ITEC 621 Exercise 2 - Foundations

Descriptive and Predictive Analytics

Dawoon (Kate) Jung

1/28/2021

Table of Contents

[General Instructions 1](#_Toc62754817)

[1. Descriptive Analytics 1](#_Toc62754818)

[2. Basic Predictive Modeling 4](#_Toc62754819)

## General Instructions

In this exercise you will do quick descriptive and predictive analytics to evaluate if the Salaries data set (with professor salaries) supports the **gender pay gap hypothesis**.

First, download the R Markdown template for this exercise **Ex1\_Foundations\_YourLastName.Rmd** and save it with your own last name **exactly**. Then open it in R Studio and complete all the exercises and answer the questions below in the template. Run the code to ensure everything is working fine. When done, upload onto blackboard, knit your R Markdown file into a Word document and upload it into Blackboard. If for some reason you can’t knit a Word file, knit an HTML file and save it as a PDF. Blackboard will not accept HTML files, but will take your PDF.

## 1. Descriptive Analytics

**1.1 Examine the data**

Is there a gender pay gap? Let’s analyze this important question using professor salaries.

Load the library **{car}**, which contains the **Salaries** data set. Then, list the first few records with head(Salaries). The display the summmary() for this dataset, which will show frequencies.

Then, load the library **{psych}** which contains the describe() function and use this function to list the descriptive statistics for the data set.

Then display the median salary grouped by gender using the aggregate() function (feed grouping variables, dataset and aggregate function, i.e., salary ~ sex, Salaries, mean)

library(car)  
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

library(psych)  
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

aggregate(salary ~ sex, Salaries, mean)

## sex salary  
## 1 Female 101002.4  
## 2 Male 115090.4

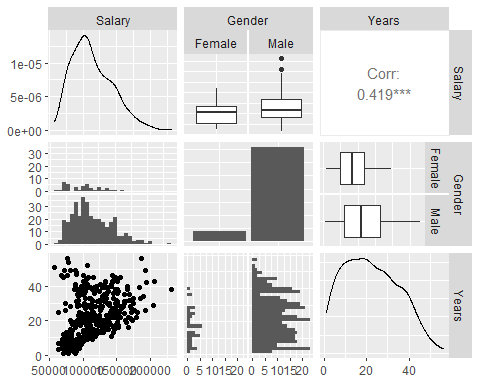
**1.2 Correlation, Boxplots and ANOVA**

Load the library **GGally** and run the **ggpairs()** function on the **salary** (notice that the **Salary** data set is capitalized, whereas the variable **salary** is not), **sex** and **yrs.since.phd** variables (only) in the **Salaries** data set to display some basic descriptives and correlation visually. Please label your variables appropriately (see graph below).

Tips: ggpairs() requires a **data frame**. So you need to use the data.frame() function to bind the necessary column vectors into a data frame (e.g., ggpairs(data.frame("Salary"=Salaries$salary, etc.). Notice the difference in the quality of the graphics and how categorical variables are labeled. Also, add the attribute upper=list(combo='box') at the end to get labels for the boxplot.

Finally, conduct an ANOVA test to evaluate if there is a significant difference between mean salaries for male and female faculty. Feed Salaries$salary ~ Salaries$sex into the aov() function. Embed the aov() function inside the summary() function to see the statistical test results.

library(GGally)  
attach(Salaries)  
  
ggpairs(data.frame("Salary" = Salaries$salary, "Gender" = Salaries$sex, "Years" = Salaries$yrs.since.phd), upper=list(combo='box'))



summary(aov(Salaries$salary ~ Salaries$sex))

## Df Sum Sq Mean Sq F value Pr(>F)   
## Salaries$sex 1 6.980e+09 6.980e+09 7.738 0.00567 \*\*  
## Residuals 395 3.563e+11 9.021e+08   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**1.3 Preliminary Interpretation**

Based on the output above, does it appear to be a gender pay gap? Why or why not. In your answer, please refer to as much of the data above to support your answer.

# ANOVA shows that there is a statistically significant difference in salary across gender. However, in the boxplot of salary and gender, since the two boxes are largely overlapped, it does not support their differences.   
  
# There is another variable that has an association with salary is years since they achieved a Ph.D. degree. The correlation between years and salary is about 0.419, which indicated the two variables are highly correlated. The barplot of Years and Gender shows that the years of Ph.D. of male faculty are higher than the female faculty's years. Therefore, it can be one possible explanation of the gender pay gap.

## 2. Basic Predictive Modeling

**2.1 Salary Gender Gap: Simple OLS Regression**

Suppose that you hypothesize that there is a salary gender pay gap. Fit a linear model function lm() to test this hypothesis by predicting salary using only **sex** as a predictor. Store the results in an object called lm.fit.1, then inspect the results using the summary() function. Do these results support the salary gender gap hypothesis? Briefly explain why.

lm.fit.1 <- lm(salary ~ sex, data = Salaries)  
summary(lm.fit.1)

##   
## Call:  
## lm(formula = salary ~ sex, data = Salaries)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -57290 -23502 -6828 19710 116455   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 101002 4809 21.001 < 2e-16 \*\*\*  
## sexMale 14088 5065 2.782 0.00567 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 30030 on 395 degrees of freedom  
## Multiple R-squared: 0.01921, Adjusted R-squared: 0.01673   
## F-statistic: 7.738 on 1 and 395 DF, p-value: 0.005667

# Yes. The linear regression model shows that on average, when the faculty is male, the salary is higher than female faculty by about $14,088.

**2.2 Multivariate OLS Regression**

Now fit a linear model with **sex** and **yrs.since.phd** as predictors and save it in an object named lm.fit.2. Then inspect the results using the summary() function. Do these results support the salary gender gap hypothesis? Briefly explain why.

lm.fit.2 <- lm(salary ~ sex + yrs.since.phd, data = Salaries)  
summary(lm.fit.2)

##   
## Call:  
## lm(formula = salary ~ sex + yrs.since.phd, data = Salaries)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -84167 -19735 -2551 15427 102033   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 85181.8 4748.3 17.939 <2e-16 \*\*\*  
## sexMale 7923.6 4684.1 1.692 0.0915 .   
## yrs.since.phd 958.1 108.3 8.845 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 27470 on 394 degrees of freedom  
## Multiple R-squared: 0.1817, Adjusted R-squared: 0.1775   
## F-statistic: 43.74 on 2 and 394 DF, p-value: < 2.2e-16

# The evidence is not conclusive. On average, male faculty get paid $7923.6 more than female faculty, holding the years since achieving a Ph.D. degree. However, the p-value is only 0.0915; this effect is only moderately significant.

**2.3 Comparing Models with ANOVA F-Test**

Run an ANOVA test using the anova() funtion to compare **lm.fit.1** to **lm.fit.2**.

anova(lm.fit.1)

## Analysis of Variance Table  
##   
## Response: salary  
## Df Sum Sq Mean Sq F value Pr(>F)   
## sex 1 6.9800e+09 6980014930 7.7377 0.005667 \*\*  
## Residuals 395 3.5632e+11 902077538   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(lm.fit.2)

## Analysis of Variance Table  
##   
## Response: salary  
## Df Sum Sq Mean Sq F value Pr(>F)   
## sex 1 6.9800e+09 6.9800e+09 9.2507 0.002512 \*\*   
## yrs.since.phd 1 5.9031e+10 5.9031e+10 78.2341 < 2.2e-16 \*\*\*  
## Residuals 394 2.9729e+11 7.5454e+08   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**2.4 Interpretation**

Provide your brief conclusions (in no **more than 3 lines**) about whether you think there is a gender pay gap based on this analysis (you will expand this analysis much further in HW2). First, which lm() model is better and why? Then, compare the best predictive model of the two against the descriptive analytics results you obtained in section 1 above. If the null hypothesis is that there is no gender pay gap, is this hypothesis supported? Why or why not?

# With lm.fit.2 model, the null hypothesis is failed to reject due to the lack of significance. In order to reject the null hypothesis, the rule of thumb number of a coefficient is p < 0.05. On the other hand, the lm.fit.1 model rejects the hull hypothesis with statistical significance. It indicates that the predictor yrs.since.phd correlates with the other predictor, gender. As the ANOVA test shows yrs.since.phd is an important prediction revealing the gender pay gap.