Computational
Social Science

Logistic Regression Fundamentals .III

Roberto Cerina

27.02.2024

University of Amsterdam

Predicted Probability & Classification

- *Point estimates* of predicted values:

➾ Logistic Regression can be used to predict the probability of an event happening, conditional on a set of covariates:

$$\hat{\pi} = \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p)$$

➾ From the probability, we can predict the *class* of an observation/subject conditional on the same set of covariates:

$$\hat{y} = \begin{cases} 0 & \text{if} < \tau \\ 1 & \text{if} \geq \tau \end{cases}$$

- where τ is an arbitrary threshold to indicate an optimal cutoff point, typically set to 0.5 as default.

Measures of Performance: Predicted Probability / Risk

- **Brier Score**:

$$\widehat{BS} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\pi}_i - y_i)^2$$

- if an event ends up happening (e.g. Y = 1) it should have an associates probability of 1, and vice-versa . . .
- . . . average difference between the predicted probability and this theoretical probability is a measure of the average error on the probability scale;
- . . . simply the RMSE, but for probability-scale predictions and binary outcomes.

Measures of Performance: Predicted Class

✏ Confusion Matrix

		Predicted		
		$\hat{y}_i = 1$	$\hat{y}_i = 0$	
Observed	$y_i = 1$	True Positives	False Negatives	TP Rate (Sensitivity/Recall): $\frac{\text{TP}}{\text{TP}+\text{FN}}$
	$y_i = 0$	False Positives	True Negatives	TN Rate (Specificity): $\frac{\text{TN}}{\text{TN}+\text{FP}}$
		Precision: $\frac{\text{TP}}{\text{TP}+\text{FP}}$	Neg. Pred. Value: $\frac{\text{TN}}{\text{TN}+\text{FN}}$	Accuracy: $\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$

Measures of Performance: Predicted Class

☛ Nominal Predictions: *Confusion Matrix*

- ✏ Sensitivity (Recall): % True Positives out of all Observed Positives ;
- ✏ Specificity: % True Negatives out of all Observed Negatives ;
- ✏ Precision (PPV): % True Positives out of all Predicted Positives ;
- ✏ NPV: % True Negatives out of all Predicted Negatives ;
- ✏ Accuracy: % Correctly Predicted over all Observations ;

Other useful metrics

- *No Information Rate* (NIR):
  - Accuracy if pred. value is the the modal category in the outcome ;
  - $\hat{y}_i = \left(\frac{1}{n}\sum_i^n y_i > 0.5\right)$ for binary outcome . . .
- *Balanced Accuracy* :
  - Accounts for imbalance in the sample ;
  - $\frac{1}{2}(\text{sensitivity} + \text{specificity})$

Other useful metrics

- Cohen's Kappa:

$$\kappa = \frac{\text{accuracy gains from our model relative to change}}{\text{accuracy gains from a perfect model relative to chance}} = \frac{p_0 - p_e}{1 - p_e}$$

$$\kappa \in [-1, 1]$$

  - $\kappa = 1$ indicates perfect agreement between the predictions and the observations;
  - $\kappa = 0$ indicates no agreement beyond that expected by chance;
  - $\kappa = -1$ indicates perfect disagreement between the predictions and the observations.

$$\text{Accuracy} = \text{Pr(Obs = Pred)}$$

$$\text{Model Accuracy} = \hat{p}_0 = \widehat{\text{Pr}}(\text{Obs} = 1, \text{Pred} = 1) + \widehat{\text{Pr}}(\text{Obs} = 0, \text{Pred} = 0) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

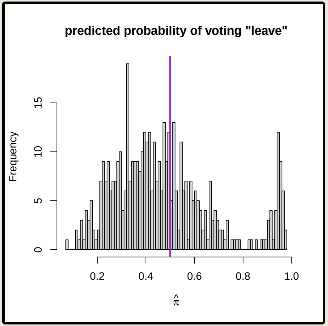
by independence of Predictions & Observations

$$\text{Chance Accuracy} = \hat{p}_e = \widehat{\text{Pr}}(\text{Obs} = 1)\widehat{\text{Pr}}(\text{Pred} = 1) + \widehat{\text{Pr}}(\text{Obs} = 0)\widehat{\text{Pr}}(\text{Pred} = 0)$$

$$= \frac{(\text{TP + FN})}{N}\frac{(\text{TP + FP})}{N} + \frac{(\text{TN + FP})}{N}\frac{(\text{TN + FN})}{N}$$

Measuring & Visualising Prediction Error

- Threshold τ is typically set to 0.5 by default...
- this is not always optimal to separate events events according to their risk...

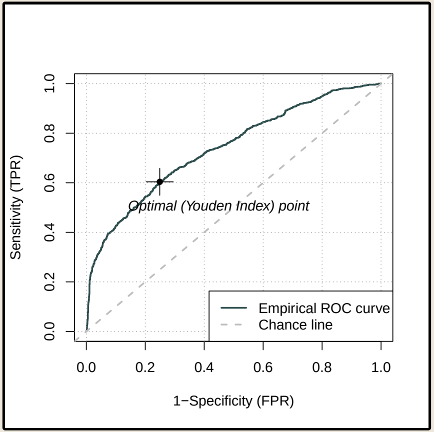


- Changing the threshold for classification in a prediction model will change the confusion matrix. How to choose the optimal threshold ?

Choosing an Optimal Threshold

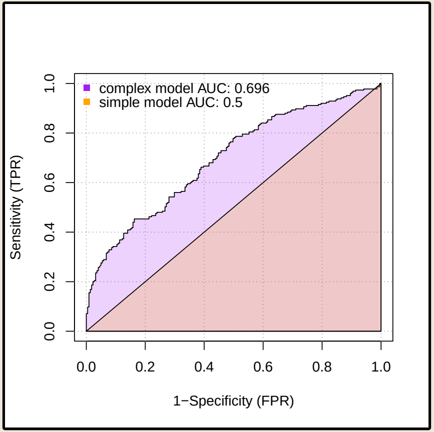
✏ ROC curve (TPR v. FPR) over different threshold values.

- The ‘Youden Index’ point identifies the threshold with the best predictive power.



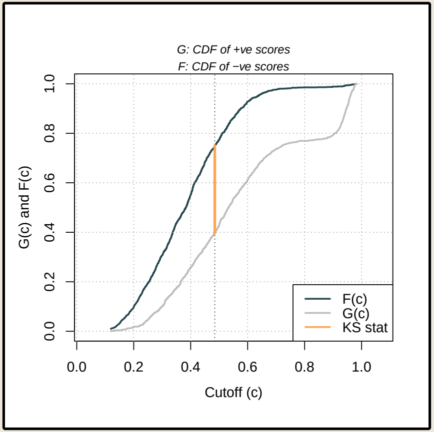
Choosing an Optimal Threshold

- Area Under the Curve (AUC) is a measure of the overall 'goodness' of a classifier.
- An optimal model will have its ROC curve go through $(0, 1)$, hence displaying AUC $= 1$. AUC can be used to discriminate between models.



Choosing an Optimal Threshold

- Kolmogorov-Smirnov method is used to identify optimal threshold value ;
- Kolmogorov-Smirnov distance is a measure of *purity* of the classification.



Choosing an Optimal Threshold

❶ Order my observations based on the predicted probability, from lowest to largest;

❷ For every candidate threshold, from lowest to highest:

i. calculate the proportion of the instances in of $y_i = 1$ below the threshold (CDF of +ve scores);
ii. calculate the proportion of the instances in of $y_i = 0$ below the threshold (CDF of -ve scores);

❸ Plot the CDFs against the threshold values;

❹ Plot the CDFs against the threshold values;

- If a threshold achieves perfect separation : $\widehat{KS} = 1$;
- Chance model: $\widehat{KS}$ = 0

Point Estimates v. Distribution of Errors

✏ Using *point estimates* for predicted values will lead to point estimates of the error...

✒ E.g. **Brier Score**:

$$\widehat{BS} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\pi}_i - y_i)^2$$

✏ This is fine if the goal is to discriminate between models:

- need to make a decision as to which model to deploy;
- the 'error which we are most likely to see' from each model is generally good enough to justify the choice.

✏ If our goal is to provide a true estimate of the *Generalisation Error*:

- Point estimate of the error is of interest;
- 'Worst-case' and 'Best-case' scenarios are also of interest for planning !
- We need to incorporate **uncertainty around the generalisation error**.

Simulating the Generalisation Error

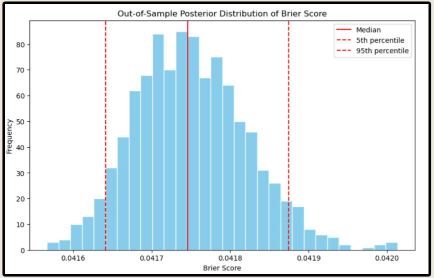
- Remember generalisation error is the expected error over a set of L 'new' or 'unseen' data points: $X_l^\star = [x_{l1}^\star, ..., x_{lp}^\star]$
- Typically we have this new data in a test-set, so we also know the respective outcomes associated with this new data: $y_l^\star$
- From the posterior distribution of $\boldsymbol{\beta}$, simulate values for predicted probabilities $\tilde{\pi}$, and predicted outcomes $\tilde{y}$:

  $$\boldsymbol{\beta}^s \sim N(\hat{\boldsymbol{\beta}}_{MLE}, \hat{\boldsymbol{\Sigma}})$$
  $$\tilde{\mu}_l^s = \beta_0^s + \beta_1^s x_{l1}^\star + ... + \beta_p^s x_{lp}^\star$$
  $$\tilde{\pi}_l^s = \frac{\exp(\tilde{\mu}_l^s)}{1+\exp(\mu_l^s)}$$
  $$\tilde{y}_l^s \sim \text{Bernoulli}(\tilde{\pi}_l^s)$$

- For each simulation, calculate the relevant error metric (e.g. Brier Score):
  $\widetilde{BS}^s = \frac{1}{L}\sum_{l=1}^{L}(\tilde{\pi}_l^s - y_l^\star)^2$

Simulating the Generalisation Error

- Histogram of $\widetilde{BS}_{1:S}$ will give you the estimated posterior distribution of the generalisation error...
- you can use Monte Carlo methods to extract summary statistics for the generalisation error, such as its median (expected error) and 90% interval (typical 'best' and 'worst' case scenarios).



Simulating the Generalisation Error

- **Final Note**: you can do this with any error – E.g. you can have a distribution of ‘confusion matrices’ so that you can derive a ‘worst and best case scenario‘ for every entry in the matrix.