STAT347: Generalized Linear Models
Lecture 6

Today's topics: Chapters 5.2, 5.3 - 5.5, 5.7

- Some applications of Binary GLM
- Binary GLM inference
- Fitting logistic regression and the infinite estimates
- Binary GLM example

1 Some applications of a Binary GLM

1.1 2×2 table

When Both the X_i and y_i are binary, the grouped data can be represented by a 2×2 table.

- Number of grouped samples: 2.
- Number of total ungrouped observations: $N = n_1 + n_2$ (Table 5.2 of the Agresti book)
- Assume that (X_i, y_i) are i.i.d. Odds ratio (OR) for the response variable Y:

$$\text{OR} = \frac{\mathbb{P}(Y = 1 \mid X = 1)/\mathbb{P}(Y = 0 \mid X = 1)}{\mathbb{P}(Y = 1 \mid X = 0)/\mathbb{P}(Y = 0 \mid X = 0)}$$

- Interpretation of the coefficient β_1 in the binary GLM with logit link: $\text{logit}(p_i) = \beta_0 + \beta_1 X_i$

$$e^{\beta_1} = \text{OR}$$

1.2 Case-control study

We want to know

$$\text{Risk factor } X \overset{\text{effect?}}{\longrightarrow} \text{ Outcome } Y$$

$X_i = 1/0$ if the person is a smoker/non-smoker and $y_i = 1/0$ if the person develops cancer/is a healthy control.

- Prospective design: randomly select smokers and non-smokers from the population and observe whether they will develop cancer in the future.
  - We can compare $\mathbb{E}(Y = 1 \mid X = 1)$ with $\mathbb{E}(Y = 1 \mid X = 0)$
  - Drawbacks: the study takes a long time; lung cancer is a rare disease, may observe very few cancer samples.

- Case-control study (retrospective): We randomly select some samples from patients who develop cancer and some samples from healthy controls. Then, we check whether the person has been a smoker or not.
  - We can now only compare $\mathbb{E}(X = 1 \mid Y = 1)$ with $\mathbb{E}(X = 1 \mid Y = 0)$
  - The study takes a shorter time, and we can obtain enough cancer cases.

Why is the case-control study popular?

$$\begin{aligned}\text{OR} &= \frac{\mathbb{P}(Y = 1 \mid X = 1)/\mathbb{P}(Y = 0 \mid X = 1)}{\mathbb{P}(Y = 1 \mid X = 0)/\mathbb{P}(Y = 0 \mid X = 0)} \\ &= \frac{\mathbb{P}(X = 1 \mid Y = 1)/\mathbb{P}(X = 0 \mid Y = 1)}{\mathbb{P}(X = 1 \mid Y = 0)/\mathbb{P}(X = 0 \mid Y = 0)}\end{aligned}$$

We can also include other covariates $\tilde{X}$:

$$\begin{aligned}\text{OR}\mid_{\tilde{X}=x} &= \frac{\mathbb{P}(Y = 1 \mid X = 1, \tilde{X} = x)/\mathbb{P}(Y = 0 \mid X = 1, \tilde{X} = x)}{\mathbb{P}(Y = 1 \mid X = 0, \tilde{X} = x)/\mathbb{P}(Y = 0 \mid X = 0, \tilde{X} = x)} \\ &= \frac{\mathbb{P}(X = 1 \mid Y = 1, \tilde{X} = x)/\mathbb{P}(X = 0 \mid Y = 1, \tilde{X} = x)}{\mathbb{P}(X = 1 \mid Y = 0, \tilde{X} = x)/\mathbb{P}(X = 0 \mid Y = 0, \tilde{X} = x)}\end{aligned}$$

Thus, we can study estimate the odds ratio of the risk factor from case-control studies.

Thus, building the logistic regression using case-control study samples is the same as building the model using prospective samples:

$$e^{\beta_1} \equiv \text{OR}\mid_{\tilde{X}=x}$$

1.3 Classification

Binary GLM models can be used for classification.

Some concepts in evaluating the classification result

- A classification table (Table 5.1 of the Agresti book): y v.s. $\hat{y}$
- Sensitivity (recall, true positive rate, tpr): $P(\hat{y} = 1 \mid y = 1)$
- Specificity: $P(\hat{y} = 0 \mid y = 0)$
- False positive rate (fpr): $1 -$ specificity $= P(\hat{y} = 1 \mid y = 0)$
- ROC curve (Figure 5.2 of the Agresti book): fpr v.s. sensitivity

2 Binary GLM model inference

We have already learnt the inference of a general GLM model, we now look what the specific forms are for a binary GLM.

2.1 Score equation in logistic regression

For logistic regression, as the logit link is the canonical link, the score equation is:

$$\frac{\partial L}{\partial \beta_j} = \sum_i (y_i - n_i p_i) x_{ij} = \sum_i \left(y_i - \frac{n_i e^{X_i^T \beta}}{1 + e^{X_i^T \beta}} \right) x_{ij} = 0$$

We have derived that as $n \to \infty$

$$\mathrm{Var}(\hat{\beta}) \to (X^T W X)^{-1}$$

where $W = D^2 V^{-1}$ is a diagonal matrix. For logistic regression where the logit link is the canonical link, we have $W = V$ so

$$W_{ii} = n_i p_i (1 - p_i), \quad \widehat{W}_{ii} = n_i \frac{e^{X_i^T \hat{\beta}}}{(1 + e^{X_i^T \hat{\beta}})^2}$$

2.2 Deviance

The total (residual) deviance for a binary GLM (the deviance between the saturated model and the fitted model) is

$$\begin{aligned} D_+(y, \hat{\mu}) &= \sum_i D(y_i, n_i \hat{p}_i) \\ &= -2 \sum_i \log \left[f(y_i, \hat{\theta}_i) / f(y_i, \theta_{y_i}) \right] \\ &= -2 \sum_i \log \left[\frac{\hat{p}_i^{y_i} (1 - \hat{p}_i)^{n_i - y_i}}{(y_i/n_i)^{y_i} (1 - y_i/n_i)^{n_i - y_i}} \right] \\ &= 2 \sum_i y_i \log \frac{y_i}{n_i \hat{p}_i} + 2 \sum_i (n_i - y_i) \log \frac{n_i - y_i}{n_i - n_i \hat{p}_i} \end{aligned}$$

- The total deviance is different for grouped data and ungrouped data as the saturated model is different.
  - Ungrouped data: the saturated model is $\hat{p}_i = y_i$ for each individual sample
  - grouped data: the saturated model is $\hat{p}_k = \tilde{y}_k$ for each group k. Thus all samples in the same group should have the same $\hat{p}_i$ even in the saturated model.

2.3 Goodness-of-fit test

The group level data can be presented by a $K \times 2$ count table, where each row is a group, and the two columns store the number of success $\tilde{y}_k$ and the number of failure $n_k - \tilde{y}_k$ respectively in each cell.

- Residual deviance for the grouped data:

$$\begin{aligned} G^2 = D_+(y, \hat{\mu}) &= 2 \sum_k \tilde{y}_k \log \frac{\tilde{y}_k}{n_k \hat{p}_k} + 2 \sum_k (n_k - \tilde{y}_k) \log \frac{n_k - \tilde{y}_k}{n_k - n_k \hat{p}_k} \\ &= 2 \sum_{2K \text{ cells}} \text{observed} \times \log \left(\frac{\text{observed}}{\text{fitted}} \right) \end{aligned}$$

- When the number of groups K is fixed while the total samples size $N = \sum_k n_k$ is large, then the residual deviance is the likelihood ratio satisfying
$$G^2 = D_+(y, \hat{\mu}) \xrightarrow{p} \chi^2_{K-p}$$
which can be used for goodness-of-fit test of the fitted model.

- Pearson's statistics for goodness of fit:
$$\begin{aligned} X^2 &= \sum_{2K \text{ cells}} \frac{(\text{observed } - \text{ fitted})^2}{\text{fitted}} \\ &= \sum_k \frac{(n_k\tilde{y}_k - n_k\hat{p}_k)^2}{n_k\hat{p}_k} + \sum_k \frac{[(n_k - \tilde{y}_k) - (n_k - n_k\hat{p}_k)]^2}{n_k - n_k\hat{p}_k} \\ &= \sum_k \frac{(\tilde{y}_k - n_k\hat{p}_k)^2}{n_k\hat{p}_k(1-\hat{p}_k)} \xrightarrow{p} \chi^2_{K-p} \end{aligned}$$

- Comparison between G^2 and X^2
  - $X^2 = \sum_k e_k^2$: sum square of Pearson residuals of grouped data. X^2 converges to χ^2_{K-p} more quickly, so it works better than G^2 for N not too large.
  - $G^2 = \sum_k d_k^2$: sum square of deviance residuals of grouped data. G^2 gives more reliable p-values than X^2 when some cells have small expected counts (≤ 5).

3 Binary GLM computation

For logistic regression, Newton's method = Fisher scoring = IRLS.
For IRLS, the tth iteration is
$$X^T W^{(t)}(z^{(t)} - X\beta) = 0$$
where
$$\begin{aligned} z_i^{(t)} &= X_i^T\beta^{(t)} + \left(D_{ii}^{(t)}\right)^{-1}(y_i - \mu_i^{(t)}) \\ &= \log\left(\frac{p_i^{(t)}}{1-p_i^{(t)}}\right) + \frac{y_i - n_i p_i^{(t)}}{n_i p_i^{(t)}(1-p_i^{(t)})} \end{aligned}$$
and
$$W_{ii}^{(t)} = V_{ii}^{(t)} = n_i p_i^{(t)}(1-p_i^{(t)})$$

3.1 Infinite parameter estimates

One may sometimes see this warning message using R to solve the logistic regression:
Warning message: glm.fit: fitted probabilities numerically 0 or 1 occurred
You may see very large estimates of β. What happened?

- Perfect separation:
  There exists β_s such that if $X_i^T\beta_s > 0$ then $y_i = 1$ otherwise $y_i = 0$.

We proof that the MLE for β does not exist. Let $\eta_i = kX_i^T\beta_s$.

When $k \to \infty$, then

$$p_i = \frac{e^{kX_i^T\beta_s}}{1+e^{kX_i^T\beta_s}} \to \begin{cases} 1 & \text{if } X_i^T\beta_s > 0, \text{ or equivalently } y_i = 1 \\ 0 & \text{else} \end{cases}$$

Thus, $\frac{\partial L}{\partial \beta} \to 0$ if $k \to \infty$ so the solution of the score equation is infinite. In other words, the MLE does not exist.

- Quasi-complete separation:

  There exists β_s such that if $X_i^T\beta_s > 0$ then $y_i = 1$, if $X_i^T\beta_s < 0$ then $y_i = 0$, and if $X_i^T\beta_s = 0$ then $y_i = 0$ or 1 (allow data points on the separation hyperplane with both outcomes).

  We can also show that the MLE for β does not exist (Albert and Anderson, *Biometrika* 1984). Any value β can be decomposed as $\beta = \beta_s + \gamma$. Denote $\beta_k = k\beta_s + \gamma$ Let $\eta_i = kX_i^T\beta_s + X_i^T\gamma$. When $k \to \infty$, then

$$p_i = \frac{e^{kX_i^T\beta_s + X_i^T\gamma}}{1+e^{kX_i^T\beta_s + X_i^T\gamma}} \to \begin{cases} 1 & \text{if } X_i^T\beta_s > 0 \\ 0 & \text{if } X_i^T\beta_s < 0 \\ \frac{e^{X_i^T\gamma}}{1+e^{X_i^T\gamma}} & \text{if } X_i^T\beta_s = 0 \end{cases}$$

  This tells us that for any β, we can find β_k with large enough k so that the log-likelihood $L(\beta_k) > L(\beta)$, so the log-likelihood function $L(\cdot)$ does not have a finite maximum point. In other words, the MLE does not exist.

- How to deal with perfect/quasi-complete separation? (Read Chapter 5.4.2)

  We can add a penalization or add a prior of the parameter to obtain finite estimates of β.

4 Data example

Chapter 5.7. Please check the R notebook 3-2.

Next time: Chapter 6.1, multivariate GLM: nominal response