MATH 282B – Homework 1
Due Monday, 01/18/2016, by 11:59 PM

Send your code to `math282ucsd@gmail.com` *with the exact subject line "MATH 282B (HW 1)" and nothing else in the body. There should only be one file attached, named* `hw1-lastname-firstname.R`. *Make sure your code is clean, commented and running. Keep your code simple, using packages only if really necessary. If your code does not run, include an explanation of what is going on.*

Problem 1. (Verifying the normal theory when the model is correct) The theory says that, under the standard assumptions, the distribution of the least squares estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ are known. Let's conduct some simulations to verify the theory. Consider the model $y = -1 + 3x + \varepsilon$ where x is one-dimensional and $\varepsilon \sim \mathcal{N}(0, 1)$. (The numbers are more or less arbitrary.)

Do the following for $n \in \{10, 100, 1000\}$, perform the following simulation:

- Generate $x_1, \dots, x_n$ IID with distribution $\mathcal{N}(0, 2)$. Keep them fixed.
- Set $B = 999$ (number of repeats). Repeat the following B times.
  - For $i = 1, \dots, n$, generate $y_i = -1 + 3x_i + \varepsilon_i$ as in the model described above.
  - Fit a linear model by least squares and record $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)$ and $\widehat{\sigma}^2$.

Each time, use the simulation results as follows:

A. Plot the B variance estimates as a histogram. Overlay the distribution they are supposed to follow according to the theory.

B. Repeat with the B intercept estimates.

C. Repeat with the B slope estimates.

NotGraded Plot these B coefficient vectors (slope versus intercept in dimension 2). Assess how well the bivariate normal distribution given by the theory fits the scatterplot. (Be creative.)

Problem 2. (Verifying the normal theory with real data) The standard assumptions can hardly be excepted to hold in practice. The difficulty is that, in practice, we do not know the truth — compare with the setting of Problem 1. Here is one way to assess how reliable the normal theory is in practice. Consider the `04cars` dataset (we focus on complete observations). This could be considered the entire 'population' of cars sold in the US in 2004 and, if this were the case, then it could be hard to justify performing statistical inference. More appropriate would be a descriptive analysis. So let's use that as the entire population. We simulate a situation where we only have access to a sample from the population. We focus on `mpg` and `hp` (see lecture notes).

- Repeat the following 1000 times. Sample at random 50 observations (without replacement). Doing as if this were the available data, perform a regression analysis as in the first set of lecture notes. Build a 95% CI for the slope. Check whether the LS slope coefficient from the entire dataset (which could be considered to be the true coefficient) is in that interval.

A. Return the proportion of times that the CI contains the LS slope coefficient from the entire dataset. Offer some very brief comment.

B. Plot the LS slope coefficients from the simulation as a histogram. What distribution would you expect? (This question is optional and a bit tricky.)