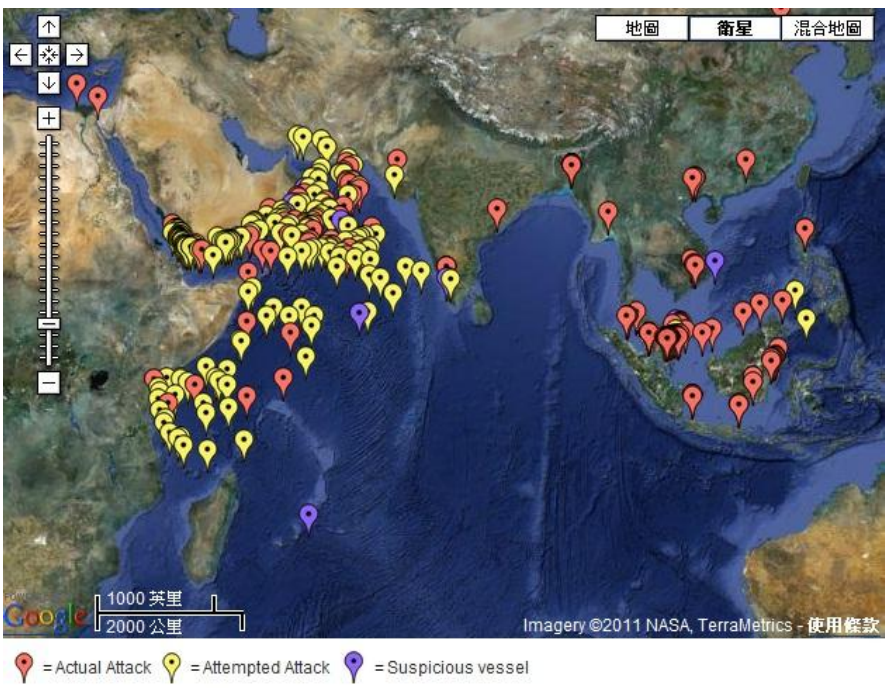
Regression Analysis Student Project
Nicholas Chu Hau Man
Winter 2011

VEE Regression Analysis Project

Model for Premium Pricing of the Maritime Kidnap and Ransom Insurance Market

Background

The pirate activities, especially those of the Somalian Pirates, has been a threat to the international shipping in recent years. Since, 2005, many international organizations such as the International Maritime Organization (i.e. IMO) has expressed concern over the rise of the piracy problems. A new insurance product called "Kidnap and Ransom Cover" (i.e. K&R) had been announced by some London underwriting leaders, such as Hiscox[1,2], since time around 2007, which mainly covers the Ransom paid to the pirate by the ship owners, as well as other relevant expenses such as negotiation expert fees etc., so as to meet the demand of the maritime market.



Piracy & Armed Robbery Map of 2011

Source: http://www.icc-ccs.org/piracy-reporting-centre/imb-live-piracy-map/imb-live-piracy-map

[1] (March 21, 2008). Hiscox USA Adds 2 Experts in Kidnap and Ransom Coverage. *Insurance Journal*, from http://www.insurancejournal.com/news/national/2008/03/21/88456.htm

[2] Wikipedia: Background of Hiscox - http://en.wikipedia.org/wiki/Hiscox

I hereby use the data of the Premium of some K&R policies written by varies maritime underwriters from the end of 2008 to early 2011, together with some other relevant statistics such as vessels' gross tonnages, as well as the piracy activities data at the policy inception month to perform regression analysis to try to find a proper model for the pricing of the K&R market. Although the underwriters used their own models to perform underwriting, their models will be of great differences and amended frequently, since the K&R product is too new to the underwriters and that the covered period of the product is mostly very short (say 7 days) similar to a travel insurance policy. As a result, it is a good idea to combine all underwriters' data together to investigate the fundamental factors behind the K&R pricing.

Data

For the premium rates and other information of the policies, I have used mainly the data of my company, an insurance/ reinsurance broker firm, for this student project. For confidential reasons, I have made the following amendments to the original data:-

1. Assigning nick names to each vessels;
2. Adding a few-days-constant to the policy inception dates but without making any amendment to the policy durations;
3. Adding a normal random variable with mean zero (with variance relatively small so that it will not distort the figure at a significant extend) to the policy premium before calculating the rates.

For the vessels' gross tonnages, the figures are obtained from Equasis[3], which is a database on the world's merchant fleet.

For the Pirate Activities Factor, I used the figures of "Monthly Global Piracy Actual and Attempt Attacks" stated in the Annual Piracy Reports issued by the International Chamber of Commerce[4] (i.e. ICC).

Factors

The original factors for this study are the Number of insured vessels of the policy (M), Average Gross Tonnage (Per vessel; G = ΣTonnage / M), Total Premium (USD; P), Total Sum Insured (USD; I), Transits Allowed (T), Premium Rate per Transit (R = P/ (I*T)), Fronting Policy (F), Nationality of the client's agents (S1 - S5), Inception Dates, Expiry Dates, Duration (Days; U), Number of Global Pirate Attacks in the Inception Month (A)

As explained above, K&R insurance is a short period insurance similar to the travel insurance. Normally the ship owners will purchase the cover just before their vessels' entering the risking area (i.e. the Somalia coast, or the Gulf of Aden etc.) and the cover will only last for a short period (i.e. say 7 days or 14 days). The factor "Transits Allowed" indicates the total transits allowed to enter the specified risky area by all insured vessels before the expiry date. For example, if the policy covers 6 vessels but the allowed transits is 2, if the same ship had

[3] Official Web site: http://www.equasis.org/EquasisWeb/public/HomePage
[4] Official Web Site: http://www.iccwbo.org/

transited through the risky area twice, the cover would be "used up" even if the other 5 vessels had not entered the risky area, and thus the ship owner need to purchase a new cover if he need to take some new transits.

Instead of using the Premium P for regression, I believe that there are some factors philosophically affecting the premium on a pro rata basis (such as Total Sum Insured), and would therefore combine several factors to produce a single index. In view of the above explanation of the meaning of Transits Allowed, I would use the Premium Rate per Transit (i.e. R = P/ (I*T)) instead of neither P/ (I*M) nor P/ (I*T*M) for the regression, because the number of vessels M appears not to be affecting the prices on pro rata basis even at philosophical aspect.

Different Gross Tonnage of the vessels will also have different risk exposures in the ocean as pointed out by ICC's piracy reports. For multi-vessels policy, I used Gross Tonnage per vessel (i.e. G = ΣTonnage / M) for the study.

I would not assign regressors to neither the inception date nor the expiry date for 2 reasons. First, the piracy activities are seasonal, and therefore the Pirate Attacks Number (A) is more effective to reflect the pricing reasoning than the inception time. Second, I will use duration (U) in the regression which already reflects the difference between the inception and expiry time.

For Fronting Policy (F), which means whether the underwriters are writing the business directly (i.e. non fronting), or they were seeking a local insurance company to write the policy on behalf of them (i.e. fronting) due to some regulation restriction. It might affect the pricing as one more party is involved in the fronting policy. I use F = 1 for fronting and F = 0 for non-fronting.

The ship owners would normally not contact my company (i.e. insurance broker) directly, instead they will contact us via their own agent (but they might change their agent as you can compare the data for vessels in same nick names) and we will therefore assist the agents to deal with the underwriters. After a few years of experience in the K&R business, I believe that the nationality of agents would affect the pricing as well. The nationality of the client's agents contains Korea, England, Japan, Hong Kong, Taiwan and Singapore, and I will use Deviation Regressors[5] S1 - S5 to code the nationalities as below:-

	S1	S2	S3	S4	S5
Korea	1	0	0	0	0
England	0	1	0	0	0
Japan	0	0	1	0	0
Hong Kong	0	0	0	1	0
Taiwan	0	0	0	0	1
Singapore	-1	-1	-1	-1	-1

[5] P. 146, Applied Regression Analysis and Generalized Linear Models (John Fox, 2008).

Analysis

The premium rate is neither counting process nor probability measure, and I would therefore prefer to use Normal Model instead of neither Poisson nor Binomial for the regression, and would use identical function as the link function g(.)[6].

I would use Excel's regression add-in to run the regression at 95% and 99% confidence interval, and analyze the results to see if I need to eliminate some of the explanatory variables for the best fit.

Regression Analysis 1 (Excel tab "Regression 1") - Overview

The regression equation is:

$R = \alpha + \beta_G G + \beta_F F + \beta_1 S_1 + \beta_2 S_2 + \beta_3 S_3 + \beta_4 S_4 + \beta_5 S_5 + \beta_U U + \beta_A A$

For the first regression, I take all 9 explanatory variables in the Excel's Regression.

The results of the regression are as below:-

Summary Output

Regression Statistics	
Multiple R	0.769735023
R Square	0.592492006
Adjusted R Square	0.556535418
Standard Error	0.001446998
Observations	112

ANOVA

	df	SS	MS	F	Significance F
Regression	9	0.000310515	3.45017E-05	16.47798203	2.07894E-16
Residual	102	0.000213568	2.0938E-06		
Total	111	0.000524083			

	Coefficients	Standard Error	t Stat	P-value
Incept	0.003857202	0.000491688	7.844823864	4.42334E-12
G	-6.19537E-09	5.142E-09	-1.204855695	0.231047076
F	0.001094182	0.000515527	2.122450996	0.036220792
S1	-0.000710453	0.000282154	-2.517957615	0.013358948
S2	-0.002013969	0.000289967	-6.945511486	3.60879E-10
S3	-0.000682527	0.000398762	-1.711614423	0.090006884
S4	-0.001813835	0.000445887	-4.067921611	9.36227E-05
S5	-0.000734734	0.000520933	-1.41041953	0.161458211
U	-5.44734E-06	1.29985E-06	-4.190745463	5.92629E-05
A	2.28691E-05	1.29906E-05	1.760436651	0.081331222

[6] P. 382, Applied Regression Analysis and Generalized Linear Models (John Fox, 2008).

So the equation now becomes:-
$R = 0.003857202 – “6.19537E-09”G + 0.001094182F - 0.000710453S_1 - 0.002013969S_2 - 0.000682527S_3 - 0.001813835S_4 - 0.000734734S_5 – “5.44734E-06”U + “2.28691E-05”A$

The resultant R^2 is 0.592 which is acceptable but not very good. The p-value of the F-test is 2.07894E-16, which suggests that this 9 explanatory variables are good predictors for R. However, the P-value for β_G is 0.23, which suggest that there is little evidence to reject the hypothesis “$\beta_G = 0$”. We will therefore take G off the model for comparison in Regression 2 to see if it would produce a better fit for us.

Before doing Regression 2, we look at the result of regression 1 at a deeper extend to investigate the data.

Regression Analysis 1 - Outliers

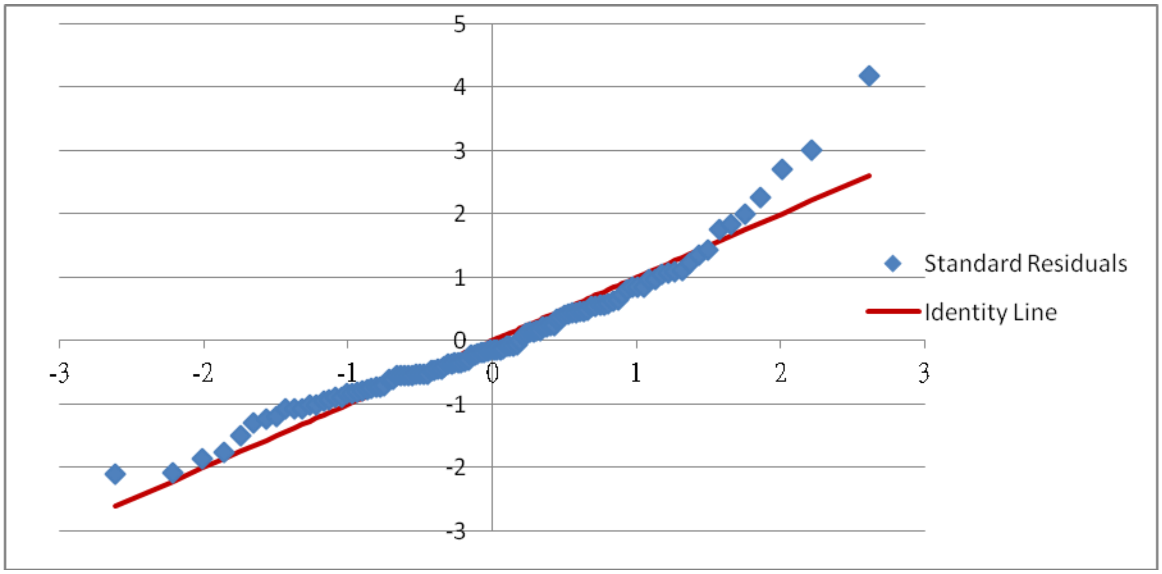
From the same tab of Regression 1, we would like to take a look at the outliers, and can see that there are some observation points with standard residuals larger than 2:-

Observation	Predicted R	Residuals	Standard Residuals	Absolute Value > 2 ?
9	0.003479675	-0.002918357	-2.103932996	> 2
32	0.004175556	0.005805397	4.185288492	>2.5
37	0.00238444	0.002777568	2.00243348	> 2
40	0.003382204	0.003151659	2.272127607	> 2
96	0.0039447	0.004173265	3.008634304	>2.5
100	0.004582231	-0.002878701	-2.075343444	> 2
103	0.004736363	0.003749382	2.703043942	>2.5

As we can see, among the 112 data points, there are 4 data points with Standard residuals between 2 and 2.5, while 3 of them are higher than 2.5. The observation no. 32 (i.e. vessel “Naara”) even have a standard residuals higher than 4.

Regression Analysis 1 – Normality Check

As can be seen from the same Excel tab, the Quantile-Comparison Plot of the Ordered Standard Residuals (i.e. y-axis) against Normal Quantiles (i.e. x-axis) is as below:-



As the points are above the identical line at both the upper and lower tail, the distribution of the standard residuals is positively skewed[7].

Regression Analysis 2 (Excel tab "Regression 2") - Overview

As mentioned above, the p-value of G of Regression 1 is not significantly small, and I would like to eliminate G to see if we can get a better model.

Thus the regression equation is now with 8 explanatory variables as below:
$R = \alpha + \beta_F F + \beta_1 S_1 + \beta_2 S_2 + \beta_3 S_3 + \beta_4 S_4 + \beta_5 S_5 + \beta_U U + \beta_A A$

Summary Output

Regression Statistics		*Compare to Regression 1*
Multiple R	0.765958419	*0.769735023*
R Square	0.586692299	*0.592492006*
Adjusted R Square	0.55459073	*0.556535418*
Standard Error	0.001450167	*0.001446998*
Observations	112	*112*

ANOVA

	df	SS	MS	F	Significance F
Regression	8	0.000307475	3.84344E-05	18.27612537	9.28245E-17
Residual	103	0.000216607	2.10298E-06		
Total	111	0.000524083			

[7] P. 37, Applied Regression Analysis and Generalized Linear Models (John Fox, 2008).

	Coefficients	Standard Error	t Stat	P-value	*Compare to Regression 1*
Incept	0.003797438	0.00049025	7.745912994	6.89002E-12	*4.42334E-12*
F	0.001007861	0.000511643	1.969851454	0.051540535	*0.036220792*
S1	-0.000738122	0.000281834	-2.61899208	0.010149831	*0.013358948*
S2	-0.001995891	0.000290213	-6.877335287	4.84171E-10	*3.60879E-10*
S3	-0.000562571	0.000386979	-1.45375023	0.14905583	*0.090006884*
S4	-0.001843329	0.00044619	-4.131265127	7.35654E-05	*9.36227E-05*
S5	-0.000888866	0.000506087	-1.756348469	0.082001098	*0.161458211*
U	-5.79976E-06	1.26928E-06	-4.569311547	1.36288E-05	*5.92629E-05*
A	2.03785E-05	1.28532E-05	1.585487639	0.11592108	*0.081331222*

The adjusted R^2 is now slightly decreased from 0.5565 to 0.5546, which means model 2 is slightly worse than model 1. The p-value of F-test is decreased from 2.07894E-16 to 9.28245E-17, which means it is easier for regression 2 to reject H_0 than regression 1. It seems that the elimination of G does not give a significant improvement for us. By looking at the p-value of the coefficient, the P-value of Incept, F, S2, S3 and A are increased, while S3 is increased from 0.090 to 0.149, A from 0.081 to 0.116.

Regression Analysis 2 - Outliers

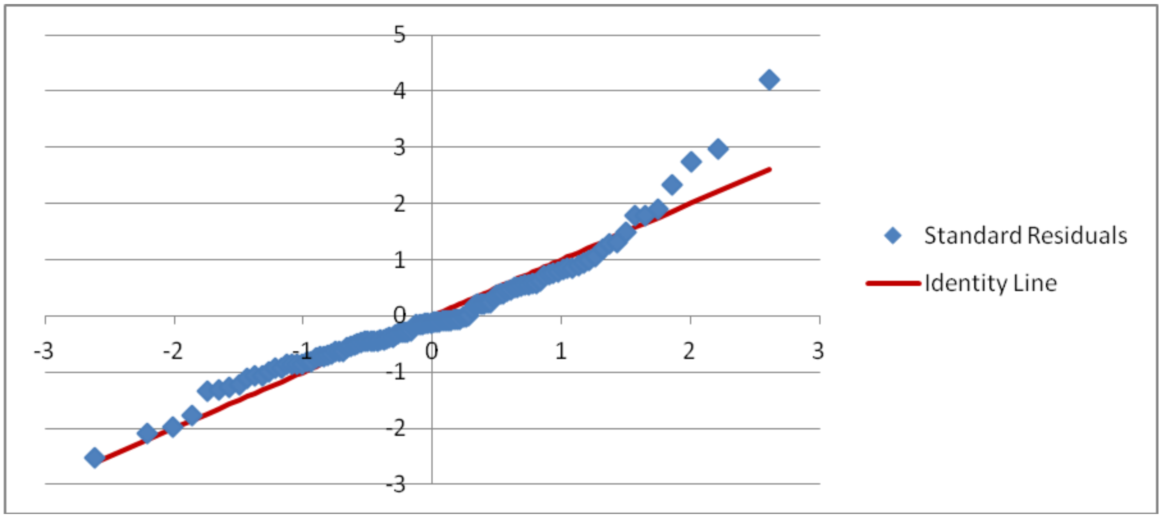
From the same tab of Regression 2, we can now see the following information about the residuals:-

Observation	Predicted R	Residuals	Standard Residuals	Absolute Value > 2 ?
9	0.003496679	-0.002935361	-2.101291223	> 2
32	0.004119158	0.005861795	4.196192786	>2.5
40	0.00326326	0.003270604	2.341276725	> 2
96	0.003970867	0.004147097	2.968718565	>2.5
100	0.005240669	-0.003537139	-2.532077152	>2.5
103	0.004661768	0.003823977	2.737411405	>2.5

For Regression 2, among the 112 data points, there are 2 data points with Standard residuals between 2 and 2.5, while 4 of them are higher than 2.5. Same as Regression 1, the observation no. 32 (i.e. vessel "Naara") have a standard residuals higher than 4 again. Besides, the observation no. 96 (i.e. Vessel "Iain") has the Standard Residuals with absolute 2^{nd} high.

Regression Analysis 2 – Normality Check

As can be seen from the same tab, the Quantile-Comparison Plot of Standard Residuals (i.e. y-axis) against Normal Quantiles (i.e. x-axis) is as below:-



As the points are above the identical line at both the upper and lower tail, the distribution of the standard residuals is positively skewed.

<u>Conclusion</u>

By comparing the adjusted R^2 and F-test values of Regression 1 and 2, Regression 2 does not give a significant improvement for us.