Regression Analysis Student Project
Fall 2012

I. Introduction and Objective

This student project applies regression analysis on simulated paid losses with non-constant coefficients. For simplicity, this study makes the assumption that claim amounts are affected by two variables, namely: an annual inflation rate, and a geometric decay factor. The analysis is developed by applying a discrete change in the annual inflation rate and examines the impact of introducing a dummy variable to differentiate between two periods.

By nature, both the decay factor and inflation rate have a multiplicative effect on the size of future claim payments, exhibiting the following mathematical relationship:

$$Y^{'} = \alpha^{'} \beta_1^{'X_1} \beta_2^{'X_2} \varepsilon^{'}$$

In the above expression, the following definitions are made:

Y' – incremental paid loss
α' – constant scalar
X_1 – development year
X_2 – calendar year
β_1' – geometric decay factor
β_2' – annual inflation rate
ε' – error term

In order to perform a linear regression analysis in the above relationship, transformation is performed by taking the logarithms of both sides, as follows:

$$\ln(Y^{'}) = \ln(\alpha^{'}) + \ln(\beta_1^{'})X_1 + \ln(\beta_2^{'})X_2 + \ln(\varepsilon^{'})$$

Introducing another set of variables, let Y = ln(Y'), α = ln(α'), β_1 = ln(β_1'), β_2 = ln(β_2'), ε = ln(ε'). This brings forth the following unrestricted model on which linear regression analysis can be performed:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

II. Calculations and Analyses

To begin the analysis, the following preliminary values are assigned to the regression parameters:

1. The intercept, α, is equal to 10.
2. The geometric decay factor, β_1, is equal to -0.10.
3. The annual inflation rate, β_2, is equal to 0.25.
4. The stochastic term, sigma, is equal to 0.01.

5. The number of development and calendar periods is projected up to a period of 15 years. That is, the development periods, X_1 and X_2, take on values ranging from 0 to 14 years.

For purposes of this study, in order to effectively isolate the effect of adding a dummy variable to the regression, the stochastic term, sigma, is set arbitrarily low (i.e., 0.01) for the control set-up. This variable, as will be seen later on, will be set to a moderate value towards the end of the study in order to see the effect of adding a dummy variable to a regression with a more real-world performance.

The expected value of Y, E(Y), is equal to $\alpha + \beta_1 X_1 + \beta_2 X_2$. The remaining error term, ε, is simulated by invoking Excel's RAND() function, which generates a random number from 0 to 1 for each observation. The NORMSINV() function, the CDF of the standard normal distribution, is then applied in order to convert the random number to a random draw from a standard normal distribution. Finally, this random draw is multiplied by the stochastic term, sigma, in order to get the error term. The simulated values of Y are calculated as E(Y) plus the corresponding error term. The detailed calculations by calendar period are located in the "Data_original" tab of the accompanying Excel workbook.

To perform the regression analysis, Excel's *Regression* add-in feature was utilized. The summary of the regression results based on the preliminary conditions are seen below. The detailed calculations of the preliminary regression analysis can be found in the "Reg_orig" and the "Residual_orig" tabs of the accompanying Excel workbook.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.999936236							
R Square	**0.999872475**							
Adjusted R Square	**0.999870295**							
Standard Error	**0.009054663**							
Observations	120							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	75.21091309	37.60545654	458676.2451	1.5041E-228			
Residual	117	0.009592471	8.19869E-05					
Total	119	75.22050556						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	**10.00259933**	0.002276943	4392.995848	4.4939E-307	9.998089962	10.00710869	9.998089962	10.00710869
X1	**-0.099889867**	0.000262482	-380.5595353	8.4334E-183	-0.100409698	-0.099370036	-0.100409698	-0.099370036
X2	**0.249740696**	0.000262482	951.4599025	2.4035E-229	0.249220865	0.250260527	0.249220865	0.250260527

Table 1: Summary output of the preliminary regression results.

It can be seen from Table 1 above that based on regression analysis, the best-fit values of the coefficients very closely approximate the assumptions. Moreover, the values of the R-square and the Adjusted R-square statistics are both around 99.99% and 99.99%, a near-perfect fit. The standard error also closely approximates the sigma assumption. All of these suggest that the above regression model is a good fit for the data.

The average residual plots by calendar year and development year are seen below. The vector of averages shows no apparent trend or pattern, indicating stability of the regression coefficients.

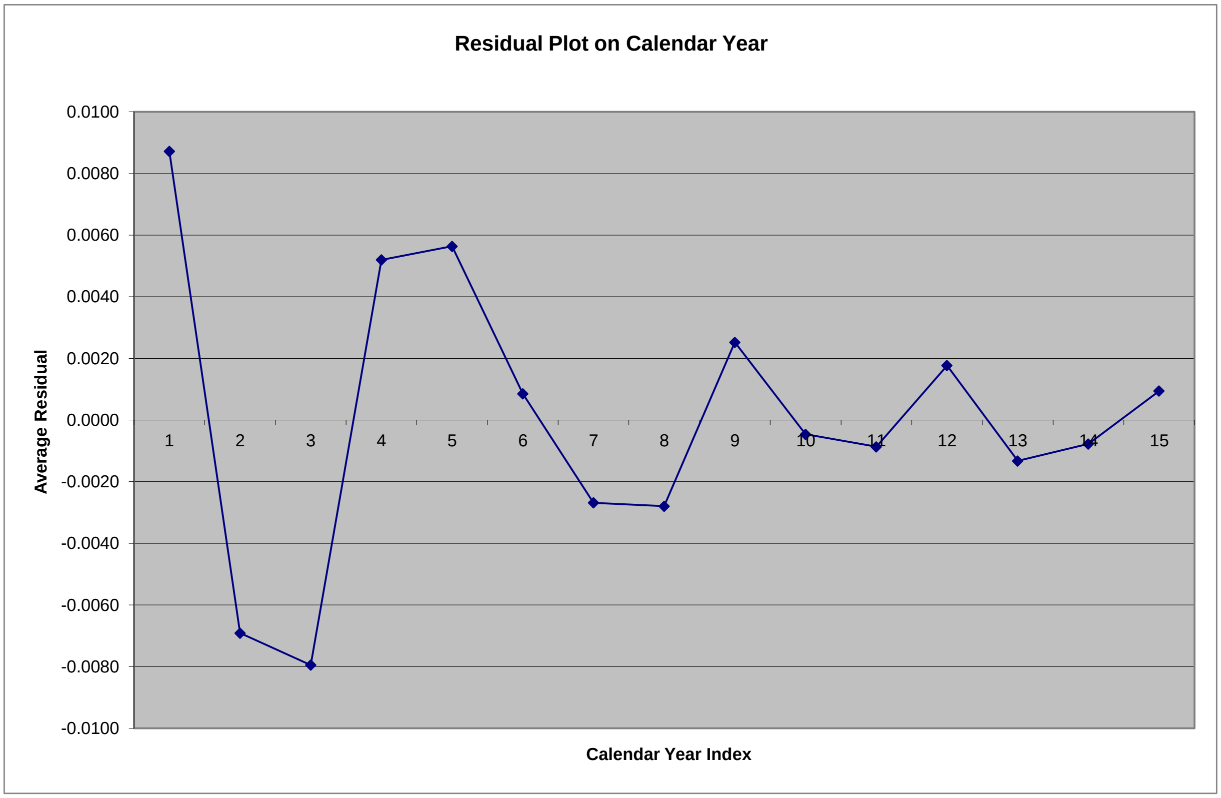

Figure 1: Average residual plot by calendar year for the equation $Y = 10.02599 – 0.09890 X_1 + 0.24741 X_2$.

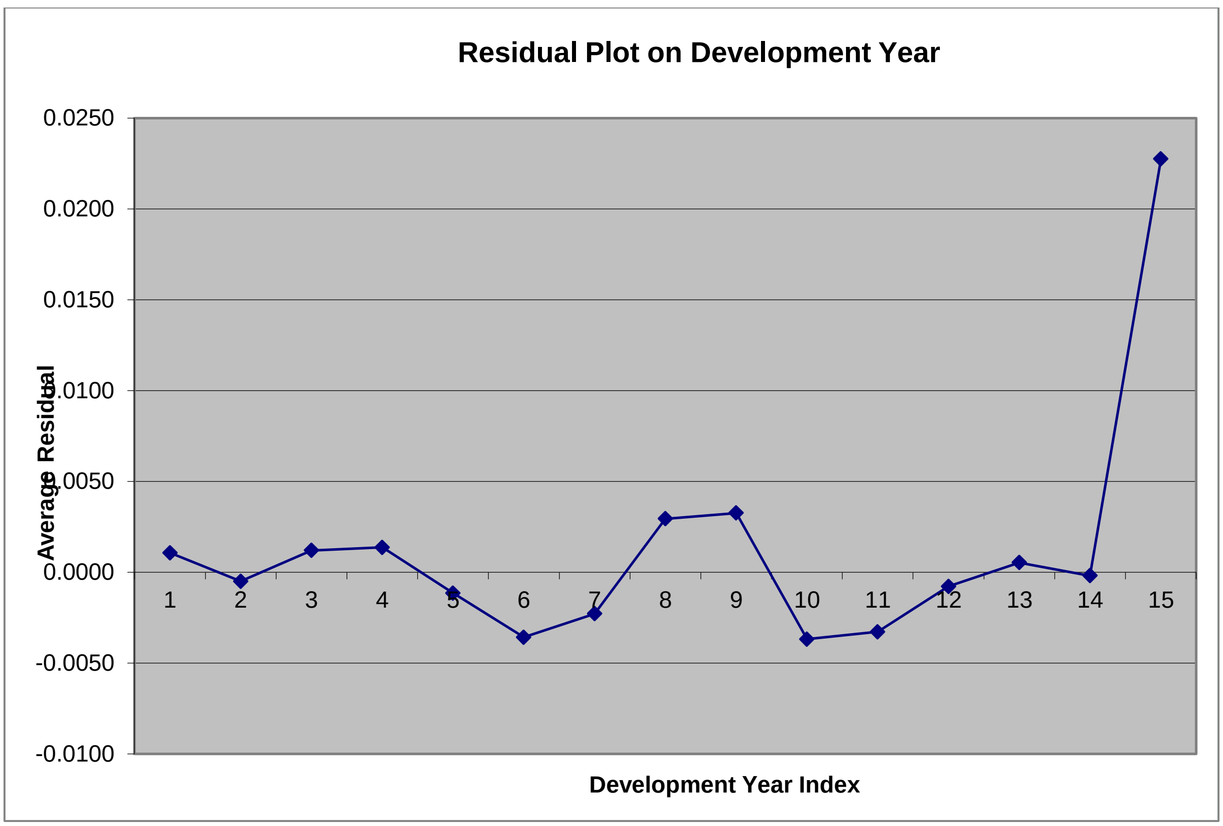

Figure 2: Average residual plot by development year for the equation $Y = 10.02599 – 0.09890 X_1 + 0.24741 X_2$.

The study is developed by performing a regression analysis on simulated paid loss amounts with a discrete change in the inflation rate. Leaving the values of all other regression parameters unchanged, the assumptions for the new model can be summarized as follows:

1. The intercept, α, is equal to 10.
2. The geometric decay factor, β_1, is equal to -0.10.
3. The annual inflation rate, β_2, is equal to **0.25 for the first 10 calendar years, and 0.04 for the last five calendar years.**
4. The stochastic term, sigma, is equal to 0.01.
5. The number of development and calendar periods is projected up to a period of 15 years. That is, the development periods, X_1 and X_2, take on values ranging from 0 to 14 years.

The model based on the above assumptions shall be called the unrestricted inflation rate model. The loss amounts are simulated in the same manner as before, and regression analysis is performed on the resulting data. The detailed calculations are located in the "Data_Chg_IR" tab of the accompanying Excel workbook. Below is the summary of the regression results for the unrestricted inflation rate model, the details of which are located in the "Reg_Chg_IR" and "Residual_Chg_IR" tabs of the Excel workbook.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.927232299							
R Square	**0.859759737**							
Adjusted R Square	**0.857362468**							
Standard Error	**0.203797837**							
Observations	120							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	29.79129542	14.89564771	358.6412604	1.2361E-50			
Residual	117	4.859426325	0.041533558					
Total	119	34.65072174						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	**10.5167905**	0.051248286	205.2125331	1.8115E-151	10.41529596	10.61828505	10.41529596	10.61828505
X1	**-0.099889867**	0.005907804	-16.90812097	3.35536E-33	-0.111589964	-0.088189771	-0.111589964	-0.088189771
X2	**0.156211285**	0.005907804	26.44151374	3.54916E-51	0.144511188	0.167911381	0.144511188	0.167911381

Table 2: Summary output of the regression analysis of the unrestricted inflation rate model.

As shown in Table 2 above, the R-square and Adjusted R-square statistics for the regression are around 85.98% and 85.74%, respectively. This is lower compared to the corresponding statistics from the initial model. Also, the standard error is around 20.38%, which is higher than the benchmark. Compared to the benchmark, the coefficients have a larger variance from the assumptions, except for the X_1, the development year coefficient, which is unaffected. This is expected because the geometric decay factor, β_1, remained unchanged. These findings are consistent with the observation on the R-square statistics and standard error. Below is the average residual plot by calendar year for the unrestricted inflation rate model:

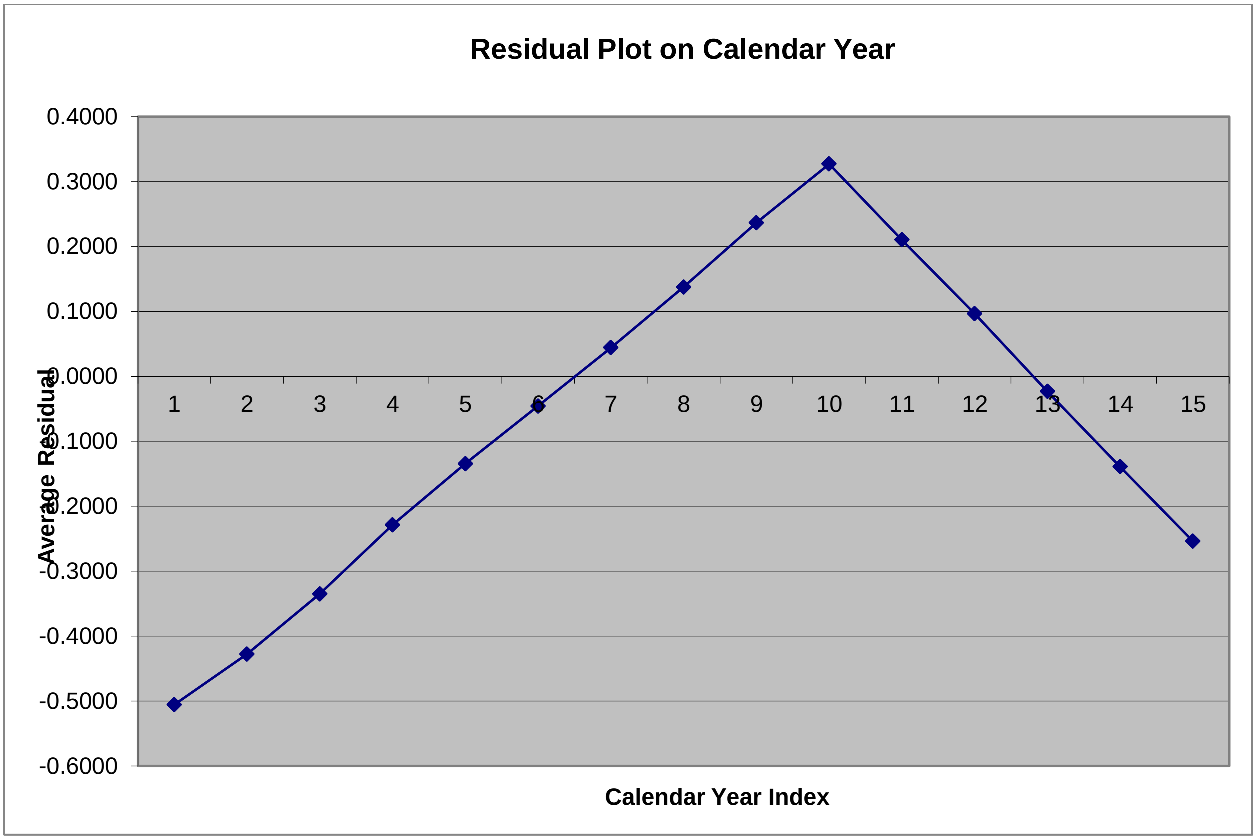

Figure 3: Average residual plot by calendar year for the equation Y = 10.51679 – 0.09890 X_1 + 0.15621 X_2.

It can be seen that the graph in Figure 3 above resembles an upside-down "V" shape that changes slope at the calendar year in which the inflation rate changes. The overall shape of the graph strongly suggests the need to introduce a dummy variable and create a restricted inflation rate model that will correct for the change in inflation rate. In order to differentiate between two periods, a dummy variable, D, is introduced, where D is **equal to 1 for the first 10 calendar years and equal to 0 for the last five calendar years.** Thus, the following restricted inflation rate model is derived:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + D\alpha_2 + D\beta_{2b} X_2 + \varepsilon$$

In the above equation, α_2 is equivalent to 9 years of the difference in the inflation rates: 9 x (β_2 – β_{2b}). The above regression introduces two additional independent variables: D and the product D x X_2. Below is the summary of the regression results for the above, the details of which are located in the "Reg_Chg_IRwDummy" and "Residual_Chg_IRwDummy" tabs of the accompanying Excel workbook.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.999988576							
R Square	**0.999977152**							
Adjusted R Square	**0.999976357**							
Standard Error	**0.009132079**							
Observations	120							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	4	419.7411174	104.9353	1258294.092	7.9575E-266			
Residual	115	0.009590411	8.34E-05					
Total	119	419.7507078						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	**10.00268245**	0.003257897	3070.288	2.1621E-284	9.996229188	10.00913572	9.996229188	10.00913572
X1	**-0.099889867**	0.000264726	-377.333	1.0408E-179	-0.100414238	-0.099365496	-0.100414238	-0.099365496
X2	**0.249733312**	0.000519838	480.4059	9.1935E-192	0.248703613	0.250763012	0.248703613	0.250763012
D	**1.888410245**	0.010382263	181.8881	2.4967E-143	1.867844979	1.908975511	1.867844979	1.908975511
D x X2	**0.040128644**	0.000949681	42.25488	6.91059E-72	0.038247509	0.042009779	0.038247509	0.042009779

Table 3: Summary output of the regression analysis of the restricted inflation rate model.

From Table 3, it can be seen that for the restricted inflation rate model, the R-square and Adjusted R-square statistics have both significantly improved to around 99.998%, a near-perfect fit. Meanwhile the standard error has been significantly reduced to around 0.91%. Similar to the unrestricted model, the development year coefficient is unaffected by the change in inflation rate. Below is the average residual plot for the restricted inflation rate model:

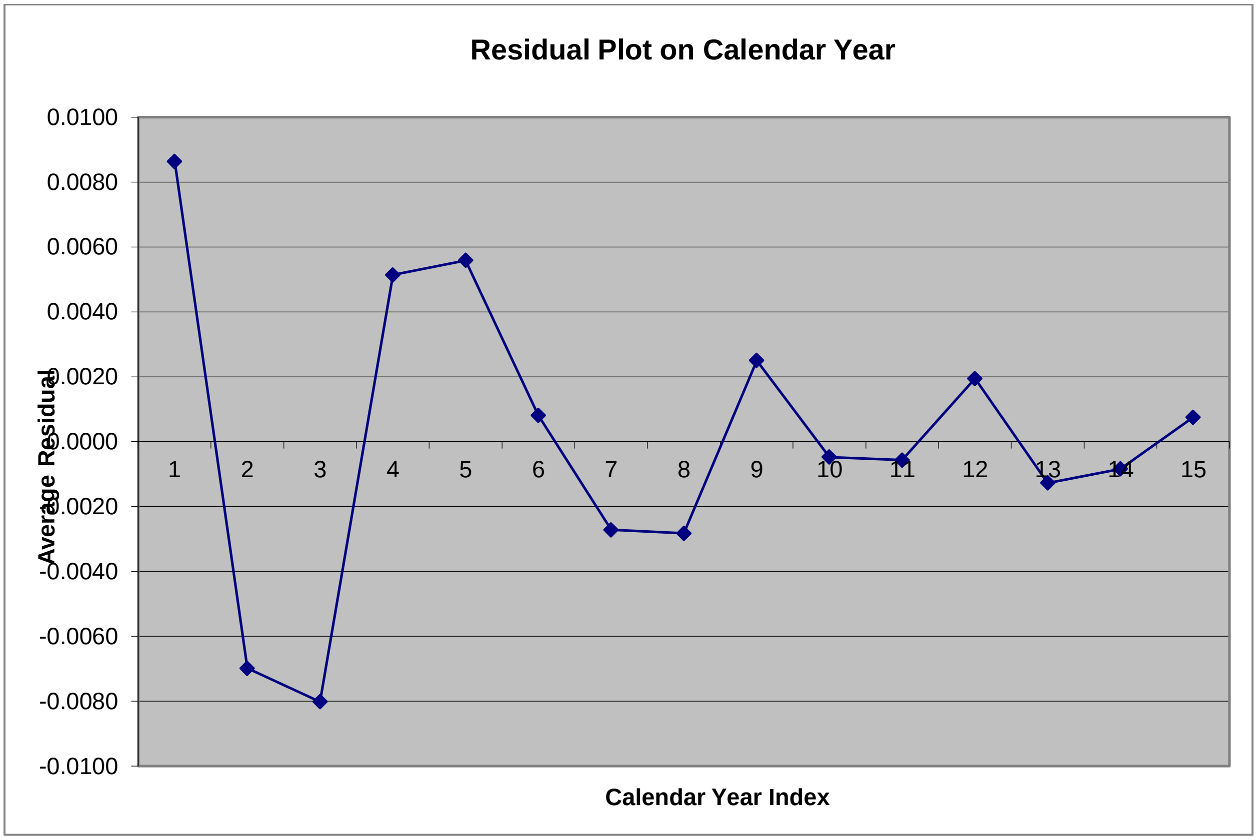

Figure 4: Average residual plot by calendar year for the restricted inflation rate model.

As seen in Figure 4, the average residual plot by calendar year is centered along the horizontal. Furthermore, there is no apparent trend in the vector of averages, indicating stability of the parameters.

As a final step in this study, the restricted inflation rate model is tested using a moderate stochastic term assumption. A **sigma value of 0.10** is selected, while all other assumptions remain unchanged. The loss amounts are simulated as usual, and regression analysis is performed on the resulting data. The detailed calculations are located in the "Data_Chg_IR_mod_stoch" tab of the accompanying Excel workbook. Below is the summary of the regression results for the restricted inflation rate model with a moderate stochastic term, the details of which are located in the "Reg_Chg_IR_mod_stoch" and "Residual_Chg_IR_mod_stoch" tabs of the Excel workbook.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.998851078							
R Square	**0.997703475**							
Adjusted R Square	**0.997623596**							
Standard Error	**0.091320795**							
Observations	120							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	4	416.6463023	104.1615756	12490.16499	1.066E-150			
Residual	115	0.95904107	0.008339488					
Total	119	417.6053434						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	**10.02682455**	0.032578971	307.7698393	1.5302E-169	9.962291879	10.09135722	9.962291879	10.09135722
X1	**-0.098898672**	0.002647258	-37.35891539	3.75925E-66	-0.10414238	-0.093654965	-0.10414238	-0.093654965
X2	**0.247333124**	0.005198381	47.57887355	1.70332E-77	0.237036131	0.257630117	0.237036131	0.257630117
D	**1.874102448**	0.103822635	18.05100064	2.39349E-35	1.66844979	2.079755106	1.66844979	2.079755106
D x X2	**0.04128644**	0.009496807	4.347402241	2.99105E-05	0.022475093	0.060097786	0.022475093	0.060097786

Table 4: Summary output of the regression analysis of the restricted inflation rate model with moderate stochastic term.

As seen in Table 4, the R-square and Adjusted R-square statistics are around 99.77% and 99.76%, respectively. The standard error of 0.0913 closely approximates the sigma assumption of 0.10. Furthermore, the values of the regression parameters closely match the values of their corresponding assumptions. All of this suggests that the above regression model is a suitable fit. Below is the average residual plot by calendar year for the said model. It can be seen that the values are centered along the horizontal and have no apparent trend. This is consistent with all other observations for this model, supporting the already strong goodness-of-fit.

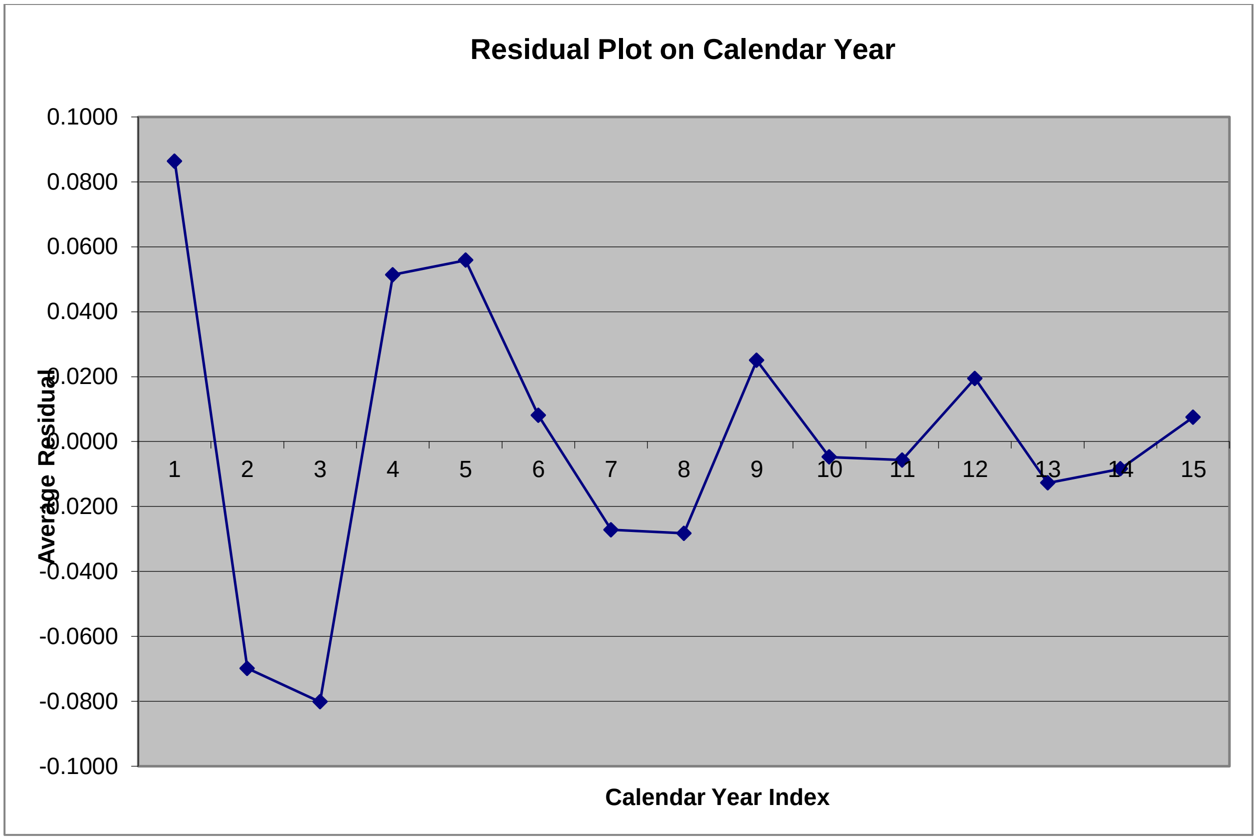

Figure 5: Average residual plot by calendar year for the restricted inflation rate model with moderate stochastic term.

III. Conclusion

Based on the above calculations, the addition of a dummy variable to correct for a discrete change in inflation rate between two periods significantly improves overall performance the linear regression.