STA 302f16 Assignment Nine[1]

Except for Problem 9, these problems are preparation for the quiz in tutorial on Thursday November 17th, and are not to be handed in. As usual, sometimes you may be asked to prove things that are false. Please bring your printout for Problem 9 to the quiz. Do not write anything on the printout in advance of the quiz, except possibly your name and student number.

1. Consider a linear regression model with $n > p$, which is always the case in practice. Since the vector of residuals $\mathbf{e} \sim N_n\left(\mathbf{0}, \sigma^2(I - H)\right)$, it is tempting to write $\frac{1}{\sigma^2}\mathbf{e}'(I - H)^{-1}\mathbf{e} \sim \chi^2(n)$. Please locate support for this idea on the formula sheet. But it only works if the $n \times n$ matrix $I - H$ has an inverse. Calculate $(I - H)\,X$, and use this to show that if $(I - H)^{-1}$ exists, the columns of X cannot be linearly independent.

2. 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?

3. In a study comparing the effectiveness of different exercise programmes, volunteers were randomly assigned to one of three exercise programmes (A, B, C) or put on a waiting list and told to work out on their own. Aerobic capacity is the body's ability to process oxygen. Aerobic capacity was measured before and after 6 months of participation in the program (or 6 months of being on the waiting list). The response variable was improvement in aerobic capacity. The explanatory variables were age (a covariate) and treatment group.

   (a) First consider a regression model with an intercept, and no interaction between age and treatment group.

      i. Make a table showing how you would set up indicator dummy variables for treatment group. Make Waiting List the reference category

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

ii. Write the regression model. Please use x for age, and make its regression coefficient β_1.

iii. In terms of β values, what null hypothesis would you test to find out whether, allowing for age, the three exercise programmes differ in their effectiveness? That's the *three* programs, not including the wait list control.

iv. Write the null hypothesis for the preceding question as $H_0 : C\boldsymbol{\beta} = \mathbf{0}$. Just give the C matrix.

v. In terms of β values, what null hypothesis would you test to find out whether Programme B was better than the waiting list?

vi. In terms of β values, what null hypothesis would you test to find out whether Programmes A and B differ in their effectiveness?

vii. Suppose you wanted to estimate the difference in average benefit between programmes A and C for a 27 year old participant. Give your answer in terms of b_j values.

viii. Is it safe to assume that age is independent of the other explanatory variables? Answer Yes or No and briefly explain.

4. Now consider a regression model with an intercept and the interaction (actually a set of interactions) between age and treatment.

(a) Write the regression model. Make it an extension of your earlier model.

(b) Suppose you wanted to know whether the slopes of the 4 regression lines were equal. In terms of β values, what null hypothesis would you test?

(c) Suppose you wanted to know whether any differences among mean improvement in the four treatment conditions depends on the participant's age. In terms of β values, what null hypothesis would you test?

(d) Write the null hypothesis for the preceding question as $H_0 : C\boldsymbol{\beta} = \mathbf{0}$. Just give the C matrix. It is $r \times p$. What is r? What is p?

(e) Suppose you wanted to know whether the difference in effectiveness between Programme A and the Waiting List depends on the participant's age. In terms of β values, what null hypothesis would you test?

(f) Suppose you wanted to *estimate* the difference in average benefit between programmes A and C for a 27 year old participant. Give your answer in terms of $\widehat{\beta}$ values.

5. A general principle is that all valid dummy variable coding schemes are equivalent. This is because they are one-to-one linear transformations of one another. Let A be a $(k+1) \times (k+1)$ nonsingular matrix. Note that $X^* = XA$ is a one-to-one linear transformation of the explanatory variables, and

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \;\Leftrightarrow\; \mathbf{y} = XA\,A^{-1}\boldsymbol{\beta} + \boldsymbol{\epsilon} = X^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}.$$

This is already interesting, because it shows how transforming the explanatory variables changes the meaning of the regression coefficients. Refer to $\mathbf{y} = X\boldsymbol{\beta}+\boldsymbol{\epsilon}$ as the "original" model, and $\mathbf{y} = X^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$ as the "transformed" model.

   (a) Just to make this more concrete, suppose you have a 3-category explanatory variable and a quantiative covariate. $Y_i = \beta_0 + \beta_1 d_{i,1} + \beta_2 d_{i,2} + \beta_3 x_i + \epsilon_i$, where $d_{i,1}$ and $d_{i,2}$ are indicator dummy variables for the first two groups. You want to switch to cell means coding, so that $Y_i = \beta_1^* g_{i,1} + \beta_2^* g_{i,2} + \beta_3^* g_{i,3} + \beta_4^* x_i + \epsilon_i$. Note that $\beta_4^* = \beta_3$. Give the matrix A; you can make tables if that helps.

   (b) Write down the least squares estimate $\mathbf{b}^*$ for the transformed model, and simplify. How is $\mathbf{b}^*$ related to $\mathbf{b}$? Give a formula.

   (c) Compare the vector of predicted values from the two models.

   (d) Compare the vector of residuals from the two models.

   (e) Which is greater, SSE or SSE^*?

   (f) Suppose you want to test $H_0: C\boldsymbol{\beta} = \boldsymbol{\gamma}$. Give the equivalent null hypothesis for the transformed model. That is, what are matrices C^*, $\boldsymbol{\beta}^*$ and $\boldsymbol{\gamma}^*$ in $H_0: C^*\boldsymbol{\beta}^* = \boldsymbol{\gamma}^*$?

   (g) Compare the F statistic for $H_0: C^*\boldsymbol{\beta}^* = \boldsymbol{\gamma}^*$ to the F statistic for $H_0: C\boldsymbol{\beta} = \boldsymbol{\gamma}$.

6. Question 5 suggests that if a regression model with no intercept is equivalent to one with an intercept, then the residuals will add to zero. This is good to know, because it means $SST = SSR + SSE$, and R^2 is meaningful; so is a, the proportion of remaining variation. Here is an easy condition to check. Let $\mathbf{1}$ denote an $n \times 1$ column of ones. Show that if there is a $(k+1) \times 1$ vector of constants $\mathbf{v}$ with $X\mathbf{v} = \mathbf{1}$, then $\sum_{i=1}^n e_i = 0$. (Another way to state this is that if there is a linear combination of the columns of X that equals a column of ones, then the sum of residuals equals zero. Clearly this applies to a model with cell means coding.)

7. 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.

8. Suppose you have a random sample from a normal distribution, say $y_1, \ldots, y_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$. If someone randomly sampled another observation from this population and asked you to guess what it was, there is no doubt you would say $\overline{y}$, and a confidence interval for μ is routine. But what if you were asked for a *prediction* interval for a *new* observation?

   Accordingly, suppose the normal model is reasonable and you observe a sample mean of $\overline{y} = 7.5$ and a sample variance (with $n-1$ in the denominator) of $s^2 = 3.82$. The sample size is $n = 14$. Give a 95% prediction interval for the next observation. The answer is a pair of numbers. Be able to show your work. You can get the distribution result you need from the formula sheet, or you can re-derive it for this special case. Be able to do it both ways. You should use R to get the critical value, but don't bother to bring your R printout for this question.

9. 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 the dummy variables were defined. Make another column giving $E(y|\mathbf{x})$ for each drug.

   (c) Predict the dressed weight of a pig getting Drug 2, whose mother weighed 140 pounds, and whose father weighed 185 pounds. Your answer is a single number.

   (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) 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. **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.