STA 302f16 Assignment Ten[1]

Except for Problem 5, these problems are preparation for the quiz in tutorial on Thursday November 24th, and are not to be handed in. Please bring your printout for Problem 5 to the quiz. Do not write anything on the printout in advance of the quiz, except possibly your name and student number.

1. Suppose you fit (estimate the parameters of) a regression model, obtaining $\mathbf{b}$, $\widehat{\mathbf{y}}$ and $\mathbf{e}$. Call this Model One.

   (a) In attempt to squeeze some more information out of the data, you fit a second regression model, using $\mathbf{e}$ from Model One as the dependent variable, and exactly the same X matrix as Model One. Call this Model Two.

      i. What is $\mathbf{b}$ for Model Two? Show your work and simplify.
      ii. What is $\widehat{\mathbf{y}}$ for Model Two? Show your work and simplify.
      iii. What is $\mathbf{e}$ for Model Two? Show your work and simplify.
      iv. What is s^2 for Model Two?

   (b) Now you fit a *third* regression model, this time using $\widehat{\mathbf{y}}$ from Model One as the dependent variable, and again, exactly the same X matrix as Model One. Call this Model Three.

      i. What is $\mathbf{b}$ for Model Three? Show your work and simplify.
      ii. What is $\widehat{\mathbf{y}}$ for Model Three? Show your work and simplify.
      iii. What is $\mathbf{e}$ for Model Three? Show your work and simplify.
      iv. What is s^2 for Model Three?

2. Data for a regression study are collected at two different locations; n_1 observations are collected at location one, and n_2 observations are collected at location two. The same independent variables are used at each location. We need to know whether the error variance σ^2 is the same at the two locations, possibly because we are concerned about data quality.

   Recall the definition of the F distribution. If $W_1 \sim \chi^2(\nu_1)$ and $W_2 \sim \chi^2(\nu_2)$ are independent, then $F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim F(\nu_1, \nu_2)$. Suggest a statistic for testing $H_0 : \sigma_1^2 = \sigma_2^2$. Using facts from the formula sheet, show it has an F distribution when H_0 is true. Don't forget to state the degrees of freedom. Assume that data coming from the two locations are independent.

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

3. Assume the usual linear model with normal errors; see the formula sheet. We know that a one-to-one linear transformation of the independent variables affects the interpretation of the β_j parameters, but otherwise it has no effect. Suppose that a model is to be used for prediction only, so that interpretation of the regression coefficients is not an issue. Here is a transformation that has interesting effects; it is also convenient for some purposes.

   Since $X'X$ is symmetric, we have the spectral decomposition $X'X = CDC'$, where D is a diagonal matrix of eigenvalues, and the columns of C are the corresponding eigenvectors. Suppose we transform X by $X^* = XC$. This also transforms $\boldsymbol{\beta}$, and the corresponding estimated $\boldsymbol{\beta}^*$ is denoted by $\mathbf{b}^*$.

   (a) Could any of the eigenvalues be negative or zero? Answer Yes or No and briefly explain. This might require some review.

   (b) Give a formula for $\mathbf{b}^*$. Simplify.

   (c) What is the distribution of $\mathbf{b}^*$? Simplify.

   (d) What is $Var(b_j^*)$?

   (e) Are the b_j^* random variables independent? Answer Yes or No. Why?

   (f) What is the variance of the linear combination $\ell_0 b_0^* + \ell_1 b_1^* + \cdots + \ell_k b_k^*$?

4. 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 $\mathbf{b}$ and s^2 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 r more trees they are planning to cut down, and they have measured the independent variables for each tree, yielding $\mathbf{x_{n+1}} \ldots, \mathbf{x_{n+r}}$. 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 $w = \sum_{j=n+1}^{n+r} y_j$, 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 $\mathbf{x_{n+1}} \ldots, \mathbf{x_{n+r}}$. Give a formula for $\widehat{w}$. Simplify.

   (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.

5. 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.

6. Regression diagnostics are mostly based on the residuals. This question compares the error terms ϵ_i to the residuals e_i. Answer True or False to each statement. For statements about the residuals, show a calculation that proves your answer. You may use anything on the formula sheet.

   (a) $E(\epsilon_i) = 0$

   (b) $E(e_i) = 0$

   (c) $Var(\epsilon_i) = 0$

   (d) $Var(e_i) = 0$

   (e) ϵ_i has a normal distribution.

   (f) e_i has a normal distribution.

   (g) $\epsilon_1, \ldots, \epsilon_n$ are independent.

   (h) $e_1, \ldots, e_n$ are independent.

7. One of these statements is true, and the other is false. Pick one, and show it is true with a quick calculation. Start with something from the formula sheet.

   - $\widehat{\mathbf{y}} = X\mathbf{b} + \mathbf{e}$
   - $\mathbf{y} = X\mathbf{b} + \mathbf{e}$
   - $\widehat{\mathbf{y}} = X\boldsymbol{\beta} + \mathbf{e}$

   As the saying goes, "Data equals fit plus residual."

8. The *deleted residual* is $e_{(i)} = y_i - \mathbf{x}_i'\mathbf{b}_{(i)}$, where $\mathbf{b}_{(i)}$ is defined as usual, but based on the $n-1$ observations with observation i deleted.

   (a) Guided by an expression on the formula sheet, write the formula for the Studentized deleted residual. You don't have to prove anything. You will need the symbols $X_{(i)}$ and $s^2_{(i)}$, which are defined in the natural way.

   (b) If the model is correct, what is the distribution of the Studentized deleted residual? Make sure you have the degrees of freedom right.

   (c) Why are numerator and denominator independent?

9. For the general linear regression model, are $\widehat{\mathbf{y}}$ and $\mathbf{e}$ independent?

   (a) Answer Yes or No and prove your answer.

   (b) What does this imply about the plot of predicted values against residuals?

10. For the general linear regression model, are $\mathbf{y}$ and $\widehat{\mathbf{y}}$ independent? Answer Yes or No and prove your answer.

11. For the general linear regression model, are $\mathbf{y}$ and $\mathbf{e}$ independent? Answer Yes or No and prove your answer.

12. For the general linear regression model, calculate $X'\mathbf{e}$ one more time. This will help with the next question.

13. For the general linear regression model in which X is a matrix of constants,

    (a) Why does it not make sense to ask about independence of the independent variable values and the residuals?

    (b) Prove that the sample correlation between residuals and independent variable values must equal exactly zero.

    (c) Does this result depend on the correctness of the model?

    (d) What does the sample correlation between residuals and independent variable values imply about the corresponding plots?

Please bring your printout for Question 5 to the quiz. **Your printout should show *all* R input and output, and *only* R input and output**. Do not write anything on your printouts except your name and student number.
