Modelling Binary Outcomes

12/12/2023

Contents

0.1 Cross-tabulation

If we are interested in the association between two binary variables, for example the presence or absence of a given disease and the presence or absence of a given exposure. Then we can simply count the number of subjects with the exposure and the disease; those with the exposure but not the disease, those without the exposure who have the disease and those without the exposure who do not have the disease. We can than put these numbers into a 2 × 2 table, as shown in Table 0.1.

	Exposed	Unexposed	Total
Cases	a	b	a + b
Controls	c	d	c + d
Total	a + c	b + d	a + b + c + d

Table 0.1: Presentation of data in a two-by-two table

If our sample has been randomly selected from a given population, then the proportions in each cell (i.e. $a/(a+b+c+d)$ etc.) will differ from the proportions in the population only by random variation. However, there are two other widely used sampling schemes for which this is not the case. Both sampling schemes can be thought of as stratified samples, in which subjects are sampled from two different strata.

The first scheme is exposure-based sampling, in which a fixed number of exposed subjects and a fixed number of unexposed subjects are sampled. In this case, the prevalence of the disease in the exposed and unexposed subjects ($a/(a+c)$ and $b/(b+d)$) are unaffected, but the proportion of exposed subjects is fixed by the sampling scheme, and need not reflect the proportion in the population[a].

The alternative is outcome-based sampling, often referred to as a case-control study, in which we sample a fixed number of cases $(a+b)$ and a fixed number of controls $(c+d)$. In this case, the prevalence of the disease in our sample $((a+b)/(a+b+c+d)$ is fixed by the design, and does not reflect the prevalence in the population.

If there is no association between the exposure and disease, we would expect the prevalence of the disease to be the same in the exposed as it is in the unexposed. Since $a+b$ subjects have the disease, the overall prevalence of disease is $\frac{a+b}{a+b+c+d}$. There are $a+c$ exposed subjects, so we would expect $\frac{(a+b)\times(a+c)}{a+b+c+d}$ subjects who are exposed and ill. This is the expected value of a, under the null hypothesis that the prevalence does not vary with the exposure, and this works for all sampling schemes.

We can calculate expected values for b, c and d in a similar fashion. If the observed values are sufficiently far from their expected values under the null hypothesis, we can conclude that the null hypothesis is unlikely to be true. This can be done by calculating $\frac{(O-E)^2}{E}$ for each cell of the table (where O is the observed value and E is the expected value) and summing them for the four cells in the table. This sum will follow a χ^2 distribution on 2 degree of freedom if the null hypothesis is true, so a P-value can be obtained as the probablity of the observing a higher value from this distribution. This hypothesis test is referred to as the χ^2-test.

To perform a χ^2-test in stata, the command to use is `tabulate`, with the `chi2` option. So if the variable `exposure` contains the exposure data and `disease` contains the disease information, the full command for a χ^2-test is

[a] In fact, it generally won't, because the reason for sampling in this way is usually that the exposure is rare, and we need to artificially increase the number of exposed subjects in our sample in order to increase our power.

```
tabulate exposure disease, chi2
```

0.1.1 Measures of Effect

Rather than simply testing the null hypothesis that the exposure does not affect the outcome, we may wish to quantify the effect. Most commonly, we are interested in the relative risk and its confidence interval, but this is not the only possibility. For example, if we are particularly interested in the absolute risk of a particular outcome, we can use the risk difference (i.e. the difference between the prevalence in the unexposed and the prevalence in the exposed.

However, if we have outcome-based sampling, the relative risk and the risk difference are meaningless, because the overall prevalence of the disease in our sample is fixed by the design, and does not reflect the prevalence in the population. However, the odds ratio will take the same value for any of the sampling schemes, and hence it can still be used with outcome-based sampling.

In stata, the relative risk and risk difference are most easily obtained using the command

```
cs disease_var exposure_var
```

The odds ratio can be obtained from the command by adding the option `or` at the end of the command.

$$\begin{aligned}
\text{Relative Risk} &= \frac{\frac{a}{a+c}}{\frac{b}{b+d}} = \frac{a(c+d)}{c(a+b)} \\
\text{Risk Difference} &= \frac{a}{a+c} - \frac{b}{b+d} \\
\text{Odds Ratio} &= \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{cb}
\end{aligned}$$

0.1.2 Limitations of Tabulation

Whilst tabulation can be useful when initially exploring data, it is not sufficient for complex analysis. This is primarily because continuous variables cannot be included in the analysis, only categorical variables. In addition, only a single variable can be analysed at a time. There are more complex tabulations that can be used to allow for more than one variable, but these also have drawbacks: the more variables are included in the model, the more “sparse” the tables become (i.e. the fewer observations in each cell of the table), and the inference from the table becomes less robust.

0.2 Modelling Approaches

0.2.1 Linear Regression and dichotomous outcomes

To understand what we are doing with logistic regression, consider an example taken from [1]. In this example, we wish to predict the probability that a person has coronary heart disease (CHD) from their age. A sccatter plot of CHD against age is, at first sight, uninformative: see Figure 0.1

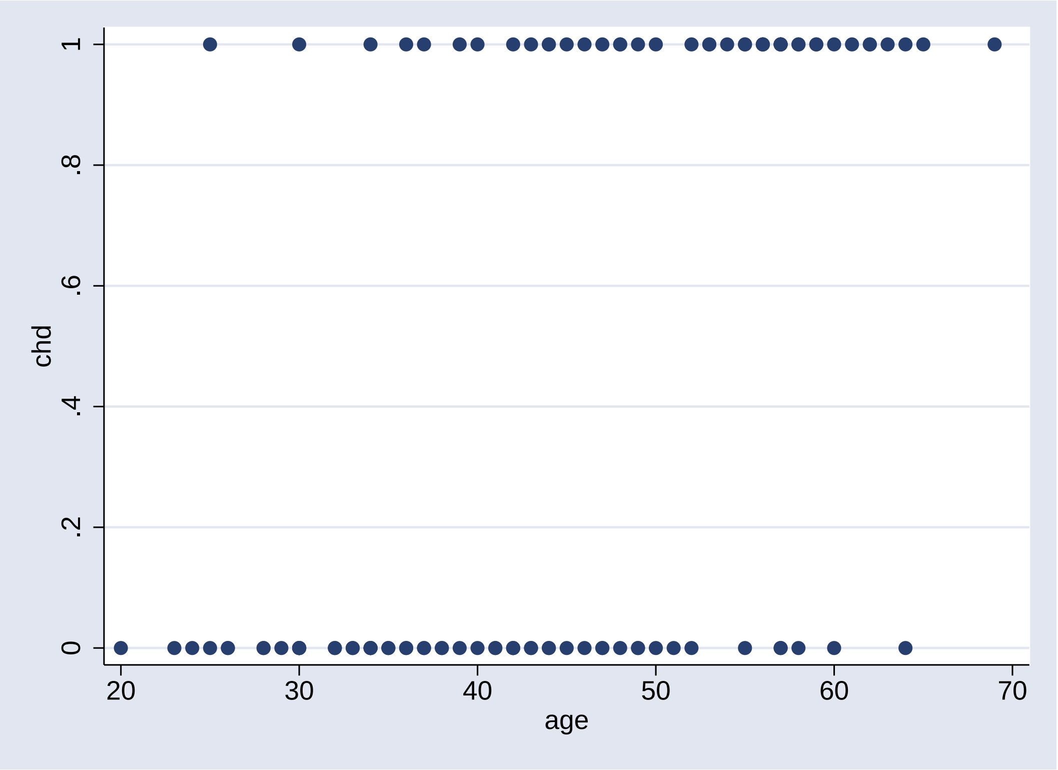


Figure 0.1: Scatter plot of CHD against against age

This illustrates one of the problems with using a linear model for a dichotomous outcome: the outcome variable Y is clearly not normally distributed. It can only take one of two values: 0 or 1. Hence, to model this data, we will need to use a different error distribution.

However, the scatter plot does suggest that those with CHD tend to be older than those without. We can confirm this by dividing subjects up into age groups, and plotting the proportion of subjects with CHD in each group against the mean age in that group. The results are illustrated in Figure 0.2, which confirms that the probability of CHD increases as age increases.

Notice that although the observed values of Y are either 0 or 1, the predicted outcome ($\hat{Y}$) can take any value between 0 and 1. For example, there are 10 subjects between 20 and 30, of whom 1 had CHD. Therefore, the proportion of subjects with CHD is 0.1, although there is nobody in the sample with a value of 0.1 for the CHD variable. $\hat{Y}$ can also be thought of as the mean of Y in the subgroup: since 1 subject has $Y = 1$ and 9 have $Y = 0$, the mean of $Y = (9 \times 0 + 1 \times 1)/10 = 0.1$.

Notice also that the estimates for each age group do not lie on a straight line: the line becomes flatter as the proportion of subjects with CHD approaches 0 or 1. This is common when there is a strong association between a continuous risk factor and a dichotomous outcome: the strong association means that the slope is steep when the risk is around 0.5, but it has to flatten out as it approaches 0 and 1 so that it does not exceed these values. If we were to fit a straight line to this data, the predicted probability of CHD would be less than 0 for the very youngest subjects in this dataset, as shown in Figure 0.3.

Clearly, we cannot use the linear regression model for this data, since this would give predicted values ranging from $-\infty$ to ∞, and even within the age range we are considering it would

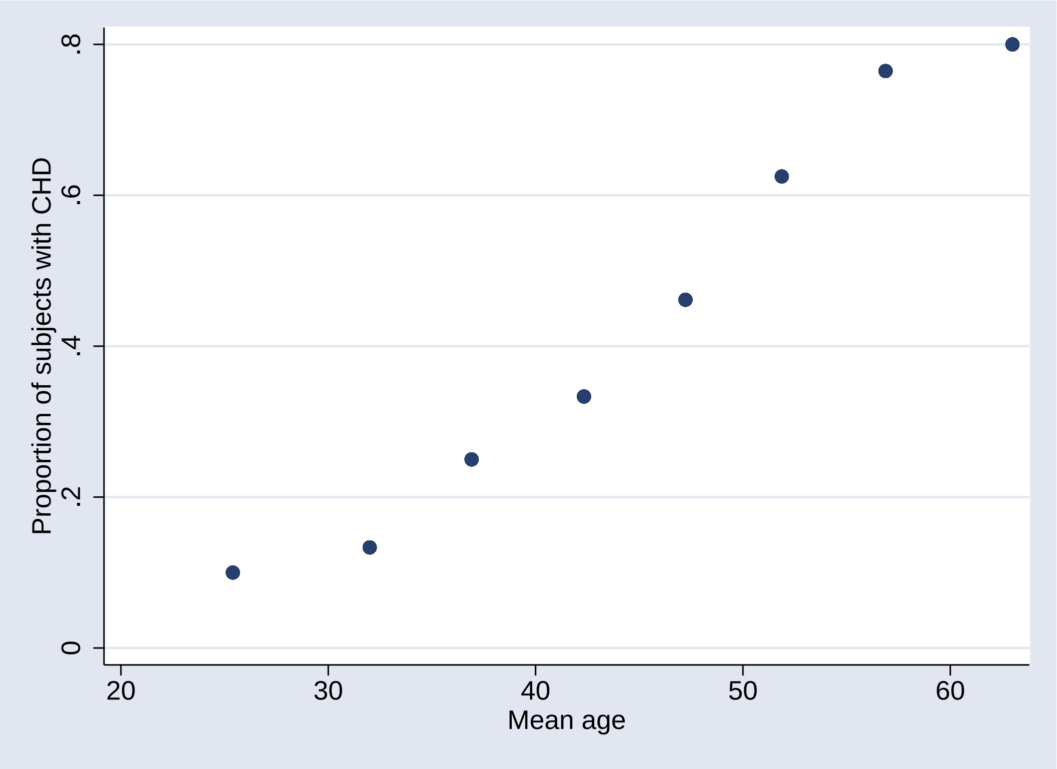


Figure 0.2: Scatter plot of proportion of CHD against against mean age

produce meaningless predicted values. In order to find a suitable model, we need to consider the relationship between probabilities and odds.

0.2.2 Probabilities and Odds

A probability is a number between 0 and 1 (inclusive): 0 means the event in question never happens, 1 means it always happens, and 0.5 means it happens half of the time. Another scale that is useful for measuring probabilities is the odds scale, familiar to those who like betting on the horses. If the probability of an event occuring in p, then the odds (Ω) of it occuring are $p : 1 - p$, which is often written as a fraction $\Omega = p/(1 - p)$. Hence if the probability of an event is 0.5, the odds are 1:1, whilst if the probability is 0.1, the odds are 1:9.

Why is the odds scale useful ? Primarily because it can take any value from 0 to ∞. Both p and $1 - p$ have to be positive, so $p/(1 - p)$ must be positive. As $p \to 0$, $p/(1 - p) \to 0$, whilst as $p \to 1$, $1 - p$ gets extremely small so $p/(1 - p)$ gets extremely large: for example, if $p = 0.99, \Omega = 0.99/0.01 = 99$, whilst if $p = 0.999, \Omega = 0.999/0.001 = 999$.

So, if p ranges from 0 to 1, $p/(1 - p)$ ranges from 0 to ∞. If we now take logs, then $\log(p/(1 - p))$ ranges from $-\infty$ to ∞. Hence, we can model

$$g(p) = \log\left(\frac{p}{(1 - p)}\right) \tag{0.1}$$

rather than p: as $g(p)$ goes from $-\infty$ to ∞, p goes from 0 to 1. Figure 0.4 shows the relationship between p and $g(p)$: you will see that the shape of this curve is very similar to that in Figure

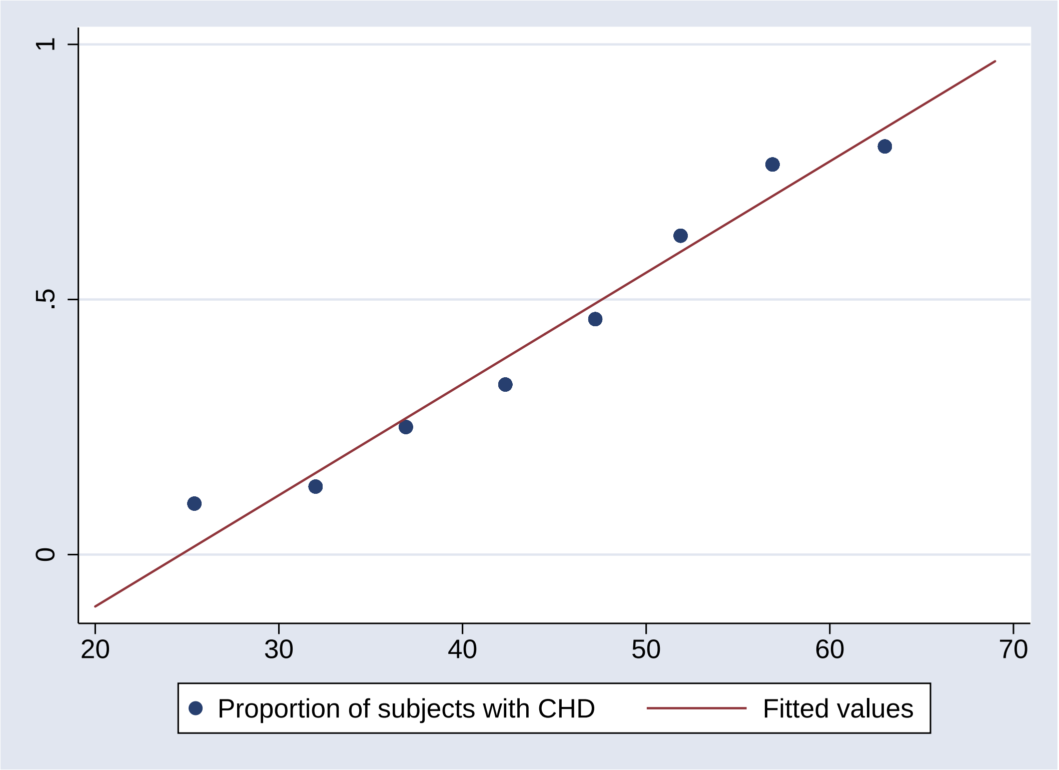


Figure 0.3: Scatter plot of proportion of CHD against against mean age with overlaid linear model

0.2, suggesting that using this link would give a good fit to that data. It is also symmetric, since if the odds of an event happening are Ω, the odds of it not happening are $1/\Omega$.

It is important to remember that there is a one-to-one correspondance between odds and probabilities: if you know the probability (p) of an event occurring, you can calculate the odds ($\Omega = p/(1-p)$), and if you know the odds you can calculated the probability ($p = \Omega/(1+\Omega)$). So the odds scale is just a different scale that can be used to measure probabilities.

0.2.3 The Binomial Distribution

Suppose you toss a coin 10 times: what is the probability that you get a) 10 heads, b) 5 heads, c) no heads. Intuitively, you think that getting 5 heads should happen more often than getting 10 or none. Getting 10 heads would be seen as being unusual, but not impossible. However, getting 4 or 6 heads would be seen as less unusual.

The binomial distribution can be used to determine how unusual getting 10 heads would be in this case. It requires two parameters: the probability of a success (in this case 0.5, assuming that the coin we use is fair), and the number of trials that we carry out (in this case 10). The number of heads we get has to be a whole number from 0 to 10 inclusive, and from the binomial distribution we can calculate how like each of these outcomes is.

The binomial distribution is appropriate whenever:

1. the outcome we are interested in is dichotomous (we can think of it as a success or a failure); and

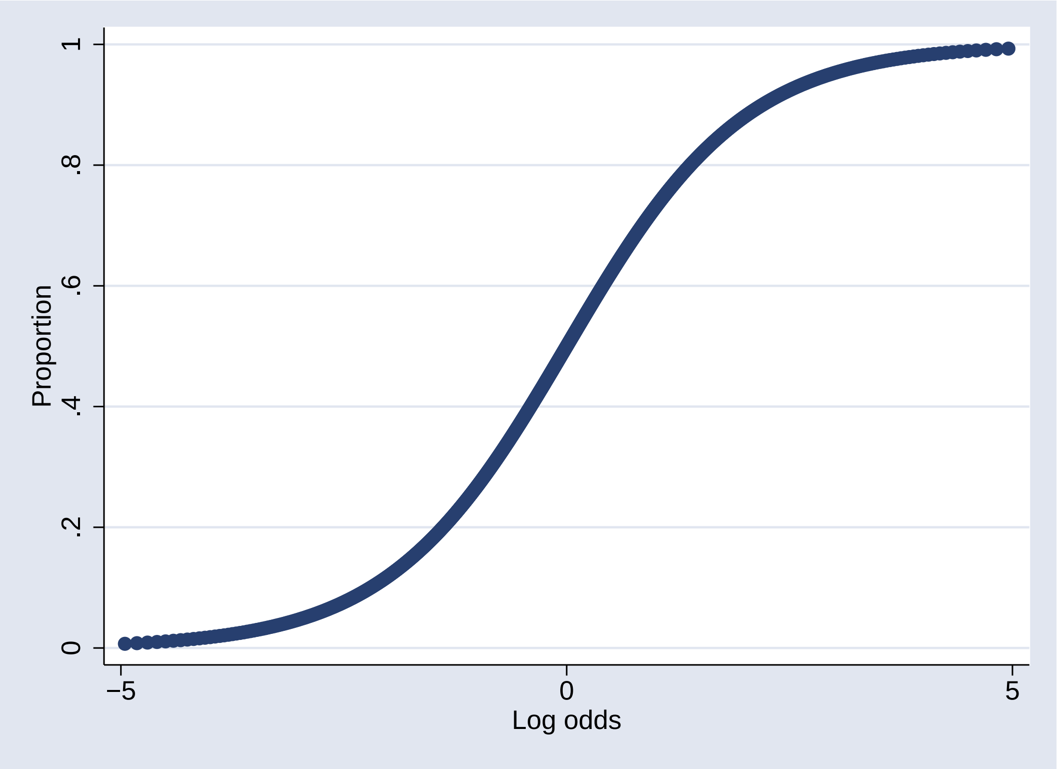


Figure 0.4: Relationship between proportion and log-odds

2. we are considering a number of independent trials.

Therefore, the binomial distribution is appropriate to use as an error distribution in logistic regression.

0.2.4 The Logistic Regression Model

So, the logistic regression model is

$$\begin{aligned} \log\left(\frac{\hat{\pi}}{(1-\hat{\pi})}\right) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \\ Y &\sim \text{Binomial}(\hat{\pi}) \end{aligned}$$

where $\hat{\pi}$ is the predicted probability that $Y = 1$, given the values of $x_1 \ldots x_p$.

Parameter Interpretation

In a simple linear model $Y = \beta_0 + \beta_1 x$, if x increases by 1, Y increases by β_1. In a logistic regression model, it is $\log(\hat{\pi}/(1-\hat{\pi})$ which increases by β_1. What does this mean in the real world ?

Suppose that the predicted probability of the event of interest is p_0 when $x = 0$ and p_1 when $x = 1$, Then we have

$$\log\left(\frac{\hat{p_0}}{1-\hat{p_0}}\right) = \beta_0$$

and

$$\log\left(\frac{\hat{p_1}}{1-\hat{p_1}}\right) = \beta_0 + \beta_1$$

So

$$\log\left(\frac{\hat{p_1}}{1-\hat{p_1}}\right) = \log\left(\frac{\hat{p_0}}{1-\hat{p_0}}\right) + \beta_1$$

If we exponentiate both sides of this equation we get

$$e^{\log\left(\frac{\hat{p_1}}{1-\hat{p_1}}\right)} = e^{\log\left(\frac{\hat{p_0}}{1-\hat{p_0}}\right)+\beta_1}$$

which simplifies to

$$\frac{\hat{p_1}}{1-\hat{p_1}} = \frac{\hat{p_0}}{1-\hat{p_0}} \times e^{\beta_1} \qquad (0.2)$$

In other words, when x increases by 1, the odds of a positive outcome increase by a factor of e^{β}_{1}. Hence e^{β}_{1} is called the *odds ratio for a unit increase in x*.

The interpretation of β_0 is slightly different, since there is no variable associated with this coefficient. However, we can say that if $x = 0$, $\log\left(\frac{\hat{\pi}}{(1-\hat{\pi})}\right) = \beta_0$, so $\frac{\hat{\pi}}{(1-\hat{\pi})} = e^{\beta_0}$. In other words, β_0 is the log of the odds of a positive outcome when all of the predictor variables take the value 0 (it is analogous to the intercept in a linear model, which is the value taken by Y when all the predictor variables are 0).

Odds Ratios and Relative risks If p is small, then $\Omega \approx p$, since $1 - p \approx 1$. Thus, odds are very similar to probabilities provided that p is small, and odds ratios are then very similar to relative risks. However, if the outcome of interest is more common, then the difference is greater. Figure 0.5shows a plot of Ω against p, showing clearly how similar they are for small p but how they diverge increasingly as p increases.

0.2.5 Logistic Regression in Stata

There are two commands for performing logistic regression in stata, `logistic` and `logit`. They are almost identical: the only difference is that by default, `logistic` produces a table of odds ratios whilst `logit` produces a table of coefficients. However, either can be used, and there are options to get coefficients from `logistic` and odds ratios from `logit`.

The basic syntax for both commands is the same:

`logistic` *depvar* [*varlist*]

where ***depvar*** is the outcome variable and ***varlist*** are the predictors. Stata assumes that *depvar* takes the value 0 if the event of interest did not take place and 1 if it did[b]. The variables in *varlist* can be categorical (if you use the construction `logistic depvar i.indepvar`), continuous or a mixture of the two.

Here is an example of performing logistic regression using stata. We will look at the same example as at the beginning of this chapter, using age to predict the presence of coronary heart disease, with the data taken from [1]. The results of running the command `logistic chd age` are given below:

[b] In fact, it will take any value not equal to 0 or missing as meaning the event took place, but that is not usually a sensible way of coding your data

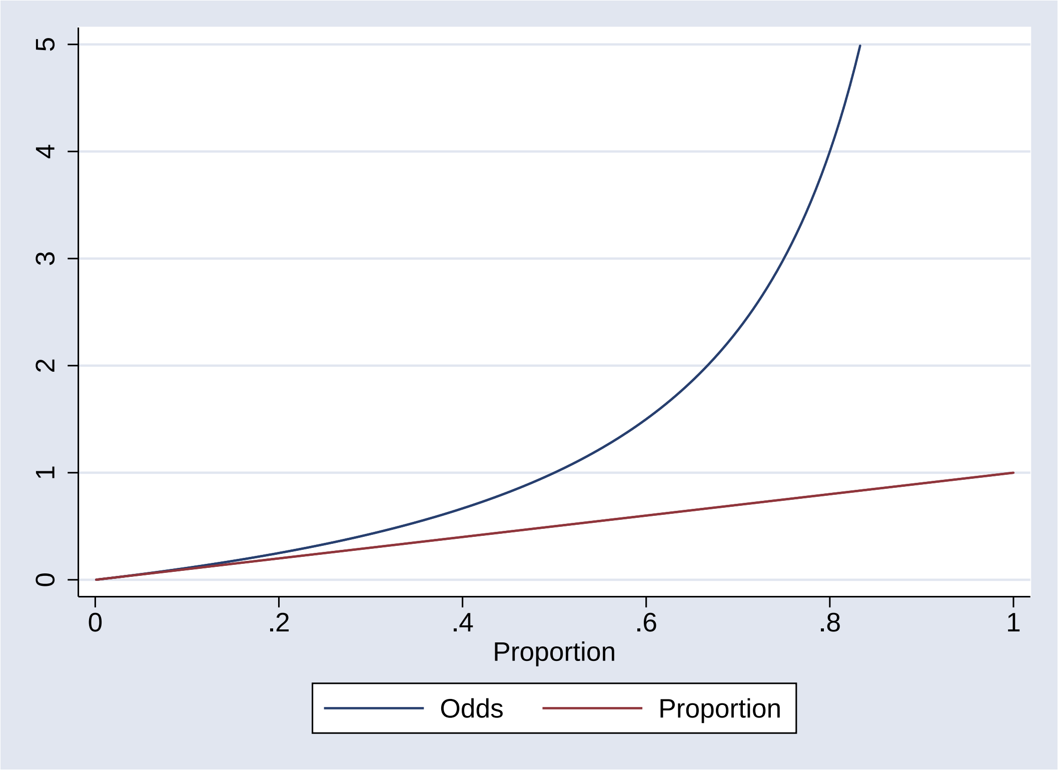


Figure 0.5: Comparison of Odds and Probability

```
. logistic chd age
Logistic regression                               Number of obs   =        100
                                                  LR chi2(1)      =      29.31
                                                  Prob > chi2     =     0.0000
Log likelihood = -53.676546                       Pseudo R2       =     0.2145

         chd   Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]

         age     1.117307   .0268822     4.61   0.000     1.065842    1.171257
```

This is very similar to the output from linear regression wich you are familiar with. There is no ANOVA table, since that is not meaningful in logistic regression. There is an overall model likelihood ratio χ^2-statistic (29.31) and its corresponding p-value at the top. R^2 cannot be calculated for a logistic regression model, but there are a number of surrogates that can be used, one of which is given by stata as the "Pseudo R2". More details can be found in section 0.3.2

Then there is a table of odds ratios. In this case, it only contains a single entry, since we only have a single predictor variable. The odds ratio is given as 1.12, with a 95% confidence interval of (1.07, 1.17). This means that for each year increase in age, the odds of CHD increase by a factor of 1.12. There is also a z-statistic for testing the null hypothesis that the odds ratio is 1, and the result of that significance test ($p = 0.000$, the effect of age is highly significant).

If we were to use the `coef` option of the logistic command, we would get the coefficients of the logistic regression model rather than odds ratios:

```
. logistic chd age, coef

Logistic regression                               Number of obs   =        100
                                                  LR chi2(1)      =      29.31
                                                  Prob > chi2     =     0.0000
Log likelihood = -53.676546                       Pseudo R2       =     0.2145

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1109211   .0240598     4.61   0.000     .0637647    .1580776
       _cons |  -5.309453   1.133655    -4.68   0.000    -7.531376   -3.087531
------------------------------------------------------------------------------
```

The top part of the output is identical: only the table of coefficients is different. We now have two coefficients: `age` and `_cons`. The reason that we did not have `_cons` in the previous is example is that this coefficient corresponds to β_0, and, as we have seen, e^{β_0} is not an odds ratio, it is simply the odds of having CHD when age is 0, and it was therefore not appropriate to put it in a column labelled "Odds Ratios". (As with linear regression, the value of β_0 is often not of interest, since it lies outside the range of the data, and corresponds to a situation in which we are not interested (in this case, the prevalence of CHD in newborn babies)).

Using `predict` after `logistic`

Having fitted our logistic regression model, we can use the `predict` command to obtain additional variables in our dataset. There are a variety of diagnostic statistics which can be obtained: we will meet some of them in Section 0.3.3. For now, we will only consider two options: the predicted probability and the linear predictor.

The predicted probability can be generated using the command `predict it varname, p`. Lets do this with the CHD data. If we enter the commands

```
logistic chd age
predict pred, p
graph twoway scatter pred age
```

we will get a scatter plot of the predicted probability of CHD at each age, which should look something like Figure 0.6 (I have overlaid the proportion of subjects in each age-band, as shown in figure 0.2)

The change in the probability of having CHD as age increases is clearly non-linear: it increases slowly at young ages, more rapidly arounnd ages 40–50 and slowly again thereafter. This echoes what we see when we look at the prevalence by ageband.

However, this non-linearity is a consequence of the link function that we have chosen. By using the logit link, we are assuming that the log of the odds of CHD *does* increase linearly with age. We can see that if we obtain the linear predictor from the predict command using the option `xb`, and plot that against age, as seen in Figure 0.7:

```
predict lp, xb
graph twoway scatter lp age
```

0.2.6 Other Possible Models for Proportions

There are many advantages to using the logit link for modelling proportions, but it is not the only possible link function.

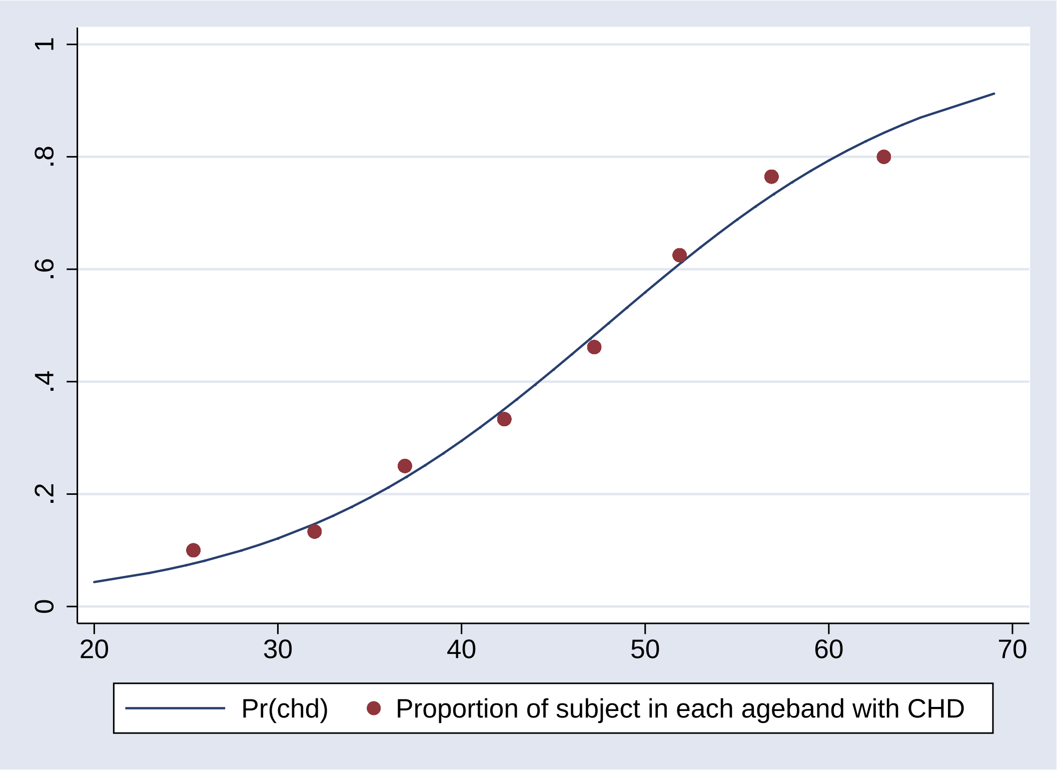


Figure 0.6: Scatter plot of predicted probability of CHD against age

0.2.7 Log-binomial

One perceived drawback of the logit link is that the model is linear in the log of the odds of the outcome, and the model coefficients can be transformed into odds ratios. People are more familiar with using probabilities than odds, and would prefer to have relative risks than odds ratios. Unfortunately, the logit link can only provide odds ratios (although, as we have seen, these are numerically close to relative risks if the prevalence is low).

In order to obtain relative risks, we need to use a log link, i.e. model

$$\log \hat{\pi} = \beta_0 + \beta_1 x$$

When x increases by one, we add β to $\log \hat{\pi}$, which is equivalent to multiplying $\hat{\pi}$ by e^{β}.

This can be done, and such a model, with a log link and and a binomial error distribution, is called a log-binomial model. Such models are gaining in popularity in epidemiology, but there are a number of concerns which should be borne in mind:

1. If $\log \hat{\pi}$ can take any value from $-\infty$ to ∞, $\hat{\pi}$ can take any value from 0 to ∞. So a log-binomial model can produce predicted values that are greater than 1 (but not less than 0)

2. A logistic regression model assumes that the probability of a positive outcome increases linearly on a logit scale, whilst a log-binomial model assumes a linear increase on a log scale. If all of the predicted probabilities are small, then there will be little difference between the logistic regression model and the log-binomial model. However, if the outcome is common,

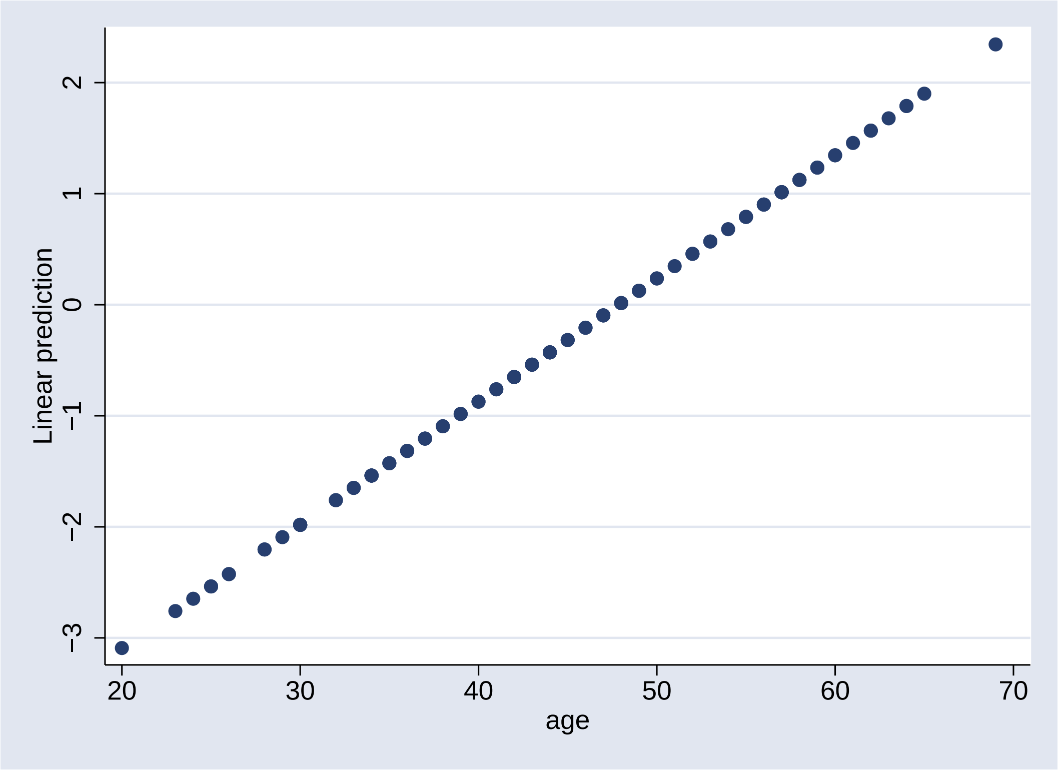


Figure 0.7: Scatter plot of linear predictor from logistic regression model for CHD against age

the can be a considerable difference in the predicted probabilities from the two models, and it will be an empirical decision which model fits better.

3. The log-binomial model is not symmetric. That is, if q is the probability that the event in question does not take place, and we estimate q using a log-binomial model, it is not true that $\hat{q} = 1 - \hat{\pi}$, which would be the case if we used logistic regression. If the outcome in question is common, it may be necessary to model q, since modelling p could produce predicted probabilities greater than 1.

The `glm` command can be used in stata to fit a log-binomial model (in fact, this command can be used to fit any Generalised Linear Model, hence the name). The stata command, and resulting output, for fitting this model are shown below

```
. glm chd age, link(log) fam(binom) eform
Iteration 0:   log likelihood = -101.27916
Iteration 1:   log likelihood = -57.187516
  (output omitted)
Iteration 50:  log likelihood = -54.740594  (backed up)
convergence not achieved
Generalized linear models                          No. of obs      =       100
Optimization     : ML: Newton-Raphson              Residual df     =        98
                                                   Scale parameter =         1
Deviance         =  109.4811885                    (1/df) Deviance =  1.117155
Pearson          =  95.42977676                    (1/df) Pearson  =  .9737732
Variance function: V(u) = u*(1-u)                  [Bernoulli]
Link function    : g(u) = ln(u)                    [Log]
Standard errors  : OIM
Log likelihood   = -54.74059424                    AIC             =  1.134812
BIC              = -341.8254897
------------------------------------------------------------------------------
         chd | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.047654     .00875     5.57   0.000     1.030644    1.064945
------------------------------------------------------------------------------
Warning: parameter estimates produce inadmissible mean estimates in one or
         more observations.
```

This produces a relative risk estimate of 1.05 per year increase in age, with a 95% confidence interval (1.03, 1.06). Notice the warning at the bottom of the output that some predicted values are inadmissable. This means that we have predicted probabilities greater than 1. We can see this by plotting the predicted probabilities against age, as seen in Figure 0.8

In this case, we clearly get a better fit from the logistic regression model: the assumption made by the log-binomial model that the probability of CHD increases exponentially with age is not borne out by the data. However, this needs to be checked for each individual model.

In particular, the problem in this example arises from the fact that age is a continuous variable. If all of the predictors are categorical, the issue of the choice of link is far less important (since all variables can only take the values either 0 or 1), and the difference between the logistic and log-binomial models will be less.

0.2.8 Other Link Functions

The log link function can be used because it produces coefficients that are more easily interpretable than the odds ratios that are produced by logistic regression. There are also other link functions that have been used historically, such as the probit function and the complementary log-log link. Both of these link functions produce predicted values in the range 0-1, but the coefficients from these models do not have a straightforward interpretation, and they are therefore not as useful as the logistic regression models. The link functions that can be used in `glm` can be found using the command `help glm`.

0.3 Logistic Regression Diagnostics

Having produced a logistic regression model, we need to check how good it is. There are two components to this: how well it fits the data overall (discussed in Section 0.3.2) and whether there are any individual observations with either a poor fit or an undue influence on the overall fit of the model (discussed in Section 0.3.3).

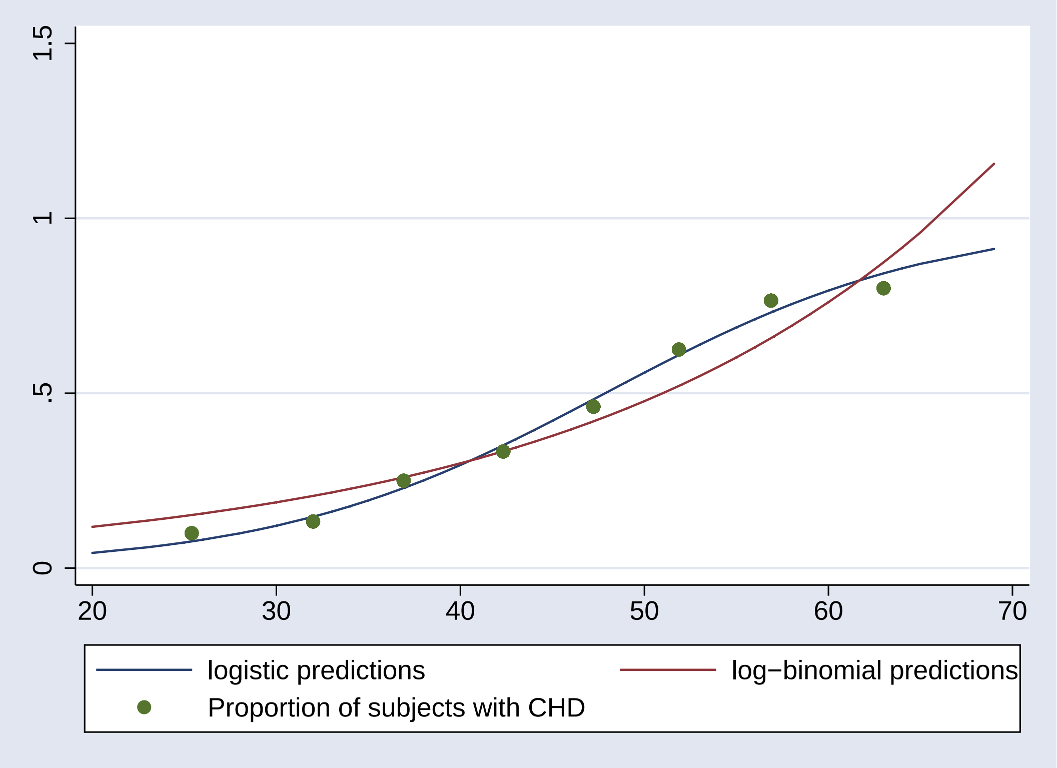


Figure 0.8: Comparison of predictions from logistic regression and log-binomial models

0.3.1 Discrimination and Calibration

Two terms often used when talking about how good a logistic regression model is are *calibration* and *discrimination*. Discrimination refers to how well the model distinguishes between subjects at high and low risk of the outcome. Calibration refers to how closely the predicted probabilities are to the true probabilities. It is possible to have good discrimination without good calibration: if a model developed in a population with high prevalence is applied to a population with low prevalence, the calibration will be poor: the predicted probabilities will be higher than the observed probabilities. However, the discrimination may still be as good: subjects with high predicted probabilities may be more likely to have the event than those with low probabilities, even if the actual predicted probabilities are all too high.

0.3.2 Goodness of Fit

R^2

Whilst R^2 is a very valuable measure of how well a linear regression model fits, it is far less useful with logistic regression. For one thing, there are a number of different ways in which it can be calculated, all of which give the same answer for linear regression but different answers for logistic regression. A comparison of 12 different statistics was made by Mittlböck and Schemper [2], who recommended using either the square of the Pearson correlation coefficient between the observed outcomes and the predicted probabilities, or an alternative measure based on the sum of the squares of the differences between the observed outcomes and the predicted probabilities. The pseudo-R^2 produced by stata is based on the log-likelihood, and was not recommended as it lacks an easily understood interpretation.

It should also be noted that R^2 values tend to be very low for logistic regression models, much lower than for linear regression models. This is because we are trying to predict the outcome, whereas the model only gives us the *probability* of the outcome. So, for example, suppose we have two groups of subjects, a high risk group and a low risk group, and we know which group each subject belongs to. If the prevalence is 0.45 in the low risk group and 0.55 in the high risk group, then the pseudo-R^2 value produced by stata for a logistic regression of group predicting outcome is less than 0.01, despite identifying the high risk and low risk groups perfectly. If the prevalences were 0.1 and 0.9 respectively, the pseudo-R^2 value would be 0.53. Hence, R^2 is a poor choice for an overall assessment of the fit of the model.

Hosmer-Lemeshow test

A better way of assessing the fit of a logistic regression model is compare the expected and observed numbers of positives for different subgroups of the data. If the observed and expected numbers are sufficiently close, then we can assume that we have an adequate model. This test assesses calibration.

How do we select these subgroups ? If all the predictor variables are categorical, then there are only a limited number of covariate patterns that are possible. For example, if we considered age in 5 age-bands and sex as our only predictors, then there are 10 possible covariate patterns, no matter how many subjects in our dataset. We could therefore compare the observed and expected numbers of positives in each of these 10 subgroups.

However, if we treated age as continous variable, it is quite possible that there are no two subjects in the dataset with exactly the same age, and therefore there are as many different covariate patterns as there are subjects. We therefore need another way to group the data.

One suggestion, now widely used, was made by Hosmer & Lemeshow [3]. They rank the subjects according to their predicted probability of a positive outcome, and then divide them into a number of equally sized groups. A χ^2-statistic can be calculated from the expected and observed numbers of positive outcomes in each group. The number of groups to use is arbitrary, but 10 is common.

If this statistic is unusually large, then the differences between the observed and expected values are greater than we would expect by chance, and our model is not adequate. This may be becasue we have missed out an important predictor variable, misspecified the association between one or more predictors and the outcome (e.g. assumed a linear association between the predictor and the logit of the outcome when this is not in fact the case) or omitted one or more important interactions between the variables in the model. We would have to go back to our original logistic regression model and see if changin the model improved the fit to an acceptable level.

This Hosmer-Lemeshow test is implemented in stata using the command `estat gof`. Without any options, this treats each covariate pattern as a distinct group, which is often not useful if there are continuous predictors. The option `group` enables you to choose the number of groups yourself. The results of applying this test to the CHD data is given below:

```
. estat gof, group(10)
Logistic model for chd, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =       100
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =          2.22
                  Prob > chi2 =          0.9734
```

In the above example, the fit of the model would be considered to be adequate, since a χ^2 value of 2.22 on 8 degrees of freedom is not large: $p = 0.97$. So our model fits the data well.

ROC Curves

Another intuitively appealing way to assess the fit of a logistic regression model is to see what proportion of true positives it classifies as being positive (the *sensitivity*) and what proportion of true negatives it classifies as being negative (the *specificity*). Unfortunately, the output of a logistic regression is not a classification as positive or negative, but a predicted probability of being positive or negative.

Of course, we can choose one particular probability as our threshold, and count how many positives and negatives are above and below the threshold, but the choice of threshold will always be arbitrary. The stata command `estat classification` will produce sensitivities and specificities for you. By default, it uses a probability of 0.5 as the threshold, but this can be changed with the option `cutoff`.

A better idea is to measure the sensitivity and specificity at *every* possible threshold. We can then plot the sensitivity and specificity against the chosen threshold, as shown in Figure 0.9 (which was produced using the stata command `lsens`.

If we say that everybody with a probability more than 0.1 counts as a positive, we will have a very sensitive test, but it will not be very specific. As the threshold increases, the sensitivity decreases and the specificity increases, until nobody at all is classed as positive (sensitivity = 0%, specificity = 100%).

However, this is not a very efficient summary of the fit of the model. We can improve matters by plotting the sensitivity against (1 - the specificity) to give a receiver operating characteristic (ROC) curve as shown in Figure 0.10 (which was produced using the stata command `lroc`.

The bottom left corner of this graph corresponds to the right hand side of the previous graph (sensitivity is 0%, specificity 100%), whilst the top right corner corresponds to the left hand side (sensitivity 100%, specificity 0%). Ideally, we would like to reach the top left corner, since there sensitivity is 100% and specificity is 100%. The closer we approach this point, the better the model.

The results of the goodness of fit can therefore be summed up in a single number: the area under the ROC curve (often abbreviated to AUC). If this area reaches 1, then the curve must pass through the top left corner, and the model is perfect. It can be shown that the area under the curve is the probability that, if one one positive subject and one negative subject are selected at random, the positive subject has a higher predicted probability than the negative subject. It is thus a measure of discrimination. The diagonal line represents an area under the curve of 0.5, which is the discrimination that you would expect if you tossed a coin to identify positive subjects, rather than use a logistic regression model.

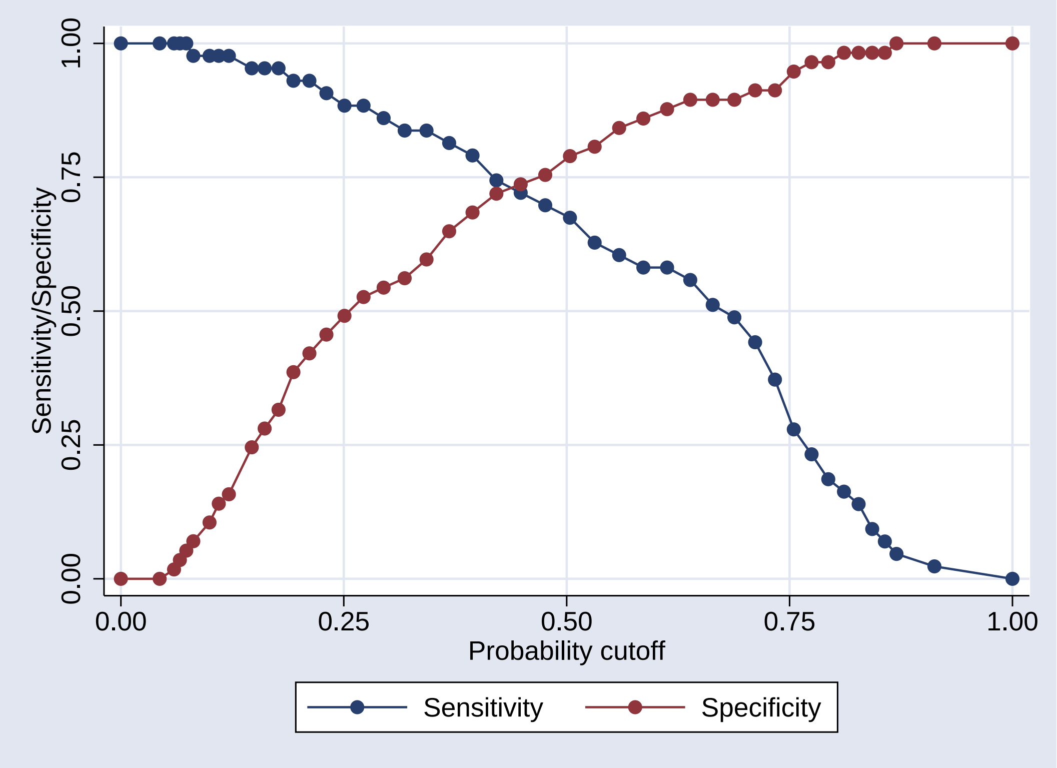


Figure 0.9: Sensitivity and specificity by predicted probability

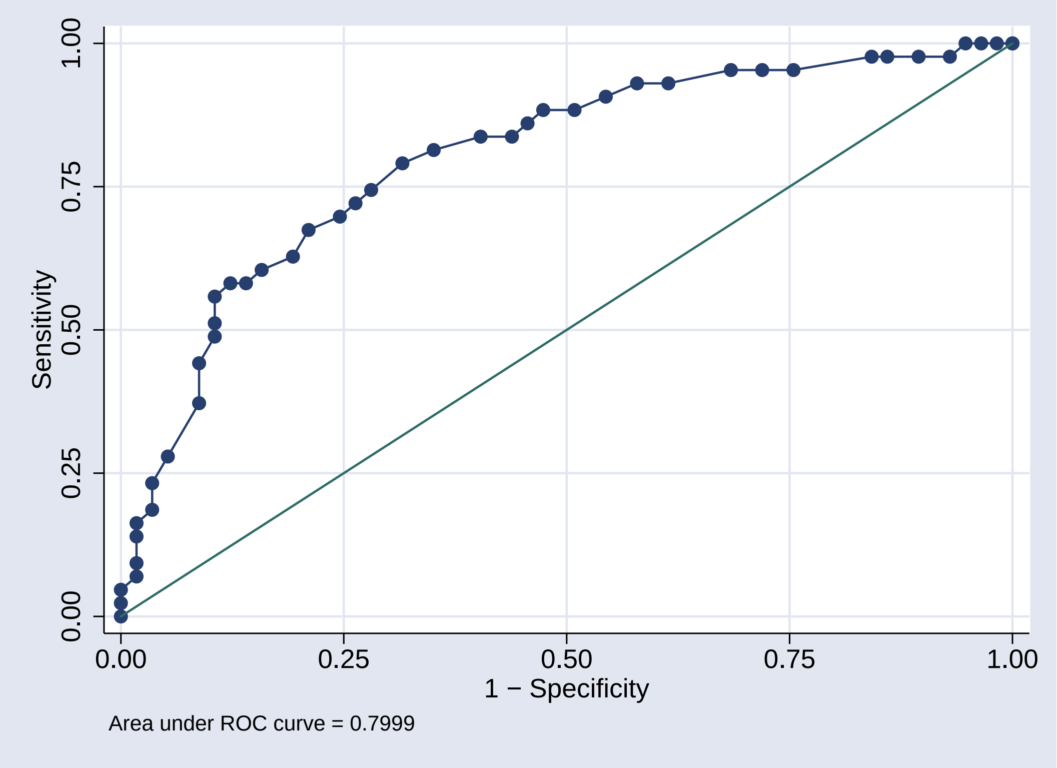
0.3.3 Assessing Fit of Individual Points

The residuals from a linear model are vital in determining whether or not the data satisfies the assumptions of a linear model. However, if the logistic regression model is based on data from individual subjects (which is what we have considered so far), the residuals from a logistic regression model are far less useful. This is because the outcome variable can only take the values 0 and 1, so the residual can only take the values $1 - \hat{\pi}$ or $-\hat{\pi}$.

However, there are some diagnostics available. We can assess how great an influence an individual point has on the coefficients of the logistic regression model and on any lack of overall fit in the model. Any points with unduly high influence should be investigated to ensure that the data has been recorded correctly. The effect of removing the influential point on the model should also be investigated, and if necessary the results presented both with and without the influential point(s).

The concept of *leverage* in a linear model does not translate directly to a logistic model. However, there is a quantity $\Delta\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ which measures the amount that the logistic regression model parameters change when the i^{th} observation is omitted from the model. This can be obtained from stata using the `dbeta` option to the `predict` command after a logistic regression model has been fitted. A plot of $\Delta\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ against $\hat{\pi}$ for each observation (or each covariate pattern) will reveal observations which have a large effect on the regression parameters. It is not possible to give a numerical value to determine which points are influential: you need to identify outliers by eye.

As an example, Figure 0.11 is a plot of $\Delta\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ against $\hat{\pi}$ for each observation in the CHD dataset.



Area under ROC curve = 0.7999

Figure 0.10: ROC curve

You will notice that there is one point which has far more influence than the others. This corresponds to subjects aged 25, 1 of whom had CHD and one did not. This gives an observed prevalence of 0.5, compared to a predicted prevalence of 0.07, which will tend to reduce the slope of the logistic regression model by pulling the regression line upwards. However, excluding these two subjects has very little effect on the logistic regression model: the odds ratio increased from 1.12 per year to 1.13 per year.

```
. logistic chd age if dbeta < 0.2
Logistic regression                               Number of obs   =         98
                                                  LR chi2(1)      =      32.12
                                                  Prob > chi2     =     0.0000
Log likelihood = -50.863658                       Pseudo R2       =     0.2400

------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.130329   .0293066     4.73   0.000     1.074324    1.189254
------------------------------------------------------------------------------
```

In addition, we can use residuals to identify individual points for which the model fits poorly. Stata provides two different kinds of residuals for this: ΔX_i^2, obtained by using the dx2 option to predict; and ΔD_i, obtained by using the ddeviance option. Again, these statistics can be plotted against $\hat{\pi}$, and outlying observations identified by eye. Such observations do not need to be omitted from the analysis, since they are not having an undue influence on the regression parameters, but it may be important to point out areas in which the predictions from the model are poor.

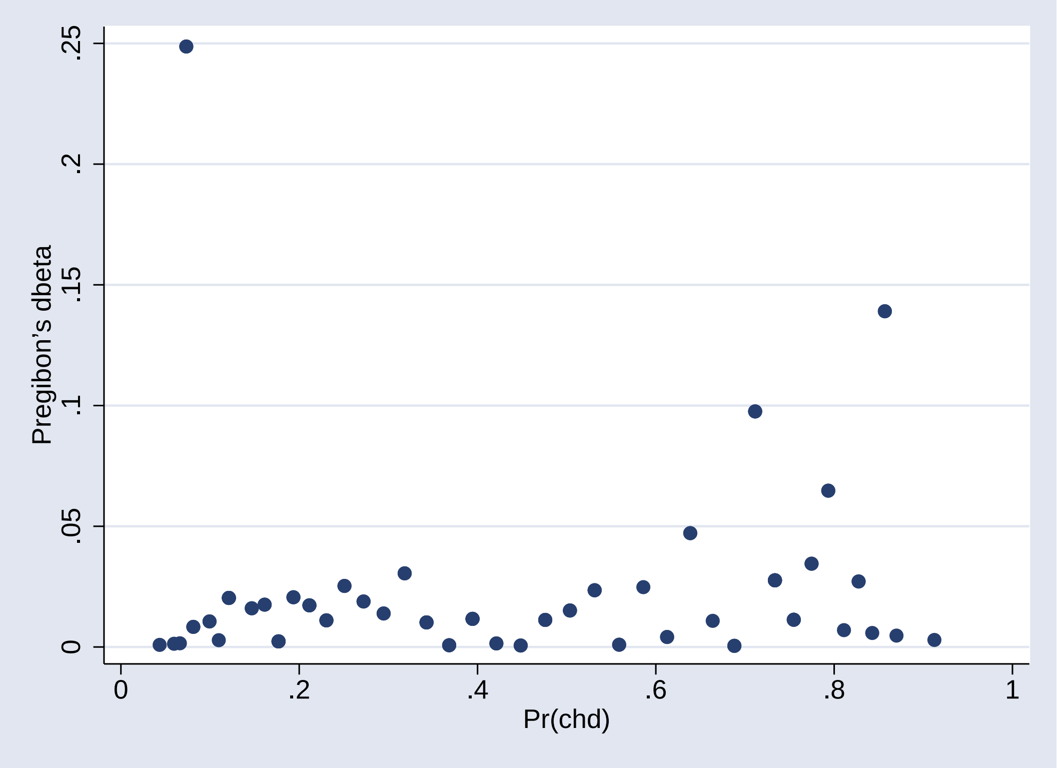


Figure 0.11: Plot of $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ against $\hat{\pi}$

0.3.4 Problems of separation

It can happen that one of the predictor variables predicts the outcome perfectly. For example, suppose that we are interested in the effect of gender on our outcome, but we only have a single woman in our sample. The odds ratio for women compared to men will be either 0 (if the one woman does not have the outcome of interest), or ∞ if she does. Unfortunately, odds of 0 and ∞ correspond to a linear predictor of $-\infty$ or ∞, and stata cannot handle infinite numbers (that should not be taken as criticism, nor can any other stats package, or indeed computer program). It will therefore report that "gender determines outcome exactly" and drop the woman from the regression model.

This problem arises when there is a particular combination of indicator variables for which there are no cases (or every subject is a case). The problem is most common when modelling interactions between predictors with several categories, since this involves dividing subjects into a large number of subgroups. The only solution is to collapse categories together until there is at least one case and one control for each combination of the indicator variables. If one of the categorical variables can be fitted as a continuous variable, this might also help.

Bibliography

[1] David W. Hosmer and Stanley Lemeshow. *Applied Logistic Regression.* John Wiley & Sons, Inc., 2 edition, 2000.

[2] M. Mittlböck and M. Schemper. Explained variation for logistic regression. *Statistics in Medicine*, 15(19):1987–1997, 1996.

[3] David W. Hosmer and Stanley Lemeshow. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, 9(10):1043–1069, 1980.