34 Gaussian regression models

Most of this book has been about events such as the incidence of disease or mortality. Although events are particularly important in epidemiology, in some studies the response of interest is a quantitative measurement such as blood pressure. The most widely used probability model for such responses is the Gaussian model, described in Chapter 8. In this chapter we show how regression models are used in conjunction with the Gaussian probability model. We shall call this combination *Gaussian regression* although it is more usual for it to be called simply regression or *multiple regression* because it was developed before other regression methods.

34.1 Models for the mean

The Gaussian probability model differs from the binary model in having two parameters instead of one. These are μ, the mean, and σ, the standard deviation. In the simplest situation changing the level of an explanatory variable changes the value of μ but leaves σ unchanged. The distributions of response for a comparison of exposed and unexposed subjects predicted by such a model is illustrated in Fig. 34.1. The effect of exposure is measured by the difference between the means, $\mu_1 - \mu_0$.

To control for confounding by age, using stratification, we would stratify by age and make the assumption that $\mu_1 - \mu_0$ is constant across age groups. This is equivalent to fitting the regression model

$$\text{Mean} = \text{Corner} + \text{Age} + \text{Exposure}.$$

The effect of exposure in this model is simply the (common) difference between mean responses for exposed and unexposed subjects within age groups.

To illustrate such models we shall use some additional data from the study of diet and coronary heart disease. These concern daily intake of fibre which is the response variable. Age and occupation are the explanatory variables, both with three levels.* Table 34.1 shows a simple summary of these data in which a separate estimate of mean and standard deviation has been calculated for each of the nine age–occupation groups. The main interest is in differences between occupations and inspection of the estimated means suggests that there is a systematic tendency for bank clerks to eat more fibre than the drivers and conductors. There is no obvious systematic variation in the standard deviation parameters, so the assumption that changing the levels of age and occupation does not affect σ is reasonable.

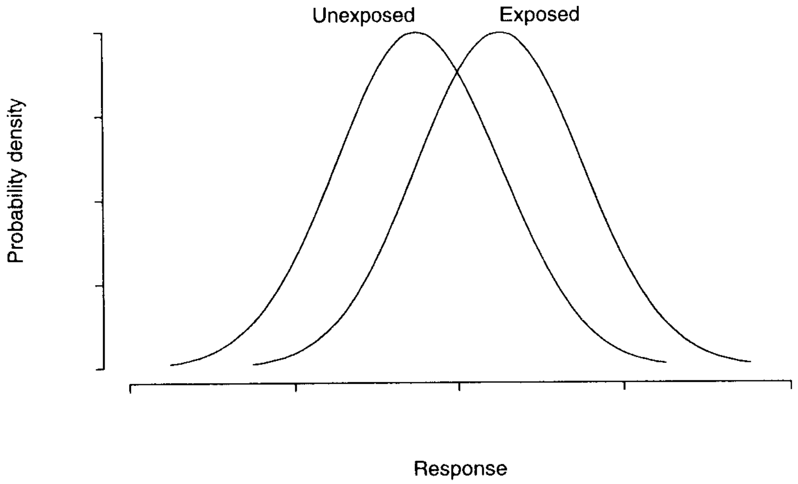

Fig. 34.1. Effect of exposure on the mean response.

The additive regression model relating the mean daily intake of fibre to the effects of age and occupation is

$$\text{Mean} = \text{Corner} + \text{Age} + \text{Work}.$$

When both age and work are treated as categorical this has five parameters in all, namely the Corner, Age(1), Age(2), Work(1), and Work(2) parameters. These are called the *regression parameters* to distinguish them from σ, the common standard deviation, which is called the *residual standard deviation.* The square of σ is called the *residual variance.*

34.2 Likelihood, sums of squares, and deviance

From Chapter 8, the log likelihood for a study of size N is

$$-N\log(\sigma) - \frac{1}{2}\sum_{\text{Subjects}}\left(\frac{x-\mu}{\sigma}\right)^2.$$

*Unpublished data

Table 34.1. Dietary fibre intake (gm/day) by age and occupation

		Occupation		
Age		Bus driver	Bus conductor	Bank clerk
< 45	N	23	16	38
	Mean	16.1	17.2	19.1
	SD	3.91	5.00	5.53
45 – 49	N	30	29	57
	Mean	16.3	17.0	18.5
	SD	4.22	5.42	6.88
50+	N	45	39	56
	Mean	16.6	14.8	17.6
	SD	6.28	4.48	5.43
All	N	98	84	151
	Mean	16.4	16.0	18.34
	SD	5.17	5.00	6.04

However, in contrast with Chapter 8, the mean parameter μ is not a single constant but can vary from subject to subject according to the regression model. In our example μ can take nine different values according to the combination of age and occupation. For estimating the regression parameters the $N\log(\sigma)$ term in the log likelihood can be ignored, and because σ is assumed to be the same for all subjects the parameter values which minimize the sum of squared differences,

$$\sum(x-\mu)^2,$$

will also maximize the log likelihood, regardless of the value of σ. Thus the most likely values of the regression parameters do not depend on σ. Because they minimize a sum of squared differences they are also called *least squares estimates.* The minimum value which this sum of squared differences takes is known as the *residual sum of squares.*

For example, Table 34.2 shows the parameter estimates for the model

$$\text{Mean} = \text{Corner} + \text{Work}$$

for the dietary fibre data. The table shows most likely values for the three parameters in this model, together with their standard deviations. The standard deviation of each regression parameter has been calculated from the profile log likelihood obtained by maximizing the log likelihood with respect to all the other regression parameters. Although the estimated values of these parameters do not depend on σ their standard deviations do, and in constructing the table σ has been taken equal to 5.5401 (we shall see where this value comes from later in the chapter).

Table 34.2. Effects of work on fibre intake (gm/day)

Parameter	Estimate	SD
Corner	16.425	0.560
Work(1)	−0.402	0.824
Work(2)	1.911	0.719

Exercise 34.1. Use the results in Table 34.2 to find the 90% confidence interval for the Work(1) parameter.

34.3 Analysis of deviance

The deviance for any fitted model is defined as minus twice the log likelihood ratio, when this compares the fitted model with a *saturated* model which has a parameter for each record. When the records refer to individual subjects the saturated model has $\mu = x$ so the deviance is

$$\sum\left(\frac{x-\mu}{\sigma}\right)^2.$$

This is proportional to the residual sum of squares for that model.[†] As before, the degrees of freedom for the deviance are equal to the the number of parameters in the saturated regression model, which is equal to the number of subjects N, less the number of parameters in the regression model which has been fitted. These are also the degrees of freedom for the residual sums of squares.

The deviance can be used to compare models in the same way as in Chapter 24, but all calculations are first done in terms of residual sums of squares and later converted to deviances by dividing by a suitable estimate of the square of σ. The residual sums of squares are obtained from the *analysis of variance* table which is usually in the output when a Gaussian regression model is fitted. For example, the analysis of variance table produced when fitting the model

$$\text{Mean} = \text{Corner} + \text{Age} + \text{Work}$$

to the data in Table 34.1 would look something like Table 34.3. The most important line in this table is the middle one labelled 'Error' which gives

[†] In the original definition of the idea of deviance, this was called the *scaled* deviance because of its dependence on the unknown scale parameter σ and the word deviance was reserved for its value when σ is taken as 1. However, this usage has not received widespread acceptance.

Table 34.3. Analysis of variance for the variable work

Source	DF	SSq
Model	2	369.891
Error	330	10128.636
Total	332	10498.527

the residual sum of squares for the model which has been fitted and its degrees of freedom. Since the number of subjects is $N = 333$ and the regression model has three parameters, the degrees of freedom here are $333 - 3 = 330$. The last line of the table, headed 'Total' gives the same information for the degenerate model

$$\text{Mean} = \text{Corner}$$

in which the mean response is the same for all subjects. This regression model has only one parameter so the degrees of freedom for its residual sum of squares and deviance are 332. The line labelled 'Model' is obtained by subtracting the degrees of freedom and the residual sum of squares for the error and total lines. When this difference in residual sum of squares is converted to a difference in deviance by division by the square of a suitable estimate of σ, it provides us with a test of the null hypothesis that all parameters in the model, other than the corner parameter, are zero. In this case this would be a test of the difference between occupations.

With more than one explanatory variable, testing the hypothesis that all the parameters in the model are zero is rarely of any interest. The only use of analysis of variance tables for such models is to obtain the residual sum of squared deviations from the second line. By fitting a series of models a more useful table can be constructed, as follows. Table 34.4 shows the residual sums of squares extracted from the analysis of variance tables for five models fitted to the fibre data. Changes in residual sums of squares from one model to another can be converted to deviances and used to test a variety of hypotheses. For example, the effects of work controlled for age can be tested by using the change in residual sum of squares between models 3 and 4.

ESTIMATING σ

Using the joint likelihood for the regression parameters and σ it can be shown, using calculus, that the most likely value of σ is

$$\sqrt{\frac{\text{Residual sum of squares}}{N}}.$$

Table 34.4. Analysis of deviance ($\sigma = 5.5445$)

Mean = Corner + ···	DF	SSq	Deviance
1. –	332	10498.527	341.510
2. Work	330	10128.636	329.478
3. Age	330	10384.702	337.807
4. Age + Work	328	10048.456	326.870
5. Age + Work + Age·Work	324	9960.268	324.000

This is the value of σ which maximizes the total likelihood and it therefore also maximizes the profile likelihood for σ. When the number of regression parameters is large compared with the number of subjects, it is preferable to use a conditional likelihood which depends only on σ, rather than the profile likelihood. The most likely value of σ is then equal to the residual sum of squares divided by its degrees of freedom. For example, the value of σ used throughout Table 34.4 was

$$\sigma = \sqrt{9960.268/324} = 5.5445$$

which is the conditional estimate obtained from model 5, although the overall most likely value is

$$\sigma = \sqrt{9960.268/333} = 5.4691$$

It can be seen that the use of the degrees of freedom in place of N has a negligible effect for a study of this size. The reason why σ is generally estimated from the conditional likelihood can be illustrated by a simple argument. If we imagine a study of 10 subjects and fit a regression model with 10 parameters it will fit the observations exactly. The overall most likely value of σ would be zero but the reality is that we have no data for estimating σ. Only when we add an eleventh subject to our study do we start collecting information about σ. It follows that the *effective* size of the study for the purposes of estimating σ is given by the N minus the number of regression parameters — the degrees of freedom — and the estimated value of σ should be

$$\sqrt{\frac{\text{Residual sum of squares}}{\text{Degrees of freedom}}}.$$

One consequence of using this estimate is that the deviance for the model used to estimate σ is equal to its degrees of freedom.

A test for interaction between work and age may be obtained by comparing the deviances for models 4 and 5. The difference in deviance is $326.870 - 324.000 = 2.870$ with $326 - 324 = 2$ degrees of freedom. Referring this to the chi-squared distribution shows this to be clearly non-

Table 34.5. Effects of age and work on fibre intake (gm/day)

Parameter	Estimate	SD
Corner	16.430	0.560
Age(1)	−0.223	0.814
Age(2)	−1.118	0.788
Work(1)	−0.387	0.824
Work(2)	1.828	0.720

significant so that we are reassured concerning our assumption of constant occupational effects over age groups.

The parameter estimates for model 4 are shown in Table 34.5. Note, however, that the value of σ used to calculate the standard deviations of the parameters is slightly different from that used in Table 34.4. This is because, whereas the estimate of σ used in Table 34.4 was obtained from model 5, Table 34.5 refers to model 4 and it is therefore logical to estimate σ using this model, that is by

$$\sigma = \sqrt{10048.456/328} = 5.5349.$$

The significance of the occupational effect, controlled for age, can be tested by comparing the deviances for models 4 and 3. However, since this test only makes sense when there is no interaction, deviances should properly be calculated using the model 4 estimate of σ rather than that used in Table 34.4.

Exercise 34.2. Carry out the test for the effect of occupation controlled for age.

Similarly, the value of σ used to calculate standard deviations of parameter estimates in Table 34.2 is obtained from model 2,

$$\sigma = \sqrt{10128.636/330} = 5.5401$$

and this is the value which would be used if we wished to compare models 1 and 2. In practice the difference between the possible estimates of σ are usually inconsequential except in very small studies.

F RATIO TESTS

The tests discussed above refer changes in deviance to the appropriate chi-squared distribution. If the value of σ were a known constant, these would be *exact tests.* However, when σ is estimated they are only approximate. Exact tests which take account of the fact that σ is estimated may be carried out using *F distributions,* tables of which are readily available. Instead of referring the change in deviance to the chi-square distribution, we divide it by the corresponding degrees of freedom to obtain the *F ratio.* For example, the change in deviance for the test for interaction was 2.870, with two degrees of freedom, so the corresponding F ratio is 1.435. To obtain the exact p-value, the F ratio is referred to the correct F distribution. However, to select the correct F distribution, we must specify two different numbers of degrees of freedom. The first, called the *numerator* degrees of freedom, is the same as the degrees of freedom for the approximate chi-squared test while the second, called the *denominator* degrees of freedom, is the number of degrees of freedom used to estimate σ. In our example these are 2 and 334 respectively.

In practice there is only a noticeable difference between F ratio tests and the approximate chi-squared test in small studies. In our example, the p-value obtained from the chi-squared distribution is 0.2381 while that obtained from the F distribution is 0.2396. Since the F ratio test is only exact if the assumptions of Gaussian distribution shape and constancy of σ are true, they are not usually worth the (admittedly slight) extra trouble.

34.4 Multiplicative models ★

A basic assumption in the Gaussian regression model is that changes in the explanatory variables affect the mean level of response but not the variability. However, it is commonly the case that as the level of response goes up, so does its variability. A simple multiplicative model acting at the individual level would explain this, for if the effect of changing the level of work is to double the values of the individual responses, then the standard deviation of these individual values will also get doubled. On a log scale, however, the effect of doubling the response will be to add log(2) to the log response, leaving the standard deviation of the log responses unchanged. This suggests that when the effects appear to act multiplicatively at an individual level, the log response should be analysed in place of the response.

There is some suggestion in Table 34.1 that standard deviation of fibre intake goes up with the mean, so that a multiplicative model may be more appropriate. This suggests analysing log fibre intakes rather than fibre intakes themselves. Inspection of the data suggests that the distribution of log fibre intake is closer to the Gaussian shape than the distribution of fibre intake, and this is another point in favour of analysing log fibre intakes. When the Gaussian regression model

$$\text{Mean} = \text{Corner} + \text{Age} + \text{Work}.$$

is fitted to the logs of the fibre intakes we obtain the parameter estimates shown in Table 34.6.

The effect parameters shown in this table are additive effects upon log fibre intake and these should be exponentiated to express them as multi-

Table 34.6. Effects of age and work on log fibre intake

Parameter	Estimate	SD
Corner	2.8039	0.0430
Age(1)	−0.0253	0.0445
Age(2)	−0.0800	0.0431
Work(1)	−0.0345	0.0451
Work(2)	0.0962	0.0394

plicative effects on fibre intake. The error factor method can be used to calculate confidence intervals for the multiplicative effects.

Exercise 34.3. Express the estimates of the Work parameters as multiplicative effects, and calculate 90% confidence intervals.

Apart from this change in the way the parameter estimates are interpreted the use of the log response in place of the response does not affect matters. Models are compared using residual sums of squares in the same way as before.

If the effect of the explanatory variables is multiplicative at a group level, but not at an individual level, so that σ is constant, a multiplicative model such as

$$\text{Mean} = \text{Corner} \times \text{Age} \times \text{Work},$$

can be fitted to the data on the original scale. Computer programs are available for fitting such models but the need for them rarely arises because the idea of an explanatory variable acting multiplicatively at a group level but not at an individual level is rather implausible.

Solutions to the exercises

34.1 The 90% confidence interval is from $-0.402-1.645\times0.824 = -1.757$ to $-0.402 + 1.645 \times 0.824 = 0.953$. The lower limit is a reduction of 1.757 gm, the upper limit is an increase of 0.953 gm.

34.2 The appropriate value for σ is 5.5349, taken from the model which includes both age and work. The deviance for this model is then 328.000, and the deviance for the model which includes age alone is

$$10384.702/5.5349^2 = 338.982.$$

The change in deviances is $338.982 - 328.000 = 10.982$ on 2 degrees of freedom, for which $p = 0.004$ (from the chi-squared distribution on two degrees of freedom.

34.3 The Work(1) parameter is estimated as −0.0345, and since

$$\exp(-0.0345) = 0.966,$$

the fibre intakes of conductors are 0.966 times those of drivers. The 90% confidence interval for this ratio is found from the error factor

$$\exp(1.645 \times 0.0451) = 1.077,$$

to be from $0.966/1.077 = 0.897$ to $0.966 \times 1.077 = 1.04$. Similarly, the multiplicative effect of Work(2) is 1.101 with 90% confidence interval from 1.032 to 1.175.