Statistics
Logistic regression

Maarten Speekenbrink *UCL Cognitive, Perceptual and Brain Sciences*
November, 2012

Logistic regression is a technique to analyse dichotomous dependent variables. For example, the dependent variable can take on the value "yes" and "no", "corect" and "incorrect", "married" and "not-married", etc. It is customary to code for these categories using the values 0 (e.g.,"correct") and 1 (e.g., "incorrect"). These notes are meant to give you some more background information on logistic regression.

1 The logit transform

The main problem with dichotomous variables is that if we use a linear regression model such as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_{p-1} X_{p-1,i} + \epsilon_i$$

then the model predictions $\hat{Y}_i$ can take "impossible" values (e.g., values which are not 0 or 1). One thing we can (and will) do, is to predict the *probability* that Y_i takes a particular value, say the value 1. We will denote this probability as $\pi_i = P(Y_i = 1)$. Using a linear model for this "transformed" dependent variable

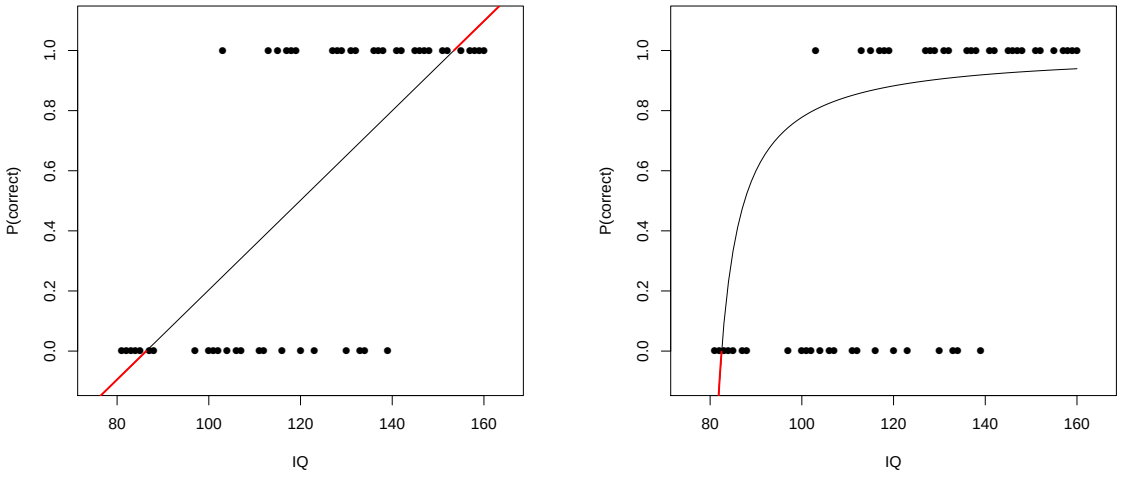
$$\pi_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_{p-1} X_{p-1,i} + \epsilon_i$$

solves part of the problem, as predictions between 0 and 1 are now allowed. However, as shown in Figure 1(a), we can still get predictions that are smaller than 0, or larger than 1, and these are not probabilities. The main trick in logistic regression is to transform the dependent variable – which has now become π_i – so that it can take any value between $-\infty$ (minus infinity) and ∞ (infinity). The transformation that accomplishes this is called the logit transformation. The first step is to divide the dependent variable π_i by its complement, which is $P(Y_i = 0) = 1 - P(Y_i = 1) = 1 - \pi_i$. This gives us the "odds" that the dependent variable is 1 rather than 0:

$$\frac{P(Y_i = 1)}{P(Y_i = 0)} = \frac{\pi_i}{1 - \pi_i}$$

The odds can be any value between 0 and ∞. If it is 1, then $P(Y = 1) = P(Y = 0)$ (i.e., the probability that Y is 1 is identical to the probability that Y is 0). A value smaller than 1 means that$P(Y = 0) > P(Y = 1)$, while a value greater than 1 means that $P(Y = 1) > P(Y = 0)$.

Transforming to odds doesn't solve all the problems, as the odds can't be smaller than 0, while predictions of the linear model can. Also, if we transform the predicted odds back to probabilities, as in Figure 1(b), we see that we can still predict probabilities less than 0



(a) $P(Y_i = 1)$ (b) $\frac{P(Y_i=1)}{P(Y_i=0)}$

Figure 1: A linear regression model for the probabilities (A) and for the odds (B). Impossible predictions indicated in red.

(when the predicted odds are less than 0). These problems are solved by taking the (natural) logarithm of the odds:

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

When the log-odds is 0, this means that $P(Y_i = 1) = P(Y_i = 0)$. A *negative* log-odds means that $P(Y_i = 1) < P(Y_i = 0)$. A *postive* log-odds means that $P(Y_i = 1) > P(Y_i = 0)$.

We can now set up a linear model to predict the log-odds as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_{p-1} X_{p-1,i} + \epsilon_i \tag{1}$$

So after transforming the dependent variable (from a binary variable, to a probability, to an odds, to a log-odds), the model looks like a normal GLM. The main trick was to transform the dependent variable. Predicted log-odds are not as natural to interpret as predicted probabilities. To get the predicted probabilities, we can transform log-odds back to probabilities. For this, you need the "inverse" of the logit transformation, which is known as the logistic transform:

$$\pi_i = \frac{e^{\beta_0+\beta_1 X_{1i}+\ldots+\beta_{p-1}X_{p-1,i}} + \epsilon_i}{1 + e^{\beta_0+\beta_1 X_{1i}+\ldots+\beta_{p-1}X_{p-1,i}} + \epsilon_i} \tag{2}$$

(if you want to see how to derive this "inverse transformation", you can look in Appendix A). If we have estimates b_j of the population parameters β_j, we can compute the predicted probabilities as

$$\widehat{\pi}_i = \frac{e^{b_0+b_1 X_{1i}+\ldots+b_{p-1}X_{p-1,i}}}{1 + e^{b_0+b_1 X_{1i}+\ldots+b_{p-1}X_{p-1,i}}}$$

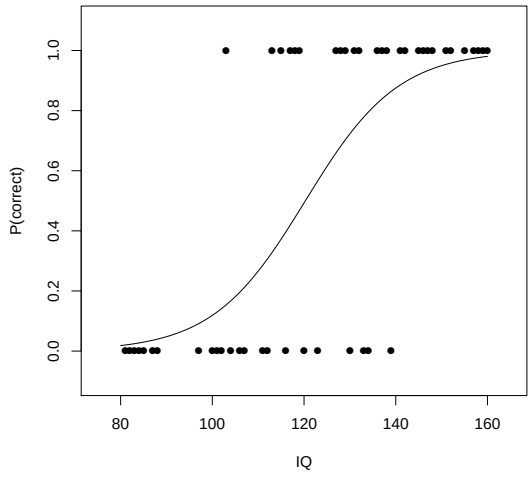
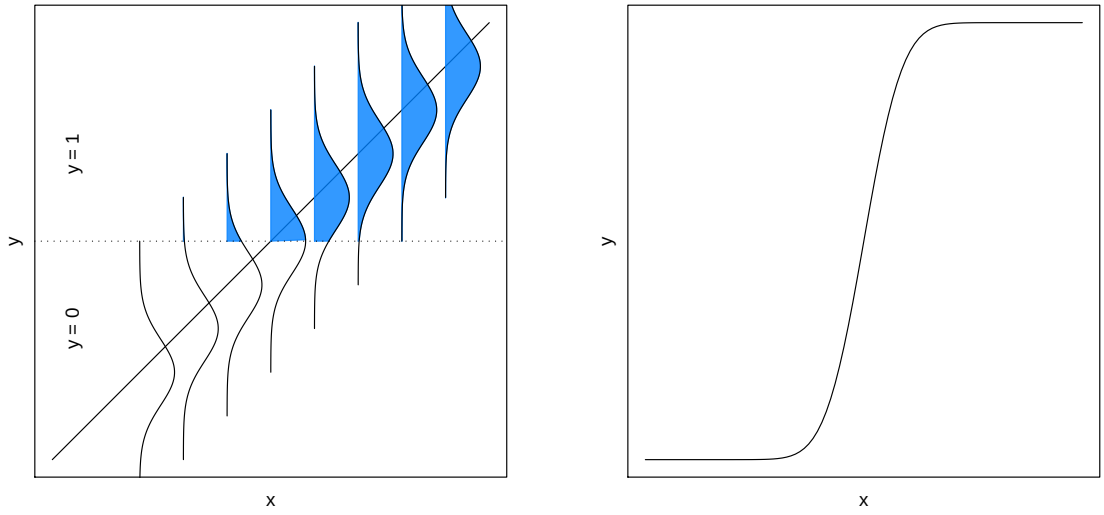


Figure 2: A linear regression model for the log odds $\log\left(\frac{P(Y_i=1)}{P(Y_i=1)}\right)$.

The resulting model predictions, for a model with a single predictor, are depicted in Figure 2.

This heuristic justification of the logistic regression model may not seem entirely convincing. All we did was transform a probability so that we could make predictions between minus and plus infinity. Another justification is in terms of a latent variable model. Assume that underlying an observed dichotomous variable is an unobserved (i.e., latent) quantity. For instance, it seems sensible to assume there is a continuous range of emotions from unhappy to happy. If we ask someone whether (s)he is happy or unhappy, (s)he will have to place a cutoff point on this continuum, such that emotions "above" it are classified as "happy", and emotions "below" it, are classified as "unhappy". A predictor of happiness, let's just call it x for now, is essentially a predictor of this underlying latent "happiness" variable. Figure 3(a) represents a linear regression model predicting the latent variable y from x. As with any regression model, we assume a normally distributed error round the linear regression line. For a given value of x, the probability that someone will respond with "happy" is given by the area of this normal distribution which is above the cutoff value (these areas are represented in blue). When we plot these probabilities as a function of x, as in Figure 3(b), we get the S-shaped curve of the logistic function. Actually, if we assume a Normal distribution for the model error, the curve is slightly different from the logistic curve (it is called a "normal ogive" if you really want to know), but the logistic curve is a very close match and is much easier to work with. If we would assume the errors follow a logistic distribution (which itself looks a lot like a Normal distribution), then the curve in Figure 3(b) would be the logistic function.



(a) Regression model of the latent variable. Values of Y above the cutoff value are classified as $Y = 1$, values below as $Y = 0$.

(b) $P(Y_i = 1)$ as a function of x.

Figure 3: Latent variable representation of logistic regression

2 Interpretation of the model parameters

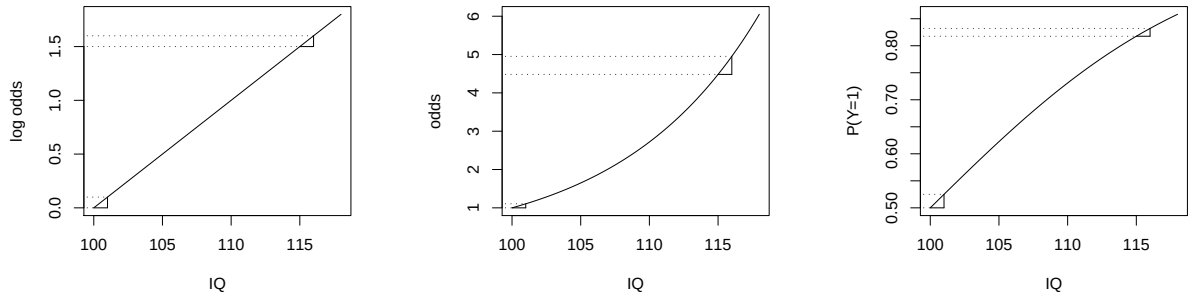
Interpretation of the parameters is the same as for the GLM. So the intercept β_0 reflects the value of the dependent variable when all predictors have the value 0 (all $X_j = 0$). The slopes β_j represent the increase in the dependent variable when the predictor X_j increases by 1 unit. But you have to remember that the dependent variable in logistic regression is the log-odds. As most people don't think in terms of log-odds, it will be helpful to assess the effect of the predictors on the odds, or on the probabilities. Take for example a model to predict the probability of answering a question correctly ($Y_i = 1$) from a person's IQ score. The estimated model is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -10 + 0.1 \times \texttt{IQ}_i + e_i$$

In the table below I've collected the model predictions for four IQ scores:

IQ	$\log\left(\frac{P(Y=1)}{P(Y=0)}\right)$	$\frac{P(Y=1)}{P(Y=0)}$	$P(Y=1)$
100	0	1	0.500
101	0.1	1.105	0.525
115	1.5	4.482	0.818
116	1.6	4.953	0.832

As the model is linear for the log-odds, the effect on the log-odds of a 1 point increase in IQ is $b_1 = 0.1$, whether we increase the IQ from 100 to 101, or from 115 to 116. But as the



(a) log odds (b) odds (c) probability

Figure 4: Effect of a 1 unit increase in IQ on the log odds, odds, and probability of giving a correct answer to a question

model is nonlinear in terms of the odds, the effect of a 1 point increase on the odds depends on where you start. As you can see in Figure 4(b), a 1 point increase has a larger effect on the odds when you start at 115 than when you start at 100. This is because the effect is actually multiplicative: for every unit increase in X_j, you have to multiply the odds by e^{β_j}. So, for the example here, every 1 point increase in IQ multiplies the odds by $e^{0.1} = 1.105$. Someone with an IQ of 100 has equal odds of answering the question correctly or incorrectly, i.e., the odds are $\frac{1}{1}$. But for someone with an IQ of 101, the odds are $1.105 \times \frac{1}{1} = 1.105$. For someone with an IQ of 115, the odds are $(1.105)^{15} \times \frac{1}{1} = 4.481$, or starting from 101, they are $(1.105)^{14} \times \frac{1.105}{1} = 4.481$. Because of this multiplicative property, the exponentiated slope, e^{b_j}, is also interpreted as the odds ratio for those who are 1 point apart on the predictor. This ratio is the same for any two people who differ by 1 IQ point. For example, using the values in the table, you can see that

$$\frac{1.105}{1} = \frac{4.953}{4.482} = 1.105$$

As for the odds, the model is also nonlinear in terms of probabilities. As shown in Figure 4(c), a 1 point increase has a larger effect on the probability when you start from 100 than when you start from 115. There is no easy way to convert the slope to a change in probability; you just have to fill in the values in the model. So, for a change from an IQ of 100 to an IQ of 101, the increase in probability is

$$\frac{\exp(-10 + .1 \times 101)}{1 + \exp(-10 + .1 \times 101)} - \frac{\exp(-10 + .1 \times 100)}{1 + \exp(-10 + .1 \times 100)} = .525 - .500 = .025$$

and you would have to work this out for each starting value of IQ.

3 Log-likelihoods and error

The model specified in Equation 1 looks a lot like the GLM. But for the logistic regression model, we should not rely on the Sum of Squared Errors (SSE) measure of error for comparing different models. The error measure of choice in logistic regression is "-2 log-likelihood". To understand what this is, let's take the predictions of Equation 2, which are probabilities. Now, if the model does well, these probabilities will correspond nicely with the observed values. For instance, let's take the simple model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \epsilon_i$$

so that

$$\widehat{\pi}_i = \frac{e^{b_0}}{1+e^{b_0}}$$

As for all models with only an intercept, we thus assume that $P(Y = 1)$ is some constant. Let's say we have observed a sample of 20 values of Y, and that out of those 20, there were 15 cases where $Y_i = 1$. Using this sample, we can estimate the probability as

$$\widehat{\pi}_i = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{15}{20} = .75$$

which will give an estimated intercept of $b_0 = \log(.75/.25) = 1.099$. So a sensible prediction for each observation is $\widehat{\pi}_i = \frac{e^{1.099}}{1+e^{1.099}} = .75$. Of course, using this prediction for the actual values of Y_i will mean that we'll always be wrong (as Y_i can only be 0 or 1, and nothing in between, unlike the *probability* $P(Y_i = 1)\ldots$). But, while incorrect for any particular observation Y_i, it is the best (constant) prediction overall. For example, it has an SSE of $\sum_{i=1}^{n}(Y_i - \widehat{\pi}_i)^2 = 15 \times (1-.75)^2 + 5 \times (0-.75)^2 = 3.75$, and taking any other value $\widehat{\pi}_i \neq .75$ will only increase the SSE. But while this is the case, I've indicated that the SSE is not a very useful error measure in logistic regression. Rather, we look at how well the predicted probabilities match the observations Y_i. Recall that $\widehat{\pi}_i = \widehat{P}(Y_i = 1)$, so the model predicts that each observation $Y_i = 1$ has a probability of $\widehat{\pi}_i = .75$, and each observation $Y_i = 0$ has a probability of $1 - \widehat{\pi}_i = .25$. These predicted probabilities are also called likelihoods; the are the probability of observations assigned by the model. Assuming independent observations, the likelihood of the whole sample of observations is

$$P(Y_1, Y_2, \ldots, Y_n) = \prod_{i=1}^{n} P(Y_i) = \widehat{\pi}_i^{15} \times (1-\widehat{\pi}_i)^5 = .000001305$$

As you can imagine, for larger samples, this value will become smaller and smaller (if you keep multiplying values between 0 and 1, the result will approach 0). It is therefore handier to work with the (natural) logarithm of the likelihood

$$\log P(Y_1, Y_2, \ldots, Y_n) = \sum_{i=1}^{n} \log P(Y_i)$$

As you can see, using a log transform turns the product into a sum of individual logs. For the example, the log likelihood of the sample data is $15 \times \log \widehat{\pi}_i + 5 \times \log(1 - \widehat{\pi}_i) = -11.247$.

The reason for taking the *negative* log-likelihood is that we want to make it a measure of prediction *error*. If we don't do this, higher values indicate better predictions. In other words, the log likelihood is a measure of model *fit*, rather than model *error*. By taking the negative log-likelihood, high values mean poor predictions, just as for the SSE. For technical reasons, we need to multiply the negative log-likelihood by 2 (this helps later on when we use the measure for model comparison). Our final resulting measure of error is thus

$$\begin{aligned} -2 \log L &= -2 \times \sum_{i=1}^{n} \log \widehat{P}(Y_i) \\ &= -2 \times \sum_{i=1}^{n} \log\{Y_i \widehat{\pi}_i + (1 - Y_i)(1 - \pi_i)\} \end{aligned}$$

The last line above is just a way to compute $\widehat{P}(Y_i)$, which is $\widehat{\pi}_i = \widehat{P}(Y_i = 1)$ if $Y_i = 1$ and $(1 - \widehat{\pi}_i) = \widehat{P}(Y_i = 0)$ if $Y_i = 0$. So we take the log of these predicted probabilities, sum them, and multiply the resulting value by -2.

Just as for linear regression, the model parameters β_j are defined as those that minimize the error in the population. We obtain estimates b_j of the parameters by minimizing the error in the sample. In this case, minimizing the error is the same as maximising the likelihood. Therefore, the estimation procedure in logistic regression is also called maximum-likelihood estimation. This is to be contrasted to Ordinary Least Squares (OLS) estimation used in linear regression.

4 Model comparison and hypothesis testing

Now we've got our measure of error, we can turn to model comparison. Remember the equation for the Proportional Reduction in Error (PRE) for the GLM:

$$\text{PRE} = \frac{\text{SSE(C)} - \text{SSE(A)}}{\text{SSE(C)}}$$

We're going to use something pretty similar here, but now using the (negative) log-likelihood measure of error:

$$\text{PRE} = \frac{2 \log L(C) - 2 \log L(A)}{2 \log L(C)} \tag{3}$$

Note that this is identical to $\frac{-2 \log L(C) - (-2 \log L(A))}{-2 \log L(C)}$; I've just multiplied the numerator and denominator by -1. When MODEL C has only an intercept, this measure is also known as McFadden's pseudo-R^2. Remember that it was useful to transform the PRE to an F value to test the significance of the PRE. For logistic regression, we also need to use a transformation of the PRE, namely $\text{PRE} \times -2 \log L(C)$. After a bit of algebra, this is equivalent to

$$\chi^2 = -2 \log L(C) - (-2 \log L(A)) \tag{4}$$

as our test statistic. This is the same statistic used in Mixed Effects models. The sampling distribution of this statistic is (approximately) chi-square. The chi-square distribution has one-parameter, the degrees of freedom, which in this case is df = PA − PC, e.g. the difference in the number of parameters between MODEL A and MODEL C (so the degrees of freedom is identical to df_1 for the F statistic).

Going back to our example, let's assume that the observed variable Y_i represents a coin toss, with $Y_i = 1$ denoting "heads", and $Y_i = 0$ denoting "tails". Someone believes the coin has been meddled with to come up heads more often than tails. To test whether this is indeed the case, we can compare MODEL C:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 0 + \epsilon_i$$

to MODEL A:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \epsilon_i$$

Note that by using an intercept of 0, MODEL C assumes heads and tails are equally likely. For MODEL C, we have $\log L(C) = 20 \times \log .5 = -13.863$, and for MODEL A we had $\log L(A) = -11.247$. So the test statistic is

$$\begin{aligned}\chi^2 &= -2 \times -13.863 - (-2 \times 11.247)) \\ &= 5.232\end{aligned}$$

and $P(\chi_1^2 \geq 5.232) = .022$, so we can reject MODEL C in favour of MODEL A and conclude that the coin is indeed biased.

Just as for the GLM, we can extend this procedure to models with more parameters than just the intercept. We can test the slope for each predictor by comparing a model in which we fix the slope to 0 to a model in which we estimate it, which is equivalent to comparing a model with the predictor included to one with just that predictor removed. We can also perform omnibus tests in which we compare a model in which we fix a set of slopes to 0 to a model in which we estimate the slopes. This model comparison procedure is the same procedure as for the GLM, the only difference is that we use a different statistic.

5 Model evaluation and fit

5.1 Coefficient of determination

To assess whether a logistic regression model gives accurate predictions of the data, we can first look at a measure like R^2. Recall that in linear regression, R^2 is interpreted as the proportion of the variance of the dependent variable Y that is accounted for by the model. There are several measures like the R^2 for logistic regression. They are "pseudo"-R^2 measures: while they are usually interpreted in terms of "proportion variance explained",

this is technically not entirely accurate. In any case, the first pseudo-R^2 measure uses our definition of the PRE in Equation 3. It is called McFadden's pseudo-R^2 when we compute this PRE for a MODEL C with only an intercept and a MODEL A with any number of predictors. Like the R^2 for linear regression, McFadden's R^2 measure takes values between 0 and 1, where higher values are better. However, the values of McFadden's R^2 are usually a lot lower than the values of the linear regression R^2, and values between .2 and .4 are usually taken to be highly satisfactory. Cox and Snell have proposed an alternative pseudo-R^2 which more closely follows the values of the linear regression R^2. It is defined as

$$\begin{aligned}\text{pseudo-}R^2(\text{Cox \& Snell}) &= 1 - \exp\left(-\frac{2}{n}(\log L(A) - \log L(C))\right) \\ &= 1 - \left(\frac{L(C)}{L(A)}\right)^{2/n}\end{aligned}$$

but a problem with this measure is that its maximum is less than 1, so that a perfectly predictive model can have an R^2 less than 1. Nagelkerke showed that the maximum value of the Cox & Snell measure is

$$1 - \exp\left(\frac{2}{N}\log L(C)\right)$$

and proposed a simple correction of the Cox & Snell R^2 which does have a maximum value of 1. The Nagelkerke pseudo-R^2 measure is simply the Cox & Snell R^2 divided by this maximum.

5.2 Testing model fit

In logistic regression, it is customary to test the overall fit of a model. This is something that we haven't done for the General Linear Model. Testing whether the model fits the data can tell you something about whether you have included enough predictors in the model (a non significant test result), or whether there is substantial variation in the data that is left unexplained (a significant test result). The general idea is to compare the model of interest to "the most complex model, which has a seperate parameter at each explanatory setting" (Agresti, 1996, p.96). This complex model is the so-called *saturated* model. There are actually different ways in which to specify the saturated model. The first one is to treat each observation as a basic unit (or "explanatory setting"). In this case, we use a seperate parameter for each observation, so that the saturated model has as many parameters as observations. This saturated model will fit the data perfectly. For an observation $Y_i = 1$, it would estimate $\widehat{\pi}_i = 1$, and for an observation $Y_i = 0$, it would estimate $\widehat{\pi}_i = 0$. This is the best possible model to predict the observed data, but it is unlikely to generalize to new data (remember the example of a polynomial regression model that fitted the data perfectly, but made rather strange predictions for not-yet observed data). Fitting the data perfectly, it has a log-likelihood value of $\sum_{i=1}^{n} \log 1 = 0$. The test to compare a model (which we'll call MODEL C, as it is more restricted than the saturated model) to this saturated model

is then

$$\begin{aligned}\chi^2 &= -2\log L(C) - (-2\log L(\text{saturated})) \\ &= -2\log L(C)\end{aligned}$$

with $\text{df} = n - \text{PC}$ degrees of freedom. This is quite an easy approach, but unfortunately, the distribution of the statistic is not well approximated by a chi-square distribution, so that the test results are likely to be biased. A second approach is to use unique predictor patterns as the basic unit. This will give different results when there are multiple observations with the same values for the predictors. The saturated model then has as many parameters as unique predictor patterns and $\log L(\text{saturated}) \leq 0$ (it will be 0 only when each observation has a different predictor pattern). This method may give better results than the first one, especially if there are many replicated predictor patterns.

A more heuristic approach to testing whether the model fits the data was proposed by Hosmer and Lemeshow (1989). They propose to group the data into a relatively small number of equally sized groups, usually $g = 10$ (g denotes the number of groups). Data are grouped according to the predicted probabilities $\widehat{\pi}_i$. Then, for each of these groups, the test compares the expected frequency of $Y_i = 1$ observations (the number of 1's as predicted by the model) to the observed frequency of $Y_i = 1$ observations, using a Chi-Square test with $\text{df} = g - 2$ degrees of freedom. The Hosmer and Lemeshow test seems to be the one of choice for many people, and is routinely produced by SPSS.

As we haven't discussed the Chi-Square test (although we have used the chi-square distribution), I'll briefly show here how the test statistic is computed. Say that we have a total of 100 observations. The first thing is to order these in terms of the predicted probabilities $\widehat{\pi}_i$. Let's make the example really easy, and assume the predicted probabilities are equally spaced as $0.01, 0.02, 0.03, \ldots, 1.00$. So in group 1, we would have the 10 predicted probabilities $0.01, 0.02, \ldots, 0.10$. The expected frequency of $Y_i = 1$ in this group is simply the sum of these probabilities. If we denote the expected frequency in group j as E_j, we thus have $E_1 = .01 + .02 + \ldots + .10 = 0.55$. The second group contains the 10 predicted probabilities $0.11, 0.12, \ldots, 0.20$, and the expected frequency is $E_2 = 1.55$. We can do the same for each of the 10 groups (of course, in reality, the predicted probabilities wouldn't be so nicely spaced as in this example). The resulting expected frequencies are given in Table 1.

Suppose the observed frequencies (the actual number of $Y = 1$ observations) in each group are as given in Table 1. The main idea is then to compare the predicted and expected frequencies to see if they are not too far apart. The error measure used is like an SSE, but each square is divided by the expected frequency. The final test statistic is then

$$\chi^2 = \sum_{j=1}^{g} \frac{(O_j - E_j)^2}{E_j} \tag{5}$$

which here results in

$$\chi^2 = \frac{(2 - 0.55)^2}{0.55} + \frac{(2 - 1.55)^2}{1.55} + \ldots + \frac{(9 - 9.55)^2}{9.55} = 8.15$$

Table 1: Hosmer-Lemeshow test

g	$\widehat{\pi}_i$	E_g	F_g	$\frac{(F_g-E_g)^2}{E_g}$
1	$0.01, 0.02, \ldots, 0.10$	0.55	2	3.82
2	$0.11, 0.12, \ldots, 0.20$	1.55	2	0.13
3	$0.21, 0.22, \ldots, 0.20$	2.55	4	0.82
4	$0.31, 0.32, \ldots, 0.30$	3.55	5	0.59
5	$0.41, 0.42, \ldots, 0.40$	4.55	8	2.62
6	$0.51, 0.52, \ldots, 0.50$	5.55	6	0.04
7	$0.61, 0.62, \ldots, 0.60$	6.55	6	0.05
8	$0.71, 0.72, \ldots, 0.70$	7.55	8	0.03
9	$0.81, 0.82, \ldots, 0.80$	8.55	9	0.02
10	$0.91, 0.92, \ldots, 1.00$	9.55	9	0.03
Σ		50.5	59	8.15

For this value and $g - 2 = 8$ degrees of freedom, $P(\chi^2_8 \geq 8.15) = 0.419$, which is not significant, and we can conclude that the model did a pretty good job.

This chi-square test to compare expected and observed frequencies is applicable to many other situations than that treated by Hosmer and Lemeshow. It is usually called the Pearson Chi-Square goodness of fit test.

References

Agresti, A. (1996). *An introduction to categorical data analysis.* John Wiley, New York.

Hosmer, D. and Lemeshow, S. (1989). *Applied logistic regression.* John Wiley, New York.

A The logit and logistic transformation

To show that the logistic function is indeed the inverse of the logit function, you need to solve the logit function for π. A little algebra shows that:

$$\begin{aligned}
\log\left(\frac{\pi}{1-\pi}\right) &= V \\
\left(\frac{\pi}{1-\pi}\right) &= e^V && \text{(by taking exponent on both sides)} \\
\pi &= e^V(1-\pi_i) && \text{(by multiplying both sides by } (1-\pi)) \\
\pi &= e^V - \pi e^V \\
1 &= \frac{e^V}{\pi} - e^V && \text{(by dividing both sides by } \pi) \\
1 + e^V &= \frac{e^V}{\pi} && \text{(by adding } e^V \text{ to both sides)} \\
\pi(1+e^V) &= e^V && \text{(by multiplying both sides by } \pi) \\
\pi &= \frac{e^V}{1+e^V} && \text{(by dividing both sides by } 1+e^V)
\end{aligned}$$