Practical Exercise 4 | Statistics for CSAI II

Burcu Ibicioglu, u986202

The goals of this exercise are to (a) to use R to run multiple linear regression models, b) check the assumptions of the model and c) report your results with a focus on interactions and regression with multiple categories.

For this part of Practical Exercise #4, Tasks indicate things that you need to complete in R/R Studio.

Task 1. Install the carData package for R and load the "Salaries" data set. This data includes the 2008-2009 nine-month academic salary for Assistant Professors, Associate Professors, and Professors at a college in the U.S..

```
install.packages("carData")
```

```
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
```

```
library(carData)
data <- ("Salaries")
```

Task 2. Inspect the data by looking at the first few entries and the last few entries in the dataset as well as the variable types. For this analysis, we are interested in predicting the salaries of professors as a function of the number of years since they obtained their Ph.D. (yrs.since.phd) and the number of years of service (yrs.service) and gender (sex).

```
head (Salaries)
```

```
##        rank discipline yrs.since.phd yrs.service  sex salary
## 1      Prof          B            19          18 Male 139750
## 2      Prof          B            20          16 Male 173200
## 3  AsstProf          B             4           3 Male  79750
## 4      Prof          B            45          39 Male 115000
## 5      Prof          B            40          41 Male 141500
## 6 AssocProf          B             6           6 Male  97000
```

```
tail (Salaries)
```

```
##         rank discipline yrs.since.phd yrs.service  sex salary
## 392     Prof          A            30          19 Male 151292
## 393     Prof          A            33          30 Male 103106
## 394     Prof          A            31          19 Male 150564
## 395     Prof          A            42          25 Male 101738
## 396     Prof          A            25          15 Male  95329
## 397 AsstProf          A             8           4 Male  81035
```

```
str (Salaries)
```

```
## 'data.frame':    397 obs. of  6 variables:
##  $ rank         : Factor w/ 3 levels "AsstProf","AssocProf",..: 3 3 1 3 3 2 3 3 3 3 ...
##  $ discipline   : Factor w/ 2 levels "A","B": 2 2 2 2 2 2 2 2 2 2 ...
```

```
##  $ yrs.since.phd: int  19 20 4 45 40 6 30 45 21 18 ...
##  $ yrs.service  : int  18 16 3 39 41 6 23 45 20 18 ...
##  $ sex          : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 1 ...
##  $ salary       : int  139750 173200 79750 115000 141500 97000 175000 147765 119250 129000 ...
```

a. Generate descriptive statistics. Evaluate these descriptives and print them here.

```
install.packages("psych")
```

```
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.3.3
```

```
describe(Salaries)
```

```
##               vars   n      mean       sd median   trimmed      mad   min
## rank*            1 397      2.50     0.77      3      2.62     0.00     1
## discipline*      2 397      1.54     0.50      2      1.55     0.00     1
## yrs.since.phd    3 397     22.31    12.89     21     21.83    14.83     1
## yrs.service      4 397     17.61    13.01     16     16.51    14.83     0
## sex*             5 397      1.90     0.30      2      2.00     0.00     1
## salary           6 397 113706.46 30289.04 107300 111401.61 29355.48 57800
##                  max  range  skew kurtosis      se
## rank*              3      2 -1.12    -0.38    0.04
## discipline*        2      1 -0.18    -1.97    0.03
## yrs.since.phd     56     55  0.30    -0.81    0.65
## yrs.service       60     60  0.65    -0.34    0.65
## sex*               2      1 -2.69     5.25    0.01
## salary        231545 173745  0.71     0.18 1520.16
```

b. Generate a correlation matrix that includes all appropriate variables in the data set and print it here. Consider if there are any variables that we are interested in that you should be concerned about multicollinearity problems. If there is a correlation that is too high, make a decision about whether to drop one of the variables or try centering both predictor variables.

```
nvars <- Salaries[, c("yrs.since.phd", "yrs.service", "salary")]

cm <- cor(nvars)

cm
```

```
##               yrs.since.phd yrs.service    salary
## yrs.since.phd     1.0000000   0.9096491 0.4192311
## yrs.service       0.9096491   1.0000000 0.3347447
## salary            0.4192311   0.3347447 1.0000000
```

Task 3. Run a multiple regression model to predict "salary' that includes the variables of interest described in Task 2 (yrs.since.phd, yrs.service, sex), but taking into account your decisions from question 2. For example, perhaps you are leaving a variable out or you are including centered versions of some variables. Generate 95% confidence intervals of the b estimates and also generate the standardized beta estimates.

```
model <- lm(salary ~ yrs.since.phd  + yrs.service + sex, data = Salaries)
confint(model, level = 0.95)
```

```
##                    2.5 %     97.5 %
## (Intercept)   73437.7833 92314.0402
## yrs.since.phd  1049.1922  2056.3219
```

```
## yrs.service   -1149.1001  -150.4215
## sexMale        -696.9875 17611.1175
```

```
summary(model)
```

```
##
## Call:
## lm(formula = salary ~ yrs.since.phd + yrs.service + sex, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -79586 -19564  -3018  15071 105898
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    82875.9     4800.6  17.264  < 2e-16 ***
## yrs.since.phd   1552.8      256.1   6.062 3.15e-09 ***
## yrs.service     -649.8      254.0  -2.558   0.0109 *
## sexMale         8457.1     4656.1   1.816   0.0701 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27280 on 393 degrees of freedom
## Multiple R-squared:  0.1951, Adjusted R-squared:  0.189
## F-statistic: 31.75 on 3 and 393 DF,  p-value: < 2.2e-16
```

a. Report the results here in APA format. Be sure to include the adjusted R2 value, the b estimates, the p-values, and the 95% confidence intervals. What can you conclude from your results?

###The multiple regression model showed that years since PhD significantly increased salary ($p < 0.001$). Being male was related to salary increase, but was not significant ($p = 0.0701$). Overall, the model was significant since $F(3, 393) = 31.75$ and $p < 0.001$ with an adjusted R^2 of 0.189.###

Task 4. Now run a multiple regression model to predict "salary" that includes the same variables from your last model, but tests for an interaction between sex and yrs.since.phd. Is there a significant interaction? If yes, then you should compare the model fit to your first model. If it is better, then generate 95% confidence intervals of the b estimates and also generate the standardized beta estimates and report the results and change in adj R2. Otherwise, proceed to answer question 4a.

```
intmodel <- lm(salary ~ yrs.since.phd * sex + yrs.service, data = Salaries)
summary(intmodel)
```

```
##
## Call:
## lm(formula = salary ~ yrs.since.phd * sex + yrs.service, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -78781 -20091  -3212  14720 106268
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            73098.6     8647.7   8.453 5.65e-16 ***
## yrs.since.phd           2124.9      492.7   4.312 2.05e-05 ***
## sexMale                19135.2     9132.5   2.095   0.0368 *
## yrs.service             -621.2      254.6  -2.440   0.0151 *
## yrs.since.phd:sexMale   -634.1      466.7  -1.359   0.1750
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27250 on 392 degrees of freedom
## Multiple R-squared:  0.1989, Adjusted R-squared:  0.1907
## F-statistic: 24.33 on 4 and 392 DF,  p-value: < 2.2e-16
```

```
confint(intmodel, level = 0.95)
```

```
##                              2.5 %     97.5 %
## (Intercept)           56096.847 90100.3889
## yrs.since.phd          1156.147  3093.6393
## sexMale                1180.387 37090.0247
## yrs.service           -1121.759  -120.7390
## yrs.since.phd:sexMale -1551.671   283.4726
```

a. What can you conclude from your results for Task 4?

###The model showed no significant interaction effect on salary since the p-value was 0.1750. The adjusted R^2 increased from 0.189 to 0.1907. There i snot much differnece since the interaction term does not significantly improve the model.###

b. What salary would you expect if you are female and have 7 years since Ph.D.?

expected = 73098.6 + (2124.9x7) = 87972.9 The expected salary for a female who has 7 years since Ph.D. is 87,972.9$

Task 5. Load the "Friendly" dataset from the carData package for R. This data includes results from a word recall experiment with three conditions: Before (recalled words presented before others); Meshed (recalled words meshed with others); SFR (standard free recall). Correct is the number of words correctly recalled, out of 40 on the final trial of the experiment.

a. Generate descriptive statistics. Evaluate these descriptives and print them here.

```
library(carData)

data("Friendly")
summary(Friendly)
```

```
##   condition     correct
##  Before:10   Min.   :21.0
##  Meshed:10   1st Qu.:30.0
##  SFR   :10   Median :37.0
##              Mean   :34.5
##              3rd Qu.:39.0
##              Max.   :40.0
```

```
d <- describeBy(Friendly$correct, group = Friendly$condition)
d
```

```
##
##  Descriptive statistics by group
## group: Before
##    vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 10 36.6 5.34     39   37.75 1.48  24  40    16 -1.4      0.4 1.69
## ------------------------------------------------------------
## group: Meshed
##    vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 10 36.6 3.03   36.5      37 2.97  30  40    10 -0.76    -0.32 0.96
```

```
## ------------------------------------------------------------
## group: SFR
##    vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 10 30.3 7.33     27   30.38 6.67  21  39    18  0.2    -1.94 2.32
```

Task 6. Run a multiple regression model to predict "correct" using dummy coding for the condition variable. Generate 95% confidence intervals of the b estimates. Be sure to consider and/or specify your reference group.

```
Friendly$condition <- relevel(Friendly$condition, ref = "Before")
model <- lm(correct ~ condition, data = Friendly)
summary(model)
```

```
##
## Call:
## lm(formula = correct ~ condition, data = Friendly)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.600  -4.625   0.900   3.400   8.700
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.660e+01  1.746e+00  20.965   <2e-16 ***
## conditionMeshed  1.034e-15  2.469e+00   0.000   1.0000
## conditionSFR    -6.300e+00  2.469e+00  -2.552   0.0167 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.521 on 27 degrees of freedom
## Multiple R-squared:  0.2433, Adjusted R-squared:  0.1873
## F-statistic: 4.341 on 2 and 27 DF,  p-value: 0.02319
```

```
confint(model, level = 0.95)
```

```
##                      2.5 %    97.5 %
## (Intercept)      33.017938 40.182062
## conditionMeshed  -5.065801  5.065801
## conditionSFR    -11.365801 -1.234199
```

a. Report the results here in APA format. Be sure to include the adjusted R2 value, the b estimates, the p-values, and the 95% confidence intervals. What can you conclude from your results?

###The results are showing that the experimental condition significantly affected word recall scores. Participants in the SFR condition recalled less words than the ones in the Before condition. The difference is significant (p=0.017). However, there was no significant difference between the Meshed and Before conditions (b=0,p=1). The R^2 was 0.19 which means that the model explained 19% of the variance in recall scores. In conclusion, the SFR condition negatively affcted the recall performance.###

Task 7. Run a multiple regression model to predict "correct" using unweighted effects coding for the condition variable. Generate 95% confidence intervals of the b estimates. Be sure to consider and/or specify your reference group.

```
Friendly$condition <- factor(Friendly$condition, levels = c("Before", "Meshed", "SFR"))
contrasts(Friendly$condition) <- contr.sum(3)
model <- lm(correct ~ condition + 0, data = Friendly)
summary(model)
```

```
##
```

```
## Call:
## lm(formula = correct ~ condition + 0, data = Friendly)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.600  -4.625   0.900   3.400   8.700
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## conditionBefore   36.600      1.746   20.96  < 2e-16 ***
## conditionMeshed   36.600      1.746   20.96  < 2e-16 ***
## conditionSFR      30.300      1.746   17.36 3.59e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.521 on 27 degrees of freedom
## Multiple R-squared:  0.9776, Adjusted R-squared:  0.9752
## F-statistic: 393.4 on 3 and 27 DF,  p-value: < 2.2e-16
```

```
confint(model, level = 0.95)
```

```
##                    2.5 %   97.5 %
## conditionBefore 33.01794 40.18206
## conditionMeshed 33.01794 40.18206
## conditionSFR    26.71794 33.88206
```

a. What are the differences in the interpretation of your intercept and b estimates when using unweighted effects code versus dummy coding as you used in Task 6?

In task 6, The intercept is the mean of 'before'. The b estimates show how each condition differs from the reference group. Here, the intercept is the overall mean across all conditions. The b estimates show how each condition differs from the overall mean.

b. What is the mean of the base group?

It's 36.6.