Logistic Regression Introduction

Paul Johnson <pauljohn@ku.edu>

July 15, 2011

1 Times Have Changed

Logistic regression is one of my areas of special interest. I first learned of it in 1981, when the only way to estimate these models was to buy a computer tape and send it to CalTech, where Richard McKelvey's students would copy some Fortran programs onto the tape and then send it back.

In those days, Logistic regression was new and unfamiliar and, somewhat bizarrely in my view, seen by many as an unnecessary complication in the regression modeling process.

In 1987, when I started teaching at KU, the SPSS statistical package did not include procedures for logistic regression. SAS offered an experimental routine called "Logistic" that was contributed by Frank Harrell, who has since moved on to write procedures for S+/R. Political scientist Doug Rivers wrote a DOS program that he called SST and it was commercialized for a while, at least until LimDep, the program written by William Greene (to go along with his massively successful econometrics textbook) hit the market. In all of these packages, logistic regression was treated as a special case.

In 1989, McCullagh and Nelder published their book *Generalized Linear Models* and the logistic regression model was recast into a family of models. In the S+/R statistical packages, the generalized linear model framework is adopted, and a logit model is fitted with a "binomial distribution" and a "logit link" function. For example,

```
myLogit <-
glm(y ~ x1+x2, data=whatever, family=binomial(link=logit))
```

The glm procedure assumes the link is logit unless you specify otherwise, so that can simply be written

```
myLogit <- glm(y ~ x1+x2, data=whatever, family=binomial)
```

If instead you want the so-called "probit" model, one simply changes the link:

```
myLogit <-
glm(y ~ x1+x2, data=whatever, family=binomial(link=probit))
```

As you might have guessed, there are many other link functions that can be used. I've always wondered if I could find a use for the option link=cauchit, but only because I think that it is fun to say that out loud. It sounds like a sneeze! I've never used link=cloglog, but it sounds fun too.

One of the main benefits of the GLM approach is that the same estimation routine can be used for many different theories; it is thus mainly a benefit to the people who write programs. To the users–like us–perhaps we don't care so much that the programmers save time or that their work is more coherent. Many of us are just as happy if they have to write completely separate programs for all of these models that are only slightly different.

But the experts who develop these models are also doing the programming, of course, and so we, "the users," have to try to understand what they are talking about. Regression for categorical dependent variables–logistic, probit, or whatever, is not conceptually difficult. But it is difficult in practice because there are many competing sets of terminology for it. It is difficult to teach because it has been developed from several different–and I mean completely different–schools of thought.

2 Funny Looking Graphs!

2.1 y_i is dichotomous

Suppose y_i is coded 0 and 1, representing answers to a Yes or No question. Consider Figure 1:

Suppose you fit a straight line through that distribution of observations. Go ahead, knock yourself out! See Figure 1. Before logistic regresion was widely known, this was the best we could do. It was called a "linear probability model". The theory is that

$$y_i = b_0 + b_1 X_i + e_i \tag{1}$$

If $E(e_i) = 0$, then the predicted value can be thought of as the probability of $y_i = 1$.

2.2 The Straight Line is Obviously wrong.

OLS predicts out of range.

As long as $b_1 \neq 0$, the line representing predicted values will go above 1 and below 0. This is evident if we construct a more extreme example. Consider Figure 3. What do you say about the bottom left and top right?

You could 'truncate' the predicted values at 0 and 1, I suppose...

To prevent the OLS model from going out of bounds, one approach is to insert "kinks" in the fitted line. That will constrain the predictions so that they can't go above 0 or 1. This is a constrained linear probability model.

It has all kinds of unattractive features, including

Figure 1: Plot a dichotomous Y

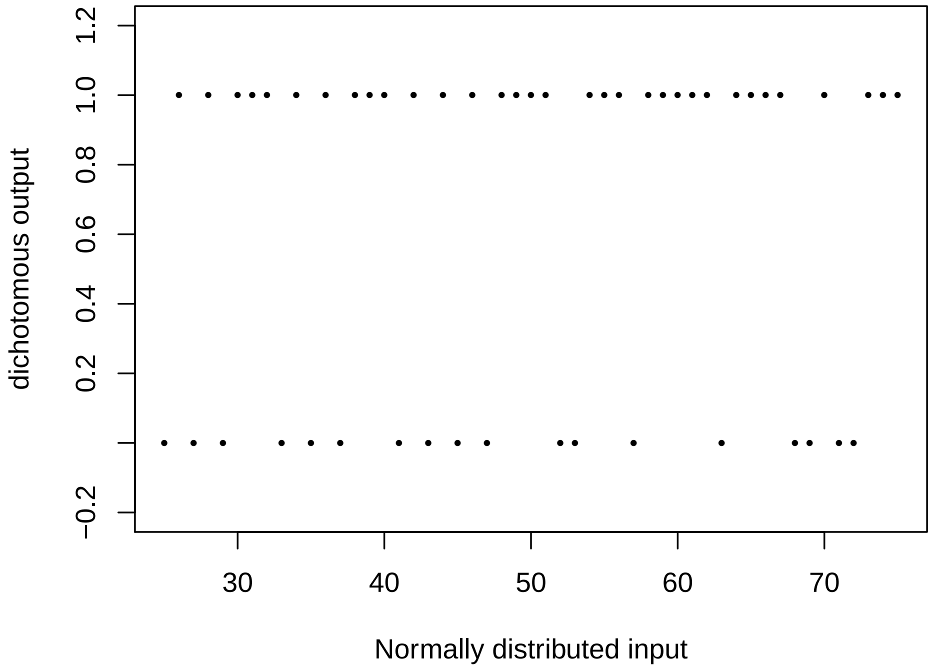


Figure 2: Fit OLS with the Dichotomous Data

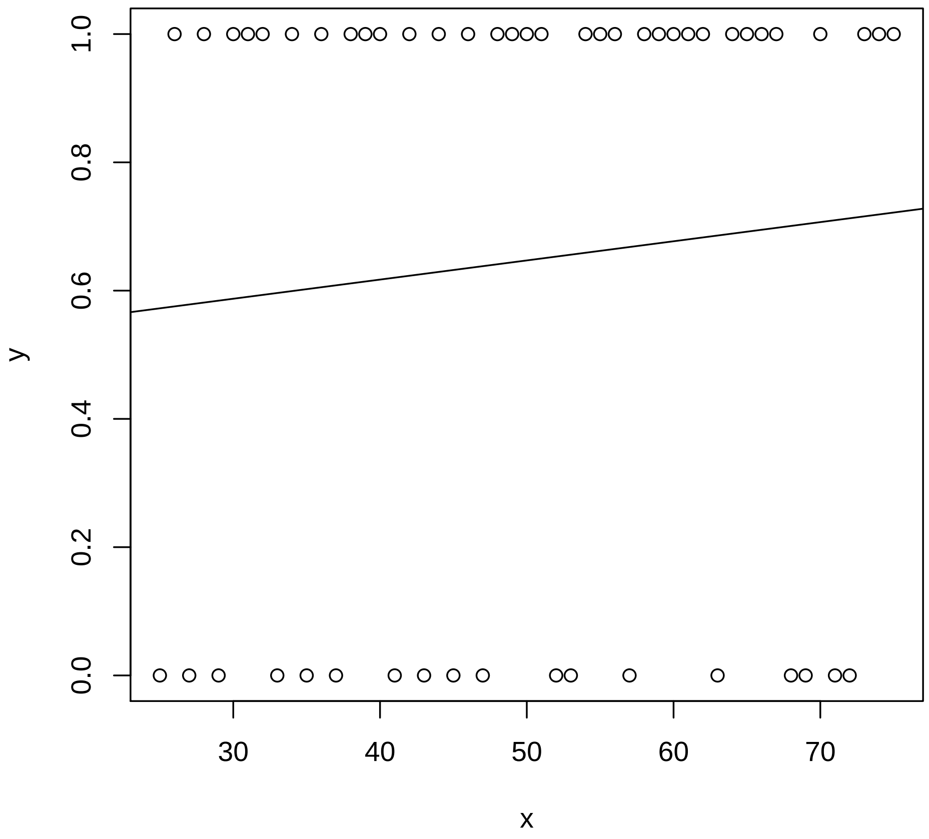
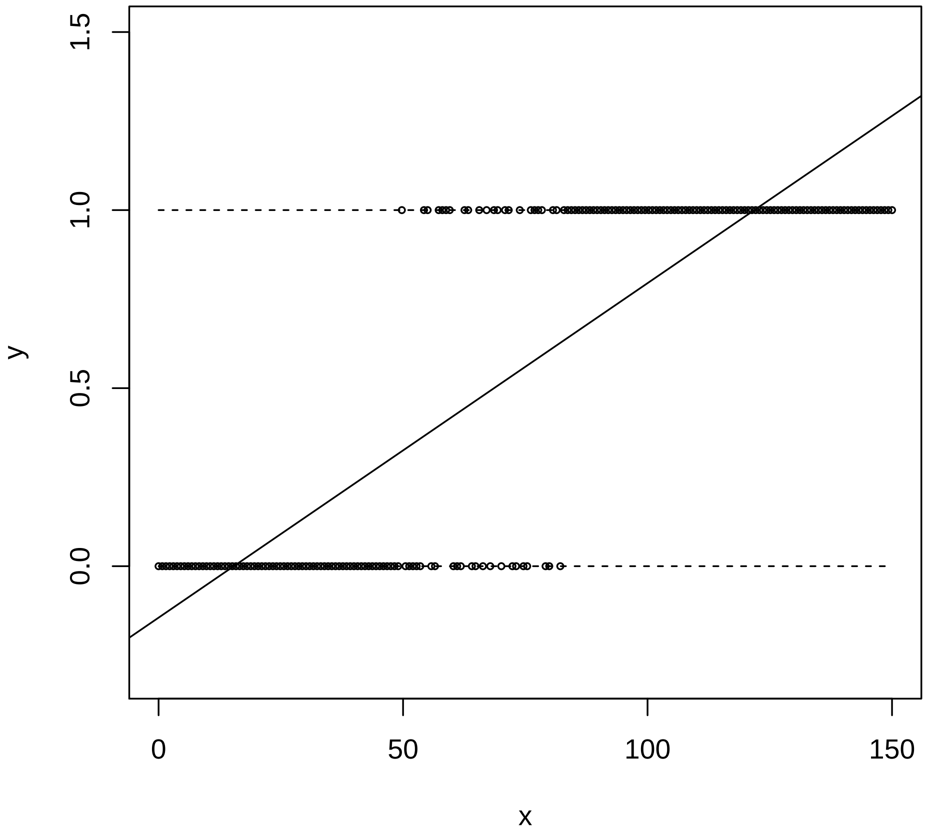


Figure 3: OLS out of range



1. Sharp kinks in the predicted value curve are theoretically unappealing. Why would it be that the input has no effect, then a linear effect, then no effect again?

2. Does $\hat{y}_i = 0$ mean something is actually impossible? Does $\hat{y}_i = 1$ mean something is certain to happen? Sometimes a model will state that a 1 is certain to happen, and yet the observation is 0. Something "impossible" happened! Does it mean the whole model is wrong?

3. I don't know of an estimation method that actually tries to account for the "kinks" in the predicted values.

Heteroskedasticity.

Newcomers should skip this section. It is included only for completeness.

Recall that OLS assumes the variance of the error term is the same for all cases. It seems obvious cannot apply in the linear probability model. If $\hat{y}_i = 0.5$, then we are very uncertain about the outcome, so it will have an error with high variance. On the other hand, if $\hat{y}_i = 0.99$, then we are very very confident the outcome will be 1, and the error term cannot have the same variance.

You may recall that a vital, indispensable assumption in regression is that $E(e_i) = 0$ is vital. For any given value of X_i, note that the error term has to make up the difference between $b_0 + b_1 \cdot X_i$ and the true value, 1 or 0. Hence the error term must be either $1 - b_0 - b_1 \cdot X_i$ or $+b_0 + b_1 \cdot X_i$.

Let P_i be the probability that y_i is 1. That is, $P_i = b_0 + b_1 \cdot X_i$. Note this is not the predicted value from an estimated equation–it is the probability from the equation with the true values of b_0 and b_1 inserted.

To repeat the story in the previous paragraph, if $y_i = 1$, then the error term must be $1 - P_i$, because this amount goes from P_i up to 1. Similarly, the probability that $y_i = 0$ is $(1 - P_i)$. And, if $y_i = 0$, that must mean the error term is $-P_i$, because that is the amount you have to take away from P_i to get down to zero. Hence, for a particular value of X_i, the expected value of the error term must be:

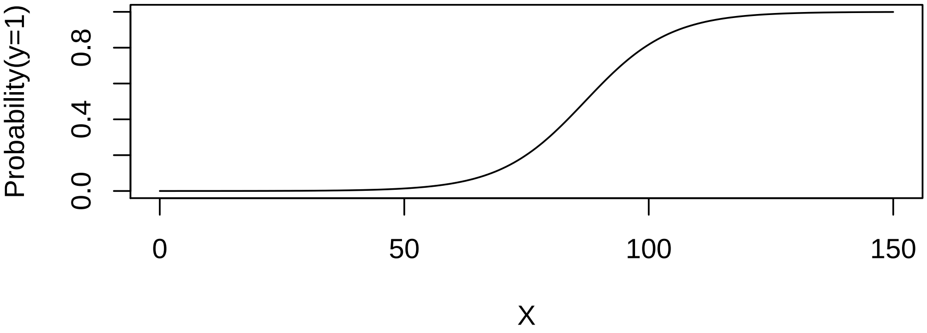
$$E[e_i] = (1 - P_i)P_i + P_i(1 - P_i) = 0 \tag{2}$$

By the same logic, the variance of the error term is

$$\begin{aligned} E(e_i^2) &= (1 - P_i)P_i^2 + P_i(1 - P_i)^2 \\ &= P_i(1 - P_i) \\ &= (b_0 + b_1 X_i)(1 - b_0 - b_1 X_i) \end{aligned} \tag{3}$$

As you can plainly see, the variance of the error term depends on X_i. There is heteroskedasticity by definition. You might treat this with WLS, but in small samples that not a very desirable proposition.

Figure 4: An S-shaped Curve



3 Just One! I Just Need One S-Shaped Curve.

Out in “the real world,” suppose that the probability that $y_i = 1$ changes “smoothly” in response to changes in x_i. That implies that the relationship between x_i and $\boldsymbol{Prob}(y_i = 1)$ should be S-shaped, as illustrated in Figure 4.

Here’s a formula for an S-shaped curve that I’ve just pulled from the the clear blue sky.

$$\frac{exp(b_0 + b_1 X_i)}{1 + exp(b_0 + b_1 X_i)} = \frac{1}{1 + exp(-1 * (b_0 + b_1 X_i))} = \frac{1}{1 + e^{-(b_0 + b_1 X_i)}} \tag{4}$$

This transformation of X_i gives back a very small value if X_i if X_i is really small, and it gives back a value that is close to 1 if X_i is really big. That is called a “logistic curve,” it was used to create Figure 4. In a Logistic regression model, we think of the combined value of the inputs being transformed through a function into a statement about probability.

3.1 Notes about this particular formula:

Here are some highlights.

1. The slope–change in probability resulting from a unit increase in X_i – is $b_1 \cdot P_i \cdot (1 - P_i)$. Hence, the effect of a unit change in X_i depends on the probability. If y_i is very likely to be a 1 or a 0, a change in X_i doesn’t make much difference.

2. The ”odds ratio” is $\frac{P_i}{(1-P_i)}$. It can be shown that

$$\ln\left[\frac{P_i}{1 - P_i}\right] = b_0 + b_1 \cdot X_i \tag{5}$$

That is to say, if you had observations on the probabilities, you could transform them on the left hand side and you would have something that you could estimate with OLS. That is the way logistic regressions were run for a long time, using observed proportions from "grouped data" to calculate the estimates of the probabilities.

3. Some people like to emphasize $exp(b_1)$. In many computer programs that estimate logistic models, the exponential is presented along with the coefficient estimates. Here's why. Note that

$$\left[\frac{P_i}{1-P_i}\right] = exp(b_0 + b_1 \cdot X_i) = exp(b_0) \cdot exp(b_1 X_i) \tag{6}$$

   Suppose that the input variable X_i is a "dummy variable" coded 0 and 1. Then the "overall effect" of X would be summarized by the difference between

$$exp(b_0) \tag{7}$$

   and

$$exp(b_0 + b_1) = exp(b_0) \cdot exp(b_1) \tag{8}$$

   So, in some sense, the difference in the outcome for the two values of X_i boils down to $exp(b_1)$.

   In my experience, people tend to overemphasize $exp(b_1)$ by applying it to input variables that are not dichotomous.

3.2 Estimation

How can the parameters b_0 and b_1 be estimated?

This is a fairly straightforward exercise in maximum likelihood. What is the likelihood that we would observe a given sample of 0's and 1's? Put the observations with 0's first and then the 1's. The first critical assumption is that the observations are statistically independent, meaning the probability of the sample equals the individual probabilities multiplied together. Hence,

$$\begin{aligned} Likelihood\, of\, Sample &= P(y_1 = 0, y_2 = 0, ..., y_m = 0, y_{m+1} = 1, y_{m+2} = 1, ..., y_N = 1) & (9) \\ &= P(y_1 = 0)P(y_2 = 0)\cdots P(y_m = 0)P(y_{m+1} = 1)P(y_{m+2} = 1)\cdots P(y_N = 1) & (10) \end{aligned}$$

This expression is the likelihood function, L, and since the probabilities depend on parameters b_0 and b_1 , we might as well write $L(b_0, b_1)$.

Remembering that the probability that $y_i = 0$ is 1 minus the probability that $y_i = 1$, we can write

$$\begin{aligned} L(b_0, b_1) = &(1 - P(y_1 = 1))(1 - P(y_2 = 1))\cdots(1 - P(y_m = 1)) \\ &\times P(y_{m+1} = 1)P(y_{m+2} = 1)\cdots P(y_N = 1) \end{aligned} \tag{11}$$

This notation can be made a little more compact. It is not necessary to keep writing down the $P(y_i = 1)$ over and over again. Instead, save a little time and effort by writing P_i for this.

The Likelihood function is an impossibly complicated formula because it is composed of numbers that are multiplied together. The multiplication means that none of the components are separable. In contrast, if we work with logarithms, then the product is convertd to a sum. It is mathematically identical to maximize L or the log of L.

$$\ln L(b_0, b_1) = \ln(1-P_1) + \ln(1-P_2) + \cdots + \ln(1-P_m) + \ln(P_{m+1}) + \cdots + \ln(P_N) \tag{12}$$

(fill in the logistic formula $P_i = \frac{1}{1+e^{-(b_0+b_1X_i)}}$ to here to get an idea of where the b_0 and b_1 fit in.)

MLE, short for Maximum Likelihood Estimate, is the choice of estimators b_0, b_1 that maximize the log of the likelihood function. This solution is also a maximizer of L.

3.3 Quick summary of MLE properties.

1. NOT unbiased.

2. MLE's are consistent, asymptotically efficient, asymptotically Normal. The asymptotic normality implies that we can conduct approximate t-tests, as long as we can get estimates of the standard errors of b_0 and b_1.

3. There are ways to calculate asymptotic standard errors. They tell you, approximately, if you had an infinite sample, what the standard error of would be. They are sometimes called approximate standard errors because you never have an infinite sample.

4. Maximum Likelihood Estimation allows an equivalent of the F test.

   A. Let L_0 be the value of the likelihood function in which the "slope" coefficient b_1 (or other coefficients if they are in the model) is 0. Hence, the L_0 is the maximized likelihood when only a constant, b_0, is estimated.

   B. Let L_{max} be the value of the likelihood function at its maximum, when all coefficients, the slope and the intercept, are estimated to maximize the likelihood.

   C. Let λ, (Greek "lambda"), be the ratio of L_0 to L_{max}:

$$\lambda \quad = \frac{L_0}{L_{max.}} \tag{13}$$

   D. It can be shown that $-2 \cdot \ln(\lambda)$ has a χ^2 distribution with k degrees of freedom, where k is the number of "slope" coefficients you estimated (equivalently, the difference in the number of coefficients estimated in calculating L_0 versus L_{max}).

In the next two sections, I describe two approaches to this.

4 GLM approach.

The previous section is the "heres an S shaped curve" approach. The GLM approach would probably be summarized as the "here is a bunch of S-shaped curves" approach. The GLM approach boils down to the the following practical reasoning:

> We need an S-shaped curve. Go get a mathematician who can give you a formula for an S-shaped curve. But be ready. Each mathematician you visit will give you a different formula for an S-shaped curve, and the data will hardly ever give you a reason to prefer one over another.

The part of the formula that blends coefficients and observed inputs, $b_0 + b_1 X_i$, is known as the "linear predictor." It is customary to call it "eta", η_i:

$$\eta_i = b_0 + b_1 X_i.$$

In ordinary least squares regression, of course, $y = b_0 + b_1 X_i + e_i$, so $y_i = \eta_i + e_i$. So the "linear predictor" is the same as "error-free value" in OLS. The predicted value is $\hat{y}_i = \hat{b}_0 + \hat{b}_1 X_i = \hat{\eta}_i$. In the Normally distributed outcome model, where we use OLS, η_i would be playing the role of μ_i in $y_i \sim N(\mu_i, \sigma^2)$.

Our observations are either 0 or 1, and the argument in the previous sections should lead you to conclude that OLS is the wrong approach. We need a formula that connects η_i to a probability value, the chance that we will observe a particular outcome. In the Generalized Linear Model, we think of η_i as a "location" parameter of a distribution (roughly, a "central tendency"). It is causing our probability distribution of outcomes to shift, somehow.

In the GLM, we think of the 0 or 1 observations as coming from a Binomial distribution. Recall if there are N "coin flips" with probability p of "success", then the probability of observing a given number of successes is said to follow the distribution $B(N, p)$. In this case, we think of gathering just one observation at a time, so really we only need $B(1, p)$, which, if you are precise, is a Bernoulli distribution.

The Logit link transforms the probabilities to give back the linear predictor:

$$\eta_i = ln\left[\frac{Prob(y_i = 1|x_i, b)}{1 - Prob(y_i = 1|x_i, b)}\right]$$

Which you can rearrange as

$$P(y_i = 1|x_i, b) = \frac{1}{1 + e^{-\eta}} = \frac{1}{1 + e^{-(b_0 + b_1 x_i)}} \tag{14}$$

From this perspective, a logistic curve is just a particular "S-shaped" curve. For example, the formula with $b_0 = 10$ and $b_1 = 0.115$ is shown in Figure 4.

One can replace the logit link with other functions. Any function that maps from the estimated probabilities back to the range of the linear predictor will do just as well.

In section 3, it was noted that a customized ML estimation routine can be used to estimate this model. One of the major elements of appeal for the GLM approach is that a single algorithm, the Iteratively-Reweighted Least Squares algorithm, can be used to fit all of the members of the GLM family. McCullagh and Nelder proved that the estimates from their algorithm are equivalent to the special-case ML approach described in the previous section.

5 Cumulative Probability Interpretation

Although the GLM approach offers various links like "logit," "cauchit", or "probit", I don't think I can explain why a researcher would use one approach or another. They don't seem to be conceptually related or differentiated. They are just S-Shaped curves.

It is possible to "change gears," however, and understand all of these S shapes from a single, simplifying perspective. And I usually start presentations on that by asking the audience to re-consider the logistic regression proposed in equation 14. Look back at that for a moment, then consider my question:

> Where did the error term go?

Hm. Its almost a Sherlock Holmes mystery, a dog that did not bark in the night. An error term that seemed to vanish. I say "seemed" to vanish, because it did not disappear, it just fulfills a different role.

This is the point at which the econometrician's view of the model becomes helpful. In the "dichotomous" or "binary" outcomes case, one can write a predictive statement that says

$$y_i = \begin{cases} 1 & if\, Z_i = b_0 + b_1 X_i - e_i > 0 \\ 0 & if\, Z_i = b_0 + b_1 X_i - e_i \le 0 \end{cases} \tag{15}$$

There's your missing error term. Think of Z_i as an underlying, unmeasured variable that is linked to the proclivity of case i to reveal an outcome of 1. If this underlying variable exceeds a threshold of 0, then $Y_i = 1$. If Z_i is less than 0, then Y_i takes on the value of 0. If you let Z_i equal the linear function above, the story is complete.

And so the probability that $y_i = 1$ is the same as the probability that

$$e_i < b_0 + b_1 X_i. \tag{16}$$

For a given person with input X_i, the probability of a $Y_i = 1$ is given by the probability that a random variable e_i is smaller than the "target" value $b_0 + b_1 \cdot X_i$.

Eventually you see that is just the cumulative probability distribution function (CDF) for the variable e_i. It is the area under the density function of the error term up to the point $b_0 + b_1 \cdot X_i$. Consider the illustration in Figure 5.

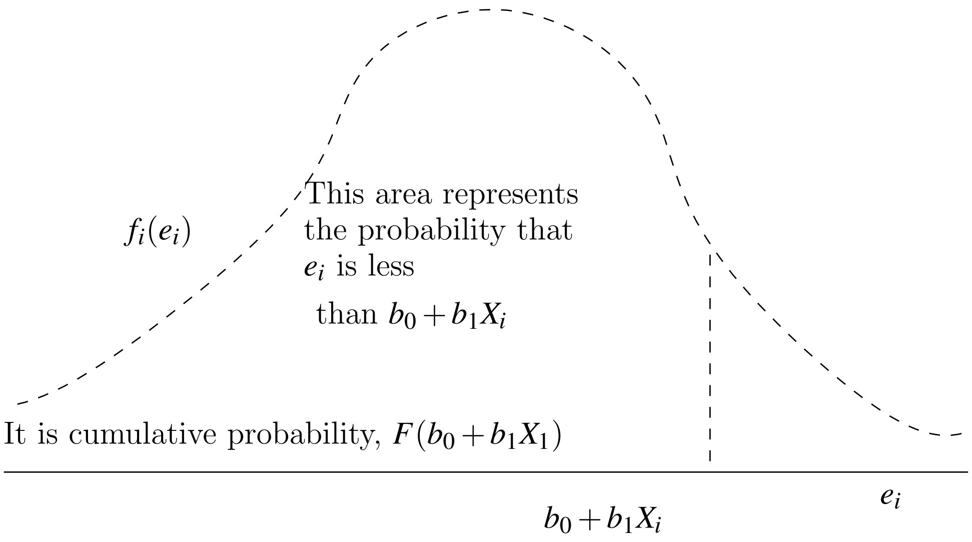
If you said the probability density function (PDF) of the error term was $f(e_i)$, it would be conventional to call the cumulative distribution function $F(\alpha)$, representing the probability that e_i is less than a number α. In this case, the "particular value" α is $b_0 + b_1 X_i$.

Generally speaking, then, we are modeling the upper limit of an integral:

$$Prob(y_i = 1 | X_i, b) = Prob(e_i \le b_0 + b_1 X_i) = F(b_0 + b_1 X_i) = \int_{-\infty}^{b_0 + b_1 X_i} f(e_i) de_i. \tag{17}$$

Any probability distribution will work within this formulation.

Figure 5: The Cumulative Distribution



$f_i(e_i)$
This area represents the probability that e_i is less than $b_0 + b_1X_i$
It is cumulative probability, $F(b_0 + b_1X_1)$
$b_0 + b_1X_i$
e_i

5.1 Logistic Regression

There is a probability distribution called the "logistic" and it happens to have a very workable mathematical form. (In R, run ?rlogis to see for yourself.)

The logistic distribution has a probability density function for a variable "x_i":

$$f(x_i) = \frac{e^{-(x_i-\mu)/\sigma}}{\sigma(1+e^{-(x_i-\mu)/\sigma})^2} \tag{18}$$

The expected value of x_i in this distribution is μ. And the variance is

$$Var(x) = \frac{1}{3}(\pi\sigma)^2$$

As a result, if you compared a Normal distribution with mean μ and variance σ^2 against this Logistic distribution, you could "adjust" the variance of the Logistic to "match up" against the variance of the Normal distribution. The value $\pi^2/3$ is needed to "rescale" the Logistic variance to equal the Normal variance. I have prepared a separate handout on the Logistic distribution. Look in my Distributions folder.

Replace the variable "x_i" by "e_i" to make the PDF match the problem, and we are almost done. Please do not be confused that I have Euler's constant e and the random variable e_i in the same expression (I used x_i above to avoid that confusion, but now I can't).

$$f(e_i) = \frac{e^{-(e_i-\mu)/\sigma}}{\sigma(1+e^{-(e_i-\mu)/\sigma})^2} \tag{19}$$

If e_i is supposed to be a "random disturbance", it seems obvious we have to assume it is "unbiased," in the sense that its expected value is 0. So we suppose that $\mu = 0$. Unlike an OLS regression, we NEVER get to estimate σ, so we just have to set it to a constant value. By custom, we suppress that by "assuming" it is equal to 1.0. That means the PDF of the logistic simplifies to

$$f(e_i) = \frac{e^{-e_i}}{(1+e^{-e_i})^2} \tag{20}$$

The definite integral of this is very simple.

$$F(\alpha) = \int_{-\infty}^{\alpha} \frac{e^{-e_i}}{(1+e^{-e_i})^2} de_i = \frac{e^{\alpha}}{1+e^{\alpha}} = \frac{1}{1+e^{-\alpha}} \tag{21}$$

In the present context, then, the logistic distribution offers a very great simplification. This expression, the integral that represents CDF of the problem under consideration:

$$Prob(y_i = 1|X_i, b) = F(b_0 + b_1 X_i) = \int_{-\infty}^{b_0+b_1X_i} f(e_i) de_i. \tag{22}$$

is easily solved with a relatively simple formula. Use the logistic PDF in place of $f(e_i)$, and $b_0 + b_1 X_i$ is used in place of K, then the whole ugly problem simplifies to the commonly used "Logistic regression" model.

$$P(y_i = 1|X_i, b_i) = \frac{e^{b_0+b_1X_i}}{1+e^{b_0+b_1X_i}} = \frac{1}{1+e^{-(b_0+b_1X_i)}} \tag{23}$$

5.2 Probit Regression

On the other hand, if you suppose that e_i is Normally distributed. In that case, the formula in expression 22 is not mathematically simplified. There is no simple formula for the definite integral:

$$P(y_i = 1|X_i, b_i) = \int_{-\infty}^{b_0+b_1X_i} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(e_i-\mu)^2}{2\sigma^2}} de_i \tag{24}$$

One assumes that $\mu = 0$. The "scale parameter" σ^2 cannot be estimated (it is "unidentified"). It is set equal to 1. Setting the value of μ or σ^2 is not thought to have substantive significance because the values of the estimated b coefficients will scale up and down accordingly. There's a theorem in mathematical statistics concerning the "Change of Variables" that deals with that.

Because the probit equation 24 is such a complicated thing to write down, articles in social science almost always avoid it, instead adopting some complicated-looking symbol, such as Φ, to refer to the cumulative probability distribution. One sees expressions such as

$$P(y_i = 1|X_i, b_i) = \Phi(b_0 + b_1 X_i) \tag{25}$$

We don't usually think about this in the two-category model, but the probability of observing a 0 is

$$P(y_i = 0|X_i, b_i) = 1 - \Phi(b_0 + b_1 X_i) \tag{26}$$

5.3 Does it matter if you use Logit or Probit?

Not very much. The PDF of the Logistic and the Normal distributions overlap to a considerable extent. If we adjust the parameter σ of the logistic so that the variance of the Logistic and the Normal distributions are the same, then we can see that we are talking about 2 distributions that barely differ. In Figure **??**, one finds a Logistic distribution with $\sigma = 1$, and so for both distributions, the expected value is 0 and the standard deviation is $\pi/\sqrt{3}$.

6 Testing Statistical Significance

6.1 z-test and t-test: two approximations

Asymptotically, $\hat{b}$ is Normal (recall fundamental ML Theory). We have the standard error. What about the quantity:

$$\frac{\hat{b} - b_{null}}{s.e.(\hat{b})}$$

it looks like a $t-test$, doesn't it? I've seen some people/programs call it a $t-test$, but that's not a widespread option these days. Instead, it is more usual to say it is a z statistic, approximately Normally distributed

$$z = \frac{\hat{b} - b_{null}}{s.e.(\hat{b})} \tag{27}$$

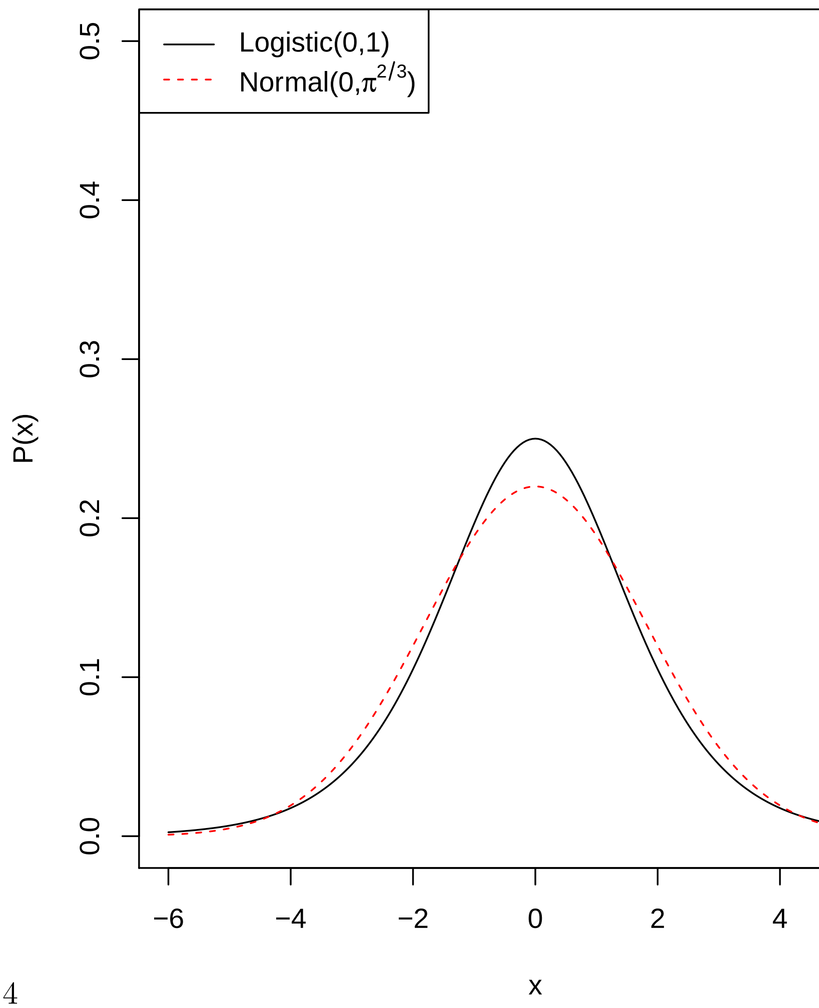
6.2 χ^2 test: another approximation

Most programs today will either report a z test or the Wald Chi-Square. The Wald Chi-square is the ratio of the squared estimate to the variance,$\hat{b}^2/Var(\hat{b})$. But alert users will note that it is simply z^2.

$$z^2 = \left(\frac{\hat{b}}{s.e.(\hat{b})}\right)^2 \tag{28}$$

Wald contended that value is distributed as a χ^2 variable. The square root $\hat{b}/se(\hat{b})$

Figure 6: Compare Logistic and Normal



is approximately Normal (and also approximately t-distributed). That explains why some programs call the statistic $\hat{b}/se(\hat{b})$ a t variable, while others call it a Z statistic. Either way, the information is the same. That is called a Wald Chi-Square statistic in most programs.

From elementary statistics, we know that the square root of a χ^2 variable with one degree of freedom is Normal(0,1), so the Wald Chi-Square test for a single parameter is actually substantively IDENTICAL to the t or z approaches.

The Wald Chi-Square can be used to simultaneously test several coefficients.

$$\hat{b}Var(\hat{b})^{-1}\hat{b}$$

Note that if we were testing only one parameter, this degenerates to the preceeding equation.

6.3 How to estimate the $s.e.(\hat{b})$?

This idea depends on your understanding of maximum likelihood theory. It is probably enough for beginners to believe that we can calculate standard errors as estimates of the standard deviation of sampling distributions.

Nevertheless, here's something worth noting and its not too complicated. This insight was gathered from literature on the Generalized Linear Model.

For the ordinary logistic regression model, the covariance-variance matrix is

$$Var(\hat{b}) = \left[(X'WX\right]^{-1} \tag{29}$$

which is quite reminiscent of the OLS formula. In fact, it would be exactly the same if not for the fact that the cases have unique variances.

The matrix in the middle, W, has the variance of the individual cases, is like this:

$$W = \begin{bmatrix} P_1(1-P_1) & 0 & 0 & 0 & 0 \\ 0 & P_2(1-P_2) & 0 & 0 & 0 \\ 0 & \cdots & & & \\ 0 & \cdots & & & \\ 0 & 0 & 0 & 0 & P_N(1-P_N) \end{bmatrix} \tag{30}$$

We don't know the "true variances" because we don't know the true probabilities. But we can use approximations from fitted models to get those values.

In the generalized linear model literature, they describe the logistic regression as a binomial-distributed dependent variable in which the probability of observing an "event" on a particular random draw is given by $P = 1/(1+e^{-Xb})$. In the days before high-speed computers were so freely available, it helped a lot to have an algorithm with which to estimate models that did not require a full maximum likelihood optimizer. The GLM folks found that an interative weighted least squares procedure could produce estimates that were equivalent to Maximum Likelihood (for the restricted class of distributions inside the GLM terminology). As one iterates, one calculates the predicted values for the cases, and so $\hat{P}$ values are obtained that can be used in the formula for the variance-covariance matrix.

7 Diagnostics: Goodness of Fit, Deviance, etc

I'm not wasting any breath on pseudo-R^2s.

7.1 Do you have the right model?

Should you remove some variables from your model? Perhaps all of them

Some programs report a "model chi-square", equivalent to $-2ln(LikelihoodRatio)$ or $-2LLR$, which is the difference between $-2ln(Likelihood\,of\,fitted\,model)$ and $-2ln(Likelihood\,of\,constant-only\,model)$.

The likelihood ratio test can be used to compare any two models that are estimated ON THE SAME DATA. Comparing a full model against a subset, the following can be used to test the hypothesis that the coefficients for a set of variables are all equal to zero.

$$= -2[lnL_{restricted} - lnL_{full}] = -2ln\left[\frac{L_{restricted}}{L_{full}}\right] \sim \chi^2_{diff} \tag{31}$$

where $diff$ is the difference in the number of fitted parameters, or "excluded variables" from the restricted model.

7.1.1 How "good" is your likelihood? Deviance

I have learned of another way that the likelihood value is put to use. This is based on the concept of **deviance**.

Deviance is a benchmark, which in most models is equal to$-2ln(Likelihood\,of\,fitted\,model)$. In the fine print, one finds out that deviance is the difference between $-2ln(Likelihood\,of\,fitted\,model)$ and $-2ln(Likelihood\,of\,saturated\,model)$. But the latter value is usually 0: R reports deviance values that are scaled to so that the saturated model has Likelihood of 1 (and hence $-2ln(L) = 0$).

A "saturated model" is one in which we are allowed to make a unique estimate of $\widehat{P_i}$ for each unique combination of the input variables. That means, for each combination of input variables, we can make a customized prediction. The likelihood obtained by such a model is surely the best we could possibly get, right? Its not possible to use more degrees of freedom. The "best possible" deviance value is obtained by fitting a parameter for each case in the dataset, which would reduce $-2ln(Likelihood)$ to 0.

Get the likelihood for the saturate model, call that L_{sat}. Then fit a model using some independent variables. Call that likelihood L_{fit}. The difference between these two likelihood values is an indicator of "how bad" your model is when compared against the saturated model. In fact, asymptotically, the deviance indicator:

$$deviance = -2[lnL_{fit} - lnL_{sat}] = -2ln\left[\frac{L_{fit}}{L_{sat}}\right] \tag{32}$$

is distributed approximately as a χ^2 statistic with degrees of freedom equal to sample size (N) minus the number of parameters.

If each observation in the dataset has a unique set of values for the inputs, then the saturated model has a log likelihood of 0. See why?

Myers, Montgomery, and Vining (2002) observe, (I'm paraphrasing notation here to match the above notation) "Formally, an insignificant value of (deviance) in a one-tailed test implies that the fit of the model is not significantly worse than that of the saturated model. ... Often the rule of thumb is applied that the quality of fit is reasonable if $\frac{deviance}{N-p}$ is not appreciably larger than 1. The rule of thumb is prompted by the fact that N-p is the mean of the χ^2_{N-p} distribution"(p. 113).

In some articles, the idea of deviance is applied mechanically as a test of the quality of a model.

7.1.2 Hosmer and Lemeshow test

Proceed as follows. Calculate predicted values, $\hat{P}_i$ for all cases. Sort them from low to high. Then subdivide the sample into subgroups. Then find out if the observed frequency of 1's and 0's matches the estimated probabilities from the model.

Pick some pleasant number of subgroups, say 10. For each subgroup, one can calculate the observed "success rate" O_i and an expected (from the model) success rate, and the traditional χ^2 test is used to find out if the model is grossly out-of-whack.

$$homer.and.lemeshow^2_{\chi} = \sum_{i=1}^{10}\left[\frac{(O_i - E_i)^2}{E_i}\right] \tag{33}$$

If the χ^2 value is extreme by that standard, it means that your predicted probabilities do not match the observations very well.

That is informative, but not too informative. It does not tell you if the model is "off" for any particular reason, and there could be many suspects in your search for the criminal.

7.1.3 Overdispersion

Suppose your deviance is very high. That means your model does not fit as well as you might like, and one of the frequent causes is that the observed variance of scores is higher than the model would predict. If the model is wrongly specified, then the deviance is big.

Another explanation is that the assumed model for the variance of outcomes is wrong. The binomial variance is $P_i(1-P_1)$, and so when the predicted value is 0.99, it means we should almost never observe a score of 0. However, in "real data", we often do. Sometimes I have seen this called "extrabinomial variation".

The effects of overdispersion are the same as heteroskedasticity in the OLS model. The estimated variance-covariance matrix of the $\hat{b}$'s does not match the true variance of the $\hat{b}$'s.

If one has “grouped data”, that is, many observations for each value of the input variables, then there is a correction which can be applied that is exactly analogous to feasible weighted least squares.

If one does not have grouped data, but rather all individual level data, then perhaps a more dramatic departure/respecification is needed. For example, instead of assuming that all cases with the same inputs should have the same predicted probability, suppose instead there is an unmeasured error term that affects the subgroups in the dataset. SO, if one started with 25, one could insert an error term, a random component that afflicts the members of group j:

$$P(y_i = 1|X_i, b_i) = \Phi(b_0 + b_1 X_i + e_j) \tag{34}$$

This model is a mixed model, one in which there is a random coefficient that should be taken into account. In R, it is covered in the lmer procedure of lme4 package or the GLMM package (which is simpler to use at the current time)

8 Calculate predicted values and use them to interpret and present results

In the past, I have calculated predicted values “manually” by taking the estimated coefficients and using them in the logistic equations,

$$P(Y_i = 1) = \frac{1}{1 + e^{-(b_0 + b_1 X_i)}}$$

It is tedious, but instructive, to calculate that. That can be done in R in various ways.