Ec. 2120, Fall 2022 Due: September 30, 2022 at 5:00 PM

PROBLEM SET 2

It is fine to discuss the questions with others. Each group should submit one solution set.

Part I. This part is based on a sample of $n = 815$ observations from the Young Men's Cohort of the National Longitudinal Survey. The variables are usual weekly earnings in 1980 (uwe), age in 1980 (age80), years of schooling completed (educ), father's education (fed), mother's education (med), and an IQ score (iq). There is an additional test score (kww) that we will not be using. The data are in the file nls.mat in the Assignments section of the course web site. Once you are in Matlab, type the command

```
load nls
```

Now the series uwe, age80, educ, fed, med, iq, kww will be available. We will be using a (potential) labor market experience variable defined as age - schooling - 6.

1. (a) Calculate the least squares regression of log(earnings) on a constant, schooling, experience, and experience squared. (Multiply the coefficients by 100 to make them easier to read.)

(b) Now add IQ to the regression and calculate the coefficients. Show how the results so far are sufficient to calculate the coefficient on schooling in a regression of IQ on a constant, schooling, experience, and experience squared. Then run this third regression to check your answer.

2. The IQ coefficient in 1(b) can be obtained from a simple regression of log(earnings) on a single variable w. What is w? Construct w and run the regression to check your answer.

3. (a) Calculate the coefficients in a regression of log(earnings) on a constant, schooling, experience, experience squared, IQ, father's education, and mother's education. Discuss the magnitudes of the coefficients. (The IQ scores are constructed to have a normal distribution with a mean of 100 and a standard deviation ($= \sqrt{V(IQ)}$) of 15.)

(b) Compare the schooling coefficient in (a) with the schooling coefficient from the regression in 1(a), which did not include family background variables or IQ. Discuss the magnitude of the change in the coefficient.

4. Derive the asymptotic distribution of the least squares coefficient assuming iid random sampling. Derive a 95% confidence interval for the returns to schooling coefficient. Does this interval change when we recompute it assuming homoskedasticity?

PART II. Suppose that Z_1 and Z_2 are binary random variables: Z_1 takes on only the values 0 and 1, and Z_2 takes on only the values 0 and 1. Consider the (population) linear predictor of Y given $1, Z_1, Z_2, Z_1 \cdot Z_2$:

$$E^*(Y \mid 1, Z_1, Z_2, Z_1 \cdot Z_2) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 \cdot Z_2.$$

(We can think of (Y, Z_1, Z_2) as corresponding to a single random draw from the population.)

(a) Does

$$E(Y \mid Z_1, Z_2) = E^*(Y \mid 1, Z_1, Z_2, Z_1 \cdot Z_2)?$$

Explain.

(b) Suppose that data (Y_i, Z_{i1}, Z_{i2}) are available from a random sample of $i = 1, \ldots, n$ individuals. The following four sample means have been tabulated:

$$\bar{Y}_{00}, \quad \bar{Y}_{01}, \quad \bar{Y}_{10}, \quad \bar{Y}_{11},$$

where

$$\bar{Y}_{lm} = \frac{\sum_{i=1}^{n} Y_i 1(Z_{i1} = l, Z_{i2} = m)}{\sum_{i=1}^{n} 1(Z_{i1} = l, Z_{i2} = m)} \qquad (l, m = 0, 1).$$

($1(B)$ is the indicator function that equals 1 if the event B occurs and equals 0 otherwise.) Use these means to provide an estimate of β_3. Is this a consistent estimator of β_3 as $n \to \infty$? Explain.

Part III. Consider the following model for measurement error:

$$\begin{aligned} E^*(Y_i \mid 1, \tilde{Z}_i, Z_{i1}, Z_{i2}) &= \beta_0 + \beta_1 \tilde{Z}_i \\ E^*(Z_{i1} \mid 1, \tilde{Z}_i, Z_{i2}) &= \tilde{Z}_i \\ E^*(Z_{i2} \mid 1, \tilde{Z}_i, Z_{i1}) &= \tilde{Z}_i, \end{aligned}$$

where Z_{i1} and Z_{i2} are two noisy measurements on the true value $\tilde{Z}_i$. $\tilde{Z}_i$ is a *latent variable*. The population model is expressed in terms of the vector of random variables

$$Q_i = (Y_i, Z_{i1}, Z_{i2}, \tilde{Z}_i).$$

Assume that the Q_i are independent and identically distributed (i.i.d.) according to some unknown distribution (for $i = 1, \ldots, n$). We have observations on

$$D_i = (Y_i, Z_{i1}, Z_{i2})$$

for $i = 1, \ldots, n$. Data on $\tilde{Z}_i$ are not available.

(a) Work out the covariance matrix for (Y_i, Z_{i1}, Z_{i2}) as a function of β_1, $\mathrm{Var}(\tilde{Z}_i)$, and some additional parameters that you will need to define. (It may be helpful to define prediction errors, with

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \tilde{Z}_i + U_i \\ Z_{i1} &= \tilde{Z}_i + V_{i1}, \\ Z_{i2} &= \tilde{Z}_i + V_{i2}, \end{aligned}$$

and to show that these prediction errors are uncorrelated with each other.)

(b) A parameter is *identified* if it is determined by the population distribution of the observable D_i. Show that β_1 is identified by expressing it as a function of the elements of the covariance matrix in (a).

(c) Suggest an estimator for β_1 and show that it is consistent.

Consider a modified version of the above measurement error problem.

In this version, we have the following:

$$E^*(Y_i|1, Z_i^*, Z_i) = \beta_0 + \beta_1 Z_i^*$$

$$E^*(Z_i|1, Z_i^*) = Z_i^*$$

where Z_i is a noisy measurement on the true value Z_i^*. The population model is expressed in terms of the vector of random variables

$$D_i = (Y_i, Z_i, Z_i^*)$$

Assume that the D_i are independent and identically distributed (i.i.d.) according to some unknown distribution. We have observations on

$$W_i = (Y_i, Z_i)$$

for $i = 1, \ldots, n$. Data on Z_i^* are not available. Assume that $\beta_1 > 0$.

(d) Work out the covariance matrix for (Y_i, Z_i) (i.e., express the two variances and one covariance in terms of the model parameters β_1, $Var(A_i)$, $Var(U_i), Var(V_i)$).

(e) Consider the linear predictors

$$E^*(Y_i|1, Z_i) = \pi_o + \pi_1 Z_i$$

$$E^*(Z_i|1, Y_i) = \alpha_o + \alpha_1 Y_i$$

Show that

$$\pi_1 \leq \beta_1 \leq 1/\alpha_1$$

Provide a procedure that uses the data on (Y_i, Z_i) for $i = 1, \ldots, n$ to provide estimates of these lower and upper bounds for β_1. Explain the motivation for your procedure.

Part IV . Show that the sample version of the variance of the least squares estimator is consistent.