Dominic LaRoche

Karen Gonzalez

Zach Peterson

576C

Project 1

Relationship of Predictive Factors and Total Cost for Severely Ill Hospitalized Adults

**Introduction:**

The SUPPORT (Study to Understand Prognoses Preferences Outcomes and Risks of Treatment) was conducted to measure survival estimates of severely ill hospitalized adults. The purpose of this model is to identify the correlation of specified predictive values on total RCC cost. The specified variables are defined as: age, sex, race, mean blood pressure, dzgroup , number of comorbidities, support coma score, partial pressure of oxygen in arterial blood, and the heart rate, temperature, and albumin serum level taken on day 3 of hospitalization.

**Methods:**

We used a linear regression to model the relationship between total cost and the set of predictors listed above. We log-transformed total cost to because the raw variable exhibited a strong right skew (Figure 1). The log-transformation appeared to be adequate in addressing the excessive variance in raw total cost (Figure 2). Since we had very little *a priori* information on the relationship between each continuous variable and log(total cost), we centered each continuous predictor at zero to alleviate problems of colinearity between the main effect and non-linear effect of the same variable.

We screened the continuous predictive variables (age, sex, dzgroup, num.co, scoma, race, meanbp, hrt, temp, pafi, and alb) to determine if any strong relationships existed among them that may lead to problems with parameter estimation (Figure 4). We used a simple scatter plot matrix in which each variable is plotted against each other variable. We also included log(total cost) in the scatterplot matrix to assess the shape of possible non-linear relationships between each predictor and log(total cost). We initially modeled all continuous predictors with quadratic terms to capture possible non-linear relationships between the predictors and total cost. We then removed non-significant quadratic terms (p<0.10) from the model while keeping all main effects in a single step. We used diagnostic plots to assess the fit and assumptions of the final model (Figure 5).

All SAS code used to perform these analyses is included in the appendix.

**Results:**

Figure 1: Total cost is skewed to the right.

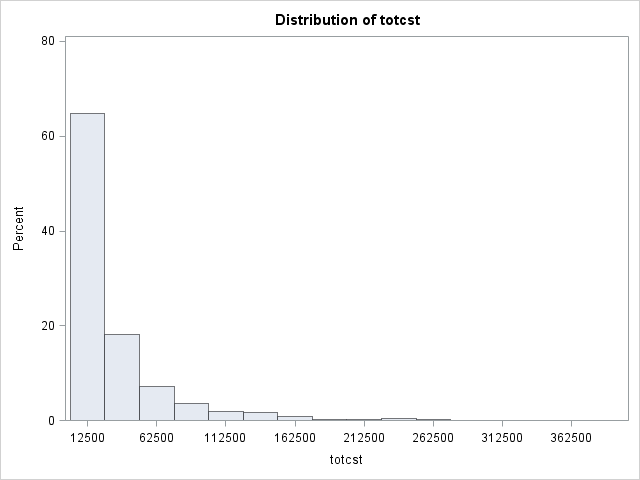


Figure 2: Log transformation of total cost.

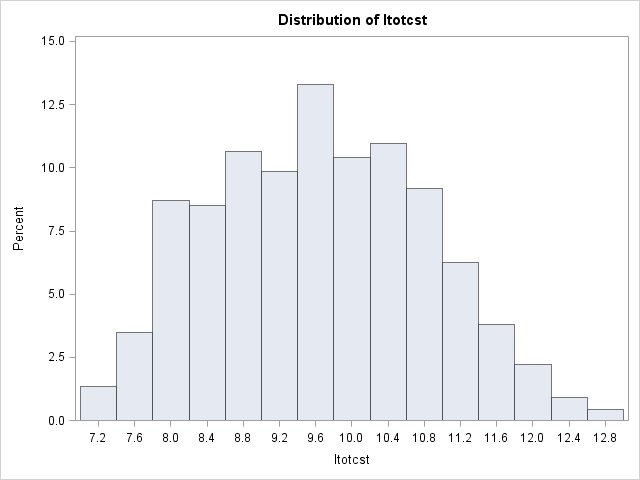


Figure 4: Scatter plots of predictive and total cost.

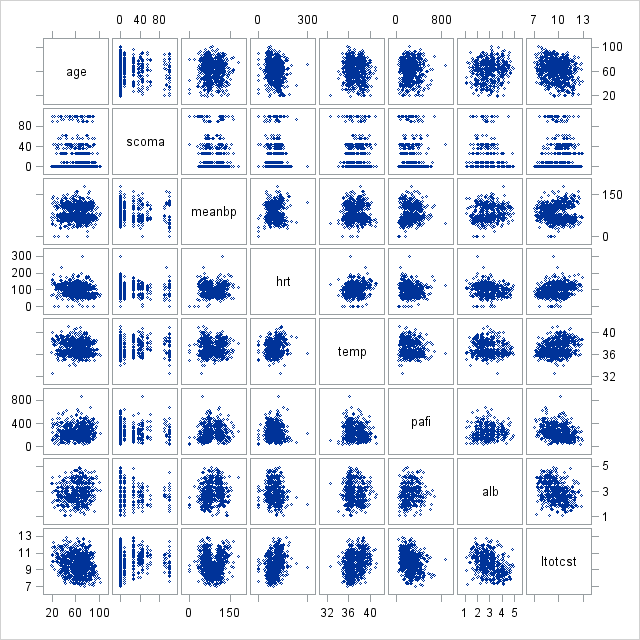


Table 1: Model parameters before variable selection. We dropped highlighted terms from the final model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Adj. R-Square** | **Coeff Var** | **Root MSE** | **ltotcst Mean** |
| 0.469 | 303.0962 | 0.841665 | 9.911084 |

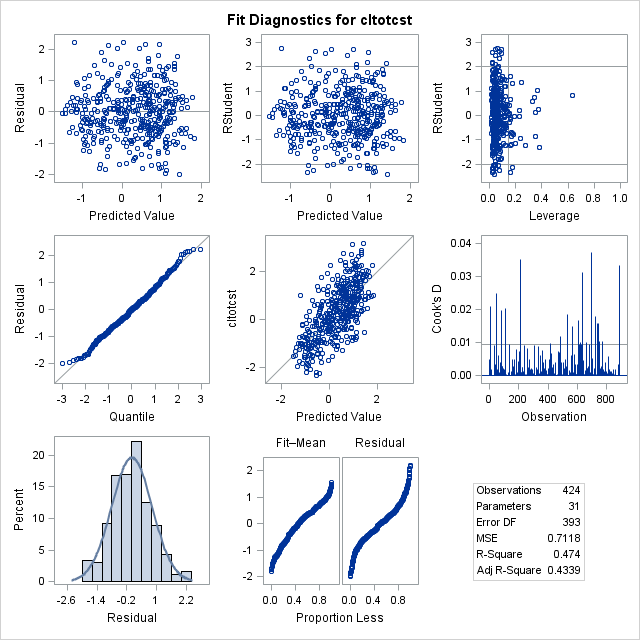
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **age\_c** | 1 | 2.06968644 | 2.06968644 | 2.92 | 0.0882 |
| **age\_c\*age\_c** | 1 | 0.31225786 | 0.31225786 | 0.44 | 0.5071 |
| **scoma\_c** | 1 | 12.39846597 | 12.39846597 | 17.50 | <.0001 |
| **scoma\_c\*scoma\_c** | 1 | 9.94880975 | 9.94880975 | 14.04 | 0.0002 |
| **meanbp\_c** | 1 | 0.30115230 | 0.30115230 | 0.43 | 0.5148 |
| **meanbp\_c\*meanbp\_c** | 1 | 2.31711110 | 2.31711110 | 3.27 | 0.0713 |
| **hrt\_c** | 1 | 5.05816517 | 5.05816517 | 7.14 | 0.0079 |
| **hrt\_c\*hrt\_c** | 1 | 6.24640329 | 6.24640329 | 8.82 | 0.0032 |
| **temp\_c** | 1 | 6.84782441 | 6.84782441 | 9.67 | 0.0020 |
| **temp\_c\*temp\_c** | 1 | 0.00127274 | 0.00127274 | 0.00 | 0.9662 |
| **pafi\_c** | 1 | 3.64795852 | 3.64795852 | 5.15 | 0.0238 |
| **pafi\_c\*pafi\_c** | 1 | 3.06628788 | 3.06628788 | 4.33 | 0.0381 |
| **alb\_c** | 1 | 14.88154986 | 14.88154986 | 21.01 | <.0001 |
| **alb\_c\*alb\_c** | 1 | 0.36178655 | 0.36178655 | 0.51 | 0.4753 |
| **sex** | 1 | 2.15645579 | 2.15645579 | 3.04 | 0.0818 |
| **race** | 4 | 6.42137455 | 1.60534364 | 2.27 | 0.0615 |
| **dzgroup** | 7 | 33.07251179 | 4.72464454 | 6.67 | <.0001 |
| **num\_co** | 7 | 9.22662585 | 1.31808941 | 1.86 | 0.0747 |

Table 2: Final model parameters. The R-squared value only decreased by .001 despite removing three parameters from the model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Adj. R-Square** | **Coeff Var** | **Root MSE** | **ltotcst Mean** |
| 0. 471 | 8.469669 | 0.839436 | 9.911084 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **age\_c** | 1 | 1.87032989 | 1.87032989 | 2.65 | 0.1041 |
| **scoma\_c** | 1 | 12.87525379 | 12.87525379 | 18.27 | <.0001 |
| **scoma\_c\*scoma\_c** | 1 | 10.47160364 | 10.47160364 | 14.86 | 0.0001 |
| **meanbp\_c** | 1 | 0.34512140 | 0.34512140 | 0.49 | 0.4844 |
| **meanbp\_c\*meanbp\_c** | 1 | 2.15649178 | 2.15649178 | 3.06 | 0.0810 |
| **hrt\_c** | 1 | 4.97302441 | 4.97302441 | 7.06 | 0.0082 |
| **hrt\_c\*hrt\_c** | 1 | 6.28504498 | 6.28504498 | 8.92 | 0.0030 |
| **temp\_c** | 1 | 7.64489664 | 7.64489664 | 10.85 | 0.0011 |
| **pafi\_c** | 1 | 3.69841319 | 3.69841319 | 5.25 | 0.0225 |
| **pafi\_c\*pafi\_c** | 1 | 3.19241089 | 3.19241089 | 4.53 | 0.0339 |
| **alb\_c** | 1 | 15.04639084 | 15.04639084 | 21.35 | <.0001 |
| **sex** | 1 | 2.07472889 | 2.07472889 | 2.94 | 0.0870 |
| **race** | 4 | 6.86807814 | 1.71701953 | 2.44 | 0.0467 |
| **dzgroup** | 7 | 34.00932989 | 4.85847570 | 6.89 | <.0001 |
| **num\_co** | 7 | 9.60220760 | 1.37174394 | 1.95 | 0.0612 |

Figure 7: General linear diagnostics of the final model.



**Discussion:**

As shown in Figure 1, the total cost of hospitalization was not normally distributed. After doing a log transformation on total cost, the data was distributed more normally. After analyzing the scatter plot for each predictive variable, we did not detect any strong relationships among the predictors so we did not feel there was any problematic redundancy (Figure 4). Our evaluation of the relationship between each predictor and total cost suggested that several non-linear relationships may exist. In particular, we observed a strong non-linear relationship between meanbp and total cost, which appeared u-shaped, as well as scoma and total cost which appeared to increase until scoma approached 80 and then decrease.

We kept the main effect of each predictor since our goal was to model the relationship between the predictors and total cost, and not necessarily to find the most parsimonious model. However, we did remove non-significant quadratic terms from the model to determine: 1) if they were necessary (added any explained variance), or 2) if the quadratic terms were modifying an important main effect. We highlight the quadratic terms removed from the final model in table 2. Parameter estimates and significance of main effects changed slightly after removing the non-significant quadratic terms. After removal of the quadratic terms the adjusted R-squared (a measure of model fit accounting for model complexity) increased by 0.002, adding further evidence that the quadratic terms were not necessary.

As we show in Figure 7 the final model met all modeling assumptions, including homoscedasticity and normality of the residuals. We identified some higher leverage points but could not justify removing them based on the available information. Also, a careful examination of the predicted vs observed values shows some bias in the predictions at both high and low values (high values are predicted higher than observed and low values are predicted lower than observed). This indicates that the model may be improved further with the addition of missing relevant predictors or with a better non-linear modeling strategy. With our available information we cannot address the former but we can use restricted cubic splines (RCS) to address the latter. Quadratic terms can give poor fit at the extremes of the observed values in the predictor and RCS may alleviate this problem with the inclusion of knots.

**Appendix:**

**proc** **import**

datafile = "C:\Classes\AppliedBiostat\project1\support\_data.csv" DBMS = CSV out = sup;

**run**;

**data** class.sup;

set sup;

if totcst <= **0** then delete;

run;

**proc** **univariate** data=class.Sup;

var totcst;

histogram;

**run**;

**proc** **means** data = class.Sup mean;

var age scoma meanbp hrt temp pafi alb;

output out=center mean=age\_m scoma\_m meanbp\_m hrt\_m temp\_m pafi\_m alb\_m;

**run**;

**data** class.sup1;

merge class.sup center;

retain age\_mn scoma\_mn meanbp\_mn hrt\_mn temp\_mn pafi\_mn alb\_mn;

if \_n\_ = **1** then do;

age\_mn = age\_m;

scoma\_mn = scoma\_m;

meanbp\_mn = meanbp\_m;

hrt\_mn = hrt\_m;

temp\_mn = temp\_m;

pafi\_mn = pafi\_m;

alb\_mn = alb\_m;

end;

drop \_freq\_ \_type\_ age\_m scoma\_m meanbp\_m hrt\_m temp\_m pafi\_m alb\_m;

age\_c = age - age\_mn;

scoma\_c = scoma - scoma\_mn;

meanbp\_c = meanbp - meanbp\_mn;

hrt\_c = hrt - hrt\_mn;

temp\_c = temp - temp\_mn;

pafi\_c = pafi - pafi\_mn;

alb\_c = alb - alb\_mn;

ltotcst = log (totcst);

run;

**proc** **univariate** data=class.Sup1;

var ltotcst;

histogram;

**run**;

**proc** **sgscatter** data=class.sup1;

matrix age scoma meanbp hrt temp pafi alb ltotcst;

**run**;

**proc** **glm** data=class.sup1 plots = (diagnostics);

class sex dzgroup num\_co race;

model ltotcst = age\_c age\_c\*age\_c scoma\_c scoma\_c\*scoma\_c meanbp\_c meanbp\_c\*meanbp\_c hrt\_c hrt\_c\*hrt\_c temp\_c temp\_c\*temp\_c pafi\_c pafi\_c\*pafi\_c alb\_c alb\_c\*alb\_c sex race dzgroup num\_co;

**run**;

**proc** **glm** data=class.sup1 plots = (diagnostics);

class sex dzgroup num\_co race;

model ltotcst = age\_c scoma\_c scoma\_c\*scoma\_c meanbp\_c meanbp\_c\*meanbp\_c hrt\_c hrt\_c\*hrt\_c temp\_c pafi\_c pafi\_c\*pafi\_c alb\_c sex race dzgroup num\_co;

**run**;