STA 302f20 Assignment Twelve[1]

The following problems are not to be handed in. They are preparation for the final exam.

1. 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 that the vector of explanatory variables $\mathbf{x}_i$ as well as ϵ_i is continuous. For the usual linear regression model with normal errors, what is the conditional distribution of ϵ_i given $\mathbf{x}_i$?

2. In a regression with one random 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. For convenience, you may assume that the distributions are continuous, so you can integrate.

3. In the usual multiple regression model, the $\mathbf{X}$ matrix is an $n \times (k+1)$ matrix of known constants. But in practice, the independent variables are often random and not fixed. Clearly, if the model holds *conditionally* upon the values of the independent variables, then all the usual results hold, again conditionally upon the particular values of the independent 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 $\mathcal{X} = \mathbf{X}$. Here, the $n \times (k+1)$ matrix $\mathcal{X}$ is used to denote the matrix containing the random explanatory variables. It does not have to be *all* random. For example the first column might contain only ones if the model has an intercept.

   (a) Show that the least-squares estimator $\widehat{\boldsymbol{\beta}} = (\mathcal{X}'\mathcal{X})\mathcal{X}'\mathbf{y}$ is unbiased, conditionally upon $\mathcal{X} = \mathbf{X}$. You've done this before with a slightly different notation.

   (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 F-test comparing full and reduced models), and f_c be the critical value. If the null hypothesis is true, then the test is size α, conditionally upon the independent variable values. That is, $P(F > f_c|\mathcal{X} = \mathbf{X}) \stackrel{H_0}{=} \alpha$. Using the Law of Total Probability (see lecture slides), find the *unconditional* probability of a Type I error. Assume that the explanatory variables are discrete, so you can write a multiple sum.

4. Consider the following model with random predictor variables. Independently for $i = 1, \ldots, n$,

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i \\ &= \alpha + \boldsymbol{\beta}' \mathbf{x}_i + \epsilon_i, \end{aligned}$$

where

$$\mathbf{x}_i = \left(\begin{array}{c} x_{i1} \\ \vdots \\ x_{ik} \end{array} \right)$$

and $\mathbf{x}_i$ is independent of ϵ_i.

Note that in this notation, α is the intercept, and $\boldsymbol{\beta}$ does not include the intercept. The "independent" variables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})'$ are not statistically independent. They have the symmetric and positive definite $k \times k$ covariance matrix $\boldsymbol{\Sigma}_x = [\sigma_{ij}]$, which need not be diagonal. They also have the $k \times 1$ vector of expected values $\boldsymbol{\mu}_x = (\mu_1, \ldots, \mu_k)'$.

(a) Let $\boldsymbol{\Sigma}_{xy}$ denote the $k \times 1$ matrix of covariances between y_i and x_{ij} for $j = 1, \ldots, k$. Calculate $\boldsymbol{\Sigma}_{xy} = cov(\mathbf{x}_i, y_i)$. Stay with matrix notation and don't expand.

(b) From the equation you just obtained, solve for $\boldsymbol{\beta}$ in terms of $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_{xy}$.

(c) Based on your answer to the last part and and letting $\widehat{\boldsymbol{\Sigma}}_x$ and $\widehat{\boldsymbol{\Sigma}}_{xy}$ denote matrices of *sample* variances and covariances, what would be a reasonable estimator of $\boldsymbol{\beta}$ that you could calculate from sample data?

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 $\epsilon_i \sim N(0, \sigma^2)$ is independent of $x_{i,1}$ and $x_{i,2}$.

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

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

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.

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i \\ &= (\beta_0 + \beta_2 \mu_2) + \beta_1 x_{i,1} + (\beta_2 x_{i,2} - \beta_2 \mu_2 + \epsilon_i) \\ &= \beta_0^* + \beta_1 x_{i,1} + \epsilon_i^*. \end{aligned}$$

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 $Cov(x_{i,1}, \epsilon_i^*)$? This is a scalar calculation.

(b) Calculate $Cov(x_{i,1}, y_i)$. This is another scalar calculation. Is it possible to have non-zero covariance between $x_{i,1}$ and y_i when $\beta_1 = 0$?

(c) Suppose we want to estimate β_1 using the usual least squares estimator $\widehat{\beta}_1$ (see formula sheet). As $n \to \infty$, does $\widehat{\beta}_1 \to \beta_1$? You may use the fact that like sample means, sample variances and covariances converge to the corresponding Greek-letter versions as $n \to \infty$ (except possibly on a set of probability zero) like ordinary limits, and all the usual rules of limits apply. So for example, defining $\widehat{\sigma}_{xy}$ as $\frac{1}{n-1}\sum_{i=1}^n (x_{i,1} - \overline{x}_1)(y_i - \overline{y})$, we have $\widehat{\sigma}_{xy} \to Cov(x_i, y_i)$.

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

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

The observable data are $\mathbf{D}_1, \ldots, \mathbf{D}_n$, where $\mathbf{D}_i = \left(\begin{array}{c} X_i \\ Y_i \end{array}\right)$.

(a) Draw a path diagram for this model.

(b) What is the distribution of $\mathbf{D}_i$?

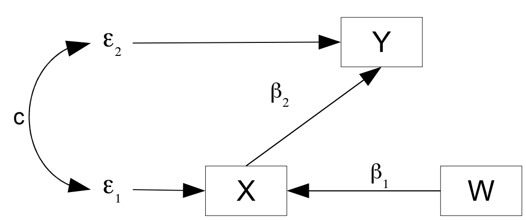
(c) What is the vector of parameters for this model? What are their possible values?

(d) Give a numerical example of two different parameter vectors that yield the same distribution of $\mathbf{D}_i$. What are the mean vector and covariance matrix of $\mathbf{D}_i$ for your example?

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

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

and the path diagram is



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

(b) Suggest an estimator for the parameter β_1. Does your $\widehat{\beta}_1 \to \beta_1$? Why?

(c) Suggest an estimator of the covariance parameter c in terms of $\widehat{\sigma}_{ij}$ values.

(d) Do you have $\widehat{c} \to c$? Why?