Econ 103: Introduction to Econometrics

Lecture 19 — Indicator (Dummy) Variables

Ryan Longmuir

UCLA

Summer Session C, 2026

Reading: Hill, Griffiths & Lim (5th ed.), §7.1–7.2, 7.4; Stock & Watson (4th ed.), §5.3, 8.3, 11.1.

Where we are

Every regressor so far has been quantitative — income, price, square feet. But many drivers of economic outcomes are **qualitative**: a house's *neighborhood*, a worker's *sex* or *region*, whether someone *got the treatment*.

We encode a qualitative factor as a **0/1 indicator** (dummy) variable, and it drops straight into OLS. **Today:**

- intercept dummies — shift the line (a group "premium"), with a *reference group*;
- slope dummies — let groups have different *slopes* (house-price UTOWN/POOL);
- flipping it around: when y itself is binary, the **linear probability model** — and its limits.

Today's plan

1. Intercept dummies
2. Slope dummies
3. Several categories & joint tests
4. The linear probability model

Part 1

Intercept dummies

A dummy shifts the intercept

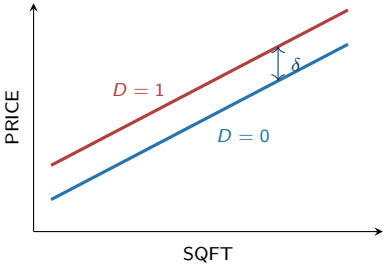
Start from a **hedonic** house-price model, $\text{PRICE} = \beta_1 + \beta_2 \text{SQFT} + e$. Does being near a university add value? Define $D = 1$ if near the university, 0 otherwise, and add it:

$$\text{PRICE} = \beta_1 + \delta D + \beta_2 \text{SQFT} + e.$$

The regression function splits in two:

$$\mathbb{E}(\text{PRICE} \mid \text{SQFT}) = \begin{cases} (\beta_1 + \delta) + \beta_2 \text{SQFT}, & D = 1 \\ \beta_1 + \beta_2 \text{SQFT}, & D = 0 \end{cases}$$

Same slope, intercept shifted by δ — a parallel shift. Here δ is the **location premium**: the price difference from being near the university, holding size fixed.



Parallel shift by the premium δ.

Reference group and the dummy-variable trap

The reference group

$D = 0$ is the **base** (reference) group — the omitted category everyone else is compared *to*. δ is the gap relative to the base. Which group is the base is your choice; pick the convenient one.

The dummy-variable trap

Don't include *both* D and its opposite $(1 - D)$ alongside the intercept — they sum to 1, so they're perfectly collinear with the constant (MR5 fails). Include only one; the omitted category becomes the base.

A dummy is treated like any regressor: δ has a standard error, a t-test ("is the premium significant?"), a confidence interval. Nothing new in the mechanics — only the interpretation as a group difference.

Part 2

Slope dummies

A dummy $\times$ a continuous variable changes the slope

Maybe location changes the *value per square foot*, not just the base level. Interact D with SQFT:

$$\text{PRICE} = \beta_1 + \beta_2 \text{SQFT} + \gamma\,(\text{SQFT} \times D) + e.$$

$$\frac{\partial\,\mathbb{E}(\text{PRICE})}{\partial \text{SQFT}} = \begin{cases} \beta_2 + \gamma, & D = 1 \\ \beta_2, & D = 0 \end{cases}$$

γ is the difference in slopes — the extra value of a square foot near the university. This is the **slope-indicator** (slope dummy) variable.

Combine *both* an intercept dummy and a slope dummy,

$$\text{PRICE} = \beta_1 + \delta D + \beta_2 \text{SQFT} + \gamma(\text{SQFT} \times D) + e,$$

and each group gets its *own* intercept *and* slope — equivalent to running two separate regressions (the basis of the **Chow test** for equal regressions).

Worked example: the university effect (HGL `utown`)

$N = 1000$ homes; regression of PRICE on UTOWN (intercept), SQFT, SQFT×UTOWN (slope), AGE, POOL, FPLACE:

variable	coeff
UTOWN (intercept)	27.45
SQFT	7.61
SQFT × UTOWN (slope)	1.30
AGE	−0.19
POOL	4.38
FPLACE	1.65

$R^2 = 0.87$; all significant (one-tail) except FPLACE borderline.

Reading the estimates (PRICE in \$1000, SQFT in 100s of ft^2):

- Location premium: \$27,453 near the university.
- Price per 100 ft^2: \$8,912 near campus vs. \$7,612 elsewhere — the slope dummy adds \$1,299.
- Each year of age: −\$190. A pool: +\$4,377. A fireplace: +\$1,649.

POOL and FPLACE are *intercept* dummies (level shifts); UTOWN enters *both* as an intercept and a slope dummy. This is the binary-interaction machinery previewed in Lecture 16.

Part 3

Several categories & joint tests

Categories with more than two levels

A factor with G categories (region: NE, South, Midwest, West) needs $G - 1$ dummies plus the intercept — not G (that's the trap again):

$$\text{WAGE} = \beta_1 + \beta_2 \text{EDUC} + \delta_1 \text{SOUTH} + \delta_2 \text{MIDWEST} + \delta_3 \text{WEST} + e.$$

- The omitted region (NORTHEAST) is the reference group.
- Each δ is that region's wage gap *relative to the Northeast*, holding education fixed (e.g. South $\approx -\$1.65$/hr).
- The choice of base is arbitrary — only the *comparisons* change, not the underlying fit.

Testing a whole categorical factor

Is there *any* regional effect? That's a joint hypothesis on all the region dummies:

$$H_0: \ \delta_1 = \delta_2 = \delta_3 = 0.$$

Use an ***F*-test** (Lecture 17), *not* three separate t's. For the wage data $F = 1.58$ ($p = 0.19$) — fail to reject; no significant regional difference in this sample.

Dummies can interact with each other, too

Want the wage gap specific to *black women*? Separate BLACK and FEMALE dummies miss it — add the product BLACK×FEMALE. Each cell (white male, black male, white female, black female) then gets its own intercept, read off as a sum of coefficients.

Part 4

The linear probability model

When the dependent variable is binary

Flip the dummy to the left-hand side: many outcomes are yes/no — a mortgage *denied*, Coke vs. Pepsi, college or not. Let $y \in \{0, 1\}$. Then

$$\mathbb{E}(y \mid X) = 1 \cdot \mathbb{P}(y = 1 \mid X) + 0 \cdot \mathbb{P}(y = 0 \mid X) = \mathbb{P}(y = 1 \mid X).$$

So the regression *is* a model of a probability — the **linear probability model** (LPM):

$$\mathbb{P}(y = 1 \mid X) = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K.$$

Each coefficient is the change in the probability that $y = 1$ for a one-unit change in that regressor. Estimate by OLS, exactly as before.

LPM example: mortgage denial (Boston HMDA)

Does race affect the chance of a mortgage denial, holding the payment-to-income ratio fixed? With $y = \text{deny}$:

$$\widehat{\text{deny}} = -0.091 + 0.559\,(\text{P/I ratio}) + 0.177\,\text{black}.$$

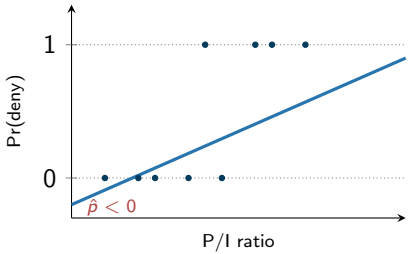
- A 0.1 rise in the P/I ratio raises the denial probability by about $0.559(0.1) \approx 5.6$ percentage points.
- Holding P/I fixed, a Black applicant's denial probability is 17.7 points higher than a white applicant's ($t = 7.1$).

Suggestive — but *not* proof of discrimination: credit history and other factors are omitted, so OVB (Lecture 18) is a live worry. The coefficient is a starting point, not a verdict.

The limits of the LPM

Linearity makes the LPM easy — and is also its flaw.

1. Probabilities can leave $[0, 1]$. A straight line eventually predicts $\hat{p} < 0$ or $\hat{p} > 1$ — nonsense for a probability.
2. Errors are heteroskedastic: $\mathrm{Var}(e \mid X) = p(1 - p)$ depends on X, so SR3/MR3 fails. Use robust standard errors.
3. R^2 is not meaningful (points can't lie on a line when y is 0/1).



The line dips below 0, rises above 1.

Still, it estimates marginal effects well when p isn't near 0 or 1, and it's transparent.

The fix — **probit**/**logit**, which use an S-curve to keep $\hat{p} \in (0, 1)$ — is beyond this course (HGL ch. 16, S&W §11.2).

Recap

Dummies as regressors

- intercept dummy δD: parallel shift (premium); $D = 0$ is the base
- slope dummy $\gamma(x \times D)$: different slope
- UTOWN: +\$27.5k premium; \$8,912 vs \$7,612 per 100 ft^2
- G categories $\to$ $G - 1$ dummies; trap if you keep all

Binary y: the LPM

- $\Pr(y{=}1 \mid X) = \beta_1 + \beta_2 x_2 + \ldots$
- coeff = change in probability
- mortgage: black +17.7pp denial
- flaws: $\hat{p} \notin [0,1]$, heterosk. (robust SE), no R^2

Next time (Lecture 20): treatment effects & difference-in-differences

The most important dummy of all is the treatment indicator. With potential outcomes, the ATE, and randomization (Project STAR), we'll see when a regression coefficient is truly *causal* — and close the loop on correlation vs. causation.

Questions?