STA 2201/442 Assignment 7

Please bring printouts for the R parts to the quiz. Some may be handed in. *Do not write anything on your printouts in advance except maybe your name (and student number, if you wish).*

1. This homework problem comes from the following excellent question by a student. Wilks' theorem for the large-sample likelihood ratio test applies to data that are independent and identically distributed. Does it also apply to a conditional model like the normal regression model (and others) where there are fixed constants that make the data independent but no longer identically distributed? Here you will see a set of reasonable conditions under which the answer is "Yes."

   For $i = 1, \ldots, n$, let the pairs $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ be independent and identically distributed. The boldface notation indicates that $\mathbf{X}_i$ and $\mathbf{Y}_i$ could be matrices, so this is quite general. The probability distribution of $(\mathbf{X}_i, \mathbf{Y}_i)$ has the following structure. The conditional distribution of $\mathbf{Y}_i$ given $\mathbf{X}_i = \mathbf{x}_i$ depends upon the parameter (vector) $\boldsymbol{\theta}_1$[1]. The marginal distribution of $\mathbf{X}_i$ depends upon the parameter (vector) $\boldsymbol{\theta}_2$. There is no functional connection between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, meaning that no parameters appear in both vectors, and no parameter in either set is a function of any parameters in the other set. Also, assume that the distributions involved have likelihoods (joint probability densities or probability mass functions), and that the modest regularity conditions of Wilks' theorem hold.

   Finally, after all this assuming, assume that the null hypothesis being tested concerns only $\boldsymbol{\theta}_1$ (one could say that $\boldsymbol{\theta}_2$ is a nuisance parameter). Show that the likelihood ratio test for the conditional model is identical to the one for the unconditional model. The answer is a lot shorter than the question.

2. In Assignment 6, Question 5 was about the regression analysis of the SAT data. Part x. of Question 5 asks how many Studentized deleted residuals you would expect to be beyond the critical value of t just by chance.

   (a) Write the number of Studentized deleted residuals beyond the critical value as a sum of random variables, then take the expected value. This shows that the non-independence of these random variables has no effect on the *expected* number beyond the critical value.

   (b) And indeed the random variables are not independent. There is one for each hypothesis test, and the test statistics have almost the same $\widehat{\beta}$ and MSE. How many are there? The answer is a number.

   (c) Suppose we want to protect all the tests against Type I error at *joint* significance level 0.05 with a Bonferroni correction? What critical value of t should we use? The answer is a number, which I computed with R.

[1] For the usual univariate regression model, $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \sigma^2)$.

(d) List all the Studentized deleted residuals that are beyond the Bonferroni critical value. That's better! It's helpful to think of detecting outliers as a multiple comparison problem.

3. In a study of remedies for lower back pain, volunteer patients at a back clinic were randomly assigned to one of seven treatment conditions:

   1. OxyContin: A pain pill in the opiate family.
   2. Ibuprofin: A non-steroidal anti-inflammatory drug (Advil, Motrin)
   3. Acupuncture: The insertion and manipulation of thin needles into specific points on the body to relieve pain or for therapeutic purposes.
   4. Chiropractic: A form of therapy that includes manipulation of the spine, other joints and soft tissue.
   5. Stress reduction training based on thinking positive thoughts, a treatment that theoretically should not be effective. This is the non-drug control condition.
   6. Placebo: A sugar pill; patients were told that it was a pain killer with few side effects. This is the drug control condition.
   7. Waiting list control: Patients were told that the clinic was overcrowded (true), and that they would were on a waiting list. This group received no treatment at all, not even a pretend treatment — until the study was over, at which point they received the most effective treatment based on the results of the study. We'll call this the "No treatment" group.

   Degree of reported pain was measured by a questionnaire before treatment began, and again after six weeks. The response variable was Before-minus-After difference in reported pain, which will be called "improvement," or "effectiveness." Each of the following questions can be answered by testing whether one or more contrasts of treatment means are different from zero. For each question below, first state the null hypothesis in terms of the population treatment means μ_1 through μ_7, and then give the contrast matrix $\mathbf{C}$ for $H_0 : \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$.

   (a) Does OxyContin work any better than the placebo[2]?

   (b) Does Ibuprofin work any better than the placebo?

   (c) Do Chiropractic treatment and Stress reduction training differ in their effectiveness?

   (d) Which results in more mean improvement, Acupuncture or Stress reduction training?

[2]This and some of the other questions might seem like recipes for one-sided tests, but think about it! If the placebo worked better, wouldn't they have a press conference?

(e) Is the average improvement from the two drug therapies different from the improvement from the placebo?

(f) Does either drug therapy differ from the placebo in its effectiveness? (This involves 2 contrasts.)

(g) Does either non-drug therapy differ in effectiveness from Stress reduction training?

(h) Is the Placebo better than no treatment at all?

(i) Is Stress reduction training better than no treatment at all?

(j) Is the average effectiveness of the drug therapies different from the average effectiveness of the non-drug therapies?

(k) Do Stress reduction training and the Placebo differ in their effectiveness?

(l) Does either control condition (Drug or Non-Drug) differ from no treatment at all?

(m) Is treatment condition (the full categorical explanatory variable) related to improvement?

4. 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 (the body's ability to process oxygen) 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 $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$. Just give the $\mathbf{C}$ matrix. Is the test statistic part of the `summary` output for an `lm` object?

      v. In terms of β values, what null hypothesis would you test to find out whether Programme B was better than the waiting list? Is the test statistic part of the `summary` output for an `lm` object?

      vi. In terms of β values, what null hypothesis would you test to find out whether Programmes A and B differ in their effectiveness? Is the test statistic part of the `summary` output for an `lm` object?

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 N 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 parallel. 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 $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$. Just give the $\mathbf{C}$ matrix.

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? Is the test statistic part of the `summary` output for an `lm` object?

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 non-parallel 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. 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.

Of course it would be possible to ask dozens of questions like the ones for the model with an intercept, but we'll leave that for later.

5. By typing `help(chickwts)` at the `R` prompt, you can find out about the `chickwts` data.

(a) First, produce a table of means, standard deviations and sample sizes for the 6 feed types. If there is a package for this, I'd like to know about it.

(b) Now how about some side-by-side boxplots? Can you do anything about those axis labels?

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

(d) As in Question 2, treat the Studentized deleted residuals as t-test statistics with a Bonferroni correction. What is the Bonferroni critical value and how many residuals exceed it?

(e) Carry out Tukey HSD tests for all pairwise differences between means. Which means are different from each other at the *joint* 0.05 level?

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

i. First, write the null hypothesis in terms of β values. For this to be helpful, you must use `R`'s default dummy variable coding scheme or change it to something you prefer.

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