Linear Regression

Lectures 24–25

QBA201 – Quantitative Business Analysis

Shi Qiu – Fall 2026

Population model: define the pieces

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

- Y_i: response for observation i.
- X_i: explanatory variable.
- β_0: intercept; β_1: population slope.
- ε_i: deviation from the population regression function.

If $\mathrm{E}(\varepsilon_i \mid X_i) = 0$, then $\mathrm{E}(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$.

The conditional mean is a line; individual observations need not lie on it.

The sample line and its residuals

$$\hat{Y}_i = b_0 + b_1 X_i, \qquad e_i = Y_i - \hat{Y}_i.$$

	Population	Sample
Intercept and slope	β_0, β_1	b_0, b_1
Deviation	ε_i	e_i
Line	Unknown conditional mean	Estimated fitted response

A residual is observed minus fitted. It is not the unobserved population error itself.

Least squares chooses the line

Ordinary least squares (OLS) selects b_0, b_1 to minimize

$$\mathrm{SSE} = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 = \sum_{i=1}^{n} e_i^2.$$

For a simple regression with an intercept,

$$b_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}, \qquad b_0 = \bar{Y} - b_1 \bar{X}.$$

The denominator must be positive: X cannot be constant.

Squared residuals prevent positive and negative misses from cancelling.

A small example we can calculate by hand

Constructed data: five observations

X	Y	$X - \bar{X}$	$Y - \bar{Y}$	Product
1	2	-2	-2	4
2	4	-1	0	0
3	5	0	1	0
4	4	1	0	0
5	5	2	1	2

$\bar{X} = 3$, $\bar{Y} = 4$, $\sum (X - \bar{X})^2 = 10$, $\sum (X - \bar{X})(Y - \bar{Y}) = 6$.

Keep the cross-product sum separate from the squared-X sum.

Calculate and interpret the fitted line

$$b_1 = \frac{6}{10} = 0.6, \qquad b_0 = 4 - 0.6(3) = 2.2.$$

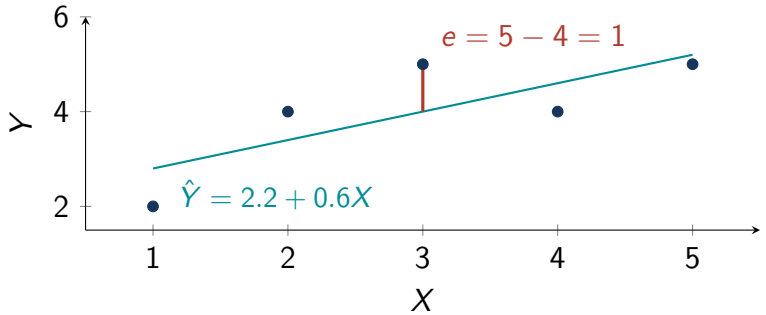
$$\boxed{\hat{Y} = 2.2 + 0.6X}$$

- An increase of one unit in X is associated with a 0.6-unit increase in fitted Y.
- At $X = 0$, the fitted value is 2.2, but zero is outside these observed X values.

Use units and context. The intercept may not have a meaningful real-world interpretation.

See the line and one residual

Five constructed observations and their fitted line:



Residuals are vertical differences, measured in units of Y.

Calculate the unexplained variation

X	Y	$\hat{Y}$	$e = Y - \hat{Y}$	e^2
1	2	2.8	-0.8	0.64
2	4	3.4	0.6	0.36
3	5	4.0	1.0	1.00
4	4	4.6	-0.6	0.36
5	5	5.2	-0.2	0.04

$\mathrm{SSE} = 2.40$. For simple regression, $s_e = \sqrt{\mathrm{SSE}/(n-2)} = \sqrt{2.4/3} = 0.8944$.

s_e estimates the error SD; the denominator allows for two estimated coefficients.

Measure fit using R-squared

For OLS with an intercept,

$$\mathrm{SST} = \sum_i (Y_i - \bar{Y})^2 = \underbrace{\sum_i (\hat{Y}_i - \bar{Y})^2}_{\mathrm{SSR}} + \mathrm{SSE}.$$

If $\mathrm{SST} > 0$,

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}.$$

In our example, $\mathrm{SST} = 6$, so $R^2 = 1 - 2.4/6 = 0.60$.

The fitted line accounts for 60% of the sample variation in Y. This is not a measure of causality.

Assumptions: what inference needs

- A correctly specified conditional mean: $\mathrm{E}(\varepsilon \mid X) = 0$.
- Independent observations, with nonconstant X.
- Constant conditional error variance for the usual SE formula.
- Normal conditional errors for exact small-sample t inference.

A least-squares line can be calculated even when its usual inferential assumptions fail.

Is the population slope different from zero?

For simple regression,

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0.$$

Under the classical normal-error assumptions,

$$t = \frac{b_1 - 0}{\mathrm{SE}(b_1)} \sim t_{n-2} \quad \text{under } H_0, \qquad \mathrm{SE}(b_1) = \frac{s_e}{\sqrt{\sum_i (X_i - \bar{X})^2}}.$$

The coefficient test concerns the population slope, not whether the fitted sample slope is exactly zero.

Complete the example's slope test

For our constructed data,

$$\mathrm{SE}(b_1) = \frac{0.8944}{\sqrt{10}} = 0.28284, \qquad t = \frac{0.6}{0.28284} = 2.1213.$$

With $n - 2 = 3$ degrees of freedom, the 5% two-sided critical value is 3.1824. Since $|2.1213| < 3.1824$, do not reject H_0.

Even $R^2 = 0.60$ need not yield a significant slope in a sample of five observations.

Prediction is not a guarantee

At $X = 4$, our fitted value is

$$\hat{Y} = 2.2 + 0.6(4) = 4.6.$$

- A fitted conditional mean is not an exact individual outcome.
- Prediction outside the observed X range is extrapolation.
- A historical cross-country income–happiness association does not identify the effect of raising one person's income.

Check the population, units, time period, and observed range before applying the line.

Check the residual pattern before trusting the line

- A curve in residuals versus X: a straight-line mean may be inadequate.
- A widening funnel: the constant-variance assumption may fail.
- Large isolated residuals: inspect outliers and possible recording errors.
- A time pattern: independence may fail for ordered observations.

A useful model needs sensible assumptions and residual behavior, not merely a large R^2.

Extend the notation, keep the interpretation

With k explanatory variables,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i.$$

- β_j: conditional mean change per unit of X_j, holding the other included variables fixed.
- A usual coefficient test uses $n - k - 1$ degrees of freedom.
- Linear means linear in the coefficients; transformed predictors can still enter a linear regression.

Multiple regression adjusts for included variables; adjustment alone does not prove causation.