**Project 1 Group 2 report**

**BY**

**Yoshira Ornelas**

**&**

**Meng Lu**

1. Be more careful with your writing: grammar, usage, capitalization, punctuation and spelling.

2. Use the actual name of the measure rather than the variable’s abbreviated name. For example, coma score, rather than scoma.

3. Where is residual scatterplot?

Assessment

\_\_5\_\_ 5% Justification of transformation (or not) of outcome

\_\_4\_\_ 5% Screening of predictors

\_\_37\_\_ 40% Fitting model

\_8\_\_\_ 10% Assess assumptions

\_\_17\_\_ 20% Interpretation of results

\_\_15\_\_ 20% Clarity/conciseness of writing

86

**INTRODUCTION**

The data set used in this project comes from the Study to Understand Prognoses and Preferences for Outcomes and Risk Treatments (SUPPORT). The original study consisted of 2 phases; which goals were to investigate outcomes of extremely ill patients and to develop predicted models. The first phase enrolled 4301 patients from June 1989 to June 1991. Phase two consisted of 4028 patients enrolled from January 1992 to January 1994. All variables in the study were pre-selected. The dataset that we will be using in this project consists of 1000 random patients from the original SUPPORT data set.

**STUDY OBJECTIVES**

The primary object of this project is to develop a regression model relating the set of predictors; age, sex, disease group, number of comorbidities, support comma score, race, mean arterial blood pressure on day 3, heart rate on day 3, temperature, PaO2 ratio, and serum albumin to the outcome of total cost for the SUPPORT study data (Appendix: Table A1 study variables). The model should be one that maximizes the R-square, reduces the residuals, and uses splines or other necessary transformations of the independent and dependent variable.

**STATISTICAL METHODS**

Observations with an outcome of $0.00 for total cost were removed from all analysis; there was only one observation that fit these criteria. For this project the decision was made to not input any data for missing variables.

The original data set consisted on 1000 subjects. One observation was deleted because it had a total cost of $0.00, and thus was not included in the model. The mean total cost was $30,524, the median was $15,119, indicating a left skewed outcome. There were a total of 438 females and 562 females. Table 1a and 1b summarize the variables for the un-transformed variables in the model. The disease group with the most patients was ARF/MOSF w/Sepsis with 391, and most patients were white. A complete summary table is available in the Appendix (Table A2: Summary of variables).

When the totcost outcome is graphed as a histogram and a normal density plot is overlaid its distribution does not appear normal, but is rather highly skewed (Figure 1a). When the totcost was log transformed the histogram improved; it was visually evident that the distribution was now normally distributed (Figure 1b).

Figure 1a: Total cost Figure 1b: Log total cost

 

*Screening of predictors*

The predictors were plotted in a matrix scatterplot to assess their relationship with the outcome and one another. Visually the scatterplot matrix does not reveal any issues with co-linearity (Figure 2: Scatterplot Matrix). A Spearman’s correlation was then used to further assess co-linearity (Table 3: Spearman’s Correlation); the coefficients of the Spearman’s correlation are not strong, but some are significant: specifically heart rate which appears to be correlated with most of the variables except meanbp and scoma. This correlation might indicate that co-linearity might be problematic when including all the variables in the model.

Figure 2: Scatterplot matrix



Table 3: Spearman’s correlation of continuous variables

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **log\_totcst** | **age** | **temp** | **alb** | **pafi** | **meanbp** | **hrt** | **numco** |
| age | -0.19 | 1.00 |  |  |  |  |  |  |
| p-value | 0.00 |  |  |  |  |  |  |  |
| temp | 0.38 | -0.15 | 1.00 |  |  |  |  |  |
| p-value | 0.00 | 0.00 |  |  |  |  |  |  |
| alb | -0.36 | 0.15 | -0.12 | 1.00 |  |  |  |  |
| p-value | 0.00 | 0.00 | 0.00 |  |  |  |  |  |
| pafi | -0.28 | 0.03 | -0.17 | 0.10 | 1.00 |  |  |  |
| p-value | 0.00 | 0.38 | 0.00 | 0.02 |  |  |  |  |
| meanbp | -0.03 | -0.03 | -0.02 | 0.12 | 0.11 | 1.00 |  |  |
| p-value | 0.35 | 0.38 | 0.45 | 0.00 | 0.00 |  |  |  |
| hrt | 0.34 | -0.21 | 0.28 | -0.18 | -0.12 | 0.05 | 1.00 |  |
| p-value | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 |  |  |
| numco | -0.21 | 0.15 | -0.25 | 0.04 | 0.11 | -0.06 | -0.10 | 1.00 |
| p-value | 0.00 | 0.00 | 0.00 | 0.37 | 0.00 | 0.06 | 0.00 |  |
| scoma | 0.19 | -0.02 | 0.17 | -0.12 | -0.02 | -0.09 | 0.01 | -0.10 |
| p-value | 0.00 | 0.58 | 0.00 | 0.00 | 0.59 | 0.01 | 0.70 | 0.00 |

*Fitting model*

While building the model it was decided that we would treat sex, race, and disease group as class variables. Due to the width of the tails of the variables meanbp and hrt were determined to be non-linear and thus cubic splines were generated for each. These cubic splines would correct for the non-linearity by generating 5 cubic splines at the 5, 27.5, 50, 72.5, and 95th percentiles (Figure 4a and 4b Cubic Splines). Originally scoma was treated as a continuous variable but upon further investigation, the researchers’ paper (cite here)??? revealed that they used ranges to categorize the variable as a categorical variable. The variable was thus transformed into a categorical variable with the following classes: normal (0-20), intermediate (21-70), and deep coma (71-100).

Here, in the methods section, is where you should mention how you assessed model assumptions.

**Results**

The final model resulted in the inclusion of 424 subjects; one subject was initially excluded because of a total cost of $0.00; while the rest were dropped by the model because of missing information. A table listing the significance of a variable to the model is listed in table 4. The assumptions of normality, linearity, and constant variance were met and discussed below. Overall the model has a R2=0.48, with an overall significance of p=0.00. Total medical cost was found to be significantly associated with the following variables PaO2/(.01\*FiO2), temperature, serum albumin, and patients with CHF, patience with COPD, Colon cancer, coma, lung cancer, and a coma score of 21-70.

With every increase in the PaO2/(.01\*FiO2) ratio the total cost is reduced by 0.1%, with one increase of temperature total cost is increased by 11.4%, with every increase in serum albumin total cost is decreased by 23.7%. Compared to the disease reference group (ARF/MOSF w/Sepsis) patients with CHF have a 50.4% lower total cost, patients with COPH have a 49.4% lower cost, patients with colon cancer have a 70.8% lower cost, patients in a coma have a 38.2% lower cost, patients with lung cancer have a 54.4% lower cost when compared to the reference group. Patients who reported having a coma score between 21-70 had a 69% higher total cost when compared to the reference group (coma score=0-20).

While not significant, variables included in the model are; age, numco, sex, race, patients with cirrhosis, patients with MOST w/Malig, and coma scores, 71-100. With every increase in age the total cost is reduced by 0.5%, with every increase in numco the total cost is reduced by 5.8%. When compared to females, males have an 8.0% lower total cost. Compared to the racial reference group (asian) blacks have a 2.7% lower total cost, hispanics have a 15.5% lower total cost, those categorized under other have a 178.4% higher total cost, whites have a 19.2% higher total cost. Compared to the disease reference group (ARF/MOSF w/Sepsis) patients with cirrhosis have a 21.3% lower total cost, patients with MOST w/Malig have a 13.1% lower cost. Patients who reported having a coma score between 71-100 had 8.8% higher total cost when compared to the reference group (coma score=0-20).

Mean blood pressure (meanbp) and heart rate (hrt) are variables included in the model as cubic splines because they had a non-linear relationship with the outcome variable. The results in table 4 for these variables are not interpretable and thus a graphical representation is show in the appendix (Figure: spline of hrt and Figure: spline of meanbp).

The final model is:

**xi: regress log\_totcst pafi temp alb age i.dzgroup i.race i.sex i.scoma\_paper numco hrtsp1 hrtsp2 hrtsp3 hrtsp4 meanbpsp1 meanbpsp2 meanbpsp3 meanbpsp4**.

Table 4: Final Model

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Source** | **SS** | **df** | **MS** | **Observations** | = | 424 |
| Model | 257.5 | 27 | 9.5 | **F (27,396)** | = | 13.77 |
| Residual | 274.3 | 396 | 0.7 | **Prob>F** | = | 0.00 |
| Total | 531.9 | 423 | 1.3 | **R2** | = | 0.48 |
|  |  |  |  |  |  |  |
| **log\_totcst** | **Coef** | **t** | **p-value** | **95% Conf** | **Interval** | **% Change** |
| pafi | -0.001 | -3.74 | 0.00 | -0.002 | -0.00 | -0.1 |
| temp | 0.108 | 3.12 | 0.00 | 0.040 | 0.176 | 11.4 |
| alb | -0.270 | -4.24 | 0.00 | -0.395 | -0.145 | -23.7 |
| age | -0.005 | -1.89 | 0.06 | -0.010 | 0.000 | -0.5 |
| ARF/MOSF w/Sepsis | Reference | - | - | - | - | - |
| CHF | -0.701 | -4.01 | 0.00 | -1.044 | -0.357 | -50.4 |
| COPD | -0.682 | -4.65 | 0.00 | -0.971 | -0.394 | -49.4 |
| Cirrhosis | -0.240 | -1.24 | 0.22 | -0.619 | 0.140 | -21.3 |
| Colon Cancer | -1.230 | -2.51 | 0.01 | -2.194 | -0.267 | -70.8 |
| Coma | -0.481 | -2.14 | 0.03 | -0.924 | -0.038 | -38.2 |
| Lung Cancer | -0.786 | -3.4 | 0.00 | -1.240 | -0.332 | -54.4 |
| MOST w/Malig | -0.140 | -0.98 | 0.33 | -0.422 | 0.142 | -13.1 |
| asian | Reference | - | - | - | - | - |
| black | -0.027 | -0.05 | 0.96 | -1.025 | 0.971 | -2.7 |
| hispanic | -0.169 | -0.3 | 0.76 | -1.280 | 0.941 | -15.5 |
| other | 1.024 | 1.62 | 0.11 | -0.218 | 2.266 | 178.4 |
| white | 0.176 | 0.35 | 0.73 | -0.813 | 1.165 | 19.2 |
| female | Reference | - | - | - | - | - |
| male | -0.083 | -0.98 | 0.33 | -0.250 | 0.084 | -8.0 |
| Scoma 0-20 | Reference | - | - | - | - | - |
| Scoma 21-70 | 0.525 | 4.5 | 0.00 | 0.295 | 0.754 | 69.0 |
| Scoma 71-100 | 0.084 | 0.36 | 0.72 | -0.371 | 0.540 | 8.8 |
| numco | -0.060 | -1.83 | 0.07 | -0.124 | 0.005 | -5.8 |
| hrtsp1\* | 0.008 | 1.4 | 0.16 | -0.003 | 0.019 | - |
| hrtsp2 | -0.045 | -1.04 | 0.30 | -0.130 | 0.040 | - |
| hrtsp3 | 0.139 | 1.1 | 0.27 | -0.110 | 0.388 | - |
| hrtsp4 | -0.239 | -1.25 | 0.21 | -0.616 | 0.137 | - |
| meanbpsp1\* | 0.014 | 1.82 | 0.07 | -0.001 | 0.030 | - |
| meanbpsp2 | -0.170 | -2.1 | 0.04 | -0.328 | -0.011 | - |
| meanbpsp3 | 0.544 | 1.83 | 0.07 | -0.041 | 1.128 | - |
| meanbpsp4 | -0.417 | -1.43 | 0.15 | -0.992 | 0.158 | - |
| \_cons | 6.233 | 4.37 | 0.00 | 3.428 | 9.038 | - |

\*Cubic splines cannot be interpreted the same way as other variables and thus are shown graphically in the appendix (Figure: spline of hrt and Figure: spline of meanbp)

*Outliers*

In order to identify outliers a series of different tests were carried out. Residuals and leverage values were used to detect outliers, while Cook’s distance identified influential observations. The outliers in x (Figure 3a: Outliers in x) were identified as having leverage greater than 0.3, these are: 118, 159, 224, 253, 675, 744, 789, and 839. The greatest outlier in x is observation number 789, with the following characteristics; a white female aged 71.8 admitted with ARF/MOSF w/sepsis. The outliers in y (Figure 3b: Outliers in y) were identified as having a residual greater than 2.5, these are: 57, 83, 363, 529, and 697. Influential observations were found to have a Cook’s distance greater than 0.02, these are: 9, 57, and 224 (Figure 3c: Influential Observations). A complete list with characteristics is available in the appendix (Table A5: Outliers and Influential Observations).

**Assessment of assumptions**

The assumption of normality was assessed using a histogram plot, linearity by a normal quantile plot, and constant variance by a Breusch-Pagan/ Cook-Weisberg test.

Linearity

Normal quantile plot



Normality

Histogram plot



*Constant variance*

A Breusch-Pagan/ Cook-Weisberg test to determine constant variance was performed, the p-value=0.42 is greater than p=0.05 indicating that constant variance is met.

**Sensitivity test**

The sensitivity analysis involved removing the outliers and influential observations and re-running the model. Removing these observations changes the model (Model 2); specifically the R2 increases from 0.48 to 0.51. Additionally the following variables are now significant age, Hispanic, black, and white. The differences are listed in the appendix (Table A6: Sensitivity Analysis). Even though the model would be improved by removing these observations, we have decided to not remove them from our final model because we do not want to bias the results. Additionally the interaction between dzgroup and race also resulted in a R2 increases from 0.48 to 0.49, this however was also not included in the model because the interaction itself was not significant.

**Interpretation of results**

The final equation of this predictive mode is:



The model has a R2=0.48, indicating that is able to account for 48% of the variation in total cost. Total cost was found to be significantly associated with the following continuous variables PaO2/(.01\*FiO2) (pafi), temperature, and serum albumin. PaO2/(.01\*FiO2) is a measure of respiratory rate. From our model every increase in pafi results in a 0.1% decrease in total cost. This indicates that those requiring treatment for respiratory failure would have a higher total cost. The increase in 1 degree Celsius in temperature results in 11.4% higher cost, indicating that patients that are feverish would have a higher total cost. Lastly serum albumin is used to evaluate nutritional status, a 1-unit increase in serum albumin results in a 23.7% lower total cost. Overall this dataset indicated that patients requiring prolonged treatment would incur higher total cost. This is most noticeable by our categorical variables. Patients in the reference group (ARF/MOSF w/Sepsis), have a total cost that is significantly higher than patients with CHF, COPD, Cirrhosis, Colon cancer, lung cancer, and coma patients. Patients with a coma score between 21-70 also have a significantly higher total cost than patients with a score below 21.

**Conclusion**

A model consisting of 11 variables; age, sex, disease group, number of comorbidities, support comma score, race, mean arterial blood pressure on day 3, heart rate on day 3, temperature, PaO2 ratio, and serum albumin was developed to predict the total cost of hospitalization. Total medical cost was found to be significantly associated with the following variables PaO2/(.01\*FiO2), temperature, serum albumin, and patients with CHF, patience with COPD, Colon cancer, coma, lung cancer, and a coma score of 21-70. The model accounts for 48% of the variation in total cost, signifying that additional research is needed to account for the remaining variation in total cost.

**Reference**

Knaus, W. A., Harrell, F. E., Lynn, J., Goldman, L., Phillips, R. S., Connors, A. F., ... & Wagner, D. P. (1995). The SUPPORT prognostic model: objective estimates of survival for seriously ill hospitalized adults. Annals of Internal Medicine, 122(3), 191-203.

Vittinghoff, E., Shiboski, S., & McCulloch, C. E. (2005). *Regression methods in biostatistics* (pp. 98-109). New York:: Springer.

**Appendices**

**Table A1: Study variables**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Coding and Meaning** | **Type of variable** |
| Age | Age of participants in the study | Continuous |
| Sex | Gender of participants | Category |
| Dzgroup | Disease group | Category |
| Num.co | Number of comorbidities | Continous |
| Scoma | SUPPORT coma score based on Glasgow D3 | Category |
| Totcst | Total RCC cost | Continuous |
| Race | Race of participant | Category |
| Meanbp | Mean arterial blood pressure on day 3 | Cubic Spline |
| Hrt | Heart rate Day 3 | Cubic Spline |
| Temp | Temperature of participant ton day 3 (Celsius) | Continuous |
| Pafi | PaO2/(.01\*FiO2) day 3 | Continuous |
| Alb | Serum albumin day 3 | Continuous |

**Table A2: Summary of variables**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Obs** | **Min** | **Mean** | **Max** |
| totcst | 895 | 0 | 30524.03 | 390460.50 |
| age | 1000 | 18.04 | 62.49 | 101.85 |
| meanbp | 1000 | 0.00 | 84.94 | 180.00 |
| hrt | 1000 | 0.00 | 97.90 | 300.00 |
| temp | 1000 | 32.50 | 37.08 | 41.20 |
| pafi | 747 | 34.00 | 244.11 | 869.38 |
| alb | 622 | 1.10 | 2.92 | 4.90 |
| scoma | 1000 | 0.00 | 11.75 | 100.00 |
| numco | 1000 | 0 | 1.88 | 7 |

**Table A2: Summary of Categorical Variables**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Freq.** | **Percent** | **Total cost Mean** |
| Sex |  |  |  |
| female | 438 | 43.84 | 32395.9 |
| male | 562 | 56.16 | 29055.7 |
| Dzgroup |  |  |  |
| ARF/MOSF w/Sepsis | 391 | 39.04 | 50939.2 |
| CHF | 143 | 14.31 | 10870.1 |
| COPD | 116 | 11.61 | 17408.3 |
| Cirrhosis | 55 | 5.51 | 20181.1 |
| Colon Cancer | 49 | 4.9 | 8241.2 |
| Coma | 60 | 6.01 | 23015.8 |
| Lung Cancer | 100 | 10.01 | 8508.5 |
| MOSF w/Malig | 86 | 8.61 | 42363.1 |
| Race |  |  |  |
| asian | 9 | 0.91 | 27684.6 |
| black | 156 | 15.69 | 24595.2 |
| hispanic | 36 | 3.62 | 32822.8 |
| other | 12 | 1.21 | 34741.1 |
| white | 781 | 78.57 | 31554.4 |

**Figure: Spline of hrt**

****

****

**Figure: Spline of meanbp**

****

****

**Figure 3a: Outliers in x**

****

**Figure 3b: Outliers in y**

****

**Figure 3c: Influential Observations**

****

**Table A5: Outliers and Influential Observations**

**Outliers in x**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **blev** | **ID** | **age** | **sex** | **dzgroup** | **totcst** | **meanbp** | **hrt** | **temp** | **pafi** | **alb** |
| 0.36 | 118 | 59 | F | ARF/MOSF w/Sepsis | 52945.78 | 57 | 115 | 37.80 | 323.3 | 4.1 |
| 0.37 | 159 | 60 | F | ARF/MOSF | 46810.66 | 143 | 135 | 35.50 | 540.0 | 3.0 |
| 0.38 | 224 | 65 | F | Coma | 5589.57 | 56 | 47 | 38.09 | 393.3 | 2.6 |
| 0.34 | 253 | 67 | M | Colon cancer | 9882.65 | 117 | 110 | 35.70 | 316.3 | 2.7 |
| 0.34 | 675 | 53 | M | Lung cancer | 45421.34 | 142 | 68 | 38.20 | 300.0 | 3.1 |
| 0.35 | 744 | 75 | M | Colon cancer | 5308.98 | 99 | 107 | 36.40 | 228.6 | 2.6 |
| 0.47 | 789 | 72 | F | ARF/MOSF w/Sepsis | 11167.88 | 122 | 300 | 36.30 | 225.0 | 2.0 |
| 0.36 | 840 | 74 | F | Colon cancer | 3107.41 | 70 | 64 | 37.59 | 485.7 | 2.5 |

**Outliers in y**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **bstd** | **ID** | **age** | **sex** | **dzgroup** | **totcst** | **meanbp** | **hrt** | **temp** | **pafi** | **alb** |
| 2.76 | 57 | 51 | M | ARF/MOSF w/Sepsis | 236295.25 | 42 | 118 | 38.90 | 270.0 | 3.60 |
| 2.54 | 83 | 66 | F | ARF/MOSF w/Sepsis | 228884.63 | 50 | 95 | 38.50 | 106.7 | 3.60 |
| 2.73 | 363 | 63 | M | ARF/MOSF w/Sepsis | 338955 | 72 | 120 | 37.80 | 267.5 | 1.80 |
| 2.87 | 529 | 71 | F | ARF/MOSF w/Sepsis | 142914.63 | 78 | 75 | 36.20 | 286.6 | 2.50 |
| 2.77 | 697 | 82 | F | CHF | 41473.844 | 103 | 64 | 36.09 | 281.9 | 4.50 |

**Influential Observations**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cook** | **ID** | **age** | **sex** | **dzgroup** | **totcst** | **meanbp** | **hrt** | **temp** | **pafi** | **alb** |
| 0.02 | 9 | 31.70 | F | Cirrhosis | 101867.94 | 77 | 80 | 37.20 | 490.4 | 2.2 |
| 0.03 | 57 | 50.75 | M | ARF/MOSF w/Sepsis | 236295.25 | 42 | 118 | 38.90 | 270.0 | 3.6 |
| 0.04 | 224 | 64.83 | F | Coma | 5589.5703 | 56 | 47 | 38.09 | 393.3 | 2.6 |

**Table A6: Sensitivity Analysis, comparing Final Model and Model 21**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Obs | 424 |  |  | Obs | 410 |  |
| F (25, 384) | 13.77 |  |  | F (25, 384) | 15.95 |  |
| P>F | 0 |  |  | P>F | 0 |  |
| R-squred | 0.48 |  |  | R-squred | 0.51 |  |
|  |  |  |  |  |  |  |
| **Final Model** |  |  |  | **Model 2** |  |  |
| **log\_totcst** | **Coef** | **t** | **P>t** | **Coef** | **t** | **P>t** |
| pafi | -0.001 | -3.74 | 0.00 | -0.001 | -3.87 | 0.00 |
| temp | 0.108 | 3.12 | 0.00 | 0.098 | 2.90 | 0.00 |
| alb | -0.270 | -4.24 | 0.00 | -0.289 | -4.67 | 0.00 |
| age\* | -0.005 | -1.89 | 0.06 | -0.005 | -2.03 | 0.04 |
| ARF/MOSF w/Sepsis | Reference | - | - | - | - | - |
| CHF | -0.701 | -4.01 | 0.00 | -0.685 | -4.05 | 0.00 |
| COPD | -0.682 | -4.65 | 0.00 | -0.622 | -4.39 | 0.00 |
| Cirrhosis | -0.240 | -1.24 | 0.22 | -0.282 | -1.50 | 0.13 |
| Colon Cancer | -1.230 | -2.51 | 0.01 | Omitted | - | - |
| Coma | -0.481 | -2.14 | 0.03 | -0.323 | -1.44 | 0.15 |
| Lung Cancer | -0.786 | -3.40 | 0.00 | -0.719 | -3.15 | 0.00 |
| MOST w/Malig | -0.140 | -0.98 | 0.33 | -0.111 | -0.81 | 0.42 |
| asian | Reference |  |  |  |  |  |
| black\* | -0.027 | -0.05 | 0.96 | -1.134 | -2.65 | 0.01 |
| hispanic\* | -0.169 | -0.30 | 0.76 | -1.154 | -2.37 | 0.02 |
| other | 1.024 | 1.62 | 0.11 | Omitted | - | - |
| white\* | 0.176 | 0.35 | 0.73 | -0.871 | -2.08 | 0.04 |
| female | Reference | - | - | - | - | - |
| male | -0.083 | -0.98 | 0.33 | -0.062 | -0.75 | 0.45 |
| numco | -0.060 | -1.83 | 0.07 | -0.051 | -1.61 | 0.11 |
| Scoma 0-20 | Reference | - | - | - | - | - |
| Scoma 21-70 | 0.525 | 4.50 | 0.00 | 0.547 | 4.87 | 0.00 |
| Scoma 71-100 | 0.084 | 0.36 | 0.72 | -0.036 | -0.15 | 0.88 |
| hrtsp1 | 0.008 | 1.40 | 0.16 | 0.004 | 0.75 | 0.45 |
| hrtsp2 | -0.045 | -1.04 | 0.30 | -0.025 | -0.59 | 0.56 |
| hrtsp3 | 0.139 | 1.10 | 0.27 | 0.092 | 0.73 | 0.47 |
| hrtsp4 | -0.239 | -1.25 | 0.21 | -0.186 | -0.94 | 0.35 |
| meanbpsp1\* | 0.014 | 1.82 | 0.07 | 0.019 | 2.49 | 0.01 |
| meanbpsp2 | -0.170 | -2.10 | 0.04 | -0.226 | -2.89 | 0.00 |
| meanbpsp3\* | 0.544 | 1.83 | 0.07 | 0.750 | 2.59 | 0.01 |
| meanbpsp4\* | -0.417 | -1.43 | 0.15 | -0.612 | -2.15 | 0.03 |
| \_cons | 6.233 | 4.37 | 0.00 | 7.629 | 5.71 | 0.00 |

1. The difference between the Final model and Model 2 is that model 2 removes the outliers and influential observations.

\* Indicates variables are significant in model 2 when outliers and influential observations are removed

**STATA Do file Code**

\*Summarize variables

summarize totcst age meanbp hrt temp pai alb scoma alb numco

tab sex

tab dzgroup

tab race

\*verify normality of totcst

\*histogram totct

histogram totcst, graphregion(fcolor(white))

(bin=29, start=1162.4287, width=13424.071)

\*logtransform totcst

generate log\_totcst = log(totcst)

\*histogram log total cost

histogram log\_totcst, graphregion(fcolor(white))

\*delete totcst 0

drop if totcst ==0

\*graphing a scatterplot matrix to visually identify multicollinearity

graph matrix log\_totcst age numco scoma meanbp hrt temp pafi alb

\*do a correlation

corr log\_totcst age numco scoma meanbp hrt temp pafi alb, mean

\*Run model with log total cost all other untrasformed

xi: regress log\_totcst pafi temp alb age i.dzgroup i.race i.sex hrt meanbp scoma numco

\*scoma not a continous variable so need to do categorical

xi: regress log\_totcst pafi temp alb age i.dzgroup i.race i.sex hrt meanbp i.scoma numco

\*Great but would exclude people with scoma values not in those groups so need to create new groups

\*create new scomma categories

twoway (scatter log\_totcst scoma)

\*generate scoma categories based on Support paper publsihed

\*score 0-43 is normal 44 to 99 is intermideate and 100+ is deep

generate scoma\_paper 0/43=1

generate scoma\_paper=scoma

recode scoma\_paper 0/43=1

recode scoma\_paper 44/99=2

recode scoma\_paper 100=3

label define scoma\_paper 1 "normal" 2 "Intermediate" 3 "Deep Coma", replace

label values scoma\_paper scoma\_paper

label variable scoma\_paper "scoma\_paper"

xi: regress log\_totcst pafi temp alb age i.dzgroup i.race i.sex i.scoma\_paper numco hrt meanbp

xi: regress log\_totcst pafi temp alb age i.dzgroup i.race i.sex i.scoma numco hrt meanbp

\*seems like improvements were not made so now lets re do scoma with different breaks

drop scoma\_paper

\*recoding scoma to see if improvement can be made based on natural breaks

generate scoma\_paper=scoma

recode scoma\_paper 0/20=1

recode scoma\_paper 21/70=2

recode scoma\_paper 71/100=3

label define scoma\_paper 1 "normal" 2 "Intermediate" 3 "Deep Coma", replace

label values scoma\_paper scoma\_paper

label variable scoma\_paper "scoma\_paper"

\* now run with new scoma

xi: regress log\_totcst pafi temp alb age i.dzgroup i.race i.sex scoma\_paper numco hrt meanbp

\*So decided not to do continues since grouping improved model but not the numbers exactly on paper

\*so redefined based on natural breaks which can be seened when graphed

\*couldnt do classes on regular data because this would exclude people who did not fall in those ranges

twoway (scatter log\_totcst scoma\_paper)

\*creating the spline for heart vairbale

mkspline hrtsp = hrt, cubic displayknots

twoway mspline log\_totcst hrt

xi: regress log\_totcst pafi temp alb age i.dzgroup i.race i.sex i.scoma\_paper numco hrtsp1 hrtsp2 hrtsp3 hrtsp4 meanbp

\*Creating the spline for meanbp variable

mkspline meanbpsp = meanbp, cubic displayknots

twoway mspline log\_totcst meanbp

xi: regress log\_totcst pafi temp alb age i.dzgroup i.race i.sex i.scoma\_paper numco hrtsp1 hrtsp2 hrtsp3 hrtsp4 meanbpsp1 meanbpsp2 meanbpsp3 meanbpsp4

\*Residuals, Linearity and Normality assumtpions

predict resid, residuals

qnorm resid

histogram resid

estat hettest

\*identifying outliers

\*outliers in x

predict blev, leverage

graph box blev, marker(1, mlabel (idnumber))

\*looks like above 0.3 is outlier

list blev idnumber age sex dzgroup totcst meanbp hrt temp pafi alb if blev >0.3

\*forgot to exlucde missing

list blev idnumber age sex dzgroup totcst meanbp hrt temp pafi alb if blev >0.3 & blev!=.

\* now outliers in y

predict bstd, rstudent

graph box bstd, marker(1, mlabel(idnumber))

\*looks like anything above 2.5

list bstd id idnumber age sex dzgroup totcst meanbp hrt temp pafi alb if bstd >2.5 & bstd!=.

\*identify influential observations based on cooks disntace

predict cook, cooksd

graph box cook, marker(1, mlabel(idnumber))

\*looks like anything above 0.02

list cook id idnumber age sex dzgroup totcst meanbp hrt temp pafi alb if cook >0.02 & cook!=.

\* A sensitivity test removing outliers and influential observations

\*Untransformed model first

xi: regress log\_totcst pafi temp alb age i.dzgroup i.race i.sex scoma numco hrt meanbp

\*first Final model again

xi: regress log\_totcst pafi temp alb age i.dzgroup i.race i.sex i.scoma\_paper numco hrtsp1 hrtsp2 hrtsp3 hrtsp4 meanbpsp1 meanbpsp2 meanbpsp3 meanbpsp4

\*removing influential and outliers

xi: regress log\_totcst pafi temp alb age i.dzgroup i.race i.sex i.scoma\_paper numco hrtsp1 hrtsp2 hrtsp3 hrtsp4 meanbpsp1 meanbpsp2 meanbpsp3 meanbpsp4 if bstd <2.5 & blev<0.3 & cook <0.02

\*Improves model but we DO not want to exclude any observations

\*Now interactions

\*sex and race \*Improved

xi: regress log\_totcst pafi temp alb age i.dzgroup i.race\*i.sex i.scoma\_paper numco hrtsp1 hrtsp2 hrtsp3 hrtsp4 meanbpsp1 meanbpsp2 meanbpsp3 meanbpsp4

\*Dzgroup and Race \*\*\*Improved the most

xi: regress log\_totcst pafi temp alb age i.dzgroup\*i.race i.sex i.scoma\_paper numco hrtsp1 hrtsp2 hrtsp3 hrtsp4 meanbpsp1 meanbpsp2 meanbpsp3 meanbpsp4

\*Dzgroup and Sex \*Improved

xi: regress log\_totcst pafi temp alb age i.dzgroup\*i.sex i.race i.scoma\_paper numco hrtsp1 hrtsp2 hrtsp3 hrtsp4 meanbpsp1 meanbpsp2 meanbpsp3 meanbpsp4

\*Reasoning even though the overall model is improved when accounting for interaction between dzgroup and race

\*The interaction did not appear to be significant and thus it was decided that the itneraction term would be excluded

\*The Final model treat dzgroup and race separately with no interaction term

\*Final Model

xi: regress log\_totcst pafi temp alb age i.dzgroup i.race i.sex i.scoma\_paper numco hrtsp1 hrtsp2 hrtsp3 hrtsp4 meanbpsp1 meanbpsp2 meanbpsp3 meanbpsp4