# Loss Reserves Regression Analysis

César Raúl Urteaga Reyesvera

## Table of Contents



## Introduction

One of the most important functions that an actuary performs in the property/casualty insurance industry is loss reserving whereby the actuary determines the present liability associated with future claim payments.

In this project, we will use classical regression analysis to model these future payments. The problem with this kind of analysis is that it assumes constancy of the regression parameters (i.e., the $\beta$'s) for the entire range of variables; however, this is not always an appropiate analysis to model the phenomenon under consideration.

In order to correct this, we will test the constancy of $\beta$'s (i.e., inflation rate and payment pattern) using three simulated data sets:

1. One with a discrete change in inflation rate
2. One with a discrete change in the payment pattern

3. One with both changes

And we will check the constancy of the regression coefficients through the use of residual plots. If a coefficient is not constant, we will add a dummy variable to the model. So, we will see that this solves the problem that a discrete change will cause in the classical regression approach.

The purpose of this project is to test different scenarios to determine an appropriate regression model, when the parameters change over time, that explains the variability of paid claims for insurance reserving through the use of dummy variables.

## Description of the data

The data was simulated using the method described in the project template "Regression project template loss development.xls" from the NEAS web site. We considered a period of 15 years (i.e., 120 simulated observations).

## Statistical workbook

The statistical analysis was done in Excel 2010, please refer to the workbook "20150721_RA_Project_CRUR.xlsm".

Excel file sheets:

| Sheet | Description |
|---|---|
| Control_S | Simulated values of the stable scenario |
| Control_M | Regression results of Control_S |
| Scenario1_S | Simulated values of scenario 1 |
| Scenario1_M | Regression results of Scenario1_S |
| Scenario1_SD | Simulated values of scenario 1 under the proposed model |
| Scenario1_MD | Regression results of Scenario1_SD |
| Scenario2_S | Simulated values of scenario 2 |
| Scenario2_M | Regression results of Scenario2_S |
| Scenario2_SD | Simulated values of scenario 2 under the proposed model |
| Scenario2_MD | Regression results of Scenario2_SD |
| Scenario3_S | Simulated values of scenario 3 |
| Scenario3_M | Regression results of Scenario3_S |
| Scenario3_SD | Simulated values of scenario 3 under the proposed model |
| Scenario3_MD | Regression results of Scenario3_SD |

# Multiplicative relation and simulation of the data

## Multiplicative model

Since the geometric decay payment pattern and the inflation rate have a multiplicative effect on future claim payments, future paid losses can be modelled with the following equation:

$$Y'|DY, CY = \alpha' \cdot ({\beta_1}')^{DY} \cdot ({\beta_2}')^{CY} \cdot \varepsilon',$$

where $\alpha'$ is a constant scalar, ${\beta_1}'$ is the geometric decay factor, $DY$ is the number of development years after the accident, ${\beta_2}'$ is the yearly inflation rate, $CY$ is the number of calendar years after the initial observation, and $\varepsilon'$ is the error term.

This model is not linear, but it is inherently linear. To transform a multiplicative relation into an additive one, we take logarithms:

$$\ln(Y'|DY, CY) = \ln(\alpha') + DY \cdot \ln({\beta_1}') + CY \cdot \ln({\beta_2}') + \ln(\varepsilon').$$

We can restate the above equation if we define $X_1 = DY$, $X_2 = CY$, $Y|X_1, X_2 = \ln(Y'|DY, CY)$, $\alpha = \ln(\alpha')$, $\beta_1 = \ln({\beta_1}')$, $\beta_2 = \ln({\beta_2}')$, and $\varepsilon = \ln(\varepsilon')$. So, the additive relation is:

$$Y|X_1, X_2 = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \varepsilon.$$

Under this transformation we can now perform linear regression analysis.

## Simulation of $Y|X_1, X_2$

In order to simulate the logarithm of paid losses (i.e., $Y|X_1, X_2$), we generated the error terms using the macro "SimulateUniformRandomNumbers" contained in the workbook.

Since the $\varepsilon \sim Normal(0, \sigma)$, we simulated the error terms with the following equation:

$$\varepsilon = Z \cdot \sigma,$$

where $Z \sim Normal(0,1)$.

We obtain a uniform random number less than 1 and greater or equal to 0 with the `Rnd` Excel's visual basic function. Then, we simulate $Z$ taking the inverse normal distribution of this random number through the `NORMSINV` Excel's function. Finally, we multiply this value by the standard error of the regression (i.e., $\sigma$).

So, the simulated values were obtained using the following equation:

$$Y|X_1, X_2 = E[Y|X_1, X_2] + \varepsilon,$$

where $E[Y|X_1, X_2] = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2$.

# Scenarios

## Stable Scenario

First, I considered a low stochasticity scenario in order to check my work.

### Model parameters

The model parameters were:

| Model Parameters | | | |
|---|---|---|---|
| $\sigma$ | = | 0.05 | Standard error of the regression |
| $\alpha$ | = | 30 | Intercept of the regression |
| $\beta_1$ | = | -0.32 | Geometric decay of incremental paid losses by development period |
| $\beta_2$ | = | 0.06 | Inflation rate by calendar year |

| Definitions | | |
|---|---|---|
| $Y \mid X_1, X_2$ | = | Logarithm of paid losses |
| $X_1$ | = | Development period |
| $X_2$ | = | Calendar year |
| $E[Y \mid X_1, X_2]$ | = | $\alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2$ |
| $\varepsilon$ | = | $Z \cdot \sigma$ |
| U | = | Simulated uniform random number between [0, 1) |
| Z | = | Simulated standar normal random number |

The simulated values were stored in the "Control_S" worksheet.

### Results

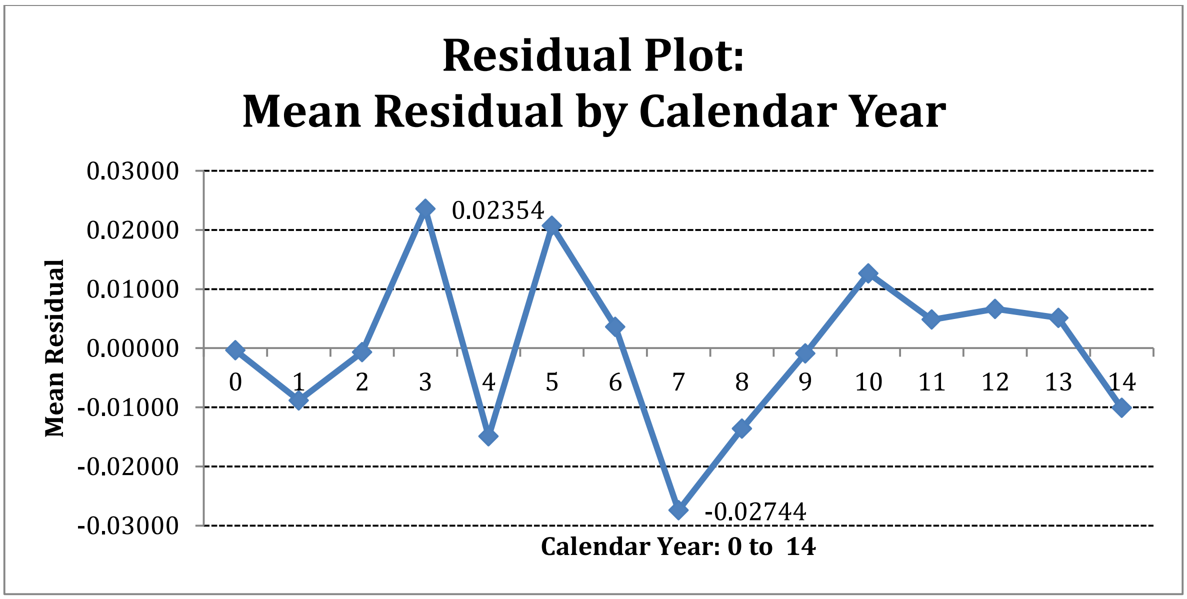
Regressing the simulated values on $X_1$ and $X_2$, using the Excel regression add-in, we obtained the following results (see the "Control_M" worksheet):

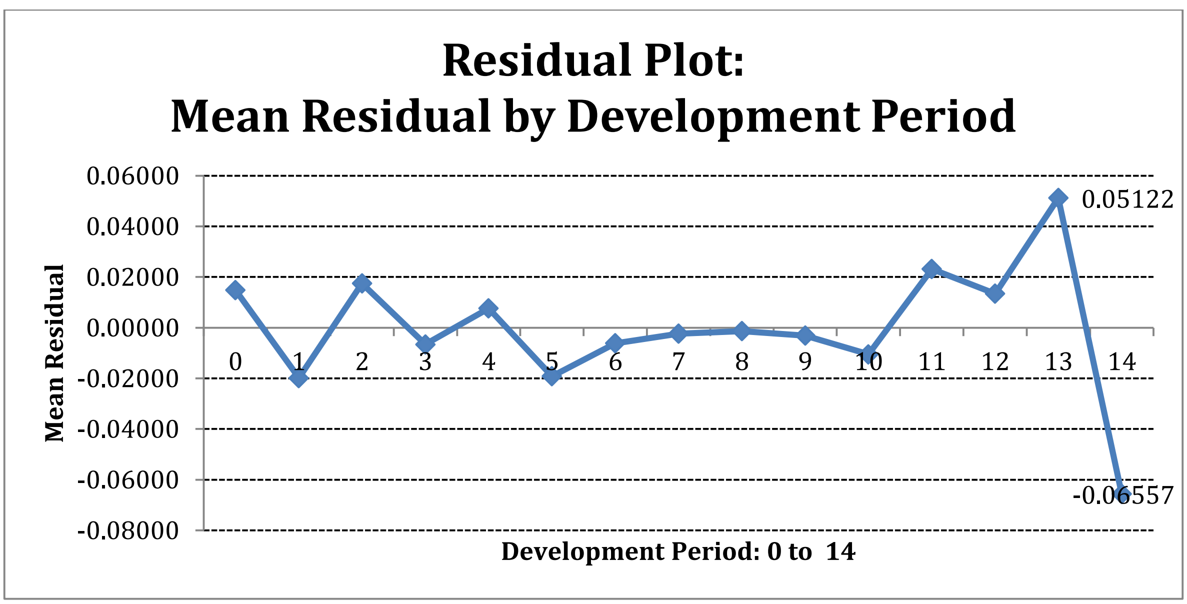
| *Regression Statistics* | |
|---|---|
| Multiple R | 0.998959 |
| R Square | 0.997919 |
| Adjusted R Square | 0.997883 |
| Standard Error | 0.049616 |
| Observations | 120 |

| | *Coefficients* | *Standard Error* |
|---|---|---|
| Intercept | 29.97888607 | 0.012476867 |
| X1 | -0.32137397 | 0.001438309 |
| X2 | 0.06278075 | 0.001438309 |

The regression estimates were close to the true parameters, while the standard errors of the estimated coefficients were very small. The proportion of the total variation of $Y|X_1, X_2$ that is "captured" by the linear regression model (i.e., $R^2$) was "high" (99.79%).

The mean residual plots by calendar year and development period are shown below. We can see that, even though there is some volatility, the average residuals are centered around zero and very small in magnitude. Thus, we can conclude that the regression analysis is working "reasonably well".



## Scenario 1: Change in the inflation rate

Because I wanted to test the scenary with a discrete change on inflation, I considered an inflation rate of 3% during the calendar years 1 through 10 and 16% in the last 5 years.

### Model parameters

The model parameters were:

Model Parameters

| | | | |
|---|---|---|---|
| $\sigma$ | = | 0.05 | Standard error of the regression |
| $\alpha$ | = | 30 | Intercept of the regression |
| $\beta_1$ | = | -0.32 | Geometric decay of incremental paid losses by development period |
| $\beta_{21}$ | = | 0.03 | Inflation rate by calendar year |
| $\beta_{22}$ | = | 0.16 | Inflation rate by calendar year |

Definitions

| | | | |
|---|---|---|---|
| $Y \mid X_1, X_2$ | = | Logarithm of paid losses | |
| $X_1$ | = | Development period | |
| $X_2$ | = | Calendar year | |
| $E[Y \mid X_1, X_2]$ | = | $\alpha + \beta_1 \cdot X_1 + \beta_{21} \cdot X_2$ | If $X_2 < 10$ |
| | | $\alpha + \beta_1 \cdot X_1 + \beta_{21} \cdot 9 + \beta_{22} \cdot (X_2 - 9)$ | If $X_2 \geq 10$ |
| $\varepsilon$ | = | $Z \cdot \sigma$ | |
| $U$ | = | Simulated uniform random number between [0, 1) | |
| $Z$ | = | Simulated standar normal random number | |

The simulated values were stored in the "Scenario1_S" worksheet.

### Results

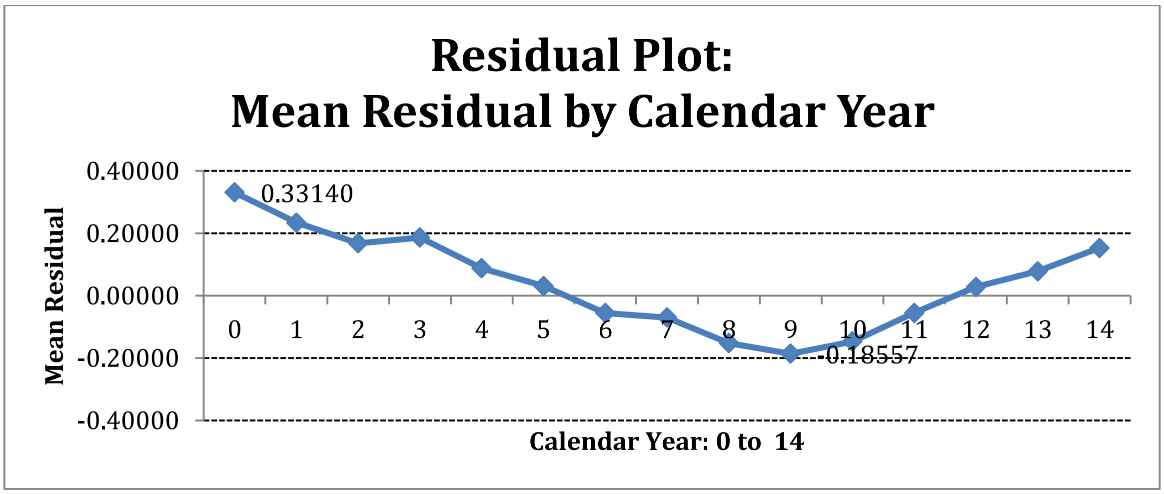
Regressing the simulated values on $X_1$ and $X_2$, using the Excel regression add-in, we obtained the following results (see the "Scenario1_M" worksheet):

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.992272824 |
| R Square | 0.984605357 |
| Adjusted R Square | 0.984342201 |
| Standard Error | 0.13161344 |
| Observations | 120 |

| | Coefficients | Standard Error |
|---|---|---|
| Intercept | 29.67905496 | 0.033096343 |
| X1 | -0.319616248 | 0.003815283 |
| X2 | 0.088543359 | 0.003815283 |

We can see from the above results that the standard errors of the estimated coefficients were greater than before. The $R^2$ is lower by 1.3% percentage points at 98.4%, and the overall standard error is much higher at 0.131613 (i.e., 1.7 times higher).

The mean residual plot by calendar year is shown below, which is the corresponding variable for inflation. The average residuals form has a V shape, which tells us that we are underestimating Y in calendar years 1 to 6 and 13 to 15, and overestimating Y in the remaining years. Therefore, we need to introduce a dummy variable to correct the inflation change in calendar year 11.




## Proposed model

If we want to take into account the change in the inflation rate, we need to introduce a dummy variable to correct for the intercept and the slope, which are both different during the last 5 years.

The proposed model was:

$$Y|X_1, X_2, D_2 = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \gamma_2 \cdot D_2 + \delta_{22} \cdot (D_2 \cdot X_2) + \varepsilon,$$

where $D_2$ is equal to 1 if $X_2 > 9$ (i.e., if the calendar year is greater than 10), 0 otherwise.

The added dummy variables ($D_2$ and $D_2 \cdot X_2$) were stored in the "Scenario1_SD" worksheet.

## Results

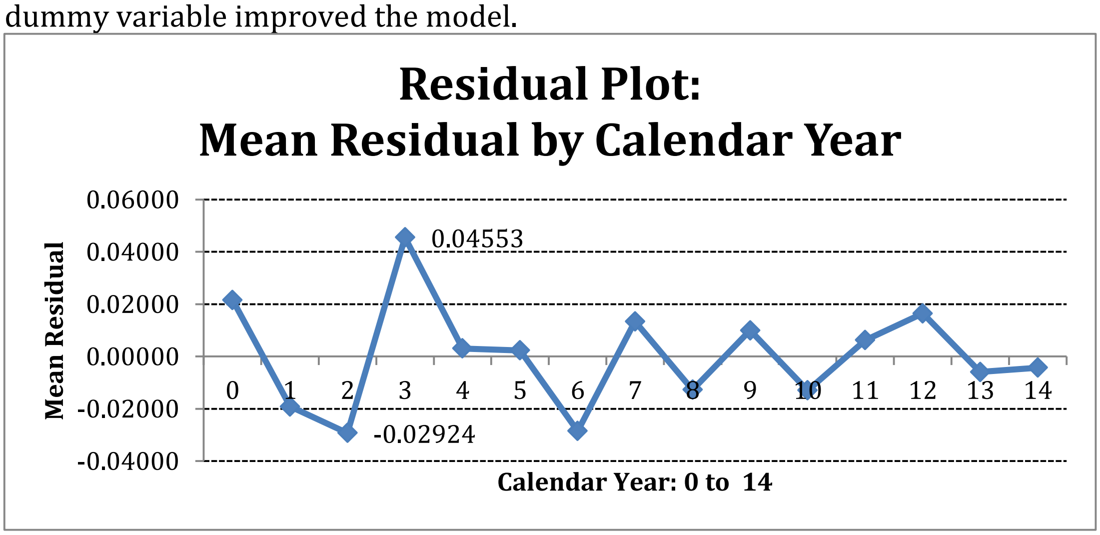
Regressing the simulated values on $X_1$, $X_2$, $D_2$ and $D_2 \cdot X_2$, using the Excel regression add-in, we obtained the following results (see the "Scenario1_MD" worksheet):

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.999125 |
| R Square | 0.998252 |
| Adjusted R Square | 0.998191 |
| Standard Error | 0.044739 |
| Observations | 120 |

| | *Coefficients* | *Standard Error* |
|---|---|---|
| Intercept | 29.98886211 | 0.015960867 |
| X1 | -0.319616248 | 0.001296926 |
| X2 | 0.032391131 | 0.002546755 |
| D2 | -1.172032674 | 0.050864077 |
| D2 · X2 | 0.128981977 | 0.004652611 |

We can see that the estimated coefficients were much better. The intercept was 29.98886 in the first 10 calendar years and (29.98886 - 1.17203) for the remaining 5 years. While the slope coefficient of the inflation rate is 0.032391 in the first 10 calendar years ($D_2$ equal to 0) and (0.032391 + 0.128982) in calendar years 11 to 15. The $R^2$ improved at 99.82%, and the overall standard error is much lower at 0.044739.

The average residuals are plotted in the chart below. The V shape has been removed. The mean residuals are centered around zero and are of low magnitude. Now, the overall shape of the mean residuals is roughly horizontal. Hence, the inclusion of the dummy variable improved the model.



## Scenario 2: Change in the payment pattern

For this scenario, I considered an decay factor of 32% in calendar years 1 to 10 and 15% over the last 5 years.

### Model parameters

The model parameters were:

| Model Parameters | | | |
|---|---|---|---|
| $\sigma$ | = | 0.05 | Standard error of the regression |
| $\alpha$ | = | 30 | Intercept of the regression |
| $\beta_{11}$ | = | -0.32 | Geometric decay of incremental paid losses by development period |
| $\beta_{12}$ | = | -0.15 | Geometric decay of incremental paid losses by development period |
| $\beta_2$ | = | 0.03 | Inflation rate by calendar year |

| Definitions | | | |
|---|---|---|---|
| $Y \mid X_1, X_2$ | = | Logarithm of paid losses | |
| $X_1$ | = | Development period | |
| $X_2$ | = | Calendar year | |
| $E[Y \mid X_1, X_2]$ | = | $\alpha + \beta_{11} \cdot X_1 + \beta_2 \cdot X_2$ | If $X_1 < 10$ |
| | | $\alpha + \beta_{11} \cdot 9 + \beta_{12} \cdot (X_1 - 9) + \beta_2 \cdot X_2$ | If $X_1 \geq 10$ |
| $\varepsilon$ | = | $Z \cdot \sigma$ | |
| U | = | Simulated uniform random number between [0, 1) | |
| Z | = | Simulated standar normal random number | |

The simulated values were stored in the "Scenario2_S" worksheet.

### Results

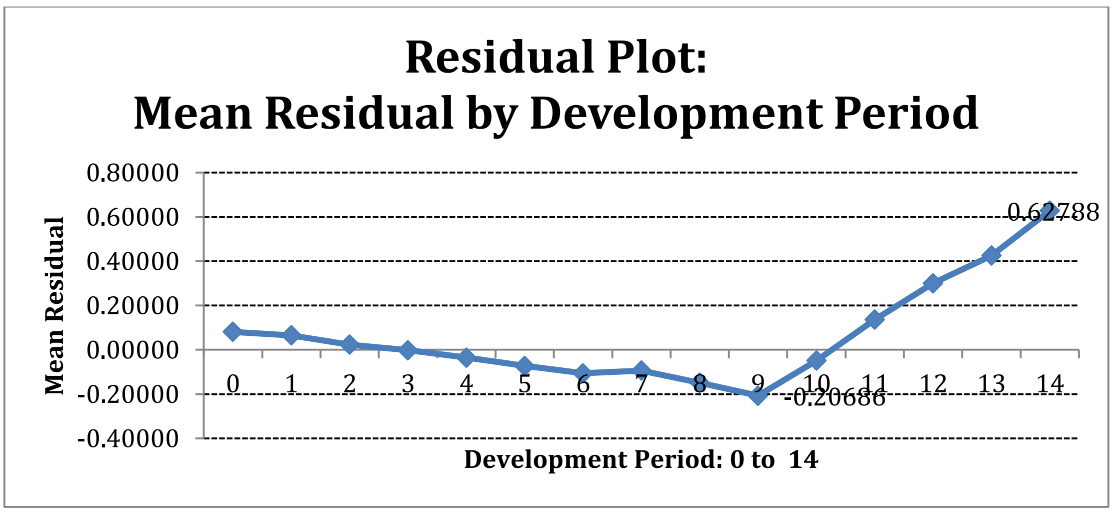
Regressing the simulated values on $X_1$ and $X_2$, using the Excel regression add-in, we obtained the following results (see the "Scenario2_M" worksheet):

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.990959164 |
| R Square | 0.982000065 |
| Adjusted R Square | 0.981692374 |
| Standard Error | 0.137679863 |
| Observations | 120 |

|  | *Coefficients* | *Standard Error* |
|---|---|---|
| Intercept | 29.92238537 | 0.034621844 |
| X1 | -0.289523123 | 0.00399114 |
| X2 | 0.029048117 | 0.00399114 |

From the results showed above, we can observe that the standard errors of the estimated coefficients were greater than the control scenario. The $R^2$ is lower by 1.6% percentage points at 98.2%, and the overall standard error is much higher at 0.137679.

The mean residual plot by development year is shown below, which is the corresponding variable for payment pattern. The average residuals form V shape, which tells us that we are underestimating Y in development years 1 to 3 and 12 to 15, and overestimating Y in the remaining years. As a result, we need to introduce a dummy variable to correct the change payment pattern in development year 11.




## Proposed model

If we want to take into account the change in the inflation rate, we need to introduce a dummy variable to correct for the intercept and the slope, which are both different in the last 5 years.

The proposed model was:

$$Y|X_1, X_2, D_1 = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \gamma_1 \cdot D_1 + \delta_{11} \cdot (D_1 \cdot X_1) + \varepsilon,$$

where $D_1$ is equal to 1 if $X_1 > 9$ (i.e., if the development year is greater than 10), 0 otherwise.

The added dummy variables ($D_1$ and $D_1 \cdot X_1$) were stored in the "Scenario2_SD" worksheet.

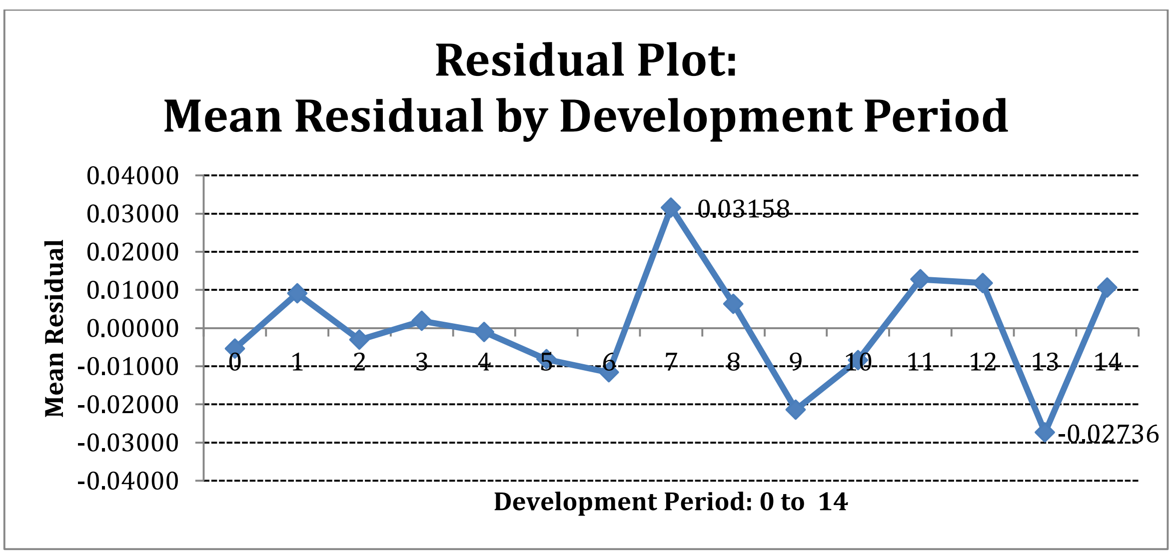
## Results

Regressing the simulated values on $X_1$, $X_2$, $D_1$ and $D_1 \cdot X_1$, using the Excel regression add-in, we obtained the following results (see the "Scenario2_MD" worksheet):

| *Regression Statistics* | | |
|---|---|---|
| Multiple R | 0.998758763 | |
| R Square | 0.997519067 | |
| Adjusted R Square | 0.997432773 | |
| Standard Error | 0.051556867 | |
| Observations | 120 | |
| | *Coefficients* | *Standard Error* |
| Intercept | 30.00926418 | 0.013435925 |
| X1 | -0.319783423 | 0.001968538 |
| X2 | 0.029048117 | 0.001494559 |
| D1 | -1.772531146 | 0.121985842 |
| D1 · X1 | 0.194754574 | 0.010827545 |

According to the calculated results, we can see that the estimated coefficients were much better. The intercept was 30.009264 in the first 10 development years and (30.009264 - 1.77253) for the remaining 5 years, while the slope coefficient of the decay factor was -0.319783 during the first 10 development years ($D_1$ equal to 0) and (-0.319783 + 0.194754) in development years 11 to 15. The $R^2$ improved at 99.75%, and the overall standard error was much lower at 0.051556.

The average residuals are plotted in the chart below. The V shape has been removed. The mean residuals are centered around zero and are of low magnitude. Now, the overall shape of the mean residuals is roughly horizontal. Consequently, the inclusion of the dummy variable improved the model.



## Scenario 3: Change in the inflation rate and in the payment pattern

The model parameters were:

Model Parameters

| | | | |
|---|---|---|---|
| $\sigma$ | = | 0.01 | Standard error of the regression |
| $\alpha$ | = | 10 | Intercept of the regression |
| $\beta_{11}$ | = | -0.20 | Geometric decay of incremental paid losses by development period |
| $\beta_{12}$ | = | -0.07 | Geometric decay of incremental paid losses by development period |
| $\beta_{21}$ | = | 0.35 | Inflation rate by calendar year |
| $\beta_{22}$ | = | 0.05 | Inflation rate by calendar year |

Definitions

| | | | |
|---|---|---|---|
| $Y \mid X_1, X_2$ | = | Logarithm of paid losses | |
| $X_1$ | = | Development period | |
| $X_2$ | = | Calendar year | |
| $E[Y \mid X_1, X_2]$ | = | $\alpha + \beta_{11} \cdot X_1 + \beta_{21} \cdot X_2$ | If $X_1 < 10$ and $X_2 < 10$ |
| | | $\alpha + \beta_{11} \cdot 9 + \beta_{12} \cdot (X_1 - 9) + \beta_{21} \cdot X_2$ | If $X_1 \geq 10$ and $X_2 < 10$ |
| | | $\alpha + \beta_{11} \cdot X_1 + \beta_{21} \cdot 9 + \beta_{22} \cdot (X_2 - 9)$ | If $X_1 < 10$ and $X_2 \geq 10$ |
| | | $\alpha + \beta_{11} \cdot 9 + \beta_{12} \cdot (X_1 - 9) + \beta_{21} \cdot 9 + \beta_{22} \cdot (X_2 - 9)$ | If $X_1 \geq 10$ and $X_2 \geq 10$ |
| $\varepsilon$ | = | $Z \cdot \sigma$ | |
| $U$ | = | Simulated uniform random number between [0, 1) | |
| $Z$ | = | Simulated standar normal random number | |

The simulated values were stored in the "Scenario3_S" worksheet.

### Results

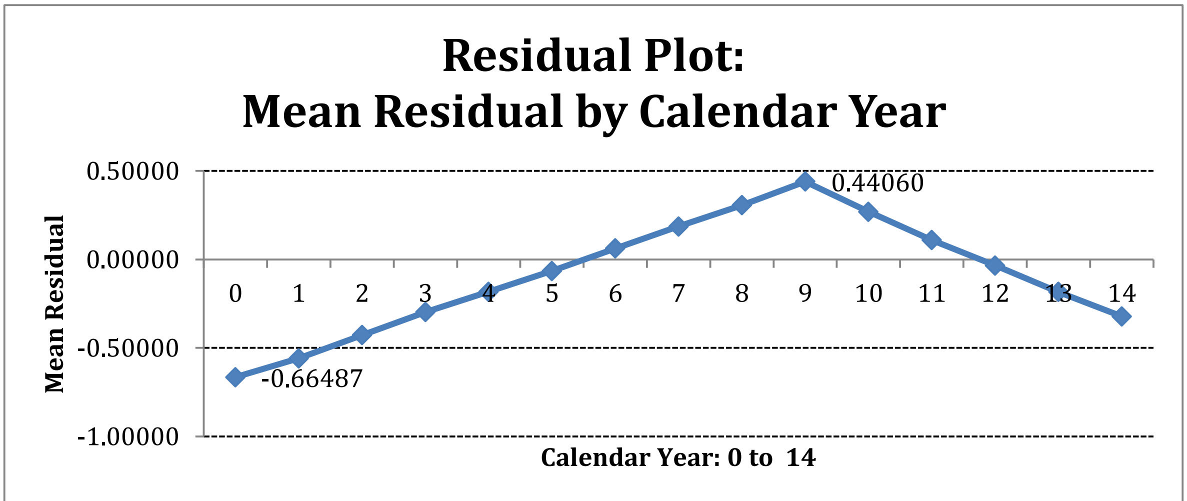
Regressing the simulated values on $X_1$ and $X_2$, using the Excel regression add-in, we obtained the following results (see the "Scenario3_M" worksheet):

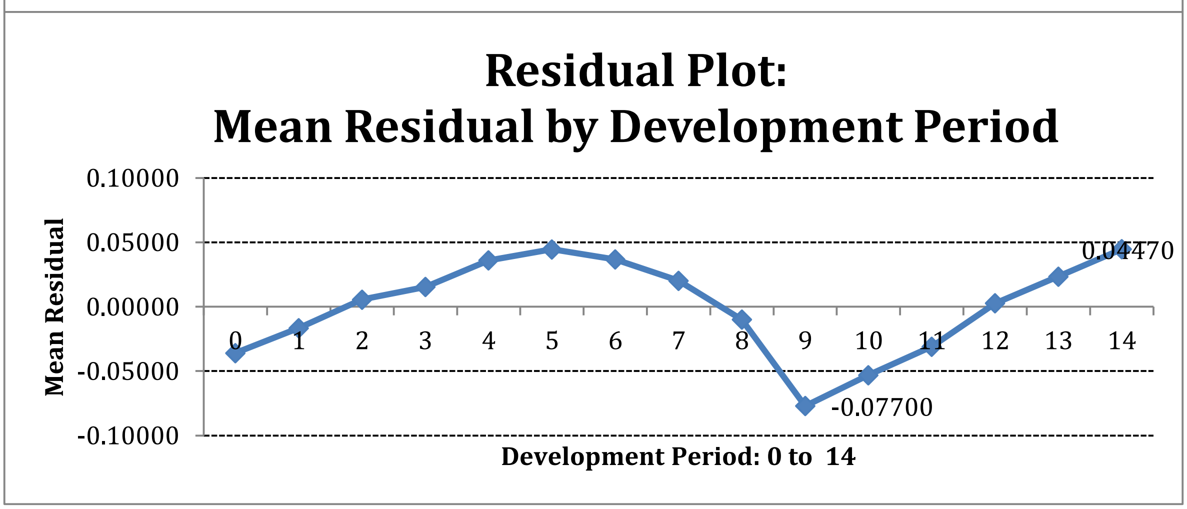
| *Regression Statistics* | |
|---|---|
| Multiple R | 0.935848463 |
| R Square | 0.875812346 |
| Adjusted R Square | 0.87368948 |
| Standard Error | 0.277777178 |
| Observations | 120 |

| | *Coefficients* | *Standard Error* |
|---|---|---|
| Intercept | 10.6744995 | 0.069851596 |
| X1 | -0.17890533 | 0.008052358 |
| X2 | 0.216419856 | 0.008052358 |

Considering the obtained results, we can see that the standard errors of the estimated coefficients were the greatest of previous scenarios. Comparing with the control scenario, the $R^2$ was lower by 12.21% percentage points at 87.58%, and the overall standard error was much higher at 0.277777 (almost 5 times greater than the obtained in the control scenario).

The mean residual plots by calendar year and development year are shown below, which show the change at year 10. So, we need to introduce a dummy variable to correct this.



## Proposed model

If we want to take into account both changes, we need to introduce two dummy variables.

The proposed model was:

$$Y|X_1, X_2, D_1, D_2 = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \gamma_1 \cdot D_1 + \gamma_2 \cdot D_2 + \delta_{11} \cdot (D_1 \cdot X_1) + \delta_{12} \cdot (D_1 \cdot X_2) + \delta_{21} \cdot (D_2 \cdot X_1) + \delta_{22} \cdot (D_2 \cdot X_2) + \varepsilon,$$

where $D_1$ (development) and $D_2$ (calendar) are the dummy variables described above.

The added dummy variables were stored in the "Scenario3_SD" worksheet.

### Results

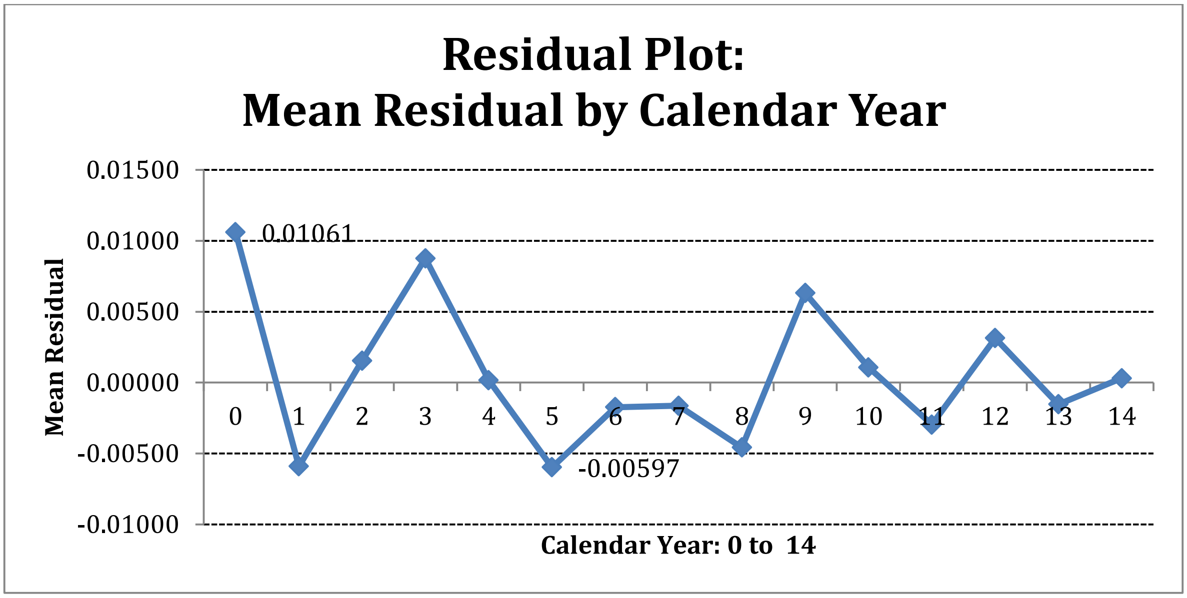
Regressing the simulated values, using the Excel regression add-in, we obtained the following results (see the "Scenario3_MD" worksheet):

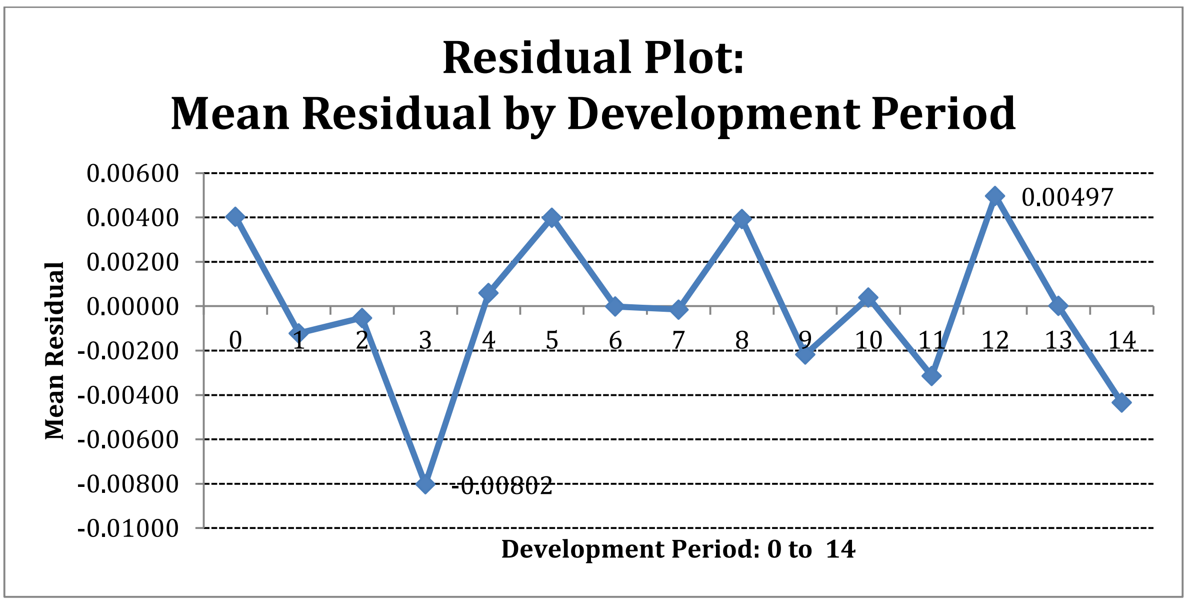
| *Regression Statistics* | |
|---|---|
| Multiple R | 0.999915453 |
| R Square | 0.999830913 |
| Adjusted R Square | 0.999818727 |
| Standard Error | 0.010523098 |
| Observations | 120 |

| | *Coefficients* | *Standard Error* |
|---|---|---|
| Intercept | 9.999018104 | 0.003754147 |
| X1 | -0.199632885 | 0.000668892 |
| X2 | 0.350090949 | 0.000668892 |
| D1 | -1.163906445 | 0.033077563 |
| D1 · X1 | 0.129863306 | 0.002568306 |
| D2 | 2.699428515 | 0.013461192 |
| D2 · X2 | -0.300001148 | 0.001246905 |
| D1 · X2 | -0.000796774 | 0.002726739 |
| D2 · X1 | 6.29093E-05 | 0.000846089 |

We can see that the estimated coefficients were much better. The $R^2$ improved at 99.98%, and the overall standard error was much lower at 0.010523.

The average residuals are plotted in the chart below. The mean residuals are centered around zero and are of low magnitude. Now, the overall shape of the mean residuals is roughly horizontal. Hence, the inclusion of the dummy variables improved the model.



## Conclusions

In sum, in order to model a more realistic situation, where both inflation and payment pattern changes are occurring at the same time, the inclusion of dummy varibles allow us to capture best the variability of the logarithm of paid losses under the proposed model using regression analysis.