Principles of Data Mining

David Hand
Imperial College

Heikki Mannila
Microsoft Research

Padhraic Smyth
University of California, Irvine

MIT Press, to appear
Draft — October 23, 1999
Do not distribute

Contents

Chapter 11

Predictive Modeling for Regression

11.1 Introduction

In Chapter 6 we noted the distinction between predictive and descriptive models. In Chapter 10 we described in detail predictive models in which the variable to be predicted (the *response variable*) was a nominal variable - it could take one of only a finite (and typically small) number of values, and these values had no numerical significance, so that they were simply *class identifiers*. In this chapter we turn to predictive models in which the response variable does have numerical significance. Examples are the amount a retail store might earn from a given customer over a ten year period, the rate of fuel consumption of a given type of car under normal conditions, the number of people who might access a particular website in a given month, and so on.

The variables to be used for prediction will be called *predictor variables*. In addition to these names, other authors sometimes use the words *dependent* and *target* for the response variable, and *independent*, *explanatory*, and *regressor* for the predictor variables. Other names used in the classification context were mentioned in Chapter 10. The predictor variables can be numerical, but they need not be.

Our aim, then, is to use a sample of objects for, which both the response variable and the predictor variables are known, to construct a model which will allow prediction of the numerical value of the response variable for a new case for which only the predictor variables are known. This is essentially the same problem as in Chapter 10, the only difference being the numerical instead of nominal nature of the response variable

Accuracy of prediction is one of the most important properties of such models, so various measures of accuracy have been devised. These measures may also be used for choosing between alternative models, and for choosing the values of parameters in models. In the terminology introduced earlier, these measures are *score* functions, by which different models may be compared.

Predictive accuracy is a critical aspect of models, but it is not the only aspect. For example, we might use the model to shed insight into which of the predictor variables are most important. We might even insist that some variables are included in the model, even though they only lead to small predictive improvement. Contrariwise, we might omit variables which we feel would enhance our predictive performance. (An example of this situation arises in credit scoring, where, in many countries, it is illegal to include sex as a predictor variable.) We might be interested in whether predictor variables interact, in the sense that the effect that one has on the response variable depends on the values taken by others. For obvious reasons, we might be interested in whether good prediction can be achieved by a simple model. Sometimes one might even be willing to sacrifice some predictive accuracy in exchange for substantially reduced model complexity. We could go on, but the essential point is that though predictive accuracy is perhaps the most important component of the performance of a predictive model, this has to be tempered by the context in which the model is to be applied.

11.2 Linear Models and Least Squares Fitting

We begin with the case in which the response variable is continuous rather than nominal. Section 6.4.1XXX introduced the idea of linear models, so-called because they are linear in the parameters. The simplest such model yields predicted values, $\hat{y}$, of the response variable y,using a linear combination of the predictor variables x_j:

$$\hat{y} = a_o + \sum_{i=1}^{p} a_j x_j \tag{11.1}$$

(so that such models are in fact linear in both the parameters a_j and the predictor variables x_j). In fact, of course, we will not normally be able to predict the response variable perfectly (life is seldom so simple) and a common aim is to predict the mean value that y takes at each vector of the predictor variables—so $\hat{y}$ is our predicted estimate of the mean value at $\mathbf{x} = (x_1, \ldots, x_p)$.

Models of this form are known as *linear regression models*. In the simplest case of a single predictor variable (so-called *simple regression*), we have a *regression line*. More generally (*multiple regression*) we have a *regression plane*. Such models are the oldest, most important, and single most widely used form of predictive model. One reason for this is their evident simplicity; a simple weighted sum is very easy both to compute and to understand. Another reason is the compelling one that they often perform very well—even in circumstances where one knows enough to be confident that the true relationship between the predictor and response variables cannot be linear. This is not altogether surprising: when one expands continuous mathematical functions in a Taylor series one often finds that the lowest order terms—the linear terms—are the most important, so that the best simple approximation is obtained by using a linear model.

It would be an extremely rare situation in which the chosen model was exactly right. This is especially true in data mining situations, where one's model is generally empirical rather than iconic (see Chapter 6XXX) so that it is not based on any underlying theory. The model may not include all of the predictor variables which would be needed for perfect prediction (many may not have been measured or even be measurable); it may not include certain functions of the predictor variables (maybe x_1^2 is needed as well as x_1, or maybe products of the predictor variables are needed because they interact in their effect on y);

and, in any case, no measurement is perfect: the y variable will have errors associated with it so that each vector $(x_1, \ldots, x_p)$ will be associated with a distribution of possible y values, as we have noted above.

All of this means that the actual y values in a sample will differ from the predicted values. The differences between observed and predicted values are called *residuals*, and we denote these by e:

$$y = \hat{y} + e = a_o + \sum_{i=1}^{p} a_j x_j + e \tag{11.2}$$

In matrix terms, if we denote the observed y measurements on the n objects in the design sample by the vector $\mathbf{y}$ and the p measurements of the predictor variables on the n objects by the n by p matrix $\mathbf{X}$, we can express the relationship between the observed response and predictor measurements, in terms of our model, as

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{e} \tag{11.3}$$

In this equation, $\mathbf{a} = (a_0, \ldots, a_p)$ represents the vector of parameter values, and the vector $\mathbf{e} = (e_1, \ldots, e_n)$ contains the residuals. The coefficient a_0, providing the intercept of the model, has been included here, so that the rows of matrix $\mathbf{X}$ include a first term which is always 1. Clearly we want to choose the parameters in our model, the values in the $p+1$ vector $\mathbf{a}$, so as to yield predictions which are as accurate as possible. Put another way, we must find estimates for the a_j which minimize the $\mathbf{e}$ discrepancies in some way. To do this, we combine the elements of $\mathbf{e}$ in some way to yield a single numerical measure which we can minimize.

Various ways of combining the e_i have been proposed, but by far the most popular method is to sum their squares. That is, we seek the values for the parameter vector $\mathbf{a}$ which minimizes

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} a_j x_{ij} \right)^2 \tag{11.4}$$

In this expression, y_i is the observed y value for the ith design sample point and $(x_{i0}, x_{i1}, \ldots, x_{ip}) = (1, x_{i1}, \ldots, x_{ip})$ is the vector of predictor variables for this point. For obvious reasons, this method is known as the *least squares method*. For simplicity, we will denote the parameter vector which minimizes this by $(a_0, \ldots, a_p)$. (It would be more correct, of course, if we used some notation to indicate that it is an estimate, such as $(\hat{a}_0, \ldots, \hat{a}_p)$, but our notation has the merit of simplicity.)

In matrix terms, the values of the parameters which minimize Equation 11.4 are given by

$$\mathbf{a} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}\mathbf{y} \tag{11.5}$$

In linear regression in general, the $\mathbf{a}$ parameters are called *regression coefficients*. Once the parameters have been estimated, they are used in Equation 11.1 to yield predictions. The predicted value of y, $\hat{y}_k$, for a vector of predictor variables $\mathbf{x}_k$, is given by $\hat{y}_k = \mathbf{x}_k'\mathbf{a} = \mathbf{a}'\mathbf{x}_k$.

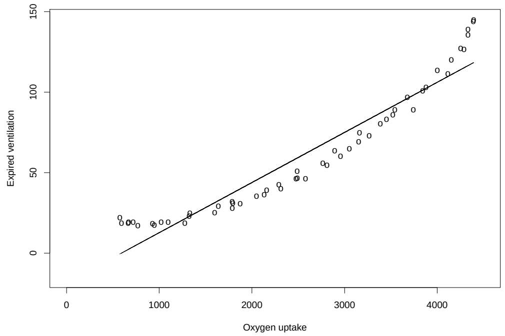

Figure 11.1: Expired ventilation plotted against oxygen uptake in a series of trials, with fitted straight line.

11.2.1 Computational Issues in Fitting the Model

Solution 11.5 requires that the matrix $\mathbf{X}^T\mathbf{X}$ be invertible. Problems will arise if the sample size n is small (rare in data mining situations) or if there are linear dependencies between the measured values of the predictor variables (not so rare). In the latter case, modern software normally issues warnings and appropriate action can be taken, such as dropping some of the predictor variables.

A rather more subtle problem arises when the measured values of the predictor variables are not exactly linearly dependent, but are almost so. Now the matrix can be inverted, but the solution will be unstable. This means that slight alterations to the observed $\mathbf{X}$ values would lead to substantial differences in the estimated values of $\mathbf{a}$. Different measurement errors or a slightly different design sample would have led to different parameter estimates. This problem is termed *multicollinearity*. The instability in the estimated parameters is a problem if these values are the focus of interest—for example, if one wants to know which of the variables is most important in the model. However, it will not normally be a problem as far as predictive accuracy is concerned: although substantially different $\mathbf{a}$ vectors may be produced by slight variations of the data, all of these vectors will lead to similar predictions for most $\mathbf{x}_k$ vectors.

In Chapter 6XXX we remarked that the additive nature of the regression model could be retained while permitting more flexible model forms by replacing the raw x_j by transformations of them. Figure 11.2.1 shows a plot of data collected in an experiment in which a subject performed a physical task at a gradually increasing level of difficulty. The vertical axis shows a measure on the gases expired from the lungs while the horizontal axis shows the oxygen uptake. The non-linearity of the relationship between these two variables

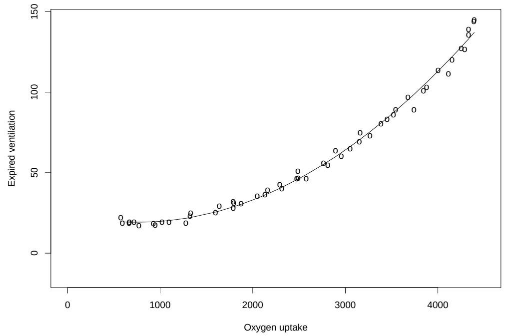

Figure 11.2: The data from Figure 11.2.1 with a model which includes a term in x^2.

is quite clear from the plot. A straight line $y = a_0 + a_1x$ provides a poor fit—as is shown in the figure. The predicted values from this model would only be accurate for x (oxygen uptake) values just above 1000 and just below 4000. (Having said that, the model is not grossly inaccurate—the point made earlier about models linear in x providing reasonable approximations is clearly true.) However, the model $y = a_0 + a_1x + a_2x^2$ gives the fitted line shown in Figure 11.2.1. This model is still linear in the parameters, so that these can be easily estimated using the standard matrix manipulation shown above. It is clear that the predictions obtained from this model are about as good as one can do. The remaining inaccuracy in the model is the irreducible measurement error associated with the variance of y about its mean at each value of x.

11.2.2 A Probabilistic Interpretation of Linear Regression

The informal data analytic route described above allows us to fit a regression model to any data involving a response variable and a set of predictor variables, and obtain a vector of estimated regression coefficients. If our aim were merely to produce a convenient summary of the design data (as, very occasionally, it is) then we could stop there. However, this chapter is concerned with predictive models. Our aim is to go beyond the design data to predict y values for other "out-of-sample" objects. Goodness of fit to the given data is all very well, but we are really interested in fit to future data which arises from the same process, so that our future predictions are as accurate as possible. In order to explore this, we need to embed the above model-building process in a more formal inferential context. To do this, we suppose that each observed value y_i is produced as a sum of weighted predictor variables $\alpha'\mathbf{x}_k$ and a random term ϵ_i which follows a $N(0, \sigma^2)$ distribution independent of other values. (Note that, implicit in this, is the assumption that the variances of the random terms are all the same - σ^2 is the same for all values of the vector of predictor variables. We shall discuss this assumption further below.) The random vector $\mathbf{Y}$ thus takes the form $\mathbf{Y} = \mathbf{X}\alpha + \epsilon$. The observed $\mathbf{y}$ vector in Equation 11.2 is a realization from this distribution. The components of ϵ are often called *errors*. Note that they are different from the residuals, $\mathbf{e}$. An "error" is a random realization from a given distribution, whereas a residual is a difference between a fitted model and an observed y value. Note also that α is different from $\mathbf{a}$. α is represents the underlying and unknown truth, whereas $\mathbf{a}$ gives the values used in a model of the truth.

It turns out that, within this framework the least squares estimate $\mathbf{a}$ above is also the maximum likelihood estimate. Furthermore, the covariance matrix of the estimate $\mathbf{a}$ obtained above is $(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$. In the case of a single predictor variable, this gives

$$\left(1 + \frac{n\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)\frac{\sigma^2}{n} \tag{11.6}$$

for the variance of the intercept term and

$$\frac{\sigma^2}{\sum (x_i - \bar{x})^2} \tag{11.7}$$

for the variance of the slope. Here $\bar{x}$ is the sample mean of the single predictor variable. The diagonal elements of the covariance matrix for $\mathbf{a}$ above give the variances of the regression coefficients—which can be used to test whether the individual regression coefficients are significantly different from zero. If v_j is the jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$, then the ratio $a_j/\sqrt{v_j}$ can be compared with a $t(n-p-1)$ distribution to see if the regression coefficient is zero. However, as we discuss below, this test only makes sense in the context of the other variables included in the model, and alternative methods, also discussed below, are available for more elaborate model-building exercises.

If $\mathbf{x}_k$ is the vector of predictor variables for a new object, with predicted y value $\hat{y}_k$, then the variance of $\hat{y}_k$ is $\mathbf{x}_k'(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_k\sigma^2$. With one predictor variable, this reduces to $\sigma^2\left(\frac{1}{n} + (x_k - \bar{x})^2/\sum(x_i - \bar{x})^2\right)$. Note that this variance is greater the further $\mathbf{x}_k$ is from the mean of the design sample—the least accurate predictions, in terms of variance, are those in the tails of the distribution of predictor variables. Note also that confidence intervals (see Chapter 5XXX) based on this variance are confidence values for the *predicted value* of y. We might also be interested in (what are somewhat confusingly called) *prediction intervals*, telling us a range of plausible values for the observed y at a given value of x, not a range of plausible values for the predicted value. Prediction intervals must include the uncertainty arising from our prediction and also that arising from the variability of y about our predicted value. This means that the variance above is increased by an extra term σ^2 , yielding $\sigma^2\left(1 + \frac{1}{n} + (x_k - \bar{x})^2/\sum(x_i - \bar{x})^2\right)$.

Display 11.1 *The most important special case of linear regression arises when there is just one predictor variable. Figure 11.2.2 shows a plot of Relative Performance on the vertical axis against Estimated Relative Performance on the horizontal axis for 209 computer CPUs. We can use regression to predict Relative Performance from Estimated Relative Performance. A simple regression of the data as shown in Figure XXX3 gives*

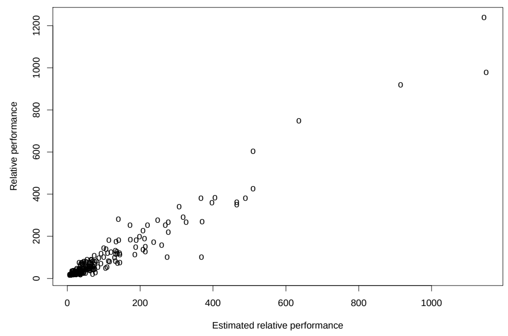

Figure 11.3: A plot of relative performance against estimated relative performance for 209 computer CPUs.

an estimated intercept value of 1.1091 and an estimated regression coefficient of 0.9300. Most modern data analytic packages will give the associated standard errors of the estimates, along with significance tests of the null hypotheses that the true parameters which led to the data are zero. In this case, the standard errors are 3.2989 and 0.0172, respectively, yielding significance probabilities of 0.7371 and 0.0000. From this we would conclude that there is strong evidence that the positive linear relationship is real, but no evidence of a non-zero intercept.

The plot in Figure 11.2.2 shows marked skewness in both variables (they become more sparsely spread towards the top and right of the figure). It is clear that the position of the regression line will be much more sensitive to the precise position of points to the right of the figure than it will be to the position of points to the left. Points which can have a big effect on the conclusion are called points of high leverage—they are points at the extreme values of Estimated Relative Performance in Figure 11.2.2. Points which actually do have a big effect are called influential points. For example, if the rightmost point in Figure 11.2.2 had Relative Performance value 200 (while still having Estimated relative performance value around 1200), it would clearly have a big effect on the regression line. The asymmetry of the leverage of the points in the figure might be regarded as undesirable. One might try to overcome this by reducing the skewness—for example by log transforming both the variables before fitting the regression line.

11.2.3 Interpreting the Fitted Model

The coefficients in a multiple regression model can be interpreted as follows. If the jth predictor variable, x_j, is increased by one unit, while all the other predictor variables are kept fixed, then the response variable y will increase by α_j. The regression coefficients thus tell us the *conditional* effect of each predictor variable, conditional on keeping the other predictor variables constant. This is an important aspect of the interpretation. In particular, the size of the regression coefficient associated with the jth variable will depend on what other variables are in the model. This is clearly especially important if one is constructing models in some kind of sequential way: add another variable and the coefficients of those already in the model will change. (There is an exception to this. If the predictor variables are orthogonal, then the regression coefficients are unaffected by the presence or absence of others in the model. However, this situation is most common in designed experiments, and is rare in the kinds of secondary data analyses encountered in data mining.) The sizes of the regression coefficients tells us the relative importance of the variables, in the sense that one can compare the effects of unit changes. Note also that the size of the effects depends on the chosen units of measurement for the predictor variables. If one measures x_1 in kilometers instead of millimeters, then its associated regression coefficient will be divided by a million. This can make comparisons between variables difficult, so people often work with standardized variables—measuring each predictor variable relative to its standard deviation.

We used the sum of squared errors between the predictions and the observed y values as a criterion through which to choose the values of the parameters in the model. This is the *residual sum of squares* or the *sum of squared residuals*, $\sum e_i^2 = \sum(y_i - \hat{y}_i)^2$. In a sense, the worst model would be obtained if we simply predicted all of the y values by the value $\bar{y}$, the mean of the sample of y values and which is constant relative to the x values (thus, effectively ignoring the inputs to the model and always guessing the output to be the mean of y). The *total sum of squares* is defined as the sum of squared errors for this worst model, $\sum(y_i - \bar{y})^2$. The difference between the residual sum of squares from a model and the total sum of squares is the sum of squares which can be attributed to the regression for that model—it is the *regression sum of squares*. This is the sum of squared differences of the predicted values, $\hat{y}_i$, from the overall mean, $\sum(\hat{y}_i - \bar{y})^2$. The symbol R^2 is often used for the "multiple correlation coefficient," the ratio of the regression sum of squares to total sum of squares:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}. \tag{11.8}$$

A value near 1 tells us that the model explains most of the y variation in the data.

The number of independent components contributing to each sum of squares is called the number of *degrees of freedom* for that sum of squares. The degrees of freedom for the total sum of squares is $n-1$ (one less than the sample size, since the components are all calculated relative to the mean). The degrees of freedom for the residual sum of squares is $n-1-p$ (although there are n terms in the summation, $p+1$ regression coefficients are calculated). The degrees of freedom for the regression sum of squares is p, the difference between the total and residual degrees of freedom.

These sums of squares and their associated degrees of freedom are usefully put together in an analysis of variance table, as in Table 11.2.3, summarizing the decomposition of the totals into components. The meaning of the final column is described below.

Source of variation	Sum of squares	Degrees of freedom	Mean square
Regression	$\sum(\hat{y}_i - \bar{y})^2$	p	$\sum(\hat{y}_i - \bar{y})^2/p$
Residual	$\sum(y_i - \hat{y}_i)^2$	$n-p-1$	$\sum(y_i - \hat{y}_i)^2/(n-p-1)$
Total	$\sum(y_i - \bar{y})^2$	$n-1$	

Table 11.1: The analysis of variance decomposition table for a regression

11.2.4 Inference and Generalization

We have already noted that our real aim in this chapter is one of inference: we want to make statements (predictions) about objects for which we do not know the y values. This means that goodness of fit to the design data is not our real objective. In particular, merely because one has obtained non-zero estimated regression coefficients does not necessarily mean that the variables are related: it could be merely that one's model has captured chance idiosyncrasies of the design sample. This is particularly relevant in the context of data mining where many models may be explored and fitted to the data in a relatively automated fashion. As explained in Chapter 5XXX, we need some way to *test* the model, to see how easily the observed data could have arisen by chance, even if there was no structure in the population the data were collected from. In this case, we need to test whether the population regression coefficients are really zero. (Of course, this is not the only test one might be interested in, but it is the one most often required.) It can be shown that if the values of α_j are all zero (and still making the assumption that the ϵ_i are independently distributed as $N(0, \sigma^2)$), then

$$\frac{\sum(\hat{y}_i - \bar{y})^2/p}{\sum(y_i - \bar{y})^2/(n-p-1)}. \tag{11.9}$$

has an $F(p, n-p-1)$ distribution. This is just the ratio of the two mean squares given in Table 1. The test is carried out by comparing the value of this ratio with the upper critical level of the $F(p, n-p-1)$ distribution. If the ratio exceeds this value the test is significant—and we would conclude that there is a linear relationship between the y and x_j variables (or that a very unlikely event has occurred). If the ratio is less than the critical value we have no evidence to reject the null hypothesis that the population regression coefficients are all zero. Alternatively, many data analysis packages give the probability that a regression sum of squares this large or larger would be observed by chance, if there really was no relationship between y and the x's in the population. (In statistics, this probability is called a p-value. This use of p is not to be confused with its use as the number of predictor variables.)

11.2.5 Model Search and Model Building

We have described an overall test to see if the regression coefficients in a given model are all zero. However, we are more often involved in a situation of searching over model spage or *model building*, in which we examine a sequence of models to find one which is 'best' in some sense. In particular, we often need to examine the effect of adding a set of predictor variables to a set we have already included. Note that this includes the special case of adding just

Source of variation	Sum of squares	Degrees of freedom	Mean square
Regression Model 1	$SS(M)$	p	$SS(M)/p$
Regression Full Model	$SS(M^*)$	q	$SS(M^*)/q$
Difference	$SS(M^*) - SS(M)$	$q - p$	$(SS(M*) - SS(M))/(q-p)$
Residual	$SS(T) - SS(M^*)$	$n - q - 1$	$(SS(T) - SS(M^*))/(n-q-1)$
Total	$SS(T)$	$n - 1$	

Table 11.2: The analysis of variance decomposition table for model building

one extra variable, and that by applying the idea in reverse, it can also handle the situation of removing variables from a model.

In order to compare models we need a measure of goodness-of-fit. Once again, the obvious one is the sum of squared errors between the predictions and the observed y values. Suppose we are comparing two models: a model with p predictor variables (model M) and the largest model we are prepared to contemplate, with say q variables (these will include all the untransformed predictor variables we think might be relevant, along with any transformations of them we think might be relevant), model M^*. Each of these models will have an associated residual sum of squares, and the difference between them will tell us how much better the larger model fits the data than the smaller model. (Equivalently, we could calculate the difference between the regression sum of squares. Since the residual and regression sum of squares sum to the total sum of squares, which is the same for both models, the two calculations will yield the same result.) The degrees of freedom associated with the difference between the residual sums of squares for the two models is $q - p$, the extra number of regression coefficients computed in fitting the larger model, M^*. The ratio between the difference of the residual sums of squares and the difference of degrees of freedom, again give us a mean square—now a mean square for the difference between the two models. Comparison of this with the residual mean square for model M^* gives us an F-test of whether the difference between the models is real or not. Table 11.2.5 illustrates this extension. From this table, the ratio $[(SS(M^*) - SS(M))/(q-p)]\ \ [(SS(T) - SS(M^*))/(n-q-1)]$ is compared with the critical value of an $F(q-p, n-q-1)$ distribution.

This is fine if one has just a few models one wants to compare, but data mining problems are such that often one needs to rely on automatic model building processes. Such methods are available in most modern data mining computer packages. There are various strategies which may be adopted. A basic form is a *forward selection* method, in which variables are added one at a time to an existing model. At each step that variable is chosen from the set of potential variables which leads to the greatest increase in predictive power (measured in terms of reduction of sum of squared residuals), provided the increase exceeds some specified threshold. In principle, the addition is made provided the increase in predictive power is statistically significant, but in practice this is not the case. The variable selection process necessarily involves carrying out many tests, not all independent, so that computing correct significance values is a non-trivial process. The simple significance level based on Table

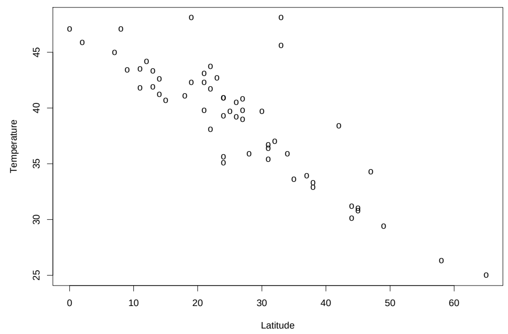

Figure 11.4: Temperature (deg F) against latitude (deg N) for 56 cities in the United States.

11.2.5 does not apply when multiple dependent tests are made.

An opposite strategy to forward selection is *backwards elimination*. One begins with the most complex model one might contemplate (the "largest model," M^*, above) and progressively eliminates variables, selecting them on the basis that they lead to the least increase in sum of squared residuals (again, subject to some threshold). Other variants include combinations of forward selection and backwards elimination. For example, one might add in two variables, eliminate one, add two, remove one, and so on. For data sets where the number of variables p is very large, it may be much more practical computationally to build the model in the forward direction than in the backwards direction.

All of these stepwise methods are attempts to restrict the search of the space of all possible sets of predictor variables, so that the search is manageable. But by restricting the search, it is possible that some highly effective combination of variables may be overlooked. Very occasionally (if the set of potential predictor variables is small), one can examine all possible sets of variables (although, with p variables, there are $2p$-1 possible subsets). The range of search for all possible subsets has been expanded by the use of strategies such as branch and bound, which rely on the monotonicity of the residual sum of squares criterion (it can only decrease when new variables are added). The model building and search strategies outlined above are described in a more general context in Chapter 8XXX.

A couple of cautionary comments are worth making here. First, recall that the coefficients of variables already in the model will change as new variables are added. A variable which is important for one model, may become less so when the model is extended. Secondly, if too elaborate a search is carried out then there is a high chance of overfitting the design set, that is, of obtaining a model which provides a good fit to the design set (small residual sum of squares) but does not predict new data very well. Such issues are discussed in Chapter XXX.

11.2.6 Diagnostics and Model Inspection

Although multiple regression is a very powerful and widely used technique, some of the assumptions might be regarded as restrictive. The assumption that the variance of the y distribution is the same at each vector $\mathbf{x}$ is often inappropriate. (The assumption of equal variances is called *homoscedasticity*. The converse is *heteroscedasticity*.) For example, Figure 11.4 shows the normal average January minimum temperature (in deg F) plotted against the latitude (deg N) for 56 cities in the United States. There is evidence that the variance of the temperature increases with increasing latitude (although the mean temperature seems to decrease). We can still apply the standard least squares algorithm above to estimate parameters in this new situation (and the resulting estimates would still be unbiased if the model form was correct), but one could do better in the sense that it is possible to find estimators with smaller variance. To do this we need to modify the basic method above. Essentially, we need to arrange things so that those values of $\mathbf{x}$ associated with y values with larger variance are weighted less heavily in the model fitting process. This makes perfect sense - it means that the estimator is more influenced by the more accurate values. Formally, this idea leads to a modification of solution Equation 11.5. Suppose that the covariance matrix of the random vector ϵ is $\sigma^2\mathbf{V}$ (previously we took $\mathbf{V} = \mathbf{I}$). The case of unequal variances means that $\mathbf{V}$ is diagonal with terms which are not all equal. Now it is possible (see any standard text on linear algebra) to find a unique nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{P} = \mathbf{V}$. We can use this to define a new random vector $\mathbf{f} = \mathbf{P}^{-1}\epsilon$, and it is easy to show that the covariance matrix of $\mathbf{f}$ is $\sigma^2\mathbf{I}$. Using this idea, we form a new model by premultiplying the old one by $\mathbf{P}^{-1}$:

$$\mathbf{P}^{-1}\mathbf{Y} = \mathbf{P}^{-1}\mathbf{X}\alpha + \mathbf{P}^{-1}\epsilon \tag{11.10}$$

or

$$\mathbf{Z} = \mathbf{W}\beta + \mathbf{f}, \tag{11.11}$$

now of the form required to apply the standard least squares algorithm. If we do this, and then convert the solution back into the original variables $\mathbf{Y}$, we obtain:

$$\mathbf{a} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)\mathbf{X}\mathbf{V}^{-1}\mathbf{y}, \tag{11.12}$$

a *weighted* least squares solution. The variance of this estimated parameter vector $\mathbf{a}$ is $\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\sigma^2$.

Unequal variances of the y distributions for different $\mathbf{x}$ vectors is one way in which the assumptions of basic multiple regression can break down. There are others. What we really need are ways to explore the quality of the model and tools which will enable us to detect where and why the model deviates from the assumptions. That is, we require *diagnostic* tools.

In simple regression, where there is only one predictor variable, one can see the quality of the model from a plot of y against x. Figure 11.4 provides an illustration. More generally, however, when there is more than one predictor variable, such a simple plot is not possible and more sophisticated methods are needed. In general, the key features for examining the quality of a regression model are the residuals, the components of the vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$. If

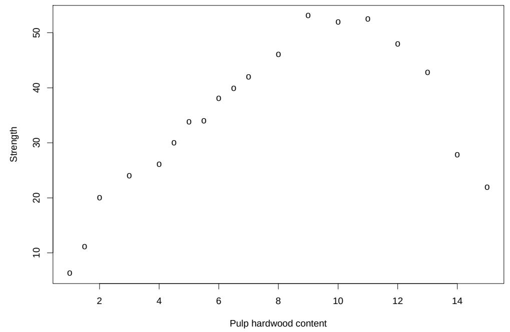

Figure 11.5: A plot of tensile strength of paper against the percentage of hardwood in the pulp.

there is a pattern to these it tells us that the model is failing to explain the distribution of the data. Various plots involving the residuals are used, including plotting the residuals against the fitted values, plotting standardized residuals (obtained by dividing the residuals by their standard errors) against the fitted values, and plotting the standardized residuals against standard normal quantiles. (The latter are 'normal probability plots'. If the residuals are approximately normally distributed, then the points in this plot should lie roughly on a straight line.) Of course, interpreting some of the diagnostic plots requires practice and experience.

At this point it is convenient to make a general cautionary comment, which applies to all predictive models. Such models are only valid within the bounds of the data. It can be very risky to extrapolate beyond the data. A very simple example is given in Figure 11.5. This shows a plot of the tensile strength of paper plotted against the percentage of hardwood in the pulp from which the paper was made. But suppose only those samples with pulp values between 1 and 9 had been measured. The figure shows that a straight line would provide quite a good fit to this subset of the data. For new samples of paper, with pulp values lying between 1 and 9, quite good prediction of the strength could legitimately be expected. But the figure also shows, strikingly clearly, that our model would produce predictions which were seriously amiss if we used it to predict the strength of paper with pulp values greater than 9. Only within the bounds of our data is the model trustworthy. Figure XXX1 in Chapter 3XXX gives another example. Here the graph shows the number of credit cards in circulation each year. A straight line fitted to years 1985 to 1990 would provide a good fit - but if predictions beyond those years was based on this model, disaster would follow.

These examples are particularly clear—but they involve just a few data points and a single predictor variable. In data mining applications, with large data sets and many variables, things may not be so clear. Caution needs to be exercised when making predictions.

11.3 Generalized linear models

Section 11.2 described the *linear model*, in which the random response variable was decomposed into two parts: a weighted sum of the predictor variables and a random component: $Y_i = \sum \alpha_i x_{ij} + \epsilon_i$. For inferential purposes we also assumed that the ϵ_i were independently distributed as $N(0, \sigma^2)$. We can write this another way, which permits convenient generalization, splitting the description of the model into three parts:

- (i) The Y_i are independent random variables, with distribution $N(\mu_i, \sigma^2)$.
- (ii) The parameters enter the model in a linear way via the sum $\nu_i = \sum \alpha_j x_{ij}$.
- (iii) The ν_i and μ_i are linked by $\nu_i = \mu_i$

This permits two immediate generalizations, while retaining the advantages of the linear combination of the parameters. Firstly, in (i) we can relax the requirement that the random variables follow a normal distribution. Secondly, we can generalize the link expressed in (iii), so that some other *link* function $g(\mu_i) = \nu_i$ relates the parameter of the distribution to the linear term $\nu_i = \sum \alpha_j x_{ij}$. These extensions result in what are called *generalized linear models*. They are one of the most important advances in data analysis of the last two decades. As we shall see, such models can also be regarded as fundamental components of feed forward neural networks.

One of the most important kinds of generalized linear model for data mining is *logistic regression*. We have already encountered this in Chapter 10, but we describe it in rather more detail here, and use it as an illustration of the ideas underlying generalised linear models. In many situations the response variable is not continuous, as we have assumed above, but is a proportion: the number of flies from a given sample which die when exposed to an insecticide, the proportion of questions people get correct in a test, the proportion of oranges in a carton which are rotten. The extreme of this arises when the proportion is out of 1, that is, the observed response is binary: whether or not an individual insect dies, whether or not a person gets a particular one of the questions right, whether or not an individual orange is rotten. This is exactly the situation we discussed in Chapter 10, though here we embed it in a more general context.

We are now dealing with a binary response variable, with the random variable taking values 0 or 1 corresponding to the two possible outcomes. We shall assume that the probability that the ith individual yields the value 1 is p_i, and that the responses of different individuals are independent. This means that the response for the ith individual follows a Bernoulli distribution:

$$p(Y_i = y_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \tag{11.13}$$

where here $y_i \in \{0, 1\}$. For logistic regression, this is the generalization of (i) above: the Bernoulli distribution is replacing the normal distribution.

Our aim is to formulate a model for the probability that an object with predictor vector $\mathbf{x}_k$ will take value 1. That is, we want a model for the mean value of the response, the probability p_k. We could use a linear model - a weighted sum of the predictor variables. However, this would not be ideal. Most obviously, a linear model can take values less than 0 and greater than 1. This suggests that we need to modify to the model to include a nonlinear aspect. We achieve this by transforming the probability, nonlinearly, so that it can be modeled by a linear combination. That is, we use a nonlinear link function in (iii).

A suitable function (not the only possible one) is a *logistic* (or *logit*) link function, in which

$$g(p(x)) = \log \frac{p(x)}{1 - p(x)}. \tag{11.14}$$

As p varies from 0 to 1, this clearly varies from $-\infty$ to ∞, matching the potential range of $\sum \alpha_j x_{ij}$.

One of the advantages of the logistic link function over alternatives (for example, over a *probit* function, which was widely used in the past), is that it permits convenient interpretation. For example:

- The ratio $\frac{p(x)}{1-p(x)}$ in the transformation is the familiar *odds* that a 1 will be observed and $\log \frac{p(x)}{1-p(x)}$ is the *log odds*.
- Given a new vector of predictor variables $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})$, the predicted probability of observing a 1 is derived from $\log \frac{p(\mathbf{x}_k)}{1-p(\mathbf{x}_k)}$. The effect on this of changing the jth predictor variable by one unit is simply α_j . Thus the coefficients tell us the difference in log odds—or, equivalently, the log odds ratio resulting from the two values. From this it is easy to see that e^{α_j} is the factor by which the odds changes when the jth predictor variable changes by one unit. (c.f. the discussion of the effect of a unit change of one variable in the multiple regression case discussed in Section 11.2).

Display 11.2 *Two minutes into its flight on January 29th, 1996 the space shuttle Challenger exploded, killing everyone on board. The two booster rockets for the shuttle are made of several pieces, with each of three joints sealed with a rubber 'O-ring', making six rings in total. It was known that these O-rings were sensitive to temperature. Records of the proportion of O-rings damaged in previous flights were available, along with the temperatures on those days. The lowest previous temperature was 53*deg*F. On the day of the flight the temperature was 31*deg*F, so there was much discussion about whether the flight should go ahead. One argument was based on an analysis of the seven previous flights which had resulted in damage to at least one O-ring. A logistic regression to predict the probability of failure from temperature led to a slope estimate of 0.0014 with a standard error of 0.0498. From this, the predicted logit of the probability of failure at 31*deg*F is 1.3466, yielding a predicted probability 0.206. The slope in this model is positive, suggesting that, if anything, the probability of failure is lower at low temperatures. However, this slope is not significantly different from zero, so that there is little evidence for a relationship between failure probability and temperature.*

*This analysis is far from ideal. First, 31*deg*F is far below 53*deg*F, so one is extrapolating beyond the data—a practice we warned against above. Secondly, there is valuable*

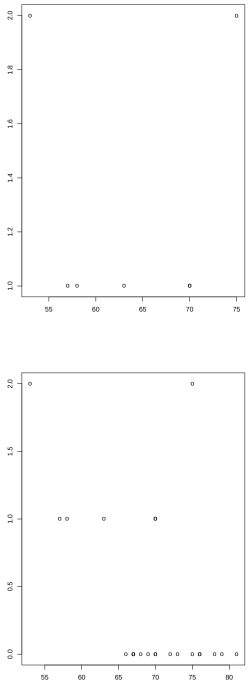

Figure 11.6: Number of O-rings damaged (vertical axis) against temperature on day of flight, (a) data examined before the flight, and (b) the complete data.

information in the 16 flights which had not resulted in O-ring damage. This is immediately obvious from a comparison of Figure 11.6(a), which shows the numbers damaged for the seven flights above (vertical axis) against temperature (horizontal axis), and Figure 11.6(b), which shows the number for all 23 flights. These 16 flights all took place at higher temperatures. The second figure suggests that the relationship might, in fact, have a negative slope. A logistic model fitted to the data in Figure XXX6(b) gave a slope estimate of -0.1156, with a standard error of -2.46 (and an intercept estimate of 5.08 with standard error of 3.05). From this the predicted probability at 31 deg$F$ *is 0.817. This gives a rather different picture, one which could have been deduced before the flight if all the data had been studied.*

Generalized linear models thus have three features: 1) The $Y_i, i = 1, \ldots, n$, are independent random variables, with the same exponential family distribution (see below). 2) The predictor variables are combined in a form $\nu_i = \sum a_i x_{ij}$, called the *linear predictor*. 3) The mean μ_i of the distribution for a given predictor vector is related to the linear combination in (ii) through the link function $g(\mu_i) = \nu_i = \sum a_j x_{ij}$

The *exponential family* of distributions is an important family which includes the normal, the Poisson, the Bernoulli, and the binomial distributions. Members of this family can be expressed in the general form

$$f(y;\theta,\phi) = e^{\frac{(y\theta - b(\theta)}{a(\phi)+c(y,\phi)}} \tag{11.15}$$

If ϕ is known, then θ is called the *natural* or *canonical* parameter. When, as is often the case, $a(\phi) = \phi$, ϕ is called the *dispersion* or *scale* parameter. A little algebra reveals that the mean of this distribution is given by $b'(\theta)$ and the variance by $a(\phi)b''(\theta)$. Note that the variance is related to the mean via $b''(\theta)$, and this, expressed in the form $V(\theta)$, is sometimes called the *variance function*. In the model as described in (i) to (iii) above, there are no restrictions on the link function. However (and this is where the exponential family comes in), things simplify if the link function is chosen to be the function expressing the canonical parameter for the distribution being used as a linear sum. For multiple regression this is simply the identity distribution and for logistic regression it is the logistic transformation presented above. For *Poisson regression*, in which the distribution in (i) is the Poisson distribution, the canonical link is the log link $g(u) = log(u)$. Prediction from a generalized linear model requires the inversion of the relationship $g(\mu_i) = \sum a_j x_{ij}$.

The algorithms in least squares estimation were very straightforward, essentially only involving matrix inversion. For generalized linear models, however, things are more complicated: the nonlinearity means that an iterative scheme has to be adopted. We will not go into details of the mathematics here, but it is not difficult to show that the maximum likelihood solution is given by solving the equations

$$\sum_{i=1}^{n} \frac{x_{ij}\left(y_i - \mu_i\right)}{a_i\left(\phi\right) V\left(\mu_i\right) g'\left(\mu_i\right)} = 0, \qquad j = 1, \ldots, p \tag{11.16}$$

where the i subscripts in $a_i(\phi)$ and μ_i are in recognition of the fact that these vary from data point to data point. Standard application of the Newton Raphson method leads to iteration of the equations

$$a^{(s)} = a^{(s-1)} - \mathbf{M}_{\mathbf{s-1}}^{-1}\mathbf{u}_{\mathbf{s-1}} \tag{11.17}$$

where $a^{(s)}$ represents the vector of values of $(a_1, \ldots, a_p)$ at the sth iteration, $\mathbf{u}_{\mathbf{s-1}}$ is the vector of first derivatives of the log likelihood, evaluated at $a^{(s-1)}$, and $\mathbf{M}_{\mathbf{s-1}}$ is the matrix of second derivatives of the log likelihood, again evaluated at $a^{(s-1)}$.

An alternative method, the method of "scoring" (this is a traditional name, and is not to be confused with our use of the word "socre", though the meaning is similar), replaces $\mathbf{M}_{\mathbf{s-1}}$ by the matrix of expected second derivatives. The iterative steps of this method can be expressed in a form similar to the weighted version, Equation 11.12, of the standard least squares matrix solution, Equation 11.5:

$$\left(\mathbf{X}'\mathbf{W}_{(\mathbf{s-1})}\mathbf{X}\right) a^{(s)} = \mathbf{X}'\mathbf{W}_{(\mathbf{s-1})}\mathbf{z}_{(\mathbf{s-1})}, \tag{11.18}$$

where $\mathbf{W}_{(\mathbf{s-1})}$ is a diagonal matrix with iith element $\partial\mu_i/\partial\nu_i)^2/var(Y_i)$ evaluated at $a^{(s-1)}$ and $\mathbf{z}_{(\mathbf{s-1})}$ is a vector with ith element $\sum_j x_{ij}a_j + (y_i - \mu_i)\partial\nu_i/\partial\mu_i$ again evaluated at $a^{(s-1)}$. Given the similarity of this to Equation 11.12 it will hardly be surprising to learn that this method is called *iteratively weighted least squares*.

We need a measure of the goodness of fit of a generalized linear model, analogous to the sum of squares used for linear regression. Such a measure is the *deviance* of a model. In fact, the sum of squares is the special case of deviance when applied to linear models. We introduced the deviance for a model M in Chapter 5XXX as $D(M) = -2\left(\log L(M;Y) - \log L(M^*;Y)\right)$, essentially the difference between the log likelihood of model M and the log likelihood of the largest model we are prepared to contemplate, M^*. Deviance can be decomposed like the sum of squares to permit exploration of classes of models.

Display 11.3 *In a study of ear infections in swimmers, 287 swimmers were asked if they were frequent ocean swimmers, whether they preferred beach or non-beach, their age, their sex, and also the number of self-diagnosed ear infections they had had in a given period. The last variable here is the response variable, and a predictive model is sought, in which this number can be predicted from the other variables. Clearly standard linear regression would be inappropriate: the response variable is discrete and, being a count, is unlikely to look remotely like a normal distribution. Likewise, it is not a proportion, it is not bounded between 0 and 1, so it would be inappropriate to model it using logistic regression. Instead, it is reasonable to assume that the response variable follows a Poisson distribution, with parameter depending on the value of the predictor variables. Fitting a generalised linear model to predict the number of infections from the other variables, with the response following a Poisson distribution and using a log function for the link, led to the analysis of deviance table shown below.*

	d.f.	deviance	mean deviance	deviance ratio
Regression	4	1.67	0.4166	0.42
Residual	282	47.11	0.1671	
Total	286	48.78	0.1706	
Change	-4	-1.67	0.4166	0.42

To test the null hypothesis of no predictive relationship between the response variable and the predictors, we compare the value of the regression deviance (1.67, from the top of the second column of numbers) with the chi-squared distribution with 4 degrees of freedom (given at the top of the first column of numbers). This gives a p-value of 0.7962. This is far from small, suggesting that there is little evidence that the response variable is related to the predictor variables. Not all data necessarily lead to a model which gives accurate predictions!

Before leaving this section, it is worth noting a property of Equations 11.16. Although these were derived on the assumption that the random variables followed an exponential family distribution, examination reveals that these estimating equations only make use of the means μ_i, the variances $a_i(\phi)V(\mu_i)$, as well as the link function and the data values. There is nothing about any other aspect of the distributions. This means that even if we are not prepared to make tighter distributional assumptions, we can still estimate the parameters in the linear predictor $\nu_i = \sum a_i x_{ij}$. Because no full likelihood has to be formulated in this approach, it is termed *quasi-likelihood estimation*. Once again, of course, iterative algorithms are needed.

11.4 Artificial neural networks

Artificial neural networks (ANNs) are one of a class of highly parameterized statistical models which have attracted considerable attention in recent years (other such models are outlined in later sections). In the present context, we will only be concerned with *feed forward neural networks* or *multi-layer perceptrons*. In this section, we can barely scratch the surface of this topic and suitable further reading is suggested below. The fact that ANNs are highly parameterised makes them very flexible, so that they can accurately model relatively small irregularities in functions. On the other hand (see Chapter XXX) this flexibility means that there is a serious danger of overfitting. Indeed, early (by which is meant during the 1980s) work was characterised by inflated claims when such networks were overfitted to design sets, and predictions of future performance based on the design set performance. In recent years strategies have been developed for overcoming this problem, resulting in a very powerful class of predictive models.

To set ANNs in context, recall that the generalised linear models of the previous section formed a linear combination of the predictor variables, and transformed this via a nonlinear transformation. Feedforward ANNs adopt this as the basic element. However, instead of using just one such element, they use layers of many. The outputs from one layer–the transformed linear combinations from each basic element–serve as inputs to the next layer. In this next layer the inputs are combined in exactly the same way–each element forms a weighted sum, which is then nonlinearly transformed.

Mathematically, for a network with just one layer of transformations between the input variables x and the final transformation f yielding the output y (one *hidden* layer), we have

$$y = \sum_k w_k^{(2)} f_k^{(2)} \left(\sum_j w_j^{(1)} x_j \right) \tag{11.19}$$

Here the w are the weights in the linear combinations and the f are the nonlinear transformations. The nonlinearity of these transformations is essential, since otherwise the model reduces to a nested series of linear combinations of linear combinations—which is simply a linear combination. The term "network" derives from a graphical representation of this structure in which the predictor variables and each weighted sum are nodes, with edges connecting the terms in the summation to the node.

There is no limit to the number of layers which can be used, though it can be proven that a single hidden layer (with enough nodes in that layer) is sufficient to model any continuous functions. Of course, the practicality of this will depend on the available data, and it might be convenient for other reasons (such as interpretability) to use more than one hidden layer. There are also generalisations, in which layers are skipped, with inputs to a node coming not only from the layer immediately below but also from other lower layers.

The earliest forms of ANN used threshold logic units as the nonlinear transformations: the output was 0 if the weighted sum of inputs was below some threshold and 1 otherwise. However, there are mathematical advantages to be gained by adopting differentiable forms for these functions. In applications, the two most common forms seem to be logistic $f(x) = e^x/(1+e^x)$ and hyperbolic tangent $f(x) = tanh(x)$ transformations of the weighted sums.

We saw, when we moved from simple linear models to generalized linear models, that estimating the parameters became more complicated. A further extra level of complication occurs when we move from simple generalized linear models to ANNs. This will probably not come as a surprise, given the number of parameters (these now being the weights in the linear combinations) in the model and the fundamental nonlinearity of the transformations. As a consequence of this, neural network models can be slow to estimate. This can limit their applicability in data mining problems involving large data sets. (But slow estimation and convergence is not all bad. There are stories within the ANN folklore relating how severe overfitting by a flexible model has been avoided by accident, simply because the estimation procedure was stopped early.)

Various estimation algorithms have been proposed. A popular approach is to minimize the sum of squared deviations (again!) between the output and predicted values by steepest descent on the weight parameters. This can be expressed as a sequence of steps in which the weights are updated, working from the output node(s) back to the input nodes. For this reason, the method is called *back-propagation*.

Other criteria have also been used. When y takes only two values (so that the problem is really one of supervised classification, as discussed in Chapter 10XXX) the sum of squared deviations is rather unnatural (being equivalent to log-likelihood for normal distributions) and a more natural one, based on log-likelihood for Bernoulli data is

$$\sum_i \left[y_i \log \frac{\hat{y}_i}{y_i} - (1-y_i) \log \frac{(1-\hat{y}_i)}{(1-y_i)} \right] \tag{11.20}$$

In practical applications, with reasonably sized data sets, the precise choice of criterion seems to make little difference.

The vast amount of work on neural networks in recent years, by a diverse range of intellectual communities, has led to the rediscovery of many concepts and phenomena already well-known and understood in other areas. It has also led to the introduction of unnecessary new terminology.

11.5 Other highly parameterized models

The characterizing feature of neural networks is that they provide a very flexible model with which to approximate functions. Partly because of this power and flexibility, but also partly because of the appeal of their name and its implied promise, they have attracted a great deal of media attention. However, they are not the only class of flexible models. Others, in some cases with an approximating power equivalent to that of neural networks, have also been developed. Some of these have advantages as far as interpretation and estimation goes. In this section we briefly outline two of the more important classes of flexible model. Others are mentioned in the penultimate section of this chapter.

11.5.1 Generalized additive models

We have seen how the generalized linear model extends the ideas of linear models. Yet further extension arises in the form of *generalized additive models*. These replace the simple weighted sums of the predictor variables by weighted sums of transformed versions of the predictor variables. To achieve greater flexibility, the relationships between the response variable and the predictor variables are estimated nonparametrically (for example, by kernel or spline smoothing, as outlined in Chapter 4XXX), so that the generalized linear model form $g(\mu_i) = \sum \alpha_j x_{ij}$ becomes $g(\mu_i) = \sum \alpha_j f_j(x_{ij})$. The right hand side here is sometimes termed the *additive predictor*. Such models take to the nonparametric limit the idea of extending the scope of linear models by transforming the predictor variables. Generalized additive models of this form retain the merits of linear and generalized linear models. In particular, how g changes with any particular predictor variable does not depend on how other predictor variables change: interpretation is eased. Of course, this is at the cost of assuming that such an additive form does provide a good approximation to the "true" surface. The model can be readily generalized by including multiple predictor variables within individual f components of the sum, but this is at the cost of relaxing the simple additive interpretation. The additive form also means that we can examine each smoothed predictor variable separately, to see how well it fits the data.

In the special case when g is the identity function (so we are merely discussing an *additive model*, rather than a *generalized* additive model) appropriate smoothing functions can be found by a *backfitting* algorithm. If the additive model $y_i = \sum \alpha_j f_j(x_{ij}) + \epsilon_i$ is correct, then $f_k(X_k) = E\left(Y - \sum_{j \neq k} \alpha_j f_j(X_{ij}) | X_k\right)$. This leads to an iterative algorithm in which, at each step the "partial residuals" $y - \sum_{j \neq k} \alpha_j f_j(x_{ij})$ for the kth predictor variable are smoothed, cycling through the predictor variables until the smoothed functions do not change. The precise details will, of course, depend on the choice of smoothing method: kernel, spline, or whatever.

To extend this from additive to generalized additive models, we make the same extension as in above, where we extended the ideas from linear to generalized linear models. In Section 11.3XXX we outlined the iteratively weighted least squares algorithm for fitting generalized linear models. We showed that this was essentially an iteration of a weighted least squares solution applied to an "adjusted" response variable, defined by $\sum_j x_{ij} a_j + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i}$. For generalized additive models, instead of the weighted linear regression we adopt an algorithm for fitting a weighted additive model.

Display 11.4

Display XXX: *Sometimes blood pressure is deliberately lowered during surgery, using drugs. Once the operation is completed, and the administration of the drug discontinued, it is desirable that the blood pressure should return to normal as soon as possible. The data in this example relate to how soon (in minutes) systolic blood pressure returned to 100 mm of mercury after discontinuing the medication. There are two predictor variables: the log of the dose of the particular drug used and the average systolic blood pressure of the patient during administration of the drug. A generalised additive model was fitted, using splines (in fact, cubic B-splines) to effect the smoothing. Figures 11.7 and 11.8 show, respectively, a plot of the transformed Log(dose) against observed Log(dose) values and a plot of the transformed blood pressure during administration against the observed values. (There is some nonlinearity evident in both these plots—although that in the Log(dose) plot seems to be attributable to a single point.). Predictions to new data points are made by adding to together the predictions from each of these components separately.*

11.6 Projection pursuit regression

Projection pursuit regression models can be proven to have the same ability to estimate arbitrary functions as neural networks, but they are not as widely used. This is perhaps unfortunate, since estimating their parameters can have advantages over the neural network situation. The additive models of the last section essentially focus on individual variables (albeit, transformed versions of these). Such models can be extended so that each additive component involves several variables, but it is not clear how best to select such subsets. If the total number of available variables is large, then one may also be faced with a combinatorial explosion of possibilities.

The basic projection pursuit regression model takes the form

$$Y = \alpha_0 + \sum f_j \left(\alpha_k^T \mathbf{X}\right) + \varepsilon \qquad (11.21)$$

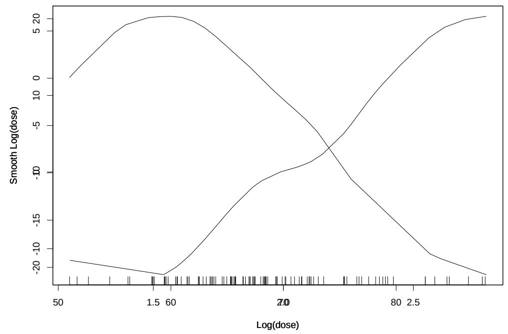

Figure 11.7: The transformation function of log(dose) in the model for predicting time for blood pressure to revert to normal.

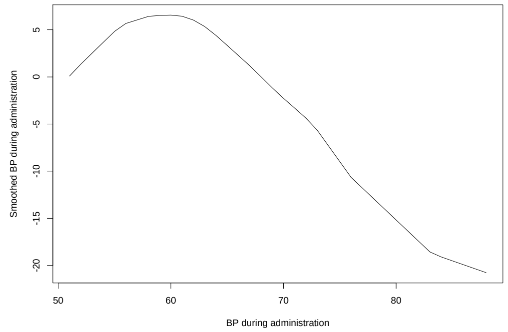

Figure 11.8: The transformation function of blood pressure during administration in the model for predicting time for blood pressure to revert to normal.

This has obvious close similarities to the neural network model–it is a linear combination of (potentially nonlinear) transformations of linear combinations of the raw variables. Here, however, the f functions are not constrained (as in neural networks) to take a particular form, but are usually found by smoothing, as in generalized additive models. This makes them a generalization of neural networks. Various forms of smoothing have been used, including spline methods, Friedman's "supersmoother" (which makes a local linear fit about the point where the smooth is required), and various polynomial functions.

The term "projection pursuit" arises from the viewpoint that one is projecting $\mathbf{X}$ in direction α_k, and then seeking directions of projection which are optimal for some purpose. (In this case, optimal as components in a predictive model.)

Various algorithms have been developed to estimate the parameters. In one, components of the sum are added sequentially up to some maximum value, and then sequentially dropped, each time selecting on the basis of least squares fit of the model to the data. For a given number of terms, the model is fitted using standard iterative procedures to estimate the parameters in the α_k

11.7 Further Reading

Traditional linear regression is covered in depth in the classic book of Draper and Smith (1981), as well as in innumerable other texts. The seminal text on generalised linear models is that of McCullagh and Nelder (1989) and a comprehensive outline of generalized additive models is given in the book by Hastie and Tibshirani (1990). Projection pursuit regression (PPR) was introduced by Friedman and Stuetzle (1981) and theoretical approximation results are given (for example) in Diaconis and Shashahani (1984). A flexible model for multivariate regression called MARS (Multivariate Adaptive Regression Splines) was introduced by Friedman (1991). This which is a good example of a class of very flexible data-driven function approximation algorithms originating in statistics in recent years.

Good introductions to neural networks are given by Bishop (1995), Ripley (1996), and Fine (1999). Ripley's text is particularly noteworthy in that it includes an integrated and extensive discussion of many techniques from the fields of neural networks, statistics, machine learning, and pattern recognition (unlike most texts which tend to focus on one or two of these areas).

The computer CPU data set, the oxygen uptake data set, the ear infections in swimmers data set, and the blood pressure after surgery data are given in Hand *et al* (1994). The temperature and latitude data are from Peixoto (1990). The space shuttle data are reproduced in Chatterjee *et al* (1995).

References

Bishop C.M. (1995) *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Chatterjee, S., Handcock, M.S., and Simonoff, J.S. (1995) *A Casebook for a First Course in Statistics and Data Analysis*, New York: John Wiley and Sons.

Diaconis, P. and Shahshahani, M. (1984) On non-linear functions of linear combinations. *SIAM Journal of Scientific Computing*, 5, 175–191.

Draper, N. R., and Smith, H. (1981) *Applied Regression Analysis*, New York: John Wiley and Sons.

Fine, T.L. (1999) *Feedforward Neural Network Methodology*, NewYork: Springer.

Friedman, J. H. (1991) Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 19, 1–141.

Friedman, J. H. and Stuetzle, W (1981) Project pursuit regression. *Journal of the American Statistical Association*, 76, 817-823.

Hand D.J., Daly F., Lunn A.D., McConway K.J., and Ostrowski E. (eds.) (1994) *A Handbook of Small Data Sets*. London: Chapman and Hall.

Hastie T.J. and Tibshirani R.J. (1990) *Generalized Additive Models*. London: Chapman and Hall.

McCullagh P. and Nelder J.A. (1989) *Generalised Linear Models*. (2nd ed.), London: Chapman and Hall.

Peixoto J.L. (1990) A property of well-formulated polynomial regression models. *American Statistician*, **44**, 26-30.

Ripley B.D. (1996) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.