Chapter 20: Logistic regression for binary response variables

In 1846, the Donner and Reed families left Illinois for California by covered wagon (87 people, 20 wagons). They attempted a new and untried crossing of the region between Ft. Bridger, Wyoming and the Sacramento Valley. After numerous problems and delays in Utah, they reached the Eastern Serria Nevada in late October. They were stranded near Lake Tahoe by a series of snowstorms that left as much as 8 feet of snow by some accounts. By the time they were rescued in April of the following year, 40 members had died. Some or perhaps all of those that survived did so by resorting to cannibalism according to newspapers reporting on the Donner party.

Grayson\textsuperscript{1} collated data on the survivorship of the party members which may now be used to gain some insight into human behavior and natural selection under extreme stress. For example, some questions of interest are whether males are better able to withstand harsh conditions than females and the extent to which the odds of survival vary with age. Grayson provides data on survival, age and gender for 45 individuals.

From a modeling standpoint, survivorship (yes, no) is the binary response variable of interest, age is a quantitative explanatory variable and gender is a categorical explanatory variable. A first attempt at modeling survivorship defines the response variable $Y_i$ to be 1 if the $i$th individual survived and 0 if they died. The Figure to the right reveals that the linear regression model is not appropriate— it’s not clear how a fitted survivorship value besides 0 and 1 should be interpreted, particularly if the value is less than 0 or greater than 1.

More formally, survivorship is a realization of a binary random variable $Y$ with the following Bernoulli distribution

$$Y = \begin{cases} 
1, & \text{if the outcome is survival,} \\
0, & \text{if the outcome is death.} 
\end{cases}$$

The probability distribution of a Bernoulli random variable $Y$ can be written as

$$P(Y = y) = \pi^y (1 - \pi)^{1-y}, \text{ for } y \in \{0, 1\},$$

where $0 < \pi < 1$ is the probability of survival. The probability distribution also may be expressed as

$$P(Y = y) = \begin{cases} \pi, & \text{if } y = 1 \\ 1 - \pi, & \text{if } y = 0. \end{cases}$$

The expected value and variance of $Y$ are

$$E(Y) = \mu = \pi$$

and

$$\text{var}(Y) = \pi(1 - \pi).$$

Recall that the linear regression model is of the mean $\mu(Y|x) = E(Y|x)$ and that the model of $Y$ is $Y = E(Y|x) + \varepsilon$. The distribution of $\varepsilon$ and hence $Y$, is assumed to be normal. $E(Y|x)$ may be any real number if $Y$ is normal; in contrast, $0 < E(Y|x) = \pi < 1$ if $Y$ is Bernoulli.

A regression model of $E(Y|x)$ for a Bernoulli random variable should not produce fitted values or predictions that are beyond the range of $Y$, that is outside the unit interval $[0, 1]$. As the figure above shows, a linear regression model almost always will produce fitted values that are inconsistent with the Bernoulli distributional model.

In addition, if formal inferences are desired, then the homogeneity of variance condition will not be satisfied unless all observations are realizations from a Bernoulli distribution with a common, single parameter $\pi$—hardly an interesting situation. An alternative to linear regression is needed.

The solution, in part, is to model the logit transformation of $\pi$ as a linear function of the explanatory variables as there are no upper or lower bounds on the transformed $\pi$.

In logistic regression, $\text{logit}(\pi)$ is modeled as a linear function of $x_1, \ldots, x_{p-1}$.

The figure below and left shows the graph of the function in red. The function is unbounded since as $\pi \to 0, \text{logit}(\pi) \to -\infty$ and as $\pi \to 1, \text{logit}(\pi) \to \infty$.

The logit function of $\pi$ is

$$\text{logit}(\pi) = \log \left( \frac{\pi}{1 - \pi} \right).$$

The logistic regression model is

$$\text{logit}(\pi|x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}.$$
The figures below show the transformation $\pi \rightarrow \logit(\pi)$ and the inverse transformation $\logit(\pi) \rightarrow \pi$.

The inverse of the logit can be computed

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} = \log \left( \frac{\pi}{1 - \pi} \right)$$

$$\Rightarrow e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}} = \frac{\pi}{1 - \pi}$$

$$\Rightarrow \frac{1}{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}}} = \frac{1 - \pi}{\pi} = 1 - \frac{1}{\pi}$$

$$\Rightarrow 1 + \frac{1}{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}}} = \frac{1}{\pi}$$

$$\Rightarrow 1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}} = \frac{1}{\pi}$$

$$\Rightarrow e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}} = \frac{1}{\pi}.$$ 

An estimate of $\pi$ can be computed given a set of values $x_1, \ldots, x_{p-1}$ using the parameter estimates $\hat{\beta}_0, \ldots, \hat{\beta}_{p-1}$ by computing

$$\hat{\pi} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_{p-1} x_{p-1})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_{p-1} x_{p-1})}$$

or

$$\hat{\pi} = \frac{1}{1 + \exp[-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_{p-1} x_{p-1})]}.$$ 

\[\text{Parameters are estimated by the method of maximum likelihood rather than the method of least squares.}\]
The odds ratio

Logistic regression parameters are interpreted through the odds ratio, a measure of how much more (or less) likely an event, say, $A$ is to occur than an event $B$. Suppose that $A$ is the outcome that a smoker has lung cancer and that the probability of $A$ is $\pi_A = P(A)$. Suppose $B$ is the outcome that a nonsmoker has lung cancer and that $\pi_B = P(B)$. Studies on the effect of smoking on health are often oriented toward comparing $\pi_A$ to $\pi_B$.

The odds that $A$ will occur are $\pi_A/(1 - \pi_A)$ and the odds of $B$ are $\pi_B/(1 - \pi_B)$. The odds ratio is

$$\frac{\pi_A/(1 - \pi_A)}{\pi_B/(1 - \pi_B)} = \frac{\pi_A(1 - \pi_B)}{\pi_B(1 - \pi_A)},$$

and it expresses how much more likely the smoker is to contract lung cancer than a nonsmoker. To illustrate, suppose that $\pi_A = 1/10$, and $\pi_B = 1/100$, so the odds of $A$ happening are 1:9 and the odds of $B$ happening are 1:99. The odds ratio, expressing how much more likely $A$ is to happen than $B$ is

$$\frac{\pi_A}{1 - \pi_A} \times \frac{1 - \pi_B}{\pi_B} = \frac{1}{9} \times \frac{99}{1} = 11.$$

The odds of $A$ happening is 11 times greater than the odds of $B$ happening.

In contrast, suppose that $\pi_A = 3/4$, and $\pi_B = 3/5$. Now the odds of $A$ are 3:1 the odds of $B$ are 3:2, and the odds of $A$ are twice that of $B$. Consequently, the odds ratio of $A$ to $B$ is 2.

Interpretation of logistic regression coefficients

Consider the logit model of $\pi = P(Y = 1)$ as a function of a single explanatory variable $x$:

$$\log \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 x$$

and let $\pi_A$ denote the probability when $x = x_A$. According to the model, the log-odds of $A$ are

$$\log \left( \frac{\pi_A}{1 - \pi_A} \right) = \beta_0 + \beta_1 x_A$$

and the odds of $A$ are

$$\frac{\pi_A}{1 - \pi_A} = e^{\beta_0 + \beta_1 x_A}.$$

Similarly, let $\pi_B$ denote the probability when $x = x_B$. The odds of $B$ are

$$\frac{\pi_B}{1 - \pi_B} = e^{\beta_0 + \beta_1 x_B}.$$
The odds of \( Y = 1 \) happening given \( x = A \) compared to the odds of \( Y = 1 \) happening given \( x = B \) is
\[
\frac{\pi_A/(1 - \pi_A)}{\pi_B/(1 - \pi_B)} = \frac{e^{\beta_0 + \beta_1 x_A}}{e^{\beta_0 + \beta_1 x_B}} = e^{\beta_0 + \beta_1 x_A - (\beta_0 + \beta_1 x_B)} = e^{\beta_1 (x_A - x_B)}.
\]

If \( x_A \) differs from \( x_B \) by one unit, i.e., \( x_A - x_B = 1 \), then
\[
e^{\beta_1 (x_A - x_B)} = e^{\beta_1},
\]
and the odds of \( Y = 1 \) happening changes by a multiplicative factor of \( e^{\beta_1} \) given a one unit change in \( x \).

In multiple regression, a one unit change in an explanatory variable \( x_i \), given that all other explanatory variables are held fixed, is estimated to produce a \( \hat{\beta}_i \) unit change in the response variable mean \( (E(Y|x)) \).

In logistic regression, a one unit change in an explanatory variable \( x_i \), given that all other explanatory variables are held fixed, is estimated to produce a \( e^{\hat{\beta}_i} \) unit change in the odds of success (i.e., \( Y = 1 \)).

For example, suppose that \( \beta_1 = .5 \); then \( e^{\beta_1} = e^{.5} = 1.649 \). So, a one unit change in \( x \) will increase the odds of success by a factor of 1.65. Suppose instead that \( \beta_1 = -.5 \); then
\[
e^{\beta_1} = e^{-\.5} = (e^{-.5})^{-1} = 1.649^{-1} = .6065,
\]
and a one unit change in \( x \) will reduce the odds of success by 100(1 - .606)\% = 39.4\%.

The algorithm for fitting logistic regression models and the statistical theory for hypothesis testing is discussed below, but to continue with the development of logistic regression, Table 1 shows the parameter estimates and approximate tests of significance for each variable.3

Table 1: Coefficients and standard errors obtained from a logistic regression of survival on sex and age for the Donner party. Residual deviance 51.256 on 42 degrees of freedom, \( n = 45 \).

| Variable        | Coefficient | Std. Error | Wald statistic (Z) | \( P(Z > |z|) \) |
|-----------------|-------------|------------|--------------------|----------------|
| Intercept       | 1.633       | 1.110      | 1.47               | 0.14           |
| Age             | -.0782      | .0373      | -2.097             | .0359          |
| Gender (female) | 1.597       | .755       | 2.114              | .0345          |

3A better significance test is developed below.
The model shown in Table 1 is

\[
\log \left( \frac{\hat{\pi}}{1 - \hat{\pi}} \right) = 1.633 - 0.0782x_{\text{age}} + 1.597x_{\text{Female}},
\]

or equivalently,

\[
\hat{\pi} = \frac{\exp(1.633 - 0.0782x_{\text{age}} + 1.597x_{\text{Female}})}{1 + \exp(1.633 - 0.0782x_{\text{age}} + 1.597x_{\text{Female}})}.
\]

According to the fitted model, the odds of survival for females are estimated to be \(e^{1.597} = 4.94\) times the odds of survival for males of the same age. An approximate 95% confidence interval for this odds ratio is 1.123 to 21.7. Also, the odds of survival for males are estimated to be

\[e^{-1.597} = .202\]

times the odds of survival for females of the same age (20.2% of the odds of survival for a comparable female). An approximate 95% confidence interval for this odds ratio is 1/21.7 = .046 to 1/1.123 = .89.

To get a direct comparison of probabilities, suppose that age and gender are fixed at 45 and female; then \(\hat{\alpha}_{\text{female},45} = .428\). For males of age 45, \(\hat{\alpha}_{\text{male},45} = .132\). The odds of survival for the female are \(.428/(1 - .428) = .748\), and for the male, \(.132/.868 = .152\) and the odds ratio is \(.748/.152 = 4.92 \approx e^{1.597} = 4.94\). The difference is rounding error.

The figure below and left shows the fitted probabilities from the additive logistic regression model (described in Table 1) and the figure below and right shows the same model with an additional term for the interaction of age and gender. The models are quite different in appearance and the drop-in-deviance test for the significance of the interaction term yielded p-value = .048.
Estimation of logistic regression coefficients

In addition from replacing the linear regression model with a logistic regression model, the
normal distribution model of responses is replaced with a Bernoulli distribution, and con-
sequently, inferential methods are different. In contrast to ordinary multiple regression, the
parameters are estimated by maximizing the likelihood of the data (instead of minimizing
the sum of the residuals). Among a large class of estimation methods, the method of maxi-
mum likelihood is best according to several important statistical criteria (such as asymptotic
unbiasedness and minimum variance).

The likelihood of the data, in the case of discrete data, is the probability of obtaining the
sample as a function of the parameters.

For example, suppose that random sample of \( n = 20 \) independent Bernoulli observations
are collected and the sample is \((y_1, \ldots, y_{20}) = (0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1)\).
Each observation is a realization from the Bin(1, \( \pi \)) distribution though \( \pi \) is unknown. The
probability of obtaining the first realization is

\[
P(Y_1 = 0) = \pi^0(1 - \pi)^{1-0} = 1 - \pi.
\]

Because the sample is random, the observations are independent, and the likelihood

\[
L(\pi; (y_1, \ldots, y_{20})) = P(Y_1 = 0) \times P(Y_2 = 0) \times \cdots \times P(Y_{20} = 1)
= (1 - \pi) \times (1 - \pi) \times \cdots \times \pi
= \pi^4(1 - \pi)^{16}.
\]
The problem is now to determine an estimate of $\pi$, and the principle underlying maximum likelihood estimation is to choose $\pi$ so that the likelihood is as large as possible (in other words, maximized). The figure to the right shows the graph of $L(\pi; (y_1, \ldots, y_{20}))$ as a function of $\pi$ (left), and the graph of the function $\log[L(\pi; (y_1, \ldots, y_{20}))]$ as a function of $\pi$ on the right.

Both functions have a unique maximum at the value $\pi = .2$, and so these data yield the maximum likelihood estimate $\hat{\pi} = .2$. The maximum likelihood estimate $\hat{\pi} = .2$ is also the conventional estimator (the sample proportion).

Maximizing the likelihood through the choice of the parameter estimates makes the probability of obtaining the sample as large as possible. Any other value of the parameter results in a lesser probability, and so any other value is less consistent with the sample data.

A more general approach to deriving the maximum likelihood estimators differentiates the likelihood with respect to $\pi$, sets the derivative to zero and solves for $\pi$. The derivative of $L(\pi; (y_1, \ldots, y_{20}))$ is

$$\frac{dL(\pi; (y_1, \ldots, y_{20}))}{d(\pi)} = 4\pi^3(1 - \pi)^{16} - 16\pi^4(1 - \pi)^{15}.$$ 

Setting this expression equal to zero and solving for $\pi$ yields

$$4\pi^3(1 - \pi)^{16} = 16\pi^4(1 - \pi)^{15} \Rightarrow 1 - \pi = 4\pi \Rightarrow \hat{\pi} = .2.$$ 

When the likelihood is more complicated (almost always), then using logarithm of the likelihood tends to be substantially easier. In this case, $\log[L(\pi; (y_1, \ldots, y_{20}))] = 4\log(\pi) + 16\log(1 - \pi)$, and

$$\frac{d \log[L(\pi; (y_1, \ldots, y_{20}))]}{d(\pi)} = \frac{4}{\pi} - \frac{16}{1 - \pi}.$$ 

Setting the derivative equal to zero, and re-arranging yields

$$\frac{1 - \pi}{\pi} = 4.$$
Solving for \( \pi \) yields the maximum likelihood estimate \( \hat{\pi} = .2 \).

For the additive logistic regression model of survival, the probability of survival for the \( i \)th individual is a function of the parameters \( \beta_0, \beta_1 \) and \( \beta_2 \). Two preliminary calculations help simplify the log-likelihood. They are

1. \[
\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})}{1 + \exp(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})} \Rightarrow 1 - \pi_i = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})}.
\]

2. The log-likelihood for the \( i \)th observation can then be expressed as

\[
\log[P(Y_i = y_i)] = \log[\pi_i^{y_i}(1 - \pi_i)^{1-y_i}]
= \log(\pi_i^{y_i}) + \log((1 - \pi_i)^{1-y_i})
= y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)
= y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + \log(1 - \pi_i)
= y_i(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}) - \log[1 + \exp(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})].
\]

Then,

\[
\log[L(\pi; (y_1, \ldots, y_{45}))] = \sum_{i=1}^{45} \log[P(Y_i = y_i)]
= \sum_{i=1}^{45} y_i(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}) - \log[1 + \exp(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})].
\]

Numerical optimization methods are used to determine the values of \( \beta_0, \beta_1 \) and \( \beta_2 \) that maximize \( \log[L(\pi; (y_1, \ldots, y_{45}))] \). Table 3 shows \( \log[L(\pi; (y_1, \ldots, y_{45}))] \) for a few values of \( \beta_0, \beta_1, \) and \( \beta_2 \).

<table>
<thead>
<tr>
<th>( \beta_0 )</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \log[L(\pi; (y_1, \ldots, y_{45}))] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.50</td>
<td>-.50</td>
<td>1.25</td>
<td>-27.7083</td>
</tr>
<tr>
<td>1.80</td>
<td></td>
<td></td>
<td>-28.7931</td>
</tr>
<tr>
<td></td>
<td>-.80</td>
<td>1.25</td>
<td>-26.1531</td>
</tr>
<tr>
<td></td>
<td>1.80</td>
<td></td>
<td>-25.7272</td>
</tr>
<tr>
<td>1.70</td>
<td>-.50</td>
<td>1.25</td>
<td>-29.0467</td>
</tr>
<tr>
<td></td>
<td>1.80</td>
<td></td>
<td>-30.3692</td>
</tr>
<tr>
<td>1.80</td>
<td>-.80</td>
<td>1.25</td>
<td>-25.7972</td>
</tr>
<tr>
<td></td>
<td>1.80</td>
<td></td>
<td>-25.6904</td>
</tr>
<tr>
<td>1.63</td>
<td>-.078</td>
<td>1.60</td>
<td>-25.6282</td>
</tr>
</tbody>
</table>

The last line in of Table 3 shows the value of \( \log[L(\pi; (y_1, \ldots, y_{45}))] \) at the maximum likelihood estimate.
Properties of maximum likelihood estimators

Maximum likelihood estimators (estimators that maximize the likelihood) have several desirable attributes. If the model is correct and the sample size is large, then

1. Maximum likelihood estimators (MLEs) are nearly unbiased.
2. The standard errors of the MLEs are nearly unbiased.
3. MLEs are more precise than nearly all other estimators.
4. The distribution of a MLE is approximately normal.4

Least squares estimators were used in conventional multiple linear regression. The MLEs are the same as the least squares estimators in the multiple linear regression situation. Least squares estimators have several other desirable attributes not possessed by the MLEs; most importantly, the least squares estimators are exactly normal in distribution if the residuals are normal in distribution.

Tests and confidence intervals involving a single parameter

Approximate normality of the MLE’s implies that if \( \hat{\beta}_i \) is the MLE of \( \beta_i \), then

\[
\hat{\beta}_i \sim N(\beta_i, \hat{\sigma}(\hat{\beta}_i)),
\]

where \( \hat{\sigma}(\hat{\beta}_i) \) is the standard error of \( \hat{\beta}_i \). This results leads to the Wald test of

\[ H_0 : \beta_i = 0 \text{ versus } H_1 : \beta_i \neq 0. \]

The test statistic is

\[ Z = \frac{\hat{\beta}_i}{\hat{\sigma}(\hat{\beta}_i)}. \]

If \( H_0 \) is true, then \( Z \sim N(0, 1) \) and the standard normal distribution is used to approximate the p-value. Informally, \( p\text{-value} = 2P(Z \geq |z|) \) where \( z \) is the observed value of \( Z \sim N(0, 1) \).

An approximate 100(1 - \( \alpha \))% confidence interval for \( \beta_i \) is

\[ \hat{\beta}_i \pm z^*\hat{\sigma}(\hat{\beta}_i) \]

where \( z^* \) is the \( \alpha/2 \) quantile from the \( N(0, 1) \) distribution.

4A consequence of the Central Limit Theorem
The normal approximation may be poor unless the sample size is large (e.g., $n > 100p$), and so the test should not be used with the Donner Party data. For illustrative purposes, Table 1 reports $\hat{\beta}_3 = .162$ and $\hat{\sigma}(\hat{\beta}_3) = .092$. The Wald statistic approximate p-value is .086 because

$$Z = \frac{\hat{\beta}_3}{\hat{\sigma}(\hat{\beta}_3)} = \frac{.162}{.092} = 1.73 \text{ and } 2P(Z \geq 1.73) = .086$$

provided that $Z \sim N(0, 1)$. To compensate for the inaccuracies associated with small samples, R computes the p-value using the $t$-distribution with $n - p$ degrees of freedom even though this distribution is, theoretically incorrect. There’s no formal argument that the $t$-distribution tail area is better than the standard normal tail area. It is, however, more conservative and less likely to lead to an incorrect assessment of statistical significance. It is also more likely to lead to an incorrect assessment of non-significance.

A better test statistic than the Wald statistic (obtained from the drop-in-deviance test) yielded p-value = .048. The estimate $\hat{\beta}_3 = .162$ implies that the difference in odds of death resulting from a one year increase in age is greater for females than males by a factor of 1.175. An approximate 95% confidence interval for this multiplicative factor is .977 to 1.141.\(^5\)

Table 1 shows that the the odds of survival is estimated to decrease by a multiplicative factor of $e^{-0.078} = .925$ for every year of life if gender is held constant. For a female at age= 20 compared to a female at age= 30, the odds of survival are reduced by a factor of $e^{-0.078 \times 10} = .925^{10} = .458$ for the older female. Said another way, the odds of survival are $e^{10\times0.078} = 2.18$ times greater for the younger female.

The drop-in-deviance test

The drop-in-deviance test compares the fit of two nested models and can be used whenever maximum likelihood is used. The test is analogous to the extra-sums-of-squares $F$ test used in linear regression and analysis of variance. The test must be used if a factor with more than two levels is to be tested for significance. In addition, the test is preferred to the Wald statistic for testing the significance of a single quantitative variable because the drop-in-deviance p-value is more accurate than the Wald statistic p-value.

The fit of the full model (or unconstrained) containing the explanatory variable (a quantitative variable or categorical factor) of interest is compared to the reduced (or constrained) model containing all explanatory variables in the full model except for the explanatory variable of interest. If the difference in fit is sufficiently large, then it is concluded that the

\(^5\)The confidence interval is obtained from the inaccurate Wald statistic explaining why the interval contains zero.

150
explanatory variable is associated with the response variable given all other variables (in both models).

If the explanatory variable is a factor with \( k \) levels, then the unconstrained model contains an additional \( k - 1 \) indicator variables representing the levels of the factor. Suppose that \( k - 1 \) variables are used to account for the explanatory variable\(^6\), and the parameters associated \( g = k - 1 \) variables are \( \beta_1, \beta_2, \ldots, \beta_g \). The null hypothesis of interest is

\[
H_0 : \beta_1 = \beta_2 = \cdots = \beta_g = 0
\]

and the alternative hypothesis is

\[
H_a : \text{at least one of } \beta_1, \beta_2, \ldots, \beta_g \text{ is not } 0.
\]

The drop-in-deviance test statistic is a likelihood ratio statistic

\[
LRT = 2 \log \left( \frac{L(\hat{\beta}_{\text{unconstrained}}; (y_1, \ldots, y_n))}{L(\hat{\beta}_{\text{constrained}}; (y_1, \ldots, y_n))} \right) = 2 \log[L(\hat{\beta}_{\text{unconstrained}}; (y_1, \ldots, y_n))] - 2 \log[L(\hat{\beta}_{\text{constrained}}; (y_1, \ldots, y_n))].
\]

Recall that the likelihood computes the probability of obtaining the sample given a set of parameters. The likelihood function is larger when an unconstrained set is used versus a constrained set, just as the linear regression residual sums-of-squares is smaller when comparing an unconstrained model to a constrained model.

Consequently, \( LRT \geq 0 \) whenever the models are nested and the unconstrained model is in the numerator. If \( H_0 \) is true, then the \( LRT \) is approximately chi-square in distribution and the degrees of freedom are \( g \), i.e.,

\[
LRT \sim \chi^2_g.
\]

Large values of \( LRT \) support \( H_a \) and contradict \( H_0 \) and so p-value = \( P(LRT \geq l) \) where \( l \) is the observed value of \( LRT \). It’s convenient to compute the contribution of a fitted model towards \( LRT \) whenever a model is fit, and the contribution is often referred to as the deviance of the model. The deviance is analogous to the residual sums-of-squares in the linear model context.

The R function call `summary(glm.1)`, where `glm.1` is a fitted logistic regression model, prints out the deviance of the fitted model and labels it as the Residual Deviance. Another deviance, called the Null Deviance is computed and reported, but it has no use when the responses are binary variables. Ramsey and Schafer use notation

\[
\text{deviance}_{\text{model}} = -2 \log[L(\hat{\beta}_{\text{model}}; (y_1, \ldots, y_n))]
\]

\(^6\)If the variable of interest is quantitative, then \( k - 1 = 1 = g \).
for the deviance. Consequently, in Ramsey and Schafer’s notation,

\[ LRT = \text{deviance}_{\text{constrained}} - \text{deviance}_{\text{unconstrained}}. \]

Notice that in formula (1) the log-likelihood (which measure the goodness-of-fit) is multiplied by \(-2\) so that the deviance measures lack-of-fit (as does the residual sums-of-squares in the linear model context) whereas the likelihood function measures goodness-of-fit.

If the null hypothesis is correct (and the \(g\) parameters in the unconstrained model are truly all 0), then the LRT statistic is approximately chi-square in distribution and the degrees of freedom are \(g\). It’s helpful to note that mean of a chi-square random variable is the degrees of freedom and the variance is twice the degrees of freedom.

To illustrate, Table 3 shows a summary of an interaction model fit to the Donner party data.

Table 3: Coefficients and standard errors obtained from a logistic regression of survival on gender and age for the Donner party. Residual deviance 47.346 on 41 degrees of freedom, \(n = 45\).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>Wald statistic ((Z))</th>
<th>(P(,Z &gt; \mid z\mid))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>.318</td>
<td>1.131</td>
<td>.281</td>
<td>.77</td>
</tr>
<tr>
<td>Age</td>
<td>-.0324</td>
<td>.0352</td>
<td>-.921</td>
<td>.36</td>
</tr>
<tr>
<td>Gender (female)</td>
<td>6.92</td>
<td>3.398</td>
<td>2.038</td>
<td>.042</td>
</tr>
<tr>
<td>Age (\times) Gender (female)</td>
<td>-.161</td>
<td>.094</td>
<td>-1.71</td>
<td>.086</td>
</tr>
</tbody>
</table>

The unconstrained model deviance (from Table 3) is \(\text{deviance}_{\text{unconstrained}} = 47.346\) and the constrained (or additive) model (Table 1) is \(\text{deviance}_{\text{constrained}} = 51.256\). Hence, the likelihood ratio statistic is

\[ LRT = \text{deviance}_{\text{constrained}} - \text{deviance}_{\text{unconstrained}} = 51.256 - 47.346 = 3.91, \]

and \(p\)-value = \(P(LRT \geq 3.91) = .0482\). Since interaction is accounted for by a single variable, the degrees of freedom are 1. The \(R\) function call for computing the \(p\)-value is \(1-p\text{chisq}(3.91,1)\). The \(R\) function for computing the statistic and the \(p\)-value is \texttt{anova(glm.2,glm.1,test ="Chisq")} where \texttt{glm.2=glm(y\sim age*gender)} is the fitted unconstrained model and \texttt{glm.1} is the fitted constrained model.

**Bird keeping and lung cancer**

A study conducted in The Hague (1972–1981) discovered an association between bird keeping
and lung cancer risk. In 1985, a retrospective\textsuperscript{8} case-control study\textsuperscript{9} was carried out to further investigate the apparent connection between bird keeping and lung cancer risk. Specifically, a central objective was to determine if there was evidence of an increased probability of lung cancer associated with bird keeping, after accounting for other important factors. The researchers\textsuperscript{10} collected data on 49 lung cancer subjects that were less than 66 years old and had resided in the city since 1965. These patients were matched with 99 controls with a similar distribution of ages. The explanatory variables are

1. Gender,

2. Age (in years),

3. Socioeconomic status (high or low) of the household’s principal wage earner, determined by occupation.

4. Years of smoking prior to diagnosis or examination.

5. Average number of cigarettes consumed per day.

6. An indicator variable identifying subjects as engaged in bird keeping. Bird keeping was defined as keeping caged birds in the home for more than 6 consecutive months from 5 to 14 years before diagnosis (cases) or examination (controls).

Of the 147 subjects, 67 kept birds and 80 did not.

\textsuperscript{7}There is some biological reason for an association as exposure to airborne particles increases macrophages in the lungs and macrophages are believed to help cancer cells proliferate.

\textsuperscript{8}Retrospective means that that the data were collected on patients that had already contracted lung cancer.

\textsuperscript{9}A case-control study is one in which patients with lung cancer were (loosely) matched with individuals with similar characteristics (age, gender, smoking habits, and socioeconomic status).

The figure to the right shows the relationship between age and smoking for two groups: bird keepers and non-bird keepers. Individuals having lung cancer are identified. There is an apparent association between smoking and lung cancer; for instance, all but one cancer patient has been smoking for more than 10 years, and all but 4 have been smoking for more than 20 years. In addition, a comparison of panels shows that there are relatively more lung cancer patients in the left panel (bird keepers) than in the right panel (non-bird keepers). Of those with lung cancer, 49.2% kept birds, and of those without lung cancer, 20% kept birds.

Bird keeping does not appear to be moderately or strongly associated with years of smoking or age as there are no substantial differences in the pattern of points in the two panels if the color of the points is ignored.

The researchers (main) objective was to determine whether there is an association between bird keeping and incidence of lung cancer. This statement implies the model fitting strategy should be to fit a rich model with all available explanatory variables besides bird keeping, and determine whether the indicator of bird keeping significantly improves the fit of the rich model. If bird keeping is significant, then the differences in the odds of contracting lung cancer effect between bird keepers and others should be estimated.

Table 4 shows the fitted model obtained from all variables. Most of the variables apparently have little evidence supporting an association with lung cancer.
Table 4: Coefficients and standard errors obtained from a logistic regression of lung cancer on years of smoking and age. The binary response variable takes on the value one if the subject contracted lung cancer and zero if not. Residual deviance 168.83 on 144 degrees of freedom, $n = 147$.

| Variable                  | Coefficient | Std. Error | Wald statistic ($Z$) | $P(Z > |z|)$ |
|---------------------------|-------------|------------|----------------------|-------------|
| Intercept                 | $-1.94$     | $1.80$     | $-1.07$              | $.282$      |
| Gender (Female)           | $.561$      | $.531$     | $1.06$               | $.290$      |
| Years of smoking          | $.0728$     | $.0264$    | $2.75$               | $.0059$     |
| Age                       | $-0.0397$   | $0.035$    | $-1.12$              | $.262$      |
| Socio-economic (High)     | $.105$      | $.468$     | $0.225$              | $.822$      |
| Cigarettes per day        | $.0260$     | $.0255$    | $1.02$               | $.308$      |
| Bird keeping (yes)        | $1.36$      | $.411$     | $3.31$               | $.0009$     |

The drop-in-deviance statistic obtained from adding bird keeping to the main-effects-only model is $LRT = 11.67$ and $p$-value = $.00063$. The coefficient associated with the bird keeping indicator variable is $\hat{\beta}_{BK} = 1.362$ and the odds associated with contracting cancer are estimated to be $e^{1.36} = 3.90$ times greater when comparing two individuals that differ according to whether they keep birds or not, given all other variables are held constant. An approximate 95% confidence interval for this increase is $[e^{1.36-1.96\times.411}, e^{1.36+1.96\times.411}] = [1.74, 8.74]$.

There may be interest in the significance of the other variables, but it ought to be kept in mind that there are other unobserved variables that are associated with the incidence of lung cancer that were not measured (genetics and workplace environment are among the obvious). It seems somewhat futile to attempt to discuss the relative importance observed variables because of these unobserved and latent variables.