Module 13: Logistic Regression

Learning outcomes

1. Compare linear and logistic regression to determine their suitability for predictive modelling tasks.
2. Interpret how changing the parameters of a logistic function affects model behaviour and performance.
3. Optimise a logistic regression model in Python by adjusting the decision threshold to balance false positive and false negative rates.
4. Analyse the advantages and trade-offs of using logistic regression compared to other classification methods in a domain-specific context.
5. Analyse logistic regression outputs and modelling decisions to evaluate performance, interpret significance and compare trade-offs across use cases.

Logistic regression

- A binary classification algorithm that estimates the probability of an outcome belonging to a specific class (e.g. diabetes: yes/no)
- Outputs probabilities between 0 and 1 using the sigmoid function
- Commonly used when interpretability and simplicity are important

Comparison: Linear vs logistic regression

Aspect	Linear regression	Logistic regression
Output	Continuous values	Probabilities (0 to 1)
Use case	Regression problems	Classification problems
Assumption	Linear relationship with output	Linear relationship with log odds
Output interpretation	Direct numeric prediction	Probability used for binary decision
Fitting method	Least squares	Maximum likelihood

Threshold selection

- Logistic regression outputs probabilities. The use of a threshold (often 0.5) can convert this into a binary decision.
- The choice of threshold affects:
  - **False positive rate (FPR):** Class 0 incorrectly classified as 1

  - **False negative rate (FNR):** Class 1 incorrectly classified as 0
- Lowering the threshold:
  - Increases recall (fewer false negatives)
  - Raises false negatives
- Raising the threshold:
  - Increases precision (fewer false positives)
  - Increases false negatives

Regularisation

- Helps prevent overfitting by penalising large coefficients
- Two types commonly used:
  - ***L1*** (lasso): shrinks some coefficients to 0 (feature selection)
  - ***L2*** (ridge): shrinks all coefficients but keeps them non-zero (simpler models)
- Strength is controlled by `C` in `LogisticRegression`. Smaller `C` = stronger regularisation

Interpreting coefficients in multi-predictor models

- Each coefficient shows the effect of a predictor while holding others constant.
- Significance may change with the addition of other predictors:
  - A variable might appear important alone, but it loses significance with the inclusion of other variables.
  - Always interpret in the context of the full model.
- Use z-scores and p-values (via `statsmodels`) to evaluate statistical significance.