MSDS 6372 : Regression Analysis

***Predictions of New Business Production***

**1. Introduction**

For any insurance provider, new business is a major component of profitability (Top line growth) and understanding the close ratio in addition to sustaining the existing book of business to renewals. The purpose of this study is to understand the key factors influencing the amount of new business (new insurance policies) production metrics for an insurance company. Considering this is **an observational study**, we will not be able to make a causal inference however, understanding the key factors and predictions will help tremendously on how the business is performing on whole, by line of business, and regions to help enable better business and strategic decisions around agency investments, marketing campaigns etc.,.

In order to help with the understanding the predictors for the response, let us look into the dataset to better know the distribution, linearity assumptions so that multiple linear regression model could be used to help with the questions of interest in this context. . The dataset used here provides the view of the new business, full quotes, line of business (package versus non-package), and region for the past three years from 2013 to 2015.

**2. Exploratory Analysis**

In order to understand the influencing factors and find answers for the questions of interest, I collected the given the dataset from the insurers that would adequately help explain the volume of the new business policies and the number of full quotes by region (grouping of states by region), line of business which are primarily product lines called "package policy" vs. "non-package policy" for the past three years. At this time, I have not included any agency or customer attributes as part of the analysis. Based on the understanding of the domain, I am choosing these explanatory variables assuming they will help predict the response significantly. There are about 200+ observations in the given dataset. Below table describes the response and predictor variables in detail.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **2.1. Attribute Name : New Business (nb)**  The total number of new policies written which would be used as a response variable, grouped by region, state and line of business.   |  |  | | --- | --- | |  |  | | | **Basic Statistical Measures** | | | | | --- | --- | --- | --- | | **Location** | | **Variability** | | | **Mean** | 751.6009 | **Std Deviation** | 1541 | | **Median** | 158.5000 | **Variance** | 2375335 | | **Mode** | 0.0000 | **Range** | 9506 | |  |  | **Interquartile Range** | 606.50000 | | | **Extreme Observations** | | | | | --- | --- | --- | --- | | **Lowest** | | **Highest** | | | **Value** | **Obs** | **Value** | **Obs** | | 0 | 224 | 7605 | 32 | | 0 | 208 | 7981 | 105 | | 0 | 201 | 8498 | 81 | | 0 | 192 | 9294 | 16 | | 0 | 176 | 9506 | 12 | |   The overall mean "new business" across all regions is 751.6 with a standard deviation of 1541. Based on the histogram and Q-Q plot, the data are highly right skewed with possible outliers. Before I get into addressing the outliers, I would like to see if a log transform of "new business" would help with distribution. Based on the histogram and QQPlot of logNB, the distribution looks much better and normally distributed. Let us move forward with using the log transformed data for further regression analysis.   |  |  | | --- | --- | |  |  | |
| **2.2. Attribute Name : Full Quote (fq)**  Full quote attribute represents the serious quotes (completed quote) for an insurance policy. Any new business usually originates from a full quote. This attribute in the given dataset represents the number of full quotes by state, region, line of business for 2013, 2014 and 2015. Like you have guessed, full quote is the quantitative predictor/explanatory variable to explain the new business growth. |
| **2.3. Attribute Name : Region**  Region is a categorical variable with three levels (east, west and northeast) simply a grouping of states in united states. All the states in U.S is categorized into one of three regions based their geographical positions as listed in the dataset. |
| **2.4. Attribute Name : Line of Business (lob)**  Line of Business is a categorical variable with two levels (package policy and non-package policy) represents the two type of products sold by this insurer for mass-affluent customers (household income of > 150K per year). The business context for package policy means it is a "one policy" for the customer which covers all their risks such as Vehicle, Property, Boat, and any rental properties they may own. This is promoted as the preferred strategic offering to attract the consumer group. Non-Package policy would mean separate policies for vehicles, property and other risks based on what customers would like to protect. |

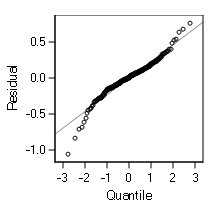
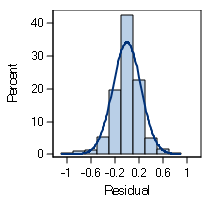
**3. Validation of assumptions**

As we explore the option on using multivariate linear regression, let us review dataset and its distribution for the following key assumptions that are critical for using linear regression :

3.1 **Linearity :** As linear relation assumption is expected of response given explanatory variable, log transformation of New business ends up having not so linear relationship with full quote. So, I will move forward with log transform of the quantitative explanatory variable (full quote) which helps with addressing the linearity assumption.

|  |  |
| --- | --- |
| Log(newbusiness) vs fullquote | Log(newbusiness) vs log(fullquote) |
|  |  |

3.2 **Normality :** The normality assumption for linear regression applies to the errors, not the outcome variable and not to the explanatory variables. However the Standard Errors influence the predictions and would cause broader prediction band even if the sample size is large. Reviewing the residuals histograms and QQ plots below, which fairly looks normal move forward with the test.



**3.3 Independence:** We assume the data are independent from one another.

**3.4 Constant variance :** Based on the residual value of the predicted observations vs. the scores, the variance seems to randomly distributed and fairly equal.

As most of the assumptions are not violated as much, let us move forward with the regression analysis to understand the relationship between the response and given predictor variables. However based on the dataset and the Cooks'D plot, I do see potential problems with extreme outliers. There is not much we could do with the dataset as this is real data and observational study. But I would run the model with and without the extreme outliers to understand the difference.

**4. Regression Analysis**

To begin the data analysis, based on the response and given attributes as explanatory variables, let us start with the regression model below:

|  |
| --- |
| **µ {log(NB) | log(fq) region lob} = βₒ + β₁(log(fq))+ β₂(east) + β₃(west) + β₄(package)** |

To find the best fit model, let us use a variable selection method of STEPWISE considering we dont have too many explanatory variables.

Parameter Estimate (REG Procedure)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Analysis of Variance** | | | | | | | --- | --- | --- | --- | --- | --- | | **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** | | **Model** | 2 | 190.11663 | 95.05831 | 1731.66 | <.0001 | | **Error** | 225 | 12.35122 | 0.05489 |  |  | | **Corrected Total** | 227 | 202.46785 |  |  |  | | | **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** | | --- | --- | --- | --- | --- | --- | | **Intercept** | -0.76274 | 0.05182 | 11.89324 | 216.66 | <.0001 | | **logfq** | 0.92613 | 0.01690 | 164.77381 | 3001.65 | <.0001 | | **package** | 0.39151 | 0.03144 | 8.51380 | 155.09 | <.0001 |  |  | | --- | | ***Variable logfq Entered: R-Square = 0.9390 and C(p) = 1.3741*** | |

The regression equation based on the parameter estimates and statistical significance, after accounting for all the predictor variables (full quote, region and lob):

|  |
| --- |
| **Log(nb) = βₒ + β₁(log(fq)) + β₄(package)**  **Log(nb) = 0.92613 (log(fq)) + .39151(package) -0.76274** |

As we now have the fitted model, let us take a brief look at the Cooks'D plot to understand the influencing and high leverage observations in the fitted model and derivation of correlation coefficients.

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
| For the partial Cooks's plot below for lognb, observation 201, 208, 224 stands out to be highly influencing compared to the other observations. (please note there are few more observations found like these below) |  |

|  |  |
| --- | --- |
| Based on the graphical review of the leverage points, and reviewing the actual observations, I would like to move forward with deleting the observations : 108, 122, 132, 185, 201, 208, 224 and refit the model to get a better estimate for the correlation coefficients. |  |

A quick review of the fit diagnostics (above) after deleting the outlier observations, the residuals and leverage points looks better and the assumptions are still holds good.

Refitting the model with the new dataset, provides the below parameter estimates:

Parameter Estimate (REG Procedure)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Analysis of Variance** | | | | | | | --- | --- | --- | --- | --- | --- | | **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** | | **Model** | 2 | 863.56615 | 431.78307 | 1991.25 | <.0001 | | **Error** | 218 | 47.27108 | 0.21684 |  |  | | **Corrected Total** | 220 | 910.83723 |  |  |  | | | **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** | | --- | --- | --- | --- | --- | --- | | **Intercept** | -1.85186 | 0.11575 | 55.50130 | 255.96 | <.0001 | | **logfq** | 0.93689 | 0.01647 | 702.08546 | 3237.81 | <.0001 | | **package** | 0.95718 | 0.06403 | 48.45134 | 223.44 | <.0001 |  |  | | --- | | ***Variable package Entered: R-Square = 0.9481 and C(p) = 3.3995*** | |

Please note that this model does provides a slightly increased R-squared value of 0.9481 compared to the previous model. However, the new model estimates does end up in increased sum of squared errors compared to the previous model and also not a significant increase in R-square. So I decide to keep the originally fitted model with all the observations, which is re-stated below .

|  |
| --- |
| **Log(nb) = βₒ + β₁(log(fq)) + β₄(package)**  **Log(nb) = 0.92613 (log(fq)) + .39151(package) -0.76274** |

**5. Conclusion**

**Statistical Conclusion** : Based on the parameter estimates for the explanatory variables and their p-values, accounting for all other plausible variables accounted**, "full quote (fq)"** and "**lob"** found to have a strong positive correlation with the "**new business(nb)"** and explains about 93% of the mean responses (Rsquare = 0.939). Regions with the three levels (east, west and northeast) does not seem to have any statistical significance on the response and has been rejected from the model.

Since we have outcome and some of the predictor variables log transformed, let us look at how to best interpret the regression coefficients.

* From the overall final model perspective.

**Log(nb) = 0.92613 (log(fq)) + .39151(package) -0.76274**

if we assume a fullquote = 100, and package is kept constant at 1, then, we can say that the response value would be:

**log(nb) = 0.92613\*(LOG(100,10)) + 0.39151(1) - 0.76274 = 1.48103**

**nb = 10^1.48103 = 30.2712**

\*\* base 10 log transformation.

This can be explained as, for a full quote volume of 100, we expect as mean response of 30.27 holding package as constant (in this case 1).

In case of package is kept constant at 0 (for nonpackage policies), then, we can say that the response value would be:

**log(nb) = 0.92613\*(LOG(100,10)) + 0.39151(0) - 0.76274 = 1.48103**

**nb = 10^1.08952 = 12.2891**

which suggests that, for a full quote volume of 100, we expect as mean response of 12.29 non-package policies (package = 0).

* For the categorical variables that are not log transformed, in this case **package** , we can interpret that to be, for 1 unit of increase in "package" policy, we expect about .47% increase in the non-package policy since (exp(0.39151) = 1.477). In other words, for every 100 unit increase in package policy we would have about 47 units of increase in non-package policy.

**Scope of inference :** The statistical association from these observational data cannot be used to establish a causal interpretation. However, based on the parameter estimates, we do see that there is a very strong positive correlation between newbusiness and the explanatory variables full quote and package. **With the R-squared value of 0.939, we can say that with 95% confidence level, that the 93% of the response variable (new business) is explained by the predictor variables new business and line of business.**

**References**

1. The Statistical Slueth - A Course in Methods of Data Analysis III Edition - by Ramsey/Schafer

2. [http://www.ats.ucla.edu/stat/mult\_pkg/faq/general/log\_transformed\_regression.htm](http://www.ats.ucla.edu/stat/mult_pkg/faq/general/log_transformed_regression.htm" \t "_blank)

3. http://support.sas.com/rnd/app/stat/procedures/CategoricalDataAnalysis.html

**APPENDIX**

**SAS CODE and DATASET**

|  |
| --- |
| /\* New Business Regression \*/  /\* Load data into SAS for testing \*/  DATA mydata;  INFILE '/folders/myfolders/sasuser.v94/NBRegression.csv' DELIMITER=',' firstobs=2;  INPUT year lob $ state $ Region $ FQ NB CloseRatio package nonpackage east northeast west;  RUN;  proc print data=mydata; run;  /\* Exploratory analysis : Histograms and QQ Plots \*/  proc univariate data=mydata;  var nb;  class package;  histogram nb;  qqplot nb;  run;  data mylogdata;  set mydata;  lognb = log(NB+1);  logfq = log(FQ);  run;  proc print data=mylogdata; run;  /\* Exploratory analysis with log data : Histograms and QQ Plots \*/  proc univariate data=mylogdata;  var lognb;  class lob;  histogram lognb;  qqplot lognb;  run;  proc sgscatter data=mylogdata;  matrix lognb logfq;  run;  /\* Linear regression model with log transformed data using stepwise variable selection technique \*/  proc reg data=mylogdata;  model lognb=logfq east west package / selection = stepwise R Influence CLI CLM ;  run;  /\* Deleting observations with high leverage to rerun the model for refittment \*/  data mylogdata\_xleverage;  set mylogdata;  if \_n\_ = 108 then delete;  if \_n\_ = 122 then delete;  if \_n\_ = 132 then delete;  if \_n\_ = 185 then delete;  if \_n\_ = 201 then delete;  if \_n\_ = 208 then delete;  if \_n\_ = 224 then delete;  run;  proc print data=mylogdata\_xleverage; run;  /\* Rerun of linear reg with mylogdata\_xleverage dataset \*/  proc reg data=mylogdata\_xleverage;  model lognb=logfq east west package / selection = stepwise R Influence CLI CLM ;  run;  proc sort data=mylogdata out=mylogsorteddata;  by descending lob region;  run;  proc print data=mylogsorteddata; run;  /\* Cross validation of the model\*/  proc GLMSELECT data=mylogsorteddata;  class lob region;  model lognb=logfq lob region/ selection = LASSO cvmethod=random;  run;  /\* End New Business Regression \*/ |
|  |