S631 HW3

Erik Parker

September 12, 2017

1. Show that $\hat{\beta}_0$ and $\hat{\beta}_1$ can be written as linear combinations $y_1, ..., y_n$

$$\hat{\beta}_0 = \sum_{i=1}^{n} a_i y_i \quad and \quad \hat{\beta}_1 = \sum_{i=1}^{n} b_i y_i \quad where \quad a_1, ..., a_n \quad and \quad b_1, ...b_n \in \mathbb{R}$$

because we know from ALR page 24 that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and $\hat{\beta}_1 = SXY/SXX$

where, from Table 2.1, page 23 we know:

$$SXX = \sum (x_i - \bar{x}) x_i \quad and \quad SXY = \sum (x_i - \bar{x}) y_i$$
$$so \quad \hat{\beta}_1 = \sum \frac{(x_i - \bar{x}) y_i}{(x_i - \bar{x}) x_i} = \sum \frac{(x_i - \bar{x})}{(x_i - \bar{x}) x_i} y_i = \sum (b_i y_i)$$
$$where \quad b_i = \frac{(x_i - \bar{x})}{(x_i - \bar{x}) x_i}$$

Furthermore, and using this $\hat{\beta}_1$:

$$\hat{\beta}_0 = \bar{y} - (\sum \frac{(x_i - \bar{x})}{(x_i - \bar{x}) x_i} y_i)\bar{x} = \sum \frac{1}{n} y_i - (\sum \frac{(x_i - \bar{x})}{(x_i - \bar{x}) x_i} y_i)\bar{x} = \sum (\frac{1}{n} - \frac{(x_i - \bar{x})}{(x_i - \bar{x}) x_i} \bar{x}) y_i = \sum (a_i y_i)$$
$$where \quad a_i = \frac{1}{n} - \frac{(x_i - \bar{x})}{(x_i - \bar{x}) x_i} \bar{x} = \frac{1}{n} - b_i \bar{x}$$

2. The bank, UBS, regularly reports on prices and earnings in major cities throughout the world. Three included measures are prices of commodities (1kg of rice, 1kg of bread, and the price of a Big Mac). Prices are measured in minutes of labor required for a typical worker to buy these goods.

1) The line $y = x$ is shown on the plot as the solid line. What is the key difference between points above and below this line?

> Points falling above the line $y = x$ saw an increase in the price, in terms of number of hours worked, from year 2003 to 2009. While points falling below the line saw a decrease in in cost from 2003 to 2009.

2) Which cities had the largest increases and decreases in rice price?

> The city showing the largest increase in rice price on this plot is Vilnius, followed closely by Budapest. Conversely, the city showing the largest decrease in rice price from 2003 to 2009 is Mumbai followed (not so closely) by Nairobi.

3) Does the dashed line, representing the OLS line, with a slope < 1 suggest that prices are lower in 2009 than in 2003?

Setting aside potential problems with the use of OLS here, the line with the of < 1 does seem to be suggesting that prices are lower in 2009 than in 2003, but only for locations where the 2003 rice price was greater than ~25. Before this point (where the majority of the data are), this model shows that rice prices were actually greater in 2009 than in 2003.

4) Give two reasons why simple linear regression is not likely appropriate for this problem.

First, there are a few outliers which are quite far from the majority of points and may be serving as strong leverage points and skewing the model. Secondly, the majority of the data used to create the model here is clustered around low x and y values, so this slope of the OLS line is generated mostly by the few outlying points to the right of the plot. This means that it is hard to generate a model that is of much explanatory use for larger x values as there are just too few of them.

3. Simulation. Assume the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, ..., n$$

where $e_i \sim N(0, \sigma^2)$ **for** $i = 1, ..., n$.

Let's set $\beta_0 = 10, \beta_1 = -2.5,$ *and* $n = 30$.

a) Set $\sigma = 100,$ *and* $x_i = i$ *for* $i = 1, ..., n$.

```
beta0 <- 10
beta1 <- -2.5
n <- 30
sigma <- 100

xi <- rep(0, n)
for (i in 1:n) {
    xi[i] = i
}
```

b) Simulation will have 10,000 iterations. Set a random seed using birthday (0330) and report the seed with responses. For each iteration, obtain and store linear regression parameter estimates: $\hat{\beta}_0$'s, $\hat{\beta}_1$'s, and $\hat{\sigma}^2$'s.

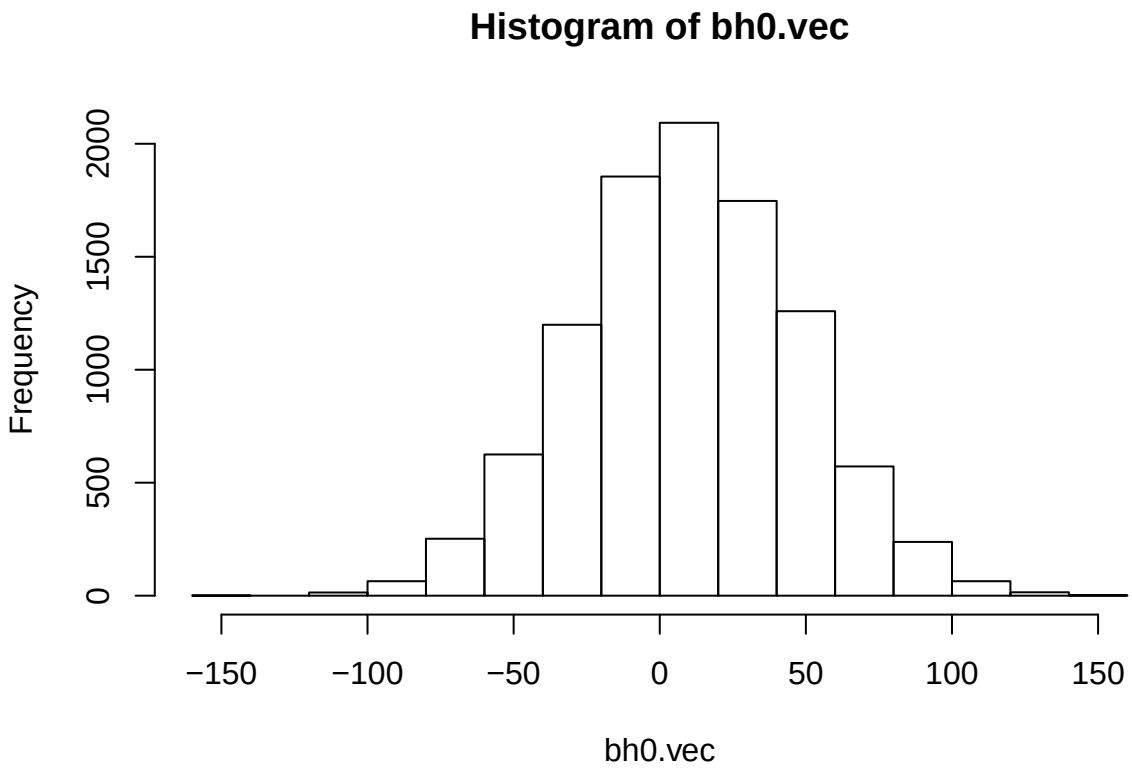
```
set.seed(330)
sim <- 10000
bh0.vec <- rep(0, sim)
bh1.vec <- rep(0, sim)
sig2.vec <- rep(0, sim)

for (i in 1:sim) {
    e <- rnorm(n, mean = 0, sd = sigma)
    y <- beta0 + beta1 * xi + e
    m1 <- lm(y ~ xi)
    bh0.vec[i] <- coef(m1)[1]
    bh1.vec[i] <- coef(m1)[2]
```

```
    sig2.vec[i] <- sum(e^2)/n - 2
}
```

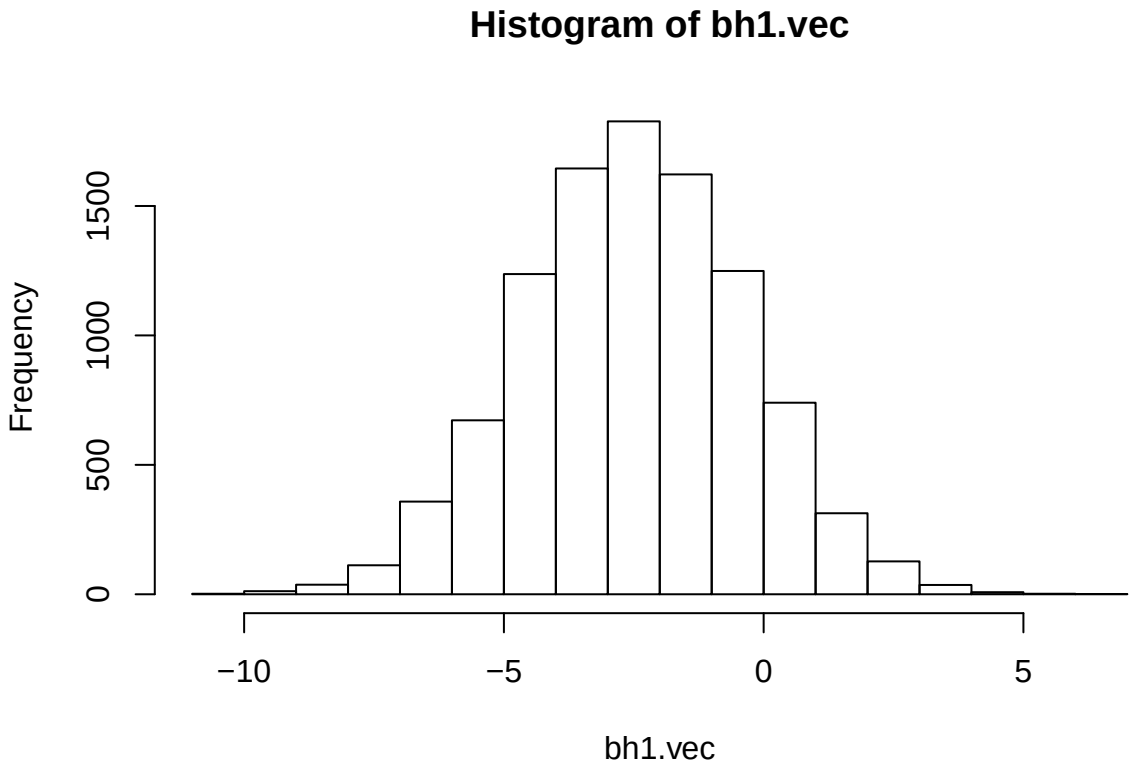
$\hat{\sigma}^2$ obtained by taking $\frac{RSS}{n-2}$ where $RSS = \sum \hat{e}_i^2$

c) Present three histograms for $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$. Describe the main charactersitics of these histograms.

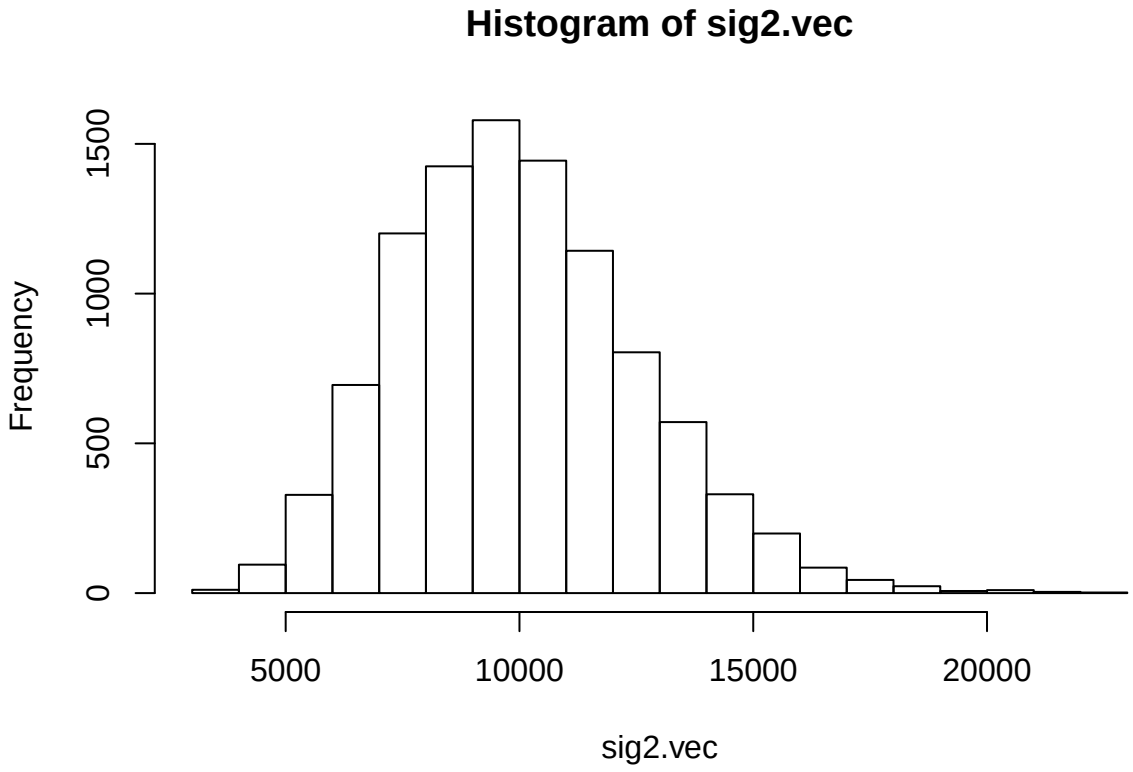
```
hist(bh0.vec)
```




```
hist(bh1.vec)
```




```
hist(sig2.vec)
```



All three of these histograms show generally bell shaped distributions, with the plotted values of $\hat{\beta}_0$ and $\hat{\beta}_1$ following close to a symmetric, normal distribution, and the values of $\hat{\sigma}^2$ following a slightly right-skewed distribution.

d) Find the averages of our estimates, how do they compare with the true parameters?

```
mean(bh0.vec)
```

```
## [1] 9.611829
```

```
mean(bh1.vec)
```

```
## [1] -2.490867
```

```
sqrt(mean(sig2.vec))
```

```
## [1] 99.94228
```

The averages of our estimated values are very close to the values of the true parameters supplied in the beginning of the simulation. This makes sense as we did 10,000 iterations of our simulation, and we expect our estimated values to move closer to the true values as the number of samples increases.

e) Find the sample variances of $\hat{\beta}_0$'s and $\hat{\beta}_1$'s. How do they compare with the true variances?

```
var(bh0.vec)
```

```
## [1] 1428.127
```

```
var(bh1.vec)
```

```
## [1] 4.581832
```

```
xbar <- sum(xi)/n

sxx <- sum(xi^2) - n * mean(xi)^2

varb0 <- sigma^2 * ((1/n) + (xbar^2)/sxx)
varb0
```

```
## [1] 1402.299
```

```
varb1 <- sigma^2 * (1/sxx)
varb1
```

```
## [1] 4.449388
```

The sample variances calculated from $\hat{\beta}_0$ and $\hat{\beta}_1$ are quite close to the true variances calculated using the formulas found in the book, and the original starting x and σ values. The variance of β_0 is larger and is off by only 26 between the estimate and parameter, while the variance β_1 is off by about 0.1 between the two.

f) Now set $\sigma = 100,\ and\ x_i = 100 * i\ for\ i = 1, ..., n$. Repeat parts b, d, and e. How does the new sample variance of $\hat{\beta}_0$'s and $\hat{\beta}_1$'s compare with the previous result?

```
xib <- rep(0, n)
for (i in 1:n) {
    xib[i] = i * 100
}

set.seed(330)
sim <- 10000
bh0b.vec <- rep(0, sim)
```

```
bh1b.vec <- rep(0, sim)
sig2b.vec <- rep(0, sim)

for (i in 1:sim) {
    e <- rnorm(n, mean = 0, sd = sigma)
    y <- beta0 + beta1 * xib + e
    m1b <- lm(y ~ xib)
    bh0b.vec[i] <- coef(m1b)[1]
    bh1b.vec[i] <- coef(m1b)[2]
    sig2b.vec[i] <- sum(e^2)/n - 2
}


mean(bh0b.vec)
```

```
## [1] 9.611829
```

```
mean(bh1b.vec)
```

```
## [1] -2.499909
```

```
sqrt(mean(sig2b.vec))
```

```
## [1] 99.94228
```

```
xbarb <- sum(xib)/n

sxxb <- sum(xib^2) - n * mean(xib)^2

varb0 <- sigma^2 * ((1/n) + (xbarb^2)/sxxb)
varb0
```

```
## [1] 1402.299
```

```
var(bh0b.vec)
```

```
## [1] 1428.127
```

```
varb1 <- sigma^2 * (1/sxxb)
varb1
```

```
## [1] 0.0004449388
```

```
var(bh1b.vec)
```

```
## [1] 0.0004581832
```

The sample variance, and true variance, of β_0 is roughly the same as it was in part e when we had different values for xi. The variance of β_1, on the other hand, is much lower than it was before. This is because of the structre of the two variance expressions. $Var(\beta_1) = \sigma^2 * \frac{1}{SXX}$ has all the x terms on the bottom of a fraction, while β_0 has an $\bar{x}^2$ above SXX. This means that the larger xi values used in part f are reducing the variance of β_1 by a lot, as they are leading to a very small $\frac{1}{SXX}$ term, while the x's are more balanced in the β_0 term.