Simple linear regression

Estimation, inference, diagnostics, and prediction

Prof. Ignacio Garrón

Preparatory Course in Statistics and Econometrics · Master in Finance · UC3M
Academic year 2026/2027

Regression targets a conditional mean

Finance often asks how an outcome changes with information:

- How does an asset return vary with the market return?
- How does credit spread vary with leverage?
- How does a portfolio loss vary with a risk factor?

The simple linear regression model summarizes the conditional mean:

$$\mathbb{E}[Y \mid X = x] = \beta_0 + \beta_1 x.$$

Regression is about the average value of Y among observations that share the same X.

What you should be able to do

1. Interpret the conditional mean, slope, intercept, and error term.
2. Derive and compute the OLS estimators.
3. Explain the assumptions behind unbiasedness, consistency, normality, and efficiency.
4. Evaluate fit with residuals, R^2, and the standard error of the regression.
5. Test coefficients, diagnose model failures, and construct predictions.

Part 1

Simple Regression Model: Conditional Means

Deterministic rule or stochastic model?

Deterministic relationship

$$Y = f(X)$$

Once X is known, Y is known exactly. For example,

$$F = 32 + 1.8C.$$

There is no prediction error to model.

Stochastic relationship

$$Y = f(X) + u$$

Observations with the same X can have different Y values because other determinants remain in u.

Regression estimates the systematic part of a stochastic relationship and quantifies the remaining uncertainty.

Regression direction follows the question

Question: How does Apple's return vary with the market return?

$$\underbrace{R_{\text{Apple},t}}_{Y_t} = \beta_0 + \beta_1 \underbrace{R_{\text{S\&P 500},t}}_{X_t} + u_t.$$

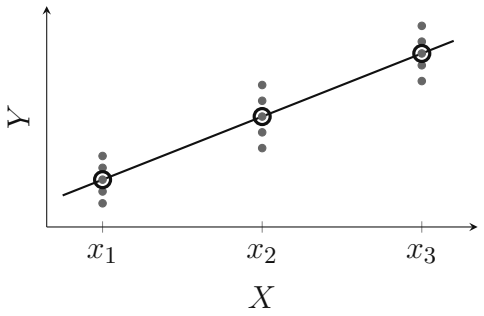
- Y_t is the response: the return to explain.
- X_t is the regressor: the market information used to explain it.
- u_t collects Apple-specific influences not summarized by X_t.

Regression is directional: choosing which variable is Y and which is X determines the question being answered.

Zero conditional mean centers every slice of Y

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X + u, \\
\mathbb{E}[u \mid X = x] &= 0, \\
\therefore \quad \mathbb{E}[Y \mid X = x] &= \beta_0 + \beta_1 x.
\end{aligned}
$$

- At a fixed x, outcomes vary because u varies.
- Their conditional average lies on the regression line.



Circles: conditional means. Vertical clouds: possible Y values at the same X.

The slope is a change in a conditional mean

$$\beta_1 = \frac{\Delta \mathbb{E}[Y \mid X = x]}{\Delta x}$$

for a linear conditional mean.

- β_1: expected change in Y from a one-unit increase in X.
- β_0: expected value of Y at $X = 0$.
- The intercept can be mathematically necessary even when $X = 0$ is not substantively interesting.
- Units matter: rescaling X or Y changes the numerical coefficients.

If X is expressed in percentage points rather than decimals, how does the slope change?

Correlation is not regression

For population random variables with finite second moments,

$$\operatorname{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]\,,$$
$$\rho_{XY} = \operatorname{Corr}(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} = \frac{\operatorname{Cov}(X, Y)}{\sigma_X \sigma_Y}, \qquad -1 \le \rho_{XY} \le 1.$$

- Correlation is unit-free and symmetric: $\operatorname{Corr}(X, Y) = \operatorname{Corr}(Y, X)$.
- It summarizes linear association but does not designate a response variable.
- $\rho_{XY} = 0$ can still hide a strong nonlinear relationship.

If the population conditional mean is $\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X$, then

$$\beta_1 = \frac{\operatorname{Cov}(X, Y)}{\operatorname{Var}(X)} = \rho_{XY} \frac{\sigma_Y}{\sigma_X}, \qquad \beta_0 = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X].$$

Regression is directional: it describes how the conditional mean of Y changes with X.

Association is not automatically causation

A regression slope can reflect:

- a causal effect of X on Y;
- reverse causality from Y to X;
- an omitted variable related to both X and Y;
- selection, measurement error, or simultaneity;
- a predictive relationship without a causal interpretation.

A causal interpretation requires a design or assumption that makes X as-good-as randomly assigned after conditioning on the model's information set.

A causal slope is an identification claim

Imagine a clinical trial for a new COVID vaccine: Let $D_i = 1$ for vaccine and 0 for placebo; let $Y_i = 1$ if participant i gets COVID and 0 otherwise.

1. **Randomize.** Assign volunteers by chance, so they do not choose D_i.
2. **Follow.** Treat both groups alike and record Y_i.
3. **Compare.** Estimate the binary-regressor model

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \qquad \beta_1 = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0].$$

Because D_i is randomly assigned, the difference in COVID rates identifies the vaccine's causal effect. A negative β_1 means the vaccine reduces risk.

Part 2

Estimating the Parameters: Least Squares Estimator

Ordinary least squares

The method of ordinary least squares (OLS) chooses the estimates to minimize the sum of squared residuals.

$$\widehat{u}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i,$$

so

$$\begin{aligned}\mathrm{SSR}(\widehat{\beta}_0, \widehat{\beta}_1) &= \sum_{i=1}^{n} \widehat{u}_i^2 \\ &= \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.\end{aligned}$$

The first-order conditions define OLS

Differentiate with respect to the candidate coefficients and evaluate at the minimizing estimates:

$$0 = \left.\frac{\partial \text{SSR}(\beta_0, \beta_1)}{\partial \beta_0}\right|_{(\widehat{\beta}_0, \widehat{\beta}_1)} = -2\sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i) = -2\sum_{i=1}^{n}\widehat{u}_i,$$

$$0 = \left.\frac{\partial \text{SSR}(\beta_0, \beta_1)}{\partial \beta_1}\right|_{(\widehat{\beta}_0, \widehat{\beta}_1)} = -2\sum_{i=1}^{n}X_i(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i) = -2\sum_{i=1}^{n}X_i\widehat{u}_i.$$

Thus $\sum_i \widehat{u}_i = 0$ and $\sum_i X_i\widehat{u}_i = 0$: the OLS normal equations.

OLS slope = covariance / variance

Let

$$S_{XX} = \sum_{i=1}^{n} (X_i - \overline{X})^2, \qquad S_{XY} = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}).$$

Then

$$\widehat{\beta}_1 = \frac{S_{XY}}{S_{XX}}, \qquad \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}.$$

Thus the fitted line passes through $(\overline{X}, \overline{Y})$.

OLS needs variation in X: if $S_{XX} = 0$, the slope cannot be estimated.

Worked example: Spanish wheat in the 1980s

The source reports ten observations on production X and price per kilogram Y (pesetas):

X	30	28	32	25	25	25	22	24	35	40
Y	25	30	27	40	42	40	50	45	30	25

From $\overline{X} = 28.6$, $\overline{Y} = 35.4$, $S_{XX} = 288.4$, and $S_{XY} = -390.4$,

$$\widehat{\beta}_1 = \frac{-390.4}{288.4} = -1.354, \qquad \widehat{\beta}_0 = 35.4 - (-1.354)(28.6) = 74.115.$$

Therefore,

$$\widehat{Y} = 74.115 - 1.354X.$$

Within the observed range, one additional unit of production is associated with about 1.35 fewer pesetas in expected price per kilogram.

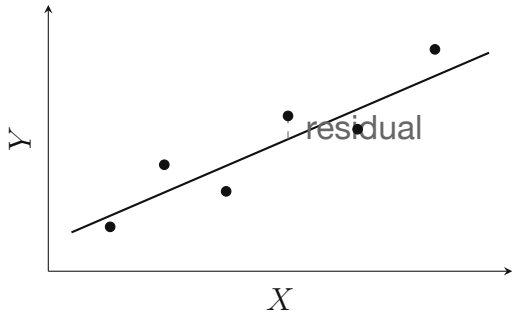
Residuals are vertical estimation errors

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

$$\widehat{u}_i = Y_i - \widehat{Y}_i$$

OLS minimizes $\sum_i \widehat{u}_i^2$.

Residuals are estimates of unobserved errors, not the errors themselves.



Schematic illustration; not empirical data.

OLS residuals satisfy useful identities

With an intercept in the model,

$$\sum_i \widehat{u}_i = 0, \qquad \sum_i X_i \widehat{u}_i = 0, \qquad \sum_i \widehat{Y}_i \widehat{u}_i = 0.$$

Consequences:

- $\overline{\widehat{u}} = 0$.
- $\overline{\widehat{Y}} = \overline{Y}$.
- Residuals are sample-uncorrelated with the regressor and fitted values.
- These are algebraic properties of OLS, not evidence that the population assumption $\mathbb{E}[u \mid X] = 0$ is true.

Part 3

Goodness of Fit

OLS decomposes total sample variation

Define

$$\text{TSS} = \sum_i (Y_i - \overline{Y})^2,$$

$$\text{ESS} = \sum_i (\widehat{Y}_i - \overline{Y})^2, \qquad \text{SSR} = \sum_i \widehat{u}_i^2.$$

With an intercept,

$$\text{TSS} = \text{ESS} + \text{SSR}.$$

Total variation equals explained variation plus residual variation. Notation varies across textbooks; here SSR means “sum of squared residuals.”

R^2 is the explained share of variation

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}}, \qquad 0 \leq R^2 \leq 1$$

for a model with an intercept.

- $R^2 = 0$: the fitted values explain none of the variation around $\overline{Y}$.
- $R^2 = 1$: every observation lies on the fitted line.
- In simple regression with an intercept, $R^2 = \text{Corr}(X, Y)^2$.

R^2 is an in-sample fit measure, not a causal score or a guarantee of good forecasts.

Regression standard error = residual scale

Under homoskedastic errors, estimate σ_u with

$$\text{SER} = s_u = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2}.$$

- SER is measured in the units of Y.
- The denominator $n-2$ reflects two estimated coefficients.
- A lower SER means observations lie closer to the fitted line on average.
- SER and R^2 answer different questions: absolute residual size vs. relative explained variation.

The wheat data: residual variance and fit

For the fitted line $\widehat{Y}_i = 74.115 - 1.354X_i$,

$$\mathrm{SSR} = \sum_i \widehat{u}_i^2 = 207.93.$$

$$R^2 = 0.718, \qquad R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SSR}/(n-2)}{\mathrm{TSS}/(n-1)} = 0.682.$$

Thus the fitted line explains about 71.8% of sample price variation; after adjusting for the estimated coefficients, the explained share is 68.2%.

A high R^2 cannot rescue a bad model

A high R^2 can coexist with:

- omitted-variable bias;
- reverse causality;
- nonlinearity hidden by a strong trend;
- unstable coefficients across periods;
- poor out-of-sample performance;
- mechanically related or nonstationary time series.

A low R^2 can still accompany a precisely estimated and economically important slope.

Part 4

Properties of the Least Squares Estimator

OLS assumptions

1. **Linear conditional mean**: $Y_i = \beta_0 + \beta_1 X_i + u_i$.
2. **Conditional exogeneity**: $\mathbb{E}[u_i \mid X_i] = 0$.
3. **Random sampling**: $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} F$.
4. **Identification and moments**: $0 < \mathrm{Var}(X_i) < \infty$ and sufficiently high moments exist for the stated limit theorem.

The needed assumptions depend on the claim: algebraic OLS identities, unbiasedness, consistency, asymptotic inference, and causality are different results.

Unbiasedness and consistency differ

Under zero conditional mean and the finite-sample OLS conditions,

$$\mathbb{E}[\widehat{\beta}_1 \mid X_1, \ldots, X_n] = \beta_1.$$

Under suitable large-sample conditions,

$$\widehat{\beta}_1 \xrightarrow{p} \beta_1.$$

- Unbiasedness is an exact repeated-sample statement.
- Consistency says the estimator approaches the target as n grows.
- If $\mathbb{E}[u \mid X] \neq 0$, more data need not remove the bias.

Asymptotic normality enables inference

Under the least squares assumptions and regularity conditions,

$$\frac{\widehat{\beta}_1 - \beta_1}{\mathrm{SE}(\widehat{\beta}_1)} \xrightarrow{d} N(0,1).$$

This approximation supports:

- confidence intervals for coefficients;
- tests of economic restrictions;
- comparisons of estimates with sampling uncertainty;
- prediction intervals after additional assumptions are stated.

Normal errors give an exact small-sample law

Add the classical assumptions

$$u_i \mid X_1, \ldots, X_n \overset{\text{ind.}}{\sim} N(0, \sigma_u^2).$$

Conditional on the regressors,

$$\widehat{\beta}_1 \mid X \sim N\left(\beta_1, \frac{\sigma_u^2}{S_{XX}}\right), \qquad \frac{\widehat{\beta}_1 - \beta_1}{s_u/\sqrt{S_{XX}}} \sim t_{n-2}.$$

- Normality is not needed to compute OLS or establish unbiasedness.
- It supplies exact finite-sample t inference under homoskedasticity; large-sample robust inference uses different conditions.

Heteroskedasticity changes standard errors

Homoskedasticity

$$\mathrm{Var}(u \mid X = x) = \sigma_u^2$$

Constant conditional variance.

Heteroskedasticity

$\mathrm{Var}(u \mid X = x)$ varies with x

Variance changes with x.

With zero conditional mean, OLS can remain unbiased and consistent, but homoskedastic standard errors are generally wrong.

Use heteroskedasticity-robust standard errors unless homoskedasticity is substantively credible.

Efficiency needs homoskedasticity

Under the Gauss–Markov conditions, including homoskedasticity, OLS is BLUE:

Best Linear Unbiased Estimator.

"Best" means that for every other linear unbiased estimator $\widetilde{\beta}$,

$$\mathrm{Var}(\widetilde{\beta} \mid X) - \mathrm{Var}(\widehat{\beta} \mid X) \quad \text{is positive semidefinite.}$$

- BLUE does not mean minimum variance among every imaginable estimator.
- Homoskedasticity is not required for OLS unbiasedness or consistency.
- With heteroskedasticity, robust inference repairs standard errors but does not make OLS efficient.

Part 5

Hypothesis Testing

Test a regression coefficient

For

$$H_0 : \beta_1 = \beta_{1,0}, \qquad H_1 : \beta_1 \neq \beta_{1,0},$$

compute

$$t = \frac{\widehat{\beta}_1 - \beta_{1,0}}{\mathrm{SE}(\widehat{\beta}_1)}.$$

In a large sample, the two-sided p-value is

$$p \approx 2\Phi(-|t|).$$

The common null $\beta_{1,0} = 0$ asks whether the linear conditional-mean relationship is statistically distinguishable from zero.

Confidence intervals show precision

A large-sample 95% confidence interval is

$$\widehat{\beta}_1 \pm 1.96\,\mathrm{SE}(\widehat{\beta}_1).$$

Read it in three layers:

1. **Direction**: are plausible values positive, negative, or both?
2. **Magnitude**: which effect sizes remain compatible with the data?
3. **Precision**: how wide is the interval?

Statistical significance asks whether zero is excluded; economic significance asks whether plausible effects matter.

Worked inference for the wheat slope

Under the classical Normal model,

$$\mathrm{SE}(\widehat{\beta}_1) = \frac{s_u}{\sqrt{S_{XX}}} = 0.300.$$

With $n - 2 = 8$ degrees of freedom and $t_{8,0.975} = 2.306$, the 95% confidence interval is

$$-1.354 \pm 2.306(0.300) = [-2.046, -0.661].$$

For $H_0 : \beta_1 = 0$,

$$t = \frac{-1.354}{0.300} = -4.509, \qquad p = 0.002.$$

The data reject zero linear association at the 5% level. Precision improves with a larger sample, greater spread in X, or smaller residual variance.

The intercept has its own standard error

Under the classical Normal model,

$$\mathrm{SE}(\widehat{\beta}_0) = s_u \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}}, \qquad \frac{\widehat{\beta}_0 - \beta_0}{\mathrm{SE}(\widehat{\beta}_0)} \sim t_{n-2}.$$

For the wheat data,

$$\widehat{\beta}_0 = 74.115, \qquad 95\%\ \mathrm{CI} = [53.970, 94.260].$$

This interval excludes zero, but $X = 0$ lies far outside the observed range $[22, 40]$.

Statistical significance does not make an extrapolated intercept economically meaningful.

Part 6

Residual Diagnostics

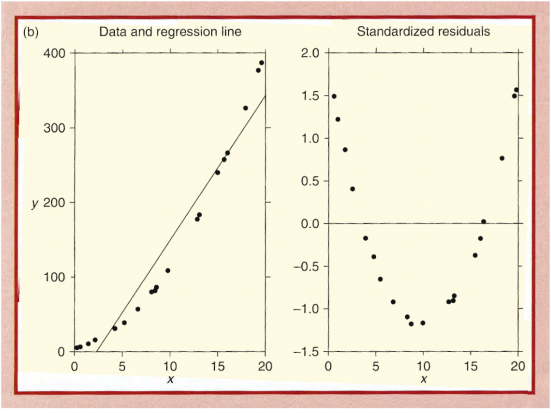
Residual diagnostics search for structure

A useful diagnostic set includes:

- residuals versus fitted values or X;
- residual scale versus fitted values;
- tail and quantile plots;
- leverage and influence measures;
- residuals over time and their autocorrelation;
- stability across subperiods or regimes.

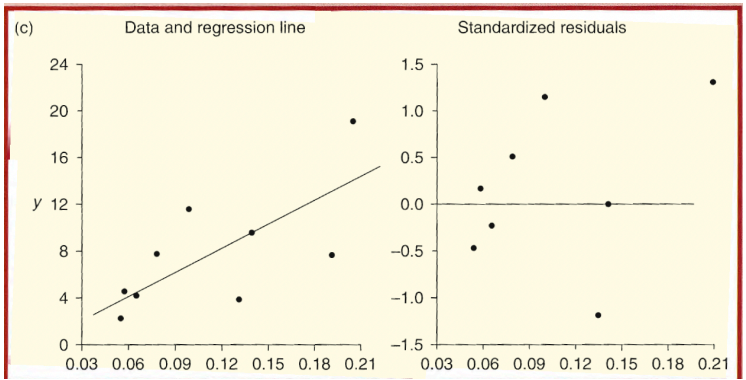
Diagnostics do not prove the model correct. They search for visible ways in which it is inadequate.

Curvature reveals nonlinearity



- The straight line misses the systematic bend in the data.
- Residuals trace a U-shape instead of scattering randomly around zero.
- Under the linear specification, $\mathbb{E}[u \mid X]$ varies with X.
- Consider transforming X or adding nonlinear terms.

Unequal spread: heteroskedasticity



- Residual variance changes with X:

$$\mathrm{Var}(u \mid X) \neq \sigma_u^2.$$

- OLS coefficients remain unbiased if $\mathbb{E}[u \mid X] = 0$.
- Homoskedastic standard errors are unreliable; report robust standard errors.

Part 7

Using the Regression Model to Predict

Mean response or new observation?

At $X = x_0$, let $\widehat{m}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$. Under homoskedastic Normal errors:

SE for conditional mean

$$s_u \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{X})^2}{S_{XX}}}$$

Uncertainty from estimating the line.

SE for new observation

$$s_u \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{X})^2}{S_{XX}}}$$

Adds a new realization of u.

The corresponding $(1 - \alpha)$ intervals are $\widehat{m}_0 \pm t_{n-2,1-\alpha/2}$ times the relevant standard error. A prediction interval for a new observation is wider because it includes a new realization of u.

Worked prediction: the same point, different uncertainty

At wheat production $x_0 = 30$,

$$\widehat{m}_0 = 74.115 - 1.354(30) = 33.50.$$

Under the classical Normal model, $t_{8,0.975} = 2.306$ gives:

Mean price at $X = 30$	**One new year's price**
33.50 ± 3.84	33.50 ± 12.37
$[29.66, 37.35]$	$[21.14, 45.87]$

Both intervals share the point estimate. The prediction interval is wider because it includes a new disturbance u_0.

Prediction is strongest within the data

- **Interpolation**: predict at x_0 within the range of observed X values.
- **Extrapolation**: predict beyond that range.

As x_0 moves away from $\overline{X}$,

$$\frac{(x_0 - \overline{X})^2}{S_{XX}}$$

increases and the estimated mean becomes less precise. The linear relationship itself may fail outside the observed range. A precise in-sample line does not justify distant extrapolation.

Example: Capital Asset Pricing Model (CAPM)

For excess returns,

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_i\big(R_{m,t} - R_{f,t}\big) + u_{i,t}.$$

- β_i: linear sensitivity of the asset's excess return to the market excess return.
- α_i: conditional mean excess return when the market excess return is zero.
- $u_{i,t}$: asset-specific component not captured by the market factor.
- R^2: fraction of sample variation associated with the market factor.

This regression describes exposure; it does not by itself prove a causal effect of market returns.

What the fitted line can—and cannot—tell us

It can summarize

- a linear conditional mean;
- in-sample fit;
- coefficient uncertainty;
- model-based predictions.

It cannot guarantee

- a causal effect;
- correct functional form;
- stable future relationships;
- valid inference under wrong SEs.

OLS is a calculation; regression analysis is the reasoning that makes the calculation credible.

Sources and chapter map

- James H. Stock and Mark W. Watson (2015), *Introduction to Econometrics*, 3rd ed. update, Pearson: Chapters 4–5; Chapters 13 and 17 for causal and projection formalisms.
- Michael B. Miller (2014), *Mathematics and Statistics for Financial Risk Management*, 2nd ed., Wiley: Chapter 10.
- Kevin Sheppard (2020), *Financial Econometrics Notes*, 17 January 2020: Chapter 3.