Simulations HW 4

Let Y = "% final grade in Intro to Stats (0 to 100)" at a certain university. Suppose Y is associated with X = "grade on diagnostic test for this course (0 to 40)" for the entire population of students who take this class as follows:

$Y = 5 + 2X + \varepsilon$ where ε is normal with $\mu = 0$ and $\sigma = 3$ (thus $\sigma^2 = 9$)

Suppose a researcher cannon access this info, and therefore, he/she decides to use linear regression to estimate the parameters of this model. Suppose the researcher decides to poll 1 student from each subpopulations of students who got the following scores in the diagnostic test: 20, 25, 30, 35 and 40 (that is decides to get Y for $n = 5$ students). Simulate the samples that 1000 researchers could get as follows:

1. Create a 1000 x 5 matrix whose first column are random numbers from a normal random variable with parameters $\mu = 5 + 2(20)$ and standard deviation $\sigma = 3$; whose second column are random numbers from a normal random variable with parameters $\mu = 5 + 2(25)$ and standard deviation $\sigma = 3$, etc.

2. Create functions to obtain $\hat{\beta}_0$ and $\hat{\beta}_1$ (formulas on p. 416 of the textbook), and $\hat{\sigma}^2$ = MSE= SSE/(n-2) (formula for SSE on p. 418).

3. Create a matrix 1000 x 3 matrix that has as columns $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$.

a. Obtain the histograms for your $\hat{\beta}_0$'s and $\hat{\beta}_1$'s. Is the shape of the histogram an indication that $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed? Why?

b. Obtain the mean values of your $\hat{\beta}_0$'s, $\hat{\beta}_1$'s and $\hat{\sigma}^2$'s. Compare these means with the true β_0, β_1, and σ , and articulate sentences on whether this empirically shows that $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$ are unbiased estimators of β_0 and β_1 and σ, respectively.

4. a If you constructed (don't do it for this homework assignment!) a 1000 x 2 matrix with the following columns, $\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2}}$ and $\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)MSE}}$, what would the histograms of this columns indicate? That is, what the difference be between these histograms? Hint: state the theoretical distribution of the population of these values.

b. a If you constructed (don't do it for this homework assignment!) a 1000 x 2 matrix with the following columns, $\frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}$ and $\frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}$, what would the histograms of this columns indicate? That is, what the difference be between these histograms? Hint: state the theoretical distribution of the population of these values.

c. Construct a 1000 x 4 matrix (or two 1000 x 2 matrices) containing the lower and upper bounds of 95% confidence intervals for β_0 and β_1 (formulas on p. 426).

d. What proportion of your intervals contain the true (population) β_0?

e. What proportion of your intervals contain the true (population) β_1?