Regression 14

Logistic Regression, Poisson Regression and Generalized Linear Models

1 Regression Models with Binary Responses

We consider regression models with binary (dichotomous) responses. For example, the responses may be alive or dead, failure or success, etc. These binary responses can be quantified as the binary random variable

$$Y_i = \begin{cases} 1 & \text{if the response is success with } P(Y_i = 1) = \pi_i \\ 0 & \text{if the response is failure with } P(Y_i = 0) = 1 - \pi_i \end{cases}$$

That is, Y_i is a Bernoulli random variable with the success probability π_i.

Simple Linear Regression Model

We can consider the simple linear regression model for the binary response

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

Taking the expectation of Y_i with the assumption of $E[\epsilon_i] = 0$, we have

$$E[Y_i] = \beta_0 + \beta_1 X_i.$$

Since Y_i is a Bernoulli random variable with the probability π_i, we have

$$\pi_i = E[Y_i] = P(Y_i = 1) = \beta_0 + \beta_1 X_i.$$

Since $0 \leq \pi \leq 1$, the mean response should satisfy the condition

$$0 \leq \beta_0 + \beta_1 X_i \leq 1.$$

In practice, this condition can not be satisfied. Thus, this model is not widely used. If we know the range of X_i, say $c_1 \leq X_i \leq c_2$, then one of estimating the parameter is to use

$$\hat{\beta}_0 = -\frac{c_1}{c_2 - c_1} \quad \text{and} \quad \hat{\beta}_1 = \frac{1}{c_2 - c_1}.$$

2 Sigmoidal Response Models

When the response is binary (dichotomous), the probability of success,

$$\pi_i = E[Y_i] = P(Y_i = 1)$$

should be between 0 and 1. As seen in the previous simple linear regression model with binary responses, the condition of the response $0 \leq \beta_0 + \beta_1 X_i \leq 1$ is not satisfied in general. One can think of using a reasonable transform of the response function $(\beta_0 + \beta_1 X_i)$ so that $0 \leq \pi_i \leq 1$ with $\pi_i = P(Y_i = 1) = F_D(\beta_0 + \beta_1 X_i)$. The function F_D should be bounded between 0 and 1, and be characterized by a S-shape curve (so-called, sigmoidal). A typical way of choosing this transform is to use a CDF function. Most popular choices are:

(i) Normal distribution (probit model).

(ii) Logistic distribution (logistic/logit model).

(iii) Gumbel (Extreme value) distribution (complementary log-log model, or extreme value model).

In statistics literature, the models using the above CDFs are also called the *probit*, *logit* and *complementary log-log* response functions. The standard normal and logistic CDFs are

Model	CDF	inverse CDF
Probit	$\pi = \Phi(\beta_0 + \beta_1 X)$	$\Phi^{-1}(\pi) = \beta_0 + \beta_1 X$
Logit	$\pi = \dfrac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$	$\log\left(\dfrac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$
c-log-log	$\pi = 1 - \exp\big(-\exp(\beta_0 + \beta_1 X)\big)$	$\log\big(-\log(1-\pi)\big) = \beta_0 + \beta_1 X$

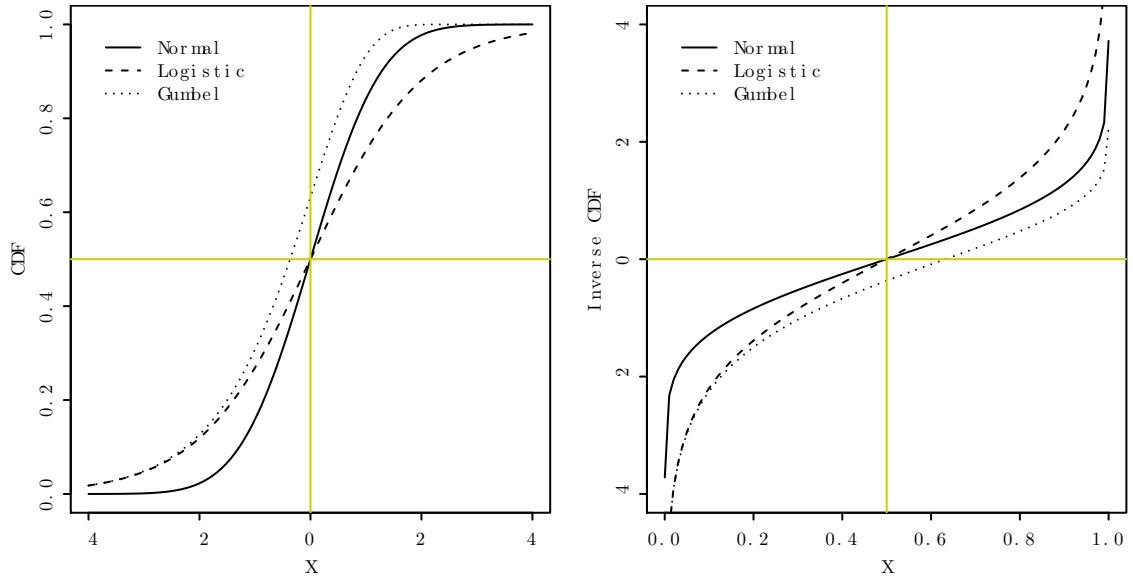

Figure 14.1: CDF and inverse CDF.

skew-symmetric at $(0, \frac{1}{2})$ so are their corresponding inverse CDFs at $(\frac{1}{2}, 0)$. This property is similar to the case of the Bernoulli distribution, for example, $P(Z = 1) = 1 - P(Z = 0)$.

For example, the probit model has $P(Y = 1) = \Phi(\beta_0 + \beta_1 X)$. Then we also have $P(Y = 0) = 1 - P(Y = 1) = 1 - \Phi(\beta_0 + \beta_1 X) = \Phi\big(-(\beta_0 + \beta_1 X)\big)$ since $\Phi(-z) = 1 - \Phi(z)$. Similarly the logit model also has this skew-symmetric property. However, the complementary log-log model does not have this property.

Especially for the logit model, the function

$$\log\left(\frac{\pi}{1-\pi}\right) \tag{14.1}$$

is called the *logit* transformation of the probability π. This logit transformation is actually

the inverse CDF of the logistic distribution. The argument

$$\frac{\pi}{1-\pi}$$

in (14.1) is called the *odds*. For instance, if $\pi = 0.8$, then the odds of success equal $0.8/(1-0.8) = 4$. This implies that a success is four times as likely as a failure, so we expect to observe one failure for every four successes. On the other hand, if the odds is $1/4$, then a failure is four times as likely as a success. For the logit model, the odds are simply given by $\exp(\beta_0 + \beta_1 X)$.

When we have two odds, the ratio of two odds is called the *odds ratio*

$$\theta_{21} = \frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)}.$$

This is a popular measure of association for 2×2 contingency tables in categorical data analysis. Note that the *relative risk* is a ratio of two probabilities, π_2/π_1.

3 Maximum Likelihood Estimation

The responses, Y_i, are Bernoulli random variables with the success probability π_i, where $P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$. Let y_i be observations of the random variables Y_i. Note that the probabilities, π_i, are function of β_0 and β_1 with a predictor X_i. The pmf of Y_i is given by

$$f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

where $y_i = 0, 1, 2, \ldots, n$ and $y_i = 0, 1$. Thus, the likelihood is given by

$$L(\beta_0, \beta_1 | y_1, \ldots, y_2) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}.$$

Taking the logarithm of the above, we have the log-likelihood

$$\begin{aligned}\ell(\beta_0, \beta_1) &= \sum_{i=1}^{n} \Big[y_i \log \pi_i + (1-y_i)\log(1-\pi_i) \Big] \\ &= \sum_{i=1}^{n} \Big[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \Big].\end{aligned}$$

When the logit model is used, the log-likelihood is given by

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \log\left\{1 + \exp(\beta_0 + \beta_1 X_i)\right\}.$$

Similarly, we can derive the log-likelihood for other models.

After the maximum likelihood estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, are calculated, we can obtain the fitted response function. The fitted probit response function is given by

$$\hat{\pi}_i = \Phi(\hat{\beta}_0 + \hat{\beta}_1 X_i),$$

and the fitted logit response function is

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)}$$

Example 1. Textbook Example. See Table 14.1 on Page 566.

A task was given to 25 programmers who have varying amount of programming experience (measured in months). We code $Y = 1$ if the task was completed successfully.

R

Read Data

```
> url = "https://raw.githubusercontent.com/AppliedStat/LM/master/CH14TA01.txt"
> Data = read.table(url)
> x = Data[,1]
> y = Data[,2]
```

GLM

```
> # Table 14.1
> GLM = glm(y~x, family=binomial("logit"))
> summary(GLM)
Call:
glm(formula = y ~ x, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8992  -0.7509  -0.4140   0.7992   1.9624

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05970    1.25935  -2.430   0.0151 *
x            0.16149    0.06498   2.485   0.0129 *
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.296  on 24  degrees of freedom
Residual deviance: 25.425  on 23  degrees of freedom
AIC: 29.425

Number of Fisher Scoring iterations: 4
```

```
>
> cbind(x, y, fitted(GLM), residuals(GLM))
     x y
1   14 0 0.31026237 -0.8619095
2   29 0 0.83526292 -1.8991601
3    6 0 0.10999616 -0.4827618
.............................
```

Plots

```
> # Figure 14.5
> idx = order(x)
> x = sort(x)
> y = y[idx]
> GLM1 = glm(y~x, family=binomial("logit"))

> plot(x,y)
> lines( x, fitted(GLM1) )
> lines( lowess(x,y), lty=2 )
```


```
> # Figure 14.5
> idx = order(x)
> x = sort(x)
> y = y[idx]
> GLM1 = glm(y~x, family=binomial("logit"))

> plot(x,y)
> lines( x, fitted(GLM1) )
> lines( lowess(x,y), lty=2 )
```


```
> # Figure 14.6
> GLM2 = glm(y~x, family=binomial("probit"))
> GLM3 = glm(y~x, family=binomial("cloglog"))

> plot(x,y)
> lines( x, fitted(GLM1) )
> lines( x, fitted(GLM2), lty=2 )
> lines( x, fitted(GLM3), col="red" )
```

It is immediate from the R program that the fitted *logistic* response function (fitted success probability) is

$$\hat{\pi}(X) = \frac{\exp(-3.0597 + 0.1615X)}{1 + \exp(-3.0597 + 0.1615X)}.$$

Note that the *logit* response function is given by $\pi' = \log\big(\pi/(1-\pi)\big)$. Thus, the fitted *logit* response function is

$$\hat{\pi}'(X) = -3.0597 + 0.1615X.$$

‖

Interpretation of $\hat{\beta}_1$

The logit response function is given by

$$\pi'(X) = \beta_0 + \beta_1 X.$$

Thus, the slope parameter β_1 means the increase in the logit response function when the predictor X increases by one,

$$\beta_1 = \frac{\pi'(X+1) - \pi'(X)}{(X+1) - X} = \pi'(X+1) - \pi'(X).$$

Note that $\exp(\pi') = \pi/(1-\pi)$ is known as the odds. Taking the exponential to β_1 above, we have

$$\exp(\beta_1) = \frac{\exp(\pi'(X+1))}{\exp(\pi'(X))} = \frac{\text{odds}_2}{\text{odds}_1}.$$

where odds_1 denotes $\exp(\pi'(X))$ and odds_2 denotes $\exp(\pi'(X+1))$. Thus, $\exp(\beta_1)$ is the ratio of two odds (*odds ratio*) when the predictor value is increased by one.

In the previous example, we have $\hat{\beta}_1 = 0.1615$ which gives the odds ratio $\hat{\theta} = \exp(0.1615) = 1.175$. The increase in the odds is easily obtained as

$$\theta - 1 = \frac{\text{odds}_2 - \text{odds}_1}{\text{odds}_1}.$$

Thus, in this example, the increase in the odds is 0.175 due to an additional month of experience. That is, the odds of success (completing task) increase by 17.5% due to one-month additional experience.

Binomial outcomes due to repeated measurements

Suppose that we observe binomial outcomes repeated at the level X. The responses of the ith binary response at X_j are denoted by Y_{ij}, where $i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, c$. If we denote the binary response to be $Y_{ij} = 1$ when the response is success, then the number of successes at level X_j is given by $Y_{\bullet j}$ where

$$Y_{\bullet j} = \sum_{i=1}^{n_j} = Y_{ij}. \tag{14.2}$$

For notational convenience, we denote $Y_{\bullet j}$ by Y_j. The proportion of successes at level X_j is also easily calculated as

$$p_j = \frac{Y_j}{n_j}.$$

If the random variable Y_{ij} is a Bernoulli random variable with the success probability π_j, then Y_j is a binomial random variable

$$Y_j \sim \text{Bin}(n_j, \pi_j). \tag{14.3}$$

The probability mass function of the binomial random variable Y_j is given by

$$p(y_j|\pi) = \binom{n}{y_j} \pi^{y_j}(1-\pi)^{n-y_j} \quad \text{if} \quad y_j = 0, 1, 2, \ldots, n_j, \tag{14.4}$$

where y_j is a realization of random variable Y_j. The likelihood function $L(\cdot)$ is given by

$$L(\pi_1, \ldots, \pi_c) = \prod_{j=1}^{c} \binom{n_j}{y_j} \pi_j^{y_j}(1-\pi_j)^{n_j-y_j}.$$

By taking the logarithm function, we have the log-likelihood function

$$\begin{aligned} \ell(\pi_1, \ldots, \pi_c) &= \sum_{j=1}^{c} \left[\log \binom{n_j}{y_j} + y_j \log \pi_j + (n_j - y_j) \log(1-\pi_i) \right] \\ &= \sum_{j=1}^{c} \left[y_j \log \pi_j + (n_j - y_j) \log(1-\pi_j) \right] + C. \end{aligned} \tag{14.5}$$

It is immediate from differentiating (14.5) that the maximum likelihood estimate for the unknown proportion parameter π_j is given by $\pi_j = y_j/n_j$. In many practical cases, however, the proportion parameter π_i in the ith subgroup should be characterized by other predictor(s). As we have studied as before, the relation between the proportion parameter π_j and predictor(s) with regression parameters is given by probit model, logistic model, complementary log-log model, etc.

Example 2. Textbook Example. See Table 14.2 on Page 569.

The predictor X is the amount of price reduction of coupons and the response Y is the number of coupons redeemed.

R

Read Data

```
> # Table 14.2 on Page 569
> url = "https://raw.githubusercontent.com/AppliedStat/LM/master/CH14TA02.txt"
> Data = read.table(url)
> Xj = Data[,1]
> nj = Data[,2]
```

```
> Yj = Data[,3]
> pj = Data[,4]
> Data
  V1  V2  V3    V4
1  5 200  30 0.150
2 10 200  55 0.275
3 15 200  70 0.350
4 20 200 100 0.500
5 30 200 137 0.685
```

GLM

```
> # Wrong results
> GLM = glm(Yj~Xj, family=binomial("logit"))
Error in eval(family$initialize) : y values must be 0 <= y <= 1

> # Correct results (convert binomial obs to Bernoulli obs)
> X1 = rep(Xj,Yj)
> X0 = rep(Xj,nj-Yj)
> Y1 = rep(1, sum(Yj))
> Y0 = rep(0, sum(nj-Yj))
> X = c(X1,X0)
> Y = c(Y1,Y0)
> GLM = glm(Y~X, family=binomial("logit"))
> summary(GLM)
Call:
glm(formula = Y ~ X, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5578  -0.9385  -0.6176   1.2235   1.8713

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.044348   0.160977  -12.70   <2e-16 ***
X            0.096834   0.008549   11.33   <2e-16 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1339.3  on 999  degrees of freedom
Residual deviance: 1192.0  on 998  degrees of freedom
AIC: 1196

> fitted(GLM) # Too many values b/c Bernoulli obs. Use predict

> pj.hat = predict(GLM, type="response", data.frame(X=Xj))
> pj.hat # The 5th col. in Table 14.2
        1         2         3         4         5
0.1736208 0.2542615 0.3562119 0.4731071 0.7027987

> cbind(Data, pj.hat) # TABLE 14.2 on Page 569

> plot(Xj, pj, xlim=c(0,40), ylim=c(0,1)) # OK
> b = coef(GLM)
> curve( exp(b[1]+b[2]*x)/(1+exp(b[1]+b[2]*x)), 0,40, add=TRUE)

> # Better version
> YjNj = cbind(Yj, nj-Yj)
> GLM2 = glm(YjNj~Xj, family=binomial("logit"))
> summary(GLM2)
Call:
glm(formula = YjNj ~ Xj, family = binomial("logit"))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.044348   0.160977  -12.70   <2e-16 ***
Xj           0.096834   0.008549   11.33   <2e-16 ***
```

4 Multiple Logistic Regression

We studied the logistic regression model with one predictor. It is natural to extend this simple model to a more complex model by replacing $\beta_0+\beta_1 X$ with $\beta_0+\beta_1 X_1+\cdots+\beta_{p-1}X_{p-1}$.

Using the matrix notation, we can simplify the multiple logistic regression formulas. We need to define the following matrices:

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \underset{p\times 1}{\mathbf{x}} = \begin{bmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_{p-1} \end{bmatrix} \qquad \underset{p\times 1}{\mathbf{x}_i} = \begin{bmatrix} 1 \\ X_{i1} \\ X_{i2} \\ \vdots \\ X_{i,p-1} \end{bmatrix}$$

It is immediate from the matrix notation that we have

$$\mathbf{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1}X_{p-1}$$

$$\mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1}X_{i,p-1}.$$

Thus, using the above notation, we can write the multiple logistic response function as

$$\pi = E[Y] = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{x}'\boldsymbol{\beta})}.$$

Using the logit transform $\pi' = \log\big(\pi/(1-\pi)\big)$, we can write the logit response function as

$$\pi' = \mathbf{x}'\boldsymbol{\beta}.$$

Thus, the multiple logistic regression model is that the response of the ith binary response (Y_i) at the level $\mathbf{x}_i$ is distributed as *iid* Bernoulli random variables with the success probability

$$\pi_i = E[Y_i] = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}.$$

Likelihood function

The log-likelihood function for multiple logistic regression is given by

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \cdot \mathbf{x}_i'\boldsymbol{\beta} - \sum_{i=1}^{n} \log\left\{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})\right\}. \tag{14.6}$$

After the parameters are estimated using the above log-likelihood function which we denote by $\hat{\boldsymbol{\beta}}$, the fitted logistic response function is also easily obtained by

$$\hat{\pi} = \frac{\exp(\mathbf{x}'\hat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{x}'\hat{\boldsymbol{\beta}})}.$$

Example 3. Textbook Example. See Table 14.3 on Page 574.

The response variable Y was coded 1 if a person contracted a disease and 0 if not. Three predictor variables were included and these are age (X_1), socioeconomic status (X_2 and X_3 dummy code), and city sector (X_4). Socioeconomic status has three levels which are upper ($X_2 = 0$ and $X_3 = 0$), middle ($X_2 = 1$ and $X_3 = 0$), and lower ($X_2 = 0$ and $X_3 = 1$). City sector has only two sectors so that $X_4 = 0$ for Sector 1 and $X_4 = 1$ for Sector 2.

R

Read Data

```
> url = "https://raw.githubusercontent.com/AppliedStat/LM/master/CH14TA03.txt"
> Data = read.table(url)
> x1 = Data[,2]
> x2 = Data[,3]
> x3 = Data[,4]
> x4 = Data[,5]
> y  = Data[,6]
```

GLM

```
> # Table 14.3
> GLM = glm(y~x1+x2+x3+x4, family=binomial("logit"))

> # Table 14.4 (a)
> summary(GLM)
Call:
glm(formula = y ~ x1 + x2 + x3 + x4, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6552  -0.7529  -0.4788   0.8558   2.0977

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.31293    0.64259  -3.599 0.000319 ***
x1           0.02975    0.01350   2.203 0.027577 *
x2           0.40879    0.59900   0.682 0.494954
x3          -0.30525    0.60413  -0.505 0.613362
```

```
x4           1.57475    0.50162   3.139 0.001693 **
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 122.32  on 97  degrees of freedom
Residual deviance: 101.05  on 93  degrees of freedom
AIC: 111.05

Number of Fisher Scoring iterations: 4

> cbind( Data[,2:6], fitted(GLM) )
   V2 V3 V4 V5 V6 fitted(GLM)
1  33  0  0  0  0  0.20896395
2  35  0  0  0  0  0.21896953
3   6  0  0  0  0  0.10579477
4  60  0  0  0  0  0.37099998
 ...................

97 11  0  1  0  0  0.09187623
98 35  0  1  0  0  0.17122984


> # Estimated Odds Ratio
> exp( coef(GLM)[-1] )
       x1        x2        x3        x4
1.0301970 1.5049960 0.7369358 4.8295304


> # Table 14.4 (a) Estimated Odds Ratio
> exp( coef(GLM)[-1] )
       x1        x2        x3        x4
1.0301970 1.5049960 0.7369358 4.8295304

> # Table 14.4 (b) Estimated variance-covariance matrix
> vcov(GLM)
             (Intercept)            x1           x2            x3
(Intercept)  0.412919186 -0.0057142121 -0.183574952 -0.2009773348
x1          -0.005714212  0.0001823259  0.001149844  0.0007319103

.............................
```

Polynomial Logistic Regression

In Chapter 6, we have studied a polynomial regression model with one predictor variable. For example,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \tag{14.7}$$

is also a special case of the general linear regression model despite the curvilinear nature of the response function in (14.7). If we define

$$X_{i1} = X_i \quad \text{and} \quad X_{i2} = X_i^2,$$

we can write (14.7) as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

which is in the form of the general linear regression model. Similarly models with higher-degree polynomial response functions are also particular cases of the general linear regression model.

We can build the kth order polynomial logistics regression model similar to the above general linear regression model. Using the logit response function, we have

$$\pi' = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k.$$

Note that the textbook asks one to center the predictor variable, but it is not necessary in general.

Example 4. Textbook Example. See Appendix C.11 and Table 14.5 on Page 576.

The response variable Y_i is the involvement of a venture capital. The original predictor X_i is the natural logarithm of the face value of a company. It seems that the simple logistic regression model is not inadequate. The second order polynomial logistic regression model looks more plausible. In the textbook example, the predictor is transformed by centering (de-meaned), $x_i = X_i - \bar{X}$. Thus, we de-meaned the predictor although it is not necessary. Note, after the transformation, the intercept estimate, β_0, can be changed, but the other estimates are unchanged.

R

Read Data

```
> url = "https://raw.githubusercontent.com/AppliedStat/LM/master/APPENC11.txt"
> Data = read.table(url)
> y = Data[,2]
> x =  log(Data[,3])
> idx = order(x)
> x = sort(x)     # sort x
> y = y[idx]      # sort y according to the order of x
> x = x - mean(x)
> x2 = x^2
```

GLM

```
> # Table 14.5 on Page 576
> GLM1 = glm( y ~ x , family=binomial("logit"))
> summary(GLM1)
Call:
glm(formula = y ~ x, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-1.5695  -1.0773  -0.8279   1.2141   1.6793

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.25229    0.09364  -2.694  0.00706 **
x            0.44407    0.10752   4.130 3.62e-05 ***
---
```

Plots

```
> par( mfrow=c(1,2) )
> # Figure 14.9 (a)
> plot(x,y)
> lines( x, fitted(GLM1) )
> lines(lowess(x,y), lty=2)
> # Figure 14.9 (b)
> # GLM2 = glm( y ~ x + x2 , family=binomial("logit")) # OK
> #        glm( y ~ x + x^2, family=binomial("logit")) # Wrong
> GLM2 = glm( y ~ x + I(x^2) , family=binomial("logit"))
> summary(GLM2)
Call:
glm(formula = y ~ x + I(x^2), family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3463  -1.1681  -0.4297   1.0590   2.8889

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3005     0.1240   2.424   0.0154 *
x             0.5516     0.1385   3.984 6.78e-05 ***
I(x^2)       -0.8615     0.1404  -6.136 8.46e-10 ***
---

> # Figure 14.9 (b)
> plot(x,y)
> lines(lowess(x,y), lty=2)
> xx = seq(min(x), max(x), l=50)
> yy = predict(GLM2, type="response", data.frame(x=xx) )
> lines( xx, yy )
```

5 Inferences about Regression Parameters

Let $\ell_{ij}(\boldsymbol{\beta})$ be the second-order partial derivatives of the log-likelihood function such as in (14.6)

$$\ell_{ij}(\boldsymbol{\beta}) = \frac{\partial^2}{\partial\beta_i\partial\beta_j}\ell(\boldsymbol{\beta}),$$

where $i, j = 1, 2, \ldots, p-1$. Let $I_o(\boldsymbol{\beta})$ be the observed Fisher information matrix

$$I_o(\boldsymbol{\beta}) = \Big[-\ell_{ij}(\boldsymbol{\beta})\Big].$$

Then the covariance-variance matrix of the estimated parameters can be approximated by the inverse of the observed Fisher information matrix

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) \approx I_o(\boldsymbol{\beta})^{-1} \approx I_o(\hat{\boldsymbol{\beta}})^{-1}.$$

Let $\left\{\text{SE}(\hat{\beta}_j)\right\}^2$ be the jth diagonal component of the above matrix. Then we have an approximate test statistic for the regression parameter

$$Z = \frac{\hat{\beta}_j - \beta_j}{\text{SE}(\hat{\beta}_j)} \overset{\bullet}{\sim} N(0,1), \tag{14.8}$$

where $j = 0, 1, \ldots, p-1$. Thus, using (14.8), we can perform a statistical hypothesis test of the null $H_0 : \beta_j = 0$ against the alternative.

Example 5. Textbook example on Page 578.

In Example 1, we estimated the regression parameters. We expect that the slope parameter β_1 is positive. Thus, the hypotheses of interest are $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 > 0$.

R

Read Data

```
> url = "https://raw.githubusercontent.com/AppliedStat/LM/master/CH14TA01.txt"
> Data = read.table(url)
>x = Data[,1]
>y = Data[,2]
```

GLM

```
># Table 14.1
>GLM = glm(y~x, family=binomial("logit"))
>summary(GLM)

Call:
glm(formula = y ~ x, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8992  -0.7509  -0.4140   0.7992   1.9624

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05970    1.25935  -2.430   0.0151 *
x            0.16149    0.06498   2.485   0.0129 *
---
```

From the R output, we have $Z = 2.485$. Thus, for $\alpha = 0.05$, the critical value is given by the 5% *upper* percentile denoted by $z_\alpha = 1.645$ and the p-value is $1 - \Phi(2.485) = 0.00647$. Thus, we reject the null hypothesis. So, we can expect that the slope is positive. Note that the p-value in the R output is given by 0.0129 which is for testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. ‖

Using (14.8), we can also obtain the approximate $(1-\alpha)$ confidence limits for β_k

$$\hat{\beta}_k \pm z_{\alpha/2} \cdot \mathrm{SE}(\hat{\beta}_k).$$

If one is interested in the corresponding confidence limits for the odds ratio $\exp(\beta_k)$, this is easily calculated by take the exponential of this limits

$$\exp\left(\hat{\beta}_k \pm z_{\alpha/2} \cdot \mathrm{SE}(\hat{\beta}_k)\right).$$

If one is interested in one-sided confidence interval, this can be easily obtained by

$$\left[\hat{\beta}_k - z_{\alpha} \cdot \mathrm{SE}(\hat{\beta}_k), \infty\right) \quad \text{or} \quad \left(-\infty, \hat{\beta}_k + z_{\alpha} \cdot \mathrm{SE}(\hat{\beta}_k)\right].$$

Example 6. In the previous example, we can also find the confidence limits. The approximate 95% confidence limits for β_1 is obtained as

$$0.16149 \pm 1.96 \cdot 0.06498 = 0.16149 \pm 0.1273608,$$

which results in the approximate 95% confidence interval,

$$[0.0341292, 0.2888508].$$

The approximate 95% confidence interval for the odds ratio $\exp(\beta_k)$ is also easily obtained by taking the exponential of $[0.0341292, 0.2888508]$ which is obtained as

$$[1.034718, 1.334893].$$

However, it should be noted that a symmetric confidence interval using the standard error may not be appropriate for some nonlinear regression models. Instead, an *asymmetric* confidence interval can be more appropriate. In nonlinear regression it is customary to invert the "extra sum of square" to provide such an asymmetric confidence interval, which is usually almost identical the confidence interval based on the likelihood ratio statistic. The R provides a function for this interval `confint()`. Using `confint`, one can obtain a 95% asymmetric confidence interval

$$[0.05002505, 0.3140397].$$

Then the approximate 95% confidence interval for the odds ratio $\exp(\beta_k)$ is obtained by taking the exponential of $0.05002505, 0.3140397]$ which is given by

$$[1.051297, 1.368944].$$

On the other hand, the approximate 95% one-sided confidence interval is given by

$$[0.16149 - 1.645 \cdot 0.06498, \infty) = [0.0545979, \infty).$$

We can also find this one-sided confidence interval using `confint()` in R. For 95% coverage, first find the *two-sided* confidence with $1 - \alpha = 0.9$ and take the lower limit from the result of `confint()`. Using `confint(..., level=0.90)`, we have the 90% two-sided confidence interval, $[0.06628211, 0.2855569]$. Then the 95% *one-sided* confidence interval is obtained as $[0.06628211, \infty)$. ‖

Test concerning several regression parameters using the likelihood ratio test statistic

We are interested in testing the significance of a group of additional predictors. This is equivalent to testing whether a group of the associated regression parameters are zero or not.

We incorporate the likelihood ratio test method of hypothesis into the logistic regression model. The likelihood ratio test statistic is a very general statistic which can be used in a variety of applications.

Definition 1. The likelihood ratio test (LRT) statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ is given by

$$\lambda = \frac{\sup_{\Theta_0} L(\theta)}{\sup_{\Theta} L(\theta)}, \tag{14.9}$$

where $\Theta = \Theta_0 \cup \Theta_1$.

A LRT is any test that has a rejection region which is in $\lambda < c$. Note that $0 \leq \lambda \leq 1$ and thus $0 \leq c \leq 1$.

Theorem 1. *The distribution of the LRT statistic, $-2\log\lambda$, is approximately distributed as a chi-square. The degrees of freedom of the chi-square distribution are the difference between the number of free parameters in $\theta \in \Theta_0$ and those in $\theta \in \Theta$.*

Thus, the rejection region for testing $H_0 : \theta \in \Theta_0$ is

$$-2\log\lambda > \chi^2(1-\alpha; \mathrm{df}),$$

where α is the significance level and df is the corresponding degrees of freedom.

We consider the *full* logistic model with response function

$$\pi = \frac{\exp(\mathbf{X}'\boldsymbol{\beta}_F)}{1+\exp(\mathbf{X}'\boldsymbol{\beta}_F)}.$$

Then the hypothesis test is equivalent to testing

$$H_0 : \beta_q = \beta_{q+1} = \ldots = \beta_{p-1} = 0$$

$$H_1 : \text{not all of } \beta_q, \beta_{q+1}, \ldots, \beta_{p-1} \text{ equal zero,}$$

where, for convenience, the regression parameters are arranged so that the last $p-q$ parameters are tested. The above test is equivalent to testing

$$H_0 : \mathbf{X}'\boldsymbol{\beta}_R = \beta_0 + \beta_1 X_1 + \cdots + \beta_{q-1}X_{q-1}$$

$$H_1 : \mathbf{X}'\boldsymbol{\beta}_F = \beta_0 + \beta_1 X_1 + \cdots + \beta_{q-1}X_{q-1} + \cdots + \beta_{p-1}X_{p-1}.$$

For convenience, let $D = -2\log\lambda$ and L_0 be the likelihood under H_0 (reduced model) and L_1 be the likelihood under H_1 (full model). The approximate rejection region is given by

$$D = -2(\log L_0 - \log L_1) = 2(\log L_1 - \log L_0) > \chi^2(1-\alpha; p-q). \tag{14.10}$$

Example 7. Textbook example on Page 581.

In Example 3 (or Textbook Table 14.3 on Page 574), we analyzed the disease outbreak data. The response variable Y was coded 1 if a person contracted a disease and 0 if not. Three predictor variables were included and these are age (X_1), socioeconomic status (X_2 and X_3 dummy code), and city sector (X_4). Socioeconomic status has three levels which are upper ($X_2 = 0$ and $X_3 = 0$), middle ($X_2 = 1$ and $X_3 = 0$), and lower ($X_2 = 0$ and $X_3 = 1$). City sector has only two sectors so that $X_4 = 0$ for Sector 1 and $X_4 = 1$ for Sector 2.

We want to test if the first predictor (X_1, age) can be dropped from the logistic model. That is, we want to test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.

R

Read Data

```
> url = "https://raw.githubusercontent.com/AppliedStat/LM/master/CH14TA03.txt"
> Data = read.table(url)
> x1 = Data[,2]
> x2 = Data[,3]
> x3 = Data[,4]
> x4 = Data[,5]
> y  = Data[,6]
```

GLM

```
> GLM0 = glm(y~   x2+x3+x4, family=binomial("logit"))
> GLM1 = glm(y~x1+x2+x3+x4, family=binomial("logit"))

> D = deviance(GLM0) - deviance(GLM1)
> D
[1] 5.149519

> qchisq(0.95, df=1)
[1] 3.841459

> # p-value
> 1-pchisq( D, df=1)
[1] 0.02325281
```

It should be noted that `deviance()` in R calculates $2(\ell_{\max} - \ell(\hat{\boldsymbol{\beta}}))$, where $\ell_{\max}$ is the log-likelihood under the "complete full" model in (14.16) (without any restriction in response function) and $\ell(\hat{\boldsymbol{\beta}})$ is the log-likelihood with the estimates. In the above R program, `deviance(GLM0)` thus calculates $2(\ell_{\max} - \ell(\hat{\boldsymbol{\beta}}_R))$ and `deviance(GLM1)` calculates $2(\ell_{\max} - \ell(\hat{\boldsymbol{\beta}}_F))$.

The null deviance, $2\big(\ell_{\max} - \ell(\emptyset)\big)$, corresponds to SSTo, $2\big(\ell_{\max} - \ell(\hat{\boldsymbol{\beta}})\big)$ to SSE, and

$2\big(\ell(\hat{\boldsymbol{\beta}}) - \ell(\emptyset)\big)$ to SSR in the ordinary regression. Here $\ell(\emptyset)$ is the log-likelihood function without any predictors (only intercept is used).

For $\alpha = 0.05$, the rejection region is given by

$$D > 3.841459.$$

Since $D = 5.149519 > 3.841459$, the null hypothesis is rejected. So we conclude that X_1 can not be dropped from the model. Note that the p-value is 0.02325281.

Next, we consider the following full model which includes all possible two-factor interactions so that we have

$$\begin{aligned}\mathbf{X}'\boldsymbol{\beta}_F = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \\ + \beta_5 X_1 X_2 + \beta_6 X_1 X_3 + \beta_7 X_1 X_4 + \beta_8 X_2 X_4 + \beta_9 X_3 X_4.\end{aligned}$$

We want to test

$$H_0 : \beta_5 = \beta_6 = \beta_6 = \beta_6 = \beta_9 = 0$$

$$H_1 : \text{not all of } \beta_5, \beta_6, \ldots, \beta_9 \text{ equal zero}$$

so that we have

$$\mathbf{X}'\boldsymbol{\beta}_R = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4.$$

The rejection region is given by Equation (14.10). Since the degrees of freedom are 5 and the chi-square critical value for $\alpha = 0.05$ is 11.0705, the null hypothesis is rejected when $D > 11.0705$. Since $D = 7.0583$ as the R program below, we can not reject the null hypothesis. Note that the p-value is 0.2163404.

GLM

```
> GLM0 = glm(y~x1+x2+x3+x4, family=binomial("logit"))
> GLM1 = glm(y~x1+x2+x3+x4+I(x1*x2)+I(x1*x3)+I(x1*x4)+I(x2*x4)+I(x3*x4),
+            family=binomial("logit"))
> D = deviance(GLM0) - deviance(GLM1)
> D
[1] 7.058277
> qchisq(0.95, df=5)
[1] 11.0705
> anova(GLM0, GLM1)
```

```
Analysis of Deviance Table

Model 1: y ~ x1 + x2 + x3 + x4
Model 2: y ~ x1 + x2 + x3 + x4 + I(x1 * x2) + I(x1 * x3) + I(x1 * x4) +
    I(x2 * x4) + I(x3 * x4)
  Resid. Df Resid. Dev Df Deviance
1        93    101.054
2        88     93.996  5   7.0583
```

6 Model Selection

There are several methods for model selection. The idea of model selection with logistic regression models is very similar to that with regular regression models in Chapter 9.

We can perform best subsets procedures with the R package, `bestglm`, but it is not so satisfactory in my opinion.

7 Tests for Goodness of Fit

First, we introduce two statistics for goodness-of-fit test which measure how well an observed contingency table (coined by Karl Pearson) to the statistical model (in general, multinomial model is frequently used). These two statistics are the Pearson goodness-of-fit and the deviance (also called the log-likelihood-ratio test statistic).

The *Pearson goodness-of-fit* statistic is defined by

$$X^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j},$$

where O_j is the observed frequency in the jth cell and E_j is the expected frequency in the jth cell under the model. The *deviance* statistic is defined by

$$G^2 = 2\sum_{j=1}^{k} O_j \log\left(\frac{O_j}{E_j}\right),$$

where log is the natural logarithm and $0 \cdot \log 0 = 0$ (for example, when $O_j = 0$).

Testing goodness of fit under the multinomial model

The two statistics, X^2 and G^2, both measure how closely the multinomial model fits the observed frequency. Let x denote a realization of the random variable X distributed as

the multinomial with size n and probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)$ where $\pi_k = 1 - \sum_{j=1}^{k-1} \pi_j$. The distributions of both X^2 and G^2 are approximated by the chi-square distribution with $k-1$ degrees of freedom. It should be noted that the result about the number of degrees of freedom is valid when the original data are multinomial and hence the estimated parameters are efficient for minimizing the chi-square statistic. However, when maximum likelihood estimation does not coincide with minimum chi-square estimation, the distribution will lie somewhere between a chi-square distribution with $n-1-p$ and $n-1$ degrees of freedom (Chernoff and Lehmann, 1954; Berkson, 1980).

This implies that we can easily test

$$H_0 : \pi = \pi_0 \quad \text{versus} \quad H_1 : \pi \neq \pi_0.$$

Example 8. Suppose that we roll a die 30 times and observe the following tally table.

Face	1	2	3	4	5	6	Total
Count	3	7	5	10	2	3	30

We want to test if the die is fair. The null hypothesis under the multinomial model is

$$H_0 : \boldsymbol{\pi}_0 = \Big(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\Big).$$

Under H_0, the expected counts are given by $E_j = n\pi_j = 30 \cdot \frac{1}{6} = 5$. Thus, we have the following table for E_j and O_j.

E_j	5	5	5	5	5	5
O_j	3	7	5	10	2	3

The Pearson goodness-of-fit statistic and the deviance statistic are

$$X^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j} = 9.2$$

and

$$G^2 = 2\sum_{j=1}^{k} O_j \log\Big(\frac{O_j}{E_j}\Big) = 8.8.$$

The p-values for the above statistics are $P(\chi^2_5 > 9.2) = 0.1013$ and $P(\chi^2_5 > 8.8) = 0.1173$, respectively. Thus, we can not reject the null hypothesis, which implies that the fit is fine enough to conclude beyond to reasonable doubt that the die is unfair. ‖

In many practical cases, we do not know the values of proportions in $\boldsymbol{\pi}$, but can only specify it up to some unknown parameters. Let $\hat{\boldsymbol{\pi}}_0$ be the estimated proportions under the null hypothesis H_0 and $\hat{\boldsymbol{\pi}}_1$ be the estimated proportions under the alternative H_1. The expected cell counts $(E_1, \ldots, E_k)$ are estimated by $n\hat{\boldsymbol{\pi}}_0$ and the observed cell counts equivalent to estimating them by $n\hat{\boldsymbol{\pi}}_1$. Thus, we can write X^2 and G^2 as $X^2(\hat{\boldsymbol{\pi}}_0, \hat{\boldsymbol{\pi}}_1)$ and $G^2(\hat{\boldsymbol{\pi}}_0, \hat{\boldsymbol{\pi}}_1)$. The distributions of both $X^2(\hat{\boldsymbol{\pi}}_0, \hat{\boldsymbol{\pi}}_1)$ and $G^2(\hat{\boldsymbol{\pi}}_0, \hat{\boldsymbol{\pi}}_1)$ are approximated by the chi-square distribution with the degrees of freedom equal to the number of unknown parameters under the alternative hypothesis minus the number of unknown parameters under the null hypothesis.

Testing goodness of fit under the logistic regression model

We apply the idea of the above goodness of fit to the logistic regression model. We want to test

$$H_0 : \pi = \frac{\exp(\mathbf{X}'\boldsymbol{\beta})}{1 + \exp(\mathbf{X}'\boldsymbol{\beta})} \tag{14.11}$$

versus H_1: the model in H_0 is not appropriate. As was the case with the lack-of-fit test in linear regression model, we denote the number of distinct values of the predictors by c, the ith binary response with the vector of the predictors $\mathbf{x}_j$ by Y_{ij} and the number of cases in the jth class by n_j, where $j = 1, 2, \ldots, c$. If we denote the binary response variable to be $Y_{ij} = 1$ with a success, then the response variable Y_{ij} is a Bernoulli random variable with the success probability π_j. As seen in (14.2) and (14.3), $Y_{\bullet j} = \sum_{i=1}^{n_j} Y_{ij}$ is a binomial random variable with size n_j and probability π_j which is equivalent to a multinomial random variable with size n_j and probabilities $\boldsymbol{\pi}_j = (\pi_j, 1 - \pi_j)$. It should be noted that there are two different outcome cells (say, cells of *success* and *failure*) under this multinomial model.

The number of cases in the jth class with outcome 1 (say, success) is denoted by $O_j^{(1)}$

and that with outcome 0 (say, failure) by $O_j^{(0)}$. The estimated expected number of cases in the jth class with outcome 1 is denoted by $E_j^{(1)}$ and that with outcome 0 (say, failure) by $E_j^{(0)}$. It should be noted that

$$O_j^{(0)} = n_j - O_j^{(1)} \quad \text{and} \quad E_j^{(0)} = n_j - E_j^{(1)}.$$

If the logistic response function is appropriate, the expected value of Y_{ij} is given by

$$\pi_j = \frac{\exp(\mathbf{x}_j'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_j'\boldsymbol{\beta})},$$

where $\mathbf{x}_j' = (1, X_{j1}, X_{j2}, \ldots, X_{j,p-1})$. The estimate of π_j is obtained by

$$\hat{\pi}_j = \frac{\exp(\mathbf{x}_j'\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}_j'\hat{\boldsymbol{\beta}})}, \tag{14.12}$$

where $\hat{\boldsymbol{\beta}}$ are the estimated values of the parameters, usually the MLE. Thus, we can obtain $E_j^{(1)}$ and $E_j^{(0)}$ using

$$E_j^{(1)} = n_j\hat{\pi}_j \quad \text{and} \quad E_j^{(0)} = n_j(1-\hat{\pi}_j).$$

The Pearson goodness-of-fit statistic is then given by

$$X^2 = \sum_{j=1}^{c} \frac{(O_j^{(0)} - E_j^{(0)})^2}{E_j^{(0)}} + \sum_{j=1}^{c} \frac{(O_j^{(1)} - E_j^{(1)})^2}{E_j^{(1)}}$$

and the deviance statistic is

$$G^2 = 2\sum_{j=1}^{c} O_j^{(0)} \log\Big(\frac{O_j^{(0)}}{E_j^{(0)}}\Big) + 2\sum_{j=1}^{c} O_j^{(1)} \log\Big(\frac{O_j^{(1)}}{E_j^{(1)}}\Big). \tag{14.13}$$

Both statistics have an approximate chi-square distribution. Under the null hypothesis, we use the p free parameters and under the alternative, we use c distinct values of predictors so the degrees of freedom are $c-p$. We reject the null hypothesis at the level α when

$$X^2 > \chi^2(1-\alpha; c-p) \quad \text{or} \quad G^2 > \chi^2(1-\alpha; c-p).$$

Example 9. Textbook Example on Page 587 (see also Table 14.2 on Page 569).

The predictor X is the amount of price reduction of coupons and the response Y is the number of coupons redeemed. The estimated values of $\boldsymbol{\pi}$ was $\hat{\boldsymbol{\pi}} = (0.1736208, 0.2542615, 0.3562119, 0.4731071, 0.7027987)$. Using these, we have the following values.

j	X_j	n_j	$\hat{\pi}_j$	$O_j^{(0)}$	$E_j^{(0)}$	$O_j^{(1)}$	$E_j^{(1)}$
1	5	200	0.1736208	170	165.28	30	34.72
2	10	200	0.2542615	145	149.15	55	50.85
3	15	200	0.3562119	130	128.76	70	71.24
4	20	200	0.4731071	100	105.38	100	94.62
5	30	200	0.7027987	63	59.44	137	140.56

The degrees of freedom of the chi-square distribution are 3 due to $c = 5$ and $p = 2$. For $\alpha = 0.05$, the rejection region is $X^2 > 7.814728$ or $G^2 > 7.814728$. Since $X^2 = 2.149$ and $G^2 = 2.167$, both can not reject the null hypothesis, which conclude that the proposed logistic regression model is appropriate. Note that the p-value is 0.5421341 using X^2 and 0.538514 using G^2. ‖

It should be noted that the deviance statistic in (14.13) is equivalent to $-2\log\lambda$ where λ is defined in (14.9). The log-likelihood function under the binomial model obtained in (14.5) is given by

$$\ell(\pi_1, \ldots, \pi_c) = \sum_{j=1}^{c} \Big[y_j \log \pi_j + (n_j - y_j) \log(1 - \pi_j) \Big] + C.$$

The above log-likelihood function is maximized with $\hat{\pi}_j^* = y_j/n_j$ without any restriction in response function. On the other hand, under the logistic regression model in (14.11), the above is maximized with $\hat{\pi}_j = \exp(\mathbf{x}_j'\hat{\boldsymbol{\beta}})/[1 + \exp(\mathbf{x}_j'\hat{\boldsymbol{\beta}})]$ which is from (14.12). Thus, we have

$$\begin{aligned} -2\log\lambda &= -2\big\{\ell(\hat{\pi}_1, \ldots, \hat{\pi}_c) - \ell(\hat{\pi}_1^*, \ldots, \hat{\pi}_c^*)\big\} \\ &= 2\big\{\ell(\hat{\pi}_1^*, \ldots, \hat{\pi}_c^*) - \ell(\hat{\pi}_1, \ldots, \hat{\pi}_c)\big\} \\ &= 2\sum_{j=1}^{c} \Big\{ y_j \log\Big(\frac{y_j/n_j}{\hat{\pi}_j}\Big) + (n_j - y_j) \log\Big(\frac{1 - y_j/n_j}{1 - \hat{\pi}_j}\Big) \Big\}. \end{aligned}$$

8 Poisson Regression

Poisson random variables can be used when the event is something that can be counted in whole numbers. Thus, the Poisson regression model can be stated as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where Y_i is distributed as Poisson with the mean given by

$$\mu_i = E[Y_i] = \beta_0 + \beta_1 X_i.$$

We can also extend the above simple regression part to the multiple regression so that we have

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

As discussed in the simple linear regression model with the binary response, the above model may have a flaw of being negative value of the mean μ_i. One easy way of fixing this problem is to take an exponential function

$$\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}).$$

This is also called the Poisson log-linear model in categorical data analysis since this relation is equivalent to

$$\log \mu_i = \mathbf{x}_i'\boldsymbol{\beta}. \tag{14.14}$$

The function that relates the mean response μ_i to $\mathbf{x}_i'\boldsymbol{\beta}$ is also called a *log link* in the generalized linear model. This log link is the most popular function for Poisson regression model. Other popular link functions for Poisson regression are

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta} \qquad \text{identity}$$

$$\sqrt{\mu_i} = \mathbf{x}_i'\boldsymbol{\beta} \qquad \text{square root.}$$

Note that $\mu_i = \log(\mathbf{x}_i'\boldsymbol{\beta})$, equivalent to $\exp(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$, is rarely used in practice.

Maximum Likelihood Estimation

The likelihood function for the Poisson regression model with mean μ_i is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!},$$

where y_i is a realization of Y_i and μ_i is a function of $\mathbf{x}_i'\boldsymbol{\beta}$ using a link function such as log link, identity, etc. Then the log-likelihood function with the link μ_i is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log \mu_i - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \log(y_i!). \tag{14.15}$$

Goodness of Fit

If there is no link function relating μ_i and $\mathbf{x}_i'\boldsymbol{\beta}$, then the log-likelihood function in (14.15) is maximized with $\mu_i = y_i$

$$\ell_{\max} = \sum_{i=1}^{n} y_i \log y_i - \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \log(y_i!). \tag{14.16}$$

Let $\hat{\mu}_i$ be the fitted values with the MLE $\hat{\boldsymbol{\beta}}$ using the Poisson regression model. Then we have

$$\ell(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} y_i \log \hat{\mu}_i - \sum_{i=1}^{n} \hat{\mu}_i - \sum_{i=1}^{n} \log(y_i!).$$

The deviance is given by

$$\begin{aligned} D &= 2(\ell_{\max} - \ell(\hat{\boldsymbol{\beta}})) \\ &= 2\sum_{i=1}^{n} y_i(\log y_i - \log \hat{\mu}_i) - 2\sum_{i=1}^{n}(y_i - \hat{\mu}_i) \\ &= 2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - 2\sum_{i=1}^{n}(y_i - \hat{\mu}_i), \end{aligned}$$

which is approximately distributed as a chi-square with degrees of freedom $n - p$. For most models, we have $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\mu}_i$. Then the deviance is very close to G^2. Denoting $O_i = y_i$ and $E_i = \hat{\mu}_i$, we can write

$$D \approx 2\sum_{i=1}^{n} O_i \log\left(\frac{O_i}{E_i}\right).$$

Example 10. Textbook Example. See Table 14.14 on Page 622.

The response variable Y_i is the number of customers who visited a store. The predictors are obtained from in-store surveys of customers. These are $X_1, \ldots, X_5$. We want to analyze the data set with the Poisson regression model with the log link.

R

Read Data

```
> url = "https://raw.githubusercontent.com/AppliedStat/LM/master/CH14TA14.txt"
> Data = read.table(url)
> y  = Data[,1]
> x1 = Data[,2]
> x2 = Data[,3]
> x3 = Data[,4]
> x4 = Data[,5]
> x5 = Data[,6]
```

GLM

```
> GLM = glm(y~x1+x2+x3+x4+x5, family=poisson("log"))
> summary(GLM)
Call:
glm(formula = y ~ x1 + x2 + x3 + x4 + x5, family = poisson("log"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.93195  -0.58868  -0.00009   0.59269   2.23441

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.942e+00  2.072e-01  14.198  < 2e-16 ***
x1           6.058e-04  1.421e-04   4.262 2.02e-05 ***
x2          -1.169e-05  2.112e-06  -5.534 3.13e-08 ***
x3          -3.726e-03  1.782e-03  -2.091   0.0365 *
x4           1.684e-01  2.577e-02   6.534 6.39e-11 ***
x5          -1.288e-01  1.620e-02  -7.948 1.89e-15 ***
---
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 422.22  on 109  degrees of freedom
Residual deviance: 114.99  on 104  degrees of freedom
AIC: 571.02

> # Goodness of fit
> O = y
> E = fitted(GLM)
> G2 = 2*sum(O*log(O/E), na.rm=TRUE)
> G2
[1] 114.9854

> qchisq(1-0.05, 110-6)  # n=110 and p=6
[1] 128.8039
```

9 Generalized Linear Models

The models we studied in this chapter can belong to a family of models called *generalized linear models*.

In some regression problems, traditional linear regression models may not work properly. Transforming response or predictor variables can improve regression models, but it is often inadequate. Generalized linear models are an extension to traditional linear regression models which can handle more complex problems.

A generalized linear model has three components: (i) the *random component* which identifies the response variable Y, (ii) the *system component* which specifies the predictor variables, and (iii) the *link* function which describes the relationship between the random and system components.

- Random Component:

  The response variable, $Y_1, Y_2, \ldots, Y_n$, are independent (not necessarily identical) random variables, each with a distribution from a specified *exponential family*. Exponential families include (i) continuous (normal, lognormal, gamma, inverse Gaussian, beta) and (ii) discrete (binomial with size known, Poisson, geometric, negative binomial with r known). Note that uniform, F, Cauchy, hypergeometric, logistic, and Weibull are not exponential families.

  This family restriction is not so serious because the exponential families include important and useful distributions.

- System Component:

  It is a model which specifies function of the predictor variables, linear in the parameters. We denote this component by

  $$\eta_i = \mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}.$$

- Link Function:

  This function links between the random and system components. Denote the expectation of Y by $\mu_i = E(Y_i)$. The link function function $g(\mu)$ has

  $$\eta_i = g(\mu_i).$$

  where $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$.

Thus, the identity link $\eta_i = g(\mu_i)$ with a normal distribution provides a traditional linear regression model.

Table 14.1: Link functions and their inverses.

Link name	$\eta = g(\mu)$	$\mu = g^{-1}(\eta)$
identity	μ	η
inverse	$1/\mu$	$1/\eta$
log	$\log \mu$	$\exp(\eta)$
logit	$\log\big(\mu/(1-\mu)\big)$	$\exp(\eta)/\big(1+\exp(\eta)\big)$
probit	$\Phi^{-1}(\mu)$	$\Phi(\eta)$
complementary log-log	$\log\big(-\log(1-\mu)\big)$	$1-\exp\big(-\exp(\eta)\big)$
square root	$\sqrt{\mu}$	η^2
inverse square	$1/\mu^2$	$1/\sqrt{\eta}$

Table 14.2: Families and link functions for `glm()` function in R. The default link is denoted by D.

Family name	`identity`	`inverse`	`log`	`logit`	`probit`	`cloglog`	`sqrt`	`1/mu^2`
`gaussian`	D	√	√					
`binomial`			√	D	√	√		
`poisson`	√		D				√	
`Gamma`	√	D	√					
`inverse.gaussian`	√	√	√					D

10 Miscellaneous

We have studied two important statistics for testing goodness of fit. These are deviance and Pearson goodness-of-fit statistics defined by

$$G^2 = 2\sum_{j=1}^{k} O_j \log\left(\frac{O_j}{E_j}\right)$$

and

$$X^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j},$$

respectively. In the previous examples, we noticed that these two statistics are quite close and both are always positive. For the Pearson goodness-of-test statistic, it is easily seen that it is always positive, but it is not trivial to show the positivity for the deviance statistic since $\log(O_j/E_j) < 0$ with $O_j < E_j$.

First, we look at the positivity of the deviance statistic. Let Y be the discrete random variable from a multinomial distribution with the pmf $f(\,)$ and y is the realization of Y. Assuming E_j is the expected value of the jth cell in the multinomial distribution, we can rewrite $E_y = nf(y)$. Let $g(x)$ be the empirical pmf. Then we have $O_y = ng(y)$. Let $\delta(y) = [g(y) - f(y)]/f(y) = g(y)/f(y) - 1$, which is known as Pearson residual. Note that $\delta(y) \geq -1$ for any y.

The deviance is expressed as

$$\begin{aligned} G^2 &= 2n \sum_y g(y) \log\Big(\frac{g(y)}{f(y)}\Big) \\ &= 2n \sum_y (\delta(y)+1) \log(\delta(y)+1) f(y) \\ &= 2nE\Big[(\delta(Y)+1)\log(\delta(Y)+1)\Big]. \end{aligned}$$

It is immediate from $(\delta+1)\log(\delta+1) \geq \delta$ that

$$E\Big[(\delta(Y)+1)\log(\delta(Y)+1)\Big] \geq E\big[\delta(Y)\big] = 0.$$

Thus, we have$= G^2 \geq 0$.

Next, we will show that G^2 is approximated by X^2. Using $E_y = nf(y)$ and $O_y = ng(y)$,

we can rewrite the Pearson goodness-of-fit statistic as

$$\begin{aligned} X^2 &= \sum_y \frac{\{ng(y) - nf(y)\}^2}{nf(y)} \\ &= n\sum_y \Big[\frac{g(y) - f(y)}{f(y)}\Big]^2 f(y) \\ &= nE\big[\delta(Y)^2\big]. \end{aligned}$$

Using the Taylor expansion at $\delta = 0$, we have

$$(\delta + 1)\log(\delta + 1) = 0 + 1 \cdot \delta + \frac{1}{2} \cdot \delta^2 + o(\delta^2) \approx \delta + \frac{1}{2}\delta^2.$$

Using $E\big[\delta(Y)\big] = 0$, we have

$$G^2 = 2nE\Big[(\delta(Y) + 1)\log(\delta(Y) + 1)\Big] \approx 2nE\Big[\delta(Y) + \frac{1}{2}\delta(Y)^2\Big] = nE\Big[\delta(Y)^2\Big] = X^2.$$

The squared Hellinger distance is also often used for goodness of fit and it is defined by

$$H^2 = 2\sum_{j=1}^{k}\Big(\sqrt{O_j} - \sqrt{E_j}\Big)^2.$$

We will take a look at the approximation of the squared Hellinger distance. Again, using $E_y = nf(y)$ and $O_y = ng(y)$, we have

$$\begin{aligned} H^2 &= 2\sum_y \Big[\sqrt{ng(y)} - \sqrt{nf(y)}\Big]^2 \\ &= 2n\sum_y \Big[g(y) + f(y) - 2\sqrt{f(y)g(y)}\Big] \\ &= 2n\sum_y \left[\frac{g(y)}{f(y)} + 1 - 2\sqrt{\frac{g(y)}{f(y)}}\right] f(y) \\ &= 2n\sum_y (\delta(y) + 2 - 2\sqrt{\delta(y) + 1}) f(y) \\ &= n\sum_y (4 - 4\sqrt{\delta(y) + 1}) f(y) \\ &= nE\Big[4 - 4\sqrt{\delta(Y) + 1}\Big]. \end{aligned}$$

Using the Taylor expansion at $\delta = 0$, we have

$$4 - 4\sqrt{\delta + 1} = 0 + (-2)\delta + \delta^2 + o(\delta^2).$$

Using this, we can easily see that

$$H^2 = nE\Big[4 - 4\sqrt{\delta(Y) + 1}\Big] \approx nE\Big[\delta(Y)^2\Big] = X^2.$$

References

Berkson, J. (1980). Minimum chi-square, not maximum likelihood! *Annals of Statistics*, 8:457–487.

Chernoff, H. and Lehmann, E. L. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Annals of Mathematical Statistics*, 25:579–586.