ITEC 621 Exercise 2 - Foundations

Descriptive and Predictive Analytics

J. Alberto Espinosa

January 15, 2023

Table of Contents

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 **Ex2_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.

Knitting and Formatting: no or improper knitting and formatting is worth up to **3 points** in this exercise. Once all your R code is working properly, **knit** your R Markdown file into a Word document and upload it into Canvas. If for some reason you can't knit a Word file, you can knit an HTML or PDF file. But please ensure that all your text narratives are fully visible (if I can't see the text I can't grade it). Also, please ensure that your **Table of Contents** is properly formatted.

Note about where to write interpretations: Please write your interpretations in the text area of R Markdown and **DO NOT** use the # or ## tags. These cause your text to appear as headings or sub-headings and show up in the table of contents. I use the # tag, but inside the R code chunks. I write my solutions inside the R code chunk rather than in the text area, so that I can suppress the solution. But you don't need to do this, so write all your narratives in the text areas.

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.

```
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
```

```
summary(Salaries)
```

```
##         rank     discipline yrs.since.phd    yrs.service       sex
##  AsstProf : 67   A:181      Min.   : 1.00   Min.   : 0.00   Female: 39
##  AssocProf: 64   B:216      1st Qu.:12.00   1st Qu.: 7.00   Male  :358
##  Prof     :266              Median :21.00   Median :16.00
##                             Mean   :22.31   Mean   :17.61
##                             3rd Qu.:32.00   3rd Qu.:27.00
##                             Max.   :56.00   Max.   :60.00
##      salary
##  Min.   : 57800
##  1st Qu.: 91000
##  Median :107300
##  Mean   :113706
##  3rd Qu.:134185
##  Max.   :231545
```

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 mean salary grouped by gender using the `aggregate()` function (feed grouping formula first, followed by the dataset **Salaries** and then the aggregate function to apply, i.e., mean.

```
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
```

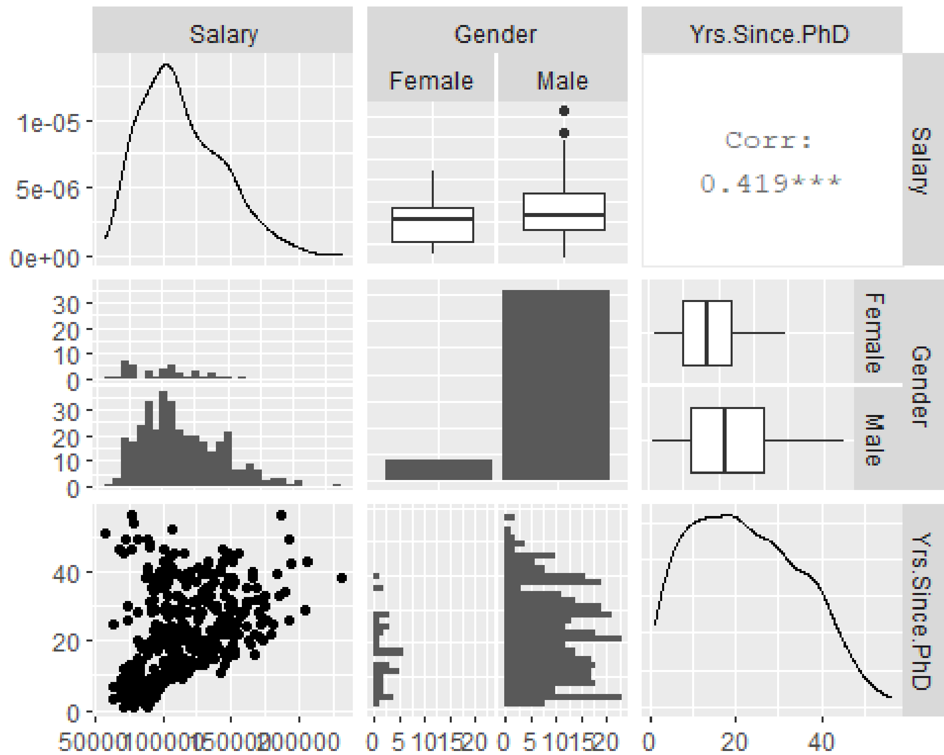
1.2 Correlation, Boxplots and ANOVA

The means by gender above suggest that there may be a gender pay gap at this institution. Let's analyze this visually and statistically. Load the library **GGally** and run the **ggpairs()** function on the **salary**, **sex** and **yrs.since.phd** variables (only) in the **Salaries** data set to display some basic descriptive statistics and correlation, visually. Please note that the **Salary** data set is **capitalized**, whereas the variable **salary** is not. Please also 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')` in the `ggpairs()` function to get labels for the boxplot.

```
require(GGally)

ggpairs(data.frame("Salary" = Salaries$salary,
                   "Gender" = Salaries$sex,
                   "Yrs Since PhD" = Salaries$yrs.since.phd),
                   upper = list(combo = 'box'))
```



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.

```
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.

```
# The mean salary for males is slightly higher than for females. The boxplot
of salary by sex does not appear to support a gender pay gap because the
boxes are largely overlapping. However, the ANOVA test is significant
supporting the gender pay gap argument. At the same time we can also observe
a high correlation between Yrs.Since.Phd and Salary, which is supported by
the corresponding scatter plot. Since male faculty have more years since
obtaining their PhD degrees than females, this may explain why males make
somewhat larger salaires.
```

2. Basic Predictive Modeling

2.1 Salary Gender Gap: Simple OLS Regression

Suppose that you hypothesized that there is a salary gender pay gap.

** Technical Note:** it is more effective to set the null hypothesis to the contrary of what you want to prove, so that you can reject it if not supported.

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.

```
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
```

Do these results support the salary gender gap hypothesis? Briefly explain why.

```
# Yes, the sexMale coefficient is positive and significant, so based on this
data set, on average, male faculty make about $14K more than female faculty.
```

2.2 Multivariate OLS Regression

Now fit a 2-predictor linear model (quantitative + dummy variable) with **yrs.since.phd** and **sex** as predictors, and save it in an object named `lm.fit.2`. Then inspect the results using the `summary()` function.

```
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
```

Do these results support the salary gender gap hypothesis? Briefly explain why.

```
# The evidence is not conclusive. Controlling for years since obtaining a PhD
degree, on average, male faculty make about $7,923 more than female faculty,
but this effect is only significant at the p=0.0915 level, not at the p <
0.05 level.
```

2.3 Comparing Models with ANOVA F-Test

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

```
anova(lm.fit.1, lm.fit.2)
```

```
## Analysis of Variance Table
##
## Model 1: salary ~ sex
## Model 2: salary ~ sex + yrs.since.phd
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    395 3.5632e+11
## 2    394 2.9729e+11  1 5.9031e+10 78.234 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.4 Interpretation

Provide your brief conclusions (in **6 lines** or so) about whether you think there is a gender pay gap based on this analysis (you will expand this analysis much further in HW2). First, based on the Anova test above, which `lm()` model is better and why? Then, compare the best predictive model of the two against the descriptive analytics results you obtained in 1.2 above. If the null hypothesis is that there is no gender pay gap, is this hypothesis supported? Why or why not?

```
# The ANOVA test is significant so the larger model, lm.fit2 is a better
model than the smaller model lm.fit.1.

# Descriptive analytics (ANOVA) suggested that there is a gender pay gap. But
these results are only preliminary. In order to have more substantive
statistical support and test the gender pay gap hypothesis properly, we need
to analyze predictive models with the appropriate control variables.

# The lm.fit.1 model supports the gender salary gap hypothesis at a
significance level of p < 0.05, with males making more than females. Since
this is a simple regression model with just one predictor, it is not
surprising that we get the same results as with the ANOVA test.

# However, the results with the better model lm.fit.2 also show a positive
coefficient, with males making more than females, but the coefficient is only
marginally significant with p = 0.0915. It looks like the number of years
since their PhD degree was earned (i.e., experience) is a stronger predictor,
which is what we observed in ggpairs().

# Insight: when 2 predictors are correlated and you omit one, it causes the
included variable to be biased (it picks up some of the effect of the omitted
predictor). It looks like there are many more male professors than female
professors with many years of tenure. The ANOVA test is consistent with the
regression results. Adding yrs.since.phd improves the explanatory power of
the model significantly.
```