Cardiac Insurance claims as a function of treatment factors and patient health

Abhirupa Sen
Haritha Peyyeti

Introduction

Most coronary diseases require lifelong monitoring and treatment. The cost of treatment is quite high and it depends on various factors. One key issue is to determine the factors that are driving up the costs. Insurance companies believe that patients with similar factors have similar risks, which tends to have an association with total costs claimed. Claims received by insurance companies can be dependent on quite some factors like age, gender, pre-existing medical conditions, severity of the disease, geographic location of the subscriber etc. Apart from these factors, one's general health also affects the treatment procedure and in turn the cost incurred due to the treatment. Each insurance company uses historic data and extensive research to understand the various factors that might help estimate the cost of claims.

The current analysis is on the data that was collected by a health insurance company regarding the total cost for services provided to 788 of its subscribers who had made claims resulting from ischemic (coronary) heart disease. In this study, we are interested in predicting the total cost of claims by a subscriber based on the medical history and nature of various services he/she receives.

Data Sample

The data set provides information on total cost of services provided, identification number and 8 other variables for each subscriber. The variables in the data set are

- Identification Number of the subscriber
- Total cost of claims by subscriber (in dollars). Costs of claims are widely spread over a range of $1.6 to $52,000. 75% of the subscribers have total costs claimed less than $2000.
- Age of the patient/subscriber (in years). Subscriber's age ranges from 24 to 70, with the average age being 59.
- Gender of the subscriber (0 for Male and 1 for Female). 77% of the subscribers in the data collected are males and the rest 23% are females.
- Interventions: The total number of treatment or procedures carried out for the patient. Every diagnostic or examination carried out collectively defines this variable.
- Drugs – The number of drugs that were prescribed in the course of the treatment.
- Number of emergency room visits by the patient during the course of the treatment. This variable can explain the severity of the illness of the patient. Higher the severity of the illness, it is more likely that the patient will have more number of emergency room visits.
- Number of other complications that arose during the heart disease treatment. Many a times, especially for older patients heart ailments are associated with other complications like hypertension and diabetes that arise during the treatments.
- Co-morbidities: This is the number of other diseases that the subscribers already had during the treatment period. Again, older patients are more likely to have existing health problems like hypertension, diabetes, or neurological ailments which, in fact, leads to cardiac ailment.
- Duration of the treatment (in number of days).

Building a model of the total cost of claims as a function of the above factors would be helpful for the insurance company to analyze the cumulative effect of multiple predictors on the cost incurred and the company could use this information to set the premiums for its future subscribers on the basis of their age, gender and health status.

Initial Analysis Steps

Scatter plot

The first step of data analysis, is checking the scatter plot matrix of all the available predictors to see if there is any clear relationship between any of the predictors and the response variable and also any correlation among the predictors. The second row of the scatter plot matrix (Figure1) depicts the plot of cost against all other predictors. Non linearity in the relationship was evident between cost and age, drugs and co morbidity. However, no clear correlation was evident between the predictors.

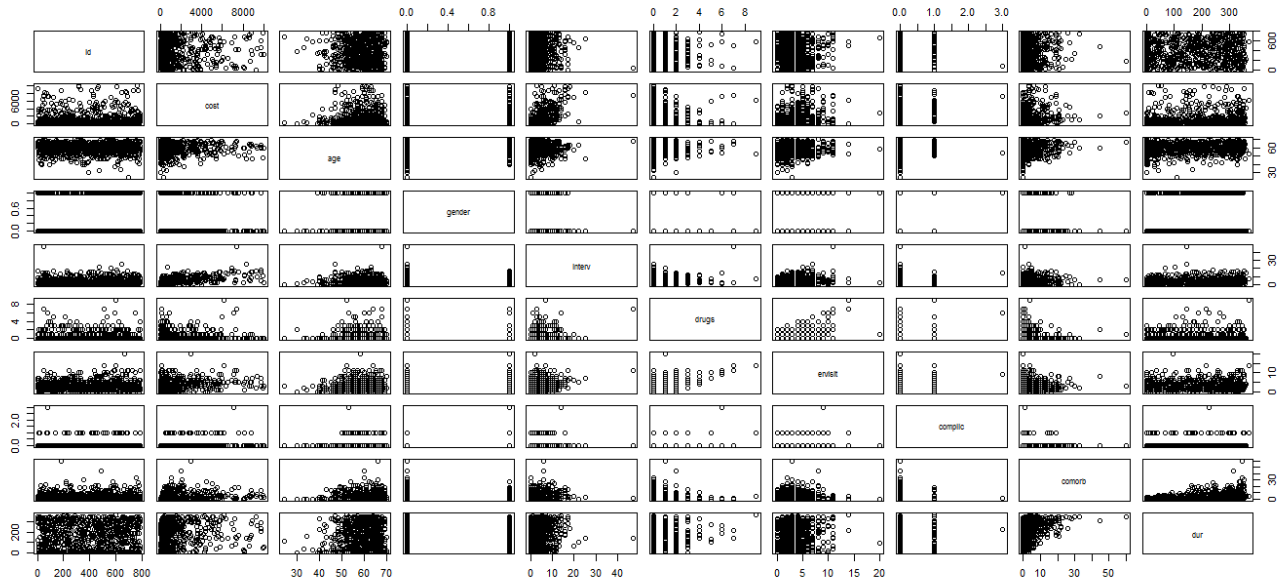

Figure 1: Scatter plot matrix of predictors and response

Sample size

The data set consists of a wide range of claims which are not evenly distributed over the range. 64% of the subscribers had total costs claimed below $1000 while 93% of the subscribers had below $10,000. On comparison, the distribution of each predictor had 2 groups: Subscribers whose total cost claimed is less than or equal to $1,000 and those with total costs above $1,000. The box-plot distribution of the predictors for the two groups is shown in Figure 2. It is interesting to note from the below plot that the distribution of age, number of complications and number of drugs looks very similar for the two groups, while the distributions are different across the two groups for other predictors like number of days of treatment and number of emergency room visits. This gives us a hint that the latter set of predictors can help predict if the costs claimed are going to be high or low. From the analysis of the full data set it was observed that dropping various data points would improve the model. However, there was no abnormality in these points apart from the fact that the cost was much higher than the average. The non-even spread of cost makes the data highly skewed and also locating the outliers becomes subjective. Hence, the subset of the original data set where cost of claims is less than $10,000 was chosen as the sample data. Discarding all the data points above $10,000 would mean loss of only 7% of the original sample. The final size of the data sample used for analysis is 731 which includes all claims less than $10,000.

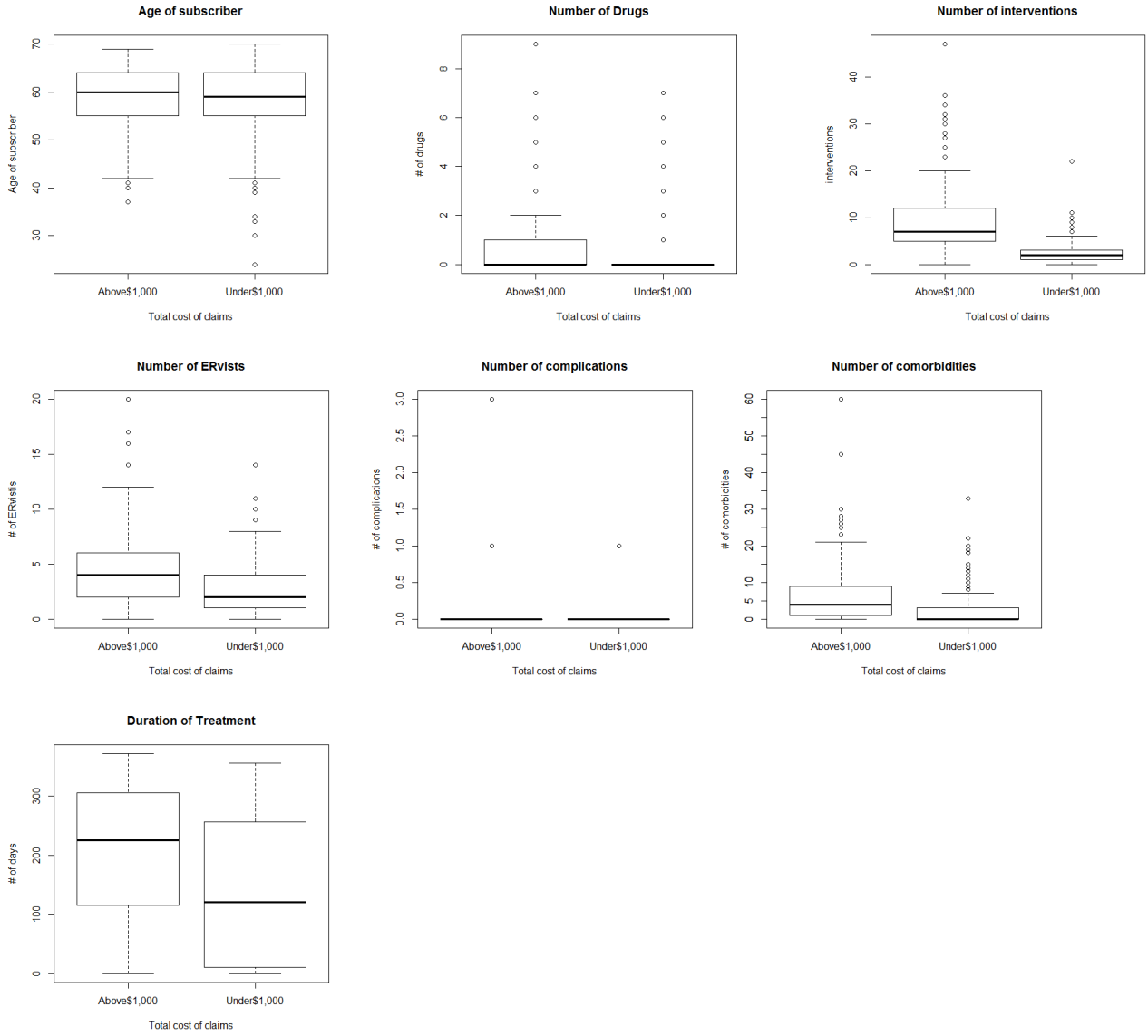

Figure 2: Box plot distribution of predictors for cost of claims above and below $1,000

Linear model with all predictors

The next step was to fit a multiple linear regression model and look at the validity of the model first. The first full model was just a linear fit between cost and all other predictors with no transformation.

Figure 3 shows the Normal QQ plot and residual plot. It is clear from the Normal QQ plot that the residuals are far from the normal distribution. Also, the outward funnel shape of the plotted residuals suggests that the variance of the residuals is not constant. This situation calls for a transformation in the response which is total cost in dollars. Several transformations like log10, ln, square-root, inverse, were tried, but none of the transformations satisfied the assumption of normality and constant variance of the residuals. So a Box-Cox was called on the linear model.

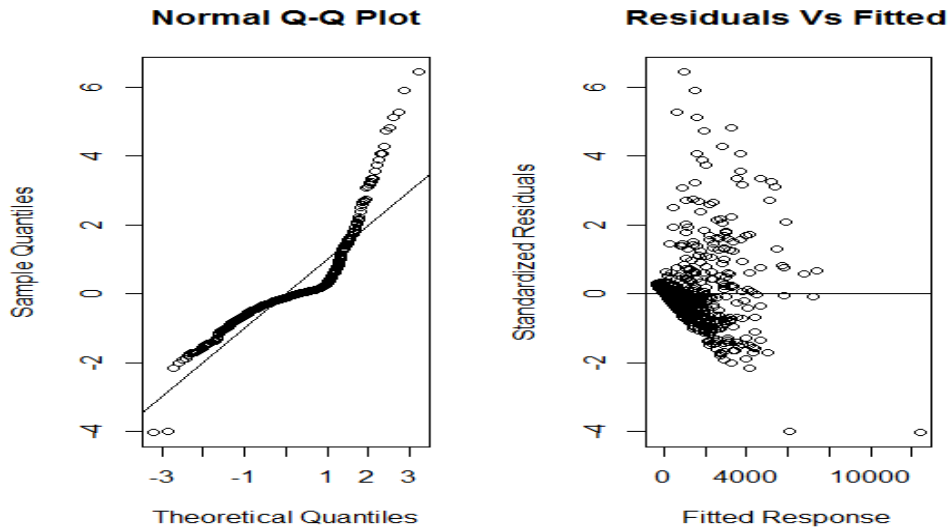

Figure 3: Normal plot and Residual plot of the model w/o transformations

Box-Cox Transformation

The confidence interval of λ as suggested by the Box-Cox transformation did not include 0 which calls for a log transformation or 0.5. The best λ as suggested by the function is λ = 0.1

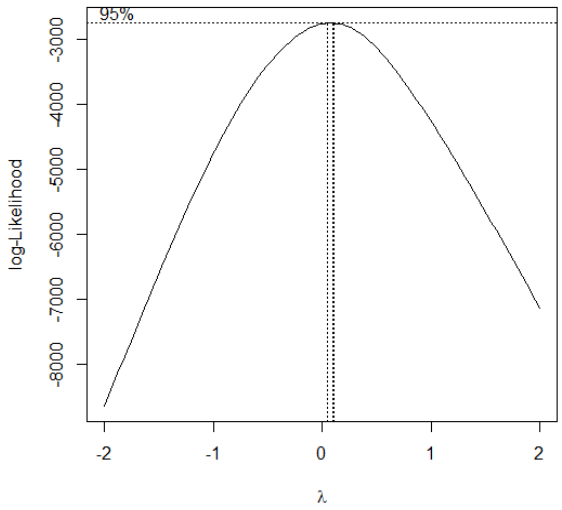

Figure 4: Box Cox plot

So the transformation applied on the response variable is Cost' = $(Cost)^{0.1}$

Residual Analysis of the fit with transformation

Figure 5 has the residual plots of the multiple linear model with transformed cost as the function of the other predictors .

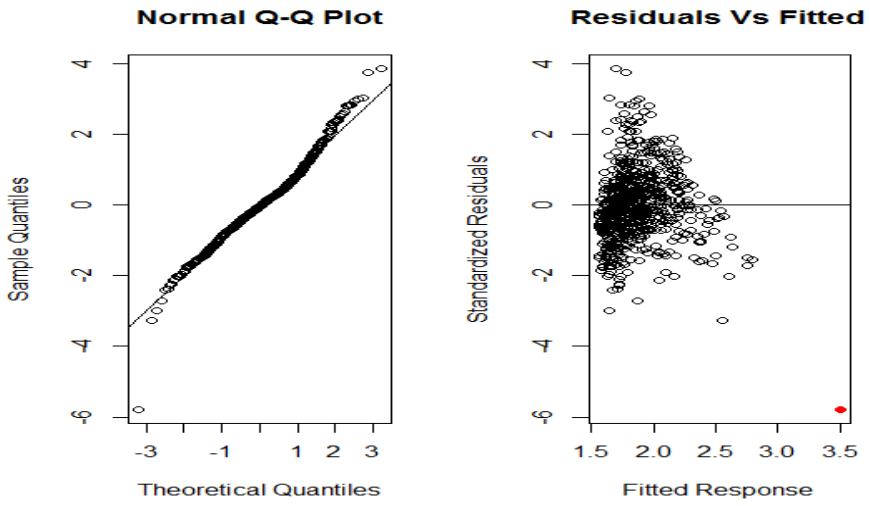

Figure 5 : Normal plot and Residual plot after transforming response variable

The Normal QQ plot suggests that the distribution of the residuals is fairly normal. Also when the residuals are plotted against the fitted response the stabilization of their variance is obvious. However, the 42nd observation (in red) looks like an outlier which will be considered in the analysis of outliers.

Notice that the transformation raised the Multiple R-squared value from 0.477 to 0.559.
It has raised Adjusted R-Squared to 0.5545 from 0.4712. There has been a drastic reduction in the SS Residuals. From 1,415,556,916, as in the initial model without any transformation, the SS Residual has gone down to 29.6285 when the transformation was applied to the response variable. Backed by all these results, the transformation Cost^0.1 was finalized for the model before getting into further analysis.

Further Statistical Analysis: Variable Selection

Now that the transformation of the response variable is done, a model with best choice of predictor is searched for. The method of Forward, Backward and Step-wise selections were taken to see the models advocated by each of these methods.

Exhaustive Search
So far, according to the summary of the full model, the p-value of the variable gender (0.813) is the only one that is insignificant. All other predictors are significant.
Table 1 gives the top three contenders with 7 predictors in the model.

Table 1: Three best models from Exhaustive Search

Predictor dropped	Multiple R-squared	Adjusted R-squared	SS Residual	Mallowe's Cp
Gender of patient (gender)	0.56	0.56	29.63	7.06
Age of the patient (age)	0.56	0.55	29.79	10.88
Age, Gender	0.56	0.55	28.92	14.16

Considering all the four statistics, the model with Mallowe's Cp = 7.06 is the best considering that this value of C8 is closest to 8 as compared to the other two models.

Forward Selection
Table 2 gives the result of the forward variable search. The α (to enter) = 0.1.

Table 2: Steps of Forward Variable Selection

Predictor Added	p-value
Number of Interventions (interv)	2.20E-016
Co-Morbidity (comorb)	2.20E-016
Duration of Treatment (dur)	3.11E-008
Number of Complications (complic)	6.64E-008
Number of Emergency Room visits (ervisist)	0.07
Number of drugs prescribed (drugs)	0.01
Age of the patient (age)	0.05

The predictor gender with p-value (0.8125) > α (to enter) cannot be added to the model. So the model suggested by the Forward selection method has 7 predictors.
Table 3 gives the summary of the model selected by the Forward selection method.

Table 3: Summary of model selected by Forward Search.

Multiple R-squared	Adjusted R-Squared	SS Residuals	Mallowe's Cp
0.56	0.56	29.63	7.7

Backward Selection
Table 4 gives the result of the backward variable search. The α (to drop)= 0.1.

Table 4: Steps of Backward variable Selection

Predictor dropped	p-value
Gender of the patient (gender)	0.812

No further predictors can be dropped from the model. The next predictor Age, with lowest F value 3.919 has p-value (0.048) < α (to drop).
Hence the model suggested by the backward variable selection method has 7 predictors, same as suggested by the forward variable search method.

Step-wise Selection
Table 5 gives the result of the step wise variable selection method

Table 5: Steps of Step-wise variable Selection

Predictor added	p-value
Number of Interventions (interv)	2.20E-016
Co-Morbidity (comorb)	2.20E-016
Duration of Treatment (dur)	3.11E-008
Number of Complications (complic)	6.64E-008
Number of Emergency Room visits (ervisist)	0.07
Number of drugs prescribed (drugs)	0.01
Age of the patient (age)	0.05

So, this is also a model with 7 predictors. The model suggested by Forward, Backward and Step-wise selections is the same.

Multicollinearity check

Now that the three method have suggested a model with seven predictor, a check was made on the assumption of linear independence among these predictors.
Table 6 gives the correlation matrix for the predictors

Table 6: Correlation matrix for the data

	Age	Interventions	Drugs	ER visits	Complications	Co-Morbidity	Duration
Age	1	0.02	0	0.06	-0.03	0.09	0.14
Interventions		1	0.22	0.25	0.16	0.11	0.19
Drugs			1	0.51	0.19	-0.04	0.05
EMR room visits				1	0.17	0.01	0.1
Complications					1	0.02	0.05
Co-morbidity						1	0.5
Duration							1

It is evident from the correlation matrix that this data does not have any serious problem of multicollinearity. The highest correlation coefficient is 0.5 which is between duration and co-morbidity.

Exhaustive Search

Based on the 7 predictors suggested by the above methods, an exhaustive search was performed to see if there is possibly a smaller model that is competitive. Table 7 gives the list of subset models. The model highlighted in blue in the simplest model with equally good R-squared and MSE values as the model suggested by Forward, Backward and Step-wise methods, even though it has high C_p statistic. The correlation between the predictors co-morbidity and duration is 0.5, which is not high enough to warrant exclusion of either of them based on multicollinearity.
The final model selected: Cost' ~ (Interventions + Co-morbidities + Duration of treatment) where
Cost' = $(Cost)^{0.1}$

Table 7: Exhaustive search based on 7 predictors

# Predictors	# Parameters	Age	Drugs	Intervention	Ervisits	Complicat	Comorbid	Duration	SSRes	R2	AdjR2	MSE	Cp
1	2	0	0	1	0	0	0	0	37.4	0.43	0.43	0.05	214
1	2	0	0	0	0	0	0	1	55.3	0.15	0.15	0.08	667
1	2	0	0	0	0	0	1	0	56.5	0.13	0.13	0.08	697
2	3	0	0	1	0	0	1	0	31.8	0.51	0.51	0.04	75
2	3	0	0	1	0	0	0	1	32.7	0.50	0.50	0.04	99
2	3	0	0	1	0	1	0	0	36.1	0.45	0.44	0.05	185
3	4	0	0	1	0	0	1	1	30.5	0.53	0.53	0.04	44
3	4	0	0	1	0	1	1	0	30.5	0.53	0.53	0.04	46
3	4	0	0	1	1	0	1	0	31.4	0.52	0.52	0.04	69
4	5	0	0	1	0	1	1	1	29.3	0.55	0.55	0.04	16
4	5	0	0	1	1	0	1	1	30.2	0.54	0.53	0.04	40
4	5	1	0	1	0	0	1	1	30.3	0.53	0.53	0.04	42
5	6	0	0	1	1	1	1	1	29.1	0.55	0.55	0.04	15
5	6	1	0	1	0	1	1	1	29.1	0.55	0.55	0.04	15
5	6	0	1	1	0	1	1	1	29.2	0.55	0.55	0.04	16
6	7	0	1	1	1	1	1	1	28.9	0.56	0.55	0.04	10
6	7	1	0	1	1	1	1	1	29.0	0.55	0.55	0.04	13
6	7	1	1	1	0	1	1	1	29.0	0.55	0.55	0.04	15
7	8	1	1	1	1	1	1	1	28.7	0.56	0.55	0.04	8

Table 8 shows that all the predictors in the final model are significant and their estimates have reasonably low standard errors.

Table 8: Summary of parameter estimates

	Estimate	Std. Error	t value	p-value	
(Intercept)	1.586	0.013600	116.607	< 2e-16	***
Comorbidity	0.011	0.001479	7.306	7.26E-13	***
Intervention	0.042	0.001834	23.11	< 2e-16	***
Duration	0.0004	0.000073	5.604	2.98E-08	***

Analysis of the outliers

Analysis of outliers starts with analysis of the residuals as well as the hat matrix for the data. The final model has three predictors and 4 parameters. The residual plots for the model indicates the obvious outliers with large residuals. Both the standardized and studentized residuals when plotted against the fitted responses, one of the data points had very large residual. Four data points with high studentized residuals have been taken up. With number of data points being 731 is would be reasonable to assume that the standardized and studentized residuals are approximately similar. However, hence forth, the analysis considers the studentized residuals which are more close to normality with 0 mean and unit variance, and hence more safe in locating outliers. Figure 6 is the plot of the studentized residuals for the model against the fitted response.

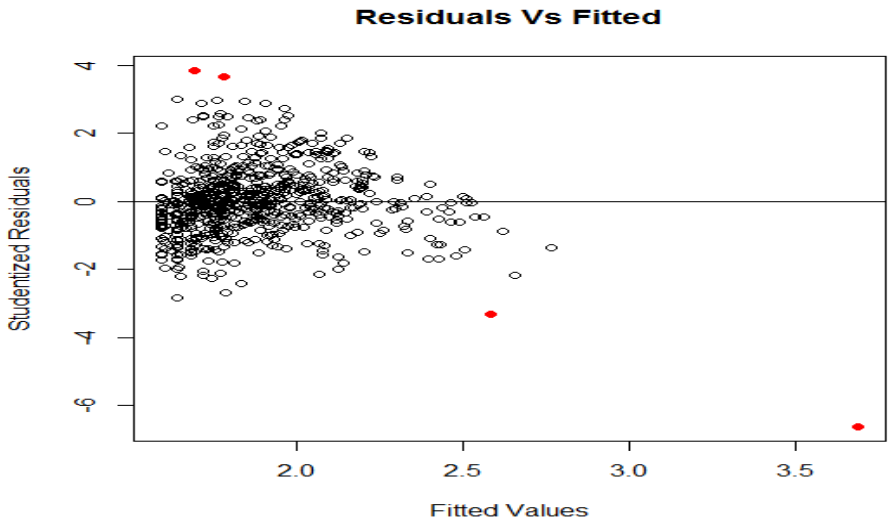

Figure 6: Plot of Studentized residuals Vs Fitted Response

The four points in red are the possible outliers with highest absolute studentized residuals. Though the the 42nd data point at the bottom corner looks like a potential outlier, and the other three fall more or less in the data space, the influence of all these four data points on the model is studied.

Table 9 gives the residuals for these four points, their standardized, studentized and PRESS residuals.

Table 9: Residuals of four potential outliers

Data point	Residual	Standardized Residual (di)	Studentized Residual (ri)	PRESS Residual (e(i))
42nd	-1.23	-6.43	-6.62	-1.45
67th	0.79	3.8	3.84	0.79
294th	0.76	3.65	3.68	0.76
430th	-0.68	-3.3	-3.32	-0.69

The Press residuals of none of the observations are very different from the actual residuals. Also, the studentized residuals are close to the standardized residuals. So, it cannot be said that any of these data points are outliers in terms of this analysis.

The Press residuals also kind of determines whether the points have high leverage. In that respect, none of the points seem to have high leverage.

For this data n (number of data points) = 731 and p (number of parameters) = 4. Hence $2p << n$ so that the cut-off for leverage points is $2p/n = 8/731 = 0.011$. The hii values of all these four points are given in Table 10.

Table 10 :Leverage measure of the four suspected outliers

Data points	hii values	Conclusion
42	0.15	A high leverage point
67	.003	Not leverage points
294	.003	Not leverage points
430	0.03	Leverage point

The PRESS statistic for this model = 32.41. SS Total for the model = 67.2501

$R^2_{prediction} = 1 – PRESS/SS\ Total = 0.5180$

So the model is capable of explaining about 51% of the variability in the predictions made by the model.

Of the four points two does not even fall into the category of leverage points. To be influential, a data point also has to be a leverage point. Since data points 67 and 294 do not fall into the category of leverage points, hence, only the points 42 and 430 will be examined as influential points.

Table 11 has the Cook's Distance, DF Fits and the Cov-Ratio for these two points.

The cut off for the Cook's distance is $F0.5,4,727 = 0.84$

The cut-off for DF Fits = $2*sqrt(p/n) = 0.147$.

The cut-off for the Cov-Ratio is (0.983, 1.016)

Table 11: Testing for influential data points

Data points	Cooks Distance	|DF Fits|	Cov Ratio
42	1.85	2.8	0.94
430	0.08	0.58	0.98

Considering the DF Fits both the points are influential. The Cook's distance suggests that only point 42 is influential. Though slightly, the Cov-Ratio of both the points are outside the valid cut off. Also, the Cov-Ratio for both the points are below 1, which implies that inclusion of these points in the data degrades precision of the model.

Since both the points are potentially influential, how they influence the estimation of the model parameters, is looked into. Table 12 contains the DF Betas for all the 4 parameters of the model, for both the points. The cut off for the DF Betas = $2/sqrt(n) = 0.07397268$

Table 12: DF Betas for the two influential data points

Data Points	Intercept	DF Betas comorb	DF Betas interv	DF Betas dur
42	0.88	0.17	-2.78	0.41
430	0.1	0.06	-0.55	0.112

For both the points the DF Betas are above the cut-off points. This implies that they do influence the estimate of the model parameters.

Finally, Table 13 compares the Standard error of the parameter estimates of the model, with all data points, one without data point 42, one without data point 430 and one without both the data points 42 and 430.

Table 13 : Comparative measure of SE(betas)

	Model with all data points	Model without data point 42	Model without data point 430	Model without data points 42 and 430
Intercept	1.38E-002	1.36E-002	1.37E-002	1.36E-002
Co -morbidity	1.50E-003	1.46E-003	1.49E-003	1.46E-003
Interventions	1.86E-003	1.97E-003	1.88E-003	1.97E-003
Duration	7.43E-005	7.23E-005	7.39E-005	7.24E-005

No significant change in the standard error of the estimates could be seen, on removing the data points. Table 13 gives the summary statistics of the model with each of the four data sets.

Table 14 : Summary Statistics of all the four data sets

	Model with all data points	Model without data point 42	Model without data point 430	Model without data points 42 and 430
Multiple R squared	0.53	0.55	0.54	0.55
Adj R squared	0.53	0.55	0.54	0.55
SS Res	31.45	29.66	30.98	29.63

There is some change when the data points 42 or 430 or both are removed from the model. Hence, considering all these results data points 42 and 430 are definitely influential but they are not erroneous data points. Hence, removing them from the data is not recommended.
Hence, both the points were kept in the data set, so that it remained intact with 731 data points.

Model Validation

The final model obtained from the analysis is

$$cost^{0.1} = 1.59 + 0.011*\text{comorb} + 0.043*\text{interv} + 0.0004*\text{dur}$$

Though the correlation matrix in Table 6 confirms that the data is without multicollinearity problem, a final check of the VIFs of the predictors is made as a part of the model validation procedure. Table 15 gives the VIFs of all the three predictors in the model.

Table 15 : VIFs of the predictors

Co-morbidity	Interventions	Duration
1.33	1.04	1.37

From Table 15 it is obvious that none of the VIFs suggest linear correlation between the predictors.

It is important to check if the model can estimate cost with considerable accuracy. The $R^2_{prediction}$ for the model is 0.52, which implies that the model can explain 52% of the variability in predicting new observations. To make sure there is no problem of over fitting, the final data set was randomly split several times into estimation and prediction data sets so that different subsets of data are used for model building. A 30% random split of the data set multiple times was used and it was found that average $R^2_{prediction}$ is negative. The $R^2_{prediction}$ value for the whole data set from is most representative of the actual data. The value from the latter method though random, is strongly influenced by the negative outliers in $R^2_{prediction.}$

Interpretation

With a power transformation of 0.1 the interpretation of the model is no longer straight forward. However, it can be obtained by taking the derivative of the transformed cost with respect to each of the predictors. Suppose cost'0 is the transformed cost of a given patient for a given combination of co-morbidities, interventions and duration. The model has b1 = 0.11, so for this patient, a increase in the number of co-**morbidities** by one (with the other predictors at the same values) would increase his transformed cost ($cost^{0.1}$) by 10*b1*(cost'0 raised to the power of 9) units.
Similar argument hold good for a change with respect to the other predictors.

Conclusion

The data set consisted of a sample of relatively skewed total claims cost and higher percentage of males than females. This could be one of the reasons why gender did not show an association with the cost claimed. Also, various factors have not been considered here that may be related to the total cost claimed for the treatment of heart disease. These factors may include geographic location of the medical center where the disease was diagnosed, severity of the illness at the time of diagnosis, income of the subscriber. It would be interesting to see if there is any association between income of the subscriber and the cost claimed by him/her, which would also give clues to check if it was a fraudulent claim. The current analysis was performed to estimate the cost claimed, the variable of interest from the insurance company point of view. We can also use this data set to understand if the relation between duration of treatment and various predictors like age, number of complications, gender, number of other diseases at the time of diagnosis.

References

Data Source : Applied Linear Statistical Models, 5th edition, by Kutner, Nachtsheim, Neter, and Li.

Notes on the use of data transformations: http://pareonline.net/getvn.asp?v=8&n=6

Further interpretation guide:
http://stats.stackexchange.com/questions/35982/how-to-interpret-regression-coefficients-when-response-was-transformed-by-the-4t