Chapter 2 Linear Continuous Models.pdf

Chapter 3 Logit Models for Binary Data.pdf

Chapter 4 Poisson Models for Count Data.pdf

Chapter 4 Supplementary.pdf

Chapter 5 Log Linear Models for Contingency Tabels.pdf

Chapter 6 Multinominal Response Models.pdf

Chapter 7 Survival Models.pdf

Chapter 7.1 Parametric Survival.pdf

Chapter 7.2 Non Parametric Survival.pdf

Chapter 7.3 Cumulative Incidence.pdf

Chapter 7.4 Unobserved Hetergeneity.pdf

Chapter 7.5 Multivariate Survival.pdf

Chapter 7.6 Competing Risks.pdf

Chapter 8 Longitudinal and Clustered Data Models.pdf

Chapter 9 Non Parametric Regression.pdf

Generalized Linear Models.pdf

Review of Likelihood Theory.pdf

The Bootstrap.pdf

Chapter 2

Linear Models for Continuous Data

The starting point in our exploration of statistical models in social research will be the classical linear model. Stops along the way include multiple linear regression, analysis of variance, and analysis of covariance. We will also discuss regression diagnostics and remedies.

2.1 Introduction to Linear Models

Linear models are used to study how a quantitative variable depends on one or more predictors or explanatory variables. The predictors themselves may be quantitative or qualitative.

2.1.1 The Program Effort Data

We will illustrate the use of linear models for continuous data using a small dataset extracted from Mauldin and Berelson (1978) and reproduced in Table 2.1. The data include an index of social setting, an index of family planning effort, and the percent decline in the crude birth rate (CBR)—the number of births per thousand population—between 1965 and 1975, for 20 countries in Latin America and the Caribbean.

The index of social setting combines seven social indicators, namely literacy, school enrollment, life expectancy, infant mortality, percent of males aged 15–64 in the non-agricultural labor force, gross national product *per capita* and percent of population living in urban areas. Higher scores represent higher socio-economic levels.

TABLE 2.1: The Program Effort Data

	Setting	Effort	CBR Decline
Bolivia	46	0	1
Brazil	74	0	10
Chile	89	16	29
Colombia	77	16	25
CostaRica	84	21	29
Cuba	89	15	40
Dominican Rep	68	14	21
Ecuador	70	6	0
El Salvador	60	13	13
Guatemala	55	9	4
Haiti	35	3	0
Honduras	51	7	7
Jamaica	87	23	21
Mexico	83	4	9
Nicaragua	68	0	7
Panama	84	19	22
Paraguay	74	3	6
Peru	73	0	2
Trinidad-Tobago	84	15	29
Venezuela	91	7	11

The index of family planning effort combines 15 different program indicators, including such aspects as the existence of an official family planning policy, the availability of contraceptive methods, and the structure of the family planning program. An index of 0 denotes the absence of a program, 1–9 indicates weak programs, 10–19 represents moderate efforts and 20 or more denotes fairly strong programs.

Figure 2.1 shows scatterplots for all pairs of variables. Note that CBR decline is positively associated with both social setting and family planning effort. Note also that countries with higher socio-economic levels tend to have stronger family planning programs.

In our analysis of these data we will treat the percent decline in the CBR as a continuous response and the indices of social setting and family planning effort as predictors. In a first approach to the data we will treat the predictors as continuous covariates with linear effects. Later we will group

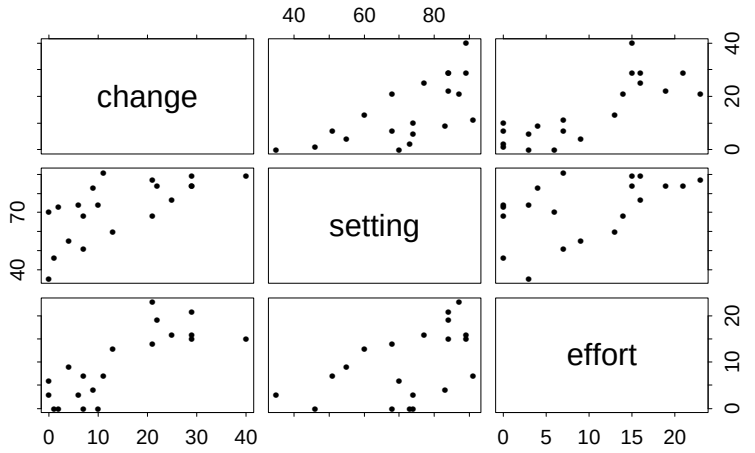


FIGURE 2.1: Scattergrams for the Program Effort Data

them into categories and treat them as discrete factors.

2.1.2 The Random Structure

The first issue we must deal with is that the response will vary even among units with identical values of the covariates. To model this fact we will treat each response y_i as a realization of a random variable Y_i. Conceptually, we view the observed response as only one out of many possible outcomes that we could have observed under identical circumstances, and we describe the possible values in terms of a probability distribution.

For the models in this chapter we will assume that the random variable Y_i has a normal distribution with mean μ_i and variance σ^2, in symbols:

$$Y_i \sim N(\mu_i, \sigma^2).$$

The mean μ_i represents the expected outcome, and the variance σ^2 measures the extent to which an actual observation may deviate from expectation.

Note that the expected value may vary from unit to unit, but the variance is the same for all. In terms of our example, we may expect a larger fertility decline in Cuba than in Haiti, but we don't anticipate that our expectation will be closer to the truth for one country than for the other.

The normal or *Gaussian* distribution (after the mathematician Karl Gauss) has probability density function

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2}\frac{(y_i - \mu_i)^2}{\sigma^2}\}. \tag{2.1}$$

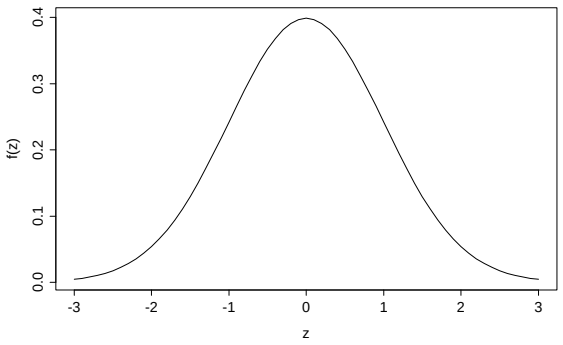


FIGURE 2.2: The Standard Normal Density

The standard density with mean zero and standard deviation one is shown in Figure 2.2.

Most of the probability mass in the normal distribution (in fact, 99.7%) lies within three standard deviations of the mean. In terms of our example, we would be very surprised if fertility in a country declined 3σ more than expected. Of course, we don't know yet what to expect, nor what σ is.

So far we have considered the distribution of one observation. At this point we add the important assumption that the observations are mutually *independent*. This assumption allows us to obtain the joint distribution of the data as a simple product of the individual probability distributions, and underlies the construction of the likelihood function that will be used for estimation and testing. When the observations are independent they are also uncorrelated and their covariance is zero, so $\text{cov}(Y_i, Y_j) = 0$ for $i \neq j$.

It will be convenient to collect the n responses in a column vector $\mathbf{y}$, which we view as a realization of a random vector $\mathbf{Y}$ with mean $\text{E}(\mathbf{Y}) = \boldsymbol{\mu}$ and variance-covariance matrix $\text{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The diagonal elements of $\text{var}(\mathbf{Y})$ are all σ^2 and the off-diagonal elements are all zero, so the n observations are uncorrelated and have the same variance. Under the assumption of normality, $\mathbf{Y}$ has a multivariate normal distribution

$$\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I}) \tag{2.2}$$

with the stated mean and variance.

2.1.3 The Systematic Structure

Let us now turn our attention to the systematic part of the model. Suppose that we have data on p predictors $x_1, \ldots, x_p$ which take values $x_{i1}, \ldots, x_{ip}$

for the i-th unit. We will assume that the expected response depends on these predictors. Specifically, we will assume that μ_i is a *linear* function of the predictors

$$\mu_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

for some unknown coefficients $\beta_1, \beta_2, \ldots, \beta_p$. The coefficients β_j are called *regression coefficients* and we will devote considerable attention to their interpretation.

This equation may be written more compactly using matrix notation as

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta}, \tag{2.3}$$

where $\mathbf{x}_i'$ is a row vector with the values of the p predictors for the i-th unit and $\boldsymbol{\beta}$ is a column vector containing the p regression coefficients. Even more compactly, we may form a column vector $\boldsymbol{\mu}$ with all the expected responses and then write

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \tag{2.4}$$

where $\mathbf{X}$ is an $n \times p$ matrix containing the values of the p predictors for the n units. The matrix $\mathbf{X}$ is usually called the *model* or design matrix. Matrix notation is not only more compact but, once you get used to it, it is also easier to read than formulas with lots of subscripts.

The expression $\mathbf{X}\boldsymbol{\beta}$ is called the *linear predictor*, and includes many special cases of interest. Later in this chapter we will show how it includes simple and multiple linear regression models, analysis of variance models and analysis of covariance models.

The simplest possible linear model assumes that every unit has the same expected value, so that $\mu_i = \mu$ for all i. This model is often called the *null* model, because it postulates no systematic differences between the units. The null model can be obtained as a special case of Equation 2.3 by setting $p = 1$ and $x_i = 1$ for all i. In terms of our example, this model would expect fertility to decline by the same amount in all countries, and would attribute all observed differences between countries to random variation.

At the other extreme we have a model where every unit has its own expected value μ_i. This model is called the *saturated* model because it has as many parameters in the linear predictor (or linear parameters, for short) as it has observations. The saturated model can be obtained as a special case of Equation 2.3 by setting $p = n$ and letting x_i take the value 1 for unit i and 0 otherwise. In this model the x's are indicator variables for the different units, and there is no random variation left. All observed differences between countries are attributed to their own idiosyncrasies.

Obviously the null and saturated models are not very useful by themselves. Most statistical models of interest lie somewhere in between, and most of this chapter will be devoted to an exploration of the middle ground. Our aim is to capture systematic sources of variation in the linear predictor, and let the error term account for unstructured or random variation.

2.2 Estimation of the Parameters

Consider for now a rather abstract model where $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ for some predictors $\mathbf{x}_i$. How do we estimate the parameters $\boldsymbol{\beta}$ and σ^2?

2.2.1 Estimation of β

The likelihood principle instructs us to pick the values of the parameters that maximize the likelihood, or equivalently, the logarithm of the likelihood function. If the observations are independent, then the likelihood function is a product of normal densities of the form given in Equation 2.1. Taking logarithms we obtain the normal log-likelihood

$$\log L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum(y_i - \mu_i)^2/\sigma^2, \tag{2.5}$$

where $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$. The most important thing to notice about this expression is that maximizing the log-likelihood with respect to the linear parameters $\boldsymbol{\beta}$ for a fixed value of σ^2 is exactly equivalent to minimizing the sum of squared differences between observed and expected values, or residual sum of squares

$$\text{RSS}(\boldsymbol{\beta}) = \sum(y_i - \mu_i)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.6}$$

In other words, we need to pick values of $\boldsymbol{\beta}$ that make the fitted values $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ as close as possible to the observed values y_i.

Taking derivatives of the residual sum of squares with respect to $\boldsymbol{\beta}$ and setting the derivative equal to zero leads to the so-called *normal equations* for the maximum-likelihood estimator $\hat{\boldsymbol{\beta}}$

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}.$$

If the model matrix $\mathbf{X}$ is of full column rank, so that no column is an exact linear combination of the others, then the matrix of cross-products $\mathbf{X}'\mathbf{X}$ is of full rank and can be inverted to solve the normal equations. This gives an explicit formula for the *ordinary least squares* (OLS) or maximum likelihood estimator of the linear parameters:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{2.7}$$

If $\mathbf{X}$ is not of full column rank one can use generalized inverses, but interpretation of the results is much more straightforward if one simply eliminates redundant columns. Most current statistical packages are smart enough to detect and omit redundancies automatically.

There are several numerical methods for solving the normal equations, including methods that operate on $\mathbf{X}'\mathbf{X}$, such as Gaussian elimination or the Choleski decomposition, and methods that attempt to simplify the calculations by factoring the model matrix $\mathbf{X}$, including Householder reflections, Givens rotations and the Gram-Schmidt orthogonalization. We will not discuss these methods here, assuming that you will trust the calculations to a reliable statistical package. For further details see McCullagh and Nelder (1989, Section 3.8) and the references therein.

The foregoing results were obtained by maximizing the log-likelihood with respect to $\boldsymbol{\beta}$ for a fixed value of σ^2. The result obtained in Equation 2.7 does not depend on σ^2, and is therefore a global maximum.

For the *null* model $\mathbf{X}$ is a vector of ones, $\mathbf{X}'\mathbf{X} = n$ and $\mathbf{X}'\mathbf{y} = \sum \mathbf{y}_i$ are scalars and $\hat{\beta} = \bar{y}$, the sample mean. For our sample data $\bar{y} = 14.3$. Thus, the calculation of a sample mean can be viewed as the simplest case of maximum likelihood estimation in a linear model.

2.2.2 Properties of the Estimator

The least squares estimator $\hat{\boldsymbol{\beta}}$ of Equation 2.7 has several interesting properties. If the model is correct, in the (weak) sense that the expected value of the response Y_i given the predictors $\mathbf{x}_i$ is indeed $\mathbf{x}_i'\boldsymbol{\beta}$, then the OLS estimator is *unbiased*, its expected value equals the true parameter value:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}. \tag{2.8}$$

It can also be shown that if the observations are uncorrelated and have constant variance σ^2, then the variance-covariance matrix of the OLS estimator is

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2. \tag{2.9}$$

This result follows immediately from the fact that $\hat{\boldsymbol{\beta}}$ is a linear function of the data $\mathbf{y}$ (see Equation 2.7), and the assumption that the variance-covariance matrix of the data is $\mathrm{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

A further property of the estimator is that it has minimum variance among all unbiased estimators that are linear functions of the data, i.e.

it is the best linear unbiased estimator (BLUE). Since no other unbiased estimator can have lower variance for a fixed sample size, we say that OLS estimators are fully *efficient*.

Finally, it can be shown that the sampling distribution of the OLS estimator $\hat{\boldsymbol{\beta}}$ in large samples is approximately multivariate normal with the mean and variance given above, i.e.

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2).$$

Applying these results to the *null* model we see that the sample mean $\bar{y}$ is an unbiased estimator of μ, has variance σ^2/n, and is approximately normally distributed in large samples.

All of these results depend only on second-order assumptions concerning the mean, variance and covariance of the observations, namely the assumption that $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$.

Of course, $\hat{\boldsymbol{\beta}}$ is also a maximum likelihood estimator under the assumption of normality of the observations. If $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ then the sampling distribution of $\hat{\boldsymbol{\beta}}$ is *exactly* multivariate normal with the indicated mean and variance.

The significance of these results cannot be overstated: the assumption of normality of the observations is required only for inference in small samples. The really important assumption is that the observations are uncorrelated and have constant variance, and this is sufficient for inference in large samples.

2.2.3 Estimation of σ^2

Substituting the OLS estimator of $\boldsymbol{\beta}$ into the log-likelihood in Equation 2.5 gives a profile likelihood for σ^2

$$\log L(\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\text{RSS}(\hat{\boldsymbol{\beta}})/\sigma^2.$$

Differentiating this expression with respect to σ^2 (not σ) and setting the derivative to zero leads to the maximum likelihood estimator

$$\hat{\sigma^2} = \text{RSS}(\hat{\boldsymbol{\beta}})/n.$$

This estimator happens to be *biased*, but the bias is easily corrected dividing by $n-p$ instead of n. The situation is exactly analogous to the use of $n-1$ instead of n when estimating a variance. In fact, the estimator of σ^2 for

the *null* model is the sample variance, since $\hat{\beta} = \bar{y}$ and the residual sum of squares is RSS $= \sum(y_i - \bar{y})^2$.

Under the assumption of normality, the ratio RSS$/\sigma^2$ of the residual sum of squares to the true parameter value has a chi-squared distribution with $n - p$ degrees of freedom and is independent of the estimator of the linear parameters. You might be interested to know that using the chi-squared distribution as a likelihood to estimate σ^2 (instead of the normal likelihood to estimate both $\boldsymbol{\beta}$ and σ^2) leads to the unbiased estimator.

For the sample data the RSS for the null model is 2650.2 on 19 d.f. and therefore $\hat{\sigma} = 11.81$, the sample standard deviation.

2.3 Tests of Hypotheses

Consider testing hypotheses about the regression coefficients $\boldsymbol{\beta}$. Sometimes we will be interested in testing the significance of a single coefficient, say β_j, but on other occasions we will want to test the joint significance of several components of $\boldsymbol{\beta}$. In the next few sections we consider tests based on the sampling distribution of the maximum likelihood estimator and likelihood ratio tests.

2.3.1 Wald Tests

Consider first testing the significance of one particular coefficient, say

$$H_0 : \beta_j = 0.$$

The m.l.e. $\hat{\beta}_j$ has a distribution with mean 0 (under H_0) and variance given by the j-th diagonal element of the matrix in Equation 2.9. Thus, we can base our test on the ratio

$$t = \frac{\hat{\beta}_j}{\sqrt{\text{var}(\hat{\beta}_j)}}. \tag{2.10}$$

Note from Equation 2.9 that var$(\hat{\beta}_j)$ depends on σ^2, which is usually unknown. In practice we replace σ^2 by the unbiased estimate based on the residual sum of squares.

Under the assumption of normality of the data, the ratio of the coefficient to its standard error has under H_0 a *Student's t* distribution with $n - p$ degrees of freedom when σ^2 is estimated, and a standard normal distribution if σ^2 is known. This result provides a basis for exact inference in samples of any size.

Under the weaker second-order assumptions concerning the means, variances and covariances of the observations, the ratio has approximately in large samples a standard normal distribution. This result provides a basis for approximate inference in large samples.

Many analysts treat the ratio as a Student's t statistic regardless of the sample size. If normality is suspect one should not conduct the test unless the sample is large, in which case it really makes no difference which distribution is used. If the sample size is moderate, using the t test provides a more conservative procedure. (The Student's t distribution converges to a standard normal as the degrees of freedom increases to ∞. For example the 95% two-tailed critical value is 2.09 for 20 d.f., and 1.98 for 100 d.f., compared to the normal critical value of 1.96.)

The t test can also be used to construct a confidence interval for a coefficient. Specifically, we can state with $100(1-\alpha)\%$ confidence that β_j is between the bounds

$$\hat{\beta}_j \pm t_{1-\alpha/2,n-p}\sqrt{\text{var}(\hat{\beta}_j)}, \tag{2.11}$$

where $t_{1-\alpha/2,n-p}$ is the two-sided critical value of Student's t distribution with $n-p$ d.f. for a test of size α.

The Wald test can also be used to test the joint significance of several coefficients. Let us partition the vector of coefficients into two components, say $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ with p_1 and p_2 elements, respectively, and consider the hypothesis

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}.$$

In this case the Wald statistic is given by the quadratic form

$$W = \hat{\boldsymbol{\beta}}_2' \,\text{var}^{-1}(\hat{\boldsymbol{\beta}}_2)\, \hat{\boldsymbol{\beta}}_2,$$

where $\hat{\boldsymbol{\beta}}_2$ is the m.l.e. of $\boldsymbol{\beta}_2$ and $\text{var}(\hat{\boldsymbol{\beta}}_2)$ is its variance-covariance matrix. Note that the variance depends on σ^2 which is usually unknown; in practice we substitute the estimate based on the residual sum of squares.

In the case of a single coefficient $p_2 = 1$ and this formula reduces to the square of the t statistic in Equation 2.10.

Asymptotic theory tells us that under H_0 the large-sample distribution of the m.l.e. is multivariate normal with mean vector $\mathbf{0}$ and variance-covariance matrix $\text{var}(\boldsymbol{\beta}_2)$. Consequently, the large-sample distribution of the quadratic form W is chi-squared with p_2 degrees of freedom. This result holds whether σ^2 is known or estimated.

Under the assumption of normality we have a stronger result. The distribution of W is exactly chi-squared with p_2 degrees of freedom if σ^2 is

known. In the more general case where σ^2 is estimated using a residual sum of squares based on $n-p$ d.f., the distribution of W/p_2 is an F with p_2 and $n-p$ d.f.

Note that as n approaches infinity for fixed p (so $n-p$ approaches infinity), the F distribution times p_2 approaches a chi-squared distribution with p_2 degrees of freedom. Thus, in large samples it makes no difference whether one treats W as chi-squared or W/p_2 as an F statistic. Many analysts treat W/p_2 as F for all sample sizes.

The situation is exactly analogous to the choice between the normal and Student's t distributions in the case of one variable. In fact, a chi-squared with one degree of freedom is the square of a standard normal, and an F with one and v degrees of freedom is the square of a Student's t with v degrees of freedom.

2.3.2 The Likelihood Ratio Test

Consider again testing the joint significance of several coefficients, say

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$$

as in the previous subsection. Note that we can partition the model matrix into two components $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ with p_1 and p_2 predictors, respectively. The hypothesis of interest states that the response does not depend on the last p_2 predictors.

We now build a likelihood ratio test for this hypothesis. The general theory directs us to (1) fit two nested models: a smaller model with the first p_1 predictors in $\mathbf{X}_1$, and a larger model with all p predictors in $\mathbf{X}$; and (2) compare their maximized likelihoods (or log-likelihoods).

Suppose then that we fit the smaller model with the predictors in $\mathbf{X}_1$ only. We proceed by maximizing the log-likelihood of Equation 2.5 for a fixed value of σ^2. The maximized log-likelihood is

$$\max \log \mathrm{L}(\boldsymbol{\beta}_1) = c - \frac{1}{2}\mathrm{RSS}(\mathbf{X}_1)/\sigma^2,$$

where $c = -(n/2)\log(2\pi\sigma^2)$ is a constant depending on π and σ^2 but not on the parameters of interest. In a slight abuse of notation, we have written $\mathrm{RSS}(\mathbf{X}_1)$ for the residual sum of squares after fitting $\mathbf{X}_1$, which is of course a function of the estimate $\hat{\boldsymbol{\beta}}_1$.

Consider now fitting the larger model $X_1 + X_2$ with all predictors. The maximized log-likelihood for a fixed value of σ^2 is

$$\max \log \mathrm{L}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = c - \frac{1}{2}\mathrm{RSS}(\mathbf{X}_1 + \mathbf{X}_2)/\sigma^2,$$

where $\text{RSS}(\mathbf{X}_1 + \mathbf{X}_2)$ is the residual sum of squares after fitting $\mathbf{X}_1$ and $\mathbf{X}_2$, itself a function of the estimate $\hat{\boldsymbol{\beta}}$.

To compare these log-likelihoods we calculate minus twice their difference. The constants cancel out and we obtain the likelihood ratio criterion

$$-2\log\lambda = \frac{\text{RSS}(\mathbf{X}_1) - \text{RSS}(\mathbf{X}_1 + \mathbf{X}_2)}{\sigma^2}. \tag{2.12}$$

There are two things to note about this criterion. First, we are directed to look at the reduction in the residual sum of squares when we add the predictors in $\mathbf{X}_2$. Basically, these variables are deemed to have a significant effect on the response if including them in the model results in a reduction in the residual sum of squares. Second, the reduction is compared to σ^2, the error variance, which provides a unit of comparison.

To determine if the reduction (in units of σ^2) exceeds what could be expected by chance alone, we compare the criterion to its sampling distribution. Large sample theory tells us that the distribution of the criterion converges to a chi-squared with p_2 d.f. The expected value of a chi-squared distribution with ν degrees of freedom is ν (and the variance is 2ν). Thus, chance alone would lead us to expect a reduction in the RSS of about one σ^2 for each variable added to the model. To conclude that the reduction exceeds what would be expected by chance alone, we usually require an improvement that exceeds the 95-th percentile of the reference distribution.

One slight difficulty with the development so far is that the criterion depends on σ^2, which is not known. In practice, we substitute an estimate of σ^2 based on the residual sum of squares of the *larger* model. Thus, we calculate the criterion in Equation 2.12 using

$$\hat{\sigma^2} = \text{RSS}(\mathbf{X}_1 + \mathbf{X}_2)/(n-p).$$

The large-sample distribution of the criterion continues to be chi-squared with p_2 degrees of freedom, even if σ^2 has been estimated.

Under the assumption of normality, however, we have a stronger result. The likelihood ratio criterion $-2\log\lambda$ has an *exact* chi-squared distribution with p_2 d.f. if σ^2 is know. In the usual case where σ^2 is estimated, the criterion divided by p_2, namely

$$F = \frac{(\text{RSS}(\mathbf{X}_1) - \text{RSS}(\mathbf{X}_1 + \mathbf{X}_2))/p_2}{\text{RSS}(\mathbf{X}_1 + \mathbf{X}_2)/(n-p)}, \tag{2.13}$$

has an exact F distribution with p_2 and $n-p$ d.f.

The numerator of F is the reduction in the residual sum of squares per degree of freedom spent. The denominator is the average residual sum of

squares, a measure of noise in the model. Thus, an F-ratio of one would indicate that the variables in $\mathbf{X}_2$ are just adding noise. A ratio in excess of one would be indicative of signal. We usually reject H_0, and conclude that the variables in $\mathbf{X}_2$ have an effect on the response if the F criterion exceeds the 95-th percentage point of the F distribution with p_2 and $n - p$ degrees of freedom.

A Technical Note: In this section we have built the likelihood ratio test for the linear parameters $\boldsymbol{\beta}$ by treating σ^2 as a nuisance parameter. In other words, we have maximized the log-likelihood with respect to $\boldsymbol{\beta}$ for fixed values of σ^2. You may feel reassured to know that if we had maximized the log-likelihood with respect to both $\boldsymbol{\beta}$ and σ^2 we would have ended up with an equivalent criterion based on a comparison of the *logarithms* of the residual sums of squares of the two models of interest. The approach adopted here leads more directly to the distributional results of interest and is typical of the treatment of scale parameters in generalized linear models.□

2.3.3 Student's t, F and the Anova Table

You may be wondering at this point whether you should use the Wald test, based on the large-sample distribution of the m.l.e., or the likelihood ratio test, based on a comparison of maximized likelihoods (or log-likelihoods). The answer in general is that in large samples the choice does not matter because the two types of tests are asymptotically equivalent.

In linear models, however, we have a much stronger result: the two tests are *identical*. The proof is beyond the scope of these notes, but we will verify it in the context of specific applications. The result is unique to linear models. When we consider logistic or Poisson regression models later in the sequel we will find that the Wald and likelihood ratio tests differ.

At least for linear models, however, we can offer some simple practical advice:

- To test hypotheses about a single coefficient, use the t-test based on the estimator and its standard error, as given in Equation 2.10.

- To test hypotheses about several coefficients, or more generally to compare nested models, use the F-test based on a comparison of RSS's, as given in Equation 2.13.

The calculations leading to an F-test are often set out in an analysis of variance (anova) table, showing how the total sum of squares (the RSS of the null model) can be partitioned into a sum of squares associated with $\mathbf{X}_1$,

a sum of squares *added by* $\mathbf{X}_2$, and a residual sum of squares. The table also shows the degrees of freedom associated with each sum of squares, and the mean square, or ratio of the sum of squares to its d.f.

Table 2.2 shows the usual format. We use ϕ to denote the null model. We also assume that one of the columns of $\mathbf{X}_1$ was the constant, so this block adds only $p_1 - 1$ variables to the null model.

TABLE 2.2: The Hierarchical Anova Table

Source of variation	Sum of squares	Degrees of freedom
$\mathbf{X}_1$	$\text{RSS}(\phi) - \text{RSS}(\mathbf{X}_1)$	$p_1 - 1$
$\mathbf{X}_2$ given $\mathbf{X}_1$	$\text{RSS}(\mathbf{X}_1) - \text{RSS}(\mathbf{X}_1 + \mathbf{X}_2)$	p_2
Residual	$\text{RSS}(\mathbf{X}_1 + \mathbf{X}_2)$	$n - p$
Total	$\text{RSS}(\phi)$	$n - 1$

Sometimes the component associated with the constant is shown explicitly and the bottom line becomes the total (also called 'uncorrected') sum of squares: $\sum y_i^2$. More detailed analysis of variance tables may be obtained by introducing the predictors one at a time, while keeping track of the reduction in residual sum of squares at each step.

Rather than give specific formulas for these cases, we stress here that *all* anova tables can be obtained by calculating differences in RSS's and differences in the number of parameters between nested models. Many examples will be given in the applications that follow. A few descriptive measures of interest, such as simple, partial and multiple correlation coefficients, turn out to be simple functions of these sums of squares, and will be introduced in the context of the applications.

An important point to note before we leave the subject is that the order in which the variables are entered in the anova table (reflecting the order in which they are added to the model) is extremely important. In Table 2.2, we show the effect of adding the predictors in $\mathbf{X}_2$ to a model that already has $\mathbf{X}_1$. This *net* effect of X_2 after allowing for X_1 can be quite different from the *gross* effect of X_2 when considered by itself. The distinction is important and will be stressed in the context of the applications that follow.

2.4 Simple Linear Regression

Let us now turn to applications, modelling the dependence of a continuous response y on a single linear predictor x. In terms of our example, we will study fertility decline as a function of social setting. One can often obtain useful insight into the form of this dependence by plotting the data, as we did in Figure 2.1.

2.4.1 The Regression Model

We start by recognizing that the response will vary even for constant values of the predictor, and model this fact by treating the responses y_i as realizations of random variables

$$Y_i \sim N(\mu_i, \sigma^2) \tag{2.14}$$

with means μ_i depending on the values of the predictor x_i and constant variance σ^2.

The simplest way to express the dependence of the expected response μ_i on the predictor x_i is to assume that it is a linear function, say

$$\mu_i = \alpha + \beta x_i. \tag{2.15}$$

This equation defines a straight line. The parameter α is called the *constant* or intercept, and represents the expected response when $x_i = 0$. (This quantity may not be of direct interest if zero is not in the range of the data.) The parameter β is called the *slope*, and represents the expected increment in the response per unit change in x_i.

You probably have seen the simple linear regression model written with an explicit error term as

$$Y_i = \alpha + \beta x_i + \epsilon_i.$$

Did I forget the error term? Not really. Equation 2.14 defines the random structure of the model, and is equivalent to saying that $Y_i = \mu_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$. Equation 2.15 defines the systematic structure of the model, stipulating that $\mu_i = \alpha + \beta x_i$. Combining these two statements yields the traditional formulation of the model. Our approach separates more clearly the systematic and random components, and extends more easily to generalized linear models by focusing on the distribution of the response rather than the distribution of the error term.

2.4.2 Estimates and Standard Errors

The simple linear regression model can be obtained as a special case of the general linear model of Section 2.1 by letting the model matrix $\mathbf{X}$ consist of two columns: a column of ones representing the constant and a column with the values of x representing the predictor. Estimates of the parameters, standard errors, and tests of hypotheses can then be obtained from the general results of Sections 2.2 and 2.3.

It may be of interest to note that in simple linear regression the estimates of the constant and slope are given by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \text{and} \qquad \hat{\beta} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}.$$

The first equation shows that the fitted line goes through the means of the predictor and the response, and the second shows that the estimated slope is simply the ratio of the covariance of x and y to the variance of x.

Fitting this model to the family planning effort data with CBR decline as the response and the index of social setting as a predictor gives a residual sum of squares of 1449.1 on 18 d.f. (20 observations minus two parameters: the constant and slope).

Table 2.3 shows the estimates of the parameters, their standard errors and the corresponding t-ratios.

TABLE 2.3: Estimates for Simple Linear Regression of CBR Decline on Social Setting Score

Parameter	Symbol	Estimate	Std.Error	t-ratio
Constant	α	-22.13	9.642	-2.29
Slope	β	0.5052	0.1308	3.86

We find that, on the average, each additional point in the social setting scale is associated with an additional half a percentage point of CBR decline, measured from a baseline of an expected 22% *increase* in CBR when social setting is zero. (Since the social setting scores range from 35 to 91, the constant is not particularly meaningful in this example.)

The estimated standard error of the slope is 0.13, and the corresponding t-test of 3.86 on 18 d.f. is highly significant. With 95% confidence we estimate that the slope lies between 0.23 and 0.78.

Figure 2.3 shows the results in graphical form, plotting observed and fitted values of CBR decline versus social setting. The fitted values are

calculated for any values of the predictor x as $\hat{y} = \hat{\alpha} + \hat{\beta}x$ and lie, of course, in a straight line.

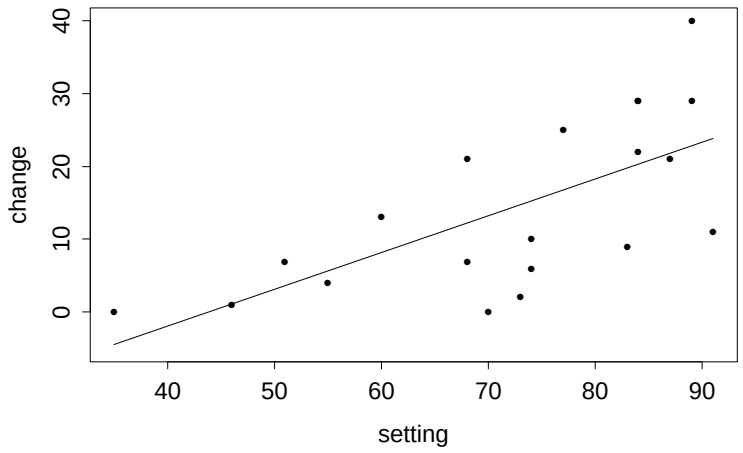


FIGURE 2.3: Linear Regression of CBR Decline on Social Setting

You should verify that the analogous model with family planning effort as a single predictor gives a residual sum of squares of 950.6 on 18 d.f., with constant 2.336 ($\pm$2.662) and slope 1.253 ($\pm$0.2208). Make sure you know how to interpret these estimates.

2.4.3 Anova for Simple Regression

Instead of using a test based on the distribution of the OLS estimator, we could test the significance of the slope by comparing the simple linear regression model with the null model. Note that these models are nested, because we can obtain the null model by setting $\beta = 0$ in the simple linear regression model.

Fitting the null model to the family planning data gives a residual sum of squares of 2650.2 on 19 d.f. Adding a linear effect of social setting reduces the RSS by 1201.1 at the expense of one d.f. This gain can be contrasted with the remaining RSS of 1449.1 on 18 d.f. by constructing an F-test. The calculations are set out in Table 2.4, and lead to an F-statistic of 14.9 on one and 18 d.f.

These results can be used to verify the equivalence of t and F test statistics and critical values. Squaring the observed t-statistic of 3.86 gives the observed F-ratio of 14.9. Squaring the 95% two-sided critical value of the

TABLE 2.4: Analysis of Variance for Simple Regression of CBR Decline on Social Setting Score

Source of variation	Degrees of freedom	Sum of squares	Mean squared	F-ratio
Setting	1	1201.1	1201.1	14.9
Residual	18	1449.1	80.5	
Total	19	2650.2		

Student's t distribution with 18 d.f., which is 2.1, gives the 95% critical value of the F distribution with one and 18 d.f., which is 4.4.

You should verify that the t and F tests for the model with a linear effect of family planning effort are $t = 5.67$ and $F = 32.2$.

2.4.4 Pearson's Correlation Coefficient

A simple summary of the strength of the relationship between the predictor and the response can be obtained by calculating a proportionate reduction in the residual sum of squares as we move from the null model to the model with x. The quantity

$$R^2 = 1 - \frac{\text{RSS}(x)}{\text{RSS}(\phi)}$$

is know as the *coefficient of determination*, and is often described as the proportion of 'variance' explained by the model. (The description is not very accurate because the calculation is based on the RSS not the variance, but it is too well entrenched to attempt to change it.) In our example the RSS was 2650.2 for the null model and 1449.1 for the model with setting, so we have 'explained' 1201.1 points or 45.3% as a linear effect of social setting.

The square root of the proportion of variance explained in a simple linear regression model, with the same sign as the regression coefficient, is *Pearson's linear correlation coefficient*. This measure ranges between -1 and 1, taking these values for perfect inverse and direct relationships, respectively. For the model with CBR decline as a linear function of social setting, Pearson's $r = 0.673$. This coefficient can be calculated directly from the covariance of x and y and their variances, as

$$r = \frac{\sum(y - \bar{y})(x - \bar{x})}{\sqrt{\sum(y - \bar{y})^2 \sum(x - \bar{x})^2}}.$$

There is one additional characterization of Pearson's r that may help in interpretation. Suppose you standardize y by subtracting its mean and dividing by its standard deviation, standardize x in the same fashion, and then regress the standardized y on the standardized x forcing the regression through the origin (i.e. omitting the constant). The resulting estimate of the regression coefficient is Pearson's r. Thus, we can interpret r as the expected change in the response in units of standard deviation associated with a change of one standard deviation in the predictor.

In our example, each standard deviation of increase in social setting is associated with an additional decline in the CBR of 0.673 standard deviations. While the regression coefficient expresses the association in the original units of x and y, Pearson's r expresses the association in units of standard deviation.

You should verify that a linear effect of family planning effort accounts for 64.1% of the variation in CBR decline, so Pearson's $r = 0.801$. Clearly CBR decline is associated more strongly with family planning effort than with social setting.

2.5 Multiple Linear Regression

Let us now study the dependence of a continuous response on two (or more) linear predictors. Returning to our example, we will study fertility decline as a function of both social setting and family planning effort.

2.5.1 The Additive Model

Suppose then that we have a response y and two predictors x_1 and x_2. We will use y_i to denote the value of the response and x_{i1} and x_{i2} to denote the values of the predictors for the i-th unit, where $i = 1, \ldots, n$.

We maintain the assumptions regarding the stochastic component of the model, so y_i is viewed as a realization of $Y_i \sim N(\mu_i, \sigma^2)$, but change the structure of the systematic component. We now assume that the expected response μ_i is a linear function of the two predictors, that is

$$\mu_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2}. \tag{2.16}$$

This equation defines a plane in three dimensional space (you may want to peek at Figure 2.4 for an example). The parameter α is the constant, representing the expected response when both x_{i1} and x_{i2} are zero. (As before, this value may not be directly interpretable if zero is not in the

range of the predictors.) The parameter β_1 is the slope along the x_1-axis and represents the expected change in the response per unit change in x_1 at constant values of x_2. Similarly, β_2 is the slope along the x_2 axis and represents the expected change in the response per unit change in x_2 while holding x_1 constant.

It is important to note that these interpretations represent abstractions based on the model that we may be unable to observe in the real world. In terms of our example, changes in family planning effort are likely to occur in conjunction with, if not directly as a result of, improvements in social setting. The model, however, provides a useful representation of the data and hopefully approximates the results of comparing countries that differ in family planning effort but have similar socio-economic conditions.

A second important feature of the model is that it is *additive*, in the sense that the effect of each predictor on the response is assumed to be the same for all values of the other predictor. In terms of our example, the model assumes that the effect of family planning effort is exactly the same at every social setting. This assumption may be unrealistic, and later in this section we will introduce a model where the effect of family planning effort is allowed to depend on social setting.

2.5.2 Estimates and Standard Errors

The multiple regression model in 2.16 can be obtained as a special case of the general linear model of Section 2.1 by letting the model matrix $\mathbf{X}$ consist of three columns: a column of ones representing the constant, a column representing the values of x_1, and a column representing the values of x_2. Estimates, standard errors and tests of hypotheses then follow from the general results in Sections 2.2 and 2.3.

Fitting the two-predictor model to our example, with CBR decline as the response and the indices of family planning effort and social setting as linear predictors, gives a residual sum of squares of 694.0 on 17 d.f. (20 observations minus three parameters: the constant and two slopes). Table 2.5 shows the parameter estimates, standard errors and t-ratios.

We find that, on average, the CBR declines an additional 0.27 percentage points for each additional point of improvement in social setting at constant levels of family planning effort. The standard error of this coefficient is 0.11. Since the t ratio exceeds 2.11, the five percent critical value of the t distribution with 17 d.f., we conclude that we have evidence of association between social setting and CBR decline net of family planning effort. A 95% confidence interval for the social setting slope, based on Student's t

TABLE 2.5: Estimates for Multiple Linear Regression of CBR Decline on Social Setting and Family Planning Effort Scores

Parameter	Symbol	Estimate	Std.Error	t-ratio
Constant	α	-14.45	7.094	-2.04
Setting	β_1	0.2706	0.1079	2.51
Effort	β_2	0.9677	0.2250	4.30

distribution with 17 d.f., has bounds 0.04 and 0.50.

Similarly, we find that on average the CBR declines an additional 0.97 percentage points for each additional point of family planning effort at constant social setting. The estimated standard error of this coefficient is 0.23. Since the coefficient is more than four times its standard error, we conclude that there is a significant linear association between family planning effort and CBR decline at any given level of social setting. With 95% confidence we conclude that the additional percent decline in the CBR per extra point of family planning effort lies between 0.49 and 1.44.

The constant is of no direct interest in this example because zero is not in the range of the data; while some countries have a value of zero for the index of family planning effort, the index of social setting ranges from 35 for Haiti to 91 for Venezuela.

The estimate of the residual standard deviation in our example is $\hat{\sigma} = 6.389$. This value, which is rarely reported, provides a measure of the extent to which countries with the same setting and level of effort can experience different declines in the CBR.

Figure 2.4 shows the estimated regression equation $\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ evaluated for a grid of values of the two predictors. The grid is confined to the range of the data on setting and effort. The regression plane may be viewed as an infinite set of regression lines. For any fixed value of setting, expected CBR decline is a linear function of effort with slope 0.97. For any fixed value of effort, expected CBR decline is a linear function of setting with slope 0.27.

2.5.3 Gross and Net Effects

It may be instructive to compare the results of the multiple regression analysis, which considered the two predictors simultaneously, with the results of the simple linear regression analyses, which considered the predictors one at a time.

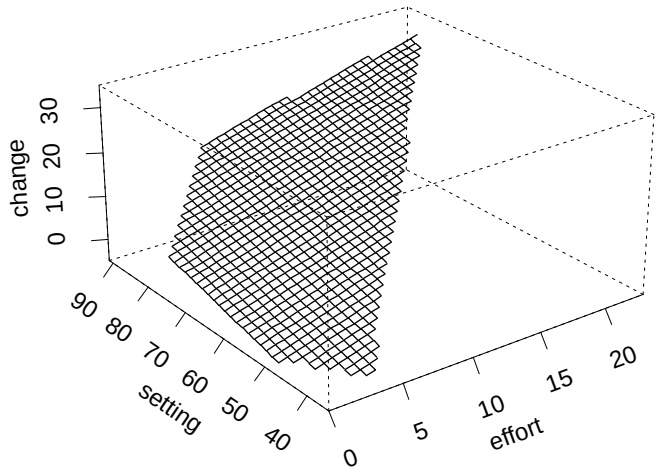


FIGURE 2.4: Multiple Regression of CBR Decline on Social Setting and Family Planning Effort

The coefficients in a simple linear regression represent changes in the response that can be associated with a given predictor, and will be called *gross* effects. In our simple linear regression analysis of CBR decline as a function of family planning effort we found that, on the average, each additional point of family planning effort was associated with an additional 1.25 percentage point of CBR decline. Interpretation of gross effects must be cautious because comparisons involving one factor include, implicitly, other measured and unmeasured factors. In our example, when we compare countries with strong programs with countries with weak programs, we are also comparing implicitly countries with high and low social settings.

The coefficients in a multiple linear regression are more interesting because they represent changes in the response that can be associated with a given predictor for fixed values of other predictors, and will be called *net* effects. In our multiple regression analysis of CBR decline as a function of both family planning effort and social setting, we found that, on the average, each additional point of family planning effort was associated with an additional 0.97 percentage points of CBR decline if we held social setting constant, i.e. if we compared countries with the same social setting. Interpretation of this coefficient as measuring the effect of family planning effort is on somewhat firmer ground than for the gross effect, because the differences have been adjusted for social setting. Caution is in order, however, because there are bound to be other confounding factors that we have not taken into account.

In my view, the closest approximation we have to a true causal effect in social research based on observational data is a net effect in a multiple regression analysis that has controlled for all relevant factors, an ideal that may be approached but probably can never be attained. The alternative is a controlled experiment where units are assigned at random to various treatments, because the nature of the assignment itself guarantees that any ensuing differences, beyond those than can be attributed to chance, must be due to the treatment. In terms of our example, we are unable to randomize the allocation of countries to strong and weak programs. But we can use multiple regression as a tool to adjust the estimated effects for the confounding effects of observed covariates.

TABLE 2.6: Gross and Net Effects of Social Setting and Family Planning Effort on CBR Decline

Predictor	Effect	
	Gross	Net
Setting	0.505	0.271
Effort	1.253	0.968

Gross and net effects may be presented in tabular form as shown in Table 2.6. In our example, the gross effect of family planning effort of 1.25 was reduced to 0.97 after adjustment for social setting, because part of the observed differences between countries with strong and weak programs could be attributed to the fact that the former tend to enjoy higher living standards. Similarly, the gross effect of social setting of 0.51 has been reduced to 0.27 after controlling for family planning effort, because part of the differences between richer and poorer countries could be attributed to the fact that the former tend to have stronger family planning programs.

Note, incidentally, that it is not reasonable to compare either gross or net effects across predictors, because the regression coefficients depend on the units of measurement. I could easily 'increase' the gross effect of family planning effort to 12.5 simply by dividing the scores by ten. One way to circumvent this problem is to standardize the response and all predictors, subtracting their means and dividing by their standard deviations. The regression coefficients for the standardized model (which are sometimes called 'beta' coefficients) are more directly comparable. This solution is particularly appealing when the variables do not have a natural unit of measurement, as is often the case for psychological test scores. On the other hand,

standardized coefficients are heavily dependent on the range of the data; they should not be used, for example, if one has sampled high and low values of one predictor to increase efficiency, because that design would inflate the variance of the predictor and therefore reduce the standardized coefficient.

2.5.4 Anova for Multiple Regression

The basic principles of model comparison outlined earlier may be applied to multiple regression models. I will illustrate the procedures by considering a test for the significance of the entire regression, and a test for the significance of the net effect of one predictor after adjusting for the other.

Consider first the hypothesis that all coefficients other than the constant are zero, i.e.

$$H_0 : \beta_1 = \beta_2 = 0.$$

To test the significance of the entire regression we start with the null model, which had a RSS of 2650.2 on 19 degrees of freedom. Adding the two linear predictors, social setting and family planning effort, reduces the RSS by 1956.2 at the expense of two d.f. Comparing this gain with the remaining RSS of 694.0 on 17 d.f. leads to an F-test of 24.0 on two and 17 d.f. This statistic is highly significant, with a P-value just above 0.00001. Thus, we have clear evidence that CBR decline is associated with social setting and family planning effort. Details of these calculations are shown in Table 2.7

TABLE 2.7: Analysis of Variance for Multiple Regression of CBR Decline by Social Setting and Family Planning Effort

Source of variation	Sum of squares	Degrees of freedom	Mean squared	F-ratio
Regression	1956.2	2	978.1	24.0
Residual	694.0	17	40.8	
Total	2650.2	19		

In the above comparison we proceeded directly from the null model to the model with two predictors. A more detailed analysis is possible by adding the predictors one at a time. Recall from Section 2.4 that the model with social setting alone had a RSS of 1449.1 on 18 d.f., which represents a gain of 1201.1 over the null model. In turn, the multiple regression model with both social setting and family planning effort had a RSS of 694.0 on 17 d.f. which represents a gain of 755.1 over the model with social setting alone. These calculation are set out in the *hierarchical* anova shown in Table 2.8.

TABLE 2.8: Hierarchical Analysis of Variance for Multiple Regression of CBR Decline by Social Setting and Family Planning Effort

Source of variation	Sum of squares	Degrees of freedom	Mean squared	F-ratio
Setting	1201.1	1	1201.1	29.4
Effort|Setting	755.1	1	755.1	18.5
Residual	694.0	17	40.8	
Total	2650.2	19		

Note the following features of this table. First, adding the sums of squares and d.f.'s in the first two rows agrees with the results in the previous table; thus, we have further decomposed the sum of squares associated with the regression into a term attributed to social setting and a term added by family planning effort.

Second, the notation Effort|Setting emphasizes that we have considered first the contribution of setting and then the additional contribution of effort once setting is accounted for. The order we used seemed more natural for the problem at hand. An alternative decomposition would introduce effort first and then social setting. The corresponding hierarchical anova table is left as an exercise.

Third, the F-test for the additional contribution of family planning effort over and above social setting (which is $F = 18.5$ from Table 2.8) coincides with the test for the coefficient of effort based on the estimate and its standard error (which is $t = 4.3$ from Table 2.5), since $4.3^2 = 18.5$. In both cases we are testing the hypothesis

$$H_0 : \beta_2 = 0$$

that the *net* effect of effort given setting is zero. Keep in mind that dividing estimates by standard errors tests the hypothesis that the variable in question has no effect *after* adjusting for all other variables. It is perfectly possible to find that two predictors are jointly significant while neither exceeds twice its standard error. This occurs when the predictors are highly correlated and either could account for (most of) the effects of the other.

2.5.5 Partial and Multiple Correlations

A descriptive measure of how much we have advanced in our understanding of the response is given by the proportion of variance explained, which was

first introduced in Section 2.4. In our case the two predictors have reduced the RSS from 2650.2 to 694.0, explaining 73.8%.

The square root of the proportion of variance explained is the *multiple correlation coefficient*, and measures the linear correlation between the response in one hand and all the predictors on the other. In our case $R = 0.859$. This value can also be calculated directly as Pearson's linear correlation between the response y and the fitted values $\hat{y}$.

An alternative construction of R is of some interest. Suppose we want to measure the correlation between a single variable y and a set of variables (a vector) $\mathbf{x}$. One approach reduces the problem to calculating Pearson's r between two single variables, y and a linear combination $z = \mathbf{c}'\mathbf{x}$ of the variables in $\mathbf{x}$, and then taking the maximum over all possible vectors of coefficients $\mathbf{c}$. Amazingly, the resulting maximum is R and the coefficients $\mathbf{c}$ are proportional to the estimated regression coefficients.

We can also calculate proportions of variance explained based on the hierarchical anova tables. Looking at Table 2.8, we note that setting explained 1201.1 of the total 2650.2, or 45.3%, while effort explained 755.1 of the same 2650.2, or 28.5%, for a total of 1956.2 out of 2650.2, or 73.8%. In a sense this calculation is not fair because setting is introduced before effort. An alternative calculation may focus on how much the second variable explains not out of the total, but out of the variation left unexplained by the first variable. In this light, effort explained 755.1 of the 1449.1 left unexplained by social setting, or 52.1%.

The square root of the proportion of variation explained by the second variable out of the amount left unexplained by the first is called the *partial correlation coefficient*, and measures the linear correlation between y and x_2 after adjusting for x_1. In our example, the linear correlation between CBR decline and effort after controlling for setting is 0.722.

The following calculation may be useful in interpreting this coefficient. First regress y on x_1 and calculate the residuals, or differences between observed and fitted values. Then regress x_2 on x_1 and calculate the residuals. Finally, calculate Pearson's r between the two sets of residuals. The result is the partial correlation coefficient, which can thus be seen to measure the simple linear correlation between y and x_2 after removing the linear effects of x_1.

Partial correlation coefficients involving three variables can be calculated directly from the pairwise simple correlations. Let us index the response y as variable 0 and the predictors x_1 and x_2 as variables 1 and 2. Then the

partial correlation between variables 0 and 2 adjusting for 1 is

$$r_{02.1} = \frac{r_{02} - r_{01}r_{12}}{\sqrt{1 - r_{01}^2}\sqrt{1 - r_{12}^2}},$$

where r_{ij} denotes Pearson's linear correlation between variables i and j. The formulation given above is more general, because it can be used to compute the partial correlation between two variables (the response and one predictor) adjusting for any number of additional variables.

TABLE 2.9: Simple and Partial Correlations of CBR Decline with Social Setting and Family Planning Effort

Predictor	Correlation	
	Simple	Partial
Setting	0.673	0.519
Effort	0.801	0.722

Simple and partial correlation coefficients can be compared in much the same vein as we compared gross and net effects earlier. Table 2.9 summarizes the simple and partial correlations between CBR decline on the one hand and social setting and family planning effort on the other. Note that the effect of effort is more pronounced and more resilient to adjustment than the effect of setting.

2.5.6 More Complicated Models

So far we have considered four models for the family planning effort data: the null model (ϕ), the one-variate models involving either setting (x_1) or effort (x_2), and the additive model involving setting and effort ($x_1 + x_2$).

More complicated models may be obtained by considering higher order polynomial terms in either variable. Thus, we might consider adding the squares x_1^2 or x_2^2 to capture *non-linearities* in the effects of setting or effort. The squared terms are often highly correlated with the original variables, and on certain datasets this may cause numerical problems in estimation. A simple solution is to reduce the correlation by centering the variables before squaring them, using x_1 and $(x_1 - \bar{x}_1)^2$ instead of x_1 and x_1^2. The correlation can be eliminated entirely, often in the context of designed experiments, by using orthogonal polynomials.

We could also consider adding the cross-product term x_1x_2 to capture a form of *interaction* between setting and effort. In this model the linear predictor would be

$$\mu_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}. \tag{2.17}$$

This is simply a linear model where the model matrix $\mathbf{X}$ has a column of ones for the constant, a column with the values of x_1, a column with the values of x_2, and a column with the products x_1x_2. This is equivalent to creating a new variable, say x_3, which happens to be the product of the other two.

An important feature of this model is that the effect of any given variable now depends on the value of the other. To see this point consider fixing x_1 and viewing the expected response μ as a function of x_2 for this fixed value of x_1. Rearranging terms in Equation 2.17 we find that μ is a linear function of x_2:

$$\mu_i = (\alpha + \beta_1 x_{i1}) + (\beta_2 + \beta_3 x_{i1}) x_{i2},$$

with both constant and slope depending on x_1. Specifically, the effect of x_2 on the response is itself a linear function of x_1; it starts from a baseline effect of β_2 when x_1 is zero, and has an additional effect of β_3 units for each unit increase in x_1.

The extensions considered here help emphasize a very important point about model building: the columns of the model matrix are not necessarily the predictors of interest, but can be any functions of them, including linear, quadratic or cross-product terms, or other transformations.

Are any of these refinements necessary for our example? To find out, fit the more elaborate models and see if you can obtain significant reductions of the residual sum of squares.

2.6 One-Way Analysis of Variance

We now consider models where the predictors are categorical variables or *factors* with a discrete number of levels. To illustrate the use of these models we will group the index of social setting (and later the index of family planning effort) into discrete categories.

2.6.1 The One-Way Layout

Table 2.10 shows the percent decline in the CBR for the 20 countries in our illustrative dataset, classified according to the index of social setting in three

categories: low (under 70 points), medium (70–79) and high (80 or more).

TABLE 2.10: CBR Decline by Levels of Social Setting

Setting	Percent decline in CBR
Low	1, 0, 7, 21, 13, 4, 7
Medium	10, 6, 2, 0, 25
High	9, 11, 29, 29, 40, 21, 22, 29

It will be convenient to modify our notation to reflect the one-way layout of the data explicitly. Let k denote the number of groups or levels of the factor, n_i denote the number of observations in group i, and let y_{ij} denote the response for the j-th unit in the i-th group, for $j = 1, \ldots, n_i$, and $i = 1, \ldots, k$. In our example $k = 3$ and y_{ij} is the CBR decline in the j-th country in the i-th category of social setting, with $i = 1, 2, 3; j = 1, \ldots, n_i; n_1 = 7, n_2 = 5$ and $n_3 = 8$).

2.6.2 The One-Factor Model

As usual, we treat y_{ij} as a realization of a random variable $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$, where the variance is the same for all observations. In terms of the systematic structure of the model, we assume that

$$\mu_{ij} = \mu + \alpha_i, \tag{2.18}$$

where μ plays the role of the constant and α_i represents the effect of level i of the factor.

Before we proceed further, it is important to note that the model as written is not identified. We have essentially k groups but have introduced $k+1$ linear parameters. The solution is to introduce a constraint, and there are several ways in which we could proceed.

One approach is to set $\mu = 0$ (or simply drop μ). If we do this, the α_i's become *cell means*, with α_i representing the expected response in group i. While simple and attractive, this approach does not generalize well to models with more than one factor.

Our preferred alternative is to set one of the α_i's to zero. Conventionally we set $\alpha_1 = 0$, but any of the groups could be chosen as the *reference cell* or level. In this approach μ becomes the expected response in the reference cell, and α_i becomes the effect of level i of the factor, compared to the reference level.

A third alternative is to require the group effects to add-up to zero, so $\sum \alpha_i = 0$. In this case μ represents some sort of overall expected response, and α_i measures the extent to which responses at level i of the factor deviate from the overall mean. Some statistics texts refer to this constraint as the 'usual' restrictions, but I think the reference cell method is now used more widely in social research.

A variant of the 'usual' restrictions is to require a weighted sum of the effects to add up to zero, so $\sum w_i\alpha_i = 0$. The weights are often taken to be the number of observations in each group, so $w_i = n_i$. In this case μ is a weighted average representing the expected response, and α_i is, as before, the extent to which responses at level i of the factor deviate from the overall mean.

Each of these parameterizations can easily be translated into one of the others, so the choice can rest on practical considerations. The reference cell method is easy to implement in a regression context and the resulting parameters have a clear interpretation.

2.6.3 Estimates and Standard Errors

The model in Equation 2.18 is a special case of the generalized linear model, where the design matrix $\mathbf{X}$ has $k+1$ columns: a column of ones representing the constant, and k columns of indicator variables, say $x_1, \ldots, x_k$, where x_i takes the value one for observations at level i of the factor and the value zero otherwise.

Note that the model matrix as defined so far is rank deficient, because the first column is the sum of the last k. Hence the need for constraints. The cell means approach is equivalent to dropping the constant, and the reference cell method is equivalent to dropping one of the indicator or dummy variables representing the levels of the factor. Both approaches are easily implemented. The other two approaches, which set to zero either the unweighted or weighted sum of the effects, are best implemented using Lagrange multipliers and will not be considered here.

Parameter estimates, standard errors and t ratios can then be obtained from the general results of Sections 2.2 and 2.3. You may be interested to know that the estimates of the regression coefficients in the one-way layout are simple functions of the cell means. Using the reference cell method,

$$\hat{\mu} = \bar{y}_1 \qquad \text{and} \qquad \hat{\alpha_i} = \bar{y}_i - \bar{y}_1 \text{ for } i > 1,$$

where $\bar{y}_i$ is the average of the responses at level i of the factor.

Table 2.11 shows the estimates for our sample data. We expect a CBR decline of almost 8% in countries with low social setting (the reference cell). Increasing social setting to medium or high is associated with additional declines of one and 16 percentage points, respectively, compared to low setting.

TABLE 2.11: Estimates for One-Way Anova Model of CBR Decline by Levels of Social Setting

Parameter	Symbol	Estimate	Std. Error	t-ratio
Low	μ	7.571	3.498	2.16
Medium (vs. low)	α_2	1.029	5.420	0.19
High (vs. low)	α_3	16.179	4.790	3.38

Looking at the t ratios we see that the difference between medium and low setting is not significant, so we accept $H_0 : \alpha_2 = 0$, whereas the difference between high and low setting, with a t-ratio of 3.38 on 17 d.f. and a two-sided P-value of 0.004, is highly significant, so we reject $H_0 : \alpha_3 = 0$. These t-ratios test the significance of two particular contrasts: medium vs. low and high vs. low. In the next subsection we consider an overall test of the significance of social setting.

2.6.4 The One-Way Anova Table

Fitting the model with social setting treated as a factor reduces the RSS from 2650.2 (for the null model) to 1456.4, a gain of 1193.8 at the expense of two degrees of freedom (the two α's). We can contrast this gain with the remaining RSS of 1456.4 on 17 d.f. The calculations are laid out in Table 2.12, and lead to an F-test of 6.97 on 2 and 17 d.f., which has a P-value of 0.006. We therefore reject the hypothesis $H_0 : \alpha_2 = \alpha_3 = 0$ of no setting effects, and conclude that the expected response depends on social setting.

TABLE 2.12: Analysis of Variance for One-Factor Model of CBR Decline by Levels of Social Setting

Source of variation	Sum of squares	Degrees of Freedom	Mean squared	F-ratio
Setting	1193.8	2	596.9	6.97
Residual	1456.4	17	85.7	
Total	2650.2	19		

Having established that social setting has an effect on CBR decline, we can inspect the parameter estimates and t-ratios to learn more about the nature of the effect. As noted earlier, the difference between high and low settings is significant, while that between medium and low is not.

It is instructive to calculate the Wald test for this example. Let $\boldsymbol{\alpha} = (\alpha_2, \alpha_3)'$ denote the two setting effects. The estimate and its variance-covariance matrix, calculated using the general results of Section 2.2, are

$$\hat{\boldsymbol{\alpha}} = \begin{pmatrix} 1.029 \\ 16.179 \end{pmatrix} \qquad \text{and} \qquad \widehat{\text{var}}(\hat{\boldsymbol{\alpha}}) = \begin{pmatrix} 29.373 & 12.239 \\ 12.239 & 22.948 \end{pmatrix}.$$

The Wald statistic is

$$W = \hat{\boldsymbol{\alpha}}' \, \widehat{\text{var}}^{-1}(\hat{\boldsymbol{\alpha}}) \, \hat{\boldsymbol{\alpha}} = 13.94,$$

and has approximately a chi-squared distribution with two d.f. Under the assumption of normality, however, we can divide by two to obtain $F = 6.97$, which has an F distribution with two and 17 d.f., and coincides with the test based on the reduction in the residual sum of squares, as shown in Table 2.12.

2.6.5 The Correlation Ratio

Note from Table 2.12 that the model treating social setting as a factor with three levels has reduced the RSS by 1456.6 out of 2650.2, thereby explaining 45.1%. The square root of the proportion of variance explained by a discrete factor is called the *correlation ratio*, and is often denoted η. In our example $\hat{\eta} = 0.672$.

If the factor has only two categories the resulting coefficient is called the *point-biserial correlation*, a measure often used in psychometrics to correlate a test score (a continuous variable) with the answer to a dichotomous item (correct or incorrect). Note that both measures are identical in construction to Pearson's correlation coefficient. The difference in terminology reflects whether the predictor is a continuous variable with a linear effect or a discrete variable with two or more than two categories.

2.7 Two-Way Analysis of Variance

We now consider models involving two factors with discrete levels. We illustrate using the sample data with both social setting and family planning effort grouped into categories. Key issues involve the concepts of main effects and interactions.

2.7.1 The Two-Way Layout

Table 2.13 shows the CBR decline in our 20 countries classified according to two criteria: social setting, with categories low (under 70), medium (70–79) and high (80+), and family planning effort, with categories weak (0–4), moderate (5–14) and strong (15+). In our example both setting and effort are factors with three levels. Note that there are no countries with strong programs in low social settings.

TABLE 2.13: CBR Decline by Levels of Social Setting and Levels of Family Planning Effort

Setting	Effort		
	Weak	Moderate	Strong
Low	1,0,7	21,13,4,7	–
Medium	10,6,2	0	25
High	9	11	29,29,40,21,22,29

We will modify our notation to reflect the two-way layout of the data. Let n_{ij} denote the number of observations in the (i, j)-th cell of the table, i.e. in row i and column j, and let y_{ijk} denote the response of the k-th unit in that cell, for $k = 1, \ldots, n_{ij}$. In our example y_{ijk} is the CBR decline of the k-th country in the i-th category of setting and the j-th category of effort.

2.7.2 The Two-Factor Additive Model

Once again, we treat the response as a realization of a random variable $Y_{ijk} \sim N(\mu_{ijk}, \sigma^2)$. In terms of the systematic component of the model, we will assume that

$$\mu_{ijk} = \mu + \alpha_i + \beta_j \tag{2.19}$$

In this formulation μ represents a baseline value, α_i represents the effect of the i-th level of the row factor and β_j represents the effect of the j-th level of the column factor. Before we proceed further we must note that the model is not identified as stated. You could add a constant δ to each of the α_i's (or to each of the β_j's) and subtract it from μ without altering any of the expected responses. Clearly we need two constraints to identify the model.

Our preferred approach relies on the reference cell method, and sets to zero the effects for the first cell (in the top-left corner of the table), so that $\alpha_1 = \beta_1 = 0$. The best way to understand the meaning of the remaining

parameters is to study Table 2.14, showing the expected response for each combination of levels of row and column factors having three levels each.

TABLE 2.14: The Two-Factor Additive Model

Row	Column		
	1	2	3
1	μ	$\mu+\beta_2$	$\mu+\beta_3$
2	$\mu+\alpha_2$	$\mu+\alpha_2+\beta_2$	$\mu+\alpha_2+\beta_3$
3	$\mu+\alpha_3$	$\mu+\alpha_3+\beta_2$	$\mu+\alpha_3+\beta_3$

In this formulation of the model μ represents the expected response in the reference cell, α_i represents the effect of level i of the row factor (compared to level 1) for any fixed level of the column factor, and β_j represents the effect of level j of the column factor (compared to level 1) for any fixed level of the row factor.

Note that the model is *additive*, in the sense that the effect of each factor is the same at all levels of the other factor. To see this point consider moving from the first to the second row. The response increases by α_2 if we move down the first column, but also if we move down the second or third columns.

2.7.3 Estimates and Standard Errors

The model in Equation 2.19 is a special case of the general linear model, where the model matrix $\mathbf{X}$ has a column of ones representing the constant, and two sets of dummy or indicator variables representing the levels of the row and column factors, respectively. This matrix is not of full column rank because the row (as well as the column) dummies add to the constant. Clearly we need two constraints and we choose to drop the dummy variables corresponding to the first row and to the first column. Table 2.15 shows the resulting parameter estimates, standard errors and t-ratios for our example.

Thus, we expect a CBR decline of 5.4% in countries with low setting and weak programs. In countries with medium or high social setting we expect CBR declines of 1.7 percentage points *less* and 2.4 percentage points more, respectively, than in countries with low setting and the same level of effort. Finally, in countries with moderate or strong programs we expect CBR declines of 3.8 and 20.7 percentage points more than in countries with weak programs and the same level of social setting.

It appears from a cursory examination of the t-ratios in Table 2.15 that the only significant effect is the difference between strong and weak pro-

TABLE 2.15: Parameter Estimates for Two-Factor Additive Model of CBR Decline by Social Setting and Family Planning Effort

Parameter		Symbol	Estimate	Std. Error	t-ratio
Baseline	low/weak	μ	5.379	3.105	1.73
Setting	medium	α_2	−1.681	3.855	−0.44
	high	α_3	2.388	4.457	0.54
Effort	moderate	β_2	3.836	3.575	1.07
	strong	β_3	20.672	4.339	4.76

grams. Bear in mind, however, that the table only shows the comparisons that are explicit in the chosen parameterization. In this example it turns out that the difference between strong and moderate programs is also significant. (This test can be calculated from the variance-covariance matrix of the estimates, or by fitting the model with strong programs as the reference cell, so the medium-strong comparison becomes one of the parameters.) Questions of significance for factors with more than two-levels are best addressed by using the F-test discussed below.

2.7.4 The Hierarchical Anova Table

Fitting the two-factor additive model results in a residual sum of squares of 574.4 on 15 d.f., and represents an improvement over the null model of 2075.8 at the expense of four d.f. We can further decompose this gain as an improvement of 1193.8 on 2 d.f. due to social setting (from Section 2.6) and a gain of 882.0, also on 2 d.f., due to effort given setting. These calculations are set out in Table 2.16, which also shows the corresponding mean squares and F-ratios.

TABLE 2.16: Hierarchical Anova for Two-Factor Additive Model of CBR Decline by Social Setting and Family Planning Effort

Source of variation	Sum of squares	Degrees of freedom	Mean squared	F-ratio
Setting	1193.8	2	596.9	15.6
Effort|Setting	882.0	2	441.0	11.5
Residual	574.4	15	38.3	
Total	2650.2	19		

We can combine the sum of squares for setting and for effort given setting to construct a test for the overall significance of the regression. This results in an F-ratio of 13.6 on four and 15 d.f., and is highly significant. The second of the F-ratios shown in Table 2.16, which is 11.5 on two and 15 d.f., is a test for the *net* effect of family planning effort after accounting for social setting, and is highly significant. (The first of the F-ratios in the table, 15.6 on two and 15 d.f., is not in common use but is shown for completeness; it can be interpreted as an alternative test for the *gross* effect of setting, which combines the same numerator as the test in the previous section with a more refined denominator that takes into account the effect of effort.)

There is an alternative decomposition of the regression sum of squares into an improvement of 2040.0 on two d.f. due to effort and a further gain of 35.8 on two d.f. due to setting given effort. The latter can be contrasted with the error sum of squares of 574.4 on 15 d.f. to obtain a test of the *net* effect of setting given effort. This test would address the question of whether socio-economic conditions have an effect on fertility decline after we have accounted for family planning effort.

2.7.5 Partial and Multiple Correlation Ratios

The sums of squares described above can be turned into proportions of variance explained using the now-familiar calculations. For example the two factors together explain 2075.8 out of 2650.2, or 78.3% of the variation in CBR decline.

The square root of this proportion, 0.885 in the example, is the *multiple correlation ratio*; it is analogous to (and in fact is often called) the multiple correlation coefficient. We use the word 'ratio' to emphasize the categorical nature of the predictors and to note that it generalizes to more than one factor the correlation ratio introduced in Section 2.4.

We can also calculate the proportion of variance explained by one of the factors out of the amount left unexplained by the other. In our example effort explained 882.0 out of the 1456.6 that setting had left unexplained, or 60.6%. The square root of this proportion, 0.778, is called the *partial correlation ratio*, and can be interpreted as a measure of correlation between a discrete factor and a continuous variable after adjustment for another factor.

2.7.6 Fitted Means and Standardization

Parameter estimates from the additive model can be translated into *fitted means* using Equation 2.19 evaluated at the estimates. The body of Table 2.17 shows these values for our illustrative example. Note that we are able to estimate the expected CBR decline for strong programs in low social settings although there is no country in our dataset with that particular combination of attributes. Such extrapolation relies on the additive nature of the model and should be interpreted with caution. Comparison of observed and fitted values can yield useful insights into the adequacy of the model, a topic that will be pursued in more detail when we discuss regression diagnostics later in this chapter.

TABLE 2.17: Fitted Means Based on Two-Factor Additive Model of CBR Decline by Social Setting and Family Planning Effort

Setting	Effort			All
	Weak	Moderate	Strong	
Low	5.38	9.22	26.05	13.77
Medium	3.70	7.54	24.37	12.08
High	7.77	11.60	28.44	16.15
All	5.91	9.75	26.59	14.30

Table 2.17 also shows column (and row) means, representing expected CBR declines by effort (and setting) after adjusting for the other factor. The column means are calculated as weighted averages of the cell means in each column, with weights given by the total number of countries in each category of setting. In symbols

$$\hat{\mu}_{.j} = \sum n_{i.}\hat{\mu}_{ij}/n,$$

where we have used a dot as a subscript placeholder so $n_{i.}$ is the number of observations in row i and $\mu_{.j}$ is the mean for column j.

The resulting estimates may be interpreted as *standardized* means; they estimate the CBR decline that would be expected at each level of effort if those countries had the same distribution of social setting as the total sample. (The column means can also be calculated by using the fitted model to predict CBR decline for each observation with the dummies representing social setting held fixed at their sample averages and all other terms kept as observed. This construction helps reinforce their interpretation in terms of predicted CBR decline at various levels of effort adjusted for setting.)

TABLE 2.18: CBR Decline by Family Planning Effort
Before and After Adjustment for Social Setting

Effort	CBR Decline	
	Unadjusted	Adjusted
Weak	5.00	5.91
Moderate	9.33	9.75
Strong	27.86	26.59

Standardized means may be useful in presenting the results of a regression analysis to a non-technical audience, as done in Table 2.18. The column labelled unadjusted shows the observed mean CBR decline by level of effort. The difference of 23 points between strong and weak programs may be due to program effort, but could also reflect differences in social setting. The column labelled adjusted corrects for compositional differences in social setting using the additive model. The difference of 21 points may be interpreted as an effect of program effort net of social setting.

2.7.7 Multiple Classification Analysis

Multiple Classification Analysis (MCA), a technique that has enjoyed some popularity in Sociology, turns out to be just another name for the two factor additive model discussed in this section (and more generally, for multi-factor additive models). A nice feature of MCA, however, is a tradition of presenting the results of the analysis in a table that contains

- the gross effect of each of the factors, calculated using a one-factor model under the 'usual' restrictions, together with the corresponding correlation ratios (called 'eta' coefficients), and
- the net effect of each factor, calculated using a two-factor additive model under the 'usual' restrictions, together with the corresponding partial correlation ratios (unfortunately called 'beta' coefficients).

Table 2.19 shows a multiple classification analysis of the program effort data that follows directly from the results obtained so far. Estimates for the additive model under the usual restrictions can be obtained from Table 2.18 as differences between the row and column means and the overall mean.

The overall expected decline in the CBR is 14.3%. The effects of low, medium and high setting are substantially reduced after adjustment for ef-

TABLE 2.19: Multiple Classification Analysis of CBR Decline by Social Setting and Family Planning Effort

Factor	Category	Gross Effect	Eta	Net Effect	Beta
Setting	Low	−6.73		−0.54	
	Medium	−5.70		−2.22	
	High	9.45		1.85	
			0.67		0.24
Effort	Weak	−9.30		−8.39	
	Moderate	−4.97		−4.55	
	Strong	13.56		12.29	
			0.88		0.78
Total		14.30		14.30	

fort, an attenuation reflected in the reduction of the correlation ratio from 0.67 to 0.24. On the other hand, the effects of weak, moderate and strong programs are slightly reduced after adjustment for social setting, as can be seen from correlation ratios of 0.88 and 0.78 before and after adjustment. The analysis indicates that the effects of effort are more pronounced and more resilient to adjustment than the effects of social setting.

2.7.8 The Model With Interactions

The analysis so far has rested on the assumption of additivity. We now consider a more general model for the effects of two discrete factors on a continuous response which allows for more general effects

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}. \tag{2.20}$$

In this formulation the first three terms should be familiar: μ is a constant, and α_i and β_j are the *main* effects of levels i of the row factor and j of the column factor.

The new term $(\alpha\beta)_{ij}$ is an *interaction* effect. It represents the effect of the *combination* of levels i and j of the row and column factors. (The notation $(\alpha\beta)$ should be understood as a single symbol, not a product; we could have used γ_{ij} to denote the interaction, but the notation $(\alpha\beta)_{ij}$ is more suggestive and reminds us that the term involves a combined effect.)

One difficulty with the model as defined so far is that it is grossly overparameterized. If the row and column factors have R and C levels, respectively,

we have only RC possible cells but have introduced $1+R+C+RC$ parameters. Our preferred solution is an extension of the reference cell method, and sets to zero all parameters involving the first row or the first column in the two-way layout, so that $\alpha_1 = \beta_1 = (\alpha\beta)_{1j} = (\alpha\beta)_{i1} = 0$. The best way to understand the meaning of the remaining parameters is to study Table 2.20, which shows the structure of the means in a three by three layout.

TABLE 2.20: The Two-Factor Model With Interactions

Row	Column		
	1	2	3
1	μ	$\mu+\beta_2$	$\mu+\beta_3$
2	$\mu+\alpha_2$	$\mu+\alpha_2+\beta_2+(\alpha\beta)_{22}$	$\mu+\alpha_2+\beta_3+(\alpha\beta)_{23}$
3	$\mu+\alpha_3$	$\mu+\alpha_3+\beta_2+(\alpha\beta)_{32}$	$\mu+\alpha_3+\beta_3+(\alpha\beta)_{33}$

Here μ is the expected response in the reference cell, just as before. The main effects are now more specialized: α_i is the effect of level i of the row factor, compared to level one, when the column factor is at level one, and β_j is the effect of level j of the column factor, compared to level one, when the row factor is at level one. The interaction term $(\alpha\beta)_{ij}$ is the additional effect of level i of the row factor, compared to level one, when the column factor is at level j rather than one. This term can also be interpreted as the additional effect of level j of the column factor, compared to level one, when the row factor is at level i rather than one.

The key feature of this model is that the effect of a factor now depends on the levels of the other. For example the effect of level two of the row factor, compared to level one, is α_2 in the first column, $\alpha_2+(\alpha\beta)_{22}$ in the second column, and $\alpha_2+(\alpha\beta)_{23}$ in the third column.

The resulting model is a special case of the general lineal model where the model matrix $\mathbf{X}$ has a column of ones to represent the constant, a set of $R-1$ dummy variables representing the row effects, a set of $C-1$ dummy variables representing the column effects, and a set of $(R-1)(C-1)$ dummy variables representing the interactions.

The easiest way to calculate the interaction dummies is as products of the row and column dummies. If r_i takes the value one for observations in row i and zero otherwise, and c_j takes the value one for observations in column $j1$ and zero otherwise, then the product $r_i c_j$ takes the value one for observations that are in row i *and* column j, and is zero for all others.

In order to fit this model to the program effort data we need to introduce one additional constraint because the cell corresponding to strong programs

in low settings is empty. As a result, we cannot distinguish β_3 from $\beta_3 + (\alpha\beta)_{23}$. A simple solution is to set $(\alpha\beta)_{23} = 0$. This constraint is easily implemented by dropping the corresponding dummy, which would be r_2c_3 in the above notation.

The final model has eight parameters: the constant, two setting effects, two effort effects, and three (rather than four) interaction terms.

TABLE 2.21: Anova for Two-Factor Model with Interaction Effect for CBR Decline by Social Setting and Family Planning Effort

Source of variation	Sum of squares	Degrees of freedom	Mean squared	F-ratio
Setting	1193.8	2	596.9	15.5
Effort|Setting	882.0	2	441.0	11.5
Interaction	113.6	3	37.9	1.0
Residual	460.8	12	38.4	
Total	2650.2	19		

Fitting the model gives a RSS of 460.8 on 12 d.f. Combining this result with the anova for the additive model leads to the hierarchical anova in Table 2.21. The F-test for the interaction is one on three and 12 d.f. and is clearly not significant. Thus, we have no evidence to contradict the assumption of additivity. We conclude that the effect of effort is the same at all social settings. Calculation and interpretation of the parameter estimates is left as an exercise.

2.7.9 Factors or Variates?

In our analysis of CBR decline we treated social setting and family planning effort as continuous *variates* with linear effects in Sections 2.4 and 2.5, and as discrete *factors* in Sections 2.6 and 2.7.

The fundamental difference between the two approaches hinges on the assumption of linearity. When we treat a predictor as a continuous variate we assume a linear effect. If the assumption is reasonable we attain a parsimonious fit, but if it is not reasonable we are forced to introduce transformations or higher-order polynomial terms, resulting in models which are often harder to interpret.

A reasonable alternative in these cases is to model the predictor as a discrete factor, an approach that allows arbitrary changes in the response from one category to another. This approach has the advantage of a simpler

and more direct interpretation, but by grouping the predictor into categories we are not making full use of the information in the data.

In our example we found that social setting explained 45% of the variation in CBR declines when treated as a variate and 45% when treated as a factor with three levels. Both approaches give the same result, suggesting that the assumption of linearity of setting effects is reasonable.

On the other hand family planning effort explained 64% when treated as a variate and 77% when treated as a factor with three levels. The difference suggests that we might be better off grouping effort into three categories. The reason, of course, is that the effect of effort is non-linear: CBR decline changes little as we move from weak to moderate programs, but raises steeply for strong programs.

2.8 Analysis of Covariance Models

We now consider models where some of the predictors are continuous variates and some are discrete factors. We continue to use the family planning program data, but this time we treat social setting as a variate and program effort as a factor.

2.8.1 The Data and Notation

Table 2.22 shows the effort data classified into three groups, corresponding to weak (0–4), moderate (5–14) and strong (15+) programs. For each group we list the values of social setting and CBR decline.

TABLE 2.22: Social Setting Scores and CBR Percent Declines by Levels of Family Planning Effort

Family Planning Effort					
Weak		Moderate		Strong	
Setting	Change	Setting	Change	Setting	Change
46	1	68	21	89	29
74	10	70	0	77	25
35	0	60	13	84	29
83	9	55	4	89	40
68	7	51	7	87	21
74	6	91	11	84	22
72	2			84	29

As usual, we modify our notation to reflect the structure of the data. Let k denote the number of groups, or levels of the discrete factor, n_i the number of observations in group i, y_{ij} the value of the response and x_{ij} the value of the predictor for the j-th unit in the i-th group, with $j = 1, \ldots, n_i$ and $i = 1, \ldots, k$.

2.8.2 The Additive Model

We keep the random structure of our model, treating y_{ij} as a realization of a random variable $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$. To express the dependence of the expected response μ_{ij} on a discrete factor we have used an anova model of the form $\mu_{ij} = \mu + \alpha_i$, whereas to model the effect of a continuous predictor we have used a regression model of the form $\mu_{ij} = \alpha + \beta x_{ij}$. Combining these two models we obtain the additive analysis of covariance model

$$\mu_{ij} = \mu + \alpha_i + \beta x_{ij}. \tag{2.21}$$

This model defines a series of straight-line regressions, one for each level of the discrete factor (you may want to peek at Figure 2.5). These lines have different intercepts $\mu + \alpha_i$, but a common slope β, so they are *parallel*. The common slope β represents the effects of the continuous variate at any level of the factor, and the differences in intercept α_i represent the effects of the discrete factor at any given value of the covariate.

The model as defined in Equation 2.21 is not identified: we could add a constant δ to each α_i and subtract it from μ without changing any of the expected values. To solve this problem we set $\alpha_1 = 0$, so μ becomes the intercept for the reference cell, and α_i becomes the difference in intercepts between levels i and one of the factor.

The analysis of covariance model may be obtained as a special case of the general linear model by letting the model matrix $\mathbf{X}$ have a column of ones representing the constant, a set of k dummy variables representing the levels of the discrete factor, and a column with the values of the continuous variate. The model is not of full column rank because the dummies add up to the constant, so we drop one of them, obtaining the reference cell parametrization. Estimation and testing then follows form the general results in Sections 2.2 and 2.3.

Table 2.23 shows the parameter estimates, standard errors and t-ratios after fitting the model to the program effort data with setting as a variate and effort as a factor with three levels.

The results show that each point in the social setting scale is associated with a further 0.17 percentage points of CBR decline at any given level of

TABLE 2.23: Parameter Estimates for Analysis of Covariance Model of CBR Decline by Social Setting and Family Planning Effort

Parameter		Symbol	Estimate	Std.Error	t-ratio
Constant		μ	-5.954	7.166	-0.83
Effort	moderate	α_2	4.144	3.191	1.30
	strong	α_3	19.448	3.729	5.21
Setting	(linear)	β	0.1693	0.1056	1.60

effort. Countries with moderate and strong programs show additional CBR declines of 19 and 4 percentage points, respectively, compared to countries with weak programs at the same social setting.

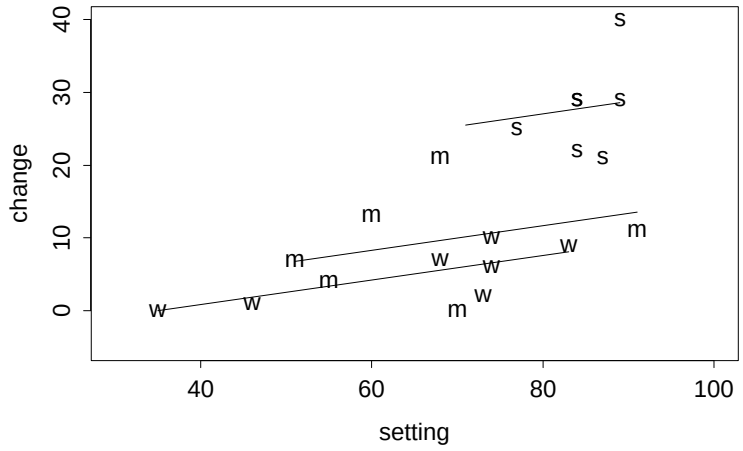


FIGURE 2.5: Analysis of Covariance Model for CBR Decline by Social Setting Score and Level of Program Effort

Figure 2.5 depicts the analysis of covariance model in graphical form. We have plotted CBR decline as a function of social setting using the letters w, m and s for weak, moderate and strong programs, respectively. The figure also shows the fitted lines for the three types of programs. The vertical distances between the lines represent the effects of program effort at any given social setting. The common slope represents the effect of setting at any given level of effort.

2.8.3 The Hierarchical Anova Table

Fitting the analysis of covariance model to our data gives a RSS of 525.7 on 16 d.f. (20 observations minus four parameters: the constant, two intercepts and one slope). Combining this result with the RSS's for the null model and for the model of Section 2.4 with a linear effect of setting, leads to the hierarchical analysis of variance shown in Table 2.24.

TABLE 2.24: Hierarchical Anova for Analysis of Covariance Model of CBR Decline by Social Setting and Family Planning Effort

Source of variation	Sum of squares	Degrees of freedom	Mean squared	F-ratio
Setting (linear)	1201.1	1	1201.1	36.5
Effort|Setting	923.4	2	461.7	14.1
Residual	525.7	16	32.9	
Total	2650.2	19		

The most interesting statistic in this table is the F-test for the net effect of program effort, which is 14.1 on two and 16 d.f. and is highly significant, so we reject the hypothesis $H_0 : \alpha_2 = \alpha_3 = 0$ of no program effects. Looking at the t-ratios in Table 2.23 we see that the difference between strong and weak programs is significant, while that between moderate and weak programs is not, confirming our earlier conclusions. The difference between strong and moderate programs, which is not shown in the table, is also significant.

From these results we can calculate proportions of variance explained in the usual fashion. In this example setting explains 45.3% of the variation in CBR declines and program effort explains an additional 34.5%, representing 63.7% of what remained unexplained, for a total of 80.1%. You should be able to translate these numbers into simple, partial and multiple correlation coefficients or ratios.

2.8.4 Gross and Net Effects

The estimated net effects of setting and effort based on the analysis of covariance model may be compared with the estimated gross effects based on the simple linear regression model for setting and the one-way analysis of variance model for effort. The results are presented in a format analogous to multiple classification analysis in Table 2.25, where we have used the reference cell method rather than the 'usual' restrictions.

TABLE 2.25: Gross and Net Effects of Social Setting Score and Level of Family Planning Effort on CBR Decline

Predictor	Category	Effect	
		Gross	Net
Setting	(linear)	0.505	0.169
Effort	Weak	–	–
	Moderate	4.33	4.14
	Strong	22.86	19.45

The effect of social setting is reduced substantially after adjusting for program effort. On the other hand, the effects of program effort, measured by comparing strong and moderate programs with weak ones, are hardly changed after adjustment for social setting.

If interest centers on the effects of program effort, it may be instructive to calculate CBR declines by categories of program effort unadjusted and adjusted for linear effects of setting. To obtain adjusted means we use the fitted model to predict CBR decline with program effort set at the observed values but social setting set at the sample mean, which is 72.1 points. Thus, we calculate expected CBR decline at level i of effort holding setting constant at the mean as $\hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}\,72.1$. The results are shown in Table 2.26.

TABLE 2.26: CBR Decline by Family Planning Effort Before and After Linear Adjustment for Social Setting

Effort	CBR Decline	
	Unadjusted	Adjusted
Weak	5.00	6.25
Moderate	9.33	10.40
Strong	27.86	25.70

Thus, countries with strong program show on average a 28% decline in the CBR, but these countries tend to have high social settings. If we adjusted linearly for this advantage, we would expect them to show only a 26% decline. Clearly, adjusting for social setting does not change things very much.

Note that the analysis in this sections parallels the results in Section 2.7. The only difference is the treatment of social setting as a discrete factor with

three levels or as a continuous variate with a linear effect.

2.8.5 The Assumption of Parallelism

In order to test the assumption of equal slopes in the analysis of covariance model we consider a more general model where

$$\mu_{ij} = (\mu + \alpha_i) + (\beta + \gamma_i)x_{ij}. \tag{2.22}$$

In this formulation each of the k groups has its own intercept $\mu + \alpha_i$ and its own slope $\beta + \gamma_i$.

As usual, this model is overparametrized and we introduce the reference cell restrictions, setting $\alpha_1 = \gamma_1 = 0$. As a result, μ is the constant and β is the slope for the reference cell, α_i and γ_i are the differences in intercept and slope, respectively, between level i and level one of the discrete factor. (An alternative is to drop μ and β, so that α_i is the constant and γ_i is the slope for group i. The reference cell method, however, extends more easily to models with more than one discrete factor.)

The parameter α_i may be interpreted as the effect of level i of the factor, compared to level one, when the covariate is zero. (This value will not be of interest if zero is not in the range of the data.) On the other hand, β is the expected increase in the response per unit increment in the variate when the factor is at level one. The parameter γ_i is the additional expected increase in the response per unit increment in the variate when the factor is at level i rather than one. Also, the product $\gamma_i x$ is the additional effect of level i of the factor when the covariate has value x rather than zero.

Before fitting this model to the program effort data we take the precaution of centering social setting by subtracting its mean. This simple transformation simplifies interpretation of the intercepts, since a value of zero represents the mean setting and is therefore definitely in the range of the data. The resulting parameter estimates, standard errors and t-ratios are shown in Table 2.27.

The effect of setting is practically the same for countries with weak and moderate programs, but appears to be more pronounced in countries with strong programs. Note that the slope is 0.18 for weak programs but increases to 0.64 for strong programs. Equivalently, the effect of strong programs compared to weak ones seems to be somewhat more pronounced at higher levels of social setting. For example strong programs show 13 percentage points more CBR decline than weak programs at average levels of setting, but the difference increases to 18 percentage points if setting is 10 points

TABLE 2.27: Parameter Estimates for Ancova Model with Different Slopes for CBR Decline by Social Setting and Family Planning Effort
(Social setting centered around its mean)

Parameter		Symbol	Estimate	Std.Error	t-ratio
Constant		μ	6.356	2.477	2.57
Effort	moderate	α_2	3.584	3.662	0.98
	strong	α_3	13.333	8.209	1.62
Setting	(linear)	β	0.1836	0.1397	1.31
Setting ×	moderate	γ_2	−0.0868	0.2326	−0.37
Effort	strong	γ_3	0.4567	0.6039	0.46

above the mean. However, the t ratios suggest that none of these interactions is significant.

To test the hypothesis of parallelism (or no interaction) we need to consider the joint significance of the two coefficients representing differences in slopes, i.e. we need to test $H_0 : \gamma_2 = \gamma_3 = 0$. This is easily done comparing the model of this subsection, which has a RSS of 497.1 on 14 d.f., with the parallel lines model of the previous subsection, which had a RSS of 525.7 on 16 d.f. The calculations are set out in Table 2.28.

TABLE 2.28: Hierarchical Anova for Model with Different Slopes of CBR Decline by Social Setting and Family Planning Effort

Source of variation	Sum of squares	Degrees of freedom	Mean squared	F-ratio
Setting (linear)	1201.1	1	1201.1	33.8
Effort (intercepts)	923.4	2	461.7	13.0
Setting × Effort (slopes)	28.6	2	14.3	0.4
Residual	497.1	14	35.5	
Total	2650.2	19		

The test for parallelism gives an F-ratio of 0.4 on two and 14 d.f., and is clearly not significant. We therefore accept the hypothesis of parallelism and conclude that we have no evidence of an interaction between program effort and social setting.

2.9 Regression Diagnostics

The process of statistical modeling involves three distinct stages: formulating a model, fitting the model to data, and checking the model. Often, the third stage suggests a reformulation of the model that leads to a repetition of the entire cycle and, one hopes, an improved model. In this section we discuss techniques that can be used to check the model.

2.9.1 Fitted Values and Residuals

The raw materials of model checking are the *residuals* r_i defined as the differences between observed and fitted values

$$r_i = y_i - \hat{y}_i, \tag{2.23}$$

where y_i is the observed response and $\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ is the fitted value for the i-th unit.

The fitted values may be written in matrix notation as $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Using Equation 2.7 for the m.l.e. of $\boldsymbol{\beta}$, we can write the fitted values as $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

The matrix $\mathbf{H}$ is called the *hat* matrix because it maps y into y-hat. From these results one can show that the fitted values have mean $E(\hat{\mathbf{y}}) = \mu$ and variance-covariance matrix $\text{var}(\hat{\mathbf{y}}) = \mathbf{H}\sigma^2$.

The residuals may be written in matrix notation as $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$, where $\mathbf{y}$ is the vector of responses and $\hat{\mathbf{y}}$ is the vector of fitted values. Since $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, we can write the raw residuals as $\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{y}$. It is then a simple matter to verify that under the usual second-order assumptions, the residuals have expected value $\mathbf{0}$ and variance-covariance matrix $\text{var}(\mathbf{r}) = (\mathbf{I} - \mathbf{H})\sigma^2$. In particular, the variance of the i-th residual is

$$\text{var}(r_i) = (1 - h_{ii})\sigma^2, \tag{2.24}$$

where h_{ii} is the i-th diagonal element of the hat matrix.

This result shows that the residuals may have different variances even when the original observations all have the same variance σ^2, because the precision of the fitted values depends on the pattern of covariate values.

For models with a constant it can be shown that the value of h_{ii} is always between $1/n$ and $1/r$, where n is the total number of observations and r is the number of replicates of the i-th observation (the number of units with

the same covariate values as the i-th unit). In simple linear regression with a constant and a predictor x we have

$$h_{ii} = 1/n + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2},$$

so that h_{ii} has a minimum of $1/n$ at the mean of x. Thus, the variance of the fitted values is smallest for observations near the mean and increases towards the extremes, as you might have expected. Perhaps less intuitively, this implies that the variance of the residuals is greatest near the mean and decreases as one moves towards either extreme.

Table 2.29 shows raw residuals (and other quantities to be discussed below) for the covariance analysis model fitted to the program effort data. Note that the model underestimates the decline of fertility in both Cuba and the Dominican Republic by a little bit more than eleven percentage points. At the other end of the scale, the model overestimates fertility change in Ecuador by ten percentage points.

2.9.2 Standardized Residuals

When we compare residuals for different observations we should take into account the fact that their variances may differ. A simple way to allow for this fact is to divide the raw residual by an estimate of its standard deviation, calculating the *standardized* (or internally studentized) residual

$$s_i = \frac{r_i}{\sqrt{1 - h_{ii}}\hat{\sigma}}, \tag{2.25}$$

where $\hat{\sigma}$ is the estimate of the standard deviation based on the residual sum of squares.

Standardized residuals are useful in detecting anomalous observations or *outliers*. In general, any observation with a standardized residual greater than two in absolute value should be considered worthy of further scrutiny although, as we shall see below, such observations are not necessarily outliers.

Returning to Table 2.29, we see that the residuals for both Cuba and the Dominican Republic exceed two in absolute value, whereas the residual for Ecuador does not. Standardizing the residuals helps assess their magnitude relative to the precision of the estimated regression.

2.9.3 Jack-knifed Residuals

One difficulty with standardized residuals is that they depend on an estimate of the standard deviation that may itself be affected by outliers, which may

TABLE 2.29: Regression Diagnostics for Analysis of Covariance Model of CBR Decline by Social Setting and Program Effort

Country	Residual			Leverage	Cook's
	r_i	s_i	t_i	h_{ii}	D_i
Bolivia	−0.83	−0.17	−0.16	0.262	0.0025
Brazil	3.43	0.66	0.65	0.172	0.0225
Chile	0.44	0.08	0.08	0.149	0.0003
Colombia	−1.53	−0.29	−0.28	0.164	0.0042
Costa Rica	1.29	0.24	0.24	0.143	0.0025
Cuba	11.44	**2.16**	**2.49**	0.149	**0.2043**
Dominican Rep.	11.30	**2.16**	**2.49**	0.168	**0.2363**
Ecuador	−10.04	−1.93	**−2.13**	0.173	**0.1932**
El Salvador	4.65	0.90	0.89	0.178	0.0435
Guatemala	−3.50	−0.69	−0.67	0.206	0.0306
Haiti	0.03	0.01	0.01	**0.442**	0.0000
Honduras	0.18	0.04	0.03	0.241	0.0001
Jamaica	−7.22	−1.36	−1.40	0.144	0.0782
Mexico	0.90	0.18	0.18	0.256	0.0029
Nicaragua	1.44	0.27	0.26	0.147	0.0032
Panama	−5.71	−1.08	−1.08	0.143	0.0484
Paraguay	−0.57	−0.11	−0.11	0.172	0.0006
Peru	−4.40	−0.84	−0.83	0.166	0.0352
Trinidad-Tobago	1.29	0.24	0.24	0.143	0.0025
Venezuela	−2.59	−0.58	−0.56	0.381	0.0510

thereby escape detection.

A solution to this problem is to standardize the i-th residual using an estimate of the error variance obtained by *omitting* the i-th observation. The result is the so-called *jack-knifed* (or externally studentized, or sometimes just studentized) residual

$$t_i = \frac{r_i}{\sqrt{1-h_{ii}}\hat{\sigma}_{(i)}}, \tag{2.26}$$

where $\hat{\sigma}_{(i)}$ denotes the estimate of the standard deviation obtained by fitting the model without the i-th observation, and is based on a RSS with $n-p-1$ d.f. Note that the fitted value and the hat matrix are still based on the model with all observations.

You may wonder what would happen if we omitted the i-th observation not just for purposes of standardizing the residual, but also when estimating the residual itself. Let $\hat{\boldsymbol{\beta}}_{(i)}$ denote the estimate of the regression coefficients obtained by omitting the i-th observation. We can combine this estimate with the covariate values of the i-th observation to calculate a predicted response $\hat{y}_{(i)} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(i)}$ based on the rest of the data. The difference between observed and predicted responses is sometimes called a *predictive* residual

$$y_i - \hat{y}_{(i)}.$$

Consider now standardizing this residual, dividing by an estimate of its standard deviation. Since the i-th unit was not included in the regression, y_i and $\hat{y}_{(i)}$ are independent. The variance of the predictive residual is

$$\text{var}(y_i - \hat{y}_{(i)}) = (1 + \mathbf{x}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{x}_i)\sigma^2,$$

where $\mathbf{X}_{(i)}$ is the model matrix without the i-th row. This variance is estimated replacing the unknown σ^2 by $\hat{\sigma}^2_{(i)}$, the estimate based on the RSS of the model omitting the i-th observation. We are now in a position to calculate a standardized predictive residual

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\sqrt{\hat{\text{var}}(y_i - \hat{y}_{(i)})}}. \tag{2.27}$$

The result turns out to be exactly the same as the jack-knifed residual in Equation 2.26 and provides an alternative characterization of this statistic.

At first sight it might appear that jack-knifed residuals require a lot of calculation, as we would need to fit the model omitting each observation in turn. It turns out, however, that there are simple updating formulas that allow direct calculation of regression coefficients and RSS's after omitting one observation (see Weisberg, 1985, p. 293). These formulas can be used to show that the jack-knifed residual t_i is a simple function of the standardized residual s_i

$$t_i = s_i\sqrt{\frac{n-p-1}{n-p-s_i^2}}.$$

Note that t_i is a monotonic function of s_i, so ranking observations by their standardized residuals is equivalent to ordering them by their jack-knifed residuals.

The jack-knifed residuals on Table 2.29 make Cuba and the D.R. stand out more clearly, and suggest that Ecuador may also be an outlier.

2.9.4 A Test For Outliers

The jack-knifed residual can also be motivated as a formal test for outliers. Suppose we start from the model $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ and add a dummy variable to allow a location shift for the i-th observation, leading to the model

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta} + \gamma z_i,$$

where z_i is a dummy variable that takes the value one for the i-th observation and zero otherwise. In this model γ represents the extent to which the i-th response differs from what would be expected on the basis of its covariate values $\mathbf{x}_i$ and the regression coefficients $\boldsymbol{\beta}$. A formal test of the hypothesis

$$H_0 : \gamma = 0$$

can therefore be interpreted as a test that the i-th observation follows the same model as the rest of the data (i.e. is not an outlier).

The Wald test for this hypothesis would divide the estimate of γ by its standard error. Remarkably, the resulting t-ratio,

$$t_i = \frac{\hat{\gamma}}{\sqrt{\text{var}(\hat{\gamma})}}$$

on $n - p - 1$ d.f., is none other than the jack-knifed residual.

This result should not be surprising in light of the previous developments. By letting the i-th observation have its own parameter γ, we are in effect estimating $\boldsymbol{\beta}$ from the rest of the data. The estimate of γ measures the difference between the response and what would be expected from the rest of the data, and coincides with the predictive residual.

In interpreting the jack-knifed residual as a test for outliers one should be careful with levels of significance. If the suspect observation had been picked in advance then the test would be valid. If the suspect observation has been selected after looking at the data, however, the nominal significance level is not valid, because we have implicitly conducted more than one test. Note that if you conduct a series of tests at the 5% level, you would expect one in twenty to be significant by chance alone.

A very simple procedure to control the overall significance level when you plan to conduct k tests is to use a significance level of α/k for each one. A basic result in probability theory known as the *Bonferroni* inequality guarantees that the overall significance level will not exceed α. Unfortunately, the procedure is conservative, and the true significance level could be considerably less than α.

For the program effort data the jack-knifed residuals have $20-4-1=15$ d.f. To allow for the fact that we are testing 20 of them, we should use a significance level of $0.05/20 = 0.0025$ instead of 0.05. The corresponding two-sided critical value of the Student's t distribution is $t_{.99875,15} = 3.62$, which is substantially higher than the standard critical value $t_{.975,15} = 2.13$. The residuals for Cuba, the D.R. and Ecuador do not exceed this more stringent criterion, so we have no evidence that these countries depart systematically from the model.

2.9.5 Influence and Leverage

Let us return for a moment to the diagonal elements of the hat matrix. Note from Equation 2.24 that the variance of the residual is the product of $1-h_{ii}$ and σ^2. As h_{ii} approaches one the variance of the residual approaches zero, indicating that the fitted value $\hat{y}_i$ is forced to come close to the observed value y_i. In view of this result, the quantity h_{ii} has been called the *leverage* or potential influence of the i-th observation. Observations with high leverage require special attention, as the fit may be overly dependent upon them.

An observation is usually considered to have high leverage if h_{ii} exceeds $2p/n$, where p is the number of predictors, including the constant, and n is the number of observations. This tolerance is not entirely arbitrary. The trace or sum of diagonal elements of $\mathbf{H}$ is p, and thus the average leverage is p/n. An observation is influential if it has more than twice the mean leverage.

Table 2.29 shows leverage values for the analysis of covariance model fitted to the program effort data. With 20 observations and four parameters, we would consider values of h_{ii} exceeding 0.4 as indicative of high leverage. The only country that exceeds this tolerance is Haiti, but Venezuela comes close. Haiti has high leverage because it is found rather isolated at the low end of the social setting scale. Venezuela is rather unique in having high social setting but only moderate program effort.

2.9.6 Actual Influence and Cook's Distance

Potential influence is not the same as actual influence, since it is always possible that the fitted value $\hat{y}_i$ would have come close to the observed value y_i anyway. Cook proposed a measure of influence based on the extent to which parameter estimates would change if one omitted the i-th observation. We define *Cook's Distance* as the standardized difference between $\hat{\boldsymbol{\beta}}_{(i)}$, the estimate obtained by omitting the i-th observation, and $\hat{\boldsymbol{\beta}}$, the estimate

obtained using all the data

$$D_i = (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})' \, \hat{\text{var}}^{-1}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})/p. \tag{2.28}$$

It can be shown that Cook's distance is also the Euclidian distance (or sum of squared differences) between the fitted values $\hat{\mathbf{y}}_{(i)}$ obtained by omitting the i-th observation and the fitted values $\hat{\mathbf{y}}$ based on all the data, so that

$$D_i = \sum_{j=1}^{n} (\hat{y}_{(i)j} - \hat{y}_j)^2/(p\hat{\sigma}^2). \tag{2.29}$$

This result follows readily from Equation 2.28 if you note that $\text{var}^{-1}(\hat{\boldsymbol{\beta}}) = \mathbf{X}'\mathbf{X}/\sigma^2$ and $\hat{\mathbf{y}}_{(i)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)}$.

It would appear from this definition that calculation of Cook's distance requires a lot of work, but the regression updating formulas mentioned earlier simplify the task considerably. In fact, D_i turns out to be a simple function of the standardized residual s_i and the leverage h_{ii},

$$D_i = s_i^2 \frac{h_{ii}}{(1 - h_{ii})p}.$$

Thus, Cook's distance D_i combines residuals and leverages in a single measure of influence.

Values of D_i near one are usually considered indicative of excessive influence. To provide some motivation for this rule of thumb, note from Equation 2.28 that Cook's distance has the form W/p, where W is formally identical to the Wald statistic that one would use to test H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ if one hypothesized the value $\hat{\boldsymbol{\beta}}_{(i)}$. Recalling that W/p has an F distribution, we see that Cook's distance is equivalent to the F statistic for testing this hypothesis. A value of one is close to the median of the F distribution for a large range of values of the d.f. An observation has excessive influence if deleting it would move this F statistic from zero to the median, which is equivalent to moving the point estimate to the edge of a 50% confidence region. In such cases it may be wise to repeat the analysis without the influential observation and examine which estimates change as a result.

Table 2.29 shows Cook's distance for the analysis of covariance model fitted to the program effort data. The D.R., Cuba and Ecuador have the largest indices, but none of them is close to one. To investigate the exact nature of the D.R.'s influence, I fitted the model excluding this country. The main result is that the parameter representing the difference between moderate and weak programs is reduced from 4.14 to 1.89. Thus, a large part

of the evidence pointing to a difference between moderate and weak programs comes from the D.R., which happens to be a country with substantial fertility decline and only moderate program effort. Note that the difference was not significant anyway, so no conclusions would be affected.

Note also from Table 2.29 that Haiti, which had high leverage or potential influence, turned out to have no actual influence on the fit. Omitting this country would not alter the parameter estimates at all.

2.9.7 Residual Plots

One of the most useful diagnostic tools available to the analyst is the residual plot, a simple scatterplot of the residuals r_i versus the fitted values $\hat{y}_i$. Alternatively, one may plot the standardized residuals s_i or the jack-knifed residuals t_i versus the fitted values. In all three cases we expect basically a rectangular cloud with no discernible trend or pattern. Figure 2.6 shows a plot of jack-knifed residuals for the analysis of covariance model fitted to the program effort data. Some of the symptoms that you should be alert for

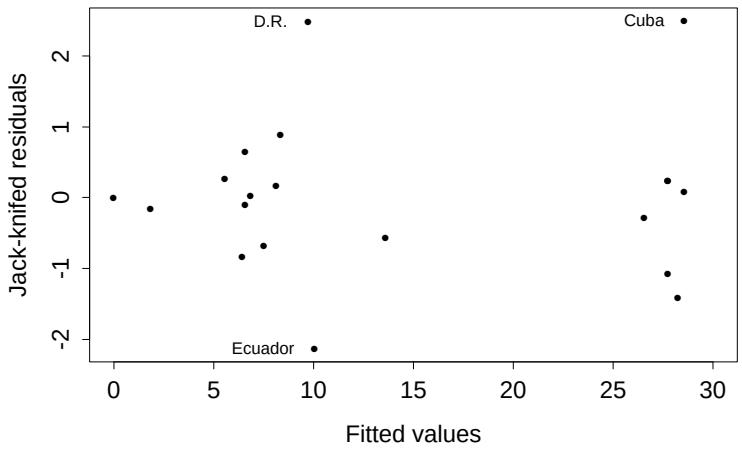


FIGURE 2.6: Residual Plot for Analysis of Covariance Model of CBR Decline by Social Setting and Program Effort

when inspecting residual plots include the following:

- Any trend in the plot, such as a tendency for negative residuals at small $\hat{y}_i$ and positive residuals at large $\hat{y}_i$. Such a trend would indicate non-linearities in the data. Possible remedies include transforming the response or introducing polynomial terms on the predictors.

- Non-constant spread of the residuals, such as a tendency for more clustered residuals for small $\hat{y}_i$ and more dispersed residuals for large $\hat{y}_i$. This type of symptom results in a cloud shaped like a megaphone, and indicates heteroscedasticity or non-constant variance. The usual remedy is a transformation of the response.

For examples of residual plots see Weisberg (1985) or Draper and Smith (1966).

2.9.8 The Q-Q Plot

A second type of diagnostic aid is the probability plot, a graph of the residuals versus the expected order statistics of the standard normal distribution. This graph is also called a *Q-Q Plot* because it plots quantiles of the data versus quantiles of a distribution. The Q-Q plot may be constructed using raw, standardized or jack-knifed residuals, although I recommend the latter.

The first step in constructing a Q-Q plot is to order the residuals from smallest to largest, so $r_{(i)}$ is the i-th smallest residual. The quantity $r_{(i)}$ is called an *order statistic*. The smallest value is the first order statistic and the largest out of n is the n-th order statistic.

The next step is to imagine taking a sample of size n from a standard normal distribution and calculating the order statistics, say $z_{(i)}$. The expected values of these order statistics are sometimes called *rankits*. A useful approximation to the i-th rankit in a sample of size n is given by

$$E(\mathbf{z}_{(i)}) \approx \Phi^{-1}[(i - 3/8)/(n + 1/4)]$$

where Φ^{-1} denotes the inverse of the standard normal distribution function. An alternative approximation proposed by Filliben (1975) uses $\Phi^{-1}[(i - 0.3175)/(n+0.365)]$ except for the first and last rankits, which are estimated as $\Phi^{-1}(1 - 0.5^{1/n})$ and $\Phi^{-1}(0.5^{1/n})$, respectively. The two approximations give very similar results.

If the observations come from a normal distribution we would expect the observed order statistics to be reasonably close to the rankits or expected order statistics. In particular, if we plot the order statistics versus the rankits we should get approximately a straight line.

Figure 2.7 shows a Q-Q plot of the jack-knifed residuals from the analysis of covariance model fitted to the program effort data. The plot comes very close to a straight line, except possibly for the upper tail, where we find a couple of residuals somewhat larger than expected. In general, Q-Q plots showing curvature indicate skew distributions, with downward concavity corresponding to negative skewness (long tail to the left) and upward

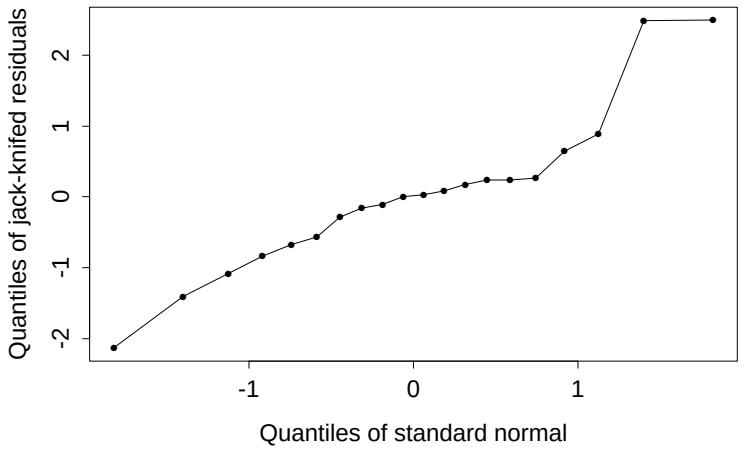


FIGURE 2.7: Q-Q Plot of Residuals From Analysis of Covariance Model of CBR Decline by Social Setting and Program Effort

concavity indicating positive skewness. On the other hand, S-shaped Q-Q plots indicate heavy tails, or an excess of extreme values, relative to the normal distribution.

Filliben (1975) has proposed a test of normality based on the linear correlation between the observed order statistics and the rankits and has published a table of critical values. The 5% points of the distribution of r for $n = 10(10)100$ are shown below. You would reject the hypothesis of normality if the correlation is *less* than the critical value. Note than to accept normality we require a very high correlation coefficient.

n	10	20	30	40	50	60	70	80	90	100
r	.917	.950	.964	.972	.977	.980	.982	.984	.985	.987

The Filliben test is closely related to the Shapiro-Francia approximation to the Shapiro-Wilk test of normality. These tests are often used with standardized or jack-knifed residuals, although the fact that the residuals are correlated affects the significance levels to an unknown extent. For the program effort data in Figure 2.7 the Filliben correlation is a respectable 0.966. Since this value exceeds the critical value of 0.950 for 20 observations, we conclude that we have no evidence against the assumption of normally distributed residuals.

2.10 Transforming the Data

We now consider what to do if the regression diagnostics discussed in the previous section indicate that the model is not adequate. The usual solutions involve transforming the response, transforming the predictors, or both.

2.10.1 Transforming the Response

The response is often transformed to achieve linearity and homoscedasticity or constant variance. Examples of *variance stabilizing* transformations are the square root, which tends to work well for counts, and the arc-sine transformation, which is often appropriate when the response is a proportion. These two solutions have fallen out of fashion as generalized linear models designed specifically to deal with counts and proportions have increased in popularity. My recommendation in these two cases is to abandon the linear model in favor of better alternatives such as Poisson regression and logistic regression.

Transformations to achieve linearity, or *linearizing* transformations, are still useful. The most popular of them is the logarithm, which is specially useful when one expects effects to be proportional to the response. To fix ideas consider a model with a single predictor x, and suppose the response is expected to increase 100ρ percent for each point of increase in x. Suppose further that the error term, denoted U, is multiplicative. The model can then be written as

$$Y = \gamma(1+\rho)^x U.$$

Taking logs on both sides of the equation, we obtain a linear model for the transformed response

$$\log Y = \alpha + \beta x + \epsilon,$$

where the constant is $\alpha = \log\gamma$, the slope is $\beta = \log(1+\rho)$ and the error term is $\epsilon = \log U$. The usual assumption of normal errors is equivalent to assuming that U has a log-normal distribution. In this example taking logs has transformed a relatively complicated multiplicative model to a familiar linear form.

This development shows, incidentally, how to interpret the slope in a linear regression model when the response is in the log scale. Solving for ρ in terms of β, we see that a unit increase in x is associated with an increase of $100(e^\beta - 1)$ percent in y. If β is small, $e^\beta - 1 \approx \beta$, so the coefficient can be interpreted directly as a relative effect. For $|\beta| < 0.10$ the absolute error of the approximation is less than 0.005 or half a percentage point. Thus, a coefficient of 0.10 can be interpreted as a 10% effect on the response.

A general problem with transformations is that the two aims of achieving linearity and constant variance may be in conflict. In generalized linear models the two aims are separated more clearly, as we will see later in the sequel.

2.10.2 Box-Cox Transformations

Box and Cox (1964) have proposed a family of transformations that can be used with non-negative responses and which includes as special cases all the transformations in common use, including reciprocals, logarithms and square roots.

The basic idea is to work with the power transformation

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

and assume that $y^{(\lambda)}$ follows a normal linear model with parameters $\boldsymbol{\beta}$ and σ^2 for some value of λ. Note that this transformation is essentially y^λ for $\lambda \neq 0$ and $\log(y)$ for $\lambda = 0$, but has been scaled to be continuous at $\lambda = 0$. Useful values of λ are often found to be in the range $(-2, 2)$. Except for scaling factors, -1 is the reciprocal, 0 is the logarithm, 1/2 is the square root, 1 is the identity and 2 is the square.

Given a value of λ, we can estimate the linear model parameters $\boldsymbol{\beta}$ and σ^2 as usual, except that we work with the transformed response $y^{(\lambda)}$ instead of y. To select an appropriate transformation we need to try values of λ in a suitable range. Unfortunately, the resulting models cannot be compared in terms of their residual sums of squares because these are in different units. We therefore use a likelihood criterion.

Starting from the normal distribution of the transformed response $y^{(\lambda)}$, we can change variables to obtain the distribution of y. The resulting log-likelihood is

$$\log L(\boldsymbol{\beta}, \sigma^2, \lambda) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum (y_i^{(\lambda)} - \mu_i)^2/\sigma^2 + (\lambda - 1)\sum \log(y_i),$$

where the last term comes from the Jacobian of the transformation, which has derivative $y^{\lambda-1}$ for all λ. The other two terms are the usual normal likelihood, showing that we can estimate $\boldsymbol{\beta}$ and σ^2 for any fixed value of λ by regressing the transformed response $y^{(\lambda)}$ on the x's. Substituting the m.l.e.'s of $\boldsymbol{\beta}$ and σ^2 we obtain the concentrated or profile log-likelihood

$$\log L(\lambda) = c - \frac{n}{2}\log \mathrm{RSS}(y^{(\lambda)}) + (\lambda - 1)\sum \log(y_i),$$

where $c = \frac{n}{2}\log(2\pi/n) - \frac{n}{2}$ is a constant not involving λ.

Calculation of the profile log-likelihood can be simplified slightly by working with the alternative transformation

$$z^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda \tilde{y}^{\lambda-1}} & \lambda \neq 0 \\ \log(y)\tilde{y} & \lambda = 0, \end{cases}$$

where $\tilde{y}$ is the geometric mean of the original response, best calculated as $\tilde{y} = \exp(\sum \log(y_i)/n)$. The profile log-likelihood can then be written as

$$\log L(\lambda) = c - \frac{n}{2}\log \mathrm{RSS}(z^{(\lambda)}), \tag{2.30}$$

where $\mathrm{RSS}(z^{(\lambda)})$ is the RSS after regressing $z^{(\lambda)}$ on the x's. Using this alternative transformation the models for different values of λ can be compared directly in terms of their RSS's.

In practice we evaluate this profile log-likelihood for a range of possible values of λ. Rather than selecting the exact maximum, one often rounds to a value such as -1, 0, 1/2, 1 or 2, particularly if the profile log-likelihood is relatively flat around the maximum.

More formally, let $\hat{\lambda}$ denote the value that maximizes the profile likelihood. We can test the hypothesis H_0: $\lambda = \lambda_0$ for any fixed value λ_0 by calculating the likelihood ratio criterion

$$\chi^2 = 2(\log L(\hat{\lambda}) - \log L(\lambda_0)),$$

which has approximately in large samples a chi-squared distribution with one d.f. We can also define a likelihood-based confidence interval for λ as the set of values that would be a accepted by the above test, i.e. the set of values for which twice the log-likelihood is within $\chi^2_{1-\alpha,1}$ of twice the maximum log-likelihood. Identifying these values requires a numerical search procedure.

Box-Cox transformations are designed for non-negative responses, but can be applied to data that have occassional zero or negative values by adding a constant α to the response before applying the power transformation. Although α could be estimated, in practice one often uses a small value such as a half or one (depending, obviously, on the scale of the response).

Let us apply this procedure to the program effort data. Since two countries show no decline in the CBR, we add 0.5 to all responses before transforming them. Figure 2.8 shows the profile log-likelihood as a function of λ for values in the range $(-1, 2)$. Note that $\lambda = 1$ is not a bad choice, indicating that the model in the original scale is reasonable. A slightly better choice appears to be $\lambda = 0.5$, which is equivalent to using $\sqrt{y + 0.5}$ as the

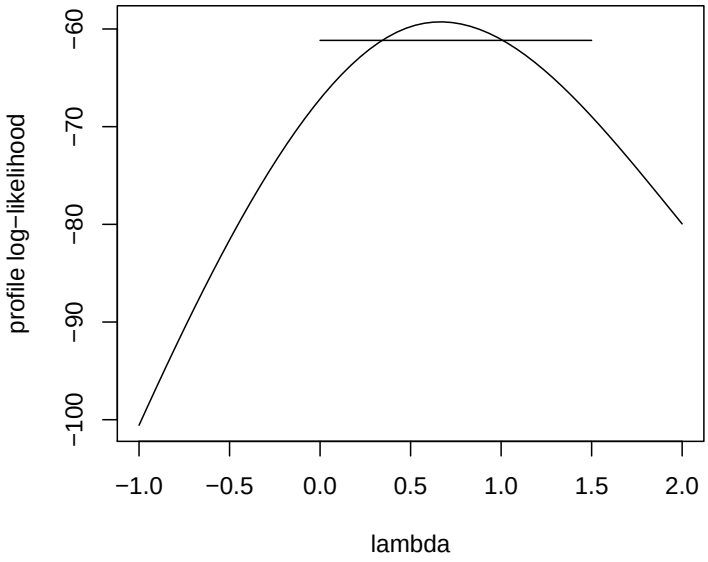


FIGURE 2.8: Profile Log-likelihood for Box-Cox Transformations for Ancova Model of CBR Decline by Setting and Effort

response. Fitting this model leads to small changes in the significance of the coefficients of setting and strong programs, but does not materially alter any of the conclusions.

More formally, we note that the profile log-likelihood for $\lambda = 1$ is -61.07. The maximum is attained at $\lambda = 0.67$ and is -59.25. Twice the difference between these values gives a chi-squared statistic of 3.65 on one degree of freedom, which is below the 5% critical value of 3.84. Thus, there is no evidence that we need to transform the response. A more detailed search shows that a 95% confidence interval for λ goes from 0.34 to 1.01. The horizontal line in Figure 2.8, at a height of -61.17, identifies the limits of the likelihood-based confidence interval.

2.10.3 The Atkinson Score

The Box-Cox procedure requires fitting a series of linear models, one for each trial value of λ. Atkinson (1985) has proposed a simpler procedure that gives

a quick indication of whether a transformation of the response is required at all. In practical terms, this technique involves adding to the model an auxiliary variable a defined as

$$a_i = y_i \left(\log(y_i/\tilde{y}) - 1\right), \tag{2.31}$$

where $\tilde{y}$ is the geometric mean of y, as in the previous subsection. Let γ denote the coefficient of a in the expanded model. If the estimate of γ is significant, then a Box-Cox transformation is indicated. A preliminary estimate of the value of λ is $1 - \hat{\gamma}$.

To see why this procedure works suppose the true model is

$$\mathbf{z}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where we have used the scale-independent version of the Box-Cox transformation. Expanding the left-hand-side using a first-order Taylor series approximation around $\lambda = 1$ gives

$$z^{(\lambda)} \approx z^{(1)} + (\lambda - 1) \left.\frac{dz^{(\lambda)}}{d\lambda}\right|_{\lambda=1}.$$

The derivative evaluated at $\lambda = 1$ is $a + \log \tilde{y} + 1$, where a is given by Equation 2.31. The second term does not depend on λ, so it can be absorbed into the constant. Note also that $z^{(1)} = y - 1$. Using these results we can rewrite the model as

$$\mathbf{y} \approx \mathbf{X}\boldsymbol{\beta} + (1 - \lambda)\mathbf{a} + \boldsymbol{\epsilon}.$$

Thus, to a first-order approximation the coefficient of the ancillary variable is $1 - \lambda$.

For the program effort data, adding the auxiliary variable a (calculated using CBR+1/2 to avoid taking the logarithm of zero) to the analysis of covariance model gives a coefficient of 0.59, suggesting a Box-Cox transformation with $\lambda = 0.41$. This value is reasonably close to the square root transformation suggested by the profile log-likelihood. The associated t-statistic is significant at the two percent level, but the more precise likelihood ratio criterion of the previous section, though borderline, was not significant. In conclusion, we do not have strong evidence of a need to transform the response.

2.10.4 Transforming the Predictors

The Atkinson procedure is similar in spirit to a procedure first suggested by Box and Tidwell (1962) to check whether one of the predictors needs to be

transformed. Specifically, to test whether one should use a transformation x^λ of a continuous predictor x, these authors suggest adding the auxiliary covariate

$$a_i = x_i \log(x_i)$$

to a model that already has x.

Let $\hat{\gamma}$ denote the estimated coefficient of the auxiliary variate $x \log(x)$ in the expanded model. This coefficient can be tested using the usual t statistic with $n - p$ d.f. If the test is significant, it indicates a need to transform the predictor. A preliminary estimate of the appropriate transformation is given by $\hat{\lambda} = \hat{\gamma}/\hat{\beta}+1$, where $\hat{\beta}$ is the estimated coefficient of x in the original model with x but not $x \log(x)$.

We can apply this technique to the program effort data by calculating a new variable equal to the product of setting and its logarithm, and adding it to the covariance analysis model with setting and effort. The estimated coefficient is -0.030 with a standard error of 0.728, so there is no need to transform setting. Note, incidentally, that the effect of setting is not significant in this model.

Chapter 3

Logit Models for Binary Data

We now turn our attention to regression models for dichotomous data, including logistic regression and probit analysis. These models are appropriate when the response takes one of only two possible values representing success and failure, or more generally the presence or absence of an attribute of interest.

3.1 Introduction to Logistic Regression

We start by introducing an example that will be used to illustrate the analysis of binary data. We then discuss the stochastic structure of the data in terms of the Bernoulli and binomial distributions, and the systematic structure in terms of the logit transformation. The result is a generalized linear model with binomial response and link logit.

3.1.1 The Contraceptive Use Data

Table 3.1, adapted from Little (1978), shows the distribution of 1607 currently married and fecund women interviewed in the Fiji Fertility Survey of 1975, classified by current age, level of education, desire for more children, and contraceptive use.

In our analysis of these data we will view current use of contraception as the response or dependent variable of interest and age, education and desire for more children as predictors. Note that the response has two categories: use and non-use. In this example all predictors are treated as categorical

TABLE 3.1: Current Use of Contraception Among Married Women by Age, Education and Desire for More Children Fiji Fertility Survey, 1975

Age	Education	Desires More Children?	Contraceptive Use		Total
			No	Yes	
<25	Lower	Yes	53	6	59
		No	10	4	14
	Upper	Yes	212	52	264
		No	50	10	60
25–29	Lower	Yes	60	14	74
		No	19	10	29
	Upper	Yes	155	54	209
		No	65	27	92
30–39	Lower	Yes	112	33	145
		No	77	80	157
	Upper	Yes	118	46	164
		No	68	78	146
40–49	Lower	Yes	35	6	41
		No	46	48	94
	Upper	Yes	8	8	16
		No	12	31	43
Total			1100	507	1607

variables, but the techniques to be studied can be applied more generally to both discrete factors and continuous variates.

The original dataset includes the date of birth of the respondent and the date of interview in month/year form, so it is possible to calculate age in single years, but we will use ten-year age groups for convenience. Similarly, the survey included information on the highest level of education attained and the number of years completed at that level, so one could calculate completed years of education, but we will work here with a simple distinction between lower primary or less and upper primary or more. Finally, desire for more children is measured as a simple dichotomy coded yes or no, and therefore is naturally a categorical variate.

The fact that we treat all predictors as discrete factors allows us to summarize the data in terms of the numbers using and not using contraception in each of sixteen different groups defined by combinations of values of the pre-

dictors. For models involving discrete factors we can obtain exactly the same results working with grouped data or with individual data, but grouping is convenient because it leads to smaller datasets. If we were to incorporate continuous predictors into the model we would need to work with the original 1607 observations. Alternatively, it might be possible to group cases with identical covariate patterns, but the resulting dataset may not be much smaller than the original one.

The basic aim of our analysis will be to describe the way in which contraceptive use varies by age, education and desire for more children. An example of the type of research question that we will consider is the extent to which the association between education and contraceptive use is affected by the fact that women with upper primary or higher education are younger and tend to prefer smaller families than women with lower primary education or less.

3.1.2 The Binomial Distribution

We consider first the case where the response y_i is binary, assuming only two values that for convenience we code as one or zero. For example, we could define

$$y_i = \begin{cases} 1 & \text{if the } i\text{-th woman is using contraception} \\ 0 & \text{otherwise.} \end{cases}$$

We view y_i as a realization of a random variable Y_i that can take the values one and zero with probabilities π_i and $1-\pi_i$, respectively. The distribution of Y_i is called a *Bernoulli* distribution with parameter π_i, and can be written in compact form as

$$\Pr\{Y_i = y_i\} = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \tag{3.1}$$

for $y_i = 0, 1$. Note that if $y_i = 1$ we obtain π_i, and if $y_i = 0$ we obtain $1-\pi_i$.

It is fairly easy to verify by direct calculation that the expected value and variance of Y_i are

$$\begin{aligned} \mathrm{E}(Y_i) &= \mu_i = \pi_i, \text{ and} \\ \mathrm{var}(Y_i) &= \sigma_i^2 = \pi_i(1-\pi_i). \end{aligned} \tag{3.2}$$

Note that the mean and variance depend on the underlying probability π_i. Any factor that affects the probability will alter not just the mean but also the variance of the observations. This suggest that a linear model that allows

the predictors to affect the mean but assumes that the variance is constant will not be adequate for the analysis of binary data.

Suppose now that the units under study can be classified according to the factors of interest into k groups in such a way that all individuals in a group have identical values of all covariates. In our example, women may be classified into 16 different groups in terms of their age, education and desire for more children. Let n_i denote the number of observations in group i, and let y_i denote the number of units who have the attribute of interest in group i. For example, let

$$y_i = \text{number of women using contraception in group } i.$$

We view y_i as a realization of a random variable Y_i that takes the values $0, 1, \ldots, n_i$. If the n_i observations in each group are *independent*, and they all have the same probability π_i of having the attribute of interest, then the distribution of Y_i is *binomial* with parameters π_i and n_i, which we write

$$Y_i \sim B(n_i, \pi_i).$$

The probability distribution function of Y_i is given by

$$\Pr\{Y_i = y_i\} = \left(\begin{array}{c} n_i \\ y_i \end{array} \right) \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \tag{3.3}$$

for $y_i = 0, 1, \ldots, n_i$. Here $\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}$ is the probability of obtaining y_i successes and $n_i - y_i$ failures in some specific order, and the combinatorial coefficient is the number of ways of obtaining y_i successes in n_i trials.

The mean and variance of Y_i can be shown to be

$$\begin{aligned} \mathrm{E}(Y_i) &= \mu_i = n_i \pi_i, \text{ and} \\ \mathrm{var}(Y_i) &= \sigma_i^2 = n_i \pi_i (1-\pi_i). \end{aligned} \tag{3.4}$$

The easiest way to obtain this result is as follows. Let Y_{ij} be an indicator variable that takes the values one or zero if the j-th unit in group i is a success or a failure, respectively. Note that Y_{ij} is a Bernoulli random variable with mean and variance as given in Equation 3.2. We can write the number of successes Y_i in group i as a sum of the individual indicator variables, so $Y_i = \sum_j Y_{ij}$. The mean of Y_i is then the sum of the individual means, and by independence, its variance is the sum of the individual variances, leading to the result in Equation 3.4. Note again that the mean and variance depend

on the underlying probability π_i. Any factor that affects this probability will affect both the mean and the variance of the observations.

From a mathematical point of view the grouped data formulation given here is the most general one; it includes individual data as the special case where we have n groups of size one, so $k = n$ and $n_i = 1$ for all i. It also includes as a special case the other extreme where the underlying probability is the same for all individuals and we have a single group, with $k = 1$ and $n_1 = n$. Thus, all we need to consider in terms of estimation and testing is the binomial distribution.

From a practical point of view it is important to note that if the predictors are discrete factors and the outcomes are independent, we can use the Bernoulli distribution for the individual zero-one data or the binomial distribution for grouped data consisting of counts of successes in each group. The two approaches are equivalent, in the sense that they lead to exactly the same likelihood function and therefore the same estimates and standard errors. Working with grouped data when it is possible has the additional advantage that, depending on the size of the groups, it becomes possible to test the goodness of fit of the model. In terms of our example we can work with 16 groups of women (or fewer when we ignore some of the predictors) and obtain exactly the same estimates as we would if we worked with the 1607 individuals.

In Appendix B we show that the binomial distribution belongs to Nelder and Wedderburn's (1972) exponential family, so it fits in our general theoretical framework.

3.1.3 The Logit Transformation

The next step in defining a model for our data concerns the systematic structure. We would like to have the probabilities π_i depend on a vector of observed covariates $\mathbf{x}_i$. The simplest idea would be to let π_i be a linear function of the covariates, say

$$\pi_i = \mathbf{x}_i'\boldsymbol{\beta}, \tag{3.5}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients. Model 3.5 is sometimes called the *linear probability model*. This model is often estimated from individual data using ordinary least squares (OLS).

One problem with this model is that the probability π_i on the left-hand-side has to be between zero and one, but the linear predictor $\mathbf{x}_i'\boldsymbol{\beta}$ on the right-hand-side can take any real value, so there is no guarantee that the

predicted values will be in the correct range unless complex restrictions are imposed on the coefficients.

A simple solution to this problem is to *transform* the probability to remove the range restrictions, and model the transformation as a linear function of the covariates. We do this in two steps.

First, we move from the probability π_i to the *odds*

$$\text{odds}_i = \frac{\pi_i}{1-\pi_i},$$

defined as the ratio of the probability to its complement, or the ratio of favorable to unfavorable cases. If the probability of an event is a half, the odds are one-to-one or even. If the probability is 1/3, the odds are one-to-two. If the probability is very small, the odds are said to be long. In some contexts the language of odds is more natural than the language of probabilities. In gambling, for example, odds of $1:k$ indicate that the fair payoff for a stake of one is k. The key from our point of view is that the languages are equivalent, i.e. one can easily be translated into the other, but odds can take any positive value and therefore have no ceiling restriction.

Second, we take logarithms, calculating the *logit* or log-odds

$$\eta_i = \text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i}, \tag{3.6}$$

which has the effect of removing the floor restriction. To see this point note that as the probability goes down to zero the odds approach zero and the logit approaches $-\infty$. At the other extreme, as the probability approaches one the odds approach $+\infty$ and so does the logit. Thus, logits map probabilities from the range $(0, 1)$ to the entire real line. Note that if the probability is 1/2 the odds are even and the logit is zero. Negative logits represent probabilities below one half and positive logits correspond to probabilities above one half. Figure 3.1 illustrates the logit transformation.

Logits may also be defined in terms of the binomial mean $\mu_i = n_i\pi_i$ as the log of the ratio of expected successes μ_i to expected failures $n_i - \mu_i$. The result is exactly the same because the binomial denominator n_i cancels out when calculating the odds.

In the contraceptive use data there are 507 users of contraception among 1607 women, so we estimate the probability as $507/1607 = 0.316$. The odds are 507/1100 or 0.461 to one, so non-users outnumber users roughly two to one. The logit is $\log(0.461) = -0.775$.

The logit transformation is one-to-one. The inverse transformation is sometimes called the *antilogit*, and allows us to go back from logits to prob-

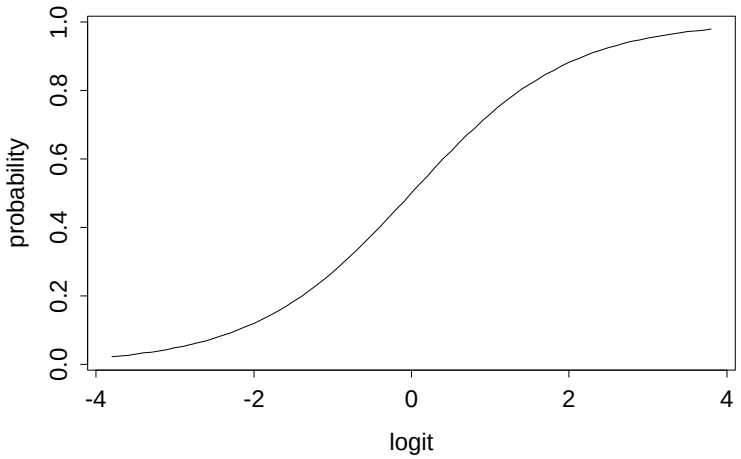


FIGURE 3.1: The Logit Transformation

abilities. Solving for π_i in Equation 3.6 gives

$$\pi_i = \text{logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}. \tag{3.7}$$

In the contraceptive use data the estimated logit was -0.775. Exponentiating this value we obtain odds of $\exp(-0.775) = 0.461$ and from this we obtain a probability of $0.461/(1 + 0.461) = 0.316$.

We are now in a position to define the logistic regression model, by assuming that the *logit* of the probability π_i, rather than the probability itself, follows a linear model.

3.1.4 The Logistic Regression Model

Suppose that we have k independent observations $y_1, \ldots, y_k$, and that the i-th observation can be treated as a realization of a random variable Y_i. We assume that Y_i has a binomial distribution

$$Y_i \sim B(n_i, \pi_i) \tag{3.8}$$

with binomial denominator n_i and probability π_i. With individual data $n_i = 1$ for all i. This defines the stochastic structure of the model.

Suppose further that the *logit* of the underlying probability π_i is a linear function of the predictors

$$\text{logit}(\pi_i) = \mathbf{x}_i'\boldsymbol{\beta}, \tag{3.9}$$

where $\mathbf{x}_i$ is a vector of covariates and $\boldsymbol{\beta}$ is a vector of regression coefficients. This defines the systematic structure of the model.

The model defined in Equations 3.8 and 3.9 is a generalized linear model with binomial response and link logit. Note, incidentally, that it is more natural to consider the distribution of the response Y_i than the distribution of the implied error term $Y_i - \mu_i$.

The regression coefficients $\boldsymbol{\beta}$ can be interpreted along the same lines as in linear models, bearing in mind that the left-hand-side is a logit rather than a mean. Thus, β_j represents the change in the *logit* of the probability associated with a unit change in the j-th predictor holding all other predictors constant. While expressing results in the logit scale will be unfamiliar at first, it has the advantage that the model is rather simple in this particular scale.

Exponentiating Equation 3.9 we find that the odds for the i-th unit are given by

$$\frac{\pi_i}{1-\pi_i} = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}. \tag{3.10}$$

This expression defines a multiplicative model for the odds. For example if we were to change the j-th predictor by one unit while holding all other variables constant, we would multiply the odds by $\exp\{\beta_j\}$. To see this point suppose the linear predictor is $\mathbf{x}_i'\boldsymbol{\beta}$ and we increase x_j by one, to obtain $\mathbf{x}_i'\boldsymbol{\beta} + \beta_j$. Exponentiating we get $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ times $\exp\{\beta_j\}$. Thus, the exponentiated coefficient $\exp\{\beta_j\}$ represents an *odds ratio*. Translating the results into multiplicative effects on the odds, or odds ratios, is often helpful, because we can deal with a more familiar scale while retaining a relatively simple model.

Solving for the probability π_i in the logit model in Equation 3.9 gives the more complicated model

$$\pi_i = \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}. \tag{3.11}$$

While the left-hand-side is in the familiar probability scale, the right-hand-side is a non-linear function of the predictors, and there is no simple way to express the effect on the probability of increasing a predictor by one unit while holding the other variables constant. We can obtain an approximate answer by taking derivatives with respect to x_j, which of course makes sense only for continuous predictors. Using the quotient rule we get

$$\frac{\mathrm{d}\pi_i}{\mathrm{d}x_{ij}} = \beta_j \pi_i (1-\pi_i).$$

Thus, the effect of the j-th predictor on the probability π_i depends on the coefficient β_j *and* the value of the probability. Analysts sometimes evaluate this product setting π_i to the sample mean (the proportion of cases with the attribute of interest in the sample). The result approximates the effect of the covariate near the mean of the response.

In the examples that follow we will emphasize working directly in the logit scale, but we will often translate effects into odds ratios to help in interpretation.

Before we leave this topic it may be worth considering the linear probability model of Equation 3.5 one more time. In addition to the fact that the linear predictor $\mathbf{x}_i'\boldsymbol{\beta}$ may yield values outside the $(0, 1)$ range, one should consider whether it is reasonable to assume linear effects on a probability scale that is subject to floor and ceiling effects. An incentive, for example, may increase the probability of taking an action by ten percentage points when the probability is a half, but couldn't possibly have that effect if the baseline probability was 0.95. This suggests that the assumption of a linear effect across the board may not be reasonable.

In contrast, suppose the effect of the incentive is 0.4 in the logit scale, which is equivalent to approximately a 50% increase in the odds of taking the action. If the original probability is a half the logit is zero, and adding 0.4 to the logit gives a probability of 0.6, so the effect is ten percentage points, just as before. If the original probability is 0.95, however, the logit is almost three, and adding 0.4 in the logit scale gives a probability of 0.97, an effect of just two percentage points. An effect that is constant in the logit scale translates into varying effects on the probability scale, adjusting automatically as one approaches the floor of zero or the ceiling of one. This feature of the transformation is clearly seen from Figure 3.1.

3.2 Estimation and Hypothesis Testing

The logistic regression model just developed is a generalized linear model with binomial errors and link logit. We can therefore rely on the general theory developed in Appendix B to obtain estimates of the parameters and to test hypotheses. In this section we summarize the most important results needed in the applications.

3.2.1 Maximum Likelihood Estimation

Although you will probably use a statistical package to compute the estimates, here is a brief description of the underlying procedure. The likelihood

function for n independent binomial observations is a product of densities given by Equation 3.3. Taking logs we find that, except for a constant involving the combinatorial terms, the log-likelihood function is

$$\log L(\boldsymbol{\beta}) = \sum \{y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i)\},$$

where π_i depends on the covariates $\mathbf{x}_i$ and a vector of p parameters $\boldsymbol{\beta}$ through the logit transformation of Equation 3.9.

At this point we could take first and expected second derivatives to obtain the score and information matrix and develop a Fisher scoring procedure for maximizing the log-likelihood. As shown in Appendix B, the procedure is equivalent to iteratively re-weighted least squares (IRLS). Given a current estimate $\hat{\boldsymbol{\beta}}$ of the parameters, we calculate the linear predictor $\hat{\boldsymbol{\eta}} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ and the fitted values $\hat{\boldsymbol{\mu}} = \text{logit}^{-1}(\boldsymbol{\eta})$. With these values we calculate the working dependent variable $\mathbf{z}$, which has elements

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i(n_i - \hat{\mu}_i)} n_i,$$

where n_i are the binomial denominators. We then regress $\mathbf{z}$ on the covariates calculating the weighted least squares estimate

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z},$$

where $\mathbf{W}$ is a diagonal matrix of weights with entries

$$w_{ii} = \hat{\mu}_i(n_i - \hat{\mu}_i)/n_i.$$

(You may be interested to know that the weight is inversely proportional to the estimated variance of the working dependent variable.) The resulting estimate of $\boldsymbol{\beta}$ is used to obtain improved fitted values and the procedure is iterated to convergence.

Suitable initial values can be obtained by applying the link to the data. To avoid problems with counts of 0 or n_i (which is always the case with individual zero-one data), we calculate empirical logits adding $1/2$ to both the numerator and denominator, i.e. we calculate

$$z_i = \log \frac{y_i + 1/2}{n_i - y_i + 1/2},$$

and then regress this quantity on $\mathbf{x}_i$ to obtain an initial estimate of $\boldsymbol{\beta}$.

The resulting estimate is consistent and its large-sample variance is given by

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \tag{3.12}$$

where $\mathbf{W}$ is the matrix of weights evaluated in the last iteration.

Alternatives to maximum likelihood estimation include weighted least squares, which can be used with grouped data, and a method that minimizes Pearson's chi-squared statistic, which can be used with both grouped and individual data. We will not consider these alternatives further.

3.2.2 Goodness of Fit Statistics

Suppose we have just fitted a model and want to assess how well it fits the data. A measure of discrepancy between observed and fitted values is the *deviance* statistic, which is given by

$$D = 2\sum\{y_i \log(\frac{y_i}{\hat{\mu}_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{\mu}_i})\}, \tag{3.13}$$

where y_i is the observed and $\hat{\mu}_i$ is the fitted value for the i-th observation. Note that this statistic is twice a sum of 'observed times log of observed over expected', where the sum is over both successes and failures (i.e. we compare both y_i and $n_i - y_i$ with their expected values). In a perfect fit the ratio observed over expected is one and its logarithm is zero, so the deviance is zero.

In Appendix B we show that this statistic may be constructed as a likelihood ratio test that compares the model of interest with a saturated model that has one parameter for each observation.

With grouped data, the distribution of the deviance statistic as the group sizes $n_i \to \infty$ for all i, converges to a chi-squared distribution with $n - p$ d.f., where n is the number of *groups* and p is the number of parameters in the model, including the constant. Thus, for reasonably large groups, the deviance provides a goodness of fit test for the model. With individual data the distribution of the deviance does not converge to a chi-squared (or any other known) distribution, and cannot be used as a goodness of fit test. We will, however, consider other diagnostic tools that can be used with individual data.

An alternative measure of goodness of fit is *Pearson's chi-squared statistic*, which for binomial data can be written as

$$\chi^2_P = \sum_i \frac{n_i(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(n_i - \hat{\mu}_i)}. \tag{3.14}$$

Note that each term in the sum is the squared difference between observed and fitted values y_i and $\hat{\mu}_i$, divided by the variance of y_i, which is $\mu_i(n_i -$

$\mu_i)/n_i$, estimated using $\hat{\mu}_i$ for μ_i. This statistic can also be derived as a sum of 'observed minus expected squared over expected', where the sum is over both successes and failures.

With grouped data Pearson's statistic has approximately in large samples a chi-squared distribution with $n-p$ d.f., and is asymptotically equivalent to the deviance or likelihood-ratio chi-squared statistic. The statistic can not be used as a goodness of fit test with individual data, but provides a basis for calculating residuals, as we shall see when we discuss logistic regression diagnostics.

3.2.3 Tests of Hypotheses

Let us consider the problem of testing hypotheses in logit models. As usual, we can calculate Wald tests based on the large-sample distribution of the m.l.e., which is approximately normal with mean $\boldsymbol{\beta}$ and variance-covariance matrix as given in Equation 3.12.

In particular, we can test the hypothesis

$$H_0 : \beta_j = 0$$

concerning the significance of a single coefficient by calculating the ratio of the estimate to its standard error

$$z = \frac{\hat{\beta}_j}{\sqrt{\hat{\text{var}}(\hat{\beta}_j)}}.$$

This statistic has approximately a standard normal distribution in large samples. Alternatively, we can treat the square of this statistic as approximately a chi-squared with one d.f.

The Wald test can be use to calculate a confidence interval for β_j. We can assert with $100(1-\alpha)\%$ confidence that the true parameter lies in the interval with boundaries

$$\hat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\hat{\text{var}}(\hat{\beta}_j)},$$

where $z_{1-\alpha/2}$ is the normal critical value for a two-sided test of size α. Confidence intervals for effects in the logit scale can be translated into confidence intervals for odds ratios by exponentiating the boundaries.

The Wald test can be applied to tests hypotheses concerning several coefficients by calculating the usual quadratic form. This test can also be inverted to obtain confidence regions for vector-value parameters, but we will not consider this extension.

For more general problems we consider the likelihood ratio test. A key to construct these tests is the deviance statistic introduced in the previous subsection. In a nutshell, the likelihood ratio test to compare two nested models is based on the *difference* between their deviances.

To fix ideas, consider partitioning the model matrix and the vector of coefficients into two components

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2) \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$$

with p_1 and p_2 elements, respectively. Consider testing the hypothesis

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0},$$

that the variables in $\mathbf{X}_2$ have no effect on the response, i.e. the joint significance of the coefficients in $\boldsymbol{\beta}_2$.

Let $D(\mathbf{X}_1)$ denote the deviance of a model that includes only the variables in $\mathbf{X}_1$ and let $D(\mathbf{X}_1 + \mathbf{X}_2)$ denote the deviance of a model that includes all variables in $\mathbf{X}$. Then the difference

$$\chi^2 = D(\mathbf{X}_1) - D(\mathbf{X}_1 + \mathbf{X}_2)$$

has approximately in large samples a chi-squared distribution with p_2 d.f. Note that p_2 is the difference in the number of parameters between the two models being compared.

The deviance plays a role similar to the residual sum of squares. In fact, in Appendix B we show that in models for normally distributed data the deviance *is* the residual sum of squares. Likelihood ratio tests in generalized linear models are based on scaled deviances, obtained by dividing the deviance by a scale factor. In linear models the scale factor was σ^2, and we had to divide the RSS's (or their difference) by an estimate of σ^2 in order to calculate the test criterion. With binomial data the scale factor is one, and there is no need to scale the deviances.

The Pearson chi-squared statistic in the previous subsection, while providing an alternative goodness of fit test for grouped data, cannot be used in general to compare nested models. In particular, differences in deviances have chi-squared distributions but differences in Pearson chi-squared statistics do not. This is the main reason why in statistical modelling we use the deviance or likelihood ratio chi-squared statistic rather than the more traditional Pearson chi-squared of elementary statistics.

3.3 The Comparison of Two Groups

We start our applications of logit regression with the simplest possible example: a two by two table. We study a binary outcome in two groups, and introduce the odds ratio and the logit analogue of the two-sample t test.

3.3.1 A 2-by-2 Table

We will use the contraceptive use data classified by desire for more children, as summarized in Table 3.2

TABLE 3.2: Contraceptive Use by Desire for More Children

Desires	Using	Not Using	All
i	y_i	$n_i - y_i$	n_i
Yes	219	753	972
No	288	347	635
All	507	1100	1607

We treat the counts of users y_i as realizations of independent random variables Y_i having binomial distributions $B(n_i, \pi_i)$ for $i = 1, 2$, and consider models for the logits of the probabilities.

3.3.2 Testing Homogeneity

There are only two possible models we can entertain for these data. The first one is the *null* model. This model assumes homogeneity, so the two groups have the same probability and therefore the same logit

$$\text{logit}(\pi_i) = \eta.$$

The m.l.e. of the common logit is -0.775, which happens to be the logit of the sample proportion $507/1607 = 0.316$. The standard error of the estimate is 0.054. This value can be used to obtain an approximate 95% confidence limit for the logit with boundaries $(-0.880, -0.669)$. Calculating the antilogit of these values, we obtain a 95% confidence interval for the overall probability of using contraception of $(0.293, 0.339)$.

The deviance for the null model happens to be 91.7 on one d.f. (two groups minus one parameter). This value is highly significant, indicating that this model does not fit the data, i.e. the two groups classified by desire for more children do not have the same probability of using contraception.

The value of the deviance is easily verified by hand. The estimated probability of 0.316, applied to the sample sizes in Table 3.2, leads us to expect 306.7 and 200.3 users of contraception in the two groups, and therefore 665.3 and 434.7 non-users . Comparing the observed and expected numbers of users and non-users in the two groups using Equation 3.13 gives 91.7.

You can also compare the observed and expected frequencies using Pearson's chi-squared statistic from Equation 3.14. The result is 92.6 on one d.f., and provides an alternative test of the goodness of fit of the null model.

3.3.3 The Odds Ratio

The other model that we can entertain for the two-by-two table is the *one-factor* model, where we write

$$\text{logit}(\pi_i) = \eta + \alpha_i,$$

where η is an overall logit and α_i is the effect of group i on the logit. Just as in the one-way anova model, we need to introduce a restriction to identify this model. We use the reference cell method, and set $\alpha_1 = 0$. The model can then be written

$$\text{logit}(\pi_i) = \begin{cases} \eta & i = 1 \\ \eta + \alpha_2 & i = 2 \end{cases}$$

so that η becomes the logit of the reference cell, and α_2 is the effect of level two of the factor compared to level one, or more simply the difference in logits between level two and the reference cell. Table 3.3 shows parameter estimates and standard errors for this model.

TABLE 3.3: Parameter Estimates for Logit Model of Contraceptive Use by Desire for More Children

Parameter	Symbol	Estimate	Std. Error	z-ratio
Constant	η	-1.235	0.077	-16.09
Desire	α_2	1.049	0.111	9.48

The estimate of η is, as you might expect, the logit of the observed proportion using contraception among women who desire more children, $\text{logit}(219/972) = -1.235$. The estimate of α_2 is the difference between the logits of the two groups, $\text{logit}(288/635) - \text{logit}(219/972) = 1.049$.

Exponentiating the additive logit model we obtain a multiplicative model for the odds:

$$\frac{\pi_i}{1-\pi_i} = \begin{cases} e^{\eta} & i = 1 \\ e^{\eta}e^{\alpha_2} & i = 2 \end{cases}$$

so that e^{η} becomes the odds for the reference cell and e^{α_2} is the ratio of the odds for level 2 of the factor to the odds for the reference cell. Not surprisingly, e^{α_2} is called the *odds ratio*.

In our example, the effect of 1.049 in the logit scale translates into an odds ratio of 2.85. Thus, the odds of using contraception among women who want no more children are nearly three times as high as the odds for women who desire more children.

From the estimated logit effect of 1.049 and its standard error we can calculate a 95% confidence interval with boundaries $(0.831, 1.267)$. Exponentiating these boundaries we obtain a 95% confidence interval for the odds ratio of $(2.30, 3.55)$. Thus, we conclude with 95% confidence that the odds of using contraception among women who want no more children are between two and three-and-a-half times the corresponding odds for women who want more children.

The estimate of the odds ratio can be calculated directly as the cross-product of the frequencies in the two-by-two table. If we let f_{ij} denote the frequency in cell i, j then the estimated odds ratio is

$$\frac{f_{11}f_{22}}{f_{12}f_{21}}.$$

The deviance of this model is zero, because the model is saturated: it has two parameters to represent two groups, so it has to do a perfect job. The reduction in deviance of 91.7 from the null model down to zero can be interpreted as a test of

$$H_0 : \alpha_2 = 0,$$

the significance of the effect of desire for more children.

An alternative test of this effect is obtained from the m.l.e of 1.049 and its standard error of 0.111, and gives a z-ratio of 9.47. Squaring this value we obtain a chi-squared of 89.8 on one d.f. Note that the Wald test is similar, but not identical, to the likelihood ratio test. Recall that in linear models the two tests were identical. In logit models they are only asymptotically equivalent.

The logit of the observed proportion $p_i = y_i/n_i$ has large-sample variance

$$\text{var}(\text{logit}(p_i)) = \frac{1}{\mu_i} + \frac{1}{n_i - \mu_i},$$

which can be estimated using y_i to estimate μ_i for $i = 1, 2$. Since the two groups are independent samples, the variance of the difference in logits is the sum of the individual variances. You may use these results to verify the Wald test given above.

3.3.4 The Conventional Analysis

It might be instructive to compare the results obtained here with the conventional analysis of this type of data, which focuses on the sample proportions and their difference. In our example, the proportions using contraception are 0.225 among women who want another child and 0.453 among those who do not. The difference of 0.228 has a standard error of 0.024 (calculated using the pooled estimate of the proportion). The corresponding z-ratio is 9.62 and is equivalent to a chi-squared of 92.6 on one d.f.

Note that the result coincides with the Pearson chi-squared statistic testing the goodness of fit of the null model. In fact, Pearson's chi-squared and the conventional test for equality of two proportions are one and the same.

In the case of two samples it is debatable whether the group effect is best measured in terms of a difference in probabilities, the odds-ratio, or even some other measures such as the relative difference proposed by Sheps (1961). For arguments on all sides of this issue see Fleiss (1973).

3.4 The Comparison of Several Groups

Let us take a more general look at logistic regression models with a single predictor by considering the comparison of k groups. This will help us illustrate the logit analogues of one-way analysis of variance and simple linear regression models.

3.4.1 A k-by-Two Table

Consider a cross-tabulation of contraceptive use by age, as summarized in Table 3.4. The structure of the data is the same as in the previous section, except that we now have four groups rather than two.

The analysis of this table proceeds along the same lines as in the two-by-two case. The null model yields exactly the same estimate of the overall logit and its standard error as before. The deviance, however, is now 79.2 on three d.f. This value is highly significant, indicating that the assumption of a common probability of using contraception for the four age groups is not tenable.

TABLE 3.4: Contraceptive Use by Age

Age i	Using y_i	Not Using $n_i - y_i$	Total n_i
<25	72	325	397
25–29	105	299	404
30–39	237	375	612
40–49	93	101	194
Total	507	1100	1607

3.4.2 The One-Factor Model

Consider now a one-factor model, where we allow each group or level of the discrete factor to have its own logit. We write the model as

$$\text{logit}(\pi_i) = \eta + \alpha_i.$$

To avoid redundancy we adopt the reference cell method and set $\alpha_1 = 0$, as before. Then η is the logit of the reference group, and α_i measures the difference in logits between level i of the factor and the reference level. This model is exactly analogous to an analysis of variance model. The model matrix $\mathbf{X}$ consists of a column of ones representing the constant and $k-1$ columns of dummy variables representing levels two to k of the factor.

Fitting this model to Table 3.4 leads to the parameter estimates and standard errors in Table 3.5. The deviance for this model is of course zero because the model is saturated: it uses four parameters to model four groups.

TABLE 3.5: Estimates and Standard Errors for Logit Model of Contraceptive Use by Age in Groups

Parameter		Symbol	Estimate	Std. Error	z-ratio
Constant		η	−1.507	0.130	−11.57
Age	25–29	α_2	0.461	0.173	2.67
	30–39	α_3	1.048	0.154	6.79
	40–49	α_4	1.425	0.194	7.35

The baseline logit of −1.51 for women under age 25 corresponds to odds of 0.22. Exponentiating the age coefficients we obtain odds ratios of 1.59, 2.85 and 4.16. Thus, the odds of using contraception increase by 59% and

185% as we move to ages 25–29 and 30–39, and are quadrupled for ages 40–49, all compared to women under age 25.

All of these estimates can be obtained directly from the frequencies in Table 3.4 in terms of the logits of the observed proportions. For example the constant is $\text{logit}(72/397) = -1.507$, and the effect for women 25–29 is logit(105/404) minus the constant.

To test the hypothesis of no age effects we can compare this model with the null model. Since the present model is saturated, the difference in deviances is exactly the same as the deviance of the null model, which was 79.2 on three d.f. and is highly significant. An alternative test of

$$H_0 : \alpha_2 = \alpha_3 = \alpha_4 = 0$$

is based on the estimates and their variance-covariance matrix. Let $\boldsymbol{\alpha} = (\alpha_2, \alpha_3, \alpha_4)'$. Then

$$\hat{\boldsymbol{\alpha}} = \begin{pmatrix} 0.461 \\ 1.048 \\ 1.425 \end{pmatrix} \qquad \text{and} \qquad \hat{\text{var}}(\hat{\boldsymbol{\alpha}}) = \begin{pmatrix} 0.030 & 0.017 & 0.017 \\ 0.017 & 0.024 & 0.017 \\ 0.017 & 0.017 & 0.038 \end{pmatrix},$$

and the Wald statistic is

$$W = \hat{\boldsymbol{\alpha}}' \, \hat{\text{var}}^{-1}(\hat{\boldsymbol{\alpha}}) \, \hat{\boldsymbol{\alpha}} = 74.4$$

on three d.f. Again, the Wald test gives results similar to the likelihood ratio test.

3.4.3 A One-Variate Model

Note that the estimated logits in Table 3.5 (and therefore the odds and probabilities) increase monotonically with age. In fact, the logits seem to increase by approximately the same amount as we move from one age group to the next. This suggests that the effect of age may actually be linear in the logit scale.

To explore this idea we treat age as a variate rather than a factor. A thorough exploration would use the individual data with age in single years (or equivalently, a 35 by two table of contraceptive use by age in single years from 15 to 49). However, we can obtain a quick idea of whether the model would be adequate by keeping age grouped into four categories but representing these by the *mid-points* of the age groups. We therefore consider a model analogous to simple linear regression, where

$$\text{logit}(\pi_i) = \alpha + \beta x_i,$$

where x_i takes the values 20, 27.5, 35 and 45, respectively, for the four age groups. This model fits into our general framework, and corresponds to the special case where the model matrix $\mathbf{X}$ has two columns, a column of ones representing the constant and a column with the mid-points of the age groups, representing the linear effect of age.

Fitting this model gives a deviance of 2.40 on two d.f. , which indicates a very good fit. The parameter estimates and standard errors are shown in Table 3.6. Incidentally, there is no explicit formula for the estimates of the constant and slope in this model, so we must rely on iterative procedures to obtain the estimates.

TABLE 3.6: Estimates and Standard Errors for Logit Model of Contraceptive Use with a Linear Effect of Age

Parameter	Symbol	Estimate	Std. Error	z-ratio
Constant	α	-2.673	0.233	-11.46
Age (linear)	β	0.061	0.007	8.54

The slope indicates that the logit of the probability of using contraception increases 0.061 for every year of age. Exponentiating this value we note that the odds of using contraception are multiplied by 1.063—that is, increase 6.3%—for every year of age. Note, by the way, that $e^{\beta} \approx 1 + \beta$ for small $|\beta|$. Thus, when the logit coefficient is small in magnitude, 100β provides a quick approximation to the percent change in the odds associated with a unit change in the predictor. In this example the effect is 6.3% and the approximation is 6.1%.

To test the significance of the slope we can use the Wald test, which gives a z statistic of 8.54 or equivalently a chi-squared of 73.9 on one d.f. Alternatively, we can construct a likelihood ratio test by comparing this model with the null model. The difference in deviances is 76.8 on one d.f. Comparing these results with those in the previous subsection shows that we have captured most of the age effect using a single degree of freedom.

Adding the estimated constant to the product of the slope by the mid-points of the age groups gives estimated logits at each age, and these may be compared with the logits of the observed proportions using contraception. The results of this exercise appear in Figure 3.2. The visual impression of the graph confirms that the fit is quite good. In this example the assumption of linear effects on the logit scale leads to a simple and parsimonious model. It would probably be worthwhile to re-estimate this model using the individual

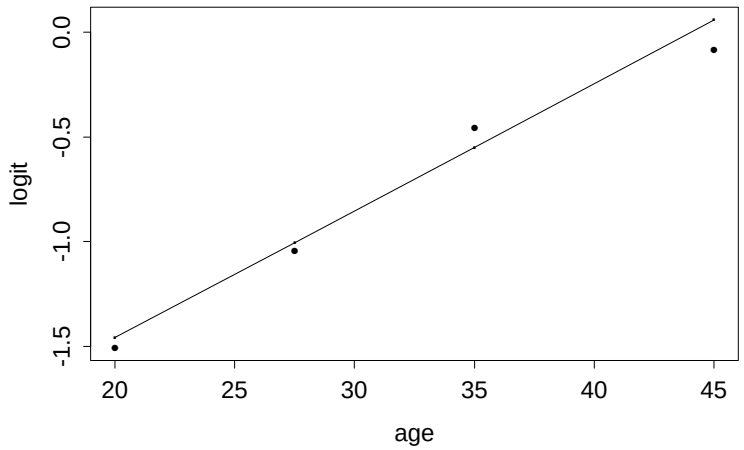


FIGURE 3.2: Observed and Fitted Logits for Model of Contraceptive Use with a Linear Effect of Age

ages.

3.5 Models With Two Predictors

We now consider models involving two predictors, and discuss the binary data analogues of two-way analysis of variance, multiple regression with dummy variables, and analysis of covariance models. An important element of the discussion concerns the key concepts of main effects and interactions.

3.5.1 Age and Preferences

Consider the distribution of contraceptive use by age and desire for more children, as summarized in Table 3.7. We have a total of eight groups, which will be indexed by a pair of subscripts i, j, with $i = 1, 2, 3, 4$ referring to the four age groups and $j = 1, 2$ denoting the two categories of desire for more children. We let y_{ij} denote the number of women using contraception and n_{ij} the total number of women in age group i and category j of desire for more children.

We now analyze these data under the usual assumption of a binomial error structure, so the y_{ij} are viewed as realizations of independent random variables $Y_{ij} \sim B(n_{ij}, \pi_{ij})$.

TABLE 3.7: Contraceptive Use by Age and Desire for More Children

Age i	Desires j	Using y_{ij}	Not Using $n_{ij} - y_{ij}$	All n_{ij}
<25	Yes	58	265	323
	No	14	60	74
25–29	Yes	68	215	283
	No	37	84	121
30–39	Yes	79	230	309
	No	158	145	303
40–49	Yes	14	43	57
	No	79	58	137
Total		507	1100	1607

3.5.2 The Deviance Table

There are five basic models of interest for the systematic structure of these data, ranging from the null to the saturated model. These models are listed in Table 3.8, which includes the name of the model, a descriptive notation, the formula for the linear predictor, the deviance or goodness of fit likelihood ratio chi-squared statistic, and the degrees of freedom.

Note first that the null model does not fit the data: the deviance of 145.7 on 7 d.f. is much greater than 14.1, the 95-th percentile of the chi-squared distribution with 7 d.f. This result is not surprising, since we already knew that contraceptive use depends on desire for more children and varies by age.

TABLE 3.8: Deviance Table for Models of Contraceptive Use by Age (Grouped) and Desire for More Children

Model	Notation	$\text{logit}(\pi_{ij})$	Deviance	d.f.
Null	ϕ	η	145.7	7
Age	A	$\eta + \alpha_i$	66.5	4
Desire	D	$\eta + \beta_j$	54.0	6
Additive	$A + D$	$\eta + \alpha_i + \beta_j$	16.8	3
Saturated	AD	$\eta + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	0	0

Introducing age in the model reduces the deviance to 66.5 on four d.f. The difference in deviances between the null model and the age model provides a test for the *gross* effect of age. The difference is 79.2 on three d.f.,

and is highly significant. This value is exactly the same that we obtained in the previous section, when we tested for an age effect using the data classified by age only. Moreover, the estimated age effects based on fitting the age model to the three-way classification in Table 3.7 would be exactly the same as those estimated in the previous section, and have the property of reproducing exactly the proportions using contraception in each age group.

This equivalence illustrate an important property of binomial models. All information concerning the gross effect of age on contraceptive use is contained in the marginal distribution of contraceptive use by age. We can work with the data classified by age only, by age and desire for more children, by age, education and desire for more children, or even with the individual data. In all cases the estimated effects, standard errors, and likelihood ratio tests based on differences between deviances will be the same.

The deviances themselves will vary, however, because they depend on the context. In the previous section the deviance of the age model was zero, because treating age as a factor reproduces exactly the proportions using contraception by age. In this section the deviance of the age model is 66.5 on four d.f. and is highly significant, because the age model does not reproduce well the table of contraceptive use by both age and preferences. In both cases, however, the difference in deviances between the age model and the null model is 79.2 on three d.f.

The next model in Table 3.8 is the model with a main effect of desire for more children, and has a deviance of 54.0 on six d.f. Comparison of this value with the deviance of the null model shows a gain of 97.1 at the expense of one d.f., indicating a highly significant *gross* effect of desire for more children. This is, of course, the same result that we obtained in Section 3.3, when we first looked at contraceptive use by desire for more children. Note also that this model does not fit the data, as it own deviance is highly significant.

The fact that the effect of desire for more children has a chi-squared statistic of 91.7 with only one d.f., whereas age gives 79.2 on three d.f., suggests that desire for more children has a stronger effect on contraceptive use than age does. Note, however, that the comparison is informal; the models are not nested, and therefore we cannot construct a significance test from their deviances.

3.5.3 The Additive Model

Consider now the two-factor additive model, denoted $A + D$ in Table 3.8. In this model the logit of the probability of using contraception in age group i

and in category j of desire for more children is

$$\text{logit}(\pi_{ij}) = \eta + \alpha_i + \beta_j,$$

where η is a constant, the α_i are age effects and the β_j are effects of desire for more children. To avoid redundant parameters we adopt the reference cell method and set $\alpha_1 = \beta_1 = 0$. The parameters may then be interpreted as follows:

- η is the logit of the probability of using contraception for women under 25 who want more children, who serve as the reference cell,
- α_i for $i = 2, 3, 4$ represents the *net* effect of ages 25–29, 30–39 and 40–49, compared to women under age 25 in the same category of desire for more children,
- β_2 represents the *net* effect of wanting no more children, compared to women who want more children in the same age group.

The model is additive in the logit scale, in the usual sense that the effect of one variable does not depend on the value of the other. For example, the effect of desiring no more children is β_2 in all four age groups. (This assumption must obviously be tested, and we shall see that it is not consistent with the data.)

The deviance of the additive model is 16.8 on three d.f. With this value we can calculate three different tests of interest, all of which involve comparisons between nested models.

- As we move from model D to $A + D$ the deviance decreases by 37.2 while we lose three d.f. This statistic tests the hypothesis $H_0 : \alpha_i = 0$ for all i, concerning the *net* effect of age after adjusting for desire for more children, and is highly significant.
- As we move from model A to $A + D$ we reduce the deviance by 49.7 at the expense of one d.f. This chi-squared statistic tests the hypothesis $H_0 : \beta_2 = 0$ concerning the *net* effect of desire for more children after adjusting for age. This value is highly significant, so we reject the hypothesis of no net effects.
- Finally, the deviance of 16.8 on three d.f. is a measure of goodness of fit of the additive model: it compares this model with the saturated model, which adds an interaction between the two factors. Since the deviance exceeds 11.3, the one-percent critical value in the chi-squared

distribution for three d.f., we conclude that the additive model fails to fit the data.

Table 3.9 shows parameter estimates for the additive model. We show briefly how they would be interpreted, although we have evidence that the additive model does not fit the data.

TABLE 3.9: Parameter Estimates for Additive Logit Model of Contraceptive Use by Age (Grouped) and Desire for Children

Parameter		Symbol	Estimate	Std. Error	z-ratio
Constant		η	−1.694	0.135	−12.53
Age	25–29	α_2	0.368	0.175	2.10
	30–39	α_3	0.808	0.160	5.06
	40–49	α_4	1.023	0.204	5.01
Desire	No	β_2	0.824	0.117	7.04

The estimates of the α_j's show a monotonic effect of age on contraceptive use. Although there is evidence that this effect may vary depending on whether women desire more children, on average the odds of using contraception among women age 40 or higher are nearly three times the corresponding odds among women under age 25 in the same category of desire for another child.

Similarly, the estimate of β_2 shows a strong effect of wanting no more children. Although there is evidence that this effect may depend on the woman's age, on average the odds of using contraception among women who desire no more children are more than double the corresponding odds among women in the same age group who desire another child.

3.5.4 A Model With Interactions

We now consider a model which includes an interaction of age and desire for more children, denoted AD in Table 3.8. The model is

$$\text{logit}(\pi_{ij}) = \eta + \alpha_i + \beta_j + (\alpha\beta)_{ij},$$

where η is a constant, the α_i and β_j are the main effects of age and desire, and $(\alpha\beta)_{ij}$ is the interaction effect. To avoid redundancies we follow the reference cell method and set to zero all parameters involving the first cell, so that $\alpha_1 = \beta_1 = 0$, $(\alpha\beta)_{1j} = 0$ for all j and $(\alpha\beta)_{i1} = 0$ for all i. The remaining parameters may be interpreted as follows:

η is the logit of the reference group: women under age 25 who desire more children.

α_i for $i = 2, 3, 4$ are the effects of the age groups 25–29, 30–39 and 40–49, compared to ages under 25, for women who want another child.

β_2 is the effect of desiring no more children, compared to wanting another child, for women under age 25.

$(\alpha\beta)_{i2}$ for $i = 2, 3, 4$ is the *additional* effect of desiring no more children, compared to wanting another child, for women in age group i rather than under age 25. (This parameter is also the *additional* effect of age group i, compared to ages under 25, for women who desire no more children rather than those who want more.)

One way to simplify the presentation of results involving interactions is to combine the interaction terms with one of the main effects, and present them as effects of one factor within categories or levels of the other. In our example, we can combine the interactions $(\alpha\beta)_{i2}$ with the main effects of desire β_2, so that

$\beta_2 + (\alpha\beta)_{i2}$ is the effect of desiring no more children, compared to wanting another child, for women in age group i.

Of course, we could also combine the interactions with the main effects of age, and speak of age effects which are specific to women in each category of desire for more children. The two formulations are statistically equivalent, but the one chosen here seems demographically more sensible.

To obtain estimates based on this parameterization of the model we have to define the columns of the model matrix as follows. Let a_i be a dummy variable representing age group i, for $i = 2, 3, 4$, and let d take the value one for women who want no more children and zero otherwise. Then the model matrix $\mathbf{X}$ should have a column of ones to represent the constant or reference cell, the age dummies a_2, a_3 and a_4 to represent the age effects for women in the reference cell, and then the dummy d and the products a_2d, a_3d and a_4d, to represent the effect of wanting no more children at ages < 25, 25–29, 30–39 and 40–49, respectively. The resulting estimates and standard errors are shown in Table 3.10.

The results indicate that contraceptive use among women who desire more children varies little by age, increasing up to age 35–39 and then declining somewhat. On the other hand, the effect of wanting no more children

TABLE 3.10: Parameter Estimates for Model of Contraceptive Use With an Interaction Between Age (Grouped) and Desire for More Children

Parameter		Estimate	Std. Error	z-ratio
Constant		−1.519	0.145	−10.481
Age	25–29	0.368	0.201	1.832
	30–39	0.451	0.195	2.311
	40–49	0.397	0.340	1.168
Desires	<25	0.064	0.330	0.194
No More	25–29	0.331	0.241	1.372
at Age	30–39	1.154	0.174	6.640
	40–49	1.431	0.353	4.057

increases dramatically with age, from no effect among women below age 25 to an odds ratio of 4.18 at ages 40–49. Thus, in the older cohort the odds of using contraception among women who want no more children are four times the corresponding odds among women who desire more children. The results can also be summarized by noting that contraceptive use for spacing (i.e. among women who desire more children) does not vary much by age, but contraceptive use for limiting fertility (i.e among women who want no more children) increases sharply with age.

3.5.5 Analysis of Covariance Models

Since the model with an age by desire interaction is saturated, we have essentially reproduced the observed data. We now consider whether we could attain a more parsimonious fit by treating age as a variate and desire for more children as a factor, in the spirit of covariance analysis models.

Table 3.11 shows deviances for three models that include a linear effect of age using, as before, the midpoints of the age groups. To emphasize this point we use X rather than A to denote age.

The first model assumes that the logits are linear functions of age. This model fails to fit the data, which is not surprising because it ignores desire for more children, a factor that has a large effect on contraceptive use.

The next model, denoted $X + D$, is analogous to the two-factor additive model. It allows for an effect of desire for more children which is the same at all ages. This common effect is modelled by allowing each category of desire for more children to have its own constant, and results in two parallel lines. The common slope is the effect of age within categories of desire for

TABLE 3.11: Deviance Table for Models of Contraceptive Use by Age (Linear) and Desire for More Children

Model	Notation	$\text{logit}(\pi_{ij})$	Deviance	d.f.
One Line	X	$\alpha + \beta x_i$	68.88	6
Parallel Lines	$X + D$	$\alpha_j + \beta x_i$	18.99	5
Two Lines	XD	$\alpha_j + \beta_j x_i$	9.14	4

more children. The reduction in deviance of 39.9 on one d.f. indicates that desire for no more children has a strong effect on contraceptive use after controlling for a linear effect of age. However, the attained deviance of 19.0 on five d.f. is significant, indicating that the assumption of two parallel lines is not consistent with the data.

The last model in the table, denoted XD, includes an interaction between the linear effect of age and desire, and thus allows the effect of desire for more children to vary by age. This variation is modelled by allowing each category of desire for more children to have its own slope in addition to its own constant, and results in two regression lines. The reduction in deviance of 9.9 on one d.f. is a test of the hypothesis of parallelism or common slope $H_0 : \beta_1 = \beta_2$, which is rejected with a P-value of 0.002. The model deviance of 9.14 on four d.f. is just below the five percent critical value of the chi-squared distribution with four d.f., which is 9.49. Thus, we have no evidence against the assumption of two straight lines.

Before we present parameter estimates we need to discuss briefly the choice of parameterization. Direct application of the reference cell method leads us to use four variables: a dummy variable always equal to one, a variable x with the mid-points of the age groups, a dummy variable d which takes the value one for women who want no more children, and a variable dx equal to the product of this dummy by the mid-points of the age groups. This choice leads to parameters representing the constant and slope for women who want another child, and parameters representing the *difference* in constants and slopes for women who want no more children.

An alternative is to simply report the constants and slopes for the two groups defined by desire for more children. This parameterization can be easily obtained by omitting the constant and using the following four variables: d and $1 - d$ to represent the two constants and dx and $(1 - d)x$ to represent the two slopes. One could, of course, obtain the constant and slope for women who want no more children from the previous parameterization

simply by adding the main effect and the interaction. The simplest way to obtain the standard errors, however, is to change parameterization.

In both cases the constants represent effects at age zero and are not very meaningful. To obtain parameters that are more directly interpretable, we can center age around the sample mean, which is 30.6 years. Table 3.12 shows parameter estimates obtained under the two parameterizations discussed above, using the mid-points of the age groups minus the mean.

TABLE 3.12: Parameter Estimates for Model of Contraceptive Use With an Interaction Between Age (Linear) and Desire for More Children

Desire	Age	Symbol	Estimate	Std. Error	z-ratio
More	Constant	α_1	−1.1944	0.0786	−15.20
	Slope	β_1	0.0218	0.0104	2.11
No More	Constant	α_2	−0.4369	0.0931	−4.69
	Slope	β_2	0.0698	0.0114	6.10
Difference	Constant	$\alpha_2 - \alpha_1$	0.7575	0.1218	6.22
	Slope	$\beta_2 - \beta_1$	0.0480	0.0154	3.11

Thus, we find that contraceptive use increases with age, but at a faster rate among women who want no more children. The estimated slopes correspond to increases in the odds of two and seven percent per year of age for women who want and do not want more children, respectively. The difference of the slopes is significant by a likelihood ratio test or by Wald's test, with a z-ratio of 3.11.

Similarly, the effect of wanting no more children increases with age. The odds ratio around age 30.6—which we obtain by exponentiating the difference in constants–is 2.13, so not wanting more children at this age is associated with a doubling of the odds of using contraception. The difference in slopes of 0.048 indicates that this differential increases five percent per year of age.

The parameter estimates in Table 3.12 may be used to produce fitted logits for each age group and category of desire for more children. In turn, these can be compared with the empirical logits for the original eight groups, to obtain a visual impression of the nature of the relationships studied and the quality of the fit. The comparison appears in Figure 3.3, with the solid line representing the linear age effects (the dotted lines are discussed below). The graph shows clearly how the effect of wanting no more children increases with age (or, alternatively, how age has much stronger effects among limiters

than among spacers).

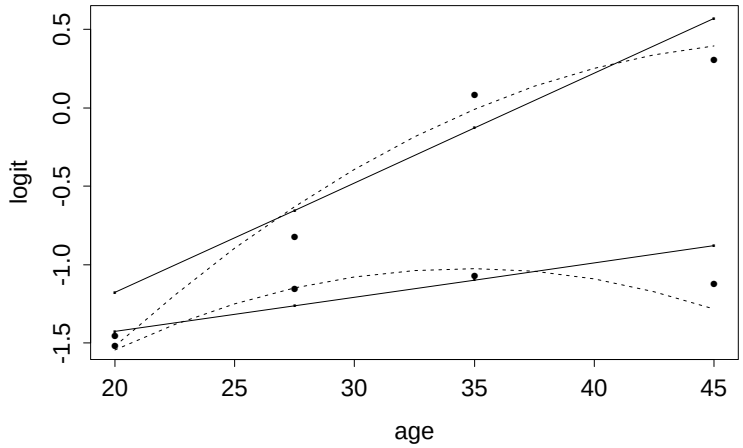


FIGURE 3.3: Observed and Fitted Logits for Models of Contraceptive Use With Effects of Age (Linear and Quadratic), Desire for More Children and a Linear Age by Desire Interaction.

The graph also shows that the assumption of linearity of age effects, while providing a reasonably parsimonious description of the data, is somewhat suspect, particularly at higher ages. We can improve the fit by adding higher-order terms on age. In particular

- Introducing a quadratic term on age yields an excellent fit, with a deviance of 2.34 on three d.f. This model consists of two parabolas, one for each category of desire for more children, but with the same curvature.

- Adding a quadratic age by desire interaction further reduces the deviance to 1.74 on two d.f. This model allows for two separate parabolas tracing contraceptive use by age, one for each category of desire.

Although the linear model passes the goodness of fit test, the fact that we can reduce the deviance by 6.79 at the expense of one d.f. indicates significant curvature. The dotted line in Figure 3.3 shows the intermediate model, where the curvature by age is the same for the two groups. While the fit is much better, the overall substantive conclusions do not change.

3.6 Multi-factor Models: Model Selection

Let us consider a full analysis of the contraceptive use data in Table 3.1, including all three predictors: age, education and desire for more children.

We use three subscripts to reflect the structure of the data, so π_{ijk} is the probability of using contraception in the (i, j, k)-th group, where $i = 1, 2, 3, 4$ indexes the age groups, $j = 1, 2$ the levels of education and $k = 1, 2$ the categories of desire for more children.

3.6.1 Deviances for One and Two-Factor Models

There are 19 basic models of interest for these data, which are listed for completeness in Table 3.13. Not all of these models would be of interest in any given analysis. The table shows the model in abbreviated notation, the formula for the linear predictor, the deviance and its degrees of freedom.

Note first that the null model does not fit the data. The assumption of a common probability of using contraception for all 16 groups of women is clearly untenable.

Next in the table we find the three possible one-factor models. Comparison of these models with the null model provides evidence of significant *gross* effects of age and desire for more children, but not of education. The likelihood ratio chi-squared tests are 91.7 on one d.f. for desire, 79.2 on three d.f. for age, and 0.7 on one d.f. for education.

Proceeding down the table we find the six possible two-factor models, starting with the additive ones. Here we find evidence of significant *net effects* of age and desire for more children after controlling for one other factor. For example the test for an effect of desire net of age is a chi-squared of 49.7 on one d.f., obtained by comparing the additive model $A + D$ on age and desire the one-factor model A with age alone. Education has a significant effect net of age, but not net of desire for more children. For example the test for the net effect of education controlling for age is 6.2 on one d.f., and follows from the comparison of the $A + E$ model with A. None of the additive models fits the data, but the closest one to a reasonable fit is $A + D$.

Next come the models involving *interactions* between two factors. We use the notation ED to denote the model with the main effects of E and D as well as the $E \times D$ interaction. Comparing each of these models with the corresponding additive model on the same two factors we obtain a test of the interaction effect. For example comparing the model ED with the additive model $E + D$ we can test whether the effect of desire for more children varies

TABLE 3.13: Deviance Table for Logit Models of Contraceptive Use by Age, Education and Desire for More Children

Model	$\text{logit}(\pi_{ijk})$	Dev.	d.f.
Null	η	165.77	15
One Factor			
*A*ge	$\eta+\alpha_i$	86.58	12
*E*ducation	$\eta+\beta_j$	165.07	14
*D*esire	$\eta+\gamma_k$	74.10	14
Two Factors			
$A+E$	$\eta+\alpha_i+\beta_j$	80.42	11
$A+D$	$\eta+\alpha_i+\gamma_k$	36.89	11
$E+D$	$\eta+\beta_j+\gamma_k$	73.87	13
AE	$\eta+\alpha_i+\beta_j+(\alpha\beta)_{ij}$	73.03	8
AD	$\eta+\alpha_i+\gamma_k+(\alpha\gamma)_{ik}$	20.10	8
ED	$\eta+\beta_j+\gamma_k+(\beta\gamma)_{jk}$	67.64	12
Three Factors			
$A+E+D$	$\eta+\alpha_i+\beta_j+\gamma_k$	29.92	10
$AE+D$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}$	23.15	7
$AD+E$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\gamma)_{ik}$	12.63	7
$A+ED$	$\eta+\alpha_i+\beta_j+\gamma_k+(\beta\gamma)_{jk}$	23.02	9
$AE+AD$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}$	5.80	4
$AE+ED$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\beta\gamma)_{jk}$	13.76	6
$AD+ED$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}$	10.82	6
$AE+AD+ED$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}$	2.44	3

with education. Making these comparisons we find evidence of interactions between age and desire for more children ($\chi^2 = 16.8$ on three d.f.), and between education and desire for more children ($\chi^2 = 6.23$ on one d.f.), but not between age and education ($\chi^2 = 7.39$ on three d.f.).

All of the results described so far could be obtained from two-dimensional tables of the type analyzed in the previous sections. The new results begin

to appear as we consider the nine possible three-factor models.

3.6.2 Deviances for Three-Factor Models

The first entry is the *additive* model $A+E+D$, with a deviance of 29.9 on ten d.f. This value represents a significant improvement over any of the additive models on two factors. Thus, we have evidence that there are significant net effects of age, education and desire for more children, considering each factor after controlling the other two. For example the test for a net effect of education controlling the other two variables compares the three-factor additive model $A+E+D$ with the model without education, namely $A+D$. The difference of 6.97 on one d.f. is significant, with a P-value of 0.008. However, the three-factor additive model does not fit the data.

The next step is to add *one interaction* between two of the factors. For example the model $AE + D$ includes the main effects of A, E and D and the $A \times E$ interaction. The interactions of desire for more children with age and with education produce significant gains over the additive model ($\chi^2 = 17.3$ on three d.f. and $\chi^2 = 6.90$ on one d.f., respectively), whereas the interaction between age and education is not significant ($\chi^2 = 6.77$ with three d.f.). These tests for interactions differ from those based on two-factor models in that they take into account the third factor. The best of these models is clearly the one with an interaction between age and desire for more children, $AD+E$. This is also the first model in our list that actually passes the goodness of fit test, with a deviance of 12.6 on seven d.f.

Does this mean that we can stop our search for an adequate model? Unfortunately, it does not. The goodness of fit test is a joint test for all terms omitted in the model. In this case we are testing for the AE, ED and AED interactions simultaneously, a total of seven parameters. This type of omnibus test lacks power against specific alternatives. It is possible that one of the omitted terms (or perhaps some particular contrast) would be significant by itself, but its effect may not stand out in the aggregate. At issue is whether the remaining deviance of 12.6 is spread out uniformly over the remaining d.f. or is concentrated in a few d.f. If you wanted to be absolutely sure of not missing anything you might want to aim for a deviance below 3.84, which is the five percent critical value for one d.f., but this strategy would lead to over-fitting if followed blindly.

Let us consider the models involving *two interactions* between two factors, of which there are three. Since the AD interaction seemed important we restrict attention to models that include this term, so we start from $AD + E$, the best model so far. Adding the age by education interaction

AE to this model reduces the deviance by 6.83 at the expense of three d.f. A formal test concludes that this interaction is not significant. If we add instead the education by desire interaction ED we reduce the deviance by only 1.81 at the expense of one d.f. This interaction is clearly not significant. A model-building strategy based on *forward selection* of variables would stop here and choose $AD + E$ as the best model on grounds of parsimony and goodness of fit.

An alternative approach is to start with the saturated model and impose progressive simplification. Deleting the *three-factor interaction* yields the model $AE + AD + ED$ with three two-factor interactions, which fits the data rather well, with a deviance of just 2.44 on three d.f. If we were to delete the AD interaction the deviance would rise by 11.32 on three d.f., a significant loss. Similarly, removing the AE interaction would incur a significant loss of 8.38 on 3 d.f. We can, however, drop the ED interaction with a non-significant increase in deviance of 3.36 on one d.f. At this point we can also eliminate the AE interaction, which is no longer significant, with a further loss of 6.83 on three d.f. Thus, a *backward elimination* strategy ends up choosing the same model as forward selection.

Although you may find these results reassuring, there is a fact that both approaches overlook: the AE and DE interactions are jointly significant! The change in deviance as we move from $AD+E$ to the model with three two-factor interactions is 10.2 on four d.f., and exceeds (although not by much) the five percent critical value of 9.5. This result indicates that we need to consider the more complicated model with all three two-factor interactions. Before we do that, however, we need to discuss parameter estimates for selected models.

3.6.3 The Additive Model: Gross and Net Effects

Consider first Table 3.14, where we adopt an approach similar to multiple classification analysis to compare the gross and net effects of all three factors. We use the reference cell method, and include the omitted category for each factor (with a dash where the estimated effect would be) to help the reader identify the baseline.

The gross or unadjusted effects are based on the single-factor models A, E and D. These effects represent overall differences between levels of each factor, and as such they have descriptive value even if the one-factor models do not tell the whole story. The results can easily be translated into odds ratios. For example not wanting another child is associated with an increase in the odds of using contraception of 185%. Having upper primary or higher

TABLE 3.14: Gross and Net Effects of Age, Education and Desire for More Children on Current Use of Contraception

Variable and category		Gross effect	Net effect
Constant		–	−1.966
Age	<25	–	–
	25–29	0.461	0.389
	30–39	1.048	0.909
	40–49	1.425	1.189
Education			
	Lower	–	–
	Upper	-0.093	0.325
Desires More			
	Yes	–	–
	No	1.049	0.833

education rather than lower primary or less appears to reduce the odds of using contraception by almost 10%.

The net or adjusted effects are based on the three-factor additive model $A+E+D$. This model assumes that the effect of each factor is the same for all categories of the others. We know, however, that this is not the case—particularly with desire for more children, which has an effect that varies by age—so we have to interpret the results carefully. The net effect of desire for more children shown in Table 3.14 represents an average effect across all age groups and may not be representative of the effect at any particular age. Having said that, we note that desire for no more children has an important effect net of age and education: on the average, it is associated with an increase in the odds of using contraception of 130%.

The result for education is particularly interesting. Having upper primary or higher education is associated with an increase in the odds or using contraception of 38%, compared to having lower primary or less, after we control for age and desire for more children. The gross effect was close to zero. To understand this result bear in mind that contraceptive use in Fiji occurs mostly among older women who want no more children. Education has no effect when considered by itself because in Fiji more educated women are likely to be younger than less educated women, and thus at a stage of their lives when they are less likely to have reached their desired family size,

even though they may want fewer children. Once we adjust for their age, calculating the net effect, we obtain the expected association. In this example age is said to act as a *suppressor* variable, masking the association between education and contraceptive use.

We could easily add columns to Table 3.14 to trace the effects of one factor after controlling for one or both of the other factors. We could, for example, examine the effect of education adjusted for age, the effect adjusted for desire for more children, and finally the effect adjusted for both factors. This type of analysis can yield useful insights into the confounding influences of other variables.

3.6.4 The Model with One Interaction Effect

Let us now examine parameter estimates for the model with an age by desire for more children interaction $AD + E$, where

$$\text{logit}(\pi_{ijk}) = \eta + \alpha_i + \beta_j + \gamma_j + (\alpha\gamma)_{ik}.$$

The parameter estimates depend on the restrictions used in estimation. We use the reference cell method, so that $\alpha_1 = \beta_1 = \gamma_1 = 0$, and $(\alpha\gamma)_{ik} = 0$ when either $i = 1$ or $k = 1$.

In this model η is the logit of the probability of using contraception in the reference cell, that is, for women under 25 with lower primary or less education who want another child. On the other hand β_2 is the effect of upper primary or higher education, compared to lower primary or less, for women in any age group or category of desire for another child. The presence of an interaction makes interpretation of the estimates for age and desire somewhat more involved:

α_i represents the effect of age group i, compared to age < 25, for women who want more children.

γ_2 represents the effect of wanting no more children, compared to desiring more, for women under age 25.

$(\alpha\gamma)_{i2}$, the interaction term, can be interpreted as the *additional* effect of wanting no more children among women in age group i, compared to women under age 25.

It is possible to simplify slightly the presentation of the results by combining the interactions with some of the main effects. In the present example, it is convenient to present the estimates of α_i as the age effects for women who

TABLE 3.15: The Estimates

Variable	Category	Symbol	Estimate	Std. Err	z-ratio
Constant		η	-1.803	0.180	-10.01
Age	25–29	α_2	0.395	0.201	1.96
	30–39	α_3	0.547	0.198	2.76
	40–49	α_4	0.580	0.347	1.67
Education	Upper	β_2	0.341	0.126	2.71
Desires	<25	γ_2	0.066	0.331	0.20
no more	25–29	$\gamma_2 + (\alpha\gamma)_{22}$	0.325	0.242	1.35
at age	30–39	$\gamma_2 + (\alpha\gamma)_{32}$	1.179	0.175	6.74
	40–49	$\gamma_2 + (\alpha\gamma)_{42}$	1.428	0.354	4.04

want another child, and to present $\gamma_2 + (\alpha\gamma)_{i2}$ as the effect of not wanting another child for women in age group i.

Calculation of the necessary dummy variables proceeds exactly as in Section 3.5. This strategy leads to the parameter estimates in Table 3.15.

To aid in interpretation as well as model criticism, Figure 3.4 plots observed logits based on the original data in Table 3.1, and fitted logits based on the model with an age by desire interaction.

The graph shows four curves tracing contraceptive use by age for groups defined by education and desire for more children. The curves are labelled using L and U for lower and upper education, and Y and N for desire for more children. The lowest curve labelled LY corresponds to women with lower primary education or less who want more children, and shows a slight increase in contraceptive use up to age 35–39 and then a small decline. The next curve labelled UY is for women with upper primary education or more who also want more children. This curve is parallel to the previous one because the effect of education is additive on age. The constant difference between these two curves corresponds to a 41% increase in the odds ratio as we move from lower to upper primary education. The third curve, labelled LN, is for women with lower primary education or less who want no more children. The distance between this curve and the first one represents the effect of wanting no more children at different ages. This effect increases sharply with age, reaching an odds ratio of four by age 40–49. The fourth curve, labelled UN, is for women with upper primary education or more who want no more children. The distance between this curve and the previous one is the effect of education, which is the same whether women want more children or not, and is also the same at every age.

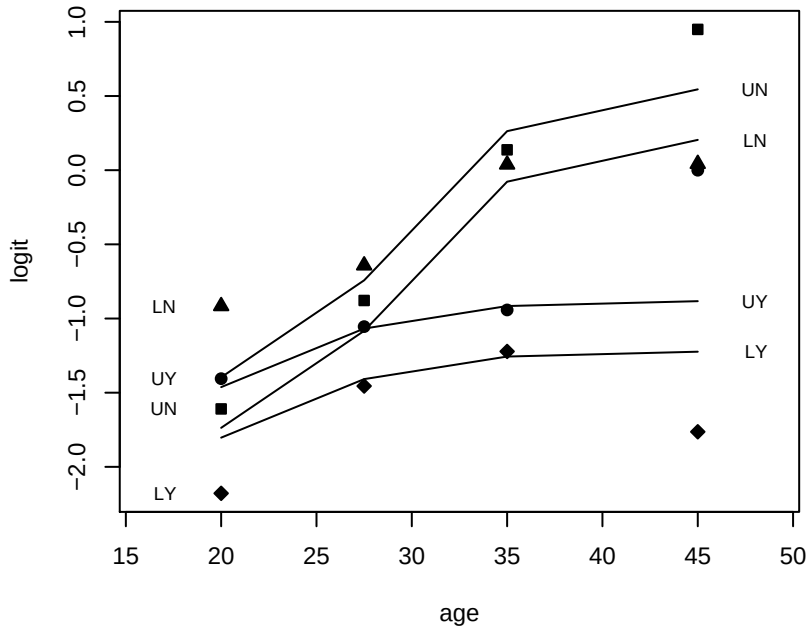


FIGURE 3.4: Logit Model of Contraceptive Use By Age, Education and Desire for Children, With Age by Desire Interaction

The graph also shows the observed logits, plotted using different symbols for each of the four groups defined by education and desire. Comparison of observed and fitted logits shows clearly the strengths and weaknesses of this model: it does a fairly reasonable job reproducing the logits of the proportions using contraception in each group *except* for ages 40–49 (and to a lesser extend the group < 25), where it seems to underestimate the educational differential. There is also some indication that this failure may be more pronounced for women who want more children.

3.6.5 Best Fitting and Parsimonious Models

How can we improve the model of the last section? The most obvious solution is to move to the model with all three two-factor interactions, $AE + AD + ED$, which has a deviance of 2.44 on three d.f. and therefore fits the data

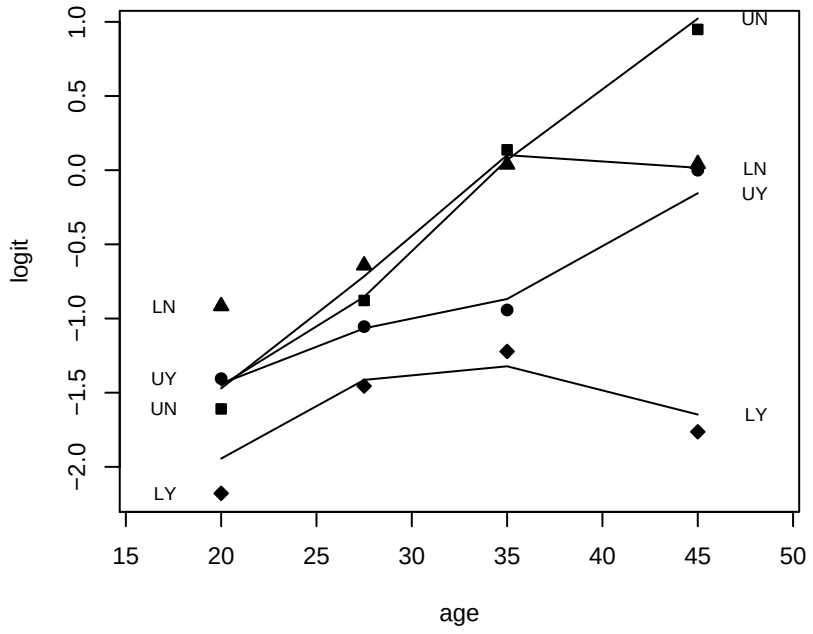


FIGURE 3.5: Observed and Fitted Logits of Contraceptive Use Based on Model with Three Two-Factor Interactions

extremely well. This model implies that the effect of each factor depends on the levels of the other two, but not on the combination of levels of the other two. Interpretation of the coefficients in this model is not as simple as it would be in an additive model, or in a model involving only one interaction. The best strategy in this case is to plot the fitted model and inspect the resulting curves.

Figure 3.5 shows fitted values based on the more complex model. The plot tells a simple story. Contraceptive use for spacing increases slightly up to age 35 and then declines for the less educated but continues to increase for the more educated. Contraceptive use for limiting increases sharply with age up to age 35 and then levels off for the less educated, but continues to increase for the more educated. The figure shows that the effect of wanting no more children increases with age, and appears to do so for both educational groups in the same way (look at the distance between the LY and LN curves, and

between the UY and UN curves). On the other hand, the effect of education is clearly more pronounced at ages 40–49 than at earlier ages, and also seems slightly larger for women who want more children than for those who do not (look at the distance between the LY and UY curves, and between the LN and UN curves).

One can use this knowledge to propose improved models that fit the data without having to use all three two-factor interactions. One approach would note that all interactions with age involve contrasts between ages 40–49 and the other age groups, so one could collapse age into only two categories for purposes of modelling the interactions. A simplified version of this approach is to start from the model $AD + E$ and add one d.f. to model the larger educational effect for ages 40–49. This can be done by adding a dummy variable that takes the value one for women aged 40–49 who have upper primary or more education. The resulting model has a deviance of 6.12 on six d.f., indicating a good fit. Comparing this value with the deviance of 12.6 on seven d.f. for the $AD + E$ model, we see that we reduced the deviance by 6.5 at the expense of a single d.f. The model $AD + AE$ includes all three d.f. for the age by education interaction, and has a deviance of 5.8 on four d.f. Thus, the total contribution of the AE interaction is 6.8 on three d.f. Our one-d.f. improvement has captured roughly 90% of this interaction.

An alternative approach is to model the effects of education and desire for no more children as smooth functions of age. The logit of the probability of using contraception is very close to a linear function of age for women with upper primary education who want no more children, who could serve as a new reference cell. The effect of wanting more children could be modelled as a linear function of age, and the effect of education could be modelled as a quadratic function of age. Let L_{ijk} take the value one for lower primary or less education and zero otherwise, and let M_{ijk} be a dummy variable that takes the value one for women who want more children and zero otherwise. Then the proposed model can be written as

$$\text{logit}(\pi_{ijk}) = \alpha + \beta x_{ijk} + (\alpha_E + \beta_E x_{ijk} + \gamma_E x_{ijk}^2)L_{ijk} + (\alpha_D + \beta_D x_{ijk})M_{ijk}.$$

Fitting this model, which requires only seven parameters, gives a deviance of 7.68 on nine d.f. The only weakness of the model is that it assumes equal effects of education on use for limiting and use for spacing, but these effects are not well-determined. Further exploration of these models is left as an exercise.

3.7 Other Choices of Link

All the models considered so far use the logit transformation of the probabilities, but other choices are possible. In fact, any transformation that maps probabilities into the real line could be used to produce a generalized linear model, as long as the transformation is one-to-one, continuous and differentiable.

In particular, suppose $F(.)$ is the cumulative distribution function (c.d.f.) of a random variable defined on the real line, and write

$$\pi_i = F(\eta_i),$$

for $-\infty < \eta_i < \infty$. Then we could use the inverse transformation

$$\eta_i = F^{-1}(\pi_i),$$

for $0 < \pi_i < 1$ as the link function.

Popular choices of c.d.f.'s in this context are the normal, logistic and extreme value distributions. In this section we motivate this general approach by introducing models for binary data in terms of latent variables.

3.7.1 A Latent Variable Formulation

Let Y_i denote a random variable representing a binary response coded zero and one, as usual. We will call Y_i the *manifest* response. Suppose that there is an unobservable continuous random variable Y_i^* which can take any value in the real line, and such that Y_i takes the value one if an only if Y_i^* exceeds a certain threshold θ. We will call Y_i^* the *latent* response. Figure 3.6 shows the relationship between the latent variable and the response when the threshold is zero.

The interpretation of Y_i and Y_i^* depends on the context. An economist, for example, may view Y_i as a binary choice, such as purchasing or renting a home, and Y_i^* as the difference in the utilities of purchasing and renting. A psychologist may view Y_i as a response to an item in an attitude scale, such as agreeing or disagreeing with school vouchers, and Y_i^* as the underlying attitude. Biometricians often view Y_i^* as a dose and Y_i as a response, hence the name dose-response models.

Since a positive outcome occurs only when the latent response exceeds the threshold, we can write the probability π_i of a positive outcome as

$$\pi_i = \Pr\{Y_i = 1\} = \Pr\{Y_i^* > \theta\}.$$

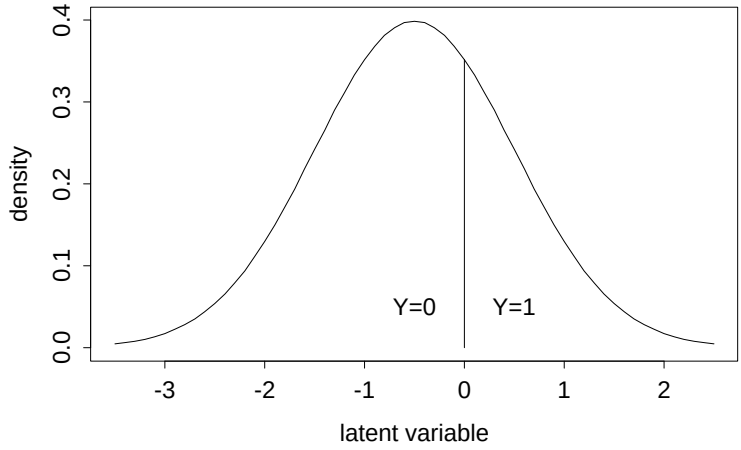


FIGURE 3.6: Latent Variable and Manifest Response

As often happens with latent variables, the location and scale of Y_i^* are arbitrary. We can add a constant a to both Y_i^* and the threshold θ, or multiply both by a constant c, without changing the probability of a positive outcome. To identify the model we take the threshold to be zero, and standardize Y_i^* to have standard deviation one (or any other fixed value).

Suppose now that the outcome depends on a vector of covariates $\mathbf{x}$. To model this dependence we use an ordinary linear model for the *latent* variable, writing

$$Y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + U_i, \tag{3.15}$$

where $\boldsymbol{\beta}$ is a vector of coefficients of the covariates $\mathbf{x}_i$ and U_i is the error term, assumed to have a distribution with c.d.f. $F(u)$, not necessarily the normal distribution.

Under this model, the probability π_i of observing a positive outcome is

$$\begin{aligned} \pi_i &= \Pr\{Y_i > 0\} \\ &= \Pr\{U_i > -\eta_i\} \\ &= 1 - F(-\eta_i), \end{aligned}$$

where $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ is the linear predictor. If the distribution of the error term U_i is symmetric about zero, so $F(u) = 1 - F(-u)$, we can write

$$\pi_i = F(\eta_i)$$

This expression defines a generalized linear model with Bernoulli response and link

$$\eta_i = F^{-1}(\pi_i). \tag{3.16}$$

In the more general case where the distribution of the error term is not necessarily symmetric, we still have a generalized linear model with link

$$\eta_i = -F^{-1}(1 - \pi_i). \tag{3.17}$$

We now consider some specific distributions.

3.7.2 Probit Analysis

The obvious choice of an error distribution is the normal. Assuming that the error term has a standard normal distribution $U_i \sim N(0, 1)$, the results of the previous section lead to

$$\pi_i = \Phi(\eta_i),$$

where Φ is the standard normal c.d.f. The inverse transformation, which gives the linear predictor as a function of the probability

$$\eta_i = \Phi^{-1}(\pi_i),$$

is called the *probit*.

It is instructive to consider the more general case where the error term $U_i \sim N(0, \sigma^2)$ has a normal distribution with variance σ^2. Following the same steps as before we find that

$$\begin{aligned}\pi_i &= \Pr\{Y_i^* > 0\} \\ &= \Pr\{U_i > -\mathbf{x}_i'\boldsymbol{\beta}\} = \Pr\{U_i/\sigma > -\mathbf{x}_i'\boldsymbol{\beta}/\sigma\} \\ &= 1 - \Phi(-\mathbf{x}_i'\boldsymbol{\beta}/\sigma) = \Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma),\end{aligned}$$

where we have divided by σ to obtain a standard normal variate, and used the symmetry of the normal distribution to obtain the last result.

This development shows that we cannot identify $\boldsymbol{\beta}$ and σ separately, because the probability depends on them only through their ratio $\boldsymbol{\beta}/\sigma$. This is another way of saying that the scale of the latent variable is not identified. We therefore take $\sigma = 1$, or equivalently interpret the β's in units of standard deviation of the latent variable.

As a simple example, consider fitting a probit model to the contraceptive use data by age and desire for more children. In view of the results in Section 3.5, we introduce a main effect of wanting no more children, a linear effect

TABLE 3.16: Estimates for Probit Model of Contraceptive Use With a Linear Age by Desire Interaction

Parameter	Symbol	Estimate	Std. Error	z-ratio
Constant	α_1	−0.7297	0.0460	−15.85
Age	β_1	0.0129	0.0061	2.13
Desire	$\alpha_2 - \alpha_1$	0.4572	0.0731	6.26
Age × Desire	$\beta_2 - \beta_1$	0.0305	0.0092	3.32

of age, and a linear age by desire interaction. Fitting this model gives a deviance of 8.91 on four d.f. Estimates of the parameters and standard errors appear in Table 3.16

To interpret these results we imagine a latent continuous variable representing the woman's motivation to use contraception (or the utility of using contraception, compared to not using). At the average age of 30.6, not wanting more children increases the motivation to use contraception by almost half a standard deviation. Each year of age is associated with an increase in motivation of 0.01 standard deviations if she wants more children and 0.03 standard deviations more (for a total of 0.04) if she does not. In the next section we compare these results with logit estimates.

A slight disadvantage of using the normal distribution as a link for binary response models is that the c.d.f. does not have a closed form, although excellent numerical approximations and computer algorithms are available for computing both the normal probability integral and its inverse, the probit.

3.7.3 Logistic Regression

An alternative to the normal distribution is the standard logistic distribution, whose shape is remarkably similar to the normal distribution but has the advantage of a closed form expression

$$\pi_i = F(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}},$$

for $-\infty < \eta_i < \infty$. The standard logistic distribution is symmetric, has mean zero, and has variance $\pi^2/3$. The shape is very close to the normal, except that it has heavier tails. The inverse transformation, which can be obtained solving for η_i in the expression above is

$$\eta_i = F^{-1}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i},$$

our good old friend, the *logit*.

Thus, coefficients in a logit regression model can be interpret not only in terms of log-odds, but also as effects of the covariates on a latent variable that follows a linear model with logistic errors.

The logit and probit transformations are almost linear functions of each other for values of π_i in the range from 0.1 to 0.9, and therefore tend to give very similar results. Comparison of probit and logit coefficients should take into account the fact that the standard normal and the standard logistic distributions have different variances. Recall that with binary data we can only estimate the ratio $\boldsymbol{\beta}/\sigma$. In probit analysis we have implicitly set $\sigma = 1$. In a logit model, by using a standard logistic error term, we have effectively set $\sigma = \pi/\sqrt{3}$. Thus, coefficients in a logit model should be standardized dividing by $\pi/\sqrt{3}$ before comparing them with probit coefficients.

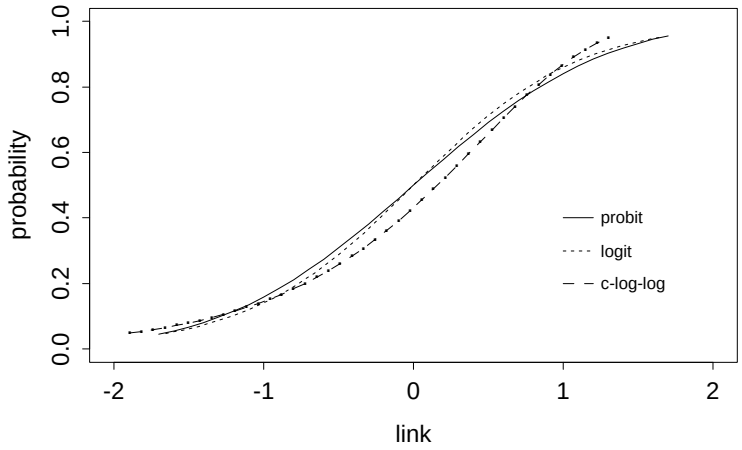


FIGURE 3.7: The Standardized Probit, Logit and C-Log-Log Links

Figure 3.7 compares the logit and probit links (and a third link discussed below) after standardizing the logits to unit variance. The solid line is the probit and the dotted line is the logit divided by $\pi/\sqrt{3}$. As you can see, they are barely distinguishable.

To illustrate the similarity of these links in practice, consider our models of contraceptive use by age and desire for more children in Tables 3.10 and 3.16. The deviance of 9.14 for the logit model is very similar to the deviance of 8.91 for the probit model, indicating an acceptable fit. The Wald tests of individual coefficients are also very similar, for example the test for the effect of wanting no more children at age 30.6 is 6.22 in the logit model and 6.26

in the probit model. The coefficients themselves look somewhat different, but of course they are not standardized. The effect of wanting no more children at the average age is 0.758 in the logit scale. Dividing by $\pi/\sqrt{3}$, the standard deviation of the underlying logistic distribution, we find this effect equivalent to an increase in the latent variable of 0.417 standard deviations. The probit analysis estimates the effect as 0.457 standard deviations.

3.7.4 The Complementary Log-Log Transformation

A third choice of link is the complementary log-log transformation

$$\eta_i = \log(-\log(1-\pi_i)),$$

which is the inverse of the c.d.f. of the extreme value (or log-Weibull) distribution, with c.d.f.

$$F(\eta_i) = 1 - e^{-e^{\eta_i}}.$$

For small values of π_i the complementary log-log transformation is close to the logit. As the probability increases, the transformation approaches infinity more slowly that either the probit or logit.

This particular choice of link function can also be obtained from our general latent variable formulation if we assume that $-U_i$ (note the minus sign) has a standard extreme value distribution, so the error term itself has a *reverse* extreme value distribution, with c.d.f.

$$F(U_i) = e^{-e^{-U_i}}.$$

The reverse extreme value distribution is asymmetric, with a long tail to the right. It has mean equal to Euler's constant 0.577 and variance $\pi^2/6 = 1.645$. The median is $-\log\log 2 = 0.367$ and the quartiles are -0.327 and 1.246.

Inverting the reverse extreme value c.d.f. and applying Equation 3.17, which is valid for both symmetric and asymmetric distributions, we find that the link corresponding to this error distribution is the complementary log-log.

Thus, coefficients in a generalized linear model with binary response and a complementary log-log link can be interpreted as effects of the covariates on a latent variable which follows a linear model with reverse extreme value errors.

To compare these coefficients with estimates based on a probit analysis we should standardize them, dividing by $\pi/\sqrt{6}$. To compare coefficients with logit analysis we should divide by $\sqrt{2}$, or standardize both c-log-log and logit coefficients.

Figure 3.7 compares the c-log-log link with the probit and logit after standardizing it to have mean zero and variance one. Although the c-log-log link differs from the other two, one would need extremely large sample sizes to be able to discriminate empirically between these links.

The complementary log-log transformation has a direct interpretation in terms of hazard ratios, and thus has practical applications in terms of hazard models, as we shall see later in the sequel.

3.8 Regression Diagnostics for Binary Data

Model checking is just as important in logistic regression and probit analysis as it is in classical linear models. The raw materials are again the residuals, or differences between observed and fitted values. Unlike the case of linear models, however, we now have to make allowance for the fact that the observations have different variances. There are two types of residuals in common use.

3.8.1 Pearson Residuals

A very simple approach to the calculation of residuals is to take the difference between observed and fitted values and divide by an estimate of the standard deviation of the observed value. The resulting residual has the form

$$p_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(n_i - \hat{\mu}_i)/n_i}}, \tag{3.18}$$

where $\hat{\mu}_i$ is the fitted value and the denominator follows from the fact that $\text{var}(y_i) = n_i\pi_i(1-\pi_i)$.

The result is called the Pearson residual because the square of p_i is the contribution of the i-th observation to Pearson's chi-squared statistic, which was introduced in Section 3.2.2, Equation 3.14.

With grouped data the Pearson residuals are approximately normally distributed, but this is not the case with individual data. In both cases, however, observations with a Pearson residual exceeding two in absolute value may be worth a closer look.

3.8.2 Deviance Residuals

An alternative residual is based on the deviance or likelihood ratio chi-squared statistic. The deviance residual is defined as

$$d_i = \sqrt{2[y_i \log(\frac{y_i}{\hat{\mu}_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{\mu}_i})]}, \tag{3.19}$$

with the same sign as the raw residual $y_i - \hat{y}_i$. Squaring these residuals and summing over all observations yields the deviance statistic. Observations with a deviance residual in excess of two may indicate lack of fit.

3.8.3 Studentized Residuals

The residuals defined so far are not fully standardized. They take into account the fact that different observations have different variances, but they make no allowance for additional variation arising from estimation of the parameters, in the way studentized residuals in classical linear models do.

Pregibon (1981) has extended to logit models some of the standard regression diagnostics. A key in this development is the weighted *hat* matrix

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2},$$

where $\mathbf{W}$ is the diagonal matrix of iteration weights from Section 3.2.1, with entries $w_{ii} = \mu_i(n_i - \mu_i)/n_i$, evaluated at the m.l.e.'s. Using this expression it can be shown that the variance of the raw residual is, to a first-order approximation,

$$\text{var}(y_i - \hat{\mu}_i) \approx (1 - h_{ii})\text{var}(y_i),$$

where $h_i i$ is the leverage or diagonal element of the weighted hat matrix. Thus, an internally studentized residual can be obtained dividing the Pearson residual by the square root of $1 - h_{ii}$, to obtain

$$s_i = \frac{p_i}{\sqrt{1 - h_{ii}}} = \frac{y_i - \hat{\mu}_i}{\sqrt{(1 - h_{ii})\hat{\mu}_i(n_i - \hat{\mu}_i)/n_i}}.$$

A similar standardization can be applied to deviance residuals. In both cases the standardized residuals have the same variance only approximately because the correction is first order, unlike the case of linear models where the correction was exact.

Consider now calculating jack-knifed residuals by omitting one observation. Since estimation relies on iterative procedures, this calculation would

be expensive. Suppose, however, that we start from the final estimates and do only one iteration of the IRLS procedure. Since this step is a standard weighted least squares calculation, we can apply the standard regression updating formulas to obtain the new coefficients and thus the predictive residuals. Thus, we can calculate a jack-knifed residual as a function of the standardized residual using the same formula as in linear models

$$t_i = s_i \sqrt{\frac{n-p-1}{n-p-s_i^2}}$$

and view the result as a one-step approximation to the true jack-knifed residual.

3.8.4 Leverage and Influence

The diagonal elements of the hat matrix can be interpreted as leverages just as in linear models. To measure actual rather than potential influence we could calculate Cook's distance, comparing $\hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}_{(i)}$, the m.l.e.'s of the coefficients with and without the i-th observation. Calculation of the later would be expensive if we iterated to convergence. Pregibon (1981), however, has shown that we can use the standard linear models formula

$$D_i = s_i^2 \frac{h_{ii}}{(1-h_{ii})p},$$

and view the result as a one-step approximation to Cook's distance, based on doing one iteration of the IRLS algorithm towards $\hat{\boldsymbol{\beta}}_{(i)}$ starting from the complete data estimate $\hat{\boldsymbol{\beta}}$.

3.8.5 Testing Goodness of Fit

With grouped data we can assess goodness of fit by looking directly at the deviance, which has approximately a chi-squared distribution for large n_i. A common rule of thumb is to require all expected frequencies (both expected successes $\hat{\mu}_i$ and failures $n_i - \hat{\mu}_i$) to exceed one, and 80% of them to exceed five.

With individual data this test is not available, but one can always group the data according to their covariate patterns. If the number of possible combinations of values of the covariates is not too large relative to the total sample size, it may be possible to group the data and conduct a formal goodness of fit test. Even when the number of covariate patterns is large, it is possible that a few patterns will account for most of the observations. In this

case one could compare observed and fitted counts at least for these common patterns, using either the deviance or Pearson's chi-squared statistic.

Hosmer and Lemeshow (1980, 1989) have proposed an alternative procedure that can be used with individual data even if there are no common covariate patterns. The basic idea is to use predicted probabilities to create groups. These authors recommend forming ten groups, with predicted probabilities of 0–0.1, 0.1–0.2, and so on, with the last group being 0.9–1. One can then compute expected counts of successes (and failures) for each group by summing the predicted values (and their complements), and compare these with observed values using Pearson's chi-squared statistic. Simulation studies show that the resulting statistic has approximately in large samples the usual chi-squared distribution, with degrees of freedom equal to $g - 2$, where g is the number of groups, usually ten. It seems reasonable to assume that this result would also apply if one used the deviance rather than Pearson's chi-squared.

Another measure that has been proposed in the literature is a pseudo-R^2, based on the proportion of deviance explained by a model. This is a direct extension of the calculations based on RSS's for linear models. These measures compare a given model with the null model, and as such do not necessarily measure goodness of fit. A more direct measure of goodness of fit would compare a given model with the saturated model, which brings us back again to the deviance.

Yet another approach to assessing goodness of fit is based on prediction errors. Suppose we were to use the fitted model to predict 'success' if the fitted probability exceeds 0.5 and 'failure' otherwise. We could then crosstabulate the observed and predicted responses, and calculate the proportion of cases predicted correctly. While intuitively appealing, one problem with this approach is that a model that fits the data may not necessarily predict well, since this depends on how predictable the outcome is. If prediction was the main objective of the analysis, however, the proportion classified correctly would be an ideal criterion for model comparison.

Chapter 4

Poisson Models for Count Data

In this chapter we study log-linear models for count data under the assumption of a Poisson error structure. These models have many applications, not only to the analysis of counts of events, but also in the context of models for contingency tables and the analysis of survival data.

4.1 Introduction to Poisson Regression

As usual, we start by introducing an example that will serve to illustrative regression models for count data. We then introduce the Poisson distribution and discuss the rationale for modeling the logarithm of the mean as a linear function of observed covariates. The result is a generalized linear model with Poisson response and link log.

4.1.1 The Children Ever Born Data

Table 4.1, adapted from Little (1978), comes from the Fiji Fertility Survey and is typical of the sort of table published in the reports of the World Fertility Survey. The table shows data on the number of children ever born to married women of the Indian race classified by duration since their first marriage (grouped in six categories), type of place of residence (Suva, other urban and rural), and educational level (classified in four categories: none, lower primary, upper primary, and secondary or higher). Each cell in the table shows the mean, the variance and the number of observations.

In our analysis of these data we will treat the number of children ever

TABLE 4.1: Number of Children Ever Born to Women of Indian Race
By Marital Duration, Type of Place of Residence and Educational Level
(Each cell shows the mean, variance and sample size)

Marr.	Suva				Urban				Rural			
Dur.	N	LP	UP	S+	N	LP	UP	S+	N	LP	UP	S+
0–4	0.50	1.14	0.90	0.73	1.17	0.85	1.05	0.69	0.97	0.96	0.97	0.74
	1.14	0.73	0.67	0.48	1.06	1.59	0.73	0.54	0.88	0.81	0.80	0.59
	8	21	42	51	12	27	39	51	62	102	107	47
5–9	3.10	2.67	2.04	1.73	4.54	2.65	2.68	2.29	2.44	2.71	2.47	2.24
	1.66	0.99	1.87	0.68	3.44	1.51	0.97	0.81	1.93	1.36	1.30	1.19
	10	30	24	22	13	37	44	21	70	117	81	21
10–14	4.08	3.67	2.90	2.00	4.17	3.33	3.62	3.33	4.14	4.14	3.94	3.33
	1.72	2.31	1.57	1.82	2.97	2.99	1.96	1.52	3.52	3.31	3.28	2.50
	12	27	20	12	18	43	29	15	88	132	50	9
15–19	4.21	4.94	3.15	2.75	4.70	5.36	4.60	3.80	5.06	5.59	4.50	2.00
	2.03	1.46	0.81	0.92	7.40	2.97	3.83	0.70	4.91	3.23	3.29	–
	14	31	13	4	23	42	20	5	114	86	30	1
20–24	5.62	5.06	3.92	2.60	5.36	5.88	5.00	5.33	6.46	6.34	5.74	2.50
	4.15	4.64	4.08	4.30	7.19	4.44	4.33	0.33	8.20	5.72	5.20	0.50
	21	18	12	5	22	25	13	3	117	68	23	2
25–29	6.60	6.74	5.38	2.00	6.52	7.51	7.54	–	7.48	7.81	5.80	–
	12.40	11.66	4.27	–	11.45	10.53	12.60	–	11.34	7.57	7.07	–
	47	27	8	1	46	45	13	–	195	59	10	–

born to each woman as the response, and her marriage duration, type of place of residence and level of education as three discrete predictors or factors.

4.1.2 The Poisson Distribution

A random variable Y is said to have a Poisson distribution with parameter μ if it takes integer values $y = 0, 1, 2, \ldots$ with probability

$$\Pr\{Y = y\} = \frac{e^{-\mu}\mu^y}{y!} \tag{4.1}$$

for $\mu > 0$. The mean and variance of this distribution can be shown to be

$$E(Y) = \text{var}(Y) = \mu.$$

Since the mean is equal to the variance, any factor that affects one will also affect the other. Thus, the usual assumption of homoscedasticity would not be appropriate for Poisson data.

The classic text on probability theory by Feller (1957) includes a number of examples of observations fitting the Poisson distribution, including data on the number of flying-bomb hits in the south of London during World War II. The city was divided into 576 small areas of one-quarter square kilometers each, and the number of areas hit exactly k times was counted. There were a total of 537 hits, so the average number of hits per area was 0.9323. The observed frequencies in Table 4.2 are remarkably close to a Poisson distribution with mean $\mu = 0.9323$. Other examples of events that fit this distribution are radioactive disintegrations, chromosome interchanges in cells, the number of telephone connections to a wrong number, and the number of bacteria in different areas of a Petri plate.

TABLE 4.2: Flying-bomb Hits on London During World War II

Hits	0	1	2	3	4	5+
Observed	229	211	93	35	7	1
Expected	226.7	211.4	98.6	30.6	7.1	1.6

The Poisson distribution can be derived as a limiting form of the binomial distribution if you consider the distribution of the number of successes in a very large number of Bernoulli trials with a small probability of success in each trial. Specifically, if $Y \sim B(n, \pi)$ then the distribution of Y as $n \to \infty$ and $\pi \to 0$ with $\mu = n\pi$ remaining fixed approaches a Poisson distribution with mean μ. Thus, the Poisson distribution provides an approximation to the binomial for the analysis of rare events, where π is small and n is large.

In the flying-bomb example, we can think of each day as one of a large number of trials where each specific area has only a small probability of being hit. Assuming independence across days would lead to a binomial distribution which is well approximated by the Poisson.

An alternative derivation of the Poisson distribution is in terms of a stochastic process described somewhat informally as follows. Suppose events occur randomly in time in such a way that the following conditions obtain:

- The probability of at least one occurrence of the event in a given time interval is proportional to the length of the interval.
- The probability of two or more occurrences of the event in a very small time interval is negligible.
- The numbers of occurrences of the event in disjoint time intervals are mutually independent.

Then the probability distribution of the number of occurrences of the event in a fixed time interval is Poisson with mean $\mu = \lambda t$, where λ is the rate of occurrence of the event per unit of time and t is the length of the time interval. A process satisfying the three assumptions listed above is called a Poisson process.

In the flying bomb example these conditions are not unreasonable. The longer the war lasts, the greater the chance that a given area will be hit at least once. Also, the probability that the same area will be hit twice the same day is, fortunately, very small. Perhaps less obviously, whether an area is hit on any given day is independent of what happens in neighboring areas, contradicting a common belief that bomb hits tend to cluster.

The most important motivation for the Poisson distribution from the point of view of statistical estimation, however, lies in the relationship between the mean and the variance. We will stress this point when we discuss our example, where the assumptions of a limiting binomial or a Poisson process are not particularly realistic, but the Poisson model captures very well the fact that, as is often the case with count data, the variance tends to increase with the mean.

A useful property of the Poisson distribution is that the sum of independent Poisson random variables is also Poisson. Specifically, if Y_1 and Y_2 are independent with $Y_i \sim P(\mu_i)$ for $i = 1, 2$ then

$$Y_1 + Y_2 \sim P(\mu_1 + \mu_2).$$

This result generalizes in an obvious way to the sum of more than two Poisson observations.

An important practical consequence of this result is that we can analyze individual or grouped data with equivalent results. Specifically, suppose we have a group of n_i individuals with identical covariate values. Let Y_{ij} denote the number of events experienced by the j-th unit in the i-th group, and let Y_i denote the total number of events in group i. Then, under the usual assumption of independence, if $Y_{ij} \sim P(\mu_i)$ for $j = 1, 2, \ldots, n_i$, then $Y_i \sim P(n_i\mu_i)$. In words, if the individual counts Y_{ij} are Poisson with mean μ_i, the group total Y_i is Poisson with mean $n_i\mu_i$. In terms of estimation, we obtain exactly the same likelihood function if we work with the individual counts Y_{ij} or the group counts Y_i.

4.1.3 Log-Linear Models

Suppose that we have a sample of n observations $y_1, y_2, \ldots, y_n$ which can be treated as realizations of independent Poisson random variables, with

$Y_i \sim P(\mu_i)$, and suppose that we want to let the mean μ_i (and therefore the variance!) depend on a vector of explanatory variables $\mathbf{x}_i$.

We could entertain a simple linear model of the form

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta},$$

but this model has the disadvantage that the linear predictor on the right hand side can assume any real value, whereas the Poisson mean on the left hand side, which represents an expected count, has to be non-negative.

A straightforward solution to this problem is to model instead the *logarithm* of the mean using a linear model. Thus, we take logs calculating $\eta_i = \log(\mu_i)$ and assume that the transformed mean follows a linear model $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$. Thus, we consider a generalized linear model with link log. Combining these two steps in one we can write the log-linear model as

$$\log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}. \tag{4.2}$$

In this model the regression coefficient β_j represents the expected change in the *log* of the mean per unit change in the predictor x_j. In other words increasing x_j by one unit is associated with an increase of β_j in the log of the mean.

Exponentiating Equation 4.2 we obtain a multiplicative model for the mean itself:

$$\mu_i = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}.$$

In this model, an exponentiated regression coefficient $\exp\{\beta_j\}$ represents a multiplicative effect of the j-th predictor on the mean. Increasing x_j by one unit multiplies the mean by a factor $\exp\{\beta_j\}$.

A further advantage of using the log link stems from the empirical observation that with count data the effects of predictors are often multiplicative rather than additive. That is, one typically observes small effects for small counts, and large effects for large counts. If the effect is in fact proportional to the count, working in the log scale leads to a much simpler model.

4.2 Estimation and Testing

The log-linear Poisson model described in the previous section is a generalized linear model with Poisson error and link log. Maximum likelihood estimation and testing follows immediately from the general results in Appendix B. In this section we review a few key results.

4.2.1 Maximum Likelihood Estimation

The likelihood function for n independent Poisson observations is a product of probabilities given by Equation 4.1. Taking logs and ignoring a constant involving $\log(y_i!)$, we find that the log-likelihood function is

$$\log L(\boldsymbol{\beta}) = \sum \{y_i \log(\mu_i) - \mu_i\},$$

where μ_i depends on the covariates $\mathbf{x}_i$ and a vector of p parameters $\boldsymbol{\beta}$ through the log link of Equation 4.2.

It is interesting to note that the log is the canonical link for the Poisson distribution. Taking derivatives of the log-likelihood function with respect to the elements of $\boldsymbol{\beta}$, and setting the derivatives to zero, it can be shown that the maximum likelihood estimates in log-linear Poisson models satisfy the estimating equations

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\hat{\boldsymbol{\mu}}, \tag{4.3}$$

where $\mathbf{X}$ is the model matrix, with one row for each observation and one column for each predictor, including the constant (if any), $\mathbf{y}$ is the response vector, and $\hat{\boldsymbol{\mu}}$ is a vector of fitted values, calculated from the m.l.e.'s $\hat{\boldsymbol{\beta}}$ by exponentiating the linear predictor $\boldsymbol{\eta} = \mathbf{X}'\hat{\boldsymbol{\beta}}$. This estimating equation arises not only in Poisson log-linear models, but more generally in any generalized linear model with canonical link, including linear models for normal data and logistic regression models for binomial counts. It is not satisfied, however, by estimates in probit models for binary data.

To understand equation 4.3 it helps to consider a couple of special cases. If the model includes a constant, then one of the columns of the model matrix $\mathbf{X}$ is a column of ones. Multiplying this column by the response vector $\mathbf{y}$ produces the sum of the observations. Similarly, multiplying this column by the fitted values $\hat{\boldsymbol{\mu}}$ produces the sum of the fitted values. Thus, in models with a constant one of the estimating equations matches the sum of observed and fitted values. In terms of the example introduced at the beginning of this chapter, fitting a model with a constant would match the total number of children ever born to all women.

As a second example suppose the model includes a discrete factor represented by a series of dummy variables taking the value one for observations at a given level of the factor and zero otherwise. Multiplying this dummy variable by the response vector $\mathbf{y}$ produces the sum of observations at that level of the factor. When this is done for all levels we obtain the so-called *marginal* total. Similarly, multiplying the dummy variable by the fitted values $\hat{\boldsymbol{\mu}}$ produces the sum of the expected or fitted counts at that level. Thus,

in models with a discrete factor the estimating equations match the observed and fitted marginals for the factor. In terms of the example introduced at the outset, if we fit a model that treats marital duration as a discrete factor we would match the observed and fitted total number of children ever born in each category of duration since first marriage.

This result generalizes to higher order terms. Suppose we entertain models with two discrete factors, say A and B. The additive model $A+B$ would reproduce exactly the marginal totals by A or by B. The model with an interaction effect AB would, in addition, match the totals in each combination of categories of A and B, or the AB margin. This result, which will be important in the sequel, is the basis of an estimation algorithm known as *iterative proportional fitting*.

In general, however, we will use the iteratively-reweighted least squares (IRLS) algorithm discussed in Appendix B. For Poisson data with link log, the working dependent variable $\mathbf{z}$ has elements

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i},$$

and the diagonal matrix $\mathbf{W}$ of iterative weights has elements

$$w_{ii} = \hat{\mu}_i,$$

where $\hat{\mu}_i$ denotes the fitted values based on the current parameter estimates.

Initial values can be obtained by applying the link to the data, that is taking the log of the response, and regressing it on the predictors using OLS. To avoid problems with counts of 0, one can add a small constant to all responses. The procedure usually converges in a few iterations.

4.2.2 Goodness of Fit

A measure of discrepancy between observed and fitted values is the deviance. In Appendix B we show that for Poisson responses the deviance takes the form

$$D = 2\sum\{y_i \log(\frac{y_i}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i)\}.$$

The first term is identical to the binomial deviance, representing 'twice a sum of observed times log of observed over fitted'. The second term, a sum of differences between observed and fitted values, is usually zero, because m.l.e.'s in Poisson models have the property of reproducing marginal totals, as noted above.

For large samples the distribution of the deviance is approximately a chi-squared with $n-p$ degrees of freedom, where n is the number of observations and p the number of parameters. Thus, the deviance can be used directly to test the goodness of fit of the model.

An alternative measure of goodness of fit is Pearson's chi-squared statistic, which is defined as

$$\chi^2_p = \sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

The numerator is the squared difference between observed and fitted values, and the denominator is the variance of the observed value. The Pearson statistic has the same form for Poisson and binomial data, namely a 'sum of squared observed minus expected over expected'.

In large samples the distribution of Pearson's statistic is also approximately chi-squared with $n - p$ d.f. One advantage of the deviance over Pearson's chi-squared is that it can be used to compare nested models, as noted below.

4.2.3 Tests of Hypotheses

Likelihood ratio tests for log-linear models can easily be constructed in terms of deviances, just as we did in logistic regression models. In general, the difference in deviances between two nested models has approximately in large samples a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the models, under the assumption that the smaller model is correct.

One can also construct Wald tests as we have done before, based on the fact that the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ has approximately in large samples a multivariate normal distribution with mean equal to the true parameter value $\boldsymbol{\beta}$ and variance-covariance matrix $\text{var}(\hat{\boldsymbol{\beta}}) = \mathbf{X'WX}$, where $\mathbf{X}$ is the model matrix and $\mathbf{W}$ is the diagonal matrix of estimation weights described earlier.

4.3 A Model for Heteroscedastic Counts

Let us consider the data on children ever born from Table 4.1. The unit of analysis here is the individual woman, the response is the number of children she has borne, and the predictors are the duration since her first marriage, the type of place where she resides, and her educational level, classified in four categories.

4.3.1 The Mean-Variance Relation

Data such as these have traditionally been analyzed using ordinary linear models with normal errors. You might think that since the response is a discrete count that typically takes values such as 0, 2 or six, it couldn't possibly have a normal distribution. The key concern, however, is not the normality of the errors but rather the assumption of constant variance.

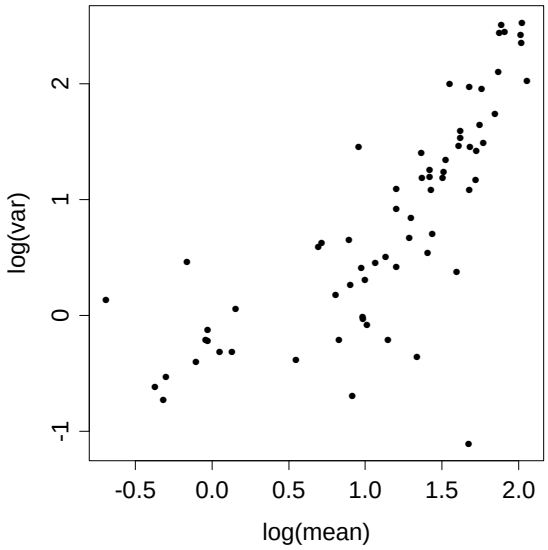


FIGURE 4.1: The Mean-variance Relationship for the CEB Data

In Figure 4.1 we explore the form of the mean-variance relationship for these data by plotting the variance versus the mean for all cells in the table with at least 20 observations. For convenience we use a log-log scale. Clearly, the assumption of constant variance is not valid. Although the variance is not exactly equal to the mean, it is not far from being proportional to it. Thus, we conclude that we can do far more justice to the data by fitting Poisson regression models than by clinging to ordinary linear models.

4.3.2 Grouped Data and the Offset

At this point you may wonder whether we need the individual observations to be able to proceed further. The answer is no; all the information we need is available in Table 4.1. To see this point let Y_{ijkl} denote the number of children borne by the l-th woman in the (i, j, k)-th group, where i denotes

marital duration, j residence and k education, and let $Y_{ijk} = \sum_l Y_{ijkl}$ denote the group total. If each of the observations in this group is a realization of an independent Poisson variate with mean μ_{ijk}, then the group total will be a realization of a Poisson variate with mean $n_{ijk}\mu_{ijk}$, where n_{ijk} is the number of observations in the (i, j, k)-th cell.

Suppose now that you postulate a log-linear model for the individual means, say

$$\log \mathrm{E}(Y_{ijkl}) = \log(\mu_{ijk}) = \mathbf{x}'_{ijk}\boldsymbol{\beta},$$

where $\mathbf{x}_{ijk}$ is a vector of covariates. Then the log of the expected value of the group total is

$$\log \mathrm{E}(Y_{ijk}) = \log(n_{ijk}\mu_{ijk}) = \log(n_{ijk}) + \mathbf{x}'_{ijk}\boldsymbol{\beta}.$$

Thus, the group totals follow a log-linear model with exactly the same coefficients $\boldsymbol{\beta}$ as the individual means, except for the fact that the linear predictor includes the term $\log(n_{ijk})$. This term, which is known beforehand, is called an *offset*, and is a frequent feature of log-linear models for counts of events. Often, when the response is a count of events the offset represents the log of some measure of exposure, in our case the number of women.

Thus, we can analyze the data by fitting log-linear models to the individual counts, or to the group totals. In the latter case we treat the log of the number of women in each cell as an offset. The parameter estimates and standard errors will be exactly the same. The deviances of course, will be different, because they measure goodness of fit to different sets of counts. Differences of deviances between nested models, however, are exactly the same whether one works with individual or grouped data. The situation is analogous to the case of individual and grouped binary data discussed in the previous chapter, with the offset playing a role similar to that of the binomial denominator.

4.3.3 The Deviance Table

Table 4.3 shows the results of fitting a variety of Poisson models to the children ever-born data. The null model has a deviance of 3732 on 69 degrees of freedom (d.f.) and does not fit the data, so we reject the hypothesis that the expected number of children is the same for all these groups.

Introducing marital duration reduces the deviance to 165.8 on 64 d.f. The substantial reduction of 3566 at the expense of only five d.f. reflects the trivial fact that the (cumulative) number of children ever born to a woman depends on the total amount of time she has been exposed to childbearing,

TABLE 4.3: Deviances for Poisson Log-linear Models Fitted to the Data on CEB by Marriage Duration, Residence and Education

Model	Deviance	d.f.
Null	3731.52	69
One-factor Models		
Duration	165.84	64
Residence	3659.23	67
Education	2661.00	66
Two-factor Models		
$D + R$	120.68	62
$D + E$	100.01	61
DR	108.84	52
DE	84.46	46
Three-factor Models		
$D + R + E$	70.65	59
$D + RE$	59.89	53
$E + DR$	57.06	49
$R + DE$	54.91	44
$DR + RE$	44.27	43
$DE + RE$	44.60	38
$DR + DE$	42.72	34
$DR + DE + RE$	30.95	28

as measured by the duration since her first marriage. Clearly it would not make sense to consider any model that does not include this variable as a necessary control.

At this stage one could add to the model type of place of residence, education, or both. The additive model with effects of duration, residence and education has a deviance of 70.65 on 59 d.f. (an average of 1.2 per d.f.) and provides a reasonable description of the data. The associated P-value under the assumption of a Poisson distribution is 0.14, so the model passes the goodness-of-fit test. In the next subsection we consider the interpretation of parameter estimates for this model.

The deviances in Table 4.3 can be used to test the significance of gross and net effects as usual. To test the gross effect of education one could compare the one-factor model with education to the null model, obtaining a remarkable chi-squared statistic of 1071 on three d.f. In this example it really

doesn't make sense to exclude marital duration, which is an essential control for exposure time. A better test of the effect of education would therefore compare the additive model $D + E$ with both duration and education to the one-factor model D with duration only. This gives a more reasonable chi-squared statistic of 65.8 on three d.f., still highly significant. Since educated women tend to be younger, the previous test overstated the educational differential.

We can also test the net effect of education controlling for type of place of residence, by comparing the three-factor additive model $D + R + E$ with the two-factor model $D + R$ with duration and residence only. The difference in deviances of 50.1 on three d.f. is highly significant. The fact that the chi-squared statistic for the net effect is somewhat smaller than the test controlling duration only indicates that part of the effect of education may be attributed to the fact that more educated women tend to live in Suva or in other urban areas.

The question of interactions remains to be raised. Does education make more of a difference in rural areas than in urban areas? To answer this question we move from the additive model to the model that adds an interaction between residence and education. The reduction in deviance is 10.8 on six d.f. and is not significant, with a P-value of 0.096. Does the effect of education increase with marital duration? Adding an education by duration interaction to the additive model reduces the deviance by 15.7 at the expense of 15 d.f., hardly a bargain. A similar remark applies to the residence by duration interaction. Thus, we conclude that the additive model is adequate for these data.

4.3.4 The Additive Model

Table 4.4 shows parameter estimates and standard errors for the additive model of children ever born (CEB) by marital duration, type of place of residence and education.

The constant represents the log of the mean number of children for the reference cell, which in this case is Suvanese women with no education who have been married 0–4 years. Since $\exp\{-0.1173\} = 0.89$, we see that on the average these women have 0.89 children at this time in their lives. The duration parameters trace the increase in CEB with duration for any residence-education group. As we move from duration 0–4 to 5–9 the log of the mean increases by almost one, which means that the number of CEB gets multiplied by $\exp\{0.9977\} = 2.71$. By duration 25–29, women in each category of residence and education have $\exp\{1.977\} = 7.22$ times as many children as

TABLE 4.4: Estimates for Additive Log-Linear Model of Children Ever Born by Marital Duration, Type of Place of Residence and Educational Level

Parameter		Estimate	Std. Error	z-ratio
Constant		−0.1173	0.0549	−2.14
Duration	0–4	–		
	5–9	0.9977	0.0528	18.91
	10–14	1.3705	0.0511	26.83
	15–19	1.6142	0.0512	31.52
	20–24	1.7855	0.0512	34.86
	25–29	1.9768	0.0500	39.50
Residence	Suva	–		
	Urban	0.1123	0.0325	3.46
	Rural	0.1512	0.0283	5.34
Education	None	–		
	Lower	0.0231	0.0227	1.02
	Upper	-0.1017	0.0310	-3.28
	Sec+	-0.3096	0.0552	-5.61

they did at duration 0–4.

The effects of residence show that Suvanese women have the lowest fertility. At any given duration since first marriage, women living in other urban areas have 12% larger families ($\exp\{0.1123\} = 1.12$) than Suvanese women with the same level of education. Similarly, at any fixed duration, women who live in rural areas have 16% more children ($\exp\{0.1512\} = 1.16$), than Suvanese women with the same level of education.

Finally, we see that higher education is associated with smaller family sizes net of duration and residence. At any given duration of marriage, women with upper primary education have 10% fewer kids, and women with secondary or higher education have 27% fewer kids, than women with no education who live in the same type of place of residence. (The last figure follows from the fact that $1 - \exp\{-0.3096\} = 0.27$.)

In our discussion of interactions in the previous subsection we noted that the additive model fits reasonably well, so we have no evidence that the effect of a variable depends on the values of other predictors. It is important to note, however, that the model is additive in the *log* scale. In the original scale the model is multiplicative, and postulates relative effects which translate into different absolute effects depending on the values of the

other predictors. To clarify this point we consider the effect of education. Women with secondary or higher education have 27% fewer kids than women with no education. Table 4.5 shows the predicted number of children at each duration of marriage for Suvanese women with secondary education and with no education, as well as the difference between these two groups.

TABLE 4.5: Fitted Values for Suvanese Women with No Education and with Secondary or Higher Education

Marital Duration	0–4	5–9	10–14	15–19	20–24	25+
No Education	0.89	2.41	3.50	4.47	5.30	6.42
Secondary+	0.65	1.77	2.57	3.28	3.89	4.71
Difference	0.24	0.64	0.93	1.19	1.41	1.71

The educational differential of 27% between these two groups translates into a quarter of a child at durations 0–4, increases to about one child around duration 15, and reaches almost one and a quarter children by duration 25+. Thus, the (absolute) effect of education measured in the original scale increases with marital duration.

If we had used an ordinary linear regression model for these data we would have ended up with a large number of interaction effects to accommodate the fact that residence and educational differentials increase with marital duration. In addition, we would have faced a substantial problem of heteroscedasticity. Taking logs of the response would ameliorate the problem, but would have required special treatment of women with no children. The Poisson log-linear model solves the two problems separately, allowing the variance to depend on the mean, and modeling the log of the mean as a linear function of the covariates.

Models for Count Data With Overdispersion

Germán Rodríguez

November 6, 2013

Abstract

This addendum to the WWS 509 notes covers extra-Poisson variation and the negative binomial model, with brief appearances by zero-inflated and hurdle models.

1 Extra-Poisson Variation

One of the key features of the Poisson distribution is that the variance equals the mean, so

$$\text{var}(Y) = E(Y) = \mu$$

Empirically, however, we often find data that exhibit over-dispersion, with a variance larger than the mean. The Stata logs show an example from Long (1990), involving the number of publications produced by Ph.D. biochemists. As the analysis shows, there's evidence that the variance is about 1.8 times the mean. We now consider models that accommodate the excess variation.

An interesting feature of the iteratively-reweighted least squares (IRLS) algorithm used in generalized linear models is that it depends only on the mean and variance of the observations. Nelder and Wedderburn (1972) proposed specifying just the mean and variance relationship and then applying the algorithm. The resulting estimates are called maximum quasi-likelihood estimates (MQLE), and have been shown to share many of the optimality properties of maximum likelihood estimates (MLE) under fairly general conditions.

In the context of count data, consider the assumption that the variance is *proportional* to the mean. Specifically,

$$\text{var}(Y) = \phi E(Y) = \phi\mu$$

If $\phi = 1$ then the variance equals the mean and we obtain the Poisson mean-variance relationship. If $\phi > 1$ then the have over-dispersion relative to Poisson. If $\phi < 1$ we would have under-dispersion, but this is relatively rare.

It turns out that applying the IRLS algorithm in this more general case involves working with weights $w^* = \mu/\phi$. These are the Poisson weights $w = \mu$ divided by ϕ, but the ϕ cancels out when we compute the weighted estimator $(X'WX)^{-1}X'Wy$, which thus reduces to the Poisson MLE. This implies that Poisson estimates are MQLE when the variance is proportional (not just equal) to the mean.

The variance of the estimator in the more general case involves ϕ, and is given by

$$\text{var}(\hat{\beta}) = \phi(X'WX)^{-1}$$

where $W = \text{diag}(\mu_1, \ldots, \mu_n)$, reducing to the Poisson variance when $\phi = 1$. This means that Poisson standard errors will be conservative in the presence of over-dispersion.

The obvious solution is to correct the standard errors using an estimate of ϕ. The standard approach relies on Pearson's χ^2 statistic, which is defined as

$$\chi^2_p = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\text{var}(y_i)} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\phi\mu_i}$$

If the model is correct the expected value of this statistic is $n - p$. Equating the statistic to its expectation and solving for ϕ gives the estimate

$$\hat{\phi} = \frac{\chi^2_p}{n - p}$$

In the biochemist example, $\chi^2_p = 1662.55$ for a model with $p = 6$ parameters on $n = 915$ observations, which leads to $\hat{\phi} = 1662.55/909 = 1.829$. We retain the Poisson estimates but multiply the standard errors by $\sqrt{1.829} = 1.352$, which inflates them by 35.2%.

A word of caution is in order. Normally one would consider a large value of Pearson's χ^2 as evidence of lack of fit. What we are doing here is treating it as pure error to inflate our standard errors. Obviously this requires a high degree of confidence in the systematic part of the model, so we can be reasonably sure that the lack of fit is not due to specification errors but just over-dispersion.

An alternative approach that often gives similar results is to use a robust estimator of standard errors in the tradition of Huber (1956) and White (1980).

2 Negative Binomial

An alternative approach to modeling over-dispersion in count data is to start from a Poisson regression model and add a multiplicative *random effect* θ to represent unobserved heterogeneity. This leads to the negative binomial regression model.

Suppose that the *conditional* distribution of the outcome Y given an unobserved variable θ is indeed Poisson with mean and variance $\theta\mu$, so

$$Y|\theta \sim P(\mu_\theta)$$

In the context of the Ph.D. biochemists, θ captures unobserved factors that increase (if $\theta > 1$) or decrease (if $\theta < 1$) productivity relative to what one would expect given the observed values of the covariates, which is of course μ where $\log \mu = x'\beta$. For convenience we take $E(\theta) = 1$, so μ represents the expected outcome for the average individual given covariates x.

In this model the data *would* be Poisson if only we could observe θ. Unfortunately we do not. Instead we make an assumption regarding its distribution and "integrate it out" of the likelihood, effectively computing the unconditional distribution of the outcome. It turns out to be mathematically convenient to assume that θ has a gamma distribution with parameters α and β. This distribution has mean α/β and variance α/β^2, so we take $\alpha = \beta = 1/\sigma^2$, which makes the mean of the unobserved effect one and its variance σ^2.

With this information we can compute the *unconditional* distribution of the outcome, which happens to be the negative binomial distribution. The density is best written in terms of the parameters α, β and μ as done below, although you must recall that in our case $\alpha = \beta = 1/\sigma^2$, so there's just one more parameter compared to a Poisson model.

$$\Pr\{Y = y\} = \frac{\Gamma(\alpha+y)}{y!\Gamma(\alpha)}\frac{\beta^\alpha \mu^y}{(\mu+\beta)^{\alpha+y}}$$

This distribution is best known as the distribution of the number of failures before the k-th success in a series of independent Bernoulli trials with common probability of success π. The density corresponding to this interpretation can be obtained from the expression above by setting $\alpha = k$ and $\pi = \beta/(\mu+\beta)$.

The negative binomial distribution with $\alpha = \beta = 1/\sigma^2$ has mean

$$E(Y) = \mu$$

and variance

$$\text{var}(Y) = \mu(1 + \sigma^2\mu)$$

If $\sigma^2 = 0$ there's no unobserved heterogeneity and we obtain the Poisson variance. If $\sigma^2 > 0$ then the variance is larger than the mean. Thus, the negative binomial distribution is over-dispersed relative to the Poisson.

Interestingly, these moments can be derived using the law of iterated expectations without assuming that the unobservable has a gamma distribution; all we need are the conditional moments $E(Y|\theta) = \text{var}(Y|\theta) = \theta\mu$ and the assumption that the unobservable has mean one and variance σ^2. The unconditional mean is simply the expected value of the conditional mean

$$E(Y) = E_\theta[E_{Y|\theta}(Y|\theta)] = E_\theta(\theta\mu) = \mu E_\theta(\theta) = \mu$$

where we used subscripts to clarify over which distribution we are taking expectations. The unconditional variance is the expected value of the conditional variance plus the variance of the conditional mean

$$\begin{aligned}
\text{var}(Y) &= E_\theta[\text{var}_{Y|\theta}(Y|\theta)] + \text{var}_\theta[E_\theta(\theta\mu)] \\
&= E_\theta(\theta\mu) + \text{var}_\theta(\theta\mu) = \mu E_\theta(\theta) + \mu^2\text{var}_\theta(\theta) \\
&= \mu + \mu^2\sigma^2 = \mu(1 + \mu\sigma^2)
\end{aligned}$$

again using a subscript to clarify over which distribution we are taking expectation or computing variance.

Stata has a command called `nbreg` that can fit the negative binomial model described here by maximum likelihood. The output uses `alpha` to label the variance of the unobservable, which we call σ^2. The model can also be fit by maximum quasi-likelihood using only the mean-variance relationship, provided σ^2 is known or estimated separately. It is possible to derive an estimate of σ^2 using Pearson's χ^2 statistic, but this requires alternating between estimating μ given σ^2 and then σ^2 given μ, so this approach loses the simplicity it has in the Poisson case.

Because the Poisson model is a special case of the negative binomial when $\sigma^2 = 0$, we can use a likelihood ratio test to compare the two models. There is, however, a small difficulty. Because the null hypothesis corresponding to the Poisson model is on a boundary of the parameter space, the likelihood ratio test statistic does not converge to a χ^2 distribution with one d.f as one might expect. Simulations suggest that the null distribution is better approximated as a 50:50 mixture of zero and a χ^2 with one d.f., and this is the approximation that Stata reports as `chibar2`. An alternative is simply

to note that treating the test statistic as χ^2 with one d.f. results in a conservative test.

For the Ph.D. Biochemist data in the Stata log the negative binomial model gives estimates that are very similar to the Poisson model, and have of course the same interpretation with the caveat that we are talking about an individual with average unobserved characteristics. The standard errors are very similar to the over-dispersed Poisson standard errors, and are both larger than the reference Poisson errors. The test that $\sigma^2 = 2$ leads to a $\chi^2_{LR} = 180.2$, which is highly significant with or without adjustment for the boundary problem.

We note in closing this section that there are alternative formulations of the negative binomial model that lead to different models, including the over-dispersed Poisson model of the previous section. The formulation given here, however, is the one in common use.

3 Zero-Inflated Models

Another common problem with count data models, including both Poisson and negative binomial models, is that empirical data often show more zeroes than would be expected under either model. In the Ph.D. Biochemist data, for example, we find that 30% of the individuals publish no papers at all. The Poisson model predicts 21%, so it underestimates the zeroes by nine percentage points.

The zero-inflated Poisson model postulates that there are two latent classes of people. The "always zero", which in our example would be individuals who never publish, and the rest, or "not always zero", for whom the number of publications has a Poisson distribution with mean and variance $\mu > 0$. The model combines a logit model that predicts which of the two latent classes a person belongs, with a Poisson model that predicts the outcome for those in the second latent class. In this model there are two kinds of zeroes: some are structural zeroes from the always zero class, and some are random zeroes from the other class.

Stata can fit this model using the `zip` or zero-inflated Poisson command. You specify the predictors for the Poisson equation in the usual fashion, and use the `inflate()` option to specify the predictors for the logit equation. The inflate part can be `inflate(_cons)` to specify that everyone has the same probability of belonging to the zero class. Stata is rather finicky in determining which models are nested, and will not allow direct comparison of the zip and Poisson models, although they are nested, with the Poisson

model corresponding to the zip model where the probability of "always zero" is zero for everybody.

Stata can also fit a zero-inflated negative binomial model using the command `zinb`, which combines a logit equation for the latent classes with a negative binomial for the counts in the not "always zero" class. One can view this model as adding unobserved heterogeneity to the Poisson equation for the counts in the second latent class.

These models are very appealing but interpretation is not always straightforward. In an analysis of number of cigarettes smoked last week the latent classes have a natural interpretation: the "always zero" are non-smokers, to be distinguished from smokers who happen to smoke no cigarettes in a week, but one would be better off ascertaining whether the respondent smokes. When analyzing publications the "always zero" could be respondents in non-academic jobs where publishing is not required or expected, but again it would be better to include the type of job as a covariate. In health research it is common to ask elderly respondents how many limitations they encounter in carrying out activities of daily living, such as getting up from a chair or climbing some stairs. It is common to observe more zeroes than expected, but it is not clear what an "always zero" class would mean.

4 Hurdle Models

Another approach to excess zeroes is to use a logit model to distinguish counts of zero from larger counts, effectively collapsing the count distribution into two categories, and then use a *truncated* Poisson model, namely a Poisson distribution where zero has been excluded, for the positive counts. This approach differs from the zip models because the classes are observed rather than latent, one consists of observed zeroes and the other of observed positive counts. The term "hurdle" is evocative of a threshold that must be exceeded before events occur, with a separate process determining the number of events.

Stata can fit a zero-truncated Poisson model using the `ztp` command. To fit a hurdle model you create an indicator variable for counts above zero and run an ordinary logit, and then run `ztp` on the respondents with positive counts. To obtain an overall log-likelihood you just add the log-likelihoods of the two parts.

Interpretation of the logit equation is more straightforward than in zip models because the binary choice is clear; you are modeling whether people smoke, publish, or have limitations in activities of daily living. Interpreta-

tion of the Poisson equation, however, is not so obvious because the quantity being modeled, μ, is the mean of the entire distribution, not the mean for those who experience events, which would be $\mu/(1-e^{-\mu})$. This means that a coefficient such as $\beta = 0.1$ cannot be interpreted as reflecting a 10% increase in the mean, and the simplicity of the Poisson multiplicative model is lost.

An alternative approach is to compute the derivative of the fitted values with respect to the predictors and interpret results in terms of marginal effects.

Chapter 5

Log-Linear Models for Contingency Tables

In this chapter we study the application of Poisson regression models to the analysis of contingency tables. This is perhaps one of the most popular applications of log-linear models, and is based on the existence of a very close relationship between the multinomial and Poisson distributions.

5.1 Models for Two-dimensional Tables

We start by considering the simplest possible contingency table: a two-by-two table. However, the concepts to be introduced apply equally well to more general two-way tables where we study the joint distribution of two categorical variables.

5.1.1 The Heart Disease Data

Table 5.1 was taken from the Framingham longitudinal study of coronary heart disease (Cornfield, 1962; see also Fienberg, 1977). It shows 1329 patients cross-classified by the level or their serum cholesterol (below or above 260) and the presence or absence of heart disease.

There are various sampling schemes that could have led to these data, with consequences for the probability model one would use, the types of questions one would ask, and the analytic techniques that would be employed. Yet, all schemes lead to equivalent analyses. We now explore several approaches to the analysis of these data.

TABLE 5.1: Serum Cholesterol and Heart Disease

Serum Cholesterol	Heart Disease		Total
	Present	Absent	
< 260	51	992	1043
260+	41	245	286
Total	92	1237	1329

5.1.2 The Multinomial Model

Our first approach will assume that the data were collected by sampling 1329 patients who were then classified according to cholesterol and heart disease. We view these variables as two responses, and we are interested in their joint distribution. In this approach the total sample size is assumed fixed, and all other quantities are considered random.

We will develop the random structure of the data in terms of the row and column variables, and then note what this implies for the counts themselves. Let C denote serum cholesterol and D denote heart disease, both discrete factors with two levels. More generally, we can imagine a row factor with I levels indexed by i and a column factor with J levels indexed by j, forming an $I \times J$ table. In our example $I = J = 2$.

To describe the joint distribution of these two variables we let π_{ij} denote the probability that an observation falls in row i and column j of the table. In our example words, π_{ij} is the probability that serum cholesterol C takes the value i and heart disease D takes the value j. In symbols,

$$\pi_{ij} = \Pr\{C = i, D = j\}, \tag{5.1}$$

for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. These probabilities completely describe the *joint* distribution of the two variables.

We can also consider the *marginal* distribution of each variable. Let $\pi_{i.}$ denote the probability that the row variable takes the value i, and let $\pi_{.j}$ denote the probability that the column variable takes the value j. In our example $\pi_{i.}$ and $\pi_{.j}$ represent the marginal distributions of serum cholesterol and heart disease. In symbols,

$$\pi_{i.} = \Pr\{C = i\} \qquad \text{and} \qquad \pi_{.j} = \Pr\{D = j\}. \tag{5.2}$$

Note that we use a dot as a placeholder for the omitted subscript.

The main hypothesis of interest with two responses is whether they are *independent*. By definition, two variables are independent if (and only if)

their joint distribution is the product of the marginals. Thus, we can write the hypothesis of independence as

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j} \tag{5.3}$$

for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$. The question now is how to estimate the parameters and how to test the hypothesis of independence.

The traditional approach to testing this hypothesis calculates expected counts under independence and compares observed and expected counts using Pearson's chi-squared statistic. We adopt a more formal approach that relies on maximum likelihood estimation and likelihood ratio tests. In order to implement this approach we consider the distribution of the counts in the table.

Suppose each of n observations is classified independently in one of the IJ cells in the table, and suppose the probability that an observation falls in the (i, j)-th cell is π_{ij}. Let Y_{ij} denote a random variable representing the number of observations in row i and column j of the table, and let y_{ij} denote its observed value. The joint distribution of the counts is then the *multinomial* distribution, with

$$\Pr\{\mathbf{Y} = \mathbf{y}\} = \frac{n!}{y_{11}!y_{12}!y_{21}!y_{22}!}\pi_{11}^{y_{11}}\pi_{12}^{y_{12}}\pi_{21}^{y_{21}}\pi_{22}^{y_{22}}, \tag{5.4}$$

where $\mathbf{Y}$ is a random vector collecting all four counts and $\mathbf{y}$ is a vector of observed values. The term to the right of the fraction represents the probability of obtaining y_{11} observations in cell (1,1), y_{12} in cell (1,2), and so on. The fraction itself is a combinatorial term representing the number of ways of obtaining y_{11} observations in cell (1,1), y_{12} in cell (1,2), and so on, out of a total of n. The multinomial distribution is a direct extension of the binomial distribution to more than two response categories. In the present case we have four categories, which happen to represent a two-by-two structure. In the special case of only two categories the multinomial distribution reduces to the familiar binomial.

Taking logs and ignoring the combinatorial term, which does not depend on the parameters, we obtain the multinomial log-likelihood function, which for a general $I \times J$ table has the form

$$\log L = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}\log(\pi_{ij}). \tag{5.5}$$

To estimate the parameters we need to take derivatives of the log-likelihood function with respect to the probabilities, but in doing so we must take into

account the fact that the probabilities add up to one over the entire table. This restriction may be imposed by adding a Lagrange multiplier, or more simply by writing the last probability as the complement of all others. In either case, we find the unrestricted maximum likelihood estimate to be the sample proportion:

$$\hat{\pi}_{ij} = \frac{y_{ij}}{n}.$$

Substituting these estimates into the log-likelihood function gives its unrestricted maximum.

Under the hypothesis of independence in Equation 5.3, the joint probabilities depend on the margins. Taking derivatives with respect to $\pi_{i.}$ and $\pi_{.j}$, and noting that these are also constrained to add up to one over the rows and columns, respectively, we find the m.l.e.'s

$$\hat{\pi}_{i.} = \frac{y_{i.}}{n} \qquad \text{and} \qquad \hat{\pi}_{.j} = \frac{y_{.j}}{n},$$

where $y_{i.} = \sum_j y_{ij}$ denotes the row totals and $y_{.j}$ denotes the column totals. Combining these estimates and multiplying by n to obtain expected counts gives

$$\hat{\mu}_{ij} = \frac{y_{i.}y_{.j}}{n},$$

which is the familiar result from introductory statistics. In our example, the expected frequencies are

$$\hat{\mu}_{ij} = \begin{pmatrix} 72.2 & 970.8 \\ 19.8 & 266.2 \end{pmatrix}.$$

Substituting these estimates into the log-likelihood function gives its maximum under the restrictions implied by the hypothesis of independence. To test this hypothesis, we calculate twice the difference between the unrestricted and restricted maxima of the log-likelihood function, to obtain the deviance or likelihood ratio test statistic

$$D = 2\sum_i \sum_j y_{ij} \log(\frac{y_{ij}}{\hat{\mu}_{ij}}). \tag{5.6}$$

Note that the numerator and denominator inside the log can be written in terms of estimated probabilities or counts, because the sample size n cancels out. Under the hypothesis of independence, this statistic has approximately in large samples a chi-squared distribution with $(I-1)(J-1)$ d.f.

Going through these calculations for our example we obtain a deviance of 26.43 with one d.f. Comparison of observed and fitted counts in terms of

Pearson's chi-squared statistic gives 31.08 with one d.f. Clearly, we reject the hypothesis of independence, concluding that heart disease and serum cholesterol level are associated.

5.1.3 The Poisson Model

An alternative model for the data in Table 5.1 is to treat the four counts as realizations of independent Poisson random variables. A possible physical model is to imagine that there are four groups of people, one for each cell in the table, and that members from each group arrive randomly at a hospital or medical center over a period of time, say for a health check. In this model the total sample size is not fixed in advance, and all counts are therefore random.

Under the assumption that the observations are independent, the joint distribution of the four counts is a product of Poisson distributions

$$\Pr\{\mathbf{Y}=\mathbf{y}\} = \prod_i \prod_j \frac{\mu_{ij}^{y_{ij}} e^{-\mu_{ij}}}{y_{ij}!}. \tag{5.7}$$

Taking logs we obtain the usual Poisson log-likelihood from Chapter 4.

In terms of the systematic structure of the model, we could consider three log-linear models for the expected counts: the null model, the additive model and the saturated model. The null model would assume that all four kinds of patients arrive at the hospital or health center in the same numbers. The additive model would postulate that the arrival rates depend on the level of cholesterol and the presence or absence of heart disease, but not on the combination of the two. The saturated model would say that each group has its own rate or expected number of arrivals.

At this point you may try fitting the Poisson additive model to the four counts in Table 5.1, treating cholesterol and heart disease as factors or discrete predictors. You will discover that the deviance is 26.43 on one d.f. (four observations minus three parameters, the constant and the coefficients of two dummies representing cholesterol and heart disease). If you print the fitted values you will discover that they are exactly the same as in the previous subsection.

This result, of course, is not a coincidence. *Testing the hypothesis of independence in the multinomial model is exactly equivalent to testing the goodness of fit of the Poisson additive model.* A rigorous proof of this result is beyond the scope of these notes, but we can provide enough information to show that the result is intuitively reasonable and to understand when it can be used.

First, note that if the four counts have independent Poisson distributions, their sum is distributed Poisson with mean equal to the sum of the means. In symbols, if $Y_{ij} \sim P(\mu_{ij})$ then the total $Y_{\cdot\cdot} = \sum_i \sum_j Y_{ij}$ is distributed Poisson with mean $\mu_{\cdot\cdot} = \sum_i \sum_j \mu_{ij}$. Further, the conditional distribution of the four counts *given* their total is multinomial with probabilities

$$\pi_{ij} = \mu_{ij}/n,$$

where we have used n for the observed total $y_{\cdot\cdot} = \sum_{i,j} y_{ij}$. This result follows directly from the fact that the conditional distribution of the counts $\mathbf{Y}$ given their total $Y_{\cdot\cdot}$ can be obtained as the ratio of the joint distribution of the counts and the total (which is the same as the joint distribution of the counts, which imply the total) to the marginal distribution of the total. Dividing the joint distribution given in Equation 5.7 by the marginal, which is Poisson with mean $\mu_{\cdot\cdot}$, leads directly to the multinomial distribution in Equation 5.4.

Second, note that the systematic structure of the two models is the same. In the model of independence the joint probability is the product of the marginals, so taking logs we obtain

$$\log \pi_{ij} = \log \pi_{i\cdot} + \log \pi_{\cdot j},$$

which is exactly the structure of the additive Poisson model

$$\log \mu_{ij} = \eta + \alpha_i + \beta_j.$$

In both cases the log of the expected count depends on the row and the column but not the combination of the two. In fact, it is only the constraints that differ between the two models. The multinomial model restricts the joint and marginal probabilities to add to one. The Poisson model uses the reference cell method and sets $\alpha_1 = \beta_1 = 0$.

If the systematic and random structure of the two models are the same, then it should come as no surprise that they produce the same fitted values and lead to the same tests of hypotheses. There is only one aspect that we glossed over: the equivalence of the two distributions holds conditional on n, but in the Poisson analysis the total n is random and we have not conditioned on its value. Recall, however, that the Poisson model, by including the constant, reproduces exactly the sample total. It turns out that we don't need to condition on n because the model reproduces its exact value anyway.

The morale of this long-winded story is that we do not need to bother with multinomial models and can always resort to the equivalent Poisson

model. While the gain is trivial in the case of a two-by-two table, it can be very significant as we move to cross-classifications involving three or more variables, particularly as we don't have to worry about maximizing the multinomial likelihood under constraints. The only trick we need to learn is how to translate the questions of independence that arise in the multinomial context into the corresponding log-linear models in the Poisson context.

5.1.4 The Product Binomial*

(On first reading you may wish to skip this subsection and the next and proceed directly to the discussion of three-dimensional tables in Section 5.2.)

There is a third sampling scheme that may lead to data such as Table 5.1. Suppose that a decision had been made to draw a sample of 1043 patients with low serum cholesterol and an independent sample of 286 patients with high serum cholesterol, and then examine the presence or absence of heart disease in each group.

Interest would then focus on the *conditional* distribution of heart disease given serum cholesterol level. Let π_i denote the probability of heart disease at level i of serum cholesterol. In the notation of the previous subsections,

$$\pi_i = \Pr\{D = 1|C = i\} = \frac{\pi_{i1}}{\pi_{i\cdot}},$$

where we have used the fact that the conditional probability of falling in column one given that you are in row i is the ratio of the joint probability π_{i1} of being in cell (i,1) to the marginal probability $\pi_{i\cdot}$ of being in row i.

Under this scheme the row margin would be fixed in advance, so we would have n_1 observations with low cholesterol and n_2 with high. The number of cases with heart disease in category y of cholesterol, denoted Y_{i1}, would then have a binomial distribution with parameters π_i and n_i independently for $i = 1, 2$. The likelihood function would then be a product of two binomials:

$$\Pr\{\mathbf{Y} = \mathbf{y}\} = \frac{n_1!}{y_{11}!y_{12}!}\pi_1^{y_{11}}(1-\pi_1)^{y_{12}}\frac{n_2!}{y_{21}!y_{22}!}\pi_2^{y_{21}}(1-\pi_2)^{y_{22}}, \qquad (5.8)$$

where we have retained double subscripts and written y_{i1} and y_{i2} instead of the more familiar y_i and $n_i - y_i$ to facilitate comparison with Equations 5.4 and 5.7.

The main hypothesis of interest would be the hypothesis of *homogeneity*, where the probability of heart disease is the same in the two groups:

$$H_o : \pi_1 = \pi_2.$$

To test this hypothesis you might consider fitting logistic regression models to the data, treating heart disease as the response and serum cholesterol as the predictor, and working with two observations representing the two groups. If you try this, you will discover that the deviance for the null model, which can be interpreted as a likelihood ratio test of the hypothesis of homogeneity, is 26.43 with one d.f., and coincides with the multinomial and Poisson deviances of the previous two subsections.

Again, this is no coincidence, because the random and systematic components of the models are equivalent. The product binomial distribution in Equation 5.8 can be obtained starting from the assumption that the four counts Y_{ij} are independent Poisson with means μ_{ij}, and then conditioning on the totals $Y_{i.} = \sum_j Y_{ij}$, which are Poisson with means $\mu_{i.} = \sum_j \mu_{ij}$, for $i = 1, 2$. Taking the ratio of the joint distribution of the counts to the marginal distribution of the two totals leads to the product binomial in Equation 5.8 with $\pi_i = \mu_{i1}/\mu_{i.}$.

Similarly, the hypothesis of homogeneity turns out to be equivalent to the hypothesis of independence and hence the additive log-linear model. To see this point note that if two variables are independent, then the conditional distribution of one given the other is the same as its marginal distribution. In symbols, if $\pi_{ij} = \pi_{i.}\pi_{.j}$ then the conditional probability, which in general is $\pi_{j|i} = \pi_{ij}/\pi i.$, simplifies to $\pi_{j|i} = \pi_{.j}$, which does not depend on i. In terms of our example, under independence or homogeneity the conditional probability of heart disease is the same for the two cholesterol groups.

Again, note that the binomial and Poisson models are equivalent conditioning on the row margin, but in fitting the additive log-linear model we did not impose any conditions. Recall, however, that the Poisson model, by treating serum cholesterol as a factor, reproduces exactly the row margin of the table. Thus, it does not matter that we do not condition on the margin because the model reproduces its exact value anyway.

The importance of this result is that the results of our analyses are in fact independent of the sampling scheme.

- If the row margin is fixed in advance we can treat the row factor as a predictor and the column factor as a response and fit a model of homogeneity using the product binomial likelihood.

- If the total is fixed in advance we can treat both the row and column factors as responses and test the hypothesis of independence using the multinomial likelihood.

- Or we can treat all counts as random and fit an additive log-linear

model using the Poisson likelihood.

Reassuringly, the results will be identical in all three cases, both in terms of fitted counts and in terms of the likelihood ratio statistic.

Note that if the total is fixed and the sampling scheme is multinomial we can always condition on a margin and use binomial models, the choice being up to the investigator. This choice will usually depend on whether one wishes to treat the two variables symmetrically, assuming they are both responses and studying their correlation, or asymmetrically, treating one as a predictor and the other as a response in a regression framework.

If the row margin is fixed and the sampling scheme is binomial then we must use the product binomial model, because we can not estimate the joint distribution of the two variables without further information.

5.1.5 The Hypergeometric Distribution*

There is a fourth distribution that could apply to the data in Table 5.1, namely the hypergeometric distribution. This distribution arises from treating both the row and column margins as fixed. I find it hard to imagine a sampling scheme that would lead to fixed margins, but one could use the following conditioning argument.

Suppose that the central purpose of the enquiry is the possible association between cholesterol and heart disease, as measured, for example, by the odds ratio. Clearly, the total sample size has no information about the odds ratio, so it would make sense to condition on it. Perhaps less obviously, the row and column margins carry very little information about the association between cholesterol and heart disease as measured by the odds ratio. It can therefore be argued that it makes good statistical sense to condition on both margins.

If we start from the assumption that the four counts are independent Poisson with means μ_{ij}, and then condition on the margins $Y_{i\cdot}$ and $Y_{\cdot j}$ as well as the total $Y_{\cdot\cdot}$ (being careful to use $Y_{1\cdot}, Y_{\cdot 1}$ and $Y_{\cdot\cdot}$ to maintain independence) we obtain the hypergeometric distribution, where

$$\Pr\{\mathbf{Y}=\mathbf{y}\} = \frac{y_{\cdot 1}!}{y_{11}!y_{21}!}\frac{y_{\cdot 2}!}{y_{21}!y_{22}!}\Big/\frac{n!}{y_{1\cdot}!y_{2\cdot}!}.$$

In small samples this distribution is the basis of the so-called Fisher's exact test for the two-by-two table. McCullagh and Nelder (1989, Sections 7.3–7.4) discuss a conditional likelihood ratio test that differs from the unconditional one. The question of whether one should use conditional or unconditional tests is still a matter of controversy, see for example Yates (1934, 1984). We will not consider the hypergeometric distribution further.

5.2 Models for Three-Dimensional Tables

We now consider in more detail linear models for three-way contingency tables, focusing on testing various forms of complete and partial independence using the equivalent Poisson models.

5.2.1 Educational Aspirations in Wisconsin

Table 5.2 classifies 4991 Wisconsin male high school seniors according to socio-economic status (low, lower middle, upper middle, and high), the degree of parental encouragement they receive (low and high) and whether or not they have plans to attend college (no, yes). This is part of a larger table found in Fienberg (1977, p. 101).

TABLE 5.2: Socio-economic Status, Parental Encouragement and Educational Aspirations of High School Seniors

Social Stratum	Parental Encouragement	College Plans		Total
		No	Yes	
Lower	Low	749	35	784
	High	233	133	366
Lower Middle	Low	627	38	665
	High	330	303	633
Upper Middle	Low	420	37	457
	High	374	467	841
Higher	Low	153	26	179
	High	266	800	1066
Total		3152	1938	4991

In our analysis of these data we will view all three variables as responses, and we will study the extent to which they are associated. In this process we will test various hypotheses of complete and partial independence.

Let us first introduce some notation. We will use three subscripts to identify the cells in an $I \times J \times K$ table, with i indexing the I rows, j indexing the J columns and k indexing the K layers. In our example $I = 4, J = 2$, and $K = 2$ for a total of 16 cells.

Let π_{ijk} denote the probability that an observation falls in cell (i, j, k). In our example, this cell represents category i of socio-economic status (S), category j of parental encouragement (E) and category k of college plans (P). These probabilities define the joint distribution of the three variables.

We also let y_{ijk} denote the observed count in cell (i, j, k), which we treat as a realization of a random variable Y_{ijk} having a multinomial or Poisson distribution.

We will also use the dot convention to indicate summing over a subscript, so $\pi_{i..}$ is the marginal probability that an observation falls in row i and $y_{i..}$ is the number of observations in row i. The notation extends to two dimensions, so $\pi_{ij.}$ is the marginal probability that an observation falls in row i and column$j$ and $y_{ij.}$ is the corresponding count.

5.2.2 Deviances for Poisson Models

In practice we will treat the Y_{ijk} as independent Poisson random variables with means $\mu_{ijk} = n\pi_{ijk}$, and we will fit log-linear models to the expected counts.

Table 5.3 lists all possible models of interest in the Poisson context that include all three variables, starting with the three-factor additive model $S + E + P$ on status, encouragement and plans, and moving up towards the saturated model SEP. For each model we list the abbreviated model formula, the deviance and the degrees of freedom.

TABLE 5.3: Deviances for Log-linear Models
Fitted to Educational Aspirations Data

Model	Deviance	d.f.
$S + E + P$	2714.0	10
$SE + P$	1877.4	7
$SP + E$	1920.4	7
$S + EP$	1092.0	9
$SE + SP$	1083.8	4
$SE + EP$	255.5	6
$SP + EP$	298.5	6
$SE + SP + EP$	1.575	3

We now switch to a multinomial context, where we focus on the joint distribution of the three variables S, E and P. We consider four different types of models that may be of interest in this case, and discuss their equivalence to one of the above Poisson models.

5.2.3 Complete Independence

The simplest possible model of interest in the multinomial context is the model of complete independence, where the joint distribution of the three variables is the product of the marginals. The corresponding hypothesis is

$$H_0 : \pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}, \tag{5.9}$$

where $\pi_{i..}$ is the marginal probability that an observation falls in row i, and $\pi_{.j.}$ and $\pi_{..k}$ are the corresponding column and layer margins.

Under this model the logarithms of the expected cell counts are given by

$$\log \mu_{ijk} = \log n + \log \pi_{i..} + \log \pi_{.j.} + \log \pi_{..k},$$

and can be seen to depend only on quantities indexed by i, j and k but none of the combinations (such as ij, jk or ik). The notation is reminiscent of the Poisson additive model, where

$$\log \mu_{ijk} = \eta + \alpha_i + \beta_j + \gamma_k,$$

and in fact the two formulations can be shown to be equivalent, differing only on the choice of constraints: the marginal probabilities add up to one, whereas the main effects in the log-linear model satisfy the reference cell restrictions.

The m.l.e.'s of the probabilities under the model of complete independence turn out to be, as you might expect, the products of the marginal proportions. Therefore, the m.l.e.'s of the expected counts under complete independence are

$$\hat{\mu}_{ijk} = y_{i..}y_{.j.}y_{..k}/n^2.$$

Note that the estimates depend only on row, column and layer totals, as one would expect from considerations of marginal sufficiency.

To test the hypothesis of complete independence we compare the maximized multinomial log-likelihoods under the model of independence and under the saturated model. Because of the equivalence between multinomial and Poisson models, however, the resulting likelihood ratio statistic is exactly the same as the deviance for the Poisson additive model.

In our example the deviance of the additive model is 2714 with 10 d.f., and is highly significant. We therefore conclude that the hypothesis that social status, parental encouragement and college plans are completely independent is clearly untenable.

5.2.4 Block Independence

The next three log-linear models in Table 5.3 involve one of the two-factor interaction terms. As you might expect from our analysis of a two-by-two table, the presence of an interaction term indicates the existence of association between those two variables.

For example the model $SE + P$ indicates that S and E are associated, but are jointly independent of P. In terms of our example this hypothesis would state that social status and parental encouragement are associated with each other, and are jointly independent of college plans.

Under this hypothesis the joint distribution of the three variables factors into the product of two blocks, representing S and E on one hand and P on the other. Specifically, the hypothesis of block independence is

$$H_0 : \pi_{ijk} = \pi_{ij.}\pi_{..k}. \tag{5.10}$$

The m.l.e.'s of the cell probabilities turn out to be the product of the SE and P marginal probabilities and can be calculated directly. The m.l.e.'s of the expected counts under block independence are then

$$\hat{\mu}_{ijk} = y_{ij.}y_{..k}/n.$$

Note the similarity between the structure of the probabilities and that of the estimates, depending on the combination of levels of S and E on the one hand, and levels of P on the other.

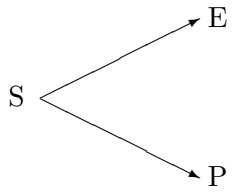
To test the hypothesis of block independence we compare the maximized multinomial log-likelihood under the restrictions imposed by Equation 5.10 with the maximized log-likelihood for the saturated model. Because of the equivalence between multinomial and Poisson models, however, the test statistic would be exactly the same as the deviance for the model $SE + P$.

In our example the deviance for the model with the SE interaction and a main effect of P is 1877.4 on 7 d.f., and is highly significant. We therefore reject the hypothesis that college plans are independent of social status and parental encouragement.

There are two other models with one interaction term. The model $SP+E$ has a deviance of 1920.4 on 7 d.f., so we reject the hypothesis that parental encouragement is independent of social status and college plans. The model $EP + S$ is the best fitting of this lot, but the deviance of 1092.0 on 9 d.f. is highly significant, so we reject the hypothesis that parental encouragement and college plans are associated but are jointly independent of social status.

5.2.5 Partial Independence

The next three log-linear models in Table 5.3 involve two of the three possible two-factor interactions, and thus correspond to cases where two pairs of categorical variables are associated. For example the log-linear model $SE + SP$ corresponds to the case where S and E are associated and so are S and P. In terms of our example we would assume that social status affects both parental encouragement and college plans. The figure below shows this model in path diagram form.



Note that we have assumed no direct link between E and P, that is, the model assumes that parental encouragement has no direct effect on college plans. In a two-way crosstabulation these two variables would appear to be associated because of their common dependency on social status S. However, *conditional* on social status S, parental encouragement E and college plans P would be independent.

Thus, the model assumes a form of partial or conditional independence, where the joint conditional distribution of EP given S is the product of the marginal conditional distributions of E given S and P given S. In symbols,

$$\Pr\{E = j, P = k|S = i\} = \Pr\{E = j|S = i\}\Pr\{P = k|S = i\}.$$

To translate this statement into unconditional probabilities we write the conditional distributions as the product of the joint and marginal distributions, so that the above equation becomes

$$\frac{\Pr\{E = j, P = k, S = i\}}{\Pr\{S = i\}} = \frac{\Pr\{E = j, S = i\}}{\Pr\{S = i\}} \frac{\Pr\{P = k, S = i\}}{\Pr\{S = i\}},$$

from which we see that

$$\Pr\{S = i, E = j, P = k\} = \frac{\Pr\{S = i, E = j\}\Pr\{S = i, P = k\}}{\Pr\{S = i\}},$$

or, in our usual notation,

$$\pi_{ijk} = \frac{\pi_{ij.}\pi_{i.k}}{\pi_{i..}}. \tag{5.11}$$

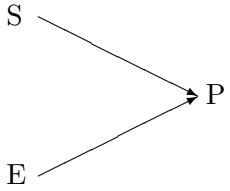
The m.l.e.'s of the expected cell counts have a similar structure and depend only on the SE and SP margins:

$$\hat{\mu}_{ijk} = \frac{y_{ij.}y_{i.k}}{y_{i..}}.$$

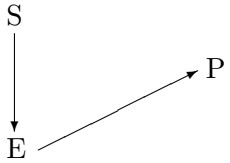
To test the hypothesis of partial independence we need to compare the multinomial log-likelihood maximized under the constraints implied by Equation 5.11 with the unconstrained maximum. Because of the equivalence between multinomial and Poisson models, however, the resulting likelihood ratio test statistic is the same as the deviance of the model $SE + SP$.

In terms of our example, the deviance of the model with SE and SP interactions is 1083.8 on 4 d.f., and is highly significant. We therefore reject the hypothesis that parental encouragement and college plans are independent within each social stratum.

There are two other models with two interaction terms. Although both of them have smaller deviances than any of the models considered so far, they still show significant lack of fit. The model $SP + EP$ has a deviance of 298.5 on 6 d.f., so we reject the hypothesis that given college plans P social status S and parental encouragement E are mutually independent. The best way to view this model in causal terms is by assuming that S and E are unrelated and both have effects on P, as shown in the path diagram below.



The model $SE + EP$ has a deviance of 255.5 on 6 d.f., and leads us to reject the hypothesis that given parental encouragement E, social class S and college plans P are independent. In causal terms one might interpret this model as postulating that social class affects parental encouragement which in turn affects college plans, with no direct effect of social class on college plans.



Note that all models consider so far have had explicit formulas for the m.l.e.'s, so no iteration has been necessary and we could have calculated all test

statistics using the multinomial likelihood directly. An interesting property of the iterative proportional fitting algorithm mentioned earlier, and which is used by software specializing in contingency tables, is that it converges in one cycle in all these cases. The same is not true of the iteratively re-weighted least squares algorithm used in Poisson regression, which will usually require a few iterations.

5.2.6 Uniform Association

The only log-linear model remaining in Table 5.3 short of the saturated model is the model involving all three two-factor interactions. In this model we have a form of association between all pairs of variables, S and E, S and P, as well as E and P. Thus, social class is associated with parental encouragement and with college plans, and in addition parental encouragement has a direct effect on college plans.

How do we interpret the lack of a three-factor interaction? To answer this question we start from what we know about interaction effects in general and adapt it to the present context, where interaction terms in models for counts represent association between the underlying classification criteria. The conclusion is that in this model the association between any two of the variables is the same at all levels of the third.

This model has no simple interpretation in terms of independence, and as a result we cannot write the structure of the joint probabilities in terms of the two-way margins. In particular

$$\pi_{ijk} \qquad \textit{is not} \qquad \frac{\pi_{ij.}\pi_{i.k}\pi_{.jk}}{\pi_{i..}\pi_{.j.}\pi_{..k}},$$

nor any other simple function of the marginal probabilities.

A consequence of this fact is that the m.l.e.'s cannot be written in closed form and must be calculated using an iterative procedure. They do, however, depend only on the three two-way margins SE, SP and EP.

In terms of our example, the model $SP + SE + EP$ has a deviance of 1.6 on three d.f., and therefore fits the data quite well. We conclude that we have no evidence against the hypothesis that all three variables are associated, but the association between any two is the same at all levels of the third. In particular, we may conclude that the association between parental encouragement E and college plans P is the same in all social strata.

To further appreciate the nature of this model, we give the fitted values in Table 5.4. Comparison of the estimated expected counts in this table with the observed counts in Table 5.2 highlights the goodness of fit of the model.

TABLE 5.4: Fitted Values for Educational Aspirations Data Based on Model of Uniform Association $SE + SP + EP$

Social Stratum	Parental Encouragement	College Plans	
		No	Yes
Lower	Low	753.1	30.9
	High	228.9	137.1
Lower Middle	Low	626.0	39.0
	High	331.0	302.0
Upper Middle	Low	420.9	36.1
	High	373.1	467.9
Higher	Low	149.0	30.0
	High	270.0	796.0

We can also use the fitted values to calculate measures of association between parental encouragement E and college plans P for each social stratum. For the lowest group, the odds of making college plans are barely one to 24.4 with low parental encouragement, but increase to one to 1.67 with high encouragement, giving an odds ratio of 14.6. If you repeat the calculation for any of the other three social classes you will find exactly the same ratio of 14.6.

We can verify that this result follows directly from the lack of a three-factor interaction in the model. The logs of the expected counts in this model are

$$\log \mu_{ijk} = \eta + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}.$$

The log-odds of making college plans in social stratum i with parental encouragement j are obtained by calculating the difference in expected counts between $k = 2$ and $k = 1$, which is

$$\log(\mu_{ij2}/\mu_{ij1}) = \gamma_2 - \gamma_1 + (\alpha\gamma)_{i2} - (\alpha\gamma)_{i1} + (\beta\gamma)_{j2} - (\beta\gamma)_{j1},$$

because all terms involving only i, j or ij cancel out. Consider now the difference in log-odds between high and low encouragement, i.e. when $j = 2$ and $j = 1$:

$$\log(\frac{\mu_{i22}/\mu_{i21}}{\mu_{i12}/\mu_{i11}}) = (\beta\gamma)_{22} - (\beta\gamma)_{21} - (\beta\gamma)_{12} + (\beta\gamma)_{11},$$

which does not depend on i. Thus, we see that the log of the odds ratio is the same at all levels of S. Furthermore, under the reference cell restrictions

all interaction terms involving level one of any of the factors would be set to zero, so the log of the odds ratio in question is simply $(\beta\gamma)_{22}$. For the model with no three-factor interaction the estimate of this parameter is 2.683 and exponentiating this value gives 14.6.

5.2.7 Binomial Logits Revisited

Our analysis so far has treated the three classification criteria as responses, and has focused on their correlation structure. An alternative approach would treat one of the variables as a response and the other two as predictors in a regression framework. We now compare these two approaches in terms of our example on educational aspirations, treating college plans as a dichotomous response and socio-economic status and parental encouragement as discrete predictors.

To this end, we treat each of the 8 rows in Table 5.2 as a group. Let Y_{ij} denote the number of high school seniors who plan to attend college out of the n_{ij} seniors in category i of socio-economic status and category j of parental encouragement. We assume that these counts are independent and have binomial distributions with $Y_{ij} \sim B(n_{ij}, \pi_{ij})$, where π_{ij} is the probability of making college plans We can then fit logistic regression models to study how the probabilities depend on social stratum and parental encouragement.

Table 5.5: Deviances for Logistic Regression Models Fitted to the Educational Aspirations Data

Model	Deviance	d.f.
Null	1877.4	7
S	1083.8	4
E	255.5	6
$S + E$	1.575	3

Table 5.5 shows the results of fitting four possible logit models of interest, ranging from the null model to the additive model on socioeconomic status (S) and parental encouragement (E). It is clear from these results that both social class and encouragement have significant gross and net effects on the probability of making college plans. The best fitting model is the two-factor additive model, with a deviance of 1.6 on three d.f. Table 5.6 shows parameter estimates for the additive model.

Exponentiating the estimates we see that the odds of making college plans increase five-fold as we move from low to high socio-economic status.

TABLE 5.6: Parameter Estimates for Additive Logit Model Fitted to the Educational Aspirations Data

Variable	Category	Estimate	Std. Err.
Constant		−3.195	0.119
Socio-economic	low	–	
status	lower middle	0.420	0.118
	upper middle	0.739	0.114
	high	1.593	0.115
Parental	low	–	
encouragement	high	2.683	0.099

Furthermore, in each social stratum, the odds of making college plans among high school seniors with high parental encouragement are 14.6 times the odds among seniors with low parental encouragement.

The conclusions of this analysis are consistent with those from the previous subsection, except that this time we do not study the association between social stratification and parental encouragement, but focus on their effect on making college plans. In fact it is not just the conclusions, but all estimates and tests of significance, that agree. A comparison of the binomial deviances in Table 5.5 with the Poisson deviances in Table 5.3 shows the following 'coincidences':

log-linear model	*logit model*
$SE + P$	Null
$SE + SP$	S
$SE + EP$	E
$SE + SP + EP$	$S + E$

The models listed as equivalent have similar interpretations if you translate from the language of correlation analysis to the language of regression analysis. Note that all the log-linear models include the SE interaction, so they allow for association between the two predictors. Also, all of them include a main effect of the response P, allowing it to have a non-uniform distribution. The log-linear model with just these two terms assumes no association between P and either S or E, and is thus equivalent to the null logit model.

The log-linear model with an SP interaction allows for an association between S and P, and is therefore equivalent to the logit model where the response depends only on S. A similar remark applies to the log-linear model with an EP interaction. Finally, the log-linear model with all three

two-factor interactions allows for associations between S and P, and between E and P, and assumes that in each case the strength of association does not depend on the other variable. But this is exactly what the additive logit model assumes: the response depends on both S and E, and the effect of each factor is the same at all levels of the other predictor.

In general, log-linear and logit models are equivalent as long as the log-linear model

- is saturated on all factors treated as predictors in the logit model, including all possible main effects and interactions among predictors (in our example SE),
- includes a main effect for the factor treated as response (in our example P), and
- includes a two-factor (or higher order) interaction between a predictor and the response for each main effect (or interaction) included in the logit model (in our example it includes SP for the main effect of S, and son on).

This equivalence extends to parameter estimates as well as tests of significance. For example, multiplying the fitted probabilities based on the additive logit model $S + E$ by the sample sizes in each category of social status and parental encouragement leads to the same expected counts that we obtained earlier from the log-linear model $SE+SP+EP$. An interesting consequence of this fact is that one can use parameter estimates based on a log-linear model to calculate logits, as we did in Section 5.2.6, and obtain the same results as in logistic regression. For example the log of the odds ratio summarizing the effect of parental encouragement on college plans within each social stratum was estimated as 2.683 in the previous subsection, and this value agrees exactly with the estimate on Table 5.6.

In our example the equivalence depends crucially on the fact that the log-linear models include the SE interaction, and therefore reproduce exactly the binomial denominators used in the logistic regression. But what would have happened if the SE interaction had turned out to be not significant? There appear to be two schools of thought on this matter.

Bishop et al. (1975), in a classic book on the multivariate analysis of qualitative data, emphasize log-linear models because they provide a richer analysis of the structure of association among all factors, not just between the predictors and the response. If the SE interaction had turned out to be not significant they would probably leave it out of the model. They would

still be able to translate their parameter estimates into fitted logits, but the results would not coincide exactly with the logistic regression analysis (although they would be rather similar if the omitted interaction is small.)

Cox (1972), in a classic book on the analysis of binary data, emphasizes logit models. He argues that if your main interest is on the effects of two variables, say S and E on a third factor, say P, then you should condition on the SE margin. This means that if you are fitting log-linear models with the intention of understanding effects on P, you would include the SE interaction even if it is not significant. In that case you would get exactly the same results as a logistic regression analysis, which is probably what you should have done in the first place if you wanted to study specifically how the response depends on the predictors.

Chapter 6

Multinomial Response Models

We now turn our attention to regression models for the analysis of categorical dependent variables with more than two response categories. Several of the models that we will study may be considered generalizations of logistic regression analysis to polychotomous data. We first consider models that may be used with purely qualitative or *nominal* data, and then move on to models for *ordinal* data, where the response categories are ordered.

6.1 The Nature of Multinomial Data

Let me start by introducing a simple dataset that will be used to illustrate the multinomial distribution and multinomial response models.

6.1.1 The Contraceptive Use Data

Table 6.1 was reconstructed from weighted percents found in Table 4.7 of the final report of the Demographic and Health Survey conducted in El Salvador in 1985 (FESAL-1985). The table shows 3165 currently married women classified by age, grouped in five-year intervals, and current use of contraception, classified as sterilization, other methods, and no method.

A fairly standard approach to the analysis of data of this type could treat the two variables as responses and proceed to investigate the question of independence. For these data the hypothesis of independence is soundly rejected, with a likelihood ratio χ^2 of 521.1 on 12 d.f.

TABLE 6.1: Current Use of Contraception By Age
Currently Married Women. El Salvador, 1985

Age	Contraceptive Method			All
	Ster.	Other	None	
15–19	3	61	232	296
20–24	80	137	400	617
25–29	216	131	301	648
30–34	268	76	203	547
35–39	197	50	188	435
40–44	150	24	164	338
45–49	91	10	183	284
All	1005	489	1671	3165

In this chapter we will view contraceptive use as the response and age as a predictor. Instead of looking at the joint distribution of the two variables, we will look at the conditional distribution of the response, contraceptive use, given the predictor, age. As it turns out, the two approaches are intimately related.

6.1.2 The Multinomial Distribution

Let us review briefly the multinomial distribution that we first encountered in Chapter 5. Consider a random variable Y_i that may take one of several discrete values, which we index $1, 2, \ldots, J$. In the example the response is contraceptive use and it takes the values 'sterilization', 'other method' and 'no method', which we index 1, 2 and 3. Let

$$\pi_{ij} = \Pr\{Y_i = j\} \tag{6.1}$$

denote the probability that the i-th response falls in the j-th category. In the example π_{i1} is the probability that the i-th respondent is 'sterilized'.

Assuming that the response categories are mutually exclusive and exhaustive, we have $\sum_{j=1}^{J} \pi_{ij} = 1$ for each i, i.e. the probabilities add up to one for each individual, and we have only $J - 1$ parameters. In the example, once we know the probability of 'sterilized' and of 'other method' we automatically know by subtraction the probability of 'no method'.

For *grouped data* it will be convenient to introduce auxiliary random variables representing counts of responses in the various categories. Let n_i denote the number of cases in the i-th group and let Y_{ij} denote the number

of responses from the i-th group that fall in the j-th category, with observed value y_{ij}.

In our example i represents age groups, n_i is the number of women in the i-th age group, and y_{i1}, y_{i2}, and y_{i3} are the numbers of women sterilized, using another method, and using no method, respectively, in the i-th age group. Note that $\sum_j y_{ij} = n_i$, i.e. the counts in the various response categories add up to the number of cases in each age group.

For *individual data* $n_i = 1$ and Y_{ij} becomes an indicator (or dummy) variable that takes the value 1 if the i-th response falls in the j-th category and 0 otherwise, and $\sum_j y_{ij} = 1$, since one and only one of the indicators y_{ij} can be 'on' for each case. In our example we could work with the 3165 records in the individual data file and let y_{i1} be one if the i-th woman is sterilized and 0 otherwise.

The probability distribution of the counts Y_{ij} given the total n_i is given by the *multinomial* distribution

$$\Pr\{Y_{i1} = y_{i1}, \ldots, Y_{iJ} = y_{iJ}\} = \binom{n_i}{y_{i1}, \ldots, y_{iJ}} \pi_{i1}^{y_{i1}} \ldots \pi_{iJ}^{y_{iJ}} \qquad (6.2)$$

The special case where $J = 2$ and we have only two response categories is the binomial distribution of Chapter 3. To verify this fact equate $y_{i1} = y_i$, $y_{i2} = n_i - y_i$, $\pi_{i1} = \pi_i$, and $\pi_{i2} = 1 - \pi_i$.

6.2 The Multinomial Logit Model

We now consider models for the probabilities π_{ij}. In particular, we would like to consider models where these probabilities depend on a vector $\mathbf{x}_i$ of covariates associated with the i-th individual or group. In terms of our example, we would like to model how the probabilities of being sterilized, using another method or using no method at all depend on the woman's age.

6.2.1 Multinomial Logits

Perhaps the simplest approach to multinomial data is to nominate one of the response categories as a baseline or reference cell, calculate log-odds for all other categories relative to the baseline, and then let the log-odds be a linear function of the predictors.

Typically we pick the *last* category as a baseline and calculate the odds that a member of group i falls in category j as opposed to the baseline as π_{i1}/π_{iJ}. In our example we could look at the odds of being sterilized rather

than using no method, and the odds of using another method rather than no method. For women aged 45–49 these odds are 91:183 (or roughly 1 to 2) and 10:183 (or 1 to 18).

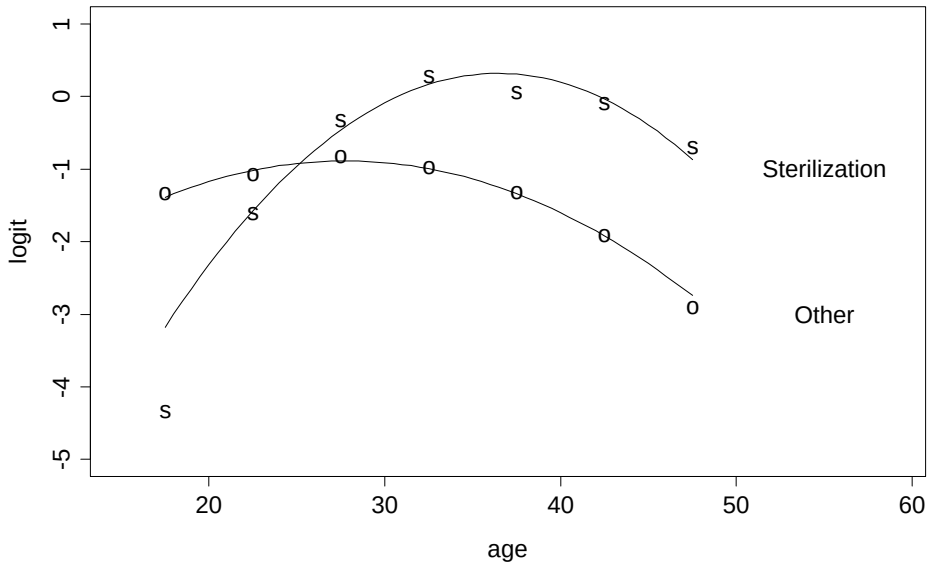


FIGURE 6.1: Log-Odds of Sterilization vs. No Method and Other Method vs. No Method, by Age

Figure 6.1 shows the empirical log-odds of sterilization and other method (using no method as the reference category) plotted against the mid-points of the age groups. (Ignore for now the solid lines.) Note how the log-odds of sterilization increase rapidly with age to reach a maximum at 30–34 and then decline slightly. The log-odds of using other methods rise gently up to age 25–29 and then decline rapidly.

6.2.2 Modeling the Logits

In the multinomial logit model we assume that the log-odds of each response follow a linear model

$$\eta_{ij} = \log \frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + \mathbf{x}_i' \boldsymbol{\beta}_j, \tag{6.3}$$

where α_j is a constant and $\boldsymbol{\beta}_j$ is a vector of regression coefficients, for $j = 1, 2, \ldots, J-1$. Note that we have written the constant explicitly, so we will

assume henceforth that the model matrix $\mathbf{X}$ does not include a column of ones.

This model is analogous to a logistic regression model, except that the probability distribution of the response is multinomial instead of binomial and we have $J-1$ equations instead of one. The $J-1$ multinomial logit equations contrast each of categories $1, 2, \ldots J-1$ with category J, whereas the single logistic regression equation is a contrast between successes and failures. If $J = 2$ the multinomial logit model reduces to the usual logistic regression model.

Note that we need only $J-1$ equations to describe a variable with J response categories and that it really makes no difference which category we pick as the reference cell, because we can always convert from one formulation to another. In our example with $J = 3$ categories we contrast categories 1 versus 3 and 2 versus 3. The missing contrast between categories 1 and 2 can easily be obtained in terms of the other two, since $\log(\pi_{i1}/\pi_{i2}) = \log(\pi_{i1}/\pi_{i3}) - \log(\pi_{i2}/\pi_{i3})$.

Looking at Figure 6.1, it would appear that the logits are a quadratic function of age. We will therefore entertain the model

$$\eta_{ij} = \alpha_j + \beta_j a_i + \gamma_j a_i^2, \tag{6.4}$$

where a_i is the midpoint of the i-th age group and $j = 1, 2$ for sterilization and other method, respectively.

6.2.3 Modeling the Probabilities

The multinomial logit model may also be written in terms of the original probabilities π_{ij} rather than the log-odds. Starting from Equation 6.3 and adopting the convention that $\eta_{iJ} = 0$, we can write

$$\pi_{ij} = \frac{\exp\{\eta_{ij}\}}{\sum_{k=1}^{J} \exp\{\eta_{ik}\}}. \tag{6.5}$$

for $j = 1, \ldots, J$. To verify this result exponentiate Equation 6.3 to obtain $\pi_{ij} = \pi_{iJ} \exp\{\eta_{ij}\}$, and note that the convention $\eta_{iJ} = 0$ makes this formula valid for all j. Next sum over j and use the fact that $\sum_j \pi_{ij} = 1$ to obtain $\pi_{iJ} = 1/\sum_j \exp\{\eta_{ij}\}$. Finally, use this result on the formula for π_{ij}.

Note that Equation 6.5 will automatically yield probabilities that add up to one for each i.

6.2.4 Maximum Likelihood Estimation

Estimation of the parameters of this model by maximum likelihood proceeds by maximization of the multinomial likelihood (6.2) with the probabilities π_{ij} viewed as functions of the α_j and $\boldsymbol{\beta}_j$ parameters in Equation 6.3. This usually requires numerical procedures, and Fisher scoring or Newton-Raphson often work rather well. Most statistical packages include a multinomial logit procedure.

In terms of our example, fitting the quadratic multinomial logit model of Equation 6.4 leads to a deviance of 20.5 on 8 d.f. The associated P-value is 0.009, so we have significant lack of fit.

The quadratic age effect has an associated likelihood-ratio χ^2 of 500.6 on four d.f. ($521.1 - 20.5 = 500.6$ and $12 - 8 = 4$), and is highly significant. Note that we have accounted for 96% of the association between age and method choice ($500.6/521.1 = 0.96$) using only four parameters.

TABLE 6.2: Parameter Estimates for Multinomial Logit Model Fitted to Contraceptive Use Data

Parameter	Contrast	
	Ster. vs. None	Other vs. None
Constant	-12.62	-4.552
Linear	0.7097	0.2641
Quadratic	-0.009733	-0.004758

Table 6.2 shows the parameter estimates for the two multinomial logit equations. I used these values to calculate fitted logits for each age from 17.5 to 47.5, and plotted these together with the empirical logits in Figure 6.1. The figure suggests that the lack of fit, though significant, is not a serious problem, except possibly for the 15–19 age group, where we overestimate the probability of sterilization.

Under these circumstances, I would probably stick with the quadratic model because it does a reasonable job using very few parameters. However, I urge you to go the extra mile and try a cubic term. The model should pass the goodness of fit test. Are the fitted values reasonable?

6.2.5 The Equivalent Log-Linear Model*

Multinomial logit models may also be fit by maximum likelihood working with an equivalent log-linear model and the Poisson likelihood. (This section

will only be of interest to readers interested in the equivalence between these models and may be omitted at first reading.)

Specifically, we treat the random counts Y_{ij} as Poisson random variables with means μ_{ij} satisfying the following log-linear model

$$\log \mu_{ij} = \eta + \theta_i + \alpha_j^* + \mathbf{x}_i' \boldsymbol{\beta}_j^*, \tag{6.6}$$

where the parameters satisfy the usual constraints for identifiability. There are three important features of this model:

First, the model includes a separate parameter θ_i for each multinomial observation, i.e. each individual or group. This assures exact reproduction of the multinomial denominators n_i. Note that these denominators are fixed known quantities in the multinomial likelihood, but are treated as random in the Poisson likelihood. Making sure we get them right makes the issue of conditioning moot.

Second, the model includes a separate parameter α_j^* for each response category. This allows the counts to vary by response category, permitting non-uniform margins.

Third, the model uses interaction terms $\mathbf{x}_i' \boldsymbol{\beta}_j^*$ to represent the effects of the covariates $\mathbf{x}_i$ on the log-odds of response j. Once again we have a 'step-up' situation, where main effects in a logistic model become interactions in the equivalent log-linear model.

The log-odds that observation i will fall in response category j relative to the last response category J can be calculated from Equation 6.6 as

$$\log(\mu_{ij}/\mu_{iJ}) = (\alpha_j^* - \alpha_J^*) + \mathbf{x}_i'(\boldsymbol{\beta}_j^* - \boldsymbol{\beta}_J^*). \tag{6.7}$$

This equation is identical to the multinomial logit Equation 6.3 with $\alpha_j = \alpha_j^* - \alpha_J^*$ and $\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^* - \boldsymbol{\beta}_J^*$. Thus, the parameters in the multinomial logit model may be obtained as differences between the parameters in the corresponding log-linear model. Note that the θ_i cancel out, and the restrictions needed for identification, namely $\eta_{iJ} = 0$, are satisfied automatically.

In terms of our example, we can treat the counts in the original 7×3 table as 21 independent Poisson observations, and fit a log-linear model including the main effect of age (treated as a factor), the main effect of contraceptive use (treated as a factor) and the interactions between contraceptive use (a factor) and the linear and quadratic components of age:

$$\log \mu_{ij} = \eta + \theta_i + \alpha_j^* + \beta_j^* a_i + \gamma_j^* a_i^2 \tag{6.8}$$

In practical terms this requires including six dummy variables representing the age groups, two dummy variables representing the method choice categories, and a total of four interaction terms, obtained as the products of

the method choice dummies by the mid-point a_i and the square of the mid-point a_i^2 of each age group. Details are left as an exercise. (But see the Stata notes.)

6.3 The Conditional Logit Model

In this section I will describe an extension of the multinomial logit model that is particularly appropriate in models of choice behavior, where the explanatory variables may include attributes of the choice alternatives (for example cost) as well as characteristics of the individuals making the choices (such as income). To motivate the extension I will first reintroduce the multinomial logit model in terms of an underlying latent variable.

6.3.1 A General Model of Choice

Suppose that Y_i represents a discrete choice among J alternatives. Let U_{ij} represent the value or *utility* of the j-th choice to the i-th individual. We will treat the U_{ij} as independent random variables with a systematic component η_{ij} and a random component ϵ_{ij} such that

$$U_{ij} = \eta_{ij} + \epsilon_{ij}. \tag{6.9}$$

We assume that individuals act in a rational way, maximizing their utility. Thus, subject i will choose alternative j if U_{ij} is the largest of $U_{i1}, \ldots, U_{iJ}$. Note that the choice has a random component, since it depends on random utilities. The *probability* that subject i will choose alternative j is

$$\pi_{ij} = \Pr\{Y_i = j\} = \Pr\{\max(U_{i1}, \ldots, U_{iJ}) = U_{ij}\}. \tag{6.10}$$

It can be shown that if the error terms ϵ_{ij} have standard Type I extreme value distributions with density

$$f(\epsilon) = \exp\{-\epsilon - \exp\{-\epsilon\}\} \tag{6.11}$$

then (see for example Maddala, 1983, pp 60–61)

$$\pi_{ij} = \frac{\exp\{\eta_{ij}\}}{\sum \exp\{\eta_{ik}\}}, \tag{6.12}$$

which is the basic equation defining the multinomial logit model.

In the special case where $J = 2$, individual i will choose the first alternative if $U_{i1} - U_{i2} > 0$. If the random utilities U_{ij} have independent

extreme value distributions, their difference can be shown to have a logistic distribution, and we obtain the standard logistic regression model.

Luce (1959) derived Equation 6.12 starting from a simple requirement that the odds of choosing alternative j over alternative k should be independent of the choice set for all pairs j, k. This property is often referred to as the axiom of *independence from irrelevant alternatives*. Whether or not this assumption is reasonable (and other alternatives are indeed irrelevant) depends very much on the nature of the choices.

A classical example where the multinomial logit model does not work well is the so-called "red/blue bus" problem. Suppose you have a choice of transportation between a train, a red bus and a blue bus. Suppose half the people take the train and half take the bus. Suppose further that people who take the bus are indifferent to the color, so they distribute themselves equally between the red and the blue buses. The choice probabilities of $\pi = (.50, .25, .25)$ would be consistent with expected utilities of $\eta = (\log 2, 0, 0)$.

Suppose now the blue bus service is discontinued. You might expect that all the people who used to take the blue bus would take the red bus instead, leading to a 1:1 split between train and bus. On the basis of the expected utilities of $\log 2$ and 0, however, the multinomial logit model would predict a 2:1 split.

Keep this caveat in mind as we consider modeling the expected utilities.

6.3.2 Multinomial Logits

In the usual multinomial logit model, the expected utilities η_{ij} are modeled in terms of characteristics of the individuals, so that

$$\eta_{ij} = \mathbf{x}_i'\boldsymbol{\beta}_j.$$

Here the regression coefficients $\boldsymbol{\beta}_j$ may be interpreted as reflecting the effects of the covariates on the odds of making a given choice (as we did in the previous section) or on the underlying utilities of the various choices.

A somewhat restrictive feature of the model is that the same attributes $\mathbf{x}_i$ are used to model the utilities of all J choices.

6.3.3 Conditional Logits

McFadden (1973) proposed modeling the expected utilities η_{ij} in terms of characteristics of the alternatives rather than attributes of the individuals. If $\mathbf{z}_j$ represents a vector of characteristics of the j-th alternative, then he

postulated the model

$$\eta_{ij} = \mathbf{z}_j'\boldsymbol{\gamma}.$$

This model is called the *conditional logit* model, and turns out to be equivalent to a log-linear model where the main effect of the response is represented in terms of the covariates $\mathbf{z}_j$.

Note that with J response categories the response margin may be reproduced exactly using any $J-1$ linearly independent attributes of the choices. Generally one would want the dimensionality of z_j to be substantially less than J. Consequently, conditional logit models are often used when the number of possible choices is large.

6.3.4 Multinomial/Conditional Logits

A more general model may be obtained by combining the multinomial and conditional logit formulations, so the underlying utilities η_{ij} depend on characteristics of the individuals as well as attributes of the choices, or even variables defined for combinations of individuals and choices (such as an individual's perception of the value of a choice). The general model is usually written as

$$\eta_{ij} = \mathbf{x}_i'\boldsymbol{\beta}_j + \mathbf{z}_{ij}'\boldsymbol{\gamma} \tag{6.13}$$

where $\mathbf{x}_i$ represents characteristics of the individuals that are constant across choices, and $\mathbf{z}_{ij}$ represents characteristics that vary across choices (whether they vary by individual or not).

Some statistical packages have procedures for fitting conditional logit models to datasets where each combination of individual and possible choice is treated as a separate observation. These models may also be fit using any package that does Poisson regression. If the last response category is used as the baseline or reference cell, so that $\eta_{iJ} = 0$ for all i, then the $\mathbf{z}_{ij}$ should be entered in the model as differences from the last category. In other words, you should use $\mathbf{z}^*_{ij} = \mathbf{z}_{ij} - \mathbf{z}_{iJ}$ as the predictor.

6.3.5 Multinomial/Conditional Probits

Changing the distribution of the error term in Equation 6.9 leads to alternative models. A popular alternative to the logit models considered so far is to assume that the ϵ_{ij} have independent standard normal distributions for all i, j. The resulting model is called the multinomial/conditional *probit* model, and produces results very similar to the multinomial/conditional logit model after standardization.

A more attractive alternative is to retain independence across subjects but allow dependence across alternatives, assuming that the vector $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iJ})'$ has a *multivariate* normal distribution with mean vector $\mathbf{0}$ and arbitrary correlation matrix $\mathbf{R}$. (As usual with latent variable formulations of binary or discrete response models, the variance of the error term cannot be separated from the regression coefficients. Setting the variances to one means that we work with a correlation matrix rather than a covariance matrix.)

The main advantage of this model is that it allows correlation between the utilities that an individual assigns to the various alternatives. The main difficulty is that fitting the model requires evaluating probabilities given by multidimensional normal integrals, a limitation that effectively restricts routine practical application of the model to problems involving no more than three or four alternatives.

For further details on discrete choice models see Chapter 3 in Maddala (1983).

6.4 The Hierarchical Logit Model

The strategy used in Section 6.2.1 to define logits for multinomial response data, namely nominating one of the response categories as a baseline, is only one of many possible approaches.

6.4.1 Nested Comparisons

An alternative strategy is to define a hierarchy of *nested* comparisons between two subsets of responses, using an ordinary logit model for each comparison. In terms of the contraceptive use example, we could consider (1) the odds of using some form of contraception, as opposed to none, and (2) the odds of being sterilized among users of contraception. For women aged 15–49 these odds are 1494:1671 (or roughly one to one) and 1005:489 (or roughly two to one).

The hierarchical or nested approach is very attractive if you assume that individuals make their decisions in a sequential fashion. In terms of contraceptive use, for example, women may first decide whether or nor they will use contraception. Those who decide to use then face the choice of a method. This sequential approach may also provide a satisfactory model for the "red/blue bus" choice.

Of course it is also possible that the decision to use contraception would

be affected by the types of methods available. If that is the case, a multinomial logit model may be more appropriate.

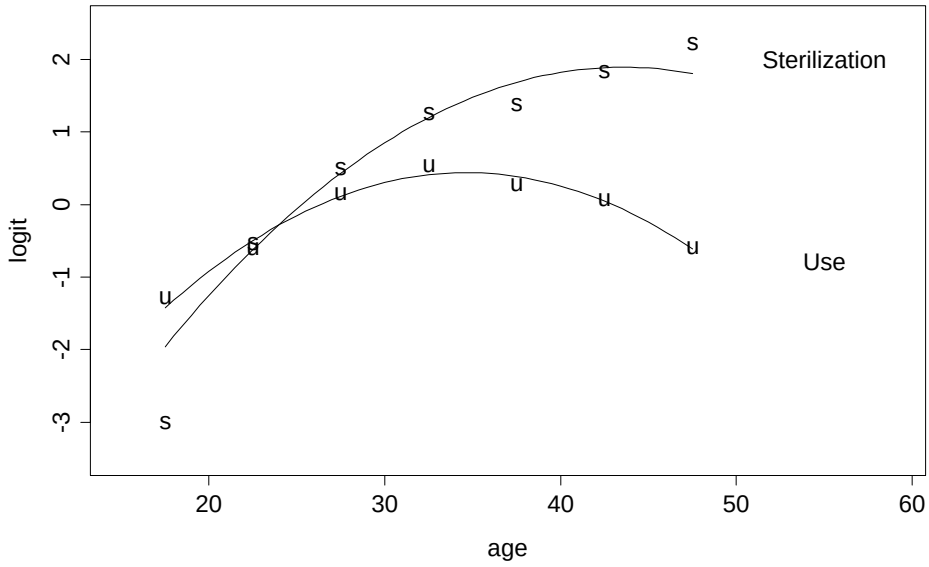


FIGURE 6.2: Log-Odds of Contraceptive Use vs. No Use and Sterilization vs. Other Method, by Age.

Figure 6.2 shows the empirical log-odds of using any method rather than no method, and of being sterilized rather than using another method among users, by age. Note that contraceptive use increases up to age 35–39 and then declines, whereas the odds of being sterilized among users increase almost monotonically with age.

The data suggest that the hierarchical logits may be modeled as quadratic functions of age, just as we did for the multinomial logits. We will therefore consider the model

$$\eta_{ij} = \alpha_j + \beta_j a_i + \gamma_j a_i^2, \tag{6.14}$$

where a_i is the mid-point of the i-th age group, $j = 1$ for the contraceptive use equation and $j = 2$ for the method choice equation.

6.4.2 Maximum Likelihood Estimation

An important practical feature of the hierarchical logit model is that the multinomial likelihood factors out into a product of binomial likelihoods, which may then be maximized separately.

I will illustrate using the contraceptive use data with 3 response categories, but the idea is obviously more general. The contribution of the i-th individual or group to the multinomial likelihood (ignoring constants) has the form

$$L_i = \pi_{i1}^{y_{i1}} \pi_{i2}^{y_{i2}} \pi_{i3}^{y_{i3}}, \tag{6.15}$$

where the π_{ij} are the probabilities and the y_{ij} are the corresponding counts of women sterilized, using other methods, and using no methods, respectively.

Multiply and divide this equation by $(\pi_{i1}+\pi_{i2})^{y_{i1}+y_{i2}}$, which is the probability of using contraception raised to the total number of users of contraception, to obtain

$$L_i = \left(\frac{\pi_{i1}}{\pi_{i1}+\pi_{i2}}\right)^{y_{i1}} \left(\frac{\pi_{i2}}{\pi_{i1}+\pi_{i2}}\right)^{y_{i2}} (\pi_{i1}+\pi_{i2})^{y_{i1}+y_{i2}} \pi_{i3}^{y_{i3}}. \tag{6.16}$$

Let $\rho_{i1} = \pi_{i1} + \pi_{i2}$ denote the probability of using contraception in age group i, and let $\rho_{i2} = \pi_{i1}/(\pi_{i1} + \pi_{i2})$ denote the *conditional* probability of being sterilized given that a woman is using contraception. Using this notation we can rewrite the above equation as

$$L_i = \rho_{i2}^{y_{i1}} (1-\rho_{i2})^{y_{i2}} \rho_{i1}^{y_{i1}+y_{i2}} (1-\rho_{i1})^{y_{i3}}. \tag{6.17}$$

The two right-most terms involving the probability of using contraception ρ_{i1} may be recognized, except for constants, as a standard binomial likelihood contrasting users and non-users. The two terms involving the conditional probability of using sterilization ρ_{i2} form, except for constants, a standard binomial likelihood contrasting sterilized women with users of other methods. As long as the parameters involved in the two equations are distinct, we can maximize the two likelihoods separately.

In view of this result we turn to Table 6.1 and fit two separate models. Fitting a standard logit model to the contraceptive use contrast (sterilization or other method vs. no method) using linear and quadratic terms on age gives a deviance of 6.12 on four d.f. and the parameter estimates shown in the middle column of Table 6.3. Fitting a similar model to the method choice contrast (sterilization vs. other method, restricted to users) gives a deviance of 10.77 on four d.f. and the parameter estimates shown in the rightmost column of Table 6.3.

The combined deviance is 16.89 on 8 d.f. (6.12+10.77 = 16.89 and 4+4 = 8). The associated P-value is 0.031, indicating lack of fit significant at the 5% level. Note, however, that the hierarchical logit model provides a somewhat better fit to these data than the multinomial logit model considered earlier, which had a deviance of 20.5 on the same 8 d.f.

TABLE 6.3: Parameter Estimates for Hierarchical Logit Model Fitted to Contraceptive Use Data

Parameter	Contrast	
	Use vs. No Use	Ster. vs. Other
Constant	-7.180	-8.869
Linear	0.4397	0.4942
Quadratic	-0.006345	-0.005674

To look more closely at goodness of fit I used the parameter estimates shown on Table 6.3 to calculate fitted logits and plotted these in Figure 6.2 against the observed logits. The quadratic model seems to do a reasonable job with very few parameters, particularly for overall contraceptive use. The method choice equation overestimates the odds of choosing sterilization for the age group 15–19, a problem shared by the multinomial logit model.

The parameter estimates may also be used to calculate illustrative odds of using contraception or sterilization at various ages. Going through these calculations you will discover that the odds of using some form of contraception increase 80% between ages 25 and 35. On the other hand, the odds of being sterilized among contraceptors increase three and a half times between ages 25 and 35.

6.4.3 Choice of Contrasts

With three response categories the only possible set of nested comparisons (aside from a simple reordering of the categories) is

{1,2} versus {3}, and
{1} versus {2}.

With four response categories there are two main alternatives. One is to contrast

{1, 2} versus {3, 4},
{1} versus {2}, and
{3} versus {4}.

The other compares

{1} versus {2, 3, 4},
{2} versus {3, 4}, and
{3} versus {4}.

The latter type of model, where one considers the odds of response $Y = j$ relative to responses $Y \geq j$, is known as a *continuation ratio* model (see Fienberg, 1980), and may be appropriate when the response categories are ordered.

More generally, any set of $J - 1$ linearly independent contrasts can be selected for modeling, but only orthogonal contrasts lead to a factorization of the likelihood function. The choice of contrasts should in general be based on the logic of the situation.

6.5 Models for Ordinal Response Data

Most of the models discussed so far are appropriate for the analysis of nominal responses. They may be applied to *ordinal* data as well, but the models make no explicit use of the fact that the response categories are ordered. We now consider models designed specifically for the analysis of responses measured on an ordinal scale. Our discussion follows closely McCullagh (1980).

6.5.1 Housing Conditions in Copenhagen

We will illustrate the application of models for ordinal data using the data in Table 6.4, which was first published by Madsen (1976) and was reproduced in Agresti (1990, p. 341). The table classifies 1681 residents of twelve areas in Copenhagen in terms of the type of housing they had, their feeling of influence on apartment management, their degree of contact with other residents, and their satisfaction with housing conditions.

In our analysis of these data we will treat housing satisfaction as an ordered response, with categories low, medium and high, and the other three factors as explanatory variables.

6.5.2 Cumulative Link Models

All of the models to be considered in this section arise from focusing on the *cumulative* distribution of the response. Let $\pi_{ij} = \Pr\{Y_i = j\}$ denote the probability that the response of an individual with characteristics $\mathbf{x}_i$ falls in the j-th category, and let γ_{ij} denote the corresponding cumulative probability

$$\gamma_{ij} = \Pr\{Y_i \leq j\} \tag{6.18}$$

TABLE 6.4: Housing Condition in Copenhagen

Housing Type	Influence	Contact	Satisfaction		
			low	medium	high
Tower block	low	low	21	21	28
		high	14	19	37
	medium	low	34	22	36
		high	17	23	40
	high	low	10	11	36
		high	3	5	23
Apartments	low	low	61	23	17
		high	78	46	43
	medium	low	43	35	40
		high	48	45	86
	high	low	26	18	54
		high	15	25	62
Atrium houses	low	low	13	9	10
		high	20	23	20
	medium	low	8	8	12
		high	10	22	24
	high	low	6	7	9
		high	7	10	21
Terraced houses	low	low	18	6	7
		high	57	23	13
	medium	low	15	13	13
		high	31	21	13
	high	low	7	5	11
		high	5	6	13

that the response falls in the j-th category *or below*, so

$$\gamma_{ij} = \pi_{i1} + \pi_{i2} + \ldots + \pi_{ij}. \tag{6.19}$$

Let $g(.)$ denote a link function mapping probabilities to the real line. Then the class of models that we will consider assumes that the transformed *cumulative* probabilities are a linear function of the predictors, of the form

$$g(\gamma_{ij}) = \theta_j + \mathbf{x}_i'\boldsymbol{\beta}. \tag{6.20}$$

In this formulation θ_j is a constant representing the baseline value of the transformed cumulative probability for category j, and $\boldsymbol{\beta}$ represents the

effect of the covariates on the transformed cumulative probabilities. Since we write the constant explicitly, we assume that the predictors do not include a column of ones. Note that there is just one equation: if x_{ik} increases by one, then *all* transformed cumulative probabilities increase by β_k. Thus, this model is more parsimonious than a multinomial logit or a hierarchical logit model; by focusing on the cumulative probabilities we can postulate a single effect. We will return to the issue of interpretation when we consider specific link functions.

These models can also be interpreted in terms of a *latent variable*. Specifically, suppose that the manifest response Y_i results from grouping an underlying continuous variable Y_i^* using cut-points $\theta_1 < \theta_2 < \ldots < \theta_{J-1}$, so that Y_i takes the value 1 if Y_i^* is below θ_1, the value 2 if Y_i^* is between θ_1 and θ_2, and so on, taking the value J if Y_i^* is above θ_{J-1}. Figure 6.3 illustrates this idea for the case of five response categories.

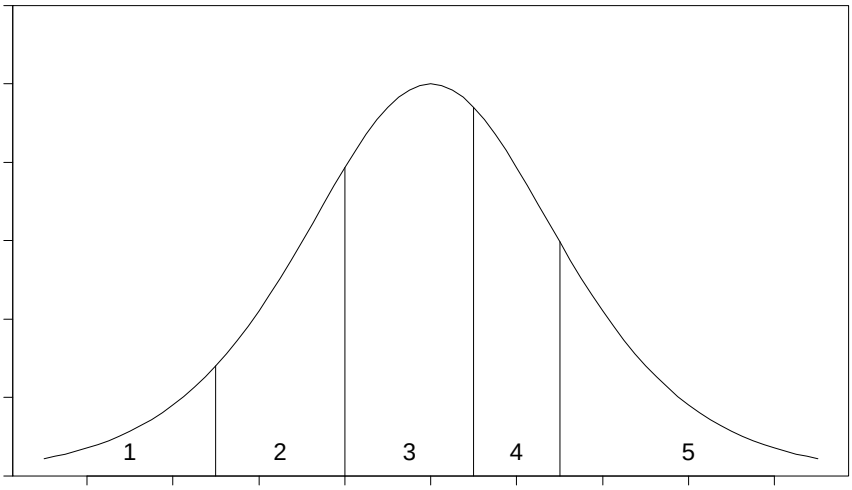


FIGURE 6.3: An Ordered Response and its Latent Variable

Suppose further that the underlying continuous variable follows a linear model of the form

$$Y_i^* = \mathbf{x}_i'\boldsymbol{\beta}^* + \epsilon_i, \tag{6.21}$$

where the error term ϵ_i has c.d.f. $F(\epsilon_i)$. Then, the probability that the response of the i-th individual will fall in the j-th category *or below*, given

$\mathbf{x}_i$, satisfies the equation

$$\gamma_{ij} = \Pr\{Y_i^* < \theta_j\} = \Pr\{\epsilon_i < \theta_j - \mathbf{x}_i'\boldsymbol{\beta}^*\} = F(\theta_j - \mathbf{x}_i'\boldsymbol{\beta}^*) \tag{6.22}$$

and therefore follows the general form in Equation (6.20) with link given by the inverse of the c.d.f. of the error term

$$g(\gamma_{ij}) = F^{-1}(\gamma_{ij}) = \theta_j - \mathbf{x}_i'\boldsymbol{\beta}^* \tag{6.23}$$

and coefficients $\boldsymbol{\beta}^* = -\boldsymbol{\beta}$ differing only in sign from the coefficients in the cumulative link model. Note that in both formulations we assume that the predictors $\mathbf{x}_i$ do not include a column of ones because the constant is absorbed in the cutpoints.

With grouped data the underlying continuous variable Y^* will have real existence and the cutpoints θ_j will usually be known. For example income data are often collected in broad categories, and all we known is the interval where an observation falls, i.e. < \$25,000, between \$25,000 and \$50,000, and so on.

With ordinal categorical data the underlying continuous variable will often represent a latent or unobservable trait, and the cutpoints will not be known. This would be the case, for example, if respondents are asked whether they support a balance budget amendment, and the response categories are strongly against, against, neutral, in favor, and strongly in favor. We could imagine an underlying degree of support Y_i^* and thresholds θ_1 to θ_4, such that when the support is below θ_1 one is strongly against, when the support exceeds θ_1 but not θ_2 one is against, and so on, until the case where the support exceeds θ_4 and one is strongly for the amendment.

While the latent variable interpretation is convenient, it is not always necessary, since some of the models can be interpreted directly in terms of the transformation $g(.)$ of the cumulative probabilities.

6.5.3 The Proportional Odds Model

The first model we will consider is a direct extension of the usual logistic regression model. Instead of applying the logit transformation to the response probabilities π_{ij}, however, we apply it to the *cumulative* response probabilities γ_{ij}, so that

$$\text{logit}(\gamma_{ij}) = \log \frac{\gamma_{ij}}{1-\gamma_{ij}} = \theta_j + \mathbf{x}_i'\boldsymbol{\beta}. \tag{6.24}$$

Some authors refer to this model as the ordered logit model, because it is a generalization of the logit model to ordered response categories. McCullagh

(1980) calls it the *proportional odds* model, for reasons that will be apparent presently. Exponentiating (6.24) we find that the odds of $Y_{ij} \leq j$, in words, the odds of a response in category j or below, are

$$\frac{\gamma_{ij}}{1-\gamma_{ij}} = \lambda_j \exp\{\mathbf{x}_i'\boldsymbol{\beta}\} \tag{6.25}$$

where $\lambda_j = \exp\{\theta_j\}$. The λ_j may be interpreted as the *baseline* odds of a response in category j or below when $x = 0$. The effect of the covariates x is to raise or lower the odds of a response in category j or below by the factor $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$. Note that the effect is a proportionate change in the odds of $Y_i \leq j$ for all response categories j. If a certain combination of covariate values doubles the odds of being in category 1, it also doubles the odds of being in category 2 or below, or in category 3 or below. Hence the name proportional odds.

This model may also be obtained from the latent variable formulation assuming that the error term ϵ_i has a standard logistic distribution. In this case the cdf is

$$F(\eta) = \frac{\exp\{\eta\}}{1+\exp\{\eta\}} \tag{6.26}$$

and the inverse cdf is the logit transformation. The $\boldsymbol{\beta}^*$ coefficients may then be interpreted as linear effects on the underlying continuous variable Y_i^*.

The proportional odds model is not a log-linear model, and therefore it can not be fit using the standard Poisson trick. It is possible, however, to use an iteratively re-weighted least squares algorithm very similar to the standard algorithm for generalized linear models, for details see McCullagh (1980).

We will illustrate this model applying it to the housing satisfaction data in Table 6.4. Let us start by noting that the log-likelihood for a saturated multinomial model that treats each of the 24 covariate patterns as a different group is -1715.71. Fitting a proportional odds model with additive effects of housing type, influence in management and contact with neighbors, yields a log-likelihood of -1739.57, which corresponds to a deviance (compared to the saturated multinomial model) of 47.73 on 40 d.f. To calculate the degrees of freedom note that the saturated multinomial model has 48 parameters (2 for each of 24 groups), while the additive proportional odds model has only 8 (2 threshold parameters, 3 for housing type, 2 for influence and one for contact). The 95% critical point of the χ^2_{40} distribution is 55.8, so you might think that this model fits the data.

To be thorough, however, we should investigate interaction effects. The models with one two-factor interaction have log-likelihoods of -1739.47 (in-

cluding contact × influence), -1735.24 (including housing × contact), and -1728.32 (including housing × influence), with corresponding deviance reductions of 0.21, 8.67 and 22.51, at the expense of 2, 3 and 6 degrees of freedom, respectively. Clearly the only interaction of note is that of housing × influence, which has a P-value of 0.001. Adding this term gives a model deviance of 25.22 on 34 d.f. and an excellent fit to the data.

Table 6.5 shows parameter estimates for the final model with all three predictors and a housing × influence interaction. The table lists the cutpoints and the regression coefficients.

TABLE 6.5: Parameter Estimates for Ordered Logit Model (Latent Variable Formulation)

Parameter	Estimate	Std. Error	z-ratio
Apartments	-1.1885	.1972	-6.026
Atrium house	-.6067	.2446	-2.481
Terraced house	-1.6062	.2410	-6.665
Influence medium	-.1390	.2125	-0.654
Influence high	.8689	.2743	3.167
Contact high	.3721	.0960	3.876
Apart × Influence med	1.0809	.2658	4.066
Apart × Influence high	.7198	.3287	2.190
Atrium × Influence med	.6511	.3450	1.887
Atrium × Influence high	-.1556	.4105	-0.379
Terrace × Influence med	.8210	.3307	2.483
Terrace × Influence high	.8446	.4303	1.963
Cutpoint 1	-.8881	.1672	
Cutpoint 2	.3126	.1657	

Note first the cutpoints: -.89 and .31, corresponding to cumulative odds of 0.41 and 1.37, or to cumulative probabilities of 0.29 and 0.58, for the reference cell. Considering residents of tower blocks with low influence in management and low contact with neighbors, we estimate that 29% have low satisfaction, 29% (58-29) have medium satisfaction, and 42% (100-58) have high satisfaction. (These values are fairly close to the observed proportions.)

Before we discuss interpretation of the remaining coefficients, we must note that I have reported the coefficients corresponding to the latent variable formulation (the $\boldsymbol{\beta}^*$'s) rather than the cumulative link coefficients (the $\boldsymbol{\beta}$'s), which have opposite sign. Thus, a positive coefficient is equivalent to a shift

to the right on the latent scale, which increases the odds of being to the *right* of a cutpoint. Equation (6.24) models the odds of being to the *left* of a cutpoint, which would then decrease. I prefer the sign used here because the interpretation is more straightforward. A positive coefficient increases one's underlying satisfaction, which makes a 'high' response more likely.

The coefficient of contact indicates that residents who have high contact with their neighbors are generally more satisfied than those who have low contact. The odds of high satisfaction (as opposed to medium or low), are 45% higher for high contact than for low contact, as are the odds of medium or high satisfaction (as opposed to low). The fact that the effect of contact on the odds is the same 45% for the two comparisons is a feature of the model.

To interpret the effects of the next two variables, type of housing and degree of influence, we have to allow for their interaction effect. One way to do this is to consider the effect of type of housing when the residents feel that they have low influence on management; then residents of apartments and houses (particularly terraced houses) are *less* satisfied than residents of tower blocks. Feeling that one has some influence on management generally increases satisfaction; the effect of having high rather than low influence is to increase the odds of medium or high satisfaction (as opposed to low) by 138% for residents of tower blocks, 390% for apartment dwellers, 104% for residents of atrium houses and 455% for those who live in terraced houses. Having medium influence is generally better than having low influence (except for tower clock residents), but not quite as good as having high influence (except possibly for residents of atrium houses).

Although we have interpreted the results in terms of odds, we can also interpret the coefficients in terms of a latent variable representing degree of satisfaction. The effect of having high contact with the neighbors, as compared to low contact, is to shift one's position on the latent satisfaction scale by 0.37 points. Similarly, having high influence on management, as compared to low influence, shifts one's position by an amount that varies from 0.71 for residents of atrium houses to 1.71 for residents of terraced houses. Interpretation of these numbers must be done by reference to the standard logistic distribution, which is depicted in Figure 6.3. This symmetric distribution has mean 0 and standard deviation $\pi/\sqrt{3} = 1.81$. The quartiles are ± 1.1, and just over 90% of the area lies between -3 and 3.

6.5.4 The Ordered Probit Model

The ordered probit model, first considered by Aitchison and Silvey (1957), results from modeling the *probit* of the cumulative probabilities as a linear function of the covariates, so that

$$\Phi^{-1}(\gamma_{ij}) = \theta_j + \mathbf{x}_i'\boldsymbol{\beta} \tag{6.27}$$

where $\Phi()$ is the standard normal cdf. The model can also be obtained from the latent-variable formulation assuming that the error term has a standard normal distribution, and this is usually the way one would interpret the parameters.

Estimates from the ordered probit model are usually very similar to estimates from the ordered logit model—as one would expect from the similarity of the normal and the logistic distributions—provided one remembers to standardize the coefficients to correct for the fact that the standard normal distribution has variance one, whereas the standard logistic has variance $\pi^2/3$.

For the Copenhagen data, the ordered probit model with an interaction between housing type and influence has a log-likelihood of -1728.67, corresponding to a deviance of 25.9 on 34 d.f., almost indistinguishable from the deviance for the ordered logit model with the same terms. Table 6.6 shows parameter estimates for this model.

The cutpoints can be interpreted in terms of z-scores: the boundary between low and medium satisfaction is at $z = -0.54$ and the boundary between medium and high satisfaction is at $z = 0.19$. These values leave $\Phi(-.54) = 0.29$ or 29% of the reference group in the low satisfaction category, $\Phi(0.19) - \Phi(-0.54) = 0.28$ or 28% in the medium satisfaction category, and $1 - \Phi(0.19) = 0.42$ or 42% in the high satisfaction category.

The remaining coefficients can be interpreted as in a linear regression model. For example, having high contact with the neighbors, compared to low contact, increases one's position in the latent satisfaction scale by 0.23 standard deviations (or increases one's z-score by 0.23), everything else being equal.

Note that this coefficient is very similar to the equivalent value obtained in the ordered logit model. A shift of 0.37 in a standard logistic distribution, where $\sigma = \pi/\sqrt{3} = 1.81$, is equivalent to a shift of $0.37/1.81 = 0.21$ standard deviations, which in turn is very similar to the ordered probit estimate of 0.23 standard deviations. A similar comment applies to the other coefficients. You may also wish to compare the Wald tests for the individual coefficients in Tables 6.5 and 6.6, which are practically identical.

TABLE 6.6: Parameter Estimates for Ordered Probit Model (Latent Variable Formulation)

Parameter	Estimate	Std. Error	z-ratio
Apartments	-.7281	.1205	-6.042
Atrium house	-.3721	.1510	-2.464
Terraced house	-.9790	.1456	-6.725
Influence medium	-.0864	.1303	-0.663
Influence high	.5165	.1639	3.150
Contact high	.2285	.0583	3.918
Apart × Influence med	.6600	.1626	4.060
Apart × Influence high	.4479	.1971	2.273
Atrium × Influence med	.4109	.2134	1.925
Atrium × Influence high	-.0780	.2496	-0.312
Terrace × Influence med	.4964	.2016	2.462
Terrace × Influence high	.5217	.2587	2.016
Cutpoint 1	-.5440	.1023	
Cutpoint 2	.1892	.1018	

6.5.5 Proportional Hazards

A third possible choice of link is the complementary log-log link, which leads to the model

$$\log(-\log(1-\gamma_{ij})) = \theta_j + \mathbf{x}_i'\boldsymbol{\beta} \tag{6.28}$$

This model may be interpreted in terms of a latent variable having a (reversed) extreme value (log Weibull) distribution, with cdf

$$F(\eta) = 1 - \exp\{-\exp\{\eta\}\} \tag{6.29}$$

This distribution is asymmetric, it has mean equal to negative Euler's constant -0.57722 and variance $\pi^2/6 = 1.6449$. The median is $\log\log 2 = -0.3665$ and the quartiles are -1.2459 and 0.3266. Note that the inverse cdf is indeed, the complementary log-log transformation in Equation (6.28).

This model can also be interpreted in terms of a proportional hazards model. The hazard function plays a central role in survival analysis, and will be discussed in detail in the next Chapter.

6.5.6 Extensions and Other Approaches

The general cumulative link model of Section 6.5.2 will work with any monotone link function mapping probabilities to the real line, but the three choices mentioned here, namely the logit, probit, and complementary log-log, are by far the most popular ones. McCullagh (1980) has extended the basic model by relaxing the assumption of constant variance for the latent continuous variable. His most general model allows a separate scale parameter for each multinomial observation, so that

$$g(\gamma_{ij}) = \frac{\theta_j + \mathbf{x}_i'\boldsymbol{\beta}}{\tau_i} \tag{6.30}$$

where the τ_i are unknown scale parameters. A constraint, such as $\tau_1 = 0$, is required for identification. More generally, τ_i may be allowed to depend on a vector of covariates.

An alternative approach to the analysis of ordinal data is to assign scores to the response categories and then use linear regression to model the mean score. Ordinary least squares procedures are not appropriate in this case, but Grizzle et al. (1969) have proposed weighted least-squares procedures that make proper allowances for the underlying independent multinomial sampling scheme. For an excellent discussion of these models see Agresti (1990, Section 9.6).

A similar approach, used often in two-way contingency tables corresponding to one predictor and one response, is to assign scores to the rows and columns of the table and to use these scores to model the interaction term in the usual log-linear model. Often the scores assigned to the columns are the integers $1, 2, \ldots, J-1$, but other choices are possible. If integer scores are used for both rows and columns, then the resulting model has an interesting property, which has been referred to as *uniform association*. Consider calculating an odds ratio for adjacent rows i and $i+1$, across *adjacent* columns or response categories j and $j+1$, that is

$$\rho_{ij} = \frac{\pi_{i,j}/\pi_{i,j+1}}{\pi_{i+1,j}/\pi_{i+1,j+1}} \tag{6.31}$$

Under the additive log-linear model of independence, this ratio is unity for all i and j. Introducing an interaction term based on integer scores, of the form $(\alpha\beta)_{ij} = \gamma ij$, makes the odds ratio constant across adjacent categories. This model often produces fits similar to the proportional odds model, but the parameters are not so easily interpreted. For further details see Haberman (1974), Goodman (1979) or Agresti (1990, Section 8.1).

A fourth family of models for ordinal responses follows from introducing constraints in the multinomial logit model. Let $\boldsymbol{\beta}_j$ denote the vector of coefficients for the j-th equation, comparing the j-th category with the last category, for $j = 1, 2, \ldots, J-1$. The most restrictive model assumes that these coefficients are the same for all contrasts, so $\boldsymbol{\beta}_j = \boldsymbol{\beta}$ for all j. A less restrictive assumption is that the coefficients have a linear trend over the categories, so that $\boldsymbol{\beta}_j = j\boldsymbol{\beta}$. Anderson (1984) has proposed a model termed the *stereotype* model where the coefficients are proportional across categories, so $\boldsymbol{\beta}_j = \gamma_j \boldsymbol{\beta}$, with unknown proportionality factors given by scalars γ_j.

One advantage of the cumulative link models considered here is that the parameter estimates refer to the cumulative distribution of the manifest response (or the distribution of the underlying latent variable) and therefore are not heavily dependent on the actual categories (or cutpoints) used. In particular, we would not expect the results to change much if we were to combine two adjacent categories, or if we recoded the response using fewer categories. If the cumulative odds are indeed proportional before collapsing categories, the argument goes, they should continue to be so afterwards.

In contrast, inferences based on log-linear or multinomial logit models apply only to the actual categories used. It is quite possible, for example, that odds ratios that are relatively constant across adjacent categories will no longer exhibit this property if some of the categories are combined. These considerations are particularly relevant if the categories used are somewhat arbitrary.

Chapter 7

Survival Models

Our final chapter concerns models for the analysis of data which have three main characteristics: (1) the dependent variable or response is the waiting *time* until the occurrence of a well-defined event, (2) observations are *censored*, in the sense that for some units the event of interest has not occurred at the time the data are analyzed, and (3) there are predictors or *explanatory* variables whose effect on the waiting time we wish to assess or control. We start with some basic definitions.

7.1 The Hazard and Survival Functions

Let T be a non-negative random variable representing the waiting time until the occurrence of an event. For simplicity we will adopt the terminology of survival analysis, referring to the event of interest as 'death' and to the waiting time as 'survival' time, but the techniques to be studied have much wider applicability. They can be used, for example, to study age at marriage, the duration of marriage, the intervals between successive births to a woman, the duration of stay in a city (or in a job), and the length of life. The observant demographer will have noticed that these examples include the fields of fertility, mortality and migration.

7.1.1 The Survival Function

We will assume for now that T is a continuous random variable with probability density function (p.d.f.) $f(t)$ and cumulative distribution function (c.d.f.) $F(t) = \Pr\{T < t\}$, giving the probability that the event has occurred by duration t.

It will often be convenient to work with the complement of the c.d.f, the *survival* function

$$S(t) = \Pr\{T \geq t\} = 1 - F(t) = \int_t^\infty f(x)dx, \tag{7.1}$$

which gives the probability of being alive just before duration t, or more generally, the probability that the event of interest has not occurred by duration t.

7.1.2 The Hazard Function

An alternative characterization of the distribution of T is given by the *hazard* function, or instantaneous rate of occurrence of the event, defined as

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt | T \geq t\}}{dt}. \tag{7.2}$$

The numerator of this expression is the conditional probability that the event will occur in the interval $[t, t + dt)$ given that it has not occurred before, and the denominator is the width of the interval. Dividing one by the other we obtain a rate of event occurrence per unit of time. Taking the limit as the width of the interval goes down to zero, we obtain an instantaneous rate of occurrence.

The conditional probability in the numerator may be written as the ratio of the joint probability that T is in the interval $[t, t + dt)$ *and* $T \geq t$ (which is, of course, the same as the probability that t is in the interval), to the probability of the condition $T \geq t$. The former may be written as $f(t)dt$ for small dt, while the latter is $S(t)$ by definition. Dividing by dt and passing to the limit gives the useful result

$$\lambda(t) = \frac{f(t)}{S(t)}, \tag{7.3}$$

which some authors give as a definition of the hazard function. In words, the rate of occurrence of the event at duration t equals the density of events at t, divided by the probability of surviving to that duration without experiencing the event.

Note from Equation 7.1 that $-f(t)$ is the derivative of $S(t)$. This suggests rewriting Equation 7.3 as

$$\lambda(t) = -\frac{d}{dt} \log S(t).$$

If we now integrate from 0 to t and introduce the boundary condition $S(0) = 1$ (since the event is sure not to have occurred by duration 0), we can solve the above expression to obtain a formula for the probability of surviving to duration t as a function of the hazard at all durations up to t:

$$S(t) = \exp\{-\int_0^t \lambda(x)dx\}. \tag{7.4}$$

This expression should be familiar to demographers. The integral in curly brackets in this equation is called the *cumulative hazard* (or cumulative risk) and is denoted

$$\Lambda(t) = \int_0^t \lambda(x)dx. \tag{7.5}$$

You may think of $\Lambda(t)$ as the sum of the risks you face going from duration 0 to t.

These results show that the survival and hazard functions provide alternative but equivalent characterizations of the distribution of T. Given the survival function, we can always differentiate to obtain the density and then calculate the hazard using Equation 7.3. Given the hazard, we can always integrate to obtain the cumulative hazard and then exponentiate to obtain the survival function using Equation 7.4. An example will help fix ideas.
Example: The simplest possible survival distribution is obtained by assuming a constant risk over time, so the hazard is

$$\lambda(t) = \lambda$$

for all t. The corresponding survival function is

$$S(t) = \exp\{-\lambda t\}.$$

This distribution is called the exponential distribution with parameter λ. The density may be obtained multiplying the survivor function by the hazard to obtain

$$f(t) = \lambda \exp\{-\lambda t\}.$$

The mean turns out to be $1/\lambda$. This distribution plays a central role in survival analysis, although it is probably too simple to be useful in applications in its own right.□

7.1.3 Expectation of Life

Let μ denote the mean or expected value of T. By definition, one would calculate μ multiplying t by the density $f(t)$ and integrating, so

$$\mu = \int_0^\infty tf(t)dt.$$

Integrating by parts, and making use of the fact that $-f(t)$ is the derivative of $S(t)$, which has limits or boundary conditions $S(0) = 1$ and $S(\infty) = 0$, one can show that

$$\mu = \int_0^\infty S(t)dt. \tag{7.6}$$

In words, the mean is simply the integral of the survival function.

7.1.4 A Note on Improper Random Variables*

So far we have assumed implicitly that the event of interest is bound to occur, so that $S(\infty) = 0$. In words, given enough time the proportion surviving goes down to zero. This condition implies that the cumulative hazard must diverge, i.e. we must have $\Lambda(\infty) = \infty$. Intuitively, the event will occur with certainty only if the cumulative risk over a long period is sufficiently high.

There are, however, many events of possible interest that are not bound to occur. Some men and women remain forever single, some birth intervals never close, and some people are happy enough at their jobs that they never leave. What can we do in these cases? There are two approaches one can take.

One approach is to note that we can still calculate the hazard and survival functions, which are well defined even if the event of interest is not bound to occur. For example we can study marriage in the entire population, which includes people who will never marry, and calculate marriage rates and proportions single. In this example $S(t)$ would represent the proportion still single at age t and $S(\infty)$ would represent the proportion who never marry.

One limitation of this approach is that if the event is not certain to occur, then the waiting time T could be undefined (or infinite) and thus not a proper random variable. Its density, which could be calculated from the hazard and survival, would be improper, i.e. it would fail to integrate to one. Obviously, the mean waiting time would not be defined. In terms of our example, we cannot calculate mean age at marriage for the entire population, simply because not everyone marries. But this limitation is of no great consequence if interest centers on the hazard and survivor functions, rather than the waiting time. In the marriage example we can even calculate a median age at marriage, provided we define it as the age by which half the population has married.

The alternative approach is to condition the analysis on the event actually occurring. In terms of our example, we could study marriage (perhaps retrospectively) for people who eventually marry, since for this group the

actual waiting time T is always well defined. In this case we can calculate not just the conditional hazard and survivor functions, but also the mean. In our marriage example, we could calculate the mean age at marriage for those who marry. We could even calculate a conventional median, defined as the age by which half the people who will eventually marry have done so.

It turns out that the conditional density, hazard and survivor function for those who experience the event are related to the unconditional density, hazard and survivor for the entire population. The conditional density is

$$f^*(t) = \frac{f(t)}{1 - S(\infty)},$$

and it integrates to one. The conditional survivor function is

$$S^*(t) = \frac{S(t) - S(\infty)}{1 - S(\infty)},$$

and goes down to zero as $t \to \infty$. Dividing the density by the survivor function, we find the conditional hazard to be

$$\lambda^*(t) = \frac{f^*(t)}{S^*(t)} = \frac{f(t)}{S(t) - S(\infty)}.$$

Derivation of the mean waiting time for those who experience the event is left as an exercise for the reader.

Whichever approach is adopted, care must be exercised to specify clearly which hazard or survival is being used. For example, the conditional hazard for those who eventually experience the event is always higher than the unconditional hazard for the entire population. Note also that in most cases all we observe is whether or not the event has occurred. If the event has not occurred, we may be unable to determine whether it will eventually occur. In this context, only the unconditional hazard may be estimated from data, but one can always translate the results into conditional expressions, if so desired, using the results given above.

7.2 Censoring and The Likelihood Function

The second distinguishing feature of the field of survival analysis is censoring: the fact that for some units the event of interest has occurred and therefore we know the exact waiting time, whereas for others it has not occurred, and all we know is that the waiting time exceeds the observation time.

7.2.1 Censoring Mechanisms

There are several mechanisms that can lead to censored data. Under censoring of *Type I*, a sample of n units is followed for a fixed time τ. The number of units experiencing the event, or the number of 'deaths', is random, but the total duration of the study is fixed. The fact that the duration is fixed may be an important practical advantage in designing a follow-up study.

In a simple generalization of this scheme, called *fixed censoring*, each unit has a potential maximum observation time τ_i for $i = 1, \ldots, n$ which may differ from one case to the next but is nevertheless fixed in advance. The probability that unit i will be alive at the end of her observation time is $S(\tau_i)$, and the total number of deaths is again random.

Under censoring of *Type II*, a sample of n units is followed as long as necessary until d units have experienced the event. In this design the number of deaths d, which determines the precision of the study, is fixed in advance and can be used as a design parameter. Unfortunately, the total duration of the study is then random and cannot be known with certainty in advance.

In a more general scheme called *random censoring*, each unit has associated with it a potential censoring time C_i and a potential lifetime T_i, which are assumed to the independent random variables. We observe $Y_i = \min\{C_i, T_i\}$, the minimum of the censoring and life times, and an indicator variable, often called d_i or δ_i, that tells us whether observation terminated by death or by censoring.

All these schemes have in common the fact that the censoring mechanism is *non-informative* and they all lead to essentially the same likelihood function. The weakest assumption required to obtain this common likelihood is that the censoring of an observation should not provide any information regarding the prospects of survival of that particular unit beyond the censoring time. In fact, the basic assumption that we will make is simply this: all we know for an observation censored at duration t is that the lifetime exceeds t.

7.2.2 The Likelihood Function for Censored Data

Suppose then that we have n units with lifetimes governed by a survivor function $S(t)$ with associated density $f(t)$ and hazard $\lambda(t)$. Suppose unit i is observed for a time t_i. If the unit died at t_i, its contribution to the likelihood function is the density at that duration, which can be written as the product of the survivor and hazard functions

$$L_i = f(t_i) = S(t_i)\lambda(t_i).$$

If the unit is still alive at t_i, all we know under non-informative censoring is that the lifetime exceeds t_i. The probability of this event is

$$L_i = S(t_i),$$

which becomes the contribution of a censored observation to the likelihood.

Note that both types of contribution share the survivor function $S(t_i)$, because in both cases the unit lived up to time t_i. A death multiplies this contribution by the hazard $\lambda(t_i)$, but a censored observation does not. We can write the two contributions in a single expression. To this end, let d_i be a death indicator, taking the value one if unit i died and the value zero otherwise. Then the likelihood function may be written as follows

$$L = \prod_{i=1}^{n} L_i = \prod_i \lambda(t_i)^{d_i} S(t_i).$$

Taking logs, and recalling the expression linking the survival function $S(t)$ to the cumulative hazard function $\Lambda(t)$, we obtain the log-likelihood function for censored survival data

$$\log L = \sum_{i=1}^{n} \{d_i \log \lambda(t_i) - \Lambda(t_i)\}. \tag{7.7}$$

We now consider an example to reinforce these ideas.

Example: Suppose we have a sample of n censored observations from an exponential distribution. Let t_i be the observation time and d_i the death indicator for unit i.

In the exponential distribution $\lambda(t) = \lambda$ for all t. The cumulative risk turns out to be the integral of a constant and is therefore $\Lambda(t) = \lambda t$. Using these two results on Equation 7.7 gives the log-likelihood function

$$\log L = \sum \{d_i \log \lambda - \lambda t_i\}.$$

Let $D = \sum d_i$ denote the total number of deaths, and let $T = \sum t_i$ denote the total observation (or exposure) time. Then we can rewrite the log-likelihood as a function of these totals to obtain

$$\log L = D \log \lambda - \lambda T. \tag{7.8}$$

Differentiating this expression with respect to λ we obtain the score function

$$u(\lambda) = \frac{D}{\lambda} - T,$$

and setting the score to zero gives the maximum likelihood estimator of the hazard

$$\hat{\lambda} = \frac{D}{T}, \tag{7.9}$$

the total number of deaths divided by the total exposure time. Demographers will recognize this expression as the general definition of a death rate. Note that the estimator is optimal (in a maximum likelihood sense) only if the risk is constant and does not depend on age.

We can also calculate the observed information by taking minus the second derivative of the score, which is

$$I(\lambda) = \frac{D}{\lambda^2}.$$

To obtain the expected information we need to calculate the expected number of deaths, but this depends on the censoring scheme. For example under Type I censoring with fixed duration τ, one would expect $n(1-S(\tau))$ deaths. Under Type II censoring the number of deaths would have been fixed in advance. Under some schemes calculation of the expectation may be fairly complicated if not impossible.

A simpler alternative is to use the observed information, estimated using the m.l.e. of λ given in Equation 7.9. Using this approach, the large sample variance of the m.l.e. of the hazard rate may be estimated as

$$\hat{\text{var}}(\hat{\lambda}) = \frac{D}{T^2},$$

a result that leads to large-sample tests of hypotheses and confidence intervals for λ.

If there are no censored cases, so that $d_i = 1$ for all i and $D = n$, then the results obtained here reduce to standard maximum likelihood estimation for the exponential distribution, and the m.l.e. of λ turns out to be the reciprocal of the sample mean.

It may be interesting to note in passing that the log-likelihood for censored exponential data given in Equation 7.8 coincides exactly (except for constants) with the log-likelihood that would be obtained by treating D as a Poisson random variable with mean λT. To see this point, you should write the Poisson log-likelihood when $D \sim P(\lambda T)$, and note that it differs from Equation 7.8 only in the presence of a term $D \log(T)$, which is a constant depending on the data but not on the parameter λ.

Thus, treating the deaths as Poisson conditional on exposure time leads to exactly the same estimates (and standard errors) as treating the exposure

times as censored observations from an exponential distribution. This result will be exploited below to link survival models to generalized linear models with Poisson error structure.

7.3 Approaches to Survival Modeling

Up to this point we have been concerned with a homogeneous population, where the lifetimes of all units are governed by the same survival function $S(t)$. We now introduce the third distinguishing characteristic of survival models—the presence of a vector of covariates or explanatory variables that may affect survival time—and consider the general problem of modeling these effects.

7.3.1 Accelerated Life Models*

Let T_i be a random variable representing the (possibly unobserved) survival time of the i-th unit. Since T_i must be non-negative, we might consider modeling its logarithm using a conventional linear model, say

$$\log T_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i,$$

where ϵ_i is a suitable error term, with a distribution to be specified. This model specifies the distribution of log-survival for the i-th unit as a simple *shift* of a standard or baseline distribution represented by the error term.

Exponentiating this equation, we obtain a model for the survival time itself

$$T_i = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}T_{0i},$$

where we have written T_{0i} for the exponentiated error term. It will also be convenient to use γ_i as shorthand for the multiplicative effect $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ of the covariates.

Interpretation of the parameters follows along standard lines. Consider, for example, a model with a constant and a dummy variable x representing a factor with two levels, say groups one and zero. Suppose the corresponding multiplicative effect is $\gamma = 2$, so the coefficient of x is $\beta = \log(2) = 0.6931$. Then we would conclude that people in group one live twice as long as people in group zero.

There is an interesting alternative interpretation that explains the name 'accelerated life' used for this model. Let $S_0(t)$ denote the survivor function in group zero, which will serve as a reference group, and let $S_1(t)$ denote the

survivor function in group one. Under this model,

$$S_1(t) = S_0(t/\gamma).$$

In words, the probability that a member of group one will be alive at age t is exactly the same as the probability that a member of group zero will be alive at age t/γ. For $\gamma = 2$, this would be half the age, so the probability that a member of group one would be alive at age 40 (or 60) would be the same as the probability that a member of group zero would be alive at age 20 (or 30). Thus, we may think of γ as affecting the passage of time. In our example, people in group zero age 'twice as fast'.

For the record, the corresponding hazard functions are related by

$$\lambda_1(t) = \lambda_0(t/\gamma)/\gamma,$$

so if $\gamma = 2$, at any given age people in group one would be exposed to half the risk of people in group zero half their age.

The name 'accelerated life' stems from industrial applications where items are put to test under substantially worse conditions than they are likely to encounter in real life, so that tests can be completed in a shorter time.

Different kinds of parametric models are obtained by assuming different distributions for the error term. If the ϵ_i are normally distributed, then we obtain a log-normal model for the T_i. Estimation of this model for censored data by maximum likelihood is known in the econometric literature as a Tobit model.

Alternatively, if the ϵ_i have an extreme value distribution with p.d.f.

$$f(\epsilon) = \exp\{\epsilon - \exp\{\epsilon\}\},$$

then T_{0i} has an exponential distribution, and we obtain the exponential regression model, where T_i is exponential with hazard λ_i satisfying the log-linear model

$$\log \lambda_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

An example of a demographic model that belongs to the family of accelerated life models is the Coale-McNeil model of first marriage frequencies, where the proportion ever married at age a in a given population is written as

$$F(a) = cF_0(\frac{a - a_0}{k}),$$

where F_0 is a model schedule of proportions married by age, among women who will ever marry, based on historical data from Sweden; c is the proportion who will eventually marry, a_0 is the age at which marriage starts, and k is the *pace* at which marriage proceeds relative to the Swedish standard.

Accelerated life models are essentially standard regression models applied to the log of survival time, and except for the fact that observations are censored, pose no new estimation problems. Once the distribution of the error term is chosen, estimation proceeds by maximizing the log-likelihood for censored data described in the previous subsection. For further details, see Kalbfleish and Prentice (1980).

7.3.2 Proportional Hazard Models

A large family of models introduced by Cox (1972) focuses directly on the hazard function. The simplest member of the family is the *proportional hazards* model, where the hazard at time t for an individual with covariates $\mathbf{x}_i$ (not including a constant) is assumed to be

$$\lambda_i(t|\mathbf{x}_i) = \lambda_0(t) \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}. \tag{7.10}$$

In this model $\lambda_0(t)$ is a baseline hazard function that describes the risk for individuals with $\mathbf{x}_i = \mathbf{0}$, who serve as a reference cell or pivot, and $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ is the relative risk, a proportionate increase or reduction in risk, associated with the set of characteristics $\mathbf{x}_i$. Note that the increase or reduction in risk is the same at all durations t.

To fix ideas consider a two-sample problem where we have a dummy variable x which serves to identify groups one and zero. Then the model is

$$\lambda_i(t|x) = \begin{cases} \lambda_0(t) & \text{if } x = 0, \\ \lambda_0(t)e^{\beta} & \text{if } x = 1. \end{cases}.$$

Thus, $\lambda_0(t)$ represents the risk at time t in group zero, and $\gamma = \exp\{\beta\}$ represents the ratio of the risk in group one relative to group zero at any time t. If $\gamma = 1$ (or $\beta = 0$) then the risks are the same in the two groups. If $\gamma = 2$ (or $\beta = 0.6931$), then the risk for an individual in group one at any given age is twice the risk of a member of group zero who has the same age.

Note that the model separates clearly the effect of time from the effect of the covariates. Taking logs, we find that the proportional hazards model is a simple additive model for the log of the hazard, with

$$\log \lambda_i(t|\mathbf{x}_i) = \alpha_0(t) + \mathbf{x}_i'\boldsymbol{\beta},$$

where $\alpha_0(t) = \log \lambda_0(t)$ is the log of the baseline hazard. As in all additive models, we assume that the effect of the covariates $\mathbf{x}$ is the same at all times or ages t. The similarity between this expression and a standard analysis of covariance model with parallel lines should not go unnoticed.

Returning to Equation 7.10, we can integrate both sides from 0 to t to obtain the cumulative hazards

$$\Lambda_i(t|\mathbf{x}_i) = \Lambda_0(t) \exp\{\mathbf{x}_i'\boldsymbol{\beta}\},$$

which are also proportional. Changing signs and exponentiating we obtain the survivor functions

$$S_i(t|\mathbf{x}_i) = S_0(t)^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}, \tag{7.11}$$

where $S_0(t) = \exp\{-\Lambda_0(t)\}$ is a baseline survival function. Thus, the effect of the covariate values $\mathbf{x}_i$ on the survivor function is to raise it to a power given by the relative risk $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$.

In our two-group example with a relative risk of $\gamma = 2$, the probability that a member of group one will be alive at any given age t is the square of the probability that a member of group zero would be alive at the same age.

7.3.3 The Exponential and Weibull Models

Different kinds of proportional hazard models may be obtained by making different assumptions about the baseline survival function, or equivalently, the baseline hazard function. For example if the baseline risk is constant over time, so $\lambda_0(t) = \lambda_0$, say, we obtain the exponential regression model, where

$$\lambda_i(t, \mathbf{x}_i) = \lambda_0 \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}.$$

Interestingly, the exponential regression model belongs to both the proportional hazards and the accelerated life families. If the baseline risk is a constant and you double or triple the risk, the new risk is still constant (just higher). Perhaps less obviously, if the baseline risk is constant and you imagine time flowing twice or three times as fast, the new risk is doubled or tripled but is still constant over time, so we remain in the exponential family.

You may be wondering whether there are other cases where the two models coincide. The answer is yes, but not many. In fact, there is only one distribution where they do, and it includes the exponential as a special case.

The one case where the two families coincide is the *Weibull* distribution, which has survival function

$$S(t) = \exp\{-(\lambda t)^p\}$$

and hazard function

$$\lambda(t) = p\lambda(\lambda t)^{p-1},$$

for parameters $\lambda > 0$ and $p > 0$. If $p = 1$, this model reduces to the exponential and has constant risk over time. If $p > 1$, then the risk increases over time. If $p < 1$, then the risk decreases over time. In fact, taking logs in the expression for the hazard function, we see that the log of the Weibull risk is a linear function of log time with slope $p - 1$.

If we pick the Weibull as a baseline risk and then multiply the hazard by a constant γ in a proportional hazards framework, the resulting distribution turns out to be still a Weibull, so the family is closed under proportionality of hazards. If we pick the Weibull as a baseline survival and then speed up the passage of time in an accelerated life framework, dividing time by a constant γ, the resulting distribution is still a Weibull, so the family is closed under acceleration of time.

For further details on this distribution see Cox and Oakes (1984) or Kalbfleish and Prentice (1980), who prove the equivalence of the two Weibull models.

7.3.4 Time-varying Covariates

So far we have considered explicitly only covariates that are fixed over time. The local nature of the proportional hazards model, however, lends itself easily to extensions that allows for covariates that change over time. Let us consider a few examples.

Suppose we are interested in the analysis of birth spacing, and study the interval from the birth of one child to the birth of the next. One of the possible predictors of interest is the mother's education, which in most cases can be taken to be fixed over time.

Suppose, however, that we want to introduce breastfeeding status of the child that begins the interval. Assuming the child is breastfed, this variable would take the value one ('yes') from birth until the child is weaned, at which time it would take the value zero ('no'). This is a simple example of a predictor that can change value only once.

A more elaborate analysis could rely on frequency of breastfeeding in a 24-hour period. This variable could change values from day to day. For example a sequence of values for one woman could be 4,6,5,6,5,4,...

Let $\mathbf{x}_i(t)$ denote the value of a vector of covariates for individual i at time or duration t. Then the proportional hazards model may be generalized to

$$\lambda_i(t, \mathbf{x}_i(t)) = \lambda_0(t) \exp\{\mathbf{x}_i(t)'\boldsymbol{\beta}\}. \tag{7.12}$$

The separation of duration and covariate effects is not so clear now, and on occasion it may be difficult to identify effects that are highly collinear with time. If all children were weaned when they are around six months old, for example, it would be difficult to identify effects of breastfeeding from general duration effects without additional information. In such cases one might still prefer a time-varying covariate, however, as a more meaningful predictor of risk than the mere passage of time.

Calculation of survival functions when we have time-varying covariates is a little bit more complicated, because we need to specify a path or trajectory for each variable. In the birth intervals example one could calculate a survival function for women who breastfeed for six months and then wean. This would be done by using the hazard corresponding to $x(t) = 0$ for months 0 to 6 and then the hazard corresponding to $x(t) = 1$ for months 6 onwards. Unfortunately, the simplicity of Equation 7.11 is lost; we can no longer simply raise the baseline survival function to a power.

Time-varying covariates can be introduced in the context of accelerated life models, but this is not so simple and has rarely been done in applications. See Cox and Oakes (1984, p.66) for more information.

7.3.5 Time-dependent Effects

The model may also be generalized to allow for *effects* that vary over time, and therefore are no longer proportional. It is quite possible, for example, that certain social characteristics might have a large impact on the hazard for children shortly after birth, but may have a relatively small impact later in life. To accommodate such models we may write

$$\lambda_i(t, \mathbf{x}_i) = \lambda_0(t) \exp\{\mathbf{x}_i'\boldsymbol{\beta}(t)\},$$

where the parameter $\boldsymbol{\beta}(t)$ is now a function of time.

This model allows for great generality. In the two-sample case, for example, the model may be written as

$$\lambda_i(t|x) = \begin{cases} \lambda_0(t) & \text{if } x = 0 \\ \lambda_0(t)e^{\beta}(t) & \text{if } x = 1 \end{cases},$$

which basically allows for two arbitrary hazard functions, one for each group. Thus, this is a form of saturated model.

Usually the form of time dependence of the effects must be specified parametrically in order to be able to identify the model and estimate the parameters. Obvious candidates are polynomials on duration, where $\beta(t)$ is a linear or quadratic function of time. Cox and Oakes (1984, p. 76) show how one can use quick-dampening exponentials to model transient effects.

Note that we have lost again the simple separation of time and covariate effects. Calculation of the survival function in this model is again somewhat complicated by the fact that the coefficients are now functions of time, so they don't fall out of the integral. The simple Equation 7.11 does not apply.

7.3.6 The General Hazard Rate Model

The foregoing extensions to time-varying covariates and time-dependent effects may be combined to give the most general version of the hazard rate model, as

$$\lambda_i(t, \mathbf{x}_i(t)) = \lambda_0(t) \exp\{\mathbf{x}_i(t)'\boldsymbol{\beta}(t)\},$$

where $\mathbf{x}_i(t)$ is a vector of time-varying covariates representing the characteristics of individual i at time t, and $\boldsymbol{\beta}(t)$ is a vector of time-dependent coefficients, representing the effect that those characteristics have at time or duration t.

The case of breastfeeding status and its effect on the length of birth intervals is a good example that combines the two effects. Breastfeeding status is itself a time-varying covariate $x(t)$, which takes the value one if the woman is breastfeeding her child t months after birth. The effect that breastfeeding may have in inhibiting ovulation and therefore reducing the risk of pregnancy is known to decline rapidly over time, so it should probably be modeled as a time dependent effect $\beta(t)$. Again, further progress would require specifying the form of this function of time.

7.3.7 Model Fitting

There are essentially three approaches to fitting survival models:

- The first and perhaps most straightforward is the *parametric* approach, where we assume a specific functional form for the baseline hazard $\lambda_0(t)$. Examples are models based on the exponential, Weibull, gamma and generalized F distributions.

- A second approach is what might be called a flexible or *semi-parametric* strategy, where we make mild assumptions about the baseline hazard

$\lambda_0(t)$. Specifically, we may subdivide time into reasonably small intervals and assume that the baseline hazard is constant in each interval, leading to a piece-wise exponential model.

- The third approach is a *non-parametric* strategy that focuses on estimation of the regression coefficients $\boldsymbol{\beta}$ leaving the baseline hazard $\lambda_0(t)$ completely unspecified. This approach relies on a partial likelihood function proposed by Cox (1972) in his original paper.

A complete discussion of these approaches in well beyond the scope of these notes. We will focus on the intermediate or semi-parametric approach because (1) it is sufficiently flexible to provide a useful tool with wide applicability, and (2) it is closely related to Poisson regression analysis.

7.4 The Piece-Wise Exponential Model

We will consider fitting a proportional hazards model of the usual form

$$\lambda_i(t|\mathbf{x}_i) = \lambda_0(t) \exp\{\mathbf{x}_i'\boldsymbol{\beta}\} \tag{7.13}$$

under relatively mild assumptions about the baseline hazard $\lambda_0(t)$.

7.4.1 A Piece-wise Constant Hazard

Consider partitioning duration into J intervals with cutpoints $0 = \tau_0 < \tau_1 < \ldots < \tau_J = \infty$. We will define the j-th interval as $[\tau_{j-1}, \tau_j)$, extending from the $(j-1)$-st boundary to the j-th and including the former but not the latter.

We will then assume that the baseline hazard is *constant* within each interval, so that

$$\lambda_0(t) = \lambda_j \quad \text{for } t \text{ in } \quad [\tau_{j-1}, \tau_j). \tag{7.14}$$

Thus, we model the baseline hazard $\lambda_0(t)$ using J parameters $\lambda_1, \ldots, \lambda_J$, each representing the risk for the reference group (or individual) in one particular interval. Since the risk is assumed to be piece-wise constant, the corresponding survival function is often called a piece-wise exponential.

Clearly, judicious choice of the cutpoints should allow us to approximate reasonably well almost any baseline hazard, using closely-spaced boundaries where the hazard varies rapidly and wider intervals where the hazard changes more slowly.

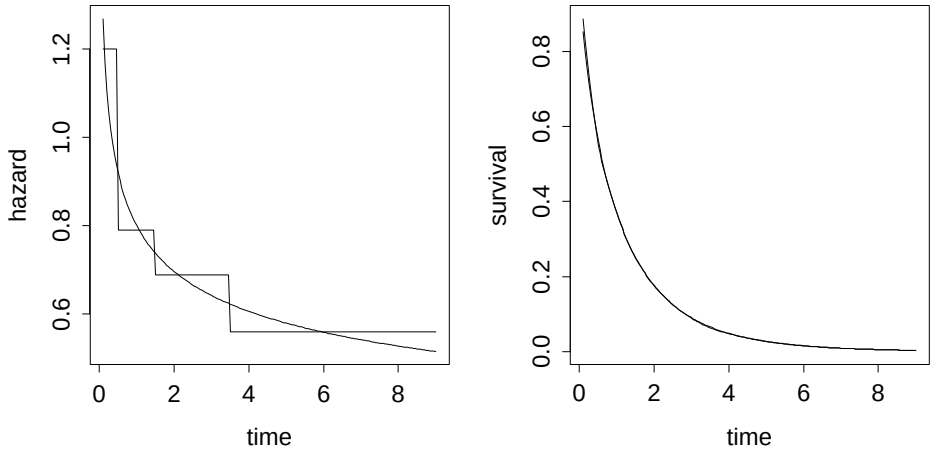


FIGURE 7.1: Approximating a Survival Curve Using a Piece-wise Constant Hazard Function

Figure 7.1 shows how a Weibull distribution with $\lambda = 1$ and $p = 0.8$ can be approximated using a piece-wise exponential distribution with boundaries at 0.5, 1.5 and 3.5. The left panel shows how the piece-wise constant hazard can follow only the broad outline of the smoothly declining Weibull hazard yet, as shown on the right panel, the corresponding survival curves are indistinguishable.

7.4.2 A Proportional Hazards Model

let us now introduce some covariates in the context of the proportional hazards model in Equation 7.13, assuming that the baseline hazard is piece-wise constant as in Equation 7.14. We will write the model as

$$\lambda_{ij} = \lambda_j \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}, \tag{7.15}$$

where λ_{ij} is the hazard corresponding to individual i in interval j, λ_j is the baseline hazard for interval j, and $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ is the relative risk for an individual with covariate values $\mathbf{x}_i$, compared to the baseline, at any given time.

Taking logs, we obtain the additive log-linear model

$$\log \lambda_{ij} = \alpha_j + \mathbf{x}_i'\boldsymbol{\beta}, \tag{7.16}$$

where $\alpha_j = \log \lambda_j$ is the log of the baseline hazard. Note that the result is a standard log-linear model where the duration categories are treated as

a factor. Since we have not included an explicit constant, we do not have to impose restrictions on the α_j. If we wanted to introduce a constant representing the risk in the first interval then we would set $\alpha_1 = 0$, as usual.

The model can be extended to introduce time-varying covariates and time-dependent effects, but we will postpone discussing the details until we study estimation of the simpler proportional hazards model.

7.4.3 The Equivalent Poisson Model

Holford (1980) and Laird and Oliver (1981), in papers produced independently and published very close to each other, noted that the piece-wise proportional hazards model of the previous subsection was equivalent to a certain Poisson regression model. We first state the result and then sketch its proof.

Recall that we observe t_i, the total time lived by the i-th individual, and d_i, a death indicator that takes the value one if the individual died and zero otherwise. We will now define analogous measures for each interval that individual i goes through. You may think of this process as creating a bunch of pseudo-observations, one for each combination of individual and interval.

First we create measures of exposure. Let t_{ij} denote the time lived by the i-th individual in the j-th interval, that is, between τ_{j-1} and τ_j. If the individual lived beyond the end of the interval, so that $t_i > \tau_j$, then the time lived in the interval equals the width of the interval and $t_{ij} = \tau_j - \tau_{j-1}$. If the individual died or was censored in the interval, i.e. if $\tau_{j-1} < t_i < \tau_j$, then the timed lived in the interval is $t_{ij} = t_i - \tau_{j-1}$, the difference between the total time lived and the lower boundary of the interval. We only consider intervals actually visited, but obviously the time lived in an interval would be zero if the individual had died before the start of the interval and $t_i < \tau_{j-1}$.

Now we create death indicators. Let d_{ij} take the value one if individual i dies in interval j and zero otherwise. Let $j(i)$ indicate the interval where t_i falls, i.e. the interval where individual i died or was censored. We use functional notation to emphasize that this interval will vary from one individual to another. If t_i falls in interval $j(i)$, say, then d_{ij} must be zero for all $j < j(i)$ (i.e. all prior intervals) and will equal d_i for $j = j(i)$, (i.e. the interval where individual i was last seen).

Then, the piece-wise exponential model may be fitted to data by treating the death indicators d_{ij}'s as if they were independent Poisson observations with means

$$\mu_{ij} = t_{ij}\lambda_{ij},$$

where t_{ij} is the exposure time as defined above and λ_{ij} is the hazard for individual i in interval j. Taking logs in this expression, and recalling that the hazard rates satisfy the proportional hazards model in Equation 7.15, we obtain

$$\log \mu_{ij} = \log t_{ij} + \alpha_j + \mathbf{x}_i'\boldsymbol{\beta},$$

where $\alpha_j = \log \lambda_j$ as before.

Thus, the piece-wise exponential proportional hazards model is equivalent to a Poisson log-linear model for the pseudo observations, one for each combination of individual and interval, where the death indicator is the response and the log of exposure time enters as an offset.

It is important to note that we do not assume that the d_{ij} have independent Poisson distributions, because they clearly do not. If individual i died in interval $j(i)$, then it must have been alive in all prior intervals $j < j(i)$, so the indicators couldn't possibly be independent. Moreover, each indicator can only take the values one and zero, so it couldn't possibly have a Poisson distribution, which assigns some probability to values greater than one. The result is more subtle. It is the likelihood functions that coincide. Given a realization of a piece-wise exponential survival process, we can find a realization of a set of independent Poisson observations that happens to have the same likelihood, and therefore would lead to the same estimates and tests of hypotheses.

The proof is not hard. Recall from Section 7.2.2 that the contribution of the i-th individual to the log-likelihood function has the general form

$$\log L_i = d_i \log \lambda_i(t_i) - \Lambda_i(t_i),$$

where we have written $\lambda_i(t)$ for the hazard and $\Lambda_i(t)$ for the cumulative hazard that applies to the i-th individual at time t. Let $j(i)$ denote the interval where t_i falls, as before.

Under the piece-wise exponential model, the first term in the log-likelihood can be written as

$$d_i \log \lambda_i(t_i) = d_{ij(i)} \log \lambda_{ij(i)},$$

using the fact that the hazard is $\lambda_{ij(i)}$ when t_i is in interval $j(i)$, and that the death indicator d_i applies directly to the last interval visited by individual i, and therefore equals $d_{j(i)}$.

The cumulative hazard in the second term is an integral, and can be written as a sum as follows

$$\Lambda_i(t_i) = \int_0^{t_i} \lambda_i(t)dt = \sum_{j=1}^{j(i)} t_{ij}\lambda_{ij},$$

where t_{ij} is the amount of time spent by individual i in interval j. To see this point note that we need to integrate the hazard from 0 to t_i. We split this integral into a sum of integrals, one for each interval where the hazard is constant. If an individual lives through an interval, the contribution to the integral will be the hazard λ_{ij} multiplied by the width of the interval. If the individual dies or is censored in the interval, the contribution to the integral will be the hazard λ_{ij} multiplied by the time elapsed from the beginning of the interval to the death or censoring time, which is $t_i - \tau_{j-1}$. But this is precisely the definition of the exposure time t_{ij}.

One slight lack of symmetry in our results is that the hazard leads to *one* term on $d_{ij(i)} \log \lambda_{ij(i)}$, but the cumulative hazard leads to $j(i)$ terms, one for each interval from $j = 1$ to $j(i)$. However, we know that $d_{ij} = 0$ for all $j < j(i)$, so we can add terms on $d_{ij} \log \lambda_{ij}$ for all prior j's; as long as $d_{ij} = 0$ they will make no contribution to the log-likelihood. This trick allows us to write the contribution of the i-th individual to the log-likelihood as a sum of $j(i)$ contributions, one for each interval visited by the individual:

$$\log L_i = \sum_{j=1}^{j(i)} \{d_{ij} \log \lambda_{ij} - t_{ij}\lambda_{ij}\}.$$

The fact that the contribution of the individual to the log-likelihood is a *sum* of several terms (so the contribution to the likelihood is a product of several terms) means that we can treat each of the terms as representing an independent observation.

The final step is to identify the contribution of each pseudo-observation, and we note here that it agrees, except for a constant, with the likelihood one would obtain if d_{ij} had a Poisson distribution with mean $\mu_{ij} = t_{ij}\lambda_{ij}$. To see this point write the Poisson log-likelihood as

$$\log L_{ij} = d_{ij} \log \mu_{ij} - \mu_{ij} = d_{ij} \log(t_{ij}\lambda_{ij}) - t_{ij}\lambda_{ij}.$$

This expression agrees with the log-likelihood above except for the term $d_{ij} \log(t_{ij})$, but this is a constant depending on the data and not on the parameters, so it can be ignored from the point of view of estimation. This completes the proof.□

This result generalizes the observation made at the end of Section 7.2.2 noting the relationship between the likelihood for censored exponential data and the Poisson likelihood. The extension is that instead of having just one 'Poisson' death indicator for each individual, we have one for each interval visited by each individual.

Generating pseudo-observations can substantially increase the size of the dataset, perhaps to a point where analysis is impractical. Note, however, that the number of distinct covariate patterns may be modest even when the total number of pseudo-observations is large. In this case one can group observations, adding up the measures of exposure and the death indicators. In this more general setting, we can define d_{ij} as the number of deaths and t_{ij} as the total exposure time of individuals with characteristics $\mathbf{x}_i$ in interval j. As usual with Poisson aggregate models, the estimates, standard errors and likelihood ratio tests would be exactly the same as for individual data. Of course, the model deviances would be different, representing goodness of fit to the aggregate rather than individual data, but this may be a small price to pay for the convenience of working with a small number of units.

7.4.4 Time-varying Covariates

It should be obvious from the previous development that we can easily accommodate time-varying covariates provided they change values only at interval boundaries. In creating the pseudo-observations required to set-up a Poisson log-likelihood, one would normally replicate the vector of covariates $\mathbf{x}_i$, creating copies $\mathbf{x}_{ij}$, one for each interval. However, there is nothing in our development requiring these vectors to be equal. We can therefore redefine $\mathbf{x}_{ij}$ to represent the values of the covariates of individual i in interval j, and proceed as usual, rewriting the model as

$$\log \lambda_{ij} = \alpha_j + \mathbf{x}'_{ij}\boldsymbol{\beta}.$$

Requiring the covariates to change values only at interval boundaries may seem restrictive, but in practice the model is more flexible than it might seem at first, because we can always further split the pseudo observations. For example, if we wished to accommodate a change in a covariate for individual i half-way through interval j, we could split the pseudo-observation into two, one with the old and one with the new values of the covariates. Each half would get its own measure of exposure and its own death indicator, but both would be tagged as belonging to the same interval, so they would get the same baseline hazard. All steps in the above proof would still hold.

Of course, splitting observations further increases the size of the dataset, and there will usually be practical limitations on how far one can push this approach, even if one uses grouped data. An alternative is to use simpler indicators such as the mean value of a covariate in an interval, perhaps lagged to avoid predicting current hazards using future values of covariates.

7.4.5 Time-dependent Effects

It turns out that the piece-wise exponential scheme lends itself easily to the introduction of non-proportional hazards or time-varying effects, provided again that we let the effects vary only at interval boundaries.

To fix ideas, suppose we have a single predictor taking the value x_{ij} for individual i in interval j. Suppose further that this predictor is a dummy variable, so its possible values are one and zero. It doesn't matter for our current purpose whether the value is fixed for the individual or changes from one interval to the next.

In a proportional hazards model we would write

$$\log \lambda_{ij} = \alpha_j + \beta x_{ij},$$

where β represents the effect of the predictor on the log of the hazard at any given time. Exponentiating, we see that the hazard when $x = 1$ is $\exp\{\beta\}$ times the hazard when $x = 0$, and this effect is the same at all times. This is a simple additive model on duration and the predictor of interest.

To allow for a time-dependent effect of the predictor, we would write

$$\log \lambda_{ij} = \alpha_j + \beta_j x_{ij},$$

where β_j represents the effect of the predictor on the hazard during interval j. Exponentiating, we see that the hazard in interval j when $x = 1$ is $\exp\{\beta_j\}$ times the hazard in interval j when $x = 0$, so the effect may vary from one interval to the next. Since the effect of the predictor depends on the interval, we have a form of interaction between the predictor and duration, which might be more obvious if we wrote the model as

$$\log \lambda_{ij} = \alpha_j + \beta x_{ij} + (\alpha\beta)_j x_{ij}.$$

These models should remind you of the analysis of covariance models of Chapter 2. Here α plays the role of the intercept and β the role of the slope. The proportional hazards model has different intercepts and a common slope, so it's analogous to the parallel lines model. The model with a time-dependent effect has different intercepts and different slopes, and is analogous to the model with an interaction.

To sum up, we can accommodate non-proportionality of hazards simply by introducing interactions with duration. Obviously we can also test the assumption of proportionality of hazards by testing the significance of the interactions with duration. We are now ready for an example.

7.5 Infant and Child Mortality in Colombia

We will illustrate the use of piece-wise exponential survival models using data from an analysis of infant and child mortality in Colombia done by Somoza (1980). The data were collected in a 1976 survey conducted as part of the World Fertility Survey. The sample consisted of women between the ages of 15 and 49. The questionnaire included a maternity history, recording for each child ever born to each respondent the sex, date of birth, survival status as of the interview and (if applicable) age at death.

7.5.1 Calculating Events and Exposure

As if often the case with survival data, most of the work goes into preparing the data for analysis. In the present case we started from tables in Somoza's article showing living children classified by current age, and dead children classified by age at death. Both tabulations reported age using the groups shown in Table 7.1, using fine categories early in life, when the risk if high but declines rapidly, and wider categories at later ages. With these two bits of information we were able to tabulate deaths and calculate exposure time by age groups, assuming that children who died or were censored in an interval lived on the average half the length of the interval.

Table 7.1: Infant and Child Deaths and Exposure Time by Age of Child and Birth Cohort, Colombia 1976.

Exact Age	Birth Cohort 1941–59		1960–67		1968-76	
	deaths	exposure	deaths	exposure	deaths	exposure
0–1 m	168	278.4	197	403.2	195	495.3
1–3 m	48	538.8	48	786.0	55	956.7
3–6 m	63	794.4	62	1165.3	58	1381.4
6–12 m	89	1550.8	81	2294.8	85	2604.5
1–2 y	102	3006.0	97	4500.5	87	4618.5
2–5 y	81	8743.5	103	13201.5	70	9814.5
5–10 y	40	14270.0	39	19525.0	10	5802.5

Table 7.1 shows the results of these calculations in terms of the number of deaths and the total number of person-years of exposure to risk between birth and age ten, by categories of age of child, for three groups of children (or cohorts) born in 1941–59, 1960–67 and 1968–76. The purpose of our

analysis will be to assess the magnitude of the expected decline in infant and child mortality across these cohorts, and to study whether mortality has declined uniformly at all ages or more rapidly in certain age groups.

7.5.2 Fitting The Poisson Models

Let y_{ij} denote the number of deaths for cohort i (with $i = 1, 2, 3$) in age group j (for $j = 1, 2, \ldots, 7$). In view of the results of the previous section, we treat the y_{ij} as realizations of Poisson random variables with means μ_{ij} satisfying

$$\mu_{ij} = \lambda_{ij} t_{ij},$$

where λ_{ij} is the hazard rate and t_{ij} is the total exposure time for group i at age j. In words, the expected number of deaths is the product of the death rate by exposure time.

A word of caution about units of measurement: the hazard rates must be interpreted in the same units of time that we have used to measure exposure. In our example we measure time in years and therefore the λ_{ij} represent rates per person-year of exposure. If we had measured time in months the λ_{ij} would represent rates per person-month of exposure, and would be exactly one twelfth the size of the rates per person-year.

To model the rates we use a log link, so that the linear predictor becomes

$$\eta_{ij} = \log \mu_{ij} = \log \lambda_{ij} + \log t_{ij},$$

the sum of two parts, $\log t_{ij}$, an *offset* or known part of the linear predictor, and $\log \lambda_{ij}$, the log of the hazard rates of interest.

Finally, we introduce a log-linear model for the hazard rates, of the usual form

$$\log \lambda_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta},$$

where $\mathbf{x}_{ij}$ is a vector of covariates. In case you are wondering what happened to the baseline hazard, we have folded it into the vector of parameters $\boldsymbol{\beta}$. The vector of covariates $\mathbf{x}_{ij}$ may include a constant, a set of dummy variables representing the age groups (i.e. the shape of the hazard by age), a set of dummy variables representing the birth cohorts (i.e. the change in the hazard over time) and even a set of cross-product dummies representing combinations of ages and birth cohorts (i.e. interaction effects).

Table 7.2 shows the deviance for the five possible models of interest, including the null model, the two one-factor models, the two-factor additive model, and the two-factor model with an interaction, which is saturated for these data.

TABLE 7.2: Deviances for Various Models Fitted to Infant and Child Mortality Data From Colombia

Model	Name	$\log \lambda_{ij}$	Deviance	d.f.
ϕ	Null	η	4239.8	20
A	Age	$\eta + \alpha_i$	72.7	14
C	Cohort	$\eta + \beta_j$	4190.7	18
$A + C$	Additive	$\eta + \alpha_i + \beta_j$	6.2	12
AC	Saturated	$\eta + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	0	0

7.5.3 The Equivalent Survival Models

The null model assumes that the hazard is a constant from birth to age ten and that this constant is the same for all cohorts. It therefore corresponds to an *exponential survival model with no covariates*. This model obviously does not fit the data, the deviance of 4239.8 on 20 d.f. is simply astronomical. The m.l.e. of η is -3.996 with a standard error of 0.0237. Exponentiating we obtain an estimated hazard rate of 0.0184. Thus, we expect about 18 deaths per thousand person-years of exposure. You may want to verify that 0.0184 is simply the ratio of the total number of deaths to the total exposure time. Multiplying 0.0184 by the amount of exposure in each cell of the table we obtain the expected number of deaths. The deviance quoted above is simply twice the sum of observed times the log of observed over expected deaths.

The age model allows the hazard to change from one age group to another, but assumes that the risk at any given age is the same for all cohorts. It is therefore equivalent to *a piece-wise exponential survival model with no covariates*. The reduction in deviance from the null model is 4167.1 on 6 d.f., and is extremely significant. The risk of death varies substantially with age over the first few months of life. In other words the hazard is clearly not constant. Note that with a deviance of 72.7 on 14 d.f., this model does not fit the data. Thus, the assumption that all cohorts are subject to the same risks does not seem tenable.

Table 7.3 shows parameter estimates for the one-factor models A and C and for the additive model $A + C$ in a format reminiscent of multiple classification analysis. Although the A model does not fit the data, it is instructive to comment briefly on the estimates. The constant, shown in parentheses, corresponds to a rate of $\exp\{-0.7427\} = 0.4758$, or nearly half a death per person-year of exposure, in the first month of life. The estimate for ages 1–3 months corresponds to a multiplicative effect of $\exp\{-1.973\} =$

0.1391, amounting to an 86 percent reduction in the hazard after surviving the first month of life. This downward trend continues up to ages 5–10 years, where the multiplicative effect is $\exp\{-5.355\} = 0.0047$, indicating that the hazard at these ages is only half-a-percent what it was in the first month of life. You may wish to verify that the m.l.e.'s of the age effects can be calculated directly from the total number of deaths and the total exposure time in each age group. Can you calculate the deviance by hand?

Let us now consider the model involving only birth cohort, which assumes that the hazard is constant from birth to age ten, but varies from one birth cohort to another. This model is equivalent to *three exponential survival models*, one for each birth cohort. As we would expect, it is hopelessly inadequate, with a deviance in the thousands, because it fails to take into account the substantial age effects that we have just discussed. It may of of interest, however, to note the parameter estimates in Table 7.3. As a first approximation, the overall mortality rate for the older cohort was $\exp\{-3.899\} = 0.0203$ or around 20 deaths per thousand person-years of exposure. The multiplicative effect for the cohort born in 1960–67 is $\exp\{-0.3020\} = 0.7393$, indicating a 26 percent reduction in overall mortality. However, the multiplicative effect for the youngest cohort is $\exp\{0.0742\} = 1.077$, suggesting an eight percent *increase* in overall mortality. Can you think of an explanation for this apparent anomaly? We will consider the answer after we discuss the next model.

TABLE 7.3: Parameter Estimates for Age, Cohort and Age+Cohort Models of Infant and Child Mortality in Colombia

Factor	Category	Gross Effect	Net Effect
Baseline			−0.4485
Age	0–1 m	(−0.7427)	–
	1–3 m	−1.973	−1.973
	3–6 m	−2.162	−2.163
	6–12 m	−2.488	−2.492
	1–2 y	−3.004	−3.014
	2–5 y	−4.086	−4.115
	5–10 y	−5.355	−5.436
Cohort	1941–59	(−3.899)	–
	1960–67	−0.3020	−0.3243
	1968–76	0.0742	−0.4784

Consider now the additive model with effects of both age and cohort, where the hazard rate is allowed to vary with age and may differ from one cohort to another, but the age (or cohort) effect is assumed to be the same for each cohort (or age). This model is equivalent to a *proportional hazards model*, where we assume a common shape of the hazard by age, and let cohort affect the hazard proportionately at all ages. Comparing the proportional hazards model with the age model we note a reduction in deviance of 66.5 on two d.f., which is highly significant. Thus, we have strong evidence of cohort effects net of age. On the other hand, the attained deviance of 6.2 on 12 d.f. is clearly not significant, indicating that the proportional hazards model provides an adequate description of the patterns of mortality by age and cohort in Colombia. In other words, the assumption of proportionality of hazards is quite reasonable, implying that the decline in mortality in Colombia has been the same at all ages.

Let us examine the parameter estimates on the right-most column of Table 7.3. The constant is the baseline hazard at ages 0–1 months for the earliest cohort, those born in 1941–59. The age parameters representing the baseline hazard are practically unchanged from the model with age only, and trace the dramatic decline in mortality from birth to age ten, with half the reduction concentrated in the first year of life. The cohort affects adjusted for age provide a more reasonable picture of the decline in mortality over time. The multiplicative effects for the cohorts born in 1960–67 and 1068–76 are $\exp\{-0.3243\} = 0.7233$ and $\exp\{-0.4784\} = 0.6120$, corresponding to mortality declines of 28 and 38 percent at every age, compared to the cohort born in 1941–59. This is a remarkable decline in infant and child mortality, which appears to have been the same at all ages. In other words, neonatal, post-neonatal, infant and toddler mortality have all declined by approximately 38 percent across these cohorts.

The fact that the gross effect for the youngest cohort was positive but the net effect is substantially negative can be explained as follows. Because the survey took place in 1976, children born between 1968 and 76 have been exposed mostly to mortality at younger ages, where the rates are substantially higher than at older ages. For example a child born in 1975 would have been exposed only to mortality in the first year of life. The gross effect ignores this fact and thus overestimates the mortality of this group at ages zero to ten. The net effect adjusts correctly for the increased risk at younger ages, essentially comparing the mortality of this cohort to the mortality of earlier cohorts when they had the same ages, and can therefore unmask the actual decline.

A final caveat on interpretation: the data are based on retrospective re-

ports of mothers who were between the ages of 15 and 49 at the time of the interview. These women provide a representative sample of both mothers and births for recent periods, but a somewhat biased sample for older periods. The sample excludes mothers who have died before the interview, but also women who were older at the time of birth of the child. For example births from 1976, 1966 and 1956 come from mothers who were under 50, under 40 and under 30 at the time of birth of the child. A more careful analysis of the data would include age of mother at birth of the child as an additional control variable.

7.5.4 Estimating Survival Probabilities

So far we have focused attention on the hazard or mortality rate, but of course, once the hazard has been calculated it becomes an easy task to calculate cumulative hazards and therefore survival probabilities. Table 7.4 shows the results of just such an exercise, using the parameter estimates for the proportional hazards model in Table 7.3.

TABLE 7.4: Calculation of Survival Probabilities for Three Cohorts Based on the Proportional Hazards Model

Age Group	Width	Baseline			Survival for Cohort		
		Log-haz	Hazard	Cum.Haz	<1960	1960–67	1968–76
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0–1 m	1/12	−0.4485	0.6386	0.0532	0.9482	0.9623	0.9676
1–3 m	2/12	−2.4215	0.0888	0.0680	0.9342	0.9520	0.9587
3–6 m	3/12	−2.6115	0.0734	0.0864	0.9173	0.9395	0.9479
6–12 m	1/2	−2.9405	0.0528	0.1128	0.8933	0.9217	0.9325
1–2 y	1	−3.4625	0.0314	0.1441	0.8658	0.9010	0.9145
2–5 y	3	−4.5635	0.0104	0.1754	0.8391	0.8809	0.8970
5–10 y	5	−5.8845	0.0028	0.1893	0.8275	0.8721	0.8893

Consider first the baseline group, namely the cohort of children born before 1960. To obtain the log-hazard for each age group we must add the constant and the age effect, for example the log-hazard for ages 1–3 months is $-0.4485 - 1.973 = -2.4215$. This gives the numbers in column (3) of Table 7.3. Next we exponentiate to obtain the hazard rates in column (4), for example the rate for ages 1–3 months is $\exp\{-2.4215\} = 0.0888$. Next we calculate the cumulative hazard, multiply the hazard by the width of the interval and summing across intervals. In this step it is crucial to express the width of the interval in the same units used to calculate exposure, in

this case years. Thus, the cumulative hazard at then end of ages 1–3 months is $0.6386 \times 1/12 + 0.0888 \times 2/12 = 0.0680$. Finally, we change sign and exponentiate to calculate the survival function. For example the baseline survival function at 3 months is $\exp\{-0.0680\} = 0.9342$.

To calculate the survival functions shown in columns (7) and (8) for the other two cohorts we could multiply the baseline hazards by $\exp\{-0.3242\}$ and $\exp\{-0.4874\}$ to obtain the hazards for cohorts 1960–67 and 1968–76, respectively, and then repeat the steps described above to obtain the survival functions. This approach would be necessary if we had time-varying effects, but in the present case we can take advantage of a simplification that obtains for proportional hazard models. Namely, the survival functions for the two younger cohorts can be calculated as the baseline survival function *raised* to the relative risks $\exp\{-0.3242\}$ and $\exp\{-0.4874\}$, respectively. For example the probability of surviving to age three months was calculated as 0.9342 for the baseline group, and turns out to be $0.9342^{\exp\{-0.3242\}} = 0.9520$ for the cohort born in 1960–67, and $0.9342^{\exp\{-0.4874\}} = 0.9587$ for the cohort born in 1968–76.

Note that the probability of dying in the first year of life has declined from 106.7 per thousand for children born before 1960 to 78.3 per thousand for children born in 1960–67 and finally to 67.5 per thousand for the most recent cohort. Results presented in terms of probabilities are often more accessible to a wider audience than results presented in terms of hazard rates. (Unfortunately, demographers are used to calling the probability of dying in the first year of life the 'infant mortality rate'. This is incorrect because the quantity quoted is a probability, not a rate. In our example the rate varies substantially within the first year of life. If the probability of dying in the first year of life is q, say, then the average rate is approximately $-\log(1-q)$, which is not too different from q for small q.)

By focusing on events and exposure, we have been able to combine infant and child mortality in the same analysis and use all available information. An alternative approach could focus on infant mortality (deaths in the first year of life), and solve the censoring problem by looking only at children born at least one year before the survey, for whom the survival status at age one is know. One could then analyze the probability of surviving to age one using ordinary logit models. A complementary analysis could then look at survival from age one to five, say, working with children born at least five years before the survey who survived to age one, and then analyzing whether or not they further survive to age five, using again a logit model. While simple, this approach does not make full use of the information, relying on cases with complete (uncensored) data. Cox and Oakes (1980) show that

this so-called reduced sample approach can lead to inconsistencies. Another disadvantage of this approach is that it focuses on survival to key ages, but cannot examine the shape of the hazard in the intervening period.

7.6 Discrete Time Models

We discuss briefly two extensions of the proportional hazards model to discrete time, starting with a definition of the hazard and survival functions in discrete time and then proceeding to models based on the logit and the complementary log-log transformations.

7.6.1 Discrete Hazard and Survival

Let T be a discrete random variable that takes the values $t_1 < t_2 < \ldots$ with probabilities

$$f(t_j) = f_j = \Pr\{T = t_j\}.$$

We define the survivor function at time t_j as the probability that the survival time T is at least t_j

$$S(t_j) = S_j = \Pr\{T \geq t_j\} = \sum_{k=j}^{\infty} f_j.$$

Next, we define the hazard at time t_j as the conditional probability of dying at that time given that one has survived to that point, so that

$$\lambda(t_j) = \lambda_j = \Pr\{T = t_j | T \geq t_j\} = \frac{f_j}{S_j}. \tag{7.17}$$

Note that in discrete time the hazard is a conditional probability rather than a rate. However, the general result expressing the hazard as a ratio of the density to the survival function is still valid.

A further result of interest in discrete time is that the survival function at time t_j can be written in terms of the hazard at all prior times $t_1, \ldots, t_{j-1}$, as

$$S_j = (1 - \lambda_1)(1 - \lambda_2) \ldots (1 - \lambda_{j-1}). \tag{7.18}$$

In words, this result states that in order to survive to time t_j one must first survive t_1, then one must survive t_2 given that one survived t_1, and so on, finally surviving t_{j-1} given survival up to that point. This result is analogous to the result linking the survival function in continuous time to the integrated or cumulative hazard at all previous times.

An example of a survival process that takes place in discrete time is time to conception measured in menstrual cycles. In this case the possible values of T are the positive integers, f_j is the probability of conceiving in the j-th cycle, S_j is the probability of conceiving in the j-th cycle or later, and λ_j is the conditional probability of conceiving in the j-th cycle given that conception had not occurred earlier. The result relating the survival function to the hazard states that in order to get to the j-th cycle without conceiving, one has to fail in the first cycle, then fail in the second given that one didn't succeed in the first, and so on, finally failing in the $(j-1)$-st cycle given that one hadn't succeeded yet.

7.6.2 Discrete Survival and Logistic Regression

Cox (1972) proposed an extension of the proportional hazards model to discrete time by working with the conditional odds of dying at each time t_j given survival up to that point. Specifically, he proposed the model

$$\frac{\lambda(t_j|\mathbf{x}_i)}{1-\lambda(t_j|\mathbf{x}_i)} = \frac{\lambda_0(t_j)}{1-\lambda_0(t_j)} \exp\{\mathbf{x}_i'\boldsymbol{\beta}\},$$

where $\lambda(t_j|\mathbf{x}_i)$ is the hazard at time t_j for an individual with covariate values $\mathbf{x}_i$, $\lambda_0(t_j)$ is the baseline hazard at time t_j, and $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ is the relative risk associated with covariate values $\mathbf{x}_i$.

Taking logs, we obtain a model on the *logit* of the hazard or conditional probability of dying at t_j given survival up to that time,

$$\text{logit}\lambda(t_j|\mathbf{x}_i) = \alpha_j + \mathbf{x}_i'\boldsymbol{\beta}, \tag{7.19}$$

where $\alpha_j = \text{logit}\lambda_0(t_j)$ is the logit of the baseline hazard and $\mathbf{x}_i'\boldsymbol{\beta}$ is the effect of the covariates on the logit of the hazard. Note that the model essentially treats time as a discrete factor by introducing one parameter α_j for each possible time of death t_j. Interpretation of the parameters $\boldsymbol{\beta}$ associated with the other covariates follows along the same lines as in logistic regression.

In fact, the analogy with logistic regression goes further: we can fit the discrete-time proportional-hazards model by running a logistic regression on a set of pseudo observations generated as follows. Suppose individual i dies or is censored at time point $t_{j(i)}$. We generate death indicators d_{ij} that take the value one if individual i died at time j and zero otherwise, generating one for each discrete time from t_1 to $t_{j(i)}$. To each of these indicators we associate a copy of the covariate vector $\mathbf{x}_i$ and a label j identifying the time point. The proportional hazards model 7.19 can then be fit by treating

the d_{ij} as independent Bernoulli observations with probability given by the hazard λ_{ij} for individual i at time point t_j.

More generally, we can group pseudo-observations with identical covariate values. Let d_{ij} denote the number of deaths and n_{ij} the total number of individuals with covariate values $\mathbf{x}_i$ observed at time point t_j. Then we can treat d_{ij} as binomial with parameters n_{ij} and λ_{ij}, where the latter satisfies the proportional hazards model.

The proof of this result runs along the same lines as the proof of the equivalence of the Poisson likelihood and the likelihood for piece-wise exponential survival data under non-informative censoring in Section 7.4.3, and relies on Equation 7.18, which writes the probability of surviving to time t_j as a product of the conditional hazards at all previous times. It is important to note that we do not assume that the pseudo-observations are independent and have a Bernoulli or binomial distribution. Rather, we note that the likelihood function for the discrete-time survival model under non-informative censoring coincides with the binomial likelihood that would be obtained by treating the death indicators as independent Bernoulli or binomial.

Time-varying covariates and time-dependent effects can be introduced in this model along the same lines as before. In the case of time-varying covariates, note that only the values of the covariates at the discrete times $t_1 < t_2 < \ldots$ are relevant. Time-dependent effects are introduced as interactions between the covariates and the discrete factor (or set of dummy variables) representing time.

7.6.3 Discrete Survival and the C-Log-Log Link

An alternative extension of the proportional hazards model to discrete time starts from the survival function, which in a proportional hazards framework can be written as

$$S(t_j|\mathbf{x}_i) = S_0(t_j)^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}},$$

where $S(t_j|\mathbf{x}_i)$ is the probability that an individual with covariate values $\mathbf{x}_i$ will survive up to time point t_j, and $S_0(t_j)$ is the baseline survival function. Recalling Equation 7.18 for the discrete survival function, we obtain a similar relationship for the complement of the hazard function, namely

$$1 - \lambda(t_j|\mathbf{x}_i) = [1 - \lambda_0(t_j)]^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}},$$

so that solving for the hazard for individual i at time point t_j we obtain the model

$$\lambda(t_j|\mathbf{x}_i) = 1 - [1 - \lambda_0(t_j)]^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}.$$

The transformation that makes the right hand side a linear function of the parameters is the complementary log-log. Applying this transformation we obtain the model

$$\log(-\log(1-\lambda(t_j|\mathbf{x}_i))) = \alpha_j + \mathbf{x}_i'\boldsymbol{\beta}, \tag{7.20}$$

where $\alpha_j = \log(-\log(1-\lambda_0(t_j)))$ is the complementary log-log transformation of the baseline hazard.

This model can be fitted to discrete survival data by generating pseudo-observations as before and fitting a generalized linear model with binomial error structure and complementary log-log link. In other words, the equivalence between the binomial likelihood and the discrete-time survival likelihood under non-informative censoring holds both for the logit and complementary log-log links.

It is interesting to note that this model can be obtained by grouping time in the continuous-time proportional-hazards model. To see this point let us assume that time is continuous and we are really interested in the standard proportional hazards model

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}.$$

Suppose, however, that time is grouped into intervals with boundaries $0 = \tau_0 < \tau_1 < \ldots < \tau_J = \infty$, and that all we observe is whether an individual survives or dies in an interval. Note that this construction imposes some constraints on censoring. If an individual is censored at some point inside an interval, we do not know whether it would have survived the interval or not. Therefore we must censor it at the end of the previous interval, which is the last point for which we have complete information. Unlike the piecewise exponential set-up, here we can not use information about exposure to part of an interval. On the other hand, it turns out that we do not need to assume that the hazard is constant in each interval.

Let λ_{ij} denote the discrete hazard or conditional probability that individual i will die in interval j given that it was alive at the start of the interval. This probability is the same as the complement of the conditional probability of surviving the interval given that one was alive at the start, and can be written as

$$\begin{aligned}
\lambda_{ij} &= 1 - \Pr\{T_i > \tau_j | T_i > \tau_{j-1}\} \\
&= 1 - \exp\{-\int_{\tau_{j-1}}^{\tau_j} \lambda(t|\mathbf{x}_i)dt\} \\
&= 1 - \exp\{-\int_{\tau_{j-1}}^{\tau_j} \lambda_0(t)dt\}^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}
\end{aligned}$$

$$= \quad 1-(1-\lambda_j)^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}},$$

where λ_j is the baseline probability of dying in interval j given survival to the start of the interval. The second line follows from Equation 7.4 relating the survival function to the integrated hazard, the third line follows from the proportional hazards assumption, and the last line defines λ_j.

As noted by Kalbfleish and Prentice (1980, p. 37), "this discrete model is then the uniquely appropriate one for grouped data from the continuous proportional hazards model". In practice, however, the model with a logit link is used much more often than the model with a c-log-log link, probably because logistic regression is better known that generalized linear models with c-log-log links, and because software for the former is more widely available than for the latter. In fact, the logit model is often used in cases where the piece-wise exponential model would be more appropriate, probably because logistic regression is better known than Poisson regression.

In closing, it may be useful to provide some suggestions regarding the choice of approach to survival analysis using generalized linear models:

- If time is truly discrete, then one should probably use the discrete model with a logit link, which has a direct interpretation in terms of conditional odds, and is easily implemented using standard software for logistic regression.

- If time is continuous but one only observes it in grouped form, then the complementary log-log link would seem more appropriate. In particular, results based on the c-log-log link should be more robust to the choice of categories than results based on the logit link. However, one cannot take into account partial exposure in a discrete time context, no matter which link is used.

- If time is continuous and one is willing to assume that the hazard is constant in each interval, then the piecewise exponential approach based on the Poisson likelihood is preferable. This approach is reasonably robust to the choice of categories and is unique in allowing the use of information from cases that have partial exposure.

Finally, if time is truly continuous and one wishes to estimate the effects of the covariates without making any assumptions about the baseline hazard, then Cox's (1972) partial likelihood is a very attractive approach.

Parametric Survival Models

Germán Rodríguez
grodri@princeton.edu

Spring, 2001; revised Spring 2005, Summer 2010

We consider briefly the analysis of survival data when one is willing to assume a parametric form for the distribution of survival time.

1 Survival Distributions

1.1 Notation

Let T denote a continuous non-negative random variable representing survival time, with probability density function (pdf) $f(t)$ and cumulative distribution function (cdf) $F(t) = \Pr\{T \leq t\}$. We focus on the *survival function* $S(t) = \Pr\{T > t\}$, the probability of being alive at t, and the hazard function $\lambda(t) = f(t)/S(t)$. Let $\Lambda(t) = \int_0^t \lambda(u)du$ denote the cumulative (or integrated) hazard and recall that

$$S(t) = \exp\{-\Lambda(t)\}.$$

Any distribution defined for $t \in [0, \infty)$ can serve as a survival distribution. We can also draft into service distributions defined for $y \in (-\infty, \infty)$ by considering $t = \exp\{y\}$, so that $y = \log t$. More generally, we can start from a r.v. W with a standard distribution in $(-\infty, \infty)$ and generate a family of survival distributions by introducing location and scale changes of the form

$$logT = Y = \alpha + \sigma W.$$

We now review some of the most important distributions.

1.2 Exponential

The exponential distribution has constant hazard $\lambda(t) = \lambda$. Thus, the survivor function is $S(t) = \exp\{-\lambda t\}$ and the density is $f(t) = \lambda \exp\{-\lambda t\}$. It

can be shown that $E(T) = 1/\lambda$ and $\text{var}(T) = 1/\lambda^2$. Thus, the coefficient of variation is 1.

The exponential distribution is related to the extreme-value distribution. Specifically, T has an exponential distribution with parameter λ, denoted $T \sim E(\lambda)$, iff

$$Y = \log T = \alpha + W$$

where $\alpha = -\log \lambda$ and W has a standard extreme value (min) distribution, with density

$$f_W(w) = e^{w - e^w}.$$

This is a unimodal density with $E(W) = -\gamma$, where $\gamma = 0.5722$ is Euler's constant, and $\text{var}(W) = \pi^2/6$. The skewness is -1.14.

The proof follows immediately from a change of variables.

1.3 Weibull

T is Weibull with parameters λ and p, denoted $T \sim W(\lambda, p)$, if $T^p \sim E(\lambda)$. The cumulative hazard is $\Lambda(t) = (\lambda t)^p$, the survivor function is $S(t) = \exp\{-(\lambda t)^p\}$, and the hazard is

$$\lambda(t) = \lambda^p p t^{p-1}.$$

The log of the Weibull hazard is a linear function of log time with constant $p \log \lambda + \log p$ and slope $p - 1$. Thus, the hazard is rising if $p > 1$, constant if $p = 1$, and declining if $p < 1$.

The Weibull is also related to the extreme-value distribution: $T \sim W(\lambda, p)$ iff

$$Y = \log T = \alpha + \sigma W,$$

where W has the extreme value distribution, $\alpha = -\log \lambda$ and $p = 1/\sigma$.

The proof follows again from a change of variables; start from W and change variables to $Y = \alpha + \sigma W$, and then change to $T = e^Y$.

1.4 Gompertz-Makeham

The Gompertz distribution is characterized by the fact that the log of the hazard is linear in t, so

$$\lambda(t) = \exp\{\alpha + \beta t\}$$

and is thus closely related to the Weibull distribution where the log of the hazard is linear in $\log t$. In fact, the Gompertz *is* a log-Weibull distribution.

This distribution provides a remarkably close fit to adult mortality in contemporary developed countries.

1.5 Gamma

The gamma distribution with parameters λ and k, denoted $\Gamma(\lambda, k)$, has density

$$f(t) = \frac{\lambda(\lambda t)^{k-1}e^{-\lambda t}}{\Gamma(k)},$$

and survivor function

$$S(t) = 1 - I_k(\lambda t),$$

where $I_k(x)$ is the incomplete gamma function, defined as

$$I_k(x) = \int_0^x \lambda^{k-1}e^{-x}dx/\Gamma(k).$$

There is no closed-form expression for the survival function, but there are excellent algorithms for its computation. (R has a function called `pgamma` that computes the cdf and survivor function. This function calls k the shape parameter and $1/\lambda$ the scale parameter.)

There is no explicit formula for the hazard either, but this may be computed easily as the ratio of the density to the survivor function, $\lambda(t) = f(t)/S(t)$. The gamma hazard

- increases monotonically if $k > 1$, from a value of 0 at the origin to a maximum of λ,
- is constant if $k = 1$
- decreases monotonically if $k < 1$, from ∞ at the origin to an asymptotic value of λ.

If $k = 1$ the gamma reduces to the exponential distribution, which can be described as the waiting time to one hit in a Poisson process. If k is an integer $k > 1$ then the gamma distribution is called the Erlang distribution and can be characterized as the waiting time to k hits in a Poisson process. The distribution exists for non-integer k as well.

The gamma distribution can also be characterized in terms of the distribution of log-time. By a simple change of variables one can show that $T \sim \Gamma(\lambda, k)$ iff

$$\log T = Y = \alpha + W,$$

where W has a *generalized* extreme-value distribution with density

$$f_w(w) = \frac{e^{kw-e^w}}{\Gamma(k)},$$

controlled by a parameter k. This density reduces to the ordinary extreme value distribution when $k = 1$.

1.6 Generalized Gamma

Stacy has proposed a generalized gamma distribution that fits neatly in the scheme we are developing, as it simply adds a scale parameter in the expression for $\log T$, so that

$$Y = \log T = \alpha + \sigma W,$$

where W has a generalized extreme value distribution with parameter k. The density of the generalized gamma distribution can be written as

$$f(t) = \frac{\lambda p (\lambda t)^{pk-1} e^{-(\lambda t)^p}}{\Gamma(k)},$$

where $p = 1/\sigma$.

The generalized gamma includes the following interesting special cases:

- gamma, when $p = 1$,
- Weibull, when $k = 1$,
- exponential, when $p = 1$ and $k = 1$.

It also includes the log-normal as a special limiting case when $k \to \infty$.

1.7 Log-Normal

T has a lognormal distribution iff

$$Y = \log T = \alpha + \sigma W,$$

where W has a standard normal distribution.

The hazard function of the log-normal distribution increases from 0 to reach a maximum and then decreases monotonically, approaching 0 as $t \to \infty$.

As $k \to \infty$ the generalized extreme value distribution approaches a standard normal, and thus the generalized gamma approaches a log-normal.

1.8 Log-Logistic

T has a log-logistic distribution iff

$$Y = \log T = \alpha + \sigma W,$$

where W has a standard logistic distribution, with pdf

$$f_W(w) = \frac{e^w}{(1+e^w)^2},$$

and cdf

$$F_W(w) = \frac{e^w}{1+e^w}.$$

The survivor function is the complement

$$S_W(w) = \frac{1}{1+e^w}.$$

Changing variables to T we find that the log-logistic survivor function is

$$S(t) = \frac{1}{1+(\lambda t)^p},$$

where we have written, as usual, $\alpha = -\log\lambda$ and $p = 1/\sigma$. Taking logs we obtain the (negative) integrated hazard, and differentiating w.r.t. t we find the hazard function

$$\lambda(t) = \frac{\lambda p(\lambda t)^{p-1}}{1+(\lambda t)^p}.$$

Note that the *logit* of the survival function $S(t)$ is linear in $\log t$. This fact provides a diagnostic plot: if you have a non-parametric estimate of the survivor function you can plot its logit against log-time; if the graph looks like a straight line then the survivor function is log-logistic.

The hazard itself is

- monotone decreasing from ∞ if $p < 1$,
- monotone decreasing from λ if $p = 1$, and
- similar to the log-normal if $p > 1$.

1.9 Generalized F

Kalbfleisch and Prentice (1980) consider the more general case where

$$Y = \log T = \alpha + \sigma W$$

and W is distributed as the log of an F-variate (which adds two more parameters).

The interesting thing about this distribution is that it includes *all* of the above distributions as special or limiting cases, and is therefore useful for testing different parametric forms.

1.10 The Coale-McNeil Model

The Coale-McNeil model of first marriage frequencies among women who will eventually marry is closely related to the extreme value and gamma distributions.

The model assumes that the density of first marriages at age a among women who will eventually marry is given by

$$g(a) = g_0\left(\frac{a-a_0}{k}\right)\frac{1}{k},$$

where a_0 and k are location and scale parameters and $g_0(.)$ is a standard schedule based on Swedish data. This standard schedule was first derived empirically, but later Coale and McNeil showed that it could be closely approximated by the following analytic expression:

$$g_0(z) = 1.946e^{-0.174(z-6.06)-e^{-0.288(z-6.06)}}.$$

It will be convenient to write a somewhat more general model with three parameters:

$$g(x) = \frac{\lambda}{\gamma(\alpha/\gamma)}e^{-\alpha(x-\theta)-e^{-\lambda(x-\theta)}}.$$

This is a form of extreme value distribution. In fact, if $\alpha = \lambda$ it reduces to the standard extreme value distribution that we discussed before. This more general case is known as a (reversed) generalized extreme value.

The mean of this distribution is

$$\mu = \theta - \frac{1}{\lambda}\psi(\alpha/\lambda),$$

where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the *digamma* function (or derivative of the log of the gamma function).

The Swedish standard derived by Coale and McNeil corresponds to the case

$$\alpha = 0.174, \qquad \lambda = 0.288, \qquad \text{and} \qquad \theta = 6.06,$$

which gives a mean of $\mu = 11.36$.

By a simple change of variables, it can be seen that the more general case with parameters a_0 and k corresponds to

$$\alpha^* = \frac{0.174}{k}, \qquad \lambda^* = \frac{0.288}{k}, \qquad \text{and} \qquad \theta^* = a_0 + 6.06k.$$

Thus, X has the (more general) Coale-McNeil distribution with parameters α, λ and θ iff

$$X = \theta - \frac{1}{\lambda} \log Y,$$

where Y has a gamma distribution with shape parameter $p = \alpha/\lambda$.

In other words, age at marriage is distributed as a linear function of the logarithm of a gamma random variable.

In particular, the Swedish standard can be obtained as

$$X = 6.06 - \frac{1}{0.288} \log Y,$$

where Y is gamma with $p = \alpha/\lambda = 0.174/0.288 = 0.604$.

The case with parameters a_0 and k can be obtained as

$$X = a_0 + 6.06k - \frac{k}{0.288} \log Y,$$

where Y is again gamma with $p = 0.604$.

The Coale-McNeil models holds the ratio $p = \alpha/\lambda$ fixed at 0.604, but along the way we have generalized the model and could entertain the notion of estimating p rather than holding it fixed.

The main significance of these results is computational:

- we can calculate marriage schedules as long as we have a function to compute the incomplete gamma function (or even chi-squared)
- we can fit nuptiality models using software for fitting gamma models.

For further details see my 1980 paper with Trussell. In that paper we used the mean and standard deviation as the parameters of interest, instead of a_0 and k. I have also written a set of R/S functions to compute marriage schedules, and these are documented separately.

2 Models With Covariates

There are four approaches to modelling survival data with covariates:

- Parametric Families
- Accelerated Life
- Proportional Hazards
- Proportional Odds

We describe each in turn.

2.1 Parametric Families

A general approach is to pick one of the parametric distributions that we have discussed and let the *parameters* of that distribution depend on covariates. For example,

- In an exponential distribution we could let the parameter λ depend on a vector of covariates x, for example using a log-linear model where

$$\log \lambda = x'\beta$$

- In a Weibull distribution we could use a similar model for λ while holding p fixed, or we could let p depend on covariates as well, for example as

$$\log p = x'\gamma$$

- In the Coale-McNeil model using the Rodríguez-Trussell parametrization, one could use a linear model for the mean

$$\mu = x'\beta$$

  while holding the standard deviation σ constant (as usually done in linear models). Alternatively, we could let the dispersion depend on covariates as well, using

$$\log \sigma = x'\gamma,$$

  with parameters γ. In the most general case, we could let the proportion that eventually marries depend on yet another set of parameters.

In general, with k groups one could give each group its own distribution in a family. This is a workable approach, but it is not exactly parsimonious and doesn't lend itself to easy interpretations.

2.2 Accelerated Life Models

Consider an ordinary regression model for log survival time, of the form

$$Y = \log T = -x'\beta + \sigma W,$$

where the error term W has a suitable distribution, e.g. extreme value, generalized extreme value, normal or logistic. This leads to Weibull, generalized gamma, log-normal or log-logistic models for T.

For example if W is extreme value then T has a Weibull distribution with

$$\log \lambda = x'\beta \qquad \text{and} \qquad p = \frac{1}{\sigma}.$$

Note that λ depends on the covariates but p is assumed the same for everyone.

This model has an *accelerated life* interpretation. In this formulation we view the error term σW as a standard or reference distribution that applies when $x = 0$. It will be convenient to translate the reference distribution to the time scale by defining $T_0 = \exp\{\sigma W\}$. The probability that a reference subject will be alive at time t, which will be denoted $S_0(t)$, is

$$S_0(t) = \Pr\{T_0 > t\} = \Pr\{W > \log t/\sigma\}.$$

Consider now the effect of the covariates x. In this model T is distributed as $T_0 e^{-x'\beta}$, so the covariates act multiplicatively on survival time. What is the probability that a subject with covariate values x will be alive at time t?

$$S(t,x) = \Pr\{T > t|x\} = \Pr\{T_0 e^{-x'\beta} > t\} = \Pr\{T_0 > te^{x'\beta}\} = S_0(te^{x'\beta}).$$

In words, the probability that a subject with covariates x will be alive at time t is the same as the probability that a reference subject will be alive at time $t\exp\{x'\beta\}$. This may be interpreted as time passing more rapidly (or people aging more quickly) by a factor $\exp\{x'\beta\}$, for example twice as fast or half as fast. (The analogy to 'dog years' should no go unnoticed.)

We can also write the density and hazard functions for any subject in terms of the baseline or reference density and hazard:

$$f(t) = f_0(te^{x'\beta})e^{x'\beta},$$

and

$$\lambda(t) = \lambda_0(te^{x'\beta})e^{x'\beta}.$$

We see that a simple relationship between the survivor functions for different x's (just a stretching of the time axis), translates into a more complex relationship when viewed in terms of the pdf or the hazard function.

Consider for example a multiplier of two for a subject with covariates x. In terms of survival, this means that the probability that the subject would be alive at any given age is the same as the probability that a reference subject would be alive at twice the age. In terms of risk, it means that our

subject is exposed at any given age to double the risk of a reference subject twice as old.

Note also that if we start with a given distribution and stretch the time axis we may well end up with a distribution in a completely different family. Stretching a Weibull produces another Weibull, but not all families are closed under acceleration of time. (Is the Coale-McNeil an accelerated life model?)

2.3 Proportional Hazards

An alternative approach to modelling survival data is to assume that the effect of the covariates is to increase or decrease the *hazard* by a proportionate amount at all durations. Thus

$$\lambda(t,x) = \lambda_0(t)e^{x'\beta},$$

where $\lambda_0(t)$ is the *baseline hazard*, or the hazard for a reference individual with covariate values 0, and $\exp\{x'\beta\}$ is the *relative risk* associated with covariate values x.

Obviously the cumulative hazards would follow the same relationship, as can be seen by integrating both sides of the previous equation. Exponentiating minus the integrated hazard we find the survivor functions to be

$$S(t,x) = S_0(t)^{e^{x'\beta}},$$

so the survivor function for covariates x is the baseline survivor raised to a power. If a subject is exposed to twice the risk of a reference subject at every age, then the probability that the subject will be alive at any given age is the square of the probability that the reference subject would be alive at the same age. In this model a simple relationship in terms of hazards translates into a more complex relationship in terms of survival functions.

These equations define a *family* of models. Picking a different parametric form for the baseline hazard leads to a different model in the proportional hazards family. Suppose we start with a Weibull baseline hazard, so

$$\lambda_0(t) = \lambda p(\lambda t)^{p-1},$$

and we then multiply this by a relative risk $e^{x'\beta}$. You should be able to show that the resulting hazard is again Weibull,

$$\lambda(t,x) = \lambda^* p(\lambda^* t)^{p-1},$$

with the same p as before but $\lambda^* = \lambda e^{x'\beta/p}$.

Thus, the Weibull family is closed under proportionality of hazards, but this is not true for other distributions. If T_0 is log-logistic, for example, and we multiply the *hazard* by a relative risk $e^{x'\beta}$, the resulting distribution is not log-logistic.

2.4 Proportional Hazards and Accelerated Life

Do the proportional hazard and accelerated life models ever coincide? More precisely, if we start with a hazard and multiply by a relative risk, and someone else starts with another hazard and stretches time, do we ever end up with the same distribution? The condition just formulated may be stated as

$$\lambda_0(t)e^{x'\beta} = \lambda_0^*(te^{x'\beta^*})e^{x'\beta^*}$$

for all x and t. The stars indicate that we do not necessarily start with the same baseline hazard or end up with the same parameters reflecting the effects of the covariates.

If this condition is to be true for all x then it must be true for $x = 0$, implying

$$\lambda_0(t) = \lambda_0^*(t),$$

so the baseline hazards must be the same. Let us try to find this hazard.

The trick here is to consider a very special value of the vector of covariates x, where we set the first element to $-\log t/\beta_1^*$ and the others to zero, so

$$x = (-\log t/\beta_1^*, 0, \ldots, 0).$$

Multiplying by β^* we find that $x'\beta^* = -\log t$, while multiplying by β gives $x'\beta = -\log t\beta_1/\beta_1^*$. Using these results on the condition we obtain

$$\lambda_0(t)e^{-\log t\beta_1/\beta_1^*} = \lambda_0(te^{-\log t})e^{-\log t}.$$

Using the fact that $e^{b\log a} = a^b$ we can simplify this expression to

$$\lambda_0(t)\left(\frac{1}{t}\right)^{\beta_1/\beta_1^*} = \lambda_0(1)\frac{1}{t},$$

or, moving the term on $1/t$ to the right-hand side,

$$\lambda_0(t) = \lambda_0(1)t^{\beta_1/\beta_1^*-1}.$$

Repeat this exercise with a covariate vector x that has $\log t/\beta_i^*$ in the i-th slot and 0 everywhere else, so that $x'\beta^* = -\log t$ and $x'\beta = -\log t\beta_i/\beta_i^*$.

We find the same result but with $\beta_i/\beta_i^* - 1$ as the exponent of t. If the condition is to be true for all x, then the ratio of the coefficients must be constant,

$$\frac{\beta_i}{\beta_i^*} = \frac{\beta}{\beta^*} = p,$$

say. This leads to the solution

$$\lambda_0(t) = \lambda_0(1)t^{p-1},$$

which can be recognized as a Weibull hazard. To see this last point write the result in the more familiar form

$$\lambda_0(t) = \lambda^p p t^{p-1} = \lambda p(\lambda t)^{p-1},$$

where I have taken the constant to be $\lambda_0(1) = \lambda^p p$, which is the same as defining $\lambda = (\lambda_0(1)/p)^{1/p}$.

This result shows that the Weibull is the *only* distribution that is closed under both the accelerated life and proportional hazards families.

Note that the accelerated life and proportional hazards parameters β^* and β are proportional to each other, with proportionality constant p. In particular, they are equal for $p = 1$.

Thus, doubling the risk in an exponential model makes time go twice as fast. But doubling the risk in a Weibull model with $p = 2$ makes time go only about 40% faster. Can you see why?

2.5 Proportional Odds

An alternative approach to survival modelling is to assume that the effect of the covariates is to increase or decrease the *odds* of dying by a given duration by a proportionate amount:

$$\frac{1 - S(t,x)}{S(t,x)} = \frac{1 - S_0(t)}{S_0(t)} e^{x'\beta},$$

where $S_0(t)$ is a baseline survivor function, taken from a suitable distribution, and $\exp\{x'\beta\}$ is a multiplier reflecting the proportionate increase in the odds associated with covariate values x.

Taking logs, we find that

$$\text{logit}(1 - S(t,x)) = \text{logit}(1 - S_0(t)) + x'\beta,$$

so the covariate effects are linear in the logit scale.

A somewhat more general version of the proportional odds model (but without covariates) is known as the *relational logit model* in demography. The idea is to allow the log-odds of dying in a given population to be a linear function of the log-odds in a reference or baseline population, so that

$$\text{logit}(1 - S(t)) = \alpha + \theta \text{logit}(1 - S_0(t)).$$

These models were popularized by Brass. The proportional odds model is the special case where $\theta = 1$ (but we let the constant α depend on covariates).

These models could be defined in terms of the odds of *surviving* to duration t, but this merely changes the sign of β. I prefer the definition in terms of the odds of *dying* because it preserves the interpretation of the β coefficients as increasing the risk, which is consistent with hazard models. This is also the reason why I used a minus sign when defining the coefficients for accelerated life models.

As an example consider a proportional odds model with a log-logistic baseline. The corresponding survival function, its complement, and the odds of dying are

$$S_0(t) = \frac{1}{1 + (\lambda t)^p}, \qquad 1 - S_0(t) = \frac{(\lambda t)^p}{1 + (\lambda t)^p}, \qquad \frac{1 - S_0(t)}{S_0(t)} = (\lambda t)^p.$$

Multiplying the odds by $\exp\{x'\beta\}$ yields another log-logistic model, this time with $\lambda^* = \lambda e^{x'\beta/p}$ and $p^* = p$. Thus, the log-logistic family is closed under proportionality of odds.

This is not true of other distributions. For example if we start with a Weibull baseline and multiply the odds of dying by a constant, the resulting distribution is not Weibull.

2.6 Proportional Odds And Accelerated Life

Do the proportional odds and accelerated life models ever coincide? The answer is yes, when (and only when) the baseline is log-logistic.

The proof follows essentially the same steps as the proof for the intersection of the proportional hazards and accelerated life models.

3 Maximum Likelihood Estimation

All parametric models may be fit by maximizing the appropriate likelihood function.

The data consist of pairs $\{t_i, d_i\}$ where

- t_i is the survival or censoring time, and
- d_i is a death indicator, taking the value 1 for deaths and 0 for censored cases

The likelihood function under general non-informative censoring has the form

$$L(\theta) = \prod_{i=1}^{n} \lambda(t_i|x_i)^{d_i} S(t_i|x_i),$$

and in general must be maximized numerically using a procedure such as Newton-Raphson.

Kalbfleisch and Prentice have a nice discussion of the procedures that need to be followed in fitting parametric models, including first and second derivatives for accelerated life models using the parametric distributions discussed here.

Stata's `streg` can fit a number of parametric models, including exponential, Weibull and Gompertz in the proportional hazards framework, and log-normal, log-logistic, and generalized gamma (as well as exponential and Weibull) in the accelerated failure-time framework. Now you know why the Weibull is included in both the PH and AFT metrics.

Non-Parametric Estimation in Survival Models

Germán Rodríguez
grodri@princeton.edu

Spring, 2001; revised Spring 2005

We now discuss the analysis of survival data without parametric assumptions about the form of the distribution.

1 One Sample: Kaplan-Meier

Our first topic is non-parametric estimation of the *survival function*. If the data were not censored, the obvious estimate would be the empirical survival function

$$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I\{t_i > t\},$$

where I is the indicator function that takes the value 1 if the condition in braces is true and 0 otherwise. The estimator is simply the proportion alive at t.

1.1 Estimation with Censored Data

Kaplan and Meier (1958) extended the estimate to *censored* data. Let

$$t_{(1)} < t_{(2)} < \ldots < t_{(m)}$$

denote the distinct *ordered* times of death (not counting censoring times). Let d_i be the number of deaths at $t_{(i)}$, and let n_i be the number alive *just before* $t_{(i)}$. This is the number exposed to risk at time $t_{(i)}$. Then the Kaplan-Meier or *product limit* estimate of the survivor function is

$$\hat{S}(t) = \prod_{i:t_{(i)}<t} \left(1 - \frac{d_i}{n_i}\right).$$

A heuristic justification of the estimate is as follows. To survive to time t you must first survive to $t_{(1)}$. You must then survive from $t_{(1)}$ to $t_{(2)}$ given

that you have already survived to $t_{(1)}$. And so on. Because there are no deaths between $t_{(i-1)}$ and $t_{(i)}$, we take the probability of dying between these times to be zero. The conditional probability of dying at $t_{(i)}$ given that the subject was alive just before can be estimated by d_i/n_i. The conditional probability of surviving time $t_{(i)}$ is the complement $1 - d_i/n_i$. The overall unconditional probability of surviving to t is obtained by multiplying the conditional probabilities for all relevant times up to t.

The Kaplan-Meier estimate is a step function with discontinuities or jumps at the observed death times. Figure 1 shows Kaplan-Meier estimates for the treated and control groups in the famous Gehan data (see Cox, 1972 or Andersen et al., 1993, p. 22-23).

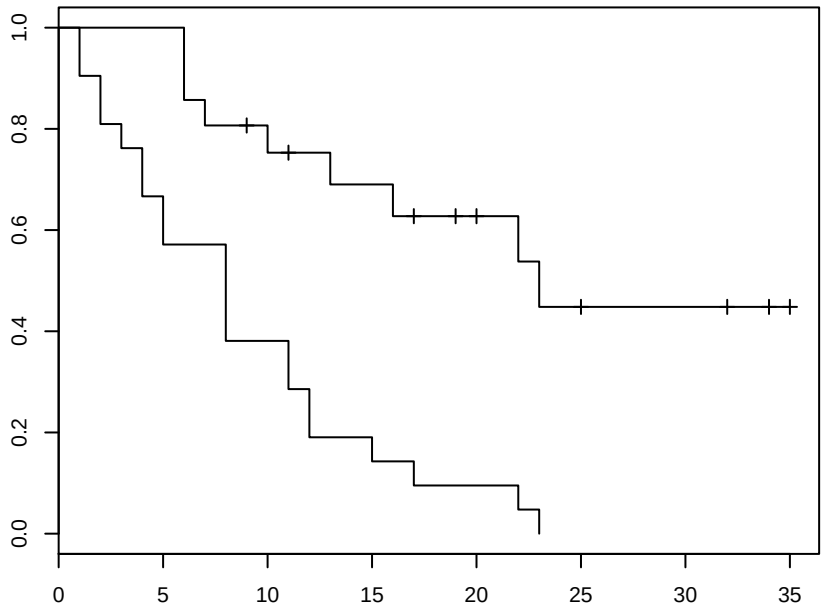


Figure 1: Kaplan-Meier Estimates for Gehan Data

If there is no censoring, the K-M estimate coincides with the empirical survival function. If the last observation happens to be a censored case, as is the case in the treated group in the Gehan data, the estimate is undefined beyond the last death.

1.2 Non-parametric Maximum Likelihood

The K-M estimator has a nice interpretation as a non-parametric maximum likelihood estimator (NPML). A rigorous treatment of this notion is beyond the scope of the course, but the original article by K-M provides a more intuitive approach. We consider the contribution to the likelihood of cases that die or are censored at time t.

- If a subject is censored at t its contribution to the likelihood is $S(t)$. In order to maximize the likelihood we would like to make this as large as possible. Because a survival function must be non-increasing, the best we can do is keep it constant at t. In other words, the estimated survival function doesn't change at censoring times.

- If a subject dies at t then this is one of the distinct times of death that we introduced before. Say it is $t_{(i)}$. We need to make the survival function just before $t_{(i)}$ as large as possible. The largest it can be is the value at the previous time of death or 1, whichever is less. We also need to make the survival at $t_{(i)}$ itself as small as possible. This means we need a discontinuity at $t_{(i)}$.

Let c_i denote the number of cases censored between $t_{(i)}$ and $t_{(i+1)}$, and let d_i be the number of cases that die at $t_{(i)}$. Then the likelihood function takes the form

$$L = \prod_{i=1}^{m} [S(t_{(i-1)}) - S(t_{(i)})]^{d_i} S(t_{(i)})^{c_i},$$

where the product is over the m distinct times of death, and we take $t_{(0)} = 0$ with $S(t_{(0)}) = 1$. The problem now is to estimate m parameters representing the values of the survival function at the death times $t_{(1)}, t_{(2)}, \ldots, t_{(m)}$.

Write $\pi_i = S(t_{(i)})/S(t_{(i-1)})$ for the conditional probability of surviving from $S(t_{(i-1)})$ to $S(t_{(i)})$. Then we can write

$$S(t_{(i)}) = \pi_1 \pi_2 \ldots \pi_i,$$

and the likelihood becomes

$$L = \prod_{i=1}^{m} (1 - \pi_i)^{d_i} \pi_i^{c_i} (\pi_1 \pi_2 \ldots \pi_{i-1})^{d_i + c_i}.$$

Note that all cases who die at $t_{(i)}$ or are censored between $t_{(i)}$ and $t_{(i+1)}$ contribute a term π_j to each of the previous times of death from $t_{(1)}$ to $t_{(i-1)}$. In addition, those who die at $t_{(i)}$ contribute $1 - \pi_i$, and the censored

cases contribute an additional π_i. Let $n_i = \sum_{j \geq i}(d_j + c_j)$ denote the total number exposed to risk at $t_{(i)}$. We can then collect terms on each π_i and write the likelihood as

$$L = \prod_{i=1}^{m} (1 - \pi_i)^{d_i} \pi_i^{n_i - d_i},$$

a binomial likelihood. The m.l.e. of π_i is then

$$\hat{\pi}_i = \frac{n_i - d_i}{n_i} = 1 - \frac{d_i}{n_i}.$$

The K-M estimator follows from multiplying these conditional probabilities.

1.3 Greenwood's Formula

From the likelihood function obtained above it follows that the large sample variance of $\hat{\pi}_i$ conditional on the data n_i and d_i is given by the usual binomial formula

$$\text{var}(\hat{\pi}_i) = \frac{\pi_i(1 - \pi_i)}{n_i}.$$

Perhaps less obviously, $\text{cov}(\hat{\pi}_i, \hat{\pi}_j) = 0$ for $i \neq j$, so the covariances of the contributions from different times of death are all zero. You can verify this result by taking logs and then first and second derivatives of the log-likelihood function.

To obtain the large sample variance of $\hat{S}(t)$, the K-M estimate of the survival function, we need to apply the delta method twice. First we take logs, so that instead of the variance of a product we can find the variance of a sum, working with

$$K_i = \log \hat{S}(t_{(i)}) = \sum_{j=1}^{i} \log \hat{\pi}_j.$$

Now we need to find the variance of the log of $\hat{\pi}_i$. This will be our first application of the delta method. The large-sample variance of a function f of a random variable X is

$$\text{var}(f(X)) = (f'(X))^2 \text{var}(X),$$

so we just multiply the variance of X by the derivative of the transformation. In our case the function is the log and we obtain

$$\text{var}(\log \hat{\pi}_i) = (\frac{1}{\hat{\pi}_i})^2 \text{var}(\hat{\pi}_i) = \frac{1 - \pi_i}{n_i \pi_i}.$$

Because K_i is a sum and the covariances of the $\pi_i's$ (and hence of the $\log \pi_i's$) are zero, we find

$$\text{var}(\log \hat{S}(t_{(i)})) = \sum_{j=1}^{i} \frac{1-\pi_j}{n_j \pi_j} = \sum \frac{d_j}{n_j(n_j - d_j)}.$$

Now we have to use the delta method again, this time to get the variance of the survivor function from the variance of its log:

$$\text{var}(\hat{S}(t_{(i)})) = [\hat{S}(t_{(i)})]^2 \sum_{j=1}^{i} \frac{1-\hat{\pi}_j}{n_j \hat{\pi}_j}.$$

This result is known as *Greenwood's formula*. You may question the derivation because it conditions on the n_j which are random variables, but the result is in the spirit of likelihood theory, conditioning on all observed quantities, and has been justified rigourously.

Peterson (1977) has shown that the K-M estimator $\hat{S}(t)$ is consistent, and Breslow and Crowley (1974) show that $\sqrt{n}(\hat{S}(t) - S(t))$ converges in law to a Gaussian process with expectation 0 and a variance-covariance function that may be approximated using Greenwood's formula. For a modern treatment of the estimator from the point of view of counting processes see Andersen et al. (1993).

1.4 The Nelson-Aalen Estimator

Consider estimating the cumulative hazard $\Lambda(t)$. A simple approach is to start from an estimator of $S(t)$ and take minus the log. An alternative approach is to estimate the cumulative hazard directly using the Nelson-Aalen estimator:

$$\hat{\Lambda}(t_{(i)}) = \sum_{j=1}^{i} \frac{d_j}{n_j}.$$

Intuitively, this expression is estimating the hazard at each distinct time of death $t_{(j)}$ as the ratio of the number of deaths to the number exposed. The cumulative hazard up to time t is simply the sum of the hazards at all death times up to t, and has a nice interpretation as the expected number of deaths in $(0, t]$ per unit at risk. This estimator has a strong justification in terms of the theory of counting processes.

The variance of $\hat{\Lambda}(t_{(i)})$ can be approximated by $\text{var}(-\log \hat{S}(t_{(i)}))$, which we obtained on our way to Greenwood's formula. Therneau and Grambsch (2000) discuss alternative approximations.

Breslow (1972) suggested estimating the survival function as

$$\hat{S}(t) = \exp\{-\hat{\Lambda}(t)\},$$

where $\hat{\Lambda}(t)$ is the Nelson-Aalen estimator of the integrated hazard. The Breslow estimator and the K-M estimator are asymptotically equivalent, and usually are quite close to each other, particularly when the number of deaths is small relative to the number exposed.

1.5 Expectation of Life

If $\hat{S}(t_{(m)}) = 0$ then one can estimate $\mu = E(T)$ as the integral of the K-M estimate:

$$\hat{\mu} = \int_0^\infty \hat{S}(t)dt = \sum_{i=1}^{m}(t_{(i)} - t_{(i-1)})\hat{S}(t_{(i)}).$$

Can you figure out the variance of $\hat{\mu}$?

2 k-Samples: Mantel-Haenszel

Consider now the problem of comparing two or more survivor functions, for example urban versus rural, or treated versus control. Let

$$t_{(1)} < t_{(2)} < \ldots < t_{(m)}$$

denote the distinct times of death observed in the *total* sample, obtained by combining all groups of interest. Let

$$\begin{aligned} d_{ij} &= \text{deaths at time } t_{(i)} \text{ in group } j, \text{ and} \\ n_{ij} &= \text{number at risk at time } t_{(i)} \text{ in group } j. \end{aligned}$$

We also let d_i and n_i denote the total number of deaths and subjects at risk at time $t_{(i)}$.

If the survival probabilities are the same in all groups, then the d_i deaths at time $t_{(i)}$ should be distributed among the k groups in proportion to the number at risk. Thus, conditional on d_i and n_{ij},

$$E(d_{ij}) = d_i\frac{n_{ij}}{n_i} = n_{ij}\frac{d_i}{n_i},$$

where the last term shows that we can also view this calculation as applying an overall failure rate d_i/n_i to the n_{ij} subjects in group j.

We now proceed beyond the mean to obtain the distribution of these counts. Imagine setting up a contingency table at each distinct failure time, with rows given by survival status and columns given by group membership. The entries in the table are d_{ij} or $n_{ij} - d_{ij}$, the row totals are d_i and n_i and the column totals are n_{ij}. The distribution of the counts conditional on both the row and column totals is *hypergeometric*. (We mentioned this distribution briefly in WWS509 when we considered contingency tables with both margins fixed.) The hypergeometric distribution has mean as given above, variance

$$\text{var}(d_{ij}) = \frac{d_i(n_i - d_i)}{n_i - 1}\frac{n_{ij}}{n_i}(1 - \frac{n_{ij}}{n_i}),$$

and covariance

$$\text{cov}(d_{ir}, d_{is}) = -\frac{d_i(n_i - d_i)}{n_i - 1}\frac{n_{ir}n_{is}}{n_i^2}.$$

Let $\vec{d_i}$ denote the vector of deaths at time $t_{(i)}$, with mean $E(\vec{d_i})$ and var-cov matrix $\text{var}(\vec{d_i})$. We sum these over all times to obtain

$$D = \sum_{i=1}^{m}[\vec{d_i} - E(\vec{d_i})] \qquad \text{and} \qquad V = \sum_{i=1}^{m}\text{var}(\vec{d_i}).$$

Mantel and Haenszel proposed testing the equality of the k survival functions

$$H_0 : S_1(t) = S_2(t) = \ldots = S_k(t)$$

by treating the quadratic form

$$Q = D'V^{-}D$$

as a χ^2 statistic with $k - 1$ degrees of freedom. Here V^{-} is any generalized inverse of V. Omitting the i-th group from the calculation of D and V will do; the test is invariant to the choice of omitted group. For $k = 2$ we get

$$z = \sqrt{Q} = \frac{\sum(d_{i1} - E(d_{i1}))}{\sqrt{\sum \text{var}(d_{i1})}}.$$

An approximation for $k \geq 2$ which does not require matrix inversion treats

$$\sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where O_{ij} denotes observed and E_{ij} expected deaths at time $t_{(i)}$ in group j, as a χ^2 statistic with $k - 1$ d.f.

The Mantel-Haenszel test can be derived as a linear rank test, and in that context is often called the *log-rank* or Savage test. Kalbfleisch and Prentice have proposed an extension to censored data of the Wilcoxon test. Other alternatives have been proposed by Gehan (1965) and Breslow (1970), but the M-H test is the most popular one.

3 Regression: Cox's Model

Let us consider the more general problem where we have a vector x of covariates. The k-sample problem can be viewed as the special case where the x's are dummy variables denoting group membership. Recall the basic model

$$\lambda(t, x) = \lambda_0(t)e^{x'\beta},$$

and consider estimation of β without making any assumptions about the baseline hazard $\lambda_0(t)$.

3.1 Cox's Partial Likelihood

In his 1972 paper Cox proposed fitting the model by maximizing a special likelihood. Let

$$t_{(1)} < t_{(2)} < \ldots < t_{(m)}$$

denote the observed distinct times of death, as before, and consider what happens at $t_{(i)}$. Let R_i denote the risk set at $t_{(i)}$, defined as the set of indices of the subjects that are alive just before $t_{(i)}$. Thus, $R_0 = \{1, 2, \ldots, n\}$.

Suppose first that there are no ties in the observation times, so one and only person subject failed at $t_{(i)}$. Let's call this subject $j(i)$. What is the conditional probability that this particular subject would fail at $t_{(i)}$ given the risk set R_i and the fact that exactly one subject fails at that time? Answer:

$$\frac{\lambda(t_{(i)}, x_{j(i)})dt}{\sum_{j\in R_i}\lambda(t_{(i)}, x_j)dt}.$$

We can write this probability in terms of the baseline hazard and relative risk as

$$\frac{\lambda_0(t_{(i)})e^{x'_{j(i)}\beta}}{\sum_{j\in R_i}\lambda_0(t_{(i)})e^{x'_j\beta}},$$

and we notice that the baseline hazard cancels, so the probability in question is

$$\frac{e^{x'_{j(i)}\beta}}{\sum_{j\in R_i} e^{x'_j\beta}}$$

and does not depend on the baseline hazard $\lambda_0(t)$.

Cox proposed multiplying these probabilities together over all distinct failure times and treating the resulting product

$$L = \prod_{i=1}^{m} \frac{e^{x'_{j(i)}\beta}}{\sum_{j\in R_i} e^{x'_j\beta}}$$

as if it was an ordinary likelihood. In his original paper Cox called this a "conditional likelihood" because it is a product of conditional probabilities, but later abandoned the name because it is misleading: L is not itself a conditional probability.

Kalbfleisch and Prentice considered the case where the covariates are fixed over time and showed that L is the *marginal likelihood* of the ranks of the observations, obtained by considering just the order in which people die and not the actual times at which they die.

In 1975 Cox provided a more general justification of L as part of the full likelihood—in fact, a part that happens to contain most of the information about β—and therefore proposed calling L a *partial likelihood*. This justification is valid even with time-varying covariates.

A more rigorous justification of the partial likelihood in terms of the theory of counting processes can be found in Andersen et al. (1993).

3.2 The Score and Information

The log of Cox's partial likelihood is

$$logL = \sum_i \{x_{(j(i)}\beta - \log \sum_{j\in R_i} e^{x'_j\beta}\}.$$

Taking derivatives with respect to β we find the score to be

$$\frac{\partial \log L_i}{\partial \beta_r} = x_{j(i)r} - \frac{\sum_{j\in R_i} e^{x'_j\beta} x_{jr}}{\sum_{j\in R_i} e^{x'_j\beta}}.$$

The term to the right of the minus sign is just a weighted average of x_r over the risk set R_i with weights equal to the relative risks $e^{x'_j\beta}$. Thus, we can

write the score as

$$U_r(\beta) = \frac{\partial \log L_i}{\partial \beta_r} = x_{j(i)r} - A_{ir}(\beta),$$

where $A_{ir}(\beta)$ is the weighted average of x_r over the risk set R_i.

Taking derivatives again and changing sign we find the observed information to be

$$-\frac{\partial^2 \log L_i}{\partial \beta_r \partial \beta_s} = \frac{\sum e^{x_j'\beta} x_{jr} x_{js} (\sum e^{x_j'\beta}) - (\sum e^{x_j'\beta} x_{jr})(\sum e^{x_j'\beta} x_{js})}{(\sum e^{x_j'\beta})^2},$$

where all sums are over the risk set R_i. The right hand side can be written as the difference of two terms. The first term can be interpreted as a weighted average of the cross-product of x_r and x_s. The second term is the product of the weighted averages $A_{ir}(\beta)$ and $A_{is}(\beta)$. Thus we can write

$$-\frac{\partial^2 \log L_i}{\partial \beta_r \partial \beta_s} = \frac{\sum e^{x_j'\beta} x_{jr} x_{js}}{\sum e^{x_j'\beta}} - A_{jr}(\beta) A_{js}(\beta).$$

You may recognize this expression as the old "desk calculator" formula for a covariance, leading to the observed information

$$I(\beta) = -\frac{\partial^2 \log L_i}{\partial \beta_r \partial \beta_s} = \sum_{i=1}^{m} C_{irs}(\beta),$$

where $C_{irs}(\beta)$ denotes the weighted covariance of x_r and x_s over the risk set R_i with weights equal to the relative risks $e^{x_j'\beta}$ for $j \in R_i$. Calculation of the score and information is thus relatively simple. In matrix notation

$$u(\beta) = \sum_i (x_{j(i)} - A_i(\beta)) \qquad \text{and} \qquad I(\beta) = \sum_i C_i(\beta),$$

where $A_i(\beta)$ is the mean and $C_i(\beta)$ is the variance-covariance matrix of x over the risk set R_i with weights $e^{x_j'\beta}$ for $j \in R_i$.

Notably, the partial log-likelihood is formally identical with the log-likelihood for a conditional logit model.

3.3 The Problem of Ties

The development so far has assumed that only one death occurs at each distinct time $t_{(i)}$. In practice we often observe several deaths, say d_i, at $t_{(i)}$. This can happen in several ways:

- The data are *discrete*, so in fact there is a positive probability of failure at $t_{(i)}$. If this is the case one should really use a discrete model. We will discuss possible approaches below.

- The data are *continuous* but have been *grouped*, so d_i represents the number of deaths in some interval around $t_{(i)}$. In this case one would probably be better off estimating a separate parameter for each interval, using a complementary-log-log binomial model or a Poisson model corresponding to a piece-wise exponential survival model, as discussed in my WWS509 notes.

- The data are *continuous* and are *not grouped*, but there are a few ties resulting perhaps from coarse measurement. In this case we can extend the argument used to build the likelihood.

Let D_i denote the set of indices of the d_i cases who failed at $t_{(i)}$. The probability that the d_i cases that actually fail would be those in D_i given the risk set R_i and the fact that d_i of them fail at $t_{(i)}$ is

$$L = \frac{\prod_{j\in D_i} e^{x_j'\beta}}{\sum_{P_i}\prod_{j\in P_i} e^{x_j'\beta}},$$

where the sum in the denominator is over all possible permutations P_i or ways of choosing d_i indices from the risk set, and the product is over the set of chosen indices, which we call a permutation. For example assume four people are at risk and two die. The deaths could be {1,2}, {1,3}, {1,4}, {2,3}, {2,4}, {3,4}. We calculate the probability of each of these outcomes. Then we divide the probability of the outcome that actually occurred by the sum of the probabilities of all possible outcomes.

This likelihood was proposed by Cox in his original paper. The numerator is easy to calculate, and has the form $\exp\{S_i'\beta\}$, where $S_i = \sum_{j\in D_i} x_j$ is the sum of the x's over the death set D_i. The denominator is difficult to calculate because the number of permutations grows very quickly with d_i.

Peto (and independently Breslow) proposed approximating the denominator by calculating the sum $\sum e^{x_j'\beta}$ over the entire risk set R_i and raising it to d_i. This leads to the much simpler expression

$$L \approx \prod_{i=1}^{m} e^{S_i'\beta}\left(\sum_{j\in R_i} e^{x_j'\beta}\right)^{d_i}.$$

The Peto-Breslow approximation is reasonably good when d_i is small relative to n_i, and is popular because of its simplicity. Efron proposed a better

approximation that requires only modest additional computational effort. Consider again our example where two out of four subjects fail. Suppose the subjects that fail are 1 and 2, and let $r_j = e^{x_j'\beta}$ denote the relative risk for the j-th subject. In continuous time one must have failed before the other, we just don't know which. The contributions to the partial likelihood would be

$$\frac{r_1}{r_1+r_2+r_3+r_4}\frac{r_2}{r_2+r_3+r_4}$$

if 1 failed before 2, or

$$\frac{r_2}{r_1+r_2+r_3+r_4}\frac{r_1}{r_1+r_3+r_4}$$

if 2 was the first to fail. In both cases the numerator is $r_1 r_2$. To compute the denominator Peto and Breslow add the risks over the complete risk set both times, using $(r_1+r_2+r_3+r_4)^2$, which is obviously conservative. Efron uses the average risk of subjects 1 and 2 for the second term, so he takes the denominator to be $(r_1+r_2+r_3+r_4)(0.5r_1+0.5r_2+r_3+r_4)$. This approximation is much more accurate unless d_i is very large relative to n_i. For more details see Therneau and Grambsch (2000, Section 3.3).

3.4 Tests of Hypotheses

As usual, we have three approaches to testing hypotheses about $\hat{\beta}$:

- *Likelihood Ratio Test:* given two nested models, we treat twice the difference in partial log-likelihoods as a χ^2 statistic with degrees of freedom equal to the difference in the number of parameters.

- *Wald Test:* we use the fact that approximately in large samples $\hat{\beta}$ has a multivariate normal distribution with mean β and variance-covariance matrix $\text{var}(\hat{\beta}) = I^{-1}(\beta)$. Thus, under $H_0 : \beta = \beta_0$, the quadratic form

  $$(\hat{\beta}-\beta_0)'\text{var}^{-1}(\hat{\beta})(\hat{\beta}-\beta_0) \sim \chi^2_p,$$

  where p is the dimension of β. This test is often used for a subset of β.

- *Score Test:* we use the fact that approximately in large samples the score $u(\beta)$ has a multivariate normal distribution with mean 0 and variance-covariance matrix equal to the information matrix. Thus, under $H_0 : \beta = \beta_0$, the quadratic form

  $$u(\beta_0)'I^{-1}(\beta_0)u(\beta_0) \sim \chi^2_p.$$

  Note that this test does not require calculating the m.l.e. $\hat{\beta}$.

One reason for bringing up the score test is that in the k-sample case the score test of $H_0 : \beta = 0$ based on Cox's model happens to be the same as the Mantel-Haenszel log-rank test.

Here is an outline of the proof. Assume no ties. If $\beta = 0$ then the weights used to calculate A_i and C_i are all 1, A_{ir} happens to be the proportion of the risk set R_i that comes from the r-th sample (which is the same as the expected number of deaths in that group) and C_i is a binomial variance-covariance matrix. If there are ties, Cox's approach leads to the test discussed in Section 2. Use of Peto's approximation is equivalent to omitting the factor $(n_i - d_i)/(n_i - 1)$ from the variance-covariance matrix.

All three tests are asymptotically equivalent. The quality of the normal approximations depends on the sample size, the distribution of the cases over the covariate space, and the extent of censoring.

3.5 Time-Varying Covariates

A nice feature of the Cox model and partial likelihood is that it extends easily to the case of time-varying covariates. Note that the partial likelihood is built by considering only what happens at each failure time, so we only need to know the values of the covariates at the distinct times of death.

One use of time-varying covariates is to check the assumption of proportionality of hazards. In his original paper Cox analyzes a two-sample problem and introduces an auxiliary covariate

$$z_i = \begin{cases} 0, \text{in group 0} \\ t - c, \text{in group 1} \end{cases}$$

where c is an arbitrary constant close to the mean of t, chosen to avoid numerical instability. If the coefficient of z is 0, the assumption of proportionality of hazards is adequate. A positive value indicates that the ratio of hazards for group 1 over group 0 increases over time. A negative coefficient suggests a declining hazard ratio, a common occurrence.

Another use of time-varying covariates is to represent variables that simply change over time. In a study of contraceptive failure, for example, one may treat frequency of intercourse as a time-varying covariate. Another example of a time-varying covariate is education in an analysis of age at marriage.

Note that $x(t)$ may represent the actual value of a variable at time t, or any index based on the individual's history up to time t. In a study of the effects of breastfeeding on post-partum amenorrhea, for example, $x(t)$ could represent the total number of suckling episodes in the week preceding t.

The partial likelihood function with time-varying covariates does not have an interpretation as a marginal likelihood of the ranks. For further details see Cox and Oakes (1984, Chapter 8).

3.6 Estimating the Baseline Survival

Interest so far has focused on the regression coefficients β. We now consider how to estimate the baseline hazard $\lambda_0(t)$, which dropped out of the partial likelihood.

Kalbfleisch and Prentice (1980, Section 4.3) use an argument similar to the derivation of the Kaplan-Meier estimate, noting that the hazard should assign mass only to the discrete times of death. Let π_i denote the conditional survival probability at time $t_{(i)}$ for a baseline subject. To obtain the conditional probability for a subject with covariates x we would need to raise π_i to $e^{x'\beta}$. This leads to a likelihood of the form

$$L = \prod_{i=1}^{m} \prod_{j \in D_i} (1-\pi_i)^{e^{x_j'\beta}} \prod_{j \in R_i - D_i} \pi_i^{e^{x_j'\beta}}.$$

Meier suggested maximizing this likelihood with respect to both the π_i and β. A simpler approach is to plug-in the estimate $\hat{\beta}$ from the partial likelihood, and maximize the resulting expression with respect to the π_i only. If there are no ties, this gives

$$\hat{\pi}_i = \left(1 - \frac{e^{x_{j(i)}'\hat{\beta}}}{\sum_{j \in R_i} e^{x_j'\hat{\beta}}}\right)^{e - x_{j(i)}'\hat{\beta}}.$$

Think of this as follows. With no covariates, our estimate would be $\hat{\pi}_i = 1 - d_i/n_i$ with $d_i = 1$, which is the K-M estimate. With covariates we do the same thing, except that we weight each case by its relative risk. The resulting survival probability is then raised to $e^{-x_{j(i)}'\beta}$ to turn it into a *baseline* probability.

If there are ties one has to solve

$$\sum_{j \in D_i} \frac{e^{x_j'\hat{\beta}}}{1 - \pi_i^{e^{x_j'\hat{\beta}}}} = \sum_{j \in R_i} e^{x_j'\hat{\beta}}$$

iteratively. A suitable starting value is

$$\log \pi_i = -\frac{d_i}{\sum_{j \in R_i} e^{-x_j'\hat{\beta}}}.$$

The estimate of the baseline survival function is then a step function

$$\hat{S}_0(t) = \prod_{i:t_{(i)}<t} \hat{\pi}_i.$$

Cox and Oakes (1984, Section 7.8) describe a simpler estimator that extends the Nelson-Aalen estimate of the cumulative hazard to the case of covariates. The estimator can be described somewhat heuristically as follows. Treat the baseline hazard as 0 except at failure times. The expected number of deaths at $t_{(i)}$ can be obtained by summing the hazards over the risk set:

$$E(d_i) = \sum_{j \in R_i} \lambda_0(t_{(i)}) e^{x_j'\beta}.$$

Equating the observed and expected number of deaths at $t_{(i)}$ leads us to estimate $\lambda_i = \lambda_0(t_{(i)})$ as

$$\hat{\lambda}_i = \frac{d_i}{\sum e^{x_j'\beta}},$$

where the sum is over the risk set R_i. The cumulative hazard and survival functions are then estimated as

$$\hat{\Lambda}_0(t) = \sum_{i:t_{(i)}<t} \hat{\lambda}_i, \qquad \text{and} \qquad \hat{S}_0(t) = e^{-\hat{\Lambda}_0(t)}.$$

If there are no covariates these reduce to the ordinary Nelson-Aalen and Breslow estimators described earlier.

Having obtained estimates of the baseline hazard and survival, we can obtain fitted hazards and survival functions for any value of x. This task is pretty straightforward when we have time-fixed covariates, as all we need to do is multiply the baseline hazard by the relative risk, or raise the baseline survival to the relative risk. With time-varying covariates things get somewhat more complicated, as we have to pick up the appropriate hazard for each distinct failure time depending on the values of the covariates at that point.

3.7 Martingale Residuals

Residuals play an important role in checking linear and generalized linear models. Not surprisingly, the concept has been extended to survival models. A lot of this work relies heavily on the terminology and notation of counting processes. We will try to convey the essential ideas in a non-technical way.

Instead of focusing on the i-th individual's survival time T_i, we will introduce a function $N_i(t)$ that counts events over time. In survival models $N_i(t)$ will be zero while the subject is alive and then it will become one. (In more general event-history models $N_i(t)$ counts the number of occurrences of the event to subject i by time t.) While survival time T_i is a random *variable*, $N_i(t)$ is a stochastic *process*, a function of time.

We will also introduce a function to track the i-th individual's exposure. $Y_i(t)$ will be one while the individual is in the risk set and zero afterwards. Note that $Y_i(t)$ can become zero due to death or due to censoring.

To complete the model we add a hazard function $\lambda_i(t)$ representing the i-th individuals risk at time t. In a Cox model $\lambda_i(t) = \lambda_0(t)\exp\{x_i'\beta\}$.

In the terminology of counting processes, the process $N_i(t)$ is said to have *intensity* $Y_i(t)\lambda_i(t)$. The intensity is just the risk if the subject is exposed, and is zero otherwise. The probability that $N_i(t)$ will jump in a small interval $[t, t+dt)$ conditional on the entire history of the process is given by $\lambda_i(t)Y_i(t)dt$, and is proportional to the intensity of the process and the width of the interval.

A key feature of this formulation is that

$$M_i(t) = N_i(t) - \int_0^t \lambda_i(t)Y_i(t)dt$$

is a *martingale*, a fact that can be used to establish the properties of tests and estimators using martingale central limit theory. A martingale is essentially a stochastic process without drift. Given two times $0 < t_1 < t_2$ the expectation $E(M(t_2))$ given the history up to time t_1 is simply $M(t_1)$. In other words martingale increments have mean zero. Also, martingale increments are uncorrelated, although not necessarily independent.

The integral following the minus sign in the above equation is called the *compensator* of $N_i(t)$. You may think of it as the conditional expected value of the counting process at time t. Subtracting the compensator turns the counting process into a martingale. This equation suggest immediately the first type of residual one can use in Cox models, the so-called *Martingale Residual*:

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(t)e^{x_i'\hat{\beta}}d\hat{\Lambda}_0(t)$$

where $\hat{\Lambda}_0(t)$ denotes the Nelson-Aalen estimator of the baseline hazard. Because this is a discrete function with jumps at the observed failure times, the integral in the above equation should be interpreted as a sum over all j such that $t_{(j)} < t$. Usually the residual is computed at $t = \infty$ (or the largest

observed survival time), in which case the martingale residual is

$$\hat{M}_i = d_i - e^{x_i'\hat{\beta}}\hat{\Lambda}_0(t_i),$$

where d_i is the usual death indicator and t_i is the observation time (to death or censoring) for the i-th individual.

One possible use of residuals is to find outliers, in this case individuals who lived unusually short or long times, after taking into account the relative risks associated with their observed characteristics.

Martingale residuals are just one of several types of residuals that have been proposed for survival models. Others include deviance, score and Schoenfeld residuals. For more details on counting processes, martingales and residuals see Therneau and Grambsch (2000), especially Section 2.2 and Chapter 4.

3.8 Models for Discrete and Grouped Data

We close with a brief review of alternative approaches for discrete and continuous grouped data, expanding slightly on the WWS509 discussion.

Cox (1972) proposed an alternative version of the proportional hazards model for *discrete* data. In this case the hazard is the conditional probability of dying at time t given survival up to that point, so that

$$\lambda(t) = \Pr\{T = t | T \geq t\}.$$

Cox's discrete logistic model assumes that the *conditional* odds of surviving $t_{(i)}$ are proportional to some baseline odds, so that

$$\frac{\lambda(t,x)}{1-\lambda(t,x)} = \frac{\lambda_0(t)}{1-\lambda_0(t)} e^{x'\beta}.$$

Note that taking logs on both sides results in a logit model, where the logit of the discrete hazard is linear on β.

Do not confuse this model with the proportional odds model, where the *unconditional* odds of survival (or odds of dying) are proportional for different values of x.

If the $\lambda_0(t)$ are small so that $1 - \lambda_0(t)$ is close to one, this model will be similar to the proportional hazards model, as the odds are close to the hazard itself.

A nice property of this model is that the partial likelihood turns out to be identical to that of the continuous-time model. To see this point note

that under this model the hazard at time t_i for an individual with covariate values x is

$$\lambda(t_i, x) = \frac{\theta_i e^{x'\beta}}{1 + \theta_i e^{x'\beta}},$$

where $\theta_i = \lambda_0(t_i)/(1 - \lambda_0(t_i))$ denotes the baseline conditional odds of surviving t_i. Suppose there are two cases exposed at time t_i, labelled 1 and 2. The probability that 1 dies and 2 does not is

$$\lambda(t_i, x_1)(1 - \lambda(t_i, x_2)) = \frac{\theta_i e^{x_1'\beta}}{1 + \theta_i e^{x_1'\beta}} \frac{1}{1 + \theta_i e^{x_2'\beta}}.$$

The probability that 2 dies and 1 does not has a similar structure, with $\theta_i e^{x_2'\beta}$ in the numerator and the same denominator. When we divide the probability of one of these outcomes by the sum of the two, the denominators cancel out, as do the θ_i. Thus, the conditional probability is

$$\frac{e^{x_{j(i)}'\beta}}{\sum_{j \in R_i} e^{x_j'\beta}},$$

which is exactly the same as in the continuous case. (You may wonder why we did not consider the $1-\lambda$'s in the continuous case. It turns out we didn't have to. In you repeat the continuous-time derivation including terms of the form λdt for deaths and $1 - \lambda dt$ for survivors you will discover that the dt's for deaths cancel out but those for survivors do not, and as $dt \to 0$ the terms $1 - \lambda dt \to 1$ and drop from the likelihood.)

There is an alternative approach to discrete data that is particularly appropriate when you have *grouped continuous times*. See Kalbfleisch and Prentice (1980), Section 2.4.2 and Section 4.6.1. Suppose we wanted a discrete time model that preserves the relationship between survivor functions in the continuous time model, namely

$$S(t, x) = S_0(t)^{e^{x'\beta}}.$$

In the discrete case we have $S(t, x) = \prod_{u<t}(1 - \lambda(u, x))$, so we must have

$$1 - \lambda(t_i, x) = (1 - \lambda_0(t_i))^{e^{x'\beta}}.$$

Solving for the hazard we get

$$\lambda(t_i, x) = 1 - (1 - \lambda_0(t_i))^{e^{x'\beta}}.$$

The linearizing transformation for this model is the complementary log-log link, $\log(-\log(1-\lambda))$.

To see why this model is uniquely appropriate for grouped data suppose we only observe failures in intervals

$$[0=\tau_0,\tau_1),[\tau_1,\tau_2),\ldots,[\tau_{k-1},\tau_k=\infty).$$

Suppose the hazard is continuous and satisfies a standard proportional hazards model. The probability of surviving interval i for a subject with covariates x is

$$\Pr\{T>\tau_i|T>\tau_{i-1},x\}=\frac{S(\tau_i,x)}{S(\tau_{i-1},x)}.$$

Writing this in terms of the baseline survival we get

$$\left(\frac{S_0(\tau_i)}{S_0(\tau_{i-1})}\right)^{e^{x'\beta}}=\left(e^{-\int_{\tau_{i-1}}^{\tau_i}\lambda_0(t)dt}\right)^{e^{x'\beta}}.$$

In view of this result, we define the baseline hazard in interval (τ_{i-1},τ_i) as

$$\lambda_{0i}=1-e^{-\int_{\tau_{i-1}}^{\tau_i}\lambda_0(t)dt}.$$

The hazard for an individual with covariates x in the same interval then becomes

$$\lambda_i(x)=1-(1-\lambda_{0i})^{e^{x'\beta}},$$

and can be linearized using the c-log-log transformation.

This is the only discrete model appropriate for grouped data from a continuous-time proportional hazards model. Unfortunately, it cannot be estimated non-parametrically using a partial likelihood argument. (If you try to construct a partial likelihood you will discover that the λ_{0i}'s do not drop out of the likelihood.)

In practice, both the logit and the complementary log-log discrete models can be estimated via ordinary likelihood techniques by treating the baseline hazard at each discrete failure time (or interval) as a separate parameter to be estimated, provided of course one has enough failures at each time (or interval). The resulting likelihood is the same as that of a binomial model with logit or c-log-log link, so either model can be easily fitted with standard software.

One difficulty with discrete models generated by grouping continuous data has to do with censored cases that may have been exposed part of the interval. The only way to handle these cases is to censor them at the

beginning of the interval, which throws away some information. This is not a problem with piece-wise exponential models based on the Poisson equivalence, because one can easily take into account partial exposure in the offset.

Cumulative Incidence

Germán Rodríguez
grodri@princeton.edu

Spring 2012

We continue our treatment of competing risks by considering estimation of the cumulative incidence function and the Fine and Gray competing risks regression model.

1 The Cumulative Incidence Function

In our earlier discussion we introduced the cause-specific densities

$$f_j(t) = \lim_{dt \downarrow 0} \Pr\{T \in (t, t+dt) \text{ and } J = j\}/dt$$

which have the property of summing to the overall density $f(t) = \sum_j f_j(t)$.

The integral

$$I_j(t) = \int_0^t f_j(u)du = \Pr\{T \leq t \text{ and } J = j\}$$

is called the *cumulative incidence function* (CIF), and represents the probability that an event of type j has occurred by time t.

Earlier we also introduced the cause-specific hazards

$$\lambda_j(t) = \lim_{dt \downarrow 0} \Pr\{T \in (t, t+dt) \text{ and } J = j | T > t\}/dt$$

representing the (conditional) rate of occurrence of events of type j at time t among survivors to that time. The cause-specific density can be written as

$$f_j(t) = S(t)\lambda_j(t)$$

reflecting the fact that to experience an event of type j at time t you first have to survive to time t, and then experience an event of type j conditional on having survived to t.

This representation leads to a non-parametric estimator of the cumulative incidence function which extends the Kaplan-Meier estimator. With distinct failure times $0 < t_{(1)} < \cdots < t_{(m)} < \infty$, the estimator is

$$\hat{I}_j(t) = \sum_{i:t_{(i)} \leq t} \hat{S}(t_{(i)}) \frac{d_{ij}}{n_i}$$

where d_{ij} is the number of events of type j at time $t_{(i)}$, n_i is the total number of observations at risk at time $t_{(i)}$, and $\hat{S}(t_{(i)}$ is the standard Kaplan-Meier estimator of survival to time $t_{(i)}$.

This is a step function with increments every time a failure of type j occurs. An interesting feature of this function is that if we add the cumulative incidence of all types of failure we obtain the complement of the Kaplan-Meier estimator:

$$\sum_j \hat{I}_j(t) = 1 - \hat{S}(t)$$

In words, at any time t the observations are either still at risk with probability $S(t)$, or have experienced an event of type j with probability $I_j(t)$ for some j. In the case of mortality you are either alive or have succumbed to cause of death j.

Standard errors for the cumulative incidence function can be obtained using the delta method, although the derivation is a bit more complicated that in the case of Greenwood's formula.

In the Stata logs we study how long U.S. Supreme Court Justices serve on the court, treating death and retirement as competing risks, with the nine justices currently serving treated as censored observations. We find, for example, that averaging over the existence of court, the probability that a justice will die on the job is 48% and the probability of retiring is 52%.

2 The Fine-Gray Model

How do you introduce covariates in the context of competing risks? There are essentially two approaches:

1. You can apply a Cox proportional-hazards model to the cause-specific hazards introduced earlier, or

2. You can use a model due to Fine and Gray that focuses on the cumulative incidence function.

The first approach is more structural, focusing on the covariates of the risk of each type of event. The second approach is more descriptive, focusing on the probability of each event type.

To understand the difference in approaches note that a covariate may appear to increase the incident of events of a certain type simply by lowering the rate of occurrence of events of other types, even if it has no effect on the rate of occurrence of the event in question.

We now describe the Fine and Gray model. Let $I_j(t,x)$ denote the cumulative incidence function for events of type j given a vector of covariates x. We can formally treat the complement of the CIF as a survival function and calculate the underlying hazard. To avoid confusion with the cause-specific and overall hazards we follow Fine and Gray in calling this a *sub*-hazard for cause j and denote it with a bar

$$\bar{\lambda}_j(t,x) = -\frac{d}{dt}\log(1 - I_j(t,x)) = \frac{f_j(t)}{1 - I_j(t)}$$

They then propose a proportional hazards model for the sub-hazard associated with type j, effectively writing it as

$$\bar{\lambda}_j(t,x) = \bar{\lambda}_{j0}(t)\exp\{x'\beta_j\}$$

where $\bar{\lambda}_{j0}(t)$ is the baseline sub-hazard for events of type j and $\exp\{x'\beta_j\}$ is the relative risk associated with covariates x.

While the formulation looks very similar to Cox regression, the present model applies to the sub-hazard underlying the cumulative incidence function, not the cause-specific hazards. One problem with this approach is that the sub-hazard is hard to interpret. From the Fine and Gray definition,

$$\begin{aligned}\bar{\lambda}_j(t) = \quad & \lim_{dt\downarrow 0}\Pr\{T \in (t, t+dt) \text{ and } J = j| \\ & T > t \textbf{ or } T \le t \text{ and } J \ne j\}/dt\end{aligned}$$

In other words, we count events of type j in a small interval $(t, t+dt)$ but treat as the risk set those alive at t *and* those who failed before t due to causes other than j.

The authors themselves recognize that this is an "un-natural" hazard because units who experienced an event of some other type before time t are not really at risk of experiencing an event of type j at t. One way to derive the sub-hazard as a standard hazard is to imagine a random variable T^* which equals T_j if an event of type j occurs and equals ∞ if an event of another type occurs, but this is also artificial.

In the end, the authors argue that their formulation is just a convenient way to model the incidence function. I agree, and tend to view their model as just a binary outcome model for the cumulative incidence function using the complementary log-log link. This is because under their model

$$\log(-\log(1 - I_j(t, x))) = \log(-\log(1 - I_{j0}(t))) + x'\beta_j$$

Thus, the effect of the covariates is to shift the transformed CIF up or down by an amount depending on the coefficients. Because the transformation is monotonic we know that positive coefficients indicate increases in the CIF and negative coefficients indicate decreases, but quantifying the effect requires conducting illustrative calculations.

In the Stata logs we study the length of service of U.S. Supreme Court justices treating death and retirement as competing risks and age at appointment and calendar year of appointment as predictors. (This is one case where estimating anything at zero values of the covariates is fraught with peril, as the court was founded in 1789 and age at appointment goes from 33 to 66.)

Fitting a Cox model to the hazard of death gives hazard ratios of 1.07 for age and 0.99 for calendar year, so the risk of death increases 7% per year of age at appointment and declines about one per cent per calendar year of appointment.

Fitting a similar Cox model to the risk of retirement gives hazard ratios of 1.10 for age and 1.00 for year, so the risk of retiring increases 10% per year of age at appointment and does not depend on the calendar year of appointment.

These two models give us a good understanding of the underlying process, and they can be used to estimate overall survival, cause-specific densities and hazards, and even the CIFs of death and retirement from their definitions.

Alternatively, we can fit a Fine-Gray model directly to the CIF for death or for retirement. Fitting a model to the CIF of death gives sub-hazard ratios (called SHR in Stata) of 1.01 for age at appointment (not significant) and 0.99 for calendar year of appointment (highly significant).

The first finding is that the probability of dying while serving on the court does *not* depend on age at appointment. You may find this result a bit surprising, as I did, but note that justices who are appointed at an older age have a higher risk of death than those appointed at younger ages in the same period, but they also have a higher risk of retirement, and these two forces are about equal so they balance out.

The second finding is that the probability of dying in the court has declined with calendar year of appointment, so justices appointed more recently are less likely to die and hence more likely to retire. The sub-hazard ratio of 0.99, however, is hard to interpret in terms other than the sign and significance without additional calculations.

The best bet here is to compute illustrative values of the CIF. In the Stata logs we show that the probability that a justice appointed at age 55 will leave the court by death is 29.7% if appointed in 1950 and 20.2% if appointed in 2000 (both figures lower than the overall mean of 48%). Note that

$$\log(-\log(1-.202)) - \log(-\log(1-.297)) = -0.446,$$

and $-0.446/50 = -0.009$. the coefficient of year, which Stata reports as an SHR of $\exp\{-0.009\} = 0.99$. Thus, the transformed CIF is declining 0.009 per calendar year.

Unobserved Heterogeneity

Germán Rodríguez
grodri@princeton.edu

Spring, 2001. Revised Spring 2005*

This unit considers survival models with a random effect representing unobserved heterogeneity of *frailty*, a term first introduced by Vaupel et al. (1979). We consider models without covariates and then move on to the more general case. These notes are intended to complement Rodríguez (1995).

1 The Statistics of Heterogeneity

Standard survival models assume homogeneity: all individuals are subject to the same risks embodied in the hazard $\lambda(t)$ or the survivor functions $S(t)$. Models with covariates relax this assumption by introducing observed sources of heterogeneity. Here we consider unobserved sources of heterogeneity that are not readily captured by covariates.

1.1 Conditional Hazard and Survival

A popular approach to modeling such sources is the *multiplicative frailty* model, where the hazard for individual i at time t is

$$\lambda_i(t) = \lambda(t|\theta_i) = \theta_i \lambda_0(t),$$

the product of an individual-specific random effect θ_i representing the individual's *frailty*, and a baseline hazard $\lambda_0(t)$. Note that this is essentially a proportional hazards model.

The individual hazard $\lambda_i(t)$ is interpreted as a *conditional* hazard given θ_i. Associated with it we have a conditional survival function

$$S_i(t) = S(t|\theta_i) = S_0(t)^{\theta_i},$$

*Minor revisions Spring 2014

representing the probability of being alive at t given the random effect θ_i.

The twist is that the random effect θ_i is not observed (perhaps not observable), but is assumed to have some a distribution with density $g(\theta)$.

1.2 Unconditional Hazard and Survival

To obtain the *unconditional* survival function we need to "integrate out" the unobserved random effect:

$$S(t) = \int_0^\infty S(t|\theta) g(\theta) d\theta.$$

We integrate from 0 to ∞ because frailty is non-negative. If frailty was discrete, taking values $\theta_1, \ldots, \theta_k$ with probabilities $\pi_1, \ldots, \pi_k$ then the integral would be replaced by a sum

$$S(t) = \sum_i S(t|\theta_i)\pi_i.$$

In both cases $S(t)$ is the average $S_i(t)$. In a demographic context $S(t)$ is often referred to as the *population* survivor function, while $S_i(t)$ is the *individual* survivor function.

To obtain the unconditional hazard we start from the unconditional survival and take negative logs to obtain the cumulative hazard

$$\begin{aligned}
\Lambda(t) &= -\log S(t) \\
&= -\log \int_0^\infty S(t|\theta) g(\theta) d\theta \\
&= -\log \int_0^\infty S_0(t)^\theta g(\theta) d\theta.
\end{aligned}$$

The next step is to take derivatives w.r.t. t. Assuming that we can take the derivative operator inside the integral we find the unconditional hazard to be

$$\lambda(t) = -\frac{\int_0^\infty \frac{d}{dt} S_0(t)^\theta g(\theta) d\theta}{\int_0^\infty S_0(t)^\theta g(\theta) d\theta} = \frac{\int_0^\infty \theta \lambda_0(t) S_0(t)^\theta g(\theta) d\theta}{\int_0^\infty S_0(t)^\theta g(\theta) d\theta},$$

where we used the fact that $S_0(t)^\theta = e^{-\theta \Lambda_0(t)}$, so that

$$\frac{d}{dt} S_0(t) = -e^{-\theta \Lambda_0(t)} \theta \lambda_0(t) = -\theta \lambda_0(t) e^{-\theta \Lambda_0(t)},$$

and the last exponential can be recognized as $S_0(t)^\theta$.

Note that the population hazard $\lambda(t)$ is a weighted average of the individual hazards $\lambda_i(t)$ with weights equal to the density of θ times the probability of surviving to t:

$$S(t|\theta)g(\theta) = S_0(t)^\theta g(\theta),$$

Why can't we calculate the population hazard as a simple average of the individual hazards, the way we calculated the population survivor function?

1.3 Expected Frailty of Survivors

We now show that the weights in the above expression represent the conditional distribution of frailty θ among survivors to age t. From first principles, the density of θ among survivors is

$$g(\theta|T \geq t) = \frac{\Pr\{T \geq t|\theta\}g(\theta)}{\Pr\{T \geq t\}} = \frac{S(t|\theta)g(\theta)}{\int_0^\infty S(t|\theta)g(\theta)d\theta} = \frac{S_0(t)^\theta g(\theta)}{\int_0^\infty S_0(t)^\theta g(\theta)d\theta},$$

which are indeed the weights in the expression for $\lambda(t)$. The expected frailty of survivors can be calculated as

$$E(\theta|T \geq t) = \int_0^\infty \theta g(\theta|T \geq t)d\theta = \frac{\int_0^\infty \theta S_0(t)^\theta g(\theta)d\theta}{\int_0^\infty S_0(t)^\theta g(\theta)d\theta}.$$

From this result is becomes clear that

$$\lambda(t) = \lambda_0(t)E(\theta|T \geq t). \tag{1}$$

In words, the unconditional (population) hazard at t is the baseline (individual) hazard times the mean frailty of survivors to t.

Typically, the mean frailty of survivors declines over time as the more frail tend to die earlier. As a result, the population hazard declines more steeply (or increases less rapidly) than the individual hazard. This result is the source of interesting paradoxes.

2 Frailty Distributions

We now specialize our results considering a few alternative assumptions about the distribution of frailty.

2.1 Gamma Frailty

A convenient assumption used by many authors is that θ has a gamma distribution. This distribution has the appropriate range $(0, \infty)$ and is mathematically tractable.

The density of a gamma distribution with parameters α and β is

$$g(\theta) = \theta^{\alpha-1} e^{-\beta\theta} \beta^{\alpha} / \Gamma(\alpha),$$

where Γ is the gamma function. The mean and variance are

$$E(\theta) = \frac{\alpha}{\beta} \qquad \text{and} \qquad \text{var}(\theta) = \frac{\alpha}{\beta^2},$$

so the coefficient of variation σ/μ is $1/\sqrt{\alpha}$.

It is often convenient to take $E(\theta) = 1$ so $\alpha = \beta = 1/\sigma^2$. This entails no loss of generality because the average level of frailty can always be absorbed into the baseline hazard.

2.1.1 Unconditional Survival and Hazard

The unconditional survivor function is

$$\begin{aligned} S(t) &= \int_0^{\infty} S_0(t)^{\theta} g(\theta) d\theta \\ &= \int_0^{\infty} e^{-\theta\Lambda_0(t)} \theta^{\alpha-1} e^{-\beta\theta} \beta^{\alpha} \frac{1}{\Gamma(\alpha)} d\theta. \end{aligned}$$

The trick now is to consolidate the coefficients of θ and complete a gamma density:

$$S(t) = \int_0^{\infty} \theta^{\alpha-1} e^{-(\beta+\Lambda_0(t))\theta} (\beta + \Lambda_0(t))^{\alpha} \frac{1}{\Gamma(\alpha)} \frac{\beta^{\alpha}}{(\beta + \Lambda_0(t))^{\alpha}} d\theta.$$

The last fraction on the right does not depend on θ and can be taken out of the integral. What's left is a gamma density with parameters α and $\beta + \Lambda_0(t)$, and therefore integrates to one. This gives

$$S(t) = \left(\frac{\beta}{\beta + \Lambda_0(t)} \right)^{\alpha}.$$

The result is known as a Pareto distribution of the second kind. If frailty has mean one and variance σ^2, we write $\alpha = \beta = 1/\sigma^2$ to obtain

$$S(t) = \frac{1}{(1 + \sigma^2 \Lambda_0(t))^{1/\sigma^2}}, \tag{2}$$

the unconditional (population) survivor function under gamma frailty.

To find the unconditional (population) hazard we first take negative logs to obtain the cumulative hazard

$$\Lambda(t) = \alpha \log(\beta + \Lambda_0(t)) - \alpha \log(\beta),$$

and then take derivatives w.r.t. t, to obtain

$$\lambda(t) = \frac{\alpha \lambda_0(t)}{\beta + \Lambda_0(t)}.$$

If frailty has mean one and variance σ^2 we obtain

$$\lambda(t) = \frac{\lambda_0(t)}{1 + \sigma^2 \Lambda_0(t)}, \tag{3}$$

the unconditional (population) hazard function under gamma frailty.

Example: Gamma mixtures of exponentials are a popular model for unobserved heterogeneity. If the hazard is constant for each individual but people are heterogenous and frailty has a gamma distribution then the population hazard is

$$\lambda(t) = \frac{\lambda}{1 + \sigma^2 \lambda t},$$

where λ is the average individual hazard and σ^2 is the variance of frailty.

2.1.2 The Frailty of Survivors

In view of our earlier result connecting $\lambda(t)$ and $E(\theta|T \geq t)$, the expected frailty of survivors to t under gamma frailty must be

$$E(\theta|T \geq t) = \frac{1}{1 + \sigma^2 \Lambda_0(t)}$$

In fact, we can obtain the whole distribution of frailty among survivors to t. The conditional density of θ given $T \geq t$ is

$$\begin{aligned} g(\theta|T \geq t) &= \frac{S(t|\theta) g(\theta)}{S(t)} \\ &= \frac{e^{-\theta \Lambda_0(t)} \theta^{\alpha-1} e^{-\beta\theta} \beta^\alpha / \Gamma(\alpha)}{\frac{\beta^\alpha}{(\beta + \Lambda_0(t))^\alpha}} \\ &= \theta^{\alpha-1} e^{-(\beta + \Lambda_0(t))^\theta} (\beta + \Lambda_0(t))^\alpha / \Gamma(\alpha), \end{aligned}$$

a gamma density with parameters α and $\beta + \Lambda_0(t)$. Thus, if frailty at birth has a gamma distribution with mean one and variance σ^2, so $\alpha = \beta = 1/\sigma^2$, then frailty of survivors to t has a gamma distribution with

$$E(\theta|T \geq t) = \frac{1}{1 + \sigma^2 \Lambda_0(t)} \qquad \text{and} \qquad \text{var}(\theta|T \geq t) = \frac{\sigma^2}{(1 + \sigma^2 \Lambda_0(t))^2}.$$

Note that $\Lambda_0(t)$ is a monotone non-decreasing function of t. As a result, the mean frailty of survivors declines over time. The variance of frailty of survivors also declines over time, so the population becomes more homogeneous in absolute terms. However, the coefficient of variation stays constant over time, so the population does not become more homogeneous in relative terms (compared to the mean).

Note also that mean frailty will decline more rapidly over time (or selectivity will operate more quickly) when (1) the population is more heterogeneous to start with (larger σ^2), or (2) the risk is higher (larger $\Lambda_0(t)$).

2.2 Inverse Gaussian Frailty

Another distribution that can be used to represent frailty is the inverse Gaussian distribution, which arises as the first passage time in Brownian motion.

Hougaard (1984) has shown that if θ has an inverse Gaussian distribution with mean and variance σ^2 then the mean frailty of survivors to time t is

$$E(\theta|T \geq t) = \frac{1}{(1 + 2\sigma^2 \Lambda_0(t))^{1/2}}.$$

It then follows from our general results that the unconditional (population) hazard is

$$\lambda(t) = \frac{\lambda_0(t)}{\sqrt{1 + 2\sigma^2 \Lambda_0(t)}}.$$

The unconditional (population) survivor function can also be obtained explicitly, and turns out to be

$$S(t) = \exp\{-\frac{1}{\sigma^2}(\sqrt{1 + 2\sigma^2 \Lambda_0(t)} - 1)\},$$

a result that can easily be verified by taking negative logs to get $\Lambda(t)$ and differentiating w.r.t. t to obtain $\lambda(t)$. Can you derive this result?

2.2.1 Notes on the Inverse Gaussian Distribution

The inverse Gaussian distribution has density

$$g(\theta) = \sqrt{\frac{\alpha}{\pi}} e^{\sqrt{4\alpha\beta}} \theta^{-\frac{3}{2}} e^{-\alpha/\theta - \beta\theta},$$

depending on parameters α and β (called ψ and θ by Hougaard (1984), who uses z for our θ). This distribution has mean and variance

$$E(\theta) = \sqrt{\frac{\alpha}{\beta}} \qquad \text{and} \qquad \text{var}(\theta) = \frac{1}{2}\alpha^{1/2}\beta^{-3/2},$$

so the coefficient of variation σ/μ is $1/\sqrt{2}(\alpha\beta)^{1/4}$. Choosing $\alpha = \beta$ gives a mean of one and variance $\text{var}(\theta) = \frac{1}{2}\alpha^{\frac{1}{2}}\beta^{-\frac{3}{2}} = \frac{1}{2\beta}$, so to get a distribution with variance σ^2 we take $\alpha = \beta = \frac{1}{2\sigma^2}$.

Hougaard shows that under the multiplicative frailty model the distribution of θ among survivors to t is also inverse Gaussian, with parameters α and $\beta + \Lambda_0(t)$. In particular, the mean frailty of survivors is

$$E(\theta|T \geq t) = \sqrt{\frac{\alpha}{\beta + \Lambda_0(t)}} = \frac{1}{(1 + 2\sigma^2\Lambda_0(t))^{1/2}}.$$

Interestingly, the distribution of frailty among those who die at t is a "generalized" inverse Gaussian.

2.3 A More General Family

If you look again at the expressions for $E(\theta|T \geq t)$, the mean frailty of survivors to t, under gamma and inverse Gaussian frailty, you will notice a certain resemblance. In fact, you could write

$$E(\theta|T \geq t) = \frac{1}{(1 + \frac{\sigma^2}{k}\Lambda_0(t))^k}, \tag{4}$$

where $k = 1$ gives the result for gamma frailty and $k = 1/2$ gives the result for inverse Gaussian frailty.

One naturally wonders whether this result makes sense more generally. Does this formula represent expected frailty of survivors under some distribution for other values of k?

Hougaard (1986) shows that Equation 4 makes sense for any $k < 1$, yielding a family based on the *stable distributions*, which includes the inverse Gaussian as a special case.

A distribution is called "stable" if the distribution of the sum of n i.i.d. r.v.'s is the same as the distribution of $n^{1/\alpha}$ times one of them for some $\alpha \in [0, 2]$. In symbols,

$$\mathcal{D}(X_1 + X_2 + \ldots + X_n) = \mathcal{D}(n^{\frac{1}{\alpha}} X_1),$$

where $\mathcal{D}$ denotes distribution. For example the normal distribution $N(\mu, \sigma^2)$ is stable with $\alpha = 1$, because $\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$.

Aalen (1988) showed that Equation 4 also makes sense for $k > 1$, showing that all the remaining cases could be obtained by assuming that θ has a *compound Poisson distribution.*

To construct this distribution suppose N is distributed Poisson and $X_1, X_2, \ldots$ are i.i.d. gamma r.v.'s, and define

$$\theta = \begin{cases} 0 & \text{if } N = 0 \\ X_1 + \ldots + X_N & \text{if } N > 0 \end{cases}$$

One way to think about this distribution is to imagine a population that has infinitely many strata, one with no frailty, one where frailty is gamma, one where frailty is the sum of two gammas, and so on, with relative stratum sizes given by a Poisson distribution.

Note that this distribution leads to improper survival functions, because for some people $\theta = 0$ and the event of interest has no risk of ever occurring.

2.4 Frailty Transforms

A very useful tool in frailty analysis is the *Laplace transform.* Given a function $f(x)$, the Laplace transform, considered as a function of a real argument s is defined as

$$\mathcal{L}(s) = \int_0^\infty e^{-sx} f(x) dx.$$

The reason why this is useful in our context is that the Laplace transform has exactly the same form as the unconditional survival function. Think of $f(x)$ as the frailty distribution $g(\theta)$ and s as the cumulative baseline hazard $\Lambda_0(t)$ and you obtain

$$\begin{aligned} S(t) &= \int_0^\infty e^{-\Lambda_0(t)\theta} g(\theta) d\theta \\ &= \mathcal{L}(\Lambda_0(t)), \end{aligned}$$

where $\mathcal{L}$ denotes the Laplace transform.

Because Laplace transforms are well-know, and many are tabulated, our task is easier. For example the Laplace transform of the gamma distribution with parameters α and β is

$$\mathcal{L}(s) = \left(\frac{\beta}{\beta + s}\right)^{\alpha}.$$

Evaluating this at $s = \Lambda_0(t)$ we obtain the same result as before, but with a lot less work.

Vaupel (1990) has defined the *frailty transform* as the function

$$\mathcal{F}(m, s) = \int_0^{\infty} \theta^m e^{-s\theta} g(\theta) d\theta.$$

Note that $\mathcal{F}(0, s)$ is the good old Laplace transform, and $\mathcal{F}(m, s)$ gives the m-th moment of the distribution of θ at birth. For the gamma distribution the frailty transform is

$$\mathcal{F}(m, s) = \frac{\Gamma(\alpha + m)}{\Gamma(m)} \frac{\beta^{\alpha}}{(\beta + s)^{\alpha+m}}.$$

The connection with Laplace transforms has practical as well as theoretical importance.

- Given a function $f(x)$, computation of the Laplace transform is a well understood problem with efficient algorithms.
- Given the Laplace transform $\mathcal{L}(s)$, recovery of the function $f(x)$ by inversion is an ill-conditioned problem, in the sense that slight changes in $\mathcal{L}(s)$ can induce huge fluctuations in $f(x)$.

3 The Inversion Formula

So far we have gone from conditional to unconditional (or if you wish from individual to population) hazard and survival by a process of "mixing". Can we go the other way? Can we infer the unconditional (individual) hazard and survival from the conditional (population) counterparts by a process of "unmixing"?

The answer is yes, provided we know the distribution of frailty (or how the mixing was done). In the next two subsections we provide inversion formulas for gamma and inverse Gaussian frailty. These results can be used to express population survival functions as gamma or inverse Gaussian mixtures of individual survival functions.

3.1 Gamma Mixtures

We have shown that under gamma frailty the unconditional hazard can be written as

$$\lambda(t) = \frac{\lambda_0(t)}{1 + \sigma^2 \Lambda_0(t)}.$$

We will integrate the left-hand side to obtain the cumulative hazard $\Lambda(t)$. In order to do this it helps to rewrite the previous equation as a derivative

$$\lambda(t) = \frac{1}{\sigma^2} \frac{d}{dt} \log(1 + \sigma^2 \Lambda_0(t)),$$

because then we can integrate tp pbtain

$$\Lambda(t) = \frac{1}{\sigma^2} \log(1 + \sigma^2 \Lambda_0(t)),$$

where we used the boundary condition $\Lambda_0(t) = 0$. This gives

$$1 + \sigma^2 \Lambda_0(t) = e^{\sigma^2 \Lambda(t)},$$

or

$$\Lambda_0(t) = \frac{1}{\sigma^2}(e^{\sigma^2 \Lambda(t)} - 1).$$

Taking derivatives w.r.t. t we obtain the conditional (individual) baseline hazard as a function of the unconditional (population) hazard

$$\lambda_0(t) = \lambda(t) e^{\sigma^2 \Lambda(t)}. \tag{5}$$

Example: We noted earlier the popularity of gamma mixtures of exponentials. We now show that the exponential distribution itself can be viewed as a gamma mixture of something else. If the population survival function is exponential then

$$\lambda(t) = \lambda \qquad \text{and} \qquad \Lambda(t) = \lambda t.$$

Plugging these functions into our inversion formula we find the conditional (individual) hazard to be

$$\lambda_0(t) = \lambda e^{\sigma^2 \lambda t},$$

which we recognize as a Gompertz or extreme value hazard, where the log of the hazard is a linear funtion of t.

Thus, we have the remarkable result that a population that shows a constant hazard over time may result from individuals with gamma-distributed heterogeneity and Gompertz hazards that increase exponentially with time.

You may begin to suspect that we have a bit of an indentification problem here, because a flat population hazard could also result from a homogeneous population where the hazard for each individual is constant over time.

No amount of data can help us distinguish between these two models because they have identical observable consequences.

3.2 Inverse Gaussian Mixtures

We can also obtain an inversion formula for inverse Gaussian frailty. Recall that the unconditional hazard was

$$\lambda(t) = \frac{\lambda_0(t)}{(1+2\sigma^2\Lambda_0(t))^{1/2}}.$$

As before, we write the right-hand side as a derivative, so integrating is simpler:

$$\lambda(t) = \frac{1}{\sigma^2}\frac{d}{dt}(1+2\sigma^2\Lambda_0(t))^{1/2}.$$

To integrate from 0 to t we impose the boundary condition $\Lambda(t) = 0$ and obtain

$$\Lambda(t) = \frac{1}{\sigma^2}((1+2\sigma^2\Lambda_0(t))^{1/2} - 1),$$

which incidentally answers the question posed earlier, on how to derive the unconditional survival for inverse Gaussian frailty (see page 2.2). Now we use this result to solve for the baseline integrated hazard:

$$\begin{aligned}(1+2\sigma^2\Lambda_0(t))^{1/2} &= 1+\sigma^2\Lambda(t)\\ 1+2\sigma^2\Lambda_0(t) &= (1+\sigma^2\Lambda(t))^2\\ \Lambda_0(t) &= \frac{(1+\sigma^2\Lambda(t))^2-1}{2\sigma^2}.\end{aligned}$$

Now take derivatives w.r.t. t to obtain

$$\begin{aligned}\lambda_0(t) &= \frac{1}{2\sigma^2}2(1+\sigma^2\Lambda(t))\sigma^2\lambda(t)\\ &= \lambda(t)(1+\sigma^2\Lambda(t)).\end{aligned}$$

This result gives us a baseline hazard $\lambda_0(t)$ that can be mixed using an inverse Gaussian distribution to obtain any given population hazard $\lambda(t)$.

Example: Using this result we should be able to produce an exponential distribution as an inverse Gaussian mixture of something else. Let's try. If the population survival function is exponential then

$$\lambda(t) = \lambda \qquad \text{and} \qquad \Lambda(t) = \lambda t,$$

and plugging these into our general result we obtain

$$\lambda_0(t) = \lambda(1 + \sigma^2\lambda t) = \lambda + \sigma^2\lambda^2 t,$$

a linear hazard.

Thus, a population with a constant hazard could consist of an inverse Gaussian mix of individuals with linearly rising hazards. Note that σ^2 is not specified, so the steepness of the individual hazards is arbitrary.

My conclusion from these results is that models with unobserved heterogeneity are not identified in the sense that we can not distinguish between competing models that have identical observable consequences.

However, it is very important to know that the data we observe could have been generated by different mechanisms. For example in the analysis of waiting time to conception the hazard typically declines over time. This could be due to the fact that the hazard actually declines for each individual. But is is also possible that the individual hazard is constant and the observed decline reflects a selection effect.

4 Models with Covariates

Many of the ideas discussed so far extend to models with covariates. Here we will summarize some of the key ideas.

4.1 The Omitted Variable Bias

We know from linear models that omitting a variable from a model introduces a bias unless the omitted variable is uncorrelated with the other predictors in the model. In hazard models it turns out that we obtain a bias *even* if the the omitted variable is uncorrelated with the predictors.

To see this point consider a simple problem with two dummy variables, x_1 and x_2. Suppose these variables are independent and $\frac{1}{4}$ of the population falls in each of the four categories defined by them.

Suppose further that both variables affect survival time. When both are zero the hazard is constant at $\lambda_0(t) = 1$, x_1 doubles the risk, so $e^{\beta_1} = 2$, and x_2 triples the risk, so $e^{\beta_2} = 3$.

Under these assumptions we have four exponentials with a proportional hazards structure. If i denotes the value of x_1 and j the value of x_2, the hazard is $\lambda_{ij}(t) = 2^i 3^j$ for $i = 0, 1; j = 0, 1$.

Now suppose we do not observe x_2. The survival functions we observe for $x_1 = i$ are a mixture of the curves for $x_2 = 0$ and $x_2 = 1$ with equal

weights, so we can write

$$\begin{aligned} S_i(t) &= \frac{1}{2}S_{i0}(t) + \frac{1}{2}S_{i1}(t) \\ &= \frac{1}{2}e^{-2^i t} + \frac{1}{2}e^{-2^i 3t} \end{aligned}$$

We can take negative logs to obtain the cumulative hazards and then differentiate to obtain the hazards. (If you do this numerically you can just difference the cumulative hazards.) Figure 1 shows the resulting hazards.

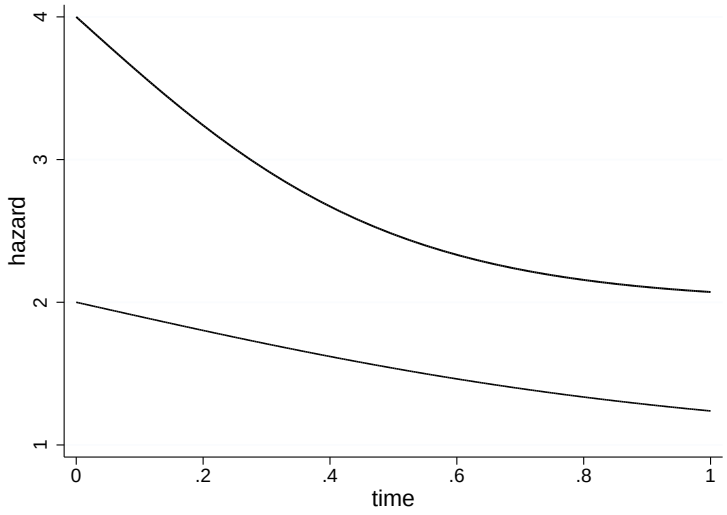


Figure 1: Proportional Hazards with Unobserved Heterogeneity

First, the hazards are *not constant*, even though we started with exponentials. This shows the effect of selection. Individuals with $x_2 = 1$ die more quickly than those with $x_2 = 0$ and are selected out of the risk set, so the observed hazard declines over time.

Second, the hazards are *not proportional*, even though we started with a proportional hazards structure. Recall that the hazard for $x_1 = 1$ was twice the hazard for $x_1 = 0$, holding everything else constant. This is now true only at $t = 0$. After that the curves come closer together and the effect of $x_1 = 1$ is less than 2. If you fitted a proportional hazards model to data generated from this model you would underestimate the effect of x_1.

The reason for this result is that selection happens more quickly when the hazard is higher (as shown before). The group with $x_1 = 1$ has a higher

hazard (in fact, twice the hazard)—and therefore can select out the frail (those with $x_2 = 1$) more quickly—than the group with $x_1 = 0$.

This happened even though x_1 and x_2 were independent. The key to understanding these results is to realize that they were independent at $t = 0$ but they are no longer independent at $t > 0$. The proportion "frail" (i.e. with $x_2 = 1$) is 50% for $x_1 = 1$ and for $x_1 = 0$ at the outset, but is 12% for $x_1 = 1$ compared to 2% for $x_1 = 0$ at $t = 1$. Can you reproduce these percents? As you can imagine, the situation is worse if x_1 and x_2 are correlated.

4.2 Models with Unobserved Heterogeneity

One possible solution to the problem of unobserved heterogeneity is to introduce a random effect θ in the hope that it will capture the effects of omitted variables that are independent of the X's in the model.

The general model we will entertain is a proportional hazards model with a frailty term, where the hazard at time t for an individual with covariates x and frailty θ is

$$\lambda(t, x, \theta) = \theta \lambda_0(t) e^{x'\beta},$$

where θ is a random effect with mean zero and a distribution that does not depend on the observed covariates.

Estimation of this model can be done by maximum likelihood using standard techniques if you can assume

- a parametric form for the baseline hazard $\lambda_0(t)$, and
- a distribution for the random effect θ.

For example Newman and McCullogh analyzed birth intervals using gamma frailty (here representing fecundability). An alternative tractable functional form is the inverse Gaussian.

It is possible to relax one of these two assumptions, but not both.

4.3 Heckman-Singer

Heckman and Singer (1984), in a very influential paper, noted some instability of parameter estimates depending on the type of assumption made about the distribution of frailty.

As a solution, they proposed using a non-parametric maximum likelihood estimator (NPMLE) of the distribution of frailty. Following on earlier work by Laird and others, they show that the NPMLE is a discrete mixing distribution that assigns positive mass to a finite (usually small) set of points of support.

Specifically, the non-parametric estimate takes values $\theta_1, \theta_2, \ldots, \theta_k$ with probabilities $\pi_1, \pi_2, \ldots, \pi_k$ for some value of k. The distribution has $2(k-1)$ parameters if one restricts θ to have mean one.

Usually one fits a model with $k = 2$ and increases k by adding an additional point of support until the likelihood fails to improve, at which point two of the points often coalesce. When one of the points has negligible risk ($\theta \approx 0$) the result can be interpreted as a mover-stayer model.

Flexibility in estimation of the frailty distribution requires parametric assumptions about the hazard. A common choice used in the program CTM (Continuous Time Models) developed by Heckman and associates is the Box-Cox specification

$$\lambda_0(t) = \lambda + \beta_1 \frac{t^{\lambda_1} - 1}{\lambda_1} + \beta_2 \frac{t^{\lambda_2} - 1}{\lambda_2},$$

where t^λ is interpreted as $\log t$ for $\lambda = 0$. This includes as special cases the exponential, Weibull, Gompertz and a log-quadratic hazard.

4.4 Trussell-Richards

Trussell and Richards (1985) wondered whether models estimated using the Heckman-Singer technique were sensitive to the choice of baseline hazard. They found that the results were indeed sensitive, a conclusion confirmed in further work by Trussell and Montgomery.

In fact, it seems clear that the results should be more sensitive to the choice of the baseline hazard than to the choice of the distribution of the unobservable. Why? Recall that the unconditional survival function $S(t)$, the only piece of the puzzle that we can actually estimate, has the structure of a Laplace transform

$$S(t, x) = \mathcal{L}_{g(\theta)}(\Lambda_0(t) e^{x'\beta}),$$

so that large variations in $g(\theta)$ tend to be "smoothed" out and result in small variations in $S(t, x)$.

As a result, I think that one is usually be better off using a flexible specification of the baseline hazard combined with a parametric assumption for the distribution of frailty.

4.5 The Identification Problem

One difficulty with these models is that the underlying assumption of proportionality of hazards is confounded with unobserved heterogeneity. Consider

again Figure 1. We know that the underlying hazards are proportional, but look non-proportional because we are missing x_2. But I could have generated the same hazards without any omitted variables by assuming that the baseline hazard declines over time and the effect of x_1 is non-proportional.

To further explore these issues we extend our earlier results on unobserved heterogeneity to the case where we have covariates. We start from a proportional hazards model where the conditional or subject-specific hazard is

$$\lambda(t, x, \theta) = \theta \lambda_0(t) e^{x'\beta},$$

and θ has density $g(\theta)$. To obtain the unconditional or population-average hazard we integrate out θ using the appropriate conditional density

$$\lambda(t, x) = \int_0^\infty \lambda(t, x, \theta) g(\theta | T \geq t, x) d\theta.$$

Using the proportional hazards tructure we can write this as

$$\lambda(t, x) = \lambda_0(t) e^{x'\beta} \int_0^\infty \theta g(\theta | T \geq t, x) d\theta.$$

The integral can be recognized as the expected frailty of survivors, so we have our first result:

$$\lambda(t, x) = \lambda_0(t) e^{x'\beta} E(\theta | T \geq t, x). \tag{6}$$

The form of the expectation can be worked out for specific distributions. From our earlier results, we can write

$$E(\theta | T \geq t, x) = \frac{1}{(1 + \frac{1}{k}\sigma^2 \lambda_0(t) e^{x'\beta})^k},$$

with $k = 1$ for gamma frailty and $k = \frac{1}{2}$ for inverse Gaussian frailty. If we substitute this result on the formula for the hazard and take logs we can write the model as

$$\log \lambda(t, x) = \alpha(t) + x'\beta + \gamma(t, x),$$

where $\alpha(t) = \log \lambda_0(t)$ is the log-baseline hazard, representing the main effect of duration, $x'\beta$ is the log-relative risk, representing the main effects of the covariates, and $\gamma(t, x) = \log E(\theta | T \geq t, x)$, the log of the expected frailty of survivors, representing a form of interaction between duration and the covariates.

In other words, the presence of unobserved heterogeneity in a subject-specific proportional hazards model results in a population-average model where the hazards are no longer proportional.

As a result, we conclude that unobserved heterogeneity is indeed confounded with proportionality of hazards. We can't test for one without assuming the other.

Example 1: My 1995 paper shows an example of a proportional hazards models that, combined with gamma heterogeneity, leads to declining non-proportional hazards.

But it also shows that exactly the same population hazards could have been generated from a model with inverse Gaussian heterogeneity where the individual hazards are non-proportional.

And of course, it is possible (though unlikely) that there is no unobserved heterogeneity and the individual hazards look just like the population hazards that we observe.

Example 2: Suppose you have found that the following proportional hazards model (with a constant baseline!) fits your data well:

$$\lambda(t, x) = \exp\{\alpha + x'\beta\} \tag{7}$$

Before you conclude that the hazard is indeed constant for each individual, consider the alternative subject-specific model

$$\lambda(t, x, \theta) = \theta \exp\{\alpha + x'\beta + \sigma^2 t e^{\alpha + x'\beta}\},$$

where heterogeneity has a gamma distribution with mean one and any variance σ^2 that you like. This is an accelerated life model with a Gompertz baseline. You should be able to verify that this model leads to exactly the same population-average hazard as Equation 7.

But there is more. Consider the following subject-specific model

$$\lambda(t, x, \theta) = \theta \exp\{\alpha + x'\beta\}(1 + \sigma^2 \exp\{\alpha + x'\beta\} t),$$

where θ has an inverse Gaussian distribution with mean one and variance σ^2. In this model the hazard is a linear function of t. Again, you should be able to verify that this model leads to the same population-average hazard as Equation 7.

Thus, the hazards in Equation 7 could represent

- homogeneous populations with constant risk,
- gamma frailty with Gompertz accelerated life, or

- inverse Gaussian frailty with linear risks.

The choice between these interpretations cannot be made on statistical grounds.

In our next unit we will see that these models are in fact identified when we have multiple observations, as we do in multivariate survival and event-history models.

Multivariate Survival Models

Germán Rodríguez
grodri@princeton.edu

Spring, 2001; revised Spring 2005

In this unit we study models for *multivariate* survival (in the statistical sense of many outcomes, not just many predictors).

1 Areas of Application

We start by reviewing four main areas of applications of these models.

1.1 Series of Events

One area of interest is processes where each individual may experience a succession of events. Examples include birth intervals and spells of unemployment.

Because the various events occur to the same individual, the waiting times will in general not be independent. Some couples tend to have short birth intervals while others have long ones. Observed covariates such as contraceptive use may explain some of the association. In general, however, there will remain some correlation due to unobserved individual traits.

Because the events occur one after the other, it will generally be the case that only the last interval can be censored. This introduces some simplification in estimation. In particular, it makes it possible to study the sequence using *successive conditioning*.

To fix ideas consider an example with three intervals. The joint density of T_1, T_2 and T_3

$$f_{123}(t_1, t_2, t_3)$$

can always be written as the product of the marginal of T_1, the conditional distribution of T_2 given T_1, and the conditional distribution of T_3 given T_1 and T_2:

$$f_1(t_1)\, f_{2|1}(t_2|t_1)\, f_{3|12}(t_3|t_1, t_2).$$

These two equations are an identity: any joint distribution can be factored in this way.

A typical contribution to the likelihood function, given the fact that T_3 is the only waiting time that can be censored, will look like

$$f_1(t_1)\, f_{2|1}(t_2|t_1)\, \lambda_{3|12}(t_3|t_1,t_2)^{d_3} S_{3|12}(t_3|t_1,t_2).$$

where the last term is, as usual, the conditional survival function for censored cases and the conditional density for deaths.

As long as we model the conditional distributions using different parameters, the likelihood will factor into separate components. Typically, one would model the conditional distributions by introducing the previous waiting times as covariates, for example we could write the three-equation model

$$\begin{aligned}
\lambda_1(t_1|x) &= \lambda_{01}(t_1)e^{x'\beta_1} \\
\lambda_2(t_2|t_1,x) &= \lambda_{02}(t_2)e^{x'\beta_1+s_1(t_1)} \\
\lambda_3(t_3|t_1,t_2,x) &= \lambda_{03}(t_2)e^{x'\beta_1+s_1(t_1)+s_2(t_2)}
\end{aligned}$$

where $s(t)$ denotes a smooth term on t, such as a smoothing or regression spline.

My own work on birth intervals (Rodríguez et al. 1984) used this approach. It turns out that only the previous interval turned out to be relevant, so our models were simplified. Some of the advantages of this approach are

- It is easy, because it breaks down into a series of univariate analyzes.
- It is consistent with sequential decision making, where the actual values of $t_1,\ldots,t_{j-1}$ may affect behavior influencing t_j.

On the other hand it has the disadvantage of using separate parameters for each spell.

1.2 Kindred Lifetimes

A second area where we may use multivariate survival models consists of related lifetimes, such as the survival of husband and wife, siblings, or other kin. Following Vaupel I will call these *kindred* lifetimes. In general there is reason to believe that these lifetimes are correlated, because of common unobserved characteristics of the couple (in the case of husband and wife survival) or the family (in the case of sibling survival).

An important feature of kindred lifetimes is that any (or all) of the waiting times may be censored. With three children, for example, you may observe 8 different patterns of censoring. This means that we cannot adopt the simple sequential approach outlined earlier for series of events, as we will often lack the information needed. For example we can't very well model T_2 as a function of T_1 when T_1 is censored, at least not in the simple way we have described. Thus, we need a more general approach.

1.3 Competing Risks

We have already encountered a third type of multivariate data in our discussion of competing risks, where $T_1, T_2, \ldots, T_k$ represent *latent* survival times to different causes of death.

As noted earlier, estimation of these models is complicated by the fact that we only observe

$$T = \min\{T_1, \ldots, T_k\}$$

and even this can be censored. Keep in mind, however, that the models that follow could be used, at least conceptually, in this context.

1.4 Event History Models

The fourth and final type of multivariate data involves transitions among several types of states. This combines elements of competing risk models with models for series of events.

Consider for example the analysis of nuptiality. You start in the single state. From there you can move to cohabiting or married. From cohabiting you can move to married or to separated. And so on. If you distinguish separations from marriage or cohabiting as well as widowhood and divorce, you probably have about 15 possible transitions of interest.

Analysts often study one type of transition, for example age at first marriage or marriage dissolution. With event history data, however, one may study the complete process, allowing for inter-dependencies among the different kinds of transitions. The nature of the data allows conditioning each move on the entire history of previous moves.

A closely related subject in demography are multi-state models. A lot of work in that area assumes a homogeneous population with constant transition rates and independent moves, and emphasizes analytic results, such as the steady-state proportion in each state. In some ways event history models are to multi-state models what Cox regression models are to the traditional life table.

2 Bivariate Survival Models

We consider first the case of only two survival times, T_1 and T_2. This section follows Cox and Oakes (1984, Chapter 10) and Guo and Rodríguez (1992).

2.1 Basic Definitions

Interest will focus on the *joint* survival

$$S_{12}(t_1, t_2) = \Pr\{T_1 \geq t_1, T_2 \geq t_2\}.$$

Note that $S_{12}(t, t)$ is the probability that both units are alive at t.

We also have the *marginal* survival function

$$S_1(t_1) = \Pr\{T_1 \geq t_1\} = S_{12}(t_1, 0),$$

and similarly for t_2. If T_1 and T_2 were independent the joint survival function would be the product of the marginals.

We might also be interested in the *conditional* survival function, which has two variants

$$S_{1|2}(t_1|T_2 = t_2) = \Pr\{T_1 \geq t_1|T_2 = t_2\},$$

giving the survival probabilities given that the other unit failed at time t_2, and

$$S_{1|2}(t_1|T_2 \geq t_2) = \Pr\{T_1 \geq t_1|T_2 \geq t_2\},$$

given that the other unit survived to just before time t_2.

Associated with each of these survival functions there will be a cumulative hazard function, which can be obtained by taking minus the log of the survival function. There will also be a hazard function, which can be obtained by taking derivatives of the cumulative hazard.

Specifically, we can define the *joint* hazard function as

$$\lambda_{12}(t_1, t_2) = \lim_{dt \to 0} \Pr\{T_1 \in [t_1, t_1 + dt), T_2 \in [t_2, t_2 + dt)|T_1 \geq t_1, T_2 \geq t_2\}/dt^2.$$

This is the instantaneous risk that one unit fails at t_1 and the other fails at t_2 given that they were alive just before t_1 and t_2, respectively.

We can also define a *marginal* hazard,

$$\lambda_1(t_1) = \lim_{dt \to 0} \Pr\{T_1 \in [t_1, t_1 + dt)|T_1 \geq t_1\}/dt.$$

Under independence, the joint hazard is the sum of the marginal hazards. Can you prove this result?

We can also define *conditional* hazards

$$\lambda_{1|2}(t_1|T_2 = t_2) = \lim_{dt \to 0} \Pr\{T_1 \in [t_1, t_1 + dt) | T_1 \geq t_1, T_2 = t_2\}/dt$$

which tracks the risk for one unit given that the other failed at t_2, and

$$\lambda_{1|2}(t_1|T_2 \geq t_2) = \lim_{dt \to 0} \Pr\{T_1 \in [t_1, t_1 + dt) | T_1 \geq t_1, T_2 \geq t_2\}/dt$$

given that the other unit survived to just before t_2.

The cause-specific hazard considered in the context of competing risks (tracking the risk of death due to cause j among survivors to time t) is a special case of the latter, namely the case where $t_1 = t_2$:

$$\lambda_{1|2}(t|T_2 \geq t) = \lim_{dt \to 0} \Pr\{T_1 \in [t, t + dt) | T_1 \geq t, T_2 \geq t\}/dt.$$

Knowledge of the two types of conditional hazards completely determines a joint distribution, see Cox and Oakes (1984, p. 157) for an expression linking the joint density to the conditional hazards.

2.2 Frailty Models

One way to model a joint survival function is to assume the existence of a random effect θ such that given θ, T_1 and T_2 are independent. Depending on the context, θ may represent traits that persist across spells or are common among kin, and which account for the lack of independence.

In symbols, we can write the assumption of conditional independence as

$$S_{12}(t_1, t_2|\theta) = S_1(t_1|\theta) S_2(t_2|\theta),$$

where all survival functions are conditional on θ. Usually the random effect is assumed to act multiplicatively on the hazard, so that

$$S_i(t_i|\theta) = S_{0i}(t_i)^\theta$$

for some baseline survival function $S_{0i}(t)$. Under this assumption the cumulative hazards are

$$\Lambda_i(t_i) = \theta \Lambda_{0i}(t_i)$$

and the individual hazards are

$$\lambda_i(t_i) = \theta \lambda_{0i}(t_i).$$

The *conditional* joint survival function is then

$$\begin{aligned} S_{12}(t_1, t_2|\theta) &= S_{01}(t_1)^\theta S_{02}(t_2)^\theta \\ &= e^{-\theta\Lambda_{01}(t_1)} e^{-\theta\Lambda_{02}(t_2)} \\ &= e^{-\theta(\Lambda_{01}(t_1)+\Lambda_{02}(t_2))}. \end{aligned}$$

This is not very useful because θ is not observed. To obtain the *unconditional* survival function we need to 'integrate out' θ. Suppose that θ has density $g(\theta)$. Then

$$\begin{aligned} S_{12}(t_1, t_2) &= \int_0^\infty S_{12}(t_1, t_2|\theta) g(\theta) d\theta \\ &= \int_0^\infty e^{-\theta(\Lambda_{01}(t_1)+\Lambda_{02}(t_2))} g(\theta) d\theta, \end{aligned}$$

and we recognize this expression as the Laplace transform of $g(\theta)$ evaluated at $s = \Lambda_{01}(t_1) + \Lambda_{02}(t_2)$. Thus

$$S_{12}(t_1, t_2) = \mathcal{L}_g(\Lambda_{01}(t_1) + \Lambda_{02}(t_2)).$$

To make further progress we need to know the distribution of θ.

2.3 Gamma Frailty

Suppose the common or persistent frailty component θ has a gamma distribution with parameters $\alpha = \beta = 1/\sigma^2$. The Laplace transform of the gamma density is $\mathcal{L}(s) = (\beta/(\beta + s))^\alpha$. Using this result,

$$S_{12}(t_1, t_2) = \left(\frac{1}{1 + \sigma^2\Lambda_{01}(t_1) + \sigma^2\Lambda_{02}(t_2)} \right)^{\frac{1}{\sigma^2}}. \tag{1}$$

Actual estimation of this model requires some assumption about the baseline hazards and will be considered in detail further below.

While we are on this subject, it will be useful to write the joint survival function $S_{12}(t_1, t_2)$ under gamma frailty as a function of the marginals $S_i(t_i)$.

We start from the expression for the joint survival and obtain a marginal by setting one of the t's to zero, thus

$$S_1(t_1) = S_{12}(t_1, 0) = \left(\frac{1}{1 + \sigma^2\Lambda_{01}(t_1)} \right)^{\frac{1}{\sigma^2}}.$$

We now use this expression to solve for $\Lambda_{01}(t_1)$, or better still $\sigma^2\Lambda_{01}(t_1)$:

$$S_1(t_1)^{\sigma^2} = \frac{1}{1 + \sigma^2\Lambda_{01}(t_1)},$$

taking reciprocals and subtracting one on both sides gives

$$S_1(t_1)^{-\sigma^2} - 1 = \sigma^2 \Lambda_{01}(t_1),$$

and using this result on Equation 1 we obtain

$$S_{12}(t_1, t_2) = (S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} - 1)^{-\frac{1}{\sigma^2}}. \tag{2}$$

Keep this handy for future reference.

2.4 Non-parametric Frailty

An alternative assumption regarding θ is to treat it as discrete, assuming values $\theta_1, \theta_2, \ldots, \theta_k$ with probabilities $\pi_1, \pi_2, \ldots, \pi_k$, where $\sum \pi_j = 1$.

Laird (1978) and Heckman and Singer (1982, 1984) show that a non-parametric maximum likelihood approach to the estimation of $g(\theta)$ leads precisely to this discrete model. Under the foregoing assumptions, the unconditional survival function is the finite mixture

$$S_{12}(t_1, t_2) = \sum_{j=1}^{k} e^{-\theta_j(\Lambda_{01}(t_1)+\Lambda_{02}(t_2))}\pi_j.$$

Again, estimation of this model requires specifying the baseline hazards $\lambda_{0i}(t)$, and will be considered below.

Other distributional assumptions are possible. Are these models identified? Fortunately yes, as we shall see presently.

2.5 Clayton's Model

Clayton (1978) proposed a continuous bivariate survival model where the two conditional hazards for T_1 given $T_2 = t_2$ and given $T_2 \geq t_2$ are proportional, namely

$$\frac{\lambda_1(t_1|T_2 = t_2)}{\lambda_1(t_1|T_2 \geq t_2)} = 1 + \phi. \tag{3}$$

In words, the risk for unit one at time t_1 given that the other unit failed at time t_2 is $1 + \phi$ times the risk for unit one at time t_1 given that the other unit survived to t_2.

Think of all families with two children whose second child survived to age one, say. Separate those families whose second child died shortly after his or her first birthday. Clearly these families have, on the average, higher risk that the original pool. In fact, the first child in these families is subject to $100\phi\%$ higher risk, at any given age.

Note that the hazard ratio $1 + \phi$ is constant over time. Conditioning on survival (and then death) to age two (instead of age one) in our example would lead to exactly the same hazard ratio.

The remarkable thing about this model is that it is exactly equivalent to a multiplicative frailty model with a gamma-distributed random effect, with $\phi = \sigma^2$.

An important implication of this result is that the model is clearly identified and can be tested, as it has an observable consequence, namely the fact that the ratio of two hazards (which are themselves estimable) is constant over time (something we can verify).

Moreover, this result gives a new interpretation to σ^2, the variance of the random effect. A variance of σ^2 means that children who lost a sibling at age t have a risk $(1 + \sigma^2)$ times the risk of children who had a sibling survive to age t.

Note back on Equation 2 that as $\sigma^2 \to 0$, $S_{12}(t_1, t_2)$ approaches the product of the marginals, as we would expect under independence. As $\sigma^2 \to \infty$, $S_{12}(t_1, t_2)$ approaches $\min\{S_1(t_1), S_2(t_2)\}$, which is known as the Fréchet bound on the maximum possible positive association between two distributions with given marginals.

In other words, the model covers the entire spectrum from independence to maximum possible *positive* association. However, the model cannot account for negative association.

2.6 Oakes's Interpretation

Oakes (1982) showed that ϕ (or σ^2) is closely related to a measure of ordinal association known as Kendall's τ (tau).

Given a bivariate sample $(T_{11}, T_{12}), (T_{21}, T_{22}), \ldots, (T_{n1}, T_{n2})$, Kendall considers all possible pairs of observations. He calls a pair concordant if the first coordinates have the same rank order as the second coordinates. Otherwise a pair is discordant. (For example in husband and wife survival two couples would be concordant if either husband A dies younger than husband B *and* wife A dies younger than wife B, or the A's outlive the corresponding B's.) Kendall's τ is then defined as

$$\tau = \frac{\text{concordant pairs} - \text{discordant pairs}}{\text{number of pairs}}.$$

Unfortunately, when the data are censored we may be unable to calculate this measure. (For example if husband A died younger than husband B, wife

A died, and wife B is still alive but has not yet reached the age at which wife A died, we would not know if the pair is concordant or discordant.)

Oakes has shown that if we restrict the calculation to pairs that can definitely be classified as either concordant or discordant (for example wife B is censored but has already outlived wife A), then the expected value of Kendall's τ under Clayton's model is

$$E(\hat{\tau}) = \frac{\phi}{\phi + 2},$$

which provides further justification for interpreting ϕ (or σ^2) as a measure of ordinal association between kindred lifetimes.

There are no known similar interpretations for frailty distributions other than the gamma. It would be interesting to explore how the ratio of hazards varies over time for other distributions, such as the inverse Gaussian.

3 Multivariate Extensions

The foregoing ideas extend easily to more than two lifetimes and models with observed covariates.

3.1 Notation and Definitions

Consider a set of *clustered* data where

$$\begin{aligned} t_{ij} &= \text{observation time} \\ d_{ij} &= \text{death indicator} \\ x_{ij} &= \text{vector of covariates} \end{aligned}$$

all for the j-th individual in the i-th group (or cluster).

We assume that given x_{ij} and a random effect θ_i the m_i lifetimes in cluster i, say $T_{i1}, T_{i2}, \ldots, T_{im_i}$ are independent.

Thus, the joint distribution of these lifetimes given θ_i is the product of the marginal distributions given θ_i. Under the multiplicative frailty model, the marginal hazards satisfy

$$\lambda_{ij}(t_{ij}|x_{ij}, \theta_i) = \theta_i \lambda_{0ij}(t_{ij}|x_{ij}).$$

Often the covariate effects will be modelled using a proportional hazards model, so that

$$\lambda_{0ij}(t_{ij}|x_{ij}) = \lambda_0(t_{ij}) e^{x'_{ij}\beta}.$$

Combining these two equations we obtain the full model:

$$\lambda_{ij}(t_{ij}|x_{ij},\theta_i) = \theta_i \lambda_0(t_{ij}) e^{x'_{ij}\beta}. \tag{4}$$

In the sections that follow we will often use t_i to denote the vector $(t_{i1}, t_{i2}, \ldots, t_{im_i})'$ of m_i survival times for cluster i, and X_i to denote an m_i by p matrix of covariates with one row for the covariates of each unit in the cluster.

3.2 Gamma Frailty

If the cluster-specific random effects θ_i have independent gamma distributions, then the unconditional survival for the m_i lifetimes in cluster i is

$$S_i(t_i, X_i) = \int_0^\infty \prod_j S_{ij}(t_{ij}|x_{ij},\theta_i) g(\theta_i) d\theta_i,$$

which can be solved easily using the Laplace transform, to give

$$S(t_i, X_i) = \left(\frac{1}{1 + \sigma^2 \Lambda_0(t_{i1}) e^{x'_{i1}\beta} + \ldots + \sigma^2 \Lambda_0(t_{im_i}) e^{x'_{im_i}\beta}} \right)^{\frac{1}{\sigma^2}}.$$

Alternatively, we can write the joint distribution as a function of the marginals. In a direct extension of our earlier result for bivariate distributions,

$$S_i(t_i|X_i) = (S_{i1}(t_{i1}|x_{i1})^{-\sigma^2} + \ldots + S_{im_i}(t_{im_i}|x_{im_i})^{-\sigma^2} - m_i + 1)^{-\frac{1}{\sigma^2}}.$$

Under a proportional hazards model

$$S_{ij}(t_{ij}|x_{ij}) = S_0(t_{ij})^{e^{X'_{ij}\beta}}.$$

3.3 Clayton's Model

Clayton's characterization extends to the multivariate case. We consider the risk for a given individual given the survival status of the others to any given set of ages. Specifically, consider the risk to individual one given that the others have survived to (if $d_{ij} = 0$) or died at (if $d_{ij} = 1$) ages t_{ij}. We will write this conditional hazard as

$$\lambda_1(t_1|t_2, t_3, \ldots, t_m; d_2, d_3, \ldots, d_m),$$

where I have suppressed the group subscript for clarity.

Consider now a similar hazard given that all other members of the group survived to the same ages, which in our notation is

$$\lambda_1(t_1|t_2, t_3, \ldots, t_m; 0, 0, \ldots, 0).$$

Then under a multiplicative frailty model the ratio of these two hazards is constant over time and equals

$$1 + \sigma^2 \sum_j d_j.$$

In the bivariate case this reduces to $1 + \sigma^2$ when the other member of the pair died. In a trivariate case, the risk for the index individual would be $1 + \sigma^2$ if one of the other two died and $1 + 2\sigma^2$ if both of them died.

To clarify this interpretation think of all families with 3 children where the second child survives to age one and the third child survives to age two. These families are subject to the baseline risk. Now consider the subset where the second child died around age one but the third child was alive at age two. These families have higher risk, and their risk is $1 + \sigma^2$ times the baseline. There is another subset where the second child was alive at age one but the third child died around age two. These families also have higher risk and the relative risk factor is also $1 + \sigma^2$. Finally, we have the families where the second child died around age one and the third child died around age two. These families have the highest risk; their relative risk factor is $1 + 2\sigma^2$.

Note that in this model the death of a child is not assumed to have a direct effect on the survival of the remaining siblings. Rather, the death of a child is an indicator that the family has higher than average risk.

3.4 Oakes's Interpretation

Clearly, $\sigma^2/(2 + \sigma^2)$ can still be interpreted as a measure of association between the lifetimes of any two members of the group. As in all models with a single random effect to account for the correlation of three or more r.v.'s, the association between any two members of the group is the same.

Note that this assumption may not always be reasonable. For example kids born closer together in time may face more similar risks than those born farther apart. The model allows independence and maximum positive correlation, but restricts intermediate cases to equal pairwise association.

4 Estimation Using the EM Algorithm

The fact that a multiplicative frailty model would be a standard proportional hazards model–and therefore relatively simple to estimate–if θ_i was observed, suggest immediately the possibility of using the EM algorithm. We now show that this leads to very simple procedures for gamma and for non-parametric frailty. This section follows closely Guo and Rodríguez (1992).

4.1 Gamma Frailty

If θ_i was observed, the likelihood function would depend on the joint distribution of frailty and the survival times T_{ij}. We can write this in terms of the density of θ_i times the conditional distribution of the survival time T_{ij} given θ_i. The contribution of the i-th cluster to the *complete* data log-likelihood would be

$$\log L_i = \log g(\theta_i) + \sum_{j=1}^{m_i} \{d_{ij} \log(\theta_i \lambda_{ij}(t_{ij})) - \theta_i \Lambda_{ij}(t_{ij})\}. \tag{5}$$

The hazard $\lambda_{ij}(t_{ij})$ and cumulative hazard $\Lambda_{ij}(t_{ij})$ will in general depend on the covariates x_{ij} and the parameters β as well as the baseline hazard. I am leaving that implicit to focus on the key aspects of the estimation procedure.

The E-step of the algorithm requires finding the expected value of the complete data log-likelihood, where expectation is taken with respect to the conditional distribution of θ_i given the data. Given the structure of $\log L_i$, this reduces to finding the expected value of θ_i and $\log \theta_i$ given (t_{ij}, d_{ij}) for $j = 1, \ldots, m_i$.

It turns out that this is not hard at all. Direct integration (like we did in the univariate case) shows that if the marginal (or prior) distribution of θ_i is gamma with parameters α and β (usually $\alpha = \beta = 1/\sigma^2$), then the conditional (or posterior) distribution of θ_i given the survival experience of the i-th cluster is also gamma, but with parameters

$$\alpha^* = \alpha + \sum_j d_{ij} \qquad \text{and} \qquad \beta^* = \beta + \sum_j \Lambda_{ij}(t_{ij}).$$

Note that α increases by the total number of deaths and β increases by the total cumulative hazard (or exposure to risk).

The expected value of θ_i given the data is then

$$\mu_i = E(\theta_i) = \frac{\alpha^*}{\beta^*} = \frac{\alpha + \sum_j d_{ij}}{\beta + \sum_j \Lambda_{ij}(t_{ij})}.$$

We could rewrite this expression in terms of σ^2, the variance of frailty, but it turns out to be easier to work with $\alpha(=\beta)$, which may be interpreted as a precision parameter.

The expected value of $\log\theta_i$ when θ_i has a gamma distribution is well known, and in this case turns out to be:

$$\xi_i = E(\log\theta_i) = \Psi(\alpha^*) - log\beta^* = \Psi(\alpha + \sum d_{ij}) - \log(\beta + \sum \Lambda_{ij}(t_{ij})),$$

where Ψ is the digamma function (the first derivative of the log of the gamma function, so $\Psi(x) = \Gamma'(x)/\Gamma(x)$).

Let $\hat{\mu}_i$ and $\hat{\xi}_i$ denote the expected values of θ_i and $\log\theta_i$ evaluated at current parameter estimates. Then the result of the E-step is

$$\begin{aligned} Q_i = \quad & (\alpha-1)\hat{\xi}_i - \alpha\hat{\mu}_i + \alpha\log\alpha - \log\Gamma(\alpha) + \\ & d_i\hat{\xi}_i + \textstyle\sum_j d_{ij}\log\lambda_{ij}(t_{ij}) - \hat{\mu}_i\sum\Lambda_{ij}(t_{ij}). \end{aligned} \tag{6}$$

The first line comes from the density of θ_i and the second line comes from the conditional survival likelihood.

The M-step requires maximizing Q_i w.r.t. $\alpha(=1/\sigma^2)$ and the parameters in $\lambda_{ij}(t_{ij})$. This step breaks neatly into two separate problems.

The part of $Q = \sum Q_i$ involving α is

$$Q_1 = (\alpha-1)\sum\hat{\xi}_i - \alpha\sum\hat{\mu}_i + n\alpha\log\alpha - n\log\Gamma(\alpha),$$

where n is the number of clusters. The first derivative is

$$\frac{\partial Q_1}{\partial\alpha} = \sum(\hat{\xi}_i - \hat{\mu}_i) + n(1 + \log\alpha - \Psi(\alpha)),$$

and the second derivative is

$$\frac{\partial^2 Q_1}{\partial\alpha^2} = n(\frac{1}{\alpha} - \Psi^{(1)}(\alpha)),$$

where $\Psi^{(1)}$ is the trigamma function. This part can be maximized using a Newton-Raphson algorithm. Calculation of the gamma, digamma and trigamma functions can be accomplished using published algorithms. (All three functions are available in R.)

The part of $Q = \sum Q_i$ involving the remaining parameters is exactly the same as the log-likelihood for a standard survival model where $\hat{\mu}_i$ is treated as an extra relative risk. To see this point note that we can add to Q_i the quantity $\sum d_{ij}\log\hat{\mu}_i$, which does not depend on unknown parameters. Then this part becomes

$$Q_2 = \sum_i\sum_j\{d_{ij}\log(\hat{\mu}_i\lambda_{ij}(t_{ij})) - \hat{\mu}_i\Lambda_{ij}(t_{ij})\},$$

which is a standard survival log-likelihood. For example, if we were using a proportional hazards model with a piece-wise constant baseline hazard, Q_2 would be equivalent to a Poisson log-likelihood.

To summarize, the EM algorithm for this problem involves the following steps. Given initial estimates (obtained, for example, by ignoring the multivariate structure of the data):

1 Estimate the expected value of the random effect θ_i and of its logarithm $\log \theta_i$ for each cluster. Call these $\hat{\mu}_i$ and $\hat{\xi}_i$.

2 Obtain new estimates of the parameters by (a) fitting the model using standard univariate procedures but including $\hat{\mu}_i$ as a known relative risk and (b) solving the Newton-Raphson equations for α.

These steps are repeated to convergence. The algorithm is slow, but extremely robust. It is also comparatively easy to implement, because you can take advantage of existing code for the univariate model.

4.2 Non-parametric Frailty

The EM algorithm for non-parametric frailty is even simpler. We assume that a cluster comes from one of K populations representing different levels of frailty $\theta_1, \ldots, \theta_K$. Let π_k denote the probability that the cluster comes from the k-th population (or level of frailty), with $\sum \pi_k = 1$.

We introduce an indicator variable Z_{ik} that takes the value one if the i-th cluster comes from the k-th population (or level of frailty) and zero otherwise. Note that $\Pr\{Z_{ik} = 1\} = \pi_k$. If the Z_{ik} were observed we would maximize the complete data log-likelihood, to which the i-th cluster contributes

$$\log L_i = \sum_{k=1}^{K} \{z_{ik}(\log \pi_k + \log L_{ik})\},$$

where $\log L_{ik}$ denotes the standard log-likelihood given that the level of frailty is θ_k, namely

$$\log L_{ik} = \sum_{j=1}^{n_i} \{d_{ij} \log(\theta_k \lambda_{ij}(t_{ij})) - \theta_k \Lambda_{ij}(t_{ij})\},$$

which of course would depend on some parameters, say β. (Note that for each cluster only one of the K terms in $\log L_i$ is non-zero.)

The E-step requires taking the expected value of $\log L_i$ given the data, which in turn requires the expected value of the indicator variable Z_{ik}. A

fairly straightforward argument shows that if the prior probability that $Z_{ik} = 1$ is π_k, and given this the likelihood of the data is L_{ik}, then the posterior probability that $Z_{ik} = 1$ given the data is

$$\rho_{ik} = E(Z_{ik}|(t_i, d_i)) = \frac{\pi_k L_{ik}}{\sum_{r=1}^{K} \pi_r L_{ir}}.$$

This follow from Bayes theorem or the definition of conditional probabilities. You just have to be careful that for some cases we are talking about the probability of dying at t_{ij} while for others (censored) we need the probability of being alive just before t_{ij}. Note that ρ_{jk} represents the posterior probability that a cluster comes from the k-th population (or level of frailty). Let $\hat{\rho}_{ik}$ denote this posterior probability evaluated at current parameter estimates. The result of the E-step is then

$$Q_i = \sum_{k=1}^{K} \hat{\rho}_{ik} \log \pi_k + \sum_{k=1}^{K} \hat{\rho}_{ik} \log L_{ik}.$$

Note that the second term is just a weighted average of the log-likelihoods given that frailty has value θ_k, with weights given by the posterior probabilities that frailty has value θ_k.

The M-step requires maximizing $Q = \sum Q_i$ w.r.t. the π_k and the parameters in $\log L_{ik}$, namely the θ_k and β. Again, this problem breaks down neatly into two separate problems.

First, maximizing w.r.t. the π_k's gives the explicit solution

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{\rho}_{ik},$$

or the average of the posterior probabilities. This follows directly from the multinomial structure of Q_i. (You may take derivatives to verify this result, but remember the restriction $\sum \pi_k = 1$. The easiest thing to do is work with only $k-1$ probabilities and write $\pi_K = 1 \sum_{i=1}^{K-1} \pi_k$.)

Second, maximizing w.r.t. the θ_k's and β is equivalent to maximizing the standard univariate log-likelihood $\log L_{ik}$ with a twist: each observation contributes to each possible level of frailty with weight equal to its posterior probability of coming from that population.

To fix ideas suppose you are fitting a model with two levels of frailty (or two points of support). Then all you have to do is duplicate all observations, introduce a factor coded 1 for the first copy and 2 for the second (this will give θ_1 and θ_2, give weight $\hat{\rho}_{i1}$ to the first copy and $\hat{\rho}_{i2} = 1 - \hat{\rho}_{i1}$ to the second, and fit as usual. Isn't that easy?

These two results apply quite generally to finite mixture models, not just frailty models; for details see the book by Everitt and Hand (1981).

Note: With gamma frailty we assumed $E(\theta_i) = 1$. Introducing a similar restriction in the present context would impose a constraint on the θ_k's and π_k's and would make life more difficult. A much simpler solution is to leave the θ_k's unrestricted and omit a constant from the baseline hazard. After the model is fit, you can calculate the mean frailty as

$$\bar{\theta} = \sum_{k=1}^{K} \hat{\pi}_k \hat{\theta}_k,$$

and then absorb this into the constant. In practice we divide $\hat{\theta}_k$ by $\bar{\theta}$ to make the new mean one and obtain results comparable with gamma frailty.

4.3 Further Notes

4.3.1 Non-Parametric Hazards

All the procedures discussed so far require a parametric model for the baseline hazard, be it exponential, Weibull, log-logistic, or piece-wise exponential survival.

Clearly it would be nice to have a robust method that, like Cox's partial likelihood, could be used to estimate the main parameters of interest, i.e. β and σ^2, without any assumptions about the form of the hazard.

Clayton and Cuzick (1985) have some interesting results along these lines in a procedure that appears to be closely related to the EM algorithm. I recommend their excellent paper and the ensuing discussion.

4.3.2 Baseline Hazards

In our discussion we allowed a different baseline hazard for each member of a cluster, and even a different vector of coefficients β. This makes sense in the study of series of events, such as birth intervals. In other cases, such as studies of siblings, particularly where the events are not distinguishable, it will probably suffice to have a common baseline and common coefficients.

4.3.3 Accelerating the Algorithm

The results of Louis (1984), described in the technical note on the EM algorithm, can be used to

- speed-up the algorithm, and

- obtain standard errors.

See my paper with Guo for details.

4.3.4 The Incomplete Data log-Likelihood

The EM algorithm is simple and stable, and sometimes it is the only way to proceed, particularly if the correct likelihood is hard to obtain or would require numerical integration.

That is *not* the case here. Both with gamma heterogeneity and a finite mixture model the correct incomplete data log-likelihood is tractable. We have already given expressions for the survival function; the hazard follows easily.

Moreover, the first and second derivatives w.r.t. the parameters can be obtained, opening the possibility of using a Newton-Raphson algorithm. This however, requires good starting values. The accelerated EM algorithm is probably more stable and just as fast.

5 Fixed-Effects Models

In the discussion so far we have treated the cluster-specific effect θ_i as *random*: we postulate a distribution and then estimate the parameters of that distribution.

An alternative approach is to treat the θ_i as *fixed* quantities to be estimated, effectively adding one parameter for each cluster.

One difficulty with this approach is that when the number of parameters to be estimated increases with the number of observations, we generally get *inconsistent* estimates, not only for the offending parameters, but also for the other parameters in the model.

There is, however, a way around these difficulties. It is often possible to eliminate the fixed effects θ_i from the likelihood by suitable *conditioning*. Usually one conditions on a statistic (a function of the data) that is minimal sufficient for θ_i (meaning that the likelihood viewed as a function of θ_i depends on the data only through this statistic, which has the same dimensionality as $theta_i$).

In the present context one can construct a *partial likelihood* that eliminates the θ_i. This approach is discussed by Kalbfleisch and Prentice (1980) under the rubric of stratification, and has been advocated among demographers and economists in a series of papers by Ridder and Tunali.

Here are the basis ideas. We assume a multivariate proportional hazards model where the risk to subject j in cluster i is

$$\lambda_{ij}(t_{ij}) = \theta_i \lambda_0(t_{ij}) e^{x'_{ij}\beta},$$

where θ_i is a fixed cluster effect, $\lambda_0(t_{ij}($ is a baseline hazard and $e^{x'_{ij}\beta}$ is a relative risk, as usual.

Next we construct a partial likelihood separately in each cluster. Let $t_{i1} < \ldots < t_{im_i}$ denote the distinct times of death observed in cluster i. Assume no ties, so only one person dies at each t_{ij}, and let R_{ij} denote the risk set in cluster i just before time t_{ij}. The partial likelihood is

$$L_= \prod_{j=1}^{m_i} \frac{\theta_i \lambda_0(t_{ij}) e^{x'_{ij}\beta}}{\sum_{k \in R_{ij}} \theta_i \lambda_0(t_{ij}) e^{x'_{ik}\beta}},$$

and as you may see, this time it is not just the baseline hazard $\lambda_0(t_{ij})$ but also the cluster-specific effect θ_i that cancels out of the likelihood. (In fact, we don't even need to assume the same baseline hazard for each cluster.)

The overall partial likelihood is obtained as the product of the cluster-specific partial likelihoods over all clusters. Ties can be handled using the standard approximations, such as Peto's or Efron's.

Some important points to note about this procedure:

- Clusters with no deaths do not contribute to the likelihood

This is an unfortunate consequence of the fact that θ_i is a fixed unknown parameter. If there are no deaths, it is conceivable that θ_i is zero. This doesn't happen in random-effects models with continuous frailty because the θ_i are supposed to come from a distribution that puts no mass at zero.

- Covariates that are constant within a cluster also drop out from the likelihood

This point is very important. In a study of child mortality, we cannot estimate the effect of mother's education if we use a family-level fixed effect. The reason is that the term $e^{\beta x_i}$ would appear both in the numerator and denominator of the cluster-specific partial likelihood and would therefore cancel out. Another way to think about this is to note that θ_i captures all influences common to members of a cluster, both observed and unobserved. So you can only estimate the effects of child-level covariates.

Moreover, if a covariate happens to have the same value for all children in a family it would also drop out of the likelihood. Imagine a family with

three girls. This family cannot contribute to estimating the effect of sex on mortality. You might think that this family would contribute to estimating the mortality of girls while other families contribute to estimating the mortality of boys, The problem is that the differences between these families could be due to their fixed effects, and have nothing to do with the sex of their children.

Here, then, lies the main advantage and disadvantage of the technique. One often ends up using only a small fraction of the original data, raising the specter that the cases selected for analysis are very different from the rest. On the other hand one may argue that the are precisely the cases that contain information. Only by looking at children within a family who differ on a trait, and such that one dies and the other doesn't, can we be sure that the apparent effect of the trait is not due to unobserved family characteristics.

Fixed-effects models control for both observed and unobserved cluster characteristics; they solve the omitted variables problem at this level, but cannot estimate the effects of included variables. Random-effects models address the problem of intra-cluster correlation, but can only capture the effects of unobserved cluster characteristics that are uncorrelated with observed covariates. They offer no solution to the omitted variables problem, but can estimate the effects of observed variables at all levels.

Competing Risks

Germán Rodríguez
grodri@princeton.edu

Spring, 2001; revised Spring 2005

In this unit we consider the analysis of multiple causes of failure in the framework of competing risk models. An excellent reference on this material is Chapter 8 in Kalbfleisch and Prentice (2002), or Chapter 7 in the 1980 edition.

1 Introduction and Notation

Consider an example involving multiple causes of failure. Women who start using an intrauterine device (IUD) are subject to several risks, including accidental pregnancy, expulsion of the device, removal for medical reasons and removal for personal reasons. K-P discuss three areas of interest in the analysis of competing risks such as IUD discontinuation:

1 Studying the relationship between a vector of covariates x and the rate of occurrence of specific types of failure; for example the covariates of IUD expulsion.

2 Analyzing whether people at high risk of one type of failure are also at high risk for others, even after controlling for covariates; for example are women who are at high-risk of expelling an IUD also at high risk of accidental pregnancy while wearing the device?

3 Estimating the risk of one type of failure after removing others; for example how long would we expect women to use an IUD if we could eliminate the risk of expulsion?

It turns out that we can answer the first of these questions, but the other two are essentially intractable. The third question can be answered under the strong assumption that the competing risks are independent, which essentially assumes away the second question.

We start by introducing some notation. Let T be a continuous r.v. representing survival time. We assume that when failure occurs it may be one of m distinct types indexed by $j \in \{1, 2, \ldots, m\}$, and we let J be a r.v. representing the type of failure. Also, we let x be a vector of covariates.

1.1 Cause-Specific Hazards

We define the overall hazard rate as usual:

$$\lambda(t, x) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt | T \geq t, x\}}{dt}.$$

We will also define a *cause-specific hazard rate*, representing the instantaneous risk of dying of cause j:

$$\lambda_j(t, x) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt, J = j | T \geq t, x\}}{dt}.$$

In words, we calculate the conditional probability that a subject with covariates x dies in the interval $[t, t + dt)$ *and* the cause of death is the j-th cause, given that the subject was alive just before time t. We turn the probability into a rate dividing by dt and then take the limit as $dt \to 0$.

By the law of total probability, we have

$$\lambda(t, x) = \sum_{j=1}^{m} \lambda_j(t, x),$$

because failure must be due to one (and only one) of the m causes. If two types of failure can occur simultaneously we define the combination of the two as a new type of failure, so we can maintain this assumption.

1.2 Integrated Hazard and Survival

The overall survival function can be defined as usual:

$$S(t, x) = e^{-\Lambda(t,x)},$$

where $\Lambda(t, x)$ is the cumulative risk obtained by integrating the overall hazard

$$\Lambda(t, x) = \int_0^t \lambda(u, x) du.$$

We have assumed that the covariates are fixed to keep the notation simple. Extension to time-varying covariates is fairly straightforward, but calculation of the survival function requires specifying the trajectory of time-varying covariates.

The function $S(t,x)$ has a clear meaning as the probability of surviving *all* types of failure up to time t.

We will also define, by analogy with $S(t,x)$, the function

$$S_j(t,x) = e^{-\Lambda_j(t,x)},$$

where $\Lambda_j(t,x)$ is the integrated or cumulative hazard for case j;

$$\Lambda_j(t,x) = \int_0^t \lambda_j(u,x)du.$$

Note, however, that

Note 1 *$S_j(t,x)$ will not, in general, have a survivor function interpretation if $m > 1$.*

1.3 Cause-Specific Densities

We can also define a *cause-specific density* of failures at time t, say

$$\begin{aligned} f_j(t,x) &= \lim_{dt\to 0} \frac{\Pr\{t \le T < t+dt, J = j|x\}}{dt} \\ &= \lambda_j(t,x)S(t,x). \end{aligned}$$

This density represents the unconditional risk that a subject dies at time t of cause j. By the law of total probability, the overall density of deaths at time t is

$$f(t,x) = \sum_{i=1}^{m} f_j(t,x).$$

2 Estimation: One Sample

Consider first the homogeneous case with no covariates.

2.1 Kaplan-Meier

The Kaplan-Meier estimator can easily be generalized to include competing risks. Let

$$t_{j1} < t_{j2} < \ldots < t_{jk_j}$$

denote the k_j distinct failure times for failures of type j. Let n_{ji} denote the number of subjects at risk just before t_{ji} and let d_{ji} denote the number of

deaths due to cause j at time t_{ji}. Then the same arguments used to derive the usual K-M estimator lead to

$$\hat{S}_j(t) = \prod_{i:t_{ji}<t} \left(1 - \frac{d_{ji}}{n_{ji}}\right).$$

It is interesting to note that $\hat{S}(t)$ is exactly the same as the standard K-M estimator that one would obtain if all failures of type other than j were treated as censored cases.

If there are no ties between different types of failure, then

$$\hat{S}(t) = \prod_{j=1}^{m} \hat{S}_j(t),$$

so the K-M estimator of the overall survival is the product of the K-M estimators of the cause-specific survivor-like functions.

2.2 Nelson-Aalen

The Nelson-Aalen estimator of the cause-specific cumulative hazard is

$$\hat{\Lambda}_j(t) = \sum_{i:t_{ji}<t} \frac{d_{ji}}{n_{ji}},$$

and corresponds to an estimate of the cause-specific hazard $\lambda_j(t)$ that takes the value d_{ji}/n_{ji} at t_{ji} and 0 elsewhere. An alternative estimate based on K-M is

$$\hat{\Lambda}_j(t) = -\log \hat{S}_j(t).$$

Of course one could also exponentiate minus the Nelson-Aalen integrated hazard to obtain an alternative estimator of the cause-specific survivor-like function $S_j(t)$.

3 Estimation: Regression Models

Suppose we have n observations consisting of four pieces of information each:

$$(t_i, d_i, j_i, x_i),$$

where t_i is the observation time, d_i is a death indicator (1 if dead, 0 if censored), j_i is a cause of death index (takes a value between 1 and m for deaths and is undefined for censored cases), and x_i is a vector of covariates.

3.1 The Likelihood Function

Under non-informative censoring the likelihood function can be written as

$$L = \prod_{i=1}^{n} \lambda_{j_i}(t_i, x_i)^{d_i} S(t_i, x_i).$$

This likelihood is constructed in the usual manner. A subject censored at time t_i contributes the probability of being alive at that time :

$$S(t_i, x_i).$$

A subject observed to die at time t_i of cause j_i contributes the density of deaths of that cause at that time $f_{j_i}(t_i, x_i)$, which can be written in terms of the hazard and survivor functions as

$$\lambda_{j_i}(t_i, x_i) S(t_i, x_i).$$

Introducing the indicator d_i allows us to write the two types of terms in a compact way.

Recalling that $S(t_i, x_i) = \prod S_j(t_i, x_i)$, we can write the likelihood as

$$L = \prod_{i=1}^{n} \lambda_{j_i}(t_i, x_i)^{d_i} \prod_{j=1}^{m} e^{-\Lambda_j(t_i, x_i)}.$$

Let d_{ij} indicate whether subject i died of cause j. Clearly $d_i = \sum_j d_{ij}$, because you can die of at most one cause. We can then write

$$L = \prod_{i=1}^{m} \prod_{j=1}^{m} \lambda_j(t_i, x_i)^{d_{ij}} e^{-\Lambda_j(t_i, x_i)}.$$

Because the order in which we multiply is immaterial, we can state two important results:

Note 2 *The overall likelihood function is a product of m likelihoods, one for each type of failure.*

This means that we can estimate the $\lambda_j(t, x)$ by maximizing separate likelihoods, as longs as they do not depend on the same parameters. Moreover,

Note 3 *The likelihood involving a specific type of failure is exactly the same likelihood you would obtain by treating all other types of failures as censored observations.*

In other words, each of these likelihoods has exactly the same form that we have studied before. Fitting models is thus a question of applying what we already know.

3.2 Weibull Regression

Suppose the j-th hazard function follows a proportional hazards model with Weibull baseline, say

$$\lambda_j(t, x) = \lambda_{j0}(t)e^{x'\beta},$$

where the baseline hazard is

$$\lambda_{j0}(t) = \lambda_j p_j(\lambda_j t)^{p_j - 1}.$$

In view of the above results, we can estimate the parameters $(p_j, \lambda_j, \beta_j)$ using the techniques discussed before, simply by treating failures for causes other than j as censored cases.

Note that we have allowed *all* parameters to depend on the cause of death. We could, if we wanted, use different x's for each type of failure.

If we wanted to restrict all the Weibulls to have the same index, for example, so $p_j = p \forall j$, then the overall likelihood function would not factor out and we would not be able to use this simplification. The same would be true if we wanted to force some β's to be equal across causes (but why?). In either case one would have to maximize the full likelihood.

3.3 Cox Regression and Partial Likelihood

We can also fit a proportional hazards model without any assumptions about the baseline hazards $\lambda_{j0}(t)$. The standard Cox argument leads to a partial likelihood

$$L = \prod_{j=1}^{m} \prod_{i=1}^{k_j} \frac{e^{x'_{ji(j)}\beta_j}}{\sum_{k \in R(t_{ji})} e^{x'_{jk}\beta_j}},$$

where k_j is the number of distinct times of death due to cause j, t_{ji} denotes the i-th such time, $R(t_{ji})$ is the risk set at time t_{ji} and $i(j)$ is the index of the case that died at t_{ji}. Again:

Note 4 *The overall partial likelihood is a product of m partial likelihoods, one for each type of failure, and each identical to the partial likelihood one would obtain by treating all other causes of death as censored cases.*

If you wanted to restrict the m baseline hazards so they are in turn proportional to a super-baseline, say

$$\lambda_{j0}(t) = \lambda_0(t)e^{\gamma_j},$$

then a different partial likelihood would be obtained; see Equations 8.15-8.16 in K-P.

3.4 Piece-wise Exponential Survival

Here's my favorite model in the context of competing risks. Following the standard argument in Holford or Laird and Olivier, we define intervals with breakpoints $0 = \tau_1 < \tau_2 < \ldots < \tau_{k_1} = \infty$, and assume that the baseline hazard for the j-th type of failure is a step function with a constant value in each interval:

$$\lambda_{j0}(t) = \lambda_{jk}, \quad \text{for} \quad t \in [\tau_k, \tau_{k+1}).$$

Then the factor in the likelihood function corresponding to failures of type j is identical to the kernel of a Poisson likelihood that treats the number of deaths of cause j in interval k to people with covariate values x_i as Poisson with mean

$$\mu_{ijk} = E_{ik}\lambda_{jk}e^{x_i'\beta_j},$$

where E_{ik} is the total exposure of people with covariates x_i in interval k. (Note that the exposure is not cause-specific, at any time each subject is at risk of dying from any cause.)

Thus, we can fit competing risk models by running a series of Poisson regressions where we treat the number of deaths due to each cause as the outcome and the exposure to all causes as the offset.

A nice feature of this model concerns the conditional probability of dying due to cause j at time t given that the subject dies (of some cause) at time t. The probability of dying of cause j at time t given survival to just before t is

$$\lambda_j(t,x) = \lambda_{j0}(t)e^{x'\beta_j} = e^{\alpha_{jk}+x'\beta_j},$$

say, where $\alpha_{jk} = \log \lambda_{jk}$ is the log baseline risk for cause j in interval k. Now the overall risk at that instant is

$$\lambda(t,x) = \sum_{j=1}^{m} \lambda_j(t,x) = \sum_{j=1}^{m} e^{\alpha_{jk}+x'\beta_j}.$$

The conditional probability of interest can then be obtained as

$$\pi_{jk} = \frac{e^{\alpha_{jk}+x'\beta_j}}{\sum_{r=1}^{m} e^{\alpha_{rk}+x'\beta_r}},$$

and can be seen to follow a multinomial logit model with the same parameters β_j as the competing risk model!

This means that we can break down the analysis of competing risks into two parts, using

1 a standard hazards model to get the overall risk, and

2 a multinomial logit model on cause of death.

The results would be exactly the same as fitting separate Poisson models to failures of each type.

4 The Identification Problem

So far we have focused on observable quantities. The literature on competing risks defines *latent* failure times

$$T_1, T_2, \ldots, T_m$$

where T_j is the time when the subject would fail due to the j-th cause.

The problem, of course, is that we only observe the *shortest* of these, as well as an index which tells us which of the T's we have observed. Formally, the data are realizations of two r.v.'s

$$\begin{aligned} T &= \min\{T_1, T_2, \ldots, T_m\} \\ J &= \{j : T_j \leq T_k \forall\ k\} \end{aligned}$$

4.1 Multivariate and Marginal Survival

Let us introduce a joint survivor function, also called the multiple decrement function

$$S_M(t_1, t_2, \ldots, t_m) = \Pr\{T_1 \geq t_1; T_2 \geq t_2; \ldots; T_m \geq t_m\}.$$

Extension to covariates is trivial, so I will keep the notation simple by omitting them.

To be alive at time t all of these potential failure times have to exceed t. This gives us a key identity relating the multivariate and marginal survival functions:

$$S(t) = S_M(t, t, \ldots, t).$$

This shows, incidentally, that $S(t)$ is well defined.

We can also define the cause-specific hazards in terms of partial derivatives of the log of the multivariate survival function:

$$\begin{aligned} \lambda_j(t) &= \lim_{dt \to 0} \frac{\Pr\{t \leq T_j < t + dt | T > dt\}}{dt}, \\ &= -\frac{\partial}{\partial t_j} \log S_M(t, t, \ldots, t). \end{aligned}$$

The multivariate survival function is a function of $t_1, t_2, \ldots, t_m$. The last line refers to the log of the partial derivative w.r.t. t_j evaluated at $t_1 = t_2 = \ldots = t_m = t$. The calculation is conditional on the overall survival time T being at least t. As a result, the event $t \leq T_j < t + dt$ is equivalent to the events $t \leq T < t + dt$ and $J = j$. Thus, the result is the same as the cause-specific hazard introduced earlier.

Note 5 *Because the likelihood of the data depends only on the cause-specific hazards $\lambda_j(t)$, it follows that only these hazards or functions of them can be estimated. Other quantities are not identifiable.*

For example the marginal distributions of the latent failure times are not identifiable. Let

$$\begin{aligned} S_j^*(t) &= \Pr\{T_j \geq t\} \\ &= S_M(0, \ldots, 0, t, 0, \ldots, 0), \end{aligned}$$

where the t appears as the j-th argument to $S_M()$, denote the marginal distribution of T_j. From this marginal survival we can define a marginal hazard

$$\lambda_j^*(t) = -\frac{d}{dt} \log S_j^*(t).$$

This marginal hazard $\lambda_j^*(t)$ is not in general the same as the cause-specific hazard $\lambda_j(t)$. In fact, it cannot be written as a function of $\lambda_j(t)$ without further assumptions. It is therefore not identifiable.

Here comes the big exception. If the latent times T_j are *mutually independent* then

$$S_M(t_1, \ldots, t_m) = \prod_{j=1}^{m} S_j^*(t_j).$$

It then follows that

$$S_j^*(t) = S_j(t),$$

so the marginal survival is the same as the cause-specific survivor-like function we introduced before, and

$$\lambda_j^*(t) = \lambda_j(t),$$

so the marginal hazard is the same as the cause-specific hazard (and is therefore identified).

There is a catch, however. Because the multivariate survival function cannot be estimated, the hypothesis of independence cannot be tested.

In other words, when all you observed is the minimum of the latent times, you cannot distinguish between independent competing risks and infinitely many dependent competing risks that produce exactly the same cause-specific hazards.

4.2 A Bivariate Example

In case you still believe that competing risks models are identified, here is a counter example. The following equation shows a bivariate survival model

$$S(t_1, t_2) = \exp\{1 - \alpha_1 t_1 - \alpha_2 t_2 - e^{\alpha_{12}(\alpha_1 t_1 + \alpha_2 t_2)}\},$$

where $\alpha_1, \alpha_2 > 0$ and α_{12} measures the dependence between T_1 and T_2.

Taking logs and differentiating w.r.t. t_j we find the cause-specific hazards to be

$$\lambda_j(t) = \alpha_j(1 + \alpha_{12} e^{\alpha_{12}(\alpha_1 + \alpha_2)t})$$

and it is clear that all three parameters can be estimated.

Consider, however, a model of independent competing risks, where the marginal (and cause-specific hazards) are given by the above equation. Integrating the marginal hazards we obtain the marginal cumulative hazards and exponentiating minus those gives the marginal survival functions. Multiplying the two survivor functions together we obtain the joint survivor function

$$S(t_1, t_2) = \exp\{1 - \alpha_1 t_1 - \alpha_2 t_2 - \frac{\alpha_1 e^{\alpha_{12}(\alpha_1 + \alpha_2)t_1} + \alpha_2 e^{\alpha_{12}(\alpha_1 + \alpha_2)t_2}}{\alpha_1 + \alpha_2}\},$$

and clearly α_{12} is not a measure of association because by construction T_1 and T_2 are independent!

The point here is that the two bivariate survivor functions are different—moreover, in one case the latent times are correlated while in the other they are independent—yet they lead to the same cause-specific hazards and thus have the same observable consequences.

Thus, if you use the first model and interpret α_{12} as a measure of association between the causes you are relying on untestable assumptions.

4.3 Discussion

The identification problem does not arise if one can observe more than one T_j, but this is usually not feasible. An exception is attrition in panel studies, where one can treat death and attrition as competing risks. It may be possible to have special follow-up studies of attriters to determine if death has

occurred. Having both time to attrition and time to death allows estimation of the correlation between these outcomes.

Heckman has proposed identifying the marginal survival functions by introducing covariates that are supposed to affect one of the latent times but not the others. The problem, again, is that these assumptions themselves are not testable. You cannot check whether a covariate really has no effect on a given type of failure, you have to assume it.

Regrettably, this means that we cannot achieve objective 2 at all:

Note 6 *Data on time to death and cause of death do not permit studying the relationship among failure modes, or even testing for independence.*

It also means that we can achieve the third objective in only a limited sense:

Note 7 *We can only estimate survival following cause-removal under the untestable assumption that the competing risks are independent.*

Of course we are talking about independence given the observed covariates x, so if you have measured every conceivable covariate the assumption of independence would not be unreasonable.

A final note on terminology. The overall probability of failure due to cause j in some interval A is

$$\int_A \lambda_j(t,x)e^{-\Lambda(t,x)}dt;$$

the subject survives all causes up to time t, then dies of cause j.

The same probability if only cause j was operating is, under the assumption of independence

$$\int_A \lambda_j(t,x)e^{-\Lambda_j(t,x)}dt;$$

the subject survives cause j up to time t, then dies of cause j.

In the statistical literature these are called crude and net probabilities, respectively. The demographic literature is not consistent. To avoid confusion it is best to refer to the latter as cause-deleted. In this example all causes other than j have been deleted.

Models for Longitudinal and Clustered Data

Germán Rodríguez

December 9, 2008, revised December 6, 2012

1 Introduction

The most important assumption we have made in this course is that the observations are *independent*. Situations where this assumption is not appropriate include

- Longitudinal data, where we have repeated observations on each individual, for example on multiple waves of a survey
- Clustered data, where the observations are grouped, for example data on mothers and their children
- Multilevel data, where we have multiple levels of grouping, for example students in classrooms in schools.

This is a large subject worthy of a separate course. In these notes I will review briefly the main approaches to the analysis of this type of data, namely fixed and random-effects models. I will deal with linear models for continuous data in Section 2 and logit models for binary data in section 3. I will describe the models in terms of clustered data, using Y_{ij} to represent the outcome for the j-th member of the i-th group. The same procedures, however, apply to longitudinal data, so Y_{ij} could be the response for the i-th individual on the j-th wave. There is no requirement that all groups have the same number of members or, in the longitudinal case, that all individuals have the same number of measurements.

The Stata section of the course website has relevant logs under 'panel data models', including an analysis of data on verbal IQ and language scores for 2287 children in 131 schools in the Netherlands, and a study of the relationship between low birth weight and participation in the Aid to Families with Dependent Children (AFDC) welfare program using state-level data for 1987 and 1990. For binary data we use an example in the Stata manual. A short do file is included at the end of this document.

2 Continuous Data

Suppose that Y_{ij} is a continuous outcome for the j-th member of group i. We are willing to assume independence across groups, but not within each group. The basic idea is that there may be unobserved group characteristics that affect the outcomes of the individuals in each group. We consider two ways to model these characteristics.

2.1 Fixed Effects

The first model we will consider introduces a separate parameter for each group, so the observations satisfy

$$Y_{ij} = \alpha_i + x'_{ij}\beta + e_{ij} \tag{1}$$

Here α_i is a group-specific parameter representing the effect of unobserved group characteristics, the β are regression coefficients representing the effects of the observed covariates, and the e_{ij} are *independent* error terms, say $e_{ij} \sim N(0, \sigma_e^2)$.

You can think of the α_i as equivalent to introducing a separate dummy variable for each group. It is precisely because we have controlled for (all) group characteristics that we are willing to assume independence of the observations. Unfortunately this implies that we cannot include group-level covariates among the predictors, as they would be collinear with the dummies. Effectively this means that we can control for group characteristics, but we cannot estimate their effects.

This model typically has a large number of parameters, and this causes practical and theoretical problems.

In terms of theory the usual OLS estimator of α_i is consistent as the number of individuals approaches infinity in every group, but is not consistent if the number of groups approaches infinity but the number of individuals per group does not, which is the usual case of interest. Fortunately the OLS estimator of β is consistent in both cases.

On the practical side, introducing a dummy variable for each group may not be feasible when the number of groups is very large. Fortunately it is possible to solve for the OLS estimator of β without having to estimate the α_i's explicitly through a process known as *absorption*.

An alternative is to remove the α_i from the model by differencing or conditioning. This is very easy to do if you have two observations per group, as would be the case for longitudinal data from a two-wave survey. Suppose

Y_{i1} and Y_{i2} follow model (1). The *difference* would then follow the model

$$Y_{i2} - Y_{i1} = (x_{i2} - x_{i1})'\beta + (e_{i2} - e_{i1})$$

which is a linear model with exactly the same regression coefficients as (1). Moreover, because the e_{ij} are independent, so are their differences. This means that we can obtain unbiased estimates of β by simply differencing the Y's and the x's and using ordinary OLS on the differences.

The same idea can be extended to more than two observations per group, and it involves working with a transformation of the data reflecting essentially differences with respect to the group means. The same estimator can also be obtained by working with the conditional distribution of the observations given the group totals $Y_i = \sum_j Y_{ij}$.

Looking at the model in terms of differences shows clearly how it can control for unobserved group characteristics. Suppose the 'true' model includes a group-level predictor z_i with coefficient γ, so

$$Y_{ij} = z_i'\gamma + x_{ij}'\beta + e_{ij}$$

When you difference the y's the term $z_i'\gamma$ drops out. Therefore you can estimate effects of the x's controlling for z even though you haven't observed z! Unfortunately, this also means that in a fixed-effects model we can't estimate γ even if we have observed z_i, as noted earlier.

2.2 Random Effects

An alternative approach writes a model that looks almost identical to the previous one:

$$Y_{ij} = a_i + x_{ij}'\beta + e_{ij} \tag{2}$$

Here a_i is a *random* variable representing a group-specific effect, β is a vector of regression coefficients and the e_{ij} are independent error terms.

You can think of the a_i and e_{ij} as two error terms, one at the level of the group and the other at the level of the individual. As usual with error terms we assign them distributions; specifically we assume that $a_i \sim N(0, \sigma_a^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. We also assume that e_{ij} is independent of a_i.

Another way to write the model is by combining the two error terms in one:

$$Y_{ij} = x_{ij}'\beta + u_{ij}$$

where $u_{ij} = a_i + e_{ij}$. This looks like an ordinary regression model, but the errors are not independent. More precisely, they are independent across

groups but not within a subgroup because the u_{ij}'s for members of group i share a_i.

We can write the correlation between any two observations in the same group as

$$\rho = \text{cor}(Y_{ij}, Y_{ij'}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}.$$

a result that follows directly from the usual definition of correlation; the covariance between Y_{ij} and $Y_{ij'}$ is σ_a^2 and the variance of either is $\sigma_a^2 + \sigma_e^2$. This coefficient is often called *the intra-class correlation coefficient.*

Because the variance of the observations has been partitioned into two components these models are also called *variance components models.* The term σ_a^2 represents variation across groups (usually called *between* groups, even if we have more than two) and the term σ_e^2 represents variation *within* groups.

If we were to use OLS estimation in the model of equation (2) we would obtain consistent estimates for the regression coefficients β, but the estimates would not be fully efficient because they do not take into account the covariance structure, and the standard errors would be biased unless they are corrected for clustering.

Fortunately maximum likelihood estimation is pretty straightforward, and yields fully-efficient estimates. We also obtain as by-products estimates of the error variances σ_a^2 and σ_e^2 and the intra-class correlation ρ. (Stata also computes these quantities for fixed-effect models, where they are best viewed as components of the total variance.)

2.3 Fixed Versus Random Effects

There is a lot of confusion regarding fixed and random-effects models. Here are five considerations that may help you decide which approach may be more appropriate for a given problem.

First let us note the obvious, in one case the α_i are fixed but unknown parameters to be estimated (or differenced out of the model), in the other the a_i are random variables and we estimate their distribution, which has mean zero and variance σ_a^2. This distinction leads to one traditional piece of advice: use random effects if you view the groups as a sample from a population, and fixed effects if you are interested in inferences for the specific groups at hand. I find this advice to be the least useful of all. (It is particularly baffling to Bayesians, who view all parameters as random.)

Second, note that the a_i are assumed to be *independent* across groups, which is another way of saying that they have to be uncorrelated with ob-

served group covariates, as all well-behaved error terms are supposed to do. In contrast, the α_i can be seen to control for all unobserved group characteristics that are shared by group members, whether or not they are correlated with the observed covariates. This is a very useful distinction. Econometricians often view fixed effects as random effects which happen to be correlated with the observed covariates.

Third, note that the fixed-effects estimator cannot estimate the effects of group-level variables, or more generally variables that are constant across group members. Otherwise you might think that all we need is the fixed-effects estimator, which is valid under more general conditions. (Incidentally there is a Hausman especification test for random effects which compares the two estimators of the effects for individual-level variables. Just bear in mind that when this test rejects the random specification it doesn't mean that the fixed specification is valid, just that the random is not.)

Fourth, fixed-effect models deal with just two levels, whereas random-effects models can be generalized easily to more than two levels. This can become an important consideration if you have three-level data, for example children, families and communities, and want to study the dependence at all levels.

Fifth, in a random-effects framework we can let any of the coefficients vary from group to group, not just the constant, moving from so-called *random-intercept* models to more interesting *random-slope* models. You can think of a random slope as interacting an individual covariate with unobserved group characteristics. Of particular interest are treatment effects that may vary from group to group. (Or from individual to individual if you have repeated measurements on each person.)

2.4 Between and Within Groups

There's one more way to look at these models. Let us start from the random-effects model and consider the group means, which follow the model

$$\bar{Y}_i = a_i + \bar{x}_i'\beta + \bar{e}_i \tag{3}$$

where we have also averaged the covariates and the error terms for all members of each group. The key fact is that the means follow a linear model with the same regression coefficients β as the individual data.

If the error terms are independent across groups then we can obtain a consistent estimator of β using OLS, or WLS if the number of observations varies by group. (If the a_i are correlated with the x's, however, we have the usual endogeneity problem.) We call this the *between*-groups estimator.

We can also look at deviations from the group means, which follow the model

$$Y_{ij} - \bar{Y}_i = (x_{ij} - \bar{x}_i)'\beta + (e_{ij} - \bar{e}_i)$$

The interesting thing here is that the deviations from the mean also follow a linear model with the same regression coefficients β. The errors are not independent across subjects, but the dependence arises just from subtracting the mean and is easily corrected. We call the resulting estimator the *within*-groups estimator.

It can be shown that the fixed-effects estimator is the same as the within-groups estimator, and that the random-effects estimator is an average or compromise between the between and within estimators, with the precise weight a function of the intra-class correlation.

In the context of multilevel models it is possible to reconcile the fixed and random-effects approaches by considering the group means as additional predictors. Specifically, consider the model

$$Y_{ij} = a_i + \bar{x}_i'\beta_B + (x_{ij} - \bar{x}_i)'\beta_W + e_{ij}$$

where the group mean and the individual's deviation from its group mean appear as predictors. The estimate of β_B, representing the effect of the group average on individual outcomes, coincides with the between-group estimator. The estimate of β_W, representing the effect of an individual's deviation from the group average, coincides with the within-groups or fixed-effects estimator. The random-effects estimator is appropriate only if both coefficients are equal, in which case it is appropriate to average the two estimates.

This more general model can be fitted by OLS to obtain consistent if not fully efficient parameters estimates, but to obtain correct standard errors you would need to correct for clustering at the group level. A much better approach is to fit the model as a random-effects model, in which case maximum likelihood will yield fully-efficient estimates and correct standard errors.

2.5 Examples

We consider two examples, one where the fixed and random-effect approaches lead to similar estimates and one where they differ substantially.

Example 1. Snijders and Bosker (1999) have data for 2287 eighth-grade children in 131 schools in the Netherlands. We are interested in the relationship between verbal IQ and the score in a language test. The table below compares OLS, fixed-effects and random-effects estimators.

```
-------------------------------------------------
Variable     |    ols           re           fe
-------------+-----------------------------------
#1           |
     iq_verb |  2.6538956     2.488094     2.4147722
       _cons |  9.5284841    11.165109    12.358285
-------------+-----------------------------------
sigma_u      |
       _cons |                3.0817186
-------------+-----------------------------------
sigma_e      |
       _cons |                6.4982439
-------------------------------------------------
```

The differences among the three approaches in this particular example are modest. The random-effects model estimates the correlation between the language scores of children in the same school as 0.18. This is equivalent to saying that 18% of the variance in language scores is across schools, and of course 82% is among students in the same school.

The Stata logs also show the regression based on school means, with a coefficient of 3.90, and separate regressions for each school, indicating that the relationship between verbal IQ and language scores varies by school.

Example 2. Wooldridge (2002) has an interesting dataset with the percentage of births classified as low birth weight and the percentage of the population in the AFDC welfare program in each of the 50 states in 1987 and 1990.

We consider models predicting low birth weight from AFDC participation and a dummy for 1990. For simplicity we ignore other controls such as physicians per capita, beds per capita, per capita income, and population (all logged), which turn out not to be needed in the fixed-effects specification. Here are the results:

```
-------------------------------------------------
Variable     |    ols           re           fe
-------------+-----------------------------------
#1           |
         d90 |  .03326679    .14854716    .21247362
     afdcprc |  .26012832   -.01566323   -.16859799
       _cons |  5.6618251    6.6946585    7.2673958
-------------+-----------------------------------
```

```
sigma_u      |
       _cons |               1.1478165
-------------+----------------------------------------
sigma_e      |
       _cons |                .19534447
------------------------------------------------------
```

The OLS estimate suggests that AFDC has a pernicious effect on low birth weight: a higher percentage of the population in AFDC is associated with increased prevalence of low birth weight. The random-effects estimator shows practically no association between AFDC participation and low birth weight. The intra- state correlation is 0.972, indicating that 97% of the variation in low birth weight is across states and only 3% is within states over time. Focusing on intra-state variation, the fixed-effects estimator shows that an increase in the percent of the population in AFDC is associated with a reduction in the percent of low birth-weight births, a much more reasonable result.

My interpretation of these results is that there are unobserved state characteristics (such as poverty) that increase both AFDC participation and the prevalence of low birth weight, inducing a (spurious) positive correlation that masks or reverses the (true) negative effect of AFDC participation on low birth weight. By controlling (implicitly) for all persistent state characteristics, the fixed-effects estimator is able to unmask the negative effect.

The Stata log expands on these analysis using all the controls mentioned above. It also shows how one can reproduce the fixed effects estimate by working with changes between 1987 and 1990 in AFDC participation and in the percent low birth weight, or by working with the original data and introducing a dummy for each state.

3 Binary Data

We now consider extending these ideas to modeling binary data, which pose a few additional challenges. In this section Y_{ij} is a binary outcome which takes only the values 0 and 1.

3.1 Fixed-Effects Logits

In a fixed-effects model we assume that the Y_{ij} have independent Bernoulli distributions with probabilities satisfying

$$\text{logit}(\pi_{ij}) = \alpha_i + x'_{ij}\beta$$

Effectively we have introduced a separate parameter α_i for each group, thus capturing unobserved group characteristics.

Introducing what may be a large number of parameters in a logit model causes the usual practical difficulties and a twist on the theory side. In the usual scenario, where we let the number of groups increase to infinity but not the number of individuals per group, it is not just the estimates of α_i that are not consistent, but the inconsisteny propagates to β as well! This means that there is not point in introducing a separate dummy variable for each group, even if we could.

There is, however, an alternative approach that leads to a consistent estimator of β. We calculate the total number of successes for each group, say $Y_i = \sum_j Y_{ij}$, and look at the distribution of each Y_{ij} given the total Y_i. It turns out that this conditional distribution does not involve the α_i but does depend on β, which can thus be estimated consistently.

In the linear case the dummy and conditioning approaches were equivalent. Here they are not. The conditioning approach requires the existence of a minimal sufficient statistic for the α_i. In logit models the totals have this property. Interestingly, in probit models there is no minimal sufficient statistic for the α_i, which is why there is no such thing as a fixed-effects probit model.

We will skip the details here except to note that conditioning means that groups were all observations are successes (or all are failures) do not contribute to the conditional likelihood. In some situations this can lead to estimating the model in a small subset of the data. This is worrying, but advocates of fixed-effects models argure that those are the only cases with revelant information.

An example may help fix ideas. Suppose one was interested in studying the effect of teenage pregnancy on high school graduation. In order to control for unobserved family characteristics, you decide to use data on sisters and fit a fixed-effects model. Consider families with two sisters. If both graduate from high school, the conditional probability of graduation is one for each sister, and hence the pair is uninformative. If neither graduates the conditional probability of graduation is zero, and thus the pair is also uninformative. It is only when one of the sisters graduates and the other doesn't that we have some information.

So far we have considered variation in the outcome but it turns out that we also need variation in the predictor. If both sisters had a teenage pregnancy the pair provides no information regarding the effect of pregnancy on graduation. The same is true if neither gets pregnant. The only families that contribute information consist of pairs where one sister get pregnant

and the other doesn't, and where one graduates from high school and the other doesn't. The question then becomes whether the one who graduates is the one who didn't get pregnant, an event that can be shown to depend on the parameter of interest and is not affected by unobserved *family* characteristics.

The concern is that very few pairs meet these conditions, and those pairs may be selected on unobserved *individual* characteristics. To see why this is a problem suppose the effect of teenage pregnancy on high school graduation varies with an unobserved individual attribute. The estimated effect can still be interpreted as an average, but the average would be over a selected subset, not the entire population.

3.2 Random-Effects Logits

In a random-effects logit model we postulate the existence of an unobserved individual effect a_i such that *given* a_i the Y_{ij} are independent Bernoulli random variables with probability π_{ij} such that

$$\text{logit}(\pi_{ij}) = a_i + x'_{ij}\beta$$

In other words the *conditional* distribution of the outcomes given the random effects a_i is Bernoulli, with probability following a standard logistic regression model with coefficients a_i and β.

Just as before we treat a_i as an error term and assume a distribution, namely $N(0, \sigma_a^2)$. One difficulty with this model is that the *unconditional* distribution of Y_{ij} involves a logistic-normal integral and does not have a closed form.

This lead several authors to propose approximations, such as marginal quasi-likelihood (MQL) or penalized quasi-likelihood (PQL), but unfortunately these can lead to substantial biases (Rodríguez and Goldman, 1995).

Fortunately it is possible to evaluate the likelihood to a close approximation using *Gaussian quadrature*, a procedure that relies on a weighted sum of conditional probabilities evaluated at selected values of the random effect. These values can be pre-determined or tailored to the data at hand in a procedure known as adaptive Gaussian quadrature, the latest Stata default.

The model can also be formulated in terms of a *latent variable* Y_{ij}^* such that $Y_{ij} = 1$ if and only if $Y_{ij}^* > 0$, by assuming that the latent variable follows a random-effects linear model

$$Y_{ij}^* = a_i + x'_{ij}\beta + e_{ij}$$

where e_{ij} has a standard logistic distribution. The unconditional distribution of Y^* is then logistic-normal and, as noted above, does not have a closed form.

Recall that the variance of the standard logistic is $\pi^2/3$. This plays the role of σ_e^2, the individual variance. We also have the group variance σ_a^2. Using these two we can compute an *intraclass correlation* for the latent variable:

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \pi^2/3}$$

Computing an intra-class correlation for the manifest outcomes is a bit more complicated, as the coefficient turns out to depend on the covariates, see Rodríguez and Elo, 2000) and their `xtrho` command.

3.3 Subject-Specific and Population-Average Models

A common mistake is to believe that all one needs to do with clustered or longitudinal data is to run ordinary regression or logit models and then correct the standard errors for clustering.

This is essentially correct in the linear case, where OLS estimators are consistent but not fully effcient, so all one sacrifices with this approach is a bit of precision. But with logit models, ignoring the random effect introduces a bias in the estimates as well as the standard errors.

To see this point consider a random-effects model, where the expected value of the outcome Y_{ij} given the random effect a_i is

$$E(Y_{ij}|a_i) = \text{logit}^{-1}(a_i + x'_{ij}\beta_{SS})$$

An analyst ignoring the random effects would fit a model where the expected value is assumed to be

$$E(Y_{ij}) = \text{logit}^{-1}(x'_{ij}\beta_{PA})$$

Note that we have been careful to use different notation for the coefficients. We call β_{SS} the subject-specific effect and β_{PA} the population-average effect, because we have effectively averaged over all groups in the population.

In the linear case (just ignore the inverse logit in the above two equations) taking expectation with respect to a_i in the first equation leads to the second, so $\beta_{SS} = \beta_{PA}$ and both approaches estimate the same parameter.

Because of the non-linear nature of the logit function, however, taking expectation in the first equation does not lead to the second. In fact, if the first model is correct the second usually isn't, except approximately.

Typically $|\beta_{PA}| < |\beta_{SS}|$, so population-average effects are smaller in magnitude than subject-specific effects, with the difference increasing with the intra-class correlation.

One could make a case for either model, the main point here is that they differ. From a policy point of view, for example, one could argue that decisions should be based on the average effect. I find this argument more persuasive with longitudinal data, where the averaging is for individuals over time, than with hierarchical data. Suppose you are evaluating a program intended to increase the probability of high school graduation and the model includes a school random effect. Are you interested in the increase in the odds of graduation for students in the school they attend or an hypothetical increase averaged over all the schools in the population?

3.4 Example

Our example comes from the Stata manual and is based on data from the National Longitudinal Survey (NLS) for 4,434 women who were 14-24 in 1968 and were observed between 1 and 12 times each. We are interested in union membership as a function of age, education (grade), and residence, represented by dummy variables for 'not a standard metropolitan area' and the south, plus an interaction between south and time (coded as zero for 1970). We fit ordinary, random-effects, and fixed-effects logit models.

```
------------------------------------------------------
Variable     |    logit        relogit      felogit
-------------+----------------------------------------
#1           |
         age |   .00999311    .00939361    .00797058
       grade |   .04834865    .08678776    .08118077
    not_smsa |  -.22149081   -.25193788    .02103677
       south |  -.71444608   -1.1637691   -1.0073178
     southXt |    .0068356    .02324502    .02634948
       _cons |  -1.8882564   -3.3601312
-------------+----------------------------------------
lnsig2u      |
       _cons |                1.7495341
------------------------------------------------------
```

Compare first the logit and random-effects logit models. We see that, except for age, the subject-specific effects are larger in magnitude than the

population-average effects, as we would expect. For example a woman living in the south in 1970 has 69% lower odds of being a union member than one living elsewhere, everything else being equal. The logit model, however, estimates the average effect as 51% lower odds in 1970. The intraclass correlation measured in a latent scale of propensity to belong to a union is 0.636.

The fixed-effects estimates are in general agreement with the random-effects results except for the indicator for living outside a standard metropolitan area, which changes from -0.252 to $+0.021$. This suggests that the negative association between living outside a SMA and belonging to a union is likely to be spurious, due to persistent unobserved characteristics of women that are associated with both SMA residence and union membership. If we estimate the effect by comparing union membership for the same women when they lived in and outside a SMA we find no association.

Note in closing that we had a total of 26,200 observations on 4,434 women. However, the fixed-effects logit analysis dropped 14,165 observations on 2,744 women because they had no variation over time in union membership.

4 Appendix: Stata Commands

Here's a copy of the do file used to produce the results in this handout.

```
// WWS 509 - Fall 2008 - G. Rodriguez <grodri@princeton.edu>
// Models for Clustered and Longitudinal Data

// Verbal IQ and language scores
use http://data.princeton.edu/wws509/datasets/snijders, clear
reg langpost iq_verb
estimates store ols
xtreg langpost iq_verb, i(schoolnr) mle
estimates store re
xtreg langpost iq_verb, i(schoolnr) fe
estimates store fe
estimates table ols re fe, eq(1 1 1)

// AFDC participation and low birth weight
use http://www.stata.com/data/jwooldridge/eacsap/lowbirth, clear
encode stateabb, gen(stateid)
reg lowbrth d90 afdcprc
estimates store ols
xtreg lowbrth d90 afdcprc, i(stateid) mle
estimates store re
xtreg lowbrth d90 afdcprc, i(stateid) fe
estimates store fe
estimates table ols re fe, eq(1 1 1)

// Union membership
use http://data.princeton.edu/wws509/datasets/union, clear
logit union age grade not_smsa south southXt
estimates store logit
xtlogit union age grade not_smsa south southXt, i(id) re
estimates store relogit
xtlogit union age grade not_smsa south southXt, i(id) fe
estimates store felogit
estimates table logit relogit felogit, eq(1 1 1)
```

Nonparametric Regression

John Fox

Department of Sociology

McMaster University

1280 Main Street West

Hamilton, Ontario

Canada L8S 4M4

jfox@mcmaster.ca

February 2004

Abstract

Nonparametric regression analysis traces the dependence of a response variable on one or several predictors without specifying in advance the function that relates the predictors to the response. This article discusses several common methods of nonparametric regression, including kernel estimation, local polynomial regression, and smoothing splines. Additive regression models and semiparametric models are also briefly discussed.

Keywords: kernel estimation; local polynomial regression; smoothing splines; additive regression; semiparametric regression

Nonparametric regression analysis traces the dependence of a response variable (y) on one or several predictors (xs) without specifying in advance the function that relates the response to the predictors:

$$E(y_i) = f(x_{1i}, \ldots, x_{pi})$$

where $E(y_i)$ is the mean of y for the ith of n observations. It is typically assumed that the conditional variance of y, $\text{Var}(y_i|x_{1i}, \ldots, x_{pi})$ is a constant, and that the conditional distribution of y is normal, although these assumptions can be relaxed.

Nonparametric regression is therefore distinguished from linear regression, in which the function relating the mean of y to the xs is linear in the parameters,

$$E(y_i) = \alpha + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

and from traditional nonlinear regression, in which the function relating the mean of y to the xs, though nonlinear in its parameters, is specified explicitly,

$$E(y_i) = f(x_{1i}, \ldots, x_{pi}; \gamma_1, \ldots, \gamma_k)$$

In traditional regression analysis, the object is to estimate the parameters of the model — the βs or γs. In nonparametric regression, the object is to estimate the regression function directly.

There are many specific methods of nonparametric regression. Most, but not all, assume that the regression function is in some sense smooth. Several of the more prominent methods are described in this article. Moreover, just as traditional linear and nonlinear regression can be extended to generalized linear and nonlinear regression models that accommodate non-normal error distributions, the same is true of nonparametric regression. There is a large literature on nonparametric regression analysis, both in scientific journals and in texts. For more extensive introductions to the subject, see in particular, Bowman and Azzalini [1], Fox [2, 3], Hastie and Tibshirani [4], Hastie, Tibshirani, and Freedman [5], and Simonoff [6].

The simplest use of nonparametric regression is in smoothing scatterplots. Here, there is a numerical

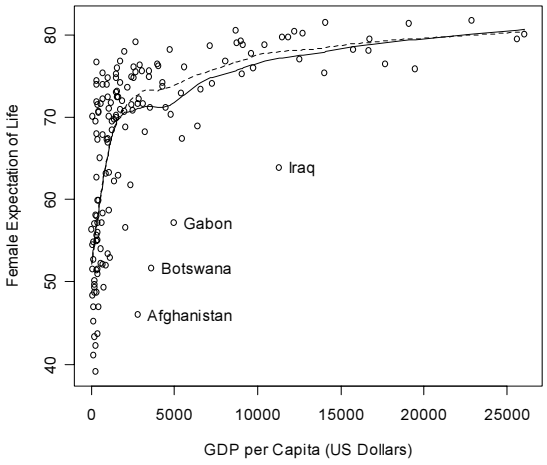


Figure 1: Female expectation of life by GDP per capita, for 154 nations of the world. The solid line is for a local-linear regression with a span of 0.5, while the broken line is for a similar fit deleting the four outlying observations that are labelled on the plot.

response y and a single predictor x, and we seek to clarify visually the relationship between the two variables in a scatterplot. Figure 1, for example, shows the relationship between female expectation of life at birth and GDP per capita for 154 nations of the world, as reported in 1998 by the United Nations. Two fits to the data are shown, both employing local linear regression (described below); the solid line represents a fit to all of the data, while the broken line omits four outlying nations, labelled on the graph, which have values of female life expectancy that are unusually low given GDP per capita. It is clear that although there is a positive relationship between expectation of life and GDP, the relationship is highly nonlinear, levelling off substantially at high levels of GDP.

Three common methods of nonparametric regression are *kernel estimation*, *local-polynomial regression* (which is a generalization of kernel estimation), and *smoothing splines*. *Nearest-neighbor* kernel estimation proceeds as follows (as illustrated for the UN data in Figure 2):

1. Let x_0 denote a focal x-value at which $f(x)$ is to be estimated; in Figure 2 (a), the focal value is the 80th ordered x-value in the UN data, $x_{(80)}$. Find the m nearest x-neighbors of x_0, where $s = m/n$ is called the *span* of the kernel smoother. In the example, the span was set to $s = 0.5$, and thus $m = 0.5 \times 154 = 77$. Let h represent the half-width of a window encompassing the m nearest neighbors

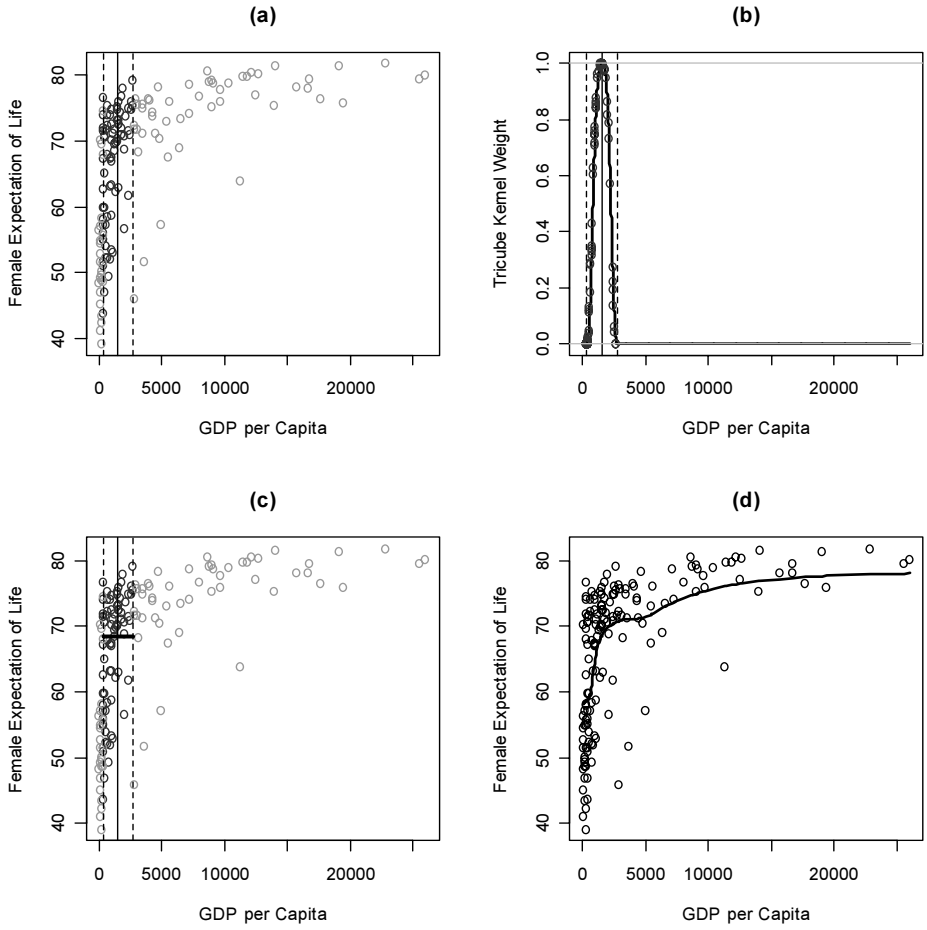


Figure 2: How the kernel estimator works: (a) A neighborhood including the 77 observations closest to $x_{(80)}$, corresponding to a span of 0.5. (b) The tricube weight function defined on this neighborhood; the points show the weights for the observations. (c) The weighted mean of the y-values within the neighborhood, represented as a horizontal line. (d) The nonparametric regression line connecting the fitted values at each observation. (The four outlying points are excluded from the fit.)

of x_0. The larger the span (and hence the value of h), the smoother the estimated regression function.

2. Define a symmetric unimodal weight function, centered on the focal observation, that goes to zero (or nearly zero) at the boundaries of the neighborhood around the focal value. The specific choice of weight function is not critical: In Figure 2 (b), the tricube weight function is used:

$$W_T(x) = \begin{cases} \left[1 - \left(\dfrac{|x - x_0|}{h}\right)^3\right]^3 & \text{for } \dfrac{|x - x_0|}{h} < 1 \\ 0 & \text{for } \dfrac{|x - x_0|}{h} \geq 1 \end{cases}$$

A Gaussian (normal) density function is another common choice.

3. Using the tricube (or other appropriate) weights, calculate the weighted average of the y-values to obtain the fitted value

$$\widehat{y}_0 = \widehat{f}(x_0) = \frac{\sum W_T(x_i) y_i}{\sum W_T(x_i)}$$

as illustrated in Figure 2 (c). Greater weight is thus accorded to observations whose x-values are close to the focal x_0.

4. Repeat this procedure at a range of x-values spanning the data — for example, at the ordered observations $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. Connecting the fitted values, as in Figure 2 (d), produces an estimate of the regression function.

Local polynomial regression is similar to kernel estimation, but the fitted values are produced by locally weighted regression rather than by locally weighted averaging; that is, $\widehat{y}_0$ is obtained in step 3 by the polynomial regression of y on x to minimize the weighted sum of squared residuals

$$\sum W_T(x_i)(y_i - a - b_1 x_i - b_2 x_i^2 - \cdots - b_k x_i^k)^2$$

Most commonly, the order of the local polynomial is taken as $k = 1$, that is, a local linear fit (as in Figure 1). Local polynomial regression tends to be less biased than kernel regression, for example at the boundaries of data — as in evident in the artificial flattening of the kernel estimator at the right of Figure 2 (d). More generally, the bias of the local-polynomial estimator declines and the variance increases with the order of

the polynomial, but an odd-ordered local polynomial estimator has the same asymptotic variance as the preceding even-ordered estimator: Thus, the local-linear estimator (of order 1) is preferred to the kernel estimator (of order 0), and the local-cubic (order 3) estimator to the local-quadratic (order 2).

Smoothing splines are the solution to the *penalized regression problem*: Find $\widehat{f}(x)$ to minimize

$$S(h) = \sum [y_i - f(x_i)]^2 + h \int_{x(1)}^{x(n)} [f''(x)]^2 \, dx$$

Here h is a *roughness penalty*, analogous to the span in nearest-neighbor kernel or local polynomial regression, and f'' is the second derivative of the regression function (taken as a measure of roughness). Without the roughness penalty, nonparametrically minimizing the residual sum of squares would simply interpolate the data. The mathematical basis for smoothing splines is more satisfying than for kernel or local polynomial regression, since an explicit criterion of fit is optimized, but spline and local polynomial regressions of equivalent smoothness tend to be similar in practice.

Local regression with several predictors proceeds as follows, for example. We want the fit $\widehat{y}_0 = \widehat{f}(\mathbf{x}_0)$ at the focal point $\mathbf{x}_0 = (x_{10}, \ldots, x_{p0})$ in the predictor space. We need the distances $D(\mathbf{x}_i, \mathbf{x}_0)$ between the observations on the predictors and the focal point. If the predictors are on the same scale (as, for example, when they represent coordinates on a map), then measuring distance is simple; otherwise, some sort of standardization or generalized distance metric will be required. Once distances are defined, weighted polynomial fits in several predictors proceed much as in the bivariate case. Some kinds of spline estimators can also be generalized to higher dimensions.

The generalization of nonparametric regression to several predictors is therefore mathematically straightforward, but it is often problematic in practice. First, multivariate data are afflicted by the so-called *curse of dimensionality*: Multidimensional spaces grow exponentially more sparse with the number of dimensions, requiring very large samples to estimate nonparametric regression models with many predictors. Second, although slicing the surface can be of some help, it is difficult to visualize a regression surface in more than three dimensions (that is, for more than two predictors).

Additive regression models are an alternative to unconstrained nonparametric regression with several predictors. The additive regression model is

$$E(y_i) = \alpha + f_1(x_{1i}) + \cdots + f_p(x_{pi})$$

where the f_j are smooth partial-regression functions, typically estimated with smoothing splines or by local regression. This model can be extended in two directions: (1) To incorporate interactions between (or among) specific predictors; for example

$$E(y_i) = \alpha + f_1(x_{1i}) + f_{23}(x_{2i}, x_{3i})$$

which is not as general as the unconstrained model $E(y_i) = \alpha + f(x_{1i}, x_{2i}, x_{3i})$. (2) To incorporate linear terms, as in the model

$$E(y_i) = \alpha + \beta_1 x_{1i} + f_2(x_{2i})$$

Such *semiparametric* models are particularly useful for including dummy regressors or other contrasts derived from categorical predictors.

Returning to the UN data, an example of a simple additive regression model appears in Figures 3 and 4. Here female life expectancy is regressed on GDP per capita and the female rate of illiteracy, expressed as a percentage. Each term in this additive model is fit as a smoothing spline, using the equivalent of four degrees of freedom. Figure 3 shows the two-dimensional fitted regression surface, while Figure 4 shows the partial regression functions, which in effect slice the regression surface in the direction of each predictor; because the surface is additive, all slices in a particular direction are parallel, and the two-dimensional surface in three-dimensional space can be summarized by two two-dimensional graphs. The ability to summarize the regresson surface with a series of two-dimensional graphs is an even greater advantage when the surface is higher-dimensional.

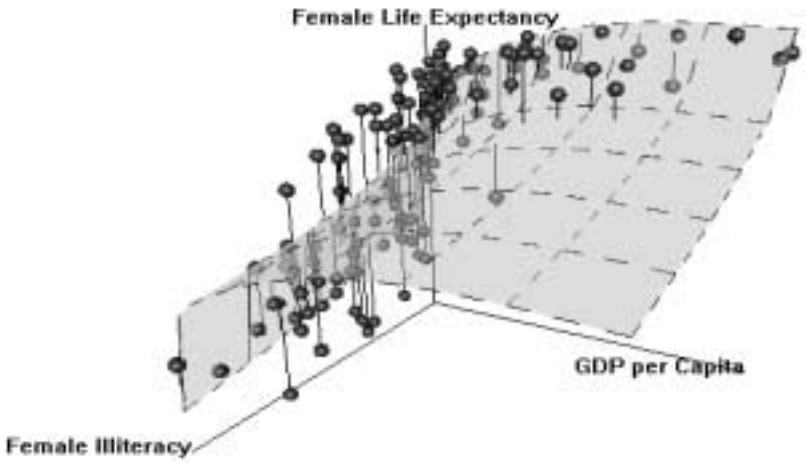


Figure 3: Fitted regression surface for the additive regression of female expectation of life on GDP per capita and female illiteracy. The vertical lines represent residuals.

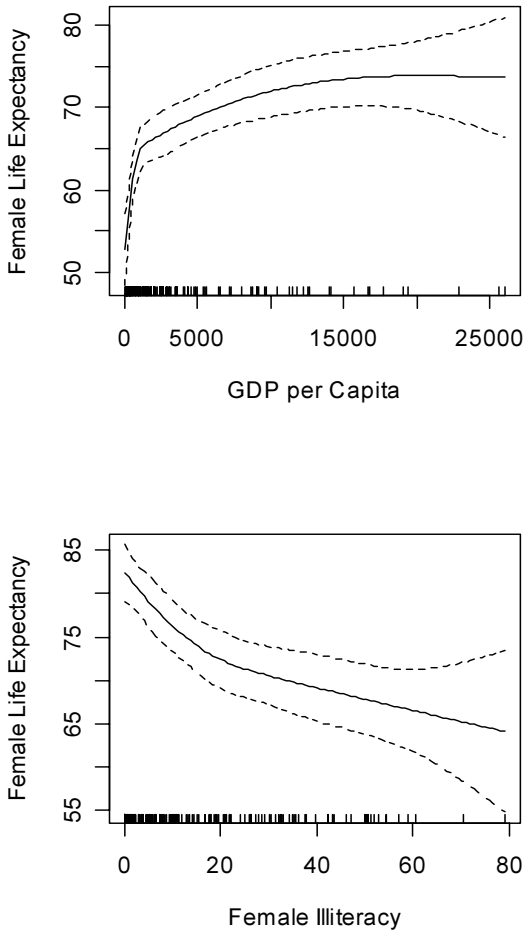


Figure 4: Partial regression functions from the additive regression of female life expectancy on GDP per capita and female illiteracy; each term is fit by a smoothing spline using the equivalent of four degrees of freedom. The "rug plot" at the bottom of each graph shows the distribution of the predictor, and the broken lines give a point-wise 95-percent confidence envelope around the fit.

A central issue in nonparametric regression is the selection of smoothing parameters — such as the span in kernel and local-polynomial regression or the roughness penalty in smoothing-spline regression (or equivalent degrees of freedom for any of these). In the examples in this article, smoothing parameters were selected by visual trial and error, balancing smoothness against detail. The analogous statistical balance is between variance and bias, and some methods (such as cross-validation) attempt to select smoothing parameters to minimize estimated mean-square error (i.e., the sum of squared bias and variance).

References

[1] A. W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press, Oxford, 1997.

[2] J. Fox. *Nonparametric Simple Regression: Smoothing Scatterplots*. Sage, Thousand Oaks CA, 2000a.

[3] J. Fox. *Multiple and Generalized Nonparametric Regression*. Sage, Thousand Oaks CA, 2000b.

[4] T. Hastie, R. Tibshirani, and J. Freedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2001.

[5] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.

[6] J. S. Simonoff. *Smoothing Methods in Statistics*. Springer, New York, 1996.

Appendix B

Generalized Linear Model Theory

We describe the generalized linear model as formulated by Nelder and Wedderburn (1972), and discuss estimation of the parameters and tests of hypotheses.

B.1 The Model

Let $y_1, \ldots, y_n$ denote n independent observations on a response. We treat y_i as a realization of a random variable Y_i. In the general linear model we assume that Y_i has a normal distribution with mean μ_i and variance σ^2

$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2),$$

and we further assume that the expected value μ_i is a linear function of p predictors that take values $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$ for the i-th case, so that

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters.

We will generalize this in two steps, dealing with the stochastic and systematic components of the model.

B.1.1 The Exponential Family

We will assume that the observations come from a distribution in the exponential family with probability density function

$$f(y_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\}. \tag{B.1}$$

Here θ_i and ϕ are parameters and $a_i(\phi)$, $b(\theta_i)$ and $c(y_i, \phi)$ are known functions. In all models considered in these notes the function $a_i(\phi)$ has the form

$$a_i(\phi) = \phi/p_i,$$

where p_i is a known *prior weight*, usually 1.

The parameters θ_i and ϕ are essentially location and scale parameters. It can be shown that if Y_i has a distribution in the exponential family then it has mean and variance

$$\begin{aligned} \mathrm{E}(Y_i) &= \mu_i = b'(\theta_i) && \text{(B.2)} \\ \mathrm{var}(Y_i) &= \sigma_i^2 = b''(\theta_i)a_i(\phi), && \text{(B.3)} \end{aligned}$$

where $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of $b(\theta_i)$. When $a_i(\phi) = \phi/p_i$ the variance has the simpler form

$$\mathrm{var}(Y_i) = \sigma_i^2 = \phi b''(\theta_i)/p_i.$$

The exponential family just defined includes as special cases the normal, binomial, Poisson, exponential, gamma and inverse Gaussian distributions.

Example: The normal distribution has density

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2}\frac{(y_i - \mu_i)^2}{\sigma^2}\}.$$

Expanding the square in the exponent we get $(y_i - \mu_i)^2 = y_i^2 + \mu_i^2 - 2y_i\mu_i$, so the coefficient of y_i is μ_i/σ^2. This result identifies θ_i as μ_i and ϕ as σ^2, with $a_i(\phi) = \phi$. Now write

$$f(y_i) = \exp\{\frac{y_i\mu_i - \frac{1}{2}\mu_i^2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\}.$$

This shows that $b(\theta_i) = \frac{1}{2}\theta_i^2$ (recall that $\theta_i = \mu_i$). Let us check the mean and variance:

$$\begin{aligned} E(Y_i) &= b'(\theta_i) = \theta_i = \mu_i, \\ \mathrm{var}(Y_i) &= b''(\theta_i)a_i(\phi) = \sigma^2. \end{aligned}$$

Try to generalize this result to the case where Y_i has a normal distribution with mean μ_i and variance σ^2/n_i for known constants n_i, as would be the case if the Y_i represented sample means.□

Example: In Problem Set 1 you will show that the exponential distribution with density

$$f(y_i) = \lambda_i \exp\{-\lambda_i y_i\}$$

belongs to the exponential family.□

In Sections B.4 and B.5 we verify that the binomial and Poisson distributions also belong to this family.

B.1.2 The Link Function

The second element of the generalization is that instead of modeling the mean, as before, we will introduce a one-to-one continuous differentiable transformation $g(\mu_i)$ and focus on

$$\eta_i = g(\mu_i). \tag{B.4}$$

The function $g(\mu_i)$ will be called the *link* function. Examples of link functions include the identity, log, reciprocal, logit and probit.

We further assume that the transformed mean follows a linear model, so that

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta}. \tag{B.5}$$

The quantity η_i is called the *linear predictor*. Note that the model for η_i is pleasantly simple. Since the link function is one-to-one we can invert it to obtain

$$\mu_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}).$$

The model for μ_i is usually more complicated than the model for η_i.

Note that we do not transform the response y_i, but rather its expected value μ_i. A model where $\log y_i$ is linear on x_i, for example, is not the same as a generalized linear model where $\log \mu_i$ is linear on x_i.

Example: The standard linear model we have studied so far can be described as a generalized linear model with normal errors and identity link, so that

$$\eta_i = \mu_i.$$

It also happens that μ_i, and therefore η_i, is the same as θ_i, the parameter in the exponential family density.□

When the link function makes the linear predictor η_i the same as the canonical parameter θ_i, we say that we have a *canonical link*. The identity is the canonical link for the normal distribution. In later sections we will see that the logit is the canonical link for the binomial distribution and the log is the canonical link for the Poisson distribution. This leads to some natural pairings:

Error	Link
Normal	Identity
Binomial	Logit
Poisson	Log

However, other combinations are also possible. An advantage of canonical links is that a minimal sufficient statistic for $\boldsymbol{\beta}$ exists, i.e. all the information about $\boldsymbol{\beta}$ is contained in a function of the data of the same dimensionality as $\boldsymbol{\beta}$.

B.2 Maximum Likelihood Estimation

An important practical feature of generalized linear models is that they can all be fit to data using the same algorithm, a form of *iteratively re-weighted least squares*. In this section we describe the algorithm.

Given a trial estimate of the parameters $\hat{\boldsymbol{\beta}}$, we calculate the estimated linear predictor $\hat{\eta}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ and use that to obtain the fitted values $\hat{\mu_i} = g^{-1}(\hat{\eta}_i)$. Using these quantities, we calculate the working dependent variable

$$z_i = \hat{\eta}_i + (y_i - \hat{\mu_i})\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i}, \tag{B.6}$$

where the rightmost term is the derivative of the link function evaluated at the trial estimate.

Next we calculate the iterative weights

$$w_i = p_i/[b''(\theta_i)(\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i})^2], \tag{B.7}$$

where $b''(\theta_i)$ is the second derivative of $b(\theta_i)$ evaluated at the trial estimate and we have assumed that $a_i(\phi)$ has the usual form ϕ/p_i. This weight is inversely proportional to the variance of the working dependent variable z_i given the current estimates of the parameters, with proportionality factor ϕ.

Finally, we obtain an improved estimate of $\boldsymbol{\beta}$ regressing the working dependent variable z_i on the predictors $\mathbf{x}_i$ using the weights w_i, i.e. we calculate the weighted least-squares estimate

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}, \tag{B.8}$$

where $\mathbf{X}$ is the model matrix, $\mathbf{W}$ is a diagonal matrix of weights with entries w_i given by (B.7) and $\mathbf{z}$ is a response vector with entries z_i given by (B.6).

The procedure is repeated until successive estimates change by less than a specified small amount. McCullagh and Nelder (1989) prove that this algorithm is equivalent to Fisher scoring and leads to maximum likelihood estimates. These authors consider the case of general $a_i(\phi)$ and include ϕ in their expression for the iterative weight. In other words, they use $w_i^* = \phi w_i$, where w_i is the weight used here. The proportionality factor ϕ cancels out when you calculate the weighted least-squares estimates using (B.8), so the estimator is exactly the same. I prefer to show ϕ explicitly rather than include it in $\mathbf{W}$.

Example: For normal data with identity link $\eta_i = \mu_i$, so the derivative is $d\eta_i/d\mu_i = 1$ and the working dependent variable is y_i itself. Since in addition $b''(\theta_i) = 1$ and $p_i = 1$, the weights are constant and no iteration is required.□

In Sections B.4 and B.5 we derive the working dependent variable and the iterative weights required for binomial data with link logit and for Poisson data with link log. In both cases iteration will usually be necessary.

Starting values may be obtained by applying the link to the data, i.e. we take $\hat{\mu}_i = y_i$ and $\hat{\eta}_i = g(\hat{\mu}_i)$. Sometimes this requires a few adjustments, for example to avoid taking the log of zero, and we will discuss these at the appropriate time.

B.3 Tests of Hypotheses

We consider Wald tests and likelihood ratio tests, introducing the *deviance* statistic.

B.3.1 Wald Tests

The Wald test follows immediately from the fact that the information matrix for generalized linear models is given by

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{W}\mathbf{X}/\phi, \tag{B.9}$$

so the large sample distribution of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is multivariate normal

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\phi). \tag{B.10}$$

with mean $\boldsymbol{\beta}$ and variance-covariance matrix $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\phi$.

Tests for subsets of $\boldsymbol{\beta}$ are based on the corresponding marginal normal distributions.

Example: In the case of normal errors with identity link we have $\mathbf{W} = \mathbf{I}$ (where $\mathbf{I}$ denotes the identity matrix), $\phi = \sigma^2$, and the *exact* distribution of $\hat{\boldsymbol{\beta}}$ is multivariate normal with mean $\boldsymbol{\beta}$ and variance-covariance matrix $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$.

B.3.2 Likelihood Ratio Tests and The Deviance

We will show how the likelihood ratio criterion for comparing any two nested models, say $\omega_1 \subset \omega_2$, can be constructed in terms of a statistic called the *deviance* and an unknown scale parameter ϕ.

Consider first comparing a model of interest ω with a *saturated* model Ω that provides a separate parameter for each observation.

Let $\hat{\mu}_i$ denote the fitted values under ω and let $\hat{\theta}_i$ denote the corresponding estimates of the canonical parameters. Similarly, let $\tilde{\mu}_O = y_i$ and $\tilde{\theta}_i$ denote the corresponding estimates under Ω.

The likelihood ratio criterion to compare these two models in the exponential family has the form

$$-2\log\lambda = 2\sum_{i=1}^{n}\frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)}{a_i(\phi)}.$$

Assume as usual that $a_i(\phi) = \phi/p_i$ for known prior weights p_i. Then we can write the likelihood-ratio criterion as follows:

$$-2\log\lambda = \frac{D(\mathbf{y}, \hat{\boldsymbol{\mu}})}{\phi}. \tag{B.11}$$

The numerator of this expression does not depend on unknown parameters and is called the *deviance*:

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} p_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]. \tag{B.12}$$

The likelihood ratio criterion $-2\log L$ is the deviance divided by the scale parameter ϕ, and is called the *scaled deviance*.

Example: Recall that for the normal distribution we had $\theta_i = \mu_i$, $b(\theta_i) = \frac{1}{2}\theta_i^2$, and $a_i(\phi) = \sigma^2$, so the prior weights are $p_i = 1$. Thus, the deviance is

$$\begin{aligned} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) &= 2\sum\{y_i(y_i - \hat{\mu}_i) - \frac{1}{2}y_i^2 + \frac{1}{2}\hat{\mu}_i{}^2\} \\ &= 2\sum\{\frac{1}{2}y_i^2 - y_i\hat{\mu}_i{}^2 + \frac{1}{2}\hat{\mu}_i{}^2\} \\ &= \sum(y_i - \hat{\mu}_i)^2 \end{aligned}$$

our good old friend, the residual sum of squares.□

Let us now return to the comparison of two nested models ω_1, with p_1 parameters, and ω_2, with p_2 parameters, such that $\omega_1 \in \omega_2$ and $p_2 > p1$.

The log of the ratio of maximized likelihoods under the two models can be written as a difference of deviances, since the maximized log-likelihood under the saturated model cancels out. Thus, we have

$$-2\log\lambda = \frac{D(\omega_1) - D(\omega_2)}{\phi} \tag{B.13}$$

The scale parameter ϕ is either known or estimated using the larger model ω_2.

Large sample theory tells us that the asymptotic distribution of this criterion under the usual regularity conditions is χ^2_ν with $\nu = p_2 - p_1$ degrees of freedom.

Example: In the linear model with normal errors we estimate the unknown scale parameter ϕ using the residual sum of squares of the larger model, so the criterion becomes

$$-2\log\lambda = \frac{\mathrm{RSS}(\omega_1) - \mathrm{RSS}(\omega_2)}{\mathrm{RSS}(\omega_2)/(n - p_2)}.$$

In large samples the approximate distribution of this criterion is χ^2_ν with $\nu = p_2 - p_1$ degrees of freedom. Under normality, however, we have an exact result: dividing the criterion by $p_2 - p_1$ we obtain an F with $p_2 - p_1$ and $n - p_2$ degrees of freedom. Note that as $n \to \infty$ the degrees of freedom in the denominator approach ∞ and the F converges to $(p_2 - p_1)\chi^2$, so the asymptotic and exact criteria become equivalent.□

In Sections B.4 and B.5 we will construct likelihood ratio tests for binomial and Poisson data. In those cases $\phi = 1$ (unless one allows overdispersion and estimates ϕ, but that's another story) and the deviance is the same as the scaled deviance. All our tests will be based on asymptotic χ^2 statistics.

B.4 Binomial Errors and Link Logit

We apply the theory of generalized linear models to the case of binary data, and in particular to logistic regression models.

B.4.1 The Binomial Distribution

First we verify that the binomial distribution $B(n_i, \pi_i)$ belongs to the exponential family of Nelder and Wedderburn (1972). The binomial probability distribution function (p.d.f.) is

$$f_i(y_i) = \left(\begin{array}{c} n_i \\ y_i \end{array} \right) \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}. \tag{B.14}$$

Taking logs we find that

$$\log f_i(y_i) = y_i \log(\pi_i) + (n_i - y_i) \log(1-\pi_i) + \log \left(\begin{array}{c} n_i \\ y_i \end{array} \right).$$

Collecting terms on y_i we can write

$$\log f_i(y_i) = y_i \log(\frac{\pi_i}{1-\pi_i}) + n_i \log(1-\pi_i) + \log \left(\begin{array}{c} n_i \\ y_i \end{array} \right).$$

This expression has the general exponential form

$$\log f_i(y_i) = \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)$$

with the following equivalences: Looking first at the coefficient of y_i we note that the canonical parameter is the logit of π_i

$$\theta_i = \log(\frac{\pi_i}{1-\pi_i}). \tag{B.15}$$

Solving for π_i we see that

$$\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}, \quad \text{so} \qquad 1-\pi_i = \frac{1}{1+e^{\theta_i}}.$$

If we rewrite the second term in the p.d.f. as a function of θ_i, so $\log(1-\pi_i) = -\log(1+e^{\theta_i})$, we can identify the cumulant function $b(\theta_i)$ as

$$b(\theta_i) = n_i \log(1+e^{\theta_i}).$$

The remaining term in the p.d.f. is a function of y_i but not π_i, leading to

$$c(y_i, \phi) = \log \left(\begin{array}{c} n_i \\ y_i \end{array} \right).$$

Note finally that we may set $a_i(\phi) = \phi$ and $\phi = 1$.

Let us verify the mean and variance. Differentiating $b(\theta_i)$ with respect to θ_i we find that

$$\mu_i = b'(\theta_i) = n_i \frac{e^{\theta_i}}{1 + e^{\theta_i}} = n_i \pi_i,$$

in agreement with what we knew from elementary statistics. Differentiating again using the quotient rule, we find that

$$v_i = a_i(\phi) b''(\theta_i) = n_i \frac{e^{\theta_i}}{(1 + e^{\theta_i})^2} = n_i \pi_i (1 - \pi_i),$$

again in agreement with what we knew before.

In this development I have worked with the binomial count y_i, which takes values $0(1)n_i$. McCullagh and Nelder (1989) work with the proportion $p_i = y_i/n_i$, which takes values $0(1/n_i)1$. This explains the differences between my results and their Table 2.1.

B.4.2 Fisher Scoring in Logistic Regression

Let us now find the working dependent variable and the iterative weight used in the Fisher scoring algorithm for estimating the parameters in logistic regression, where we model

$$\eta_i = \text{logit}(\pi_i). \tag{B.16}$$

It will be convenient to write the link function in terms of the mean μ_i, as:

$$\eta_i = \log(\frac{\pi_i}{1 - \pi_i}) = \log(\frac{\mu_i}{n_i - \mu_i}),$$

which can also be written as $\eta_i = \log(\mu_i) - \log(n_i - \mu_i)$.

Differentiating with respect to μ_i we find that

$$\frac{d\eta_i}{d\mu_i} = \frac{1}{\mu_i} + \frac{1}{n_i - \mu_i} = \frac{n_i}{\mu_i(n_i - \mu_i)} = \frac{1}{n_i \pi_i (1 - \pi_i)}.$$

The working dependent variable, which in general is

$$z_i = \eta_i + (y_i - \mu_i) \frac{d\eta_i}{d\mu_i},$$

turns out to be

$$z_i = \eta_i + \frac{y_i - n_i\pi_i}{n_i\pi_i(1-\pi_i)}. \tag{B.17}$$

The iterative weight turns out to be

$$\begin{aligned} w_i &= 1/\left[b''(\theta_i)(\frac{d\eta_i}{d\mu_i})^2\right], \\ &= \frac{1}{n_i\pi_i(1-\pi_i)}[n_i\pi_i(1-\pi_i)]^2, \end{aligned}$$

and simplifies to

$$w_i = n_i\pi_i(1-\pi_i). \tag{B.18}$$

Note that the weight is inversely proportional to the variance of the working dependent variable. The results here agree exactly with the results in Chapter 4 of McCullagh and Nelder (1989).

Exercise: Obtain analogous results for Probit analysis, where one models

$$\eta_i = \Phi^{-1}(\mu_i),$$

where $\Phi()$ is the standard normal cdf. *Hint:* To calculate the derivative of the link function find $d\mu_i/d\eta_i$ and take reciprocals.□

B.4.3 The Binomial Deviance

Finally, let us figure out the binomial deviance. Let $\hat{\mu_i}$ denote the m.l.e. of μ_i under the model of interest, and let $\tilde{\mu_i} = y_i$ denote the m.l.e. under the saturated model. From first principles,

$$\begin{aligned} D &= 2\sum[y_i\log(\frac{y_i}{n_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i}) \\ &\quad -y_i\log(\frac{\hat{\mu_i}}{n_i}) - (n_i - y_i)\log(\frac{n_i - \hat{\mu_i}}{n_i})]. \end{aligned}$$

Note that all terms involving $\log(n_i)$ cancel out. Collecting terms on y_i and on $n_i - y_i$ we find that

$$D = 2\sum[y_i\log(\frac{y_i}{\hat{\mu_i}}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \mu_i})]. \tag{B.19}$$

Alternatively, you may obtain this result from the general form of the deviance given in Section B.3.

Note that the binomial deviance has the form

$$D = 2\sum o_i \log(\frac{o_i}{e_i}),$$

where o_i denotes observed, e_i denotes expected (under the model of interest) and the sum is over both "successes" and "failures" for each i (i.e. we have a contribution from y_i and one from $n_i - y_i$).

For grouped data the deviance has an asymptotic chi-squared distribution as $n_i \to \infty$ for all i, and can be used as a goodness of fit test.

More generally, the difference in deviances between nested models (i.e. the log of the likelihood ratio test criterion) has an asymptotic chi-squared distribution as the number of groups $k \to \infty$ or the size of each group $n_i \to \infty$, provided the number of parameters stays fixed.

As a general rule of thumb due to Cochrane (1950), the asymptotic chi-squared distribution provides a reasonable approximation when all *expected* frequencies (both $\hat{\mu}_i$ and $n_i - \hat{\mu}_i$) under the *larger* model exceed one, and at least 80% exceed five.

B.5 Poisson Errors and Link Log

Let us now apply the general theory to the Poisson case, with emphasis on the log link function.

B.5.1 The Poisson Distribution

A Poisson random variable has probability distribution function

$$f_i(y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \tag{B.20}$$

for $y_i = 0, 1, 2, \ldots$. The moments are

$$E(Y_i) = \text{var}(Y_i) = \mu_i.$$

Let us verify that this distribution belongs to the exponential family as defined by Nelder and Wedderburn (1972). Taking logs we find

$$\log f_i(y_i) = y_i \log(\mu_i) - \mu_i - \log(y_i!).$$

Looking at the coefficient of y_i we see immediately that the canonical parameter is

$$\theta_i = \log(\mu_i), \tag{B.21}$$

and therefore that the canonical link is the log. Solving for μ_i we obtain the inverse link

$$\mu_i = e^{\theta_i},$$

and we see that we can write the second term in the p.d.f. as

$$b(\theta_i) = e^{\theta_i}.$$

The last remaining term is a function of y_i only, so we identify

$$c(y_i, \phi) = -\log(y_i!).$$

Finally, note that we can take $a_i(\phi) = \phi$ and $\phi = 1$, just as we did in the binomial case.

Let us verify the mean and variance. Differentiating the cumulant function $b(\theta_i)$ we have

$$\mu_i = b'(\theta_i) = e^{\theta_i} = \mu_i,$$

and differentiating again we have

$$v_i = a_i(\phi)b''(\theta_i) = e^{\theta_i} = \mu_i.$$

Note that the mean equals the variance.

B.5.2 Fisher Scoring in Log-linear Models

We now consider the Fisher scoring algorithm for Poisson regression models with canonical link, where we model

$$\eta_i = \log(\mu_i). \tag{B.22}$$

The derivative of the link is easily seen to be

$$\frac{d\eta_i}{d\mu_i} = \frac{1}{\mu_i}.$$

Thus, the working dependent variable has the form

$$z_i = \eta_i + \frac{y_i - \mu_i}{\mu_i}. \tag{B.23}$$

The iterative weight is

$$\begin{aligned} w_i &= 1/\left[b''(\theta_i)(\frac{d\eta_i}{d\mu_i})^2\right] \\ &= 1/\left[\mu_i \frac{1}{\mu_i^2}\right], \end{aligned}$$

and simplifies to

$$w_i = \mu_i. \tag{B.24}$$

Note again that the weight is inversely proportional to the variance of the working dependent variable.

B.5.3 The Poisson Deviance

Let $\hat{\mu_i}$ denote the m.l.e. of μ_i under the model of interest and let $\tilde{\mu_i} = y_i$ denote the m.l.e. under the saturated model. From first principles, the deviance is

$$\begin{aligned} D &= 2\sum[y_i \log(y_i) - y_i - \log(y_i!) \\ &\quad -y_i \log(\hat{\mu_i}) + \hat{\mu_i} + \log(y_i!)]. \end{aligned}$$

Note that the terms on $y_i!$ cancel out. Collecting terms on y_i we have

$$D = 2\sum[y_i \log(\frac{y_i}{\hat{\mu_i}}) - (y_i - \hat{\mu_i})]. \tag{B.25}$$

The similarity of the Poisson and Binomial deviances should not go unnoticed. Note that the first term in the Poisson deviance has the form

$$D = 2\sum o_i \log(\frac{o_i}{e_i}),$$

which is identical to the Binomial deviance. The second term is usually zero. To see this point, note that for a canonical link the score is

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \mathbf{X}'(\mathbf{y} - \boldsymbol{\mu}),$$

and setting this to zero leads to the estimating equations

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\hat{\boldsymbol{\mu}}.$$

In words, maximum likelihood estimation for Poisson log-linear models—and more generally for any generalized linear model with canonical link—requires equating certain functions of the m.l.e.'s (namely $\mathbf{X}'\hat{\boldsymbol{\mu}}$) to the same functions of the data (namely $\mathbf{X}'\mathbf{y}$). If the model has a constant, one column of $\mathbf{X}$ will consist of ones and therefore one of the estimating equations will be

$$\sum y_i = \sum \hat{\mu_i} \qquad \text{or} \qquad \sum(y_i - \hat{\mu_i}) = 0,$$

so the last term in the Poisson deviance is zero. This result is the basis of an alternative algorithm for computing the m.l.e.'s known as "iterative proportional fitting", see Bishop *et al.* (1975) for a description.

The Poisson deviance has an asymptotic chi-squared distribution as $n \to \infty$ with the number of parameters p remaining fixed, and can be used as a goodness of fit test. Differences between Poisson deviances for nested models (i.e. the log of the likelihood ratio test criterion) have asymptotic chi-squared distributions under the usual regularity conditions.

Appendix A

Review of Likelihood Theory

This is a brief summary of some of the key results we need from likelihood theory.

A.1 Maximum Likelihood Estimation

Let $Y_1, \ldots, Y_n$ be n independent random variables (r.v.'s) with probability density functions (pdf) $f_i(y_i; \boldsymbol{\theta})$ depending on a vector-valued parameter $\boldsymbol{\theta}$.

A.1.1 The Log-likelihood Function

The joint density of n independent observations $\mathbf{y} = (y_1, \ldots, y_n)'$ is

$$f(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i; \boldsymbol{\theta}) = \mathrm{L}(\boldsymbol{\theta}; \mathbf{y}). \tag{A.1}$$

This expression, viewed as a function of the unknown parameter $\boldsymbol{\theta}$ given the data $\mathbf{y}$, is called the *likelihood* function.

Often we work with the natural logarithm of the likelihood function, the so-called *log-likelihood* function:

$$\log \mathrm{L}(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{n} \log f_i(y_i; \boldsymbol{\theta}). \tag{A.2}$$

A sensible way to estimate the parameter $\boldsymbol{\theta}$ given the data $\mathbf{y}$ is to maximize the likelihood (or equivalently the log-likelihood) function, choosing the parameter value that makes the data actually observed as likely as possible. Formally, we define the *maximum-likelihood estimator* (mle) as the value $\hat{\boldsymbol{\theta}}$ such that

$$\log \mathrm{L}(\hat{\boldsymbol{\theta}}; \mathbf{y}) \geq \log \mathrm{L}(\boldsymbol{\theta}; \mathbf{y}) \text{ for all } \boldsymbol{\theta}. \tag{A.3}$$

Example: The Log-Likelihood for the Geometric Distribution. Consider a series of independent Bernoulli trials with common probability of success π. The distribution of the number of *failures* Y_i before the first success has pdf

$$\Pr(Y_i = y_i) = (1-\pi)^{y_i}\pi. \tag{A.4}$$

for $y_i = 0, 1, \ldots$. Direct calculation shows that $\mathrm{E}(Y_i) = (1-\pi)/\pi$.

The log-likelihood function based on n observations $\mathbf{y}$ can be written as

$$\begin{aligned} \log \mathrm{L}(\pi; \mathbf{y}) &= \sum_{i=1}^{n} \{y_i \log(1-\pi) + \log \pi\} & \text{(A.5)} \\ &= n(\bar{y} \log(1-\pi) + \log \pi), & \text{(A.6)} \end{aligned}$$

where $\bar{y} = \sum y_i/n$ is the sample mean. The fact that the log-likelihood depends on the observations only through the sample mean shows that $\bar{y}$ is a *sufficient* statistic for the unknown probability π.

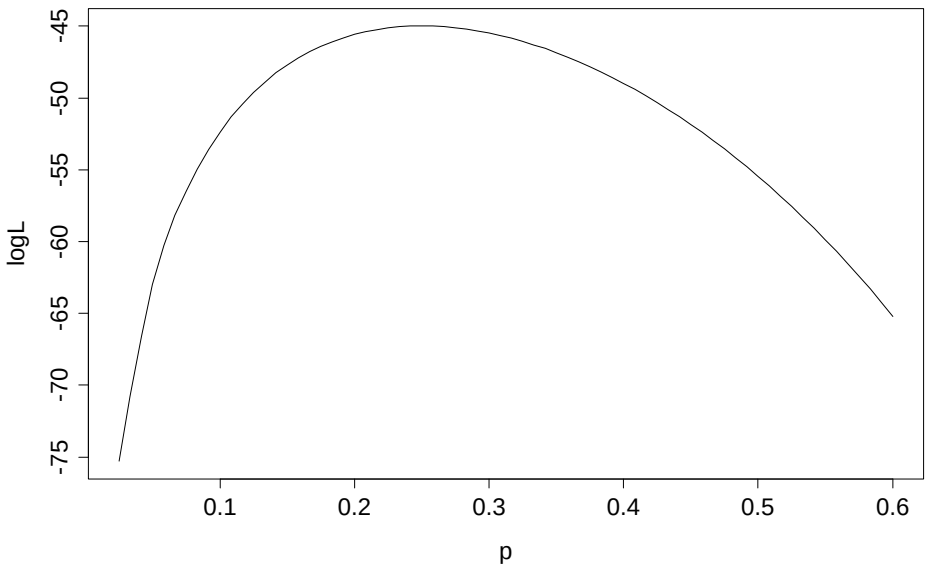


FIGURE A.1: The Geometric Log-Likelihood for $n = 20$ and $\bar{y} = 3$

Figure A.1 shows the log-likelihood function for a sample of $n = 20$ observations from a geometric distribution when the observed sample mean is $\bar{y} = 3$.□

A.1.2 The Score Vector

The first derivative of the log-likelihood function is called Fisher's *score function*, and is denoted by

$$\mathbf{u}(\boldsymbol{\theta}) = \frac{\partial \log \mathrm{L}(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}}. \tag{A.7}$$

Note that the score is a vector of first partial derivatives, one for each element of $\boldsymbol{\theta}$.

If the log-likelihood is concave, one can find the maximum likelihood estimator by setting the score to zero, i.e. by solving the system of equations:

$$\mathbf{u}(\hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{A.8}$$

Example: The Score Function for the Geometric Distribution. The score function for n observations from a geometric distribution is

$$u(\pi) = \frac{\mathrm{d} \log \mathrm{L}}{\mathrm{d}\pi} = n(\frac{1}{\pi} - \frac{\bar{y}}{1-\pi}). \tag{A.9}$$

Setting this equation to zero and solving for π leads to the maximum likelihood estimator

$$\hat{\pi} = \frac{1}{1+\bar{y}}. \tag{A.10}$$

Note that the m.l.e. of the probability of success is the reciprocal of the number of trials. This result is intuitively reasonable: the longer it takes to get a success, the lower our estimate of the probability of success would be.

Suppose now that in a sample of $n = 20$ observations we have obtained a sample mean of $\bar{y} = 3$. The m.l.e. of the probability of success would be $\hat{\pi} = 1/(1+3) = 0.25$, and it should be clear from Figure A.1 that this value maximizes the log-likelihood.

A.1.3 The Information Matrix

The score is a random vector with some interesting statistical properties. In particular, the score evaluated at the true parameter value θ has mean zero

$$\mathrm{E}[\mathbf{u}(\boldsymbol{\theta})] = \mathbf{0}$$

and variance-covariance matrix given by the *information matrix*:

$$\mathrm{var}[\mathbf{u}(\boldsymbol{\theta})] = \mathrm{E}[\mathbf{u}(\boldsymbol{\theta})\mathbf{u}'(\boldsymbol{\theta})] = \mathbf{I}(\boldsymbol{\theta}). \tag{A.11}$$

Under mild regularity conditions, the information matrix can also be obtained as minus the expected value of the second derivatives of the log-likelihood:

$$\mathbf{I}(\boldsymbol{\theta}) = -\mathrm{E}[\frac{\partial^2 \log \mathrm{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}]. \tag{A.12}$$

The matrix of negative observed second derivatives is sometimes called the *observed* information matrix.

Note that the second derivative indicates the extent to which the log-likelihood function is peaked rather than flat. This makes the interpretation in terms of information intuitively reasonable.

Example: Information for the Geometric Distribution. Differentiating the score we find the observed information to be

$$-\frac{\mathrm{d}^2 \log \mathrm{L}}{\mathrm{d}\pi^2} = -\frac{\mathrm{d}u}{\mathrm{d}\pi} = n(\frac{1}{\pi^2} + \frac{\bar{y}}{(1-\pi)^2}). \tag{A.13}$$

To find the expected information we use the fact that the expected value of the sample mean $\bar{y}$ is the population mean $(1-\pi)/\pi$, to obtain (after some simplification)

$$\mathrm{I}(\pi) = \frac{n}{\pi^2(1-\pi)}. \tag{A.14}$$

Note that the information increases with the sample size n and varies with π, increasing as π moves away from $\frac{2}{3}$ towards 0 or 1.

In a sample of size $n = 20$, if the true value of the parameter was $\pi = 0.15$ the expected information would be $I(0.15) = 1045.8$. If the sample mean turned out to be $\bar{y} = 3$, the observed information would be 971.9. Of course, we don't know the true value of π. Substituting the mle $\hat{\pi} = 0.25$, we estimate the expected and observed information as 426.7. □

A.1.4 Newton-Raphson and Fisher Scoring

Calculation of the mle often requires iterative procedures. Consider expanding the score function evaluated at the mle $\hat{\boldsymbol{\theta}}$ around a trial value $\boldsymbol{\theta}_0$ using a first order Taylor series, so that

$$\mathbf{u}(\hat{\boldsymbol{\theta}}) \approx \mathbf{u}(\boldsymbol{\theta}_0) + \frac{\partial \mathbf{u}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0). \tag{A.15}$$

Let $\mathbf{H}$ denote the Hessian or matrix of second derivatives of the log-likelihood function

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2 \log \mathrm{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \frac{\partial \mathbf{u}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{A.16}$$

Setting the left-hand-size of Equation A.15 to zero and solving for $\hat{\boldsymbol{\theta}}$ gives the first-order approximation

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - \mathbf{H}^{-1}(\boldsymbol{\theta}_0)\mathbf{u}(\boldsymbol{\theta}_0). \tag{A.17}$$

This result provides the basis for an iterative approach for computing the mle known as the *Newton-Raphson* technique. Given a trial value, we use Equation A.17 to obtain an improved estimate and repeat the process until differences between successive estimates are sufficiently close to zero. (Or until the elements of the vector of first derivatives are sufficiently close to zero.) This procedure tends to converge quickly if the log-likelihood is well-behaved (close to quadratic) in a neighborhood of the maximum and if the starting value is reasonably close to the mle.

An alternative procedure first suggested by Fisher is to replace minus the Hessian by its expected value, the information matrix. The resulting procedure takes as our improved estimate

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + \mathbf{I}^{-1}(\boldsymbol{\theta}_0)\mathbf{u}(\boldsymbol{\theta}_0), \tag{A.18}$$

and is known as *Fisher Scoring*.

Example: Fisher Scoring in the Geometric Distribution. In this case setting the score to zero leads to an explicit solution for the mle and no iteration is needed. It is instructive, however, to try the procedure anyway. Using the results we have obtained for the score and information, the Fisher scoring procedure leads to the updating formula

$$\hat{\pi} = \pi_0 + (1 - \pi_0 - \pi_0\bar{y})\pi_0. \tag{A.19}$$

If the sample mean is $\bar{y} = 3$ and we start from $\pi_0 = 0.1$, say, the procedure converges to the mle $\hat{\pi} = 0.25$ in four iterations. □

A.2 Tests of Hypotheses

We consider three different types of tests of hypotheses.

A.2.1 Wald Tests

Under certain regularity conditions, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ has approximately in large samples a (multivariate) normal distribution with mean equal to the true parameter value and variance-covariance matrix given by the inverse of the information matrix, so that

$$\hat{\boldsymbol{\theta}} \sim N_p(\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta})). \tag{A.20}$$

The regularity conditions include the following: the true parameter value $\boldsymbol{\theta}$ must be interior to the parameter space, the log-likelihood function must be thrice differentiable, and the third derivatives must be bounded.

This result provides a basis for constructing tests of hypotheses and confidence regions. For example under the hypothesis

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \tag{A.21}$$

for a fixed value $\boldsymbol{\theta}_0$, the quadratic form

$$W = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \text{var}^{-1}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{A.22}$$

has approximately in large samples a chi-squared distribution with p degrees of freedom.

This result can be extended to arbitrary linear combinations of $\boldsymbol{\theta}$, including sets of elements of $\boldsymbol{\theta}$. For example if we partition $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$, where $\boldsymbol{\theta}_2$ has p_2 elements,then we can test the hypothesis that the last p_2 parameters are zero

$$H_o : \boldsymbol{\theta}_2 = 0,$$

by treating the quadratic form

$$W = \hat{\boldsymbol{\theta}}_2{}' \, \text{var}^{-1}(\hat{\boldsymbol{\theta}}_2) \, \hat{\boldsymbol{\theta}}_2$$

as a chi-squared statistic with p_2 degrees of freedom. When the subset has only one element we usually take the square root of the Wald statistic and treat the ratio

$$z = \frac{\hat{\theta}_j}{\sqrt{\text{var}(\hat{\theta}_j)}}$$

as a z-statistic (or a t-ratio).

These results can be modified by replacing the variance-covariance matrix of the mle with any consistent estimator. In particular, we often use the inverse of the expected information matrix evaluated at the mle

$$\widehat{\text{var}}(\hat{\boldsymbol{\theta}}) = \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}).$$

Sometimes calculation of the expected information is difficult, and we use the observed information instead.

Example: Wald Test in the Geometric Distribution. Consider again our sample of $n = 20$ observations from a geometric distribution with sample mean $\bar{y} = 3$. The mle was $\hat{\pi} = 0.25$ and its variance, using the estimated expected information, is $1/426.67 = 0.00234$. Testing the hypothesis that the true probability is $\pi = 0.15$ gives

$$\chi^2 = (0.25 - 0.15)^2/0.00234 = 4.27$$

with one degree of freedom. The associated p-value is 0.039, so we would reject H_0 at the 5% significance level. □

A.2.2 Score Tests

Under some regularity conditions the score itself has an asymptotic normal distribution with mean 0 and variance-covariance matrix equal to the information matrix, so that

$$\mathbf{u}(\boldsymbol{\theta}) \sim N_p(0, \mathbf{I}(\boldsymbol{\theta})). \tag{A.23}$$

This result provides another basis for constructing tests of hypotheses and confidence regions. For example under

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$$

the quadratic form

$$Q = \mathbf{u}(\boldsymbol{\theta}_0)' \, \mathbf{I}^{-1}(\boldsymbol{\theta}_0) \, \mathbf{u}(\boldsymbol{\theta}_0)$$

has approximately in large samples a chi-squared distribution with p degrees of freedom.

The information matrix may be evaluated at the hypothesized value $\boldsymbol{\theta}_0$ or at the mle $\hat{\boldsymbol{\theta}}$. Under H_0 both versions of the test are valid; in fact, they are asymptotically equivalent. One advantage of using $\boldsymbol{\theta}_0$ is that calculation of the mle may be bypassed. In spite of their simplicity, score tests are rarely used.

Example: Score Test in the Geometric Distribution. Continuing with our example, let us calculate the score test of $H_0 : \pi = 0.15$ when $n = 20$ and $\bar{y} = 3$. The score evaluated at 0.15 is $u(0.15) = -62.7$, and the expected information evaluated at 0.15 is $\mathbf{I}(0.15) = 1045.8$, leading to

$$\chi^2 = 62.7^2/1045.8 = 3.76$$

with one degree of freedom. Since the 5% critical value is $\chi^2_{1,0.95} = 3.84$ we would accept H_0 (just). □

A.2.3 Likelihood Ratio Tests

The third type of test is based on a comparison of maximized likelihoods for nested models. Suppose we are considering two models, ω_1 and ω_2, such that $\omega_1 \subset \omega_2$. In words, ω_1 is a subset of (or can be considered a special case of) ω_2. For example, one may obtain the simpler model ω_1 by setting some of the parameters in ω_2 to zero, and we want to test the hypothesis that those elements are indeed zero.

The basic idea is to compare the maximized likelihoods of the two models. The maximized likelihood under the smaller model ω_1 is

$$\max_{\boldsymbol{\theta}\in\omega_1} \mathrm{L}(\boldsymbol{\theta}, \mathbf{y}) = \mathrm{L}(\hat{\boldsymbol{\theta}}_{\omega_1}, \mathbf{y}), \tag{A.24}$$

where $\hat{\boldsymbol{\theta}}_{\omega_1}$ denotes the mle of $\boldsymbol{\theta}$ under model ω_1.

The maximized likelihood under the larger model ω_2 has the same form

$$\max_{\boldsymbol{\theta}\in\omega_2} \mathrm{L}(\boldsymbol{\theta}, \mathbf{y}) = \mathrm{L}(\hat{\boldsymbol{\theta}}_{\omega_2}, \mathbf{y}), \tag{A.25}$$

where $\hat{\boldsymbol{\theta}}_{\omega_2}$ denotes the mle of $\boldsymbol{\theta}$ under model ω_2.

The ratio of these two quantities,

$$\lambda = \frac{\mathrm{L}(\hat{\boldsymbol{\theta}}_{\omega_1}, \mathbf{y})}{\mathrm{L}(\hat{\boldsymbol{\theta}}_{\omega_2}, \mathbf{y})}, \tag{A.26}$$

is bound to be between 0 (likelihoods are non-negative) and 1 (the likelihood of the smaller model can't exceed that of the larger model because it is *nested* on it). Values close to 0 indicate that the smaller model is not acceptable, compared to the larger model, because it would make the observed data very unlikely. Values close to 1 indicate that the smaller model is almost as good as the large model, making the data just as likely.

Under certain regularity conditions, minus twice the log of the likelihood ratio has approximately in large samples a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models. Thus,

$$-2\log\lambda = 2\log \mathrm{L}(\hat{\theta}_{\omega_2}, y) - 2\log \mathrm{L}(\hat{\theta}_{\omega_1}, y) \to \chi^2_\nu, \tag{A.27}$$

where the degrees of freedom are $\nu = \dim(\omega_2) - \dim(\omega_1)$, the number of parameters in the larger model ω_2 minus the number of parameters in the smaller model ω_1.

Note that calculation of a likelihood ratio test requires fitting two models (ω_1 and ω_2), compared to only one model for the Wald test (ω_2) and sometimes no model at all for the score test.

Example: Likelihood Ratio Test in the Geometric Distribution. Consider testing $H_0 : \pi = 0.15$ with a sample of $n = 20$ observations from a geometric distribution, and suppose the sample mean is $\bar{y} = 3$. The value of the likelihood under H_0 is $\log \mathrm{L}(0.15) = -47.69$. Its unrestricted maximum value, attained at the mle, is $\log \mathrm{L}(0.25) = -44.98$. Minus twice the difference between these values is

$$\chi^2 = 2(47.69 - 44.99) = 5.4$$

with one degree of freedom. This value is significant at the 5% level and we would reject H_0. Note that in our example the Wald, score and likelihood ratio tests give similar, but not identical, results. □

The three tests discussed in this section are asymptotically equivalent, and are therefore expected to give similar results in large samples. Their small-sample properties are not known, but some simulation studies suggest that the likelihood ratio test may be better that its competitors in small samples.

21:59 Thursday 18th February, 2016

Chapter 6

The Bootstrap

We are now several chapters into a statistics class and have said basically nothing about uncertainty. This should seem odd, and may even be disturbing if you are very attached to your p-values and saying variables have "significant effects". It is time to remedy this, and talk about how we can quantify uncertainty for complex models. The key technique here is what's called **bootstrapping**, or **the bootstrap**.

6.1 Stochastic Models, Uncertainty, Sampling Distributions

Statistics is the branch of mathematical engineering which studies ways of drawing inferences from limited and imperfect data. We want to know how a neuron in a rat's brain responds when one of its whiskers gets tweaked, or how many rats live in Pittsburgh, or how high the water will get under the 16th Street bridge during May, or the typical course of daily temperatures in the city over the year, or the relationship between the number of birds of prey in Schenley Park in the spring and the number of rats the previous fall. We have some data on all of these things. But we know that our data is incomplete, and experience tells us that repeating our experiments or observations, even taking great care to replicate the conditions, gives more or less different answers every time. It is foolish to treat any inference from the data in hand as certain.

If all data sources were totally capricious, there'd be nothing to do beyond piously qualifying every conclusion with "but we could be wrong about this". A mathematical discipline of statistics is possible because while repeating an experiment gives different results, some kinds of results are more common than others; their relative frequencies are reasonably stable. We thus model the data-generating mechanism through probability distributions and stochastic processes. When and why we can use stochastic models are very deep questions, but ones for another time. If we *can* use them in our problem, quantities like the ones I mentioned above are represented as functions of the stochastic model, i.e., of the underlying probability distribution.

Since a function of a function is a "functional", and these quantities are functions of the true probability distribution function, we'll call these **functionals** or **statistical functionals**[1]. Functionals could be single numbers (like the total rat population), or vectors, or even whole curves (like the expected time-course of temperature over the year, or the regression of hawks now on rats earlier). Statistical inference becomes estimating those functionals, or testing hypotheses about them.

These estimates and other inferences are functions of the data values, which means that they inherit variability from the underlying stochastic process. If we "re-ran the tape" (as the late, great Stephen Jay Gould used to say), we would get different data, with a certain characteristic distribution, and applying a fixed procedure would yield different inferences, again with a certain distribution. Statisticians want to use this distribution to quantify the uncertainty of the inferences. For instance, the standard error is an answer to the question "By how much would our estimate of this functional vary, typically, from one replication of the experiment to another?" (It presumes a particular meaning for "typically vary", as the root-mean-square deviation around the mean.) A confidence region on a parameter, likewise, is the answer to "What are all the values of the parameter which *could* have produced this data with at least some specified probability?", i.e., all the parameter values under which our data are not low-probability outliers. The confidence region is a promise that *either* the true parameter point lies in that region, *or* something very unlikely under any circumstances happened — or that our stochastic model is wrong.

[[Cross-ref hypothesis testing appendix, when it's written]]

To get things like standard errors or confidence intervals, we need to know the distribution of our estimates around the true values of our functionals. These **sampling distributions** follow, remember, from the distribution of the data, since our estimates are functions of the data. Mathematically the problem is well-defined, but actually *computing* anything is another story. Estimates are typically complicated functions of the data, and mathematically-convenient distributions may all be poor approximations to the data source. Saying anything in closed form about the distribution of estimates can be simply hopeless. The two classical responses of statisticians were to focus on tractable special cases, and to appeal to asymptotics.

Your introductory statistics courses mostly drilled you in the special cases. From one side, limit the kind of estimator we use to those with a simple mathematical form — say, means and other linear functions of the data. From the other, assume that the probability distributions featured in the stochastic model take one of a few forms for which exact calculation *is* possible, analytically or via tabulated special functions. Most such distributions have origin myths: the Gaussian arises from averaging many independent variables of equal size (say, the many genes which contribute to height in humans); the Poisson distribution comes from counting how many of a large number of independent and individually-improbable events have occurred (say, radioactive nuclei decaying in a given second), etc. Squeezed from both ends, the sampling distribution of estimators and other functions of the data becomes exactly calculable in terms of the aforementioned special functions.

That these origin myths invoke various limits is no accident. The great results

[1] Most writers in theoretical statistics just call them "parameters" in a generalized sense, but I will try to restrict that word to actual parameters specifying statistical models, to minimize confusion. I may slip up.

of probability theory — the laws of large numbers, the ergodic theorem, the central limit theorem, etc. — describe limits in which *all* stochastic processes in broad classes of models display the same asymptotic behavior. The central limit theorem, for instance, says that if we average more and more independent random quantities with a common distribution, and that common distribution isn't too pathological, then the average becomes closer and closer to a Gaussian[2]. Typically, as in the CLT, the limits involve taking more and more data from the source, so statisticians use the theorems to find the asymptotic, large-sample distributions of their estimates. We have been especially devoted to re-writing our estimates as averages of independent quantities, so that we can use the CLT to get Gaussian asymptotics.

Up through about the 1960s, statistics was split between developing general ideas about how to draw and evaluate inferences with stochastic models, and working out the properties of inferential procedures in tractable special cases (especially the linear-and-Gaussian case), or under asymptotic approximations. This yoked a very broad and abstract theory of inference to very narrow and concrete practical formulas, an uneasy combination often preserved in basic statistics classes.

The arrival of (comparatively) cheap and fast computers made it feasible for scientists and statisticians to record lots of data and to fit models to it, so they did. Sometimes the models were conventional ones, including the special-case assumptions, which often enough turned out to be detectably, and consequentially, wrong. At other times, scientists wanted more complicated or flexible models, some of which had been proposed long before, but now moved from being theoretical curiosities to stuff that could run overnight[3]. In principle, asymptotics might handle either kind of problem, but convergence to the limit could be unacceptably slow, especially for more complex models.

By the 1970s, then, statistics faced the problem of quantifying the uncertainty of inferences without using either implausibly-helpful assumptions or asymptotics; all of the solutions turned out to demand *even more* computation. Here we will examine what may be the most successful solution, Bradley Efron's proposal to combine estimation with simulation, which he gave the less-than-clear but persistent name of "the bootstrap" (Efron, 1979).

6.2 The Bootstrap Principle

Remember (from baby stats.) that the key to dealing with uncertainty in parameters and functionals is the sampling distribution of estimators. Knowing what distribution we'd get for our estimates on repeating the experiment would give us things like standard errors. Efron's insight was that we can *simulate* replication. After all, we have already fitted a model to the data, which is a guess at the mechanism which generated the data. Running that mechanism generates simulated data which, by hypothesis, has the same distribution as the real data. Feeding the simulated data through

[2]The reason is that the non-Gaussian parts of the distribution wash away under averaging, but the average of two Gaussians is another Gaussian.

[3]Kernel regression (§1.5.2), kernel density estimation (Ch. 14), and nearest-neighbors prediction (§1.5.1) were all proposed in the 1950s or 1960s, but didn't begin to be widely used until about 1980.

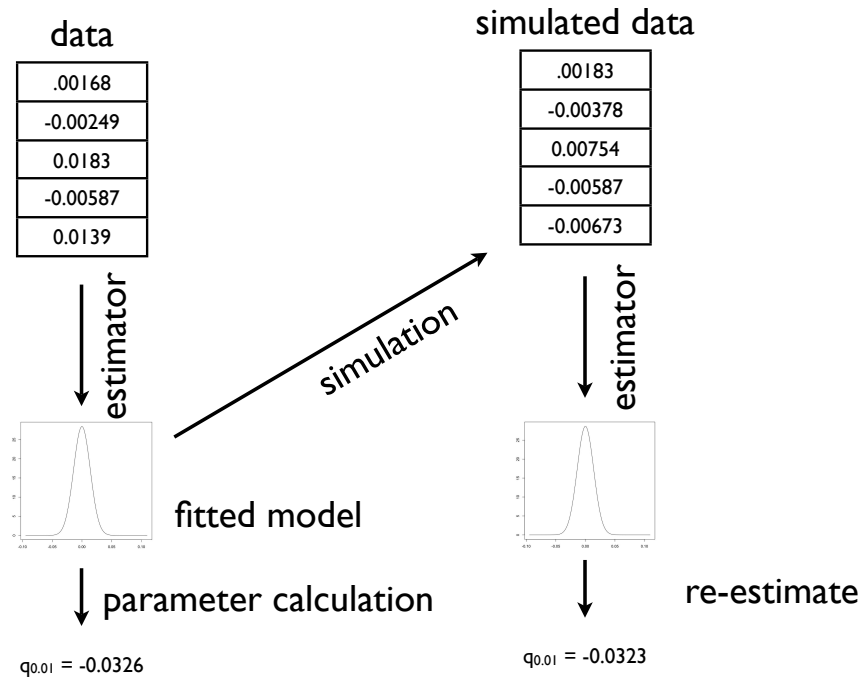


FIGURE 6.1: *Schematic for model-based bootstrapping: simulated values are generated from the fitted model, then treated like the original data, yielding a new estimate of the functional of interest, here called* $q_{0.01}$.

our estimator gives us one draw from the sampling distribution; repeating this many times yields the sampling distribution. Since we are using the model to give us its own uncertainty, Efron called this "bootstrapping"; unlike the Baron Munchhausen's plan for getting himself out of a swamp by pulling on his own bootstraps, it works.

Figure 6.1 sketches the over-all process: fit a model to data, use the model to calculate the functional, then get the sampling distribution by generating new, synthetic data from the model and repeating the estimation on the simulation output.

To fix notation, we'll say that the original data is x. (In general this is a whole data frame, not a single number.) Our parameter estimate from the data is $\hat{\theta}$. Surrogate data sets simulated from the fitted model will be $\tilde{X}_1, \tilde{X}_2, \ldots \tilde{X}_B$. The corresponding re-estimates of the parameters on the surrogate data are $\tilde{\theta}_1, \tilde{\theta}_2, \ldots \tilde{\theta}_B$. The functional of interest is estimated by the statistic[4] T, with sample value $\hat{t} = T(x)$, and values of the surrogates of $\tilde{t}_1 = T(\tilde{X}_1)$, $\tilde{t}_2 = T(\tilde{X}_2)$, $\ldots \tilde{t}_B = T(\tilde{X}_B)$. (The statistic T may be a direct function of the estimated parameters, and only indirectly a function of x.) Everything which follows applies without modification when the functional of interest *is* the parameter, or some component of the parameter.

[4] T is a common symbol in the literature on the bootstrap for a *generic* function of the data. It may or may not have anything to do with Student's t test for difference in means.

In this section, we will assume that the model is correct for *some* value of θ, which we will call θ_0. This means that we are employing a parametric **model-based bootstrap**. The true (population or ensemble) values of the functional is likewise t_0.

6.2.1 Variances and Standard Errors

The simplest thing to do is to get the variance or standard error:

$$\widehat{\mathrm{Var}}\left[\hat{t}\right] = \mathbb{V}\left[\tilde{t}\right] \tag{6.1}$$

$$\widehat{\mathrm{se}}(\hat{t}) = \mathrm{sd}(\tilde{t}) \tag{6.2}$$

That is, we approximate the variance of our estimate of t_0 under the true but unknown distribution θ_0 by the variance of re-estimates $\tilde{t}$ on surrogate data from the fitted model $\widehat{\theta}$. Similarly we approximate the true standard error by the standard deviation of the re-estimates. The logic here is that the simulated $\tilde{X}$ has about the same distribution as the real X that our data, x, was drawn from, so applying the same estimation procedure to the surrogate data gives us the sampling distribution. This assumes, of course, that our model is right, and that $\hat{\theta}$ is not too far from θ_0.

A code sketch is provided in Code Example 6. Note that this may not work *exactly* as given in some circumstances, depending on the syntax details of, say, just what kind of data structure is needed to store $\hat{t}$.

6.2.2 Bias Correction

We can use bootstrapping to correct for a biased estimator. Since the sampling distribution of $\tilde{t}$ is close to that of $\widehat{t}$, and $\widehat{t}$ itself is close to t_0,

$$\mathbb{E}\left[\widehat{t}\right] - t_0 \approx \mathbb{E}\left[\tilde{t}\right] - \widehat{t} \tag{6.3}$$

The left hand side is the bias that we want to know, and the right-hand side the was what we can calculate with the bootstrap.

In fact, Eq. 6.3 remains valid so long as the sampling distribution of $\widehat{t} - t_0$ is close to that of $\tilde{t} - \widehat{t}$. This is a weaker requirement than asking for $\widehat{t}$ and $\tilde{t}$ themselves to have similar distributions, or asking for $\widehat{t}$ to be close to t_0. In statistical theory, a random variable whose distribution does not depend on the parameters is called a **pivot**. (The metaphor is that it stays in one place while the parameters turn around it.) A sufficient (but not necessary) condition for Eq. 6.3 to hold is that $\widehat{t} - t_0$ be a pivot, or approximately pivotal.

6.2.3 Confidence Intervals

A confidence interval is a random interval which contains the truth with high probability (the confidence level). If the confidence interval for g is C, and the confidence level is $1-\alpha$, then we want

$$\Pr(t_0 \in C) = 1 - \alpha \tag{6.4}$$

```
rboot <- function(statistic, simulator, B) {
    tboots <- replicate(B, statistic(simulator()))
    if (is.null(dim(tboots))) {
        tboots <- array(tboots, dim = c(1, B))
    }
    return(tboots)
}
bootstrap <- function(tboots, summarizer, ...) {
    summaries <- apply(tboots, 1, summarizer, ...)
    return(t(summaries))
}
bootstrap.se <- function(statistic, simulator, B) {
    bootstrap(rboot(statistic, simulator, B), summarizer = sd)
}
```

CODE EXAMPLE 6: *Code for calculating bootstrap standard errors. The function* `rboot` *generates* `B` *bootstrap samples (using the* `simulator` *function) and calculates the statistic on them (using* `statistic`*).* `simulator` *needs to be a function which returns a surrogate data set in a form suitable for* `statistic`*. (How would you modify the code to pass arguments to* `simulator` *and/or* `statistic`*?) Because every use of bootstrapping is going to need to do this, it makes sense to break it out as a separate function, rather than writing the same code many times (with many chances of getting it wrong). The* `bootstrap` *function takes the output of* `rboot` *and applies a summarizing function.* `bootstrap.se` *just calls* `rboot` *and makes the summarizing function* `sd`*, which takes a standard deviation.* IMPORTANT NOTE: *This is just a code* sketch*, because depending on the data structure which the statistic returns, it may not (e.g.) be feasible to just run* `sd` *on it, and so it might need some modification. See detailed examples below.*

```
bootstrap.bias <- function(simulator, statistic, B, t.hat) {
    expect <- bootstrap(rboot(statistic, simulator, B), summarizer = mean)
    return(expect - t.hat)
}
```

CODE EXAMPLE 7: *Sketch of code for bootstrap bias correction. Arguments are as in Code Example 6, except that* `t.hat` *is the estimate on the original data.* IMPORTANT NOTE: *As with Code Example 6, this is just a code* sketch*, because it won't work with all data types that might be returned by* `statistic`*, and so might require modification.*

no matter what the true value of t_0. When we calculate a confidence interval, our inability to deal with distributions exactly means that the true confidence level, or **coverage** of the interval, is not quite the desired confidence level $1-\alpha$; the closer it is, the better the approximation, and the more accurate the confidence interval.[5]

When we simulate, we get samples of $\tilde{t}$, but what we really care about is the distribution of $\hat{t}$. When we have enough data to start with, those two distributions will be approximately the same. But at any given amount of data, the distribution of $\tilde{t} - \hat{t}$ will usually be closer to that of $\hat{t} - t_0$ than the distribution of $\tilde{t}$ is to that of $\hat{t}$. That is, the distribution of fluctuations around the true value usually converges quickly. (Think of the central limit theorem.) We can use this to turn information about the distribution of $\tilde{t}$ into accurate confidence intervals for t_0, essentially by re-centering $\tilde{t}$ around $\hat{t}$.

Specifically, let $q_{\alpha/2}$ and $q_{1-\alpha/2}$ be the $\alpha/2$ and $1-\alpha/2$ quantiles of $\tilde{t}$. Then

$$
\begin{aligned}
1-\alpha &= \Pr\left(q_{\alpha/2} \leq \tilde{T} \leq q_{1-\alpha/2}\right) && (6.5)\\
&= \Pr\left(q_{\alpha/2} - \hat{T} \leq \tilde{T} - \hat{T} \leq q_{1-\alpha/2} - \hat{T}\right) && (6.6)\\
&\approx \Pr\left(q_{\alpha/2} - \hat{T} \leq \hat{T} - t_0 \leq q_{1-\alpha/2} - \hat{T}\right) && (6.7)\\
&= \Pr\left(q_{\alpha/2} - 2\hat{T} \leq -t_0 \leq q_{1-\alpha/2} - 2\hat{T}\right) && (6.8)\\
&= \Pr\left(2\hat{T} - q_{1-\alpha/2} \leq t_0 \leq 2\hat{T} - q_{\alpha/2}\right) && (6.9)
\end{aligned}
$$

The interval $C = [2\hat{T} - q_{\alpha/2}, 2\hat{T} - q_{1-\alpha/2}]$ is random, because $\hat{T}$ is a random quantity, so it makes sense to talk about the probability that it contains the true value t_0. Also, notice that the upper and lower quantiles of $\tilde{T}$ have, as it were, swapped roles in determining the upper and lower confidence limits. Finally, notice that we do not actually know those quantiles exactly, but they're what we approximate by bootstrapping.

This is the **basic bootstrap confidence interval**, or the **pivotal** CI. It is simple and reasonably accurate, and makes a very good default choice for finding confidence intervals.

[[ATTN: Add subsection specifically on confidence bands?]]

6.2.3.1 Other Bootstrap Confidence Intervals

The basic bootstrap CI relies on the distribution of $\tilde{t} - \hat{t}$ being approximately the same as that of $\hat{t} - t_0$. Even when this is false, however, it can be that the distribution of

$$\tau = \frac{\hat{t} - t_0}{\widehat{\mathrm{se}}(\hat{t})} \tag{6.10}$$

[5] You might wonder why we'd be unhappy if the coverage level was *greater* than $1-\alpha$. This is certainly better than if it's *less* than the nominal confidence level, but it usually means we could have used a smaller set, and so been more precise about t_0, without any more real risk. Confidence intervals whose coverage is greater than the nominal level are called **conservative**; those with less than nominal coverage are **anti-conservative** (and not, say, "liberal").

```
equitails <- function(x, alpha) {
    lower <- quantile(x, alpha/2)
    upper <- quantile(x, 1 - alpha/2)
    return(c(lower, upper))
}
bootstrap.ci <- function(statistic = NULL, simulator = NULL, tboots = NULL,
    B = if (!is.null(tboots)) {
        ncol(tboots)
    }, t.hat, level) {
    if (is.null(tboots)) {
        stopifnot(!is.null(statistic))
        stopifnot(!is.null(simulator))
        stopifnot(!is.null(B))
        tboots <- rboot(statistic, simulator, B)
    }
    alpha <- 1 - level
    intervals <- bootstrap(tboots, summarizer = equitails, alpha = alpha)
    upper <- t.hat + (t.hat - intervals[, 1])
    lower <- t.hat + (t.hat - intervals[, 2])
    CIs <- cbind(lower = lower, upper = upper)
    return(CIs)
}
```

CODE EXAMPLE 8: *Sketch of code for calculating the basic bootstrap confidence interval. See Code Example 6 for* `rboot` *and* `bootstrap`*, and cautions about blindly applying this to arbitrary data-types. See online for comments.*

is close to that of

$$\tilde{\tau} = \frac{\tilde{t} - \hat{t}}{\mathrm{se}(\tilde{t})} \tag{6.11}$$

This is like what we calculate in a t-test, and since the t-test was invented by "Student", these are called **studentized** quantities. If τ and $\tilde{\tau}$ have the same distribution, then we can reason as above and get a confidence interval

$$\left(\hat{t} - \widehat{\mathrm{se}}(\hat{t}) Q_{\tilde{\tau}}(1-\alpha/2), \hat{t} - \widehat{\mathrm{se}}(\hat{t}) Q_{\tilde{\tau}}(\alpha/2)\right) \tag{6.12}$$

This is the same as the basic interval when $\widehat{\mathrm{se}}(\hat{t}) = \mathrm{se}(\tilde{t})$, but different otherwise. To find $\mathrm{se}(\tilde{t})$, we need to actually do a *second* level of bootstrapping, as follows.

1. Fit the model with $\hat{\theta}$, find $\hat{t}$.
2. For $i \in 1 : B_1$
   (a) Generate $\tilde{X}_i$ from $\hat{\theta}$
   (b) Estimate $\tilde{\theta}_i$, $\tilde{t}_i$
   (c) For $j \in 1 : B_2$
      i. Generate $X_{ij}^{\dagger}$ from $\tilde{\theta}_i$
      ii. Calculate $t_{ij}^{\dagger}$
   (d) Set $\tilde{\sigma}_i =$ standard deviation of the $t_{ij}^{\dagger}$
   (e) Set $\tilde{\tau}_{ij} = \frac{t_{ij}^{\dagger} - \tilde{t}_i}{\tilde{\sigma}_i}$ for all j
3. Set $\widehat{\mathrm{se}}(\hat{t}) =$ standard deviation of the $\tilde{t}_i$
4. Find the $\alpha/2$ and $1-\alpha/2$ quantiles of the distribution of the $\tilde{\tau}$
5. Plug into Eq. 6.12.

The advantage of the studentized intervals is that they are more accurate than the basic ones; the disadvantage is that they are more work! At the other extreme, the **percentile method** simply sets the confidence interval to

$$\left(Q_{\tilde{t}}(\alpha/2), Q_{\tilde{t}}(1-\alpha/2)\right) \tag{6.13}$$

This is definitely easier to calculate, but not as accurate as the basic, pivotal CI.

All of these methods have many variations, described in the monographs referred to at the end of this chapter (§6.8).

```
boot.pvalue <- function(test, simulator, B, testhat) {
    testboot <- rboot(B = B, statistic = test, simulator = simulator)
    p <- (sum(testboot >= testhat) + 1)/(B + 1)
    return(p)
}
```

CODE EXAMPLE 9: *Bootstrap p-value calculation.* `testhat` *should be the value of the test statistic on the actual data.* `test` *is a function which takes in a data set and calculates the test statistic, presuming that large values indicate departure from the null hypothesis. Note the* $+1$ *in the numerator and denominator of the p-value — it would be more straightforward to leave them off, but this is a little more stable when B is comparatively small. (Also, it keeps us from ever reporting a p-value of exactly 0.)*

6.2.4 Hypothesis Testing

For hypothesis tests, we may want to calculate two sets of sampling distributions: the distribution of the test statistic under the null tells us about the size of the test and significance levels, and the distribution under the alternative tells us about power and realized power. We can find either with bootstrapping, by simulating from either the null or the alternative. In such cases, the statistic of interest, which I've been calling T, is the test statistic. Code Example 9 illustrates how to find a p-value by simulating under the null hypothesis. The same procedure would work to calculate power, only we'd need to simulate from the alternative hypothesis, and `testhat` would be set to the critical value of T separating acceptance from rejection, not the observed value.

6.2.4.1 Double bootstrap hypothesis testing

When the hypothesis we are testing involves estimated parameters, we may need to correct for this. Suppose, for instance, that we are doing a goodness-of-fit test. If we estimate our parameters on the data set, we adjust our distribution so that it matches the data. It is thus not surprising if it seems to fit the data well! (Essentially, it's the problem of evaluating performance by looking at in-sample fit, which gave us so much trouble in Chapter 3.)

Some test statistics have distributions which are not affected by estimating parameters, at least not asymptotically. In other cases, one can analytically come up with correction terms. When these routes are blocked, one uses a **double bootstrap**, where a second level of bootstrapping checks how much estimation improves the apparent fit of the model. This is perhaps most easily explained in pseudo-code (Code Example 10).

```
doubleboot.pvalue <- function(test, simulator, B1, B2, estimator, thetahat,
    testhat, ...) {
    for (i in 1:B1) {
        xboot <- simulator(theta = thetahat, ...)
        thetaboot <- estimator(xboot)
        testboot[i] <- test(xboot)
        pboot[i] <- boot.pvalue(test, simulator, B2, testhat = testboot[i],
            theta = thetaboot)
    }
    p <- (sum(testboot >= testhat) + 1)/(B1 + 1)
    p.adj <- (sum(pboot <= p) + 1)/(B1 + 1)
    return(p.adj)
}
```

CODE EXAMPLE 10: *Code sketch for "double bootstrap" significance testing. The inner or second bootstrap is used to calculate the distribution of nominal bootstrap p-values. For this to work, we need to draw our second-level bootstrap samples from $\tilde{\theta}$, the bootstrap re-estimate, not from $\hat{\theta}$, the data estimate. The code presumes the* `simulator` *function takes a* `theta` *argument allowing this. Exercise: replace the* `for` *loop with* `replicate`.

6.2.5 Model-Based Bootstrapping Example: Pareto's Law of Wealth Inequality

The **Pareto** or **power-law** distribution[6], is a popular model for data with "heavy tails", i.e. where the probability density $f(x)$ goes to zero only very slowly as $x \to \infty$. The probability density is

$$f(x) = \frac{\theta - 1}{x_0}\left(\frac{x}{x_0}\right)^{-\theta} \tag{6.14}$$

where x_0 is the minimum scale of the distribution, and θ is the **scaling exponent** (exercise 1). The Pareto is highly right-skewed, with the mean being much larger than the median.

If we know x_0, one can show that the maximum likelihood estimator of the exponent θ is

$$\hat{\theta} = 1 + \frac{n}{\sum_{i=1}^{n} \log \frac{x_i}{x_0}} \tag{6.15}$$

and that this is consistent (Exercise 3), and efficient. Picking x_0 is a harder problem (see Clauset *et al.* 2009) — for the present purposes, pretend that the Oracle tells us. The file `pareto.R`, on the book website, contains a number of functions related to the Pareto distribution, including a function `pareto.fit` for estimating it. (There's an example of its use below.)

Pareto came up with this density when he attempted to model the distribution of personal wealth. Approximately, but quite robustly across countries and time-

[6] Named after Vilfredo Pareto (1848–1923), the highly influential economist, political scientist, and proto-Fascist.

```
sim.wealth <- function() {
    rpareto(n = n.tail, threshold = wealth.pareto$xmin, exponent = wealth.pareto$exponent)
}
est.pareto <- function(data) {
    pareto.fit(data, threshold = x0)$exponent
}
```

CODE EXAMPLE 11: *Simulator and estimator for model-based bootstrapping of the Pareto distribution.*

periods, the upper tail of the distribution of income and wealth follows a power law, with the exponent varying as money is more or less concentrated among the very richest individuals and households[7]. Figure 6.2 shows the distribution of net worth for the 400 richest Americans in 2003. [[TODO: Permanent URL]]

```
source("http://www.stat.cmu.edu/~cshalizi/uADA/16/lectures/pareto.R")
wealth <- scan("http://www.stat.cmu.edu/~cshalizi/uADA/16/lectures/wealth.dat")
x0 <- 9e+08
n.tail <- sum(wealth >= x0)
wealth.pareto <- pareto.fit(wealth, threshold = x0)
```

Taking $x_0 = 9 \times 10^8$ (again, see Clauset *et al.* 2009), the number of individuals in the tail is 302, and the estimated exponent is $\hat{\theta} = 2.34$.

How much uncertainty is there in this estimate of the exponent? Naturally, we'll bootstrap. We need a function to generate Pareto-distributed random variables; this, along with some related functions, is part of the file `pareto.R` on the course website. With that tool, model-based bootstrapping proceeds as in Code Example 11.

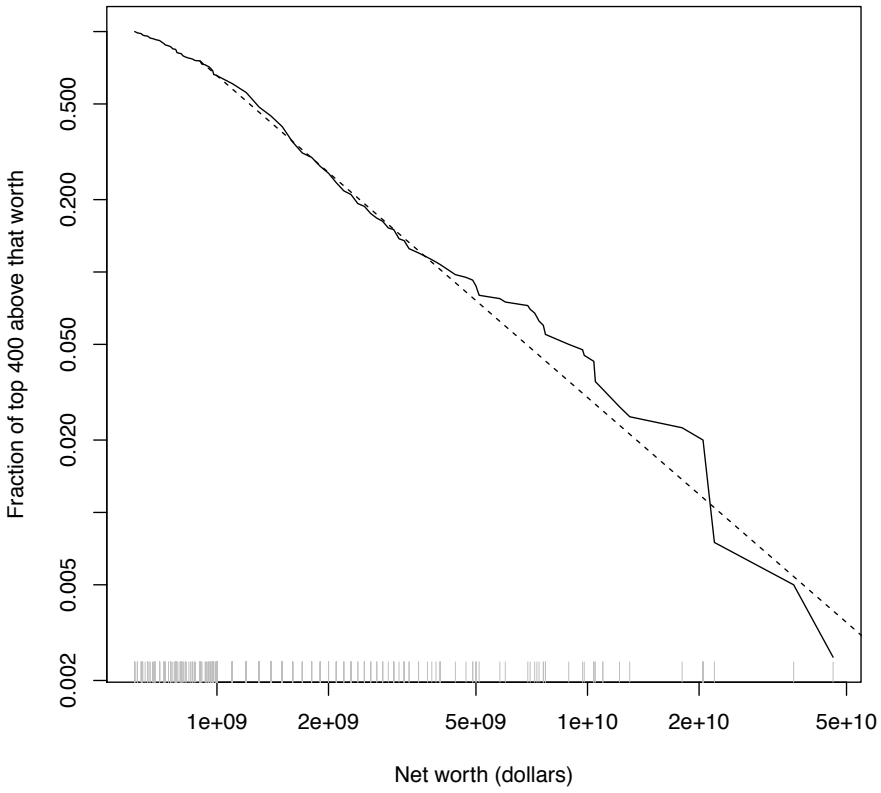
Using these functions, we can now calculate the bootstrap standard error, bias and 95% confidence interval for $\hat{\theta}$, setting $B = 10^4$:

```
pareto.se <- bootstrap.se(statistic = est.pareto, simulator = sim.wealth, B = 10000)
pareto.bias <- bootstrap.bias(statistic = est.pareto, simulator = sim.wealth,
    t.hat = wealth.pareto$exponent, B = 10000)
pareto.ci <- bootstrap.ci(statistic = est.pareto, simulator = sim.wealth, B = 10000,
    t.hat = wealth.pareto$exponent, level = 0.95)
```

This gives a standard error of ± 0.077, matching the asymptotic approximation reasonably well[8], but not needing asymptotic assumptions.

[7] Most of the distribution, for ordinary people, roughly conforms to a log-normal.

[8] "In Asymptopia", the variance of the MLE should be $\frac{(\hat{\theta}-1)^2}{n}$, in this case 0.076. The intuition is that this variance depends on how sharp the maximum of the likelihood function is — if it's sharply peaked, we can find the maximum very precisely, but a broad maximum is hard to pin down. Variance is thus inversely proportional to the second derivative of the negative log-likelihood. (The minus sign is because the second derivative has to be negative at a maximum, while variance has to be positive.) For one sample, the expected second derivative of the negative log-likelihood is $(\theta - 1)^{-2}$. (This is called the **Fisher information** of the model.) Log-likelihood adds across independent samples, giving us an over-all factor




```
plot.survival.loglog(wealth, xlab = "Net worth (dollars)", ylab = "Fraction of top 400 above tha
rug(wealth, side = 1, col = "grey")
curve((n.tail/400) * ppareto(x, threshold = x0, exponent = wealth.pareto$exponent,
    lower.tail = FALSE), add = TRUE, lty = "dashed", from = x0, to = 2 * max(wealth))
```

FIGURE 6.2: *Upper cumulative distribution function (or "survival function") of net worth for the 400 richest individuals in the US (2000 data). The solid line shows the fraction of the 400 individuals whose net worth W equaled or exceeded a given value w, $\Pr(W \geq w)$. (Note the logarithmic scale for both axes.) The dashed line is a maximum-likelihood estimate of the Pareto distribution, taking $x_0 = \$9 \times 10^8$. (This threshold was picked using the method of Clauset* et al. *2009.) Since there are 302 individuals at or above the threshold, the cumulative distribution function of the Pareto has to be reduced by a factor of (302/400).*

```
ks.stat.pareto <- function(x, exponent, x0) {
    x <- x[x >= x0]
    ks <- ks.test(x, ppareto, exponent = exponent, threshold = x0)
    return(ks$statistic)
}
ks.pvalue.pareto <- function(B, x, exponent, x0) {
    testhat <- ks.stat.pareto(x, exponent, x0)
    testboot <- vector(length = B)
    for (i in 1:B) {
        xboot <- rpareto(length(x), exponent = exponent, threshold = x0)
        exp.boot <- pareto.fit(xboot, threshold = x0)$exponent
        testboot[i] <- ks.stat.pareto(xboot, exp.boot, x0)
    }
    p <- (sum(testboot >= testhat) + 1)/(B + 1)
    return(p)
}
```

CODE EXAMPLE 12: *Calculating a p-value for the Pareto distribution, using the Kolmogorov-Smirnov test and adjusting for the way estimating the scaling exponent moves the fitted distribution closer to the data.*

Asymptotically, the bias is known to go to zero; at this size, bootstrapping gives a bias of 0.0051, which is effectively negligible.

We can also get the confidence interval; with the same 10^4 replications, the 95% CI is 2.17, 2.48. In theory, the confidence interval could be calculated exactly, but it involves the inverse gamma distribution (Arnold, 1983), and it is quite literally faster to write and do the bootstrap than go to look it up.

A more challenging problem is goodness-of-fit; we'll use the Kolmogorov-Smirnov statistic.[9] Code Example 12 calculates the p-value. With ten thousand bootstrap replications,

```
signif(ks.pvalue.pareto(10000, wealth, wealth.pareto$exponent, x0), 4)
## [1] 0.008999
```

Ten thousand replicates is enough that we should be able to accurately estimate probabilities of around 0.01 (since the binomial standard error will be $\sqrt{\frac{(0.01)(0.99)}{10^4}} \approx 9.9 \times 10^{-4}$); if it weren't, we might want to increase B.

Simply plugging in to the standard formulas, and thereby ignoring the effects of estimating the scaling exponent, gives a p-value of 0.171, which is not outstanding but not awful either. Properly accounting for the flexibility of the model, however, the

of n. In the large-sample limit, the actual log-likelihood will converge on the expected log-likelihood, so this gives us the asymptotic variance. (See also §H.4.1.)

[9] The `pareto.R` file contains a function, `pareto.tail.ks.test`, which does a goodness-of-fit test for fitting a power-law to the tail of the distribution. That differs somewhat from what follows, because it takes into account the extra uncertainty which comes from having to estimate x_0. Here, I am pretending that an Oracle told us $x_0 = 9 \times 10^8$.

discrepancy between what it predicts and what the data shows is so large that it would take a big (one-in-a-hundred) coincidence to produce it. We have, therefore, detected that the Pareto distribution makes systematic errors for this data, but we don't know much about what they are. In Chapter 15, we'll look at techniques which can begin to tell us something about *how* it fails.

[[TODO: Revisit this example in that chapter, cross-ref]]

6.3 Bootstrapping by Resampling

The bootstrap approximates the sampling distribution, with three sources of approximation error. First, **simulation error**: using finitely many replications to stand for the full sampling distribution. Clever simulation design can shrink this, but brute force — just using enough replicates — can also make it arbitrarily small. Second, **statistical error**: the sampling distribution of the bootstrap re-estimates under our estimated model is not exactly the same as the sampling distribution of estimates under the true data-generating process. The sampling distribution changes with the parameters, and our initial estimate is not completely accurate. But it often turns out that distribution of estimates *around* the truth is more nearly invariant than the distribution of estimates themselves, so subtracting the initial estimate from the bootstrapped values helps reduce the statistical error; there are many subtler tricks to the same end. Third, **specification error**: the data source doesn't exactly follow our model at all. Simulating the model then never quite matches the actual sampling distribution.

Efron had a second brilliant idea, which is to address specification error by replacing simulation from the model with re-sampling from the data. After all, our initial collection of data gives us a lot of information about the relative probabilities of different values. In a sense the empirical distribution is the least prejudiced estimate possible of the underlying distribution — anything else imposes biases or preconceptions, possibly accurate but also potentially misleading[10]. Lots of quantities can be estimated directly from the empirical distribution, without the mediation of a model. Efron's **resampling bootstrap** (a.k.a. the **non-parametric bootstrap**) treats the original data set as a complete population and draws a new, simulated sample from it, picking each observation with equal probability (allowing repeated values) and then re-running the estimation (Figure 6.3, Code Example 13). In fact, this is usually what people mean when they talk about "the bootstrap" without any modifier.

Everything we did with model-based bootstrapping can also be done with resampling bootstrapping — the only thing that's changing is the distribution the surrogate data is coming from.

The resampling bootstrap should remind you of k-fold cross-validation. The analog of leave-one-out CV is a procedure called the **jack-knife**, where we repeat the estimate n times on $n-1$ of the data points, holding each one out in turn. It's historically important (it dates back to the 1940s), but generally doesn't work as well as resampling.

An important variant is the **smoothed bootstrap**, where we re-sample the data points and then perturb each by a small amount of noise, generally Gaussian[11].

[10] See §14.6 in Chapter 14.

[11] We will see in Chapter 14 that this corresponds to sampling from a kernel density estimate.

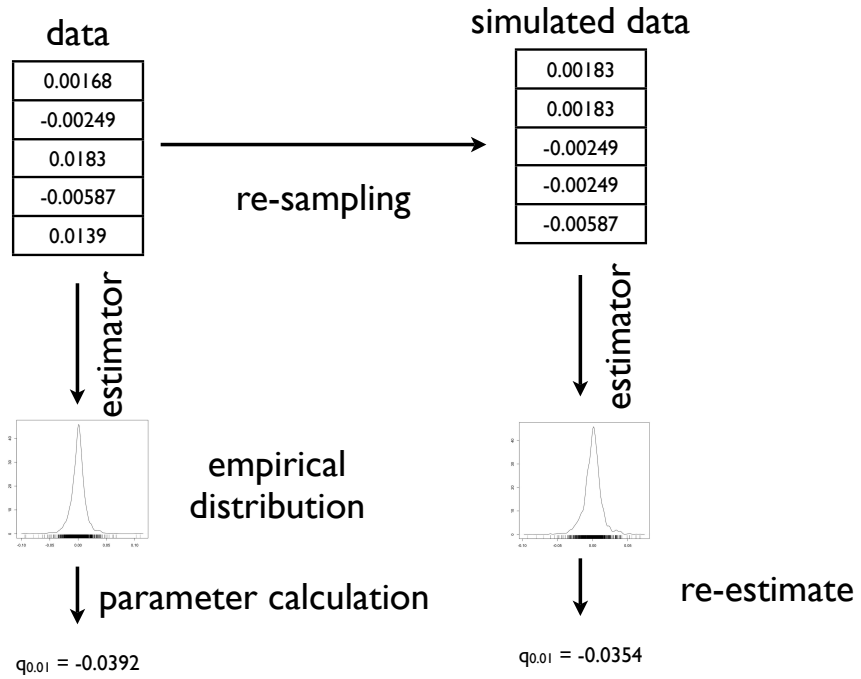


FIGURE 6.3: *Schematic for the resampling bootstrapping. New data is simulated by re-sampling from the original data (with replacement), and functionals are calculated either directly from the empirical distribution, or by estimating a model on this surrogate data.*

```
resample <- function(x) {
    sample(x, size = length(x), replace = TRUE)
}
resample.data.frame <- function(data) {
    sample.rows <- resample(1:nrow(data))
    return(data[sample.rows, ])
}
```

CODE EXAMPLE 13: *A utility function to resample from a vector, and another which resamples from a data frame. Can you write a single function which determines whether its argument is a vector or a data frame, and does the right thing in each case/*

Back to the Pareto example Let's see how to use re-sampling to get a 95% confidence interval for the Pareto exponent[12].

```
wealth.resample <- function() {
    resample(wealth[wealth >= x0])
}
pareto.CI.resamp <- bootstrap.ci(statistic = est.pareto, simulator = wealth.resample,
    t.hat = wealth.pareto$exponent, level = 0.95, B = 10000)
```

The interval is 2.16, 2.48; this is very close to the interval we got from the model-based bootstrap, which should actually reassure us about the latter's validity.

6.3.1 Model-Based vs. Resampling Bootstraps

When we have a properly specified model, simulating from the model gives more accurate results (at the same n) than does re-sampling the empirical distribution — parametric estimates of the distribution converge faster than the empirical distribution does. If on the other hand the model is mis-specified, then it is rapidly converging to the *wrong* distribution. This is of course just another bias-variance trade-off, like those we've seen in regression.

Since I am suspicious of most parametric modeling assumptions, I prefer re-sampling, when I can figure out how to do it, or at least until I have convinced myself that a parametric model is a good approximation to reality.

6.4 Bootstrapping Regression Models

Let's recap what we're doing estimating regression models. We want to learn the regression function $\mu(x) = \mathbb{E}[Y|X = x]$. We estimate the model on a set of predictor-response pairs, $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$, resulting in an estimated curve (or surface) $\widehat{\mu}(x)$, fitted values $\widehat{\mu}_i = \widehat{\mu}(x_i)$, and residuals, $\epsilon_i = y_i - \widehat{\mu}_i$. For any such model, we have a choice of several ways of bootstrapping, in decreasing order of reliance on the model.

- Simulate new X values from the model's distribution of X, and then draw Y from the specified conditional distribution $Y|X$.
- Hold the x fixed, but draw $Y|X$ from the specified distribution.
- Hold the x fixed, but make Y equal to $\widehat{\mu}(x)$ plus a randomly re-sampled ϵ_j.
- Re-sample (x, y) pairs.

[12] Even if the Pareto model is wrong, the estimator of the exponent will converge on the value which gives, in a certain sense, the best approximation to the true distribution from among all power laws. Econometricians call such parameter values the **pseudo-truth**; we are getting a confidence interval for the pseudo-truth. In this case, the pseudo-true scaling exponent can still be a useful way of summarizing *how* heavy tailed the income distribution is, despite the fact that the power law makes systematic errors.

The first case is pure model-based bootstrapping. (So is the second, sometimes, when the regression model is agnostic about X.) The last case is just re-sampling from the joint distribution of (X, Y). The next-to-last case is called **re-sampling the residuals** or **re-sampling the errors**. When we do that, we rely on the regression model to get the conditional expectation function right, but we don't count on it getting the distribution of the noise around the expectations.

The specific procedure of re-sampling the residuals is to re-sample the ϵ_i, with replacement, to get $\tilde{\epsilon}_1, \tilde{\epsilon}_2, \ldots \tilde{\epsilon}_n$, and then set $\tilde{x}_i = x_i$, $\tilde{y}_i = \widehat{\mu}(\tilde{x}_i) + \tilde{\epsilon}_i$. This surrogate data set is then re-analyzed like new data.

6.4.1 Re-sampling Points: Parametric Model Example

[[ATTN: Replace with more modern data example?]]

A classic data set contains the time between 299 eruptions of the Old Faithful geyser in Yellowstone, and the length of the subsequent eruptions; these variables are called `waiting` and `duration`. (We saw this data set already in §5.4.2.1, and will see it again in §7.3.2.) We'll look at the linear regression of `waiting` on `duration`. We'll re-sample (`duration`, `waiting`) pairs, and would like confidence intervals for the regression coefficients. This is a confidence interval for the coefficients of *the best linear predictor*, a functional of the distribution, which, as we saw in Chapters 1 and 2, exists no matter how nonlinear the process really is. It's only a confidence interval for the *true regression parameters* if the real regression function is linear.

Before anything else, look at the model:

```
library(MASS)
data(geyser)
geyser.lm <- lm(waiting ~ duration, data = geyser)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	99.3	1.960	50.7	0
duration	-7.8	0.537	-14.5	0

The first step in bootstrapping this is to build our simulator, which just means sampling rows from the data frame:

```
resample.geyser <- function() {
    resample.data.frame(geyser)
}
```

We can check this by running `summary(geyser.resample())`, and seeing that it gives about the same quartiles and mean for both variables as `summary(geyser)`[13], but that the former gives different numbers each time it's run.

Next, we define the estimator:

[13] The minimum and maximum won't match up well — why not?

```
est.geyser.lm <- function(data) {
    fit <- lm(waiting ~ duration, data = data)
    return(coefficients(fit))
}
```

We can check that this function works by seeing that `coefficients(geyser.lm)` matches `est.geyser.lm(geyser)`, but that `est.geyser.lm(resample.geyser()` is different every time we run it.

Put the pieces together:

```
geyser.lm.ci <- bootstrap.ci(statistic=est.geyser.lm,
                             simulator=resample.geyser,
                             level=0.95,
                             t.hat=coefficients(geyser.lm),
                             B=1e4)
```

	lower	upper
(Intercept)	96.5	102.00
duration	-8.7	-6.91

Notice that we do not have to assume homoskedastic Gaussian noise — fortunately, because that's a very bad assumption here[14].

[14] We have calculated 95% confidence intervals for the intercept β_0 and the slope β_1 separately. These intervals cover their coefficients all but 5% of the time. Taken together, they give us a rectangle in (β_0, β_1) space, but the coverage probability of *this* rectangle could be anywhere from 95% all the way down to 90%. To get a confidence *region* which simultaneously covers both coefficients 95% of the time, we have two big options. One is to stick to a box-shaped region and just increase the confidence level on each coordinate (to 97.5%). The other is to define some suitable metric of how far apart coefficient vectors are (e.g., ordinary Euclidean distance), find the 95% percentile of the distribution of this metric, and trace the appropriate contour around $\hat{\beta}_0, \hat{\beta}_1$. [[TODO: Example.]]

```
main.curve <- npr.geyser(geyser)

# We already defined this in a previous example, but it doesn't hurt
resample.geyser <- function() { resample.data.frame(geyser) }

geyser.resampled.curves <- rboot(statistic=npr.geyser,
                                 simulator=resample.geyser,
                                 B=800)
```

CODE EXAMPLE 14: *Generating multiple kernel-regression curves for the geyser data, by resampling that data frame and re-estimating the model on each simulation.* `geyser.resampled.curves` *stores the predictions of those 800 models, evaluated at a* common *set of values for the predictor variable. The vector* `main.curve`*, which we'll use presently to get confidence intervals, stores predictions of the model fit to the whole data, evaluated at that same set of points.*

6.4.2 Re-sampling Points: Non-parametric Model Example

Nothing in the logic of re-sampling data points for regression requires us to use a parametric model. Here we'll provide 95% confidence bounds for the kernel smoothing of the geyser data. Since the functional is a whole curve, the confidence set is often called a **confidence band**.

We use the same simulator, but start with a different regression curve, and need a different estimator.

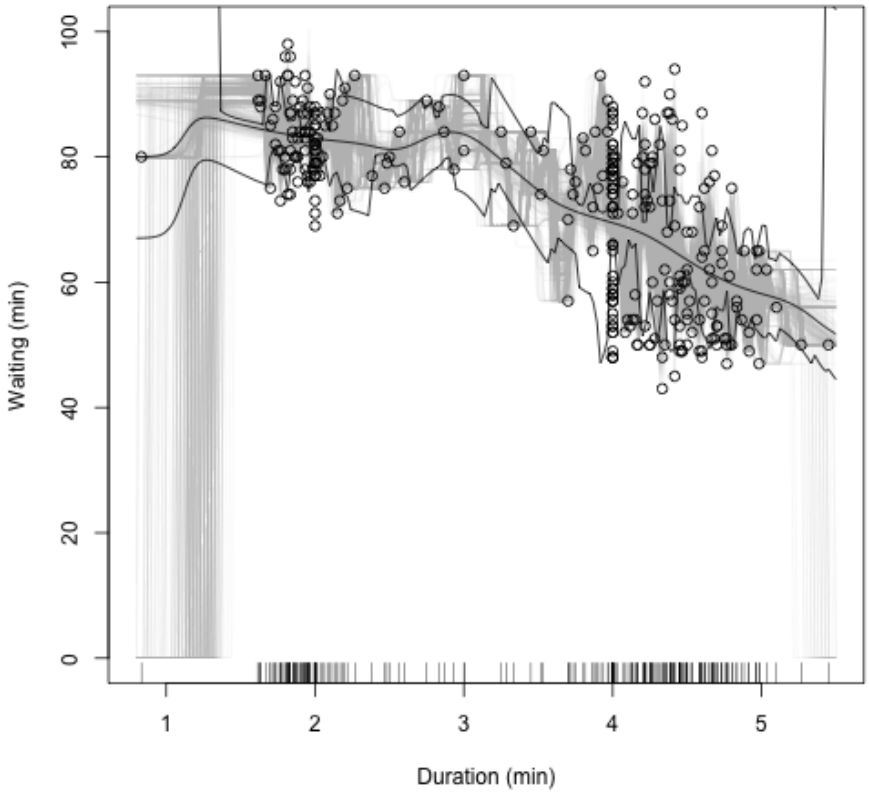
```
evaluation.points <- data.frame(duration = seq(from = 0.8, to = 5.5, length.out = 200))
library(np)
npr.geyser <- function(data, tol = 0.1, ftol = 0.1, plot.df = evaluation.points) {
    bw <- npregbw(waiting ~ duration, data = data, tol = tol, ftol = ftol)
    mdl <- npreg(bw)
    return(predict(mdl, newdata = plot.df))
}
```

Now we construct pointwise 95% confidence bands for the regression curve. For this end, we don't really need to keep around the whole kernel regression object — we'll just use its predicted values on a uniform grid of points, extending slightly beyond the range of the data (Code Example 14). Observe that this will go through bandwidth selection again for each bootstrap sample. This is slow, but it is the most secure way of getting good confidence bands. Applying the bandwidth we found on the data to each re-sample would be faster, but would introduce an extra level of approximation, since we wouldn't be treating each simulation run the same as the original data.

Figure 6.4 shows the curve fit to the data, the 95% confidence limits, and (faintly) all of the bootstrapped curves. Doing the 800 bootstrap replicates took 4 minutes on my laptop[15].

[[TODO: Talk about the bias issue (here? further reading?), hacks vs. living with it]]

[15] Specifically, I ran `system.time(geyser.resampled.curves <- rboot(statistic=npr.geyser,`




```
plot(0, type = "n", xlim = c(0.8, 5.5), ylim = c(0, 100), xlab = "Duration (min)",
    ylab = "Waiting (min)")
for (i in 1:ncol(geyser.resampled.curves)) {
    lines(evaluation.points$duration, geyser.resampled.curves[, i], lwd = 0.1,
        col = "grey")
}
geyser.npr.cis <- bootstrap.ci(tboots = geyser.resampled.curves, t.hat = main.curve,
    level = 0.95)
lines(evaluation.points$duration, geyser.npr.cis[, "lower"])
lines(evaluation.points$duration, geyser.npr.cis[, "upper"])
lines(evaluation.points$duration, main.curve)
rug(geyser$duration, side = 1)
points(geyser$duration, geyser$waiting)
```

FIGURE 6.4: *Kernel regression curve for Old Faithful (central black line), with 95% confidence bands (other black lines), the 800 bootstrapped curves (thin, grey lines), and the data points. Notice that the confidence bands get wider where there is less data.* Caution: *doing the bootstrap took 4 minutes to run on my computer.*

`simulator=resample.geyser, B=800))`, which not only did the calculations and stored them in `geyser.resampled.curves`, but told me how much time it took R to do all that.

```
resample.residuals.penn <- function() {
    new.frame <- penn
    new.growths <- fitted(penn.lm) + resample(residuals(penn.lm))
    new.frame$gdp.growth <- new.growths
    return(new.frame)
}
penn.estimator <- function(data) {
    mdl <- lm(penn.formula, data = data)
    return(coefficients(mdl))
}
penn.lm.cis <- bootstrap.ci(statistic = penn.estimator, simulator = resample.residuals.penn,
    B = 10000, t.hat = coefficients(penn.lm), level = 0.95)
```

CODE EXAMPLE 15: *Re-sampling the residuals to get confidence intervals in a linear model.*

6.4.3 Re-sampling Residuals: Example

As an example of re-sampling the residuals, rather than data points, let's take a linear regression, based on the data-analysis assignment in §A.13.[16] We will regress `gdp.growth` on `log(gdp)`, `pop.growth`, `invest` and `trade`:

```
penn <- read.csv("http://www.stat.cmu.edu/~cshalizi/uADA/13/hw/02/penn-select.csv")
penn.formula <- "gdp.growth ~ log(gdp) + pop.growth + invest + trade"
penn.lm <- lm(penn.formula, data = penn)
```

(Why make the formula a separate object here?) The estimated parameters are

(Intercept)	5.71e-04
log(gdp)	5.07e-04
pop.growth	-1.87e-01
invest	7.15e-04
trade	3.11e-05

Code Example 15 shows the new simulator for this set-up (`resample.residuals.penn`)[17], the new estimation function (`penn.estimator`)[18], and the confidence interval calculation (`penn.lm.cis`):

	lower	upper
(Intercept)	-1.58e-02	1.70e-02
log(gdp)	-1.46e-03	2.45e-03
pop.growth	-3.56e-01	-1.92e-02
invest	4.93e-04	9.42e-04
trade	-1.86e-05	8.35e-05

[16] Note to 2016 students: This is an old problem set related to our homework 2, also on the determinants of economic growth across countries and years, but using a different set of variables. `gdp.growth` and `gdp` are obvious, `pop.growth` is the country's rate of population growth, `invest` is the fraction of GDP devoted to investment, and `trade` is the ratio of imports plus exports to GDP. The data repository is the "Penn World Tables". See pp. 664 of the full draft for details and the assignment.

[17] How would you check that this worked?

[18] How would you check that this worked?

Doing ten thousand linear regressions took 45 seconds on my computer, as opposed to 4 minutes for eight hundred kernel regressions.

6.5 Bootstrap with Dependent Data

If the data points we are looking at are vectors (or more complicated structures) with dependence between components, but each data point is independently generated from the same distribution, then dependence isn't really an issue. We re-sample vectors, or generate vectors from our model, and proceed as usual. In fact, that's what we've done so far in several cases.

If there is dependence *across* data points, things are more tricky. If our model incorporates this dependence, then we can just simulate whole data sets from it. An appropriate re-sampling method is trickier — just re-sampling individual data points destroys the dependence, so it won't do. We will revisit this question when we look at time series and spatial data in Chapters 21–22.

6.6 Things Bootstrapping Does Poorly

The principle behind bootstrapping is that sampling distributions under the true process should be close to sampling distributions under good estimates of the truth. If small perturbations to the data-generating process produce huge swings in the sampling distribution, bootstrapping will not work well, and may fail spectacularly. For model-based bootstrapping, this means that small changes to the underlying parameters must produce small changes to the functionals of interest. Similarly, for resampling, it means that adding or removing a few data points must change the functionals only a little[19].

Re-sampling in particular has trouble with extreme values. Here is a simple example: Our data points X_i are IID, with $X_i \sim Unif(0, \theta_0)$, and we want to estimate θ_0. The maximum likelihood estimate $\hat{\theta}$ is just the sample maximum of the x_i. We'll use resampling to get a confidence interval for this, as above — but I will fix the true $\theta_0 = 1$, and see how often the 95% confidence interval covers the truth.

```
max.boot.ci <- function(x, B) {
    max.boot <- replicate(B, max(resample(x)))
    return(2 * max(x) - quantile(max.boot, c(0.975, 0.025)))
}
boot.cis <- replicate(1000, max.boot.ci(x = runif(100), B = 1000))
(true.coverage <- mean((1 >= boot.cis[1, ]) & (1 <= boot.cis[2, ])))
## [1] 0.878
```

That is, the actual coverage probability is not 95% but about 88%.

If you suspect that your use of the bootstrap may be setting yourself up for a similar epic fail, your two options are (1) learn some of the theory of the bootstrap from the references in the "Further Reading" section below, or (2) set up a simulation experiment like this one.

[19] More generally, moving from one distribution function f to another $(1-\epsilon)f + \epsilon g$ mustn't change the functional very much when ϵ is small, no matter in what "direction" g we perturb it. Making this idea precise calls for some fairly deep mathematics, about differential calculus on spaces of functions (see, e.g., van der Vaart 1998, ch. 20).

6.7 Which Bootstrap When?

This chapter has introduced a bunch of different bootstraps, and before it closes it's worth reviewing the general principles, and some of the considerations which go into choosing among them in a particular problem.

When we bootstrap, we try to approximate the sampling distribution of some statistic (mean, median, correlation coefficient, regression coefficients, smoothing curve, difference in MSEs...) by running simulations, and calculating the statistic on the simulation. We've seen three major ways of doing this:

- The model-based bootstrap: we estimate the model, and then simulate from x the estimated model;
- Resampling residuals: we estimate the model, and then simulate by resampling residuals to that estimate and adding them back to the fitted values;
- Resampling cases or whole data points: we ignore the estimated model completely in our simulation, and just re-sample whole rows from the data frame.

Which kind of bootstrap is appropriate depends on how much trust we have in our model.

The model-based bootstrap trusts the model to be completely correct for *some* parameter value. In, e.g., regression, it trusts that we have the right shape for the regression function *and* that we have the right distribution for the noise. When we trust our model this much, we could in principle work out sampling distributions analytically; the model-based bootstrap replaces hard math with simulation.

Resampling residuals doesn't trust the model as much. In regression problems, it assumes that the model gets the *shape* of the regression function right, and that the noise around the regression function is independent of the predictor variables, but doesn't make any further assumption about how the fluctuations are distributed. It is therefore more secure than model-based bootstrap.[20]

Finally, resampling cases assumes nothing at all about either the shape of the regression function or the distribution of the noise, it just assumes that each data point (row in the data frame) is an independent observation. Because it assumes so little, and doesn't depend on any particular model being correct, it is very safe.

The reason we do not always use the safest bootstrap, which is resampling cases, is that there is, as usual, a bias-variance trade-off. Generally speaking, if we compare three sets of bootstrap confidence intervals on the same data for the same statistic, the model-based bootstrap will give the narrowest intervals, followed by resampling residuals, and resampling cases will give the loosest bounds. If the model really *is* correct about the shape of the curve, we can get more precise results, without any loss of accuracy, by resampling residuals rather than resampling cases. If the model is also correct about the distribution of noise, we can do even better with a model-based bootstrap.

[20] You could also imagine simulations where we presume that the noise takes a very particular form (e.g., a t-distribution with 10 degrees of freedom), but are agnostic about the shape of the regression function, and learn that non-parametrically. It's harder to think of situations where this is really plausible, however, except *maybe* Gaussian noise arising from central-limit-theorem considerations.

To sum up: resampling cases is safer than resampling residuals, but gives wider, weaker bounds. If you have good reason to trust a model's guess at the shape of the regression function, then resampling residuals is preferable. If you don't, or it's not a regression problem so there are no residuals, then you prefer to resample cases. The model-based bootstrap works best when the over-all model is correct, and we're just uncertain about the exact parameter values we need.

6.8 Further Reading

Davison and Hinkley (1997) is both a good textbook, and the reference I consult most often. Efron and Tibshirani (1993), while also very good, is more theoretical. Canty *et al.* (2006) has useful advice for serious applications.

All the bootstraps discussed in this chapter presume IID observations. For bootstraps for time series, see §21.5.

Software For professional purposes, I strongly recommend using the R package `boot` (Canty and Ripley, 2013), based on Davison and Hinkley (1997). I deliberately do *not* use it in this chapter, or later in the book, for pedagogical reasons; I have found that forcing students to write their own bootstrapping code helps build character, or at least understanding.

The bootstrap vs. robust standard errors For linear regression coefficients, econometricians have developed a variety of "robust" standard errors which are valid under weaker conditions than the usual assumptions. Buja *et al.* (2014) shows their equivalence to resampling cases. (See also King and Roberts 2015.)

Historical notes The original paper on the bootstrap, Efron (1979), is extremely clear, and for the most part presented in the simplest possible terms; it's worth reading. His later small book (Efron, 1982), while often cited, is not in my opinion so useful nowadays[21].

As the title of that last reference suggests, the bootstrap is in some ways a successor to an older method, apparently dating back to the 1940s if not before, called the "jackknife", in which each data point is successively held back and the estimate is re-calculated; the variance of these re-estimates, appropriately scaled, is then taken as the variance of estimation, and similarly for the bias[22]. The jackknife is appealing in its simplicity, but is only valid under much stronger conditions than the bootstrap.

6.9 Exercises

1. Show that x_0 is the mode of the Pareto distribution.

[21] It seems to have done a good job of explaining things to people who were already professional statisticians in 1982.

[22] A "jackknife" is a knife with a blade which folds into the handle; think of the held-back data point as the folded-away blade.

2. Derive the maximum likelihood estimator for the Pareto distribution (Eq. 6.15) from the density (Eq. 6.14).

3. Show that the MLE of the Pareto distribution is consistent.

   (a) Using the law of large numbers, show that $\hat{\theta}$ (Eq. 6.15) converges to a limit which depends on $\mathbb{E}[\log X/x_0]$.

   (b) Find an expression for $\mathbb{E}[\log X/x_0]$ in terms of θ and from the density (Eq. 6.14). *Hint:* Write $\mathbb{E}[\log X/x_0]$ as an integral, change the variable of integration from x to $z = \log(x/x_0)$, and remember that the mean of an exponential random variable with rate λ is $1/\lambda$.

4. Find confidence bands for the linear regression model of §6.4.1 using

   (a) The usual Gaussian assumptions (*hint*: try the `intervals="confidence"` option to `predict`);

   (b) Resampling of residuals; and

   (c) Resampling of cases.

5. (Computational) Writing new functions to simulate every particular linear model is somewhat tedious.

   (a) Write a function which takes, as inputs, an `lm` model and a data frame, and returns a new data frame where the response variable is replaced by the model's predictions plus Gaussian noise, but all other columns are left alone.

   (b) Write a function which takes, as inputs, an `lm` model and a data frame, and returns a new data frame where the response variable is replaced by the model's predictions plus resampled residuals.

   (c) Will your functions work with `npreg` models, as well as `lm` models? If not, what do you have to modify?

   Hint: See Code Example 2 in Chapter 3 for some R tricks to extract the name of the response variable from the estimated model.