Chapter 10

Multiple Regression—Dummy Variables, Contrasts, and Analysis of Covariance

Any analysis of variance model (for example, anything in Chapters 6, 12, 13, or 14) can be expressed as a regression with dummy variables. The dummy variables are usually based on a set of contrasts. The algebra of individual contrast vectors is discussed in Section 6.9. Many software procedures and functions make explicit use of this form of expression. Here we explore this equivalence of different representations of the contrasts associated with a factor. The notation in Chapter 10 is that used in Sections I.4.2, 9.3, and 9.4.1.

Section 10.1 introduces dummy variables. Section 10.3 looks at the equivalence of different sets of dummy variable codings for factors. Section 13.5 shows how the R and SAS languages express the dummy variable coding schemes. Table 13.18 shows the notation for applying them to describe models with two or more factors.

10.1 Dummy (Indicator) Variables

Dummy variables, also called indicator variables, are a way to incorporate qualitative predictors into a regression model. If we have a qualitative predictor A with a distinct values, we will need $a - 1$ distinct dummy variables to code it. For example, suppose we believe that the gender of the subject may impact the response. We could define $X_{\text{female}} = 1$ if the subject is female and $X_{\text{female}} = 0$ if the subject is male. Then we interpret the estimated regression coefficient $\hat{\beta}_{\text{female}}$ as the estimated average amount by which responses for females exceed responses for males, assuming the values of all other predictors are unchanged. If $\hat{\beta}_{\text{female}} > 0$, then on average females will tend to have a higher response than males; if $\hat{\beta}_{\text{female}} < 0$, then the average male response will exceed the average female response. There are $g = 2$ levels to the classification variable `gender`, hence we defined $g - 1 = 1$ dummy variable to code that information. We pursue this example in Section 10.2.

R.M. Heiberger, B. Holland, *Statistical Analysis and Data Display*,
Springer Texts in Statistics, DOI 10.1007/978-1-4939-2122-5_10

As another example, suppose one of the predictor variables in a model is the nominal variable `ResidenceLocation`, which can take one of $r = 3$ values: `urban`, `suburban`, or `rural`. If a qualitative predictor has r categories, we must assign $r - 1$ dummy variables to represent it adequately. Otherwise, we may be imposing an unwarranted implicit constraint. It would be incorrect to code this with a single numeric variable $X_{\text{RL}} = 0$ for `urban`, 1 for `suburban`, and 2 for `rural`, as that would imply that the difference between average `urban` and `suburban` responses must equal the difference between average `suburban` and `rural` responses, which is probably not justifiable.

One correct coding is to let $X_{\text{RLu}} = 1$ if `urban` and 0 otherwise and let $X_{\text{RLs}} = 1$ if `suburban` and 0 otherwise. Then the coefficient $\hat{\beta}_{\text{RLu}}$ of X_{RLu} is interpreted as the average difference between `urban` and `rural` response, and the coefficient $\hat{\beta}_{\text{RLs}}$ of X_{RLs} is interpreted as the average difference between `suburban` and `rural` response. The difference between the coefficients $\hat{\beta}_{\text{RLu}}$ and $\hat{\beta}_{\text{RLs}}$ is the average difference between the `urban` and `suburban` response. Here we used `rural` as the reference response. The results of the analysis would have been the same had we used either `urban` or `suburban` as the reference response. See Section 10.3 for the justification of this statement. See Exercise 10.3 to apply the justification to this example.

This type of coding is done automatically in R's linear modeling functions (`lm` and `aov` when variables have been defined as factors with the `factor()` function.

The `PROC ANOVA` and `PROC GLM` in SAS require use of the `CLASSES` command within the `PROC` specification. SAS's `PROC REG` requires explicit coding to construct the dummy variables in the `DATA` step.

Any pair of independent linear combinations of X_{RLu} and X_{RLs} would be equally as valid. R gives the user choice with the `contrasts()` and related functions. SAS gives the user choice with the `estimate` and `test` statements on the `PROC ANOVA` and `PROC GLM` commands.

10.2 Example—Height and Weight

10.2.1 Study Objectives

In the fall of 1998, one of us (RMH) collected the height, weight, and age of the 39 students in one of his classes. The data appear in file `data(htwt)`. While this example does give information on the comparative height distributions of men and women, the primary intent then, and now, is to use this example to illustrate how the techniques of statistics give us terminology and notation for discussing ordinary observations.

10.2.2 Data Description

`feet:` height in feet rounded down to an integer

`inches:` inches to be added to the height in feet

`lbs:` weight in pounds

`months:` age in months

`sex:` `m` or `f`

`meters:` height in meters

10.2.3 Data Problems

From the stem-and-leaf in Table 10.1 we see that even in this small dataset, collected with some amount of care, there are data problems. There are 39 observations, yet only 38 made it to the stem-and-leaf and one of those has a missing value. Further investigation of the data file shows that one student reported her height in meters and another didn't indicate sex. For the remaining figures and tables in this chapter we converted meters to inches for the one. For the other we had the good fortune to have access to the sample population at the next class meeting and were able to fill in the missing value (`m` in this case) by checking the data forms directly with the students. We were lucky in this example that the data file was investigated soon enough after collection that the data anomalies could be resolved. That is not always possible. We describe techniques for dealing with missing data in Section 2.4.

We show a splom of the completed data in Figure 10.1. The age range in our class was 18–28 for women and 19–24 for men. There is no visible relation between age and either height or weight. There is a clear difference in height ranges between men and women and a visible, but less strong, difference in weight ranges. We investigate this further by expanding the `lbs ~ ht` panel in Figure 10.2.

Table 10.1 Stem-and-leaf of Heights from class observation. We used this display to detect the two missing values. Note that this is an edited version of the output. We placed the two distributions adjacent to each other and added additional lines to the high end of the female distribution and to the low end of the male distribution to make the two stem-and-leaf displays align correctly.

```
> data(htwt)

> levels(factor(htwt$sex, exclude=NULL))
[1] "f" "m" NA

> any(is.na(htwt$ht))
[1] TRUE

> for (h in tapply(htwt$ht, factor(htwt$sex, exclude=NULL), c))
+   stem(h, scale=1.5)

  The decimal point is at the |

     Female              Male
  58 | 0              58 |
  60 |                60 |
  62 | 00000          62 | 0
  64 | 000000000      64 | 0
  66 | 000008         66 | 000
  68 | 0              68 | 00
  70 |                70 |
  72 |                72 | 000000
  74 |                74 | 00
```

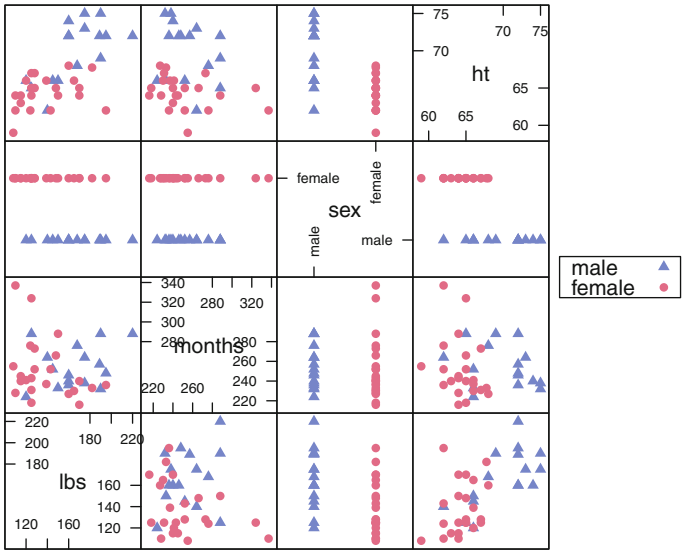

Fig. 10.1 Scatterplot matrix of completed height and weight data from example collected in class.

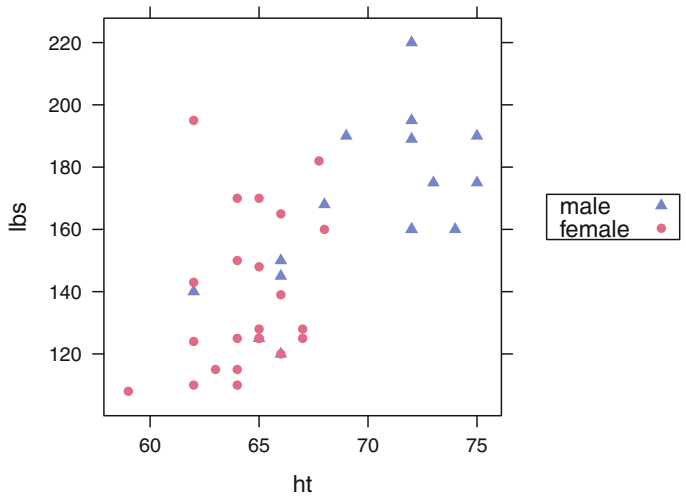

Fig. 10.2 Expansion of `lbs ~ ht` panel of Figure 10.1. There is visibly less overlap in the range for the heights of men and women than for their weights.

Table 10.2 One-way analysis of variance of heights from class observation.

```
> ## one-way analysis of variance
> htwt.aov <- aov(ht ~ sex, data=htwt)

> summary(htwt.aov)
             Df Sum Sq Mean Sq F value  Pr(>F)
sex           1    282   282.3    30.8 2.5e-06 ***
Residuals    37    339     9.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> model.tables(htwt.aov, type="means")
Tables of means
Grand mean

66.71

 sex
       f     m
    64.47 69.94
rep 23.00 16.00
```

10.2.4 Three Variants on the Analysis

Table 10.2 uses the techniques of Chapter 6 to compare the means of two distributions. The specific features that we will look at are the various values in the ANOVA table and the mean heights for each of the groups. We will follow by using regression on two different sets of dummy variables to duplicate those numbers.

We initially use the $g - 1 = 1$ dummy variable X_{female} with the $(1, 0)$ coding scheme suggested above, with value 1 for females and value 0 for males. We display the results of an ordinary linear regression of height on the dummy variable X_{female} in Table 10.3. The estimated intercept $\hat{\beta}_0 = 69.9375$ is the mean height for males. The

Table 10.3 Regression analysis of heights from class observation on the dummy variable coding sex as `female`=1 for female and `female`=0 for male.

```
> ## dummy variable
> htwt$female <- as.numeric(htwt$sex == "f")

> htwt.lm <- lm(ht ~ female, data=htwt)

> summary(htwt.lm, corr=FALSE)

Call:
lm(formula = ht ~ female, data = htwt)

Residuals:
   Min     1Q Median     3Q    Max
-7.938 -2.202  0.533  2.062  5.062

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   69.938      0.757   92.42  < 2e-16 ***
female        -5.470      0.985   -5.55  2.5e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.03 on 37 degrees of freedom
Multiple R-squared:  0.454,Adjusted R-squared:  0.44
F-statistic: 30.8 on 1 and 37 DF,  p-value: 2.54e-06


> anova(htwt.lm)
Analysis of Variance Table

Response: ht
          Df Sum Sq Mean Sq F value  Pr(>F)
female     1    282   282.3    30.8 2.5e-06 ***
Residuals 37    339     9.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

estimated regression coefficient for the X_{female} predictor, $\hat{\beta}_{\text{female}} = -5.4701$, is the increment to the intercept that produces the mean height for females. The ANOVA table in Table 10.3 is identical to the ANOVA table in Table 10.2.

There are many other dummy variable coding schemes that we could use to get exactly the same ANOVA table and the same estimated mean heights for the two groups. We show another in Table 10.4. In this coding, the dummy variable X_{treat} has the value 1 for females and the value -1 for males. The estimated intercept $\hat{\beta}_0 = 67.2024$ is the average of the mean heights for females and males. The estimated regression coefficient for the X_{treat} predictor, $\hat{\beta}_{\text{treat}} = -2.7351$, is the amount that

Table 10.4 Regression analysis of heights from class observation on the dummy variable coding sex as `treat`=1 for female and `treat`=−1 for male.

```
> ## dummy variable
> htwt$treat <- (htwt$sex == "f") - (htwt$sex == "m")

> htwtb.lm <- lm(ht ~ treat, data=htwt)

> summary(htwtb.lm, corr=FALSE)

Call:
lm(formula = ht ~ treat, data = htwt)

Residuals:
   Min     1Q Median     3Q    Max
-7.938 -2.202  0.533  2.062  5.062

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   67.202      0.493  136.40  < 2e-16 ***
treat         -2.735      0.493   -5.55  2.5e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.03 on 37 degrees of freedom
Multiple R-squared:  0.454,Adjusted R-squared:  0.44
F-statistic: 30.8 on 1 and 37 DF,  p-value: 2.54e-06


> anova(htwtb.lm)
Analysis of Variance Table

Response: ht
          Df Sum Sq Mean Sq F value  Pr(>F)
treat      1    282   282.3    30.8 2.5e-06 ***
Residuals 37    339     9.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

added to the intercept produces the mean height for females and subtracted from the intercept produces the mean height for males. The ANOVA table in Table 10.4 is also identical to the ANOVA table in Table 10.2.

10.3 Equivalence of Linear Independent X-Variables (such as Contrasts) for Regression

It is not an accident that the ANOVA tables in Tables 10.2, 10.3, and 10.4 are identical. We explore here why that is the case.

Please review the definition of linear dependence in Section I.4.2.

The $\underset{n\times c}{X}$ matrix in the linear regression presentation of the one-way analysis of variance model with one factor with a categories must have a leading column of ones $X_0 = \mathbf{1}$ for the intercept and at least $a - 1$ additional columns, for a total of $c \geq a$ columns. The entire X matrix can be summarized by a *contrast matrix* $\underset{a\times c}{W}$ consisting of a unique rows, one for each level of the factor.

We explore the relationship between several different contrast matrices W in the case $a = 4$. The principles work for any value a. The matrix X of dummy variables itself consists of n_i copies of the i^{th} row of W (where $n = \sum_{i=1}^{a} n_i$):

$$\underset{n\times c}{X} = \underset{n\times 4}{\begin{pmatrix} n_1\{(1\,0\,0\,0) \\ n_2\{(0\,1\,0\,0) \\ n_3\{(0\,0\,1\,0) \\ n_4\{(0\,0\,0\,1) \end{pmatrix}} \underset{4\times c}{W} = \underset{n\times 4}{N} \quad \underset{4\times c}{W} \tag{10.1}$$

Any contrast matrix W with $a = 4$ rows and with rank 4 (which means it must have at least 4 columns) is equivalent for linear regression in the senses that

1. Any two such matrices W_1 and W_2 with dimensions $(4 \times c_1)$ and $(4 \times c_2)$ where $c_i \geq 4$ are related by postmultiplication of the first matrix by a full-rank matrix $\underset{c_1\times c_2}{A}$, that is,

$$\underset{4\times c_1}{W_1} \quad \underset{c_1\times c_2}{A} = \underset{4\times c_2}{W_2}$$

   Equivalently, any two such dummy variables matrices X_1 and X_2 with dimensions $(n \times c_1)$ and $(n \times c_2)$ are similarly related by

$$\underset{n\times c_1}{X_1} \quad \underset{c_1\times c_2}{A} = \underset{n\times c_2}{X_2}$$

   Examples (R code for all the contrast types in these examples is included in file `HHscriptnames(10)`):

1a. A simple overparameterized matrix (5 columns with rank=4) (this is the SAS default):

$$W_{\underset{4\times(1+4)}{\mathbf{simple}}} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

1b. Treatment contrasts (4 columns with rank=4) (R `contr.treatment`. This is the R default for factors. These are not 'contrasts' as defined in the standard theory for linear models as they are not orthogonal to the intercept.):

$$W_{\underset{4\times(1+4)}{\mathbf{simple}}} \underset{(1+4)\times(1+3)}{A} = W_{\underset{4\times(1+3)}{\texttt{treatment}}}$$

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

1c. Helmert contrasts (4 columns with rank=4) (R `contr.helmert`):

$$W_{\underset{4\times(1+4)}{\mathbf{simple}}} \underset{(1+4)\times(1+3)}{A} = W_{\underset{4\times(1+3)}{\texttt{helmert}}}$$

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & -2 & -2 & -2 \\ 0 & 0 & -2 & -2 \\ 0 & -1 & 1 & -2 \\ 0 & -1 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 0 & 2 & -1 \\ 1 & 0 & 0 & 3 \end{pmatrix}$$

1d. Sum contrasts (4 columns with rank=4) (R `contr.sum`):

$$W_{\underset{4\times(1+4)}{\mathbf{simple}}} \underset{(1+4)\times(1+3)}{A} = W_{\underset{4\times(1+3)}{\texttt{sum}}}$$

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -1 & -1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{pmatrix}$$

1e. Polynomial contrasts (4 columns with rank=4) (R `contr.poly`. This is the R default for ordered factors.):

$$\underset{4\times(1+4)}{W_{\textbf{simple}}} \qquad \underset{(1+4)\times(1+3)}{A} \qquad = \underset{4\times(1+3)}{W_{\texttt{polynomial}}}$$

$$\begin{pmatrix} 1&1&0&0&0\\ 1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&1 \end{pmatrix} \begin{pmatrix} 0.8 & 0.0000 & 0.0 & 0.0000\\ 0.2 & -0.6708 & 0.5 & -0.2236\\ 0.2 & -0.2236 & -0.5 & 0.6708\\ 0.2 & 0.2236 & -0.5 & -0.6708\\ 0.2 & 0.6708 & 0.5 & 0.2236 \end{pmatrix} =$$

$$\begin{pmatrix} 1 & -0.6708 & 0.5 & -0.2236\\ 1 & -0.2236 & -0.5 & 0.6708\\ 1 & 0.2236 & -0.5 & -0.6708\\ 1 & 0.6708 & 0.5 & 0.2236 \end{pmatrix}$$

2. The hat matrices are the same.

$$H_1 = (X_1(X_1'X_1)^{-1}X_1') = (X_2(X_2'X_2)^{-1}X_2') = H_2$$

An equivalent statement is that both X matrices span the same column space.

Proof. For the special case that $c = a$, hence the $X'X$ and A matrices are invertible:

$$\begin{aligned} H_2 &= \\ X_2(X_2'X_2)^{-1}X_2' &= \\ (X_1A)\Big((X_1A)'(X_1A)\Big)^{-1}(X_1A)' &= \\ (X_1A)(A'X_1'X_1A)^{-1}(A'X_1') &= \\ X_1(X_1'X_1)^{-1}X_1' &= \\ H_1 & \end{aligned}$$

When $c > a$, the step from line 4 to line 5 involves matrix algebra manipulations that we do not discuss here. Effectively, we are dropping any redundant columns.

3. The predicted values are the same.

$$\hat{Y} = H_1Y = H_2Y$$

4. The regression coefficients are related by premultiplication of the second set of coefficients by the same matrix A,

$$\beta_1 = A\beta_2$$

Proof.

$$E(Y) = X_2\beta_2 = (X_1A)\beta_2 = X_1(A\beta_2) = X_1\beta_1$$

5. The ANOVA (analysis of variance) table is the same:

Source	Sum of Squares
Regression	$SS_{Reg} = Y'H_1Y = Y'H_2Y$
Residual	$SS_{Res} = Y'(I - H_1)Y = Y'(I - H_2)Y$

Exercise 10.1 gives you the opportunity to explore the equivalence of the two coding schemes in Section 10.2.

As a consequence of the equivalence up to multiplication by a matrix A, the regression coefficients in regression analyses with factors (which means most experiments) are uninterpretable unless the definitions of the dummy variables have been provided.

10.4 Polynomial Contrasts and Orthogonal Polynomials

Ott (1993) reports an experiment that uses an abrasives testing machine to test the wear of a new experimental fabric. The machine was run at six different speeds (measured in revolutions per minute). Forty-eight identical square pieces of fabric were prepared, 8 of which were randomly assigned to each of the 6 machine speeds. Each square was tested for a three-minute period at the appropriate machine setting. The order of testing was appropriately randomized. For each square, the amount of wear was measured and recorded. The data from file `data(fabricwear)` are displayed in Figure 10.3. The initial ANOVA is in Table 10.5.

From Figure 10.3 we see that the assumption in Equation (6.3) of approximately constant variance across groups is satisfied by this dataset, hence ANOVA is an appropriate technique for investigating the data. We also note one outlier at speed=200. We will return to that data point later.

The ANOVA table in Table 10.5 shows that speed is significant. From the table of means we see that the means increase with speed and the increase is also faster as speed increases. Figure 10.3 shows the same and suggests that the means are increasing as a quadratic polynomial in speed.

There are several essentially identical ways to check this supposition. We start with the easiest to do and then expand by illustrating the arithmetic behind it. When we defined `speed` as a factor in Table 10.5, we actually did something more specific,

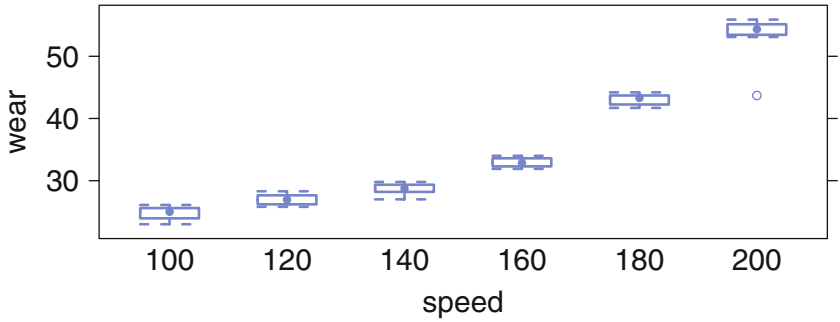

Fig. 10.3 Fabric wear as a function of speed. We see constant variance and a curved uphill trend. There is one outlier

Table 10.5 ANOVA and means for wear as a function of speed.

```
> fabricwear.aov <- aov(wear ~ speed, data=fabricwear)

> summary(fabricwear.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
speed        5   4872     974     298 <2e-16 ***
Residuals   42    137       3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> model.tables(fabricwear.aov, "mean")
Tables of means
Grand mean

34.93

 speed
speed
  100   120   140   160   180   200
24.78 26.96 28.68 32.93 43.05 53.19
```

we declared it to be an *ordered factor*. This means that the dummy variables are the orthogonal polynomials for six levels. We display the orthogonal polynomials in Figure 10.4 and Table 10.6. See the discussion in Section I.4 for an overview of orthogonal polynomials and their construction.

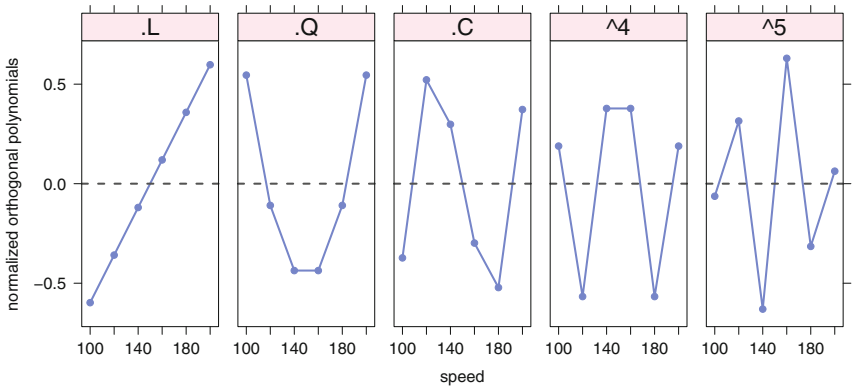

Fig. 10.4 Orthogonal polynomials for speed.

From the panels in Figure 10.4 we see that the linear polynomial plots as a straight line against the speed. The quadratic polynomial plots as a discretization of a parabola. The higher-order polynomials are rougher discretizations of their functions. In Table 10.6 we see that the orthogonal polynomials are scaled so their cross product is the identity matrix, that is, it is a diagonal matrix with 1s on the diagonal. Compare this (in Exercise 10.2) to a matrix of the simple powers of the integers $(1, 2, 3, 4, 5, 6)$. The columns of the simple powers span the same linear space as the orthogonal properties. Because they are not orthogonal (their cross product is not diagonal), their plots are harder to interpret and they may show numerical difficulties when used as predictor variables in a regression. See Appendix G for further discussion on the numerical issues.

In Table 10.7 we show two variants of an expanded display of the ANOVA from Table 10.5. The top of the table shows the regression coefficients for the regression against the orthogonal polynomials used as the dummy variables. Here we see that the linear and quadratic terms are highly significant. The cubic term is not significant. Based on our reading of the graph, and the comparison of the p-value for the quartic term to that of the quadratic term, we will interpret the quartic term as not significant and do all continuing work with the quadratic model.

In the bottom of Table 10.7 we show the partitioned ANOVA table with the linear, quadratic, and cubic terms isolated. By dint of the orthogonality the F-values are the square of the t-values for the coefficients ($36.3580^2 = 1321.903$) and the p-values are identical.

What happens when we redo the analysis without the outlier noted in Figure 10.3? The residual mean square goes down by a factor of 4; consequently, all the t-values go up. While the p-values for the cubic and quartic terms now show significance at .0001, we will continue to exclude them from our recommended model because the p-values for the linear and quadratic terms are orders of magnitude smaller ($< 10^{-16}$). See Exercise 10.8.

Table 10.6 Orthogonal polynomials for speed. The slightly complex algorithm shown here for scaling the orthogonal polynomials, with attention paid to computational precision by use of the `zapsmall` function, is necessary for factors with an odd number of levels. See Appendix G for further discussion on the numerical issues.

```
> tmp.c <- zapsmall(contrasts(fabricwear$speed), 14)

> dimnames(tmp.c)[[1]] <- levels(fabricwear$speed)

> tmp.c
         .L      .Q      .C      ^4       ^5
100 -0.5976  0.5455 -0.3727  0.1890 -0.06299
120 -0.3586 -0.1091  0.5217 -0.5669  0.31497
140 -0.1195 -0.4364  0.2981  0.3780 -0.62994
160  0.1195 -0.4364 -0.2981  0.3780  0.62994
180  0.3586 -0.1091 -0.5217 -0.5669 -0.31497
200  0.5976  0.5455  0.3727  0.1890  0.06299

> zapsmall(crossprod(tmp.c), 13)
   .L .Q .C ^4 ^5
.L  1  0  0  0  0
.Q  0  1  0  0  0
.C  0  0  1  0  0
^4  0  0  0  1  0
^5  0  0  0  0  1

> min.nonzero <- function(x, digits=13) {
+   xx <- zapsmall(x, digits)
+   min(xx[xx != 0])
+ }

> tmp.min <- apply(abs(tmp.c), 2, min.nonzero)

> sweep(tmp.c, 2, tmp.min, "/")
    .L .Q    .C ^4  ^5
100 -5  5 -1.25  1  -1
120 -3 -1  1.75 -3   5
140 -1 -4  1.00  2 -10
160  1 -4 -1.00  2  10
180  3 -1 -1.75 -3  -5
200  5  5  1.25  1   1
```

Table 10.7 Regression coefficients on dummy variables, and partitioned ANOVA table.

```
> summary(fabricwear.aov,
+         split=list(speed=list(speed.L=1, speed.Q=2,
+                     speed.C=3, rest=4:5)))
                 Df Sum Sq Mean Sq F value  Pr(>F)
speed             5   4872     974  297.70 < 2e-16 ***
  speed: speed.L  1   4327    4327 1321.90 < 2e-16 ***
  speed: speed.Q  1    513     513  156.76 9.1e-16 ***
  speed: speed.C  1      7       7    2.10   0.154
  speed: rest     2     25      13    3.88   0.028 *
Residuals        42    137       3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary.lm(fabricwear.aov)

Call:
aov(formula = wear ~ speed, data = fabricwear)

Residuals:
   Min     1Q Median     3Q    Max
-9.487 -0.653  0.181  0.825  2.712

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   34.929      0.261  133.76  < 2e-16 ***
speed.L       23.256      0.640   36.36  < 2e-16 ***
speed.Q        8.009      0.640   12.52  9.1e-16 ***
speed.C        0.928      0.640    1.45    0.154
speed^4       -1.677      0.640   -2.62    0.012 *
speed^5       -0.600      0.640   -0.94    0.354
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.81 on 42 degrees of freedom
Multiple R-squared:  0.973,Adjusted R-squared:  0.969
F-statistic:  298 on 5 and 42 DF,  p-value: <2e-16
```

10.4.1 Specification and Interpretation of Interaction Terms

Example—consider a model

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{34} X_3 X_4$$

to "explain" determinants of annual salary Y in dollars for workers in some population. Here X_1 is age in years, X_2 is gender (1 if female, 0 if male), X_3 is race (1 if white, 0 if nonwhite), and X_4 is number of years of schooling. (Discussion: What other variables might such a model include to explain salary?)

The existence of the interaction terms allows for the possibility that the degree of enhancement of education on schooling differs for whites and nonwhites.

Consider a white and a nonwhite of the same age and gender and having the same amount of schooling. Then:

- β_4 is the expected increase in annual salary for a nonwhite attributable to an additional year of schooling.
- $\beta_4 + \beta_{34}$ is the expected increase in annual salary for a white attributable to an additional year of schooling.
- β_{34} is the expected amount by which a white's salary increase as a result of an additional year of schooling exceeds a nonwhite's salary increase as a result of an additional year of schooling.

Also, still assuming the same age and gender,

- $\beta_3 + \beta_{34} X_4$ is the difference between white and nonwhite expected salary.
- β_3 is the component of this difference that does not depend on years of schooling and is attributable only to difference in race.

We examine this model further in Exercise 10.7.

10.5 Analysis Using a Concomitant Variable (Analysis of Covariance—ANCOVA)

In some situations where we seek to compare the differences in the means of a continuous response variable across levels of a factor A, we have available a second continuous variable that can be used to improve our ability to distinguish among the levels. Historically this extended model has been called the *analysis of covariance* model because the second variable varies along with the first. To avoid confusion with the concept of covariance introduced in Chapter 3, we prefer to call this approach *analysis using a concomitant variable*. Nevertheless we will retain use of

the term covariate as a shorthand term for concomitant variable and the acronym ANCOVA as an abbreviation for this method.

If X_{ij} denotes the j^{th} observation of the covariate at the i^{th} level of factor A, our original ANOVA model in Equation (6.1) generalizes to

$$Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \bar{\bar{X}}) + \epsilon_{ij} \quad \text{for} \quad i = 1, \ldots, a \quad \text{and} \quad j = 1, \ldots, n_i \tag{10.2}$$

where $\bar{\bar{X}}$ is the grand mean of the X_{ij}'s and all other terms are as defined in Equation (6.1). The model in Equation (10.2) has separate intercepts α_i for each level of A but retains a common slope. The differences between the intercepts α_i are identical to the vertical differences between the parallel lines (to be illustrated in Figure 10.8). Equation (10.2) is the classical ANCOVA model.

The logic of this approach is that if X_{ij} is related to Y_{ij} then the ϵ's of the model in Equation (10.2) will be measured from a different regression line for each level of A rather than from a different horizontal line as in model (6.1). This will give the ϵ's less variability than those of Equation (6.1), thereby sharpening our inferences on the α_i's. The α_i's estimated from Equation (10.2) are said to be *adjusted* for the covariate. Quite frequently the range of observed X_{ij} differs for each level of A_i and therefore the $\bar{Y}_i$ means from Equation (6.1) reflect the difference in the X-values more than the differences attributable to the change in levels of A.

The next level of generalization allows the slopes to differ, i.e., replace the common β in Equation (10.2) with β_i:

$$Y_{ij} = \mu + \alpha_i + \beta_i(X_{ij} - \bar{\bar{X}}) + \epsilon_{ij} \quad \text{for} \quad i = 1, \ldots, a \quad \text{and} \quad j = 1, \ldots, n_i \tag{10.3}$$

We illustrate models Equations (10.2) and (10.3) in Section 10.6. In Section 10.6.5 we will use the model in Equation (10.3) to test the assumption that the lines are parallel. Formally, we will test whether the lines have the same slope

$$H_0: \beta_1 = \beta_2 = \beta_3 \tag{10.4}$$
$$H_1: \text{Not all } \beta_i \text{ are identical}$$

or the same intercept

$$H_0: \alpha_1 = \alpha_2 = \alpha_3 \tag{10.5}$$
$$H_1: \text{Not all } \alpha_i \text{ are identical}$$

or both (in which case the lines coincide). We illustrate each model by an appropriate graph. We construct a single meta-graph in Figure 10.12 to illustrate the comparison of all the models we consider.

These ideas can be extended to situations with more than one covariate variable and to more complicated experimental designs such as those discussed in Chapters 12 through 14.

10.6 Example—Hot Dog Data

10.6.1 Study Objectives

Hot dogs based on poultry are said to be healthier than ones made from either meat (beef and pork) or all beef. A basis for this claim may be the lower-calorie (fat) content of poultry hot dogs. Is this advantage of poultry hot dogs offset by a higher sodium content than meat hot dogs?

Researchers for *Consumer Reports* analyzed three types of hot dog: beef, poultry, and meat (mostly pork and beef, but up to 15% poultry meat). The data in file `data(hotdog)` come from Consumer Reports (1986) and were later used by Moore and McCabe (1989).

10.6.2 Data Description

`Type:` Type of hot dog (beef, meat, or poultry)

`Calories:` Calories per hot dog

`Sodium:` Milligrams of sodium per hot dog

10.6.3 One-Way ANOVA

We start by comparing the `Sodium` content of the three hot dog `Types` by the methods of Chapter 6 in Figure 10.5 and in Table 10.8. We see that the three `Types` have similar `Sodium` content.

Figure 10.6 shows the response `Sodium` plotted against the covariate `Calories` by `Type`. Within each panel we plot a horizontal line at the mean of the `Sodium`

values for that Type. The analysis of variance in Table 10.8 compares the vertical distance between these horizontal lines. It ignores the most evident feature of this plot, that the three Types have very different fat contents with Poultry low, Beef intermediate, and Meat high. We wish to see if knowledge about Calories affects our understanding about Sodium.

10.6.4 Concomitant Explanatory Variable—ANCOVA

It is possible that our finding of similar Sodium content is attributable in part to a need to add sodium to enhance the flavor of higher-fat hot dogs. The Calories information can be incorporated into the analysis by adding Calories to the model as a concomitant explanatory variable. Then in this revised model, comparisons between the mean Sodium contents of the three Types will have been *adjusted for* differing Calories contents. In this way, comparisons between the three Types will be made on the basis that each Type has the mean Calories content of all Types.

We illustrate this revised analysis in two steps. Initially, in Figure 10.7 and Table 10.9, we show the regression (Chapter 8) of Sodium on Calories ignoring the Types. The common regression line makes some sense in the Superpose panel but very clearly has the wrong slope and wrong intercept in all three of the individual panels.

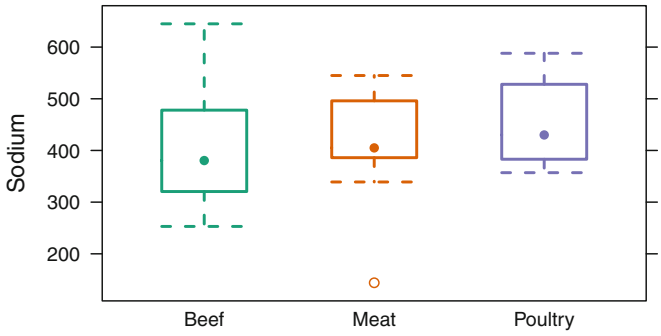

Fig. 10.5 Boxplots comparing the Sodium content of three Types of hot dogs. See Table 10.8.

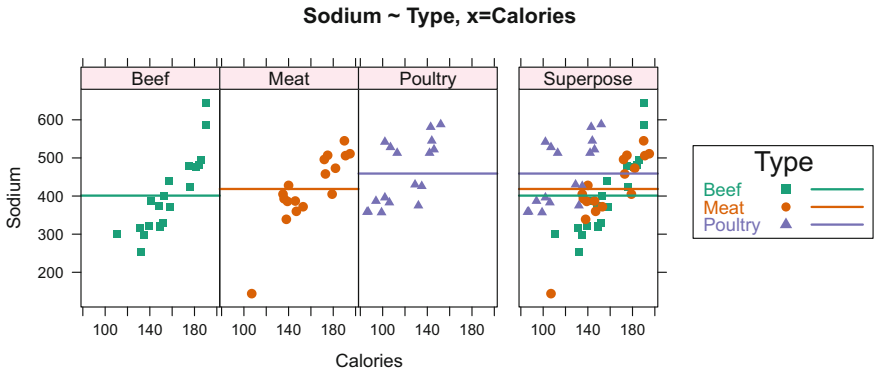

Fig. 10.6 `Sodium ~ Type, x=Calories`. Horizontal lines at `Sodium` means for each `Type`. $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$. See Table 10.8. The intent of the notation is twofold: The arithmetic of the analysis is based on the one-way ANOVA of `Sodium ~ Type`. The graph is more complex. The points in the graph show `y=Sodium` plotted against `x=Calories` separately for each level of `Type`. The horizontal line in each panel is the mean of the levels of `Sodium` at each level of `Type`.

Table 10.8 Hot dog ANOVA and means. This is the one-way ANOVA of Chapter 6. See Figures 10.5 and 10.6.

```
> aovStatementAndAnova(TxC)
> anova(aov(Sodium ~ Type, data = hotdog))
Analysis of Variance Table

Response: Sodium
          Df Sum Sq Mean Sq F value Pr(>F)
Type       2  31739   15869    1.78   0.18
Residuals 51 455249    8926

> model.tables(TxC, type="means")
Tables of means
Grand mean

424.8

 Type
     Beef  Meat Poultry
    401.1 418.5     459
rep  20.0  17.0      17
```

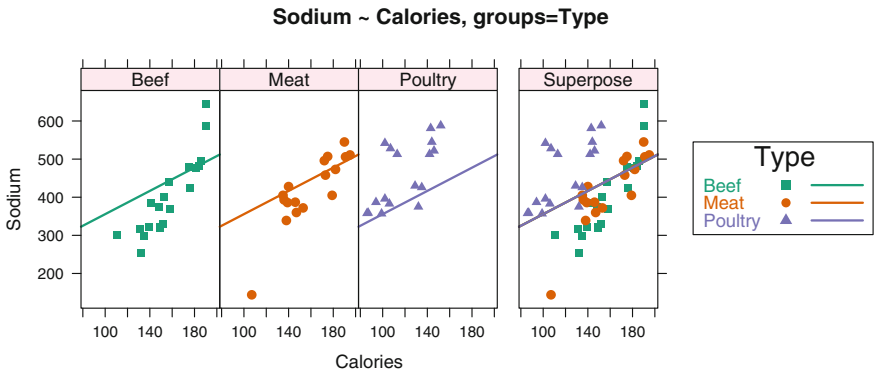

Fig. 10.7 `Sodium ~ Calories, groups=Type`. Common regression line that ignores `Type`. $Y_{ij} = \mu + \beta(X_{ij} - \bar{\bar{X}}) + \epsilon_{ij}$. See Table 10.9. The intent of the notation is twofold: The arithmetic of the analysis is based on the simple linear regression of `Sodium ~ Calories`. The graph is more complex. The points in the graph show y=`Sodium` plotted against x=`Calories` separately for each level of `Type`. The common regression line in all panels ignores `Type`.

Table 10.9 Hot dog ANCOVA with a common regression line that ignores `Type`. See Figure 10.7.

```
> aovStatementAndAnova(CgT, warn=FALSE)
> anova(aov(Sodium ~ Calories, data = hotdog))
Analysis of Variance Table

Response: Sodium
           Df Sum Sq Mean Sq F value  Pr(>F)
Calories    1 106270  106270    14.5 0.00037 ***
Residuals 52 380718    7321
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 10.8 and Table 10.10 show parallel regression lines for each type. They have separate intercepts and a common slope. This model is the standard *analysis of covariance* model. We are interested in the vertical distance between the parallel lines. Equivalently, we are interested in the distance between the intercepts. We see from the $F = 37.07433$ with $p = 1.3_{10^{-10}}$ in the first part of Table 10.10 that the vertical distance is significant.

The original preliminary conclusion based on Table 10.8 was misleading because it left out the critical dependence of y=`Sodium` on the x=`Calories` variable.

It is possible (see Exercise 10.5 for an example) for the covariate to be significant and not the grouping factor. In this example both are significant.

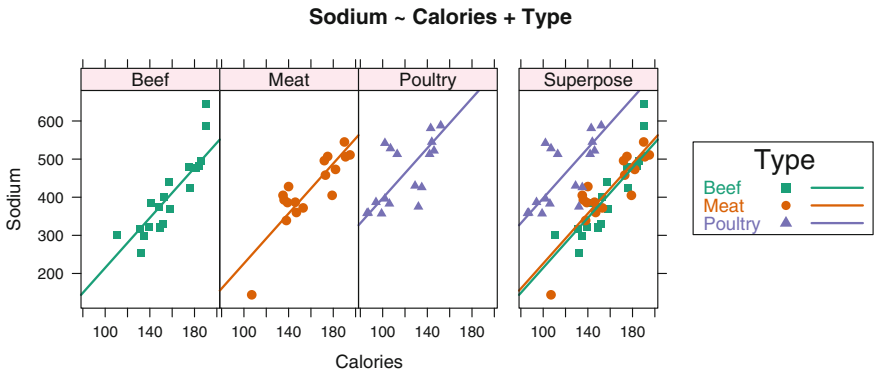

Fig. 10.8 `Sodium ~ Calories + Type`. Parallel lines. $Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \bar{\bar{X}}) + \epsilon_{ij}$. See Table 10.10. This illustrates the standard ANCOVA model.

Table 10.10 Hot dog ANCOVA with parallel lines and separate intercepts. See Figure 10.8.

```
> aovStatementAndAnova(CpT)
> anova(aov(Sodium ~ Calories + Type, data = hotdog))
Analysis of Variance Table

Response: Sodium
           Df Sum Sq Mean Sq F value  Pr(>F)
Calories    1 106270  106270    34.6 3.3e-07 ***
Type        2 227386  113693    37.1 1.3e-10 ***
Residuals 50 153331    3067
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We construct Figure 10.9 and Table 10.11 to show the means for the response `Sodium` adjusted for the covariate `Calories`. The adjustment maintains the same vertical distance between the fitted lines that we observe in Figure 10.8. From the ANOVA table in Table 10.11 we see that the adjusted means have the same residual sum of squares as the unadjusted means. The residual degrees of freedom are wrong because the analysis doesn't know that the effect of the `Calories` variable has already been removed. The `Type` sum of squares is not what we anticipated because we did not adjust the `Type` dummy variables for the covariate; we only adjusted the response variable.

Now that we have shown the factor `Type` to be important, we show in Table 10.12 and Figure 10.10 the results of multiple comparisons analysis using the Tukey procedure. These show that `Meat` and `Beef` are indistinguishable and that `Poultry` differs from both `Meat` and `Beef`.

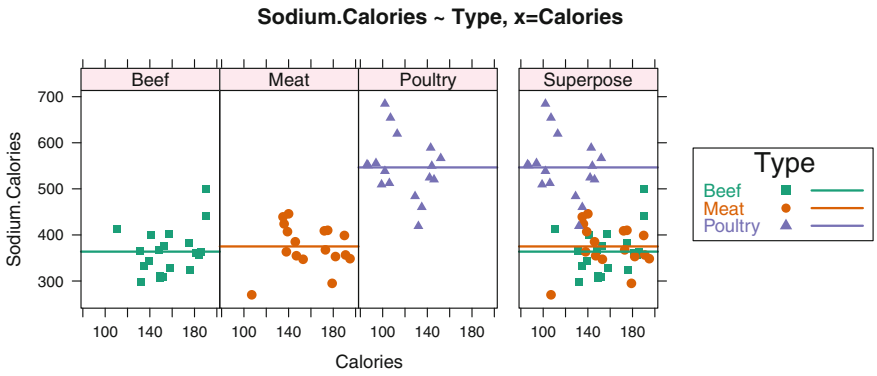

Fig. 10.9 `Sodium.Calories ~ Type, x=Calories`. Horizontal lines after adjustment for the covariate. $\left(Y_{ij}|X_{ij}\right) = \mu + \alpha_i + \epsilon_{ij}$. See Table 10.11. The vertical distance from each point to its line is identical in this figure to the vertical distances shown in Figure 10.8.

Table 10.11 Horizontal lines after adjustment for the covariate. See Figure 10.9.

```
> aovStatementAndAnova(T.C)
> anova(aov(Sodium.Calories ~ Type, data = hotdog))
Analysis of Variance Table

Response: Sodium.Calories
          Df Sum Sq Mean Sq F value  Pr(>F)
Type       2 368463  184232    61.3 2.7e-14 ***
Residuals 51 153331    3006
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

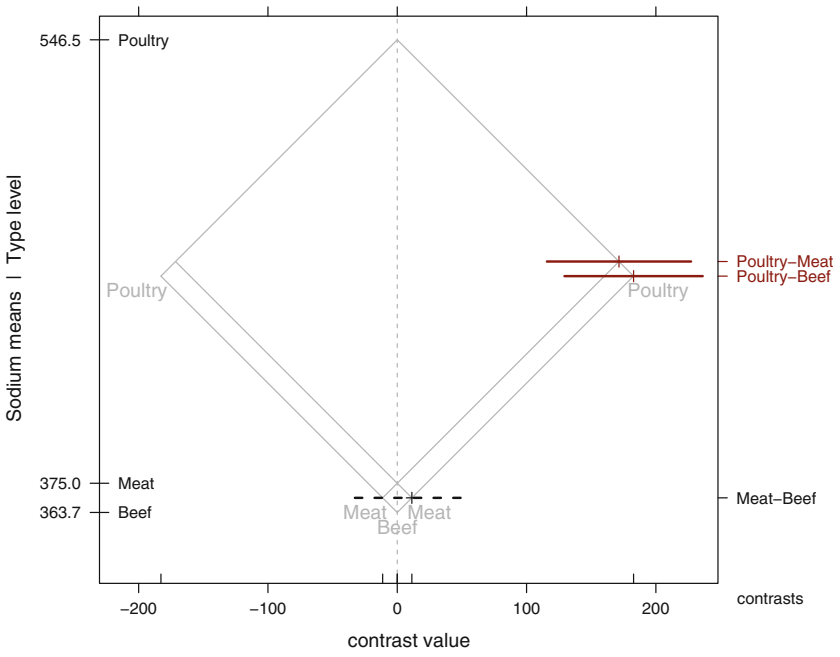

Fig. 10.10 Multiple comparisons by Tukey's method of the ANCOVA model in Figure 10.8 and Table 10.10 comparing the mean `Sodium` content of three `Types` of hot dogs adjusted for `Calories`. See also Figures 10.8 and 10.9 and Table 10.12.

Table 10.12 Multiple comparisons by Tukey's method of the ANCOVA model in Figure 10.8 and Table 10.10 comparing the mean `Sodium` content of three `Types` of hot dogs adjusted for `Calories`. See also Figure 10.10.

```
> CpT.mmc <- mmc(aov.trellis(CpT))

> CpT.mmc
Tukey contrasts
Fit: aov(formula = Sodium ~ Calories + Type, data = hotdog)
Estimated Quantile = 2.41
95% family-wise confidence level
$mca
             estimate stderr  lower  upper height
Poultry-Meat   171.47  23.13 115.73 227.21  460.8
Poultry-Beef   182.76  22.19 129.30 236.22  455.1
Meat-Beef       11.29  18.28 -32.75  55.34  369.4
$none
        estimate stderr lower upper height
Poultry    546.5  16.07 507.8 585.2  546.5
Meat       375.0  14.13 341.0 409.1  375.0
Beef       363.7  12.94 332.6 394.9  363.7
```

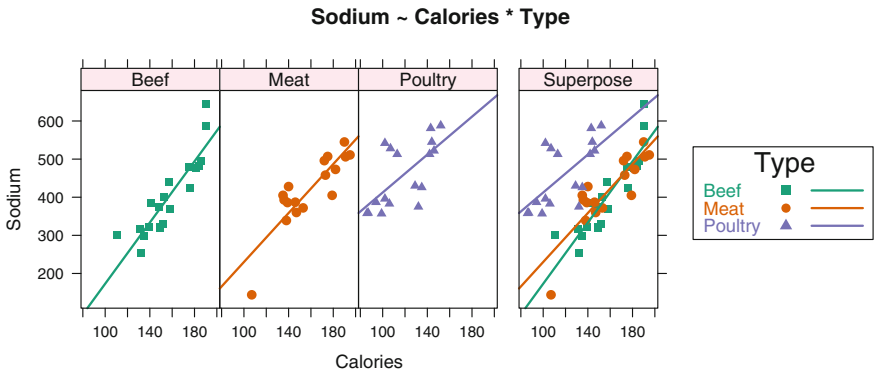

Fig. 10.11 `Sodium ~ Calories * Type`. Separate regression lines. $Y_{ij} = \mu + \alpha_i + \beta_i(X_{ij} - \bar{\bar{X}}) + \epsilon_{ij}$. See Table 10.13.

Table 10.13 Hot dog ANCOVA with separate regression lines (slopes and intercepts). See Figure 10.11.

```
> aovStatementAndAnova(CsT)
> anova(aov(Sodium ~ Calories * Type, data = hotdog))
Analysis of Variance Table

Response: Sodium
             Df Sum Sq Mean Sq F value  Pr(>F)
Calories      1 106270  106270   35.69 2.7e-07 ***
Type          2 227386  113693   38.18 1.2e-10 ***
Calories:Type 2  10402    5201    1.75    0.19
Residuals    48 142930    2978
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

10.6.5 Tests of Equality of Regression Lines

In Section 10.6.4 we assume the constant slope model (10.2) and test whether the intercepts differed by testing (10.5) about α_i. We can also work with the separate slope model (10.3) and test (10.4) about β_i.

Figure 10.11 and Table 10.13 show separate regression lines for each group. These have separate intercepts and slopes. The F-test of `Calories:Type` in Table 10.13 having p-value = .185 addresses the null hypothesis that the regression lines for predicting `Sodium` from `Calories` are parallel.

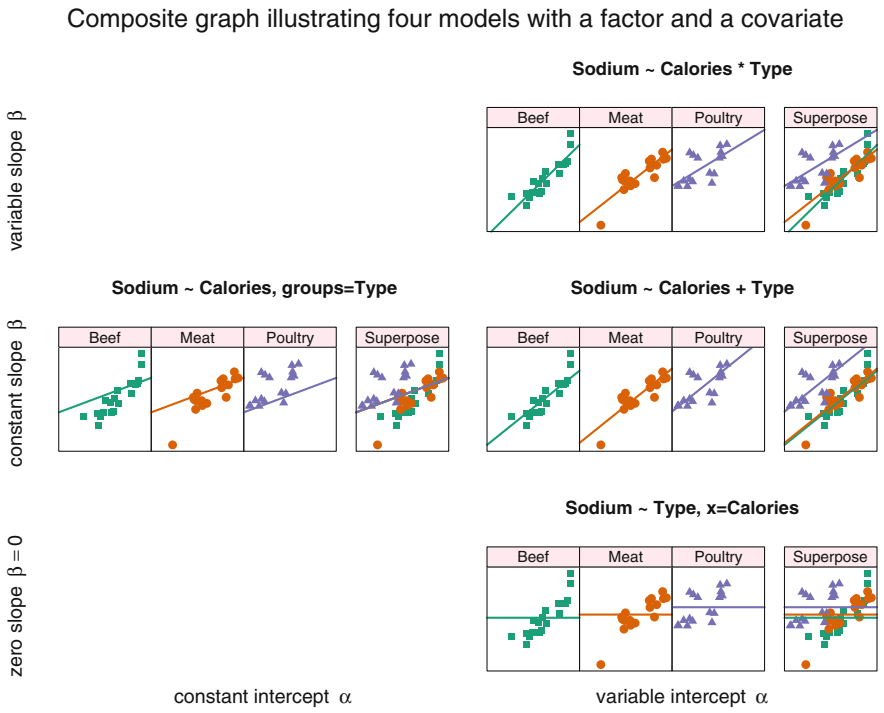

Fig. 10.12 Four models for the hot dog data, arranged in two columns corresponding to the two possibilities for the intercept in the model and three rows corresponding to the three possibilities for the slope in the model. The models are often described as

	Constant intercept α	Variable intercept α
Variable slope β		Analysis of covariance with interaction of the factor and the covariate.
Constant slope β	Linear regression, ignoring the factor.	Standard analysis of covariance with constant slope and variable intercept.
Zero slope $\beta = 0$		Analysis of variance, ignoring the covariate.

Observe in Figure 10.11 that the slopes of the lines for the regressions of Sodium on Calories appear to differ for the three Types of hot dog. This null hypothesis is expressed as two equalities in Equation (10.4) and is tested in Table 10.13 using the two degree-of-freedom sum of squares for the interaction Calories:Type. The p-value for this test, 0.185, implies that the null hypothesis cannot be rejected and therefore that the three slopes are homogeneous. Any difference among them is too small to detect with the sample sizes in this data set.

Conditional on the homogeneity of the three slopes, the two degree-of-freedom sum of squares for Type in Table 10.10 tests the hypothesis that the three regression lines have a common intercept, a null hypothesis expressed in Equation (10.5). The zero p-value for this test implies that the intercepts are not identical.

10.7 ancovaplot Function

The ANCOVA plot has been calculated with the ancovaplot function, one of the functions that we provide in the **HH** package. The ancovaplot function constructs the appropriate trellis graphics commands for the plot. The specific feature that requires a separate function is its handling of the x= and groups= arguments respectively for the one-way ANOVA and the simple regression models. The result of the function is an ancovaplot object, which is essentially an ordinary trellis object with a different class. We have provided methods for ancova and related functions that will operate directly on the ancovaplot object.

The four basic options are shown in Table 10.14. Output from each is shown in Figures 10.7, 10.6, 10.8, and 10.11 and Tables 10.9, 10.8, 10.10, and 10.13. Figure 10.12 shows the graphs from all four in a single coordinated display.

Table 10.14 Four ways to use the `ancovaplot` function. See Figure 10.12 for a coordinated placement of all four of these plots on the same page.

```
data(hotdog, package="HH")
data(col3x2, package="HH")

## constant line across all groups
## y ~ x
ancovaplot(Sodium ~ Calories, groups=Type, data=hotdog, col=col3x2)

## different horizontal line in each group
## y ~ a
ancovaplot(Sodium ~ Type, x=Calories, data=hotdog, col=col3x2)

## constant slope, different intercepts
## y ~ x + a  or  y ~ a + x
ancovaplot(Sodium ~ Calories + Type, data=hotdog, col=col3x2)

## different slopes, and different intercepts
## y ~ x * a  or  y ~ a * x
ancovaplot(Sodium ~ Calories * Type, data=hotdog, col=col3x2)
```

10.8 Exercises

We recommend that for all exercises involving a data set, you begin by examining a scatterplot matrix of the variables.

10.1. Demonstrate that the two coding schemes

$$W_{\text{female}} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad W_{\text{treat}} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

in Section 10.2 are equivalent for regression in the sense of Section 10.3 by finding the A matrix that relates them.

10.2. Demonstrate that the orthogonal polynomials in Table 10.6 span the same column space as the matrix whose columns are the simple polynomials $x = (1, 2, 3, 4, 5, 6)$, x^2, x^3, x^4, x^5. Plot the columns of the matrix and compare the plot to Figure 10.4.

10.3. Demonstrate that the two coding schemes for the `ResidenceLocation` example in Section 10.1 are equivalent by defining the corresponding W variables and finding the A matrix that relates them.

10.4. We first investigated the dataset `data(water)` in Exercise 4.4.

a. Plot `mortality` vs `calcium`, using separate plot symbols for each value of `derbynor`. Does it appear from this plot that `derbynor` would contribute to explaining the variation in mortality?

b. Perform separate regressions of `mortality` on `calcium` for each value of `derbynor`. Compare these to the estimated coefficients in a multiple regression of `mortality` on both `calcium` and `derbynor`.

c. Interpret the regression coefficients in the multiple regression in terms of the model variables.

d. Suggest the public health conclusions of your analysis.

10.5. Do an analysis of covariance with model (10.2) of the simple dataset

y	x	a
1	1	1
2	2	1
3	3	2
4	4	2
5	5	3
6	6	3

Show that covariate x is significant and the grouping factor a is not.

10.6. The Erie house-price data `data(hpErie)` is introduced in Exercise 9.3. That exercise invites examination of the impact of two high-priced houses by comparing analyses with these houses included or omitted. Revisit these data, adding a dummy variable `highprice` defined as 1 if one of the two high-priced houses and 0 otherwise. Perform a stepwise regression analysis including this new variable and compare your results with those in Exercise 9.3.

10.7. Reconsider the salary model in Section 10.4.1.

a. Interpret, in terms of the model variables salary, age, gender, etc., the finding that β_2 is significantly less than zero.

b. Write the null hypothesis in terms of the β_j's:

> $E(Y)$ for whites with 12 years of schooling is the same as $E(Y)$ for nonwhites with 16 years of schooling.

c. Write the null hypothesis in terms of the β_j's:

> $E(Y)$ increases at the rate of \$2,000 per year of schooling for whites and at the rate of \$2,500 per year of schooling for nonwhites.

d. If the gender and race are interpreted as factors, rather than as arbitrarily coded dummy variables, then the generated dummy variables differ from the 0 and 1 coding used in Section 10.4.1. Therefore, the estimated $\hat{\beta}_j$ will differ. Explain why the t-tests and the F-test will remain the same.

10.8. Rerun the polynomial contrasts for the `data(fabricwear)` example in Table 10.7 without the outlier noted in Figure 10.3.