Chapter 15

Analysis of Covariance

Predictors that are qualitative in nature, like for example eye color, are sometimes called *categorical* or *factors*. How can these predictors be incorporated into a regression analysis? Analysis of Covariance refers to regression problems where there is a mixture of quantitative and qualitative predictors.

Suppose we are interested in the effect of a medication on cholesterol level - we might have two groups - one of which receives the medication and the other which does not. However, we could not treat this as a simple two sample problem if we knew that the two groups differed with respect to age and this would affect the cholesterol level. See Figure 15 for a simulated example. For the patients who received the medication, the mean reduction in cholesterol level was 0% while for those who did not the mean reduction was 10%. So superficially it would seem that it would be better not to be treated. However, the treated group ranged in age from 50 to 70 while those who were not treated ranged in age between 30 and 50. We can see that once age is taken into account, the difference between treatment and control is again 10% but this time in favor of the treatment.

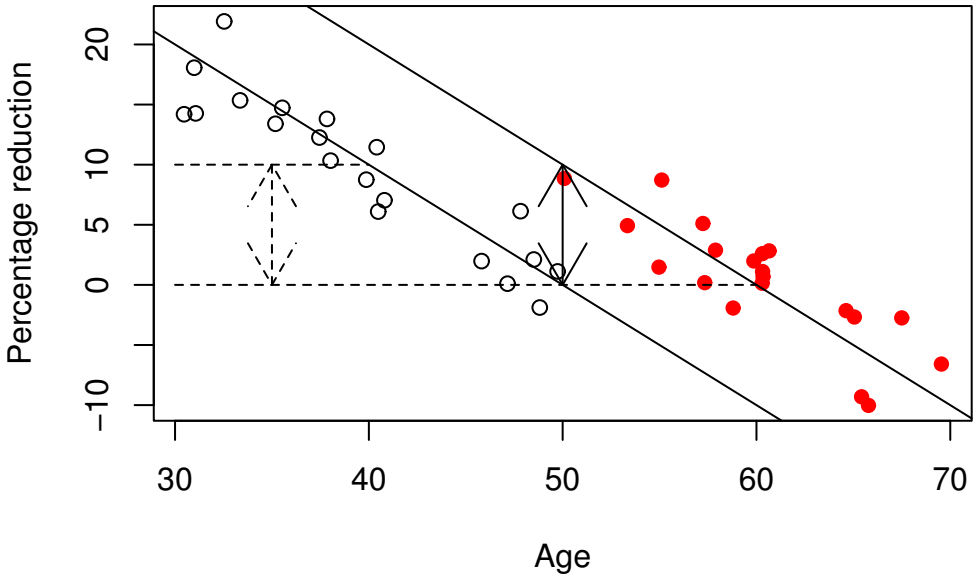

Figure 15.1: Simulated example showing the confounding effect of a covariate. The patients who took the medication are marked with a solid dot while those who did not are marked with an empty dot

Analysis of covariance is a method for adjusting the groups for the age difference and then seeing the

effect of the medication. It can also be used when there are more than two groups and more than one covariate.

> Our strategy is to incorporate the qualitative predictors within the $y = X\beta + \varepsilon$ framework. We can then use the estimation, inferential and diagnostic techniques that we have already learnt.

This avoids have to learn a different set of formulae for each new type of qualitative predictor configuration which is the approach taken by many texts. To put qualitative predictors into the $y = X\beta + \varepsilon$ form we need to code the qualitative predictors. Let's consider a specific example:

- y = change in cholesterol level
- x = age
- $$d = \begin{cases} 0 & \text{did not take medication} \\ 1 & \text{took medication} \end{cases}$$

A variety of linear models may be considered here:

1. The same regression line for both groups — $y = \beta_0 + \beta_1 x + \varepsilon$ or in R `y ~ x`
2. Separate regression lines for each group but with the same slope — $y = \beta_0 + \beta_1 x + \beta_2 d + \varepsilon$ or in R `y ~ x + d`. In this case β_2 represents the distance between the regression lines i.e. the effect of the drug.
3. Separate regression lines for each group $y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 x.d + \varepsilon$ or in R `y ~ x + d + d:x` or `y ~ x*d`. Any interpretation of the effect of the drug will now depend on age also. To form the slope interaction term $x.d$ in the X-matrix, simply multiply x by d elementwise.

Estimation and testing works just as it did before. Interpretation is much easier if we can eliminate the slope interaction term.

Other codings of d are possible, for instance

$$d = \begin{cases} -1 & \text{did not take medication} \\ 1 & \text{took medication} \end{cases}$$

is used by some. This coding enables β_2 and β_3 to be viewed as differences from a response averaged over the two groups. Any other coding that assigned a different number to the two groups would also work but interpretation of the estimated parameters might be more difficult.

15.1 A two-level example

The data for this example consist of `x`= nave height and `y` = total length in feet for English medieval cathedrals. Some are in the Romanesque (r) style and others are in the Gothic (g) style. Some cathedrals have parts in both styles and are listed twice. We wish to investigate how the length is related to height for the two styles. Read in the data and make a summary on the two styles separately.

```
> data(cathedral)
> cathedral
             style   x   y
Durham           r  75 502
Canterbury       r  80 522
....etc....
Old.St.Paul      g 103 611
Salisbury        g  84 473
> lapply(split(cathedral,cathedral$style),summary)
$g
  style        x               y
 g:16   Min.   : 45.0   Min.   :182
 r: 0   1st Qu.: 60.8   1st Qu.:299
        Median : 73.5   Median :412
        Mean   : 74.9   Mean   :397
        3rd Qu.: 86.5   3rd Qu.:481
        Max.   :103.0   Max.   :611

$r
  style       x              y
 g:0   Min.   :64.0   Min.   :344
 r:9   1st Qu.:70.0   1st Qu.:425
       Median :75.0   Median :502
       Mean   :74.4   Mean   :475
       3rd Qu.:80.0   3rd Qu.:530
       Max.   :83.0   Max.   :551
```

Now plot the data — see Figure 15.1.

```
> plot(cathedral$x,cathedral$y,type="n",xlab="Nave height",ylab="Length")
> text(cathedral$x,cathedral$y,as.character(cathedral$s))
```

Now fit the separate regression lines model. `y ~ x*style` is equivalent.

```
> g <- lm(y ~ x+style+x:style, data=cathedral)
> summary(g)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    37.11      85.68    0.43   0.6693
x               4.81       1.11    4.32   0.0003
style         204.72     347.21    0.59   0.5617
x.style        -1.67       4.64   -0.36   0.7227

Residual standard error: 79.1 on 21 degrees of freedom
Multiple R-Squared: 0.541,      Adjusted R-squared: 0.476
F-statistic: 8.26 on 3 and 21 degrees of freedom,       p-value: 0.000807
```

Because style is non-numeric, R automatically treats it as a qualitative variables and sets up a coding - but which coding?

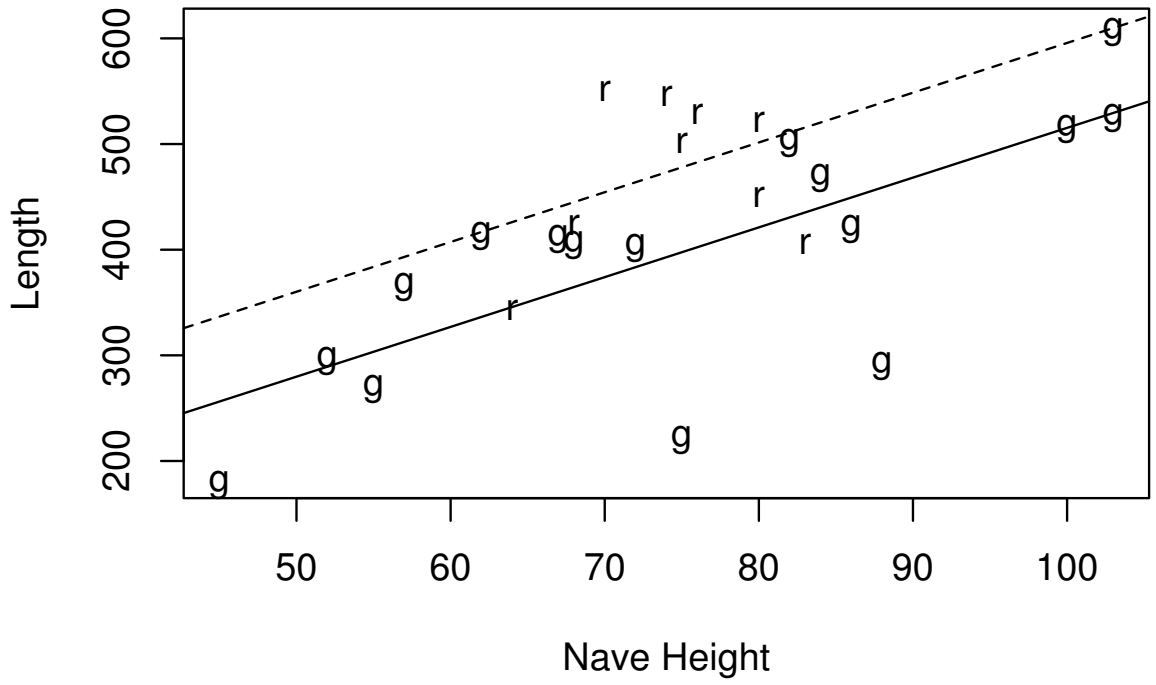

Figure 15.2: A comparison of Romanesque (r) and Gothic (g) Cathedrals

```
> model.matrix(g)
             (Intercept)   x style x.style
Durham                 1  75     1      75
Canterbury             1  80     1      80
...etc...
Old.St.Paul            1 103     0       0
Salisbury              1  84     0       0
```

We see that the model can be simplified to

```
> g <- lm(y ~ x+style, cathedral)
> summary(g)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    44.30      81.65    0.54   0.5929
x               4.71       1.06    4.45   0.0002
style          80.39      32.31    2.49   0.0209

Residual standard error: 77.5 on 22 degrees of freedom
Multiple R-Squared: 0.538,      Adjusted R-squared: 0.496
F-statistic: 12.8 on 2 and 22 degrees of freedom,       p-value: 0.000203
```

Put the two parallel regression on the plot:

```
> abline(44.30,4.71)
> abline(44.30+80.39,4.71,lty=2)
```

A check on the diagnostics reveals no particular problems.

Our conclusion is that for cathedrals of the same height, Romanesque ones are 80.39 feet longer. For each extra foot in height, both types of cathedral are about 4.7 feet longer. Gothic cathedrals are treated as the *reference level* because "g" comes before "r" in the alphabet. We can change this:

```
> cathedral$style <- relevel(cathedral$style,ref="r")
> g <- lm(y ~ x+style, cathedral)
> summary(g)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   124.69      82.92    1.50   0.1469
x               4.71       1.06    4.45   0.0002
style         -80.39      32.31   -2.49   0.0209

Residual standard error: 77.5 on 22 degrees of freedom
Multiple R-Squared: 0.538,      Adjusted R-squared: 0.496
F-statistic: 12.8 on 2 and 22 degrees of freedom,       p-value: 0.000203
```

Although the coefficients have different numerical values, this coding leads to the same conclusion as before.

Notice that in this case the two groups have about the same average height — about 74 feet. The difference in the lengths is 78 feet on average which is similar to the 80 feet from the fit. This is in contrast to the cholesterol example above where the two groups had very different means in their predictors.

15.2 Coding qualitative predictors

There is no unique coding for a two-level factor — there are even more choices with multi-level predictors. For a k-level predictor, $k-1$ *dummy* variables are needed for the representation. One parameter is used to represent the overall mean effect or perhaps the mean of some reference level and so only $k-1$ variables are needed rather than k.

These dummy variables cannot be exactly collinear but otherwise there is no restriction. The choice should be based on convenience.

Treatment coding

Consider a 4 level factor that will be coded using 3 dummy variables. This table describes the coding:

	Dummy coding		
	1	2	3
1	0	0	0
levels 2	1	0	0
3	0	1	0
4	0	0	1

This treats level one as the standard level to which all other levels are compared so a control group, if one exists, would be appropriate for this level. R assigns levels to a factor in alphabetical order by default. The columns are orthogonal and the corresponding dummies will be too. The dummies won't be orthogonal to the intercept. Treatment coding is the default choice for R

Helmert Coding

	Dummy coding		
	1	2	3
1	-1	-1	-1
levels 2	1	-1	-1
3	0	2	-1
4	0	0	3

If there are equal numbers of observations in each level (a balanced design) then the dummy variables will be orthogonal to the each other and the intercept. This coding is not so nice for interpretation. It is the default choice in S-PLUS.

There are other choices of coding — anything that spans the $k-1$ dimensional space will work. The choice of coding does not affect the R^2, $\hat{\sigma}^2$ and overall F-statistic. It does effect the $\hat{\beta}$ and you do need to know what the coding is before making conclusions about $\hat{\beta}$.

15.3 A Three-level example

Here's an example with a qualitative predictor with more than one level. The data for this example come from a 1966 paper by Cyril Burt entitled "The genetic determination of differences in intelligence: A study of monozygotic twins reared apart". The data consist of IQ scores for identical twins, one raised by foster parents, the other by the natural parents. We also know the social class of natural parents (high, middle or low). We are interested in predicting the IQ of the twin with foster parents from the IQ of the twin with the natural parents and the social class of natural parents. Let's read in and take a look at the data:

```
> data(twins)
> twins
   Foster Biological Social
 1     82         82   high
 2     80         90   high
etc.
26    107        106    low
27     98        111    low
> plot(twins$B,twins$F,type="n",xlab="Biological IQ",ylab="Foster IQ")
> text(twins$B,twins$F,substring(as.character(twins$S),1,1))
```

See Figure 15.3 — what model seems appropriate? The most general model we'll consider is the seperate lines model:

```
> g <- lm(Foster ~ Biological*Social, twins)
> summary(g)
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.8720    17.8083   -0.11     0.92
Biological               0.9776     0.1632    5.99    6e-06
Sociallow                9.0767    24.4487    0.37     0.71
Socialmiddle             2.6881    31.6042    0.09     0.93
Biological.Sociallow    -0.0291     0.2446   -0.12     0.91
Biological.Socialmiddle -0.0050     0.3295   -0.02     0.99

Residual standard error: 7.92 on 21 degrees of freedom
```

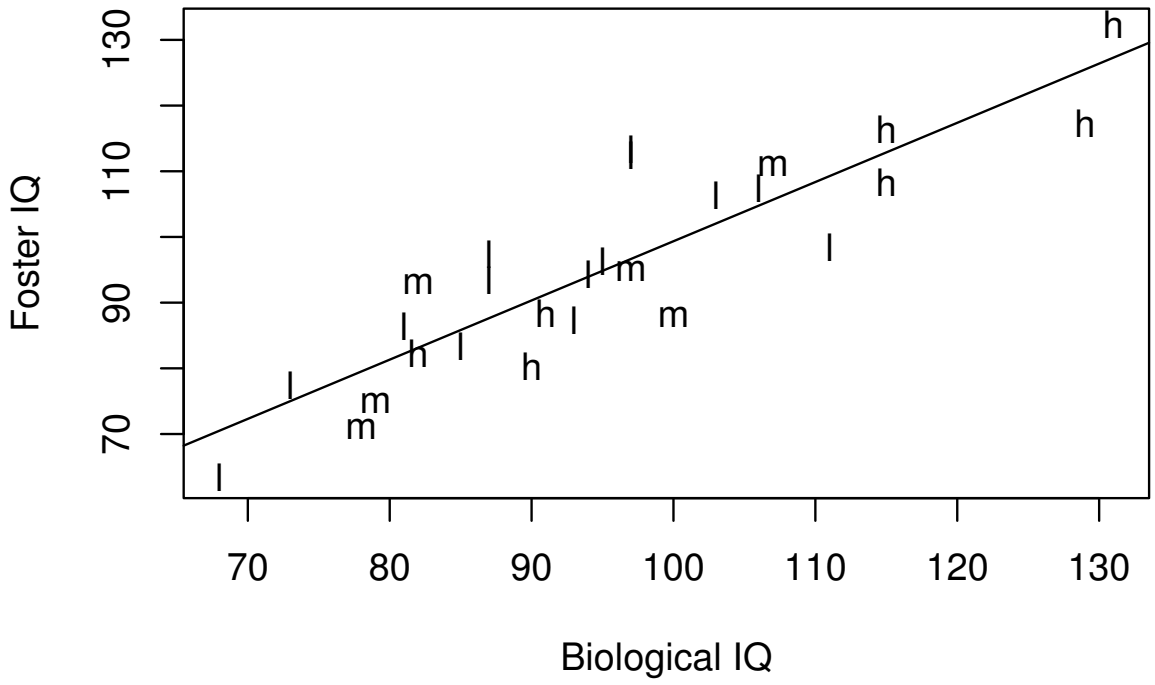

Figure 15.3: Burt twin data, l=low class, m=middle class and h= high class, regression fit shown

```
Multiple R-Squared: 0.804,      Adjusted R-squared: 0.757
F-statistic: 17.2 on 5 and 21 degrees of freedom,       p-value: 8.31e-07
```

The reference level is high class, being first alphabetically. We see that the intercept for low class line would be -1.872+9.0767 while the slope for the middle class line would be 0.9776-0.005. Check the design matrix for the gory details

```
> model.matrix(g)
```

Now see if the model can be simplified to the parallel lines model:

```
> gr <- lm(Foster ~ Biological+Social, twins)
> anova(gr,g)
Analysis of Variance Table

Model 1: Foster ~ Biological + Social
Model 2: Foster ~ Biological + Social + Biological:Social
  Res.Df Res.Sum Sq Df Sum Sq F value Pr(>F)
1     23       1318
2     21       1317  2      1    0.01    1.0
```

Yes it can. The sequential testing can be done in one go:

```
> anova(g)
Analysis of Variance Table
```

```
Response: Foster
                  Df Sum Sq Mean Sq F value  Pr(>F)
Biological         1   5231    5231   83.38 9.3e-09
Social             2    175      88    1.40    0.27
Biological:Social  2      1 4.7e-01    0.01    0.99
Residuals         21   1317      63
```

We see that a further reduction to a single line model is possible:

```
> gr <- lm(Foster ~ Biological, twins)
```

Plot the regression line on the plot:

```
> abline(gr$coef)
```

A check of the diagnostics shows no cause for concern. The (almost) final model:

```
> summary(gr)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.2076     9.2999    0.99     0.33
Biological    0.9014     0.0963    9.36  1.2e-09

Residual standard error: 7.73 on 25 degrees of freedom
Multiple R-Squared: 0.778,      Adjusted R-squared: 0.769
F-statistic: 87.6 on 1 and 25 degrees of freedom,       p-value: 1.2e-09
```

The icing on the cake would be a further simplification of this model to the line y=x (the IQ's are equal). The model has no parameters at all so it has $RSS = \sum_i (y_i - x_i)^2$ and degrees of freedom equal to the sample size. We compute the F-test and p-value:

```
> sum(gr$res^2)
[1] 1493.5
> sum((twins$F-twins$B)^2)
[1] 1557
> ((1557-1493.5)/2)/(1493.5/25)
[1] 0.53147
> 1-pf(0.53147,2,25)
[1] 0.59423
```

So the null is not rejected.

Burt was interested in demonstrating the importance of heredity over environment in intelligence and this data certainly point that way. (Although it would be helpful to know the social class of the foster parents)

However, before jumping to any conclusions, you may be interested to know that there is now considerable evidence that Cyril Burt invented some of his data on identical twins. In light of this, can you see anything in the above analysis that might lead one to suspect this data?