STA 108 Project II

Junyao Lu, Fengshuo Song
STA108 SectionB
Instructor:JoAnna Whitener

11/20/2019

Part I:

The data is used from the file salary1.csv.

The summary of data is shown as follows:

```
##        sl              yd                dg
##  Min.   :15000   Min.   : 1.00   doctorate:34
##  1st Qu.:18247   1st Qu.: 6.75   masters  :18
##  Median :23719   Median :15.50
##  Mean   :23798   Mean   :16.12
##  3rd Qu.:27259   3rd Qu.:23.25
##  Max.   :38045   Max.   :35.00
```

(a)

```
##
## Call:
## lm(formula = sl ~ ., data = salary1.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8589.8 -2724.4  -682.9  2391.1  9486.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17422.01    1066.60  16.334  < 2e-16 ***
## yd            483.99      63.98   7.565 8.89e-10 ***
## dgmasters   -4113.89    1361.44  -3.022  0.00399 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4090 on 49 degrees of freedom
## Multiple R-squared:  0.541,  Adjusted R-squared:  0.5222
## F-statistic: 28.87 on 2 and 49 DF,  p-value: 5.19e-09
```

Our estimated linear regression model is $\hat{Y}$=17422.01+483.99X1-4113.89X2, where

Y represents the estimated three month salary in dollars,

X1 represents the number of years since the subject earned their highest degree,

X2 represents the highest degree (0 for doctorate and 1 for masters).

(b) Interpret b1 and b2

b1: When the number of years after earning the highest degree increases by 1 year, the average increase in the 3-month salary is 483.99 dollars, holding the highest degree earned constant. b2:The subject whose highest degree is masters has less 3-month salary than the subject whose highest degree is doctorate by 4113.89 dollars, on average, holding the number of years after earning highest degree constant.

(c)

```
##        1
## 22261.91
```

The predicted 3-month salary of a subject who has 10 years of experience and has earned their doctorate is 22261.91 dollars.

(d)

```
##     2.5 %   97.5 %
## yd 355.42 612.5604
```

We are 95% confident that when the number of years after earning the highest degree increases by 1 year, the average increase in the 3-month salary is between 355.42 dollars and 612.56 dollars, holding the highest degree earned constant.

(e)

```
##                0.833 %   99.167 %
## (Intercept) 14777.9575 20066.0536
## yd            325.3897   642.5906
## dgmasters   -7488.8303  -738.9551
```

Interpretation for β 1:

We are overall 95% confident that when the number of years after earning the highest degree increases by 1 year, the average increase in the 3-month salary is between 325.3897 dollars and 642.5906 dollars, holding the highest degree earned constant.

Interpretation for β 2:

We are overall 95% confident that the subject whose highest degree is masters has less 3-month salary than the subject whose highest degree is doctorate by between 738.9551 dollars and 7488.8303 dollars, on average, holding the number of years after earning highest degree constant.

(f)

```
##            lwr      upr
## [1,]  6159.973 25296.15
## [2,] 13136.101 31387.71
```

We are overall 90% confident that the exact 3-month salary of a subject who has 5 years of experience and whose highest degree is masters is between 6159.97 dollars and 25296.15 dollars, and the exact 3-month salary of a subject who has 10 years of experience and whose highest degree is doctorate is between 13136.10 dollars and 31387.71 dollars.

Part II:

The data is used from the file salary2.csv.

The summary of data is shown as follows:

```
##        sl              yd               dg           sx            rk
##  Min.   :15000   Min.   : 1.00   doctorate:34   female:14   assistant:18
##  1st Qu.:18247   1st Qu.: 6.75   masters  :18   male  :38   associate:14
##  Median :23719   Median :15.50                              full     :20
##  Mean   :23798   Mean   :16.12
##  3rd Qu.:27259   3rd Qu.:23.25
##  Max.   :38045   Max.   :35.00
```

(a)

Our full regression model for salary is shown as follows:

```
## (Intercept)          yd   dgmasters      sxmale rkassociate      rkfull
## 16454.06785   107.72990   -39.03539  1153.77112  3718.83503  9819.22279
```

$\hat{Y}$ = 16454.07 + 107.73X1 - 39.04X2 + 1153.77X3 + 3718.84X41 + 9819.22X42, where

Y represents the estimated three month salary in dollars,

X1 represents the number of years since the subject earned their highest degree,

X2 represents the highest degree (0 for doctorate and 1 for masters),

X3 represents the gender (0 for female and 1 for male),

X4 represents the rank (X41=0 X42=0 for assistant, X41=1 X42=0 for associate and X41=0 X42=1 for full).

The reduced model is shown as follows:

```
## (Intercept)          yd   dgmasters      sxmale
##  15594.8591    476.0263  -4228.3524   2730.1492
```

$\hat{Y}$ = 15594.9 + 476.0X1 - 4228.4X2 + 2730.2X3

```
## Analysis of Variance Table
##
## Model 1: sl ~ yd + dg + sx + rk
## Model 2: sl ~ yd + dg + sx
##   Res.Df       RSS Df  Sum of Sq      F    Pr(>F)
## 1     46 403725768
## 2     48 744165729 -2 -340439961 19.395 7.791e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With the information above, we can then do a hypothesis test to see if X4(rank) can be dropped or not.

i) H0: β 4 = β 5 = 0; Ha: at least one of the β 4 or β 5 $\neq$ 0.

ii) Fs = 19.395

iii) p-value = 7.791e-07

iv) As $\alpha = 0.01$, p-value $< \alpha$. So, we reject H0. We conclude that we cannot drop X4(rank) from the model.

(b)

The reduced model is shown as follows:

```
## (Intercept)          yd rkassociate      rkfull
## 17166.46499    95.08447  4209.65030 10310.29631
```

$\hat{Y} = 17166.46499 + 95.08447X1 + 4209.65030X41 + 10310.29631X42$

```
## Analysis of Variance Table
##
## Model 1: sl ~ yd + dg + sx + rk
## Model 2: sl ~ yd + rk
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     46 403725768
## 2     48 415783967 -2 -12058199 0.6869 0.5082
```

With the information above, we can then do a hypothesis test to see if X2(highest degree) and X3(gender) can be dropped or not.

i) H0: β 2 = β 3 = 0; Ha: at least one of the β 2 or β 3 $\neq$ 0.

ii) Fs = 0.6869

iii) p-value = 0.5082

iv) As $\alpha = 0.01$, p-value $> \alpha$. So, we fail to reject H0. We conclude that we can drop X2(highest degree) and X3(gender) from the model.

(c)

Based on our observations from (a) and (b), our "best" model is Y ~ X1 + X4. A summary of the best model is shown as follows:

```
## (Intercept)          yd rkassociate      rkfull
## 17166.46499    95.08447  4209.65030 10310.29631
```

So, the estimated linear equation is:

$\hat{Y} = 17166.46 + 95.08X1 + 4209.65X41 + 110310.30X42$

(d)

```
## [1] 0.5724403
```

57.24% of the error for a model including only X1(number of years since the subject earned their highest degree) is reduced when we add X4(rank).

(e)

```
## [1] 0.02900112
```

2.90% of the error for a model including only X1(number of years since the subject earned their highest degree),X4(rank) is reduced when we add X2(highest degree),X3(gender).

(f)

We can conclude from (d) and (e) that the effect of adding X2(highest degree) and X3(gender) to the given model is 2.90%, which is very small compared to the effect of adding X4(rank). Therefore, it is not necessary to include X2(highest degree) and X3(gender) in our model. So, the above values agree with our "best model" from part (c).

Appendix Code

```
library(ggplot2)
salary1.data = read.csv("C:/Users/songf/Documents/FQ2019/STA 108/project/salary1.csv")
summary(salary1.data)
lm.fit = lm(sl~., data=salary1.data)
summary(lm.fit)
predict(lm.fit, data.frame(yd = 10, dg = 'doctorate'))
confint(lm.fit, "yd")
confint(lm.fit, level = (1 - 0.05/3))
p = 3
g = 2
n = dim(salary1.data)[1]
X = model.matrix(lm.fit)
middle.part=solve(t(X)%*%X)
X.new = cbind(1, c(5,10), c(1,0))
sigma.hat = summary(lm.fit)$sigma
width.half = sqrt(g*qf(0.90, g, n-p)*(1+apply(X.new, 1, function(x){t(x)%*%middle.part%*%x})))*sigma.ha
Schef.pred = matrix(rep(predict(lm.fit, data.frame(yd=c(5,10), dg=c('masters','doctorate'))), 2), ncol =
Schef.pred[, 1] = Schef.pred[, 1] - width.half
Schef.pred[, 2] = Schef.pred[, 2] + width.half
colnames(Schef.pred) = c('lwr', 'upr')
print(Schef.pred)
salary2.data = read.csv("C:/Users/songf/Documents/FQ2019/STA 108/project/salary2.csv")
summary(salary2.data)
full.model = lm(sl~., data=salary2.data)
full.model$coefficients
reduced.model1 = lm(sl~yd+dg+sx, data=salary2.data)
reduced.model1$coefficients
anova(full.model, reduced.model1)
reduced.model2 = lm(sl~yd+rk, data=salary2.data)
reduced.model2$coefficients
anova(full.model, reduced.model2)
best.model = lm(sl~yd+rk, data=salary2.data)
best.model$coefficients
orig.model = lm(sl~yd, data=salary2.data)
SSE.before=sum(orig.model$residuals^2)
SSE.after=sum(best.model$residuals^2)
partialR2=(SSE.before-SSE.after)/SSE.before
print(partialR2)
full.model = lm(sl~yd+dg+sx+rk, data=salary2.data)
SSE.before=sum(best.model$residuals^2)
SSE.after=sum(full.model$residuals^2)
partialR2=(SSE.before-SSE.after)/SSE.before
print(partialR2)
```