Lorraine Jiang

Analysis of employee compensation

The `Sleuth3` package contains a dataset of salaries and other information for clerical employees at Harris Trust and Savings Bank in 1977. The first few rows of this data are shown below.

```
# give the data a descriptive name
salaries <- Sleuth3::case1202

# preview
head(salaries)
```

```
##   Bsal Sal77  Sex Senior Age Educ Exper
## 1 5040 12420 Male     96 329   15  14.0
## 2 6300 12060 Male     82 357   15  72.0
## 3 6000 15120 Male     67 315   15  35.5
## 4 6000 16320 Male     97 354   12  24.0
## 5 6000 12300 Male     66 351   12  56.0
## 6 6840 10380 Male     92 374   15  41.5
```

You can find variable descriptions by querying the help file:

```
# check documentation
?Sleuth3::case1202
```

Your objective is to construct a linear model of employee salaries (`Sal77`) and use the model to answer the following questions:

1. Do the data provide evidence of discrimination on the basis of sex?
2. How do mean salaries appear to change with age, education, experience, and seniority?

You will be guided through the data analysis sequentially, much as in the 'Applications' sections of your homework assignments, in the questions below.

A0. Preprocessing Notice that age, experience, and seniority are all measured in months. This is a somewhat odd unit of measurement, and model coefficients will likely have more intuitive interpretations if they are converted instead to years.

i. Construct new variables named `age` (lowercase 'a'), `experience`, and `seniority` that report these quantities in years rather than months. Ensure that these are stored in the `salaries` dataframe for later use. Show your codes only.

```
salaries$age = salaries$Age/12
salaries$experience = salaries$Exper/12
salaries$seniority = salaries$Senior/12
salaries
```

ii. Follow the example below to rename `Sex`, `Bsal`, `Educ`, and `Sal77` as follows: `sex` (lowercase 's'), `base`, `education`, and `salary`. Show only your codes. (*Hint*: `rename(newname = oldname)`.)

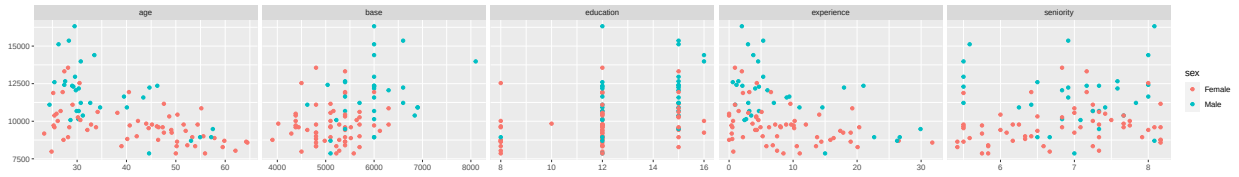
```
# example
salaries %>% rename(education = Educ)
```

iii. Now select the newly defined and renamed columns by running the chunk below. If you followed the naming instructions in (i) – (ii) correctly, this should run without error.

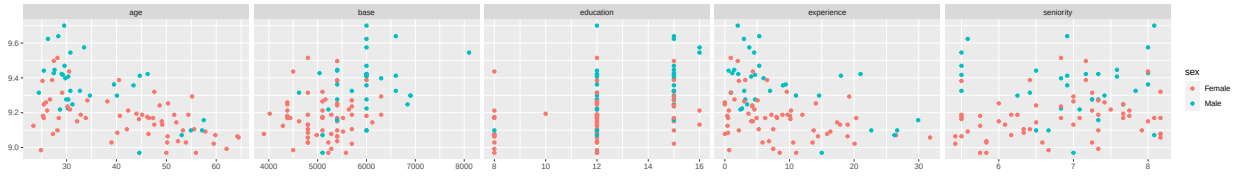
```
# select columns
salaries <- salaries %>%
  select(salary, base, age, sex, education, experience, seniority)

# preview
head(salaries)
```

A1. Data visualization

i. Construct a 1 x 5 panel of scatterplots of salary against each predictor *except* sex, and color the points according to sex. Show only the graphic, and be sure to adjust the figure sizing in the code chunk options so that the graphic renders well. Also be sure that labels are legible and appropriate; you may need to rotate the value labels to avoid overlap (see lab 4, *Detecting unusual observations* for an example).



ii. Repeat (i) but with log-salary shown on the y axes.



iii. On which scale do you think the relationships look closer to linear?

The relationships seems to be closer to linear if y axes is log(salary).

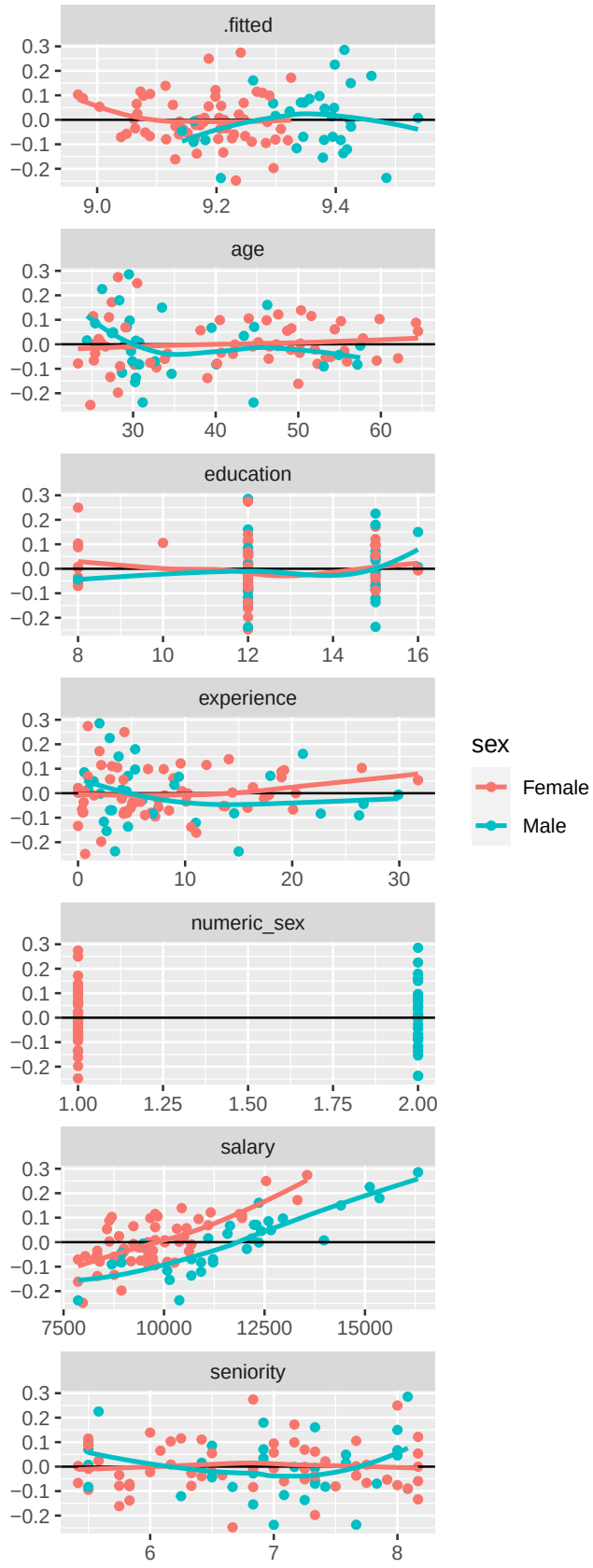
iv. Overall, does it appear from the plots that salaries differ between male and female employees? Yes, the salaries differ between male and female employees. From these plots, we get to know mean salary of male is higher than that of female.

A2. Model fitting and checking

i. Fit a model with log-salary as the response that is linear in all predictors. Show only your codes.

```
##
## Call:
## lm(formula = log(salary) ~ base + age + education + experience +
##     seniority + numeric_sex, data = salaries)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.247898 -0.070468 -0.007108  0.068992  0.285636
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.688e+00  1.842e-01  47.166  < 2e-16 ***
## base         7.631e-05  2.334e-05   3.270  0.00155 **
## age         -5.506e-03  1.891e-03  -2.912  0.00457 **
## education    9.639e-03  5.826e-03   1.655  0.10166
## experience  -2.241e-03  2.760e-03  -0.812  0.41903
## seniority    2.432e-02  1.521e-02   1.599  0.11356
## numeric_sex  6.169e-02  3.331e-02   1.852  0.06745 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1106 on 86 degrees of freedom
## Multiple R-squared:  0.5718, Adjusted R-squared:  0.542
## F-statistic: 19.14 on 6 and 86 DF,  p-value: 4.837e-14
```

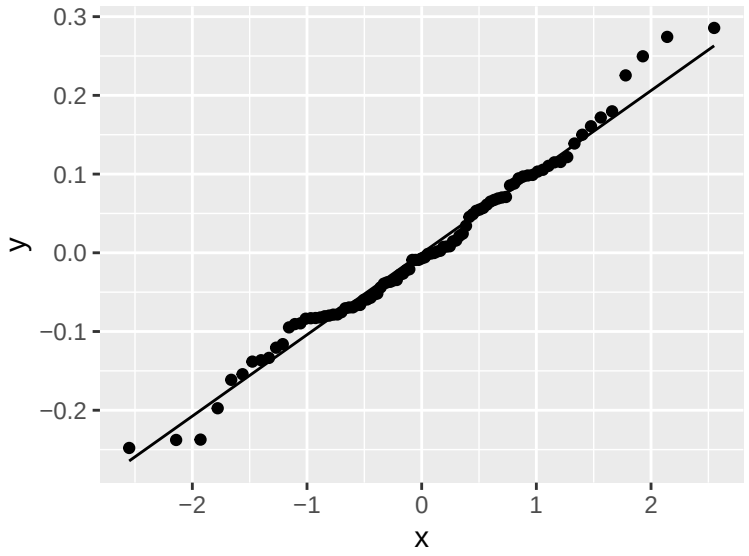
ii. Construct a 7 x 1 panel of residual scatterplots showing residuals on the y axis against: fitted values; age; base salary; education; experience; seniority; and sex. Include a LOESS smooth to help visualize any trends with a smoothing span of your choosing. Ensure the labels and knit options are organized so that the figure is legible when rendered.


.fitted
age
education
experience
numeric_sex
salary
seniority
0.3
0.2
0.1
0.0
−0.1
−0.2
9.0
9.2
9.4
30
40
50
60
8
10
12
14
16
0
10
20
30
1.00
1.25
1.50
1.75
2.00
7500
10000
12500
15000
6
7
8
sex
Female
Male

iii. Do you see any problems with model assumptions based on the residual scatterplots? If so, identify the assumption(s) and describe what you see in the plots. Answer in 1-3 sentences. No. From these residual plots, the linearity assumption still holds. The only problem is the constant variance as the variance does not spread out evenly, but it will no affect profoundly our identification.

iv. If you identified any problems, are they important given your goals? (If not, skip this question.) No. Skip this question.

v. Construct a quantile-quantile plot of the residuals. Show only the graphic.

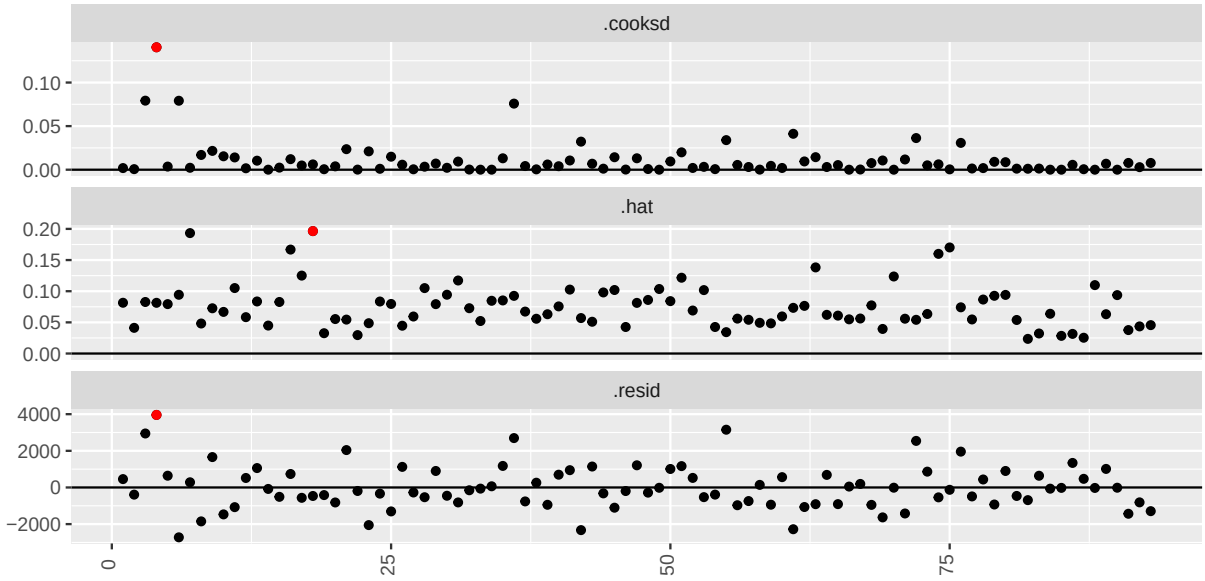


vi. Do you see any issues with model assumptions based on the Q-Q plot? If so, identify the assumption(s) and describe what you see in the plot. Answer in 1-2 sentences.

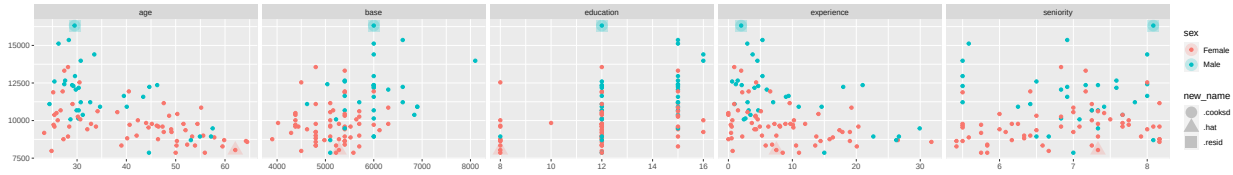
I think the normality still holds even though there are some outliers. For the sample size is large, we can ignore the effect of residuals.

A3. Outlier and influential point detection

i. Construct a 3 x 1 panel of plots of the case influence statistics for each observation. Highlight any unusual observations in red (if there are no unusual observations, there is no need to highlight any points). Show only the graphic, and ensure it is sized and labeled appropriately.



ii. Show a scatterplot of the data (modify one of the two figures from A1) with the unusual points highlighted. Show only the graphic, and ensure it is sized and labeled appropriately.




```
## [1] 16320  8040 16320
```

iii. In what way, if any, do the highlighted points seem unusual? Answer in 1-2 sentences.

The unusual observation with salary 16320 is an outlier as its salary is extremely large. The unusual leverage with salary 8040 is leverage point for its high t value.

iv. Assess the fit of the model without the observations you highlighted (if any). Are the points, in fact, influential? Answer in 1 sentence and show any codes (but not output) you used to check.

```
unusual_idx <- augment(dat, salaries) %>%
  mutate(idx = row_number()) %>%
  slice_max(order_by = abs(.resid), n = 1)  %>%
  pull(idx)

unusual_idx2 <- augment(dat, salaries) %>%
  mutate(idx2 = row_number()) %>%
  slice_max(order_by = abs(.hat), n = 1)  %>%
  pull(idx2)

fit1 <- lm((salary)~experience + age + education + base + seniority + numeric_sex, data = salaries[-unus
summary(fit1)
```

Those points are not influential.

A4. Questions of interest Answer the questions of interest. You should provide both a verbal answer and quantitative support for that answer. You are free to choose *how* you support your answers with quantitative evidence, but should make use of the model that was fit above and provide some display of R output or graphics. For example, you might choose to support an answer with a confidence interval; in that case, you should show the code and output for the calculation and interpret the interval.

i. Do the data provide evidence of discrimination on the basis of sex?

```
fit_model <- lm(formula = log(salary) ~ education + age + base + seniority+ experience+numeric_sex, data
summary(fit_model)
```

No. Because the p value of sex is 0.06745 which is obviously larger than 0.05, so the data provides sufficient evidence against that there is discrimination on the basis of sex.

ii. How do median salaries appear to change with age, education, experience, and seniority?

The mean salaries would decrease with the increase of age. They will increase with the increase of education. They will decrease with the increase of experience. They will increase with the increase of seniority.

iii. Do you have any concerns about the model that was used to answer (i) - (ii)? No, because the constant variance assumption and normality are accepted at this time.

Submission instructions

1. Clear your environment and run all codes to check for errors. Resolve any if detected.
2. Input your name in the author information, remove the instructions at the beginning and end of the document, and knit to pdf.
3. Inspect the pdf and fix any display issues.
4. Once the pdf looks good, upload a copy to Gradescope.
5. Download a backup copy of your work and store locally.

Code appendix

```
# knit options
knitr::opts_chunk$set(echo = F,
                      results = 'markup',
                      fig.width = 4,
                      fig.height = 3,
                      fig.align = 'center',
                      message = F,
                      warning = F)

# packages
library(tidyverse)
library(tidymodels)
library(modelr)
# give the data a descriptive name
salaries <- Sleuth3::case1202

# preview
head(salaries)
# check documentation
?Sleuth3::case1202
salaries$age = salaries$Age/12
salaries$experience = salaries$Exper/12
salaries$seniority = salaries$Senior/12
salaries

# example
salaries %>% rename(education = Educ)
salaries <- salaries %>%rename(sex = Sex, base = Bsal, education = Educ, salary = Sal77)

# select columns
salaries <- salaries %>%
  select(salary, base, age, sex, education, experience, seniority)

# preview
head(salaries)
# plotting codes here
salaries %>%
  pivot_longer(cols = c(base, age, education, experience, seniority)) %>%
  ggplot(aes(x = value, y = salary, color = sex))+
  facet_wrap( ~ name, scales = 'free_x', nrow=1)+
  geom_point() +
  labs(x='', y='')

# plotting codes here
# plotting codes here
salaries %>%
  pivot_longer(cols = c(base, age, education, experience, seniority)) %>%
  ggplot(aes(x = value, y = log(salary), color = sex))+
  facet_wrap( ~ name, scales = 'free_x', nrow=1)+
  geom_point() +
  labs(x='', y='')
```

```r
# fitting codes here
## I encode the female to be 1 and male to be 0
salaries <- mutate(.data=salaries, numeric_sex=as.numeric(sex))
model <- lm(log(salary)~ base+ age+ education+experience+ seniority+ numeric_sex,data = salaries)
summary(model)
# plotting codes here
augment(model, salaries) %>%
  pivot_longer(cols = c(.fitted, age, salary, education, experience, seniority, numeric_sex)) %>%
  ggplot(aes(y = .resid, x = value, color = sex)) +
  facet_wrap(~ name, scales = 'free_x',nrow = 7) +
  geom_point() +
  geom_hline(aes(yintercept = 0)) +
  geom_smooth(method = 'loess', formula = 'y ~ x', se = F, span = 1)+
  labs(x='',y='')
# plotting codes here
augment(model, salaries) %>%
  ggplot(aes(sample=.resid))+
  geom_qq()+geom_qq_line()

# plotting codes here

dat <- lm(formula = salary ~ experience + age + education + base + seniority + numeric_sex, data = sala
unusual_obs <- augment(dat, salaries) %>%
  mutate(obs_index = row_number()) %>%
  pivot_longer(cols = c(.resid, .hat, .cooksd)) %>%
  group_by(name) %>%
  slice_max(order_by = abs(value), n = 1) %>%
  ungroup()

p_caseinf <- augment(dat, salaries) %>%
  mutate(obs_index = row_number()) %>%
  pivot_longer(cols = c(.resid, .hat, .cooksd)) %>%
  ggplot(aes(x = obs_index, y=value)) +
  facet_wrap(~ name, scales = 'free_y', nrow=3) +
  geom_point() +
  geom_hline(aes(yintercept = 0)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.25)) +
  labs(x='', y='')

p_caseinf + geom_point(data = unusual_obs, color = 'red')
unusual_obs2 <- unusual_obs %>%
  rename(new_name = name) %>%
  select(salary, base, age, education, experience, seniority, obs_index, new_name, sex) %>%
  pivot_longer(cols=c(base, age, education, experience, seniority))

salaries %>%
  pivot_longer(cols=c(base, age, education, experience, seniority)) %>%
  ggplot(aes(x= value, y = salary, color = sex)) +
  facet_wrap(~name, scales='free_x', nrow = 1) +
  geom_point()+
  labs(x='', y='')+
  geom_point(data = unusual_obs2, aes(shape=new_name), size = 6, alpha = 0.2)
```

```r
unusual_obs %>% pull(salary)

unusual_idx <- augment(dat, salaries) %>%
  mutate(idx = row_number()) %>%
  slice_max(order_by = abs(.resid), n = 1)  %>%
  pull(idx)

unusual_idx2 <- augment(dat, salaries) %>%
  mutate(idx2 = row_number()) %>%
  slice_max(order_by = abs(.hat), n = 1)  %>%
  pull(idx2)

fit1 <- lm((salary)~experience + age + education + base + seniority + numeric_sex, data = salaries[-unus
summary(fit1)



fit_model <- lm(formula = log(salary) ~ education + age + base + seniority+ experience+numeric_sex, data
summary(fit_model)
```