STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee

UNIVERSITY OF TORONTO

February 27, 2018

Using Logistic Regression for Classification

Using Logistic Regression for Classification

- Want: predict outcome as

$$y^*|(x_1^*, x_2^*, \ldots, x_p^*) = \begin{cases} 1 \\ 0 \end{cases}$$

- Do: calculate $\widehat{\pi}_M^*$- the estimated probability that $y^* = 1$ based on the fitted model given $X_1 = x_1^*, X_2 = x_2^*, \ldots, X_p = x_p^*$. From this we want to predict that

$$y^* = \begin{cases} 1 & \text{if } \widehat{\pi}_M^* \text{ is large} \\ 0 & \text{if } \widehat{\pi}_M^* \text{ is small} \end{cases}$$

- Need: choose a cut-off probability to distinguish between large and small.

Classification: Approaches to choosing a threshold

Approach 1 - Set cut-off probability as 0.5

- If $\widehat{\pi}^*_M > 0.5$, classify y^* as 1
- Useful if there are equal numbers of 1's and 0's
- Useful if false negatives and false positives are equally bad.

Classification: Approaches to choosing a threshold

Approach 2- Find "best" cut-off probability from data.

- Try different cut-offs and see which gives fewest incorrect classifications
- Useful if proportions of 1's and 0's in data reflect their relative proportions in the population
- Likely to overestimate the proportions of correct predictions that model makes. Then, one should assess model correct classification rates on different data than was used to fit the model.

Confusion Matrix

	Truth		
Prediction	Positive ($Y = 1$)	Negative ($Y = 0$)	Prop (row)
Positive	TP	FP	PPV=$\frac{TP}{TP+FP}$
Negative	FN	TN	NPV=$\frac{TN}{TN+FN}$
Prop (column)	Sensitivity= TPR=$\frac{TP}{TP+FN}$	Specificity= TNR=$\frac{TN}{TN+FP}$	

- TP: true positive; TN: true negative
- FP: false positive (type I error); FN: false negative (type II error)
- PPV: precision or positive predictive value; false discovery rate=1-PPV
- NPV: negative predictive value; false omission rate=1-NPV

1. Sensitivity (True Positive Rate, TPR)- hit rate
2. Specificity (True Negative Rate, TNR)- prop. of correctly classified negatives
3. False Positive Rate, FPR=1-TNR, fall-out rate
4. False Negative Rate, FNR=1-TPR, miss rate
5. Classification rate=(TN+TP)/(TP+FN+FP+TN); accuracy

Diagnostic Accuracy

- Choose a cut-off probability based on one of the 5 criteria for success of classification that is most important to you.

- High Sensitivity (TPR) makes good screening test.
- High Specificity (TNR) makes a good confirmatory test.
- A screening test followed by a confirmatory test is a good (but expensive) diagnostic procedure.

Confusion Matrix

- From Wikipedia

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$

`https://en.wikipedia.org/wiki/Confusion_matrix`

Least Squares Regression vs Logistic Regression

	(Ordinary) Least Squares	(Binomial) Logistic
Response, Y	Normal	# of successes in m trials
Variance	Equal for each level of X	$mp(1-p)$ for each level of X
Model	$\mu_y = \beta_0 + \beta_1 X$	$\log(\frac{\mu}{1-\mu}) = \beta_0 + \beta_1 X$
Model fitting	Least Squares	MLE
Exploratory plot	X vs Y (add line)	logit vs X
Comparing models	Partial F-test AIC/BIC Residuals	LRT/Deviance tests AIC/ BIC (Pearson, Deviance) Residuals
Interpreting	β_1: **change in μ_y for unit change in X**	e^{β_1}: **% change in odds for unit change in X**