MATH 282B – Homework 2
Due Monday, 02/01/2016, by 11:59 PM

Send your code to math282ucsd@gmail.com*. Follow the following format exactly. For Homework 1, in subject line write "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. (WLS versus OLS) Repeat the following for $n \in \{10^2, 10^3, 10^4, 10^5\}$. Generate $x_1, \dots, x_n$ iid from the uniform distribution on $(0, 1)$. Keep them fixed. Repeat the following $B = 200$ times:

- Generate $y_1, \dots, y_n$ as follows $y_i = 1 + 2x_i + \varepsilon_i$ where $\varepsilon_1, \dots, \varepsilon_n$ are independent normal with mean 0 and respective variance $\mathrm{Var}(\varepsilon_i) = x_i^2/10$.
- Fit a line explaining y as a function of x by ordinary least squares (OLS). Compute the usual 95% CI for the slope and record whether it contains the true slope.
- Do the same, now fitting the model by weighted least squares (WLS) with the correct weights.

Do your simulations show a difference between OLS and WLS? Note that the latter is the MLE in this context.

Problem 2. (Vadermonde matrices are ill-conditioned) We saw in lecture that fitting a polynomial model using the canonical basis for polynomials (meaning $1, x, x^2, x^3, \dots$) leads to the regression matrix being a Vandermonde matrix. It was mentioned in lecture that such a matrix is ill-conditioned even if the sample size is large. Although this may be shown by direct mathematical calculations, we do this with the help of the computer.

Repeat the following for $p \in \{1, \dots, 30\}$. Let $n = 10^5$ (fairly large) and let $x_1, \dots, x_n$ span a regular grid from 0 to 1 (on the interval $[0, 1]$). You can do this with the function `seq`. Then generate regression matrix described above for fitting a polynomial of degree p. Make sure to include an intercept. Each time, compute the condition number of this regression matrix. Finally, plot the condition number as a function of the degree. Offer some brief comments.

Problem 3. (Confidence band for polynomial regression) Write a function that does the following. It takes in data $(x_1, y_1), \dots, (x_n, y_n)$ (both x and y are real-valued) and a degree p. It fits a polynomial of degree p and outputs a scatterplot of y vs x, with the regression function added in red and the 95% confidence band (upper and lower) in blue. (You might need to get some hints in section to do this.) Test your function on `mpg` versus `hp`.

On your own. Learn how to shade the confidence band using the function `polygon`. The package `ggplot2` offers other ways to do that.