Stats 401 Lab 9_Solutions

401 GSI team

3/15/2018 and 3/16/2018

Solutions: Q1

```
model.matrix(ses_lm)[1:3,]
```

```
##   (Intercept)   ses
## 1           1 -0.13
## 2           1 -0.39
## 3           1 -0.80
```

```
model.matrix(ses_edu_lm)[1:3,]
```

```
##   (Intercept)   ses pareducollege pareduhs paredulesshs pareduma pareduphd
## 1           1 -0.13             0        1            0        0         0
## 2           1 -0.39             0        1            0        0         0
## 3           1 -0.80             0        1            0        0         0
```

$H_0 : Y = \mathbb{X}\beta + \epsilon$ where $\mathbb{X} = [1 \quad X_1]$ where 1 is the intercept column containing all 1's and X_1 is the parents' socio-economic status.

$H_a : Y = \mathbb{X}\beta + \epsilon$where $\mathbb{X} = [1 \;\; X_1 \;\; X_2 \;\; X_3 \;\; X_4 \;\; X_5 \;\; X_6]$ where 1 and X_1 are as defined above. X_2 is an indicator variable which equals 1 if the parents' education level is college. X_3 equals 1 if 'paredu' = high school, 0 otherwise X_4 equals 1 if 'paredu' = below high school, 0 otherwise X_5 equals 1 if 'paredu' = ma, 0 otherwise X_6 equals 1 if 'paredu' = phd, 0 otherwise

Solutions: Q2a

Fit the lm models:

```
lm0 <- lm(math ~ ses + paredu, data = nels88)
lm1 <- lm(math ~ ses + paredu + sex, data = nels88)
```

Get the design matrix:

```
model.matrix(lm1)[1:3,]
```

```
##   (Intercept)   ses pareducollege pareduhs paredulesshs pareduma pareduphd
## 1           1 -0.13             0        1            0        0         0
## 2           1 -0.39             0        1            0        0         0
## 3           1 -0.80             0        1            0        0         0
##   sexMale
## 1       0
## 2       1
## 3       1
```

$H_0 : Y = \mathbb{X}\beta + \epsilon$where $\mathbb{X} = [1 \;\; X_1 \;\; X_2 \;\; X_3 \;\; X_4 \;\; X_5 \;\; X_6]$ where X is the same as the H_a for Q1.

$H_a : Y = \mathbb{X}\beta + \epsilon$where $\mathbb{X} = [1 \;\; X_1 \;\; X_2 \;\; X_3 \;\; X_4 \;\; X_5 \;\; X_6 \;\; X_7]$ where the additional variable X_7 is the indicator for the sex of the student which equals 1 if male and 0 if female.

Solutions: Q2b

First, we need RSS_0 and RSS_a

```
rss_0 <- sum(residuals(lm0)^2); rss_0
```

```
## [1] 18082.73
```

```
rss_a <- sum(residuals(lm1)^2); rss_a
```

```
## [1] 18075.71
```

$$
\begin{aligned}
f &= \frac{(RSS_0 - RSS_a)/d}{RSS_a/(n-q)} \\
f &= \frac{(18082.73 - 18075.71)/(253-252)}{18075.71/252} \\
f &= \frac{(7.02)}{71.72901} \\
f &= 0.09786835
\end{aligned}
\tag{1}
$$

Solutions: Q2b contd

p-value

```
pf(0.09786835,1, 252, lower.tail = FALSE)
```

```
## [1] 0.7546616
```

Since p-value $= 0.7546616 > 0.05$ we fail to reject the null hypothesis. That is, the sex of the student does not affect their test score! :)

Solutions: Q2c

Confirm using ANOVA

```
anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: math
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## ses         1 12391.4 12391.4 172.7532 < 2.2e-16 ***
## paredu      5  1642.4   328.5   4.5796 0.0005132 ***
## sex         1     7.0     7.0   0.0979 0.7546221
## Residuals 252 18075.7    71.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we see, the p-value is 0.7546221 which matches our p-value from 2b and thus our conclusion is confirmed.