stat151A HW1

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

(b) $E(y_i|x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$E(x_i|y_i) = \frac{1}{\beta_1}y_i - \frac{\beta_0}{\beta_1} = \alpha_1 y_i + \alpha_0$

thus. $\alpha_1 = \frac{1}{\beta_1} = \frac{\Sigma(x_i-\bar{x})^2}{\Sigma(x_i-\bar{x})\cdot y_i}$

(c) yes, As can be seen in (b) $\hat{\alpha}_1 = \frac{1}{\hat{\beta}_1}$

(d). See R code.

3. (b). $math10 = \beta_0 + \beta_1 \log_{10}(expand) + \varepsilon$

if expand increases by 10%.

$math10 = \beta_0 + \beta_1 \log_{10}(1.1\, expand) + \varepsilon$

$= math10 + 0.1\beta_1$

4. (a) $\hat{\beta_0} + \hat{\beta_1} a = \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} a + \bar{y} - \hat{\beta_1}\bar{x}$

$= \bar{y} - \hat{\beta_1}\bar{x} + \hat{\beta_1} a$

$= \bar{y} - \hat{\beta_1}(a-\bar{x}) = \frac{\sum y_i}{n} + \frac{\sum (a-\bar{x})(x_i-\bar{x})y_i}{\sum (x_i-\bar{x})^2}$

(b) $var(\hat{\beta_0} + \hat{\beta_1} a) = var \sum_{i=1}^{n} y_i \left(\frac{1}{n} + \frac{(a-\bar{x})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right)$

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

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

$= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} + \frac{2(a-\bar{x})(x_i-\bar{x})}{n\sum_{i=1}^{n}(x_i-\bar{x})^2} + \frac{(a-\bar{x})(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^4} \right)$

$= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} + 0 + \frac{(a-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right)$

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

(c). let $t = var(\hat{\beta_0} + \hat{\beta_1} a)$

when $a = \bar{x}$, $var(\hat{\beta_0} + \hat{\beta_1} a)$ can acheive its smallest value.

since $\frac{\sigma^2(a-\bar{x})}{\sum (x_i-\bar{x})^2} = 0$.

and $var(\hat{\beta_0} + \hat{\beta_1} a) = \frac{\sigma^2}{n}$

6. (a) $\cancel{\hat{\beta_1}} = y = \hat{\beta}_1 x$

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

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i) \cdot x_i = 0.$$

$$\Rightarrow \sum (y_i x_i - \hat{\beta}_1 x_i^2) = 0.$$

$$\sum y_i x_i = \sum \hat{\beta}_1 x_i^2$$

$$\sum y_i x_i = \hat{\beta}_1 \sum x_i^2$$

$$\hat{\beta}_1 = \frac{\sum y_i x_i}{\sum x_i^2}$$

(b). To show $\hat{\beta}_1$ is unbiased. $E(\hat{\beta}_1 | x) = \beta_1$.

$$E(\hat{\beta}_1 | x) = E\left(\frac{\sum y_i x_i}{\sum x_i^2}\right).$$

$$= \cancel{\sum x_i E} \; \frac{\sum x_i E(y_i)}{\sum x_i^2} = \frac{\sum x_i E(\beta_1 x_i)}{\sum x_i^2} = \frac{\beta_1 \sum x_i x_i}{\sum x_i^2}$$

$$\cancel{= \frac{\sum x_i E(\hat{\beta}_1 x_i)}{\sum x_i^2}} \qquad = \frac{\beta_1 \cancel{\sum x_i^2}}{\cancel{\sum x_i^2}} = \beta_1$$

thus proved.

(c). $\cancel{var}$ $var(y|x) = \sigma^2$ by assumption.

$$var(\hat{\beta}_1) = var(\hat{\beta}_1 | x) = var\left(\frac{\sum y_i x_i}{\sum x_i^2} \Big| x\right)$$

$$= \frac{\sum x_i^2 var(y_i)}{(\sum x_i^2)^2} = \frac{\sum x_i^2}{(\sum x_i^2)^2} \cdot \sigma^2 = \frac{1}{\sum x_i^2} \cdot \sigma^2.$$

7. (a). $E(\hat{\beta_0}|x) = E(\bar{y} - \hat{\beta_1}\bar{x})$.

$$E(\hat{\beta_1}|x) = E\left(\frac{\sum(x_i - \bar{x}) y_i}{\sum(x_i - \bar{x})^2} \middle| x \right) = \frac{\sum(x_i - \bar{x}) E(y_i|x_i)}{\sum(x_i - \bar{x})^2}$$

$E(y_i|x_i) = \beta_0 + \beta_1 X$

$$E(\hat{\beta_1}|x) = \frac{\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum(x_i - \bar{x})^2} = \beta_1$$

thus $\hat{\beta_1}$ is unbiased estimator.

$$E(\hat{\beta_0}|x) = E\left(\frac{1}{n}\sum y_i\right) - \hat{\beta_1}\bar{x}$$
$$= \frac{1}{n}\sum E(y_i) - \beta_1\bar{x}$$
$$= \frac{1}{n}\sum(\beta_0 + \beta_1 x_i) - \beta_1 \bar{x}$$
$$= \beta_0 + \beta_1 \cdot \frac{1}{n}\sum x_i - \beta_1 \bar{x}$$
$$= \beta_0 + \beta_1 \cdot \frac{1}{n}\sum x_i - \beta_1 \cdot \frac{1}{n}\sum x_i = \beta_0.$$

thus $\hat{\beta_0}$ is unbiased estimator.

(c) $\begin{cases} N(0,25) & x_i \leq 65 \\ T_3(0,10) & 65 < x_i \leq 70 \\ \text{uniform}[-8,8] & x_i > 70 \end{cases}$

variance from the three intervals are not consistent.
The assumption of homoskedasticity suggests that the variance is consistent for all x_is. Thus in this case, the assumption of homoskedasticity is not valid.

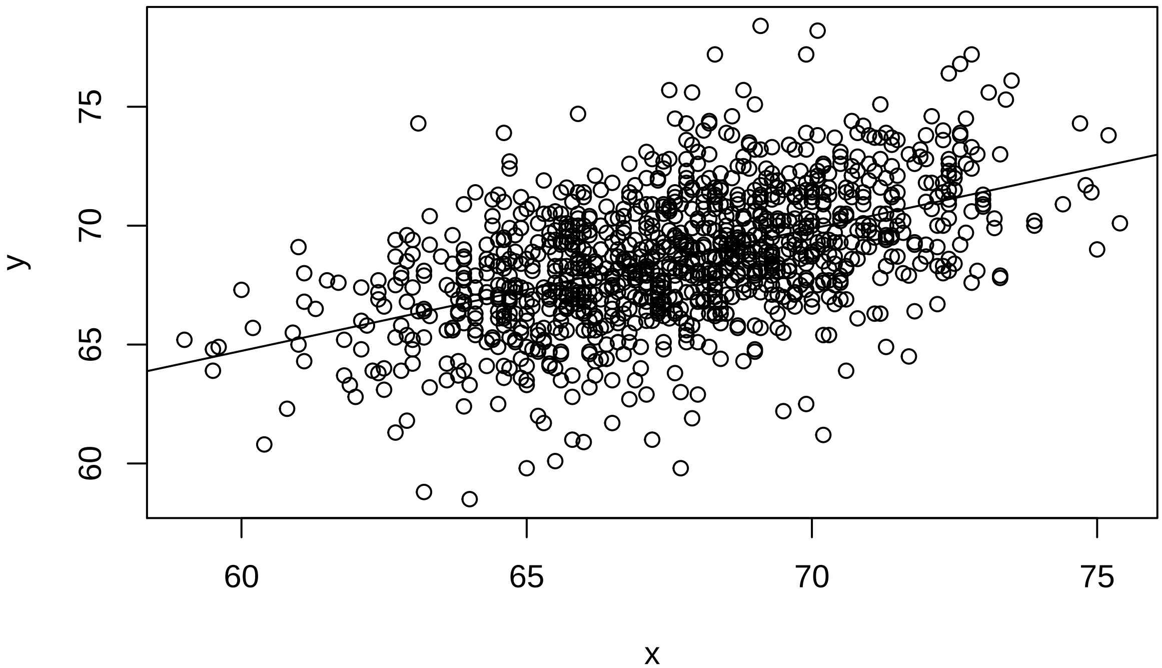
stat151a HW1 coding part

1.

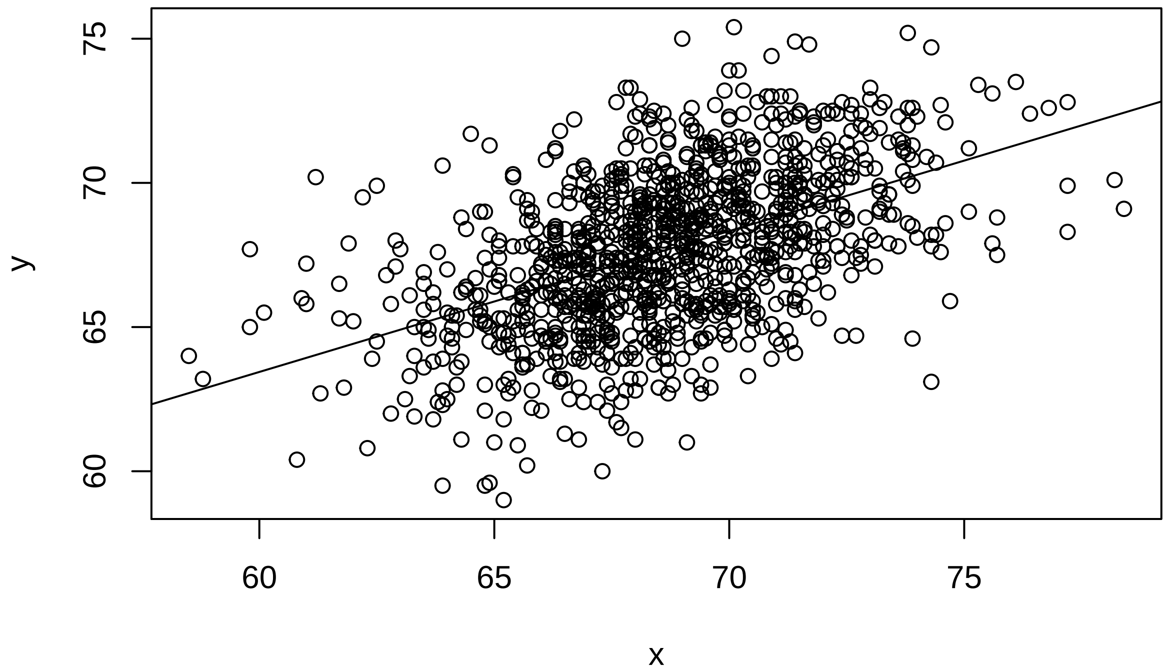
(d)

```
pearson<- read.delim("PearsonHeightData.txt")
x=pearson[,1] #father's height
y=pearson[,2]#son's height

slm=lm(y ~x,data=pearson)
plot(x,y,abline(slm))
```




```
x=pearson[,2] #son's height
y=pearson[,1] #father's height
slm=lm(y ~x,data=pearson)
plot(x,y,abline(slm))
```



2.meap93 data(meap93)

```
meap93=get(load("/Users/xiaoyingliu/Desktop/meap93.RData"))
head(meap93)
```

```
##   lnchprg enroll staff expend salary benefits droprate gradrate math10
## 1     1.4   1862 112.6   5765  37498     7420      2.9     89.2   56.4
## 2     2.3  11355 101.2   6601  48722    10370      1.3     91.4   42.7
## 3     2.7   7685 114.0   6834  44541     7313      3.5     91.4   43.8
## 4     3.4   1148  85.4   3586  31566     5989      3.6     86.6   25.3
## 5     3.4   1572  96.1   3847  29781     5545      0.0    100.0   15.3
## 6     3.4   2496 101.1   5070  36801     5895      2.7     89.2   46.0
##   sci11 totcomp ltotcomp  lexpend  lenroll   lstaff    bensal  lsalary
## 1  67.9   44918 10.71259 8.659560 7.529407 4.723842 0.1978772 10.53204
## 2  65.3   59092 10.98685 8.794976 9.337414 4.617099 0.2128402 10.79389
## 3  54.3   51854 10.85619 8.829665 8.947025 4.736198 0.1641858 10.70417
## 4  60.0   37555 10.53356 8.184793 7.045776 4.447346 0.1897295 10.35984
## 5  65.8   35326 10.47237 8.255049 7.360104 4.565389 0.1861925 10.30163
## 6  60.5   42696 10.66186 8.531096 7.822445 4.616110 0.1601859 10.51328
```

```
typeof(meap93)
```

```
## [1] "list"
```

```
y=meap93$math10
x=meap93$lnchprg
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

a) Fit a simple linear regression model for y on x.

Report the estimates of

```
slm1=lm(y ~x,data=meap93)
beta0=summary(slm1)$coefficient[1,1]
beta1=summary(slm1)$coefficient[2,1]
beta0_se=summary(slm1)$coefficient[1,2]
beta1_se=summary(slm1)$coefficient[2,2]
beta1
```

```
## [1] -0.3188643
```

```
beta0
```

```
## [1] 32.14271
```

```
beta0_se
```

```
## [1] 0.9975824
```

```
beta1_se
```

```
## [1] 0.03483933
```

b) We would expect the lunch program to have a positive effect on student perfor- mance. Does your model support such a positive relationship?

no,negative relationship. Since beta1 is negative which is not of our expectation. I am assuming that all the explanatory varibles might have some dependency with each other. For example, the lunch program might affect the droprate, and droprate might affect the math10 performance.Overall, lunch program has a negative effect on student performance of math10.

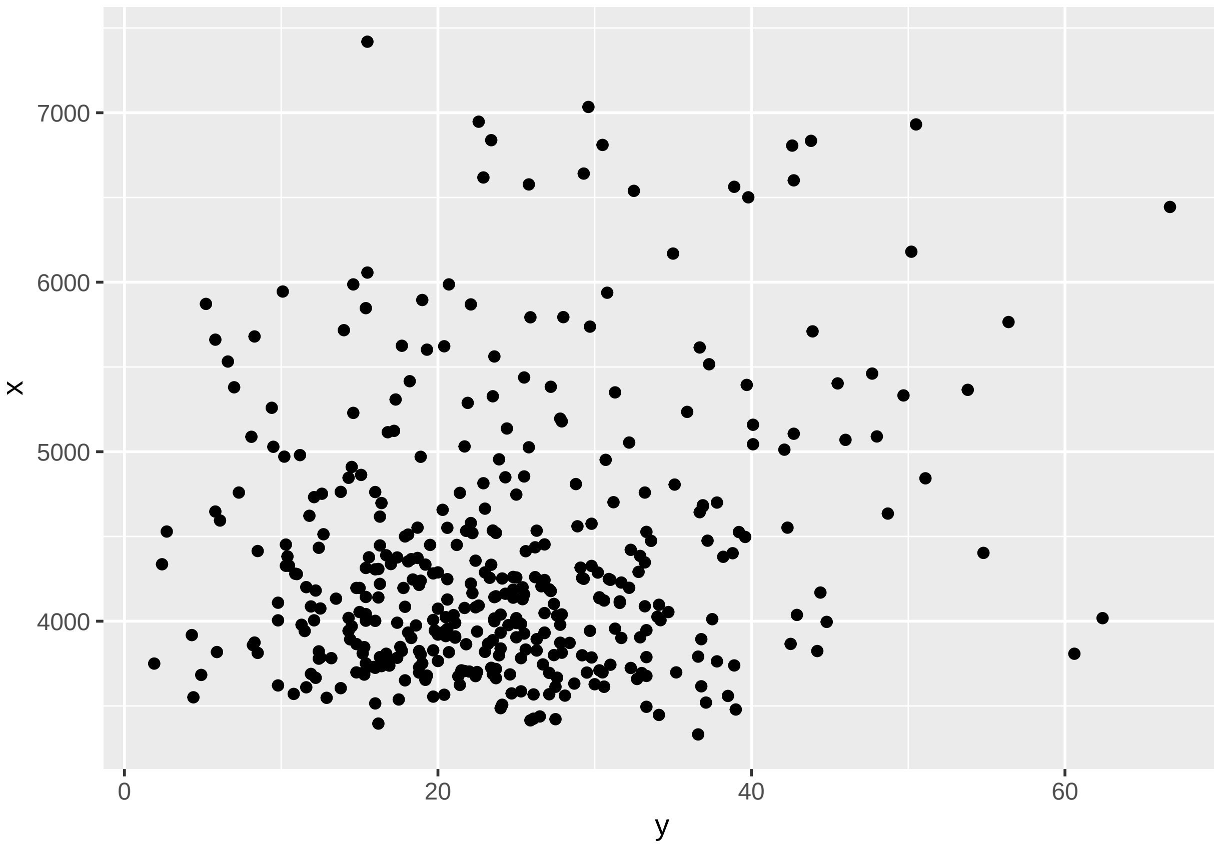
3.

```
head(meap93)
```

```
##   lnchprg enroll staff expend salary benefits droprate gradrate math10
## 1     1.4   1862 112.6   5765  37498     7420      2.9     89.2   56.4
## 2     2.3  11355 101.2   6601  48722    10370      1.3     91.4   42.7
## 3     2.7   7685 114.0   6834  44541     7313      3.5     91.4   43.8
## 4     3.4   1148  85.4   3586  31566     5989      3.6     86.6   25.3
## 5     3.4   1572  96.1   3847  29781     5545      0.0    100.0   15.3
## 6     3.4   2496 101.1   5070  36801     5895      2.7     89.2   46.0
##   sci11 totcomp ltotcomp  lexpend  lenroll   lstaff    bensal  lsalary
```

```
## 1  67.9   44918 10.71259 8.659560 7.529407 4.723842 0.1978772 10.53204
## 2  65.3   59092 10.98685 8.794976 9.337414 4.617099 0.2128402 10.79389
## 3  54.3   51854 10.85619 8.829665 8.947025 4.736198 0.1641858 10.70417
## 4  60.0   37555 10.53356 8.184793 7.045776 4.447346 0.1897295 10.35984
## 5  65.8   35326 10.47237 8.255049 7.360104 4.565389 0.1861925 10.30163
## 6  60.5   42696 10.66186 8.531096 7.822445 4.616110 0.1601859 10.51328
```

```r
y=meap93$math10
x=meap93$expend
ggplot(data=meap93,aes(y,x))+geom_point()
```



(a)Does the additional dollar have same effect, or it has

a diminishing effect?

beta1 seems really small, from the scatter plot, we can conclude that even if x is very large, y tends not to exceed a certain value, and there is no obvious linear relationship shown from the scatter plot. To conclude, expand has a diminishing effect rather than the same positive linear effect on math10 passing rate.

(c)

```
y=meap93$math10
x=log(meap93$expend)
slm3=lm(y ~x,data=meap93)
beta1=summary(slm3)$coefficient[2,1]
beta0=summary(slm3)$coefficient[1,1]
beta0_se=summary(slm3)$coefficient[1,2]
beta1_se=summary(slm3)$coefficient[2,2]
beta1
```

```
## [1] 11.1644
```

```
beta0
```

```
## [1] -69.34116
```

```
beta0_se
```

```
## [1] 26.53013
```

```
beta1_se
```

```
## [1] 3.169011
```

(d)

as shown in (b), beta1/10 is the percentage point increase in math10 if spending increases by 10 percent.

(e)fitted value of math10 might be bigger than 100, why is that

not a concern in this dataset?

```
max(meap93$math10)
```

```
## [1] 66.7
```

in order for the math10 value greater than 100, the student might need to spend a huge amount, which is not possible. Thus,it is not much of a worry in the math10 ~log(expend) model.

5.

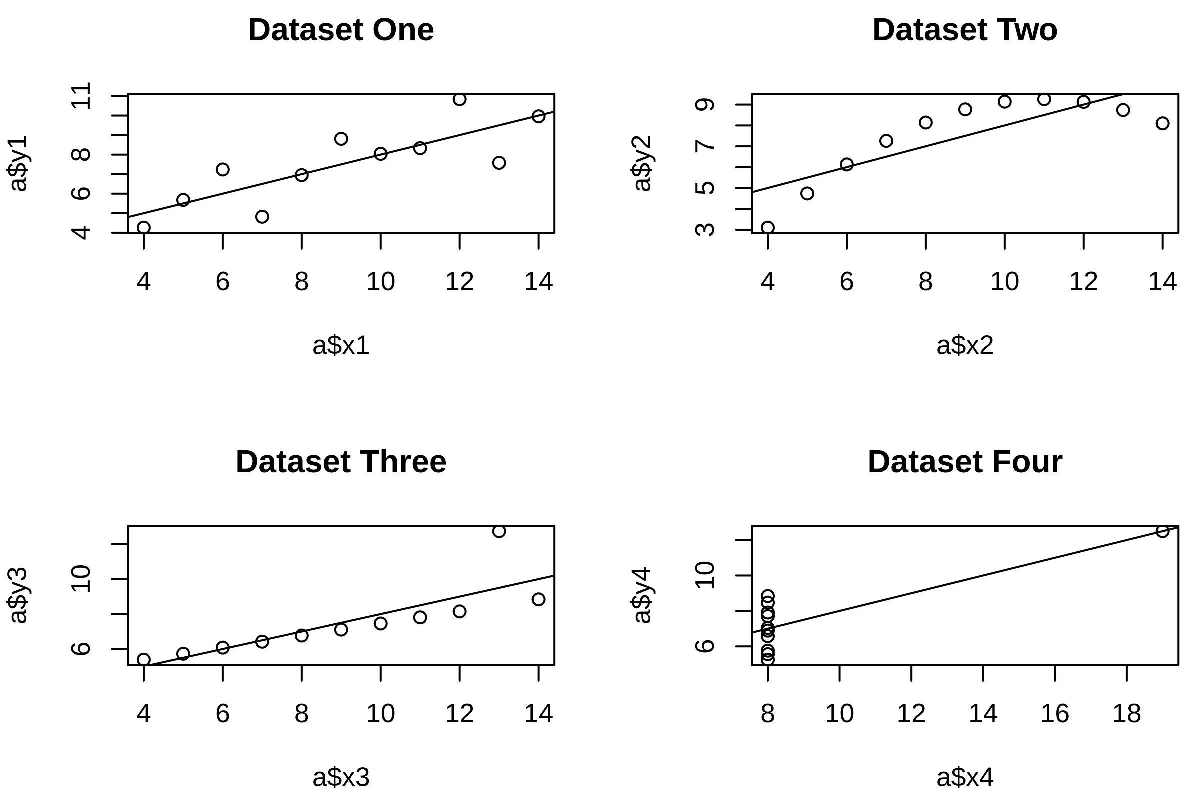
(a)

```
library(datasets)
a=anscombe
head(a)
```

```
##   x1 x2 x3 x4   y1   y2    y3   y4
## 1 10 10 10  8 8.04 9.14  7.46 6.58
## 2  8  8  8  8 6.95 8.14  6.77 5.76
## 3 13 13 13  8 7.58 8.74 12.74 7.71
## 4  9  9  9  8 8.81 8.77  7.11 8.84
## 5 11 11 11  8 8.33 9.26  7.81 8.47
## 6 14 14 14  8 9.96 8.10  8.84 7.04
```

```
par(mfrow=c(2,2))

plot(a$x1,a$y1, main=paste("Dataset One"),abline(lm(y1 ~x1,data=a)))
plot(a$x2,a$y2, main=paste("Dataset Two"),abline(lm(y2 ~x2,data=a)))
plot(a$x3,a$y3, main=paste("Dataset Three"),abline(lm(y3 ~x3,data=a)))
plot(a$x4,a$y4, main=paste("Dataset Four"),abline(lm(y4 ~x4,data=a)))
```



from the scatter plot and the linear model, the linear models do make sense on dataset1, and 3. However, in dataset4, all the x values are too close to each other, and thus it is unreasonable to generate a linear model based on such dataset. In dataset2, from the scatter plot, it seems that (x,y) follows quadratic function which is obviously not linear.

(b)predict y when x is 10.Does it make sense?

```
slm2=lm(y2 ~x2,data=a)
slm3=lm(y3 ~x3,data=a)
slm4=lm(y4 ~x4,data=a)

y1=summary(slm1)$coefficient[1,1]+summary(slm1)$coefficient[2,1]*10
y2=summary(slm2)$coefficient[1,1]+summary(slm2)$coefficient[2,1]*10
y3=summary(slm3)$coefficient[1,1]+summary(slm3)$coefficient[2,1]*10
y4=summary(slm4)$coefficient[1,1]+summary(slm4)$coefficient[2,1]*10
y1
```

```
## [1] 28.95407
```

```
y2
```

```
## [1] 8.000909
```

```
y3
```

```
## [1] 7.999727
```

```
y4
```

```
## [1] 8.000818
```

it does not make sense for dataset4 and dataset2. The y value corresponding to x=10 is far from the value predicted obviously. It does make sense for dataset1 and dataset3.

7.

(b)

```
n=100
beta0=32
beta1=0.5
x=seq(59,76,length.out=100)
M=10000
y1=c()
y2=c()
```

```r
y3=c()
beta_matrix=matrix(0,nrow=0,ncol=4)
```

generate one trial y data

```r
one_trial=function(x){
for(i in x){
  if(i<=65){
  y1=c(y1,rnorm(1,beta0+beta1*i,5))
  }
  if((i>65)&&(i<=70)){
  y2=c(y2,10*rt(n=1,df=3)+beta0+beta1*i)
  }
  if(i>70){
  y3=c(y3,(runif(1,min=-8,max=8)+beta1*i+beta0))
  }
}
  y=c(y1,y2,y3)
  slm=lm(y ~x)
  beta0_hat=summary(slm)$coefficient[1,1]
  beta1_hat=summary(slm)$coefficient[2,1]
  #slm=lm(y ~x)
  #summary(slm)$coefficient[1,2]
  se_beta0=summary(slm)$coefficient[1,2]
  se_beta1=summary(slm)$coefficient[2,2]
  beta_vector=c(beta0_hat,beta1_hat,se_beta0,se_beta1)
  return(beta_vector)
}



for(i in 1:M){
  beta_matrix=rbind(beta_matrix,one_trial(x))
}


beta0_bias=mean(beta_matrix[,1])-beta0
beta1_bias=mean(beta_matrix[,2])-beta1
beta0_bias
```

```
## [1] -0.006666817
```

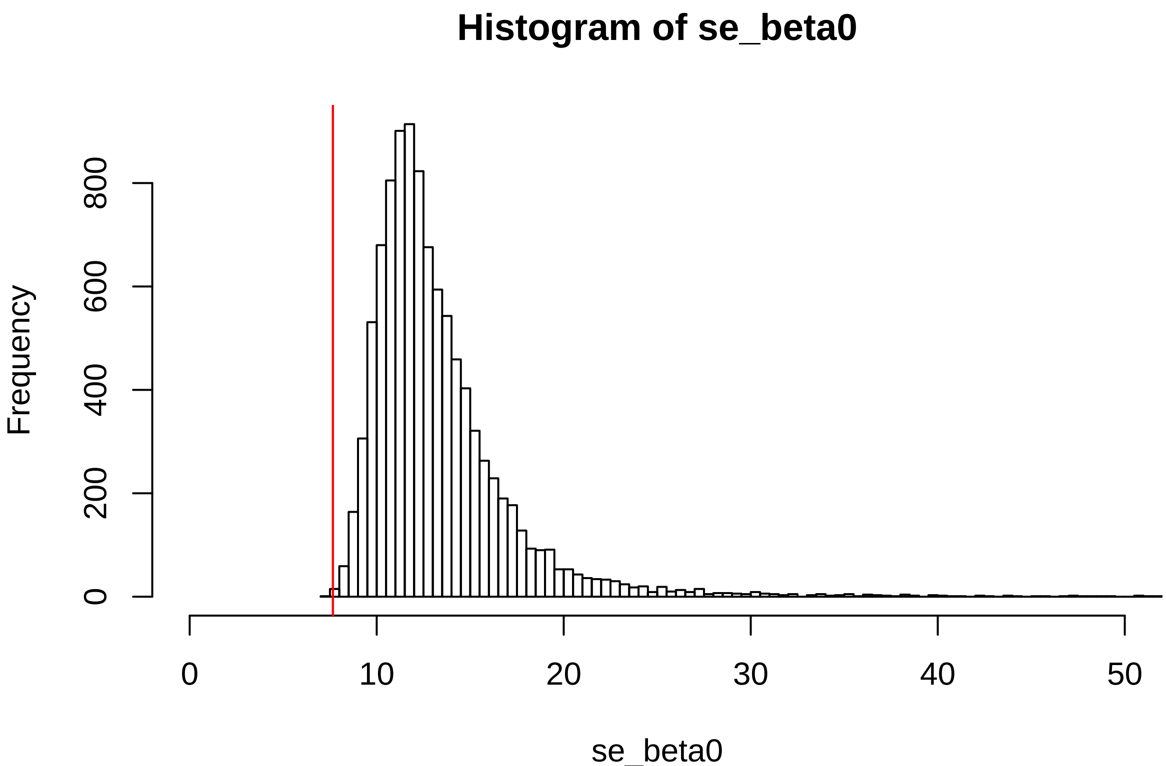
```r
beta1_bias
```

```
## [1] -7.307511e-05
```

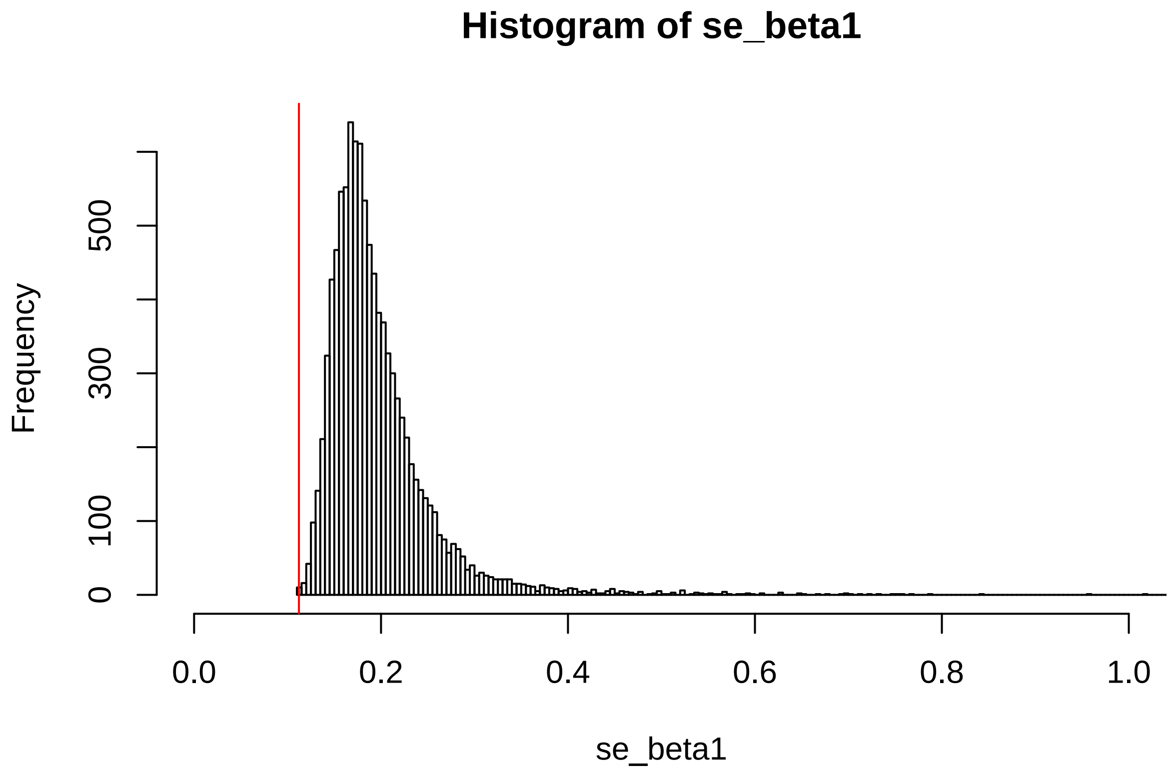
Both estimates are close enough to 0, which indicates that the bias is close to zero(unbiasedness is proved)

(d)

```
sd_beta0hat=sd(beta_matrix[,1])
sd_beta1hat=sd(beta_matrix[,2])
se_beta0=beta_matrix[,3]
se_beta1=beta_matrix[,4]
hist(se_beta0,breaks=1000,xlim=c(0,50))
abline(v=sd_beta0hat,col="red")
abline(v=sd_beta0hat,col="red")
```




```
hist(se_beta1,breaks=1000,xlim=c(0,1))
abline(v=sd_beta1hat,col="red")
```



the se reported by R is not reliable when homoskedasticity is violated. Because in the assumptions of our linear model,we assume that the noise would have same variance. In this case, the noise follows different distribution and have different variance.