PSTAT 126 Final

Aaron Armbruster

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 <- mutate(salaries,
       age = Age / 12,
       experience = Exper / 12,
       seniority = Senior / 12
       )
```

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)`.)

```
salaries <- mutate(salaries,
                   sex = Sex, education = Educ,base = Bsal,salary = Sal77)
```

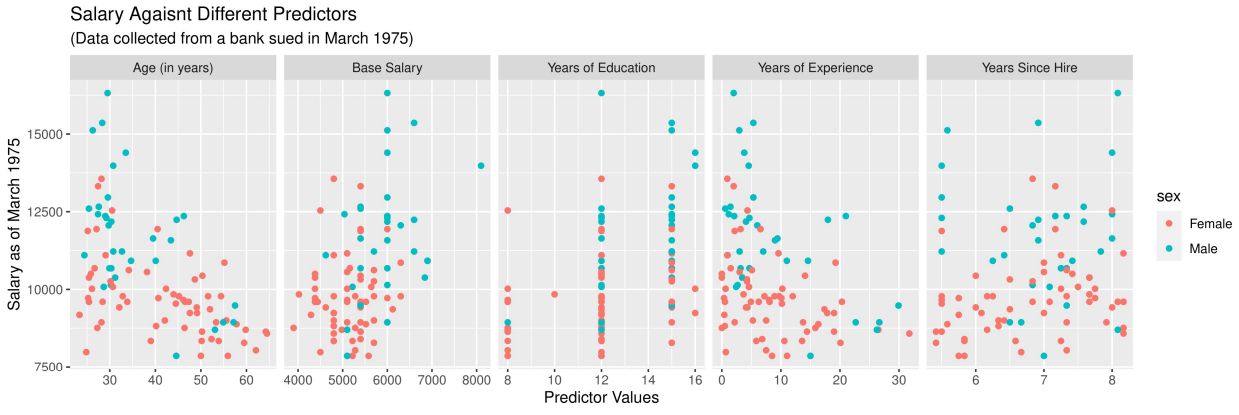
iii. Now select the newly defined and renamed columns by running the chunk below. If you followed the naming instructions in (i) – (ii) correctly, this should run without error.

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


# preview
head(salaries)
```

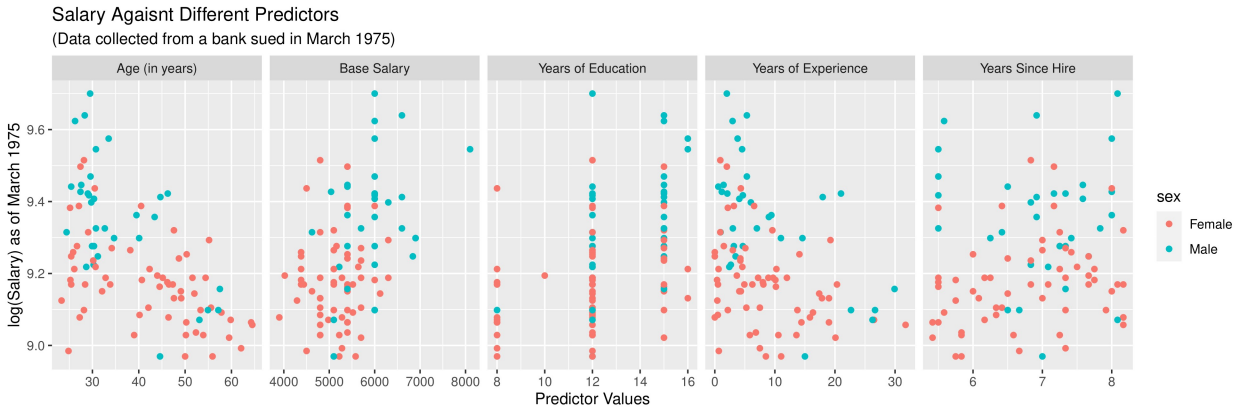
```
##   salary base      age  sex education experience seniority
## 1  12420 5040 27.41667 Male        15   1.166667  8.000000
## 2  12060 6300 29.75000 Male        15   6.000000  6.833333
## 3  15120 6000 26.25000 Male        15   2.958333  5.583333
## 4  16320 6000 29.50000 Male        12   2.000000  8.083333
## 5  12300 6000 29.25000 Male        12   4.666667  5.500000
## 6  10380 6840 31.16667 Male        15   3.458333  7.666667
```

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 log scale seems to appear slightly more linear, but not a very noticeable difference between the two.

iv. Overall, does it appear from the plots that salaries differ between male and female employees?

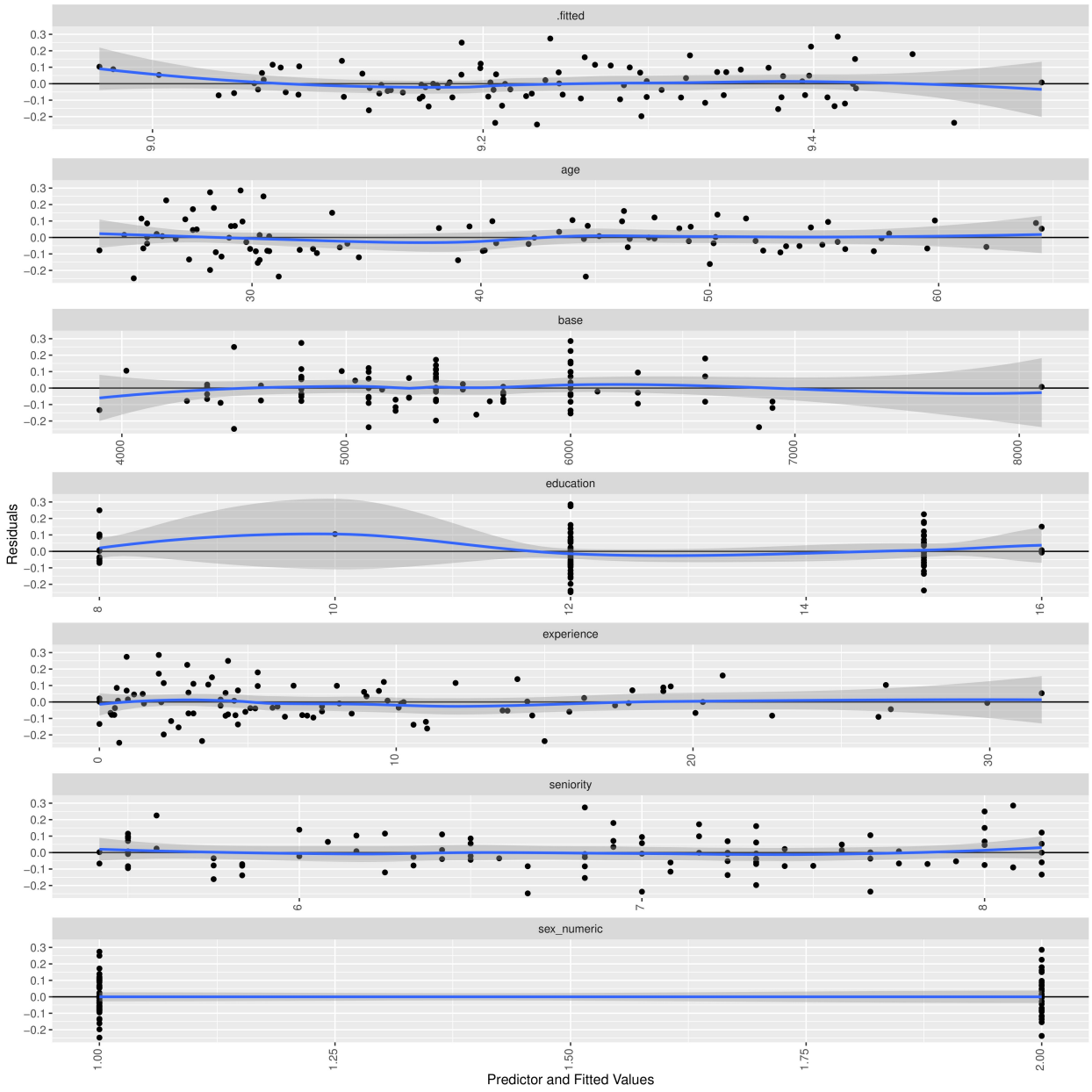
It is hard to state if the salaries differ on average but it does seem that male employees maximum salary is higher in all predictors compared to females.

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.

```
fit <- lm(log(salary) ~ age + base + sex + education + experience + seniority, data = salaries)
```

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.



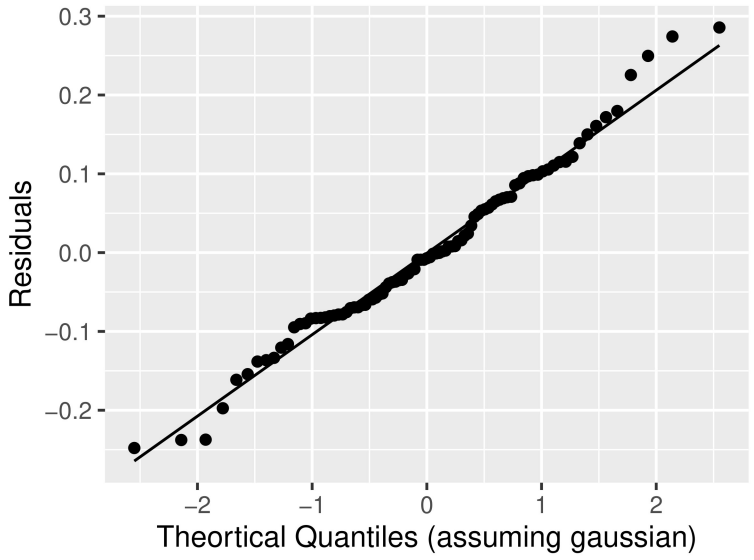
iii. Do you see any problems with model assumptions based on the residual scatterplots? If so, identify the assumption(s) and describe what you see in the plots. Answer in 1-3 sentences.

The distribution of points for experience and age might violate constant variance and if so only slightly, it is difficult to come to a definite conclusion on any violations of the model assumptions from these residual plots.

iv. If you identified any problems, are they important given your goals? (If not, skip this question.)

They are not important

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

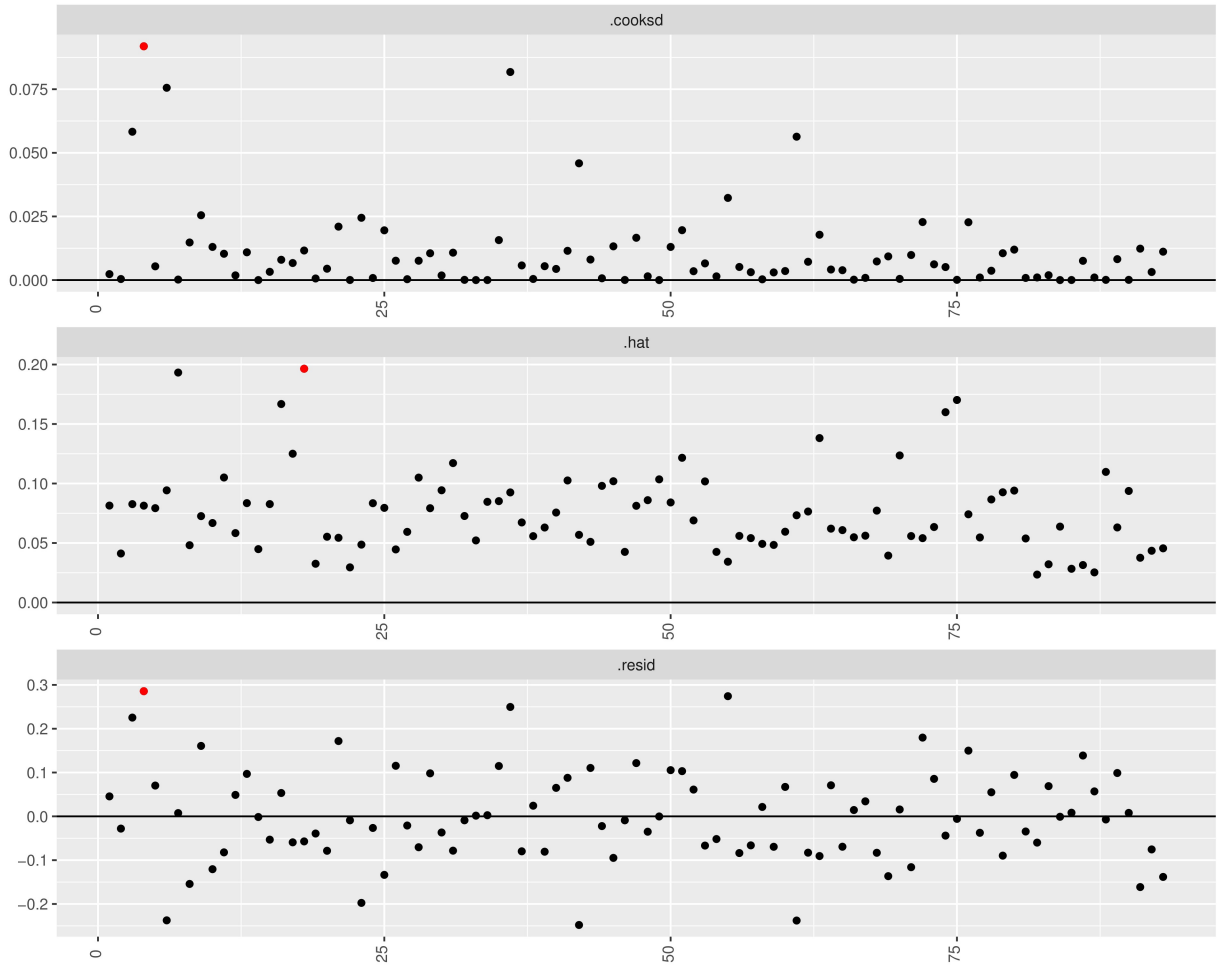


vi. Do you see any issues with model assumptions based on the Q-Q plot? If so, identify the assumption(s) and describe what you see in the plot. Answer in 1-2 sentences.

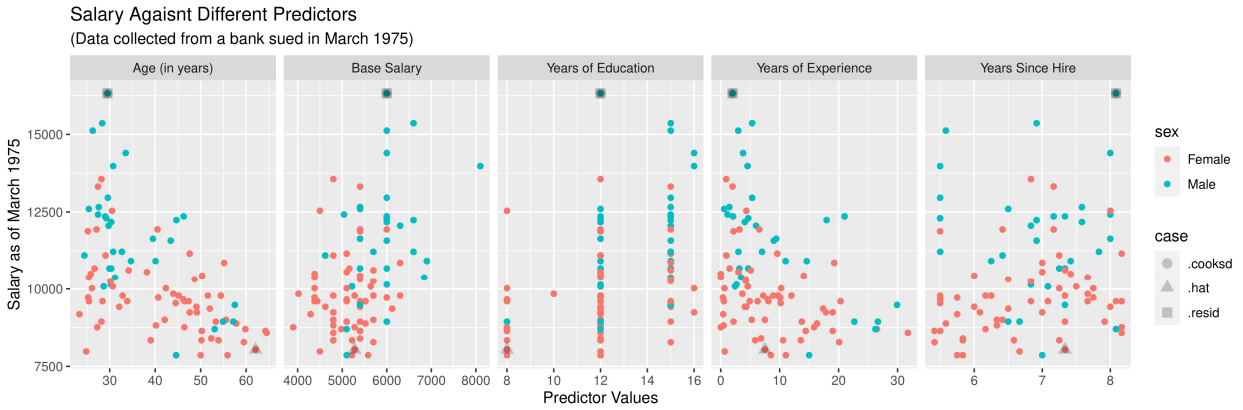
Based on the Q-Q plot the model lines up well.

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.



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

The points seem unusual because there are points that are on the low end of salary yet on the high end of

attributes normally associated with a higher salary. For example there are points with the highest years of education and a large variance in salary.

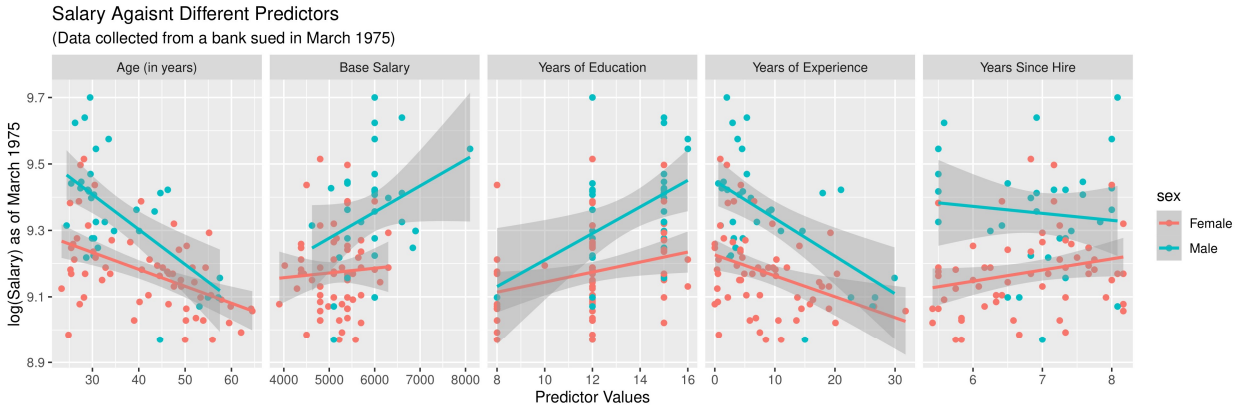
iv. Assess the fit of the model without the observations you highlighted (if any). Are the points, in fact, influential? Answer in 1 sentence and show any codes (but not output) you used to check.

```
unusual_idx <- augment(fit, salaries) %>%
  mutate(idx = row_number()) %>%
  pivot_longer(cols = c(.resid, .hat, .cooksd)) %>%
  group_by(name) %>%
  slice_max(order_by = abs(value), n = 3) %>%
  pull(idx)


fit_drop <- lm(log(salary) ~ age + base + sex + education + experience + seniority,
                  data = salaries[-unusual_idx,])

summary(fit)
summary(fit_drop)
```

Checking the summary for each fit removing the unusual observations had very little changes on the model, the coefficients and the R-Squared changed very little. Because the changes were very insignificant the highlighted 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?



Examining the scatter plots created earlier it can be seen that in all predictors the top earners are males. Even in cases where predictors are equal the top earner is male. The data does provide evidence to show that all the top earners are male, but it can not be said that they are the top earners only because they are male. For example in the age plot it can be seen that men tend to have higher salary, but as men got older there salary went down faster compared to women as their salary stayed more constant as they got older.

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

Going off the plots created in A4i the slopes of each predictor show what happens with changes in the predictor. For age it can be seen that as males get older their salary decreases sharply where as women salary with age has a slow decline. For years of education there is a sharper increase in salary for men than for women. Experience much like age has sharper decrease for men than for women. Then for seniority there is a decrease for men and a increase for women.

iii. Do you have any concerns about the model that was used to answer (i) - (ii)?

No I don't have any concerns.

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)
library(gridExtra)
# give the data a descriptive name
salaries <- Sleuth3::case1202

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

salaries <- mutate(salaries,
                   sex = Sex, education = Educ,base = Bsal,salary = Sal77)


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


# preview
head(salaries)
labels <- c(age = 'Age (in years)',
            base = 'Base Salary',
            education = 'Years of Education',
            experience = 'Years of Experience',
            seniority = 'Years Since Hire')

s_scatter <- salaries %>%
  pivot_longer(col = c(base,age,education,experience,seniority)) %>%
  ggplot(aes(x = value, y = salary)) +
  facet_wrap(~ name, scales = 'free_x',nrow = 1,labeller = labeller(name = labels)) +
  geom_point(mapping = aes(color= sex)) +
  labs(title = "Salary Agaisnt Different Predictors",
       subtitle = "(Data collected from a bank sued in March 1975)",
```

```
       y = "Salary as of March 1975", x ="Predictor Values")

s_scatter



s_scatter1 <- salaries %>%
  pivot_longer(col = c(base,age,education,experience,seniority)) %>%
  ggplot(aes(x = value, y = log(salary))) +
  facet_wrap(~ name, scales = 'free_x',nrow = 1,labeller = labeller(name = labels) ) +
  geom_point(mapping = aes(color = sex)) +
  labs(title = "Salary Agaisnt Different Predictors",
       subtitle = "(Data collected from a bank sued in March 1975)",
       y = "log(Salary) as of March 1975", x ="Predictor Values")

s_scatter1

fit <- lm(log(salary) ~ age + base + sex + education + experience + seniority, data = salaries)


scatter_resid <- augment(fit,salaries) %>%
  mutate(sex_numeric = as.numeric(sex)) %>%
  pivot_longer(cols = c(.fitted,age,base,education,experience,seniority,sex_numeric)) %>%
  ggplot(aes(x = value, y = .resid)) +
  facet_wrap(~ name, scales = 'free_x',nrow=7) +
  geom_point() +
  geom_hline(aes(yintercept = 0)) +
  geom_smooth(method = 'loess',formula =  'y~x') +
  theme(axis.text.x = element_text(angle = 90,vjust = 0.25)) +
  # geom_smooth(method = 'loess', formula = 'y ~ x', se = T, span = 1, mapping = aes(color = sex),level
  labs(x ="Predictor and Fitted Values",y = "Residuals" )

scatter_resid


augment(fit,salaries) %>%
  ggplot(aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line() +
  labs(y = 'Residuals',x= 'Theortical Quantiles (assuming gaussian)')


unusual <- augment(fit, salaries) %>%
  mutate(index = row_number()) %>%
  pivot_longer(cols = c(.resid, .hat, .cooksd)) %>%
  group_by(name) %>%
  slice_max(order_by = abs(value), n = 1) %>%
  ungroup()

case_influence <- augment(fit,salaries) %>%
  mutate(index = row_number()) %>%
  pivot_longer(cols = c(.resid,.cooksd,.hat)) %>%
  ggplot(aes(x = index, y =value)) +
```

```
  facet_wrap(~name, scales = 'free',nrow =3) +
  geom_point() +
  geom_hline(aes(yintercept=0)) +
  theme(axis.text.x = element_text(angle = 90,vjust = 0.25)) +
  labs(x= '',y='')


case_influence + geom_point(data = unusual, color = 'red')



unusual_obs <- unusual %>%
  rename(case = name) %>%
  select(age,base,education,experience,seniority,salary,case,sex) %>%
  pivot_longer(col = c(base,age,education,experience,seniority))

s_scatter + geom_point(data = unusual_obs, size = 3,alpha = 0.2,aes(x=value,y = salary, shape =case))



unusual_idx <- augment(fit, salaries) %>%
  mutate(idx = row_number()) %>%
  pivot_longer(cols = c(.resid, .hat, .cooksd)) %>%
  group_by(name) %>%
  slice_max(order_by = abs(value), n = 3) %>%
  pull(idx)


fit_drop <- lm(log(salary) ~ age + base + sex + education + experience + seniority,
                  data = salaries[-unusual_idx,])

summary(fit)
summary(fit_drop)



re_scatter <- salaries %>%
  pivot_longer(col = c(base,age,education,experience,seniority)) %>%
  ggplot(aes(x = value, y = log(salary),group = sex)) +
  facet_wrap(~ name, scales = 'free_x',nrow = 1,labeller = labeller(name = labels) ) +
  geom_point(mapping = aes(color = sex)) +
  labs(title = "Salary Agaisnt Different Predictors",
       subtitle = "(Data collected from a bank sued in March 1975)",
       y = "log(Salary) as of March 1975", x ="Predictor Values") +
  geom_smooth(method = 'lm',aes(color = sex), formula = 'y~x')

re_scatter
```