Economics 361
Problem Set #7

Jun Ishii [*]
Department of Economics
Amherst College

Fall 2022

Question 1: Returns from Education

Social scientists and policy-makers have for years strived to evaluate the importance of formal education. One measure adopted by some is the impact of formal education on subsequent earnings.

Linear regressions have been used to estimate this "return" from formal education. Usually, the regressions are of the following form:

$$ln(\text{Wage})_i \quad = \quad \beta_0 + X_{1i}'\beta_1 + X_{2i}'\beta_2 + \epsilon_i$$

X_{1i} represents the various educational attainment of worker i and X_{2i} other characteristics of worker i that may determine their wage. The returns to higher education are the coefficients in β_1 associated with the variables in X_{1i} related to higher education.

For example, a researcher many believe that wages were largely determined by (i) years of formal education ($EDUC_i$) (ii) whether you attended a selective "elite" college ($ELITE_i$) (iii) years of experience (EXP_i) (iv) some measure of innate ability ($ABIL_i$)

$$ln(\text{Wage})_i \quad = \quad \gamma_0 + \gamma_1 EDUC_i + \gamma_2 ELITE_i + \gamma_3 EXP_i + \gamma_4 ABIL_i + \epsilon_i$$

In this example, $X_{1i} = \{EDUC_i, ELITE_i\}$ and $X_{2i} = \{EXP_i, ABIL_i\}$. γ_1 would represent the return from an additional year of formal schooling and γ_2 the return from attending a selective college (after controlling for years of formal education).

Read the following academic article

> "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables," by S. Dale & A. Krueger, *Quarterly Journal of Economics*, November 2002, pp.1491-1527 (dk02.pdf)

Answer problems on the following page(s) related to the first part of the article (p.1491-1500).

[*] Office: Converse Hall 315 Phone: (413) 542-2901 E-mail: jishii@amherst.edu

Consider the log wage regressions introduced on pages 1495 and 1496. On the top p.1496, the authors write that " ... and ϵ_i is an idiosyncratic error term that is uncorrelated with the other variables on the right hand side of (2)." This is the authors' way of saying that they assume

$$E[lnW_i|SAT_{j*}, X_{1i}, X_{2i}] \quad = \quad \beta_0 + \beta_1 SAT_{j*} + \beta_2 X_{1i} + \beta_3 X_{2i}$$

(a) Use the above and the Law of Iterated Expectations to derive $E[lnW_i|SAT_{j*}, X_{1i}]$. Your answer should not be identical to the one proposed by the authors near the bottom of p.1496

On p.1496, the authors argue that

$$E[lnW_i|SAT_{j*}, X_{1i}] \quad = \quad \beta_0 + \beta_1 SAT_{j*} + \beta_2 X_{1i} + E[u_i|X_{1i}, \gamma_1 X_{1i} + \gamma_2 X_{2i} + e_{ij*} > C_{j*}]$$

But applying conditional expectation to both sides of equation (3) yields

$$E[lnW_i|SAT_{j*}, X_{1i}] \quad = \quad \beta_0 + \beta_1 SAT_{j*} + \beta_2 X_{1i} + E[u_i|SAT_{j*}, X_{1i}]$$

(b) The authors equate $E[u_i|SAT_{j*}, X_{1i}]$ with $E[u_i|X_{1i}, \gamma_1 X_{1i} + \gamma_2 X_{2i} + e_{ij*} > C_j]$. What must be true about the statistical relationship among $\{X_{1i}, X_{2i}, SAT_{j*}, e_{ij}, \epsilon_i\}$ in order for the authors to make this claim? Briefly explain.

(c) Near the bottom of p.1496, the authors assert that "the coefficient on school-average SAT score is biased upward in this situation." The bias is due to "omitted variables," as discussed in lecture. But why do the authors claim that the bias is **upward**? What, if anything, are the authors assuming about the values of $\{\gamma_1, \gamma_2, \beta_0, \beta_1, \beta_2, \beta_3\}$?

Suppose we observed $E[u_i|SAT_{j*}, X_{1i}]$ up to some unknown scalar multiple. In other words, suppose we observed $X_{3i} = \alpha \, E[u_i|SAT_{j*}, X_{1i}]$ for each observation i. α is the unknown scalar multiple.

(d) Derive $E[lnW_i|SAT_{j*}, X_{1i}, X_{3i}]$

(e) Briefly explain how one could use this additional piece of data, $\{X_{3i}\}_{i=1}^{N}$, to obtain unbiased estimates of $(\beta_0, \beta_1, \beta_2)$

(f) The authors propose to mitigate the omitted variables bias by including "an unrestricted set of dummy variables indicating groups of students who received the same admissions decisions (i.e., the same combination of acceptances and rejections) from the same set of colleges" to wage equation (3). Evaluate this proposal.

Question 2: Measurement Error and the Permanent Income Model

Let us consider the Permanent Income Model as described in Goldberger Chapter 31.2. The model, attributed to Nobel Laureate Milton Friedman, is based on the idea that people *smooth* their consumption over their lifetime. While young, they borrow money against future earnings in order to consume more now (think college loans and mortgage). When in middle age, they consume less than their full contemporary income, paying off debts from their youth and saving for retirement. When old, they consume their savings. Consequently, the consumption decisions of people are determined not so much by income *today* but rather the *lifetime* stream of income. The amortized share of this lifetime income can be thought of as *permanent income.* Therefore, ideally, in order to estimate income elasticity of consumption, we would want to regress consumption (y) not on current income (x) but perceived permanent income (z).

$$y \;=\; \alpha + \beta z + v$$

Unfortunately, we rarely, if ever, observe z. At best, we observe the person's current income, x. Suppose we believe that the relationship between z and x can be described as follows:

$$x \;=\; z + u$$

To complete the Goldberger formulation, further assume that (z, u, v) are jointly distributed according to the following trivariate normal distribution:

$$\begin{pmatrix} z \\ u \\ v \end{pmatrix} \;\sim\; N\left[\begin{pmatrix} \mu \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_z^2 & 0 & 0 \\ 0 & \sigma_u^2 & 0 \\ 0 & 0 & \sigma_v^2 \end{pmatrix} \right]$$

For the following questions, you may want to refer to Goldberger Chapters 7 and 18.

(a) Explain briefly why $\begin{pmatrix} x \\ y \end{pmatrix}$ is distributed bivariate normal

(b) Explain why $E(y \mid x) = \alpha^* + \beta^* x$ where

$$\beta^* = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \qquad \alpha^* = E(Y) - \beta^* E(X)$$

HINT: Goldberger Chapter 18.2

Based on your answer in **(b)** you should be able to convince yourself that the best predictor and best linear predictor (BLP) of y given x under MSE is the same when (x, y) is distributed bivariate normal.

(c) Show *explicitly* that $\text{Cov}(x, y) = \beta\sigma_z^2$ and $\text{Var}(x) = \sigma_z^2 + \sigma_u^2$

From here on, assume that you are given a random sample of (y, x).

(d) Goldberger claims that using OLS to regress y on x under the above formulation and sample leads to **biased** estimates of (α, β) but **unbiased** estimates of (α^*, β^*). Explain why this is true (do not just repeat Goldberger verbatim – use arguments we have discussed (repeatedly) in class).

(e) Goldberger further claims that this "bias" does not depend on (x, y) being distributed bivariate Normal. Explain. (Here, you can use Goldberger's words – but make sure you understand them!)

(f) The result in **(d)** is sometimes referred to as the "attenuation bias due to measurement error." Explain why. **HINT:** See β and β^*

(g) Suppose you know the true values of $(\sigma_z^2, \sigma_u^2, \sigma_v^2)$ but not μ. Can you construct unbiased estimates of α and β? If so, state the estimate. If not, explain why not. How does your answer change if you know μ?

HINT: If you know $(\sigma_z^2, \sigma_u^2, \sigma_v^2)$, you also know $\theta = \frac{\sigma_z^2}{\sigma_z^2 + \sigma_u^2}$. Also, you will need to use the OLS coefficients from regressing y on x. See **(d)**.

(h) Suppose it is known that consumption is proportional to permanent income, in the sense that $\alpha = 0$. How does this alter your answers in **(g)**?

(i) Suppose that $\alpha \neq 0$. Additionally, $E[u] = \gamma \neq 0$. So

$$\begin{pmatrix} z \\ u \\ v \end{pmatrix} \sim N\left[\begin{pmatrix} \mu \\ \gamma \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_z^2 & 0 & 0 \\ 0 & \sigma_u^2 & 0 \\ 0 & 0 & \sigma_v^2 \end{pmatrix} \right]$$

How does this unknown γ alter your answers in **(g)**?

(j) The above difficulties stem from the desire to estimate (α, β) rather than (α^*, β^*). For which types of statistical inference may an estimate of (α^*, β^*) suffice? For which do you need an estimate of (α, β)?

Question 3: Simultaneity and the Keynesian Model

Let us consider the Keynesian Model as described in Golberger Chapter 31.3. Goldberger presents a simplified (and linearized) version of the model taught in introductory and intermediate macroeconomics courses. The model consists of two equations: a consumption equation that relates the amount of consumption to income and an equation reflecting national income identity (*sans* government spending and net exports).

$$\begin{aligned} y &= \alpha + \beta x + u \\ x &= y + z \end{aligned}$$

where y is consumption, x income (or output), z private investment, and u a "consumption shock." α is the subsistence level of consumption and β the income proclivity to consume. The second equation, national income identity, can be considered an equilibrium condition, much like quantity supplied equals quantity demanded. In fact, it can be re-arranged to give the familiar "saving = investment" equilibrium condition: $x - y = z$.

To complete the Goldberger formulation, further assume that (z, u) are jointly distributed according to the following bivariate normal distribution:

$$\begin{pmatrix} z \\ u \end{pmatrix} \sim N\left[\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_z^2 & 0 \\ 0 & \sigma_u^2 \end{pmatrix} \right]$$

Lastly, you *directly* observe (y, x) but not (z, u). The challenge is to estimate (α, β) using a sample of (y, x).

(a) Under what *economic* condition do you indirectly observe z as well?

(b) Show *explicitly* how the *reduced form* equations

$$\begin{aligned} y &= (\alpha + \beta z + u)/(1 - \beta) \\ x &= (\alpha + z + u)/(1 - \beta) \end{aligned}$$

can be derived from the *structural* equations given above (the consumption and national income identity equations).

(c) Briefly explain why **(b)** implies that (y, x) is joint distributed multivariate normal. **HINT:** Goldberger Chapter 18.2

(d) Show *explicitly* the following results concerning the moments of x and y

$$\begin{aligned} \text{Cov}(x, y) &\equiv \sigma_{xy} = \frac{\beta\sigma_z^2 + \sigma_u^2}{(1-\beta)^2} \qquad & \text{Var}(x) &\equiv \sigma_x^2 = \frac{\sigma_z^2 + \sigma_u^2}{(1-\beta)^2} \\ E[y] &\equiv \mu_y = \frac{\alpha + \beta\mu}{1-\beta} \qquad & E[x] &\equiv \mu_x = \frac{\alpha + \mu}{1-\beta} \end{aligned}$$

From here on, assume that you are given a random sample of (y, x).

(e) Show *explicitly* that the bias from "estimating" β using OLS on (y, x) is

$$E[b^{ols} - \beta \mid x] \quad = \quad (1-\theta)(1-\beta)$$

where $\theta = \sigma_z^2/(\sigma_z^2 + \sigma_u^2)$.

HINT: See **(3b)** and **(3d)**

For the remaining problems, assume that a dollar increase in income (output) leads to a less than dollar increase in consumption: $0 < \beta < 1$.

(f) How does the bias resulting from "estimating" β via OLS on (y, x) change with [1] an increase in the variance of z [2] an increase in the variance of u? Which variance do you prefer to be bigger, σ_z^2 or σ_u^2?

(g) Derive the variance of y in terms of $\{\alpha, \beta, \sigma_z^2, \sigma_u^2, \sigma_{xy}\}$

HINT: See **(b)**

(h) Use **(f)** and **(g)** to evaluate the following statement

> "The simultaneity bias of the naïve OLS estimate of β falls as the variation in z accounts for a larger share of the variation in y. In other words, the simultaneity bias is mitigated to the extent that private investment shocks drive consumption fluctuation more than consumption shocks."

Question 4: Supply & Demand

Consider the Supply & Demand model discussed in class

$$\begin{aligned} \text{Supply :} \quad & P_t = \alpha_0 + \alpha_1 Q_t^s + \delta C_t + \eta_t \\ \text{Demand :} \quad & Q_t^d = \beta_0 + \beta_1 P_t + \gamma I_t + \nu_t \\ \text{Assume} \quad & \eta_t \overset{i.i.d.}{\sim} N(0, \sigma_\eta^2) \quad \nu_t \overset{i.i.d.}{\sim} N(0, \sigma_\nu^2) \end{aligned}$$

P_t is the price of gasoline, (Q_t^s, Q_t^d) quantity supplied and quantity demanded, C_t the cost of providing gasoline, and I_t consumer income. In addition to the assumptions above, assume that I_t and C_t are **exogenous** variables. In other words, (I_t, C_t) are independent of (η_t, ν_t). Moreover, assume that η and ν are completely independent of each other.

You are given a data set that contains the **market equilibrium** price (P_t) and quantity (Q_t) for 100 different markets $(t = 1, \ldots, 100)$ as well as the corresponding (C_t, I_t). The data set is called gasoline.csv and can be downloaded from the course website. The dataset should be read using the following command:

```
infile p q c i using gasoline.csv
```

If you stored gasoline.csv is a directory other than the STATA data directory, put the directory address in front of gasoline.csv. e.g. if the file is in C:\projects\data then use

```
infile p q c i using C:\projects\data\gasoline.csv
```

If you are using the Economics Department Computer Lab, remember to delete any STATA files you have created (including gasoline.csv) after you are done working on the assignment.

You can consider the data set to be a **random** sample following the data generating process described above.

(a) Given the assumptions above, do the **structural** equations individually satisfy the classical OLS assumptions? Given the assumptions above, do the **reduced form** equations satisfy the classical OLS assumptions? Explain *briefly.*

(b) Use OLS to calculate unbiased estimates of the reduced form parameters:
$(\Pi_{11}, \Pi_{12}, \Pi_{13}, \Pi_{21}, \Pi_{22}, \Pi_{23})$

(c) Derive Var$(\epsilon_{1t} \mid I_t, C_t)$,Var$(\epsilon_{2t} \mid I_t, C_t)$, and Cov$(\epsilon_{1t}, \epsilon_{2t} \mid I_t, C_t)$ in terms of the structural parameters $(\alpha_0, \alpha_1, \delta, \beta_0, \beta_1, \gamma)$, σ_η^2, and σ_ν^2. Note: $(\epsilon_{1t}, \epsilon_{2t})$ are the implied reduced form disturbance/error terms. (See lecture notes).

(d) In light of the answer in **(c)**, propose a way of estimating the reduced form parameters that might achieve lower MSE than the method used in **(b)**. **HINT:** Think SUR Model

OPTIONAL: Estimate the reduced form parameters using the method proposed in (d). Good practice to see if you understand GLS and SUR.

(e) Use your estimates from **(b)** to calculate consistent estimates of the structural parameters, via **indirect least squares** (ILS).

(f) Suppose you ran OLS on the structural equations, ignoring the simultaneity "bias" problem. Although these OLS "estimates" are neither unbiased nor consistent for the structural parameters, these "estimates" are consistent for (BLANK) ?

HINT: Recall the Keynesian Model and (α, β) versus (α^*, β^*)

(g) Calculate consistent estimates of the structural parameters, via **two stage least squares** (2SLS) *without* using `ivreg` or `ivregress`. i.e. do it "old school" and run both regressions. Show your steps, clearly

Note: You can use the STATA command `predict` to save the predicted values from a regression.

(h) Now compare your answers in **(g)** with the results from running the following commands:

- ivregress 2sls p (q = i) c
- ivregress 2sls q (p = c) i

You should find that the standard errors in **(h)** are different than the standard errors obtained from the second stage regression. In fact, one set should be larger than the other by a given scale (divide them and see). This is due to the fact that the estimate of σ^2 used by STATA in `reg` is different from that in `ivregress`.

(i) Use your answer in **(h)** to test the following economic hypotheses:

- Cost Pass-through: Producers pass along the entire burden of cost changes to the consumers $(\delta = 1)$
- Demand for the first observation is unit price elastic $(\frac{\partial Q_1^d}{\partial P_1}\frac{P_1}{Q_1^d} = -1)$

Keep in mind that you are using a consistent estimate and asymptotic standard errors. So your hypothesis tests should be asymptotically valid tests: i.e. you should be looking up critical values in the standard Normal table (not t-table)

For the unit price elastic hypothesis, note the following

- You can use the Data Browser (under Top Menu, Data) to find the values of P_1 and Q_1^d. Note that $Q_1^d = Q_1^s = Q_1$.
- You should be able to tell me why $\frac{\partial Q_1^d}{\partial P_1} = \beta_1$ (Think exogenous variables and earlier discussion on partial derivatives and interpreting regression coefficients)