STAT347: Generalized Linear Models
Lecture 6

Winter, 2024
Jingshu Wang

Today’s topics:

- Binary GLM inference
- Fitting logistic regression and the infinite estimates
- Some applications of Binary GLM
- Binary GLM example (part II)

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 $n \to \infty$

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

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

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

Residual deviance is different for grouped and ungroup data

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

- For the grouped data

$$
D_+(y,\hat{\mu}) = 2\sum_k \tilde{y}_k \log\frac{\tilde{y}_k}{n_k\hat{p}_k} + 2\sum_k (n_k-\tilde{y}_k)\log\frac{n_k-\tilde{y}_k}{n_k-n_k\hat{p}_k}
$$

Much smaller

- For the ungrouped data

$$
\begin{aligned}
D_+(y,\hat{\mu}) &= 2\sum_k\sum_{i\in I_k} y_i \log\frac{y_i}{\hat{p}_k} + 2\sum_k\sum_{i\in I_k}(1-y_i)\log\frac{1-y_i}{1-\hat{p}_k} \\
&= 2\textstyle\sum_k \tilde{y}_k \log\frac{1}{\hat{p}_k} + 2\sum_k (n_k-\tilde{y}_k)\log\frac{1}{1-\hat{p}_k}
\end{aligned}
$$

Residual deviance is different for grouped and ungroup data

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

- For the ungrouped data, each observation is y_i
  - The saturated model is $\hat{p}_i = y_i$ for each individual sample
- For the grouped data each observation is $\tilde{y}_k$
  - The saturated model is $\hat{p}_k = \tilde{y}_k$ for each group (so that $\hat{p}_i$ for each individual sample in the saturated model is $\tilde{y}_k$ instead of the birary y_i)

Residual deviance for grouped data

- 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 group data

$$G^2 = D_+(y, \hat{\mu}) = 2\sum_k \tilde{y}_k \log \frac{\tilde{y}_k}{n_k \hat{p}_k} + 2\sum_k (n_k - \tilde{y}_k) \log \frac{n_k - \tilde{y}_k}{n_k - n_k \hat{p}_k}$$

$$= 2 \sum_{2K \text{ cells}} \text{observed} \times \log \left(\frac{\text{observed}}{\text{fitted}} \right)$$

- 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

$$G^2 = D_+(y, \hat{\mu}) \xrightarrow{p} \chi^2_{K-p}$$

Goodness-of-fit test of the fitted model

- Residual deviance for goodness of fit

$$G^2 = D_+(y, \hat{\mu}) \xrightarrow{p} \chi^2_{K-p}$$

- 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 in general 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).

Infinite parameter estimates in logistic regression

```
-----------------------------------------------------------------
> x <- c(1,2,3,4,5,6);  y <- c(1,1,1,0,0,0)  # complete separation
> fit <- glm(y ~ x, family = binomial(link = logit))
> summary(fit)
Coefficients:
            Estimate  Std. Error  z value Pr(>|z|)
(Intercept)   165.32   407521.43        0        1 # x estimate is
x             -47.23   115264.41        0        1 # actually -infinity

Number of Fisher Scoring iterations: 25   # unusually large
> logLik(fit)
'log Lik.' -1.107576e-10 (df=2) # maximized log-likelihood = 0
-----------------------------------------------------------------
```

Or sometimes one may see the following warning message:

Warning message: glm.fit: fitted probabilities numerically 0 or 1 occurred

Perfect (complete) separation

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

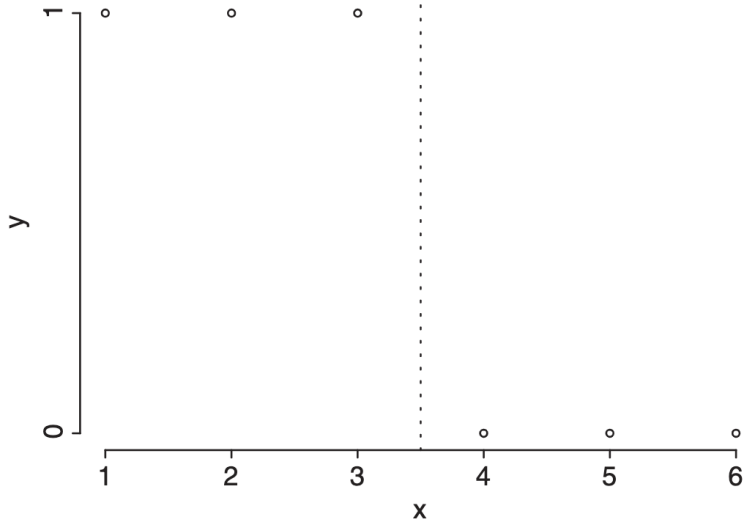

Figure 5.3 Complete separation of explanatory variable values, such as $y = 1$ when $x < 3.5$ and $y = 0$ when $x > 3.5$, causes an infinite ML effect estimate.

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 $y_i = 1$,
$X_i^T \beta_s < 0$ then $y_i = 0$,
$X_i^T \beta_s = 0$ then $y_i = 0$ or 1

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.

2 X 2 table

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

		Event	
		Yes	No
Exposure	Yes	a	b
	No	c	d

- 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: $\text{logit}(p_i) = \beta_0 + \beta_1 X_i$

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

Prospective V.S. retrospective design

- We want to know the effect of a risk factor (say smoking) on an outcome (say lung cancer)

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

- Retrospective design (case-control study): 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.
  - Only compare $\mathbb{E}(X = 1|Y = 1)$ with $\mathbb{E}(X = 1|Y = 0)$
  - The study takes a shorter time, and we can obtain enough cancer cases.

Case-control study

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 $\tilde{X}$:

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

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:

$$e^{\beta_1} \equiv \text{OR}\,|_{\tilde{X}=x}$$

Classification

Table 5.1 A Classification Table

	Prediction $\hat{y}$	
y	0	1
0		
1		

Cell counts in such tables yield estimates of sensitivity = $P(\hat{y} = 1 \mid y = 1)$ and specificity = $P(\hat{y} = 0 \mid y = 0)$.

- 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

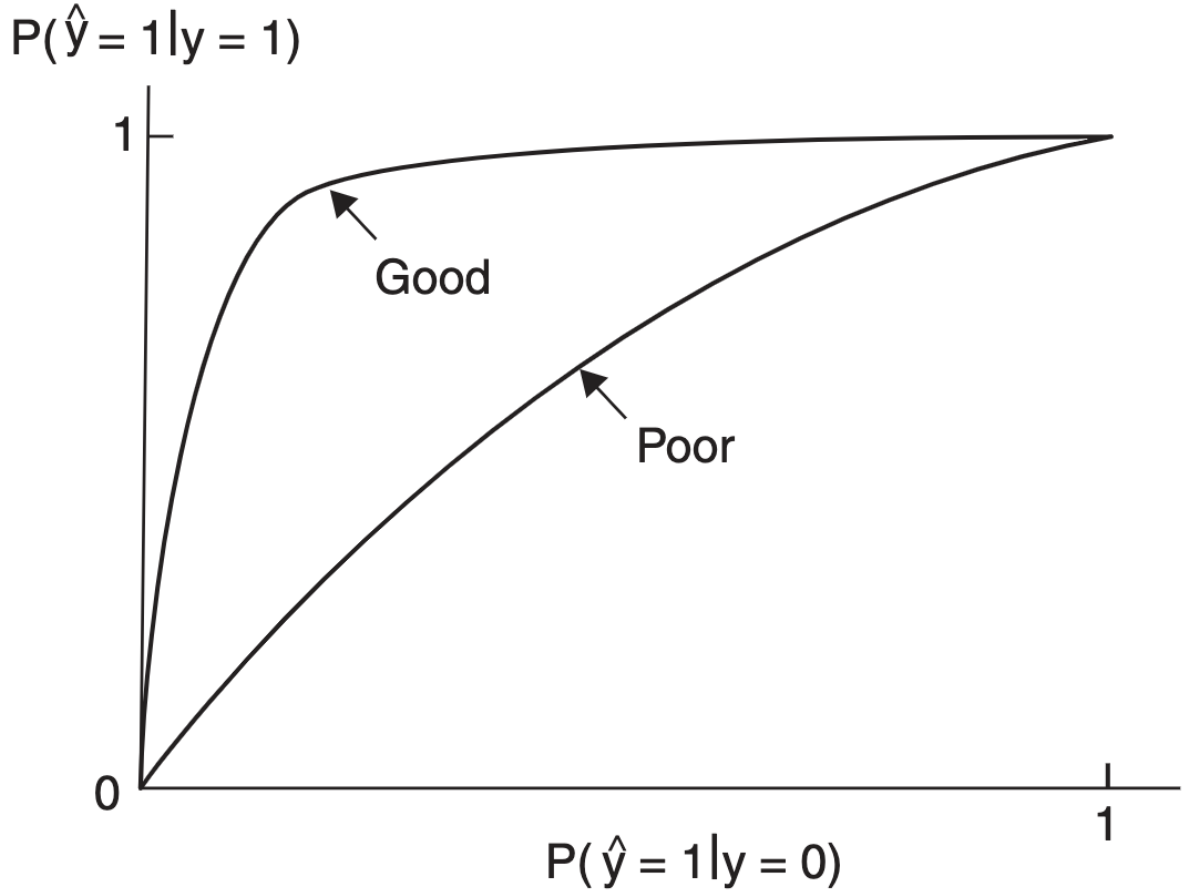

Figure 5.2 ROC curves for a binary GLM having good predictive power and for a binary GLM having poor predictive power.

R data example for binary / binomial GLM (part II)

- Check Example3_2 R notebook