Cost of Healthcare Insurance

Yin T. Ho

Department of Statistics

University of Illinois at Chicago

STAT 481 Applied Statistical Methods II Midterm Project

Dr. Jennifer Pajda-De La O

April 1, 2019

Contents

1 Problem Setup **2**

2 Data Description **2**
2.1 Response Variable 2
2.2 Explanatory Variable 3

3 Multiple Linear Regression Analysis **4**
3.1 The Default Model 5
3.1.1 Multicollinearity Check 5
3.1.2 Assumptions Check 5
3.1.3 BoxCox Transformation 7
3.2 The Log Linear Model of Cost 8
3.2.1 Assumptions Check 8
3.2.2 General Information 10
3.3 Forward Selection 11

4 Conclusion **12**

A SAS Code **13**

1 Problem Setup

A health insurance company collected information from its subscribers who ho have filed a claim for ischemic (coronary) heart disease. This statistical approach is used to predict the total cost of services as a function of age, gender, drugs, emergency, comorbidities, and duration.

2 Data Description

This data is collected from January 1, 1998 to December 31, 1999 without any missing values. It has 383 observations.

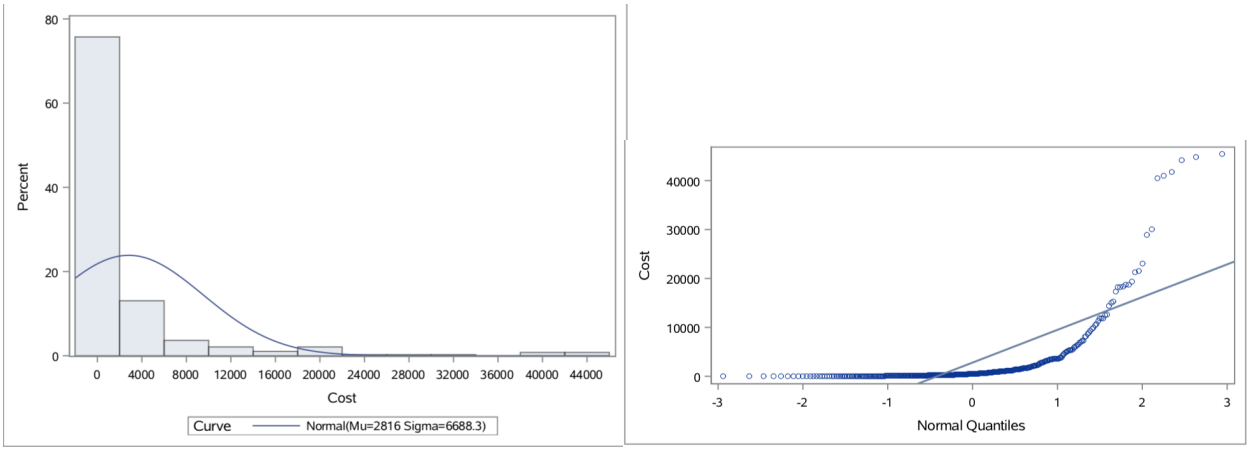
2.1 Response Variable

The variable we want to predict is total cost of claims by subscriber.

Table 2.1.1: Summary Statistics for Cost

	N	Mean	Std Deviation	Median	Min	Max	Skewness	Kurtosis
Cost	383	2815.98851	6688.32247	528.6	10.60	45447.70	4.27061796	20.5824039

The total cost had a maximum of $45447.70 and a minimum of $10.60. The median of the cost is $528.6 and the mean is $2815.99. The cost series is right skewed. According to Figure 2.1.1 and Table 2.1.2 below, the histogram, Q-Q plot, and Shapiro-Wilk test show that the cost is not normally distributed.



(a) Histogram **(b)** Q-Q Plot

Figure 2.1.1: Normality Tests for Cost

Table 2.1.2: Tests for Normality of Cost

Test	Statistic		p Value	
Shapiro-Wilk	W	0.4389	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.337445	Pr > D	<0.0100
Cramer-von Mises	W-Sq	15.22638	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	75.18602	Pr > A-Sq	<0.0050

2.2 Explanatory Variable

Here are six explanatory variable is collected:

1. **Age**: Age of subscriber (years)
2. **Gender**: Gender of subscriber: 1 presents males, and 0 represents females
3. **Drugs**: Number of tracked drugs prescribed
4. **Emergency**: Number of emergency room visits
5. **Comorbidities**: Number of other diseases that the subscriber had during period
6. **Duration**: Number of days of duration of treatment condition

The six variables are summarized in Table 2.2.1 and Figure 2.2.1.

Table 2.2.1: Summary Statistics for Each Explanatory Variable

	N	Mean	Std Deviation	Median	Min	Max	Skewness	Kurtosis	Sum
Age	383	58.5091384	6.5629689	59	39	70	-0.6574969	0.06500731	22409
Gender	383	0.2245431	0.4178269	0	0	1	1.32544313	-0.2445047	86
Drugs	383	0.4151436	0.9748032	0	0	6	2.90773004	9.41211977	159
Emergency	383	3.3002611	2.3009615	3	0	12	0.97051408	0.95770145	1264
Comorbidities	383	3.5013055	5.3106998	1	0	30	2.17794552	5.0966346	1341
Duration	383	160.5639687	122.5815273	150	0	359	0.089628	-1.4610687	61496

The youngest subscriber is 39 years old and the oldest subscriber is at the age of 70. The average age of the subscribers are 58.5091384 years old, which approximates to the median of 59 years old. The age series is weakly left skewed, so there are quite a few middle- and old-aged subscribers in this sample.

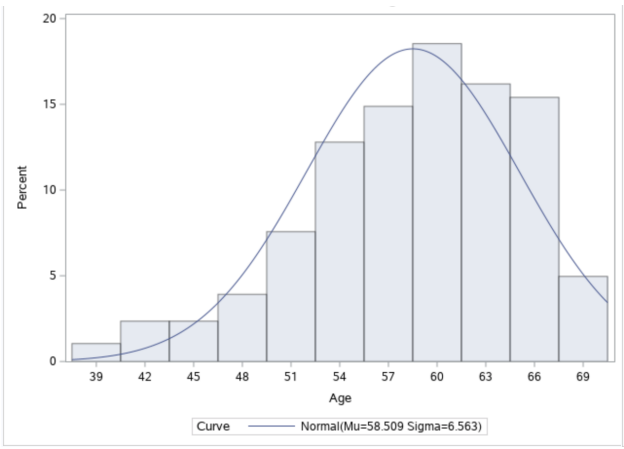
Gender is an indicator variable. We use 1 represents males and 0 represents females. There are 86 male subscribers and 297 female subscribers. The gender series is also moderately right skewed. Thus, most of the subscribers are female.

The number of tracked drugs prescribed has a maximum of 6 and a minimum of 0, with a median of 0 and a mean of 0.4151436. The drugs series is strongly right skewed. Therefore, the majority of the subscribers had less prescribed drugs.

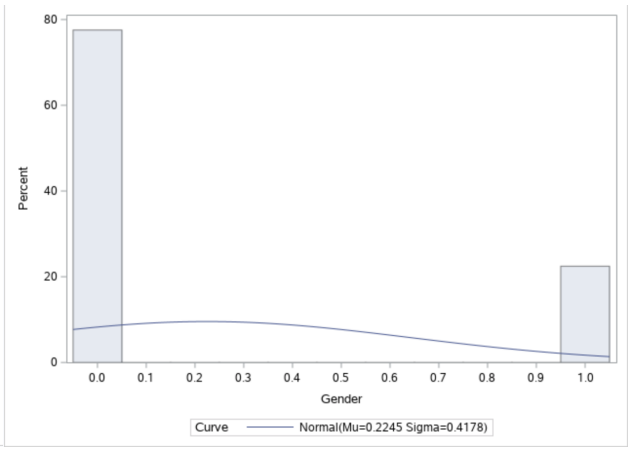
The maximum number of emergency room visits is 12 with a minimum of 0. The subscribers visited emergency room about 3 to 4 times on average, which is also close to a median of 3. Thus, the number of emergency room visits has a weak right skewness.

The subscribers had up to 30 diseases that were not ischemic (coronary) heart disease, a minimum of zero other diseases, and a median of one other disease. The average of the other diseases that they had is about 3 to 4 diseases, which leads to a strong right skewness in comorbidities.

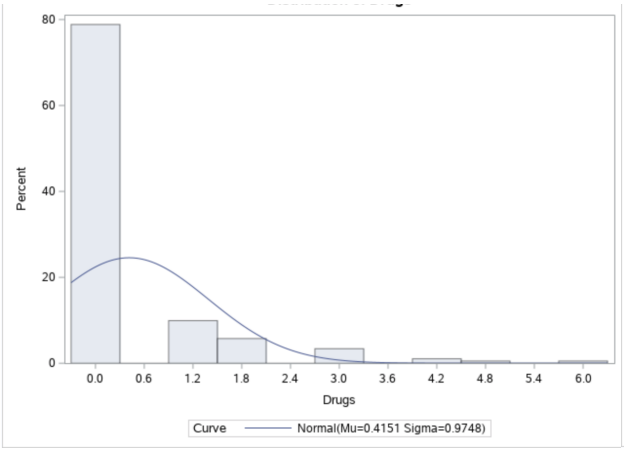
The subscribers had treatment conditions from zero to 359 days. The average of duration of treatment condition is about 160.56 days with a mean of 150. The duration series is a weakly right skewed.



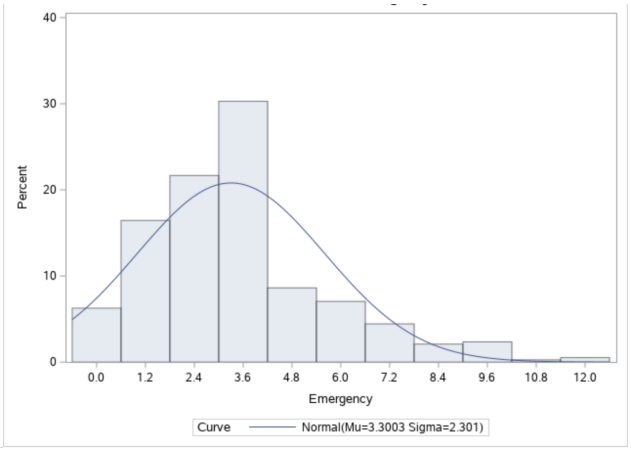
(a) Age



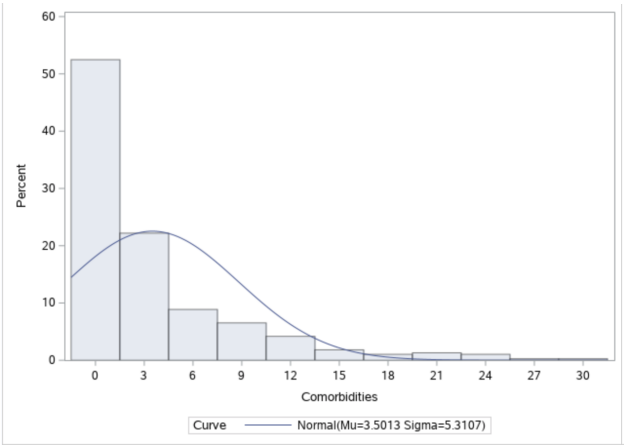
(b) Gender



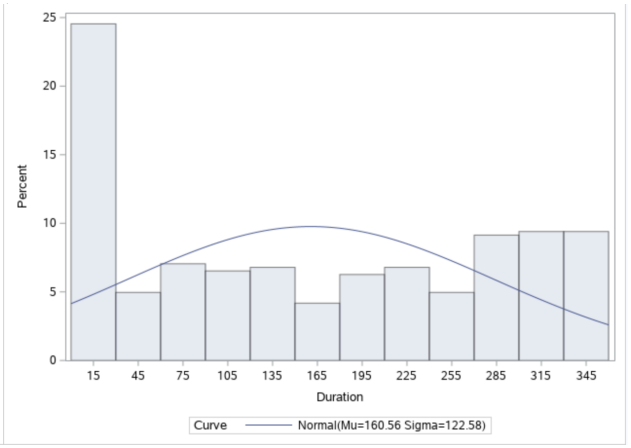
(c) Drugs



(d) Emergency



(e) Comorbidities



(f) Duration

Figure 2.2.1: Histograms for Each Explanatory Variable

3 Multiple Linear Regression Analysis

The relationship between each explanatory variable and *Cost* are studied in multiple linear regressions through data transformation and elimination of the unnecessary variables at 5% significance level (α).

3.1 The Default Model

Model:

$$\text{Cost}_i = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Gender}_i + \beta_3 \text{Drugs}_i + \beta_4 \text{Emergency}_i + \beta_5 \text{Comorbidities}_i + \beta_6 \text{Duration}_i + \varepsilon_i$$

where $i = 1, 2, \ldots, n$.

3.1.1 Multicollinearity Check

Variance of inflation factors (VIF) of all the explanatory variables in Table 3.1.1 are between 1 and 1.5, so none of the explanatory variables will be removed when VIF $\not>$10. Thus, multicollinearity and potential dependence among residual are not a concern.

Table 3.1.1: Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	5427.74260	2789.03286	1.95	0.0524	.	0	-56.31382	10912
Age	Age	1	-124.94805	47.13758	-2.65	0.0084	0.98009	1.02032	-217.63435	-32.26175
Gender	Gender	1	-1304.82503	737.85256	-1.77	0.0778	0.98689	1.01329	-2755.65953	146.00946
Drugs	Drugs	1	-83.74046	363.23545	-0.23	0.8178	0.74815	1.33663	-797.96788	630.48695
Emergency	Emergency	1	1105.95522	154.48483	7.16	<.0001	0.74235	1.34707	802.19275	1409.71770
Comorbidities	Comorbidities	1	220.07886	69.77925	3.15	0.0017	0.68304	1.46405	82.87239	357.28533
Duration	Duration	1	3.77481	3.03412	1.24	0.2142	0.67808	1.47474	-2.19117	9.74078

3.1.2 Assumptions Check

1. **Linearity**

   All the plots in Figure 3.1.1 have different patterns, such as a fan or horn shape, between residuals and the explanatory variables. Consequently, the expected value of the total cost at each value of the explanatory variables is not a linear function of the independent variables.

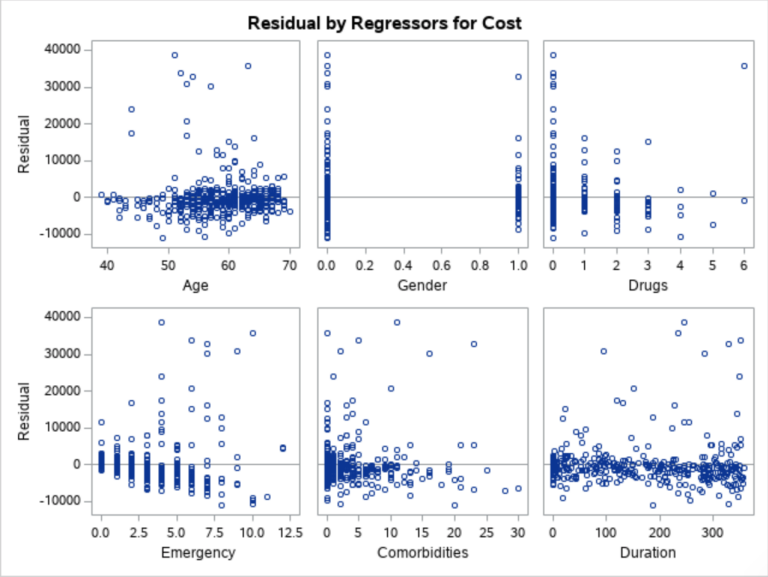


Figure 3.1.1

2. **Independence of Errors**

The lag-1 autocorrelation value in Table 3.1.2 is 0.004 with the sample size (n) of 383. We calculate $2/\sqrt{n}$ to be 0.1022, which is larger than 0.004. Therefore, it is possible the errors could be independent. Moreover, the Durbin-Watson statistic is 1.992, and the associated 2-sided p-value is 0.9342. Since the p-value is greater than the significance level, we fail to reject the null hypothesis that the true autocorrelation is 0. There is evidence that the errors are independent.

Table 3.1.2: Independence Tests for Errors

Durbin-Watson D	1.992
Pr < DW	0.4671
Pr > DW	0.5329
Number of Observations	383
1st Order Autocorrelation	0.004

3. **Residual Normality**

The p-value of the Shapiro-Wilk in Table 3.1.3 is smaller than 0.05, and there is a curve in residual Q-Q plot in Figure 3.1.2, so residuals are not normally distributed.

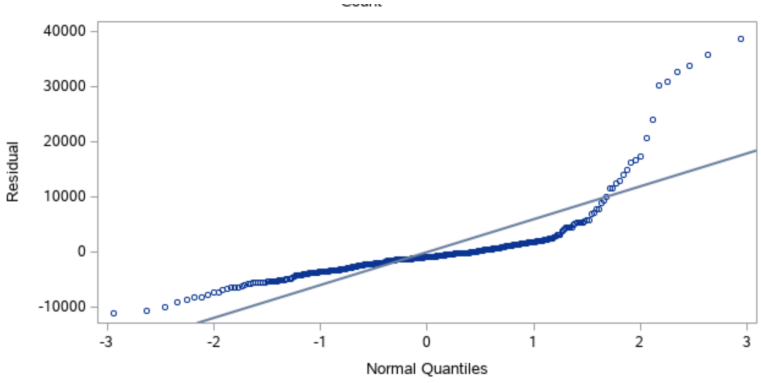


Figure 3.1.2: Q-Q plot for Residuals

Table 3.1.3: Normality Tests for Residuals

Test	Statistic		p Value	
Shapiro-Wilk	**W**	0.669731	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.221981	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	5.809893	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	31.67013	**Pr > A-Sq**	<0.0050

4. **Equal Varainces**

The equal variance condition does not appear to be met because there is a pattern when Residual vs $\widehat{\text{Cost}_i}$ is plotted in Figure 3.1.3.

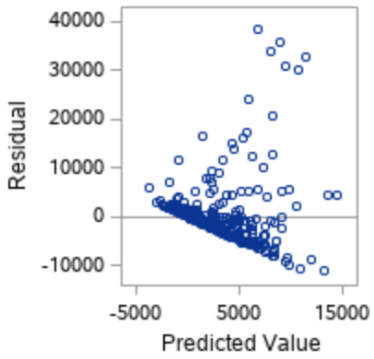


Figure 3.1.3: Residual vs $\widehat{\text{Cost}_i}$

3.1.3 BoxCox Transformation

From previous checks, residual linearity, equal variance, and normality are violated, so we need to transform the dependent variable first and then re-check the new model assumptions after the transformation. If there is something wrong with model assumptions in residual linearity and independence of errors, the the relevant dependent variables have to be transformed as well.

In order to transform the variables, the Boxcox transformation is applied to determine an appropriate and convenient lambda value (λ), which is zero. It is found from Figure 3.1.4 and Table 3.1.4. Hence, the dependent variable, Cost, will be transformed through a log transformation to fix the data.

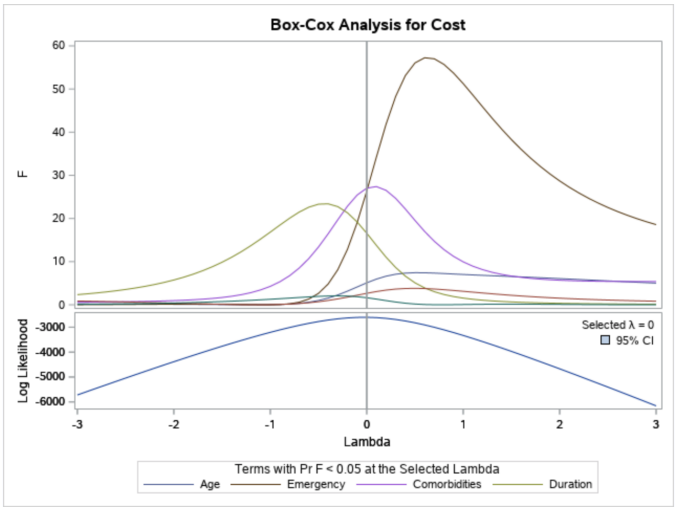


Figure 3.1.4

Table 3.1.4: Model Statement Specification Details

Type	DF	Variable	Description	Value
Dep	1	BoxCox(Cost)	Lambda Used	0
			Lambda	0
			Log Likelihood	-2602.2
			Conv. Lambda	0
			Conv. Lambda LL	-2602.2
			CI Limit	-2604.1
			Alpha	0.05
			Options	Convenient Lambda Used
			Label	Cost
Ind	1	Identity(Age)	Label	Age
Ind	1	Identity(Gender)	Label	Gender
Ind	1	Identity(Drugs)	Label	Drugs
Ind	1	Identity(Emergency)	Label	Emergency
Ind	1	Identity(Comorbidities)	Label	Comorbidities
Ind	1	Identity(Duration)	Label	Duration

3.2 The Log Linear Model of Cost

After transforming the dependent variable, Cost, into $\log_e Cost$, here is the new model:

$$\ln(\text{Cost}_i) = \beta_0 + \beta_1 \cdot \text{Age}_i + \beta_2 \cdot \text{Gender}_i + \beta_3 \cdot \text{Drugs}_i + \beta_4 \cdot \text{Emergency}_i + \beta_5 \cdot \text{Comorbidities}_i + \beta_6 \cdot \text{Duration}_i + \varepsilon_i$$

where $i = 1, 2, \ldots, n$.

Then we start checking whether the characteristics of this model meet our assumptions.

3.2.1 Assumptions Check

1. **Linearity**

   In Figure 3.2.1, the plots of Residual vs Age and Residual vs Duration have no patterns in the graphs between residuals and the explanatory variables. However, the Residual vs Comorbidities plot still has a horn shape, and there are still have some patterns in the rest of the plots. Consequently, the expected value of Cost at each value of the relevant explanatory variables is not a linear function of the independent variables.

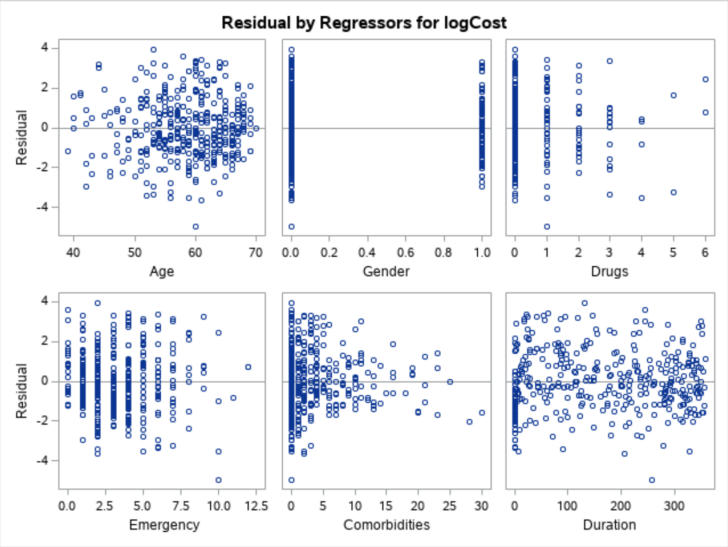


Figure 3.2.1

2. **Independenceof Errors**

The lag-1 autocorrelation value in Table 3.2.1 is -0.015 with the sample size (n) of 383. We calculate $2/\sqrt{n}$ to be 0.1022, which is larger than -0.015. Therefore, it is possible the errors could be independent. Moreover, the Durbin-Watson statistic is 2.024, and the associated 2-sided p-value is 1.1872. Since the p-value is greater than the significance level, we fail to reject the null hypothesis that the true autocorrelation is 0. There is evidence that the errors are independent.

Table 3.2.1: Independence Tests for Errors

Durbin-Watson D	2.024
Pr < DW	0.5936
Pr > DW	0.4064
Number of Observations	383
1st Order Autocorrelation	-0.015

3. **Residual Normality**

The p-value of the Shapiro-Wilk in Table 3.2.2 is greater than 0.05, and the residual Q-Q plot in Figure 3.2.2 looks like a straight line, so residuals are normally distributed.

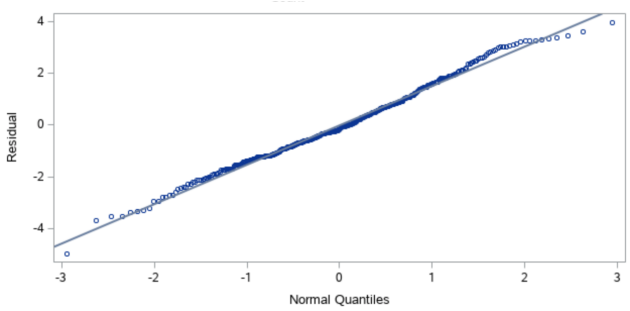


Figure 3.2.2: Q-Q plot for Residuals

Table 3.2.2: Normality Tests for Residuals

Test	Statistic		p Value	
Shapiro-Wilk	W	0.992703	Pr < W	0.0589
Kolmogorov-Smirnov	D	0.046953	Pr > D	0.0394
Cramer-von Mises	W-Sq	0.140829	Pr > W-Sq	0.0329
Anderson-Darling	A-Sq	0.874049	Pr > A-Sq	0.0247

4. **Equal Varainces**

   The equal variance condition appears to be met because there is no pattern when Residual vs $\widehat{\ln \text{Cost}}$ is plotted in Figure 3.2.3.

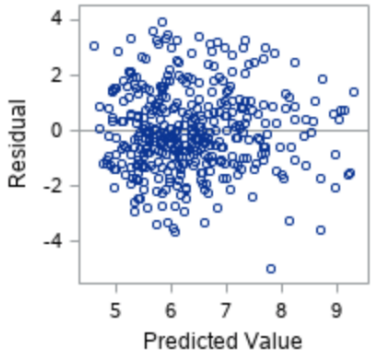


Figure 3.2.3: Residual vs $\ln \hat{C}ost_i$

3.2.2 General Information

Table 3.2.3a shows there is a significant regression relationship in this model because the p-value is smaller than the significance level. However, R^2 in Table 3.2.3b is equal to 29.13%, which means that 29.13% of $\widehat{\ln \text{Cost}}$ is explained by each explanatory variable.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	362.64319	60.44053	25.76	<.0001
Error	376	882.09566	2.34600		
Corrected Total	382	1244.73885			

(a)

Root MSE	1.53167	R-Square	0.2913
Dependent Mean	6.36798	Adj R-Sq	0.2800
Coeff Var	24.05262		

(b)

Table 3.2.3: Analysis of Variance

In Table 3.2.4, Age and Gender have a negative coefficient and the other explanatory variables have a positive coefficient. This means Age and Gender are negatively related to $\ln Cost$, but neither the others do. The p-values of Gender and Drugs are greater than the significant level, so these independent variables do not have a significant impact on $\ln Cost$.

Table 3.2.4: Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	6.47751	0.71365	9.08	<.0001	.	0	5.07426	7.88076
Age	Age	1	-0.02721	0.01206	-2.26	0.0247	0.98009	1.02032	-0.05092	-0.00349
Gender	Gender	1	-0.30789	0.18880	-1.63	0.1038	0.98689	1.01329	-0.67912	0.06335
Drugs	Drugs	1	0.11881	0.09294	1.28	0.2019	0.74815	1.33663	-0.06395	0.30156
Emergency	Emergency	1	0.20289	0.03953	5.13	<.0001	0.74235	1.34707	0.12516	0.28061
Comorbidities	Comorbidities	1	0.09283	0.01785	5.20	<.0001	0.68304	1.46405	0.05772	0.12794
Duration	Duration	1	0.00316	0.00077636	4.07	<.0001	0.67808	1.47474	0.00163	0.00469

3.3 Forward Selection

I choose Forward Section to improve the log model in Section 3.2. Each explanatory variable is be tested whether it is significant with $\widehat{\ln \text{Cost}}$ when α=0.10. If it is significant, the explanatory variable will be added in the simplest model (i.e.the model only involves one explanatory variable) to form a new model. As a result in Table 3.3.1, only one variable, Drugs, is dropped, which does not have a significant impact on $\widehat{\ln \text{Cost}}$ and vice versa. Based on Table 3.3.3, we know the new model will become

$$\widehat{\ln \text{Cost}}_i = 6.44231 - 0.02697 \cdot \text{Age}_i - 0.31268 \cdot \text{Gender}_i + 0.22724 \cdot \text{Emergency}_i + 0.08939 \cdot \text{Comorbidities}_i + 0.00318 \cdot \text{Duration}_i + \varepsilon_i$$

where $i = 1, 2, \ldots, n$.

Table 3.3.1: Summary of Forward Selection

Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Duration	Duration	1	0.1533	0.1533	70.2162	69.01	<.0001
2	Emergency	Emergency	2	0.0716	0.2249	34.2276	35.10	<.0001
3	Comorbidities	Comorbidities	3	0.0495	0.2744	9.9706	25.85	<.0001
4	Age	Age	4	0.0087	0.2831	7.3779	4.56	0.0333
5	Gender	Gender	5	0.0052	0.2883	6.6340	2.74	0.0987

In this model, there is a significant regression relationship because the p-value in Table 3.3.2a is smaller than the significance level. Although the log model is reconstructed, R^2 in the new model becomes smaller, according to Table 3.2.3b and Table 3.3.2b, after Drugs is removed. 28.83% of $\widehat{\ln \text{Cost}}$ is explained by each explanatory variables.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	358.80989	71.76198	30.54	<.0001
Error	377	885.92897	2.34994		
Corrected Total	382	1244.73885			

(a)

Root MSE	1.53295	R-Square	0.2883
Dependent Mean	6.36798	Adj R-Sq	0.2788
Coeff Var	24.07284		

(b)

Table 3.3.2: Analysis of Variance

Moreover, Age and Gender have a negative coefficient, and the other explanatory variables have a positive coefficient, based on Table 3.3.3. This means Age and Gender are negatively related to $\widehat{\ln \text{Cost}}$, and Emergency, Comorbidities, and Duration are positively related to $\widehat{\ln \text{Cost}}$. The p-values of Age and Gender

are greater than the significant level, so only these independent variables do not have a significant impact on $\widehat{\ln \text{Cost}}$.

Table 3.3.3: Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	6.44231	0.71372	9.03	<.0001	.	0	5.03894	7.84568
Age	Age	1	-0.02697	0.01207	-2.23	0.0260	0.98032	1.02007	-0.05070	-0.00324
Gender	Gender	1	-0.31268	0.18892	-1.66	0.0987	0.98728	1.01288	-0.68416	0.05879
Emergency	Emergency	1	0.22724	0.03466	6.56	<.0001	0.96695	1.03418	0.15908	0.29540
Comorbidities	Comorbidities	1	0.08939	0.01767	5.06	<.0001	0.69886	1.43090	0.05466	0.12413
Duration	Duration	1	0.00318	0.00077685	4.10	<.0001	0.67838	1.47409	0.00165	0.00471

4 Conclusion

In this project, the data is being improved through log transformation and forward section since R^2 is increasing around 8% after the default model is reconstructed. All the explanatory variables except Gender have a significant impact on $\widehat{\ln \text{Cost}}$. If Age and Gender increase by one unit, $\widehat{\ln \text{Cost}}$ will decrease by the corresponding parameter estimates in Table 3.3.3, holding the other explanatory variable fixed. Similarly, if Emergency, Comorbidities, and Duration increase by one unit, $\widehat{\ln \text{Cost}}$ will increase by the corresponding parameter estimates ceteris paribus.

A SAS Code

```
FILENAME REFFILE '/folders/myfolders/Project_1/P1_Dataset1F(1).XLSX';
PROC IMPORT DATAFILE=REFFILE
        DBMS=XLSX
        OUT=TotalCost;
        GETNAMES=YES;
RUN;

PROC CONTENTS DATA=TotalCost; RUN;

PROC MEANS DATA=TotalCost NMISS N; RUN;

title "Summary Statistics for Cost";
PROC MEANS DATA=TotalCost;
VAR Cost;
RUN;

** boxplots, qqplots, histograms by proc univariate **;
PROC UNIVARIATE DATA=TotalCost Normal Plot;
        Var Cost;
        qqplot Cost;
        histogram Cost / normal;
RUN;

title "Summary Statistics for Each Explanatory Variable";
PROC MEANS DATA=TotalCost (DROP=ID Cost) ;
RUN;

PROC MEANS DATA=TotalCost (DROP=ID Cost) SUM;
RUN;

** boxplots, qqplots, histograms by proc univariate **;
PROC UNIVARIATE DATA=TotalCost Normal Plot;
        Var Age Gender Drugs Emergency Comorbidities Duration;
        qqplot Age Gender Drugs Emergency Comorbidities Duration;
        histogram Age Gender Drugs Emergency Comorbidities Duration / normal;
RUN;


/* Full Model with All Variables */
PROC REG DATA=TotalCost;
MODEL Cost = Age Gender Drugs Emergency Comorbidities Duration / dwprob clb corrb tol vif
OUTPUT OUT = result1 residual = residual;
TITLE 'Full_Model';
RUN;

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

/* Transformation on Y required. Use Box-Cox Transformation. */
/* Note: Box-Cox only works for independent variable(s) */
PROC TRANSREG DATA=TotalCost DETAIL;
```

```
MODEL BOXCOX(Cost / convenient lambda = -3 to 3 by 0.1)
        = identity(Age Gender Drugs Emergency Comorbidities Duration);
TITLE 'Boxcox Transformation';
RUN;

/* Perform a transformation on Y */
DATA Logy;
SET TotalCost;
logCost = log(Cost);
RUN;

/* Full Model with All Variables After Transformation */
PROC REG DATA=Logy;
MODEL logCost = Age Gender Drugs Emergency Comorbidities Duration /dwprob clb corrb tol vi
OUTPUT OUT = result2 residual = residual;
TITLE 'Full Model After Transformation';
RUN;

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

/* Run FORWARD Selection on Model - with logy */
PROC REG DATA=Logy;
MODEL logCost = Age Gender Drugs Emergency Comorbidities Duration
        / selection = forward clb corrb tol vif collin CP SLENTRY = 0.10;
TITLE 'FORWARD SELECTION';
OUTPUT OUT = result5 residual = residual;
RUN;

PROC UNIVARIATE NORMAL PLOT DATA = result5;
        VAR residual;
RUN;
```