Project 2

Oliver Hannaoui & Will Khouri, STA 108: Dr. JoAnna C. Whitener, section B01 & section B02

11/24/2019

Part I.

(a)

```
##
## Call:
## lm(formula = sl ~ yd + dg, data = salary1)
##
## Coefficients:
## (Intercept)           yd    dgmasters
##       17422          484        -4114
```

The estimated linear regression model is displayed below.

Y = 17422 + 484X1 - 4114X2

(b)

For b1 we say that when the number of years since the subject earned their highest degree increases by 1 year the average increase in the three month salary of the subject is $484, holding the highest degree earned by the subject constant.

For b2 we say that, on average, we expect the 3 month salary of a subject whose highest degree earned is a masters to be $4,114 less than a subject whose highest degree earned is a doctorate, holding the number of years since the subject earned their highest degree constant.

(c)

```
## [1] 22262
```

The predicted 3-month salary of a subject who has 10 years of experience and has earned their doctorate is $22,262.

(d)

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

We are 95% confident that the average 3 month salary increase for a 1 year increase in the number of years since the subject earned their highest degree is between $355.42 and $612.56, holding highest degree earned constant.

(e)

```
##                  2.5 %     97.5 %
## (Intercept) 15278.595 19565.4159
## yd            355.420   612.5604
## dgmasters   -6849.802 -1377.9832
```

We found the simultaneous 95% intervals for b1 and b2 using Bonferonni’s multiplier.

We expect the 3 month salary of a subject whose highest degree earned is a masters to be $4,114 less than a subject whose highest degree earned is a doctorate, holding the number of years since the subject earned their highest degree constant.

For b1, the expected increase in 3-month salary for every additional year of experience holding highest degree earned constant, we are 95% family-wise confident the expected increase in 3-month salary is between $355.42 and $612.56.

For b2, the expected difference in 3-month salary between a subject whose highest degree earned is a masters and doctorate, holding the years of experience constant, we are 95% family-wise confident that, on average, masters degree subjects’ 3-month salary is between $6,849.80 and $1,377.98 less than that of a doctorate degree subject.

Simply put, the 95% simultaneous confidence interval using Bonferonni’s multiplier for b1 has a lower bound of 355.420 and an upper bound of 612.5604.

The 95% simultaneous confidence interval using Bonferonni’s multiplier for b2 has a lower bound of -6849.802 and an upper bound of -1377.9832.

(f)

```
##           [,1]
## [1,]  6159.973
## [2,] 13136.101
```

```
##          [,1]
## [1,] 25296.15
## [2,] 31387.71
```

Above we created two prediction intervals for the 3-month salary at values of 5 and 10 years of experience and highest degree earned of masters and doctorate respectively using Scheffe’s multiplier. The first block of output is the lower bound for our prediction intervals and the second block of output is the upper bounds for our prediction intervals.

For the prediction of a 3-month salary for 5 years of experience and a masters degree we are family-wise 90% confident that the 3-month salary is between $6,159.973 and $25,296.15.

For the prediction of a 3-month salary for 10 years of experience and a doctorate degree we are family-wise 90% confident that the 3-month salary is between $13,136.10 and $31,387.71.

Part II.

(a)

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

Our null hypothesis for this hypothesis test is that B4 = 0.

Our alternate hypothesis for this hypothesis test is that B4 does not equal 0.

We are testing to see if we should drop the predictor X4, the rank of the subject, from our model.

Conducting our test we find a F-statistic of 19.395 with a corresponding p-value of 7.791 x 10^-7. Since the p-value is statistically significant and signficiantly lower than our signficance level alpha = 0.01 we reject the null hypothesis. This conclusion means that we decide that B4 does not equal 0 and we do not drop X4, the rank of the subject, from our model.

(b)

```
## Analysis of Variance Table
##
## Model 1: sl ~ sx
## Model 2: sl ~ yd + dg + sx + rk
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1     50 1671623638
## 2     46  403725768  4 1267897870 36.116 1.185e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our null hypothesis for this hypothesis test is that B2 = B3 = 0.

Our alternate hypothesis for this hypothesis test is that at least B2 or B3 does not equal 0.

We are testing to see if we should drop at least one of the predictors X2, highest degree earned, or X3, gender, from our model.

Conducting our test we find a F-statistic of 36.116 with a corresponding p-value of 1.185 x 10^-13. Since the p-value is statistically significant and signficiantly lower than our signficance level alpha = 0.01 we reject the null hypothesis.

This conclusion means that we decide that at least X2 or X3 does not equal 0 and at least one of the variables, highest degree earned or gender, should remain in our model.

(c)

```
##
## Call:
## lm(formula = sl ~ yd + dg + sx + rk, data = salary2)
##
## Coefficients:
## (Intercept)           yd    dgmasters       sxmale  rkassociate
##    16454.07       107.73       -39.04      1153.77      3718.84
##      rkfull
##     9819.22
```

Based on our observations from the hypothesis tests we conducted we conclude that the full model is the “best” model compared to a reduced model.

The estimated linear equation is as follows:

Y = 1654.07 + 107.73X1 -39.04X2 + 1153.77X3 + 3718.84X4C1 + 9819.22X4C2

These values are all found from the calculations which can be found in the appendix code and output below.

(d)

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

The expected reduction in error we expect to see when we add X4 to a model with only X1 in it is approximately 57.24%.

That is, when we add the rank of the subject, (assistant, associate, or full) to a model with only the number of years of experience, we reduce the error in the 3-month salary of the subject in dollars by approximately 57.24%.

(e)

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

The expected reduction in error we expect to see when we add X2 and X3 to a model with only X1 and X4 in it is approximately 3%.

That is, when we add highest degree earned and gender of the subject to a model with only the number of years of experience and the rank of the subject (assistant, associate, or full), we reduce the error in the 3-month salary of the subject in dollars by approximately 3%.

(f)

Based on the information from above, we conclude that although adding X4, rank of subject, to a reduced model of just X1, years of experience, reduces the error in the 3-month salary of the subject in dollars almost 20 times more than adding X2, advanced degree earned, and X3, gender of subject, to a model with only X1 and X4.

We would still keep the full model since each variable reduces the overall error in 3-month salary.

In summation, based on the partial R-squares we calculated and the hypothesis test conducted we conclude that the full model calculated in (c) of Part II is the “best” model.

Appendix:

```
# Loading the data
salary1 <- read.csv("~/Desktop/salary1.csv")
salary2 <- read.csv("~/Desktop/salary2.csv")
salary_model = lm(sl ~ yd + dg, data = salary1)
salary_model
prediction = 17422 + 484*10 - 4114*0
prediction
confint(object=salary_model, parm="yd", level=0.95)
# Calculating 95% simulatenoeus CI for b1 and b2 of the model
confint(salary_model, adjust.method = "bonferroni", level=0.95)
Cstar_Scheffe = sqrt(2 * qf(.90, 2, 49)) # Scheffe multiplier

Xnew_mat = cbind(c(1,1), c(5,10), c(1,0))
Ystar = Xnew_mat %*% salary_model$coefficients

X = model.matrix(salary_model)
mse = sum(salary_model$residuals^2)/(49)
s_Ystar = sqrt(diag(mse + mse*(Xnew_mat %*% solve(t(X)%*%X)) %*% t(Xnew_mat)))
LB_Scheffe = Ystar - Cstar_Scheffe * s_Ystar
UB_Scheffe = Ystar + Cstar_Scheffe * s_Ystar

LB_Scheffe # Lower bounds for both prediction intervals
UB_Scheffe # Upper bounds for both prediction intervals


reduced_model = lm(sl~yd+dg+sx, data=salary2)
full_model = lm(sl~yd+dg+sx+rk, data=salary2)
anova(reduced_model, full_model)
reduced_model = lm(sl~sx, data=salary2)
full_model = lm(sl~yd+dg+sx+rk, data=salary2)
anova(reduced_model, full_model)
salary_model2 = lm(sl ~ yd + dg + sx + rk, data = salary2)
salary_model2
fit_after = lm(sl~yd+rk, data=salary2)
SSE_after = sum(fit_after$residuals^2)

fit_before = lm(sl~yd, data=salary2)
SSE_before = sum(fit_before$residuals^2)

partialR2 = (SSE_before-SSE_after)/(SSE_before)
partialR2
fit_after = lm(sl~yd+dg+sx+rk, data=salary2)
SSE_after = sum(fit_after$residuals^2)

fit_before = lm(sl~yd+rk, data=salary2)
SSE_before = sum(fit_before$residuals^2)

partialR2 = (SSE_before-SSE_after)/(SSE_before)
partialR2
```