Least Squares Regression Lab

Mei Maddox

2/22/2022

Setup

This lab illustrates the theoretical properties of linear regression coefficients using the following regression equation:

$$y = 2 + 3X + \epsilon,\ \epsilon \sim N(0, \sigma^2),$$

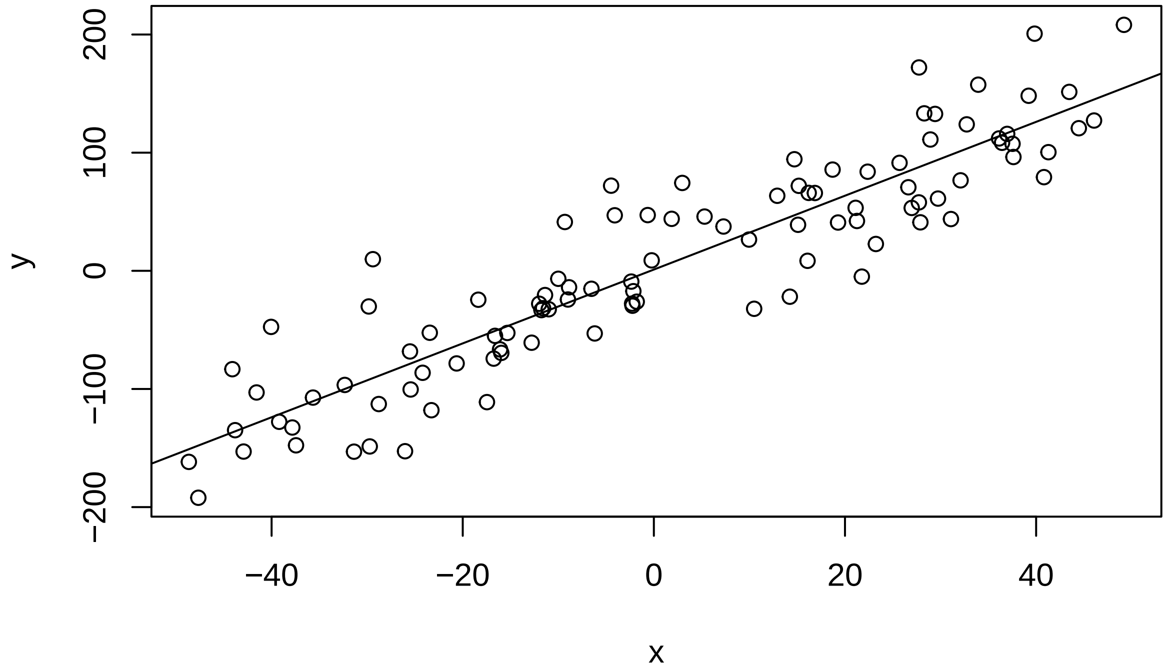
where the fixated X-values are obtained from a uniform distribution from -50 to 50

```
# Establish population constants
beta.0 <- 2
beta.1 <- 3

# Establish epsilon standard deviation
sigma <- 40

## Generate a 100 values for explanatory variable x, uniformly distributed from -50 to 50.
set.seed(1)
x <- runif(100, -50, 50)
```

Single random sample

Because X is fixated, randomness is introduced by ϵ.

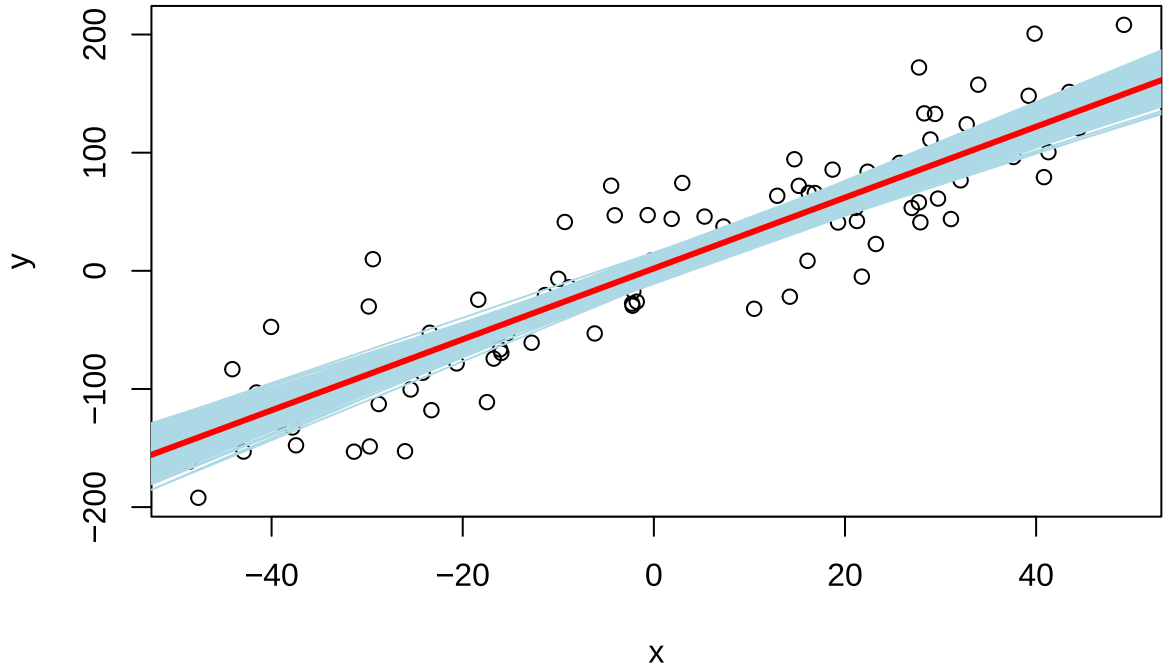
```
## Generate y values from the model
y <- beta.0 + beta.1*x + rnorm(100, 0, sigma)

## Fit least squares regression y ~ x, plot the resulting fit.
plot.default(y~x)
abline( lm(y~x) )
```



Multiple random samples

Conducting 1,000 simulations of the earlier sampling method illustrates how sampling variety can affect the least squares regression line. In the plot below, the red line represents the population regression line.

```r
plot.default(y~x)

beta.est <- NULL
for (i in 1:1000){
  y <- beta.0 + beta.1*x + rnorm(100, 0, sigma)
  lm.obj <- lm(y~x)
  abline( lm.obj, col = 'lightblue' )
  beta.est <- rbind( beta.est, lm.obj$coefficients )
}

## Overlay a thick red population regression line over.
abline( beta.0, beta.1, col = 'red', lwd = 3 )
```



Coeficient Calculations

The collection of coefficients obtained from sampling from the population multiple times can be used to estimate the population coefficients.

```
# Beta_1
mean.1 <- mean(beta.est[,2])
var.theo.1 <- sigma^2 / sum( (x-mean(x))^2 )
var.1 <- var(beta.est[,2])


# Beta_0
mean.0 <- mean(beta.est[,1])
var.theo.0 <- sigma^2 * (1/100 + mean(x)^2/sum( (x-mean(x))^2 ))
var.0 <- var(beta.est[,1])
```

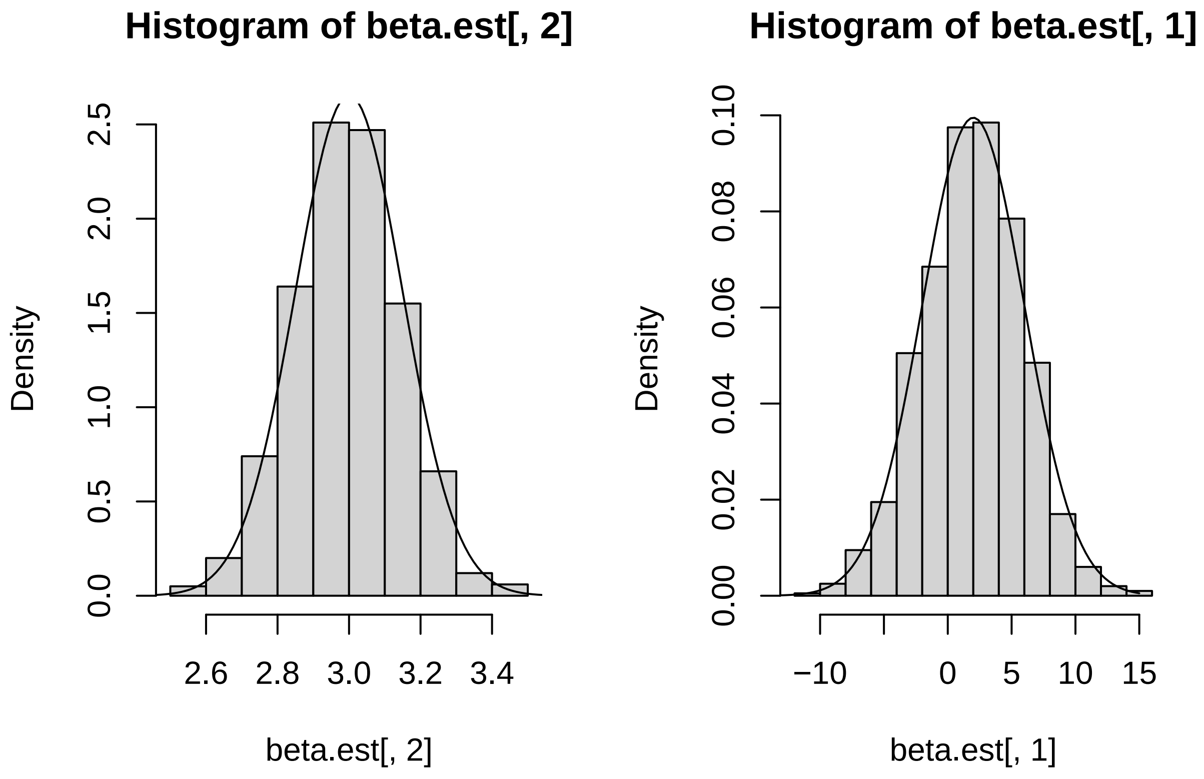
Both coefficients are normally distributed and centered about the respective population parameter ($\hat{\beta}_j \sim N(\beta_j,\ V[\hat{\beta}_j]),\ j = 0, 1$). The following two histograms illustrate this property.

```
par(mfrow=c(1,2))

hist( beta.est[,2], freq = F )
curve( dnorm(x, beta.1, sqrt(var.theo.1)), from=2.4, to=3.6, add = T )
```

Table 1: y = 2 + 3X + ϵ

	Theoretical		
	Equation	Value	Practical
Expected Value (Mean)			
$E[\hat{\beta}_1]$	β_1	3.0000000	2.9940515
$E[\hat{\beta}_0]$	β_0	2.0000000	1.9224617
Variance			
$V[\hat{\beta}_1]$	$\frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$	0.0225716	0.0225886
$V[\hat{\beta}_0]$	$\sigma^2 \times (\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2})$	16.0718945	15.5101762

```
hist( beta.est[,1], freq = F )
curve( dnorm(x, beta.0, sqrt(var.theo.0)), from=-15, to=15, add = T )
```


Histogram of beta.est[, 2]
Density
beta.est[, 2]
Histogram of beta.est[, 1]
Density
beta.est[, 1]