Applied Statistical Methods II

Chapter #14

Logistic Regression, Poisson Regression, and Generalized Linear Models

Part IV

Prediction with logistic regression:

- In linear regression, we encountered two types of problems: inference for mean function/parameters and prediction of a new observation.
- Same issues hold in logistic regression:
  - Thus far we have focused on inference for mean function/parameters.
  - Today: prediction of a new observation.
- For binary responses, prediction takes the form of classification.
  - Given some covariates values, classify the predicted response into one of two categories
  - There is an entire subfield for classification analysis.
  - A fitted logistic regression model provides one simple approach to the problem.

Idea Behind Classification

- Formulate some rule so that, based on some covariates **X**, we predict either $\hat{Y} = 1$ or $\hat{Y} = 0$.
- Using logistic regression: fit a model from some test data and use the predicted probabilities to determine a prediction rule.
  - Exact prediction rule depends on loss function
  - Predict $\hat{Y}_i = 1$ if $\hat{\pi}_i > C$ for some cut-off-value $0 < C < 1$.
- Then evaluate the classification performance with some criterions.

“Good" Prediction

	Classified	
True	G0	G1
G0	n_{00}	n_{01}
G1	n_{10}	n_{11}

- Consider two aspects of prediction.
  - Sensitivity = $P(\hat{Y} = 1 | Y = 1) \approx n_{11}/(n_{10} + n_{11})$ (maximize true positives)
  - Specificity = $P(\hat{Y} = 0 | Y = 0) \approx n_{00}/(n_{00} + n_{01})$ (minimize false positives)
- In an ideal world, you want both of these to be big.
- But big values of one leads to small values of the other.
  - $C \to 0$ means that sensitivity $\to 1$ but specificity $\to 0$.
  - $C \to 1$ means that sensitivity $\to 0$ but specificity $\to 1$.

"Good" Prediction Cont.

- Misclassification rate for G_1
  $P(\hat{Y} = 0|Y = 1) \approx n_{10}/(n_{10} + n_{11})$ (i.e. type II errors)
- Misclassificaiton rate for G_0
  $P(\hat{Y} = 1|Y = 0) \approx n_{01}/(n_{00} + n_{01})$ (i.e. type I errors)
- Overall misclassification rate, let $n = n_{10} + n_{11} + n_{00} + n_{01}$

$$
\begin{aligned}
& P(\hat{Y} \neq Y) \\
=& P(\hat{Y} = 0|Y = 1)P(Y = 1) + P(\hat{Y} = 1|Y = 0)P(Y = 0) \\
\approx& \frac{n_{10}}{n_{10} + n_{11}} \frac{n_{10} + n_{11}}{n} + \frac{n_{01}}{n_{00} + n_{01}} \frac{n_{00} + n_{01}}{n} \\
=& \frac{n_{10} + n_{01}}{n}
\end{aligned}
$$

Minimize the overall misclassification rate

- If the criterion is to minimize the overall misclassification rate, the optimal solution only depends on the ratio of the conditional probabilities.
- The optimal classification rule is: If $\frac{\hat{P}(Y=1|X)}{\hat{P}(Y=0|X)} > 1$, classify as G_1, otherwise G_0.
- In logistic regression, predict that $Y_h = 1$ if $\hat{\pi}_h > .5$, or equivalently, $\log \frac{\hat{\pi}_h}{1-\hat{\pi}_h} > 0$, or equivalently $X_h^T \beta > 0$.

Other Classification Rules

- Is the $C = 0.5$ always good?
- Theoretically $C = 0.5$ is the optimal solution to **minimize overall misclassification error rate**, if we assume that 1) We have a simple random sample 2) and the logistic regression model for the conditional probability is correct.
- In practice, it also makes sense to calibrate C using the empirically estimated misclassification rate. Implicitly, this serves as a model adjustment.

Other Classification Rules Cont.

- Overall misclassification rate might not be the best criterion to use in some applications.
- In some rare event studies, $\pi_1 = P(Y = 1)$ could be small. The approaches targeted at overall misclassification criterion tends to ignore the minority class and have low sensitivity.
- In some applications, we want to have high sensitivity
  - Cancer screening: willing to accept some false positives
- In some applications, we want to have high specificity.
  - Fraudulent sellers on eBay: eBay does not want to falsely accuse someone of nefarious activity.
- Minimize a loss function: $P(\hat{Y} = 1|Y = 0)P(Y = 0)C(1|0) + P(\hat{Y} = 0|Y = 1)P(Y = 1)C(0|1)$, where $C(1|0)$ and $C(0|1)$ are the cost functions for misclassification in G_0 and G_1.

Other Classification Rules Cont.

- Overall misclassification rate might not be the best criterion to use in some applications.
- In some rare event studies, $\pi_1 = P(Y = 1)$ could be small. The approaches targeted at overall misclassification criterion tends to ignore the minority class and have low sensitivity.
- In some applications, we want to have high sensitivity
  - Cancer screening: willing to accept some false positives
- In some applications, we want to have high specificity.
  - Fraudulent sellers on eBay: eBay does not want to falsely accuse someone of nefarious activity.
- Minimize a loss function: $P(\hat{Y} = 1|Y = 0)P(Y = 0)C(1|0) + P(\hat{Y} = 0|Y = 1)P(Y = 1)C(0|1)$, where $C(1|0)$ and $C(0|1)$ are the cost functions for misclassification in G_0 and G_1.

Other Classification Rules Cont.

- Overall misclassification rate might not be the best criterion to use in some applications.
- In some rare event studies, $\pi_1 = P(Y = 1)$ could be small. The approaches targeted at overall misclassification criterion tends to ignore the minority class and have low sensitivity.
- In some applications, we want to have high sensitivity
  - Cancer screening: willing to accept some false positives
- In some applications, we want to have high specificity.
  - Fraudulent sellers on eBay: eBay does not want to falsely accuse someone of nefarious activity.
- Minimize a loss function: $P(\hat{Y} = 1|Y = 0)P(Y = 0)C(1|0) + P(\hat{Y} = 0|Y = 1)P(Y = 1)C(0|1)$, where $C(1|0)$ and $C(0|1)$ are the cost functions for misclassification in G_0 and G_1.

ROC Curves

- ROC = Receiver Operating Characteristic
- ROC curve plots estimated Sensitivity (y-axis) vs. 1 - Specificity (x-axis)
- ROC curve plots true positive rate vs. false positive rate
- Each point on the ROC curve corresponds to one cut-off value C
  - If $C = 0$, then we always set $\hat{Y} = 1$. Perfect sensitivity, poor specificity
  - If $C = 1$, then we always set $\hat{Y} = 0$. Perfect specificity, poor sensitivity
- The point (0,1) corresponds to a perfect classification.

ROC-based selection of the cut-off C

- If you care for sensitivity more at the cost of sacrificing specificity, you can choose a small value of C to achieve the level of sensitivity.
- If you care for specificity more, a larger value of C may be a good choice.
- If you do not have any strong preference, you can choose the point on the ROC curve that is closest to $(0, 1)$.

Areas under ROC curves

- The non-discriminant line: the diagonal line. The Area under is 0.5. Basically X has no discriminant power for the value of Y.
- Area under the curve describes predictive power
  - Area can be interpreted as the concordant probability: $P(\hat{\pi}_i > \hat{\pi}_j | Y_i = 1, Y_j = 0)$
  - A one number summary for ROC curves. Interpretation needs to be very careful.
  - A coin flip has oncordant probability 0.5.
- The graph of ROC is often more useful. It reflects the trade-off between sensitivity and specificity.

Validation and Evaluation of Performance

- Assume that you fit a model to determine a classification rule from Data Set A.
- You then receive Data Set B which is sampled from the same population.
- In practice, you will have worse classification on Dataset B.
  - The logistic model was trained on Data Set A, and the cut-off value C might also be calibrated by the empirical performance on data set A.
  - Especially if the dimension of covariates is high, we could have over fitting problem.
- Often would like to have a separate validation data set to determine the predictive power.
- To compare different classification methods, it has to be on a testing data set or use cross-validation at least.

Multinomial Regression

- What if Y_i is polytomous (i.e. > 2 categories)?
- Y_i can take on $1, \ldots, J$.
- The distribution for Y_i is now a multinomial:
  - $P(Y_i = j) = \pi_j$ for $j = 1, \ldots, J$.
  - $\sum_{i=1}^{J} \pi_j = 1$
- Can think of two types of polytomous data: nominal and ordinal.
- Nominal: No metric for possible values of Y.
  Is a car a American, European, or Asian?
- Ordinal: values of Y have an ordering.
  Severity of head injury on a scale from 1-4.

Categorical Data: baseline odds model

- Given some covariates, let $\pi_{ij} = P(Y_i = j|X_i)$
- Will choose a reference group (lets say J), and model for j=1,...,J-1

$$\log(\frac{\pi_{ij}}{\pi_{iJ}}) = X\beta_j$$

- From these J-1 log odds, you can get any log odds.
- Note that there are J-1 different parameter estimates for each covariate.
- Estimate through maximum likelihood.

Interpretation of the parameters

- Let β_{jk} be kth element of β_j.
- If all other covariates are held fixed, an increase in X_{ik} by one unit will result in an increase in the log odds of $Y_i = j$ vs $Y_i = J$ by β_{jk}.
  - aka the log $\frac{P(Y_i=j)}{P(Y_i=J)}$.
- If all other covariates are held fixed, an increase in X_{ik} by one unit will result in an increase in the log odds of $Y_i = j$ vs $Y_i = \ell \neq J$ by $\beta_{jk} - \beta_{\ell k}$.
- How do you think we can do inference?

Ordinal Data

- The previous categorical regression did not take into account that Y=1 might be closer to Y=2 than to Y=3.
- For ordinal data, we can fit a proportional odds model.
- Remove the intercept from the design matrix and fit:

$$\log\left\{\frac{P(Y_i \leq j|X_i)}{P(Y_i > j|X_i)}\right\} = \alpha_j + X_i^T\beta$$

- Each j=1,...,J-1 has a different intercept. In order for this model to make sense, what is the condition on the intercepts?
- There is one estimate for each covariate effect.
  - β_k is the log of the odds ratio of Y less than or equal to a value for $X_k = x$ vs. $X_k = x - 1$
- Link is know as the cumulative logit.

Example

- We will use the crab example with the variable mates:
  - mates = 0 if there are no satellites
  - mates =1 if there are 1-5 satellites
  - mates =2 if there are >5 satellites