Section 7 Model Assessment

Internal vs. external validity

- Internal validity refers to whether the analysis is valid for the population and sample being studied. External validity refers to whether these results can be generalized to other populations: is the population from which the sample is drawn representative of a larger population about which inference is sought?

External validity

- External validity is related to Assumption #0.
    - But in this case, the question is not whether all sample observations follow the same model but rather do the sample observations follow the same model as the more general population.
    - Or, alternatively, are they drawn from a sub-population that has characteristics that would make the coefficients (or specification) different?
- All populations have sub-populations that vary in their characteristics.
- If our sampling process is based on a particular sub-population, we must worry about the generalizability of our results, which is external validity:
    - Can perform an internally valid analysis of an idiosyncratic sub-population that would not generalize to others.
    - Example: Noel's work measuring the value of tree canopy or walkability in Portland. Do results generalize to other cities or do Portlanders value these characteristics more (or less) than people in other cities.
- There are no direct statistical tests for external validity (unless you have data drawn from a broader population, in which case you probably should have used it to begin with).
    - It is a usually a matter of judgment.
    - One way that some people try to assess external validity is to split the sample in half, estimate over one sample, then assess the predictions for the other sample. If predictions are good, then both halves of the sample may follow same model.
        - This is useless if both halves of the sample are drawn from a sub-population that is idiosyncratic, though.

Internal validity

- Given the population from which the sample is drawn, are the assumptions underlying the estimators valid?
- **Omitted variables**
    - They are always there.

        - Omitted variables bias the coefficient estimators for any included variables that are correlated with them.
        - In a strict sense, nearly every econometric regression is biased because of this.
    - What variables are most obviously omitted?
    - What variables in the equation would be correlated with them?
    - How does this omission bias the included coefficients?
    - Proxy variables are observable variables that are correlated with unobserved variables that should be included.
        - Proxy variables are legitimate if we are not particularly interested in the effect of the variable for which they proxy.
            - Can't interpret the coefficient on the proxy directly as the coefficient on the omitted variable.
        - OK if the difference between the true variable and the proxy is uncorrelated with included variables.
    - Panel data can help if unobserved variables vary across units but not over time or over time but not across units.
- **Misspecification of functional form**
    - Can use RESET test to explore whether quadratics are useful.
    - If you know what alternative functional forms might be more appropriate, you can test them.
- **Measurement error** (errors-in-variables bias)
    - Measurement error in dependent variable
        - Suppose that the true dependent variable is Y but that we instead observe $\tilde{Y}_i = Y_i + \varepsilon_i$, where ε_i is a random measurement error.
        - The estimated model, then is $\tilde{Y}_i = \beta_0 + \beta_1 X_i + (u_i + \varepsilon_i)$.
        - As long as the measurement error in Y (ε) is uncorrelated with X, there is no bias in the estimator of β_1. The SER will be an estimate of the standard deviation of the composite error term $u + \varepsilon$, but otherwise OLS is fine.
    - Measurement error in regressor
        - Suppose that the dependent variable is measured accurately but that we measure X with error: $\tilde{X}_i = X_i + \eta_i$.
        - The estimated model is $Y_i = \beta_0 + \beta_1 \tilde{X}_i + (u_i - \beta_1 \eta_i)$.
        - Because η is part of $\tilde{X}$ and therefore correlated with it, the composite error term is now correlated with the actual regressor, meaning that $\hat{\beta}_1$ is biased and inconsistent.

- If u and η are independent and normal, then

$$\text{plim}\,\hat{\beta}_1 = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_\eta^2}\beta_1.$$

- The estimator is biased toward zero.
- If most of the variation in $\tilde{X}$ comes from X, then the bias will be small.
- As the variance of the measurement error grows in relation to the variation in the true variable, the magnitude of the bias increases.
- As a worst-case limit, if the true X doesn't vary across our sample of observations and all of the variation in our measure $\tilde{X}$ is random noise, then the expected value of our coefficient is zero.

  - Best solution is getting a better measure.
  - Alternatives are instrumental variables or direct measurement of degree of measurement error.
    - For example, if an alternative, precise measure is available for some arguably random sub-sample of observations, then we can calculate the variance of the true variable and the variance of the measurement error and correct the estimate.

- **Sample selection bias**
  - Few samples are truly random draws from full population. Instead, they are draws (random or not) from some sub-population:
    - Many homeless are uncounted in census
    - No wage data on those who do not work
    - Polls miss people with no listed phone number
    - Cross-country regressions are often limited to the countries for which good data are available (which is not a random sample of countries)
  - If sample selection is related to X, then we have issues of external validity (do estimates apply to missed sub-population) but not internal validity. Results may be valid for the sub-population for which they are estimated.
  - If sample selection is related to Y (or, specifically, to u), then we are not drawing randomly from the population distribution of the error term (as we assume) and our results will be biased.
  - There are methods of coping with sample-selection bias.
    - Imputing values for missing wage data to allow inclusion of full sample
- **Simultaneity bias (reverse or bidirectional causality)**
  - If changes in Y (presumably due to changes in u) cause X to change, then X and u will be correlated and OLS estimates will be biased and inconsistent.
    - For example, for many years macroeconomists estimated Keynesian consumption functions by OLS: $C_t = \beta_0 + \beta_1 GDP_t + u_t$.

- (There are time-series problems with this regression that we will study later.)
- For now, note that if aggregate demand affects output, then *GDP* in each year is $C + I + G + NX$, so a positive shock to consumption (a positive u) increases *GDP*. Because the regression is correlated with the error term, OLS estimates of β_1 were biased and inconsistent. (But they looked good and had ridiculously high R^2 values, so they persisted for many years despite the protests of econometricians.)

  - The usual correction is to use an instrumental-variables (two-stage least squares) estimator.

- **Heteroskedasticity**
  - Recall that heteroskedasticity causes OLS to be inefficient (relative to WLS), but it is still unbiased and consistent.
  - The classical standard errors will be biased under heteroskedasticity, but we can use White's "robust" covariance matrix estimator, which we've talked about earlier.
  - Using robust errors is the most common correction for heteroskedasticity.
- **Autocorrelation**
  - If error terms of different observations are correlated, then OLS is also inefficient (relative to a corrected GLS estimator), but is unbiased and consistent.
    - Autocorrelation can be spatial: Unmeasured neighborhood characteristics (omitted variables) that cause houses that are close together to be more or less valuable.
    - Autocorrelation is ubiquitous in time-series data: This period's error term is nearly always related to last period's. (Unmeasured omitted variables are themselves correlated over time.)
  - Again, standard errors are biased, but White's heteroskedastic-consistent standard errors don't help here.
  - There are estimated standard errors that are robust to autocorrelation. (Use "hac" option in Stata.)
  - Alternatively, one can try to model the autocorrelation and transform the model into one that has no autocorrelation (GLS).
    - Examples include AR(1) models in time series and modeling spatially correlated errors in cross-section models.

Validity in forecasting/prediction

- Regression models may be valid for forecasting even if their coefficients are not unbiased or consistent.
  - Suppose that we know that X is measured with error.

- We can still use a regression of Y on $\tilde{X}$ to predict the outcome of a particular measured $\tilde{X}$ even though the estimated coefficient is a biased estimator for the effect of X. That is because we have correctly estimated the relationship between the noisy $\tilde{X}$ and Y.
    - We would not get reliable estimates if our prediction question relied on the true X rather than the noisy $\tilde{X}$.

- We often build models with noisy data or proxy variables to get predictions of another variable.
- The biggest question in forecasting is external validity: does the model that applies to the sample you used for estimation also apply to the observation for which you want a forecast?
- Measuring prediction error: What is the variance of $\hat{Y}$?
    - $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

      $Y = \beta_0 + \beta_1 X + u$

      $Y - \hat{Y} = \beta_0 - \hat{\beta}_0 + \left(\beta_1 - \hat{\beta}_1\right) X + u$

      $\text{var}\left(\hat{Y}\right) = E\left(Y - \hat{Y}\right)^2 = \text{var}\left(\hat{\beta}_0\right) + X^2 \,\text{var}\left(\hat{\beta}_1\right) + 2X \,\text{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) + \text{var}(u).$
    - For simple regression under homoskedasticity,

$$\text{cov}\left(\hat{\beta}\right) = \sigma_u^2 \left(X'X\right)^{-1} = \sigma_u^2 \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix}^{-1}$$

$$= \frac{\sigma_u^2}{n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2} \begin{bmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{bmatrix}$$

$$= \frac{\sigma_u^2}{n\sum_{i=1}^n \left(X_i - \bar{X}\right)^2} \begin{bmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{bmatrix}.$$

- So

$$\operatorname{var}\left(\hat{Y}\right)=\sigma_u^2\left[1+\frac{nX^2-2Xn\bar{X}+\sum_{i=1}^{n}X_i^2}{n\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\right]$$

$$=\sigma_u^2\left[1+\frac{\sum_{i=1}^{n}X_i^2-n\bar{X}^2+nX^2-2Xn\bar{X}+n\bar{X}^2}{n\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\right]$$

$$=\sigma_u^2\left[1+\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2+n\left(X-\bar{X}\right)^2}{n\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\right]$$

$$=\sigma_u^2\left[1+\frac{1}{n}+\frac{\left(X-\bar{X}\right)^2}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\right].$$

- Prediction error is smaller for:
  - Smaller error variance
  - Larger sample size (through both second and third terms)
  - Greater sample variation in X
  - Observations closer (X) to the mean