Data Exercise 3

Devin Judge-Lord

April 11, 2019

Women at the Deer Valley Utility Company claim that their job performances are not rewarded to the same degree as the job performances of men. Is there statistical evidence to support this complaint? This summary analysis for the Director of the Office of Equal Opportunity includes findings and a brief discussion of other factors we may want to investigate before issuing a final report.

We have the following data for 60 people:

`Salary`: thousands of dollars.

`Gender`: "1" for men and "0" for women.

`Rating`: The employee's average performance rating over the last two years. The scale has a top score of 100. The company claims that performance rating is the primary factor in the determination of salary.

`Credits` earned either in college courses or company programs.

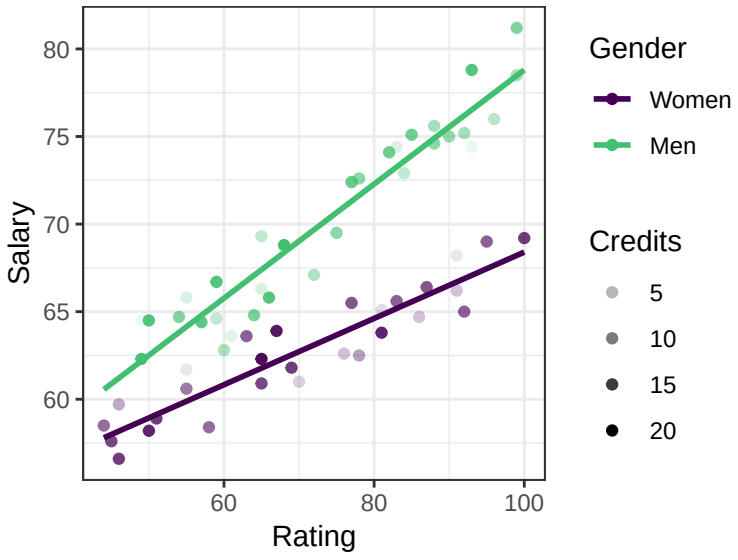
Retrieve data from *Stata* and load them into **R**.

```
net install PS813_EX3, from(https://faculty.polisci.wisc.edu/weimer)
PS813_EX3 1234
save "data/EX3.dta"
```

```
d <- read_dta("data/EX3.dta") %>% zap_formats() %>%
  mutate(Gender = factor(Sex, labels = c("Women","Men")))
```

First, let us examine the raw data.

```
# scatterplot
p <- ggplot(d) +
  aes(x = Rating, y = Salary, color = Gender) +
  geom_point(aes(alpha = Credits)) + scale_color_discrete()

# quick y ~ mx + b linear regression per group
p + geom_smooth(method = lm, se = F, fullrange = T)
```



Hypotheses

H1: Job performances of women are rewarded differently than the job performances of men. That is, the relationship between salary and performance differs by gender.

H0: There is no difference in how men's performance and women's performance are rewarded. That is, the relationship between salary and performance does not differ by gender.

(There are least two other ways to write this hypothesis and at least one slightly different hypothesis that might better address the question.)

Model

The dependent variable is salary. For employee i, let their salary be y_i in the model $y_i = \beta_0 + ... + \epsilon_i$. β_0 is the predicted salary, $\hat{y}$, when all other variables in the model are 0.

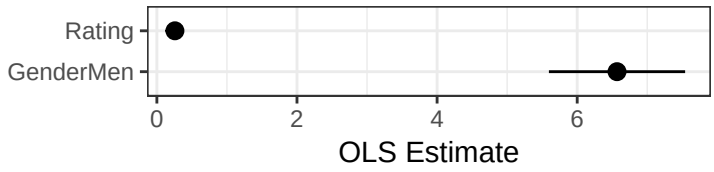
Note: The model, $y_i = \beta_0 + \beta_1 * Gender_i + \beta_2 * Rating_i + \epsilon_i$, does *not* test the relationship of interest.

```
model_1 <- lm(Salary ~ Gender + Rating, data = d)
summary(model_1)
```

```
##
## Call:
## lm(formula = Salary ~ Gender + Rating, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8982 -1.5364  0.0581  1.3178  4.4643
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.68752    1.10151   40.57   <2e-16 ***
## GenderMen    6.56843    0.48546   13.53   <2e-16 ***
## Rating       0.25737    0.01485   17.33   <2e-16 ***
```

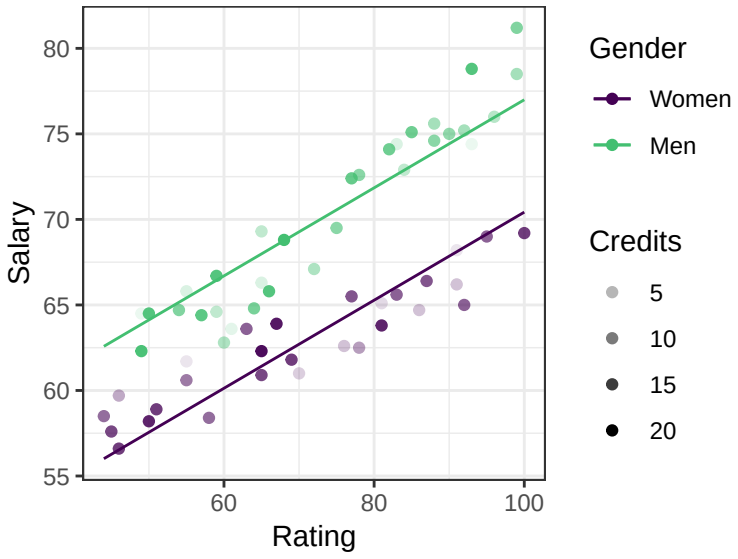
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.866 on 57 degrees of freedom
## Multiple R-squared:  0.904,  Adjusted R-squared:  0.9006
## F-statistic: 268.4 on 2 and 57 DF,  p-value: < 2.2e-16
```

```r
m1 <- model_1 %>%
  tidy(conf.int = TRUE)

ggplot(m1 %>% filter(term != "(Intercept)")) +
  aes(x = term,
      y = estimate,
      ymin = conf.low,
      ymax = conf.high) +
  geom_pointrange() +
  coord_flip() +
  labs(x="", y="OLS Estimate")
```



As always, let's plot the results against our data!

```r
# illustrating with yhat formula; more easily done with augment()
b0 <- m1$estimate[1]
b1 <- m1$estimate[2]
b2 <- m1$estimate[3]

p +
  geom_line(aes(color = "Men", # yhat for men
                y = b0 + b1*1 + b2*Rating) ) +
  geom_line(aes(color = "Women", # yhat for women
                y = b0 + b1*0 + b2*Rating) )
```



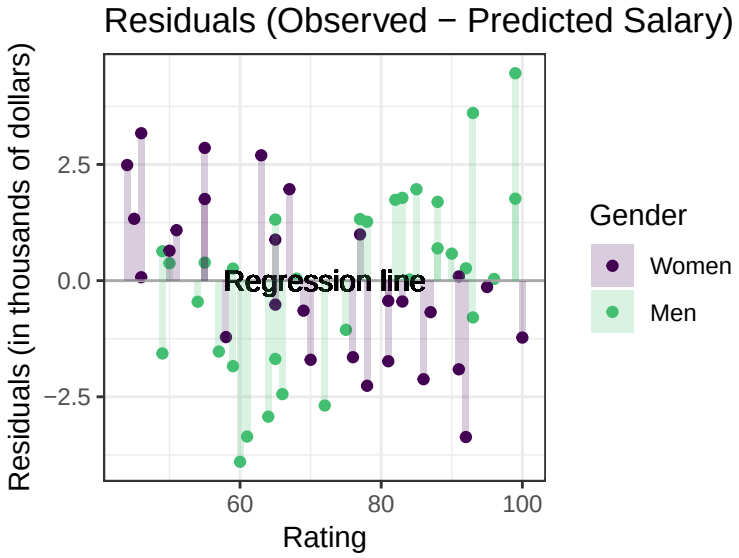
Interpretation: Why does this model fail to test the hypothesis? What hypothesis did it test? How should we interpret the coefficient of 6.6 on Gender?

Fit: Let's also plot the residuals. Aside from interpretation, we want to know where (e.g. for what performance ratings) our model fits the data and where it does not, especially if residuals seem to vary systematically over the range of our data.

`augment` computs tidy residuals, among other cool things.

```
m1 <- lm(Salary ~ Gender + Rating, data = d) %>%
  augment()

ggplot(m1) +
  aes(y = .resid, x = Rating) +
  geom_point(aes(color = Gender)) +
  scale_color_discrete() +
  ## to show how risiduals are the distance between an
  ## observation and the regression line:
  geom_hline(yintercept = 0, color = "dark grey") +
  geom_text(x= mean(m1$Rating), y = 0,
            label = "Regression line") +
  geom_col(aes(fill = Gender), alpha = .2, position ="identity") +
  ## + labels:
  labs(title = "Residuals (Observed - Predicted Salary)",
```

```
       y = "Residuals (in thousands of dollars)")
```



Hypothesis test

Lorem ipsum $\beta_? = 0$

Lorem ipsum $\beta_? \neq 0$

Findings

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.