# PSTAT 126 Final

Aidan Baker

## 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.

```
# solution

salaries <- mutate(.data = salaries, age = Age/12)
salaries <- mutate(.data = salaries, experience = Exper/12)
salaries <- mutate(.data = salaries, 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)`.)

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

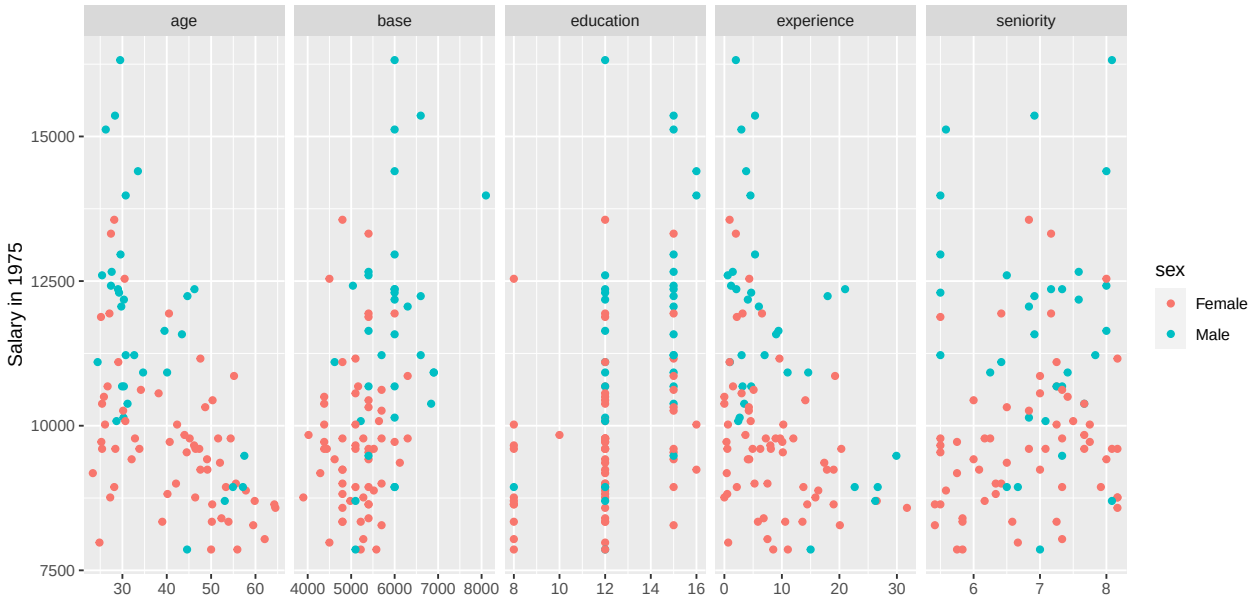
iii. Now select the newly defined and renamed columns by running the chunk below. If you followed the naming instructions in (i) – (ii) correctly, this should run without error.

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

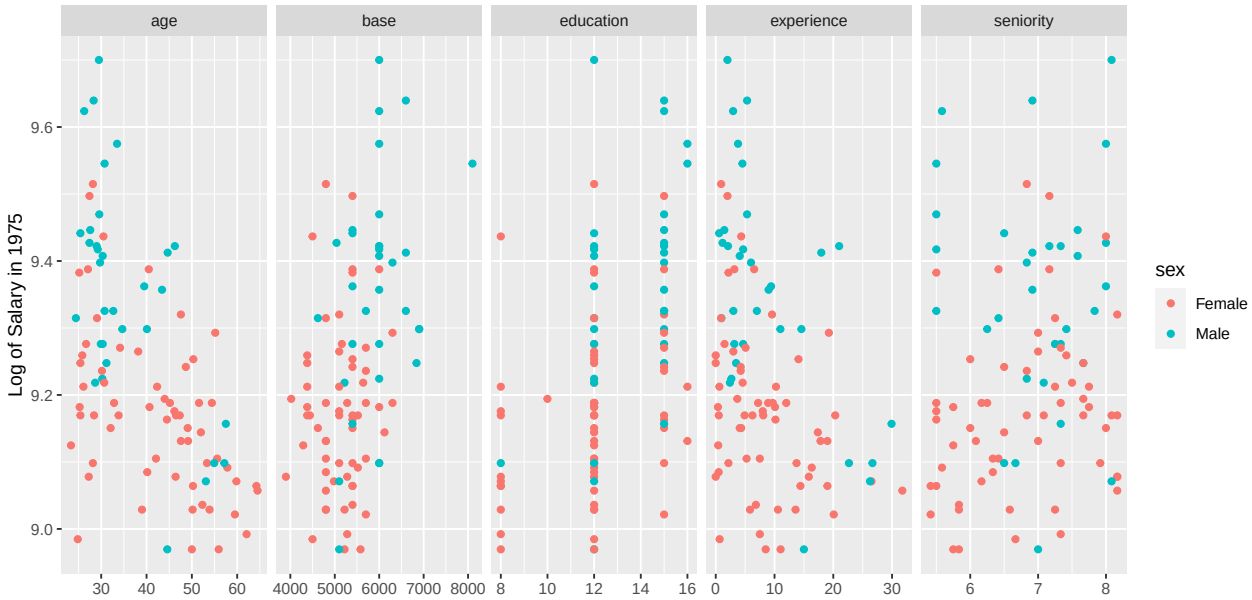
```
# preview
head(salaries)
```

### 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?

Both the relationships appear to look very similar. Once we took the log of the salary, it changed the scale of the y-axis.

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

Yes. Salary is represented by the y axis and after coloring the plots, we see that on average the male salaries are higher.

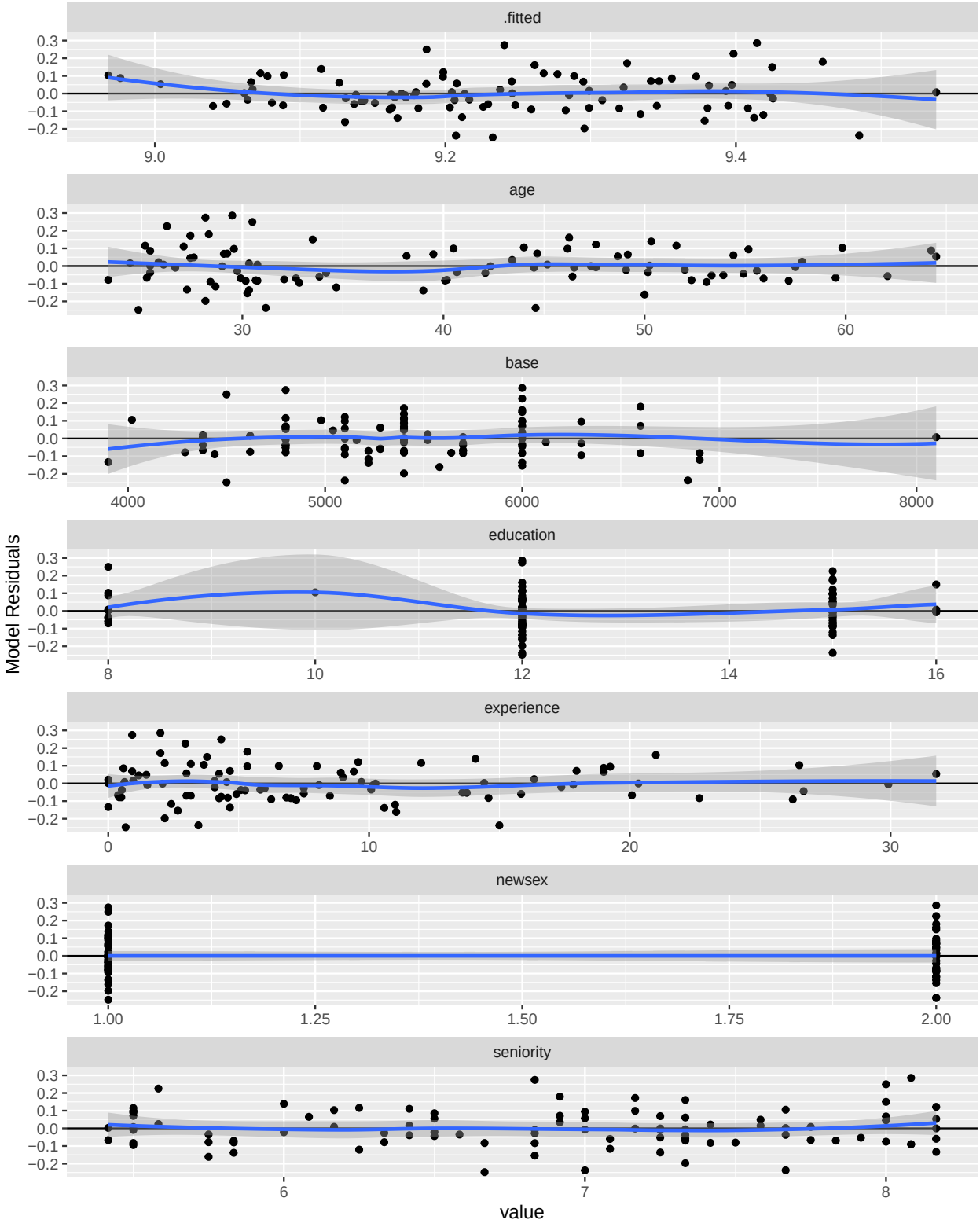
### A2. Model fitting and checking

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

```
# fitting codes here

fit_model <- lm(log(salary) ~ base + age + 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.

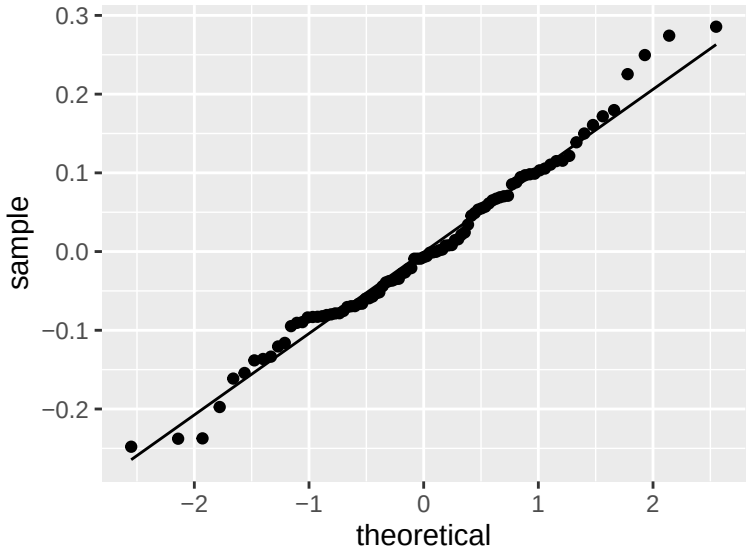
No I don't see any big problems with the model assumptions based on the residual scatterplots. The only potential problem could occur with the education and how it seems like the model could be entered as a

cubic but this is caused by a singular point so I dont think its a problem that needs to be dealt with.

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

Skip

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

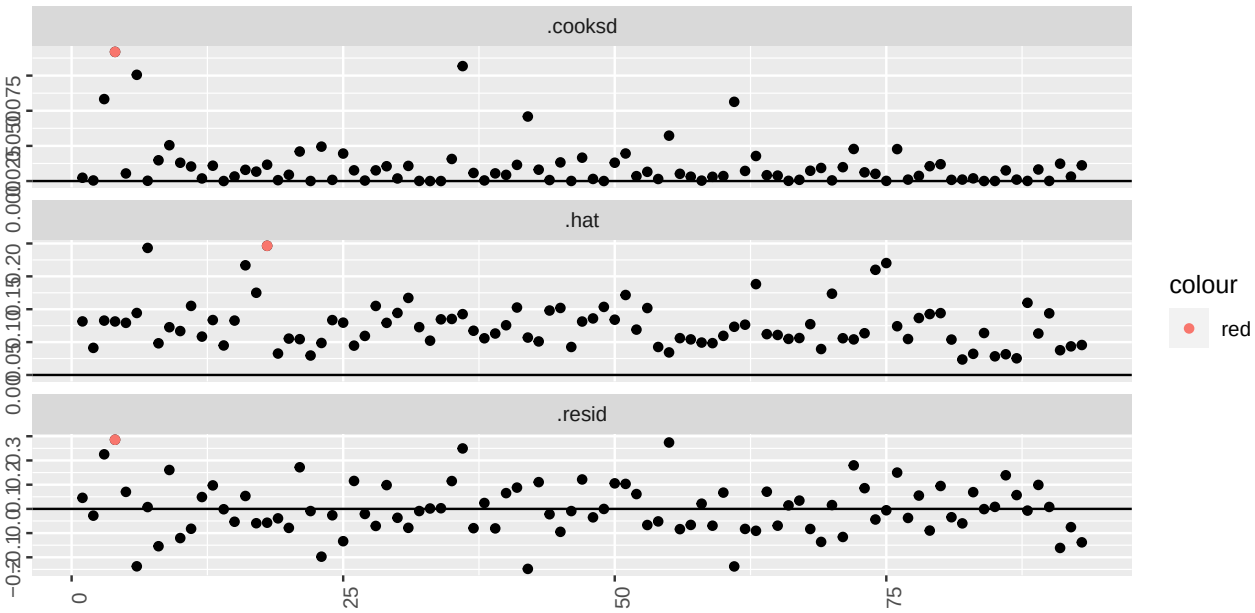



vi. Do you see any issues with model assumptions based on the Q-Q plot? If so, identify the assumption(s) and describe what you see in the plot. Answer in 1-2 sentences.

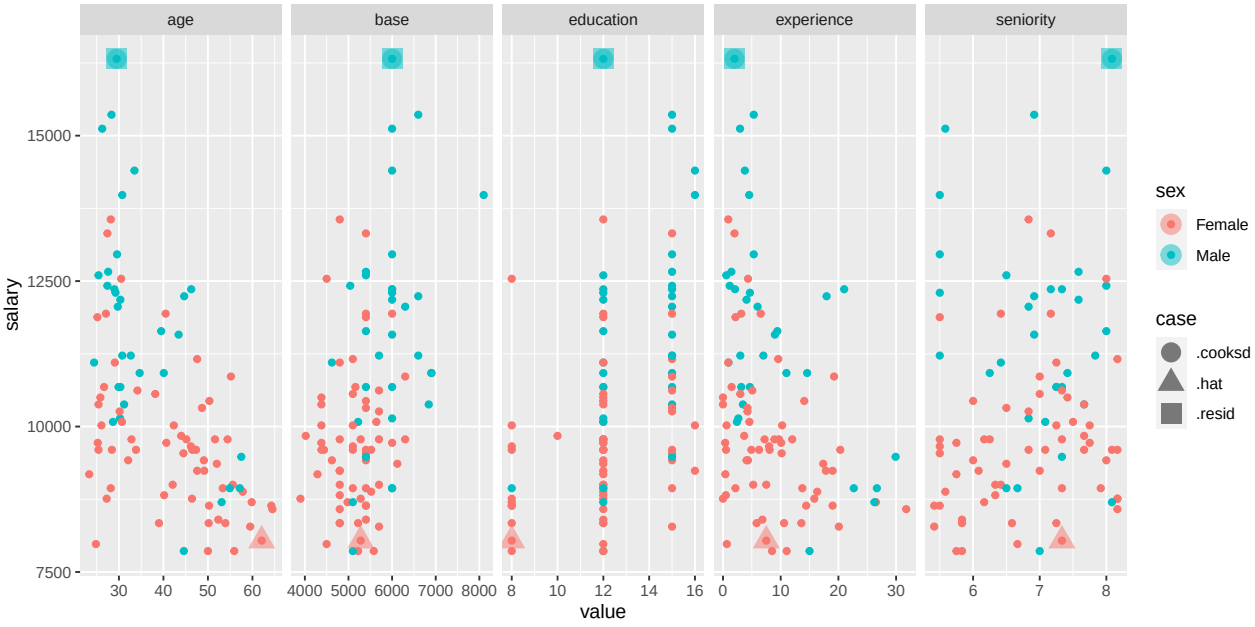
I do not see any any issues with the model assumptions on the Q-Q plot. It appears to be quite normal. Although it is not perfect, it does not look to be skewed in any way such as heavier tails or off in one direction such as a gamma distribution.

**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.

Some of these points appear to be a little unusual, such as someone with very little experience and a young age having some of the highest salaries.

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.

```
unusual1 <- augment(fit,salaries) %>%
  mutate(index1 = row_number()) %>%
  slice_max(order_by = abs(.resid), n=1) %>%
  pull(index1)
```

```
unusual2 <- augment(fit,salaries) %>%
  mutate(index2 = row_number()) %>%
  slice_max(order_by = abs(.hat), n=1) %>%
  pull(index2)

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

fit_withoutbad <- lm(salary ~ base + age + education + experience + seniority, data = salaries[c(-unusua

summary(fit)
summary(fit_withoutbad)
```

By looking at the summaries of the two models, one without the unusual data point, we see that this point in not influential to the model. By pulling up the summaries of these two plots, we can see that the R-squared value differs by less than 1%.

**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?

```
unusual1 <- augment(fit,salaries) %>%
  mutate(index1 = row_number()) %>%
  slice_max(order_by = abs(.resid), n=1) %>%
  pull(index1)

unusual2 <- augment(fit,salaries) %>%
  mutate(index2 = row_number()) %>%
  slice_max(order_by = abs(.hat), n=1) %>%
  pull(index2)

newfit <- lm(salary ~ base + age + sex + education + experience + seniority, data = salaries[c(-unusual

newfit %>% confint()
```

```
##                    2.5 %       97.5 %
## (Intercept) 1925.6425555 10243.848512
## base           0.2278312     1.225931
## age          -99.1378305   -13.701201
## sexMale      -40.2241190  1368.898924
## education     -8.0855925   244.686685
## experience   -82.5313487    41.574997
## seniority   -182.9346601   480.336598
```

We see that the 95% confidence interval of Male sex contains the value 0. Thus the data does not provide any evidence of discrimination on the basis of sex.

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

```
median <- lm(salary ~ age + education + experience + seniority, data = salaries)
summary(median)
```

```
##
```

```
## Call:
## lm(formula = salary ~ age + education + experience + seniority,
##     data = salaries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3355.5  -913.4  -172.4   765.2  5316.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10423.93    1814.65   5.744  1.3e-07 ***
## age           -86.16      22.13  -3.894 0.000191 ***
## education     238.82      67.23   3.552 0.000616 ***
## experience     25.63      32.96   0.778 0.438845
## seniority      25.25     177.73   0.142 0.887363
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1421 on 88 degrees of freedom
## Multiple R-squared:  0.3969, Adjusted R-squared:  0.3695
## F-statistic: 14.48 on 4 and 88 DF,  p-value: 4.004e-09
```

Based off of the summary of the model, we see that salary increases with respect to education, experience, and seniority, and decreases with respect to age.

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

We can see from the residual plots in A2ii that there are not any concerns about the model that was used to answer (ii). For part i, I do not have any concerns with the interpretation of the confidence interval.

## 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
# solution

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

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

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

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

# preview
head(salaries)
# plotting codes here
fit <- lm(salary ~ base + age + education + experience + seniority, data = salaries)

plot <- augment(fit,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(y="Salary in 1975", x= '')

plot
```

```r
# plotting codes here

fit2 <- lm(log(salary) ~ base + age + education + experience + seniority, data = salaries)

plot2 <- augment(fit2,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(y="Log of Salary in 1975", x= '')

plot2


# fitting codes here

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



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

augment(fit_model,salaries) %>%
  mutate(newsex = as.numeric(sex)) %>%
  pivot_longer(cols = c(.fitted,age,base,education,experience,seniority,newsex)) %>%
  ggplot(aes(x=value,y=.resid)) +
  labs(y = 'Model Residuals') +
  facet_wrap(~name, scales = 'free_x', nrow = 7) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_smooth(method = 'loess')


# plotting codes here

augment(fit_model,salaries) %>%
  ggplot(aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line() +
  labs(x = 'theoretical',y = 'sample')



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

plot1 <- augment(fit_model,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 = element_text(angle = 90, vjust = .25)) +
```

```
  labs(x = '', y = '')

unusual <- augment(fit_model,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()

plot1 + geom_point(data = unusual, aes(color = 'red'))

# plotting codes here
#Using A1i

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

A1iplot <- augment(fit,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()


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


A1iplot + geom_point(data = unusual_obs_long,
                     aes(shape = case),
                     size = 5, alpha = .5)


unusual1 <- augment(fit,salaries) %>%
  mutate(index1 = row_number()) %>%
  slice_max(order_by = abs(.resid), n=1) %>%
  pull(index1)

unusual2 <- augment(fit,salaries) %>%
  mutate(index2 = row_number()) %>%
  slice_max(order_by = abs(.hat), n=1) %>%
  pull(index2)

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

fit_withoutbad <- lm(salary ~ base + age + education + experience + seniority, data = salaries[c(-unusua

summary(fit)
summary(fit_withoutbad)


unusual1 <- augment(fit,salaries) %>%
  mutate(index1 = row_number()) %>%
```

```
  slice_max(order_by = abs(.resid), n=1) %>%
  pull(index1)

unusual2 <- augment(fit,salaries) %>%
  mutate(index2 = row_number()) %>%
  slice_max(order_by = abs(.hat), n=1) %>%
  pull(index2)

newfit <- lm(salary ~ base + age + sex + education + experience + seniority, data = salaries[c(-unusual

newfit %>% confint()


median <- lm(salary ~ age + education + experience + seniority, data = salaries)
summary(median)
```