STA 2101 Assignment 4[1]

The questions on this assignment are not to be handed in. They are practice for Quiz Four on Friday October 11th. There is a posted formula sheet that will be provided with the quiz. Please bring the printouts from Questions 1, 2 and 5. It would be best to keep the printouts separate. You will not hand all of them in.

Your printouts should have *only* R input and output. It is okay to have the questions in comment statements, but NO ANSWERS! If you understand what you have done, the answers will be in your head.

1. 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 available here. We seek to predict GPA from the two test scores. Throughout, please use the usual $\alpha = 0.05$ significance level.

   (a) First, fit a model using just the Math score as a predictor. "Fit" means estimate the model parameters. Does there appear to be a relationship between Math score and grade point average?

      i. Answer Yes or No.

      ii. Fill in the blank. Students who did better on the Math test tended to have _____ first-year grade point average.

      iii. Do you reject $H_0 : \beta_1 = 0$?

      iv. Are the results statistically significant? Answer Yes or No.

      v. What is the p-value? The answer can be found in *two* places on your printout.

      vi. What proportion of the variation in first-year grade point average is explained by score on the SAT Math test? The answer is a number from your printout.

      vii. Give a predicted first-year grade point average for a student who got 700 on the Math SAT. The answer is a number you could get with a calculator from your printout.

   (b) Now fit a model with both the Math and Verbal sub-tests.

      i. Give the test statistic, the degrees of freedom and the p-value for each of the following null hypotheses. The answers are numbers from your printout.

         A. $H_0 : \beta_1 = \beta_2 = 0$

         B. $H_0 : \beta_1 = 0$

         C. $H_0 : \beta_2 = 0$

         D. $H_0 : \beta_0 = 0$

[1]This assignment was prepared by Jerry Brunner, Department of Statistics, 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/2101f19

ii. Controlling for Math score, is Verbal score related to first-year grade point average?
   A. Give the value of the test statistic. The answer is a number from your printout.
   B. Give the p-value. The answer is a number from your printout.
   C. Do you reject the null hypothesis?
   D. Are the results statistically significant? Answer Yes or No.
   E. In plain, non-statistical language, what do you conclude? The answer is something about test scores and grade point average.

iii. Allowing for Verbal score, is Math score related to first-year grade point average?
   A. Give the value of the test statistic. The answer is a number from your printout.
   B. Give the p-value. The answer is a number from your printout.
   C. Do you reject the null hypothesis?
   D. Are the results statistically significant? Answer Yes or No.
   E. In plain, non-statistical language, what do you conclude? The answer is something about test scores and grade point average.

iv. Give a predicted first-year grade point average for a student who got 650 on the Verbal and 700 on the Math SAT.

v. Let's do one more test. We want to know whether expected GPA increases faster as a function of the Verbal SAT, or the Math SAT. That is, we want to compare the regression coefficients, testing $H_0 : \beta_1 = \beta_2$.
   A. Express the null hypothesis in matrix form as $\mathbf{L}\boldsymbol{\beta} = \mathbf{h}$.
   B. Carry out an F test. Feel free to use my `ftest` function (from lecture) if you wish.
   C. State your conclusion in plain, non-technical language. It's something about first-year grade point average.

2. Let $X_1, \ldots, X_n$ be a random sample from a distribution with density

$$f(x) = \frac{\theta e^{\theta(x-\mu)}}{(1+e^{\theta(x-\mu)})^2}$$

for x real, where $-\infty < \mu < \infty$ and $\theta > 0$. Here are some numerical data:

```
4.82  3.66  4.39  1.66  3.80  4.69  1.73  4.50  9.29  4.05  4.50 -0.64  1.40
4.18  2.70  5.65  5.47  0.55  4.64  1.19  2.28  7.16  4.80  3.19  2.33  2.57
2.31  0.35  2.81  2.35  2.52  3.44  2.71 -1.43  7.61  0.93  2.52  6.86  6.14
4.37  3.79  5.04  4.50  1.92  3.25 -0.06  2.81  3.09  2.95  3.69
```

You can read the data from
http://www.utstat.toronto.edu/~brunner/data/legal/mystery.data.txt.

(a) Find the maximum likelihood estimates of μ and θ.

(b) Obtain an approximate 95% confidence interval for θ.

(c) Test $H_0 : \mu = 0$ at the $\alpha = 0.05$ significance level with a large-sample Z test.

3. Looking at the expression for the multivariate normal likelihood on the formula sheet, how can you tell that for *any* fixed positive definite $\boldsymbol{\Sigma}$, the likelihood is greatest when $\boldsymbol{\mu} = \overline{\mathbf{y}}$?

4. Based on a random sample of size n from a p-dimensional multivariate normal distribution, derive a formula for the large-sample likelihood ratio test statistic G^2 for the null hypothesis that $\boldsymbol{\Sigma}$ is diagonal (all covariances between variables are zero). You may use the likelihood function on the formula sheet. You may also use without proof the fact that the unrestricted MLE is $\widehat{\theta} = (\overline{\mathbf{y}}, \widehat{\boldsymbol{\Sigma}})$.

   Hint: Because zero covariance implies independence for the multivariate normal, the joint density is a product of marginals under H_0. To be direct, I am suggesting that you *not* use the likelihood function on the formula sheet to calculate the numerator of the likelihood ratio. You'll eventually get the right answer if you insist on doing it that way, but it's a lot more work.

5. The file http://www.utstat.toronto.edu/~brunner/data/illegal/bp.data.txt has diastolic blood pressure, education, cholesterol, number of cigarettes per day and weight in pounds for a sample of middle-aged men. There are missing values; `summary` on the data frame will tell you what they are.

   Assuming multivariate normality and using R, carry out a large-sample likelihood ratio test to determine whether there are any non-zero covariances among just these three variables: education, number of cigarettes, and weight. Guided by the usual $\alpha = 0.05$ significance level, what do you conclude? Is there evidence that the three variables are related (non-independent)? Answer Yes or No. For this question, let's agree that we will base the sample covariance matrix only on *complete observations*. That is, there will be no missing values on any variable. Don't forget that $\widehat{\boldsymbol{\Sigma}}$, like $\widehat{\sigma}^2_j$, has n in the denominator and not $n-1$. What is n?

6. Here is a useful variation on Problem 4. Suppose n independent and identically data vectors $\mathbf{d}_i = \begin{pmatrix} \mathbf{x}_i \\ \mathbf{y}_i \end{pmatrix}$ are multivariate normal. The notation means that $\mathbf{d}_i$ is $\mathbf{x}_i$ stacked on top of $\mathbf{y}_i$. For example, $\mathbf{x}_i$ could be physical measurements and $\mathbf{y}_i$ could be psychological measurements. Derive a likelihood ratio test to determine whether $\mathbf{x}_i$ and $\mathbf{y}_i$ are independent. Your answer is a formula for G^2 and a formula for the degrees of freedom. Part of the job here is to make up good, simple notation.

7. I said there would be just one more question, but this one does not count. I'ts just set-up for Question 8.

   (a) Let X and Y be scalar random variables and let a be a constant. Using the definition $Cov(X,Y) = E\{(X-\mu_x)(Y-\mu_y)\}$, show $Cov(aX,Y) = a\,Cov(Y,Y)$.

   (b) Let X_1, X_2, Y_1 and Y_2 be scalar random variables. Show

   $$Cov(X_1+X_2,\, Y_1+Y_2) = Cov(X_1,Y_1) + Cov(X_1,Y_2) + Cov(X_2,Y_1) + Cov(X_2,Y_2)$$

   Clearly, this rule extends to larger numbers of terms. Don't hesitate to use it.

   (c) Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be a random sample from a bivariate distribution with $E(X_i) = \mu_x$, $E(Y_i) = \mu_y$, $Var(X_i) = \sigma^2_x$, $Var(Y_i) = \sigma^2_y$, and $Cov(X_i,Y_i) = \sigma_{xy}$. Find $Cov(\bar{X},\bar{Y})$.

8. Independently for $i = 1, \ldots, n$, let $Y_i = \beta X_i + \epsilon_i$, where $E(X_i) = \mu \neq 0$, $E(\epsilon_i) = 0$, $Var(X_i) = \sigma_x^2$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and ϵ_i is independent of X_i. Some data from this model are at http://www.utstat.toronto.edu/~brunner/data/legal/xy.data.txt.

   Let
   $$\widehat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} \text{ and } \widehat{\beta}_2 = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i}.$$
   You know that both of these estimators are consistent; there is no need to prove it again.

   (a) $\widehat{\beta}_1$ is just the ordinary least-squares estimator, though's not technically least squares anymore because the explanatory variable is random. Using R's `lm` function, fit a regression line through the origin, obtain a numerical $\widehat{\beta}_1$, and calculate a 95% confidence interval for β. It's not easy to justify the confidence interval formally, because the error terms are definitely not normal. But please do it anyway for comparison.

   (b) Now calculate $\widehat{\beta}_2$ for these data, and obtain the corresponding large-sample 95% confidence interval for β. Here is a bit of discussion to help you along.

   You're definitely going to use the delta method, but please don't over-think it. It's possible to walk down a dark path here by finding asymptotically normal estimators of *all* the parameters, and seeking an estimate of their asymptotic covariance matrix. Instead, retreat to the model of Question 7c, and realize that for reasons of your own, you are looking at a function $g(\mu_x, \mu_y)$. The asymptotic covariance matrix of $(\bar{x}, \bar{y})$ is actually easy, because you can obtain the exact covariance matrix for any n. That's why you did Question 7. I used the `var` function to get an estimate of $cov \begin{pmatrix} x_i \\ y_i \end{pmatrix}$. Finally, the lower limit of my confidence interval was 1.474.