Material for 8070 Review Quiz

Properties of Estimators

Sampling Distribution The distribution of an estimator with respect to a repeated sampling thought experiment where we imagine repeatedly drawing new samples of size n from the underlying population and re-computing the estimator for each new sample.

Bias The difference between the expected value of an estimator and its actual value, i.e., $\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$, where the expectation is with respect to the sampling distribution of the estimator. An estimator is said to be *unbiased* if $\text{Bias}(\hat{\theta}) = 0$.

Sampling Variance The variance of an estimator with respect to its sampling distribution, i.e., $\text{Var}(\hat{\theta})$.

In general, we prefer estimators with low (or 0) bias and low sampling variance.

Consistency In large samples, if $\hat{\theta}$ is consistent, then it is guaranteed to be close to θ if we have enough data, i.e., $\hat{\theta} \xrightarrow{p} \theta$.

Asymptotic Normality In large samples, a centered and scaled version of our estimator will follow a normal distribution, typically, $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V)$. Alternatively, $\hat{\theta} \overset{a}{\sim} \mathcal{N}(\theta, V/n)$ (i.e., $\hat{\theta}$ approximately follows a normal distribution with mean θ and variance V/n).

Inference

Null Hypothesis The hypothesis to be tested, e.g., $H_0 : \theta = \theta_0$, where θ_0 is a researcher-specified specific value.

Alternative Hypothesis Competing hypothesis to the null, most commonly, $H_1 : \theta \neq \theta_0$ (two-sided)

Type I Error To reject H_0 when it is true.

Type II Error To Fail to reject H_0 when H_1 is true.

Significance Level Researcher-specified willingness to make Type I errors, usually denoted by α, where the most common value is $\alpha = 0.05$

Size The rate at which a particular test makes Type I errors (ideally, size = significance level).

Power The probability of correctly rejecting H_0 when H_1 is true.

Standard Error An estimate of the standard deviation of an estimator with respect to its sampling distribution. If $\hat{\theta}$ is asymptotically normal with asymptotic variance V, then the standard error is $\text{se}(\hat{\theta}) = \sqrt{\hat{V}/n}$, where $\hat{V}$ is an estimate of V.

Common test-statistics:

- **t-statistic** If θ is scalar, then $t = \dfrac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})}$.
- **Wald statistic** - If θ is a k-dimensional vector, then $W = n(\hat{\theta} - \theta_0)'\hat{V}^{-1}(\hat{\theta} - \theta_0)$, where $\hat{V}$ is an estimate of the asymptotic variance matrix of $\hat{\theta}$.

Behavior of t-statistic

- Under H_0,

$$t = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{\hat{V}}} = \frac{1}{\sqrt{V}}\sqrt{n}(\hat{\theta} - \theta_0) + o_p(1) \xrightarrow{d} \frac{1}{\sqrt{V}}\mathcal{N}(0, V) = \mathcal{N}(0, 1)$$

.

- Under H_1,

$$t = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{\hat{V}}} = \frac{1}{\sqrt{V}}\sqrt{n}(\hat{\theta} - \theta_0) + o_p(1) \text{ diverges because } \hat{\theta} \xrightarrow{p} \theta_1 \neq \theta_0.$$

i.e., under H_0, t should behave like a draw from $\mathcal{N}(0, 1)$; under H_1, it should diverge. This strongly differing behavior provides the basis for hypothesis testing, where we typically use the decision rule: if $|t| > c_{1-\alpha/2} = 1.96$, reject H_0, otherwise fail to reject.

Behavior of Wald statistic

- Under H_0, $W = \sqrt{n}(\hat{\theta} - \theta_0)'\hat{V}^{-1}\sqrt{n}(\hat{\theta} - \theta_0) = Z'V^{-1}Z + o_p(1) \xrightarrow{d} \chi^2_k$, where $Z \sim \mathcal{N}(0, V)$.
- Under H_1, W diverges because $\hat{\theta} \xrightarrow{p} \theta_1 \neq \theta_0$.

Confidence Interval The set of values of θ that are "compatible" with the observed data. That is, a $100(1-\alpha)\%$ confidence interval is the set of values of θ that would not be rejected by a two-sided test at significance level α. If θ is scalar, this CI is given by

$$\left[\hat{\theta} - c_{1-\alpha/2} \times \text{se}(\hat{\theta}),\ \hat{\theta} + c_{1-\alpha/2} \times \text{se}(\hat{\theta})\right]$$

where $c_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of $\mathcal{N}(0, 1)$ (e.g., $c_{1-\alpha/2} = 1.96$).

p-value The probability of getting an estimate (or, equivalently, a test-statistic) as extreme as the one we got if the null hypothesis were true. $p = 2\,(1 - \Phi(|t|))$ where $\Phi(\cdot)$ is the cdf of $\mathcal{N}(0, 1)$.

Asymptotic Tools

Law of Large Numbers If $\{Y_i\}_{i=1}^n$ is an iid sample with $\mathbb{E}\big[|Y|\big] < \infty$, then $\bar{Y} \xrightarrow{p} \mathbb{E}[Y]$.

Central Limit Theorem If $\{Y_i\}_{i=1}^n$ is an iid sample with $\mathbb{E}[Y^2] < \infty$, then $\sqrt{n}(\bar{Y} - \mathbb{E}[Y]) \xrightarrow{d} \mathcal{N}(0, \text{Var}(Y))$.

Continuous Mapping Theorem If $X_n \xrightarrow{p} c$ and $g(\cdot)$ is continuous at c, then $g(X_n) \xrightarrow{p} g(c)$. Similarly, if $X_n \xrightarrow{d} X$ and $g(\cdot)$ is continuous, then $g(X_n) \xrightarrow{d} g(X)$.

Slutsky's Theorem Collects the most common uses of the CMT. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then

- $X_n + Y_n \xrightarrow{d} X + c$
- $X_n Y_n \xrightarrow{d} cX$
- $X_n / Y_n \xrightarrow{d} X/c$, provided $c \neq 0$.

Estimating $\mathbb{E}[\mathbf{Y}]$

The natural estimator of $\mathbb{E}[Y]$ is $\hat{\mu} := \dfrac{1}{n}\displaystyle\sum_{i=1}^n Y_i$.

Assumptions

(1) iid sample: $\{Y_i\}_{i=1}^n$ is independent and identically distributed.
(2a) existence of moments: $\mathbb{E}[Y^2] < \infty$
(2b) existence of moments: $\mathbb{E}[Y^4] < \infty$

Bias

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n Y_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i] = \mathbb{E}[Y]$$

where the first equality holds by the definition of $\hat{\mu}$, the second equality holds by the linearity of expectations, and the third equality holds if we have an iid sample (particularly, identically distributed). This implies $\text{Bias}(\hat{\mu}) = 0$, i.e., $\hat{\mu}$ is unbiased.

Sampling Variance

$$\text{Var}(\hat{\mu}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^n Y_i\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^n Y_i\right) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}(Y_i) = \frac{\text{Var}(Y)}{n}$$

Thus, $\text{Var}(\hat{\mu})$ depends on $\text{Var}(Y)$ and is shrinking in n.

Consistency

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n Y_i \xrightarrow{p} \mathbb{E}[Y]$$

which holds by the weak law of large numbers. Invoking the weak law of large numbers requires Assumption 1 and Assumption 2a.

Asymptotic Normality

$$\sqrt{n}(\hat{\mu} - \mathbb{E}[Y]) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Y_i - \mathbb{E}[Y]\right) \xrightarrow{d} \mathcal{N}(0, V)$$

where $V = \text{Var}(Y)$. This holds by the central limit theorem, which requires Assumption 1 and Assumption 2b.

The above discussion applied in the case where Y is scalar. If Y is a k-dimensional vector, all of the results above hold, but now $\text{Var}(Y)$ is a $k \times k$ variance matrix, with diagonal elements equal to the variance of each component of Y, and off diagonal elements equal to the covariance between the components of Y.

Conditional Expectations

Conditional Expectation Function (CEF) $m(x) := \mathbb{E}[Y \mid X = x]$ is a function of X that gives the expected value of Y given $X = x$.

Law of Iterated Expectations

$$\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y|X]\big] \quad \text{or} \quad \mathbb{E}[Y|X_1] = \mathbb{E}\big[\mathbb{E}[Y|X_1, X_2]\big|X_1\big]$$

CEF Error $e := Y - m(X)$. $\mathbb{E}[e \mid X] = 0$ by construction.

CEF as Best Predictor Among all prediction functions $g(X)$, the CEF minimizes mean squared error, i.e.,

$$\mathbb{E}\big[(Y - g(X))^2\big] \text{ is minimized when } g(X) = m(X)$$

This provides a strong justification for using the CEF to make predictions of Y.

Regression Derivative Often, in economics, we are interested in derivatives of the CEF, such as

$$\nabla_1 m(x) := \begin{cases} \frac{\partial m(x)}{\partial x_1} & \text{if } x_1 \text{ is continuous} \\ \mathbb{E}[Y \mid X_1 = x_1 + 1, X_{-1} = x_{-1}] \\ \quad - \mathbb{E}[Y \mid X_1 = x_1, X_{-1} = x_{-1}] & \text{if } x_1 \text{ is discrete} \end{cases}$$

- $\nabla_1 m(x)$ should be interpreted as how much the average outcome increases for a 1 unit increase in X_1, holding other regressors constant. For example, if X_1 is a person's years of education and Y is earnings, $\nabla_1 m(x)$ is how much higher average earnings are for people with one more year of education holding other regressors constant. Importantly, this is not generally equal to how much an extra year of education *causes* education to increase; rather, you should think of this descriptively, i.e., as just a statement of fact that, on average, people with one more year of education earned some amount more than those with one year less holding other regressors constant.
- Regression derivatives are sometimes called *marginal contrasts*, *marginal effects*, or *partial effects*
- $\nabla_1 m(x)$ holds other regressors constant, but not "all else" constant
- $\nabla_1 m(x)$ is a function of x and can change for different values of the regressors. For this reason, it is common to average them into a single number $AMC = \mathbb{E}[\nabla_1 m(X)]$

Interpret Linear CEFs/Regressions Be able to interpret any linear regression model in terms of predicted values or regression derivatives, including logarithms and interaction terms. Example: $\log(Wage) = \beta_1 Educ + \beta_2 Educ \times Female + \beta_3 Female + \beta_4 Exper + \beta_5 + e$ with $\mathbb{E}[e|Educ, Female, Exper] = 0$.

Linear Projection Model

Often, we do not know that the CEF is linear, however, that does not stop us from running a regression. This idea gives rise to the linear projection model:

$$Y = X'\beta + e, \text{ where } \beta := \arg\min_b \mathbb{E}\big[(Y - X'b)^2\big]$$

This amounts to choosing the best prediction function among the class of linear prediction functions. This problem can be solved to give the formula $\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$. The projection error $e = Y - X'\beta$ satisfies $\mathbb{E}[Xe] = 0$. $X'\beta$ is called the best linear predictor.

Estimating β

The analogy principle immediately suggests an estimator for β:

$$\hat\beta = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n X_i Y_i = \big(\mathbf{X}'\mathbf{X}\big)^{-1}\mathbf{X}'\mathbf{Y}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are $n \times k$ and $n \times 1$ data matrices.

Assumptions

(1) iid sample: $\{X_i, Y_i\}_{i=1}^n$ are iid
(2a) existence of moments: $\mathbb{E}[Y^2] < \infty$ and $\mathbb{E}\big[||X||^2\big] < \infty$
(2b) existence of moments: $\mathbb{E}[Y^4] < \infty$ and $\mathbb{E}\big[||X||^4\big] < \infty$
(3) no full multi-collinearity: $\mathbb{E}[XX']$ is invertible

The discussion below about bias and sampling variance holds under the linear CEF model, the discussion about consistency and asymptotic normality holds for both the linear CEF and linear projection models.

Bias

$$\mathbb{E}[\hat\beta] = \mathbb{E}\big[\mathbb{E}[\hat\beta \mid \mathbf{X}]\big] = \mathbb{E}\big[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{Y} \mid \mathbf{X}]\big] = \mathbb{E}\big[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta\big] = \beta$$

and, hence, $\hat\beta$ is unbiased for β.

Sampling Variance

$$\mathrm{Var}(\hat\beta) = \mathrm{Var}\big(\mathbb{E}[\hat\beta \mid \mathbf{X}]\big) + \mathbb{E}\big[\mathrm{Var}(\hat\beta \mid \mathbf{X})\big] = \mathbb{E}\big[\mathrm{Var}(\hat\beta \mid \mathbf{X})\big]$$

and

$$\mathrm{Var}(\hat\beta \mid \mathbf{X}) = \big(\mathbf{X}'\mathbf{X}\big)^{-1}\mathbf{X}'\mathrm{Var}(\mathbf{e} \mid \mathbf{X})\mathbf{X}\big(\mathbf{X}'\mathbf{X}\big)^{-1}$$

Homoskedasticity: $\mathrm{Var}(\mathbf{e} \mid \mathbf{X}) = \sigma^2\mathbf{I}$, which implies that

$$\mathrm{Var}(\hat\beta \mid \mathbf{X}) = \sigma^2\big(\mathbf{X}'\mathbf{X}\big)^{-1}$$

Gauss-Markov Theorem: Under homoskedasticity, $\mathrm{Var}(\hat\beta \mid \mathbf{X}) \leq \mathrm{Var}(\tilde\beta \mid \mathbf{X})$ where $\tilde\beta$ is any other linear unbiased estimator of β.

Consistency

$$\frac{1}{n}\sum_{i=1}^n X_i X_i' \xrightarrow{p} \mathbb{E}[XX']$$

$$\frac{1}{n}\sum_{i=1}^n X_i Y_i \xrightarrow{p} \mathbb{E}[XY]$$

which holds by the law of large numbers under Assumptions (1) and (2a). Then, $\hat\beta \xrightarrow{p} \beta$ by the continuous mapping theorem, which applies under Assumption 3.

Asymptotic Normality

$$\sqrt{n}(\hat\beta - \beta) = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i e_i$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i e_i \xrightarrow{d} \mathcal{N}(0, \mathbf{\Omega}) \text{ where } \mathbf{\Omega} = \mathbb{E}[XX'e^2]$$

which holds by the central limit theorem under Assumptions 1 and 2b, and

$$\sqrt{n}(\hat\beta - \beta) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}) \text{ where } \mathbf{V} = \mathbb{E}[XX']^{-1}\mathbf{\Omega}\mathbb{E}[XX']^{-1}$$

which holds by the continuous mapping theorem, which applies under Assumption 3.

Omitted Variable Bias

Consider the following two linear projections, a long regression and short regression

$$Y = X_1'\beta_1 + X_2'\beta_2 + e \text{ and } Y = X_1'\gamma_1 + u$$

We are interested in β_1 but suppose the first regression is infeasible (e.g., X_2 is unobserved) so we run the second regression instead. Then,

$$\begin{aligned} \gamma_1 &= \mathbb{E}[X_1X_1']^{-1}\mathbb{E}[X_1Y] = \mathbb{E}[X_1X_1']^{-1}\mathbb{E}[X_1(X_1'\beta_1 + X_2'\beta_2 + e)] \\ &= \beta_1 + \mathbb{E}[X_1X_1']^{-1}\mathbb{E}[X_1X_2']\beta_2 = \beta_1 + \Gamma_{21}'\beta_2 \end{aligned}$$

where Γ_{21} is a $k_2 \times k_1$ matrix that contains the coefficients from running a regression of each element of X_2 on X_1. Notice that $\gamma_1 = \beta_1$ if either (i) $\Gamma_{21} = 0$ (i.e., if you run a regression of X_2 on X_1, the coefficients are all equal to 0) or (ii) $\beta_2 = 0$ (i.e., the coefficient on X_2 in the long regression is equal to 0).

Frisch-Waugh-Lovell Theorem

Suppose we are interested in β_1 from the following long linear projection

$$Y = X_1'\beta_1 + X_2'\beta_2 + e$$

and consider the following auxiliary linear projections

$$Y = X_2'\gamma_2 + u \text{ and } X_1 = \Lambda_{12}X_2 + v$$

where Λ_{12} is a $k_1 \times k_2$ matrix of coefficients from projecting each element of X_1 on X_2. Then, consider the linear projection of u on v:

$$\begin{aligned} \mathbb{E}[vv']^{-1}\mathbb{E}[vu] &= \mathbb{E}[vv']^{-1}\mathbb{E}[v(Y - X_2'\gamma_2)] = \mathbb{E}[vv']^{-1}\mathbb{E}[vY] \\ &= \mathbb{E}[vv']^{-1}\mathbb{E}[v(X_1'\beta_1 + X_2'\beta_2 + e)] = \beta_1 \end{aligned}$$

This says that β_1 from the long linear projection can equivalently be obtained from the following steps: (i) project Y on X_2 and recover the projection error u, (ii) project X_1 on X_2 and recover the projection error v, (iii) run the regression of u on v and recover the coefficients on v. This three-step procedure provides a rationalization for a "partialling out" interpretation of β_1 (i.e., that it captures the relationship between Y and X_1 after "removing" the relationship between these variables and X_2). This version of FWL holds for the population quantity β_1, but similar arguments can be used for the sample analogue.

Special cases:
(1) $k_1 = 1$ (i.e., X_1 includes a single regressor), then $\beta_1 = \frac{\mathbb{E}[vu]}{\mathbb{E}[v^2]} = \frac{\mathbb{E}[vY]}{\mathbb{E}[v^2]}$

(2) $X_2 = 1$ (i.e., X_2 includes only an intercept), then β_1 is equivalent to running a regression using a de-meaned outcome $(Y - \mathbb{E}[Y])$ on de-meaned regressors $(X - \mathbb{E}[X])$.