OLS Asymptotics

OLS is consistent, even if we replace the assumption $E(u \mid x_1, \ldots, x_n) = 0$ by the assumption that $E(u) = 0$ and $Cov(x_j, u) = 0$ for all j. (If this does not hold, OLS is biased and inconsistent.)

Asymptotic normality: Under the standard assumptions, including homoskedasticity, but not normally distributed error terms:

- $\sqrt{n}\,(\beta_j^{\wedge} - \beta_j) \sim N(0, \sigma^2/a_j^2)$, asymptotically, where $a_j^2 = \text{plim}(\sum r^{\wedge}{}_{ij}^{2}/n)$, and $r^{\wedge}{}_{ij}$ are the residuals from regressing x_j on the other independent variables
- $\sigma^{\wedge 2}$ is consistent for σ^2
- $(\beta_j^{\wedge} - \beta_j)/se(\beta_j^{\wedge}) \sim N(0, 1)$, asymptotically

$se(\beta_j^{\wedge})$ is proportional to $1/\sqrt{n}$, since $Var(\beta_j^{\wedge}) = \sigma^{\wedge 2}/SST_j(1 - R_j^2)$ and $SST_j \approx n\sigma_j^2$, while all the other factors converge to fixed numbers.

LM Statistic:

- Testing: $\beta_{k-q+1} = \ldots = \beta_k = 0$, in the model $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$.
- Find the residuals from $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-q} x_{k-q} + u\sim$.
- Regress $u\sim$ on $x_1, \ldots, x_k$ and obtain R_u^2 from this regression.
- Under the null hypothesis, $LM = nR_u^2 \sim \chi_q^2$.
- This is equivalent to testing $(SSR_{\text{restricted}} - SSR_{\text{unrestricted}})/\sigma^{\wedge 2} \sim \chi_g^2$, where $\sigma^{\wedge 2} = SSR_r/(T - k + q)$

Likelihood Ratio Test: $LR = 2(l(\theta^{\wedge}) - l(\theta_0)) \sim \chi^2{}_g$.

- θ_0 is the restricted estimate of θ (under H_0); $\theta^{\wedge}$ is the unrestricted estimate.
- l( ) is the likelihood of an estimate.

Wald Test: $W = (SSR_{\text{restricted}} - SSR_{\text{unrestricted}}) / \sigma^{\wedge 2} \sim \chi_g^2$.

- $\sigma^{\wedge 2} = SSR_u / (t - k)$ – the estimate from the unrestricted model.

Specification and Data Problems

Functional Form Misspecification: Tests that the functional form is wrong, assuming that there are no omitted variables.

- F Test: Add quadratics ort other functions of the independent variables to the model and see if they explain significantly more.
- RESET Test:
  - Save $y^{\wedge}$ from the original regression.
  - Regress Y on the original variables together with $y^{\wedge 2}$ and $y^{\wedge 3}$.
  - Test the joint restriction that the coefficients on $y^{\wedge 2}$ and $y^{\wedge 3}$ are zero. If they are insignificant, the functional form is not rejected.
- Non-nested models:
  - Combine the independent variables from the two models. Test the combined model against each alternative model. (Both might be rejected, or neither might be rejected. Hopefully this doesn't happen.)
- Davidson-MacKinnon
  - Add the predicted values from one regression as the regressor in the other regression and test their significance.

Proxy Variables: Used in place of *unobservable* variables.

- Necessary Assumptions
  - The error term in the model is uncorrelated with the proxy variable.

    - The error in the proxy's relation to the unobservable variable is uncorrelated with the other independent variables.

Measurement Error

- In the dependent variable, this just increases the variance (assuming that measurement error is not correlated with the error term).
- In the explanatory variables:
    - Classical Errors in Variables (CEV): Assume that measurement error (e_1) is uncorrelated with the true x-value, x*. Then the observed value, x, is correlated with the error. In one variable, this means $|\text{plim}(\beta_1^{\wedge})| \leq |\beta_1|$, which is called attenuation bias.
    - If we assume e_1 and x are uncorrelated, then e_1 and x* are correlated, which is less likely but causes no attenuation bias.

Outliers: Observations that are very different from the rest of the population. Their inclusion causes regression results to change. (Looking at observations with large residuals is not enough.)

- Drop them and regress again (reporting all results).
- Use a robust regression method.

Instrumental Variables and Two-Stage Least Squares

Model: $y = \beta_0 + \beta_1 x + u$, where x is endogenous ($\text{Cov}(x, u) \neq 0$).

- Choose an instrumental variable, z, which is correlated with x ($\text{Cov}(x, z) \neq 0$) but not with the error term ($\text{Cov}(z, u) = 0$). If z is not exogenous, then the estimators are biased.
- Then, $\beta_1 = \text{Cov}(z, y) / \text{Cov}(x, z)$. Recall that we estimate covariance by $\sum(z_i - \text{z-bar})(y_i - \text{y-bar})$. The resulting estimator for $\beta_1^{\wedge}$ is consistent but biased.
- Homoskedasticity Assumption: $E(u^2 \mid z) = \sigma^2$.
- Standard Error: $se(\beta_1^{\wedge}) = \sqrt{(\sigma^{\wedge 2}/SST_x R_{x,z}^2)}$; this means that a small relationship between x and z makes the standard error bigger. This is another reason to use large samples.
- Functional form: Keep things linear. In particular, any variable that appears in one form in the second stage regression should appear in the same form in the first stage regression.

Two Stage Least Squares: Suppose we have exogenous variables $z_1, \ldots, z_k$ which are either instruments for x or exogenous variables in the structural equation.

- Estimate $x = \pi_0 + \pi_1 z_1 + \ldots + \pi_k z_k$. Use the predicted values for x in the original regression.
- We need x to be significantly correlated with at least one instrument that is not in the original equation. (Otherwise, x is not identified.)
- This worsens multicollinearity, because we have less variation in x.

Tests

- To test whether x is really endogenous: Find the residuals from the first stage regression. Add the residuals into the original equation (including the original potentially-endogenous variable - its original values). If they are significant, the variable really was endogenous.
- To test whether one of the instruments is correlated with the error term (overidentification): Obtain the residuals from 2SLS. Regress the residuals on

ALL exogenous variables (obtain R_1^2 from this). Under the null hypothesis that all the IV's are uncorrelated with the errors, $nR_1^2 \sim \chi_q^2$, where q = #IV's - #exogenous variables

Some other examples of using IV's:

- Measurement error: If there are two measurements, use one as the IV for the other to pull out the true value.
- Natural experiments: When randomness happens exogenously (like draft numbers).
- Mostly-randomized experiments: When people are randomly assigned to groups but might change, the original assignment is an IV for the actual assignment.

Limited Response Models

Latent Variable Models: We assume a model where the unobserved y* is a linear function of **x** and y is observed in various ways depending on the value of y*.

Binary Response: $P(y = 1 \mid \mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)$

- Logit Model: $G(z) = \exp(z)/(1 + \exp(z))$
- Probit Model: G(z) is the cumulative density function for the normal distribution.
- Tests:
    - Percent correctly predicted (for either outcome)
    - Pseudo-R^2: $1 - L_{ur}/L_0$, where L_{ur} is the likelihood ratio of the regression, and L_0 is the likelihood ratio is the regression is only on an intercept.

Tobit Model: When the observed value might be zero or some number (that is, the observed values are constrained, so some people might choose a "corner solution")

- $y^* = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u$, $u \sim N(0, \sigma^2)$, $y = \max(0, y^*)$.
- $E(y \mid \mathbf{x}) = P(y > 0 \mid \mathbf{x})\, E(y \mid y > 0, \mathbf{x})$
- This assumes that $P(y > 0 \mid \mathbf{x})$ and $E(y \mid y > 0, \mathbf{x})$ vary together. (This could be tested by estimating a probit model on whether P is zero or not. If the two models are close, this is good.)

Poisson Regression: For count data

- Model: $E(y \mid \mathbf{x}) = \exp(\beta_0 + \mathbf{x}\boldsymbol{\beta})$. Y ~ Poisson(E(y | **x**).

Censored Regression: When the highest values are assigned to one category.

- Model: $y_i^* = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u_i$. $y_i = \min(y_i, c_i)$

Truncated Regression: When some values are omitted. (Usually the highest values are left out.)

Sample Selection in General: Let s_i indicate whether y_i is observed. Then, we can estimate $s_i y_i = s_i x_i + s_i u_i$.

- Exogenous Sample Selection: When s_i is a function only of the $\mathbf{x}_i$.. In this case, OLS is unbiased and consistent
- Otherwise, OLS is biased and inconsistent.

Incidental Truncation: When $\mathbf{x}_i$ is always observed, but y_i might not be. Then, we can use the Heckit method (which combines probit to estimate s_i and then use this to correct the regression on the observed y_i's).

Asymptotic Time Series

Types of Processes:

- Stationary: The joint distribution of $(x_{t1+h}, x_{t2+h}, \ldots, x_{tm+h})$ for fixed $1 \le t_1 < \ldots < t_m$ is the same for any h.
- Covariance Stationary: $E(X_t)$ and $Var(X_t)$ is constant, and $Cov(X_t, X_{t+h})$ depends on h but not on t. (And the variance converges.)
- Weakly Dependent: $Corr(x_t, x_{t+h})$ approaches 0 as h approaches ∞. (x_t and x_{t+h} are asymptotically uncorrelated.)
  - MA(1): $x_t = e_t + \alpha e_{t-1}$, and $e_i \sim [0, \sigma_e^2]$ identically and independently
  - AR(1): $y_t = \rho y_{t-1} + e_t$, $|\rho| < 1$
- Trend Stationary Process: A process that is stationary once a time trend is removed.
  - Example: $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 t + e_t$, $|\beta_1| < 1$.
  - Removing trends from variables helps avoid spurious regressions. Top do this:
    - Regress x_t and y_t on time trends. Use the residuals from these regressions in the new regression.
    - Alternately, put a time trend in the regression.

Asymptotic Assumptions:

- The model is linear in the parameters
- $\{(\mathbf{x}_t, y_t)\}$ is weakly dependent (for LLN and asymptotic normality)
- $E(u \mid \mathbf{x}_t) = 0$
- No perfect multicollinearity
- Homoskedasticity: $Var(u_t \mid \mathbf{x}_t) = 0$
- No serial correlation: $E(u_t u_s \mid \mathbf{x}_t, \mathbf{xs}) = \mathbf{0}$

Highly Persistent Time Series

- These include random walks ($y_t = y_{t-1} + e_t = \sum e_i$), unit root processes, and random walks with drift ($y_t = \alpha_0 + y_{t-1} + e_t = t\alpha_0 + \sum e_i$).
  - I(0) = weakly dependent process
  - I(1) = process whose first difference is weakly dependent.
- These can be fixed by taking first differences or differencing the logs.

Dynamically Complete Models: A model is dynamically complete when adding more lags will not explain any more variation. (In particular, not dynamically complete model has serial correlation.)

More Time Series Stuff

Infinite Distributed Lag Model: $y_t = \alpha + \delta_0 z_t + \delta_1 z_{t-1} + \ldots + u_t$ (with infinitely many lags)

- Impact Propensity: δ_n
- Long Run Propensity: $\sum \delta_i$ (this should converge)
- One model: $\delta_j = \gamma \rho^j$, $|\rho| < 1$
- Another model: $\delta_j = \rho^{j-1}(\rho_0 \gamma_0 + \gamma_1)$

Dynamically Complete: When no more lags are needed in the model.

- This assumes there is no autocorrelation in the error term (since all of it is controlled for)
- E(u | *all lags*) = E(u | *included lags*)
- Lagged dependent variables cause bias but are still consistent

Cointegration: When $\{x_t\}$ and $\{y_t\}$ are I(1), but some linear combination, $\{y_t - \beta x_t\}$is I(0).

- Testing this: Regress $\{y_t\}$ on $\{x_t\}$. Find a t-statistic for the coefficient on x_t; use Dickey-Fuller critical values (these are bigger to correct for spurious regression).
- If $\{x_t\}$ and $\{y_t\}$ are cointegrated, then there exists an error correction model that explains the short run dynamics (Granger Representation Theorem):
    - Error correction model: $\Delta y_t = \beta_0 + \textit{other terms} + \beta_k(y_{t-1} - \alpha_0 - \alpha_1 x_{t-1}) + u_t$
    - This model suggests that when y_t strays too far from $\alpha_0 + \alpha_1 x_t$, it will correct for that.

Stationary but Serial Correlation: Lags and Leads Estimator

- Instead of estimating $y_t = \beta_0 + \beta_1 x_t + u_t$ (u_t stationary & serially correlated), augment this regression with $\Delta x_{t-1}, \ldots, \Delta x_{t-k}, \Delta x_{t+1}, \ldots, \Delta x_{t+k}$. (This cleans up serial correlation.)

Forecasting: Trying to predict future y_t from x_t and previous values.

- Evaluation: Root Mean Squared Error = $\sqrt{}(\sum(y_i - y_i^{\wedge})^2/n)$ [finding the difference between the predictions and the actual values
    - In sample evaluation: Look at the fitted values and the actual values. (This gives the model to have more information, but it is also circular.)
    - Out of sample evaluation: Estimate the model based on half the data and predict the other half; test these values.

Testing for Unit Roots: Regress Δy_t on y_t (test whether the coefficient is 0). To clean up serial correlation, add in Δy_{t-1} and other lags.

Pooled Cross Sections

Model: Distinct, independent random samples are taken in each period.

- Use a year dummy for each time period. Interactions can be used as well.

Chow Test

- SSR_U = overall residual sum of squares from regressions in each time period
- SSR_R = residual sum of squares from combining all the cross sections with either no time dummy or just a dummy but not interactions
- $SSR_U / SSR_R \sim F$

Difference-in-Differences Estimator: (to use natural experiments)

- Model: $y = \beta_0 + \beta_1 \textit{year2} + \beta_2 \textit{policy} + \beta_3 \textit{year2*policy}$
- $\beta_3 = (\mu_{\text{after, treatment}} - \mu_{\text{after, no}}) - (\mu_{\text{before, treatment}} - \mu_{\text{before, no}})$ = change between years due to treatment

Panel Data Analysis

Model: One random sample of individuals is observed for multiple time periods

- $y_{it} = \beta_0 + \beta_1 x_{it} + \ldots + \beta_k x_{ik} + a_i + u_{it}$
    - a_i is the unobservable but fixed effect for each individual
    - u_{it} is the idiosyncratic effect for an individual at a certain time

Heterogeneity Bias: The bias introduced by the correlation of a_i with $\mathbf{x}_i$. If panel data methods are not used, $a_i + u_{it}$ is correlated with $\mathbf{x}_i$.

First Difference Estimator:

- $\Delta y_t = \delta_0 + \beta_1 \Delta x_i + \ldots + \beta_k \Delta x_k + \Delta u_i$ (for two time periods)

- $\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \ldots + \alpha_T dT_t + \ldots + \beta_1 \Delta x_{it1} + \ldots + \beta_k \Delta x_{iyk} + \Delta u_{it}$
  - dk_t is 1 if t = k and 0 otherwise
- Assumptions:
  - Δx_i varies
  - $E(u_i \mid x_{i1}, x_{i2}) = E(u_i)$ [uncorrelated in both time periods!]
  - $Cov(x_{itj}, u_{is}) = 0$
  - homoskedasticity

Fixed Effects (Within) Estimator: Regress time-demeaned values of variables on each other.

- Let $x_i' = x_i$ – x-bar; this is called <u>time-demeaning</u>, and removes effects that are fixed over time (so $a_i' = 0$).
- Regress y_{it}' on $x_{it1}', \ldots, x_{itk}'$.
- Time constant variables cannot be used (but they can be interacted).

Dummy Variable Regression: Add a dummy for each individual, instead of time-demeaning. This estimates the a_i, and otherwise gives the same results as fixed effect.

Fixed Effects vs. First Differences:

- Same when T = 2.
- If the u_{it} are serially uncorrelated, use fixed effects
- If the u_{it} are positively serially correlated, use first differences
- For large T and small N, fixed effects is more sensitive to unit root processes and other violations of assumptions.
- Fixed effects handle endogeneity better.

Random Effects Estimator:

- Assumption: $Corr(x_{itj}, a_i) = 0$. (the unobserved effects are not related to the other variables)
- This means the errors have mean zero, but are serially correlated.
- This means we use GLS:
  - $\lambda^{\wedge} = 1 - \sqrt{}(1/(1 + T(\sigma_a^{\wedge 2}/\sigma^{\wedge}{}_u{}^2)))$
  - Quasi-demean: $x_{itj}' = x_{it} - \lambda^{\wedge} x_i$-bar
- Hausman Test: To test the hypothesis that random effects can be used ($Corr(a_i, \mathbf{x}) = 0$).
  - Let $W = (\beta^{\wedge}{}_{RE} - \beta^{\wedge}{}_{FE})^T Var(\beta^{\wedge}{}_{RE} - \beta^{\wedge}{}_{FE})^{-1}(\beta^{\wedge}{}_{RE} - \beta^{\wedge}{}_{FE}) = (\beta^{\wedge}{}_{RE} - \beta^{\wedge}{}_{FE})^2/\sigma_{RE\text{-}FE}{}^2$
  - Under the null hypothesis, $W \sim \chi^2{}_k$, where k is the number of estimated coefficients (except the constant).