James Pilarski
Regression Analysis Student Project
Spring 2010 Exam Semester

Introduction

Loss development triangles have long been relied upon to estimate ultimate loss and expense in property casualty insurance. Two commonly employed techniques are chain ladder development and the Bornhuetter-Ferguson method. Ultimate paid and incurred loss and expense amounts may be calculated using either of these approaches. Separate estimates of ultimate losses and expenses may also be determined. It is possible, if not practical, to produce up to 12 different sets of ultimate values from which a selection is made for each year in the experience period. Having numerous methods from which to choose does not increase an actuary's ability to make an informed judgment as to which estimate or combination of estimates is likely to be correct. Nor do any of these methods provide any type of diagnostic information that can used to determine how well the projected ultimate loss and expense amounts describe the trends affecting the underlying data. At best, an actuary can retrospectively assess the accuracy of their selections. This practice is unresponsive to changes in the nature of loss development and trend. It can also be misleading since a method that was the most reliable for one evaluation of a particular year is not necessarily representative of that year's performance at subsequent evaluations. A loss estimation process that relies on linear least-squares regression analysis provides a single value for the ultimate loss in each year and a set of diagnostic tools for assessing how well the data fit the proposed model.

Loss Development Model

Zehnwirth (1994) described a statistical model for loss development, which seeks to estimate the trends that affect how incremental paid losses change. Trends represent the percentage change over time along the three dimensions of a loss development triangle: accident year, development year, and calendar year. Losses aggregated according to the year in which a claim occurred are considered to be on an accident year basis. The trend affecting this dimension is generated by changes in the number of exposures insured. Regular historical evaluations of each accident year represent the development years, which are typically measured at the end of the accident year and every 12 months thereafter. The additional amount of loss paid at each evaluation of an accident year is the development year trend. This is also known as loss development and typically produces a decreasing amount of incremental paid loss as development years increase. Finally, a calendar year is comprised of the accident year plus the number of development years over which it has been evaluated. Inflation is the factor that causes incremental paid losses to change along this dimension. An equation describing this model is as follows:

Equation 1
$Y = \alpha * (\beta_1)^\wedge X_1 * (\beta_2)^\wedge X_2 * (\beta_3)^\wedge X_3 * e$

In this formula Y is the incremental paid loss, X_1 is the development year, X_2 the calendar year, and accident year X_3. The amount of incremental paid loss in the first accident year and development period is shown as α with the parameters β_1, β_2, and β_3 representing the trends of loss development, inflation, and exposure growth, respectively. Finally, a residual term, e, is added to the model to account for the fact that it cannot perfectly describe the underlying data. It is assumed the residuals are normally distributed, an assumption which may be subject to scrutiny once the data are fit to the model.

The purpose of this project is to use linear least-squares regression, an additive model, to estimate the trends in the incremental paid loss triangle. It is clear that the model proposed in Equation 1 is multiplicative. However, the natural logarithm of this model is additive and transforming it in this manner yields the following revised model.

Equation 2

$$\ln(Y) = \ln(\alpha) + X_1 * \ln(\beta_1) + X_2 * \ln(\beta_2) + X_3 * \ln(\beta_3) + \ln(e)$$

It will be necessary to calculate the natural logarithm of the incremental paid losses and understand that this transformation will need to be undone if one were to estimate incremental paid loss amounts beyond the evaluation date of the loss triangles. With this in mind, the data will be fit to this model using the R software environment provided by The R Project for Statistical Computing.

Model Fitting

The initial attempt to fit the data to the model outlined in Equation 2 and estimate the coefficients $\ln(\alpha)$, $\ln(\beta_1)$, $\ln(\beta_2)$, and $\ln(\beta_3)$ produced the results shown in Table 1.

Table 1: R Output for Equation 2

```
Call:
lm(formula = ln.incr.paid ~ DY + CY + AY, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0239 -0.4651  0.2093  0.5259  1.0711

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.03220    0.24936  60.283  < 2e-16 ***
DY          -0.19287    0.04443  -4.341 6.57e-05 ***
CY           0.09729    0.04443   2.190    0.033 *
AY                NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.699 on 52 degrees of freedom
Multiple R-squared: 0.266, Adjusted R-squared: 0.2378
F-statistic: 9.422 on 2 and 52 DF, p-value: 0.0003222

The coefficient associated with accident year is undefined. As was mentioned above, the calendar year equals accident year plus development year. Calendar year is clearly a linear combination of the other two explanatory variables. This is known as collinearity and a perfect linear relationship between variables will result in least-squares coefficients that do not have a unique solution. One means for coping with collinearity is model respecification. The model as it is currently specified attempts to estimate three kinds of trend that affect incremental paid losses. Restating the model to only estimate two of these trends will eliminate the perfect linear relationship. As it turns out, one of these trends is already known. The amount of incremental paid loss across accident years is related to exposure volume, assuming the exposure base is adequately descriptive of and responsive to the underlying risk. The exposures in each year are known and can be used to index the losses and remove the affect of this trend, where Y' denotes the adjusted incremental loss. The new model, with the accident year variable and coefficient removed, is outlined in Equation 3 with results displayed in Table 2.

Equation 3
$\ln(Y') = \ln(\alpha) + X_1 * \ln(\beta_1) + X_2 * \ln(\beta_2) + \ln(e)$

Table 2: R Output for Equation 3

```
Call:
lm(formula = ln.adj.incr.paid ~ DY + CY, data = df)

Residuals:
    Min      1Q   Median      3Q     Max
-1.66914 -0.43375  0.08704  0.52881  0.97994

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.67746    0.23845  61.555  < 2e-16 ***
DY          -0.15540    0.04248  -3.658 0.000594 ***
CY           0.05983    0.04248   1.408 0.165035
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6684 on 52 degrees of freedom
Multiple R-squared:  0.2075,	Adjusted R-squared:  0.177
F-statistic: 6.808 on 2 and 52 DF,  p-value: 0.002366
```

Removing accident year from the model has allowed for a unique solution to found and several conclusions may be drawn from the output of the regression model. The residual standard error of 0.6684 is small relative to the magnitude of response variable ln(Y') and the p-value of the F-statistic is rather close to zero. There is cause to reject the omnibus null hypothesis of $\ln(\beta_1) = \ln(\beta_2) = 0$ in favor of the alternative where at least one of the coefficients influences the model. While these metrics suggest the model is performing well, R^2 indicates that only 20.75% of the variation in the response variable is captured by its regression on development and calendar years. Examining the regression output at a more granular level is necessary in light of this observation.

The t-distribution statistics for the regression coefficients provide a more detailed look at how well this model fits the data. The null hypotheses of $\ln(\beta_1) = 0$ and $\ln(\beta_2) = 0$, if not rejected, would suggest the explanatory variables do not impact the regression. The p-values for the coefficients associated with $\ln(\alpha)$ and $\ln(\beta_1)$ are nearly zero indicating it is very unlikely they do not influence the response variable. Additionally, these parameters for the intercept and development year appear to be rather precisely estimated since they are at least several times greater than their standard errors in absolute terms. Calendar year does not yield statistics that are nearly as promising. A standard error that is large relative to the coefficient suggests the estimate of this parameter is not very precise. Also, the usefulness of this parameter in the model is suspect given that its p-value of 0.165035 is quite far from zero. Closer examination of the calendar year coefficient will be necessary to determine if it is possible to improve the fit and stability of this parameter. Various graphical representations of the data can be useful in this investigation.

A useful visual tool for evaluating the fit of a linear model is the residual plot. More specifically, examining the mean residual by year will highlight any inconsistencies that might be present in the data. Plots of the mean residual by development and calendar year are shown below in Figures 1 and 2. Both of these plots show that coefficients estimated for development and calendar year are not constant. A regression equation that accounts for differing trends is needed in order to improve the fit of the model.

Figure 1:

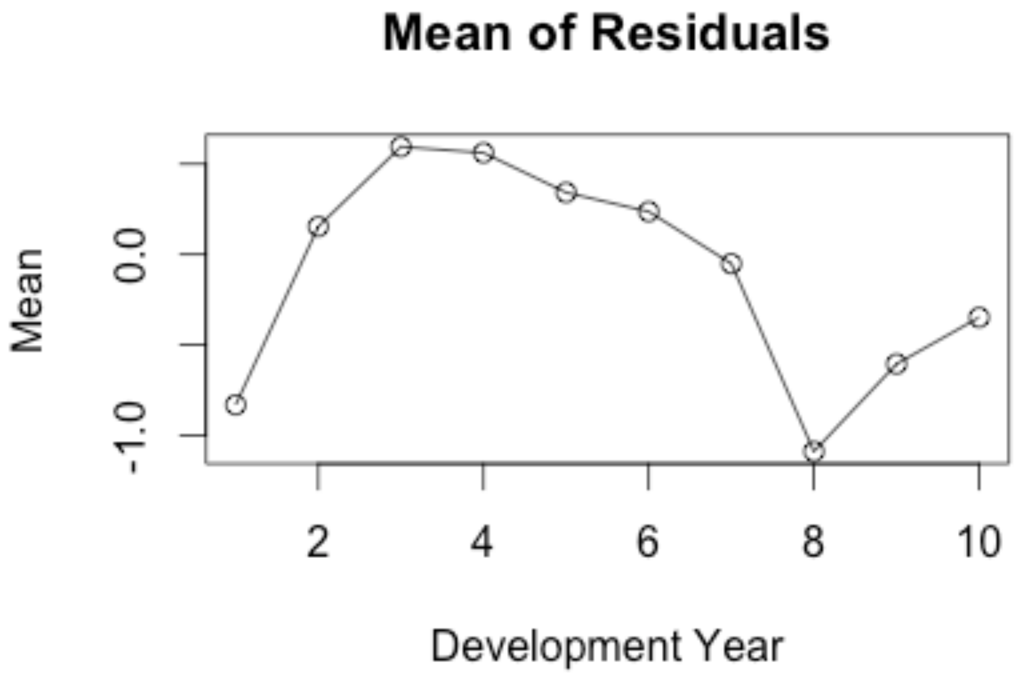
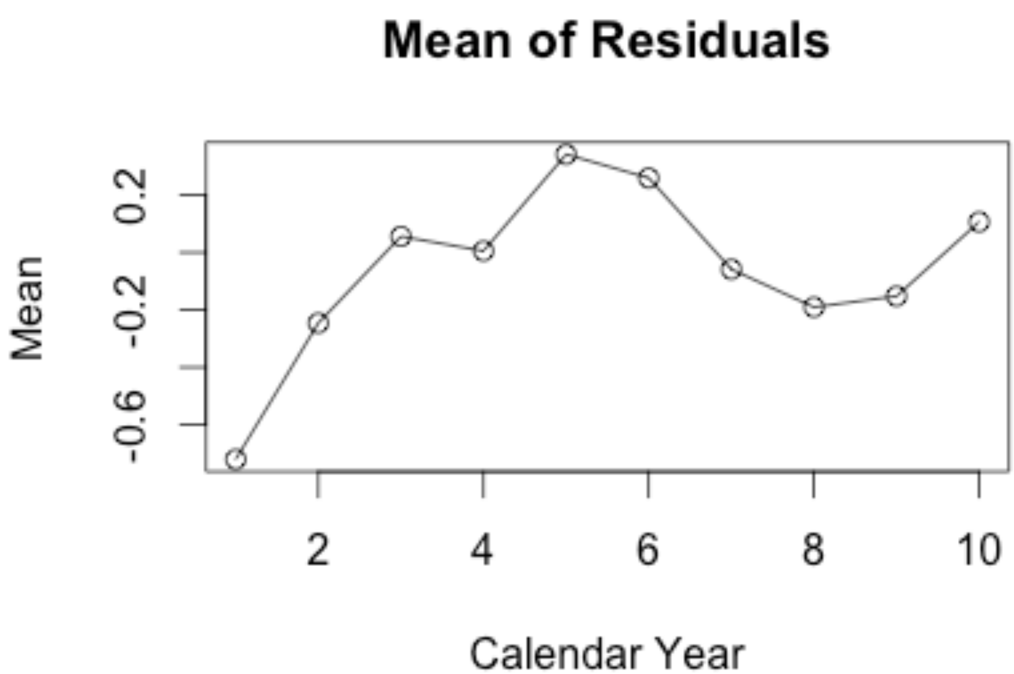


Figure 2:



Dummy Variables

Introducing a dummy variable makes it possible to control for qualitative differences in the data. Development year trend appears to change discretely after period 2 while 4 or 7 may be transition points for calendar year. A dummy variable for development year will be introduced first since the mean residual plot is more clearly defined. The variability in development years 9 and 10 is likely due to a lack of data points. This regression equation will be of the form:

Equation 4

$$\ln(Y') = \ln(\alpha) + X_1 * \ln(\beta_1) + \{(X_1 - 2) * [\ln(\gamma_1) - \ln(\beta_1)]\} * D_1 + X_2 * \ln(\beta_2) + \ln(e)$$

Here, D_1 is the dummy variable regressor and γ_1 is its coefficient. Development years 0 to 2 will be considered the baseline category for which D_1 will be assigned a value of 0 and yield Equation 3 defined above. Performing a regression with this model results in the statistics shown in Table 3. Introducing a dummy variable to account for the assumed discrete change in trend seems to improve the fit of the model. The residual standard error decreased 0.2711 points to 0.3973. The F-statistic increased markedly and its p-value is effectively zero. Examining the t-statistics for the individual regressors reveals encouraging results as well. The standard errors of the development year and dummy variable coefficients are relatively small and their p-values are essentially zero. These statistics also improved for the calendar year coefficient but not enough to suggest the model fits the data along this dimension very well.

Table 3: R Output for Equation 4

```
Call:
lm(formula = ln.adj.incr.paid ~ DY1 + DD1 + CY, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.25226 -0.25207  0.03093  0.23365  0.90913

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.91620    0.16162  86.102  < 2e-16 ***
DY1          0.60586    0.08165   7.421 1.18e-09 ***
DD1         -0.39103    0.03486 -11.216 2.23e-15 ***
CY           0.05983    0.02526   2.369   0.0217 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3973 on 51 degrees of freedom
Multiple R-squared:  0.7253,       Adjusted R-squared:  0.7091
```

F-statistic: 44.89 on 3 and 51 DF, p-value: 2.434e-14

Introducing a baseline qualitative category for calendar years 0 to 4 yields the following regression equation.

Equation 5

$$\ln(Y') = \ln(\alpha) + X_1 * \ln(\beta_1) + X_2 * \ln(\beta_2) + \{(X_2 - 4) * [\ln(\gamma_1) - \ln(\beta_2)]\} * D_1 + \ln(e)$$

The results of this regression are shown in Table 4 do not differ much from the regression model without dummy variables. The residual standard error decreased 0.0064 points and the p-value of the F-statistic increased 0.000589, offering slightly less support for rejecting the omnibus null hypothesis. The baseline category appears to impact the regression somewhat more than the unqualified calendar year variable from Equation 3 as the p-value of the t-distribution statistic for the former is 0.07385 versus 0.165035 for the latter. However, a p-value of 0.99727 for the dummy variable representing calendar years subsequent to period 4 suggests one would not reject the null hypothesis of $\ln(\gamma_1) = 0$.

<u>Table 4: R Output for Equation 5</u>

```
Call:
lm(formula = ln.adj.incr.paid ~ DY + CY1 + CD1, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.68916 -0.41391  0.08645  0.49059  0.99436 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 14.2372332  0.3901035  36.496  < 2e-16 ***
DY          -0.1553988  0.0420782  -3.693  0.00054 ***
CY1          0.2199073  0.1204935   1.825  0.07385 .  
CD1         -0.0002052  0.0596932  -0.003  0.99727    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.662 on 51 degrees of freedom
Multiple R-squared:  0.2376,	Adjusted R-squared:  0.1927 
F-statistic: 5.297 on 3 and 51 DF,  p-value: 0.002955
```

Performing another regression that bifurcates the calendar years at period 7 requires a revised equation.

Equation 6

$$\ln(Y') = \ln(\alpha) + X_1 * \ln(\beta_1) + X_2 * \ln(\beta_2) + \{(X_2 - 7) * [\ln(\gamma_1) - \ln(\beta_2)]\} * D_1 + \ln(e)$$

The residual standard error and the p-value of the F-statistic in Table 5 increased 0.0124 and 0.004236 points relative to the regression statistics displayed in Table 4. The R-squared values each decreased by approximately 0.03 points, indicating this model describes even less of the variation in the response variable. Furthermore, the p-values of the t-distribution tests for calendar year and dummy variable coefficients are non-zero.

Table 5: R Output for Equation 6

```
Call:
lm(formula = ln.adj.incr.paid ~ DY + CY2 + CD2, data = df)

Residuals:
    Min       1Q   Median       3Q      Max
-1.63130 -0.44125  0.08704  0.54323  0.98264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.72161    0.28550  51.564  < 2e-16 ***
DY          -0.15540    0.04286  -3.625 0.000666 ***
CY2          0.04811    0.05917   0.813 0.419915
CD2          0.09767    0.13857   0.705 0.484115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6744 on 51 degrees of freedom
Multiple R-squared:  0.2088,      Adjusted R-squared:  0.1622
F-statistic: 4.486 on 3 and 51 DF,  p-value: 0.007191
```

As before, reviewing plots of the mean residuals will offer additional insights. Figure 3 shows the mean residuals by development year, while Figures 4 and 5 plot the mean residuals by calendar year assuming discrete changes in trend at periods 4 and 7, respectively. The line of mean residuals in these three plots would be horizontal if the introduction of qualitative variables adequately accounted for a discrete change in inflation across development or calendar years.

Figure 3:

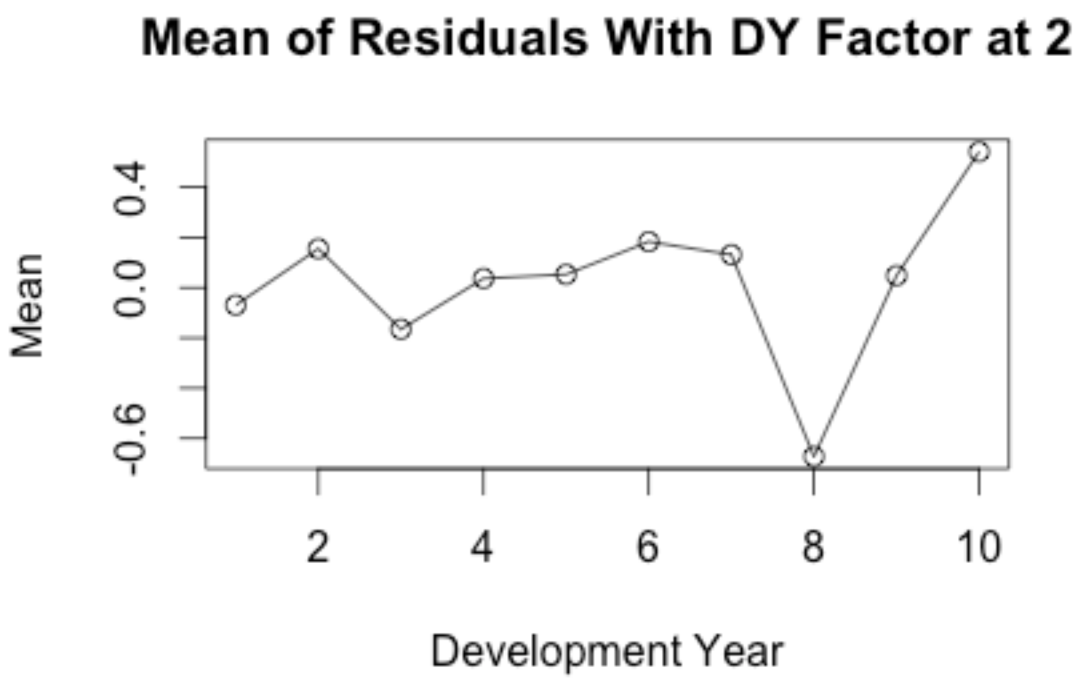


Figure 4:

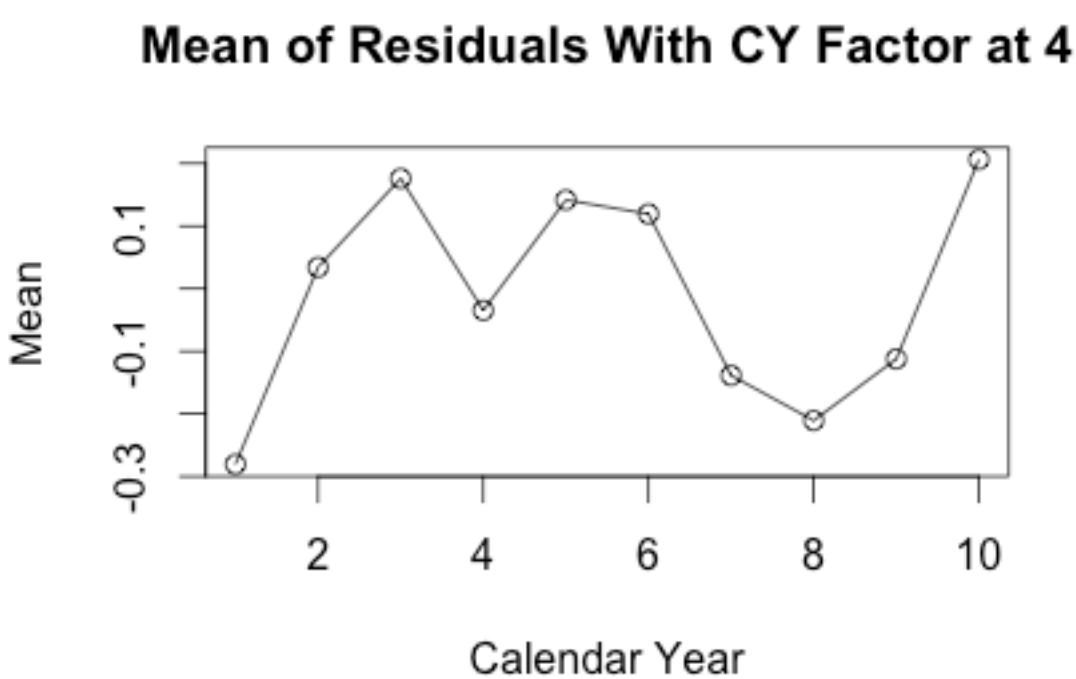
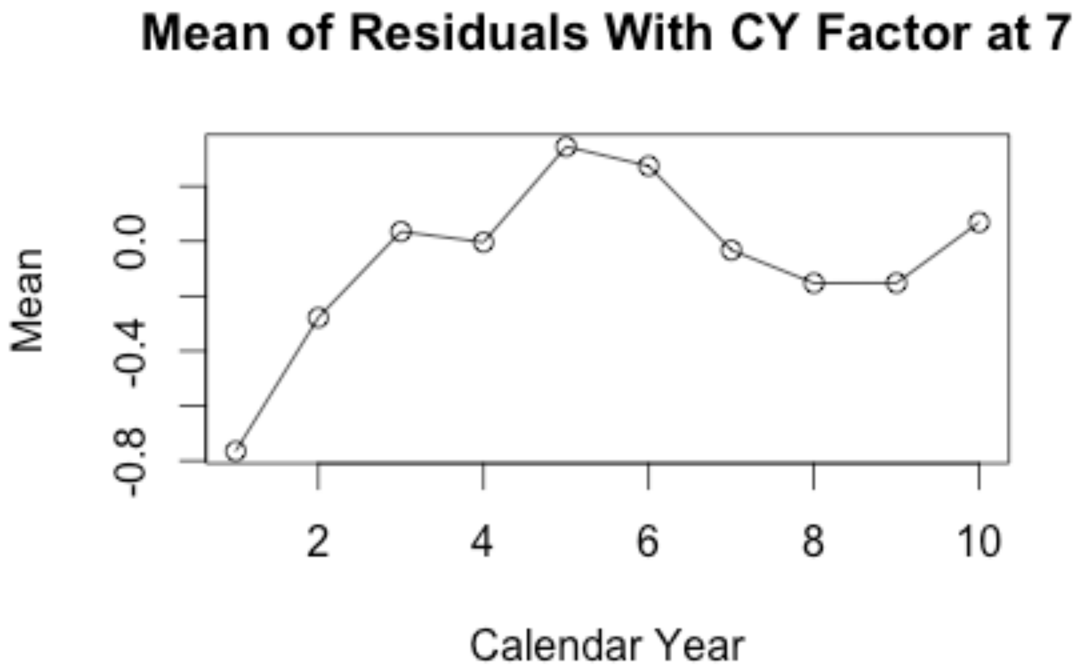


Figure 5:



Conclusion

The models proposed in this project sought to quantify the trends in a triangle of incremental paid loss data. The impact of accident year trend was removed from the data through the use of an exposure index. This adjustment eliminated a perfect linear relationship among the explanatory variables and was represented a model respecified to cope with multicollinearity. The remaining development and calendar year trends formed a multiplicative model that was made additive and linear by taking the natural logarithm of the regression equation.

The initial regression appeared to describe the development year trend well, if not the trend along the calendar year dimension of the triangle of incremental paid loss. A visual inspection of the mean residuals by year revealed that neither trend was sufficiently described by the models as they were not constant across years. Dummy variable regressors were introduced in an effort to account for discrete changes in the trends. However, these qualified models did not adequately describe the influence of the explanatory variables on the response variable. The annual change in trend is clearly not discrete and the proposed models would not be useful for estimating ultimate losses. Performing separate analyses on loss and expense using one of the regression equations described above may lead to models that effectively quantify the impact of the explanatory variables on the response. However, it is also possible that a completely different model is needed to properly describe the underlying data.