Endogeneity Notes

Questions

- Would we get full credit for stating the omitted variable bias formula without proof? What about for the measurement error least squares attenuation bias?
  - In questions that ask about the effect of omitting a variable on the remaining coefficient estimates, it is generally not required to prove either formula, unless asked to "show", "prove" or "derive".
- What happens if we think some control is both causal and a proxy?
  - Tactically, simply choose to interpret the control as a proxy.
- If endogeneity in one regressor is sufficient to render OLS estimators of the other parameters inconsistent, are endogenous controls in the broader sense (rather than the narrower sense of being a post-treatment characteristic) also problematic? (See 200529 Q7a).
  - No, these additional controls can be treated as proxies.
- Regarding Problem Set 6 Q2a, we establish that p is correlated with u and v by expressing p in terms of t, u, v and finding $cov(p,u), cov(p,v)$. If we do the same for t, i.e. express t in terms of p, u, v, we find some complex equations for $cov(t,u), cov(t,v)$. Why do we not then also conclude that it is not plausible to treat t as exogenous?
  - "Simultaneity is easy, you have two equations, you solve, then you show that the regressor of interest is correlated with the error in the causal equation (with that regressor)."
  - We do not also conclude that t is endogenous because we are not given an equation or model that determines t. So it remains open to argument that t is determined in a way that is uncorrelated with the unobserved determinants.

Endogeneity

- In the causal model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + u$, regressor X_I (where $I \in \{1, \ldots, k\}$) is endogenous iff orthogonality fails and X_I is the offending regressor, i.e. $cov(X_I, u) \neq 0$.
- The three sources of endogeneity studied in this course are (1) omitted variables, (2) measurement error, and (3) simultaneity.
- In general, endogeneity of one regressor is sufficient to render OLS estimators of every causal parameter inconsistent.

Omitted Variables

- Consider the causal model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$. The causal parameter of interest is β_1. Suppose that orthogonality holds in this causal model, i.e. $\mathbb{E}u = cov(X_1, u) = cov(X_2, u) = 0$. Then the parameters of this causal model coincide with those of the population regression model (of Y on X_1, X_2) and can be consistently estimated by OLS regression. The causal model can be rewritten as $Y = \beta_0 + \beta_1 X_1 + \epsilon$, where $\epsilon := \beta_2 X_2 + u$. The OLS estimator for β_1 is consistent only if orthogonality holds for X_1, i.e. $cov(X_1, \epsilon) = cov(X_1, \beta_2 X_2 + u) = \beta_2 cov(X_1, X_2) = 0$ where the second equality holds by the supposition of orthogonality in the "long" causal model. In general, neither $\beta_2 = 0$ nor $cov(X_1, X_2) = 0$, hence the OLS estimator for β_1 is not consistent.
- Consider the population regression model $Y = \gamma_0 + \gamma_1 X_1 + e$ corresponding to the "short" regression model. By construction, $\gamma_1 = \frac{cov(Y, X_1)}{var(X_1)}$. The consistent OLS estimator for γ_1 is $\hat{\gamma}_1 = \frac{c\hat{o}v(Y, X_1)}{v\hat{a}r(X_1)}$.
  - $$\begin{aligned} cov(Y, X_1) &= cov(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + u, X_1) \\ &= \beta_1 var(X_1) + \beta_2 cov(X_1, X_2) \end{aligned}.$$
  - Then,
  - $$\begin{aligned} \gamma_1 &= \frac{cov(Y, X_1)}{var(X_1)} \\ &= \beta_1 + \beta_2 \frac{cov(X_1, X_2)}{var(X_1)}, \\ &= \beta_1 + \beta_2 \pi_1 \end{aligned}$$
  - where π_1 is the population regression coefficient in the regression of X_2 on X_1.
  - This is the omitted variable bias formula. Even in large samples, $\hat{\gamma}_1$ yields a biased (by $\beta_2 \gamma_1$) estimate of β_1, $\hat{\gamma}_1$ is an inconsistent estimator for β_1.
- If the magnitude of $\beta_2 \pi_1$ is thought to be small (which is if either the causal effect of X_2 on Y or the correlation between X_2 and X_1 is thought to be small), particularly relative to the magnitude of sampling variability in $\hat{\gamma}_1$ (that is measured by the standard error), it is reasonable to neglect the omitted variable bias.

- The omitted variable bias formula is useful for determining the direction of omitted variable bias. If the direction of omitted variable bias is positive, the estimate of the causal effect can be regarded as an upper bound on the causal effect. The reverse is true if the direction of bias is negative.
- The above argument generalises as follows. Given the causal model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + 0$, supposing that orthogonality holds in this causal model, and the population regression model $Y = \gamma_0 + \gamma_1 X_1 + \ldots + \gamma_{k-1} X_{k-1} + e$, the OLS estimator $\hat{\gamma}_1$ is consistent for $\gamma_1 = \beta_1 + \beta_k \pi_1$, where π_1 is the coefficient on X_1 in the population linear regression of X_k on $X_1, \ldots, X_k$.

Proxying for Omitted Variables

- Consider again the causal model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$, where the causal parameter of interest again is β_1. Continue to suppose that orthogonality holds in this causal model. Suppose that X_2 is unobservable, but the proxy W_1 is observable, and W_1 is plausibly correlated with X_2. Suppose further that W_1 is orthogonal to u, i.e. $cov(u, W_1) = 0$, i.e. W_1 is not itself a determinant of Y.
- Let $X_2 = \delta_0 + \delta_1 W_1 + e$ denote the auxiliary population linear regression of X_2 on W_1 (where $\mathbb{E}e = cov(W_1, e) = 0$ by construction).
- By substitution,
  - $$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2(\delta_0 + \delta_1 W_1 + e) + u \\ &= (\beta_0 + \beta_2 \delta_0) + \beta_1 X_1 + \beta_2 \delta_1 W_1 + (u + \beta_2 e). \\ &=: \gamma_0 + \beta_1 X_1 + \gamma_2 W_1 + \eta \end{aligned}$$
  - This population regression model is the estimable model.
- The OLS regression coefficient on X_1 in the regression of Y on X_1, W_1 is consistent for β_1 iff $cov(X_1, \eta) = cov(W_1, \eta) = 0$.
  - $cov(W_1, \eta) = cov(W_1, u + \beta_2 e) = cov(W_1, u) + \beta_2 cov(W_1, e) = 0$, where the final equality holds by supposition and by construction of the auxiliary population linear regression.
  - $cov(X_1, \eta) = cov(X_1, u + \beta_2 e) = \beta_2 cov(X_1, e)$.
- Supposing that $\beta_2 \neq 0$, the necessary condition (for the consistency of the OLS estimator for β_1) holds iff $cov(X_1, e) = 0$, which is iff the component of X_2 not predicted by the instrument W_1 is entirely uncorrelated with X_1. Informally, this is iff W_1 controls for the correlation between X_1 and X_2.
- If this further condition is satisfied, W_1 is a valid proxy for X_2. This means that β_1 is recoverable by a population linear regression of Y on X_1 and W_1 in place of X_2. This causal effect is consistently estimated by OLS regression.
- The above argument generalises as follows. $W_1, \ldots, W_l$ are valid proxies for X_k iff (1) $W_1, \ldots, W_l$ are plausibly correlated with X_k, (2) each of $W_1, \ldots, W_l$ is orthogonal to u, (3) the residual e in the regression of X_k on $W_1, \ldots, W_l$ is uncorrelated with $X_1, \ldots, X_{k-1}$, (where the causal effects of these regressors are the causal effects of interest).
- The OLS regression coefficients on the proxy variables generally cannot be given a causal interpretation.

Measurement Error

Measurement Error in the Dependent Variable

- Consider the causal model $Y = \beta_0 + \beta_1 X + u$. Suppose that the dependent variable Y is measured with error such that only $Y^* = Y + e_y$ is observed.
- By substitution, $Y^* = \beta_0 + \beta_1 X + (u + e_y)$. Then, if $cov(X, e_y) = 0$, X is uncorrelated with the error term $u + e_y$ and β_1 coincides with the coefficient on X in the population linear regression model of Y on X and β_1 can be consistently estimated by OLS regression of Y on X.
- Recalling that the precision (under homoskedasticity) of the OLS estimator for β_1 is decreasing with increasing population variance of the residual (in this case $u + e_y$), supposing that u and e_y are uncorrelated, $\sigma^2_{u+e_y} = var(u) + var(e_y) > var(u) = \sigma^2_u$, hence measurement error in the dependent variable, even if uncorrelated with X, causes a decrease in the precision of the OLS estimator for β_1.

Measurement Error in an Independent Variable

- Suppose instead that the independent variable X is measured with error such that only $X^* = X + e_x$ is observed.
- By substitution, $Y = \beta_0 + \beta_1(X^* - e_x) + u = \beta_0 + \beta_1 X^* + (u - \beta_1 e_x) =: \beta_0 + \beta_1 X^* + \epsilon$. Suppose (favourably) that e_x is uncorrelated with Y, X, which implies also that e_x is uncorrelated with $u = Y - \beta_0 + \beta_1 X$. Even then, orthogonality fails in the above causal model because $cov(X^*, \epsilon) = cov(X + e_x, u - \beta_1 e_x) = -\beta_1 var(e_x) \neq 0$. Hence OLS regression of Y and X^* does not consistently estimate β_1.
- Let β_1^* denote the coefficient on X^* in a population linear regression of Y on X^*. Then,

  - $$\begin{aligned}\beta_1^* &= \frac{cov(Y, X^*)}{var(X^*)} \\ &= \frac{cov(\beta_0 + \beta_1 X^* + \epsilon, X^*)}{var(X^*)} \\ &= \frac{\beta_1 var(X^*) + cov(\epsilon, X^*)}{var(X^*)} \\ &= \frac{\beta_1 var(X^*) + cov(u - \beta_1 e_x, X + e_x)}{var(X^*)} \\ &= \frac{\beta_1 var(X^*) - \beta_1 var(e_x)}{var(X^*)} \\ &= \beta_1 \frac{var(X^*) - var(e_x)}{var(X^*)} \\ &= \beta_1 \frac{var(X)}{var(X) + var(e_x)}\end{aligned}.$$
  - OLS regression of Y on X^* yields a coefficient on X^* that is consistent for β_1^*.
- Measurement error in an independent variable imposes a negative bias on the OLS estimate of the coefficient on that independent variable. This bias is the least squares attenuation bias. The magnitude of attenuation is proportionate to the magnitude of the measurement error.
- If one independent variable is measured with error, the estimate of the coefficient on that variable is biased by the least squares attenuation bias, the estimates of the remaining coefficients will be biased in directions that depend on the directions of correlations between regressors.

Simultaneity

- Simultaneity occurs when one regressor in a model is dependent on the dependent variable. Then, this regressor and the dependent variable are jointly or simultaneously determined, and it is not the case that one is determined as a function of the other.
- The problem of simultaneity is illustrated in the context of estimation of demand and supply. Suppose that the causal model for demand is $Q^d = \beta_0 + \beta_1 P + \beta_2 X^d + u$ and the causal model for supply is $Q^s = \gamma_0 + \gamma_1 P + \gamma_2 X^s + v$, and Q^d, Q^s are causally related by the market clearing condition $Q^d = Q^s = Q$. Suppose that the causal effect of interest is the marginal effect of a change in price on quantity demanded, β_1.
  - X^d and X^s can be interpreted as some other non-price determinant of demand and supply respectively.
- The causal setup above implies $\beta_0 + \beta_1 P + \beta_2 X^d + u = \gamma_0 + \gamma_1 P + \gamma_2 X^s$, which in turn implies $P = \frac{\beta_0 - \gamma_0}{\gamma_1 - \beta_1} + \frac{\beta_2}{\gamma_1 - \beta_1} X^d + \frac{-\gamma^2}{\gamma_1 - \beta_1} X^s + \frac{u - v}{\gamma_1 - \beta_1}$, which in turn implies $cov(P, u) \neq 0$. Then, orthogonality fails in the causal model $Q = \beta_0 + \beta_1 P + \beta_2 X^d + u$, and the causal effect β_1 cannot be consistently estimated by OLS regression.
  - Economically, a change in some unobserved determinant of demand, collected in u causes a shift in the demand curve, which has some effect on price, hence orthogonality fails.

Randomised Control Trials

- A randomised control trial is a study in which a researcher is able to randomly assign treatment to participants in a way that is independent of participants' other characteristics. Then, treatment is uncorrelated with other unobserved determinants of outcome, and OLS regression of the outcome on treatment consistently estimates the causal effect. The inclusion of other regressors improves the precision of the estimate.

Binary Regressor

- Where treatment is binary, the OLS estimate of the causal effect of treatment coincides with the difference in the (sample) mean outcome between the treatment group and the control group.
- In the population,
  - $$\begin{aligned}\mathbb{E}[Y|D] &= \begin{cases}\mathbb{E}[Y|D=0] & \text{if } D = 0 \\ \mathbb{E}[Y|D=1] & \text{if } D = 1\end{cases} \\ &= \mathbb{E}[Y|D=0] + \{\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]\} \\ &=: \rho_0 + \rho_1 D\end{aligned}.$$
- Then, by definition of conditional expectation, $\rho_0 + \rho_1 D$ is the best predictor of Y given D, hence it is the best linear predictor of Y given D. ρ_0, ρ_1 solve the population linear regression problem, hence coincide with the coefficients of the population linear regression model.
- The sample analogue of this result is that the difference in the sample means between the treatment group and the control group coincides with the OLS regression coefficient on D.

Conditional Random Assignment

- Under some treatment protocols, treatment is conditionally randomly assigned. Suppose that treatment X is conditionally randomly assigned on $W_1, \ldots, W_k$.
- Consider the causal model $Y = \beta_0 + \beta_1 X + u$. Orthogonality does not necessarily hold in this causal model because X is correlated with $W_1, \ldots, W_k$, and in general, it is the case that $W_1, \ldots, W_k$ have a causal effect on Y.
- Let $X = \gamma_0 + \gamma_1 W_1 + \ldots + \gamma_k W_k + \tilde{X}$ denote the population linear regression model of X on $W_1, \ldots, W_k$. Then, by construction, $\tilde{X}$ is uncorrelated with each of $W_1, \ldots, W_k$. Given that X is uncorrelated with other (than $W_1, \ldots, W_k$) determinants of Y, $\tilde{X}$ is uncorrelated with every other (than X) determinant of Y.
- Consider a population linear regression of Y on X and $W_1, \ldots, W_k$. By the Frisch-Waugh-Lovell theorem, the coefficient on X in this regression is $\frac{cov(Y,\tilde{X})}{var(\tilde{X})}$.
  - $$\begin{aligned} cov(Y,\tilde{X}) &= cov(\beta_0 + \beta_1 X + u, \tilde{X}) \\ &= \beta_1 cov(X,\tilde{X}) + cov(u,\tilde{X}) \\ &= \beta_1 cov(\gamma_0 + \gamma_1 W_1 + \ldots + \gamma_k W_k + \tilde{X}, \tilde{X}) + cov(u,\tilde{X}), \\ &= \beta_1 cov(\tilde{X},\tilde{X}) + cov(u,\tilde{X}) \\ &= \beta_1 var(\tilde{X}) \end{aligned}$$
  - where the first equality follow by substitution of the causal model for Y, the second equality follows by linearity of covariance, the third equality follows by substitution of the auxiliary population regression for X, the fourth equality follows by linearity of covariance, and the fifth equality follows by X being uncorrelated with the unobserved determinants of Y and $\tilde{X}$ being the component of X uncorrelated with the observed determinants $W_1, \ldots, W_k$, hence being the component of X uncorrelated with any other determinants of Y.
  - Then, the Frisch-Waugh-Lovell coefficient on X in the population linear regression of Y on X and $W_1, \ldots, W_k$ coincides with β_1.
- Then, the OLS regression coefficient on X in the regression of Y on X and $W_1, \ldots, W_k$ is consistent for β_1.

Endogenous Controls

- Recall that the inclusion of controls as regressors improves the precision of the estimate of the causal effect of treatment. The problem of endogenous controls entails that only pre-treatment characteristics are appropriate controls.
- An endogenous control is a control that is causally determined by treatment. The practice of including an endogenous control is "over controlling".
- Suppose that the causal effect of interest is the causal effect of D on Y. Consider the following causal models:
  - $Y = \beta_0 + \beta_1 D + u$,
  - $Y' = \gamma_0 + \gamma_1 D + v$,
  - $Y = \delta_0 + \delta_1 D + \delta_2 Y' + \epsilon$.
- The interpretation of these causal models is as follows. The first causal model contains the causal effect of interest, the second causal model is the causal model for the endogenous control Y', which can be understood as an intermediate outcome. The third causal model treats Y' as a control, and can be understood as attempting to isolate the effect of D on Y holding Y' constant.
- By substitution of the second into the third,
  - $$\begin{aligned} Y &= \delta_0 + \delta_1 D + \delta_2(\gamma_0 + \gamma_1 D + v) + u \\ &= (\delta_0 + \delta_2\gamma_0) + (\delta_1 + \delta_2\gamma_1)D + (\epsilon + \delta_2 v) \end{aligned}.$$
  - This recovers the first causal model.
- Orthogonality does not plausibly hold in the third causal model because Y' is likely correlated with unobserved determinants of Y collected in ϵ.
  - For example, where D is a randomly assigned student-to-teacher ratio, Y is a student's end-of-year test result, and Y' is a student's mid-year test result, each of Y and Y' is determined by a student's academic ability. Hence, Y' is correlated with this unobserved determinant of Y.
  - Analytically, $cov(Y',\epsilon) = cov(\gamma_0 + \gamma_1 D + v, \epsilon) = cov(v,\epsilon) \neq 0$.
  - Then, δ_1 cannot be consistently estimated by OLS regression of Y on D, Y', and only β_1 can be consistently estimated by OLS regression of Y on D (supposing that D is successfully randomly assigned).
- Note that β_1 and δ_1 are different causal effects. The former is the causal effect of D on Y, the latter is the causal effect of D on Y, holding Y' constant.

Heterogenous Causal Effects

- Consider the causal model $Y_i = \beta_0 + \beta_{1i} X_i + u_i$, where β_{1i} is not constant across observations i. The causal effect of X on Y in such a model is heterogenous. Causal effects vary across observations because the causal effect is a function of observed variables (including X) or because the causal effect is a function of unobservable variables. The former case is the familiar case under which the causal effects can be consistently estimated by OLS regression of an appropriate model that is nonlinear in the variables.
- In the latter case, the OLS coefficient on X in the regression of Y on X, supposing that orthogonality holds in the causal model, and that the causal effect β_{1i} is independent of X_i, is consistent for the average causal effect.
  - Mean independence and X_i's independence of β_{1i} are entailed by successful random assignment of X.
  - $$\begin{aligned}\mathbb{E}[Y_i|X_i] &= \mathbb{E}[\beta_0 + \beta_{1i} X_i + u_i|X_i] \\ &= \beta_0 + X_i \mathbb{E}[\beta_{1i}|X_i] + \mathbb{E}[u_i|X_i]. \\ &= \beta_0 + (\mathbb{E}\beta_{1i}) X_i\end{aligned}$$
  - Then, by definition of conditional expectation, $\beta_0 + (\mathbb{E}\beta_{1i})X_i$ is the best predictor of Y_i given only X_i. Noting that $\beta_0 + (\mathbb{E}\beta_{1i})X_i$ is linear, it is the best linear predictor of Y_i given only X_i, then its coefficients coincide with those of the population linear regression of Y_i on X_i, and are consistently estimated by the OLS regression coefficients.
- The average causal effect is denoted as ACE. In the context of a binary treatment, the average causal effect is referred to as the average treatment effect ATE. Both are equal to $\mathbb{E}\beta_{1i}$.
- In the context of a binary treatment, the average effect of treatment on the treated is denoted as TOT, and is equal to $\mathbb{E}[\beta_{1i}|D_i = 1]$. Consistent estimation of TOT is non-straightforward and non-examinable.
- Whether ATE or TOT is the object of interest depends on the application.

Internal and External Validity

- A study is internally valid iff the estimates it yields are consistent for the causal effects of interest.
- The threats to external validity are heterogeneity between the study population and the population of policy interest, potential spillover effects (which are when the assumption of individualistic treatment response fails), and the failure of surrogate outcomes. A study is externally valid iff its findings can be credibly extrapolated to the population or policy of interest. If a study is externally valid, then the outcomes of the population or policy of interest are expected to be identical to those in the study.