CHARTA LEARNING · LESSON NOTES

Simple linear regression

CFA Level I · Quantitative Methods

You will learn: Fit and read a simple regression line, check its assumptions, measure fit and typical error, make a forecast with a range, estimate a stock's alpha and beta, and read the ANOVA table and F-statistic.

Running example: Examples use made-up figures. Five months of data: market excess return X = 1, 2, 3, 4, 5 (%) and fund return Y = 3, 5, 6, 9, 7 (%).

1. The regression line

- The dependent variable Y is what you explain or predict (the fund); the independent variable X is what you use (the market). Simple means one X; linear means a straight line.
- A residual is a dot's miss: what happened minus what the line said. The best line squares the misses, adds them up, and makes that total as small as possible: ordinary least squares.
- b_0 is the intercept (Y when X is zero); b_1 is the slope (change in Y per 1 of X); the error is everything else that moves Y. The least squares line always passes through the point of the two averages.

> **$Y = b_0 + b_1 \times X + \text{error}$. $b_1 = \text{Cov}(X, Y) \div \text{Var}(X)$. $b_0 =$ mean of Y − $b_1 \times$ mean of X**

FIVE MONTHS: X = 1, 2, 3, 4, 5 AND Y = 3, 5, 6, 9, 7

Averages of X and Y	15 ÷ 5; 30 ÷ 5	**3; 6**
Sum of squared X gaps	4 + 1 + 0 + 1 + 4	**10**
Sum of X gap × Y gap	6 + 1 + 0 + 3 + 2	**12**
Slope b_1	12 ÷ 10	**1.2**
Intercept b_0	6 − 1.2 × 3	**2.4**
Fitted line; check at X = 3	Y = 2.4 + 1.2X; 2.4 + 1.2 × 3	**6**

2. Assumptions and what breaks them

- A regression will fit a line to any data, so checking the four assumptions is your job: linearity (the pattern is a straight line); constant spread of the errors; independent errors, not linked to each other or to X; normally distributed errors.
- Unequal spread (the misses fan out like a trumpet) is heteroskedasticity. Errors linked to each other (runs of positive then negative) is serial correlation, or autocorrelation. A curved pattern means the wrong shape.
- Heteroskedasticity and serial correlation do not usually bend the slope but make standard errors too small, so tests look more reliable than they are.
- Plot the residuals first. A shapeless band is fine; a curve means not linear; a fan means unequal spread; runs mean serial correlation; huge outliers cast doubt on normality.

Watch out: Quoting coefficients before looking at the residuals.

3. How good is the line? R squared

- The total sum of squares (SST) is the movement of Y around its average. The part the line explains is the regression sum of squares (SSR); the leftover squared misses are the sum of squared errors (SSE). SST = SSR + SSE.
- R squared, the coefficient of determination, is the explained share and runs from 0 to 1. In a simple regression it is the square of the correlation between X and Y.
- Read the numbers with units. Slope 1.2: when the market is 1 point higher, the fund is on average 1.2 points higher. Intercept 2.4: the fitted Y when X is zero, meaningful only if zero is a sensible X. R^2 never proves cause.

R^2 = SSR ÷ SST = 1 − SSE ÷ SST

SAME FIVE MONTHS

Total, SST	9 + 1 + 0 + 9 + 1	**20**
Explained, SSR	$1.2^2 \times 10$	**14.4**
Unexplained, SSE	20 − 14.4	**5.6**
R squared	14.4 ÷ 20	**0.72**

Watch out: Reading R squared as proof of cause or of a good model; quoting a slope without units.

4. Standard error of estimate and forecasting

- R squared has no units, so it cannot show how big the errors are. The standard error of estimate (SEE) is the typical miss in the units of Y. Divide by n − 2 because the slope and intercept were estimated. A smaller SEE means the dots hug the line.
- A point forecast puts X into the line. The prediction interval is the forecast ± critical t × forecast error, with n − 2 degrees of freedom. The forecast error is a little above the SEE and grows the further X is from the average X.
- Do not forecast outside the range of your data. The interval allows for uncertainty in the line, not for the line being the wrong shape out there.

SEE = √(SSE ÷ (n − 2)). Predicted Y = b_0 + b_1 × X. Interval = predicted Y ± critical t × forecast error

LINE Y = 2.4 + 1.2X, SSE = 5.6, N = 5

Standard error of estimate	√(5.6 ÷ 3) = √1.867	**1.37**
Forecast, market 4.5%	2.4 + 1.2 × 4.5	**7.8%**
Forecast, market 3%	2.4 + 1.2 × 3	**6.0%**
Critical t, 3 df; forecast error	from a t table	**3.182; 1.63**
95% interval at X = 4.5	7.8 ± 3.182 × 1.63 = 7.8 ± 5.19	**2.61 to 12.99**
Market 10%?	2.4 + 1.2 × 10, but data ran from 1 to 5	**14.4, do not trust**

Watch out: Dividing by n instead of n − 2, and forecasting for an X far outside the data.

5. Alpha and beta

- Regress a stock's excess return (minus the risk-free rate) on the market's excess return. The slope, beta, shows how strongly the stock reacts to the market. The intercept, alpha, is the average excess return left after the market has done its work.
- Beta is the correlation times the stock's standard deviation divided by the market's. Alpha comes from the averages, since the line passes through them. Alpha is only meaningful if it passes a significance test.

Stock excess return = alpha + beta × market excess return + error

MARKET: SD 4%, MEAN 0.5%. STOCK: SD 7%, MEAN 1.5%. CORRELATION 0.8

Beta	0.8 × 7 ÷ 4	**1.4**
Alpha	1.5 − 1.4 × 0.5	**0.8% a month**
R squared	0.8^2	**0.64**

Watch out: Swapping alpha and beta: beta is the slope, alpha is the intercept.

6. The ANOVA table and F-statistic

- Each sum of squares divided by its degrees of freedom gives a mean square. In a simple regression the explained part has 1 degree of freedom and the unexplained part has n − 2.
- F tests whether the slope is really zero. If it were, F would be near 1. A big F means the line explains far more than chance would. Compare F with the critical value from an F table. In simple regression $F = t^2$, so the two tests agree.
- A bigger sample can make a lower R^2 more convincing.

F = MSR ÷ MSE, with MSR = SSR ÷ 1 and MSE = SSE ÷ (n − 2)

TWO REGRESSIONS

n = 5: MSR; MSE	14.4 ÷ 1; 5.6 ÷ 3	**14.4; 1.867**
F-statistic	14.4 ÷ 1.867	**7.71**
Critical F, 1 and 3 df	7.71 < 10.13	**Do not reject**
Slope t-statistic	√7.71 vs critical 3.182	**2.78, not significant**
n = 32, SST 200, SSR 90: SSE; R^2	200 − 90; 90 ÷ 200	**110; 0.45**
MSE; F	110 ÷ 30 = 3.67; 90 ÷ 3.67	**24.5 vs 4.17: reject**

Key takeaways

- The slope is covariance ÷ variance of X, and the intercept is mean of Y − slope × mean of X.
- Check linearity, constant spread, independent errors and normality; plot the residuals first.
- R^2 = SSR ÷ SST is the explained share, and equals the correlation squared in simple regression.
- SEE = √(SSE ÷ (n − 2)) is the typical miss in units of Y; forecast only inside the range of the data.
- Beta is the slope and alpha is the intercept when you regress excess returns.
- F = MSR ÷ MSE tests whether the slope is zero, and $F = t^2$ in simple regression.

Quick check

1. Fitted line Y = 1 + 0.8X. What is the forecast when X = 5?
   Answer: 0.8 × 5 = 4, plus 1 = 5.
2. n = 22 and SSE = 40. What is the standard error of estimate?
   Answer: √(40 ÷ 20) = √2 ≈ 1.41.
3. Simple regression, $R^2 = 0.49$, slope positive. What is the correlation?
   Answer: √0.49 = +0.7 (positive because the slope is).