esm-206-assignment-5-q4

Sean Denny

11/29/2018

Question 4

```
library(tidyverse)
prof_sal <- read_csv("Faculty salary data (2008 - 2009 survey).csv")
```

Rename discipline column:

```
prof_sal <- prof_sal %>%
  rename(discipline = "discipline_(A=theoretical_B=applied)")
```

1. Explore the data

```
#rank

salary_rank <- prof_sal %>%
  group_by(rank) %>%
  summarize(
    mean = mean(salary)
)

#As expected.

#discipline; note: A = theoretical, B=applied

salary_discipline <- prof_sal %>%
  group_by(discipline) %>%
  summarize(
    mean = mean(salary)
  )

#Average of applied salaries is higher.

#years_since_phd

salary_years_phd <- prof_sal %>%
  group_by(years_since_phd) %>%
  summarize(
    mean = mean(salary)
  )

#Generally increases, but not consistent.

#years_faculty

salary_years_faculty <- prof_sal %>%
```

```
  group_by(years_faculty) %>%
  summarize(
    mean = mean(salary)
  )

#Appears to genereally increase, but not consistent.

#sex

salary_sex <- prof_sal %>%
  group_by(sex) %>%
  summarize(
    mean = mean(salary)
  )

#Average of males is higher.
```

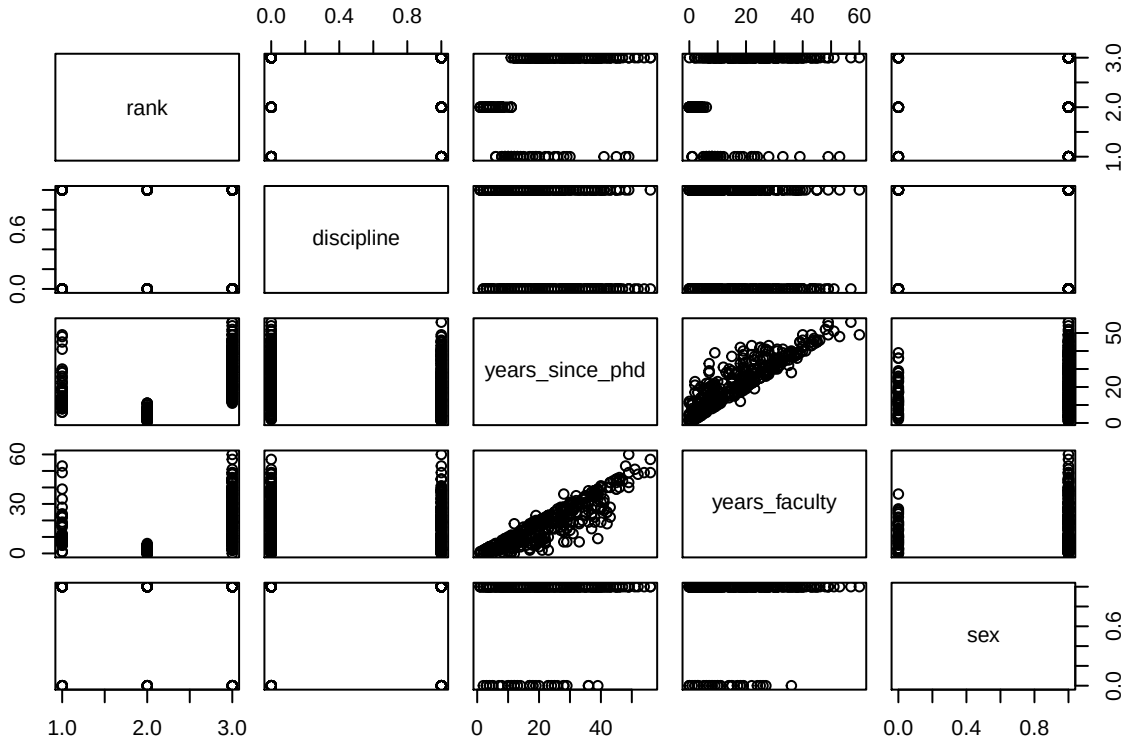
2. Test for corellations

1. Coerce character variables into factors then numeric (i.e. change the class of the character data)
2. Test for correlations among all of the explanatory variables. Can we do this if the variables are factors, or do we absolutely have to change them to numeric? **Remember to create a new dataset without the response variable (salary) so you can run pairs() or cor() for the explanatory variables only.**

```
cor_data <- select(prof_sal, -salary) %>%
  mutate(rank = as.factor(rank)) %>%
  mutate(rank = as.numeric(rank)) %>%
  mutate(discipline = as.factor(discipline)) %>%
  mutate(discipline = as.numeric(discipline)-1) %>%
  mutate(sex = as.factor(sex)) %>%
  mutate(sex = as.numeric(sex)-1)

cor(cor_data)
```

```
##                       rank   discipline years_since_phd years_faculty
## rank             1.00000000 -0.086266163       0.5255004     0.4474990
## discipline      -0.08626616  1.000000000      -0.2180873    -0.1645987
## years_since_phd  0.52550037 -0.218087325       1.0000000     0.9096491
## years_faculty    0.44749898 -0.164598697       0.9096491     1.0000000
## sex              0.13249244  0.003723739       0.1487878     0.1537396
##                         sex
## rank            0.132492439
## discipline      0.003723739
## years_since_phd 0.148787792
## years_faculty   0.153739575
## sex             1.000000000
```

```
pairs(cor_data)
```




```
#Correlation for years_since_phd and rank = 0.53
#years_since_phd and years_faculty = 0.91
```

Before running the model, re-level the levels in rank so that Assistant Professor is the reference level. This isn't necessary but slightly easier for interpretation.

```
prof_sal <- prof_sal %>%
  mutate(rank = as.factor(rank))

prof_sal$rank <- fct_relevel(prof_sal$rank, "AsstProf")
```

3. Build the model

We're going to remove years_since_phd, which removes correlations with both years_faculty and rank. Also, these two variables (the ones we're keeping) are more interesting.

Use lm() with the following syntax: lm(y ~ x1 + x2 + x3..., data = df_name)

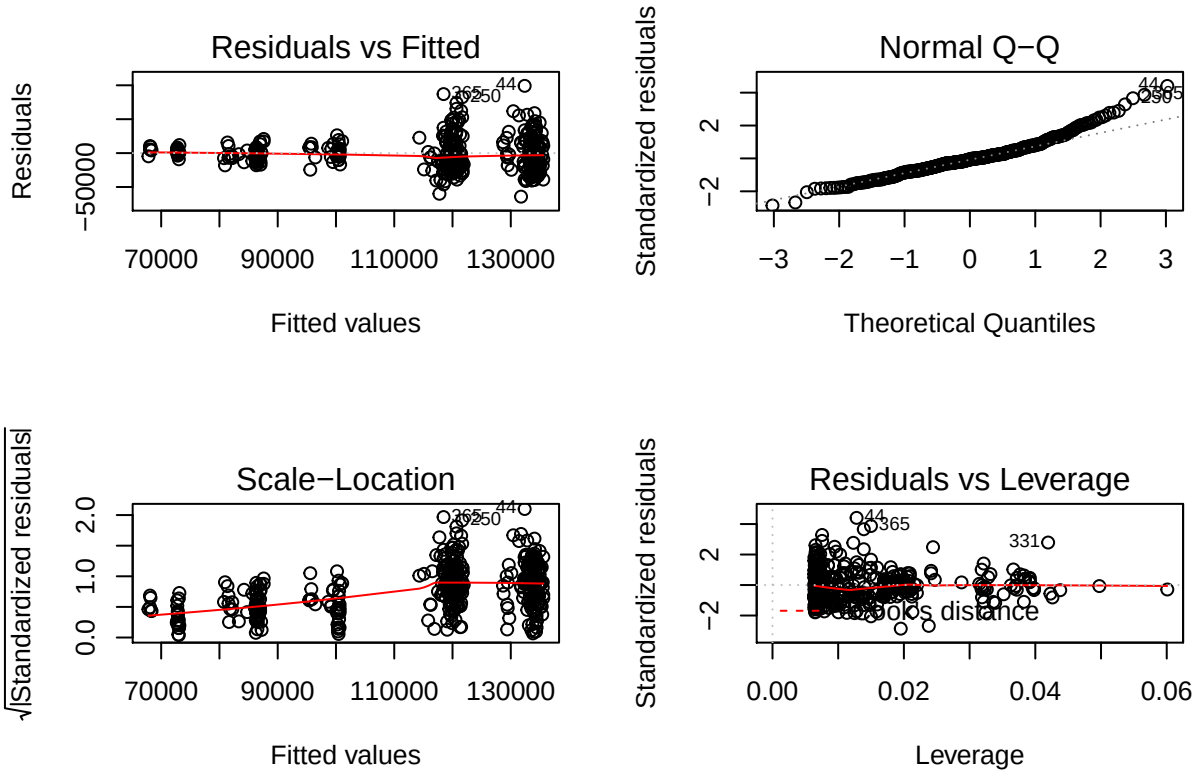
```
prof_lm_1 <- lm(salary ~ sex + discipline + rank + years_faculty, data = prof_sal)

summary(prof_lm_1)
```

```
##
## Call:
## lm(formula = salary ~ sex + discipline + rank + years_faculty,
##     data = prof_sal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -64202 -14255  -1533  10571  99163
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   68351.67    4482.20  15.250  < 2e-16 ***
## sexMale        4771.25    3878.00   1.230 0.219311
## disciplineB   13473.38    2315.50   5.819 1.24e-08 ***
## rankAssocProf 14560.40    4098.32   3.553 0.000428 ***
## rankProf      49159.64    3834.49  12.820  < 2e-16 ***
## years_faculty   -88.78     111.64  -0.795 0.426958
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22650 on 391 degrees of freedom
## Multiple R-squared:  0.4478, Adjusted R-squared:  0.4407
## F-statistic: 63.41 on 5 and 391 DF,  p-value: < 2.2e-16
```

Adjusted R-squared is 0.4407. Meaning the variance in salary is not particularly well explained by variance in the... model output? explanatory variables?

```
par(mfrow = c(2,2))
plot(prof_lm_1)
```



Problems with homescedasticity. The data appear heteroscedastic. *Also, data isn't normally distributed at upper values?*

Is there a need to try other models? The variables we included are all interesting and make sense to include.

4. Figures

```
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```