CHAPTER 14

"Dummy" Variables

What Are "Dummy" Variables?

The variables considered in regression equations usually can take values over some continuous range. Occasionally we must introduce a factor that has two or more distinct levels. For example, data may arise from three machines, or two factories, or six operators. In such a case we cannot set up a continuous scale for the variable "machine" or "factory" or "operator." We must assign to these variables some levels in order to take account of the fact that the various machines or factories or operators may have separate deterministic effects on the response. Variables of this sort are usually called *dummy variables.* They are usually (but not always) unrelated to any physical levels that might exist in the factors themselves.

One example of a dummy variable is found in the attachment of a variable X_0 (whose value is always unity) to the term β_0 in a regression model. The X_0 is unnecessary but provides a notational convenience at times. Other dummy variables are somewhat more than a mere convenience, as we shall see.

An Infinite Number of Choices

The suggestions we make for setting up dummy variable systems are not unique. Typically, there are an infinite number of alternative ways to set up a system to cover any particular type of situation. Given a particular selection of dummy variable vectors that "works," that is, represents the factors as desired, we can derive other sets by taking linear combinations of the vectors in the first set. As long as the second set is chosen so that their vectors are not linearly dependent on one another, all will be well. In general, the most useful dummy variable setups are simple in form, employing levels of 0 and 1, for example, or −1 and 1. Usefulness, however, lies in the eye of the user (as "beauty lies in the eye of the beholder").

14.1. DUMMY VARIABLES TO SEPARATE BLOCKS OF DATA WITH DIFFERENT INTERCEPTS, SAME MODEL

Suppose we wish to introduce into a model the idea that there are two types of machines (types A and B, say) that produce different levels of response, in addition to the variation that occurs due to other predictor variables. One way of doing this is to add to the model a dummy variable Z and regression coefficient α (say) so that an additional term αZ appears in the model. The coefficient α must be estimated at the same time the β's are estimated. Values can be assigned to Z as follows:

$$Z = 0 \text{ if the observation is from machine } A,$$

$$Z = 1 \text{ if the observation is from machine } B.$$

If a is the least squares estimate of α, and if $\hat{f}$ represents the rest of the fitted model, we have

$$\hat{Y} = \hat{f} + aZ \tag{14.1.1}$$

Thus machine A data are estimated by putting $Z = 0$ to get $\hat{Y} = \hat{f}$, while machine B data are predicted by setting $Z = 1$ to give $\hat{Y} = \hat{f} + a$. The value "a" simply estimates the difference in levels between the responses of group B compared to group A and all other factors fitted are represented in $\hat{f}$.

Other Possibilities

Any two distinct values of Z would actually be suitable, though the above is usually best. However, other assignments are sometimes convenient; for example, suppose that of a total of n observations, n_1 come from type A machines and $n_2 = n - n_1$ from type B machines. If we choose levels

$$\begin{aligned} Z &= \frac{-n_2}{\sqrt{n_1 n_2(n_1 + n_2)}} \quad \text{for machine } A, \\ Z &= \frac{n_1}{\sqrt{n_1 n_2(n_1 + n_2)}} \quad \text{for machine } B, \end{aligned} \tag{14.1.2}$$

it will be found that the corresponding column of the $\mathbf{X}$ matrix is orthogonal to the "β_0 column" and is "normalized," that is, has sum of squares unity, which may be convenient. (We can also omit the denominators if the normalization of the column is of no consequence.) When the column is normalized, the $\hat{Y}$(group B) $-$ $\hat{Y}$(group A) difference is $a(n_1 + n_2)^{1/2}/(n_1 n_2)^{1/2}$. Or, if we choose $Z = -1$ for machine A, $Z = 1$ for machine B, the difference $\hat{Y}$(group B) $-$ $\hat{Y}$(group A) is $2a$. Obviously all of the choices above achieve the same purpose, to provide a difference in levels between the two groups. Thus it is sensible to use a representation that is convenient to the user.

To see how one representation is derived by linear combination from another we *must* count in the dummy $\mathbf{X}_0$ column of the $\mathbf{X}$ matrix. The first representation is covered by the vectors

$$(\mathbf{X}_0, \mathbf{Z}) = \left[\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ \cdots & \\ 1 & 0 \\ \hline 1 & 1 \\ 1 & 1 \\ \cdots & \cdots \\ 1 & 1 \end{array}\right] \begin{array}{l} \\ \\ n_1 \\ \\ \\ \\ n_2 \\ \\ \end{array}. \tag{14.1.3}$$

The second representation (14.1.2) has columns $(\mathbf{X}_0, \mathbf{U})$, where

$$\begin{aligned}\mathbf{U} &= \{-n_2/[n_1n_2(n_1+n_2)]^{1/2}\}(\mathbf{X}_0 - \mathbf{Z}) + \{n_1/[n_1n_2(n_1+n_2)]^{1/2}\}\mathbf{Z} \\ &= (n_1+n_2)^{1/2}/(n_1n_2)^{1/2}\mathbf{Z} - \{n_2/[n_1n_2(n_1+n_2)^{1/2}\}\mathbf{X}_0\end{aligned} \qquad (14.1.4)$$

and the third has columns $(\mathbf{X}_0, \mathbf{W})$, where

$$\mathbf{W} = (\mathbf{Z} - \mathbf{X}_0) + \mathbf{Z} = 2\mathbf{Z} - \mathbf{X}_0. \qquad (14.1.5)$$

How Many Dummies?

In the above example of two categories (machines A and B) we see we need to construct one dummy column *in addition to* $\mathbf{X}_0$. So two groups require two dummies *including* $\mathbf{X}_0$. This essentially enables us to have the same linear model with two different intercepts. For the first illustration above, the estimated intercepts are, conveniently, b_0 and $(b_0 + a)$. For the third they are $(b_0 - a)$ and $(b_0 + a)$ with a different value of a. (We leave the second to the reader.)

Three Categories, Three Dummies

If we wish to take account of three different categories, two extra dummies (besides $\mathbf{X}_0$) would be needed. The simplest way is to use

$$\begin{aligned}(Z_1, Z_2) &= (1, 0) \quad \text{for machine } A, \\ &= (0, 1) \quad \text{for machine } B, \\ &= (0, 0) \quad \text{for machine } C,\end{aligned} \qquad (14.1.6)$$

and the model would include extra terms $\alpha_1 Z_1 + \alpha_2 Z_2$, with coefficients α_1, α_2 to be estimated. Thus the $\mathbf{X}$ matrix for such data would take the form below, assuming that all the A-data were listed first, then all the B-data, then all the C-data.

$$\begin{array}{ccccccc} & \mathbf{X}_0 & \text{Other } X\text{'s} & \mathbf{Z}_1 & \mathbf{Z}_2 & & \\ & \left[\begin{array}{cccc} 1 & \cdots & 1 & 0 \\ 1 & & 1 & 0 \\ \cdots & & \cdots & \\ 1 & & 1 & 0 \\ \hline 1 & \cdots & 0 & 1 \\ 1 & & 0 & 1 \\ \cdots & & \cdots & \\ 1 & & 0 & 1 \\ \hline 1 & \cdots & 0 & 0 \\ 1 & & 0 & 0 \\ \cdots & & \cdots & \\ 1 & & 0 & 0 \end{array}\right. & & & \left.\begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array}\right] & & \end{array}$$

$$\mathbf{X} = \left[\begin{array}{cccc} 1 & \cdots & 1 & 0 \\ 1 & & 1 & 0 \\ \cdots & & \cdots & \\ 1 & & 1 & 0 \\ \hline 1 & \cdots & 0 & 1 \\ 1 & & 0 & 1 \\ \cdots & & \cdots & \\ 1 & & 0 & 1 \\ \hline 1 & \cdots & 0 & 0 \\ 1 & & 0 & 0 \\ \cdots & & \cdots & \\ 1 & & 0 & 0 \end{array}\right] \begin{array}{l} \left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right\} \text{Group } A \\ \left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right\} \text{Group } B. \\ \left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right\} \text{Group } C \end{array} \qquad (14.1.7)$$

Again, many different allocations of levels are possible. If desired, columns that are orthogonal to the $\mathbf{X}_0$ column and that have sum of squares unity can be achieved by setting

$$\begin{aligned}(Z_1, Z_2) &= \left(\frac{-n_3}{\sqrt{n_1 n_3(n_1+n_3)}},\ 0\right) && \text{for machine } A,\\ &= \left(0,\ \frac{-n_3}{\sqrt{n_2 n_3(n_2+n_3)}}\right) && \text{for machine } B, \qquad (14.1.8)\\ &= \left(\frac{n_1}{\sqrt{n_1, n_3(n_1+n_3)}},\ \frac{n_2}{\sqrt{n_2 n_3(n_2+n_3)}}\right) && \text{for machine } C,\end{aligned}$$

where n_1, n_2, and n_3 are, respectively, the numbers of observations from machines A, B, and C. These $\mathbf{Z}_1$, $\mathbf{Z}_2$ columns are not orthogonal to each other but two orthogonal columns could be constructed. Again, all denominators could be dropped in (14.1.8).

r Categories, *r* Dummies

In general, by an extension of this procedure, we can deal with r levels by the introduction of $(r - 1)$ dummy variables in addition to $\mathbf{X}_0$. The basic allocation pattern is obtained by writing down an $(r - 1) \times (r - 1)$ $\mathbf{I}$ matrix and adding a row of $(r - 1)$ zeros. The case $r = 6$ is illustrated by the $X_1, X_2, \dots, X_5$ columns in the third display of Example 2 below.

We now give an example of the use of dummy variables in this manner.

Example 1. The data in Table 14.1 show turkey weights (Y) in pounds, and ages (X) in weeks, of 13 Thanksgiving turkeys. Four of these turkeys were reared in Georgia (G), four in Virginia (V), and five in Wisconsin (W). We would like to relate Y to X via a simple straight line model, but the different origins of the turkeys may cause a problem. If they do, how do we handle it?

Suppose we first regress Y against X to give the fitted equation $\hat{Y} = 1.98 + 0.4167X$. The residuals from this fit are, in order, -0.4, -1.4, -0.2, -0.8, -1.0, -0.8, -0.5, -1.0, 0.8, 1.0, 1.3, 1.5, 1.4. When plotted in sets according to origin, they give rise to

TABLE 14.1. Turkey Data (X, Y, Origin) and Dummy Variables (Z_1, Z_2)

X	Y	Origin	Z_1	Z_2
28	13.3	G	1	0
20	8.9	G	1	0
32	15.1	G	1	0
22	10.4	G	1	0
29	13.1	V	0	1
27	12.4	V	0	1
28	13.2	V	0	1
26	11.8	V	0	1
21	11.5	W	0	0
27	14.2	W	0	0
29	15.4	W	0	0
23	13.1	W	0	0
25	13.8	W	0	0

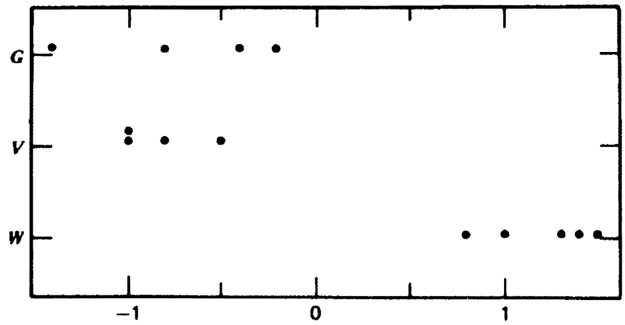

Figure 14.1. Turkey data; residuals from the fitted equation $\hat{Y} = 1.98 + 0.4167X$ plotted against origin of turkey.

Figure 14.1, which clearly signals the need to take account of the different levels of response. To do this, we select the dummy variables Z_1, Z_2 shown in Table 14.1 and fit the model

$$Y = \beta_0 + \beta_1 X + \alpha_1 Z_1 + \alpha_2 Z_2 + \epsilon \tag{14.1.9}$$

by least squares. The fitted equation is

$$\hat{Y} = 1.43 + 0.4868X - 1.92Z_1 - 2.19Z_2. \tag{14.1.10}$$

The estimates $a_1 = -1.92$ and $a_2 = -2.19$ estimate the differences in response levels between (1) sources G and W, and (2) sources V and W, respectively. By substituting for the three sets of values for (Z_1, Z_2) we obtain, for the three different origins,

$$\begin{aligned} \hat{Y} &= -0.49 + 0.4868X, && \text{for } G; \\ \hat{Y} &= -0.76 + 0.4868X, && \text{for } V; \\ \hat{Y} &= 1.43 + 0.4868X, && \text{for } W. \end{aligned} \tag{14.1.11}$$

The original data and the three fitted straight lines are shown in Figure 14.2. The three lines are all parallel but have different intercepts. The analysis of variance for this fitted model can be written as shown in Table 14.2. Both F-values are highly significant, implying that use of the dummies is clearly worthwhile and that the lines appear to have a definite nonzero slope. Of the variation of the data about the mean, 97.94% has been explained by this equation. (Without the dummies, only 66.47% is explained.)

If desired, t-tests can be constructed to test for differences between the intercepts. For example, the true $W - G$ difference is estimated by $-a_1 = 1.92$ and this, divided by its standard error, namely, the square root of the appropriate diagonal term of the $(\mathbf{X}'\mathbf{X})^{-1}s^2$ matrix, gives a t-value whose modulus (positive value) is compared to the percentage point $t(9, 1 - \frac{1}{2}\alpha)$ for a two-sided test of the null hypothesis $H_0: \alpha_1 = 0$ versus $H_1: \alpha_1 \neq 0$. We find, for our data, $t = 1.92/0.201 = 9.55$, which is significant at 0.1%. An alternative and equivalent test is given by using

$$F = \{\text{SS}(a_1|b_0, b_1, a_2)/1\}/s^2 = 8.145/0.090 = 90.50.$$

This is compared with $F(1, 9, 1 - \alpha)$ for a test at the same level. The result is identical, because the F-value is, theoretically, the square of the t-value above; here $t^2 = 91.20$ would be the same as $F = 90.50$ were it not for rounding differences. A test for $H_0: \alpha_2 = 0$, α_2 being the true $V - W$ difference, can be carried out in a similar manner. The t-value is $-2.19/0.21 = -10.43$, which is also significant at 0.1%. The estimated

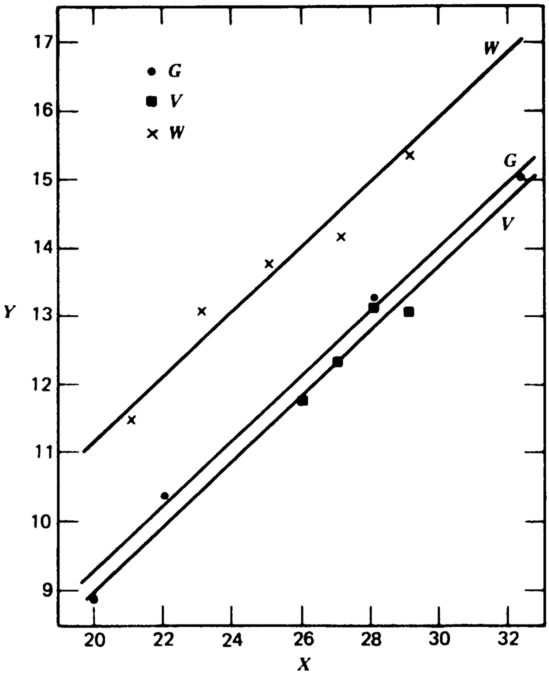

Figure 14.2. Plot of the turkey data and the three fitted straight lines.

$G - V$ difference is given by $a_1 - a_2 = 0.27$, which has an estimated variance of Est. $V(a_1)$ + Est. $V(a_2)$ − 2 Est. cov (a_1, a_2), all three terms of which can be obtained from the $(\mathbf{X}'\mathbf{X})^{-1}s^2$ matrix. We find Est. $V(a_1 - a_2) = 0.040369 + 0.044310 - 2(0.018690) = 0.047299 = (0.217)^2$. The t-value is thus $0.27/0.217 = 1.24$, not significant. Thus overall, real differences appear to exist between the G and W levels and between the V and W levels but not between the G and V levels.

An Alternative Analysis of Variance Sequence

The order indicated in the source column of Table 14.2 removes the differences between intercepts first and leaves the remaining variation to $(b_1|b_0, a_1, a_2)$ and residual. It would also be possible to use the order b_0; $(b_1|b_0)$; $(a_1, a_2|b_0, b_1)$; and residual, and so test the hypothesis $H_0: \alpha_1 = \alpha_2 = 0$ (which asks "Do the three sets of data here have different intercepts?") using an extra sum of squares F-test based on $F(2, 9)$. One would then still have to check where those differences were, if a significant result were obtained.

TABLE 14.2. Analysis of Variance for Turkey Example

Source	df	SS	MS	F
b_0	1	2124.803		
$a_1, a_2\|b_0$	2	6.382	3.191	35.46
$b_1\|b_0, a_1, a_2$	1	32.224	32.224	358.04
Residual	9	0.811	$s^2 = 0.090$	
Total	13	2164.220		

Will My Selected Dummy Setup Work?

The vectors of (14.1.3) can be described by writing down the components in a two by two matrix as

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \tag{14.1.12}$$

where the first row denotes the values of (X_0, Z) for a group A piece of data while the second row applies to a group B piece of data. If this matrix has a nonzero determinant (it is, of course, easy to see that it does) the setup will work. For the turkey data, the corresponding matrix is

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \tag{14.1.13}$$

and the determinant is 1, also. Let us examine the six-group case.

Example 2. Is the dummy variable scheme below a workable one for dealing with possible level differences among six groups?

Group	Z_0	Z_1	Z_2	Z_3	Z_4	Z_5
1	1	1	1	1	1	1
2	1	0	1	1	1	1
3	1	0	0	1	1	1
4	1	0	0	0	1	1
5	1	0	0	0	0	1
6	1	0	0	0	0	0

The determinant value is −1, so the system will work. Evaluating determinants is not difficult, and rules for quick reduction may be found in matrix algebra books. For our purposes, it is usually simplest to use the computer to find the determinant via the so-called eigenvalues, the values $\lambda_1, \lambda_2, \ldots, \lambda_q$, say, which are the roots of the equation $\det(\mathbf{M} - \lambda\mathbf{I}) = 0$, where $\mathbf{M}$ is a (square) matrix of interest and det $(\cdot)$ means the determinant of $(\cdot)$. The reason this works is that

$$\det \mathbf{M} = \lambda_1\lambda_2 \cdots \lambda_q$$

so that, if all the eigenvalues are nonzero, so is the determinant. Problems with "almost-dependent" columns can arise here too, but use of 1's and 0's as dummy levels will usually avoid such problems.

In the MINITAB system, for example, we can write:

```
read 6 6 m1
1 1 1 1 1 1
1 0 1 1 1 1
1 0 0 1 1 1
1 0 0 0 1 1
1 0 0 0 0 1
1 0 0 0 0 0
transpose m1 m2
multiply m2 m1 m3
eigen m3 c3
print c3
end
stop
```

The output from this routine is the column of numbers (which we write as a row here)

$$17.2069 \quad 1.9882 \quad 0.7747 \quad 0.4462 \quad 0.3189 \quad 0.2652;$$

the product of these is 1 not -1. The reason for this is that the MINITAB eigen program requires a symmetric matrix, and we have created one by evaluating $\mathbf{M} = \mathbf{m}_1'\mathbf{m}_1$ before calling for eigenvalues. Because $\mathbf{m}_1$ is square, $\det(\mathbf{m}_1'\mathbf{m}_1) = (\det \mathbf{m}_1')(\det \mathbf{m}_1) = (\det \mathbf{m}_1)^2$. Even though we do not get the sign from our result, we know det $\mathbf{m}_1$ is nonzero. Evaluation of the eigenvalues of $\mathbf{m}_1\mathbf{m}_1'$ would give identical results because $\mathbf{m}_1$ is square. Note that we do not actually need to take the product of the eigenvalues but can just look at the smallest one to see that it is nonzero (and not almost zero).

Other Verification Methods

A second way of checking a setup like that of Example 2 is to relate it to the basic scheme. Recall that our basic scheme of vectors for this situation, written down with the X_0 column, was as follows:

X_0	X_1	X_2	X_3	X_4	X_5
1	1	0	0	0	0
1	0	1	0	0	0
1	0	0	1	0	0
1	0	0	0	1	0
1	0	0	0	0	1
1	0	0	0	0	0

We see immediately that

$$\begin{aligned} Z_0 &= X_0, & Z_3 &= X_1 + X_2 + X_3, \\ Z_1 &= X_1, & Z_4 &= X_1 + X_2 + X_3 + X_4, \\ Z_2 &= X_1 + X_2, & Z_5 &= X_1 + X_2 + X_3 + X_4 + X_5. \end{aligned}$$

This establishes the Z's as linear combinations of the X's. Moreover, each Z_j, $j = 2$, 3, 4, 5, in turn introduces an additional X_j so that none of these Z_j can be linearly dependent on prior Z's, and vice versa. It follows that the system will work for separating the intercept levels of six groups of data.

A third, somewhat more tedious, verification is to write down the condition that the columns are dependent and then solve the resulting equations for the constants that form the linear combination. For Example 2, we write that

$$aZ_0 + bZ_1 + cZ_2 + dZ_3 + eZ_4 + fZ_5 = 0,$$

and this must hold row by row, which implies that

$$\begin{aligned} a + b + c + d + e + f &= 0, \\ a \quad\; + c + d + e + f &= 0, \\ a \quad\quad\quad + d + e + f &= 0, \\ a \quad\quad\quad\quad\quad + e + f &= 0, \\ a \quad\quad\quad\quad\quad\quad\quad + f &= 0, \\ a &= 0. \end{aligned}$$

Working from the bottom equation up, it is obvious that all the coefficients $a, \ldots, f$ are zero. That means *no* linear combination gives zero, so that the Z's are linearly independent and the system works. If a *nonzero* numerical solution emerges, however, dependence of the Z's is established, making the setup useless.

Example 3. Another workable system in the same context as Example 2 would consist of columns $Z_0 = X_0$, $Z_i = X_0 + X_i$, $i = 1, 2, \ldots, 5$. This would lead to the following scheme:

Group	Z_0	Z_1	Z_2	Z_3	Z_4	Z_5
1	1	2	1	1	1	1
2	1	1	2	1	1	1
3	1	1	1	2	1	1
4	1	1	1	1	2	1
5	1	1	1	1	1	2
6	1	1	1	1	1	1

Note: The dummy variable columns not only must form a linearly independent set themselves but must also form a linearly independent set when they are united with other predictors variable columns in the full regression. In most regressions, the occurrence of such a dependence is unlikely, but it would show up when the full regression is performed, requiring the dummies to be selected differently.

14.2. INTERACTION TERMS INVOLVING DUMMY VARIABLES

The way we used dummy variables in the foregoing section allowed us to fit the same basic model with different intercepts to several sets of data. By an extension of those ideas, we can also allow changes in the way the predictors enter.

Two Sets of Data, Straight Line Models

Suppose A and B denote two data sets, and we are considering fits involving straight lines. There are four possibilities, shown in Figure 14.3.

(a) Two distinct lines, $\beta_0 + \beta_1 X$, $\gamma_0 + \gamma_1 X$, four parameters.

(b) Two parallel lines, $\beta_0 + \beta_1 X$, $\gamma_0 + \beta_1 X$, three parameters.

(c) Two lines with the same intercept $\beta_0 + \beta_1 X$, $\beta_0 + \gamma_1 X$, three parameters.

(d) One line, $\beta_0 + \beta_1 X$, two parameters.

We can take care of all four possibilities at once by choosing two dummies (we count the column of 1's associated with the intercept as one of these):

X_0	Z		
1	0	for the A data	(14.2.1)
1	1	for the B data	

and then fitting the model

$$Y = X_0(\beta_0 + \beta_1 X) + Z(\alpha_0 + \alpha_1 X) + \epsilon$$

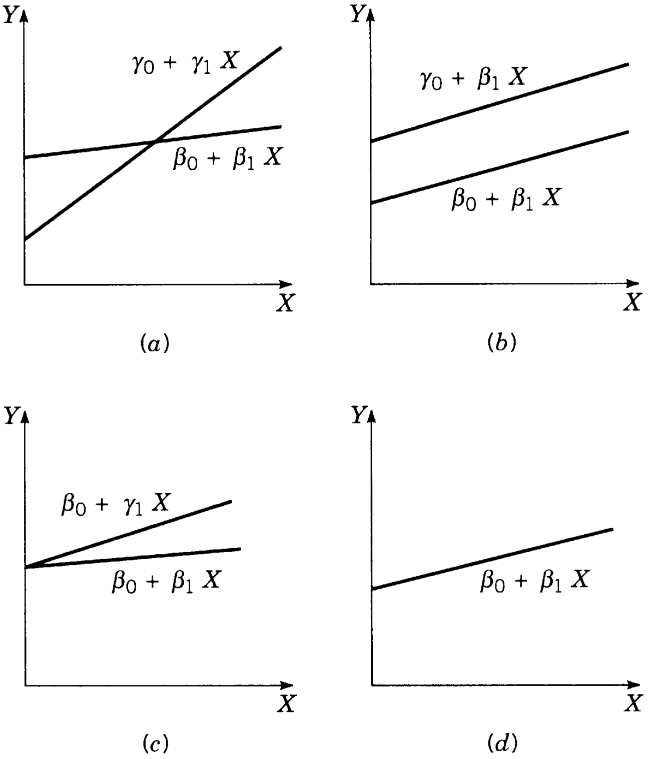

Figure 14.3. Four possibilities for two straight lines.

or

$$Y = \beta_0 + \beta_1 X + \alpha_0 Z + \alpha_1 XZ + \epsilon \tag{14.2.2}$$

when the parentheses are removed. We see that the model contains not only Z but an interaction term involving Z as well. The separate models for the A and B lines are given by setting $Z = 0$ and $Z = 1$ to produce the

$$\left.\begin{aligned} &A \text{ model function:} \quad \beta_0 + \beta_1 X, \\ &B \text{ model function:} \quad (\beta_0 + \alpha_0) + (\beta_1 + \alpha_1)X. \end{aligned}\right. \tag{14.2.3}$$

Thus in Figure 14.3*a*, $\gamma_0 = \beta_0 + \alpha_0$ and $\gamma_1 = \beta_1 + \alpha_1$. The parameters α_0 and α_1 represent changes needed to get from the A model function to the B model function. To test whether the two parallel lines of Figure 14.3*b* will do, we would fit (14.2.2) and then test $H_0: \alpha_1 = 0$ versus $H_1: \alpha_1 \neq 0$. To test for the appropriateness of the case in Figure 14.3*c*, the null hypothesis used would be $H_0: \alpha_0 = 0$, while for the case in Figure 14.3*d* it would be $H_0: \alpha_0 = \alpha_1 = 0$. In all cases, the null hypotheses reduce the two model functions in (14.2.3) to the various reduced forms shown in Figure 14.3, with the identification $\gamma_0 = \beta_0 + \alpha_0$ and $\gamma_1 = \beta_1 + \alpha_1$.

Hierarchical Models

In discussing extra sums of squares in various earlier contexts, for example, Chapter 6, we saw that the terms of the smaller model always formed a part of the larger model. Sequential sums of squares calculations involve a hierarchy of models, a sequence in which each model contains all the terms in the model below it in the hierarchical order (as well as one more term). These hierarchical aspects arise in our current discussion also. Look at Figure 14.3. We can write

$$(a) \underset{\alpha_1}{>} (b) \underset{\alpha_0}{>} (d)$$

to denote that model (d) is contained in (b) when $\alpha_0 = 0$, which is itself contained in (a) when $\alpha_1 = 0$. Similarly,

$$(a) \underset{\alpha_0}{>} (c) \underset{\alpha_1}{>} (d)$$

Appropriate sequences of tests on the parameters quickly become clear when one arranges the various models into hierarchical patterns in this way.

Three Sets of Data, Straight Line Models

Before we discuss the general case, let us extend (14.2.2) to three straight lines and illustrate with the turkey data of Table 14.1. To allow the fitting of three separate straight lines, we form the model

$$Y = Z_0(\beta_0 + \beta_1 X) + Z_1(\gamma_0 + \gamma_1 X) + Z_2(\delta_0 + \delta_1 X) + \epsilon, \tag{14.2.4}$$

where $Z_0 = 1$ and Z_1 and Z_2 are two additional dummy variables. We shall choose these as follows:

Z_1	Z_2	Line
1	0	G (Georgia)
0	1	V (Virginia)
0	0	W (Wisconsin)

We rewrite the model as

$$Y = \beta_0 + \beta_1 X + \gamma_0 Z_1 + \gamma_1(Z_1 X) + \delta_0 Z_2 + \delta_1(Z_2 X) + \epsilon \tag{14.2.5}$$

and note that there are two "interaction with dummies" terms, Z_1X and Z_2X. The fitted equation is

$$\begin{aligned}\hat{Y} &= 2.475 + 0.4450X - 3.454Z_1 - 2.775Z_2 \\ &\quad + 0.06104(Z_1X) + 0.02500(Z_2X).\end{aligned} \tag{14.2.6}$$

The three separate straight lines are then

$$\begin{aligned}\hat{Y} &= -0.979 + 0.5060X && (\text{setting } Z_1 = 1, Z_2 = 0), \\ \hat{Y} &= -0.300 + 0.4700X && (\text{setting } Z_1 = 0, Z_2 = 1), \\ \hat{Y} &= 2.475 + 0.4450X && (\text{setting } Z_1 = 0, Z_2 = 0).\end{aligned} \tag{14.2.7}$$

These lines, which are exactly what one would find if one fitted each subset of data separately, are slightly displaced from the lines shown in Figure 14.2, as the reader can confirm by plotting, or by simply comparing the fitted equations above with those in Eq. (14.1.11). The analysis of variance table for this fit takes the form

ANOVA

Source	df	SS	MS	F
b_0	1	2124.803		
$b_1, c_0, c_1, d_0, d_1 \| b_0$	5	38.711	7.742	76.6
Residual	7	0.706	0.101	
Total	13	2164.220		

The three fitted lines would be identical if $H_0: \gamma_0 = \gamma_1 = \delta_0 = \delta_1 = 0$ were true. To test this hypothesis versus $H_1: H_0$ not true, the extra sum of squares for c_0, c_1, d_0, and d_1, given by

$$\text{SS}(b_1, c_0, c_1, d_0, d_1|b_0) - \text{SS}(b_1|b_0) = 38.71 - 26.20 = 12.51$$

with 4 df is needed. (The figure 26.20 is the regression sum of squares when a common line is fitted; it was not given earlier.) The appropriate F-statistic is

$$F = (12.51/4)/(0.101) = 30.97,$$

which exceeds $F(4, 7, 0.99) = 7.85$, so that H_0 is rejected. This, of course, is not surprising, as we have already seen when the data were given initially.

We can test the hypothesis that there are three parallel lines, that is, $H_0: \gamma_1 = \delta_1 = 0$ versus $H_1: H_0$ not true, by finding the extra sum of squares for c_1 and d_1 via

$$\text{SS}(b_1, c_0, c_1, d_0, d_1|b_0) - \text{SS}(b_1, c_0, d_0|b_0) = 38.71 - 38.61 = 0.10,$$

the figure 38.61 being the sum of the second and third entries in Table 14.2. This value 0.10 has 2 df and provides the nonsignificant F-ratio $(0.10/2)/0.101 = 0.50$. We do not reject H_0, and so the fit illustrated in Figure 14.2 is clearly a satisfactory one.

As our example shows, the use of interaction terms involving dummy variables makes it easy to formulate appropriate tests and to obtain the right test statistics. This may be the method's greatest virtue.

Two Sets of Data: Quadratic Model

Suppose we have two sets of similar data on a response Y and a predictor X and we have in mind a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \epsilon \tag{14.2.8}$$

for each set. We want to check if we can use the same fitted model for both sets and, if so, what the fitted coefficients should be. We fit, to both sets of data at one time, the model

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \alpha_0 Z + \alpha_1 XZ + \alpha_{11} X^2 Z + \epsilon, \tag{14.2.9}$$

where Z is a dummy variable with levels 0 for one set of data and 1 for the other. Extra sums of squares tests then enable us to check the various possibilities, as follows, for example:

1. $H_0: \alpha_0 = \alpha_1 = \alpha_{11} = 0$ versus H_1: not so. If this null hypothesis is rejected, we conclude the models are not the same; if not rejected, we take them to be the same.

2. If H_0 in (1) is rejected, we would look at subsets of the α's. For example, we could test $H_0: \alpha_1 = \alpha_{11} = 0$ versus H_1: not so. If H_0 were not rejected, we would

conclude that the two sets of data exhibited only a difference in response levels but had the same slope and curvature.

3. If H_0 in (2) is rejected, we could test $H_0: \alpha_{11} = 0$ versus $H_1: \alpha_{11} \neq 0$ to see if the models differed only in zero and first-order terms, indicated by nonrejection of H_0.

Other sequences of tests could be used if the were sensible in the context of the problem under study. The sequence chosen above is a natural hierarchical buildup that is often sensible.

General Case: *r* Sets, Linear Model

In principle, there is no difficulty extending this setup to situations with more sets of data, and other models involving more predictors, $X_1, X_2, \ldots, X_k$. If there were r sets of data, we would specify, in addition to $Z_0 = 1$, $(r - 1)$ dummy variables $Z_1, Z_2, \ldots, Z_{r-1}$ with levels given by writing down an $\mathbf{I}_{r-1}$ matrix with a line of $(r - 1)$ zeros below it. The rows then designate the groups, and the columns the dummies. If the basic model were to be

$$Y = f(\mathbf{X}, \boldsymbol{\beta}) + \epsilon$$

for one set of data, we would fit, to all the data, the model

$$Y = f(\mathbf{X}, \boldsymbol{\beta}) + \sum_{j=1}^{r-1} Z_j f(\mathbf{X}, \boldsymbol{\alpha}_j) + \epsilon, \tag{14.2.10}$$

where $\boldsymbol{\alpha}_j$ is a vector of parameters of the same size as $\boldsymbol{\beta}$, as in the $r = 2$ example above. This can also be written as

$$Y = \sum_{j=0}^{r-1} Z_j f(\mathbf{X}, \boldsymbol{\alpha}_j), \tag{14.2.11}$$

where $\boldsymbol{\alpha}_0$ replaces $\boldsymbol{\beta}$.

We would get the same answers if we fitted the various sets of data individually. For example, if $\mathbf{X}_i$ is the "$\mathbf{X}$ matrix" for the ith set of data and we have two sets, then the model is

$$E(\mathbf{Y}) = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix}, \tag{14.2.12}$$

which we can regard as

$$E(\mathbf{Y}) = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} + \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\theta} \end{bmatrix}, \tag{14.2.13}$$

say. The advantage of using interaction terms with dummy variables instead is that it enables a single formulation and allows appropriate extra sums of squares tests to be set up in a straightforward way, once the hierarchical aspects of the various submodels have been recognized.

14.3. DUMMY VARIABLES FOR SEGMENTED MODELS

It is sometimes appropriate to fit a model relating a response Y to a single predictor X in terms of segments. (Often, this X is a time variable, or a space variable.) For

TABLE 14.3. Parity Price (¢) per Pound of Live Weight of Chickens

Date	Parity Price, Y	X	or	X'
Jan. 1955	29.1	1		−10
May 1955	29.0	2		−9
Sept. 1955	28.6	3		−8
Jan. 1956	28.1	4		−7
May 1956	28.6	5		−6
Sept. 1956	28.7	6		−5
Jan. 1957	28.2	7		−4
May 1957	28.6	8		−3
Sept. 1957	28.6	9		−2
Jan. 1958	28.1	10		−1
May 1958	28.7	11		0
Sept. 1958	28.6	12		1
Jan. 1959	26.9	13		2
May 1959	27.0	14		3
Sept. 1959	26.8	15		4
Jan. 1960	25.7	16		5
May 1960	25.9	17		6
Sept. 1960	25.6	18		7
Jan. 1961	25.1	19		8
May 1961	25.2	20		9
Sept. 1961	25.1	21		10

example, we might believe that, up to a certain point (rarely known, usually estimated) on the X-scale, a straight line fit is appropriate but, beyond that point, a *different* straight line is needed. Or, we might believe that the initial straight line was followed by a quadratic curve. A variation on the latter case could be that, where the straight line and the quadratic curve intersect, the slope of the model did not change. There could also be several segments rather than two. In addition, there could be other predictor variables, the effects of which were *not* segmented. The main topic of this section will be that of a pair of segmented straight lines in one X with no other predictor variables. This should enable the case of several segments to be tackled using similar ideas. Moreover, adding other predictors typically poses no problems. One simply adds these to the model and **X** matrix in the usual way and fits all terms simultaneously by least squares. We can fit segmented models by using suitably defined dummy variables. As always, there is an infinite choice of ways to select the dummy variable levels. We edge into this topic by first thinking about a single straight line segment in this context.

One Segment

Example. Table 14.3 shows data on the parity price in cents per pound of live weights of chickens at equal intervals of time. Two alternative dummy variables are shown as columns X and X'. Either will do, although the centered column X' may be preferred since it is orthogonal to the column of 1's in the **X** matrix. The appropriate models in the two cases are

$$Y = \beta_0 + \beta_1 X + (\text{other terms with predictor variables}) + \epsilon \qquad (14.3.1)$$

and (since $X_i' = X_i - \overline{X}$, $i = 1, 2, \ldots, n$)

$$\begin{aligned} Y &= (\beta_0 + \beta_1\overline{X}) + \beta_1 X' + (\text{other terms}) + \epsilon \\ &= \beta_0' + \beta_1 X' + (\text{other terms}) + \epsilon. \end{aligned} \tag{14.3.2}$$

Note: Here, since $n = 21$ is odd, the quantities $X_i' = X_i - \overline{X}$ are all integers. When n is even we can use instead $X_i' = 2(X_i - \overline{X})$ to avoid fractions. For example,

$$\begin{array}{rcccc} X_i = & 1 & 2 & 3 & 4 \quad (\overline{X} = 2\tfrac{1}{2}) \\ X_i - \overline{X} = & -1\tfrac{1}{2} & -\tfrac{1}{2} & \tfrac{1}{2} & 1\tfrac{1}{2} \\ 2(X_i - \overline{X}) = & -3 & -1 & 1 & 3 \end{array}$$

(This is a simple orthogonal polynomial. See Chapter 22.)

For a quadratic model, terms $\beta_0 + \beta_1 X + \beta_{11}X^2$ (or $\beta_0' + \beta_1 X' + \beta_{11}X'^2$) would be added, or the trend could be expressed through the first- and second-order orthogonal polynomials described in Chapter 22. Higher-order time trends would be handled in similar fashion with higher-order terms.

The data in Table 14.3 are equally spaced in time, and so the X's are chosen there as equally spaced. If the data were unequally spaced in time, the X's would be chosen accordingly. For example, if the dates were January 1955, February 1955, April 1955, June 1955, . . . , the X's would be 1, 2, 4, 6, . . . , and so on. In such a case, use of the column $X - \overline{X}$ might be inconvenient because it might lead to noninteger values. In that case, either one would use a multiplier to convert the numbers to integers, or, if this were not reasonable, one might use values $X_i - A$, where A was some convenient integer close to $\overline{X}$; however, there is usually little if any advantage to be gained by using $X_i - A$, except that the sizes of the numbers used are reduced somewhat. Because of the unequal spacings, special orthogonal polynomials would have to be evaluated if that kind of approach were desired; consequently, orthogonal polynomials are hardly ever used for unequally spaced data.

Two Segments

When there are two straight line segments, a dummy variable must be set up for each. The problem divides up into two main levels of complication:

1. When it is known which data points lie on which segments, with subcases:
   a. when the abscissa of intersection of the two lines can be assumed to be at a specific value at which one or more observations exists; and
   b. when the abscissa of intersection of the two lines is unknown.
2. When it is not known which data points lie on which segments.

Case 1: When It Is Known Which Points Lie on Which Segments

Example 1a. The equally spaced data plotted in Figure 14.4 fall into case (1a). We suppose that it is known that the first five data points lie (apart from random error) on the first line and that the last five data points lie (apart from random error) on the second line; the fifth point is thus common to both lines. We can (for example) set up two dummy variables X_1 and X_2 for the two lines as follows. Both are set to zero at the known intersection point, namely, the fifth observation, X_1 is stepped back for the first line, X_2 is stepped forward for the second line, and both variables are

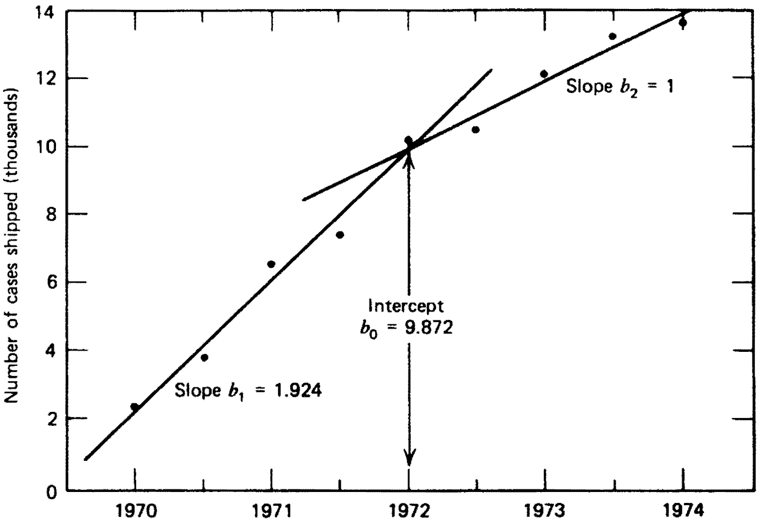

Figure 14.4. Use of dummy variables; two lines, abscissa of point of intersection known.

otherwise zero. (The steps are equally spaced here because the data are equally spaced. If they were not, other appropriate step levels would be chosen instead.) The resulting data matrix, assuming no other predictor variables are involved, is shown in Table 14.4. If we now fit the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \tag{14.3.3}$$

the estimates obtained have these roles:

b_0 = value of $\hat{Y}$ at point of intersection, $X_1 = X_2 = 0$,

b_1 = slope of first trend line,

b_2 = slope of second trend line.

For the data shown, the normal equations become

$$\begin{aligned} 9b_0 - 10b_1 + 10b_2 &= 79.6, \\ -10b_0 + 30b_1 &= -41.0, \\ 10b_0 + 30b_2 &= 128.7, \end{aligned} \tag{14.3.4}$$

TABLE 14.4. Dummy Variables for Example of Two Straight Lines Whose Abscissa of Intersection Is Known

Observation Number	Date	X_0	X_1	X_2	Y
1	1970	1	−4	0	2.3
2		1	−3	0	3.8
3	1971	1	−2	0	6.5
4		1	−1	0	7.4
5	1972	1	0	0	10.2
6		1	0	1	10.5
7	1973	1	0	2	12.1
8		1	0	3	13.2
9	1974	1	0	4	13.6

TABLE 14.5. Alternative Dummy Variable Setup for Example of Two Straight Lines Whose Point of Intersection Is Known

Observation Number	Date	X_0	X_1	X_2	Y
1	1970	1	1	0	2.3
2		1	2	0	3.8
3	1971	1	3	0	6.5
4		1	4	0	7.4
5	1972	1	5	0	10.2
6		1	5	1	10.5
7	1973	1	5	2	12.1
8		1	5	3	13.2
9	1974	1	5	4	13.6

with solutions $b_0 = 9.871$, $b_1 = 1.924$, and $b_2 = 1.000$ as drawn in Figure 14.4. If other predictor variables were involved in the problem, appropriate terms would be added on the right-hand side of Eq. (14.3.3).

As with all dummy variable situations, the representation is not unique. For example, an alternative setup is shown in Table 14.5, in which (new X_1) = (previous X_1) + 5. This setup will provide estimates of the slopes as before but the constant term b_0, the value of $\hat{Y}$ when $X_1 = X_2 = 0$, will now be the intercept of the first line at the abscissa $1969\frac{1}{2}$.

Straight Line and Quadratic Curve

Higher-order models would be accommodated by adding higher-order terms. For a straight line followed by a quadratic curve, for example, we would fit

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{22} X_2^2 + \epsilon. \tag{14.3.5}$$

If, in addition, we wanted the model to have continuous derivative at the join point, we would require $\partial Y/\partial X_1 = \partial Y/\partial X_2$ at the point of intersection, that is, at the fifth data point where $X_2 = 0$,

$$\beta_1 = \beta_2 + 2\beta_{22} X_2. \tag{14.3.6}$$

Thus we would set $\beta_1 = \beta_2 = \beta$, say, in Eq. (14.3.5) and fit the reduced model $Y = \beta_0 + \beta(X_1 + X_2) + \beta_{22} X_2^2 + \epsilon$.

Example 1b. The equally spaced data in Figure 14.5 fall into case (1b). We assume that it is known that the first four data points lie (apart from random error) on one line and that the last five data points lie (apart from random error) on a second line. However, the point of intersection is unknown. A third dummy variable X_3 is needed to take care of the unknown point of intersection. It is set to zero for all points on the first line and then goes to 1 for all points on the second line to allow for a jump (positive or negative) from the first line to the second. The dummies X_1 and X_2 are chosen in either of the ways indicated in Example 1a. Table 14.6 shows an appropriate data matrix. If no other predictor variables are involved we can fit the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon. \tag{14.3.7}$$

The parameter β_3 is the step change that comes into effect at the fifth observation

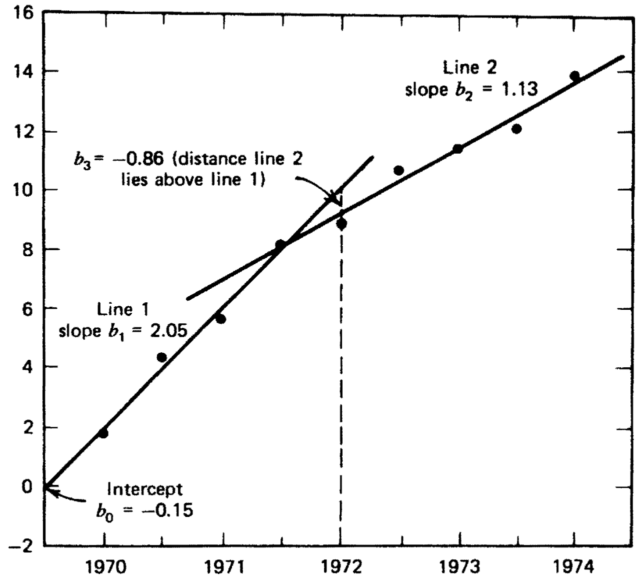

Figure 14.5. Use of dummy variables; two lines, abscissa of point of intersection unknown.

point and is the vertical distance the second line lies *above* the first at this point. (If the second line lies below the first, β_3 is negative.) For the data of Table 14.6, the normal equations are

$$\begin{aligned} 9b_0 + 35b_1 + 10b_2 + 5b_3 &= 77.4, \\ 35b_0 + 155b_1 + 50b_2 + 25b_3 &= 347.5, \\ 10b_0 + 50b_1 + 30b_2 + 10b_3 &= 126.3, \\ 5b_0 + 25b_1 + 10b_2 + 5b_3 &= 57.5, \end{aligned} \qquad (14.3.8)$$

with solutions

$b_0 = -0.15$ intercept of line 1, when $X_1 = 0$,

$b_1 = 2.05$ (slope of line 1),

$b_2 = 1.13$ (slope of line 2),

$b_3 = -0.86$ (the vertical distance between line 2 and line 1 at the fifth observation point).

The situation is shown graphically in Figure 14.5. The negative sign of b_3 and the fact that $b_1 > b_2$ indicate that the point of intersection of the two lines is to the left of the fifth observation point. In fact, it occurs when $X_1 = 4.065$. This point of intersection can be found by writing both lines in terms of the X_1 scale. The first line is given by

$$\hat{Y} = -0.15 + 2.05X_1, \qquad (14.3.9)$$

and the second by

$$\hat{Y} = -0.15 + 2.05(5) + 1.13X_2 - 0.86, \qquad (14.3.10)$$

that is,

$$\hat{Y} = 9.24 + 1.13X_2. \qquad (14.3.11)$$

TABLE 14.6. Dummy Variables for Example of Two Straight Lines Whose Abscissa of Intersection Is Unknown

Observation Number	Date	X_0	X_1	X_2	X_3	Y
1	1970	1	1	0	0	1.8
2		1	2	0	0	4.3
3	1971	1	3	0	0	5.6
4		1	4	0	0	8.2
5	1972	1	5	0	1	9.1
6		1	5	1	1	10.7
7	1973	1	5	2	1	11.5
8		1	5	3	1	12.2
9	1974	1	5	4	1	14.0

Looking at the X_1 and X_2 scales in relation to the scale on Figure 14.5, we see that $X_2 = 0$ at $X_1 = 5$ so that we can substitute $X_2 = X_1 - 5$ into the equation of the second line to reduce it to

$$\hat{Y} = 3.59 + 1.13X_1. \qquad (14.3.12)$$

Setting the two right-hand sides of Eqs. (14.3.9) and (14.3.12) equal gives $X_1 = 4.065$ as the point of intersection. These calculations could also be made using X_2 to give $X_2 = -0.935$.

Case 2: When It Is Not Known Which Points Lie on Which Segments

In the previous cases, we estimated the parameters of the composite model by fitting a given model by linear least squares. In the present case, the solution could be obtained by looking at every possible division of the points to the first and second lines, estimating the parameters by linear least squares and evaluating the residual sum of squares for each division, and then picking that division and set of estimates that give rise to the smallest of all the residual sums of squares. (In practice, one would usually not need to look at every division of the points, because a few computations usually show the "ballpark" in which the best division lies and the computations are then confined to those divisions only.) Alternatively, the problem could be cast into the form of a nonlinear estimation problem and solved by the methods discussed in Chapter 24. (Local minima can occur, so some care is needed at times.) Using nonlinear methods is complicated and usually not worthwhile, however.

Remark

In the above illustrations we have assumed that the points belonging to the two segments did not overlap. This would most often be the case. Where overlap occurs, the situation is more complex, but the methods remain valid. Inspection of a plot of the data is usually helpful in determining how to divide the points. When other predictors are present, these can be regressed out first, and the residuals plotted against the variable to be segmented.

EXERCISES FOR CHAPTER 14

A. A response variable is measured on the following dates: November 17, 19, 20, 22, 26, 29, 30, December 1, 2, 3, and 5. It is believed that the response depends on two factors X_1 and

TABLE B. Yields Y from Three Tiller Types for Two Varieties of Wheat Subjected to Different Nitrogen Applications

Y	Tiller Subscript	Waldron/Ciano	N Rate	N per Tiller at Tillering, mg (X)
370	1	W	0	4.2
659	1	W	50	7.2
935	1	W	270	9.8
390	1	C	0	3.6
753	1	C	50	7.6
733	1	C	270	10.3
182	2	W	0	3.1
417	2	W	50	5.2
686	2	W	270	7.8
188	2	C	0	2.8
632	2	C	50	6.0
538	2	C	270	7.7
27	3	W	0	2.0
141	3	W	50	2.8
262	3	W	270	3.6
34	3	C	0	2.7
222	3	C	50	3.1
242	3	C	270	4.4

X_2 whose values have been recorded (but are not given here) and that, in addition, there is a quadratic time trend in the data. How would you take account of it? (If you mention any new variables, you should specify their actual levels.)

B. (*Source:* "Tiller development and yield of standard and semidwarf spring wheat varieties as affected by nitrogen fertilizer," by J. F. Power and J. Alessi, *Journal of Agricultural Science, Cambridge*, **90,** 1978, 97–108. Adapted with the permission of Cambridge University Press.) Table B shows, as Y, grain yields in kg/ha (kilograms per hectare), which resulted from growing two varieties, Waldron and Ciano, of hard red spring wheat, at three different nitrogen rates $N = 0, 50, 270$, representing deficient, adequate, and excessive nitrogen supply. The tiller subscripts refer to tillers (ear-bearing growths from the leaf axils of main stems) developed from the first, second, and third or later leaves of the main stem. (See p. 98 of the source reference.) The last column shows nitrogen content in milligrams (mg) per tiller at tillering, X.

Fit a quadratic model in X to the response Y with additional terms added to the model for tiller subscript, wheat variety, and N rate. Examine your results and provide conclusions.

C. Bars of soap are scored for their appearance in a manufacturing operation. These scores are on a 1–10 scale, and the higher the score the better. The difference between operator performance and the speed of the manufacturing line is believed to measurably affect the quality of the appearance. The following data were collected on this problem:

Operator	Line Speed	Appearance (Sum for 30 Bars)
1	150	255
1	175	246
1	200	249
2	150	260
2	175	223
2	200	231
3	150	265
3	175	247
3	200	256

Requirements

1. Using dummy variables, fit a multiple regression model to these data.
2. Using $\alpha = 0.05$, determine whether operator differences are important in bar appearance. Using the regression model, demonstrate that the average appearance score for operator No. 1 is 250, operator No. 2 is 238, and operator No. 3 is 256.
3. Does line speed affect appearance? (Use $\alpha = 0.05$.)
4. What model would you use to predict bar appearance?

D. An experimenter suggests the following dummy variable scheme to separate possible level differences among six groups. Is it a workable one?

Z_0	Z_1	Z_2	Z_3	Z_4	Z_5
1	1	−1	−1	−1	−1
1	−1	2	−1	−1	−1
1	−1	−1	3	−1	−1
1	−1	−1	−1	4	−1
1	−1	−1	−1	−1	5
1	−1	−1	−1	−1	−1

E. Here is another six-group scheme. Will it work?

Z_1	Z_2	Z_3	Z_4	Z_5
1	1	1	1	1
2	3	3	2	1
3	2	3	3	2
3	3	2	3	3
2	3	3	2	3
1	2	3	3	1

F. 1. An experimenter suggests the following dummy variable scheme to separate possible level differences among six groups. If $u = v = 0$, is it a workable one?

Z_1	Z_2	Z_3	Z_4	Z_5
−1	−1	−1	1	2
1	−1	−1	1	−1
−1	1	−1	1	u
1	1	−1	−1	v
−1	−1	1	−1	−1
1	−1	1	−1	2

2. If you said it *is* workable, are there any nonzero values of u and v that make it unworkable?
3. If you said it is *not* workable, are there any nonzero values of u and v that would make it work?

G. An experimenter says he feels the need to fit two straight lines to ten equally spaced points, the first five of which he believes are on one line, and the second five on another line. He proposes to use dummy system A, below. The statistician on the project suggests system B. Who is right?

System A				System B			
X_0	X_1	X_2	X_3	X_0	X_1	X_2	X_3
1	1	0	−1	1	1	0	0
1	2	0	−1	1	2	0	0
1	3	0	−1	1	3	0	0
1	4	0	−1	1	4	0	0
1	5	0	−1	1	5	0	0
1	0	0	0	1	5	1	1
1	0	1	0	1	5	2	1
1	0	2	0	1	5	3	1
1	0	3	0	1	5	4	1
1	0	4	0	1	5	5	1

H. An experimenter seeks your help in fitting a "two-piece" relationship to seven observations equally spaced in time. He wants to fit a quadratic followed by a straight line. Four data points lie, apart from error, on the quadratic, three on the line. He does not know where the curve and line will intersect, and different slopes are allowable at the point of intersection. Set up his **X** matrix for him, and write down the model he must fit.

I. The following nine values of Y: 1, 4, 6, 7, 9.5, 11, 11.5, 13, 13.5, are observed at successive equally spaced time intervals. Suppose that one straight line is a suitable model for the first four observations and that a second straight line is suitable to represent the last five observations. Estimate the slopes of the two lines and find where the lines intersect. Is the fit satisfactory?

J. (*Source*: "Nutrition of infants and preschool children in the north central region of the United States of America," by E. S. Eppright, H. M. Fox, B. A. Fryer, G. H. Lamkin, V. M. Vivian, and E. S. Fuller, *World Review of Nutrition and Dietetics*, **14**, 1972, 269–332.) Below are given 72 observations of the response Y = boy's weight/height ratio (W/H), for equally spaced values of the corresponding predictor variable X = age in months. Assume that the observations fall into two groups: (1) the first seven observations and (2) the remaining 65 observations. Assume, further, that the two groups of data can be explained by two straight line time trends. Using the methods of Section 14.3, find the slopes of the two trend lines and their point of intersection. Plot the data and the fitted lines. Also provide an analysis of variance table, and find and analyze the residuals.

W/H	Age	W/H	Age	W/H	Age
0.46	0.5	0.88	24.5	0.92	48.5
0.47	1.5	0.81	25.5	0.96	49.5
0.56	2.5	0.83	26.5	0.92	50.5
0.61	3.5	0.82	27.5	0.91	51.5
0.61	4.5	0.82	28.5	0.95	52.5
0.67	5.5	0.86	29.5	0.93	53.5
0.68	6.5	0.82	30.5	0.93	54.5
0.78	7.5	0.85	31.5	0.98	55.5
0.69	8.5	0.88	32.5	0.95	56.5
0.74	9.5	0.86	33.5	0.97	57.5
0.77	10.5	0.91	34.5	0.97	58.5
0.78	11.5	0.87	35.5	0.96	59.5
0.75	12.5	0.87	36.5	0.97	60.5
0.80	13.5	0.87	37.5	0.94	61.5
0.78	14.5	0.85	38.5	0.96	62.5
0.82	15.5	0.90	39.5	1.03	63.5
0.77	16.5	0.87	40.5	0.99	64.5
0.80	17.5	0.91	41.5	1.01	65.5

W/H	Age	W/H	Age	W/H	Age
0.81	18.5	0.90	42.5	0.99	66.5
0.78	19.5	0.93	43.5	0.99	67.5
0.87	20.5	0.89	44.5	0.97	68.5
0.80	21.5	0.89	45.5	1.01	69.5
0.83	22.5	0.92	46.5	0.99	70.5
0.81	23.5	0.89	47.5	1.04	71.5

K. (This is a shorter version of the foregoing exercise.) Make use of only the first 32 observations. Assume that these 32 observations fall into two groups: (1) the first seven observations and (2) the remaining 25 observations. Assume, further, that the two groups of data can be explained by two straight line time trends. Using the methods of Section 14.3, find the slopes of the two trend lines and their point of intersection. Plot the data and the fitted lines. Also, provide an analysis of variance table, and find and analyze the residuals.

L. "Look at these data," a friend moans. "I don't know whether to fit two straight lines, one straight line, or what." You look at his notes and see that he has two sets of (X, Y) data, given below, which both cover the same X-range. How do you resolve his dilemma? Describe, and give model details, and "things he needs to do."

Set A: X	Y	Set B: X	Y
8	5.3	9	5.1
0	0.9	7	4.4
12	7.1	8	5.2
2	2.4	6	3.8

M. An experimenter has two sets of data, of (X, Y) type, and wishes to fit a quadratic equation to each set. She also wishes (later) to test if the two quadratic fits might be identical in "location" and "curvature" but have different intercept values. Explain how you would set this up for her.

N. You have two sets of data involving values of X and Y, but you are unsure whether to fit the data separately or together. You consider and fit the six-parameter model

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + Z(\alpha_0 + \alpha_1 X + \alpha_{11} X^2) + \epsilon,$$

where Z is a dummy variable whose value is -1 for "set A" and 1 for "set B."

1. What hypothesis would you test to answer the question: "Will a single quadratic model fit all the data?"
2. What hypothesis would you test to answer the question: "Will a single straight line model fit all the data?"
3. How would you obtain *separate* quadratic fits to the two data sets?
4. If a data point in set A and a data point in set B had the same X-value, would those two points be "repeat points" in the fit of the full model written out above?

O. Data on a response Y and k predictor variables $X_1, X_2, \ldots, X_k$ arise from two factories A and B. It is desired to fit a model of the form

$$\hat{Y}_Q = b_{0Q} + b_{1Q} X_1 + b_2 X_2 + \cdots + b_k X_k,$$

where $Q = A$ or B denotes the factory in which the prediction will be made. In other words, the effects of $X_2, \ldots, X_k$ are the same in both factories, but the intercept, and the slope with respect to X_1, are different for each factory. Show that this problem can be handled by using one new dummy variable Z, which takes the value 1 for A, and 0 for B, and then fitting the model

$$Y = \beta_0 + \beta_z Z + \beta_{1z} X_1 Z + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

and that, in this case,

$$b_{0A} = b_0 + b_z, \qquad b_{1A} = \mathrm{b}_1 + \mathrm{b}_{1z},$$
$$b_{0B} = b_0, \qquad b_{1B} = b_1.$$

The same idea can be extended to other Xs by adding other $\beta_{jz}ZX_j$ cross-product terms. Show that, if the idea is extended to *all* the X's, we are essentially fitting separate models to the A data and to the B data.

P. (*Source*: "Application of the principle of least squares to families of sraight lines," by S. Ergun, *Industrial and Engineering Chemistry*, **48**, November 1956, 2063–2068.)

1. Show that the least squares estimates $a_1, a_2, \ldots, a_m, b$, of the parameters $\alpha_1, \alpha_2, \ldots, \alpha_m, \beta$ in the family of straight lines

$$E(Y_i) = \alpha_i + \beta X_i, \qquad i = 1, 2, \ldots, m,$$

are given by

$$b = \frac{\sum_{i=1}^{m} \sum_{u=1}^{n_i} (X_{iu} - \overline{X}_i)(Y_{iu} - \overline{Y}_i)}{\sum_{i=1}^{m} \sum_{u=1}^{n_i} (X_{iu} - \overline{X}_i)^2},$$
$$a_i = \overline{Y}_i - b\overline{X}_i,$$

where

$$X_{i1}, X_{i2}, \ldots, X_{iu}, \ldots, X_{in_i},$$
$$Y_{i1}, Y_{i2}, \ldots, Y_{iu}, \ldots, Y_{in_i}$$

denote the observed values of X_i and Y_i, which relate to the ith line, $i = 1, 2, \ldots, m$. Show also that the residual sum of squares is

$$S^2 = \sum_{i=1}^{m} \sum_{u=1}^{n_i} (Y_{iu} - \overline{Y}_i)^2 - b^2 \sum_{i=1}^{m} \sum_{u=1}^{n_i} (X_{iu} - \overline{X}_i)^2$$

with $(\sum_{i=1}^{m} n_i - m - 1)$ degrees of freedom, that

$$\sigma_b^2 = \frac{\sigma^2}{\sum_{i=1}^{m} \sum_{u=1}^{n_i} (X_{iu} - \overline{X}_i)^2},$$

and that

$$\sigma_{a_i}^2 = \frac{\sigma^2}{n_i}\left\{1 + \frac{n_i \overline{X}_i^2}{\sum_{i=1}^{m} \sum_{u=1}^{n_i} (X_{iu} - \overline{X}_i)^2}\right\}.$$

2. Show that the least squares estimates $a, b_1, b_2, \ldots, b_m$, of the parameters $\alpha, \beta_1, \beta_2, \ldots, \beta_m$ in the family of straight lines

$$E(Y_i) = \alpha + \beta_i X_i, \qquad i = 1, 2, \ldots, m,$$

are given by

$$a = \frac{\sum_{i=1}^{m} n_i(\overline{Y}_i - \overline{X}_i\{\sum_{u=1}^{n_i} X_{iu} Y_{iu} / \sum_{u=1}^{n_i} X_{iu}^2\})}{\sum_{i=1}^{m} n_i(1 - n_i \overline{X}_i^2 / \sum_{u=1}^{n_i} X_{iu}^2)},$$

$$b_i = \frac{\{\sum_{u=1}^{n_i} X_{iu} Y_{iu} - a \sum_{u=1}^{n_i} X_{iu}\}}{\sum_{u=1}^{n_i} X_{iu}^2},$$

where X_{iu}, Y_{iu} are as above. Also show that the residual sum of squares is

$$S^2 = \sum_{i=1}^{m}\sum_{u=1}^{n_i} Y_{iu}^2 - \sum_{i=1}^{m} b_i^2 \sum_{u=1}^{n_i} X_{iu}^2 + a^2 \sum_{i=1}^{m} n_i - 2a \sum_{i=1}^{m} n_i \overline{Y}_i$$

with $(\Sigma_{i=1}^{m} n_i - m - 1)$ degrees of freedom, that

$$\sigma_a^2 = \sigma^2 / \Sigma_{i=1}^{m} n_i(1 - n_i \overline{X}_i^2 / \Sigma_{u=1}^{n_i} X_{iu}^2),$$

and that

$$\sigma_{b_i}^2 = \left\{ \frac{1}{\Sigma_{u=1}^{n_i} X_{iu}^2} + \frac{(\Sigma_{u=1}^{n_i} X_{iu})^2/(\Sigma_{u=1}^{n_i} X_{iu}^2)^2}{\Sigma_{i=1}^{m} n_i(1 - n_i \overline{X}_i^2/\Sigma_{u=1}^{n_i} X_{iu}^2)} \right\} \sigma^2.$$

Q. A test for the equality of the slopes β_i of m lines represented by the first-order models

$$Y_{iu} - \overline{Y}_i = \beta_i(X_{iu} - \overline{X}_i) + \epsilon_{iu}, \qquad i = 1, 2, \ldots, m,$$

can be conducted as follows. Suppose that

$$\begin{aligned} &X_{i1}, X_{i2}, \ldots, X_{iu}, \ldots, X_{in_i} \quad \text{(fixed)} \\ &Y_{i1}, Y_{i2}, \ldots, Y_{iu}, \ldots, Y_{in_i} \quad (\epsilon_{iu} \sim N(0, \sigma^2), \text{independent}) \end{aligned}$$

are available for estimation of the parameters of the ith line. The least squares estimate of β_i is

$$b_i = \left\{ \frac{\Sigma_{u=1}^{n_i} (X_{iu} - \overline{X}_i)(Y_{iu} - \overline{Y}_i)}{\Sigma_{u=1}^{n_i} (X_{iu} - \overline{X}_i)^2} \right\}$$

with sum of squares (1 df)

$$\mathrm{SS}(b_i) = b_i^2 \left\{ \sum_{u=1}^{n_i} (X_{iu} - \overline{X}_i)^2 \right\}$$

and residual sum of squares ($n_i - 2$ df)

$$S_i = \sum_{u=1}^{n_i} (Y_{iu} - \overline{Y}_i)^2 - \mathrm{SS}(b_i).$$

If we assume $\beta_i = \beta$, all i, then the least squares estimate of β is

$$b = \left\{ \frac{\Sigma_{i=1}^{m} \Sigma_{u=1}^{n_i} (X_{iu} - \overline{X}_i)(Y_{iu} - \overline{Y}_i)}{\Sigma_{i=1}^{m} \Sigma_{u=1}^{n_i} (X_{iu} - \overline{X}_i)^2} \right\}$$

with sum of squares (1 df),

$$\mathrm{SS}(b) = b^2 \left\{ \sum_{i=1}^{m} \sum_{u=1}^{n_i} (X_{iu} - \overline{X}_i)^2 \right\}$$

and residual sum of squares ($\Sigma\, n_i - 2m$ df)

$$S = \sum_{i=1}^{m} \sum_{u=1}^{n_i} (Y_{iu} - \overline{Y}_i)^2 - \mathrm{SS}(b).$$

We can form an analysis of variance table as follows:

ANOVA

Source	df	SS	MS	F
b	1	SS(b)	M_1	$F_1 = \dfrac{M_1}{s^2}$
All $b_i \vert b$	$m - 1$	$\sum_{i=1}^{m} \text{SS}(b_i) - \text{SS}(b)$	M_2	$F_2 = \dfrac{M_2}{s^2}$
Residual	$\sum_{i=1}^{m} n_i - 2m$	by subtraction	s^2 (estimates σ^2 if first-order models are correct)	
Total	$\sum_{i=1}^{m} n_i - m$	$\sum_{i=1}^{m} \sum_{u=1}^{n_i} (Y_{iu} - \bar{Y}_i)^2$		

The hypothesis $H_0: \beta_i = \beta$ is tested by comparing F_2 with an appropriate percentage point of the $F\{(m - 1), (\sum_{i=1}^{m} n_i - 2m)\}$ distribution. If H_0 is not rejected, b is used as the common slope of the lines. (This is a special case of testing a linear hypothesis. A test for the equality of intercepts of two lines can also be constructed.) F_1 is used to test $H_0: \beta = 0$.

Apply the above procedure to the data below.

u	X_1	Y_1	X_2	Y_2	X_3	Y_3
1	3.5	24	3.2	22	3.0	32
2	4.1	32	3.9	33	4.0	36
3	4.4	37	4.9	39	5.0	47
4	5.0	40	6.1	44	6.0	49
5	5.5	43	7.0	53	6.5	55
6	6.1	51	8.1	57	7.0	59
7	6.6	62			7.3	64
8					7.4	64

R. The table shows data from a test to compare two types of outdoor boots coded A and B. Six subjects wore pairs of each type of the boot on each of four different occasions; thus there were eight tests per subject and 48 tests in all. For each test, the subject was placed in a room held at the ambient temperature shown and the drop in temperature in degrees Fahrenheit after 90 minutes at one (randomly selected) little toe was recorded. (The rest of the clothing worn was the same for each subject.) Analyze the data and answer the question: Which boot is better and is it significantly so?

	Boot A				Boot B			
	Ambient Temperature (°F)				Ambient Temperature (°F)			
Subject Number	20	0	−5	−22	20	0	−5	−22
1	4.5	10.3	8.4	12.6	8.3	9.0	9.9	8.6
2	2.1	7.6	7.3	13.6	11.6	10.6	11.2	17.0
3	7.9	11.9	11.9	14.5	11.3	12.0	15.1	16.0
4	7.9	15.2	10.8	16.5	5.9	18.2	15.3	9.0
5	5.0	5.9	14.1	12.3	8.6	11.3	15.6	16.1
6	5.3	10.2	12.8	10.2	6.0	10.1	13.3	13.0

S. (*Source*: "Productivity of field-grown soybeans exposed to simulated acidic rain," by L. S. Evans, K. F. Lewin, M. J. Patti, and E. A. Cunningham, *New Phytologist*, **93**, 1983, 377–388.) The data shown in the table result from some 1981 experiments performed to determine the effects of simulated acidic rain on soybean yields. The lower pH levels denote "more acidic" simulated rain. Some plants were shielded from ambient rainfall, and some were not, as indicated. The response values shown are y = seed mass per plant in grams.

Plants Were Shielded from Ambient Rainfall?	x = pH Level of Simulated Rainfall			
	2.7	3.3	4.1	5.6
Yes	10.1	10.9	11.7	13.1
No	10.55	10.62	11.11	11.42

Fit the straight line model $Y = \beta_0 + \beta_1 X + \epsilon$ to each set of data. Test H_0: "the slopes of the two lines are equal" versus H_1: "not so." What models would you adopt for the two sets of data and why?

T. (*Source*: "Effects of population density on sex expression in *Onoclea sensibilis* L. on agar and ashed soil," by G. Rubin, D. S. Robson, and D. J. Paolillo, Jr., *Annals of Botany*, **55**, 1985, 205–215.) The 1983 data in the accompanying table have been adapted from Figures 1 and 2 of the source reference. The symbol X denotes time in days and Y denotes either proportion female attained at that time when gametophytes of the fern *Onoclea sensibilis* L. are cultured on agar soil (denoted by f), or proportion male attained at that time when cultured on ashed soil (denoted by m). The data of the sex opposite to that indicated were sparse and are not given. The Y data have already been translated to probit values (a transformation not described here). In adapting the data we have (a) ignored all zero Y values in days prior to those indicated, (b) ignored all 1982 data, and (c) read off the data from the graphs to a crude accuracy of two figures. Please see the original source reference for a more complete presentation.

Define a suitable numerical dummy variable to separate the two groups (f and m) of data and investigate whether two parallel straight lines will provide a good fit in either of these two cases:

(*a*) Predictor variable X, response Y

(*b*) Predictor variable $U = \ln X$, response Y.

X	Y	f or m	X	Y	f or m
18	4.8	f	40	3.3	m
20	5.5	f	42	3.8	m
24	5.8	f	45	4.1	m
28	6.0	f	62	5.0	m
30	6.5	f	71	5.5	m
33	6.6	f	75	6.0	m
36	6.7	f			
48	7.0	f			
60	7.3	f			

U. Two types of tomatoes, cherry (six plants) and yellow oval (nine plants), were grown from seeds indoors in Madison, Wisconsin, in 1996 and then replanted at three separate planting times. The numbers of tomatoes produced were recorded as follows:

	Planting Time		
	Early	Middle	Late
Cherry	47, 36	19, 20, 27, 50	—
Yellow oval	39, 50	—	4, 8, 8, 10, 12, 14, 19

Set up one dummy variable to distinguish types of tomatoes and a pair of dummies to separate the planting times, and analyze the data. What do you conclude?