STA 302f15 Assignment Ten[1]

Problem 3 uses R. Please bring your printout for Problem 3 to the quiz. **Do not write anything on the printout in advance of the quiz, except possibly your name and student number.** The other questions are preparation for the quiz, and are not to be handed in.

1. Based on the general linear model with normal error terms,
   (a) Prove the t distribution given on the formula sheet for a new observation y_0. Use earlier material on the formula sheet. For example, how do you know numerator and denominator are independent?
   (b) Derive the $(1-\alpha) \times 100\%$ prediction interval for a new observation from this population, in which the independent variable values are given in $\mathbf{x}_0$. "Derive" means show the High School algebra.

2. A forestry company has developed a regression equation for predicting the amount of useable wood that they will get from a tree, based on a set of measurements that can be taken without cutting the tree down. They are convinced that a model with normal error terms is right. They have $\widehat{\boldsymbol{\beta}}$ and MSE based on a set of n trees they measured first and then cut down, and they know how to calculate a predicted y and a prediction interval for the amount of wood they will get from a single tree.

   But that's not what they want. They have a set of m more trees they are planning to cut down, and they have measured several independent variables for each tree, yielding $\mathbf{x_{n+1}} \ldots, \mathbf{x_{n+m}}$. What they want is a prediction of the *total* amount of wood they will get from these trees, along with a 95% prediction interval for the total.
   (a) The quantity they want to predict is $W = \sum_{j=n+1}^{n+m} y_j$, where $y_j = \mathbf{x}_j'\boldsymbol{\beta} + \epsilon_j$. What is the distribution of W? You can just write down the answer without showing any work.
   (b) Let $\widehat{W}$ denote the prediction of W. It is calculated using the company's regression data along with $\mathbf{x_{n+1}} \ldots, \mathbf{x_{n+k}}$. Give a formula for $\widehat{W}$.
   (c) What is the distribution of $W - \widehat{W}$? Show your work, but don't use moment-generating functions. Just write down expected value and calculate the variance.
   (d) Now standardize $W - \widehat{W}$ to obtain a standard normal. Call it Z.
   (e) Divide Z by the square root of a chi-squared random variable, divided by its degrees of freedom, and simplify. Call it T. What are the degrees of freedom?
   (f) How do you know that numerator and denominator are independent?
   (g) Using your formula for T, derive the $(1-\alpha) \times 100\%$ prediction interval for W. Please use the symbol $t_{\alpha/2}$ for the critical value.

[1] Copyright information is at the end of the last page.

3. This question uses the `trees` data you saw in the R lecture ("Least squares with R"). Start by fitting a model with just `Girth` and `Height`.

   The forestry company wants to predict the volume of wood they would obtain if they cut down three particular trees. The first tree has a girth of 11.0 and a height of 75. The second tree has a girth of 14.8 and a height of 80. The third tree has a girth of 10.5 and a height of 65. Using R,

   (a) Calculate a predicted amount of wood the company will obtain by cutting down these trees. The answer is a number.

   (b) Calculate a 95% prediction interval for the total amount of wood. The answer is a pair of numbers, a lower prediction limit and an upper prediction limit.

4. In a study comparing the effectiveness of different weight loss diets, volunteers were randomly assigned to one of two diets (A or B) or put on a waiting list and advised to lose weight on their own. Participants were weighed before and after 6 months of participation in the program (or 6 months of being on the waiting list). The response variable is weight *loss*. The explanatory variables are age and treatment group.

   (a) Write the regression equation, starting with $y_i = \beta_0 + \ldots$. Your model should have parallel regression lines for the different treatment groups, and notice that it *does have an intercept*. Please use x for age. You don't have to say how your dummy variables are defined. You'll do that in the next part.

   (b) Make a table with three rows, showing how you would set up indicator dummy variables for treatment group. Also give $E(Y|x)$ in the last column. See lecture slide shows for examples.

   (c) In terms of β values, what null hypothesis would you test to find out whether, allowing for age, the three diets (including Wait List) differ in their effectiveness?

   (d) In terms of β values, what null hypothesis would you test to find out whether, allowing for age, diets A and B differ in their effectiveness?

   (e) In terms of β values, what null hypothesis would you test to find out whether 25 year old participants who spend six months on the Wait list "diet" have any average tendency to gain or lose weight? Remember, Y is weight loss.

5. Now please repeat Question 4 with *cell means coding*. That's the dummy variable coding scheme with no β_0, and an indicator dummy variable for each diet (including Wait List).
