2018 USC Marshall Statistics PhD Screening Exam
Part II: Take Home (Due at 6PM, June 27, 2018)

Note: You may NOT consult with anyone about this test. However, if you have any clarifying questions about the exam you may contact Gareth. He will answer emails and will be in the office on June 26.

Question 1

You are going to calculate some tennis probabilities and then perform a simulation analysis to confirm your results. Each game of tennis is won by the first person to win four points (denoted 15, 30, 40 and game). However, a player must win by at least two points, so if the first six points are split evenly (denoted Deuce) then two more points are played. If one person wins both points the game is over, but if they are split we are back to Deuce and two more points are played. This continues until someone wins the game.

a In tennis the person serving in a particular game is generally considered to have the advantage (unfortunately the same does not apply when I am serving!). Let p denote the probability that the server will win any individual point. Suppose that a particular game has reached Deuce (i.e. the two players split the first six points). What is the probability that the server will still win the game? Provide your answer BOTH for general p and for $p = 0.7$.

b Now compute the probability that the server will win a game from the start (i.e. before any points have been scored). Provide your answer BOTH for general p and for $p = 0.7$.

c Perform a simulation analysis where you simulate 1,000 games each for a grid of 100 values of p ranging from 0 to 1 (i.e. a total of 100,000 games) to compute the fraction of games won by the server for a given value of p. Provide a plot with p on the x-axis and the probability of winning the game on the y-axis. Provide two lines. A black solid line for the true probability (from part b), and a red solid line for the estimated probability (from your simulation). In addition provide two dotted lines to indicate 95% confidence intervals on your probability estimate. Provide your R code (it should be compact!). Note, if you wish, you may use your answer from part a to handle games that reach Deuce. My code took a couple of seconds to run all 100,000 games so this should not take long to run.

Question 2

After your tennis success you decide to take a vacation in the South Pacific. Unfortunately, your cruise ship sinks and you end up marooned on an isolated Pacific island with no way to get home. Fortunately, you are well educated and figure out how to construct an emergency transmitter from the parts out of your laptop (naturally you grabbed that as the ship was going down—if only you had hopped on the life raft instead...). There is only one problem. It turns out that in order to correctly operate the transmitter you need to compute the density function for X, a random variable (the story behind why you need this expression is very interesting but would take too long to tell here :)). It turns out that $X|\alpha \sim Beta(\alpha, \beta = 2)$ i.e.

$$f_X(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

and that $\alpha \sim Exp(\lambda = .5)$ i.e.

$$f_\alpha(\alpha) = \lambda e^{-\lambda\alpha}.$$

You need the marginal density of X, $f_X(x)$. However, you rapidly realize that there is no simple closed form solution for this expression. Nevertheless you are not daunted by this difficulty and decide that you can use simulation modeling on your laptop to compute a very accurate estimate for the marginal density. You start programming (it only requires a couple of lines) but then discover, to your horror, that in pulling parts off your laptop to build the transmitter it is no longer operating quite right! For some strange reason the only random number generator that is operating correctly is the `runif()` command in R. Of course even this last challenge does not prove insurmountable. You finish your transmitter and are rescued in no time (unfortunately it is by the Australian navy so you end up stuck in Australia but that is a whole other story!)

Use simulation modeling, with random variables generated using the `runif()` function, to estimate $f_X(x)$ over a grid of 100 x ranging from 0 to 1. Provide a plot of the estimated density function along with 95% confidence intervals.

Question 3

You have just been hired as a statistical consultant by the CEO of Babies R Hard Work. The company sells car seats throughout the US. Babies R Hard Work has grown very rapidly in the last few years and has never performed a careful analysis of the factors that affect their profitability. Your task is to perform a statistical analysis based on the data they provide. The data is available in `Carseats` as part of the `ISLR` library in R. Type `?Carseats` for information on the variables.

Babies R Hard Work have a number of questions they would like answered. Make sure to include any relevant computer output to validate your conclusions. However, points will be deducted for including irrelevant results without any explanation!

a The Model : Produce and describe a linear regression model relating Unit sales to the other variables. Note you may include quadratic, and interaction, terms if they seem appropriate. Some questions to consider answering (this is not an exhaustive list!)

- What is the final model?
- Which variables do not appear to be important?
- Do the relationships appear to be linear?
- What do the estimated coefficients tell you about the various variables?
- What do the coefficients for the categorical variables tell you?
- How good is the fit to the data?
- Are there any problems with the fit i.e. violations of the regression assumptions, multicollinearity problems?

b Advertising Effect : Using the model from 1. is there clear evidence that advertising has a (positive) effect on unit sales? If so at what price levels does it seem to be a profitable operation?[1] In answering this question you should take into account not only the estimated coefficient but also the confidence bound on this value. (Assume that advertising is the only fixed cost and variable costs per unit are \$80). There has been some anecdotal evidence that advertising might have different effects on sales in different regions. For example regions with different income levels, age or years of education may respond differently to advertising. Is there statistical evidence to support this belief? If so give some examples of the effect of advertising for different levels of the other variables.

c Optimal Price to Charge : Babies R Hard Work have generally tried to charge \$10 below their competitor's price. However, this has only been an adhoc rule and has often been violated in practice. Assuming their goal is to maximize profit and they are free to adjust price at any time, give an equation for optimal price as a function of competitor's price, unit cost of production and any other variables that were included in the model from 1. Briefly explain how price should be adjusted for any given change in one of the other variables. Give examples of price and corresponding profit for a few combinations of the variables.

d Go Wild : Play with alternative, more modern, regression methods beyond linear regression. Carefully explain the approaches you test out and how well they work (both in terms of predictive accuracy and interpretability). Provide results for the best fit that you can find to the data.

[1] Note you may find it easier to answer this part of the question using a model with no interaction terms.