P8130_final_project

Michael Yan

11/15/2019

```
# tidying data and adding a column with the average salary
law_data = read_csv("./data/Lawsuit.csv") %>%
  janitor::clean_names() %>%
  mutate(dept = recode_factor(dept,
                              "1" = "Biochemistry/Molecular Biology",
                              "2" = "Physiology",
                              "3" = "Genetics",
                              "4" = "Pediatrics",
                              "5" = "Medicine",
                              "6" = "Surgery"),
         gender = recode_factor(gender,
                                "1" = "Male",
                                "0" = "Female"),
         clin = recode_factor(clin,
                              "1" = "Primarily clinical emphasis",
                              "0" = "Primarily research emphasis"),
         cert = recode_factor(cert,
                              "1" = "Board certified",
                              "0" = "Not certified"),
         rank = recode_factor(rank,
                              "1" = "Assistant",
                              "2" = "Associate",
                              "3" = "Full professor"),
         avg_salary = (sal94 + sal95) / 2)
```

```
# data exploration pt. 1
my_labels  =  list(dept = "Dept, n(%)",
                   clin = "Clin, n(%)",
                   cert = "Clin, n(%)",
                   prate = "Prate",
                   exper = "Exper",
                   rank = "Rank, n(%)",
                   sal94 = "Sal94",
                   sal95 = "Sal95",
                   avg_salary = "Average Salary")

my_controls = tableby.control(
               total = F,
               test = F,
               numeric.stats = c("meansd", "medianq1q3"),
               digits = 2,
               digits.pct = 2)

table1 = tableby(gender ~ dept + clin + cert + prate + exper + rank + sal94 + sal95 + avg_salary, data =

summary(table1, labelTranslations = my_labels,
        title = "Demographics and co-morbidities ", text = T) %>%
```

```
knitr::kable()
```

	Male (N=155)	Female (N=106)
Dept, n(%)		
- Biochemistry/Molecular Biology	30 (19.35%)	20 (18.87%)
- Physiology	20 (12.90%)	20 (18.87%)
- Genetics	10 (6.45%)	11 (10.38%)
- Pediatrics	10 (6.45%)	20 (18.87%)
- Medicine	50 (32.26%)	30 (28.30%)
- Surgery	35 (22.58%)	5 (4.72%)
Clin, n(%)		
- Primarily clinical emphasis	100 (64.52%)	60 (56.60%)
- Primarily research emphasis	55 (35.48%)	46 (43.40%)
Clin, n(%)		
- Board certified	118 (76.13%)	70 (66.04%)
- Not certified	37 (23.87%)	36 (33.96%)
Prate		
- Mean (SD)	4.65 (1.94)	5.35 (1.89)
- Median (Q1, Q3)	4.00 (3.10, 6.70)	5.25 (3.73, 7.27)
Exper		
- Mean (SD)	12.10 (6.70)	7.49 (4.17)
- Median (Q1, Q3)	10.00 (7.00, 15.00)	7.00 (5.00, 10.00)
Rank, n(%)		
- Assistant	43 (27.74%)	69 (65.09%)
- Associate	43 (27.74%)	21 (19.81%)
- Full professor	69 (44.52%)	16 (15.09%)
Sal94		
- Mean (SD)	177338.76 (85930.54)	118871.27 (56168.01)
- Median (Q1, Q3)	155006.00 (109687.00, 231501.50)	108457.00 (75774.50, 143096.00)
Sal95		
- Mean (SD)	194914.09 (94902.73)	130876.92 (62034.51)
- Median (Q1, Q3)	170967.00 (119952.50, 257163.00)	119135.00 (82345.25, 154170.50)
Average Salary		
- Mean (SD)	186126.43 (90397.11)	124874.09 (59089.62)
- Median (Q1, Q3)	162987.00 (114612.50, 244332.25)	113706.00 (79059.88, 148401.12)

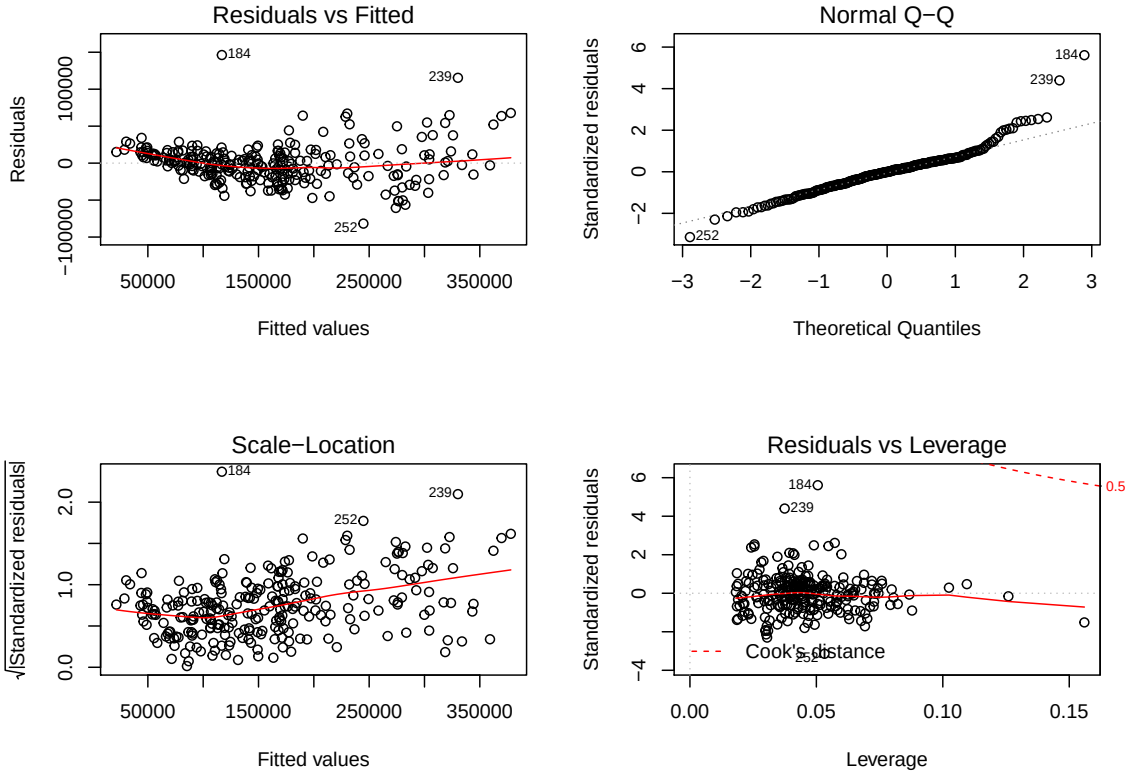
Assumptions about residuals: 1. Normally Distributed 2. They have the same variance at every predictor (Homoscedasticity) 3. They are independent of one another

```
# data exploration pt. 2
# investigate the shape of the distribution for variable 'avg_salary' and try different transformations
law_model = lm(avg_salary ~ dept + clin + cert + prate + exper + rank + avg_salary, data = law_data)
summary(law_model)
```

```
##
## Call:
## lm(formula = avg_salary ~ dept + clin + cert + prate + exper +
##     rank + avg_salary, data = law_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -81806 -15581   -201  12484 146200
##
```

```
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       85421.0    19981.1   4.275 2.72e-05 ***
## deptPhysiology                   -12363.0     5821.4  -2.124 0.034679 *
## deptGenetics                      24053.5     7728.1   3.112 0.002072 **
## deptPediatrics                    23648.9    10461.9   2.260 0.024656 *
## deptMedicine                      77483.5     8984.0   8.625 7.65e-16 ***
## deptSurgery                      179580.1    12289.2  14.613  < 2e-16 ***
## clinPrimarily research emphasis -18736.8     8057.7  -2.325 0.020858 *
## certNot certified               -20185.1     4250.2  -4.749 3.45e-06 ***
## prate                            -2621.2     3371.4  -0.777 0.437617
## exper                             3133.7      360.7   8.687 5.01e-16 ***
## rankAssociate                    17412.3     4645.8   3.748 0.000222 ***
## rankFull professor               34684.7     5151.8   6.732 1.14e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26750 on 249 degrees of freedom
## Multiple R-squared:  0.9043, Adjusted R-squared:  0.9001
## F-statistic: 213.9 on 11 and 249 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))
plot(law_model)
```



Based on the residual vs. fitted values plot, there is a pattern and one can't say that the points are evenly distributed along the residuals = 0 dash line. This indicates that there is going to be a transformation of the average salary.

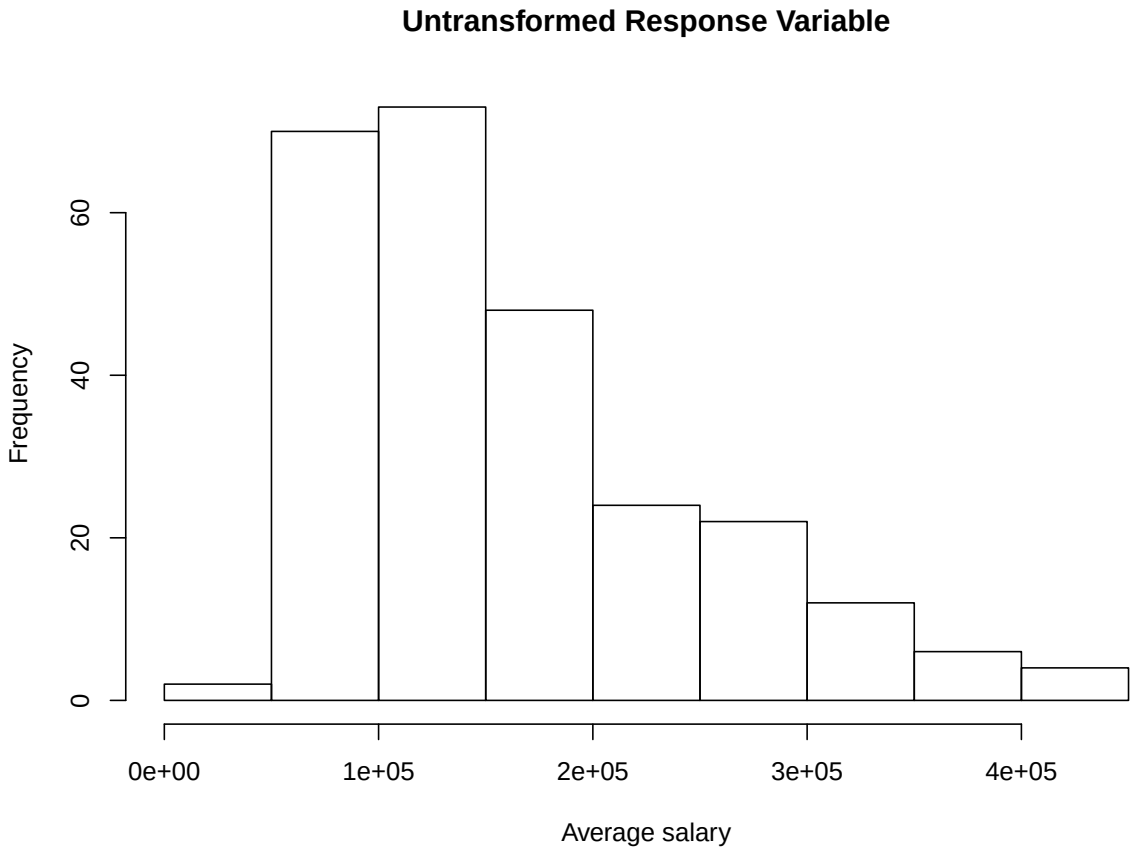
The normal Q-Q plot further emphasizes this decision since a number of data points on the near the tail on the right side are not aligned with the dash line, indicating the residuals are not normal.

The scale-location plot is used to test for homoscedasticity and since points are approximately evenly spread around the line, we can conclude that variance of residuals are approximately constant across the range of all predictor variables.

There is presence of point close to the boundary, dashed-line of the Cook's distance, those have to be tested as potential outliers and a decision have to be made about their relevance to the overal population.

- Yeo-Johnson: Test to Determine the best Transformation for average salary.

```
par(mfrow = c(1, 1))
hist(law_data$avg_salary,
     main = "Untransformed Response Variable",
     xlab = "Average salary")
```



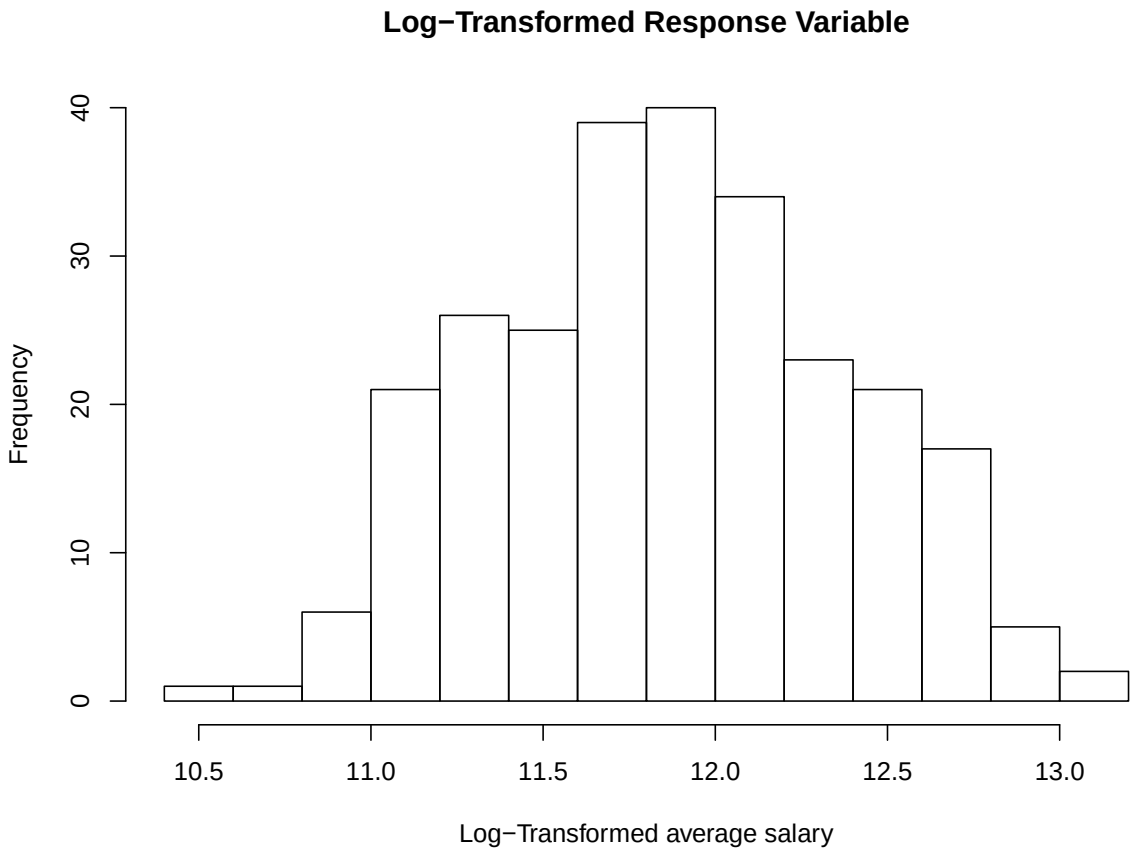
Based on the graph we see an apparent right skewness. Therefore, we are not able to use BoxCox as it does not transform negative response variables. Hence we will use the yeojohnson function from the bestNormalize package to determine the power (Lamda) at which the outcome variable needs to be raised.

```
yeojohnson(law_data$avg_salary)
```

```
## Standardized Yeo-Johnson Transformation with 261 nonmissing obs.:
##  Estimated statistics:
##  - lambda = -0.06327105
##  - mean (before standardization) = 8.339307
##  - sd (before standardization) = 0.2400873
```

Therefore, I will round up the Lamda and use the Log transformation of the response variable.

```
hist(log(law_data$avg_salary),
main = "Log-Transformed Response Variable",
xlab = "Log-Transformed average salary")
```



After doing log-transformation, the histogram shows normal distribution of the response variable, in this case, the log-transformed average salary.

- Building a new Multiple Regression Model

```
law_log_model = lm(log(avg_salary) ~ dept + clin + cert + prate + exper + rank + avg_salary, data = law

summary(law_log_model)
```

```
##
## Call:
## lm(formula = log(avg_salary) ~ dept + clin + cert + prate + exper +
##     rank + avg_salary, data = law_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39664 -0.05235  0.00430  0.06132  0.38146
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1.122e+01  7.156e-02 156.734  < 2e-16 ***
## deptPhysiology                   -1.321e-01  2.030e-02  -6.504 4.28e-10 ***
## deptGenetics                      7.874e-02  2.723e-02   2.892  0.00417 **
## deptPediatrics                    5.615e-02  3.653e-02   1.537  0.12558
## deptMedicine                      2.107e-01  3.539e-02   5.953 8.92e-09 ***
## deptSurgery                       2.171e-01  5.790e-02   3.750  0.00022 ***
## clinPrimarily research emphasis -8.879e-02  2.815e-02  -3.153  0.00181 **
## certNot certified                -1.188e-01  1.534e-02  -7.746 2.43e-13 ***
## prate                            -1.640e-02  1.167e-02  -1.406  0.16111
```

```
## exper                           7.008e-03  1.423e-03   4.924 1.55e-06 ***
## rankAssociate                   7.383e-02  1.651e-02   4.473 1.18e-05 ***
## rankFull professor              9.961e-02  1.936e-02   5.145 5.45e-07 ***
## avg_salary                      3.608e-06  2.191e-07  16.472  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09246 on 248 degrees of freedom
## Multiple R-squared:  0.9685, Adjusted R-squared:  0.967
## F-statistic: 635.6 on 12 and 248 DF,  p-value: < 2.2e-16
```