STA 2101/442 Assignment 7[1]

The non-computer questions on this assignment are practice for the quiz on Friday October 27th, and are not to be handed in. Please do the problems using the formula sheet as necessary. A copy of the formula sheet will be distributed with the quiz. As usual, you may use anything on the formula sheet unless you are directly asked to prove it.

1. 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 $\boldsymbol{\mu} = \overline{\mathbf{y}}$?

2. 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.

3. 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` 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 the five variables; guided by the usual $\alpha = 0.05$ significance level, what do you conclude? Are the five variables all independent of one another? 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? *Bring your printout to the quiz.*

4. Here is a useful variation on Problem 2. Suppose n independent and identically data vectors $\mathbf{d}_i = \begin{pmatrix} \mathbf{x}_i \\ \hline \mathbf{y}_i \end{pmatrix}$ 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.

5. Dead pixels are a big problem in manufacturing computer and cell phone screens. The physics of the manufacturing process dictates that dead pixels happen according to a spatial Poisson process, so that the numbers of dead pixels in cell phone screens are independent Poisson random variables with parameter λ, the expected number of dead pixels. Naturally, λ depends on details of how the screens are manufactured.

   In an effort to reduce the expected number of dead pixels, six assembly lines were set up, each with a different version of the manufacturing process. A random sample of 50 phones was taken from each assembly line and sent to the lab for testing. Mysteriously, three phones from one assembly line disappeared in transit, and 15 phones from another assembly line disappeared. Sample sizes and sample mean numbers of dead pixels appear in the table below.

	Manufacturing Process					
	1	2	3	4	5	6
ybar	10.68	9.87234	9.56	8.52	10.48571	9.98
n	50	47	50	50	35	50

   (a) The first task is to carry out a large sample likelihood ratio test to see whether the expected numbers of dead pixels are different for the six manufacturing processes. Using R, calculate the test statistic and the p-value. Also report the degrees of freedom.

   You are being asked for a computation, but *most of the task is thinking and working things out on paper*. I got away with only five lines of code: One line to enter the means, one line to enter the sample sizes, one line to compute G^2, one line to compute the p-value, and one other line. Here are some little questions to get you started.

      i. Is this a between-cases design or a within-cases design?
      ii. Denote the parameter vector by $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^\top$. What is p?
      iii. What is the null hypothesis?
      iv. What is the distribution of a sum of independent Poisson random variables?
      v. What is the distribution of $n_j \overline{Y}_j$?
      vi. What is the likelihood function? Write it down and simplify.
      vii. What is the unrestricted MLE $\widehat{\boldsymbol{\lambda}}$? It's a vector. Work it out if you need to.
      viii. What is the restricted MLE $\widehat{\boldsymbol{\lambda}}_0$? It's a vector. Work it out if you need to.
      ix. Now you are ready to write the test statistic. There are a lot of cancellations. Keep simplifying!
      x. Now use R to compute the test statistic and p-value. For comparison, my p-value is 0.01169133. *Bring your printout to the quiz.*

   (b) Clearly we need to follow up this result to see where it came from, but we'll do Wald tests because they are a little easier. As preparation (and to get some exercise), carry out a Wald test of the overall null hypothesis you just tested above.

You'll need an estimated asymptotic covariance matrix of $\widehat{\boldsymbol{\lambda}}$, and in this case expressing it in closed form is easier than obtaining and inverting the Hessian. So please do it the easy way. The questions below serve as a guide.

i. The asymptotic variance of $\overline{Y}_j$ is just its variance. What is $Var(\overline{Y}_j)$?

ii. The *estimated* asymptotic variance of $\overline{Y}_j$ is the most natural thing you can imagine. What is it?

iii. So what's the estimated asymptotic variance-covariance matrix of the random vector $\widehat{\boldsymbol{\lambda}}$?

Now you can carry out the Wald test. Do it with R, obtaining the test statistic, the degrees of freedom and the p-value. *Bring your printout to the quiz.*

(c) Finally carry out all pairwise comparisons with a Bonferroni correction, protecting the entire family of tests against Type I error at the *joint* $\alpha = 0.05$ significance level. In plain language, what do you conclude? Remember, with pairwise comparisons you can always draw directional conclusions. *Bring your printout to the quiz.*

6. If two events have equal probability, the odds ratio equals ____.

7. For a multiple logistic regression model, if the value of the kth explanatory variable is increased by c units and everything else remains the same, the odds of Y=1 are ____ times as great. Prove your answer.

8. For a multiple logistic regression model, let $P(Y_i = 1|x_{i,1}, \ldots, x_{i,p-1}) = \pi(\mathbf{x}_i)$. Show that a linear model for the log odds is equivalent to

$$\pi(\mathbf{x}_i) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}} = \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}$$

9. Write the log likelihood for a general logistic regression model, and simplify it as much as possible. Of course use the result of the last question.

10. A logistic regression model with no explanatory variables has just one parameter, β_0. It also the same probability $\pi = P(Y = 1)$ for each case.

(a) Write π as a function of β_0; show your work.

(b) The *invariance principle* of maximum likelihood estimation says the MLE of a function of the parameter is that function of the MLE. It is very handy. Now, still considering a logistic regression model with no explanatory variables,

i. Suppose $\overline{y}$ (the sample proportion of $Y = 1$ cases) is 0.57. What is $\widehat{\beta}_0$? Your answer is a number.

ii. Suppose $\widehat{\beta}_0 = -0.79$. What is $\overline{y}$? Your answer is a number.

11. Consider a logistic regression in which the cases are newly married couples with both people from the same religion. The explanatory variables are total family income and religion. Religion is coded A, B, C and None (let's call "None" a religion), and the response variable is whether the marriage lasted 5 years (1=Yes, 0=No).

    (a) Write a linear model for the log odds of a successful[2] marriage. You do not have to say how the dummy variables are defined. You will do that in the next part.

    (b) Make a table with four rows, showing how you would set up indicator dummy variables for Religion, with None as the reference category.

    (c) Add a column showing the odds of the marriage lasting 5 years. The *symbols* for your dummy variables should not appear in your answer, because they are zeros and ones, and different for each row. But of course your answer contains β values. Denote income by x.

    (d) For a constant value of income, what is the ratio of the odds of a marriage lasting 5 years or more for Religion C to the odds of lasting 5 years or more for No Religion? Answer in terms of the β symbols of your model.

    (e) Holding income constant, what is the ratio of the odds of lasting 5 years or more for religion A to the odds of lasting 5 years or more for Religion B? Answer in terms of the β symbols of your model.

    (f) You want to test whether controlling for income, Religion is related to whether the marriage lasts 5 years. State the null hypothesis in terms of one or more β values.

    (g) You want to know whether marriages from Religion A are more likely to last 5 years than marriages from Religion C, allowing for income. State the null hypothesis in terms of one or more β values.

    (h) You want to test whether marriages between people of No Religion with an average income have a 50-50 chance of lasting 5 years. State the null hypothesis in symbols. To hold income to an "average" value, just set $x = \overline{x}$.

[2] I agree, this may be a modest definition of success.