Econometrics
Honor’s Exam Review Session

Spring 2012
Eunice Han
Topics

1. OLS
   • The Assumptions
   • Omitted Variable Bias
   • Conditional Mean Independence
   • Hypothesis Testing and Confidence Intervals
   • Homoskedasticity vs Heteroskedasticity
   • Nonlinear Regression Models: Polynomials, Log Transformation, and Interaction Terms

2. Panel Data:
   • Fixed Effects: Entity FE and Time FE
   • Serial Correlation and Clustered HAC SE

3. Internal Validity and External Validity
4. Binary Dependent Variables:
   - Linear Probability Model
   - Probit Model
   - Logit Model
   - Ordered Probit Model
5. Instrumental Variables Regression
   - Conditions for Valid Instruments: Relevance and Exogeneity
   - 2SLS estimation: The First and the Second Stage Regression
   - Tests of Instrumental Validity: F-test and J-test
6. Time Series Data
   - Stationarity
   - Forecasting Models: AR and ADL Model
   - Dynamic Causal Effects: Distributed Lag Model
   - Serial Correlation and Newey-West HAC SE
Least Squares Assumptions

1. \((X_i, Y_i)\) are iid; random sampling
2. Large outliers are rare; If this condition fails, OLS estimator is not consistent.
3. \(E(u_i | X_i) = 0\); Conditional Mean Zero assumption. Xs are exogenous. This assumption fails if X and u are correlated.
4. No Perfect Multicollinearity (In multivariate regression): The regressors are said to be perfectly multicollinear if one of the regressors is a perfect linear function of the other regressor(s).
   - If all the assumptions are satisfied, the OLS estimates are unbiased and consistent.
Univariate Regression Model

• Do married men make more money than single men?

\[ lwage_i = \beta_0 + \beta_1 married_i + u_i \ ; \ i=1,2,\ldots,n \]

\[ lwage = \log(\text{monthly wage}) \]

\[ married = 1 \text{ if married, 0 if single} \]

Q1. How would you interpret \( \beta_0 \) and \( \beta_1 \)?

\( \beta_0 \) (intercept) is the average \( lwage \) of single men.

\( \beta_1 \) (slope) represents the difference in average \( lwage \) between single and married men.

Q2. In our model, which factors are likely to be in the error term?

The error term \( u_i \) captures all other factors except marital status that might affect the log of monthly wage. Education, experience, and ability are potential factors which are omitted here.
Omitted variable Bias

- Population regression equation (True world)
  \[ Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \]

- Suppose we omitted \( X_{1i} \) and estimated the following regression.
  \[ Y_i = \hat{\beta}_2 X_{2i} + \hat{\varepsilon}_i \]
  \[ E(\hat{\beta}_2) = \beta_2 + \delta \beta_1, \quad \text{where} \quad \delta = \frac{\text{Corr}(X_{1i}, X_{2i})}{\text{Var}(X_{2i})} \]

Now, the OLS estimator is no longer unbiased, and \( \text{OVB} = \delta \beta_1 \)

Q1. Under what condition, the OLS estimator suffers from OVB?
   1) The omitted variable \( X_{1i} \) is a determinant of \( Y_i \) (\( \beta_i \neq 0 \)) and
   2) \( X_{1i} \) is correlated with the regressor \( X_{2i} \) (\( \delta \neq 0 \))

Q2. Can you predict the sign of this OVB?

Q3. What can you do to solve OVB problem?
Multivariate Regression Model

\[ lwage_i = \beta_0 + \beta_1 married_i + \beta_2 edu_i + \beta_3 exp_i + \beta_4 job_i + u_i \]

Q1. How would you interpret \( \beta_0 \), \( \beta_1 \), and \( \beta_2 \)?

\( \beta_0 \): Predicted value of \( lwage \) when all the regressors are zero. Meaningless in this case.

\( \beta_1 \): Effect of unit change in “married” on \( lwage \), holding other regressors constant.

Q2. Suppose that we are interested in the effect of marriage on \( lwage \). Can we be sure that there is no OVB after we include more variables in the regression?

Q3. Supposed we are only interested in the effect of marriage on \( lwage \). Do we really need the strong assumption of exogeneity of \( Xs \), \( E(u_i | X_i) = 0 \), to obtain unbiased estimator for \( \beta_1 \)?
Conditional Mean Independence

\[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + ... + \beta_{1+r} W_{ri} + u_i \]

- X: treatment variable  
  W: control variables.
- If we are only interested in the causal effect of X on Y, we can use a weaker assumption of **Conditional Mean Independence:**
  \[ E(u \mid X, W) = E(u \mid W) \]
- The conditional expectation of u does not depend on X if control for W. Conditional on W, X is as if randomly assigned, so X becomes uncorrelated with u, but W can still be correlated with u.
- Under the conditional mean independence assumption, OLS can give us the unbiased and consistent estimator for \( \beta_1 \), but the coefficients for W can be biased.
Measure of Fit

Q. What is $R^2$ and what does it tell us?
   
   $R^2$ is the fraction of variation of dependent variable which is explained by our model.

Q. Why is Adjusted $R^2$?

   Unlike $R^2$, adjusted $R^2$ adjusts for the number of regressors in a model because it increases only if the new regressor improves the model more than would be expected by chance.

Q. What is SER (Standard Error of Regression)?

   SER estimates the standard deviation of the error term. A large SER implies that the spread of the observations around the regression is large, and there could be other important factors that we have not included in the regression.
Hypothesis Testing

- H₀: β₁ = 0
- H₁: β₁ ≠ 0 (two-sided)

**T-test:** By the Central Limit Theorem, t-statistics is normally distributed when n is large enough.

\[
t\text{-statistics} = \frac{\hat{\beta}_1 - 0}{\sqrt{\text{Var}(\hat{\beta}_1)}} \sim N(0,1)
\]

If |t| > 1.96, we reject null hypothesis at 5% significance level.

**P-value:** The p-value is the probability of drawing a value of \( \hat{\beta}_1 \) that differs from 0, by at least as much as the value actually calculated with the data, if the null is true. If p-value is less than 0.05, we reject the null hypothesis at 5% significance level.

\[
P\text{-value} \approx \Pr(|z| > |t|) = 2\Phi(-|t|)
\]
Confidence Interval

- **Confidence Interval**: An interval that contains the true population parameter with a certain pre-specified probability. (usually 95%)

- Confidence Interval is useful because it is impossible to learn the exact value of population mean using only the information in one sample due to the random sampling error. However, it is possible to use this data to construct a set of values that contains the true population mean with certain probability.

- How do we construct the 95% confidence interval?

  \[
  \Pr(-1.96 \leq t \leq 1.96) = 0.95
  \]

  \[
  \left[ \hat{\beta}_1 - 1.96 \text{SE} (\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + 1.96 \text{SE} (\hat{\beta}_1) \right]
  \]

  If this confidence interval contains 0, we fail to reject \( H_0 : \beta_1 = 0 \)
Homoskedasticity vs Heteroskedasticity

• **Homoskedasticity**: The error term $u_i$ is homoskedastic if the variance of the conditional distribution of $u_i$ given $X_i$ is constant for $i=1,2,...n$.

• **Heteroskedasticity**: The variance of the conditional distribution of $u_i$ given $X_i$ is different across $X_i$.

Q. What happens if we have heteroskedasticity problem?

The OLS estimator is still unbiased and consistent, as long as the OLS assumptions are met (esp. $E(u_i|X_i) = 0$). However, our SE calculated using homoskedasticity-only formula gives us a wrong answer, so the hypothesis testing and confidence intervals based on homoskedasticity-only formula are no longer valid. We have to use **heteroskedasticity-robust SE**.
Joint Hypothesis Testing

- For joint hypothesis testing, we use **F-test**.
- Under the null hypothesis, in large samples, the F-statistic has a sampling distribution of $F_{q,\infty}$. That is,
  \[
  \text{F-statistic} \sim F_{q,\infty}
  \]
  where $q$ is the number of coefficients that you are testing.
- If F-statistics is bigger than the critical value or p-value is smaller than 0.05, we reject the null hypothesis at 5% significance level. (ex. The critical value for $F_{2,\infty}$ at 5% significance level is 3)
- If $H_0: \beta_2 = \beta_3 = 0$ and F-stat > 3, we reject the null and conclude that at least one of the coefficients is not zero,
Nonlinear Regression (1)

1. Polynomials in X

\[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_r X_i^r + u_i \]

- The coefficients do not have a simple interpretation because it is impossible to change X holding \( X^2 \) constant.
- Which polynomial model to choose?
  1. Linear vs. Quadratic? T-test on \( H_0: \beta_2 = 0 \)
  2. Quadratic vs. Cubic? T-test on \( H_0: \beta_3 = 0 \)
  3. Linear vs. non-linear? F-test on \( H_0: \beta_2 = \beta_3 = 0 \)

If fails to reject, linear model is preferred. If rejects, at least one of the coefficients are not zero.
Nonlinear Regression (2)

2. Log transformation

<table>
<thead>
<tr>
<th>Case</th>
<th>Regression Specification</th>
<th>Interpretation of $\beta_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear-Log</td>
<td>$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$</td>
<td>1% change in $X \Rightarrow 0.01 \beta_1$ change in $Y$</td>
</tr>
<tr>
<td>Log-Linear</td>
<td>$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$</td>
<td>1 unit change in $X \Rightarrow 100 \beta_1$ % change in $Y$</td>
</tr>
<tr>
<td>Log-Log</td>
<td>$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$</td>
<td>1% change in $X \Rightarrow \beta_1$ % change in $Y$</td>
</tr>
</tbody>
</table>

- In log-log specification, $\beta_1$ has elasticity implication.
Nonlinear Regression (3)

3. The interaction terms

\[ Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \]

- If we believe that the effect of \( X_{2i} \) on \( Y \) depends on \( X_{1i} \), we can include the “interaction term” \( X_{1i} * X_{2i} \) as a regressor.

\[ Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} * X_{2i}) + u_i \]

- Suppose we want to plot the regression line of \( Y \) on \( X_{2i} \) when \( X_1 = a \) and \( X_1 = b \). Assume that all coefficients are positive and \( a \) and \( b \) are also positive with \( a < b \).

- When \( X_1 = a \), \( Y_i = \beta_0 + \beta_1 a + (\beta_2 + \beta_3 a) X_{2i} \)
- When \( X_1 = b \), \( Y_i = \beta_0 + \beta_1 b + (\beta_2 + \beta_3 b) X_{2i} \)
• The intercepts of the bottom regression line represent the predicted value of \( Y \) when \( X_2 = 0 \) for \( X_1 = a \).
• The slope of the bottom regression line represents the effect on \( Y \) of a unit change in \( X_2 \) given \( X_1 = a \).
Example of the Interaction Terms

$X_{\text{black}} = 1$ if observation is black and 0 otherwise. $X_{\text{exp}} = \text{years of experience}$. If we believe that the effect of experience on wage depends on individual’s race, we can add the interaction term, $X_{\text{black}} \times X_{\text{exp}}$ to the regression.

$$lwage_i = \beta_0 + \beta_{\text{black}} X_{\text{black}} + \beta_{\text{exp}} X_{\text{exp}} + \beta_{\text{black}_\text{exp}} (X_{\text{black}} \times X_{\text{exp}}) + u_i$$

- If $X_{\text{black}} = 1 \rightarrow Y_i = \beta_0 + \beta_{\text{black}} + (\beta_{\text{exp}} + \beta_{\text{black}_\text{exp}}) X_{\text{exp}} + u_i$
- If $X_{\text{black}} = 0 \rightarrow Y_i = \beta_0 + \beta_{\text{exp}} X_{\text{exp}} + u_i$

- To see whether the average wage for no experience (intercept) is the same for two groups, test $H_0 : \beta_{\text{black}} = 0$
- To see whether effect of experience on wage (slope) is the same for two groups, test $H_0 : \beta_{\text{black}_\text{exp}} = 0$
Types of Data

- **Panel Data**: $N$ different entities are observed at $T$ different time periods
  1. **Balanced Panel**: All variables are observed for each entity and each time period.
  2. **Unbalanced Panel**: There are missing data for at least one time period for at least one entity.

- **Cross-Sectional Data**: $N$ different entities are observed at the same point in time

- **Time Series Data**: 1 entity is observed at $T$ different time periods
Panel Data

- Since each entity is observed multiple times, we can use fixed effect to get rid of the OVB, which results from the omitted variables that are invariant within an entity or within a period.

- **Entity Fixed Effects** control for omitted variables that are constant within the entity and do not vary over time
  - ex. gender, race, or cultural and religious characteristics of each State

- **Time Fixed Effects** control for omitted variables that are constant across entities but vary over time
  - ex. national level anti-crime policy or national safety standard each year
Fixed Effects Regression Model

\[ Y_{it} = \beta_0 + \beta_1 X_{1it} + \ldots + \beta_k X_{kit} + (\gamma_1 D_1 + \ldots + \gamma_{N-1} D_{N-1}) + (\delta_1 P_1 + \ldots + \delta_{T-1} P_{T-1}) + u_{it} \]

\[ Y_{it} = \beta_1 X_{1it} + \ldots + \beta_k X_{kit} + \gamma_i + \delta_t + u_{it} \]

- \( i=1,2,\ldots,N \) and \( t=1,2,\ldots,T \). So, \# of observation = \( i \times t \).
- \( X_{kit} \) means \( k^{th} \) variable for individual \( i \) at time \( t \)
- \( D_1, D_2, \ldots, D_{N-1} \) are entity dummy variables
- \( P_1, P_2, \ldots, P_{T-1} \) and time dummy variables
- \( \gamma_i \) are entity fixed effect, and \( \delta_t \) are time fixed effect.
- When \( T=2 \), before-and-after regression with entity FE and without an intercept produce the same OLS estimates for \( \beta \):
Special case: FE Regression and Difference Regression for T=2

- Consider the regression with entity FE:

  (1) \( Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it} \quad \text{for } t=1,2 \)

- Change regression for entity FE regression:

  (2) \( Y_{i2} - Y_{i1} = \beta_1 (X_{i2} - X_{i1}) + (u_{i2} - u_{i1}) \)

  \( \Rightarrow \beta_1 \text{ in (1)} = \beta_1 \text{ in (2)} \) without intercept.

- Consider the regression with entity FE and time FE:

  (3) \( Y_{it} = \alpha_i + \beta_1 X_{it} + \delta D_{2i} + u_{it} \quad \text{for } t=1,2 \)

- Change regression for entity FE and time FE regression:

  (4) \( Y_{i2} - Y_{i1} = \delta + \beta_1 (X_{i2} - X_{i1}) + (u_{i2} - u_{i1}) \)

  \( \Rightarrow \beta_1 \text{ in (3)} = \beta_1 \text{ in (4)} \) with intercept.
Differences-in-Differences Estimator

\[ Y_{it} = a + \alpha(Treatment_{it}) + \beta(\text{After}_{it}) + \gamma(Treatment_{it} \ast \text{After}_{it}) + u_{it} \]

- Treatment\(=\) 1 for treatment group, 0 for control group
- After\(=\) 1 after the program, 0 before the program

• For treatment group,
  - After the program \(\Rightarrow Y = a + \alpha + \beta + \gamma\)
  - Before the program \(\Rightarrow Y = a + \alpha\)
  - The effect of the program: After-before \(\Rightarrow \beta + \gamma\)

• For control group,
  - After the program \(\Rightarrow Y = a + \beta\)
  - Before the program \(\Rightarrow Y = a\)
  - The effect of the program: After-before \(\Rightarrow \beta\)

• Differences-in-Differences estimator: \((\beta + \gamma) - \beta = \gamma\)
Serial Correlation in Panel Data

- If there is a correlation of the error terms from year to year for a given entity, called **serial correlation**, conditional on the regressors and entity FE, OLS estimate of $\beta$ in panel data is still *unbiased*.

- However, the estimated standard errors will be wrong, which implies that we can’t carry on the usual hypothesis tests. In this case, we have to use Heteroskedasticity-Autocorrelation-Consistent (HAC) standard errors.

- **Clustered HAC SE** is robust to both heteroskedasticity and serial correlation of the error terms in panel data. It allows the errors to be correlated within a cluster, but not across the clusters.
Internal and External Validity

- Threats to internal validity
  1. OVB
  2. Misspecification of the Functional Form
  3. Measurement Error
  4. Sample Selection Bias
  5. Simultaneous Causality Bias
    \[ E(u | X) \neq 0, \] so OLS estimates are biased.

- Threats to external validity
  1. Differences in population studied and population of interest
  2. Difference in setting studied and setting of interest
Models for Binary Dependent Variable

- Linear Regression Model
  - Linear Probability Model
- Non-Linear Regression Models
  - Probit Model
  - Logit Model
  - Ordered Probit Model
- The main idea of the model with a binary dependent variable is to interpret the population regression as the probability of success given X, \( \Pr(Y = 1 \mid X) \).
- If \( Y \) is binary, \( \mathbb{E}(Y \mid X) = 1 \times P(Y = 1 \mid X) + 0 \times P(Y = 0 \mid X) = P(Y = 1 \mid X) \)
Linear Probability Model (LPM)

\[ Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i \]

\[ E(Y_i \mid X_i) = \Pr(Y_i = 1 \mid X_{1i}, \ldots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} \]

- This is a probability model. \( \beta_1 \) tells you the change in probability that \( Y=1 \) for a unit change in \( X_1 \), holding other regressors constant.

- The predicted value from the regression is the predicted probability that \( Y=1 \) for the given values of the regressors.

- Ex. For regression model \( Y_i = \beta_0 + \beta_1 X_i + u_i \)

\[ E(Y_i = 1 \mid X = 5) = \Pr(Y_i = 1 \mid X = 5) = \hat{\beta}_0 + \hat{\beta}_1 \ast 5 \]
Advantage and Disadvantages of LPM

- **Advantage:**
  - Simple to estimate and to interpret. Statistical inference is the same as for multivariate regression model.
  - One unit increase in X increases Pr(Y=1) by 100* $\beta_1$ percentage point.

- **Disadvantages:**
  - The predicted probabilities can be < 0 or >1, which does not make sense in probability.
  - These disadvantages can be solved by introducing nonlinear probability model: probit and logit regression.
Probit Regression Model

- **Probit regression** models the probability that $Y=1$ using cumulative standard normal distribution, $\Phi(z) = P(Z \leq z)$ where $Z \sim N(0,1)$

\[
E(Y_i \mid X_1, X_2) = \Pr(Y_i = 1 \mid X_1, X_2) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2) = \Phi(z)
\]

- Predicted probability of $Y=1$ given $X$ is calculated by computing the “z-score”, $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, and looking up this z-value in the standard normal distribution table.

- $\beta_1$ measures the change in the z-score for a unit change in $X_1$ holding $X_2$ constant, not the change in $\Pr(Y = 1 \mid X_1, X_2)$.
Logit Regression Model

- **Logit regression** models the probability that \( Y=1 \) using the logistic function evaluated at \( z = \beta_0 + \beta_1 X_1 \)

\[
E(Y_i \mid X_i) = \Pr(Y_i = 1 \mid X_i) = F(\beta_0 + \beta_1 X_1) = \frac{1}{1 + \exp(- (\beta_0 + \beta_1 X_1))}
\]

- Predicted probability of \( Y=1 \) given \( X \) is calculated by computing \( z-value, \ z = \beta_0 + \beta_1 X_1 \), and looking up this \( z \)-value in the standard logistic distribution table

- \( \beta_1 \) measures the change in the \( z \)-score for a unit change in \( X_1 \), not the change in \( \Pr(Y = 1 \mid X_1) \)
Issues in Probit/Logit Model

• The “S” shape of the curves of probit and logit models ensures:
  • $0 \leq \Pr(Y = 1 \mid X) \leq 1$, for all $X$
  • $\Pr(Y = 1 \mid X)$ to be increasing in $X$ for positive coefficients.

• We can interpret the sign of the coefficients in the usual way, but we cannot interpret the magnitude of the coefficients in the usual way. The coefficient itself is meaningless.

• To compute the effect on the probability of $Y=1$ of change from $X_1=a$ to $X_1=b$, holding other regressors constant (say $X_2=c$),
  \[
  \Pr(Y_i = 1 \mid X_1 = b, X_2 = c) - \Pr(Y_i = 1 \mid X_1 = a, X_2 = c) \\
  = \Phi(\beta_0 + \beta_1 \ast b + \beta_2 \ast c) - \Phi(\beta_0 + \beta_1 \ast a + \beta_2 \ast c)
  \]
Ordered Probit Model

- Ordered probit model is used when the dependent variable consists of ordered categories. For example,
  
  \[ Y_i = 1 \text{ if } Y_i^* \leq \text{cutoff1} \]
  
  \[ Y_i = 2 \text{ if } \text{cutoff1} < Y_i^* \leq \text{cutoff2} \]
  
  \[ Y_i = 3 \text{ if } \text{cutoff2} < Y_i^* \]

  where \( Y_i^* \) is the latent variable (usually true values).

- Suppose \( Y_i^* = \beta_0 + \beta_1 X_i + u_i \) where \( u_i \sim N(0,1) \).

  \[
  \Pr(Y_i = 1 \mid X) = \Phi[\text{cutoff1} - (\beta_0 + \beta_1 X_i)]
  
  \Pr(Y_i = 2 \mid X) = \Phi[\text{cutoff2} - (\beta_0 + \beta_1 X_i)] - \Phi[\text{cutoff1} - (\beta_0 + \beta_1 X_i)]
  
  \Pr(Y_i = 3 \mid X) = 1 - \Phi[\text{cutoff2} - (\beta_0 + \beta_1 X_i)]
  
- To find the effect of changing \( X \) from \( a \) to \( b \) on probability of being in category \( y \), we need to compute the “before-and-after” change:

  \[
  \Pr(Y_i = y \mid X = b) - \Pr(Y_i = y \mid X = a)
  \]
Example of Probit Model:

- Suppose we regressed mortgage denial rate on race and the ratio of debt payments to income (P/I), and obtained following result.

\[
Deny = -2.26 + 2.74 \times P/I + 0.71 \times black
\]

\[
(0.16) \quad (0.44) \quad (0.08)
\]

\[
\text{Pr}(\text{deny}=1 | \text{P/I ratio, black}) = \Phi(-2.26 + 2.74 \times P/I ratio + 0.71 \times black)
\]

Q. Calculate the estimated effect of being black when P/I ratio = 0.3.

\[
\text{Pr}(\text{deny}=1 | 0.3, \text{black}=1) = \Phi(-2.26+2.74\times0.3+.71\times1) = 0.233
\]

\[
\text{Pr}(\text{deny}=1 | 0.3, \text{black}=0) = \Phi(-2.26+2.74\times0.3+.71\times0) = 0.075
\]

\[
\Rightarrow \text{Pr}(\text{deny}=1 | 0.3, \text{black}=1) - \text{Pr}(\text{deny}=1 | 0.3, \text{black}=0) = 0.158
\]

The rejection probability increases by .158. Thus, the chance of being denied on mortgage is 15.8 percentage point higher for the black applicants compared to the non-black.
Endogenous independent variable

- The OLS assumption of exogeneity, $E(u | X) = 0$, is violated if there is correlation between $X$ and $u$. Then, $X$ is said to be endogenous variable.
- If the exogeneity assumption is violated, the coefficient of $X$ in OLS regression is biased and inconsistent. Thus, the causal interpretation of the coefficients are not valid.
- IV regression helps us get an unbiased and consistent estimator of the causal effect of changes in $X$ on $Y$.
- When should you use IV regression? When you believe that $X$ is correlated with error term $u$ and you have valid instruments.
Instrumental Variable (IV) Regression

Regression Model: \( Y_i = \beta_0 + \beta_1 X_i + u_i \)

- Suppose \( E(u \mid X) \neq 0 \); \( X \) is endogenous.
- IV regression breaks \( X \) into two parts by using the instrumental variable \( Z \); a part that is correlated with \( u \), and a part that is not correlated with \( u \) (this part is correlated with \( Z \)).
- By focusing on the variation in \( X \) that is not correlated with \( u \) using the help from the instrument \( Z \), it is possible to obtain an unbiased and consistent estimate of \( \beta_1 \).
Conditions for a Valid instrumental Variable

• For Z to be a valid instrument, it must satisfy two conditions:

  1) **Instrument relevance**: $\text{Corr}(Z, X) \neq 0$
     The instrument should be correlated with X since we will use this correlation to take out the part of X, which is uncorrelated with the regression error term.

  2) **Instrument exogeneity**: $\text{Corr}(Z, u) = 0$
     Z must be exogenous. Z should affect Y only through X, not through $u_i$. In other words, Z has an indirect effect on Y through X, but has no direct effect on Y. Z is correlated with X but uncorrelated with any other variables that can affect Y.
Two Stage Least Squares (TSLS)

- Suppose we have a model: \( Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i \)
where \( X \) is endogenous variable and \( W \) is included exogenous variable.

- **First stage regression**: \( X_i = \pi_0 + \pi_1 Z_i + \pi_2 W_i + v_i \)
  - Regress \( X_i \) on \( Z_i \) and \( W_i \) to isolate the part of \( X_i \) that is uncorrelated with \( u_i \)
  - Compute the predicted values of \( X_i \): \( \hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i + \hat{\pi}_2 W_i \)
  - Now, \( \hat{X}_i \) is uncorrelated with \( u_i \) by construction.
  - A regression is called **reduced form** if it regresses an endogenous variable on all the available exogenous variables (both \( Zs \) and \( Ws \)).
- The first stage regression and the regression of \( Y \) on \( Zs \) and \( Ws \) are in reduced form.
Second stage regression: $Y_i = \beta_0 + \beta_1 \hat{X}_i + \beta_2 W_i + u_i$

- Regress $Y_i$ on $\hat{X}_i$ and $W$, using OLS.
- Because $\hat{X}_i$ is uncorrelated with $u_i$, all the regressors are exogenous in the second stage regression.
- The resulting estimator of $\beta_1$ is called TSLS estimator.
- With some algebra, we can show:

$$\hat{\beta}_{1}^{2SLS} = \frac{\text{Cov}(\hat{X}_i, Y_i)}{\text{Var}(\hat{X}_i)} = \frac{\sigma(Z, Y)}{\sigma(Z, X)} = \frac{\alpha}{\delta}$$

where $\sigma$ is sample covariance, $\alpha$ is the coefficient for regression of $Y$ on $Z$, and $\delta$ is the coefficient for the regression of $X$ on $Z$. 
The General IV Regression Model

\[ Y = \beta_0 + \beta_1 X + ... + \beta_k X_k + \beta_{k+1} W_1 + ... + \beta_{k+r} W_r + u \]

- \( X_1, \ldots, X_k \) are **endogenous regressors** (correlated with \( u \))
- \( W_1, \ldots, W_r \) are **included exogenous regressors** (uncorrelated with \( u \))
- \( Z_1, \ldots, Z_m \) are \( m \) **instrumental variables**
- The coefficients \( \beta_1, \ldots, \beta_k \) are said to be:
  1) **exactly identified** if \( m = k \)
     - The number of instruments = the number of endogenous regressors.
  2) **over-identified** if \( m > k \)
     - The number of instruments > the number of endogenous regressors.
  3) **under-identified** if \( m < k \)
     - The number of instruments < the number of endogenous regressors.
     In this case, you need to find more instruments.
Testing the Relevance of Instruments

**Step 1:** Run the first stage regression with all Zs and Ws;

\[ X_i = \pi_0 + \pi_1 Z_{1i} + \pi_2 Z_{2i} + \ldots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \ldots + \pi_{m+r} W_{ri} + \nu_i \]

**Step 2:** Run the joint hypothesis to test \( H_0: \pi_1 = \pi_2 = \ldots = \pi_m = 0 \)

**Step 3:** The instruments are relevant if at least one of the m coefficient is nonzero. If the F-statistic > 10, reject the null and conclude that our instrument(s) are highly relevant. If the F-statistic < 10, our instruments are weak.

- Weak instruments explain very little of the variation in \( X \). If instruments are weak, the sampling distribution of TSLS estimator and its t-statistic are not normally distributed even with large sample, so \( \hat{\beta}_{1sLS}^2 \) is inconsistent.
Testing the Exogeneity of Instruments

- We can use the J-test if the model is over-identified \((m>k)\);

If instruments are exogenous, they are uncorrelated with \(u_i\), and also (approximately) uncorrelated with the residual from TSLS regression.

**Step 1:** Run the TSLS regression to find the residuals \(\hat{u}_{i}^{TSLS}\)

**Step 2:** Regress the residuals on the Zs and Ws;

\[
\hat{u}_{i}^{TSLS} = \delta_0 + \delta_1 Z_{1i} + \ldots + \delta_m Z_{mi} + \delta_{m+1} W_{1i} + \ldots + \delta_{m+r} W_{ri} + \varepsilon_i
\]

**Step 3:** Compute the F-statistic testing that all instruments are exogenous. \(H_0: \delta_{1i} = \delta_{2i} = \ldots = \delta_{mi} = 0\)

**Step 4:** Compute the J-statistic = \(m\times F\). J-statistic has a chi-squared distribution with \(m\) (number of instruments) – \(k\) (number of endogenous variables) degrees of freedom; \(J \sim \chi^2_{m-k}\)

**Step 5:** If the J-statistic < the critical value, then we fail to reject the null and conclude that the instruments are exogenous.
Time Series Data

- Data for the same entity observed over time. There is one observation at one point in time.
- When we analyze time series data, we are mainly interested in “impulse(shock)-response” relationship. For example, we can examine the impact of rise in oil price on stock market index or the impact of monetary policy on aggregate demand.
- The purpose in time series study is often to make a good forecast.
- The first lag of $Y_t$ is $Y_{t-1}$, the second lag is $Y_{t-2}$, etc.
- The correlation of a series with its own lagged values is called **autocorrelation** or **serial correlation**
Stationarity

- Time series $Y_t$ is **stationary** if its distribution does not change over time; $E(Y_t)$ and $\text{Var}(Y_t)$ is constant over time and $\text{Cov}(Y_t, Y_{t-j})$ does not depend on $t$.

- Stationarity implies that history is relevant for forecasting the future.

- If $Y_t$ is not stationary, our normal regression theory breaks down: The $t$-statistics don’t have standard normal distributions even in a large sample, so we can’t check significance of the coefficients in the usual way.

- If data is not stationary, we can transform it usually by taking **first differences** to make it stationary before the estimation.
Forecasting Model (1)

1. **Autoregressive Model (AR)**
   
   The $p^{th}$-order autoregression, $\text{AR}(p)$, is a linear regression in which $Y_t$ is regressed on its first $p$ lagged values:
   
   $$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + u_t$$

   • Assumption 1: $\text{E}(u_t \mid Y_{t-1}, Y_{t-2}, \ldots ) = 0$: The error terms are serially uncorrelated.

   • Assumption 2: $Y_t$ needs to be stationary.

   • The best forecast of $Y_{T+1}$ based on its entire history depends on only the most recent $p$ past values.

   $$Y_{T+1 \mid T} = \beta_0 + \beta_1 Y_T + \beta_2 Y_{T-1} + \ldots + \beta_p Y_{T-p+1}$$

   • The coefficients do **not** have a causal interpretation.
2. Autoregressive Distributed Lag (ADL) Model

The autoregressive distributed lag model, ADL(p,r), is a linear regression in which \( Y_t \) is regressed on \( p \) lags of \( Y \) and \( r \) lags of \( X \):

\[
Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \delta_1 X_{t-1} + \delta_2 X_{t-2} + \ldots + \delta_r X_{t-r} + u_t
\]

- Assumption 1: \( E(u_t | Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}, X_{t-1}, X_{t-2}, \ldots, X_{t-r}) = 0 \).
  No serial correlation and no additional lags beyond \( p \) and \( r \) belong to ADL(p,r) model
- Assumption 2: The random variable \( (Y_t, X_t) \) are stationary.
- The coefficients do not have a causal interpretation.
Model Selection

- Tradeoff: If including too few lags, the coefficients could be biased. If adding too many lags, forecast errors could be too large.
- When we decide the optimal length of lag (p), use Information Criterion: BIC or AIC
- Bayesian Information Criterion: \( BIC = \ln\left(\frac{RSS(p)}{T}\right) + (p + 1)\frac{\ln T}{T} \)
  - First term: Always decreasing in p (larger p, better fit)
  - Second term: Always increasing in p. (Penalty term for increased forecast error due to increased p)
  \( \rightarrow \) Choose p that minimizes BIC.
Dynamic Causal Effects

- **Dynamic causal effect**: effect of a change in $X$ on $Y$ over time. A shock to $X_{t-2}$ affects $Y_t$ directly and indirectly through $X_{t-1}$.

- The **Distributed Lag model (DL)**:

  $$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 X_{t-2} + \ldots + \beta_{r+1} X_{t-r} + u_t$$

  - $\beta_1$ = **impact effect of a change in $X$** = effect of a change in $X_t$ on $Y_t$, holding past $X_t$s constant
  - $\beta_2$ = **1-period dynamic multiplier** = effect of a change in $X_{t-1}$ on $Y_t$, holding constant $X_t$, $X_{t-2}$, $X_{t-3}$,\ldots
  - $\beta_3$ = **2-period dynamic multiplier** = effect of a change in $X_{t-2}$ on $Y_t$, holding constant $X_t$, $X_{t-1}$, $X_{t-3}$,\ldots
  - **Cumulative dynamic multipliers** = the 2-period cumulative dynamic multiplier $= \beta_1 + \beta_2 + \beta_3$
Serial Correlation and HAC Standard Errors

- When $u_t$ is serially correlated, the OLS coefficients are consistent, but the usual SE without correcting for serial correlation is wrong. So t-statistic based on this SE is also wrong.
- In panel data, we use clustered HAC SE to correct for serial correlation. It allows the errors to be correlated within a cluster but not across the clusters. This approach requires the number of entities to be $> 1$.
- In time series data, the number of entities $= 1$ for each period, so clustering is not available. Instead, we use Newey-West HAC SE, which estimates correlations between lagged values in time series.
Newey-West HAC SE

- Too little lags ➔ some of the autocorrelation that might be important for estimation could be missing from the regression
- Too many lags ➔ there could be a large estimation error
- The rule of thumb to decide how many lags we will have to estimate is to use the truncation parameter \( m \) (number of lags) to employ Newey-West HAC SE: \( m = 0.75T^{1/3} \), where \( T \) is the number of observation (or the number of periods)
- In AR(\( p \)) model or ADL(\( p, q \)) model, the error terms are serially uncorrelated if we include enough lags. Thus, we do not need to use Newey-West HAC SE if the model includes the optimal number of lags. In DL model, however, serial correlation is unavoidable and Newey-West HAC SE is preferred.
Good Luck on the Exam!