Linear Models Lecture 10: Asymptotics

Robert Gulotty

University of Chicago

February 26, 2026

Big Picture

- We began with a model in which $Y = \boldsymbol{x}'\beta + e$
- OLS finds the projection of $\boldsymbol{y}$ in the vector space spanned by $\boldsymbol{x}$:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}$$
$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta} = \boldsymbol{y} - \boldsymbol{P}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y}$$

- We showed that the projection coefficient $\hat{\beta}$ is an unbiased estimator of β.
- With the additional assumption of normality, we derived t and F distributions (Lecture 7).

In Lecture 9, we used asymptotic arguments **informally**: Taylor expansion of the score, LLN for the Hessian, CLT for the score. Today we develop these tools rigorously.

Standards for Inference

- Unbiased estimators ($E[\hat{\theta}] = \theta$) are "correct on average" even in small samples.
  - Unbiasedness tells us that we are not better off with a different guess (we would lose money in expectation).
  - A biased estimator is biased even with infinite data.
- Unbiasedness is not always what we want.
  - Unbiasedness doesn't rule out pathological behavior: the "first impression" estimator is unbiased.
  - Many popular estimators are biased: Probit MLE (Lecture 9), 2SLS, ridge regression.
- A biased estimator may be a good approximation if it gets **closer to the truth with more data**.
- What is a "good" approximation?

Why Consistency Matters: A Concrete Example

U.S. hourly wages: $\mu = 24$, $\sigma^2 = 430$.

"First Impression" estimator

$\hat{\mu} = X_1$ (use only the first observation)

- Unbiased: $\mathbb{E}[X_1] = \mu = 24$
- $\text{Var}(\hat{\mu}) = 430$ for *any n*
- 95% interval: 24 ± 40.7

Sample mean $\bar{X}_n$

- Unbiased: $\mathbb{E}[\bar{X}_n] = \mu = 24$
- $\text{Var}(\bar{X}_n) = 430/n \to 0$
- 95% interval at $n = 100$: 24 ± 4.1

Both estimators are unbiased. Only $\bar{X}_n$ is **consistent**: its distribution collapses onto μ. When forced to choose, **consistency** $>$ **unbiasedness**.

It has been customary to assume (somewhat loosely) that when a quantity is calculated from a random sample to estimate a parameter of a hypothetical frequency distribution, the accuracy of the determination will increase without limit as the number in the sample increases. A *consistent statistic* is a function of the observations actually possessing this property.

- Hotelling 1930

Consistency

- Consistent estimators are "more likely correct with more data".
- The analog to expectations in large samples is called the probability limit: plim.
- Claim of Consistency: plim$\hat{\beta} = \beta$, or $\hat{\beta}_j$ is *consistent* for β_j

Convergence of Random Variables

Consider the sample mean $\bar{X}_n$ as a sequence indexed by sample size n. Its distribution changes with n—how do we describe its limit?

There is a hierarchy of convergence concepts for random variables:

$$X_n \xrightarrow{\text{a.s.}} X \quad \Rightarrow \quad X_n \xrightarrow{p} X \quad \Rightarrow \quad X_n \xrightarrow{d} X$$

- **Almost sure (a.s.) convergence** is the strongest form—pathwise convergence except on a probability-zero set. We won't need it operationally.
- **Convergence in probability** ($\xrightarrow{p}$): for all $\varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$.
  - This is the concept behind **consistency** (WLLN, plim, CMT).
- **Convergence in distribution** ($\xrightarrow{d}$): $F_{X_n}(t) \to F_X(t)$ at all continuity points.
  - This is the concept behind the **CLT** and asymptotic normality.

Today we develop each in turn: $\xrightarrow{p}$ (next), then $\xrightarrow{d}$ (CLT section).

Convergence in Probability

A sequence of random variables X_n converges *in probability* to c if, for all $\delta > 0$,

$$\lim_{n\to\infty} Pr(|X_n - c| \leq \delta) = 1$$

This is written as $X_n \underset{p}{\rightarrow} c$ or $\text{plim} X_n = c$

We can also say

$$\lim_{n\to\infty} Pr(|X_n - c| > \delta) = 0$$

That is, the random variable concentrates about a point in a certain interval.

Plim example

- Define a random variable Z_n that takes on values 0 and a_n with probability $1 - p_n$ and p_n respectively,
- Intuitively Z_n should converge to 0 if either $p_n \to 0$ or $a_n \to 0$.
- Consider any $\delta > 0$. If $a_n \to 0$, then for n large enough, $a_n < \delta$. If so $Pr(|Z_n| \leq \delta) = 1$.
- If $p_n \to 0$ then $1 - p_n \to 1$.
- $1 - p_n$ of the time $Z_n = 0 < \delta$, so $Pr(|Z_n| \leq \delta) \geq 1 - p_n$
- Therefore $Pr(|Z_n| \leq \delta) \to 1$ as $p_n \to 0$

Consistency in Practice: OLS with Increasing n

DGP: $Y_i = 2 + 3\,X_i + \varepsilon_i, \quad X_i \sim N(0,1), \quad \varepsilon_i \sim \text{Exp}(1) - 1$ (skewed, non-normal errors).

```
set.seed(42)
par(mfrow = c(2, 2))
for (n in c(20, 50, 200, 2000)) {
  betas <- replicate(5000, {
    x <- rnorm(n); y <- 2 + 3*x + (rexp(n)-1)
    coef(lm(y ~ x))[2]
  })
  hist(betas, breaks = 40, prob = TRUE,
       main = paste("n =", n), xlim = c(2, 4),
       xlab = expression(hat(beta)[1]))
  abline(v = 3, col = "blue", lwd = 2)
}
```

Errors are **skewed and non-normal**, yet $\hat{\beta}_1$ concentrates around the true value $\beta_1 = 3$ as n grows. This is consistency in action.

Why Useful : The continuous mapping theorem

- Unlike expectations, plim survives continuous functions (like division!).

The continuity theorem, the first of Slutsky's theorems, the continuous mapping theorem.
Given a random variable x_n that converges in probability to x and a continuous function $g(x_n)$, we have that:

$$\text{plim} g(x_n) = g(\text{plim} x_n)$$

If $\text{plim} x_n = c$ and $\text{plim} y_n = d$, then

$$\text{plim}(x_n y_n) = (\text{plim} x_n)(\text{plim} y_n) = cd$$
$$\text{plim}(x_n / y_n) = (\text{plim} x_n)/(\text{plim} y_n) = c/d$$
$$\text{plim}(x_n + y_n) = (\text{plim} x_n) + (\text{plim} y_n) = c + d$$

CMT in Action: Why Division Works

The t-statistic divides one random quantity by another. Why is this valid?

- WLLN: $\bar{X}_n \xrightarrow{p} \mu$ and $s_n^2 \xrightarrow{p} \sigma^2$
- CMT (continuous function $g(a, b) = a/\sqrt{b}$):

$$\frac{\bar{X}_n}{s_n} \xrightarrow{p} \frac{\mu}{\sigma}$$

- More generally, any smooth function of consistent estimators is itself consistent.

> Every time you divide an estimate by its standard error, you rely on CMT. Without it, ratios of random variables have no guarantee of converging to the right thing.

Weak Law of Large Numbers

Proof plan:

- Markov's Inequality
- Chebyshev's Inequality
- Weak Law of Large Numbers

Markov's Inequality

Proposition

Suppose X is a random variable that takes on non-negative values. Then, for all $a > 0$,

$$P(X \geq a) \leq \frac{E[X]}{a}$$

Markov's Inequality

Proof.

For $a > 0$,

$$
\begin{aligned}
E[X] &= \int_0^\infty xf(x)dx \\
&= \int_0^a xf(x)dx + \int_a^\infty xf(x)dx \\
E[X] &\geq \int_a^\infty xf(x)dx \geq \int_a^\infty af(x)dx = aP(X \geq a) \qquad (X \geq 0) \\
\frac{E[X]}{a} &\geq P(X \geq a)
\end{aligned}
$$

$\square$

Chebyshev's Inequality

Proposition

If X is a random variable with mean μ and variance σ^2, then, for any value $k > 0$,

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

Chebyshev's Inequality

Proof.

Define the random variable

$$Y = (X - \mu)^2$$

Where $\mu = E[X]$.
Then we know Y is a non-negative random variable. Set $a = k^2$ in Markov's inequality

$$P(Y \geq k^2) \leq \frac{E[Y]}{k^2}$$

$$P((X - \mu)^2 \geq k^2) \leq \frac{E[(X-\mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$$

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

Example of Markov and Chebyshev

Suppose that we know that the number of items produced in a factory is a random variable with mean 500.

- How likely are we to see a production of 1000?
- If the variance is 50, how likely is the production to be between 400 and 600?

Solution

- **How likely are we to see a production of 1000?**
  Markov's Inequality:

$$\mathbb{P}(X \geq 1000) \leq \frac{\mathbb{E}[X]}{1000} = \frac{500}{1000} = 0.5$$

- **How likely is production between 400 and 600, given variance = 50?**
  Chebyshev's Inequality:

$$\mathbb{P}(|X - 500| \geq 100) \leq \frac{50}{100^2} = \frac{1}{200}$$

$$\mathbb{P}(400 \leq X \leq 600) = \mathbb{P}(|X - 500| < 100) \geq 1 - \frac{1}{200} = \frac{199}{200}$$

Chebyshev Sample Size Calculation

- U.S. hourly wages are distributed $\mu = 24$ and $\sigma^2 = 430$,
- How large does n have to be so that $\bar{X}_n$ is within \$1 of the true mean with probability 99%?

$$\mathbb{P}\left[|\bar{X}_n - \mu| \geq 1\right] = \mathbb{P}\left[|\bar{X}_n - \mathbb{E}\left[\bar{X}_n\right]| \geq 1\right] \leq var\left[\bar{X}_n\right] = 430/n$$

$$\frac{430}{n} \leq .01$$

$$\mathbb{P}\left[|\bar{X}_n - \mu| \geq 1\right] \leq .01 \text{ if } n = 43{,}000$$

Weak Law of Large Numbers

Proposition

Suppose $X_1, X_2, \ldots, X_n$ is iid random sample from a distribution with expectation μ and $Var(X_i) = \sigma^2$. Then, for all $\varepsilon > 0$,

$$P\left\{\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu\right| \geq \varepsilon\right\} \to 0 \text{ as } n \to \infty$$

or

$$\bar{X}_n \underset{p}{\to} \mathbb{E}[X]$$

Weak Law of Large Numbers

Proof.

$$\frac{\mathrm{E}[X_1 + X_2 + \cdots + X_n]}{n} = \frac{\sum_{i=1}^{n} \mathrm{E}[X_i]}{n} = \mu$$

$$\mathrm{E}\left[\left(\frac{\sum_{i=1}^{n} X_i - \mu}{n}\right)^2\right] = \frac{\mathrm{Var}(X_1 + X_2 + \cdots + X_n)}{n^2} = \frac{\sum_{i=1}^{n} \mathrm{Var}(X_i)}{n^2} = \frac{\sigma^2}{n}$$

$$\mathbb{P}\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0 \quad \text{as } n \to \infty$$

□

WLLN interpretation

- The sample mean approaches the population mean as $n \to \infty$
- General version only assumes $\mathbb{E}[X] < \infty$
- This notion that a sequence of the same sample quantity approaches a constant is called consistency.

Immediate Consequence

Any sample moment converges to its population counterpart:

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

This is the foundation for OLS consistency (Lecture 11) and for the Probit Hessian convergence (Lecture 9).

WLLN Counterexample

- WLLN needs to have independent observations:
- For example, assume $X_i = Z + U_i$, like in the random effects model.
- Suppose Z and U_i are random and independent, $\mathbb{E}(U_i) = 0$.
- $\bar{X}_n = Z + \bar{U}_n \underset{p}{\rightarrow} Z$ not a constant.

Convergence in Distribution

- We want the sampling distribution of the sample mean $\bar{X}_n$.
- The sampling distribution is a function of the distribution of observations F and the sample size n.
- We will get an *asymptotic approximation* by standardizing $\bar{X}_n$ and taking the limit as $n \to \infty$

Definition: Convergence in Distribution

- Let X_n be a sequence of random variables, and X a target random variable.
- We say that X_n **converges in distribution** to X if:

$$\lim_{n\to\infty} \mathbb{P}(X_n \leq t) = \mathbb{P}(X \leq t) \quad \text{for all } t \text{ where } F_X(t) \text{ is continuous}$$

- **Implications**
  - The cumulative distribution functions converge: $F_{X_n}(t) \to F_X(t)$
  - If the target distribution is a constant, then this is just convergence in probability
- **Notation:** $X_n \underset{d}{\to} X$

Standardizing the Sample Mean

- The WLLN tells us that $\bar{X}_n \underset{p}{\rightarrow} \mu$, so $\bar{X}_n \underset{d}{\rightarrow} \mu$.
- But the distribution of μ is degenerate, not useful.
- So rescale by the variance: $var[\bar{X}_n - \mu] = \frac{\sigma^2}{n}$, so $var[\sqrt{n}(\bar{X}_n - \mu)] = \sigma^2$

$$Z_n = \sqrt{n}(\bar{X}_n - \mu)$$

- As $n \rightarrow \infty$, $\mathbb{E}[Z_n] \rightarrow 0$, $\mathbb{E}[Z_n^2] \rightarrow \sigma^2$, $\mathbb{E}[Z_n^3] \rightarrow 0$, $\mathbb{E}[Z_n^4] \rightarrow 3\sigma^4$
- The Central Limit Theorem states that $Z_n \underset{d}{\rightarrow} Z$ where $Z \sim N(0, \sigma^2)$.
- This means that $\bar{X}_n \underset{a}{\sim} N(\mu, \ \frac{\sigma^2}{n})$.

The Central Limit Theorem

Theorem (Lindeberg–Lévy CLT)

Let $X_1, X_2, \ldots$ *be iid with* $\mathbb{E}[X_i] = \mu$ *and* $\mathrm{Var}(X_i) = \sigma^2 < \infty$. *Then:*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Equivalently: $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$.

Key conditions:

- **iid** — can be relaxed (Lindeberg–Feller allows heterogeneity)
- **Finite variance** — essential; without it, convergence may fail or be to a non-normal limit

Remarkable fact: The limit distribution is $N(0, \sigma^2)$ **regardless** of the distribution of X_i. Whether the data are Bernoulli, exponential, or uniform, the sample mean is approximately normal for large n.

CLT Sample Size Calculation

- U.S. wages are distributed $\mu = 24$ and $\sigma^2 = 430$, so $\bar{X}_n \underset{a}{\rightarrow} N(24, .43)$ for n=1000.
- How large does n have to be so that $\bar{X}_n$ is within \$1 of the true mean with 99% probability?

$$\mathbb{P}\left[|\bar{X}_n - \mu| \geq 1\right] \underset{a}{\sim} \mathbb{P}\left[|N(0,1)| \geq 1/\sqrt{var(\bar{X}_n)}\right] = \mathbb{P}\left[|N(0,1)| \geq 1/\sqrt{\frac{430}{n}}\right] = .01$$

$$\mathbb{P}(|Z| \geq z) = .01 \rightarrow qnorm(.005) = -2.576$$

$$\frac{1}{\sqrt{430/n}} = 2.576$$

$$n = \frac{430}{(1/2.576)^2}$$

$$n \geq 2{,}862$$

Compare with Chebyshev: $n = 43{,}000$. The CLT is far more efficient because it uses the shape

CLT in Action: R Simulation

```
# CLT: sample means of Exponential(1) converge to Normal
set.seed(42)
par(mfrow = c(2, 2))
for (n in c(5, 30, 100, 1000)) {
  xbar <- replicate(10000, mean(rexp(n, rate = 1)))
  hist(xbar, breaks = 40, prob = TRUE,
       main = paste("n =", n),
       xlab = expression(bar(X)[n]))
  curve(dnorm(x, 1, 1/sqrt(n)), add = TRUE,
        col = "red", lwd = 2)
}
```

The red curve is the normal approximation $N(\mu, \sigma^2/n)$. Even for **skewed** distributions (Exponential), the normal approximation is excellent by $n = 30$.

When Do Asymptotics Kick In?

Data shape	**Rule of thumb**	**Example**
Symmetric, light tails	$n \geq 30$	Heights, test scores
Moderate skew	$n \geq 100$	Wages, income
Heavy tails / outliers	$n \geq 500+$	Financial returns, conflict
Rare binary outcome	Events ≥ 10/covariate	Coups, wars

There is no universal rule. The CLT convergence rate depends on the **shape** of the underlying distribution. When in doubt, **simulate**.

Multivariate CLT

The CLT generalizes to vectors. If $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$ are iid k-vectors with $\mathbb{E}[\boldsymbol{W}_i] = \boldsymbol{\mu}$ and $\text{Var}(\boldsymbol{W}_i) = \boldsymbol{\Sigma}$, then:

$$\sqrt{n}(\bar{\boldsymbol{W}}_n - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$$

Applications:

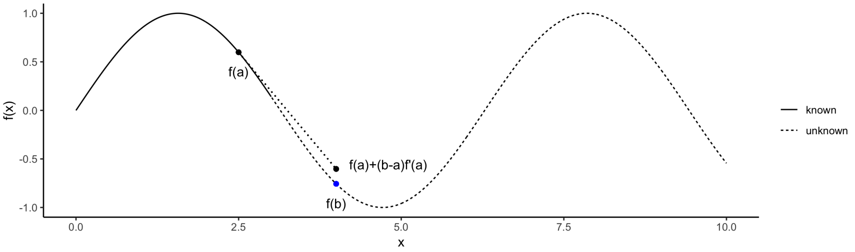
- **OLS** (Lecture 11): $\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i e_i \xrightarrow{d} N(\mathbf{0}, \sigma^2 Q_{XX})$
- **Probit MLE** (Lecture 9): $\frac{1}{\sqrt{n}} S_n(\beta_0) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}(\beta_0))$
- **GMM** (coming later): $\frac{1}{\sqrt{n}} \sum_{i=1}^n g(W_i, \theta_0) \xrightarrow{d} N(\mathbf{0}, \Omega)$

In each case, a sum of mean-zero random vectors converges to a multivariate normal. The covariance matrix differs, but the **structure** is identical.

Math Aside: Linear Approximation.

- We want to know $f(b)$, we know $f(a)$, $f'(a)$
- Math fact: linear Taylor series lets us approximate a value of a function $f(b) \approx f(a) + \frac{\partial f(a)}{\partial a}(b-a) + \ldots$.
- With the slope and the distance between a and b we can calculate the position of $f(b)$:

$$f(b) \approx f(a) + f'(a)(b-a)$$



Big Picture Example

- Suppose we want to estimate some function of parameters $f(b_1, b_2) = b_1/b_2$.
- What is the distribution of $\boldsymbol{f}(\boldsymbol{b})$?
- We can use the Delta Method (Asymptotic Normality + Taylor's Expansion) to approximate it.

Asymptotic Normality

If the errors are independently distributed with mean zero and finite variance and $\boldsymbol{X}$ is of full rank and well behaved, then the central limit theorem gives us that

$$\boldsymbol{b} \overset{a}{\sim} N[\beta, \ \frac{\sigma^2}{n}[\text{plim}\frac{\boldsymbol{X}'\boldsymbol{X}}{n}]^{-1}]$$

That is:

$$\text{Asy. Var}(\boldsymbol{b} - \beta) = \sigma^2 \frac{[\text{plim}\frac{\boldsymbol{X}'\boldsymbol{X}}{n}]^{-1}}{n}$$

Which we will estimate with:

$$s^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-K}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

The Delta Method

- Define a $J \times K$ matrix of partial derivatives across J functions and K variables.

$$\begin{aligned} \boldsymbol{C}(\boldsymbol{b}) &= \frac{\partial \boldsymbol{f}(\boldsymbol{b})}{\partial \boldsymbol{b}'} \\ \text{plim}\, \boldsymbol{C}(\boldsymbol{b}) &= \text{plim} \frac{\partial \boldsymbol{f}(\boldsymbol{b})}{\partial \boldsymbol{b}'} \\ &= \frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} \equiv \boldsymbol{\Gamma} \qquad \text{(Slutsky)} \end{aligned}$$

- The Taylor approximation from β to $\boldsymbol{b}$ is then

$$\boldsymbol{f}(\boldsymbol{b}) = \boldsymbol{f}(\boldsymbol{\beta}) + \boldsymbol{\Gamma} \times (\boldsymbol{b} - \boldsymbol{\beta}) + \text{ Higher Order things}$$

- The asymptotic distribution is: $\boldsymbol{f}(\boldsymbol{b}) \overset{a}{\sim} N(\boldsymbol{f}(\boldsymbol{\beta}), \; \boldsymbol{\Gamma}[\text{Asy. Var}(\boldsymbol{b} - \boldsymbol{\beta})]\boldsymbol{\Gamma}')$
- We can estimate this approximation with: $\boldsymbol{C}[s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]\boldsymbol{C}'$

Application: Greene Example 4.4

- Suppose we think effects persist through time:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \gamma y_{it-1} + \epsilon_{it}$$

- In the long run we expect stability:

$$\begin{aligned}
\bar{y} &= \beta_0 + \beta_1 \bar{x} + \gamma \bar{y} \\
\bar{y} - \gamma \bar{y} &= \beta_0 + \beta_1 \bar{x} \\
\bar{y}(1-\gamma) &= \beta_0 + \beta_1 \bar{x} \\
\bar{y} &= \frac{\beta_0}{(1-\gamma)} + \frac{\beta_1}{(1-\gamma)} \bar{x}
\end{aligned}$$

Long run effects with Lagged DV

- Suppose we raise x_{it} by 1:

$$
\begin{aligned}
y^*_{it} &= \beta_0 + \beta_1(\bar{x} + 1) + \gamma\bar{y} & &= \bar{y} + \beta_1 \\
y^*_{it+1} &= \beta_0 + \beta_1\bar{x} + \gamma(\bar{y} + \beta_1) & &= \bar{y} + \gamma\beta_1 \\
y^*_{it+2} &= \beta_0 + \beta_1\bar{x} + \gamma(\bar{y} + \gamma(\beta_1)) & &= \bar{y} + \gamma^2\beta_1
\end{aligned}
$$

Long run effects with Lagged DV

- The long run effect of a temporary one unit increase in x is thereby:

$$
\begin{aligned}
&= \beta_1 + \gamma\beta_1 + \gamma^2\beta_1 + \gamma^3\beta_1 + \ldots \\
&= \beta_1(1 + \gamma + \gamma^2 + \gamma^3 + \ldots) \\
&= \beta_1 \frac{1}{1-\gamma}
\end{aligned}
$$

Example: Gas Consumption

- Demand for gasoline is based on the kind of vehicle consumers choose to buy and how much they drive.
- We can model consumer demand for gas (G=gas expenditures/gas price) as a function of gas prices, income (Y) and past car purchasing decisions, which are themselves a function of gas prices and income.

$$ln(G/pop)_t = \hat{\beta}_1 + \hat{\beta}_2 lnP_{G,t} + \hat{\beta}_3 ln(Y/pop)_t + \hat{\beta}_4 ln(G/Pop)_{t-1} + e_t$$

- $\hat{\beta}_2$ is the short run price elasticity. $\hat{\beta}_3$ is the short run income elasticity.
- $\frac{\hat{\beta}_2}{1-\hat{\beta}_4}$ is the long run price elasticity, $\frac{\hat{\beta}_3}{1-\hat{\beta}_4}$ is the long run income elasticity.

Delta Method

- The Delta Method tells us that we can estimate the variance of a function f of our parameters by calculating the matrix of derivatives.
- Going back to gas consumption, call $\boldsymbol{b} = (\hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\beta}_3,\ \hat{\beta}_4)$, $\boldsymbol{f}(\boldsymbol{b}) = \begin{bmatrix} \frac{\hat{\beta}_2}{1-\hat{\beta}_4} \\ \frac{\hat{\beta}_3}{1-\hat{\beta}_4} \end{bmatrix}$

$$\boldsymbol{C}(\boldsymbol{b}) = \frac{\partial \boldsymbol{f}(\boldsymbol{b})}{\partial \boldsymbol{b}'} = \begin{bmatrix} \frac{\partial \frac{\hat{\beta}_2}{1-\hat{\beta}_4}}{\partial \hat{\beta}_1} & \frac{\partial \frac{\hat{\beta}_2}{1-\hat{\beta}_4}}{\partial \hat{\beta}_2} & \frac{\partial \frac{\hat{\beta}_2}{1-\hat{\beta}_4}}{\partial \hat{\beta}_3} & \frac{\partial \frac{\hat{\beta}_2}{1-\hat{\beta}_4}}{\partial \hat{\beta}_4} \\ \frac{\partial \frac{\hat{\beta}_3}{1-\hat{\beta}_4}}{\partial \hat{\beta}_1} & \frac{\partial \frac{\hat{\beta}_3}{1-\hat{\beta}_4}}{\partial \hat{\beta}_2} & \frac{\partial \frac{\hat{\beta}_3}{1-\hat{\beta}_4}}{\partial \hat{\beta}_3} & \frac{\partial \frac{\hat{\beta}_3}{1-\hat{\beta}_4}}{\partial \hat{\beta}_4} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{1-\hat{\beta}_4} & 0 & \frac{\hat{\beta}_2}{(1-\hat{\beta}_4)^2} \\ 0 & 0 & \frac{1}{1-\hat{\beta}_4} & \frac{\hat{\beta}_3}{(1-\hat{\beta}_4)^2} \end{bmatrix}$$

- The rows of this matrix are called gradient vectors.

```
USGasG <- read_csv("http://people.stern.nyu.edu/wgreene/Text/
                Edition7/TableF2-2.csv")
USGasG <- USGasG %>%
        mutate( logG     = log(GASEXP/(GASP*POP)),
                logGlag = lag(logG),
                logGASP = log(GASP))
fm <- lm(logG ~ logGASP+log(INCOME)+log(PNC)+log(PUC)+logGlag,
              data= USGasG )
xprimex <- vcov(fm)
sigmasq <- sigma(fm)^2
bs <- coef(fm)
ggas <-         c(0, 1/(1-bs[6]), 0, 0, 0, bs[2]/(1-bs[6])^2)
gincome <-      c(0, 0, 1/(1-bs[6]), 0, 0, bs[3]/(1-bs[6])^2)
seGpricelongrun <- sqrt(t(ggas)%*%(xprimex)%*%ggas)
seIncomelongrun <- sqrt(t(gincome)%*%(xprimex)%*%gincome)
```

```
> cat("Long run elasticity to Gas prices (95% conf. int.):",
        round(bs[2]/(1-bs[6]), 2), "+/-",
        round(seGpricelongrun*qnorm(.975),2))
Long run elasticity to Gas prices (95% conf. int.): -0.41 +/- 0.3

> cat("Long run elasticity to Income (95% conf. intervals):",
        round(bs[3]/(1-bs[6]), 2),"+/-",
        round( seIncomelongrun*qnorm(.975),2))
Long run elasticity to Income (95% conf. int.): 0.97 +/- 0.32
```

```
Canned Package Methods

In R:
> car :: deltaMethod(fm, "logGASP/(1-logGlag)")
                     Estimate      SE     2.5 %   97.5 %
logGASP/(1 - logGlag) -0.41136 0.15230 -0.70985 -0.1129

In STATA:
nlcom (_b[x2]/(1-_b[x6]))
```

Delta Method for Probit: Average Marginal Effects

Recall from Lecture 9: the Probit average marginal effect is a **nonlinear function** of $\hat{\beta}$:

$$\widehat{\text{AME}}_j = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i'\hat{\beta})\,\hat{\beta}_j$$

This is exactly the kind of problem the delta method solves. Define $h(\beta) = \text{AME}_j(\beta)$. Then:

$$\text{Asy. Var}(\widehat{\text{AME}}_j) = \nabla h(\hat{\beta})' \cdot \widehat{\text{Var}}(\hat{\beta}) \cdot \nabla h(\hat{\beta})$$

where ∇h accounts for both the direct effect $(\hat{\beta}_j)$ and the indirect effect through $\phi(X_i'\hat{\beta})$.

Common pattern: Estimate $\hat{\theta}$ by MLE or OLS $\rightarrow$ compute a nonlinear function $h(\hat{\theta}) \rightarrow$ use the delta method for its standard error. Same logic for long-run elasticities, marginal effects, or any transformation of coefficients.

What We Proved Today

Tool	Statement	Application
WLLN	$\bar{X}_n \xrightarrow{p} \mu$	Consistency
CLT	$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$	Asymptotic normality
CMT	$\operatorname{plim} g(X_n) = g(\operatorname{plim} X_n)$	Functions of estimators
Delta Method	$h(\hat{\beta}) \overset{a}{\sim} N(h(\beta), \Gamma V \Gamma')$	SEs for transformations

These four tools, combined with the convergence hierarchy

$$X_n \xrightarrow{\text{a.s.}} X \quad \Rightarrow \quad X_n \xrightarrow{p} X \quad \Rightarrow \quad X_n \xrightarrow{d} X,$$

form the complete asymptotic toolkit for this course.

Looking Ahead

Lecture 11 applies today's tools to OLS:

- **Consistency**: WLLN $\Rightarrow \frac{1}{n}\boldsymbol{X'X} \xrightarrow{p} Q_{XX}$, CMT $\Rightarrow \hat{\beta} \xrightarrow{p} \beta$
- **Asymptotic normality**: Multivariate CLT $\Rightarrow \sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, Q_{XX}^{-1}\Omega Q_{XX}^{-1})$
- **Wald tests**: Delta method $\Rightarrow$ test nonlinear hypotheses about β

These tools also justify what we did informally in Lecture 9 (Probit):

- Score CLT: $\frac{1}{\sqrt{n}}S_n(\beta_0) \xrightarrow{d} N(0, \mathcal{I}(\beta_0))$ (multivariate CLT)
- Hessian convergence: $\frac{1}{n}H_n(\beta_0) \xrightarrow{p} -\mathcal{I}(\beta_0)$ (WLLN)
- Combining via Slutsky: $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \mathcal{I}^{-1})$