SAMENVATTING · OP TENTAMENNIVEAU

Module 5

Lineaire regressie

Alles wat je voor de open vragen moet kunnen definiëren, uitleggen, toepassen en beoordelen. Met de modellen als figuur, de vergelijkingen als tabel en de valkuilen gemarkeerd.

WAT ERIN ZIT

01	Ordinary least squares en residuals	2
02	Coëfficiënten lezen: intercept, slope en ceteris paribus	2
03	Standard error, t-value, p-value en confidence interval	3
04	R-squared en adjusted R-squared	3
05	F-test en partial F-test	3
06	Dummy variables en de reference category	3
07	Omitted variable bias	4
08	Multicollinearity en VIF	4
09	Gestandaardiseerde coëfficiënten	4
10	Niet-lineaire effecten: kwadraat en log	4
11	Interaction term	5
12	Mediation	5
13	Voorspellen en extrapolation	5
14	Uitschieters en steekproef	5
15	Vraag 7.2: opbouw van het antwoord (250 woorden)	5
→	Zo schrijf je het antwoord	6
→	Checklist: kan ik dit uitleggen?	6
→	Overhoor jezelf	7

TENTAMENKANS

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

Inschatting per onderwerp, op basis van het voorbeeldtentamen en de nadruk in de stof.

KERNBEGRIPPEN

ordinary least squares · residual · error term · intercept · slope · ceteris paribus · R-squared · adjusted R-squared · standard error · t-value · p-value · confidence interval · dummy variable · reference category

OP HET TENTAMEN

Blok 5

17.5 punten · 180 min totaal

MC	7 meerkeuzevragen	2.5 punten per vraag
7	Blok 7 (25 punten) bestaat uit drie open vragen over R-output.	200 / 250 / 150 woorden

Module 5 levert 7 meerkeuzevragen op (17,5 punten) en de open vraag 7.2 (9 punten, maximaal 250 woorden). Op het formuleblad staan alleen R-squared, adjusted R-squared en de t-toets; alle andere formules moet je kennen. De voorbeelden hieronder komen uit echte R-uitvoer; gesimuleerde casussen zijn als zodanig gemarkeerd.

01 Ordinary least squares en residuals



FIGUUR 5.1 **Aannames van lineaire regressie.** Controleer ze aan de residuen.

OLS kiest de lijn met de kleinste som van gekwadrateerde **residuals** (SSR). Een **residual** is waargenomen y min voorspelde y. De **error term** hoort bij het populatiemodel en bevat alles wat y beïnvloedt en niet in het model zit. De som van de residuals is bij OLS met een intercept altijd nul, dus je kwadrateert ze.

Tentamentip: "OLS minimaliseert de som van de absolute residuals" is fout; het zijn de gekwadrateerde residuals.

02 Coëfficiënten lezen: intercept, slope en ceteris paribus

De **intercept** is de voorspelde y als alle x nul zijn (bij dummies: de voorspelling voor de reference category). De **slope** b1 is de gemiddelde verandering in y bij een stijging van x met 1 eenheid. In een meervoudige regressie voeg je **ceteris paribus** toe: de overige voorspellers constant gehouden.

Voorbeeld (gesimuleerde data, 90 winkels, omzet in duizend euro): **omzet = 43,27 + 2,66 reclame + 6,31 personeel – 7,08 parkeer**. Per extra 1000 euro reclame stijgt de omzet gemiddeld met 2,66 duizend euro, ceteris paribus. Voorspelling voor reclame 10, personeel 8 en parkeer 1: 43,27 + 26,63 + 50,47 – 7,08 = 113,28 (afgerond).

Regressie op observationele data toont samenhang, geen causaliteit. Formuleer als "is geassocieerd met".

03 Standard error, t-value, p-value en confidence interval

De **t-value** is (b – beta0)/SE; bij H0: beta = 0 is dat b/SE. Reclame: 2,6625/0,6187 = 4,30, p = 4,4e-05, dus significant. Parkeer: -7,0823/4,8682 = -1,46, p = 0,149, dus niet significant op 5%.

De **p-value** is de kans op een zo extreme schatting als H0 waar is. Het is niet de kans dat H0 waar is. Het **confidence interval** is b +/- ongeveer 2 x SE. Staat nul erin, dan niet significant op 5%. Statistisch significant is niet hetzelfde als praktisch belangrijk: kijk ook naar de grootte.

04 R-squared en adjusted R-squared

R-squared = SSE/SST = 1 – SSR/SST, met SST = SSE + SSR. Winkelmodel: R-squared 0,6644 betekent dat 66,4% van de variatie in omzet verklaard wordt. Adjusted R-squared = 1 – (1 – R2)(n – 1)/(n – k – 1) = 1 – 0,3356 × 89/86 = 0,6527.

R-squared stijgt of blijft gelijk bij elke extra voorspeller; **adjusted R-squared** kan dalen en is de maat om modellen met verschillend aantal voorspellers te vergelijken. Het bepaalt niet of een voorspeller significant is: dat doet de p-value. Een hoge R-squared zegt niets over causaliteit of bias.

05 F-test en partial F-test

De **F-test** onderaan **summary(lm)** toetst H0: alle slopes zijn tegelijk nul. F = (R2/k)/((1 – R2)/(n – k – 1)) = (0,6644/3)/(0,3356/86) = 56,75. Een significante F zegt dat het model als geheel iets verklaart, niet dat elke voorspeller significant is.

De **partial F-test** (**anova(klein, groot)**) vergelijkt twee geneste modellen. Huizendata Long Island: het grote model met vier stijl-dummies extra (beide modellen hebben **taxes** en **rooms**) verlaagt de residuale kwadratensom van 793.934 naar 607.120; F = ((793.934 – 607.120)/4)/(607.120/83) = 6,38, p = 0,00016. De vier dummies voegen samen iets toe. R-squared stijgt van 0,0785 naar 0,2953, adjusted R-squared van 0,0573 naar 0,2444.

06 Dummy variables en de reference category

Een categorie met k niveaus wordt k – 1 **dummy variables**. De overgebleven categorie is de **reference category**; het intercept hoort bij die groep en elke dummycoëfficiënt is het verschil met die groep, ceteris paribus. Valkuil: k dummies tegelijk geeft perfecte multicollineariteit (de dummy-trap).

Voorbeeld (huizen op Long Island, 90 huizen): **valuem2 = 208,11 + 0,029 taxes – 157,98 f.regionSuffolk**; R-squared 0,427. Een huis in Suffolk is bij gelijke taxes gemiddeld 157,98 dollar per m2 minder waard dan in Nassau. Verander je de

referentie naar Suffolk, dan krijgt de dummy +157,98 en verandert het intercept; R-squared en voorspellingen blijven gelijk.

07 Omitted variable bias

Ontbreekt een variabele die samenhangt met zowel y als een opgenomen x, dan is die coëfficiënt vertekend (**omitted variable bias**): hij pikt ook het effect van de weggelaten variabele op. Je herkent het doordat de coëfficiënt sterk verandert zodra je de variabele toevoegt. Het tegenovergestelde is niet waar: een variabele die niet met x samenhangt, vertekent niets.

08 Multicollinearity en VIF

Multicollinearity is sterke samenhang tussen voorspellers. Gevolg: grotere standard errors en onstabiele, vaak niet-significante coëfficiënten; de schatting blijft unbiased. De **VIF** van variabele j is 1/(1 – R2_j), met R2_j uit de regressie van x_j op de andere voorspellers. Bij R2_j = 0,8 is de VIF 5. Huizendata: rooms 1,713, bedrooms 1,532, bath 1,229: geen probleem. Oplossing: een van de variabelen weglaten of samenvoegen.

09 Gestandaardiseerde coëfficiënten

Een **standardized coefficient** drukt het effect uit in standaarddeviaties: 1 sd meer in x hoort bij b sd meer in y. Daarmee vergelijk je de relatieve sterkte van voorspellers met verschillende eenheden. De bijbehorende p-values veranderen niet.

10 Niet-lineaire effecten: kwadraat en log

Een **quadratic term**: y = b0 + b1 x + b2 x^2 geeft een marginaal effect b1 + 2 b2 x en een minimum of maximum bij -b1/(2 b2) (een minimum als b2 > 0). Huizendata: lotsize -0,6455 en lotsize^2 0,0001853; minimum 0,6455/(2 × 0,0001853) = 1742 bij een bereik van 381 tot 2044; R-squared 0,7204. Daarbuiten geldt de kromme niet: **extrapolation** is onbetrouwbaar.

Een **log transformation**: in log(y) = b0 + b1 x is b1 ongeveer een relatieve verandering van 100 x b1 procent per eenheid x (exact: e^b1 – 1); in log-log is b1 een elasticity (procent per procent). Voorbeeld (gesimuleerde data, 240 alumni): in **log(salaris)** is de coëfficiënt voor vrouw -0,056; dat is ongeveer 5,6% lager salaris, ceteris paribus (p = 0,081, niet significant op 5%).

11 Interaction term

Met **y ~ x1 * x2** is het effect van x1 gelijk aan b1 + b3 ×2. De hoofdcoëfficiënt b1 is het effect bij x2 = 0. Bij een dummy en een getal heeft elke groep een eigen helling. Voorbeeld (gesimuleerde data, 150 bezorgingen): de interactie afstand x regen is 0,774 minuten extra per kilometer bij regen. Huizendata **rooms * f.region**: interactie 6,92 met p = 0,584, dus geen bewijs voor een verschil in helling.

12 Mediation

Mediation betekent dat x via een tussenvariabele m op y werkt. Je schat drie regressies: y op x (totale effect), m op x en y op x en m (direct effect). Wordt het directe effect van x nul zodra m in het model zit, dan is er volledige mediation. Moderation is iets anders: daar verandert een variabele de sterkte van het effect (interaction term).

13 Voorspellen en extrapolation

Voorspellen doe je met **predict(model, newdata = data.frame(...))**. Een **prediction interval** is breder dan een confidence interval, omdat het ook de spreiding van een individuele y bevat. Energiedata: 36,961 + 0,14 × 5 = 37,66. Voorspellen buiten het bereik van x (**extrapolation**) is onbetrouwbaar.

14 Uitschieters en steekproef

Twee uitschieters kunnen een resultaat omkeren: in de energiedata is de p-value 0,216 met alle waarnemingen en 0,047 zonder de twee waarnemingen met pCool boven 35 (R-squared 0,7% naar 1,9%). Dat is een reden om de data te bekijken, niet om waarnemingen te schrappen zonder reden.

15 Vraag 7.2: opbouw van het antwoord (250 woorden)

1. Coëfficiënten: lees b met eenheden, ceteris paribus, en zeg wat het voor de casus betekent (2 punten).
2. P-values: significant op 5% ja of nee, wat H0 zegt, geen kans op H0; significantie zegt niets over de grootte of het belang van het effect (2 punten).
3. Modelfit: R-squared of adjusted R-squared als aandeel verklaarde variatie, en RSE/RMSE als typische voorspelfout; een goede fit op deze data garandeert geen goede voorspelling voor nieuwe data (2 punten).
4. Steekproef: wie zit erin, bias, representativiteit, interne en externe validiteit (3 punten).

VOORBEELDANTWOORD

Zo schrijf je het antwoord

Vier bouwstenen, in deze volgorde. Elke steen levert punten op; laat er geen weg.

1. Term
Noem het model of begrip met de Engelse vakterm.

2. Definitie
Geef de definitie of de kenmerken, precies.

3. Casusfeit
Koppel aan een concreet feit uit de casus.

4. Conclusie
Sluit af met een oordeel, advies of gevolg.

OPDRACHT 1 · WOORDLIMIET 75 WOORDEN (WOORDLIMIET AANGENOMEN)

In een regressie is de coëfficiënt van een dummy variable 4,2 met $p < 0{,}05$. Wat betekent dat?

TERM De coëfficiënt van een dummy variable vergelijkt een groep met de referentiegroep. DEFINITIE Hij geeft het verwachte verschil in Y tussen groep 1 en de referentiegroep (0), bij gelijke andere variabelen. CASUSFEIT Hier is dat verschil 4,2 eenheden, en omdat $p < 0{,}05$ is het significant verschillend van nul. CONCLUSIE Groep 1 scoort gemiddeld 4,2 hoger dan de referentiegroep; dit is een verband, geen bewezen oorzaak.

63 / 75 woorden

KERNBEGRIPPEN

Checklist: kan ik dit uitleggen?

Vink af wat je uit je hoofd in één zin kunt uitleggen, met de Engelse term.

- [ ] ordinary least squares
- [ ] residual
- [ ] error term
- [ ] intercept
- [ ] slope
- [ ] ceteris paribus
- [ ] R-squared
- [ ] adjusted R-squared
- [ ] standard error
- [ ] t-value
- [ ] p-value
- [ ] confidence interval
- [ ] dummy variable
- [ ] reference category
- [ ] omitted variable bias
- [ ] multicollinearity
- [ ] VIF
- [ ] standardized coefficient
- [ ] interaction term
- [ ] F-test
- [ ] partial F-test
- [ ] mediation
- [ ] extrapolation

FLASHCARDS

Overhoor jezelf

Dek de rechterkolom af en zeg het antwoord hardop.

Vraag	Antwoord
1 Wat minimaliseert OLS?	De som van de gekwadrateerde residuals (SSR) rond de regressielijn.
2 Wat betekent het intercept?	De voorspelde y wanneer alle x-variabelen nul zijn (of de reference category).
3 Wanneer is een coefficient significant op het 5%-niveau?	Wanneer de p-value kleiner is dan 0,05 (|t| ongeveer groter dan 2).
4 Welke relatie geldt tussen SST, SSE en SSR?	SST = SSE + SSR: totale variatie = verklaard + residual.
5 Hoe interpreteer je een dummycoefficient?	Het gemiddelde verschil in y met de reference category, ceteris paribus.
6 Wat gebeurt er met de coefficient van x als je een relevante confounder weglaat?	Hij is vertekend (biased): hij pikt ook het effect van de confounder op.
7 Welke VIF-waarde wijst op een probleem?	Ruwweg een VIF boven 5 tot 10 (R2_j boven 0,8 tot 0,9).
8 Hoe interpreteer je b1 in log(y) = b0 + b1 x?	Een stijging van x met 1 eenheid hoort bij ongeveer 100 x b1 procent verandering in y.
9 Wat doet een interaction term x1 * x2?	Het effect van x1 op y hangt af van de waarde van x2 (en andersom).
10 Wat toetst de F-test onderaan summary(lm)?	H0: alle slopecoefficienten tegelijk nul; het model verklaart niets.
11 Wat is mediation?	x beinvloedt y via een tussenvariabele m (x naar m naar y).
12 Hoe voorspel je in R voor een nieuwe waarneming?	predict(model, newdata = data.frame(...)), met alle voorspellers erin.