Chapter 6
Regression Analysis

A regression model is any general linear model $Y = X\beta + e$ in which $X'X$ is nonsingular. $X'X$ is nonsingular if and only if the $n \times p$ matrix X has rank p. In regression models, the parameter vector β is estimable. Let $P' = (X'X)^{-1}X'$, then $\beta = P'X\beta$.

The simple linear regression model considered in Section 1 is similar to the one-way ANOVA model considered in Chapter 4 in that the theory of estimation and testing simplifies to the point that results can be presented in simple algebraic formulae. For the general regression model of Section 2 there is little simplification. Section 2 also contains brief introductions to nonparametric regression and generalized additive models as well as an analysis of a partitioned model. The partitioned model is important for two reasons. First, the partitioned model appears in many discussions of regression. Second, the results for the partitioned model are used to establish the correspondence between the standard regression theory presented in Section 2 and an alternative approach to regression, based on best prediction and best linear prediction, presented in Section 3. The approach given in Section 3 assumes that the rows of the model matrix are a random sample from a population of possible row vectors. Thus, in Section 3, X is a random matrix, whereas in Section 2, X is fixed. Section 3 presents the alternative approach which establishes that best linear prediction also yields the least squares estimates. Sections 4 and 5 discuss some special correlation coefficients related to best predictors and best linear predictors. It is established that the natural estimates of these correlation coefficients can be obtained from standard regression results. Section 6 examines testing for lack of fit. Finally, Section 7 establishes the basis of the relationship between polynomial regression and polynomial contrasts in one-way ANOVA.

There is additional material, spread throughout the book, that relates to the material in this chapter. Section 2 examines a partitioned model. Partitioned models are treated in general in Sections 9.1 and 9.2. Chapter 9 also contains an exercise that establishes the basis of the sweep operator used in regression computations. The results of Section 7 are extended in Section 7.3 to relate polynomial regression with polynomial contrasts in a two-way ANOVA. Section 12.2 is an extension of Section 3. Finally, Chapters 13, 14, and 15 are concerned with topics that are traditionally considered part of regression analysis.

R. Christensen, *Plane Answers to Complex Questions: The Theory of Linear Models*,
Springer Texts in Statistics, DOI 10.1007/978-1-4419-9816-3_6,

There are a number of fine books available on regression analysis. Those that I refer to most often are Cook and Weisberg (1999), Daniel and Wood (1980), Draper and Smith (1998), and Weisberg (1985).

6.1 Simple Linear Regression

The model for simple linear regression is $y_i = \beta_0 + \beta_1 x_i + e_i$ or $Y = X\beta + e$, where $\beta' = [\beta_0, \beta_1]$ and

$$X' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}.$$

Often it is easier to work with the alternative model $y_i = \gamma_0 + \gamma_1(x_i - \bar{x}.) + e_i$ or $Y = X_*\gamma + e$, where $\gamma' = [\gamma_0, \gamma_1]$ and

$$X_*' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 - \bar{x}. & x_2 - \bar{x}. & \cdots & x_n - \bar{x}. \end{bmatrix}.$$

Note that $C(X) = C(X_*)$. In fact, letting

$$U = \begin{bmatrix} 1 & -\bar{x}. \\ 0 & 1 \end{bmatrix},$$

we have $X_* = XU$ and $X_* U^{-1} = X$. Moreover, $\mathrm{E}(Y) = X\beta = X_*\gamma = XU\gamma$. Letting $P' = (X'X)^{-1}X'$ leads to

$$\beta = P'X\beta = P'XU\gamma = U\gamma.$$

In particular,

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \gamma_0 - \gamma_1 \bar{x}. \\ \gamma_1 \end{bmatrix}.$$

(See also Example 3.1.2.)

To find the least squares estimates and the projection matrix, observe that

$$X_*'X_* = \begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x}.)^2 \end{bmatrix},$$

$$(X_*'X_*)^{-1} = \begin{bmatrix} 1/n & 0 \\ 0 & 1/\sum_{i=1}^n (x_i - \bar{x}.)^2 \end{bmatrix},$$

and, since the inverse of $X_*'X_*$ exists, the estimate of γ is

$$\hat{\gamma} = (X_*'X_*)^{-1}X_*'Y = \begin{bmatrix} \bar{y}. \\ \sum_{i=1}^n (x_i - \bar{x}.)y_i / \sum_{i=1}^n (x_i - \bar{x}.)^2 \end{bmatrix}.$$

Moreover,

$$\hat{\beta} = U\hat{\gamma},$$

so the least squares estimate of β is

$$\hat{\beta} = \begin{bmatrix} \bar{y}_\cdot - \hat{\gamma}_1 \bar{x}_\cdot \\ \hat{\gamma}_1 \end{bmatrix}.$$

The projection matrix $M = X_*(X_*'X_*)^{-1}X_*'$ is

$$\begin{bmatrix} \frac{1}{n} + \frac{(x_1-\bar{x}_\cdot)^2}{\sum_{i=1}^n (x_i-\bar{x}_\cdot)^2} & \cdots & \frac{1}{n} + \frac{(x_1-\bar{x}_\cdot)(x_n-\bar{x}_\cdot)}{\sum_{i=1}^n (x_i-\bar{x}_\cdot)^2} \\ \vdots & & \vdots \\ \frac{1}{n} + \frac{(x_1-\bar{x}_\cdot)(x_n-\bar{x}_\cdot)}{\sum_{i=1}^n (x_i-\bar{x}_\cdot)^2} & \cdots & \frac{1}{n} + \frac{(x_n-\bar{x}_\cdot)^2}{\sum_{i=1}^n (x_i-\bar{x}_\cdot)^2} \end{bmatrix}.$$

The covariance matrix of $\hat{\gamma}$ is $\sigma^2(X_*'X_*)^{-1}$; for $\hat{\beta}$ it is $\sigma^2(X'X)^{-1}$. The usual tests and confidence intervals follow immediately upon assuming that $e \sim N(0, \sigma^2 I)$.

A natural generalization of the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + e_i$ is to expand it into a polynomial, say

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_{p-1} x_i^{p-1} + e_i.$$

Although polynomial regression has some special features that will be discussed later, at a fundamental level it is simply a linear model that involves an intercept and $p-1$ predictor variables. It is thus a special case of multiple regression, the model treated in the next section.

Exercise 6.1 For simple linear regression, find the MSE, $\mathrm{Var}(\hat{\beta}_0)$, $\mathrm{Var}(\hat{\beta}_1)$, and $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)$.

Exercise 6.2 Use Scheffé's method of multiple comparisons to derive the Working–Hotelling simultaneous confidence band for a simple linear regression line $\mathrm{E}(y) = \beta_0 + \beta_1 x$.

6.2 Multiple Regression

Multiple regression is any regression problem with $p \geq 2$ that is not simple linear regression. If we take as our model $Y = X\beta + e$, we have

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y, \\ \mathrm{Cov}\left(\hat{\beta}\right) &= \sigma^2(X'X)^{-1}X'IX(X'X)^{-1} = \sigma^2(X'X)^{-1}, \\ SSR(X) &= Y'MY = \hat{\beta}'(X'X)\hat{\beta}, \end{aligned}$$

$$SSE = Y'(I-M)Y,$$
$$dfE = r(I-M) = n-p.$$

Since β is estimable, Any linear function $\lambda'\beta$ is estimable. If $Y \sim N(X\beta, \sigma^2 I)$, tests and confidence intervals based on

$$\frac{\lambda'\hat{\beta} - \lambda'\beta}{\sqrt{MSE\ \lambda'(X'X)^{-1}\lambda}} \sim t(dfE)$$

are available.

Suppose we write the regression model as

$$Y = [X_1, \ldots, X_p] \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + e.$$

If we let $\lambda_j' = (0, \ldots, 0, 1, 0, \ldots, 0)$, with the 1 in the jth place, we have

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{MSE\ \lambda_j'(X'X)^{-1}\lambda_j}} \sim t(dfE),$$

where $\lambda_j'(X'X)^{-1}\lambda_j$ is the jth diagonal element of $(X'X)^{-1}$. This yields a test of the hypothesis $H_0 : \beta_j = 0$. It is important to remember that this t test is equivalent to the F test for testing the reduced model

$$Y = [X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p] \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{j-1} \\ \beta_{j+1} \\ \vdots \\ \beta_p \end{bmatrix} + e$$

against the full regression model. The t and F tests for $\beta_j = 0$ depend on all of the other variables in the regression model. Add or delete any other variable in the model and the tests change.

$SSR(X)$ can be broken down into single degree of freedom components:

$$\begin{aligned} SSR(X) &= SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2) + \cdots + SSR(X_p|X_1, \ldots, X_{p-1}) \\ &= R(\beta_1) + R(\beta_2|\beta_1) + R(\beta_3|\beta_1, \beta_2) + \cdots + R(\beta_p|\beta_1, \ldots, \beta_{p-1}). \end{aligned}$$

Of course, any permutation of the subscripts $1, \ldots, p$ gives another breakdown. The interpretation of these terms is somewhat unusual. For instance, $SSR(X_3|X_1, X_2)$ *is not* the sum of squares for testing any very interesting hypothesis about the full regression model. $SSR(X_3|X_1, X_2)$ is the sum of squares needed for testing the model

$$Y = [X_1, X_2] \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + e$$

against the larger model

$$Y = [X_1, X_2, X_3] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + e.$$

This breakdown is useful in that, for instance,

$$SSR(X_{p-1}, X_p | X_1, \ldots, X_{p-2}) = SSR(X_p | X_1, \ldots, X_{p-1}) + SSR(X_{p-1} | X_1, \ldots, X_{p-2}).$$

$SSR(X_{p-1}, X_p | X_1, \ldots, X_{p-2})$ is the sum of squares needed to test $H_0 : \beta_p = \beta_{p-1} = 0$.

Often, multiple regression models are assumed to have a column of 1s in the model matrix. In that case, the model can be written

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{ip-1} + e_i.$$

An analysis of variance table is often written for testing this model against the model

$$y_i = \beta_0 + e_i,$$

for which the model matrix consists only of a column of 1s. The table is given below.

ANOVA

Source	df	SS
β_0	1	$Y'\left(\frac{1}{n}J_n^n\right)Y$
Regression	$p-1$	$Y'\left(M - \frac{1}{n}J_n^n\right)Y$
Error	$n-p$	$Y'(I-M)Y$
Total	n	$Y'Y$

The $SSReg$ from the table can be rewritten as $\hat{\beta}'(X'X)\hat{\beta} - C$, where C is the correction factor, i.e., $C = Y'([1/n]J_n^n)Y = n(\bar{y}_\cdot)^2$.

EXAMPLE 6.2.1. Consider the data given in Table 6.1 on the heating requirements for a factory. There are 25 observations on a dependent variable y (the number of pounds of steam used per month) and 2 independent variables, x_1 (the average atmospheric temperature for the month in $^\circ F$) and x_2 (the number of operating days in the month). The predictor variables are from Draper and Smith (1998).

The parameter estimates, standard errors, and t statistics for testing whether each parameter equals zero are

Table 6.1 Steam Data

Obs. no.	x_1	x_2	y	Obs. no.	x_1	x_2	y
1	35.3	20	17.8270	14	39.1	19	19.0198
2	29.7	20	17.0443	15	46.8	23	20.6128
3	30.8	23	15.6764	16	48.5	20	20.7972
4	58.8	20	26.0350	17	59.3	22	28.1459
5	61.4	21	28.3908	18	70.0	22	33.2510
6	71.3	22	31.1388	19	70.0	11	30.4711
7	74.4	11	32.9019	20	74.5	23	36.1130
8	76.7	23	37.7660	21	72.1	20	35.3671
9	70.7	21	31.9286	22	58.1	21	25.7301
10	57.5	20	24.8575	23	44.6	20	19.9729
11	46.4	20	21.0482	24	33.4	20	16.6504
12	28.9	21	15.3141	25	28.6	22	16.5597
13	28.1	21	15.2673				

Parameter	Estimate	SE	t
β_0	−1.263	2.423	−0.052
β_1	0.42499	0.01758	24.18
β_2	0.1790	0.1006	1.78

As will be seen from the ANOVA table below, the t statistics have 22 degrees of freedom. There is a substantial effect for variable x_1. The P value for β_2 is approximately 0.10. The estimated covariance matrix for the parameter estimates is $MSE\ (X'X)^{-1}$. MSE is given in the ANOVA table below. The matrix $(X'X)^{-1}$ is

	β_0	β_1	β_2
β_0	2.77875	−0.01124	−0.10610
β_1	−0.01124	0.00015	0.00018
β_2	−0.10610	0.00018	0.00479

The analysis of variance table is

ANOVA

Source	df	SS	MS	F
β_0	1	15270.78	15270.78	
Regression	2	1259.32	629.66	298
Error	22	46.50	2.11	
Total	25	16576.60		

The F statistic is huge. There is a very significant effect due to fitting the regression variables (after fitting a mean value to the data). One breakdown of the sum of squares for regression is

$$SSR(X_1|J) = 1252.62$$
$$SSR(X_2|X_1, J) = 6.70.$$

We now partition the model matrix and parameter vector in order to get a multiple regression analogue of the alternative model for simple linear regression. This alternative model is often discussed in regression analysis and is necessary for establishing, in the next section, the relationship between multiple regression and best linear prediction. We can write the regression model as

$$Y = [J, Z] \begin{bmatrix} \beta_0 \\ \beta_* \end{bmatrix} + e,$$

where $\beta_* = [\beta_1, \ldots, \beta_{p-1}]'$ and

$$Z = \begin{bmatrix} x_{11} & \cdots & x_{1p-1} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np-1} \end{bmatrix}.$$

An alternative way to write the model that is often used is

$$Y = \left[J_n, \left(I - \frac{1}{n} J_n^n \right) Z \right] \begin{bmatrix} \gamma_0 \\ \gamma_* \end{bmatrix} + e.$$

The models are equivalent because $C[J, Z] = C[J, (I - [1/n]J_n^n)Z]$. The second model is correcting all of the variables in Z for their means, i.e., the second model is

$$y_i = \gamma_0 + \gamma_1 (x_{i1} - \bar{x}_{\cdot 1}) + \gamma_2 (x_{i2} - \bar{x}_{\cdot 2}) + \cdots + \gamma_{p-1} (x_{ip-1} - \bar{x}_{\cdot p-1}) + e_i.$$

This is analogous to the alternative model considered for simple linear regression. We now find formulae for sums of squares and estimation of parameters based on the adjusted variables $(I - [1/n]J_n^n)Z$. The formulae derived will be used in subsequent sections for demonstrating the relationship between regression and best linear prediction. The formulae are also of interest in that some multiple regression computer programs use them.

The parameters in the two models are related by $\gamma_* = \beta_*$ and $\beta_0 = \gamma_0 - (1/n)J_1^n Z\gamma_*$, cf. Exercise 6.8.10. Since $I - [1/n]J_n^n$ is the ppo onto $C(J_n)^\perp$, it is easily seen that $C[(I - [1/n]J_n^n)Z]$ is the orthogonal complement of $C(J_n)$ with respect to $C(X)$; therefore

$$\begin{aligned} SSReg &= SS(Z|J_n) = Y'(M - [1/n]J_n^n)Y \\ &= Y'(I - [1/n]J_n^n)Z[Z'(I - [1/n]J_n^n)Z]^- Z'(I - [1/n]J_n^n)Y. \end{aligned}$$

In order to parallel the theory of best linear prediction, we use the normal equations to obtain least squares estimates of γ_0 and γ_*. Since

$$\left[J_n, \left(I - \frac{1}{n} J_n^n \right) Z \right]' \left[J_n, \left(I - \frac{1}{n} J_n^n \right) Z \right] = \begin{bmatrix} n & 0 \\ 0 & Z' \left(I - \frac{1}{n} J_n^n \right) Z \end{bmatrix}$$

and

$$\left[J_n, \left(I - \frac{1}{n}J_n^n\right)Z\right]' Y = \begin{bmatrix} \sum_{i=1}^n y_i \\ Z'\left(I - \frac{1}{n}J_n^n\right)Y \end{bmatrix},$$

the least squares estimates of γ_0 and γ_* are solutions to

$$\begin{bmatrix} n & 0 \\ 0 & Z'\left(I - \frac{1}{n}J_n^n\right)Z \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_* \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ Z'\left(I - \frac{1}{n}J_n^n\right)Y \end{bmatrix}.$$

Equivalently $\hat{\gamma}_0 = \bar{y}_\cdot$ and $\hat{\gamma}_*$ is a solution to

$$Z'\left(I - \frac{1}{n}J_n^n\right)Z\gamma_* = Z'\left(I - \frac{1}{n}J_n^n\right)Y.$$

Since $\gamma_* = \beta_*$, we have that $\hat{\gamma}_* = \hat{\beta}_*$. Since $\beta_0 = \gamma_0 - (1/n)J_1^n Z\gamma_* = \gamma_0 - \beta_1\bar{x}_{\cdot 1} - \cdots - \beta_{p-1}\bar{x}_{\cdot p-1}$, we have $\hat{\beta}_0 = \hat{\gamma}_0 - \hat{\beta}_1\bar{x}_{\cdot 1} - \cdots - \hat{\beta}_{p-1}\bar{x}_{\cdot p-1}$.

Finally, from the formula for *SSReg* developed earlier, the normal equations, and the fact that $\hat{\gamma}_* = \hat{\beta}_*$, we get

$$\begin{aligned} SSReg &= Y'\left(I - \frac{1}{n}J_n^n\right)Z\left[Z'\left(I - \frac{1}{n}J_n^n\right)Z\right]^- Z'\left(I - \frac{1}{n}J_n^n\right)Y \\ &= \hat{\beta}_*'Z'\left(I - \frac{1}{n}J_n^n\right)Z\left[Z'\left(I - \frac{1}{n}J_n^n\right)Z\right]^- Z'\left(I - \frac{1}{n}J_n^n\right)Z\hat{\beta}_* \\ &= \hat{\beta}_*'Z'\left(I - \frac{1}{n}J_n^n\right)Z\hat{\beta}_*, \end{aligned}$$

where the last equality follows from the definition of a generalized inverse.

6.2.1 Nonparametric Regression and Generalized Additive Models

In general, regression models assume $y_i = m(x_i) + e_i$, $\mathrm{E}(e_i) = 0$, where x_i is a p vector of known predictor variables and $m(\cdot)$ is some function. If $m(x_i) = x_i'\beta$, the model is linear. If $m(x_i) = h(x_i'\beta)$ for a known function h, the model is generalized linear as discussed in Section 1.4. If $m(x_i) = h(x_i;\beta)$, for a known h depending on the predictor variables x_i and some parameters β, the model is nonlinear regression, see Christensen (1996a, Chapter 18).

In *nonparametric regression*, m is not assumed to fall into any such parametric family. However, by making weak assumptions about $m(x)$, one can often write it as $m(x) = \sum_{j=1}^\infty \beta_j\phi_j(x)$ for some class of known functions $\{\phi_j\}$. When x is a scalar, examples of such classes include polynomials, cosines, and wavelets. In practice, one fits a *linear model*

$$y_i = \sum_{j=1}^p \beta_j\phi_j(x_i) + e_i, \tag{1}$$

where p is large enough to capture the interesting behavior of $m(\cdot)$, see Christensen (2001, Chapter 7). Note that in the linear model, each of the terms $\phi_j(x_i)$ is a predictor variable, i.e., the jth column of the model matrix is $[\phi_j(x_1), \cdots, \phi_j(x_n)]'$.

This "basis function" approach to nonparametric regression involves fitting one large, complicated linear model. Other approaches involve fitting many simple linear models. For example, the *lowess fit (locally weighted scatterplot smoother)* begins estimation of $m(x)$ by fitting a weighted linear regression to some fraction of the x_is that are nearest x with weights proportional to the distance from x. The linear model is used only to estimate one point, $m(x)$. One performs a separate fit for a large number of points x_g, and to estimate all of $m(x)$ just connect the dots in the graph of $[x_g, \hat{m}(x_g)]$. Kernel estimation is similar but the weights are determined by a kernel function and the estimate is just a weighted average, not the result of fitting a weighted line or plane.

Unfortunately, these methods quickly run into a "curse of dimensionality." With the basis function approach and only one predictor variable, it might take, for example, $p = 8$ functions to get an adequate approximation to $m(\cdot)$. No problem! With 2 predictor variables, we could expect to need approximately $p = 8^2 = 64$ predictor variables in the linear model. Given a few hundred observations, this is doable. However, with 5 predictor variables, we could expect to need about $p = 8^5 = 32,768$ functions.

One way to get around this curse of dimensionality is to fit *generalized additive models*. For example, with 3 predictor variables, $x = (x_1, x_2, x_3)'$, we might expect to need $p = 8^3 = 512$ terms to approximate $m(\cdot)$. To simplify the problem, we might assume that $m(\cdot)$ follows a generalized additive model such as

$$m(x) = f_1(x_1) + f_{23}(x_2, x_3) \tag{2}$$

or

$$m(x) = f_{12}(x_1, x_2) + f_{23}(x_2, x_3). \tag{3}$$

If we need 8 terms to approximate $f_1(\cdot)$ and 64 terms to approximate each of the $f_{jk}(\cdot,\cdot)$s, the corresponding linear model for (2) involves fitting only 72 predictor variables, and for model (3) only $128 - 8 = 120$ rather than $8^3 = 512$. Fitting f_{12} and f_{23} typically duplicates fitting of an f_2 term. With the same 8 term approximations and 5 predictor variables, a generalized additive model that includes all of the possible $f_{jk}(\cdot,\cdot)$s involves only 526 terms, rather than the 32,768 required by a full implementation of a nonparametric regression.

I should repeat that my use of 8 terms per dimension is merely an illustration. One might need 5 terms, or 10 terms, or 15 terms. And in two dimensions, one might need more or less than 8^2 terms. But these computations illustrate the magnitude of the problem. It should also be noted that with alternative (nonlinear) methods of fitting generalized additive models it may be necessary to fit lower order terms, i.e., instead of model (3), fit

$$m(x) = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_{12}(x_1, x_2) + f_{23}(x_2, x_3).$$

Another problem with the basis function approach is that it can lead to very strange results if the number of functions p being fitted is too large. It is well known, see for example Christensen (1996a, Section 7.11), that fitting high order polynomials can be a bad thing to do. For x a scalar predictor, by a high order polynomial we mean a polynomial in which the number of polynomial terms p is close to the number of distinct x_i values in the data. If the x_i values are all distinct, a high order polynomial will fit every observed data point almost perfectly. The price of this almost perfect fit to the data is that *between the x_i values* the polynomial can do very weird and inappropriate things. Thus, we have a model that will work poorly for future predictions. This behavior is not unique to polynomials. It occurs with all the standard classes of functions (with the possible exception of wavelets). Similar problems occur when the x_i values are not all distinct.

6.3 General Prediction Theory

General prediction theory provides another approach to regression analysis. It ties in closely with linear model theory but also with generalized linear models, nonlinear regression, and nonparametric regression.

One of the great things about writing a technical book is that if you do enough good mathematics, people will put up with you spouting off once in a while. In this section, we take up the subject of prediction and its application to linear model theory. To me, prediction is what science is all about, and I cannot resist the urge to spout off. If you just want to get to work, skip down to the subsection headed "General Prediction."

6.3.1 Discussion

There is a fundamental philosophical conflict in statistics between Bayesians (who incorporate subjective beliefs into their analyses) and non-Bayesians. On the philosophical issues, the Bayesians seem to me to have much the better of the argument. (Yes, hard as it may be to believe, the author of this book is a Bayesian.) However, in the practice of statistics, the non-Bayesians have carried the day. (Although I believe that the trend is shifting.) Perhaps the most difficult aspect of Bayesian statistics for non-Bayesians to accept is the incorporation of subjective beliefs. Many scientists believe that objectivity is of paramount importance, and classical statistics maintains the semblance of objectivity. (In fact, classical statistics is rife with subjective inputs. Choosing an experimental design, choosing independent variables to consider in regression, and any form of data snooping such as choosing contrasts after looking at the data are examples.)

As Smith (1986) has pointed out, this concept of objectivity is very elusive. Objectivity is almost indistinguishable from the idea of consensus. If all the "clear

thinking" people agree on something, then the consensus is (for the time being) "objective" reality.

Fortunately, the essence of science is not objectivity; it is repeatability. The object of scientific statistical inference is not the examination of parameters (too often created by and for statisticians so that they have something to draw inferences about). The object of scientific statistical inference is the (correct) prediction of future observable events. (I bet you were wondering what all this had to do with prediction.) Parameters are at best a convenience, and parameters are at their best when they are closely related to prediction (e.g., probabilities of survival). Geisser (1971, 1993) gives excellent discussions of the predictive approach.

In this book the emphasis has been placed on models rather than on parameters. Now you know why. Models can be used for prediction. They are an endproduct. Parameters are an integral part of most models, but they are a tool and not an end in themselves. Christensen (1995) gives a short discussion on the relation among models, prediction, and testing. Having now convinced a large portion of the statistical community of the unreliability of my ideas, I shall return to the issue at hand.

6.3.2 General Prediction

Suppose we have random variables $y, x_1, x_2, \ldots, x_{p-1}$. Regression can be viewed as the problem of predicting y from the values of $x_1, \ldots, x_{p-1}$. We will examine this problem and consider its relationship to the linear model theory approach to regression. Let x be the vector $x = (x_1, \ldots, x_{p-1})'$. A reasonable criterion for choosing a predictor of y is to pick a predictor $f(x)$ that minimizes the mean squared error, $\mathrm{E}[y - f(x)]^2$. (The *MSE* defined in linear model theory is a function of the observations that estimates the theoretical mean squared error defined here.) Note that, unlike standard linear model theory, the expected value is taken over the joint distribution of y and x.

The use of an expected squared error criterion presupposes the existence of first and second moments for y and $f(x)$. Let $\mathrm{E}(y) = \mu_y$, $\mathrm{Var}(y) = \sigma_y^2 \equiv \sigma_{yy}$ and let $\mathrm{E}[f(x)] = \mu_f$, $\mathrm{Var}[f(x)] = \sigma_{ff}$, and $\mathrm{Cov}[y, f(x)] = \sigma_{yf}$, with similar notations for other functions of x, e.g., $m(x)$ has $\sigma_{ym} = \mathrm{Cov}[y, m(x)]$.

The remainder of this section consists of three subsections. The next examines best predictors, i.e., those functions $f(x)$ that do the best job of predicting y. Without knowing (or being able to estimate) the joint distribution of x and y, we cannot find the best predictor, so in Subsection 4 we examine the best predictors among functions $f(x)$ that are restricted to be linear functions of x. These best linear predictors depend only on the means and covariances of the random variables and are thus relatively easy to estimate. Subsection 4 also examines the relationship between best linear predictors and least squares estimation. Both the best predictor and the best linear predictor can be viewed as orthogonal projections in an appropriate vector space, a subject that is commented on earlier but is amplified in Subsection 5.

6.3.3 Best Prediction

We now establish that the best predictor is the conditional expectation of y given x. See Appendix D for definitions and results about conditional expectations.

Theorem 6.3.1. Let $m(x) \equiv \mathrm{E}(y|x)$. Then for any other predictor $f(x)$, $\mathrm{E}[y - m(x)]^2 \leq \mathrm{E}[y - f(x)]^2$; thus $m(x)$ is the best predictor of y.

PROOF.

$$\begin{aligned}\mathrm{E}[y - f(x)]^2 &= \mathrm{E}[y - m(x) + m(x) - f(x)]^2 \\ &= \mathrm{E}[y - m(x)]^2 + \mathrm{E}[m(x) - f(x)]^2 + 2\mathrm{E}\{[y - m(x)][m(x) - f(x)]\}.\end{aligned}$$

Since both $\mathrm{E}[y - m(x)]^2$ and $\mathrm{E}[m(x) - f(x)]^2$ are nonnegative, it is enough to show that $\mathrm{E}\{[y - m(x)][m(x) - f(x)]\} = 0$. Consider this expectation conditional on x.

$$\begin{aligned}\mathrm{E}\{[y - m(x)][m(x) - f(x)]\} &= \mathrm{E}\left(\mathrm{E}\{[y - m(x)][m(x) - f(x)]|x\}\right) \\ &= \mathrm{E}\left([m(x) - f(x)]\mathrm{E}\{[y - m(x)]|x\}\right) \\ &= \mathrm{E}\left([m(x) - f(x)]0\right) = 0\end{aligned}$$

where $\mathrm{E}\{[y - m(x)]|x\} = 0$ because $\mathrm{E}(y|x) = m(x)$. □

The goal of most predictive procedures is to find, or rather estimate, the function $\mathrm{E}(y|x) = m(x)$. Suppose we have a random sample (x_i', y_i), $i = 1, \ldots, n$. In linear regression, $m(x_i) = \alpha + x_i'\beta$ with α and β unknown. A generalized linear model assumes a distribution for y given x and that $\mathrm{E}(y_i|x_i) = m(x_i) = h(\alpha + x_i'\beta)$ for known h and unknown α and β. Here h is just the inverse of the link function. The standard nonlinear regression model is a more general version of these. It uses $m(x_i) = h(x_i; \alpha, \beta)$, where h is some known function but α and β are unknown. The conditional mean structure of all three parametric models is that of the nonlinear regression model: $m(x) = h(x; \alpha, \beta)$, h known. We then become interested in estimating α and β. Evaluating whether we have the correct "known" form for h is a question of whether lack of fit exists, see Section 6. Nonparametric regression is unwilling to assume a functional form $h(x; \alpha, \beta)$. The standard nonparametric regression model is $y_i = m(x_i) + e_i$ where, conditional on the x_is, the e_is are independent with mean 0 and variance σ^2. In nonparametric regression, m is completely unknown.

All of these versions of regression involve estimating whatever parts of $m(x)$ are not assumed to be known. On the other hand, best prediction *theory* treats $m(x)$ as a known function, so for models involving α and β it treats them as known.

In Theorem 6.3.3, we present a result that does two things. First, it provides a justification for the residual plots used in Chapter 13 to identify lack of fit. Second, as discussed in Subsection 5, it establishes that $\mathrm{E}(y|x)$ can be viewed as the perpendicular projection of y into the space of random variables, say $f(x)$, that are functions

of x alone, have mean $\mathrm{E}[f(x)] = E[y]$, and a finite variance, see also deLaubenfels (2006). Before doing this, we need to establish some covariance and correlation properties.

Proposition 6.3.2. $\mathrm{Cov}[y, f(x)] = \mathrm{Cov}[m(x), f(x)]$. In particular, $\mathrm{Cov}[y, m(x)] = \mathrm{Var}[m(x)] = \sigma_{mm}$ and $\mathrm{Corr}^2[y, m(x)] = \sigma_{mm}/\sigma_{yy}$.

PROOF. Recall that, from the definition of conditional expectation, $\mathrm{E}[m(x)] = \mu_y$.

$$\begin{aligned}\mathrm{Cov}[y, f(x)] &= \mathrm{E}_{yx}[(y-\mu_y)f(x)] \\ &= \mathrm{E}_x\mathrm{E}_{y|x}[(y-m(x)+m(x)-\mu_y)f(x)] \\ &= \mathrm{E}_x[(m(x)-\mu_y)f(x)] \\ &= \mathrm{Cov}[m(x), f(x)].\end{aligned}$$

□

Now consider an arbitrary predictor $\tilde{y}(x)$.

Theorem 6.3.3. Let $\tilde{y}(x)$ be any predictor with $\mathrm{E}[\tilde{y}(x)] = \mu_y$, then $\mathrm{Cov}[f(x), y - \tilde{y}(x)] = 0$ for any function f if and only if $\mathrm{E}(y|x) = \tilde{y}(x)$ almost surely.

PROOF. $\Leftarrow$ If $\mathrm{E}(y|x) \equiv m(x) = \tilde{y}(x)$, the fact that $\mathrm{Cov}[f(x), y - m(x)] = 0$ for any function f is an immediate consequence of Proposition 2.

$\Rightarrow$ Now suppose that $\mathrm{E}[\tilde{y}(x)] = \mu_y$ and $\mathrm{Cov}[f(x), y - \tilde{y}(x)] = 0$ for any function f. To show that $\tilde{y}(x) = m(x)$ a.s., it is enough to note that $\mathrm{E}[\tilde{y}(x) - m(x)] = \mu_y - \mu_y = 0$ and to show that $\mathrm{Var}[\tilde{y}(x) - m(x)] = 0$. Thinking of $f(x) = \tilde{y}(x) - m(x)$ in the covariance conditions, observe that

$$\begin{aligned}\mathrm{Var}[\tilde{y}(x) - m(x)] &= \mathrm{Cov}[\tilde{y}(x) - m(x), \tilde{y}(x) - m(x)] \\ &= \mathrm{Cov}\{[y - m(x)] - [y - \tilde{y}(x)], [\tilde{y}(x) - m(x)]\} \\ &= \mathrm{Cov}\{[y - m(x)], [\tilde{y}(x) - m(x)]\} - \mathrm{Cov}\{[y - \tilde{y}(x)], [\tilde{y}(x) - m(x)]\} \\ &= 0 - 0.\end{aligned}$$

□

In fitting linear models, or any other regression procedure, we typically obtain fitted values $\hat{y}_i$ corresponding to the observed data y_i, from which we can obtain residuals $y_i - \hat{y}_i$. According to Theorem 6.3.3, if the fitted values are coming from the best predictor, plotting the residuals against any function of the predictor vector x_i should display zero correlation. Thus, if we plot the residuals against some function $f(x_i)$ of the predictors and observe a correlation, we obviously do not have the best predictor, so we should try fitting some other model. In particular, regardless of how nonlinear the original regression model might have been, adding a linear term to the model using $f(x_i)$ as the predictor should improve the fit of the model. Unfortunately, there is no reason to think that adding such a linear term will get you to the best predictor. Finally, it should be noted that a common form of lack

of fit detected in residual plots is a parabolic shape, which does not necessarily suggest a nonzero correlation with the predictor used in the plot. However, when the residual plot is a parabola, a plot of the residuals versus the (suitably centered) squared predictor will display a nonzero correlation.

Finally, if we plot the residuals against some predictor variable z that is not part of x, the fact that we would make such a plot suggests that we really want the best predictor $m(x,z)$ rather than $m(x)$, although if we originally left z out, we probably suspect that $m(x) = m(x,z)$. A linear relationship between the residuals and z indicates that the estimated regression function $\hat{m}(x)$ is not an adequate estimate of the best predictor $m(x,z)$.

6.3.4 Best Linear Prediction

The ideas of best linear prediction and best linear unbiased prediction (see Section 12.2) are very important. As will be seen here and in Chapter 12, best linear prediction theory has important applications in standard linear models, mixed models, and the analysis of spatial data. The theory has traditionally been taught as part of multivariate analysis (cf. Anderson, 1984). It is important for general stochastic processes (cf. Doob, 1953), time series (cf. Shumway and Stoffer, 2000; Brockwell and Davis, 1991; or Christensen, 2001, Chapters 4 and 5), principal component analysis (cf. Christensen, 2001, Chapter 3), and it is the basis for linear Bayesian methods (cf. Hartigan, 1969). For applications to spatial data, see also Christensen (2001, Chapter 6), Cressie (1993), and Ripley (1981).

In order to use the results on best prediction, one needs to know $\mathrm{E}(y|x)$, which generally requires knowledge of the joint distribution of $(y,x_1,x_2,\ldots,x_{p-1})'$. If the joint distribution is not available but the means, variances, and covariances are known, we can find the best linear predictor of y. We seek a linear predictor $\alpha + x'\beta$ that minimizes $\mathrm{E}[y-\alpha-x'\beta]^2$ for all scalars α and $(p-1)\times 1$ vectors β.

In addition to our earlier assumptions that first and second moments for y exist, we now also assume the existence of $\mathrm{E}(x) = \mu_x$, $\mathrm{Cov}(x) = V_{xx}$, and $\mathrm{Cov}(x,y) = V_{xy} = V_{yx}' = \mathrm{Cov}(y,x)'$.

Let β_* be a solution to $V_{xx}\beta = V_{xy}$, then we will show that the function

$$\hat{E}(y|x) \equiv \mu_y + (x-\mu_x)'\beta_*$$

is a best linear predictor of y based on x. $\hat{E}(y|x)$ is also called *the linear expectation of y given x*. (Actually, it is the linear expectation of y given x and a random variable that is constant with probability 1.) Note that when V_{xx} is singular, there are an infinite number of vectors β_* that solve to $V_{xx}\beta = V_{xy}$, but by using Lemma 1.3.5 we can show that all such solutions give the same best linear predictor with probability 1. The idea is that since $(x-\mu_x) \in C(V_{xx})$ with probability 1, for some random b, $(x-\mu_x) = V_{xx}b$ with probability 1, so

$$(x-\mu_x)'\beta_* = b'V_{xx}\beta_* = b'V_{xy},$$

which does not depend on the choice of β_* with probability 1.

Theorem 6.3.4. $\hat{E}(y|x)$ is a best linear predictor of y.

PROOF. Define η so that $\alpha = \eta - \mu_x'\beta$. An arbitrary linear predictor is $f(x) = \alpha + x'\beta = \eta + (x-\mu_x)'\beta$.

$$\begin{aligned}\mathrm{E}[y-f(x)]^2 &= \mathrm{E}[y-\hat{E}(y|x)+\hat{E}(y|x)-f(x)]^2\\ &= \mathrm{E}[y-\hat{E}(y|x)]^2+\mathrm{E}[\hat{E}(y|x)-f(x)]^2\\ &\quad +2\mathrm{E}\{[y-\hat{E}(y|x)][\hat{E}(y|x)-f(x)]\}.\end{aligned}$$

If we show that $\mathrm{E}\{[y-\hat{E}(y|x)][\hat{E}(y|x)-f(x)]\}=0$, the result follows almost immediately. In that case,

$$\mathrm{E}[y-f(x)]^2 = \mathrm{E}[y-\hat{E}(y|x)]^2+\mathrm{E}[\hat{E}(y|x)-f(x)]^2.$$

To find $f(x)$ that minimizes the left-hand side, observe that both terms on the right are nonnegative, the first term does not depend on $f(x)$, and the second term is minimized by taking $f(x)=\hat{E}(y|x)$.

We now show that $\mathrm{E}\{[y-\hat{E}(y|x)][\hat{E}(y|x)-f(x)]\}=0$.

$$\begin{aligned}&\mathrm{E}\{[y-\hat{E}(y|x)][\hat{E}(y|x)-f(x)]\}\\ &= \mathrm{E}(\{y-[\mu_y+(x-\mu_x)'\beta_*]\}\{[\mu_y+(x-\mu_x)'\beta_*]-[\eta+(x-\mu_x)'\beta]\})\\ &= \mathrm{E}(\{(y-\mu_y)-(x-\mu_x)'\beta_*\}\{(\mu_y-\eta)+(x-\mu_x)'(\beta_*-\beta)\})\\ &= \mathrm{E}\big[(y-\mu_y)(\mu_y-\eta)-(x-\mu_x)'\beta_*(\mu_y-\eta)\\ &\quad +(y-\mu_y)(x-\mu_x)'(\beta_*-\beta)-(x-\mu_x)'\beta_*(x-\mu_x)'(\beta_*-\beta)\big]\\ &= (\mu_y-\eta)\mathrm{E}[(y-\mu_y)]-\mathrm{E}[(x-\mu_x)']\beta_*(\mu_y-\eta)\\ &\quad +\mathrm{E}[(y-\mu_y)(x-\mu_x)'](\beta_*-\beta)-\mathrm{E}[\beta_*'(x-\mu_x)(x-\mu_x)'(\beta_*-\beta)]\\ &= 0-0+V_{yx}(\beta_*-\beta)-\beta_*'\mathrm{E}[(x-\mu_x)(x-\mu_x)'](\beta_*-\beta)\\ &= V_{yx}(\beta_*-\beta)-\beta_*'V_{xx}(\beta_*-\beta).\end{aligned}$$

However, by definition, $\beta_*'V_{xx}=V_{yx}$; so

$$V_{yx}(\beta_*-\beta)-\beta_*'V_{xx}(\beta_*-\beta) = V_{yx}(\beta_*-\beta)-V_{yx}(\beta_*-\beta)=0. \qquad \square$$

It is of interest to note that if $(y,x')'$ has a multivariate normal distribution, then the best linear predictor is the best predictor. Morrison (1976) contains a discussion of conditional expectations for multivariate normals.

The following proposition will be used in Section 12.2 to develop the theory of best linear unbiased prediction.

Proposition 6.3.5. $\mathrm{E}[y-\alpha-x'\beta]^2 = \mathrm{E}[y-\hat{E}(y|x)]^2 + \mathrm{E}[\hat{E}(y|x)-\alpha-x'\beta]^2$.

PROOF. The result is part of the proof of Theorem 6.3.4. □

We show that best linear predictors are essentially unique. In other words, we show that to minimize $\mathrm{E}[\hat{E}(y|x)-\eta-(x-\mu_x)'\beta]^2$, you need $\mu_y = \eta$ and β must be a solution to $V_{xx}\beta = V_{xy}$. It is not difficult to show that

$$\mathrm{E}[\hat{E}(y|x)-\eta-(x-\mu_x)'\beta]^2 = (\mu_y-\eta)^2 + \mathrm{E}[(x-\mu_x)'\beta_* - (x-\mu_x)'\beta]^2.$$

Clearly, minimization requires $\mu_y = \eta$ and $\mathrm{E}[(x-\mu_x)'\beta_* - (x-\mu_x)'\beta]^2 = 0$. The best linear predictors will be essentially unique if we can show that $\mathrm{E}[(x-\mu_x)'\beta_* - (x-\mu_x)'\beta]^2 = 0$ implies that β must be a solution to $V_{xx}\beta = V_{xy}$. Observe that

$$\begin{aligned}\mathrm{E}[(x-\mu_x)'\beta_* - (x-\mu_x)'\beta]^2 &= \mathrm{E}[(x-\mu_x)'(\beta_*-\beta)]^2 \\ &= \mathrm{Cov}[(x-\mu_x)'(\beta_*-\beta)] \\ &= (\beta_*-\beta)'V_{xx}(\beta_*-\beta).\end{aligned}$$

Write $V_{xx} = QQ'$ with $C(V_{xx}) = C(Q)$. Then $(\beta_*-\beta)'V_{xx}(\beta_*-\beta) = 0$ if and only if $(\beta_*-\beta)'QQ'(\beta_*-\beta) = 0$ if and only if $Q'(\beta_*-\beta) = 0$ if and only if $(\beta_*-\beta) \perp C(Q) = C(V_{xx})$ if and only if $V_{xx}(\beta_*-\beta) = 0$ if and only if $V_{xx}\beta = V_{xx}\beta_* = V_{xy}$. So β must be a solution.

The variance of the prediction error $y-\hat{E}(y|x)$ is given in Section 5. (Actually, the covariance matrix for a bivariate prediction is given.)

Next, we examine the correspondence between this theory and linear model regression theory. Suppose we have n observations on the vector $(y,x')' = (y,x_1,x_2,\ldots,x_{p-1})'$. We can write these as $(y_i,x_i')' = (y_i,x_{i1},x_{i2},\ldots,x_{i,p-1})'$, $i = 1,\ldots,n$. In matrix notation write $Y = (y_1,y_2,\ldots,y_n)'$ and $Z = [x_{ij}]$, $i = 1,\ldots,n$, $j = 1,\ldots,p-1$. (Z plays the same role as Z did in Section 2 on multiple regression.) The usual estimates for V_{xx} and V_{xy} can be written as

$$\begin{aligned}S_{xx} &= \frac{1}{n-1}Z'\left(I-\frac{1}{n}J_n^n\right)Z = \frac{1}{n-1}\sum_{i=1}^n (x_i-\bar{x}_\cdot)(x_i-\bar{x}_\cdot)' \\ S_{xy} &= \frac{1}{n-1}Z'\left(I-\frac{1}{n}J_n^n\right)Y = \frac{1}{n-1}\sum_{i=1}^n (x_i-\bar{x}_\cdot)(y_i-\bar{y}_\cdot).\end{aligned}$$

The usual estimates of μ_y and μ_x are

$$\bar{y}_\cdot = \frac{1}{n}J_1^nY = \frac{1}{n}\sum_{i=1}^n y_i$$

$$\bar{x}_\cdot' = \frac{1}{n} J_1^n Z = \left(\frac{1}{n}\sum_{i=1}^n x_{i1}, \ldots, \frac{1}{n}\sum_{i=1}^n x_{ip-1} \right)'.$$

The natural predictor of y is

$$\hat{y} = \bar{y} + (x - \bar{x})'\hat{\beta}_*,$$

where $\hat{\beta}_*$ is a solution to $S_{xx}\beta_* = S_{xy}$, i.e., it solves $Z'\left(I - \frac{1}{n}J_n^n\right)Z\beta_* = Z'\left(I - \frac{1}{n}J_n^n\right)Y$. From the results of Section 2, $\bar{y} = \hat{\gamma}_0$ and any solution of $Z'\left(I - \frac{1}{n}J_n^n\right)Z\beta_* = Z'\left(I - \frac{1}{n}J_n^n\right)Y$ is a least squares estimate of $\beta_* = \gamma_*$. Thus, the natural estimates of the parameters in the best linear predictor are the least squares estimates from the mean corrected regression model considered in the previous section.

Finally, we include a result for best linear predictors that is analogous to Theorem 6.3.3 for best predictors. First, if we have residuals from the best linear predictor, they will be uncorrelated with any linear combination of the predictor variables. Second, $\hat{E}(y|x)$ can be viewed as the perpendicular projection of y into the space of random variables $f(x)$ that are linear functions of x and have $\mathrm{E}[f(x)] = E[y]$, cf. Subsection 5.

Theorem 6.3.6. Suppose $\tilde{y}(x)$ is any linear predictor with $\mathrm{E}[\tilde{y}(x)] = \mu_y$, then $\mathrm{Cov}[f(x), y - \tilde{y}(x)] = 0$ for any linear function f if and only if $\hat{E}(y|x) = \tilde{y}(x)$ almost surely.

PROOF. Let $f(x) = \eta + (x - \mu_x)'\beta$ and let $\tilde{y} = \mu_y + (x - \mu_x)'\delta$.
$\Leftarrow$ For $\hat{E}(y|x) = \tilde{y}(x)$,

$$\begin{aligned}\mathrm{Cov}[\eta + (x-\mu_x)'\beta, (y-\mu_y) - (x-\mu_x)'\beta_*] &= \beta' V_{xy} - \beta' V_{xx}\beta_* \\ &= \beta' V_{xy} - \beta' V_{xy} = 0.\end{aligned}$$

$\Rightarrow$ If

$$\begin{aligned}0 &= \mathrm{Cov}[\eta + (x-\mu_x)'\beta, (y-\mu_y) - (x-\mu_x)'\delta] \\ &= \beta' V_{xy} - \beta' V_{xx}\delta\end{aligned}$$

for any β, then $V_{xy} = V_{xx}\delta$, so $\tilde{y}(x) = \hat{E}(y|x)$ with probability 1. □

As mentioned earlier, the theory of best linear predictors also comes up in developing the theory of best linear unbiased predictors (BLUPs), which is an important subject in models that have random effects. In random effects models, part of the β vector in $Y = X\beta + e$ is assumed to be random and unobservable. Random effects models and BLUPs are discussed in Chapter 12.

The result of the next exercise will be used in Section 5 on partial correlations.

Exercise 6.3 For predicting $y = (y_1, \ldots, y_q)'$ from $x = (x_1, \ldots, x_{p-1})'$ we say that

a predictor $f(x)$ is best if the scalar $\mathrm{E}\{[y-f(x)]'[y-f(x)]\}$ is minimized. Show that with simple modifications, Theorems 6.3.1 and 6.3.4 hold for the extended problem, as does Proposition 6.3.5.

6.3.5 Inner Products and Orthogonal Projections in General Spaces

In most of this book, we define orthogonality and length using the Euclidean inner product in $\mathbf{R}^n$. For two vectors x and y, the Euclidean inner product is $x'y$, so $x \perp y$ if $x'y = 0$ and the length of x is $\|x\| = \sqrt{x'x}$. In Section B.3 we discussed, in detail, perpendicular projection operators relative to this inner product. We established that the projection of a vector Y into $C(X)$ is $\hat{Y} \equiv MY$. It also follows from Theorems 2.2.1 and 2.8.1 that $\hat{Y}$ is the unique vector in $C(X)$ that satisfies $(Y-\hat{Y}) \perp C(X)$. This last property is sometimes used to define what it means for $\hat{Y}$ to be the perpendicular projection of Y into $C(X)$. We use this concept to extend the application of perpendicular projections to more general vector spaces

More generally in $\mathbf{R}^n$, we can use any positive definite matrix B to define an inner product between x and y as $x'By$. As before, x and y are orthogonal if their inner product $x'By$ is zero and the length of x is the square root of its inner product with itself, now $\|x\|_B \equiv \sqrt{x'Bx}$. As argued in Section B.3, any idempotent matrix is always a projection operator, but which one is the perpendicular projection operator depends on the inner product. As can be seen from Proposition 2.7.2 and Exercise 2.5, the matrix $A \equiv X(X'BX)^{-}X'B$ is an oblique projection operator onto $C(X)$ for the Euclidean inner product, but it is the perpendicular projection operator onto $C(X)$ with the inner product defined using the matrix B. It is not too difficult to see that AY is the unique vector in $C(X)$ that satisfies $(Y-AY) \perp_B C(X)$, i.e., $(Y-AY)'BX = 0$.

These ideas can be applied in very general spaces. In particular, they can be applied to the concepts of prediction introduced in this section. For example, we can define the inner product between two random variables y and x with mean 0 and finite variance as the $\mathrm{Cov}(x,y)$. In this case, $\mathrm{Var}(x)$ plays the role of the squared length of the random variable and the standard deviation is the length. Two random variables are orthogonal if they are uncorrelated.

Now consider a vector of random variables $x = (x_1, \ldots, x_{p-1})'$ and the space of all functions $f(x)$ into $\mathbf{R}^1$ that have mean 0 and variances. We showed in Theorem 6.3.3 that $m(x) \equiv \mathrm{E}(y|x)$ is the unique function of x having mean μ_y for which $\mathrm{Cov}[y - m(x), f(x)] = 0$ for any $f(x)$. Thus, as alluded to above, $m(x) - \mu_y$ satisfies a property often used to *define* the perpendicular projection of $y - \mu_y$ into the space of functions of x that have mean 0 and variances. Alternatively, we can think of $m(x)$ as the perpendicular projection of y into the space of functions of x that have mean μ_y and variances.

We can also consider a reduced space of random variables, the linear functions of x, i.e., $f(x) = \alpha + x'\beta$. In Theorem 6.3.6 we show that $\mathrm{Cov}[y - \hat{E}(y|x), \alpha + x'\beta] = 0$

for any linear function of x, so once again, the best linear predictor is the perpendicular projection of y into the linear functions of x with mean μ_y.

We now generalize the definitions of an inner product space, orthogonality, and orthogonal projection.

Definition A.11 (Alternate). A vector space $\mathscr{X}$ is an inner product space if for any $x, y \in \mathscr{X}$, there exists a symmetric bilinear function $[x,y]$ into $\mathbf{R}$ with $[x,x] > 0$ for any $x \neq 0$. A bilinear function has the properties that for any scalars a_1, a_2 and any vectors x_1, x_2, y, $[a_1x_1 + a_2x_2, y] = a_1[x_1,y] + a_2[x_2,y]$ and $[y, a_1x_1 + a_2x_2] = a_1[y,x_1] + a_2[y,x_2]$. A symmetric function has $[x_1,x_2] = [x_2,x_1]$. The vectors x and y are orthogonal if $[x,y] = 0$ and the squared length of x is $[x,x]$. The perpendicular projection of y into a subspace $\mathscr{X}_0$ of $\mathscr{X}$ is defined as the unique vector $y_0 \in \mathscr{X}_0$ with the property that $[y - y_0, x] = 0$ for any $x \in \mathscr{X}_0$.

Note that the set of mean zero, finite variance, real-valued functions of x and y form a vector space under Definition A.1 and an inner product space using the covariance of any two such functions as the inner product. Both the set of mean 0, finite variance functions of x and the set of mean 0 linear functions of x are subspaces, so y can be projected into either subspace.

We now relate Definition A.11 (Alternate) to the concept of a perpendicular projection operator.

Exercise 6.4 Consider an inner product space $\mathscr{X}$ and a subspace $\mathscr{X}_0$. Suppose that any vector $y \in \mathscr{X}$ can be written uniquely as $y = y_0 + y_1$ with $y_0 \in \mathscr{X}_0$ and $y_1 \perp \mathscr{X}_0$. Let $M(x)$ be a linear operator on $\mathscr{X}$ in the sense that for any $x \in \mathscr{X}$, $M(x) \in \mathscr{X}$ and for any scalars a_1, a_2 and any vectors x_1, x_2, $M(a_1x_1 + a_2x_2) = a_1M(x_1) + a_2M(x_2)$. $M(x)$ is defined to be a perpendicular projection operator onto $\mathscr{X}_0$ if for any $x_0 \in \mathscr{X}_0$, $M(x_0) = x_0$, and for any $x_1 \perp \mathscr{X}_0$, $M(x_1) = 0$. Using Definition A.11 (Alternate), show that for any vector y, $M(y)$ is the perpendicular projection of y into $\mathscr{X}_0$.

6.4 Multiple Correlation

The coefficient of determination, denoted R^2, is a commonly used measure of the predictive ability of a model. Computationally, it is most often defined as

$$R^2 = \frac{SSReg}{SSTot - C},$$

so it is the proportion of the total variability explained by the independent variables. The greater the proportion of the total variability in the data that is explained by the model, the better the ability to predict with that model. The use and possible abuse of R^2 as a tool for comparing models is discussed in Chapter 14, however it

should be noted here that since R^2 is a measure of the predictive ability of a model, R^2 does not give direct information about whether a model fits the data properly. Demonstrably bad models can have very high R^2s and perfect models can have low R^2s.

We now define the multiple correlation, characterize it in terms of the best linear predictor, and show that R^2 is the natural estimate of it. Subsection 6.4.1 generalizes the idea of R^2 from best linear prediction to best prediction.

Recall that the correlation between two random variables, say x_1 and x_2, is

$$\mathrm{Corr}(x_1,x_2) = \mathrm{Cov}(x_1,x_2) \Big/ \sqrt{\mathrm{Var}(x_1)\mathrm{Var}(x_2)}.$$

The multiple correlation of y and $(x_1,x_2,\ldots,x_{p-1})' = x$ is the maximum of $\mathrm{Corr}(y, \alpha + x'\beta)$ over all α and β. Note that

$$\mathrm{Cov}(y, \alpha + x'\beta) = V_{yx}\beta = \beta_*' V_{xx}\beta,$$

where β_* is defined as in Subsection 6.3.4 and

$$\mathrm{Var}(\alpha + x'\beta) = \beta' V_{xx}\beta.$$

In particular, $\mathrm{Cov}(y, \alpha + x'\beta_*) = \beta_*' V_{xx}\beta_* = \mathrm{Var}(\alpha + x'\beta_*)$. The Cauchy–Schwarz inequality says that

$$\left(\sum_{i=1}^{t} r_i s_i\right)^2 \le \sum_{i=1}^{t} r_i^2 \sum_{i=1}^{t} s_i^2.$$

Since $V_{xx} = RR'$ for some matrix R, the Cauchy–Schwarz inequality gives

$$\left(\beta_*' V_{xx}\beta\right)^2 \le \left(\beta' V_{xx}\beta\right)\left(\beta_*' V_{xx}\beta_*\right).$$

Considering the squared correlation gives

$$\begin{aligned}
\mathrm{Corr}^2(y, \alpha + x'\beta) &= \left(\beta_*' V_{xx}\beta\right)^2 / \left(\beta' V_{xx}\beta\right)\sigma_y^2 \\
&\le \beta_*' V_{xx}\beta_* / \sigma_y^2 \\
&= \left(\beta_*' V_{xx}\beta_*\right)^2 / \left(\beta_*' V_{xx}\beta_*\right)\sigma_y^2 \\
&= \mathrm{Corr}^2(y, \alpha + x'\beta_*)
\end{aligned}$$

and the squared multiple correlation between y and x equals

$$\mathrm{Corr}^2(y, \alpha + x'\beta_*) = \beta_*' V_{xx}\beta_* / \sigma_y^2.$$

If we have observations $(y_i, x_{i1}, x_{i2}, \ldots, x_{i,p-1})'$, $i = 1, \ldots, n$, the usual estimate of σ_y^2 can be written as

$$s_y^2 = \frac{1}{n-1} Y' \left(I - \frac{1}{n} J_n^n\right) Y = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Using equivalences derived earlier, the natural estimate of the squared multiple correlation coefficient between y and x is

$$\frac{\hat{\beta}_*' S_{xx} \hat{\beta}_*}{s_y^2} = \frac{\hat{\beta}_*' Z' \left(I - \frac{1}{n} J_n^n\right) Z \hat{\beta}_*}{Y' \left(I - \frac{1}{n} J_n^n\right) Y} = \frac{SSReg}{SSTot - C}.$$

It is worth noticing that $SSTot - C = SSReg + SSE$ and that

$$\begin{aligned} SSReg/SSE &= SSReg/[(SSE + SSReg) - SSReg] \\ &= SSReg/[SSTot - C - SSReg] \\ &= \frac{SSReg}{[SSTot - C][1 - R^2]} \\ &= R^2/[1 - R^2]. \end{aligned}$$

For normally distributed data, the α level F test for $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ is to reject H_0 if

$$\frac{n-p}{p-1} \frac{R^2}{1-R^2} > F(1-\alpha, p-1, n-p).$$

EXAMPLE 6.4.1. From the information given in Example 6.2.1, the coefficient of determination can be found. From the ANOVA table in Example 6.2.1, we get

$$SSReg = 1259.32$$

$$SSTot - C = SSTotal - SS(\beta_0) = 16576.60 - 15270.78 = 1305.82$$

and

$$R^2 = 1259.32/1305.82 = 0.964.$$

This is a very high value for R^2 and indicates that the model has very substantial predictive ability. However, before you conclude that the model actually fits the data well, see Example 6.6.3.

Exercise 6.5 Show that for a linear model with an intercept, R^2 is simply the square of the correlation between the data y_i and the predicted values $\hat{y}_i$.

This discussion has focused on R^2, which estimates the squared multiple correlation coefficient, a measure of the predictive ability of the best linear predictor. We now consider an analogous measure for the best predictor.

6.4.1 Squared Predictive Correlation

As in Subsection 6.3.2, consider an arbitrary predictor $\tilde{y}(x)$. This is a function of x alone and not a function of y.

The *squared predictive correlation* of $\tilde{y}(x)$ is $\text{Corr}^2[y, \tilde{y}(x)]$. The highest squared predictive correlation is obtained by using the best predictor. Note that in the special case where the best predictor $m(x)$ is also the best linear predictor, the highest squared predictive correlation equals the squared multiple correlation coefficient.

Theorem 6.4.1. $\text{Corr}^2[y, \tilde{y}(x)] \leq \text{Corr}^2[y, m(x)]$.

PROOF. By Cauchy-Schwarz, $(\sigma_{m\tilde{y}})^2 \leq \sigma_{mm}\sigma_{\tilde{y}\tilde{y}}$, so $(\sigma_{m\tilde{y}})^2/\sigma_{\tilde{y}\tilde{y}} \leq \sigma_{mm}$. Using Proposition 6.3.2,

$$\text{Corr}^2[y, \tilde{y}(x)] = \frac{(\sigma_{y\tilde{y}})^2}{\sigma_{yy}\sigma_{\tilde{y}\tilde{y}}} = \frac{(\sigma_{m\tilde{y}})^2}{\sigma_{yy}\sigma_{\tilde{y}\tilde{y}}} \leq \frac{\sigma_{mm}}{\sigma_{yy}} = \text{Corr}^2[y, m(x)].$$

The result follows from the last part of Proposition 6.3.2. □

Theorem 6.4.1 is also established in Rao (1973, Section 4g.1). From Theorem 6.4.1, the best regression function $m(x)$ has the highest squared predictive correlation. When we have perfect prediction, the highest squared predictive correlation is 1. In other words, if the conditional variance of y given x is 0, then $y = m(x)$ a.s., and the highest squared predictive correlation is the correlation of $m(x)$ with itself, which is 1. On the other hand, if there is no regression relationship, i.e., if $m(x) = \mu_y$ a.s., then $\sigma_{mm} = 0$, and the highest squared predictive correlation is 0.

We would now like to show that as the squared predictive correlation increases, we get increasingly better prediction. First we need to deal with the fact that high squared predictive correlations can be achieved by bad predictors. Just because $\tilde{y}(x)$ is highly correlated with y does not mean that $\tilde{y}(x)$ is actually close to y. Recall that $\tilde{y}(x)$ is simply a random *variable* that is being used to predict y. As such, $\tilde{y}(x)$ is a linear predictor of y, that is, $\tilde{y}(x) = 0 + 1\tilde{y}(x)$. We can apply Theorem 6.3.4 to this random variable to obtain a linear predictor that is at least as good as $\tilde{y}(x)$, namely

$$\hat{y}(x) = \mu_y + \frac{\sigma_{y\tilde{y}}}{\sigma_{\tilde{y}\tilde{y}}}[\tilde{y}(x) - \mu_{\tilde{y}}].$$

We refer to such predictors as *linearized predictors*. Note that $\text{E}[\hat{y}(x)] \equiv \mu_{\hat{y}} = \mu_y$,

$$\sigma_{\hat{y}\hat{y}} \equiv \text{Var}[\hat{y}(x)] = \left(\frac{\sigma_{y\tilde{y}}}{\sigma_{\tilde{y}\tilde{y}}}\right)^2 \sigma_{\tilde{y}\tilde{y}} = \frac{(\sigma_{y\tilde{y}})^2}{\sigma_{\tilde{y}\tilde{y}}},$$

and

$$\sigma_{y\hat{y}} \equiv \text{Cov}[y, \hat{y}(x)] = \frac{\sigma_{y\tilde{y}}}{\sigma_{\tilde{y}\tilde{y}}}\sigma_{y\tilde{y}} = \frac{(\sigma_{y\tilde{y}})^2}{\sigma_{\tilde{y}\tilde{y}}}.$$

In particular, $\sigma_{\hat{y}\hat{y}} = \sigma_{y\hat{y}}$, so the squared predictive correlation of $\hat{y}(x)$ is

$$\text{Corr}^2[y, \hat{y}(x)] = \frac{\sigma_{\hat{y}\hat{y}}}{\sigma_{yy}}.$$

In addition, the direct measure of the goodness of prediction for $\hat{y}(x)$ is

$$\text{E}[y - \hat{y}(x)]^2 = \sigma_{yy} - 2\sigma_{y\hat{y}} + \sigma_{\hat{y}\hat{y}} = \sigma_{yy} - \sigma_{\hat{y}\hat{y}}.$$

This leads directly to the next result.

Theorem 6.4.2. For two linearized predictors $\hat{y}_1(x)$ and $\hat{y}_2(x)$, the squared predictive correlation of $\hat{y}_2(x)$ is higher if and only if $\hat{y}_2(x)$ is a better predictor.

PROOF. $\sigma_{\hat{y}_1\hat{y}_1}/\sigma_{yy} < \sigma_{\hat{y}_2\hat{y}_2}/\sigma_{yy}$ if and only if $\sigma_{\hat{y}_1\hat{y}_1} < \sigma_{\hat{y}_2\hat{y}_2}$ if and only if $\sigma_{yy} - \sigma_{\hat{y}_2\hat{y}_2} < \sigma_{yy} - \sigma_{\hat{y}_1\hat{y}_1}$. □

It should be noted that linearizing $m(x)$ simply returns $m(x)$.

For any predictor $\tilde{y}(x)$, no matter how one arrives at it, to estimate the squared predictive correlation from data y_i, simply compute the squared sample correlation between y_i and its predictor of $\tilde{y}(x_i)$.

6.5 Partial Correlation Coefficients

Many regression programs have options available to the user that depend on the values of the sample partial correlation coefficients. The partial correlation is defined in terms of two random variables of interest, say y_1 and y_2, and several auxiliary variables, say $x_1, \ldots, x_{p-1}$. The partial correlation coefficient of y_1 and y_2 given $x_1, \ldots, x_{p-1}$, written $\rho_{y \cdot x}$, is a measure of the linear relationship between y_1 and y_2 after taking the effects of $x_1, \ldots, x_{p-1}$ out of both variables.

Writing $y = (y_1, y_2)'$ and $x = (x_1, \ldots, x_{p-1})'$, Exercise 6.3 indicates that the best linear predictor of y given x is

$$\hat{E}(y|x) = \mu_y + \beta_*'(x - \mu_x),$$

where β_* is a solution of $V_{xx}\beta_* = V_{xy}$ and V_{xy} is now a $(p-1) \times 2$ matrix. We take the effects of x out of y by looking at the prediction error

$$y - \hat{E}(y|x) = (y - \mu_y) - \beta_*'(x - \mu_x),$$

which is a 2×1 random vector. The partial correlation is simply the correlation between the two components of this vector and is readily obtained from the covariance matrix. We now find the prediction error covariance matrix. Let $\text{Cov}(y) \equiv V_{yy}$.

$$\begin{aligned}\mathrm{Cov}[(y-\mu_y)-\beta_*'(x-\mu_x)] &= \mathrm{Cov}(y-\mu_y)+\beta_*'\mathrm{Cov}(x-\mu_x)\beta_* \\ &\quad -\mathrm{Cov}(y-\mu_y,x-\mu_x)\beta_* \\ &\quad -\beta_*'\mathrm{Cov}(x-\mu_x,y-\mu_y) \\ &= V_{yy}+\beta_*'V_{xx}\beta_*-V_{yx}\beta_*-\beta_*'V_{xy} \\ &= V_{yy}+\beta_*'V_{xx}\beta_*-\beta_*'V_{xx}\beta_*-\beta_*'V_{xx}\beta_* \\ &= V_{yy}-\beta_*'V_{xx}\beta_*.\end{aligned}$$

Since $V_{xx}\beta_* = V_{xy}$ and, for any generalized inverse, $V_{xx}V_{xx}^-V_{xx} = V_{xx}$,

$$\begin{aligned}\mathrm{Cov}[(y-\mu_y)-\beta_*'(x-\mu_x)] &= V_{yy}-\beta_*'V_{xx}V_{xx}^-V_{xx}\beta_* \\ &= V_{yy}-V_{yx}V_{xx}^-V_{xy}.\end{aligned}$$

If we have a sample of the ys and xs, say $y_{i1}, y_{i2}, x_{i1}, x_{i2}, \ldots, x_{i,p-1}$, $i = 1,\ldots,n$, we can estimate the covariance matrix in the usual way. The relationship with the linear regression model of Section 2 is as follows: Let

$$Y = \begin{bmatrix} y_{11} & y_{12} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{bmatrix} = [Y_1, Y_2]$$

and

$$Z = \begin{bmatrix} x_{11} & \cdots & x_{1p-1} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np-1} \end{bmatrix}.$$

The usual estimate of $V_{yy} - V_{yx}V_{xx}^-V_{xy}$ is $(n-1)^{-1}$ times

$$Y'\left(I-\frac{1}{n}J_n^n\right)Y - Y'\left(I-\frac{1}{n}J_n^n\right)Z\left[Z'\left(I-\frac{1}{n}J_n^n\right)Z\right]^-Z'\left(I-\frac{1}{n}J_n^n\right)Y.$$

From Section 2, we know that this is the same as

$$Y'\left(I-\frac{1}{n}J_n^n\right)Y - Y'\left(M-\frac{1}{n}J_n^n\right)Y = Y'(I-M)Y,$$

where M is the perpendicular projection operator onto $C([J,Z])$. Remembering that $Y'(I-M)Y = [(I-M)Y]'[(I-M)Y]$, we can see that the estimate of $\rho_{y\cdot x}$, written $r_{y\cdot x}$ and called the *sample partial correlation coefficient*, is just the sample correlation coefficient between the residuals of fitting $Y_1 = [J,Z]\beta + e$ and the residuals of fitting $Y_2 = [J,Z]\beta + e$, i.e.,

$$r_{y\cdot x} = \frac{Y_1'(I-M)Y_2}{\sqrt{Y_1'(I-M)Y_1 Y_2'(I-M)Y_2}}.$$

The square of the sample partial correlation coefficient, often called the coefficient of partial determination, has a nice relationship to another linear model. Consider fitting $Y_1 = [J,Z,Y_2]\gamma + e$. Because $C[(I-M)Y_2]$ is the orthogonal complement of $C([J,Z])$ with respect to $C([J,Z,Y_2])$, the sum of squares for testing whether Y_2 adds to the model is

$$SSR(Y_2|J,Z) = Y_1'(I-M)Y_2[Y_2'(I-M)Y_2]^{-1}Y_2'(I-M)Y_1.$$

Since $Y_2'(I-M)Y_2$ is a scalar, it is easily seen that

$$r^2_{y\cdot x} = \frac{SSR(Y_2|J,Z)}{SSE(J,Z)},$$

where $SSE(J,Z) = Y_1'(I-M)Y_1$, the sum of squares for error when fitting the model $Y_1 = [J,Z]\beta + e$.

Finally, for normally distributed data we can do a t test of the null hypothesis $H_0 : \rho_{y\cdot x} = 0$. If H_0 is true, then

$$\sqrt{n-p-1}\, r_{y\cdot x} \Big/ \sqrt{1-r^2_{y\cdot x}} \sim t(n-p-1).$$

See Exercise 6.6 for a proof of this result.

EXAMPLE 6.5.1. Using the data of Example 6.2.1, the coefficient of partial determination (squared sample partial correlation coefficient) between y and x_2 given x_1 can be found:

$$SSR(X_2|J,X_1) = 6.70,$$

$$SSE(J,X_1) = SSE(J,X_1,X_2) + SSR(X_2|J,X_1) = 46.50 + 6.70 = 53.20,$$

$$r^2_{y\cdot x} = 6.70/53.20 = 0.1259.$$

The absolute value of the t statistic for testing whether $\rho_{y2\cdot 1} = 0$ can also be found. In this application we are correcting Y and X_2 for only one variable X_1, so $p-1 = 1$ and $p = 2$. The formula for the absolute t statistic becomes

$$\sqrt{25-2-1}\sqrt{0.1259}\Big/\sqrt{1-0.1259} = 1.78.$$

Note that this is precisely the t statistic reported for β_2 in Example 6.2.1.

Exercise 6.5 Assume that V_{xx} is nonsingular. Show that $\rho_{y\cdot x} = 0$ if and only if the best linear predictor of y_1 based on x and y_2 equals the best linear predictor of y_1 based on x alone.

Exercise 6.6 If $(y_{i1}, y_{i2}, x_{i1}, x_{i2}, \ldots, x_{i,p-1})'$, $i = 1,\ldots,n$ are independent $N(\mu, V)$, find the distribution of $\sqrt{n-p-1}\, r_{y\cdot x} \Big/ \sqrt{1-r^2_{y\cdot x}}$ when $\rho_{y\cdot x} = 0$.

Hint: Use the linear model $\mathrm{E}(Y_1|X,Y_2) \in C(J,X,Y_2)$, i.e., $Y_1 = [J,X,Y_2]\gamma + e$, to find a distribution conditional on X,Y_2. Note that the distribution does not depend on the values of X and Y_2, so it must also be the unconditional distribution. Note also that from Exercise 6.5 and the equality between the conditional expectation and the best linear predictor for multivariate normals that we have $\rho_{y \cdot x} = 0$ if and only if the regression coefficient of Y_2 is zero.

Finally, the usual concept of partial correlation, which looks at the correlation between the components of $y - \hat{E}(y|x)$, i.e., the residuals based on best linear prediction, can be generalized to a concept based on examining the correlations between the components of $y - \mathrm{E}(y|x)$, the residuals from the best predictor.

6.6 Testing Lack of Fit

Suppose we have a linear model $Y = X\beta + e$ and we suspect that the model is an inadequate explanation of the data. The obvious thing to do to correct the problem is to add more variables to the model, i.e, fit a model $Y = Z\gamma + e$, where $C(X) \subset C(Z)$. Two questions present themselves: 1) how does one choose Z, and 2) is there really a lack of fit? Given a choice for Z, the second of these questions can be addressed by testing $Y = X\beta + e$ against $Y = Z\gamma + e$. This is referred to as a *test for lack of fit*. Since Z is chosen so that $Y = Z\gamma + e$ will actually fit the data (or at least fit the data better than $Y = X\beta + e$), the error sum of squares for the model $Y = Z\gamma + e$, say $SSE(Z)$, can be called the *sum of squares for pure error*, $SSPE$. The difference $SSE(X) - SSE(Z)$ is used for testing lack of fit, so $SSE(X) - SSE(Z)$ is called the *sum of squares for lack of fit*, $SSLF$.

In general, there are few theoretical guidelines for choosing Z. The most common situation is where there are a variety of other variables that are known and it is necessary to select variables to include in the model. Variable selection techniques are discussed in Chapter 14. In this section, we discuss the problem of testing lack of fit when there are no other variables available for inclusion in the model. With no other variables available, the model matrix X must be used as the basis for choosing Z. We will present four approaches. The first is the traditional method based on having a model matrix X in which some of the rows are identical. A second approach is based on identifying clusters of rows in X that are nearly identical. A third approach examines different subsets of the data. Finally, we briefly mention a nonparametric regression approach to testing lack of fit.

One final note. This section is in the chapter on regression because testing lack of fit has traditionally been considered as a topic in regression analysis. Nowhere in this section do we assume that $X'X$ is nonsingular. *The entire discussion holds for general linear models.*

6.6.1 The Traditional Test

To discuss the traditional approach that originated with Fisher (1922), we require notation for identifying which rows of the model matrix are identical. A model with replications can be written

$$y_{ij} = x_i'\beta + e_{ij},$$

where β is the vector of parameters, $x_i' = (x_{i1}, \ldots, x_{ip})$, $i = 1, \cdots, c$, and $j = 1, \cdots, N_i$. We will assume that $x_i' \neq x_k'$ for $i \neq k$. Using the notation of Chapter 4 in which a pair of subscripts is used to denote a row of a vector or a row of the model matrix, we have $Y = [y_{ij}]$ and

$$X = [w_{ij}'], \qquad \text{where } w_{ij}' = x_i'.$$

The idea of pure error, when there are rows of X that are replicated, is that if several observations have the same mean value, the variability about that mean value is in some sense pure error. The problem is to estimate the mean value. If we estimate the mean value in the ith group with $x_i'\hat{\beta}$, then estimate the variance for the group by looking at the deviations about the estimated mean value, and finally pool the estimates from the different groups, we get $MSE(X)$. Now consider a more general model, $Y = Z\gamma + e$, where $C(X) \subset C(Z)$ and Z is chosen so that

$$Z = [z_{ij}'], \qquad \text{where } z_{ij}' = v_i'$$

for some vectors v_i', $i = 1, \ldots, c$. Two rows of Z are the same if and only if the corresponding rows of X are the same. Thus, the groups of observations that had the same mean value in the original model still have the same mean value in the generalized model. If there exists a lack of fit, we hope that the more general model gives a more accurate estimate of the mean value.

It turns out that there exists a most general model $Y = Z\gamma + e$ that satisfies the condition that two rows of Z are the same if and only if the corresponding rows of X are the same. We will refer to the property that rows of Z are identical if and only if the corresponding rows of X are identical as X and Z having the same *row structure*. X was defined to have c distinct rows, therefore $r(X) \leq c$. Since Z has the same row structure as X, we also have $r(Z) \leq c$. The most general matrix Z, the one with the largest column space, will have $r(Z) = c$. We need to find Z with $C(X) \subset C(Z)$, $r(Z) = c$, and the same row structure as X. We also want to show that the column space is the same for any such Z.

Let Z be the model matrix for the model $y_{ij} = \mu_i + e_{ij}$, $i = 1, \ldots, c$, $j = 1, \ldots, N_i$. If we let $z_{ij,k}$ denote the element in the ijth row and kth column of Z, then from Chapter 4

$$Z = [z_{ij,k}], \qquad \text{where } z_{ij,k} = \delta_{ik}.$$

Z is a matrix where the kth column is 0 everywhere except that it has 1s in rows that correspond to the y_{kj}s. Since the values of $z_{ij,k}$ do not depend on j, it is clear that Z has the same row structure as X. Since the c columns of Z are linearly independent,

we have $r(Z) = c$, and it is not difficult to see that $X = M_Z X$, where M_Z is the perpendicular projection operator onto $C(Z)$; so we have $C(X) \subset C(Z)$.

In fact, because of the form of Z and M_Z, it is clear that any matrix Z_1 with the same row structure as X must have $Z_1 = M_Z Z_1$ and $C(Z_1) \subset C(Z)$. If $r(Z_1) = c$, then it follows that $C(Z_1) = C(Z)$ and the column space of the most general model $Y = Z\gamma + e$ does not depend on the specific choice of Z.

If one is willing to assume that the lack of fit is not due to omitting some variable that, if included, would change the row structure of the model (i.e., if one is willing to assume that the row structure of the true model is the same as the row structure of X), then the true model can be written $Y = W\delta + e$ with $C(X) \subset C(W) \subset C(Z)$. It is easily seen that the lack of fit test statistic based on X and Z has a noncentral F distribution and if $Y = X\beta + e$ is the true model, the test statistic has a central F distribution.

The computations for this lack of fit test are quite simple. With the choice of Z indicated, $C(Z)$ is just the column space for a one-way ANOVA.

$$SSPE = SSE(Z) = Y'(I - M_Z)Y = \sum_{i=1}^{c} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{i\cdot})^2.$$

With $M_Z Y = (\bar{y}_1, \ldots \bar{y}_1, \bar{y}_2, \ldots \bar{y}_2, \ldots, \bar{y}_c, \ldots \bar{y}_c)'$, $\hat{y}_i = x_i'\hat{\beta}$ and $MY = (\hat{y}_1, \ldots \hat{y}_1, \hat{y}_2, \ldots \hat{y}_2, \ldots, \hat{y}_c, \ldots \hat{y}_c)'$, the sum of squares for lack of fit is

$$\begin{aligned} SSLF = Y'(M_Z - M)Y &= [(M_Z - M)Y]'[(M_Z - M)Y] \\ &= \sum_{i=1}^{c} \sum_{j=1}^{N_i} (\bar{y}_{i\cdot} - \hat{y}_i)^2 = \sum_{i=1}^{c} N_i (\bar{y}_{i\cdot} - \hat{y}_i)^2. \end{aligned}$$

Exercise 6.7 Show that if M is the perpendicular projection operator onto $C(X)$ with

$$X = \begin{bmatrix} w_1' \\ \vdots \\ w_n' \end{bmatrix} \quad \text{and} \quad M = \begin{bmatrix} T_1' \\ \vdots \\ T_n' \end{bmatrix},$$

then $w_i = w_j$ if and only if $T_i = T_j$.

Exercise 6.8 Discuss the application of the traditional lack of fit test to the problem where $Y = X\beta + e$ is a simple linear regression model.

As we have seen, in the traditional method of testing for lack of fit, the row structure of the model matrix X completely determines the choice of Z. Now, suppose that none of the rows of X are identical. It is still possible to have lack of fit, but the traditional method no longer applies.

6.6.2 Near Replicate Lack of Fit Tests

Another set of methods for testing lack of fit is based on mimicking the traditional lack of fit test. With these methods, rows of the model matrix that are nearly replicates are identified. One way of identifying near replicates is to use a hierarchical clustering algorithm (see Gnanadesikan, 1977) to identify rows of the model matrix that are near one another. Tests for lack of fit using near replicates are reviewed by Neill and Johnson (1984). The theory behind such tests is beautifully explained in Christensen (1989, 1991). (OK, so I'm biased in favor of this particular author.) Miller, Neill, and Sherfey (1998, 1999) provide a theoretical basis for choosing near replicate clusters.

Christensen (1991) suggests that a very good all-purpose near replicate lack of fit test was introduced by Shillington (1979). Write the regression model in terms of c clusters of near replicates with the ith cluster containing N_i cases, say

$$y_{ij} = x_{ij}'\beta + e_{ij}, \tag{1}$$

$i = 1,\ldots,c$, $j = 1,\ldots,N_i$. Note that at this point we have done nothing to the model except play with the subscripts; model (1) is just the original model. Shillington's test involves finding means of the predictor variables in each cluster and fitting the model

$$y_{ij} = \bar{x}_{i\cdot}'\beta + e_{ij}. \tag{2}$$

The numerator for Shillington's test is then the numerator mean square used in comparing this model to the one-way ANOVA model

$$y_{ij} = \mu_i + e_{ij}. \tag{3}$$

However, the denominator mean square for Shillington's test is the mean squared error from fitting the model

$$y_{ij} = x_{ij}'\beta + \mu_i + e_{ij}. \tag{4}$$

It is not difficult to see that if model (1) holds, then Shillington's test statistic has a central F distribution with the appropriate degrees of freedom. Christensen (1989, 1991) gives details and explains why this should be a good all-purpose test — even though it is not the optimal test for either of the alternatives developed by Christensen. The near replicate lack of fit test proposed in Christensen (1989) is the test of model (1) against model (4). The test proposed in Christensen (1991) uses the same numerator as Shillington's test, but a denominator sum of squares that is the *SSE* from model (1) minus the numerator sum of squares from Shillington's test. Both of Christensen's tests are optimal for certain types of lack of fit. If the clusters consist of exact replicates, then all of these tests reduce to the traditional test.

EXAMPLE 6.6.1. Using the data of Example 6.2.1 we illustrate the near replicate lack of fit tests. Near replicates were chosen visually by plotting x_1 versus x_2. The near replicates are presented below.

Near Replicate Clusters for Steam Data

Obs. no.	x_1	x_2	Near rep.	Obs. no.	x_1	x_2	Near rep.
1	35.3	20	2	14	39.1	19	11
2	29.7	20	2	15	46.8	23	12
3	30.8	23	9	16	48.5	20	3
4	58.8	20	4	17	59.3	22	13
5	61.4	21	5	18	70.0	22	7
6	71.3	22	7	19	70.0	11	1
7	74.4	11	1	20	74.5	23	8
8	76.7	23	8	21	72.1	20	14
9	70.7	21	10	22	58.1	21	5
10	57.5	20	4	23	44.6	20	3
11	46.4	20	3	24	33.4	20	2
12	28.9	21	6	25	28.6	22	15
13	28.1	21	6				

Fitting models (1) through (4) gives the following results:

Model	(1)	(2)	(3)	(4)
dfE	22	22	10	9
SSE	46.50	50.75	12.136	7.5

Shillington's test is

$$F_S = \frac{[50.75 - 12.136]/[22-10]}{7.5/9} = 3.8614 > 3.073 = F(0.95, 12, 9).$$

Christensen's (1989) test is

$$F_{89} = \frac{[46.50 - 7.5]/[22-9]}{7.5/9} = 3.6000 > 3.048 = F(0.95, 13, 9).$$

Christensen's (1991) test is

$$F_{91} = \frac{[50.75 - 12.136]/[22-10]}{[46.5 - (50.75 - 12.136)]/[22-(22-10)]} = 4.0804 > 2.913 = F(0.95, 12, 10).$$

All three tests indicate a lack of fit.

In this example, all of the tests behaved similarly. Christensen (1991) shows that the tests can be quite different and that for the single most interesting type of lack

of fit, the 91 test will typically be more powerful than Shillington's test, which is typically better than the 89 test.

Christensen's 89 test is similar in spirit to "nonparametric" lack of fit tests based on using orthogonal series expansions to approximate general regression models, see Subsection 6.2.1. In particular, it is similar to adding Haar wavelets as additional predictor variables to model (1) except that Haar wavelets amount to adding indicator variables for a predetermined partition of the space of predictor variables while the near replicate methods use the observed predictors to suggest where indicator variables are needed. See Christensen (2001, Section 7.8) and Subsection 4 for additional discussion of these nonparametric approaches.

6.6.3 Partitioning Methods

Another way to use X in determining a more general matrix Z is to partition the data. Write

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}.$$

The model $Y = Z\gamma + e$ can be chosen with

$$Z = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}.$$

Clearly, $C(X) \subset C(Z)$. We again refer to the difference $SSE(X) - SSE(Z)$ as the $SSLF$ and, in something of an abuse of the concept "pure," we continue to call $SSE(Z)$ the $SSPE$.

Exercise 6.9 Let M_i be the perpendicular projection operator onto $C(X_i)$, $i = 1,2$. Show that the perpendicular projection operator onto $C(Z)$ is

$$M_Z = \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix}.$$

Show that $SSE(Z) = SSE(X_1) + SSE(X_2)$, where $SSE(X_i)$ is the sum of squares for error from fitting $Y_i = X_i\beta_i + e_i$, $i = 1,2$.

If there is no lack of fit for $Y = X\beta + e$, since $C(X) \subset C(Z)$, the test statistic will have a central F distribution. Suppose that there is lack of fit and the true model is, say, $Y = W\delta + e$. It is unlikely that W will have the property that $C(X) \subset C(W) \subset C(Z)$, which would ensure that the test statistic has a noncentral F distribution. In general, if there is lack of fit, the test statistic has a doubly noncentral F distribution. (A doubly noncentral F is the ratio of two independent noncentral chi-squareds divided by their degrees of freedom.) The idea behind the lack of fit test based on partitioning the data is the hope that X_1 and X_2 will be chosen so that the combined

fit of $Y_1 = X_1\beta_1 + e_1$ and $Y_2 = X_2\beta_2 + e_2$ will be qualitatively better than the fit of $Y = X\beta + e$. Thus, it is hoped that the noncentrality parameter of the numerator chi-squared will be larger than the noncentrality parameter of the denominator chi-squared.

EXAMPLE 6.6.2. Let $Y = X\beta + e$ be the simple linear regression $y_i = \beta_0 + \beta_1 x_i + e_i$, $i = 1, \ldots, 2r$, with $x_1 \le x_2 \le \cdots \le x_{2r}$. Suppose that the lack of fit is due to the true model being $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$, so the true curve is a parabola. Clearly, one can approximate a parabola better with two lines than with one line. The combined fit of $y_i = \eta_0 + \eta_1 x_i + e_i$, $i = 1, \ldots, r$, and $y_i = \tau_0 + \tau_1 x_i + e_i$, $i = r+1, \ldots, 2r$, should be better than the unpartitioned fit.

EXAMPLE 6.6.3. We now test the model used in Example 6.2.1 for lack of fit using the partitioning method. The difficulty with the partitioning method lies in finding some reasonable way to partition the data. Fortunately for me, I constructed this example, so I know where the lack of fit is and I know a reasonable way to partition the data. (The example was constructed just like Example 13.4.4. The construction is explained in Chapter 13.) I partitioned the data based on the variable x_1. Any case that had a value of x_1 less than 24 went into one group. The remaining cases went into the other group. This provided a group of 12 cases with small x_1 values and a group of 13 cases with large x_1 values. The sum of squares for error for the small group was $SSE(S) = 2.925$ with 9 degrees of freedom and the sum of squares for error for the large group was $SSE(L) = 13.857$ with 10 degrees of freedom. Using the error from Example 6.2.1, we get

$$\begin{aligned} SSPE &= 13.857 + 2.925 = 16.782, \\ dfPE &= 10 + 9 = 19, \\ MSPE &= 0.883, \end{aligned}$$

$$\begin{aligned} SSLF &= 46.50 - 16.78 = 29.72, \\ dfLF &= 22 - 19 = 3, \\ MSLF &= 9.91, \end{aligned}$$

$$F = 9.91/.883 = 11.22.$$

This has 3 degrees of freedom in the numerator, 19 degrees of freedom in the denominator, is highly significant, and indicates a definite lack of fit. But remember, I knew that the lack of fit was related to x_1, so I could pick an effective partition.

Recall from Example 6.4.1 that the R^2 for Example 6.2.1 is 0.964, which indicates a very good predictive model. In spite of the high R^2, we are still able to establish a lack of fit using both the partitioning method and the near replicate method.

The partitioning method can easily be extended, cf. Atwood and Ryan (1977). For example, one could select three partitions of the data and write

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}, \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}, \quad Z = \begin{bmatrix} X_1 & 0 & 0 \\ 0 & X_2 & 0 \\ 0 & 0 & X_3 \end{bmatrix}.$$

The lack of fit test would proceed as before. Note that the partitioning method is actually a generalization of the traditional method. If the partition of the data consists of the different sets of identical rows of the model matrix, then the partitioning method gives the traditional lack of fit test. The partitioning method can also be used to give a near replicate lack of fit test with the partitions corresponding to the clusters of near replicates. As mentioned earlier, it is not clear in general how to choose an appropriate partition.

Utts (1982) presented a particular partition to be used in what she called the Rainbow Test (for lack of fit). She suggests selecting a set of rows from X that are centrally located to serve as X_1, and placing each row of X not in X_1 into a separate set that consists only of that row. With this partitioning, each of the separate sets determined by a single row corresponds to p columns of Z that are zero except for the entries in that row. These p columns are redundant. Eliminating unnecessary columns allows the Z matrix to be rewritten as

$$Z = \begin{bmatrix} X_1 & 0 \\ 0 & I \end{bmatrix}.$$

From Exercise 6.9, it is immediately seen that $SSE(Z) = SSE(X_1)$; thus, the Rainbow Test amounts to testing $Y = X\beta + e$ against $Y_1 = X_1\beta + e_1$. To select the matrix X_1, Utts suggests looking at the diagonal elements of $M = [m_{ij}]$. The smallest values of the m_{ii}s are the most centrally located data points, cf. Section 13.1. The author's experience indicates that the Rainbow Test works best when one is quite selective about the points included in the central partition.

EXAMPLE 6.6.3 CONTINUED. First consider the Rainbow Test using half of the data set. The variables x_1, x_2, and an intercept were fitted to the 12 cases that had the smallest m_{ii} values. This gave a $SSE = 16.65$ with 9 degrees of freedom. The Rainbow Test mean square for lack of fit, mean square for pure error, and F statistic are

$$MSLF = (46.50 - 16.65)/(22 - 9) = 2.296,$$

$$MSPE = 16.65/9 = 1.850,$$

$$F = 2.296/1.850 = 1.24.$$

The F statistic is nowhere near being significant. Now consider taking the quarter of the data with the smallest m_{ii} values. These 6 data points provide a $SSE = 0.862$ with 3 degrees of freedom.

$$MSLF = (46.50 - 0.862)/(22 - 3) = 2.402,$$

$$MSPE = .862/3 = 0.288,$$

$$F = 2.402/0.288 = 8.35.$$

In spite of the fact that this has only 3 degrees of freedom in the denominator, the F statistic is reasonably significant. $F(0.95, 19, 3)$ is approximately 8.67 and $F(0.90, 19, 3)$ is about 5.19.

6.6.4 Nonparametric Methods

As discussed in Subsection 6.2.1, one approach to nonparametric regression is to fit very complicated linear models. One particular application of this approach to nonparametric regression is the fitting of moderate (as opposed to low or high) order polynomials. Christensen (2001, Chapter 7) provides more details of this general approach to nonparametric regression and in particular his Section 7.8 discusses testing lack of fit. Fundamentally, the idea is to test the original linear model $y_i = x_i'\beta + e_i$ against a larger model that incorporates the nonparametric regression components, i.e., $y_i = x_i'\beta + \sum_{j=1}^{q} \gamma_j \phi_j(x_i) + e_i$. For high dimensional problems, the larger model may need to involve generalized additive functions.

Exercise 6.10 Test the model $y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_{ij}$ for lack of fit using the data:

x_i	1.00	2.00	0.00	−3.00	2.50
y_{ij}	3.41	22.26	−1.74	79.47	37.96
	2.12	14.91	1.32	80.04	44.23
	6.26	23.41	−2.55	81.63	
		18.39			

Exercise 6.11 Using the following data, test the model $y_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_{ij}$ for lack of fit. Explain and justify your method.

X_1	X_2	Y	X_1	X_2	Y
31	9.0	122.41	61	2.2	70.08
43	8.0	115.12	36	4.7	66.42
50	2.8	64.90	52	9.4	150.15
38	5.0	64.91	38	1.5	38.15
38	5.1	74.52	41	1.0	45.67
51	4.6	75.02	41	5.0	68.66
41	7.2	101.36	52	4.5	76.15
57	4.0	74.45	29	2.7	36.20
46	2.5	56.22			

6.7 Polynomial Regression and One-Way ANOVA

Polynomial regression is the special case of fitting a model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_{p-1} x_i^{p-1} + e_i,$$

i.e.,

$$Y = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{p-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + e.$$

All of the standard multiple regression results hold, but there are some additional issues to consider. For instance, one should think very hard about whether it makes sense to test $H_0 : \beta_j = 0$ for any j other than $j = p-1$. Frequently, the model

$$y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_{j-1} x_i^{j-1} + \beta_{j+1} x_i^{j+1} + \cdots + \beta_{p-1} x_i^{p-1} + e_i$$

is not very meaningful. Typically, it only makes sense to test the coefficient of the highest order term in the polynomial. One would only test $\beta_j = 0$ if it had already been decided that $\beta_{j+1} = \cdots = \beta_{p-1} = 0$.

Sometimes, polynomial regression models are fitted using orthogonal polynomials. This is a procedure that allows one to perform all the appropriate tests on the β_js without having to fit more than one regression model. The technique uses the Gram–Schmidt algorithm to orthogonalize the columns of the model matrix and then fits a model to the orthogonalized columns. Since Gram–Schmidt orthogonalizes vectors sequentially, the matrix with orthogonal columns can be written $T = XP$, where P is a nonsingular upper triangular matrix. The model $Y = X\beta + e$ is equivalent to $Y = T\gamma + e$ with $\gamma = P^{-1}\beta$. P^{-1} is also an upper triangular matrix, so γ_j is a linear function of $\beta_j, \beta_{j+1}, \ldots, \beta_{p-1}$. The test of $H_0 : \gamma_{p-1} = 0$ is equivalent to the test of $H_0 : \beta_{p-1} = 0$. If $\beta_{j+1} = \beta_{j+2} = \cdots = \beta_{p-1} = 0$, then the test of $H_0 : \gamma_j = 0$ is equivalent to the test of $H_0 : \beta_j = 0$. In other words, the test of $H_0 : \gamma_j = 0$ is equivalent to the test of $H_0 : \beta_j = 0$ in the model $y_i = \beta_0 + \cdots + \beta_j x_i^j + e_i$. However, because the columns of T are orthogonal, the sum of squares for testing $H_0 : \gamma_j = 0$ depends only on the column of T associated with γ_j. It is not necessary to do any additional model fitting to obtain the test.

An algebraic expression for the orthogonal polynomials being fitted is available in the row vector

$$[1, x, x^2, \ldots, x^{p-1}]P. \tag{1}$$

The $p-1$ different polynomials that are contained in this row vector are orthogonal only in that the coefficients of the polynomials were determined so that XP has columns that are orthogonal. As discussed above, the test of $\gamma_j = 0$ is the same as the test of $\beta_j = 0$ when $\beta_{j+1} = \cdots = \beta_{p-1} = 0$. The test of $\gamma_j = 0$ is based on the $(j+1)$st column of the matrix T. β_j is the coefficient of the $(j+1)$st column of X,

i.e., the jth degree term in the polynomial. By analogy, *the* $(j+1)$*st column of* T *is called the* j*th degree orthogonal polynomial*.

Polynomial regression has some particularly interesting relationships with the problem of estimating pure error and with one-way ANOVA problems. It is clear that for all values of p, the row structure of the model matrices for the polynomial regression models is the same, i.e., if $x_i = x_{i'}$, then $x_i^k = x_{i'}^k$; so the i and i' rows of X are the same regardless of the order of the polynomial. Suppose there are q distinct values of x_i in the model matrix. The most general polynomial that can be fitted must give a rank q model matrix; thus the most general model must be

$$y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_{q-1} x_i^{q-1} + e_i.$$

It also follows from the previous section that the column space of this model is exactly the same as the column space for fitting a one-way ANOVA with q treatments.

Using double subscript notation with $i = 1, \ldots, q$, $j = 1, \ldots, N_i$, the models

$$y_{ij} = \beta_0 + \beta_1 x_i + \cdots + \beta_{q-1} x_i^{q-1} + e_{ij} \tag{2}$$

and

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

are equivalent. Since both β_0 and μ are parameters corresponding to a column of 1s, the tests of $H_0 : \beta_1 = \cdots = \beta_{q-1} = 0$ and $H_0 : \alpha_1 = \cdots = \alpha_q$ are identical. Both tests look at the orthogonal complement of J_n with respect to $C(X)$, where $C(X)$ is the column space for either of the models. Using the ideas of Section 3.6, one way to break this space up into $q-1$ orthogonal one-degree-of-freedom hypotheses is to look at the orthogonal polynomials for $i = 1, \ldots, q-1$. As seen in Section 4.2, any vector in $C(X)$ that is orthogonal to J determines a contrast in the α_is. In particular, each orthogonal polynomial corresponds to a contrast in the α_is.

Finding a set of $q-1$ orthogonal contrasts amounts to finding an orthogonal basis for $C(M_\alpha)$. If we write $T = [T_0, \ldots, T_{q-1}]$, then $T_1, \ldots, T_{q-1}$ is an orthogonal basis for $C(M_\alpha)$. Given these vectors in $C(M_\alpha)$, we can use Proposition 4.2.3 to read off the corresponding contrasts. Moreover, the test for dropping, say, T_j from the model is the test of $H_0 : \gamma_j = 0$, which is just the test that the corresponding contrast is zero. Note that testing this contrast is not of interest unless $\beta_{j+1} = \cdots = \beta_{q-1} = 0$ or, equivalently, if $\gamma_{j+1} = \cdots = \gamma_{q-1} = 0$ or, equivalently, if all the higher order polynomial contrasts are zero.

Definition 6.7.1. The orthogonal contrasts determined by the orthogonal polynomials are called the *polynomial contrasts*. The contrast corresponding to the first degree orthogonal polynomial is called the *linear contrast*. The contrasts for higher degree orthogonal polynomials are called the *quadratic*, *cubic*, *quartic*, etc., contrasts.

Using Proposition 4.2.3, if we identify an orthogonal polynomial as a vector $\rho \in C(M_\alpha)$, then the corresponding contrast can be read off. For example, the second

column of the model matrix for (2) is $X_1 = [t_{ij}]$, where $t_{ij} = x_i$ for all i and j. If we orthogonalize this with respect to J, we get the linear orthogonal polynomial. Letting

$$\bar{x}_\cdot = \sum_{i=1}^{q} N_i x_i \Big/ \sum_{i=1}^{q} N_i$$

leads to the linear orthogonal polynomial

$$T_1 = [w_{ij}], \qquad \text{where } w_{ij} = x_i - \bar{x}_\cdot .$$

From Section 4.2, this vector corresponds to a contrast $\sum \lambda_i \alpha_i$, where $\lambda_i / N_i = x_i - \bar{x}_\cdot$. Solving for λ_i gives

$$\lambda_i = N_i (x_i - \bar{x}_\cdot).$$

The sum of squares for testing $H_0 : \beta_1 = \gamma_1 = 0$ is

$$\left[\sum_i N_i (x_i - \bar{x}_\cdot) \bar{y}_{i\cdot} \right]^2 \Big/ \left[\sum_i N_i (x_i - \bar{x}_\cdot)^2 \right].$$

And, of course, one would not do this test unless it had already been established that $\beta_2 = \cdots = \beta_{q-1} = 0$.

As with other directions in vector spaces and linear hypotheses, orthogonal polynomials and orthogonal polynomial contrasts are only of interest up to constant multiples. In applying the Gram–Schmidt theorem to obtain orthogonal polynomials, we really do not care about normalizing the columns. It is the sequential orthogonalization that is important. In the example of the linear contrast, we did not bother to normalize anything.

It is well known (see Exercise 6.12) that, if $N_i = N$ for all i and the quantitative levels x_i are equally spaced, then the orthogonal polynomial contrasts (up to constant multiples) depend only on q. For any value of q, the contrasts can be tabled; see, for example, Snedecor and Cochran (1980) or Christensen (1996a).

Although it is difficult to derive the tabled contrasts directly, one can verify the appropriateness of the tabled contrasts. Again we appeal to Chapter 4. Let $[J, Z]$ be the model matrix for the one-way ANOVA and let X be the model matrix for model (2). The model matrix for the orthogonal polynomial model is $T = XP$. With $C(X) = C(Z)$, we can write $T = ZB$ for some matrix B. Writing the $q \times q$ matrix B as $B = \left[b_0, \ldots, b_{q-1} \right]$ with $b_k = \left(b_{1k}, \ldots, b_{qk} \right)'$, we will show that for $k \geq 1$ and $N_i = N$, the kth degree orthogonal polynomial contrast is $b_k' \alpha$, where $\alpha = (\alpha_1, \ldots, \alpha_q)'$. To see this, note that the kth degree orthogonal polynomial is $T_k = Zb_k$. The first column of X is a column of 1s and T is a successive orthogonalization of the columns of X; so $J_n = Zb_0$ and for $k \geq 1$, $Zb_k \perp J_n$. It follows that $b_0 = J_q$ and, from Chapter 4, for $k \geq 1$, $Zb_k \in C(M_\alpha)$. Thus, for $k \geq 1$, Zb_k determines a contrast $(Zb_k)' Z\alpha \equiv \sum_{i=1}^{q} \lambda_i \alpha_i$. However, $(Zb_k)' Z\alpha = b_k' Z' Z \alpha = b_k' \text{Diag}(N_i) \alpha$. The contrast coefficients $\lambda_1, \ldots, \lambda_q$ satisfy $b_{ik} N_i = \lambda_i$. When $N_i = N$, the contrast is $\sum_i \lambda_i \alpha_i = N \sum_i b_{ik} \alpha_i =$

$Nb_k'\alpha$. Orthogonal polynomial contrasts are defined only up to constant multiples, so the kth degree orthogonal polynomial contrast is also $b_k'\alpha$.

Given a set of contrast vectors $b_1,\ldots,b_{q-1}$, we can check whether these are the orthogonal polynomial contrasts. Simply compute the corresponding matrix $T = ZB$ and check whether this constitutes a successive orthogonalization of the columns of X.

Ideas similar to these will be used in Section 9.4 to justify the use of tabled contrasts in the analysis of balanced incomplete block designs. These ideas also relate to Section 7.3 on Polynomial Regression and the Balanced Two-Way ANOVA. Finally, these ideas relate to nonparametric regression as discussed in Subsection 6.2.1. There, polynomials were used as an example of a class of functions that can be used to approximate arbitrary continuous functions. Other examples mentioned were cosines and wavelets. The development given in this section for polynomials can be mimicked for any approximating class of functions.

For completeness, an alternative idea of orthogonal polynomials should be mentioned. In equation (1), rather than using the matrix P that transforms the columns of X into T with orthonormal columns, one could instead choose a P_0 so that the transformed functions are orthogonal in an appropriate function space. The Legendre polynomials are such a collection. The fact that such orthogonal polynomials do not depend on the specific x_is in the data is both an advantage and a disadvantage. It is an advantage in that they are well known and do not have to be derived for each unique set of x_is. It is a disadvantage in that, although they display better numerical properties than the unadjusted polynomials, since $T_0 = XP_0$ typically does not have (exactly) orthonormal columns, these polynomials will not display the precise features exploited earlier in this section.

Exercise 6.12

(a) Find the model matrix for the orthogonal polynomial model $Y = T\gamma + e$ corresponding to the model

$$y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_{ij},$$

$i = 1,2,3,4$, $j = 1,\ldots,N$, where $x_i = a + (i-1)t$.

Hint: First consider the case $N = 1$.

(b) For the model $y_{ij} = \mu + \alpha_i + e_{ij}$, $i = 1,2,3,4$, $j = 1,\ldots,N$, and for $k = 1,2,3$, find the contrast $\sum \lambda_{ik}\alpha_i$ such that the test of $H_0 : \sum \lambda_{ik}\alpha_i = 0$ is the same as the test of $H_0 : \gamma_k = 0$, i.e., find the polynomial contrasts.

Exercise 6.13 Repeat Exercise 6.11 with $N = 2$ and $x_1 = 2$, $x_2 = 3$, $x_3 = 5$, $x_4 = 8$.

6.8 Additional Exercises

The first three exercises involve deriving *Fieller's method* of finding confidence intervals.

Exercise 6.8.1 *Calibration.*
Consider the regression model $Y = X\beta + e$, $e \sim N(0, \sigma^2 I)$ and suppose that we are interested in a future observation, say y_0, that will be independent of Y and have mean $x_0'\beta$. In previous work with this situation, y_0 was not yet observed but the corresponding vector x_0 was known. The calibration problem reverses these assumptions. Suppose that we have observed y_0 and wish to infer what the corresponding vector x_0 might be.

A typical calibration problem might involve two methods of measuring some quantity: y, a cheap and easy method, and x, an expensive but very accurate method. Data are obtained to establish the relationship between y and x. Having done this, future measurements are made with y and the calibration relationship is used to identify what the value of y really means. For example, in sterilizing canned food, x would be a direct measure of the heat absorbed into the can, while y might be the number of bacterial spores of a certain strain that are killed by the heat treatment. (Obviously, one needs to be able to measure the number of spores in the can both before and after heating.)

Consider now the simplest calibration model, $y_i = \beta_0 + \beta_1 x_i + e_i$, e_is i.i.d. $N(0, \sigma^2)$, $i = 1, 2, 3, \ldots, n$. Suppose that y_0 is observed and that we wish to estimate the corresponding value x_0 (x_0 is viewed as a parameter here).

(a) Find the MLEs of β_0, β_1, x_0, and σ^2.

Hint: This is a matter of showing that the obvious estimates are MLEs.

(b) Suppose now that a series of observations $y_{01}, \ldots, y_{0r}$ were taken, all of which correspond to the same x_0. Find the MLEs of β_0, β_1, x_0, and σ^2.

Hint: Only the estimate of σ^2 changes form.

(c) Based on one observation y_0, find a $(1-\alpha)100\%$ confidence interval for x_0. When does such an interval exist?

Hint: Use an $F(1, n-2)$ distribution based on $(y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0)^2$.

Comment: Aitchison and Dunsmore (1975) discuss calibration in considerable detail, including a comparison of different methods.

Exercise 6.8.2 *Maximizing a Quadratic Response.*
Consider the model, $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$, e_is i.i.d. $N(0, \sigma^2)$, $i = 1, 2, 3, \ldots, n$. Let x_0 be the value at which the function $\mathrm{E}(y) = \beta_0 + \beta_1 x + \beta_2 x^2$ is maximized (or minimized).

(a) Find the maximum likelihood estimate of x_0.

(b) Find a $(1-\alpha)100\%$ confidence interval for x_0. Does such an interval always exist?

Hint: Use an $F(1, n-3)$ distribution based on $(\hat{\beta}_1 + 2\hat{\beta}_2 x_0)^2$.

Comment: The problem of finding values of the independent variables that maximize (or minimize) the expected y value is a basic problem in the field of response surface methods. See Box, Hunter, and Hunter (1978) or Christensen (2001, Chapter 8) for an introduction to the subject, or Box and Draper (1987) for a detailed treatment.

Exercise 6.8.3 *Two-Phase Linear Regression.*
Consider the problem of sterilizing canned pudding. As the pudding is sterilized by a heat treatment, it is simultaneously cooked. If you have ever cooked pudding, you know that it starts out soupy and eventually thickens. That, dear reader, is the point of this little tale. Sterilization depends on the transfer of heat to the pudding and the rate of transfer depends on whether the pudding is soupy or gelatinous. On an appropriate scale, the heating curve is linear in each phase. The question is, "Where does the line change phases?"

Suppose that we have collected data (y_i, x_i), $i = 1, \ldots, n+m$, and that we know that the line changes phases between x_n and x_{n+1}. The model $y_i = \beta_{10} + \beta_{11} x_i + e_i$, e_is i.i.d. $N(0, \sigma^2)$, $i = 1, \ldots, n$, applies to the first phase and the model $y_i = \beta_{20} + \beta_{21} x_i + e_i$, e_is i.i.d. $N(0, \sigma^2)$, $i = n+1, \ldots, n+m$, applies to the second phase. Let γ be the value of x at which the lines intersect.

(a) Find estimates of β_{10}, β_{11}, β_{20}, β_{21}, σ^2, and γ.

Hint: γ is a function of the other parameters.

(b) Find a $(1-\alpha)100\%$ confidence interval for γ. Does such an interval always exist?

Hint: Use an $F(1, n+m-4)$ distribution based on

$$\left[(\hat{\beta}_{10} + \hat{\beta}_{11}\gamma) - (\hat{\beta}_{20} + \hat{\beta}_{21}\gamma) \right]^2.$$

Comment: Hinkley (1969) has treated the more realistic problem in which it is not known between which x_i values the intersection occurs.

Exercise 6.8.4 Consider the model $y_i = \beta_0 + \beta_1 x_i + e_i$, e_is i.i.d. $N(0, \sigma^2 d_i)$, $i = 1, 2, 3, \ldots, n$, where the d_is are known numbers. Derive algebraic formulas for $\hat{\beta}_0$, $\hat{\beta}_1$, $\mathrm{Var}(\hat{\beta}_0)$, and $\mathrm{Var}(\hat{\beta}_1)$.

Exercise 6.8.5 Consider the model $y_i = \beta_0 + \beta_1 x_i + e_i$, e_is i.i.d. $N(0, \sigma^2)$, $i = 1, 2, 3, \ldots, n$. If the x_is are restricted to be in the closed interval $[-10, 15]$, determine how to choose the x_is to minimize

(a) $\mathrm{Var}(\hat{\beta}_0)$.

(b) $\mathrm{Var}(\hat{\beta}_1)$.

(c) How would the choice of the x_is change if they were restricted to the closed interval $[-10, 10]$?

Exercise 6.8.6 Find $\mathrm{E}[y - \hat{E}(y|x)]^2$ in terms of the variances and covariances of x and y. Give a "natural" estimate of $\mathrm{E}\left[y - \hat{E}(y|x)\right]^2$.

Exercise 6.8.7 Test whether the data of Example 6.2.1 indicate that the multiple correlation coefficient is different from zero.

Exercise 6.8.8 Test whether the data of Example 6.2.1 indicate that the partial correlation coefficient $\rho_{y1\cdot 2}$ is different from zero.

Exercise 6.8.9 Show that

(a) $$\rho_{12\cdot 3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}}$$

(b) $$\rho_{12\cdot 34} = \frac{\rho_{12\cdot 4} - \rho_{13\cdot 4}\rho_{23\cdot 4}}{\sqrt{1-\rho_{13\cdot 4}^2}\sqrt{1-\rho_{23\cdot 4}^2}}.$$

Exercise 6.8.10 Show that in Section 2, $\gamma_* = \beta_*$ and $\beta_0 = \gamma_0 - (1/n)J_1^n Z\gamma_*$.

Hint: Examine the corresponding argument given in Section 1 for simple linear regression.