Lineaire regressie

CHEATSHEET

Tentamen ma 7 dec 2026

TENTAMENKANS Regressiecoëfficiënten interpreteren (ceteris paribus, eenheden) **97%** · R^2 en modelfit (goodness of fit) **95%** · Dummyvariabelen en referentiecategorie **92%** · Kleinste-kwadratenmethode (OLS) **92%** · Weggelaten variabele (omitted variable bias) **92%** · Storingsterm (error term) en OLS-aannames **88%**

Regressie-output lezen

Coëfficiënt	p-value	R^2
Hoeveel verandert Y per eenheid X, bij gelijke andere variabelen?	Is het effect significant verschillend van nul?	Welk deel van de variatie in Y verklaart het model?

Aannames van lineaire regressie

1. Lineair verband
2. Onafhankelijke waarnemingen
3. Normaal verdeelde residuen
4. Constante spreiding

M 01 Lijn en coëfficiënten

- **Ordinary least squares**: minimaliseert de som van gekwadrateerde **residuals** (niet de absolute). **Error term** = alles buiten het model.
- Intercept = y als alle x nul zijn. Slope b1 = verandering in y per +1 x, **ceteris paribus**. Zeg "is geassocieerd met", niet "veroorzaakt".

M 02 Toetsen

- **t-value** = b/standard error. **p-value** = kans op een schatting minstens zo extreem als de gevonden, als H0 waar is; niet de kans op H0.
- **Confidence interval** = b +/- ca. 2 x standard error; bevat het 0, dan niet significant op 5%.
- **Prediction interval** is breder dan het confidence interval.

M 03 Fit

- **R-squared** = SSE/SST = 1 – SSR/SST. Stijgt bij elke extra voorspeller.
- **Adjusted R-squared** = 1 – (1 – R2)(n – 1)/(n – k – 1). Kan dalen; vergelijkt modellen. Beide zeggen niets over causaliteit.
- **F-test**: H0 alle slopes nul. F = (R2/k)/((1 – R2)/(n – k – 1)).
- **Partial F-test**: F = ((SSR_klein – SSR_groot)/q)/(SSR_groot/(n – k – 1)); voor een groep variabelen tegelijk.

M 04 Dummy variables

- k categorieën geeft k – 1 **dummy variables**. De weggelaten groep is de **reference category**; de coëfficiënt is het verschil met die groep. k dummies is de dummy trap.

M 05 Vormen van effecten

- **Quadratic term**: marginaal effect b1 + 2 b2 x, minimum of maximum bij -b1/(2 b2) (minimum als b2 > 0).
- **Log transformation**: log(y) op x: ca. 100 x b procent per eenheid (benadering bij kleine b; exact e^b – 1); log-log: elasticity.
- **Interaction term**: effect van x1 = b1 + b3 ×2; b1 is het effect bij x2 = 0.
- **Standardized coefficient**: effect in standaarddeviaties, om voorspellers te vergelijken.
- **Mediation**: x werkt via m op y. Drie regressies: totaal, x naar m, direct. Moderation = interaction term.

M 06 Vraag 7.2 (9 p, 250 woorden)

1. Coëfficiënten met eenheden, ceteris paribus, betekenis voor de casus (2 p)
2. p-values: significant op 5%, voorwaardelijk onder H0; significantie is niet grootte of belang (2 p)
3. Modelfit: R-squared als verklaarde variatie; RSE als typische fout; goede fit is geen garantie voor nieuwe data (2 p)
4. Steekproef: bias, representativiteit, interne en externe validiteit (3 p)

M 07 Rekenvoorbeeld

Winkel (gesimuleerd): R2 0,6644, n = 90, k = 3: adjusted = 1 – 0,3356 × 89/86 = 0,6527.

! Valkuilen

- **Omitted variable bias**: ontbrekende variabele hangt samen met y én x; de coëfficiënt verandert sterk als je haar toevoegt.
- **Multicollinearity**: grotere standard errors, geen bias. **VIF** = 1/(1 – R2_j); vanaf ca. 5 opletten. Oplossing: een variabele weglaten of combineren.
- **Extrapolation** buiten het bereik van x is onbetrouwbaar.