Econ 103: Introduction to Econometrics

Lecture 11 — Prediction & Goodness of Fit

Ryan Longmuir

UCLA

Summer Session C, 2026

Reading: Hill, Griffiths & Lim (5th ed.), §4.1–4.2; Stock & Watson (4th ed.), §4.3.

Where we are

Welcome back. Before the midterm we built the full simple-regression toolkit: estimate (b_1, b_2), assess (BLUE), quantify precision (se), and test (t).

Two loose ends remain from that toolkit:

- In Lecture 8 we made a point prediction $\hat{y}_0 = b_1 + b_2 x_0$ but never said *how uncertain* it is. Today: the **prediction interval**.
- We never gave a single number for how well the line fits. Today: the sum-of-squares decomposition and R^2.

The link between them

Both come from the same split of y into an *explained* part (the fitted line) and an *unexplained* part (the residual). Prediction asks how big the unexplained part can be; R^2 asks how small it is on average.

Today's plan

1. Prediction intervals
2. Decomposing the variation
3. The coefficient of determination

Part 1

Prediction intervals

Recall: the point prediction and its error

To forecast y at a new x_0, the household obeys the same model $y_0 = \beta_1 + \beta_2 x_0 + e_0$, and our point predictor is the fitted value

$$\hat{y}_0 = b_1 + b_2 x_0.$$

We judge it by the **forecast error**

$$f = y_0 - \hat{y}_0 = (\beta_1 + \beta_2 x_0 + e_0) - (b_1 + b_2 x_0), \qquad \mathbb{E}(f \mid x) = 0.$$

So $\hat{y}_0$ is unbiased (the BLUP). But "unbiased" only centers the forecast — to put a band around it we need the variance of f.

The forecast variance: two sources of error

A forecast misses for two independent reasons, and the variance adds them up (HGL Appendix 4A):

$$\mathrm{Var}(f \mid x) = \sigma^2 \left[\underbrace{1}_{\text{new shock } e_0} + \underbrace{\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}_{\text{estimating the line}} \right].$$

- The "1" is the *irreducible* part: household 0 has its own fresh e_0 we can never anticipate. Even if we knew β_1, β_2 *exactly*, this term remains.
- The other two terms are the line-estimation error — exactly the variance of the mean estimate $\hat{\lambda} = b_1 + b_2 x_0$ from Lecture 9.

Replace σ^2 by $\hat{\sigma}^2$ and take the root to get the **standard error of the forecast**, $\mathrm{se}(f) = \sqrt{\widehat{\mathrm{Var}}(f \mid x)}$.

The prediction interval

With $\text{se}(f)$ in hand, the $100(1-\alpha)\%$ **prediction interval** for the outcome y_0 is

$$\hat{y}_0 \pm t_c \,\text{se}(f), \qquad t_c = t_{(1-\alpha/2,\, N-2)}$$

Example (Food expenditure, $x_0 = 20$)

$\hat{y}_0 = 287.61$. *With* $\hat{\sigma}^2 = 8013.29$, $N = 40$, $\bar{x} = 19.60$, *the standard error of the forecast is* $\text{se}(f) = \sqrt{8214.31} = 90.63$. *Using* $t_c = 2.024$,

$$287.61 \pm 2.024(90.63) = [104.13,\ 471.09].$$

A household with \$2,000 income spends somewhere between \$104 and \$471 — enormously wide. The point forecast \$287.61 is not, by itself, very useful.

Why so wide? The "1" dominates

Compare the two interval widths at $x_0 = 20$:

	variance term	95% interval
CI for the mean $\mathbb{E}(y \mid x_0)$	$\hat{\sigma}^2[\frac{1}{N} + \frac{(x_0-\bar{x})^2}{\sum(\cdot)^2}] = 201$	[258.91, 316.31]
PI for the outcome y_0	$\hat{\sigma}^2[1 + \frac{1}{N} + \frac{(x_0-\bar{x})^2}{\sum(\cdot)^2}] = 8214$	[104.13, 471.09]

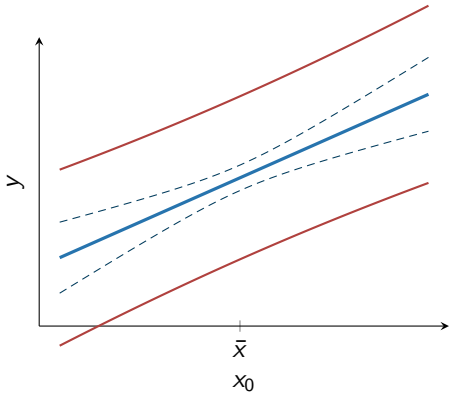
The prediction variance (8214) is almost all the "$1 \cdot \hat{\sigma}^2 = 8013$" term. Individual household behavior is intrinsically noisy: income explains the average well, but one family's spending swings for a hundred unmodeled reasons.

Collecting more data shrinks the line-estimation terms ($1/N \to 0$) but *not* the "1." To predict individuals better you need better *variables*, not just more rows — a motive for multiple regression (Lecture 13).

The prediction band fans out

Because $\mathrm{Var}(f)$ grows with $(x_0 - \bar{x})^2$, the bands are narrowest at $\bar{x}$ and widen as x_0 moves away — we predict best where we have the most data.

- Both bands pinch in at the point of the means $(\bar{x}, \bar{y})$.
- Extrapolating far beyond the data is doubly risky: wider bands, and the linear model may not even hold out there.
- The mean band (inner) always sits *inside* the prediction band (outer) — the gap is the irreducible σ^2.



Solid red: prediction band. Dashed: mean band.

Part 2

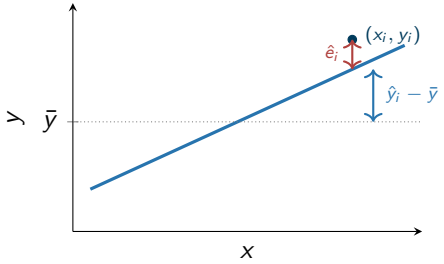
Decomposing the variation

Splitting each deviation in two

Now switch from *prediction* to *fit*: how much of the variation in y does the line account for? Start from the estimated decomposition $y_i = \hat{y}_i + \hat{e}_i$ and subtract $\bar{y}$ from both sides:

$$\underbrace{(y_i - \bar{y})}_{\text{total deviation}} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{explained by the line}} + \underbrace{\hat{e}_i}_{\text{unexplained}} .$$

- $\hat{y}_i - \bar{y}$: how far the *fitted* value moved from the mean — the part the regression explains.
- $\hat{e}_i = y_i - \hat{y}_i$: the leftover the line could not capture.
- Holds because the OLS line passes through $(\bar{x}, \bar{y})$.



From deviations to sums of squares

Square both sides and sum over i. The cross-product vanishes ($\sum(\hat{y}_i - \bar{y})\hat{e}_i = 0$), leaving a clean decomposition:

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum \hat{e}_i^2.$$

$$\boxed{\text{SST} = \text{SSR} + \text{SSE}}$$

- **SST** $= \sum(y_i - \bar{y})^2$ — total sample variation in y.
- **SSR** $= \sum(\hat{y}_i - \bar{y})^2$ — variation explained by the regression (a.k.a. explained sum of squares, ESS).
- **SSE** $= \sum \hat{e}_i^2$ — unexplained (residual) variation; the very thing OLS minimized.

Warning: notation collides across books. S&W call the explained part ESS and the residual part SSR. We follow HGL: SSR = regression (explained), SSE = error.

Part 3

The coefficient of determination

R^2: the fraction explained

Divide the decomposition by SST to get a unit-free fit measure, the **coefficient of determination**:

$$\boxed{R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}} \qquad 0 \le R^2 \le 1.$$

- $R^2 = 1$: all points on the line, $\text{SSE} = 0$ — perfect fit.
- $R^2 = 0$: the line is flat at $\bar{y}$, $\text{SSR} = 0$ — x explains nothing.
- In between: the proportion of the variation in y about its mean that the model explains.

Example (Food expenditure)

$$\text{SST} = 495132, \quad \text{SSE} = 304505,$$

$$R^2 = 1 - \frac{304505}{495132} = 0.385.$$

Income explains 38.5% of the variation in food expenditure; the other 61.5% is everything else.

R^2 is a squared correlation

Two neat identities in simple regression:

$$R^2 = r_{xy}^2 = r_{y\hat{y}}^2.$$

- r_{xy} is the sample correlation between x and y. Food: $r_{xy} = 0.62$, and $0.62^2 = 0.385 = R^2$. ✓
- $r_{y\hat{y}}$ is the correlation between the actual y and the fitted $\hat{y}$ — "how well do predictions track outcomes?" This version generalizes to multiple regression, where there is no single x to correlate with.

So R^2 and the **standard error of the regression** $\hat{\sigma}$ are two complementary fit summaries: R^2 is a *relative* share (unit-free), $\hat{\sigma}$ is an *absolute* typical miss (in the units of y).

Using R^2 wisely

"Is $R^2 = 0.385$ good?" — usually not a useful question.

- Microeconomic cross-sectional data: R^2 of 0.10–0.40 is completely normal — human behavior is hard to explain.
- Macro time-series data that trend together: R^2 of 0.90+ is routine. A high R^2 is not a badge of a better model.
- R^2 never falls when you add a regressor — so it can't, by itself, tell you whether a variable belongs (we fix this with *adjusted* R^2 in Lecture 18).

Judge a model by more than fit

Signs and magnitudes of coefficients, statistical *and* economic significance, whether the assumptions hold, and how it predicts out-of-sample data — all matter more than a high in-sample R^2.

Recap

Prediction intervals

- $\mathrm{Var}(f) = \sigma^2[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}]$
- PI: $\hat{y}_0 \pm t_c \,\mathrm{se}(f)$
- food: [104, 471] — the "1" (new shock) dominates
- bands pinch at $\bar{x}$, fan out; mean band $\subset$ prediction band

Goodness of fit

- $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$
- $R^2 = \dfrac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \dfrac{\mathrm{SSE}}{\mathrm{SST}}$
- food: $R^2 = 0.385$ (38.5% explained)
- $R^2 = r^2_{xy} = r^2_{y\hat{y}}$; don't over-read it

Next time (Lecture 12): functional forms

The line need not be straight in the *variables*. By logging or squaring x or y we model curves, elasticities, and growth rates — all still "linear regression," because it is linear in the parameters.

Questions?