LM10 · QUANTITATIVE METHODS · CFA LEVEL I

Simple Linear Regression

OLS estimation, assumptions, R^2, ANOVA, hypothesis tests of coefficients, functional forms

What Is Simple Linear Regression?

Simple Linear Regression (SLR) describes a **linear relationship** between one **independent variable** (X, explanatory) and one **dependent variable** (Y, explained). It fits the line with the *smallest sum of squared errors*.

Learning Outcome Statements

LOS a	Describe SLR, how least squares is used to estimate coefficients, and their interpretation
LOS b	Explain the assumptions; describe how residuals and plots indicate violations
LOS c	Calculate and interpret measures of fit; tests of fit and of coefficients
LOS d	Describe ANOVA in regression; interpret ANOVA results; calculate/interpret SEE
LOS e	Calculate and interpret the predicted value and a prediction interval
LOS f	Describe different functional forms (log-lin, lin-log, log-log)

LOS A · THE REGRESSION MODEL

The SLR Model & Ordinary Least Squares

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

b_0 = intercept; **b_1** = slope coefficient; **ε** = error term (residual) — the portion of Y that cannot be explained by X.

Least Squares Criterion

Choose $\hat{b}_0$ and $\hat{b}_1$ to **minimize** the sum of squared deviations between observed Y and predicted Y:

$$min \Sigma(Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2 = \Sigma \varepsilon_i^2 = SSE$$

$$\hat{b}_1 = \mathrm{Cov}(X, Y)/\sigma_X^2 = \Sigma(X_i - \bar{X})(Y_i - \bar{Y})/\Sigma(X_i - \bar{X})^2$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \cdot \bar{X}$$

Key Properties

- The point ($\bar{X}$, $\bar{Y}$) always lies on the regression line.
- The denominator of $\hat{b}_1$ can never be negative, so the **sign of $\hat{b}_1$ is determined by Cov(X, Y)**.
- If $\hat{b}_1 > 0$, then $r_{XY} > 0$.
- $\Sigma\varepsilon_i = 0$ and $E(\varepsilon) = 0$.

LOS A · INTERPRETING COEFFICIENTS

Interpreting $\hat{b}_0$ and $\hat{b}_1$

Intercept ($\hat{b}_0$)

$\hat{Y} = b_0$ when $X_i = 0$. Only makes sense if the IV has meaning at $X = 0$.

Slope ($\hat{b}_1$)

The **change in Y for a one-unit change in X**. If $\hat{b}_1 > 0$, X and Y move in the same direction; if $\hat{b}_1 < 0$, they move in opposite directions.

WORKED EXAMPLE **ROA vs CAPEX**

GIVEN

ROA(%) = 4.875% + 1.25 · CAPEX(%)

INTERPRETATION

At CAPEX = 0%, predicted ROA = **4.875%**.
If CAPEX increases by 1 percentage point, ROA increases by **1.25 percentage points**.

Data Types

- **Cross-sectional:** many observations on X and Y for the same time period
- **Time-series:** many observations on Y (and sometimes X) from different time periods

LOS B · FOUR ASSUMPTIONS

SLR Assumptions

1. **Linearity:** the relationship between X and Y is linear in the parameters b_0 and b_1 — neither is multiplied or divided by another parameter. Implies X is not random.
2. **Homoskedasticity:** Var(ε) is the same for all observations. A violation (heteroskedasticity) indicates data may come from two different populations (cross-section) or regimes (time-series).
3. **Independence:** pairs (X, Y) are independent; ε is uncorrelated across observations (no serial correlation). Needed to correctly estimate variances of b_0 and b_1.
4. **Normality:** ε is normally distributed — required for valid tests of regression coefficients.

Exam Tip: A residual plot should show **randomness** with no pattern. Funnel shapes signal heteroskedasticity; curvature signals non-linearity; a trend signals autocorrelation.

CHECK YOUR UNDERSTANDING · QUIZ 1

Quiz 1

Select the best answer. Feedback appears after each click.

Q1. In $Y_i = b_0 + b_1X_i + \varepsilon_i$, OLS chooses $\hat{b}_0$ and $\hat{b}_1$ to MINIMIZE:

Q2. The sign of the estimated slope $\hat{b}_1$ is determined by:

Q3. The point $(\bar{X}, \bar{Y})$:

Q4. Residual plot shows a funnel (spread increases with X). This suggests a violation of:

LOS D · ANALYSIS OF VARIANCE

ANOVA in Regression

$$SST = SSR + SSE \quad (Total = Explained + Unexplained)$$

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE
Regression (explained)	$SSR = \Sigma(\hat{Y}_i - \bar{Y})^2$	$k = 1$	MSR = SSR / k
Error (unexplained)	$SSE = \Sigma(Y_i - \hat{Y}_i)^2$	$n - 2$	MSE = SSE / (n − k − 1)
Total	$SST = \Sigma(Y_i - \bar{Y})^2$	$n - 1$	

Coefficient of Determination (R^2)

$$R^2 = SSR/SST = explainedvariation/totalvariation$$

The fraction of total variation in Y explained by X. With one IV, **R² = (correlation of X and Y)²**. *Measure of fit, not a statistical test.*

F-statistic

$$F = MSR/MSE = (SSR/k)/[SSE/(n-k-1)]$$

Tests H_0: $b_1 = 0$ vs H_a: $b_1 \neq 0$. $df_1 = k$, $df_2 = n - k - 1$. In SLR, k = 1. Higher F ⇒ better model.

Standard Error of Estimate (SEE)

$$SEE = \quad MSE = \quad [\Sigma(Y_i - \hat{Y}_i)^2/(n-2)]$$

Standard deviation of the residuals. **Lower SEE ⇒ more accurate regression.** Also called standard error of the regression or root mean square error.

LOS D, F · COMPLETE EXAMPLE

Hypothesis Test of the Slope

Test Statistic for b_1

$$t = (\hat{b}_1 - b_{1,hypothesized})/S_{\hat{b}_1} \quad df = n - (k+1)$$

$$S_{\hat{b}_1} = SEE/ \quad \Sigma(X_i - \bar{X})^2$$

WORKED EXAMPLE **Slope Significance Test**

GIVEN

$\hat{b}_1 = 1.25$, $\Sigma(X - \bar{X})^2 = 122.64$, MSE = 11.9678, n = 6, r = 0.8945.

H_0: $b_1 = 0$; H_a: $b_1 \neq 0$; $\alpha = 5\%$.

SOLUTION

SEE = √11.9678 = **3.4594**

$S_{\hat{b}_1}$ = 3.4594 / √122.64 = **0.312398**

t = (1.25 − 0) / 0.312398 = **4.00131**

Critical value: T.INV(0.05, 4) = **2.776**

Also via r: t = 0.8945 √4 / √(1 − 0.8945^2) = **4.00131** (SLR only)

F-equivalent: $t^2 = 4.00131^2$ = **16.0104**, compare F-critical = F.INV(0.95,1,4) = 7.71

DECISION

|4.00| > 2.776 ⇒ **Reject H_0**. Slope is statistically significant.

Exam Tip: In SLR, the F-test and t-test of the slope are equivalent: $F = t^2$. Both test H_0: $b_1 = 0$.

LOS F · TESTS OF B_0 AND DUMMY VARIABLES

Intercept Test and Indicator Variables

Test of $\hat{b}_0$

$$t = (\hat{b}_0 - B_0)/S_{\hat{b}_0}, df = n - (k+1)$$

$$S_{\hat{b}_0} = SEE \quad [1/n + \bar{X}^2/\Sigma(X_i - \bar{X})^2]$$

Indicator (Dummy) Variables

If the IV is a dummy (0 or 1), $Y = b_0 + b_1 \cdot IND$.

- If IND = 0 ⇒ $\hat{Y} = b_0$ = mean of all "0" obs ($\bar{X}_0$)
- If IND = 1 ⇒ $\hat{Y} = b_0 + b_1$ = mean of all "1" obs ($\bar{X}_1$)
- $b_1 = \bar{X}_1 - \bar{X}_0$ = **difference in means**

The test of b_1 is equivalent to a **test of differences in means**.

p-values & Software Output

Most software reports p-values against H_0: parameter = 0 at α = 5% as default.

$$p-value = (1 - T.DIST(+t, df, 1)) \ \times \ 2 \quad (two-tailed)$$

CHECK YOUR UNDERSTANDING · QUIZ 2

Quiz 2

Select the best answer. Feedback appears after each click.

Q1. In SLR, SST = 400 and SSR = 260. R^2 is:

Q2. For an SLR with n = 40, degrees of freedom for the MSE are:

Q3. The standard error of estimate (SEE):

Q4. In SLR, the F-test of the slope is equivalent to:

Q5. If b_1 = 2.0, $S_{\hat{b}_1}$ = 0.5, n = 20, testing H_0: b_1 = 0 at α = 5% (two-tailed), the t-statistic is:

LOS E · PREDICTION INTERVAL

Predicted Values & Prediction Intervals

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 \cdot X_f$$

Two Sources of Error

- The equation itself is **estimated with error** (b_0, b_1 sampling error).
- The residuals $Y_i - \hat{Y}_i = \varepsilon$ are **estimated with error**.

$$s_f^2 = s_e^2 \cdot [1 + 1/n + (X_f - \bar{X})^2 / \Sigma(X_i - \bar{X})^2]$$

$$s_f = \quad s_f^2 \quad (adjusted SE for the forecast)$$

Prediction Interval Steps

1. Determine $\hat{Y}$
2. Select α
3. Determine t_c (critical value, df = n−2)
4. Determine s_f
5. $\hat{Y} \pm t_c \cdot s_f$

Prediction Accuracy Drivers

1. Better fit (lower s_e^2) ⇒ lower s_f^2
2. Larger n ⇒ smaller s_f^2
3. Closer X_f to $\bar{X}$ ⇒ smaller s_f^2 (extrapolation inflates error)

LOS F · FUNCTIONAL FORMS

Three Log Transformations

FORM	EQUATION	INTERPRETATION OF B_1	WHEN USED
Log-Lin	$\ln Y = b_0 + b_1X$	Relative change in Y for absolute change in X	Y grows exponentially (e.g., revenue growth rate)
Lin-Log	$Y = b_0 + b_1\ln(X)$	Absolute change in Y for relative change in X	Y and X differ greatly in scale (Y=%, X=$billions)
Log-Log	$\ln Y = b_0 + b_1\ln(X)$	Relative change in Y for relative change in X (elasticity)	Both variables grow proportionally

Model Selection Criteria

Select the functional form based on **goodness of fit**: R^2, F-stat, SEE.

- Residual plot should show randomness; distribution should be normal.
- If not, consider transforming the DV, IV, or both.

Exam Tip: A log-lin model (ln Y on X) is the classic choice for modeling a **growth rate** (e.g., revenue). The slope b_1 is then the approximate growth rate per unit X.

CHECK YOUR UNDERSTANDING · QUIZ 3

Quiz 3

Select the best answer. Feedback appears after each click.

Q1. A prediction interval for Y at X_f narrows when:

Q2. A log-log regression ln Y = b_0 + $b_1$ln(X) has slope coefficient interpreted as:

Q3. If we want to model a company's revenue that is growing at a compound rate, we would typically use a:

MODULE COMPLETE

LM10 · Simple Linear Regression Summary

Final Score

0 / 0

Cheat Sheet

$$Y_i = b_0 + b_1 X_i + \varepsilon_i; \hat{b}_1 = \mathrm{Cov}(X, Y)/\sigma_X^2$$

$$SST = SSR + SSE; R^2 = SSR/SST$$

$$SEE = \quad MSE; MSE = SSE/(n-2)$$

$$t = (\hat{b}_1 - b_{1,0})/S_{\hat{b}_1}, df = n - 2$$

$$F = MSR/MSE = t^2 (in SLR)$$

$$s_f^2 = s_e^2[1 + 1/n + (X_f - \bar{X})^2/\Sigma(X - \bar{X})^2]$$

Last-Minute Exam Checklist

OLS criterion: min $\Sigma\varepsilon^2_i$ = SSE

ROA example: ROA=4.875%+1.25·CAPEX

Slope test example: $\hat{b}_1$=1.25, t=4.00, reject at 5%

Four assumptions: Linear, Homo, Independ, Normal

Prediction interval: wider far from $\bar{X}$

Three log forms: log-lin (growth), lin-log (scale), log-log (elasticity)

Step 1 of 13