Chapter 12

Estimating the Odds Ratio

Overweight and sleep apnea

Many otherwise healthy people do not breathe normally when they sleep, and most of them are not even aware that something goes wrong every night. Instead of breathing peacefully and orderly they struggle breathing through collapsed upper airways, sometimes unable to take a breath at all (apnea) and sometimes taking only a shallow breath (hypopnea). This common disorder has several names— apnea-hypopnea syndrome, sleep-disordered breathing, obstructive sleep apnea—depending on the frequency and severity of the abnormality. For pedagogical reasons, I will ignore the heterogeneity; use the label "sleep apnea"; and treat the condition as a binary variable (present, absent) even though the pathology is measured on a continuous scale: from infrequent hypopneas to near-chocking every minute. Keep in mind, however, that dichotomizing a continuous effect is almost always a bad idea in causal inquiry—as I argued at the end of chapter 9.

When the effect takes on a binary form and the data set does not contain information on the person-time at risk, we cannot compute rates and have to rely on probabilities or odds and their derived measures of associations. In multi-variable analysis, the linear probability model may be used to estimate conditional probability differences or modified probability differences (chapters 9 and 10). Nonetheless, for reasons that were explained in chapter 3, we should always prefer to estimate ratio measures of effect—the probability ratio or the odds ratio. In this chapter we will estimate various odds ratios for sleep apnea, using data from a cross-sectional sample of about 1,000 people. Methods to estimate the probability ratio will be the topic of chapter 16.

Only a few important causes of sleep apnea have been identified so far. Of these, WEIGHT will draw our attention here as the exposure variable. Again, for pedagogical reasons, I first dichotomized that continuous variable at an arbitrary cutoff value of 100kg: anyone in the sample whose weight exceeded 100kg was labeled "overweight". In a second approach, I categorized weight into four groups—applying the cutoff points of 70kg, 80kg, and 90kg—and created an ordinal WTGROUP variable (1,2,3,4) as well as three dummy variables called WEIGHT2, WEIGHT3, and WEIGHT4. Other variables of interest, which will play the role of confounders or effect modifiers, are sex and age (dichotomized at 60 years). Tables12–1 and 12–2 show my coding rules. Please refer back to these tables later, when you need to interpret some of the printout.

Table 12–1. Variables and their coded values

Variable Name	Variable Values*	
WEIGHT	50-143 kg (continuous)	
OVERWEIGHT	1 = overweight	0 or 2 = normal weight
APNEA	1 = sleep apnea	0 or 2 = no sleep apnea
SEX	1 = male	0 or 2 = female
AGEBIN (age, BINary)	1 = "old" (age>60)	0 or 2 = "young" (age≤60)

* For SAS-related technical reasons, I used the value of 2 in tabular analysis and the value of 0 in regression models

Table 12–2. Four categories of weight and their coding scheme

Weight Category (kg)	WEIGHT2	WEIGHT3	WEIGHT4	WTGROUP
< 70	0	0	0	1
70 – 79	1	0	0	2
80 – 89	0	1	0	3
≥ 90	0	0	1	4

Tabular methods

Two methods are used to estimate odds ratios: tabular analysis, also known as *contingency tables* or *cross-classification tables*, and logistic regression, a commonly used regression model. We will start with the simpler method: tables.

SAS code

```
PROC FREQ;
 TABLES overweight*apnea;
run;
```

PROC FREQ (short for frequency) is a SAS procedure for displaying a cross-classification table of the "multiplied" variables, written to the right of the **TABLES** statement. The variable on the left (overweight) will be listed as the row variable.

Selected SAS printout

The FREQ Procedure

Table of overweight by apnea

overweight	apnea		
Frequency Percent Row Pct Col Pct	1	2	Total
1	61 5.68 39.10 36.75	95 8.85 60.90 10.46	156 14.53
2	105 9.78 11.44 63.25	813 75.70 88.56 89.54	918 85.47
Total	166 15.46	908 84.54	1074 100.00

The printout contains not only counts per cell and per margin, but also row percentages, column percentages, and percentages of the total sample of 1,074 observations. Because our causal theory dictates the order OVERWEIGHT→APNEA, column percentages (the distribution of OVERWEIGHT in each category of APNEA) and percentages of the total sample are irrelevant, and will be suppressed from now on.

To compute the marginal ("crude") odds ratio for sleep apnea, divide the odds of sleep apnea in overweight people (61/95) by the odds of sleep apnea in their normal-weight counterparts (105/813).

$OR_{MARGINAL}$ = (61/95) / (105/813) = 0.6421 / 0.1292 = 4.97

Many introductory courses and textbooks teach you to compute the odds ratio from a 2x2 table by diagonal multiplication of the cells (61x813; 95x105) followed by division of the two products: OR = (61x813) / (95x105) = 4.97. I advise you to never use that method. First, you can't see any ratio of two odds in that formula, so you will be following technical computation instead of thinking about two odds of the outcome and dividing them. Second and more important: if the four cells are not organized in the order above (overweight as the row variable; overweight and apnea in the left upper cell), diagonal multiplication and division might lead you to compute the odds ratio for *not* having sleep apnea.

As we saw in chapter 3, the odds are defined as the ratio of two complementary probabilities (or percentages): Odd (sleep apnea)=Pr (sleep apnea)/Pr (no sleep apnea). Therefore, the marginal odds ratio may also be computed from the table's four row percentages:

$OR_{MARGINAL}$ = (39.10/60.90) / (11.44/88.56) = 4.97

If you use the code below, adding **CMH** as an option after the slash, SAS will compute the odds ratio (and a lot more, which I deleted).

```
PROC FREQ;
 TABLES overweight*apnea/NOCOL NOPERCENT CMH;
run;
```

Selected SAS printout

The FREQ Procedure
Table of overweight by apnea

overweight	apnea		
Frequency Row Pct	1	2	Total
1	61 39.10	95 60.90	156
2	105 11.44	813 88.56	918
Total	166	908	1074

Summary Statistics for overweight by apnea

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	4.9717	3.3985	7.2732

Under the misnomer "Common Relative Risk" and to the right of the misnomer "Case-Control", you find the odds ratio of interest, identical to what we have computed (4.97). In the context of our example, that number is simply the odds ratio for prevalent sleep apnea: the odds of prevalent sleep apnea in overweight people divided by the odds of prevalent sleep apnea in their normal-weight counterparts, as defined here. Because the odds of prevalent disease are sometimes called the "prevalence odds" (to distinguish them from the odds of incident disease, which are called the "incidence odds"), the number 4.97 may also be called the "prevalence odds ratio".

The standard error of the (log) odds ratio

Assuming that confounders are lurking in the background, the estimator behind 4.97 is biased and, therefore, the standard error of that estimator is of no interest. Nor do we learn anything useful from any standard error-based computation, such as a confidence interval or a confidence limit ratio (chapter 8). Nonetheless, for pedagogical reasons let's assume that 4.97 emerged from an unbiased estimator and follow the method by which SAS has calculated the standard error and the 95% confidence interval.

You have already seen (chapter 8) that the sampling (or replication) distribution of ratio measures of effect is not bell-shaped, so the standard error of that distribution does not describe well the spread of point estimates around the expected value. You may also recall that the problem was solved by switching to the log (OR) scale and focusing on the bell-shaped distribution of the log (OR). (I will use the natural logarithm throughout, but keep the symbol "log" rather than "ln".) What, then, is the standard error of the log (OR)? Or, what is the variance of the log (OR)?

Using the notation of Table 12–3, the following formula is typically used for that variance:

$$Var[\log(OR)] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

The standard error is, of course, the square root of this expression.

$$SE[\log(OR)] = \sqrt{\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)}$$

Notice that the standard error is a function of the counts in the four cells, so it's crucial to have "enough" observations in each cell. In fact, sparse data in just one cell could

markedly inflate the variance even if others contain large counts. For example, if one cell contains just one observation (say, a=1), the inverse of that cell count is 1/1=1 and the variance would be at least 1, no matter how large are the counts in the other three cells.

Table 12–3. General notation of the 2x2 table

overweight	apnea	
Frequency Row Pct	1	2
1	a	b
2	c	d
		T

Plugging in the values from the SAS output for obesity and sleep apnea, we get:

$$SE[\log(OR)] = \sqrt{\left(\frac{1}{61} + \frac{1}{95} + \frac{1}{105} + \frac{1}{813}\right)} = 0.194$$

Using the standard error, we can compute three kinds of 95% confidence limits:

CI for the log(OR): $\log(4.97) \pm 1.96 \times 0.194 = 1.6034 \pm 0.3802 =$ [1.2232, 1.9836]

CI for the OR: [exp(1.2232), exp(1.9836)] = [3.4, 7.3]

Confidence limit ratio (CLR) for the OR: 7.3/3.4 = 2.1

The 95% confidence limits for the OR (3.4, 7.3) match the numbers on the SAS output (3.3985, 7.2732), after rounding.

Deconfounding the odds ratio

On the assumptions of the naïve diagram below (Figure 12–1), the marginal association between overweight status and sleep apnea status contains not only the effect of weight but also confounding by age. As was explained in chapter 7, you can deconfound the odds ratio by following three steps:

1. Stratify the sample on age (a binary variable in this example)
2. Estimate the odds ratio for the effect of overweight on sleep apnea in each age group
3. On the assumption of no effect modification, compute a weighted average of the two estimates

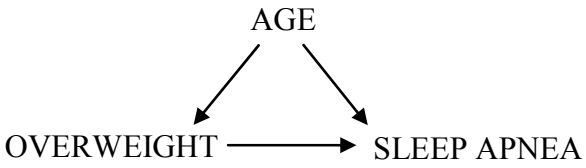

Figure 12–1. A causal diagram relating age, overweight, and sleep apnea.

To generate stratified tables by SAS, we simply add the stratification variable to the **TABLES** statement as the left-most variable. SAS will display a table for each stratum and even a weighted average of the odds ratio and its confidence interval. And in general: The SAS statement **TABLES var1*var2*var3*...*overweight*apnea** will replicate the cross-classification of OVERWEIGHT and APNEA in all of the possible combinations of the preceding variables.

SAS code

```
PROC FREQ;
 TABLES agebin*overweight*apnea/NOCOL NOPERCENT CMH;
run;
```

Selected SAS printout

The FREQ Procedure

Table 1 of overweight by apnea
Controlling for agebin=1

overweight apnea

Frequency Row Pct	1	2	Total
1	36 46.15	42 53.85	78
2	71 14.34	424 85.66	495
Total	107	466	573

OR=(36/42)/(71/424)=5.12

Table 2 of overweight by apnea
Controlling for agebin=2

overweight apnea

Frequency / Row Pct	1	2	Total
1	25 32.05	53 67.95	78
2	34 8.04	389 91.96	423
Total	59	442	501

OR=(25/53)/(34/389)=5.40

Summary Statistics for overweight by apnea
Controlling for agebin

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	5.2324	3.5541	7.7032
	Logit	5.2360	3.5572	7.7070

SAS does not compute the stratum-specific odds ratios, so I added the computation to the right of each table. Since the stratum-specific odds ratios are fairly similar, we don't have to worry about effect modification by age. Their weighted average is shown at the bottom of the printout under the heading "value". We find two, albeit similar, estimates (5.2324 and 5.2360) and, again, two similar 95% confidence limits.

How are the estimates computed?

Neither is a simple average, for reasons that will become clear shortly. The estimate computed by the Mantel-Haenszel method (5.2324) is a weighted average of the age-specific odds ratios:

$$OR_{Mantel-Haenszel} = \frac{5.12w_1 + 5.40w_2}{w_1 + w_2}$$

whereas the estimate computed by the logit method (5.2360) is computed by first taking a weighted average of the *log* odds ratios

$$\log(OR_{\log it}) = \frac{\log(5.12)w_1 + \log(5.40)w_2}{w_1 + w_2}$$

followed by exponentiation of the result to return to the odds ratio scale:

$$OR_{\log it} = \exp\left(\frac{\log(5.12)w_1 + \log(5.40)w_2}{w_1 + w_2}\right)$$

What are those weights?

On the assumption of homogeneity of the causal parameter across the strata of age, our weights should reflect the spread of the sampling distribution from which each estimate arose. An estimate that arose from a tighter distribution should carry more weight. Because the standard error (or the variance) measures the spread of a sampling distribution, the inverse of the variance (1/variance) may be a good choice for our weights: the smaller the variance (of the age-specific estimator), the larger is the inverse of the variance and the greater is the weight we would assign to the estimate.

The Mantel-Haenszel Method

Historically, the Mantel-Haenszel method preceded the logit method, and the formula for the weight in that approach only approximates the inverse of the variance (when the odds ratio is not too far from 1…) Using the layout and notation of Table 12–3, the weight (w) in the Mantel-Haenszel formula is "bc/T". Therefore,

For the old age group (AGEBIN=1): $w_1 = 42 \times 71 / 573 = 5.204$
For the young age group (AGEBIN=2): $w_2 = 53 \times 34 / 501 = 3.597$

There is no surprise here. Examining the age-specific 2x2 tables for "old"(AGEBIN=1) and "young"(AGEBIN=2), we generally find larger counts in the table for "old" people. More data means a tighter sampling distribution behind their odds ratio, which justifies a larger weight.

After entering the two weights into the formula, the result matches the number on the printout:

$$OR_{Mantel-Haenszel} = \frac{5.12x5.204 + 5.40x3.597}{5.204 + 3.597} = 5.23$$

To derive a general formula for $OR_{Mantel\text{-}Haenszel}$ for any number of confounders and any number of strata, I displayed the data for the *i*-th stratum by adding the subscript "*i*", using generic terms for "exposure status" (exposed; unexposed) and "disease status" (diseased; disease-free). The notation is shown in Table 12–4.

Table 12–4. General notation of the 2x2 table for the *i*-th stratum

	Diseased	Disease-free	
Exposed	a_i	b_i	
Unexposed	c_i	d_i	
			T_i

$OR_i = (a_i/b_i)/(c_i/d_i)$

$w_i = b_i c_i / T_i$

The weighted average of all OR_i would be

$$\text{OR}_{\textit{Mantel-Haenszel}} = \frac{\sum OR_i w_i}{\sum w_i} = \frac{\sum (a_i / b_i)/(c_i / d_i)_i (b_i c_i / T_i)}{\sum b_i c_i / T_i} = \frac{\sum (a_i d_i / b_i c_i)_i (b_i c_i / T_i)}{\sum b_i c_i / T_i}$$

$$\text{OR}_{\textit{Mantel-Haenszel}} = \frac{\sum a_i d_i / T_i}{\sum b_i c_i / T_i} \qquad \text{(Equation 12–1)}$$

If you compute $\text{OR}_{\textit{Mantel-Haenszel}}$ by hand, I suggest that you calculate each OR_i and each w_i, and then enter the numbers into the original formula of a weighted average rather than taking a shortcut via Equation 12–1. (It is always a good idea to see the stratum-specific estimates and their relative weights.) Like any estimator, $\text{OR}_{\textit{Mantel-Haenszel}}$ (actually, the log of it) carries along a standard error, but the formula requires heavy notation and is not shown here.

The "logit" method

In the "logit" method, we switch to the log scale and use the (inverse) of the variance of the log odds ratio as our weight. As we saw earlier, that variance is the sum of the inverse of the cell's counts:

$$Var[\log(OR)] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

So, the weight of the estimated OR from the *i*-th stratum is

$$w_i = 1/Var[\log(OR_i)] = 1/\left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}\right)$$

For the old age group (AGEBIN=1): $w_1 = 1/\left(\frac{1}{36} + \frac{1}{42} + \frac{1}{71} + \frac{1}{424}\right) = 14.699$

For the young age group (AGEBIN=2): $w_2 = 1/\left(\frac{1}{25} + \frac{1}{53} + \frac{1}{34} + \frac{1}{389}\right) = 11.007$

And the weighted average of the log (OR) is

$$\log(OR_{\log it}) = \frac{\log(5.12)x14.699 + \log(5.40)x11.007}{14.699 + 11.007} = 1.656$$

Transforming back to the odds ratio scale:

$$OR_{logit} = e^{1.656} = 5.24$$

Extension to any number of strata is straightforward, but there is no short version of the formula:

$$OR_{logit} = \exp\left(\frac{\sum \log(OR_i) w_i}{\sum w_i}\right) \qquad \text{(Equation 12–2)}$$

where $w_i = 1/\left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}\right)$, as we saw earlier

The standard error of $\log(OR_{logit})$ is a function of the stratum-specific weights:

$$SE[\log(OR_{logit})] = \frac{1}{\sqrt{\sum w_i}}$$

*

Notice that in our example the two methods used different numbers for the weights, but the relative weights themselves were similar. Compare, for example, the ratio of the two Mantel-Haenszel weights (5.204/3.597=1.4) to the ratio of the "logit" weights (14.699/11.007=1.3). They are similar.

After conditioning on age, the odds ratio (5.2) is a little larger than the marginal value (4.97), implying weak *negative confounding*: the marginal association underestimated the (presumed) effect. Theoretically, we might have expected to observe positive confounding (attenuation after conditioning on age) because age should have a positive association with both sleep apnea and weight. In this sample, however, there was a weak inverse association between overweight and age, which has caused negative confounding.

Again, if you wish to condition on several categorical confounders simultaneously, simply add the conditioning variables to the **TABLES** statement:

```
TABLES var1*var2*var3*...*overweight*apnea/NOCOL NOPERCENT CMH;
```

SAS will create a 2x2 table of "overweight x apnea" in all of the possible joint strata of the confounders, and compute a weighted average according to the two methods: Mantel-Haenszel and logit. The confounders don't have to be binary variables, but they must be categorical, of course.

Keep in mind that the more variables you add, the thinner you stratify and the smaller is the amount of data in each stratum. Eventually, our precious goal of deconfounding might succumb to the force of randomness, because the weighted average of too many low-weighted estimates may be an average of poor estimates. Logistic regression, which will be presented later, offers a partial remedy to this problem and also opens the door to the modeling of continuous variables.

Dose-response analysis

A binary OVERWEIGHT variable leaves no room for exploring the dose-response function whereas a continuous weight variable cannot be studied by tabular methods. In between, we may learn more about the effect of weight on sleep apnea from categorization of weight into four groups, using the ordinal variable WTGROUP (Table 12–2). A similar approach was described in chapter 9 when we explored the effect of age on blood pressure.

SAS code

```
PROC FREQ;
 TABLES wtgroup*apnea/NOCOL NOPERCENT;
run;
```

Selected SAS printout

The FREQ Procedure

Table of WTGROUP by apnea

WTGROUP	apnea		
Frequency Row Pct	1	2	Total
1	10 3.83	251 96.17	261
2	17 7.02	225 92.98	242
3	32 13.01	214 86.99	246
4	107 32.92	218 67.08	325
Total	166	908	1074

Reference

OR = (17/225)/(10/251) = 1.90

OR = (32/214)/(10/251) = 3.75

OR =(107/218)/(10/251)=12.32

SAS does not display odds ratios on the printout, so I added the estimates to the right of the 4x2 table. Any category of WTGROUP may serve as the reference category, but the lowest (WTGROUP=1) should be a natural choice. The estimates I computed also allow you to derive an odds ratio for any other causal contrast of interest, simply by dividing two of these numbers. For example, the effect of WTGROUP=4 versus WTGROUP=2 on sleep apnea may be computed as 12.32 /1.90 = 6.5, because

$$12.32/1.90 = [(107/218)/(\underline{10/251})] \,/\, [(17/225)/(\underline{10/251})] = \underbrace{(107/218)}_{\text{Odds(apnea) WTGROUP=4}} / \underbrace{(17/225)}_{\text{Odds(apnea) WTGROUP=2}} = 6.5$$

Displaying the dose-response function

When the effect was a continuous variable, such as blood pressure, we explored the dose-response relation by displaying the mean values of the dependent variable across the exposure categories—the so-called step function. For example, in chapter 9 mean systolic blood pressure in four age groups was depicted as four horizontal lines, later connected by vertical lines to create a continuous graph (Figure 9–5.)

When the effect resides in a binary variable, such as APNEA, the natural "response" is rate, probability or odds, so we may depict the odds of sleep apnea in each weight category as a horizontal line, analogous to mean blood pressure in each age category. Nonetheless, there is one important difference between a step function for the mean and a step function for the odds: Whenever we choose to measure the effect on a ratio scale (here, odds ratio), the dose-response function should be displayed on a logarithmic scale, rather than on an arithmetic scale. Alternatively, we may compute the *log* of the odds and display log values on an arithmetic scale. The same rule applies to probability and rate, whenever ratios serve as their derived measures of effect.

To understand why a log-scale graph should follow the computation of a ratio measure of effect, consider the following example: Suppose that the odds of sleep apnea have doubled between adjacent categories of weight, such as 0.1, 0.2, 0.4, and 0.8 for the four ascending categories. If these numbers were depicted as horizontal lines on an arithmetic scale, the vertical distance between adjacent lines would not correspond to the constant effect, which is an odds ratio of 2. For instance, the effect of WTGROUP=4 versus WTGROUP=3 on sleep apnea (0.8–0.4=0.4) would seem twice as large as the effect of WTGROUP=3 versus WTGROUP=2 (0.4–0.2=0.2), even though these causal contrasts show an identical effect on a ratio scale (0.8/0.4 = 0.4/0.2 = 2). By switching to the log scale, we prevent such a false visual impression because the difference between the logs of two numbers mirrors the ratio of the original numbers. For example:

(1) $\log(0.8)-\log(0.4) = \log(0.4)-\log(0.2)$
(2) $\log(0.8/0.4) = \log(0.4/0.2)$
(3) $0.8/0.4 = 0.4/0.2$

The second equality is derived from the first by an arithmetic rule for logarithms, which you will find in the Appendix: $\log(a)-\log(b) = \log(a/b)$

If the odds ratios for adjacent categories are not identical, the log scale provides correct visual impression of a ratio-based, dose-response function. To illustrate, I computed the "response" (odds and log-odds of sleep apnea) for each category of weight (Table 12–5) and depicted these numbers as a step function (Figure 12–2): the odds are displayed on a logarithmic Y-axis (left panel) whereas the log-odds are displayed on an arithmetic Y-axis (right panel). As you see, both approaches led to the same graphical shape. It is not entirely clear, however, how to draw a smooth line through this graph and whether a straight line would capture the true dose-response function behind our arbitrary categorization of weight. Logistic regression will provide other means to address this issue.

Table 12–5. Odds and log odds of sleep apnea, by weight category

Weight Category (kg)	Odds (sleep apnea)	Log odds (sleep apnea)
< 70	10/251= 0.0398	–3.223
70 – 79	17/225 =0.0755	–2.583
80 – 89	32/214 =0.1495	–1.900
≥ 90	107/218 =0.4908	–0.712

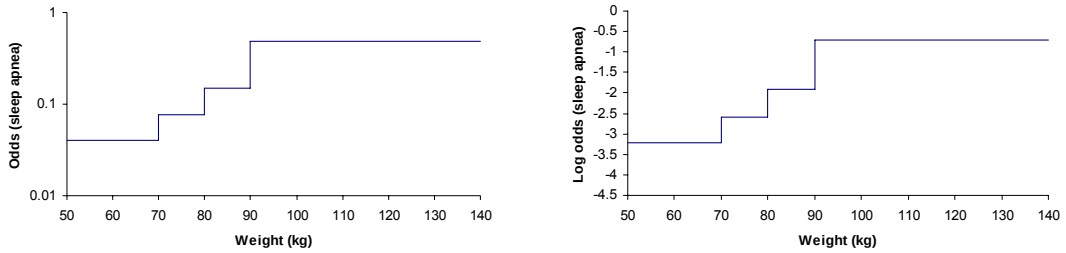

Figure 12–2. Two methods to display a ratio-based, dose-response function for the effect of weight on sleep apnea: logarithmic scale (odds, left panel) and arithmetic scale (log-odds, right panel).

Logistic regression

Just as linear regression allowed us to estimate the mean difference—whether marginal, conditional or modified (chapters 9 and 10) —logistic regression will allow us to estimate marginal, conditional and modified odds ratios. The analogy gets even closer when we compare logistic regression to linear regression of log-transformed continuous variable (chapter 11). In both models the dependent variable is the log of something and in both models the final product of interest is a ratio measure of effect: a geometric mean ratio (linear regression after log transformation) or an odds ratio (logistic regression).

A simple logistic regression model for the marginal association between OVERWEIGHT (1=overweight; 0=normal weight) and APNEA (1=sleep apnea; 0=no sleep apnea) takes the following form:

$$\text{Log odds (APNEA=1)} = \beta_0 + \beta_1 \text{ OVERWEIGHT}$$

To understand the meaning of the coefficient β_1, compare the model above to the linear regression model below:

$$\text{Mean Y} = \beta_0 + \beta_1 \text{ OVERWEIGHT}$$

In linear regression, β_1 estimates the mean difference in Y for the causal contrast between overweight and normal weight. In logistic regression, β_1 estimates the difference in the *log odds* of sleep apnea for the same causal contrast. Since the difference between the logs of two numbers may also be expressed as the log of their ratio, log(a)–log(b)=log(a/b), the difference between two log odds of sleep apnea is also the *log of the*

odds ratio for sleep apnea: $\beta_1 = \log OR$

Therefore, the odds ratio for sleep apnea, contrasting overweight and normal weight, is computed by exponentiation of β_1: $\exp(\beta_1) = OR$

We see again the analogy to linear regression of a log-transformed continuous variable (chapter 11) where the exponentiated coefficient delivered a ratio measure of effect. In that model it was the geometric mean ratio; in logistic regression it is the odds ratio.

Another way to derive the meaning of β_1 is shown in Table 12–6.

Table 12–6. Computing the effect of overweight (1=overweight; 0=normal weight) on sleep apnea from a simple logistic regression model

Causal assignments	log odds (APNEA=1) $= \beta_0 + \beta_1$ OVERWEIGHT
OVERWEIGHT = 1	log odds (APNEA=1) $= \beta_0 + \beta_1 \times 1$
OVERWEIGHT = 0	log odds (APNEA=1) $= \beta_0 + \beta_1 \times 0$
Effect of 1 unit increment	difference in log odds or $=$ $\boldsymbol{\beta_1}$ log of the odds ratio

In this approach, which mimics similar tables for the mean difference (chapters 9 and 10), we compute two predicted log odds under two causal assignments and calculate the difference between them, which turns out to be β_1. Notice that the intercept in logistic regression, just as in linear regression, helps to predict the value of the dependent variable but does not deliver any knowledge about causal effects. Estimating an effect and predicting the value of a dependent variable are different tasks.

The equation Log odds (APNEA=1) $= \beta_0 + \beta_1$ OVERWEIGHT
is my preferred expression of the logistic regression model, because the right hand side is a simple linear term and we can see the analogy to linear regression and to other regression models. Nonetheless, there are three alternative equations, which may be easily derived from the equation above:

(1) Odds (APNEA=1) $= \exp(\beta_0 + \beta_1 \text{ OVERWEIGHT})$

(2) $$\Pr(\text{APNEA=1}) = \frac{\exp(\beta_0 + \beta_1 \text{ OVERWEIGHT})}{1 + \exp(\beta_0 + \beta_1 \text{ OVERWEIGHT})}$$

(3) $$\Pr(\text{APNEA=1}) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 \text{ OVERWEIGHT})]}$$

Expression (1) follows my preferred equation: if "Log odds (A=1) = B", then

"Odd(A=1)=exp(B)". Expression (2) follows the first because probability and odds are connected by the following equality (in simplified notation): Pr = Odds/ (1 + Odds). Expression (3) is derived from the second after dividing the numerator and denominator by the numerator and reversing the order of the terms in the new denominator.

Having understood the arithmetic principles of a simple logistic regression model, you might have realized that extensions to the multi-variable setting are analogous to what you have seen in linear regression. In fact, the right hand side of every model we fit in chapters 9 and 10 may be exported into the right hand side of logistic regression—including dummy variables, continuous variables, quadratic terms, confounders, and interaction terms. Ample examples will follow, but the only difference would be the scale on which the effect is estimated. Instead of estimating mean differences, the coefficients will be estimating log odds ratios. And instead of estimating effect modification on the additive scale, we will be estimating effect modification on the multiplicative scale. Again, the analogy would seem even closer if you recall the computation of various geometric mean ratios (chapter 11).

The binomial likelihood function

An equation such as "log odds (APNEA=1)= $\beta_0 + \beta_1$ OVERWEIGHT" from which we can compute the odds ratio does not help much—unless we find a method to estimate the coefficients from the data. To explain the estimation method as simply as possible, let's pretend that our sample includes only 5 people, instead of 1,074, whose names are Jeff, John, Sandy, Sara, and Jack. Each of these five people has contributed a value of OVERWEIGHT and a value of APNEA, and the relation between the two variables is summarized in Table 12–7. In this small sample the odds ratio for sleep apnea is (2/1)/(1/1)=2, for the contrast between overweight and normal weight.

Table 12–7. Cross-classification of overweight and apnea in a small hypothetical sample (N=5)

	apnea	
overweight	1	0
1	**2** Jeff, John	**1** Sandy
0	**1** Sara	**1** Jack

How can we estimate that number by fitting the logistic regression model "log odds (APNEA=1)= $\beta_0 + \beta_1$ OVERWEIGHT" ?

Or in other words, how can we estimate β_1, which is the log of the odds ratio?

The process requires several steps, collectively called "maximum likelihood estimation", a fairly scary term, which need not intimidate anyone with knowledge of high school

algebra and one simple rule of probability: the probability of observing a series of independent events is the product of the probability of observing each event in that series.

In the first step, we construct a function called "likelihood" (often denoted by the letter L), which corresponds to the probability of observing our data. Simplistically, we may write:

Likelihood = Pr("data")

Next, we should state explicitly what we mean by Pr("data"). The "data" here are 5 observations of 5 people, some of whom have apnea (APNEA=1) and some of whom do not (APNEA=0); some of whom are overweight (OVERWEIGHT=1) and some of whom are not (OVERWEIGHT=0). According to our causal theory, the probability of the apnea status of each person depends on the person's weight status. In statistical notation we may write, for example, Pr(APNEA=1/OVERWEIGHT=1). Read: "the probability that a person has apnea, given that the person is overweight." Or another example: Pr(APNEA=0/OVERWEIGHT=0). Read: "the probability that a person does not apnea, given that the person is not overweight." Table 12–8 shows the five observations and their associated probabilities, in notation.

Table 12–8. Apnea status, overweight status, and the associated probability (N=5)

Name (Observation #)	Observed apnea status	Observed overweight status	Probability of the observation ("data")
Jeff (1)	APNEA=1	OVERWEIGHT=1	Pr(APNEA=1/OVERWEIGHT=1)
John (2)	APNEA=1	OVERWEIGHT=1	Pr(APNEA=1/OVERWEIGHT=1)
Sandy (3)	APNEA=0	OVERWEIGHT=1	Pr(APNEA=0/OVERWEIGHT=1)
Sara (4)	APNEA=1	OVERWEIGHT=0	Pr(APNEA=1/OVERWEIGHT=0)
Jack (5)	APNEA=0	OVERWEIGHT=0	Pr(APNEA=0/OVERWEIGHT=0)

On the assumption of independence between the observations (for instance, Jeff's apnea status does not depend on Jack's apnea status), the probability of observing what we have observed should be the product of the five individual probabilities, shown in the right column of Table 12–7. This product is called the likelihood of (observing) the data.

Likelihood =	Pr(APNEA=1/OVERWEIGHT=1)	Jeff
	x	
	Pr(APNEA=1/OVERWEIGHT=1)	John
	x	
	Pr(APNEA=0/OVERWEIGHT=1)	Sandy
	x	
	Pr(APNEA=1/OVERWEIGHT=0)	Sara
	x	
	Pr(APNEA=0/OVERWEIGHT=0)	Jack

We'll keep this product term in mind for a moment, and turn back to the logistic regression model (using expression 2). That model has specified the relation between a person's weight and his or her probability of having sleep apnea as follows:

$$\Pr(\text{APNEA}=1) = \frac{\exp(\beta_0 + \beta_1 \text{ OVERWEIGHT})}{1 + \exp(\beta_0 + \beta_1 \text{ OVERWEIGHT})}$$

We may use this expression for each person, namely, enter the values of APNEA and OVERWEIGHT for Jeff, John, Sandy, Sara, and Jack, and derive 5 expressions for their 5 probabilities. Each expression will contain one or both of the unknown coefficients, β_0 and β_1. Here are the five expressions.

Jeff: $$\Pr(\text{APNEA}=1/\text{OVERWEIGHT}=1) = \frac{\exp(\beta_0 + \beta_1 x1)}{1 + \exp(\beta_0 + \beta_1 x1)}$$

John: $$\Pr(\text{APNEA}=1/\text{OVERWEIGHT}=1) = \frac{\exp(\beta_0 + \beta_1 x1)}{1 + \exp(\beta_0 + \beta_1 x1)}$$

Sandy: $$\Pr(\text{APNEA}=0/\text{OVERWEIGHT}=1) = 1 - \Pr(\text{APNEA}=1/\text{OVERWEIGHT}=1)$$

$$= 1 - \frac{\exp(\beta_0 + \beta_1 x1)}{1 + \exp(\beta_0 + \beta_1 x1)}$$

Sara: $$\Pr(\text{APNEA}=1/\text{OVERWEIGHT}=0) = \frac{\exp(\beta_0 + \beta_1 x0)}{1 + \exp(\beta_0 + \beta_1 x0)}$$

Jack: $$\Pr(\text{APNEA}=0/\text{OVERWEIGHT}=0) = 1 - \Pr(\text{APNEA}=1/\text{OVERWEIGHT}=0)$$

$$= 1 - \frac{\exp(\beta_0 + \beta_1 x0)}{1 + \exp(\beta_0 + \beta_1 x0)}$$

If we substitute the last five expressions for the five probabilities in the likelihood formula, the likelihood will be expressed as a function of two unknown values: β_0 and β_1. In mathematical notation: $L = f(\beta_0, \beta_1)$. Read: "the likelihood of observing the data is a function of β_0 and β_1, the regression coefficients."

Obviously, we may enter different pairs of values of β_0 and β_1 into the function we have just derived and compute different values of the likelihood, some larger than others. But

one pair (and only one) will yield the largest value of L. This pair is called the maximum likelihood estimates (MLE) of the likelihood function. Moreover, it turns out that

The maximum likelihood estimate of β_1 = The log of the odds ratio

How do we actually solve the likelihood function and find the maximum likelihood estimates?

In this simple example, a formal step-by-step solution is possible, just as we solve a set of equations with two unknown quantities. In more complex equations, with several covariates, the method requires iteration ("trial and error"), as well as an iteration algorithm.

SAS PROC LOGISTIC

The SAS code below displays the hypothetical data from our small sample of five people and shows how to fit a simple logistic regression model using SAS. As was the case in linear regression, the **MODEL** statement is capturing the essence of the regression model: the dependent variable (here, binary apnea status) is written to the left of the equality sign whereas the predictors (here, only one) are written to the right. The option **RL** after the slash requests the computation of a 95% confidence interval.

SAS code

```
DATA one;
 INPUT name $ APNEA OVERWEIGHT;
 DATALINES;

Jeff   1 1
John   1 1
Sandy  0 1
Sara   1 0
Jack   0 0
 ;
run;

PROC LOGISTIC DESCENDING;
MODEL apnea=overweight/RL;
run;
```

Selected SAS printout

```
                          The LOGISTIC Procedure

                            Model Information

Data Set                      WORK.ONE
Response Variable             APNEA
```

```
Number of Response Levels     2
Model                         binary logit

   Number of Observations Used           5


              Response Profile

      Ordered                       Total
        Value        APNEA     Frequency

            1            1             3
            2            0             2

       Probability modeled is APNEA=1.


           Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.


            Model Fit Statistics

                                  Intercept
                Intercept             and
    Criterion          Only     Covariates

    -2 Log L                        6.592


   Analysis of Maximum Likelihood Estimates

                                       Standard
    Parameter     DF    Estimate         Error

    Intercept      1    2.343E-7        1.4142
    OVERWEIGHT     1      0.6931        1.8708


          Odds Ratio Estimates

                Point          95% Wald
    Effect      Estimate      Confidence Limits

    OVERWEIGHT     2.000       0.051      78.248
```

At the beginning of the printout we find information about the sample and the dependent variable. Technically, either category of a binary dependent variable may be considered the "event", so it is important to know what was modeled: the probability (or odds) of having sleep apnea or the probability (or odds) of *not* having sleep apnea? Here the printout states "`Probability modeled is APNEA=1`", as we prefer. Notice the term "logit" at the top of the printout, which is a synonym of "log-odds".

Statements about "model convergence" refer to the search for the maximum likelihood estimates through a sequence of iterations. This process generates a series of estimates on the way to those that maximize the likelihood function. While approaching the final numbers, the results of successive iterations do not differ much anymore, and

when the differences are "small enough", the convergence criterion is satisfied. At that point, the process stops and the estimated coefficients from the last iteration are reported. In this example, the maximum likelihood estimates were found to be β_0=2.343E-7 and β_1=0.6931. Therefore, the regression model takes the following form:

$$\text{Log odds (APNEA=1)} = 2.343E\text{-}7 + 0.6931 \times \text{OVERWEIGHT}$$

Or alternatively,

$$\Pr(\text{APNEA=1}) = \frac{\exp(2.343E\text{-}7 + 0.6931 \times \text{OVERWEIGHT})}{1 + \exp(2.343E\text{-}7 + 0.6931 \times \text{OVERWEIGHT})}$$

You may wonder what was the maximal value of the likelihood function itself, namely, the largest value of that function as generated by plugging in the maximum likelihood estimates, β_0=2.343E-7 and β_1=0.6931. That number is not shown on the printout, but we do find a derivation, `-2 Log L`, in the section "model fit statistics". The number 6.592 under the column heading "Intercept and Covariates" is equal to (−2) times the log of the maximum value of the likelihood. Therefore,

$$-2 \text{ Log L} = 6.592$$
$$\text{Log L} = -3.296$$
$$\text{L} = \exp(-3.296) = 0.037$$

Notice how small is the likelihood—the probability of observing our data—given the regression model "log odds (APNEA=1)= 2.343E-7 + 0.6931 x OVERWEIGHT". Yet of all possible values of the likelihood function, 0.037 is the maximum. Again, that maximum is reached when β_0=2.343E-7 and β_1=0.6931 are entered into the likelihood function.

Last but not least, we find the key result at the bottom of the printout: the odds of sleep apnea in this hypothetical sample of 5 people were twice as high for overweight people as compared with their normal weight counterparts: $\exp(0.6931) = 2$. A logistic regression model and maximum likelihood estimation converged with the results we had computed earlier by hand from a 2x2 table. I find myself wondering whether that was a miracle.

More on maximum likelihood estimation

I have described above the principles of maximum likelihood estimation using a simple situation: a tiny sample and a logistic regression model with a single predictor. The very same principles, however, are followed when the sample is much larger; when the regression equation contains several predictors; and even when other regression models are used. It is possible, for example, to write a likelihood function for a multiple linear regression model and search for its largest value. Interestingly, in that case the maximum likelihood estimates would be identical to the estimates we would get by the method of least-squares regression (chapter 9). In other words, the estimates from ordinary least-square regression are actually maximum likelihood estimates! Indeed, maximum likelihood estimation is intimately connected to almost every regression model.

Different likelihood functions for different kinds of regression models are developed along different mathematical trails, and often require heavy notations and complex equations. (You will see another example in chapter 17.) Nonetheless, they all share a common goal: to express the likelihood of observing the data as a function of unknown regression coefficients. In generic notation: $L = f(\beta_0, \beta_1, \beta_2, \ldots, \beta_n)$. Then, we search for the set of values of these coefficients that will generate the largest value of L—the so-called maximum likelihood estimates. Not every likelihood function has a maximum, but when it does exist we can find the maximum likelihood estimates through an iterative process of trial and error.

You have already seen that a log-derivation of the likelihood is reported on the SAS printout (`-2 Log L`). For technical reasons (ease of computation), we often prefer to work with the log of the likelihood from the very beginning—that is, look for the maximum likelihood estimates of the following function:

$$\log\text{–}L = f(\beta_0, \beta_1, \beta_2, \ldots, \beta_n)$$

Of course, the set of coefficients that yield the largest value of the log-likelihood function will also yield the largest value of the likelihood function itself. The scale does not matter.

Overweight and sleep apnea: logistic regression models

Earlier in this chapter we used tabular methods to compute odds ratios for the effect of weight on sleep apnea. First, we studied the marginal association between APNEA and the binary exposure, OVERWEIGHT, in a 2x2 table. Next, we assumed confounding by age, and deconfounded by stratifying on the binary variable, AGEBIN. Finally, we studied the dose-response function using a 4x2 table: cross-classification of four categories of weight and two categories of apnea. All of these odds ratios will be computed again from logistic regression models. Flip back and fourth between the tables and the regression output to convince yourself that the odds ratios are identical.

The first model takes the following form:

$$\text{Log odds (APNEA=1)} = \beta_0 + \beta_1 \text{ OVERWEIGHT}$$

SAS code

```
PROC LOGISTIC DESCENDING;
 MODEL apnea = overweight;
run;
```

Selected SAS printout

```
                    The LOGISTIC Procedure

                    Model Information

Data Set                          WORK.TWO
Response Variable                 apnea
```

```
Number of Response Levels     2
Model                         binary logit


Number of Observations Used        1074


        Response Profile

Ordered                    Total
Value        apnea     Frequency

1            1           166
2            0           908

Probability modeled is apnea=1.


Analysis of Maximum Likelihood Estimates

                                  Standard
Parameter     DF    Estimate       Error

Intercept      1     -2.0468      0.1037
overweight     1      1.6038      0.1941


        Odds Ratio Estimates

                Point          95% Wald
Effect         Estimate      Confidence Limits

overweight       4.972       3.399       7.273
```

log odds (APNEA=1) = –2.0468 + 1.6038 OVERWEIGHT

The sample size (1,074 observations) on this output matches the grand total in the first 2x2 table, and so does the count of people with apnea (166). The coefficient of OVERWEIGHT is the log (OR) for the contrast between OVERWEIGHT=1 and OVERWEIGHT=0. After exponentiating that coefficient, we get an odds ratio of 4.97, identical to the odds ratio from the 2x2 table. Keep in mind that the standard error of the coefficient of OVERWEIGHT is the standard error of the log (OR), not the standard error of the OR.

To compute the 95% confidence interval, start on the log scale and compute the lower and upper limit for the log(OR). Then, exponentiate the results to return to the OR scale.

	Log OR	OR
Lower limit:	1.6038 – 1.96 x 0.1941 = 1.2234	exp(1.2234) = 3.399
Upper limit:	1.6038 + 1.96 x 0.1941 = 1.9842	exp(1.9842) = 7.273

You might have noticed that the confidence intervals are slightly different from those found in tabular methods. That's because the standard error is not computed according to the Mantel-Haenszel formula. More important, notice that the log odds ratio (1.6038) resides in the middle of the confidence interval, at the midpoint between 1.2234 and 1.9842, but the odds ratio (4.97) does not. The distance between 4.97 and 3.399 (the lower limit) is not equal to the distance between 4.97 and 7.273 (the upper limit), because the symmetry on the log scale disappears after exponentiation. So, if you see a symmetrical confidence interval around the odds ratio (or around another ratio measure of effect), someone might have made a mistake.

Analogous to stratification on age in tabular analysis, we will simply add AGEBIN as a covariate. Again, in all regression models, adding covariates should deconfound the association of interest, assuming we have a causal diagram in mind, which supports the need to deconfound.

The model takes the following form:

$$\text{Log odds (APNEA=1)} = \beta_0 + \beta_1 \text{ OVERWEIGHT} + \beta_2 \text{ AGEBIN}$$

SAS code

```
PROC LOGISTIC DESCENDING;
 MODEL apnea = overweight agebin;
run;
```

Selected SAS printout

```
                The LOGISTIC Procedure

                Model Information

Data Set                      WORK.TWO
Response Variable             apnea
Number of Response Levels     2
Model                         binary logit


Number of Observations Used        1074


                Response Profile

Ordered                  Total
Value        apnea     Frequency

1            1           166
2            0           908

Probability modeled is apnea=1.
```

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error
Intercept	1	-2.4266	0.1593
overweight	1	1.6555	0.1973
agebin	1	0.6340	0.1833

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
overweight	5.236	3.557	7.708
agebin	1.885	1.316	2.700

log odds (APNEA=1) = –2.4266 + 1.6555 OVERWEIGHT + 0.6340 AGEBIN

Focus again on the coefficient of OVERWEIGHT (1.6555), and ignore the coefficient of AGEBIN. After exponentiation, we get an estimate of the conditional odds ratio: exp(1.6555) = 5.236. The confidence interval is [3.557, 7.708]. Below, I copied comparable results from tabular analysis, which you can find yourself by flipping to the beginning of this chapter. Ignoring rounding-related differences, the numbers on the "logit" line are identical. Indeed, the method used in tabular analysis is called "logit" because the point estimate and the standard error are computed just as they are computed in logistic regression.

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	5.2324	3.5541	7.7032
	Logit	5.2360	3.5572	7.7070

To replicate the tabular dose-response analysis, we will model three dummy variables for the four weight categoris (Table 12–2.). I decided to label the intercept β_1 rather than β_0 to match the subscripts of coefficients and dummy variables.

$$\text{Log odds (APNEA=1)} = \beta_1 + \beta_2 \text{ WEIGHT2} + \beta_3 \text{ WEIGHT3} + \beta_4 \text{ WEIGHT4}$$

SAS code

```
PROC LOGISTIC DESCENDING;
```

```
 MODEL apnea = weight2 weight3 weight4;
run;
```

Selected SAS printout

The LOGISTIC Procedure

Model Information

Data Set	WORK.TWO
Response Variable	apnea
Number of Response Levels	2
Model	binary logit

Number of Observations Used 1074

Response Profile

Ordered Value	apnea	Total Frequency
1	1	166
2	0	908

Probability modeled is apnea=1.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error
Intercept	1	-3.2228	0.3225
weight2	1	0.6399	0.4090
weight3	1	1.3226	0.3740
weight4	1	2.5112	0.3434

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
weight2	1.896	0.851	4.227
weight3	3.753	1.803	7.812
weight4	12.319	6.285	24.147

log odds (APNEA=1) = –3.2228 + 0.6399 WEIGHT2 + 1.3226 WEIGHT3 + 2.5112 WEIGHT4

The intercept estimates the log odds of sleep apnea for the reference group (<70 kg), whereas each coefficient of a dummy variable estimate the log (OR) for the causal

contrast between that category and the reference. After exponentiating the coefficients, we get the following odds ratios (rounded): 1.90, 3.75, and 12.32. Please check that identical numbers were computed earlier from a 4x2 table.

It is just as easy to display the dose-response function from the regression model as it was from tabular methods. Plug in the values of the dummy variables for each category of weight and compute the predicted log odds of sleep apnea (Table 12–9).

Table 12–9. Computing the dose-response function from the regression model

Weight category (kg)	Values of dummy variables	Log odds (APNEA=1) = –3.2228 + 0.6399 WEIGHT2 + 1.3226 WEIGHT3 + 2.5112 WEIGHT4
< 70	WEIGHT2=0 WEIGHT3=0 WEIGHT4=0	**–3.2228** + 0.6399 x 0 + 1.3226 x 0 + 2.5112 x 0 = **–3.223**
70-79	WEIGHT2=1 WEIGHT3=0 WEIGHT4=0	**–3.2228 + 0.6399 x 1** + 1.3226 x 0 + 2.5112 x 0 = **–2.583**
80-89	WEIGHT2=0 WEIGHT3=1 WEIGHT4=0	**–3.2228** + 0.6399 x 0 **+ 1.3226 x 1** + 2.5112 x 0 = **–1.900**
≥ 90	WEIGHT2=0 WEIGHT3=0 WEIGHT4=1	**–3.2228** + 0.6399 x 0 + 1.3226 x 0 **+ 2.5112 x 1** = **–0.712**

Compare the results to those we computed by tabular analysis (Table 12–5). They are identical. Therefore, there is no need to display the function again. It was already shown in Figure 12–2.

More on deconfounding by logistic regression

So far we saw no advantage of using logistic regression over tabular analysis, and no clear reason to substitute complex maximum likelihood estimation for contingency tables. Assume, however, that we wish to deconfound the marginal association between OVERWEIGHT and APNEA from the age effect, yet retain the continuous form of the age variable. Tabular analysis cannot handle that situation but the task is trivial in logistic regression as it was in linear regression. Simply place the continuous variable, AGE, next to the variable OVERWEIGHT on the right hand side of the model:

$$\text{Log odds (APNEA=1)} = \beta_0 + \beta_1 \text{ OVERWEIGHT} + \boldsymbol{\beta_2}\ \textbf{AGE}$$

Similarly, if you wish to model the four weight categories by three dummy variables and deconfound at the same time, add AGE to the model:

$$\text{Log odds (APNEA=1)} = \beta_1 + \beta_2 \text{ WEIGHT2} + \beta_3 \text{ WEIGHT3} + \beta_4 \text{ WEIGHT4} + \boldsymbol{\beta_5}\ \textbf{AGE}$$

Just like in linear regression, you may simultaneously model several confounders, both continuous and categorical, and obtain conditional odds ratios for the weight effect. To illustrate, here is SAS code and selected printout from regression of the log odds of apnea on WEIGHT2, WEIGHT3, WEIGHT4, AGE, and SEX.

SAS code

```
PROC LOGISTIC DESCENDING;
 MODEL apnea = weight2 weight3 weight4 age sex;
run;
```

Selected SAS printout

```
                The LOGISTIC Procedure

                Model Information

Data Set                          WORK.TWO
Response Variable                 apnea
Number of Response Levels         2
Model                             binary logit


Number of Observations Used        1074


                Response Profile

        Ordered                     Total
        Value        apnea      Frequency

        1            1            166
        2            0            908

        Probability modeled is apnea=1.

Analysis of Maximum Likelihood Estimates

                                          Standard
Parameter     DF     Estimate       Error

Intercept      1      -6.4293      1.0801
weight2        1       0.4586      0.4200
weight3        1       1.0427      0.3956
weight4        1       2.1959      0.3736
age            1       0.0503      0.0162
sex            1       0.5296      0.2126


        Odds Ratio Estimates

                        Point          95% Wald
```

Effect	Estimate	Confidence Limits	
weight2	1.582	0.694	3.603
weight3	2.837	1.306	6.160
weight4	8.988	4.322	18.695
age	1.052	1.019	1.086
sex	1.698	1.119	2.576

Assuming that the rules of a causal diagram dictate deconfounding from the effects of age and sex, we obtained deconfounded estimates of the effect of weight groups on sleep apnea. The causal contrasts between three ascending weight groups and the reference (<70kg) translates into odds ratios of 1.6, 2.8, and 9.0 (after rounding), all of which are smaller than the marginal odds ratios, which were 1.9, 3.7, and 12.3.

Linear and quadratic dose-response functions

Tabular methods have forced us to categorize the weight variable. By contrast, in logistic regresssion, just as in linear regression, we may also model the continuous WEIGHT variable and even a quadratic function, regressing the log odds of sleep apnea on WEIGHT and WEIGHT^2. (Covariates may always be added, of course, if deconfounding is needed.) Let's take a closer look at two logistic regression models that allow for the continuous variable WEIGHT: linear and quadratic.

Model 1: $\text{Log odds (APNEA=1)} = \beta_0 + \beta_1 \text{ WEIGHT}$

Model 2: $\text{Log odds (APNEA=1)} = \beta_0 + \beta_1 \text{ WEIGHT} + \beta_2 \text{ WEIGHT}^2$

Model (1) is imposing a straight line relation between weight and the log odds of sleep apnea, analogous to a straight line relation between age and mean blood pressure (chapter 9). Again, the coefficient, β_1, is the slope of the regression line—regression of the log odds of sleep apnea on weight.

As you may recall, when we fit the linear regression model "$\text{Mean SBP} = \beta_0 + \beta_1 \text{ AGE}$", we implicitly assumed that the effect of one year of aging on mean blood pressure is constant (identical) for all causal contrasts of 1 year difference, such as [40, 41], [45.5, 46.5], and [68, 69]. In the context of logistic regression, the causal assumption behind model (1) dictates a constant difference *in the log odds*, rather a constant difference in the mean. By fitting model (1) we assume that the effect of gaining 1kg of weight confers the same difference in the log odds of sleep apnea, regardless of the baseline weight: gaining 1 kg on top of 45kg would change the log odds of apnea as much as gaining 1kg on top of 77kg, or on top of any other baseline weight. Notice, however, that a constant difference in the log odds also implies a *constant log odds ratio* [because log(a)−log(b) = log (a/b).] But "log(OR)=constant" implies "OR=constant", too! To sum up, model (1) rests on the causal assumption that a causal contrast between two weights that differ by 1kg has the same effect on sleep apnea, regardless of which weights are compared. The magnitude of that effect is a *constant odds ratio* whose value is $\exp(\beta_1)$. Not a trivial assumption at all.

SAS code

```
PROC LOGISTIC DESCENDING;
 MODEL apnea = weight;
run;
```

Selected SAS printout

```
        The LOGISTIC Procedure

        Model Information

Data Set                      WORK.TWO
Response Variable             apnea
Number of Response Levels     2
Model                         binary logit


Number of Observations Used        1074


               Response Profile

        Ordered                    Total
        Value        apnea     Frequency

        1            1           166
        2            0           908

        Probability modeled is apnea=1.

Analysis of Maximum Likelihood Estimates

                                    Standard
Parameter    DF    Estimate       Error

Intercept     1     -7.0318      0.5342
weight        1      0.0607     0.00571


        Odds Ratio Estimates

             Point          95% Wald
Effect    Estimate      Confidence Limits

weight       1.063       1.051       1.075
```

log odds (APNEA=1) = –7.0318 + 0.0607 WEIGHT

Again, the coefficient of WEIGHT estimates the difference in the log odds of sleep apnea per 1kg weight gain, which is also the log(OR) for sleep apnea per 1kg weight gain. After exponentiation ($e^{0.0607}$), we find that the estimated odds ratio is 1.06. Given the linear

model, that number describes the effect of *any* causal contrast that differs by 1kg: [54, 55], [63.2, 64.2], [98, 99], and so on.

A side note, which I often find helpful: In math, the following approximation of e^x holds when x is small: $e^x \approx 1+x$. Therefore, before formal computation you can easily guess the odds ratio you would get after exponentiating a small coefficient: $e^{0.0607} \approx 1 + 0.0607 = 1.0607$.

To compute the 95% confidence limits around the odds ratio, you have to start on the log scale. First, compute the confidence limits around the log odds ratio:

95% confidence limits for the log(OR): $0.0607 \pm 1.96 \times 0.00571 = [0.0495, 0.0719]$

Then, exponentiate the results:

95% confidence limits for the OR: $[\exp(0.0495), \exp(0.0719)] = [1.051, 1.075]$

These numbers are found at the bottom of the printout.

Sometimes, a one-unit increment on the exposure scale may be too small to convey a meaningful causal contrast. Suppose, for example, that we wish to compute the odds ratio for sleep apnea (and the 95% confidence interval) for 10kg weight gain, rather than 1kg weight gain. A similar challenge was presented in chapter 11, when we computed geometric mean ratios, and the solution is similar, too. Start on the log(OR) scale. If the estimated difference in the log odds (namely, the log odds ratio) is 0.0607 per 1kg weight gain, then the estimated difference per 10kg weight gain should be 0.0607 x 10 = 0.607. Last, return to the odds ratio scale: the estimated odds ratio is exp(0.607) = 1.83

Or in general: if the coefficient of a continuous exposure in a logistic regression mode is β_1, then:

β_1 = log(OR) per 1 unit exposure increment

$\beta_1 \times \Delta$ = log(OR) per Δ unit exposure increment

$\exp(\beta_1 \times \Delta)$ = OR per Δ unit exposure increment

To compute a 95% confidence interval, start again on the log scale.

The 95% CI for any Δ is computed according to the following formula:

$$(\beta_1 \times \Delta) \pm 1.96 \times SE(\beta_1 \times \Delta)$$

A rule of arithmetic for standard errors tells us that $SE(\beta_1 \times \Delta) = SE(\beta_1) \times \Delta$. Therefore, the 95% CI may also be written as

$$(\beta_1 \times \Delta) \pm 1.96 \times SE(\beta_1) \times \Delta$$

For example, for Δ=10 kg, the 95% CI for the log(OR)

$$0.0607 \times 10 \pm 1.96 \times 0.00571 \times 10 = [0.4951, 0.7189]$$

Last, exponentiate the results to get the 95% CI for the OR:

$$[\exp(0.4951), \exp(0.7189)] = [1.64, 5.23]$$

To summarize, for a causal contrast of 10kg weight gain, the estimated odds ratio for sleep apnea is 1.83, and the 95% confidence limit ratio is 5.23/1.64 = 3.2. Assuming that a straight line indeed captures the true dose-response function. That line is depicted in Figure 12–3.

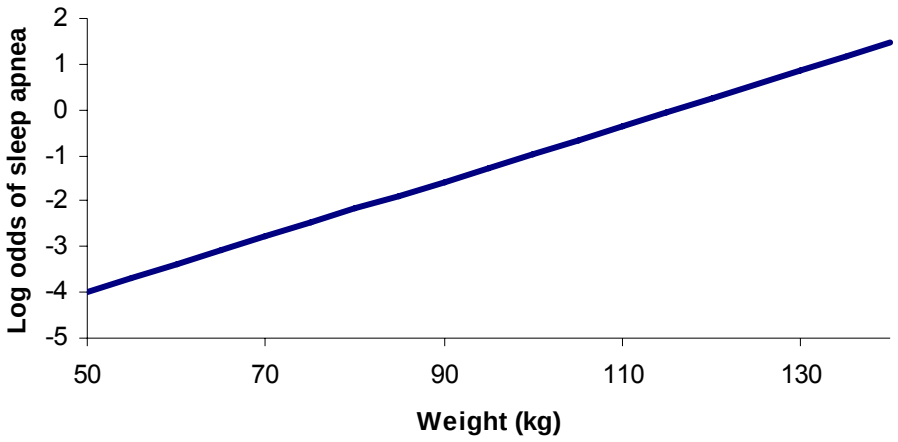

Figure 12–3. Log-linear dose-response function for the effect of weight on sleep apnea

Are we willing to make the assumption of a constant odds ratio per 1kg weight gain or should we continue to explore the dose-response function? The step function (Figure 12–2) does not unequivocally support a straight line. At a minimum, we might examine a quadratic function as well.

SAS code

```
PROC LOGISTIC DESCENDING;
 MODEL apnea = weight weight*weight;
run;
```

Selected SAS printout

```
                    The LOGISTIC Procedure

                     Model Information

Data Set                          WORK.TWO
Response Variable                 apnea
Number of Response Levels         2
Model                             binary logit
```

```
Number of Observations Used        1074

        Response Profile

Ordered                      Total
Value        apnea      Frequency

1            1            166
2            0            908

Probability modeled is apnea=1.

Analysis of Maximum Likelihood Estimates

                                    Standard
Parameter          DF    Estimate      Error

Intercept          1      -8.8478      2.3641
weight             1       0.1007      0.0507
weight*weight      1     -0.00021    0.000268
```

$$\text{log odds (APNEA=1)} = -8.8478 + 0.1007\ \text{WEIGHT} + (-0.00021)\ \text{WEIGHT}^2$$

The regression line, depicted in Figure 12–4, shows some departure from a straight line over the observed range of weight (compare Figure12–4 to Figure 12–3). Therefore, the quadratic function seems to support the curvature suggested by the step function (Figure 12–2). Keep in mind, however, that all three dose-response functions do not account for any confounders. Both the shape and the inference might change after adding confounders to the models.

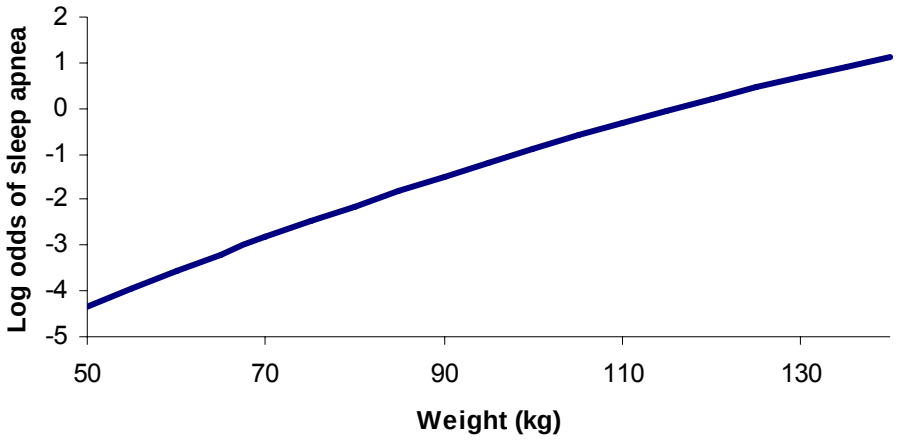

Figure 12–4. Log-quadratic dose-response function for the effect of weight on sleep apnea

How do we compute estimates of the odds ratio from a quadratic dose-response function?

Just like in linear regression with a quadratic term (chapter 9), the coefficients of WEIGHT and $WEIGHT^2$ are not interpretable individually. But if we work on the log-odds scale, we can easily replicate the steps we followed in chapter 9 (Table 9–7.) Table 12–10 shows an example for the causal contrast [50kg, 60kg].

Table 12–10. Computing the effect of a causal contrast between the weights of 60kg and 50kg, based on the quadratic function

Causal assignment	Values of weight variables	log odds (APNEA=1) = $-8.8478 + 0.1007\ WEIGHT - 0.00021\ WEIGHT^2$
WEIGHT = 60	WEIGHT = 60 $WEIGHT^2 = 3600$	log odds (APNEA=1) = $-8.8478 + 0.1007 \times 60 - 0.00021 \times 3600 = -3.5618$
WEIGHT = 50	WEIGHT = 50 $WEIGHT^2 = 2500$	log odds (APNEA=1) = $-8.8478 + 0.1007 \times 50 - 0.00021 \times 2500 = -4.3378$
Effect (difference between two log odds)		log odds ratio: 0.776
		Odd ratio: exp(0.776) = 2.17

Based on the quadratic dose-response function, the estimated odds ratio for this causal contrast is 2.17. The estimate we computed from the linear function was smaller: 1.83 for *any* causal contrast of 10kg difference, including [50kg, 60kg]. Unlike the linear function, the quadratic function prescribes a different odds ratio for each contrast of 10kg difference. For example, the odds ratio for the contrast [58kg, 68kg] differs from the odds ratio for the contrast [80kg, 90kg], and neither is equal to 2.17. Notice, again, that the model claims a special kind of effect modification—weight modifies the effect of weight—but this time we assume effect modification on the multiplicative scale, not the additive scale. The odds ratio for sleep apnea per Δ weight gain is not constant. It depends on the starting weight.

To sum up, logistic regression models are constructed and interpreted along the principles of linear regression—with one important difference: the dependent variable is log odds (Y=1) rather than the mean of Y. As a result, the coefficients are interpreted as log odds ratios and their exponential form turns out to be odds ratios. The analogy carries to effect modification, as well (chapter 13).