report

Xinyu Shen xs2384

2019/12/7

```
lawsuit =
read_csv("data/Lawsuit.csv") %>%
   janitor::clean_names() %>%
  mutate(dept = factor(dept,levels = c(1:6),
                       labels =
                     c("Biochemistry","Physiology","Genetics",
                       "Pediatrics","Medicine","Surgery")),
         gender = factor(gender,levels = c(0:1),
                       labels =
                     c("Female","Male")),
         clin = factor(clin,levels = c(0:1),
                       labels =
                     c("Research","Clinical")),
         cert = factor(cert,levels = c(0:1),
                       labels =
                     c("Not certified","Broad certified")),
         rank = factor(rank,levels = c(1:3),
                       labels =
                     c("Assistant","Associate","Full professor")))
```

Summarize all variables by gender

```
 sum_data  <-  arsenal::tableby( gender ~ dept + clin + cert +
                                 prate + exper + rank + sal94   +
                                 sal95,
                                data  = lawsuit,
                                test  = FALSE,
                                total = FALSE,
                                numeric.stats =
                                  c("meansd","medianq1q3","range"))
summ = summary(sum_data,text = TRUE)
summ
```

```
##
##
## |                  |          Female (N=106)          |            Male (N=155)            |
## |:-----------------|:---------------------------------:|:----------------------------------:|
## |dept              |                                   |                                    |
## |-  Biochemistry   |            20 (18.9%)             |             30 (19.4%)             |
## |-  Physiology     |            20 (18.9%)             |             20 (12.9%)             |
## |-  Genetics       |            11 (10.4%)             |              10 (6.5%)             |
## |-  Pediatrics     |            20 (18.9%)             |              10 (6.5%)             |
## |-  Medicine       |            30 (28.3%)             |             50 (32.3%)             |
## |-  Surgery        |             5 (4.7%)              |             35 (22.6%)             |
## |clin              |                                   |                                    |
## |-  Research       |            46 (43.4%)             |             55 (35.5%)             |
## |-  Clinical       |            60 (56.6%)             |             100 (64.5%)            |
## |cert              |                                   |                                    |
```

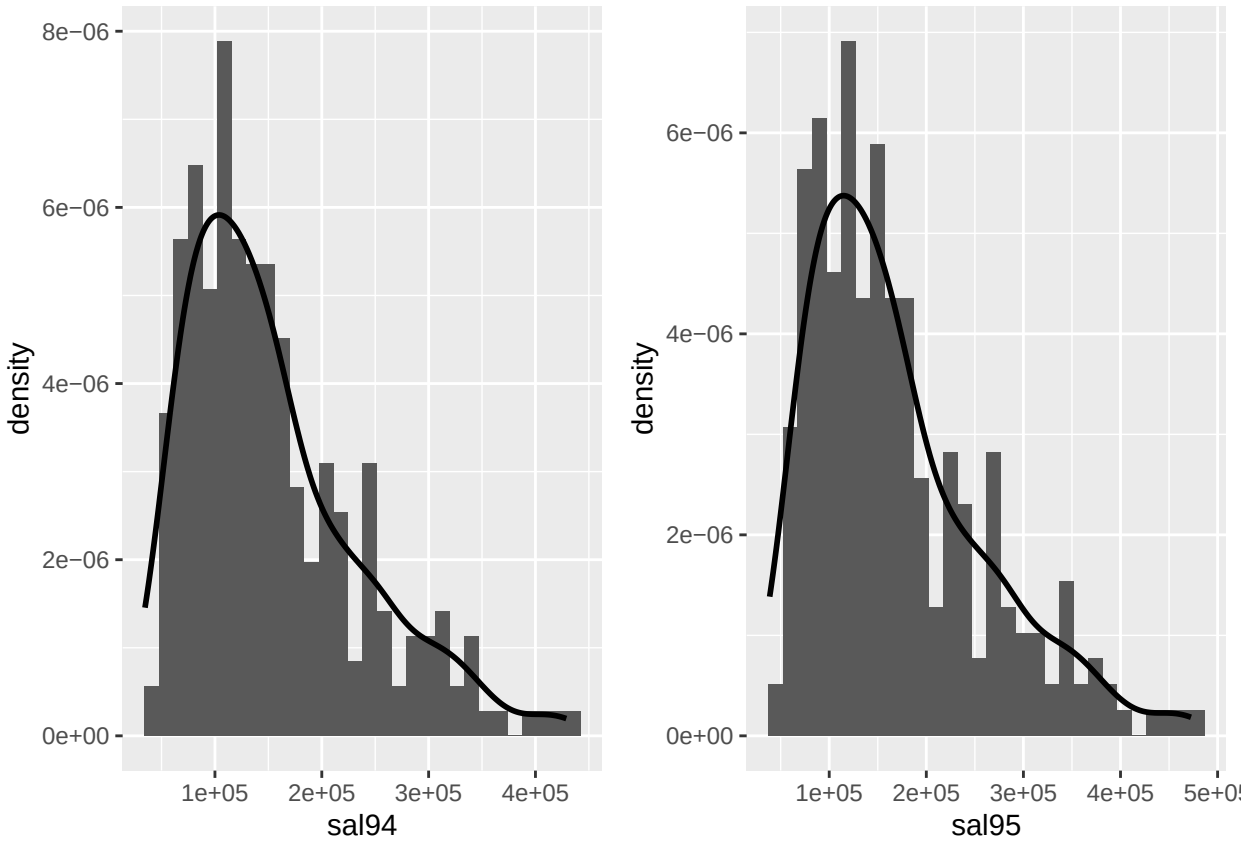
```
## |-  Not certified   |             36 (34.0%)             |             37 (23.9%)             |
## |-  Broad certified |             70 (66.0%)             |             118 (76.1%)            |
## |prate              |                                    |                                    |
## |-  Mean (SD)       |           5.350 (1.886)            |            4.646 (1.938)           |
## |-  Median (Q1, Q3) |        5.250 (3.725, 7.275)        |        4.000 (3.100, 6.700)        |
## |-  Range           |           2.400 - 8.700            |            1.300 - 8.600           |
## |exper              |                                    |                                    |
## |-  Mean (SD)       |           7.491 (4.166)            |           12.103 (6.704)           |
## |-  Median (Q1, Q3) |       7.000 (5.000, 10.000)        |       10.000 (7.000, 15.000)       |
## |-  Range           |           1.000 - 23.000           |            2.000 - 37.000          |
## |rank               |                                    |                                    |
## |-  Assistant       |             69 (65.1%)             |             43 (27.7%)             |
## |-  Associate       |             21 (19.8%)             |             43 (27.7%)             |
## |-  Full professor  |             16 (15.1%)             |             69 (44.5%)             |
## |sal94              |                                    |                                    |
## |-  Mean (SD)       |       118871.274 (56168.006)       |       177338.761 (85930.540)       |
## |-  Median (Q1, Q3) | 108457.000 (75774.500, 143096.000) | 155006.000 (109687.000, 231501.500) |
## |-  Range           |       34514.000 - 308081.000       |       52582.000 - 428876.000       |
## |sal95              |                                    |                                    |
## |-  Mean (SD)       |       130876.915 (62034.507)       |       194914.090 (94902.728)       |
## |-  Median (Q1, Q3) | 119135.000 (82345.250, 154170.500) | 170967.000 (119952.500, 257163.000) |
## |-  Range           |       38675.000 - 339664.000       |       58923.000 - 472589.000       |
```

Distributions

```
gg_94 =
lawsuit %>%
 ggplot(aes(sal94,..density..))+
 geom_histogram()+
 geom_line(stat = 'density',size = 1)+
    labs(x = "sal94")

gg_95 =
lawsuit %>%
 ggplot(aes(sal95,..density..))+
 geom_histogram()+
 geom_line(stat = 'density',size = 1)+
    labs(x = "sal95")

gg_94 + gg_95
```



The distribution for outcome is right skew. So we may want to try the log transformation.

Possible transformation

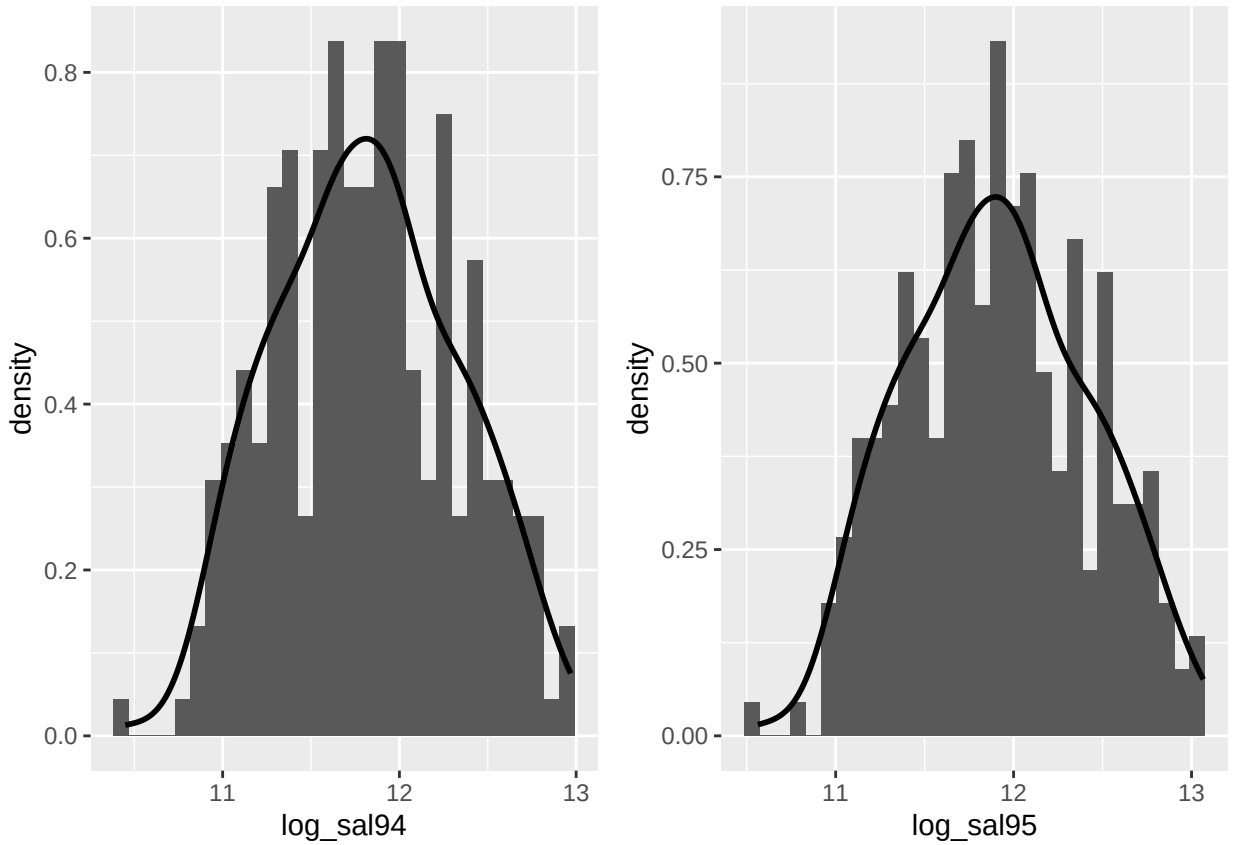
```
lawsuit_log = lawsuit %>% mutate(
        log_sal94 = log(sal94),
        log_sal95 = log(sal95)) %>% dplyr::select(-sal94,-sal95)

gg_94 =
lawsuit_log %>%
 ggplot(aes(log_sal94,..density..))+
 geom_histogram()+
 geom_line(stat = 'density',size = 1)+
    labs(x = "log_sal94")

gg_95 =
lawsuit_log %>%
 ggplot(aes(log_sal95,..density..))+
 geom_histogram()+
 geom_line(stat = 'density',size = 1)+
    labs(x = "log_sal95")

gg_94 + gg_95
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



After using log transformation, the outcome almost follow a normal distribution.

Interaction

Interaction for 94

```
lawsuit_log_94 =
lawsuit_log %>%
  dplyr::select(-log_sal95,-id)

lawsuit_log_95 = lawsuit_log %>%
  dplyr::select(-log_sal94,-id)

bind_rows(lm(log_sal94 ~ gender*dept, data = lawsuit_log_94) %>% summary() %>% .$coefficients %>% as.da
lm(log_sal94 ~ gender*clin, data = lawsuit_log_94) %>% summary() %>% .$coefficients %>% as.data.frame()
lm(log_sal94 ~ gender*cert, data = lawsuit_log_94) %>% summary() %>% .$coefficients %>% as.data.frame()
lm(log_sal94 ~ gender*prate, data = lawsuit_log_94) %>% summary() %>% .$coefficients %>% as.data.frame(
lm(log_sal94 ~ gender*exper, data = lawsuit_log_94) %>% summary() %>% .$coefficients %>% as.data.frame(
lm(log_sal94 ~ gender*rank, data = lawsuit_log_94) %>% summary() %>% .$coefficients %>% as.data.frame()
```

```
##                          rowname   Estimate Std. Error   t value   Pr(>|t|)
## 1 genderMale:rankFull professor -0.4055985  0.1577672 -2.570867 0.01071249
```

Interaction for 95

```
bind_rows(lm(log_sal95 ~ gender*dept, data = lawsuit_log_95) %>% summary() %>% .$coefficients %>% as.da
lm(log_sal95 ~ gender*clin, data = lawsuit_log_95) %>% summary() %>% .$coefficients %>% as.data.frame()
lm(log_sal95 ~ gender*cert, data = lawsuit_log_95) %>% summary() %>% .$coefficients %>% as.data.frame()
lm(log_sal95 ~ gender*prate, data = lawsuit_log_95) %>% summary() %>% .$coefficients %>% as.data.frame()
lm(log_sal95 ~ gender*exper, data = lawsuit_log_95) %>% summary() %>% .$coefficients %>% as.data.frame()
lm(log_sal95 ~ gender*rank, data = lawsuit_log_95) %>% summary() %>% .$coefficients %>% as.data.frame()
```

```
##                          rowname   Estimate Std. Error   t value  Pr(>|t|)
## 1 genderMale:rankFull professor -0.4057257   0.158384 -2.561658 0.0109932
```

From the result above, we can see that "rank" is the interaction term for gender in 1994 and 1995.

Confounders

Confounders for 94

```
con = lm(log_sal94 ~ gender, data = lawsuit_log_94) %>% summary()
con_1 = lm(log_sal94 ~ gender + dept, data = lawsuit_log_94) %>% summary()
con_2 = lm(log_sal94 ~ gender + clin, data = lawsuit_log_94) %>% summary()
con_3 = lm(log_sal94 ~ gender + cert, data = lawsuit_log_94) %>% summary()
con_4 = lm(log_sal94 ~ gender + prate, data = lawsuit_log_94) %>% summary()
con_5 = lm(log_sal94 ~ gender + exper, data = lawsuit_log_94) %>% summary()
con_6 = lm(log_sal94 ~ gender + rank, data = lawsuit_log_94) %>% summary()

con_tab_94 = tibble("variables" = c("gender", "gender + dept", "gender + clin", "gender + cert", "gende

con_tab_94 %>% mutate(
  diff = abs((coef[1]-coef)/coef[1]),
  confounder = ifelse(diff>=0.1, "Y", "N")
)
```

```
## # A tibble: 7 x 4
##   variables       coef   diff confounder
##   <chr>          <dbl>  <dbl> <chr>
## 1 gender         0.386 0      N
## 2 gender + dept  0.206 0.466  Y
## 3 gender + clin  0.338 0.124  Y
## 4 gender + cert  0.334 0.136  Y
## 5 gender + prate 0.253 0.344  Y
## 6 gender + exper 0.309 0.201  Y
## 7 gender + rank  0.351 0.0922 N
```

Confounder 95

```
con = lm(log_sal95 ~ gender, data = lawsuit_log_95) %>% summary()
con_1 = lm(log_sal95 ~ gender + dept, data = lawsuit_log_95) %>% summary()
con_2 = lm(log_sal95 ~ gender + clin, data = lawsuit_log_95) %>% summary()
con_3 = lm(log_sal95 ~ gender + cert, data = lawsuit_log_95) %>% summary()
con_4 = lm(log_sal95 ~ gender + prate, data = lawsuit_log_95) %>% summary()
con_5 = lm(log_sal95 ~ gender + exper, data = lawsuit_log_95) %>% summary()
con_6 = lm(log_sal95 ~ gender + rank, data = lawsuit_log_95) %>% summary()

con_tab_95 = tibble("variables" = c("gender", "gender + dept", "gender + clin", "gender + cert", "gende
```

```
con_tab_95 %>% mutate(
  diff = abs((coef[1]-coef)/coef[1]),
  confounder = ifelse(diff>=0.1, "Y", "N")
)
```

```
## # A tibble: 7 x 4
##   variables       coef   diff confounder
##   <chr>          <dbl>  <dbl> <chr>
## 1 gender         0.384 0      N
## 2 gender + dept  0.204 0.468  Y
## 3 gender + clin  0.336 0.126  Y
## 4 gender + cert  0.332 0.136  Y
## 5 gender + prate 0.250 0.348  Y
## 6 gender + exper 0.307 0.202  Y
## 7 gender + rank  0.348 0.0936 N
```

From above result, we can see that all the variables except rank is the confounder for gender.

Global F-test

94

```
fit_94 = lm(log_sal94 ~ . + gender:rank, data = lawsuit_log_94)
anova(fit_94)
```

```
## Analysis of Variance Table
##
## Response: log_sal94
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## dept          5 48.472  9.6943 551.5163 < 2.2e-16 ***
## gender        1  2.424  2.4239 137.8974 < 2.2e-16 ***
## clin          1  2.321  2.3206 132.0184 < 2.2e-16 ***
## cert          1  2.504  2.5035 142.4266 < 2.2e-16 ***
## prate         1  0.001  0.0011   0.0603   0.80616
## exper         1  5.745  5.7452 326.8464 < 2.2e-16 ***
## rank          2  1.253  0.6264  35.6352 2.566e-14 ***
## gender:rank   2  0.118  0.0592   3.3675   0.03607 *
## Residuals   246  4.324  0.0176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the global F-test, we can see that prate is not significant. In order to be parsimony, I may delete it from our model.

Partial Ftest

```
model1_94 = lm(log_sal94~ dept + gender + clin + cert + exper + rank + gender:rank, data = lawsuit_log_
model2_94 = lm(log_sal94~ dept + gender + clin + cert + exper + rank + gender:rank + prate, data = lawsu


anova(model1_94, model2_94)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: log_sal94 ~ dept + gender + clin + cert + exper + rank + gender:rank
## Model 2: log_sal94 ~ dept + gender + clin + cert + exper + rank + gender:rank +
##     prate
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    247 4.3554
## 2    246 4.3241  1  0.031299 1.7806 0.1833
```

Since the P-value > 0.05, we can conclude that the model 2 is not better and we decide to exclude the "prate" for 94.

95

```
fit_95 = lm(log_sal95 ~ . + gender:rank, data = lawsuit_log_95)
anova(fit_95)
```

```
## Analysis of Variance Table
##
## Response: log_sal95
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## dept          5 48.737  9.7473 553.2395 < 2.2e-16 ***
## gender        1  2.381  2.3814 135.1652 < 2.2e-16 ***
## clin          1  2.431  2.4306 137.9585 < 2.2e-16 ***
## cert          1  2.421  2.4205 137.3847 < 2.2e-16 ***
## prate         1  0.001  0.0007   0.0422   0.83744
## exper         1  5.812  5.8121 329.8856 < 2.2e-16 ***
## rank          2  1.276  0.6378  36.1975  1.66e-14 ***
## gender:rank   2  0.110  0.0552   3.1304   0.04544 *
## Residuals   246  4.334  0.0176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Partial Ftest

```
model1_95 = lm(log_sal95~ dept + gender + clin + cert + exper + rank + gender:rank, data = lawsuit_log_
model2_95 = lm(log_sal95~ dept + gender + clin + cert + exper + rank + gender:rank + prate, data = lawsu


anova(model1_95, model2_95)
```

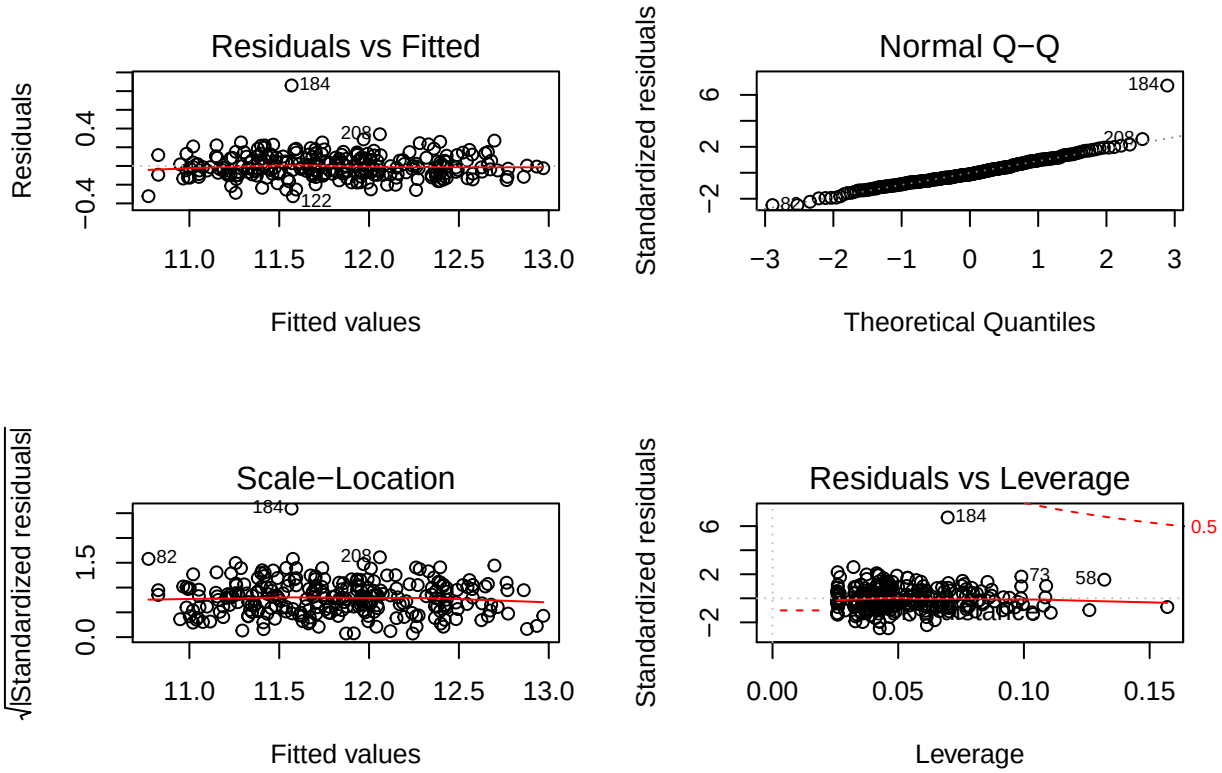
```
## Analysis of Variance Table
##
## Model 1: log_sal95 ~ dept + gender + clin + cert + exper + rank + gender:rank
## Model 2: log_sal95 ~ dept + gender + clin + cert + exper + rank + gender:rank +
##     prate
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    247 4.3636
## 2    246 4.3342  1  0.029443 1.6711 0.1973
```

Since the P-value > 0.05, we can conclude that the model 2 is not better and we decide to exclude the "prate" for 95.

Model Diagnostics

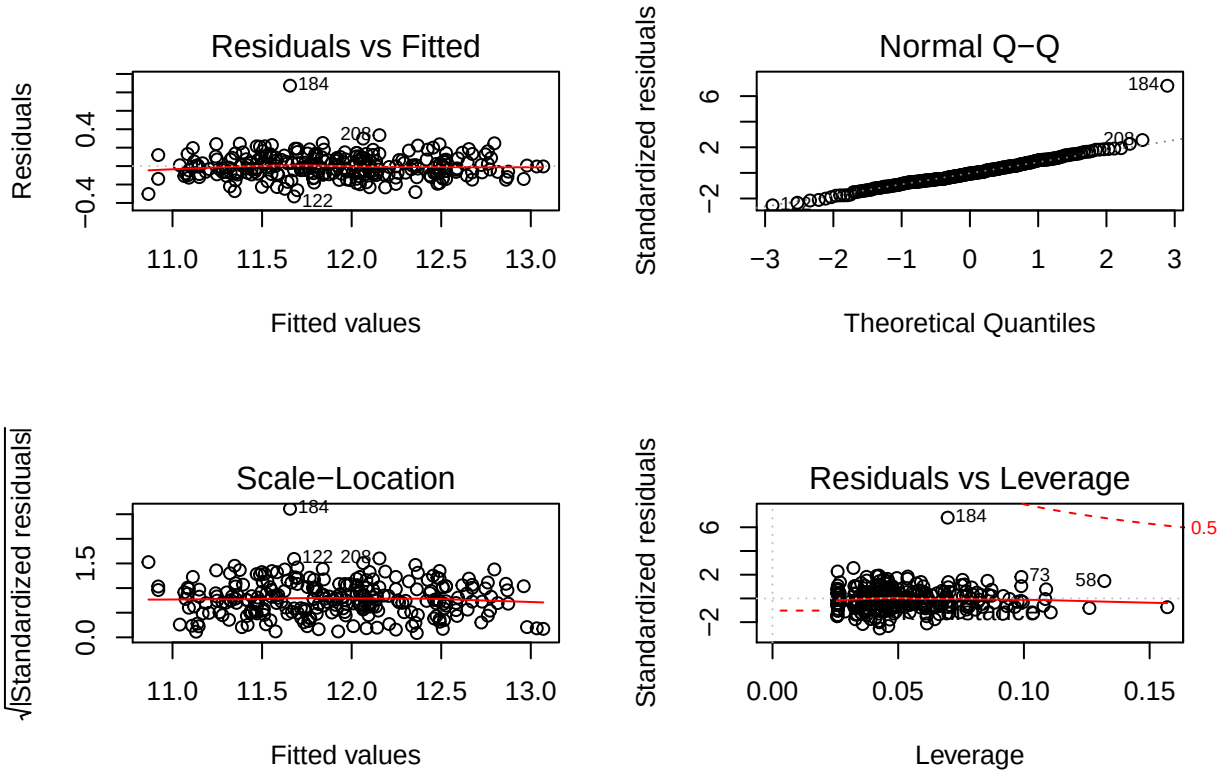
94

```
par(mfrow=c(2,2))
plot(model1_94)
```



95

```
par(mfrow=c(2,2))
plot(model1_95)
```



Generally, the assumption hold for both 94 and 95. The scale-location plot shows that the data has constant variance except for some outliers, which means heteroscedasticity assumption holds. Also, the qq plot shows that the data are followed the normal line, which means the normality assumption holds.

Multicollinearity

94

```
vif_94 = vif(model1_94) %>% as.data.frame() %>% rownames_to_column()
names(vif_94) = c("variable", "vif")
vif_94 %>% .[which(.$vif > 5),]
```

```
##                          variable      vif
## 11            rankFull professor 5.115323
## 13 genderMale:rankFull professor 6.289418
```

95

```
vif_95 = vif(model1_95) %>% as.data.frame() %>% rownames_to_column()
names(vif_95) = c("variable", "vif")
vif_95 %>% .[which(.$vif > 5),]
```

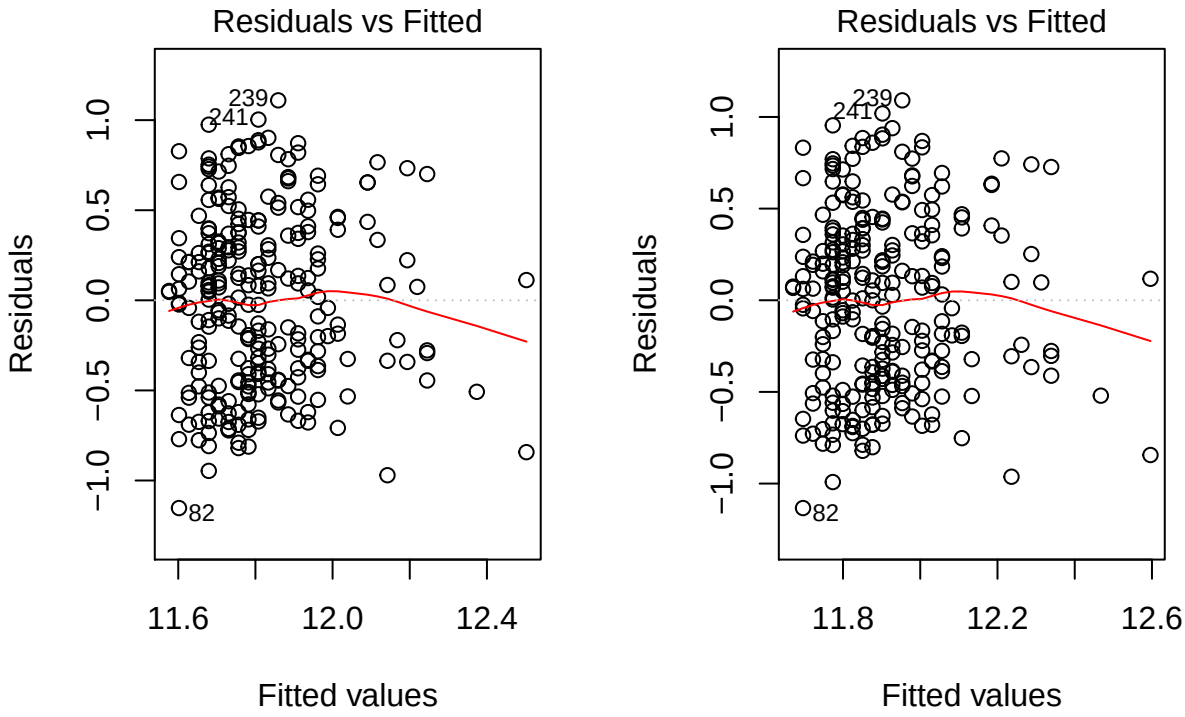
```
##                          variable      vif
## 11            rankFull professor 5.115323
## 13 genderMale:rankFull professor 6.289418
```

The VIF suggest that rank and the interaction term for rank and gender may have collinearity. However, the interaction term will always has collinearity with main effect itself. We would not drop the interaction term and will keep it in the model for both 94 and 95.

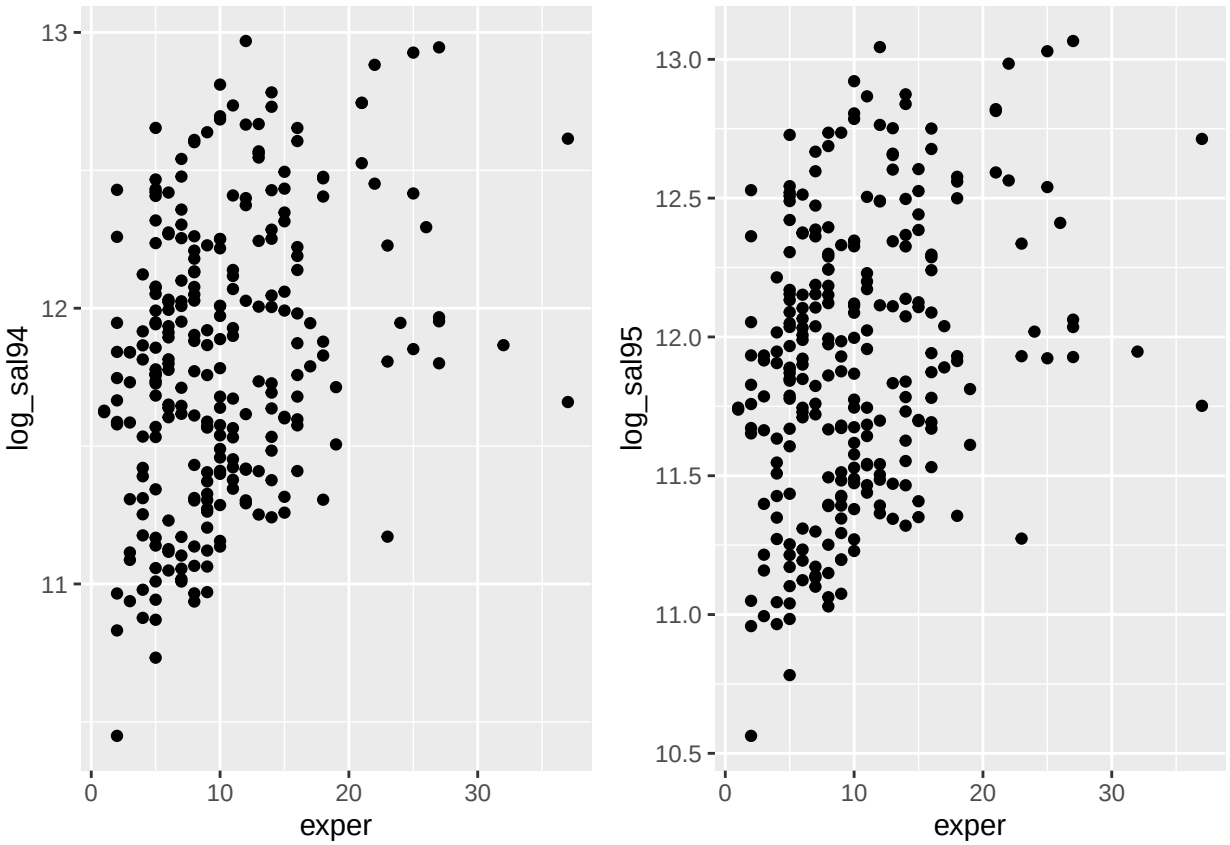
Functional forms for continuous variables

94 vs 95

```
fit1 = lm(log_sal94 ~ exper, data = lawsuit_log_94)
fit2 = lm(log_sal95 ~ exper, data = lawsuit_log_95)
par(mfrow=c(1,2))
plot(fit1, which = 1)
plot(fit2, which = 1)
```




```
lawsuit_log_94 %>% ggplot(aes(x=exper, y = log_sal94)) + geom_point() +
  lawsuit_log_95 %>% ggplot(aes(x=exper, y = log_sal95)) + geom_point()
```



We can see that for both 94 and 95, the residual vs fitted plots does not suggest a curvilinear trend and the scatter plot shows a potenrial increasing linear relationship between exper and outcome. Thus, for continuous variables "exper", the function form may be linear.

Outliers/Influential points

Outliers in Y

94

```
rs_94 = rstandard(model1_94)
out_y_94 = rs_94[abs(rs_94)>2.5]
out_y_94
```

```
##      184      208
## 6.718456 2.595229
```

For year 94, the data 184 and 208 are outliers in X.

95

```
rs_95 = rstandard(model1_95)
out_y_95 = rs_95[abs(rs_95)>2.5]
out_y_95
```

```
##       122       184       208
## -2.533667  6.807405  2.566378
```

For year 95, the data 122, 184 and 208 are outliers in X.

Outliers in X

94

```
hat_94 = lm.influence(model1_94)$hat
hat_94[hat_94>0.2]
```

```
## named numeric(0)
```

There is no outlier in Y for 94.

94

```
hat_95 = lm.influence(model1_95)$hat
hat_95[hat_95>0.2]
```

```
## named numeric(0)
```

There is no outlier in Y for 95.

Influential point

94

```
dffits_94 = dffits(model1_94)
abs(dffits_94[c(184,208)])>(sqrt(5/nrow(lawsuit_log_94))*2)
```

```
##  184  208
## TRUE TRUE
```

The dffits suggest that data 184 and 208 are both influential points for 94.

```
cooks.distance(model1_94)[c(184,208)] > (4/nrow(lawsuit_log_94))
```

```
##  184  208
## TRUE TRUE
```

The cook's distance also suggest that data 184 and 208 are both influential points for 94.

95

```
dffits_95 = dffits(model1_95)
abs(dffits_95[c(122,184,208)]) > (sqrt(5/nrow(lawsuit_log_95))*2)
```

```
##  122  184  208
## TRUE TRUE TRUE
```

The dffits suggest that data 122, 184 and 208 are both influential points for 95.

```
cooks.distance(model1_95)[c(122,184,208)] > (4/nrow(lawsuit_log_95))
```

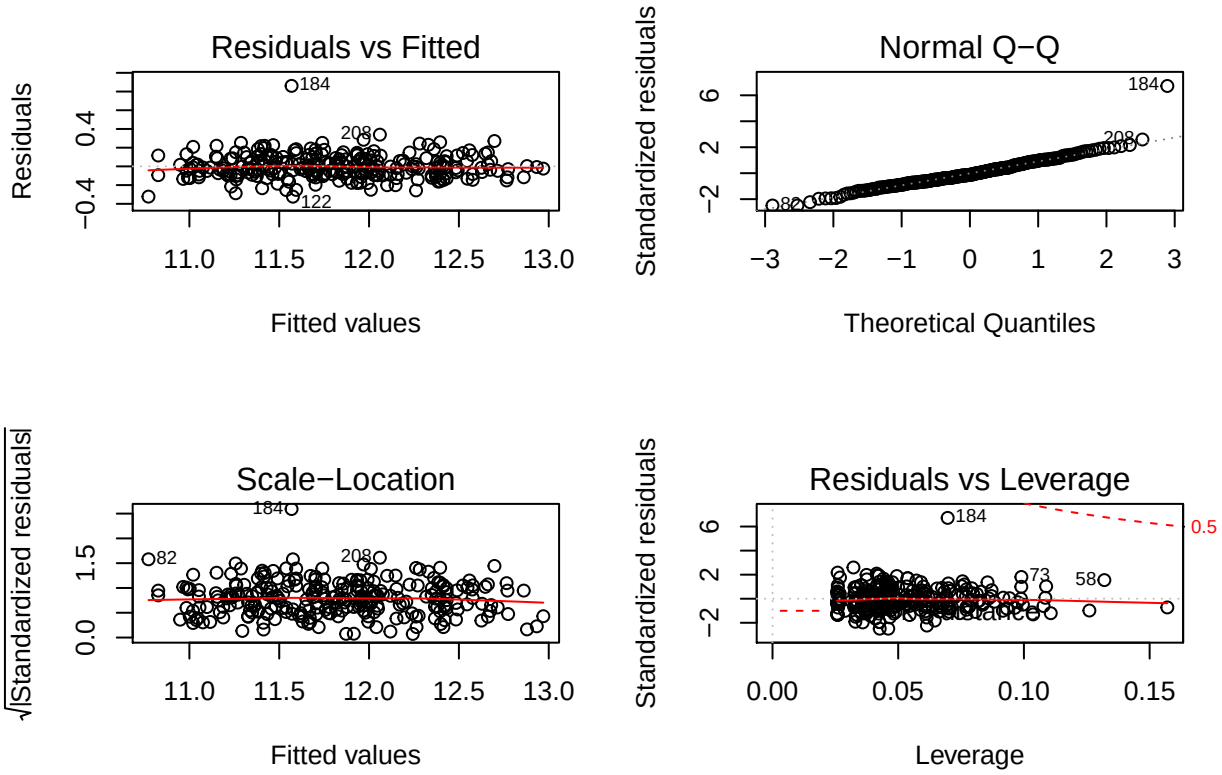
```
##  122  184  208
## TRUE TRUE TRUE
```

The cook's distance also suggest that data 122, 184 and 208 are both influential points for 95.

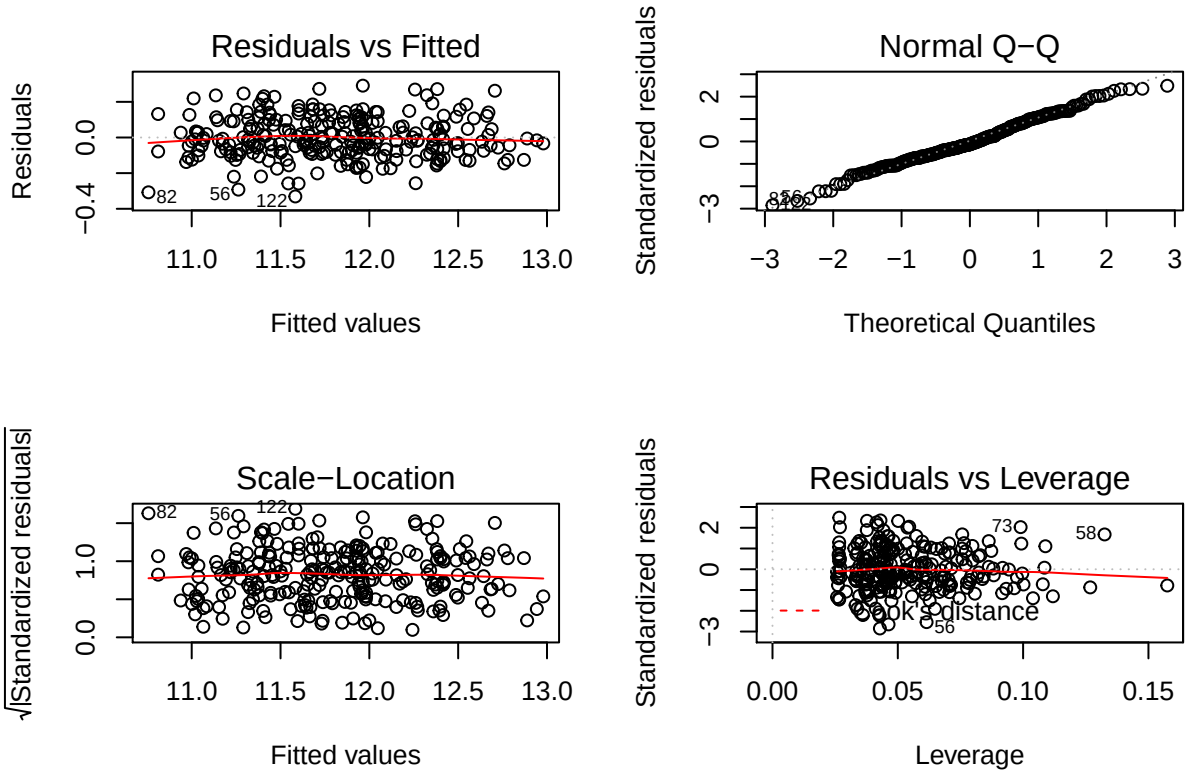
Removing influential points

94

```
fit.model_94 = lm(log_sal94~ dept + gender + clin + cert + exper + rank + gender:rank, data = lawsuit_l
par(mfrow = c(2,2))
plot(fit.model_94)
```



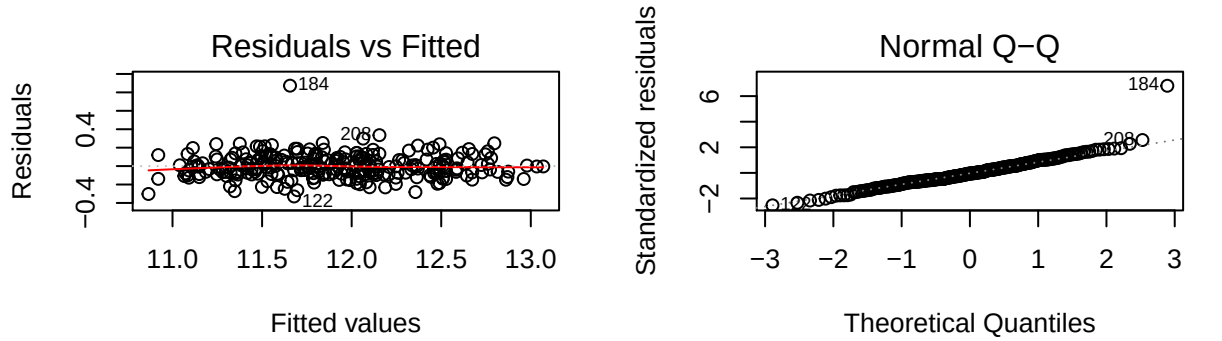

```
fit.model_94_nooutlier = lm(log_sal94~ dept + gender + clin + cert + exper + rank + gender:rank, data =
par(mfrow = c(2,2))
plot(fit.model_94_nooutlier)
```



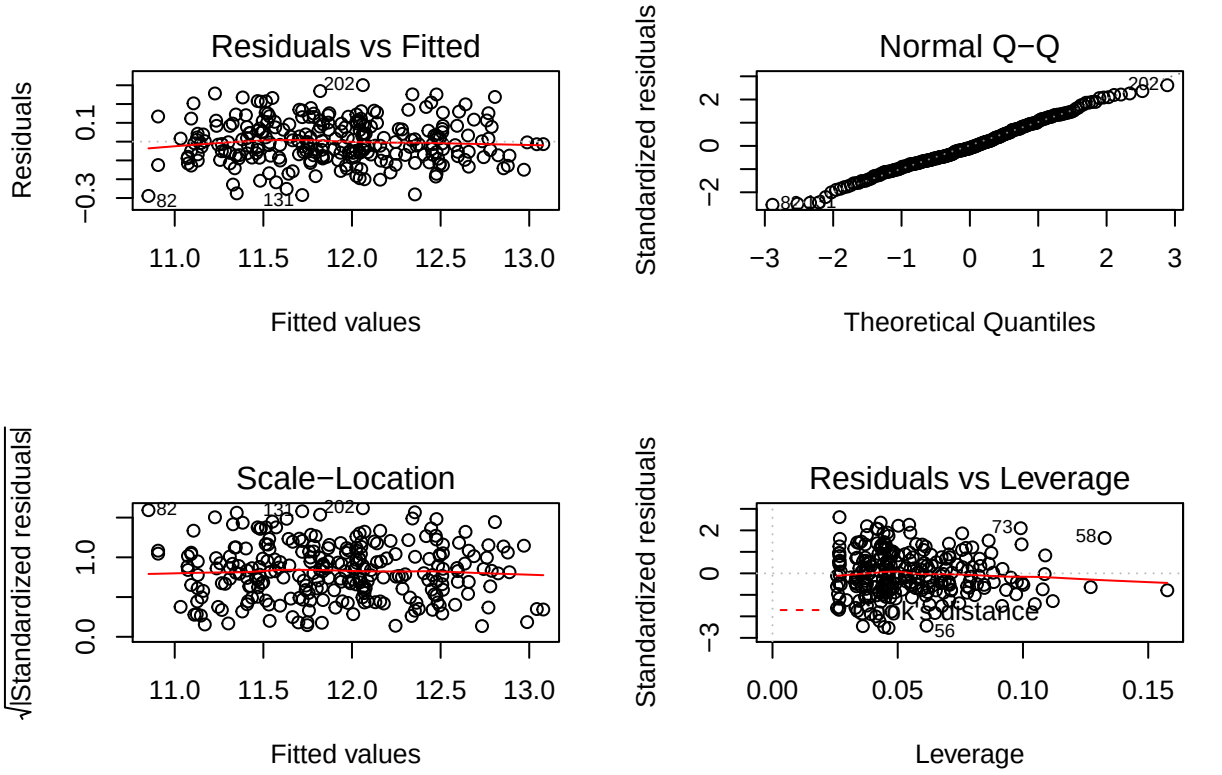
After removing the influential points from 94, we can see that the assumption holds well. The residual vs fitted plot shows that data are more evenly distributed on two side of zero line. The qq plot does not have outlier far away from the line, and there are no outliers close to the cook's distance.

95

```
fit.model_95 = lm(log_sal95~ dept + gender + clin + cert + exper + rank + gender:rank, data = lawsuit_l
par(mfrow = c(2,2))
plot(fit.model_95)
```




```
fit.model_95_nooutlier = lm(log_sal95~ dept + gender + clin + cert + exper + rank + gender:rank, data =
par(mfrow = c(2,2))
plot(fit.model_95_nooutlier)
```



After removing the influential points from 95, we can see that the assumption holds well. The residual vs

fitted plot shows that data are more evenly distributed on two side of zero line. The qq plot does not have outlier far away from the line, and there are no outliers close to the cook's distance.