STA 302f17 Assignment Nine[1]

Except for Problem 12, these problems are preparation for the quiz in tutorial on Thursday November 23d, and are not to be handed in. Please bring your printouts for Problem 12 to the quiz. Do not write anything on the printouts 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 and the columns of X linearly independent; 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 (call them $\lambda_0, \lambda_1, \ldots, \lambda_k$), 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. This question will be a lot easier if you remember that if $X \sim \chi^2(\nu)$, then $E(X) = \nu$ and $Var(X) = 2\nu$. You don't have to prove these facts; just use them.

   For the usual linear regression model with normal errors, σ^2 is usually estimated with $s^2 = \mathbf{e}'\mathbf{e}/(n-k-1)$.

   (a) Show that s^2 is an unbiased estimator of σ^2. You did this the hard way in an earlier assignment. It's much easier when the errors are normal.

   (b) What is the distribution of $\sum_{i=1}^n \left(\frac{\epsilon_i - 0}{\sigma}\right)^2$?

   (c) Here is another estimate of σ^2. Define $v = \frac{1}{n}\sum_{i=1}^n \epsilon_i^2$. What is $E(v)$?

   (d) Show that $Var(v) < Var(s^2)$.

   (e) So it would appear that v is a better estimator of σ^2 than s^2 is, since they are both unbiased and the variance of v is lower. So why do you think s^2 is used in regression analysis instead of v?

5. 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.

   (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.

6. One of these statements is true, and the others are 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."

7. 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 e_i^*. 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?

8. You know that for the general linear regression model, $\widehat{\mathbf{y}}$ and $\mathbf{e}$ are independent, meaning that they are *always* independent, for every regression model with normal error terms.

   (a) Are $\mathbf{y}$ and $\widehat{\mathbf{y}}$ independent? Answer Yes or No and prove your answer.

   (b) Are $\mathbf{y}$ and $\mathbf{e}$ independent? Answer Yes or No and prove your answer.

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

10. In the regression model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, let $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\Omega)$, with Ω a *known* symmetric positive definite matrix.

    (a) Is $\mathbf{b}$ still an unbiased estimator of $\boldsymbol{\beta}$ for this problem?

    (b) What is $cov(\mathbf{b})$ for this problem?

    (c) Multiply $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ on the left by $\Omega^{-1/2}$, obtaining $\mathbf{y}^* = X^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$. What is the distribution of $\boldsymbol{\epsilon}^*$?

    (d) Substituting X^* and $\mathbf{y}^*$ into the formula for $\mathbf{b}$, obtain the generalized least squares estimate $\mathbf{b}_{gls} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\mathbf{y}$ on page 133 of the textbook. If you look in the textbook, I think you will appreciate the notation we are using.

    (e) What is the distribution of $\mathbf{b}_{gls}$? Show your work. All you have to do is calculate the expected value and covariance matrix, but *why*? What specific fact on the formula sheet are you using?

    (f) Just realize that there's more. You could obtain formulas for the starred versions of H, $\widehat{\mathbf{y}}$, $\mathbf{e}$, and F statistic for the general linear test.

11. For a very simple aggregated data set, our data are a collection of sample means $\overline{y}_1, \ldots, \overline{y}_n$ based on n independent random samples from a common population. Data values in the *unaggregated* data set come from a distribution with common mean μ and common variance σ^2. Sample mean i is based on m_i observations, so that (approximately by the Central Limit Theorem), $\overline{y}_i \sim N(\mu, \frac{\sigma^2}{m_i})$.

    (a) One could estimate μ with the arithmetic mean of the sample means. Is this estimator unbiased? What is its variance?

    (b) Start with the regression-like equation $\overline{y}_i = \mu + \epsilon_i$, where $\epsilon_i \sim N(\mu, \frac{\sigma^2}{m_i})$. Multiply both sides by $\sqrt{m_i}$, obtaining a starred version of the regression equation. What is $Var(e_i^*)$?

    (c) Give the generalized (weighted) least squares estimate of μ. Call it $\widehat{\mu}_{gls}$.

    (d) If you had access to the unaggregated data (that is, all the y_{ij} values), how would you estimate μ? Yes, that's you personally. What is the connection of your statistic to $\widehat{\mu}_{gls}$?

12. Pigs are routinely given large doses of antibiotics even when they show no signs of illness, to protect their health under unsanitary conditions. Pigs were randomly assigned to one of three antibiotic drugs. Dressed weight (weight of the pig after slaughter and removal of head, intestines and skin) was the dependent variable. Independent variables are Drug type, Mother's live adult weight and Father's live adult weight.

    Data are in the file pigweight.data.txt. You can get a copy with

```
oink = read.table("http://www.utstat.toronto.edu/~brunner/data/legal/pigweight.data.txt").
```

    (a) Write the regression equation for the full model, including ϵ_i.

    (b) Make a table with one row for every drug, with columns showing how R would define the dummy variables by default. Make another column giving $E(y|\mathbf{x})$ for each drug.

    (c) Fit the model with R, and predict the dressed weight of a pig getting Drug 2, whose mother weighed 140 pounds, and whose father weighed 185 pounds. Use the `predict` function to obtain the prediction and a 95% prediction interval. Your answer is a set of three numbers, a prediction, a lower prediction limit and an upper prediction limit.

    (d) This parallel planes regression model specifies that the differences in expected weight for the different drug treatments are the same for every possible combination of mother's weight and father's weight. Give a 95% confidence interval for the difference in expected weight between drug treatments 2 and 3. The final answer is a pair of numbers, a lower confidence limit and an upper confidence limit. There is an easy way and a less easy way.

    (e) In symbols, give the null hypotheses you would test to answer the following questions. Your answers are statements involving the β values from your regression equation.

        i. Controlling for mother's weight and father's weight, does type of drug have an effect on the expected weight of a pig?

        ii. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 1 or Drug 2?

        iii. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 1 or Drug 3?

        iv. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 2 or Drug 3?

(f) For each of the questions below, give the value of the t or F statistic (a number from your printout), and indicate whether or not you reject the null hypothesis. The numbers may or may not be part of the default output from `summary`.

i. Controlling for mother's weight and father's weight, does type of drug have an effect on the expected weight of a pig?

ii. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 1 or Drug 2?

iii. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 1 or Drug 3?

iv. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 2 or Drug 3?

v. Allowing for which drug they were given, does expected weight of a pig increase faster as a function of the mother's weight, or does it increase faster as a function of the father's weight?

(g) An accepted rule of thumb about influential observations is that if the maximum diagonal value of the H matrix is bigger than 0.2, one or more observations might be having too much influence on the results. Is there evidence of this kind of trouble? Your printout should have the one number you need to answer the question.

(h) To check the residuals for possible outliers, treat the Studentized deleted residuals as t statistics with a Bonferroni correction. Is there evidence of outliers? Of course the evidence for your conclusion (including the critical value) should be on your printout.

(i) We can assume that farmers want their pigs to weigh a lot. In plain, non-statistical language, can you offer some advice to a farmer based on these data? Remember, the farmer must be able to understand your answer or it is worthless.

Please bring your printout to the quiz.
