistic A^2 and the Cramér-von Mises statistic W^2 to test for normality of the power transformed data in one-sample problems. Through simulation studies Linnet concluded empirically that the null distributions of A^2 and W^2 do not depend on parameter values for the transformation parameter λ, the mean μ and the variance ν. A table was provided for A^2 and W^2 for finite samples in which the asymptotic critical points were obtained by extrapolation.

Since the Box-Cox transformation procedure is often used in analysis of variance and regression analysis, linear models have been examined in this chapter. Moreover, tests based on A^2 and W^2 are regarded here as overall goodness-of-fit tests, instead of tests for normality only. This is reasonable because the theory leading to the applications of A^2 and W^2 assumes not only normality of error distribution but other standard linear model assumptions as well.

In order to treat the theory of power transformations rigorously, it is necessary to complete model (5.1.2) into model (5.1.4). This completion changes little of the usual estimation process for application as discussed in section 5.2.1, and provides a quantitative guide for assessing transformation potentials as mentioned in section 5.2.1, too. However, theory for inference becomes more difficult for model (5.1.4) than for model (5.1.2). For example, some of the numerical examples in Bickel and Doksum (1981) are only approximately correct, because (in the terminology of this chapter) the δ_i's are not large enough to ignore the term $-\sum_{i=1}^{n} \log \Phi(\delta_i)$ which should be in the log-likelihood function and should show up in a way in the asymptotic variance-covariance matrix.

One analytically tractable case is when $\lambda = 0$. In this case, Hinkley (1975) obtained the asymptotic variance-covariance matrix of $\sqrt{n}(\hat{\theta} - \theta)$ for the one-sample problem. There is a misprint in his derivation because the asymptotic covariance of $\hat{\mu}$ and $\hat{\nu}$ should be $2\mu(\nu + \mu^2)/3$, instead of $2\mu(\nu + \mu^2)$. The asymptotic variance-covariance matrix for the

case of linear models is given in Theorem 5.3.1. For the case where $\lambda \neq 0$, it does not seem possible to find an explicit expression for the Fisher information matrix, but the properties of the asymptotic variance-covariance matrix are explored numerically (see Theorem 5.3.2, too). One finding worth mentioning is the large increase of the variances for $\hat{\beta}$ and $\hat{\nu}$ when λ is estimated compared to the corresponding variances when λ is known. However, the point of view of this chapter is to model the Y_i with model (5.1.4) as mentioned after the introduction of model (5.1.4). For more information concerning the variance inflation problem, see Bickel and Doksum (1981), Box and Cox (1982), and Hinkley and Runger (1984). It is conjectured that under mild conditions on the model matrix X, maximum likelihood estimates of the parameters in model (5.1.4) are asymptotically normal for general λ values and have variance-covariance matrices with the usual Fisher structure.

Following the above discussion, it can be seen that a rigorous treatment of a general distribution theory for A^2 and W^2 needs much work to build. As a first step, the $\lambda = 0$ case has been done rigorously here. In this case, $Y_i(0) = \log Y_i$ is normally distributed, so β and ν are essentially location and scale parameters and it follows that the asymptotic distributions of A^2 and W^2 do not depend on β and ν. For the case where $\lambda \neq 0$, part (1) of Theorem 5.3.2 outlines the covariance function $\rho_G(s,t)$ of the weak limit of the estimated empirical process of (5.3.1). The form of $\rho_G(s,t)$ has been guessed along the lines of proving part (2) of Theorem 5.3.1 (see section 5.6). A rigorous proof of part (1) of Theorem 5.3.2 is feasible, although many details are difficult to supply. The relationship between $\rho_G(s,t)$ and $\rho(s,t)$ is studied in two ways.

First, numerical computations are carried out. The results (not presented here) show that $\rho(s,t)$ of (5.3.2) and $\rho_G(s,t)$ of (5.3.3) give very close results when L of (5.2.1) can be well approximated by l_{BC} of (5.2.11), a situation which occurs very often in application. Since the limiting distributions of A^2 and W^2 depend on the eigenvalues of $\rho_G(s,t)$ (for $\lambda \neq 0$) and $\rho(s,t)$ (for $\lambda = 0$), it should be interesting to compare the limiting distributions of W^2 for various parameter values numerically, based on $\rho(s,t)$ and $\rho_G(s,t)$. For model

Upper Percentiles	Parameter Values				
	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.6$	$\lambda = -0.5$
	any μ	$\mu = 10$	$\mu = 0.0$	$\mu = .5$	$\mu = -13$
	any σ	$\sigma = 0.5$	$\sigma = 0.8$	$\sigma = 1.0$	$\sigma = 1.0$
	$(\delta = \infty)$	$(\delta = 24)$	$(\delta = 3.125)$	$(\delta = 2.167)$	$(-\delta = 15)$
0.0428 (0.50)	0.50043	0.50037	0.52790	0.52063	0.50028
0.0488 (0.40)	0.39992	0.39991	0.40027	0.41042	0.39990
0.0562 (0.30)	0.30010	0.30014	0.29880	0.30039	0.30018
0.0663 (0.20)	0.20082	0.20089	0.19849	0.19240	0.20098
0.0736 (0.15)	0.14972	0.14980	0.14720	0.13839	0.14989
0.0836 (0.10)	0.09996	0.10003	0.09757	0.08769	0.10001
0.1007 (0.05)	0.05005	0.05101	0.04828	0.04003	0.05017
0.1406 (0.01)	0.01002	0.01004	0.00943	0.00649	0.01006

Table 5.8: A comparison of asymptotic significance levels for different λ, μ and σ values based on statistic W^2 when testing goodness-of-fit of Box-Cox transformations.

(5.1.4) with grand mean $\mu = \beta_1$ only, the results are given in Table 5.8.

The entries in the last five columns in above table are the probabilities of observing values greater than the upper percentiles listed in the first column. The numbers inside the parentheses in the first column are from Table 5.1 and are supposed to be the correct asymptotic significance levels. 100 estimated eigenvalues are used in each case to compute the above asymptotic significance levels. The case with $\lambda = 0$ is based on covariance function $\rho(s,t)$ of (5.3.2). In this case, the asymptotic significance levels do not depend on μ and σ. The other four cases are based on $\rho_G(s,t)$ and numerical computations. See section 5.6.

Secondly, the limit of $\rho_G(s,t)$ for the general regression model (5.1.4) is obtained as $\lambda \to 0$, or $\sigma \to 0$, or $\beta_1 \to +\infty$. The results of Theorem 5.3.2 say that $\rho(s,t)$ is indeed the limit of $\rho_G(s,t)$ for small λ, or small σ, or large β_1 values. These cases are common in application, and are indicated by large values of $\delta_i = (\mu_i + 1/\lambda)/\sigma$, as discussed before.

Together, it is fairly clear from Table 5.8 and Theorem 5.3.2 that the limiting distribution of W^2 does depend on unknown parameter values (the same can be said about A^2), but it is conjectured that the covariance function $\rho(s,t)$ of (5.3.2) is approximately valid for general λ, σ and mean values encountered in applications, when model (5.1.4) is taken as the basic underlying model; that A^2, W^2 and any other statistics based on $\rho(s,t)$ are practically parameter-free; and that Table 5.1 of this chapter is applicable for practical purposes.

The upper percentage points in Table 5.1, which are different from those obtained by Linnet (1988), are computed here using the covariance function $\rho(s,t)$ of (5.3.2) and the method of Imhof (1961). A drawback of Linnet's method of generating random samples is that he simulated genuine normal samples first and then transformed the normal samples using the inverse of (5.1.1). Some simulations are done in the present research using the inverse of (5.2.15) to generate random samples and using the Box-Cox transformation procedure to estimate parameters. This latter approach is closer to the way the Box-Cox transformation is usually applied. Fortunately, the simulation results using the above two approaches are close. The asymptotic points in Table 5.1 are calculated theoretically, whereas Linnet's are the results of extrapolation of finite sample simulations. However, the differences are very slight, so Linnet's modified forms W^{**} and A^{**} are used in Table 5.1, which are based on more extensive simulations.

5.6 Proof of Theorem 5.3.1 and Theorem 5.3.2

5.6.1 Proof of Theorem 5.3.1

Proof of (1). Let $\mu = (\mu_1, \ldots, \mu_n)^t = X\beta$. When $\lambda = 0$, $W_i = Y_i(0) = \log Y_i \sim N(\mu_i, \nu)$. Denote $dY_i(\lambda)/d\lambda$ by $\dot{Y}_i(\lambda)$ and $d^2Y_i(\lambda)/d\lambda^2$ by $\ddot{Y}_i(\lambda)$, then $\dot{Y}_i(0) = W_i^2/2$, $\ddot{Y}_i(0) = W_i^3/3$. Straightforward calculations show that the inverse of the Fisher information matrix for

$\theta = (0, \beta^t, \nu)^t$ is given by

$$\Gamma_n = \begin{pmatrix} A_n & B_n \\ B_n^t & C_n \end{pmatrix}^{-1},$$

where

$$n^{-1}A_n = (4\nu)^{-1}(7\nu^2 + 10\nu\beta^t(n^{-1}X^tX)\beta + n^{-1}1_n^t\mu^4),$$

$$n^{-1}B_n = (-D/2 - (n^{-1}X^t\mu^2)^t/(2\nu),\ -\beta_1/\nu),$$

$$n^{-1}C_n = \begin{pmatrix} (n^{-1}X^tX)/\nu & 0 \\ 0^t & 1/(2\nu^2) \end{pmatrix},$$

where $D^t = (1, 0, \ldots, 0)^t$ is a $p \times 1$ vector with its first component equal to 1 and all the other components equal to 0. Therefore, as $n \to \infty$, $n^{-1}\Gamma_n \to \Gamma$ as desired.

Proof of (2). The proof is based on Loynes (1980). In the present case, the null hypothesis $H_n(\gamma)$ in Loynes (1980) specifies nothing and all the parameters $\theta = (\lambda, \beta^t, \nu)^t$ are to be estimated. Since $Y_i(0) = \log Y_i \sim N(\mu_i, \nu)$, without loss of generality, it is assumed that $\beta = 0$ and $\nu = 1$. Then the inverse of the asymptotic variance-covariance matrix for $\theta = (0, 0^t, 1)^t$ is found to be

$$\Gamma = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{7}{6} & 0 & 0 \\ 0 & 0^t & G^{-1} & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix},$$

where $G^{-1} = (\lim_{n\to\infty} n^{-1}U^tU)^{-1}$, $X = (1_n\ U)$. It is readily checked that assumptions $A1$ and $A2$ of Loynes (1980) are satisfied naturally by model (5.1.4). Under assumption (D), it can be checked that assumptions $A4$ and $A5$ of Loynes are also satisfied; the truth of assumptions $A7$ and $A9(b)$ of Loynes can be checked by direct calculations. Therefore, by Theorem 1 of Loynes (1980), the estimated empirical process $\hat{Y}_n(t)$ of (5.3.1) converges weakly to a Gaussian process $Y(t)$. By Corollary 1 of Loynes (1980), the mean of $Y(t)$ is zero and the covariance function of $Y(t)$ is

$$\rho(s,t) = \min(s,t) - st - \Psi^t(s)\Gamma\Psi(t),$$

where $\Psi(t)$ is found to be

$$\Psi(t) = (-\frac{1}{2}J_3(t),\ J_1(t)D^t,\ \frac{1}{2}J_2(t))^t, \qquad (s,t \in [0,1]).$$

Direct computations then show that the expression given in (5.3.2) follows. □

5.6.2 Proof of Theorem 5.3.2

Some Technical Details for (1). The $(p+2)\times(p+2)$ matrix Γ_G and the $(p+2)\times 1$ function $\Psi_G(t)$ mentioned in Theorem 5.3.2 are given below.

Let I_n be the Fisher information matrix for a random sample $Y_1,\ldots,Y_n$ from model (5.1.4). Then

$$\Gamma_G = \frac{1}{n}\lim_{n\to\infty} I_n^{-1},$$

and I_n has the following components:

$$E\left\{-\frac{\partial^2 L}{\partial\lambda\partial\lambda}\right\} = \begin{cases} \sum_{i=1}^n \left\{\frac{1}{\nu}J_{1i} - \frac{\phi(\delta_i)\Phi(\delta_i)(\delta_i-2\lambda\sqrt{\nu})+\phi^2(\delta_i)}{\nu\lambda^4\Phi^2(\delta_i)}\right\}, & \text{if } \lambda>0, \\ \sum_{i=1}^n \left\{\frac{1}{\nu}J_{1i} - \frac{\phi(-\delta_i)\Phi(-\delta_i)(-\delta_i-2\lambda\sqrt{\nu})+\phi^2(-\delta_i)}{\nu\lambda^4\Phi^2(-\delta_i)}\right\}, & \text{if } \lambda<0, \end{cases}$$

$$E\left\{-\frac{\partial^2 L}{\partial\lambda\partial\beta^t}\right\} = \begin{cases} -\frac{1}{\nu}X^tE\{\dot{Y}(\lambda)\} + X^t diag\left(\frac{\delta_i\phi(\delta_i)\Phi(\delta_i)+\phi^2(\delta_i)}{\nu\lambda^2\Phi^2(\delta_i)}\right)1_n, & \text{if } \lambda>0, \\ -\frac{1}{\nu}X^tE\{\dot{Y}(\lambda)\} + X^t diag\left(\frac{-\delta_i\phi(-\delta_i)\Phi(-\delta_i)+\phi^2(-\delta_i)}{\nu\lambda^2\Phi^2(-\delta_i)}\right)1_n, & \text{if } \lambda<0, \end{cases}$$

$$E\left\{-\frac{\partial^2 L}{\partial\lambda\partial\nu}\right\} = \begin{cases} -\frac{1}{\nu^2}E\{(Y(\lambda)-X\beta)^t\dot{Y}(\lambda)\} - \sum_{i=1}^n \frac{\phi(\delta_i)\Phi(\delta_i)(\delta_i^2-1)+\delta_i\phi^2(\delta_i)}{2\nu\sqrt{\nu}\lambda^2\Phi^2(\delta_i)}, & \text{if } \lambda>0, \\ -\frac{1}{\nu^2}E\{(Y(\lambda)-X\beta)^t\dot{Y}(\lambda)\} - \sum_{i=1}^n \frac{\phi(-\delta_i)\Phi(-\delta_i)(\delta_i^2-1)-\delta_i\phi^2(-\delta_i)}{2\nu\sqrt{\nu}\lambda^2\Phi^2(-\delta_i)}, & \text{if } \lambda<0, \end{cases}$$

$$E\left\{-\frac{\partial^2 L}{\partial\beta\partial\beta^t}\right\} = \begin{cases} \frac{1}{\nu}X^tX - \frac{1}{\nu}X^t diag\left(\frac{\delta_i\phi(\delta_i)\Phi(\delta_i)+\phi^2(\delta_i)}{\nu\Phi^2(\delta_i)}\right)X, & \text{if } \lambda>0, \\ \frac{1}{\nu}X^tX - \frac{1}{\nu}X^t diag\left(\frac{-\delta_i\phi(-\delta_i)\Phi(-\delta_i)+\phi^2(-\delta_i)}{\nu\Phi^2(-\delta_i)}\right), & \text{if } \lambda<0, \end{cases} \tag{5.6.1}$$

$$E\left\{-\frac{\partial^2 L}{\partial\beta\partial\nu}\right\} = \begin{cases} X^t diag\left(\frac{\phi(\delta_i)}{\nu\sqrt{\nu}\Phi(\delta_i)} + \frac{\phi(\delta_i)\Phi(\delta_i)(\delta_i^2-1)+\delta_i\phi^2(\delta_i)}{2\nu\sqrt{\nu}\Phi^2(\delta_i)}\right)1_n, & \text{if } \lambda>0, \\ -X^t diag\left(\frac{\phi(-\delta_i)}{\nu\sqrt{\nu}\Phi(-\delta_i)} + \frac{\phi(-\delta_i)\Phi(-\delta_i)(\delta_i^2-1)-\delta_i\phi^2(-\delta_i)}{2\nu\sqrt{\nu}\Phi^2(-\delta_i)}\right)1_n, & \text{if } \lambda<0, \end{cases}$$

$$E\left\{-\frac{\partial^2 L}{\partial\nu\partial\nu}\right\} = \begin{cases} \frac{n}{2\nu^2} - \sum_{i=1}^{n}\left\{\frac{\delta_i\phi(\delta_i)}{\nu^2\Phi(\delta_i)} - \frac{\delta_i\phi(\delta_i)\Phi(\delta_i)(\delta_i^2-3)+\delta_i^2\phi^2(\delta_i)}{4\nu^2\Phi^2(\delta_i)}\right\}, & \text{if } \lambda > 0, \\ \frac{n}{2\nu^2} - \sum_{i=1}^{n}\left\{\frac{-\delta_i\phi(-\delta_i)}{\nu^2\Phi(-\delta_i)} - \frac{-\delta_i\phi(-\delta_i)\Phi(-\delta_i)(\delta_i^2-3)+\delta_i^2\phi^2(-\delta_i)\}}{4\nu^2\Phi^2(-\delta_i)}\right\}, & \text{if } \lambda < 0, \end{cases}$$

where $diag(d_i)$ denotes $n \times n$ diagonal matrices with d_i as its $(i,i)^{th}$ element, 1_n denotes an $n \times 1$ column vector of 1's, J_{1i} is given by, for $\lambda > 0$,

$$\begin{aligned} J_{1i} &= E\{\dot{Y}_i^2(\lambda) + (Y_i(\lambda) - \mu_i)\ddot{Y}_i(\lambda)\} \\ &= (\lambda^4\Phi(\delta_i))^{-1}\int_{-\delta_i}^{+\infty}\phi(v)[\{(1+\lambda\mu_i+\lambda\sigma v)\log(1+\lambda\mu_i+\lambda\sigma v) - \lambda\mu_i - \lambda\sigma v\}^2 \\ &\quad +\lambda\sigma v(1+\lambda\mu_i+\lambda\sigma v)\log^2(1+\lambda\mu_i+\lambda\sigma v) \\ &\quad -2\lambda\sigma v\{(1+\lambda\mu_i+\lambda\sigma v)\log(1+\lambda\mu_i+\lambda\sigma v) - \lambda\mu_i - \lambda\sigma v\}]\,dv, \end{aligned} \tag{5.6.2}$$

where $\mu_i = x_i^t\beta$, $\delta_i = (\mu_i + 1/\lambda)/\sqrt{\nu}$, $\dot{Y}_i(\lambda)$ and $\ddot{Y}_i(\lambda)$ are the first and second derivatives of $Y_i(\lambda)$ with respect to λ, respectively; for the case $\lambda < 0$, the above integrals should be done for the range $-\infty$ to $-\delta_i$ and $\Phi(\delta_i)$ should be replaced by $\Phi(-\delta_i)$.

Similarly, $E\{\dot{Y}(\lambda)\}$ has components J_{2i} given by, for $\lambda > 0$,

$$\begin{aligned} J_{2i} &= E\{\dot{Y}_i(\lambda)\} \\ &= (\lambda^2\Phi(\delta_i))^{-1}\int_{-\delta_i}^{+\infty}\phi(v)[(1+\lambda\mu_i+\lambda\sigma v)\log(1+\lambda\mu_i+\lambda\sigma v) - \lambda\mu_i - \lambda\sigma v]\,dv, \end{aligned} \tag{5.6.3}$$

and $E\{(Y(\lambda) - X\beta)^t\dot{Y}(\lambda)\}$ has components J_{3i} given by, for $\lambda > 0$,

$$\begin{aligned} J_{3i} &= E\{(Y_i(\lambda) - \mu_i)\dot{Y}_i(\lambda)\} \\ &= (\lambda^2\Phi(\delta_i))^{-1}\int_{-\delta_i}^{+\infty}\sigma\phi(v)v[(1+\lambda\mu_i+\lambda\sigma v)\log(1+\lambda\mu_i+\lambda\sigma v) - \lambda\mu_i - \lambda\sigma v]\,dv. \end{aligned} \tag{5.6.4}$$

In the case where $\lambda < 0$, the above two integrals should be done for the range $-\infty$ to $-\delta_i$ and $\Phi(\delta_i)$ should be replaced by $\Phi(-\delta_i)$.

For function $\Psi_G(t)$, there is

$$\Psi_G(t) = \frac{1}{n}\lim_{n\to\infty}\sum_{i=1}^{n}\Psi^{(ni)}(t),$$

where $\Psi^{(ni)}(t)$ is a $(p+2)\times 1$ column vector function with the following components, where $j = 1,\ldots,p$ corresponds to the components associated with β:

$$
\begin{aligned}
\Psi_1^{(ni)}(t) &= \begin{cases} -(\sigma\lambda^2\Phi^2(\delta_i))^{-1}[\phi(w_i)\Phi(\delta_i)\{(1+\lambda\mu_i+\lambda\sigma w_i)\log(1+\lambda\mu_i \\ +\lambda\sigma w_i)-\lambda\mu_i-\lambda\sigma w_i\}+\phi(\delta_i)\Phi(w_i)-\phi(\delta_i)], & \text{if } \lambda>0, \\ -(\sigma\lambda^2\Phi^2(-\delta_i))^{-1}[\phi(v_i)\Phi(-\delta_i)\{(1+\lambda\mu_i+\lambda\sigma v_i)\log(1+\lambda\mu_i \\ +\lambda\sigma v_i)-\lambda\mu_i-\lambda\sigma v_i\}-\phi(-\delta_i)\Phi(v_i)], & \text{if } \lambda<0, \end{cases} \\
\Psi_{2j}^{(ni)}(t) &= \begin{cases} \{x_{ij}/(\sigma\Phi^2(\delta_i))\}\{\phi(w_i)\Phi(\delta_i)+\phi(\delta_i)\Phi(w_i)-\phi(\delta_i)\}, & \text{if } \lambda>0, \\ \{x_{ij}/(\sigma\Phi^2(-\delta_i))\}\{\phi(v_i)\Phi(-\delta_i)-\phi(-\delta_i)\Phi(v_i)\}, & \text{if } \lambda<0, \end{cases} \\
\Psi_3^{(ni)}(t) &= \begin{cases} (2\sigma^2\Phi^2(\delta_i))^{-1}\{\phi(w_i)w_i\Phi(\delta_i)-\delta_i\phi(\delta_i)\Phi(w_i)+\delta_i\phi(\delta_i)\}, & \text{if } \lambda>0, \\ (2\sigma^2\Phi^2(-\delta_i))^{-1}\{\phi(v_i)v_i\Phi(-\delta_i)+\delta_i\phi(-\delta_i)\Phi(v_i)\}, & \text{if } \lambda<0, \end{cases}
\end{aligned}
\tag{5.6.5}
$$

where $w_i = w_i(t) = \Phi^{-1}(1+\Phi(\delta_i)(t-1))$, $v_i = v_i(t) = \Phi^{-1}(t\Phi(-\delta_i))$, and $t\in[0,1]$.

Proof of (2). The $\lambda\to 0$ case is straightforward; the $\sigma\to 0$ case and the $\beta_1\to+\infty$ case can be handled using symbolic computing softwares such as Maple. Ready-to-run Maple files are available from the author. □

Chapter 6

EDF Tests of Composite Hypotheses

The problem of testing composite goodness-of-fit hypothesis $H_0^G : X_1, \ldots, X_n$ is a random sample from a continuous distribution with cumulative distribution function $F(x;\theta)$, where $\theta \in \Theta \subset R^p$ is unknown, is revisited in this chapter. Section 6.1 reviews the EDF test approach to the above mentioned composite hypothesis H_0^G. Through the idea of approximate normality and imaginary parameters, section 6.2 presents a new test procedure based on an approximate link between this general composite hypothesis and the problem of testing normality when both mean and variance are unknown. Heuristic justification of the new procedure is given in section 6.3, and a simulation study of the new procedure is provided in section 6.4. Section 6.5 applies the new procedure to four real data sets, and Section 6.6 ends the chapter with a few comments.

6.1 Introduction

Suppose $X_1, \ldots, X_n$ is a random sample from a continuous distribution. It is often desirable to test or confirm that the random sample is from some particular distribution. Of most

practical interest is the composite hypothesis H_0^G : $X_1, \ldots, X_n$ come from a continuous distribution with cumulative distribution function $F(x;\theta)$, where $\theta \in \Theta \subset R^p$ is unknown. Many authors have addressed this problem and a huge literature under the title goodness-of-fit is now available. See D'Agostino and Stephens (1986) for both general principles and practical techniques.

This chapter, however, will concentrate on the empirical distribution function (EDF) approach to the above mentioned goodness-of-fit problem. Since θ is an unspecified (column) parameter vector and therefore needs to be estimated, complications arise in studying the weak convergence of the underlying empirical processes which are the basis of all EDF statistics, such as Kolmogorov-Smirnov's D, Cramér-von Mises' W^2, Anderson-Darling's A^2 and Watson's U^2. See section 4.1 for definitions of these statistics. Durbin (1973a) made serious investigations on the effect of estimating unknown parameters; Loynes (1980) generalized Durbin's results to the independent but not identically distributed case; Stephens (1986), in a large part, worked out a great deal of details, provided very useful tables and studied the performances of a family of EDF statistics, including the above four.

As far as weak convergence of the underlying empirical processes is concerned, Durbin and Loynes' results are very general, provided the usual requirements on Fisher's information matrix are met and, in particular, the estimator $\hat{\theta}$ of θ is efficient in the sense that $n^{\frac{1}{2}}(\hat{\theta} - \theta)$ can be written as a sum of independent random variables with zero mean, plus an $o_p(1)$ term. When θ consists of only location and scale parameters, the limiting Gaussian processes of the underlying empirical processes do not depend on θ. This nice feature is however no longer present when θ involves a shape parameter, making tabulation of limiting distributions a heavy job. In this last case, few tables have been produced for EDF statistics. See Stephens (1986).

Because there are so many interesting distributions being employed in application, for each one a table needs to be created. As EDF tests for the case with unknown parameters are mostly asymptotic tests, a great deal of effort have been made through extensive simulations

to handle the realistic finite sample size situations. This is manifested by the large number of tables in D'Agostino and Stephens (1986). It is not bad that tables are created for each and every single distribution, but it would be nice if one table could serve most interesting distributions in a practical manner, especially when such a table has been created.

6.2 A New Procedure for Testing Goodness-of-Fit

The hypothesis is H_0^G: random sample $X_1, \ldots, X_n$ come from a continuous distribution with cumulative distribution function $F(x;\theta)$, where $\theta \in \Theta \subset R^p$ is unknown. For a fixed θ, applying the probability integral transformation to the X-sample gives a U-sample $U_1, \ldots, U_n$, where $U_i = F(X_i;\theta)$ $(i = 1, 2, \ldots, n)$. When θ is the true parameter for the X-sample, $U_1, \ldots, U_n$ will be an independent and identically distributed sample from the uniform $U(0,1)$.

Now let $F_n(x)$ denote the empirical distribution function of the X-sample, that is,

$$F_n(x) = n^{-1} \sum_{i=1}^{n} 1[x_i \leq x],$$

where $1[a \leq b] = 1$ if $a \leq b$, and $1[a \leq b] = 0$ if $a > b$. Any statistic that measures the difference between F_n and F will be called an EDF statistic. See section 4.1 for definitions of supremum and integral EDF statistics.

For the ease of reference and comparison, the computational formulas for D, W^2, U^2 and A^2 are reproduced here together with Stephens' procedure for testing normality.

For a given X-sample $x_1, \ldots, x_n$, let $u_i = F(x_i;\hat{\theta})$ $(i = 1, 2, \ldots, n)$. Without loss of generality, suppose x_i's and u_i's have been arranged into ascending order. Then

$$D = \max\left\{ \max_{1 \leq i \leq n} (i/n - u_i), \max_{1 \leq i \leq n} (u_i - (i-1)/n) \right\}, \tag{6.2.1}$$

$$W^2 = \sum_{i=1}^{n} [u_i - \{(2i-1)/(2n)\}]^2 + 1/(12n), \tag{6.2.2}$$

$$U^2 = W^2 - n\{0.5 - n^{-1} \sum_{i=1}^{n} u_i\}^2, \tag{6.2.3}$$

$$A^2 = -n - n^{-1}\sum_{i=1}^{n}\{(2i-1)\ln u_i + (2n+1-2i)\ln(1-u_i)\}. \qquad (6.2.4)$$

To test H_0^N: $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$, where μ and σ^2 are both unknown, Stephens' procedure proceeds as below:

(a) Compute $w_i = (x_i - \bar{x})/s_x$, where $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$ and $s_x^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$,

(b) Compute $u_i = \Phi(w_i)$, where $\Phi(w)$ is the cdf of a $N(0,1)$ random variable,

(c) Calculate D, W^2, U^2 and A^2 according to (6.2.1) to (6.2.4),

(d) Modify D into $D^* = D(n^{\frac{1}{2}} - 0.01 + 0.85/n^{\frac{1}{2}})$, W^2 into $W^* = W^2(1 + 0.5/n)$, U^2 into $U^* = U^2(1 + 0.5/n)$, and A^2 into $A^* = A^2(1 + 0.75/n + 2.25/n^2)$, where n is the sample size, and reject H_0^N at significance level α if the modified statistics exceed the upper tail significance points as given in Table 4.7, D'Agostino and Stephens (1986), page 123. See Table 4.1 in Chapter 4 for quick reference.

To test H_0^G: $X_1, \ldots, X_n$ is a random sample from a continuous distribution with cumulative distribution function $F(x;\theta)$, where $\theta \in \Theta \subset R^p$ is unknown, proceed as below:

(1) Estimate θ efficiently by $\hat{\theta}$ and compute $v_i = F(x_i;\hat{\theta})$, where the x_i's are in ascending order,

(2) Compute $y_i = \Phi^{-1}(v_i)$,

(3) Compute $u_i = \Phi\{(y_i - \bar{y})/s_y\}$, where $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ and $s_y^2 = (n-1)^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$,

(4) Calculate D, W^2, U^2 and A^2 according to (6.2.1) to (6.2.4),

(5) Modify D into $D^* = D(n^{\frac{1}{2}} - 0.01 + 0.85/n^{\frac{1}{2}})$, W^2 into $W^* = W^2(1 + 0.5/n)$, U^2 into $U^* = U^2(1 + 0.5/n)$, and A^2 into $A^* = A^2(1 + 0.75/n + 2.25/n^2)$, where n is the sample size, and reject H_0^G at significance level α if the modified statistics exceed the upper tail significance points as given in Table 4.7, D'Agostino and Stephens (1986), page 123. Also see Table 4.1 in Chapter 4 for quick reference.

6.3 Heuristic Justification of the New Procedure

The empirical process related to testing normality as stated in H_0^N is constructed by letting, for $i = 1, 2, \ldots, n$,

$$d_i = \Phi(W_i) = \Phi\{(X_i - \bar{X})/s_x\},$$

where $s_x^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ and defining, for $0 \leq t \leq 1$,

$$X_n(t) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \{1[d_i \leq t] - t\}. \tag{6.3.1}$$

It is well known (Loynes, 1980) that $X_n(t)$ converges weakly to a Gaussian process $X(t)$ with zero mean and covariance function

$$\rho(s,t) = \mathrm{Cov}(X(s), X(t)) = \min(s,t) - st - J_1(s)J_1(t) - \frac{1}{2}J_2(s)J_2(t), \tag{6.3.2}$$

where $s, t \in [0,1]$, $J_1(t) = \phi(\Phi^{-1}(t))$, $J_2(t) = \phi(\Phi^{-1}(t))\Phi^{-1}(t)$, and $\phi(x)$ is the density of a standard normal random variable.

Let $\theta = (\mu, \sigma)^t$ be the true parameter for the X_i's under H_0^N. Let $Z_1, \ldots, Z_n$ denote independent and identically distributed standard normal random variables. Then, $X_i = \sigma Z_i + \mu$ in distribution and

$$d_i = \Phi(W_i) = \Phi\{(Z_i - \bar{Z})/s_z\},$$

where $s_z^2 = (n-1)^{-1} \sum_{i=1}^n (Z_i - \bar{Z})^2$. In this version of the d_i's, it is clear that the distribution of $X(t)$ does not depend on the unknown parameter $\theta = (\mu, \sigma)^t$.

Now consider H_0^G. Suppose θ is the true parameter for the X_i's under H_0^G, and suppose $F(x;\theta)$ has continuous partial derivatives with respect to θ. For any estimator $\hat{\theta}$ of θ, expanding $F(x;\hat{\theta})$ about θ gives

$$\begin{aligned} V_i &= F(X_i;\hat{\theta}) \\ &= F(X_i;\theta) + F'(X_i;\eta_i)(\hat{\theta} - \theta) \\ &= U_i + F'(X_i;\eta_i)(\hat{\theta} - \theta), \end{aligned}$$

where $U_1, \ldots, U_n$ are independent and identically distributed as uniform $U(0,1)$, η_i lies between θ and $\hat{\theta}$, and the prime denotes partial derivative with respect to θ (a row vector). According to step (2) of the new procedure, $\Phi^{-1}(.)$ is applied to transform V_i's; this gives

$$\begin{aligned} Y_i &= \Phi^{-1}(V_i) \\ &= \Phi^{-1}(U_i) + F'(X_i;\eta_i)(\hat{\theta}-\theta)/J_1(\delta_i) \\ &= Z_i + F'(X_i;\eta_i)(\hat{\theta}-\theta)/J_1(\delta_i), \end{aligned}$$

where δ_i lies between V_i and U_i, $J_1(t) = \phi(\Phi^{-1}(t))$ and Z_i are iid $N(0,1)$ random variables.

Let $E_i = F'(X_i;\eta_i)(\hat{\theta}-\theta)/J_1(\delta_i)$ $(i = 1,2,\ldots,n)$. It can be seen that the Y_i's are the true normal random variables Z_i's contaminated by random quantities E_i's. In this sense, the Y_i's have approximate normality. The new procedure is to treat the Y_i's as if they were independent and identically distributed normal variables with unknown mean μ and unknown variance σ^2. The parameters in this treatment are of course imaginary. However, the following heuristic arguments attempt to justify, in an approximate sense, the use of normality and parameters in the way just described.

For the Y_i defined above, let $s_y^2 = (n-1)^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$. Then the weak limit of

$$B_n(t) = n^{-1/2}\sum_{i=1}^{n}\left\{1\left[\Phi\left(\frac{Y_i-\bar{Y}}{s_y}\right)\le t\right] - t\right\} \tag{6.3.3}$$

needs to be found. However, $B_n(t)$ can be rewritten into

$$\begin{aligned} B_n(t) &= n^{-1/2}\textstyle\sum_{i=1}^n \left\{1\left[\Phi\left(\frac{Y_i-\bar{Y}}{s_y}\right)\le t\right] - t\right\} \\ &= n^{-1/2}\textstyle\sum_{i=1}^n \{1[Y_i \le \bar{Y} + s_y\Phi^{-1}(t)] - t\} \\ &= n^{-1/2}\textstyle\sum_{i=1}^n \{1[V_i \le \Phi(\bar{Y} + s_y\Phi^{-1}(t))] - t\} \\ &= n^{-1/2}\textstyle\sum_{i=1}^n \{1[X_i \le F^{-1}(\Phi(\bar{Y} + s_y\Phi^{-1}(t));\hat{\theta})] - t\} \\ &= n^{-1/2}\textstyle\sum_{i=1}^n \{1[U_i \le F(F^{-1}(\Phi(\bar{Y} + s_y\Phi^{-1}(t));\hat{\theta});\theta)] - t\} \end{aligned}$$

and this can be further decomposed into

$$\begin{aligned} B_n(t) &= n^{-1/2}\sum_{i=1}^{n}\{1[U_i \le F(F^{-1}(\Phi(\bar{Y}+s_y\Phi^{-1}(t));\hat{\theta});\theta)] \\ &\quad -F(F^{-1}(\Phi(\bar{Y}+s_y\Phi^{-1}(t));\hat{\theta});\theta) - 1[U_i \le \Phi(\bar{Z}+s_z\Phi^{-1}(t))] \\ &\quad +\Phi(\bar{Z}+s_z\Phi^{-1}(t))\} \\ &+ \sqrt{n}\{\Phi(\bar{Z}+s_z\Phi^{-1}(t)) - t\} \\ &+ \sqrt{n}\{F(F^{-1}(\Phi(\bar{Y}+s_y\Phi^{-1}(t));\hat{\theta});\theta) - \Phi(\bar{Z}+s_z\Phi^{-1}(t))\} \\ &+ n^{-1/2}\sum_{i=1}^{n}\{1[U_i \le \Phi(\bar{Z}+s_z\Phi^{-1}(t))] - \Phi(\bar{Z}+s_z\Phi^{-1}(t)) \\ &= B_{1n}(t) + B_{2n}(t) + B_{3n}(t) + B_{4n}(t), \text{ say.} \end{aligned}$$

Because

$$\sup_{0\le t\le 1}|\Phi(\bar{Z}+s_z\Phi^{-1}(t)) - t| = o_p(1),$$

$$B_{2n}(t) = J_1(t)n^{-1/2}\sum_{i=1}^{n} Z_i + 2^{-1/2}J_2(t)n^{-1/2}\sum_{i=1}^{n}\left\{\frac{Z_i^2-1}{\sqrt{2}}\right\} + o_p(1),$$

it follows that $B_{2n}(t) + B_{4n}(t)$ converges weakly to the Gaussian process with zero mean and covariance function (6.3.2), that is, the weak limit for testing normality of a random sample with unknown mean and variance is reached.

From the identity $F(F^{-1}(x;\theta);\theta) = x$, it follows that for $\hat{\theta}$ close to θ, there is

$$F(F^{-1}(\Phi(\bar{Y}+s_y\Phi^{-1}(t));\hat{\theta});\theta) \approx \Phi(\bar{Y}+s_y\Phi^{-1}(t)).$$

Also, $\hat{\theta}$ close to θ implies $\bar{Y} \approx \bar{Z}$ and $s_y \approx s_z$. Therefore, for estimator $\hat{\theta}$ such that $\sqrt{n}(\hat{\theta}-\theta) = O_p(1)$, it follows, according to section 2.4.3, that $B_{1n}(t) = o_p(1)$.

The last term left is $B_{3n}(t)$. In general, $B_{3n}(t) = o_p(1)$ does not hold, therefore, the weak limit of $B_n(t)$ can not be the same as the weak limit for testing normality when both mean and variance are unknown. However, simulation results (shown in next section) indicate that ignoring the contribution from $B_{3n}(t)$ does not cause serious problems for some commonly used distributions.

6.4 A simulation study of the New Procedure

Simulations have been carried out, using the new procedure, for the following distributions whose densities $f(.)$ or distribution functions $F(.)$ are given by:

(1) Normal: $\Phi\{(x-\mu)/\sigma\}$, $\mu \in R$, $\sigma > 0$;

(2) Exponential: $f(x;\alpha,\beta) = \beta^{-1}\exp\{-(x-\alpha)/\beta\}$, $\alpha \in R$, $\beta > 0$;

(3) Extreme-value: $F(x;\alpha,\beta) = \exp\{-\exp\{-(x-\alpha)/\beta\}\}$, $\alpha \in R$, $\beta > 0$;

(4) Weibull: $F(x;\beta,m) = 1 - \exp\{-(x/\beta)^m\}$, $\beta > 0$, $m > 0$;

(5) Gamma: $f(x;\beta,m) = \beta^{-m}\Gamma^{-1}(m)x^{m-1}\exp\{-x/\beta\}$, $\beta > 0$, $m > 0$;

(6) Lognormal: $F(x;\mu,\sigma^2) = \Phi\{(\log(x)-\mu)/\sigma\}$, $\mu \in R$, $\sigma > 0$;

(7) Inverse Gaussian: $f(x;\mu,\lambda) = (\lambda/(2\pi))^{\frac{1}{2}}x^{-\frac{3}{2}}\exp[-\{\lambda(x-\mu)^2\}/\{2\mu^2 x\}]$, $\lambda > 0$, $\mu > 0$.

The method of maximum likelihood is used to estimate the unknown parameters in all cases except in (1) where σ^2 is estimated by the unbiased version of the sample variance and in (2) where α is estimated by the minimum variance unbiased estimator $(nX_{(1)}-\bar{X})/(n-1)$. Three sample sizes are studied for each distribution: $n = 10$, $n = 20$, and $n = 30$. For each sample size and each distribution, 1000 random samples were simulated and the new procedure was applied to each random sample at significance level 0.05. The entries in Table 6.1 are the proportions of rejections based on the simulated random samples.

It can be seen from Table 6.1 that all the four EDF statistics behave well under the null hypothesis H_0^G. The new procedure gives about the right significance level 5%, except for testing the exponential distribution, where the new procedure is a bit conservative. An interesting thing to notice here is that the new procedure works very well for sample size as small as ten. Notice also that under the null hypothesis H_0^G, the four EDF statistics behave equally well. On the whole, it is fair to say that the performance of the new procedure is acceptable under the null hypothesis H_0^G.

Distribution	Sample Size	D	W^2	U^2	A^2
Normal	n=10	0.051	0.046	0.049	0.049
$\mu = 0,\ \sigma = 1$	n=20	0.053	0.048	0.046	0.049
	n=30	0.050	0.052	0.054	0.053
Exponential	n=10	0.050	0.041	0.041	0.049
$\alpha = 2,\ \beta = 3$	n=20	0.046	0.038	0.044	0.034
	n=30	0.039	0.039	0.041	0.043
Extreme-value	n=10	0.052	0.047	0.049	0.044
$\alpha = 2,\ \beta = 3$	n=20	0.056	0.056	0.055	0.055
	n=30	0.047	0.045	0.040	0.041
Weibull	n=10	0.055	0.045	0.047	0.048
$\beta = 3,\ m = 2.5$	n=20	0.052	0.052	0.053	0.053
	n=30	0.046	0.053	0.050	0.053
Gamma	n=10	0.047	0.043	0.049	0.043
$\beta = 1,\ m = 2.5$	n=20	0.048	0.049	0.047	0.054
	n=30	0.055	0.051	0.054	0.053
Lognormal	n=10	0.048	0.048	0.049	0.051
$\mu = 2,\ \sigma = 3.5$	n=20	0.042	0.048	0.049	0.049
	n=30	0.051	0.042	0.045	0.047
Inverse Gaussian	n=10	0.039	0.034	0.042	0.045
$\mu = 1.5,\ \lambda = 2$	n=20	0.049	0.054	0.051	0.054
	n=30	0.049	0.047	0.051	0.051

Table 6.1: Simulation study of testing goodness-of-fit for seven distributions using the new procedure when all parameters are unknown and estimated from the data. 1000 samples were simulated for each sample size and distribution combination, and three sample sizes 10, 20 and 30 were considered. The entries are the proportions of rejections when the new procedure is applied at significance level $\alpha = 5\%$.

Simulated as	Estimated as	Sample Size	D	W^2	U^2	A^2
Weibull $\beta = 3,\ m = 2.5$	Gamma	n=20	0.101	0.113	0.104	0.125
		n=40	0.144	0.167	0.143	0.182
	Lognormal	n=20	0.221	0.267	0.240	0.299
		n=40	0.371	0.470	0.411	0.520
	Inverse G	n=20	0.263	0.323	0.230	0.361
		n=40	0.456	0.533	0.479	0.584
Gamma $\beta = 2,\ m = 3$	Weibull	n=20	0.077	0.082	0.075	0.085
		n=40	0.093	0.088	0.076	0.101
	Lognormal	n=20	0.084	0.101	0.097	0.102
		n=40	0.149	0.182	0.160	0.195
	Inverse G	n=20	0.149	0.165	0.155	0.176
		n=40	0.234	0.273	0.242	0.307
Lognormal $\mu = 2,\ \sigma = 3$	Weibull	n=20	0.194	0.233	0.192	0.257
		n=40	0.311	0.382	0.329	0.430
	Gamma	n=20	0.089	0.088	0.084	0.091
		n=40	0.108	0.133	0.113	0.145
	Inverse G	n=20	0.487	0.603	0.565	0.624
		n=40	0.820	0.905	0.874	0.926
Inverse Gaussian $\mu = 2,\ \lambda = 5$	Weibull	n=20	0.162	0.221	0.198	0.254
		n=40	0.347	0.430	0.356	0.477
	Gamma	n=20	0.105	0.121	0.120	0.121
		n=40	0.152	0.191	0.157	0.226
	Lognormal	n=20	0.040	0.042	0.041	0.042
		n=40	0.051	0.042	0.043	0.044

Table 6.2: Simulated powers in testing goodness-of-fit using the new procedure for four distributions when all parameters are unknown and estimated from the data. 1000 samples were simulated for each combination of sample size and alternative. Two sample sizes 20 and 40 were considered. The entries are the proportions of rejections when the new procedure is applied at significance level $\alpha = 5\%$.

Before going into power studies, it is noted that Stephens (1986) has provided tables to test distributions (1) to (5) using EDF statistics. It is interesting to see that the asymptotic points for extreme-value, Weibull and Gamma (with large shape parameter, say, ≥ 8) distributions are very close to the asymptotic points for the normal distribution.

Some power simulations are also carried out. Given that the EDF statistics behave well under the null hypothesis, the goal here is to see whether these same EDF statistics are sensitive enough to alternatives. To this end, it is decided to focus on Weibull, Gamma, Lognormal and Inverse Gaussian distributions. This choice is made on purpose, because it is known that with small samples it is difficult to distinguish these four distributions, therefore the EDF statistics are in this case under severe tests

It can be seen from Table 6.2 that in all cases the power increases when the sample size changes from 20 to 40. This is expected. Also, in all cases the EDF statistics show some power in telling the differences among Weibull, Gamma, Lognormal and Inverse Gaussian distributions, except for one case where Inverse Gaussian random samples are treated as Lognormal random samples and the EDF statistics do not find anything wrong. But it is interesting to note that when Lognormal random samples are treated as Inverse Gaussian random samples, the EDF statistics give the strongest warnings in Table 6.2. This is, of course, an interesting case only—the result might depend, among other things, on the parameter values used here; see Figure 6.1 for an explanation. Notice that as expected the Anderson-Darling statistic A^2 is the most powerful among the four EDF statistics. In general, the EDF statistics can tell differences between the Inverse Gaussian and the other three distributions best.

Attempts have also been made to simulate three-parameter distributions. In this case, all the EDF statistics tend to be conservative. For example, when $n = 30$, the following three-parameter lognormal distribution with $\alpha = -1$, $\mu = 0.5$ and $\sigma = 1.5$ was simulated 1000 times:

$$F(x; \alpha, \mu, \sigma^2) = \Phi[\{\log(x - \alpha) - \mu\}/\sigma].$$

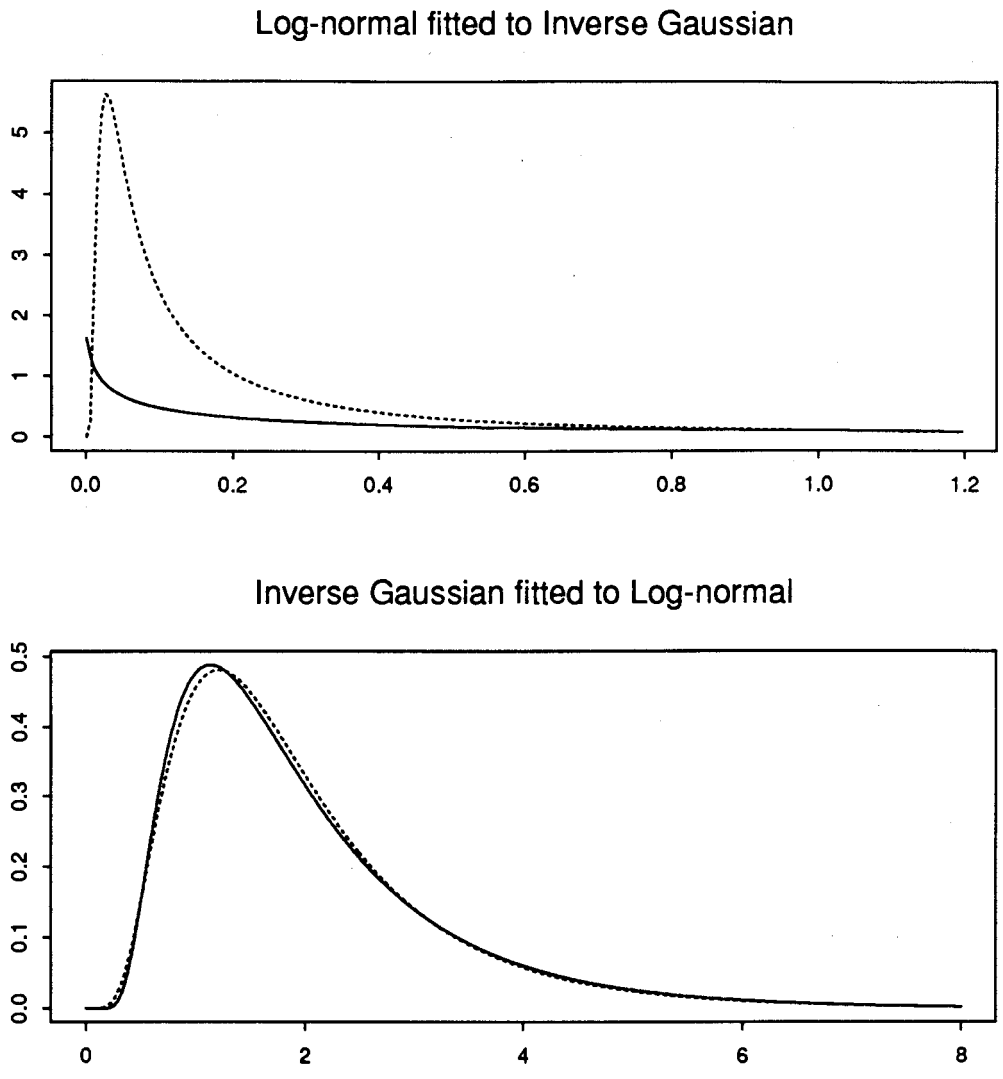


Figure 6.1: The solid lines are the densities of the simulated populations. The dashed lines are the densities from the fitted families of distributions, chosen in such a way that they have the same means and variances as those of the corresponding simulated populations. As can be seen, there is no density from the Inverse Gaussian family that can mimic the simulated log-normal density with $\mu = 2$ and $\sigma = 3$, but there is a density from the log-normal family that can mimic the simulated Inverse Gaussian density with $\mu = 2$ and $\lambda = 5$ very well. This explains the differences in power studies noticed in Table 6.2.

The new procedure was then applied to the simulated samples and the simulated 5% levels are 0.034 for D, 0.042 for W^2, 0.021 for U^2 and 0.024 for A^2. Because a rather involved iterative search is necessary to find maximum likelihood estimators of the three parameters, the conservative behaviour of the EDF statistics needs to be investigated further.

6.5 Examples

Next, the new procedure is applied to four real data sets. Table 4.1 of Chapter 4 should be consulted for P-values.

Example 6.1. Proschan, F. (1963) gave 15 intervals between failures of air conditioning equipment in aircraft. The data values are: 12, 21, 26, 27, 29, 29, 48, 57, 59, 70, 74, 153, 326, 386, 502. To test $H_0 : X_1, X_2, \ldots, X_{15}$ is a random sample from exponential distribution $F(x; \alpha, \beta) = 1 - \exp\{-(x - \alpha)/\beta\}$, where α and β are unknown, one obtains $\hat{\alpha} = 4.195$, and $\hat{\beta} = 117.1$. It follows that the modified EDF statistics are $D^* = 0.9774$, $W^* = 0.1547$, $U^* = 0.1389$ and $A^* = 0.8991$. The p-values are all less than 0.05, therefore the exponential assumption is not appealing. See Figure 6.2.

Example 6.2. Gumbel, E. J. (1964)[1] studied the yearly maximum water discharges of the Ocmulgee River measured at a location called Macon. The data look like this: 4.8, 7.3, 7.9, 8.5, 10.7, 14.2, 14.3, 16.9, 19.0, 19.1, 19.6, 21.0, 22.7, 24.0, 25.4, 28.3, 28.3, 28.8, 31.0, 31.0, 32.6, 33.3, 33.9, 37.0, 40.0, 44.8, 47.1, 47.8, 50.2, 51.0, 57.6, 64.4, 65.3, 66.2, 72.5, 73.4, 73.4, 98.6, 84.0. To test $H_0 : X_1, X_2, \ldots, X_{40}$ is a random sample from extreme-value distribution $F(x; \alpha, \beta) = \exp[-\exp\{-(x - \alpha)/\beta\}]$, where α and β are unknown, the maximum likelihood estimates are found as $\hat{\alpha} = 26.7130$ and $\hat{\beta} = 17.6061$. The modified EDF statistics are $D^* = 0.5396$, $W^* = 0.03795$, $U^* = 0.03792$ and $A^* = 0.3114$. No statistic exceeds its corresponding 15% significance level, therefore, the extreme-value distribution assumption is reasonable. Also see Figure 6.2.

[1]The data are taken from E. Castillo (1988).

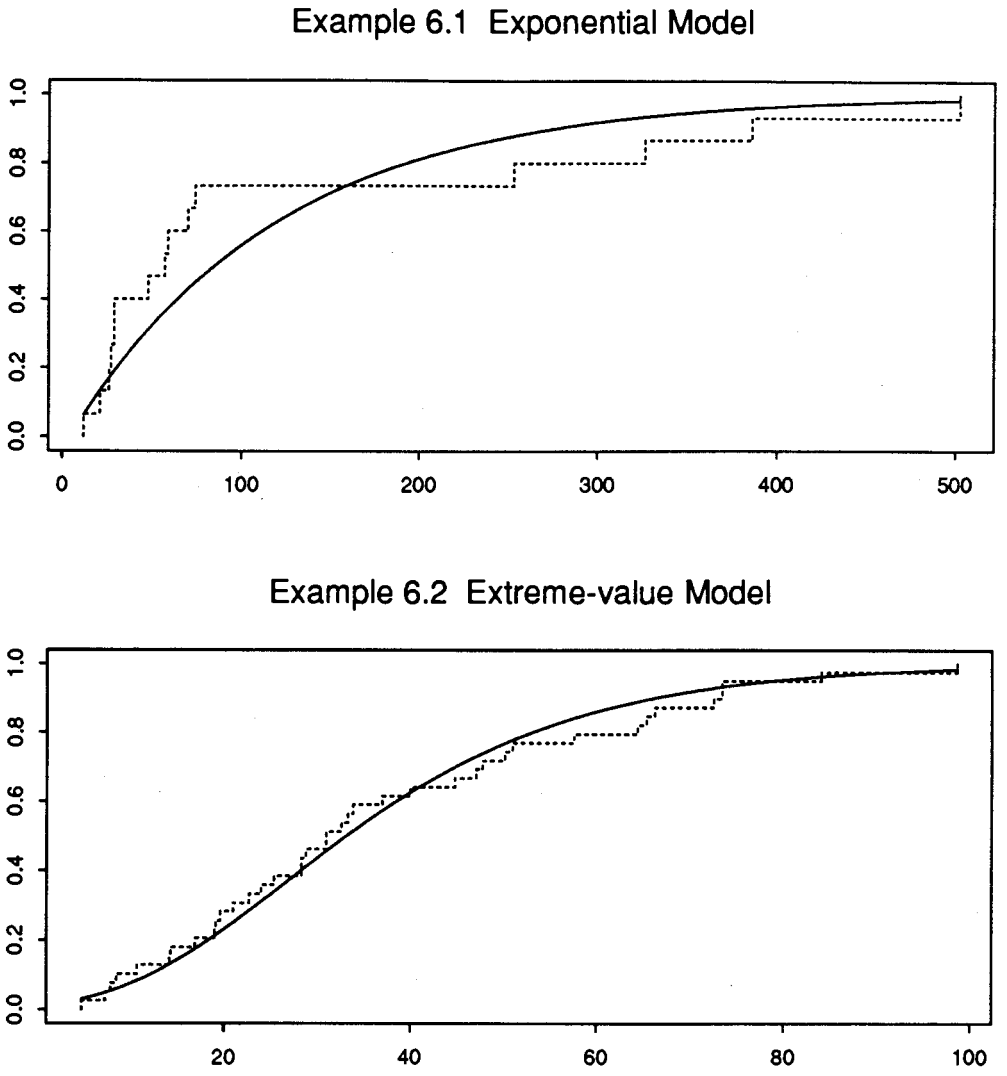


Figure 6.2: Plots of empirical (dotted lines) and estimated (solid lines) distribution functions for Example 6.1 (Exponential model) and Example 6.2 (Extreme-value model).

Example 6.3. Dumonceaux, R. and C. E. Antle (1973) cited data of maximum flood levels in millions of cubic feet per second for Susquehanna River at Harrisburg, Pennsylvania, over 20 four-year periods from 1890 to 1969: 0.654, 0.613, 0.315, 0.449, 0.297, 0.402, 0.379, 0.423, 0.379, 0.3235, 0.269, 0.740, 0.418, 0.412, 0.494, 0.416, 0.338, 0.392, 0.484, 0.265. Four three-parameter models are considered in this example. They are

Weibull Model:

$$F(x;\alpha,\beta,m) = 1 - \exp\left[-\left\{\frac{x-\alpha}{\beta}\right\}^m\right],$$

where $\alpha < x$, $0 < \beta$, $0 < m$;

Gamma Model:

$$f(x;\alpha,\beta,m) = \frac{1}{\beta^m\Gamma(m)}(x-\alpha)^{m-1}\exp\left\{-\frac{x-\alpha}{\beta}\right\},$$

where $\alpha < x$, $0 < \beta$, $0 < m$;

Lognormal Model:

$$F(x;\alpha,\mu,\sigma^2) = \Phi\left[\frac{\log(x-\alpha)-\mu}{\sigma}\right],$$

where $\alpha < x$, $\mu \in R$, $0 < \sigma$;

Inverse Gaussian Model:

$$f(x;\alpha,\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\left\{\frac{\mu}{x-\alpha}\right\}^{\frac{3}{2}}\exp\left[-\frac{1}{2}\left\{\frac{\mu}{x-\alpha}\right\}\left\{\frac{x-\alpha-\mu}{\sigma}\right\}^2\right],$$

where $\alpha < x$, $0 < \mu$, $0 < \sigma$.

Parameter estimates and modified EDF statistics are summarized in Table 6.3. It can be seen that the Gamma distribution gives the worst fit and the Lognormal distribution provides the best fit among the four competing models. The Inverse Gaussian distribution gives a fit that is very close to that given by the Lognormal distribution. See Figures 6.3 and 6.4.

Example 6.4. Steen and Stickler (1976)[2] reported the pollution data, measured in number of coliform per m, on 20 days over a five-week period at Cold Knap Beach, South

[2]The data are taken from R. C. H. Cheng and N. A. K. Amin (1981).

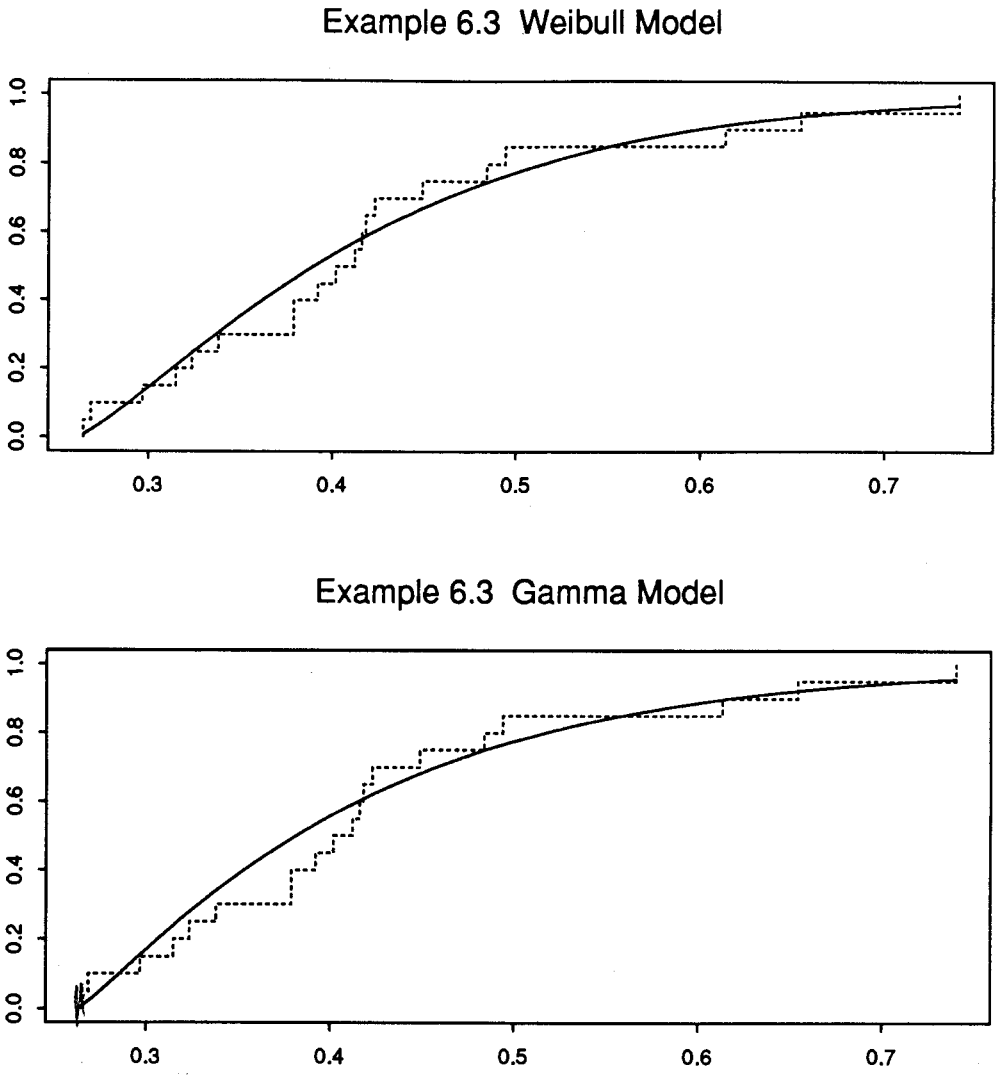
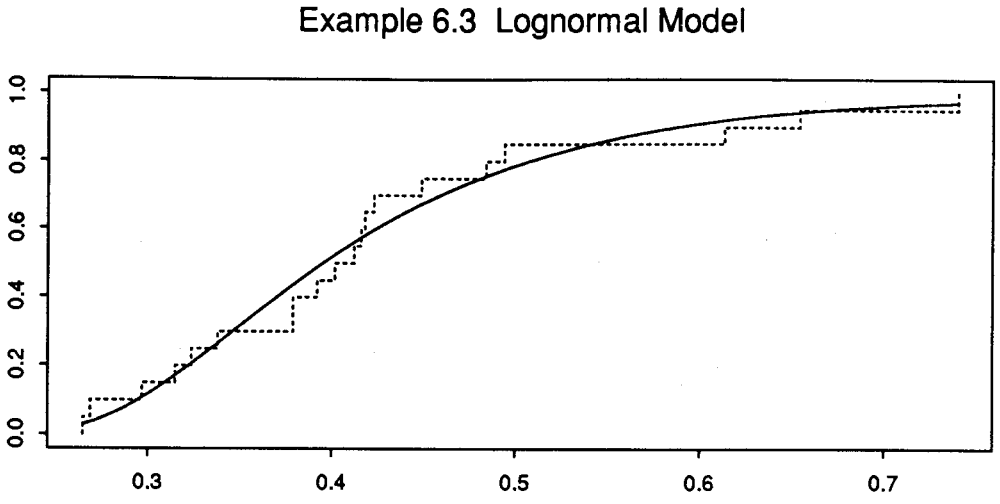


Figure 6.3: Plots of empirical (dotted lines) and estimated (solid lines) distribution functions for Example 6.3, Weibull and Gamma models.



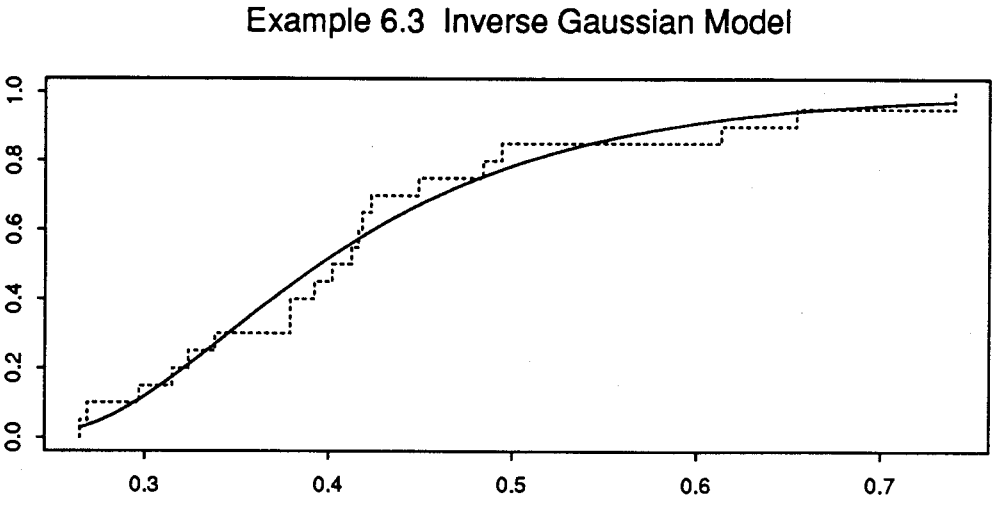


Figure 6.4: Plots of empirical (dotted lines) and estimated (solid lines) distribution functions for Example 6.3, Lognormal and Inverse Gaussian models.

Wales. The data are: 200, 6091, 336, 327, 154, 109, 111, 282, 2120, 1082, 918, 718, 482, 1345, 53600, 5900, 1918, 900, 1045, 1454.

Distribution	Parameter Estimates	Modified EDF Statistics			
		D^*	W^*	U^*	A^*
Weibull	$\hat{\alpha} = .2611\ \hat{\beta} = .1727\ \hat{m} = 1.2445$	.7777	.0752	.0721	.4283
Gamma	$\hat{\alpha} = .2628\ \hat{\beta} = .1343\ \hat{m} = 1.1943$	.8631	.0954	.0876	.5532
Lognormal	$\hat{\alpha} = .1850\ \hat{\mu} = -1.5608\ \hat{\sigma} = .5073$	.6505	.0490	.0488	.2833
IGaussian	$\hat{\alpha} = .1782\ \hat{\mu} = .2450\ \hat{\sigma} = .1268$	.6573	.0506	.0505	.2905

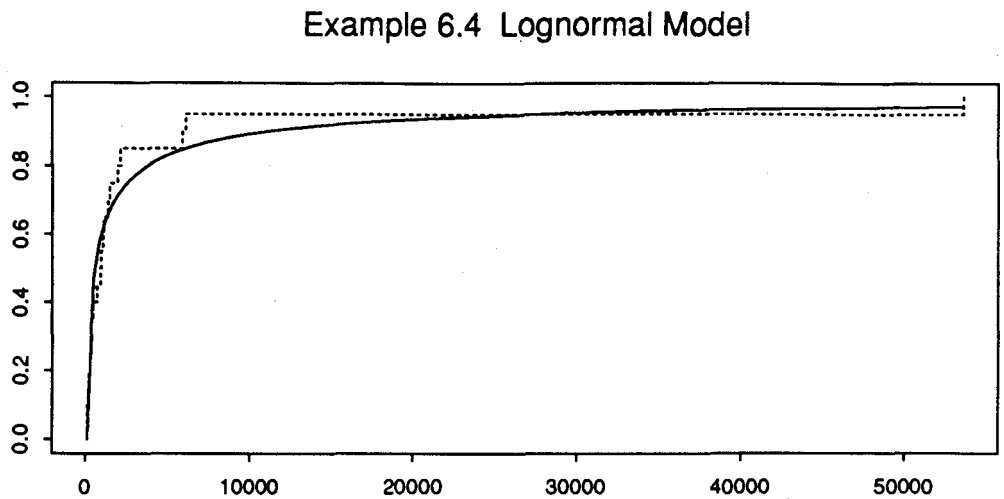
Table 6.3: Fits of Weibull, Gamma, Lognormal and Inverse Gaussian distributions to Susquehanna River flood levels data, Example 6.3. Parameter estimates and the modified EDF statistics are shown in the table.

In this case, both the Weibull and the Gamma distributions can not be fitted to the data by the method of maximum likelihood, while the Lognormal and the Inverse Gaussian distributions can be fitted, giving results in Table 6.4. Note that the Inverse Gaussian

Distribution	Parameter Estimates	Modified EDF Statistics			
		D	W^2	U^2	A^2
Lognormal	$\hat{\alpha} = 108.4746\ \hat{\mu} = 6.0849\ \hat{\sigma} = 2.5150$	.7429	.1337	.1211	.8026
IGaussian	$\hat{\alpha} = 44.33\ \hat{\mu} = 3910.27\ \hat{\sigma} = 13718.81$	.5138	.0391	.0387	.2703

Table 6.4: Fits of Lognormal and Inverse Gaussian distributions to Cold Knap Beach pollution data, Example 6.4. Parameter estimates and the modified EDF statistics are shown in the table.

distribution fits the data a lot better than the Lognormal distribution does and clearly shows its potential applicability in modeling long tailed distributions. See Figure 6.5.



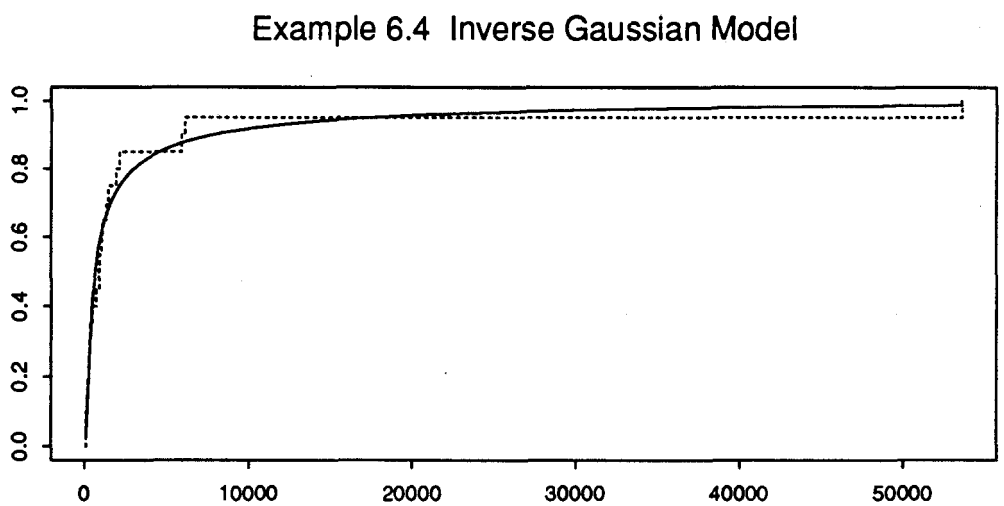


Figure 6.5: Plots of empirical (dotted lines) and estimated (solid lines) distribution functions for Example 6.4, Lognormal and Inverse Gaussian models.

6.6 Comments

It has been demonstrated that the new procedure for testing composite goodness-of-fit hypotheses given in section 6.2 works fairly well. It is easy to apply and does not require extensive tables. This second feature is especially pleasant when the distributions to be tested contain shape parameters. It is expected that the new procedure can be adapted to the situations where only censored data are available. Another potential application is to check goodness-of-fit for continuous mixture models, where even the large sample likelihood ratio method is difficult to apply.

A normality assumption is usually needed to derive exact results. Afterwards, various approximations can be made. The idea of approximate normality used in this chapter points to a general way of linking a problem about non-normal distributions to a problem essentially about normal distributions.

The idea of imaginary parameters reflects the need to adjust for the transition from a non-normal problem to a normal problem. Although it is difficult to formulate a general statement, it seems that there is an approximate equivalence between estimating parameters of a non-normal continuous distribution $F(x;\theta)$ and estimating the mean and the variance of a normal distribution, as long as $F(x;\theta)$ is continuously differentiable with respect to θ. The simulation results shown in section 6.4 indicate that the new procedure works approximately when $\sqrt{n}$-consistent estimator $\hat{\theta}$ is available.

It should be pointed out that the limited simulation results presented in this chapter serve only the purpose of demonstration. Larger simulation studies are needed, especially for distributions with three unknown parameters.

Chapter 7

Proposals

7.1 Generalized Linear Models (GLIM)

When the likelihood ratio statistic, or in GLIM terminology, the deviance statistic, is used to test for goodness-of-fit when fitting generalized linear regression models (GLIM), the test checks whether the difference in likelihood between two nested models is significant or not (except when testing for goodness-of-fit of the full model). This, however, depends on whether the model one is working with is fitted fairly well or not. So there is, at least in concept, a circularity in the way GLIM's are fitted routinely. This section studies the possibility of using the Anderson-Darling statistic and the Cramér-von Mises statistic to test for goodness-of-fit when fitting GLIM's. The Anscombe and deviance residuals are reviewed and some distribution theory of the empirical processes defined using the Anscombe or deviance residuals is studied. The major goal is to build up a link between the EDF tests for normality of a random sample with unknown mean and/or unknown variance and the problem of testing for goodness-of-fit when fitting GLIM's so that the former can be used to carry out the latter approximately.

7.1.1 Introduction

A standard linear regression model (SLIM) can be viewed as consisting of three components:

1. The random component: observations $Y_1, \ldots, Y_n$ are independent normal random variables with $E(Y_i) = \mu_i$, $V(Y_i) = \sigma^2$;

2. The systematic component: covariates $x_1^t, \ldots, x_n^t$ (row vectors of known constants) produce linear predictors $\eta_i = x_i^t\beta$, where $\beta \in R^p$ is unknown;

3. The identity link between the systematic component and the random component: $\eta_i = \mu_i$.

Generalized linear models (GLIM) are models which generalize component 1 and component 3 explicitly. Formally, a generalized linear model assumes that

(a) The observations $Y_1, \ldots, Y_n$ are independent with means μ_i $(i = 1, \ldots, n)$ and Y_i is distributed according to density

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\}, \tag{7.1.1}$$

where θ_i and ϕ are real parameters with $\phi > 0$;

(b) For each i $(i = 1, \ldots, n)$, the known covariate $x_i^t = (x_{i1}, \ldots, x_{ip})$ provides a linear predictor $\eta_i = x_i^t\beta$ of $E(Y_i) = \mu_i$, where $\beta = (\beta_1, \ldots, \beta_p)^t \in R^p$ is unknown;

(c) The η_i's are related to the μ_i's $(i = 1, \ldots, n)$ through a monotone differentiable function $h(.)$, called the link function, by $\eta_i = h(\mu_i)$.

Clearly, if $Y_1, \ldots, Y_n$ are independent $N(\mu_i, \sigma^2)$ and $h(x) = x$, then such a GLIM is a SLIM, but a general GLIM can assume error distributions other than the normal distribution and can assume link functions other than the identity link. Notice that a general GLIM assumes that the variance of Y_i depends on x_i^t (a SLIM does not assume this dependence), but this dependence is only through the mean μ_i of Y_i. If a link function $h(.)$ makes $\theta_i = \eta_i$

true for $i = 1, \ldots, n$, then such an $h(.)$ is called the canonical link function. In this thesis, the inverse of $h(.)$ is denoted by $g(.)$.

In the exponential error structure given by (7.1.1), the mean and the variance of Y_i can be easily obtained as

$$\mu_i = E(Y_i) = b'(\theta_i), \tag{7.1.2}$$

$$\sigma_i^2 = V(Y_i) = b''(\theta_i)\phi. \tag{7.1.3}$$

The parameter ϕ is called the dispersion parameter.

7.1.2 Definitions of residuals for GLIM

For standard linear regression models, standardized residuals are defined naturally as $e_i = (y_i - x_i^t\hat{\beta})/\hat{\sigma}$, that is, $y_i = \hat{\mu}_i + \hat{\sigma}e_i$, which mimics $y_i = \mu_i + \sigma\varepsilon_i$ closely. When the same idea is applied to a GLIM, one gets the so-called Pearson residual

$$r_i^P = \frac{Y_i - \hat{\mu}_i}{\hat{SD}(Y_i)}, \tag{7.1.4}$$

where $\hat{\mu}_i$ and $\hat{SD}(Y_i)$ are estimates of the mean and standard deviation of Y_i, respectively.

In a SLIM, the exact distribution of e_i's is known and is related to the t-distribution. This is not true, however, for a general GLIM, except it has been known that the distribution of r_i^P's can be highly skewed and/or highly non-normal. Anscombe (1961) suggested using transformed Y_i's to define residuals which are more like normal random variables in distribution. Suppose $t(.)$ is a transformation, chosen in view of the distribution of Y_i, such that the distribution of $t(Y_i)$ is as normal as possible. Then the so-called Anscombe residual is defined as

$$r_i^A = \frac{t(Y_i) - \hat{E}(t(Y_i))}{\hat{SD}(t(Y_i))}, \tag{7.1.5}$$

where $\hat{E}(t(Y_i))$ and $\hat{SD}(t(Y_i))$ are the estimated mean and standard deviation of $t(Y_i)$, respectively.

There is yet another type of residuals; these residuals are based not on consideration of symmetry or normality, but are on the likelihood principle. When there are two models under consideration, let $\hat{\mu}_{i1}$ and $\hat{\mu}_{i2}$ be the corresponding maximum likelihood estimators of μ_i under these two models, then the so-called deviance residual is defined as

$$r_i^D = \text{sgn}(\hat{\mu}_{i1} - \hat{\mu}_{i2}) \left[2\{l(\hat{\mu}_{i1}, y_i) - l(\hat{\mu}_{i2}, y_i)\}\right]^{1/2}, \tag{7.1.6}$$

where $l(\mu_i, y_i) = \log f_{Y_i}(y_i, \mu_i, \phi)$ is the log-likelihood function in terms of μ_i and ϕ (ϕ is fixed constant), instead of in terms of θ_i and ϕ.

Following McCullagh and Nelder (1989), five distributions will be employed as error distributions. The definitions of these five distributions, together with the corresponding Anscombe and deviance residuals are given below. Some notation changes are made to facilitate the statement of some general results to be discussed later. The rule used here is that whenever μ is used, it refers to the mean of a certain random variable, and λ is used to denote the parameter that is expected to be large.

1. Binomial $B(\lambda, \mu)/\lambda$, where λ is the total number of trials, and μ is the success probability. Note that the proportion of successes is taken as the basic observation so that the mean is μ;

2. Poisson $P(\lambda)$, where λ is the mean parameter;

3. Gamma $G(\nu, \lambda)$, where ν is a scale parameter, λ is a shape parameter. The associated density takes the following form

$$f(y; \nu, \lambda) = \{\nu^\lambda/\Gamma(\lambda)\}^{-1} y^{\lambda-1} \exp\{-\nu y\},\ y > 0,\ \nu > 0,\ \lambda > 0$$

so that the mean is λ/ν and the variance is λ/ν^2;

4. Inverse Gaussian $IG(\mu, \lambda)$, where the density is

$$g(y; \mu, \lambda) = (2\pi)^{-1/2}(\lambda/y^3)^{1/2} \exp\{-\lambda(y-\mu)^2/(2y\mu^2)\},\ y > 0,\ \mu > 0,\ \lambda > 0.$$

The mean is μ, the variance is μ^3/λ;

5. Normal $N(\mu, \sigma^2)$.

The Anscombe residuals are obtained by approximating the mean to the second order and approximating the variance to the first order in the Taylor expansions of $\hat{E}(t(Y_i))$ and $\hat{SD}(t(Y_i))$, respectively. In each particular case, the transformation $t(.)$ is chosen in such a way that the third central moment of the transformed random variable $t(Y_i)$ is approximately zero. For the exponential family given by (7.1.1), $t(.)$ is determined by the integral

$$t(y) = \int^y \{b''(\mu)\}^{-1/3} d\mu, \tag{7.1.7}$$

where μ is the mean parameter. See Barndorff-Nielsen (1978, p179). Thus, define $t(u) = \int_0^u v^{-1/3}(1-v)^{-1/3}dv$ for the binomial distribution and let $g(.)$ denote the inverse of the link function $h(.)$, the Anscombe residuals are then given by

Binomial:

$$r_i^{AB} = \frac{t(y_i/\lambda) - t(\hat{\mu}_i) - (2\hat{\mu}_i - 1)/[6\lambda\{\hat{\mu}_i(1-\hat{\mu}_i)\}^{1/3}]}{\{\hat{\mu}_i(1-\hat{\mu}_i)\}^{1/6}/\sqrt{\lambda}},$$
$$\hat{\mu}_i = g(x_i^t\hat{\beta}),\ \lambda = 1/\phi, \tag{7.1.8}$$

Poisson:

$$r_i^{AP} = \frac{y_i^{2/3} - \hat{\lambda}_i^{2/3} + \hat{\lambda}_i^{-1/3}/9}{2\hat{\lambda}_i^{1/6}/3},\ \hat{\lambda}_i = g(x_i^t\hat{\beta}),\ \phi = 1, \tag{7.1.9}$$

Gamma:

$$r_i^{AG} = \frac{(\hat{\nu}_i y_i)^{1/3} - \hat{\lambda}^{1/3} + \hat{\lambda}_i^{-2/3}/9}{\hat{\lambda}^{-1/6}/3},\ \hat{\nu}_i = \{\hat{\phi}g(x_i^t\hat{\beta})\}^{-1},\ \hat{\lambda} = 1/\hat{\phi}, \tag{7.1.10}$$

IGaussian:

$$r_i^{AI} = \frac{\log(y_i) - \log(\hat{\mu}_i) + \hat{\mu}_i/(2\hat{\lambda})}{\sqrt{\hat{\mu}_i/\hat{\lambda}}},\ \hat{\mu}_i = g(x_i^t\hat{\beta}),\ \hat{\lambda} = 1/\hat{\phi}, \tag{7.1.11}$$

Normal:

$$r_i^{AN} = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}},\ \hat{\mu}_i = g(x_i^t\hat{\beta}),\ \hat{\sigma}^2 = \hat{\phi}. \tag{7.1.12}$$

Note that the dispersion parameter ϕ is assumed the same for all i ($i = 1, \ldots, n$) in (7.1.8) to (7.1.12), but this restriction can be relaxed. Note also that for Binomial and

Poisson distributions, there is essentially only one quantity to be estimated in order to construct Anscombe residuals, while for Gamma and Inverse Gaussian distributions, there are essentially two quantities to be estimated in order to construct Anscombe residuals. This difference will be noticed and used further in section 7.1.3.

The deviance residuals used in McCullagh and Nelder (1989) are defined as

Binomial:

$$r_i^{DB} = \text{sgn}(y_i - \hat{\mu}_i)\sqrt{2}|y_i \log(y_i/\hat{\mu}_i) + (\lambda - y_i)\log\{(\lambda - y_i)/(\lambda - \hat{\mu}_i)\}|^{1/2},$$

$$\hat{\mu}_i = \exp(\hat{\theta}_i)/\{1 + \exp(\hat{\theta}_i)\}, \tag{7.1.13}$$

Poisson:

$$r_i^{DP} = \text{sgn}(y_i - \hat{\mu}_i)\sqrt{2}|y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)|^{1/2},\ \hat{\mu}_i = \exp(\hat{\theta}_i), \tag{7.1.14}$$

Gamma:

$$r_i^{DG} = \text{sgn}(y_i - \hat{\mu}_i)\sqrt{2}|-\log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i|^{1/2},\ \hat{\mu}_i = -1/\hat{\theta}_i, \tag{7.1.15}$$

IGaussian:

$$r_i^{DI} = \text{sgn}(y_i - \hat{\mu}_i)|(y_i - \hat{\mu}_i)^2/(\hat{\mu}_i^2 y_i)|^{1/2},\ \hat{\mu}_i = (-2\hat{\theta}_i)^{-1/2} \tag{7.1.16}$$

Normal:

$$r_i^{DN} = \text{sgn}(y_i - \hat{\mu}_i)|y_i - \hat{\mu}_i|^{1/2},\ \hat{\mu}_i = x_i^t\hat{\beta}. \tag{7.1.17}$$

Notice that the deviance residuals defined above do not contain the parameter ϕ or the parameter λ explicitly, and for Gamma and Inverse Gaussian distributions, this implies that the deviance residuals are not properly scaled yet. Scaled deviance residuals will be introduced shortly. The relationship between each $\hat{\mu}_i$ and $\hat{\beta}$ can be obtained through the expression $\hat{\theta}_i = (b')^{-1} \circ h(x_i^t\hat{\beta})$.

It was found in Pierce and Schafer (1986), McCullagh and Nelder (1989) that Anscombe residuals and deviance residuals are numerically very similar, despite their seemingly different functional forms. Moreover, after deviance residuals are further adjusted to give the

so-called adjusted deviance residuals as given below,

Binomial:

$$r_{i,adjusted}^{DB} = r_i^{DB} + \frac{1 - 2\hat{\mu}_i}{6\{\lambda\hat{\mu}_i(1-\hat{\mu}_i)\}^{1/2}}, \tag{7.1.18}$$

Poisson:

$$r_{i,adjusted}^{DP} = r_i^{DP} + \frac{1}{6\sqrt{\hat{\mu}_i}}, \tag{7.1.19}$$

Gamma:

$$r_{i,adjusted}^{DG} = \sqrt{\lambda}\, r_i^{DG} + \frac{1}{3\sqrt{\lambda}}, \tag{7.1.20}$$

IGaussian:

$$r_{i,adjusted}^{DI} = \sqrt{\lambda}\, r_i^{DI} + \frac{1}{2}\left\{\frac{\hat{\mu}_i}{\lambda}\right\}^{1/2}, \tag{7.1.21}$$

Normal:

$$r_{i,adjusted}^{DN} = r_i^{DN}/\sqrt{\sigma}, \tag{7.1.22}$$

both Anscombe residuals and adjusted deviance residuals are surprisingly good in terms of approximate normality when binomial, poisson and gamma distributions are taken as error distributions. The same conclusion can be drawn for the inverse Gaussian distribution. Note however that these findings are discussed when the various residuals are expressed in terms of true parameter(s), which will be called the theoretical Anscombe or (adjusted) deviance residuals in this thesis.

At this point, it is perhaps appropriate to discuss briefly the background of introducing the various residuals. Facing the lack of exact theory once non-normal error distributions are allowed for, and driven by the beauty and simplicity of the theory of standard linear regression models, statisticians have been working very hard to develop techniques similar to those used for SLIM. In this big picture, residual analysis has been given a great deal of attention. As implied by its name, residual analysis naturally requires a closer look at individual residuals, and if one wants to do for GLIM the same types of diagnostics that one can do for SLIM, one must find ways of creating residuals that behave like normal random

variables. The Anscombe residuals are directly aimed at this purpose, while the adjusted deviance residuals are found to be fairly suitable, too, as discussed in above paragraph.

However, it is important to note that "behave like normal random variables" is a rather vague concept. For the Anscombe residuals defined in (7.1.8) to (7.1.11), it can be shown that for any real number r, if the residuals are expressed in terms of true parameter(s) (that is, the theoretical Anscombe residuals), then

$$P(r_i^{A\cdot} \leq r) \longrightarrow \Phi(r), \text{ as } \lambda \to \infty, \tag{7.1.23}$$

where $\Phi(.)$ is the cumulative distribution function of a standard normal random variable and "$\cdot$" indicates Binomial, Poisson, Gamma, or Inverse Gaussian error distribution. A similar conclusion holds for the theoretic (adjusted) deviance residuals. Note that the error occurred in (7.1.23) for finite λ is usually $O(\lambda^{-1/2})$. For adjusted theoretical deviance residuals, the error can be reduced to $O(\lambda^{-1})$. See McCullagh (1984), McCullagh and Nelder (1989).

7.1.3 Residual empirical processes in GLIM

The observation that the theoretical Anscombe or adjusted deviance residuals are close to normal random variables encourages the following half-rigorous and half-heuristic development. Let $r_i = r_i(y_i; \theta_i, \lambda)$ be the i^{th} theoretical Anscombe or adjusted deviance residual, let $\hat{r}_i = \hat{r}_i(y_i; \hat{\theta}_i, \hat{\lambda}_i)$ be the i^{th} (estimated) Anscombe or adjusted deviance residual. For $t \in [0, 1]$, define residual empirical processes $Y_{n,\lambda}(t)$ and $Y_{n,\hat{\lambda}}(t)$ by

$$Y_{n,\lambda}(t) = n^{-1/2} \sum_{i=1}^{n} \{I[\Phi(r_i) \leq t] - t\}, \tag{7.1.24}$$

$$Y_{n,\hat{\lambda}}(t) = n^{-1/2} \sum_{i=1}^{n} \{I[\Phi(\hat{r}_i) \leq t] - t\}. \tag{7.1.25}$$

Since $r_1, \ldots, r_n$ are independent and as $\lambda \to \infty$, $r_i \to_d N(0, 1)$, it is readily verified that for any fixed n and for any fixed s, $t \in [0, 1]$,

$$E\{Y_{n,\lambda}(t)\} \to 0, \qquad \text{cov}\{Y_{n,\lambda}(s), Y_{n,\lambda}(t)\} \to \min(s, t) - st,$$

as $\lambda \to \infty$. In fact, in light of McCullagh (1984), and McCullagh and Nelder (1989), the empirical process $Y_{n,\lambda}(t)$ based on the theoretical Anscombe residuals converges weakly to the Brownian bridge if $\sqrt{n/\lambda} \to 0$; the empirical process $Y_{n,\lambda}(t)$ based on the theoretical adjusted deviance residuals converges weakly to the Brownian bridge if $\sqrt{n}/\lambda \to 0$.

In applications, however, it is more realistic to expect large n than to expect large λ. Nevertheless, the above discussion encourages the following heuristic arguments:

Because the theoretical Anscombe or adjusted deviance residuals become standard normal random variables only when $\lambda \to \infty$, for finite λ, it can only be expected that the theoretical Anscombe or adjusted deviance residuals have distributions which are fairly symmetric and nearly properly scaled. In other words, the process of constructing the theoretical Anscombe or adjusted deviance residuals introduces, for each λ value, a mean and a variance to the residuals constructed, and this mean will go to zero and this variance will go to one as λ goes to infinity. Viewed from this angle and recall the remarks made after introducing the various residuals, it is observed that

1. Estimating μ_i or λ_i in the theoretical Anscombe residuals with Binomial and Poisson error structures is like estimating the mean for a random normal sample with known variance;

2. Estimating ν_i and λ_i, or μ_i and λ_i in the theoretical Anscombe residuals with Gamma and Inverse Gaussian error structures is like estimating the mean and the variance for a random normal sample;

3. Estimating parameters in adjusted deviance residuals carries a parallel pattern to the above, that is, for Binomial and Poisson error structures, it is like estimating the mean for a random normal sample with known variance, and for Gamma and Inverse Gaussian error structures, it is like estimating the mean and the variance for a random normal sample.

As the effect of estimating the mean and/or variance in the normal sample situation

is well-understood, it is hoped that the estimated Anscombe residuals and the adjusted deviance residuals can be approximately treated in the way the residuals from a SLIM are treated. In particular, the residual empirical process $Y_{n,\hat{\lambda}}$ of (7.1.25) is expected to have a weak limit that can be approximated by the Gaussian process $\{Y(t) : t \in [0,1]\}$ with zero mean and covariance function

$$\min(s,t) - st - J_1(s)J_1(t), \tag{7.1.26}$$

or with zero mean and covariance function

$$\min(s,t) - st - J_1(s)J_1(t) - \frac{1}{2}J_2(s)J_2(t), \tag{7.1.27}$$

where $s,t \in [0,1]$, $J_1(t) = \phi(\Phi^{-1}(t))$, $J_2(t) = \phi(\Phi^{-1}(t))\Phi^{-1}(t)$, that is, EDF tests for normality of a random sample with unknown mean only, or with unknown mean and unknown variance can be used to perform (approximate) goodness-of-fit tests when fitting GLIM's.

Simulation results (not shown here) support the above intuitive ideas.

7.2 Transform-Both-Sides (TBS) Models

Very often in regression analysis, a particular functional form connecting known covariates and unknown parameters is either suggested by previous work or demanded by theoretical considerations so that the deterministic part of the responses has a known form. However, the underlying error structure is often less well understood. In this case, the transform-both-sides (TBS) models are appropriate. This section proposes to generalize the usual TBS models studied in details by Carroll and Ruppert (1984, 1988), among others, into generalized transform-both-sides (GTBS) models, and to study the possibility of using EDF tests to assess goodness-of-fit when fitting TBS or GTBS models. Parameter estimation for the generalized TBS models is discussed and EDF tests based on the Cramér-von Mises statistic and the Anderson-Darling statistic are suggested.

7.2.1 Introduction

Let $Y_1, \ldots, Y_n$ be independent random variables generating responses in an experiment. Let $x_i = (x_{i1}, \ldots, x_{ip})$ $(i = 1, \ldots, n)$ be known covariates associated with the Y_i's. When the standard normal theory linear regression models do not seem to be appropriate to summarize the relationship between the Y_i's and the x_i's, Box and Cox (1964) proposed to transform the Y_i's through a monotone function indexed by a parameter λ, say $h(\cdot, \lambda)$, so that the transformed responses $h(Y_i, \lambda)$'s can be fitted by a standard linear model, that is,

$$h(Y_i, \lambda) = x_i\beta + \sigma\varepsilon_i, \tag{7.2.1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^t \in R^p$ and $\sigma > 0$ are unknown and the ε_i's are independent standard normal random variables. There are three goals aimed at by the Box-Cox transformation approach: (1) a simple model which is linear in β, (2) the errors in the model have constant variances, and (3) the errors in the model are normally distributed. See Box and Cox (1964).

Note that in model (7.2.1) the covariates are not transformed and they enter the model through creating linear combinations with the unknown parameter β. This is perhaps the simplest way that covariates enter a statistical model. In some applications, however, it is necessary and possible to specify the way that covariates and unknown parameters are connected. This is the case, for example, when previous work has suggested a particular model, or when theoretical considerations demand a specific combination. Denote this (known) specific combination (or functional form) by $f(x_i, \beta)$, Carroll and Ruppert (1984, 1988), among others, introduced the so-called transform-both-sides (TBS) models as below:

$$h(Y_i, \lambda) = h[f(x_i, \beta), \lambda] + \sigma\varepsilon_i. \tag{7.2.2}$$

There are two goals for the transform-both-sides (TBS) approach: (1) the errors in the model are homoscedastic, and (2) the errors in the model are normal. TBS models are appropriate when the functional form $f(x_i, \beta)$ is well understood but the underlying error structure is not quite so. Carroll and Ruppert (1988) provide a detailed account of the theory and

application of TBS models, except that the problem of checking for goodness-of-fit when fitting TBS models is not discussed explicitly.

This section proposes to study the possibility of using tests based on empirical distribution functions (EDF) to assess goodness-of-fit when TBS models are fitted.

7.2.2 Generalized TBS models

Suppose that random variables $Y_1, \ldots, Y_n$ are independent, and that Y_i generates a known functional form $f(x_i, \beta)$ plus error, where $x_i = (x_{i1}, \ldots, x_{ip})$ is known and $\beta = (\beta_1, \ldots, \beta_p)^t \in R^p$ is unknown. Let $h(\cdot, \lambda)$ be a monotone transformation indexed by an m-dimensional parameter λ, let $g(\cdot, \psi)$ be a monotone transformation indexed by a k-dimensional parameter ψ. The functional forms of h and g are known. Then, a generalized transform-both-sides (GTBS) model is defined by supposing that there are $\beta \in R^p$, $\alpha \in R$, $\lambda \in \Lambda$, $\psi \in \Psi$, where $\Lambda \subset R^m$ and $\Psi \subset R^k$, and $\sigma > 0$ such that

$$h(Y_i, \lambda) = \alpha + g[f(x_i, \beta), \psi] + \sigma\varepsilon_i. \tag{7.2.3}$$

Clearly, if $\alpha = 0$ and $h = g$, then a GTBS model becomes a TBS model. Note that if denote $h(Y_i, \lambda)$ by Z_i and denote $g[f(x_i, \beta), \psi]$ by μ_i, then model (7.2.3) takes the form $Z_i = \alpha + \mu_i + \sigma\varepsilon_i$. This form looks like a linear regression model with intercept and inspires the EDF tests to be given in next section. The reader should be warned that in general care must be taken in selection of g if ψ and β are both to be identifiable.

Although much of the discussion in this section can be carried out using general h and g, it is useful and instructive to work with some specific transformations. In the following discussion, $h(y, \lambda)$ is taken as the modified power transformation given by

$$h(y, \lambda) = y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, \\ \log y & \text{if } \lambda = 0. \end{cases} \tag{7.2.4}$$

Since $h(y, \lambda) = y^{(\lambda)}$ is the modified power transformation now, all the observed responses y_i's must be positive. With this restriction, model (7.2.3) can not be correct exactly (see

section 5.1). The correct model in this case is given by

$$Y_i^{(\lambda)} \sim \frac{1}{\sigma}\phi\left(\frac{y_i^{(\lambda)} - \alpha - \mu_i}{\sigma}\right)\frac{1}{\Phi(\delta_i^*)}, \tag{7.2.5}$$

in terms of the transformed variables, where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and cumulative distribution function of a standard normal random variable, $\mu_i = g[f(x_i, \beta), \psi]$, and δ_i^* equals $(\alpha + \mu_i + 1/\lambda)/\sigma$, $+\infty$, or $-(\alpha + \mu_i + 1/\lambda)/\sigma$, according to $\lambda > 0$, $= 0$, or < 0, respectively. In terms of the original variables, the correct model is

$$Y_i \sim \frac{1}{\sigma}\phi\left(\frac{y_i^{(\lambda)} - \alpha - \mu_i}{\sigma}\right) y_i^{\lambda-1}\frac{1}{\Phi(\delta_i^*)}. \tag{7.2.6}$$

See Chapter 5 for more details.

From model (7.2.6), the log-likelihood function in terms of the original variables y_i is, apart from an additive constant,

$$\begin{aligned} L &= L(\alpha, \beta, \lambda, \psi, \sigma^2) \\ &= -(n/2)\log\sigma^2 - (2\sigma^2)^{-1}\sum_{i=1}^{n}\{y_i^{(\lambda)} - \alpha - g[f(x_i, \beta), \psi]\}^2 \\ &\quad +(\lambda - 1)\sum_{i=1}^{n}\log y_i - \sum_{i=1}^{n}\log\Phi(\delta_i^*). \end{aligned} \tag{7.2.7}$$

Note that the term $\sum_{i=1}^{n}\log\Phi(\delta_i^*)$ is needed to make the statement in (7.2.7) correct exactly. However, this term is often very small compared to the rest of the log-likelihood function. More specifically, if (1) λ is small, or (2) $\alpha + \mu_i$ is large, or (3) σ is small, then δ_i^* will be large, so $\log\Phi(\delta_i^*)$ will be close to zero. Therefore, in the following discussion, this term is omitted to simplify the presentation of various expressions. It is noted however that if by any means the term $\sum_{i=1}^{n}\log\Phi(\delta_i^*)$ is found to be not negligible, then (7.2.7) must be treated as a whole.

There are at least four methods to find parameter estimates based on (7.2.7) (ignoring $\sum_{i=1}^{n}\log\Phi(\delta_i^*)$):

Method 1. Simultaneous estimation, using a Newton or quasi-Newton program.

Method 2. Box-Cox method. For a fixed λ, maximizing L with respect to α, β, ψ and σ^2 is equivalent to minimizing

$$\sum_{i=1}^{n} \left\{ y_i^{(\lambda)} - \alpha - g[f(x_i, \beta), \psi] \right\}^2$$

to obtain $\tilde{\alpha}(\lambda)$, $\tilde{\beta}(\lambda)$ and $\tilde{\psi}(\lambda)$, then setting

$$\tilde{\sigma}^2(\lambda) = n^{-1} \sum_{i=1}^{n} \left\{ y_i^{(\lambda)} - \tilde{\alpha}(\lambda) - g[f(x_i, \tilde{\beta}(\lambda), \tilde{\psi}(\lambda)] \right\}^2.$$

This can be done with a nonlinear least squares program. An estimate $\tilde{\lambda}$ for λ is obtained by maximizing

$$L_{max}(\lambda) = -(n/2) \log \tilde{\sigma}^2(\lambda) + (\lambda - 1) \sum_{i=1}^{n} \log y_i$$

over λ either graphically or through a grid search.

Method 3. Pseudo-model method. For fixed α, β, ψ and λ, L is maximized over σ^2 by

$$\hat{\sigma}^2(\alpha, \beta, \psi, \lambda) = n^{-1} \sum_{i=1}^{n} \{ y_i^{(\lambda)} - \alpha - g[f(x_i, \beta), \psi] \}^2.$$

Then estimates $\hat{\alpha}$, $\hat{\beta}$, $\hat{\psi}$ and $\hat{\lambda}$ are obtained by maximizing

$$L_{max}(\alpha, \beta, \psi, \lambda) = -(n/2) \log[\hat{\sigma}^2(\alpha, \beta, \psi, \lambda) / \dot{y}^{2(\lambda - 1)}],$$

where $\dot{y} = (\prod_{i=1}^{n} y_i)^{1/n}$ is the geometric mean of $y_1, \ldots, y_n$. This is equivalent to minimizing

$$\sum_{i=1}^{n} \left[\left\{ y_i^{(\lambda)} - \alpha - g[f(x_i, \beta), \psi] \right\} / \dot{y}^{\lambda} \right]^2.$$

Let $d_i = 0$, $e_i = \{ y_i^{(\lambda)} - \alpha - g[f(x_i, \beta), \psi] \} / \dot{y}^{\lambda}$ $(i = 1, \ldots, n)$, the pseudo-model method consists of fitting d_i to e_i, where the "dependent" variable is d_i, the "independent" variables are x_i and y_i, with parameters α, β, ψ and λ. This can be done using a nonlinear least squares program. After $\hat{\alpha}$, $\hat{\beta}$, $\hat{\psi}$ and $\hat{\lambda}$ are obtained, an estimate $\hat{\sigma}^2$ for σ^2 is given by

$$\hat{\sigma}^2 = \hat{\sigma}^2(\hat{\alpha}, \hat{\beta}, \hat{\psi}, \hat{\lambda}).$$

Method 4. M-estimation method. See Carroll and Ruppert (1988).

7.2.3 EDF tests of goodness-of-fit for GTBS models

For model (7.2.3), define residuals r_i and estimated residuals $\tilde{r}_i$ and $\hat{r}_i$ by

$$r_i = (h(y_i,\lambda) - \alpha - \mu_i)/\sigma = \{h(y_i,\lambda) - \alpha - g[f(x_i,\beta),\psi]\}/\sigma, \quad (7.2.8)$$

$$\tilde{r}_i = (h(y_i,\lambda) - \tilde{\alpha} - \tilde{\mu}_i)/\tilde{\sigma} = \{h(y_i,\lambda) - \tilde{\alpha} - g[f(x_i,\tilde{\beta}),\tilde{\psi}]\}/\tilde{\sigma}, \quad (7.2.9)$$

$$\hat{r}_i = (h(y_i,\hat{\lambda}) - \hat{\alpha} - \hat{\mu}_i)/\hat{\sigma} = \{h(y_i,\hat{\lambda}) - \hat{\alpha} - g[f(x_i,\hat{\beta}),\hat{\psi}]\}/\hat{\sigma}, \quad (7.2.10)$$

where $\mu_i = g[f(x_i,\beta),\psi]$, $\tilde{\mu}_i = g[f(x_i,\tilde{\beta}),\tilde{\psi}]$ and $\hat{\mu}_i = g[f(x_i,\hat{\beta}),\hat{\psi}]$. Note that in (7.2.9) λ is supposed known. Both " $\hat{}$ " and " $\tilde{}$ " indicate estimated quantities obtained, say, by the method of maximum likelihood. The corresponding empirical processes are defined for $t \in [0,1]$ by

$$Y_n(t) = n^{-1/2}\sum_{i=1}^{n}\{I[\Phi(r_i \le t] - t\}, \quad (7.2.11)$$

$$\tilde{Y}_{n(t)} = n^{-1/2}\sum_{i=1}^{n}\{I[\Phi(\tilde{r}_i) \le t] - t\}, \quad (7.2.12)$$

$$\hat{Y}_n(t) = n^{-1/2}\sum_{i=1}^{n}\{I[\Phi(\hat{r}_i) \le t] - t\}. \quad (7.2.13)$$

If model (7.2.3) holds exactly, and if all parameters in model (7.2.3) are known, then it is well known that as $n \to \infty$, Y_n converges weakly to the Brownian bridge with zero mean and covariance function

$$\rho_0(s,t) = \min(s,t) - st. \quad (7.2.14)$$

Suppose now that λ is known and the rest of the parameters in model (7.2.3), denoted by $\theta = (\alpha,\beta,\psi,\sigma^2)^t$, need to be estimated. Let $Z_i = h(Y_i,\lambda)$, then the cumulative distribution function of Z_i is given by

$$F_i(z,\theta) = \Phi\left(\frac{z - \alpha - g[f(x_i,\beta),\psi]}{\sigma}\right).$$

This is an independent but not identically distributed case to which the results of Loynes (1980) can be applied to get

Conjecture 7.2.1 *Under some smoothness conditions on f and g, the empirical process $\tilde{Y}_n$ of (7.2.12) converges weakly to a Gaussian process with zero mean and covariance function*

$$\tilde{\rho}(s,t) = \rho_0(s,t) - c\rho_1(s,t) - \frac{1}{2}\rho_2(s,t), \tag{7.2.15}$$

where $\rho_0(s,t) = \min(s,t) - st$ is given by (7.2.14), c is a constant, and

$$\rho_1(s,t) = J_1(s)J_1(t), \tag{7.2.16}$$

$$\rho_2(s,t) = J_2(s)J_2(t), \tag{7.2.17}$$

where $J_1(t) = \phi(\Phi^{-1}(t))$, $J_2(t) = \phi(\Phi^{-1}(t))\Phi^{-1}(t)$, $t \in [0,1]$.

The additive structure of (7.2.15) is due to the normality assumption and the asymptotic independence between $(\tilde{\alpha}, \tilde{\beta}, \tilde{\psi})^t$ and $\tilde{\sigma}^2$. The constant c will in general depend on unknown parameters $\theta = (\alpha, \beta, \psi, \sigma^2)^t$, on f and g, and on the x_i's. In the cases where $g(u,\psi) = 1$, or $g(u,\psi) = u$ and $f(x_i,\beta) = x_i\beta$, c can be found to be equal to 1. In these cases, one recovers the results that hold for normal theory linear regression models. However, since the collective effect of estimating α, β and ψ is to estimate $\alpha + \tilde{\mu}_i$, it is expected that the value of c will be close to 1, that is, it is hoped that for applications, $\rho_1(s,t)$ can be used as an approximation to $c\rho_1(s,t)$ for large n and smooth f and g.

Next, suppose that all parameters in model (7.2.3) need to be estimated. As in section 7.2.2, let $h(y,\lambda)$ be the modified power transformation defined by (7.2.4). Then model (7.2.3) takes the form given by (7.2.6) in terms of the directly observable variables Y_i's. Note that in principle it is not possible to assume exact normality in model (7.2.6) for the errors unless $\lambda = 0$. But very often in applications, $\alpha + \mu_i$ is large and/or σ is small. For these interesting cases, assuming exact normality to modify model (7.2.6) will cause little trouble for a range of λ values, say, $-1 \leq \lambda \leq 1$, because $\Phi(\delta_i^*)$ is close to 1. With this observation, there is

Conjecture 7.2.2 *If $h(y,\lambda) = y^{(\lambda)}$ as defined in (7.2.4) and under some smoothness conditions on f and g, the empirical process $\hat{Y}_n$ of (7.2.13) converges weakly to a Gaussian*

process which can be approximated by the Gaussian process $\{Y(t) : t \in [0,1]\}$ *with zero mean and covariance function*

$$\rho(s,t) = \rho_0(s,t) - \frac{1}{1!}\rho_1(s,t) - \frac{1}{2!}\rho_2(s,t) - \frac{1}{3!}\rho_3(s,t), \tag{7.2.18}$$

where $\rho_0(s,t)$ *is given by (7.2.14),* $\rho_1(s,t)$ *is given by (7.2.16),* $\rho_2(s,t)$ *is given by (7.2.17) and*

$$\rho_3(s,t) = J_3^*(s)J_3^*(t), \tag{7.2.19}$$

where $J_3^*(t) = \phi(\Phi^{-1}(t))[(\Phi^{-1}(t))^2 - 1]$.

The reason $\{Y(t) : t \in [0,1]\}$ only approximates the weak limit of $\hat{Y}_n(t)$ is that $\{Y(t) : t \in [0,1]\}$ is the weak limit of $\hat{Y}_n(t)$ for $\lambda = 0$. See section 5.6 for more details.

Based on Conjecture 7.2.2, EDF tests can be performed to check goodness-of-fit of fitting TBS or GTBS models when responses y_i's are transformed using the modified power transformation (7.2.4).

7.3 Comments

Asymptotic results obtained when sample size goes to infinity have been used as guides when the distributions of sensible statistics are not readily available for finite sample size. However, this approach does not meet all the needs raised in applications. In the context of GLIM, there are more choices than in the framework of SLIM for a data analyst to choose an error distribution and a link, but (when judged according to the available GLIM packages) residual analysis and goodness-of-fit techniques available for SLIM's do not carry over to GLIM's completely.

Since a unified and working (for small to moderate samples) asymptotic theory for GLIM is not available, and since techniques for SLIM's are widely known, defining residuals that behave like residuals from fitting SLIM's seems to be a promising way to start the process of model checking when fitting GLIM's. This approach asks for another type of asymptotics

in which certain parameter of the error distribution is expected to be large. This idea is tried in this chapter.

Fitting nonlinear models is in general difficult. Any technique that can make the fitting and/or interpretation of nonlinear models easier should be worth trying. Since the theory of linear models with normal errors is best known, and since techniques for fitting such models are well developed, transforming nonlinear models in ways that will allow linear model techniques to be applied to nonlinear models seems to be a useful approach.

One possibility is to reparametrize a nonlinear model into a linear model. However, this is not always possible, especially for nonlinear models with many parameters. In this case, if the model expectation function is suggested by previous study or is based on theoretical work, the TBS approach would be suitable, because the known relationship is reserved when monotone transformations are used. In addition, analysis in the transformed scale does not seem to depend strongly on estimating parameters of the transformation. See Carroll and Ruppert (1984).

If the problem is to build an empirical model which summarizes the data well, the GTBS approach would be more flexible than the Box-Cox approach and the TBS approach.

Once leaving the theory of linear models with normal errors, things become hard to handle. This is certainly so for the EDF tests presented in this thesis. Nevertheless, intuitively speaking, nonlinear expectation functions do not present a very new problem for one to construct EDF tests, because while nonlinearity presents most problems to obtain good estimates and standard errors for individual parameters, nonlinearity does not affect residuals as much. This is the rationale behind Conjecture 7.2.1 and Conjecture 7.2.2 of section 7.2. In general, EDF tests based on the ideas of section 7.1 and section 7.2 should be used as guiding tools, such as suggesting the types of encountered problems like heavy or light tail(s) in error distribution and misspecified model expectation function.

The smoothness conditions needed to make Conjecture 7.2.1 and Conjecture 7.2.2 work are essentially the types of conditions that guarantee that the maximum likelihood estimates

be $\sqrt{n}$-consistent, and that the Fisher information matrix exist and be positive definite. See Seber and Wild (1989, Chapter 12).

Bibliography

[1] Abramowitz, M. and Stegun, I.A. (1970). *Handbook of Mathematical Functions*, U.S Department of Commerce, National Bureau of Standards, Applied Mathematics Series 55.

[2] Atkinson, A.C. (1985). *Plots, Transformations, and Regression*, Clarendon Press, Oxford.

[3] Anderson, T.W. and Darling, D.A. (1952). Asymptotic theory of certain 'goodness-of-fit' criteria based on stochastic processes. *The Annals of Mathematical Statistics*, **23**, 193–212.

[4] Anscombe, F.J. (1961). Examination of residuals. *Proc. Fourth Berkeley Symposium*, **1**, 1-36.

[5] Anscombe, F.J. (1973). Graphs in statistical analysis. *American Statistician*, **27**, 17–21.

[6] Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*, John Wiley & Sons, New York.

[7] Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley & Sons, New York.

[8] Bergström, H. (1982). *Weak Convergence of Measures*, Academic Press, New York.

[9] Bickel, P.J. and Doksum, K.A. (1981). An analysis of transformations revisited. *Journal of American Statistical Association*, **76**, 296–311.

[10] Billingsley, P. (1968). *Convergence of Probability Measures*, John Wiley & Sons, New York.

[11] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations (with discussions). *Journal of Royal Statistical Society*, **B, 26**, 211–252.

[12] Box, G.E.P. and Cox, D.R. (1982). An analysis of transformations revisited, rebutted. *Journal of American Statistical Association*, **77**, 209–210.

[13] Breiman, L. (1968). *Probability*, Addison-Wesley, Reading, Massachusetts.

[14] Carroll, R.J. and Ruppert, D. (1984). Power transformation when fitting theoretical models to data. *Journal of American Statistical Association*, **79**, 321–328.

[15] Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*, Chapman and Hall, London.

[16] Castillo, E. (1988) *Extreme Value Theory in Engineering*, Academic Press, Inc.

[17] Chatterjee, S. and Hadi, A.S. (1988). *Sensitivity Analysis In Linear Regression*, John Wiley & Sons, New York.

[18] Cheng, R. C. H. and N. A. K. Amin (1981) Maximum likelihood estimation of parameters in the inverse Gaussian distribution, with unknown origin. *Technometrics,* **5**, 257–263.

[19] Chhikara, R. S. and J. L. Folks (1989). *The Inverse Gaussian Distribution*, Marcel Dekker, Inc. New York and Basel.

[20] Cohen, A. C. and B. J. Whitten (1988). *Parameter Estimation in Reliability and Life Span Models*, Marcel Dekker, Inc. New York and Basel.

[21] Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman and Hall, London.

[22] Dagpunar, J. (1988). *Principles of Random Variate Generation*, Clarendon Press, Oxford.

[23] Darling, D.A. (1955). The Cramér-von Mises test in the parametric case. *The Annals of Mathematical Statistics*, **26**, 1–20.

[24] Davies, R.B. (1973). Numerical inversion of a characteristic function. *Biometrika*, **60**, 415–417.

[25] D'Agostino, R. and M. Stephens (1986). *Goodness-of-Fit Techniques*, Marcel Dekker, Inc. New York and Basel.

[26] Donsker, M.D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of Mathematical Statistics*, **23**, 277–281.

[27] Doob, J.L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of Mathematical Statistics*, **20**, 393–403.

[28] Dumonceaux, R. and C. E. Antle (1973). Discrimination between the lognormal and Weibull distributions. *Technometrics*, **15**, 923–926.

[29] Durbin, J. (1973a). Weak convergence of the sample distribution function when parameters are estimated. *The Annals of Statistics*, **1**, 279–290.

[30] Durbin, J. (1973b). *Distribution Theory for Tests Based on the Sample Distribution Function*, Regional Conference Series in Applied Mathematics, 9. Philadelphia: SIAM.

[31] Fröberg, C. (1985). *Numerical Mathematics*, The Benjamin/Cummings Publishing Company, Inc., Menlo Park, California.

[32] Gikhman, I.I. and Skorokhod, A.V. (1965). *Introduction to the Theory of Random Processes*, W.B. Saunders Company.

[33] Gumbel, E. L. (1964). Technische anwendung der statistischen theorie der extremwerte. *Schweizer Archiv,* **30**, 33–47.

[34] Hall, W.J. and Loynes, R.M. (1977). On the concept of contiguity. *The Annals of Probability*, **5**, 278–282.

[35] Hinkley, D.V. (1975). On power transformation to symmetry. *Biometrika*, **62**, 101–112.

[36] Hinkley, D.V. and Runger, G. (1984). The Analysis of transformed data (with discussions). *Journal of American Statistical Association*, **79**, 302–309.

[37] Hjort, N.L. (1990). Goodness-of-fit tests in models for life history data based on cumulative hazard rates. *The Annals of Statistics*, **18**, 1221–1258.

[38] Huber, P. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, **1**, 799–821.

[39] Huber, P. (1981). *Robust Statistics*, John Wiley & Sons, New York.

[40] Imhof, J.P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, **48**, 419–426.

[41] Kac, M. and Siegert, A.J.F. (1947). An explicit representation of a stationary Gaussian process. *The Annals of Mathematical Statistics*, **18**, 438–442.

[42] *IMSL, Problem-Solving Software Systems*, softcover edition, (1987), IMSL, Inc, Houston, Texas.

[43] Kendall, M. and Stuart, A. (1977). *The Advanced Theory of Statistics*, Vol. **1**, 4th edition, Charles Griffin & Company Limited, London.

[44] Khmaladze, E.V. (1981). Martingale approach in the theory of goodness-of-fit tests. *Theory of Probability and its Applications*, **26**, 240–257.

[45] Kac, M., Kieffer, J. and Wolfowitz, J. (1955). On tests of normality and other tests of goodness-of-fit based on distance method. *The Annals of Mathematical Statistics*, **26**,189–211.

[46] Kotz, S., Johnson, N.L. and Boyd, D.W. (1967). Series representations of distributions of quadratic forms in normal variables. I Central case. *The Annals of Mathematical Statistics*, **38**, 823–837.

[47] Koul, H.L. (1984). Tests of goodness-of-fit in linear regression. *Goodness-of-fit*, edited by K. Sarkadi, Hungarian.

[48] Koul, H.L. and Levental, S. (1989). Weak convergence of the residual empirical process in explosive autoregression. *The Annals of Statistics*, **17**, 1784–1794.

[49] Linnet, K. (1988). Testing normality of transformed data. *Applied Statistics*, **37**,180–186.

[50] Loève, M. (1977). *Probability Theory*, 4th edition, Springer-Verlag, New york.

[51] Loynes, R. M. (1980). The empirical distribution function of residuals from generalized regression. *The Annals of Statistics*, **8**, 285–298.

[52] Manoukian, E.B. (1986). *Modern Concepts and Theorems of Mathematical Statistics*, Springer-Verlag, New York.

[53] McCullagh, P. (1984). Local sufficiency. *Biometrika*, **71**, 233–244.

[54] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, second edition, Chapman and Hall, London.

[55] Meester, S.G. and Lockhart, R.A. (1988). Testing for normal errors in designs with many blocks. *Biometrika*, **75**, 569–575.

[56] Mukantseva, L.A. (1977). Testing normality in one-dimensional and multi-dimensional linear regression. *Theory of Probability and its Applications*, **22**, 591–602.

[57] Myers, R.H. (1986). *Classic and Modern Regression With Applications*, Duxbury Press, Boston.

[58] Neter, J. and Wasserman, W. (1974). *Applied Linear Regression Models*, Irwin, Homewood, Illinois.

[59] Neuhaus, G. (1976). Weak convergence under contiguous alternatives of the empirical processes when parameters are estimated: The D_k approach. *Lecture Notes in Mathematics*, **566**, Springer-Verlag, Heidelberg.

[60] Pierce, D.A. (1985). Testing normality in autoregressive models. *Biometrika*, **72**, 293–297.

[61] Pierce, D.A. and Kopecky, K.J. (1979). Testing goodness-of-fit for the distribution of errors in regression models. *Biometrika*, **66**, 1–6.

[62] Pierce, D.A. and Schafer, D.W. (1986). Residuals in generalized linear models. *Journal of American Statistical Association*, **81**, 977–986.

[63] Pitman, E. (1979). *Some Basic Theory for Statistical Inference*, Chapman and Hall, London.

[64] Pollard, D. (1984). *Convergence of Stochastic Processes*, Springer-Verlag, New York.

[65] Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics*, **5**, 375–383.

[66] Rao, B.L.S.P. (1987). *Asymptotic Theory of Statistical Inference*, John Wiley & Sons, New York.

[67] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, second edition, John Wiley & Sons, New York.

[68] Rao, J.S. and Sethuraman, J. (1975). Weak convergence of empirical distribution functions of random variables subject to perturbations and scale factors. *The Annals of Statistics*, **3**, 299–313.

[69] Ryan, T., Joiner, B. and Ryan, B. (1976). *Minitab Student Handbook*, Duxbury Press, North Scituate, Massachusetts.

[70] *SAS User's Guide: Statistics*, 1982 edition, SAS Institute, Inc.

[71] Seber, G.A.F. (1977). *Linear Regression Analysis*, John Wiley & Sons, New York.

[72] Seber, G.A.F. and Wild, C.J. (1989). *Nonlinear Regression*. Wiley, New York.

[73] Schilling, M.F. (1983). An infinite-dimensional approximation for nearest neighbor goodness of fit tests. *The Annals of Statistics*, **11**, 13–24.

[74] Shorack, G.R. (1984). Empirical and rank processes of observations and residuals. *Canadian Journal of Statistics*, **12**, 319–332.

[75] Shorack, G.R. and J.A. Wellner (1986). *Empirical Processes With Applications to Statistics*, John Wiley & Sons, New York.

[76] Steen, P. J. and D, J. Stickler (1976). A sewage pollution study of beaches from Cardiff to Ogmore. UWIST, Dept. of Applied Biology Report, January, Cardiff.

[77] Stephens, M.A. (1974). EDF statistics for goodness-of-fit and some comparisons. *Journal of American Statistical Association*, **69**, 730–737.

[78] Stephens, M.A. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. *The Annals of Statistics*, **4**, 367–369.

[79] Stephens, M.A. (1986). *Goodness-of-Fit Techniques*, Chapter 4, edited by A'Dgostino, R.B. and Stephens, M.A. Marcel Dekker, Inc. New York and Basel.

[80] Sukhatme, S. (1972). Fredholm determinant of a positive kernel of a special type and its applications. *The Annals of Mathematical Statistics*, **43**, 1914–1926.

[81] Terry, M. *et al.* (1990). Martingale-based residuals for survival models. *Biometrika*, **77**, 147–160.

[82] Weisberg, S. (1985). *Applied Linear Regression*, John Wiley & Sons, New York.

[83] Yohai, V.J. and Maronna, R.A. (1979). Asymptotic behaviour of M-estimators for the linear model. *The Annals of Statistics*, **7**, 258–268.