SUMMARY · AT EXAM LEVEL

Module 6

Logistic regression, overfitting and validation

Everything you must be able to define, explain, apply and assess for the open questions. With the models as figures, the comparisons as tables and the pitfalls marked.

WHAT IS INSIDE

01	Binary dependent variable and logistic regression	2
02	Odds, log-odds and the logit	2
03	Interpreting coefficients and the odds ratio	2
04	Maximum likelihood	3
05	Model fit: deviance, AIC and McFadden pseudo-R^2	4
06	Threshold and confusion matrix	4
07	Accuracy, sensitivity, specificity and precision	5
08	ROC curve and AUC	5
09	Overfitting, train/test split and validation	6
10	Approaching exam question 7.1	6
→	How to write the answer	7
→	Checklist: can I explain this?	7
→	Test yourself	8

EXAM LIKELIHOOD

Logistic regression (logit model) CERTAIN	98%
Confusion matrix, computing accuracy and sensitivity CERTAIN	97%
Accuracy is misleading with imbalanced classes CERTAIN	95%
A logit coefficient is a change in log-odds, not in probability CERTAIN	94%
Odds and odds ratio (e^b) CERTAIN	93%
Overfitting and generalization CERTAIN	92%
Choosing precision, sensitivity and specificity for a business question CERTAIN	90%

Estimate per topic, based on the sample exam and the emphasis in the material.

KEY CONCEPTS

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

ON THE EXAM

Block 6

17.5 points · 180 min in total

MC	7 multiple-choice questions	2.5 points per question
7	Block 7 (25 points) consists of three open questions about R output.	200 / 250 / 150 words

Exam: Mon, 7 Dec 2026 · 180 min · Technical terms stay in English

*Numbers come from simulated data (n = 500, **set.seed(21)**) unless stated otherwise.*

01 Binary dependent variable and logistic regression

Definition. A binary dependent variable has two outcomes, 0 or 1 (acceptance, churn, default). You model the probability p that Y = 1. The practice exam asks (6.5) why logit or probit is needed for this.

Why not OLS (linear probability model)?

- The relation between predictors and probability is non-linear (S-shape).
- Predictions can fall outside [0, 1]. In the data: 16 of the 500 fitted values.
- The variance of the error term depends on the predicted probability (heteroscedasticity).

R. **glm(Accepteert ~ Inkomen + Koophuis + Leeftijd, family = binomial, data = d)**. Without **family = binomial** you get an ordinary linear regression.

Recognising it in output. Columns **z value** and **Pr(>|z|)**, then null deviance, residual deviance and AIC. No R^2 and no F-statistic.

> **PITFALL**
> **lm** with a 0/1 outcome is not a logit; **predict()** without **type = "response"** gives log-odds, not probabilities.

02 Odds, log-odds and the logit

Definition. odds = p/(1 – p). log-odds (logit) = ln(p/(1 – p)). The model (formula sheet):

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

Calculating. z = -1.5 gives p = 1/(1 + exp(1.5)) = 0.18. p = 0.20 gives odds 0.25. p = 0.75 gives log-odds ln 3 = 1.10.

> **PITFALLS**

- Confusing odds and probability: 0.25 here is the odds, the probability is 0.20.
- Misreading the sign: a negative coefficient lowers the odds (and so the probability), a positive one raises them. How many percentage points the probability changes depends on the values of the other variables, not on b alone.

03 Interpreting coefficients and the odds ratio

Definition. A logit coefficient is a change in the log-odds per unit. The **odds ratio** is exp(b): the factor by which the odds multiply, ceteris paribus.

Example from the data.

	Estimate	exp(b)	Pr(> z)
Inkomen (per €1,000)	0.0432	1.044	6.2e-06
Koophuis (1 = owner)	1.0102	2.746	1.2e-05
Leeftijd (per year)	-0.0141	0.986	0.060

In business language. "At equal income and age, the odds of acceptance are about 2.75 times as large for owners as for renters." Leeftijd is not significant at α = 0.05 (p = 0.06): no evidence of a relation, but also no evidence that it is zero.

Several units. Odds ratios multiply. For 10 units: exp(10 b), not 10 times the odds ratio. For €10,000 extra income: exp(0.43241) = 1.54.

Reference category. For a factor with three levels you get two dummy variables; each coefficient compares with the reference category.

PITFALLS

- Saying "2.75 times the chance". It is about odds; odds ratio and probability ratio are only close if the outcome is rare.
- Seeing the size of a logit coefficient as a marginal effect. R_Functions says "sign and significance only", but the official rubric of 7.1 gives 1 point for direction and size (e^-0.0182 ≈ 0.982): the rubric wins. So also for a non-significant coefficient state direction and size (exp(b)).
- An odds ratio of 1 means no effect. If 1 is in the 95% interval, the effect is not significant.

04 Maximum likelihood

Definition. The coefficients come from maximum likelihood (MLE): choose the b's under which the observed data have the largest likelihood. The log-likelihood is on the formula sheet: $\ell = \Sigma\,[y \ln p + (1 - y) \ln(1 - p)]$.

Properties (practice exam 6.4). As the sample grows, MLE is consistent, asymptotically normal and efficient. These are properties for large n, not for small samples. The z-test is therefore an approximation.

Recognising it. The iterations (Fisher scoring) are shown below the output; the deviance is -2 times the log-likelihood.

05 Model fit: deviance, AIC and McFadden pseudo-R^2

Null and residual deviance. Null deviance (intercept-only model): 577.20 on 499 df. Residual deviance (full model): 527.77 on 496 df. A larger drop means the predictors add more.

AIC (formula sheet) = 2k – 2 ln(L). Lower is better. With k = 4 parameters, AIC = 2 × 4 + 527.77 = 535.77. The 2k penalises complexity: adding a variable always lowers the deviance a little, but not always the AIC. In the data the deviance drops by 0.33 for Stedelijk and the AIC rises from 535.77 to 537.44.

McFadden pseudo-R^2 (formula sheet) = 1 – ln L(model)/ln L(null). Here 1 – 527.77/577.20 = 0.086.

Comparing models. `anova(m1, m2, test = "Chisq")` tests whether the extra variable lowers the deviance significantly (p = 0.57: no).

> **PITFALLS**
> Pseudo-R^2 is not a percentage of explained variance. Compare AICs only on the same data.

06 Threshold and confusion matrix

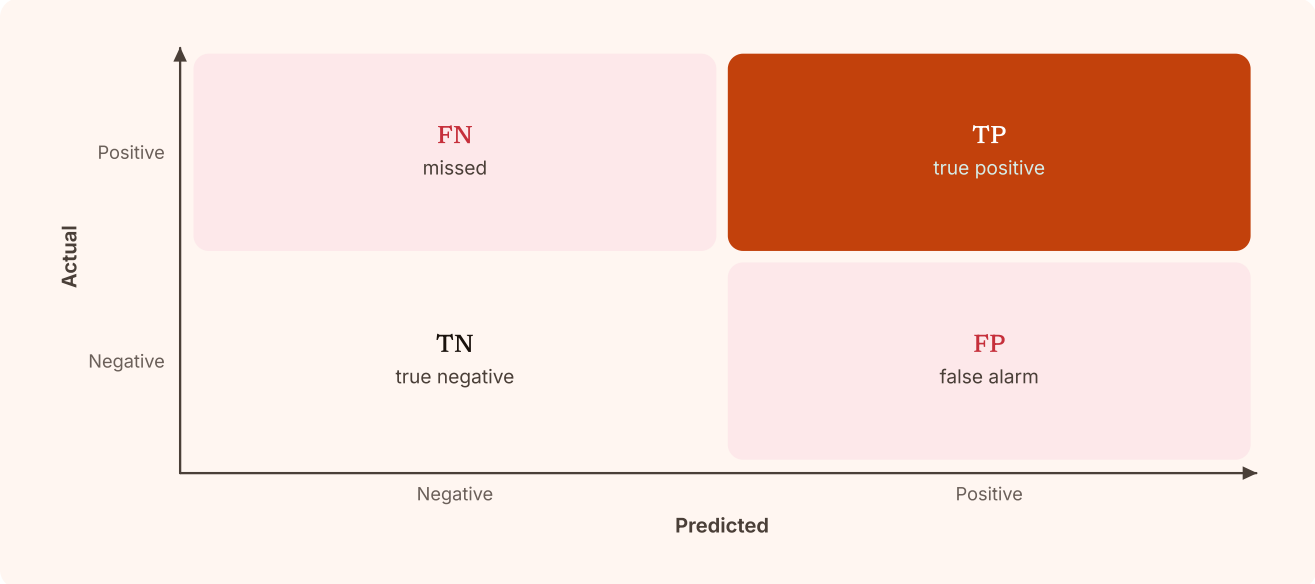


FIGURE 6.1 **Confusion matrix.** Accuracy = (TP + TN) / n. Sensitivity = TP / (TP + FN). Specificity = TN / (TN + FP). Precision = TP / (TP + FP).

Definition. You choose a threshold τ (often 0.5): predicted probability ≥ τ gives "positive". The **confusion matrix** counts TN, FN, FP and TP.

	Actual 0	Actual 1
Predicted 0	TN	FN
Predicted 1	FP	TP

Example (τ = 0.5): TN 356, FN 115, FP 12, TP 17.

Watch the order. R (**table**) puts 0 first, so TN is top left (table above). The practice exam 7.1 puts 1 first: then TP is top left. The same numbers in exam form:

	Actual 1	Actual 0
Predicted 1	17 (TP)	12 (FP)
Predicted 0	115 (FN)	356 (TN)

Always read the row and column headings first; if you expect TN top left you will point at the wrong cell.

Role of the threshold. A lower τ gives more positive predictions: sensitivity up, specificity down. At τ = 0.3 sensitivity is 0.58 against 0.13 at 0.5. If all predicted probabilities lie below τ (here max 0.66), there is no positive prediction at all and precision cannot be calculated (0/0). Choose τ on the costs of FN and FP; the coefficients and the AUC do not change.

07 Accuracy, sensitivity, specificity and precision

Measure	Formula	Example
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

Question and measure. "What share of the real positives do you find?" is sensitivity. "What share of the alerts is right?" is precision. "What share of the real negatives do you recognise?" is specificity.

Imbalanced classes. Accuracy is misleading when one class is large (practice exam 7.1 part 3). A model that always says "no" has 97% accuracy at 3% fraud and sensitivity 0. Compare with the baseline (share of the majority class; here 0.736) and look at sensitivity and precision.

08 ROC curve and AUC

The ROC curve plots sensitivity (vertical) against 1 – specificity (horizontal) for all thresholds. The **AUC** is the area under it: the probability that a random positive observation gets a higher predicted probability than a random negative one. AUC 0.5 is guessing, 1 is perfect separation. In R: **prediction()** and **performance()** from ROCR.

In the data: AUC 0.689 (train) and 0.714 (test). A model with a higher AUC on the same test set separates the classes better (practice exam 6.3). A lower AUC is not overfitting in itself.

PITFALL

AUC is not the share classified correctly and does not depend on a single threshold.

09 Overfitting, train/test split and validation

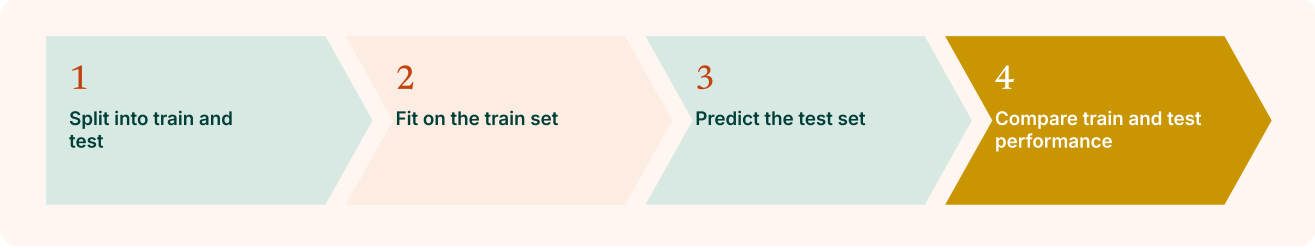


FIGURE 6.2 **Preventing overfitting with validation.** Judge a model on data it has not seen.

Definition. Overfitting: the model fits the noise in the training data and predicts new data worse. You split to judge how well the model generalises to new data (practice exam 6.6).

Procedure. Draw training rows at random (**sample()**), estimate on the training set, judge on the test set.

Example. Acceptance model: accuracy 0.729 (train) and 0.793 (test). That does not point to overfitting; the test set is small (150) and the naive baseline is 0.76. Fitness example (simulated): large model train 0.85, test 0.58; small model train 0.67, test 0.62. The large model is overfit.

Cross-validation. k-fold: k parts, each part is the test once. LOOCV is k = n. A better estimate of the test error than one split.

Bias-variance trade-off. More flexible: less bias, more variance. The training error always falls, the test error is U-shaped.

PITFALLS

Using the test set to choose the model (and then reporting that same score). Too few predictors is called underfitting, not overfitting.

10 Approaching exam question 7.1

Question 7.1 (9 points, 200 words) gives glm output and a confusion matrix and asks three things:

1. **Coefficients in business language** (about 4 points): sign, significance, odds ratio, ceteris paribus.
2. **Calculation** (about 2 points): accuracy and sensitivity or precision, with the formula and the numbers.

3 **Why accuracy misleads** (about 3 points): imbalanced outcome (baseline), low sensitivity, and the consequence for the company (which customers or cases you miss). Mention threshold or overfitting only if the case asks for it.

Write briefly: one sentence per point, numbers included.

MODEL ANSWER

→ How to write the answer

Four building blocks, in this order. Each one earns points; leave none out.

1. Term
Name the model or concept with the English term.

2. Definition
Give the definition or features, precisely.

3. Case fact
Link it to a concrete fact from the case.

4. Conclusion
End with a judgement, advice or consequence.

QUESTION 1 · WORD LIMIT 75 WORDS (WORD LIMIT ASSUMED)

A model scores 99% accuracy on the train set and 70% on the test set. What is going on?

TERM This is overfitting. DEFINITION The model adapts to the noise in the train set and therefore generalises badly to new data. CASE FACT The large gap between 99% on train and 70% on test shows it. CONCLUSION Judge the model on the test set or with cross-validation, and choose a simpler model.

48 / 75 words

KEY CONCEPTS

✓ Checklist: can I explain this?

Tick what you can explain from memory in one sentence, using the English 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

Test yourself

Cover the right-hand column and say the answer out loud.

Question	Answer
1 What is a binary dependent variable?	A dependent variable with two outcomes (0 or 1), such as cancelling or not cancelling.
2 How do you calculate odds from a probability p?	odds = p/(1 – p); at p = 0.20 the odds are 0.25.
3 What does an odds ratio of 1 mean?	No effect: the odds do not change when the predictor rises.
4 How does R estimate a logit model?	With maximum likelihood: the b's under which the observed data have the largest likelihood, iteratively.
5 What is the formula of the McFadden pseudo-R^2?	1 – ln L(model)/ln L(null), also calculated as 1 – residual deviance/null deviance.
6 How do you test whether an extra variable significantly improves the model?	With anova(m1, m2, test = "Chisq"): a likelihood-ratio test on the drop in deviance.
7 How do you calculate the precision?	TP/(TP + FP): of everything the model calls positive, the share that is really positive.
8 What is the baseline with imbalanced classes?	The share of the majority class: the accuracy of a model that always predicts that class.
9 When do you choose a low threshold?	When a missed case costs much more than a false alarm.
10 What does AUC = 0.5 mean?	The model separates the classes no better than chance.
11 Which measures help against overfitting?	Simplify the model, collect more data and test the choice with cross-validation.
12 How does k-fold cross-validation work?	The data goes into k parts; each part is the test set once, the other k – 1 train. The scores are averaged.