HW4.

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").

Problem #2

Finish the lab, compiling it into a nice *R markdown* report.

Problem #3

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

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

by generating a vector $X = (X_1, X_2, \ldots, X_5) \sim Uniform(-50, 50)$, and subsequently running a 1000 replicates of the following:

- Generate response values $Y_1, Y_2, \ldots, Y_5$ from $Y_i = 2 + 3X_i + \epsilon_i,\ \epsilon_i \sim N(0, 10^2)$ relationship;

- Calculate $TS = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$, **record the value**,

and subsequently checking the "practical" distribution of $1,000$ resulting TS values. Specifically, proceed to check whether this distribution has

- Mean 0,
- Shape of an appropriate t distribution (make sure to overlay t-density over your practical sampling distribution).

Notes:

- Make sure to use $set.seed()$ for consistency.
- Google how to extract standard error from $lm()$ object.

Problem #4

Provide the code that will generate a vector $X = (X_1, X_2, \ldots, X_{200}) \sim Uniform(-50, 50)$, and subsequently conduct a 1000 replicates of the following:

- Generate a sample $n = 200$ response values $Y_1, Y_2, \ldots, Y_{200}$ from $Y_i = 2 + 3X_i + \epsilon_i,\ \epsilon_i \sim N(0, 10^2)$ relationship;
- Calculate 90% confidence intervals resulting from $\hat{\beta}_0$ and $\hat{\beta}_1$ (just use $confint()$), **keep track of them**.

and afterwards 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.

Code to get you started:

```
set.seed(2)

n <- 200
X <- runif(n, mean=0, sd=1)


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){
  # .... That's where you generate your data,
  #      fit the linear model,
  #      calculate conf intervals for b0,b1, and save them into conf_int objects
}

# Here is where you calculate the %'es.
print(mean(conf_int_b0[,1] < 2 & 2 < conf_int_b0[,2]))
# ...
```