10

Analysis of Covariance

Analysis of covariance involves a combination of regression and analysis of variance. The response variable is continuous, and there is at least one continuous explanatory variable and at least one categorical explanatory variable. Typically, the maximal model involves estimating a slope and an intercept (the regression part of the exercise) for each level of the categorical variable(s) (the Anova part of the exercise). Let's take a concrete example. Suppose we are modelling weight (the response variable) as a function of gender and age. Gender is a factor with two levels (male and female) and age is a continuous variable. The maximal model therefore has four parameters: two slopes (a slope for males and a slope for females) and two intercepts (one for males and one for females) like this:

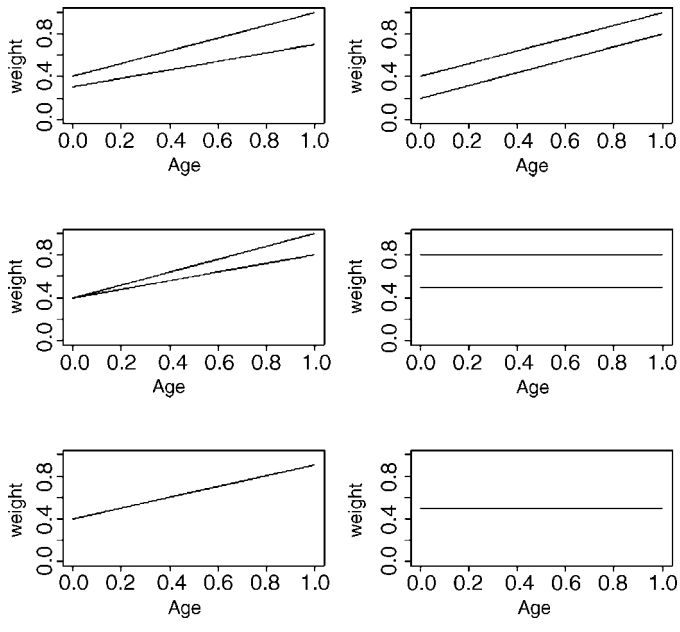
$$weight_{\text{male}} = a_{\text{male}} + b_{\text{male}} \times age$$
$$weight_{\text{female}} = a_{\text{female}} + b_{\text{female}} \times age.$$

Model simplification is an essential part of analysis of covariance, because the principle of parsimony requires that we keep as few parameters in the model as possible.

There are six possible models in this case, and the process of model simplification begins by asking whether we need all four parameters (top left). Perhaps we could make do with two intercepts and a common slope (top right). Or a common intercept and two different slopes (centre left). There again, age may have no significant effect on the response, so we may only need two parameters to describe the main effects of gender on weight; this would show up as two separated, horizontal lines in the plot (one mean weight for each gender; centre right). Alternatively, there may be no effect of gender at all, in which case we only need two parameters (one slope and one intercept) to describe the effect of age on weight (bottom left). In the limit, neither the continuous nor the categorical explanatory variables might have any significant effect on the response, in which case, model simplification will lead to the one-parameter null model $\hat{y} = \bar{y}$ (a single, horizontal line – bottom right).

Decisions about model simplification are based on the explanatory power of the model: if the simpler model does not explain significantly less of the variation in the response, then the simpler model is preferred. Tests of explanatory power are carried out using

Statistics: An Introduction using R M. J. Crawley
 ISBNs: 0-470-02298-1 (PBK); 0-470-02297-3 (PPC)



anova to compare two models: we only retain the more complicated model if the p value from the Anova comparing the two models is less than 0.05.

Let's see how this all works by investigating a realistic example. The dataframe concerns an experiment on a plant's ability to regrow and produce seeds following grazing. The initial, pre-grazing size of the plant is recorded as the diameter of the top of its rootstock. Grazing is a two-levels factor: grazed or ungrazed (protected by fences). The response is the weight of seeds produced per plant at the end of the growing season. Our expectation is that big plants will produce more seeds than small plants and that grazed plants will produce fewer seeds than ungrazed plants. Let's see what actually happened:

```
compensation<-read.table("c:\\temp\\compensation.txt",header=T)
attach(compensation)
names(compensation)
```

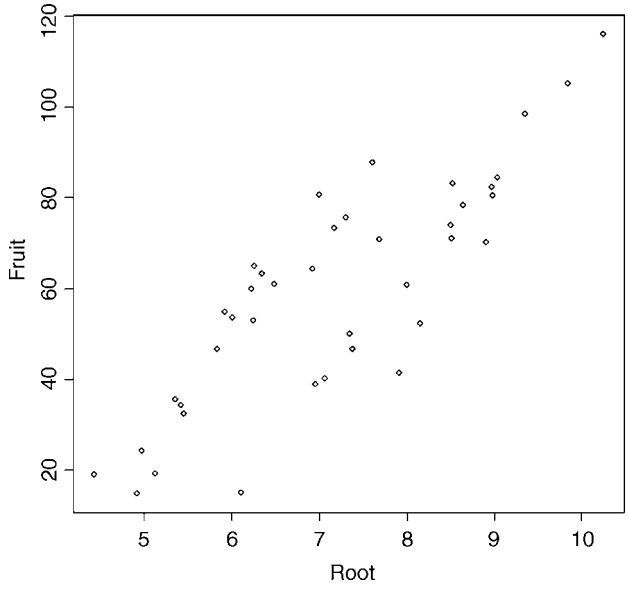
```
[1] "Root"    "Fruit"    "Grazing"
```

We begin with data inspection. First, did initial plant size matter?

Yes it did. Plants which were bigger to begin with produced more seeds at the end of the growing season. What about grazing?

```
plot(Grazing,Fruit)
```

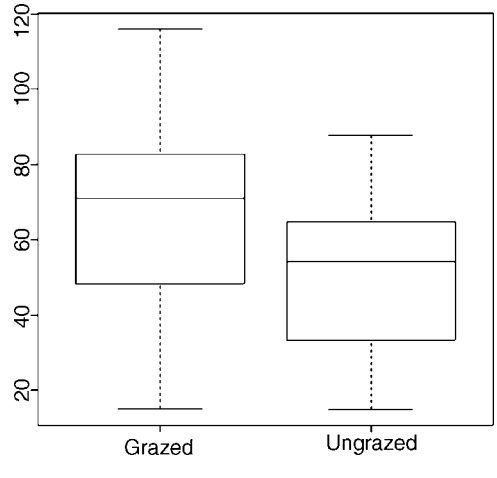
This is not at all what we expected to see. Apparently, the grazed plants produced **more** seeds, not less than the ungrazed plants. We shall return to this after we have carried out



the statistical modelling. Analysis of covariance is done in the familiar way – it is just that the explanatory variables are a mixture of continuous and categorical variables. We start by fitting the most complicated model, with different slopes and intercepts for the grazed and ungrazed plants. For this, we use the asterisk operator:

```
model <-lm(Fruit ~ Root*Grazing)
```

An important thing to realize about analysis of covariance is that '**order matters**'. Look at the regression sum of squares in the Anova table when we fit root first:



```
summary.aov(model)
```

```
              Df  Sum Sq  Mean Sq  F value    Pr(>F)
Root           1 16795.0  16795.0 359.9681  <2.2e-16  ***
Grazing        1  5264.4   5264.4 112.8316 1.209e-12  ***
Root:Grazing   1     4.8      4.8   0.1031      0.75
Residuals     36  1679.6     46.7
```

and when we fit root second:

```
model < -lm(Fruit ~ Grazing*Root)
summary.aov(model)
```

```
              Df  Sum Sq  Mean Sq  F value    Pr(>F)
Grazing        1  2910.4   2910.4  62.3795 2.262e-09  ***
Root           1 19148.9  19148.9 410.4201  <2.2e-16  ***
Grazing:Root   1     4.8      4.8   0.1031      0.75
Residuals     36  1679.6     46.7
```

In both cases, the error sum of squares (1679.6) and the interaction sum of squares (4.8) are the same, but the regression sum of squares (labelled 'root') is much greater when root is fitted to the model after grazing (19 148.9), than when it is fitted first (16 795.0). This is because the data for Ancova are typically non-orthogonal. Remember, **with non-orthogonal data, order matters** (Box 10.1).

Box 10.1. Corrected sums of squares in analysis of covariance

The total sum of squares, *SSY*, and the treatment sums of squares, *SSA*, are calculated in the same way as in a straightforward analysis of variance (Box 9.1). The sums of squares for the separate regressions within the individual factor levels, i, are calculated as shown in Box x.x: $SSXY_i$, SSX_i, SSR_i, and SSE_i.

$$SSXY_{\text{total}} = \sum SSXY_i$$
$$SSX_{\text{total}} = \sum SSX_i$$
$$SSR_{\text{total}} = \sum SSR_i.$$

Then the overall regression sum of squares, *SSR*, is calculated from the total corrected sums of products and the total corrected sums of squares of x:

$$SSR = \frac{(SSXY_{\text{total}})^2}{SSX_{\text{total}}}.$$

The difference in the two estimates, SSR and SSR_{total} is called $SSR_{\text{difference}}$ and is a measure of the significance of the differences between the regression slopes. Now we can compute SSE by difference:

$$SSE = SSY - SSA - SSR - SSR_{\text{difference}},$$

but SSE is defined for the k levels in which the regressions were computed as

$$SSE = \sum_{i=1}^{k} \sum (y - a_i - b_i x)^2.$$

Back to the analysis. The interaction, $SSR_{\text{difference}}$ representing differences in slope between the grazed and ungrazed treatments, appears to be insignificant, so we remove it:

```
model2 <- lm(Fruit ~ Grazing + Root)
```

Notice the use of + rather than * in the model formula. This says 'fit different intercepts for grazed and ungrazed plants, but fit the same slope to both graphs'. Does this simpler model have significantly lower explanatory power? We use Anova to find out:

```
anova(model,model2)
```

```
Analysis of Variance Table

Model 1: Fruit ~ Grazing + Root + Grazing:Root
Model 2: Fruit ~ Grazing + Root

  Res.Df      RSS   Df   Sum of Sq       F   Pr(>F)
1     36  1679.65
2     37  1684.46   -1      -4.81  0.1031     0.75
```

The simpler model does not have significantly lower explanatory power ($p = 0.75$), so we adopt it. Note that we did not have to do the anova in this case: the p value given in the summary.aov(model) table gave the correct, deletion p value. Here are the parameter estimates from our minimal adequate model:

```
summary.lm(model2)
```

```
Coefficients:
                  Estimate Std. Error    t value    Pr(>|t|)
(Intercept)       -127.829      9.664     -13.23    1.33e-15    ***
GrazingUngrazed     36.103      3.357      10.75    6.11e-13    ***
Root                23.560      1.149      20.51      <2e-16    ***

Residual standard error: 6.747 on 37 degrees of freedom
Multiple R-Squared: 0.9291,      Adjusted R-squared: 0.9252
F-statistic: 242.3 on 2 and 37 DF, p-value:         0
```

The model has high explanatory power, accounting for more than 90% of the variation in seed production (multiple r^2). The hard thing about analysis of covariance is understanding what the parameter estimates mean. Starting at the top, the first row, as labelled, contains an intercept. It is the intercept for the graph of seed production against initial rootstock size for the grazing treatment **whose factor level comes first in the alphabet**. To see which one this is, we can use levels:

```
levels(Grazing)
```

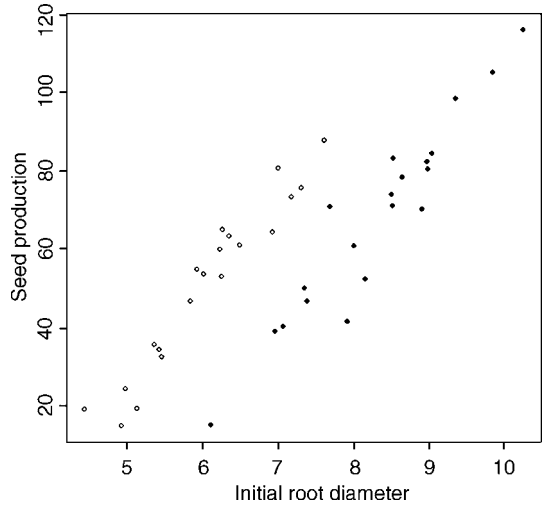
```
[1] "Grazed"    "Ungrazed"
```

So the intercept is the intercept for the grazed plants. The second row, labelled 'GrazingUngrazed' is **a difference between two intercepts**. To get the intercept for the ungrazed plants, we need to add 36.103 to the intercept for the grazed plants ($-127.829 + 36.103 = -91.726$). The third row, labelled Root, is **a slope**: it is the gradient of the graph of seed production against initial rootstock size, and it is the same for both grazed and ungrazed plants. If there had been a significant interaction term, this would have appeared in row four as **a difference between two slopes**.

We can now plot the fitted model through the scatterplot. It will be useful to have different plotting symbols for the grazed and ungrazed plants, and the function called split comes into its own in such cases:

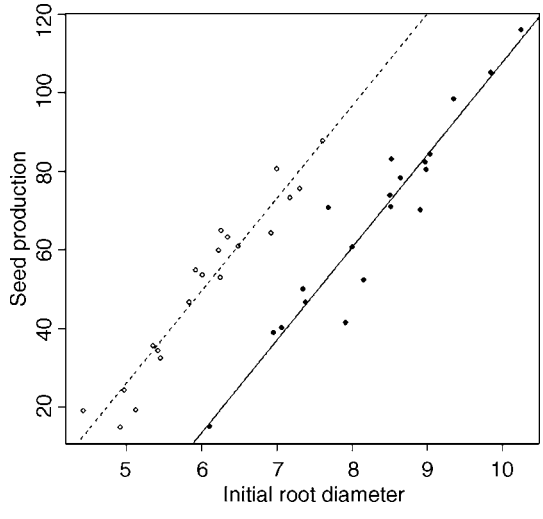
```
sf <- split(Fruit,Grazing)
sr <- split(Root,Grazing)
plot(Root,Fruit,type = "n",ylab = "Seed production",xlab = "Initial root
    diameter")
points(sr[[1]],sf[[1]],pch = 16)
points(sr[[2]],sf[[2]])
```

The double-bracketed subscripts on *sr* and *sf* are used because these two objects are lists rather than vectors. They are lists into which the points relating to the two grazing levels were separated by the split function.



With this plot, it becomes clear why we got the curious result at the beginning (in which grazing appeared to increase seed production). The truth is that the majority of big plants ended up in the grazed treatment (the solid symbols). If you compare like with like (e.g. plants at 7mm initial root diameter) it is clear that the ungrazed plants (open symbols) produced more seed than the grazed plants (36.103 more, to be precise). This will become clearer when we fit the lines predicted by model2:

```
abline(–127.829,23.56)
abline(–127.829 + 36.103,23.56,lty = 2)
```



This example shows the great strength of analysis of covariance. By controlling for initial plant size, we have completely reversed the interpretation. The naïve first impression was that grazing increased seed production:

```
tapply(Fruit,Grazing,mean)
```

```
 Grazed  Ungrazed
67.9405   50.8805
```

and this was significant if we were rash enough to fit grazing on its own ($p = 0.027$):

```
summary(aov(Fruit ~ Grazing))
```

```
            Df   Sum Sq  Mean Sq  F value    Pr(>F)
Grazing      1   2910.4   2910.4   5.3086  0.02678  *
Residuals   38 20833.4    548.2
```

However, when we do the correct analysis of covariance, we find the opposite result: grazing significantly **reduces** seed production for plants of comparable initial size, e.g. from 77.46 to 41.36 at mean rootstock size:

```
-127.829 + 36.103 + 23.56*mean(Root)
```

```
[1] 77.4619
```

```
-127.829 + 23.56*mean(Root)
```

```
[1] 41.35889
```

The moral is clear. When you have covariates (like initial size in this example), then use them. This can do no harm, because if the covariates are not significant, they will drop out during model simplification. Also remember that in Ancova, **order matters**. So always start model simplification by removing the highest-order interaction terms first. In Ancova, these interaction terms are **differences between slopes** for different factor levels (recall that in multi-way Anova, the interaction terms were differences between means). Other Ancovas are described in Chapters 13, 14 and 16 in the context of count data, proportion data and binary response variables.