Chapter 14: Logistic regression

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 Sierra 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\(^1\) 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 \(y\) to be 1 if an 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 survivorship value besides 0 and 1 should be interpreted.

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 the 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. Consequently,

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

The expected value and variance of $Y$ are

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

and

$$\text{Var}(Y) = \sigma^2 = \pi(1 - \pi).$$

Recall that the linear regression model is of the mean of $Y$ (or expected value $E(Y)$), and the distribution of $Y = E(Y|X) + \epsilon$ is assumed to be normal (given that the mean is correctly modeled). Theoretically, $E(Y)$ may be any real number if $Y$ is normal; in contrast, $0 < E(Y) = \pi < 1$ 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 will produce fitted values that are inconsistent with the response variable.

The solution is to model the logit transformation of $\pi$, as there are no upper or lower bounds on the transformation of $\pi \in (0, 1)$; thus, in logistic regression, logit($\pi$) is modeled as a linear function of $x_1, \ldots, x_p$. The figure below and left shows the graph of the function in red. The function is unbounded since as $\pi \rightarrow 0$, logit($\pi$) $\rightarrow -\infty$ and as $\pi \rightarrow 1$, logit($\pi$) $\rightarrow \infty$. The logit function has the effect of stretching the possible values of $\pi$ from 0 to 1, to $-\infty$ to $\infty$.

The logit function of $\pi$ is

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

The logistic regression model is

$$\text{logit}(\pi|x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$
The inverse of the logit can be computed

\[ \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \log \left( \frac{\pi}{1 - \pi} \right) \]

\[ \Rightarrow e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p} = \frac{\pi}{1 - \pi} \]

\[ \Rightarrow \frac{1}{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}} = \frac{1 - \pi}{\pi} = \frac{1}{\pi} - 1 \]

\[ \Rightarrow 1 + \frac{1}{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}} = \frac{1}{\pi} \]

\[ \Rightarrow \frac{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}} = \frac{1}{\pi} \]

\[ \Rightarrow \frac{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}} = \pi. \]

An estimate of \( \pi \) can be computed given a set of values \( x_1, \ldots, x_p \) using the parameter estimates \( \hat{\beta}_0, \ldots, \hat{\beta}_p \) by computing

\[ \hat{\pi} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p)} \]

or \( \hat{\pi} = \frac{1}{1 + \exp[-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p)]} \).

\(^2\)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 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 agents on health are often based on a comparison of $\pi_A$ to $\pi_B$ through the odds ratio. The odds that $A$ will occur are $\pi_A/(1 - \pi_A)$. For example, if the odds of $A$ are 5 (to 1), then the probability of $A$ is 5 times that of $A^c$; hence,

$$P(A) = 5[1 - P(A)] = 5 - 5P(A)$$

$$\Rightarrow 6P(A) = 5$$

$$\Rightarrow P(A) = 5/6 \text{ and } P(A^c) = 1/6.$$  

Given a second event $B$, 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 non-smoker. To illustrate, suppose that $\pi_A = 3/4$, and $\pi_B = 1/4$; then

$$\frac{\pi_A(1 - \pi_B)}{\pi_B(1 - \pi_A)} = \frac{3/4 \times (1 - 1/4)}{1/4 \times (1 - 3/4)} = 9.$$  

The odds of $A$ happening is 9 times greater than the odds of $B$ happening.

In contrast, suppose that $\pi_A = 3/4$, and $\pi_B = 3/5$. The odds of $A$ are 3 the odds of $B$ are 1.5; hence, the odds of $A$ are twice that of $B$ and the odds ratio of $A$ to $B$ is 2.

When $A$ and $B$ are rare events, then the odds ratio is much closer to the ratio of $\pi_A$ to $\pi_B$. For instance, suppose that $\pi_A = .01$ and $\pi_B = .005$. Then,

$$\frac{\pi_A(1 - \pi_B)}{\pi_B(1 - \pi_A)} = \frac{.01 \times .995}{.005 \times .99} = 2.010.$$  

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 odds of $A$ are

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

Similarly, let $\pi_B$ denote the probability when $x = x_B$; then

$$ \frac{\pi_B}{1 - \pi_B} = e^{\beta_0 + \beta_1 x_B}. $$

The odds of $Y = 1$ given $x = A$ relative to the odds of $Y = 1$ 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 $Y = 1$ 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 $P(Y = 1)$.

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 $P(Y = 1)$ 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 $P(Y = 1)$ by $100(1 - .6065)\% = 39.35\%$.

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.\(^3\)

\(^3\)A better significance test is developed below.
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               | −0.0782     | 0.0373     | −2.097                   | 0.0359            |
| Gender (female)   | 1.597       | 0.755      | 2.114                    | 0.0345            |

The model shown in Table 1 is

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

or equivalently,

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

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

\[
\frac{1}{e^{1.597}} = e^{-1.597} = 0.202
\]

times the odds of survival for females of the same age. 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 \( \hat{\pi}_{female,45} = .428 \). For males of age 45, \( \hat{\pi}_{male,45} = .132 \). 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 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 maximum likelihood is best according to several general and important statistical criteria (such as asymptotic unbiasedness and minimum variance). It’s useful then to examine the likelihood function and maximum likelihood estimation.

The likelihood of the data, in the case of discrete data\(^4\), is the probability of obtaining the sample as a function of the parameters.

For example, suppose that a random sample of \(n = 20\) independent Bernoulli observations are collected and the sample is \((y_1, \ldots, y_{20}) = (0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1)\). Each observation is a realization from the \(\text{Bin}(1, \pi)\) 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 \(L[\pi; (y_1, \ldots, y_{20})]\) of observing this sample of 20 observations is the product of the individual

\(^4\)In the case of continuous data, the definition is somewhat more complicated
probabilities, that is

\[ L[\pi; (y_1, \ldots, y_{20})] = P(Y_1 = 0, Y_2 = 0, \ldots, Y_{20} = 1) \]
\[ = P(Y_1 = 0) \times P(Y_2 = 0) \times \cdots \times P(Y_{20} = 1) \]
\[ = (1 - \pi) \times (1 - \pi) \times \cdots \times \pi \]
\[ = \pi^4 (1 - \pi)^{16}. \]

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 \( \log\{L(\pi; (y_1, \ldots, y_{20}))\} \) 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 for the parameter results in a lesser probability, and so any other value is less consistent with the sample data.

The maximum likelihood estimate can be determined by differentiating the likelihood, setting the derivative to zero and solving 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 \( \hat{\pi} = 0.2 \).

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

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

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 2: \( \log[L(\pi; (y_1, \ldots, y_{45}))] \) as a function of \( \beta_0, \beta_1, \beta_2 \).

<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>-0.50</td>
<td>1.25</td>
<td>-27.7083</td>
</tr>
<tr>
<td>1.80</td>
<td>-0.80</td>
<td>1.25</td>
<td>-28.7931</td>
</tr>
<tr>
<td>1.70</td>
<td>-0.50</td>
<td>1.25</td>
<td>-29.0467</td>
</tr>
<tr>
<td>1.80</td>
<td>-0.80</td>
<td>1.25</td>
<td>-30.3692</td>
</tr>
<tr>
<td>1.63</td>
<td>-0.078</td>
<td>1.60</td>
<td>-25.6282</td>
</tr>
</tbody>
</table>

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

**Properties of maximum likelihood estimators**

If the model is correct and the sample size is large, then

1. Maximum likelihood estimators are nearly unbiased.

2. The standard errors of maximum likelihood estimators are nearly unbiased.

3. Maximum likelihood estimators are more precise than nearly all other estimators.

4. The distribution of a maximum likelihood estimator is approximately normal\(^5\).

**Tests and confidence intervals involving a single parameter**

Approximate normality implies that if \( \hat{\beta}_i \) is the maximum likelihood estimator of \( \beta_i \), then

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

The standard deviation of the MLE \( \hat{\beta}_i \) is estimated by \( \hat{\sigma}(\hat{\beta}_i) \), the standard error of \( \hat{\beta}_i \). The normal approximation of the distribution of \( \hat{\beta}_i \) leads to the Wald test of

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

The test statistic is

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

\(^5\)A consequence of the Central Limit Theorem
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-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.

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 $\hat{\beta}_3 = 1.597$ and $\hat{\sigma}(\hat{\beta}_3) = 0.755$. The Wald statistic approximate p-value is .0345 because

$$Z = \frac{\hat{\beta}_3}{\hat{\sigma}(\hat{\beta}_3)} = \frac{1.597}{0.755} = 2.114$$

and $2P(Z \geq 2.114) \approx .034$, based upon $Z \sim N(0, 1)$. A better test statistic (obtained from the drop-in-deviance test) yielded p-value = .025. The estimate $\hat{\beta}_3 = 1.597$ implies that the difference in odds of survival comparing females to males is a factor of 1.597. An approximate 95% confidence interval for this multiplicative factor is 1.195 to 3.228.

Table 1 shows that the the odds of survival is estimated to decrease by a multiplicative factor of $e^{-0.978} = .925$ 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 .925$^{10} = .458$ for the older female. Said another way, the odds of survival are $e^{10 \times 0.978} = 2.18$ 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 for linear regression (and is very similar in several respects). The test is commonly 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 as the drop-in-deviance p-value is more accurate (almost always) than the Wald statistic p-value.

The fit of the full (or unconstrained) model 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 except for the explanatory variable of interest. If the difference in fit is sufficiently large, then it is concluded that the variable of interest is associated with the response variable given all other variables (in both models).
If the explanatory variable is a factor with $k$ levels, then compared to the reduced model, the full 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 and the parameters associated with the $k - 1$ variables are $\beta_1, \beta_2, \ldots, \beta_{k-1}$. The null hypothesis of interest is

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$$

and the alternative hypothesis is

$$H_a : \text{at least one of } \beta_1, \beta_2, \ldots, \beta_{k-1} \text{ is not 0.}$$

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

$$D = 2 \log \left( \frac{L(\hat{\beta}_{\text{full}}; (y_1, \ldots, y_n))}{L(\hat{\beta}_{\text{reduced}}; (y_1, \ldots, y_n))} \right)$$

$$= 2 \log[L(\hat{\beta}_{\text{full}}; (y_1, \ldots, y_n))] - 2 \log[L(\hat{\beta}_{\text{reduced}}; (y_1, \ldots, y_n))].$$

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, $D \geq 0$ whenever the models are nested and the unconstrained model is in the numerator. If the $H_0$ is true, then $D$ is approximately chi-square in distribution and the degrees of freedom are $k - 1$, (if the variable of interest is quantitative, then $k - 1 = 1$). In either case,

$$D \sim \chi^2_{k-1}.$$

Large values of $D$ support $H_a$ and contradict $H_0$, and so $p$-value $= P(D \geq d)$ where $d$ is the observed value of $D$. It’s convenient to compute the contribution of a fitted model towards $D$ whenever a model is fit, and the contribution is often referred to as the deviance of the model.$^6$ Let

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

Then,

$$D = \text{deviance}_{\text{reduced}} - \text{deviance}_{\text{full}}.$$

Notice that in formula (1) the log-likelihood (which measure the goodness-of-fit) is multiplied by $-1$ so that the deviance measures lack-of-fit whereas the likelihood function measures goodness-of-fit.

$^6$The R function call `summary(glm)` 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.
The chi-square distribution describes the distribution of a non-negative random variable. The mean of the random variable is the degrees of freedom and the variance is twice the degrees of freedom. Thus, if \( H_0 \) is correct, the expectation of \( D \) is approximately \( k - 1 \).

To illustrate, Table 3 shows the results of fitting the interaction model using the Donner party data.

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

| Variable | Coefficient | Std. Error | Wald statistic (\( Z \)) | \( P(Z > |z|) \) |
|----------|-------------|------------|--------------------------|----------------|
| Intercept| .318        | 1.131      | .281                     | .77            |
| Age      | -.0324      | .0352      | -.921                    | .36            |
| Gender (female) | 6.92 | 3.398 | 2.038 | .042 |
| Age\( \times \)Gender (female) | -.161 | .094 | -1.71 | .086 |

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

\[
D = \text{deviance}_{\text{reduced}} - \text{deviance}_{\text{full}} = 51.256 - 47.346 = 3.91.
\]

Since interaction is accounted for by a single variable, the degrees of freedom are 1 and under \( H_0, D \sim \chi^2_1 \Rightarrow p\text{-value} = P(D \geq 3.91) = .0482 \). The R function call for computing the p-value is \( 1-p\text{chisq}(3.90,1) \). The R function for computing the statistic and the p-value is \( \text{anova(glm.2,glm.1,test \"Chisq\")} \) where \( \text{glm.2=glm(y}\sim\text{age}\times\text{gender}) \) is the fitted full model and \( \text{glm.1=glm(y}\sim\text{age+gender}) \) is the fitted reduced model.

**Bird keeping and lung cancer**

A study conducted in The Hague (1972–1981) discovered an association between bird keeping and lung cancer risk.\(^7\) In 1985, a retrospective\(^8\) case-control study\(^9\) 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

\(^7\)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.

\(^8\)Retrospective means that the data were collected on patients that had already contracted lung cancer.

\(^9\)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).
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.

The figure to the right shows the relationship between age and smoking for two groups: bird keepers and non-bird keepers.

The red points identify individuals having lung cancer. There is an apparent association between smoking and lung cancer; for instance, all but one cancer patient had 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 5 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 \((y = 1)\) on years of smoking and age. 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                       | -.0397      | .035       | -1.12                   | .262           |
| Socio-economic (High)     | .105        | .468       | .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 \(D = 11.67\) 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.

Logistic discrimination

Logistic regression models may be used for predicting the class membership of an observation for which only the explanatory variables are observed. For instance, formal diagnostic procedures examine one or more symptoms of a patient and attempt to correctly assign a
cause to the symptom. A number of diseases are not readily diagnosed without invasive procedures such as biopsy; some, for which biopsies are not informative (e.g., Lyme disease) are often very difficult to diagnose. Since the variable of interest is binary (disease present or absent), it may be possible to develop an analytical prediction function using a fitted logistic regression model. This approach is often called logistic discrimination.

Logistic discrimination assumes that the response variable (or target variable in the language of statistical learning) is binary (identifying the class membership of the observation). The observation is best viewed as a pair \((y_0, x_0)\) where \(y_0\) identifies the membership of the input vector \(x_0\). For example, the \texttt{SAheart} data set in the \texttt{ElemStatLearn} library describes a retrospective sample of \(n = 462\) males in a heart-disease high-risk region of the Western Cape, South Africa. There are roughly two controls per case of coronary heart disease. The classes are coronary heart disease and no coronary heart disease, and the objective is to predict the coronary heart disease status of an similar male based on a subset of the input variables recorded in the course of the study. The input variables are inexpensive to measure and so may be used to screen individuals before a series of more expensive diagnostic procedures are undertaken.

The input variables are

<table>
<thead>
<tr>
<th>Input variables</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>systolic blood pressure</td>
<td>quantitative</td>
</tr>
<tr>
<td>tobacco use</td>
<td>quantitative</td>
</tr>
<tr>
<td>low density lipoprotein cholesterol</td>
<td>quantitative</td>
</tr>
<tr>
<td>adiposity</td>
<td>quantitative measure of body fat</td>
</tr>
<tr>
<td>family history of heart disease</td>
<td>factor with levels absent and present</td>
</tr>
<tr>
<td>type-A behavior</td>
<td>quantitative</td>
</tr>
<tr>
<td>obesity</td>
<td>quantitative (possibly body mass index)</td>
</tr>
<tr>
<td>current alcohol consumption</td>
<td>quantitative</td>
</tr>
<tr>
<td>age</td>
<td>years</td>
</tr>
</tbody>
</table>

For the first training observation, \(x_1 = (160, 12.00, 5.73, 23.11, \text{Present}, 49, 25.30, 97.20, 52)\) and \(y_1 = 1\), identifying the individual as a member of the coronary heart disease group. All the input variables have been implicated in the development of coronary heart disease. The primary focus is on determining which set of variables is best for predicting the class membership of another individual, and how accurate is a prediction function constructed from these variables?

The logistic prediction function is constructed by finding a good logistic regression model of \(\text{logit}[\pi/(1 - \pi)]\) where \(\pi = P(Y = 1|x)\) is the probability that an individual will belong...
to the coronary heart disease group \((Y = 1)\) given a set of input variables collected as the input vector \(x\). Given \(x_0\), the prediction of \(y_0\) is determined by estimating \(\text{logit}[\pi_0/(1 - \pi_0)]\) by plugging the input variables into the fitted model. If
\[
\hat{\text{logit}}[\pi_0/(1 - \pi_0)] = \hat{\beta}_0 + \hat{\beta}_1 x_{0,1} + \cdots + \hat{\beta}_p x_{0,p} \geq 0,
\]
then \(\hat{\pi}_0 \geq .5\), and the prediction of \(Y_0\) is 1 (the individual is predicted to belong to the coronary heart disease group.) The choice of \(\hat{\pi}_0 \geq .5\) as predicting membership in coronary heart disease group might be modified; setting \(\hat{\pi}_0 \geq .25\) (or some other value less than .5) as an indicator of membership in coronary heart disease risk group may be a better strategy for preventing worsening health in patients.

A logit prediction model was found by beginning with a model with tobacco use and family history in the model (my intuition) and determining which other variables improve the 10-fold cross-validation rate of accuracy. The final model was

\[
\begin{align*}
\text{Intercept} & \quad -4.20 \quad .4983 \quad -8.437 \quad < .0001 \\
tobacco & \quad .0807 \quad .0255 \quad 3.16 \quad .0015 \\
famhistPresent & \quad .924 \quad .2231 \quad 4.14 \quad < .0001 \\
ldl & \quad .167 \quad .0541 \quad 3.09 \quad .0020 \\
age & \quad .0440 \quad .0097 \quad 4.521 \quad < .0001
\end{align*}
\]

Table 5: Coefficients and standard errors obtained from a logistic regression of lung cancer \((y = 1)\) on years of smoking and age. Residual deviance 168.83 on 144 degrees of freedom, \(n = 147\).

Accuracy assessment results are often presented in the form of a confusion matrix (below). The table is the result of 10-fold cross-validation using the R function CVbinary from the DAAG library. The cross-validation predictions were computed using the function \(\text{pred} \leftarrow \text{as.integer(CVbinary(glm.obj)$cv>.5)}\). The code below constructs a confusion matrix.

\[
A \leftarrow \text{table(pred,SAheart$chd)}
\]
\[
\text{addmargins}(A)
\]

where \(\text{glm.obj}\) is the fitted logistic regression model.
The misclassification rate is estimated to be $(80 + 49)/462 = .279$. Two other important statistics are the *false positive rate*—the proportion of individuals that are positive (CHD present) given they are predicted not to be positive (predicted to be CDH free); the estimate is $80/333 = .240$. The false negative rate is the proportion of individuals with CHD present that are predicted to be free of CHD and the estimate is $80/129 = .380$. Using the matrix $A$ above, the code is

\[
\text{c("False positive rate = ", A[1,2]/sum(A[1,]))}
\]
\[
\text{c("False negative rate = ", A[2,1]/sum(A[2,]))}.
\]

The false positive and false negative rate provide information on the accuracy of a prediction from the standpoint of someone that is predicted to belong to one of the two classes. The error is conditional upon the predicted class.

*Residual analysis*

The residuals from a logit regression are

\[
y_i - \hat{\pi}_i = \begin{cases} 
1 - \hat{\pi}_i, & \text{if } y_i = 1, \\
-\hat{\pi}_i, & \text{if } y_i = 0.
\end{cases}
\]

Consequently, a plot of the residuals against the fitted values $\hat{y}_i = \hat{\pi}_i, i = 1, \ldots, n$ is uninformative.