Chapter 9: Logistic Regression

CZ

Fall, 2014

Categorical Response

- So far: We learned how to fit simple/multiple regression model with categorical/indicator variables as predictors.
- In this chapter: Consider the case that our response is a categorical variable.

Binary Logistic Regression

We first introduce the **Binary** Logistic Regression model, where the response has only two levels. The objective is to

- determine how one or more predictors affect the probability that an observation falls into one category of the response;
- and predict the probabilities (and label) for a new observation.

y_i	Probability
1	$P(y_i = 1) = \pi_i$
0	$P(y_i = 0) = 1 - \pi_i$

Examples

- We may want to predict **the probability** that a student will go to the graduate school using data on college GPA, SAT score and major.
- Response variable: "go to graduate school" and "not go to graduate school" (can be coded as 1 and 0).
- Predictors: college GPA, SAT score, and major (categorical predictor).

Binary Response

Note: What we are interested in is the **probability** that the response y taking some value, e.g., $\pi = P(y = 1)$.

Compare this to the continuous regression model: We are interested in predicting $E(y)$.

Regression: estimate $E(y)$, where y is the (transformed) response variable.

Binary Response

For a given predictor vector $\boldsymbol{x}$, if response variable y falls into one category (coded as 1) with probability $\pi(\boldsymbol{x})$ and into the other (coded as 0) with probability $1 - \pi(\boldsymbol{x})$, then y follows a Bernoulli distribution with parameter $\pi(\boldsymbol{x})$.

- In this case, the mean/expectation of Y is:

$$E(y) = 1 \cdot \pi(\boldsymbol{x}) + 0 \cdot (1 - \pi(\boldsymbol{x})) = \pi(\boldsymbol{x}).$$

  That is, the probability of the response falling into the category of interest just equals to the mean of response.

- Here $\pi(\boldsymbol{x})$ depends on $\boldsymbol{x}$.

Problems

- Is the following ordinary regression model still reasonable if the response is binary?

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

$$\text{or } E(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- Two main problems of fitting ordinary regression model for binary response:
  1. $E(y) = \pi(\boldsymbol{x}) \in (0, 1)$. But fitting ordinary regression models cannot guarantee this.
  2. $Var(y) = \pi(\boldsymbol{x})(1 - \pi(\boldsymbol{x})) = E(y)(1 - E(y))$
     The error ϵ is no longer normally distributed and the variance of error terms are not constant (depends on the mean of y).

New Response Variable

Solution: Model the probability $\pi(\boldsymbol{x})$, instead of the 0-1 label vector!

- $\boldsymbol{\pi(x)}$: **probability** of the unit falling into one category of interest. $\pi(\boldsymbol{x}) = E(y) = P(y = 1) \in (0, 1)$ for given $\boldsymbol{x}$.
- $\boldsymbol{\frac{\pi(x)}{1-\pi(x)}}$: **odds**, refers to the fraction of the probability of falling into one category versus not. $\pi(\boldsymbol{x}) \in (0, 1) \Rightarrow \frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} \in (0, +\infty)$.
- $\boldsymbol{\log\frac{\pi(x)}{1-\pi(x)}}$: **log odds**. This transformation is called **logit link**, denoted by $logit(\pi(\boldsymbol{x}))$.
$\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} \in (0, \infty) \Rightarrow \log(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}) \in (-\infty, +\infty)$.

Model Setup

The multiple binary logistic regression model is the following:

$$logit(\pi(\boldsymbol{x})) = \log(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- In terms of odds:

$$\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p},$$

  which describes the **odds** of being in the category of interest.

- In terms of probability:

$$\pi(\boldsymbol{x}) = E(y) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}$$

  Different from before, now the relation between $E(y) = \pi(\boldsymbol{x})$ and $\boldsymbol{x}$ is **non-linear**!

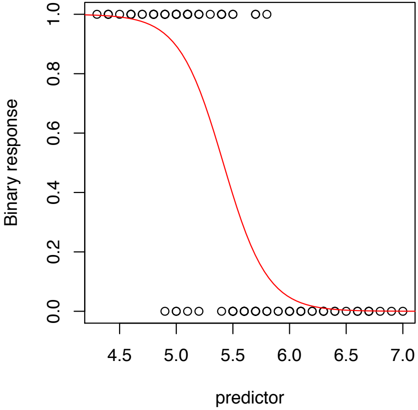
Important

We are NOT assuming that

$$logit(\pi(\boldsymbol{x})) = \log(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon.$$

Where does the randomness come from?

Relationship between $E(y)$ and x



Interpretation

- β_j: the change in the log odds when x_j is increased by one unit (and other predictors are held constant).
- e^{β_j}: the multiplicative factor on the odds when x_j is increased by one unit (and other predictors are held constant).

Estimation

After fitting the logistic model, we can estimate π_i for subject i:

$$\hat{\pi}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}}} \in (0, 1).$$

Coefficients Estimation

How do we estimate the coefficients?

- Before: Least squares criterion:

$$\hat{\beta} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y}$$

- Now: for logistic regression model, β is estimated by **Maximum Likelihood Method**.

Coefficients Estimation

Let $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^T$. The likelihood function for Bernoulli distribution is:

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\beta}}\right)^{y_i}\left(\frac{1}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\right)^{1-y_i}. \end{aligned}$$

$\hat{\boldsymbol{\beta}}$ is solved by maximizing $L(\boldsymbol{\beta})$ or $\log L(\boldsymbol{\beta})$ via iterative numerical algorithms (Newton-Raphson method).

Significance Test

To test whether a specific predictor x_j, $j = 1, \ldots, p$ is important to predict the probability of y falling into the category of interest,

$$H_0: \ \beta_j = 0 \quad H_a: \ \beta_j \neq 0.$$

Wald Test:

$$Z = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim N(0, 1) \text{ approximately, under } H_0.$$

- If $|Z_0| > Z_{\alpha/2}$, reject H_0.
- If p-value$= P(|Z| > |Z_0|) < \alpha$, reject H_0.

Note: How to test individual coefficient is another difference between logistic regression model and ordinary regression model.

Example

A group of people were asked if they have ever DUI (y). They also were asked "How many days per month do you drink at least two beers?"

Define

$$y = \begin{cases} 1, & \text{if the person says "yes";} \\ 0, & \text{if the person says "no".} \end{cases}$$

$\pi(x) = P(y = 1)$, $x =$ days per month of drinking at least two beers.

Example

```
> fit=glm(DrivDrnk~DaysBeer,family="binomial",
data=drinking)
> summary(fit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.55136    0.26605  -5.831 5.51e-09 ***
DaysBeer     0.19031    0.02946   6.459 1.05e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Example

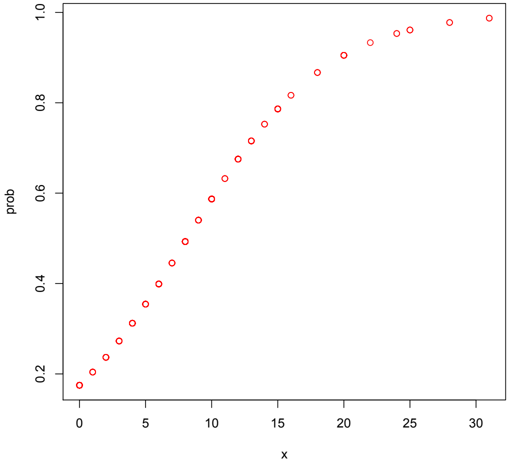
- $\hat{\beta}_0 = -1.55136$, and $\hat{\beta}_1 = 0.190306$.
- The model for estimating $\pi =$probability of ever having driven after drinking is

$$\hat{\pi}_i = \frac{e^{-1.55136+0.190306x_i}}{1 + e^{-1.55136+0.190306x_i}}$$

  Q: What is the probability of never driving after drinking?

- The variable $x =$DaysBeer is statistically significant.

Plot $\hat{\pi}_i$ versus x_i



Example

Some estimated probabilities calculated from the fitted model:

DaysBeer	4	20	28
$\hat{\pi}(x)$	0.312	0.905	0.978

- For example, if $x = 4$ days per month of drinking two beer,

$$\hat{\pi}(4) = \frac{e^{-1.55136+0.190306\times 4}}{1 + e^{-1.55136+0.190306\times 4}} = 0.312$$

Odds Ratios in Logistic Regression

- When $x = 4$, the predicted odds of ever driving after drinking is $0.312/(1-0.312) = 0.453$.
- Recall $e^{\hat{\beta}_1}$ is interpreted as the predicted multiplicative factor on the odds when that predictor is increased by one unit (and other predictors are held constant).
- Thus, when $x = 6$, the predicted odds of ever driving after drinking is $0.453 \times e^{0.190306 \times (6-4)} = 0.663$

Example: Multiple Logistic Regression

We now include **Gender** (male or female) as an predictor (along with **DaysBeer**). "Gender" is an indicator variable with a value$= 1$ if male and$= 0$ if female.

Example

```
> fit=glm(DrivDrnk~DaysBeer+Gender,family="binomial",
data=drinking)
> summary(fit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.77356    0.29446  -6.023 1.71e-09 ***
DaysBeer     0.18693    0.03004   6.223 4.87e-10 ***
Gendermale   0.61724    0.29538   2.090   0.0366 *
```

Example

- The p-values are less than 0.05 for both **DaysBeer** and **Gender**. This is evidence that both x-variables are useful for predicting the probability of ever having driven after drinking.
- For **DaysBeer**, the odds ratio is estimated to equal 1.21 (calculated as $e^{0.18693}$).
- For **Gender**, odds ratio$=1.85$ (calculated as $e^{0.6172}$). For males, the odds of ever having driven after drinking is 1.85 times the odds for females, assuming **DaysBeer** is held constant.

$100(1-\alpha)\%$ Confidence Interval

- Approximate $100(1-\alpha)\%$ confidence interval for β_j (log odds ratio) is:

$$(\hat{\beta}_j - Z_{\alpha/2} se(\hat{\beta}_j), \quad \hat{\beta}_j + Z_{\alpha/2} se(\hat{\beta}_j))$$

  where $Z_{\alpha/2}$ is the upper $100(\alpha/2)$th percentile of $N(0,1)$.

- Approximate $100(1-\alpha)\%$ confidence interval for e^{β_j} (odds ratio) is:

$$(e^{\hat{\beta}_j - Z_{\alpha/2} se(\hat{\beta}_j)}, \quad e^{\hat{\beta}_j + Z_{\alpha/2} se(\hat{\beta}_j)})$$

Likelihood Ratio Test

To test the significance of two or more predictors,

- Ordinary Regression: General Linear F-test
- Now:

$$\text{LR test statistic} = 2\{\log L(\mathit{Full}) - \log L(\mathit{Restricted})\}$$

Under H_0, the test statistic follows $\chi^2(l)$, where l is the number of independent constraints.

Likelihood Ratio Test

- A special case: to test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

- Ordinary Regression:

$$F = \frac{MSR}{MSE}$$

  Under $H_0 : F \sim F(p, n-p-1)$

- Now: Likelihood ratio test

Likelihood Ratio Test

- Under H_0,

$$\pi(\boldsymbol{x}) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

- The maximum likelihood estimator of π and β_0 under H_0 is:

$$\hat{\pi}(\boldsymbol{x}) = \frac{\sum_i y_i}{n} = \bar{y}$$

$$\hat{\beta}_0 = \log \frac{\bar{y}}{1 - \bar{y}}$$

- The maximum log-likelihood under H_0 is
  $\log L_0 = \sum y_i \log \bar{y} + \sum (1 - y_i) \log(1 - \bar{y})$
- The likelihood ratio test statistic is: $2(\log L(full) - \log L_0)$
- We reject H_0 if $LR > \chi^2_\alpha(p)$

Example

```
> fit=glm(DrivDrnk~DaysBeer+Gender,family="binomial",
data=drinking)
> summary(fit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.77356    0.29446  -6.023 1.71e-09 ***
DaysBeer     0.18693    0.03004   6.223 4.87e-10 ***
Gendermale   0.61724    0.29538   2.090   0.0366 *

Null deviance: 345.09  on 248  degrees of freedom
Residual deviance: 279.96  on 246  degrees of freedom
AIC: 285.96
```

Example

The 95% confidence interval for β_1 is:

$$\begin{aligned}\hat{\beta}_1 \pm Z_{\alpha/2} se(\hat{\beta}_1) &= 0.18693 \pm 1.96 \times 0.03004 \\ &= (0.128, 0.246)\end{aligned}$$

The 95% confidence interval for e^{β_1} is:

$$(e^{0.128}, e^{0.246}) = (1.137, 1.279)$$

Example

To test $H_0 : \beta_1 = \beta_2 = 0$

- Null Deviance: $-2\log L(Restricted) = 345.09$
- Residual Deviance: $-2\log L(Full) = 279.96$
- LR test statistic$=2\{\log L(Full) - \log L(Restricted)\} = 345.09 - 279.96 = 65.13$
- We reject H_0 because $\chi^2_{0.05}(2) = 5.99$ and $65.13 > 5.99$.

Logistic Regression with More than Two Categories

- $y \sim Multinomial(\pi_1, \pi_2, ..., \pi_k)$, the response is a categorical variable with k levels (y has probability π_i to fall into category i). In this case, we consider a form similar as the logit link:

$$
\begin{aligned}
\log(\frac{\pi_1}{\pi_k}) &= \beta_{1,0} + \beta_{1,1}x_1 + ... + \beta_{1,p}x_p \\
\log(\frac{\pi_2}{\pi_k}) &= \beta_{2,0} + \beta_{2,1}x_1 + ... + \beta_{1,p}x_p \\
&\cdots \\
\log(\frac{\pi_{k-1}}{\pi_k}) &= \beta_{k-1,0} + \beta_{k-1,1}x_1 + ...\beta_{k-1,p}x_p
\end{aligned}
$$

Logistic Regression with More than Two Categories

1. By estimating all the β-coefficients in the above equations, we can have estimated probability for y falling into each category.
2. Note that there is a hidden restriction: $\sum_{i=1}^{k} \pi_i(\boldsymbol{x}) = 1$. And the above formulation assumes level k as the baseline.
3. Note that when $k = 2$, the formulation is the same as the binary logistic regression model.