STAT 481 Project

By: Alex Bandurin

Summary:

An insurance company that specializes in addressing claims regarding coronary heart disease is interested in predicting the total cost of services using data collected from their subscribers in the following categories: Age, Gender, Interventions, Drugs, Emergency, Complications, Comorbidities, and Duration.

Right away, any missing data detected has been deleted, and descriptive statistics have been provided for the data in the categories/variables listed above for a better understanding of client information. These are presented below:

Sample Size (of all the variables): 665

Cost:

- Minimum: 8
- Maximum: 41803.2
- Median: 423.2
- Mean: 1664.46
- Standard Deviation: 4172.58
- Variance: 17410391.42

Age:

- Minimum: 24
- Maximum: 70
- Median: 60
- Mean: 58.86
- Standard Deviation: 6.57
- Variance: 43.21

Gender:

- Number of Females: 523
- Number of Males: 142

Interventions:

Minimum: 0

Maximum: 20

Median: 2

Mean: 3.79

Standard Deviation: 3.92

Variance: 15.38

Drugs:

Minimum: 0

Maximum: 9

Median: 0

Mean: 0.313

Standard Deviation: 0.796

Variance: 0.634

Emergency:

Minimum: 0

Maximum: 14

Median: 3

Mean: 3

Standard Deviation: 2.17

Variance: 4.72

Complications:

Minimum: 0

Maximum: 1

Median: 0

Mean: 0.036

Standard Deviation: 0.187

Variance: 0.0348

Comorbidities:

Minimum: 0

Maximum: 33

Median: 1

Mean: 3.55

Standard Deviation: 5.17

Variance: 26.74

Duration:

Minimum: 0

Maximum: 372

Median: 161

Mean: 162.82

Standard Deviation: 121.21

Variance: 14691.23

Furthermore, all these variables have been analyzed in a regression model using SAS software.

Goal:

Construct a regression model to analyze subscribers' data through regression analysis.

Results and Discussion:

The first step in regression analysis was to check for multicollinearity, and exclude any variables that affect cost with a variance inflation factor (VIF) greater than 10. This ensures that the model does not contain unnecessary/redundant information. For example, if "emergency" conveyed a lot of the information already provided by "drugs", then it would have been deleted.

This dataset did not include any such variables however, as could be seen in the table below:

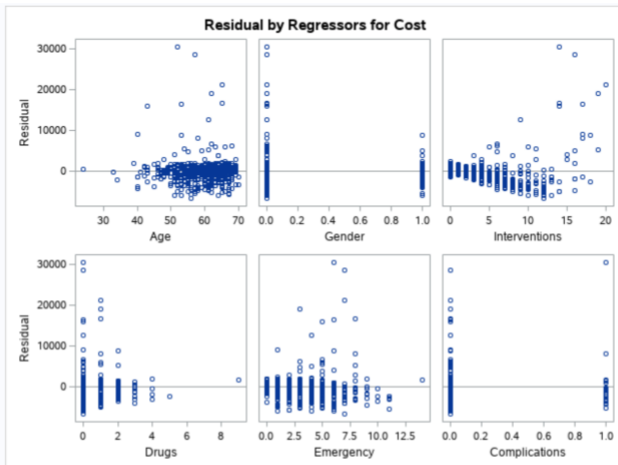
Parameter Estimates

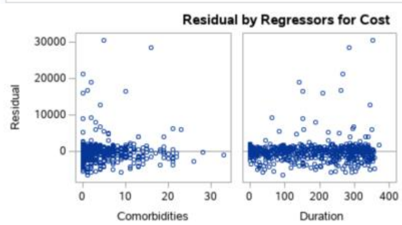
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	-600.77617	1063.02513	-0.57	0.5722	.	0	-2688.11831	1486.56597
Age	Age	1	-18.33468	18.03264	-1.02	0.3096	0.96911	1.03188	-53.74333	17.07397
Gender	Gender	1	-683.04425	288.26888	-2.37	0.0181	0.97422	1.02646	-1249.08523	-117.00327
Interventions	Interventions	1	676.11213	31.99518	21.13	<.0001	0.86471	1.15645	613.28681	738.93744
Drugs	Drugs	1	-446.22094	166.92726	-2.67	0.0077	0.77081	1.29734	-773.99711	-118.44478
Emergency	Emergency	1	249.49000	63.43864	3.93	<.0001	0.71752	1.39369	124.92272	374.05728
Complications	Complications	1	1583.90662	642.64296	2.46	0.0140	0.94632	1.05673	322.02137	2845.79187
Comorbidities	Comorbidities	1	103.92174	26.73901	3.89	0.0001	0.71215	1.40419	51.41737	156.42610
Duration	Duration	1	-0.63742	1.16706	-0.55	0.5851	0.68047	1.46956	-2.92904	1.65420

Therefore, nothing needed to be removed.

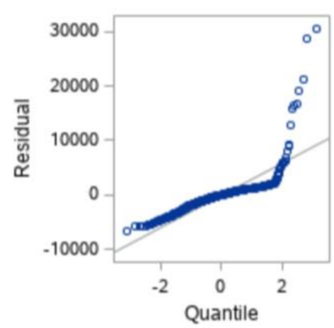
Next, the model assumptions were checked. These include linearity, normality or residuals, and equal variance or residuals. Independence of residuals did not need to be checked as this is not time-series data.

Linearity: This criterion tells whether a chart of cost plotted against each of the 8 variables mentioned earlier forms a straight line or not. If it does form a line, that means that the relationship between the two variables is linear. Based on the charts of residual vs independent variable, this assumption was **not met** as there were noticeable patterns detected for every independent variable:



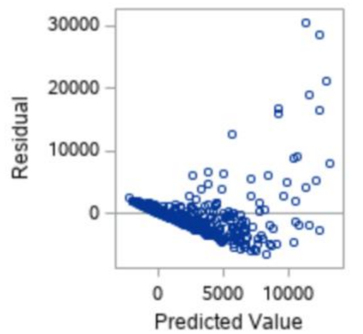


Normality: This shows if the errors (difference between actual and predicted values) are distributed in a way that agrees with normal distribution. **Met.** The Shapiro-Wilk test had a very low P-value (below 0.05 significance level), and the Q-Q plot looked mostly normal:



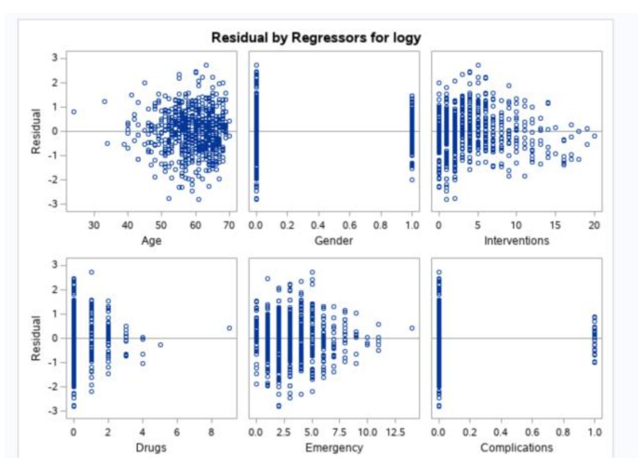
Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.64003	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.212909	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	7.551294	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	43.56842	**Pr > A-Sq**	<0.0050

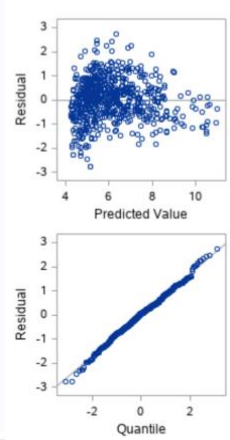
Equal Variance: This shows how much the errors vary for every subscriber. **Not met.** There is a pattern in the graph of residual vs predicted value:

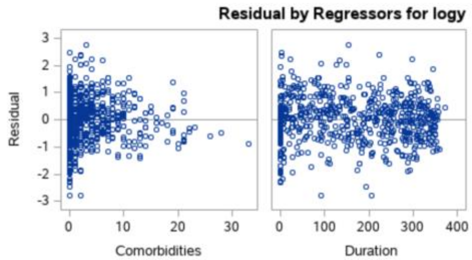


Since the linearity and equal variance assumptions have not been met, the model was adjusted using BoxCox transformation. The λ value suggested was 0.1, which has been rounded to 0. The Y variable was transformed to $Y' = \log_e(Y)$.

After this transformation, the "residual vs x" plots have become a little less patterned, while the "residual vs predicted value" plot has become much less patterned than before. The qq plot became more normal than before the transformation also. These results are visualized below:

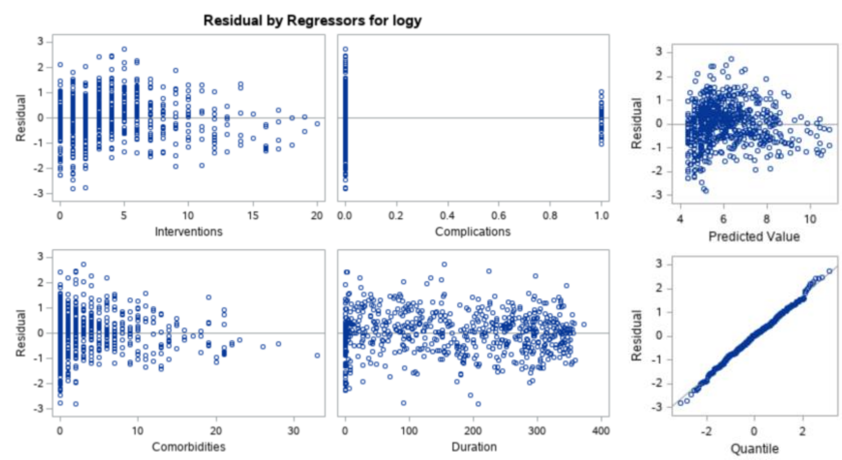






Thus, the equality of variance criteria and linearity criteria are closer to being met now, however there is still some work to be done.

Backward selection was performed on the transformed model in order to omit insignificant variables. The variables that were kept had a significance of 0.10 or lower. So only 4 of the 8 independent variables were kept, including “interventions”, “complications”, “comorbidities”, and “duration”. This means that “age”, “gender”, “drugs”, and “emergency” variables were eliminated. While the equal variance criteria did not change, the linearity criteria became much closer to being met as the four “residual vs x” plots now have little pattern.



Conclusion:

The R^2 before changes to the model were made was 0.4870. After Y transformation and backward elimination, R^2 went up to 0.7239. This means that the independent variables are able to explain the variability, or any change in cost almost 50% better than before any changes to the model were made. It could be said that the four variables used in the final model for the insurance company (“interventions”, “complications”, “comorbidities”, and “duration”) are significant in explaining the variation in predicting the total cost of services provided to subscribers, while the other variables (“age”, “gender”, “drugs”, and “emergency”) are insignificant. Keeping only these four variables ensures a higher linearity of the model and a higher r-squared value.

Code:

```
* STAT481 Insurance Project */
/** Import an XLSX file.  **/

PROC IMPORT DATAFILE="/home/u59825025/sasuser.v94/Project1Dataset (1).xlsx"
            OUT=insurance
            DBMS=XLSX
            REPLACE;
RUN;

PROC CONTENTS DATA = insurance; RUN;
/* Full Model with All Variables */
PROC REG DATA= insurance;
MODEL Cost = Age Gender Interventions Drugs Emergency Complications Comorbidities Duration / clb corrb tol vif collin;
OUTPUT OUT = result1 residual = residual;
TITLE 'Full Model';
RUN;

PROC UNIVARIATE DATA=result1 NORMAL PLOT;
VAR residual;
RUN;

/* Transformation on Y (Cost) required.  Use Box-Cox Transformation. */
/* Note: Box-Cox only works for dependent variable(s) */
PROC TRANSREG DATA=insurance DETAIL;
MODEL BOXCOX(Cost / convenient lambda = -2 to 2 by 0.5)
    = identity(Age Gender Interventions Drugs Emergency Complications Comorbidities Duration );
TITLE 'Boxcox Transformation';
RUN;

/* Perform a transformation on Y */
DATA insurance;
SET insurance;
logy = log(Cost);
RUN;

/* Full Model with All Variables After Transformation */
PROC REG DATA=insurance;
MODEL logy = Age Gender Interventions Drugs Emergency Complications Comorbidities Duration  / clb corrb tol vif collin;
OUTPUT OUT = result2 residual = residual;
TITLE 'Full Model After Transformation';
RUN;

PROC UNIVARIATE DATA=result2 NORMAL PLOT;
VAR residual;
RUN;

/* Run Backward Selection on Model - with logy and logx1 */
PROC REG DATA=insurance;
MODEL logy = Age Gender Interventions Drugs Emergency Complications Comorbidities Duration
    / selection = backward clb corrb tol vif collin CP SLSTAY = 0.10;
TITLE "BACKWARD SELECTION";
OUTPUT OUT = result4 residual = residual;
RUN;

PROC UNIVARIATE NORMAL PLOT DATA = result4;
    VAR residual;
    RUN;
```