Advanced Econometric hw 2

LEO LIN

2025-03-12

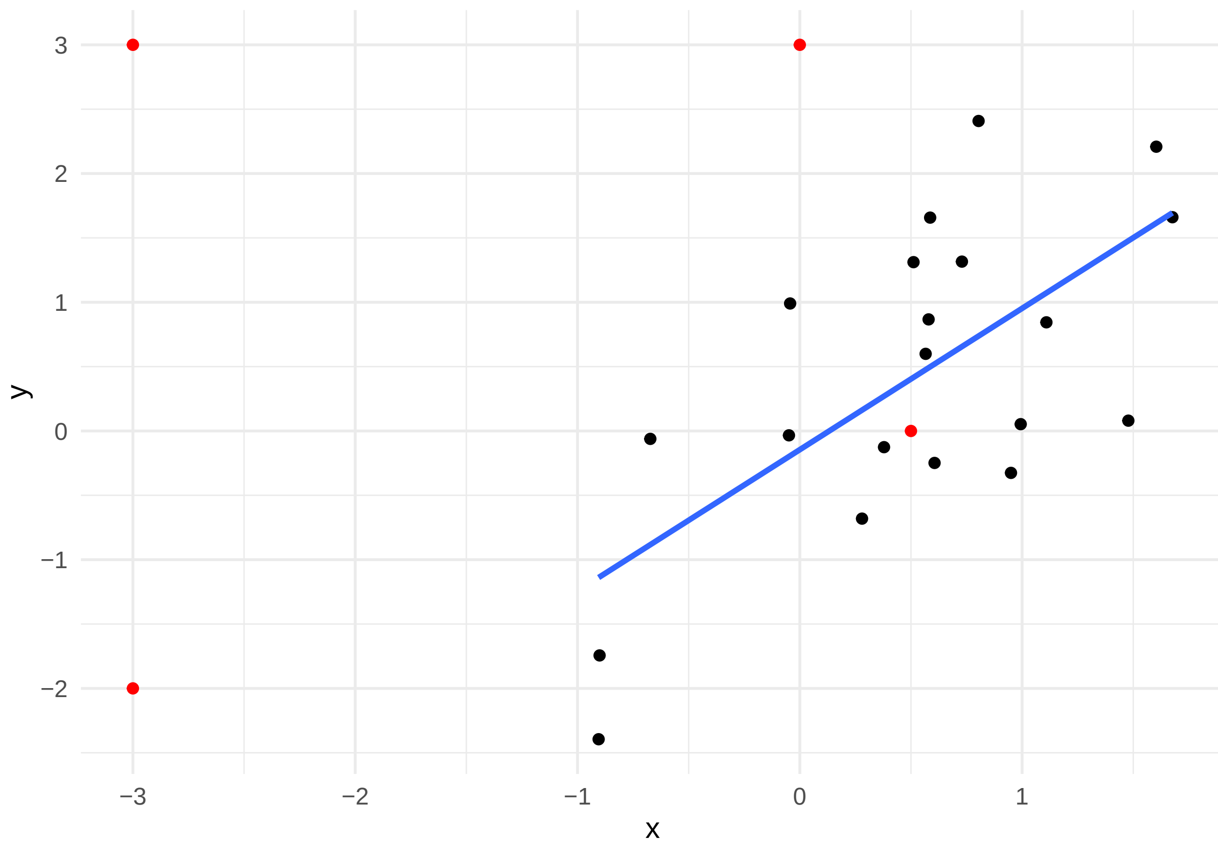
1

preamble

```
library(tidyverse)
library(magrittr)
library(readxl)
setwd("F:\\ \\  \\   \\  2\\hw 2\\data")
```

(a)

```
#simulate a random sample of size 20 , x~N(0,1) and e~N(0,0.75^2)
set.seed(0438)
x <- rnorm(20, 0, 1)
e <- rnorm(20, 0, 0.75)
y <-1*x + e
df <- data.frame(x, y)

#construct the data frame with A(-3,3)B(0,3)C(-3,-2)D(0.5,0)
df2 <- data.frame(x = c(-3, 0, -3, 0.5), y = c(3, 3, -2, 0))

# produce the scatter plot with the regression line of y on x(including the #intercept) based on the 20
a<-ggplot(df,aes(x = x, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal()

# add the 4 new points
a + geom_point(data = df2, aes(x = x, y = y), color = "red")
```



(b)

(i)

```
#simulate a random sample of size 20 , x~N(0,1) and e~N(0,0.75^2)
set.seed(0438)
x <- rnorm(20, 0, 1)
e <- rnorm(20, 0, 0.75)
y <-1*x + e
df <- data.frame(x, y)

# add the point A to df
n<-data.frame(x =-3 , y =3)

df <- rbind(df,n ) #adjust the parameters of the point

# compute the leverage value h_{ii} for point A
X <- cbind(1, df$x)
X_i <- X[21,]
h_ii <- t(X_i) %*% solve(t(X) %*% X) %*% X_i

# compute the leave one out residual foe point A
y_hat <- X %*% solve(t(X) %*% X) %*% t(X) %*% df$y
e_i_hat <- df$y[21] - y_hat[21]
```

```r
e_i_telda <- e_i_hat / (1 - h_ii)
print(h_ii)
```

```
##           [,1]
## [1,] 0.5504155
```

```r
print(e_i_telda)
```

```
##          [,1]
## [1,] 6.438592
```

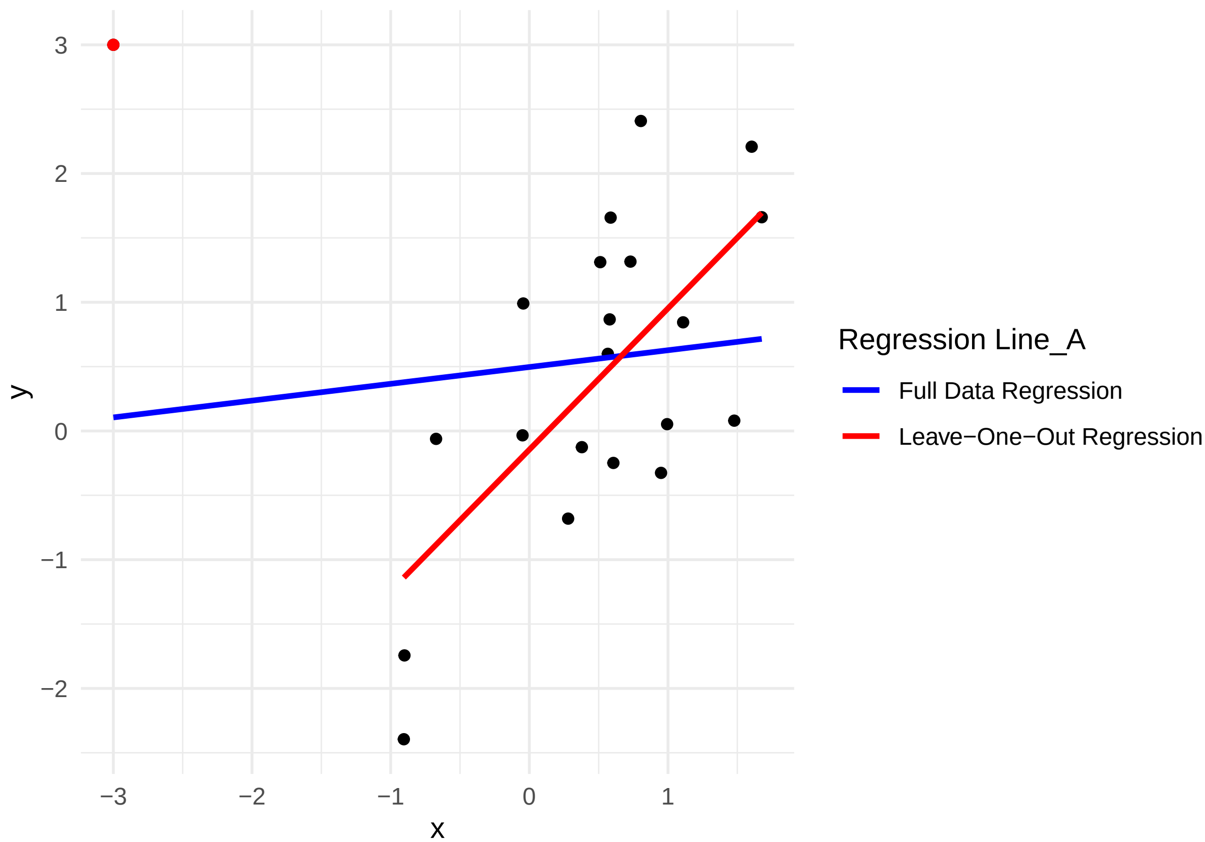
(ii)

The slope of $\tilde{\beta}$ is not necessarily larger the farther the point is, because it is influenced by both h_{ii} and $\tilde{e}$.

```r
#simulate a random sample of size 20 , x~N(0,1) and e~N(0,0.75^2)
set.seed(0438)
x <- rnorm(20, 0, 1)
e <- rnorm(20, 0, 0.75)
y <-1*x + e
df <- data.frame(x, y)

# add the point A to df
p<-data.frame(x =-3 , y =3)

df <- rbind(df,p ) #adjust the parameters of the point

b <-ggplot(df,aes(x = x, y = y)) +
  geom_point() +
  theme_minimal()

df3 <- p
df4 <- df[-21,]

b + geom_point(data = df3, aes(x = x, y = y), color = "red") +
  geom_smooth(data = df, aes(x = x, y = y, color = "Full Data"), method = "lm", se = FALSE) +
  geom_smooth(data = df4, aes(x = x, y = y, color = "Leave-One-Out"), method = "lm", se = FALSE) +
  scale_color_manual(values = c("Full Data" = "blue", "Leave-One-Out" = "red"),
                     labels = c("Full Data" = "Full Data Regression ",
                                "Leave-One-Out" = "Leave-One-Out Regression ")) +
  labs(color = "Regression Line_A")
```

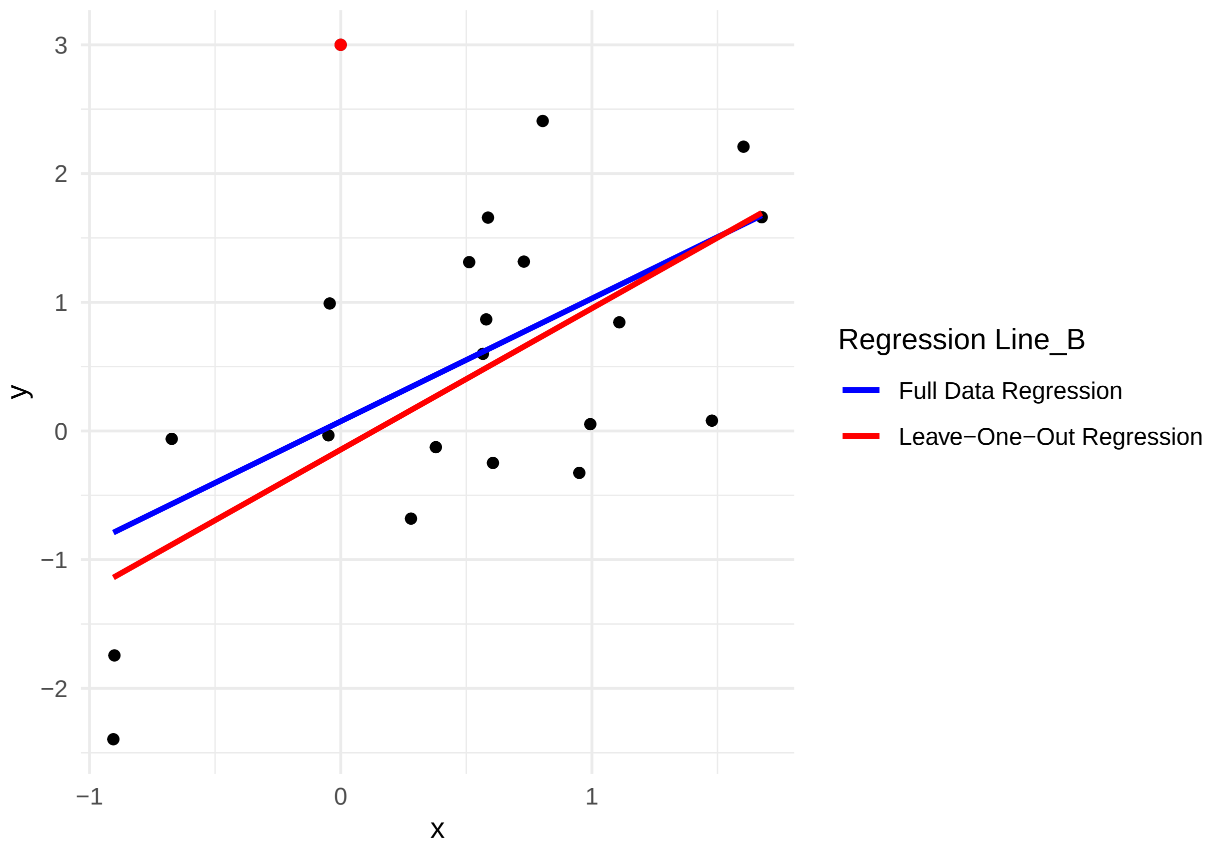
Point A

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



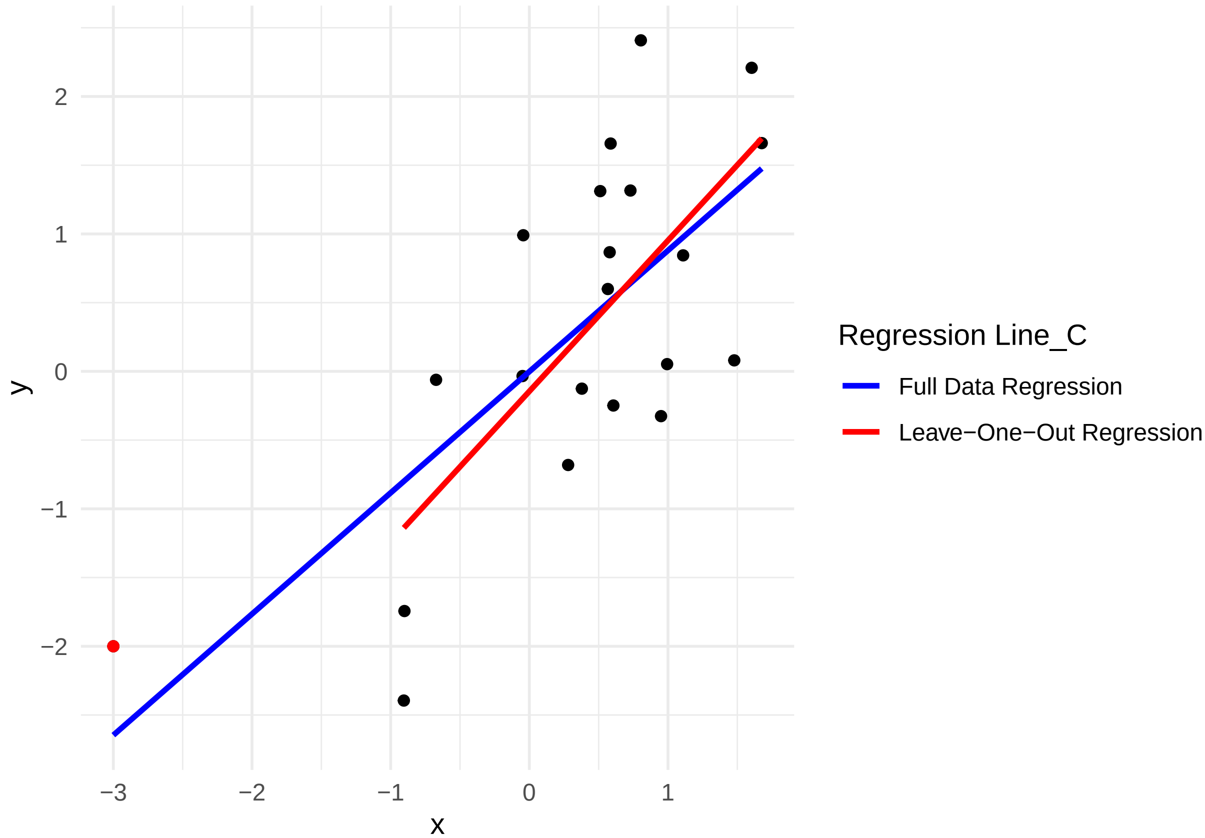
Point B

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



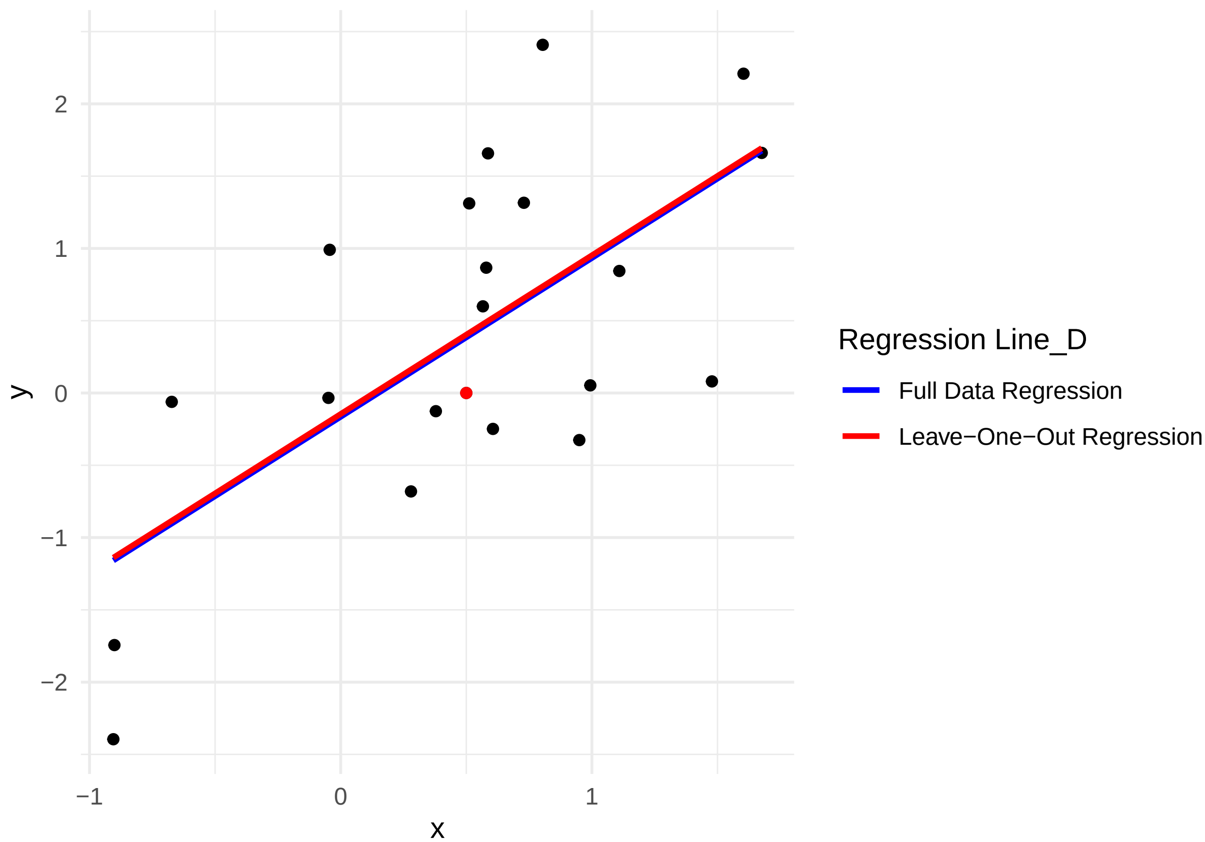
Point C

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



Point D

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



2

(a)

```r
# data cleaning
# import the data
df <- read_xlsx(".\\data\\cps09ma_rename.xlsx")
head(df)
```

```
## # A tibble: 6 x 12
##     age female  hisp education earnings hours weeks union uncov region  Race
##   <dbl>  <dbl> <dbl>     <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1    52      0     0        12   146000    45    52     0     0      1     1
## 2    38      0     0        18    50000    45    52     0     0      1     1
## 3    38      0     0        14    32000    40    51     0     0      1     1
## 4    41      1     0        13    47000    40    52     0     0      1     1
## 5    42      0     0        13   161525    50    52     1     0      1     1
## 6    66      1     0        13    33000    40    52     0     0      1     1
## # i 1 more variable: marital <dbl>
```

```r
# clean the data
df1<- df %>% filter(Race == 4,marital == 7,female == 0) %>%
```

```r
  select(education,age,earnings,hours,weeks) %>%
  mutate(lwage = log(earnings/hours/weeks),exper = age - education - 6
         ,exper2 = exper^2/100) %>%
  select(lwage,education,exper,exper2)
head(df1)
```

```
## # A tibble: 6 x 4
##   lwage education exper exper2
##   <dbl>     <dbl> <dbl>  <dbl>
## 1  1.85        13     2   0.04
## 2  3.87        18    14   1.96
## 3  2.82        14     5   0.25
## 4  3.56        20    29   8.41
## 5  3.30        18     3   0.09
## 6  2.53        12     9   0.81
```

(b)

compute the influence

$$\text{Influence} = \max_{1 \le i \le n} \left| \hat{Y}_i - \tilde{Y}_i \right|$$

```r
# lm model
X <- model.matrix(lwage ~ education + exper + exper2, data = df1)

# residual
e_hat <- resid(lm(lwage ~ education + exper + exper2, data = df1))

# projection matrix
H <- X %*% solve(t(X) %*% X) %*% t(X)

# diagonal elements of the hat matrix
h_ii <- diag(H)

# e_telda
e_telda <- e_hat / (1 - h_ii)

# absoulte value of y_hat - y_telda
influence_candidate <- abs(h_ii * e_telda)

print(which.max(influence_candidate))
```

```
## 35
## 35
```

```r
influence_candidate[35]
```

```
##        35
## 0.2926396
```

(c)

The estimated growth rate of wage with and without the 35th observation will be different significantly, there is a gap of 29.26396% between the two estimated growth rates.

(d)

```r
#recompute the estimated growth rate of wage withdraw the 35th observation
df2 <- df1[-35,]
X2 <- model.matrix(lwage ~ education + exper + exper2, data = df2)
e_hat2 <- resid(lm(lwage ~ education + exper + exper2, data = df2))
# projection matrix
H2 <- X2 %*% solve(t(X2) %*% X2) %*% t(X2)

# diagonal elements of the hat matrix
h_ii2 <- diag(H2)
# 0.253801542 %in% e_hat2
# e_telda2
e_telda2 <- e_hat2 / (1 - h_ii2)
# absoulte value of y_hat - y_telda
influence_candidate2 <- abs(h_ii2 * e_telda2)
influence_candidate2[34]
```

```
##        34
## 0.1065732
```

After removing the 35th observation, the 34th observation becomes the most influential point, but compared to the 35th observation, the 34th observation has a smaller influence value.