Metrics Assignment 2

Bobing, Cassandra, Max, Prema, Rajdev, and Yazen

March 2023

NB: Please grade sections 1, 2, 4, and 5.

1 Identifying Assumptions for Regression

1. Note that $E(u|X) = 0$ implies $E(h(X) \cdot u) = 0$ iff $h(x)E(u|x) = E[h(x) \cdot u|x]$. This later statement is only true if $h(x)$ is x measurable. If $h(x)$ is x measurable, we then can show that:

$$\begin{array}{rcc} E(u|X) & = & 0 \\ \implies h(X)E(u|X) & = & h(X) \cdot 0 \\ \implies E[E(h(X)u|X)] & = & E[0] \\ \implies E(h(X) \cdot u) & = & 0 \end{array}$$

   Where the last equality follows from the law of iterated expectations.

2. Consider the case where we include a constant term in X and can thus remove the explicit α term. We can substitute in $Y - X\beta$ for U to get:

$$E[(Y - X\beta)X^3] = 0$$
$$E[YX^3] = E[XX^3]\beta$$
$$E[XX^3]^{-1}E[YX^3] = \beta$$

   As compared to the OLS estimator, this estimator will place greater weight on outliers. Whether or not this is desirable depends upon the research context. If one thinks outliers are the result of, say, data entry problems, this is not a desirable property. If there is reason to be outliers are genuine, however, this property may be desirable.

3. We want to show that independence implies mean independence but that the converse does not hold.

   First, I'll prove that independence implies mean independence, i.e. that $P[XY] = P[X]P[Y]$ implies that $E[X|Y] = E[X]$.

   Note that:

$$\begin{array}{rcc} E(X|Y=y) & = & \sum_x xP(X=x|Y=y) \\ & = & \sum_x x\frac{P(X=x,Y=y)}{P(Y=y)} \\ & = & \sum_x x\frac{P(X=x)P(Y=y)}{P(Y=y)} \\ & = & \sum_x xP(X=x) \\ & = & E(X) \end{array}$$

   Where the third line follows from independence.

   Note that mean independence does not necessarily imply independence. Suppose we have some two variables (X, Y) that are uniformly distributed around a unit circle such that only values satisfying $X^2+Y^2 = 1$ have positive measure. Clearly, $P(X > .5) = \frac{1}{3}$ and $P(Y > .5) = \frac{1}{3}$. Notably, however, $P(X > .5 \cap Y > .5) \neq \frac{1}{9} = P(X > .5)P(Y > .5)$. So X and Y are not independent. They are, however, mean independent, since $E[X|Y] = E[X] = 0$.

4. We want to show if U is mean independent of X then E[Uh(X)]=E(U)=0.

   If U is mean independent of X, then $E[U|X] = E[U]$. By exercise (1.1), we know that $E[Uh(X)|X] = E[U|X]h(X) = E[U]h(x) = 0$. It is my understanding that we must assume $E[U] = 0$ here, but that if we do, the rest of the result follows.

   Note that if we have independence of U and X, then $E[UX] = E[U]E[X]$, since independent variables must be uncorrelated. By the linearity of the expectation, it must also be the case then that $E(g(U)X) = E[X]E[g(U)]$.

5. Note that to minimize $E[(y-\hat{y}(X))^2|X] = E[y^2 - 2y\hat{y} + \hat{y}^2|X]$, we take the FOC with respect to $\hat{y}$, which produces:

$$E[y|X] = \hat{y}.$$

   We can also say $\hat{y}$ minimizes the expected square of u.

6. To be a valid instrument, D must satisfy the exclusion restriction ($E[uD = 0]$) and relevance assumptions ($E[XD] \neq 0$). For the exclusion restriction, note that $E(u|D) = 0 \implies E(uD|D) = 0$, which, by the law of iterated expectations, implies $E(uD) = 0$.

   For relevance, note that X is not mean independent from D. We proved in (3) that independence implies mean independence. This means the contrapositive of this statement is also true, namely that not mean independent implies not independent. From this, we know that X and D are not independent, which means that $E[XD] \neq 0$.

   From section 4.1.2 of Mostly Harmless Econometrics, we know the IV estimator for a binary random variable is:

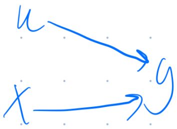
   $$\beta_{IV} = \frac{E[Y|D=1] - E[Y|D=0]}{E[X|D=1] - E[X|D=0]}$$

   In order to understand the impact of X on Y, we intuitively divide a reduced form treatment impact by a first stage treatment impact.

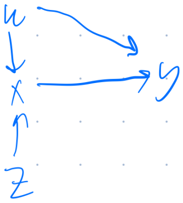
7. In order for OLS to be justified, we need to have that for some model $y = \beta x + u$, x direclty impacts y and u has no impact on y through x. If x mediated the relationship between u and y, we would then violate the assumption that $E[ux] = 0$. In the diagram titled "OLS," there is no need for an IV regression. In the diagram titled "IV Problem," we would need to use an IV estimator and would run into issues if we used an OLS estimator. This is because some component of the error term, which we do not observe, influences x which in turn influences y. If we had some instrument z, we could examine which of the variation in y is driven by changes in x by examining only the variation in x induced by z, which has no direct impact on y.

   You could test this which model is appropriate using a Hausman test. Intuitively, this test assesses whether the coefficient on x is similar when using an instrument or a simple OLS regression. If the two coefficients are sufficiently different, one can conclude that x is endogeneous (i.e. the second model is correct).

OLS



IV
problem



8.

2 Wright (1928)

In this question, we consider the canonical supply and demand model[1]. Quantity supplied is a function of price and a set of "supply shifters" (equation 2.1), quantity demanded is a function of price and a set of "demand shifters" (equation 2.2), and market clearing implies that at some price, quantity demanded is equal to quantity supplied (equation 2.3).

$$q_S = S(p, v) \tag{2.1}$$

$$q_D = D(p, u) \tag{2.2}$$

$$D(p^*(u,v), u) = S(p^*(u,v), v) = q^*(u,v) \tag{2.3}$$

Following wright34.ipynb, in the case of linear demand and supply, we have the following demand-supply system: $q_D = \alpha p + u$, $q_S = \beta p + v$, and $q_D = q_S$. Further, $(u, v) \sim N(0, \sigma^2)$ are unobserved shocks to demand and supply.

2.1 Control: Set price $p = p_0$

The expected demand when price is set to p_0 is:

$$\mathbb{E}(q_D(p_0)) = \int q_D(p_0, u) dF_u(u) \tag{2.4}$$

In this case, price is no longer a random variable. Setting the price does not affect the distribution of u. When we set price to p_0, it need not be the case that markets clear (i.e. expected quantity demanded may be greater than or less than expected quantity supplied).
In the case of the linear demand as in wright34.ipynb, we have $\mathbb{E}(q_D(p_0)) = \mathbb{E}(\alpha p_0 + u) = \alpha p_0$.

2.2 Conditional: Observe price $p = p_0$

When we observe a realization of $p = p_0$, expected demand is as follows:

$$\mathbb{E}(q^* | p = p_0) = \mathbb{E}[q^*(u, v) | q_D(p_0, u) = q_S(p_0, v)] \tag{2.5}$$

[1] The answer for this question closely follows lecture slides on "Causality and Correlation"

If we observe a price p_0 that is high, we will also see a higher *conditional* expected value of u and a lower conditional value of v (assuming that q_D is monotonically increasing in u and q_S is monotonically increasing in v).
In contrast to the case in Section 2.1, where it is possible that market is not in equilibrium when the price is *set* to p_0, in this case, we *observe* that price is p_0, meaning the market is in equilibirum at this point.
In the case of the linear demand as in wright34.ipynb, we have:

$$\mathbb{E}[q^*(u,v)|q_D(p_0,u)=q_S(p_0,v)]=\mathbb{E}[q^*(u,v)|p_0=(v-u)/(\alpha-\beta)]$$

2.3 Counterfactual: Change from (p_0, q_0) to new price p_1

If we observe equilibrium price and quantity as (p_0, q_0), and then we were to change the price to p_1, the change in demand, *ceteris paribus*, is given by:

$$\Delta = q_D(p_1, u_0) - q_D(p_0, u_0) \tag{2.6}$$

Equation 2.6 assumes that we can observe (u_0, v_0) by making assumptions such as monotonicity, so these are not random anymore.
We refer to this case as the counterfactual as it compares (in theory) quantity demanded in two states of the world, where only the price has changed from one state of the world to the other as everything else is held constant (*ceteris paribus* assumption). In the case of the linear demand as in wright34.ipynb, we have:

$$\Delta = q_D(p_1, u_0) - q_D(p_0, u_0) = \alpha p_1 - u_0 - (\alpha p_0 - u_0) = \alpha(p_1 - p_0) \tag{2.7}$$

3 “Plausibly Exogenous” (please ignore when grading)

3.1 Wright (1934)

3.1.1 IV Estimation

If $y = X\beta + u$ and X^+u is unknown, then the least squares estimator, b, is no longer unbiased. That is,

$$\begin{aligned} b &= X^+y \\ &= X^+(X\beta + u) \\ &= \beta + X^+u \\ \implies E[b|X] &= \beta + X^+E[u|X] \neq \beta. \end{aligned}$$

Furthermore, we say β is not identified in terms of the moment equations of y and X (dimension $n \times k$) alone because the corresponding moment equations have k equations in $2k$ unknowns. To see this, note that

$$\begin{aligned} \text{plim}(X'y/n) &= \text{plim}(X'X/n)\beta + \text{plim}(X'u/n) \\ Q_{Xy} &= Q_{XX}\beta + \text{plim}(X'u/n) \end{aligned}$$

where Q_{XX} is a positive definite matrix by the full rank assumption (Gauss Markov, A.2 (Greene, 2012, p. 246)) and the limit exists because we assume the conditions for the Central Limit Theorem are satisfied (e.g. all terms have finite variances).

As the question states, with an instrumental variable, Z, satisfying the following conditions, it becomes possible to identify β:

(1) Exogeneity of Z: $E[u|Z] = 0$

(2) Full rank/invertibility of Z^TZ: $E[Z^TZ] > 0$

(3) Relevance of Z for X: rank $(E[Z^TX]) = k$.

(1) implies $Cov(u^TZ) = 0 \implies E[u^TZ] = 0 \implies \text{plim}(Z^Tu/n) = 0$ if all terms have finite variances such that the Central Limit Theorem applies. By the same token, (2) implies Q_{ZZ} exists and is positive definite. (3) implies $l \geq k$ where Z comprises l instruments and X comprises k independent variables. See Hansen (2022) Definition 11.3.1.

These conditions allow for the estimation of the standard IV estimator when $l = k$:

$$b_{IV} = (Z'X)^{-1}(Z'Y)$$

which is unbiased and consistent because

$$\begin{aligned} \text{plim}(Z^T y/n) &= \text{plim}(Z^T X/n)\beta + \text{plim}(Z^T u/n) \\ &= \text{plim}(Z^T X/n)\beta \\ \implies Q_{Zy} &= Q_{ZX}\beta. \end{aligned}$$

If $l > k$ (there are more instruments than independent variables), the two stage least squares estimator can be readily computed as:

$$b_{2SLS} = (\hat{X}^T X)^{-1}(\hat{X}^T Y).$$

where

$$\hat{X} = Z(Z^T Z)^{-1} Z^T X.$$

3.2 Conley et al. (2012)

3.2.1 $\gamma \neq 0$

3.2.2 Test: $b(\gamma) = b(0)$

3.2.3 Same test under new Cov(Z, X)

Refer to are212.pset2.q3.ipynb

4 Weak Instruments

Refer to pset2.problem4.ipynb

5 A Simple Approach to Inference with Weak Instruments

Refer to PS2_Q5.ipynb.

6 Angrist and Krueger (1991) Replication

6.1

The identifying assumption made in this paper is that quarter of birth is exogenous to unmeasurable factors affecting both schooling and also future wages; that is, the instrument is uncorrelated with the error terms in the model.

This might not be the case given that some jobs of parents work on a yearly cycle, resulting in some correlation between QOB and parents' income, which can be a predictor of child's income. Additionally, there is some seasonality to when babies are conceived, which may be caused by something else endogenous to future wealth.

Coincidentally, most of our cohort are summer babies. Take from this what you will.

6.2

See PSet 2 6.2.ipynb and PSet 2 6.2.2.ipynb

6.3

See PSet 2 6.3.ipynb

6.4

See PSet 2 6.4 and 6.5.ipynb and PSet 2 6.4 and 6.5 table 7.ipynb

References

Conley, T. G., Hansen, C. B., and Rossi, P. E. (2012). Plausibly exogenous. *Review of Economics and Statistics*, 94(1):260–272.

Greene, W. (2012). *Econometric Analysis*.

Hansen, B. (2022). *Econometrics*. Princeton University Press.

Wright, S. (1934). The method of path coefficients. *The annals of mathematical statistics*, 5(3):161–215.