STA 2101/442 Assignment Seven[1]

The questions are just practice for the quiz, and are not to be handed in. Use R for Questions 6 and 7, and 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.

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

   (a) Then just for fun, you fit a second regression model, using $\widehat{\mathbf{y}}$ from Model One as the response variable, and exactly the same $\mathbf{X}$ matrix as Model One. Call this Model Two.

      i. What is $\widehat{\boldsymbol{\beta}}$ 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 MSE for Model Two?

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

      i. What is $\widehat{\boldsymbol{\beta}}$ 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 MSE for Model Three?

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

3. 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 MSE; see the formula sheet.

   (a) Show that MSE is an unbiased estimator of σ^2.

   (b) Show that MSE is a consistent estimator of σ^2. I used the Variance Rule for this one.

   (c) Under the usual regression model what is the joint distribution of $\epsilon_1, \ldots, \epsilon_n$?

   (d) Let $T_n = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2$. What is $E(T_n)$?

   (e) How do you know that $T_n \stackrel{a.s.}{\to} \sigma^2$?

   (f) Show that $Var(T_n) < Var(MSE)$.

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

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

4. Suppose 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 explanatory 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.

5. 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

      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?

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

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

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

      i. Write the regression model. Make it an extension of your earlier model.

      ii. 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?

      iii. 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?

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

      v. 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?

      vi. 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.

(c) Now consider a regression model *without* an intercept, but *with* possibly unequal slopes. Make a table to show how the dummy variables could be set up, and write the regression model. Again, please use x for age and make its regression coefficient β_1. This model needs to have the *same number of regression coefficients as the model of Question 5b*, so you have to think about this a little.

For each treatment condition, what is the conditional expected value of Y? The answer is in terms of x and the β values. Please put these values as the last column of your table.

i. 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?

ii. 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?

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

iv. 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?

v. 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.

6. Suppose you have a random sample from a normal distribution, say $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. If someone randomly sampled another observation from this population and asked you to guess what it was, there is no doubt you would say $\overline{Y}$. But what if you were asked for a prediction *interval*?

(a) 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.

(b) This suggests using Bonferroni-corrected deleted Studentized residuals for detecting outliers in simple independent random sampling, as long as a normal model seems reasonable. Try it with these data: `10.17 9.69 8.99 6.45 13.81 9.82 4.07 7.62 10.84 9.97`. Any outliers? Use R. Bring your printout to the quiz.

7. In the Chick Weights study, newly hatched chickens were randomly assigned to one of six different feed supplements, and their weight in grams after 6 weeks was recorded. The Chick Weights data are in an R dataset called `chickwts`. Type `chickwts` to see it.

   (a) Make sure a table of means, standard deviations and sample sizes for the 6 feed types is part of your output.

   (b) Test whether the six mean weights are different. Get the F statistic, degrees of freedom, p-value and proportion of explained variation.

   (c) Check the Studentized deleted residuals with a Bonferroni correction. Do there seem to be any outliers?

   (d) You want to know which means are different from which other means. Carry out the multiple comparison procedure likely to be the most powerful in this situation. Base your conclusions on the usual $\alpha = 0.05$ *joint* significance level for the family of tests. Of course when you state your conclusions in plain language, you would not mention the significance level or joint significance level. But to be honest, stating the conclusions in plain language isn't easy in this case. The pattern is complicated.

   (e) Test for differences among mean weights for the five feed types *excluding* horsebean.

      i. First, write the null hypothesis in terms of μ values.

      ii. Now obtain the F statistic, degrees of freedom and p-value. Do you reject H_0 at $\alpha = 0.05$?

      iii. What is the "reduced" (restricted) model for this problem? You should not do this test by fitting a full and a reduced model, but be aware that you could.

      iv. What proportion of the remaining variation does the effect explain? The answer is a number between zero and one that you could obtain from your printout with a calculator, though you may choose to do it with R and have it on your printout.

   (f) Obtain a 95% confidence interval for the difference between the expected weight for chicks fed horsebean, versus the mean of the other expected values. Your answer is a pair of numbers: a lower limit and an upper limit.

   (g) Would you advise a chicken farmer to purchase the Horsebean feed supplement if she wanted big fat chickens?

   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.
