2019

WPI

Major Qualifying Project

A Review of a CAS Monograph on Stochastic Loss Reserving Using Generalized Linear Models

Submitted by: Hui-Xin Chen, Zhentian Ren, Junhong Zhang

Project Advisors: Jon Abraham and Barry Posterro

Date: April 6, 2019

Abstract

The Casualty Actuarial Society (CAS) has a series of monograph publications related to the work of property-casualty insurance. This project is based on the monograph *Stochastic Loss Reserving Using Generalized Linear Models* by Greg Taylor and Gráinne McGuire, which discusses the application of generalized linear models (GLMs) to loss reserving, with an emphasis on the chain ladder algorithm. For this project, the team reviewed and explained the concepts presented in the monograph, supported the explanations with additional examples, and recreated the numerical examples in the monograph using the provided dataset and SAS software. Due to time constraints and considering the relevance of topics, the project concentrates on the first four chapters and part of Chapter 5 of the monograph, which include topics on the chain ladder algorithm, over-dispersed Poisson distributions, GLMs, the Mack and cross-classified models for loss reserving, prediction errors, and the bootstrap method for estimating outstanding losses.

During the project, we contacted the authors of the monograph, Greg Taylor and Gráinne McGuire. They graciously clarified some points of confusion for us and offered advice in SAS coding to enable us to reproduce some tables from their paper. In addition, we were able to alert them to some errors we found in the paper for which they were appreciative.

Contents

Introduction

The Casualty Actuarial Society (CAS) is a professional organization of actuaries specializing in property-casualty insurance. The CAS has published a series of monographs on related topics, and the basis of this project is the monograph *Stochastic Loss Reserving Using Generalized Linear Models* by Greg Taylor and Gráinne McGuire. The team reviewed and explained the monograph in an accessible approach to assist readers with less background knowledge on loss reserving or generalized linear models in understanding the original monograph. To improve accessibility, the team explained the notations and definitions from the monograph, clarified derivations of formulas, and provided additional examples. To supplement the theoretical content, the team also reproduced data analysis of numerical examples in the monograph using SAS and Excel, and provided SAS code for readers to experiment with.

The monograph starts by introducing the chain ladder algorithm for loss reserving. While the chain ladder method itself is non-stochastic, a stochastic version of the model with distribution and error prediction also exists. The original monograph concentrates on the two families of stochastic models that generate the chain ladder algorithm, the Mack model and the cross-classified model. The monograph introduces these stochastic models and their respective GLM formulations, with an emphasis on using statistical software to implement the formulations. Our project specifically concentrates on interpreting the content of the first four chapters and the bootstrapping portion from Chapter 5 of the monograph with supplement numerical examples. The topics are as follows.

In Chapter 1, aligning with the monograph, our paper introduces the chain ladder algorithm for loss reserving and the associated notations of the paper, with numerical examples of how to apply the chain ladder method to estimate future loss development. The dataset of this chapter will be used and frequently referenced in following chapters.

In Chapter 2, the monograph provides the theoretical background of the exponential dispersion family (EDF) of distributions, and the generalized linear model (GLM). In our paper, referencing the monograph, we first introduce the EDF distributions and its two sub-families, the Tweedie sub-family (sub-family of EDF) and the over-dispersed Poisson (ODP) sub-family (sub-family of the Tweedie Sub-family). The ODP distribution is a crucial assumption for the incremental loss dataset, and will be used in application of GLMs to loss reserving in later chapters. The chapter then defines GLM, and discusses the two types of covariates, categorical and continuous, and certain aspects of goodness-of-fit of a GLM, which will be discussed in more detail in Chapter 4.

In Chapter 3, our paper defines and explains the two types of stochastic model for the chain ladder algorithm, the ODP Mack Model and ODP cross-classified Model. These two models produce the same maximum likelihood estimates as the chain ladder

algorithm, with additional estimation for distributions of each estimate. Moreover, the estimators also possess certain minimum variance properties, which are summarized in three theorems. We first introduce the theoretical background of algorithms, with numerical examples showing the procedures of how to manually apply the algorithm to derive the parameters. Then the chapter explains the concept of data input and output in SAS to apply the algorithms, with numerical examples and associated coding provided for illustration.

In Chapter 4, our paper introduces the concept of prediction error, which can be decomposed into three components: parameter error, process error and model error. The chapter discusses these types of error and provides examples to explain the definitions. The first example is independent of the chain ladder algorithm to lead into the definitions, and later examples involve the chain ladder algorithm to explain the definitions within the context of the topic. This chapter also introduces mean square error of prediction and information criterion that measure the reliability of models, as well as cross validation of model fitness, which involves using a training and test set from the observations.

In Chapter 5, the original monograph introduces two types of estimation methods of prediction errors for the chain ladder algorithm and associated forecasts. The two methods are the delta method and the bootstrap method. Our paper focuses solely on the bootstrap method. The chapter explains the procedures of resampling the residuals to eventually obtain a distribution of outstanding losses, and illustrates the concept with diagrams. Bootstrapping in the context of SAS application is also discussed with a numerical example to demonstrate the idea.

After reading this paper, the readers should have a clear understanding of the chain ladder algorithm, and loss reserving using GLMs. Also, with the additional explanations of numerical examples, readers can further assess their understanding of the models by reproducing the examples with the given algorithms and data. Finally, based on the understanding of the stochastic models gained from this paper and the original monograph, readers can further research and modify the models introduced to improve accuracy and efficiency for their own purposes.

1. The Chain Ladder Algorithm

1.1. Introduction

The chain ladder algorithm (or the development method) is a technique to estimate future claims (also known as outstanding claims or outstanding losses) according to the selected age-to-age factors. This chapter explains the steps in the chain ladder algorithm using a numerical example. The data and notations introduced in this chapter will also be used throughout the paper.

In later chapters, we will also show how to use GLMs (generalized linear models) to apply the chain ladder method.

1.2. Framework and Notation

Consider the **incremental** claim observations. There are two indices that determine the position of an observation: the accident period and the development period. The periods can be weeks, months, years, etc. **Accident periods** are time periods in which accidents occurred, and **development periods** are periods in which incurred losses develop. Denote the incremental claim observations as Y_{kj}. If we arrange all past and future observations into a table, we obtain a $K \times J$ rectangle of data, where k represents the accident periods

$$k = 1, 2, \ldots, K$$

For example, in Table 1.1 we have incremental losses for different accident years *k*:

Incremental Paid Losses in Development Year 1 ($000)		
Accident Year	K (accident periods)	Losses (j=1)
1988	1	$41,821
1989	2	$48,167
1990	3	$52,058
1991	4	$57,251
1992	5	$59,213
1993	6	$59,475
1994	7	$65,607
1995	8	$56,748
1996	9	$52,212
1997	10	$43,962

Table 1.1

j represents development periods of losses, where

$$j = 1, 2, \ldots, J$$

And j denotes the columns in paid losses matrices.

Incremental Paid Losses in Development Year ($000)				
Accident Year		j=1	j=2	j=3
1988	k=1	$41,821	$34,729	$20,147
1989	k=2	$48,167	$39,495	$24,444
1990	k=3	$52,058	$47,459	$27,359

Table 1.2

For example, in Table 1.2, Y_{32} is the incremental paid loss for the second development period of accident year 1990. Note the data is incremental, meaning this is the amount of claim paid in during the year of 1992 only.

Claim observations consist of past and future observations. The **past observations** form a development trapezoid**,** which can be written as a subset

$$\mathfrak{D}_K = \{Y_{kj}: 1 \leq k \leq K \text{ and } 1 \leq j \leq min\,(J, K-k+1)\}$$

This is illustrated in Table 1.3, where $\mathfrak{D}_K$ is highlighted by yellow below

Incremental Paid Losses in Development Year ($000)				
Accident Year		1	2	3
1988	1	41,821	34,729	20,147
1989	2	48,167	39,495	
1990	3	52,058		

Table 1.3

Notice that the development trapezoid becomes a triangle when $K = J$. When $K \neq J$, the development matrix will look more like a trapezoid:

(K=3, J=4, Current time is 1992)

Incremental Paid Losses in Development Year ($000)					
Accident Year	K \ J	1	2	3	4
1988	1	41,821	34,729	20,147	15,965
1989	2	48,167	39,495	24,444	
1990	3	52,058	47,459		

Or

(K=4, J=3, Current time is 1992)

Incremental Paid Losses in Development Year ($000)				
Accident Year	K \ J	1	2	3
1988	1	41,821	34,729	20,147
1989	2	48,167	39,495	24,444
1990	3	52,058	47,459	
1991	4	57,251		

Similarly, for future losses trapezoid, which also becomes a triangle when $K = J$)

Note the past observations Y_{kj} can be of accident period from 1 to *K*, but of development period only from 1 to the main diagonal of the rectangle. This is because the diagonals refer to the calendar years, and the main diagonal represents the current calendar year, where the latest losses we observed are (see Table 1.4)

(Losses occur in 1990)

Incremental Paid Losses in Development Year ($000)				
Accident Year	K \ J	1	2	3
1988	1	41,821	34,729	20,147
1989	2	48,167	39,495	
1990	3	52,058		

Table 1.4

The **future observations**, which are unknown and to be estimated, form the complement of the above set, and can be written as:

$$\mathfrak{D}_K^c = \{Y_{kj}: 1 \leq k \leq K \text{ and } min\,(J, K-k+1) < j \leq J\}$$

$$= \{Y_{kj}: K-J+1 < k \leq K \text{ and } K-k+1 < j \leq J\}$$

Where $\mathfrak{D}_K^c$ is highlighted below:

Incremental Paid Losses in Development Year ($000)				
Accident Year	K \ J	1	2	3
1988	1	41,821	34,729	20,147
1989	2	48,167	39,495	
1990	3	52,058		

Table 1.5

The future observations Y_{kj} can be of accident period from the 1 to *K*, and of development period from the diagonal to *J*.

The set of both past and future claim observations is thus denoted

$$\mathfrak{D}_K^+ = \mathfrak{D}_K \cup \mathfrak{D}_K^c$$

Which is illustrated in Table 1.6:

Incremental Paid Losses in Development Year ($000)				
Accident Year	K \ J	1	2	3
1988	1	41,821	34,729	20,147
1989	2	48,167	39,495	
1990	3	52,058		

Table 1.6

By adding the incremental observations of the same accident period from 1 to the development period *j*, we can obtain the cumulative row sums up to development period *j*, i.e.,

$$X_{kj} = \sum_{i=1}^{j} Y_{ki}$$

(1-1)

For example, $X_{1,3} = \$41{,}821 + \$34{,}729 + \$20{,}147 = \$96{,}697$

This is also known as the **cumulative** claim observation, denoted X_{kj}.

Table 1.7 below shows the cumulative observations computed from the incremental paid loss table above:

Cumulative Paid Losses in Development Year ($000)				
Accident Year		1	2	3
1988	1	41,821	76,550	96,697
1989	2	48,167	87,662	
1990	3	52,058		

Table 1.7

The same notations for incremental losses and cumulative losses will be used throughout the paper, where

$$Y: Incremental\ Loss$$

$$X: Cumulative\ loss$$

For a fixed accident year, Year k, we can calculate the summation of the entire k-th row in $\mathfrak{D}_K$ (all past observations) as $\sum^{\mathcal{R}(k)}$, where

$$\sum^{\mathcal{R}(k)} = \sum_{j=1}^{min(J,K-k+1)}$$

Incremental Paid Losses in Development Year ($000)					
Accident Year	K \ J	1	2	3	$\mathcal{R}(k)$
1988	1	41,821	34,729	20,147	96,697
1989	2	48,167	39,495		87,662
1990	3	52,058			52,058

Table 1.8

Similarly, for a fixed j, we can calculate the column sum as

$$\sum^{\mathcal{C}(k)} = \sum_{k=1}^{K-j+1}$$

Incremental Paid Losses in Development Year ($000)				
Accident Year	K \ J	1	2	3
1988	1	41,821	34,729	20,147
1989	2	48,167	39,495	
1990	3	52,058		
Column Sum	$\mathcal{C}(j)$	142,046	74,224	20,147

Table 1.9

For Year k, denote the amount of outstanding losses as R_k, which is the summation of all future claim observations in row k , or equivalently, the ultimate loss subtracting the last known cumulative observation:

$$R_k = \sum_{j=K-k+2}^{J} Y_{kj} = X_{kJ} - X_{k,K-k+1} \quad (1\text{-}2)$$

For example, suppose we know the future observations $Y_{2,3} = 18{,}000, Y_{3,2} = 35{,}000, Y_{3,3} = 15{,}000$ (highlighted in green), then, R_k, the total outstanding losses of Year k can be found (highlighted in blue)

Incremental Paid Losses in Development Year ($000)					
Accident Year	K \ J	1	2	3	R_k
1988	1	41,821	34,729	20,147	0
1989	2	48,167	39,495	18,000	18,000
1990	3	52,058	35,000	15,000	50,000

Table 1.10

By summing the outstanding losses of all accident years, we obtain the sum of all future observations, i.e. the total outstanding losses, in $\mathfrak{D}_K^c$ as

$$R = \sum_{k=2}^{K} R_k$$

(1-3)

for $k = K - J + 2, \ldots, K$.

Note that k starts from 2 because Year 1 has completed development.

Using the previous example, we get $R = R_2 + R_3 = 18{,}000 + 50{,}000 = 68{,}000$

1.3. Data for Numerical Examples

Incremental Paid Losses in Development Year ($000)											
Accident Year	K \ J	1	2	3	4	5	6	7	8	9	10
1988	1	41,821	34,729	20,147	15,965	11,285	5,924	4,775	3,742	3,435	2,958
1989	2	48,167	39,495	24,444	18,178	10,840	7,379	5,683	4,758	3,959	
1990	3	52,058	47,459	27,359	17,916	11,448	8,846	5,869	5,391		
1991	4	57,251	49,510	27,036	20,871	14,304	10,552	7,742			
1992	5	59,213	54,129	29,566	22,484	14,114	10,000				
1993	6	59,475	52,076	26,836	22,332	14,756					
1994	7	65,607	44,648	27,062	22,655						
1995	8	56,748	39,315	26,748							
1996	9	52,212	40,030								
1997	10	43,962									

Table 1.11

Table 1.11 presents the example data set used throughout the paper. It references the database from the Meyers and Shi (2011), of the worker compensations of the New Jersey Manufacturers Group. The data set is an **incremental paid loss triangle** consisting of past incremental claim observations Y_{kj}, with accident periods of Year 1 to Year 10, and development periods of 1 year to 10 years.

However, the chain ladder algorithm uses cumulative losses instead of incremental losses. Thus we need to first transform the table into cumulative paid loss table by calculating the row sums using equation (1-1), $X_{kj} = \sum_{i=1}^{j} Y_{ki}$, and obtain Table 1.12 from the data in Table 1.11:

Cumulative Paid Losses in Development Year ($000)											
Accident Year	J	1	2	3	4	5	6	7	8	9	10
1988	1	41,821	76,550	96,697	112,662	123,947	129,871	134,646	138,388	141,823	144,781
1989	2	48,167	87,662	112,106	130,284	141,124	148,503	154,186	158,944	162,903	
1990	3	52,058	99,517	126,876	144,792	156,240	165,086	170,955	176,346		
1991	4	57,251	106,761	133,797	154,668	168,972	179,524	187,266			
1992	5	59,213	113,342	142,908	165,392	179,506	189,506				
1993	6	59,475	111,551	138,387	160,719	175,475					
1994	7	65,607	110,255	137,317	159,972						
1995	8	56,748	96,063	122,811							
1996	9	52,212	92,242								
1997	10	43,962									

Table 1.12

This is known as the **cumulative loss triangle**, which consists of cumulative paid claim observations X_{kj}.

Beside incremental or cumulative paid loss triangles, we could also have incurred loss triangles, or claim counts triangles. These datasets differ from paid loss triangles by using reported claims and numbers of claims instead of paid claims. However, these datasets are not used in this paper.

1.4. The Chain Ladder Algorithm

The chain ladder algorithm or development method is a technique to estimate future losses according to selected age-to-age factors. In the chain ladder algorithm, for each development periods from 1 to $J - 1$, an age-to-age factor is selected to estimate the growth in the cumulative loss.

Example of Chain Ladder Algorithm

We use a subset of Table 1.12 as an example (see Table 1.13).

Cumulative Paid Losses in Development Year ($000)				
Accident Year		1	2	3
1988	1	41,821	76,550	96,697
1989	2	48,167	87,662	
1990	3	52,058		

Table 1.13

We can calculate the age-to-age factors by dividing the cumulative loss of the next development period by the cumulative loss of the current development period, which generates table 1.14.

Age-to-age Factors in Development Year			
Accident Year		1	2
1988	1	$f_{11} = 76{,}550/41{,}821 \approx 1.83$	$f_{12} = 96{,}697/76{,}550 \approx 1.26$
1989	2	$f_{21} = 87{,}662/48{,}167 \approx 1.82$	

Table 1.14

The main purpose to calculate the age-to-age factors is to select or estimate one age-to-age factor for each development period. With age-to-age factors corresponding to each development period, we can then use the known losses to estimate future claim losses.

From the data in table 1.14, we can use $\hat{f}_2 = f_{12} = 1.26$ to be the age-to-age factor for the second development period, and select a number between 1.82 and 1.83 to be our f_1. Or we could use the weighted average of the two as f_1, with the claim amount being the weight, i.e.,

$$\hat{f}_1 = \frac{1.83 * 41{,}821 + 1.82 * 48{,}167}{41{,}821 + 48{,}167} \approx 1.825$$

Notice that this method can be simplified by working out the equivalent calculation

$$\hat{f}_1 = \frac{76{,}550 + 87{,}662}{41{,}821 + 48{,}167} \approx 1.825$$

Then, we can estimate future losses with our selected age-to-age factors by multiplying the cumulative paid loss and corresponding age-to-age factors (results shown in table 1.15)

Cumulative Paid Losses in Development Year ($000)				
Accident Year		1	2	3
1988	1	41,821	76,550	96,697
1989	2	48,167	87,662	$87{,}662 * 1.26 \approx 110{,}454$
1990	3	52,058	$52{,}058 * 1.825 \approx 95{,}006$	$95{,}006 * 1.26 \approx 119{,}707$

Table 1.15

To summarize, the first step in the chain ladder method is to calculate the **age-to-age factors**, which is done by dividing the next cumulative observation by the targeting cumulative observation of the same accident year.

In general, this is represented as

$$\hat{f}_{kj} = \frac{X_{k,j+1}}{X_{kj}}, k = 1, 2, \ldots, K-1; j = 1, 2, \ldots, min(J-1, K-k)$$

(1-4)

We use formula (1-4) to calculate age-to-age factors between two development periods.

To calculate the weighted average age-to-age factors between two consecutive development periods, we use the following formula

$$\hat{f}_j = \sum_{k=1}^{K-j} \omega_{kj} \hat{f}_{kj} \ , j = 1 \ldots J-1$$

(1-5)

The weights, ω_{kj}, are usually calculated using the size of corresponding cumulative losses. Note that the sum of all weights for one age-to-age factor should be 1, i.e.

$$\sum_{k=1}^{K-j} \omega_{kj} = 1$$

(1-6)

To choose weights, we use

$$\omega_{kj} = X_{kj} \Big/ \sum_{k=1}^{K-j} X_{kj}$$

(1-7)

where we divide the cumulative loss at position *k, j* by the total known cumulative losses for the development period. Combining equations (1-4), (1-5) and (1-7), we obtain the following:

$$\hat{f}_j = \sum_{k=1}^{K-j} w_{kj} \, \hat{f}_{kj}$$

$$= \sum_{k=1}^{K-j} \left(\frac{X_{kj}}{\sum_{k=1}^{K-j} X_{kj}} \right) \times \left(\frac{X_{k,j+1}}{X_{kj}} \right)$$

$$= \sum_{k=1}^{K-j} \frac{X_{k,j+1}}{\sum_{k=1}^{K-j} X_{kj}}$$

$$= \frac{\sum_{k=1}^{K-1} X_{k,j+1}}{\sum_{k=1}^{K-1} X_{kj}}$$

Note from the procedure that the calculations of weighted average age-to-age factors can be simplified into

$$\hat{f}_j = \frac{\sum_{k=1}^{K-j} X_{k,j+1}}{\sum_{k=1}^{K-j} X_{kj}}, j = 1 \dots J-1$$

(1-8)

Intuitively, this means

$$\hat{f}_j = \frac{Sum\ of\ all\ cumulative\ losses\ for\ development\ period\ j+1}{Sum\ of\ cumulative\ losses\ for\ development\ period\ j\ except\ the\ latest\ one}$$

So, when we compute $\hat{f}_j$ with weighted average method, we take out the latest observation of development period *j*, and use the remaining rectangle to generate the weighted average age-to-age factor.

After calculating and selecting the age-to-age factors, we can estimate the outstanding losses. Define the estimated cumulative value as the last known cumulative losses multiplied by corresponding age-to-age factors:

$$\hat{X}_{kj} = X_{k,K-k+1} \hat{f}_{K-k+1} \dots \hat{f}_{j-1}$$

(1-9)

In equation (1-9), we use the latest known observation of the accident year k to predict the cumulative losses for each of the next development period by multiplying the latest observed cumulative loss with the age-to-age factors corresponding to those steps.

Using the estimated cumulative losses, we can estimate the incremental losses:

$$\hat{Y}_{kj} = X_{k,K-k+1} \hat{f}_{K-k+1} \dots \hat{f}_{j-2} (\hat{f}_{j-1} - 1)$$

(1-10)

The derivation is shown in the following procedures:

$$\hat{Y}_{kj} = \hat{X}_{kj} - \hat{X}_{k,j-1}$$

$$= X_{k,K-k+1}\hat{f}_{K-k+1} \cdots \hat{f}_{j-2}\hat{f}_{j-1} - X_{k,K-k+1}\hat{f}_{K-k+1} \cdots \hat{f}_{j-2}$$

$$= X_{k,K-k+1}\hat{f}_{K-k+1} \cdots \hat{f}_{j-2}(\hat{f}_{j-1} - 1)$$

The sum of these incremental future losses are the outstanding loses, $\hat{R}_k$, which can also be calculated as

$$\hat{R}_k = \hat{X}_{kJ} - X_{k,K-k+1} = X_{k,K-k+1}(\hat{f}_{K-k+1} \cdots \hat{f}_{J-1} - 1)$$

(1-11)

Because outstanding losses is the sum of all losses that have not occurred for the accident year, we can use the predicted ultimate loss $\hat{X}_{kJ}$ minus the latest observed loss $X_{k,K-k+1}$ to get the estimated outstanding loss.

The total outstanding losses across all accident years can be calculated by summing up the outstanding losses for each accident year, i.e.

$$\hat{R} = \sum_{k=1}^{K-1} \hat{R}_k$$

(1-12)

1.5. Numerical Example

With the necessary background illustrated earlier in the chapter, we can use the chain ladder method with weighted average age-to-age factors to predict future cumulative losses based on given data.

The results are as follows:

Cumulative Paid Losses in Development Year ($000)											
Accident Year	K \ J	1	2	3	4	5	6	7	8	9	10
1988	1	41,821	76,550	96,697	112,662	123,947	129,871	134,646	138,388	141,823	144,781
1989	2	48,167	87,662	112,106	130,284	141,124	148,503	154,186	158,944	162,903	166,301
1990	3	52,058	99,517	126,876	144,792	156,240	165,086	170,955	176,346	180,731	184,501
1991	4	57,251	106,761	133,797	154,668	168,972	179,524	187,266	192,924	197,721	201,845
1992	5	59,213	113,342	142,908	165,392	179,506	189,506	196,828	202,774	207,817	212,151
1993	6	59,475	111,551	138,387	160,719	175,475	185,209	192,364	198,176	203,104	207,340
1994	7	65,607	110,255	137,317	159,972	174,108	183,766	190,866	196,632	201,522	205,725
1995	8	56,748	96,063	122,811	142,227	154,795	163,381	169,693	174,820	179,168	182,904
1996	9	52,212	92,242	116,312	134,700	146,603	154,735	160,713	165,569	169,686	173,225
1997	10	43,962	79,788	100,608	116,513	126,809	133,843	139,014	143,214	146,775	149,836

Age to age factors	f1	f2	f3	f4	f5	f6	f7	f8	f9
values	1.8149	1.2609	1.1581	1.0884	1.0555	1.0386	1.0302	1.0249	1.0209

Table 1.16

Try to work out the triangle and see if your answer matches table 1.16.

2. Stochastic Models

This chapter provides the background for Generalized Linear Models (GLMs). In GLMs, the response variables are expressed as a linear combination of the predictors. GLMs generalize linear regression, allowing for error distributions other than a normal distribution. Response variables for GLMs can have any distribution from the Exponential Dispersion Family (EDF), which include the normal distribution.

GLMs will be discussed in more detail in the next chapter. After introducing EDF, the later parts of this chapter focus instead on the family of distributions for the response variables of GLMs. Other aspects of GLMs, such as covariates and goodness-of-fit, are also discussed in this chapter.

2.1. Exponential Dispersion Family

The response variables of a GLM can take on a distribution that belong to the family of distributions called the **exponential dispersion family (EDF)**. In this section, we discuss the definition of distributions of EDF, and sub-families of EDF that relate to the topic of this paper.

2.1.1. The Exponential Dispersion Family in General

Introduced by Nelder and Wedderburn (1972), the distributions that belong to EDF must have the probability density function (pdf) of the following form:

$$ln\,\pi(y;\theta,\phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)$$

(2-1)

where

$\pi(.) =$ the actual probability density function

$y =$ an observation/predictor

$\theta =$ location parameter; also called canonical parameter

$\phi =$ dispersion parameter/scale parameter

$b(.) =$ cumulant function that determines the shape of the distribution

$exp\big(c(y,\phi)\big) =$ normalizing factor, which make $\sum f(y;\theta,\phi) = 1$

For the distribution of an EDF, we need to make the following assumptions:

1. Functions $a(.)$, $b(.)$ and $c(.)$ are continuous.
2. $b(.)$ is one-to-one and twice differentiable, with the first derivative also one-to-one.

There are many well-known distributions that are from the EDF. By selecting specific functions of $a(.)$, $b(.)$ and $c(.)$ dependent on the observations and parameters θ and ϕ, we can obtain a distribution of this family, denoted $EDF(\theta, \phi; a, b, c)$.

Table 2-1 below (also found in Table 2-1 of the monograph) contains examples of the distributions from the EDF

Distribution	$b(\theta)$	$a(\phi)$	$c(y, \phi)$
Normal	$\frac{1}{2}\theta^2$	ϕ	$-\frac{1}{2}\left[\frac{y^2}{\phi} + ln(2\pi\phi)\right]$
Poisson	$exp\,\theta$	1	$-ln\,y!$
Binomial	$ln\left(1 + e^{\theta}\right)$	n^{-1}	$ln\binom{n}{ny}$
Gamma	$-ln(-\theta)$	v^{-1}	$v\,ln(vy) - ln\,y - ln(\Gamma v)$
Inverse Gaussian	$-(-2\theta)^{-\frac{1}{2}}$	ϕ	$-\frac{1}{2}\left[ln\left(2\pi\phi y^3 + \frac{1}{\phi}y\right)\right]$

Table 2-1

Proof for the normal distribution as a member of EDF

Recall that the normal distribution has the following pdf

$$f(x;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Using the notations for EDF, for the normal distribution we select the following functions of $a(.)$, $b(.)$ and $c(.)$

$$a(\phi)=\phi$$

$$b(\theta)=\frac{1}{2}\theta^2$$

$$c(y,\phi)=-\frac{1}{2}\left[\frac{y^2}{\phi}+ln(2\pi\phi)\right]$$

If we write the pdf in EDF form (2-1) with the above functions, we obtain

$$ln\,\pi(y;\theta,\phi)=\frac{y\theta-\frac{1}{2}\theta^2}{\phi}-\frac{1}{2}\left[\frac{y^2}{\phi}+ln(2\pi\phi)\right]$$

$$=-\frac{1}{2}\left(\frac{-2y\theta+\theta^2+y^2}{\phi}\right)-\frac{1}{2}ln(2\pi\phi)$$

$$=-\frac{1}{2}\left(\frac{y-\theta}{\sqrt{\phi}}\right)^2-\frac{1}{2}ln(2\pi\phi)$$

which is equivalent to

$$\pi(y;\theta,\phi)=\frac{1}{\sqrt{2\pi\phi}}e^{-\frac{1}{2}(\frac{y-\theta}{\sqrt{\phi}})^2}$$

Note that if we let $\pi(.)=f(.)$, $\theta=\mu$, $\phi=\sigma^2$, and $y=x$, this is the same as the first pdf we have for the normal distribution. ■

Proof for Poisson distribution as a member of EDF

Recall 2-1:

$$ln\,\pi(y;\theta,\phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)$$

For Poisson distribution, use (see table 2-1)

$$a(\phi) = 1, b(\theta) = exp\,\theta\,, c(y,\phi) = -\,ln\,y!$$

Substitute function a, b, and c in (2-1), we have:

$$ln\,\pi(y;\theta,\phi) = \frac{y\theta - e^{\theta}}{1} - ln\,y!$$

$$=>\ \pi(y;\theta,\phi) = \frac{e^{y\theta}e^{-e^{\theta}}}{y!}$$

$$= \frac{(e^{\theta})^{y}e^{-(e^{\theta})}}{y!}$$

which is the same as

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

when $e^{\theta} = \lambda\,, y = x.$ ■

Proof for binomial distribution as a member of EDF

Recall 2-1:

$$ln\,\pi(y;\theta,\phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)$$

For Binomial distribution, use (see table 2-1)

$$a(\phi) = \mathrm{n}^{-1}, b(\theta) = \ln\left(1 + e^{\theta}\right), c(y,\phi) = \ln\binom{n}{ny}$$

Substitute function a, b, and c in 2-1:

$$ln\,\pi(y;\theta,\phi) = \frac{y\theta - \ln\left(1 + e^{\theta}\right)}{\mathrm{n}^{-1}} + \ln\binom{n}{ny}$$

$$= ny\theta - nln\left(1 + e^{\theta}\right) + \ln\binom{n}{ny}$$

$$=>\ \pi(y;\theta,\phi) = e^{ny\theta}\left(1 + e^{\theta}\right)^{-n}\binom{n}{ny}$$

$$= (e^{\theta})^{ny}\left(\frac{1}{1+e^{\theta}}\right)^{n+ny-ny}\binom{n}{ny}$$

$$= (e^{\theta})^{ny}\left(\frac{1}{1+e^{\theta}}\right)^{ny}\left(\frac{1}{1+e^{\theta}}\right)^{n-ny}\binom{n}{ny}$$

$$= \left(\frac{e^{\theta}}{1+e^{\theta}}\right)^{ny}\left(\frac{1}{1+e^{\theta}}\right)^{n-ny}\binom{n}{ny}$$

which is the same as

$$f(k;n,p) = p^{k}(1-p)^{n-k}\binom{n}{k}$$

for $k = ny$, $p = \frac{e^{\theta}}{1+e^{\theta}}$, $and\ n = n$ ■

Proof for Gamma distribution as a member of EDF

Recall 2-1:

$$ln\,\pi(y;\theta,\phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)$$

For Gamma distribution, use (see table 2-1)

$$a(\phi) = \mathrm{v}^{-1}, b(\theta) = -\ln(-\theta)\,, c(y,\phi) = \mathrm{v}\ln(vy) - \ln y - \ln(\Gamma v)$$

Substitute function a, b, and c in 2-1:

$$ln\,\pi(y;\theta,\phi) = \frac{y\theta + \ln(-\theta)}{\mathrm{v}^{-1}} + \mathrm{v}\ln(vy) - \ln y - \ln(\Gamma v)$$

$$= vy\theta + vln(-\theta) + vln(vy) - \ln y - \ln(\Gamma v)$$

$$=>\ \pi(y;\theta,\phi) = \frac{e^{vy\theta}(-\theta)^{v}(vy)^{v}}{vy\Gamma}$$

$$= \frac{e^{vy\theta}(-\theta vy)^{v}}{vy\Gamma}$$

which is same as

$$f(x;\alpha,\theta') = \frac{(\frac{x}{\theta'})^{\alpha}e^{-\frac{x}{\theta'}}}{x\Gamma(\alpha)}$$

when $x = vy, \Gamma(\alpha) = \Gamma, \theta' = -\frac{1}{\theta}\,, \alpha = v$ ■

For a random variable Y that follows an EDF distribution, it can be shown that

$$E[Y] = b'(\theta)$$

(2-2)

Take the normal distribution as an example. We know that $b(\theta) = \frac{1}{2}\theta^2$ for a normal distribution. Thus its derivative is

$$b'(\theta) = \theta$$

Recall that we had $\theta = \mu$ for the normal distribution, so the normal distribution satisfies (2-2).

The variance of the same variable Y should also satisfy

$$Var[Y] = a(\phi)b''(\theta)$$

(2-3)

Using the normal distribution again, we know that $Var(Y) = \sigma^2$. Because $a(\phi) = \phi = \sigma^2$ and $b''(\theta) = 1$,

$$Var[Y] = \sigma^2 \cdot 1 = a(\phi)b''(\theta)$$

Moreover, because $b(.)$ is one-to-one by definition, we can take the inverse of $E[Y] = b'(\theta)$ and isolate θ in equation (2-2) as

$$\theta = (b')^{-1}(E[Y])$$

Denote $E[Y]$ as μ, we obtain

$$\theta = (b')^{-1}(\mu)$$

(2-4)

This explains the description of θ being a location parameter, as it is a function of the center, μ, of the distribution.

For the variance of *Y*, we can rewrite (2-3) using the **variance function** derived from (2-4)

$$V(\mu) = b''\big((b')^{-1}(\mu)\big)$$

(2-6)

Equation (2-3) becomes

$$Var[Y] = \alpha(\phi)V(\mu)$$

(2-5)

In this form, we can express the variance of Y on μ and ϕ. This means that we have variance in a form that depends on the mean and the scale parameter.

In addition, for practical purposes, we make the following restriction to $\alpha(.)$ in this paper

$$\alpha(\phi) = \frac{\phi}{w}$$

(2-7)

where we usually assume $w = 1$, so $\alpha(\phi) = \phi$.

2.1.2. The Tweedie Sub-Family

Introduced by Tweedie (1984), the **Tweedie sub-family** belongs to the EDF with the following restriction to the variance function

$$V(\mu) = \mu^p \text{ , } p \leq 0 \text{ or } p \geq 1$$

(2-8)

Using the relations in (2-5) and (2-7) (where we assume $w = 1$), we have that

$$Var[Y] = \alpha(\phi)V(\mu) = \frac{\phi}{w}V(\mu) = \phi V(\mu) = \phi\mu^p$$

Thus we can see in the Tweedie Sub-Family, the variance of Y is proportional to the power of the mean.

Using the relation between μ, θ, and $V(\mu)$, we can further show that

$$b(\theta) = (2-p)^{-1}[(1-p)\theta]^{\frac{2-p}{1-p}}$$

(2-9)

Note that using (2-9) and the relation of $\mu = b'(\theta)$, we can derive μ as

$$\mu = b'(\theta)$$

$$= (2-p)^{-1}(1-p)^{\frac{2-p}{1-p}}\left(\frac{2-p}{1-p}\right)\theta^{\frac{1}{1-p}}$$

$$= (1-p)^{-1}(1-p)^{\frac{2-p}{1-p}}\theta^{\frac{1}{1-p}}$$

$$= (1-p)^{\frac{1}{1-p}}\theta^{\frac{1}{1-p}}$$

and

$$b''(\theta) = (1-p)^{\frac{1}{1-p}}(1-p)^{-1}\theta^{\frac{1}{1-p}-1} = (1-p)^{\frac{p}{1-p}}\theta^{\frac{p}{1-p}}$$

From (2-6) we know $V(\mu) = b''\left((b')^{-1}(\mu)\right) = b''(\theta)$, therefore

$$V(\mu) = b''(\theta) = (1-p)^{\frac{p}{1-p}}\theta^{\frac{p}{1-p}} = \left((1-p)^{\frac{1}{1-p}}\theta^{\frac{1}{1-p}}\right)^{p} = \mu^{p}$$

which is consistent with (2-8).

From the above derivations, we also showed that

$$\mu = [(1-p)\theta]^{\frac{1}{1-p}}$$

(2-10)

And from here, we obtain that for distributions of Tweedie Sub-Family,

$$\theta = \left(\mu(1-p)^{-\frac{1}{1-p}}\right)^{1-p} = \frac{\mu^{1-p}}{1-p}$$

Using the above relations, we can rewrite the pdf of Tweedie Sub-Family distributions as

$$ln\,\pi(y;\mu,\phi) = \frac{\left[\frac{y\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right]}{\phi} + c(y,\phi)$$

(2-11)

denoted $Tw(\mu,\phi;p)$.

This is derived from the definition of EDF general formula (2-1), where for

$$ln\,\pi(y;\theta,\phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)$$

we know that $\theta = \frac{\mu^{1-p}}{1-p}$ and

$$b(\theta) = (2-p)^{-1}[(1-p)\theta]^{\frac{2-p}{1-p}} = (2-1)^{-1}\left[(1-p)\frac{\mu^{1-p}}{1-p}\right]^{\frac{2-p}{1-p}} = \frac{\mu^{2-p}}{2-p}$$

In addition, because we restrained $a(\phi) = \phi$, by substituting $\frac{\mu^{1-p}}{1-p}$ for θ, $\frac{\mu^{2-p}}{2-p}$ for $b(\theta)$, and ϕ for $a(\phi)$, we can thus obtain (2-11).

Example distribution from Tweedie sub-family – normal distribution

In addition to being a distribution from the EDF, the normal distribution also belongs to the Tweedie sub-family. As we know of the normal distribution, for $Y \sim Normal(\mu, \sigma^2)$,

$$Var[Y] = \sigma^2 = \phi = a(\phi)\cdot 1 = \alpha(\phi)V(\mu)$$

And because we also know that for a normal distribution, $\theta = \mu$.

Thus $b(\theta) = \frac{1}{2}\theta^2 = \frac{1}{2}\mu^2$, which satisfies (2-9), and

$$b''(\theta) = \frac{d^2}{d^2\mu}\left(\frac{1}{2}\mu^2\right) = \frac{d}{d\mu}(\mu) = 1 = V(\mu)$$

where we have $p = 0$, which satisfies (2-10).

Using μ, ϕ, $c(y,\phi) = -\frac{1}{2}\left[\frac{y^2}{\phi} + ln(2\pi\phi)\right]$, with $p = 0$, we obtain from (2-11)

$$ln\,\pi(y;\mu,\phi) = \frac{\left[\frac{y\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right]}{\phi} + c(y,\phi)$$

$$= \frac{\left[\frac{y\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right]}{\phi} - \frac{1}{2}\left[\frac{y^2}{\phi} + ln(2\pi\phi)\right]$$

$$= \frac{\left[y\mu - \frac{\mu^2}{2}\right]}{\phi} - \frac{1}{2}\left[\frac{y^2}{\phi} + ln(2\pi\phi)\right]$$

$$= -\frac{1}{2}\frac{[y^2 - 2y\mu + \mu^2]}{\phi} - \frac{1}{2}[\ln(2\pi\phi)]$$

$$= -\frac{1}{2}\frac{(y-\mu)^2}{\phi} - \frac{1}{2}[\ln(2\pi\phi)]$$

So for $\pi(.)$, we obtain

$$\pi(y;\mu,\phi) = \frac{1}{\sqrt{2\pi\phi}} e^{-\frac{1}{2}(\frac{y-\mu}{\sqrt{\phi}})^2}$$

Beside the normal distribution, Table 2-2 below contains more examples of the Tweedie sub-family with different p values:

Distribution	p	$b(\theta)$	μ	$ln\,\pi(y;\mu,\phi)$
Over-dispersed Poisson	1	$exp\,\theta$	$exp(\theta)$	$\frac{[y\,ln\,\mu - \mu]}{\phi}$
Gamma	2	$ln(-\theta)$	$-\frac{1}{\theta}$	$\frac{\left[-\frac{y}{\mu} - ln\,\mu\right]}{\phi}$
Inverse Gaussian	3	$-(-2\theta)^{\frac{1}{2}}$	$(-2\theta)^{-\frac{1}{2}}$	$\frac{\left[-\left(\frac{y}{2}\mu^2\right) + \frac{1}{\mu}\right]}{\phi}$

Table 2-2

2.1.3. The Over-Dispersed Poisson Sub-Family

The **Over-Dispersed Poisson (ODP)** distribution was introduced in Table 2-2 at the end of the previous section. This distribution plays a central role in the rest of the paper, particularly in the stochastic models that support the chain ladder algorithms. Thus we introduce ODP distribution in more details here.

As noted in the table, the ODP distribution is part of the Tweedie sub-family with $p = 1$. We will thus denote it as $ODP(\mu, \phi)$, because p is fixed.

The pdf of ODP (Table 2-2) is as follows:

$$\pi(y;\mu,\phi) = \mu^{\frac{y}{\phi}} e^{\left[-\frac{\mu}{\phi}+c(y,\phi)\right]}$$

(2-14)

for $y = 0, \phi, 2\phi, \ldots$ and $\mu = e^{\theta}$.

We can rewrite (2-14) in the general form of distribution from EDF

$$ln\,\pi(y;\mu,\phi) = \frac{y\,ln\,\mu - \mu}{\phi} + c(y,\phi)$$

where in this case, $\theta = ln\,\mu \leftrightarrow \mu = e^{\theta}$, $b(\theta) = \mu = e^{\theta}$.

The unit total probability mass, $exp\,c(y,\phi)$ is obtained if

$$e^{c(y,\phi)} = \left[\left(\frac{y}{\phi}\right)!\right]^{-1}$$

(2-15)

If we substitute (2-15) into the pdf of ODP (2-14), we can obtain

$$\pi(y;\mu,\phi) = \frac{\mu^{\frac{y}{\phi}} e^{-\frac{\mu}{\phi}}}{\left(\frac{y}{\phi}\right)!}$$

(2-16)

for $y = 0, \phi, 2\phi, \ldots$

the monograph claims from here that from this pdf, we can actually observe that the Poisson distribution can be represented by ODP as

$$\frac{\mathrm{Y}}{\phi} \sim Poiss\left(\frac{\mu}{\phi}\right)$$

(2-17)

However, this is not true, as the form does not match exactly the distribution of the Poisson distribution.

Recall the Poisson distribution as the following form:

$$f_X(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

However, from (2-17), we obtain

$$\pi\left(\frac{y}{\phi};\mu,\phi\right) = \frac{\left(\frac{\mu}{\phi}\right)^{\frac{y}{\phi}} e^{-\frac{\mu}{\phi}}}{\left(\frac{y}{\phi}\right)!}$$

which does not match the form in (2-16).

On the other hand, if we let $\phi = 1$ in (2-16), the obtained pdf is

$$(y;\mu,1) = \frac{\mu^y e^{-\mu}}{y!}$$

Which does reduce to simple Poisson distribution denoted

$$Y \sim Poiss(\mu)$$

(2-20)

2.2. Generalized Linear Models (GLMs)

2.2.1. Definition

Let $\pi(.;\ \mu,\phi)$ denote a distribution of the EDF, and denote Y_i, $i = 1,2,\dots,n$ as a sample of observation.

Suppose that each Y_i has a known q-vector of predictors (or **covariates**, which is an independent variable of a model), $x_{i1}, x_{i2}, \dots, x_{iq}$. Denote its transpose as $x_i = \left(\mathrm{x}_{i1}, \mathrm{x}_{i2}, \dots, \mathrm{x}_{iq}\right)^T$. Then a model is called **a generalized linear model (GLM)** if it satisfies the following 3 conditions:

1. $Y_i \sim \pi(.;\ \mu_i,\phi_i)$ where μ_i are unknown parameters.
2. $h(\mu_i) = x_i^T\beta$, where $h(.)$ it is a one-to-one link function in $(-\infty,+\infty)$; β is a q-vector of unknown parameters, where $\beta = \left(\beta_1,\beta_2,\dots,\beta_q\right)^T$.
3. Observations Y_i are stochastically independent.

In the GLM, the variate Y_i is called the response, and the $x_i^T \beta$ is called the linear response.

Denote the dispersion parameter ϕ_i, where

$$\phi_i = \phi / w_i$$

(2-21)

The ϕ is the overall dispersion parameter, and w_i are the weights associated with each ϕ_i that corresponds to the variates Y_i. Usually it is assumed that the overall ϕ unknown but w_i are known.

While GLM is more generalized than a linear regression, the GLM is a regression model. Its relation with linear regression can be seen in the following example:

if we let the density function $\pi(.;\ \mu_i, \phi_i)$ be the normal density, $n(.;\ \mu_i, \phi_i)$, and let the link function $h = identity$, then we can rewrite condition (1) and (2) as

$$Y_i = x_i^T \beta + \varepsilon_i \text{ with } \varepsilon_i \sim N(0, \phi_i)$$

(2-22)

which is a weighted linear regression model.

For simplicity, we can also express condition (2) in vector and matrix form, which will be used frequently in the following chapters. The matrix X is called the design matrix of regression. Let

Y – the $n \times 1$ vector with i-th component Y_i

μ – the $n \times 1$ vector with i-th component μ_i

X – the $n \times q$ matrix with i-th row x_i^T:

$$\begin{bmatrix} x_{11} & \cdots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nq} \end{bmatrix}$$

Then we can rewrite condition (2) as

$$\mu = h^{-1}(X\beta)$$

(2-23)

where μ is an $n \times 1$ vector like Y, and β is a $q \times 1$ vector.

In GLM, each variate will have one canonical parameter (or location parameter). Thus combining all canonical parameters, we obtain an n-vector $(\theta_1, \dots, \theta_n)$ for the GLM. Denote this vector from now on as θ, then recall from (2-2), which states that $E[Y] = \mu = b'(\theta)$, and combine with (2-23), we obtain the following

$$b'(\theta_i) = E[Y_i] = \mu_i = h^{-1}(x_i^T \beta)$$

(2-24)

2.2.2. Categorical and Continuous Covariates

Covariates can be divided into two types: **categorical** and **continuous** covariates. Categorical covariates are covariates that are discrete, such as possible numerical values from rolling a dice, or non-numerical values such as genders. Continuous covariates are, as the name suggests, numerical within a continuous range, such as age and height.

Categorical Variates

Suppose a categorical variate has m possible values, where m is usually referred to as the levels of the variate. Denote these possible values as $\xi_1, \dots, \xi_m$. In the GLM, we can represent this as 0-1 variates for a total of m variates, denoted $x_{k+1}, x_{k+2}, \dots, x_{k+m}$, with other regression covariates denoted as $x_1, x_2, \dots, x_k, x_{k+m+1}, \dots$. The 0-1 variates are defined as follows:

$$x_{k+r} = \begin{cases} 1, & if\ the\ categorical\ variate\ assumes\ the\ value\ \xi_r \\ 0, & otherwise \end{cases}$$

(2-25)

For example, if we have gender as our categorical variates in a model, we obtain a level of variate of 2, where

$$x_{k+r} = \begin{cases} 1, & if\ the\ person\ selected\ is\ a\ male \\ 0, & if\ the\ person\ selected\ is\ a\ female \end{cases}$$

Note that for $r = 1,2,\dots,m$,

$$\sum_{r=1}^{m} x_{k+r} = 1$$

(2-26)

which means that only one category can be selected at a time.

Applying this concept to loss reserve, for example, if we want to include development years as a covariate in our model, we need to treat the development years as categorical variates ξ with J levels. Specifically, we have the following 0-1 variates

$$x_{k+j} = \begin{cases} 1, & if\ \xi = j \\ 0, & otherwise \end{cases}$$

Continuous Variates

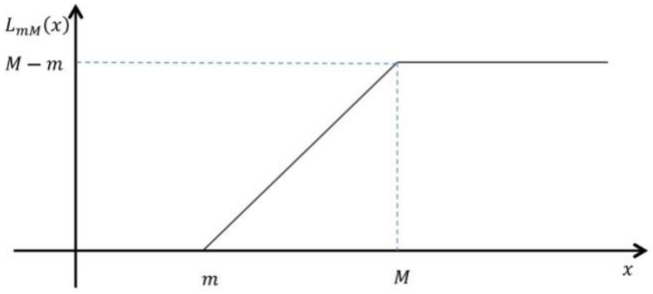
Because continuous variates take numerical values, they simply represent themselves in a regression model. For example,

$$L_{mM}(x) = min[M - m, max\ (0, x - m)] \text{ with } m < M$$

(2-27)

has unit gradient between m and M and constant outside this range. Visually this has the following graph from the monograph

Figure 2-1. Illustration of the Function $L_{mM}(x)$



A specific form of $L_{mM}(x)$, $L_{m_k m_{k+1}}(x)$, are basis functions, which are usually used to construct piecewise linear functions, called linear splines. An example of linear splines is as follows:

$$\sum_{k=1}^{K} \beta_k \, L_{m_k m_{k+1}}(x)$$

(2-28)

which is constructed as a linear combination of the basis functions, and has knots of $x = m_1, \dots, m_{K+1}$ and gradient β_k.

2.2.3. Goodness-of-Fit and Deviance

If we fit a model with parameters (arranged in a vector) β to a set of observations (also arranged in a vector) *Y*, and the parameter estimates $\hat{\beta}$ are **maximum likelihood estimates (MLEs)** of β, then the vector of fitted values $\hat{Y}$ is the MLE of μ, written as

$$\hat{Y} = h^{-1}(X\hat{\beta})$$

(2-29)

To measure how well the MLE parameter estimates $\hat{\beta}$ model the observations means, we need to test the goodness-of-fit of the model and MLEs. A common measure of goodness-of-fit of a GLM is by calculating its scaled deviance, which is defined as

$$\begin{aligned} D(Y, \hat{Y}) &= 2\left[ln\, \pi\left(Y;\, \hat{\theta}^{(s)}, \phi\right) - ln\, \pi\left(Y;\, \hat{\theta}, \phi\right)\right] \\ &= 2\sum_{i=1}^{n} \left[ln\, \pi\left(Y_i;\, \hat{\theta}^{(s)}, \phi\right) - ln\, \pi\left(Y_i;\, \hat{\theta}, \phi\right)\right] \end{aligned}$$

(2-30)

where θ is the location parameter vector and $\hat{\theta}$ is the MLE of θ, $\hat{\theta}^{(s)}$ is the estimate of θ in the saturated model (which means that each observation of the model has a corresponding parameter so that $\hat{Y} = Y$).

For simplicity, refer to the unscaled deviance as simply deviance. For a deviance calculation, the scale parameter ϕ is ignored (equivalently, set to 1) and the equation is defined as follows

$$D^*(Y, \hat{Y}) = 2\sum_{i=1}^{n}[ln\ \pi(Y_i;\ \hat{\theta}^{(s)}, 1) - ln\ \pi(Y_i;\ \hat{\theta}, 1)]$$

(2-31)

The MLE minimizes $D^*(Y, \hat{Y})$ with respect to $\hat{\theta}$.

2.2.4. Residuals

Pearson Residuals

The **standardized Pearson residuals** for associated observations Y_i are defined as

$$R_i^P = (Y_i - \hat{Y}_i)/\hat{\sigma}_i$$

(2-33)

with $\hat{\sigma}_i$ being the estimate of σ_i and $\sigma_i^2 = Var[Y_i]$. We will also use this in Chapter 5 for Bootstrapping residuals for re-sampling.

Assuming that $\hat{Y}_i$ is approximately unbiased as an estimator of μ_i, and

$$Var[\hat{Y}_i - Y_i] \cong Var[Y_i]$$

then we have the following properties for the standardized Pearson residuals:

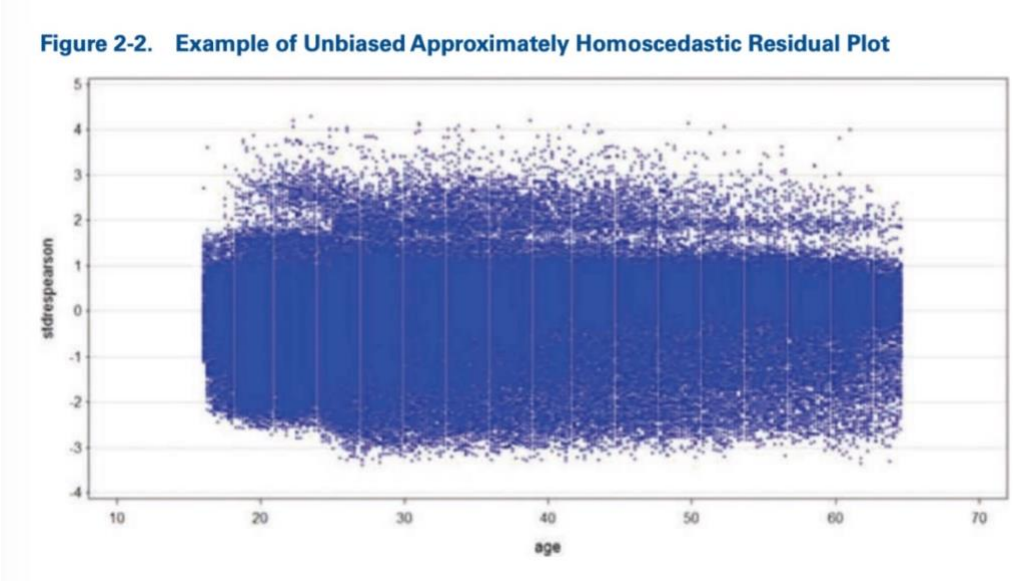
$$E[R_i^P] = 0$$

$$Var[R_i^P] = 1$$

(2-34)

Visually, this means if we plot the residuals, they should scatter evenly about the line $y = 0$, as shown in the following figure from the monograph

Figure 2-2. Example of Unbiased Approximately Homoscedastic Residual Plot



Beside the unbiasedness, note that the scatterplot also has a uniform dispersion from left to right. This feature is called homoscedasticity, and together with unbiasedness, these are crucial parts for model validation.

Deviance Residuals

The **standardized deviance residual** is defined as

$$R_i^D = sgn(Y_i - \hat{Y}_i)\left(\frac{d_i}{\hat{\Phi}}\right)^{\frac{1}{2}}$$

(2-35)

Where *sgn* is the sign function defined as

$$sgn(x) = \begin{cases} -1, & x < 0 \\ 0, & x = 0 \\ 1, & x > 0 \end{cases}$$

and d_i is the i-th observation of deviance $D^*(Y, \hat{Y})$.

The deviance residual is useful because it is not affected by non-normality in the observations as the Pearson residuals are, and thus are more applicable when handling non-normal distributions. Figure 2-4 and 2-5 from the monograph show an example of

plotting the standardized Pearson residuals and deviance residuals for the same dataset and model. The figures show that while the histogram of the standardized Pearson residuals is heavily skewed toward the right, the deviance residuals greatly reduce the skewness and are more normally distributed.

Figure 2-4. Histogram of Standardized Pearson Residuals

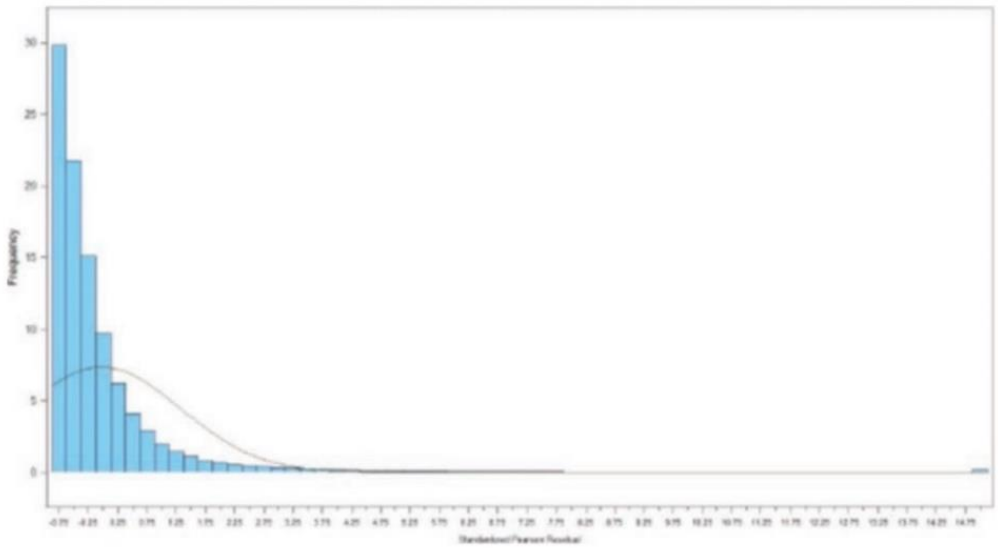
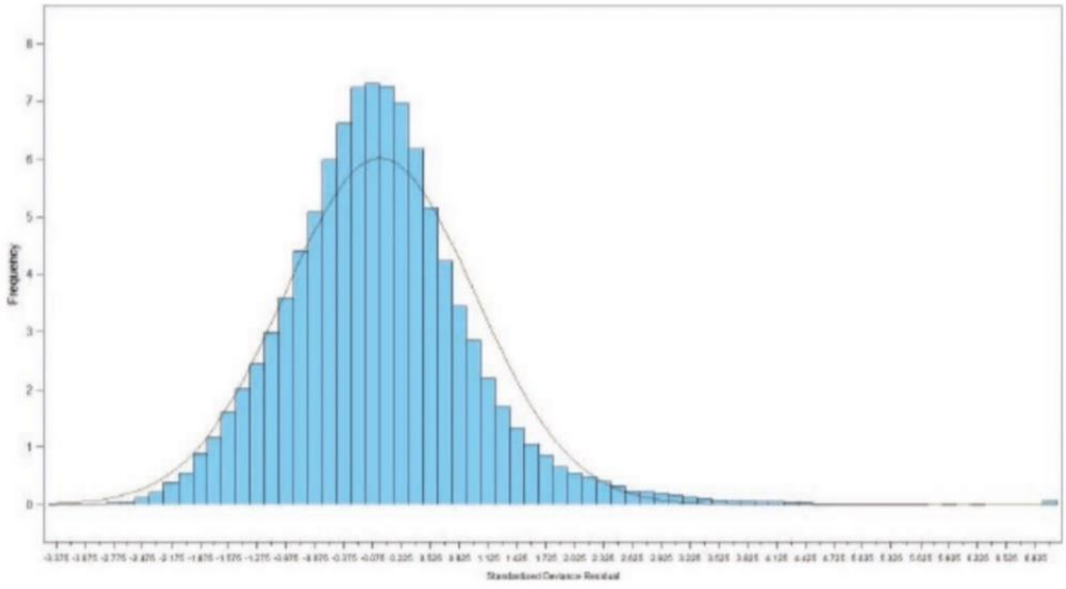


Figure 2-5. Histogram of Standardized Deviance Residuals



2.2.5 Outliers and the Use of Weights

Use of weights in the case of heteroscedasticity

Suppose a GLM has the property of homoscedasticity, specifically

$$Var[Y_i] = \phi V(\mu_i)$$

(2-36)

This means the variance of an observation Y_i depends on both the dispersion parameter and the variance of its mean.

Then, suppose the standardized Pearson residual we observed shows heteroscedasticity. For example, that the residuals above age 55 has standard deviation twice as large as those below age 55.

If we express the standardized Pearson residual using equations (2-5), (2-7), (2-33), which are

$$Var[Y] = \alpha(\phi)\mathrm{V}(\mu)$$

(2-5)

$$\alpha(\phi) = \phi/\omega$$

(2-7)

$R_i^P = \frac{Y_i - \hat{Y}_i}{\hat{\sigma}_i}$, where $\hat{\sigma}_i^2 = Var[Y_i]$

(2-33)

Then the standardized Pearson residual can be expressed as

$$\begin{aligned} R_i^P &= \frac{Y_i - \hat{Y}_i}{\hat{\sigma}_i} \\ &= \frac{Y_i - \hat{Y}_i}{Var[Y_i]^{\frac{1}{2}}} \\ &= \frac{Y_i - \hat{Y}_i}{\left(\hat{\phi}\mathrm{V}(\hat{\mu}_i)\right)^{\frac{1}{2}}} \end{aligned}$$

(2-37)

The observed heteroscedasticity suggests that ϕ takes different values for those with age above 55 and those with age below 55. Specifically, because the standard deviation is twice as large, the value of ϕ for above 55 is 4 times as large as for ϕ below 55.

Therefore, if we want to remove the heteroscedasticity in this model, we can use weights to reflect the variation in ϕ over age. Specifically, for the formula of

$$\alpha(\phi) = \phi/\omega$$

which is for the ϕ that apply to all age, we can adjust it to

$$\alpha(\phi_i) = \phi/\omega_i$$

with ϕ being a constant, and ϕ_i and ω_i are for the corresponding ith observation.

Then, let

$$\omega_i = 1 \ , age \leq 55$$

$$\omega_i = \frac{1}{4} \ , age > 55$$

such that

$$\alpha(\phi_i) = \phi, age \leq 55$$

$$\alpha(\phi_i) = 4\phi, age > 55$$

In this way, the model reflects the differences in ϕ for age below and above 55, where the value of ϕ is now 4 times as large for age above 55 as that below age 55. Thus we can eventually achieve homoscedasticity.

From the above example we can conclude the following: in the default setting with no specific introduction of weights, all observations are equally weighted for the purpose of parameter estimation. However, if certain groups of observations have variance larger than others, they should be weighted less.

Also, estimation efficiency will be optimized when each observation is weighted inversely proportional to its ϕ. In the above example, where the value of ϕ is 4 times larger, the assigned weight of $\omega_i = 1/4$ is the inverse of the coefficient for $\alpha(\phi_i)$.

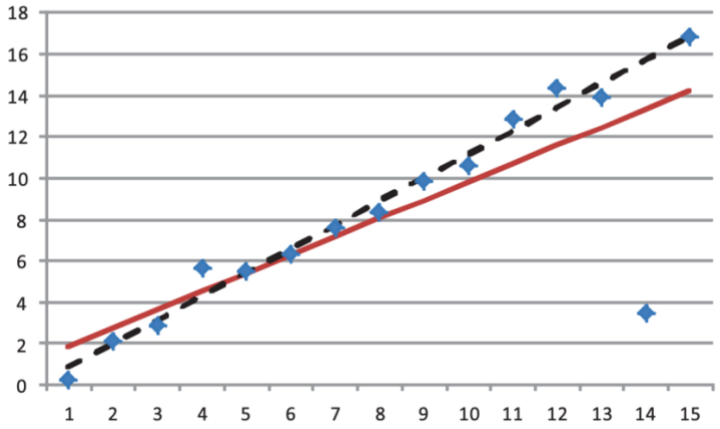
Therefore, when we see patterns of heteroscedasticity in the residual plot, we should adjust the weights of observations so that the weights are inversely proportional to the variance of their residuals.

Outliers

In residual plots, we may identify observations with very large residuals, called **outliers**. These observations influence the accuracy of our regression analysis because they can move our fitted model away from its main body.

When we observe an outlier, we can exclude it from our analysis. However, we should be careful because outliers could be the representation of a major change in the environment of the population. In this case, the exclusion of outlier would be inappropriate.

Figure 2-6. Illustration of Distortion of Regression by Outlier



2.2.6 Forecasts

Recall that

$$E[Y_i] = \mu_i = h^{-1}(x_i^T \beta)$$

(2-38)

In this model, covariates x_i includes factors that can influence the values of losses.

When we estimate the losses for future, the difference in covariates x_i is that it includes time variates related to the future.

To distinguish the difference, for future observations, we use notation * to suggest its purpose of forecast, or more commonly referred to in this paper as future estimation. For example, (2-38) should be then written as

$$E[Y_i^*] = \mu_i^* = h^{-1}\left(x_i^{*T} \beta\right)$$

or using vector form

$$\mu^* = h^{-1}(X^*\beta)$$

(2-39)

where X^* is the matrix with rows being x_i^{*T} and called the forecast design matrix.

Then, the future estimates of Y^* can be expressed as

$$\hat{Y}^* = \hat{\mu}^* = h^{-1}\left(X^*\hat{\beta}\right)$$

(2-40)

This notation will be used for the rest of the paper to identify future observations.

3. Stochastic Models Supporting the Chain Ladder Method

3.1. Mack Models

The **Mack model** is a stochastic chain ladder model introduced by Mack (1993). Section 3.1. provides the theoretical support, and distributional characteristics for the chain ladder that Chapter 1 has explained.

3.1.1. Non-Parametric Mack Model

There are 3 conditions for the Mack model:

(M1) For different accident years, i.e. $k_1 \neq k_2$, the incremental losses such as Y_{k1j1} and Y_{k2j2} are stochastically independent.

(M2) For each $k = 1, 2, \dots, K$ (i.e. for each row), the X_{kj} (j varying) form a Markov chain. A Markov chain is a chain of observations in which the probability and size of the j-th observations is only affected by the previous, $(j-1)$-th observations, i.e.

$$P\left(X_j \middle| X_1 = x_1, \dots, X_{j-1} = x_{j-1}\right) = P\left(X_j \middle| X_{j-1} = x_{j-1}\right).$$

(M3) For each $k = 1, 2, \dots, K$ and $j = 1, 2, \dots, J-1$,

(a) $E[X_{k,j+1} | X_{kj}] = f_j X_{kj}$ for some parameter $f_j > 0$
(b) $Var[X_{k,j+1} | X_{kj}] = \sigma_j^2 X_{kj}$ for some parameter $\sigma_j > 0$

Recall here that f is the age-to-age factor, and the expectation of X for the next development period is calculated by multiplying the current observation with the f for the current development period.

σ_j is the standard deviation and the variance of X of next development period is calculated by multiplying the current observation by the σ_j^2 of the current column.

The Mack model is stochastic because it considers both expected values and the variances of the observations. But it is non-parametric as it does not consider the distribution of observations.

Mack derived the following result for this model:

Result 1: The conventional chain ladder estimators $\hat{f}_j$ of the age-to-age factors f_j according to (1-8) are:

(a) unbiased
(b) minimum variance among estimators that are unbiased linear combinations of the $\hat{f}_{kj}$ defined by (1-4)

Recall (1-4)

$$\hat{f}_{kj} = \frac{X_{k,j+1}}{X_{kj}}, \qquad k = 1, 2, \ldots, K-1; j = 1, 2, \ldots, min(J-1, K-k)$$

where the $\hat{f}_{kj}$ are the age-to-age factors for each cumulative observation X_{kj} to the next $X_{k,j+1}$, where as in (1-8)

$$\hat{f}_j = \frac{\sum_{k=1}^{K-j} X_{k,j+1}}{\sum_{k=1}^{K-j} X_{kj}}$$

where the $\hat{f}_j$ are the weighted age-to-age factors for the development period j, and can also be calculated by summing the weighted $\hat{f}_{kj}$.

Result 2: The conventional chain ladder estimators $\hat{R}_k$ for the total outstanding loss R_k of accident year k from (1-11) is unbiased

(Recall (1-11):

$$\hat{R}_k = \hat{X}_{kJ} - X_{k,K-k+1} = X_{k,K-k+1}(\hat{f}_{K-k+1} \cdots \hat{f}_{J-1} - 1)$$

where the total outstanding loss is obtained by subtracting the last known observation from the estimated ultimate loss.)

3.1.2. Parametric Mack Models

EDF Mack model a parametric version of the Mack model, which means the model assumes the observations follow a distribution. Parametric versions of the Mack model were studied by Taylor (2011). Thus for EDF Mack model, the last condition of the model needs to be changed, and are as follows:

(EDFM1) For different accident years, i.e. $k_1 \neq k_2$, the incremental losses such as Y_{k1j1} and Y_{k2j2} are stochastically independent.

(EDFM2) For each $k = 1, 2, \dots, K$ (i.e. for each row), the X_{kj} (j varying) form a Markov chain. A Markov chain is a chain of observations that the probability and size of the j-th observations is only affected by the previous, $(j-1)$-th observations, i.e.

$$P\left(X_j \middle| X_1 = x_1, \dots, X_{j-1} = x_{j-1}\right) = P\left(X_j \middle| X_{j-1} = x_{j-1}\right).$$

(EDFM3) For each $k = 1, 2, \dots, K$ and $j = 1, 2, \dots, J-1$,

(a) $Y_{k,j+1} | X_{kj} \sim EDF\ (\theta_{kj}, \phi_{kj}; a, b, c)$
(b) $E[X_{k,j+1} | X_{kj}] = f_j X_{kj}$ for some parameter $f_j > 0$

Here, (EDFM3a) provides the distributional assumptions for the observations to some specific member of the EDF.

(EDFM3b) retains the assumption for expected values in (M3a).

Note that for the parametric form of Mack model, there is no specific condition for the form of variance, which allows for a more general form of variance for the model than the non-parametric model. However, there is the additional restriction of observations following a distribution from the EDF.

Recall from Chapter 2 that Tweedie and ODP are 2 sub-families of EDF, so the parametric Mack models for these families of distributions satisfy the same conditions as EDF Mack model, but with the replacement of (EDFM3a) by:

Tweedie Mack model:

Replace (EDFM3a) with $Y_{k,j+1} | X_{kj} \sim Tw\ (\mu_{kj}, \phi_{kj}; p)$

ODP Mack model:

Replace (EDFM3a) with $Y_{k,j+1} | X_{kj} \sim ODP(\mu_{kj}, \phi_{kj})$

Taylor derived the following result for the EDF Mack model:

Theorem 3.1.

Suppose the dataset of past observations, $\mathfrak{D}_K$, is a triangle, i.e. $K = J$, which has observations that satisfy the conditions (EDFM1-3) for the EDF Mack model.

(a) If a specific assumption of variance for the non-parametric version of the Mack model, (M3b), also stands in addition to (EDFM1-3), then the model's MLEs of f_j and the conventional chain ladder estimators $\hat{f}_j$ from (1-8) are the same, and are both unbiased estimators of f_j. Thus Result 1 from Section 3.1.1 holds.

(b) If the model assumption is restricted to an ODP Mack model and the dispersion parameters ϕ_{kj} are only column dependent, i.e. $\phi_{kj} = \phi_j$ (note that the condition (M3b) holds in this case), then the $\hat{f}_j$ from (1-8) are minimum variance unbiased estimators (**MVUE**) of the f_j.

(c) If the assumptions in (b) hold, then the estimators $\hat{X}_{kj}$ and $\hat{R}_k$ for cumulative outstanding losses and total outstanding losses X_{kj} and R_k from (1-9) and (1-11) are also MVUEs.

These results and theorems also extend to some cases when Y_{kj} follows a binomial distribution or negative binomial distribution.

Numerical Example:

In this section, we use the data set in Table 1-1 to illustrate the manual process of the Mack model.

Recall that the parameters for the Mack are f_j, where by condition (M3a) and (EDF3b),

$$E\left[X_{k,j+1}\middle|X_{kj}\right] = f_j X_{kj}$$

Note that this is the identical to the chain ladder algorithm. For known age-to-age factors f_j and past observation X_{kj}, the expected value of observation of the next development period $X_{k,j+1}$, should be the product of f_j and X_{kj}. Also recall part (a) of Theorem 3.1: for an EDF Mack model with additional assumption (M3b) for variance, the estimated parameters $\hat{f}_j$ are the same for Mack model and conventional chain ladder algorithm. Thus, to manually calculate the parameters, we simply apply the conventional chain ladder algorithm.

Recall from Chapter 1, for the conventional chain ladder algorithm, estimates of future cumulative losses are obtained by means of (1-9), where

$$\hat{X}_{kj} = X_{k,K-k+1}\hat{f}_{K-k+1} \dots \hat{f}_{j-1}$$

And we can use the cumulative estimates to obtain the incremental future losses, where

$$\hat{Y}_{kj} = \hat{X}_{kj} - \hat{X}_{k,j-1}$$

The age-to-age factors (calculated by taking the weighted average) by conventional chain ladder algorithm, $\hat{f}_j$, were defined in Chapter 1, (1-8) as

$$\hat{f}_j = \frac{\sum_{k=1}^{K-j} X_{k,j+1}}{\sum_{k=1}^{K-j} X_{kj}}, j = 1 \dots J-1$$

Using the dataset of Table 1-1 (adjusted to cumulative observations, which is Table 1-2), we can obtain the following age-to-age factors

Average age-to-age factor for development year								
1	2	3	4	5	6	7	8	9
1.815	1.261	1.158	1.088	1.055	1.039	1.030	1.025	1.021

Table 3-1

And the cumulative outstanding losses in Table 3-2 from the monograph

Table 3-2. Estimation of Outstanding Losses

Accident Year		Forecast Cumulative Paid Losses to and Including Development Year ($000)										Estimated Outstanding Claims ($000)
		1	2	3	4	5	6	7	8	9	10	
1988	1										**144781**	
1989	2									**162903**	166301	3398
1990	3								**176346**	180731	184501	8155
1991	4							**187266**	192924	197721	201845	14579
1992	5						**189506**	196828	202774	207817	212151	22645
1993	6					**175475**	185209	192364	198176	203104	207340	31865
1994	7				**159972**	174108	183766	190866	196632	201522	205725	45753
1995	8			**122811**	142227	154795	163381	169693	174820	179168	182904	60093
1996	9		**92242**	116312	134700	146603	154735	160713	165569	169686	173225	80983
1997	10	**43962**	79788	100608	116513	126809	133843	139014	143214	146775	149836	105874
												373346

Example of obtaining the age-to-age factor $\hat{f}_2$, the cumulative and incremental observations $\hat{X}_{1996,3}$ and $\hat{Y}_{1996,3}$

For this example, because the observations used for the age-to-age factors are cumulative, we need to refer not to the original incremental dataset, but the cumulative dataset shown in Table 1-2, derived from Table 1-1, shown partially below

Cumulative Paid Losses ($000)										
k\j	1	2	3	4	5	6	7	8	9	10
1	41,821	76,550	96,697	112,662	123,947	129,871	134,646	138,388	141,823	144,781
2	48,167	87,662	112,106	130,284	141,124	148,503	154,186	158,944	162,903	
3	52,058	99,517	126,876	144,792	156,240	165,086	170,955	176,346		

The age-to-age factor for development year 8 in the table is obtained using the equation (1-8) shown above as follows:

$$1.025 = \frac{\sum_{k=1}^{2} X_{k,9}}{\sum_{k=1}^{2} X_{k,8}} = \frac{141{,}823 + 162{,}903}{138{,}388 + 158{,}944}$$

The estimated cumulative observation $\hat{X}_{3,9}$ can thus be calculated using (1-9) as

$$\hat{X}_{3,9} = X_{3,8} \times \hat{f}_8 = 176{,}346 \times 1.025 = 180{,}731$$

and the estimated incremental observation $\hat{Y}_{3,9}$ is

$$\hat{Y}_{3,9} = \hat{X}_{3,9} - X_{3,8} = 180{,}731 - 176{,}346 = 4{,}385$$

3.2. Cross-Classified Models

Unlike the Mack Model which uses the cumulative observations X_{kj}, the **cross-classified (CC) model** uses incremental observations Y_{kj} for estimating parameters and future losses. The **EDF CC model** of the past and future observations in $D_K^+ = D_K \cup D_K^c$ satisfy the following condition:

(**EDFCC1**) The random variables $Y_{kj} \in D_K^+$ are stochastically independent.

(**EDFCC2**) For $k = 1,2,\dots,K$ and $j = 1,2,\dots,J$,

a) $Y_{kj} \sim EDF(\theta_{kj}, \phi_{kj}; a, b, c)$;
b) $E[Y_{kj}] = \alpha_k \beta_j$ for some parameters α_k, $\beta_j > 0$; and
c) $\sum_{j=1}^{J} \beta_j = 1$.

Note that because CC models are subject to (EDFCC1) and (EDFCC2b), they are non-recursive, which is different from the Mack model. Intuitively this aligns with the dataset assumptions of Mack and CC models, because Mack models uses cumulative data, where data of each development period is dependent on the data of the previous, while the CC model uses incremental data, which is independent both by accident year and development year. The EDF CC model also consists of parameters for both the rows k (accident periods), and columns j (development periods), whereas the parameters of the Mack model f_j only concerns with the columns j, because the condition (M2) of X_{kj} varying j form a Markov chain already plays the role of parametrizing the observations in each row. The last condition, (EDFCC2c) is placed to remove the excessive parameters that can occur, by scaling all the α_k and β_j with the standard $\sum_{j=1}^{J}\beta_j = 1$. This restriction will ensure the uniqueness of the model parameters, and that the parameter estimates $\hat{\alpha}_k$ are the estimated ultimate losses.

As with the Mack model, there exist the Tweedie and ODP sub-families of the EDF CC family, which are called the Tweedie CC family and ODP CC family respectively. For the Tweedie CC model and ODP CC model, there would only be change to the condition (EDFCC2a), which would become:

Tweedie CC model – replace (EDFCC2a) by $Y_{kj} \sim Tw(\mu_{kj}, \phi_{kj}; p)$.

ODP CC model – replace (EDFCC2a) by $Y_{kj} \sim ODP(\mu_{kj}, \phi_{kj})$.

Denote the MLEs of the parameters α_k and β_j as $\hat{\alpha}_k$ and $\hat{\beta}_j$, and denote the fitted values of $Y_{kj} \in D_K^+$ as $\hat{Y}_{kj} = \hat{\alpha}_k\hat{\beta}_j$. Then the following theorem by England & Verrall (2002) hold true for the ODP CC model:

Theorem 3.2. Suppose that the data array D_K is a triangle, i.e. $K = J$, with observations subject to the ODP CC model defined by:

(EDFCC1-2)

(EDFCC3a) restrict the Y_{kj} in (EDFCC2a) to ODP distribution, i.e., $Y_{kj} \sim ODP(\mu_{kj}, \phi_{kj})$;

(EDFCC3b) the dispersion parameters ϕ_{kj} are identical for all cells in D_K^+, i.e., $\phi_{kj} = \phi$.

Then the MLE fitted values and estimates $\hat{Y}_{kj}$ are the same as those given by the conventional chain ladder from (1-10).

[recall (1-10): $\hat{Y}_{kj} = X_{k,K-k+1}\hat{f}_{K-k+1} \dots \hat{f}_{j-2}(\hat{f}_{j-1} - 1)$, where $\hat{f}_j$ are the age-to-age factors by the conventional chain ladder method]

However, the same result does not hold for more general distributions, such as the Tweedie sub-family and EDF distributions.

The MLEs $\hat{Y}_{kj}$ is not unbiased in most cases for ODP CC model. But bias can be corrected, and according to Taylor (2011) the following result holds true for the bias corrected situations:

Theorem 3.3. Suppose that the data array D_K^+ is subject to the same conditions as in Theorem 3.2., and that the current and future fitted values $\hat{Y}_{kj}$ and $\hat{R}_k$ are corrected for bias. Then they are MVUEs of Y_{kj} and R_k respectively.

Recall that in Theorem 3.1 for the ODP Mack model, there was a similar statement of the $\hat{X}_{kj}$ and $\hat{R}_k$ being MVUEs of X_{kj} and R_k with some additional restrictions. Thus Theorems 3.1 and 3.2 conclude that the future estimates obtained from the ODP Mack and ODP CC models are both identical to the conventional chain ladder, despite the models having different formulations.

Numerical example

In this section, we use the data set in Table 1-1 to illustrate manual process of the CC model.

Because the ODP CC model uses the incremental dataset Y_{kj}, the parameters involve both the accident periods and development periods. The parameters of ODP CC model are α_k and β_j, which represent the ultimate losses for accident period k, and incremental observations as a proportion of ultimate losses for each development period j, respectively. They are estimated using the marginal sum estimation equations (Schmidt and Wünsche, 1998), which calculate the row sum observations and column sum observations and use these values to find the MLEs for the parameters, $\hat{\alpha}_k$ and $\hat{\beta}_j$ by equating the values with the corresponding sum of MLEs. Mathematically, this is expressed as

$$\sum^{R(k)} Y_{kj} = \sum^{R(k)} \hat{\alpha}_k \hat{\beta}_j = \hat{\alpha}_k \sum^{R(k)} \hat{\beta}_j = \hat{\alpha}_k \sum_{j=1}^{J-k+1} \hat{\beta}_j = \hat{\alpha}_k \left[1 - \sum_{j=J-k+2}^{J} \hat{\beta}_j \right]$$

(3-1)

Similarly, for column sums we have

$$\sum^{C(j)} Y_{kj} = \sum^{C(j)} \hat{\alpha}_k \hat{\beta}_j = \hat{\beta}_j \sum^{C(j)} \hat{\alpha}_k$$

(3-2)

The following data come from Table 1-1

Incremental Paid Losses ($000)										
k\j	1	2	3	4	5	6	7	8	9	10
1	41,821	34,729	20,147	15,965	11,285	5,924	4,775	3,742	3,435	2,958
2	48,167	39,495	24,444	18,178	10,840	7,379	5,683	4,758	3,959	
3	52,058	47,459	27,359	17,916	11,448	8,846	5,869	5,391		

The procedure to compute $\hat{\alpha}_k$ and $\hat{\beta}_j$ is using formula (3-1) and (3-2) alternately.

To apply the formulas, we first calculate the value for $\hat{\alpha}_1$ using (3-1)

$$\sum^{\mathrm{R}(1)} Y_{1,j} = 41{,}821 + 34{,}729 + \cdots + 2958 = 144{,}781 = \hat{\alpha}_1 \sum^{R(1)} \hat{\beta}_j = \hat{\alpha}_1$$

Incremental Paid Losses ($000)										
k\j	1	2	3	4	5	6	7	8	9	10
1	41,821	34,729	20,147	15,965	11,285	5,924	4,775	3,742	3,435	2,958
2	48,167	39,495	24,444	18,178	10,840	7,379	5,683	4,758	3,959	
3	52,058	47,459	27,359	17,916	11,448	8,846	5,869	5,391		

And thus we get $\hat{\alpha}_1 = 144{,}781$. Note that in the above calculation we have $\sum^{R(1)} \hat{\beta}_j = 1$ by the condition (EDFCC2c) of the CC model.

With value of $\hat{\alpha}_1$, we can proceed to the second step to compute $\hat{\beta}_{10}$ by applying (3-2):

$$\sum^{C(10)} Y_{k,10} = 2{,}958 = \hat{\beta}_{10} \sum^{C(10)} \hat{\alpha}_k = \hat{\alpha}_1 \hat{\beta}_{10}$$

Incremental Paid Losses ($000)										
k\j	1	2	3	4	5	6	7	8	9	10
1	41,821	34,729	20,147	15,965	11,285	5,924	4,775	3,742	3,435	2,958
2	48,167	39,495	24,444	18,178	10,840	7,379	5,683	4,758	3,959	
3	52,058	47,459	27,359	17,916	11,448	8,846	5,869	5,391		

Note that we just found $\hat{\alpha}_1 = 144{,}781$, and thus we can find $\hat{\beta}_{10}$ as

$$\hat{\beta}_{10} = \frac{2{,}958}{144{,}781} = 0.020$$

Starting from the third step, we have more than one incremental observation to consider for the row and column sums. For the third step, we apply (3-1) again as the following:

$$\sum^{\mathrm{R}(2)} Y_{2,j} = 48{,}167 + 39{,}495 + \cdots + 3{,}959 = 162{,}903 = \hat{\alpha}_2 \sum^{R(2)} \hat{\beta}_j$$
$$= \hat{\alpha}_2(\hat{\beta}_1 + \hat{\beta}_2 + \cdots + \hat{\beta}_9)$$
$$= \hat{\alpha}_2(1 - \hat{\beta}_{10})$$

Incremental Paid Losses ($000)										
k\j	1	2	3	4	5	6	7	8	9	10
1	41,821	34,729	20,147	15,965	11,285	5,924	4,775	3,742	3,435	2,958
2	48,167	39,495	24,444	18,178	10,840	7,379	5,683	4,758	3,959	
3	52,058	47,459	27,359	17,916	11,448	8,846	5,869	5,391		

Because we have found $\hat{\beta}_{10}$ from the second step, for $\hat{\alpha}_2$ we obtain

$$\hat{\alpha}_2 = \frac{\sum^{\mathrm{R}(2)} Y_{2,j}}{(1 - \hat{\beta}_{10})} = \frac{162{,}903}{1 - 0.020} = 166{,}301$$

For the following fourth step, we will need to use both $\hat{\alpha}_1$ and $\hat{\alpha}_2$. For this step, we apply (3-2) again for the following relation of column sum:

$$\sum^{C(9)} Y_{k,9} = 3{,}435 + 3{,}959 = \hat{\beta}_9 \sum^{C(9)} \hat{\alpha}_k = \hat{\beta}_9(\hat{\alpha}_1 + \hat{\alpha}_2)$$

Incremental Paid Losses ($000)										
k\j	1	2	3	4	5	6	7	8	9	10
1	41,821	34,729	20,147	15,965	11,285	5,924	4,775	3,742	3,435	2,958
2	48,167	39,495	24,444	18,178	10,840	7,379	5,683	4,758	3,959	
3	52,058	47,459	27,359	17,916	11,448	8,846	5,869	5,391		

And thus for $\hat{\beta}_9$ we obtain

$$\hat{\beta}_9 = \frac{\sum^{C(9)} Y_{k,9}}{(\hat{\alpha}_1 + \hat{\alpha}_2)} = \frac{3{,}435 + 3{,}959}{144{,}781 + 166{,}301} = 0.024$$

To obtain $\hat{\alpha}_3$, we apply (3-1) for the sum of row 3, and get

$$\sum^{\mathrm{R}(3)} Y_{3,j} = 52{,}058 + 47{,}459 + \cdots + 5{,}391 = 176{,}346 = \hat{\alpha}_3 \sum^{R(3)} \hat{\beta}_j$$
$$= 3(\hat{\beta}_1 + \hat{\beta}_2 + \cdots + \hat{\beta}_8) = \hat{\alpha}_2(1 - \hat{\beta}_9 - \hat{\beta}_{10})$$

Incremental Paid Losses ($000)										
k\j	1	2	3	4	5	6	7	8	9	10
1	41,821	34,729	20,147	15,965	11,285	5,924	4,775	3,742	3,435	2,958
2	48,167	39,495	24,444	18,178	10,840	7,379	5,683	4,758	3,959	
3	52,058	47,459	27,359	17,916	11,448	8,846	5,869	5,391		

and

$$\hat{\alpha}_3 = \frac{\sum^{\mathrm{R}(3)} Y_{3,j}}{(1 - \hat{\beta}_9 - \hat{\beta}_{10})} = \frac{176{,}346}{1 - 0.024 - 0.020} = 184{,}501$$

And to obtain $\hat{\beta}_8$, we apply (3-2) once again:

$$\sum^{C(8)} Y_{k,8} = 3{,}742 + 4{,}758 + 5{,}391 = \hat{\beta}_8 \sum^{C(8)} \hat{\alpha}_k = \hat{\beta}_8(\hat{\alpha}_1 + \hat{\alpha}_2 + \hat{\alpha}_3)$$

Incremental Paid Losses ($000)										
k\j	1	2	3	4	5	6	7	8	9	10
1	41,821	34,729	20,147	15,965	11,285	5,924	4,775	3,742	3,435	2,958
2	48,167	39,495	24,444	18,178	10,840	7,379	5,683	4,758	3,959	
3	52,058	47,459	27,359	17,916	11,448	8,846	5,869	5,391		

and

$$\hat{\beta}_8 = \frac{\sum^{C(8)} Y_{k,8}}{(\hat{\alpha}_1 + \hat{\alpha}_2 + \hat{\alpha}_3)} = \frac{3{,}742 + 4{,}758 + 5{,}391}{144{,}781 + 166{,}301 + 184{,}501} = 0.028$$

Repeating the steps, we will get the results in Table 3-3 on the following page:

Parameter Estimates for ODP CC Model		
j/k	$\hat{\alpha}_k$	$\hat{\beta}_J$
1	144,781	0.293
2	166,301	0.239
3	184,501	0.139
4	201,845	0.106
5	212,151	0.069
6	207,340	0.047
7	205,725	0.035
8	182,904	0.028
9	173,225	0.024
10	149,836	0.020

Table 3-3

After obtaining the estimated parameters, one can find the future estimates using these parameters. For example, for the same future loss we calculated in the previous example, $\hat{Y}_{9,3}$, using ODP CC model this would simply be

$$\hat{Y}_{3,9} = \hat{\alpha}_3\hat{\beta}_9 = 184{,}501 \times 0.024 = 4{,}385$$

Incremental Paid Losses ($000)										
k\j	1	2	3	4	5	6	7	8	9	10
1	41,821	34,729	20,147	15,965	11,285	5,924	4,775	3,742	3,435	2,958
2	48,167	39,495	24,444	18,178	10,840	7,379	5,683	4,758	3,959	
3	52,058	47,459	27,359	17,916	11,448	8,846	5,869	5,391		

which is consistent with the result from the conventional chain ladder algorithm and the ODP Mack model. Similarly, we can also check other future estimates of Y_{kj} and find them all in agreement with the results from chain ladder algorithm and ODP Mack model, which reinstates that ODP Mack and ODP CC models yield the same estimates for outstanding losses.

By comparing the parameters in ODP Mack and ODP CC models, we can identify the special one-to-one relation between the two models (Verrall 2000), which is

$$\hat{f}_J = \frac{\sum_{i=1}^{j+1}\hat{\beta}_i}{\sum_{i=1}^{j}\hat{\beta}_i}$$

(3-3)

or equivalently,

$$\hat{\beta}_{j+1} = \left(\hat{f}_j - 1\right)\frac{\prod_{r=1}^{j-1}\hat{f}_r}{\prod_{r=1}^{J-1}\hat{f}_r}$$

(3-4)

For example, for $\hat{f}_3 = 1.158$, we would get the following equation from (3-3):

$$\hat{f}_3 = \frac{\sum_{i=1}^{4}\hat{\beta}_i}{\sum_{i=1}^{3}\hat{\beta}_i} = \frac{0.293 + 0.239 + 0.139 + 0.106}{0.293 + 0.239 + 0.139} = 1.158$$

For $\hat{\beta}_4 = 0.106$, we would get the following from (3-4):

$$\hat{\beta}_4 = \left(\hat{f}_3 - 1\right)\frac{\prod_{r=1}^{2}\hat{f}_r}{\prod_{r=1}^{9}\hat{f}_r} = (1.158 - 1)\frac{1.815 \times 1.261}{1.815 \times 1.261 \times \ldots \times 1.021} = 0.106$$

Where the portion $\left(\hat{f}_3 - 1\right)\prod_{r=1}^{2}\hat{f}_r$ can be understood as the proportion corresponding to Y_4 and $\prod_{r=1}^{9}\hat{f}_r$ as the proportion corresponding to X_{10}, or the ultimate cumulative loss.

3.3 GLM Representation of Chain Ladder Models

3.3.1 ODP Mack Model

In section 3.1.2, we mentioned that the ODP Mack Model is a specific case of Parametric Mack Models with the condition $Y_{k,j+1}|X_{kj} \sim EDF(\theta_{kj}, \phi_{kj}, a, b, c)$ replaced by

$$Y_{k,j+1}|X_{kj} \sim ODP(\mu_{kj}, \phi_{kj})$$

Consider the following ODP Mack model:

$$Y_{k,j+1}|X_{kj} \sim ODP\left(\left(f_j - 1\right)X_{kj}, \phi_{kj}\right)$$

(3-5)

Note that in this model $E\left[Y_{k,j+1}\right] = \mu_{kj} = \left(f_j - 1\right)X_{kj}$. This is derived from the following:

$$Y_{kj} = X_{k,j+1} - X_{kj} = f_j X_{kj} - X_{kj} = \left(f_j - 1\right)X_{kj}$$

(Note: Recall from Chapter 2 that the distribution of ODP model is (2-14):

$$\pi(y;\mu,\phi) = \mu^{\frac{y}{\phi}} e^{\left[-\frac{\mu}{\phi}+c(y,\phi)\right]}, for\ y = 0, \phi, 2\phi, \dots and\ \mu = e^{\theta}$$

And the Over-Dispersed Poisson Sub-Family is in the Tweedie Sub-Family with p=1.)

On top of (3-5), we add the following condition

$$\phi_{kj} = \phi_j, independent\ of\ k$$

(3-6)

so that the dispersion parameter ϕ doesn't depend on k, which was a pre-requisite to ensure that the MLEs of f_j in this ODP Mack Model are chain ladder estimates.

With (3-5) and (3-6), we obtain

$$Y_{k,j+1}|X_{kj} \sim ODP\left((f_j - 1)X_{kj}, \phi_j\right)$$

(3-7)

Where we replaced ϕ_{kj} in (3-5) with ϕ_j.

From formula (1-4) in Chapter 1, we know that

$$E\left[Y_{k,j+1}|X_{kj}\right] = (f_j - 1)X_{kj}$$

If we replace $Y_{k,j+1}$ with $\hat{f}_{kj} - 1$ to be the variable, with the relation $\hat{f}_{kj} - 1 = Y_{k,j+1}/X_{kj}$, then we obtain

$$E\left[\hat{f}_{kj} - 1|X_{kj}\right] = f_j - 1$$

(3-8)

(This is true because the expected value equation still holds for dividing both sides by X_{kj})

Also,

$$Var\left[\hat{f}_{kj} - 1|X_{kj}\right] = \frac{Var\left[Y_{k,j+1}\middle|X_{kj}\right]}{X_{kj}^2} = \frac{\phi_j(f_j - 1)X_{kj}}{X_{kj}^2} = \frac{\phi_j(f_j - 1)}{X_{kj}}$$

(3-9)

Where ϕ_j is the dispersion parameter. Note that this suggests $\hat{f}_{kj} - 1|X_{kj}$ follows an ODP distribution with $\mu = f_j - 1$ and $\phi = \frac{\phi_j}{X_{kj}}$ because the variance of a ODP model is $Var[Y] = \phi\mu$. (This is from Chapter 2 when we discussed the Tweedie sub-family and $Var[Y] = \phi\mu^p$, while in ODP situation, $p = 1$.)

Because ODP family is known to be closed under scaling, which means an ODP variate is still an ODP variate after it's divided by some constant, therefore,

$$f_{kj} - 1|X_{kj} \sim ODP(f_j - 1, \frac{\phi_j}{X_{kj}})$$

(3-10)

For the purpose of developing the GLM, the expected values of estimated $\hat{f}_{kj} - 1|X_{kj}$ are sometimes expressed in the following form:

$$E\left[\hat{f}_{kj} - 1\middle|X_{kj}\right] = \sum_{i=1}^{9} (f_i - 1)\delta_{ji}$$

(3-11)

This expression is usually used for GLM software calculation, where δ_{ji} is called the Kronecker delta which has the value of 1 when $i = j$, and equals 0 otherwise.

(In words, 3-11 is the summation of multiple 0s, and an $f_j - 1$. The purpose of this complex model is to get all the f_i involved in the GLM formula, and the regression can thus estimate all the parameters at once. An example will be shown in section 3.3.3)

Note that with the setting of (3-10), the model includes ϕ_j with unknown values. The following argument will show that the values of ϕ_j are not required for the purpose of estimating f_{i-1}

To obtain the MLE of f_j, we start with the log-likelihood of the claims trapezoid $\mathfrak{D}_K$

$$\ell(\mathfrak{D}_K) = \sum_{\mathfrak{D}_K, j \neq 1} \ell(\hat{f}_{k,j-1} - 1)$$

(log-likelihood of the trapezoid equals the sum of the log-likelihood for all entries)

Recall for ODP,

$$\pi(y;\mu,\phi) = \mu^{\frac{y}{\phi}} e^{\left[-\frac{\mu}{\phi}+c(y,\phi)\right]}, for\ y = 0,\phi,2\phi,\dots and\ \mu = e^{\theta}$$

where $e^{c(y,\phi)} = \left[\left(\frac{y}{\phi}\right)!\right]^{-1}$

here in this example, $\hat{f}_{k,j-1} - 1 \sim ODP(f_{j-1} - 1, \frac{\phi_{j-1}}{X_{k,j-1}})$

also, $\hat{f}_{k,j-1} - 1 = \frac{Y_{kj}}{X_{k,j-1}}$

Thus, for variable $(\hat{f}_{k,j-1} - 1)$

$$\ell(\mathfrak{D}_K) = \sum_{\mathfrak{D}_K, j\neq 1} \ell\left(\hat{f}_{k,j-1} - 1\right)$$

$$= \sum_{\mathfrak{D}_K, j\neq 1} \ell\left((f_{j-1}-1)^{\frac{\hat{f}_{k,j-1}-1}{\phi_{j-1}/X_{k,j-1}}} e^{-\frac{f_{j-1}-1}{\phi_{j-1}/X_{k,j-1}}} \times \left[\left(\frac{\hat{f}_{k,j-1}-1}{\frac{\phi_{j-1}}{X_{k,j-1}}}\right)!\right]^{-1}\right)$$

(Plug in $y = f_{k,j-1} - 1, \mu = f_{j-1} - 1, \phi = \phi_{j-1}/X_{k,j-1}$)

$$= \sum_{\mathfrak{D}_K, j\neq 1} \ell\left((f_{j-1}-1)^{\frac{Y_{kj}/X_{k,j-1}}{\phi_{j-1}/X_{k,j-1}}} e^{-\frac{f_{j-1}-1}{\phi_{j-1}/X_{k,j-1}}} \times \left[\left(\frac{Y_{kj}/X_{k,j-1}}{\frac{\phi_{j-1}}{X_{k,j-1}}}\right)!\right]^{-1}\right)$$

(Replaced $\hat{f}_{k,j-1} - 1\ with\ Y_{kj}/X_{k,j-1}$)

$$= \sum_{\mathfrak{D}_K, j\neq 1} \ell\left((f_{j-1}-1)^{\frac{Y_{kj}}{\phi_{j-1}}} * e^{-\frac{f_{j-1}-1}{\phi_{j-1}/X_{k,j-1}}} \times \left[\left(\frac{Y_{kj}}{\phi_{j-1}}\right)!\right]^{-1}\right)$$

$$= \sum_{\mathfrak{D}_K, j\neq 1} \left(\frac{Y_{kj}}{\phi_{j-1}} \ln(f_{j-1}-1) - \frac{f_{j-1}-1}{\frac{\phi_{j-1}}{X_{k,j-1}}} - \ln\left[\left(\frac{Y_{kj}}{\phi_{j-1}}\right)!\right]\right)$$

$$= \sum_{\mathfrak{D}_K, j\neq 1} \left(\frac{Y_{kj}}{\phi_{j-1}} \ln(f_{j-1}-1) - \frac{f_{j-1}-1}{\frac{\phi_{j-1}}{X_{k,j-1}}} - \ln\left[\left(\frac{Y_{kj}}{\phi_{j-1}}\right)!\right]\right)$$

$$= \sum_{\mathfrak{D}_K, j \neq 1} \left(\frac{\frac{Y_{kj}}{X_{k,j-1}} \ln(f_{j-1} - 1) - (f_{j-1} - 1)}{\frac{\phi_{j-1}}{X_{k,j-1}}} - \ln\left[\left(\frac{Y_{kj}}{\phi_{j-1}}\right)!\right] \right)$$

(3-12)

Then, to find MLE of f_j, we take the partial derivative of f_{i-1}, replace *j* with *i*:

$$\frac{\partial \ell(\mathfrak{D}_K)}{\partial f_{i-1}} = \frac{\partial}{\partial f_{i-1}} \left(\sum_{\mathfrak{D}_{K, i \neq 1}} \left(\frac{\frac{Y_{ki}}{X_{k,i-1}} \ln(f_{i-1} - 1) - (f_{i-1} - 1)}{\frac{\phi_{i-1}}{X_{k,i-1}}} - \ln\left[\left(\frac{Y_{ki}}{\phi_{i-1}}\right)!\right] \right) \right)$$

$$= \frac{\partial}{\partial f_{i-1}} \left(\sum_{\mathfrak{D}_{K, i \neq 1}} \left(\frac{\frac{Y_{ki}}{X_{k,i-1}} \ln(f_{i-1} - 1) - (f_{i-1} - 1)}{\frac{\phi_{i-1}}{X_{k,i-1}}} \right) \right)$$

$$= \phi_{i-1} \sum_{\mathfrak{D}_{(k,i) \epsilon \mathcal{C}(i)}} \frac{\partial}{\partial f_{i-1}} \left(Y_{ki} \ln(f_{i-1} - 1) - X_{k,i-1}(f_{i-1} - 1) \right)$$

$$= \phi_{i-1} \sum_{\mathfrak{D}_{(k,i) \epsilon \mathcal{C}(i)}} \left(\frac{Y_{ki}}{f_{i-1} - 1} - X_{k,i-1} \right)$$

($\frac{1}{\phi_{i-1}}$ is moved out because it doesn't change with respect to f_{i-1}, use $\mathcal{C}(i)$ instead of $\mathfrak{D}_K$ to use only the column that depends on f_{i-1}, drop $\ln\left[\left(\frac{Y_{ki}}{\phi_{i-1}}\right)!\right]$ because it doesn't depend on f_{i-1})

$$set = 0$$

Then,

$$\sum_{(k,i) \epsilon \mathcal{C}(i)} \left(\left[\frac{Y_{ki}}{f_{i-1} - 1} - X_{k,i-1} \right] \right) = 0$$

Which is true because $Y_{ki}/f_{i-1} - 1 = X_{k,i-1}$, for all $(k, i) \epsilon \mathcal{C}(i)$.

Then, separate the summation into 2 parts

$$\sum_{(k,i)\epsilon\mathcal{C}(i)}\left(\frac{Y_{ki}}{f_{i-1}-1}\right)-\sum_{(k,i)\epsilon\mathcal{C}(i)}\left(X_{k,i-1}\right)=0$$

$$\sum_{(k,i)\epsilon\mathcal{C}(i)}\left(\frac{Y_{ki}}{f_{i-1}-1}\right)=\sum_{(k,i)\epsilon\mathcal{C}(i)}\left(X_{k,i-1}\right)$$

For a specific column *i* and changing *k*, $f_{i-1}-1$ can be taken out from the summation:

$$\frac{1}{f_{i-1}-1}\sum_{(k,i)\epsilon\mathcal{C}(i)}\left(Y_{kj}\right)=\sum_{(k,i)\epsilon\mathcal{C}(i)}\left(X_{k,i-1}\right)$$

More specifically

$$f_{i-1}-1=\frac{\sum_{k=1}^{k-i+1}Y_{ki}}{\sum_{k=1}^{k-i+1}X_{k,i-1}}$$

$$\hat{f}_{i-1_{MLE}}=\frac{\sum_{k=1}^{k-i+1}Y_{ki}}{\sum_{k=1}^{k-i+1}X_{k,i-1}}+1$$

Which makes sense because these are the weighted average $\hat{f}_{i-1}$ we get from the chain ladder method. Also, this argument shows that the dispersion function ϕ doesn't affect the estimation of f_{i-1}

GLM of ODP Mack Model:

With the help of SAS, we can easily obtain parameter estimates for the Mack model and CC model. We use the same dataset as analytical computation for SAS algorithm. For ODP Mack model, let

$$\mu = the\ vector\ of\ Y_{kj}$$

Here, μ is not a matrix, because we are using each Y_{kj} as observations of the dependent variable Y for the purpose of GLM, such that

$$\mu=\left(Y_{1,1},Y_{1,2},\dots,Y_{1,J},Y_{2,1}\dots,Y_{2,J-1},\dots Y_{K,1}\right)^T$$

And the vector of $f_1\ to\ f_9$, denoted by β, is what we want to estimate through the GLM

$$\beta=(f_1,f_2,\dots,f_9)^T$$

Then, for the purpose to include all f_1 to f_9 in the regression equation, while only utilizing one of them in each observation, we use the design matrix X to satisfy this purpose.

Appendix A shows a completed design matrix X of ODP Mack GLM when *K*=10 and *J*=10.

With all the setup, we can write the regression function and run the GLM:

$$\mu = h^{-1}(X\beta)$$

(3-14)

Here, please note that there are two methods to bring about the regression. The first one uses calculated f_{kj} as μ, and design matrix X with only 1s and 0s. The second one uses X_{kj} as μ, and design matrix X with values of $X_{k,j-1}$ in place of the 1s. They produce slightly different results but the idea is very similar.

The idea behind the first method is, with the setup, we have 45 observations, or 45 equations, which are:

$$f_{1,1} = f_1 \times 1 + f_2 \times 0 + \cdots + f_9 \times 0$$

$$f_{1,2} = f_1 \times 0 + f_2 \times 1 + \cdots + f_9 \times 0$$

…

$$f_{1,9} = f_1 \times 0 + f_2 \times 0 + \cdots + f_9 \times 1$$

…

$$f_{8,1} = f_1 \times 1 + f_2 \times 0 + \cdots + f_9 \times 0$$

$$f_{8,2} = f_1 \times 0 + f_2 \times 1 + \cdots + f_9 \times 0$$

$$f_{9,1} = f_1 \times 1 + f_2 \times 0 + \cdots + f_9 \times 0$$

Such that the GLM will generate its best estimate of f_1 to f_9 for us. (In this case, the numerical average of $f_{i,j}$ $\forall j$ that applies, to be the estimate of f_i)

The idea behind the second method is, with the setup, we have 45 observations, or 45 equations, which are:

$$X_{1,2} = f_1 \times X_{1,1} + f_2 \times 0 + \cdots + f_9 \times 0$$

$$X_{1,3} = f_1 \times 0 + f_2 \times X_{1,2} + \cdots + f_9 \times 0$$

…

$$X_{1,10} = f_1 \times 0 + f_2 \times 0 + \cdots + f_9 \times X_{1,9}$$

…

$$X_{8,2} = f_1 \times X_{8,1} + f_2 \times 0 + \cdots + f_9 \times 0$$

$$X_{8,3} = f_1 \times 0 + f_2 \times X_{8,2} + \cdots + f_9 \times 0$$

$$X_{9,2} = f_1 \times X_{9,1} + f_2 \times 0 + \cdots + f_9 \times 0$$

And the GLM will generate its best estimate of f_1 to f_9 for us (In this case, an estimate from $\frac{X_{ij}}{X_{i,j-1}} = f_{i,j-1}$ for $j \geq 2$ and all $i \epsilon \mathcal{C}(j)$, to be the estimate of f_{j-1}).

3.3.2 ODP Cross-Classified Model

Recall that a Cross-Classified Model has the condition of $Y_{kj} \sim EDF(\theta_{kj}, \phi_{kj}; a, b, c)$

Here, we modify it to its ODP form, with $\mu_{kj} = \alpha_k \beta_j$ being the expected value, and $\mu_{kj}\phi_{kj} = a_k b_j \phi_{kj}$ being the variance, such that

$$Y_{kj} \sim ODP\left(\alpha_k \beta_j, \phi_{kj}\right)$$

(3-15)

If we add the further condition to set the dispersion function to be a constant

$$\phi_{kj} = \phi$$

(3-16)

Then,

$$Y_{kj} \sim ODP(\alpha_k \beta_j, \phi) = ODP(\mu_{kj}, \phi)$$

(3-17)

where

$$\mu_{kj} = \alpha_k \beta_j = \exp(\ln \alpha_k + \ln \beta_j)$$

(3-18)

The exponential function and ln function are used to convert multiplication to summation for the purpose of GLM estimation, because GLM uses the summation of $coefficient \times variable$, not multiplications.

Somewhat similar to the ODP Mack Model, we use a design matrix X for a GLM estimation of CC model.

The idea behind this design matrix is to

1. Include all α_k and β_j in the regression
2. Use only <u>one</u> $\alpha_k$and <u>one</u> β_j for each observation

A completed design matrix X of ODP CC GLM when K=10 and J=10 is shown in **Appendix B**.

The idea behind this GLM is that, with the setup, we have 55 observations, or 55 equations, they are:

$$Y_{1,1} = \exp(\ln \alpha_1 + \ln \beta_1) = \alpha_1 \beta_1$$

$$Y_{1,2} = \exp(\ln \alpha_1 + \ln \beta_2) = \alpha_1 \beta_2$$

…

$$Y_{1,10} = \exp(\ln \alpha_1 + \ln \beta_{10}) = \alpha_1 \beta_{10}$$

$$Y_{2,1} = \exp(\ln \alpha_2 + \ln \beta_1) = \alpha_2 \beta_1$$

…

$$Y_{10,1} = \exp(\ln \alpha_{10} + \ln \beta_1) = \alpha_{10} \beta_1$$

Such that the GLM will generate its best estimate of all α_k and β_j for us.

Note that β_j are essentially the proportions of ultimate losses that occur in each development period. However, because most software does not automatically normalize $\hat{\beta}_j$ to make $\sum_{j=1}^{J}\hat{\beta}_j = 1$, we will first use $\hat{\beta}_1 = 1$, or $\ln(\hat{\beta}_1) = 0$ as a standard to generate the other β_j, and then normalize them. The normalizing can be done by replacing each $\hat{\beta}_j$ with $\hat{\beta}_j / \sum_{j=i}^{J}\hat{\beta}_i$. After normalizing, $\hat{\alpha}_k$ becomes the expected values of ultimate losses for year k.

With our estimates for $\hat{\alpha}_k$ and $\hat{\beta}_j$, we can then estimate future incremental losses using

$$\hat{Y}_{kj} = \left[\hat{\alpha}_k \sum_{i=1}^{J}\hat{\beta}_i\right]\left[\frac{\hat{\beta}_j}{\sum_{j=i}^{J}\hat{\beta}_i}\right] = \hat{\alpha}_k\hat{\beta}_j$$

The middle step is what we can get directly from the GLM software, which returns the $\hat{\beta}_j$ without normalizing, and those ratios are multiplied to $\hat{\alpha}_k$.

After the GLM estimation, we can estimate any future incremental losses with our estimated $\hat{\alpha}_k$ and $\hat{\beta}_j$.

3.3.3 Numerical Example

To align with the monograph *Stochastic Loss Reserving Using Generalized Linear Models*, we use the GLM procedure GENMOD in SAS to generate our estimation.

ODP Mack Model:

For ODP Mack Model, we use two different ways to generate f_j:

The first one use f_{kj} as dependent observations, and 0-1 design matrix (see **Appendix C**).

The second one use Y_{kj} as dependent observations, and in the design matrix, we use the corresponding previous cumulative losses $X_{k,j-1}$, instead of 1, to be the values of the x variates (see **Appendix C**).

For the first method, note that among the inputs, only the 3 lines

```
proc genmod data=ODPMackModelOneZero;

model Y = f1 f2 f3 f4 f5 f6 f7 f8 f9 / NOINT SCALE = PEARSON;

run;
```

are calculation of f_1 to f_9, the previous parts are all inputs for data and the design matrix.

Notice that we use the options NOINT to remove intercept from our regression equation, as our model does not include an intercept.

Also, be careful that the f1 to f9 used in the regression formula are actually the x variates that correspond to f_1 to f_9, and f_1 to f_9 are actually the coefficients going with these x variates. We use f1 to f9 here to make our result easier to read.

With the above input, we obtain the following result:

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	0	0.0000	0.0000	0.0000	0.0000	.	.
f1	1	1.8174	0.0140	1.7900	1.8448	16912.4	<.0001
f2	1	1.2619	0.0148	1.2329	1.2910	7248.17	<.0001
f3	1	1.1583	0.0158	1.1272	1.1894	5343.26	<.0001
f4	1	1.0887	0.0171	1.0551	1.1222	4045.87	<.0001
f5	1	1.0550	0.0187	1.0182	1.0917	3166.02	<.0001
f6	1	1.0384	0.0210	0.9973	1.0795	2454.00	<.0001
f7	1	1.0301	0.0242	0.9826	1.0775	1810.96	<.0001
f8	1	1.0249	0.0296	0.9668	1.0830	1195.16	<.0001
f9	1	1.0209	0.0419	0.9387	1.1030	592.91	<.0001
Scale	0	0.0419	0.0000	0.0419	0.0419		

This result corresponds to Table 3-4 in the monograph. Notice that the results are slightly different from the results in the monograph. The reason is that our code uses the numerical average of f_{kj} in column j to be our estimated $\hat{f}_j$, while the monograph uses weighted average. Therefore, the difference comes from different weighting methods for each observation in column j, but the idea is the same.

For the second method, the idea behind this method is similar to the first method. The major difference is we use X_{kj} instead of f_{kj}, and use $X_{k,j-1}$ to replace the 1s in the design matrix.

Result is as follows:

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	0	0.0000	0.0000	0.0000	0.0000	.	.
f1	1	1.8121	0.0175	1.7778	1.8465	10702.2	<.0001
f2	1	1.2600	0.0101	1.2401	1.2798	15420.6	<.0001
f3	1	1.1580	0.0086	1.1412	1.1747	18269.7	<.0001
f4	1	1.0881	0.0081	1.0723	1.1040	18022.7	<.0001
f5	1	1.0464	0.0083	1.0302	1.0627	15997.5	<.0001
f6	1	1.0388	0.0092	1.0207	1.0569	12654.1	<.0001
f7	1	1.0304	0.0109	1.0091	1.0516	8997.88	<.0001
f8	1	1.0249	0.0137	0.9979	1.0518	5558.85	<.0001
f9	1	1.0209	0.0204	0.9808	1.0609	2497.73	<.0001
Scale	0	2896.933	0.0000	2896.933	2896.933		

This result also corresponds to Table 3-4 in the monograph. Again, the $\hat{f}_j$ estimated here are slightly different from what the monograph has, because of different weights used for the entries in column j.

ODP Cross-Classified Model:

The SAS coding for ODP CC model is shown in **Appendix D**.

Here, the options dist = poisson, and SCALE = PEARSON are all options to give the model the properties of ODP.

The option link = log is to use $\ln y$ instead of y such that the underlying equation is

$$\ln Y_{kj} = \ln \alpha_k + \ln \beta_j$$

Also, notice that in the following line of the codes

model y = a1 a2 a3 a4 a5 a6 a7 a8 a9 a10 b2 b3 b4 b5 b6 b7 b8 b9 b10 /

b1 is not used in the regression formula. The reason behind is to make ln β_1 equals 0 such that $\beta_1 = 1$. Be careful that the b1 in our code has values 1 or 0, but that's not actually β_1, but the x variate corresponds to β_1.

Again, we use the option NOINT to remove the intercept from our regression equation because ODP CC model doesn't include an intercept.

The result is as follows:

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	0	0.0000	0.0000	0.0000	0.0000	.	.
a1	1	10.6454	0.4830	9.6988	11.5920	485.84	<.0001
a2	1	10.7931	0.4830	9.8465	11.7397	499.41	<.0001
a3	1	10.8875	0.4901	9.9269	11.8481	493.50	<.0001
a4	1	11.0165	0.5021	10.0324	12.0006	481.40	<.0001
a5	1	11.0387	0.5196	10.0203	12.0570	451.37	<.0001
a6	1	11.0166	0.5446	9.9492	12.0841	409.15	<.0001
a7	1	10.9999	0.5816	9.8599	12.1399	357.67	<.0001
a8	1	10.8906	0.6401	9.6360	12.1452	289.48	<.0001
a9	1	10.8337	0.7454	9.3728	12.2946	211.26	<.0001
a10	1	10.6911	1.0000	8.7311	12.6510	114.30	<.0001
b2	1	-0.2069	0.4714	-1.1309	0.7170	0.19	0.6607
b3	1	-0.7450	0.4930	-1.7112	0.2213	2.28	0.1308
b4	1	-1.0156	0.5166	-2.0281	-0.0032	3.87	0.0493
b5	1	-1.4512	0.5446	-2.5187	-0.3838	7.10	0.0077
b6	1	-1.8452	0.5802	-2.9823	-0.7080	10.11	0.0015
b7	1	-2.1486	0.6285	-3.3803	-0.9168	11.69	0.0006
b8	1	-2.3462	0.7001	-3.7184	-0.9739	11.23	0.0008
b9	1	-2.5065	0.8232	-4.1200	-0.8930	9.27	0.0023
b10	1	-2.6532	1.1105	-4.8297	-0.4766	5.71	0.0169
Scale	0	1.0000	0.0000	1.0000	1.0000		

The result generated matches the result in Table 3-5 of the monograph. Try it out and see if you can reproduce these 3 GLMs.

3.4 Minor Variations of Chain Ladder

The chain ladder algorithm we used contains no flexibility. In this section, we will discuss some variation of the chain ladder method.

3.4.1 Reliance on Only Recent Experience Years

Because recent observations can represent the current situation better than observations from many years ago, we can adjust our model to put more weights on more recent observations, or in this example, only give weights to observations in the recent m years.

If we only use observations in the recent m years, the observations we use are

$$\hat{f}_{kj} - 1 | X_{kj}$$

With k and j that satisfies:

$$k < K, \qquad k \epsilon \mathbb{N}$$

$$j < J, \qquad j \epsilon \mathbb{N}$$

and

$$K + 1 - m \leq k + j \leq K$$

The first part of the third inequality $K + 1 - m \leq k + j$ ensures that only those data after the calendar year $K + 1 - m$ are used in our model.

We can also write it as the following:

$$\omega_{kj} = X_{kj} I(K + 1 - m \leq k + j \leq K)$$

(3-20)

where $I(.)$ is an indicator function which equals 1 when the condition is satisfied, and 0 otherwise:

$$I(c) = \begin{cases} 1, & if\ the\ logical\ condition\ c\ is\ true \\ 0, & otherwise \end{cases}$$

(3-21)

With the indicator function, while also omitting the $c(y,\phi)$ member because it vanishes when partial derivative is taken with respect to $f_{j-1}-1$, the log-likelihood function of $\mathfrak{D}_K$ becomes:

$$\ell(\mathfrak{D}_K)=\sum_{\mathfrak{D}_K,j\neq 1} I(K+1-m\leq k+j\leq K)\left(\frac{\frac{Y_{kj}}{X_{k,j-1}}\ln\left(f_{j-1}-1\right)-\left(f_{j-1}-1\right)}{\frac{\phi_{j-1}}{X_{k,j-1}}}\right)$$

(3-22)

(recall in 3-12:

$$\ell(\mathfrak{D}_K)=\sum_{\mathfrak{D}_K,j\neq 1}\left(\frac{\frac{Y_{kj}}{X_{k,j-1}}\ln\left(f_{j-1}-1\right)-\left(f_{j-1}-1\right)}{\frac{\phi_{j-1}}{X_{k,j-1}}}-\ln\left[\left(\frac{Y_{kj}}{\phi_{j-1}}\right)!\right]\right)$$

The member $-\ln\left[\left(\frac{Y_{kj}}{\phi_{j-1}}\right)!\right]$ comes from the $c(y,\phi)$ function of the ODP model.)

3.4.2 Outlier Observations

Similar to giving weights of 0 to observations from many years ago, we can also give a weight of 0 to any **outlier observations** we don't want to include for our purpose of GLM estimation. By giving a weight of 0 to outlier observations, they are excluded from the model fitting process.

4. Prediction Error

Estimations with GLMs usually contain errors. The errors can be broken down into three components: parameter errors, process errors, and model errors.

In this chapter, we introduce these three types of error, with a focus on parameter error and process error, which are usually more tractable than model error.

Mean square error of prediction, goodness-of-fit of a model, and information criteria are also discussed in this chapter. A key takeaway from this section of the chapter is that an increase in goodness-of-fit does not imply reduced forecast error, and penalties are applied for an increase in the number of parameters.

The introduction to prediction error in this chapter is related to, but not limited to loss reserve application. In the next chapter, we will introduce methods related to estimating prediction error for outstanding loss.

4.1. Parameter Error and Process Error

In order to demonstrate the concepts for the different components of a prediction error, we will start with the following example, unrelated to loss reserve:

Example: Suppose we want to predict the probability of getting heads when flipping a fair coin, and assume the true probability, ½, is unknown. We can achieve this by flipping the coin multiple times for multiple trials and compute the average. Suppose we flip the coin 1000 times for each trial, for a total of 6 trials, and we get the following result:

Trial #	1	2	3	4	5	6
# of heads	496	533	521	499	498	513

Which returns an average probability of 0.51.

Process error: For the 6 trials of flipping a coin, denote the observations, the number of heads, as Y, which is a function of the total number of flips, dependent on the true probability of getting a head for each flip. Denote our parameter, the true probability of getting a head when flipping a fair coin as θ, which we know intuitively is ½. Our model can thus be written as

$$Y = n \times \theta$$

However, note that in actual trials the number is usually not exactly as expected. For trial 1, for example, the actual number of heads is 496, whereas using the model, we should expect $Y = 1000 \times \frac{1}{2} = 500$. The difference between the actual number of heads and expected number of heads, $(496 - 500)$, is our **process error**, or **noise**.

Parameter estimation: Suppose we do not know the true probability of getting a head for a flip. We can estimate it using the trials, by dividing the number of heads by the total number of flips for each trial, and obtain the following result:

Trial #	1	2	3	4	5	6
Prob. of head	.496	.533	.521	.499	.498	.513

By taking the average, we estimate that the probability is

$$\hat{\theta} = .51$$

The $\hat{\theta}$ is thus our **parameter estimate**. Using this estimated probability, suppose we want to estimate the number of heads if we flip 500 times. Then the estimated value would be

$$\hat{Y} = .51 \times 500 = 255$$

Parameter error and prediction error: Now we have the estimated parameter $\hat{\theta} = .51$. Using this parameter and trial 1 as an example, we should get $1000 \times 0.51 = 510$ heads in trial 1. Instead, we have 496 heads for trial one. The difference between the actual number of heads and the number of heads we would get in theory using $\hat{\theta}$ is called the **prediction error** associated with trial 1, which can be written as

$$e = Y - \hat{Y} = 496 - 510 = (500 - 510) + (496 - 500)$$

The first part prediction error,$(500 - 510)$, is the difference between the number of heads we should get using the true value of parameter, 1/2, and the parameter estimate of 0.51. Thus, it is called the **parameter error**.

4.1.1. Individual Observations

We now introduce the concept for the prediction error and each component.

Suppose the model used for estimating future claims is loosely defined as follows:

$$Y_{kj} = u(k, j; \theta) + \varepsilon_{kj}, \qquad for\ Y_{kj} \in D_K^+$$

(4-1)

where u is some function of accident period k and development period j, dependent on a parameter vector $\theta = [\theta_1\ \theta_2\ ...]^T$, with stochastic error or noise, ε_{kj} for each observation Y_{kj}. The expected value or center of the noise should be 0, i.e.,

$$E[\varepsilon_{kj}] = 0$$

(4-2)

Recall in the example, for trial 1 the actual number of heads is 496, where we should expect 500. In this case the difference of $(496 - 500)$ is our noise.

Suppose that the model has been calibrated against the data set D_K by some method, and a vector of parameter estimate $\hat{\theta}$ is returned. Then we can define our fitted values and future estimates as

$$\hat{Y}_{kj} = u(k, j; \hat{\theta}), \qquad for\ Y_{kj} \in D_K^+$$

(4-3)

where observations in past dataset, $\hat{Y}_{kj} \in D_K$, are the fitted values, and estimated observations associated with future dataset, $\hat{Y}_{kj} \in D_K^c$, are the estimated outstanding losses.

Recall in the example, we estimated the parameter to be $\hat{\theta} = .51$. Using this approximated probability our fitted value for each trial of 1000 flips should be 510 heads. When we also want to estimate the number of heads if we flip 500 times, our $\hat{Y}$ becomes $\hat{Y} = .51 \times 500 = 255$.

Prediction error is the difference between the actual observation and the associated fitted value, i.e.

$$e_{kj} = Y_{kj} - \hat{Y}_{kj} = [u(k, j; \theta) - u(k, j; \hat{\theta})] + \varepsilon_{kj}$$

(4-4)

Note that from (4-1) and (4-2), we obtain the following result

$$E\left[Y_{kj}\right] = E\left[u(k,j;\theta) + \varepsilon_{kj}\right] = E[u(k,j;\theta)] + E\left[\varepsilon_{kj}\right] = E[u(k,j;\theta)] + 0 = u(k,j;\theta)$$

which summarizes to

$$E\left[Y_{kj}\right] = u(k,j;\theta)$$

(4-5)

Thus we can derive (4-4) into the following form

$$e_{kj} = \left[\mu_{kj} - \hat{Y}_{kj}\right] + \varepsilon_{kj}$$

(4-6)

where $\mu_{kj} = E\left[Y_{kj}\right]$.

In this format we are representing the prediction error as the sum of parameter error and process error, i.e.,

$$prediction\ error = parameter\ error + process\ error$$

The parameter error associated with forecast $\hat{Y}_{kj}$ is the first term $\left[\mu_{kj} - \hat{Y}_{kj}\right]$, and the remaining term ε_{kj} is the associated process error, or noise.

In our example, the prediction error associated with trial 1 using (4-6) is written as

$$e = Y - \hat{Y} = 496 - 510 = (500 - 510) + (496 - 500)$$

where the first part of $(500 - 510)$ is the parameter error, and $(496 - 500)$ is the noise.

Usually parameter error and process error are stochastically independent, because parameter errors depend on past data, while process error are components of the future data. Intuitively, this is because our parameter errors are caused by the parameter estimates $\hat{\theta}$ that we obtain using the past data $Y_{kj} \in D_K$, whereas process errors ε_{kj} are caused by the stochastic nature of future observations.

Note that in our definition and example demos, we are assuming that the precise form for the model function $u(.)$ is known. However, in practice this is not always true, and an incorrect model may be used to make future estimates. Denoting this function incorrectly selected as $v(.)$, the difference between the expected outcome of the selected model, $E\left[Y_{kj}\right] = v(k,j;\theta)$,and the expected outcome of the true model, $E\left[Y_{kj}\right] = u(k,j;\theta)$, is referred to as the model error, which will be discussed in detail in later sections.

4.1.2. Loss Reserves

With the example demonstrations, we have an understanding of prediction errors, parameter errors and process errors for GLMs. To understand it in terms of loss reserves, we need an example in that context. In the following example, we will explain prediction error in the context of loss reserve using the chain ladder algorithm.

Example loss reserve: suppose we have the following data for cumulative past observations of paid loss:

Cumulative Paid Loss ($000)					
K \ J	1	2	3	4	5
1	200	380	470	500	510
2	210	375	482	503	
3	195	363	486		
4	190	376			
5	204				

Using the weighted averages, we can calculate the age-to-age factors as follows:

Age-to-age factors for development year j				
Development year j	1	2	3	4
$\hat{f}_j$	1.879	1.286	1.054	1.020

The weighted averages calculated are thus our estimated parameters for the cumulative observations

$$X_{kj} \sim u(k, j; f_j)$$

Using the weighted averages, we can estimate future paid losses and the ultimate losses for accident years 2 to 5 as

Cumulative Paid Loss ($000)					
k\j	1	2	3	4	5
1	200	380	470	500	510
2	210	375	482	503	**513.06**
3	195	363	486	512.04	522.28
4	190	376	483.62	509.53	519.72
5	204	383.37	493.10	519.51	529.90

Suppose that the true age-to-age factors are 1.9, 1.3, 1.05, 1.01, with tail factor 1 (meaning there is no more claim development after the 5th year).

Consider accident year 2 as an example, using the cumulative paid loss at development year 4 and our estimated age-to-age factor at year 4, we get the expected ultimate loss as

$$\hat{X}_{2,5} = X_{2,4} \times \hat{f}_4 = 503 \times 1.020 = 513.06$$

However, assuming a true age-to-age factor of the development year at 1.01, the expected ultimate loss should have been

$$E\left[X_{2,5}\right] = X_{2,4} \times f_4 = 503 \times 1.01 = 508.03$$

Thus the parameter error caused by parameter estimation is

$$parameter\ error = 508.03 - 513.06 = -5.03$$

Suppose that another year pass and our ultimate loss for accident year 2 is actually 510. The difference between the expected value and this actual value is the process error caused by the stochastic nature of future observations. Thus our process error is

$$process\ error = 510 - 508.03 = 1.97$$

And our prediction error for the cell is

$$e_{2,5} = X_{2,5} - \hat{X}_{2,5} = 510 - 513.06 = -3.06 = parameter\ error + process\ error$$

For simplicity, we can present the prediction errors in vector form. Without concerning about the order of the components, denote

- Y – vector formed by the past observations $Y_{kj} \in D_K$
- Y^* – vector formed by the future observations $Y_{kj} \in D_K^c$
- μ^* – vector formed by the expected future observations $E\left[Y_{kj}\right]$, for $Y_{kj} \in D_K^c$
- e^* – vector formed by the prediction errors associated with $Y_{kj} \in D_K^c$
- ε^* – vector formed by the process errors associated with $Y_{kj} \in D_K^c$

Recall the star symbol denotes any elements of the future dataset. Then (4-6) becomes

$$e^* = \left[\mu^* - \hat{Y}^*\right] + \varepsilon^*$$

(4-7)

We can also consider linear combinations of the components of vector Y^* for future observations. Denote some vector *r* as a vector of constants that has the same dimension as Y^* and the other vectors. Then we can represent the linear combination of Y^* as $r^T Y^*$, which would return a scalar that is the linear combinations of the components of Y^*. For example, to calculate the total outstanding claims, we need to let $r = [1\ 1\ ...\ 1]^T$, so we can have

$$r^T Y^* = [1\ 1\ ...\ 1]\begin{bmatrix} Y^*_{k_1 j_1} \\ Y^*_{k_1 j_2} \\ \vdots \\ Y^*_{k_n j_n} \end{bmatrix} = \sum_{all\ k,j\ for\ D_K^C} Y^*_{kj} = total\ outstanding\ loss$$

Or, let *r* be a vector with 1 in positions for Y^*_{kj} of some row *k* and 0 everywhere else, we can get the outstanding loss for accident year by computing $r^T Y^*$

$$r^T Y^* = [0\ 0\ ...\ 1\ 1\ ...\ 0\ 0]\begin{bmatrix} Y^*_{k_1 j_1} \\ Y^*_{k_1 j_2} \\ \vdots \\ Y^*_{k_i j_1} \\ Y_{k_i j_2} \\ \vdots \\ Y^*_{k_i j_n} \\ \vdots \\ Y^*_{k_n j_n} \end{bmatrix} = \sum_{all\ j\ for\ Y^*_{k_i j} \in D_K^C} Y^*_{kj}$$
$$= total\ outstanding\ loss\ for\ accident\ year\ k_i$$

Denote the prediction error associated with the linear combination $r^T Y^*$ as $e^*_{(r)}$, then by (4-7) we obtain

$$e^*_{(r)} = r^T e^* = \left[r^T \mu^* - r^T \hat{Y}^*\right] + r^T \varepsilon^*$$

(4-8)

where

$e^*_{(r)}$ – a scalar which is the prediction error associated with $r^T Y^*$
$r^T \mu^*$ – a scalar that represent the expected outstanding losses
$r^T \hat{Y}^*$ – a scalar that represent the estimated outstanding losses
$r^T \varepsilon^*$ – a scalar of associated noise

In (4-8) note that we can also express the prediction error as the sum of parameter error and process error, where the term $\left[r^T\mu^* - r^T\hat{Y}^*\right]$ represents the parameter error and $r^T\varepsilon^*$ represents the process error.

4.2. Mean Square Error of Prediction

4.2.1. Definition

The **mean square error of prediction** (**MSEP**) for prediction error, denoted $MSEP\left[e^*_{(r)}\right]$, measures the magnitude of prediction error $e^*_{(r)}$. It is defined as

$$MSEP\left[e^*_{(r)}\right] = E\left\{\left[e^*_{(r)}\right]^2\right\}$$

(4-9)

which is the expected value (or mean) of the sum of squares of prediction errors.

When parameter error and process error are stochastically independent, we can substitute (4-8) into (4-9) to calculate the MSEP of prediction error in terms of parameter and process errors. This means, for $MSEP\left[e^*_{(r)}\right] = E\left\{\left[e^*_{(r)}\right]^2\right\}$, we can rewrite it as

$$MSEP\left[e^*_{(r)}\right] = E\left\{\left(\left[r^T\mu^* - r^T\hat{Y}^*\right] + r^T\varepsilon^*\right)^2\right\} = E\left\{\left[e^*_{(r)param} + e^*_{(r)proc}\right]^2\right\}$$

$$= E\left\{\left[e^*_{(r)param}\right]^2 + 2e^*_{(r)param}\cdot e^*_{(r)proc} + \left[e^*_{(r)proc}\right]^2\right\}$$

$$= E\left\{\left[e^*_{(r)param}\right]^2\right\} + 2E\left\{e^*_{(r)param}\right\}\cdot E\left\{e^*_{(r)proc}\right\} + E\left\{\left[e^*_{(r)proc}\right]^2\right\}$$

From (4-2) we know the expected value of process error is zero, thus $E\left\{e^*_{(r)proc}\right\} = 0$, so the term $2E\left\{e^*_{(r)param}\right\}\cdot E\left\{e^*_{(r)proc}\right\}$ is zero, and we obtain

$$MSEP\left[e^*_{(r)}\right] = E\left\{\left[e^*_{(r)param}\right]^2\right\} + E\left\{\left[e^*_{(r)proc}\right]^2\right\}$$

(4-10)

where

$$e^*_{(r)param} = r^T\mu^* - r^T\hat{Y}^* = parameter\ error$$

(4-11)

and

$$e^*_{(r)proc} = r^T \varepsilon^* = process\ error$$

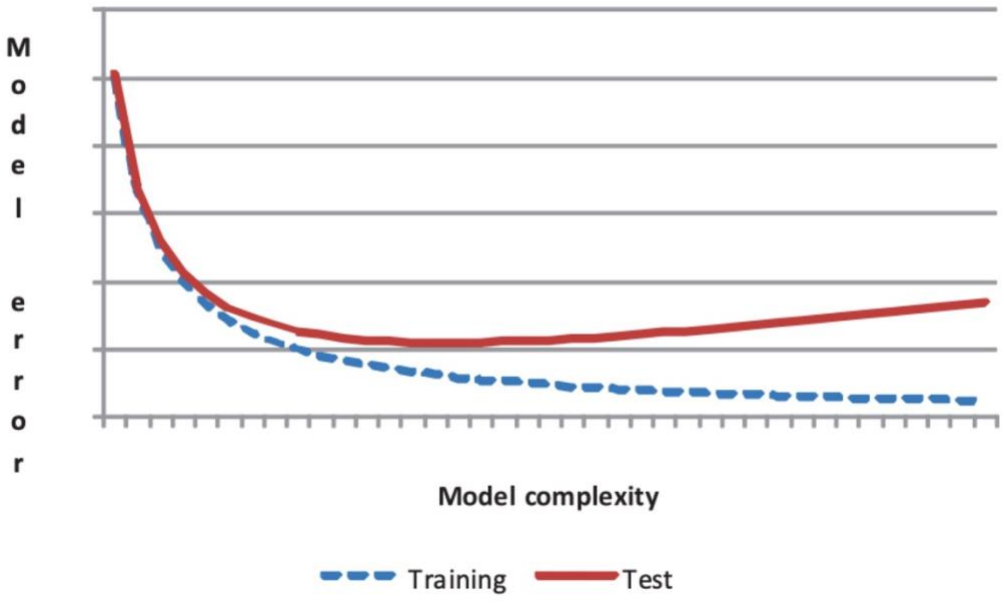
(4-12)

4.2.2. Goodness-of-Fit and Prediction Error

As stated in the previous section, the MSEP measures the magnitude of prediction error. In other words, it measures the tightness of future estimates around the target. Thus usually the smaller a model's MSEP is, the more preferred the model is. However, MSEP is not equivalent to goodness-of-fit of a model, so improving a model's goodness-of fit does not necessarily mean improving the MSEP.

The goodness-of-fit of a model can be increased by including excessive parameters, but this inclusion can destabilize model's estimations, and thus amounts to over-fitting and thus increase the value of MSEP. Therefore, an effective model needs to take into account both the goodness-of-fit and the complexity of the model. Figure 4-1 from the monograph summarizes the relationship between model error and model complexity.

Figure 4-1. Goodness-of-Fit and Prediction Error



Suppose we divide the available data set has 2 subsets, a **training set** and a **test** (or **holdout**) **set**.

First the model is fitted to the training set. We can use this to select a form of error, such as squared error and deviance, and plot the error against model complexity. Plotting this in the graph we can see that the fit of the model is improving (model error decreasing) as model complexity increases (Figure 4-1).

However, as we use the model on the test set to generate the fitted values, we can see the fit of the model as an estimator of the test data does not improve monotonically as for the training set. When the parameter number is small, the model produces a poor fit in both cases, and as model complexity is increased, the model fits both sets better. However, after a certain point, the increase of complexity results in over-fitting, where we observe a still increasing fit on the training set, but a decrease in fit on estimating the test set, as excessive parameters start to destabilize the estimation.

Thus we can conclude, as model complexity increases, both the fit and estimation of the training set and test set can be improved to a certain point, but afterwards, detraction appears. Intuitively, we can think of model complexity at the extreme case. If we have a model that fit the data perfectly, then this model has as many parameters as the data points in the training set, and can produce zero error. But at this point the model is not a model in the usual sense anymore. It is only a list of outcome and input with no formulaic meaning behind the values, and has lost its predictive value.

Therefore, it is obvious that the point where model error is at minimum for both the training and test sets is the optimal model complexity. Visually, this is the minimum point on the test curve in Figure 4-1, which produces the model that has the best predictive value.

4.3. Information Criteria

The **information criteria** are the statistics for measuring model fit error relative to a test data set. It is defined as

$$\begin{aligned} &information\ criterion \\ &\quad = measure\ of\ model\ fit\ error\ (relative\ to\ training\ data\ set) \\ &\quad + penalty\ for\ number\ of\ parameters \end{aligned}$$

(4-13)

The information criteria behave similarly to the model fit error relative to a test data set, as shown in Figure 4-1. Recall that while initially the model error relative to a test set decreases as model complexity increases, after a certain point the model error starts to increase again as the model loses predictive value. Similar for information criterion, when the model complexity increases, the model fit error for the training set decreases

monotonically, but the penalty for number of parameters increases. Thus there will also be a point of model complexity where the increase of penalty starts to overwhelm the decrease of model fit error.

For a GLM, (4-13) can be written as following, with $\hat{Y}$ being the fitted value of observation Y

$$IC(Y, \hat{Y}) = D(Y, \hat{Y}) + f(p)$$

(4-14)

where

$IC(Y, \hat{Y})$ is the information criterion;
$D(Y, \hat{Y})$ is the scaled deviance from (2-30);
p is the number of the model parameters;
$f(.)$ is a monotonically increasing function.

Recall from (2-30), the scaled deviance has the following formula:

$$D(Y, \hat{Y}) = 2\left[ln\, \pi(Y;\, \hat{\theta}^{(s)}, \phi) - ln\, \pi(Y;\, \hat{\theta}, \phi)\right]$$
$$= 2 \sum_{i=1}^{n} [ln\, \pi(Y_i;\, \hat{\theta}^{(s)}, \phi) - ln\, \pi(Y_i;\, \hat{\theta}, \phi)]$$

Table 4-1 from the monograph shows 2 of the most common information criteria, the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC). Note in BIC, n is the number of observations Y used in the model, so in both cases the penalty functions are linear functions of p.

Table 4-1. Information Criteria

Information Criterion	Function $f(p)$
Akaike Information Criterion (AIC)	$2p$
Bayes Information Criterion (BIC)	$p\ ln\ n$

AIC is independent of the number of observations n used in the model, but there is a modified version, AICc, which has a correction for finite sample size n. The last c stands for correlation.

AICc has the form

$$f(p) = 2p\left[1 + \frac{p+1}{n-p-1}\right]$$

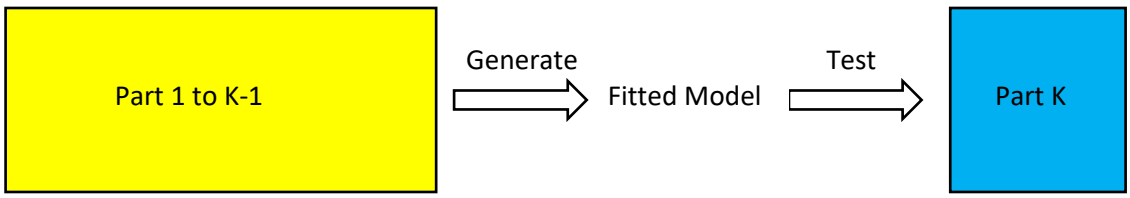
So that as n/p increases to infinity, $f(p)$ approaches $2p$:

$$f(p) = 2p\left[1 + \frac{p+1}{n-p-1}\right] = 2p\left[1 + \frac{1+\frac{1}{p}}{\frac{n}{p} - 1 - \frac{1}{p}}\right] \to 2p$$

The information criteria are used to compare the loss of information from different models of the same data set. For example, if the AIC indicates a smaller number for model 1 than for model 2, then model 1 has minimized information loss better, and model 1 would be favored.

4.4 Generalized Cross-Validation

Cross-Validation is a method commonly used in regression and non-regression models to estimate prediction error. An example of cross-validation would be to divide the data into K parts, so that the fitted model can be generated by the first K-1 parts, and tested by the Kth part. This is also called the leave-one-out cross-validation.



For linear models, the fitted value can be expressed as $\hat{y} = Hy$, where H is called the hat matrix (because it gives a "hat" to y after the multiplication). An approximation to leave-one-out validation is the **generalized cross-validation (GCV)** measure, with formula

$$GCV = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n\left[1 - \frac{trace(H)}{n}\right]^2}$$

(4-15)

where:

Y_i is the i^{th} observed value (Not referred to incremental losses)
$\hat{Y}_i$ is the i^{th} fitted value
n is the number of observations

In the numerator of the GCV formula, we have the sum of the squared error for fitted values, which is divided by the number of observations n in the denominator. We want this value to be as small as possible for a good-fitting model. However, we also need to take over-fitting into consideration. Thus we have the term $\left[1 - \frac{trace(H)}{n}\right]^2$ in the denominator. Here the hat matrix H is an n × n diagonal matrix that maps the n × 1 vector of observations Y to the n × 1 vector of fitted values $\hat{Y}$. The trace of the hat matrix, $trace(H)$, is the sum of the diagonal calculated as

$$trace(A) = \sum_{i=1}^{n} a_{ii}$$

and defined as the **effective number of parameters** in a model. Because the hat matrix maps y to $\hat{y}$, we want $trace(H)$ close to n for a good-fitting model. However, as this could also result in over-fitting of the model, we have $\left[1 - \frac{trace(H)}{n}\right]^2$ as a penalty, which decreases in value as $trace(H)$ gets closer to n, and thus increase the value of GCV.

Therefore, from the overall formula of GCV, we can tell that a smaller GCV suggests a better model for the observations not only in terms of the goodness-of-fit of the model, but also considering the number of parameters used in the model.

4.5 Model Error

Model Error is the error associated with using an inaccurate model to fit the data. Model Error is common when the accurate function to estimate the data is unknown or unknowable. In previous sections of this chapter, we only discuss the parameter error and process error with the assumption that the function $\mu(k,j;\theta)$ underlying the data is correctly identified. In this section, we will recognize that the selected modeling function can also have errors and affect its fit to the data.

We assume $u(k,j;\theta)$ is still the correct function for the data. Suppose we incorrectly choose function $v(k,j;\xi)$ as our modeling function, with some parameter ξ:

$$Y_{kj} = v(k,j;\xi) + \varepsilon_{kj} \; for \; Y_{kj} \epsilon \mathfrak{D}_K^+$$

(4-16)

Then, the fitted values for this model would be:

$$\hat{Y}_{kj} = v\left(k,j;\hat{\xi}\right) \; for \; Y_{kj} \epsilon \mathfrak{D}_K^+$$

(4-17)

In this case, the prediction error e_{kj} would be:

$$e_{kj} = Y_{kj} - \hat{Y}_{kj} = \underbrace{\left[\, v(k,j;\xi) - \; v\left(k,j;\hat{\xi}\right)\right]}_{\text{Parameter Error}} \underbrace{+ \; \varepsilon_{kj}}_{\text{Process Error}} + \underbrace{\left[\, \mu(k,j;\theta) - (k,j;\xi)\right]}_{\text{Model Error}}$$

This decomposition of prediction error includes parameter error and process error as in (4-4), but now it also includes the term for **model error**. This term measures the error incur by selecting an incorrect modeling function to fit the data.

5. The Bootstrap

In future estimation, we often have limited number of data sets to generate future estimations. The purpose of bootstrap is to generate synthetic data sets with the same stochastic properties as the original one, and produce estimates of outstanding losses from each dataset. With large number of future estimates, we can have a clearer picture for the full distribution of our target prediction such that we can set loss reserves with certain confidence levels. This chapter focuses on the two ways of bootstrapping for loss reserving purpose: semi-parametric bootstrap and parametric bootstrap.

5.1. Background

In Chapter 3, we used GLMs to generate the parameter estimates for both ODP Mack Model and ODP Cross-Classified Model. Although we showed only the parameter estimates in Chapter 3, the associated standard errors for parameter estimates and the estimated correlations between each pair of them are also reported by SAS.

We are interested in these standard errors of parameter estimates and their correlations, because with knowledge of the distribution for parameter estimates, we can randomly draw **pseudo-parameter estimates** to form **pseudo-data sets**. Because we often have limited number of data sets to generate future estimates, we cannot determine the distribution for our target prediction, which in loss reserving is the total outstanding losses. To resolve this problem, we use pseudo-data sets with the same stochastic properties as the original one to generate a large number of future estimates so that we can estimate the distribution of our target prediction.

Table 5-1 from the monograph shows the parameter estimates and their standard errors for ODP Cross-Classified Model:

Table 5-1. GLM Parameter Estimates and Standard Errors for ODP Cross-Classified Model

	$ln\ \hat{\alpha}_k$		$ln\ \hat{\beta}_j$	
j or *k*	Estimate	Standard Error	Estimate	Standard Error
1	10.657	0.0316	0.000	
2	10.795	0.0299	−0.205	0.0228
3	10.899	0.0289	−0.747	0.0282
4	10.989	0.0281	−1.017	0.0328
5	11.039	0.0278	−1.452	0.0421
6	11.016	0.0285	−1.833	0.0547
7	11.008	0.0295	−2.140	0.0715
8	10.891	0.0327	−2.348	0.0931
9	10.836	0.0367	−2.513	0.1267
10	10.691	0.0510	−2.664	0.1993

Table 5-1

Note that standard error of $\ln(\beta_1)$ is not included in the table because we set β_1=1 as the scale.

Table 5-2 from the table shows the correlation between the parameter estimates:

Table 5-2. Estimated Correlation Matrix of GLM Parameter Estimates for ODP Cross-Classified Model

	Parameter									
Parameter	$ln\ \hat{\alpha}_1$	$ln\ \hat{\alpha}_2$	$ln\ \hat{\alpha}_3$	$ln\ \hat{\alpha}_4$	$ln\ \hat{\alpha}_5$	$ln\ \hat{\alpha}_6$	$ln\ \hat{\alpha}_7$	$ln\ \hat{\alpha}_8$	$ln\ \hat{\alpha}_9$	$ln\ \hat{\alpha}_{10}$
$ln\ \hat{\alpha}_1$	1.00									
$ln\ \hat{\alpha}_2$	0.20	1.00								
$ln\ \hat{\alpha}_3$	0.20	0.21	1.00							
$ln\ \hat{\alpha}_4$	0.20	0.21	0.22	1.00						
$ln\ \hat{\alpha}_5$	0.19	0.20	0.21	0.22	1.00					
$ln\ \hat{\alpha}_6$	0.18	0.19	0.20	0.20	0.20	1.00				
$ln\ \hat{\alpha}_7$	0.16	0.17	0.18	0.18	0.18	0.18	1.00			
$ln\ \hat{\alpha}_8$	0.13	0.14	0.14	0.15	0.15	0.14	0.14	1.00		
$ln\ \hat{\alpha}_9$	0.09	0.10	0.10	0.10	0.10	0.10	0.10	0.09	1.00	
$ln\ \hat{\alpha}_{10}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Parameter	Parameter									
	$ln\,\hat{\alpha}_1$	$ln\,\hat{\alpha}_2$	$ln\,\hat{\alpha}_3$	$ln\,\hat{\alpha}_4$	$ln\,\hat{\alpha}_5$	$ln\,\hat{\alpha}_6$	$ln\,\hat{\alpha}_7$	$ln\,\hat{\alpha}_8$	$ln\,\hat{\alpha}_9$	$ln\,\hat{\alpha}_{10}$
$ln\,\hat{\beta}_2$	–0.32	–0.34	–0.35	–0.36	–0.37	–0.36	–0.35	–0.31	–0.28	0.00
$ln\,\hat{\beta}_3$	–0.28	–0.29	–0.30	–0.31	–0.32	–0.31	–0.30	–0.27	–0.10	0.00
$ln\,\hat{\beta}_4$	–0.25	–0.27	–0.28	–0.29	–0.29	–0.28	–0.27	–0.12	–0.09	0.00
$ln\,\hat{\beta}_5$	–0.21	–0.22	–0.23	–0.24	–0.24	–0.24	–0.12	–0.10	–0.07	0.00
$ln\,\hat{\beta}_6$	–0.18	–0.19	–0.20	–0.20	–0.20	–0.10	–0.09	–0.07	–0.05	0.00
$ln\,\hat{\beta}_7$	–0.16	–0.17	–0.17	–0.18	–0.09	–0.08	–0.07	–0.06	–0.04	0.00
$ln\,\hat{\beta}_8$	–0.14	–0.15	–0.16	–0.07	–0.07	–0.06	–0.06	–0.04	–0.03	0.00
$ln\,\hat{\beta}_9$	–0.14	–0.15	–0.05	–0.05	–0.05	–0.04	–0.04	–0.03	–0.02	0.00
$ln\,\hat{\beta}_{10}$	–0.16	–0.03	–0.03	–0.03	–0.03	–0.03	–0.03	–0.02	–0.01	0.00

Parameter	Parameter								
	$ln\,\hat{\beta}_2$	$ln\,\hat{\beta}_3$	$ln\,\hat{\beta}_4$	$ln\,\hat{\beta}_5$	$ln\,\hat{\beta}_6$	$ln\,\hat{\beta}_7$	$ln\,\hat{\beta}_8$	$ln\,\hat{\beta}_9$	$ln\,\hat{\beta}_{10}$
$ln\,\hat{\beta}_2$	1.00								
$ln\,\hat{\beta}_3$	0.36	1.00							
$ln\,\hat{\beta}_4$	0.31	0.27	1.00						
$ln\,\hat{\beta}_5$	0.24	0.21	0.19	1.00					
$ln\,\hat{\beta}_6$	0.19	0.16	0.15	0.12	1.00				
$ln\,\hat{\beta}_7$	0.14	0.12	0.11	0.09	0.08	1.00			
$ln\,\hat{\beta}_8$	0.11	0.09	0.09	0.07	0.06	0.05	1.00		
$ln\,\hat{\beta}_9$	0.08	0.07	0.06	0.05	0.04	0.04	0.04	1.00	
$ln\,\hat{\beta}_{10}$	0.05	0.04	0.04	0.03	0.03	0.02	0.02	0.02	1.00

With the information in these table, we are able to implement a parametric bootstrap to estimate the full distribution for outstanding losses. Detail steps for parametric bootstrap will be discussed later in section 5.3.2, followed by numerical example in 5.4.

The SAS codes to reproduce table 5-1 and table 5-2 is included in **Appendix D**.

Reproduced result is shown below:

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	0	0.0000	0.0000	0.0000	0.0000	.	.
a1	1	10.6568	0.0316	10.5947	10.7188	113431	<.0001
a2	1	10.7953	0.0299	10.7366	10.8540	129966	<.0001
a3	1	10.8992	0.0289	10.8426	10.9558	142481	<.0001
a4	1	10.9890	0.0281	10.9340	11.0441	153137	<.0001
a5	1	11.0388	0.0278	10.9843	11.0934	157335	<.0001
a6	1	11.0159	0.0285	10.9599	11.0719	148895	<.0001
a7	1	11.0081	0.0295	10.9504	11.0658	139678	<.0001
a8	1	10.8905	0.0327	10.8265	10.9545	111198	<.0001
a9	1	10.8361	0.0367	10.7642	10.9080	87231.2	<.0001
a10	1	10.6911	0.0510	10.5910	10.7911	43871.3	<.0001
b2	1	-0.2047	0.0228	-0.2493	-0.1601	80.88	<.0001
b3	1	-0.7474	0.0282	-0.8027	-0.6922	702.87	<.0001
b4	1	-1.0167	0.0328	-1.0810	-0.9523	958.18	<.0001
b5	1	-1.4516	0.0421	-1.5342	-1.3690	1186.56	<.0001
b6	1	-1.8325	0.0547	-1.9398	-1.7253	1121.89	<.0001
b7	1	-2.1403	0.0715	-2.2804	-2.0001	895.96	<.0001
b8	1	-2.3483	0.0931	-2.5308	-2.1658	635.93	<.0001
b9	1	-2.5132	0.1267	-2.7615	-2.2648	393.28	<.0001
b10	1	-2.6645	0.1993	-3.0551	-2.2739	178.73	<.0001

(Reproduction of table 5-1)

1.0000	0.2043	0.2017	0.1991	0.1932	0.1797	0.1626	0.1296	0.0899	0.0000
0.2043	1.0000	0.2131	0.2104	0.2041	0.1898	0.1719	0.1370	0.0950	0.0000
0.2017	0.2131	1.0000	0.2182	0.2117	0.1969	0.1782	0.1421	0.0986	0.0000
0.1991	0.2104	0.2182	1.0000	0.2177	0.2024	0.1833	0.1461	0.1013	0.0000
0.1932	0.2041	0.2117	0.2177	1.0000	0.2043	0.1849	0.1474	0.1023	0.0000
0.1797	0.1898	0.1969	0.2024	0.2043	1.0000	0.1803	0.1437	0.0997	0.0000
0.1626	0.1719	0.1782	0.1833	0.1849	0.1803	1.0000	0.1393	0.0966	0.0000
0.1296	0.1370	0.1421	0.1461	0.1474	0.1437	0.1393	1.0000	0.0871	0.0000
0.0899	0.0950	0.0986	0.1013	0.1023	0.0997	0.0966	0.0871	1.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

(Reproduction of table 5-2: correlations between $\ln(\alpha_i)$)

-0.3229	-0.3412	-0.3539	-0.3639	-0.3672	-0.3579	-0.3469	-0.3129	-0.2785	0.0000
-0.2772	-0.2929	-0.3038	-0.3124	-0.3152	-0.3072	-0.2978	-0.2686	-0.1009	0.0000
-0.2530	-0.2674	-0.2773	-0.2851	-0.2877	-0.2804	-0.2718	-0.1249	-0.0866	0.0000
-0.2126	-0.2246	-0.2330	-0.2395	-0.2417	-0.2356	-0.1221	-0.0973	-0.0675	0.0000
-0.1797	-0.1899	-0.1969	-0.2025	-0.2043	-0.1039	-0.0941	-0.0750	-0.0520	0.0000
-0.1564	-0.1653	-0.1714	-0.1763	-0.0855	-0.0795	-0.0720	-0.0574	-0.0398	0.0000
-0.1446	-0.1528	-0.1585	-0.0677	-0.0656	-0.0610	-0.0553	-0.0441	-0.0306	0.0000
-0.1420	-0.1501	-0.0504	-0.0497	-0.0482	-0.0449	-0.0406	-0.0324	-0.0225	0.0000
-0.1588	-0.0324	-0.0320	-0.0316	-0.0307	-0.0285	-0.0258	-0.0206	-0.0143	0.0000

(Reproduction of table 5-2: correlations between $\ln(\alpha_i)$ and $\ln(\beta_j)$)

1.0000	0.3625	0.3111	0.2425	0.1868	0.1429	0.1097	0.0806	0.0513
0.3625	1.0000	0.2671	0.2081	0.1603	0.1227	0.0942	0.0692	0.0440
0.3111	0.2671	1.0000	0.1900	0.1463	0.1120	0.0860	0.0632	0.0402
0.2425	0.2081	0.1900	1.0000	0.1229	0.0941	0.0722	0.0531	0.0337
0.1868	0.1603	0.1463	0.1229	1.0000	0.0795	0.0611	0.0449	0.0285
0.1429	0.1227	0.1120	0.0941	0.0795	1.0000	0.0532	0.0391	0.0248
0.1097	0.0942	0.0860	0.0722	0.0611	0.0532	1.0000	0.0361	0.0230
0.0806	0.0692	0.0632	0.0531	0.0449	0.0391	0.0361	1.0000	0.0225
0.0513	0.0440	0.0402	0.0337	0.0285	0.0248	0.0230	0.0225	1.0000

(Reproduction of table 5-2: correlations between $\ln(\beta_j)$)

Check out **Appendix D** and try to reproduce table 5-1 and table 5-2 with your code.

5.2. The Bootstrap

The bootstrap method provides a distribution of target estimates, instead of a point estimate. In loss reserving, for example, when insurance companies estimate future losses and set up loss reserves, it is usually necessary to set up loss reserves with some confidence level of covering for the potential loss. This requires for calculating probability of adequacy (**PoA**) of the reserve, and adjust loss reserve based on the probability. Mathematically, this means we want the true total outstanding loss R to be less than an estimated outstanding loss $\hat{R}_p$ for some given probability p, i.e.

$$Prob\left[R < \hat{R}_p\right] = p$$

(5-14)

And once we have $\hat{R}_p$, we can set our loss reserve to meet the requirement. Note that in order to calculate the probability, we need the distribution of the total outstanding loss, which requires the use of bootstrap method.

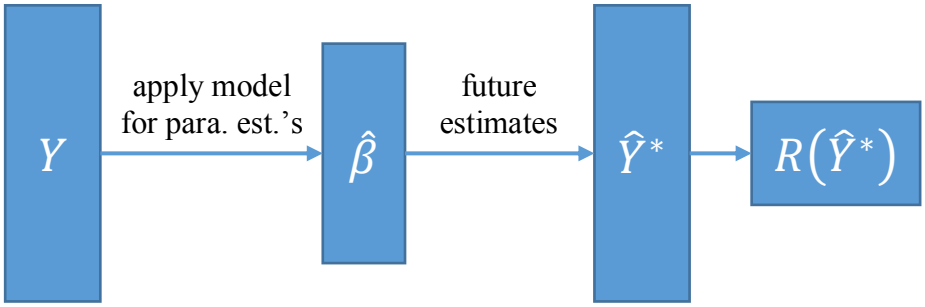
There are many approaches to the bootstrap method, which are categorized into "non-parametric", "semi-parametric", and "parametric" bootstrap methods by Shibata (1997). This classification involves the level of reliance of prediction error on model and distributional assumptions. The sub-sections below will discuss "semi-parametric" and "parametric" bootstrap methods in details.

5.2.1. Semi-Parametric Bootstrap

The original form of the bootstrap by Efron (1979) falls within the general family of **re-sampling**, which involves repeated sampling of available data and constructing pseudo datasets and fitted values.

Let Y be an n-dimensional data vector. Suppose we fit a model to Y, and obtain an n-dimensional vector of estimations $\hat{Y}^*$ of future observations Y^* with parameter estimates $\hat{\beta}$.

Let $R(Y^*)$ be the target prediction, where $R(.)$ is some function of Y^* that produces the target estimation. In the case of loss reserve, where Y^* are the future incremental losses, the function $R(.)$ is simply a summation function, as our target prediction is the total outstanding loss, computed by summing all future losses. Because Y^* are unknown, we can estimate $R(Y^*)$ using estimated future observations $\hat{Y}^*$, which gives $R(\hat{Y}^*)$.



To find the PoA described by (5-14), we need the distribution of the estimated $R(\hat{Y}^*)$, which is the objective of the Bootstrap method and re-sampling procedure.

We can use the known observations for the purpose. Let $\hat{Y}$ denote the n-dimensional vector of the fitted values of Y using the model and estimated parameters $\hat{\beta}$, and let

$S(Y;\hat{Y})$ denote the vector of associated standardized residuals where the inverse $S^{-1}(.;\hat{Y})$ exists.

Recall from Chapter 2 we introduced Pearson residuals, which has the following formula from equation (2-33):

$$R_i^P = \frac{Y_i - \hat{Y}_i}{\hat{\sigma}_i}$$

Assuming we use the Pearson residuals, then the *i*-th component of the residual vector corresponding to the data vector Y, denoted $S(Y;\hat{Y})$, $i = 1, \dots, n$, can be written as

$$S_i(Y;\hat{Y}) = \frac{Y_i - \hat{Y}_i}{\hat{\sigma}_i}$$

(5-15)

where $\hat{\sigma}_i^2$ is the estimator of the variance $Var[Y_i]$. We can then derive (5-15) to and represent the *i*-th component of Y as the inverse of the residual function, i.e.,

$$Y_i = S^{-1}(S_i;\hat{Y}) = \hat{Y}_i + \hat{\sigma}_i S_i$$

(5-16)

Suppose S_i are approximately independent and identically distributed (iid). We can then perform **data re-sampling** of the residuals. We can draw a random *n*-sample from $S(Y;\hat{Y})$ with or without replacement. Denote the sample residual components as $\tilde{S}_i$, $i = 1, \dots, n$, and denote the corresponding vector as $\tilde{S}$.

Using (5-16), we can form a sample of observations $\tilde{Y}$, where the *i*-th component is defined as

$$\tilde{Y}_i = S^{-1}(\tilde{S}_i;\hat{Y})$$

(5-17)

Note that the index *i* in (5-17) corresponds to the index in $\tilde{S}$ instead of the original residual set S. Thus the *i*-th component of $\tilde{Y}$ uses the same $\hat{Y}_i$ but a usually different S_i because orders of the residuals are changed during the random drawing process. So, $\tilde{Y}$ doesn't usually have the same value as the *i*-th component of the original data vector Y, i.e.,

$$\tilde{Y}_i \neq Y_i, \qquad for\ \tilde{S}_i \neq S_i$$

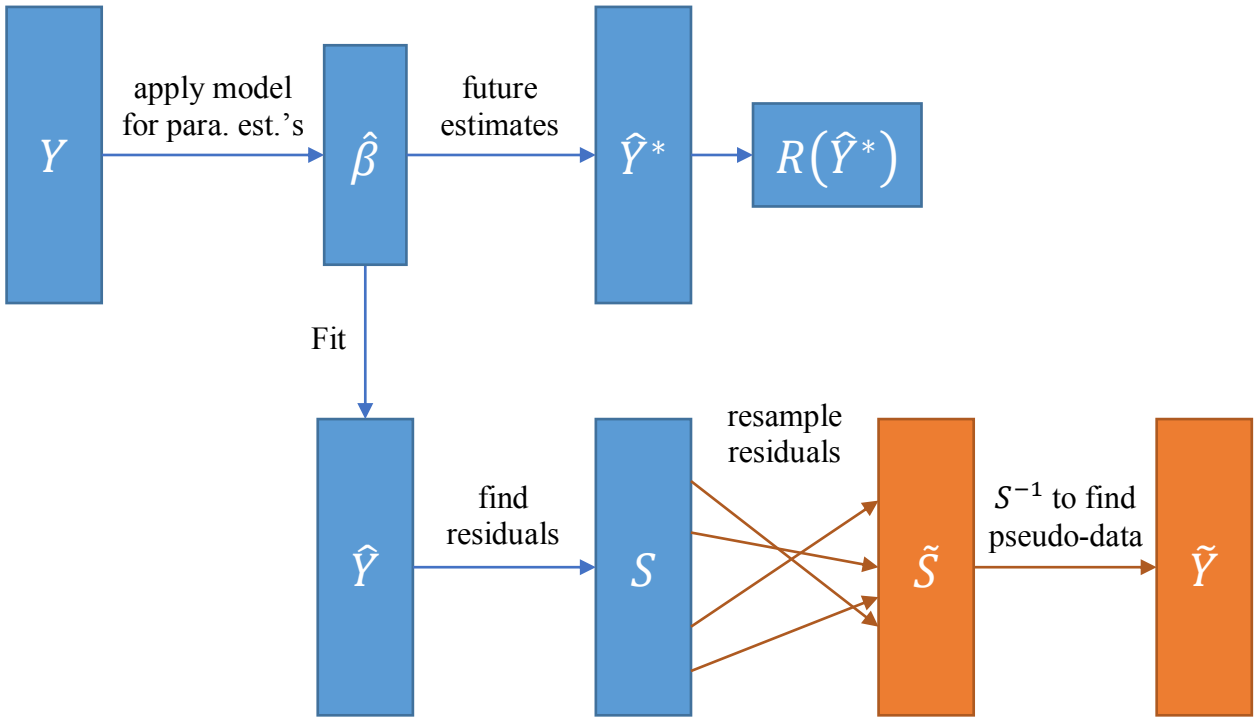
Because S_i are approximately iid, S and $\tilde{S}$ have the same stochastic properties, and thus Y and $\tilde{Y}$ have the same stochastic properties by (5-16) and (5-17). Thus through re-

sampling we have obtained an alternative data set of observations $\tilde{Y}$ that has the same stochastic properties as the original one, which is also known as a **pseudo-data set**. In our case with Pearson residuals, the components of the pseudo-data set $\tilde{Y}$ are defined as

$$\tilde{Y}_i = S^{-1}(\tilde{S}_i; \hat{Y}) = \hat{Y}_i + \hat{\sigma}_i \tilde{Y}_i$$

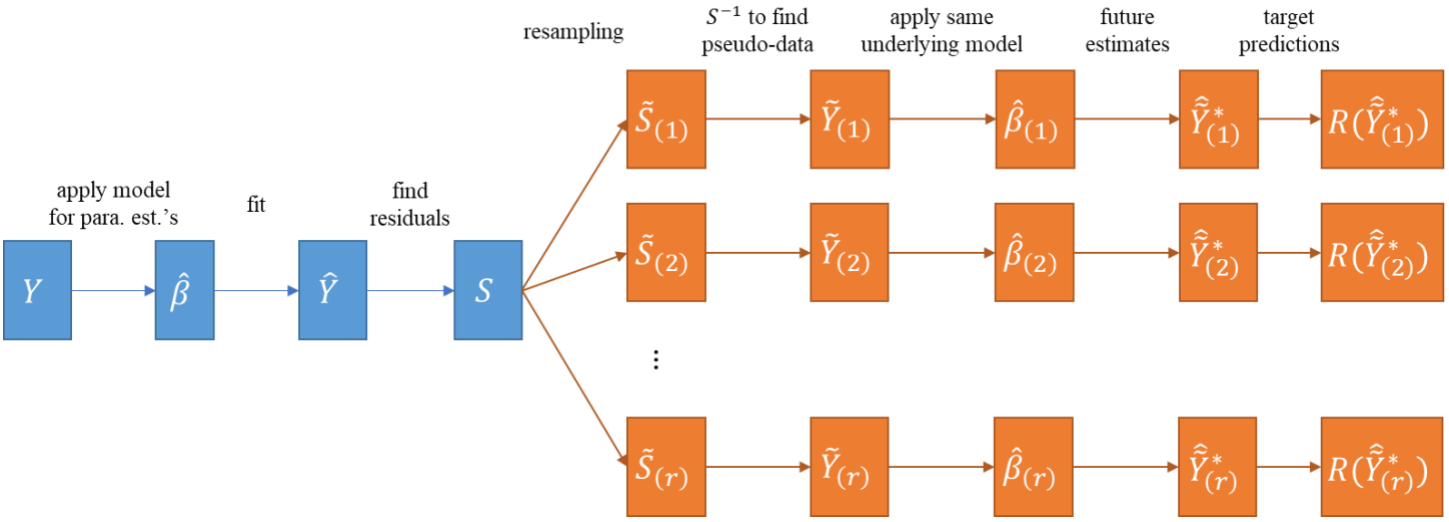
(5-18)

Which is obtained by combining substituting S_i with $\tilde{S}_i$ in (5-16).



We can draw in total of $n!$ pseudo-data sets if we sample without replacement and n^n for sampling with replacement. Suppose we draw r pseudo-data sets, where r is sufficiently large. Denote these sets as vectors $\tilde{Y}_{(1)}$, $\tilde{Y}_{(2)}$,…, $\tilde{Y}_{(r)}$, and model each of these sets with the same model applied to Y originally. That is, the model applied to each $\tilde{Y}_{(j)}$, $j = 1, \ldots, r$ has the same algebraic structure as the model applied to Y. However, because $\tilde{Y}_{(j)}$ do not contain exactly the same components as the original Y, the parameters of the model will change as the data inputs have changed, and will be different for each of the pseudo-data set $\tilde{Y}_{(j)}$.

For these $\tilde{Y}_{(j)}$, we can find the corresponding estimated parameters, called **pseudo-estimates**. Arrange the parameter estimates into a vector denoted as $\hat{\beta}_{(j)}$ for the associated $\tilde{Y}_{(j)}$. We can then apply the model with $\hat{\beta}_{(j)}$ for the corresponding j-th dataset and find the estimated future observations $\hat{\tilde{Y}}^*_{(j)}$, and the estimated target prediction $R\left(\hat{\tilde{Y}}^*_{(j)}\right)$ for the corresponding $R\left(\tilde{Y}^*_{(j)}\right)$. Because we had total of r pseudo-data sets, we now have r $R\left(\hat{\tilde{Y}}^*_{(j)}\right)$ for the pseudo-data sets.



Similar to $\tilde{S}$ and $\tilde{Y}$, the **pseudo-forecast** denoted $R\left(\tilde{Y}^*_{(j)}\right)$ have the same stochastic properties as $R(Y^*)$. Note that because the algebraic structure of the underlying model is always the same as the model applied to the original dataset Y, and the only differences are the parameter estimates, the variation between the $R(Y^*)$ and $R\left(\tilde{Y}^*_{(j)}\right)$ thus reflect parameter error introduced in Chapter 4. Recall that parameter errors are errors caused by inaccurate model parameter estimates, that is, variations that result from the differences between β and $\hat{\beta}$. Mathematically this is expressed as

$$\varepsilon_{parameter} = E[Y;\beta] - E\left[Y;\hat{\beta}\right]$$

Recall also from Chapter 4, prediction error is composed of both parameter error and process error (assuming there is no model error). Therefore to create pseudo-forecasts that reflect prediction error, we need to add noise to the estimated target predictions $R\left(\hat{\tilde{Y}}^*_{(j)}\right)$. We can find the noise, or process error, of $R(Y^*)$ using re-sampling as well.

Recall the process error is the difference between the observations and fitted values (or future estimates). In this case with future estimates, denote the process error for the *i*-th component of Y^* as

$$\varepsilon_i^* = Y_i^* - E[Y_i^*]$$

(5-19)

Which can be rewritten as

$$Y_i^* = E[Y_i^*] + \varepsilon_i^*$$

(5-20)

For the *i*-th component of $\tilde{Y}_{(j)}^*$, we can have $E[\tilde{Y}_i^*]$ estimated using parameters $\hat{\tilde{\beta}}$. In order to obtain a set of process errors for $\tilde{Y}_{(j)}^*$ that has the same properties as the set of $\{\varepsilon_i^*\}$, we draw a second vector $\tilde{S}_{proc}$ the same way $\tilde{S}$ was drawn, and form the pseudo-observation vector $\tilde{Y}_{proc}$ similar to (5-17), i.e., for the *i*-th component of $\tilde{S}_{proc}$, we have

$$\tilde{Y}_{proc,i} = S^{-1}\left(\tilde{S}_{proc,i}; \hat{Y}\right)$$

We then define the vector of process error as

$$\varepsilon_{proc}^* = \tilde{Y}_{proc} - \hat{Y}$$

(5-21)

Note this is different from the monograph, which has $\hat{Y}_{proc} - \hat{Y}$ instead of $\tilde{Y}_{proc} - \hat{Y}$. However, relating the context of prior and following discussion, this is a typo in the monograph, and (5-21) should have the form described here.

From (5-21), we can conclude that the components of the vector ε_{proc}^* have the same properties as the collection $\{\varepsilon_i^*\}$. We can repeat the procedure of drawing $\tilde{S}_{proc}$ and (5-21) to obtain *r* replicates of ε_{proc}^*. Note that as ε_{proc}^* reflect the process error for future estimates, the dimension of ε_{proc}^* should reflect that of the future estimates Y^*, not necessarily the past data vectors Y.

When we work with Pearson residuals, recall from (5-18), for the *i*-th component of $\tilde{Y}$,

$$\tilde{Y}_i = S^{-1}\left(\tilde{S}_i; \hat{Y}\right) = \hat{Y}_i + \hat{\sigma}_i \tilde{S}_i$$

Therefore after drawing *r* samples of residual vectors, $\tilde{S}_{proc}$, we obtain *r* samples of pseudo data sets of $\tilde{Y}_{proc}$, where the *i*-th component of each $\tilde{Y}_{proc}$ is defined as

$$\tilde{Y}_{proc,i} = \hat{Y}_i - \hat{\sigma}_i \tilde{S}_{proc,i}$$

So using (5-21), the i-th component of ε^*_{proc} becomes

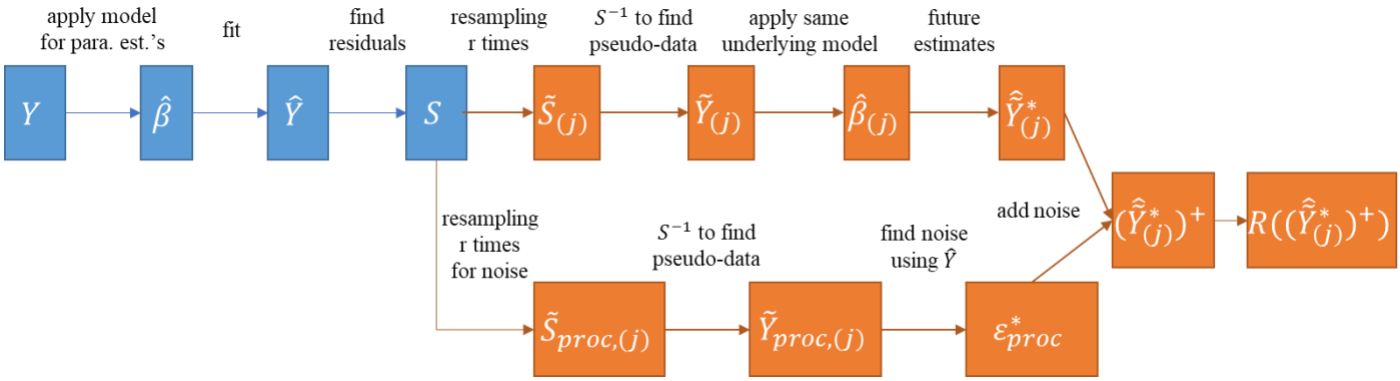
$$\varepsilon^*_{proc,i} = \tilde{Y}_{proc,i} - \hat{Y}_i = (\hat{Y}_i + \hat{\sigma}_i \tilde{S}_i) - \hat{Y}_i = \hat{\sigma}_i \tilde{S}_i$$

Now we can add noise to the future estimates for the pseudo data sets, which are simply the addition of our original estimates and the process errors:

$$\left(\hat{\tilde{Y}}^*_{(j)}\right)^+ = \hat{\tilde{Y}}^*_{(j)} + \varepsilon^*_{proc(j)}$$

(5-23)

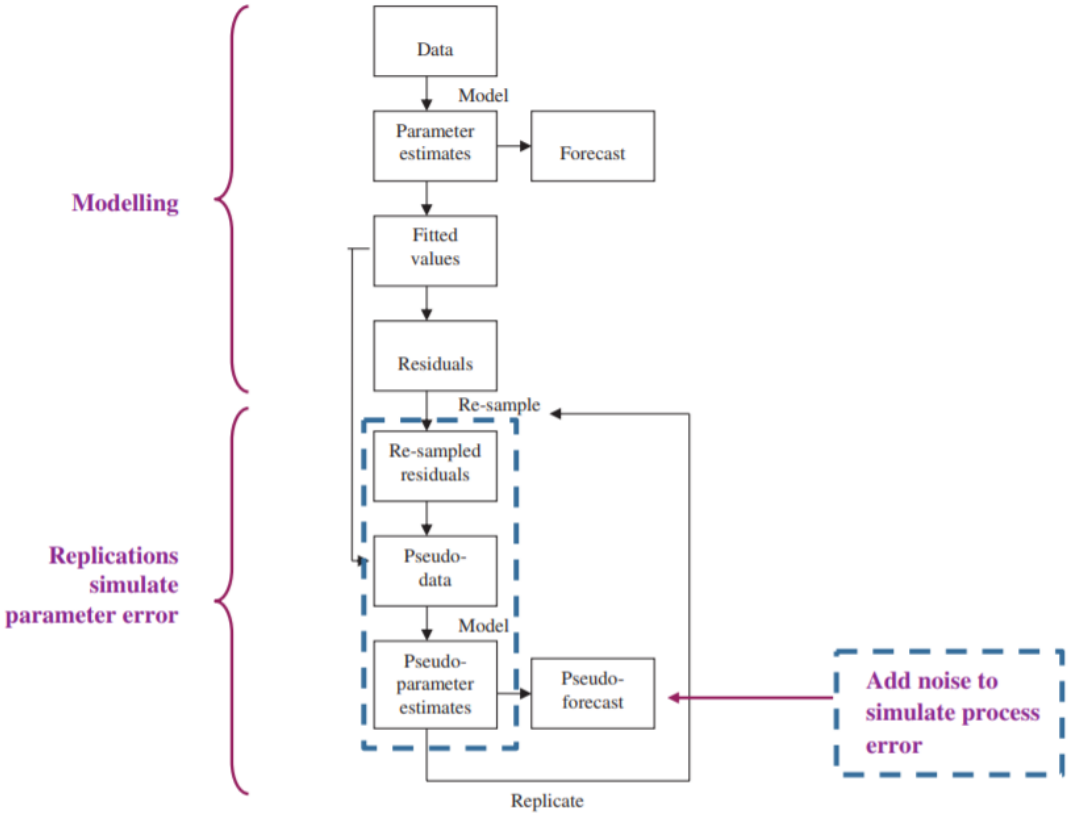
Where $\left(\hat{\tilde{Y}}^*_{(j)}\right)^+$ is a pseudo-forecast that contains both process and parameter errors as prediction error. We can then obtain the target prediction $R\left(\left(\hat{\tilde{Y}}^*_{(j)}\right)^+\right)$ that include process error.



These $R\left(\left(\hat{\tilde{Y}}^*_{(j)}\right)^+\right), j = 1, \ldots, r$ are iid, with the same distribution as $R(Y^*)$. Thus the r replicates form an empirical distribution of $R(Y^*)$, and we can achieve PoA or find other stochastic properties such as MSEP from the distribution.

Figure 5-1 from the monograph (shown below) also summarizes procedures discussed above. Like the semi-parametric bootstrap, parametric bootstrapping also involves re-sampling, but using a different approach, with assumptions of not only the underlying model for observations, but known distributions of observations and parameter estimates as well.

Figure 5-1. Diagrammatic Representation of the Semi-Parametric Bootstrap



5.2.2. Parametric Bootstrap

While semi-parametric bootstrapping is based on empirical residuals and resamples via inverse transform of the residuals, parametric bootstrapping is based on the assumption of parameter estimates with an underlying distribution with appropriate variance and known distribution of the dataset.

Parametric Estimates

Because the parameter estimates $\hat{\beta}$ for GLMs are usually MLEs, for parametric bootstrapping we assume the original $\hat{\beta}$ to be MLEs. Also, it is known that an MLE is an asymptotically normal unbiased estimator for indefinitely increasing sample size. In other words, when the number of parameter estimates generated approaches infinity, the distribution of all parameter estimates $\hat{\beta}$ becomes closer to normal. i.e.,

$$\hat{\beta} \sim N\left(\beta, Var\left[\hat{\beta}\right]\right), \qquad \text{asymptotically} \tag{5-24}$$

Assume the asymptotic relation above holds for finite data sample, then we can assume $\hat{\beta}$ approximately normal, which means

$$\hat{\beta} \sim N(\beta, \hat{C})$$

(5-25)

where C denotes the variance of the parameter estimates, $Var[\hat{\beta}]$. Recall $\hat{\beta}$ is a vector of all parameter estimates, and thus the variance estimate $\hat{C}$ is a $p \times p$ matrix that contains all the estimated correlations between each pair of parameter estimates $\hat{\beta}_i$ and $\hat{\beta}_j$, for $i, j = 1, \dots p$. Therefore $\hat{C}$ has diagonal entries of 1 (correlations between $\hat{\beta}_i$ and $\hat{\beta}_i$ are always 1), and the rest elements being the correlations. An example is the correlation matrix for ODP CC model of Table 5-2, section 5.1. with the assumed distribution of $\hat{\beta}$, we can sample the parameter estimate replicates $\hat{\beta}_{(j)}$ directly.

The sampling process follows 3 steps:

- apply a linear transformation M to $\hat{\beta}$ such that the components of $M\hat{\beta}$ are uncorrelated; in other words, we obtain a variance matrix for the linear transformation of $\hat{\beta}$, $Var[M\hat{\beta}]$, such that the correlations between each pair of $(M\hat{\beta})_i$ and $(M\hat{\beta})_j$ for $i \neq j$ are 0.

- sample each of these components from a univariate normal distribution to obtain a random vector γ;

- apply the inversion of linear transform M to the sampled vector γ to obtain the required sampling from $N(\hat{\beta}, \hat{C})$, i.e. the re-sampled $\hat{\beta}_{(j)}$.

Mathematically, for step 1 we need to find the linear transformation matrix M such that $Var[M\hat{\beta}] = \Lambda$, where Λ is a diagonal matrix:

$$M\hat{C}M^T = \Lambda = diag(\lambda_1, \dots, \lambda_p)$$

(5-26)

After step 1, the multivariate normal distribution of vector $\hat{\beta}$ is transformed to the normal distributions for each element of the linear transformation $M\hat{\beta}$. Thus in step 2 we can make random drawings of γ_i, which satisfy

$$\gamma_i \sim N\left((M\hat{\beta})_i, \lambda_i\right), i = 1, 2, \dots, p$$

(5-27)

After the step, we have a vector of $\gamma = (\gamma_1, \dots, \gamma_p)^T$.

Now we apply step 3, which uses the inverse transform to obtain parameter estimate replicates $\hat{\beta}_{(j)}$:

$$\hat{\beta}_{(j)} = M^{-1}\gamma$$

(5-28)

We can verify that $\hat{\beta}_{(j)} \sim N(\hat{\beta}, \hat{C})$ using the (5-27) and (5-28), where we derive the mean and variance as

$$E\left[\hat{\beta}_{(j)}\right] = E[M^{-1}\gamma] = M^{-1}E[\gamma] = M^{-1}M\hat{\beta} = \hat{\beta}$$

(5-29)

and

$$Var\left[\hat{\beta}_{(j)}\right] = Var[M^{-1}\gamma] = M^{-1}Var[\gamma](M^{-1})^T = M^{-1}[M\hat{C}M^T](M^{-1})^T = \hat{C}$$

(5-30)

Note here the operation for variance is $Var[MA] = MAM^T$, where A is a matrix, and M is a linear transformation applied to A.

The process of identifying M can be done by conventional statistical software by decomposition of $\hat{C}$. Namely 2 tools are **Cholesky decomposition** and **spectral decomposition**.

Cholesky decomposition of $\hat{C}$:

$$\hat{C} = LL^T$$

(5-31)

where L is the lower triangular matrix, this is equivalent to (5-26) with $M = L^{-1}$ and $\Lambda = I$.

Spectral decomposition of $\hat{C}$:

$$\hat{C} = P\Lambda P^T$$

(5-32)

where P is an orthogonal matrix and $\gamma_1, \ldots, \gamma_p$ are the eigenvalues of $\hat{C}$, this is equivalent to (5-26) with $M = P^{-1} = P^T$.

Process Error

In the parametric bootstrap method, the pseudo-data sets for process errors $\tilde{Y}_{proc}$ can be obtained by random drawings from the distribution that the original dataset Y assumes. For example, if we assume Y_i follow ODP distribution, then we can obtain each component of $\tilde{Y}_{proc}$ by random drawings from an ODP distribution with known mean $\tilde{Y}_{\mathrm{i}}$ and scale $\hat{\phi}/w_i$. and process error can be obtained similar to semi-parametric bootstrapping using $\varepsilon^*_{proc} = \tilde{Y}_{proc} - \hat{Y}$.

Discussion

The parametric bootstrapping is simpler to implement than semi-parametric bootstrapping due to shorter computational times. However, with the underlying distributional assumptions, the validity of parametric bootstrapping can decrease when:

- the sample size is so small that (5-24) is not asymptotic; and/or
- the error structure assumed within the GLM becomes a poor representation of data.

Appendix A. 3.3.1. Design matrix X of ODP Mack GLM for K=10 and J=10

1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0

47 rows in total
Columns 1-9 represents the x variates corresponding to variables f_1 to f_9

Appendix B. 3.3.2. Design matrix X of ODP CC GLM when K=10 and J=10

X=

1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0

55 rows in total

Columns 1-10 represents the x variates corresponding to variables from $\ln(\alpha_1)$ to $\ln(\alpha_{10})$

Columns 11-20 represents the x variates corresponding to variables from $\ln(\beta_1)$ to $\ln(\beta_{10})$

Appendix C 3.3.3. ODP Mack Model GENMOD codes

First Method:

Codes:

```
data ODPMackModelOneZero;
input Y f1 f2 f3 f4 f5 f6 f7 f8 f9;
datalines;
1.830420124 1 0 0 0 0 0 0 0 0
1.263187459 0 1 0 0 0 0 0 0 0
1.165103364 0 0 1 0 0 0 0 0 0
1.100166871 0 0 0 1 0 0 0 0 0
1.047794622 0 0 0 0 1 0 0 0 0
1.036767254 0 0 0 0 0 1 0 0 0
1.027791394 0 0 0 0 0 0 1 0 0
1.024821516 0 0 0 0 0 0 0 1 0
1.020856984 0 0 0 0 0 0 0 0 1
1.819959723 1 0 0 0 0 0 0 0 0
1.278843741 0 1 0 0 0 0 0 0 0
1.162150108 0 0 1 0 0 0 0 0 0
1.083202849 0 0 0 1 0 0 0 0 0
1.05228735 0 0 0 0 1 0 0 0 0
1.038268587 0 0 0 0 0 1 0 0 0
1.030858833 0 0 0 0 0 0 1 0 0
1.024908144 0 0 0 0 0 0 0 1 0
1.91165623 1 0 0 0 0 0 0 0 0
1.274917853 0 1 0 0 0 0 0 0 0
1.141208739 0 0 1 0 0 0 0 0 0
1.079065142 0 0 0 1 0 0 0 0 0
1.056618024 0 0 0 0 1 0 0 0 0
1.035551167 0 0 0 0 0 1 0 0 0
1.031534614 0 0 0 0 0 0 1 0 0
1.864788388 1 0 0 0 0 0 0 0 0
1.253238542 0 1 0 0 0 0 0 0 0
1.155990045 0 0 1 0 0 0 0 0 0
1.092481961 0 0 0 1 0 0 0 0 0
1.062448216 0 0 0 0 1 0 0 0 0
1.043125153 0 0 0 0 0 1 0 0 0
1.914140476 1 0 0 0 0 0 0 0 0
1.260856523 0 1 0 0 0 0 0 0 0
1.15733199 0 0 1 0 0 0 0 0 0
1.085336655 0 0 0 1 0 0 0 0 0
1.055708444 0 0 0 0 1 0 0 0 0
1.875594788 1 0 0 0 0 0 0 0 0
1.240571577 0 1 0 0 0 0 0 0 0
1.161373539 0 0 1 0 0 0 0 0 0
1.091812418 0 0 0 1 0 0 0 0 0
1.680537138 1 0 0 0 0 0 0 0 0
1.245449186 0 1 0 0 0 0 0 0 0
1.164983214 0 0 1 0 0 0 0 0 0
1.692799746 1 0 0 0 0 0 0 0 0
```

```
1.278442272     0       1       0       0       0       0       0       0       0
1.766681989     1       0       0       0       0       0       0       0       0
;
run;
proc genmod data=ODPMackModelOneZero;
model Y = f1 f2 f3 f4 f5 f6 f7 f8 f9 / NOINT SCALE = PEARSON;
run;
```

[Codes End]

Second Method:

Codes:

```
data ODPMackModelxy;
input Y f1 f2 f3 f4 f5 f6 f7 f8 f9;
datalines;
76550 41821 0 0 0 0 0 0 0 0
96697 0 76550 0 0 0 0 0 0 0
112662 0 0 96697 0 0 0 0 0 0
123947 0 0 0 112662 0 0 0 0 0
129871 0 0 0 0 123947 0 0 0 0
134646 0 0 0 0 0 129871 0 0 0
138388 0 0 0 0 0 0 134646 0 0
141823 0 0 0 0 0 0 0 138388 0
144781 0 0 0 0 0 0 0 0 141823
87662 48167 0 0 0 0 0 0 0 0
112106 0 87662 0 0 0 0 0 0 0
130284 0 0 112106 0 0 0 0 0 0
141124 0 0 0 130284 0 0 0 0 0
148503 0 0 0 0 148503 0 0 0 0
154186 0 0 0 0 0 148503 0 0 0
158944 0 0 0 0 0 0 154186 0 0
162903 0 0 0 0 0 0 0 158944 0
99517 52058 0 0 0 0 0 0 0 0
126876 0 99517 0 0 0 0 0 0 0
144792 0 0 126876 0 0 0 0 0 0
156240 0 0 0 144792 0 0 0 0 0
165086 0 0 0 0 156240 0 0 0 0
170955 0 0 0 0 0 165086 0 0 0
176346 0 0 0 0 0 0 170955 0 0
106761 57251 0 0 0 0 0 0 0 0
133797 0 106761 0 0 0 0 0 0 0
154668 0 0 133797 0 0 0 0 0 0
168972 0 0 0 154668 0 0 0 0 0
179524 0 0 0 0 168972 0 0 0 0
187266 0 0 0 0 0 179524 0 0 0
113342 59213 0 0 0 0 0 0 0 0
142908 0 113342 0 0 0 0 0 0 0
165392 0 0 142908 0 0 0 0 0 0
179506 0 0 0 165392 0 0 0 0 0
189506 0 0 0 0 179506 0 0 0 0
111551 59475 0 0 0 0 0 0 0 0
138387 0 111551 0 0 0 0 0 0 0
160719 0 0 138387 0 0 0 0 0 0
175475 0 0 0 160719 0 0 0 0 0
110255 65607 0 0 0 0 0 0 0 0
137317 0 110255 0 0 0 0 0 0 0
159972 0 0 137317 0 0 0 0 0 0
96063 56748 0 0 0 0 0 0 0 0
122811 0 96063 0 0 0 0 0 0 0
92242 52212 0 0 0 0 0 0 0 0
;
```

```
run;
proc genmod data=ODPMackModelxy;
model Y = f1 f2 f3 f4 f5 f6 f7 f8 f9 / NOINT SCALE = PEARSON;
run;
```

[Codes End]

Appendix D. 3.3.3. ODP CC Model GENMOD codes

Code:

```
data ODPCCModel;
input Y a1 a2 a3 a4 a5 a6 a7 a8 a9 a10 b1 b2 b3 b4 b5 b6 b7 b8 b9 b10;
datalines;
41821 1 0 0 0 0 0 0 0 0 0 1
      0 0 0 0 0 0 0 0 0
34729 1 0 0 0 0 0 0 0 0 0 0
      1 0 0 0 0 0 0 0 0
20147 1 0 0 0 0 0 0 0 0 0 0
      0 1 0 0 0 0 0 0 0
15965 1 0 0 0 0 0 0 0 0 0 0
      0 0 1 0 0 0 0 0 0
11285 1 0 0 0 0 0 0 0 0 0 0
      0 0 0 1 0 0 0 0 0
5924  1 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 1 0 0 0 0
4775  1 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 1 0 0 0
3742  1 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 1 0 0
3435  1 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 1 0
2958  1 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 1
48167 0 1 0 0 0 0 0 0 0 0 1
      0 0 0 0 0 0 0 0 0
39495 0 1 0 0 0 0 0 0 0 0 0
      1 0 0 0 0 0 0 0 0
24444 0 1 0 0 0 0 0 0 0 0 0
      0 1 0 0 0 0 0 0 0
18178 0 1 0 0 0 0 0 0 0 0 0
      0 0 1 0 0 0 0 0 0
10840 0 1 0 0 0 0 0 0 0 0 0
      0 0 0 1 0 0 0 0 0
7379  0 1 0 0 0 0 0 0 0 0 0
      0 0 0 0 1 0 0 0 0
5683  0 1 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 1 0 0 0
4758  0 1 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 1 0 0
3959  0 1 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 1 0
52058 0 0 1 0 0 0 0 0 0 0 1
      0 0 0 0 0 0 0 0 0
47459 0 0 1 0 0 0 0 0 0 0 0
      1 0 0 0 0 0 0 0 0
27359 0 0 1 0 0 0 0 0 0 0 0
      0 1 0 0 0 0 0 0 0
17916 0 0 1 0 0 0 0 0 0 0 0
      0 0 1 0 0 0 0 0 0
11448 0 0 1 0 0 0 0 0 0 0 0
      0 0 0 1 0 0 0 0 0
```

8846	0	0	1	0	0	0	0	0	0	0	0
	0	0	0	0	1	0	0	0	0		
5869	0	0	1	0	0	0	0	0	0	0	0
	0	0	0	0	0	1	0	0	0		
5391	0	0	1	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	1	0	0		
57251	0	0	0	1	0	0	0	0	0	0	1
	0	0	0	0	0	0	0	0	0		
49510	0	0	0	1	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0		
27036	0	0	0	1	0	0	0	0	0	0	0
	0	1	0	0	0	0	0	0	0		
20871	0	0	0	1	0	0	0	0	0	0	0
	0	0	1	0	0	0	0	0	0		
14304	0	0	0	1	0	0	0	0	0	0	0
	0	0	0	1	0	0	0	0	0		
10552	0	0	0	1	0	0	0	0	0	0	0
	0	0	0	0	1	0	0	0	0		
7742	0	0	0	1	0	0	0	0	0	0	0
	0	0	0	0	0	1	0	0	0		
59213	0	0	0	0	1	0	0	0	0	0	1
	0	0	0	0	0	0	0	0	0		
54129	0	0	0	0	1	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0		
29566	0	0	0	0	1	0	0	0	0	0	0
	0	1	0	0	0	0	0	0	0		
22484	0	0	0	0	1	0	0	0	0	0	0
	0	0	1	0	0	0	0	0	0		
14114	0	0	0	0	1	0	0	0	0	0	0
	0	0	0	1	0	0	0	0	0		
10000	0	0	0	0	1	0	0	0	0	0	0
	0	0	0	0	1	0	0	0	0		
59475	0	0	0	0	0	1	0	0	0	0	1
	0	0	0	0	0	0	0	0	0		
52076	0	0	0	0	0	1	0	0	0	0	0
	1	0	0	0	0	0	0	0	0		
26836	0	0	0	0	0	1	0	0	0	0	0
	0	1	0	0	0	0	0	0	0		
22332	0	0	0	0	0	1	0	0	0	0	0
	0	0	1	0	0	0	0	0	0		
14756	0	0	0	0	0	1	0	0	0	0	0
	0	0	0	1	0	0	0	0	0		
65607	0	0	0	0	0	0	1	0	0	0	1
	0	0	0	0	0	0	0	0	0		
44648	0	0	0	0	0	0	1	0	0	0	0
	1	0	0	0	0	0	0	0	0		
27062	0	0	0	0	0	0	1	0	0	0	0
	0	1	0	0	0	0	0	0	0		
22655	0	0	0	0	0	0	1	0	0	0	0
	0	0	1	0	0	0	0	0	0		
56748	0	0	0	0	0	0	0	1	0	0	1
	0	0	0	0	0	0	0	0	0		
39315	0	0	0	0	0	0	0	1	0	0	0
	1	0	0	0	0	0	0	0	0		
26748	0	0	0	0	0	0	0	1	0	0	0
	0	1	0	0	0	0	0	0	0		

```
52212   0   0   0   0   0   0   0   0   1   0   1
        0   0   0   0   0   0   0   0   0
40030   0   0   0   0   0   0   0   0   1   0   0
        1   0   0   0   0   0   0   0   0
43962   0   0   0   0   0   0   0   0   0   1   1
        0   0   0   0   0   0   0   0   0
;
run;
proc genmod data=ODPCCModel;
model y = a1 a2 a3 a4 a5 a6 a7 a8 a9 a10 b2 b3 b4 b5 b6 b7 b8 b9 b10 /
NOINT
link = log
dist = poisson
SCALE = PEARSON
CORRB
;
```

[Codes End]

References:

Friedland, J., FCAS, FCIA,MAAA, FCA, and KPMG LLP. 2010. *Estimating Unpaid Claims Using Basic Techniques*. Casualty Actuarial Society, Arlington VA.

Taylor, G. 2000. *Loss Reserving: An Actuarial Perspective.* Dordrecht, Netherlands: Kluwer Academic Publishers.

Taylor, G. 2009. "The Chain Ladder and Tweedie Distributed Claims Data." *Variance* 3: 96–104.

Taylor, G. 2015. "Bayesian Chain Ladder Models." *ASTIN Bulletin* 45(1): 75–99.

Taylor, G., and G. McGuire. 2004. "Loss Reserving with GLMs: A Case Study." Casualty Actuarial Society 2004 Discussion Paper Program, 327–392.

Taylor, G., G. McGuire, and J. Sullivan. 2008. "Individual Claim Loss Reserving Conditioned by Case Estimates," *Annals of Actuarial Science* 3(1&2): 215–256.

Taylor, G., and G. McGuire. 2016. "Stochastic Loss Reserving Using Generalized Linear Models". CAS Monograph Series, Number 3. Casualty Actuarial Society, Arlington VA.

Taylor, G., and J. Xu. 2016. "An Empirical Investigation of the Value of Finalisation Count Information to Loss Reserving," Variance (in press).