STAT 545: Categorical Data Analysis (Part II)

Liang Li

Department of Biostatistics
The University of Texas MD Anderson Cancer Center

LLi15@mdanderson.org

Fall 2015 at Rice University

Overview of Part II of this class

Oct 19, 2015 to December 2, 2015. There will be homework/projects and a final exam

- Regression model for binary data
- Regression model for counts data
- Regression model for ordinal data
- Extensions of standard regression models for categorical data
- Marginal models for longitudinal categorical data
- Conditional models for longitudinal categorical data

Logistic Regression Model

$$\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$$
$$= \text{expit}\,(\alpha + \beta x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \in (0, 1)$$
$$\text{logit}\,[\pi(x)] = \log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x \in (-\infty, \infty)$$

1. Binary outcome; for binomial outcome, the model is similar
2. Interpretation of β (log odds ratio)
3. Simple visual model checking by grouping (§ 5.1.2)
4. Logistic regression with retrospective studies (§ 5.1.4)
5. Model fitting through maximum likelihood estimation (§ 5.5)
6. Inference about model parameters and probabilities (§ 5.2.1)
7. Checking goodness of fit (§ 5.2.5)

The (log) odds ratio and its interpretation

$$\text{logit}\left[\pi(x)\right] = \alpha + \beta x$$

$$\text{logit}\left[\pi(\mathbf{x})\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p$$

Simple visual model checking by grouping

1. Group the (continuous) covariate into 10 categories by cutoffs at the quantiles, with n_i subjects in each group ($i = 1, 2, ..., 10$)
2. Calculate the average covariate within each group ($\bar{x}_i$)
3. Calculate the proportion of $Y = 1$ within each group ($\bar{y}_i$)
4. Plot logit of $\bar{y}_i$ vs. $\bar{x}_i$. It should be approximately a straight line
5. Note: may need correction when $\bar{y}_i = 0$ or 1.

$$\log \frac{\bar{y}_i}{n_i - \bar{y}_i} \Rightarrow \log \frac{\bar{y}_i + 0.5}{n_i - \bar{y}_i + 0.5}$$

6. Only work with a single covariate

Logistic regression with retrospective studies (§ 5.1.4)

Model fitting through maximum likelihood estimation (§ 5.5)

Test $H_0 : \beta = 0$ in logistic model $\text{logit}\,[\pi(x)] = \alpha + \beta x$

1. Wald, Likelihood ratio, and Score tests are applicable (§ 1.3.3)
2. The predicted probability and its confidence interval

Checking goodness of fit (§ 5.2.3)

$\text{logit}\,[\pi(\mathbf{x})] = \alpha + \beta_1 x_1 + \beta_2 x_2$

1. Visual checking through grouping (works best with a single covariate)
2. Adding interactions, quadratic terms, etc., and testing for significance or looking at AIC/BIC: problematic but widely used
3. Making the model more flexible by using splines
4. Global goodness of fit checking by *Hosmer & Lemeshow test*

$$\sum_{i=1}^{g} \frac{\left(\sum_j y_{ij} - \sum_j \hat{\pi}_{ij}\right)^2}{n_i \left(\sum_j \hat{\pi}_{ij}/n_i\right)\left[1 - \left(\sum_j \hat{\pi}_{ij}\right)/n_i\right]} \quad \sim \quad \chi^2_{g-2}$$

- A large value of any global fit statistic merely indicates *some* lack of fit but provides no insight about its nature

Logistic models with categorical predictors (§ 5.3)

- When there is a single categorical predictor, the data can be arranged in an $I \times 2$ contingency table (e.g., Table 5.3)
- When the categories are unordered (e.g., nominal data), the (saturated) model is $\text{logit}(\pi_i) = \beta_i$ ($i = 1, 2, ..., I$), with I unknown parameters.
- We may write the model as $\text{logit}(\pi_i) = \alpha + \beta_i$ with set-to-zero constraint $\beta_1 = 0$ or sum-to-zero constraint $\sum_i \beta_i = 0$
- The model for subject j ($j = 1, 2, ..., n$) is $\text{logit}(\pi_j) = \alpha + \sum_{i=1}^{I} \beta_i 1\{j \in \text{group } i\}$
- When the categories are ordered (e.g., ordinal data), we may assume that $\text{logit}(\pi_i) = \alpha + \beta x_i$
  - The number of parameters reduced with the linear assumption.
  - Be careful about coding x_i (i=1,2,...,I): (1,2,3) or (1,4,9)?
  - Treat the x_i like a continuous variable.

Cochran-Armitage Trend Test (§ 5.3.5)

- Developed by Armitage (1955) and Cochran (1954) for $I \times 2$ tables with ordered rows
- They used a linear probability model $\pi = \alpha + \beta x_i$
- It is a chi-square test of the independence between rows and columns under the linear assumption. $H_0 : \beta = 0$.
- This test is equivalent to the score statistic for testing $H_0 : \beta = 0$ in the linear logit model.
- Using directed models can improve inferential power
  - If the trend is indeed linear, making use of the linear trend (as in Cochran-Armitage test) is more powerful than not making use of the linear trend (as in $\mathrm{logit}(pi_i) = \beta_i$)

Model Selection (§ 6.1)

The data set is $\{Y_i, X_{1i}, X_{2i}, ..., X_{pi}; i = 1, 2, ..., n\}$. The logistic regression model is

$$\pi(\mathbf{X}_i) = \text{expit}\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}\right)$$

The p covariates include interactions, quadratic terms, etc. We want to retain only the predictive covariates in the model.

- Model selection is both science and art
- The same principles that you learned in linear model class still apply
- Two goals: (1) complex enough to fit the data well; (2) relatively simple to interpret (avoid overfitting)
- Confirmatory studies vs. exploratory studies

How many covariates can be included in the model?

$$Y_i \sim \text{Bernoulli with } \pi(\mathbf{X}_i) = \text{expit}\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi}\right)$$

- The effective sample size of a logistic regression is either $\sum_i Y_i$ or $n - \sum_i Y_i$, whichever is smaller
- **The rule of thumb**: no more than the effective sample size divided by 10 (or, 10 events per covariate)
- Including too many covariates may cause non-convergence
- Avoid multicollinearity, as in linear regression (Page 209, Table 6.1)
  - The overall test is highly significant ($p < 0.0001$)
  - The individual covariates are, in general, not very significant due to the multicollinearity between the horseshoe crab's width and weight ($r = 0.887$)

Forward, backward, and stepwise model selection

$$Y_i \sim \text{Bernoulli with } \pi(\mathbf{X}_i) = \text{expit}\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}\right)$$

- Forward procedure: (1) start with just the intercept (2) at each step, add the covariate with the smallest p-value in likelihood ratio or Wald test (3) stop when no more significant covariate is available (However, it can stop prematurely due to lack of power)
- Stepwise procedure: at each step, retest the significance of the terms added at previous stages
- Backward procedure: (1) start with full model (2) at each step, remove the covariate with the largest p-value (3) stop when all remaining covariates are significant. (However, full model may not be stable)
- The dummy variables for a single categorical covariate should be added or removed together (likelihood ratio test); do not place an interaction in the model without the main effect terms
- SAS PROC LOGISTIC offers additional entry and exit p-value criteria

Further comment on forward, backward, and stepwise model selection

$$Y_i \sim \text{Bernoulli with } \pi(\mathbf{X}_i) = \text{expit}\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi}\right)$$

-  Page 211, Table 6.2 illustrates
  - three-way interaction is usually not significant (e.g., lack of power) and not desirable (hard to interpret)
  - dropping multiple covariates at once using likelihood ratio test (LRT) or dropping them one at a time (Wald or LRT)
- All these procedures are not rigorously justified (*ad hoc*); use with caution!
- Modern approaches are available (LASSO, bagging, etc.)
- Philosophically, there is no such thing as "the correct model" or "the true model": ALL MODELS ARE WRONG, SOME ARE USEFUL — George Box

Akaike Information Criterion (AIC)

Select the model with smaller AIC or BIC (L: maximized log likelihood; m: number of parameters in the model; n: sample size)

$$
\begin{aligned}
AIC &= -2L + 2m \\
BIC &= -2L + \log(n)m
\end{aligned}
$$

- **Rationale**: Including more covariates will always include the log likelihood, but may cause overfitting; so we put a "penalty" by adjusting for the size of the model. There are mathematical reasons why the penalty must take this form.
- Other penalties are available: HQ, DIC, etc.
- BIC puts more penalty on larger model, and therefore tends to select the simpler model Page 213
- Like scatter plot smoothing, the "desired" amount of penalty is a somewhat subjective choice
- Need a comprehensive assessment of AIC/BIC, significance, residuals, scientific rationale, parsimony and interpretability, etc.

Residuals: Pearson, Deviance, Standardized

Let y_i denote the binomial outcome for n_i trials at setting i of the explanatory variables, $i = 1, 2, ..., N$. Let $\hat{\pi}_i$ denote the model estimate of $P(Y = 1)$.

- Pearson residual is like the residual for linear regression, but with standardization
- Deviance residual is motivated from the likelihood and deviance (which resembles the sum of squares in linear regression)
- Standardized residual has an approximate $N(0, 1)$ distribution and is the one that we usually use, BUT:
  - use it with grouped data (binomial instead of binary). Page 217, Table 6.5

Influence diagnosis for logistic regression

- A single observation can have a much more exorbitant influence in linear regression than in logistic regression, since linear regression has no bound on the distance of y_i from the expected value.
- Points that have extreme predictor values need not have high leverage. In fact, the leverage can be relatively small if $\hat{\pi}_i$ is close to 0 or 1.

Predictive power of a logistic regression model: pseudo R^2

- For linear regression $Y_i = \boldsymbol{X}_i^T \beta + \epsilon_i$, the R^2 is

$$R^2 = 1 - \frac{\sum_i \left(Y_i - \boldsymbol{X}_i^T \hat{\beta}\right)^2}{\sum_i \left(Y_i - \bar{Y}\right)^2}$$

- For logistic regression, the analog

$$1 - \frac{\sum_i \left(Y_i - \hat{\pi}_i\right)^2}{\sum_i \left(Y_i - \bar{Y}\right)^2}$$

may not be nondecreasing as the model gets more complex (undesirable)

Predictive power of a logistic regression model: pseudo R^2

For logistic regression, a more widely used measure is the pseudo R^2 of McFadden (1974): $\dfrac{L_M - L_0}{L_S - L_0} = 1 - \dfrac{L_M}{L_0}$

$$L = \log \prod_{i=1}^{N} \left[\pi_i^{y_i}(1-\hat{\pi}_i)^{1-y_i}\right] = \sum_{i=1}^{N} [y_i \log \hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i)]$$

- L_M is the log likelihood evaluated at the MLE $\hat{\pi}_i = \mathrm{expit}(\boldsymbol{X}_i^T\hat{\beta})$
- L_0 is the log likelihood evaluated under the MLE of the null model: $\hat{\pi}_i = N^{-1}\sum_i y_i$
- L_S is the log likelihood evaluated under the saturated model with $\hat{\pi}_i = y_i$. $L_S = 0$

Receiver Operative Characteristics (ROC) curve

- $y_i = 0$ or 1. $\hat{\pi}_i \in (0, 1)$. We classify the subject as a case ($Y = 1$) when $\hat{\pi} > c$ and control ($Y = 0$) when $\hat{\pi} \leq c$.
- Sensitivity, specificity
- ROC curve p225
- The area under the ROC curve (AUC) is reported as c-statistic in SAS PROC LOGISTIC. It is a number between 0 and 1. AUC $= 0.5$ is like flipping a coin. So AUC < 0.5 is unlikely. Good classification requires AUC > 0.80 (excellent, > 0.9).