Senior Thesis
R Assignment 3

Background:
In the second assignment we began generating linear models. We need to take this one step further and look at some characteristics of the resulting estimates from fits of a linear model. Specifically, we will consider the following model:

$$y_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$$

where y_i is the response for the i-th subject, $\mathbf{x}_i$ is a 5×1 vector containing the values of the covariates for the i-th subject, $\boldsymbol{\beta}$ is the vector of the 5 unknown coefficients corresponding to the covariates, β_0 is the unknown intercept, and ϵ_i is an error term. In generation of this model, the following is true:

- $\beta_0 = 0$ and $\boldsymbol{\beta}^\top = (4, 2, 1, 0.5, 0)$.
- $x_{i,j} \overset{IID}{\sim} N(0, 1)$
- $\epsilon_i \overset{IID}{\sim} N(0, 3)$ and is independent of the covariates.

From this discussion it should be clear that the variables are presented in order of decreasing importance in the model (variable 1 is more important than variable 2, and so on). We have discussed that there are several ways to order predictors; here we will consider two methods:

- Fit a full least squares model, and order the predictors by the magnitude of the estimated coefficients.
- Order the variables by the magnitude of the correlation of each with the response (which is the same as ordering by the magnitude of the marginal estimates if each variable was put in a simple linear regression model for the response); this is the idea behind SIS.

We want to compare these two methods in the above setting via the C-index (also known as the Mann-Whitney U-statistic).

Assignment:
You have been provided with an R script that outlines my general approach to this problem, but nearly all code has been removed.

1. For the first problem, write an Rfunction that takes three parameters — a vector y (the response), a matrix X of the covariates — and returns an ordering of the variables. The third parameter of the function dictates whether that order is based on the least squares fit or based on the marginal correlations. In particular, your code should have a form like:

   ```
   fct.name <- function(y,X,method){ some code here }
   ```

   You might find the following Rfunctions helpful: `order`, `cor`, and `lm`.

2. Generate $m = 1000$ replicate datasets, each with $n = 15$ observations involving a response vector (determined by the above mentioned model) and 5 covariate vectors (which can be stored as a matrix $\mathbf{X}$). For each replicate dataset, do the following:

   - Fit the model using least squares and record the ordering.
   - Determine the marginal correlations and record the ordering.

3. For the orderings generated above, compute Kendall's τ between the ordering and the order of the true coefficients.

4. Make a boxplot of the τ values for each of the two methods. Does one method appear to give a better ranking for this problem?