Chapter 6

Logistic regression

In this chapter we introduce the logistic regression model and illustrate its use by two examples.

6.1 The definition of the logistic regression model

So far we have considered only the case where the outcome of interest is a continuous variable. Often the outcome of interest is a binary variable Y. In this case, we can use a logistic regression model. In a logistic regression model we model the probability that Y takes the value 1 as a function of the covariates $X_1, X_2, \ldots, X_p$. We will denote the conditional probability to observe $Y = 1$ given the covariate values $x_1, x_2, \ldots, x_p$ by

$$\pi(x_1, x_2, \ldots, x_p)$$

Now one may start with considering a model for π, for example

$$\pi(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

This is possible, but it has the disadvantage, that the quantity on the left side is a number between 0 and 1, and the right side can take any value between $-\infty$ and ∞. To overcome this problem, we can transform the left side of the equation. The most popular transformation is the so called logit transformation, i.e. we apply the logit function

$$\text{logit}\, p = \log \frac{p}{1-p} \quad ,$$

such that we obtain the so called logistic regression model

$$\text{logit}\, \pi(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \quad .$$

The logit function is illustrated in Figure 6.1. The transformation from the probability scale (i.e. the interval $[0, 1]$) to the logit scale (i.e. the interval $[-\infty, \infty]$) is illustrated in Figure 6.2.

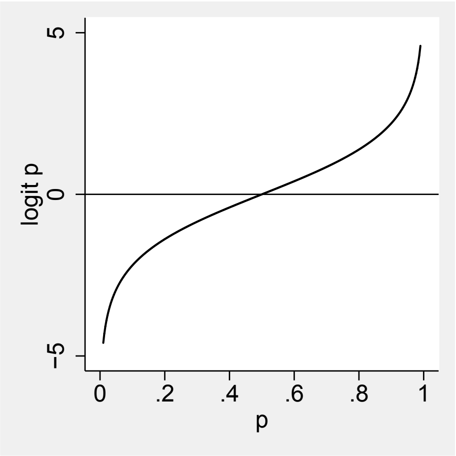

Figure 6.1: The logit function

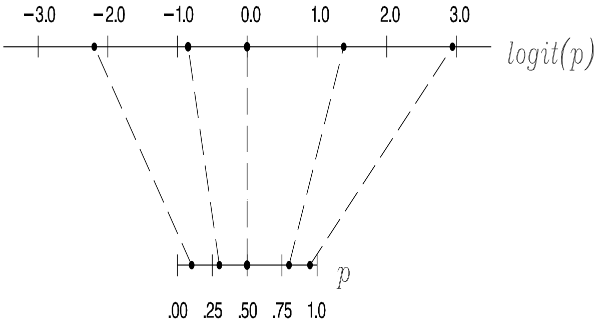

Figure 6.2: The logit transformation

The middle of the probability scale, i.e. 0.5, is transformed to the middle of the logit scale, i.e. 0.0, a probability of 0.3 is transformed to -0.85 on the logit scale, and a probability of 0.95 is transformed to 2.94.

The interpretation of the regression parameters in a logistic regression model is similar to that in the classical regression model. However, instead of describing a change in the expected value of Y, they describe now the change in the probability to observe $Y = 1$ expressed on the logit scale. So if we now compare subjects with identical covariate values except of a difference of Δ in covariate x_j, then the conditional probability to observe $Y = 1$ differs on the logit scale by $\Delta \times \beta_j$.

6.2 Analysing a dose response experiment by logistic regression

In a dose response experiment we typically expose units (like animals, cell cultures or patients) to different doses of a substance and investigate, how the probability of a response increases with increasing dose. Table 6.1 and Figure 6.3 show the results of such an experiment, where the toxic effect of a substance A and a possible protective effect of oxygen were investigated. For each of 7 different doses of substance A, 200 cells were exposed to this dose and the number of cells with a damage were counted. 100 of the 200 cells were exposed to an increased level of oxygen. From the results in Table 6.1 and Figure 6.3 we can see, that the toxicity increases with increasing dose but that exposure to oxygen has a protective effect. It remains to quantify the effects in a useful manner.

	dose(mg)						
	10	20	30	40	50	60	70
normal oxygen	.1	.28	.53	.77	.91	.98	.99
increased oxygen	.03	.09	.17	.45	.74	.91	.97

Table 6.1: The results of a dose response experiment: Relative frequency of damaged cells in each dose group stratified by oxygen exposure level. (Relative frequencies are expressed as fractions, not percentages.)

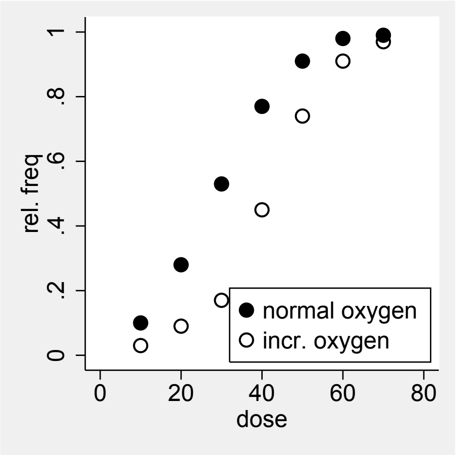

Figure 6.3: The data of Table 6.1 visualised

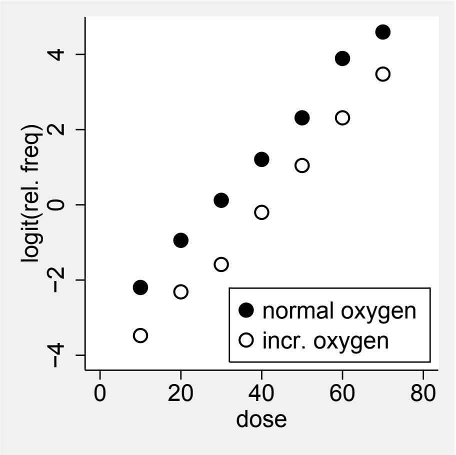

Figure 6.4: The same visualisation, but now on the logit scale

We cannot use the classical regression model here, because our lines are far away from being linear. This is due to the fact, that they are forced to stick to the interval [0,1]. However, if we transform the y-axis by a logit transformation, we obtain two rather parallel lines (Figure 6.4), suggesting that we can describe this data by a logistic regression model. We consider variables defined for each single cell in the experiment, i.e. the outcome

$$Y_i = \begin{cases} 1 & \text{if cell } i \text{ is damaged} \\ 0 & \text{if cell } i \text{ is not damaged} \end{cases}$$

and the covariates

$$X_{i1} = \text{dose of substance A applied to cell } i \text{ (in mg)}$$

and

$$X_{i2} = \begin{cases} 1 & \text{if cell } i \text{ is exposed to increased oxygen level} \\ 0 & \text{if cell } i \text{ is exposed to normal oxygen level} \end{cases} .$$

With $\pi(x_1, 0)$ we denote the probability that a cell exposed to dose level x_1 is damaged at normal oxygen level and with $\pi(x_1, 1)$ we denote the probability that a cell exposed to dose level x_1 is damaged at increased oxygen level. The results of Table 6.1 suggest that for example $\pi(30, 0)$ is about 0.53 and $\pi(30, 1)$ is about 0.17.

Then we can formulate the logistic model

$$\operatorname{logit} \pi(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Fitting this model to the data yields estimates of $\hat{\beta}_1 = 0.117$ and $\hat{\beta}_2 = -1.46$. So increasing the dose level by 10 mg increases the probability of cell damage by 1.17 on the logit scale (if we keep the oxygen level fixed), and exposing a cell to an increased oxygen level decreases the probability of cell damage by 1.46 (if we keep the dose of substance A fixed). Since $12.49 \times \beta_1 = -\beta_2$ we can also say, that increasing the level of oxygen has the same effect as decreasing the dose of substance A by 12.49.

Of course, using a standard statistical package to fit a logistic model we will also obtain standard errors, confidence intervals and p-values. So the output may look like

variable	beta	SE	95%CI	p-value
intercept	-3.374	0.212	[-3.790,-2.958]	<0.001
dose	0.117	0.006	[0.105,0.128]	<0.001
oxygen	-1.456	0.168	[-1.786,-1.126]	<0.001

So we can conclude here, that we have a rather precise knowledge about the regression coefficients (because the confidence intervals are rather small) and that there is no doubt about, that both the dose of substance A and exposure to increased oxygen levels have an effect on the probability of cell damage (because both p-values are very small).

We can use the parameter estimates also to estimate the probability of cell damage for a given dose. For example logit $\pi(25, 0)$ describes the probability (on the logit scale) of cell damage, if the cell is exposed to 25mg of substance A at normal oxygen level. We can estimate this number as

$$\operatorname{logit} \hat{\pi}(25, 0) = \hat{\beta}_0 + \hat{\beta}_1 \times 25 + \hat{\beta}_2 \times 0 = -0.46 \ .$$

To obtain this probability on the probability scale, we can just apply the inverse of the logit transformation, i.e.

$$\text{logit}^{-1}(t) = \frac{1}{1+e^{-t}} \; .$$

So we can estimate the probability of cell damage after exposition to 25mg of substance A and normal oxygen level as

$$\text{logit}^{-1}(-0.46) = \frac{1}{1+e^{-(-0.46)}} = 0.39 \; .$$

Remark: Our example is a little bit artificial, because in such an experiment there is no need to use so many cells at each dose level. We have used this large number to obtain results which allow to visualize the logit transformation. In practice in a dataset with a continuous covariate, usually all values of the covariate are different, and it is difficult to visualise the dataset by a scatter plot, because *Y* takes only the value 0 and 1. Fortunately, a logistic regression model do not require that we can draw a figure like Figure 6.3. We can apply it even if we have only one cell for each dose level. Such a situation is visualised in Figure 6.5 for the case of a single covariate *X*. The figure shows both the raw data as well as the fitted logistic model.

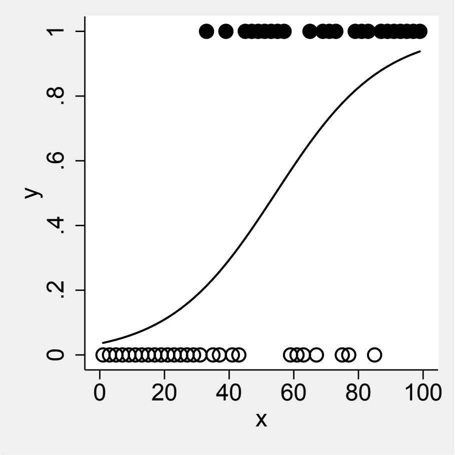

Figure 6.5: Visualisation of a dataset with a single covariate *X* (dots) and a binary outcome *Y* (empty and filled dots) and the fitted logistic regression model (dotted line).

Remark: Although it is possible to back-transform an estimated probability on the logit scale to the probability scale, it is not possible to back-transform a regression parameter estimate to the probability scale. This is illustrated in Figure 6.6: The same difference on the logit scale can result in different differences on the probability scale.

6.3 How to fit such a dose response model with Stata

We start with looking at the data set:

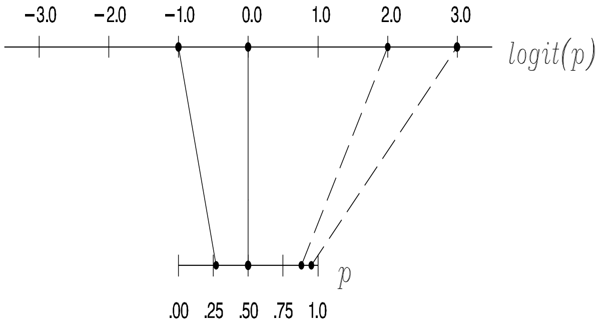

Figure 6.6: The same difference on the logit scale of 1.0 can correspond to different differences on the probability scale

```
. use toxic, clear

. list in 1/10

     +--------------------------------+
     | dose   oxygen   cell   damage |
     |--------------------------------|
  1. |   10        0      1        0 |
  2. |   10        0      2        0 |
  3. |   10        0      3        0 |
  4. |   10        0      4        0 |
  5. |   10        0      5        0 |
     |--------------------------------|
  6. |   10        0      6        0 |
  7. |   10        0      7        0 |
  8. |   10        0      8        0 |
  9. |   10        0      9        0 |
 10. |   10        0     10        0 |
     +--------------------------------+
```

We use Stata's table command to get a quick overview about the data:

```
. table dose damage oxygen, c(freq)

------------------------------------
          |    oxygen and damage
          | ---- 0 ---     ---- 1 ---
     dose |    0     1        0     1
----------+-------------------------
       10 |   90    10       97     3
       20 |    72    28      91     9
       30 |    47    53      83    17
       40 |    23    77      55    45
       50 |     9    91      26    74
       60 |     2    98       9    91
       70 |     1    99       3    97
------------------------------------
```

We fit the logistic regression model by Stata's `logit` command. The `logit` command expects the outcome as the first argument and then the covariates:

```
. logit damage dose oxygen

Iteration 0:   log likelihood = -958.27957
Iteration 1:   log likelihood = -569.58618
Iteration 2:   log likelihood = -525.72626
Iteration 3:   log likelihood = -520.93341
Iteration 4:   log likelihood = -520.83619
Iteration 5:   log likelihood = -520.83614

Logit estimates                                   Number of obs   =       1400
                                                  LR chi2(2)      =     874.89
                                                  Prob > chi2     =     0.0000
Log likelihood = -520.83614                       Pseudo R2       =     0.4565

------------------------------------------------------------------------------
      damage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |    .116534   .0058846    19.80   0.000     .1050003    .1280677
      oxygen |  -1.455875   .1682545    -8.65   0.000    -1.785647   -1.126102
       _cons |  -3.373973   .2120248   -15.91   0.000    -3.789534   -2.958413
------------------------------------------------------------------------------
```

Again the reader should focus only on the numbers of interest, i.e. the regression coefficients for the covariates *dose* and *oxygen* and the corresponding standard errors, confidence intervals and p-values.

6.4 Estimating odds ratios and adjusted odds ratios using logistic regression

Let us start with an example of an epidemiological study, in which we investigate the influence of risk factors on the development of allergies among children. The outcome variable of interest is the allergy status of the child at age 6, i.e.

$$Y_i = \begin{cases} 1 & \text{if child } i \text{ develops an allergy until age 6} \\ 0 & \text{otherwise} \end{cases}$$

and the first risk factor we look at is the maternal allergy status, i.e.

$$X_{1i} = \begin{cases} 1 & \text{if the mother of child } i \text{ is suffering from allergies} \\ 0 & \text{otherwise} \end{cases}$$

If we are interested in analysing the association between the maternal allergy status and the allergy status of the child, we can look at a cross tabulation of these variables:

		Y = allergy child					
		0 =no	1 =yes	n	rel.freq. (%)	odds	OR
X_1 = allergy mother	1 =yes	264	142	406	35.0	$\frac{35.0}{65.0} = 0.54$	$\frac{0.54}{0.39} = 1.37$
	0 =no	516	203	719	28.2	$\frac{28.2}{71.8} = 0.39$	

We can observe, that out of 406 children from mothers with allergies 142, i.e. 35.0%, suffer from an allergy, whereas out of 719 children from mothers without an allergy 203, i.e. 28.2%, suffer from an allergy. Now there is a tradition in epidemiology to express the difference between the two groups of children, i.e. the difference between 35.0% and 28.2%, as an odds ratio. The first step in computing an odds ratio is to express relative frequencies as empirical odds by dividing the relative frequency with 100 minus the relative frequency. Odds are just numbers which takes expressions like fifty:fifty or 90:10 literally, i.e. fifty:fifty gives an odds of 1.0 and 90:10 gives an odds of 9.0. In our example we obtain odds of 0.54 and 0.39, as shown in the table above. Then we have just to take the ratio of these two odds, which gives here an odds ratio of 1.37, indicating that the odds for having a child with an allergy is 1.37 times higher in mothers with an allergy compared to mothers without an allergy.

Now let us define odds ratios a little bit more formally. Let $\pi(x)$ denote the probability of $Y = 1$ given $X = x$, such that $\pi(0)$ is the probability that a mother without allergies has a child with an allergy and $\pi(1)$ is the probability that a mother with an allergy has a child with an allergy. Then the odds for having a child with an allergy are $\frac{\pi(1)}{1-\pi(1)}$ for a mother with an allergy and $\frac{\pi(0)}{1-\pi(0)}$ for a mother without allergies. So the odds ratio is defined as

$$\text{OR} = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} \quad .$$

Now there exist a strong connection between odds ratios and logistic regression models, because if we take the logarithm of the odds ratio (and apply the mathematical rule $\log \frac{a}{b} = \log a - \log b$), then we obtain

$$\log \text{OR} = \log \frac{\pi(1)}{1-\pi(1)} - \log \frac{\pi(0)}{1-\pi(0)} = \text{logit}\,\pi(1) - \text{logit}\,\pi(0) \ .$$

Hence the logarithm of an odds ratio is nothing else but the difference between the probabilities $\pi(1)$ and $\pi(0)$ on the logit scale. Now it is not surprising, that we can obtain odds ratios also from a logistic regression model. If we consider a simple logistic regression model for our data, this reads

$$\text{logit}\,\pi(x) = \beta_0 + \beta_1 x$$

and β_1 is the increase in $\pi(x)$ on the logit scale, if we go from $x = 0$ to $x = 1$, i.e.

$$\beta_1 = \text{logit}\,\pi(1) - \text{logit}\,\pi(0)$$

and hence

$$\beta_1 = \log \text{OR} \ .$$

Or with other words, if we exponentiate β_1, we obtain the odds ratio:

$$e^{\beta_1} = \text{OR} \ .$$

Indeed, is we fit a logistic regression model to our data, we obtain $\hat{\beta}_1 = 0.313$ and $e^{0.313} = 1.37$, i.e. we obtain the same result as above.

Now analysing the association between maternal atopy and allergy of the child without taking other variables into account can be misleading, because atopic mothers may tend to reduce the risk of the child by avoiding exposures like smoking. Hence for a more substantiated analysis we have to take smoking into account. One way to do this is to eliminate the effect of smoking by restricting the analysis to smoking women or non-smoking women, respectively, i.e. to stratify by the smoking status of the mother. This results in one table for the smoking women:

		Y = allergy child					
		0 =no	1 =yes	n	rel.freq. (%)	odds	OR
X_1 = allergy mother	1 =yes	78	62	140	44.3	$\frac{44.3}{55.7} = 0.79$	$\frac{0.79}{0.50} = 1.60$
	0 =no	297	148	445	33.3	$\frac{33.3}{66.7} = 0.50$	

and one for the non-smoking women:

		Y = allergy child					
		0 =no	1 =yes	n	rel.freq. (%)	odds	OR
X_1 = allergy mother	1 =yes	186	80	266	30.1	$\frac{30.1}{69.9} = 0.43$	$\frac{0.43}{0.25} = 1.71$
	0 =no	219	55	274	20.1	$\frac{20.1}{79.9} = 0.25$	

Among smoking mothers with an allergy 44.3% of the children suffer from allergies, whereas among smoking mothers without an allergy 33.3% of the children suffer from allergies, resulting in an odds ratio of 1.60. Among non smoking mothers with an allergy 30.1% of the children suffer from allergies, whereas among non smoking mothers without an allergy 20.1% of the children suffer from allergies, resulting in an odds ratio of 1.71. So in both subgroups we obtain odds ratios which are larger than those in the analysis of the whole population.

The explanation for this is, that in the first analysis of the whole population we mix the effects of maternal smoking and maternal allergy status. Among mothers without an allergy we find 61.9% smokers, whereas among mothers with an allergy we find only 34.5% smokers, i.e. mothers with an allergy tend to smoke less frequent. However, smoking itself seems to be unfavorable for the children, because we can find more children with an allergy among the smoking women. So the effect of maternal allergy status is bigger within each subgroup than in the whole population, because in the latter analysis the unfavorable effect of having a mother with an allergy is partially compensated by the the fact, that mothers with an allergy tend to smoke less. So we have here the a situation similar to those considered in Chapter 4: The effects of maternal smoking and maternal allergy are confounded.

It remains the task to summarize the two odds ratios of 1.60 and 1.71 into one number, i.e. to do some type of averaging. Here we can again use logistic regression. To see this, let us start to define the odds ratio in each subgroup in a more formal way. Let $\pi(x_1, x_2)$ denote the probability to have a child with an allergy for a mother with allergy status x_1 and smoking status x_2, where the latter refers to a second covariate

$$X_{i2} = \begin{cases} 1 & \text{if mother of child } i \text{ is smoking} \\ 0 & \text{if mother of child } i \text{ is not smoking} \end{cases} .$$

Then for a smoking mother with an allergy the odds for having a child with an allergy is $\frac{\pi(1,1)}{1-\pi(1,1)}$ and for a smoking mother without an allergy it is $\frac{\pi(0,1)}{1-\pi(0,1)}$, such that the odds ratio is

$$\text{OR}_s = \frac{\frac{\pi(1,1)}{1-\pi(1,1)}}{\frac{\pi(0,1)}{1-\pi(0,1)}} .$$

Similar, for a non smoking mother the odds ratio is

$$\text{OR}_{ns} = \frac{\frac{\pi(1,0)}{1-\pi(1,0)}}{\frac{\pi(0,0)}{1-\pi(0,0)}} .$$

Now in a logistic regression model with the two covariates X_1 and X_2 we assume

$$\text{logit}\,\pi(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

such that β_1 is the change of the probability to have a child with an allergy on the logit scale if we switch from $x_1 = 0$ (no maternal allergy) to $x_1 = 1$ (maternal allergy), keeping the smoking status x_2 fixed. With other words

$$\beta_1 = \text{logit}\,\pi(1, x_2) - \text{logit}\,\pi(0, x_2) = \log\frac{\pi(1, x_2)}{1-\pi(1, x_2)} - \log\frac{\pi(0, x_2)}{1-\pi(0, x_2)} = \log\frac{\frac{\pi(1,x_2)}{1-\pi(1,x_2)}}{\frac{\pi(0,x_2)}{1-\pi(0,x_2)}}$$

or

$$e^{\beta_2} = \frac{\frac{\pi(1,x_2)}{1-\pi(1,x_2)}}{\frac{\pi(0,x_2)}{1-\pi(0,x_2)}}$$

So for $x_2 = 0$ and $x_2 = 1$ we obtain

$$e^{\beta_2} = \frac{\frac{\pi(1,0)}{1-\pi(1,0)}}{\frac{\pi(0,0)}{1-\pi(0,0)}} = \mathrm{OR}_s \text{ and } e^{\beta_2} = \frac{\frac{\pi(1,1)}{1-\pi(1,1)}}{\frac{\pi(0,1)}{1-\pi(0,1)}} = \mathrm{OR}_{ns}$$

So in a logistic regression model we just assume, that the true odds ratios in the two subgroups are identical. This is of course a reasonable assumption, because if one wants to average the two empirical odds ratios, one implicitly assumes, that the true effects are at least of similar magnitude. Fitting a logistic regression model to our data results in an output like

variable	beta	SE	95%CI	p-value
intercept	-1.362	0.127	[-1.610,-1.113]	<0.001
allergym	0.502	0.141	[0.226,0.778]	<0.001
smokem	0.656	0.139	[0.384,0.928]	<0.001

By exponentiating the effect estimate $\hat{\beta}_1 = 0.502$ for maternal allergy status, we obtain $e^{\hat{\beta}_1} = 1.65$, i.e. an odds ratio lying in between the two odds ratios we have observed in the subgroups. Since this odds ratio takes the confounding effect of smoking into account, we call it an adjusted odds ratio, or more specific the odds ratio between maternal allergy and allergy of the child adjusted for smoking.

Since the logistic regression model allows to include several covariates, we can use it to adjust simultaneously for several potential confounders by adding them as covariates. Such a confounder might be the smoking of the father, because this may increase the risk of allergies in a child and women suffering from allergies might tend to choose a non-smoking partner. We will consider such an example with several confounders in the next exercise. Logistic regression allows of course also to adjust for confounders measured as continuous variables.

Remark: Logistic regression is only one method to average odds ratios across different subgroups. The Mantel Haenszel procedure is an alternative, which us also used frequently in the medical literature. It gives typically results similar to that of a logistic regression.

Remark: It can of course happen that odds ratios in different subgroups are rather different, such that it is questionable to average them. We will discuss in Chapter 18 techniques to compare odds ratios in different subgroups.

6.5 How to compute (adjusted) odds ratios using logistic regression in Stata

Stata has actually two commands for logistic regression: `logit` and `logistic`. They do exactly the same with the only difference, that `logit` present the regression parameters on the logit scale, whereas `logistic` presents the results as odds ratios. This affects also the standard errors and the confidence intervals, but the p-values are of course identical.

Now let us take a look on our dataset:

```
. use allergy1, clear

. list in 1/10

     +---------------------------------------+
     | childnr   allergyc   allergym   smokem |
     |---------------------------------------|
  1. |     838          0          0        1 |
  2. |      50          0          0        0 |
  3. |     584          0          0        0 |
  4. |     704          1          0        0 |
  5. |     152          1          0        0 |
     |---------------------------------------|
  6. |     722          0          1        0 |
  7. |     991          0          0        1 |
  8. |     559          1          0        1 |
  9. |     563          0          1        1 |
 10. |     423          0          0        1 |
     +---------------------------------------+
```

We have the binary outcome variable `allergyc` and the two covariates `allergym` and `smokem`.

To compute the unadjusted odds ratio for the association between the maternal allergy status and the allergy status of the child we can use

```
. logistic allergyc allergym

Logistic regression                               Number of obs   =       1125
                                                  LR chi2(1)      =       5.49
                                                  Prob > chi2     =     0.0191
Log likelihood = -690.71179                       Pseudo R2       =     0.0040

------------------------------------------------------------------------------
    allergyc | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
    allergym |   1.367219   .1818668     2.35   0.019     1.053444    1.774453
------------------------------------------------------------------------------
```

To obtain the odds ratio adjusted for maternal smoking we can use

```
. logistic allergyc allergym smokem

Logistic regression                               Number of obs   =       1125
                                                  LR chi2(2)      =      28.48
                                                  Prob > chi2     =     0.0000
Log likelihood =  -679.2193                       Pseudo R2       =     0.0205

------------------------------------------------------------------------------
    allergyc | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    allergym |   1.651714   .2328059     3.56   0.000     1.253024    2.177259
      smokem |   1.927018   .2672664     4.73   0.000     1.468348    2.528964
------------------------------------------------------------------------------
```

6.6 Exercise *Allergy in children*

In the above mentioned study on the development of allergies in early childhood also information on the smoking and allergy status of the father was recorded. This information can be found in the dataset `allergy2`.

a) Compute the odds ratio between the paternal allergy status and the allergy status of the child and adjust this odds ratio for the effect of paternal smoking. Can you explain, why the two odds ratios are so similar?
b) What happens, if you adjust for maternal allergy status instead of paternal smoking? Can you explain the difference between the unadjusted and adjusted odds ratio?
c) If you make a cross tabulation between maternal allergy status and paternal smoking (do it!), you will find a tendency that mothers with an allergy tend to have a non-smoking partner. If you fit a logistic regression model with *Maternal allergy*, *Maternal smoking* and *Paternal smoking* (do it!), you can see, that paternal smoking has an influence on the allergy status of the child. This suggest, that paternal smoking is a confounder for the association between the maternal allergy status and the allergy status of the child. However, if we remove in the model above the covariate *Paternal smoking* (do it!), we can observe, that the odds ratio for *Maternal allergy* does not change. Can you explain this?

d) Fit a logistic regression model with all four covariates.

   d.1) What can we conclude from this analysis with respect to the effect of the single covariates?

   d.2) Can we say something about the covariate with the smallest and the biggest effect?

   d.3) How big is the difference in the risk to develop an allergy between a child of smoking parents both suffering from allergies and a child of non smoking, allergy-free parents?

   d.4) Can we now make a final judgement of the magnitude of the effect of maternal and paternal allergy status on the development of allergies in early childhood?

6.7 More on logit scale and odds scale

Using odds ratios to express the effect of covariates on a binary outcome involes two steps. The first is to transform probabilities to odds, i.e. to consider the transformation $p \to \text{odds}\,(p) = \frac{p}{1-p}$ of the probability scale (i.e. the interval $[0, 1]$) to the odds scale (i.e. the interval $[0, \infty]$), cf. Figure 6.7. The second is to use ratios to measure the ”distance” between two odds. This we can also express by saying that the odds scale is a multiplicative scale. The transformation from the odds scale to the logistic scale is a simple logarithmic transformation, i.e. $\text{logit}\, p = \log \text{odds}\,(p)$, which just expresses that exponentiated differences on the logit scale correspond to ratios on the odds scale.

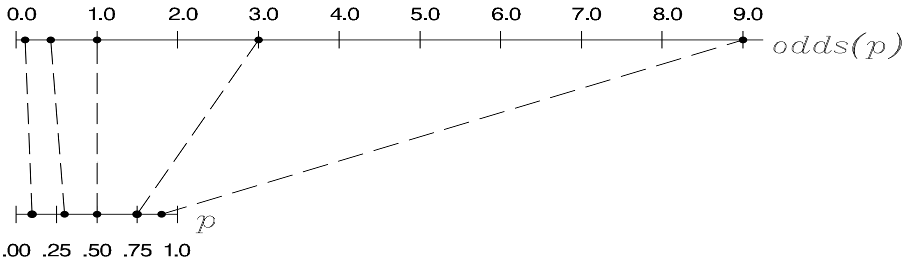

Figure 6.7: The odds transformation

In some areas of medical research it is widespread to present results from logistic regression model as odds ratios, in other areas it is more common to report effect estimates on the logit scale. One may argue that odds ratios are easier to interpret than differences on the logit scale, because an odds ratio of 2.1 means that something (namely the odds of $Y = 1$) is increased by a factor 2.1, if we compare subjects differing by 1 in the corresponding covariate. However, odds ratios suffer from the disadvantage to be asymmetric with respect to positive and negative effects. For example if we recode a binary covariate X_j by exchanging 0 and 1, the corresponding effect estimate $\hat{\beta}_j$ switches to $-\hat{\beta}_j$, whereas the corresponding estimated odds ratio switches from $\hat{\text{OR}}_j$ to $\frac{1}{\hat{\text{OR}}_j}$. So if we just look on a table with effect estimates for different covariates, it is not so simple to see that two covariates with estimated odds ratios of 2.5 and 0.4, respectively, have actually the same magnitude of the effect, but just in opposite directions, whereas the corresponding values $\hat{\beta}_j = 0.92$ and $\hat{\beta}_j = -0.92$ clearly indicate this. For this reason many authors try to code or recode their covariates in a manner, such that at the end all estimated odds ratios are above 1.0.

Odds ratios are mainly used to express the effect of binary covariates, but it is also possible to use them for continuous covariates. Then they just describe the factor by which the odds of $Y = 1$ changes when we compare two subjects differing by 1 unit in the covariate of interest, keeping all other covariates fixed.

Odds ratios and effect estimates on the logit scale share the problem that they do not refer to differences on the probability scale, which makes it (especially for beginners) difficult to get an impression, whether an estimated effect is big or small. For this reason we have summarized in Table 6.2 how some selected differences on the probability scale translate to differences on the logit scale or odds ratios, respectively. If the probabilities coincide with the probability of $Y = 1$ for two subjects differing only in the covariate X_j by 1, then the difference on the logit scale corresponds to the regression coefficient β and the odds ratio is equal to $\exp(\beta)$. The values in Table 6.2 may allow some orientation about what is a big and what is a small odds ratio or effect on the logit scale. In the long run most users of logistic regression develop a feeling for this, because they see in the literature the effect sizes which are typical in applications in their specific field.

p_1	p_2	logit p_2 − logit p_1	OR=$\frac{p_2}{1-p_2}/\frac{p_1}{1-p_1}$
50%	60%	0.41	1.50
65%	75%	0.48	1.62
80%	90%	0.81	2.25
40%	60%	0.81	2.25
55%	75%	0.90	2.45
70%	90%	1.35	3.86
30%	70%	1.69	5.44
40%	80%	1.79	6.00
50%	90%	2.20	9.00

Table 6.2: Selected pairs of probabilities and their corresponding differences on the logit scale and odds ratios.

Remark: If we consider a logistic regression model with for example three covariates, we have

$$\text{logit}\,\pi(x_1, x_2, x_3) = \log\frac{\pi(x_1, x_2, x_3)}{1 - \pi(x_1, x_2, x_3)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

and if we exponentiate both sides of the equation we obtain

$$\begin{aligned}\text{odds}\,(\pi(x_1, x_2, x_3)) = \frac{\pi(x_1, x_2, x_3)}{1 - \pi(x_1, x_2, x_3)} &= e^{\beta_0} e^{\beta_1 x_1} e^{\beta_2 x_2} e^{\beta_3 x_3} \\ &= e^{\beta_0} (\text{OR}_1)^{x_1} (\text{OR}_2)^{x_2} (\text{OR}_3)^{x_3}\ .\end{aligned}$$

Here OR_j is the odds ratio between Y and the covariate X_j adjusted for the two other covariates. So the logistic regression model can be also expressed as an multiplicative model on the odds scale, and sometimes it is described in this way in the literature. However, in any case we obtain the same model, we use just different – but equivalent – mathematical representations. Especially p-values referring to hypotheses test of no effect of an covariate are independent of the chosen representation.

This chapter in a nutshell

Logistic regression is a simple tool to analyse the effect of covariates on a binary outcome. Covariate effects can be expressed as differences on the logit scale or as odds ratios.