VEE Regression Analysis Project

Jianwei Xie
Submitted 7/10/2011

Introduction

There are many factors, which are believed to be influential to the ROE (Return on Equity) of an insurance company. These factors impact an insurer's earnings at different degrees. This analysis is prepared to examine the most influential factors to an insurer's ROE and hopefully provides some useful information for interested parties.

The factors that are tested in this analysis are listed below:

- Combined Ratio
- Net Yield on Invested Assets
- Age of Insurer
- Net Total Assets
- Direct Premiums Written Growth
- Capital & Surplus/Assets

Linear regression analysis will be used to determine if specific fators listed above can be used to predict the expected ROE for an insurance company. The proper number of explanatory variables will be determined through examining various models. A 95% confident interval is used for our linear regression analysis and the analysis is performed with the Data Analysis Package that is included in Microsoft Excel 2007. The raw data for our analysis is from SNL and 119 data records are utilized in our analysis.

Overview of Hypothesis and Methodologies

After my initial review of the data set, I set up my hypothesis based on my actuarial judgment and experiences.

- Hypothesis - the combined ratio and return on invested assets will be the factors with the most significant impacts on an insurer's ROE.

To test my hypothesis, I will start running regression analysis with all the six explanatory variables and then revising the model based on the testing results. The best model will be

selected based on the statistic significance of the explanatory variables and the correlations among the variables will be also tested.

First Model: $Y=\alpha+\beta1X1+\beta2X2+\beta3X3+\beta4X4+\beta5X5+\beta6X6$

Response Variables:

Y = ROE, Return on Equity

Explanatory Variables:

X_1 = Combined Ratio
X_2 = Net Yield on Invested Assets
X_3 = Age of Insurer
X_4 = Net Total Asset
X_5 = Direct Premium Written Growth
X_6 = Capital & Surplus / Assets

Resulted Equation from Regression in MS-Excel:

$$Y = 25.6208 - 0.2251X_1 + 1.7592X_2 - 0.0209X_3 + 0.0000X_4 + 0.0157X_5 - 0.0209X_6$$

Regression Statistics	
Multiple R	0.7618
R Square	0.5803
Adjusted R Square	0.5578
Standard Error	7.3561
Observations	119

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	25.6208	3.4968	7.3269	0.0000
X1 - Combined Ratio (%)	(0.2251)	0.0244	(9.2281)	0.0000
X2 - Net Yield on Invested Assets (%)	1.7592	0.4183	4.2059	0.0001
X3 – Age	(0.0209)	0.0147	(1.4271)	0.1563
X4- Net Total Assets ($000)	0.0000	0.0000	0.8310	0.4077
X5 - Direct Premiums Written Growth (%)	0.0157	0.0365	0.4300	0.6680
X6 - Capital & Surplus/ Assets (%)				

	(0.0209)	0.0372	(0.5619)	0.5753

By examining the above regression results, we can tell that the Combined Ratio and Net Yield on Invested Assets are significant predictors of insurance company's ROE. This finding is consistent with our hypothesis and intuitively makes senses. The Combined Ratio should be negatively related to the ROE while the Net Yield on Invested Assets should be positively related to the ROE. The results for the other four explanatory variable are also following our common senses. It is not necessary that more aged, more sizeable, faster growing, or lower leveraged companies have higher profitability.

The Adjusted R-square of 0.5578 is relatively high for testing that only involves limited amount of data. This Adjusted R-square indicates that the model is appropriate for the purpose of predicting an insurance company's ROE.

By looking at the t-statistics and p-values, we found that the X_4, X_5 and X_6 are not significant for predicting ROE. I would like to improve the current model by removing the three explanatory variables. In addition, I would like to keep the X_4 to see how its t-statistic and p-value changes in the second model.

Multi-Collinearity Regression Test on the First Model

A multi-variable regression model may include significantly correlated explanatory variables. Those correlated explanatory variables can be misleading to the users of the model. It is necessary to remove the highly correlated explanatory variables and only include the most significant ones. This will allow the users of the model to focus on the most important factors and help the users make sound business/investment decisions based on key influential factors. Therefore, I decided to test the multi-collinearity of the six variables in the above model.

The correlations between every two variables are shown in the table below.

Correlations	*X1*	*X2*	*X3*	*X4*	*X5*	*X6*
X1	1.0000					
X2	(0.3287)	1.0000				
X3	0.0211	0.0128	1.0000			
X4	0.0175	0.2336	(0.0318)	1.0000		
X5	0.1548	(0.3057)	(0.1758)	(0.1169)	1.0000	
X6	(0.0774)	(0.0474)	0.2832	(0.2578)	(0.0470)	1.0000

From the above table, I noticed that there are weak correlations between most of the variables. Therefore, I conclude that it is not necessary to remove any explanatory variables for multi-collinearity reason.

Second Model: $Y=\alpha+\beta 1X1+\beta 2X2+\beta 3X3$

Response Variables:

Y = ROE, Return on Equity

Explanatory Variables:

X_1 = Combined Ratio
X_2 = Net Yield on Invested Assets
X_3 = Age of Insurer

Resulted Equation from Regression in MS-Excel:

$$Y = 24.5548 - 0.2207X_1 + 1.8194X_2 - 0.0249X_3$$

Regression Statistics	
Multiple R	0.7582
R Square	0.5748
Adjusted R Square	0.5637
Standard Error	7.3068
Observations	119

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	24.5548	2.9222	8.4027	0.0000
X1 - Combined Ratio (%)	(0.2207)	0.0239	(9.2162)	0.0000
X2 - Net Yield on Invested Assets (%)	1.8194	0.3887	4.6802	0.0000
X3 - Age	(0.0249)	0.0137	(1.8136)	0.0723

By examining the above regression results, we can tell that the Combined Ratio and Net Yield on Invested Assets are still the most significant predictors of insurance company's ROE. Again, this finding is consistent with our hypothesis and intuitively makes senses.

The Adjusted R-square of 0.5637 is improved from the First Model, indicating the Second Model is slightly better. The X3, Age of Insurer, in my opinion, is not as significant as X1 and X2. I would like to look at the regression result of X1 and X2 only to see if I should include X3 into my predictive model for insurance company's ROE.

Third Model: $Y=\alpha+\beta 1X1+\beta 2X2$

Response Variables:

Y = ROE, Return on Equity

Explanatory Variables:

X_1 = Combined Ratio
X_2 = Net Yield on Invested Assets

Resulted Equation from Regression in MS-Excel:

$$Y = 23.4128 - 0.2218X_1 + 1.8046X_2$$

Regression Statistics	
Multiple R	0.7501
R Square	0.5627
Adjusted R Square	0.5551
Standard Error	7.3785
Observations	119

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	23.4182	2.8823	8.1249	0.0000
X1 - Combined Ratio (%)	(0.2218)	0.0242	(9.1780)	0.0000
X2 - Net Yield on Invested Assets (%)	1.8046	0.3925	4.5981	0.0000

By examining the above regression results, we can tell that the Combined Ratio and Net Yield on Invested Assets are significant predictors of insurance company's ROE. Again, this finding is consistent with our hypothesis and intuitively makes senses.

However, I find that the Adjusted R-square of 0.5551 is now smaller than the Second Model, indicating the Second Model is slightly better. Therefore, I should consider including the X3, Age of Insurer, in my predictive model for insurer's ROE, even though the regression results for X3 is not so significant as X1 and X2.

Conclusion

The variables X_1, X_2 and X_3 appear to have relatively high t statistics and low p-values signifying strong predictable power. It might be arguable whether we should use the Third Model with only X_1 and X_2 or the Second Model with all of the three variables. Purely based on the Adjusted R-square values, the Second Model seems to be the best one to use. However, taking into account the fact that the Third Model is much easier to follow intuitively and has one less explanatory variable, I conclude that the best predictive model for insurance company's ROE is the Third Model:

$$Y = 23.4128 - 0.2218X_1 + 1.8046X_2$$