SS9055B: Generalized Linear Models

Section 5: Inference for Generalized Linear Models

Objectives

This reading discusses different ways of interpreting the results from a logistic regression model. By the end of the lecture you should be able to:

- interpret the slope parameter in a logistic regression model in relation to:

  1) a single indicator variable,
  2) relative risks,
  3) latent tolerance models, and
  4) linear approximations.

- comment on alternative link functions for the binomial model including the log, probit, and log-log.

Introduction

So far in the course we have looked at generalized models (GLM) in an abstract sense.We will now turn our attention to the most important specific examples of GLM, starting with logistic regression for binomial data. Fortunately, most of the work has already been done. Through the previous (or coming) exercises, you have already shown that the logistic regression model fits within the GLM framework, derived the steps of the IWLS algorithm needed to compute estimates and their standard errors, developed methods to compare models with different sets of covariates either through hypothesis tests or via the AIC, constructed confidence intervals for the coefficients and, by extension, the fitted values, and identified methods to assess the goodness-of-fit (Phew!). That covers all of the mechanical aspects of logistic regression, and so what really remains is to discuss how we interpret the models we have fit. That's what you will cover in this reading. There are a few other wrinkles, but we'll examine those through the in-class activities.

Interpreting the Logistic Regression Model

Logistic regression is used to study how the probability of success for some binary variable varies as a function of measured covariates. The term success is used arbitrarily to refer to whichever outcome is labelled 1 and does not need to be a good thing. It seems rather morbid to discuss successfully dying from some horrible, rare disease. Alas, that is the terminology we are stuck with. Mathematically, we suppose that the random variables $Y_1, \ldots, Y_N$ represent the observed proportion of successes in separate trials of an experiment such that $n_i Y_i | \boldsymbol{x}_i \sim$ Binomial(n_i, π_i), $i = 1, \ldots, N$, and

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

where $\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = x_i'\boldsymbol{\beta}$ is the linear predictor constructed from covariates $x_{i1}, \ldots, x_{ip}$. In this section, we are going to consider how to interpret the regression coefficients, $\beta_1, \ldots, \beta_p$.

Odds are Odd

The simplest way to interpret the coefficients is to consider that we are modelling the log-odds of success, $\log(\pi_i/(1-\pi_i))$, as a linear function of the predictors. In a simple linear regression model the coefficient β_j represents the increase in the mean of the response per unit increase in x_{ij} when all of the remaining predictors are fixed. In the logistic model, $\hat{\beta}_j$ is an estimate of the increase in the log-odds of success per unit increase in x_{ij} and $\exp(\hat{\beta}_j)$ is an estimate of the multiplicative increase in the odds itself when x_{ij} increases by one unit and all of the remaining predictors are fixed. That is if

$$x_{i_2 k} = \begin{cases} x_{i_1 k} & k \neq j \\ x_{i_1 k} + 1 & k = j \end{cases}$$

then

$$\beta_j = \log\left(\frac{p_{i_2}}{1-p_{i_2}}\right) - \log\left(\frac{p_{i_1}}{1-p_{i_1}}\right)$$

and

$$\exp(\beta_j) = \left(\frac{p_{i_2}}{1-p_{i_2}}\right) \Big/ \left(\frac{p_{i_1}}{1-p_{i_1}}\right).$$

Note that you can construct confidence intervals for $\exp(\beta_j)$ by transforming the confidence interval for β_j. This is preferable to using the delta method to obtain an approximate normal distribution for β_j because it will ensure that the lower limit of the confidence interval remains above 0.

A specific case of particular importance is to consider the effect of a single, binary covariate. Suppose that $N = 2$ and

$$\eta_i = \beta_0 + \beta_1 x_i$$

where $x_1 = 0$ and $x_2 = 1$. Then

$$\beta_1 = \eta_2 - \eta_1 = \log\left(\frac{\pi_2}{1-\pi_2}\right) - \log\left(\frac{\pi_1}{1-\pi_1}\right) = \log\left[\frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)}\right].$$

The quantity inside the brackets of the final expression is called the odds ratio for the two groups and β_1 represents the log odds-ratio for the two groups. In this model, $\hat{\beta}_1$ is an estimate of the log odds-ratio and $\exp(\hat{\beta}_1)$ is an estimate of the odds-ratio between the two groups.

Unfortunately, it is not intuitive to work with odds-ratios and log odds-ratios. One of the reasons for this is that you can't really interpret what the odds ratio means without knowing one of the probabilities. Suppose $\pi_1 = .5$ and we double the odds ratio (i.e., $\beta_1 = \log(2) = .69$) then $\pi_2 = .67$, which is 1.33 times bigger. However, if $\pi_1 = .90$ and we double the odds ratio then $\pi_2 = .95$, which is only 1.05 time bigger. This makes it difficult to understand what effects on the odds-ratios really mean and, because of this, we will consider different ways to interpret the β_1 that are useful in specific situations.

Relative Risk

A more intutive way to compare probabilities between two groups is to consider the relative risk π_2/π_1. A relative risk of 2, for example, means that the probability of the outcome in one group is twice as high as the probability of the outcome in the other group – regardless of the actual values of the probabilities. The relative risk arises if we consider the model with the log link instead of the logistic link function. Again consider the model with a single binary predictor so that $\log(\pi_1) = \beta_0$ and $\log(\pi_2) = \beta_0 + \beta_1$. Then

$$\beta_1 = \log(\pi_2) - \log(\pi_1) = \log\left(\frac{\pi_2}{\pi_1}\right)$$

and

$$\exp(\beta_1) = \frac{\pi_2}{\pi_1}.$$

Hence, if we consider the binomial model with log link function then $\hat{\beta}_1$ and $\exp(\hat{\beta}_1)$ are estimates of the log relative risk and relative risk. Relative risks are much easier to interpret because they tell us exactly how the probability changes. If the relative risk is 2 then $p_2 = 2p_1$, regardless of the value of p_1.

However, working with relative risks has two problems. First, the log link requires that the linear predictor be negative so that the fitted probability remains less than 1. This constraint is difficult to work, particularly when there are multiple predictors in the model. Second, estimates of the relative risk are not appropriate in retrospective studies, like case-control studies which are very common in health research. To see why, consider the following example:

A retrospective case-control study conducted in the 1950's examined the correspondence between smoking and lung cancer rates. The study identified 709 hospital patients with lung cancer and matched them with 709 control subjects. The subjects were then classified as either smokers or non-smokers.

	Lung Cancer	
	Yes	No
Non-smoker	21	59
Smoker	688	650
Total	709	709

Based on this data the estimated probabilities of lung cancer for smokers and non-smokers are:

a) Non-smokers: $21/(21+59) = .26$ b) Smokers: $688/(688+650) = .51$

Now compare the effects of smoking on the rate of lung cancer:

a) The estimated relative risk is:

$$\frac{.51}{.26} = 1.96$$

b) The estimated odds ratio is:

$$\frac{.51/.49}{.26/.74} = 2.97$$

Suppose now that the researchers had found 2 controls for every lung cancer case but the observed proportions of smokers and non-smokers in the two groups remained exactly the same. The new data would be:

	Lung Cancer	
	Yes	No
Non-smoker	21	118
Smoker	688	1300
Total	709	1418

Intuitively our inference should not change. All we have done is to double the number of controls, but the rate of smoking within the controls has remained the same. However, based on this data the estimated probabilities of lung cancer for smokers and non-smokers are:

a) Non-smokers: $21/(21+118) = .15$ b) Smokers: $688/(688+1300) = .35$

Now compare the effects of smoking on the rate of lung cancer:

a) The estimated relative risk is:

$$\frac{.35}{.15} = 2.29$$

b) The estimated odds ratio is:

$$\frac{.35/.65}{.15/.85} = 2.97$$

Notice that the estimated probabilities and estimated relative risk both changed when we doubled the number of controls for each case, but the odds ratio remains the same. This makes the odds ratio preferable in retrospective studies like this.

Fortunately, all is not lost with respect to relative risks. Notice that if both $(1 - \pi_2)$ and $(1 - \pi_1)$ are close to 1 (i.e., if π_1 and π_2 are small) then the odds-ratio is approximately equal to the relative risk. I.e.,

$$\frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)} \approx \frac{\pi_2}{\pi_1}.$$

This means that if the outcome of interest is rare, as in the case of lung cancer, then we can interpret $\hat{\beta}_1$ from the logistic regression model as an estimate of the log relative risk and $\exp(\hat{\beta}_1)$ as an estimate of the relative risk, regardless of how the study was designed. You will often see researchers interpret odds ratios as if they were relative risks, but be careful to do this only if the outcome of interest is rare.

Latent Tolerance Model

An alternative interpretation of the parameters in a logistic regression model is to consider what is called a latent tolerance model. Suppose that we conduct an experiment in which we apply a toxin to individuals sampled from some population. Let N be the number of different doses of the drug tested, x_i be the $i-th$ dose, n_i be the number of insects treated with this dose, and Y_i the proportion of these insects that live. This is the setup of the picloram example that you have already seen.

The tolerance is the amount of the toxin that an individual can withstand before it dies. Let T_{ij} denote the tolerance of the j^{th} individual in the i^{th} group. The individual will live if the dose is less than its tolerance, $x_i < T_{ij}$, and will die if the dose is greater than the tolerance, $x_i \geq T_{ij}$. Suppose now that the tolerances are randomly drawn from a logistic distribution with mean μ and scale parameter τ. Then each T_{ij} has pdf

$$f(x|\mu,\tau) = \frac{\exp\left(\frac{x-\mu}{\tau}\right)}{\tau\left(1+\exp\left(\frac{x-\mu}{\tau}\right)\right)^2},$$

cdf

$$F(x|\mu,\tau) = \frac{\exp\left(\frac{x-\mu}{\tau}\right)}{1+\exp\left(\frac{x-\mu}{\tau}\right)}.$$

and variance $\tau^2\pi^2/3$. Given this assumption, the probability that the j^{th} individual in the i^{th} group dies is

$$P(T_{ij} \leq x_i) = \pi(x_i) = \frac{\exp\left(\frac{x_i-\mu}{\tau}\right)}{1+\exp\left(\frac{x_i-\mu}{\tau}\right)}$$

or equivalently:

$$\text{logit}(\pi(x_i)) = \frac{x_i-\mu}{\tau} = \beta_0 + \beta_1 x_i$$

where $\beta_0 = -\mu/\tau$ and $\beta_1 = 1/\tau$.

This shows that fitting a logistic regression model to the data $Y_1, \ldots, Y_n$ with intercept β_0 and slope β_1 is equivalent to modeling the tolerance of the population with a logistic distribution having mean $\mu = -\beta_0/\beta_1$ and scale parameter $\tau = 1/\beta_1$. This analogy only holds if $\beta_1 > 0$, though it is possible to create similar analogies when $\beta_1 < 0$. Once again, we can make inference about μ and τ after fitting the logistic regression model by using the delta method.

Note that we could generate alternative models by assuming a different distribution for the tolerances. Generally, if the CDF of the tolerance is $F(x)$ then we could set $g(\pi) = F^{-1}(\pi)$. The most common alternative is to assume that the tolerance is normally distributed, $T_{ij} \sim N(\mu, \sigma^2)$. This is equivalent to fitting a binary GLM with the link function:

$$g(\pi_i) = \Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_i$$

where $\Phi^{-1}(\cdot)$ is the inverse CDF of the standard normal distribution, $\beta_0 = -\mu/\sigma$ and $\beta_1 = 1/\sigma$. The model with a binomial error distribution

$$n_i Y_i | x_i \sim \text{Binomial}(n_i, \pi_i)$$

and the inverse normal CDF link function

$$\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_i$$

is called the probit regression model.

You may think that the probit model is actually more natural than the logistic model, and this is true, but it is also more complicated to fit because the inverse link function is more difficult to compute than the inverse logistic CDF[1]. Note that the shape of the logistic distribution is very close to the shape of the normal distribution if we set the parameters correctly. The variance of the logistic distribution with parameters μ and τ^2 is $\sigma^2 = \pi^2\tau^2/3$. Equating the means and variances we find that a logistic distribution with parameters μ and $\tau^2 = 3\sigma^2/\pi^2$ very closely approximates the normal distribution with mean μ and variance σ^2. For this reason we can interpret the logistic regression model as an approximation to a probit model (i.e., an approximation to a tolerance model with normally distributed tolerances).

[1] This is really a historical problem. Computation for the probit model is simple with modern computers.

Linear Approximation

A final way to interpret β_1 when x is continuous is to consider a linear approximation to $\pi(x)$. If

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

then

$$\pi'(x) = \beta_1 \pi(x)(1-\pi(x)).$$

The derivative is maximized when $\pi(x) = 1/2$. This occurs when $\beta_0 + \beta_1 x = 0$ which implies $x = -\beta_0/\beta_1$. At this point, the slope is $\beta_1/4$. For x close to $-\beta_0/\beta_1$ we can approximate $\pi(x)$ by the linear function:

$$\pi(x) \approx .5 + \frac{\beta_1}{4}(x + \beta_0/\beta_1) = \left(.5 + \frac{\beta_0}{4}\right) + \frac{\beta_1}{4}x.$$

The point $x = -\beta_0/\beta_1$ at which $\pi(x) = .5$ is called the median effect level or the LD_{50} where LD stands for lethal dose. Thinking of the latent tolerance model with $T_{ij} \sim \text{logistic}(-\beta_0/\beta_1, 1/\beta_1)$, the LD_{50} represents the median tolerance over the population (i.e., the dose which kills 50% of the individuals in the population).

We could also use this linear function to approximate specific probabilities close to the LD_{50}. For example, according to this linear function:

$$\pi(-\beta_0/\beta_1 - 1/\beta_1) \approx .25 \text{ and } \pi(-\beta_0/\beta_1 + 1/\beta_1) \approx .75.$$

The points $x_{.25} = -\beta_0/\beta_1 - 1/\beta_1$ and $x_{.75} = -\beta_0/\beta_1 + 1/\beta_1$ are called the LD_{25} and LD_{75} values, and the logistic curve is always close to linear between these two points. If all of the values of the predictor lie between $x_{.25}$ and $x_{.75}$ then we can say that the probability of success increases by, approximately, β_1 for each unit increase in x within this range. The linear approximation for the model with $\beta_0 = -2.5$ and $\beta_1 = .5$ is shown in Figure 1.

Alternative Link Functions

We have now encountered three alternative link functions for the binomial model:

a) $g(\pi) = \text{logit}(\pi)$ – canonical link that relates the log-odds to the linear predictor and is equivalent to the logistic model of tolerance. Coefficients in this model represent the change in the log-odds as one predictor increases while the remaining predictors are fixed.
b) $g(\pi) = \log(\pi)$ – relates the log probability to the linear predictor. Coefficients in this model represent the change in the log relative risk as one predictor increases while the remaining predictors are fixed.
c) $g(\pi) = \Phi^{-1}(\pi)$ – probit link which is equivalent to the normal model of tolerance.

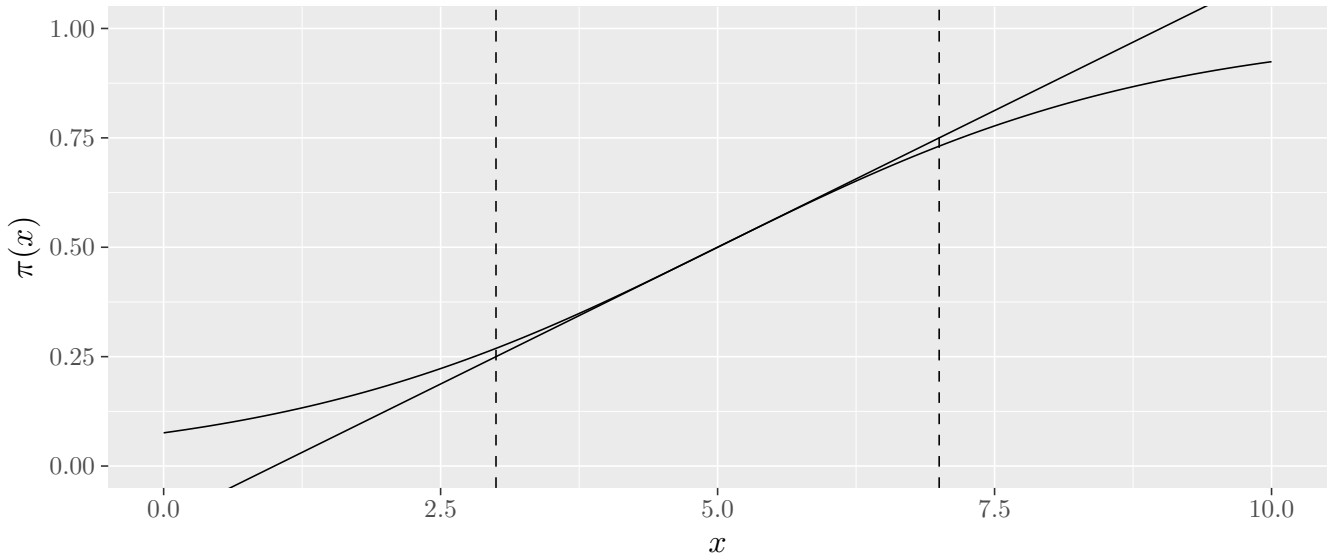

Figure 1: Linear approximation to the logistic regression curve with $\beta_0 = -2.5$ and $\beta_1 = .5$. The solid line indicates the approximation which has intercept $.5+\beta_0/4 = -.125$ and slope $\beta_1/4 = .125$. The dashed vertical lines represent the LD_{25} and LD_{75} values ($(-\beta_0 - 1)/\beta_1 = 3$ and $(-\beta_0 + 1)/\beta_1 = 7$ respectively).

Another function that is sometimes used is the log-log link:

$$g(\pi) = \log(-\log(1 - \pi)).$$

Like the logit link, the complentary log-log has the advantage that it ranges over the entire real line, so we don't have to worry about the value of η. However, it does not have an easy interpretation. If $\log(-\log(1 - \pi_1)) = \beta_0$ and $\log(-\log(1 - \pi_2)) = \beta_0 + \beta_1$ then

$$\exp(\beta_1) = \frac{\log(1 - \pi_2)}{\log(1 - \pi_1)}.$$

The ratio of log probabilities is not simple to work with. It is also important to note that the log-log function is not symmetric, $g(\pi) \neq -g(-\pi)$. While this is not necessarily a problem, it is a difference to keep in mind. The four link functions are compared in Figure 2.

Conclusion

The regression coefficients from a logistic regression model are strictly interpreted in terms of the effect of the covariates on the (log) odds-ratios. However, the scale of odds-ratios is not natural and makes this makes it difficult to understand what these effects really mean. Because

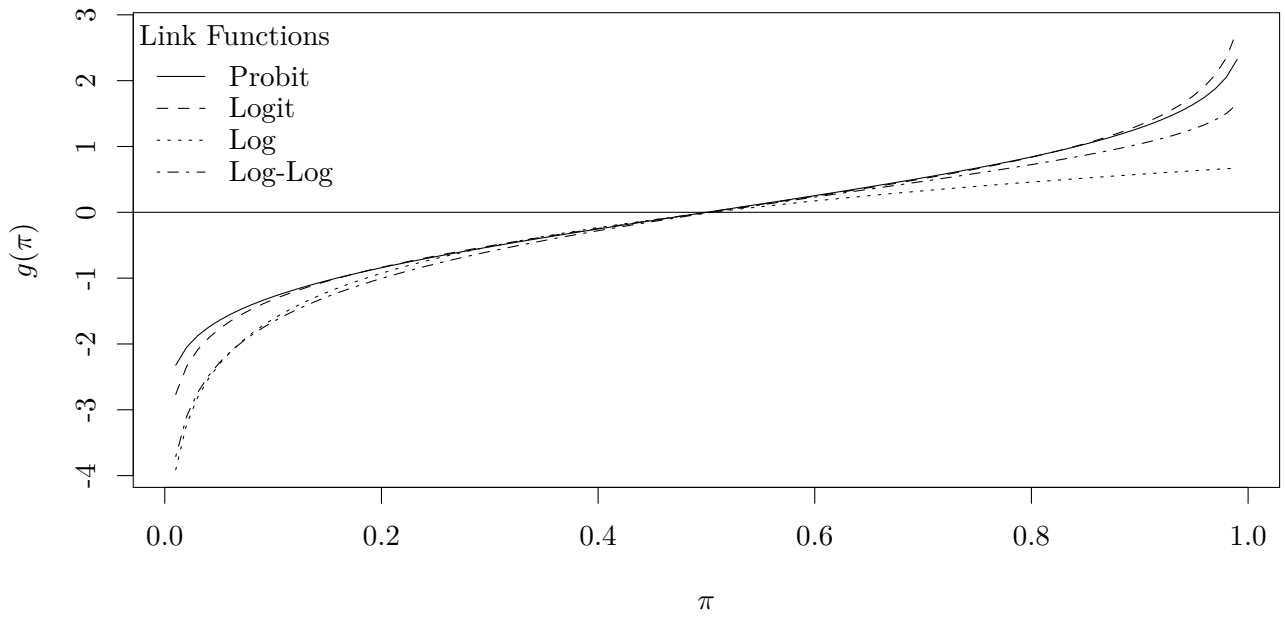

Figure 2: Four common link functions for binomial regression centered and scaled to match the probit model.

of this, it is often more useful to interpret the coefficients in terms of some sort of approximate effect. If the probability of the outcome is rare, as is true for many diseases, then the odds-ratios approximate relative risks. In this case, we can interpret the coefficients as approximations to the change in risk as a covariate increases in value while the other covariates remain fixed. Alternatively, we may interpret the model in terms of a latent tolerance distribution. The logistic model corresponds directly to a logistic distribution latent tolerance, but this is also very close to a normal distribution. If the fitted probabilities are all reasonably far from the extremes (approximately between .25 and .75) then the logistic function is approximately linear and in this case we can interpret the effects of the covariates as having an approximately linear effect on the probability of success. Finally, there is no specific reason that we must use the logistic link function for binomial data. The logistic link function is the canonical link, but this is mostly of historical importance. Modern computers make it very easy to fit models with different link functions and in practice some thought should be given to choosing the link function that is most appropriate for the problem at had.

Goodness-of-Fit

Deviance Goodness-of-Fit Test

One important challenge in fitting logistic regression models is the determination of model it. The asymptotic distribution of the deviance goodness-of-fit test described in Section 3.3 of the previous set of notes relies on the fact that the mean of the observed response values for each unique combination of the predictors is approximately normally distributed. This is true for the normal model under all conditions. However, if the response variable is not normally distributed then we need to appeal to the central limit theorem with the rule of thumb that the number of observations (for each combination of the predictors) is greater than 30.

This means that if the number of observations for any combination of the predictor variables is small then the deviance goodness-of-fit statistic may no longer follow a chi-square distribution under the null hypothesis. Tests conducted using this statistic may then provide the wrong result, leading us to believe that a model fits the data when it doesn't or vice versa. This is particularly a problem if one of the predictors is a continuous value that is observed instead of being controlled in the design of the experiment. In this case, it is likely that each value of the covariate will be represented only once in the observed data.

Pearson Chi-square Test

An alternative test for goodness-of-fit of a generalized linear model is the Pearson chi-square test. Suppose that the data contain G unique combinations of the predictor values (which I'll refer to henceforth as the groups). Let n_g denote the number of observations in the g-th group, y_g the proportion of successes, and $\hat{p}_g$ the predicted probability of success based on the model of interest. The Pearson chi-square statistic for a logistic regression model is

$$X^2 = \sum_{g=1}^{G} \left(\frac{(n_g y_g - n_g \hat{p}_g)^2}{n_g \hat{p}_g} + \frac{(n_g(1-y_g) - n_g(1-\hat{p}_g))^2}{n_g(1-\hat{p}_g)} \right) \qquad = \sum_{g=1}^{G} \frac{(n_g y_g - n_g \hat{p}_g)^2}{n_g \hat{p}_g (1-p_g)}.$$

Heuristically, the model compares the observed numbers of successes and failures, $n_g y_g$ and $n_g(1-y_g)$, with their predicted values, $n_g p_g$ and $n_g(1-p_g)$, for each of the groups. If the observed values are close to their expected values then X^2 will be small. If the observed values are far from their expected values then X^2 will be big.

You will not be surprised to hear that Pearson's chi-square statistics is asymptotically distributed according to a chi-square distribution under the same conditions that ensure that the deviance statistic also has an asymptotic chi-square distribution. That's where it's name comes from! In fact, the two statistics are asymptotically equivalent (i.e., they are almost equal in large samples) and so the distribution under the null hypothesis shares the same degrees of freedom, $G-p-1$ for a model with p predictor variables. However, this also means that Pearson's chi-square statistic has the same shortcomings. It relies on the number of observations

for each group being large enough when the response is not normally distributed, and may lead to the wrong conclusion if this is not the case.

Hosmer-Lemeshow Test

The Hosmer-Lemeshow (HL) test is a modification of the Pearson chi-square test that can be applied when the number of observations for some combinations of the predictors are small, including when the model includes a continuous covariate. Essentially, the HL test works by combining the data into larger groups and then computing the Pearson chi-square test statistic. In fact, the expression for the HL test statistic is exactly as above with G representing the number of combined groups and the test statistic is distributed approximately according to a chi-square distribution with $G - p - 1$ degrees of freedom, if the model fits the data.

However, the HL test still presents a problem in that you need to decide on the number of groups. On one hand, you want to choose a small number of groups so that the number of observations per group is as high as possible. This ensures that the distribution of the mean in each group will be closer to the approximate normal distribution and hence that the distribution of the overall test statistic will be closer to the chi-square distribution. On the other hand, you also want the range of the covariate covered by each group to be small, in order that the predicted values are similar to each other, which means that you want more groups with less observations each. This presents a problem and there is no real solution. Changing the number of groups will change the value of the test statistic, and the p-value, and may lead to different conclusions regarding the fit of the model.

le Cessie - van Houwelingen - Copas - Hosmer

There are more modern tests which avoid these problems, and the one that we will focus on is the le Cessie - van Houwelingen - Copas - Hosmer. This test does not require the user to divide the data into groups, but still results in a test statistic with a valid chi-square distribution with (always) 1 degree of freedom – also equivalent to a two-sided z-test. Essentially, this test works by applying the Pearson chi-square test within a sliding window that covers the data. I won't go into details, but you can find them in Hosmer, Cessie, and Lemeshow (1997).

Example

I will consider a simulated data set as an example so that we know the truth exactly. The code to run the simulation is in the file `hl_example.R`. Briefly, the code generates 500 observations from the logistic regression model

$$Y_i \sim \text{Bernoulli}(p_i)$$

with

$$\text{logit}(p_i) = -1.0 + 0.2x_i$$

where each x_i is generated from a uniform distribution over $(0, 10)$. The true probabilities vary between .27 and .73. If we fit the model to the data then we find that the estimated parameters are -1.03 (95%CI = -1.40, .67) and .22 (95%CI= .15, .28) which is very reasonable. However, the residual deviance is 647.74 on 498 degrees of freedom which provides a p-value $< .0001$ (6.54×10^{-6} to be exact) which would lead us to reject the fit of the model and conclude that we need a more complicated model to fit the data. Clearly this isn't true because we know that the data fit the model. We generated them!

We can also compute the Pearson goodness-of-fit test statsitic. In this case, the value of the test statistics is $X^2 = 500.17$ also on 498 degrees of freedom. The p-value associated with this test would be .54 which does not provide evidence to reject the fit of the model. This is the right conclusion, but we know that this p-value isn't reliable because the asymptotics required to approximate the distribution of the test statistic are not valid.

Instead, we can conduct the Hosmer-Lemeshow test using the aptly named function `HosmerLemeshowTest()` from the `DescTools` package. This test actually performs two variants of the HL test, called C and H. I will focus on test C which is the test defined above. In my experience the results of the two tests are similar. By default this function divides the data into 10 groups so that the test statistic has $10 - 2 = 8$ degrees of freedom. The value of the test statistic is 4.03 and the resulting p-value is .854 leading us to conclude, correctly, that there is no evidence to reject the fit of the model. Note that changing the number of groups (`ngr`) does affect the value of the test statistic and the degrees of freedom, but not the overall conclusion in this case. For example, with `ngr=100` the test statistic becomes 103.06 on 98 degrees of freedom so that the p-value is .34. This is smaller, but still provides no reason to question the fit of the model (which is good).

Finally, the le Cessie - van Houwelingen - Copas - Hosmer test can also be compute by the `HosmerLemeshowTest()` function. To do this, you need to set the argument `X` to be the matrix of all covariates in the model. Doing this for the example data, we find that the p-value is .9967, indicating that there is absolutely no reason at all to doubt the fit of the model to the data. In fact, the model fits the data almost surprisingly well. Note that changing the number of groups does not affect the results of this test.

References

Hosmer, T, S Le Cessie, and S Lemeshow. 1997. "A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model." *Statistics in Medicine* 16 (9): 965–80.