STAT 408
Applied Regression Analysis

Nan Miles Xi

Department of Mathematics and Statistics

Loyola University Chicago

Fall 2022

Logistic Regression

Binary Response

- All previous linear models have one common characteristic: the response Y is continuous variable
- In many applications, Y is categorical
  - cancel diagnostics (yes or no), spam email detection (spam or normal), image classification (dem, gop)
- The regular linear regression doesn’t work in those applications, because regular linear regression will generate response in $(-\infty, +\infty)$
- In this chapter, we will briefly introduce how to model the binary response using <u>logistic regression (logistic model)</u>

Logistic Regression

- Suppose we have a response variable Y which takes the values zero or one (binary)

- We also have q predictors $X_1, X_2, \dots, X_q$ on which we want to build a linear model

- We treat Y as a binomial random variable, that is, Y has probability p to take value one and probability $1 - p$ to take value zero

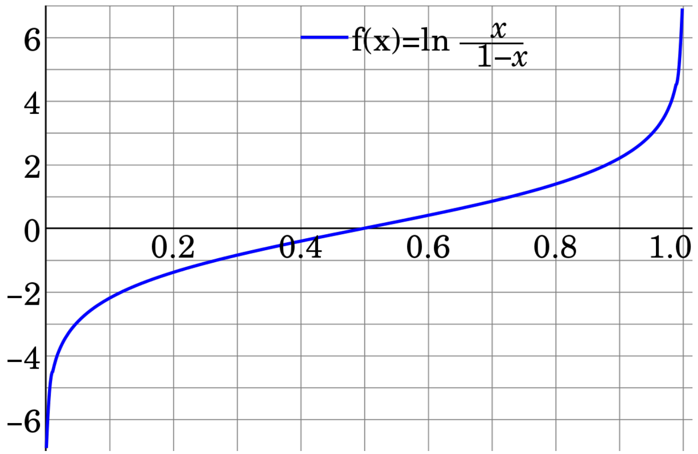
$$Y \sim Bin(p) \quad \begin{cases} P(Y = 1) = p \\ P(Y = 0) = 1 - p \end{cases}$$

Logistic Regression

- Since probability p is from 0 to 1, we use logit function

$$logit(p) = \log \frac{p}{1-p}$$

to transform (0, 1) to $(-\infty, +\infty)$



- The logistic regression takes the following form:

$$\log \frac{p}{1-p} = \underbrace{\beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q}_{-\infty \quad +\infty}$$

- The logit function $logit(p) = \log \frac{p}{1-p}$ is called link function

Logistic Regression

- In logistic regression, the linear part no longer directly relates to response Y, but to the “log-ratio” between $P(Y = 1)$ and $P(Y = 0)$

$$log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q$$

- Logistic regression explicitly assumes that Y follows a binomial distribution – it is a probabilistic model

- We can show that

$$p = P(Y = 1) = \frac{1}{1 + e^{-\underbrace{(\beta_0+\beta_1 X_1+\cdots+\beta_q X_q)}_{-\infty}}}$$

Logistic Regression

- By treating Y as a binomial random variable, logistic regression transforms the output of linear regression from $(-\infty, +\infty)$ to a probability $\in (0, 1)$

- With $p = P(Y = 1)$, we obtain the final binary model output by

$$Y = \begin{cases} 0 & if (P = 1) < 0.5 \\ 1 & if\ (P = 1) \geq 0.5 \end{cases}$$

- The logistic regression still uses a linear combination of all predictors, but maps them to a binary output

Estimation of Logistic Model

- We cannot minimize RSS to estimate the logistic model, because there is no "residual" any more

- We will use maximum likelihood estimation (MLE) to estimate the parameters in logistic model
  - MLE is a standard estimation method for probabilistic model

- The idea of MLE is to find the parameters that maximize the likelihood function, which is the "probability" of observing our data

Estimation of Logistic Model

- Suppose that the ith observation in the data is $(x_{i1}, x_{i2}, \dots, x_{iq}, y_i)$

- In logistic regression, the likelihood function for single observation i is the probability to observe y_i

$$l_i = p^{y_i}(1-p)^{1-y_i}$$

where $l_i = p$ if $y_i = 1$, and $l_i = 1 - p$ if $y_i = 0$

Estimation of Logistic Model

- Suppose there are n observations in the data

$$\text{observation 1:} \quad (x_{11}, x_{12}, \dots , x_{1q}, y_1)$$
$$\text{observation 2:} \quad (x_{21}, x_{22}, \dots , x_{2q}, y_2)$$
$$\dots$$
$$\text{observation n:} \quad (x_{n1}, x_{n2}, \dots , x_{nq}, y_n)$$

- Under observation independence, the likelihood of all observations is the joint probability of observing all responses

$$L = \prod_{i=1}^{n} l_i = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

Estimation of Logistic Model

- Since $p = P(Y = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_q X_q)}}$, the likelihood is a function of parameters β

$$L(\beta) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

- The likelihood function $L(\beta)$ is the probability of observing all observations
- A “good model” should make this probability as large as possible, because we do observe the data

Estimation of Logistic Model

- To obtain the β that maximizes likelihood function, we take derivative of $L(\beta)$ with respective to β and let it equal to zero

$$\frac{dL(\beta)}{\beta} = 0$$

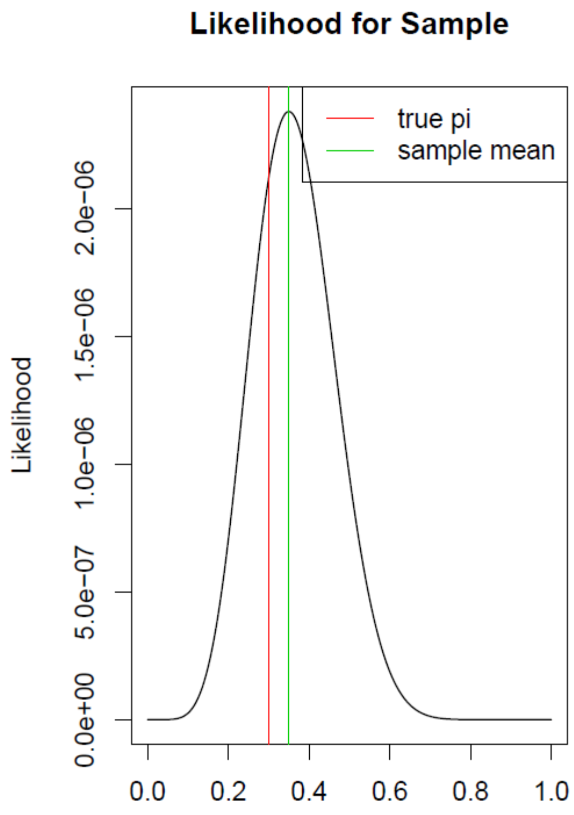
- The product $\prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$ is difficult to take derivative; In practice, we take log to change the product to summation, which is called <u>log-likelihood function</u>

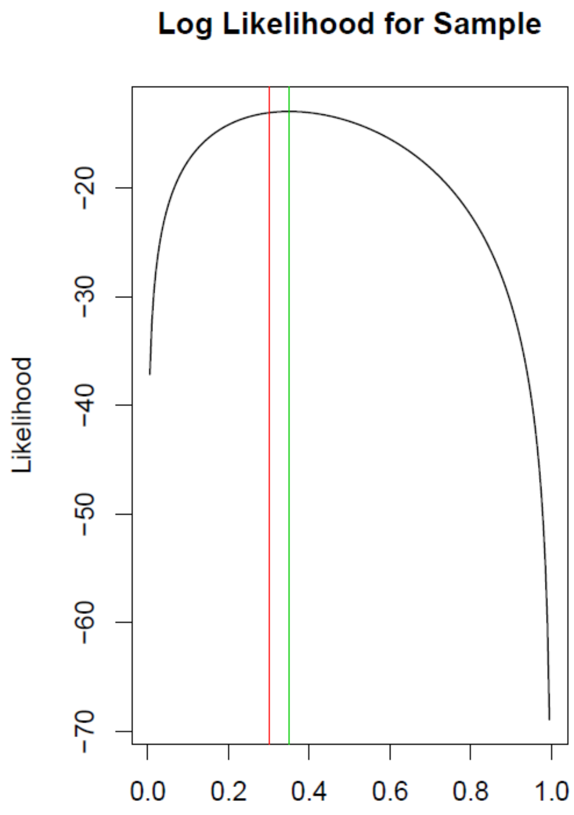
$$logL(\beta) = \log\left[\prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}\right] = \sum_{i=1}^{n}[y_i \log(p) + (1-y_i)\log(1-p)]$$

$$\frac{dlogL(\beta)}{\beta} = 0$$

Estimation of Logistic Model

- Taking log does not change the monotone of any function, so likelihood and log-likelihood would achieve their maximum at the same β





Practice

- Suppose we have a dataset with one predictor X and one binary response Y

- The dataset (x_i, y_i) has three data points

(1, 1)   (2, 1)   (3, 0)

- We use a logistic regression to model the relationship between X and Y

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

- Question: show the likelihood and log-likelihood function

Example

- The wcgs dataset records 3154 men about whether they suffer from heart disease along with many other variables that might be related to the disease

- We are interested in three variables: chd (heart disease, yes or no, response), height, cigs (number of cigarettes smoked per day)

```
lmod <- glm(chd ~ height + cigs, family = binomial, wcgs)
summary(lmod)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -4.50161    1.84186  -2.444   0.0145 *  
height       0.02521    0.02633   0.957   0.3383    
cigs         0.02313    0.00404   5.724 1.04e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Interpretation

- We can write the fitted model as

$$log\frac{p}{1-p} = -4.501 + 0.025 * height + 0.023 * cigs$$

- Taking exponential on both sides gives

$$\frac{p}{1-p} = e^{-4.501}e^{0.025*height}e^{0.023*cigs}$$

- $\frac{p}{1-p}$ is called “odd ratio”, which is the ratio between $P(Y = 1)$ and $P(Y = 0)$

Model Interpretation

- After taking exponential, the model interpretation is
    1. One unit increase of cigarettes smoked per day will increase the odd of getting disease by a factor of $e^{0.023} = 1.023$
    2. Heigh is insignificant

- Recall the exponential parameter is interpreted as percentage change
    - One additional cig will increase the disease odd by 2.3%

- The lmod$fitted.values keeps all fitted probabilities

```
> lmod$fitted.values[1:100]
      2001       2002       2003       2004       2005       2006       2007 
0.11073449 0.09325523 0.05939705 0.08907868 0.09325523 0.06376553 0.06376553 
      2019       2020       2021       2022       2023       2025       2027 
0.06490724 0.06884645 0.06843192 0.06528707 0.10156723 0.08705431 0.06376553 
      2037       2039       2041       2042       2043       2044       2045 
0.07982448 0.06376553 0.06528707 0.06684233 0.14039985 0.05400770 0.06376553 
```

Inference in Logistic Regression

- Because logistic regression is fitted by maximizing the log-likelihood function, the inference of model comparison is based on the likelihood function

- Suppose we want to test two nested models:

H_0: We prefer smaller model S

H_a: We prefer larger model L

- The test statistic we use is the difference between two log-likelihood functions

$$2 * (logL_L - logL_S)$$

where L_L is the likelihood of larger model and L_S is the likelihood of smaller model

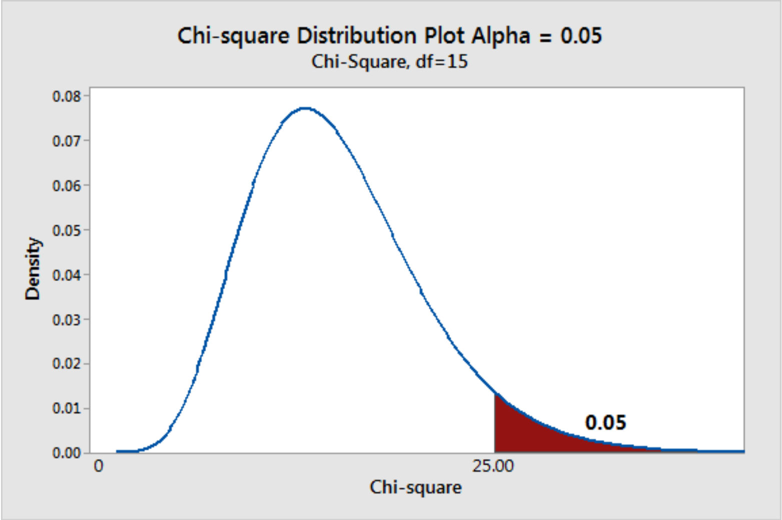
Inference in Logistic Regression

- Under H_0, the test statistic

$$2 * (logL_L - logL_S)$$

follows a Chi-square distribution with a degree of freedom $l - s, (\chi^2_{l-s})$, where l and s are the number of parameters in larger and smaller models

- Since Chi-square distribution is always positive, the p-value is the right area under the curve



Inference in Logistic Regression

- We use anova function to compare nested models by Chi-square test

```
lmod <- glm(chd ~ height + cigs, family = binomial, wcgs)
lmodc <- glm(chd ~ cigs, family = binomial, wcgs)
anova(lmodc,lmod, test="Chi")
```

```
Analysis of Deviance Table

Model 1: chd ~ cigs
Model 2: chd ~ height + cigs
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      3152       1750
2      3151       1749  1  0.92025   0.3374
```

- We fail to reject null hypothesis
  - There is no significant difference between the smaller model with only cigs and the larger model with both cigs and height

- The process of Chi-square test is similar to F-test in the linear regression

Inference in Logistic Regression

- To test the significance of single predictor, our hypothesis is

$$\mathrm{H}_0: \beta_i = 0$$
$$\mathrm{H}_a: \beta_i \neq 0$$

- Again, we use a test statistic base on signal – noise ratio

$$z = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$$

- Under H_0, the z statistic follows a standard normal distribution $N(0, 1)$

- The $se(\hat{\beta}_i)$ is obtained as the inverse of $\frac{d^2 logL(\beta_i)}{\beta_i^2}$ or bootstrap

Inference in Logistic Regression

- We can follow the same method in regular linear regression to construct a confidence interval for single parameter β

$$\hat{\beta}_i \sim N\left(\beta_i, se\left(\hat{\beta}_i\right)\right) \rightarrow \frac{\hat{\beta}_i - \beta_i}{se\left(\hat{\beta}_i\right)} \sim N(0, 1)$$

$$P\left(-z_\alpha < \frac{\hat{\beta}_i - \beta_i}{se\left(\hat{\beta}_i\right)} < z_\alpha\right) = \alpha$$

$$P\left(\hat{\beta}_i - z_\alpha * se\left(\hat{\beta}_i\right) < \beta_i < \hat{\beta}_i - z_\alpha * se\left(\hat{\beta}_i\right)\right) = \alpha$$

- Therefore, the α confidence interval for β_i is

$$\hat{\beta}_i \pm z_\alpha * se\left(\hat{\beta}_i\right)$$

Inference in Logistic Regression

- We can manually construct a 95% CI based on the output of model fitting

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.50161    1.84186  -2.444   0.0145 *
height       0.02521    0.02633   0.957   0.3383
cigs         0.02313    0.00404   5.724 1.04e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Note that now $\alpha = 0.95$, so $z_{0.95} = 1.96$
- The 95% CI for height is

0.02521 + c(-1,1) * 1.96 * 0.02633 → (-0.0263968, 0.0768168)

- The 95% CI for cigs is

0.02313 + c(-1,1) * 1.96 * 0.00404 → (0.0152116, 0.0310484)

Inference in Logistic Regression

- The height's 95% confidence interval covers zero, so we fail to reject H_0: $\beta_i = 0$, height is not significant
- The cigs's 95% confidence interval doesn't cover zero, so we reject H_0: $\beta_i = 0$, height is significant
- The confint function can automatically construct CI for all parameters

```
> confint(lmod)
                  2.5 %      97.5 %
(Intercept) -8.13475465 -0.91297018
height      -0.02619902  0.07702835
cigs         0.01514949  0.03100534
```

Model Selection

- We can still use the model selection methods for regular linear regression, here we take two examples: backward selection and AIC

- We start a logistic model by using all continuous predictors

summary(glm(chd ~ .-behave-dibep-chd-typechd-arcus, family = binomial, wcgs))

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -9.433e+00  2.661e+00  -3.545 0.000393 ***
age          6.767e-02  1.373e-02   4.928 8.32e-07 ***
height       3.038e-02  3.850e-02   0.789 0.430082    
weight       5.892e-03  4.348e-03   1.355 0.175376    
sdp          2.005e-02  7.329e-03   2.736 0.006220 ** 
dbp         -1.548e-02  1.228e-02  -1.261 0.207360    
chol         1.134e-02  1.732e-03   6.548 5.83e-11 ***
cigs         1.639e-02  4.966e-03   3.300 0.000966 ***
timechd     -1.547e-03  8.418e-05 -18.380  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We remove the height variable with largest p-value (>0.05) and refit the model

Model Selection

summary(glm(chd ~ .-behave-dibep-chd-typechd-arcus-height, family = binomial, wcgs))

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -7.510e+00  1.057e+00  -7.104 1.21e-12 ***
age          6.680e-02  1.369e-02   4.881 1.05e-06 ***
weight       7.796e-03  3.613e-03   2.158 0.030946 *  
sdp          2.012e-02  7.338e-03   2.742 0.006101 ** 
dbp         -1.646e-02  1.222e-02  -1.347 0.177991    
chol         1.123e-02  1.727e-03   6.504 7.83e-11 ***
cigs         1.678e-02  4.937e-03   3.399 0.000676 ***
timechd     -1.545e-03  8.404e-05 -18.380  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We remove the dbp variable with largest p-value (>0.05) and refit the model

Model Selection

summary(glm(chd ~ .-behave-dibep-chd-typechd-arcus-height-dbp, family = binomial, wcgs))

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -7.699e+00  1.049e+00  -7.338 2.16e-13 ***
age          6.668e-02  1.370e-02   4.869 1.12e-06 ***
weight       6.792e-03  3.545e-03   1.916 0.055357 .  
sdp          1.243e-02  4.663e-03   2.665 0.007706 ** 
chol         1.113e-02  1.716e-03   6.488 8.68e-11 ***
cigs         1.756e-02  4.892e-03   3.589 0.000331 ***
timechd     -1.536e-03  8.361e-05 -18.369  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We remove the weight variable with largest p-value (>0.05) and refit the model

Model Selection

summary(glm(chd ~ .-behave-dibep-chd-typechd-arcus-height-dbp-weight, family = binomial, wcgs))

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -6.681e+00  8.981e-01  -7.440 1.01e-13 ***
age          6.425e-02  1.362e-02   4.716 2.40e-06 ***
sdp          1.474e-02  4.476e-03   3.294 0.000988 ***
chol         1.108e-02  1.706e-03   6.494 8.36e-11 ***
cigs         1.675e-02  4.885e-03   3.429 0.000606 ***
timechd     -1.543e-03  8.345e-05 -18.486  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Now all predictors are significant (p-value > 0.05)

Model Selection: Step AIC

- Similar to regular linear regression, we define the AIC for logistic regression as

$$AIC = -2logL + 2q$$

- The “best” model should minimize the AIC, because it is negative log-likelihood

- Again, AIC seeks a balance between model fitting (measured by likelihood) and model complexity (number of parameters q)

- The step AIC method starts with a full model; in each step, it removes one predictor which can decrease the current AIC most, until there is no predictor can be removed to decrease AIC

Model Selection: Step AIC

lmod <- glm(chd ~ age + height + weight + sdp + dbp + chol + cigs, family=binomial, wcgs)

lmodr <- step(lmod, trace=T)

```
Start:  AIC=1617.96
chd ~ age + height + weight + sdp + dbp + chol + cigs

         Df Deviance    AIC
- dbp     1   1602.0 1616.0
- height  1   1602.3 1616.3
<none>        1602.0 1618.0
- weight  1   1606.5 1620.5
- sdp     1   1609.9 1623.9
- cigs    1   1629.8 1643.8
- age     1   1634.5 1648.5
- chol    1   1658.5 1672.5
```

```
Step:  AIC=1615.98
chd ~ age + height + weight + sdp + chol + cigs

         Df Deviance    AIC
- height  1   1602.3 1614.3
<none>        1602.0 1616.0
- weight  1   1606.9 1618.9
- sdp     1   1621.5 1633.5
- cigs    1   1630.1 1642.1
- age     1   1634.6 1646.6
- chol    1   1658.7 1670.7
```

```
Step:  AIC=1614.28
chd ~ age + weight + sdp + chol + cigs

         Df Deviance    AIC
<none>        1602.3 1614.3
- weight  1   1611.2 1621.2
- sdp     1   1621.5 1631.5
- cigs    1   1631.2 1641.2
- age     1   1634.6 1644.6
- chol    1   1658.7 1668.7
```

Response with More than Two Levels

- We can generalize the logistic regression to multiple-class case, which is called <u>multinomial logistic regression</u>

- Suppose we have three classes in response variable Y
  - Denote the classes by 1, 2, 3 (Dem, GOP, Ind)

- We can assume that Y is a <u>multinomial</u> random variable

$$Y \sim \begin{cases} P(Y=1) = p_1 \\ P(Y=2) = p_2 \\ P(Y=3) = p_3 = 1 - p_1 - p_2 \end{cases}$$

Multinomial Logistic Regression

- Suppose there are q predictors $X_1, X_2, \dots, X_q$, then we model the probabilities as

$$p_1 = P(Y = 1)$$
$$= \frac{e^{(\beta_{01}+\beta_{11}X_1+\cdots+\beta_{q1}X_q)}}{e^{(\beta_{01}+\beta_{11}X_1+\cdots+\beta_{q1}X_q)} + e^{(\beta_{02}+\beta_{12}X_1+\cdots+\beta_{q2}X_q)} + e^{(\beta_{03}+\beta_{13}X_1+\cdots+\beta_{q3}X_q)}}$$

$$p_2 = P(Y = 2)$$
$$= \frac{e^{(\beta_{02}+\beta_{12}X_1+\cdots+\beta_{q2}X_q)}}{e^{(\beta_{01}+\beta_{11}X_1+\cdots+\beta_{q1}X_q)} + e^{(\beta_{02}+\beta_{12}X_1+\cdots+\beta_{q2}X_q)} + e^{(\beta_{03}+\beta_{13}X_1+\cdots+\beta_{q3}X_q)}}$$

$$p_3 = P(Y = 3) = 1 - p_1 - p_2$$

Multinomial Logistic Regression

- With p_1, p_2, p_3, we can obtain the final three-class model output by

$$Y = \begin{cases} 1 & if\ p_1 = \max(p_1, p_2, p_3) \\ 2 & if\ p_2 = \max(p_1, p_2, p_3) \\ 3 & if\ p_3 = \max(p_1, p_2, p_3) \end{cases}$$

- To estimate the model parameter, we still use the likelihood function

$$L(\beta) = \prod_{i=1}^{n} l_i = \prod_{i=1}^{n} p_1^{I(y_i=1)} p_2^{I(y_i=2)} p_3^{I(y_i=3)}$$

where $I()$ is indication function (if…)

Multinomial Logistic Regression

- Taking log gives the log-likelihood function

$$logL(\beta) = \log\left[\prod_{i=1}^{n} p_1^{I(y_i=1)} p_2^{I(y_i=2)} p_3^{I(y_i=3)}\right]$$

$$= \sum_{i=1}^{n} [I(y_i = 1)\log(p_1) + I(y_i = 2)\log(p_2) + I(y_i = 3)\log(p_3)]$$

- The maximum likelihood estimation of multinomial logistic regression is obtained by

$$\frac{dlogL(\beta)}{\beta} = 0$$

Multinomial Logistic Regression

- In general, if the response variable has K levels, then the multinomial logistic regression model is

$$P(Y = k) = \frac{e^{(\beta_{0k}+\beta_{1k}X_1+\cdots+\beta_{qk}X_q)}}{\sum_{i=1}^{K} e^{(\beta_{0i}+\beta_{1i}X_1+\cdots+\beta_{qi}X_q)}}$$

- The K-class model output is obtained by

$$Y = k \quad if \quad p_k = \max(p_1, p_2, \ldots, p_K)$$