Logistic regression — modeling the probability of success

Regression models are usually thought of as only being appropriate for target variables that are continuous. Is there any situation where we might be interested in prediction of a categorical target variable? The answer is a most definite yes. Consider a study designed to investigate risk factors for cancer. Attributes of people are recorded, including age, gender, packs of cigarettes smoked, and so on. The target variable is whether or not the person has lung cancer (a 0/1 variable, with 0 for no lung cancer and 1 for the presence of lung cancer). A natural question is then "What factors can be used to predict whether or not a person will have lung cancer?" Substitute businesses for people, financial characteristics for medical risk factors, and whether a company went bankrupt for whether a person has cancer, and this becomes an investigation of the question "What financial characteristics can be used to predict whether or not a business will go bankrupt?"

There is a fundamental difference between this question and the kind of regression question we're used to asking. Rather than modeling the value of a target variable y, we are trying to model a probability. What is a sensible way to do that? We could simply do an ordinary least squares regression, treating the 0/1 variable as the target, but does this make sense? Define "success" to be the occurrence of the outcome coded 1, and let $p|\mathbf{x}$ be the probability of success given a set of predictor values. A linear regression model is consistent with

$$p|\mathbf{x} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k;$$

that is, the probability of success is linear in the predictors. Is this reasonable? Consider the following situation. We wish to model the probability of being admitted to a somewhat select college as a function of SAT score. There are three distinct kinds of situations:

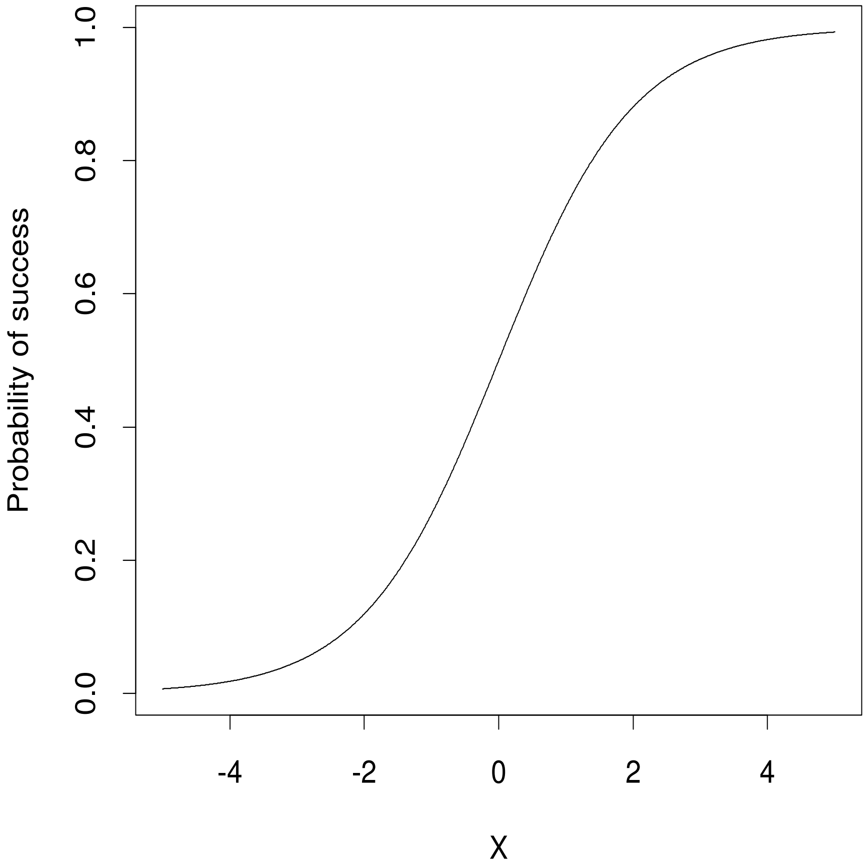
(1) For SAT scores around 1050, we might believe that each additional point on the SAT is associated with a fixed (constant) increase in the probability of being admitted. That is, a linear relationship is reasonable.

(2) However, for SAT score around 1400, this is no longer true. At that point the probability of being admitted is so high that an additional point on the SAT adds little to the probability of being admitted. In other words, the probability curve

as a function of SAT levels off for high values of SAT.

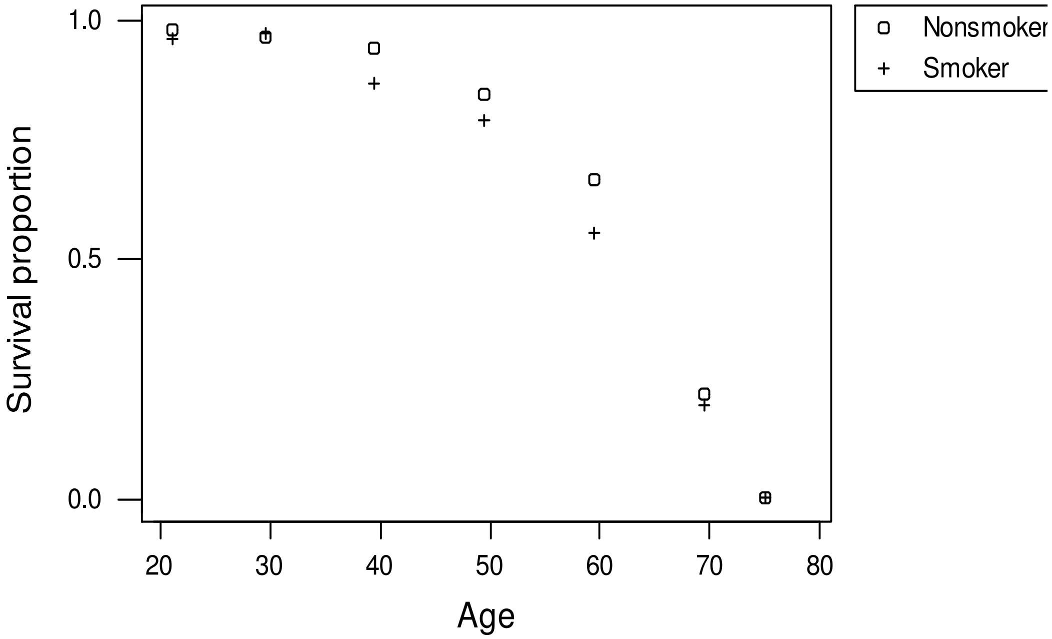
(3) The situation is similar for low SAT (say around 600). At that point the probability of being admitted is so low that one point lower on the SAT subtracts little from the probability of being admitted. In other words, the probability curve as a function of SAT levels off for low values of SAT.

These three patterns suggest that the probability curve is likely to have an S–shape, as in the following picture.



Hopefully this informal argument is convincing, but this characteristic S–shape also comes up clearly empirically. Consider, the following example, which we will look at more carefully a bit later. The data are from a survey of people taken in 1972–1974 in Whickham, a mixed urban and rural district near Newcastle upon Tyne, United Kingdom. Twenty years later a followup study was conducted. Among the information obtained originally was the age of the respondent, and whether they were a smoker or not, and it was recorded if the person was still alive twenty years later.

Here is a scatter plot of the observed survival proportion by the midpoint of each age interval, separated by smoking status. Note that the proportions do not follow a straight line, but rather an S–shape.



S–shaped curves can be fit using the *logit* (or *logistic*) function:

$$\eta \equiv \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x, \qquad (1)$$

where p is the probability of a success. The term $p/(1-p)$ are the *odds* of success:

$$\begin{array}{lcl} p = \frac{1}{2} & \Rightarrow & \frac{p}{1-p} = 1 \\ p = \frac{2}{3} & \Rightarrow & \frac{p}{1-p} = 2 \\ p = \frac{3}{4} & \Rightarrow & \frac{p}{1-p} = 3 \\ p = \frac{4}{5} & \Rightarrow & \frac{p}{1-p} = 4 \end{array}$$

Equation (1) is equivalent to

$$p = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}, \qquad (2)$$

which gives the desired S–shaped curve. Different choices of $\boldsymbol{\beta}$ give curves that are steeper (larger $|\beta_1|$) or flatter (smaller $|\beta_1|$), or shaped as an inverted (backwards) S (negative β_1). Equations (1) and (2) generalize in the obvious way to allow for multiple predictors.

The logistic model says that the log–odds follows a linear model. Another way of saying this is that $\hat\beta_1$ has the following interpretation: a one unit increase in x is estimated to be associated with multiplying the odds of success by $e^{\hat\beta_1}$, holding all else in the model fixed. The estimated odds of success when all predictors equal zero is obtained from the constant term as $e^{\hat\beta_0}$. The fact that these values are based on exponentiation reinforces that it is a good idea to use units that are meaningful for your data, as otherwise you can get estimated odds ratios that are either extremely large, almost exactly 1, or virtually zero. Note that since the logit is based on natural logs, there is a clear advantage to using the natural logarithm (base e) rather than the common logarithm (base 10) if you need to log a predictor: if this is done, the coefficient $\hat\beta_1$ represents an elasticity of the odds. So, for example, a coefficient $\hat\beta_1 = 2$ means that a 1% increase in x is associated with a (roughly) 2% increase in the odds of success.

The logistic regression model is an example of a *generalized linear model*. The model is that $y_i \sim \text{Binomial}(1, p_i)$, with p_i satisfying the logistic model (2). The parameters are estimated using maximum likelihood (OLS, WLS, and GLS are versions of maximum likelihood in Gaussian error regression models), which implies that the resultant estimated probabilities of success are the maximum likelihood estimates of the conditional probabilities of success given the observed values of the predictors.

The logit function is not the only function that yields S–shaped curves, and it would seem that there is no reason to prefer the logit to other possible choices. For example, another function that generates S–shaped curves is the cumulative distribution function for the normal distribution; analysis using that function is called *probit analysis*. Logistic regression (that is, use of the logit function) has several advantages over other methods, however. The logit function is what is called the *canonical link function*, which means that parameter estimates under logistic regression are fully efficient, and tests on those parameters are better behaved for small samples.

Even more importantly, the logit function is the only choice with a very important property related to the way the data were sampled. Let's say we were interested in building a model for the probability that a business will go bankrupt as a function of the initial debt carried by the business. There are two ways that we might imagine constructing a sample (say of size 200) of businesses to analyze:

(1) Randomly sample 200 businesses from the population of interest. Record the initial debt, and whether or not the business eventually went bankrupt. This is conceptually consistent with following the 200 businesses through time until they either go bankrupt or don't go bankrupt, and is called a *prospective* sampling scheme for this reason. In the biomedical literature this is often called a *cohort* study.

(2) First consider the set of all businesses in the population that didn't go bankrupt; randomly sample 100 of them and record the initial debt. Then consider the set of all businesses in the population that did go bankrupt; randomly sample 100 of them and record the initial debt. This is conceptually consistent with seeing the final state of the businesses first (bankrupt or not bankrupt), and then going backwards in time to record the initial debt, and is called a *retrospective* sampling scheme for this reason. In the biomedical literature this is often called a *case–control* study.

Each of these sampling schemes has advantages and disadvantages. The prospective study is more consistent with the actual physical process that we are interested in; for example, the observed sample proportion of businesses that go bankrupt is an estimate of the actual probability of a randomly chosen business from this population going bankrupt, a number that cannot be estimated using data from a retrospective study (recall that we arbitrarily decided that half the sample would be bankrupt businesses, and half would be non–bankrupt businesses). Cohort studies also have the advantage that they can be used to study multiple outcomes; for example, a single cohort study using doctors has been used to examine factors related to different types of cancer, emphysema, heart disease, stroke, and other diseases. On the other hand, if bankruptcy rates are low (say 15%), a sample of size 200 is only going to have about 30 bankrupt businesses in it, which makes it more difficult to model the probability of bankruptcy. Note that one way to distinguish between these two sampling approaches is that in a retrospective study the *sampling rate* is different for successes and for failures; that is, you deliberately oversample one group and undersample the other so as to get a "reasonable" number of observations in each group (it is **not** required that the two groups have the same number of observations, only that one group is oversampled while the other is undersampled).

To simplify things, let's assume that initial debt is recorded only as Low or High. This implies that the data take the form of a 2 × 2 contingency table (whatever the sampling scheme):

		Bankrupt		
		Yes	*No*	
Debt	*Low*	n_{LY}	n_{LN}	n_L
	High	n_{HY}	n_{HN}	n_H
		n_Y	n_N	n

Even though the data have the same form, whatever the sampling scheme, the way these data are generated is very different. Here are two tables that give the expected counts in the four data cells, depending on the sampling scheme. The π values are conditional probabilities (so, for example, $\pi_{Y|L}$ is the probability of a business going bankrupt given that it has low initial debt):

PROSPECTIVE SAMPLE

		Bankrupt	
		Yes	*No*
Debt	*Low*	$n_L \pi_{Y|L}$	$n_L \pi_{N|L}$
	High	$n_H \pi_{Y|H}$	$n_H \pi_{N|H}$

RETROSPECTIVE SAMPLE

		Bankrupt	
		Yes	*No*
Debt	*Low*	$n_Y \pi_{L|Y}$	$n_N \pi_{L|N}$
	High	$n_Y \pi_{H|Y}$	$n_N \pi_{H|N}$

Note that there is a fundamental difference between the probabilities that can be estimated using the two sampling schemes. For example, what is the probability that a business goes bankrupt given that it has low initial debt? As noted above, this is $\pi_{Y|L}$. It is easily estimated from a prospective sample ($\hat{\pi}_{Y|L} = n_{LY}/n_L$), as can be seen from the left table, but it is impossible to estimate it from a retrospective sample. On the other hand, given that a business went bankrupt, what is the probability that it had low initial debt? That is $\pi_{L|Y}$, which is estimable from a retrospective sample ($\hat{\pi}_{L|Y} = n_{LY}/n_Y$), but not from a prospective sample.

This is where the advantage of logistic regression comes in. While these different probabilities are not estimable for one sampling scheme or the other, the comparison between odds for the groupings that are relevant for that sampling scheme are identical in either scheme, and is always estimable. Consider first a prospective study. The odds are $p/(1-p)$, where here p is the probability of a business going bankrupt. For the low debt businesses, $p = \pi_{Y|L}$, so the odds are

$$\frac{\pi_{Y|L}}{1-\pi_{Y|L}} = \frac{\pi_{Y|L}}{\pi_{N|L}}.$$

For the high debt businesses, the odds are

$$\frac{\pi_{Y|H}}{1-\pi_{Y|H}} = \frac{\pi_{Y|H}}{\pi_{N|H}}.$$

Recall that the logistic regression model says that the logarithm of odds follows a linear model; that is, for this model, the ratio of the odds is a specified value. This ratio of odds (called the *cross–product ratio*) equals

$$\frac{\pi_{Y|L}/\pi_{N|L}}{\pi_{Y|H}/\pi_{N|H}}.$$

By the definition of conditional probabilities, this simplifies as follows:

$$\begin{aligned}\frac{\pi_{Y|L}/\pi_{N|L}}{\pi_{Y|H}/\pi_{N|H}} &= \frac{\pi_{Y|L}\pi_{N|H}}{\pi_{Y|H}\pi_{N|L}} \\ &= \frac{\frac{\pi_{LY}}{\pi_L}\frac{\pi_{HN}}{\pi_H}}{\frac{\pi_{HY}}{\pi_H}\frac{\pi_{LN}}{\pi_L}} \\ &= \frac{\pi_{LY}\pi_{HN}}{\pi_{HY}\pi_{LN}}.\end{aligned}$$

This last value is easily estimated using the sample cross–product ratio,

$$\frac{n_{LY}n_{HN}}{n_{HY}n_{LN}}.$$

Now consider a retrospective study. The odds are $p/(1-p)$, where here p is the probability of a business having low debt (remember that the proportion that are bankrupt is set by the design). For the bankrupt businesses, $p = \pi_{L|Y}$, so the odds are

$$\frac{\pi_{L|Y}}{1-\pi_{L|Y}} = \frac{\pi_{L|Y}}{\pi_{H|Y}}.$$

For the non–bankrupt businesses, the odds are

$$\frac{\pi_{L|N}}{1-\pi_{L|N}} = \frac{\pi_{L|N}}{\pi_{H|N}}.$$

The cross–product ratio equals

$$\frac{\pi_{L|Y}/\pi_{H|Y}}{\pi_{L|N}/\pi_{H|N}}.$$

By the definition of conditional probabilities, this simplifies as follows:

$$\begin{aligned}\frac{\pi_{L|Y}/\pi_{H|Y}}{\pi_{L|N}/\pi_{H|N}} &= \frac{\pi_{L|Y}\pi_{H|N}}{\pi_{L|N}\pi_{H|Y}} \\ &= \frac{\frac{\pi_{LY}}{\pi_Y}\frac{\pi_{HN}}{\pi_N}}{\frac{\pi_{LN}}{\pi_N}\frac{\pi_{HY}}{\pi_Y}} \\ &= \frac{\pi_{LY}\pi_{HN}}{\pi_{LN}\pi_{HY}}.\end{aligned}$$

This, of course, is identical to the cross–product ratio from the prospective study. That is, while the type of individual probability that can be estimated from data depends on the sampling scheme, the cross–product ratio is unambiguous whichever sampling scheme is appropriate. The logit link function is the only choice where effects are determined by the cross–product ratio, so it is the only link function with this property.

It might seem that a retrospective sampling scheme is a decidedly inferior choice, since we cannot answer the question of most interest; that is, what is the probability that a business goes bankrupt given its initial debt level? There is, however, a way that we can construct an answer to this question, if we have additional information. Let π_Y and π_N be the true probabilities that a randomly chosen business goes bankrupt or does not go bankrupt, respectively. These numbers are called *prior probabilities.* Consider the probability of bankruptcy given a low initial debt level, $\pi_{Y|L}$. By the definition of conditional probabilities,

$$\begin{aligned}\pi_{Y|L} &= \frac{\pi_{LY}}{\pi_L} \\ &= \frac{\pi_{L|Y}\pi_Y}{\pi_L} \\ &= \frac{\pi_{L|Y}\pi_Y}{\pi_{LY}+\pi_{LN}} \\ &= \frac{\pi_{L|Y}\pi_Y}{\pi_{L|Y}\pi_Y+\pi_{L|N}\pi_N}.\end{aligned}$$

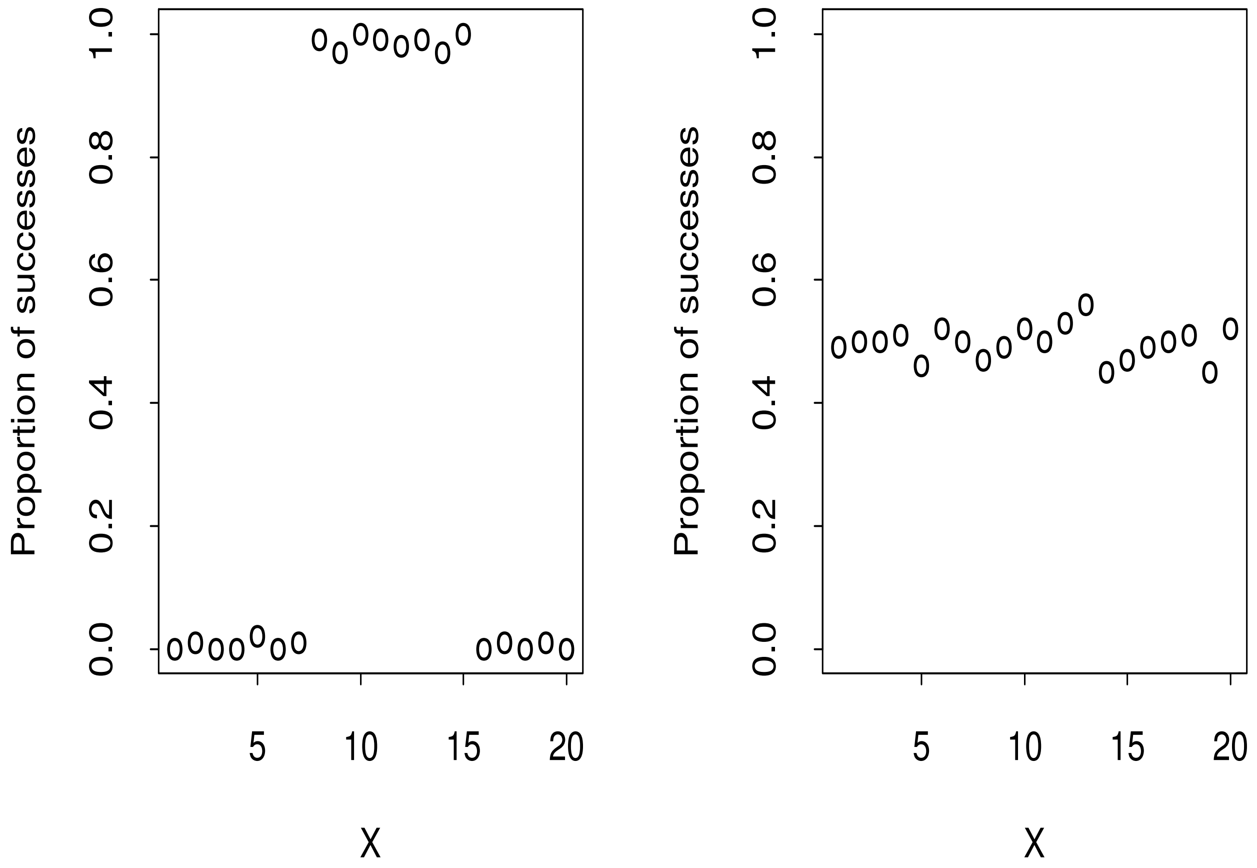
The probabilities $\pi_{L|Y}$ and $\pi_{L|N}$ are estimable in a retrospective study, so if we have a set of prior probabilities (π_Y, π_N) we can estimate the probability of interest. This point comes up again in other analyses involving group classification, such as discriminant analysis.

Since odds ratios are uniquely defined for both prospective and retrospective studies, the slope coefficients $\{\beta_1, \beta_2, \ldots\}$ are also uniquely defined. The constant term β_0, however, is driven by the observed proportions of successes and failures, so it is affected by the construction of the study. Thus, as noted above, in a retrospective study, the results of a logistic regression fit cannot be used to estimate the probability of success. If prior probabilities of success (π_Y) and failure (π_N) are available, however, the constant term can be adjusted, so that probabilities can be estimated. The adjusted constant term has the form

$$\tilde{\beta}_0 = \hat{\beta}_0 + \ln\left(\frac{\pi_Y n_N}{\pi_N n_Y}\right).$$

Thus, probabilities of success can be estimated for a retrospective study using equation (2) (allowing for multiple predictors, of course), with β_0 estimated using $\tilde{\beta}_0$ and β_1 estimated using $\hat{\beta}_1$. Note that this adjustment is **only** done when the sample is a retrospective one, **not** if (for example) observations in a prospective study have been omitted because they are outliers. If you're not sure about what (π_Y, π_N) should be, you can try a range of values to assess the sensitivity of your estimated probabilities based on the adjusted intercept to the specific choice you make; one way or another, however, you should be producing reasonable prospective probability estimates as part of a full analysis.

The logistic model (1) or (2) is a sensible one for probabilities, but is not necessarily appropriate for a particular data set. This is **not** the same thing as saying that the predicting variables are not good predictors for the probability of success. Consider the following two plots. The variable `X` is a potential predictor for the probability of success, while the vertical axis gives the observed proportion of successes in samples taken at those values of `X`. So, for example, `X` could be the dosage of a particular drug, and the target variable is the proportion of people in a trial that were cured when give that dosage.



In the plot on the left, `X` is very useful for predicting success, but the logistic regression model does **not** fit the data, since the probability of success is not a monotone function of `X`. In the plot on the right, `X` is not a useful predictor for success (the probability of success appears to be unrelated to `X`), but the logistic regression model fits the data, as a very flat S–shaped curve goes through the observed proportions of success reasonably well. `Minitab` provides three goodness–of–fit tests that test the hypotheses

$$H_0 : \text{the logistic model fits the data}$$

versus

$$H_a : \text{the logistic model does not fit the data}$$

that can be used to determine if the correct model is being fit. What if it isn't? Then the model (1) should be enriched to allow for the observed pattern. For the nonmonotonic pattern in the left plot above, obvious things to try are to fit a quadratic model by including $\texttt{X}^2$

as a predictor, or a model where X is considered nominal with three levels corresponding to low, medium, and high. A variation on this theme is to find and use additional predictors. Note that one reason why the logistic model might not fit is that the relationship between logits and predictors might be multiplicative, rather than additive; so, for example, if a predictor is long right–tailed, you might find that using a logged predictor is more effective than using an unlogged one. Unusual observations (outliers, leverage points) are also still an issue for these models, and Minitab provides diagnostics for them.

Another source of lack of fit of a logistic regression model is that the binomial distribution isn't appropriate. This can happen, for example, if there is correlation among the responses, or if there is heterogeneity in the success probabilities that hasn't been modeled. Both of these violations can lead to *overdispersion*, where the variability of the probability estimates is larger than would be implied by a binomial random variable. In that case you would need to use methods that correct for the overdispersion, either implicitly (such as through what is known as *quasi-likelihood*) or explicitly (fitting a regression model using a response distribution that incorporates overdispersion, such as the beta-binomial distribution). All of these issues are covered in my book *Analyzing Categorical Data*, which was published by Springer-Verlag in 2003.

Logistic regression also can be used for prediction of group membership, or *classification*. A fitted model (with estimated $\hat{\boldsymbol{\beta}}$) can be applied to new data to give an estimate of the probability of success for that observation using equation (2). If a simple success/failure prediction is desired, this can be done by calling an estimated $\hat{p}$ less than .5 a predicted failure, and an estimated $\hat{p}$ greater than .5 a success. Other cutoff values might make sense also. For example, the prior probability of success, often called the *base rate* (if one is available) or the observed sample proportion of successes is often more effective if those values are far from .5 (that is, if only 30% of the observations are successes, say, it's reasonable to set the cutoff in the classification matrix as .3). Another way to classify is to choose a cutoff that takes into account the costs of misclassification into success and failure. If this is done to the same data used to fit the logistic regression model, a *classification table* results. Here is a hypothetical example of such a table:

		Predicted group		
		Success	Failure	
Actual	Success	8	3	11
group	Failure	2	7	9
		10	10	

In this example 75.0% of the observations (15 out of 20) were correctly classified to the appropriate group. Is that a lot? More formally, is that more than we would expect just by random chance? The answer to this question is not straightforward, because we have used the same data to both build the model (getting $\hat{\boldsymbol{\beta}}$) and evaluate its ability to do predictions (that is, we've used the data twice). For this reason, the observed proportion correctly classified can be expected to be biased upwards compared to if the model was applied to completely new data.

What can we do about this? The best thing to do is what is implied by the last line in the previous paragraph — get some new data, and apply the fitted model to it to see how well it does (that is, validate the model on new data). What if no new data are available? Two diagnostics that have been suggested can help here. One approach we could take would be to say the following: if we simply predicted every observation to have come from the larger group, what proportion would we have gotten right? For these data, that is

$$C_{max} = \max\left(\frac{11}{20}, \frac{9}{20}\right) = 55.0\%.$$

We certainly hope to do better than this naive rule, so C_{max} provides a useful lower bound for what we would expect the observed proportion correctly classified to be. The observed 75.0% correctly classified is quite a bit larger than C_{max}, supporting the usefulness of the logistic regression.

A more sophisticated argument is as follows. If the logistic regression had no power to make predictions, the actual group that an observation falls into would be independent of the predicted group. That is, for example,

$$P(\text{Actual group a success and Predicted group a success}) =$$
$$P(\text{Actual group a success}) \times P(\text{Predicted group a success}).$$

The right side of this equation can be estimated using the marginal probabilities from the classification table, yielding

$$\begin{aligned} \widehat{P(\text{Actual group a success and Predicted group a success})} &= \left(\frac{11}{20}\right)\left(\frac{10}{20}\right) \\ &= .275. \end{aligned}$$

That is, we would expect to get 27.5% of the observations correctly classified as successes just by random chance. A similar calculation for the failures gives

$$\begin{aligned} \widehat{P(\text{Actual group a failure and Predicted group a failure})} &= \left(\frac{9}{20}\right)\left(\frac{10}{20}\right) \\ &= .225. \end{aligned}$$

Thus, assuming that the logistic regression had no predictive power for the actual group (success or failure), we would expect to correctly classify 27.5% + 22.5% = 50% of the observations. This still doesn't take into account that we've used the data twice, so this number is typically inflated by 25% before being reported; that is,

$$C_{pro} = 1.25 \times 50\% = 62.5\%.$$

The observed 75% is still considerably higher than C_{pro}, supporting the usefulness of the logistic regression. Note that if the misclassification rate for one of the groups is greater than 50% it is possible that a different cutoff value could be found that better trades off classification accuracy in the two groups.

At the beginning of this handout I mentioned the modeling of bankruptcy as an example of a potential logistic regression problem. There is a long history of at least 35 years of trying to use publicly available data to predict whether or not a company will go bankrupt, and it is clear at this point from many empirical studies that it is possible to do so, at least up to four years ahead. That is, classification tables show better classificatory performance than would have been expected by chance. Interestingly, however, this does not translate into the ability to make money, as apparently stock price reaction precedes the forecast, providing yet another argument in support of the efficient market hypothesis.

It is possible that a logistic regression data set constitutes a time series, and in that case autocorrelation can be a problem (recall that one of the assumptions here is that the Bernoulli observations be independent of each other). Time series models for binary response data are beyond the scope of this course, but there are at least a few things you can do to assess whether autocorrelation might be a problem. If the data form a time series, it is sensible to plot the standardized Pearson residuals in time order to see if there are any apparent patterns. You can also construct an ACF plot of the residuals, and construct a runs test for them. You should recognize, however, that the residuals are only roughly normally distributed (if that), and often have highly asymmetric distributions, so these tests can only be viewed as suggestive of potential problems. If you think that you might have autocorrelation, one potential corrective procedure to consider is to use a lagged version of the response as a predictor.

Logistic regression can be generalized to more than two groups. *Polytomous* (or *nominal*) logistic regression, the generalization of a binomial response with two groups to a multinomial response with multiple groups, proceeds by choosing (perhaps arbitrarily) one group as the "control" or "standard." Say there are G groups, and group G is the one chosen as the standard. The logistic model states that the probability of falling into group j given the set of predictor values $\mathbf{x}$ is

$$P(\text{group } j|\mathbf{x}) = \frac{\exp(\beta_{0j} + \beta_{1j}x_1 + \cdots + \beta_{kj}x_k)}{1 + \sum_{\ell=1}^{G-1} \exp(\beta_{0\ell} + \beta_{1\ell}x_1 + \cdots + \beta_{k\ell}x_k)}.$$

Equivalently,

$$\ln\left(\frac{p_j}{p_G}\right) = \beta_{0j} + \beta_{1j}x_1 + \cdots + \beta_{kj}x_k.$$

If group G can be thought of as a control group, for example, this is a natural model where the log–odds compared to the control is linear in the predictors (with different coefficients for each of the groups). So, for example, in a clinical trial context for a new drug, the control group might be "No side effects," while the other groups are "Headache," "Back pain," and "Dizziness." If group G is chosen arbitrarily, it is harder to interpret the coefficients of the model. Each of these models could be fit separately using ordinary (binary) logistic regression, but the resultant coefficient estimates would not be exactly the same. More importantly, polytomous logistic regression allows the construction of a test of the statistical significance of the variables in accounting for the *simultaneous* difference in probabilities of *all* of the response categories, not just individual groups versus the control group. Further, the estimated probability vector (of the probabilities of falling into each group) for an observation is guaranteed to be the same no matter which group is chosen as the reference group when fitting a polytomous logistic regression model. If it was desired to classify observations, they would then be assigned to the group with largest estimated probability.

There is one additional assumption required for the validity of the multinomial logistic regression model — the assumption of *independence of irrelevant alternatives*, or IIA. This assumption states that the odds of one group relative to another is unaffected by the presence or absence of other groups. A test for this condition exists (the so-called Hausman-McFadden test), although it is not provided in `Minitab`. An alternative approach that does not assume IIA is the multinomial probit model (also not provided in `Minitab`), although research indicates that the multinomial logit model is a more effective model even if the IIA assumption is violated.

A classification matrix can be formed when there are more than two groups the same way that it is when there are two groups. The two benchmarks C_{max} and C_{pro} are formed in an analogous way. C_{max} is the proportion of the sample that comes from the largest of the groups in the sample. To calculate C_{pro}, calculate the proportions of the sample that come from each group and the proportions of the sample that are predicted to be from each group; then, multiply the actual group proportion times the predicted group proportion for each group and sum over all of the groups; finally, multiply by 1.25.

In many situations there is a natural ordering to the groups, which we would like to take into account. For example, respondents to a survey might be asked their opinion on a political issue, with responses coded on a 5–point Likert scale of the form Strongly disagree – Disagree – Neutral – Agree – Strongly agree. Is there some way to analyze this type of data set? There are two affirmative answers to that question — a common, but not very correct approach, and a not–so–common, but much more correct approach. The common approach is to just use ordinary least squares regression, with the group membership identifier as the target variable. This often doesn't work so badly, but there are several problems with its use in general:

(1) An integral target variable clearly is inconsistent with continuous (and hence Gaus-

sian) errors, violating one of the assumptions when constructing hypothesis tests and confidence intervals.

(2) Predictions from an ordinary regression model are of course nonintegral, resulting in difficulties in interpretation: what does a prediction to group 2.763 mean, exactly?

(3) The least squares regression model doesn't answer the question that is of most interest here — what is the estimated probability that an observation falls into a particular group given its predictor values?

(4) The regression model implicitly assumes that the groups are "equally distant" from each other, in the sense that the lowest group is as far from the second lowest as the highest is from second highest. It could easily be argued that this is not sensible for a given data set.

The solution to these problems is a generalization of the logistic regression model to *ordinal* data. The model is still based on logits, but now they are *cumulative* logits. The most common cumulative logit model (and the one fitted by `Minitab`) is called the *proportional odds model.* For simplicity, consider a model with one predictor x and target group identifier Y that is a measure of agreement rated from lowest to highest. Let

$$L_j(x) = \text{logit}[F_j(x)] = \ln\left[\frac{F_j(x)}{1-F_j(x)}\right], \qquad j = 1, \ldots, J-1,$$

where $F_j(x) = P(Y \leq j|x)$ is the cumulative probability for response category j, and J is the total number of response categories (so Y takes on values $\{1, \ldots, J\}$, corresponding to strong disagreement to strong agreement). The proportional odds model says that

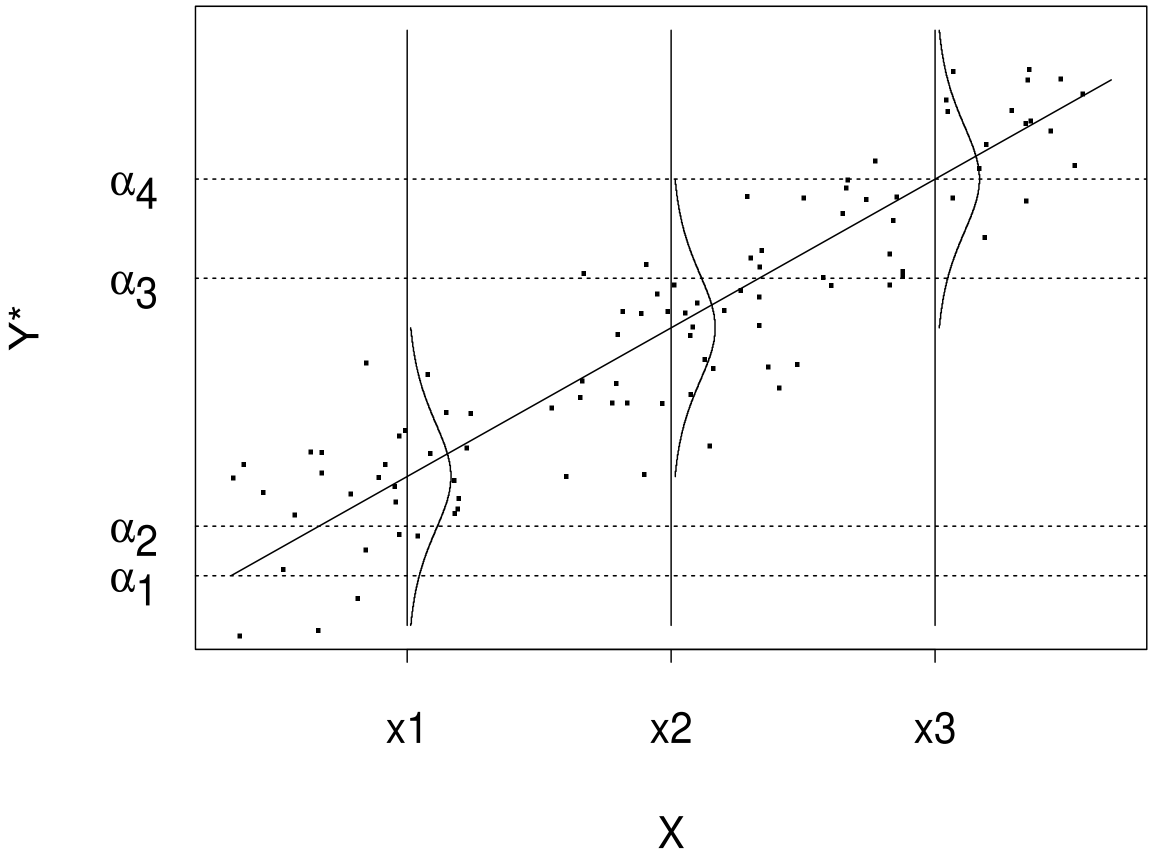
$$L_j(x) = \alpha_j + \beta x, \qquad j = 1, \ldots, J-1.$$

Note that this means that a positive β is associated with increasing odds of being less than a given value j; that is, a positive coefficient implies increasing probability of being in lower–numbered categories with increasing x.

This model can be motivated using an underlying continuous response variable Y^* representing, in some sense, agreement. Say Y^* follows the logistic distribution with location $\beta_0 + \beta x$ for given value of x. Suppose $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_J = \infty$ are such that the observed response (university rating) Y satisfies

$$Y = j \text{ if } \alpha_{j-1} < Y^* \leq \alpha_j.$$

That is, we observe a response in category j when the underlying Y^* falls in the jth interval of values. Then, if the underlying response is linearly related to x, the proportional odds model is the representation for the underlying probabilities. Thus, the model allows for the same idea of linearity that a least squares regression model does, but allows for unequally spaced groups through the estimates of $\boldsymbol{\alpha}$. That is, for example, for a 5–point Likert scale target the model allows for the possibility that people who strongly disagree with a statement are "more like" people who disagree (responses 1 and 2), than are people who disagree and people who are neutral (responses 2 and 3). Here is a graphical representation of what is going on (this is actually Figure 10.3 from my *Analyzing Categorical Data* book):



You might wonder about the logistic distribution mentioned before. This distribution is similar to the normal distribution, being symmetric but slightly longer–tailed. A version of the proportional odds model that assumes a normal distribution for the underlying variable Y^* would be based on probits, rather than logits.