Name: **Carlos Felippe Rostand Koetz**
VEE: **Regression Analysis – Winter 2014**

Regression Analysis – Stu. Project: PAID LOSS TRIANGLES

1 INTRODUCTION

The goal of this student project is to use Regression Analysis techniques learned in the online course to predict the incremental paid losses of the lower part of the diagonal of the Paid Loss Triangle, allowing to calculate the IBNR[1] reserve (the computation of the reserve will be done only to test the efficiency of the model).

The first model to be used is a classical one, that is, we assume in this part of the project that the coefficients are constant. As will be seen, it is assumed that value gaps between development years are better modeled by dummy variable and thus a new model is proposed with it. After adding the dummy variable, we re-evaluate the regression performance and also test whether the dummy variable is significant.

At the end, as a complement of the work, it is calculated the IBNR using Regression Analysis methods. The selected model is evaluated to see how well their values are when compared with the classical Chain Ladder Method and to see if their values would be enough to cover all future paid losses.

2 DATA

2.1 BASIC DATA

The data examined is from a workers compensation insurance, taken from the following website (http://www.casact.org/research/index.cfm?fa=loss_reserves_data). The data set used was from "**California Cas Grp**", available inside of the following file disposed in the website above: "**Workers Compensation Data Set (.csv)**". According to the information disposed in the website, the material was updated in 2011 and the data of the paid losses is from accidental year 1988 to 1997 (10 years). The last development year in the material is from 2006, so the data is all complete (is a square of paid losses). The data is summarized in the following table:

[1] The true IBNR uses the values of the paid losses and the case reserves, because those are the known claims by the insurer. Here, will be assumed by simplification, that the insurer uses only the paid losses.

Table 1: Full Cumulative Paid Loss "Triangle"

Full Cumulative Paid Loss "Triangle"		Development Year									
Accident Year		0	1	2	3	4	5	6	7	8	9
1988	0	9,558	22,778	33,298	40,348	45,146	48,048	49,782	50,623	51,812	51,939
1989	1	7,913	19,472	29,622	36,816	40,975	43,302	44,707	45,871	46,229	46,483
1990	2	8,744	24,302	35,406	43,412	48,057	50,897	52,879	53,956	54,440	54,857
1991	3	13,301	32,950	47,201	56,394	61,650	65,039	66,566	67,783	68,323	68,965
1992	4	11,424	29,086	42,034	50,910	56,406	59,437	61,029	62,354	63,037	63,406
1993	5	11,792	27,161	38,229	46,722	50,742	53,480	55,960	56,826	57,810	57,917
1994	6	11,194	26,893	38,488	45,580	48,836	50,559	52,119	53,426	54,666	55,255
1995	7	12,550	31,604	44,045	52,539	57,122	60,526	62,882	64,470	65,799	67,011
1996	8	13,194	31,474	44,070	51,693	57,120	60,453	63,499	66,205	67,423	68,225
1997	9	9,372	23,735	34,191	39,726	44,685	48,438	50,775	52,694	54,217	55,377

For the purposes of this student project, it will be assumed that is the end of the year 1997, so the data after this date will be excluded. Thus, the data used is the following:

Table 2: Cumulative Paid Loss Triangle at the end of year 1997

Cumulative Paid Loss Triangle @ end 1997		Development Year									
Accident Year		0	1	2	3	4	5	6	7	8	9
1988	0	9,558	22,778	33,298	40,348	45,146	48,048	49,782	50,623	51,812	51,939
1989	1	7,913	19,472	29,622	36,816	40,975	43,302	44,707	45,871	46,229	
1990	2	8,744	24,302	35,406	43,412	48,057	50,897	52,879	53,956		
1991	3	13,301	32,950	47,201	56,394	61,650	65,039	66,566			
1992	4	11,424	29,086	42,034	50,910	56,406	59,437				
1993	5	11,792	27,161	38,229	46,722	50,742					
1994	6	11,194	26,893	38,488	45,580						
1995	7	12,550	31,604	44,045							
1996	8	13,194	31,474								
1997	9	9,372									

Therefore, one of the goals of this student project is fill the lower part of the diagonal in the above triangle using a linear model to do it. We present below the Incremental Paid Loss Triangle, that is the first differences of the development years:

Table 3: Incremental Paid Loss Triangle at the end of year 1997

Incremental Paid Loss Triangle @ end 1997		Development Year									
Accident Year		0	1	2	3	4	5	6	7	8	9
1988	0	9,558	13,220	10,520	7,050	4,798	2,902	1,734	841	1,189	127
1989	1	7,913	11,559	10,150	7,194	4,159	2,327	1,405	1,164	358	
1990	2	8,744	15,558	11,104	8,006	4,645	2,840	1,982	1,077		
1991	3	13,301	19,649	14,251	9,193	5,256	3,389	1,527			
1992	4	11,424	17,662	12,948	8,876	5,496	3,031				
1993	5	11,792	15,369	11,068	8,493	4,020					
1994	6	11,194	15,699	11,595	7,092						
1995	7	12,550	19,054	12,441							
1996	8	13,194	18,280								
1997	9	9,372									

Thus, the Incremental Paid Loss Triangle has 55 observations. The above information will be displayed in a friendly way, which the uses of regression analysis becomes simple. The first 15 lines from this format are presented below:

Table 4: Friendly data display

i	Incremental Paid Loss (Yi)	ln(Yi)	Accident Year (Xi1)	Development Year (Xi2)	Calendar Year (Xi3)
1	9,558	9.165134	0	0	0
2	13,220	9.489486	0	1	1
3	10,520	9.261033	0	2	2
4	7,050	8.860783	0	3	3
5	4,798	8.475954	0	4	4
6	2,902	7.973155	0	5	5
7	1,734	7.458186	0	6	6
8	841	6.734592	0	7	7
9	1,189	7.080868	0	8	8
10	127	4.844187	0	9	9
11	7,913	8.976262	1	0	1
12	11,559	9.35522	1	1	2
13	10,150	9.225229	1	2	3
14	7,194	8.881003	1	3	4
15	4,159	8.33303	1	4	5

3 MODEL

3.1 MODEL INTUITION

The Incremental Paid Losses can be projected by Accidental Year (X_1) and by Development Year (X_2) in a simple way (It will be explained in section 5.1 why calendar years was not used as explanatory variable). There is a multiplicative relation between those two independent explanatory variables. It was used the following model representing that relationship:

$$Y' = \alpha' \beta'^{X_1}_1 \beta'^{X_2}_2 \varepsilon'$$

Where:

Y': is the projected incremental paid loss

α': is the incremental paid loss at base year

β'_1: is the exposure growth pattern

X_1: is the accident year

β'_2: is the payment pattern

X_2: is the development year

ε': is the error

Because the above equation is not linear, it is necessary to transform it into an additive expression. Thus, by taking logarithms on both sides, the model can be re-written as a linear relation, as follows:

$$ln(Y') = ln(\alpha') + ln(\beta'_1)X_1 + ln(\beta'_2)X_2 + ln(\varepsilon')$$

Therefore, the above equation can be redefined as:

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

Where, $Y = ln(Y')$, $\alpha = ln(\alpha')$, $\beta_1 = ln(\beta'_1)$, $\beta_2 = ln(\beta'_2)$, and $\varepsilon = ln(\varepsilon')$. For practical purposes those new variables will keep the names of the original variables. For example, Y will be kept as the projected incremental paid loss.

The model described so far will be called Restricted Model. If we pay a special attention to the data in Table 3, it is possible to see a strong additional pattern in the development year that breaks the values in well distinct intervals by which the above model cannot capture in a satisfactory way.

Thus, will be introduced dummy variables to incorporate those gaps and bring them to the model. It will be used polytomous factors (as described in section 7.2 of Fox's book). It can be shown that 5 categories (and 4 dummy regressor) will be enough to improve satisfactorily the model. This new model will be called of Full Model because it has more elements inside. The following table summarizes the structure of the polytomous factors used:

Table 5: Structure of the Polytomous Factors

Category	Development Year	Maximum Value	Minimum Value	D1	D2	D3	D4
1	0, 2, and 3	7,050	14,251	1	0	0	0
2	4	4,020	5,496	0	1	0	0
3	5 and 6	1,405	3,389	0	0	1	0
4	7, 8, and 9	127	1,189	0	0	0	1
5	1	11,559	19,649	0	0	0	0

The categories represent well the gaps in the development year of the Table 3. This will lead to an increase in the accuracy of the model.

3.2 Full Model (Model 1)

The Full Model can be expressed as the following way:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 + \gamma_4 D_4 + \varepsilon$$

To calculate the model parameters and statistics it was used the Add-In tool Data Analysis of the MS Excel. It is shown below the summary output of the Full Model (Model 1), which contains the dummy variables:

Table 6: Summary Output Full Model (Model 1)

Regression Statistics	
Multiple R	**0.959**
R Square	**0.919**
Adjusted R Square	**0.909**
Standard Error	**0.322**
Observations	**55**

ANOVA

	df	*SS*	*MS*	*F*	*Signif. F*
Regression	**6**	**56.740**	**9.457**	**91.280**	**1.55E-24**
Residual	**48**	**4.973**	**0.104**		
Total	**54**	**61.713**			

	Coef.	*St. Error*	*t Stat*	*P-value*	*Lo. 95%*	*Up. 95%*
Intercept	**9.722**	**0.149**	**65.177**	**0.000**	**9.422**	**10.022**
Accident Year (Xi1)	**0.026**	**0.020**	**1.260**	**0.214**	**-0.015**	**0.067**
Develop. Year (Xi2)	**-0.144**	**0.049**	**-2.957**	**0.005**	**-0.242**	**-0.046**
D1	**-0.381**	**0.127**	**-2.998**	**0.004**	**-0.637**	**-0.126**
D2	**-0.756**	**0.222**	**-3.411**	**0.001**	**-1.202**	**-0.310**
D3	**-1.268**	**0.260**	**-4.873**	**0.000**	**-1.791**	**-0.745**
D4	**-2.206**	**0.360**	**-6.135**	**0.000**	**-2.929**	**-1.483**

The summary shows that all coefficients are statistically significant by the exception of the Accident Year (the exposure growth pattern). All of them do not contains the 0 (zero) inside their 95% confidence interval by the exception of the $\widehat{\beta_1}$. This last one is not statistically significant, but it plays a special role in the model, so it will not be removed.

The R^2 of the Full Model is 0.919, that is, 91.9% of the variation in the Y (the natural logarithm of the incremental paid loss) is captured by the regression model. The $\tilde{R}^2 = 0.909$. This is the adjustment made in the R^2 by the number of variables in the model.

3.3 Restricted Model (Model 2)

The Restricted Model is the same introduced in the Topic 3, and does not contains any dummy variable.

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

To calculate the model parameters and statistics it was used the Add-In tool Data Analysis of the MS Excel. It is shown below the summary output of the Restricted Model (Model 2).

Table 7: Summary Output Restricted Model (Model 2)

Regression Statistics	
Multiple R	**0.922**
R Square	**0.851**
Adjusted R Square	**0.845**
Standard Error	**0.421**
Observations	**55**

ANOVA

	df	*SS*	*MS*	*F*	*Signif. F*
Regression	**2**	**52.514**	**26.257**	**148.428**	**3.22E-22**
Residual	**52**	**9.199**	**0.177**		
Total	**54**	**61.713**			

	Coef.	*St. Error*	*t Stat*	*P-value*	*Lo. 95%*	*Up. 95%*
Intercept	**9.743**	**0.150**	**64.931**	**0.000**	**9.442**	**10.044**
Accident Year (Xi1)	**0.026**	**0.027**	**0.964**	**0.340**	**-0.028**	**0.079**
Develop. Year (Xi2)	**-0.385**	**0.027**	**-14.416**	**0.000**	**-0.439**	**-0.332**

The summary shows that the intercept and the coefficient of the Development Year are statistically significant, but that one of the Accident Year (the exposure growth pattern) is not. The two significant coefficients do not contains the 0 (zero) inside their 95% confidence interval.

The R^2 of the Restricted Model is 0.851, that is, 85.1% of the variation in the Y (the natural logarithm of the incremental paid loss) is captured by the regression model. The $\tilde{R}^2 = 0.845$. This is the adjustment made in the R^2 by the number of variables in the model. Both R^2 and $\tilde{R}^2$ are less representatives in the Restricted Model than in the Full Model: this is due to inclusion of the dummy variables. To confirm that, the next section will be devoted to examine the importance of the dummy variables included in the Full Model.

4 MODEL TESTING AND CHOICE

We need to choose between two models, one with the presence of dummy variables (Full Model) and another one without them (Restricted Model), to represent the equation that gives the incremental paid loss in the triangle to calculate the IBNR reserve. Both Full and Restricted models have good adherence to the data, $R^2 = 0.919$ and $R^2 = 0.851$ respectively. Here, only one of them will be used, and to decide which one, the following test will be processed:

$$H_0: \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0$$

The equation that results in the F-Test is the following:

$$F_0 = \frac{n-k-1}{q} \times \frac{R^2_{Full} - R^2_{Restricted}}{1 - R^2_{Full}}$$

With q and $n - k - 1$ degrees of freedom. The table above summarizes the calculation:

Table 8: Testing the null hypothesis

Testing The Null Hypothesis:	
$H_0 = \gamma 1 = \gamma 2 = \gamma 3 = \gamma 4 = 0$	
n	**55**
k	**6**
q	**4**
R^2 - Full	**0.919**
R^2 - Restricted	**0.851**
F_0 =	**10.198**
p-value =	**0.0000047**

Therefore, with that p-value it can be concluded that that all dummy-regressors are statistically significant, that is, the gap-effect in the Development Year is significant to the model. This result guarantee the choice of the Full Model to predict the Incremental Paid Loss of the Triangle.

Before we finish the choice of the model, we will return the discussion of the use of the explanatory variable Accident Year (X_1) into the model. The F-Test was ran to the hypothesis that the coefficient of X_1 is not significant:

$$H_0 : \beta_1 = 0$$

The equation that results in the following F-Test statistic:

$$F_0 = \frac{n - k - 1}{q} \times \frac{R_1^2 - R_0^2}{1 - R_1^2}$$

With q and $n - k - 1$ degrees of freedom and 0 (zero) representing the model without X_1 and 1 (one) representing the Full Model. The table above summarizes the calculation:

Table 9: Testing the null hypothesis to the coefficient of X_1

Testing The Null Hypothesis:	
$H_0 : \beta 1 = 0$	
n	**55**
k	**5**
q	**1**
R^2_1	**0.919**
R^2_0	**0.917**
F_0 =	**1.620**
p-value =	**0.2091215**

By the result of the test, we cannot reject the null hypothesis that the true coefficient of X_1 is equal zero. However, using of my personal analytical skills, I will not exclude it from the model. The reason why is that in a data set like this, that there are only two very-well characterized variables (one representing the growth exposure of the line business (X_1) and the other one representing the payment/knowing pattern (X_2) of the claims), that specific variable can aggregate some information to the reserve actuarial analyst in a general way.

5 MODEL DIAGNOSTICS

5.1 COLLINEARITY

Both models Full and Restricted uses only two of the three variables because there is Collinearity between them. When the models were run with three variables some issues related with that made the use of all explanatory variables impossible. The relation is:

$$CY = AY + DY$$

Thus, I chose to no use the explanatory variable CY (Calendar Year) because it is a combination of the other two independent explanatory variables of the model.

5.2 RESIDUAL ANALYSIS

I will analyse the residuals of the Full Model and compare it with the residuals of the Restricted Model to present the differences between them, when necessary.

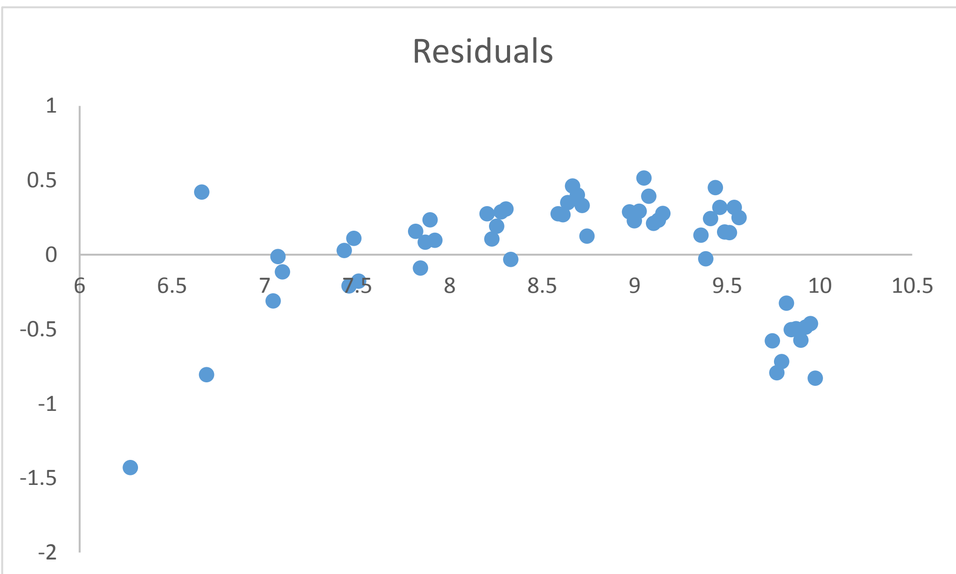
The residual standard error (standard error of the regression) is lower in the Full Model:

$$S_E^{Full} = 0.322$$

$$S_E^{Restricted} = 0.421$$

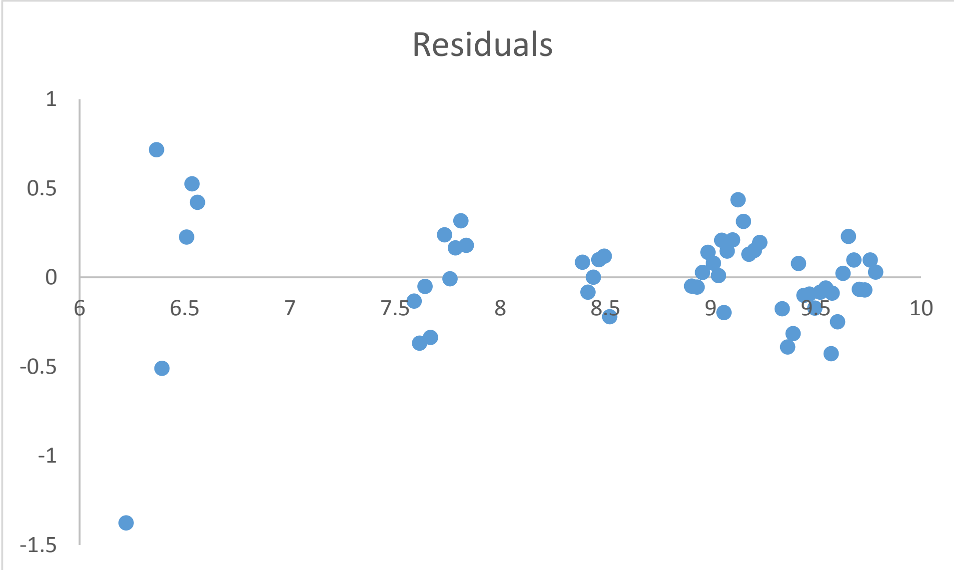
That means that the errors are also less dispersed as we can see in the following charts:

Graph 1: Residuals VS Fitted values of the Restricted Model



The residuals of the Restricted Model has an inverted V shape and is not present a random patter of their elements.

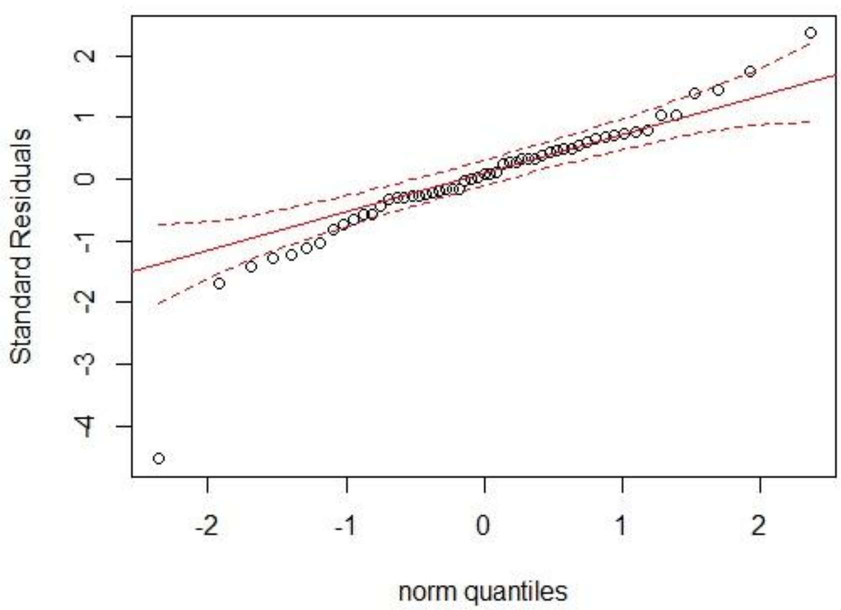
Graph 2: Residuals VS Fitted values of the Full Model



The residuals of the Full Model are more stable around zero and they have 5 points of concentration because this model uses 5 dummy-categories and the effect can be seen here. Another point is that the residuals are less dispersed, which leads to a lower residual standard error, as presented before.

To conclude this section, the Q-Q Plot of the standard residuals is presented below:

Graph 3: Q-Q Plot of the standard residuals of the Full Model



It is possible to see that the residuals has a symmetric shape but probably will not has a normal distribution (can be checked by Normality Shapiro-Wilk test), because of the lower tail, which is out of the interval of confidence of the normal quantiles.

6 FITTED VALUES

To see how useful this model is to calculate the IBNR reserve, I calculated the same reserve using the classical Chain Ladder Method and I compared both with the real value needed to pay all filled claims (using the information of the Table 1: Full Cumulative Paid Loss "Triangle"). That is, I checked if the reserve calculated at the end of 1997 by Full Model would be enough to cover all future claim payments.

Thus, the projection lead us to the following Incremental Paid Loss Triangle, where the blue numbers represent the forecasted values using the Full Model:

Table 10: Forecasted Lower Diagonal of the Incremental Paid Loss Triangle at the end of 1997

Forec. Lower Diag. Incr. Paid Loss Triangle @ end 1997			Development Year									
			0	1	2	3	4	5	6	7	8	9
Accident Year	1988	0	9,558	13,220	10,520	7,050	4,798	2,902	1,734	841	1,189	127
	1989	1	7,913	11,559	10,150	7,194	4,159	2,327	1,405	1,164	358	517
	1990	2	8,744	15,558	11,104	8,006	4,645	2,840	1,982	1,077	612	530
	1991	3	13,301	19,649	14,251	9,193	5,256	3,389	1,527	725	628	544
	1992	4	11,424	17,662	12,948	8,876	5,496	3,031	2,195	744	644	558
	1993	5	11,792	15,369	11,068	8,493	4,020	2,601	2,253	763	661	573
	1994	6	11,194	15,699	11,595	7,092	5,142	2,669	2,312	783	678	588
	1995	7	12,550	19,054	12,441	8,863	5,277	2,739	2,372	804	696	603
	1996	8	13,194	18,280	10,501	9,094	5,414	2,810	2,434	825	714	619
	1997	9	9,372	18,219	10,776	9,332	5,556	2,884	2,497	846	733	635

The table above lead us to an IBNR value of $131,964 (the sum of all blue values). The comparison is presented in the table below:

Table 11: IBNR Comparison with the Type 1 Projection

Method	Value ($)	Diff %
Real Value	**130,095**	**-**
Chain Ladder Method	**127,514**	**-1.98%**
Full Model Projection	**131,964**	**1.44%**

The projection made by the Full Model predicted a value of reserve that is enough to cover all future filled paid claims: 1.44% above of the true value and around 3.5% difference of the Chain Ladder Method value.

7 CONCLUSION

The transformation on the original data was needed because its works with money and the cumulative display presentation. To adjust that, all models tested used the natural logarithm of the Increment Paid Loss.

The two models presented were evaluated. The Restricted Model presented a good fit to the data (R^2 = 0.851), using only two of the three explanatory variables; The Full Model presented a good and a better fit to the data (R^2 = 0.919), which uses, besides the same explanatory variables, the dummy variables. The introduction of 5 dummy categories in the

Full Model allowed it to capturing a visible gap in the development payment pattern, which the regular regressor of the related variable didn't made it.

The ran tests show that dummy variables were statistically significant, so the Full Model was chosen to calculate the incremental paid claims to form the triangle.

The model diagnostics explained why only two of three explanatory variables were used: the presence of collinearity. Moreover, the analysis of the residuals expose the reasons why the Full Model was chosen compared with the Restricted Model: a better behavior of its residuals, which leads to a better fit.

Finally, the forecast of the Full Model, which generated a value of IBNR reserve, was compared with the classical Chain Ladder Method and the true value needed to cover all incremental paid losses at the end of 1997. The model developed on this student project proved its effectiveness and excellence.

8 ATTACHMENTS

The workbook attached to this student project ("**RA – Sproj – Paid Loss Triangle – CARLOS FELIPPE ROSTAND KOETZ – Winter 2014**") contains the following sheets:

- Data Base
- Paid Loss Triangle – Loss
- Chain Ladder Method
- Restricted Model
- Full Model
- Poly Factors
- Testing Models
- Testing AY
- Prediction

Some of those are calculation sheets and others are support sheets.

The workbook attached to this student project called "**standard_residuals**" contains the data used to do the QQ Plot in R, in the file "**QQ Plot Residuals**"