output

xj2249

12/15/2019

Data cleaning

```
# import
df = read_csv("Lawsuit.csv") %>%
  janitor::clean_names() %>%
  mutate(dept = factor(dept, levels = c(1:6), labels = c("Biochemistry/Molecular Biology", "Physiology
         gender = factor(gender, levels = c(0, 1), labels = c("Female","Male")),
         clin = factor(clin, levels = c(1, 0), labels = c("Clinical","Research")),
         cert = factor(cert, levels = c(1, 0), labels = c("Certified","Not certified")),
         rank = factor(rank, levels = c(1, 2, 3), labels = c("Assistant", "Associate", "Full professor"
         sal = (sal94 + sal95)/2)
```

```
## Parsed with column specification:
## cols(
##   ID = col_double(),
##   Dept = col_double(),
##   Gender = col_double(),
##   Clin = col_double(),
##   Cert = col_double(),
##   Prate = col_double(),
##   Exper = col_double(),
##   Rank = col_double(),
##   Sal94 = col_double(),
##   Sal95 = col_double()
## )
```

```
# consider log mean sal only
df_sal = df %>%
  mutate(log_sal = log(sal)) %>%
  dplyr::select(-sal95, -id, -sal94, -sal)
```

interaction term (not sure whether to test)

```
## general test

# no prate
fit_conf = lm(log_sal ~gender+dept+clin+cert+exper+rank, data = df_sal)
summary(fit_conf)
```

Call: lm(formula = log_sal ~ gender + dept + clin + cert + exper + rank, data = df_sal)

Residuals: Min 1Q Median 3Q Max -0.34605 -0.07696 -0.01873 0.07596 0.90393

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.373862 0.034398 330.651 < 2e-16 ***genderMale 0.025763 0.019624 1.313 0.19 deptPhysiology -0.175749 0.029122 -6.035 5.73e-09*** deptGenetics 0.185970 0.036501 5.095 6.90e-07 ***deptPediatrics 0.203345 0.035712 5.694 3.48e-08*** deptMedicine 0.539304 0.029515 18.272 < 2e-16 ***deptSurgery 0.933820 0.035533 26.280 < 2e-16*** clinResearch -0.208340 0.021885 -9.520 < 2e-16 ***certNot certified -0.189749 0.021244 -8.932 < 2e-16*** exper 0.017726 0.001812 9.783 < 2e-16 ***rankAssociate 0.134663 0.023557 5.716 3.10e-08*** rankFull professor 0.222214 0.026249 8.466 2.22e-15 *** — Signif. codes: 0 '' ***0.001*** '' *0.01* " 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1337 on 249 degrees of freedom Multiple R-squared: 0.9339, Adjusted R-squared: 0.931 F-statistic: 319.7 on 11 and 249 DF, p-value: < 2.2e-16

```
fit_int = lm(log_sal ~gender+dept+clin+cert+exper+rank+gender*exper, data = df_sal)
```

Interaction term *gender* ∗ *exper* is significant, thus we may conside it in our model.

stratified regression

```
stratified_dept = df_sal %>%
  group_by(dept) %>%
  summarize(
      n = n(),
      coef =  lm(log_sal ~ gender+clin+cert+exper+rank)$coef["genderMale"],
      p = summary(lm(log_sal ~ gender+clin+cert+exper+rank))$coefficients["genderMale",4]
            )
stratified_dept %>%
    knitr::kable()
```

dept	n	coef	p
Biochemistry/Molecular Biology	50	-0.0187106	0.6600235
Physiology	40	-0.0052950	0.9224663
Genetics	21	0.0754572	0.2339215
Pediatrics	30	0.0115277	0.8453661
Medicine	80	0.0366927	0.3660339
Surgery	40	0.0416427	0.4947262

```
stratified_clin = df_sal %>%
  group_by(clin) %>%
  summarize(
      n = n(),
      coef =  lm(log_sal ~ gender + dept + cert + exper +
    rank)$coef["genderMale"],
      p = summary(lm(log_sal ~ gender + dept + cert + exper +
    rank))$coefficients["genderMale",4]
            )
stratified_clin %>%
    knitr::kable()
```

clin	n	coef	p
Clinical	160	0.0083165	0.7108663
Research	101	0.0465115	0.2948187

```
stratified_cert = df_sal %>%
  group_by(cert) %>%
  summarize(
      n = n(),
      coef =  lm(log_sal ~ gender + dept + clin + exper +
    rank)$coef["genderMale"],
      p = summary(lm(log_sal ~ gender + dept + clin + exper +
    rank))$coefficients["genderMale",4]
            )
stratified_cert %>%
    knitr::kable()
```

cert	n	coef	p
Certified	188	0.0126811	0.5584154
Not certified	73	0.0265111	0.5547041

```
stratified_rank = df_sal %>%
  group_by(rank) %>%
  summarize(
      n = n(),
      coef =  lm(log_sal ~ gender + dept + clin + cert + exper)$coef["genderMale"],
      p = summary(lm(log_sal ~ gender + dept + clin + cert + exper))$coefficients["genderMale",4]
            )
stratified_rank %>%
    knitr::kable()
```

rank	n	coef	p
Assistant	112	0.0826555	0.0213160
Associate	64	-0.0132771	0.6702516
Full professor	85	-0.0404129	0.2680458

```
df_exper = df_sal %>%
  mutate(exper_fct = case_when(
      exper < 6 ~ "0",
      exper >= 6 & exper < 9 ~ "1",
      exper >= 9 & exper < 14 ~ "2",
      exper >= 14 ~ "3",
      TRUE ~ ""
  )) %>%
    mutate(exper = factor(exper_fct)) %>%
    dplyr::select(-exper_fct)

stratified_exper = df_exper %>%
  group_by(exper) %>%
  summarize(
      n = n(),
      coef =  lm(log_sal ~ gender + dept + clin + cert + rank)$coef["genderMale"],
      p = summary(lm(log_sal ~ gender + dept + clin + cert + rank))$coefficients["genderMale",4]
            )
stratified_exper %>%
    knitr::kable()
```

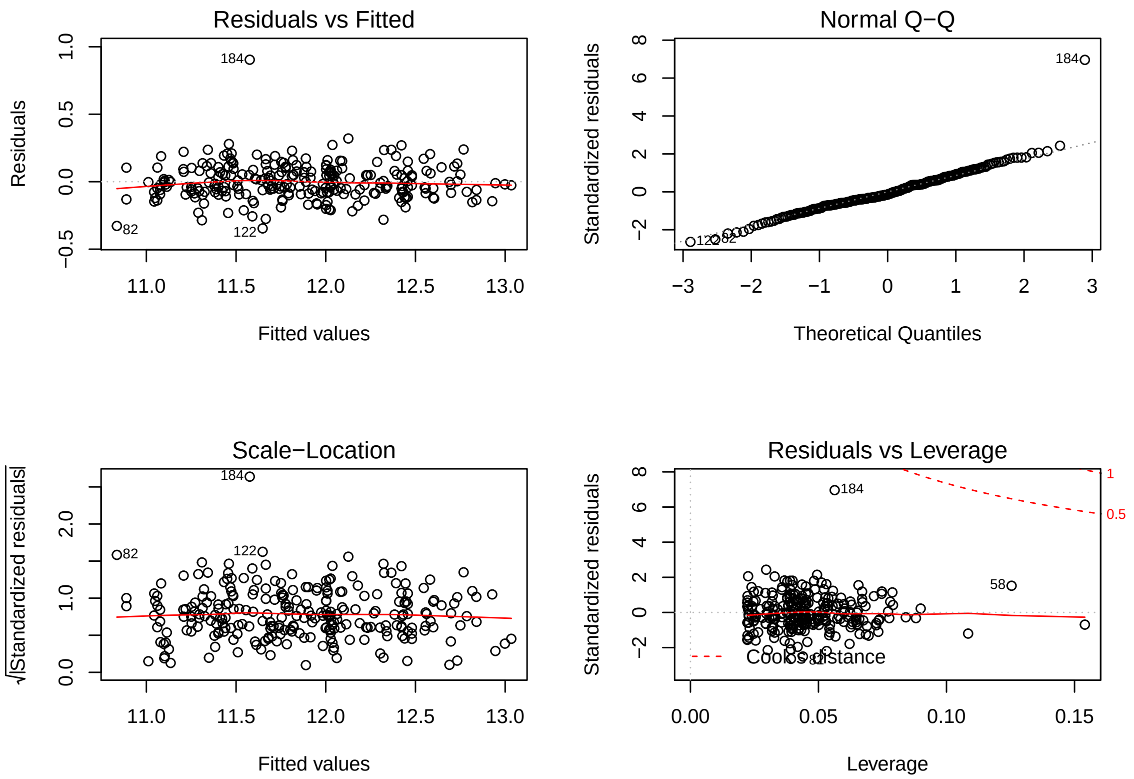
exper	n	coef	p
0	64	0.1257741	0.0238410
1	57	0.0340942	0.2757676
2	74	-0.0005508	0.9876466
3	66	-0.0034961	0.9439975

```
df_exper  %>%
  group_by(exper, rank) %>%
  summarize(
      n = n())
```

```
## # A tibble: 11 x 3
## # Groups:   exper [4]
##    exper rank              n
##    <fct> <fct>         <int>
##  1 0     Assistant        59
##  2 0     Associate         5
##  3 1     Assistant        36
##  4 1     Associate        15
##  5 1     Full professor    6
##  6 2     Assistant        12
##  7 2     Associate        31
##  8 2     Full professor   31
##  9 3     Assistant         5
## 10 3     Associate        13
## 11 3     Full professor   48
```

Model diagnostics

```
final_model = lm(log_sal ~gender+dept+clin+cert+exper+rank, data = df_sal)
par(mfrow = c(2,2))
plot(final_model)
```



Outliers/influential points

```
stu_res<-rstandard(final_model)
stu_res[abs(stu_res)>2.5]
```

```
##         82        122        184
## -2.507753 -2.640742  6.960085
```

```
influence.measures(final_model) %>%
  summary()
```

```
## Potentially influential observations of
##   lm(formula = log_sal ~ gender + dept + clin + cert + exper +      rank, data = df_sal) :
##
##     dfb.1_ dfb.gndM dfb.dptPh dfb.dptG dfb.dptPd dfb.dptM dfb.dptS
## 19   0.08  -0.01     0.04      0.04     0.00      0.03     0.03
## 82   0.01   0.07    -0.31      0.00    -0.05     -0.08    -0.11
## 91   0.04  -0.01    -0.01     -0.08    -0.02     -0.02    -0.01
## 109  0.00  -0.02     0.00      0.05     0.00      0.01     0.01
## 122 -0.14   0.07     0.03      0.04    -0.27      0.08     0.06
## 184 -0.62   0.75     0.22      0.16     0.53      1.00     0.53
## 208  0.06  -0.19    -0.01     -0.01    -0.04      0.11    -0.02
##     dfb.clnR dfb.crNc dfb.expr dfb.rnkA dfb.rnFp dffit   cov.r   cook.d
## 19  -0.01     -0.09    -0.26     0.05     0.17    -0.30    1.21_*  0.01
## 82  -0.13     -0.18     0.08     0.12     0.07    -0.54    0.81_*  0.02
## 91  -0.04     0.03    -0.04     0.02     0.01    -0.10    1.15_*  0.00
## 109  0.02    -0.03     0.00     0.03     0.00     0.07    1.15_*  0.00
## 122  0.11     0.02     0.02     0.10     0.05    -0.54    0.78_*  0.02
## 184  0.94     0.84    -0.36    -0.51    -0.26     1.89_*  0.08_*  0.24
## 208 -0.06    -0.04     0.21    -0.14    -0.18     0.43    0.81_*  0.02
##      hat
## 19   0.15_*
## 82   0.04
## 91   0.09
## 109  0.09
## 122  0.04
## 184  0.06
## 208  0.03
```

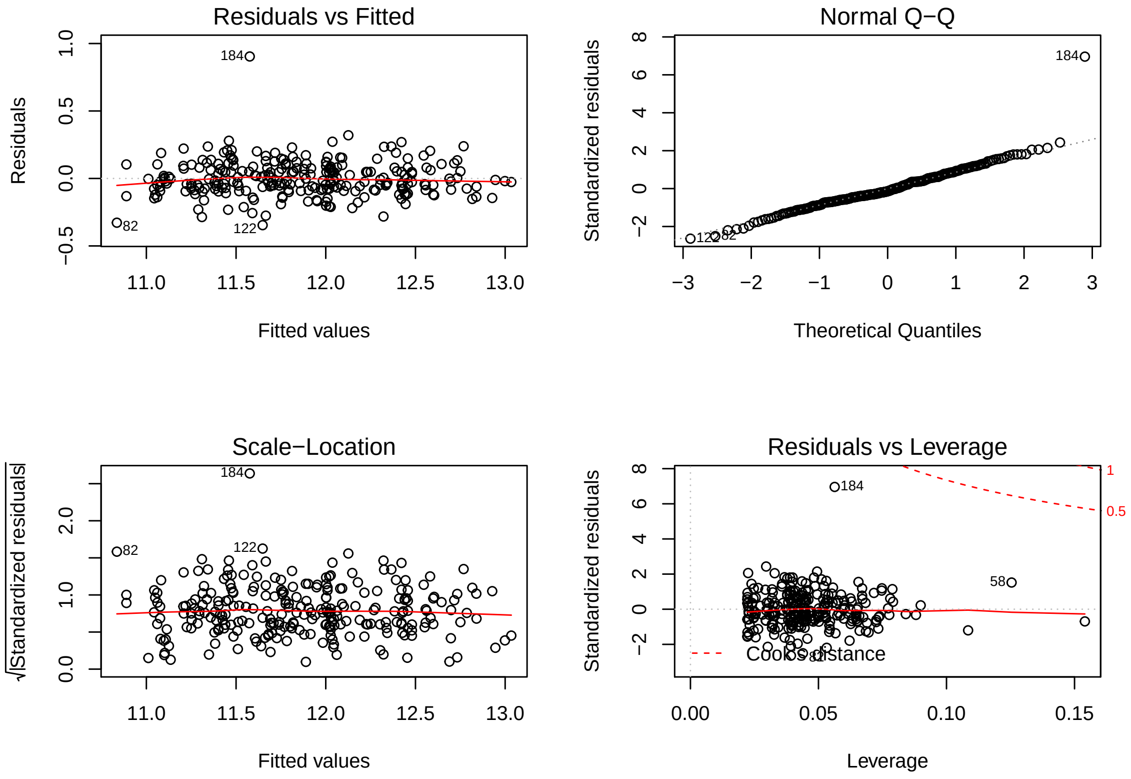
```
df[184,]
```

```
## # A tibble: 1 x 11
##       id dept   gender clin   cert    prate exper rank    sal94  sal95    sal
##   <dbl> <fct>  <fct>  <fct>  <fct>   <dbl> <dbl> <fct>  <dbl>  <dbl>  <dbl>
## 1   184 Medic~ Male   Resea~ Not ce~   5.1     2 Assi~ 250000 276163 2.63e5
```

Using studentized residuals, id 184 is an outlier in Y. using leverage values, 19 and 216 are outliers in X. Using DFFIT, 8, 184 and 216 are influential points. Using main effects only, 184 is influential.

```
# consider the data without influential points
df_sal_noinflu = df_sal[-184, ]

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



```
temp = lm(log_sal ~gender+dept+clin+cert+exper+gender*rank+gender*exper, data = df_sal_noinflu)
summary(temp)
```

```
##
## Call:
## lm(formula = log_sal ~ gender + dept + clin + cert + exper +
##     gender * rank + gender * exper, data = df_sal_noinflu)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32895 -0.07173 -0.01277  0.08089  0.28179
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   11.324438   0.039263 288.429  < 2e-16 ***
## genderMale                     0.100527   0.034725   2.895  0.00413 **
## deptPhysiology                -0.172382   0.026182  -6.584 2.77e-10 ***
## deptGenetics                   0.183611   0.032499   5.650 4.45e-08 ***
## deptPediatrics                 0.200148   0.032276   6.201 2.36e-09 ***
## deptMedicine                   0.520522   0.026654  19.529  < 2e-16 ***
## deptSurgery                    0.922849   0.031852  28.973  < 2e-16 ***
## clinResearch                  -0.225913   0.020356 -11.098  < 2e-16 ***
## certNot certified             -0.198053   0.019681 -10.063  < 2e-16 ***
## exper                          0.026606   0.003887   6.845 6.11e-11 ***
## rankAssociate                  0.138212   0.033015   4.186 3.95e-05 ***
## rankFull professor             0.213618   0.044342   4.817 2.55e-06 ***
## genderMale:rankAssociate      -0.011198   0.043888  -0.255  0.79882
## genderMale:rankFull professor  0.002157   0.054723   0.039  0.96858
## genderMale:exper              -0.009676   0.004265  -2.269  0.02416 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1188 on 245 degrees of freedom
## Multiple R-squared:  0.9483, Adjusted R-squared:  0.9454
## F-statistic: 321.2 on 14 and 245 DF,  p-value: < 2.2e-16
```

```
df_exper_noinflu = df_sal_noinflu %>%
  mutate(exper_fct = case_when(
      exper < 6 ~ "0",
      exper >= 6 & exper < 9 ~ "1",
      exper >= 9 & exper < 14 ~ "2",
      exper >= 14 ~ "3",
      TRUE ~ ""
  )) %>%
    mutate(exper = factor(exper_fct)) %>%
    dplyr::select(-exper_fct)

stratified_exper_noinflu = df_exper_noinflu %>%
  group_by(exper) %>%
  summarize(
      n = n(),
      coef =  lm(log_sal ~ gender + dept + clin + cert + rank)$coef["genderMale"],
      p = summary(lm(log_sal ~ gender + dept + clin + cert + rank))$coefficients["genderMale",4]
```

```
            )
stratified_exper_noinflu %>%
    knitr::kable()
```

exper	n	coef	p
0	63	0.0577437	0.2066119
1	57	0.0340942	0.2757676
2	74	-0.0005508	0.9876466
3	66	-0.0034961	0.9439975

```
stratified_rank_noinflu = df_sal_noinflu %>%
  group_by(rank) %>%
  summarize(
      n = n(),
      coef =  lm(log_sal ~ gender + dept + clin + cert + exper)$coef["genderMale"],
      p = summary(lm(log_sal ~ gender + dept + clin + cert + exper))$coefficients["genderMale",4]
            )
stratified_rank_noinflu %>%
    knitr::kable()
```

rank	n	coef	p
Assistant	111	0.0390298	0.2009195
Associate	64	-0.0132771	0.6702516
Full professor	85	-0.0404129	0.2680458

Not significant now, -184 usinf main effects model or -216, -184, -8 using interaction model.

model output

```
stargazer(fit_int,temp, title ="",
          dep.var.labels = c("Log Salary"),
          column.labels   = c("Final model", "Final model without 184"),
          covariate.labels = c("Male","Physiology","Genetics","Pediatrics","Medicine", "Surgery",
                             "Research emphasis","Not board certified","Experience","Associate",
                             "Full professor","Male:Associate","Male:Full professor","Male:Experience")
          )
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: , 12 15, 2019 - 17 46 41

Table 8:

	Dependent variable:	
	Log Salary	
	Final model	Final model without 184
	(1)	(2)
Male	0.129***	0.101***
	(0.037)	(0.035)
Physiology	−0.165***	−0.172***
	(0.029)	(0.026)
Genetics	0.190***	0.184***
	(0.036)	(0.032)
Pediatrics	0.219***	0.200***
	(0.035)	(0.032)
Medicine	0.547***	0.521***
	(0.029)	(0.027)
Surgery	0.940***	0.923***
	(0.035)	(0.032)
Research emphasis	−0.208***	−0.226***
	(0.021)	(0.020)
Not board certified	−0.182***	−0.198***
	(0.021)	(0.020)
Experience	0.028***	0.027***
	(0.004)	(0.004)
Associate	0.118***	0.138***
	(0.024)	(0.033)
Full professor	0.208***	0.214***
	(0.026)	(0.044)
Male:Associate		−0.011
		(0.044)
Male:Full professor		0.002
		(0.055)
Male:Experience	−0.012***	−0.010**
	(0.004)	(0.004)
Constant	11.294***	11.324***
	(0.042)	(0.039)
Observations	261	260
R^2	0.937	0.948
Adjusted R^2	0.934	0.945
Residual Std. Error	0.131 (df = 248)	0.119 (df = 245)
F Statistic	305.449*** (df = 12; 248)	321.180*** (df = 14; 245)
Note:		*p<0.1; **p<0.05; ***p<0.01