STA 431s17 Assignment Four[1]

This assignment is based mostly on lecture units Six (SAS Example 2) and Seven (Omitted Variables and Instrumental Variables). Also see Sections 0.5 and 0.6 in Chapter Zero of the text. The non-computer parts of this assignment are just practice for the quiz. They are not to be handed in.

1. This question is a deliberate repeat from last week. Starting with the multivariate normal density on the formula sheet, derive the multivariate normal likelihood, also on the formula sheet. You will use $tr(\mathbf{AB}) = tr(\mathbf{AB})$ and other properties of the trace.

2. The `statclass` data consist of Quiz average, Computer assignment average, Midterm score and Final Exam score from a statistics class, long ago. The first three variables are explanatory, and final exam score is the response variable. Data are in the plain text file

   http://www.utstat.utoronto.ca/~brunner/data/legal/LittleStatclassdata.txt.

   Fit a standard regression model using `proc calis`. Please make sure to use the `vardef=n` and `nostand` options. All the variables are observed, and the explanatory variables could be correlated with one another but they are independent of the error term. So that our regression coefficients will mean the same thing, please use the order of explanatory variables in the data file. Be able to answer questions like the following.

   (a) What is $\widehat{\beta}_0$? The answer is a number on your printout.

   (b) What is the test statistic for test $H_0 : \beta_2 = 0$? What is the p-value? The answers are numbers on your printout.

   (c) What is the predicted Final Exam score for a student with a Quiz average of 8.5, a Computer average of 5, and a Midterm mark of 60%? The answer is a number. Be able to do this kind of thing on the quiz with a calculator.

   (d) For any fixed Quiz Average and Computer Average, a score one point higher on the Midterm yields a predicted mark on the Final Exam that is _____ higher.

   (e) From your `proc calis` output, what is the estimated covariance between Quiz average and Computer average? The answer is a number on your printout.

   (f) Your regression model should have an error term. What is its estimated variance? The answer is a number from your `proc calis` output.

   Please bring printouts of your log file and results file to the quiz; you may be asked to hand them in. Make sure your name and student number appears on both files, preferably using a `title` or `title2` statement. **Do not write anything on your printouts** except possibly your name and student number if you forgot to put them in your code.

3. Ordinary least squares is often applied to data sets where the explanatory variables are best modeled as random variables. 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. What is the conditional distribution of ϵ_i given $\mathbf{X}_i = \mathbf{x}_i) = 0$?

4. In a regression with one explanatory variable, show that $E(\epsilon_i|X_i = x_i) = 0$ 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. Hint: the matrix version of this calculation is in the text.

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

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i,$$

independently for $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, \sigma^2)$ and is independent of $X_{i,1}$ and $X_{i,2}$.

The explanatory variables have a bivariate normal distribution with

$$E\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad cov\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}$$

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.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i \\ &= (\beta_0 + \beta_2\mu_2) + \beta_1 X_{i,1} + (\beta_2 X_{i,2} - \beta_2\mu_2 + \epsilon_i) \\ &= \beta_0' + \beta_1 X_{i,1} + \epsilon_i'. \end{aligned}$$

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

   (b) Calculate $Cov(X_{i,1}, Y_i)$. Is it the same under the true model and the re-parameterized model?

   (c) 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 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.

   (d) What is the parameter space under the true model?

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

6. If a parameter is a function of the distribution of the observable data, it is said to be *identifiable*. You know that if $X \sim N(\mu, \sigma^2)$, then $E(X) = \mu$ and $Var(X) = \sigma^2$. Why does this tell you that the parameters of the normal model are identifiable?

7. Men and women are calling a technical support line according to independent Poisson processes with rates λ_1 and λ_2 per hour. Data for 144 hours are available, but unfortunately the sex of the caller was not recorded. All we have is the number of callers for each hour, which is distributed Poisson$(\lambda_1 + \lambda_2)$; just use this; you don't have to show it. The parameter in this problem is $\boldsymbol{\theta} = (\lambda_1, \lambda_2)$.

   (a) Try to find the MLE by differentiating. Show your work. Are there any points in the parameter space where both partial derivatives are zero? Why did estimation fail for this fairly realistic model?

   (b) To show that the parameters of a model are not identifiable, all you need to do is find two different sets of parameter values that yield the same distribution of the observable data. Then, the parameter vector cannot possibly be a function of the distribution of the observable data. Use this to show that the parameters in this problem are not identifiable. A simple numerical example is enough, and in fact it is best.

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

$$\begin{pmatrix} X_i \\ \epsilon_i \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_x \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & c \\ c & \sigma_\epsilon^2 \end{pmatrix}\right)$$

   The observable data are $\mathbf{D}_1, \ldots, \mathbf{D}_n$, where $\mathbf{D}_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix}$.

   (a) Draw a path diagram for this model.

   (b) What is the distribution of $\mathbf{D}_i$? Hint: If $\mathbf{w} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, what is the distribution of $\mathbf{A}\mathbf{w}$? What is $\mathbf{A}$ for this problem?

   (c) What is the parameter space? Note that c cannot be just anything.

   (d) Give the formula for a Method of Moments estimate of β. Is it consistent? That is, does it converge in probability to the right answer *everywhere* in the parameter space?

   (e) For this model, showing parameter identifiability consists of solving five equations in five unknowns. What are the equations? The lecture slide entitled "Five equations in six unknowns" should help.

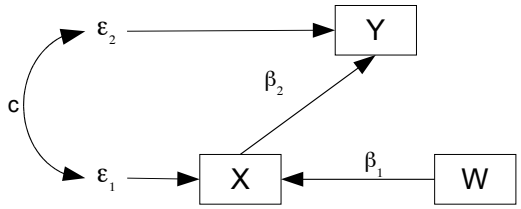
   (f) For some points in the parameter space these equations can be solved, and for others they cannot. Where can they be solved? You don't have to literally give the solutions.

   (g) As in Question 7, give a numerical example of two different parameter vectors that yield the same distribution of $\mathbf{D}_i$. Stay in the parameter space. What are the mean vector and covariance matrix of $\mathbf{D}_i$ for your example? This shows that the parameter is not technically identifiable, even though things are okay in most of the parameter space. It suggests that we need a pointwise definition of parameter identifiability, which is given later in Chapter Zero.

9. For a simple instrumental variables model, the model equations are

$$\begin{aligned} X_i &= \alpha_1 + \beta_1 W_i + \epsilon_{i1} \\ Y_i &= \alpha_2 + \beta_2 X_i + \epsilon_{i2} \end{aligned}$$

and the path diagram is



(a) Calculate the expected value vector and covariance matrix of the observable data.

(b) Is the parameter β_1 identifiable? Answer Yes or No and prove it.

(c) Give the formula for a Method of Moments estimate of the covariance parameter c in terms of $\widehat{\sigma}_{ij}$ values.

(d) This is also the maximum likelihood estimate. Why?