Handout 10

STA 138

Logistic Regression.

Heart attack data (Example 3 in Handout 9) [Ungrouped, Response 0-1 valued]

For the i^{th} subject, let Y_i=1 if there is a second heart attack within a year of the first attack and 0 otherwise. We look at an independent variable X, which is anxiety. At this moment we will ignore the other variable if the patient has gone through treatment for anger management. Thus we have observations, $(Y_i, X_i), i = 1, \ldots, n$, and if we write $\pi_i = P(Y_i = 1)$, then let π' be the vector of logit transformations of π_i, i.e., $\pi_i' = \log[\pi_i/(1-\pi_i)]$, then we we are modeling $\pi_i' = \beta_0 + \beta_1 X_i, i = 1, ..., n$, or $\boldsymbol{\pi' = X\beta}$ in the matrix notation, where X is a $n \times 2$ with first column consisting of 1's and the second column consists of the values $X_1, \ldots, X_n$.

We now fit a logistic model for this data using the R command "attack=glm(Y~X,family='binomial')". The command "summary(attack)" will give us the following output

Call:

glm(formula = Y ~X, family = binomial)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.62461	-0.83983	-0.01232	0.64540	2.10801

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.0925	3.1709	-2.237	0.0253 *
X	0.1246	0.0553	2.254	0.0242 *

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 27.726 on 19 degrees of freedom

Residual deviance: 19.601 on 18 degrees of freedom

AIC: 23.601

Number of Fisher Scoring iterations: 4

Inference. Unlike in the linear regression or ANOVA cases, all the inference in logistic regression model are approximate since it is not possible to find the exact distribution of the parameter estimates $\hat{\beta}$'s. For instance, in order to construct confidence intervals for β_1 or carrying out tests for it, we will use the normal table. Mathematically, the normal approximation is justified if the sample size n is large.

An approximate 95% confidence interval for β_1 is $\hat{\beta}_1 \pm 1.96 s(\hat{\beta}_1)$, i.e., $0.1246 \pm (1.96)(0.0553)$ i.e., 0.1246 ± 0.1084, i.e., $(0.016, 0.233)$. Similarly we can carry out a test for the hypothesis $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$, by using the z-statistic $z^* = \hat{\beta}_1/s(\hat{\beta}_1) = 2.254$.The p-value is 0.0242 as given above.

Estimating the probability of a second heart attack at $X = 71$. This is simply done by plugging in the value of of $X = 71$ in the fitted logistic regression. However, a package like R will also do this.

Concept of Deviance

Whenever you use R or any other package, you will get two quantities as part of the output: null deviance and residual deviance.

In the regression analysis we get SSE and SSTO (total sum of squares) as part of the computer output. The corresponding analgues here are: the residual deviance (an analogue of SSE), and null deviance (an analogue of SSTO).

In the Heart Attack data, we are modeling π_i as a linear function of X_i, ie, $\pi_i' = \beta_0 + \beta_1 X_i$, where π_i' is the logit of π_i. Whenever we are fitting this model, the computer looks at the following three models.

Model 1: $\pi_i' = \beta_0 + \beta_1 X_i$.

The Maximum Likelihood (ML) estimates $\hat{\beta}_0$ and $\hat{\beta}_1$of β_0 and β_1 are obtained by maximizing the lilelihood function L. Expression for the likehood function is given in the Appendix. Denote the estimate π_i by $\hat{\pi}_i^{(1)} = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)/[1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)]$. We will denote the value of the likehood function when we plug in $\hat{\pi}_i^{(1)}$ in the expression of the likelihood by L_1. Note that there are no explicit expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$, the computer obtains the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ by solving the likelihood equatuion.

Model 0 (Null model): $\pi_i' = \beta_0$, ie, π_i does not depend on X_i.

The ML estimate of β_0 is simple, and it is $\hat{\beta}_0 = \log(p/(1-p))$, where $p = \sum Y_i/n$, the sample proportion of sencond heart attacks (within a year of the first attack). Denote the estimate value of π_i by $\hat{\pi}_i^{(0)} = \exp(\hat{\beta}_0)/[1 + \exp(\hat{\beta}_0)] = p$, Let L_0 be the value of the liklehood function when $\hat{\pi}_i^{(0)}$ is plugged in the expression of the likelihood L.

Saturated Model:

Note that we have n observations, and $Y_i \sim binomial(1, \pi_i)$. The **saturated model** is defined to be the model whereby all the the π_i's are allowed to be arbitrary values between 0 and 1. Thus in the satuared model, there are n parameters to be estimated: $\pi_1, \ldots, \pi_n$, and the ML estimates are $Y_1, \ldots, Y_n$. Denote these estimates as $\hat{\pi}_i^{(S)} = Y_i$. When these estimates of $\pi_1, \ldots, \pi_n$ are plugged in the likelihhod function, we call it L_S.

Deviance associated with the model $\pi_i' = \beta_0 + \beta_1 X_i$ is defined to be

$$\begin{aligned} G^2 &= \text{Residual Deviance} = -2[\log(L_1) - \log(L_S)] \\ &= -2\sum_{i=1}^{n}[Y_i \log(\hat{\pi}_i^{(1)}) + (1 - Y_i)\log(1 - \hat{\pi}_i^{(1)})] \\ &= 19.601. \end{aligned}$$

The degrees of freedom associated with deviance G^2 is

$$\begin{aligned} &\#\text{ number of parameters estimated under the saturated model} \\ &\text{- }\#\text{ of parameters estimated under model 1} \\ = \;& n - 2 = 18. \end{aligned}$$

Let us define the null deviance. Consider the logit model $\pi_i' = \beta_0$, ie, π_i does not depend on the independent variable(s). Let $\hat{\beta}_0$ be the ML estimate of β_0. Incidentally, the estimate of β_0 is $\hat{\beta}_0 = \log(p/(1-p))$, where

$p = \sum Y_i/n$, which the sample proprtion of second heart attacks (within a year of the first one). Note that $\pi_i^{(0)} = p$. Let L_N be the value of the likelihood when we plug in the likelihood function Null devinace is defined to be

$$\begin{aligned}
\text{Null Deviance} &= -2[\log(L_0) - \log(L_S)] \\
&= -2\sum_{i=1}^{n}[Y_i \log(\hat{\pi}_i^{(0)}) + (1-Y_i)\log(1-\hat{\pi}_i^{(0)})] \\
&= -2\sum_{i=1}^{n}[Y_i \log(p) + (1-Y_i)\log(1-p)] \\
&= -2n[p\log(p) + (1-p)\log(1-p)] \\
&= 27.726.
\end{aligned}$$

The degrees of freedom associated with null deviance is

$$\begin{aligned}
&\#\text{ number of parameters estimated under the saturated model} \\
&\text{- }\#\text{ of parameters estimated under model 0} \\
= \quad & n - 1 = 19.
\end{aligned}$$

Residuals: The plot of the fitted $\hat{\pi}_i^{(1)}$ against X_i is plotted. When it comes to residuals, there are multiple definitions of residuals for the logistic regression. Two most common ones are:

Pearson Residual: $e_i = (Y_i - \hat{\pi}_i^{(1)})/\sqrt{\hat{\pi}_i^{(1)}(1-\hat{\pi}_i^{(1)})}$

Deviance Residual: $dev_i = sign(Y_i - \hat{\pi}_i^{(1)})\sqrt{-2[Y_i \log \hat{\pi}_i^{(1)} + (1-Y_i)\log(1-\hat{\pi}_i^{(1)})]}$.

[R commands: resid(attack), resid(attack,type="pear"), will yield deviance residuals and Pearson residuals, respectively. Recall that 'attack' is the R object when we give the command "attack=glm(Y~X,family='binomial')"].

Here are two measure that are somewhat similar to SSE in the regression models discussed in STA 106 and STA 108.

(a) Pearson: $X^2 = \sum e_i^2$, and (ii) Deviance: $G^2 = \sum dev_i^2$.

Even though, neither X^2 nor G^2 has exactly a χ^2 distribution with $n-2$ df, we may still use the χ^2 table to see if the calculated values of X^2 or G^2 are too high. High values indicate inadequacy of the model. For instance, area to the right of 28.87 under the χ^2_{18} curve is 0.05. The value of $G^2 = 19.601$ which is clearly smaller than 28.87. Thus we may conclude that there does not seem to be a departure from the linear model fitted here.

We may plot the residuals against the fitted $\hat{\pi}_i$'s, and usually there are parallel curves. It is not very clear how much information these plots convey. To extract more information, we may plot the loess of the residuals, if the loess seem to be quite close to the horizontal line at 0, it would indicate the model may be adequate in estimating π, the probability of a second hear attack. Another plot known as the half normal plot allows us to check adequacy of the model and identify outliers.[R command for half-normal plot is "halfnorm" (package "faraway"). Half normal-plot here does not seems to show any serious outliers, nor any inadequacy of the logistic linear model.

Beetle data (Example 4 in Handout 9) [Grouped Data]

Note that for this data, we have the number exposed and killed at $c = 8$ different doses. For instance, $n_2 = 60$ were exposed to dose $X_2 = 53$ out of which 13 were killed. This is an example of Binomial regression which is not really much different from the previous example. Here, the total number of observations is $n = 420$. One may think at dose $X_2 = 53$, we have $n_2 = 60$ observations $Y_{2j}, i = 1, \ldots, n_2 = 60$. Each Y_{ij} is 0-1 valued, i.e., Y_{ij} is Bernoulli with probability of getting killed is π_i. So $Y_{2\cdot} = \sum_j Y_{2j}$ has a Binomial(n_2, π_2) distribution. In general we have c distinct doses $X_1, ..., X_c$; n_i beetles were exposed to dose X_i and $Y_{i\cdot}$ were killed out of n_i flies. So $Y_{i\cdot}$ has a Binomial(n_i, π_i) distribution. We now want to model π_i', the logit of π_i, as a linear function of X_i. Thus we model $\pi_i' = \beta_0 + \beta_1 X_i, i = 1, \ldots, c = 8$. If this is not adequate we may also try a polynomial model such as quadratic.

We will describe the linear modeling here. The R command is "beetle=glm(cbind(killed,not killed)~X,family='binomial')" Here is a summary of the output.

Call:

glm(formula = cbind(kill, surv) ~dose, family = binomial)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.2746	-0.4668	0.7688	0.9544	1.2990

Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-14.82300	1.28959	-11.49	<2e-16
dose	0.24942	0.02139	11.66	<2e-16

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.2024 on 7 degrees of freedom

Residual deviance: 7.3849 on 6 degrees of freedom

AIC: 37.583

Number of Fisher Scoring iterations: 4

This output clearly shows that the dose is a significant variable, i.e., if we test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$, then the z-statistic is $z = \hat{\beta}_1 / s(\hat{\beta}_1) = 11.66$ and the p-value≈ 0. Thus we can reject H_0 quite safely.

Deviance

As in the Heart Attack data, we have three models to consider.

Note that we have $Y_{i\cdot} \sim binomial(n_i, \pi_i), i = 1, \ldots, c = 8$.

Model 1: $\pi_i' = \beta_0 + \beta_1 X_i$.

The Maximum Likelihood (ML) estimates $\hat{\beta}_0$ and $\hat{\beta}_1$of β_0 and β_1 are obtained by maximizing the lilelihood function L. Expression for the likehood function is given in the Appendix. Denote the estimate π_i by $\hat{\pi}_i^{(1)} = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)/[1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)]$. We will denote the value of the likehood function when we plug in $\hat{\pi}_i^{(1)}$ in the expression of the likelihood by L_1. Note that there are no explicit expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$, the computer obtains the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ by solving the likelihood equatuion.

Model 0: $\pi_i' = \beta_0$, ie, π_i does not depend on X_i.

The ML estimate of β_0 is simple, and it is $\hat{\beta}_0 = \log(p/(1-p))$, where $p = \sum_i Y_{i\cdot} / \sum n_i$, the sample proportion of second heart attacks (within a year of the first attack). Denote the estimate value of π_i by

$\hat{\pi}_i^{(0)} = \exp(\hat{\beta}_0)/[1+\exp(\hat{\beta}_0)] = p$, Let L_0 be the value of the liklehood function when $\hat{\pi}_i^{(0)}$ is plugged in the expression of the likelihood L.

Saturated Model:

Note that the **saturated model** is defined to be the model whereby all the the π_i's are allowed to be arbitrary values between 0 and 1. Thus in the satuared model, there are $c = 8$ parameters to be estimated: $\pi_1, \ldots, \pi_c$, and the ML estimates are $p_1, \ldots, p_c$, where $p_i = Y_{i\cdot}/n_i$. Estimated value of π_i under the saturated model is $\hat{\pi}_i^{(S)} = p_i$ When these estimates of $\pi_1, \ldots, \pi_c$ are plugged in the likelihhod function, we call it L_S.

The deviance for model 1 is called the Residual Deviance,

$$\begin{aligned}
G^2 &= \text{Residual Deviance} = - - 2[\log(L_1) - \log(L_S)] \\
&= -2\sum_{i=1}^{c}\left[Y_{i\cdot}\log\left(\frac{\hat{\pi}_i^{(1)}}{\hat{\pi}_i^{(S)}}\right) + (n_i - Y_{i\cdot})\log\left(\frac{1-\hat{\pi}_i^{(1)}}{1-\hat{\pi}_i^{(S)}}\right)\right] \\
&= -2\sum_{i=1}^{c}\left[Y_{i\cdot}\log\left(\frac{n_i\hat{\pi}_i^{(1)}}{Y_{i\cdot}}\right) + (n_i - Y_{i\cdot})\log\left(\frac{n_i(1-\hat{\pi}_i^{(1)})}{n_i - Y_{i\cdot}}\right)\right] \\
&= 2\sum_{i=1}^{c}\left[Y_{i\cdot}\log\left(\frac{Y_{i\cdot}}{n_i\hat{\pi}_i^{(1)}}\right) + (n_i - Y_{i\cdot})\log\left(\frac{n_i - Y_{i\cdot})}{n_i(1-\hat{\pi}_i^{(1)})}\right)\right] \\
&= 7.3849.
\end{aligned}$$

The degrees of freedom associated with the residual deviance is

$$\begin{aligned}
&\#\text{ number of parameters estimated under the saturated model} \\
&\text{- \# of parameters estimated under model 1} \\
= \quad & c - 2 = 6.
\end{aligned}$$

The null deviance is

$$\begin{aligned}
&- 2[\log(L_0) - \log(L_S)] \\
&= -2\sum_{i=1}^{c}\left[Y_{i\cdot}\log\left(\frac{\hat{\pi}_i^{(0)}}{\hat{\pi}_i^{(S)}}\right) + (n_i - Y_{i\cdot})\log\left(\frac{1-\hat{\pi}_i^{(0)}}{1-p_i}\right)\right] \\
&= -2\sum_{i=1}^{c}\left[Y_{i\cdot}\log\left(\frac{n_i\hat{\pi}_i^{(0)}}{Y_{i\cdot}}\right) + (n_i - Y_{i\cdot})\log\left(\frac{n_i(1-\hat{\pi}_i^{(0)})}{n_i - Y_{i\cdot}}\right)\right] \\
&= 2\sum_{i=1}^{c}\left[Y_{i\cdot}\log\left(\frac{Y_{i\cdot}}{n_i\hat{\pi}_i^{(0)}}\right) + (n_i - Y_{i\cdot})\log\left(\frac{n_i - Y_{i\cdot})}{n_i(1-\hat{\pi}_i^{(0)})}\right)\right] \\
&= .284.2024
\end{aligned}$$

The degrees of freedom associated with the null deviance is

$$\begin{aligned}
&\#\text{ number of parameters estimated under the saturated model} \\
&\text{- \# of parameters estimated under model 0} \\
= \quad & c - 1 = 7.
\end{aligned}$$

Goodness-of-fit tests.

In Example 4, we may wish to investigate if fitting a logistic linear model is appropriate, i.e., test $H_0 : \pi_i = \beta_0 + \beta_1 X_i$ vs. $H_1 : \{\pi_i\}$ do not lie on a straight line.

There are two popular tests for this:

(a) Pearson Chi-square test,

(b) Likelihood ratio test (Deviance goodness-of-fit test).[Note at the end of this handout.]

Let us discuss (b) first. This test statistic G^2 is given by R. Here $G^2 = 7.3849$ with df= 6. The p-value is 0.287 and hence we cannot reject H_0. Conclusion: a linear logistic fit seems to be reasonable.

(a) Pearson Chi-square goodness-of-fit test.

Let $\hat{\pi}_i$ be the estimated value of π_i under the reduced model. Then the expected number of killed at X_i is $E_{i1} = n_i \hat{\pi}_i^{(1)}$ and expected number of not-killed is $E_{i2} = n_i(1 - \hat{\pi}_i^{(1)})$. Denote O_{i1} is the observed number of killed (which is Y_i. in our notation) and the observed number of not-killed is $O_{i2} = n_i - Y_i$.. The Pearson statistics is

$$X^2 = \sum_{i=1}^{c} \sum_{l=1}^{2} \frac{(O_{il} - E_{il})^2}{E_{il}}.$$

Under H_0, $X^2 \sim \chi^2_{c-p}$, under the assumption that all the expected frequencies $\{E_{il}\}$ are large. A rule of thumb for this "largeness" is to have all (or almost all) the expected frequencies 5 or larger. Otherwise, frequencies for neighboring X_i are merged together. In our example, we have the following table (the expected frequencies in the brackets):

Dose	Exposed	Killed	Not-killed
X_i	n_i	Count	Count
49.1	59	6 (4.171)	53 (54.829)
53.0	60	13 (10.044)	47 (49.958)
56.9	62	18 (21.526)	44 (40.474)
60.8	56	28 (32.732)	28 (23.268)
64.8	63	52 (49.915)	11 (13.085)
68.7	59	53 (53.684)	6 (5.316)
72.6	62	61 (59.762)	1 (2.238)
76.5	60	60 (59.160)	0 (0.840)

Note that the expected number of not-killed are rather small for dose X_7 and X_8. It may be worthwhile to merge the doses X_6, X_7 and X_8 together. Here is the table after merging these.

Dose	Exposed	Killed	Not-killed
X_i	n_i	Count	Count
49.1	59	6 (4.171)	53 (54.829)
53.0	60	13 (10.044)	47 (49.956)
56.9	62	18 (21.526)	44 (40.474)
60.8	56	28 (32.732)	28 (23.268)
64.8	63	52 (49.915)	11 (13.085)
$\geq$68.7	181	174 (172.606)	7 (8.438)

Now note that we have $c-2 = 6$ dose categories, so the Pearson statistic X^2 will have $df = c-2-p = 4$.

The value of the statistics is $X^2 = 5.101$ and the p-value is 0.277, and we cannot reject H_0. Thus a linear logistic model may reasonable here.

Remark: We should note in passing the requirement that all the expected frequencies (i.e., $E_{i1} = n_i\hat{\pi}_i$ and $E_{i2} = n_i(1-\hat{\pi}_i)$) is also valid for the Deviance goodness-of-fit test. Even though, we have not done it here, same merging of categories are also needed for this example.

Residuals:

The Pearson and deviance residuals are

$$\begin{aligned} e_i &= \frac{Y_{i\cdot} - n_i\hat{\pi}_i^{(1)}}{\sqrt{n_i\hat{\pi}_i^{(1)}(1-\hat{\pi}_i^{(1)})}}, \\ dev_i &= sign(Y_{i\cdot} - n_i\hat{\pi}_i^{(1)})\sqrt{-2\left[Y_{i\cdot}\log\left(\frac{n_i\hat{\pi}_i^{(1)}}{Y_{i\cdot}}\right) + (n_i - Y_{i\cdot})\log\left(\frac{n_i(1-\hat{\pi}_i^{(1)})}{n_i - Y_{i\cdot}}\right)\right]} \\ &= sign(Y_{i\cdot} - n_i\hat{\pi}_i^{(1)})\sqrt{2\left[Y_{i\cdot}\log\left(\frac{Y_{i\cdot}}{n_i\hat{\pi}_i^{(1)}}\right) + (n_i - Y_{i\cdot})\log\left(\frac{n_i - Y_{i\cdot}}{n_i(1-\hat{\pi}_i^{(1)})}\right)\right]}. \end{aligned}$$

Note that

$$X^2 = \sum_i e_i^2, \text{ and } G^2 = \sum dev_i^2.$$

The following table given the Pearson as well as deviance residuals

Dose	Exposed	Killed	Pearson	Deviance	Standardized Pearson
X_i	n_i	Count	Residual	Residual	Residual
49.1	59	6 (4.171)	0.9305	0.8758	1.0313
53.0	60	13 (10.044)	1.0216	0.9868	1.2055
56.9	62	18 (21.526)	-0.9411	-0.9543	-1.1438
60.8	56	28 (32.732)	-1.2839	-1.2739	-1.4604
64.8	63	52 (49.915)	0.6467	0.6613	0.7829
68.7	59	53 (53.684)	-0.3097	-0.3053	-0.3515
72.6	62	61 (59.762)	0.8431	0.9436	1.0514
76.5	60	60 (59.160)	0.9218	NaN	1.3826

[R Command for standardized Pearson residuals is rstandard(beetle), where beetle is the R glm object. For studentized deviance residuals use rstudent(beetle).]

Note that if the the null is true (ie, $\pi' = \beta_0 + \beta_1 X$), then each Pearson residual is approximately normal with mean zeo, but the variance is smaller than 1. The same is also true for deviance residuals. For this reason, it is desirable to use the standarddized residuals.

If the logistic linear model is appropriate, then each standardized residual has an approximate $N(0,1)$ distribution. Note that all the standardized Prearosn residuals are between ± 1.5 for the Beetle data, suggesting that modeling logit of π as a linear function of X is quite reasonable.

Appendix.

Case I. Observed Y's are 0-1 valued
Estimation of parameters (Maximum likelihood).

For logistic regression, one may use the maximum likelihood (ML) method. The likelihood is

$$L = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}, \text{ and}$$
$$\log L = \sum [Y_i \log \pi_i + (1-Y_i)\log(1-\pi_i)].$$

Estimates for π_i under Model 0 and the saturated model have explicit expressions, but not under Model 1. Under model 0, $\pi_1 = \cdots = \pi_n$, and the estimate of π_i is $\hat{\pi}_i^{(0)} = p$, where $p = \sum Y_i/n$. This also implies that the estimate of β_0 under model 0 is $\log(p/(1-p))$.

Under the saturated model, estimate estimate of π_i is Y_i and $\hat{\pi}_i^{(S)} = Y_i$. Before we discusss maximimum likelihood estimation under model 1, let us write down the expressions of the log-likelihood of the various models.

Expressions for $\log(L_1), \log(L_0)$ **and** $\log(L_S)$ are

$$\begin{aligned}
\log(L_1) &= \sum_{i=1}^{n} [Y_i \log(\hat{\pi}_i^{(1)}) + (1-Y_i)\log(1-\hat{\pi}_i^{(1)})], \\
\log(L_0) &= \sum_{i=1}^{n} [Y_i \log(\hat{\pi}_i^{(0)}) + (1-Y_i)\log(1-\hat{\pi}_i^{(0)})] \\
&= \sum Y_i \log(p) + \left(n - \sum Y_i\right)\log(1-p), \text{ [with } p = \sum Y_i/n] \\
&= np\log(p) + n(1-p)\log(1-p), \\
\log(L_S) &= \sum_{i=1}^{n} [Y_i \log(\hat{\pi}_i^{(S)}) + (1-Y_i)\log(1-\hat{\pi}_i^{(S)})] \\
&= \sum_{i=1}^{n} [Y_i \log(Y_i) + (1-Y_i)\log(1-Y_i)] = 0.
\end{aligned}$$

So the residual deviance and the null deviance are

$$\begin{aligned}
\text{Residual Deviance} \quad &: \quad G^2 = -2[\log(L_1) - \log(L_S)] \\
&= -2\log(L_S) \\
&= -2\sum_{i=1}^{n} [Y_i \log(\hat{\pi}_i^{(1)}) + (1-Y_i)\log(1-\hat{\pi}_i^{(1)})], \\
\text{Null Deviance} \quad &: \quad -2[\log(L_0) - \log(L_S)] \\
&= -2\log(L_0) \\
&= -2\left[\sum Y_i \log(p) + \left(n - \sum Y_i\right)\log(1-p)\right].
\end{aligned}$$

Likelihood Equations for Model 1.

If π_i involves the parameters β_0 and β_1, i.e. $\pi_i = \exp(\beta_0 + \beta_1 X_i)/[1 + \exp(\beta_0 + \beta_1 X_i)]$. Then

$$\begin{aligned}
\log L(\beta_0, \beta_1) &= \sum_{i=1}^{n} [Y_i \log \pi_i + (1 - Y_i) \log(1 - \pi_i)] \\
&= \sum_{i=1}^{n} [Y_i \log\{\pi_i/(1 - \pi_i)\} + \log(1 - \pi_i)] \\
&= \sum_{i=1}^{n} [Y_i(\beta_0 + \beta_1 X_i) + \log(1 - \pi_i)\} \\
&= \sum_{i=1}^{n} [Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \log[1 + \exp(\beta_0 + \beta_1 X_i)].
\end{aligned}$$

In order to maximize $\log L(\beta_0, \beta_1)$ one differentiates it with respect to β_0 and β_1 and equates the derivatives to zero and that leads to the equations

$$\begin{aligned}
\sum \pi_i &= \sum Y_i, \ \ \sum X_i \pi_i = \sum X_i Y_i, \text{ where} \\
\pi_i &= \exp(\beta_0 + \beta_1 X_i)/[1 + \exp(\beta_0 + \beta_1 X_i)].
\end{aligned}$$

A solution of these equations lead to estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1. No explicit forms of $\hat{\beta}_0$ and $\hat{\beta}_1$ are available. They are usually solved via iterative methods. Thus the estimated probabilities are $\hat{\pi}_i^{(1)} = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)/[1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)]$.

If we write the model $\pi' = X\beta$, where X is $n \times 2$, then the likelihood equations can be written as

$$X^T(Y - \hat{\pi}) = 0.$$

Incidentally, this form for the likelihood equation holds even when X is a $n \times p$ matrix (i.e., there are multiple predictors). How good are the estimators? Here the results are only approximate unlike in the usual linear regression case. It can be shown that

$$E(\hat{\beta}) \approx \beta, \ \sigma^2(\hat{\beta}) = Cov(\hat{\beta}) \approx (X^T W X)^{-1},$$

where W is a $n \times n$ diagonal matrix whose diagonal entries are $\pi_1(1 - \pi_1), \ldots, \pi_n(1 - \pi_n)$. Since W can be estimated by a diagonal matrix $\hat{W}$ whose diagonal entries are $\hat{\pi}_1(1 - \hat{\pi}_1), \ldots, \hat{\pi}_n(1 - \hat{\pi}_n)$, we can therefore estimate $Cov(\hat{\beta})$ by

$$s^2(\hat{\beta}) = (X^T \hat{W} X)^{-1}.$$

Unlike in the linear regression case, we do not know the exact distribution of b_j. It turns however out that for n large, the distribution of $(\hat{\beta}_j - \beta_j)/s(\hat{\beta}_j)$ is approximate $N(0, 1)$. Thus an approximate $(1 - \alpha)100$ confidence interval for β_j is $\hat{\beta}_j \pm z(1 - \alpha/2)s(\hat{\beta}_j)$, where $z(1 - \alpha/2)$ is obtained from the normal table and $s^2(\hat{\beta}_j)$ is the j^{th} diagonal entry of $s^2(\hat{\beta})$. We can similarly carry out a hypothesis test for β_j.

Case II. Grouped Data (Observed Y's are counts).

Deviance Goodness-of-fit test

Note that $Y_{i\cdot} \sim$binomial(n_i, π_i) and $Y_{i\cdot}$'s are independent. Thus the likelihood is

$$\begin{aligned} & \prod_{i=1}^{c} \binom{n_i}{Y_{i\cdot}} \pi_i^{Y_{i\cdot}} (1-\pi_i)^{n_{ij}-Y_{i\cdot}}, \\ \log(L) &= \sum_{i=1}^{c} \log\left(\binom{n_i}{Y_{i\cdot}}\right) + \sum_{i=1}^{c} \left[Y_{i\cdot}\log(\pi_i) + (n_i - Y_{i\cdot})\log(1-\pi_i)\right]. \end{aligned}$$

Under model 1, ie, $\pi_i' = \beta_0 + \beta_1 X_i$, ML estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1 are obtained by maximizing L (or maximizing $\log L$). Since $\hat{\pi}_i^{(1)} = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)/[1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)]$, we have

$$\log(L_1) = \sum_{i=1}^{c} \log\left(\binom{n_i}{Y_{i\cdot}}\right) + \sum_{i=1}^{c} \left[Y_{i\cdot}\log(\hat{\pi}_i^{(1)}) + (n_i - Y_{i\cdot})\log(1-\hat{\pi}_i^{(1)})\right].$$

Under model 0, ie, $\pi_i' = \beta_0$, The ML estimate $\hat{\beta}_0$ of β_0 is obtained by maximizing L (or maximizing $\log L$). This estimate turns to be $\hat{\beta}_0 = \log(p/(1-p))$, where $p = \sum_i Y_{i\cdot} / \sum_i n_i$. Since $\hat{\pi}_i^{(0)} = \exp(\hat{\beta}_0)/[1+\exp(\hat{\beta}_0)] = p$, we have

$$\log(L_0) = \sum_{i=1}^{c} \log\left(\binom{n_i}{Y_{i\cdot}}\right) + \sum_{i=1}^{c} \left[Y_{i\cdot}\log(\hat{\pi}_i^{(0)}) + (n_i - Y_{i\cdot})\log(1-\hat{\pi}_i^{(0)})\right].$$

Under the saturated model, i.e., when π_i 's are arbitrary, then the MLE for π_i is $p_i = Y_{i\cdot}/n_i$. Under the saturated model, estimate of π_i is $\hat{\pi}_i^{(S)} = p_i$. Thus we have

$$\begin{aligned} \log(L_S) &= \sum_{i=1}^{c} \log\left(\binom{n_i}{Y_{i\cdot}}\right) + \sum_{i=1}^{c} \left[Y_{i\cdot}\log(\hat{\pi}_i^{(S)}) + (n_i - Y_{i\cdot})\log(1-\hat{\pi}_i^{(S)})\right] \\ &= \sum_{i=1}^{c} \log\left(\binom{n_i}{Y_{i\cdot}}\right) + \sum_{i=1}^{c} \left[Y_{i\cdot}\log(p_i) + (n_i - Y_{i\cdot})\log(1-p_i)\right]. \end{aligned}$$

Thus

$$\begin{aligned} G^2 &= \text{Residual Deviance} = -2[\log(L_1) - \log(L_S)] \\ &= -2\sum_{i=1}^{c}[\{Y_{i\cdot}\log\hat{\pi}_i^{(1)} + (n_i - Y_{i\cdot})\log(1-\hat{\pi}_i^{(1)})\} - \{Y_{i\cdot}\log p_i + (n_i - Y_{i\cdot})\log(1-p_i)\}] \\ &= -2\sum_{i=1}^{c}\left[Y_{i\cdot}\log\left(\frac{\hat{\pi}_i}{pi}\right) + (n_i - Y_{i\cdot})\log\left(\frac{1-\hat{\pi}_i}{1-p_i}\right)\right] \\ &= -2\sum_{i=1}^{c}\left[Y_{i\cdot}\log\left(\frac{n_i\hat{\pi}_i}{Y_{i\cdot}}\right) + (n_i - Y_{i\cdot})\log\left(\frac{n_i(1-\hat{\pi}_i)}{n_i - Y_{i\cdot}}\right)\right] \\ &= 2\sum_{i=1}^{c}\left[Y_{i\cdot}\log\left(\frac{Y_{i\cdot}}{n_i\hat{\pi}_i}\right) + (n_i - Y_{i\cdot})\log\left(\frac{n_i - Y_{i\cdot})}{n_i(1-\hat{\pi}_i)}\right)\right]. \end{aligned}$$

Likelihood Equations for Model 1.

Note that

$$\begin{aligned}
\log(L) &= \sum_{i=1}^{c} \log\left(\binom{n_i}{Y_{i\cdot}}\right) + \sum_{i=1}^{c} \left[Y_{i\cdot}\log(\pi_i) + (n_i - Y_{i\cdot})\log(1-\pi_i)\right] \\
&= \sum_{i=1}^{c} \log\left(\binom{n_i}{Y_{i\cdot}}\right) + \sum_{i=1}^{c} [Y_{i\cdot}\log\{\pi_i/(1-\pi_i)\} + n_i\log(1-\pi_i)] \\
&= \sum_{i=1}^{c} \log\left(\binom{n_i}{Y_{i\cdot}}\right) + \sum_{i=1}^{c} \left[Y_{i\cdot}(\beta_0 + \beta_1 X_i) - n_i\log(1+\exp(\beta_0+\beta_1 X_i)\right].
\end{aligned}$$

In order to maximize $\log L$ one differentiates it with respect to β_0 and β_1 and equates the derivatives to zero and that leads to the equations

$$\begin{aligned}
\sum n_i\pi_i &= \sum Y_{i\cdot}, \quad \sum X_i n_i \pi_i = \sum X_i Y_{i\cdot}, \text{ where} \\
\pi_i &= \exp(\beta_0+\beta_1 X_i)/[1+\exp(\beta_0+\beta_1 X_i)].
\end{aligned}$$

These equation are called the likelihood equations, and iterative methods are employed to solve them in order to obtain estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1.