STA 431s15 Assignment Three[1]

The non-computer questions on this assignment are practice for Term Test One on Monday February 2nd. There will be no SAS on Term Test One. The SAS part of this assignment (Question 12) is for Quiz Three on Friday February 6th. Please bring your log file and your output file to the quiz. There will be one or more questions about them, and you will be asked to hand them in with the quiz.

1. For each of these problems, $Y_1, \ldots, Y_n$ are a random sample from a distribution with the given density of probability mass function. For each one, obtain the formula for a Method of Moments estimator, and then calculate your formula for the given data. There is more than one estimator for most distributions, but try to find a nice simple one.

   (a) $f(x) = \theta x^{\theta-1}$ for $0 < x < 1$, where $\theta > 0$. Data: `0.04, 0.69, 0.86, 0.24, 0.99`

   (b) $p(x) = \binom{4}{x}\theta^x(1-\theta)^{4-x}$ for $x = 0, \ldots, 4$. Data: `1, 2, 3, 2, 0, 2`. Use the expected value of a Binomial without proof.

   (c) $f(x) = \frac{1}{2\theta}$ for $-\theta < x < \theta$ where $\theta > 0$. Data: `-1.43 1.89 1.58 -0.62 0.72 -1.75`.

2. Independently for $i = 1, \ldots, n$, let $\mathbf{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\mathbf{X}_i + \boldsymbol{\epsilon}_i$, where

   - $\mathbf{Y}_i$ is an $q \times 1$ random vector of observable response variables; there are q response variables.
   - $\mathbf{X}_i$ is a $p \times 1$ observable random vector; there are p explanatory variables. $E(\mathbf{X}_i) = \boldsymbol{\mu}_x$ and $V(\mathbf{X}_i) = \boldsymbol{\Phi}_{p\times p}$. The positive definite matrix $\boldsymbol{\Phi}$ is unknown.
   - $\boldsymbol{\beta}_0$ is a $q \times 1$ matrix of unknown constants.
   - $\boldsymbol{\beta}_1$ is a $q \times p$ matrix of unknown constants.
   - $\boldsymbol{\epsilon}_i$ is a $q \times 1$ random vector with expected value zero and unknown positive definite variance-covariance matrix $V(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}_{q\times q}$.
   - $\boldsymbol{\epsilon}_i$ is independent of $\mathbf{X}_i$.

   Letting $\mathbf{D}_i = \left(\begin{array}{c}\mathbf{X}_i \\ \hline \mathbf{Y}_i\end{array}\right)$, we have $V(\mathbf{D}_i) = \boldsymbol{\Sigma} = \left(\begin{array}{c|c}\boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy} \\ \hline \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_y\end{array}\right)$, and $\widehat{\boldsymbol{\Sigma}} = \left(\begin{array}{c|c}\widehat{\boldsymbol{\Sigma}}_x & \widehat{\boldsymbol{\Sigma}}_{xy} \\ \hline \widehat{\boldsymbol{\Sigma}}_{yx} & \widehat{\boldsymbol{\Sigma}}_y\end{array}\right)$.

   (a) Start by writing $\boldsymbol{\Sigma}$ in terms of the unknown parameter matrices.

   (b) Give a Method of Moments Estimator for $\boldsymbol{\Phi}$. Just write it down.

   (c) Obtain formulas for the Method of Moments Estimators of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_0$ and $\boldsymbol{\Psi}$. Show your work. You may give $\widehat{\boldsymbol{\beta}}_0$ in terms of $\widehat{\boldsymbol{\beta}}_1$, but simplify $\widehat{\boldsymbol{\Psi}}$.

[1] This assignment was prepared by Jerry Brunner, Department of Statistical Sciences, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/~brunner/oldclass/431s15

3. Independently for $i = 1, \ldots, n$, let $Y_i = \beta X_i + \epsilon_i$, where

   - $E(X_i) = \mu_x$, $Var(X_i) = \sigma^2_x$
   - $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2_\epsilon$
   - X_i and ϵ_i are independent.

   (a) Obtain formulas for the Method of Moments Estimators of β and σ^2_ϵ. Show your work.

   (b) Calculate your estimates for the following data:

   ```
   x  0.0  1.3  3.2 -2.5 -4.6 -1.6  4.5  3.8
   y -0.8 -1.3  7.4 -5.2 -6.5 -4.9  9.9  7.2
   ```

   To make it easier, use these calculations:

   ```
           x         x^2         y        y^2        xy
         0.0        0.00      -0.8       0.64      0.00
         1.3        1.69      -1.3       1.69     -1.69
         3.2       10.24       7.4      54.76     23.68
        -2.5        6.25      -5.2      27.04     13.00
        -4.6       21.16      -6.5      42.25     29.90
        -1.6        2.56      -4.9      24.01      7.84
         4.5       20.25       9.9      98.01     44.55
         3.8       14.44       7.2      51.84     27.36
   --------------------------------------------------------
   Mean 0.5125    9.57375     0.725      37.53      18.08
   ```

   Doing $\widehat{\beta}$ two different ways, I got 1.888497 and 1.414634. Doing $\widehat{\sigma}^2$ two different ways, I got 3.385971 and 3.797089. There are other correct answers.

4. In the text, the material on maximum likelihood and likelihood ratio test tends to emphasize numerical MLEs, and is a little more theoretical in places than we are going to be. However, parts may be quite useful. Please start reading on Page 128, and then do Exercises 1 and 2 starting on Page 130. If you did not need this review, please accept my apologies.

5. The formula sheet has a useful expression for the multivariate normal likelihood.

   (a) Show that you understand the notation by giving the univariate version, in which $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$. Your answer will have no matrix notation for the trace, transpose or inverse.

   (b) Now starting with the univariate normal density (also on the formula sheet), show that the univariate normal likelihood is the same as your answer to the previous question. Hint: Add and subtract $\overline{X}$.

   (c) How does this expression allow you to see *without differentiating* that the MLE of μ is $\overline{X}$?

6. Let $Y_1, \ldots, Y_n$ be a random sample from a distribution with density $f(y) = \frac{1}{\theta} e^{-\frac{y}{\theta}}$ for $y > 0$, where the parameter $\theta > 0$. We are interested in testing $H_0 : \theta = \theta_0$.

   (a) What is Θ?

   (b) What is Θ_0?

   (c) Derive a general expression for the large-sample likelihood ratio statistic.

   (d) A sample of size $n = 100$ yields $\overline{Y} = 1.37$ and $S^2 = 1.42$. One of these quantities is unnecessary and just provided to irritate you. Well, actually it's a mild substitute for reality, which always provides you with a huge pile of information you don't need. Anyway, we want to test $H_0 : \theta = 1$. You can do this with a calculator. When I did it a long time ago I got $G^2 = 11.038$.

   (e) What is the critical value at $\alpha = 0.05$? The answer is a number from the formula sheet.

   (f) Do you reject H_0? Answer Yes or No.

   (g) Is there evidence that $\theta = 1$? Answer Yes or No.

7. The label on the peanut butter jar says peanuts, partially hydrogenated peanut oil, salt and sugar. But we all know there is other stuff in there too. In the United States, the Food and Drug administration requires that a shipment of peanut butter be rejected if it contains an average of more than 8 rat hairs per pound (well, I'm not sure if it's exactly 8, but let's pretend). There is very good reason to assume that the number of rat hairs per pound has a Poisson distribution with mean λ, because it's easy to justify a Poisson process model for how the hairs get into the jars. We will test $H_0 : \lambda = \lambda_0$.

   (a) What is Θ?

   (b) What is Θ_0?

   (c) Derive a general expression for the large-sample likelihood ratio statistic.

   (d) We sample 100 1-pound jars, and observe a sample mean of $\overline{Y} = 8.57$. Should we reject the shipment? We want to test $H_0 : \lambda = 8$. What is the value of G^2? You can do this with a calculator. When I did it a long time ago I got $G^2 = 3.97$.

   (e) Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.

   (f) Do you reject the shipment of peanut butter? Answer Yes or No.

8. Let $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$.

   (a) Derive a general expression for the large-sample likelihood ratio test statistic for testing $H_0 : \sigma^2 = \sigma_0^2$ versus $\sigma^2 \neq \sigma_0^2$.

   (b) A random sample of size $n = 50$ yields $\overline{X} = 9.91$ and $\widehat{\sigma}^2 = 0.92$.

   (c) What is the critical value at $\alpha = 0.05$? The answer is a number from the formula sheet.

   (d) Do you reject $H_0 : \sigma^2 = 1$ at $\alpha = 0.05$?

   (e) What, if anything, do you conclude?

9. You might want to look again at the coffee taste test example from lecture before starting this question. An email spam company designs k different emails, and randomly assigns email addresses (from a huge list they bought somewhere) to receive the different email messages. So, this is a true experiment, in which the message a person receives is the experimental treatment. n_1 email addresses receive message 1, n_2 email addresses receive message 2, ..., and n_k email addresses receive message k.

   The response variable is whether the recipient clicks on the link in the email message: $Y_{ij} = 1$ if recipient i in Treatment j clicks on the link, and zero otherwise. According to our model, all these observations are independent, with $P(Y_{ij}) = \theta_j$ for $i = 1, \ldots, n_j$ and $j = 1, \ldots, k$. We want to know if there are any differences in the effectiveness of the treatments.

   (a) What is H_0?

   (b) What is Θ?

   (c) What is Θ_0?

   (d) Write the likelihood function.

   (e) What is $\widehat{\boldsymbol{\theta}}$? If you think about it you can write down the answer without doing any work.

   (f) What is $\widehat{\boldsymbol{\theta}}_0$? If you think about it you can write down the answer without doing any work.

   (g) Write down and simplify a general expression for the large-sample likelihood ratio statistic G^2. What are the degrees of freedom?

   (h) Comparing three spam messages with $n_1 = n_2 = n_3 = 1,000$, the company obtains $\overline{Y}_1 = 0.044$, $\overline{Y}_2 = 0.050$ and $\overline{Y}_3 = 0.061$.

   (i) What is the test statistic G^2? The answer is a number.

   (j) What is the critical value at $\alpha = 0.05$? The answer is a number from the formula sheet.

   (k) Do you reject H_0? Answer Yes or No.

   (l) Is there evidence that the messages differ in their effectiveness? Answer Yes or No.

10. You may think of this as a continuation of Question 2 of Assignment 2. Let $Y_i = \beta x_i + \epsilon_i$ for $i = 1, \ldots, n$, where $\epsilon_1, \ldots, \epsilon_n$ are a random sample from a normal distribution with expected value zero and variance σ^2. The parameters β and σ^2 are unknown constants. The numbers $x_1, \ldots, x_n$ are known, observed constants.

   (a) What is Θ?

   (b) If the null hypothesis is $H_0 : \beta = \beta_0$, what is Θ_0?

   (c) What is $\widehat{\beta}$? Just use your answer from Assignment 2 (but be able to do it again.)

   (d) What is $\widehat{\sigma}^2$? Again just use your answer from Assignment 2, but be able to do it again.

   (e) What is $\widehat{\sigma}_0^2$?

   (f) Show $G^2 = n \ln \frac{\sum_{i=1}^n (Y_i - \beta_0 x_i)^2}{\sum_{i=1}^n (Y_i - \widehat{\beta} x_i)^2}$

11. Let $\mathbf{D}_1, \ldots, \mathbf{D}_n$ be a random sample from a multivariate normal population with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Write $\mathbf{D}_i = \left(\frac{\mathbf{X}_i}{\mathbf{Y}_i} \right)$, where $\mathbf{X}_i$ is $q \times 1$, $\mathbf{Y}_i$ is $r \times 1$, and $p = q + r$, we have $V(\mathbf{D}_i) = \boldsymbol{\Sigma} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy} \\ \hline \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_y \end{array} \right)$, and $\widehat{\boldsymbol{\Sigma}} = \left(\begin{array}{c|c} \widehat{\boldsymbol{\Sigma}}_x & \widehat{\boldsymbol{\Sigma}}_{xy} \\ \hline \widehat{\boldsymbol{\Sigma}}_{yx} & \widehat{\boldsymbol{\Sigma}}_y \end{array} \right)$.

    (a) Calculate and simplify the large-sample likelihood ratio statistic G^2 for testing $H_0 : \boldsymbol{\Sigma}_{xy} = \mathbf{0}$, which is equivalent to $\mathbf{X}_i$ and $\mathbf{Y}_i$ independent. Start with the likelihood and MLEs on the formula sheet. Your answer is a formula. What are the degrees of freedom?

    (b) For the Twins Data, $\mathbf{X}_i$ could be the vector of three mental measurements and $\mathbf{Y}_i$ could be the vector of six physical measurements. For $n = 74$, I calculated $\ln |\widehat{\boldsymbol{\Sigma}}| = 40.814949$, $\ln |\widehat{\boldsymbol{\Sigma}}_x| = 14.913525$ and $\ln |\widehat{\boldsymbol{\Sigma}}_y| = 26.33133$.

        i. Calculate G^2 for these data. Your answer is a number.

        ii. What are the degrees of freedom? Your answer is a number.

        iii. SAS `proc calis` gave us Chi-square = 31.3831 for this problem (see lecture notes). Multiply $\frac{n-1}{n}G^2$ to get this number.

12. In the United States, admission to university is based partly on high school marks and recommendations, and partly on applicants' performance on a standardized multiple choice test called the Scholastic Aptitude Test (SAT). The SAT has two sub-tests, Verbal and Math. A university administrator selected a random sample of 200 applicants, and recorded the Verbal SAT, the Math SAT and first-year university Grade Point Average (GPA) for each student. The data are given in the file `openSAT.data.txt`.

    Your job is to test whether the variance of the Verbal SAT scores is different from the variance of the Math SAT scores. Start by estimating the variances; I did it with `proc corr`.

    Produce a large-sample Chi-squared test. Either G^2 or $\frac{n-1}{n}G^2$ is okay. On your printout, you should be able to locate

    - Unrestricted *maximum likelihood* estimates of the two variances (meaning divide by n, not $n - 1$), allowing them to be unequal. These are numbers.
    - The value of the test statistic (a number).
    - The degrees of freedom (a number).
    - The p-value (a number or range of numbers).

    Using if H_0 is rejected at the $\alpha = 0.05$ significance level, you should be able to state a conclusion like "The variance of the Verbal SAT scores is greater," or "The variance of the Math SAT scores is greater."

    Bring your log file and output file to the quiz. You may be asked for numbers from your printouts, and you will definitely be asked to hand them in. For full marks, **there must be no warnings, error messages or notes about missing data on your log file.**

    Please follow these guidelines. Marks will be deducted if you do not.

    - Put your name and student number in a `title` or `title2` statement.
    - Do not write anything on your printouts in advance of the quiz.

- Bring your log file to the quiz, *not* just a listing of the program file.
- The log file and the output file must be from the same run of SAS.
- Your output file must have a date stamp. This is automatically generated if you save a pdf file or print from SAS Studio.
- You must use *your* installation of SAS, not the installation on someone else's computer.