639 Midterm Reference Sheet

francisco brady

Omitted Variable Bias

OBV occurs when two criteria are met:
* the omitted variable is correlated with the included regressor
* the omitted variable is a determinant of the dependent variable (outcome)

Note: Omitted variable bias means that the first least squares assumption for causal inference &emdash; that $E(u_i|X_i) = 0$, **does not hold**, resulting in a biased estimator! This cannot be fixed with large samples!
Use this table:
* α_1: simple model regressor
* β_1: regressor after adding second variable

		How is X_2 related to Y?		
$Corr(X_1, X_2)$		$\beta_2 < 0$	$\beta_2 = 0$	$\beta_2 > 0$
Or: How is X_1	$\gamma < 0$	Positive Bias ($\alpha_1 > \beta_1$)	No Bias	Negative Bias ($\alpha_1 < \beta_1$)
related to X_2	$\gamma = 0$	No Bias	No Bias	No Bias
	$\gamma > 0$	Negative Bias ($\alpha_1 < \beta_1$)	No Bias	Positive Bias ($\alpha_1 > \beta_1$)

R^2 is the ratio of the sample variance of $\hat{Y}$ to the sample variance of Y.

$$R^2 = \frac{\text{Explained Sum of Squares (ESS)}}{\text{Total Sum of Squares (TSS)}} = 1 - \frac{\text{Sum of Squared Residuals (SSR)}}{\text{Total Sum of Squares (TSS)}}$$

Where $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$, $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, and $SSR = \sum_{i=1}^{n} {\hat{u}_i}^2$

Special case:

If $\hat{\beta}_1 = 0$, then X_i explains *none* of the variation of Y_i, and the predicted value of Y_i is $\hat{Y}_i = \hat{\beta}_0 = \bar{Y}$. **In this case, the explained sum of squares is 0 and the sum of squared residuals equals the total sum of squares; thus the R^2 is 0.**

Adjusted R^2

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \cdot \frac{SSR}{TSS} = 1 - \frac{s^2_{\hat{u}}}{s^2_Y}$$

Where k is the number of variables in the model. The adjustment accounts for the mechanical increase in R^2 from additional regressors.

OLS Assumptions:

Assumption 1: The conditional distribution of u_i given X_i has a mean of 0. The formal statement is: $E(u_i|X_i) = 0$, which implies that X_i and u_i are **uncorrelated**.
Assumption 2: $(X_i, Y_i), i = 1, ..., n$ are independently and identically distributed (i.i.d). If X_i and Y_i are drawn from the same population, they will have the same distribution, and if they are drawn randomly, then the selection of any X_i or Y_i into the sample should be independent.
Assumption 3: Large outliers are unlikely. Large outliers can make the model results misleading!

Homoskedasticity and Heteroskedasticity

Heteroskedasticity: The variance of the error term (u) is **not** constant across the sample. **This affects the standard errors of the model.**
Homoskedasticity: The variance of the error term (u) is constant across the sample, this implies that $var(u|X = x)$ does *not* depend on the value of X, i.e. errors are uncorrelated.
Impact: Heteroskedastic errors affect: standard errors, which affect t-statistics, p-values, and confidence intervals. **Effect on coefficients:** Because the least squares assumption places no restrictions on the conditional variance, they apply to both the general case of heteroskedasticity and the special case of homoskedasticity. **Therefore, the OLS estimators remain unbiased and consistent even if the errors are homoskedastic.**

tldr: use (heteroskedasticity-) robust standard errors in your models

---

The Population and Sample Regression

Population Regression line (no hats!)

$$Y_i = \beta_0 + \beta_1 X_1 i + \beta_2 X_2 i + \ ... \ \beta_k 1 X_k i + u_i, i = 1, \ ... \ , n$$

Sample Regression/OLS Regression line:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i, \ \text{i} = 1, \ ... \ , n$$

Causal Inference and Experimental Design

Fundamental problem of Causal Inference: We **cannot** observe two outcomes of a variable in the **same** unit/individual at the **same** time. **This is addressed (solved?) by randomization.**

The Ideal Experiment
- Create two groups that would be identical in the absence of treatment
- Apply treatment to one group, observe the difference in outcomes
- An RCT enforces **no** correlation between treatment and other factors

Internal validity - Did we estimate an unbiased causal effect for our sample?
- Fails when treatment and control groups are different in ways (beside the treatment) that may affect the outcome of interest - Non-compliance, attrition, evaluation-driven effects **External validity** - Can we extrapolate these estimates to other populations? To whom can we generalize? - Fails when the treatment effect is different outside the evaluation

---

Indicator Variables

Interpret regression results by multiplying all the category coefficients by 0. That is the estimate for the reference group, and all coefficients are interpreted relative to the reference group. To test, use the F-test. The null for the F-test is that all of the coefficients are zero.