Logistic Regression your first classifier

Andrew McCarren
L235 ext 8456
Andrew.mccarren@dcu.ie

What is Logistic Regression?

- Form of regression that allows the prediction of discrete variables by a mix of continuous and discrete predictors.

- Addresses the same questions that discriminant function analysis and multiple regression do but with no distributional assumptions on the predictors (the predictors do not have to be normally distributed, linearly related or have equal variance in each group)

What is Logistic Regression?

- Logistic regression is often used because the relationship between the DV (a discrete variable) and a predictor is non-linear

  - The probability of heart disease changes very little with a ten-point difference among people with low-blood pressure, but a ten point change can mean a drastic change in the probability of heart disease in people with high blood-pressure.

Background

- Odds – like probability. Odds are usually written as “4 to 1 odds” which is equivalent to 1 out of five or .2 probability or 20% chance, etc.
  - The problem with probabilities is that they are non-linear
  - Going from .10 to .20 doubles the probability but going from .80 to .90 barely increases the probability.

Background

- Odds ratio – the ratio of the odds over 1 – the odds. The probability of winning over the probability of losing. 4 to 1 odds equates to an odds ratio of .20/.80 = .25.

Background

- Logit – this is the natural log of an odds ratio; often called a log odds even though it really is a log odds ratio. The logit scale is linear and functions much like a z-score scale.

Background

LOGITS ARE CONTINOUS, LIKE Z SCORES

p = 0.50, then logit = 0

p = 0.70, then logit = 0.84

p = 0.30, then logit = -0.84

Plain old regression

- Y = A BINARY RESPONSE (DV)
  - 1 POSITIVE RESPONSE (Success) →P
  - 0 NEGATIVE RESPONSE (failure) →Q = (1-P)
- MEAN(Y) = P, observed proportion of successes
- VAR(Y) = PQ, maximized when P = .50, variance depends on mean (P)
- X_J = ANY TYPE OF PREDICTOR → Continuous, Dichotomous, Polytomous

Plain old regression

$$Y \mid X = B_0 + B_1 X_1 + \varepsilon$$

- and it is assumed that errors are normally distributed, with mean=0 and constant variance (i.e., homogeneity of variance)

Plain old regression

$$E(\hat{Y} \mid X) = B_0 + B_1 X_1$$

- an expected value is a mean, so

$$(\hat{Y} = \hat{\pi}) = P_{Y=1} \mid X$$

- The predicted value equals the proportion of observations for which Y|X = 1; P is the probability of Y = 1(A SUCCESS) given X, and Q = 1- P (A FAILURE) given X.

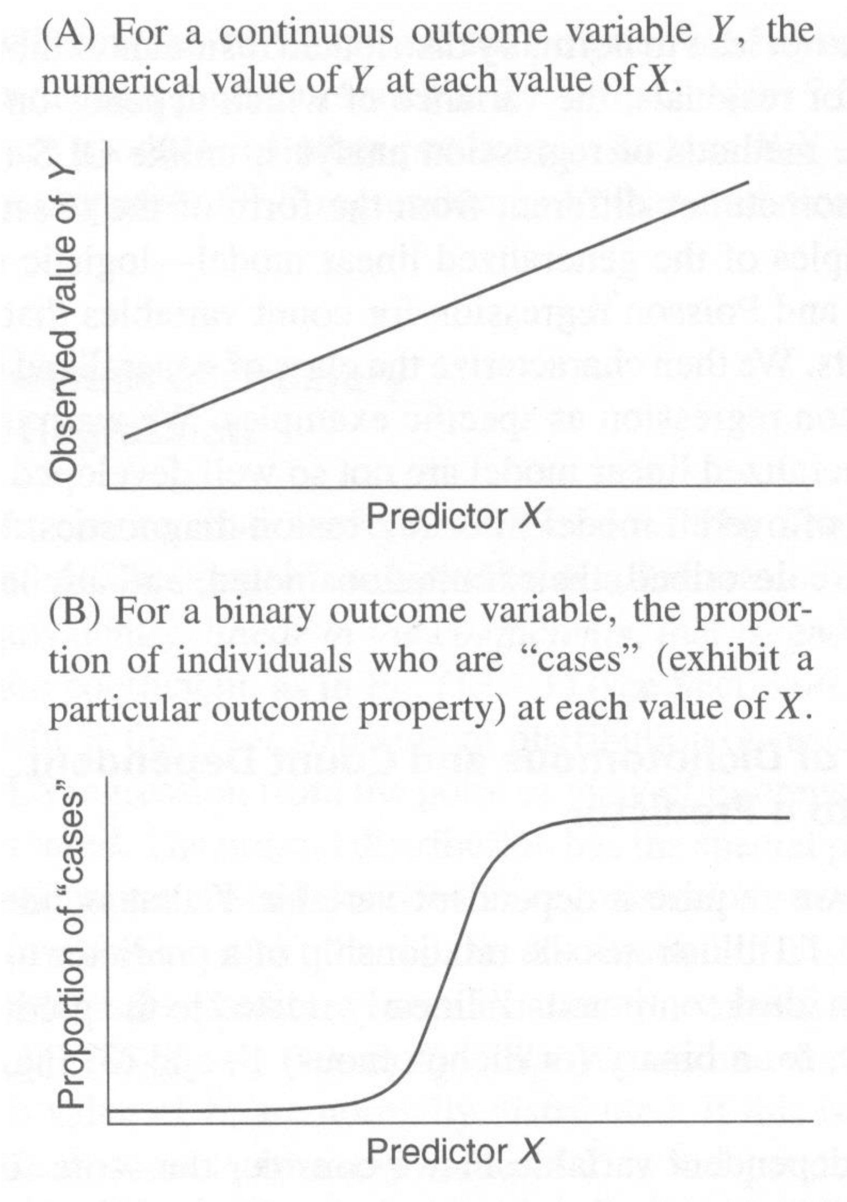
Plain old regression

- Every respondent is given a probability of success and failure which leads to every person having drastically different variances (because they depend on the mean in discrete cases) causing a violation of the homoskedasticity assumption.

Plain old regression

- Long story short – you can’t use regular old regression when you have discrete outcomes because you don’t meet homoscedasticity rules.

The logistic function



(A) For a continuous outcome variable Y, the numerical value of Y at each value of X.

(B) For a binary outcome variable, the proportion of individuals who are "cases" (exhibit a particular outcome property) at each value of X.

The logistic function

$$\widehat{Y_i} = \frac{e^u}{1+e^u}$$

- Where Y-hat is the estimated probability that the ith case is in a category and u is the regular linear regression equation:

$$u = A + B_1 X_1 + B_2 X_2 + \cdots + B_K X_K$$

The logistic function

$$\hat{\pi}_i = \frac{e^{b_0 + b_1 X_1}}{1 + e^{b_0 + b_1 X_1}}$$

The logistic function

- Change in probability is not constant (linear) with constant changes in X

- This means that the probability of a success (Y = 1) given the predictor variable (X) is a non-linear function, specifically a logistic function

The logistic function

- It is not obvious how the regression coefficients for X are related to changes in the dependent variable (Y) when the model is written this way

- Change in Y(in probability units)|X depends on value of X. Look at S-shaped function

The Logit

- By algebraic manipulation, the logistic regression equation can be written in terms of an <u>odds ratio for success</u>:

$$\left[\frac{P(Y=1|X_i)}{(1-P(Y=1|X_i))}\right]=\left[\frac{\hat{\pi}}{(1-\hat{\pi})}\right]=\exp(b_0+b_1X_{1i})$$

The Logit

- Finally, taking the natural log of both sides, we can write the equation in terms of logits (log-odds):

$$\ln\left[\frac{P(Y=1|X)}{(1-P(Y=1|X))}\right]=\ln\left[\frac{\hat{\pi}}{(1-\hat{\pi})}\right]=b_0+b_1X_1$$

For a single predictor

The Logit

$$\ln\left[\frac{\hat{\pi}}{(1-\hat{\pi})}\right]=b_0+b_1X_1+b_2X_2\ldots+b_kX_k$$

- For multiple predictors

Assumptions

- The only “real” limitation on logistic regression is that the outcome must be discrete.

Assumptions

- Ratio of cases to variables – using discrete variables requires that there are enough responses in every given category
  - If there are too many cells with no responses parameter estimates and standard errors will likely blow up

Assumptions

- Linearity in the logit – the regression equation should have a linear relationship with the logit form of the DV. There is no assumption about the predictors being linearly related to each other.

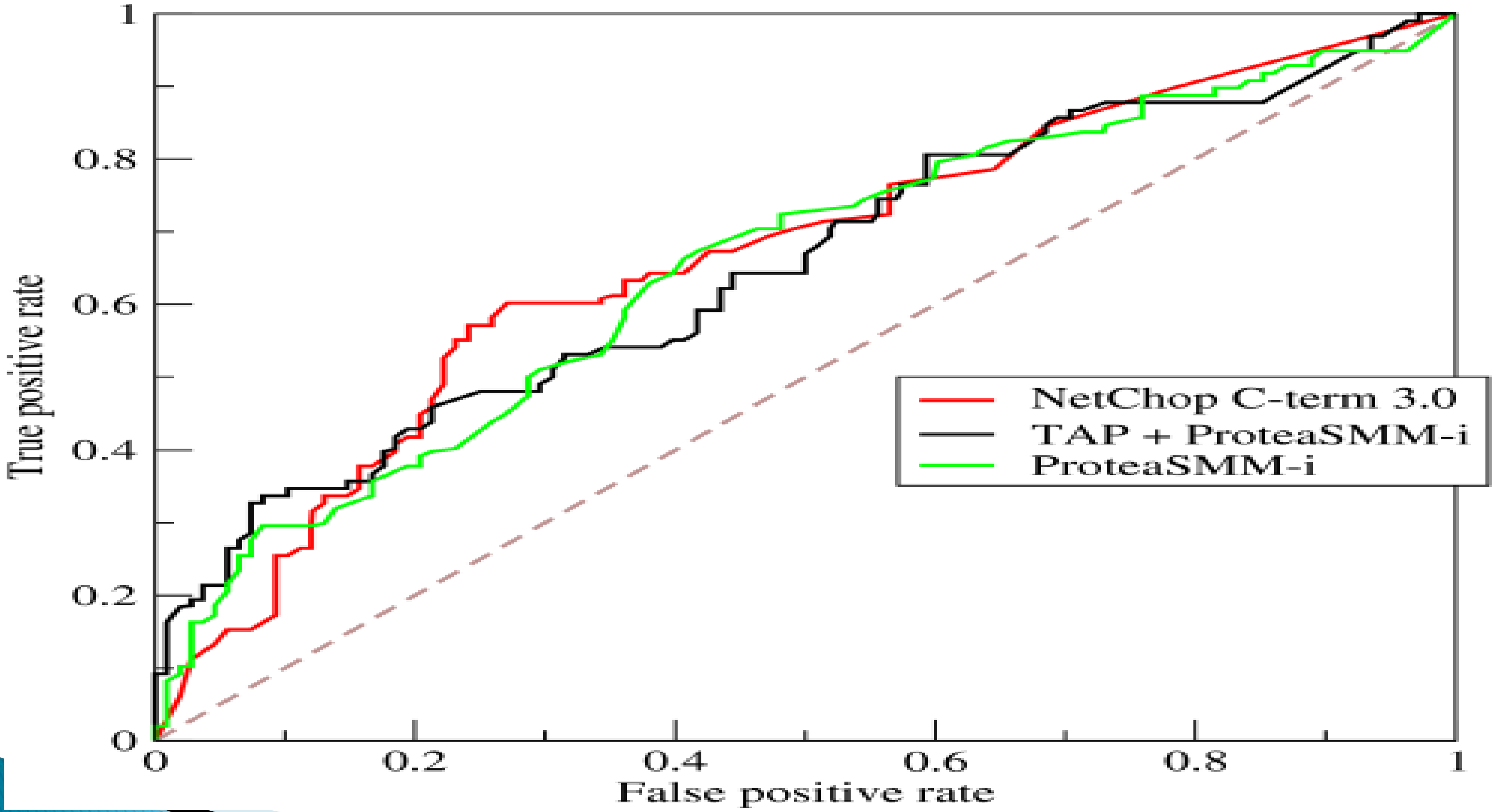
Assumptions

- Absence of multicollinearity
- No outliers
- Independence of errors – assumes a between subjects design. There are other forms if the design is within subjects.

Assessing logistic regression

- Check for Multi-Collinearity
- R-Squared Adjusted
- Other assessments such as:
  - AIC
  - BIC
- True Positive rate vs False Positive rate
  - ROC curve
  - Area under the curve

- ROC Curve Cont.
  - Measure the area under the ROC curve
    - Poor fit – area under the ROC curve approximately equal to 0.5
    - Good fit – area under the ROC curve approximately equal to 1.0


True positive rate
False positive rate
NetChop C-term 3.0
TAP + ProteaSMM-i
ProteaSMM-i