SAMENVATTING · OP TENTAMENNIVEAU

Module 6

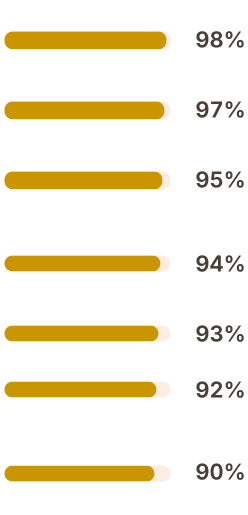
Logistische regressie, overfitting en validatie

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	Binary dependent variable en logistic regression	2
02	Odds, log-odds en de logit	2
03	Coëfficiënten interpreteren en odds ratio	3
04	Maximum likelihood	3
05	Modelfit: deviance, AIC en McFadden pseudo-R^2	4
06	Threshold en confusion matrix	4
07	Accuracy, sensitivity, specificity en precision	5
08	ROC-curve en AUC	6
09	Overfitting, train/test split en validatie	6
10	Tentamenvraag 7.1 aanpakken	7
→	Zo schrijf je het antwoord	7
→	Checklist: kan ik dit uitleggen?	7
→	Overhoor jezelf	8

TENTAMENKANS



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

KERNBEGRIPPEN

logistic regression · binary dependent variable · log-odds · odds ratio · maximum likelihood · AIC · McFadden pseudo-R^2 · confusion matrix · accuracy

OP HET TENTAMEN

Blok 6

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

*Getallen komen uit gesimuleerde data (n = 500, **set.seed(21)**), tenzij anders vermeld.*

01 Binary dependent variable en logistic regression

Definitie. Een binary dependent variable heeft twee uitkomsten, 0 of 1 (acceptatie, opzegging, wanbetaling). Je modelleert de kans p dat Y = 1. Het oefententamen vraagt (6.5) waarom hiervoor logit of probit nodig is.

Waarom niet OLS (lineair kansmodel)?

- Het verband tussen voorspellers en kans is niet-lineair (S-vorm).
- Voorspellingen kunnen buiten [0, 1] vallen. In de data: 16 van de 500 fitted values.
- De variantie van de storingsterm hangt af van de voorspelde kans (heteroscedasticiteit).

R. **glm(Accepteert ~ Inkomen + Koophuis + Leeftijd, family = binomial, data = d)**. Zonder **family = binomial** krijg je een gewone lineaire regressie.

Herkennen in output. Kolommen **z value** en **Pr(>|z|)**, daarna null deviance, residual deviance en AIC. Geen R^2 en geen F-statistic.

> **VALKUIL**
> **lm** met 0/1-uitkomst is geen logit; **predict()** zonder **type = "response"** geeft log-odds, geen kansen.

02 Odds, log-odds en de logit

Definitie. odds = p/(1 – p). log-odds (logit) = ln(p/(1 – p)). Het model (formuleblad):

ln(p/(1 – p)) = b0 + b1 ×1 + … + bk xk = z, dus p = 1/(1 + exp(-z)).

Rekenen. z = -1,5 geeft p = 1/(1 + exp(1,5)) = 0,18. p = 0,20 geeft odds 0,25. p = 0,75 geeft log-odds ln 3 = 1,10.

> **VALKUILEN**

- Log-odds lezen als kans.
- Odds en kans verwarren: 0,25 is hier de odds, de kans is 0,20.
- Het teken verkeerd lezen: een negatieve coëfficiënt verlaagt de odds (en dus de kans), een positieve verhoogt ze. Hoeveel procentpunten de kans verandert hangt af van de waarden van de andere variabelen, niet van b alleen.

03 Coëfficiënten interpreteren en odds ratio

Definitie. Een logit-coëfficiënt is een verandering in de log-odds per eenheid. De **odds ratio** is exp(b): de factor waarmee de odds vermenigvuldigen, ceteris paribus.

Voorbeeld uit de data.

	Estimate	exp(b)	Pr(> z)
Inkomen (per €1.000)	0,0432	1,044	6,2e-06
Koophuis (1 = koper)	1,0102	2,746	1,2e-05
Leeftijd (per jaar)	-0,0141	0,986	0,060

In bedrijfstaal. "Bij gelijk inkomen en gelijke leeftijd zijn de odds op acceptatie voor kopers ongeveer 2,75 keer zo groot als voor huurders." Leeftijd is bij α = 0,05 niet significant (p = 0,06): geen bewijs voor een verband, maar ook geen bewijs dat het nul is.

Meerdere eenheden. Odds ratios vermenigvuldigen. Bij 10 eenheden: exp(10 b), niet 10 maal de odds ratio. Voor €10.000 extra inkomen: exp(0,43241) = 1,54.

Referentiecategorie. Bij een factor met drie niveaus krijg je twee dummy variables; elke coëfficiënt vergelijkt met de referentiecategorie.

VALKUILEN

- "2,75 keer zoveel kans" zeggen. Het gaat om odds; odds ratio en kansverhouding liggen pas dicht bij elkaar als de uitkomst zeldzaam is.
- De grootte van een logit-coëfficiënt zien als marginaal effect. R_Functions zegt "sign and significance only", maar het officiële rubric van 7.1 geeft 1 punt voor richting én grootte (e^-0,0182 ≈ 0,982): het rubric wint. Noem dus ook bij een niet-significante coëfficiënt richting en grootte (exp(b)).
- Een odds ratio van 1 is geen effect. Staat 1 in het 95%-interval, dan is het effect niet significant.
- De p-value als kans op H0 lezen.

04 Maximum likelihood

Definitie. De coëfficiënten volgen uit maximum likelihood (MLE): kies de b's waarbij de waargenomen data de grootste likelihood hebben. De log-likelihood staat op het formuleblad: $\ell = \Sigma\,[y \ln p + (1 - y)\ln(1 - p)]$.

Eigenschappen (oefententamen 6.4). Bij groeiende steekproef is MLE consistent, asymptotisch normaal en efficiënt. Dat zijn eigenschappen voor grote n, niet voor kleine steekproeven. De z-toets is daarom een benadering.

Herkennen. De iteraties (Fisher scoring) staan onder de output; de deviance is -2 maal de log-likelihood.

05 Modelfit: deviance, AIC en McFadden pseudo-R^2

Null en residual deviance. Null deviance (model met alleen intercept): 577,20 op 499 df. Residual deviance (volledig model): 527,77 op 496 df. Een grotere daling betekent dat de voorspellers meer toevoegen.

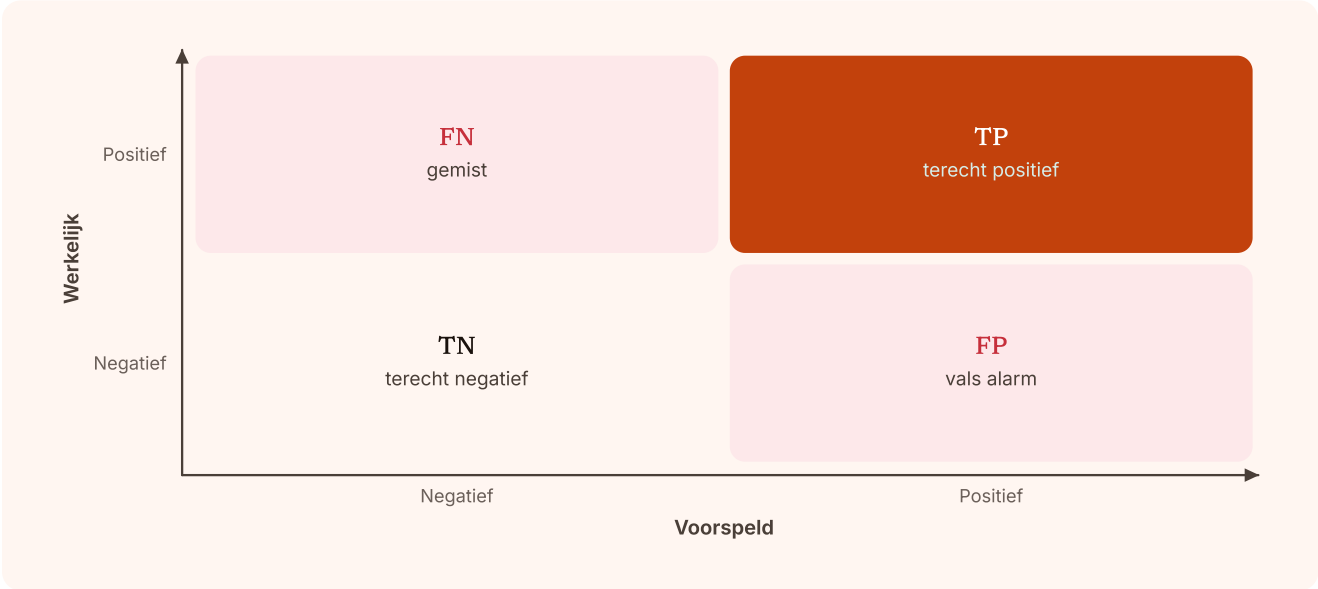
AIC (formuleblad) = 2k – 2 ln(L). Lager is beter. Met k = 4 parameters is AIC = 2 × 4 + 527,77 = 535,77. De 2k straft complexiteit: een variabele erbij verlaagt de deviance altijd een beetje, maar niet altijd de AIC. In de data daalt de deviance bij Stedelijk met 0,33 en stijgt de AIC van 535,77 naar 537,44.

McFadden pseudo-R^2 (formuleblad) = 1 – ln L(model)/ln L(null). Hier 1 – 527,77/577,20 = 0,086.

Modellen vergelijken. `anova(m1, m2, test = "Chisq")` toetst of de extra variabele de deviance significant verlaagt (p = 0,57: nee).

> **VALKUILEN**
> Pseudo-R^2 is geen percentage verklaarde variantie. AIC's alleen vergelijken bij dezelfde data.

06 Threshold en confusion matrix



FIGUUR 6.1 **Confusion matrix.** Accuracy = (TP + TN) / n. Sensitivity = TP / (TP + FN). Specificity = TN / (TN + FP). Precision = TP / (TP + FP).

Definitie. Je kiest een threshold τ (vaak 0,5): voorspelde kans ≥ τ geeft "positief". De **confusion matrix** telt TN, FN, FP en TP.

	Werkelijk 0	Werkelijk 1
Voorspeld 0	TN	FN
Voorspeld 1	FP	TP

Voorbeeld (τ = 0,5): TN 356, FN 115, FP 12, TP 17.

Let op de volgorde. R (**table**) zet 0 eerst, dus TN linksboven (tabel hierboven). Het oefententamen 7.1 zet 1 eerst: dan staat TP linksboven. Dezelfde cijfers in tentamenvorm:

	Werkelijk 1	Werkelijk 0
Voorspeld 1	17 (TP)	12 (FP)
Voorspeld 0	115 (FN)	356 (TN)

Lees altijd eerst de kop van rijen en kolommen; wie TN linksboven verwacht, wijst anders de verkeerde cel aan.

Rol van de threshold. Een lagere τ geeft meer positieve voorspellingen: sensitivity omhoog, specificity omlaag. Bij τ = 0,3 is de sensitivity 0,58 tegen 0,13 bij 0,5. Liggen alle voorspelde kansen onder τ (hier max 0,66), dan is er geen enkele positieve voorspelling en is precision niet te berekenen (0/0). Kies τ op basis van de kosten van FN en FP; de coëfficiënten en de AUC veranderen niet.

07 Accuracy, sensitivity, specificity en precision

Maat	Formule	Voorbeeld
accuracy	(TP + TN)/n	373/500 = 0,746
sensitivity	TP/(TP + FN)	17/132 = 0,129
specificity	TN/(TN + FP)	356/368 = 0,967
precision	TP/(TP + FP)	17/29 = 0,586

Vraag en maat. "Welk deel van de echte positieven vind je?" is sensitivity. "Welk deel van de meldingen klopt?" is precision. "Welk deel van de echte negatieven herken je?" is specificity.

Scheve klassen. Accuracy is misleidend als één klasse groot is (oefententamen 7.1 deel 3). Een model dat altijd "nee" zegt heeft bij 3% fraude 97% accuracy en sensitivity 0. Vergelijk met de baseline (aandeel van de meerderheidsklasse; hier 0,736) en kijk naar sensitivity en precision.

08 ROC-curve en AUC

De ROC-curve zet sensitivity (verticaal) tegen 1 – specificity (horizontaal) voor alle thresholds. De **AUC** is de oppervlakte eronder: de kans dat een willekeurige positieve waarneming een hogere voorspelde kans krijgt dan een willekeurige negatieve. AUC 0,5 is gokken, 1 is perfecte scheiding. In R: **prediction()** en **performance()** uit ROCR.

In de data: AUC 0,689 (train) en 0,714 (test). Een model met een hogere AUC op dezelfde testset onderscheidt de klassen beter (oefententamen 6.3). Een lagere AUC is op zichzelf geen overfitting.

> **VALKUIL**
> AUC is niet het aandeel goed geclassificeerd en hangt niet af van één threshold.

09 Overfitting, train/test split en validatie



FIGUUR 6.2 **Overfitting voorkomen met validatie.** Beoordeel een model op data die het nog niet gezien heeft.

Definitie. Overfitting: het model past op het toeval in de trainingsdata en voorspelt nieuwe data slechter. Je splitst om te beoordelen hoe goed het model generaliseert naar nieuwe data (oefententamen 6.6).

Werkwijze. Trek trainingsrijen willekeurig (**sample()**), schat op de training set, beoordeel op de test set.

Voorbeeld. Acceptatiemodel: accuracy 0,729 (train) en 0,793 (test). Dat wijst niet op overfitting; de test set is klein (150) en de naïeve baseline is 0,76. Fitnessvoorbeeld (gesimuleerd): groot model train 0,85, test 0,58; klein model train 0,67, test 0,62. Het grote model is overfit.

Cross-validation. k-fold: k delen, elk deel één keer test. LOOCV is k = n. Betere schatting van de testfout dan één split.

Bias-variance trade-off. Flexibeler: minder bias, meer variantie. De trainingsfout daalt altijd, de testfout is U-vormig.

> **VALKUILEN**
> De test set gebruiken om het model te kiezen (en dan dezelfde score rapporteren). Te weinig voorspellers heet underfitting, niet overfitting.

10 Tentamenvraag 7.1 aanpakken

Vraag 7.1 (9 punten, 200 woorden) geeft glm-output en een confusion matrix en vraagt drie dingen:

1. **Coëfficiënten in bedrijfstaal** (ongeveer 4 punten): teken, significantie, odds ratio, ceteris paribus.
2. **Berekening** (ongeveer 2 punten): accuracy en sensitivity of precision, met de formule en de getallen.
3. **Waarom accuracy misleidt** (ongeveer 3 punten): scheve verdeling van de uitkomst (baseline), lage sensitivity, en het gevolg voor het bedrijf (welke klanten of gevallen mis je). Threshold of overfitting noem je alleen als de casus erom vraagt.

Schrijf kort: één zin per punt, getallen erbij.

VOORBEELDANTWOORD

→ Zo schrijf je het antwoord

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

1. Term	2. Definitie	3. Casusfeit	4. Conclusie
Noem het model of begrip met de Engelse vakterm.	Geef de definitie of de kenmerken, precies.	Koppel aan een concreet feit uit de casus.	Sluit af met een oordeel, advies of gevolg.

OPDRACHT 1 · WOORDLIMIET 75 WOORDEN (WOORDLIMIET AANGENOMEN)

Een model scoort 99% accuracy op de train-set en 70% op de test-set. Wat is er aan de hand?

TERM Dit is overfitting. DEFINITIE Het model past zich aan de ruis in de train-set aan en generaliseert daardoor slecht naar nieuwe data. CASUSFEIT Het grote verschil tussen 99% op train en 70% op test laat dat zien. CONCLUSIE Beoordeel het model op de test-set of met cross-validation en kies een eenvoudiger model.

49 / 75 woorden

KERNBEGRIPPEN

✓ Checklist: kan ik dit uitleggen?

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

- ☐ logistic regression
- ☐ binary dependent variable
- ☐ log-odds
- ☐ odds ratio

- ☐ maximum likelihood
- ☐ AIC
- ☐ McFadden pseudo-R^2
- ☐ confusion matrix
- ☐ accuracy
- ☐ sensitivity
- ☐ specificity
- ☐ precision
- ☐ threshold
- ☐ ROC-curve
- ☐ AUC
- ☐ overfitting
- ☐ cross-validation
- ☐ bias-variance trade-off

FLASHCARDS

Overhoor jezelf

Dek de rechterkolom af en zeg het antwoord hardop.

Vraag	Antwoord
1 Wat is een binary dependent variable?	Een afhankelijke variabele met twee uitkomsten (0 of 1), zoals wel of niet opzeggen.
2 Hoe reken je odds uit een kans p uit?	odds = p/(1 – p); bij p = 0,20 zijn de odds 0,25.
3 Wat betekent een odds ratio van 1?	Geen effect: de odds veranderen niet als de voorspeller stijgt.
4 Hoe schat R een logitmodel?	Met maximum likelihood: de b's waarbij de waargenomen data de grootste likelihood hebben, iteratief.
5 Wat is de formule van de McFadden pseudo-R^2?	1 – ln L(model)/ln L(null), ook te berekenen als 1 – residual deviance/null deviance.
6 Hoe toets je of een extra variabele het model significant verbetert?	Met anova(m1, m2, test = "Chisq"): een likelihood-ratiotoets op de daling van de deviance.
7 Hoe reken je de precision uit?	TP/(TP + FP): van alles wat het model positief noemt, het aandeel dat echt positief is.
8 Wat is de baseline bij scheve klassen?	Het aandeel van de meerderheidsklasse: de accuracy van een model dat die klasse altijd voorspelt.
9 Wanneer kies je een lage threshold?	Als een gemist geval veel meer kost dan een vals alarm.
10 Wat betekent AUC = 0,5?	Het model scheidt de klassen niet beter dan toeval.
11 Welke maatregelen helpen tegen overfitting?	Het model vereenvoudigen, meer data verzamelen en de keuze toetsen met cross-validation.
12 Hoe werkt k-fold cross-validation?	De data gaat in k delen; elk deel is één keer testset, de overige k – 1 trainen. De scores worden gemiddeld.