Module 10: Logistic Regression 2

Jeffrey Woo

MSDS, University of Virginia

Diagnostics for Logistic Regression

Model diagnostics are not as available for logistic regression as with linear regression.

- The Pearson and deviance goodness of fit tests do not work with ungrouped data.
- Plotting the log odds versus the predictor only works with 1 predictor (or maybe 2 if you want to produce a 3d plot). But cannot do in higher dimensions.
- Examining a plot of residuals does not help due to the discrete nature (0,1) of the response variable in logistic regression.

Next best thing: evaluate the performance of the logistic regression in terms of how well it does in predicting outcome (model validation).

Comments on Confusion Matrix

A decision rule is how we decide to classify predictions. For example, if $\hat{\pi} > 0.5$, classify as 1. Otherwise classify as 0. The value 0.5 is called the threshold.

- The values in the confusion matrix depend on the cutoff value.
- In a binary setting, a cutoff value of 0.5 gives the highest accuracy / lowest overall error rate, on average.
- Depending on the context, we may be more interested in ensuring a high sensitivity or high specificity, so we may change the cutoff value.

ROC Curve

- Plots true positive rate against false positive rate for varying cutoff values.
- If logistic regression does no better than random guessing (i.e. does not use info provided by predictors), then the true positive rate is equal to the false positive rate, i.e.
  - If you randomly guess $k \times 100\%$ of observations to be positive, then among all true positives, $k\%$ of them will be classified as positive: $P(\hat{Y} = 1|Y = 1) = k \times 100\%$.
  - If you randomly guess $k \times 100\%$ of observations to be positive, then among all true negatives, $k\%$ of them will be classified as positive: $P(\hat{Y} = 1|Y = 0) = k \times 100\%$.
  - $P(\hat{Y} = 1) = P(\hat{Y} = 1|Y = 1) = P(\hat{Y} = 1|Y = 0) = k \times 100\%$.
- Hence ROC curve along the diagonal indicates a model that does no better than random guessing.

Cautions

- Accuracy can be a misleading measure especially when you have unbalanced sample sizes of the two classes.
- The ROC curve shows the true positive and false positive rates as the threshhold is varied. It does not immediately inform you of the true positive and false positive rates for your specific threshold.
- The AUC, just like the ROC, is a summary of the predictive performance for all possible values of the threshold. It does not inform you of the accuracy for your specific threshold.

Multinomial Logistic Regression

- Important to note which class is the reference class.
- The log relative risk is modeled as a linear combination of coefficients and predictors, i.e. $\log \frac{\pi_c}{\pi_{ref}} = \beta_{c,0} + \beta_{c,1} X_1 + \beta_{c,k} X_k$.
- We have separate equations for the relative risk of each class versus the reference class. For a response variable with $m + 1$ classes, there will be m logits. Each logit models the the log relative risk of belonging to a class vs the reference class as a linear combination of the coefficients and predictors.

Generalized Linear Models (GLMs)

- GLMs are a family of regression models. Their coefficients are estimated using the method of maximum likelihood.
- We have estimated the linear regression model using ordinary least squares. The linear regression model can also be estimated via the method of maximum likelihood. Both methods converge asymptotically.
- Logistic regression is a specific kind of GLM.
- Another common GLM is Poisson regression, which is used to model count data (response variable is a nonnegative integer).
- For any GLM, we can use
    - the Wald test to drop a single term
    - the difference in null and residual deviances to test if a model is useful
    - the difference in residual deviances to compare nested models