

Linear Regressions

Li Gan

(September, 2009)

Consider a linear specification: $E[y|X] = X\beta$, where y is $n \times 1$, X is $n \times K$, and β is $K \times 1$. Adding an error term:

$$y = E[y|X] + u = X\beta + u \quad (1)$$

Assumption 1: $E(X'u) = 0$

Assumption 2: $\text{rank}(E(X'X)) = K$

To solve for (1), take expectation at both sides:

$$E(X'y) = E(X'X)\beta + E(X'u)$$

Empirically, expectations are approximated by averages:

$$\begin{aligned} \hat{\beta} &= \left(\frac{1}{N} \sum_i x_i' x_i \right)^{-1} \left(\frac{1}{N} \sum_i x_i' y_i \right) \\ &= \beta + \left(\frac{1}{N} \sum_i x_i' x_i \right)^{-1} \left(\frac{1}{N} \sum_i x_i' u_i \right) \end{aligned} \quad (2)$$

Note x_i is $1 \times K$.

Discussion: Unbiasedness and consistency

If Assumption 1 holds:

$$\begin{aligned} \beta &= E(X'y) = [E(X'X)]^{-1} E(X'y) \\ &= [E(X'X)]^{-1} E(X'(X\beta + u)) \\ &= [E(X'X)]^{-1} [E(X'X)]\beta + [E(X'X)]^{-1} E(X'u) \\ &= \beta + [E(X'X)]^{-1} E(X'u) \\ &= \beta \quad \text{by Assumption 1.} \end{aligned}$$

As $N \rightarrow \infty$. Let $A = E(x_i' x_i)$. Then:

$$p \lim \hat{\beta} = \beta + A^{-1} E(x_i' u_i)$$

Obviously, A has to have full rank to have $p \lim \hat{\beta} = \beta$.

Therefore, if *Assumptions* 1 and 2 hold, as expected, we have:

$$E(\hat{\beta}) = \beta, \text{ unbiased; and } \text{plim } \hat{\beta} = \beta, \text{ consistent.}$$

Discussions: Asymptotic distribution of $\hat{\beta}$

Repeat (2) here:

$$\hat{\beta} = \beta + \left(\frac{1}{N} \sum_i^N x_i' x_i \right)^{-1} \left(\frac{1}{N} \sum_i^N x_i' u_i \right)$$

Rearrange equation (2), and multiply by \sqrt{N} :

$$\sqrt{N}(\hat{\beta} - \beta) = \sqrt{N} \left(\frac{1}{N} \sum_i^N x_i' x_i \right)^{-1} \left(\frac{1}{N} \sum_i^N x_i' u_i \right)$$

$$\text{As } N \rightarrow \infty, \sqrt{N}(\hat{\beta} - \beta) = A^{-1} \sqrt{N} \left(\frac{1}{N} \sum_i^N x_i' u_i \right)$$

Applying the central limit theorem to the term: $\sqrt{N} \left(\frac{1}{N} \sum_i^N x_i' u_i \right)$:

$$\text{Let } z_i = x_i' u_i, \text{ and } \bar{z}_N = \frac{1}{N} \sum_i^N x_i' u_i$$

By *Assumption 1*, $E(z_i) = 0$. Applying *CLT*,

$$\sqrt{N} \bar{z}_N \rightarrow N \left(0, \frac{1}{N} \sum_i^N \text{Var}(z_i) \right), \text{ where}$$

$$\begin{aligned} \text{Var}(z_i) &= E(z_i z_i') \\ &= E(x_i' u_i u_i' x_i) \\ &= X_i' E(u_i u_i') X_i \\ &= x_i' x_i \sigma^2 \end{aligned}$$

In which, $\sigma^2 = E(u_i u_i')$. Note that the sufficient condition that equation $E(x_i' u_i u_i' x_i) = X_i' E(u_i u_i') X_i$ holds is *Assumption 1*.

Therefore,

$$\begin{aligned} A \sqrt{N}(\hat{\beta} - \beta) &= \sqrt{N} \left(\frac{1}{N} \sum_i^N x_i' u_i \right) \\ &\rightarrow N(0, \sigma^2 A) \end{aligned}$$

Furthermore,

$$\sqrt{N}(\hat{\beta} - \beta) \rightarrow N(0, \sigma^2 A^{-1})$$

Assumption 3: u_i is IID, or $\Omega = E(u_i u_i') = \sigma^2 I_{NXN}$.

If Assumption 3 holds, then we must have: $\sqrt{N}(\hat{\beta} - \beta) \rightarrow N(0, \sigma^2 A^{-1})$

Now, what happens if some of the assumptions do not hold?

Assumption 1 $E(X'u) = 0$

Assumption 2 $E(X'X)$ has full rank

Assumption 3 $E(uu') = \sigma^2 I$

Case 1: If Assumption 2 does not hold:

If assumption 2 does not hold, then at least one of x_i 's is the linear combination of others. Get rid of this x_i 's (reduce the dimension of x_i). This is called multicollinearity!

Example: The Longley data (on the website)

Data: *GNP deflator*, *GNP*, *Armed forces*, and *Total employment*. 1947-1962.

$$total = \beta_0 + \beta_1 * year + \beta_2 * GNPdeflator + \beta_3 * GNP + \beta_4 * Armedforces$$

Estimation results:

Parameters	1947-1961	1947-1962
β_0	1459400	1169090
β_1	-721.76	-576.464
β_2	-181.12	-19.761
β_3	0.091068	0.064394
β_4	-0.074937	-0.01014

Symptoms of near multicollinearity:

(1) Small change in data produces wide swing in parameter values

(2) Large standard errors despite joint significant and high R^2

To fix the near-multicollinearity: ridge regression:

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'y$$

$$\hat{\beta}_r = (X'X + rD)^{-1} X'y$$

where D - diagonal elements of $X'X$, and r is a scalar to be chosen such that the

resulting estimates are “stable”. In practice, let

$$r(k) = k * .01; k = 1, \dots, K.$$

$$\left| \hat{\beta}_{r(k)} - \hat{\beta}_{r(k-1)} \right| < tolerance$$

It can be shown that $\hat{\beta}_r$ is biased but may have a smaller mean square error than OLS. However, in order to construct an estimate with smaller mean square error, it is necessary to know $\hat{\beta}_r$. So this is irrelevant.

Case 2: If Assumption 3 does not hold:

$$E(uu') \neq \sigma^2 I$$

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'y$$

$$= \beta + (X'X)^{-1} X'u$$

$$Var(\hat{\beta}_{OLS}) = E[(X'X)^{-1} X'uu'X(X'X)^{-1}]$$

$$= (X'X)^{-1} E[X'uu'X](X'X)^{-1}$$

It is easy to estimate $(X'X)^{-1}$. If we can estimate $E(X'uu'X)$, we can obtain a consistent estimate of the covariance matrix of the OLS estimate of β . This is possible because the dimension of $X'uu'X$ is only $K \times K$, as n goes to infinity.

In particular, if $E(uu') = \Omega = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}$

Then: $E(X'uu'X) \approx \frac{1}{N} \sum x_i' x_i u_i^2$. Let $E(X'uu'X) = B$, which can be approximated by:

$$\frac{1}{n} \sum_i u_i^2 x_i' x_i \rightarrow B$$

Since u_i is not observed, we replace u_i by the OLS residuals: $\hat{u}_i = y_i - X_i \hat{\beta}$:

$$\hat{B} \equiv \frac{1}{n} \sum_i \hat{u}_i^2 x_i' x_i$$

Standard error calculated by $(X'X)^{-1} \hat{B} (X'X)^{-1}$ is called heteroscedasticity-robust standard error.

Note if $E(u_i^2) = \sigma_i^2$, can we use \hat{u}_i^2 to approximate σ_i^2 ? In other words, if $E(uu')$ = Ω , can we find consistent estimate of Ω ?

The difference that $\hat{B} \xrightarrow{p} B$ is that the dimension of B is only $K \times K$, which does not change as the number of observations increases. However, the dimension of Ω is $n \times n$. So there is no way we can consistently estimate Ω without further assumptions of the covariance structure of Ω .

For example, in the time series data, it is possible that Ω has some particular structure, the following is an example:

$$\Omega = \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & & \\ \vdots & & \ddots & \\ \rho^{n-1} & & & 1 \end{pmatrix}$$

In this example, $\Omega = \Omega(\rho)$. It is then possible to estimate such a covariance matrix consistently.

Case 3: If Assumption 1 does not hold:

If Assumption 1 does not hold: $E(X'u) \neq 0$.

If one of X_i such that: $E(X_i'u) \neq 0$, then the estimate is biased and inconsistent

Question (homework question): is it true that only $\hat{\beta}_i$ is biased & inconsistent?

Suppose only x_1 and x_2 , $E(x_1) = E(x_2) = 0$. If $\text{Cov}(x_1, x_2) = 0$ vs $\text{Cov}(x_1, x_2) \neq 0$?

$\hat{\beta}_2$ will be biased and inconsistent?

Reasons that may cause: $E(X'u) \neq 0$

1. Omitted variables.

Suppose $E(y|X, q)$ is the conditional expectation of interest (true model). Let the true model be:

$$y = X\beta + \gamma q + u$$

Since q is unobserved, the model to be estimated is:

$$y = X\beta + u^0$$

Where $u^0 = \gamma q + u$. Then we have $\text{Cov}(X', u^0) \neq 0$ if $\text{Cov}(X', q) \neq 0$.

Example (wage equation):

The true model is:

$$\ln(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{ability} + \beta_3 \text{expr} + \dots$$

We don't observe "ability" → omitted variable problem. It is very likely that:
 $\text{Cov}(\text{"ability"}, \text{educ}) > 0$.

Therefore, OLS estimate for β_1 will be biased, and biased upward, i.e., $\hat{\beta}_{1OLS} > \beta_1$.

2. Measurement errors:

True model:

$$y = X\beta + \gamma z^* + u,$$

where z^* is unobserved but we can find a proxy for z^* , z . Or z^* is measured with error. Suppose $z = z^* + \xi_z$, where $E(\xi_z) = 0$. z is observed and z^* is unobserved. Assume $\text{Cov}(z^*, \xi_z) = 0$, then $\text{Cov}(z, \xi_z) \neq 0$.

$$\begin{aligned} y &= X\beta + \gamma z^* + u \\ &= X\beta + \gamma(z - \xi_z) + u \\ &= X\beta + \gamma z - \gamma \xi_z + u \\ &= X\beta + \gamma z + u^0 \end{aligned}$$

where $u^0 = -\gamma \xi_z + u$. We must have: $\text{Cov}(z, u^0) \neq 0$ since $\text{Cov}(z, \xi_z) \neq 0$.

Example (i): If "ability" is measured by *IQ*

$$\ln(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{IQ} + \beta_3 \text{expr} + \dots$$

Example (ii): We are interested in how corporate donation is determined. Two alternative hypotheses – *donation* as an advertising mechanism or *donation* as a way to maximize management utility.

A typical way in the literature is:

$$\text{Donation} = X\beta + \gamma \text{Advertising} + u$$

However, this will have an simultaneity problem.

Gan and Shan (2009) suggest using corporate product type – whether the firm is directly consumer-oriented or not.

$$\text{Donation} = X\beta + \gamma \text{Direct} + u$$

However, *Direct* would potentially be measured with error.

3. *Simultaneity*:

At least one of the explanatory variables is determined simultaneously along with y . This is what we call in economics that both y and X are endogenous variables:

Examples of Simultaneity:

Non-economics examples:

- (i) The effect of stay-home moms

A very important question in the US is whether stay-home mothers have positive contribution to their children.

$$\text{Children performance} = a + b * \text{stay-home-mommy} + Z\eta + v$$

Sociologists run this type of regressions. For economists, this is simply wrong, because of the endogeneity: whether a mommy stays at home will be affected by the behavior of their children.

- (ii) The relationship between money spent on elections and the outcomes of elections. A typical empirical model is:

$$\text{Winning percentage} = a + b * \text{difference in campaign funds} + X\beta + u$$

The problem of this analysis is that campaign fund raising is very much dependent on the perceived (or expected difference) in a possibly non-linear way.

Economics examples:

Example (i): We are interested in studying how price affects demand of a good. We have time series data:

q_t : quantity of orange juice consumed in a city at time t .

p_t : price of orange juice in a city at time t .

$$q_t = \beta_0 + \beta_1 p_t + z_t \gamma + u_t$$

p_t is endogenous. Suppose for whatever reason, the demand curve shifts up for orange juice, both p_t and q_t would increase.

However, if we have household data:

$$q_i = \beta_0 + \beta_1 p_i + z_i \gamma + u_i,$$

We can estimate the model since individual households should not be able to affect price of orange juice. Problem of this approach: p_i does not have enough variations.

Example (ii): Female labor supply.

We are interested how after-tax wage affect female labor supply. A typical empirical model is given by:

$$Hours_i = \alpha w_i(1-t_i) + \beta y_i + Z_i \eta + u_i$$

where w_i is the before-tax wage, t_i is the tax rate, and y_i is the non-labor income, and Z_i is a set of control variables.

It is well-known in the literature that t_i , the marginal tax rate is dependent on $Hours_i$, since more working hours \rightarrow higher household income \rightarrow higher marginal tax rate. Therefore, both t_i and $Hours_i$ move simultaneously and they both are jointly determined. So it is $(1-t_i)$ is endogenous, and so is w_i . This problem can be solved though by maximum likelihood proposed by Hausman.

Example (iii): Transaction volume and change of housing prices.

It is widely observed that transaction volumes and housing prices are positively correlated at the aggregate level: a 1% drop in housing prices is associated with 4% drop in transaction volumes.

Two previous models,

(a) Down payment: this model suggests that that a decrease in price would lower the equity of the existing homes. So people who want to trade-up of their homes (sell their current home to buy a different and often a larger home) cannot do so – this reduces the sellers in the market, and hence reduces the total transaction volumes and housing prices.

(b) Loss aversion: according to the Prospect Theory of Kahneman and Tversky, people hate to lose. Therefore, if the price is lower than the purchase price, people do not want to sell. As a consequence, a lower price leads to a lower transaction volume.

In both models,

$$Q_t = \beta_0 + \beta_1 p_t + Z_t \gamma + u$$

Or:

$$\Delta Q_t = \beta_0 + \beta_1 \Delta p_t + Z_t \gamma + u$$

where Q_t is the transaction volumes, p_t is the house price, Z_t are control variables.

In both models, a change in prices causes a change in transaction volumes.

(c) Thin-thick model:

unemployment $\uparrow \Rightarrow$ market thinner $\Rightarrow P \downarrow$ & transaction volume \downarrow

If (c) is correct, then one cannot run regression of p_t on Q_t . In fact, a large part of empirical work is all about this testing alternative theories. Different theories may imply different causality.

Example (iv): testing personal bankruptcy

$$B_i = X_i\beta + \gamma D_i + u_i$$

where D_i is debt, and it is correlated with u_i .

Two alternative models: “*Strategic timing*” and “*Adverse Events*”. One model implies that the *Debt*, D_i , is exogenous (*Adverse Events*), while another model implies that D_i is endogenous (*Strategic Timing*). Endogeneity of D_i becomes the key to testing alternative theories.

Example (v): Fertility or spacing and labor supply.

$$F_i = 1(X_i\beta + \gamma LS_i + u_i > 0)$$

$$S_i = X_i\beta + \gamma LS_i + u_i$$

F_i is fertility decision of a mother, S_i is the spacing between two children, and $LS_i = 1$ if work, 0 otherwise.

Direction of the Bias of $\hat{\beta}_{OLS}$ when $\text{Cov}(X, u) \neq 0$:

Given that:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'u$$

The direction of the bias depends the correlations between X and u .

- (1) For the *omitted variable* case, one can obtain the direction of bias by considering the correlation between X and the omitted variable q .

Examples:

- (i) In the case of returns to schooling with the missing “Ability” term, since $\text{Cov}(\text{Schooling-years}, \text{“Ability”}) > 0 \rightarrow \hat{\beta}_{OLS} > 0$ overestimates the returns.
- (ii) Example 4.4 of the Wooldridge book, since the estimate of the coefficient for “grant” is reduced after considering the productivity of the firm, i.e.,

$$\hat{\beta}_{OLS} > \hat{\beta}(\text{with productivity})$$

We must have:

$$\text{Cov}(\text{grant}, \text{“productivity”}) > 0.$$

(2) For the *measurement error* case:

Let the true X be denoted as X^* , the observed $X = X^* + v$

Then the true model is:

$$\begin{aligned} y &= X^* \beta + u = (X - v) \beta + u \\ &= X \beta + (u - v \beta) \end{aligned}$$

Since $\text{Cov}(X, u - v\beta) < -\beta \text{Cov}(X, v)$, and $\text{Cov}(X, v) > 0$, we have the so-called *attenuation effect*:

- (i) If $\beta > 0$, then $\text{Cov}(X, u - v\beta) < 0$, then $\hat{\beta}_{OLS} < \beta$, i.e. OLS estimate of β would underestimate the true β .
- (ii) If $\beta < 0$, then $\text{Cov}(X, u - v\beta) > 0$, then $\hat{\beta}_{OLS} > \beta$, i.e. OLS estimate of β would overestimate the true β .
- (iii) In summary, with classical measurement error, we would have: $|\hat{\beta}_{OLS}| < |\beta|$
The OLS estimate would be smaller in magnitude. In other words, it is more likely that the OLS estimates would be statistically insignificant.

Instrumental Variable Estimation

The basic model: $y = X\beta + u$, and $\text{Cov}(X, u) \neq 0$.

Possible reasons: omitted variables; measurement error in explanatory variables; and simultaneity.

How to solve this problem? Assume there exists a vector of random variable Z , such that, $\text{Cov}(Z, X) \neq 0$, and $\text{Cov}(Z, u) = 0$. Pre-multiply z on (2):

$$Z' y = Z' X \beta + Z' u$$

Taking expectations on both sides: $E(Z'y) = E(Z'X)\beta + E(Z'u)$

If $E(Z'X) \neq 0$, then we have: $\beta = \frac{E(Z'y)}{E(Z'X)}$

Empirically, we use sample average to approximate the expectation:

$$\begin{aligned}\hat{\beta}_{IV} &= \left(\frac{1}{N} \sum_i z_i' x_i \right)^{-1} \left(\frac{1}{N} \sum_i z_i' y_i \right) \\ &= \beta + \left(\frac{1}{N} \sum_i z_i' x_i \right)^{-1} \left(\frac{1}{N} \sum_i z_i' u_i \right)\end{aligned}$$

Basic requirements of *IV*:

- (1) $Z'x$ must have full rank.
- (2) $\text{Cov}(Z, x) \neq 0$ (testable), and $E(Z'u) = 0$ (not really testable).

A large part of empirical studies have been concentrated on searching for appropriate IVs. Following are several examples. We divide the examples into two categories: “Natural Experiment” vs “Regular examples”

Example (*Natural Experiment*):

(a) Omitted variable: “*Ability*” in the wage regression.

$$\ln(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{“Ability”} + \beta_3 \text{expr} + \dots + u$$

“*Ability*” is missing, so we have $\varepsilon = u + \beta_2 \text{“Ability”}$. Note it is obvious that $\text{Cov}(\text{educ}, \varepsilon) \neq 0$, since $\text{Cov}(\text{educ}, \text{Ability}) \neq 0$.

Can we find Z such that: $\text{Cov}(\text{educ}, z) \neq 0$, and $\text{Cov}(\text{ability}, z) = 0$

Angrist and Krueger (1991, November, *QJE*): *IV* is born in the first quarter in the year:

$$\text{educ} = \delta_0 + \delta_1 \text{Born-First-Quarter} + X\beta + v$$

This is because of the compulsory school law: that each person has to stay in school until age 16. In the meantime, many states have a cutoff date of 9/1 or 12/1. If a person was born before the cutoff date, he/she can go to school this year; otherwise, he/she has to wait until next year. Therefore, because the compulsory schooling law and the cutoff date for entrance, on average people who were born at the first quarter (more likely to be before the cutoff date) would have more education than those who were born the last quarter (more likely to be after the cutoff date).

Card (1995)

$$\text{educ} = \delta_0 + \delta_1 \text{Distance-to-Community-College} + X\beta + v$$

Example (b) (*Natural Experiment*) : serving military on long-term earning wages.

Angrist (1990, *AER*): the effect of serving the Vietnam War on the earnings of men. He is interested on how the participation in military would affect a person's long term wage.

$$\ln(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_3 \text{expr} + \beta_4 \text{Military} \dots + u$$

The problem of this regression, as in the previous regressions, is that "Ability" is missing. It is very likely that $\text{Cov}(\text{Ability}, \text{Military}) \neq 0$. People with higher earning ability are less likely to join military (counter example: Pat Tillman)

Angrist (1990) considers the example of the Draft number. One December 1, 1969 each day of the year has a capsule randomly picked up. The draft continued yearly until 1973 (born between January 1st, 1944 and December 31, 1950). All numbers would have equal chance of being drafted.

$$\text{Military} = \delta_0 + \delta_1 \text{Draft-Order} + X\beta + v$$

However, there still is a problem in this. Being drafted was not an automatic ticket to military, only half went (rejected on physical, mental, or legal reasons.) Draft is random but draftees choose more education as a way of increasing the chance of deferment.

Example (c) (*Natural Experiment*): Colonialism and modern income on islands (80 observations)

James Feyrer and Bruce Sacerdote (2006, working paper)

$$\text{Income}_i = X_i\beta + \gamma \text{length-of-colonization}_i + c_i(\text{unobserved}) + u_i$$

Problem: length of colonization is related to unobserved characteristics c_i .

IV regressions:

$$\text{length-of-colonization} = f(\text{wind-direction}, \text{wind speed}) + X$$

Wind pattern which matters a great deal during the sail do not have much effect on modern income.

Results: date of democratization is NOT a predictor of current income;

length-of-colonization is strongly positively correlated with current income. (45% higher per capita income if 100 years more colonization)

Example (d) (Natural Experiment):

$$\text{Performance of children} = a + b * \text{size of class} + X\beta + u$$

Obviously size of class is endogenous. However, Angrist and Lavy (1999, *QJE*, Vol 114, No.2, pp. 533-575) suggest an IV:

In Israel, the great twelfth century Rabbinic scholar, Maimonides, suggested that class size should not be more than 40.¹ Angrist and Lavy (1999) compare those schools with only 120 students at one grade (three classes are sufficient) vs those schools with 121 students at one grade (has to be divided into four classes): Regression discontinuity – we will talk about it later.

Example (d) (*regular*): The effect of village elections on village resident income and consumption insurance. (Gan, Xu, and Yao, 2006):

Basic question:

$$\log(\text{income}) = a + b1 * \text{health shock} + b2 * \text{village election} * \text{health shock} + X\beta + u$$

The interests are b1 and b2: it is expected b1 is negative, and b2 is positive.

the timing of village election, however, maybe endogenous – related to average income of the households.

Instruments variables:

Timing of the Provincial election law * percentage of the largest surnames

Timing of the Provincial election law * number of surnames in the village

Example (e) Simultaneity of X

Example: personal bankruptcy

$$\text{Bankruptcy}_i = X_i\beta + \gamma D_i + u_i$$

where D_i is debt, and it is correlated with u_i . The *IV* for this problem includes: medical problem, divorce, and unemployed.

¹ In fact, the precise wording of Maimonides is, according to Angrist and Lavy (1999), is, “Twenty-five children may be put in charge of one teacher. If the number in the class exceeds twenty-five but is not more than forty, he should have an assistant to help with the instruction. If there are more than forty, two teachers must be appointed.”

Example (e): Measurement error:

$$\begin{aligned}
 y &= X\beta + \gamma z^* + u \\
 &= X\beta + \gamma(z - \zeta_z) + u \\
 &= X\beta + \gamma z - \gamma \zeta_z + u \\
 &= X\beta + \gamma z + u^0
 \end{aligned}$$

One solution: measured twice.

$$\begin{aligned}
 z_1 &= z^* + v_1 \\
 z_2 &= z^* + v_2
 \end{aligned}$$

$\text{Cov}(v_1, v_1) = 0 \rightarrow \text{Cov}(v_1, z_2) = 0$, and $\text{Cov}(z_1, v_2) = 0$. Therefore, z_1 can be IV for z_2 , and z_2 can be IV for z_1 .

Example: Ashenfelter and Krueger (*AER*, 1994 84(5), 1157-73) twin study. They ask each of the twins their own education level, and their brother/sister education level. Therefore, for each person, they have two measurement of their education. The one from their twin brother/sister can serve as IV for the self-reported education.

Example (f): we are interested in estimating the effect of nutrition on income and labor supply:

$$\log(\text{income}) = a + b^* \text{health} + X\beta + u$$

$$\log(\text{income}) = a + b_1^* \text{carbon} + b_2^* \text{fat} + b_3^* \text{protein} + X\beta + u$$

b_1 is negative, b_2 is insignificant, and b_3 is positive.

Obviously there will be the problem of endogeneity.

Standard IVs: prices variations across regions.

Other IVs: number of people in the family (would affect effective prices of food consumption).

Natural experiment IVs: public policy change in China, such as establishing health insurance in rural China (would nutrition), an increase in university enrollment, etc.

Example (g): we are interested in estimating the following model:

$$\text{Labor supply} = a + b^* \text{number of kids} + X\beta + u$$

“*number of kids*” is obviously endogenous. Other potential IVs include age difference between the husband and the wife.

Standard IVs: whether the first two kids are in the same sex.

Example (h): fertility and labor supply

Gan and Noelia (2009) suggest using the variations in tax schedules across states and across time as *IV*. In particular, they calculate each person's marginal tax rates if she works full time and if she works half time (regardless of the actual working hours), and then using these marginal rates as instrumental variables.

Example (i): demand and supply.

In the example earlier about demand and supply,

$$Q_t = \beta_0 + \beta_1 p_t + Z_t \gamma + u$$

One potential IV is the price of the material. For example, if we are interested in orange juice' price effect on quantity sold, we may use the price of oranges as the IV for the price of the orange juice.

Optimal IV $\hat{\beta}_{IV}$: Two Stage Least Square (2SLS)?

The IV estimate is given by:

$$\hat{\beta}_{IV} = (z' x)^{-1} z' y$$

Question: can we do better?

- (1) Constancy: $\hat{\beta}_{IV} \xrightarrow{p} \beta$. It is consistent, so we can not do better than consistency.
- (2) Efficiency:

Let z be multiplied by a vector or a matrix $z\Gamma$, where Γ could be a matrix or a vector:

$$\begin{aligned} \tilde{\beta}_{IV} &= (\Gamma' z' x)^{-1} \Gamma' z' y \\ &= \beta + (\Gamma' z' x)^{-1} \Gamma' z' u \end{aligned}$$

So, given $\text{Cov}(z'u) = 0$, for any matrix of constants Γ ,

$$\begin{aligned} \tilde{\beta}_{IV} &\rightarrow \beta \\ E(\tilde{\beta}_{IV}) &= \beta \end{aligned}$$

Therefore, there are many *IV* estimators that are unbiased and consistent. Among all estimators, consider a regression:

$$x = z\delta + \varepsilon. \text{ The OLS of this equation yields: } \hat{x}_{OLS} = z\hat{\delta}$$

Let $\Gamma = \hat{\delta}$, we have: $\hat{\beta}_{2SLS} = (\hat{x}'x)^{-1}\hat{x}'y$.

Proposition: This *IV* estimator, $\hat{\beta}_{2SLS} = (\hat{x}'x)^{-1}\hat{x}'y$ turns out to be the same as an OLS estimator of:

$$y = \hat{x}\beta + \eta, \text{ where } \hat{x} = \hat{x}_{OLS} = z\hat{\delta}.$$

Proof:

$$\begin{aligned} \hat{x} &= z(z'z)^{-1}z'x = P_z x \text{ where } P_z = z(z'z)^{-1}z'. P_z \text{ is idempotent and symmetric} \\ P_z'P_z &= z(z'z)^{-1}z'z(z'z)^{-1}z' = z(z'z)^{-1}z' = P_z \\ P_z' &= P_z. \end{aligned}$$

Therefore,

$$\begin{aligned} \hat{x}'x &= x'P_z x \\ &= x'P_z'P_z x \\ &= \hat{x}'\hat{x} \end{aligned}$$

Therefore, $\hat{\beta}_{2SLS} = (\hat{x}'\hat{x})^{-1}\hat{x}'y$, which is equivalent to running the regression:

$$y = \hat{x}\beta + \eta,$$

Proposition: $Var(\hat{\beta}_{2SLS}) \leq Var(\tilde{\beta}_{IV})$:

Proof (for homoscedastic): Define $\tilde{x} = z\Gamma$:

$$\begin{aligned} \tilde{\beta}_{IV} &= (\Gamma'z'x)^{-1}\Gamma'z'y \\ &= (\tilde{x}'x)^{-1}\tilde{x}'y \end{aligned}$$

$$\tilde{\beta}_{IV} = (\tilde{x}'x)^{-1}\tilde{x}'y$$

Asymptotic variance for a generic *IV* estimator is given by:

$$Var(\sqrt{n}\hat{\beta}_{IV}) = \sigma^2[E(\tilde{x}'x)]^{-1}[E(\tilde{x}'\tilde{x})][E(x'\tilde{x})]^{-1}$$

Asymptotic variance for a $\hat{\beta}_{2SLS} = (\hat{x}'x)^{-1}\hat{x}'y$ is given by:

$$\text{Var}(\sqrt{n}\hat{\beta}_{2SLS}) = \sigma^2 [E(\hat{x}'\hat{x})]^{-1}, \text{ where } \hat{x} = z\hat{\delta}.$$

It is sufficient to show that: $[E(\hat{x}'\hat{x})] - [E(\tilde{x}'x)][E(\tilde{x}'\tilde{x})]^{-1}[E(x'\tilde{x})]$ is P.S.D.

Since \hat{x} is a prediction of x , we can let $x = \hat{x} + r$, where r is the residual.

We have: $E(\hat{x}'r) = E(\hat{\delta}'z'r) = 0$, which leads to $E(z'r) = 0$.

Therefore $E(\tilde{x}'r) = E(\Gamma'z'r) = 0$, given $\tilde{x} = z\Gamma$. Therefore,

$$E(\tilde{x}'x) = E(\tilde{x}'(\hat{x} + r)) = E(\tilde{x}'\hat{x})$$

Therefore,

$$\begin{aligned} & [E(\hat{x}'\hat{x})] - [E(\tilde{x}'x)][E(\tilde{x}'\tilde{x})]^{-1}[E(x'\tilde{x})] \\ &= [E(\hat{x}'\hat{x})] - [E(\tilde{x}'\hat{x})][E(\tilde{x}'\tilde{x})]^{-1}[E(\tilde{x}'\hat{x})] \end{aligned}$$

Consider a regression: $\hat{x} = \tilde{x}\eta + s \rightarrow \hat{\eta} = (\tilde{x}'\tilde{x})^{-1}\tilde{x}'\hat{x}$

$$\hat{s} = \hat{x} - \tilde{x}\hat{\eta} = \hat{x} - \tilde{x}(\tilde{x}'\tilde{x})^{-1}\tilde{x}'\hat{x}.$$

Finally,

$$\hat{s}'\hat{s} = \hat{x}'\hat{x} - \hat{x}'\tilde{x}(\tilde{x}'\tilde{x})^{-1}\tilde{x}'\hat{x},$$

which is P.S.D. Note there are four terms when calculating $\hat{s}'\hat{s}$. The other two terms cancel each other.

IV or Regular Model

One may for an alternative model,

$$y = X\beta + Z\gamma + u$$

Instead of using Z as IV for X . In addition to the economic argument of Z should not be entering the equation, there is a different argument. In the first stage regression,

$$X = Z\eta + v = Z_1\eta_1 + Z_2\eta_2 + v$$

If we do not observe Z_2 but only observe Z_1 , and Z_1 and Z_2 are correlated, the first stage regression would be biased, the OLS parameter estimates for η_1 will be biased. One big advantage of the IV regression is that the consistently estimated η_1 will NOT affect the consistency of the parameter β .

For example, consider a return to education model,

$$\ln(\text{wage}) = x\beta + \gamma \text{educ} + u$$

IV regression,

$$\text{educ} = x\delta + a_1 * \text{mother-educ} + a_2 * \text{father-educ} + v$$

The excluded IVs are *mother-educ* and *father-educ*. However, if we do not observe *father-educ*, and given *mother-educ* and *father-educ* are likely to positively correlated, our estimate of a_1 will be biased upward. In another word, one may not say that an increase of one year of *mother-educ* will result in an increase of a_1 year of kid education, because a_1 is overestimated. However, this biased estimated will not affect the consistency of gamma in the main equation.

More discussions of 2SLS²

$$y_1 = x_1\beta_1 + y_2\beta_2 + u, \text{ where } \text{Cov}(x_1, u) = 0, \text{ but } \text{Cov}(y_2, u) \neq 0.$$

In other words, x_1 exogenous, and y_2 endogenous.

Assumption 1: There is a set of *IV*, denoted as z_2 , $E(z_2'u) = 0$.

Assumption 2: (rank condition)

(a) rank $E(z_2'z_2) = L_2$ (b) rank $E(z_2'y_2) = k_2$. It is important to note that $L_2 \geq k_2$.

Condition (b) is critically important. It is equivalent to $\text{Cov}(z_2, y_2) \neq 0$.

For 2SLS, we typically include x_1 in the set of *IV*. So the *IV* is $z = (x_1, z_2)$, and rank $E(z'z) = L$, where $L = k_1 + L_2$.

At the 1st stage: $x = z\delta + \varepsilon$, where $z = (x_1, z_2)$.

For x_1 part of z , $\hat{x}_1 = x_1$. For \hat{y}_2 part of z :

$$\hat{y}_2 = x_1\hat{\delta}_1 + z_2\hat{\delta}_2.$$

At the 2nd stage, run regression of:

$$y = x_1\beta_1 + \hat{y}_2\beta_2 + u$$

$$\rightarrow \hat{\beta}_{2SLS} = (\hat{\beta}_{1,2SLS}, \hat{\beta}_{2,2SLS})$$

² STATA Command: ivreg y Xvars (Yvars = Zvars)

Note: $\hat{\beta}_{1,2SLS} \neq \hat{\beta}_{1,OLS}$, and $\hat{\beta}_{2,2SLS} \neq \hat{\beta}_{2,OLS}$. $\hat{\beta}_{1,2SLS} \neq \hat{\beta}_{1,OLS}$ if $\text{Cov}(x_1, y_2) = 0$.

Assumption 3: (homoscedasticity)

$$E(u^2 z z') = \sigma^2 E(z z'), \text{ where } \sigma^2 = E(u^2)$$

Given *Assumption 3*, define $\hat{u}_{i2SLS} = y_i - X_i \hat{\beta}_{2SLS}$. A consistent estimator of σ^2 under *Assumption 3* is:

$$\hat{\sigma}_{2SLS}^2 = \frac{1}{N-K} \sum_i^N \hat{u}_{i2SLS}^2$$

The $K \times K$ matrix $\hat{\sigma}_{2SLS}^2 (\hat{x}' x)^{-1}$ is a valid estimator of the asymptotic variance of $\hat{\beta}_{2SLS}$.

Heteroskedasticity-Robust inference for 2SLS:

As before, asymptotic variance of $\hat{\beta}_{2SLS}$ can be estimated as

$$\text{Var}(\sqrt{n} \hat{\beta}_{IV}) = (\hat{x}' \hat{x})^{-1} \left[\sum_i^n \hat{u}_{i2SLS}^2 \hat{x}_i' \hat{x}_i \right] (\hat{x}' \hat{x})^{-1}$$

This heteroskedastic-robust estimator can be used anywhere the estimator $\hat{\sigma}^2 (\hat{x}' \hat{x})^{-1}$ is.

Forbidden Regressions:

Consider a model,

$$y_1 = z_1 \delta_1 + \alpha_1 y_2 + \alpha_2 y_2^2 + u_1$$

The model is nonlinear in endogenous variables.

Step 1: $y_2 = z_1 \pi_{21} + z_2 \pi_{22} + v_2$. Let the predicted value be \hat{y}_2

Step 2: run regression $y_1 = z_1 \delta + \alpha_1 \hat{y}_2 + \alpha_2 (\hat{y}_2)^2 + e$.

This regression is sometimes called *forbidden regression*. It is wrong. The reason for this is easy:

$$E(y^2) \neq (E(y))^2$$

What we need is to have the predicted value of $E(y^2)$ in the model, not $(E(y))^2$. This is a common mistake in the empirical literature. We have to treat y^2 as a separate variable from y . If the instrument variable for y is z , then the instrumental variables for y^2 should often include z , and z^2 .

The correct method should be:

Step 1: run two regressions:

$y_2 = z_1\pi_{21} + z_2\pi_{22} + v_2$. Let the predicted value be \hat{y}_2
 $y_2^2 = g(z_1, z_2) + e$. Let the predicted value be \hat{y}_2^2 , and $g(z_1, z_2)$ could be a nonlinear function of z_1 and z_2 . For example,

$$y_2^2 = z_1\eta_1 + z_2\eta_2 + z_1^2\eta_3 + z_2^2\eta_4 + z_1z_2\eta_5 + v$$

And use the predicted value from this regression.

Step 2: run 2SLS as in:

$$y_1 = z_1\delta + \alpha_1\hat{y}_2 + \alpha_2\hat{y}_2^2 + e$$

Example: consider a regression:

$$\ln(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{exper}^2 + \dots + u$$

Suppose we suspect that the experience variable, *exper*, is endogenous. Let the z be a set of IVs for this variable (z could include the local unemployment rate, local industry composition change, etc.). For 2SLS:

Stage 1: run two separate regressions:

$$\begin{aligned} \text{exper} &= a_0 + a_1 \text{educ} + z a_2 + \dots + e_1 & \text{(a)} \\ \text{exper}^2 &= b_0 + b_1 \text{educ} + b_2 \text{educ}^2 + z b_3 + z^2 b_4 + b_5 \text{educ} * z + \dots + e_2 & \text{(b)} \end{aligned}$$

Stage 2: the predicted value from the first stage will be used in the second stage.

It is wrong, however, if one only runs equation (a) and used the square of the predicted *exper* from (a).

A Simple Review of Statistical Test

A test is a statement. For example, consider a simple t-test of $\beta = 0$. Given $H_0: \beta = 0$, and

$H_0: \beta \neq 0$.

Statement 1: accept H_0 if and only if $\left| \frac{\hat{\beta}}{stdc(\hat{\beta})} \right| \leq 1.95$.

However, since the estimated value of $\hat{\beta}$ has random errors, the statement will not right or wrong. Consider the following table:

		Truth	
		H_0	H_1
Action	Accept H_0	No Error	Type II Error
	Accept H_1	Type I Error	No Error

Or, Type I error is equivalent to: *Accept $H_1|H_0$ is true*; and Type II error is equivalent to *Accept $H_0|H_1$ is true*.

In terms of Statement 1, $\Pr(\text{Type I error})$ is given by:

$$\begin{aligned} & \Pr(\text{Accept } H_1 | H_0 \text{ is true}(\beta = 0)) \\ &= \Pr\left(\left| \frac{\hat{\beta}}{sd(\hat{\beta})} \right| > 1.95 \mid H_0 \text{ is true}(\beta = 0)\right) = 0.05 \end{aligned}$$

since when $\beta=0$, then $\frac{\hat{\beta}}{sd(\hat{\beta})}$ has a Student-t distribution.

Two important points may be mentioned here:

- (1) It is possible to have either $\Pr(\text{Type I error}) = 0$ or $\Pr(\text{Type II error}) = 0$, but not both. In fact:

$$\begin{aligned} \Pr(\text{Type I error}) = 0 &\rightarrow \Pr(\text{Type II error}) = 1 \\ \Pr(\text{Type II error} = 0) &\rightarrow \Pr(\text{Type I error}) = 1 \end{aligned}$$

Statement 2: always accept H_0 regardless of the truth.

This statement implies that there will be no Type I error. However, it also implies that $\Pr(\text{Type II error}) = 1$.

Statement 3: always accept H_1 regardless of the truth.

This statement implies that there will be no Type II error. However, it also implies

that $\Pr(\text{Type I error}) = 1$.

In fact, it is generally true that a lower type I error would imply a higher Type II error, and a vice versa. Therefore, a test is actually a compromise between Type I error and Type II error.

It is the default or tradition that we let:

$$\Pr(\text{Type I error}) = 0.01, 0.05, \text{ or } 0.10.$$

There are often many different tests that can achieve $\Pr(\text{Type I error}) = 0.01, 0.05,$ or 0.10 . The goal is to find a test that may have the minimum Type II error, or equivalently, the maximum power at a given size.

(2) How to determine H_0 and H_1 :

It is often NOT arbitrary to assign H_0 and H_1 . The most important factor to determine which hypothesis is H_0 is the distribution of the test statistic under H_0 . Such distribution of the test statistic should be generic and standard, and not parameter-dependent. The most often used distributions include:

- (a) standard normal distribution.
- (b) Student-t distribution.
- (c) F-distribution
- (d) $\chi^2(k)$ distribution.

For example, consider a regression:

$$y = x_1\beta_1 + x_2\beta_2 + u = x\beta + u$$

If we are interested in testing: $H_0: R\beta = 0$ where R is a matrix of constants.

$$\text{Var}(R\hat{\beta}) = R\text{Var}(\hat{\beta})R'$$

Therefore, asymptotically, and under H_0 : $R\hat{\beta} \sim N(0, R\text{Var}(\hat{\beta})R')$. It is typical to transform a k -dimension multivariate normal distribution to a $\chi^2(k)$ distribution:

The test statistic is given by: $(R\hat{\beta})[R\text{Var}(\hat{\beta})R']^{-1}(R\hat{\beta}) \sim \chi_k^2$

Based on this test statistic, we can construct a test. An example is:

Statement 4: Reject H_0 if the test statistic $(R\hat{\beta})[R\text{Var}(\hat{\beta})R']^{-1}(R\hat{\beta}) > \text{critical value}$. The critical value is obtained using the distribution of $\chi^2(k)$.

Testing for Endogeneity

Consider a model,

$$y_1 = z_1\delta_1 + \alpha_1 y_2 + u_1 \quad (*)$$

where y_2 is potentially endogenous, and z_2 is a set of instrument variables. One can use the Hausman test.

(1) Hausman test:

Consider H_0 : y_2 is exogenous.

Under H_0 , both $\hat{\alpha}_{OLS}$ and $\hat{\alpha}_{2SLS}$ are consistent. The difference between these two estimators are their covariance matrix. Asymptotically, both estimators should have normal distributions. Therefore, it is necessary that the difference of the two estimates will have zero mean and normal distributions, i.e.,

$$\hat{\alpha}_{OLS} - \hat{\alpha}_{2SLS} \sim N(0, \text{Var}(\hat{\alpha}_{OLS} - \hat{\alpha}_{2SLS}))$$

Therefore, we can construct the test statistic under H_0 :

(i) Normalize to a vector of standard normal:

$$[\text{Var}(\hat{\alpha}_{OLS} - \hat{\alpha}_{2SLS})]^{-\frac{1}{2}}(\hat{\alpha}_{OLS} - \hat{\alpha}_{2SLS}) \sim N(0, \text{Var}(\hat{\alpha}_{OLS} - \hat{\alpha}_{2SLS}))$$

(ii) Construct the test statistic with $\chi^2(k_2)$, where k_2 is the dimension of α_1 .

$$(\hat{\alpha}_{OLS} - \hat{\alpha}_{2SLS})' [\text{Var}(\hat{\alpha}_{OLS} - \hat{\alpha}_{2SLS})]^{-1} (\hat{\alpha}_{OLS} - \hat{\alpha}_{2SLS}) \sim \chi^2(k_2)$$

However, the difficulty of the previous procedure is to calculate the variance of the difference, $\text{Var}(\hat{\alpha}_{OLS} - \hat{\alpha}_{2SLS})$.

Note that the variance of the difference do not equal to the difference of the variances, nor the sum of the variance, because the covariance of the two estimators not zero.

(a) A regression based test:

Alternatively, one can try a regression-based test. To derive this test, consider the IV regression,

$$y_2 = z_1\pi_1 + v_2 = z_1\pi_{21} + z_2\pi_{22} + v_2$$

The endogeneity of y_2 is equivalent to $\text{Cov}(u_1, v_2) \neq 0$. Let

$$u_1 = \rho_1 v_2 + \varepsilon_2$$

Then the endogeneity of y_2 is equivalent to $\rho_1 \neq 0$. Plug this into (*):

$$y_1 = z_1 \delta_1 + \alpha_1 y_2 + \rho_1 v_2 + \varepsilon_2$$

Although we do not observe v_2 , we can use the residual from the IV regression instead.

Step 1: *IV* regression, $y_2 = z_1 \pi_{21} + z_2 \pi_{22} + v_2$

Let the residual from this regression be: \hat{v}_2 .

Step 2: Regression:

$$y_1 = z_1 \delta_1 + \alpha_1 y_2 + \rho_1 \hat{v}_2 + \text{error} \quad (**)$$

One can use t -statistic in the usual sense to test $H_0: \rho_1 = 0$ as a test of endogeneity of y_2 .

However, if H_0 is rejected, i.e., $\rho_1 \neq 0$, then the standard error calculated from (**) is not valid. The reason is because we use \hat{v}_2 instead of v_2 .

Testing Overidentifying Restrictions

Again, consider the model

$$y_1 = z_1 \delta_1 + \alpha_1 y_2 + u_1$$

with z_2 being a set of instrument variables.

Let $\text{dimension}(y_2) = K_2$, and $\text{dimension}(z_2) = L_2$. If $L_2 > K_2$, we have more IVs than necessary. Can we use this to test if additional *IVs* are valid *IVs*?

The basic idea is: Suppose we divide the set of IVs (z_2) into two groups, z_{21} and z_{22} . If $\text{dimension}(z_{21}) = K_2$, then we can obtain different *IV* estimates, $\hat{\beta}_{2SLS}(z_{21})$ if $\text{dimension}(z_{21}) \geq K_2$ and $\hat{\beta}_{2SLS}(z_{22})$ if $\text{dimension}(z_{22}) \geq K_2$. Intuitively, under H_0 : z_{21} and z_{22} both are sets of valid IVs, the estimates $\hat{\beta}_{2SLS}(z_{21}) \approx \hat{\beta}_{2SLS}(z_{22})$.

Alternatively, one can compare $\hat{\beta}_{2SLS}(z_{21})$ with the estimates using all available

IVs, $\hat{\beta}_{2SLS}(z_2)$. Under H_0 , we must have: $\hat{\beta}_{2SLS}(z_{21}) \approx \hat{\beta}_{2SLS}(z_2)$. The only difference is their sampling errors.

We can construct a test for this easily:

$$(\hat{\beta}_{2SLS}(z_{21}) - \hat{\beta}_{2SLS}(z_{22}))' [Var(\hat{\beta}_{2SLS}(z_{21}) - \hat{\beta}_{2SLS}(z_{22}))]^{-1} (\hat{\beta}_{2SLS}(z_{21}) - \hat{\beta}_{2SLS}(z_{22})) \sim \chi^2$$

The difficulty of this type of test is that inverse of the covariance matrix of the difference of the estimates. Note that it typically does not equal to the difference of the inverse. More particularly, suppose we have two estimates, under H_0 , both are consistent:

$$[Var((\hat{\beta}_{2SLS}(z_{21}) - \hat{\beta}_{2SLS}(z_{22})))^{-1} \neq [Var(\hat{\beta}_{2SLS}(z_{21}))]^{-1} - [Var(\hat{\beta}_{2SLS}(z_{22}))]^{-1}$$

Alternatively, we can conduct a regression-based test:

Step 1: run 2SLS using all IVs, let the residual be \hat{u}_1

Step 2: run OLS of \hat{u}_1 on all instruments z . Under H_0 , the R^2 from this regression has a distribution:

$$NR_u^2 \xrightarrow{d} \chi_{L_2 - K_2}^2$$

A larger value rejects the H_0 . $E(z'u_1) = 0$.

Testing for Functional Form

Consider a model,

$$y = x\beta + u, \quad \text{and} \quad E(u|x) = 0.$$

However, it is possible that the quadratic form of x or even higher order of x may be more appropriate:

$$y = x\beta_1 + x^2\beta_2 + \dots + u, \quad \text{and} \quad E(u|x) = 0.$$

A test includes two steps:

Step 1: a linear regression of $y = x\beta + u$. The predicted value of y is given by: $\hat{y} = x\hat{\beta}$,

and $\hat{u} = y - x\hat{\beta}$.

Step 2: Regress \hat{u} on $x, \hat{y}, \hat{y}^2, \hat{y}^3, \hat{y}^4$ as a test of neglected nonlinearity.

Testing for Heteroskedasticity

Consider a model,

$$y = x\beta + u, \quad \text{and} \quad E(u|x) = 0$$

$$H_0: E(u^2|x) = \sigma^2. \quad H_1: E(u^2|x) = h(x).$$

A general way to test heteroskedasticity is to conduct a regression:

$$u_i^2 = h(x_i)\delta + v_i \quad (*)$$

We can apply an F test for the null that $\delta = 0$. In practice, since u_i^2 is not observed, one may use the residual \hat{u}_i^2 .

One thing to notice here is that v_i cannot be normally distributed under H_0 . The OLS estimates of (*) is consistent.

Two popular tests are special cases of (*).

- (i) $h(x_i) = (1, x_i)$
- (ii) $h(x_i) = (1, \hat{y}_i, \hat{y}_i^2)$

The Difference-in-difference method

We are interested in average changes in outcome y after a policy change. As in the case of the medical field, we are interested in the outcomes of a new medicine, including its effectiveness and its side effect. To do that, it is typical that we randomly divide patients into two groups, the treatment group, and the control group. The treatment group is given the medicine while the control group is given the “placebo”.

Consider an economic example. An important policy question is how to help needy families. Income transferring programs, such as Aid to Families with Children (AFDC) creates disincentives of working. One alternative method is the Earned Income Tax Credit (EITC).

Eligibility for EITC: Gross income below a specified amount (in 2007, the

amount is \$39,783 if you children and \$14,590 if you do not have children).

Benefits (in 2007): maximum benefits: \$428 if no children; \$2,853 if one child; and \$4,716 if two children.

The Tax Reform Act of 1986 includes an expansion of earned income tax credit. We are interested if expansion of EITC helps increasing labor supply.

Denote 1 if with treatment (experiences expansions of EITC), and 0 without treatment (not affected by expansions of EITC). Average Treatment Effect (*ATE*) is defined as:

$$ATE = E(y_1 - y_0) \quad (*)$$

The difficulty in estimating (1) is that we observe either y_1 or y_0 , not both, for each person. However, we potentially can observe outcomes before the treatment and after the treatment for the same person or for different persons. We have two time periods, say year 0 and year 1. One would say that we can simply apply (*). In the case of EITC expansion, year 0 is before 1986, and year 1 is after 1986.

However, this is not entirely appropriate since there may be other factors that affect treated people as well. The difference in labor supply before 1986 and after 1986 may be due to overall economic environment.

Therefore, we need a control group. There are two groups, the control group (denoted as group A), and the treatment group (denoted as B). At period 0, no treatment for both groups. At period 1, the treatment group experiences policy change (treatment) while the control group does not. Let D_1 denote a dummy variable for time period 1, and D_B denote the treatment group. The simplest regression for analyzing the impact of the policy change is:

$$y = \beta_0 + \delta_0 D_1 + \beta_1 D_B + \delta_1 D_1 * D_B + u \quad (2)$$

It is easy to show that:

$$\hat{\delta}_1 = (\bar{y}_{B,1} - \bar{y}_{B,0}) - (\bar{y}_{A,1} - \bar{y}_{A,0}) \quad (3)$$

This is why the regression of (2) is often called the difference-in-difference.

In the example of the expansions of EITC, Eissa and Liebman use “*single women without children*” as the control group, and “*single women with children*” as the treatment group. Their time periods are: 1984-1986 as time 0, and 1988-1990 as time 1.

The regression, therefore, is:

$$\Pr(lfp_{it} = 1) = \Phi(\alpha + \beta Z_{it} + \gamma_0 ChildrenDummy_i + \gamma_0 post86_t + \gamma_2 (ChildrenDummy_i \times post86_t))$$

Eissa and Liebman (1996) find that single women with children increased their relative labor force participation by up to 2.8% percentage points.

Spatial Dependence

Consider a model,

$$y_{is} = x_{is}\beta + z_s\gamma + q_s + e_{is}$$

where i is for individual and s is for stratum. The covariates in x_{is} change with the individual, while z_s change only at the strata level. If q_s is not observed, then the presence of q_s induces correlation in the composite error $u_{is} = q_s + e_{is}$ within each stratum.

This is often called clustered sample.

Weak IV problem

1. The Problem of Weak IVs:

Now consider the more general model:

$$y = Y\beta + X\gamma + u \quad (3)$$

$$Y = Z\Pi + X\Phi + V \quad (4)$$

In (3), $Y_{N \times n}$ is endogenous, i.e., there are n endogenous variables in (3). Note the set of variables X is exogenous.

In (4), $Z_{N \times K_2}$ is a set of excluded exogenous instrumental variables. $K_2 \geq n$.

Most of the empirical work, including the “natural experiment,” concentrates on the requirement of the *IV* that $\text{Cov}(Z, u) = 0$. However, when $\text{Cov}(Z, Y)$ is small – we have a weak *IV* problem. As pointed in Stock, Wright, and Yogo (2002), weak IVs create serious problems.

- The sampling distributions of the estimates are in general non-normal. Hypothesis tests based on standard methods are not reliable.
- It is not useful to think of weak instruments as a “small sample” problem. Bound, Jaeger, and Baker (1995) provided an empirical example of weak IV despite having 329,000 observations.
- There are methods that are more robust to weak *IV* than conventional methods.

2. A Test of Weak IVs

Consider the first stage regression. Let $M_X = I - X(X'X)^{-1}X'$, and let: $Z^\perp = M_X Z$, which is the residual of the regression of Z on X . Similarly, $Y^\perp = M_X Y = Y - X(X'X)^{-1}X'Y$ is the residual of the regression of Y on X .

In addition, let $P_X = X(X'X)^{-1}X'$, then: $P_{Z^\perp} = Z^\perp (Z^{\perp\prime} Z^\perp)^{-1} Z^{\perp\prime}$. According to the partitioned regression of (4), we have:

$$\hat{\Pi}_{OLS} = (Z^{\perp\prime} Z^\perp)^{-1} Z^{\perp\prime} Y$$

Therefore,

$$\begin{aligned} Y^{\perp\prime} P_{Z^\perp} Y^\perp &= Y' M_X Z^\perp (Z^{\perp\prime} Z^\perp)^{-1} Z^{\perp\prime} M_X Y \\ &= Y' M_X M_X Z (Z' M_X M_X Z)^{-1} Z' M_X M_X Y \\ &= Y' Z^\perp (Z^{\perp\prime} Z^\perp)^{-1} Z^{\perp\prime} Y \\ &= Y' Z^\perp (Z^{\perp\prime} Z^\perp)^{-1} (Z^{\perp\prime} Z^\perp) (Z^{\perp\prime} Z^\perp)^{-1} Z^{\perp\prime} Y \\ &= \hat{\Pi}'_{OLS} (Z^{\perp\prime} Z^\perp) \hat{\Pi}_{OLS} \end{aligned}$$

Intuitively, a test of the correlation between Z and Y should be a test of $\Pi_{OLS} = 0$. Therefore, parallel to the F -test of testing the restriction of $\Pi_{OLS} = 0$, the proposed test for weak instruments is based on the eigenvalue of the matrix analog of the F -statistic from the 1st-stage regression:

$$G_T = \hat{\Sigma}_{VV}^{-1/2\prime} Y^{\perp\prime} P_{Z^\perp} Y^\perp \hat{\Sigma}_{VV}^{-1/2} / K_2, \quad \text{where } \hat{\Sigma}_{VV} = \frac{Y' M_Z Y}{n - K_1 - K_1}$$

The test statistic is the minimum eigenvalue of G_T :

$$g_{min} = \text{min-eigenvalue}(G_T)$$

To find eigenvalues, just solve the equation: $\det(G_T - \lambda I) = 0$ to get values of λ . A larger value of g_{min} would reject the H_0 (weak instrument).

If g_{min} is close to zero, then the model is unidentified (this is what Cragg-Donald statistics was for originally)

For G_T , for the case of one endogenous variable and one exogenous variables, X is a constant. $Y^\perp = Y$, $Z^\perp = Z$, and $\hat{\Sigma}_{VV}^{-1/2} = \hat{\sigma}_v^{-1}$

$$G_T = \hat{\Sigma}_{VV}^{-1/2\prime} Y' P_Z Y \hat{\Sigma}_{VV}^{-1/2} / K_2 = \frac{\hat{Y}' \hat{Y}}{\hat{\sigma}_v^2},$$

where \hat{Y} is the predicted value from the 1st stage regression of Y on Z .

A test of weak IVs is to compare the *Cragg-Donald* Statistic g_{min} with the critical values listed in following tables (from Stock and Yogo, 2004).

Example: suppose we have one endogenous variables ($n = 1$) and 3 IVs ($K_2 = 2$) at 5% significance level.

The critical values are:

2SLS, tolerance of bias 5% :	13.91 (Table 1)
2SLS, tolerance of bias 10%:	9.08 (Table 1)
2SLS, size = 10%:	22.30 (Table 2)
LIML, size = 10%:	6.46 (Table 4)

3. Robust Estimators to Weak IVs

If weak IVs are detected using 2SLS, there are alternative estimators that are more robust against the weak IVs.

Consider the k -class estimators. Note again by partitioned regression, the OLS estimator of β in (3) is given by:

$$\begin{aligned}\hat{\beta}_{OLS} &= (Y^\perp{}' Y^\perp)^{-1} Y^\perp{}' y \\ &= (Y' M_X Y)^{-1} Y' M_X y\end{aligned}$$

In other words, $\hat{\beta}_{OLS}$ is obtained by running a regression of y on the Y^\perp , which is the residual of the regression of Y on X .

The k -class estimator of β is given by:

$$\hat{\beta}(k) = (Y^\perp{}' (I - kM_{Z^\perp}) Y^\perp)^{-1} Y^\perp{}' (I - kM_{Z^\perp}) y^\perp$$

where $Z^\perp = M_X Z$, which is the residual of the regression of Z on X .
and

$$\begin{aligned}M_{Z^\perp} &= I - Z^\perp (Z^\perp{}' Z^\perp)^{-1} Z^\perp{}' \\ &= I - M_X Z (Z' M_X Z)^{-1} Z' M_X\end{aligned}$$

It has been shown that:

OLS:	$k = 0$.
2SLS:	$k = 1$.
LIML:	$k = \hat{k}_{LIML}$, the smallest root of $\det(Y' M_X Y - k Y' M_Z Y) = 0$

$$\text{Fuller - } k: \quad \hat{k}_{LIML} = \frac{c}{N - K_1 - K_2}, \text{ where } c \text{ is a positive constant}$$

$$\text{B2SLS:} \quad k = T / (T - K_2 + 2)$$

From the tables, it is important to notice that the critical value is much smaller for *LIML* than for *2SLS*. Therefore, it is much easier for *LIML* to be absent from weak IV problem than *2SLS* is. Therefore, *LIML* is more robust than *2SLS* against the weak IV problems.

Limited information maximum likelihood (LIML) estimator:

This estimator is based on a single equation under the assumption of normally distributed disturbances; LIML is efficient among single-equation estimators.

One of the results emerges from the derivation is that the LIML estimator has the same asymptotic distribution as the 2SLS estimator, and the latter does not rely on an assumption of normality.

STATA Deviation:

The command in STATA to do this is:

```
IVREG2 y X (Y = Z), ffirst
```

The options `ffirst` produces Cragg-Donald statistics g_{min} . If the weak IV problem exists, one may try:

```
IVREG2 y X (Y = Z), liml
```

Table 1.
Critical Values for the Weak Instrument Test Based on TSLS Bias
Significance level is 5%

K_2	$n = 1, b =$				$n = 2, b =$				$n = 3, b =$			
	0.05	0.10	0.20	0.30	0.05	0.10	0.20	0.30	0.05	0.10	0.20	0.30
3	13.91	9.08	6.46	5.39
4	16.85	10.27	6.71	5.34	11.04	7.56	5.57	4.73
5	18.37	10.83	6.77	5.25	13.97	8.78	5.91	4.79	9.53	6.61	4.99	4.30
6	19.28	11.12	6.76	5.15	15.72	9.48	6.08	4.78	12.20	7.77	5.35	4.40
7	19.86	11.29	6.73	5.07	16.88	9.92	6.16	4.76	13.95	8.50	5.56	4.44
8	20.25	11.39	6.69	4.99	17.70	10.22	6.20	4.73	15.18	9.01	5.69	4.46
9	20.53	11.46	6.65	4.92	18.30	10.43	6.22	4.69	16.10	9.37	5.78	4.46
10	20.74	11.49	6.61	4.86	18.76	10.58	6.23	4.66	16.80	9.64	5.83	4.45
11	20.90	11.51	6.56	4.80	19.12	10.69	6.23	4.62	17.35	9.85	5.87	4.44
12	21.01	11.52	6.53	4.75	19.40	10.78	6.22	4.59	17.80	10.01	5.90	4.42
13	21.10	11.52	6.49	4.71	19.64	10.84	6.21	4.56	18.17	10.14	5.92	4.41
14	21.18	11.52	6.45	4.67	19.83	10.89	6.20	4.53	18.47	10.25	5.93	4.39
15	21.23	11.51	6.42	4.63	19.98	10.93	6.19	4.50	18.73	10.33	5.94	4.37
16	21.28	11.50	6.39	4.59	20.12	10.96	6.17	4.48	18.94	10.41	5.94	4.36
17	21.31	11.49	6.36	4.56	20.23	10.99	6.16	4.45	19.13	10.47	5.94	4.34
18	21.34	11.48	6.33	4.53	20.33	11.00	6.14	4.43	19.29	10.52	5.94	4.32
19	21.36	11.46	6.31	4.51	20.41	11.02	6.13	4.41	19.44	10.56	5.94	4.31
20	21.38	11.45	6.28	4.48	20.48	11.03	6.11	4.39	19.56	10.60	5.93	4.29
21	21.39	11.44	6.26	4.46	20.54	11.04	6.10	4.37	19.67	10.63	5.93	4.28
22	21.40	11.42	6.24	4.43	20.60	11.05	6.08	4.35	19.77	10.65	5.92	4.27
23	21.41	11.41	6.22	4.41	20.65	11.05	6.07	4.33	19.86	10.68	5.92	4.25
24	21.41	11.40	6.20	4.39	20.69	11.05	6.06	4.32	19.94	10.70	5.91	4.24
25	21.42	11.38	6.18	4.37	20.73	11.06	6.05	4.30	20.01	10.71	5.90	4.23
26	21.42	11.37	6.16	4.35	20.76	11.06	6.03	4.29	20.07	10.73	5.90	4.21
27	21.42	11.36	6.14	4.34	20.79	11.06	6.02	4.27	20.13	10.74	5.89	4.20
28	21.42	11.34	6.13	4.32	20.82	11.05	6.01	4.26	20.18	10.75	5.88	4.19
29	21.42	11.33	6.11	4.31	20.84	11.05	6.00	4.24	20.23	10.76	5.88	4.18
30	21.42	11.32	6.09	4.29	20.86	11.05	5.99	4.23	20.27	10.77	5.87	4.17

Notes: The test rejects if g_{\min} exceeds the critical value. The critical value is a function of the number of included endogenous regressors (n), the number of instrumental variables (K_2), and the desired maximal bias of the IV estimator relative to OLS (b).

Table 2.
Critical Values for the Weak Instrument Test Based on TSLS Size
Significance level is 5%

K_2	$n = 1, r =$				$n = 2, r =$			
	0.10	0.15	0.20	0.25	0.10	0.15	0.20	0.25
1	16.38	8.96	6.66	5.53				
2	19.93	11.59	8.75	7.25	7.03	4.58	3.95	3.63
3	22.30	12.83	9.54	7.80	13.43	8.18	6.40	5.45
4	24.58	13.96	10.26	8.31	16.87	9.93	7.54	6.28
5	26.87	15.09	10.98	8.84	19.45	11.22	8.38	6.89
6	29.18	16.23	11.72	9.38	21.68	12.33	9.10	7.42
7	31.50	17.38	12.48	9.93	23.72	13.34	9.77	7.91
8	33.84	18.54	13.24	10.50	25.64	14.31	10.41	8.39
9	36.19	19.71	14.01	11.07	27.51	15.24	11.03	8.85
10	38.54	20.88	14.78	11.65	29.32	16.16	11.65	9.31
11	40.90	22.06	15.56	12.23	31.11	17.06	12.25	9.77
12	43.27	23.24	16.35	12.82	32.88	17.95	12.86	10.22
13	45.64	24.42	17.14	13.41	34.62	18.84	13.45	10.68
14	48.01	25.61	17.93	14.00	36.36	19.72	14.05	11.13
15	50.39	26.80	18.72	14.60	38.08	20.60	14.65	11.58
16	52.77	27.99	19.51	15.19	39.80	21.48	15.24	12.03
17	55.15	29.19	20.31	15.79	41.51	22.35	15.83	12.49
18	57.53	30.38	21.10	16.39	43.22	23.22	16.42	12.94
19	59.92	31.58	21.90	16.99	44.92	24.09	17.02	13.39
20	62.30	32.77	22.70	17.60	46.62	24.96	17.61	13.84
21	64.69	33.97	23.50	18.20	48.31	25.82	18.20	14.29
22	67.07	35.17	24.30	18.80	50.01	26.69	18.79	14.74
23	69.46	36.37	25.10	19.41	51.70	27.56	19.38	15.19
24	71.85	37.57	25.90	20.01	53.39	28.42	19.97	15.64
25	74.24	38.77	26.71	20.61	55.07	29.29	20.56	16.10
26	76.62	39.97	27.51	21.22	56.76	30.15	21.15	16.55
27	79.01	41.17	28.31	21.83	58.45	31.02	21.74	17.00
28	81.40	42.37	29.12	22.43	60.13	31.88	22.33	17.45
29	83.79	43.57	29.92	23.04	61.82	32.74	22.92	17.90
30	86.17	44.78	30.72	23.65	63.51	33.61	23.51	18.35

Notes: The test rejects if g_{\min} exceeds the critical value. The critical value is a function of the number of included endogenous regressors (n), the number of instrumental variables (K_2), and the desired maximal size (r) of a 5% Wald test of $\beta = \beta_0$.

Table 4.
Critical Values for the Weak Instrument Test Based on LIML Size
Significance level is 5%

K_2	$n = 1, r =$				$n = 1, r =$			
	0.10	0.15	0.20	0.25	0.10	0.15	0.20	0.25
1	16.38	8.96	6.66	5.53
2	8.68	5.33	4.42	3.92	7.03	4.58	3.95	3.63
3	6.46	4.36	3.69	3.32	5.44	3.81	3.32	3.09
4	5.44	3.87	3.30	2.98	4.72	3.39	2.99	2.79
5	4.84	3.56	3.05	2.77	4.32	3.13	2.78	2.60
6	4.45	3.34	2.87	2.61	4.06	2.95	2.63	2.46
7	4.18	3.18	2.73	2.49	3.90	2.83	2.52	2.35
8	3.97	3.04	2.63	2.39	3.78	2.73	2.43	2.27
9	3.81	2.93	2.54	2.32	3.70	2.66	2.36	2.20
10	3.68	2.84	2.46	2.25	3.64	2.60	2.30	2.14
11	3.58	2.76	2.40	2.19	3.60	2.55	2.25	2.09
12	3.50	2.69	2.34	2.14	3.58	2.52	2.21	2.05
13	3.42	2.63	2.29	2.10	3.56	2.48	2.17	2.02
14	3.36	2.57	2.25	2.06	3.55	2.46	2.14	1.99
15	3.31	2.52	2.21	2.03	3.54	2.44	2.11	1.96
16	3.27	2.48	2.18	2.00	3.55	2.42	2.09	1.93
17	3.24	2.44	2.14	1.97	3.55	2.41	2.07	1.91
18	3.20	2.41	2.11	1.94	3.56	2.40	2.05	1.89
19	3.18	2.37	2.09	1.92	3.57	2.39	2.03	1.87
20	3.21	2.34	2.06	1.90	3.58	2.38	2.02	1.86
21	3.39	2.32	2.04	1.88	3.59	2.38	2.01	1.84
22	3.57	2.29	2.02	1.86	3.60	2.37	1.99	1.83
23	3.68	2.27	2.00	1.84	3.62	2.37	1.98	1.81
24	3.75	2.25	1.98	1.83	3.64	2.37	1.98	1.80
25	3.79	2.24	1.96	1.81	3.65	2.37	1.97	1.79
26	3.82	2.22	1.95	1.80	3.67	2.38	1.96	1.78
27	3.85	2.21	1.93	1.78	3.74	2.38	1.96	1.77
28	3.86	2.20	1.92	1.77	3.87	2.38	1.95	1.77
29	3.87	2.19	1.90	1.76	4.02	2.39	1.95	1.76
30	3.88	2.18	1.89	1.75	4.12	2.39	1.95	1.75

See the notes to Table 2.