Analysis of a Paid Loss Triangle

Regression Analysis, Summer 2010

Ronald Guida

1 Introduction

In 2010, Simulated Auto Insurance, Inc. has requested an analysis of the payout patterns resulting from accident claims. A single car accident, depending on its severity, can generate claims for several years. Payment data is collected over a period of several years and categorized according to the year of each car accident and the year in which losses were paid. The resulting table of aggregate data is called a paid loss triangle. Simulated Auto Insurance has a paid loss triangle that needs to be analyzed to estimate the inflation patterns associated with car accident claims.

For this project, I simulated and analyzed a paid loss triangle. I performed my simulation and analysis using the free statistical package R.

2 Simulating a Paid Loss Triangle

Losses can be described according to three dimensions: the calendar year in which losses were paid, the year of the accident that caused the losses (the accident year), and the number of years elapsed since the accident (the development year). These three dimensions are perfectly collinear; the calendar year is always the sum of the accident year and the development year. A paid loss triangle is constructed by collecting payment data over a period of several years and then categorizing it according to the year of each car accident and the year in which losses were paid.

By selecting any two dimensions as a pair of axes, the aggregate losses from all claims can be displayed in a table. The entries in this table will form a triangle; the "missing" entries correspond to future losses due to past accidents. Any pair of dimensions can be used to uniquely select a cell in the paid loss table. For an incremental paid loss table, each cell will list the total losses paid during one particular calendar year for accidents that happened in one particular accident year. These losses can be described in terms of an inflation pattern, a payout pattern, and an exposure pattern.

The exposure pattern determines the number of insured cars for each accident year, their risk classes, and the distribution of accidents for the year. Each accident generates a set of latent loss payments, where each latent payment corresponds to a single real-valued payment to be made at some future time after the accident. For each latent payment, a payout pattern determines an associated cost distribution and an associated time-of-occurrence distribution; the cost will have to be multiplied by a price index to incorporate the effect of inflation.

Based on these ideas, it is possible to construct a fine-grained simulation of auto accidents and their losses. For a single accident year, each risk class of drivers would require a counting process (e.g. a Poisson process) to determine the number of accidents that occur. For each accident, another series of random draws would determine the characteristics of the accident (e.g. the speed of collision, number of passengers, etc), and then these characteristics would determine the parameters of several more random draws to compute the number, timing, and magnitude of paid losses.

A software-based simulator with this level of complexity is not hard to construct, if the appropriate models and parameters are readily available. On the other hand, selecting the appropriate models and parameters for each simulated random variable is likely to be arduous.

Since a fine-grained simulator, at this level of complexity, is far beyond the scope and purpose of this project, massive simplifications need to be made. The simulator will be simplified by throwing away almost the entire structure of accidents and loss payments. Instead, the simulator will compute the expectation of each incremental loss and then incorporate random noise to generate a loss triangle.

The simplified simulator is defined by an inflation pattern and a payout pattern.

An inflation pattern is defined by a price index, or a discount factor: let v^m equal the discount factor for calendar year m. It is assumed that the price index, $1/v^m$, is constant for each year, with a discrete jump at the end of the year. If the annual inflation rates are $i_1, i_2, \ldots$, then the discount factors are

$$v^0 = 1,\ v^m = \prod_{k=1}^{m} (1 + i_k)^{-1}$$

A payout pattern is defined by a cumulative distribution function $F_n(t)$ for each accident year, where $F_n(t)$ is the expected inflation-corrected cumulative proportion of losses incurred within the first t years after an accident during year n. This loosely corresponds to the distribution function of the time-of-loss random variable $T(0)$ for a randomly selected latent payment due to an accident in year n, except that the distribution is defined to incorporate the magnitude of latent payments in addition to their timing.

It is assumed that the payout pattern is the same for all accident years, $F_n(t) = F(t)$. The expected incremental proportion of inflation-corrected losses incurred during calendar year m due to an accident in year n is $F(m-n+1) - F(m-n)$, or ${}_{m-n|1}q_0$.

An exposure pattern is avoided by assuming that the number of exposure units is a known quantity for each accident year. Real loss data would be adjusted by dividing the losses for each accident year by some measure of exposure for that year. For simulation purposes, exposure is assumed to be a constant; let S equal the the expectation of inflation-corrected cumulative future losses for one "unit" of exposure.

With these definitions, the expected nominal loss paid during calendar year m, for an accident that happens in year n, is given by

$$E[L_{n,m}] = S \, \frac{{}_{m-n|1}q_0}{v^m}$$

To simulate an incremental paid loss triangle, two additional pieces of information are needed: the number of years, ω, to simulate, and the amount of noise, or volatility, σ, to incorporate into each observed incremental paid loss. Each observed loss is computed by multiplying the expected loss by a zero-mean log-normal variate whose underlying normal distribution has a variance of σ^2. Letting $Z_{n,m}$ denote a set of independent standard normal variates, the observed incremental losses are given by

$$L_{n,m} = S \, \frac{{}_{m-n|1}q_0}{v^m} \, \exp\left(-\frac{\sigma^2}{2} + \sigma Z_{n,m}\right)$$

3 Data to be Analyzed

I simulated a 10 by 10 paid loss triangle with the following parameters: I used a geometric payout pattern with a 5% decay rate, ${}_{k|1}q_0 = (0.05)(0.95)^k$, I used a linearly increasing inflation rate of 3% for the first year, increasing by 0.5% each year thereafter, and I used a volatility of 2% to generate observed losses. The inflation rate for the last year is 7%.

$$i_k = 0.03 + 0.005(k-1) \quad (k = 1, 2, \ldots, 9) \qquad \sigma = 0.02$$

The R script `generator.R` generates a new data-set with these parameters. The simulated data is in the file `data.txt`.

Having generating this data-set, I need to perform the role of an analyst. A real-world analyst can never discover the true values of statistical parameters, but I happen to know what they are because I generated my data through simulation. Therefore, I will need to pretend that I don't know the true parameters.

I am giving myself the following assumptions:

- The data represent incremental losses.
- The data have been corrected for business exposure.
- The payout pattern is geometric, with an unknown decay rate.
- The inflation pattern is unknown.
- Each incremental loss datum incorporates log-normal multiplicative noise with a constant variance.
- The goal of analysis is to estimate the inflation pattern and the payout decay rate.

4 Initial Analysis

My goal is to estimate the inflation pattern and the payment decay rate. A constant payment decay rate of d produces the payout pattern ${}_{k|1}q_0 = d(1-d)^k$.

I initially assumed that the inflation rate is constant, so $v^m = (1+i)^{-m}$. With this assumption, observed losses are

$$L_{n,m} = S\ d(1-d)^{m-n}\ (1+i)^m\ \exp\left(-\frac{\sigma^2}{2} + \sigma Z_{n,m}\right)$$

By taking logarithms, this equation can be transformed for linear regression. I will refer to this as model M_0.

$$\log L_{n,m} = \underbrace{\log S + \log d - \frac{\sigma^2}{2}}_{\alpha} + (m-n)\underbrace{\log(1-d)}_{\beta_1} + m\underbrace{\log(1+i)}_{\beta_2} + \underbrace{\sigma Z_{n,m}}_{\epsilon_{n,m}}$$

The response variable Y and the regressors X_1, X_2 are given by

$$\left.\begin{array}{rcl} Y_i & = & \log L_{n,m} \\ X_{i1} & = & m-n \\ X_{i2} & = & m \end{array}\right\} \text{ where } \left\{\begin{array}{l} 0 \le n \le m \le \omega - 1 \\ i = \omega n - \frac{n(n+1)}{2} + m + 1 \end{array}\right.$$

I estimated the parameters of model M_0 using linear regression, and obtained the following results. A few diagnostic tests will cast doubt on this estimated model.

$$\underset{R^2_{adj}=0.9574}{\hat{Y}} = \underset{(0.009755)}{-3.055181} + \underset{(0.001738)}{-0.051675}\ X_1 + \underset{(0.001738)}{0.053293}\ X_2$$

The estimated payout decay rate is $\hat{d} = 1 - e^{-0.051675}$, which is about 5.4%. To compute the endpoints of a 95% confidence interval for this estimate, I calculated the endpoints

$$1 - e^{-0.051675 \pm (1.96)\cdot 0.001738}$$

and obtained the interval $4.7\% \le \hat{d} \le 5.4\%$.

The estimated inflation rate is $\hat{i} = e^{0.053293} - 1$, which is about 5.5%. A 95% confidence interval for this estimate is $5.1\% \le \hat{i} \le 5.8\%$.

5 Diagnostics for M_0

To check the quality of model M_0, I examined its residuals and I performed an F-test of the assumption of a constant inflation rate.

5.1 Residuals for M_0

The first diagnostic test is an examination of the residuals of M_0. The upper-left panel of figure 5.1 shows a plot of studentized residuals against calendar year. The smoothed residual plot has a clear parabolic shape with a minimum around year 5 or 6. Estimated incremental losses are below actual losses at the beginning and end of the decade, and above actual losses in the middle of the decade. This suggests that the inflation rate is increasing over time.

When the residuals for each regression are plotted against development year, accident year, or fitted losses, the plots all look similar, and all of them show a weak "V" relationship. In all three cases, there are fewer data points toward the right-had side of the plot, thus the right-hand end of the lowess curve will be overly influenced by noise. Since all of these plots show weak relationships, I will investigate the relationship between incremental losses and calendar year first.

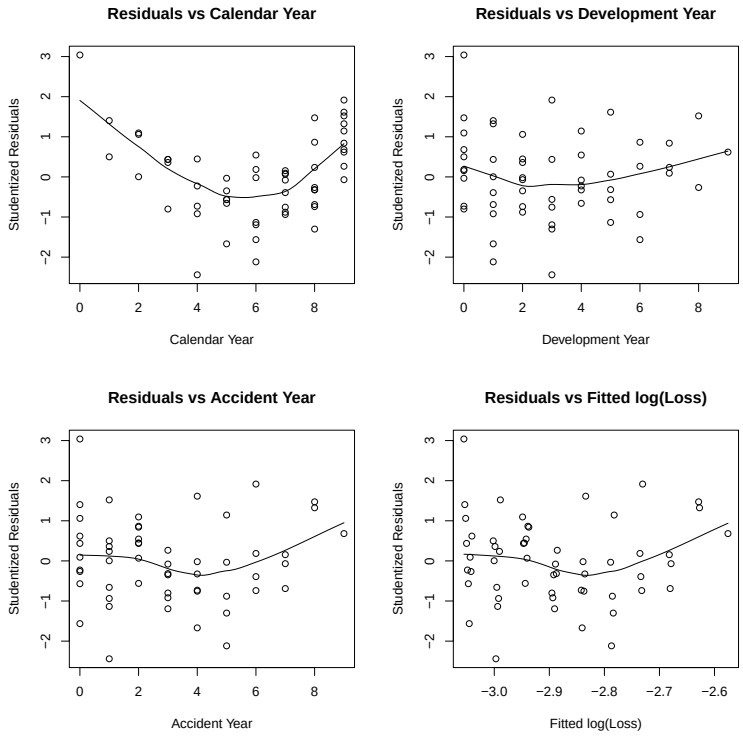

Figure 5.1: Residual Plots for Model M_0

5.2 F-Test for Constant Inflation

The second diagnostic test is an F-test to determine whether the rate of inflation is constant or changing. The null hypothesis is that the inflation rate is the same for all years, and the alternative is that the inflation rate is changing. The null hypothesis H_0 corresponds to model M_0, and the alternative hypothesis H_1 corresponds to model M_1. Model M_1 is a linear regression model that allows a different inflation rate for each year. This model is derived in Appendix A. Model M_0 is clearly nested within M_1.

$$\begin{aligned} H_0: &\quad i_1 = i_2 = \cdots = i_{\omega-1} \\ H_1: &\quad i_l \neq i_m \text{ for some pair of years } l, m \end{aligned}$$

When linear regression is used to estimate the parameters of model M_1, eight of the nine estimated annual inflation rates are statistically significant. Moreover, when the estimated inflation rates are plotted, there appears to be a clear upward trend, as shown in Figure 5.2.

The residual vector from model M_0 is a 52-dimensional vector with a squared-length of 0.038877. Model M_1 splits this residual vector into an 8-dimensional explained component with a squared-length of 0.021478, and a 44-dimensional unexplained component with a squared-length of 0.017399. Therefore, model M_1 explains 55% of the residuals from M_0, using 15% of the residual dimensions.

I would like to evaluate the probability that 15% of the residual dimensions would explain at least 55% of M_0's residuals, given the assumption that M_0's residuals have no remaining unexplained pattern (i.e. the residuals are independent normal with zero mean and constant variance). To evaluate this probability, I use the fact that the ratio of unexplained residuals per unexplained degree of freedom to explained residuals per explained degree of freedom,

$$\frac{\mathrm{ESS}_1/\mathrm{df}_1^{\text{residual}}}{(\mathrm{ESS}_0 - \mathrm{ESS}_1)/(\mathrm{df}_0^{\text{residual}} - \mathrm{df}_1^{\text{residual}})} = \frac{0.021478/8}{0.017399/44} = 6.7893,$$

is a realization of a distribution that doesn't depend on any unknowns, such as the unknown residual variance. The corresponding distribution is an F-distribution with 8 and 44 degrees of freedom.

The test statistic of 6.7893 has a p-value of about $9 \cdot 10^{-6}$, which is highly significant. There is chance of less than one in one-hundred-thousand that a model such as M_1 would explain as much as it did of M_0's residuals, if there is no residual pattern to be explained. Based on this result, the F-test rejects the null hypothesis H_0 and I conclude that the assumption of constant inflation is not adequate.

6 An Improved Model

Consider model M_1. This model has an adjusted R^2 of 97.75%, and eight of the nine estimated annual inflation rates are statistically significant. Moreover,

the residual plots for model M_1, shown in Figure 6.1, don't seem to show any significant patterns.

Keeping in mind that the adjusted R^2 already incorporates a penalty for having nine estimated inflation rates instead of one, it would appear that model M_1 is a very good model for this data-set. Nevertheless, the estimated inflation rates seem to follow a linear trend, and it would be prudent to try fitting a model that assumes a linearly increasing inflation rate.

To derive a model based on this assumption, I will start by assuming that inflation is continuous. With a constant annual inflation rate of i, the continuous inflation rate is $\delta_t = log(1+i)$. The corresponding discount factor at time t is

$$v^t = \exp\left[-\int_0^t \delta_t \, dt\right] = (1+i)^{-t}$$

and in particular, the discount factor for calendar year m, at the beginning of the year, is $v^m = (1+i)^{-m}$.

If inflation is increasing linearly, then the continuous inflation rate is $\delta_t = \beta_2 + \beta_3 t$ for some constants β_2 and β_3. The discount factor at time t is

$$v^t = \exp\left[-\int_0^t \delta_t \, dt\right] = e^{-\beta_2 t - \beta_3 t^2/2}$$

and in particular, the discount factor for calendar year m, at the beginning of the year, is $v^m = e^{-\beta_2 m - \beta_3 m^2/2}$.

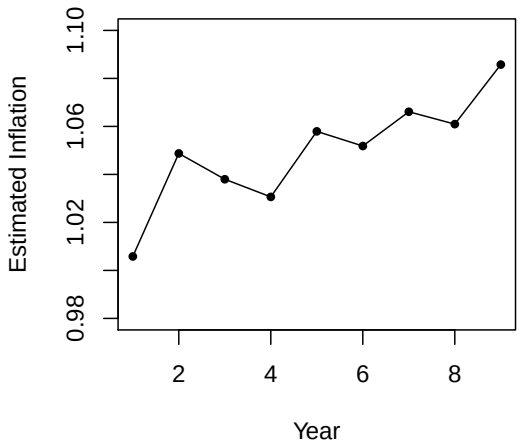

Figure 5.2: Estimated Annual Inflation Rates from Model M_1

Now, my incremental loss data-set does not include any information on how losses were distributed over the course of a year, so it would seem that I have to make an assumption here. I will assume that for each calendar year, all losses for that year are incurred at the beginning of the year in one lump sum; I'll call this "Assumption A". This is not a realistic assumption, but it is essentially equivalent to the assumption that the price index is constant during each year, with a discrete change at the end of the year. Moreover, any assumption that I might make, such as a uniform distribution of losses, will ultimately translate into a scale factor to be applied to the incremental losses. Since my response variable is the logarithm of incremental losses, that scale factor will show up in the intercept term, α, of my linear regression, where it will have no effect on my estimates of inflation or payout rates. Therefore, "Assumption A" is acceptable for my purposes.

A linearly increasing inflation rate is easily modeled by augmenting model M_0 with the square of the calendar year as an additional regressor. I will refer

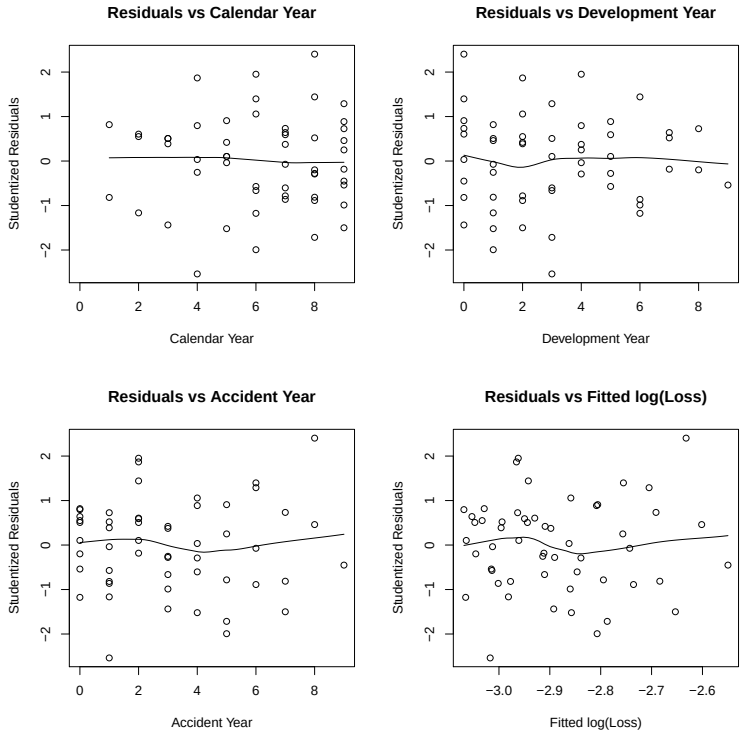

Figure 6.1: Residual Plots for Model M_1

to the following as model M_2.

$$\log L_{n,m} = \underbrace{\log S + \log d - \frac{\sigma^2}{2}}_{\alpha} + (m-n)\underbrace{\log(1-d)}_{\beta_1} + m\beta_2 + m^2\beta_3 + \underbrace{\sigma Z_{n,m}}_{\epsilon_{n,m}}$$

The response variable Y and the regressors X_1, X_2, X_3 are given by

$$\left.\begin{array}{rcl} Y_i & = & \log L_{n,m} \\ X_{i1} & = & m-n \\ X_{i2} & = & m \\ X_{i3} & = & m^2 \end{array}\right\} \text{ where } \left\{\begin{array}{l} 0 \le n \le m \le \omega - 1 \\ i = \omega n - \frac{n(n+1)}{2} + m + 1 \end{array}\right.$$

I estimated the parameters of model M_2 using linear regression, and obtained the following results.

$$\underset{R^2_{adj}=0.9792}{\hat{Y}} = \underset{(0.0113729)}{-2.9874711} + \underset{(0.0012153)}{-0.0516752}\ X_1 + \underset{(0.0046284)}{0.0200647}\ X_2 + \underset{(0.0004313)}{0.0031347}\ X_3$$

The estimated payout decay rate is $\hat{d} = 1 - e^{-0.0516752}$, which is about 5.0%. A 95% confidence interval for this estimate is $4.8\% \le \hat{d} \le 5.3\%$.

The estimated inflation rate for each year depends on a linear combination of the two coefficients β_2 and β_3. I can determine the estimated inflation rate for each year, but I cannot easily determine the corresponding confidence intervals. The estimated discount factor for year m is

$$\hat{i}_m = e^{\beta_2 + \beta_3(2m-1)} - 1$$

For year 1, $\hat{i}_1 = 2.3\%$. For year 9, $\hat{i}_9 = 7.6\%$

7 Diagnostics for M_2

The adjusted R^2 for model M_2 is 97.92%, which is only marginally better than the adjusted R^2 of 97.75% for model M_1. To check the quality of model M_2, I examined its residuals and I performed a pair of F-tests.

The first diagnostic test is an examination of the residuals of M_2. The residual plots, shown in Figure 7.1, don't seem to show any significant patterns. This is a good thing.

The second diagnostic test is an F-test of the hypothesis of a linear inflation rate. Consider the following three hypotheses:

$$\begin{array}{ll} H_0: & i_1 = i_2 = \cdots = i_{\omega-1} \\ H_2: & i_1, \ldots, i_{\omega-1} \text{ follow a linear relationship} \\ H_1: & i_l \ne i_m \text{ for some pair of years } l, m \end{array}$$

Each hypothesis (H_0, H_2, H_1) corresponds to the same-numbered model (M_0, M_2, M_1). The models are nested; M_0 is nested within M_2, which is nested within M_1. Model M_0 has 3 explanatory dimensions: two regressors and

an intercept. Model M_2 adds one explanatory dimension to M_0, and model M_1 adds 7 explanatory dimensions to M_2, leaving model M_1 with 44 residual dimensions.

Since I want to test the null hypothesis H_2 (linear inflation), I need to compare it to the the alternative H_1 (annually varying inflation). An F-test between models M_2 and M_1 yields a p-value of 0.8654; this means that there is an 87% chance that a model such as M_1 would explain as much of M_2's residuals. This F-test fails to reject the null hypothesis of linear inflation.

As a separate F-test, I want to compare model M_0 with M_2. Since I have already rejected H_0, there is no point in rejecting it again. An F-test between models M_0 and M_2 yields a p-value of about $7 \cdot 10^{-9}$. This would reject the null hypothesis H_0, compared to the alternative hypothesis H_2, but I do not intend to conduct a hypothesis test.

This p-value of $7 \cdot 10^{-9}$ is three orders of magnitude smaller than the p-value of $9 \cdot 10^{-6}$ obtained when comparing M_0 with M_1. Model M_2 actually explains 52% of M_0's residuals, using just one additional explanatory dimension. Model M_1 requires 8 additional explanatory dimensions to explain 55% of of M_0's

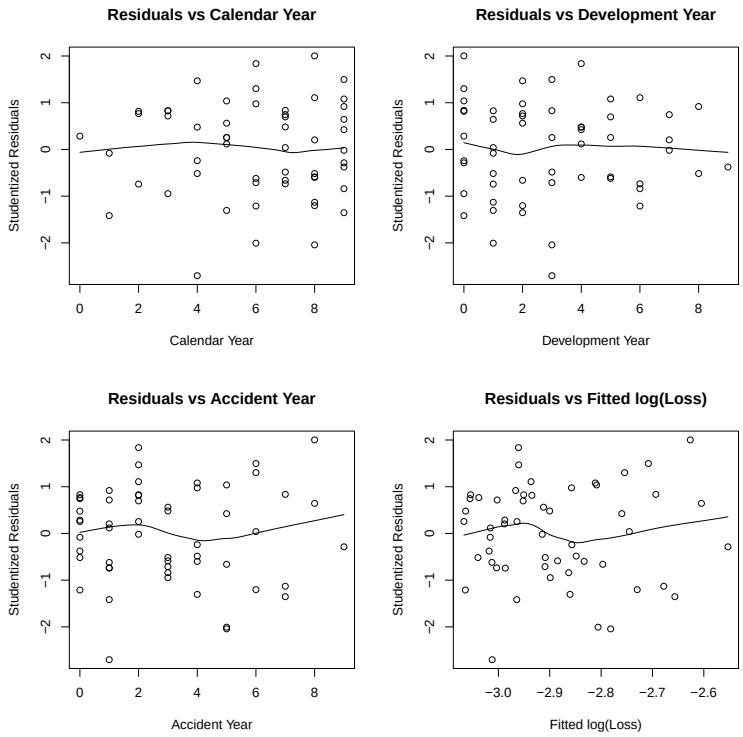

Figure 7.1: Residual Plots for Model M_2

residuals. Since model M_2 explains so much with so little, models such as M_2 are over one-thousand times less likely to occur by chance than models such as M_1.

8 Conclusion

If I compare model M_2 (linear inflation) with model M_1 (annually varying inflation), I notice there are few differences between them. Both models have about the same adjusted R^2, and neither model has unusual residuals. The main feature that makes model M_2 (linear inflation) "better" than model M_1 (annually varying inflation) is the fact that M_2 has fewer parameters and yet it is still just as good of a fit as M_1. Therefore, I conclude that model M_2 is the best model for my data-set, out of the models that I considered.

$$\hat{Y} = \underset{(0.0113729)}{-2.9874711} + \underset{(0.0012153)}{-0.0516752}\ X_1 + \underset{(0.0046284)}{0.0200647}\ X_2 + \underset{(0.0004313)}{0.0031347}\ X_3$$

9 Evaluation

In my simulation, the true payout decay rate is 5%, and the true inflation rate increases linearly from 3% to 7%. My final model, M_2, estimated a 5% decay rate, which is correct. It also estimated inflation rates that increase linearly from 2.3% to 7.6%; these estimates turned out to be fairly close to their true values.

I would like to note that I tried increasing the simulated volatility from 2% to 3%. With more volatility, the results became less obvious. I realized that I would tend to specifically look for evidence of a linear inflation trend, or I would bias my interpretation of data and plots in favor of such a trend, even though I'm not supposed to know in advance that such a trend exists. For this reason, I decided to stick with 2% volatility for my simulation.

Appendix A Derivation of Model M_1

I need to formulate a linear regression model that can handle a different inflation rate for each year. If the annual inflation rates are $i_1, i_2, \ldots$, then the price index for each year is $1/v^m$ where

$$v^0 = 1,\ v^m = \prod_{k=1}^{m} (1+i_k)^{-1},\ m = 1, 2, \ldots, \omega - 1$$

With this assumption, observed losses are

$$\begin{aligned} L_{n,m} &= S\, d(1-d)^{m-n}\, \tfrac{1}{v^m}\, \exp\left(-\tfrac{\sigma^2}{2} + \sigma Z_{n,m}\right) \\ &= S\, d(1-d)^{m-n}\, \textstyle\prod_{k=1}^{m}(1+i_k)\, \exp\left(-\tfrac{\sigma^2}{2} + \sigma Z_{n,m}\right) \end{aligned}$$

Since there is only one observation for calendar year zero, and more observations for later calendar years, I rearranged this equation to correct for inflation by working backwards from year $\omega - 1$. Letting $\mathrm{I}[c]$ denote the indicator function for condition c,

$$\begin{aligned} L_{n,m} &= S\, d(1-d)^{m-n}\, \tfrac{1}{v^{\omega-1}} \tfrac{v^{\omega-1}}{v^m}\, \exp\left(-\tfrac{\sigma^2}{2} + \sigma Z_{n,m}\right) \\ &= \tfrac{S}{v^{\omega-1}}\, d(1-d)^{m-n}\, \textstyle\prod_{k=m+1}^{\omega-1}(1+i_k)^{-1}\, \exp\left(-\tfrac{\sigma^2}{2} + \sigma Z_{n,m}\right) \\ &= \tfrac{S}{v^{\omega-1}}\, d(1-d)^{m-n}\, \textstyle\prod_{k=1}^{\omega-1}(1+i_k)^{-\mathrm{I}[k \geq m+1]}\, \exp\left(-\tfrac{\sigma^2}{2} + \sigma Z_{n,m}\right) \end{aligned}$$

Taking logarithms yields an equation suitable for linear regression. This equation will be referred to as model M_1.

$$\begin{aligned} \log L_{n,m} &= \underbrace{\log S + \log(v^{\omega-1}) + \log d - \frac{\sigma^2}{2}}_{\alpha} + (m-n)\underbrace{\log(1-d)}_{\beta_1} \\ &+ \textstyle\sum_{k=1}^{\omega-1} \underbrace{\log(1+i_k)}_{\beta_2,\ldots,\beta_\omega} (-\mathrm{I}[k \geq m+1]) + \underbrace{\sigma Z_{n,m}}_{\epsilon_{n,m}} \end{aligned}$$

The response variable Y and the regressors X_1, X_2 are given by

$$\left.\begin{array}{rcl} Y_i &=& L_{n,m} \\ X_{i1} &=& m-n \\ X_{ik} &=& -\mathrm{I}[m \leq k-2] \end{array}\right\} \text{ where } \left\{\begin{array}{l} 0 \leq n \leq m \leq \omega - 1 \\ i = \omega n - \frac{n(n+1)}{2} + m + 1 \\ 2 \leq k \leq \omega \end{array}\right.$$