Mini Assignment

Shuxin Tan 1007625447

2023-07-20

```
setwd("/Users/tanshuxin/Desktop/Second Year s/STA302/Mini Assignment")

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
model=read.csv('multiple_linear_regression_dataset.csv', header=TRUE)

attach(model)
```

Question 1:
Age and Income: $Y_i = \beta_{01} + \beta_{11} X_i + \epsilon_i$
Y_i is the observed value of income on unit i and X_i is the observed value of age on unit i. ϵ_i is the error for each pair of X_i and Y_i, which are independent and identically distributed with normal distribution with mean 0 and variance σ^2. β_{01} is the intercept and β_{11} is the slope, which are all parameters.

Experience and Income: $Y_i = \beta_{02} + \beta_{12} X_i + \epsilon_i$
Similarly, Y_i is the observed value of income on unit i and X_i is the observed value of experience on unit i. ϵ_i is the error for each observed pair of X_i and Y_i, which are independent and identically distributed with normal distribution with mean 0 and variance σ^2. β_{02} is the intercept and β_{12} is the slope, which are all parameters.

Question 2:

```
simple.fit_1 = lm(income~age, data=model)
summary(simple.fit_1)
```

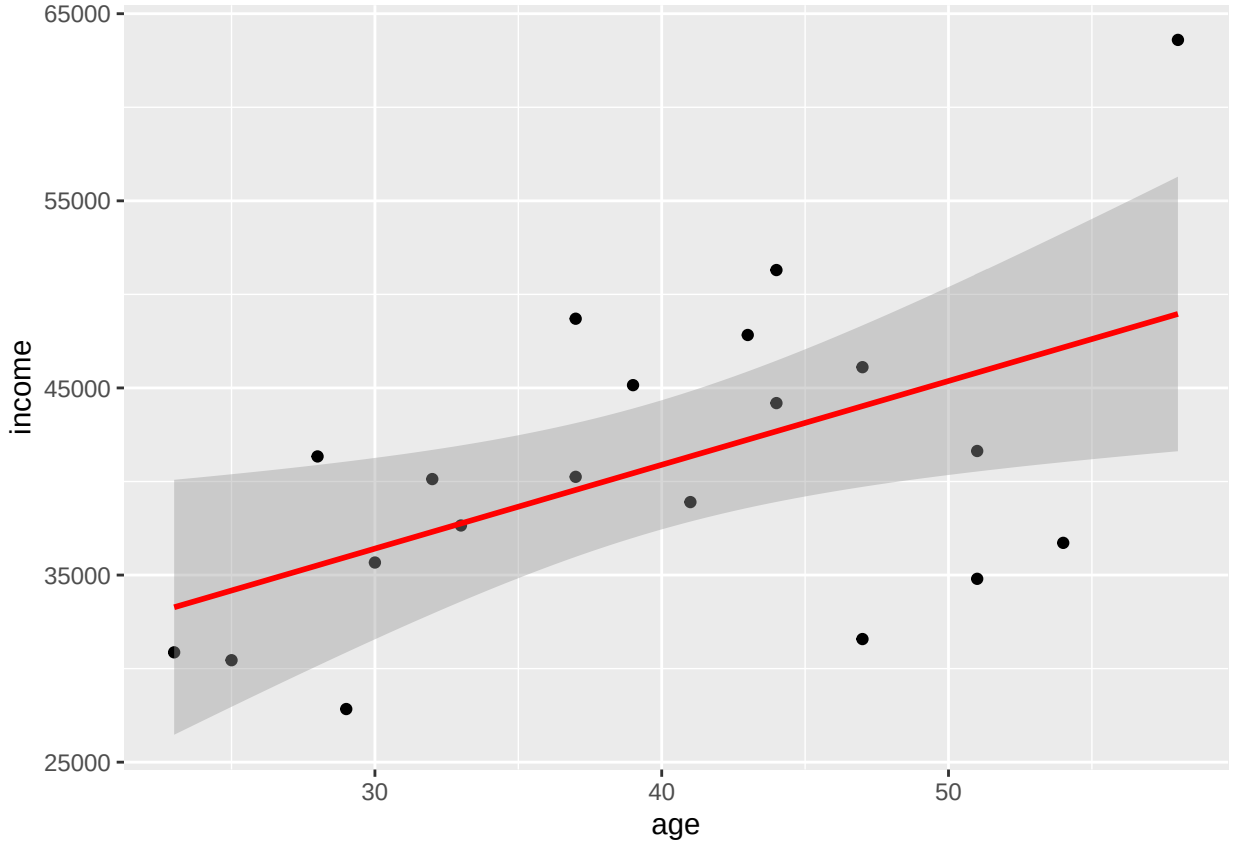
```
##
## Call:
## lm(formula = income ~ age, data = model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12447.8  -3839.9    297.4   4927.7  14645.0
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22975.2     6858.5   3.350  0.00357 **
## age            447.9      167.9   2.667  0.01571 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7341 on 18 degrees of freedom
## Multiple R-squared:  0.2832, Adjusted R-squared:  0.2434
## F-statistic: 7.113 on 1 and 18 DF,  p-value: 0.01571
```

```
library(ggplot2)

ggplot(model, aes(x = age, y = income)) +
  geom_point() +
  stat_smooth(method = "lm", col = "red")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
b0_1 <- summary(simple.fit_1)$coefficients[1, 1]
b1_1 <- summary(simple.fit_1)$coefficients[2, 1]

b0_1
```

```
## [1] 22975.16
```

```
b1_1
```

```
## [1] 447.9278
```

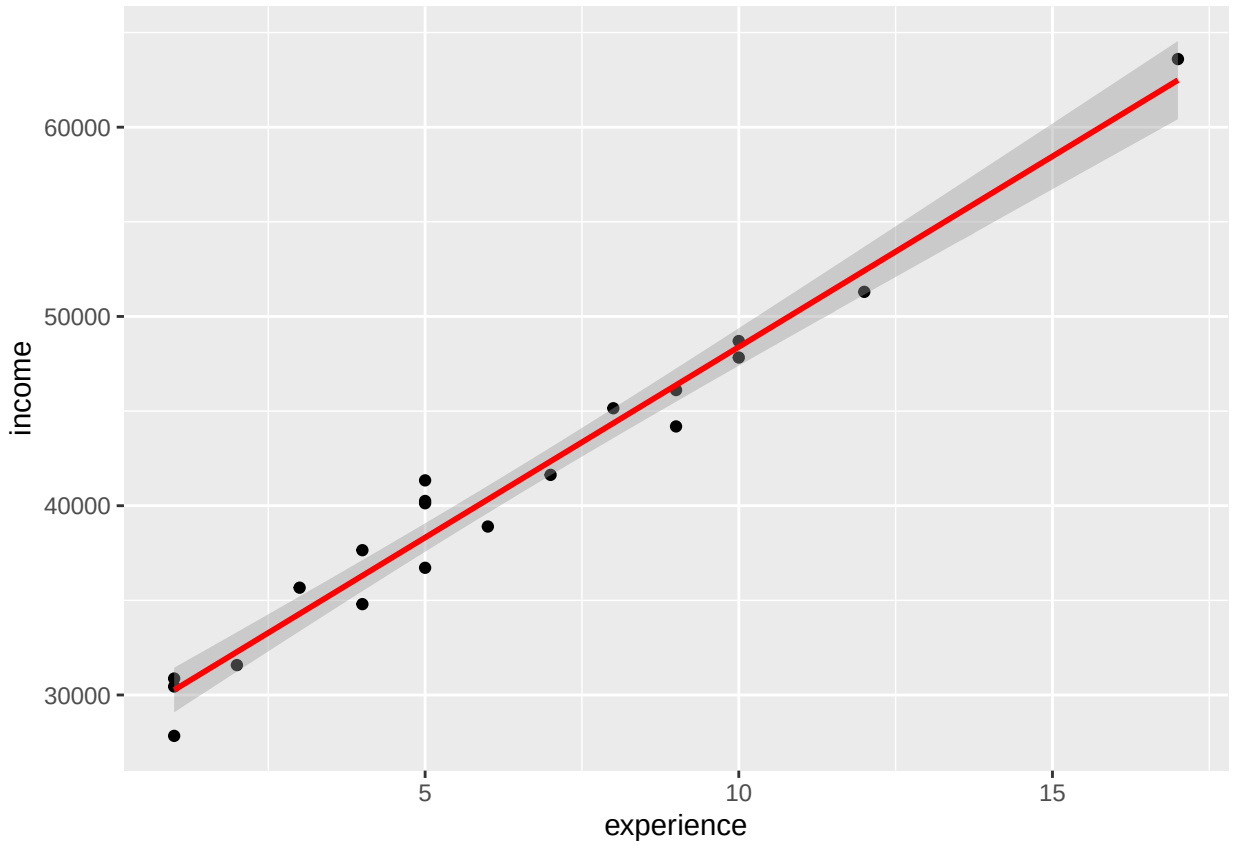
```r
simple.fit_2 = lm(income~experience, data=model)
summary(simple.fit_2)
```

```
##
## Call:
## lm(formula = income ~ experience, data = model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2422.49 -1195.88   -38.65  1170.99  3021.35
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28248.45     630.51    44.8  < 2e-16 ***
## experience   2014.04      85.33    23.6 5.42e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1534 on 18 degrees of freedom
## Multiple R-squared:  0.9687, Adjusted R-squared:  0.967
## F-statistic: 557.1 on 1 and 18 DF,  p-value: 5.422e-15
```

```r
library(ggplot2)

ggplot(model, aes(x = experience, y = income)) +
  geom_point() +
  stat_smooth(method = "lm", col = "red")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
b0_2 <- summary(simple.fit_2)$coefficients[1, 1]
b1_2 <- summary(simple.fit_2)$coefficients[2, 1]

b0_2
```

```
## [1] 28248.45
```

```
b1_2
```

```
## [1] 2014.041
```

By using the lm method in R, we know the value of $\hat{\beta}_0$ and $\hat{\beta}_1$ for both estimated linear regression lines from the summary of the lm method.
Then,
Age and Income: $\hat{Y} = 22975.16 + 447.9278X$
Experience and Income: $\hat{Y} = 28248.45 + 2014.041X$

Question 3:

```
ssreg_2 <- b1_2^2 * sum((experience - mean(experience))^2)
sstot_2 <- sum((income - mean(income))^2)
R_2 <- sqrt(ssreg_2/sstot_2)
R_2
```

```
## [1] 0.9842266
```

The correlation coefficient r for the model examining the relationship between income and experience is 0.9842266.

Question 4:

```r
ssreg_1 <- b1_1^2 * sum((age - mean(age))^2)
sstot_1 <- sum((income - mean(income))^2)
R_1 <- sqrt(ssreg_1/sstot_1)
R_1
```

```
## [1] 0.5322043
```

The correlation coefficient r for the model examining the relationship between income and age is 0.5322043.

Question 5:
Since r also equals $\frac{Cov(X,Y)}{\sigma_X \sigma_Y}$, then $\hat{\beta}_1 = \frac{\sigma_Y}{\sigma_X}\frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sigma_Y}{\sigma_X} r$. So, as correlation coefficient r for income and age is 0.5322043 which is considered to be moderate, the regression coefficient $\hat{\beta}_1$ also becomes relatively less meaningful.

Question 6:

```r
library("gridExtra")
```

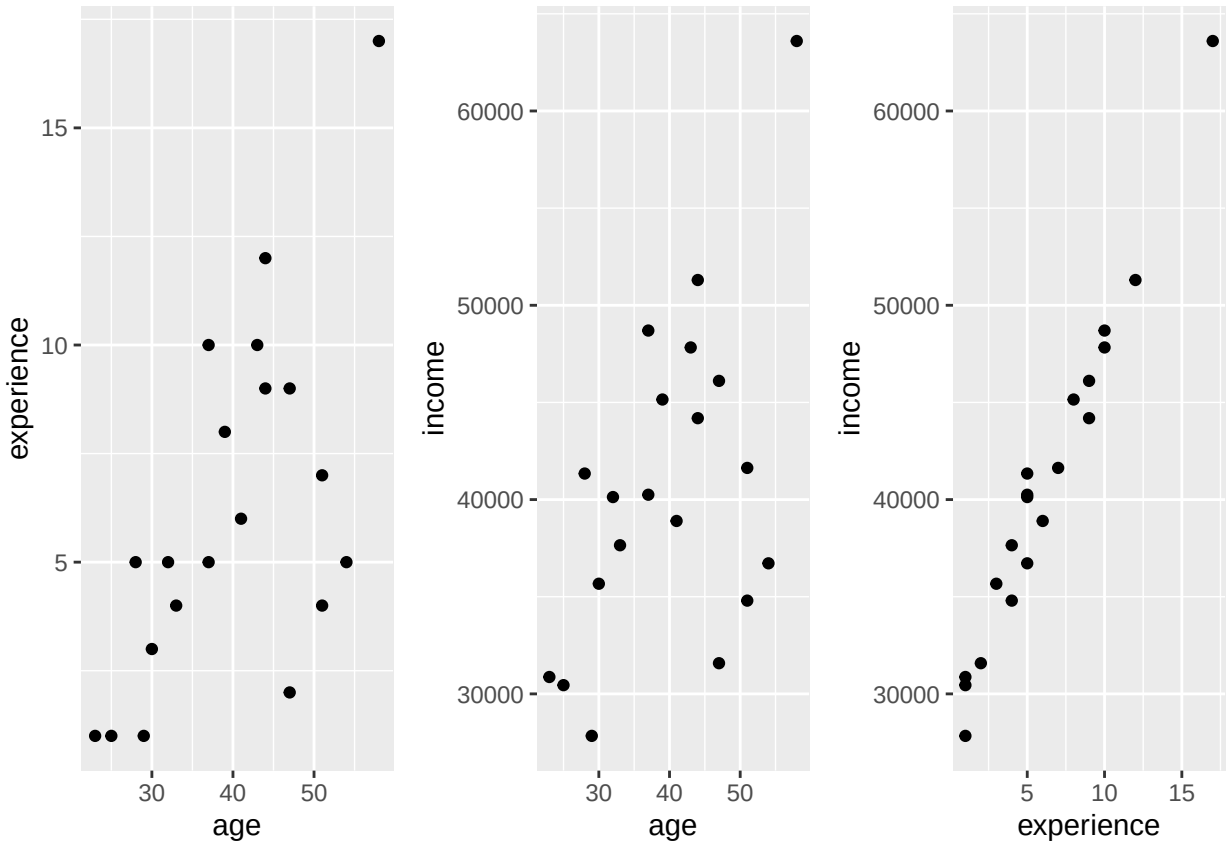
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(ggplot2)

ggp1 <- ggplot(model, aes(x = age, y = experience)) +
  geom_point()
ggp2 <- ggplot(model, aes(x = age, y = income)) +
  geom_point()
ggp3 <- ggplot(model, aes(x = experience, y = income)) +
  geom_point()

grid.arrange(ggp1, ggp2, ggp3, ncol = 3)
```



From the three scatter plots, we can see that generally as X_i increases, Y_i also increases for these three scatter plots. However, the relation between experience and age, as well as the relation between income and age are relatively weak, as the plots in these two graphs are more scattered. The relation between income and experience is much stronger, showing a more linear correlation. Also, in question 3, we know the correlation coefficient for income and experience is 0.9842266 which is very closed to 1. This suggests there is a strong positive linear relationship between experience and income. Therefore, a simple linear regression is more appropriate. The final best model is the simple linear regression model with experience being the predictor variable and income being the response variable. And since the relation between experience and age, as well as the relation between income and age are relatively weak, it's hard to tell if there is any multiple linear regression.

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

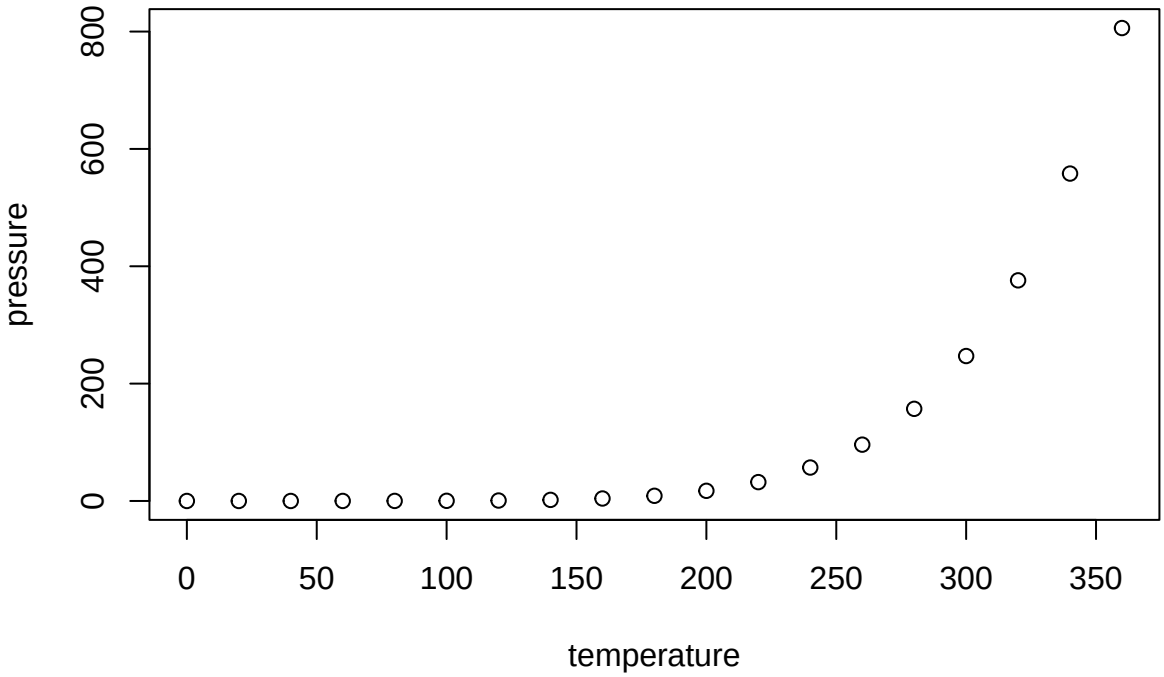
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

1. Complete the ANOVA table using the given results. Show all your work below the table.

Source of variation	Sum of squares	Degrees of freedom	Mean squares
Regression	29.23	1	29.23
Residual	3.5	2	1.75
Total	32.73	3	———

$n=4$

$df\ Reg = 1 \quad df\ Res = n-2 = 2 \quad df\ Tot = df\ Reg + df\ Res = 2+1 = 3$

$\hat{\beta}_0 = 11.5,\ \hat{\beta}_1 = -1.5$

Since $MS\ Res = \frac{SS\ Res}{n-2} = \frac{SS\ Res}{2} = 1.75$,

then $SS\ Res = 1.75 \times 2 = 3.5$

$$t = \frac{\hat{\beta}_1 - \beta_1^{\circ}}{S(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{S(\hat{\beta}_1)} = -4.087$$

$$S(\hat{\beta}_1) = \frac{\hat{\beta}_1}{-4.087} = \frac{-1.5}{-4.087} = \frac{1500}{4087}$$

$$S^2(\hat{\beta}_1) = \left(\frac{1500}{4087}\right)^2 \approx 0.1347$$

Since $S^2[\hat{\beta}_1] = \frac{S^2}{\sum_{j=1}^{4}(x_j-\bar{x})^2} = \frac{MS\ Res}{\sum_{j=1}^{4}(x_j-\bar{x})^2} = \frac{1.75}{\sum_{j=1}^{4}(x_j-\bar{x})^2} = 0.1347$

then $\sum_{j=1}^{4}(x_j-\bar{x})^2 \approx 12.99166$

So, $SS\ Reg = \hat{\beta}_1^2 \sum_{i=1}^{4}(x_i-\bar{x})^2 = (-1.5)^2 \cdot 12.99166 \approx 29.23$

So, $SS\ Tot = SS\ Reg + SS\ Res$

$= 29.23 + 3.5$

$= 32.73$

2. Compute the F statistic using the ANOVA table. Is there something special about this statistic?

$$F^* = \frac{MS\,Reg}{MS\,Res} = \frac{29.23}{1.75} = 16.70$$

$$t^2 = 4.087^2 \approx 16.70$$

Thus, we can see that $F^* = t^2$ and it only holds for simple linear regression.