Lab 2: Simulating SLR

STAT 632, Spring 2020

In this lab we consider simulating data from the the following simple linear regression (SLR) model:

$$Y = \beta_0 + \beta_1 x + e = 2 + 3x + e, \text{ where } e \sim N(0, 25),$$

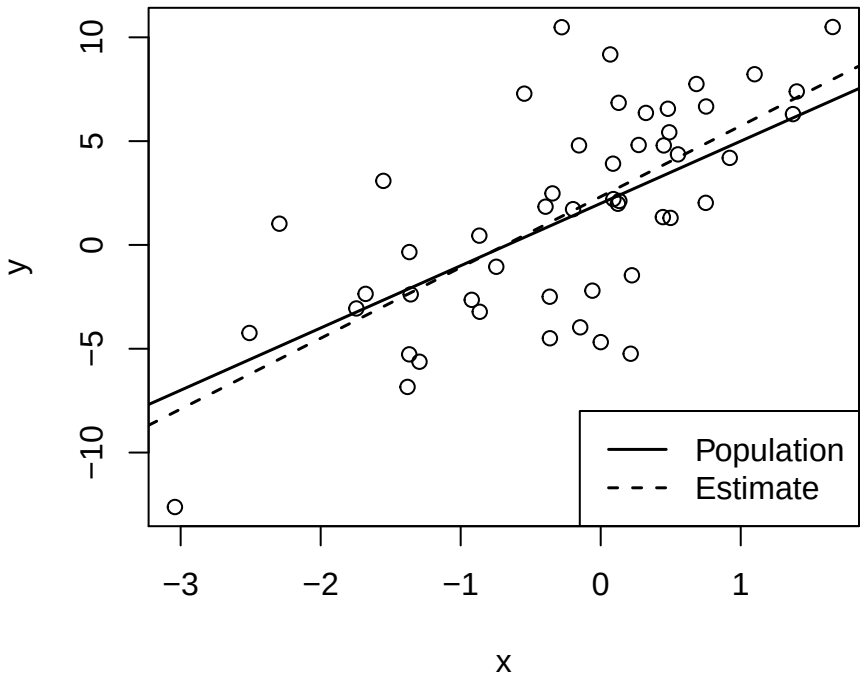
that is $Var(e) = \sigma^2 = 25$. This is the equation for the population regression line, which, in practice, is unknown. For the simulation, we can generate data from this model, and then obtain estimates of the parameters using the least squares method. This will enable us to investigate and gain insight into properties of the SLR model.

```
n <- 50 # sample size
beta0 <- 2 # population intercept
beta1 <- 3 # population slope
sigma <- 5

set.seed(99)
x <- rnorm(n)
e <- rnorm(n, mean=0, sd=sigma)
y <- beta0 + beta1 * x + e

# estimate SLR model from simulated data
lm1 <- lm(y ~ x)
summary(lm1)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.296 -2.693  0.224  1.991  9.112
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3255     0.5447   4.269 9.21e-05 ***
## x             3.4105     0.5234   6.515 4.07e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.737 on 48 degrees of freedom
```

```
## Multiple R-squared:  0.4693,Adjusted R-squared:  0.4583
## F-statistic: 42.45 on 1 and 48 DF,  p-value: 4.072e-08
```

The scatterplot below shows the population regression line (solid) and the least squares estimate (dashed).

```
par(mar=c(4.5,4.5,2,2)) #adjust margins
plot(y ~ x)
abline(beta0, beta1, lwd=1.5, lty=1) # population regression line
abline(lm1, lwd=1.5, lty=2) # least squares estimate
legend('bottomright', lwd=1.5, lty=c(1, 2), c('Population', 'Estimate'))
```



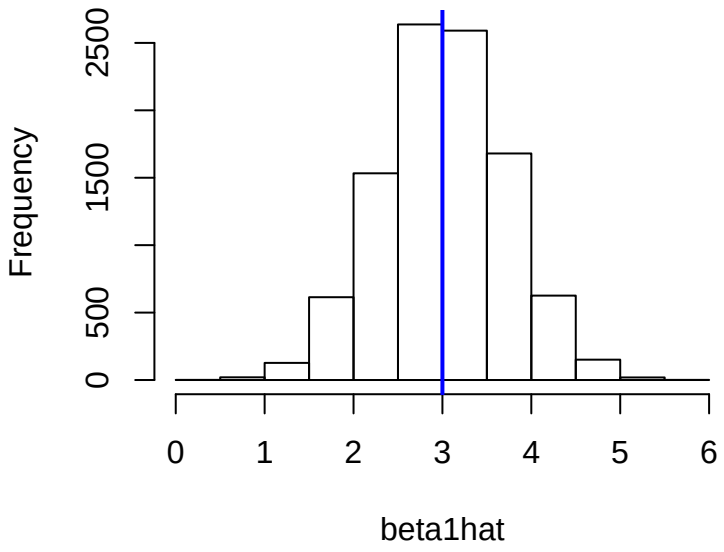
Your turn: Generate another simulated data set of size $n = 50$ from the SLR model $Y = 2 + 3x + e$, where $e \sim N(0, 25)$. Make a scatterplot with your simulated data, and add the least squares line and population regression line.

We can repeatedly simulate data from the SLR model, $Y = 2 + 3x + e$. Each simulated data set will give a slightly different least squares estimate of the slope and intercept of the line. The `for loop` below generates 10,000 simulated data sets from the SLR model. The least squares regression line is estimated from each data set. This gives 10,000 estimates of the population regression line; that is, 10,000 least squares estimates of the slope and intercept. Note that the values of the explanatory variable x are considered fixed in SLR, so the x values are only specified once (before running the loop).

```
set.seed(99)
beta1hat <- rep(0, 10000) # initialize vector of slope estimates
x <- rnorm(n) # simulate x values
for(i in 1:10000) {
  e <- rnorm(n, mean=0, sd=sigma)
  y <- beta0 + beta1 * x + e
  lm_i <- lm(y ~ x)
  beta1hat[i] <- as.numeric(coef(lm_i)[2])
}
```

Below is a histogram of the 10,000 least squares estimates of the slope (i.e., the simulated sampling distribution). Note that the estimates of the slope are centered around the population slope, $\beta_1 = 3$, and normally distributed.

```
par(mar=c(4,4,1,1)) #adjust margins
hist(beta1hat, main='')
abline(v=3, col='blue', lwd=2)
```



The variance of the 10,000 estimates of the slope is also approximately equal to true variance given by the formula $Var(\hat{\beta}_1) = \sigma^2/SXX$ where $SXX = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

```
# variance of the slope estimates
var(beta1hat)

## [1] 0.487108

# true (analytic) variance
SXX <- sum((x - mean(x))^2)
sigma^2 / SXX

## [1] 0.4905309
```