MAS programmes - Statistical Analysis (Autumn Term)

5 Regression using Dummy Variables

In all the examples we have seen to date, we have had a quantitative response variable, necessarily due to the normality assumption, but our explanatory variables have been quantitative (continuous) measurements also. In this Chapter we consider (multiple) linear regression, where one or more explanatory variables are qualitative *factors*. We begin by considering the case of a dichotomous variable, i.e. one which can take only two values, say 0 or 1.

5.1 Example 1: Mortality and Water Hardness

Consider the following example, of the annual mortality rate per 100,000 males, averaged over the years 1958-1964, and the calcium concentration (parts per million) in the drinking water supply for 61 large towns in England and Wales. (The higher the calcium concentration, the harder the water). The data are available in the data file `water1.csv` on Moodle, and can be entered into R using `read.csv()` as the data frame `water`. The first five rows of the resulting data frame are reproduced below.

```
> water[c(1:5), ]
        town mortality calcium north
1       Bath      1247     105     0
2 Birkenhead      1668      17     1
3 Birmingham      1466       5     0
4  Blackburn      1800      14     1
5  Blackpool      1609      18     1
```

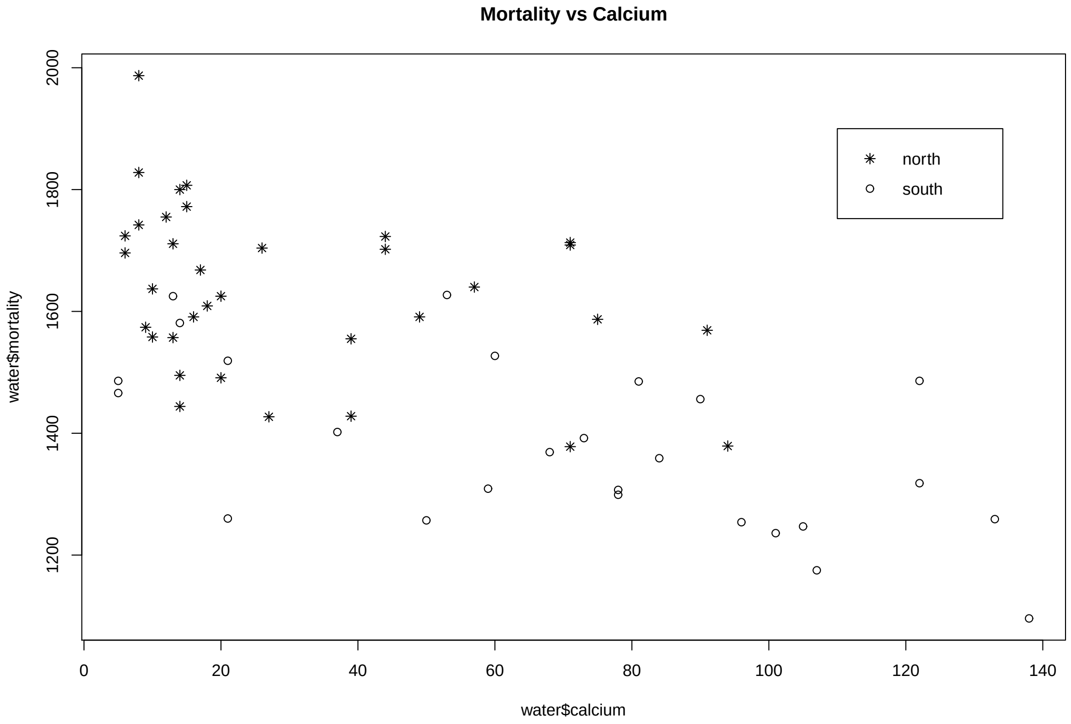
The data can be plotted as follows:

```
> plot(water$calcium, water$mortality, type = "n", main = "Mortality vs Calcium")
> points(water$calcium[water$north==0], water$mortality[water$north==0], pch=1)
> points(water$calcium[water$north==1], water$mortality[water$north==1], pch=8)
> legend(110, 1900, c("north", "south"), pch =c(8,1))
```

Notice from the plot (on the following page) that the observations can be split into two, depending on whether the town is in the ‘North’ (at least as far north as Derby, `north=1`) or the ‘South’ (`north=0`). There does appear to be a clear (possibly linear) relationship between mortality rate and calcium levels (the mortality rate appears to reduce with increasing calcium concentration). Notice also that the mortality rate is generally higher in the North than the South, which is associated with harder water. [Of course we are not suggesting that the increased mortality in the north is due to areas of soft water, but that there simply appears to be an association between mortality and calcium concentration].

Model A: Fitting a simple linear regression

```
> water.lmA <- lm(mortality ˜ calcium, data = water)
> summary(water.lmA)

Call:
lm(formula = mortality ˜ calcium, data = water)

Residuals:
```

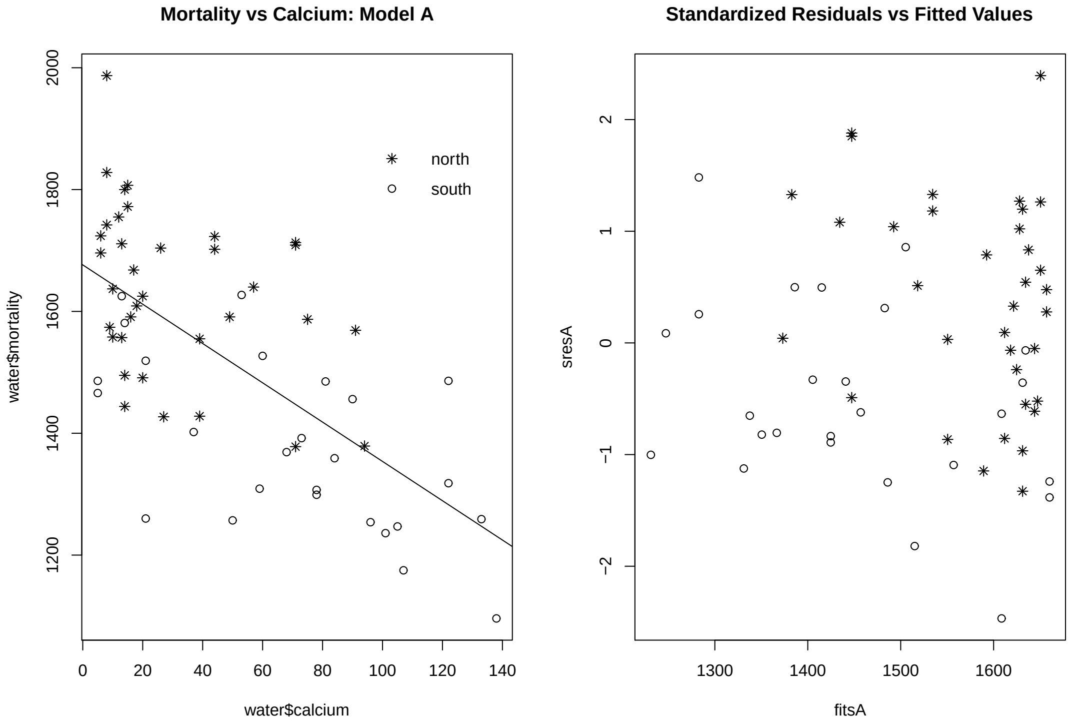


```
    Min      1Q  Median      3Q     Max
-348.61 -114.52   -7.09  111.52  336.45

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1676.3556    29.2981  57.217  < 2e-16 ***
calcium       -3.2261     0.4847  -6.656 1.03e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 143 on 59 degrees of freedom
Multiple R-squared:  0.4288,    Adjusted R-squared:  0.4191
F-statistic:  44.3 on 1 and 59 DF,  p-value: 1.033e-08
```

Plots showing the fitted line and the standardized residuals against fitted values are shown opposite.

```
> par(mfrow = c(1, 2))
> plot(water$calcium, water$mortality, type = "n", main = "Mortality vs Calcium: Model A")
> points(water$calcium[water$north == 0], water$mortality[water$north == 0], pch = 1)
> points(water$calcium[water$north == 1], water$mortality[water$north == 1], pch = 8)
> legend(90, 1900, c("north", "south"), pch =c(8,1), bty="n")
> abline(water.lmA)
> library(MASS)
> sresA <-  stdres(water.lmA)
> fitsA <- fitted(water.lmA)
>
> plot(fitsA, sresA, type = "n", main = "Standardized Residuals vs Fitted Values")
> points(fitsA[water$north == 0], sresA[water$north == 0], pch = 1)
> points(fitsA[water$north == 1], sresA[water$north == 1], pch = 8)
```

Notice from the residual plot that the model does not appear to adequately fit both groups of towns (North and South). Since the towns in the North are likely to have greater mortality, so



their residuals are more likely to be positive than those from the South. That is, on average, this simple linear regression model appears to underestimate the mortality rate in the North, whilst overestimating that in the South.

A natural question to ask is whether there are actually *separate* regression relationships (between `mortality` and `calcium`) for towns in the North and in the South. If so, is there a means by which we can test for the need for separate intercept and slope parameters.

We can achieve this by including the 'dummy variable' `north` in the model which takes the value 0 if the town is in the South (`north=0`), and 1 if the town is in the North (`north=1`).

5.2 Dummy Variable Regression: A single Dichotomous Factor

Consider the following model:

$$y_i = \alpha + \beta x_i + \gamma z_i + \epsilon_i \quad i = 1, \ldots, n \tag{5.1}$$

with $\epsilon_i \sim \text{NID}(0, \sigma^2)$. Note that (x_i, y_i) are the usual observations of the ith individual on the explanatory and response variables x and y respectively. The variable z_i is an *indicator* or *dummy* variable, which takes the following values

$$z_i = \begin{cases} 0, & \text{subject/unit } i \text{ belongs to group A} \\ 1, & \text{subject/unit } i \text{ belongs to group B} \end{cases}$$

The model (5.1) can hence be separated into two models, according to the group membership (or *category*) of i. For group/category A ($z_i = 0$), we have

$$y_i = \alpha + \beta x_i + \epsilon_i$$

and, for group/category B ($z_i = 1$)

$$y_i = (\alpha + \gamma) + \beta x_i + \epsilon_i$$

Hence, the parameter γ represents the difference in intercept value between groups/categories A and B. That is, model (5.1) has separate intercepts (but common slope) for the two groups, and this is achieved by adding the qualitative explanatory variable (or *factor*) z_i to the model, which has two categories relating to the group membership, A ($z_i = 0$) or B ($z_i = 1$). If the term γ is found to be significant in the model, then the need for a separate intercept for each group is evidenced.

Model (5.1) above allows for separate intercepts between groups, but an obvious extension allows us to include separate (i.e. *non-parallel*) slopes for each group.

$$y_i = \alpha + \beta x_i + \gamma z_i + \delta x_i z_i + \epsilon_i \qquad i = 1, \ldots, n \tag{5.2}$$

where again z_i is a 0/1 indicator variable representing group membership. In this setting we have, for group A ($z_i = 0$)

$$y_i = \alpha + \beta x_i + \epsilon_i$$

and for group B ($z_i = 1$)

$$y_i = (\alpha + \gamma) + (\beta + \delta) x_i + \epsilon_i$$

i.e. we have simple linear regressions of y on x for each of the two groups. In this case the regressor $x_i z_i$ represents the *interaction* term between x_i and z_i. We say that two variables *interact* if the partial effect of one depends on the value of the other. In this case the parameter δ represents the difference in slopes between group A (control, $z_i = 0$) and group B in the combined regression model.

Aside: Note that although we have discussed this in the setting of the simple regression model, we have in each of the cases (5.1) and (5.2) a multiple regression model. For example, model (5.2) can be written

$$y_i = \beta_0 + \beta_1 \texttt{expl}_{i1} + \beta_2 \texttt{ind}_{i2} + \beta_3 \texttt{ind}_{i2}\texttt{expl}_{i1} + \epsilon_i \quad i = 1, \ldots, n$$

or

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i \quad i = 1, \ldots, n$$

where x_{i1} is the single (quantitative) explanatory variable, x_{i2} is the indicator variable, and $x_{i3} = x_{i1} x_{i2}$ is the *interaction* term, so that $\beta_0 = \alpha$, $\beta_1 = \beta$, $\beta_2 = \gamma$ and $\beta_3 = \delta$. Extensions to allow for further explanatory variables and factors are obvious. This approach is the basis of the *general linear model*, which allows a multiple regression model with both quantitative and qualitative explanatory variables (factors).

5.3 Example 1 continued ...

We return now to the water mortality data and consider models which accommodate individual regression lines for towns in the North and South. Models B and C represent models with separate intercepts (model (5.1)) and non-parallel slopes (model (5.2)) respectively.

Model B

```
> water.lmB <- lm(mortality ~ calcium + north, data = water)
> summary(water.lmB)

Call:
lm(formula = mortality ~ calcium + north, data = water)

Residuals:
     Min       1Q   Median       3Q      Max
-222.959  -77.281    7.143   90.751  307.836

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1518.7263    41.3350  36.742  < 2e-16 ***
calcium       -2.0341     0.4829  -4.212 8.93e-05 ***
north        176.7108    36.8913   4.790 1.19e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 122.1 on 58 degrees of freedom
Multiple R-squared:  0.5907,    Adjusted R-squared:  0.5766
F-statistic: 41.86 on 2 and 58 DF,  p-value: 5.601e-12
```

Again the fitted regression lines and (standardized) residual plots can be assessed:

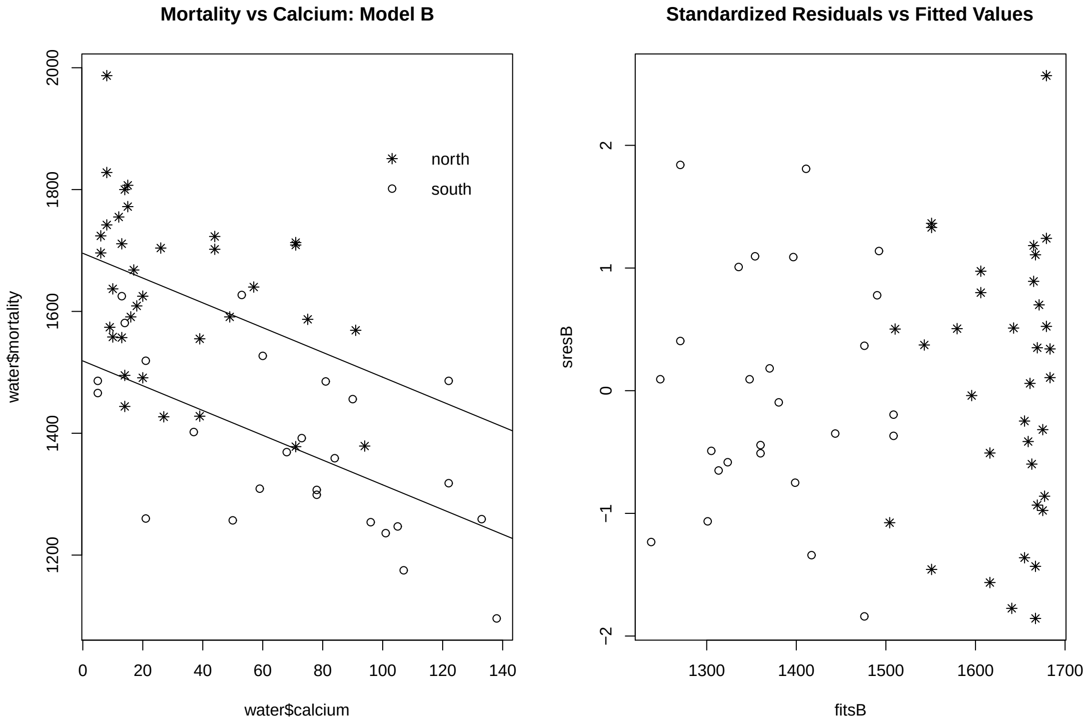
```
> plot(water$calcium, water$mortality, type = "n", main = "Mortality vs Calcium: Model B")
> points(water$calcium[water$north == 0], water$mortality[water$north == 0], pch = 1)
> points(water$calcium[water$north == 1], water$mortality[water$north == 1], pch = 8)
> legend(90, 1900, c("north", "south"), pch =c(8,1), bty="n")
> ld <- seq(0, 145, 0.1)
> lines(ld, predict(water.lmB, data.frame(calcium = ld, north = rep(0, length(ld))),
                    type = "response"))
> lines(ld, predict(water.lmB, data.frame(calcium = ld, north = rep(1, length(ld))),
                    type = "response"))

> sresB <- stdres(water.lmB)
> fitsB <- fitted(water.lmB)
> plot(fitsB, sresB, type = "n", main = "Standardized Residuals vs Fitted Values")
> points(fitsB[water$north == 0], sresB[water$north == 0], pch = 1)
> points(fitsB[water$north == 1], sresB[water$north == 1], pch = 8)
```

Here the model appears to be more appropriate, with mortality levels predicted to be higher in the North, for a similar level of water hardness (calcium concentration). The residual plot is much more acceptable with standardized residuals for towns from both the North and South evenly spread about zero. The better fit of the model is confirmed by the R^2 values from the output. Model B gives a multiple R^2 value of 0.5907 which is better than the value of 0.4288 returned by model A, which also has a higher value of model standard deviation (143 versus 122.1). The coefficient of `north` in Model B is also highly significant, indicating that the intercepts between the two groups of towns are different.

Finally, we consider whether there is a case for separate slope parameters too (non-parallel regression lines). Note that the interaction term is specified in R by adding `north:calcium` to the model formula.

Model C

```
> water.lmC <- lm(mortality ~ calcium + north + north:calcium, data = water)
```



```
> summary(water.lmC)

Call:
lm(formula = mortality ~ calcium + north + north:calcium, data = water)

Residuals:
    Min      1Q  Median      3Q     Max
-221.27  -75.45    8.52   87.48  310.14

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1522.8150    48.9500  31.110  < 2e-16 ***
calcium         -2.0927     0.6103  -3.429  0.00113 **
north          169.4978    58.5916   2.893  0.00540 **
calcium:north    0.1614     1.0127   0.159  0.87395
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 123.2 on 57 degrees of freedom
Multiple R-squared:  0.5909,    Adjusted R-squared:  0.5694
F-statistic: 27.44 on 3 and 57 DF,  p-value: 4.102e-11
```

The output shows that there is no evidence for non-parallel lines, since the coefficient corresponding to the interaction term `north:calcium` is not significant ($p = 0.874$), so that Model B is preferred.

We now consider the case of regression where more than two groups are represented within the observations.

5.4 Dummy Variable Regression: Polytomous Factors

The process described above of using an indicator (or dummy) variable to represent group membership (or *factor level*) is easily extended to situations where a factor has more than two levels. Let us first of all consider the case of *three* levels. The separate intercepts regression model may now be written

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{i1} + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 z_{i3} + \epsilon_i \qquad i = 1, \ldots, n, \\ &= (\alpha + \gamma_j) + \beta_1 x_{i1} + \epsilon_i \qquad i = 1, \ldots, n;\ j = 1, 2, 3. \end{aligned} \tag{5.3}$$

where z_{ij}, $j = 1, 2, 3$ is the dummy regressor indicating whether an observation belongs to the jth category or factor level.

$$z_{ij} = \begin{cases} 1 & \text{if obs } i \text{ belongs to category } j \\ 0 & \text{otherwise} \end{cases}$$

One immediate problem is that this model is over-parameterized, with four parameters (α, γ_1, γ_2 and γ_3) to represent three group intercepts, and we would not be able to find unique estimates for these four parameters. This is because the γ_j parameters are associated with the dummy variables z_1, z_2 and z_3 which are perfectly collinear, e.g. $z_3 = 1 - z_1 - z_2$. To accommodate this, we rely on only two dummy variables and select one of the categories to be considered as a baseline (c/f the water mortality example where the South `north=0` was automatically considered as a baseline).

In general, for a polytomous factor with m categories, we code $m - 1$ dummy regressors, selecting the *first* [1] category as the baseline and code $z_{ij} = 1$ when observation i falls into category j, and 0 otherwise.

Category	z_2	z_3	z_4	$\ldots$	z_m
1	0	0	0	$\ldots$	0
2	1	0	0	$\ldots$	0
3	0	1	0	$\ldots$	0
4	0	0	1	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
m	0	0	0	$\ldots$	1

In this way a polytomous factor can be entered into a (multiple) regression by simply coding a set of 0/1 dummy variables, one fewer than the number of categories for that factor. The 'omitted' category, coded 0 for all the dummy regressors in the set, serves as a baseline to which all other categories are compared. Such a model represents parallel regression lines (surfaces), one for each category of factor.

Interactions can be incorporated into the model by simply forming the product terms between the relevant explanatory variable and each of the $m - 1$ dummy variables representing the m-level factor. (The interaction between two factors of k levels and ℓ levels respectively will thus be modelled by the entering of $(k - 1)(\ell - 1)$ terms in the regression model).

[1] This is the default in R, but note that the choice of baseline category is arbitrary. *SAS*, for example, uses the last category as the 'control'. Note also that in both these packages the choice of baseline category is made *alphabetically* (or in numeric order if a factor is described in that way).

5.5 Example 2: Fisher's Iris data

A data set which has been analysed and re-analysed many times concerns 50 plants from each of the three species of iris: *iris setosa*, *iris versicolor* and *iris virginica*. Measurements taken on each plant include the sepal length, sepal width, petal length and petal width (all in mm). The data are available in the file `iris1.csv` and can be entered into R as the data frame `iris`. The first five observations for each species are extracted from the data frame below.

```
iris[c(1:5,51:55,101:105),]
       species sepalL sepalW petalL petalW
1       setosa    5.1    3.5    1.4    0.2
2       setosa    4.9    3.0    1.4    0.2
3       setosa    4.7    3.2    1.3    0.2
4       setosa    4.6    3.1    1.5    0.2
5       setosa    5.0    3.6    1.4    0.2
51  versicolor    7.0    3.2    4.7    1.4
52  versicolor    6.4    3.2    4.5    1.5
53  versicolor    6.9    3.1    4.9    1.5
54  versicolor    5.5    2.3    4.0    1.3
55  versicolor    6.5    2.8    4.6    1.5
101  virginica    6.3    3.3    6.0    2.5
102  virginica    5.8    2.7    5.1    1.9
103  virginica    7.1    3.0    5.9    2.1
104  virginica    6.3    2.9    5.6    1.8
105  virginica    6.5    3.0    5.8    2.2
```

Note that the variable `species` is automatically treated by R as a *factor* with three levels, indicating the species to which each observation belongs. It could alternatively have been added using the following command:

```
> species <- factor(c(rep("setosa",50),rep("versicolor",50),rep("virginica",50)))
```

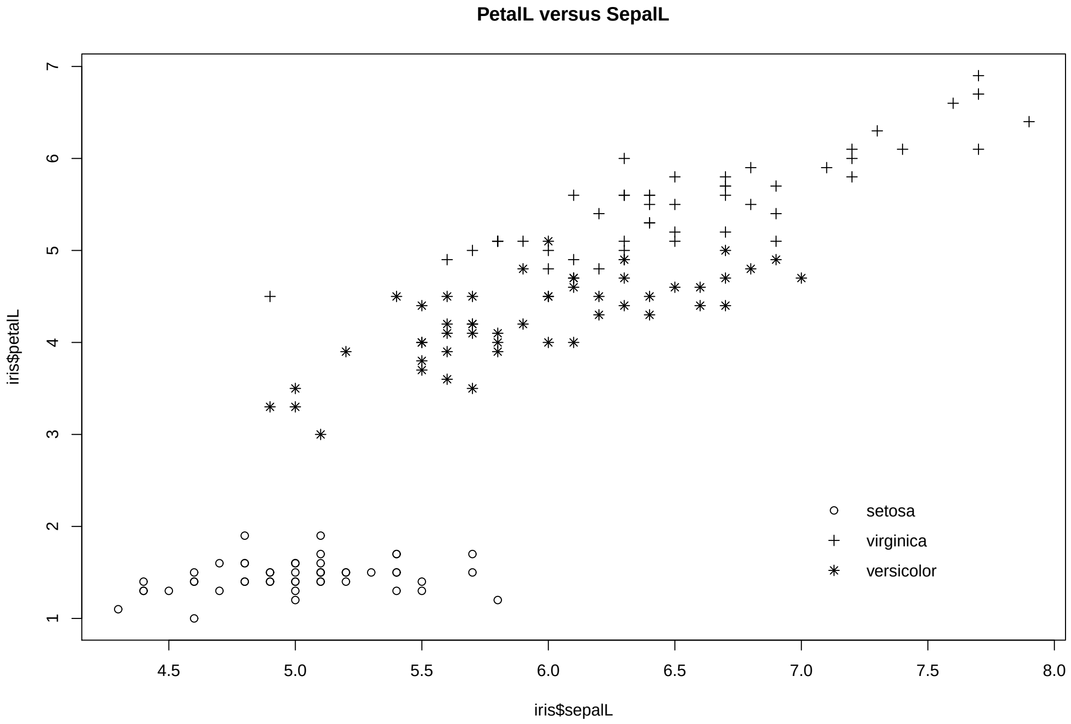
We are concerned with the relationship between petal length (`petalL`) and sepal length (`sepalL`) and whether this relationship is different for each of the species. An initial plot is shown below.

```
> plot(iris$sepalL, iris$petalL, type = "n", main = "PetalL versus SepalL")
> points(iris$sepalL[iris$species == "setosa"], iris$petalL[iris$species == "setosa"], pch = 1)
> points(iris$sepalL[iris$species == "virginica"], iris$petalL[iris$species == "virginica"], pch = 3)
> points(iris$sepalL[iris$species == "versicolor"], iris$petalL[iris$species == "versicolor"], pch = 8)
> legend(7, 2.5, c("setosa", "virginica", "versicolor"), pch =c(1, 3, 8), bty="n")
```

It looks likely that separate intercepts and slopes will be required to provide an adequate fit to these data, and a hierarchy of such models are investigated in R below, as in the case of the water mortality example.

Model A

```
> iris.lmA <- lm(petalL ~ sepalL, data = iris)
> summary(iris.lmA)

Call:
lm(formula = petalL ~ sepalL, data = iris)

Residuals:
```



```
     Min       1Q   Median       3Q      Max
-2.47747 -0.59072 -0.00668  0.60484  2.49512

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.10144    0.50666  -14.02   <2e-16 ***
sepalL       1.85843    0.08586   21.65   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8678 on 148 degrees of freedom
Multiple R-squared:   0.76,     Adjusted R-squared:  0.7583
F-statistic: 468.6 on 1 and 148 DF,  p-value: < 2.2e-16
```

Model B

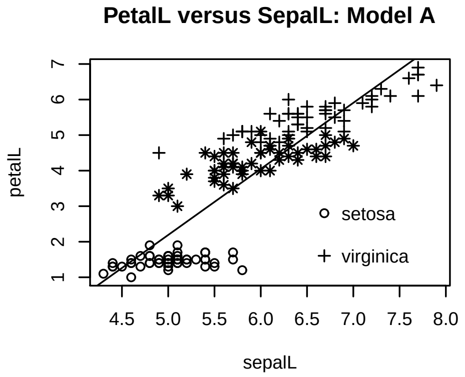
```
> iris.lmB <- lm(petalL ~ sepalL + species, data = iris)
> summary(iris.lmB)

Call:
lm(formula = petalL ~ sepalL + species, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.76390 -0.17875  0.00716  0.17461  0.79954

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.70234    0.23013  -7.397 1.01e-11 ***
sepalL             0.63211    0.04527  13.962  < 2e-16 ***
speciesversicolor  2.21014    0.07047  31.362  < 2e-16 ***
speciesvirginica   3.09000    0.09123  33.870  < 2e-16 ***
---
```

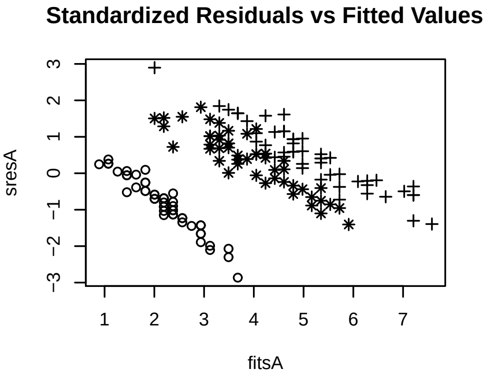
```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2826 on 146 degrees of freedom
Multiple R-squared:  0.9749,    Adjusted R-squared:  0.9744
F-statistic:  1890 on 3 and 146 DF,  p-value: < 2.2e-16
```

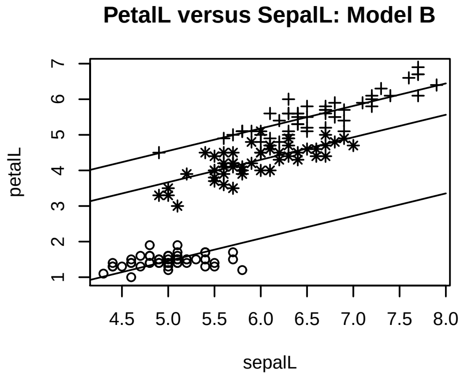
Model C

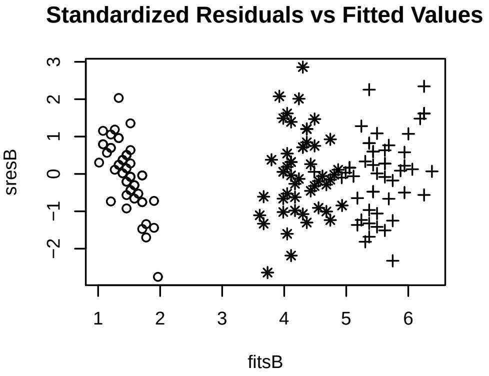
```
> iris.lmC <- lm(petalL ~ sepalL + species + species:sepalL, data = iris)
> summary(iris.lmC)

Call:
lm(formula = petalL ~ sepalL + species + species:sepalL, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.68611 -0.13442 -0.00856  0.15966  0.79607 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                0.8031     0.5310   1.512    0.133    
sepalL                     0.1316     0.1058   1.244    0.216    
speciesversicolor         -0.6179     0.6837  -0.904    0.368    
speciesvirginica          -0.1926     0.6578  -0.293    0.770    
sepalL:speciesversicolor   0.5548     0.1281   4.330 2.78e-05 ***
sepalL:speciesvirginica    0.6184     0.1210   5.111 1.00e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2611 on 144 degrees of freedom
Multiple R-squared:  0.9789,    Adjusted R-squared:  0.9781
F-statistic:  1333 on 5 and 144 DF,  p-value: < 2.2e-16
```

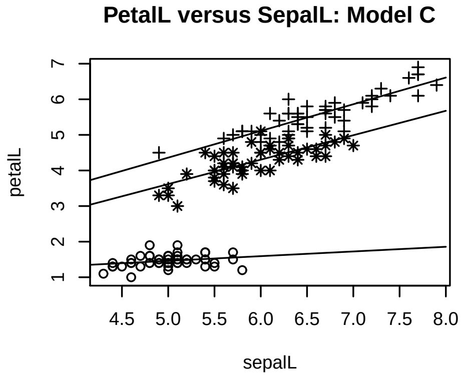
From the summary outputs it appears that Model C is preferred, in that *both* the interaction terms are significant, showing that the slopes for the species `versicolor` and `virginica` are both different to that of the baseline category `setosa`. For each model the fitted line(s) and residual versus fitted value plots are shown on the following page.

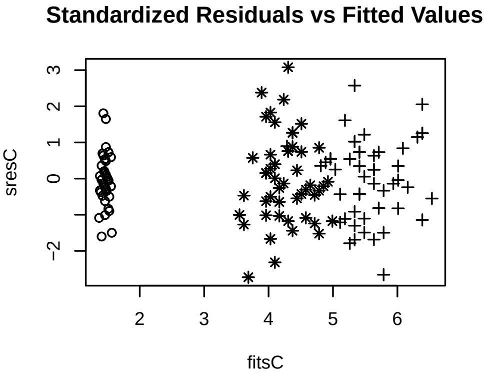

PetalL versus SepalL: Model A
petalL
sepalL
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
setosa
virginica


Standardized Residuals vs Fitted Values
sresA
fitsA
−3
−2
−1
0
1
2
3
1
2
3
4
5
6
7


PetalL versus SepalL: Model B
petalL
sepalL
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

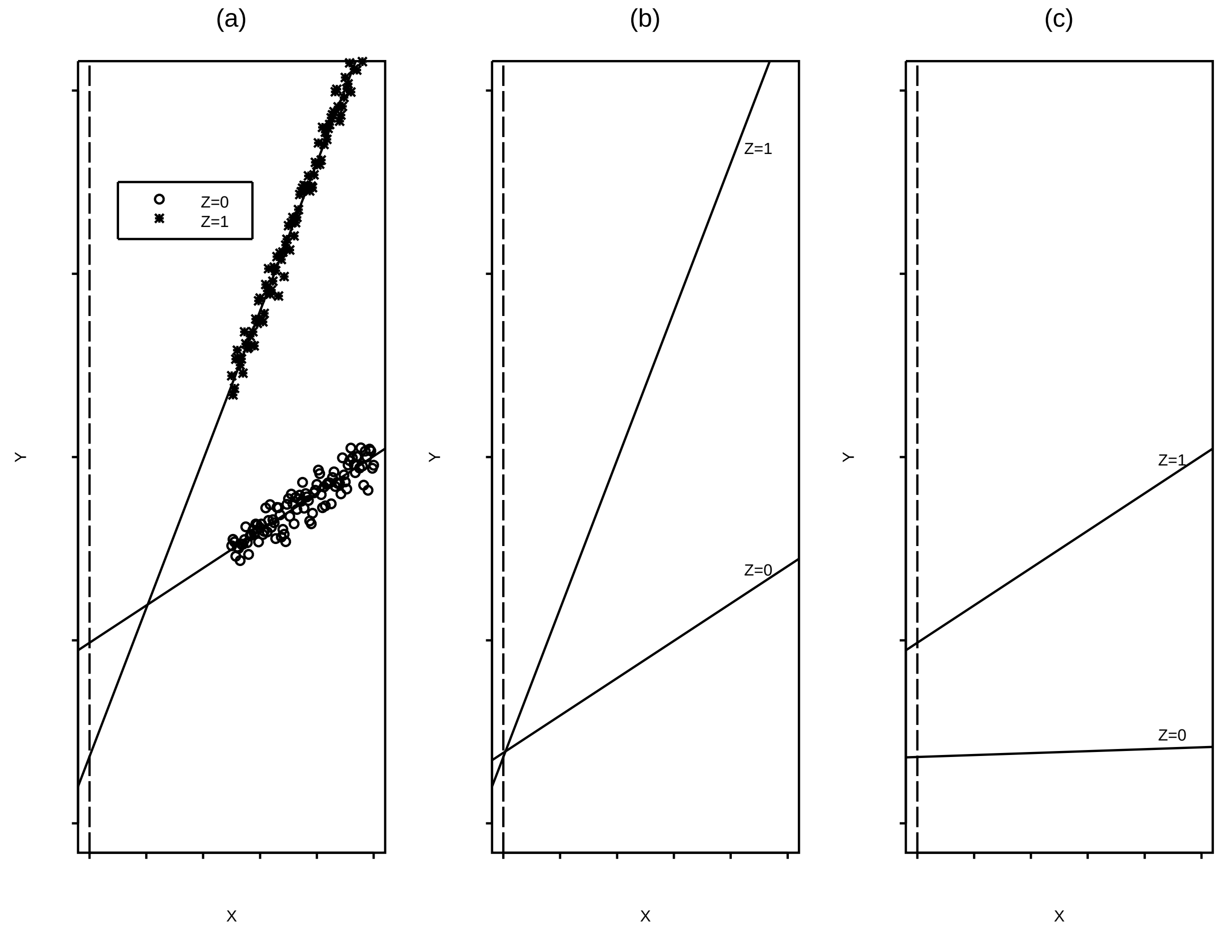

Standardized Residuals vs Fitted Values
sresB
fitsB
−2
−1
0
1
2
3
1
2
3
4
5
6


PetalL versus SepalL: Model C
petalL
sepalL
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0


Standardized Residuals vs Fitted Values
sresC
fitsC
−2
0
1
2
3
2
3
4
5
6

5.6 The Principle of Marginality

An important consideration, when dealing with models containing interactions is that of *marginality*. That is, we describe the separate partial effects, or *main effects* of `species` and `sepalL` say, in the previous `iris` example, to be *marginal* to the `species` by `sepalL` (`species:sepalL`) *interaction*. In general, we do not test or interpret the main effects of explanatory variables that interact. As a corollary to this principle, it does not generally make sense to specify and fit models that include interaction terms (regressors), but omit main effects terms that are marginal to them. That is, the *principle of marginality* specifies that a model including a higher order term (such as an interaction) should normally also include the lower order relatives (i.e. the main effects that compose the interaction).



Consider the diagrams below, each concerning the model $\mathrm{E}(y) = \alpha + \beta x + \gamma z + \delta xz$ (c/f (5.2)).

(a) shows why a difference in intercepts does not represent a meaningful partial effect in the presence of an interaction. The difference γ is negative even though within the range of the data the regression line for the group coded `z=1` is above that for group `z=0`.

Models (b) and (c) both violate the principle of marginality since they include the interaction term xz, but have one of the terms x or z omitted.

(b) The dummy regressor z is omitted from the model, $\mathrm{E}(y) = \alpha + \beta x + \delta xz$.

(c) The quantitative explanatory variable x is omitted, $\mathrm{E}(y) = \alpha + \gamma z + \delta xz$.