Homework 4

Please submit the solution in the form of R Markdown report, knitted into either of the available formats (HTML, pdf or Word). Provide all relevant code and output. Goal of this homework is to have you familiarized with 1) modeling assumptions of linear regression; 2) properties of least squares estimates; 3) confidence intervals; and 4) practice your R coding.

Problem #1

Show that $Y_i = \beta_0 + \epsilon_i,\ \epsilon_i \sim_{ind} N(0, \sigma^2)$ leads to $Y_i \sim_{ind} N(\beta_0, \sigma^2)$.

More precisely, make sure to derive:

A. Formula for $E[Y_i]$. Explain what it means in plain English
B. Formula for $V[Y_i]$. In plain English, explain what $V[Y_i]$ describes
C. Normality of Y_i.

No need to show independence ("ind")

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

We know that β_0 and β_1 are constants; X is fixed, and therefore can be treated similarly to $\beta_0,\ \beta_1$ in terms of mathematical properties. Therefore, $\beta_0 + \beta_1 X_i$ can be treated like a constant in all the following derivations. However, the given equation leads me to assume $\beta_1 = 0$ OR that $X_i = 0$.

A. $E[Y_i] = E[\beta_0 + \beta_1 X_i + \epsilon_i]$
$E[\beta_0 + \beta_1 X_i] + E[\epsilon_i] \iff E[constant] + E[\epsilon_i]$
The expected value of a constant, is just the constant.
The expected value of epsilon is 0 by the definition above.
β_1 or X_i is 0 by the assumption from above.
$\beta_0 + \beta_1 X_i + 0 \implies \beta_0 + 0 + 0$
$E[Y_i] = \beta_0$:The expected value for any observation being drawn from the population is centered at the y-intercept value β_0.

B. $V[Y_i] = V[\beta_0 + \beta_1 X_i + \epsilon_i]$
$V[constant + \epsilon_i]$
constant only shifts center; doesn't affect variability
$V[\epsilon_i]$
The variance of epsilon is σ^2 by the definition above.
$V[Y_i] = \sigma^2$:The variance for any observation being drawn from the population is σ^2.

C. Normality of Y_i.
ϵ_i is normally distributed by the definition above.
$\beta_0 + \beta_1 X_i$ simply shifts the ϵ_i distribution by the constant value.

Problem #2

Finish the lab, compiling it into a nice R markdown report.
(see attached Least Squares Regression Lab .Rmd and .pdf files)

Problem #3

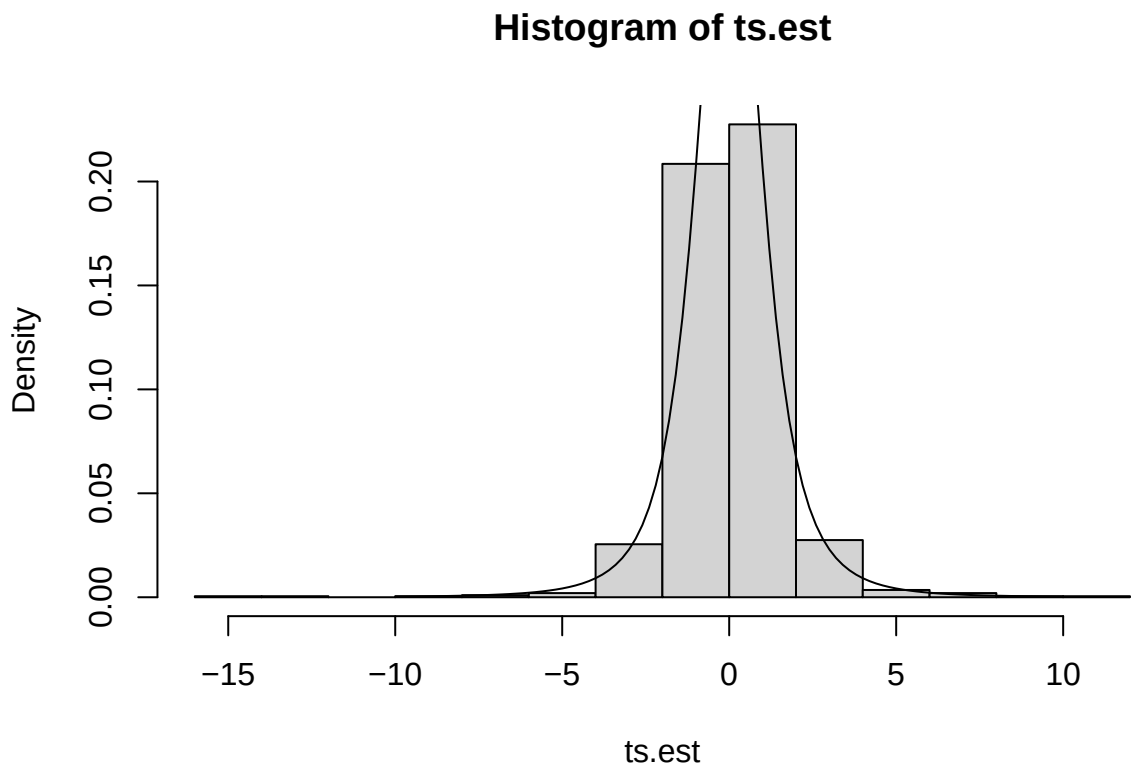
Verify that, for simple linear regression $Y = \beta_0 + \beta_1 X + \epsilon$, we have

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

```
set.seed(1)

n <- 5
X <- runif(n, -50, 50)
beta.0 <- 2
beta.1 <- 3
n.rep <- 1000

ts.est <- sapply(1:n.rep, function(v){
  Y <- beta.0 + beta.1*X + rnorm(n, 0, 10)
  lm.obj <- lm(Y~X)
  (lm.obj$coefficients["X"] - beta.1) / summary(lm.obj)$coefficients["X", 2]
})
```



TS distribution is centered at 0 (mean=0.0651359$\approx$ 0) and appears to approximately follow the respective t-distribution.

Problem #4

Calculate the % of times (out of 1000 generated confidence intervals) that the true population values $\beta_0 = 2$ and $\beta_1 = 3$ ended up within their respective confidence intervals. Are those %'es equal to what we expected? Why? Hint: Recall the practical interpretation of a 90% confidence interval.

```
set.seed(2)

n <- 200
X <- runif(n, -50, 50)
n.rep <- 1000

conf_int_b0 <- matrix(0, nrow=n.rep, ncol=2)
conf_int_b1 <- matrix(0, nrow=n.rep, ncol=2)
for (r in 1:n.rep){
  Y <- beta.0 + beta.1*X + rnorm(n, 0, 10)
  ci <- confint(lm(Y~X), level = 0.9)
  conf_int_b0[r,] <- ci[1,]
  conf_int_b1[r,] <- ci[2,]
}
```

Because we are calculating 90% confidence intervals, we would expect to see approximately 90% coverage - 90% of the confidence intervals contain the true population parameter. The β_0 coverage (0.883) and the β_1 coverage (0.893) are both close to 0.9 which is what we would expect from a 90% confidence level