Linear regression

CHEATSHEET

Exam Mon, 7 Dec 2026

EXAM LIKELIHOOD Interpreting regression coefficients (ceteris paribus, units) **97%** · R-squared and goodness of fit **95%** · Dummy variables and reference category **92%** · Ordinary least squares (OLS) **92%** · Omitted variable bias **92%** · Error term and OLS assumptions **88%**

Reading regression output

Coefficient

How much does Y change per unit of X, other variables held constant?

p-value

Is the effect significantly different from zero?

R^2

What share of the variation in Y does the model explain?

Assumptions of linear regression

1. Linear relationship
2. Independent observations
3. Normally distributed residuals
4. Constant variance

M 01 Line and coefficients

- **Ordinary least squares**: minimises the sum of squared **residual**s (not the absolute ones). **Error term** = everything outside the model.
- Intercept = y when all x are zero. Slope b1 = change in y per +1 x, **ceteris paribus**. Say "is associated with", not "causes".

M 02 Tests

- **t-value** = b/standard error. **p-value** = probability of an estimate at least as extreme as the one found, if H0 is true; not the probability of H0.
- **Confidence interval** = b +/- about 2 x standard error; if it contains 0, not significant at 5%.
- **Prediction interval** is wider than the confidence interval.

M 03 Fit

- **R-squared** = SSE/SST = 1 – SSR/SST. Rises with every extra predictor.
- **Adjusted R-squared** = 1 – (1 – R2)(n – 1)/(n – k – 1). Can fall; compares models. Neither says anything about causality.
- **F-test**: H0 all slopes zero. F = (R2/k)/((1 – R2)/(n – k – 1)).
- **Partial F-test**: F = ((SSR_small – SSR_large)/q)/(SSR_large/(n – k – 1)); for a group of variables at once.

M 04 Dummy variables

- k categories gives k – 1 **dummy variable**s. The omitted group is the **reference category**; the coefficient is the difference with that group. k dummies is the dummy trap.

M 05 Forms of effects

- **Quadratic term**: marginal effect b1 + 2 b2 x, minimum or maximum at -b1/(2 b2) (minimum if b2 > 0).
- **Log transformation**: log(y) on x: about 100 x b percent per unit (approximation for small b; exactly e^b – 1); log-log: elasticity.
- **Interaction term**: effect of x1 = b1 + b3 ×2; b1 is the effect at x2 = 0.
- **Standardized coefficient**: effect in standard deviations, to compare predictors.
- **Mediation**: x works on y through m. Three regressions: total, x to m, direct. Moderation = interaction term.

M 06 Question 7.2 (9 p, 250 words)

1. Coefficients with units, ceteris paribus, meaning for the case (2 p)
2. p-values: significant at 5%, conditional on H0; significance is not size or importance (2 p)
3. Model fit: R-squared as explained variation; RSE as typical error; a good fit does not guarantee new data (2 p)
4. Sample: bias, representativeness, internal and external validity (3 p)

M 07 Worked example

Retail (simulated): R2 0.6644, n = 90, k = 3: adjusted = 1 – 0.3356 × 89/86 = 0.6527.

! Pitfalls

- **Omitted variable bias**: the missing variable is related to both y and x; the coefficient changes strongly when you add it.
- **Multicollinearity**: larger standard errors, no bias. **VIF** = 1/(1 – R2_j); watch out from about 5. Solution: drop or combine a variable.
- **Extrapolation** outside the range of x is unreliable.