**HAP 823 | Spring 2018 | Assignment 01 | February 7, 2018**

**Question 2**

This regression model predicted total cost from comorbidities (CCS variables), age, gender, medical foster home participation, and functional disabilities. Note that the number of observations in this regression model is only 12,990 because the assignment instructions asked to "exclude patients who have 0 cost within the organization." In addition, variable ccs179\_indicator, the indicator variable for comorbidity CCS179 was excluded from the analysis because it had a value of 0 for all retained observations. Notice that the dependent variable was transformed using the natural log transformation in order to reduce the skewness in total cost.



Multiple regression results from SPSS are presented below.



multiple regression output continues on next page . . .



multiple regression output continues on next page . . .



multiple regression output continues on next page . . .



multiple regression output continues on next page . . .



multiple regression output continues on next page . . .



multiple regression output continues on next page . . .



**2.1** Data cleaning details are provided in the following list:

1. Removed all ccs variables [columns AQ through TT in Excel]
2. Removed all \_30 and \_365 variables between column K and AK
3. Created the MFH dummy variable [IF U2="MFH",1,0] in column Y
4. Did a copy/paste special in order to remove formula from column Y
5. Deleted empty column T, cohort column U, and race column W
6. Created a dummy variable for Male in column W based on column T gender [IF(T2="M",1,IF(T2="F",0,"NULL"))]
7. Did a copy/paste special in order to remove formula from column W
8. Deleted column T gender
9. Recoded column I dayssurvived into column W Days\_survived based on column G DaysFollowedInOrd and column H DaysFollowedOutOrg [IF(I2="NULL",MAX(G2:H2),I2)]
10. Did a copy/paste special in order to remove formula from column W
11. Deleted column I dayssurvived
12. Created column W Total\_cost based on column D Daily\_In\_Org, column E Daily\_Out\_Org, column G DaysFollowedInOrg, and column H DaysFollowedOutOrg [IF(H2="NULL",0,H2)\*E2+IF(G2="NULL",0,G2)\*D2]
13. Did a copy/paste special in order to remove formula from column W
14. Deleted column C MFH
15. Recoded column V Total\_cost into column W Total\_cost in order to remove Total\_cost for patients with a value of 0 in column C Daily\_In\_org [IF(C2>0,V2,"NULL")]
16. Did a copy/paste special in order to remove formula from column W
17. Deleted column V Total\_cost (old version)
18. From original dataset created dummy variables for all ccs variables. Dummy=1 if ccsX\_min or ccsX\_max (X represents a number) are both not NULL [For example, for ccs99\_indicator the formula is: IF(KE2="NULL",IF(TT2="NULL",0,1),1)].
19. Copied ccs indicator variables into working file
20. Deleted column C Daily\_In\_org, column D Daily\_Out\_Org, column E Daily\_All\_Cost, column F DaysFollowedInOrg, column G DaysFollowedOutOrg
21. Moved TotalCost to column C
22. Moved Male, age, and MFH to colimns E, F, and G respectively
23. Created a new variable in column A in order to identify rows with one or more NULL values [IF(COUNTIF(B2:JG2,"NULL")>0,1,0)]
24. Deleted all rows with one or more NULL values
25. Removed all duplicate cases (kept the first instance) based on scrssn
26. Created the age x gender interaction term

**2.2**. Descriptive statistics for all variables included in the multiple regression model except comorbidity indicators are presented in the table below. For indicator (dummy) variables the mean can be interpreted as the proportion of cases that fall in the indicator (non-reference) category.



Descriptive statistics for comorbidities are presented in the table below. Note that in the interest of saving space only means are provided. For all of these indicators the possible values are 0 (absence of comorbidity) and 1 (presence of comorbidity), and the sample size is 12,990.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *CCS code* | *M* | *SD* |  | *CCS code* | *M* | *SD* |  | *CCS code* | *M* | *SD* |
| **001** | 0.01 | 0.08 |  | **044** | 0.17 | 0.37 |  | **097** | 0.09 | 0.29 |
| **002** | 0.08 | 0.27 |  | **045** | 0.03 | 0.16 |  | **098** | 0.83 | 0.38 |
| **003** | 0.04 | 0.19 |  | **046** | 0.00 | 0.04 |  | **099** | 0.19 | 0.39 |
| **004** | 0.48 | 0.50 |  | **047** | 0.34 | 0.47 |  | **100** | 0.08 | 0.27 |
| **005** | 0.00 | 0.06 |  | **048** | 0.17 | 0.38 |  | **101** | 0.49 | 0.50 |
| **006** | 0.03 | 0.16 |  | **049** | 0.50 | 0.50 |  | **102** | 0.41 | 0.49 |
| **007** | 0.00 | 0.02 |  | **050** | 0.34 | 0.47 |  | **103** | 0.08 | 0.27 |
| **008** | 0.06 | 0.25 |  | **051** | 0.14 | 0.35 |  | **104** | 0.17 | 0.37 |
| **010** | 0.78 | 0.42 |  | **052** | 0.25 | 0.43 |  | **105** | 0.13 | 0.34 |
| **011** | 0.03 | 0.17 |  | **053** | 0.63 | 0.48 |  | **106** | 0.41 | 0.49 |
| **012** | 0.00 | 0.06 |  | **054** | 0.11 | 0.31 |  | **107** | 0.01 | 0.10 |
| **013** | 0.00 | 0.07 |  | **055** | 0.42 | 0.49 |  | **108** | 0.27 | 0.45 |
| **014** | 0.03 | 0.18 |  | **056** | 0.00 | 0.02 |  | **109** | 0.28 | 0.45 |
| **015** | 0.01 | 0.11 |  | **057** | 0.00 | 0.06 |  | **110** | 0.10 | 0.30 |
| **016** | 0.00 | 0.05 |  | **058** | 0.54 | 0.50 |  | **111** | 0.10 | 0.30 |
| **017** | 0.00 | 0.05 |  | **059** | 0.44 | 0.50 |  | **112** | 0.10 | 0.30 |
| **018** | 0.00 | 0.06 |  | **060** | 0.02 | 0.15 |  | **113** | 0.20 | 0.40 |
| **019** | 0.02 | 0.14 |  | **061** | 0.00 | 0.04 |  | **114** | 0.28 | 0.45 |
| **020** | 0.00 | 0.06 |  | **062** | 0.12 | 0.32 |  | **115** | 0.07 | 0.25 |
| **021** | 0.01 | 0.08 |  | **063** | 0.10 | 0.30 |  | **116** | 0.03 | 0.16 |
| **022** | 0.03 | 0.17 |  | **064** | 0.04 | 0.19 |  | **117** | 0.48 | 0.50 |
| **023** | 0.14 | 0.35 |  | **076** | 0.01 | 0.08 |  | **118** | 0.13 | 0.33 |
| **024** | 0.01 | 0.08 |  | **077** | 0.00 | 0.06 |  | **119** | 0.09 | 0.28 |
| **025** | 0.00 | 0.02 |  | **078** | 0.01 | 0.07 |  | **120** | 0.18 | 0.39 |
| **026** | 0.00 | 0.07 |  | **079** | 0.09 | 0.29 |  | **121** | 0.15 | 0.36 |
| **027** | 0.00 | 0.01 |  | **080** | 0.03 | 0.17 |  | **122** | 0.24 | 0.43 |
| **028** | 0.00 | 0.02 |  | **081** | 0.23 | 0.42 |  | **123** | 0.03 | 0.17 |
| **029** | 0.11 | 0.32 |  | **082** | 0.12 | 0.33 |  | **124** | 0.01 | 0.08 |
| **030** | 0.00 | 0.04 |  | **083** | 0.17 | 0.37 |  | **125** | 0.19 | 0.39 |
| **031** | 0.00 | 0.06 |  | **084** | 0.18 | 0.38 |  | **126** | 0.37 | 0.48 |
| **032** | 0.03 | 0.16 |  | **085** | 0.08 | 0.27 |  | **127** | 0.46 | 0.50 |
| **033** | 0.01 | 0.11 |  | **086** | 0.61 | 0.49 |  | **128** | 0.09 | 0.29 |
| **034** | 0.00 | 0.05 |  | **087** | 0.37 | 0.48 |  | **129** | 0.04 | 0.20 |
| **035** | 0.01 | 0.09 |  | **088** | 0.26 | 0.44 |  | **130** | 0.10 | 0.30 |
| **036** | 0.00 | 0.05 |  | **089** | 0.70 | 0.46 |  | **131** | 0.09 | 0.29 |
| **037** | 0.00 | 0.05 |  | **090** | 0.26 | 0.44 |  | **132** | 0.01 | 0.11 |
| **038** | 0.01 | 0.11 |  | **091** | 0.51 | 0.50 |  | **133** | 0.50 | 0.50 |
| **039** | 0.01 | 0.10 |  | **092** | 0.09 | 0.28 |  | **134** | 0.32 | 0.47 |
| **040** | 0.00 | 0.05 |  | **093** | 0.26 | 0.44 |  | **136** | 0.62 | 0.49 |
| **041** | 0.03 | 0.16 |  | **094** | 0.52 | 0.50 |  | **137** | 0.10 | 0.30 |
| **042** | 0.02 | 0.14 |  | **095** | 0.67 | 0.47 |  | **138** | 0.44 | 0.50 |
| **043** | 0.01 | 0.12 |   | **096** | 0.17 | 0.38 |   | **139** | 0.10 | 0.29 |
| *CCS code* | *M* | *SD* |  | *CCS code* | *M* | *SD* |  | *CCS code* | *M* | *SD* |
| **140** | 0.12 | 0.32 |  | **190** | 0.00 | 0.02 |  | **236** | 0.13 | 0.34 |
| **141** | 0.11 | 0.32 |  | **193** | 0.00 | 0.02 |  | **237** | 0.13 | 0.33 |
| **142** | 0.01 | 0.07 |  | **195** | 0.00 | 0.05 |  | **238** | 0.17 | 0.37 |
| **143** | 0.18 | 0.39 |  | **196** | 0.00 | 0.02 |  | **239** | 0.28 | 0.45 |
| **144** | 0.01 | 0.12 |  | **197** | 0.35 | 0.48 |  | **240** | 0.04 | 0.18 |
| **145** | 0.06 | 0.24 |  | **198** | 0.23 | 0.42 |  | **241** | 0.01 | 0.11 |
| **146** | 0.18 | 0.39 |  | **199** | 0.23 | 0.42 |  | **242** | 0.07 | 0.26 |
| **147** | 0.05 | 0.23 |  | **200** | 0.62 | 0.49 |  | **243** | 0.01 | 0.12 |
| **148** | 0.01 | 0.10 |  | **201** | 0.08 | 0.27 |  | **244** | 0.37 | 0.48 |
| **149** | 0.07 | 0.26 |  | **202** | 0.05 | 0.22 |  | **245** | 0.16 | 0.37 |
| **151** | 0.14 | 0.35 |  | **203** | 0.54 | 0.50 |  | **246** | 0.10 | 0.30 |
| **152** | 0.05 | 0.21 |  | **204** | 0.64 | 0.48 |  | **247** | 0.03 | 0.16 |
| **153** | 0.21 | 0.41 |  | **205** | 0.56 | 0.50 |  | **248** | 0.04 | 0.19 |
| **154** | 0.10 | 0.30 |  | **206** | 0.09 | 0.28 |  | **249** | 0.01 | 0.11 |
| **155** | 0.61 | 0.49 |  | **207** | 0.02 | 0.15 |  | **250** | 0.15 | 0.36 |
| **156** | 0.06 | 0.24 |  | **208** | 0.21 | 0.41 |  | **251** | 0.27 | 0.44 |
| **157** | 0.20 | 0.40 |  | **209** | 0.09 | 0.29 |  | **252** | 0.22 | 0.41 |
| **158** | 0.19 | 0.39 |  | **210** | 0.01 | 0.10 |  | **253** | 0.28 | 0.45 |
| **159** | 0.37 | 0.48 |  | **211** | 0.66 | 0.47 |  | **254** | 0.62 | 0.48 |
| **160** | 0.07 | 0.25 |  | **212** | 0.15 | 0.35 |  | **255** | 0.97 | 0.18 |
| **161** | 0.20 | 0.40 |  | **213** | 0.03 | 0.16 |  | **256** | 0.77 | 0.42 |
| **162** | 0.17 | 0.38 |  | **214** | 0.02 | 0.13 |  | **257** | 0.66 | 0.47 |
| **163** | 0.50 | 0.50 |  | **215** | 0.04 | 0.19 |  | **258** | 0.68 | 0.47 |
| **164** | 0.41 | 0.49 |  | **216** | 0.01 | 0.07 |  | **259** | 0.82 | 0.39 |
| **165** | 0.11 | 0.31 |  | **217** | 0.09 | 0.28 |  | **650** | 0.19 | 0.40 |
| **166** | 0.24 | 0.43 |  | **219** | 0.00 | 0.03 |  | **651** | 0.47 | 0.50 |
| **167** | 0.04 | 0.20 |  | **220** | 0.00 | 0.04 |  | **652** | 0.02 | 0.14 |
| **168** | 0.01 | 0.10 |  | **221** | 0.00 | 0.01 |  | **653** | 0.47 | 0.50 |
| **169** | 0.00 | 0.03 |  | **222** | 0.00 | 0.01 |  | **654** | 0.04 | 0.19 |
| **170** | 0.00 | 0.06 |  | **223** | 0.00 | 0.02 |  | **655** | 0.01 | 0.08 |
| **171** | 0.01 | 0.08 |  | **224** | 0.01 | 0.08 |  | **656** | 0.02 | 0.14 |
| **172** | 0.00 | 0.06 |  | **225** | 0.10 | 0.30 |  | **657** | 0.59 | 0.49 |
| **173** | 0.02 | 0.14 |  | **226** | 0.06 | 0.24 |  | **658** | 0.11 | 0.31 |
| **174** | 0.00 | 0.02 |  | **227** | 0.02 | 0.14 |  | **659** | 0.37 | 0.48 |
| **175** | 0.03 | 0.17 |  | **228** | 0.02 | 0.12 |  | **660** | 0.26 | 0.44 |
| **176** | 0.01 | 0.09 |  | **229** | 0.09 | 0.28 |  | **661** | 0.21 | 0.41 |
| **178** | 0.00 | 0.01 |  | **230** | 0.10 | 0.30 |  | **662** | 0.05 | 0.22 |
| **179** | 0.00 | 0.00 |  | **231** | 0.14 | 0.34 |  | **663** | 0.57 | 0.50 |
| **181** | 0.00 | 0.03 |  | **232** | 0.20 | 0.40 |  | **670** | 0.26 | 0.44 |
| **183** | 0.00 | 0.02 |  | **233** | 0.08 | 0.28 |  |  |  |  |
| **187** | 0.00 | 0.02 |  | **234** | 0.03 | 0.16 |  |  |  |  |
| **188** | 0.00 | 0.01 |   | **235** | 0.15 | 0.36 |   |  |   |   |

**2.3.** The distribution of Total\_cost was found to be positively skewed. Natural log transformation was applied to reduce the skewness. The original and transformed distributions of Total\_cost are shown in the following figure.

****



**2.4.** In order to check the relationship between age and total cost, the standardized residuals from the multiple regression model were plotted against age. This scatter plot did not indicate any unusual pattern suggesting a linear relationship between age and total cost (Note: In the estimated regression model, total cost was transformed by taking the natural log in order to reduce skewness in the dependent variable).



**2.5.** Partial output from the multiple regression results presented earlier for this question is shown below. This output suggests that at 5% level of significance the *age x gender* interaction is not statistically significant, *p* = .222. In other words, after controlling for all other variables in the regression model, the effect of age on total cost is the same for males and females.





Partial multiple regression output

**2.6.** Partial output from the multiple regression results presented earlier for this question is shown below. This output suggests that at 5% level of significance there is no significant difference in total cost between patients who survived and those who died, *B* = -0.536, *p* = .147. The effect of number of days survived (days\_survived) on total cost was significant, *p* < .001. However, after controlling for the effect of all other variables the practical effect was very small, *B* < .001. The main reason why this effect is statistically significant seems to be the very small standard associated with the partial slope coefficient for this effect which results in an inflated test statistic value.



Partial multiple regression output

**2.7.** The multiple regression model predicting total cost was presented earlier for this question. The regression ANOVA table is presented below. The small p value in this table, p < .001 suggests that at least one of the predictors in the multiple regression model has a statistically significant effect on total cost while adjusting for the effect of all remaining predictors in the model.



**2.8.** We do not have any time-varying variables present in the regression variable, therefore the autocorrelation assumption is not relevant here. In order to check the normality assumption a histogram of residuals was constructed. The shape of this histogram suggests that the residuals are approximately normally distributed.



 In order to check for heteroskedasticity the residuals were plotted against predicted values. One or two observations may be candidates for outlier, but other than that the resulting scatter plot does not shows any trend or pattern suggesting that heteroskedasticity is not an issue.



In order to check the linearity assumption scatter plots of residuals versus each predictor should be constructed. However, given the large number of predictors in the estimated regression model (k = 263 predictors) these plots are not presented here. A sample plot was presented earlier in part 2.4 of this question and is reproduced below.



**2.9.** The estimated multiple regression model explained approximately 13.9% of the total variation in total cost. Based on Cohen's (1992) guidelines for *R2* (Small: 2%; Medium, 13%; Large, 26%) this represents a small effect size.



**2.10.** The top 10 predictors of cost based on *p* value are presented in the table below.



**2.11.** Partial output from the multiple regression results presented earlier for this question is shown below. This output suggests that at 5% level of significance there is no significant difference in total cost between patients who participated in medical foster homes and those who did not, *B* = 0.025, *p* = .427.

