Dimension Reduction Methods with Application in Automobile Insurance

by

Yaning Zhang

B.Sc., University of Victoria, 2016

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2023

Declaration of Committee

Name:	**Yaning Zhang**
Degree:	**Master of Science**
Thesis title:	**Dimension Reduction Methods with Application in Automobile Insurance**
Committee:	**Chair:** Jean-François Bégin Assistant Professor, Statistics and Actuarial Science
	Yi Lu Supervisor Professor, Statistics and Actuarial Science
	Himchan Jeong Committee Member Assistant Professor, Statistics and Actuarial Science
	Jiguo Cao Examiner Professor, Statistics and Actuarial Science

Abstract

Motivated by the data explosion in the automobile industry due to technological innovations, this report aims to provide an outline of how dimension reduction methods can be used in modelling automobile insurance claim amounts. The framework is based on a generalized linear model (GLM) with Tweedie distribution. Three popular methods are discussed in detail, the stepwise method, the principal component analysis (PCA) using the nonlinear iterative partial least squares (NIPALS) method, and the partial least squares method. The effectiveness and predictability of the three methods are compared using a car insurance data example. The results show that a small number of latent variables can effectively capture sufficient information in the explanatory variables, and can be utilized to build a decent predictive model for loss costs. Our study confirms that when multicollinearity exists in the dataset, using orthogonal latent variables can generally result in better modelling performance than ordinary variable selection methods.

Keywords: auto insurance; principal component analysis; partial least squares; generalized linear model

Acknowledgements

First I would like to express my gratitude to my supervisor, Professor Yi Lu for her continuous support and patience throughout my graduate studies. She encouraged me all the time in my academic research and daily life. I also would like to thank the committee members Professor Jean-Francois Begin, Professor Himchan Jeong and Professor Jiguo Cao for their time and invaluable suggestions. I am sincerely grateful to all the faculty members and staff in the Department of Statistics and Actuarial Science, who made this journey most enjoyable. Finally, I would like to thank my parents for believing in me and for their endless love.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

In light of the rapid growth of data collection in every aspect of people's lives, business analysts are now challenged with utilizing mass data for better and more efficient modelling. In the automobile insurance industry, we see that emerging technologies such as vehicle telematics provides actuaries with an abundance of raw data. The newly launched Tesla Insurance introduced by American electrical car giant Tesla calculates policyholders' premiums using metrics such as percentage of aggressive turning, hard breaking and unsafe following [1] . These new metrics, combined with the traditional vehicular information (vehicle value, automatic emergency braking system, etc.) and drivers' information (geographical area, driving experience, etc.) pose a new opportunity for improving model accuracy but also challenge actuaries to control the number of predictors in their models for efficiency.

There are numerous methods that help actuaries to reduce the number of variables. They can be generalized into two types: supervised learning methods and unsupervised learning methods. Some popular supervised dimension reduction methods include least absolute shrinkage and selection operator (LASSO), stepwise methods and partial least square algorithm, and unsupervised dimension reduction methods include principal component analysis (PCA) and clustering. The unsupervised methods focus on finding a low-dimensional representation that captures the most variance in the input variables, whereas the supervised methods allow researchers to project the relationship between the input data and output data, and build predictive models. Stepwise methods and LASSO are typical methods of feature selection. The goal of feature selection is to preserve the most informative variables and eliminate the redundant ones. Stepwise methods are greedy approaches as they try to find the subset of features using the least amount of computing power, even though they may only find local optimal subsets instead of global optimal subsets. LASSO allows analysts to regulate the models to avoid overfitting; however, the results can be unstable and

[1] https://www.tesla.com/insurance

less intuitive. Partial least squares and principal components both utilize latent variables to represent the original variables. Each component is a linear combination of the original variables. By assigning different weights (loadings), the methods can give more weight to significant variables and less weight to noises. The advantages of using these latent variables are that it removes collinearity among the predictors, and can reduce dimension at a low cost of model accuracy.

Loss cost, also known as pure premium, is the proportion of the premiums collected from policyholders that is used to cover the claim amounts incurred. This equals to the expected cost of all future claims from the policyholder. It is the primary interest of pricing actuaries. Combined with profit loading and overhead costs, the loss cost is the basis of the premium calculation. In an ideal situation, the price of an insurance product should be high enough to cover the losses, but also low enough to stay desirable compared to its competitors. Thus, an accurate estimation of the loss cost is the key to maintaining insurance companies' solvency and competitiveness. In the automobile insurance industry, a common practice involves dividing policyholders into homogeneous subgroups based on their risk characteristics, and a tariff for each group is calculated using regression methods. Each insurance company has its own models for calculating appropriate tariffs to neutralize the risks; however, they usually involve using statistical models to predict claim frequency and claim severity. Some novel studies explored other modelling approaches, for example, Guelman (2012) used gradient boosting trees for auto insurance loss cost modelling and prediction. Some insurance companies use variations of experience rating, which calculates premiums based on policyholders' claim history. It is also possible to use some form of the combination of risk classification and experience rating.

This report aims to provide a few dimension reduction methods for actuaries in modelling lost costs and compare them using an application in automobile insurance.

1.2 Outline

The report is organized as follows. Chapter 2 provides a literature review on modelling the loss cost in the automobile industry, and a number of dimensionality reduction methods and their applications. Chapter 3 presents the generalized linear model (GLM) framework and a special case of Tweedie GLM. Chapter 4 explains the modern dimensionality reduction methods extensively. This chapter is divided into two parts: the unsupervised learning methods and the supervised learning methods. In the unsupervised learning methods section, we mainly discuss principal component analysis (PCA), including different methods to conduct PCA. In the supervised learning methods section, we discuss two dimensionality reduction methods: variable selection and partial least squares (PLS). Chapter 5 provides an application using an automobile dataset to demonstrate the strengths and weaknesses of each method, and Chapter 6 draws actuarial implications from this application. Finally, Chapter 7 concludes the key findings of this report.

Chapter 2

Literature review

2.1 Generalized linear model and actuarial ratemaking

The generalized linear model (GLM) is an extension of the Gaussian linear models. It loosens the normality assumption and homoscedasticity assumption of the error; thus it allows researchers to fit a variety of models of response variables with different distributions. The process of model selection, parameter estimation and prediction in GLM is discussed in detail in McCullagh and Nelder (2019).

Modelling the claims is critical in insurance ratemaking. The claim frequency and claim severity are two interests of pricing actuaries. Yip and Yau (2005) used six regression models to explore which could successfully capture the zero inflation in the distribution of claim frequency. The same dataset used by Yip and Yau (2005) is discussed in this report. While Yip and Yau (2005) focused on modelling the claim frequency, in property and casualty insurance it is common to model the severity of claims too. The aggregate claim amounts can be modelled by combining the two models, assuming that the frequency and severity are independent. Poisson distribution or negative binomial distribution is often used for the mean frequency estimation and a gamma distribution is used for the mean severity estimation. Jørgensen and Paes De Souza (1994) suggested a compound Poisson distribution with a gamma random variable called Tweedie to estimate the aggregate loss. The Tweedie distribution can be viewed as a re-parameterization of the compound Poisson-gamma distribution (Quijano Xacur and Garrido, 2015). As the Tweedie distribution belongs to the linear exponential family, we thus can use it directly to estimate the mean aggregate loss in a GLM setting. Dunn and Smyth (2018) demonstrated the estimation of Tweedie GLM parameters using R. Numerous studies have tried to use the Tweedie GLM for loss prediction and insurance ratemaking; see, for example, Smyth and Jørgensen (2002) and Yang et al. (2018). Denuit et al. (2007) discussed in detail the ratemaking process in the property and casualty industry, including risk classification, credibility and bonus-malus systems. Although the focus of this book is the modelling of claim counts, it provides a broader perspective on how statistical modelling is applied in the automobile insurance industry.

2.2 Development on regression with latent variables

Principal component analysis (PCA) was initially developed in the field of chemometrics before it gained popularity in many other scientific fields. Wold (1968) outlined the nonlinear iterative partial least squares (NIPALS) algorithm to derive the principal components, while singular value decomposition (SVD) of the data matrix and eigen-decomposition of the covariance matrix are also two commonly used methods. Abdi and Williams (2010) showed that how PCA can be obtained from the SVD of the data matrix. They also illustrated how to find the principal components geometrically by first finding the main direction (first component) where the data points exhibit the largest variance, and then the second principal component orthogonal to the first can be found.

Although PCA is initially developed for numeric variables, a few methods are proposed to circumvent the existence of categorical variables. Filmer and Pritchett (2001) proposed to use dummy variables for each categorical variable level, as an analogy for how categorical variables are used in regression models. Although this method can be simple and efficient for PCA, a shortcoming of the dummy variables is that they cannot incorporate ordering. By using a large simulation study, Kolenikov and Angeles (2004) showed that when non-ordering of categorical variables cannot be assumed, the method proposed by Filmer and Pritchett (2001) is inferior to using ordinal variables or polychoric correlations.

The purpose of PCA is to reduce the dimensionality of the data and the multicollinearity problem that resulted from it. PCA is an unsupervised method as the derivation of the principal components does not involve any response variable. Principal component regression (PCR) is a regression model that uses a selected number of principal components as the independent variables instead of the original variables; however, it has a critical disadvantage: the derivation of the components and the regressing are two separate steps; thus, we may risk discarding principal components that contain useful information or keeping principal components that consist mainly of noise (Wold et al., 1987). To remedy this disadvantage, a supervised version of the PCA, partial least squares (PLS), is developed. The PLS model consists of two relations: the inner relation that links both the explanatory variables (or blocks) $\boldsymbol{X}$ and the response variables (or blocks) $\boldsymbol{Y}$, and the outer relation that treats the explanatory block and response block individually (Wold et al., 1987). Similar to PCA, a selected number of partial least square components can be used to model one or multiple response variables, known as partial least squares regression (PLSR).

Bastien et al. (2005) extended PLSR to generalized linear regression. The PLS generalized linear regression (PLS-GLR) algorithm has the same rationale as the PLSR, but when building the inner link between $\boldsymbol{X}$ and $\boldsymbol{Y}$, the algorithm uses an assumed GLM instead of a linear model (Bastien et al., 2005). The PLS-GLR presents many possibilities for empirical studies.

Chapter 3

Generalized linear model

This section discusses the theoretical framework of the generalized linear model, including the assumptions of GLM, and also the estimation and testing of GLM. We then narrow our focus on the GLM that is used in modelling the aggregate claim amounts in auto insurance, specifically, the Tweedie GLM.

3.1 General framework

This report focuses on modelling the generalized linear model (GLM) with an underlying Tweedie distribution. The GLM is an extension of the Gaussian linear models and consists of three components: probability distribution, linear predictor and link function.

The probability distribution used in the GLM is the random component of the model. Compared to the linear model, GLM loosens the assumption of normal distribution of the response variable Y. Instead, we assume that Y follows a distribution from the linear exponential family (see, for example, Frees, 2009 and Goldburd et al., 2016); it includes the normal, exponential, Poisson, Bernoulli and Tweedie distributions as special cases. A probability distribution is a member of the linear exponential family if its density function can be expressed as:

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + S(y,\phi)\right), \tag{3.1}$$

where y is the response (or dependent) variable (discrete or continuous), θ is the natural parameter and ϕ is a scale parameter, Note here that function $b(\theta)$ depends only on the parameter θ, while $S(y,\phi)$ is a function of the response variable y and the scale parameter ϕ. The mean and variance of Y can be easily obtained from (3.1) as

$$\mathrm{E}[Y] = b'(\theta), \qquad \mathrm{Var}(Y) = \phi\, b''(\theta).$$

The linear predictors of a GLM are the systematic component of the model. A linear predictor expressed as a linear combination of the explanatory variables $x_1, x_2, \dots, x_p$:

$$\eta = \sum_{j=0}^{p} x'_j \beta_j = \boldsymbol{x}'\boldsymbol{\beta},$$

where $\boldsymbol{x} = (1, x_1, x_2, \dots, x_p)'$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)'$ is a column vector of (unknown) parameters to be estimated.

In the context of GLM, a link function represents how the mean response is linked with the predictors. It is customary to use $\mu_i = \mathrm{E}[Y_i]$ to represent the mean of the i^{th} response variable Y_i, and it is associated with the linear predictors through the expression

$$\eta_i = \boldsymbol{x}'_i\boldsymbol{\beta} = g(\mu_i),$$

where function g is known as the link function. Accordingly, the mean function can be expressed as $\mu_i = g^{-1}(\boldsymbol{x}'_i\boldsymbol{\beta})$. Recall that $\eta_i = g(\mu_i)$ and $\mu_i = b'(\theta_i)$. When function g is chosen in terms of the inverse of $b'(\theta_i)$ (i.e., $g^{-1} = b'$, implying that $\eta_i = g(b'(\theta_i)) = \theta_i$), the link function g is called the canonical link.

Table 3.1 presents some special cases of GLM along with their corresponding name and formula of the (canonical) link function $g(\mu_i)$, mean function $b'(\theta_i)$, and the range of response random variable. Note that the

Table 3.1: Some special cases of GLM

Distribution	Link function	Formula	Mean function	Range
Normal	identity	μ_i	θ_i	$(-\infty, +\infty)$
Bernoulli	Logit	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$	$\frac{e^{\theta_i}}{1+e^{\theta_i}}$	$\{0, 1\}$
Poisson	Log	$\log(\mu_i)$	e^{θ_i}	$\mathbb{N}^+$
Inverse Gaussian	Inverse-square	$-\frac{1}{2\mu_i^2}$	$(-2\theta_i)^{-1/2}$	$(0, +\infty)$
Gamma	Inverse	$-\frac{1}{\mu_i}$	$-\frac{1}{\theta_i}$	$(0, +\infty)$

Before we present the general estimation method for the model parameters of GLMs, we state below the assumptions used in GLMs. They are:

(A1) The response variables are independently distributed conditional on the observed explanatory variables;

(A2) Explanatory variables are non-stochastic and exogenous;

(A3) The response variables follow a specified distribution in the linear exponential family described by (3.1);

(A4) The mean of the response variable is linked to the linear predictors through the link function, while the variance of the response variable can be written as a function of the mean of the response variable, i.e., $\mathrm{Var}(Y_i) = \phi\, v(\mu_i)$, where ϕ is a constant known as the scale parameter and $v(\mu_i)$ is the variance function.

The maximum likelihood estimation method can be used for estimating the GLM parameters. The parameters are estimated by maximizing the likelihood, or equivalently, the log-likelihood of the parameters from the observed data (McCullagh and Nelder, 2019).

The log-likelihood of independent observation $y_1, y_2, \ldots, y_n$ is the sum of the individual log-likelihood based on (3.1), i.e.,

$$l(\vartheta; \boldsymbol{y}) = \sum_{i=1}^{n} \left(\frac{y_i \theta_i - b(\theta_i)}{\phi_i} + S(y_i, \phi_i) \right), \tag{3.2}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ is a vector of n observations, and $\vartheta = \{\theta_1, \ldots, \theta_n, \phi_1, \ldots, \phi_n\}$ is a set of parameters to be estimated.

When the distribution parameter varies by known weight factors, we can write $\phi_i = \phi / w_i$. Recall that $\theta_i = \eta_i = \boldsymbol{x}_i' \boldsymbol{\beta}$, then equation (3.2) can be rewritten as a function of parameter $\boldsymbol{\beta}$ and ϕ, namely,

$$l(\boldsymbol{\beta}, \phi; \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} \left\{ w_i \frac{y_i \boldsymbol{x}_i' \boldsymbol{\beta} - b(\boldsymbol{x}_i' \boldsymbol{\beta})}{\phi} + S\left(y_i, \frac{\phi}{w_i}\right) \right\}, \tag{3.3}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots, \boldsymbol{x}_n')'$ is the design matrix containing explanatory variables.

To find $\boldsymbol{\beta}$ that maximizes log-likelihood function (3.3), the score function is calculated by taking the partial derivative of equation (3.3) with respect to $\boldsymbol{\beta}$; this gives

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}, \phi; \boldsymbol{y}, \boldsymbol{X}) = \frac{1}{\phi} \sum_{i=1}^{n} \left(y_i - b'(\boldsymbol{x}_i' \boldsymbol{\beta}) \right) w_i \boldsymbol{x}_i. \tag{3.4}$$

By setting the score function (3.4) to zero we obtain

$$\boldsymbol{0} = \sum_{i=1}^{n} w_i \left(y_i - b'(\boldsymbol{x}_i' \boldsymbol{\beta}) \right) \boldsymbol{x}_i. \tag{3.5}$$

The parameters estimated by maximizing the log-likelihood are known as the maximum likelihood estimators (MLE), denoted $\hat{\boldsymbol{\beta}}$. Apart from a few special cases, there is no closed-form solution to finding the MLE in GLM. Instead, they can be calculated using an iterative weighted least squares procedure (Frees, 2009). A variation of the Newton-Raphson method called Fisher scoring can be used. The algorithm initiates by a guess of $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}_{old}$, and then updates it by using the following iteration relationship:

$$\boldsymbol{\beta}_{new} = \boldsymbol{\beta}_{old} - \boldsymbol{I}(\boldsymbol{\beta}_{old})^{-1} \boldsymbol{U}(\boldsymbol{\beta}_{old}), \tag{3.6}$$

where $\boldsymbol{I}(\boldsymbol{\beta}_{old})$ is the observed information matrix valued at $\boldsymbol{\beta}_{old}$ and $\boldsymbol{U}(\boldsymbol{\beta}_{old})$ is the score function given by (3.4) valued at $\boldsymbol{\beta}_{old}$. The expression of the information matrix is derived from the negative expected value of the second derivative of the log-likelihood function with respect to $\boldsymbol{\beta}$; it is given by

$$\boldsymbol{I}(\boldsymbol{\beta}) = -\mathrm{E}\left[\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} l(\boldsymbol{\beta}, \phi; \boldsymbol{y})\right].$$

This gives

$$\boldsymbol{I}(\boldsymbol{\beta}_{old}) = \frac{1}{\phi}\sum_{i=1}^{n} w_i b''(\boldsymbol{x}_i'\boldsymbol{\beta}_{old})\boldsymbol{x}_i\boldsymbol{x}_i'$$

Thus, equation (3.6) can be re-written as

$$\boldsymbol{\beta}_{new} = \boldsymbol{\beta}_{old} - \left(\sum_{i=1}^{n} w_i b''(\boldsymbol{x}_i'\boldsymbol{\beta}_{old})\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n} w_i(y_i - b'(\boldsymbol{x}_i'\boldsymbol{\beta}_{old}))\boldsymbol{x}_i\right). \tag{3.7}$$

When the difference between $\boldsymbol{\beta}_{new}$ and $\boldsymbol{\beta}_{old}$ is smaller than a predetermined criterion (usually very small), the algorithm converges and reports $\hat{\boldsymbol{\beta}}$. This method is math-intensive and time-consuming if solved by hand. Luckily, most software today can produce the solution in a few seconds. Asymptotically, the MLE $\hat{\boldsymbol{\beta}}$ is consistent and the variance-covariance matrix of $\boldsymbol{\beta}$ are given by $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$. The variance of $\boldsymbol{\beta_j}$ is the j^{th} diagonal of $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$, and the off-diagonal element is the covariance of $\boldsymbol{\beta_i}$ and $\boldsymbol{\beta_j}$, where $i \neq j$. The parameter ϕ can be estimated by the Pearson estimate as indicated in Dunn and Smyth (2018), where the estimation of the parameters in Tweedie GLMs using R is presented with details.

The strong correlation that exists in pairs of explanatory variables can cause GLM regression to be highly unstable (Goldburd et al., 2016). When similar information is entered into GLM twice, it may cause coefficients to be extremely high or low, and also have large standard errors. Another issue associated with correlation is multicollinearity. It is caused when the combination of two or more predictors is strongly predictive of another variable. It is harder to detect multicollinearity from a correlation matrix because the variable may not be highly correlated with the other variables individually (Goldburd et al., 2016).

3.2 Tweedie GLM

In the application of automobile insurance, the pure premium, also known as loss cost, is the total claim amount incurred in a policy period (normally a year). The modelling of pure premium is the primary interest of pricing actuaries since it is the proportion of the collected premium that is used to cover the loss. Correct modelling of pure premium helps to price the insurance product so enough premium is collected to cover the loss while keeping the product competitive in the market. The distribution of the loss costs is composed of two components: the frequency component and the severity component. The claim frequency denotes the number of claims per policy period. The claim severity is the average size of the claim, which equals to the total claim amount incurred in a policy period divided by the number of claims. The total claim amount during a policy period is also called the aggregate loss amount; it is calculated by adding all the claim amounts incurred during the policy period. The duration of a policy is called the exposure. In automobile insurance, one full year of policy duration from one policyholder is one exposure. In this project, we consider policyholders who keep the policy until it expires after a full year. For policyholders with various exposures, an offset can be added to modify the model, this will be discussed later in this section.

Because of the nature of the insurance claims, the distribution of aggregate loss amount is highly skewed. There is generally a large mass at point zero that represents policies with no claims throughout the policy period, and the remaining mass is the distribution of non-zero claims. Generally, the aggregate loss can be modelled using two alternative methods. The separated Poisson-gamma approach (SPGA) models the frequency component and the severity component separately, and Tweedie GLM models the aggregated loss using only one distributional model. The comparison between these two approaches is discussed by Quijano Xacur and Garrido (2015). Since Tweedie GLM is a simpler model, it is preferred over SPGA when the parsimony principle applies.

To introduce Tweedie GLM, we first introduce the Tweedie family of distributions. In addition to the mean parameter μ and the scale parameter ϕ, a distribution that belongs to the Tweedie family has a third parameter p, called the power parameter. A distribution belongs to the Tweedie family if the variance function mentioned in assumption (A4) can be written as

$$v(\mu) = \mu^p,$$

for some real number p that is not in the interval 0 to 1 (non-inclusive).

A distribution that belongs to the Tweedie family is characterized by the value of p (Quijano Xacur and Garrido, 2015). Table 3.2 lists some popular distributions that belong to the Tweedie family. A Tweedie distribution with a power parameter between 1 and 2 can be used to represent the characteristics of aggregate loss. It is also known as a compound Poisson-Gamma distribution.

Value of p	Distribution
$p = 0$	normal
$p = 1$	Poisson
$p \in (1, 2)$	compound Poisson-gamma
$p = 2$	gamma
$p = 3$	inverse Gaussian

Table 3.2: Tweedie distributions with different power parameter p

Let N be the count variable representing the number of claims in a policy period. We assume that N follows a Poisson distribution with parameter (mean) λ. Its probability function is given by

$$p_n = \mathrm{P}(N = n) = \frac{e^{\lambda}\lambda^n}{n!}, \qquad n = 0, 1, 2, \dots .$$

Let X_i be a random variable representing the i^{th} claim amount. Suppose that $X_1, X_2, ...$ are identically and independently distributed with a Gamma(α, τ) distribution, and is independent of N. The density function of the Gamma(α, τ) distribution is given by

$$f_X(x) = \frac{\tau^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\tau x}, \qquad x > 0,\ \alpha > 0,\ \tau > 0,$$

where α and τ are the shape and rate parameters, respectively. The mean of X is α/τ and the variance of X is α/τ^2.

Then, the aggregate loss S defined as

$$S = X_1 + X_2 + ...X_N$$

follows a compound Poisson-gamma distribution. Its probability density function can be expressed as

$$f_S(x) = \begin{cases} e^{-\lambda}, & y = 0 \\ \sum\limits_{k=1}^{\infty} \left(\dfrac{e^{-\lambda}\lambda^k}{k!} \cdot \dfrac{\tau^{k\alpha} y^{k\alpha-1} e^{-\tau y}}{\Gamma(k\alpha)} \right), & y > 0 \end{cases}. \tag{3.8}$$

Note that the probability of the aggregate loss $S = 0$ is equal to the probability of zero claims, i.e., $N = 0$. The distribution of S with its density function given by (3.8) is parameterized by the three parameters λ, α and τ.

The mean and the variance of S can be easily calculated as

$$\mathrm{E}[S] = \frac{\lambda\alpha}{\tau}, \qquad \mathrm{Var}(S) = \frac{\lambda\alpha}{\tau^2}(1 + \alpha).$$

Frees (2009) states that the Tweedie distribution can be shown as a member of the linear exponential family by defining three parameters μ, ϕ and p by the relations

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \qquad \alpha = \frac{2-p}{p-1}, \qquad \frac{1}{\tau} = \phi(p-1)\mu^{p-1}. \tag{3.9}$$

Then (3.8) can be written, for $y \geq 0$, as

$$f_S(y) = \exp\left\{ \frac{-1}{\phi} \left(\frac{\mu^{2-p}}{2-p} + \frac{y}{(p-1)\mu^{p-1}} \right) + S(y, \phi) \right\}. \tag{3.10}$$

Note here the distribution of S with its density function given by (3.10) is parameterized by the three parameters μ, ϕ and p with $p \in (1, 2)$. The derivation of an explicit expression of $S(y, \phi)$ can be found in Wüthrich (2003). If we further set $\theta = \mu^{1-p}/(1-p)$ and $b(\theta) = \mu^{2-p}/(2-p)$, we can see that the density of S given by (3.10) is of the form of (3.1) for the linear exponential family. It is worth mentioning that the Tweedie distribution when $p \notin (0, 1)$ also belongs to the exponential dispersion family because it extends the linear exponential family (Jørgensen, 1992). In this sense, parameter p is also called the dispersion parameter.

From (3.9), we get the power parameter of the Tweedie distribution $p = (2+\alpha)/(\alpha+1)$, which is a function of the shape parameter α of the gamma distribution. If p is close to 2 (implying α tends to zero), there is little variance in the gamma distribution, so the Poisson distribution is the main source of the randomness in the Tweedie distribution. The shape of the probability density function of this Tweedie distribution is close to the probability density function of a Poisson distribution. On the other hand, if p is close to 1 (implying α tends to be very large), the probability density function of the Tweedie distribution resembles a gamma distribution, but with a mass at point zero (Goldburd et al., 2016).

In application, a number of possible p candidates are used for testing and the optimal value of p is chosen using the maximum likelihood method when data is available. In insurance modelling, the value of p is generally between 1.5 and 1.8, many modellers simply choose a value when fine-tuning the value of p becomes costly. Some common choices are 1.6, 1.67 and 1.7 (Goldburd et al., 2016).

The Tweedie distribution is a member of the linear exponential family and thus can be used as an underlying distribution in GLM. It is typical to use a log link when modelling with Tweedie GLM as it captures the non-negativity of the response variable, namely,

$$\log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta},$$

where $\mu_i = \mathrm{E}[S_i]$, $\boldsymbol{x_i}$ represents the explanatory variable and $\boldsymbol{\beta}$ is a vector of corresponding coefficients.

The aggregate loss amounts are expected to vary directly with exposure. With everything else held fixed, a policy with one exposure is expected to have twice the claim amounts than a policy with half an exposure. This expectation can be reflected in a GLM as an exposure offset (Goldburd et al., 2016). When we consider policyholders with different exposure in the Tweedie GLM model, it becomes

$$\log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \text{offset},$$

where offset=log(number of exposures).

Chapter 4

Dimension reduction methods

When modelling with a large number of predictors, two major problems one is likely to encounter are high-dimension and in turn, multicollinearity. The word *curse of dimensionality* describes the situation that the sample size needed to maintain model accuracy grows exponentially with the number of variables. Dimension reduction methods refer to techniques that transform high-dimensional data into low-dimensional representations (Chao et al., 2019). After these representations are obtained, they can be used for classification and regression applications. The dimension reduction methods can be classified into two types, supervised learning methods and unsupervised learning methods. The main difference between the two is that the supervised learning methods incorporate both input and output variables when deriving the low-dimensional representations and the unsupervised learning methods only try to capture the information in the input variables. In this report, three methods are discussed: the forward stepwise selection, the principal component analysis and the partial least squares. The forward stepwise selection is an example of feature selection where only significant variables are kept. The principal component analysis is a popular learning method that tries to find a few latent variables to capture most information in the original inputs. The partial least squares method is similar to the principal component analysis but incorporates output variables in the algorithm. Among these three methods, principal component analysis is an unsupervised method and forward stepwise selection and partial least squares are supervised methods.

4.1 Unsupervised learning method

4.1.1 Principal component analysis

Principal component analysis (PCA) is an unsupervised learning method that utilizes a few orthogonal components that capture sufficient variability in all the explanatory variables in a dataset. These components are latent variables. They cannot be observed in the data, and can only be calculated. PCA is "unsupervised" as the derivation of the principal components does not involve the response variable. There are three popular methods that are used to calculate the principal components: eigenvalue decomposition, singular value decomposition (SVD) and non-linear iterative partial least squares (NIPALS).

This section briefly reviews the eigenvalue decomposition (eigen-decomposition) method and SVD method, and focuses on the NIPALS algorithm. The algorithms are well documented by Wold et al. (1987) and Abdi and Williams (2010).

Let $\boldsymbol{X}$ be an $I \times J$ data matrix that we are interested in, where I is the number of observations and J is the number of variables. If we fit a regression model or classification model, then $\boldsymbol{X}$ represents the explanatory variable matrix, where $\boldsymbol{X} = (x_{i,j})$, It is customary to centre the columns of $\boldsymbol{X}$. This is done by subtracting the mean of the column from each element in the column. After centring, the mean of each column equals zero. In this case, matrix $\boldsymbol{X'X}$ is the variance-covariance matrix. When the variables have different units, it is also important to normalize the columns. This is obtained by dividing each column by the standard deviation of this column. After centring and scaling, matrix $\boldsymbol{X'X}$ becomes the correlation matrix of $\boldsymbol{X}$. We further assume that the rank of $\boldsymbol{X}$, r, satisfies $r \leq \{I, J\}$ in general, or specifically, $r < J$ because we suppose that our data matrix contains correlated variables that we aim to reduce.

PCA is used to provide an approximation of the data matrix $\boldsymbol{X}$, in terms of the product of two matrices $\boldsymbol{T}$ and $\boldsymbol{P'}$, namely,

$$\boldsymbol{X} = \boldsymbol{TP'} + \boldsymbol{E}, \tag{4.1}$$

where $\boldsymbol{P} = (p_{j,m}) = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_M)$ is a $J \times M$ matrix and $\boldsymbol{T} = (t_{i,m}) = (\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_M)$ is a $I \times M$ matrix. If $\boldsymbol{X}$ has full rank, then $J = M$ and $\boldsymbol{P}$ is a square matrix. In expression (4.1), the matrix $\boldsymbol{T}$ is known as the score matrix, and its columns are called score vectors. In this report, the words scores and principal components are used interchangeably. The matrix $\boldsymbol{P}$ is known as the loading matrix and its columns are called loading vectors. The residual matrix $\boldsymbol{E}$ represents the part that is not explained by the principal components. By design, the loading vectors are all orthogonal to each other, i.e., $\boldsymbol{p}_i'\boldsymbol{p}_j = 0$, for $i \neq j$, implying $\boldsymbol{P'P} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

Once the scores and loadings are found with the data, we can use them to estimate new data or "test data", denoted by $\boldsymbol{X}_{new}$. It can be used to estimate a new score matrix, $\boldsymbol{T}_{new}$, through the relationship

$$\boldsymbol{T}_{new} = \boldsymbol{X}_{new}\boldsymbol{P}.$$

It can be seen that each component is a linear combination of the explanatory variables, and $\boldsymbol{P}$ contains the weights of $\boldsymbol{X}_{new}$ in this linear relationship. A scree plot is used to show the proportion of variance explained (also called variance accounted for (VAF) in some literature) by each component in both the eigenvalue and SVD decomposition methods that are presented below with details. The variance is explained by the m^{th} component, denoted by σ_m^2. The proportion of the variance explained by the m^{th} component is then given by

$$\frac{\sigma_m^2}{\sum\limits_{k=1}^{M} \sigma_k^2} = \frac{\sum\limits_{i=1}^{I} \left(\sum\limits_{j=1}^{J} x_{i,j} p_{j,m} \right)^2}{\sum\limits_{i=1}^{I} \sum\limits_{j=1}^{J} x_{i,j}^2}. \tag{4.2}$$

If one is interested in the cumulative variance explained by the first m components, denoted by ρ_m, it can be obtained by summing up the proportion of the variance explained from the first component to the m^{th} component using (4.2); that is,

$$\rho_m = \frac{\sum\limits_{k=1}^{m} \sigma_k^2}{\sum\limits_{k=1}^{M} \sigma_k^2}.$$

We now present the basic idea of these three methods.

Eigen-decomposition method

The first method to find the loadings and scores is by the eigen-decomposition of the correlation matrix. A matrix is said to be positive semi-definite when it can be obtained as the product of a matrix by its transpose. A positive semi-definite square matrix can be decomposed into eigenvectors and eigenvalues. This process is known as the eigen-decomposition. A vector $\boldsymbol{u}$ is said to be an eigenvector and a scalar λ is said to be the corresponding eigenvalue of square matrix $\boldsymbol{A}$ if it satisfies

$$\boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u},$$

or alternatively,

$$(\boldsymbol{A} - \lambda \boldsymbol{I}) = \boldsymbol{0}.$$

The eigen-decomposition method used in PCA can be considered an optimization method. The goal is to maximize the variance explained by the principal components (recall that $\boldsymbol{X}'\boldsymbol{X}$ is the correlation matrix), under the constraint that the loading matrix is an orthogonal matrix, i.e.,

$$\begin{aligned} \max \quad & \boldsymbol{T}'\boldsymbol{T} = \boldsymbol{P}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{P} \\ \text{s.t.} \quad & \boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}. \end{aligned} \tag{4.3}$$

Equation (4.3) can be solved using the Lagrangian multiplier $\boldsymbol{\lambda}$, i.e., solve for $\boldsymbol{P}$:

$$\max \quad \mathcal{L} = \text{trace}(\boldsymbol{P}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{P} - \boldsymbol{\lambda}(\boldsymbol{P}'\boldsymbol{P} - \boldsymbol{I})),$$

where $\boldsymbol{\lambda}$ is a diagonal $M \times M$ matrix, and the trace operator sums the diagonal elements of a square matrix.

To solve for $\boldsymbol{P}$ that maximizes $\mathcal{L}$, we can take the derivative of $\mathcal{L}$ with respect to $\boldsymbol{P}$ and set it to zero. This gives

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{P}} = 2(\boldsymbol{X}'\boldsymbol{X}\boldsymbol{P} - \boldsymbol{P}\boldsymbol{\lambda}) = \mathbf{0},$$

and we then have $\boldsymbol{X'XP} = \boldsymbol{P\lambda}$. By right multiplying $\boldsymbol{P'}$ on both sides, we get

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{P}\boldsymbol{\lambda}\boldsymbol{P}'. \tag{4.4}$$

Equation (4.4) is the eigen-decomposition of the matrix $\boldsymbol{X'X}$. The diagonal matrix $\boldsymbol{\lambda}$ consists of eigenvalues and the columns of the matrix $\boldsymbol{P}$ are the eigenvectors of the correlation matrix $\boldsymbol{X'X}$ paired to the eigenvalues in $\boldsymbol{\lambda}$.

Finally, we can find the score matrix $\boldsymbol{T}$ by $\boldsymbol{T} = \boldsymbol{XP}$. It is easy to see that

$$\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{P}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{P} = \boldsymbol{P}'\boldsymbol{P}\boldsymbol{\lambda}\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{\lambda}$$

as $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}$. Denote the m^{th} diagonal element of $\boldsymbol{\lambda}$ as λ_m, and then $\lambda_m/(n-1)$ is the variance explained by the m^{th} component. When using eigen-decomposition, the cumulative proportion of variance explained by the first m components can be expressed as

$$\rho_m = \frac{\sum\limits_{k=1}^{m} \lambda_k}{\sum\limits_{k=1}^{M} \lambda_k}. \tag{4.5}$$

<u>**SVD method**</u>

The singular vector decomposition method is closely related to the eigen-decomposition method. Under the SVD method, equation (4.1) becomes

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}' + \boldsymbol{E}. \tag{4.6}$$

We call $\boldsymbol{UDV'}$ as the SVD of the matrix $\boldsymbol{X}$, where $\boldsymbol{U}$ is the normalized eigenvectors of the matrix $\boldsymbol{XX'}$ and the columns in $\boldsymbol{U}$ are called left singular vectors of $\boldsymbol{X}$. The matrix $\boldsymbol{V}$ is the normalized eigenvectors of the matrix $\boldsymbol{X'X}$, and the columns in $\boldsymbol{V}$ are called right singular vectors of $\boldsymbol{X}$. The matrix $\boldsymbol{D}$ is the diagonal matrix with the singular values of $\boldsymbol{X}$ and $\boldsymbol{D}^2$ is the diagonal matrix of (non-zero) eigenvalues of the matrix $\boldsymbol{X'X}$ and matrix $\boldsymbol{XX'}$. As it can be seen that instead of decomposing the correlation matrix $\boldsymbol{X'X}$, we now work on the data matrix $\boldsymbol{X}$ directly.

When we use the SVD method, $\boldsymbol{V'}$ plays the same role as $\boldsymbol{P'}$, and the matrix $\boldsymbol{D}$ is a diagonal matrix with each diagonal element equal to the lengths of the column vectors of $\boldsymbol{T}$. The matrix $\boldsymbol{U}$ is the same as $\boldsymbol{T}$, but each column is normalized to length one. Thus the product $\boldsymbol{UD}$ is equivalent to $\boldsymbol{T}$. Unlike eigen-decomposition, the singular vector decomposition is not limited to square matrices; this is, in (4.6), $\boldsymbol{X}$ can be of any dimension.

NIPALS algorithm

The algorithm of NIPALS is well demonstrated by Wold (1968), Wold et al. (1987) and Geladi and Kowalski (1986). Unlike the first two methods, the NIPALS algorithm does not calculate all principal components at once, rather it is an iterative method. It is an important building block for a good understanding of partial least squares methods introduced in Section 4.2.2. The NIPALS algorithm calculates the first loading vector $\boldsymbol{p_1}$ and score vector $\boldsymbol{t_1}$, subtracts their product from $\boldsymbol{X}$ and uses the residual matrix to calculate the next component. The NIPALS algorithm is composed of the following steps.

Step 1: Start with a random column in $\boldsymbol{X}$ and denote it as $\boldsymbol{t}$; some people choose the column with largest absolute sum, but the choice of different initial $\boldsymbol{t}$ will arrive at the same result.

Step 2: Calculate loading vector $\boldsymbol{p}' = \boldsymbol{t}'\boldsymbol{X}/\boldsymbol{t}'\boldsymbol{t}$; this is the ordinary least squares solution for regressing all the columns of $\boldsymbol{X}$ onto $\boldsymbol{t}$.

Step 3: Normalize $\boldsymbol{p}$ to length one $\boldsymbol{p}/\|\boldsymbol{p}\|$.

Step 4: Calculate score vector $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{p}/\boldsymbol{p}'\boldsymbol{p}$, with $\boldsymbol{p}'\boldsymbol{p} = 1$; this is the ordinary least squares solution for regressing all the rows of $\boldsymbol{X}$ onto normalized $\boldsymbol{p}'$.

Step 5: Check for convergence of score vector $\boldsymbol{t}$ (check if the difference between $\boldsymbol{t}$ used in Step 2 and obtained in Step 4 is smaller than some predetermined criterion); if convergence did not happen, go to Step 2.

Step 6: Deflate $\boldsymbol{X}$ by subtracting the variability capture by this component: $\boldsymbol{E} = \boldsymbol{X} - \boldsymbol{t}\boldsymbol{p}'$, and use the residual matrix $\boldsymbol{E}$ as $\boldsymbol{X}$ in the next iteration.

The vectors $\boldsymbol{t}$ and $\boldsymbol{p}$ from each iteration form the columns of the score matrix $\boldsymbol{T}$ and loading matrix $\boldsymbol{P}$. Note in Step 3, the vector $\boldsymbol{p}$ is normalized by dividing its length $\|\boldsymbol{p}\|$. After normalizing, the length of $\boldsymbol{p}$ is equal to one. The length of vector $\boldsymbol{p}$ is defined as the square root of the sum of its squared elements, or the square root of $\boldsymbol{p}'\boldsymbol{p}$ (Strang, 2006).

It can be shown that at convergence, the loading matrix and the score matrix calculated using the NIPALS method are the same as the ones calculated using eigen-decomposition: by denoting the scalar $\boldsymbol{t}'\boldsymbol{t}$ in Step 2 as C and by substituting $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{p}$ in Step 4 into $\boldsymbol{p}' = \boldsymbol{t}'\boldsymbol{X}/C$ in Step 2, we get $C\boldsymbol{p}' = \boldsymbol{p}'\boldsymbol{X}'\boldsymbol{X}$ or $\boldsymbol{p}C\boldsymbol{p}' = \boldsymbol{X}'\boldsymbol{X}$. Therefore, $C = \boldsymbol{t}'\boldsymbol{t}$ is an eigenvalue of the matrix $\boldsymbol{X}'\boldsymbol{X}$, which is denoted as a diagonal element of $\boldsymbol{\lambda}$ earlier in this section, and $\boldsymbol{p}$ is the eigenvector paired with the eigenvalue. Although the eigen-decomposition of the correlation matrix and the NIPALS algorithm produces the same results, one important difference between the two methods should be noted. If $\boldsymbol{X}$ contains missing values, the eigen-decomposition method is unfeasible as the square matrix $\boldsymbol{X}'\boldsymbol{X}$ cannot be calculated, whereas the NIPALS method is still functional. To see this, observe the loading vector $\boldsymbol{p}'$ can still be estimated in Step 2 in the NIPALS algorithm. The regressions of columns of $\boldsymbol{X}$ onto $\boldsymbol{t}$ can still be estimated with a few missing values.

It is worth noting that PCA is also applicable when some or all the explanatory variables are categorical variables. A common treatment is to transform these categorical variables into dummy variables as proposed by Filmer and Pritchett (2001).

The score matrix $\boldsymbol{T}$ is a representation of the original matrix $\boldsymbol{X}$, and thus can be used to build multivariate linear regression (MLR) models on the response variable $\boldsymbol{Y}$. In the linear regression model, the MLR can be written as:

$$E[\boldsymbol{Y}] = \sum_{h=1}^{M} c_h \boldsymbol{t}_h,$$

where the parameter c_h can be estimated using the ordinary least squares method. In the GLM framework, we can compose

$$g(\boldsymbol{\mu}) = \sum_{h=1}^{M} c_h \boldsymbol{t}_h,$$

where g is the link function used in GLM, $\boldsymbol{\mu}$ is the mean vector of $\boldsymbol{Y}$ and the parameter c_h can be estimated using the maximum likelihood method. The regression models using principal components are also called principal component regression (PCR).

To choose the number of principal components to be included in a model, two common methods are used: 1) find the number of components that explains enough (usually at least 80% or 90%) variance in the explanatory variables, and 2) find the number of components that gives the least cross-validation error.

4.2 Supervised learning method

4.2.1 Stepwise selection

This section discusses variable selection using a stepwise method in the generalized linear model framework. There are three types of stepwise methods: forward selection, backward elimination and stepwise regression (Keith, 2019). The last one is often referred to as the hybrid method because it can be considered a combination of forward selection and backward elimination.

The forward selection method can be summarized in the following steps: step 0) construct the null model by using only the intercept; step 1) add one more predictor each time to the model, using a specified criterion, the best predictor from all the remaining predictors is kept; step 2) repeat step 1) until a specific stop criterion is reached. The criterion of which predictor is entered into the model in step 1) is subjective. Some commonly used criteria are the p-value of the predictor, R^2 (coefficient of determination), AIC (Akaike information criterion) and BIC (Bayesian information criterion).

Using the p-value for the addition criterion and stopping criterion requires one to choose a specific p-value, usually 0.01 or 0.05. In step 1), the predictor with the smallest p-value will be added. When the smallest p-value is greater than the pre-set value, the algorithm stops. When the data contains categorical variables, using the p-value directly might lead to erroneous results. As Cohen (1991) points out that when dummy variables are used to map the categorical variables, treating them as different predictors provide meaningless results in the forward selection procedure. Moreover, when the reference level of the variable changes in the model, it will lead to different conclusions. Thus, when the dataset contains categorical variables, at each step of the forward selection, all levels of the same categorical variable should be tested at the same time, with the degree of freedom equal to the number of levels minus 1.

Adding the predictor that produces the largest coefficient of determination R^2 is also a common practice; however, R^2 that is calculated based on the residual sum of squares in the linear model cannot be extended to the cases under the generalized linear model framework. Instead, some equivalent measures can be calculated, known as the pseudo-R^2. The most popular pseudo R^2 is introduced by McFadden (1973), later known as Mcfadden's R^2. Mcfadden's R^2, denoted as R^2_{mf}, is based on the likelihood theory, and is given by

$$R^2_{mf} = 1 - \frac{L_p}{L_0}, \tag{4.7}$$

where L_p is the log-likelihood of the model with p predictors and L_0 is the log-likelihood of the null model. When the saturated model is fitted, the model explains the data perfectly and the log-likelihood of the model is 0. We can easily see that the maximum value of R^2_{mf} is 1. It is worth noting that even when the model is well-fitted, the pseudo R^2 can be really small (Hosmer Jr et al., 2013).

The AIC is another popular selection criterion. Akaike (1974) first introduced this information criterion as an extension to the maximum likelihood principle. The AIC is given by

$$\text{AIC} = 2k - 2L_p, \tag{4.8}$$

where k is the number of independent parameters. The AIC statistic provides an overview of the performance of the models that take the number of model parameters into consideration. Throughout this report, AIC is used as a goodness-of-fit measure of models. A smaller AIC statistic implies a better fitted model so that small AIC values are desired. The algorithm stops when AIC values start to increase. Using (4.7) in equation (4.8), we get

$$\begin{aligned}\mathrm{AIC} &= 2k - 2L_p \\ &= 2k - 2L_0\left(1 - R^2_{mf}\right).\end{aligned}$$

With a negative L_0 (almost always true), we can see that AIC is minimized only when R^2_{mf} is maximized, and vice versa. This shows that the decision of whether or not to include a predictor using either AIC or Mcfadden's R^2 criterion is the same.

The forward selection is greedy and reduces the computing power needed for variable selection significantly, because it only adds one predictor at a time, rather than comparing every possible subset of predictors.

The drawbacks of stepwise methods are well documented. Since the stepwise methods do not consider all possible subsets of predictors, it is well likely that significant predictors do not make it to the optimal model because of the stopping criterion. Using AIC as an example, the forward selection procedure may choose a local minimum instead of the global minimum. Moreover, one of the most critical concerns is that the stepwise methods will not select the best subset of predictors when multicollinearity exists (Fox, 2019).

4.2.2 Partial least squares

The partial least squares method is closely related to the NIPALS methods discussed in Section 4.1.1. In fact, some refer to the PLS method as a supervised version of the NIPALS method (Gareth et al., 2013). Geladi and Kowalski (1986) summarized the algorithm in a concise manner. Bastien et al. (2005) extended the method from linear regression to generalized linear regression. This section discusses the details of the derivation of PLS components and PLS generalized linear regression (PLS-GLR).

The complete PLS procedure consists of the following steps.

Step 1: Compute all the PLS components $\boldsymbol{t}_h$ and select the number of components wished to use in the regression model using cross-validation.

Step 2: Regress $\boldsymbol{y}$ on the selected PLS components using a generalized linear model.

Step 3: Express the PLS-GLR in terms of original explanatory variables.

To compute the PLS components, we first assume explanatory vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ are centred, and scaled if they have different units. If the PLS is used with multiple linear regression, it is also customary to centre and scale the response variables $\boldsymbol{y}$; however, for our purpose of constructing GLM, we leave $\boldsymbol{y}$ as is. The following algorithm shows the steps of calculating the 1^{st} PLS component and the h^{th} PLS component for $h = 2, \ldots, M$.

Computation of the 1^{st} PLS component

Step 1: Regress $\boldsymbol{y}$ on $\boldsymbol{x}_j$ for each j, $j = 1, 2, \ldots, J$, using generalized linear model and obtain the regression coefficients $a_{11}, a_{12}, \ldots, a_{1J}$.

Step 2: Normalize the vector $\boldsymbol{a}_1 = (a_{11}, a_{12}, \ldots, a_{1J})'$ and obtain the first loading vector $\boldsymbol{w}_1 = \boldsymbol{a}_1/\|\boldsymbol{a}_1\|$.

Step 3: Compute the component $\boldsymbol{t}_1 = \boldsymbol{X}\boldsymbol{w}_1/\boldsymbol{w}_1'\boldsymbol{w}_1$, where $\boldsymbol{w}_1'\boldsymbol{w}_1 = 1$.

Computation of the h^{th} PLS component

Step 1: Regress $\boldsymbol{y}$ on $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_{h-1}$ and $\boldsymbol{x}_j$ for each j, $j = 1, 2 \ldots, J$, using generalized linear model and obtain the regression coefficients for $\boldsymbol{x}_j$, $a_{h1}, a_{h2}, \ldots, a_{hJ}$.

Step 2: Normalize the vector $\boldsymbol{a}_h = (a_{h1}, a_{h2}, \ldots, a_{hJ})'$ and obtain the h^{th} loading vector $\boldsymbol{w}_h = \boldsymbol{a}_h/\|\boldsymbol{a}_h\|$.

Step 3: Compute the residual matrix $\boldsymbol{X}_{h-1}$ from the regression of $\boldsymbol{X}$ on $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_{h-1}$.

Step 4: Compute the component $\boldsymbol{t}_h = \boldsymbol{X}_{h-1}\boldsymbol{w}_h/\boldsymbol{w}_h'\boldsymbol{w}_h$, where $\boldsymbol{w}_h'\boldsymbol{w}_h = 1$.

Step 5: Express $\boldsymbol{t}_h$ in terms of original variables: $\boldsymbol{t}_h = \boldsymbol{X}\boldsymbol{w}_h$.

By construction, the algorithm ensures the orthogonality of the PLS components because they are computed from the residual matrix from the last iteration. Step 5 and Step 4 are equivalent, but $\boldsymbol{t}_h$ is expressed in different ways. This is because

$$\begin{aligned}
\boldsymbol{t}_h &= \boldsymbol{X}_{h-1}\frac{\boldsymbol{w}_h}{\boldsymbol{w}_h'\boldsymbol{w}_h} && \text{(Step 4)}\\
&= \left(\boldsymbol{X} - \sum_{i=1}^{h-1}\boldsymbol{t}_i\boldsymbol{w}_i'\right)\boldsymbol{w}_h && (\boldsymbol{w}_h'\boldsymbol{w}_h = 1)\\
&= \boldsymbol{X}\boldsymbol{w}_h - \sum_{i=1}^{h-1}\boldsymbol{t}_i\boldsymbol{w}_i'\boldsymbol{w}_h && (\boldsymbol{w}_i'\boldsymbol{w}_j = 0,\, i \neq j)\\
&= \boldsymbol{X}\boldsymbol{w}_h. && \text{(Step 5)}
\end{aligned}$$

By comparing Step 1 of the PLS algorithm and Step 2 in the NIPALS algorithm, we can see that they demonstrate a similar idea. Instead of regressing the selected column of $\boldsymbol{X}$ onto other columns of $\boldsymbol{X}$ using simple linear regression, the PLS algorithm uses GLM to regress $\boldsymbol{y}$ instead. The distribution of the response variable can be chosen by researchers to fit their assumptions.

Similar to PCR, the PLS-GLR model with m components can be expressed as

$$g(\boldsymbol{\mu}) = \sum_{h=1}^{M} c_h\boldsymbol{t}_h$$

$$= \sum_{h=1}^{M} c_h \left(\sum_{j=1}^{J} w_{hj} \boldsymbol{x}_j \right), \tag{4.9}$$

where g is the link function used in GLM, $\boldsymbol{\mu}$ is the mean vector of $\boldsymbol{y}$, $\boldsymbol{t_h} = \left(\sum_{j=1}^{J} w_{hj} \boldsymbol{x}_j \right)$ are the orthogonal PLS components that can be expressed as a linear combination of $\boldsymbol{x_j}$ for $j = 1, 2, \ldots, J$, c_h are the coefficients to be estimated in the GLM, and $\{w_{hj}\}$ are the loadings calculated in the PLS algorithm.

By changing the order of the summation in equation (4.9), we can see that

$$\begin{aligned} g(\boldsymbol{\mu}) &= \sum_{j=1}^{J} \left(\sum_{h=1}^{M} c_h w_{hj} \right) \boldsymbol{x_j} \\ &= \sum_{j=1}^{J} \beta_j \boldsymbol{x_j}, \end{aligned} \tag{4.10}$$

where $\beta_j = \sum_{h=1}^{M} c_h w_{hj}$, $j = 1, 2, \ldots, J$, are the coefficients of multiple generalized linear regression.

For predictive analysis using PLS components, pre-process the new matrix $\boldsymbol{X}_{new}$ so that the variables are treated the same way as $\boldsymbol{X}$; then obtain the new score matrix by

$$\boldsymbol{T}_{new} = \boldsymbol{X}_{new} \boldsymbol{W},$$

where $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M)$ is the weight matrix obtained from the PLS algorithm using $\boldsymbol{X}$. The new score matrix $\boldsymbol{T}_{new}$ can then be used for predictive analysis using PLS-GLR.

Chapter 5

Application

To compare the proposed dimension reduction methods presented in Chapter 4, we apply those methods to an automobile insurance dataset. After the number of predictors is reduced, the retained predictors are used to fit a Tweedie GLM that models aggregate claim amounts. The goal is to evaluate the forward selection, NIPALS, and PLS methods by model accuracy and the ability to reduce the number of variables.

5.1 Data description

In this study, we consider an auto insurance dataset studied by Yip and Yau (2005). The dataset is from SAS Enterprise Miner and can also be retrieved from R package 'cplm'. The original dataset contains 10,296 records over the period of 1993-1999. We consider only the policyholders who held the policy in the latest year with complete information. This is consistent with Yip and Yau (2005). There are 2,341 records left after the selection. There are twenty variables of driver characteristics and vehicle characteristics such as distance to work, the value of the vehicle and the driver's violation record. These variables are treated as explanatory variables for the GLM. The descriptions of the variables are listed in Table 5.1. Out of twenty variables, ten are numerical variables and ten are categorical variables (including binary variables). The categorical variables are followed by their levels. For example, policyholders need to specify the primary use of their vehicle when purchasing or renewing their policy: either for private use or for commercial use. Thus there are two levels for variable CAR_USE: Commercial and Private. All the numerical variables such as BLUEBOOK (the value of the vehicle, in thousands) and MVR_PTS (Motor Vehicle Record violation records) have integer values. The empirical distributions of these variables are shown in Figure 5.1. The distribution of HOME_VAL has a mass at point zero. Although not specified in the original data description, it is assumed the policyholders with zero HOME_VAL value do not own a home but, for example, renting or living with parents.

In this study, we use the total claim amount in a policy year as the response variable for the model. Because the dataset includes only the total claim amounts for the past five years,

Variable	Levels	Description
CLM_AMT5		the total claim amount in the past 5 years
KIDSDRIV		the number of driving children
TRAVTIME		the distance to work
CAR_USE	Commercial	the primary use of the vehicle
	Private	
BLUEBOOK		the value of the vehicle, in thousands
CAR_TYPE	Panel Truck	the type of the car
	Pickup	
	Sedan	
	Sports Car	
	SUV	
	Van	
RED_CAR	No	whether the color of the car is red
	Yes	
REVOKED	No	whether the driver's license was invoked in the past 7 years
	Yes	
MVR_PTS		MVR violation records
AGE		the age of the driver
HOMEKIDS		the number of children at home
YOJ		years at current job
INCOME		annual income
GENDER	No	the gender of the driver
	Yes	
MARRIED	No	married or not
	Yes	
PARENT1	No	single parent
	Yes	
JOBCLASS	Unknown	the profession of driver
	Blue Collar	
	Clerical	
	Doctor	
	Home Maker	
	Lawyer	
	Manager	
	Professional	
	Student	
MAX_EDUC	<High School	maximum education level
	High School	
	Bachelors	
	Masters	
	PhD	
HOME_VAL		the value of the insured's home
SAMEHOME		years in the current address
AREA	Rural	home/work area
	Urban	

Note. Variables without levels are numeric variables.

Table 5.1: Variable Descriptions

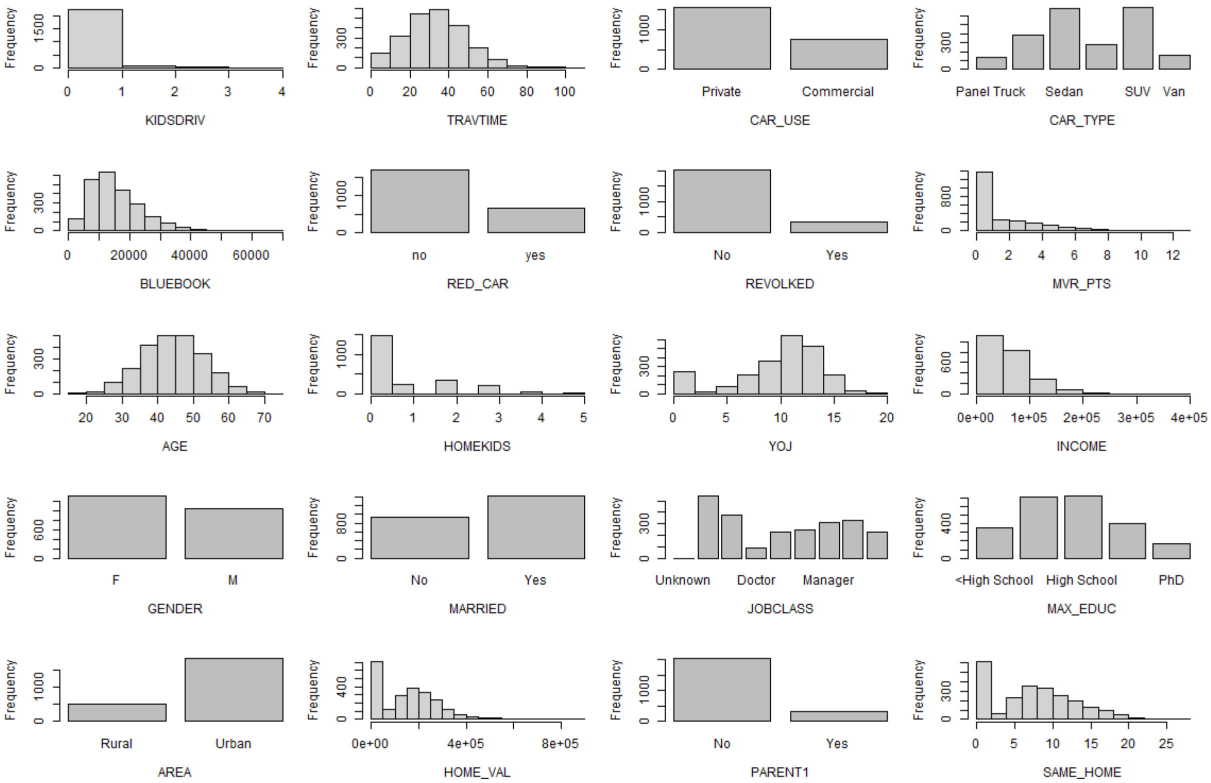

Figure 5.1: Distribution of Independent Variables

the response variable is calculated by dividing the total claim amount by five. As expected, there is a large number of policies with zero claims: out of 2,341 recorded policies, 1,599 of them have zero claim amounts. A histogram of the annual claim amount is illustrated in Figure 5.2. The distribution of the non-zero claim amounts is highly right-skewed, as shown in Figure 5.3: the majority of them concentrate between $1 and $2000, and there are also a few observations of extreme claim amounts that exceed $10,000. The summary statistics of the non-zero claim amounts are shown in Table 5.2.

Min.	1^{st} Quantile	Median	Mean	3^{rd} Quantile	Max.
100.8	758.0	1225.6	2136.2	1921.8	11407.4

Table 5.2: Claim Amount Summary

Some of the explanatory variables exhibit correlation. Figure 5.4 shows the correlation matrix of the numeric variables. The strongest correlation occurs between INCOME and HOME_VAL. This is unsurprising considering people with high income tend to purchase homes of high value. There is also a strong negative correlation between AGE and HOME-

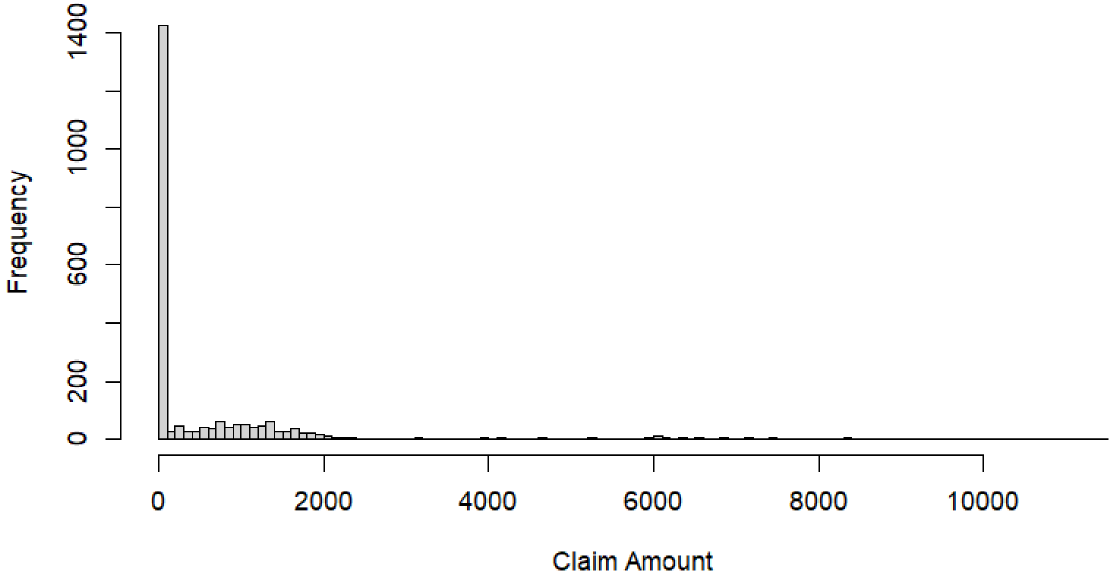

Figure 5.2: Distribution of Claim Amount

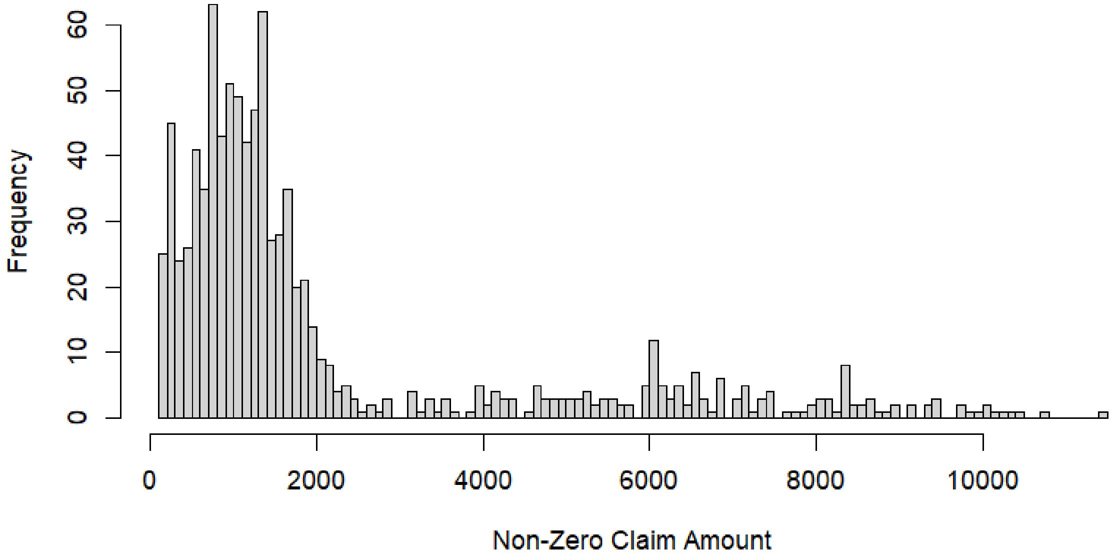

Figure 5.3: Distribution of Non-Zero Claims

KIDS. As the drivers get older, their kids move out for school or work resulting in a negative relationship between the two variables.

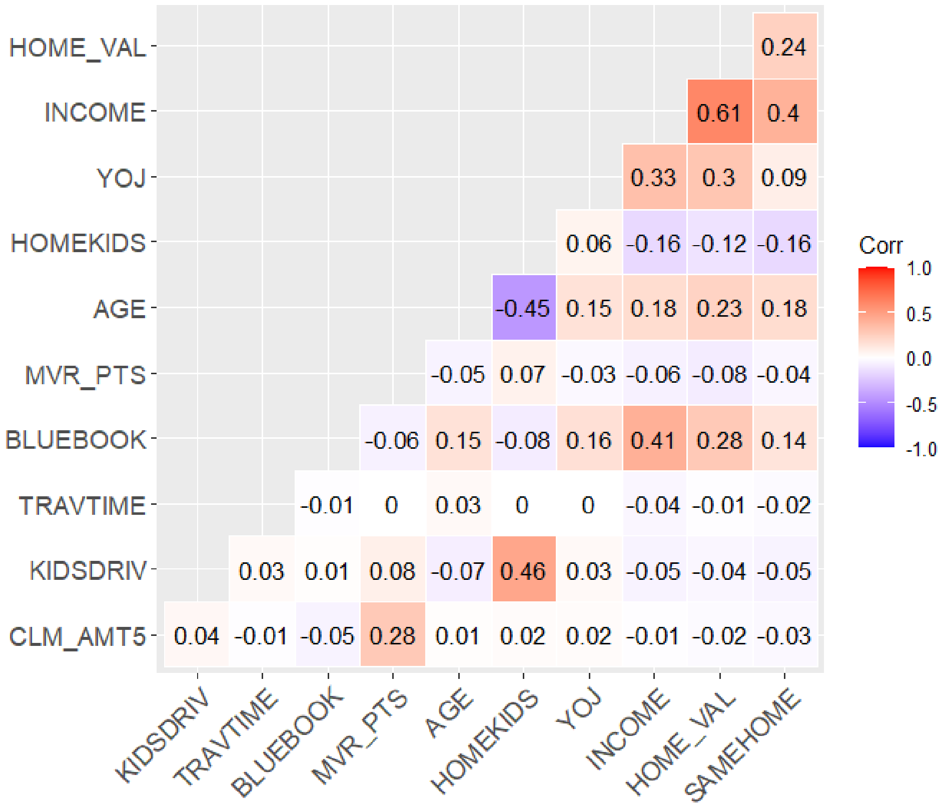

Figure 5.4: Correlation among numeric variables

5.2 Forward selection

The forward stepwise variable selection method initializes with the null model, which only includes the intercept, denoted β_0, and no other predictors.

$$\log(\mu_i) = \beta_0, \qquad i = 1, 2, \ldots, 2, 341,$$

where μ_i is the expected claim amount from the i^{th} policyholder.

The estimated β_0 is denoted $\hat{\beta}_0$, which is equal to 6.7296 when fitted to the automobile insurance dataset. The null model implies all the policyholders have the same expected pure premium equalling to $\exp\{6.7296\} = \$836.78$, which is also the empirical mean of the claim amounts in this dataset.

The process then adds one predictor to the model of the logarithm of the expected aggregate claim amount. By adding the j^{th} variable we have

$$\log(\mu_i) = \beta_0 + \beta_1 x_{ij}, \qquad i = 1, 2, \ldots, 2,341,$$

for $j = 1, 2, \ldots, 20$.

The twenty models are then compared, using the AIC criterion; the model with the variable REVOKED decreases the AIC value of the null model the most. Accordingly, REVOKED is chosen and carried forward to the next step. This is not surprising because REVOKED is the indicator variable of whether a driver's license has been revoked before. It indicates whether the driver exhibits poor judgement on the road, for example, impaired driving and reckless driving, and may carry it even after reinstatement.

The variable selection process continues by adding one more variable at a time until the AIC does not decrease anymore. In each step, the model that decreases the AIC value the most is the optimal model and the variables in the optimal model are retained. Each time a model is fitted, the power parameter for the Tweedie GLM is re-estimated using the maximum likelihood method. The selected variable from each step is shown in column 2 of Table 5.3. The AIC values from the optimal models are shown in column 3 of Table 5.3. In Figure 5.8, the red line shows the change in AIC as we add more variables to fit the model. The lowest AIC value acquired is 18402.51. The number of variables retained in the model after forward selection is eight. A heuristic method that can be applied to the model selection based on AIC values is the *elbow method*. The elbow method suggests that the selection should stop at the 'elbow point' of the AIC curve, instead of the lowest point. Although not used in this study, it is worth mentioning that using the elbow method may result in a more parsimonious model.

Number of Steps	New Variable Added	AIC
0	(INTERCEPT)	19031.14
1	REVOKED	18669.32
2	MVR_PTS	18474.40
3	AREA	18413.22
4	BLUEBOOK	18404.56
5	GENDER	18403.90
6	CAR_USE	18403.31
7	AGE	18402.59
8	MARRIED	18402.51

Table 5.3: Forward Stepwise Selection

The change in the AIC values is obvious in the first four steps; however, the changes are minimum thereafter, suggesting that the additional variables add little improvement to the goodness-of-fit of the model. After the eighth variable MARRIED is added, the model cannot be improved further by adding more variables. The forward stepwise variable selection method suggests that the best model is the one with REVOKED, MVR_PTS, AREA, BLUEBOOK, GENDER, CAR_USE, AGE and MARRIED as explanatory variables with

$$\begin{aligned}\log(\hat{\mu}_i) = \hat{\beta}_0 &+ \hat{\beta}_1 \cdot \text{REVOKED\_Yes} + \hat{\beta}_2 \cdot \text{MVR\_PTS} + \hat{\beta}_3 \cdot \text{AREA\_Urban} \\ &+ \hat{\beta}_4 \cdot \text{BLUEBOOK} + \hat{\beta}_5 \cdot \text{GENDER\_M} + \hat{\beta}_6 \cdot \text{CAR\_USE\_Commercial} \\ &+ \hat{\beta}_7 \cdot \text{AGE} + \hat{\beta}_8 \cdot \text{MARRIED\_Yes}.\end{aligned}$$

The model results are shown in Table 5.4. To find whether an individual variable is significant, we need to test the hypothesis of its coefficient equal to zero using the Wald test (Wald, 1943). The null hypothesis of the Wald test for the j^{th} variable selected is that the coefficient is equal to zero, i.e., $\text{H}_0 : \beta_j = 0$. The Wald statistics can be calculated by

$$W_j = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)},$$

where $\text{SE}(\beta_j)$ denotes the standard error of β_j.

Wald (1943) showed that when the number of observations approaches infinity, under the null hypothesis, the Wald statistic is asymptotically standard normal. Thus we can calculate the p-value using the standard normal distribution. A small p-value indicates that the variable selected is statistically significant. The p-values show that variables REVOKED, MVR_PTS, AREA, and BLUEBOOK are significant and each variable of GENDER, CAR_USE, AGE and MARRIED improves little model performance, if any. This is consistent with what we observed earlier in the AIC values. In fact, if we use the p-value > 0.05 as the stopping criterion, then the forward selection method chooses only four variables. For completeness and later comparison, the model with all the available variables is also considered. It is referred to as the full model later in this report.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.0050	0.2716	18.43	0.0000***
REVOKED_Yes	1.4884	0.0949	15.69	0.0000***
MVR_PTS	0.1888	0.0172	10.95	0.0000***
AREA_Urban	0.9360	0.1459	6.41	0.0000***
BLUEBOOK	-0.0150	0.0058	-2.60	0.0094**
GENDER_M	-0.1699	0.0939	-1.81	0.0707
CAR_USE_Commercial	0.1433	0.0979	1.46	0.1434
AGE	0.0076	0.0051	1.48	0.1389
MARRIED_Yes	-0.1054	0.0902	-1.17	0.2431
p	1.3469			
ϕ	332.8252			

Table 5.4: Model summary after forward selection

5.3 Principal component regression using NIPALS

In our data, variables CAR_USE, CAR_TYPE, RED_CAR, REVOKED, MARRIED, GENDER, AREA, PARENT1, JOBCLASS and MAX_EDUC are considered nominal categorical variables. To find the principal components, we first transform these categorical variables into dummy variables. Thirty-three indicator variables are created from the ten categorical variables. There is no data with JOB_CLASS level Unknown, so no indicator variable is created for this level. The rank of the explanatory variable matrix is 33. This can be viewed as the sum of the 23 indicator variables and 10 numerical variables in the dataset, since one level from each of the 10 categorical variables is used as a reference level. The maximum number of principal components that can be calculated is 33. All variables are standardized by subtracting their means and scaled by their corresponding standard deviations. We denote these standardized variables as x_{ij} and its standard deviation as s_j for $j = 1, 2, \ldots, J$ and $J = 33$.

Since the NIPALS is an iterative method, the algorithm initiates by finding the first principal component and then calculates the other components sequentially. The algorithm produces a loading matrix of dimension 33×33 and a score matrix of dimension $2,341 \times 33$ (i.e., $I = 2,341,\ J = M = 33$).

The scores t_{im} can be written as a linear combination of the variables with loadings p_{jm} as the weights of the variables; that is,

$$\begin{aligned} t_{i1} &= p_{11}x_{i1} + p_{21}x_{i2} + \ldots + p_{33,1}x_{i33}, \\ &\cdots\cdots \\ t_{i33} &= p_{1,33}x_{i1} + p_{2,33}x_{i2} + \ldots + p_{33,33}x_{i33}, \end{aligned}$$

for $i = 1, 2, \ldots, 2, 341$.

Table 5.5 reports the loadings (i.e., $p_{11}, \ldots, p_{33,1}$) from the principal component analysis for the first component t_{i1}. The interpretation of these loadings for the dummy variables is that a change of dummy variable x_{ij} from 0 to 1 adds the score by p_{j1}/s_j. For the numeric variables, an increase of 1 unit adds the score by p_{j1}/s_j.

A glimpse of the first three components is shown in Table 5.6. These principal components can be seen as a series of arbitrary numbers created to represent the original explanatory variables. It is worth noting that the signs of all the loadings and principal components can be changed simultaneously without losing their authenticity. This is because if we multiply -1 to both $\boldsymbol{T}$ and $\boldsymbol{P}$ in $\boldsymbol{X} = \boldsymbol{TP}'$, the relationship still holds.

The principal component regression can be built with the principal component scores calculated using the NIPALS method. The generalized linear model with principal compo-

Variable	Loading	Variable	Loading
KIDSDRIV	0.065	MARRIED_No	0.088
TRAVTIME	0.026	MARRIED_Yes	-0.088
GENDER_F	0.343	GENDER_M	-0.343
BLUEBOOK	-0.161	AREA_Rural	0.101
MVR_PTS	0.034	AREA_Urban	-0.101
AGE	-0.156	PARENT1_No	-0.170
HOMEKIDS	0.150	PARENT1_Yes	0.170
YOJ	-0.138	JOBCLASS_Blue Collar	-0.034
INCOME	-0.236	JOBCLASS_Clerical	0.095
HOME_VAL	-0.212	JOBCLASS_Doctor	-0.069
SAMEHOME	-0.132	JOBCLASS_Home Maker	0.176
CAR_USE_Private	0.124	JOBCLASS_Lawyer	-0.059
CAR_USE_Commercial	-0.124	JOBCLASS_Manager	-0.112
CAR_TYPE_Panel Truck	-0.156	JOBCLASS_Professional	-0.074
CAR_TYPE_Pickup	-0.058	JOBCLASS_Student	0.078
CAR_TYPE_Sedan	-0.140	MAX_EDUC_<High School	0.089
CAR_TYPE_Sports Car	0.134	MAX_EDUC_Bachelors	-0.056
CAR_TYPE_SUV	0.234	MAX_EDUC_High School	0.109
CAR_TYPE_Van	-0.111	MAX_EDUC_Masters	-0.094
RED_CAR_No	0.299	MAX_EDUC_PhD	-0.080
RED_CAR_Yes	-0.299	REVOKED_Yes	0.030
REVOKED_No	-0.030		

Table 5.5: Loadings of the first principal components

nents can be expressed as

$$\log(\mu_i) = c_0 + c_1 t_{i1} + c_2 t_{i2} + ... + c_T t_{iT},$$

where T is the number of components included in the model. To determine the appropriate number of components T to be included, the cumulative variance explained and the cross-validation errors are calculated.

The cumulative variance explained by the principal components can be calculated using Equation (4.5). This can be done because the dataset does not contain missing values, and the variance explained calculated in the NIPALS algorithm is the same as the eigenvalues in the eigen-decomposition method. The cumulative variance explained is illustrated in Figure 5.5. The red line represents 90% mark. In the figure, the cumulative variance explained exhibits concavity. This is because, by construction, the preceding principal component always captures a larger variation than the following component. We observe that 22 principal components are needed to explain more than 90% of the variance in the independent variables.

The K-fold cross-validation (Hastie et al., 2009) is also a popular choice when choosing the number of components. The general procedure can be described as follows.

	PC1	PC2	PC3
1	-2.645	2.532	0.374
2	-2.631	0.426	1.806
3	1.790	1.311	3.061
4	2.265	-0.496	0.496
5	0.633	-3.041	1.327
6	2.955	-0.906	-2.083
7	-2.149	0.705	1.044
8	0.554	0.358	-2.420
9	2.436	-0.383	-1.725
10	-0.694	-0.183	-1.004
⋮	⋮	⋮	⋮
2341	1.310	2.418	-0.726

Table 5.6: A glimpse of the first three components calculated using NIPALS algorithm

(1) Randomly divide dataset into K roughly equal sized groups.

(2) Fit the model using $K - 1$ groups and calculate the root mean square error of the fitted model predicting the remaining part of the data.

(3) Repeat (2) K times so that every group is used once as a test group.

(4) Obtain the cross-validation error by averaging the root mean square error across all groups.

The root mean square error is calculated as the square root of the sum of the squared difference between the actual response value and the predicted value; that is,

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}.$$

In this report $K = 10$ is used. The cross-validation shows a similar result as the scree plot, as shown in Figure 5.6. The lowest RMSE occurs when including 20 principal components in the model, compared to 22 components chosen using the scree plot (see Figure 5.5). For comparison purposes, the RMSE for the full model is indicated by the red line in Figure 5.6. The full model is the model that includes all twenty variables. The graph shows only when 20 components or 25 components are included, the PCR is better than the full model; otherwise the full model has a smaller RMSE.

Using 22 principal components, the result of the principal component regression is shown in Table 5.7. The interpretation of these coefficients can be baffling. It is one of the disadvantages of PCR. One way to understand these coefficients is to combine them with the loading matrix, as shown in Equation (4.10), and then interpret the resulting coefficient in terms of the original variables.

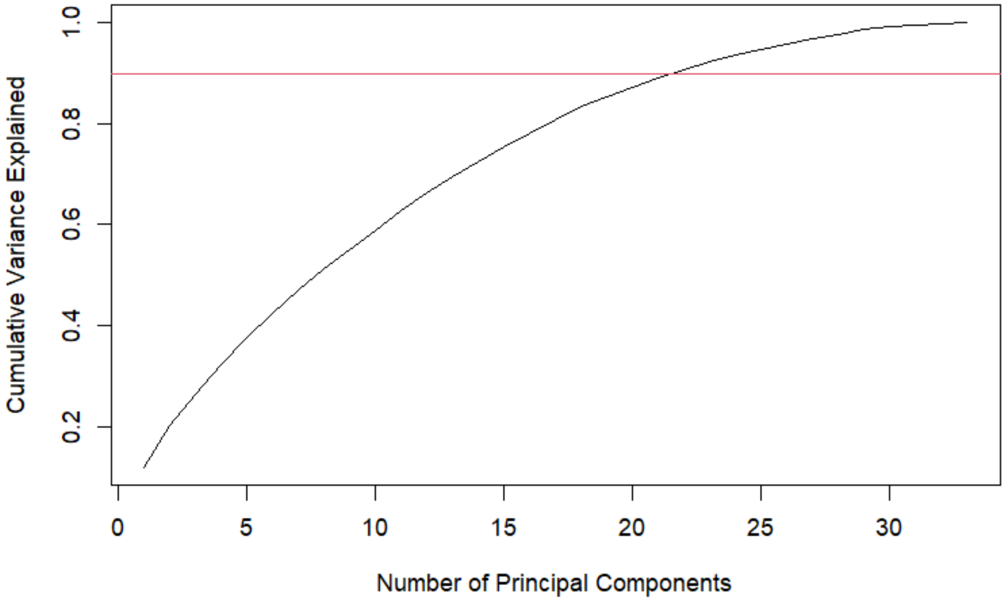

Figure 5.5: Cumulative variance explained by components

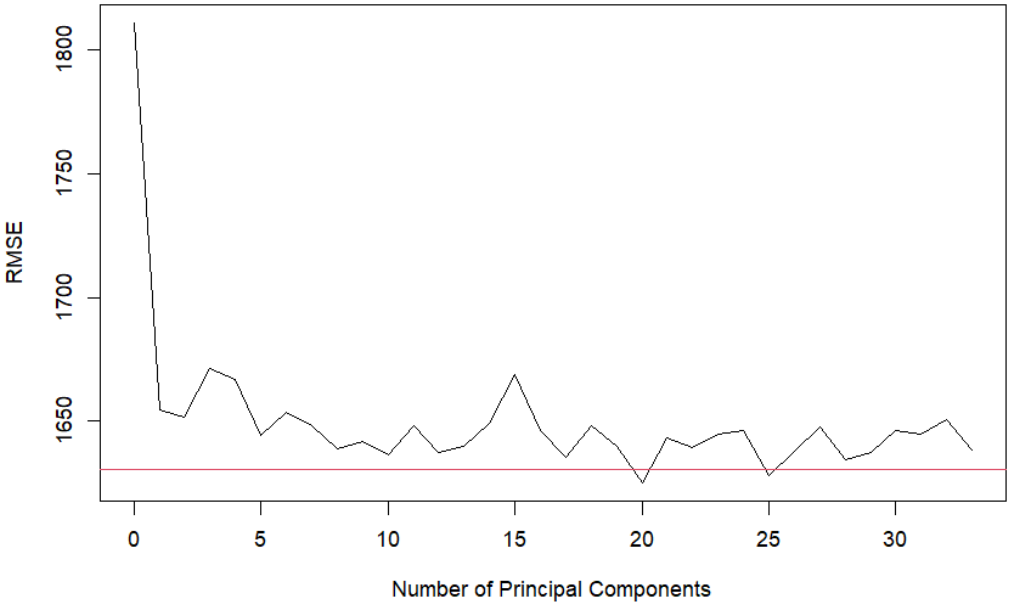

Figure 5.6: Cross-validation result

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.2936	0.0504	124.98	0.0000***
PC1	0.0238	0.0207	1.15	0.2503
PC2	-0.0037	0.0215	-0.17	0.8618
PC3	0.1388	0.0241	5.76	0.0000***
PC4	0.1165	0.0274	4.26	0.0000***
PC5	-0.4896	0.0303	-16.14	0.0000***
PC6	-0.0685	0.0304	-2.26	0.0241*
PC7	-0.0353	0.0328	-1.08	0.2819
PC8	0.0217	0.0327	0.66	0.5066
PC9	0.0431	0.0342	1.26	0.2088
PC10	-0.0698	0.0351	-1.99	0.0471*
PC11	-0.0005	0.0375	-0.01	0.9902
PC12	-0.0674	0.0398	-1.70	0.0901
PC13	0.0596	0.0376	1.59	0.1129
PC14	0.0712	0.0384	1.86	0.0637
PC15	0.0124	0.0401	0.31	0.7566
PC16	0.1309	0.0418	3.13	0.0018**
PC17	0.0444	0.0447	0.99	0.3206
PC18	0.2321	0.0435	5.33	0.0000***
PC19	-0.0029	0.0450	-0.06	0.9488
PC20	0.1242	0.0439	2.83	0.0047**
PC21	0.0137	0.0447	0.31	0.7597
PC22	-0.1335	0.0465	-2.87	0.0041**
p	1.3469			
ϕ	320.1748			

Table 5.7: Estimates for PCR

5.4 Partial least squares regression

The processing of the original data matrix is the same as what was done in the NIPALS section. The loading vector $\boldsymbol{w}_1$ for the first partial least squares component is calculated using $\boldsymbol{w}_1 = \boldsymbol{a}_1/\|\boldsymbol{a}_1\|$, where $\boldsymbol{a}_1 = (a_{11}, a_{12}...a_{1J})'$. The regression coefficient a_{1j} is calculated in the Tweedie GLM of claim amounts on each predictor x_j, for $j = 1, 2, \ldots, J$. The result of $\boldsymbol{w}_1$ is shown in Table 5.8. A large absolute value of loading indicates that the (standardized) predictor has a large contribution in calculating the first component. From Table 5.8, the largest three contributors of the first component is REVOKED, MVR_PTS and AREA. These three variables are also the most significant variables suggested in the forward selection regression.

The first component $\boldsymbol{t}_1$ can then be calculated using $\boldsymbol{t}_1 = \boldsymbol{X}\boldsymbol{w}_1/\boldsymbol{w}_1'\boldsymbol{w}_1$. The first ten rows of the first component are shown as column "PLSC1" in Table 5.9.

Variable	Loading	Variable	Loading
KIDSDRIV	0.059	MARRIED_No	0.048
TRAVTIME	-0.017	MARRIED_Yes	-0.048
GENDER_F	0.050	GENDER_M	-0.050
BLUEBOOK	-0.090	AREA_Rural	-0.396
MVR_PTS	0.407	AREA_Urban	0.396
AGE	0.017	PARENT1_No	-0.050
HOMEKIDS	0.032	PARENT1_Yes	0.050
YOJ	0.034	JOBCLASS_Blue Collar	0.070
INCOME	-0.019	JOBCLASS_Clerical	-0.016
HOME_VAL	-0.041	JOBCLASS_Doctor	-0.030
SAMEHOME	-0.050	JOBCLASS_Home Maker	-0.046
CAR_USE_Private	-0.042	JOBCLASS_Lawyer	-0.067
CAR_USE_Commercial	0.042	JOBCLASS_Manager	0.037
CAR_TYPE_Panel Truck	-0.063	JOBCLASS_Professional	-0.026
CAR_TYPE_Pickup	0.027	JOBCLASS_Student	0.023
CAR_TYPE_Sedan	-0.033	MAX_EDUC_<High School	0.032
CAR_TYPE_Sports Car	0.100	MAX_EDUC_Bachelors	-0.003
CAR_TYPE_SUV	-0.009	MAX_EDUC_High School	0.024
CAR_TYPE_Van	-0.070	MAX_EDUC_Masters	-0.060
RED_CAR_No	0.041	MAX_EDUC_PhD	-0.001
RED_CAR_Yes	-0.041	REVOKED_Yes	0.468
REVOKED_No	-0.468		

Table 5.8: Loadings for the first Partial Least Sqaures component

For the second PLS component, the regression coefficients a_{2j} is calculated using

$$\boldsymbol{y} = c_1\boldsymbol{t}_1 + a_{2j}\boldsymbol{x}_{1j},$$

where $\boldsymbol{x}_{1j}$ is the residual from the regressions of each $\boldsymbol{x}_j$ on $\boldsymbol{t}_1$ to ensure the orthogonality of $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$, satisfying

$$\boldsymbol{x}_{1j} = \boldsymbol{x}_j - c_{1j}\boldsymbol{t}_1.$$

The second loading vector equals to $\boldsymbol{w}_2 = \boldsymbol{a}_2/\|\boldsymbol{a}_2\|$ and then the second PLS component $\boldsymbol{t}_2 = \boldsymbol{X}\boldsymbol{w}_2/\boldsymbol{w}_2'\boldsymbol{w}_2$ is calculated. The process continues by including the score vector in the GLM and calculating the next loading vector by using the residual matrix from the preceding step until all the components are found. The first ten rows of the first three PLS components are shown in Table 5.9.

To determine the number of components to be included in the PLS regression model, we use AIC to compare the goodness-of-fit of the Tweedie GLM when different numbers of components are included. In Figure 5.8, the green line shows the AIC values of the PLS regression using different numbers of partial least square components. The AIC is minimized when two components are included in the regression model. The cross-validation error is

	PLSC1	PLSC2	PLSC3
1	-0.189	0.407	-0.973
2	-0.365	1.129	-0.073
3	0.792	-2.740	-0.373
4	-0.099	-1.589	-0.541
5	-0.187	0.009	-0.412
6	0.188	-0.314	0.499
7	0.416	2.572	1.838
8	0.515	0.210	-0.691
9	-0.223	-1.393	-1.070
10	-0.468	0.484	-0.853
⋮	⋮	⋮	⋮
2341	-1.950	-1.010	1.504

Table 5.9: A glance of the first three components calculated using PLS algorithm

shown in Figure 5.7. The RMSE is minimized when two PLS components are included in the regression, which gives us the same conclusion as using the AIC. The RMSE of the full model is indicated by the red line in the figure, we observe that the PLS regression is almost always better than the full model. The result of the regression model that includes two PLS

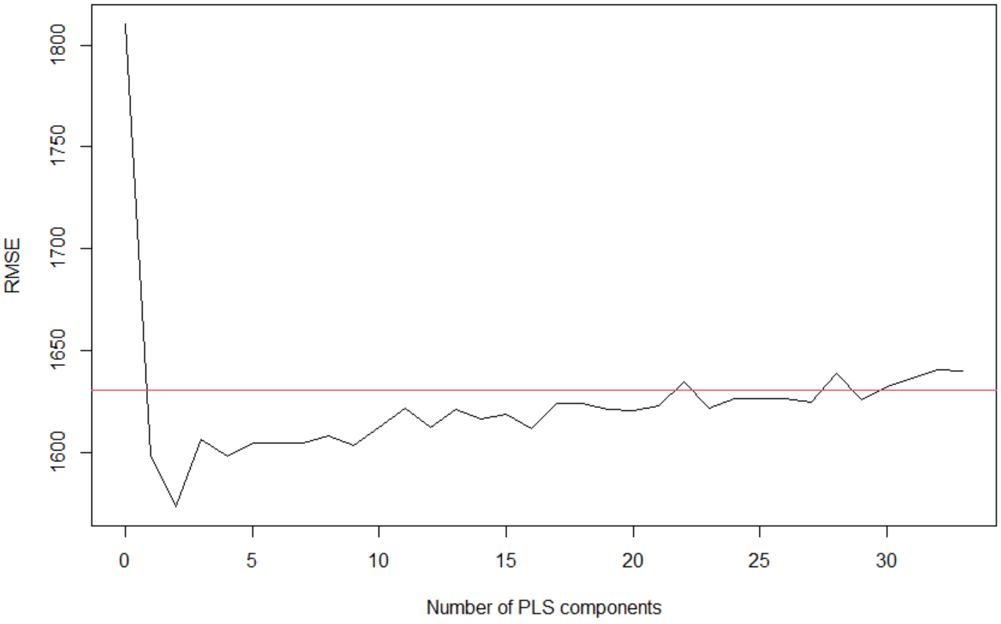

Figure 5.7: Cross-validation for PLS regression

components is shown in Table 5.10. The corresponding model is given by

$$\log(\boldsymbol{\mu}) = c_0 + c_1\boldsymbol{t_1} + c_2\boldsymbol{t_2},$$

where the estimation of c_0, c_1, c_2 is listed in the second column of Table 5.10. The result shows that both components are significant, with p-values close to zero.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.2942	0.0513	122.63	0.0000***
PLSC1	0.5810	0.0265	21.94	0.0000***
PLSC2	0.1411	0.0289	4.88	0.0000***
p	1.3469			
ϕ	334.8544			

Table 5.10: Model summary from PLS regression model

5.5 Model comparison

This section compares the three dimension reduction methods. The AIC values shown in Figure 5.8 correspond to regression models with different numbers of predictors using the three dimension reduction methods. The largest possible number of predictors is 33, which is the sum of 10 numeric variables and 23 levels of categorical variables, where each level is represented by a dummy variable (1 level from each categorical variable is used as a reference level, and thus not included here).

We observe that the PLS regression model with two components has the smallest AIC value among all models. With a given number of predictors, PLS regression always has a smaller AIC value than that from GLM with forward selection and PCR in this application. Table 5.11 summarises the RMSEs from 10-fold cross-validation of the regression model using eight variables chosen by the forward stepwise selection, the PCR with 22 principal components and the PLS regression with two partial least square components. The PLS regression has the smallest cross-validation error and is also the model with the smallest number of predictors.

The PCR performs poorly compared to the other two methods. Given the number of predictors, the PCR has the worst goodness-of-fit based on AIC. The principal component analysis concludes 22 components are needed to build the optimal PCR model, which is the largest number of predictors needed among all three dimension reduction methods. Even with the optimal PCR model, the cross-validation error is the largest compared to that for GLM after forward selection and PLS regression using two PLS components. The cross-validation error of the full model is also listed in Table 5.11 for reference. We can see that the model with the smallest RMSE is PLS regression, followed by regression with forward selection, the full model and then the PCR. This result shows that the PCR does not improve model performance from the full model. Some critiques (see, for example, Kolenikov and Angeles, 2004) are against the use of dummy variables in PCA. When the data contains categorical variables, instead of making dummy variables for each level of these categorical variables, there are different ways that are considered to be better approaches, for example, functional PCA (Segovia-Gonzalez et al., 2009), nested neural networks (Schelldorfer and Wuthrich, 2019), and PCA with embedded categorical variables (Jeong, 2022).

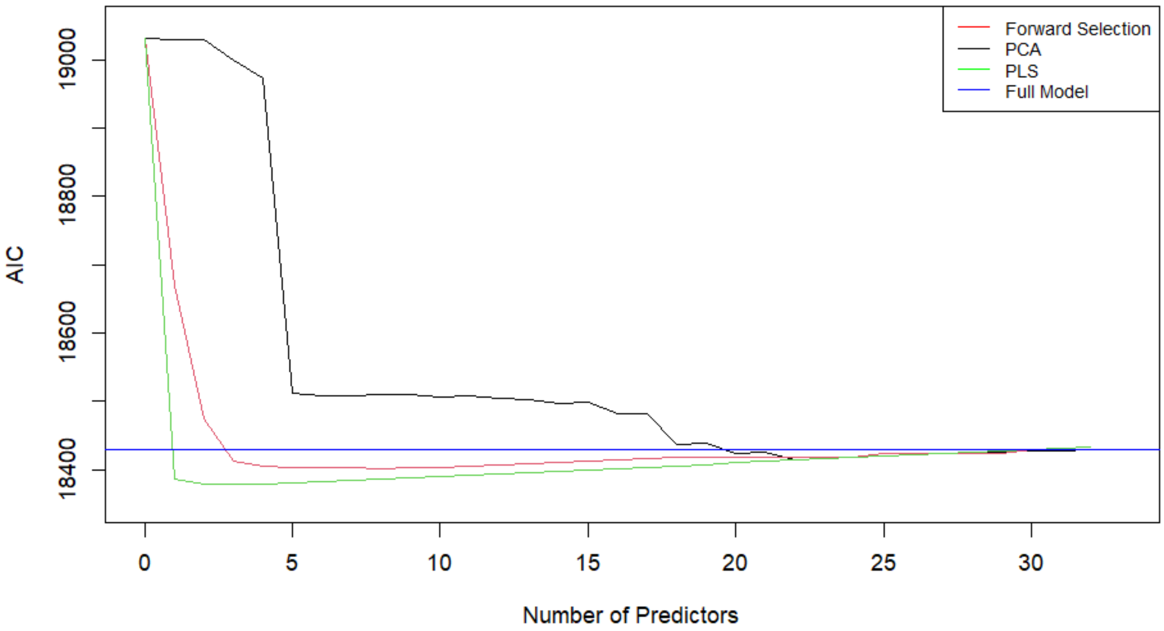

Figure 5.8: AIC comparison among four different models

	Full Model	Forward Selection	PCR	PLSR
Cross-Validation Error	1630.67	1614.86	1637.89	1568.29

Table 5.11: Cross-Validation Error

Chapter 6

Actuarial implication

Actuaries aim to balance the premium collected to cover the aggregate loss (claim amounts), operating cost and cost of capital. Thus, the estimation of all future claims helps in determining the price of insurance products and is also a crucial part of maintaining the solvency of insurance companies. The actuarial ratemaking principle is based on the cost-based pricing of individual risks (Denuit et al., 2007). The observable variables used for classification are known as a priori variables. Traditionally, these variables are information collected when policies are purchased or renewed. Some examples of a priori variables include years of driving experience, location of residence, and age of the driver. The pure premium that is estimated based on the data collected for these classified groups plays a significant part in the pricing of insurance products.

The ratemaking in property and casualty insurance is based on claim frequency distribution and loss distribution. Under the independence assumption, the pure premium is the product of average claim frequency and average loss severity. When modelling discrete count data, the Poisson distribution plays a prominent role when the underlying population is homogeneous. When this assumption is not reasonable, we can divide the population into finite homogeneous sub-populations, In this case, a mixture of Poisson distributions can be useful in modelling claim frequency (Denuit et al., 2007). The modelling of claim costs is more complicated in the real world. One reason is that in most cases, the cost of an accident cannot be determined by the policyholder. The care exercised by drivers mostly influences the number of accidents but not the costs of accidents (Denuit et al., 2007). Thus, in most statistical modelling, the observable variables are much less relevant in predicting the severity. Nevertheless, different GLM models such as gamma, inverse Gaussian and lognormal are used by actuaries to model the claim sizes, and hence the unobservable risk characteristics, such as the aggressiveness behind the wheel, can be captured through the observed response variables based on drivers' experience.

This report compares the prediction of the pure premium using Tweedie GLM with the forward stepwise selection, NIPALS and PLS. While Chapter 5 shows in detail how these GLMs are estimated, this chapter focuses on the comparison of ratemaking using these estimated GLMs. The ratemaking process in automobile insurance involves classifying policies according to their risk characteristics. Thus, we divide the data presented in Chapter 5 into subgroups according to their characteristics and estimate the pure premium for each subgroup.

6.1 Forward selection

The pure premium estimated using the GLM with forward selection utilizes eight variables in total. The variables are listed in Table 5.3 and the GLM estimation results are shown in Table 5.4. We can produce a tariff table based on each risk group with different risk characteristics. To limit the number of groups created, the numeric variables BLUEBOOK (the value of the vehicle) and AGE (the age of the driver) are binned. The the risk classes for BLUEBOOK are manually set as $0-$14,444, $14,445-$27,300, $27,301-$40,200, $40,201-$53,100, and $53,101-$66,000 (the maximum value of BLUEBOOK in the dataset is $65,970). When the value of BLUEBOOK falls into an interval, the middle point of the interval is used to estimate the pure premium. For example, if the true value of BLUEBOOK is $12,080, this belongs to the risk group $0-$14,444, and the pure premium is calculated using $7,222 (7.222 thousand). Similarly, AGE is grouped into five risk classes: 18-28, 29-39, 40-50, 51-61, and 62-73 (the oldest driver in the dataset is 73 years old). After binning BLUEBOOK and AGE, there are 11,200 risk groups in total. Table 6.1 shows nine of these risk groups. The first group is the group with the most exposures, so is used as a reference group in the discussion. The estimated pure premium using the GLM with forward selection is shown in column 4 of Table 6.2; the relative changes compared to the reference group are shown in column 5.

GROUP	REVOKED	MVR_PTS	AREA	BLUEBOOK	GENDER	CAR_USE	AGE	MARRIED
1	No	0	Urban	0-14444	F	Private	51-61	Yes
2	No	0	Urban	0-14444	F	Commercial	51-61	Yes
3	Yes	0	Urban	0-14444	F	Private	51-61	Yes
4	No	0	Rural	0-14444	F	Private	51-61	Yes
5	No	0	Urban	14445-27300	F	Private	51-61	Yes
6	No	0	Urban	0-14444	M	Private	51-61	Yes
7	No	0	Urban	0-14444	F	Private	62-73	Yes
8	No	0	Urban	0-14444	F	Private	51-61	No
9	No	5	Urban	0-14444	F	Private	51-61	Yes

Table 6.1: Risk Groups

The largest pure premium estimation among all nine rating groups appears when REVOKED is Yes. A driver's license is revoked when one or more serious offence(s) is committed by the driver; examples include impaired driving, or when the driver has too many traffic points. It is possible for drivers who had revoked driver's licenses to apply for new ones. In these cases, drivers with previously revoked driver's licenses are considered high risk. By comparing the pure premium estimation between group 1 and group 3, we observe when holding other factors fixed as risk group 1, a policyholder with a previously revoked driver's license has an estimated pure premium that is \$1916.84 − \$432.69 = \$1484.15 more than a policyholder without a previously revoked driver's license. Another risk factor that causes an increased pure premium estimation is MVR_PTS. Similar to REVOKED, MVR_PTS is a good indicator of drivers' riskiness and driving habits. As shown in Table 6.2, holding other factors fixed, drivers with 5 violation points have a pure premium estimation that is 157% more than drivers with 0 violation points.

The location of the driver's residency also has a considerable impact on the pure premium estimation. The smallest pure premium estimation among these nine groups occurs in group 4. In risk group 4, the location of residency of the driver is a rural area. The pure premium of risk group 4 is \$432.69 − \$169.70 = \$262.99 less compared to the reference group. Because of the low population density in rural areas, it is less likely for drivers to encounter an accident, and when they do, the costs of accidents are likely to be lower compared to accidents that happened in urban areas. Consequently, the pure premium estimation is lower compared to drivers from urban areas, holding other factors fixed.

If the car is used for commercial purposes (e.g., trucking, Uber), the pure premium estimation is 15.4% higher than if used for private purposes. This is not surprising as commercial cars are on the road more often than private cars.

Our results also show that younger drivers have a lower pure premium estimate. A driver who is between 62-73 years old has a pure premium estimate that is 18.2% higher than a driver who is between 51-61 years old. In addition, the younger the driver is, the lower the pure premium estimates. This is arguable to some degree. Intuitively, drivers tend to cause fewer accidents when they have certain years of driving experience and have good reflex speed. A quadratic relationship between age and claim amounts may fit better in this case. Our results also show that the value of BLUEBOOK is negatively related to pure premium estimation. This is counter-intuitive as more expensive cars are expected to have higher repair costs. It can be caused by the limited number of data records.

6.2 Principal component regression

In this section, the pre-processed predictors from the risk groups and the loadings calculated in Section 5.3 are used to find the updated score matrix. Then the updated score matrix is used to predict the pure premium using the PCR with 22 components discussed in Section 5.3. The summary of the fitted model is shown in Table 5.7.

Since the PCR method utilizes all the variables, the number of risk groups is extremely large. For comparison purposes, the pure premiums of the same nine risk groups are calculated and the results are shown in column 6 of Table 6.2; the relative changes compared to the reference group is shown in column 7. The predictors that are not listed in Table 6.1 are held fixed during the calculation. The fixed variables take the value or the level with the highest frequency. From the result, the pure premium estimated using PCR is higher than the one estimated using the forward selection method across all nine groups. This is mainly caused by the value of the fixed variables used in the PCR, which are not selected using forward selection. When different values of these variables are tried, the pure premium estimates using PCR can be lower or higher than the pure premium estimates using GLM with forward selection. The patterns of the change in pure premium estimation relative to the reference group discussed in Section 6.1 can also be observed here: the signs of the relative changes in column 7 are the same as in column 5, just with different magnitudes. Overall, the relative changes from the reference group are smaller compared to that from Section 6.1.

6.3 Partial least squares

When calculating pure premium using PLS regression, the setup is the same as the PCR. After pre-processing the risk characteristics and multiplying by the loading matrix, the updated score matrix is used to predict the pure premium using the PLSR with two components discussed in Section 5.4. The summary of the fitted model is shown in Table 5.10. The pure premium estimates using PLS regression are shown in column 8 of Table 6.2; the relative changes from the reference group are shown in column 9. Like the other two dimension reduction methods, group 4 (driver from rural area) has the lowest pure premium estimates and group 3 (driver with previously revoked driver's license) has the highest pure premium estimates.

An illustration of the pure premium estimated using three methods is shown in Figure 6.1. Compared to the forward selection method and PCR, PLSR "punishes" the drivers with revoked driver's licences and violation points less (see the bar graph of group 3 and group 9) but charges a little more to other groups (except group 4); therefore, the results in general balance out.

For comparison purposes, the pure premium estimations using all the explanatory variables are shown in column 2 and the relative changes are shown in column 3 in Table 6.2. The relative changes from the reference group mostly are consistent with the findings earlier; however, group 2 receives a lower pure premium estimation than that for group 1 under the full model, which contradicts the other regression models and our intuition. Generally, all three methods are able to capture the riskiness of drivers from their characteristics and use them to estimate the pure premium; however, the PLSR requires the least predictors when fitting the GLM.

GROUP	Full Model		Forward Selection		PCR		PLS	
1	474.46		432.69		474.44		514.09	
2	470.62	-0.8%	499.37	15.4%	517.26	9.0%	561.24	9.2%
3	2092.04	440.93%	1916.84	343.0%	2076.05	337.6%	1962.98	281.8%
4	184.54	-61.11%	169.70	-60.8%	186.29	-60.7%	173.71	-66.2%
5	431.11	-9.2%	393.02	-9.2%	448.12	-5.5%	488.27	-5.0%
6	385.40	-18.77%	365.09	-15.6%	449.51	-5.3%	488.61	-5.0%
7	550.65	16.06%	511.54	18.2%	556.21	17.2%	580.10	12.8%
8	560.21	18.07%	480.76	11.1%	508.57	7.2%	538.87	4.8%
9	1216.87	156.47%	1111.86	157.0%	1199.59	152.8%	1084.76	111.0%

Table 6.2: Pure premium estimation using four methods

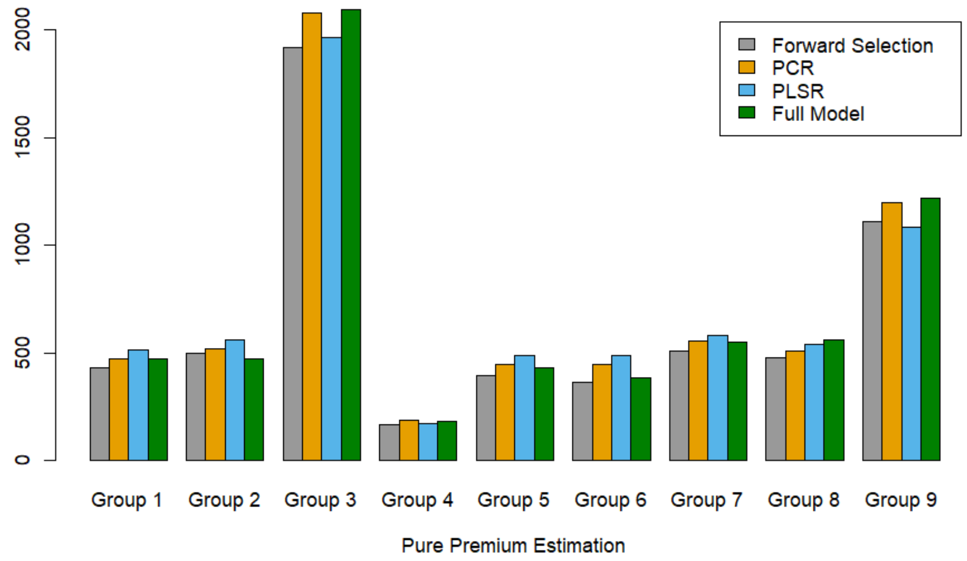

Figure 6.1: Pure premium estimation using four methods

Chapter 7

Conclusion

The booming technological innovation brings both opportunities and challenges. Traditionally in automobile insurance, the observable risk characteristics for modelling are limited to the vehicular information such as the make and use of the car, and drivers' personal information such as the location of residency and occupation. With the development of telematics, now information such as speed, number of lane changes, and frequency of emergency brakes are also available. It is challenging to use all the observable risk characteristics for a few reasons: over-fitting, computational limitation, multicollinearity and loss of interpretability. With a large amount of information available, it becomes a pressing issue for actuaries in the property and casualty industry to utilize the data effectively and efficiently. For example, recently Jeong (2022) discussed dimension reduction techniques specifically for telematics data.

In this report, three dimension reduction methods are discussed in detail: the forward stepwise selection, principal component analysis using nonlinear iterative partial least squares algorithm and partial least squares algorithm. The forward selection and the partial least squares method are supervised learning methods, whereas the NIPALS is an unsupervised method. To evaluate these dimension reduction methods, we fit a generalized linear model for aggregate claim amounts using the reduced variables under each method and compare the model performance.

By using an application of an automobile dataset, we observe that the partial least squares regression model performs better than the other two methods in terms of both model accuracy and the number of variables reduced. Compared to unsupervised methods such as principal component analysis, the partial least squares algorithm incorporates the response variable, and thus provides better performance when regression is the objective. The forward selection has the advantage of interpretability over the methods using latent variables. It is easier to interpret regression parameters using forward selection to policyholders who may not have been exposed to statistical modelling methods. The claim amounts also depend on unobservable characteristics such as drinking behaviour and the reflex speed of the driver. Thus when using unsupervised methods, these unobservable characteristics are completely omitted as the response variable is not used to generate the components. Overall, principal component analysis is not highly recommended for property and casualty pricing modelling, because it provides neither good model performance nor easy interpretability based on our study on the automobile dataset. In future studies, feature selection methods can be applied before PCA to remedy the disadvantages of PCA.

Quijano Xacur and Garrido (2015) has shown that the Tweedie GLM and the separate frequency and severity GLMs perform equally well; however, the Tweedie GLM is a simpler model, and thus should be preferred when possible. In situations when actuaries are interested in either claim frequency or severity, separate models should be applied. In real applications, it is often the case that there exists a very large proportion of zero claims, so zero-inflated models can be a better choice for modelling these situations. Zhou et al. (2022) presented a boosting-assisted zero-inflated Tweedie model for extremely unbalanced zero-inflated data. Further research could be done on applying dimension reduction methods for zero-inflated parametric Tweedie models.

This study investigates parametric dimension reduction methods. However, some literature has proposed interesting non-parametric methods such as gradient boosting trees for automobile insurance loss cost modelling (see, for example, Guelman, 2012 and Yang et al., 2018). Further research could be done on comparing the modelling performance of parametric dimension reduction methods and non-parametric ones. In application, another issue worth discussing is the presence of missing data. In real scenarios, it is unreasonable to assume that the data collected from policyholders are complete. Thus, when studying dimension reduction methods, further studies can be done on how the loss cost regression models with reduced variables perform in the presence of missing data.

Bibliography

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Bastien, P., Vinzi, V. E., and Tenenhaus, M. (2005). PLS generalised linear regression. *Computational Statistics & data analysis*, 48(1):17–46.

Chao, G., Luo, Y., and Ding, W. (2019). Recent advances in supervised dimension reduction: A survey. *Machine learning and knowledge extraction*, 1(1):341–358.

Cohen, A. (1991). Dummy variables in stepwise regression. *The American Statistician*, 45(3):226–228.

Denuit, M., Maréchal, X., Pitrebois, S., and Walhin, J.-F. (2007). *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems.* John Wiley & Sons.

Dunn, P. K. and Smyth, G. K. (2018). *Generalized linear models with examples in R*, volume 53. Springer.

Filmer, D. and Pritchett, L. H. (2001). Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of India. *Demography*, 38(1):115–132.

Fox, J. (2019). *Regression diagnostics: An introduction.* Sage publications.

Frees, E. W. (2009). *Regression modeling with actuarial and financial applications.* Cambridge University Press.

Gareth, J., Daniela, W., Trevor, H., and Robert, T. (2013). *An introduction to statistical learning: with applications in R.* Spinger.

Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17.

Goldburd, M., Khare, A., Tevet, D., and Guller, D. (2016). Generalized linear models for insurance rating. *Casualty Actuarial Society, CAS Monographs Series*, 5.

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, 39(3):3659–3667.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Jeong, H. (2022). Dimension reduction techniques for summarized telematics data. *The Journal of Risk Management, Forthcoming.*

Jørgensen, B. (1992). Exponential dispersion models and extensions: A review. *International Statistical Review/Revue Internationale de Statistique*, pages 5–20.

Jørgensen, B. and Paes De Souza, M. C. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 1994(1):69–93.

Keith, T. Z. (2019). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling.* Routledge.

Kolenikov, S. and Angeles, G. (2004). The use of discrete data in PCA: theory, simulations, and applications to socioeconomic indices. *Chapel Hill: Carolina Population Center, University of North Carolina*, 20:1–59.

McCullagh, P. and Nelder, J. A. (2019). *Generalized linear models.* Routledge.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. *Institute of Urban and Regional Development, University of California.*

Quijano Xacur, O. A. and Garrido, J. (2015). Generalised linear models for aggregate claims: to tweedie or not? *European Actuarial Journal*, 5(1):181–202.

Schelldorfer, J. and Wuthrich, M. V. (2019). Nesting classical actuarial models into neural networks. Available at SSRN 3320525. *https://ssrn.com/abstract=3320525.*

Segovia-Gonzalez, M., Guerrero, F., and Herranz, P. (2009). Explaining functional principal component analysis to actuarial science with an example on vehicle insurance. *Insurance: Mathematics and Economics*, 45(2):278–285.

Smyth, G. K. and Jørgensen, B. (2002). Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. *ASTIN Bulletin: The Journal of the IAA*, 32(1):143–157.

Strang, G. (2006). *Linear algebra and its applications.* Belmont, CA: Thomson, Brooks/Cole.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482.

Wold, H. O. A. (1968). *Nonlinear estimation by iterative least square procedures.*

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Wüthrich, M. V. (2003). Claims reserving using Tweedie's compound Poisson model. *ASTIN Bulletin: The Journal of the IAA*, 33(2):331–346.

Yang, Y., Qian, W., and Zou, H. (2018). Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models. *Journal of Business & Economic Statistics*, 36(3):456–470.

Yip, K. C. and Yau, K. K. (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, 36(2):153–163.

Zhou, H., Qian, W., and Yang, Y. (2022). Tweedie gradient boosting for extremely unbalanced zero-inflated data. *Communications in Statistics-Simulation and Computation*, 51(9):5507–5529.