Thomas Haggerty

NEAA VEE Regression Analysis

Student project – Winter 2015

Pricing Insurance for natural disasters is challenging because the majority of losses are low frequency, high severity. The entire of industry loss data is still not sufficiently credible, so catastrophe models attempt to bridge this gap by simulating tens of thousands of potential events. The output from these models, as well as old-fashioned COPE (construction, occupancy type, etc) classification of exposures, provide underwriters with information to price risks. This study explores the relationship between premiums and modelled losses.

Catastrophe models output two important pieces of information on a risk – Average Annual Loss (AAL), and Uncertainty (Standard Deviation of Losses, hereby abbreviated SD). This study applied Regression Analysis to financial information on 3800 bound accounts (disguised by multiplicative factors to protect privacy). AAL and SD serve as quantitative explanatory variables, peril serves as a qualitative explanatory variable, and premium serves as the response variable (i.e. the values we aim to predict).

First, the disguised data was transformed for the sake of a better match with the assumptions of classical statistical models, namely transforming skewness and nonlinearity. The untransformed data suggested the need to descend the ladder of powers:

Hinges:	Premium	AAL	SD
Median	34,206	5,613	143,564
25th percentile	15,048	1,577	44,993
75th percentile	71,545	16,647	345,443
$(M_U - \text{Median})/(\text{Median} - M_L)$	1.949	2.734	2.048

The ratio of the largest to smallest values for each variable are sufficiently large, and all values are positive, so there is no need to add a start to the data.

The following transformations achieved the desired corrections, becoming more normal as indicated by the proportion $(M_U - \text{Median})/(\text{Median} - M_L)$ being close to 1.

Hinges:	ln(Premium)	AAL^(1/5)	SD^(1/3)
Median	10.440	5.621	52.362
25th percentile	9.619	4.361	35.567
75th percentile	11.178	6.988	70.191
$(M_U - \text{Median})/(\text{Median} - M_L)$	0.899	1.084	1.062

The data are fit to four different regression models.

1) ln(Premium) = Intercept + B1*AAL^(1/5)

This is the simple-regression model. The Regression function in Microsoft Excel's Data Analysis package produces this output:

Regression Statistics	
Multiple R	0.77212272
R Square	0.596173494
Adjusted R Square	0.596067808
Standard Error	0.731649395
Observations	3823

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3019.679955	3019.679955	5640.984161	0
Residual	3821	2045.422709	0.535310837		
Total	3822	5065.102664			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	7.956520823	0.034652536	229.6086171	0	7.888581581	8.024460066	7.888581581	8.024460066
AAL^(1/5)	0.421848615	0.005616674	75.10648548	0	0.410836649	0.432860581	0.410836649	0.432860581

The statistic "R Square" is the Coefficient of Determination. It indicates what portion of the variation in premium is explained by the regression. The value of ~0.6 indicates good but not great predictive power for the model.

Regarding the calculations in Excel's ANOVA output: A regression can be decomposed into the explained and unexplained portions as TSS = RegSS + RSS, defined as:

TSS = total sum of squares
RegSS = regression sum of squares
RSS = residual sum of squares

R Square = RegSS/TSS

Residuals represent the difference between observed value of the response variable and its regression-fitted value of the. The aim of classical regression analysis is to minimize RSS and maximize RegSS.

MS stands for "mean square", which is the sum of squares (SS) divided by the degrees of freedom (df).

Using the multiple-regression model:

2) ln(Premium) = Intercept + B1*AAL^(1/5) + B2*SD^(1/3)

Regression Statistics	
Multiple R	0.840841059
R Square	0.707013687
Adjusted R Square	0.706860291
Standard Error	0.62328419
Observations	3823

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	3581.096911	1790.548455	4609.075865	0
Residual	3820	1484.005753	0.388483181		
Total	3822	5065.102664			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	8.468441919	0.032446527	260.9968689	0	8.404827739	8.532056099	8.404827739	8.532056099
AAL^(1/5)	-0.005247778	0.012211353	-0.429745777	0.667404818	-0.029189176	0.018693621	-0.029189176	0.018693621
SD^(1/3)	0.035209102	0.000926186	38.01514505	1.7283E-268	0.033393235	0.037024969	0.033393235	0.037024969

This model suggests a better fit than the first model, as evidenced by the higher value of R Square. The value "Multiple R" is the correlation coefficient. It is the square root of R Square, and measures the strength of a linear relationship. That it is approaching the value 1 indicates a strong positive relationship. (Zero would indicate no relationship, and approaching the value -1 would indicate a strong negative relationship.)

The statistic Adjusted R Square accounts for degrees of freedom and gives a more realistic indication of goodness of fit, since R Square could increase even if adding spurious explanatory variables.

Adjusted R Square = 1 – [RSS/(n-k-1)]/[TSS/(n-1)]

Incremental F-tests between models (for nested models, and adhering to the principal of marginality) test the significance of the slope coefficients. The high p-value for the coefficient B1 is concerning. It represents the chance that the value of B1 is as observed, or more extreme, due to random fluctuations. A low p-value would indicate that the coefficient is significant.

An additional problem with this model is collinearity between the two explanatory variables. Excel returns CORREL(AAL, SD) = 0.86. For this multiple-regression model, the variance of slope coefficients is increased by a variance-inflation factor (VIF), which in this case is 7.12. The square root of this value comes to 2.67, suggesting that this collinearity cuts the precision of the coefficient estimates by more than half.

3) ln(Premium) = Intercept + B1*AAL^(1/5) + B2*SD^(1/3) + γ1*peril1 + γ2*peril2

Regression Statistics	
Multiple R	0.842275086
R Square	0.70942732
Adjusted R Square	0.709122896
Standard Error	0.620874115
Observations	3823

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	3593.322208	898.330552	2330.39244	0
Residual	3818	1471.780456	0.385484666		
Total	3822	5065.102664			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	8.537334502	0.035308073	241.7955409	0	8.468110006	8.606558998	8.468110006	8.606558998
AAL^(1/5)	-0.028564608	0.013350321	-2.139619654	0.032448738	-0.054739053	-0.002390163	-0.054739053	-0.002390163
SD^(1/3)	0.037741565	0.001117304	33.77915057	4.3843E-219	0.035550996	0.039932134	0.035550996	0.039932134
Peril1	-0.141728343	0.02730267	-5.191006696	2.20058E-07	-0.195257563	-0.088199124	-0.195257563	-0.088199124
Peril2	-0.051382915	0.035442137	-1.449769086	0.147205058	-0.120870256	0.018104426	-0.120870256	0.018104426

This model makes use of polytomous dummy regressors for peril. Since there are three types of perils (A, B, and C), two dummy variables get coded as such:

	peril1	peril2
perilA	1	0
perilB	0	1
perilC	0	0

This model provides very little additional goodness of fit (as measured by Adjusted R Square) compared with model #2. Additionally, the high p-value for the coefficient Peril2 directs us not to reject the null hypothesis that γ1 = 0. In other words, γ1 is not statistically significant at levels of alpha ≈ 15% or less. Alpha signifies the chance of a Type I error, which is rejecting the null hypothesis when it's true.

4) ln(Premium) = Intercept + B1*AAL^(1/5) + B2*SD^(1/3) + γ1*peril1 + γ2*peril2 + δ11*(AAL^(1/5)*peril1) + δ12*(AAL^(1/5)*peril2) + δ21 *(SD^(1/3) *peril1) + δ22*(SD^(1/3) *peril2)

This model contains interaction regressors. It states that the effects of AAL and SD both vary by peril. In other words, the regressions surfaces for each peril are not parallel.

Regression Statistics	
Multiple R	0.873319313
R Square	0.762686623
Adjusted R Square	0.762188849
Standard Error	0.561402769
Observations	3823

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	3863.257874	482.9072343	1532.197011	0
Residual	3814	1202.070086	0.315173069		
Total	3822	5065.32796			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	8.054092605	0.066486509	121.1387494	0	7.923740075	8.184445136	7.923740075	8.184445136
AAL^(1/5)	-0.247787644	0.039531174	-6.268158023	4.06021E-10	-0.325291917	-0.170283372	-0.325291917	-0.170283372
SD^(1/3)	0.079894496	0.00379133	21.0729448	2.42937E-93	0.072461266	0.087327725	0.072461266	0.087327725
Peril1	0.706161227	0.079076517	8.930100321	6.48702E-19	0.551124902	0.861197552	0.551124902	0.861197552
Peril2	1.119526334	0.095262906	11.75196495	2.38089E-31	0.932755198	1.30629747	0.932755198	1.30629747
AAL^(1/5)*Peril1	0.148296499	0.042395477	3.497932085	0.000474235	0.065176513	0.231416485	0.065176513	0.231416485
AAL^(1/5)*Peril2	0.138118294	0.047046272	2.935796802	0.003346803	0.045880024	0.230356563	0.045880024	0.230356563
SD^(1/3)*Peril1	-0.041021148	0.003987944	-10.28628865	1.69298E-24	-0.048839857	-0.033202439	-0.048839857	-0.033202439
SD^(1/3)*Peril2	-0.044011063	0.004302655	-10.22881444	3.0209E-24	-0.052446789	-0.035575336	-0.052446789	-0.035575336

Model #4 satisfactorily has the highest value of Adjusted R Square, while also having low p-values for all its coefficients. However, the implication that AAL has a negative effect on premium is nonsensical, so this model is rejected.

For all these models, the omnibus null hypothesis that all coefficients are equal to zero is easily rejected per the high F value. The value of nil for Significance of F means that there's essential no probability that the coefficients' values are due to chance alone.

Thus the accepted model is:

ln(Premium) = 7.96 + 0.42*AAL^(1/5)

These models actually include an error term as well, on the right: ε_i. The linearity assumption is that this term's expected value is zero. Additionally, the error term is assumed to have constant variance, and be independent from data point to data point. If an important explanatory variable (i.e. one structurally related to, i.e. a causative factor of, the response variable) is not captured in a statistical model, then it is absorbed into the error term. This introduces bias into the model, and the assumptions of classical least-squares estimation are compromised. Additional causative categorical variables such as geography or construction could be explored; however, the catastrophe modelling software accounts for these when simulating losses. A key variable, portfolio aggregation at the time of quoting, is not easily available and thus we would likely violate another key assumption of regression analysis—that the observed values for the explanatory variables are measured without error.

It’s worth noting that many in the catastrophe insurance industry express pricing in terms of a basic average annual loss ratio (i.e. premium = loss / [target AALR]), akin to model #1. Though the chosen model is not the best-fitting, it is simple to apply and understand in the field.

Due to the infrequency of catastrophe events, market softening can be prolonged. This exerts downward pressure on pricing, a temporal effect. Over the course of a year, brokers could recite the depression of the target AALR that the market bears. Perhaps we can revisit this subject the in the NEAA Time Series course, which deals with observations that are not independently distributed. In fact, the author recently fit small datasets from the Florida windstorm market. Comparing 2011 with 2015 data, continued softening has depressed the basic market pricing model from 3*AAL to 2*AAL. The regression fit for the 2015 data is not as ideal (lower R square), presumably because the data set is not yet complete. Logarithmic transformation of the limited 2015 data produced a better fit (higher R square) than the simple, easy-to-understand single-coefficient, no intercept model, but the predicted premiums actually presented a nonsensical trend. The fit implied a lower loss ratio for higher AAL accounts, whereas the opposite is apparent due to greater competition between brokers for higher-premium (and thus higher AAL) accounts. This shows that a well-fitting regression may still not be useful if the data has problems.