Homework 1 Solution

Stephanie DeGraaf

September 6, 2018

Problem 1

a)

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

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

b)

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y}) x_i}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

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

c)

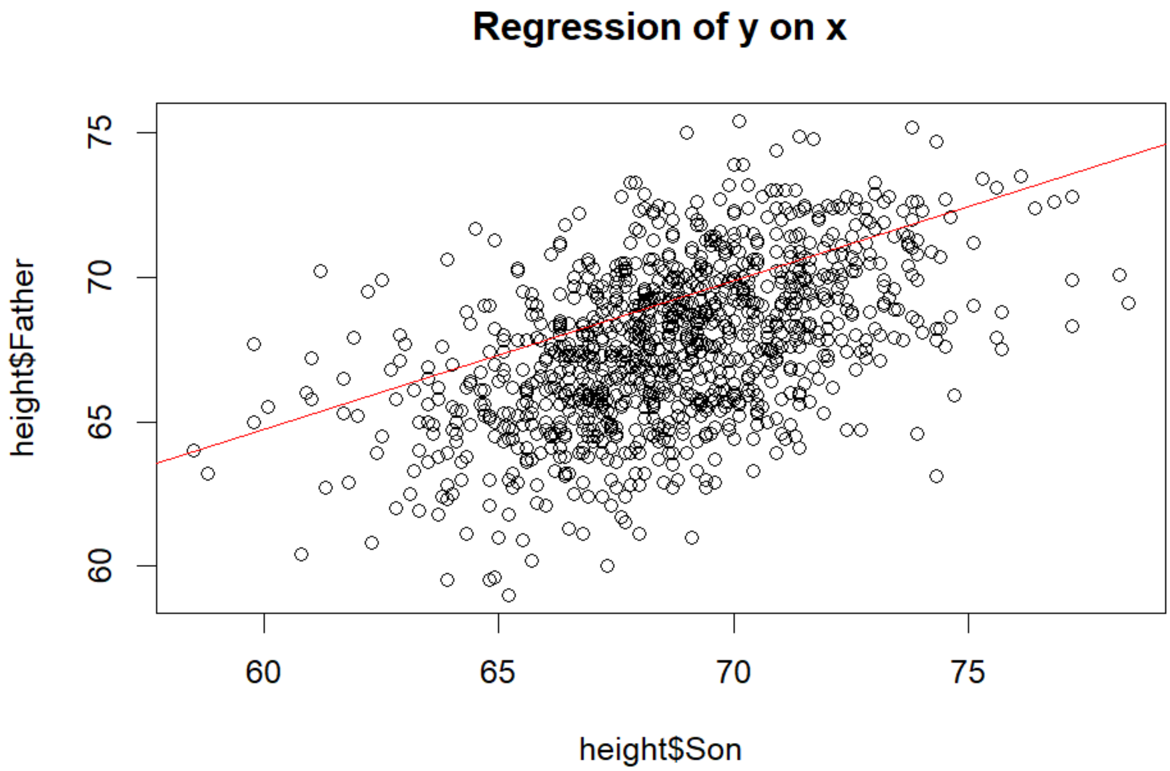
No, it is not always true that $\hat{\alpha}_1 = 1/\hat{\beta}_1$. One way to see this is to write the regression coefficients in terms of the correlation and standard deviations. Let s_x and s_y denote the sample standard deviations of x and y respectively, and let r denote the sample correlation coefficient between x and y. Then

$$\hat{\beta}_1 = r\left(\frac{s_y}{s_x}\right), \qquad \text{and} \qquad \hat{\alpha}_1 = r\left(\frac{s_x}{s_y}\right).$$

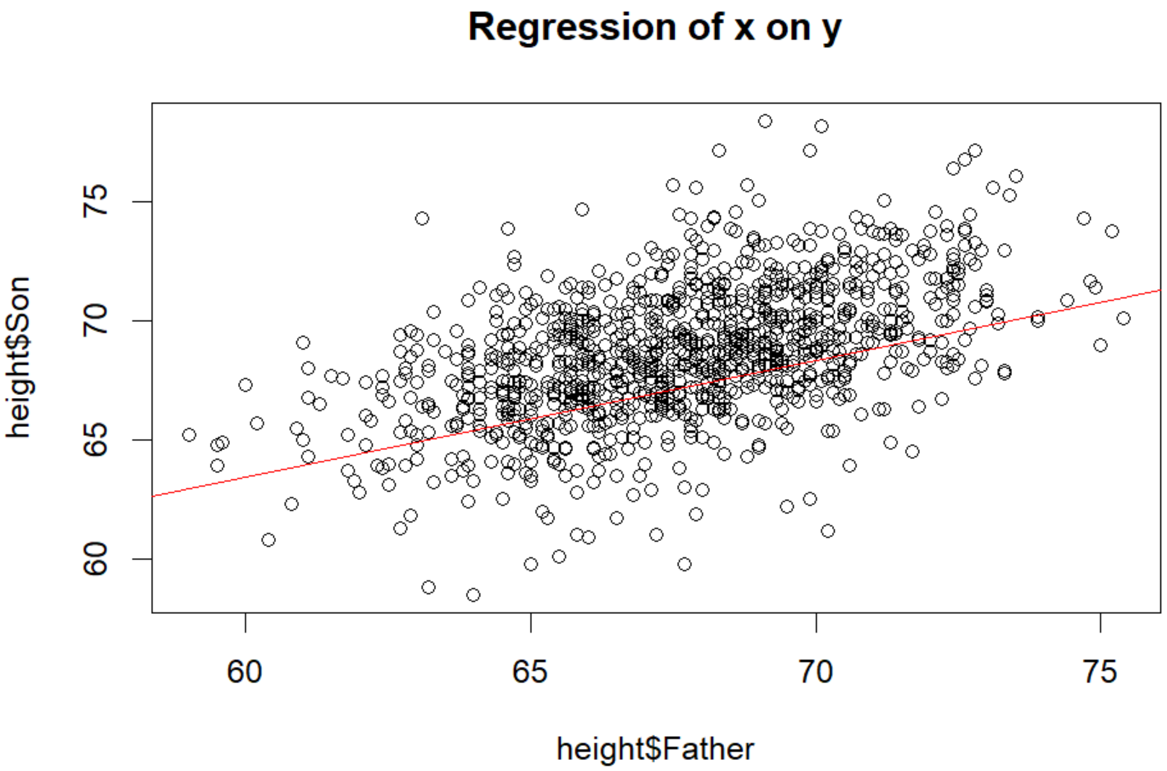
This implies $\hat{\alpha}_1 = r^2/\hat{\beta}_1$. Thus, when $r^2 = 1$, the equality holds, but in general it is not true that $\hat{\alpha}_1 = 1/\hat{\beta}_1$.

d)

```r
height<- read.table("PearsonHeightData.txt", header = T)
lm_yx<- lm(Son ~ Father, data = height)
plot(height$Son, height$Father, main = "Regression of y on x")
abline(lm_yx, col = "red")
```



```
lm_xy<- lm(Father ~ Son, data = height)
plot(height$Father, height$Son, main = "Regression of x on y")
abline(lm_xy, col = "red")
```



Problem 2

a)

```r
load("meap93.Rdata")
model<- lm(math10 ~ lnchprg, data = data)
kable(summary(model)$coefficients)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	32.1427116	0.9975824	32.220609	0
lnchprg	-0.3188643	0.0348393	-9.152422	0

b)

No, the coefficient of the lunch program variable is negative, indicating a negative effect on student performance. This could be because we are not accounting for other variables that influence math scores.

Problem 3

a)

A diminishing effect seems more appropriate. Increasing the spending in a low intial budget would have more impact than increasing the spending in a budget where spending is already high. (If the spending is already high, students likely have already have access to useful resources, and increasing the spending would probably not provide much additional value to students that would increase their math pass rate.)

b)

Let y represent the *math10* variable and x represent the *expend* variable. We are interested in the difference $y_2 - y_1$, where

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 \log(x_1) + e \\
y_2 &= \beta_0 + \beta_1 \log(x_1 * 1.10) + e.
\end{aligned}
$$

Taking the difference,

$$y_2 - y_1 = \beta_1 \log(1.10) \approx \frac{\beta_1}{10}.$$

Thus, a 10% increase in *expend* corresponds to a change in *math10* of $\beta_1/10$.

c)

```r
model<- lm(math10 ~ lexpend, data = data)
kable(summary(model)$coefficients)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-69.34110	26.530129	-2.613673	0.0092904
lexpend	11.16439	3.169011	3.522991	0.0004752

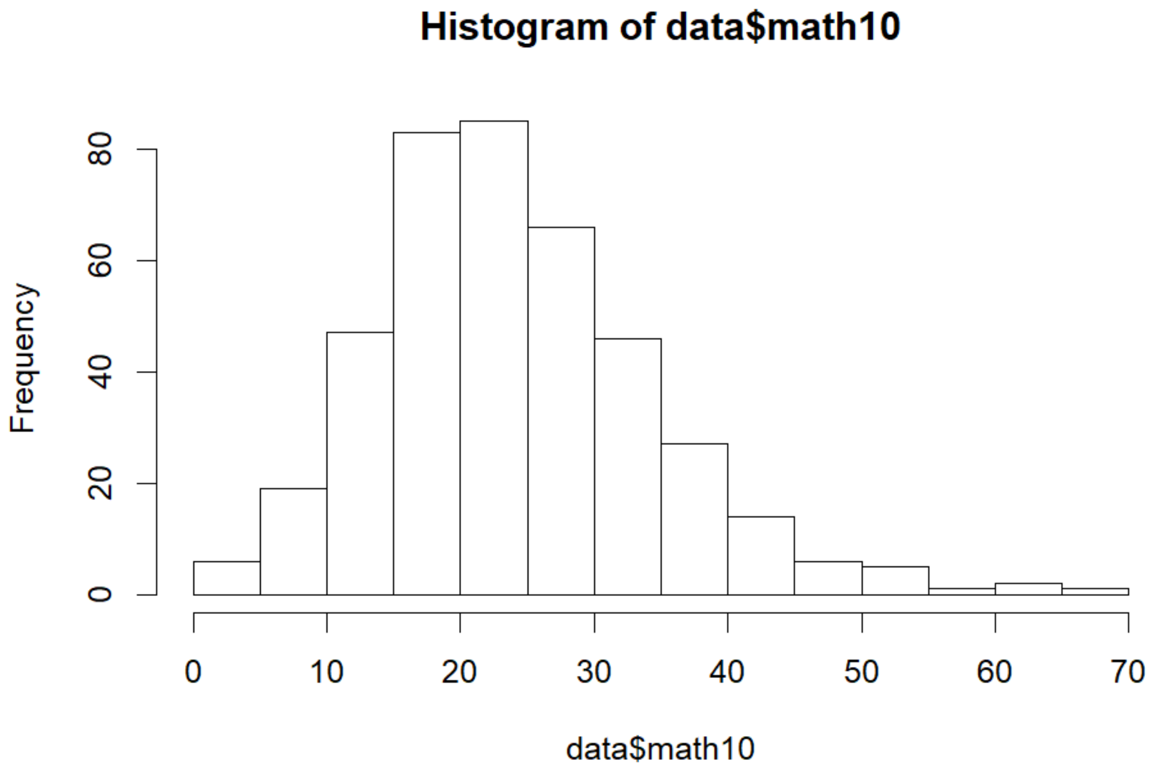
d)

If spending increases by 10 percent, *math10* will increase by an estimated 1.1164395 percentage points.

e)

In this dataset, the *math10* scores are all much less than 100: they are centered around 24 with a maximum value of 66.7.

```r
hist(data$math10)
```



Thus, our model is designed to measure the effects of variables within this range of y values. Since our model is based on this lower range of y values, we have no reason to believe that our model would be sensible for predictions of y that are much higher than this range.

Problem 4

a)

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

b)

Using the formula from part a and under the assumption of independence, we can rewrite the variance as

$$\begin{aligned} var(\hat{\beta}_0 + \hat{\beta}_1 a | x_1, ..., x_n) &= var\left(\sum_{i=1}^{n} y_i \left(\frac{1}{n} + \frac{(x_i - \bar{x})(a - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) | x_1, ..., x_n\right) \\ &= \sum_{i=1}^{n} var\left(y_i \left(\frac{1}{n} + \frac{(x_i - \bar{x})(a - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) | x_1, ..., x_n\right). \end{aligned}$$

Since we are conditioning on the x values, this further simplifies to

$$\begin{aligned} &= \sum_{i=1}^{n} \left[\left(\frac{1}{n}\right)^2 var(y_i | x_1, ..., x_n) + \left(\frac{(x_i - \bar{x})(a - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^2 var(y_i | x_1, ..., x_n)\right] \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2 (a - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \end{aligned}$$

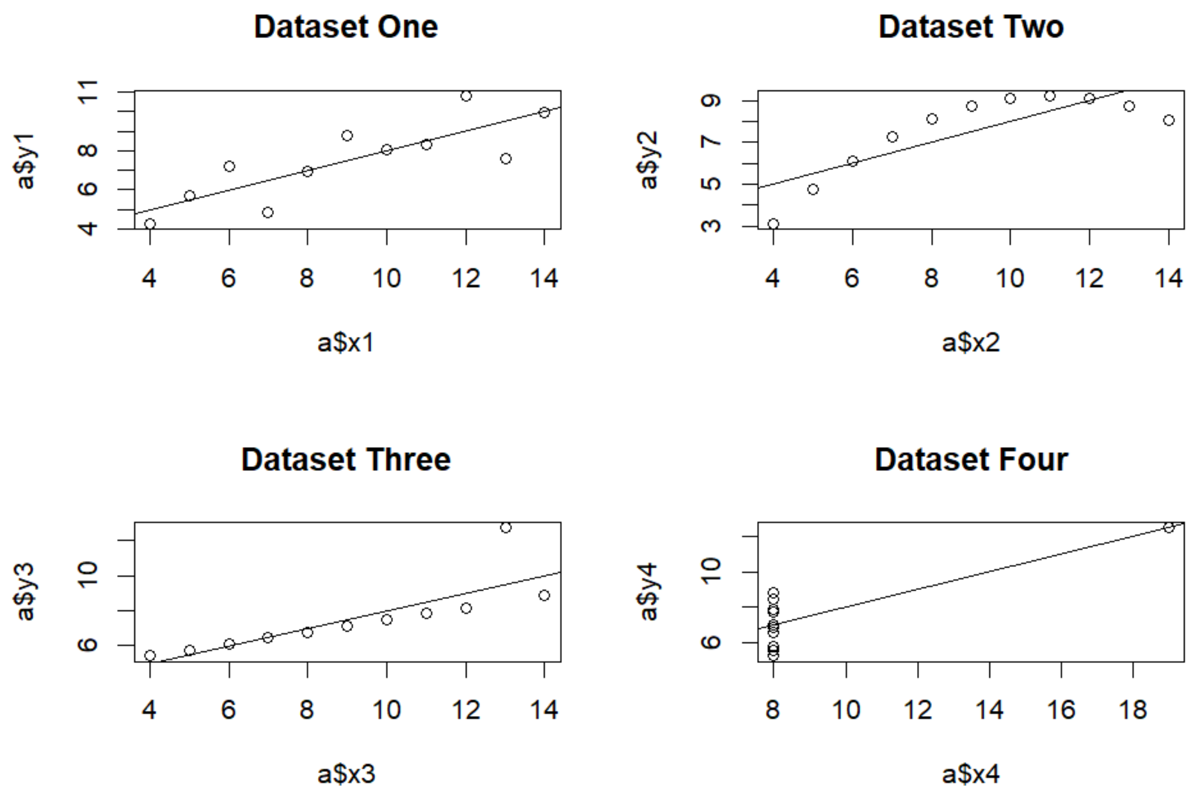
c)

If we minimize the formula in part b, we notice that the first term is constant and does not depend on a. The second term only depends on a through the term $(a - \bar{x})^2$, which is minimized when $(a - \bar{x})^2 = 0$, or $a = \bar{x}$.

Problem 5

a)

```
library(datasets)
a<- anscombe
lm1<- lm(a$y1 ~ a$x1)
lm2<- lm(a$y2 ~ a$x2)
lm3<- lm(a$y3 ~ a$x3)
lm4<- lm(a$y4 ~ a$x4)
par(mfrow=c(2,2))
plot(a$x1,a$y1, main=paste("Dataset One"))
abline(lm1)
plot(a$x2,a$y2, main=paste("Dataset Two"))
abline(lm2)
plot(a$x3,a$y3, main=paste("Dataset Three"))
abline(lm3)
plot(a$x4,a$y4, main=paste("Dataset Four"))
abline(lm4)
```



The linear model seeems appropriate for Dataset One and Dataset Three: these datasets show a strong linearly positive correlation between x and y. However, the linear model does not seem appropriate for Datasets Two and Four. Dataset Two has a clear curve to it, suggesting a quadratic trend. Dataset Four has only two distinct values of x, and there is nothing to suggest a linear relationship between y and x.

b)

```
unname(lm1$coefficients[1] + lm1$coefficients[2]*10)
```

```
## [1] 8.001
```

```
unname(lm2$coefficients[1] + lm2$coefficients[2]*10)
```

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

```
unname(lm3$coefficients[1] + lm3$coefficients[2]*10)
```

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

```
unname(lm4$coefficients[1] + lm4$coefficients[2]*10)
```

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

The predictions for $x = 10$ make sense for the datasets where the linear model is appropriate, namely in Dataset One and Dataset Three. The prediction in Dataset Two would likely underestimate the truth; the prediction in Dataset Four is meaningless.

Problem 6

a)

We can compute the least squares estimate $\hat{\beta}_1$ by minimizing the function

$$Q(\beta_1) = \sum_{i=1}^{n}(y_i - \beta_1 x_i)^2.$$

Differentiating with respect to β_1 and setting the derivate equal to zero implies

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

b)

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_1|X) &= \mathbb{E}\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}|X\right) \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})\mathbb{E}(y_i|X)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_1 x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\beta_1(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\beta_1(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)}{\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2} \\
&= \frac{\beta_1(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
&= \beta_1
\end{aligned}$$

c)

$$\begin{aligned}
Var(\hat{\beta}_1|x) &= Var\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}|x\right) \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} Var(y_i|x) \\
&= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}$$

Problem 7

a)

Yes. If we condition on the values of x_i, the least squares estimates are unbiased. The distribution of y can be arbitrary in the simple linear model, and it will not affect the unbiasedness of the coefficients as long as the mean of y equals the linear model's expectation. In the model here, the mean of y will always equal $\beta_0 + \beta_1 x_i$, as required, since the error terms each have mean zero. Since $\mathbb{E}(y_i|x) = \beta_0 + \beta_1 x_i$ for any value of x_i, it follows that $\mathbb{E}(\hat{\beta}_0) = \beta_0$ and $\mathbb{E}(\hat{\beta}_1) = \beta_1$.

b)

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

generate_y<- function(xi){
  if (xi <= 65) {
    yi = rnorm(1, mean = beta0+beta1*xi, sd = 5)
  } else if(xi <= 70){
    yi = beta0+beta1*xi + 10*rt(1, df = 3)
  } else {
    yi = beta0+beta1*xi + runif(1, min = -8, max = 8)
  }
  return(yi)
}
intercepts<- rep(NA, M)
slopes<- rep(NA, M)
for (j in 1:M){
  y_j<- sapply(x, generate_y)
  model_j<- lm(y_j ~ x)
  intercepts[j]<- model_j$coefficients[1]
  slopes[j]<- model_j$coefficients[2]
}

beta0_bias<- mean(intercepts-beta0)
beta1_bias<- mean(slopes-beta1)
beta0_bias
```

```
## [1] -0.05712906
```

```
beta1_bias
```

```
## [1] 0.0007315579
```

Yes, the estimates of bias are very close to 0.

c)

Homoskedasticity assumes that the error terms are constant and do not depend on the x-value. The model here directly violate this assumption, because the variance of each y_i depends on the value of x_i. For $x_i \leq 65$, the variance of y_i is 25. For $65 < x_i \leq 70$, the variance of y_i is 3. For $x_i > 70$, the variance of y_i is $16^2/12 = 21.33$. Thus, the variance of y is not constant: it changes depending on the x value.

d)

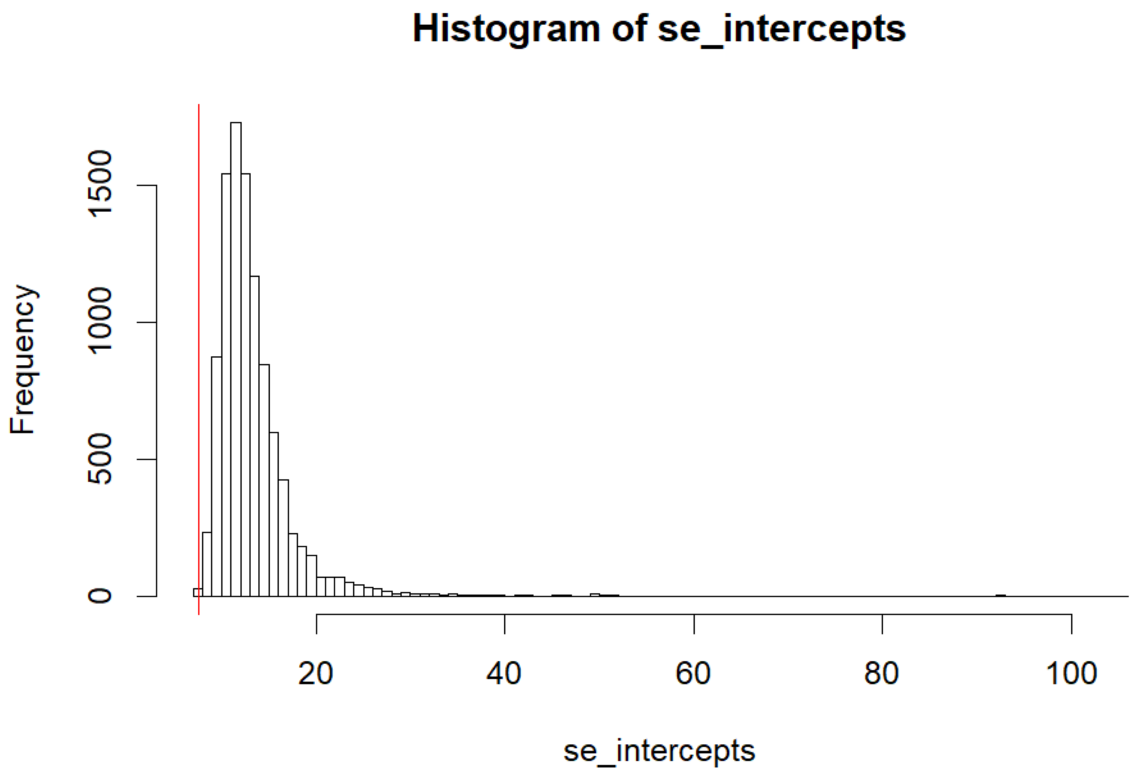
```
se_intercepts<- rep(NA, M)
se_slopes<- rep(NA, M)
for (j in 1:M){
  y_j<- sapply(x, generate_y)
  model_j<- lm(y_j ~ x)
  summary_j<- summary(model_j)
  intercepts[j]<- model_j$coefficients[1]
  slopes[j]<- model_j$coefficients[2]
  se_intercepts[j]<- summary_j$coefficients[1,2]
  se_slopes[j]<- summary_j$coefficients[2,2]
}
sd(intercepts)
```

```
## [1] 7.569255
```

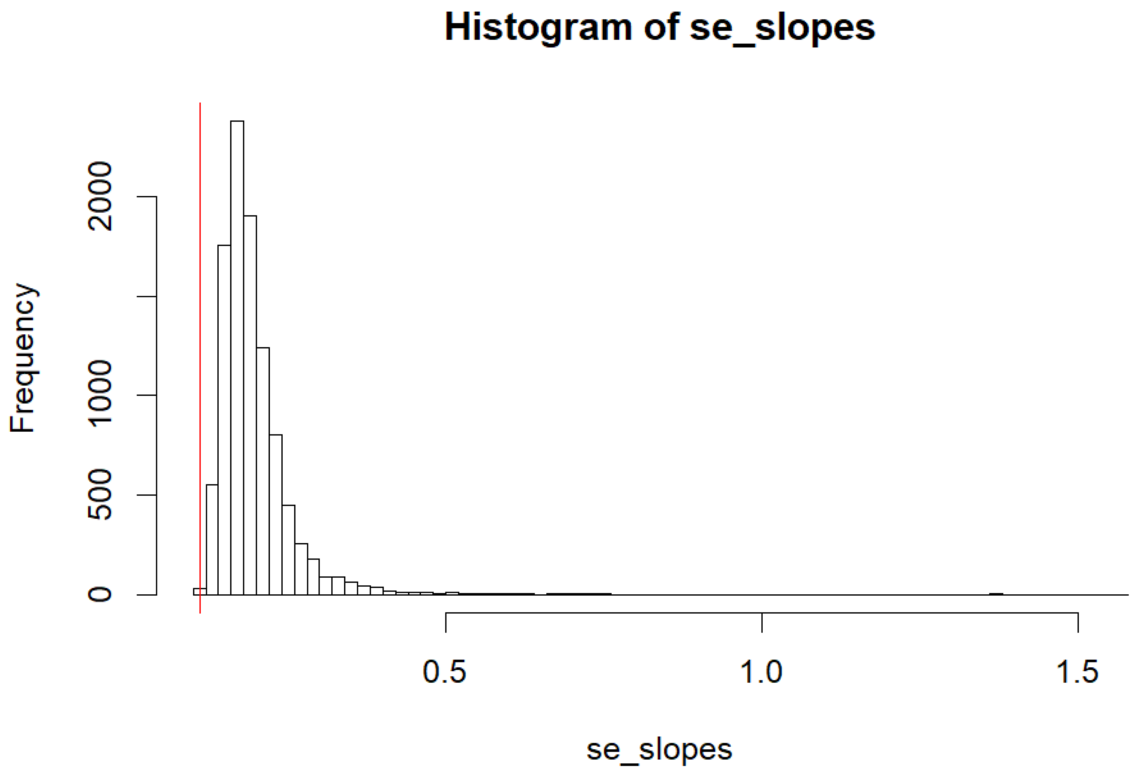
```
sd(slopes)
```

```
## [1] 0.1111177
```

```
hist(se_intercepts, breaks = 100)
abline(v=sd(intercepts), col = "red")
```



```
hist(se_slopes, breaks = 100)
abline(v=sd(slopes), col = "red")
```



The standard errors reported by R are not reliable when homoskedasticity is violated. As shown in the histograms, the standard errors are centered much higher than the standard deviations under the simulation. Moreover, there is a heavy right skew to these standard errors, suggesting that R will overestimate the standard errors when homoskedasticity is violated.