Homework 1

Greg DePaul

2023-03-28

Problem 7 - Simulation by R

You need to submit your codes alongside with the answers, plots, outputs, etc. For this homework, you are encouraged (though not required) to use R Markdown: Please submit a .rmd file and its corresponding .html file. Later in the quarter, you will be required to use R Markdown for some homework or quiz problems. (Hint: Use the help function if needed)

```
x <- seq(1, 100)
```

(a) Create a sequence of consecutive integers ranging from 1 to 100. Record these in a vector x. (Hint: use the seq function)

```
w <- 2 + 0.5 * x
```

(b) Create a new vector w by the formula: w = 2 + 0.5 ∗x.

```
Y <- rnorm(100, mean = 0, sd = 5)
```

```
mean(Y)
```

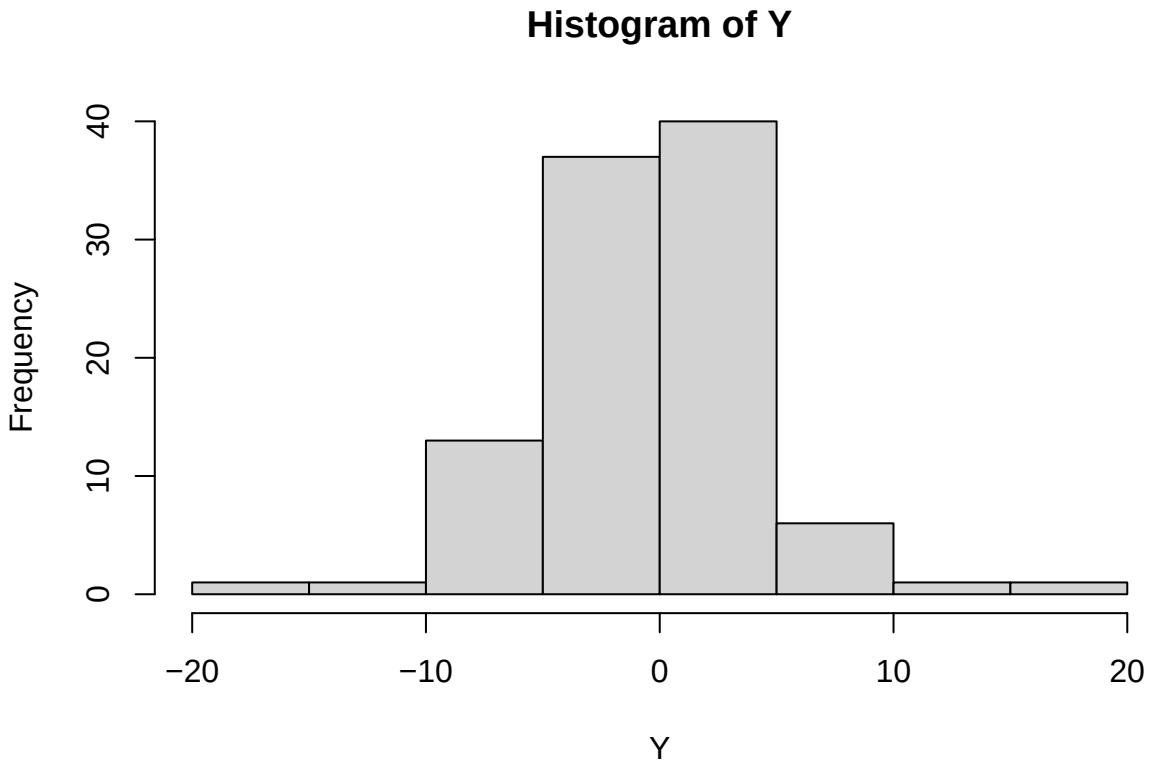
(c) Randomly sample 100 numbers from a Normal distribution with mean zero and standard deviation 5. Calculate the sample mean and sample variance and draw a histogram. What do you observe?

```
## [1] -0.5420766
```

```
var(Y)
```

```
## [1] 24.12202
```

```
hist(Y)
```



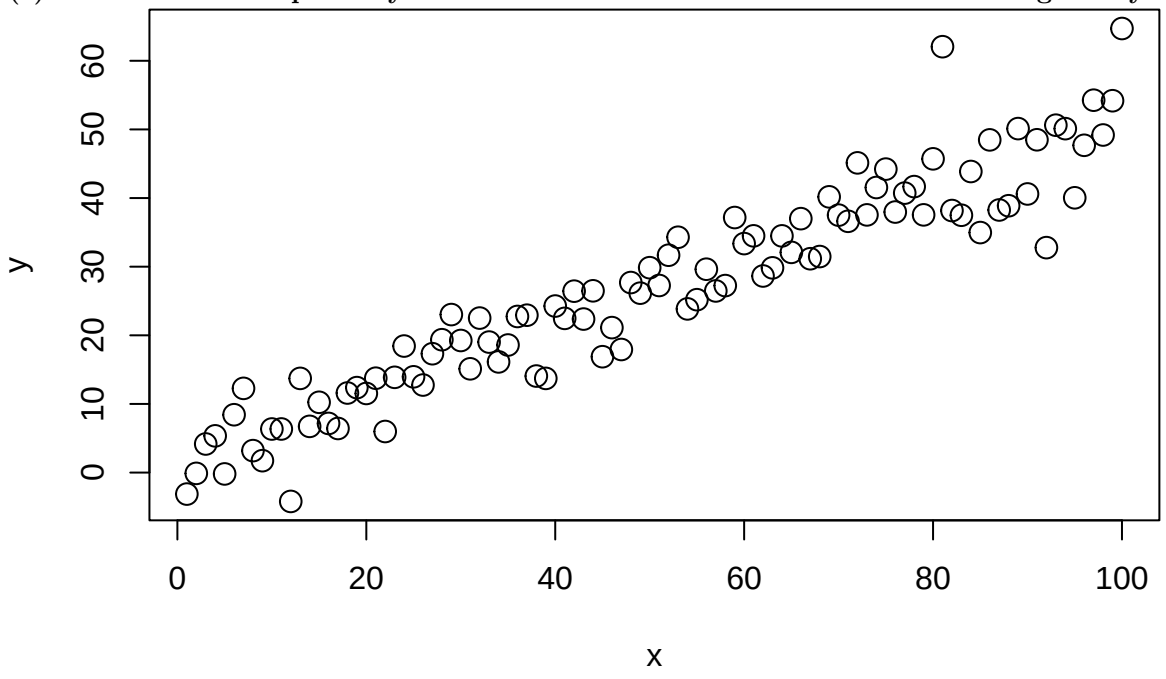
We see that the sample mean and sample variance are very different from the null distribution mean and variance.

```
y <- Y + w
```

(d) Add (element-wise) the numbers created in part (c) to the vector w. Record the new vector as y.

```
plot(x,y, cex=1.5)
```

(e) Draw the scatter plot of y versus x. Make the axes labels 1.5 times as large as by default.

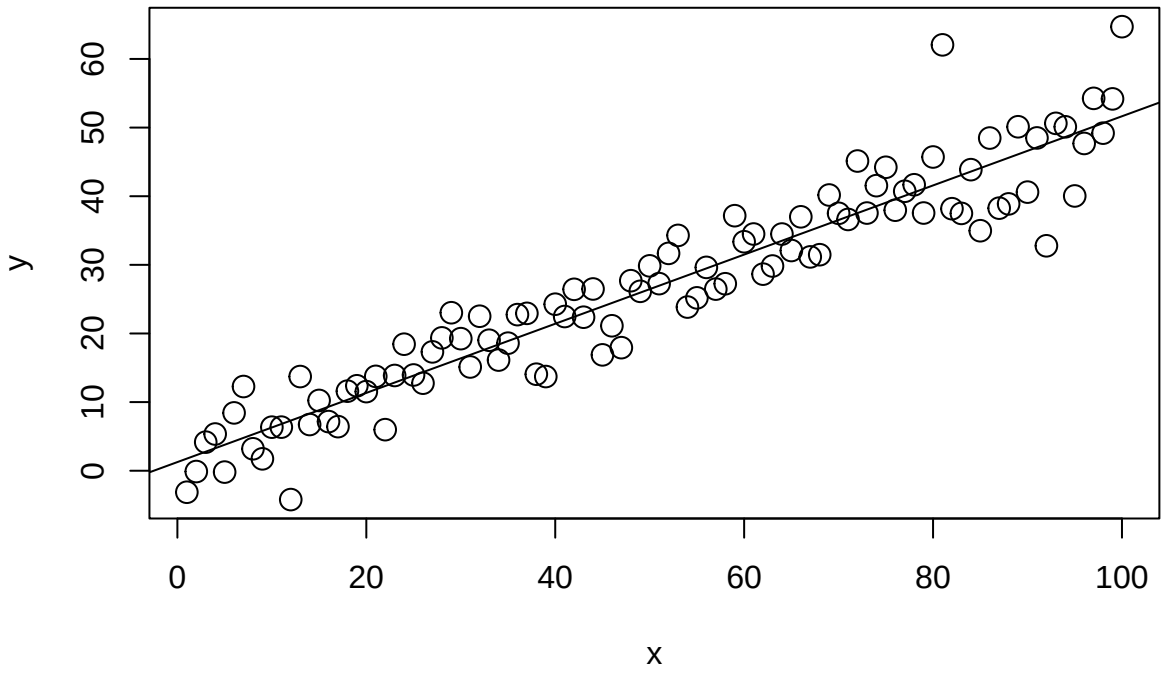



```
regression_data <- data.frame(x, y)
fit = lm(y~x, data=regression_data)
fit$coefficients
```

(f) Estimate the regression coefficients of y on x. Add the fitted regression line to the scatter plot in part (e). What do you observe?

```
## (Intercept)           x
##   1.2611662   0.5038962
```

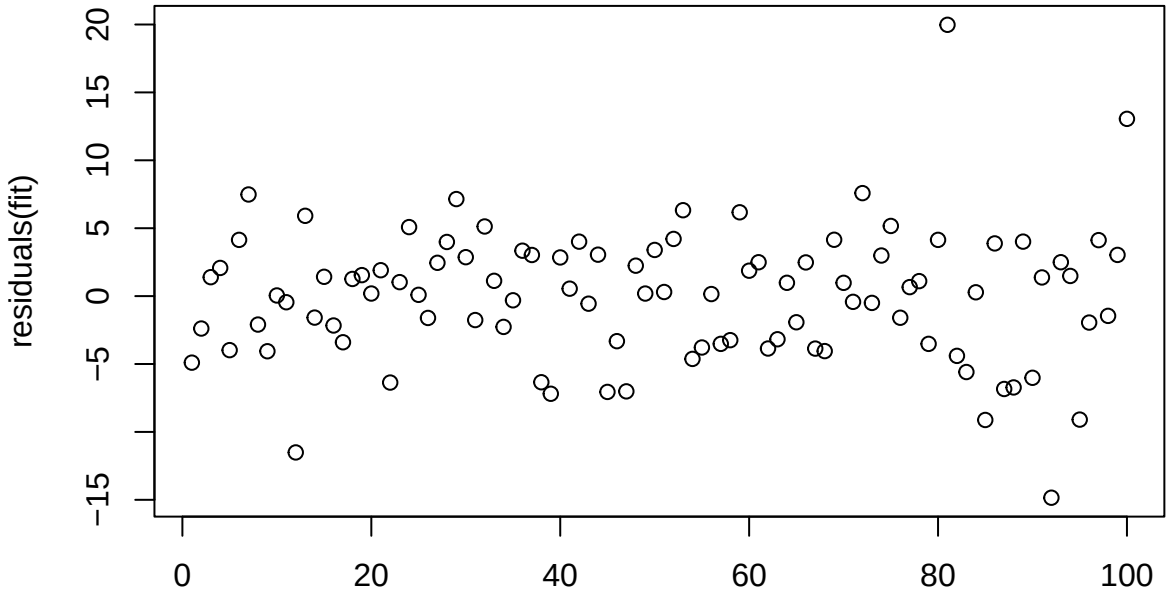
```
plot(x,y, cex=1.5)
abline(lm(y~x))
```



We see that the line general models the trend of the data.

```
plot(x,residuals(fit))
```

(g) Calculate the residuals and draw a scatter plot of residuals versus x. What do you observe?



Derive MSE.

It looks like a normal distribution with variance of 5 getting sampled over time. Calculating the mean squared error:

```
mean(fit$residuals^2)
```

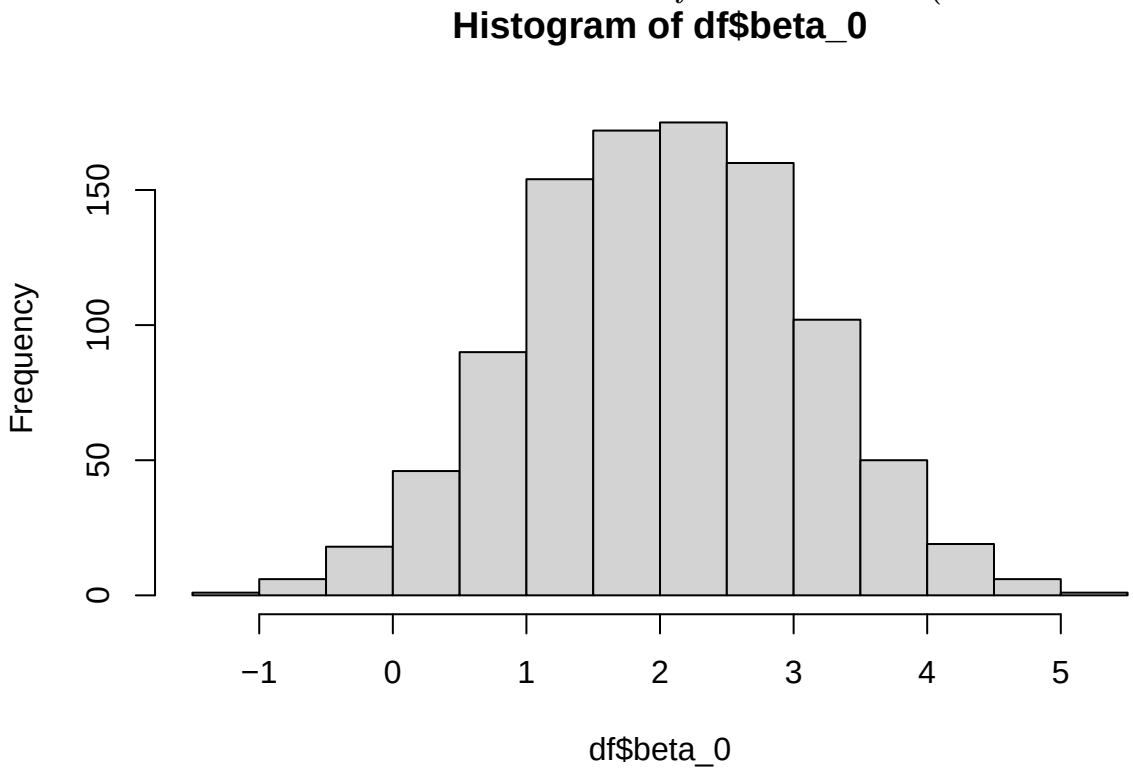
```
## [1] 23.86815
```

(h) Repeat parts (c) – (g) a couple of times. What do you observe? The numbers will change with every run of the markdown file. But the results are very consistent.

```
df <- data.frame(matrix(ncol = 3, nrow = 0))
colnames(df) <-c("beta_0", "beta_1", "mse")
for (i in 1:1000) {
  Y <- rnorm(100, mean = 0, sd = 5)
  y <- Y + w
  regression_data <- data.frame(x, y)
  fit = lm(y~x, data=regression_data)
  betas <- fit$coefficients
  mse <- mean(fit$residuals^2)
  df[i,] <- c(betas[1],betas[2],mse)
}
```

```
hist(df$beta_0)
```

(i) (Optional Problem) Repeat parts (c) – (d) 1000 times. Each time, derive the fitted regression coefficients and MSE and record them. Draw histogram and calculate sample mean and sample variance for each of the three estimators. Summarize your observations. (Hint: use the for loop)



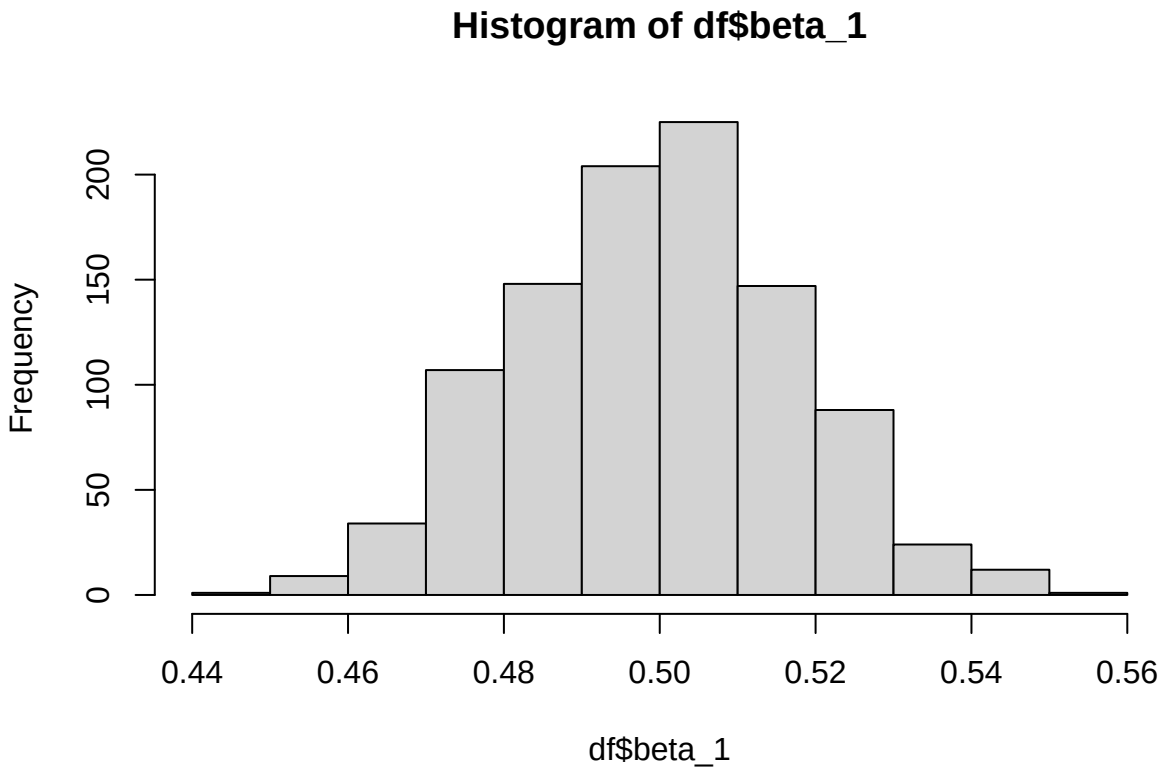

```
mean(df$beta_0)
```

```
## [1] 2.028066
```

```
var(df$beta_0)
```

```
## [1] 1.069907
```

```
hist(df$beta_1)
```



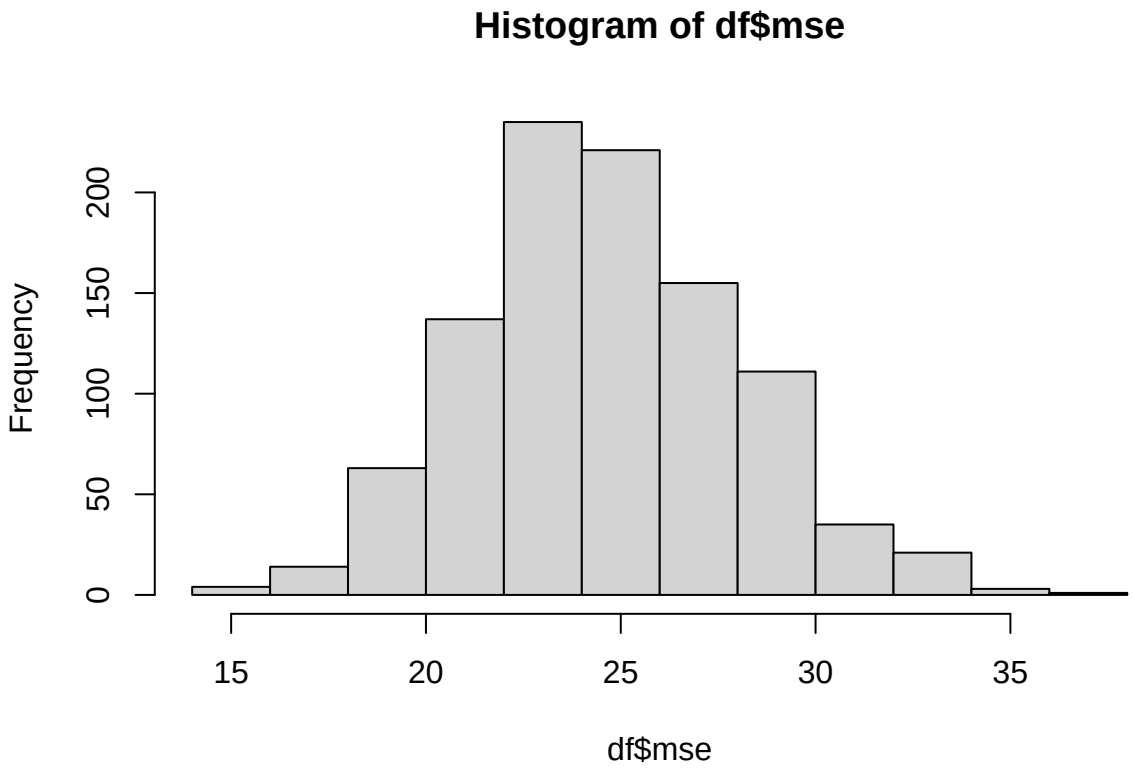

```
mean(df$beta_1)
```

```
## [1] 0.499388
```

```
var(df$beta_1)
```

```
## [1] 0.0003088876
```

```
hist(df$mse)
```




```
mean(df$mse)
```

```
## [1] 24.66128
```

```
var(df$mse)
```

```
## [1] 11.5559
```

Depending on the sample size, the estimators can vary greatly between separate runs.