The models we build are only approximations to the truth (if there is any absolute truth). Every model building procedure starts off by making some assumptions about the unknown population which we cannot verify in practice. Thus, it is impossible to evaluate the properties of the statistical procedures based on real life examples. Of course when we have the freedom of designing a study for collecting data we try our best to ensure that the data satisfy some of the assumptions. The best way to evaluate the properties of a statistical method is to get analytical results about well defined mathematical quantities that can be used to judge and compare procedures. Unfortunately, analytical derivation of theoretical properties are not always possible. In such situations one can rely on laws of probability and simulated experiments. A critical part of a graduate program nowadays is to learn how to simulate experiments and how to draw conclusions from them. The properties of the least squares estimators of the parameters of a simple linear regression model can be derived analytically (completely if one is ready to assume that the errors are normal). The following is a simulation experiment that will give you some idea about the finite sample properties of the least squares estimators and how the different assumptions affect the properties.

[MONTE CARLO EXPERIMENT]:

The emu (Dromaius novaehollandiae), is the largest bird native to Australia and, after the ostrich, the second-largest bird that survives today. The soft-feathered, brown birds reach 1.5 to 2 m in height and weigh up to 60 kg, with the male marginally smaller. The following observations are heights (rounded to nearest 0.05 of a meter) of 10 female and 10 male emus.

Female: 1.65 1.95 1.85 1.70 1.75 1.60 1.90 1.95 1.80 1.65
Male: 1.55 1.65 1.60 1.70 1.65 1.80 1.60 1.55 1.85 1.70

Suppose a simple linear regression model:

\[ y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, \ldots, 20, \] (1)

is used to model the heights of males and females where \( x_i \) is 0 for male and 1 for female. Suppose from years of experience the scientists have figured out the population mean values, 1.65 m for male and 1.8 m for female and suppose the distribution of heights is normal around the mean with a standard deviation of 0.1 m. Thus the population model for the errors is: \( e_i \) are independent and identically distributed as normal with mean zero and standard deviation, \( \sigma = 0.1 \).

Estimate the least squares estimators of \( \beta_0, \beta_1 \) and \( \sigma^2 \) from the given data and test whether there is enough evidence in the data to reject the hypothesis that the mean height of female emus is \( \mu_F = 1.8 \).

Now suppose we want to simulate heights of 10 female and 10 male Emus given the population values of the parameters. Generate 1000 replicates of the error vector \( \epsilon = \)
Let’s call them $\epsilon_1, \ldots, \epsilon_{1000}$. Now simulate 1000 replicates of the observation vector $Y = (y_1, y_2, \ldots, y_{20})'$ from model (1) as

$$Y_j = X \beta + \epsilon_j, \quad j = 1, 2, \ldots, 1000,$$

where $X$ is $20 \times 2$ with a column of ones and the second column with 0’s for the first 10 observations and 1’s for the next 10 and $\beta = (\beta_0, \beta_1)'$. We now have 1000 samples of $Y$’s. For each sample compute:

1. The least squares estimate of $\beta$:

$$\hat{\beta}_j = (\hat{\beta}_{0,j}, \hat{\beta}_{1,j})' = (X'X)^{-1}X'Y_j.$$

2. The regression sum of squares $SS_{reg,j} = \hat{Y}_j'\hat{Y}_j$ where

$$\hat{Y}_j = X\hat{\beta}_j.$$

3. The residual sum of squares $SS_{res,j} = Y_j'Y_j - \hat{Y}_j'\hat{Y}_j$.

4. Estimate of $\sigma^2$

$$s^2_j = SS_{res,j}/(20 - 2).$$

Compute the Monte Carlo expectations of the parameter estimates:

$$\hat{E}(\hat{\beta}_k) = \left(\sum_{j=1}^{1000} \hat{\beta}_{k,j}\right)/(1000), \quad k = 0, 1.$$

$$\hat{E}(\hat{\sigma}^2) = \left(\sum_{j=1}^{1000} s^2_j\right)/(1000).$$

**How do they compare with the theoretical quantities?**

Compute the Monte Carlo variance covariance matrix of the least squares estimator of $\beta$.

$$\hat{V}(\hat{\beta}) = ((v_{lk}))_{l=0,1 \ k=0,1}$$

where

$$v_{lk} = \hat{Cov}(\hat{\beta}_l, \hat{\beta}_k) = \left[\sum_{i=1}^{1000} \sum_{j=1}^{1000} \{\hat{\beta}_{l,i} - \hat{E}(\hat{\beta}_l)\}\{\hat{\beta}_{k,j} - \hat{E}(\hat{\beta}_k)\}\right]/(999).$$

**How do they compare with the theoretical quantities?**

Compute the correlation coefficients between the sample of $s^2_j$ and the samples of $\hat{\beta}_{k,j}, \ k = 0, 1$.

**How do they compare with the theoretical quantities?**
Let $C = (X'X)^{-1} = ((c_{lk}))_{l=0,1, k=0,1}$.

Construct histograms of the samples of $(\hat{\beta}_{k,j} - \beta_k)/(2\sqrt{c_{kk}})$, $k = 0, 1$.

**How do the distributions look?**

Construct histograms of the samples of $(\hat{\beta}_{k,j} - \beta_k)/(s_j \sqrt{c_{kk}})$, $k = 0, 1$.

**How do the distributions look?**

Construct histograms of the sample of $(s_j^2)/(4)$.

**How do the distributions look?**

Now suppose we want to simulate the situation the all 20 Emus are siblings and hence their heights are correlated with a constant correlation of $\rho$ between the heights of any two Emus in the groups of 20. Suppose all other facts about the population remain the same. Then in order to simulate from the population we need to generate the error vector as mean zero but with covariance matrix

$$
\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\
\rho & 1 & \cdots & \rho \\
\vdots & \vdots & \ddots & \vdots \\
\rho & \cdots & \cdots & 1
\end{pmatrix}
$$

Repeat the simulation experiment with $\rho = -0.1, 0.6, 0.99$ and summarize your findings.

Do problems 2.1, 2.2, 2.3 from the book.

*DO NOT GIVE ME UNNECESSARY OUTPUT. WRITE A PROFESSIONAL REPORT*