Chapter 6

Logistic Regression

In logistic regression, there is a categorical dependent variables, often coded 1=Yes and 0=No. Many important phenomena fit this framework. The patient survives the operation, or does not. The accused is convicted, or is not. The customer makes a purchase, or does not. The marriage lasts at least five years, or does not. The student graduates, or does not.

As usual, we assume that there is a huge population, with a sizable sub-population at each x value or configuration of x values. And as in ordinary regression, we want a regression surface that consists of the estimated sub-population mean (conditional expected value) at each x value or configuration of x values. It turns out that for any dependent variable coded zero or one, this conditional mean is exactly the conditional *probability* that $Y = 1$ given that set of x values. Again, for binary data, the population mean is just the probability of getting a one. And since it's a probability, it must lie between zero and one inclusive.

Consider the scatterplot of a single quantitative independent variable and a dependent variable Y equal to zero or one. The left panel of Figure 6.1 shows what happens when we fit a least squares line to such data. It may be reasonable in some sense, but because it is sometimes less than zero and sometimes greater than one, it can't be a probability and it's *not* yielding a sensible estimate of the conditional population mean. However, the logistic regression curve in the right panel stays nicely between zero and one. And like the least-squares line, it indicates a positive relationship for this particular data set.

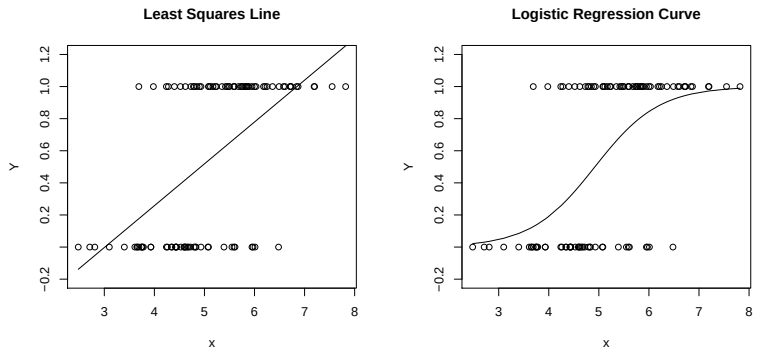
6.1 A linear model for the log odds

The logistic regression curve arises from an indirect representation of the probability of $Y = 1$ for a given set of x values. Representing the probability of an event by π (it's a probability, not 3.14159...), we define the *odds* of the event as

$$\text{Odds} = \frac{\pi}{1-\pi}.$$

Implicitly, we are saying the odds are $\frac{\pi}{1-\pi}$ "to one." That is, if the probability of the event is $\pi = 2/3$, then the odds are $\frac{2/3}{1/3} = 2$, or two to one. Instead of saying the odds are 5 to

Figure 6.1: Scatterplots with a binary dependent variable

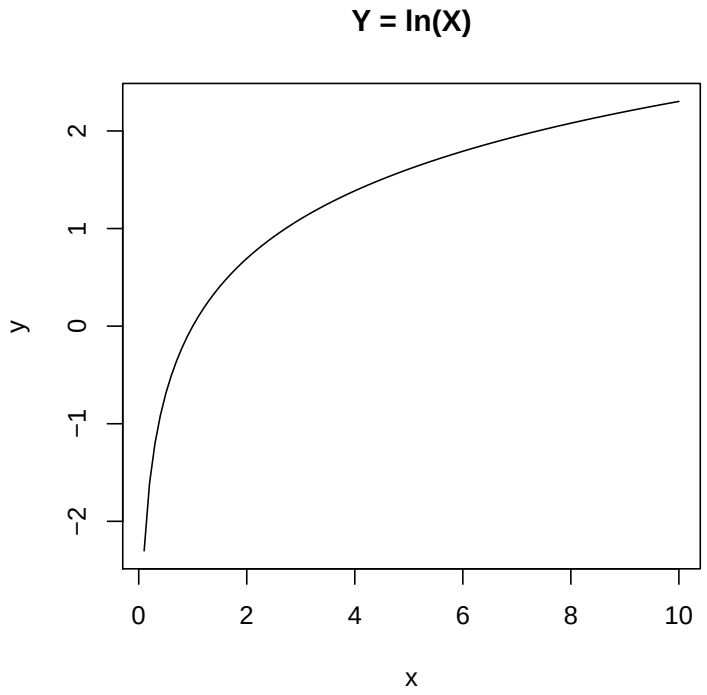


2, we'd say 2.5 to one. Instead of saying 1 to four, we'd say 0.25 to one.

The higher the probability, the greater the odds. And as the probability of an event approaches one, the denominator of the odds approaches zero. This means the odds can be anything from zero to an arbitrarily large positive number. Logistic regression adopts a regression-like linear model not for the probability of the event $Y = 1$ nor for the odds, but for the *log odds*. By log we mean the natural or Napierian log, designated by ln on scientific calculators – not the common log base 10. Here are a few necessary facts about the natural log function.

- Figure 6.2 shows that the natural log increases from minus infinity when the odds are zero, to zero when the odds equal one (fifty-fifty), and then it keeps on increasing as the odds rise, but more and more slowly.
- The fact that the log function is increasing means that if $P(A) > P(B)$, then $\text{Odds}(A) > \text{Odds}(B)$, and therefore $\ln(\text{Odds}(A)) > \ln(\text{Odds}(B))$. That is, the bigger the probability, the bigger the log odds.
- Notice that the natural log is only defined for positive numbers. This is usually fine, because odds are always positive or zero. But if the odds are zero, then the natural log is either minus infinity or undefined – so the methods we are developing here will not work for events of probability exactly zero or exactly one. What's wrong with a probability of one? You'd be dividing by zero when you calculated the odds.
- The natural log is the inverse of exponentiation, meaning that $\ln(e^x) = e^{\ln(x)} = x$, where e is the magic non-repeating decimal number 2.71828.... The number e really is magical, appearing in such seemingly diverse places as the mathematical theory of epidemics, the theory of compound interest, and the normal distribution.
- The log of a product is the sum of logs: $\ln(ab) = \ln(a) + \ln(b)$, and $\ln(\frac{a}{b}) = \ln(a) - \ln(b)$. This means the log of an odds *ratio* is the difference between the two log odds quantities.

Figure 6.2: Graph of the natural log function



To get back to the main point, we adopt a linear regression model for the log odds of the event $Y = 1$. As in normal regression, there is a conditional distribution of the dependent variable Y for every configuration of independent variable values. Keeping the notation consistent with ordinary regression, we have $p - 1$ independent variables, and the conditional distribution of the binary dependent variable Y is completely specified by the log odds

$$\ln\left(\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1}x_{p-1}. \tag{6.1}$$

This is equivalent to a *multiplicitive* model for the odds

$$\begin{aligned}\left(\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})}\right) &= e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}} \\ &= e^{\beta_0}e^{\beta_1 x_1}\cdots e^{\beta_{p-1}x_{p-1}},\end{aligned} \tag{6.2}$$

and to a distinctly non-linear model for the conditionl probability of $Y = 1$ given $\mathbf{X} = (x_1, \ldots, x_{p-1})$:

$$P(Y=1|x_1,\ldots,x_{p-1}) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}. \tag{6.3}$$

6.2 The meaning of the regression coefficients

In the log odds world, the interpretation of regression coefficients is similar to what we have seen in ordinary regression. β_0 is the intercept. It's the log odds of $Y = 1$ when all independent variables equal zero. And β_k is the increase in log odds of $Y = 1$ when x_k is increased by one unit, and all other independent variables are held constant.

This is on the scale of log odds. But frequently, people choose to think in terms of plain old odds rather than log odds. The rest of this section is an explanation of the following statement: *When x_k is increased by one unit, and all other independent variables are held constant, the odds of $Y = 1$ are multiplied by e^{β_k}.* That is, e^{β_k} is an **odds ratio** — the ratio of the odds of $Y = 1$ when x_k is increased by one unit, to the odds of $Y = 1$ when x_k is left alone. As in ordinary regression, this idea of holding all the other variables constant is what we mean when we speak of "controlling" for them.

Odds ratio with a single dummy variable Here is statement that makes sense and seems like it should be approximately true: "Among 50 year old men, the odds of being dead before age 60 are three times as great for smokers." We are talking about an odds ratio.

$$\frac{\text{Odds of death given smoker}}{\text{Odds of death given nonsmoker}} = 3$$

The point is not that the true odds ratio is exactly 3. The point is that this is a reasonable way to express how the chances of being alive might depend on whether you smoke cigarettes.

Now represent smoking status by an indicator dummy variable, with $X = 1$ meaning Smoker, and $X = 0$ meaning nonsmoker; let $Y = 1$ mean death within 10 years and $Y = 0$ mean life. The logistic regression model (6.1) for the log odds of death given x are

$$\text{Log odds} = \beta_0 + \beta_1 x,$$

and from (6.2), the odds of death given x are

$$\text{Odds} = e^{\beta_0} e^{\beta_1 x}.$$

The table below shows the odds of death for smokers and non-smokers.

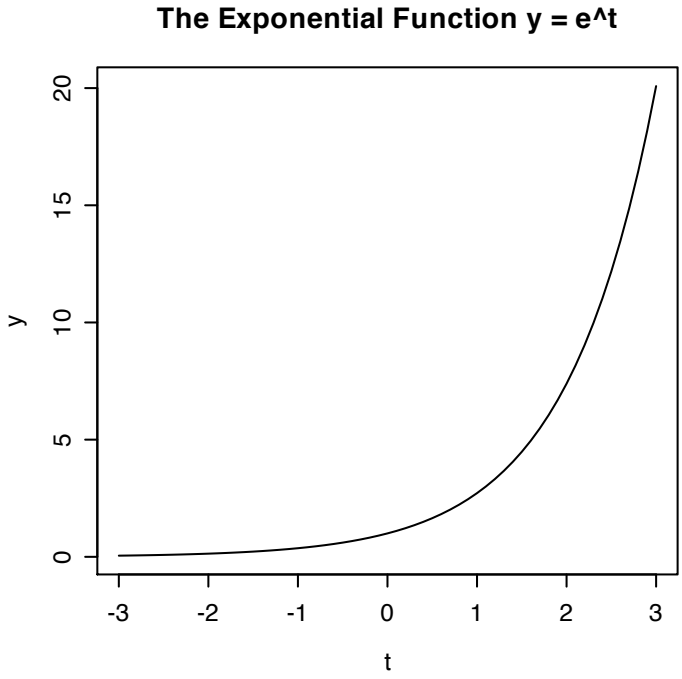
Group	x	**Odds of Death**
Smokers	1	$e^{\beta_0} e^{\beta_1}$
Non-smokers	0	e^{β_0}

Now it's easy to see that the odds ratio is

$$\frac{\text{Odds of death given smoker}}{\text{Odds of death given nonsmoker}} = \frac{e^{\beta_0} e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1}.$$

Our understanding of the regression coefficient β_1 follows from several properties of the function $f(t) = e^t$.

- e^t is always positive. This is good because odds are non-negative, but the fact that e^t is never zero reminds us that the logistic regression model cannot accomodate events of probability zero or one.
- $e^0 = 1$. So when $\beta_1 = 0$, the odds ratio is one. That is, the odds of $Y = 1$ (and hence the probability that $Y = 1$) are the same when $X = 0$ and $X = 1$. That is, the conditional distribution of Y is identical for both values of X, meaning that X and Y are unrelated.
- $f(t) = e^t$ is an increasing function. So, when β_1 is negative, $e^{\beta_1} < 1$. Therefore, the probability of $Y = 1$ would be *less* when $X = 1$. But if β_1 is positivethen the odds ratio is greater than one, and the probability of $Y = 1$ would be greater when $X = 1$, as in our example. In this sense, the sign of β_1 tells us the direction of the relationship between X and Y — just as in ordinary regression.



It should be clear that all this discussion applies when *any* single independent variable is increased by one unit; the increase does not have to be from zero to one. Now suppose that there are several independent variables. We hold all variables constant except x_k, and form an odds ratio. In the numerator is the odds of $Y = 1$ when x_k is increased by one unit, and in the denominator is the odds of $Y = 1$ when x_k is left alone. Both numerator and denominator are products (see Equation 6.2) and there is a lot of cancellationn in numerator and denominator. We are left with e^{β_k}. These calculations are a lot like the ones shown in (5.3) for regular regression; they will not be repeated here. But the conclusion is this. *When x_k is increased by one unit, and all other independent variables are held constant, the odds of $Y = 1$ are multiplied by e^{β_k}.*

"Analysis of covariance" with a binary outcome Here is one more example. Suppose the cases are patients with cancer, and we are comparing three treatments – radiation, chemotherapy and both. There is a single quantitative variable X, representing severity of the disease (a clinical judgement by the physician). The dependent variable is $Y = 1$ if the patient is alive 12 months later, zero otherwise. The question is which treatment is most effective, controlling for severity of disease.

Treatment will be represented by two indicator dummy variables: $d_1 = 1$ if the patient receives chemotherapy only, and $d_2 = 1$ if the patient receives radiation only. Odds of survival are shown in the table below.

Treatment	d_1	d_2	**Odds of Survival** $= e^{\beta_0}e^{\beta_1 d_1}e^{\beta_2 d_2}e^{\beta_3 x}$
Chemotherapy	1	0	$e^{\beta_0}e^{\beta_1}e^{\beta_3 x}$
Radiation	0	1	$e^{\beta_0}e^{\beta_2}e^{\beta_3 x}$
Both	0	0	$e^{\beta_0}e^{\beta_3 x}$

For any given disease severity x,

$$\frac{\text{Survival odds with Chemo}}{\text{Survival odds with Both}} = \frac{e^{\beta_0}e^{\beta_1}e^{\beta_3 x}}{e^{\beta_0}e^{\beta_3 x}} = e^{\beta_1}$$

and

$$\frac{\text{Survival odds with Radiation}}{\text{Survival odds with Both}} = \frac{e^{\beta_0}e^{\beta_2}e^{\beta_3 x}}{e^{\beta_0}e^{\beta_3 x}} = e^{\beta_2}.$$

If $\beta_1 = \beta_2 = 0$, then for any given level of disease severity, the odds of survival are the same in all three experimental conditions. So the test of $H_0 : \beta_1 = \beta_2 = 0$ would tell us whether, controlling for severity of disease, the three treatments differ in their effectiveness.

Sample Question 6.2.1 *What would $\beta_1 > 0$ mean?*

Answer to Sample Question 6.2.1 *Allowing for severity of diease, chemotherapy alone yields a higher one-year survival rate than the combination treatment. This could easily happen. Chemotherapy drugs and radiation are both dangerous poisons.*

This example shows that as in ordinary regression, categorical independent variables may be represented by collections of dummy variables. But parallel slopes on the log odds scale translates to *proportional* odds – like the odds of $Y = 1$ for Group 1 are always 1.3 times the odds of $Y = 1$ for Group 2, regardless of the value of x. How realistic this is will depend upon the particular application.

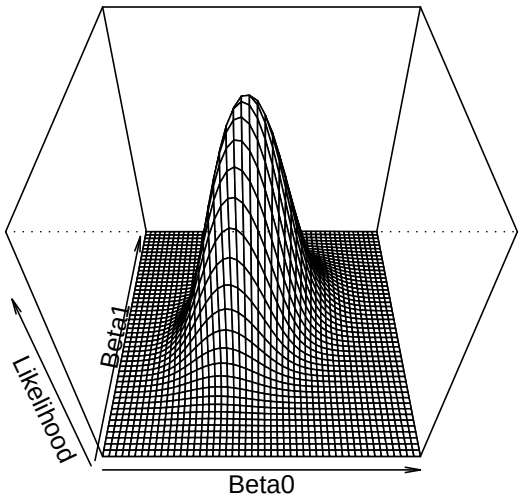
6.3 Parameter Estimation by Maximum likelihood

Using formula 6.4 for the probability of $Y = 1$ given the independent variable values, it is possible to calculate the probability of observing the data we did observe, for any set of β values. One of R. A. Fisher's many good suggestions was to take as our estimates

of β_0, β_1 and so forth, those values that made the probability of getting the data we actually observed as large as possible. Viewed as a function of the parameter values, the probability that we will get the data we actually did get is called the *likelihood*. The parameter values that make this thing as big as possible are called *maximum likelihood estimates*.

Figure 6.3 is a picture of this for one independent variable. The β_0, β_1 values located right under the peak is our set of maximum likelihood estimates. Of course it's hard to visualize in higher dimension, but the idea is the same.

Figure 6.3: Graph of the Likelihood Function for Simple Logistic Regression



In regular regression, maximum likelihood estimates are identical to least squares estimates, but not here (though they are close for large samples). Also, the $\widehat{\beta}$ quantities can be calculated by an explicit formula for regular regression, while for logistic regression they need to be found numerically. That is, a program like SAS must calculate the likelihood function for a bunch of sets of β values, and somehow find the top of the mountain. Numerical routines for maximum likelihood estimation essentially march uphill until they find a place where it is downhill in every direction. Then they stop.

For some statistical methods, the place you find this way could be a so-called "local maximum," something like the top of a foothill. You don't know you're not at the top of the highest peak, because you're searching blindfolded, just walking uphill and hoping for the best. Fortunately, this cannot happen with logistic regression. There is only one

peak, and no valleys. Start anywhere, walk uphill, and when it levels off you're at the top. This is true regardless of the particular data values and the number of independent variables.

6.4 Chi-square tests

As in regular regression, you can test hypotheses by comparing a full, or unrestricted model to a reduced, or restricted model. Typically the reduced model is the same as the full, except that's it's missing one or more independent variables. But the reduced model may be restricted in other ways, for example by setting a collection of regression coefficients equal to one another, but not necessarily equal to zero.

There are many ways to test hypotheses in logistic regression; most are large-sample chi-square tests. Two popular ones are likelihood ratio tests and Wald tests.

6.4.1 Likelihood ratio tests

Likelihood ratio tests are based on a direct comparison of the likelihood of the observed data assuming the full model to the likelihood of the data assuming the reduced model. Let $\mathcal{L}_F$ stand for the maximum probability (likelihood) of the observed data under the full model, and $\mathcal{L}_R$ stand for the maximum probability of the observed data under the reduced model. Dividing the latter quantity by the former yields a *likelihood ratio*: $\frac{\mathcal{L}_R}{\mathcal{L}_F}$. It is the maximum probability of obtaining the sample data under the reduced model (null hypothesis), *relative* to the maximum probability of obtaining the sample data under the null hypothesis under the full, or unrestricted model.

As with regular regression, the model cannot fit the data better when it is more restricted, so the likelihood of the reduced model is always less than the likelihood of the full model. If it's a *lot* less – that is, if the observed data are a lot less likely assuming the reduced model than assuming the full model – then this is evidence against the null hypothesis, and perhaps the null hypothesis should be rejected.

Well, if the likelihood ratio is small, then the natural log of the likelihood ratio is a big negative number, and minus the natural log of the likelihood ratio is a big positive number. So is twice minus the natural log of the likelihood ratio. It turns out that if the null hypothesis is true and the sample size is large, then the quantity

$$G = -2 \ln \left(\frac{\mathcal{L}_R}{\mathcal{L}_F} \right)$$

has an approximate chi-square distribution, with degrees of freedom equal to the number of non-redundant restrictions that the null hypothesis places on the set of β parameters. For example, if three regression coefficients are set to zero under the null hypotheses, the degrees of freedom equal three.

6.4.2 Wald tests

You may recall that the Central Limit Theorem says that even when data come from a non-normal distribution, the sampling distribution of the sample mean is approximately normal for large samples. The Wald tests are based on a kind of Central Limit Theorem for maximum likelihood estimates. Under very general conditions that include logistic regression, a collection of maximum likelihood estimates has an approximate multivariate normal distribution, with means approximately equal to the parameters, and variance covariance matrix that has a complicated form, but can be calculated (or approximated as a by-product of the most common types of numerical maximum likelihood).

This was discovered and proved by Abraham Wald, and is the basis of the Wald tests. It is pretty remarkable that he was able to prove this even for maximum likelihood estimates with no explicit formula. Wald was quite a guy. Anyway, if the null hypothesis is true, then a certain sum of squares of the maximum likelihood estimates has a large sample chi-square distribution. The degrees of freedom are the same as for the likelihood ratio tests, and for large enough sample sizes, the numerical values of the two tests statistics get closer and closer.

SAS makes it convenient to do Wald tests and inconvenient to do most likelihood ratio tests, so we'll stick to the Wald tests in this course.

Bibliography

[1] Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, **187**, 398-403.

[2] K. A. Bollen (1989). *Structural equations with latent variables,* New York: Wiley.

[3] Brunner, J. and Austin, P. C. (2009) Inflation of Type I error rate in multiple regression when independent variables are measured with error. *Canadian Journal of Statistics*, **37**, 33-46

[4] R. J. Carroll, D. Ruppert, L. A. Stefanski, & C. M. Crainiceanu (2006). *Measurement error in nonlinear models: a modern perspective. (2nd. ed.)* Boca Raton, FL : Chapman & Hall/CRC.

[5] Cody, R. P. and Smith, J. K. (1991). *Applied statistics and the SAS programming language.* (4th Edition) Upper Saddle River, New Jersey: Prentice-Hall.

[6] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* (2nd. Edition) Hillsdale, New Jersey: Erlbaum.

[7] Cook, T. D. and Campbell, D. T. (1979). *Quasi-experimentation: design and analysis issues for field settings.* New York: Rand McNally.

[8] Feinberg, S. (1977) *The analysis of cross-classified categorical data.* Cambridge, Massachusetts: MIT Press.

[9] Fisher, R. A. (1925) *Statistical methods for research workers.* London: Oliver and Boyd.

[10] W. A. Fuller (1987). *Measurement error models,* New York: Wiley.

[11] Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology - Impacts and Bayesian Adjustments.* Chapman & Hall/CRC, Boca Raton, USA.

[12] K. G. Jöreskog (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, 43, 443–477.

[13] Moore, D. S. and McCabe, G. P. (1993). *Introduction to the practice of statistics.* New York: W. H. Freeman.

[14] Neter, J., Kutner, M. H., Nachhtscheim, C. J. and Wasserman, W. (1996) *Applied linear statistical models.* (4th Edition) Toronto: Irwin.

[15] Roethlisberger, F. J. (1941). *Management and morale.* Cambridge, Mass.: Harvard University Press.

[16] Rosenthal, R. (1966). *Experimenter effects in behavioral research.* New York: Appleton-Century-Croft.

[17] Rosenthal, R. and Jacobson, L. (1968). *Pygmalion in the classroom: teacher expectation and pupils' intellectual development.* New York: Holt, Rinehart and Winston.

[18] Student (1908). "The probable error of a mean," *Biometrika* **6**, 1-25.