

# HW1

Sohaib Syed

09/04/2022

## Contents

<b>Problem 1</b>	<b>1</b>
a . . . . .	1
b . . . . .	2
<b>Problem 2</b>	<b>4</b>
a . . . . .	4
b . . . . .	5
<b>Problem 3</b>	<b>6</b>
a . . . . .	6
<b>Problem 4</b>	<b>6</b>
a . . . . .	6
b . . . . .	6
c . . . . .	9
d . . . . .	10
<b>Problem 5</b>	<b>11</b>
<b>Problem 6</b>	<b>11</b>

## Problem 1

a

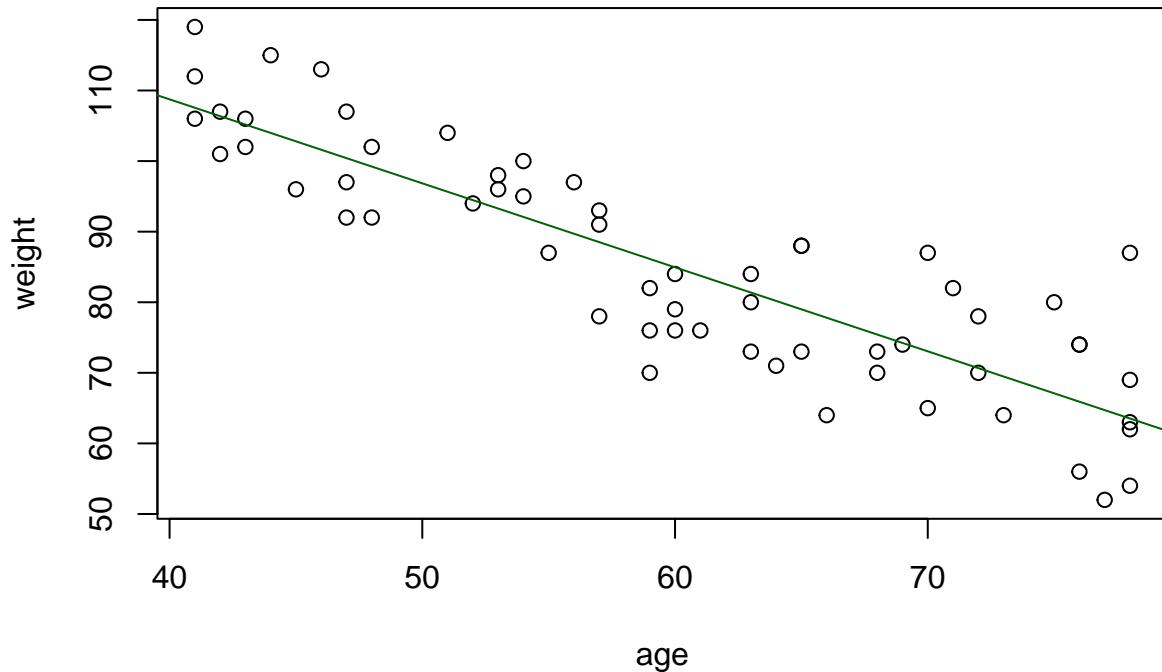
```
my_data <- read.delim("CH01PR27.txt", sep = "", header=FALSE)
colnames(my_data) <- c("Y", "X")
my_ordereddata <- my_data[order(my_data$X, decreasing = FALSE), ]
reg <- lm(Y ~ X, data=my_data)
reg
```

```

## Call:
## lm(formula = Y ~ X, data = my_data)
##
## Coefficients:
## (Intercept)          X
##       156.35      -1.19

plot(Y ~ X, data=my_data, xlab="age", ylab="weight")
abline(reg, col = "darkgreen")

```



The regression function is  $y=156.35-1.19x$ . The regression function that I found does appear to fit well to the data. The data supports the anticipation that as age increases muscle mass decreases.

b

1

```

count=0
difference=0
for(i in 1:nrow(my_ordereddata)) { # for-loop over rows
  for(j in i:nrow(my_ordereddata)){
    if (abs(my_ordereddata[i,2]-my_ordereddata[j,2])==1){
      count=count+1

```

```

        difference=difference+((my_ordereddata[i,1]-my_ordereddata[j,1]))
    }
}
mean=difference/count
mean

## [1] 2.205128

```

Point estimate for women differing in age by one year is 2.205128

**2**

```

count60s=0
sum60s=0
for (i in 1:nrow(my_ordereddata)){
  if (my_ordereddata[i,2]==60){
    count60s=count60s+1
    sum60s=sum60s+my_ordereddata[i,1]
  }
}
sum60s/count60s

```

```
## [1] 79.66667
```

Point estimate for women aged X=60 years is 79.66667

**3**

the eighth data point in the data set is [41,112]

```
my_data[8,]
```

```
##      Y   X
## 8 112 41
```

According to the regression function at 41 years old, weight should be 107.56. Thus the residual is 4.44.

```
y=156.35-1.19*(41)
my_data[8,1]-y
```

```
## [1] 4.44
```

**4**

```
var(my_data$Y)

## [1] 262.7446

Varince point estimate is 262.7446
```

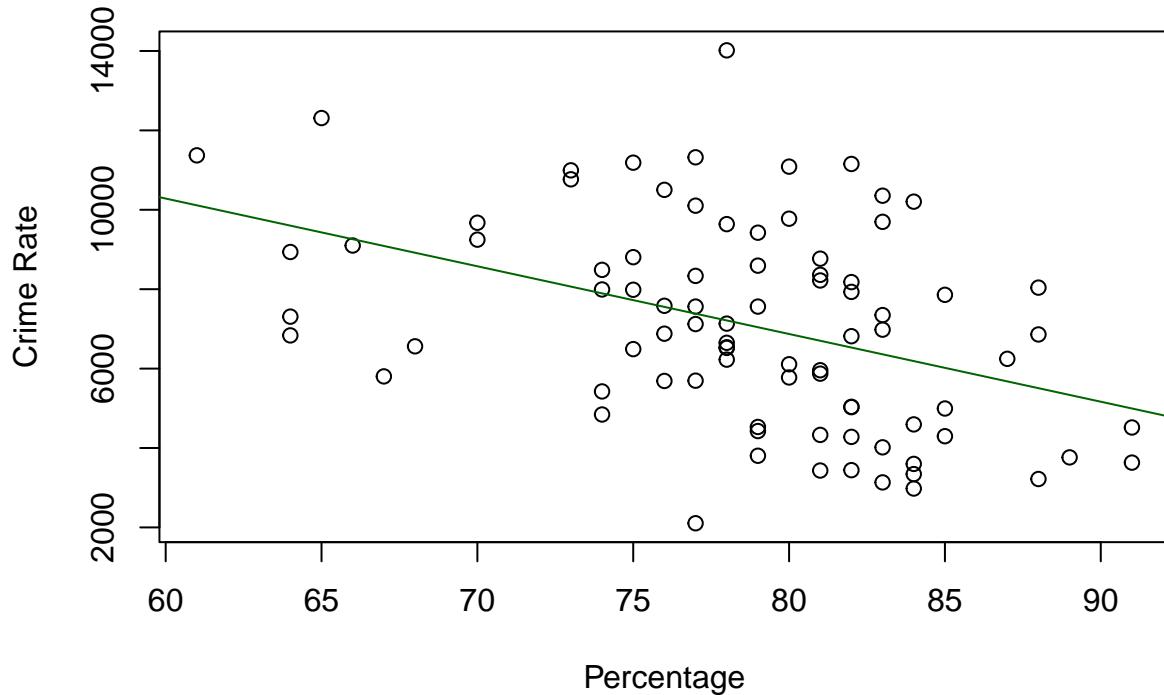
## Problem 2

a

```
my_data2 <- read.delim("CH01PR28.txt", sep = "", header=FALSE)
colnames(my_data2) <- c("Y", "X")
my_ordereddata2 <- my_data2[order(my_data2$X, decreasing = FALSE), ]
reg2 <- lm(Y ~ X, data=my_data2)
reg2

##
## Call:
## lm(formula = Y ~ X, data = my_data2)
##
## Coefficients:
## (Intercept)          X
##       20517.6        -170.6

plot(Y ~ X, data=my_data2, xlab="Percentage", ylab="Crime Rate")
abline(reg2, col = "darkgreen")
```



While the line does seem to capture how the crime rate lowers as percentage of people with at least high-school diploma increases, it doesn't capture how packed data points are between 75-85 percentage, but very scarce and spread outside of that range.

b

1

The average difference in crime rate between counties that have a high school graduation rate separated by 1%

```
count2=0
difference2=0
for(i in 1:nrow(my_ordereddata2)) { # for-loop over rows
  for(j in i:nrow(my_ordereddata2)){
    if (abs(my_ordereddata2[i,2]-my_ordereddata2[j,2])==1){
      count2=count2+1
      difference2= difference2+((my_ordereddata2[i,1]-my_ordereddata2[j,1]))
    }
  }
}
mean2=difference2/count2
mean2
```

```
## [1] 250.0081
```

The difference in crime rate between counties whose high school graduation differs by 1% is 250.0081

**2**

mean crime rate in counties when high school graduation is 80%

```
count80s=0
sum80s=0
for (i in 1:nrow(my_ordereddata2)){
  if (my_ordereddata2[i,2]==80){
    count80s=count80s+1
    sum80s=sum80s+my_ordereddata2[i,1]
  }
}
sum80s/count80s
```

```
## [1] 8187.25
```

Mean crime rate when high school graduation is 80% is 8187.25

**3**

```
reg2$residuals[10] #The residual of the 10th data point
```

```
##      10
## 1401.566
```

**4**

```
var(my_data2$Y) #Variance of the crime rate data
```

```
## [1] 6611278
```

## Problem 3

**a**

## Problem 4

**a**

**b**

Problem 3 (1.39) a)

$$x=5, x=10, x=15$$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

assuming  $\varepsilon_i$  iid

normal dist.

$$N(0, \sigma^2)$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

First 6 points:  $(5, Y_{15}), (5, Y_{25}), (10, Y_{10}), (10, Y_{210}), (15, Y_{115}), (15, Y_{215})$

$$\bar{x} = \frac{60}{6} = 10 \quad \bar{y} = \frac{Y_{15} + Y_{25} + Y_{10} + Y_{210} + Y_{115} + Y_{215}}{6}$$

Solving for  $\hat{\beta}_1$  for 6 points:

$$(5-10)(Y_{15}-\bar{y}) + (5-10)(Y_{25}-\bar{y}) + (10-\bar{x})(Y_{10}-\bar{y}) + (10-\bar{x})(Y_{210}-\bar{y}) \\ + (15-10)(Y_{115}-\bar{y}) + (15-10)(Y_{215}-\bar{y})$$

$$S_{xy} = -5(Y_{15}-\bar{y}) - 5(Y_{25}-\bar{y}) + 5(Y_{115}-\bar{y}) + 5(Y_{215}-\bar{y})$$

$$S_{xx} = (5-10)^2 + (5-10)^2 + (10-10)^2 + (10-10)^2 + (15-10)^2 + (15-10)^2 \\ = 100$$

$$-5(Y_{15} - 5)$$

$$-5Y_{15} - 5Y_{25} + 5Y_{115} + 5Y_{215} \\ - 5(Y_{15} + Y_{25}) + 5(Y_{115} + Y_{215}) = S_{xy} \\ 100 - 20 = S_{xy}$$

$$-(Y_{15} + Y_{25}) + (Y_{115} + Y_{215}) \leftarrow \hat{\beta}_1 \text{ for 6 points}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

FOR 3 points  $\rightarrow$

Figure 1: Written solution to problem 3a- part 1

3 Points:  $(5, \bar{Y}_1), (10, \bar{Y}_2), (15, \bar{Y}_3)$

$$\hat{B}_{13} = \frac{s_{xy}}{s_{xx}} \quad \bar{x} = \frac{5+10+15}{3} = 10 \quad \bar{Y} = \frac{\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3}{3}$$

$$\bar{Y}_1 = \frac{Y_{15} + Y_{25}}{2} \quad \bar{Y}_2 = \frac{Y_{10} + Y_{20}}{2} \quad \bar{Y}_3 = \frac{Y_{115} + Y_{215}}{2}$$

$$S_{xy} = (5-10)(\bar{X}, -\bar{Y}) + (10-10)(\bar{Y}_2 - \bar{Y}) + (15-10)(\bar{Y}_3 - \bar{Y}) \\ = -5(\bar{Y}_1 - \bar{Y}) + 5(\bar{Y}_3 - \bar{Y})$$

$$S_{xx} = (5-10)^2 + (10-10)^2 + (15-10)^2 = 50$$

$$\hat{B}_{13} = \frac{-5\bar{Y}_1 + 5\bar{Y}_3}{50} = \frac{-\bar{Y}_1 + \bar{Y}_3}{10} \quad \leftarrow$$

Since  $\bar{Y}_1 = \frac{Y_{15} + Y_{25}}{2}$  and  $\bar{Y}_3 = \frac{Y_{115} + Y_{215}}{2}$  and Subbing here

$$= \frac{-(Y_{15} + Y_{25}) + (Y_{115} + Y_{215})}{20} = \hat{B}_{13}$$

$$\begin{aligned} \hat{B}_{03} &= \bar{Y} - \hat{B}_{13}\bar{X} \\ \hat{B}_0 &= \bar{Y} - \hat{B}_1\bar{X} \end{aligned} \quad \left( \hat{B}_{13} = \hat{B}_1 \right)$$

Since slopes are the same  
and the  $X$  and  $Y$  also, so  
must the intercepts

3b) Yes  $\sigma^2$

can be obtained using  $s^2 = \frac{s_{xx}}{n-1}$  which  
estimates sample variance if you have all  $X$  data.

Figure 2: Written solution to the rest of 3a and 3b

Problem 4 Assume model  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

a) Likelihood function  $\rightarrow \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right)$

$$\boxed{\left(\frac{1}{\sqrt{2\pi}\sqrt{16}}\right)^6 \exp\left(-\frac{1}{32} \sum_{i=1}^6 (y_i - 0 - \beta_1 x_i)^2\right)}$$

Figure 3: Written solution to 4a

```

my_data3 <- read.delim("CH01PR42.txt", sep = "", header=FALSE)
colnames(my_data3) <- c("Y", "X")
slopes <- c(17, 18, 19)
sum17=0
for (i in 1:nrow(my_data3)){
  sum17=sum17+((my_data3[i,1]-(slopes[1]*my_data3[i,2]))^2)
}

((1/(sqrt(16)*sqrt(2*pi)))^6)*exp(sum17*(-1/32))

## [1] 9.45133e-30

sum18=0
for (i in 1:nrow(my_data3)){
  sum18=sum18+((my_data3[i,1]-(slopes[2]*my_data3[i,2]))^2)
}

((1/(sqrt(16)*sqrt(2*pi)))^6)*exp(sum18*(-1/32))

## [1] 2.649043e-07

sum19=0
for (i in 1:nrow(my_data3)){
  sum19=sum19+((my_data3[i,1]-(slopes[3]*my_data3[i,2]))^2)
}

((1/(sqrt(16)*sqrt(2*pi)))^6)*(exp(1)^(-1/32))

## [1] 9.539569e-07

```

The largest value is when  $B1=18$  with the value  $2.649e-7$

c

```

xy=0
x_square=0
for (i in 1:nrow(my_data3)){
  xy=xy+(my_data3[i,2]*my_data3[i,1])
}
for (i in my_data3[,2]){
  x_square=x_square+i^2
}
xy/x_square

```

```
## [1] 17.9285
```

The value is very close to 18, which the likelihood function in part b also was the highest for

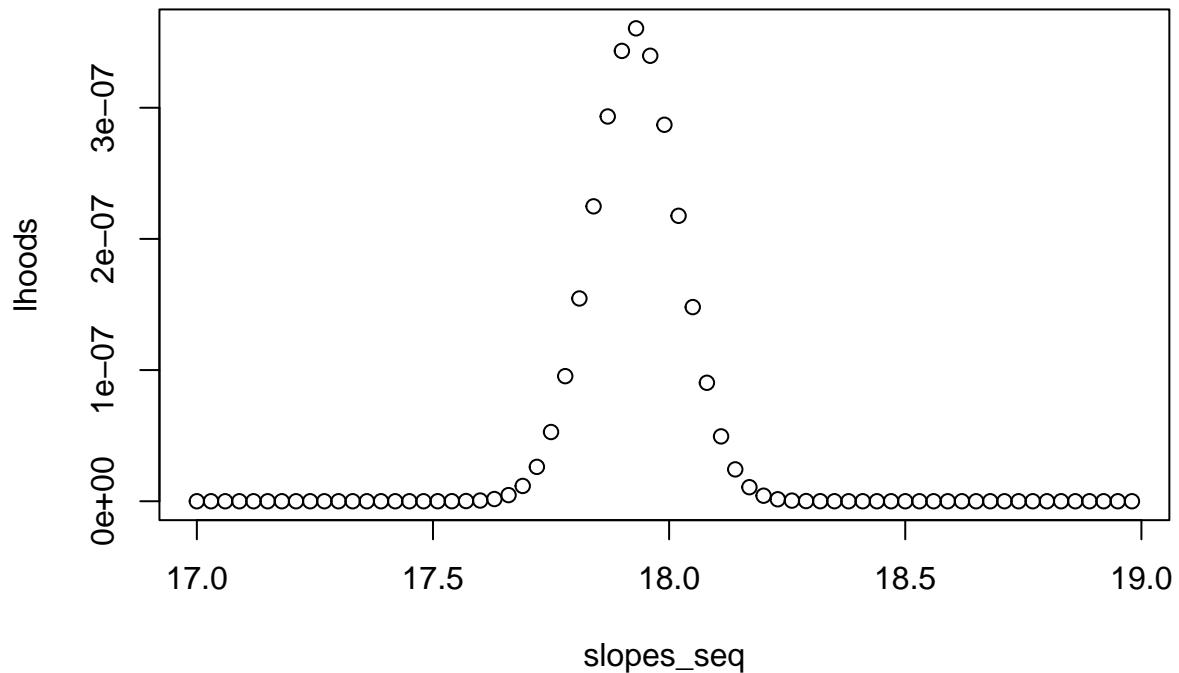
d

```

slopes_seq<-seq(17,19,0.03)
sums<-c(rep(0,length(slopes_seq)))
lhoods<-c(rep(0,length(slopes_seq)))
for (i in 1:length(sums)){
  for (j in 1:nrow(my_data3)){
    sums[i]=sums[i]+((my_data3[j,1]-(slopes_seq[i]*my_data3[j,2]))^2)
  }
}
for(i in 1:length(sums)){
  lhoods[i]<-((1/(sqrt(16)*sqrt(2*pi)))^6)*exp(sums[i]*(-1/32))
}

plot(x=slopes_seq,y=lhoods)

```



```

max(lhoods)

## [1] 3.605343e-07

which.max(lhoods)

## [1] 32

slopes_seq[32]

## [1] 17.93

```

the likelihood plot shows that indeed the point from part c is where the likelihood was the greatest

## Problem 5

## Problem 6

Proof:

$$\sum e_i x_i = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) x_i$$

$$\sum (x_i y_i - x_i (\hat{\beta}_0 + \hat{\beta}_1 x_i))$$

$$\sum (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2)$$

$$\rightarrow \sum x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2$$

$$\sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 \leftarrow \text{Norway equation}$$

online source

$$\hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0$$

QED

Figure 4: Problem 5- Written solution to prove sum of product of  $x_i$  and  $e_i$  equals 0

Problem 5

Proof  $\sum_{i=1}^n e_i = 0$

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$\hat{\beta}_0 =$$

$$\sum_{i=1}^n y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$\sum_{i=1}^n y_i - (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i)$$

$$\sum_{i=1}^n y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i$$

$$\sum_{i=1}^n y_i - \bar{y} + \hat{\beta}_1 (\bar{x} - x_i) = \sum y_i \cancel{- \bar{y}} + \sum \hat{\beta}_1 (\bar{x} - x_i)$$

$$+ \hat{\beta}_1 \sum (\bar{x} - x_i)$$

$$= 0 + \hat{\beta}_1 (0)$$

= 0 QED

Figure 5: Problem 5- Written solution to prove sum of  $e_i$  equals 0

Problem 6

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} \quad \text{prove } \hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2})$$

$$\begin{aligned} ① \quad E(\hat{\beta}_1) &= E\left(\frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}\right) = \frac{1}{\sum (x_i - \bar{x})^2} E\left(\sum (x_i - \bar{x}) y_i\right) \\ &= \frac{1}{\sum (x_i - \bar{x})^2} \sum (x_i - \bar{x}) E(y_i) \quad \rightarrow y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \\ &= \frac{1}{\sum (x_i - \bar{x})^2} \sum (x_i - \bar{x}) E(\beta_0 + \beta_1 x_i + \varepsilon_i) \\ &= \frac{1}{\sum (x_i - \bar{x})^2} \sum (x_i - \bar{x}) (\beta_0 + \beta_1 x_i + E(\varepsilon_i)) \\ &= \frac{1}{\sum (x_i - \bar{x})^2} \left( \sum (x_i - \bar{x}) \beta_0 + \sum (x_i - \bar{x}) \beta_1 x_i \right) \end{aligned}$$

Figure 6: Problem 6 - Written solution to prove  $\hat{\beta}_1$  has mean  $\beta_1$  and Variance:  $\sigma^2 / (S_{xx})$

$$= \frac{1}{\sum(x_i - \bar{x})^2} (\beta_0 \sum(x_i - \bar{x})^0 + \beta_1 \sum(x_i - \bar{x})(x_i))$$

$$= \frac{\beta_1}{\sum(x_i - \bar{x})^2} \sum(x_i - \bar{x})x_i$$

$$= \underline{\beta_1} \quad \text{assume } x_i \text{ constant}$$

used problem 5 proof that  
 $E(e_i) = 0$

$$\textcircled{2} \quad \text{var}(\hat{\beta}_1) = \text{var}\left(\frac{\sum(x_i - \bar{x})v_i}{\sum(x_i - \bar{x})^2}\right)$$

$$= \left(\frac{1}{\sum(x_i - \bar{x})^2}\right)^2 \text{var}(\sum(x_i - \bar{x})v_i)$$

$$= \left(\frac{1}{\sum(x_i - \bar{x})^2}\right)^2 \text{var}(\sum(x_i - \bar{x})^0(\beta_0 + \beta_1 x_i) + \sum(x_i - \bar{x})e_i)$$

$$= \left(\frac{1}{\sum(x_i - \bar{x})^2}\right)^2 \sum \text{var}((x_i - \bar{x})e_i)$$

$$= \left(\frac{1}{\sum(x_i - \bar{x})^2}\right)^2 \sum (x_i - \bar{x})^2 \text{var}(e_i) \quad \begin{matrix} \text{assure} \\ \text{Variance of} \\ \text{error is } \sigma^2 \end{matrix}$$

$$= \left(\frac{1}{\sum(x_i - \bar{x})^2}\right)^2 \sum (x_i - \bar{x})^2 \sigma^2$$

$$= \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \cancel{\sum(x_i - \bar{x})^2} = \boxed{\frac{\sigma^2}{\sum(x_i - \bar{x})^2}} \checkmark$$

assuming  $x_i$  is  
constant

Figure 7: Problem 6 - Continuation of written solution to prove  $\hat{\beta}_1$  has mean  $\beta_1$  and Variance:  $\sigma^2 / (\sum(x_i - \bar{x})^2)$