Regression analysis

Science, Statistics

R-STATISTICS PROJECT ID ID Lecturer Model Specification The goal of this problem is to build amodel to predict the total cost of claims for coronary heart disease based on the other variables. To build a model that would predict the dependent variable we need to have an generalized linear model where we use multiple regression. The following equation model is thus estimated;

Looking for outliers

Descriptive statistics

The summary statistics presented below shows that the data set does contain any outliers and as such there is no such need to remove them.

```
cost int duration age
Min. : 0. 0 Min. : 0. 000 Min. : 0. 00 Min. : 24. 00
1st Qu.: 161. 1 1st Qu.: 1. 000 1st Qu.: 41. 75 1st Qu.: 55. 00
Median : 507. 2 Median : 3. 000 Median : 165. 50 Median : 60. 00
Mean : 2800. 0 Mean : 4. 707 Mean : 164. 03 Mean : 58. 72
3rd Qu.: 1905. 5 3rd Qu.: 6. 000 3rd Qu.: 281. 00 3rd Qu.: 64. 00
Max. : 52664. 9 Max. : 47. 000 Max. : 372. 00 Max. : 70. 00
```

Normality test

The most important assumption for the OLS model is that the dat set need to follow a normal distribution. We therefore had to check whether our data set followed a normal distribution. It is clear that the data seems to come from a normal distribution since the p-value for the Shapiro-wilk normality test is 0. 000 (avalue less than 5% significance level)

```
Shapiro-Wilk normality test
data: resid(fit)
```

W = 0. 7456, p-value < 2. 2e-16

Model selection

In order to identify the best model, a series of regressions were conducted. Model (1) has int as the only explanatory variable, model (2) has int and duration as the explanatory variables, model (3) has int and age as the explanatory variables, model (4) has duration and age as the explanatory variables and lastly, model (5) has all the three explanatory variables included in the model.

Model (1)

Model (2)

Model (3)

Model (4)

Model (5)

int

869. 1

(0. 000***)

865. 89

(0. 000***)

867. 72

(0. 000***)

862. 88

(0. 000***)

duration

0. 6507

(0. 64)

10. 351

(0. 000***)

0. 9690

(0. 491)

age

-35. 42

(0. 145)

-83. 305

(0. 01755**)

-37. 9015

(0. 123)

R-squared

0. 5282

0. 5283

0. 5295

0. 0378

0. 5297

Adjusted R-squared

0. 5276

0. 5271

0. 5283

0. 0354

0. 5279

p-value

0. 000

0. 000

0. 000

0. 000

0. 000

From the above table, it is clear that the variable duration is an irrelevant variable in the model since addition of this particular variable in model (2) resulted to a decrease in the value of the adjusted R-squared from 0. 5276 to 0. 5271; this is a clear indication that variable, duration is not worth being included in the model. It is therefore evident that the best model is model (3) with the highest value of the adjusted R-squared.

Appendix

R-code

```
> heartlayout(matrix(c(1, 2, 3, 4), 2, 2))

> plot(fit)

> summary(heart)

> qqnorm(resid(fit))

> qqnorm

> qqline(resid(fit))

> shapiro. test(resid(fit))

> fit fit

> summary(fit)

> fit fit

> summary(fit)

> fit fit

> summary(fit)
> fit summary(fit)

> fit summary(fit)
```