ECON0019 T2 Lec 1: Potential Outcomes and Experiments

1 Overview and Motivation

Why running regression?

1. To predict an outcome Y from a series of covariates $(X_1, \ldots, X_n)$.
2. To estimate parameters of "structural" equations (e.g., price elasticity of demand in a demand equation).
3. To estimate causal effects of policies/treatments (e.g., the effect of a job-training program on wages).

2 Formalization of Causality

2.1 Potential Outcomes Framework

Assume a binary treatment status X, where $X = 1$ if treated and $X = 0$ if untreated.

- Y_1 = outcome *with* the intervention.
- Y_0 = outcome *without* the intervention.
- $Y_1 - Y_0$ = causal effect for a particular unit (which may vary across units).

Only one of $\{Y_1, Y_0\}$ is observed in reality, but both are defined conceptually. Observed outcome Y can be written via the *switching equation*:

$$Y = \begin{cases} Y_1, & \text{if } X = 1, \\ Y_0, & \text{if } X = 0 \end{cases} \quad \Longleftrightarrow \quad Y = XY_1 + (1 - X)Y_0.$$

2.2 Key Parameters of Interest

Average Treatment Effect (ATE):

$$\text{ATE} \;=\; \mathbb{E}[\,Y_1 - Y_0\,].$$

This is the average causal effect *across all individuals*. Note that:

- We do *not* directly observe $\{Y_1,\, Y_0\}$ for the same unit simultaneously.
- It generally differs from a simple comparison of means of those who receive treatment versus those who do not.

Average Treatment Effect on the Treated (ATT):

$$\text{ATT} \;=\; \mathbb{E}[\,Y_1 - Y_0 \;\mid X = 1].$$

This is the average causal effect specifically among those who *received* treatment.

3 What Does a Regression Estimate?

Consider a simple linear regression of Y on the binary variable X. The population OLS coefficient satisfies:

$$\beta_{\text{OLS}} \;=\; \mathbb{E}[\,Y \mid X = 1\,] \;-\; \mathbb{E}[\,Y \mid X = 0\,].$$

That difference is the limit of the sample OLS estimator under standard assumptions (SLR.2, SLR.3, etc.).

3.1 Proof (Binary X)

Recall:

$$\beta_{\text{OLS}} \;=\; \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad\longrightarrow\quad \frac{\text{Cov}(X,\, Y)}{\text{Var}(X)}.$$

For a binary X with $\mathbb{P}(X=1)=p$, we have $\mathbb{E}[X]=p$ and $\mathrm{Var}(X)=p(1-p)$. Then,

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathbb{E}[YX] - \mathbb{E}[Y]\mathbb{E}[X] \\
&= \mathbb{E}\{\mathbb{E}[YX \mid X]\} - \mathbb{E}\{\mathbb{E}[Y \mid X]\} \cdot \mathbb{E}[X] \\
&= p \cdot \mathbb{E}[YX \mid X=1] + (1-p)\underbrace{\mathbb{E}[YX \mid X=0]}_{=0} - p \cdot \mathbb{E}[Y \mid X=1] \cdot \mathbb{E}[X] \\
&\quad - (1-p)\mathbb{E}[X]\mathbb{E}[Y \mid X=0] \\
&= p\,\mathbb{E}[Y \mid X=1] - p^2\mathbb{E}[Y \mid X=1] - (1-p) \cdot p \cdot \mathbb{E}[Y \mid X=0] \\
&= p(1-p)\mathbb{E}[Y \mid X=1] - p(1-p)\mathbb{E}[Y \mid X=0] \\
&= p(1-p)\left[\mathbb{E}[Y \mid X=1] - \mathbb{E}[Y \mid X=0]\right]
\end{aligned}$$

Hence,

$$\beta_{\text{OLS}} \;=\; \frac{p(1-p)\left(\mathbb{E}[\,Y \mid X=1\,] - \mathbb{E}[\,Y \mid X=0\,]\right)}{p(1-p)} \;=\; \mathbb{E}[\,Y \mid X=1\,] - \mathbb{E}[\,Y \mid X=0\,].$$

Thus the OLS coefficient with a binary regressor is exactly the difference in conditional means.

3.2 Causation vs. Selection

When $X=1$, $Y=Y_1$; when $X=0$, $Y=Y_0$. Hence,

$$\beta_{\text{OLS}} \;=\; \mathbb{E}[\,Y \mid X=1\,] \;-\; \mathbb{E}[\,Y \mid X=0\,] \;=\; \mathbb{E}[\,Y_1 \mid X=1\,] \;-\; \mathbb{E}[\,Y_0 \mid X=0\,].$$

We can rewrite:

$$\beta_{\text{OLS}} \;=\; \underbrace{\left(\mathbb{E}[\,Y_1 \mid X=1\,] - \mathbb{E}[\,Y_0 \mid X=1\,]\right)}_{\text{ATT}} \;+\; \underbrace{\left(\mathbb{E}[\,Y_0 \mid X=1\,] - \mathbb{E}[\,Y_0 \mid X=0\,]\right)}_{\text{Selection Bias}},$$

highlighting that the difference in means can deviate from the true treatment effect by a *selection bias* term.

4 Experimental Protocol

4.1 Randomized Control Trials (RCTs)

Under RCTs, the assignment X is independent of $\{Y_0, Y_1\}$. Then:

$$\mathbb{E}[\,Y_0 \mid X = 1\,] = \mathbb{E}[\,Y_0 \mid X = 0\,], \quad \beta_{\text{OLS}} = \text{ATT}.$$

Moreover, ATT = ATE if $\mathbb{E}[\,Y_1 - Y_0 \mid X = 1\,] = \mathbb{E}[\,Y_1 - Y_0\,]$.

4.2 Examples

Job-Training Program:

- If treatment status is *randomly assigned*, then simple difference-in-means identifies the ATE.
- If individuals *self-select*, there may be nonzero selection bias, invalidating the direct difference-in-means as an ATE measure.

5 Control Variables After an Experiment

Consider a regression of y on x and another variable W.

Case I: W is a pre-treatment variable, and $X \perp (Y_0, Y_1, W)$.

- The regression still identifies $\mathbb{E}[Y_1 - Y_0]$ (the ATE).
- Proof by "partialling-out": in population, regressing X on W yields a slope of zero under independence, so residuals are simply $X - \mathbb{E}[X]$. Controlling for W can reduce variance if Y_0 correlates with W.

Case II: W is causally affected by X, so $X \not\perp (Y_0, Y_1, W)$.

- This is "overcontrolling" and no longer yields the (unconditional) ATE.

6 Linear Regression Connection

Rewriting the observed outcome in terms of potential outcomes:

$$Y_i \;=\; Y_{0,i} + (\,Y_{1,i} - Y_{0,i}\,)\,X_i \;=\; Y_{0,i} \;+\; \big(Y_{1,i} - Y_{0,i}\big)\,X_i.$$

In a linear form:

$$Y_i \;=\; \beta_0 + \beta_1 X_i \;+\; U_i,$$

where

$$\beta_0 \;=\; \mathbb{E}[Y_{0,i}], \quad U_i \;=\; Y_{0,i} \;-\; \mathbb{E}[Y_{0,i}], \quad \beta_1 \;=\; Y_{1,i} - Y_{0,i}.$$

If β_1 is constant across units and $\mathbb{E}[\,U_i \mid X_i\,] = 0$ (no selection bias), then a simple linear regression of Y on X identifies β_1.

7 Randomization Based on Covariates

Sometimes experiments are conducted by stratified sampling (random sampling within groups). Let W be the covariate defining the groups.

Selection Bias if We Ignore W: If the sampling probability for $X = 1$ depends on group membership (W)—and the baseline potential outcome Y_0 also relates to W—then $\mathbb{E}[Y_0 \mid X = 1] \neq \mathbb{E}[Y_0 \mid X = 0]$ in the pooled sample. One must condition on W to remove this bias.

(a) Population Average Causal Effect (ATE)

Partition a population into n groups. Then the population ATE is

$$\mathbb{E}[\,Y_1 - Y_0\,] \;=\; \sum_{k=1}^{n} \big(\mathbb{E}[\,Y_1 - Y_0 \mid W = k\,]\big)\; \Pr(W = k).$$

(b) Regression with Controls [Dummy Method]

One can include group dummies in a regression:

$$Y_i \;=\; \sum_{k=1}^{K} \beta_{0k}\, 1[\,W_i = k\,] \;+\; \beta_1 X_i \;+\; U_i.$$

The group-specific intercepts absorb all baseline group differences, allowing β_1 to capture treatment effects net of group-level confounding.

(c) Weights of OLS with Group Fixed Effects

Denote $\Pr(W = k) = p_k$ and $\Pr(X = 1 \mid W = k) = q(k)$. Then the OLS estimator for β_1 in the above fixed-effects regression is a weighted average of the group-specific causal effects

$\mathbb{E}[\,Y_1 - Y_0 \mid W = k\,]$:

$$\beta_1 \;=\; \frac{\displaystyle\sum_{k=1}^{K} p_k\, q(k)\,\big[\,1 - q(k)\big]\;\mathbb{E}\big[Y_1 - Y_0 \mid W = k\big]}{\displaystyle\sum_{k=1}^{K} p_k\, q(k)\,\big[\,1 - q(k)\big]}.$$

This need not equal the *population* ATE, which is

$$\mathbb{E}[\,Y_1 - Y_0\,] \;=\; \sum_{k=1}^{K} p_k\;\mathbb{E}[\,Y_1 - Y_0 \mid W = k\,].$$

Groups that take treatment frequently ($q(k) \to 1$) or rarely ($q(k) \to 0$) receive smaller OLS weight because of little within-group variation in X.

8 Internal and External Validity

Internal Validity: A statistical analysis is internally valid if its inference about causal effects is correct for the population actually studied.

External Validity: An analysis is externally valid if its results can be generalized to other populations or settings. External validity fails if population composition, treatment intensity, or large-scale spillovers differ in new contexts.

9. Pros and Cons of Experiments in Econometrics — Internal Validity

Benefits of Random Assignment

Under proper randomization, $X \perp (Y_0, Y_1)$ and thus:

- No selection bias: $\mathbb{E}[\,Y_0 \mid X = 1\,] = \mathbb{E}[\,Y_0 \mid X = 0\,]$.
- ATT = ATE if $\mathbb{E}[Y_1 - Y_0 \mid X = 1] = \mathbb{E}[Y_1 - Y_0]$.
- OLS difference-in-means is internally valid for that study sample.

Threats to Internal Validity

Threat #1: Randomization Failure E.g. poor design can make certain locations systematically assigned to treatment. One tests for randomization by regressing X on predetermined variables and checking whether the coefficients are jointly zero (via F-test).

Threat #2: Partial Compliance If the actual treatment taken deviates from assigned status, the difference between groups diminishes. One remedy, if compliance is observable, is instrumental variables.

Threat #3: Spillover Effect (SUTVA Violation) The Stable Unit Treatment Value Assumption requires each unit's outcome to depend only on its own treatment, not on others'. Violations can arise when one unit's treatment alters another's environment.

Threat #4: Attrition Some units may leave the sample in ways correlated with their potential outcomes, reintroducing selection bias.

5. Pros and Cons of Experiments — External Validity

RCT results are specific to a particular treatment and population. Scaling up can fail to replicate the same effects if:

- The population changes (small town vs. entire country).
- The treatment changes (e.g. quality control weakens).
- Large equilibrium spillovers occur when almost everyone is treated.

6. Why Don't Economists Stick with RCT?

- They can be expensive and logistically complicated.
- They raise ethical and political concerns (certain treatments cannot be randomly assigned).
- They might not capture realistic self-selection into programs.

ECON0019 T2 Lec2: Instrumental Variable and LATE

1. Endogeneity Issue

Consider the classical linear regression model:

$$y = \beta_0 + \beta_1 x + u \quad (\equiv \beta' X + u).$$

Ordinary Least Squares (OLS) is consistent when

$$\mathrm{Cov}(x, u) = 0$$

(SLR.4). If this assumption is violated, then x is said to be *endogenous*, and the OLS estimator becomes biased and inconsistent.

Example: Earning Returns to Education

$$\ln(\text{wage}) = \beta_0 + \beta_1 \,\text{educ} + u.$$

If $\mathrm{Cov}(u, \text{educ}) \neq 0$ (due, for instance, to unobserved ability), OLS is inconsistent.

a. Reasons for Endogeneity

- Omitted variables.
- Simultaneity (reverse causality), e.g. police and crime rates.
- Measurement error (errors-in-variables).

b. Unbiasedness $\neq$ Consistency

Unbiasedness is a small-sample concept, while consistency is an asymptotic concept (plim). For consistency, one needs:

$$\mathbb{E}[u \mid x] = 0 \quad \Longrightarrow \quad \mathrm{Cov}(x, u) = 0.$$

However, $\mathrm{Cov}(x, u) = 0$ does not necessarily imply $\mathbb{E}[u \mid x] = 0$. The proof uses the law of iterated expectations.

c. Summary

If $\mathrm{Cov}(x, u) \neq 0$, OLS is inconsistent. One needs an alternative estimation approach if exogeneity is violated.

2. Instrumental Variable (IV)

1) The Basic Idea

Instrumental Variable (IV) methods can help when endogeneity arises, for example due to omitted variable bias. An IV is a variable that:

- Correlates with x, but
- Is uncorrelated with the error u.

a. IV Assumptions

Consider:

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

where $\text{Cov}(u, x) \neq 0$. Suppose there exists an instrumental variable z_i satisfying:

$$\text{Cov}(z_i, u_i) = 0 \quad (\textbf{Exogeneity / Validity}),$$

$$\text{Cov}(z_i, x_i) \neq 0 \quad (\textbf{Relevance}).$$

b. Testing IV Relevance Assumption

To verify the relevance of z, one can regress x_i on z_i and test:

$$H_0 : \pi_1 = 0 \quad \text{vs.} \quad H_1 : \pi_1 \neq 0.$$

If $\pi_1 \neq 0$, the instrument is relevant.

3. IV Estimator

Still with

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

we note that:

$$\text{Cov}(z_i, y_i) = \text{Cov}(z_i, \beta_0 + \beta_1 x_i + u_i) = \beta_1 \,\text{Cov}(z_i, x_i),$$

given $\text{Cov}(z_i, u_i) = 0$. Rearranging,

$$\beta_1 = \frac{\text{Cov}(z_i, y_i)}{\text{Cov}(z_i, x_i)} = \frac{\frac{\text{cov}(z_i, y_i)}{\text{var}(z_i)}}{\frac{\text{cov}(z_i, x_i)}{\text{var}(z_i)}}.$$

In sample analogue form:

$$\widehat{\beta}_1^{\text{IV}} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}.$$

And

$$\widehat{\beta}_0^{\text{IV}} = \bar{y} - \widehat{\beta}_1^{\text{IV}}\,\bar{x}.$$

Under the usual IV assumptions plus random sampling, the IV estimator is consistent.

4. Special Case: Wald Estimator

Suppose $z_i \in \{0, 1\}$. Then

$$\mathbb{E}[y_i \mid z_i = 1] - \mathbb{E}[y_i \mid z_i = 0] = \beta_1\big(\mathbb{E}[x_i \mid z_i = 1] - \mathbb{E}[x_i \mid z_i = 0]\big).$$

Hence the Wald estimator is:

$$\widehat{\beta}_1 = \frac{\bar{y}_{z=1} - \bar{y}_{z=0}}{\bar{x}_{z=1} - \bar{x}_{z=0}}.$$

5. Inference Under IV Model

Under homoskedasticity, the asymptotic variance for the IV slope is

$$\mathrm{Var}(\widehat{\beta}_1^{\mathrm{IV}}) = \frac{\hat{\sigma}^2}{SST_x\, R_{x,z}^2},$$

where $R_{x,z}^2$ is the R^2 from regressing x on z. Consequently, $\mathrm{Var}(\widehat{\beta}_1^{\mathrm{IV}}) \geq \mathrm{Var}(\widehat{\beta}_1^{\mathrm{OLS}})$.

Advantages IV is consistent even if $\mathrm{Cov}(u, x) \neq 0$.

Disadvantages If $\mathrm{Cov}(u, x) = 0$, then OLS is more efficient than IV. IV can also be biased in finite samples.

6. Weak Instruments and Bias

A weak instrument is one for which $\mathrm{Cov}(x, z)$ is small. Then the denominator in the IV formula can be close to zero, making the estimator unstable. A rule of thumb is an F-statistic > 10 in the first stage to avoid weak instruments.

7. IV in the Multiple Linear Regression (MLR) Model

Consider

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i,$$

where x_{i1} is endogenous, and x_{i2} is exogenous. The instrument z_i must now satisfy

$$\mathrm{Cov}(z_i, u_i \mid x_{i2}) = 0, \quad \mathrm{Cov}(z_i, x_{i1}) \neq 0.$$

We can still define an IV estimator by the same logic, but must partial out x_{i2}.

Testing Instrument Relevance

Regress x_{i1} on z_i and x_{i2}, then test whether the instrument(s) explain a significant portion of the variation in x_{i1}.

8. Two-Stage Least Squares (2SLS)

2SLS (or TSLS) is a widely used procedure in the MLR framework when there may be multiple endogenous and exogenous variables as well as multiple instruments. For a single endogenous regressor, 2SLS coincides with IV, but operationally proceeds in two OLS regressions:

First Stage

Regress the endogenous regressor(s) on the instruments plus any other exogenous variables. For example, in the single-endogenous case:

$$x_{i1} = \pi_0 + \pi_1 z_{i1} + \cdots + \pi_M z_{iM} + \pi_{M+1} x_{i2} + \cdots + \nu_i.$$

Obtain the fitted values $\widehat{x}_{i1}$.

Second Stage

Regress y_i on $\widehat{x}_{i1}$ and any other exogenous regressors:

$$y_i = \beta_0 + \beta_1 \widehat{x}_{i1} + \beta_2 x_{i2} + \cdots + e_i.$$

The 2SLS slope estimates from this second stage match the IV estimates that treat z_i (and possibly others) as instruments for x_{i1}. 2SLS is consistent under the assumptions that the instruments are *valid* and *relevant*.

9. Testing for Endogeneity – Hausman Test

The *Hausman test* addresses whether an explanatory variable x is endogenous ($\text{Cov}(x, u) \neq 0$) or exogenous ($\text{Cov}(x, u) = 0$). The basic idea:

- Under $\text{Cov}(x, u) = 0$: both OLS and IV are consistent, and with large samples, their estimates should coincide.
- Under $\text{Cov}(x, u) \neq 0$: OLS is inconsistent, but IV is consistent. A large difference between the two estimates suggests endogeneity.

Practical Procedure (Auxiliary Regression Version)

1. **First stage**: Regress x on all exogenous variables and any instruments. Let $\hat{v}_i = x_i - \widehat{x}_i$ be the residual. This captures the portion of x not explained by exogenous variation and instruments.
2. **Auxiliary regression**:
$$y_i = \beta_0 + \beta_1 x_i + \theta\, \hat{v}_i + e_i.$$
3. **Test**: A simple t-test on θ. If $\theta \neq 0$ significantly, reject exogeneity of x. If θ is insignificant, one fails to reject exogeneity.

Alternatively, one can compare the OLS and IV estimates directly (another variant of the Hausman test); if their difference is statistically significant, that points to endogeneity.

10. Testing Overidentification Restrictions

Overidentification arises when there are more instruments than endogenous regressors, giving extra moment conditions that can be tested.

Sargan-Hansen or J-test Procedure (Homoskedastic Case)

1. Estimate coefficients by 2SLS/IV. Let $\hat{\beta}$ be the estimated parameters and collect the fitted residuals $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots$.
2. Regress $\hat{u}_i$ on *all* instruments (and exogenous regressors). Let R^2 be the *uncentered* (or appropriately handled) R^2 from that regression.[1]
3. Under the null hypothesis that all instruments are valid ($\mathrm{Cov}(z, u) = 0$), the statistic
$$n\, R^2$$
is asymptotically χ^2_{M-k}, where M is the total number of instruments and k is the number of endogenous regressors.
4. If $n\, R^2$ is large, reject the null that all instruments are valid (they appear to correlate with the residuals). If it is small, fail to reject and proceed under the assumption that the instruments are valid.

This overidentification test cannot definitively prove validity, but it can detect certain forms of instrument invalidity.

11. IV and LATE

Heterogeneous Treatment Effects If individuals differ in how they respond to a "treatment" regressor X, an IV approach using a binary or other instrument identifies the **Local Average Treatment Effect (LATE)**. That is, it recovers the effect for the subgroup of "compliers" whose treatment status X is actually influenced by the instrument Z.

Example: Job Training Program

Suppose 300 are invited (instrument $Z = 1$) and 300 are not (instrument $Z = 0$). Among those invited, some self-select to participate ($X = 1$), some do not. Because of such self-selection, OLS of wage on participation might be biased. If the invitation is random and only affects outcomes through actual participation, it is a valid instrument.

Wald Estimator in This Setting

$$\widehat{\beta}_1 = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[X \mid Z = 1] - \mathbb{E}[X \mid Z = 0]},$$

which is the LATE for those who comply with the invitation.

[1] In some expositions, the second stage residual is used differently; the general logic is to see how much the instruments can "explain" the final residual.

Monotonicity Assumption

Often it is assumed there are no *defiers* (the instrument always moves the treatment in the same direction for everyone), and that the instrument is as-if randomly assigned. Under these conditions, 2SLS recovers an average treatment effect specifically for the compliers.

12. IV Framework with Heterogeneous Effects and LATE

General Setup

$$Y_i = \beta_0 + \beta_{1i} X_i + U_i,$$
$$X_i = \pi_0 + \pi_{1i} Z_i + V_i,$$

allowing for the possibility that β_{1i} and π_{1i} vary across individuals. With suitable independence and monotonicity conditions, the IV estimate

$$\beta_{\text{IV}} = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, X)}$$

identifies a *weighted average of the individual effects* β_{1i}, known as the Local Average Treatment Effect (LATE).

When Does LATE = ATE? If either (i) all individuals have the same treatment effect or (ii) the instrument shifts treatment by the same amount for all individuals and that shift is unrelated to the magnitude of the effect, then the LATE coincides with the overall Average Treatment Effect. Generally, LATE may differ from the average effect for the entire population.

13. Proof: Rewriting the IV Estimator (Recap)

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})} = \beta_1 + \frac{\sum_{i=1}^n (z_i - \bar{z})\, u_i}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})}.$$

If $\text{Cov}(z, u) = 0$ and $\text{Cov}(z, x) \neq 0$ holds strongly enough, $\widehat{\beta}_1$ is consistent for β_1.

In Summary:

- Two-Stage Least Squares (2SLS) or IV methods address endogeneity by exploiting valid instruments.
- The Hausman test can check whether a suspected regressor is actually endogenous by comparing OLS and IV or by an auxiliary-regression approach.
- Overidentification tests (like the Sargan-Hansen J-test) can check whether extra instruments appear uncorrelated with the error term.
- In the presence of heterogeneous treatment effects, 2SLS recovers a Local Average Treatment Effect (LATE), focusing on those whose treatment status is influenced by the instrument.

ECON0019 T2 Lec3 - Simultaneous Equation Models

1 Simultaneity

Definition: Simultaneity occurs when one or more regressors is jointly determined with the dependent variable, typically through some equilibrium mechanism. In such cases, conventional Ordinary Least Squares (OLS) estimates are biased and inconsistent because one of the regressors is not exogenous.

1.1 Market Equilibrium Example

Consider a simple setting with two structural equations representing supply and demand:

$$q_S(p) = \alpha_1\, p + \nu_1, \quad \text{[Supply]}$$
$$q_D(p) = \alpha_2\, p + \nu_2, \quad \text{[Demand]}$$

where ν_1 and ν_2 are supply and demand "shifters," respectively. Suppose $\mathbb{E}[\nu_1] = \mathbb{E}[\nu_2] = 0$ and $\operatorname{cov}(\nu_1, \nu_2) = 0$. Also assume $\alpha_1 \neq \alpha_2$.

Only the equilibrium price and quantity are observed, where equilibrium satisfies

$$q = q_S = q_D \quad \Longrightarrow \quad p \;=\; \frac{\nu_1 - \nu_2}{\alpha_2 - \alpha_1} \quad \text{and} \quad q \;=\; \frac{\alpha_2\,\nu_1 - \alpha_1\,\nu_2}{\alpha_2 - \alpha_1}.$$

Because

$$\operatorname{cov}(p, \nu_1) \;=\; \frac{\operatorname{var}(\nu_1)}{\alpha_2 - \alpha_1} \;\neq\; 0 \quad \text{and} \quad \operatorname{cov}(p, \nu_2) \;=\; -\,\frac{\operatorname{var}(\nu_2)}{\alpha_2 - \alpha_1} \;\neq\; 0,$$

the variable p is endogenous in both the demand and the supply equations, illustrating the simultaneity problem.

1.1.1 Introducing a Unit Tax τ

If a tax τ is introduced into the demand equation, for instance by considering

$$q_D(p) = \alpha_2\,(p + \tau) + \nu_2,$$

the new equilibrium outcomes become

$$p \;=\; \frac{\nu_1 - \nu_2}{\alpha_2 - \alpha_1} \;-\; \frac{\alpha_2}{\alpha_2 - \alpha_1}\,\tau, \quad q \;=\; \frac{\alpha_2\,\nu_1 - \alpha_1\,\nu_2}{\alpha_2 - \alpha_1} \;-\; \frac{\alpha_1\,\alpha_2}{\alpha_2 - \alpha_1}\,\tau.$$

To analyze the impact of this policy, α_1 and α_2 must be identified and estimated in a way that avoids simultaneity bias.

2 Using an Instrument to Remove Simultaneous Equation Bias

Suppose there is a supply shifter (observed variable) z, which is exogenous with respect to the unobserved component in the supply function and with respect to the demand shifters. Concretely,

$$\nu_1 \;=\; u_1 + \beta_1\, z, \quad \mathrm{Cov}(z, u_1) = 0, \quad \mathrm{Cov}(z, \nu_2) = 0.$$

Then, the supply equation can be written as

$$q_S(p) = \alpha_1\, p + \beta_1\, z + u_1.$$

Equilibrium in this new setting gives

$$p \;=\; \frac{u_1 - \nu_2}{\alpha_2 - \alpha_1} + \frac{\beta_1}{\alpha_2 - \alpha_1}\, z, \quad q \;=\; \frac{\alpha_2\, u_1 - \alpha_1\, \nu_2}{\alpha_2 - \alpha_1} \;+\; \frac{\alpha_2\, \beta_1}{\alpha_2 - \alpha_1}\, z.$$

One obtains:

$$\mathrm{Cov}(p, z) \;=\; \frac{\beta_1}{\alpha_2 - \alpha_1}\, \mathrm{Var}(z), \quad \mathrm{Cov}(q, z) \;=\; \frac{\alpha_2\, \beta_1}{\alpha_2 - \alpha_1}\, \mathrm{Var}(z).$$

Note that

$$\mathrm{Cov}(q, z) \;=\; \alpha_2\, \mathrm{Cov}(p, z).$$

Hence,

$$\alpha_2 \;=\; \frac{\mathrm{Cov}(q, z)}{\mathrm{Cov}(p, z)}.$$

This naturally motivates an IV estimator:

$$\widehat{\alpha}_2 \;=\; \frac{\sum_{i=1}^{N} (z_i - \bar{z})(q_i - \bar{q})}{\sum_{i=1}^{N} (z_i - \bar{z})(p_i - \bar{p})}.$$

Exogenous variables that appear only in the supply function (excluded from the demand function) identify the demand equation. By the same principle, exogenous variables appearing only in the demand equation (excluded from the supply function) can identify the supply equation.

3 Two-Equation Simultaneous Equation Model (SEM)

3.1 Requirements

Each equation in the system is supposed to have a clear causal interpretation, often guided by economic theory.

3.2 SEM Model

A two-equation SEM takes the form:

$$y_1 = \alpha_1 y_2 + \beta_1^T z_1 + u_1, \quad \text{[Eq. 1]}$$
$$y_2 = \alpha_2 y_1 + \beta_2^T z_2 + u_2, \quad \text{[Eq. 2]}$$

where y_1 and y_2 are endogenous variables. The vectors z_1 and z_2 are exogenous variables satisfying $\text{Cov}(z_j, u_k) = 0$. The errors u_1 and u_2 are called structural shocks. Equations (1) and (2) are referred to as *structural equations*, and the parameters $\alpha_1, \alpha_2, \beta_1, \beta_2$ are called *structural parameters*.

3.3 Examples

- Demand and Supply
- Crime and Policing (Example 16.1 in Wooldridge)
- Peer Effects

3.4 Reduced Form Equation (for y_2)

Substitute Eq. (1) into Eq. (2):

$$y_2 = \alpha_2 \left(\alpha_1 y_2 + \beta_1^T z_1 + u_1\right) \; + \; \beta_2^T z_2 \; + \; u_2.$$

Thus,

$$(1 - \alpha_1 \alpha_2)\, y_2 = \alpha_2 \beta_1^T z_1 \; + \; \beta_2^T z_2 \; + \; u_2 \; + \; \alpha_2 u_1.$$

Providing $\alpha_1 \alpha_2 \neq 1$, one can write the reduced form:

$$y_2 = \pi_1^T z_1 \; + \; \pi_2^T z_2 \; + \; v_2,$$

where

$$\pi_1 = \frac{\alpha_2\,\beta_1}{(1-\alpha_1\,\alpha_2)}, \quad \pi_2 = \frac{\beta_2}{(1-\alpha_1\,\alpha_2)}, \quad v_2 = \frac{u_2+\alpha_2\,u_1}{(1-\alpha_1\,\alpha_2)}.$$

Parameters π_1 and π_2 are called *reduced form parameters*, and v_2 is the *reduced form error*.

3.5 Simultaneity Bias in OLS

If one tries to estimate Eq. (1),

$$y_1 = \alpha_1\,y_2 + \beta_1^T\,z_1 + u_1$$

by OLS, note that y_2 itself is a function of u_1 and u_2. Because

$$y_2 = \pi_1^T\,z_1 + \pi_2^T\,z_2 + v_2,$$

and v_2 depends on u_1 as

$$v_2 = \frac{u_2+\alpha_2\,u_1}{(1-\alpha_1\,\alpha_2)},$$

there is typically correlation between y_2 and u_1. Hence

$$\mathrm{Cov}(y_2, u_1) \neq 0,$$

leading to simultaneity bias in OLS estimation.

3.6 Estimation Using 2SLS

If there are exogenous variables excluded from an equation—i.e., they affect the other endogenous variable but not the dependent variable of interest—that equation can be identified and consistently estimated by **2SLS** (Two-Stage Least Squares). (A related, more efficient approach for systems of equations is **3SLS**, but that is beyond these notes.)

4 Exclusion Restriction and Identification

4.1 Exclusion Restriction (Necessary but Not Sufficient)

For 2SLS to estimate α_1, the first equation must exclude at least one exogenous variable that appears in the second equation. Symbolically, if z_2 contains an exogenous variable not in z_1, it can serve as an instrument to help identify α_1.

4.2 Rank Condition (Necessary and Sufficient)

The first equation in a two-equation SEM is identified if and only if the second equation contains at least one exogenous regressor (with non-zero coefficient) that is excluded from the first equation.

4.3 Order Condition (Necessary but Not Sufficient)

An equation satisfies the order condition if the number of excluded exogenous variables from that equation (E) is at least the number of endogenous regressors on the right-hand side (M). Specifically:

$$\begin{aligned} E > M: &\quad \text{Over-identified (use 2SLS)}, \\ E = M: &\quad \text{Just-identified (use 2SLS)}, \\ E < M: &\quad \text{Unidentified (cannot estimate)}. \end{aligned}$$

5 General SEM and 2SLS

5.1 Generalization

A two-equation SEM can be extended to multiple equations. For each equation, a similar logic applies: it can be identified if and only if there are sufficient excluded exogenous variables (that affect the other endogenous variables but not this particular equation).

5.1.1 Example of Over-, Just-, and Under-Identification

- If an equation has two endogenous regressors ($M = 2$) but three excluded exogenous variables ($E = 3$), it is over-identified.
- If another equation has one endogenous regressor ($M = 1$) and one excluded exogenous variable ($E = 1$), it is just-identified.
- If a third equation has one endogenous regressor ($M = 1$) and no excluded exogenous variables ($E = 0$), it is not identified.

5.2 Rank Condition

In addition to the order condition, one must also check that the excluded exogenous variables actually appear with non-zero coefficients in other parts of the system. If a variable is excluded from an equation but ends up having a zero coefficient elsewhere, the equation will fail the rank condition and thus remain unidentified.

6 General Methodology for 2SLS

Given the system:

$$y_1 = \alpha_1 y_2 + \beta_1^T z_1 + u_1, \quad \text{[Eq. 1]}$$
$$y_2 = \alpha_2 y_1 + \beta_2^T z_2 + u_2, \quad \text{[Eq. 2]},$$

with exogenous variables z_1 and z_2 (and possibly more in a bigger system), the steps to estimate each equation by 2SLS are:

1. **Check Identification:** There must be enough excluded exogenous variables in each equation relative to the endogenous regressors in that equation.

2. **First Stage:** Regress the endogenous regressor(s) on *all* exogenous variables in the entire system. For example, to estimate Eq. (1), regress y_2 on all exogenous variables (both z_1 and z_2). Obtain the fitted values $\hat{y}_2$.

3. **Second Stage:** Substitute $\hat{y}_2$ in place of y_2 in Eq. (1) and run the OLS regression:

$$y_1 = \alpha_1 \hat{y}_2 \; + \; \beta_1^T z_1 \; + \; \widehat{u}_1.$$

   The coefficient $\widehat{\alpha}_1$ from this second stage is the 2SLS estimator for α_1. The same procedure applies to Eq. (2) if one also wishes to estimate α_2.

Under standard assumptions of exogeneity and correct exclusion restrictions, 2SLS yields consistent estimates of the structural parameters.

End of Notes. These summaries cover all main points and formulas from the lecture on Simultaneous Equation Models. The crucial takeaways include the nature of simultaneity bias, the conceptual framework of structural versus reduced form equations, the necessity of identification, and the mechanics of 2SLS estimation.

ECON0019 T2 Lec4: Probit, Logit, Censoring, and Tobit Models

Ambrose W

April 11, 2025

Introduction to Limited Dependent Variable (LDV) Models

Definition. A limited dependent variable (LDV) model is one in which the dependent variable y has a restricted range. Typical reasons for restriction include:

- **Binary Response** (e.g. Probit, Logit).
- **Censored** (e.g. top-coding, duration models).
- **Tobit** (mixed continuous-discrete outcomes).
- **Truncated samples**.
- **Sample selection models**.

1. Binary Response: Logit and Probit

1.1. Latent Utility Interpretation

Consider an individual deciding whether to work ($y = 1$) or not ($y = 0$). Suppose there is a latent (unobserved) utility of working:

$$u(y; x, e_y) \;=\; \beta_y^\top x \;+\; e_y,$$

where x represents observed personal characteristics (e.g. age, experience, partner's income), and e_y is an unobserved taste shifter.

An individual chooses $y = 1$ if and only if

$$u(1; x, e_1) \;>\; u(0; x, e_0) \quad \Longleftrightarrow \quad \beta_1^\top x + e_1 \;>\; \beta_0^\top x + e_0.$$

Define

$$(\beta_1 - \beta_0)^\top x + (e_1 - e_0) \;\equiv\; \beta^\top x + e \;>\; 0.$$

This leads to the threshold-crossing formulation:

$$y \;=\; \begin{cases} 1, & \text{if } \beta^\top x + e > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Or more compactly,

$$y \;=\; 1\{\, \beta^\top x + e > 0\}.$$

The probability that $y = 1$ then depends on the CDF of e, which we need to manually specify.

1.2. Probit and Logit Specifications

Two common choices for the cumulative distribution function (CDF) of e are:

- **Probit Model:** $G(z) = \Phi(z)$, the standard normal CDF. That is,

$$\Phi(z) \;=\; \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\Big(-u^2/2\Big)\, du.$$

- **Logit Model:** $G(z) = \Lambda(z)$, the logistic CDF,

$$\Lambda(z) \;=\; \frac{\exp(z)}{1 + \exp(z)},$$

  with PDF $\lambda(z) = \Lambda(z)\,\Big[1 - \Lambda(z)\Big]$.

Hence,

$$\Pr(y = 1 \mid x) \;=\; G(\beta^\top x).$$

For the Probit model, $G(\beta^\top x) = \Phi(\beta^\top x)$. For the Logit model, $G(\beta^\top x) = \Lambda(\beta^\top x)$. However, since $G(\cdot)$ is a non-linear function, we CANNOT use OLS to estimate β.

1.3. Why Not Use a Linear Probability Model (LPM)?

A standard linear regression for a binary outcome can yield predicted probabilities outside $[0, 1]$. By contrast, Probit/Logit probabilities stay in $[0, 1]$ for all x.

2. Maximum Likelihood Principle and Estimation

Key idea: Select parameter values (e.g. β) that maximize the likelihood of observing the data.

MLE finds $\hat{\beta}_{MLE}$ by solving:

$$\hat{\beta}_{MLE} = \arg\max_{\beta} \mathcal{L}(\beta)$$

- `Stata` calculates the maximum likelihood using a numeric gradient-based method. Therefore, the solution obtained is a local maximum rather than a global maximum.
- **Note that the maximization of $\mathcal{L}$ (usually) does not have a closed-form (analytical) solution.**
  And the maximum likelihood figure calculated is not directly interpretable *(We don't really care about the max. likelihood number).*

2.1. Log-Likelihood for Binary Response (Logit/Probit)

For a single observation i, the probability mass function is:

$$f(y_i \mid x_i, \beta) \;=\; \begin{cases} G(\beta^\top x_i), & \text{if } y_i = 1, \\ 1 - G(\beta^\top x_i), & \text{if } y_i = 0, \end{cases}$$

or more compactly,

$$f(y_i \mid x_i, \beta) \;=\; \Big[G(\beta^\top x_i)\Big]^{y_i} \, \Big[1 - G(\beta^\top x_i)\Big]^{1-y_i}.$$

The log-likelihood for observation i is

$$\ell_i(\beta) \;=\; y_i \, \ln\Big[G(\beta^\top x_i)\Big] \;+\; (1 - y_i) \, \ln\Big[1 - G(\beta^\top x_i)\Big].$$

Summing over $i = 1, \ldots, n$ gives:

$$\mathcal{L}(\beta) \;=\; \sum_{i=1}^{n} \ell_i(\beta).$$

The MLE $\hat{\beta}_{\text{MLE}}$ is obtained by maximizing this sum. There is no simple closed-form solution in general, so numerical routines are typically used.

3. Properties of the MLE

Under regular conditions, the MLE has:

- **Consistency:** $\hat{\beta}_{\text{MLE}} \to \beta$ as $n \to \infty$.
- **Asymptotic Normality:**

  $$\sqrt{n}\,\Big(\hat{\beta}_{\text{MLE}} - \beta\Big) \;\xrightarrow{d}\; \mathcal{N}\Big(0,\, I^{-1}(\beta)\Big),$$

  where $I^{-1}(\beta)$ is the inverse Fisher information matrix.
- **Asymptotic Efficiency:** MLE achieves the smallest possible asymptotic variance among unbiased estimators.

4. The Trinity of Tests

A common hypothesis test is the exclusion restriction (e.g. $\beta_j = 0$ for some set of j). Three major test frameworks exist:

4.1. Lagrange Multiplier (LM / Score) Test

Only the restricted model is estimated. The gradient (score) of the likelihood is evaluated at the restricted estimates. If the score at those estimates is large, the null is rejected.

4.2. Wald Test

Only the unrestricted model is estimated. The test checks whether the unrestricted estimates of particular coefficients are sufficiently far from zero.

4.3. Likelihood Ratio (LR) Test

Both restricted and unrestricted models are estimated. The statistic is

$$LR \;=\; 2\Big(\,\mathcal{L}_{\text{ur}} \;-\; \mathcal{L}_{\text{r}}\Big),$$

where $\mathcal{L}_{\text{ur}}$ is Log-likelihood of the unrestricted model while $\mathcal{L}_{\text{r}}$ is the Log-likelihood of the restricted model.

Test statistics is asymptotically χ^2_q under the null, where q is the number of restrictions.

5. Marginal (Partial) Effects in Logit/Probit

In Probit/Logit, the coefficient β_j itself does *not* directly give the change in $\Pr(y = 1 \mid x)$ from increasing x_j, because the probability is

$$p(x) \;=\; G(\beta^\top x).$$

For continuous x_j, the partial effect is

$$\frac{\partial\, p(x)}{\partial\, x_j} \;=\; g(\beta^\top x)\,\beta_j,$$

where $g(\cdot) = G'(\cdot)$ is the PDF of the chosen CDF G. For the Probit model, $g(z) = \varphi(z)$, the standard normal PDF; for the Logit model, $g(z) = \Lambda(z)[1 - \Lambda(z)]$.

5.1. Discrete Regressor Case

If x_j is discrete (e.g. a dummy), the effect can be found by evaluating:

$$\Delta \;=\; G\Big(\beta^\top x_{\text{with } x_j=1}\Big) \;-\; G\Big(\beta^\top x_{\text{with } x_j=0}\Big).$$

5.1 Partial Effect at the Average (PEA)

» Estimating the marginal effect when the covariates are set to their sample mean
i.e., What is the effect at the average covariate value; Derivative of the average

$$\text{Estimate } \frac{\partial p(\mathbb{E}[\mathbf{x}])}{\partial x_j} \text{ by } g(\hat{\boldsymbol{\beta}}^\top \bar{\mathbf{x}})\hat{\beta}_j$$

i. Issues

- When the variable is binary, interpretation of the average becomes problematic.
- Ambiguity between averaging functions or taking functions of averages.

5.2 Average Partial Effect (APE)

» Captures the overall average impact of x_j on the probability of success across the entire sample
i.e., What is the average of the effect across the sample; Average of the derivative

$$\text{Estimate } \mathbb{E}\left[\frac{\partial p(\mathbf{x})}{\partial x_j}\right] \text{ by } \hat{\beta}_j \cdot \frac{1}{n}\sum_{i=1}^{n} g(\hat{\boldsymbol{\beta}}^\top \mathbf{x}_i)$$

6. Goodness of Fit for Binary Response Models

6.1. Percent Correctly Predicted

Predict $\hat{y}_i = 1$ if $\hat{p}_i = G(\hat{\beta}^\top x_i) \geq 0.5$, and 0 otherwise, then measure the fraction of correct predictions. A *confusion matrix* compares actual vs. predicted classes:

6.2. Pseudo R-Squared

Standard R^2 from OLS does not apply here. Alternatives include:

- **Efron's** R^2:

$$R^2_{\text{Efr}} \;=\; 1 \;-\; \frac{\sum_{i=1}^{n}\big(y_i - \hat{\pi}_i\big)^2}{\sum_{i=1}^{n}(\,y_i - \bar{y}\,)^2},$$

  where $\hat{\pi}_i$ is the predicted probability $\Pr(y = 1 \mid x_i)$.

- **McFadden's** R^2:

$$R^2_{\text{McF}} \;=\; 1 \;-\; \frac{\ln \mathcal{L}_{\text{ur}}}{\ln \mathcal{L}_0},$$

  where $\mathcal{L}_{\text{ur}}$ is the unrestricted log-likelihood and $\mathcal{L}_0$ is the log-likelihood of the model with only an intercept.

7. Censoring and the Censored Regression Model

Censoring occurs when the exact value of y is unobserved beyond a certain cut-off. For instance, top-coding at c: observed $y_i = c$ if the true $y_i^* > c$. One may model

$$y_i^* \;=\; \beta^\top x_i \;+\; u_i, \quad u_i \mid x_i, c_i \;\sim\; \mathcal{N}(0, \sigma^2),$$

but the observed outcome is

$$y_i \;=\; \min(c_i,\, y_i^*).$$

(Or a left-censoring variant with $y_i = \max(c_i,\, y_i^*)$.)

7.1. Marginal Effect

In censored models, β_j measures the effect on y_i^*, the latent variable. The actual observation y_i may not vary above (or below) the censoring boundary.

7.2. Likelihood for Censoring

For each observation i:

- Probability that $y_i = c_i$ (censored):

$$\Pr(y_i = c_i \mid x_i) \;=\; \Pr(y_i^* \geq c_i \mid x_i) \;=\; 1 - \Phi\Big(\frac{c_i - \beta^\top x_i}{\sigma}\Big).$$

- For uncensored $y_i < c_i$, the density is

$$f(y_i \mid x_i,\, y_i < c_i) \;=\; \frac{1}{\sigma}\,\varphi\Big(\frac{y_i - \beta^\top x_i}{\sigma}\Big).$$

The log-likelihood sums these components across censored and uncensored observations.

Likelihood function

$$\prod_{i=1}^{n}\left[\left(1-\Phi\left(\frac{c-\boldsymbol{\beta}^\top\mathbf{x}}{\sigma}\right)\right)^{\mathbb{I}[y_i=c_i]} \times \left(\frac{1}{\sigma}\phi\left(\frac{y-\boldsymbol{\beta}^\top\mathbf{x}}{\sigma}\right)\right)^{\mathbb{I}[y_i<c_i]}\right]$$

Log-Likelihood Function

$$\ell(\boldsymbol{\beta},\sigma) = \mathbb{I}[y_i = c_i]\ln\left(1-\Phi\left(\frac{c_i-\boldsymbol{\beta}^\top\mathbf{x}}{\sigma}\right)\right) + \mathbb{I}[y_i < c_i]\ln\left(\frac{1}{\sigma}\phi\left(\frac{y_i-\boldsymbol{\beta}^\top\mathbf{x}}{\sigma}\right)\right)$$

8. Tobit Model

Motivation: Suppose the observed y can be zero for some fraction of individuals (e.g. hours worked can be zero or positive). A *corner solution* arises: $y = 0$ with positive probability, and $y > 0$ continuously. The classical Tobit model sets

$$y_i^* \;=\; \beta^\top x_i + u_i, \quad u_i \sim \mathcal{N}(0, \sigma^2),$$

and

$$y_i \;=\; \max\Big(0,\, y_i^*\Big).$$

Hence $y_i = 0$ if $y_i^* < 0$, and $y_i = y_i^*$ if $y_i^* \geq 0$. The MLE for the Tobit uses a similar logic to the censoring case:

- Probability $y_i = 0$ is $\Pr(y_i^* < 0) = \Phi\Big(-\beta^\top x_i/\sigma\Big)$.
- For $y_i > 0$, the density is $\frac{1}{\sigma}\,\varphi\Big(\frac{y_i - \beta^\top x_i}{\sigma}\Big)$.

8.1. Log-Likelihood Function for Tobit

$$\ell_i(\beta, \sigma) \;=\; 1\{y_i = 0\}\,\ln\Big(1 - \Phi\Big(\tfrac{\beta^\top x_i}{\sigma}\Big)\Big) \;+\; 1\{y_i > 0\}\,\ln\Big(\tfrac{1}{\sigma}\,\varphi\Big(\tfrac{y_i - \beta^\top x_i}{\sigma}\Big)\Big).$$

Summing across i yields the full log-likelihood to be maximized.

8.2. Expected Value of y in Tobit

Although β_j measures the partial effect on y_i^*, the outcome of interest is y_i. The conditional expectation is:

$$\mathbb{E}[y \mid x] \;=\; \Phi\Big(\frac{\beta^\top x}{\sigma}\Big)\,\beta^\top x \;+\; \sigma\,\varphi\Big(\frac{\beta^\top x}{\sigma}\Big),$$

where

$$\varphi(z) \;=\; \frac{1}{\sqrt{2\pi}}\,\exp\Big(-z^2/2\Big), \quad \Phi(z) \;=\; \int_{-\infty}^{z} \varphi(u)\,du, \quad \lambda(z) \;=\; \frac{\varphi(z)}{\Phi(z)} \quad \text{(the Inverse Mills Ratio)}.$$

This follows from the law of iterated expectations, partitioning over $\{y = 0\}$ and $\{y > 0\}$.

8.3. Partial Effects

- **Conditional partial effect** (conditional on $y > 0$):

$$\frac{\partial\,\mathbb{E}[\,y \mid x,\, y > 0\,]}{\partial\, x_j} \;=\; \beta_j\Big[\,1 \;+\; \lambda'\Big(\tfrac{\beta^\top x}{\sigma}\Big)\Big],$$

  where additional derivatives of the inverse Mills ratio may appear.

- **Unconditional partial effect**:

$$\frac{\partial\,\mathbb{E}[\,y \mid x\,]}{\partial\, x_j} \;=\; \frac{\partial}{\partial x_j}\Big[\,\Phi\Big(\tfrac{\beta^\top x}{\sigma}\Big)\,\Big(\beta^\top x\Big) \;+\; \sigma\,\varphi\Big(\tfrac{\beta^\top x}{\sigma}\Big)\Big].$$

  In practice, partial effects can be evaluated at mean covariates (PEA) or averaged across the sample (APE).

8.4. Limitations of Tobit

Tobit imposes that x_j affects both $\Pr(y > 0)$ and the conditional mean of y (given $y > 0$) in proportion. This can be restrictive. For instance, age might increase the probability of owning life insurance while decreasing the amount purchased, conditional on having any. Such a pattern contradicts the Tobit assumption unless model extensions are introduced.

Conclusion

These notes summarized the Probit and Logit frameworks for binary response, the maximum likelihood estimation technique (including its properties and hypothesis testing procedures), as well as models involving censoring and the Tobit approach for corner-solution outcomes. Key formulas and derivations for partial effects, probability statements, and likelihood functions were presented, showing how each model addresses different data limitations or outcome distributions.

1 ECON0019 T2 Lec5: Truncation

Overview: Truncation arises when certain observations are not selected into the sample. For example, a subset of the population may be targeted due to cost or other considerations. This lecture examines how ordinary least squares (OLS) results are affected by such truncation and outlines conditions under which OLS remains consistent as well as scenarios (particularly selection based on the dependent variable) where alternative methods, such as the Heckman correction, become necessary.

2 Linear Regression Model and Sample Selection

2.1 Basic Model

$$y = \beta^T x + u, \quad \mathbb{E}[u \mid x] = 0.$$

Truncation implies that certain observations are *not* included in the sample. Let s_i be an indicator for whether observation i is observed:

$$s_i = 1 \text{ if observed}, \quad s_i = 0 \text{ if not observed}.$$

The new regression form is:

$$s_i y_i = s_i \beta^T x_i + s_i u_i.$$

2.2 OLS Estimation with Missing Observations

Before the missing data issue, one would have:

$$s_i y_i = s_i \beta^T x_i + s_i u_i, \quad i = 1, \ldots, n.$$

After the research assistant discards data (even if done randomly or without specific knowledge), in the simplified case of simple linear regression (SLR):

$$s_i y_i = s_i \alpha + s_i \beta \, x_i + u_i s_i.$$

OLS on this model essentially uses only the observations for which $s_i = 1$.

2.3 Conditions for Consistency under Truncation

Consistency requires:

$$\mathbb{E}[s\,u] = 0, \quad \mathbb{E}\Big[s\,x\,(s\,u)\Big] = \mathbb{E}[s\,x\,u] = 0.$$

In other words,

$$\mathrm{Cov}(s_i u_i,\ s_i x_i) = \mathbb{E}[(s_i u_i)(s_i x_i)] - \mathbb{E}[s_i u_i]\ \mathbb{E}[s_i x_i] = 0.$$

Hence unbiasedness under truncation requires:

$$\mathbb{E}[s\,u \mid s\,x] = 0.$$

3 Truncation Scenarios and Linearity Assumption

All conclusions about consistency and unbiasedness assume

$$\mathbb{E}[y \mid x] \text{ is linear in } x, \quad y = \beta^T x + u, \quad \mathbb{E}[u \mid x] = 0.$$

If $\mathbb{E}[y \mid x]$ is not linear, such as in Simpson's paradox, the OLS-based results do not hold.

3.1 Case I: Random Truncation Independent of (x, u)

If sample selection s is completely independent of (x, u), then

$$\mathbb{E}[s\, x\, u] = \mathbb{E}[s]\ \mathbb{E}[x\, u] = 0,$$

so OLS remains consistent and unbiased.

3.2 Case II: Truncation Based on x Only

Here, s_x depends on x only. Then

$$\mathbb{E}[u \mid s\, x] = \mathbb{E}\Big[\mathbb{E}[u \mid x]\Big] = 0.$$

Hence OLS is still consistent and unbiased. This does *not* constitute p-hacking because the selection is not tinkering with the model itself.

3.3 Case III: Combination of Randomness and x-Based Selection

This merges Case I and Case II. Truncation is partly random and partly based on x but remains independent of u. As a result,

$$\mathbb{E}[u \mid x, s] = \mathbb{E}[u \mid x] = 0.$$

OLS is again consistent and unbiased.

3.4 Case IV: Truncation Based on y

If $s = 1$ only when $y \leq c$, for some random or non-random cutoff c, then

$$s = 1 \quad \text{iff} \quad u \leq c - \beta \cdot x.$$

Since s depends on u, s and u can be correlated, producing

$$\mathbb{E}[s\, x\, u] \neq 0.$$

OLS thus becomes inconsistent and biased.

4 Truncation Based on the Dependent Variable (Case IV)

4.1 Model Statement

$$y = \beta^T x + u, \quad u \mid x, c \sim \mathcal{N}(0, \sigma^2).$$

An observation (x_i, y_i) is observed if and only if $y_i \leq c_i$.

4.2 Truncated Distribution

The probability density of y, given x and the cutoff c, is:

$$\text{PDF}(y \mid x = x_i, c = c_i) = \frac{f(y \mid x_i, \beta, \sigma)}{F(c_i \mid x_i, \beta, \sigma)}, \quad \text{if } y < c_i,$$

where $f(\cdot)$ is the normal PDF with mean $\beta \cdot x_i$ and variance σ^2, and $F(\cdot)$ is the corresponding CDF. Maximum likelihood estimation (MLE) can be employed here to perform truncated regression.

5 Incidental Truncation (Self-Selection Truncation)

Sometimes, the outcome y is unobserved for a subset of individuals, but the covariates and other variables *are* observed for the non-respondents. The model is:

$$y = \beta^T x + u, \quad \mathbb{E}[u \mid x, z] = 0,$$

$$s = 1\Big[\gamma^T z + v \geq 0\Big].$$

We observe y only when $s = 1$. Covariates x and z are always observed. Assume x is a sub-vector of z, and z is independent of (u, v). The task is to estimate β consistently.

5.1 Framework

In this self-selection model, partial effects are in $\mathbb{E}[y \mid x]$. However, under selection, the conditional expectation

$$\mathbb{E}[y \mid z, v] = \beta^T x + \mathbb{E}[u \mid v].$$

Suppose (u, v) are jointly normally distributed with mean 0, variance of v equal to 1, and covariance $\text{Cov}(u, v) = \rho$. Then

$$\mathbb{E}[u \mid v] = \rho\, v.$$

Hence:

$$\mathbb{E}[y \mid z, v] = \beta^T x + \rho\, v.$$

Though v is not observed, s is observed, and $s = 1$ if and only if $\gamma^T z + v \geq 0$. A well-known result is:

$$\mathbb{E}[y \mid z,\ s = 1] = \beta^T x + \rho\, \mathbb{E}[v \mid s = 1].$$

Given $s = 1 \iff v \geq -\gamma^T z$, the conditional expectation of v is:

$$\mathbb{E}[v \mid v \geq -\gamma^T z] = \frac{\phi(\gamma^T z)}{\Phi(\gamma^T z)} = \lambda\Big(\gamma^T z\Big),$$

where $\phi(\cdot)$ is the standard normal PDF and $\Phi(\cdot)$ the standard normal CDF. The ratio $\phi(\gamma^T z)/\Phi(\gamma^T z)$ is known as the *Inverse Mills' Ratio* $\lambda(\cdot)$. Thus:

$$\mathbb{E}[y \mid z,\ s = 1] = \beta^T x + \rho\, \lambda\Big(\gamma^T z\Big).$$

5.2 Estimating the Parameters

Case I: $\rho = 0$. If $\rho = 0$, then u and v are independent. Selection depends only on x and randomness separate from u. This is analogous to Case III before, and OLS on the truncated sample is consistent.

Case II: $\rho \neq 0$. If $\rho \neq 0$, then OLS is biased because $\lambda(\gamma^T z)$ is an omitted variable. One can recover consistency by estimating $\lambda(\gamma^T z)$ from a *Probit* model:

$$\Pr(s = 1 \mid z) = \Phi(\gamma^T z).$$

This yields an estimate $\hat{\gamma}$, which allows one to compute

$$\widehat{\lambda_i} = \lambda\Big(\hat{\gamma}^T z_i\Big)$$

for each observation in the truncated subsample. The second stage then includes $\widehat{\lambda_i}$ as an additional regressor:

$$y_i = \beta^T x_i + \rho\, \widehat{\lambda_i} + \text{error}.$$

This two-step procedure is known as the *Heckman Correction* (Heckit). It provides a consistent estimate of β under joint normality. In practice, software procedures such as the "heckman" command in Stata automatically adjust the standard errors to account for the first-step estimation.

Caveats:

- Standard errors must be corrected for the first-step Probit estimation.
- If $x = z$, collinearity with $\lambda(\gamma^T z)$ can be problematic. Including extra variables in z that are excluded from x helps mitigate multicollinearity and large variance in the second step.
- If z is high-dimensional relative to x, the R^2 in the selection equation is not large, reducing the variance of the estimated ρ.

6 Summary

- **Truncation** means certain observations are not selected into the sample. It may be random or systematically based on x or y.
- **Consistency conditions**: OLS is unbiased if selection is independent of the error term u. Random or x-based selection (and combinations thereof) typically preserve OLS consistency. Selection based on y introduces bias.
- **Truncated regression**: When y is censored by a threshold (e.g. $y \leq c$), MLE methods can be employed, using the truncated normal PDF and CDF.
- **Incidental truncation and self-selection**: If the outcome is unobserved for some, but x, z are observed, it is a "Heckman selection model." Under joint normality of (u, v), the *Heckman two-step* (Probit followed by OLS with the Inverse Mills' Ratio) yields consistent estimates.

ECON0019 T2 Lec 6: Time Series I

Ambrose W

Introduction and Technical Challenges

These notes focus on basic concepts of time series analysis and how ordinary least squares (OLS) methods adapt when the data are observed over time rather than drawn as independent samples.

- **Repeated Observation of the Same Sample Over Time:** In time series, the same observational unit is recorded repeatedly, which differs from independent and identically distributed (i.i.d.) cross-sectional data.
- **Correlation Among Observations:** Consecutive observations of a single process tend to be correlated. This violates certain standard OLS assumptions (which require independence).
- **Lagged Variables as Regressors:** Past values of variables (*lagged* variables) are often used to predict the present or future observations.
- **Serially Correlated Errors:** The error terms may exhibit correlation across time, influencing how standard errors are computed.
- **Short Samples:** Macroeconomic data, for example, may have only a few dozen annual observations, posing challenges to asymptotic properties.

Notations and Differences

- For a time series $\{y_t\}_{t=1}^{T}$, y_t denotes the value of the variable of interest in period t.
- The first lag of y_t is y_{t-1}, the j-th lag is y_{t-j}.
- The first difference captures change from period $t-1$ to t:

$$\Delta y_t = y_t - y_{t-1}.$$

- The first difference of the natural log is

$$\Delta \ln y_t \;=\; \ln y_t \;-\; \ln(y_{t-1}).$$

- A key approximation states that:

$$100\,\Delta \ln(y_t)$$

  is approximately the percentage change of y_t from period $t-1$ to t. This follows from a Taylor approximation for small relative changes.

Percentage Change vs. Percentage Point Change:

- Moving from 1 to 2 is a "50%" increase if interpreting percentage change.
- Moving from 1 to 2 is a "1 percentage point" change if looking at a percent scale shifting by exactly 1 point.

Auto-covariance and Auto-correlation

Definitions

- The **autocorrelation** (aka. serial correlation) of a time series is the correlation between a series and its own lagged values.
- The first autocovariance of y_t is

$$\text{Cov}(y_t,\, y_{t-1}) \;=\; \mathbb{E}\Big[\big(\, y_{t-1} - \mu_{t-1}\,\big)\,(y_t - \mu_t)\Big].$$

  Typically, when the process is stationary, $\mu_{t-1} = \mu_t = \mu$.

- The first autocorrelation is

$$\rho_1 \;=\; \frac{\text{Cov}(y_t,\, y_{t-1})}{\sqrt{\text{Var}(y_t)\,\text{Var}(y_{t-1})}}.$$

Sample Autocovariance and Autocorrelation

The sample analog of autocovariance for lag j is

$$\widehat{\text{Cov}}(y_t,\, y_{t-j}) = \;\frac{1}{T}\sum_{t=j+1}^{T}\Big(y_t - \overline{y}_{j+1,T}\Big)\,\Big(y_{t-j} - \overline{y}_{1,T-j}\Big),$$

where $\overline{y}_{t,j}$ is the sample average of y over the observations that line up with lag j. The sample autocorrelation is then

$$\hat{\rho}_j \;=\; \frac{\widehat{\text{Cov}}(y_t,\, y_{t-j})}{\widehat{\text{Var}}(y_t)}.$$

OLS Assumptions in Time Series

Recall the Multiple Linear Regression (MLR) assumptions for cross-sectional data:

MLR.1 Linearity in parameters:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u.$$

MLR.2 Random Sampling (i.i.d.).

MLR.3 No perfect collinearity.

MLR.4 Exogeneity: $\mathbb{E}[u \mid x_1, \ldots, x_k] = 0$.

In time series:

- We often define $X_t = (x_{1t}, x_{2t}, \ldots, x_{kt})'$ and allow for potentially lagged terms in the regression.
- We retain **TS.1** (linearity in parameters) and **TS.3** (no perfect collinearity), but **TS.2** changes the exogeneity assumption to a *strict exogeneity* form:

  $$\mathbb{E}[u_t \mid \ldots, X_{t-1}, X_t, X_{t+1}, \ldots] = 0.$$

  This means u_t is orthogonal to regressors in *all* time periods. It is often violated when there is feedback from y into future X.

Stationarity (Time-Series Counterpart of i.i.d.)

Strict Stationarity

A time series $\{y_t\}$ is *strictly stationary* if its **probability distribution does not change over time**. Equivalently, the distribution of $(y_t, y_{t+1}, \ldots, y_{t+m})$ is the same as $(y_s, y_{s+1}, \ldots, y_{s+m})$ for any shifts in time index.

Second-order (Weak) Stationarity

A weaker condition requiring:

- $\mathbb{E}[y_t]$ is constant over time.
- $\mathrm{Var}(y_t)$ is constant over time.
- $\mathrm{Cov}(y_t, y_{t-j})$ depends only on j, not on t.

If these conditions hold (and $\mathbb{E}[y_t^2] < \infty$), the series is said to be *covariance stationary* or *weakly stationary*.

Weak Dependence and Ergodicity

- **Weak Dependence:** Observations spaced far enough apart in time are approximately independent. This is weaker than assuming no correlation at all.
- **Ergodicity:** A stationary process is ergodic if long-run time averages of a function of $\{x_t\}$ converge almost surely to the expected value of that function.

$$\frac{1}{T}\sum_{t=1}^{T} g\big(x_t\big) \quad \xrightarrow[T\to\infty]{a.s.} \quad \mathbb{E}[g(x_t)].$$

  This property helps guarantee that sample moments converge to population moments.

Time Series Assumptions (TS.1 to TS.5)

- **TS.1** (Linearity in parameters):

$$y_t \;=\; \beta_0 \;+\; \beta_1 x_{1t} \;+\; \cdots + \;\beta_k x_{kt} \;+\; u_t.$$

- **TS.2** (Strict exogeneity):

$$\mathbb{E}[u_t \mid \ldots, X_{t-1}, X_t, X_{t+1}, \ldots] = 0.$$

- **TS.3** (No perfect collinearity).
- **TS.4** (Stationarity — e.g. covariance stationarity).
- **TS.5** (Weak dependence / Ergodicity).

Implications:

- Under TS.1, TS.2, and TS.3, the OLS estimator is *unbiased*.
- Under TS.1, TS.2' (a slightly weaker exogeneity), TS.3, TS.4, and TS.5, the OLS estimator is *consistent* and *asymptotically normal*.

Unbiasedness of OLS in Time Series

Strict exogeneity must hold for unbiasedness:

$$\mathbb{E}[u_t \mid \ldots, X_{t-1}, X_t, X_{t+1}, \ldots] = 0.$$

If future regressors X_{t+1} or so are affected by u_t, then TS.2 fails. For instance, in dynamic models or feedback systems (x_t is a function of past errors), u_t can be correlated with regressors in future periods, destroying strict exogeneity.

Matrix Notation (Outline)

Let each observation vector be

$$\tilde{x}_t \;=\; \begin{pmatrix} 1 \\ X_t \\ X_{t-1} \\ \vdots \\ X_{t-p} \end{pmatrix}, \quad X \;=\; \begin{pmatrix} \tilde{x}_1' \\ \tilde{x}_2' \\ \vdots \\ \tilde{x}_T' \end{pmatrix}.$$

Then in matrix form, the OLS estimate can be written:

$$\hat{\beta} \;=\; (X'X)^{-1}X'y.$$

One can show

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

Consistency of OLS

Under TS.1, TS.2' (mean independence with past and current regressors), TS.3, TS.4, and TS.5, OLS is consistent. In words:

$$\hat{\beta} \;\xrightarrow[T\to\infty]{}\; \beta.$$

Sketch of Proof:

$$\hat{\beta} = \underbrace{\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{x}_t\tilde{x}_t'\right)^{-1}}_{\to Q^{-1}} \underbrace{\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{x}_t\, y_t\right)}_{\to Q\beta + 0},$$

where $Q = \mathbb{E}[\tilde{x}_t\tilde{x}_t']$ is assumed invertible, and $\sum \tilde{x}_t u_t/T \to 0$ by ergodic law of large numbers, provided $\mathbb{E}[\tilde{x}_t\, u_t] = 0$. Thus $\hat{\beta} \to Q^{-1}(Q\beta) = \beta$.

Asymptotic Normality

CLT for Ergodic and Stationary Sequences

If $\{Z_t\}$ is ergodic and stationary, then

$$\frac{1}{\sqrt{T}}\;\sum_{t=1}^{T} Z_t \;\;\xrightarrow{d}\;\; \mathcal{N}\Big(0,\; V\Big),$$

where

$$V = \;\lim_{T\to\infty}\frac{1}{T}\,\mathbb{E}\Big[\Big(\sum_{t=1}^{T}Z_t\Big)\,\Big(\sum_{t=1}^{T}Z_t\Big)'\Big].$$

OLS Asymptotic Normality

Combining the previous result with Slutsky's theorem, under stationarity and appropriate moment conditions,

$$\sqrt{T}\,\Big(\,\hat{\beta}\;-\;\beta\Big)\;\;\xrightarrow{d}\;\;\mathcal{N}\Big(\,0,\;Q^{-1}\,\Omega\,Q^{-1}\Big),$$

where

$$Q=\;\mathbb{E}\Big[\tilde{x}_t\,\tilde{x}_t'\Big]\quad\text{and}\quad\Omega=\;\lim_{T\to\infty}\frac{1}{T}\,\mathbb{E}\Big[\Big(\sum_{t=1}^{T}\tilde{x}_t\,u_t\Big)\,\Big(\sum_{t=1}^{T}\tilde{x}_t\,u_t\Big)'\Big].$$

Summary Remarks

- Time series regression requires a shift from the i.i.d. assumption to stationarity and weak dependence.
- Strict exogeneity is often difficult to fulfill if X_t depends on past error terms or if there is feedback from y_t to future X_{t+1}.
- When assumptions **TS.1**–**TS.5** hold, OLS estimators are consistent and asymptotically normal. If **TS.2** (strict exogeneity) additionally holds, they are unbiased in finite samples.
- Stationarity and ergodicity make it possible to treat time averages like sample averages, enabling large-sample arguments (law of large numbers and central limit theorem) to hold for time series data.

[INSERT IMAGE HERE: Any final relevant figure about CLT or the geometry of time series OLS if provided]

Note: Each formula above is presented as found, without rearrangement or further simplification. For exam revision, remember to check carefully:

- The definitions of strict and weak stationarity.
- The difference between Δy_t and $\Delta \ln(y_t)$.
- The idea that $\Delta \ln(y_t) \times 100$ is an approximate percent change.
- The role of **strict exogeneity** vs. **weaker exogeneity** in time series.
- The logic of consistency and asymptotic normality using ergodicity and stationarity assumptions.

ECON0019 T2 Lec 7 - Time Series Regression

Ambrose W

1 First-Order Autoregressive Model (AR(1))

Model Setup:

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t, \quad u_t \overset{i.i.d.}{\sim} (0, \sigma_u^2).$$

Key Points:

- Recent data (y_{t-1}) serves as a good predictor of y_t.
- Model can be estimated using OLS, but is biased in time-series contexts because strict exogeneity is violated.
- An AR(p) generalizes this to p lags:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + u_t.$$

1.1 Mean of AR(1)

Starting from

$$\mathbb{E}[y_t] \;=\; \beta_0 \;+\; \beta_1\, \mathbb{E}[y_{t-1}] \;+\; \mathbb{E}[u_t],$$

and assuming $\mathbb{E}[u_t] = 0$, one gets a recursive relationship:

$$\mathbb{E}[y_t] \;=\; \beta_0 + \beta_1\, \mathbb{E}[y_{t-1}],$$

which unfolds as an infinite series if $|\beta_1| > 1$, leading to divergence. For stationarity,

$$|\beta_1| < 1.$$

Under stationarity, $\mathbb{E}[y_t] = \mathbb{E}[y_{t-1}]$, so

$$(1 - \beta_1)\, \mathbb{E}[y_t] \;=\; \beta_0 \quad \Longrightarrow \quad \mathbb{E}[y_t] \;=\; \frac{\beta_0}{1 - \beta_1}.$$

1.2 Variance of AR(1)

Using

$$\mathrm{Var}(y_t) \;=\; \mathrm{Var}(\beta_0 + \beta_1 y_{t-1} + u_t) \;=\; \beta_1^2\,\mathrm{Var}(y_{t-1}) \;+\; \sigma_u^2 \;+\; 2\,\beta_1\,\mathrm{Cov}(u_t, y_{t-1}),$$

and assuming exogeneity so $\mathrm{Cov}(u_t, y_{t-1}) = 0$, one obtains

$$\mathrm{Var}(y_t) \;=\; \beta_1^2\,\mathrm{Var}(y_{t-1}) \;+\; \sigma_u^2.$$

Under stationarity, $\mathrm{Var}(y_t) = \mathrm{Var}(y_{t-1})$, so

$$\mathrm{Var}(y_t) \;=\; \frac{\sigma_u^2}{1-\beta_1^2}.$$

1.3 First-Order Auto-Covariance

$$\mathrm{Cov}(y_t, y_{t-1}) \;=\; \mathrm{Cov}(\beta_0 + \beta_1 y_{t-1} + u_t,\; y_{t-1}) \;=\; \beta_1\,\mathrm{Var}(y_{t-1}) \;=\; \frac{\beta_1\,\sigma_u^2}{1-\beta_1^2}.$$

1.4 First-Order Auto-Correlation

$$\rho_1 \;=\; \frac{\mathrm{Cov}(y_t, y_{t-1})}{\mathrm{Var}(y_t)} \;=\; \frac{\beta_1\,\sigma_u^2/(1-\beta_1^2)}{\sigma_u^2/(1-\beta_1^2)} \;=\; \beta_1.$$

2 First-Order Moving Average Model (MA(1))

Definition:

$$y_t \;=\; \theta_0 + u_t + \theta_1\,u_{t-1}, \quad u_t \overset{i.i.d.}{\sim} (0, \sigma_u^2).$$

2.1 Generalization: MA(q)

$$y_t \;=\; \theta_0 \;+\; u_t \;+\; \theta_1 u_{t-1} \;+\; \cdots + \;\theta_q u_{t-q}.$$

2.2 Comparison to AR Models

- **AR Model:** Persistence is through y_{t-1} itself; theoretically an infinite horizon effect of shocks.
- **MA Model:** Persistence stems from the error term. Shock influence decays and vanishes after q periods.

3 Autoregressive Moving Average Model (ARMA(p, q))

$$y_t \;=\; \mu \;+\; \sum_{i=1}^{p} \beta_i\,y_{t-i} \;+\; u_t \;+\; \sum_{j=1}^{q} \theta_j\,u_{t-j}, \quad u_t \overset{i.i.d.}{\sim} (0, \sigma_u^2).$$

4 Distributed Lag Models (DL)

Basic Distributed Lag (DL):

$$y_t \;=\; \alpha_0 \;+\; \alpha_1 x_{t-1} \;+\; \alpha_2 x_{t-2} \;+\; \cdots + \;\alpha_k x_{t-k} \;+\; u_t.$$

4.1 Autoregressive Distributed Lag (ADL(p, q))

$$y_t \;=\; \mu \;+\; \sum_{I=1}^{p} \beta_I \, y_{t-I} \;+\; \sum_{j=1}^{k} \alpha_j \, x_{t-j} \;+\; u_t.$$

OLS is generally biased here because the presence of y_{t-1} on the right side violates strict exogeneity in time-series contexts.

5 Choosing the Number of Lags

5.1 Rules of Thumb

- **Monthly data:** use 6 or 12 lags.
- **Quarterly data:** use 4 lags.

5.2 Bayesian Information Criterion (BIC)

$$\mathrm{BIC}(n) \;=\; \ln\!\Big(\frac{\mathrm{SSR}(n)}{T}\Big) \;+\; \frac{n\,\ln(T)}{T},$$

where

$$n = p + k + 1,$$

and p is the number of autoregressive coefficients, k is the number of distributed-lag coefficients, plus 1 for the constant term. T is the sample size. BIC is consistent in selecting the true model as $T \to \infty$.

5.3 Akaike Information Criterion (AIC)

$$\mathrm{AIC}(p) \;=\; \ln\!\Big(\frac{\mathrm{SSR}(n)}{T}\Big) \;+\; \frac{2\,n}{T}.$$

AIC generally chooses more lags than BIC in finite samples because the penalty term is smaller.

6 Violations of Stationarity

6.1 Seasonality

Regular, predictable changes that recur every calendar year, possibly altering mean or variance over time.

Possible Solutions:

- **Fixed Effects**: include month (or season) dummies for known recurring events.
- **Trigonometric Controls**: add sine and cosine terms to capture seasonal frequencies.
- **Annual Percentage Change**: transform the data if appropriate.

6.2 Deterministic Trends (Time Trends)

A persistent, long-term movement in the data. Sometimes modeled as:

$$y_t = \delta_1 t + \dots \quad \text{or} \quad y_t = \delta_1 t + \delta_2 t^2 + \dots$$

These do not reflect random drift but a deterministic function of time t.

7 Stochastic Trends (Unit Roots)

7.1 Random Walk Model

$$y_t = y_{t-1} + u_t, \quad u_t \text{ serially uncorrelated,}$$

implies non-stationarity if the coefficient on y_{t-1} is 1. The variance grows linearly over time.

7.2 Unit Roots

When an AR(1) has $\beta_1 = 1$, it is said to have a *unit root*. This implies excessive persistence and non-stationarity. OLS estimates are strongly biased toward zero, and standard t-tests are invalid.

7.3 Proof Sketch: Bias of OLS under Unit Root

Consider

$$y_t = y_{t-1} + u_t,$$

where u_t is serially uncorrelated. Because a positive shock u_t raises y_t but not y_{t-1} (already fixed), a negative correlation emerges between y_{t-1} and u_t. This causes a downward (toward zero) bias in the OLS estimator.

7.4 Testing for Stochastic Trends: Dickey-Fuller Test

Given

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t,$$

we want to test:

$$H_0 : \beta_1 = 1 \quad \text{vs.} \quad H_1 : \beta_1 < 1.$$

Rewriting in first differences:

$$\Delta y_t \;=\; \beta_0 \;+\; (\beta_1 - 1)\, y_{t-1} \;+\; u_t,$$

define $\delta = \beta_1 - 1$. Then test

$$H_0 : \delta = 0 \quad \text{vs.} \quad H_1 : \delta < 0.$$

Dickey-Fuller distributions are used instead of the standard normal distribution.

7.5 Solutions for Trends

- **Attention is All You Need**: Visual inspection of whether there is a persistent movement over time.
- **First-Differencing**: if y_t follows a random walk or contains a unit root, then Δy_t is often stationary. But avoid differencing if it is not truly needed.

8 Heteroskedasticity

8.1 Unconditional Heteroskedasticity

Overall variance changes over time in a predictable way, violating stationarity.

8.2 Conditional Heteroskedasticity

Variance of the error term varies with past information but does *not* violate stationarity. Standard OLS inference can be invalidated by time-varying conditional variances, though the OLS estimator remains unbiased and consistent.

9 Structural Breaks

9.1 Definition

Regression coefficients may change abruptly at a certain time $t = \tau$. For instance, in an ADL(1,1):

$$y_t = \beta_0 + \beta_1 y_{t-1} + \delta_1 x_{t-1} + u_t.$$

A known break date can be tested via a Chow Test by allowing a dummy for post-break and interacting that dummy with regressors.

9.2 Chow Test (Known Break Date)

Include

$$D_t(\tau) = \begin{cases} 1, & t > \tau, \\ 0, & \text{otherwise}, \end{cases}$$

and run the augmented regression:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \delta_1 x_{t-1} + \gamma_0 D_t(\tau) + \gamma_1 \big[D_t(\tau) \times y_{t-1}\big] + \gamma_2 \big[D_t(\tau) \times x_{t-1}\big] + u_t.$$

Test $H_0 : \gamma_0 = \gamma_1 = \gamma_2 = 0$ via an F-test.

9.3 Quandt Likelihood Ratio (QLR) Test (Unknown Break Date)

The QLR test is an extension of the Chow test designed to detect unknown structural breaks. Instead of assuming a known break date, it systematically examines a range of potential break points and evaluates the corresponding Chow test statistics.

QLR Test Procedure:

1. Choose trimming values to avoid small sample sizes on either side of the break. Typically,

   $$\tau_0 = 0.15T, \quad \tau_1 = 0.85T.$$

   This ensures that only the middle 70% of the sample is tested.

2. For each candidate break point $\tau \in [\tau_0, \tau_1]$, estimate the Chow F-statistic:

   $$F(\tau) = \frac{(SSR_r - SSR_u(\tau))/q}{SSR_u(\tau)/(T - 2k)},$$

   where SSR_r is the restricted sum of squared residuals (assuming no break), and $SSR_u(\tau)$ is the unrestricted SSR allowing a break at τ. Here, q is the number of restrictions tested (e.g., interaction terms), and k is the number of parameters in each subsample.

3. The QLR statistic is defined as:

   $$QLR = \max_{\tau \in [\tau_0, \tau_1]} F(\tau).$$

4. Compare the computed QLR value against the critical values from a non-standard distribution (see table below).

Key Properties:

- The QLR test can detect general forms of instability such as multiple breaks or slow changes in regression coefficients.
- The distribution of QLR is not the standard F-distribution due to the maximization process across multiple potential break points.

- Critical values for the QLR test are generally higher than standard F-distribution values because of the multiple comparisons problem.
- It is strongly recommended to **always run a QLR test**, even if a specific break date is suspected, as prior knowledge may be incorrect.

Critical Values for QLR Test (15% Trimming)

Number of Restrictions (q)	10%	5%	1%
1	7.12	8.68	12.16
2	5.00	5.86	7.78
3	4.09	4.71	6.02
4	3.59	4.09	5.12
5	3.26	3.66	4.53
6	3.02	3.37	4.12
7	2.84	3.15	3.82
8	2.69	2.98	3.57
9	2.58	2.84	3.38
10	2.48	2.71	3.23

Table 1: Critical values of QLR statistics with 15% trimming. Here, q denotes the number of restrictions (i.e., number of interaction terms tested).

10 Robust (HAC) Standard Errors and Long-Run Variance

10.1 Cross-Sectional Asymptotic Variance (Review)

In a purely i.i.d. sample, the variance of the OLS estimator simplifies because cross-observations are uncorrelated. Covariances vanish except when indices match.

10.2 Time Series Case

Under autocorrelation, $\mathrm{Cov}(v_t, v_s)$ may be nonzero for $t \neq s$. Define

$$\gamma_j \;=\; \mathrm{Cov}(v_t, v_{t-j}),$$

assumed time-invariant under stationarity. The "long-run variance" Ω arises from summations of all covariances. Then

$$\mathrm{Var}\Big(\frac{1}{T}\sum_{t=1}^{T} v_t\Big) = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathrm{Cov}(v_t, v_s).$$

Collecting terms yields

$$\Omega \;=\; \frac{1}{T}\Big(T\,\gamma_0 \;+\; 2\sum_{j=1}^{T-1}(T-j)\,\gamma_j\Big).$$

10.3 Newey-West Estimator

One popular way to estimate Ω is via Newey-West:

$$\hat{\Omega} \;=\; \frac{1}{T}\Big(\hat{\gamma}_0 \;+\; 2\sum_{j=1}^{m}\big(1-\frac{j}{m+1}\big)\,\hat{\gamma}_j\Big),$$

where m is a truncation parameter balancing bias and variance. Common choices include $m \approx 0.75T$ or $m \approx 1.3\sqrt{T}$.

ECON0019 T2 Lec 8: Forecasting, Prediction, and Big Data Methods

Ambrose W

1 Lecture 8A: Forecasting and Prediction (Time Series)

1.1 Forecasting Terminology

- A forecast is an out-of-sample prediction about the future, generally denoted as

$$y_{T+h|T} = \text{Forecast of period } (T+h) \text{ given information at } T.$$

- In contrast, an in-sample prediction is sometimes called a "prediction," but not a true forecast because it uses the sample period to assess accuracy.
- A **forecast error** is

$$y_{T+h} - y_{T+h|T}.$$

  For one-step-ahead forecasting, the forecast error is

$$Y_{T+1} - Y_{T+1|T}.$$

- The bias of estimated coefficients is not of particular concern for pure forecasting, because the primary goal is predictive accuracy rather than interpretability.

1.2 Mean Squared Forecast Error (MSFE)

- For one-step-ahead forecasting (generally extendable to multi-step-ahead), the MSFE is

$$\text{MSFE} = \mathbb{E}\Big[\big(y_{T+1} - y_{T+1|T}\big)^2\Big].$$

- Its square root is the Root-MSFE (RMSFE):

$$\text{RMSFE} = \sqrt{\mathbb{E}\Big[\big(y_{T+1} - y_{T+1|T}\big)^2\Big]}.$$

- RMSFE is used to measure the spread of the forecast error distribution. A new lag or regressor is often included if it reduces RMSFE, instead of using coefficient significance tests.

Sources of Randomness in Forecast Errors

1. Uncertainty from the future value itself (even if we knew the true model parameters).
2. Additional uncertainty because parameters are estimated rather than known.

1.3 Example: ADL(1,1) One-step Forecast and Error

Consider a one-step forecast:

$$y_{T+1|T} \;=\; \hat{\beta}_0 \;+\; \hat{\beta}_1 y_T \;+\; \hat{\beta}_2 x_T.$$

The forecast error is

$$y_{T+1} \;-\; y_{T+1|T} \;=\; u_{T+1} \;-\; \big[(\hat{\beta}_0 - \beta_0) \;+\; (\hat{\beta}_1 - \beta_1) y_T \;+\; (\hat{\beta}_2 - \beta_2) x_T\big].$$

The term u_{T+1} is the intrinsic error of the model. The bracketed part is the estimation error. The MSFE can then be broken into:

$$\mathbb{E}\big[u_{T+1}^2\big] \;+\; \mathbb{E}\Big[\big((\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) y_T + (\hat{\beta}_2 - \beta_2) x_T\big)^2\Big].$$

1.4 Methods to Estimate or Approximate the RMSFE

- **Exact MSFE formula (for AR(p) models):**

$$\text{MSFE} \;=\; \sigma_u^2 \;+\; \text{Var}\big(\hat{\beta}_0 + \hat{\beta}_1 Y_T + \cdots + \hat{\beta}_p Y_{T-p+1}\big).$$

- **SER Approximation (large sample):**

$$\text{RMSFE}_{\text{SER}} \;=\; \sqrt{\frac{\text{SSR}}{T-p-1}} \;=\; \sqrt{\hat{\sigma}_u^2}.$$

  Under stationarity, with a large number of observations T relative to the number of lags p, estimation error is relatively small, and

$$\text{MSFE} \;\approx\; \sigma_u^2.$$

- **Final Prediction Error (FPE) (small sample):** when T is not large relative to p,

$$\text{RMSFE}_{\text{FPE}} \;=\; \sqrt{\frac{T+p+1}{T}}\,\hat{\sigma}_u \;=\; \sqrt{\frac{T+p+1}{T-p-1}\,\frac{\text{SSR}}{T}}.$$

  This approach accounts for the extra error from parameter estimation.

1.5 Pseudo Out-of-sample (POOS) Forecasting

One can simulate a genuine out-of-sample exercise to avoid strong assumptions about stationarity or homoskedasticity:

1. Split the sample at $(T - P)$, reserving P observations for out-of-sample testing.
2. Estimate the model with the first $(T - P)$ observations, forecast the $(T - P + 1)$-th observation, compute the forecast error.
3. Re-estimate the model including the newly used observation, forecast the next reserved observation, and compute the next error.
4. Repeat until the entire reserved set is used.
5. RMSFE is then

$$\text{RMSFE}_{\text{POOS}} \;=\; \sqrt{\frac{1}{P}\sum_{\omega=T-P+1}^{T}\big(\hat{u}_{\omega}\big)^2}.$$

1.6 Forecast Intervals

A 95% forecast interval is typically given by

$$y_{T+1|T} \;\pm\; 1.96\,\text{RMSFE}.$$

This is *not* a confidence interval for a parameter; it is an interval for a future random variable. Normality of the error term u_{T+1} is assumed (or approximated).

Forecast Intervals under Transformations

When, for instance, the regression is on $\Delta \ln(IP_t)$:

- Forecast the change in logs (percentage change).
- Convert that forecast to a forecast for the level of IP_{t+1} if the percentage change is small.
- Construct the interval boundaries similarly, but remember the relation between logs and levels.

$$\mathbb{E}[IP_{t+1}] \;=\; IP_t \exp\Big(\mathbb{E}\big[\Delta \ln(IP_{t+1})\big] + \tfrac{1}{2}\,\text{Var}\big[\Delta \ln(IP_{t+1})\big]\Big).$$

1.7 Forecasting Oil Prices (TUT Wk8 Exercise)

- **Model Selection:** Use BIC, or other criteria, to choose model specification.
- **Checking for Structural Breaks:**
  - Middle-of-sample: Use Chow or QLR tests.

  - End-of-sample: Plot forecast error and realized vs. forecasted values to spot outliers or unusual patterns.

- **Point Forecast:** Pay attention to transformations from logs to differences, etc.
- **Forecast Intervals:** Often assume normal errors for simplicity.
- **Choice of RMSFE Sample:**
  - If no structural break, use entire sample.
  - If break is in mid-sample, truncate the data.
  - If break is at the end, more sophisticated methods may be needed.

2 Lecture 8B: Forecasting & Prediction (Big Data)

2.1 Challenges of Big Data

- Large numbers of observations and regressors (possibly non-standard data, complicated functional forms).
- Models can be highly nonlinear.
- Potentially $k > N$.

2.2 Standardized Predictive Regression Model

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \; + \; u_i,$$

where x_{ji} are standardized regressors and y_i is de-meaned. We use one sample for estimation and hold out another sample for out-of-sample evaluation. The *first least squares assumption for prediction* requires that in-sample and out-of-sample data come from the same population distribution so that

$$\mathbb{E}[\, u_i \mid X_i \,] = 0$$

and $\mathbb{E}[Y \mid X]$ is the same in-sample and out-of-sample.

2.3 Mean-Squared Prediction Error (MSPE)

MSPE is defined similarly to the time-series MSFE, but in a cross-sectional or big-data context:

$$\text{MSPE} \;=\; \mathbb{E}\big[\, Y_{\text{OOS}} - \hat{Y}(X_{\text{OOS}}) \big]^2.$$

We can decompose MSPE into the "unavoidable" part (variance of the oracle predictor) plus the estimation error from not knowing the true coefficients. With k regressors and N observations, under homoskedastic errors,

$$\text{MSPE}_{\text{OLS}} \;\approx\; \big(1 + \tfrac{k}{N}\big)\, \sigma_u^2,$$

when k/N is relatively small or N is large.

2.4 Principal of Shrinkage

Allowing a small bias in the estimates can reduce variance and thus improve *out-of-sample* performance (lower MSPE). This is a classic bias-variance trade-off.

2.5 Cross Validation (CV) for MSPE Estimation and Model Selection

A common approach is m-fold cross validation, where the sample is partitioned into m folds. Each fold is used as a hold-out test set while the other $m-1$ folds are used for fitting. The overall prediction error across all folds is then aggregated:

$$\widehat{\mathrm{MSPE}}_{\mathrm{CV}} \;=\; \frac{1}{m}\sum_{j=1}^{m}\frac{n_j}{n}\,\widehat{\mathrm{MSPE}}_j.$$

This helps detect overfitting and select model specifications or tuning parameters that minimize out-of-sample error.

2.6 Ridge Regression (Tikhonov Regularization)

Ridge regression shrinks coefficients by adding an L^2 penalty:

$$SR_{\mathrm{Ridge}}\big(\boldsymbol{b};\lambda_{\mathrm{Ridge}}\big) \;=\; \sum_{i=1}^{N}\big(y_i - b_1x_{1i} - \cdots - b_kx_{ki}\big)^2 \;+\; \lambda_{\mathrm{Ridge}}\sum_{j=1}^{k}b_j^2,$$

where $\lambda_{\mathrm{Ridge}} \geq 0$ is the shrinkage parameter. Ridge does not set coefficients exactly to zero but shrinks large coefficients more aggressively. The best λ_{Ridge} can be chosen via cross validation by comparing out-of-sample errors.

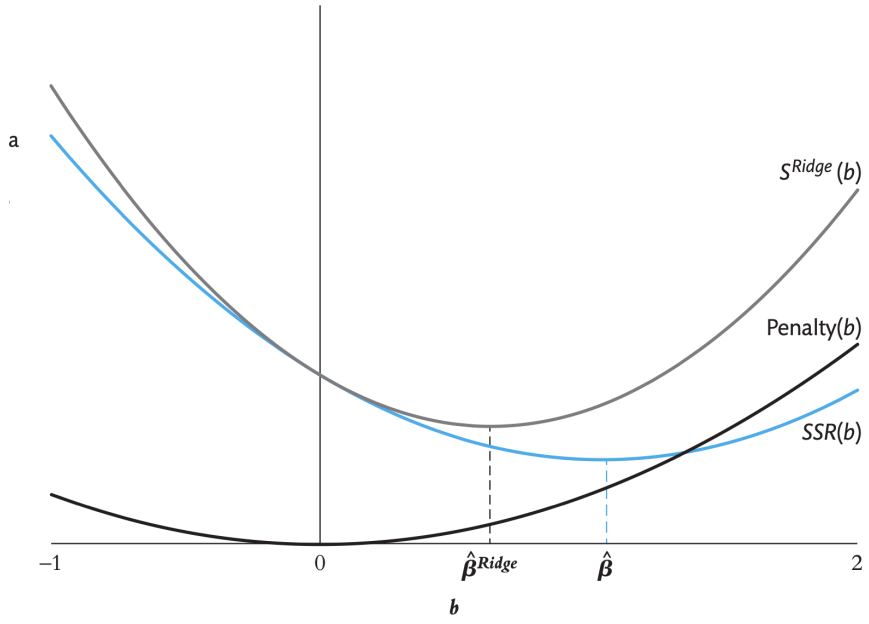

Figure 1: RIdge Regression, Penalty and SSR

2.7 Lasso (Least Absolute Shrinkage and Selection Operator)

Lasso uses an L^1 penalty:

$$S_{\text{Lasso}}(\boldsymbol{b}; \lambda_{\text{Lasso}}) \;=\; \sum_{i=1}^{N} \left(y_i - b_1 x_{1i} - \cdots - b_k x_{ki}\right)^2 \;+\; \lambda_{\text{Lasso}} \sum_{j=1}^{k} \left| \, b_j \, \right|.$$

Because the penalty is linear in $|b_j|$, Lasso can force some coefficients to be exactly zero. This is useful under a *sparse model* assumption (in which only a small fraction of regressors truly matter). The shrinkage parameter λ_{Lasso} is again typically chosen by cross validation.

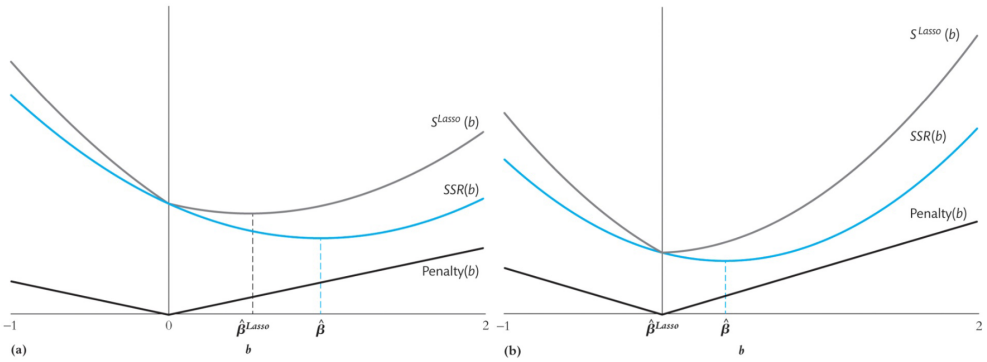

Figure 2: L1 Regularization

Warning on Reparameterization

In ordinary OLS, reparameterizing the same variable set (e.g. using male+intercept vs. male+female dummies without an intercept) yields the same fitted values. This equivalence does not hold in Lasso or Ridge because the penalty depends on the coefficient magnitudes and sign. Hence rewriting regressors can change which coefficients get shrunk toward zero.

2.8 Principal Components

Principal components (PCs) are linear combinations of correlated regressors X that are themselves uncorrelated and retain as much variation (information) as possible. Formally, the first principal component PC_1 is the linear combination $a_{1i}X_i$ that maximizes the variance subject to a normalization constraint. Then the second principal component is chosen to be orthogonal to the first, and so on. In predictive contexts, one might regress y on a subset of principal components instead of the large set of original regressors to avoid high-dimensional problems, while retaining most of the variability in X.

ECON0019 T2 Lecture 9 – Dynamic Causal Effect

Overview and Definition

Dynamic Causal Effect refers to the effect on y of a change in x over time. In an ideal setting, one would have a randomized controlled trial with clear treatment assignments. However, time series settings pose challenges:

- It may be impossible to control the "treatment" (e.g., weather).
- There may be only one unit of observation (e.g., a single market observed over many periods), which induces serial correlation.

Another conceptual approach is to treat a single market at different times as though it can be both treatment and control group, provided stationarity holds. In that scenario, a time series regression can potentially capture the dynamic causal effect.

Estimation with a Distributed-Lagged (DL) Model

A common way to estimate a dynamic causal relationship is the distributed-lag model:

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \cdots + \beta_r x_{t-r} + u_t.$$

- β_0 is the **impact effect** of a change in x: it measures the contemporaneous effect on y_t when x_t changes, holding lagged values of x constant.
- β_1 is the **1-period dynamic multiplier**: effect of x_{t-1} on y_t (again, holding other terms constant).
- β_2 is the **2-period dynamic multiplier**, etc.

Cumulative Dynamic Multiplier (CDM) is defined as the total effect of a one-time change in x on y over multiple periods. For instance, the 2-period CDM is

$$\beta_0 + \beta_1 + \beta_2.$$

Impulse Response Function

An **Impulse Response Function** (IRF) traces out how a shock to x_t affects y over various horizons. An example in macroeconomics is the effect of an exogenous monetary policy shock (change in the Federal Funds Rate):

- The horizontal axis represents time horizons (e.g., 0 to 48 months).
- The vertical axis represents the magnitude of the effect.
- A dotted line might indicate confidence intervals.

Typical plots might show:

1. Federal Funds Rate responding immediately (*strong impact effect*).
2. Industrial Production with a delay or hump-shaped response.
3. Consumer Price Index reacting slowly over time, reflecting price persistence.
4. 3-Month Treasury Bills shifting closely and quickly with the Federal Funds Rate.

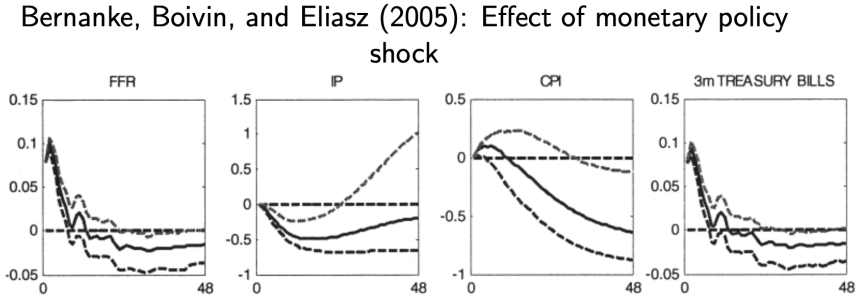

Figure 1: IRF of change in FFR

Exogeneity Reminder

For a DL model to produce unbiased estimates of the dynamic effect, one typically requires **strict exogeneity**:

$$\mathbb{E}[u_t \mid X_T, \ldots, X_t, \ldots, X_1] = 0,$$

where X_t denotes all observations of x up to time t. If only *contemporaneous exogeneity* holds, OLS might be consistent but not unbiased. Furthermore, an **autoregressive distributed-lag (ADL) model** is generally biased when strict exogeneity fails.

Standard Errors in Time Series

In practice, one may use Newey–West standard errors to correct for potential serial correlation. If Newey–West adjustments differ little from ordinary standard errors, it might indicate minimal serial correlation in u_t.

Inference on Cumulative Multipliers

Linear Combination of Variances

Cumulative multipliers often appear as sums of coefficients. For instance, to find a 1-period cumulative dynamic multiplier $\beta_0 + \beta_1$, one needs its variance:

$$\mathrm{Var}(\beta_0 + \beta_1) \;=\; \mathrm{Var}(\beta_0) \;+\; \mathrm{Var}(\beta_1) \;+\; 2\,\mathrm{Cov}(\beta_0, \beta_1).$$

A convenient way to obtain this in practice is via commands such as `lincom` in Stata (repeatedly applied for multiple horizons).

Estimating Cumulative Multipliers Directly

1-Period CDM. Consider the two-lag DL:

$$y_t \;=\; \alpha \;+\; \beta_0 x_t \;+\; \beta_1 x_{t-1} \;+\; u_t.$$

Rewriting,

$$y_t \;=\; \alpha \;+\; \beta_0 x_t \;-\; \beta_0 x_{t-1} \;+\; \beta_1 x_{t-1} \;+\; \beta_1 x_{t-1} \;+\; u_t,$$

which rearranges to

$$y_t \;=\; \alpha \;+\; \beta_0 \left(x_t - x_{t-1}\right) \;+\; \left(\beta_0 + \beta_1\right) x_{t-1} \;+\; u_t.$$

Notice that $\beta_0(x_t - x_{t-1})$ is a first difference of x. Then one may write

$$y_t \;=\; \alpha \;+\; \beta_0 \,\Delta x_t \;+\; \left(\beta_0 + \beta_1\right) x_{t-1} \;+\; u_t.$$

Hence, the coefficient on x_{t-1} in this rearranged specification is the 1-period CDM, $\beta_0 + \beta_1$.

General Case. In the general r-lag DL,

$$y_t \;=\; \alpha \;+\; \beta_0 x_t \;+\; \beta_1 x_{t-1} \;+\; \cdots \;+\; \beta_r x_{t-r} \;+\; u_t,$$

one can rewrite to isolate consecutive first differences of x. The reparameterization leads to:

$$y_t \;=\; \alpha \;+\; \delta_0 \,\Delta x_t \;+\; \delta_1 \,\Delta x_{t-1} \;+\; \cdots \;+\; \delta_{r-1} \,\Delta x_{t-r+1} \;+\; \delta_r \, x_{t-r} \;+\; u_t,$$

where

$$\delta_0 = \beta_0, \quad \delta_1 = \beta_0 + \beta_1, \quad \delta_r = \beta_0 + \beta_1 + \cdots + \beta_r.$$

In this way, each δ_j measures a cumulative sum of the β coefficients up to that point.

Exogeneity in the DL Model and Bias vs. Efficiency

Even though an ADL might be biased if strict exogeneity does not hold, it may provide more efficient estimates under certain conditions. In practice, one often attempts to incorporate sufficient control variables $w_{t-1}, w_{t-2}, \ldots$ so that the regressor x_t attains **conditional mean independence**. Symbolically, we want:

$$\mathbb{E}[u_t \mid x_t, w_{t-1}, \ldots] = \mathbb{E}[u_t \mid w_{t-1}, \ldots].$$

Nevertheless, if x_{t-j} influences the controls themselves, then a simultaneity-like problem can emerge. This possibility complicates attempts to include too many controls within a single DL specification.

Local Projections

A more flexible approach to dynamic analysis is **local projections**, which runs separate regressions for each horizon h. Suppose one is interested in horizons $h = 0, 1, \ldots, H$. A local projection framework might be:

$$y_{t+h} \; - \; y_{t-1} \; = \; \alpha_h \; + \; \beta_h \, x_t \; + \; \Gamma_h \, w_{t-1} \; + \; u_{t,h},$$

where w_{t-1} are controls included to bolster the exogeneity of x_t. Key attributes of local projections include:

- **Flexible inclusion of controls**: The specification can add any relevant w_{t-1} that helps satisfy exogeneity without the entanglements seen in large distributed-lag models.
- **Direct interpretation**: β_h directly captures the effect of x_t on changes in y from $t-1$ to $t+h$.
- **Efficiency–Robustness tradeoff**: A correctly-specified DL might be more efficient, but local projections can be more robust when exogeneity is not entirely assured.

Summary of Key Takeaways:

- Dynamic causal effects require understanding how y responds to changes in x across multiple time periods.
- DL models with sufficient lags of x (and possibly additional control variables) can quantify immediate and delayed effects.
- Cumulative multipliers sum the individual lag coefficients to assess the total impact.
- Strict exogeneity is generally needed for unbiasedness in DL models. When only contemporaneous exogeneity holds, results can be consistent but biased.
- Local projections provide an alternative by estimating separate regressions for each forecast horizon, often allowing simpler inclusion of controls.
- In practice, one must weigh the efficiency of a well-specified DL model against the robustness of local projections.

ECON0019 T2 Lecture 10: Internal & External Validity

Ambrose W

1 Conditional Exogeneity

Consider a model of the form:

$$y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \; + \; \gamma_1 w_{i1} + \cdots + \gamma_l w_{il} + u_i,$$

where the variables w are included solely to satisfy a conditional mean independence condition:

$$\mathbb{E}[u_i \mid x_{i1}, \ldots, x_{ik}, \, w_{i1}, \ldots, w_{il}] = \mathbb{E}[u_i \mid w_{i1}, \ldots, w_{il}] \neq 0.$$

Hence, conditional on w, the x's do not provide extra information about u. OLS estimates of $\beta_1, \ldots, \beta_k$ can be consistent under this condition, even if the coefficients on w (namely $\gamma_1, \ldots, \gamma_l$) themselves may not be unbiased or consistent.

2 Internal Validity

Internal validity concerns whether the statistical inferences for a causal effect in the sample and population studied are valid. Violation of basic assumptions (MLR.1 to MLR.4) causes bias and invalidates internal validity. This typically arises from:

- Omitted variable bias
- Misspecification of functional form
- Errors-in-variables bias
- Sample selection bias
- Simultaneous causality bias

2.1 Omitted Variable Bias (OVB)

An omitted variable (OV) is relevant for determining y but is absent from the regression and is also correlated with at least one regressor of interest. Formally, if

$$\mathbb{E}[u_i \mid x_i] \neq 0,$$

then OLS estimates are biased and inconsistent.

Under Different Models

- **OLS**: The OLS estimator β^{OLS} is biased and inconsistent if a relevant variable is omitted.
- **IV (Instrumental Variables)**: If an instrument z_i fails the exogeneity condition $\text{Cov}(z_i, u_i) \neq 0$, then OVB also contaminates the IV estimator.
- **Panel Data**: With fixed effects, OVB arises when relevant omitted factors vary at a different level or over time in a manner not captured by the fixed-effect transformation.

Sign of the OVB The sign of the bias can be inferred by considering how the omitted factor affects u and how x correlates with that factor. One expression is:

$$\hat{\beta_1} - \beta_1 \quad = \quad \frac{\sum_i (x_i - \bar{x})\, u_i}{\sum_i (x_i - \bar{x})^2} \quad \longrightarrow \quad \rho_{xu}\, \frac{\sigma_u}{\sigma_x},$$

where ρ_{xu} is the correlation between x and u, σ_u is the standard deviation of u, and σ_x the standard deviation of x. The product of the correlation signs indicates the direction of the bias.

Solutions to OVB

- Include the omitted variable if measurable.
- Include additional controls w if conditional exogeneity is plausibly satisfied.
- Use an IV method if direct measurement or good controls are unavailable.
- Conduct a randomized control trial if feasible.
- Use panel data and fixed effects where relevant.

2.2 Functional Form Misspecification

Non-linear relationships or missed interaction terms also introduce bias. Examples include:

- Transformations (logarithms, polynomials, interactions).
- Specific functional forms (Probit, Logit for binary outcomes; Tobit for censored data).

Ensuring the correct functional form requires diagnostic checks and possibly shrinkage techniques (e.g., LASSO) for higher-order terms.

2.3 Errors-in-Variables (Measurement Error) Bias

Real data often have measurement error in regressors. Suppose the true value is x_i^*, but only $x_i' = x_i^* + e_i$ is observed. Then:

$$y_i = \beta_0 + \beta_1 x_i^* + u_i = \beta_0 + \beta_1 (x_i' - e_i) + u_i = \beta_0 + \beta_1 x_i' + (\, u_i - \beta_1 e_i).$$

Define $\tilde{u}_i = u_i - \beta_1 e_i$. If e_i is non-trivial, then $\text{Cov}(x_i', \tilde{u}_i) \neq 0$ typically holds, leading to biased OLS estimates.

Classical Error-in-Variables (CEV) Assumption If e_i is purely random noise with

$$\text{Cov}(x_i^*, e_i) = 0 \quad \text{and} \quad \text{Cov}(u_i, e_i) = 0,$$

then

$$\hat{\beta_1} \quad \longrightarrow \quad \beta_1 \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2} \right),$$

showing *attenuation bias* (the estimated coefficient is closer to zero than the true one).

Best-Guess Measurement Error If x_i' is the best guess of x_i^* given some other w_i (e.g., $\mathbb{E}[x_i^* \mid w_i]$), and $\text{Cov}(u_i, e_i) = 0$, the error need not cause bias.

Solutions

- Obtain better or alternative data sources to reduce measurement error.
- Model the error process if appropriate (e.g., censored regression).
- Use an IV that is correlated with the mismeasured variable but uncorrelated with the error.
- Note that measurement error in y (the dependent variable) is generally less of a problem if it is random and uncorrelated with x.

2.4 Sample Selection Bias

Data may be missing or truncated for non-random reasons. If data are omitted based on y, selection bias arises.

Example To gauge performance of managed funds, one might collect data only on existing funds, thus excluding those that disappeared. This non-random selection based on past returns y causes biased estimates.

Truncation and Incidental Truncation

- **Truncation**: Observations are selected based on $y_i < c$, for instance. One may parameterize the truncated distribution and estimate via Maximum Likelihood (MLE).
- **Incidental Truncation**: y_i is observed only for some observations that satisfy another condition (possibly correlated with y). The Heckman two-stage (selection) model is commonly used, introducing the inverse Mills ratio as a control for selection.

Solutions

- Collect a sample in a way that avoids selection on y.
- Use randomized controlled experiments if feasible.
- Construct a parametric model for selection (Heckman or truncation models).

2.5 Simultaneous Causality Bias

Bias arises when the regressor x is jointly determined with the dependent variable y. An example is interest rate and inflation, where each can affect the other.

Solutions

- Randomized controlled trials (where random assignments neutralize endogeneity).
- Structural simultaneous equation models.
- Instrumental variables (provided the instrument is exogenous and relevant).

2.6 A Note on Instrumental Variables

IV methods can solve:

- Omitted variable bias.
- Measurement error bias.
- Simultaneous causality bias.

An instrument must be both *valid* (uncorrelated with error u) and *relevant* (strongly correlated with the endogenous regressor). Weak or invalid instruments lead to biased Two-Stage Least Squares (2SLS) results.

3 Problems of RCTs

While a well-run randomized controlled trial (RCT) can eliminate many biases, several pitfalls still exist if implementation is flawed:

3.1 Failure to Randomize

Participants must be randomly assigned to treatment and control groups. Non-random assignment violates exogeneity.

3.2 Non-compliance Bias

Subjects in the control group may receive treatment (or vice versa). If initial assignment is known, it can serve as an instrument for actual treatment received.

3.3 Attrition Bias

If subjects drop out based on their outcomes, a form of sample selection bias occurs.

3.4 Experimental Effects

- **Experimenter Bias**: If not double-blind, the researcher may treat groups differently.
- **Hawthorne Effect**: Participants alter behavior because they know they are studied.

4 Quasi-Experiments: Threats to Internal Validity

Quasi-experiments exploit variations in policy or environment as if randomly assigned. Nonetheless, one must ensure that the variation truly approximates randomness and that instruments are valid. Attrition or cross-boundary commuting (as in the Card and Krueger minimum wage study) can complicate identification.

5 External Validity

External validity considers whether valid results for one population or setting extend to another. It requires internal validity as a precondition.

5.1 Threats to External Validity in Experiments

- **Non-representative Sample**: Many psychology or economics experiments use students, possibly limiting generalizability.
- **Non-representative Treatment**: The experimental treatment may not be feasible in a broader context.
- **General Equilibrium Effects**: Scaling up an intervention can change overall conditions (e.g., limited number of jobs), altering the effect size.

Quasi-experiments often have an advantage in external validity compared to small-scale RCTs.

Conclusion

These notes emphasize diagnosing biases that threaten internal validity and understanding how to mitigate them. From omitted variables and measurement errors to sample selection and simultaneity, each bias compromises the reliability of causal inference. Even perfectly implemented studies must address issues of external validity before results can be generalized. Instrumental variable methods, randomized controlled trials, and quasi-experiments provide frameworks to handle these concerns, but each method imposes specific conditions that researchers must verify.