An immediate generalization is to increase the number of explanatory variables, setting

$$y_j = \beta_1 x_{j1} + \cdots + \beta_p x_{jp} + \varepsilon_j = x_j^T \beta + \varepsilon_j,$$

where $x_j^T = (x_{j1}, \ldots, x_{jp})$ is a $1 \times p$ vector of explanatory variables associated with the jth response, β is a $p \times 1$ vector of unknown parameters and ε_j is an unobserved error accounting for the discrepancy between the observed response y_j and $x_j^T \beta$. In matrix notation,

$$y = X\beta + \varepsilon, \tag{8.1}$$

where y is the $n \times 1$ vector whose jth element is y_j, X is an $n \times p$ matrix whose jth row is x_j^T, and ε is the $n \times 1$ vector whose jth element is ε_j. The data on which the investigation is to be based are y and X, and the aim is to disentangle systematic changes in y due to variation in X from the haphazard scatter added by the errors ε. Model (8.1) is known as a *linear regression model* with *design matrix* X.

Example 8.1 (Straight-line regression) For the straight-line regression model, (8.1) becomes

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

so X is an $n \times 2$ matrix and β a 2×1 vector of parameters. ■

Example 8.2 (Polynomial regression) Suppose that the response is a polynomial function of a single covariate,

$$y_j = \beta_0 + \beta_1 x_j + \cdots + \beta_{p-1} x_j^{p-1} + \varepsilon_j.$$

For example, we might wish to fit a quadratic or cubic trend in the Venice sea level data, in which case we would have $p = 3$ or $p = 4$ respectively. Then

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{p-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

where X has dimension $n \times p$. ■

A key point is that (8.1) is linear in the parameters β. Polynomial regression can be written in form (8.1) because of its linearity, not in x, but in β.

Example 8.3 (Cement data) Table 8.1 contains data on the relationship between the heat evolved in the setting of cement and its chemical composition. Data on heat evolved, y, for each of $n = 13$ independent samples are available, and for each

Table 8.1 Cement data (Woods *et al.*, 1932): y is heat evolved in calories per gram of cement, and x_1, x_2, x_3, and x_4 are percentage weight of clinkers, with x_1, $3CaO.Al_2O_3$, x_2, $3CaO.SiO_2$, x_3, $4CaO.Al_2O_3.Fe_2O_3$, and x_4, $2CaO.SiO_2$.

Case	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

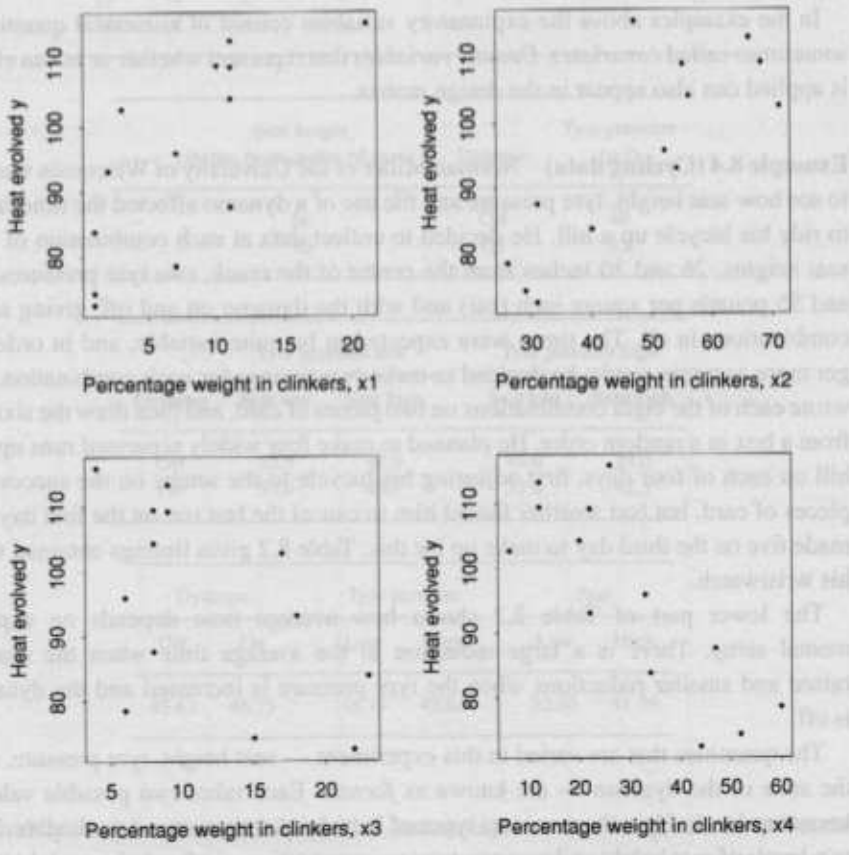

Figure 8.1 Plots of cement data. The variables are heat evolved in calories per gram, y, percentage weight in clinkers of x_1, $3CaO.Al_2O_3$, x_2, $3CaO.SiO_2$, x_3, $4CaO.Al_2O_3.Fe_2O_3$, and x_4, $2CaO.SiO_2$.

sample the percentage weight in clinkers of four chemicals, x_1, $3CaO.Al_2O_3$, x_2, $3CaO.SiO_2$, x_3, $4CaO.Al_2O_3.Fe_2O_3$, and x_4, $2CaO.SiO_2$, is recorded.

Figure 8.1 shows that although the response y depends on each of the covariates $x_1, \ldots, x_4$, the degrees and directions of the dependencies differ.

In this case we might fit the model

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4j} + \varepsilon_j,$$

where Figure 8.1 suggests that β_1 and β_2 are positive, and that β_3 and β_4 are negative. The design matrix has dimension 13×5, and is

$$X = \begin{pmatrix} 1 & 7 & 26 & 6 & 60 \\ 1 & 1 & 29 & 15 & 52 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 10 & 68 & 8 & 12 \end{pmatrix};$$

the vectors y and ε have dimension 13×1 and β has dimension 5×1. ■

In the examples above the explanatory variables consist of numerical quantities, sometimes called *covariates*. *Dummy variables* that represent whether or not an effect is applied can also appear in the design matrix.

Example 8.4 (Cycling data) Norman Miller of the University of Wisconsin wanted to see how seat height, tyre pressure and the use of a dynamo affected the time taken to ride his bicycle up a hill. He decided to collect data at each combination of two seat heights, 26 and 30 inches from the centre of the crank, two tyre pressures, 40 and 55 pounds per square inch (psi) and with the dynamo on and off, giving eight combinations in all. The times were expected to be quite variable, and in order to get more accurate results he decided to make two timings for each combination. He wrote each of the eight combinations on two pieces of card, and then drew the sixteen from a box in a random order. He planned to make four widely separated runs up the hill on each of four days, first adjusting his bicycle to the setups on the successive pieces of card, but bad weather forced him to cancel the last run on the first day; he made five on the third day to make up for this. Table 8.2 gives timings obtained with his wristwatch.

The lower part of Table 8.2 shows how average time depends on experimental setup. There is a large reduction in the average time when the seat is raised and smaller reductions when the tyre pressure is increased and the dynamo is off.

The quantities that are varied in this experiment — seat height, tyre pressure, and the state of the dynamo — are known as *factors*. Each takes two possible values, known as *levels*. Here there are two types of factors: quantitative and qualitative. The two levels of seat height and tyre pressure are quantitative — other values might have been chosen, and more than two levels could have been used — but the dynamo factor has only two possible levels and is qualitative.

An experiment like this, in which data are collected at each combination of a number of factors, is known as a *factorial experiment*. Such designs and their variants

Table 8.2 Data and experimental setup for bicycle experiment (Box *et al.*, 1978, pp. 368–372). The lower part of the table shows the average times for each of the eight combinations of settings of seat height, tyre pressure, and dynamo, and the average times for the eight observations at each setting, considered separately.

Here θ_l is the parameter for model M_l, under which the prior is $\pi(\theta_l \mid M_l)$ and the prior probability of M_l is $\Pr(M_l)$. Formally, (11.25) is just a re-expression of (11.6) in which the parameter splits into two parts, one a model indicator, M_l, and the other the parameters conditional on M_l. In using (11.25) it is crucial that z is the same quantity under all models considered, rather than one whose interpretation depends on the model.

In practice the main obstacle to model averaging is computational. For each model, the integrations involved must usually be done numerically using ideas described in Section 11.3. Furthermore there can be many models in some applications — for example, selecting among 15 covariates in a regression problem gives $2^{15} = 32,768$ models, corresponding to inclusion or exclusion of each covariate separately, without considering outliers, transformations, and so forth. Thus it may be difficult to find the most plausible models, quite apart from the calculations conditional on each model and the difficulties of specifing a prior over model space — giving the same weight to all combinations of covariates will rarely be sensible.

Example 11.18 (Cement data) We fit linear models to the data in Table 8.1 with $n = 13$ observations and four covariates. There are 2^4 possible subsets of the covariates, giving us models $M_1, \ldots, M_{16}$, which for sake of illustration we regard as equally probable *a priori*, though in practice we should hope that a small number of covariates is more likely than a large number. The models are on different parameter spaces, so the discussion in Section 11.2.2 implies that proper, preferably weak, priors should be used. We use the conjugate prior described in Example 11.17, and without loss of generality centre and scale each covariate vector to have average zero and unit variance. We then set V to be the 5×5 matrix with diagonal elements $\phi^2(v, 1, 1, 1, 1)$, where v is the sample variance of y, $\gamma^{\mathrm{T}} = (\overline{y}, 0, 0, 0, 0)$, $\nu = 2.58$, $\tau^2 = 0.28$, and $\phi = 2.85$. This choice implies that the elements of β are independent *a priori*, and should give a weak but proper prior that is consistent between different models and invariant to location and scale changes of the response and explanatory variables.

The marginal density of y under this model is (11.22); for each subset of covariates we use the corresponding submatrix of V. Table 11.6 shows the quantities $2 \log B_{10}$, where $B_{10} = \Pr(y \mid M_1)/\Pr(y \mid M_0)$ is the Bayes factor in favour of a subset of covariates relative to the model with none, the posterior probabilities of each subset, and, for comparison, the residual sums of squares under the usual linear models, which are broadly in line with the probabilities.

Let us try and predict the value of a new response y_+ with covariates $x_+^{\mathrm{T}} = (1, 10, 40, 20, 30)$. Conditional on a particular subset of covariate vectors, the predictive distribution for y_+ is given by (11.23). Figure 11.3 shows these densities for the six models shown in Table 11.6 to have non-negligible support, and the model-averaged predictive density. ■

A different approach to dealing with model uncertainty is to find a plausible model, $f(y \mid \psi)\pi(\psi)$, and then add further parameters λ whose variation allows for the most uncertain aspects of the model, together with a prior that expresses belief about them.

Model	RSS	$2\log B_{10}$	$\Pr(M \mid y)$	a	b
– – – –	2715.8	0.0	0.0000		
1 – – –	1265.7	7.1	0.0000		
– 2 – –	906.3	12.2	0.0000		
– – 3 –	1939.4	0.6	0.0000		
– – – 4	883.9	12.6	0.0000		
1 2 – –	57.9	45.7	0.2027	93.77	2.31
1 – 3 –	1227.1	4.0	0.0000		
1 – – 4	74.8	42.8	0.0480	99.05	2.58
– 2 3 –	415.4	19.3	0.0000		
– 2 – 4	868.9	11.0	0.0000		
– – 3 4	175.7	31.3	0.0002		
1 2 3 –	48.11	43.6	0.0716	95.96	2.80
1 2 – 4	47.97	47.2	0.4344	95.88	2.45
1 – 3 4	50.84	44.2	0.0986	94.66	2.89
– 2 3 4	73.81	33.2	0.0004		
1 2 3 4	47.86	45.0	0.1441	95.20	2.97

Table 11.6 Bayesian prediction using model averaging for the cement data. For each of the 16 possible subsets of covariates, the table shows the log Bayes factor in favour of that subset compared to the model with no covariates and gives the posterior probability of each model. The values of the posterior mean and scale parameters a and b are also shown for the six most plausible models. $(y_+ - a)/b$ has a posterior t density. For comparison, the residual sums of squares are also given.

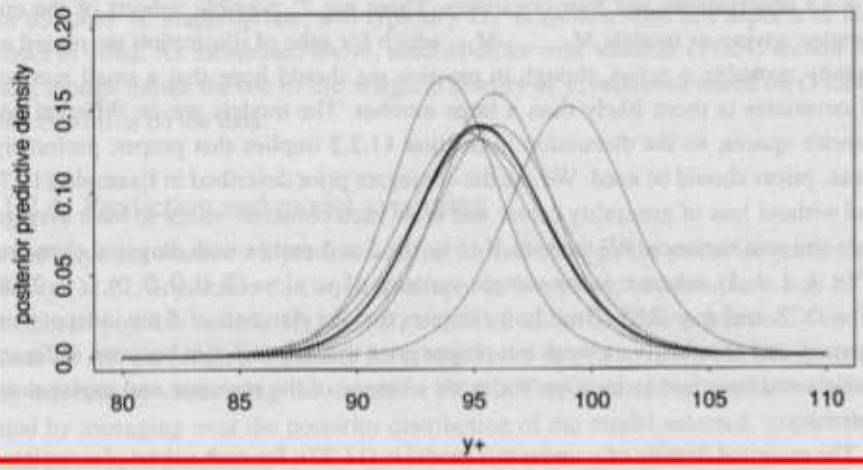

Figure 11.3 Posterior predictive densities for cement data. Predictive densities for y_+ based on individual models are given as dotted curves, and the heavy curve is the averaged prediction from all 16 models.

This gives an expanded model $f(y \mid \psi, \lambda)\pi(\psi, \lambda)$, to which (11.6) is then applied with $\theta = (\psi, \lambda)$.

Exercises 11.2

1 Find elements $\tilde{\theta}$ and $\tilde{J}(\tilde{\theta})$ of the normal approximation to a beta density, and hence check the formulae in Example 11.11. Find also the posterior mean and variance of θ. Give an approximate 0.95 credible interval for θ. How does this differ from a 0.95 confidence interval? Comment.

2 Let $Y_1, \ldots, Y_n$ be a random sample from the uniform distribution on $(0, \theta)$, and take as prior the Pareto density with parameters β and λ,

$$\pi(\theta) = \beta\lambda^{\beta}\theta^{-\beta-1}, \quad \theta > \lambda, \quad \beta, \lambda > 0.$$

(a) Find the prior distribution function and quantiles for θ, and hence give prior one- and two-sided credible intervals for θ. If $\beta > 1$, find the prior mean of θ.