STAT 151-A HW 1

Cao jilin

1.

a) $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i-\bar{X})y_i}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$ $\qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{X}$

b) $\hat{a}_1 = \dfrac{\sum_{i=1}^{n}(y_i-\bar{y})x_i}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$ $\qquad \hat{a}_0 = \bar{X} - \hat{a}_1\bar{y}$

c) No $\quad \dfrac{1}{\hat{\beta}_1} = \dfrac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})y_i} \neq \hat{a}_1$

d) stored as a picture on the next page.

2. (a) $\hat{Y} = \underbrace{-0.31886}_{\hat{\beta}_1}X + \underbrace{32.14271}_{\hat{\beta}_0}$ $\qquad SD(\hat{\beta}_1) = 0.03484$

$SD(\hat{\beta}_0) = 0.99758$

(b) our model shows a negative relationship, so the assumption was wrong. It turns out the lunch program has a negative effect on student performance.

(2)

3\.

a) Yes. More money spent means that a student can get more resources, thus increasing grade. In R we get $\beta_0 = 1.34 \times 10^1$
$\beta_1 = 2.456 \times 10^{-3}$ which also suggests a positive relationship

b) percentage change in math10

model: $math10 = \beta_0 + \beta_1 \log(expend) + e$

$\Delta math10 = \beta_1 \Delta \log(expend)$, we observe that change in log(expend)

$\Delta \log(expend) \approx \frac{1}{100} (\% \Delta expend)$, so if $\% \Delta expend = 10$

$\Delta math10 = \frac{1}{10} \beta_1$

c) $\widehat{Math10} = \underset{\beta_0}{-69.341} + \underset{\beta_1}{11.164} \log(expend)$

$SD(\beta_0) = 26.53$

$SD(\beta_1) = 3.169$

d) $\Delta expend = 10\%$ $\Delta math10 = 11.164 \times 10\% = 1.1164$, so $\widehat{math10}$ increases by 1.1164

e) inspect the data set, we found $\max(math10) = 66.7$, $\max(expend) = 7419$

$66.7 < 100$ and if we plug 7419 into $\widehat{math10} = -69.34 + 11.164 \times \log(7419) = 30.15 < 100$

the largest fitted value

So we don't need to worry that.

4. a)

$$\hat{\beta}_0+\hat{\beta}_1 a=\left(\overbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}^{\bar{y}}-\overbrace{\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}^{\hat{\beta}_1}\overbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}^{\bar{x}}\right)+\overbrace{\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}^{\hat{\beta}_1}a$$

$$=\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}(a-\bar{x})+\frac{1}{n}\sum_{i=1}^{n} y_i$$

$$=\sum_{i=1}^{n} y_i\left\{\frac{1}{n}+\frac{(a-\bar{x})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right\}$$

b) we know $\mathrm{var}_x(\hat{\beta}_1)=\dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ and $\bar{y}=\hat{\beta}_0+\hat{\beta}_1\bar{x}\Rightarrow\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$

$\mathrm{var}_x(\bar{y}-\hat{\beta}_1\bar{x}+\hat{\beta}_1 a)=\mathrm{Var}_x(\bar{y}+\hat{\beta}_1(a-\bar{x}))=\mathrm{Var}_x(\bar{y})+(a-\bar{x})^2\mathrm{var}_x(\hat{\beta}_1)+2(a-\bar{x})\mathrm{cov}(\bar{y},\hat{\beta}_1)$

we have $\mathrm{cov}(\hat{\beta}_0,\hat{\beta}_1)=\dfrac{-\sigma^2 n\bar{x}}{n^2\,\mathrm{Var}(x)}=\mathrm{cov}(\bar{y}-\hat{\beta}_1\bar{x},\hat{\beta}_1)=\mathrm{cov}(\bar{y},\hat{\beta}_1)-\bar{x}\,\mathrm{cov}(\hat{\beta}_1,\hat{\beta}_1)\Rightarrow$

$\mathrm{cov}(\bar{y},\hat{\beta}_1)=0$, $\boxed{\text{so } \mathrm{Var}_x(\hat{\beta}_0+\hat{\beta}_1 a)=\dfrac{\sigma^2}{n}+\dfrac{\sigma^2(a-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

c) given the formula in part (b), we observe that $\dfrac{\sigma^2(a-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\geq 0$ and when $a=\bar{x}$ it equals to 0, which minimizes the $\mathrm{Var}(\hat{\beta}_0+\hat{\beta}_1 a\mid x_1,\cdots,x_n)$ to $\dfrac{\sigma^2}{n}$

6. a) given $\mathbb{E}(y|x) = \beta_1 x$ and assume $E(e_i|x_i) = 0$

We can write $y_i = \beta_1 x_i + e_i$

LSE of β_1 is value of b_1 that minimize $Q(b_1) = \sum_{i=1}^{n} (y_i - b_1 x_i)^2$

$$\frac{\partial Q}{\partial b_1} = -2\sum_{i=1}^{n} x_i(y_i - b_1 x_i) = 0 \Rightarrow b_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

b) $$\mathbb{E}(\hat{\beta}_1|X) = \mathbb{E}\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \middle| X\right) = \frac{\sum_{i=1}^{n} x_i \mathbb{E}(y_i|X)}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i(\beta_1 x_i)}{\sum_{i=1}^{n} x_i^2} = \beta_1$$

and by law of iterated expectations $\mathbb{E}(\hat{\beta}_1) = \mathbb{E}(\mathbb{E}(\hat{\beta}_1|X)) = \mathbb{E}(\beta_1) = \beta_1$

Therefore, $\hat{\beta}_1$ is unbiased

c) $$\mathrm{Var}_X(\hat{\beta}_1) = \mathrm{Var}_X\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right) = \frac{\sum_{i=1}^{n} x_i \mathrm{Var}_X(y_i)}{\sum_{i=1}^{n} x_i^2} = \frac{\sigma^2 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

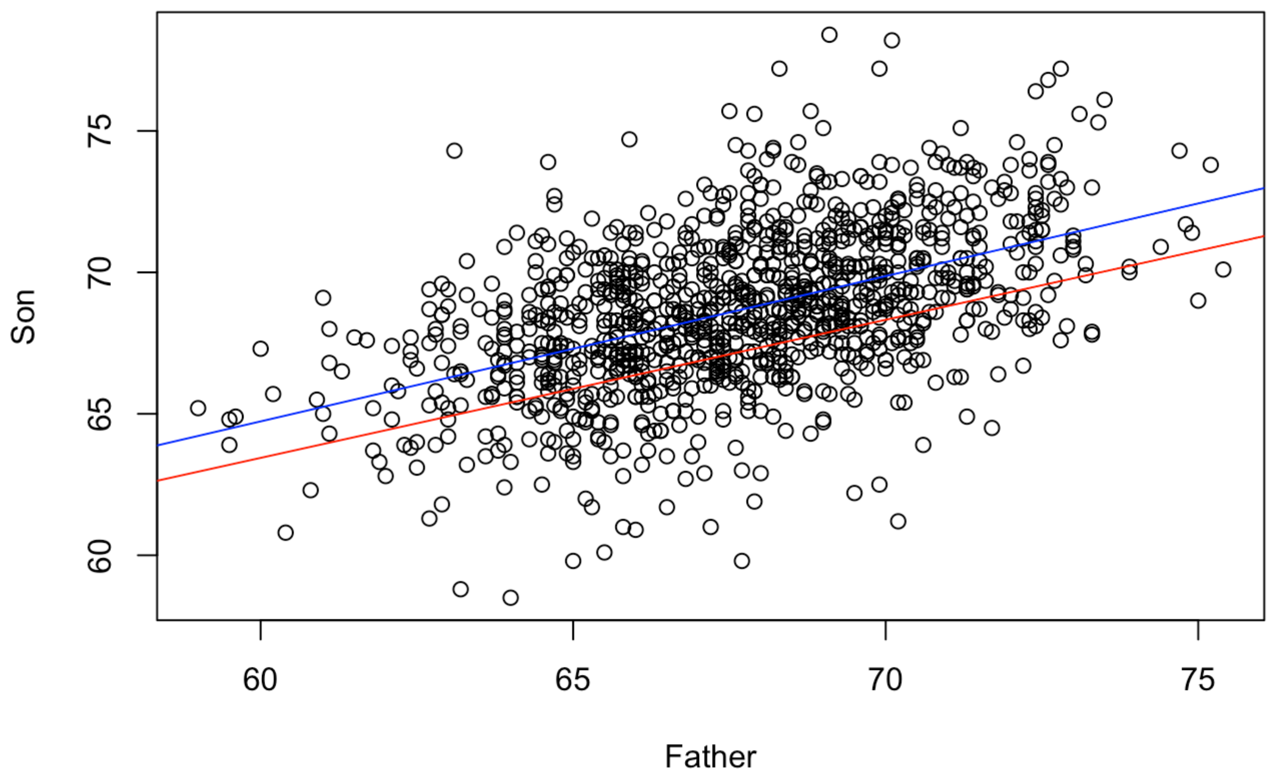
HW01

caojilin

9/5/2018

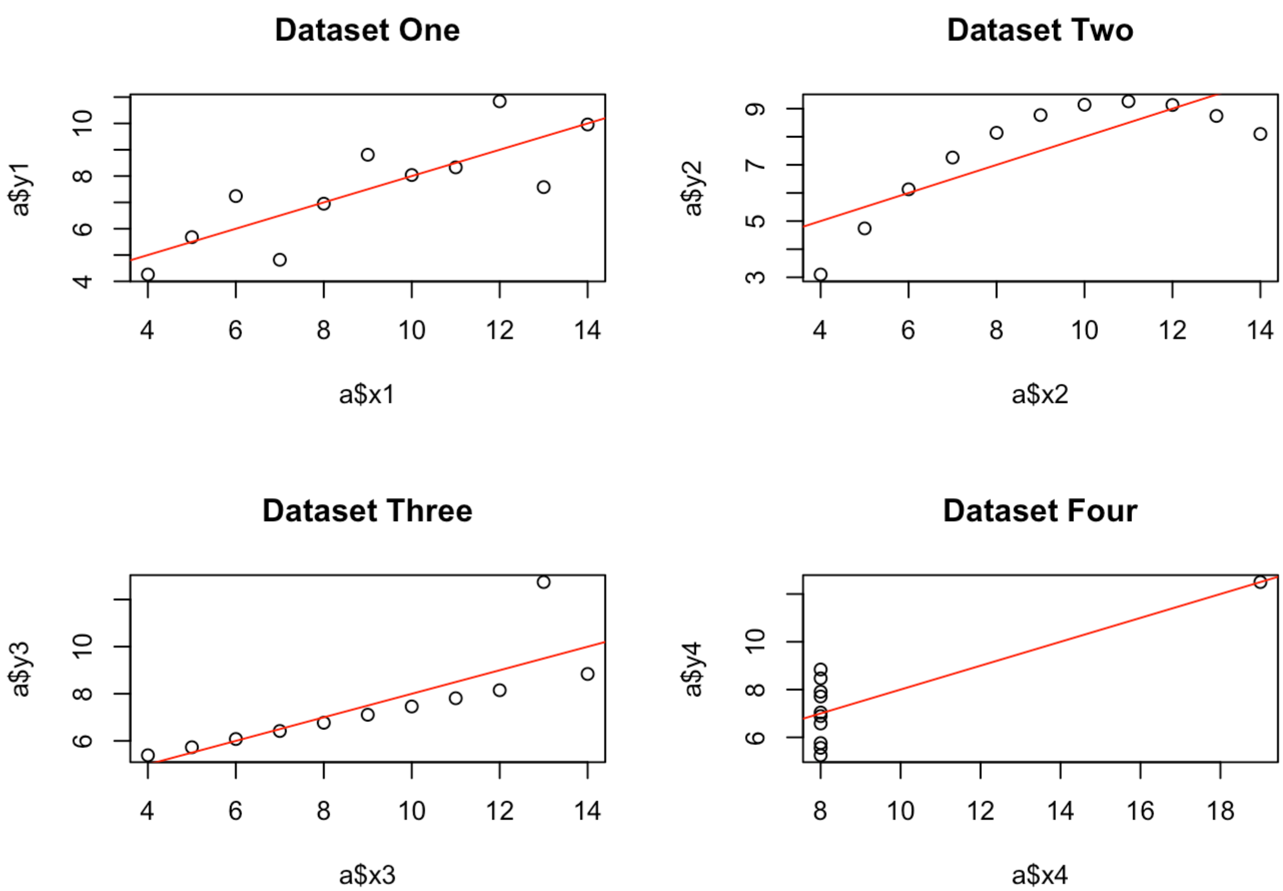
Problem 1d)

```
dat = read.table("PearsonHeightData.txt", header = T)
lmod1 = lm(Son ~ Father, data = dat)
lmod2 = lm(Father ~ Son, data = dat)
plot(Son ~ Father, data = dat)
abline(lmod1, col = "blue")
abline(lmod2, col = "red")
```



Problem 5

```r
library(datasets)
a <- anscombe
par(mfrow=c(2,2))


lm1 = lm(a$y1 ~ a$x1)

lm2 = lm(a$y2 ~ a$x2)

lm3 = lm(a$y3 ~ a$x3)

lm4 = lm(a$y4 ~ a$x4)

plot(a$x1,a$y1, main=paste("Dataset One"))
abline(lm1,col="red")
plot(a$x2,a$y2, main=paste("Dataset Two"))
abline(lm2,col="red")
plot(a$x3,a$y3, main=paste("Dataset Three"))
abline(lm3,col="red")
plot(a$x4,a$y4, main=paste("Dataset Four"))
abline(lm4,col="red")
```



dataset 1, and 3 make sense. Although dataset 3 has a outlier, it doesn't matter.
dataset 2 looks like a parabola, linear model is not accurate as dataset gets larger, but it's ok for small dataset like this.
dataset 4 doesn't make sense because most of the data are clustered on the line x = 8

predictions are:

```
lm1$coefficients[1] + 10 * lm1$coefficients[2]
```

```
## (Intercept)
##       8.001
```

```
lm2$coefficients[1] + 10 * lm2$coefficients[2]
```

```
## (Intercept)
##    8.000909
```

```
lm3$coefficients[1] + 10 * lm3$coefficients[2]
```

```
## (Intercept)
##    7.999727
```

```
lm4$coefficients[1] + 10 * lm4$coefficients[2]
```

```
## (Intercept)
##    8.000818
```

again, predictions for dataset 1,2,3 are close to real data values. So they make sense. but dataset 4 doesn't make sense.

Problem 7

a)

we assume simple linear regression model for

$$(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$$

to be

$$y_i = \beta_0 + \beta_1 x_1 + e_i$$

where

$$\mathbb{E}(e_i \mid X_i) = 0$$

That means we have to make sure

$$\mathbb{E}(e_i \mid X_i) = 0$$

in order to assume a linear model.
To verify that: for

$$x_i <= 65$$

$$y_i$$

can be wirrten as

$$y_i = N(\beta_0 + \beta_1 x_i, 25) + N(0, 25)$$

where

$$e_i = N(0, 25)$$

this satisfies that

$$\mathbb{E}(e_i \mid X_i) = 0$$

similarly, for

$$65 < x_i <= 70$$

,

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

where

$$e_i = 10T_i$$

whose mean is 0
for

$$x_i > 70$$

,

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

where

$$e_i = U_i$$

, and

$$\mathbb{E}(U_i) = 0$$

Therefore, the condition

$$\mathbb{E}(e_i \mid X_i) = 0$$

for simple linear regression model is satisfied. Since we have proved in the class and in the homework that lease square estimators are unbiased. we can conclude that

$$\hat{\beta}_0$$

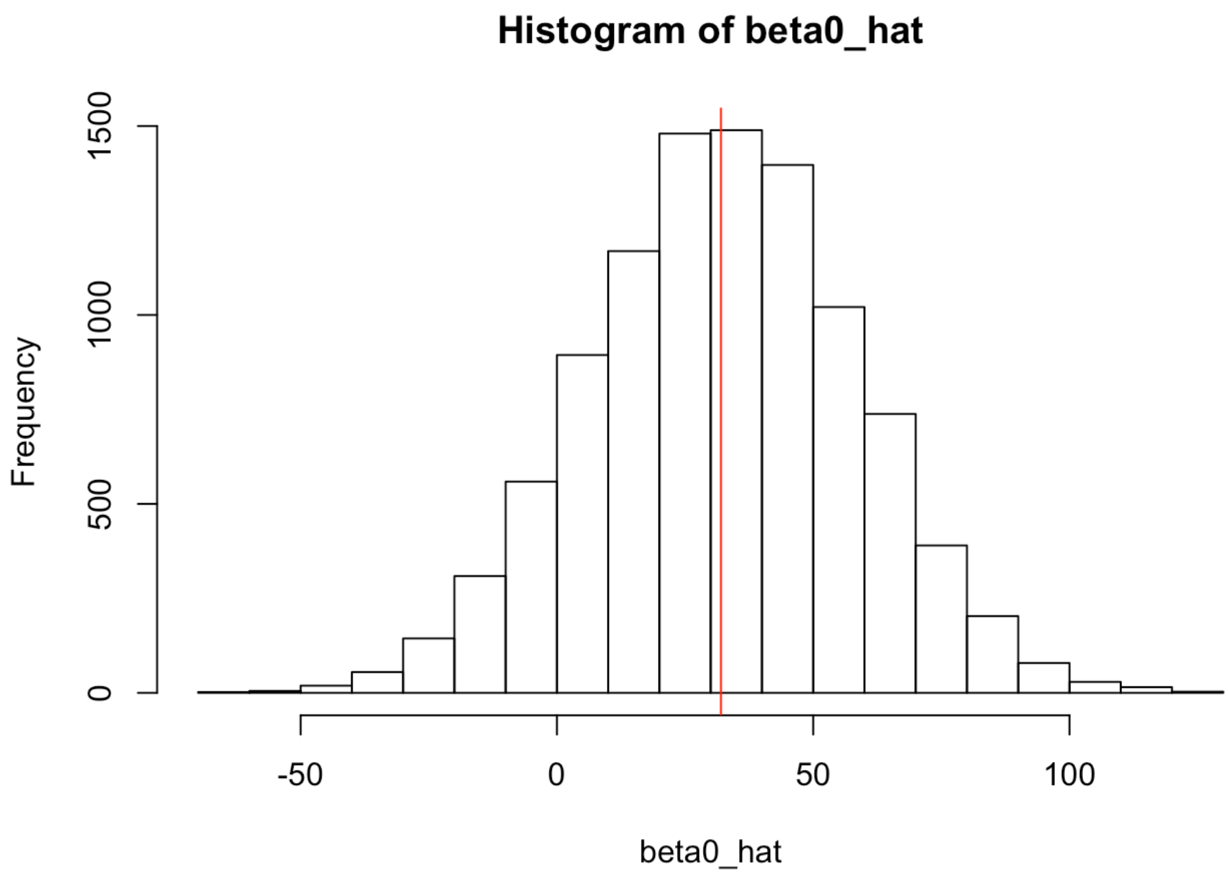
and

$$\hat{\beta}_1$$

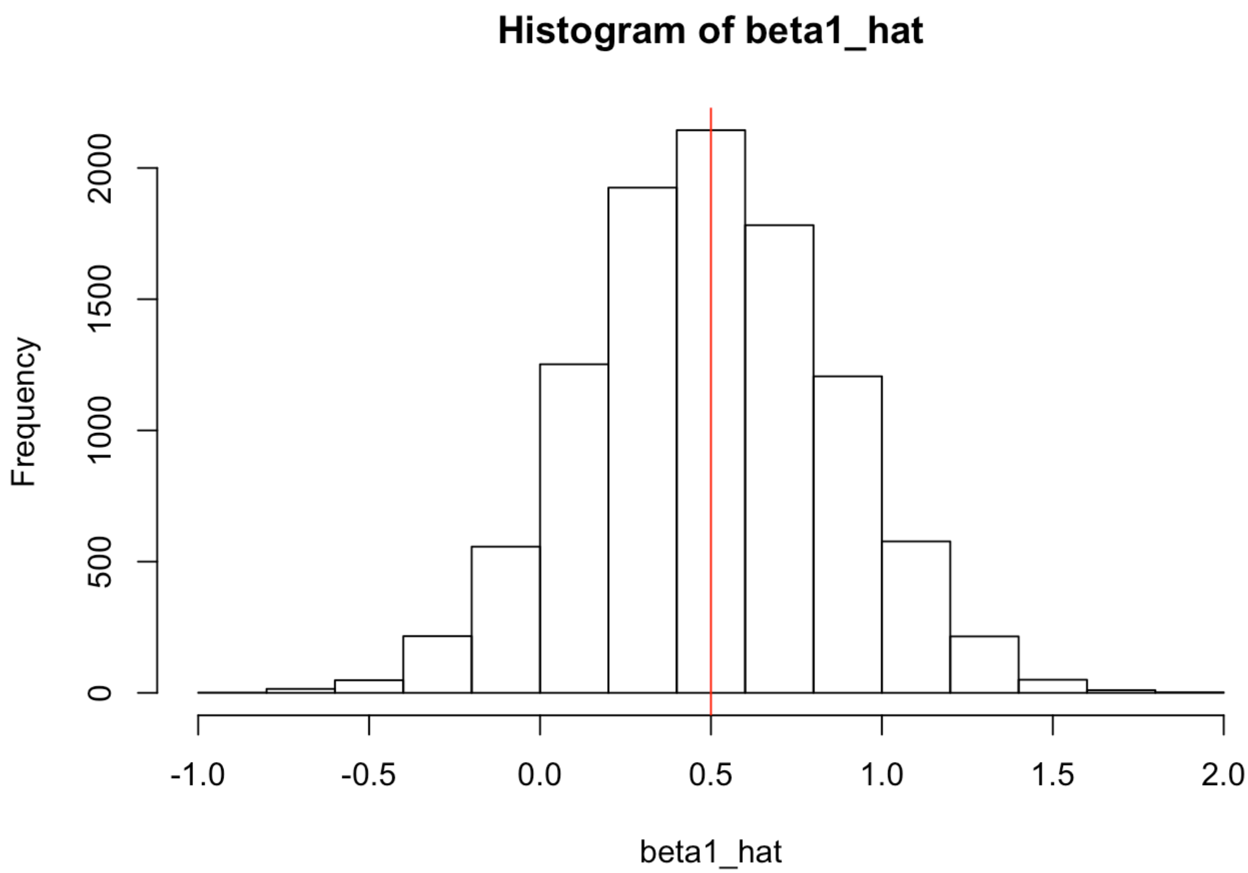
are unbiased.

b)

```
M = 1:10000

samp = function(){
  x = seq(59, 76, length.out = 100)
  x1=x[x<=65]
  y1 = rnorm(length(x1),mean = 32+0.5*x1,sd = 25)
  x2=x[x>65 & x<=70]
  y2 = 32+0.5*x2+10*rt(n = length(x2),df = 3)
  x3=x[x>70]
  y3=32+0.5*x3+runif(length(x3),min = -8,max = 8)
  y = c(y1,y2,y3)
  return(list(x,y))
}
beta0_hat = c()
beta1_hat = c()
sd_beta0 = c()
sd_beta1 = c()

for (i in M) {
  sample = samp()
  x = sample[[1]]
  y = sample[[2]]
  lmod = lm(y ~ x)
  beta0_hat = c(beta0_hat, lmod$coefficients[1])
  beta1_hat = c(beta1_hat, lmod$coefficients[2])
  sd_beta0 = c(sd_beta0,summary(lmod)$coefficients[3])
  sd_beta1 = c(sd_beta1,summary(lmod)$coefficients[4])

}

hist(beta0_hat)
abline(v=32,col="red")
```




```
hist(beta1_hat)
abline(v=0.5,col="red")
```




```
#bias of beta 0
sum(beta0_hat -32)/10000
```

```
## [1] 0.2965467
```

```
#bias of beta 1
sum(beta1_hat -0.5)/10000
```

```
## [1] -0.004216398
```

From histogram, the estimates of the bias are close to 0 and we found that the center for both graph are close to real

$$\beta_0 = 32$$

and

$$\beta_1 = 0.5$$

This verifies unbiasedness

c)

homoskedasticity means that same variance of the errors,

$$Var(e_i \mid X) = \sigma^2$$

for each i. However, in this problem, we see that

$$Var(e_i \mid X) = \sigma^2$$

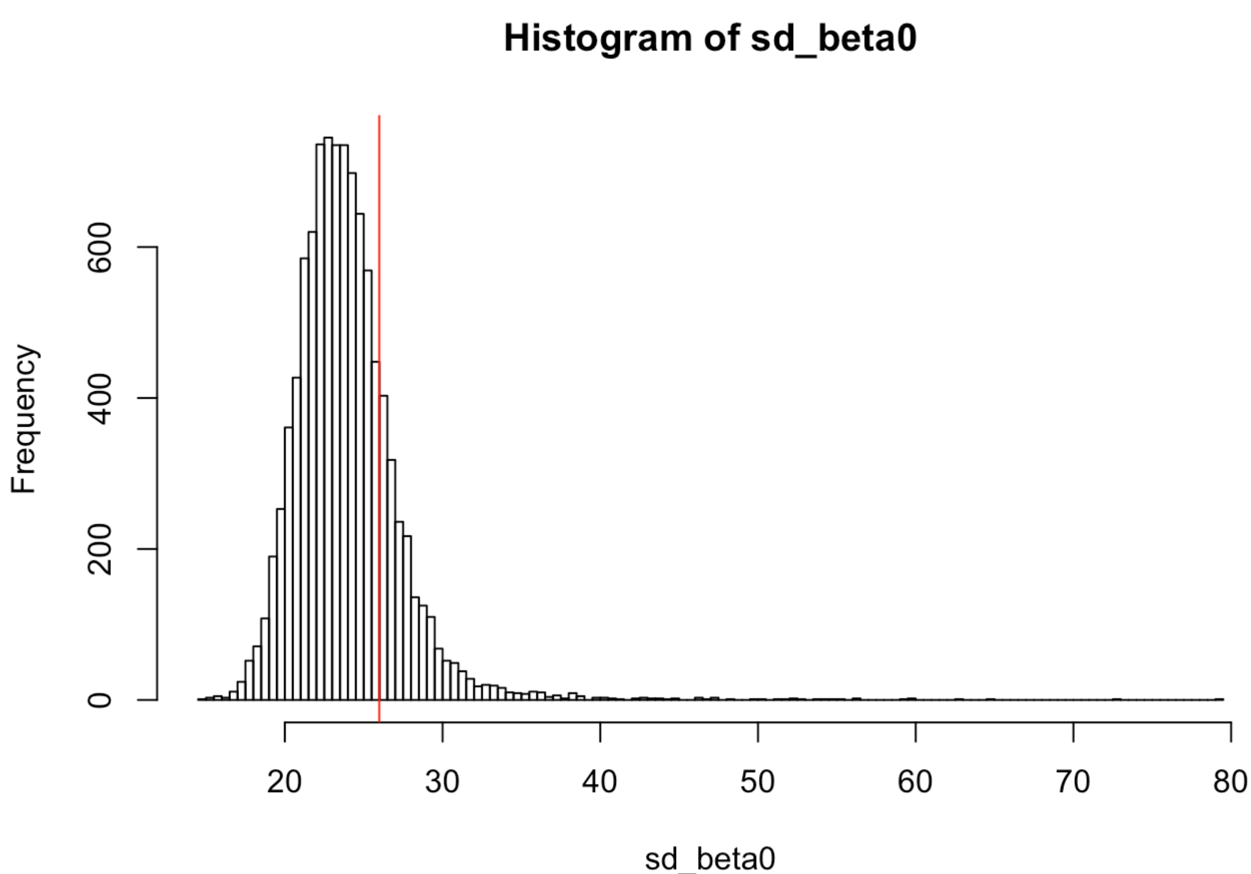
are not same for each i. Thus, homoskedasticity is not valid here.

d)

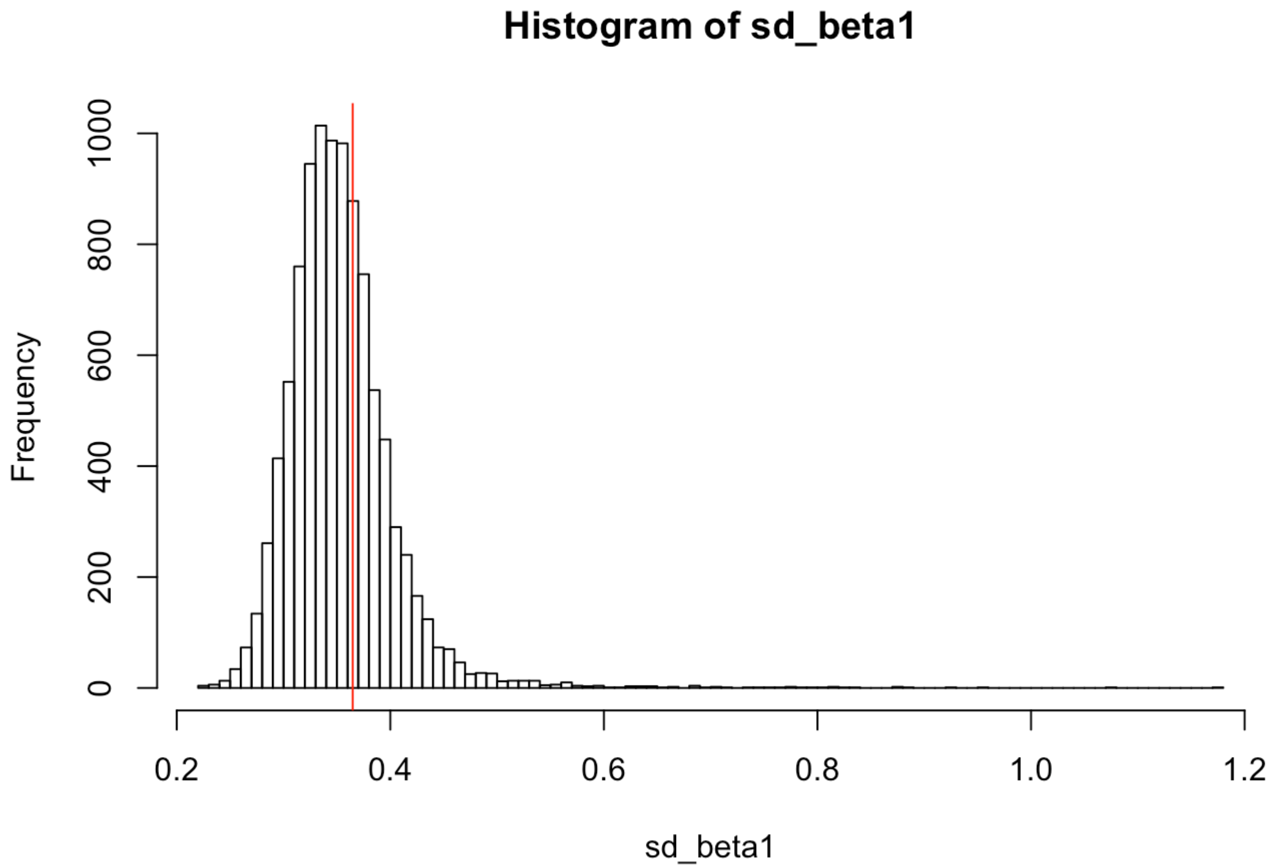
we would like to check unbiasedness of

$$\mathbb{E}(\widehat{Var}(\hat{\beta})) = Var(\hat{\beta})$$

```
sd0 = sd(beta0_hat)
sd1 = sd(beta1_hat)
hist(sd_beta0,breaks = 100)
abline(v=sd0,col="red")
```




```
hist(sd_beta1,breaks=100)
abline(v=sd1,col="red")
```



As we see from histograms, the centers are not close to sd0 and sd1, meaning that

$$\mathbb{E}(\widehat{Var(\hat{\beta})}) \neq Var(\hat{\beta})$$

this verifies that homoskedasticity is not valid