STA 2053 Assignment 2 (Large-sample and Regression)[1]

The paper and pencil questions are not to be handed in. They are practice for the quiz on October 17th. Bring hard copy of your input and output for Question 10d to the quiz. It may be handed in.

1. Let $X_1, \ldots, X_n$ be a random sample from a Binomial distribution with parameters 3 and θ. That is,
$$P(X_i = x_i) = \binom{3}{x_i} \theta^{x_i}(1-\theta)^{3-x_i},$$
for $x_i = 0, 1, 2, 3$. Choose a reasonable estimator of θ, and prove that it is strongly consistent. Where you get your estimator does not really matter, but please state how you thought of it. This question can be quick.

2. Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with density
$$f(x;\tau) = \frac{\tau^{1/2}}{\sqrt{2\pi}}\, e^{-\frac{\tau x^2}{2}},$$
where the parameter $\tau > 0$. Let
$$\widehat{\tau} = \frac{n}{\sum_{i=1}^n X_i^2}.$$
Is $\widehat{\tau}$ consistent for τ? Answer Yes or No and prove your answer. Hint: You can just write down $E(X^2)$ by inspection. This is a very familiar distribution; have confidence!

3. Let $X_1, \ldots, X_n$ be a random sample from a Gamma distribution with $\alpha = \beta = \theta > 0$. That is, the density is
$$f(x;\theta) = \frac{1}{\theta^\theta \Gamma(\theta)} e^{-x/\theta} x^{\theta-1},$$
for $x > 0$. Let $\widehat{\theta} = \overline{X}_n$. Is $\widehat{\theta}$ consistent for θ? Answer Yes or No and prove your answer.

4. Independently for $i = 1, \ldots, n$, let
$$Y_i = \beta X_i + \epsilon_i,$$
where $E(X_i) = \mu$, $E(\epsilon_i) = 0$, $Var(X_i) = \sigma^2_x$, $Var(\epsilon_i) = \sigma^2_\epsilon$, and ϵ_i is independent of X_i. Let
$$\widehat{\beta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}.$$
Is $\widehat{\beta}$ consistent for β? Answer Yes or No and prove your answer.

5. Another Method of Moments estimator for Problem 4 is $\widehat{\beta}_2 = \frac{\overline{Y}_n}{\overline{X}_n}$.
   (a) Show that $\widehat{\beta}_2 \stackrel{p}{\rightarrow} \beta$ in most of the parameter space.
   (b) However, consistency means that the estimator converges to the parameter in probability *everywhere* in the parameter space. Where does $\widehat{\beta}_2$ fail, and why?

[1]This assignment was prepared by Jerry Brunner, Department of Statistical Sciences, University of Toronto. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/brunner/oldclass/2053f22

6. Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be a random sample from a bivariate distribution with $E(X_i)=\mu_x$, $E(Y_i)=\mu_y$, $Var(X_i)=\sigma_x^2$, $Var(Y_i)=\sigma_y^2$, and $Cov(X_i,Y_i)=\sigma_{xy}$.

   (a) Show that the sample covariance $S_{xy}=\frac{\sum_{i=1}^n(X_i-\overline{X})(Y_i-\overline{Y})}{n-1}$ is a consistent estimator of σ_{xy}.

   (b) Show that the sample covariance (with n in the denominator, just for convenience) has a large-sample normal distribution. Give the asymptotic mean and covariance. Cite the Slutsky lemmas as you use them. They will be supplied with the quiz if necessary.

7. The usual univariate multiple regression model with independent normal errors is

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon},$$

   where $\mathbf{X}$ is an $n\times p$ matrix of known constants, $\boldsymbol{\beta}$ is a $p\times 1$ vector of unknown constants, and $\boldsymbol{\epsilon}$ is multivariate normal with mean zero and covariance matrix $\sigma^2\mathbf{I}_n$, with $\sigma^2>0$ an unknown constant. But of course in practice, the explanatory variables are random, not fixed. Clearly, if the model holds *conditionally* upon the values of the explanatory variables, then all the usual results hold, again conditionally upon the particular values of the explanatory variables. The probabilities (for example, p-values) are conditional probabilities, and the F statistic does not have an F distribution, but a conditional F distribution, given $\mathbf{X}=\mathbf{x}$.

   (a) Show that the least-squares estimator $\widehat{\boldsymbol{\beta}}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is conditionally unbiased.

   (b) Show that $\widehat{\boldsymbol{\beta}}$ is also unbiased unconditionally.

   (c) A similar calculation applies to the significance level of a hypothesis test. Let F be the test statistic (say for an extra-sum-of-squares F-test), and f_c be the critical value. If the null hypothesis is true, then the test is size α, conditionally upon the explanatory variable values. That is, $P(F>f_c|\mathbf{X}=\mathbf{x})=\alpha$. Find the *unconditional* probability of a Type I error. Assume that the explanatory variables are discrete, so you can write a multiple sum.

8. In the following regression model, the explanatory variables X_1 and X_2 are random variables. The true model is

$$Y_i=\beta_0+\beta_1X_{i,1}+\beta_2X_{i,2}+\epsilon_i,$$

   independently for $i=1,\ldots,n$, where $\epsilon_i\sim N(0,\sigma^2)$.

   The mean and covariance matrix of the explanatory variables are given by

$$E\left(\begin{array}{c}X_{i,1}\\X_{i,2}\end{array}\right)=\left(\begin{array}{c}\mu_1\\\mu_2\end{array}\right)\quad\text{and}\quad Var\left(\begin{array}{c}X_{i,1}\\X_{i,2}\end{array}\right)=\left(\begin{array}{cc}\phi_{11}&\phi_{12}\\\phi_{12}&\phi_{22}\end{array}\right)$$

   Unfortunately $X_{i,2}$, which has an impact on Y_i and is correlated with $X_{i,1}$, is not part of the data set. Since $X_{i,2}$ is not observed, it is absorbed by the intercept and error term, as follows.

$$\begin{aligned}Y_i &= \beta_0+\beta_1X_{i,1}+\beta_2X_{i,2}+\epsilon_i\\ &= (\beta_0+\beta_2\mu_2)+\beta_1X_{i,1}+(\beta_2X_{i,2}-\beta_2\mu_2+\epsilon_i)\\ &= \beta_0'+\beta_1X_{i,1}+\epsilon_i'.\end{aligned}$$

   The primes just denote a new β_0 and a new ϵ_i. It was necessary to add and subtract $\beta_2\mu_2$ in order to obtain $E(\epsilon_i')=0$. And of course there could be more than one omitted variable. They would all get swallowed by the intercept and error term, the garbage bins of regression analysis.

   (a) Make a path diagram of this model.

(b) What is $Cov(X_{i,1}, \epsilon_i')$?

(c) Calculate the variance-covariance matrix of $(X_{i,1}, Y_i)$ under the true model.

(d) Suppose we want to estimate β_1. The usual least squares estimator is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_{i,1} - \overline{X}_1)(Y_i - \overline{Y})}{\sum_{i=1}^n (X_{i,1} - \overline{X}_1)^2}.$$

You may just use this formula; you don't have to derive it. Is $\widehat{\beta}_1$ a consistent estimator of β_1 (meaning for all points in the parameter space) if the true model holds? Answer Yes or No and show your work. Remember, X_2 is not available, so you are doing a regression with one explanatory variable. You may use the consistency of the sample variance and covariance without proof.

(e) Are there *any* points in the parameter space for which $\widehat{\beta}_1 \stackrel{p}{\rightarrow} \beta_1$ when the true model holds?

9. Ordinary least squares is often applied to data sets where the explanatory variables are best modeled as random variables.

   (a) In the usual regression model with normal errors, what is the conditional distribution of ϵ_i given $\mathbf{X}_i = \mathbf{x}_i$?

   (b) In what way does the usual conditional linear regression model imply that (random) explanatory variables have zero covariance with the error term? Hint: Assume $\mathbf{X}_i$ as well as ϵ_i continuous to make the notation easier.

   (c) Show that for simple regression (one explanatory variable), $E(\epsilon_i | X_i = x_i) = 0$ for all x_i implies $Cov(X_i, \epsilon_i) = 0$, so that a standard regression model without the normality assumption still implies zero covariance (though not necessarily independence) between the error term and explanatory variables. I did a double expectation conditioning on X_i.

   (d) Given the results of Problem 8, is it ever safe to assume that random explanatory variables have zero covariance with the error term?

10. Women and men are coming into a store according to independent Poisson processes with rates λ_1 for women and λ_2 for men. You don't have to know anything about Poisson processes to do this question. We have that the number of women and the number of men entering the store in a given time period are independent Poisson random variables, with expected values λ_1 for women and λ_2 for men. Because the Poisson process is an independent increments process, we can treat the numbers from n time periods as a random sample.

    Management wants to know the expected number of male customers and the expected number of female customers. Unfortunately, the total numbers of customers were recorded, but not their sex. Let $y_1, \ldots, y_n$ denote the total numbers of customers who enter the store in n time periods. That's all the data we have.

    (a) What is the distribution of y_i? If you know the answer, just write it down without proof.

    (b) What is the parameter space?

    (c) Find the MLE of the parameter vector (λ_1, λ_2) analytically. Show your work.

    (d) For the data in `https://www.utstat.toronto.edu/brunner/openSEM/data/poisson.data.txt`, find the MLE numerically. *Try two different starting values.* Does your answer agree with your answer to 10c?

Please bring your *complete* R printout from Question 10d to the quiz, showing all input and output. It may be handed in.