Discrete Predictors and Dummy Variables

1 Introduction

In the multiple regression model

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad \forall\, i = 1, \ldots, n$$

it is not necessary for the X's to have any particular distribution. This is a very important observation, for it means that we can always include discrete, or categorical, predictors (X's) in our models for the observations Y. Some discrete x's that might make sense are: gender (M/F) in studies of salaries, region (S, N, MW, W) in studies of urban moral integration, automobile type (compact, midsize, full size, truck) in studies of fuel efficiency, etc.

The main idea of including these predictors in a regression is to introduce one new variable for each distinct value or category of the discrete predictor. For example for the "region" example above, I would construct the following[1] variables

$$\begin{aligned}
D_{i1} &= \begin{cases} 1, & \text{if observation } i \text{ is from the (S)outh} \\ 0, & \text{otherwise} \end{cases} \\
D_{i2} &= \begin{cases} 1, & \text{if observation } i \text{ is from the (N)orth} \\ 0, & \text{otherwise} \end{cases} \\
D_{i3} &= \begin{cases} 1, & \text{if observation } i \text{ is from the midwest (MW)} \\ 0, & \text{otherwise} \end{cases} \\
D_{i4} &= \begin{cases} 1, & \text{if observation } i \text{ is from the (W)est} \\ 0, & \text{otherwise} \end{cases}
\end{aligned}$$

These variables are called "dummy variables" or "indicator variables". I like the second name better, since each variable "indicates" the category of the discrete predictor that was observed.

In this handout we will explore a particular example, to see how indicator variables can be used in practice to explore the possibility that the data should be separated into different groups, and different regression lines fit within each group (although it is often hidden from view, this is a primary reason to use indicator variables).

2 The Data

This data was originally used by Cyril Burt (1966) to explore the "nature vs. nurture" question for intelligence. $n = 27$ twins were found in which one member of the pair was raised in a foster home and one member was raised by the twins' natural parents.

The goal is to use linear regression to explore the relationships between nature (IQ separate from environmental influences such as parental guidance) and nurture (IQ as influenced by parental guidance, by natural parents' SES, etc. The data here are Burt's data as reported by Weisberg (1985) *Applied Linear Regression, 2nd. Ed.*, pp. 180–181.

[1] Although I have constructed four variables for the four levels of "region", in general one only needs $\ell - 1$ such variables for a predictor with ℓ levels.

```
> twins _ read.table("twins.dat")
> twins
   fos.IQ nat.IQ SES
 1     82     82   H
 2     80     90   H
 3     88     91   H
 4    108    115   H
 5    116    115   H
 6    117    129   H
 7    132    131   H
 8     71     78   M
 9     75     79   M
10     93     82   M
11     95     97   M
12     88    100   M
13    111    107   M
14     63     68   L
15     77     73   L
16     86     81   L
17     83     85   L
18     93     87   L
19     97     87   L
20     87     93   L
21     94     94   L
22     96     95   L
23    112     97   L
24    113     97   L
25    106    103   L
26    107    106   L
27     98    111   L
```

`fos.IQ` and `nat.IQ` are the respective IQ's of each pair member. The SES variable is coded "H" for High SES, "M" for medium SES, "L" for low SES. Since we will be doing many regressions and creating a variety of new variables for our regression analysis, it will be useful to attach the data set for the entire session.

```
> attach(twins)
```

Before we charge ahead, let's look at the data. We want to plot all the data, but label the data pairs from different SES's differently. One way to do this is

```
> plot(nat.IQ,fos.IQ,type="n")
> text(nat.IQ,fos.IQ,labels=as.vector(SES))
```

One might have expected a command like `text(nat.IQ,fos.IQ,labels=SES)` to work, but Splus treats non-numeric data specially (as "factor" data), so we cannot get away with that. In the resulting figure, Figure 1, we see that there is a generally increasing relationship among the natural and foster IQ's but that there is some grouping according to SES.

How strong is this grouping of IQ's by SES? As a first step, we can look at boxplots of the IQ's within each SES group.

```
> par(mfrow=c(1,2))
> boxplot(split(nat.IQ,SES),style.bxp="old",ylim=c(65,130))
> boxplot(split(fos.IQ,SES),style.bxp="old",ylim=c(65,130))
```

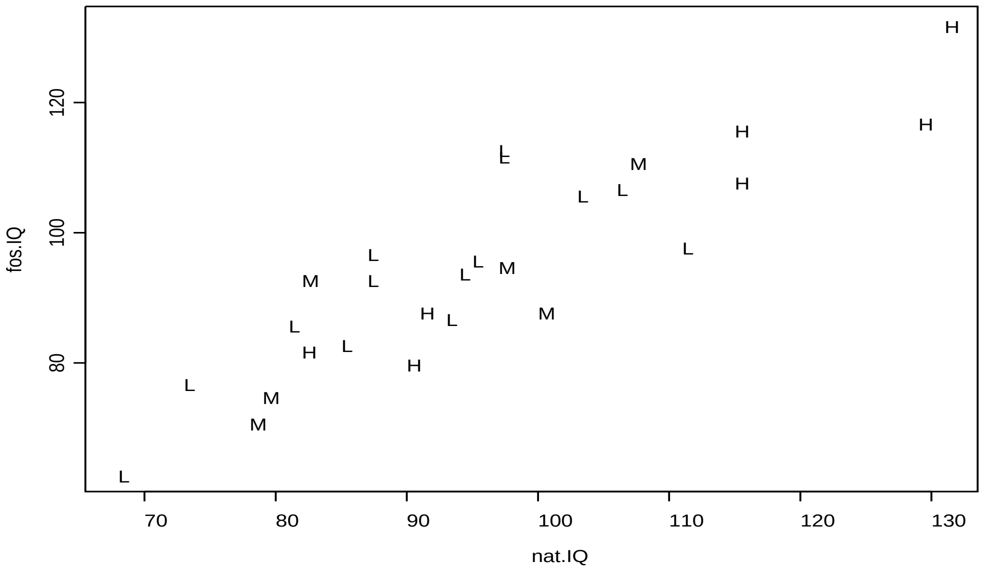

Figure 1: Initial scatter plot of the twins data.

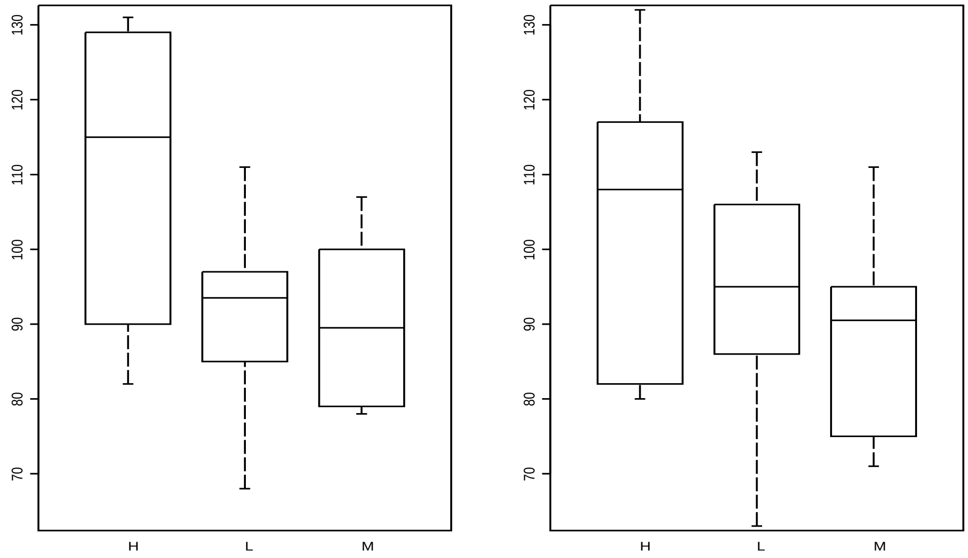

Figure 2: One-Way ANOVA's using Boxplots: Natural IQ vs SES (left) and Foster IQ vs SES (right).

Both sets of boxplots in Figure 2 show a clear separation *between* the median IQ's of the children raised by their natural parents, vs. their foster parents. However the spread *within* each boxplot is large, and it's not clear that the separations between the medians can be explained as genuine separation of the SES groups, or merely effects of variation within a single homogeneous group that does not depend on SES.

3 Exploring the relationship between IQ and SES

This comparison of "between-group" variation in data to "within-group" variation, to determine whether the groups differ substantially from one another, is called *analysis of variance*, or *ANOVA* (not to be confused with the ANOVA tables we have already been using!). Boxplots can be very helpful illustrating an analysis of variance, especially if the IQR's are small enough to visually separate the medians. However boxplots do not take into account the sample sizes within the groups (roughly we should think of dividing each IQR by $\sqrt{\text{group sample size}}$), and a regression of IQ onto the categories of the SES variable can do this for us automatically (this is called a "one-way analysis of variance", since there is one discrete predictor; two discrete predictors lead to two-way ANOVA, three predictors to three-way, etc.).

3.1 One-way Analysis of Variance: Regression on Dummies

Our first problem in conducting regressions analogous to Figure 2 is that SES is a categorical variable and we need to create indicator variables for each of the categories H, L, and M of SES.

```
> DH _ ifelse(SES=="H",1,0)
> DH
 [1] 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> DM _ ifelse(SES=="M",1,0)
> DM
 [1] 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> DL _ ifelse(SES=="L",1,0)
> DL
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Now we conduct the regressions "as usual",

```
> oneway.nat _ lm(nat.IQ ~ DH + DM + DL)
Error in lm.fit.qr(x, y): computed fit is singular, rank 3
```

Ouch—what happened!?

The regression equation we are fitting is

$$nat.IQ = \alpha + \beta_1 DH + \beta_2 DM + \beta_3 DL + \varepsilon$$

and we know that collinearity among the regressors is bad news. For example, the SE's of the $\hat{\beta}_j$'s depend on the variance inflation factors $1/(1 - R_j^2)$, which are large exactly when the multiple correlations

$$R_j^2 = \max_{A,B_1,\ldots,B_k} \{\text{Corr}(X_j, A + B_1 X_1 + \cdots + B_{j-1} X_{j-1} + B_{j+1} X_{j+1} + \cdots + B_k X_k)\}$$

between each X_j and the other X's is large (near one). In our case, each of the dummies can be expressed *exactly* as a constant plus a linear combination of the other dummies, e.g. $DH = 1 - DM - DL$, so all three $R_j^2 \equiv 1$!!!

Since the regressors are perfect linear combinations of each other, it is redundant to include them all so we might as well drop one out of the model. We will drop DL, the indicator of the low-SES group. This is called *imposing a linear constraint* on the regression, since it essentially forces the linear condition $\beta_3 = 0$ on the regression coefficients.

```
> oneway.nat _ lm(nat.IQ ~ DH + DM)
> summary(oneway.nat)

Call: lm(formula = nat.IQ ~ DH + DM)
Residuals:
    Min     1Q Median    3Q   Max
 -25.57 -10.86  2.786 8.464 23.43

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  91.2143   3.8620    23.6181   0.0000
         DH  16.3571   6.6893     2.4453   0.0222
         DM  -0.7143   7.0511    -0.1013   0.9202

Residual standard error: 14.45 on 24 degrees of freedom
Multiple R-Squared: 0.2215
F-statistic: 3.415 on 2 and 24 degrees of freedom, the p-value is 0.04954

Correlation of Coefficients:
   (Intercept)      DH
DH -0.5774
DM -0.5477      0.3162
```

To answer the question "Does SES matter for IQ of twins raised by their natural parents", we may look at the omnibus F-test, $F_{2,24} = 3.415$, which has an observed p-vaule of almost exactly 0.05, so we are moderately confident that there are IQ differences among the SES groups.

3.2 Interpreting the Coefficients

Our estimated regression equation is

$$nat.IQ = 91.214 + (16.3571)DH - (0.7143)DL$$

So the predicted IQ's for children in each SES group raised by their natural parents are

$$\begin{array}{rcll} & = & 91.2143 & \text{if SES=L, i.e. DH=0, DM=0, DL=1} \\ 91.214 - 0.7143 & = & 90.5 & \text{if SES=M, i.e. DH=0, DM=1, DL=0} \\ 91.214 + 16.3571 & = & 107.5714 & \text{if SES=H, i.e. DH=1, DM=0, DL=0} \end{array}$$

There are two interesting things to note here. First, omitting DL from the regression forces a particular interpretation on the coefficients: $\hat{\alpha} = 91.214$ is the predicted IQ among low-SES children, $\hat{\beta}_1 = -0.7143$ is the increment (or decrement) in IQ moving from low to medium SES, and $\hat{\beta}_2 = 16.3571$ is the increment in IQ moving from low to high SES. (We sometimes say SES=L is a *baseline category* and the increments β_1 and β_2 are *contrasts* for the other categories.)

The second thing to note is that the group means within each SES group,

```
> sum(nat.IQ*DH)/sum(DH)
[1] 107.5714
> sum(nat.IQ*DM)/sum(DM)
[1] 90.5
> sum(nat.IQ*DL)/sum(DL)
[1] 91.21429
```

are *the same* as the predicted IQ's from the regression. This is always true in analysis of variance (regression onto categorical variables). What might be a little surprising is that even though the boxplots in Figure 2 overlap quite a bit, the F-test for the regression suggests that the predicted values for the groups should be different (it just barely rejects the null hypothesis model $y = (const) + \varepsilon$). To think about why that it, it is worthwhile to think about the difference between variability in the sample (which causes the boxplots to overlap) and standard error for the mean (which is reduced by a factor of $1/\sqrt{group\ size}$ for each group).

Returning to the first point (imposing a linear constraint on the regression also imposes an interpretation on the parameters) we could have imposed a different constraint, like setting $\alpha = 0$ (dropping the intercept) in the regression:

```
> summary(lm(nat.IQ ~ DH + DM + DL - 1))

Call: lm(formula = nat.IQ ~ DH + DM + DL - 1)
Residuals:
    Min      1Q Median     3Q   Max
 -25.57 -10.86  2.786 8.464 23.43

Coefficients:
      Value Std. Error  t value Pr(>|t|)
DH 107.5714   5.4618    19.6954   0.0000
DM  90.5000   5.8994    15.3406   0.0000
DL  91.2143   3.8620    23.6181   0.0000

Residual standard error: 14.45 on 24 degrees of freedom
Multiple R-Squared: 0.9801
F-statistic: 393.7 on 3 and 24 degrees of freedom, the p-value is 0

Correlation of Coefficients:
   DH DM
DM 0
DL 0   0
```

Now $\hat{\beta}_1$ is the mean of the high SES IQ's, $\hat{\beta}_2$ is the mean of the medium SES IQ's and $\hat{\beta}_3$ is the mean of the low SES IQ's.

There are two important notes to make in comparing our two regressions, `nat.IQ   DH + DM` and `nat.IQ   DH + DM + DL - 1`:

- We used the omnibus F-statistic, 3.415 on 2 and 24 degrees of freedom (p-value is 0.04954), to conclude that SES may matter for nat.IQ from the first regression. Recall that this F-test compares $H_0 : \beta$'s $= 0$ (i.e. model with intercept only, y's = (their grand mean) + (noise)) to $H_1 : \beta$'s free (i.e. the regressors DH and DM matter). The second regression does not have an intercept in it, so the hypotheses that are being compared are $H_0 :$ y's = 0 + (noise) to $H_1 :$ the regressors DH, DM and DL matter. The latter can be true just as long as the grand mean of the y's is significantly different from 0, even if the group means are the same. Consequently *the omnibus F-test from the regression with the intercept omitted is not a good test of the need for the regressors in predicting outcomes.*

- We haven't talked much about the "Correlation of Coefficients" part of the `summary()` display for `lm()`, but compare the two sets of correlations for these two regressions. In the first regression, when DL was dropped, the correlations are relatively high, around ± 0.5. This reflects the fact that there is still some (though not perfect) linear relationship (collinearity) among the regressors DH, DM, and the intercept. In the second regression, where the intercept was dropped, the regressors DH, DM and

DL are perfectly orthogonal—no collinearity at all!—and the correlations between the corresponding regression coefficients are 0, reflecting this fact.

3.3 The effect of SES on IQ's of twins raised by natural and by foster parents

We have determined that SES may matter for the IQ's the twin raised by their natural parents

```
> summary(oneway.nat)
Call: lm(formula = nat.IQ ~ DH + DM)
[...]
F-statistic: 3.415 on 2 and 24 degrees of freedom, the p-value is 0.04954
```

How about for the twin raised by foster parents?

```
> oneway.fos _ lm(fos.IQ ~ DH + DM)
> summary(oneway.fos)

Call: lm(formula = fos.IQ ~ DH + DM)
Residuals:
    Min     1Q Median   3Q   Max
 -30.71 -12.27  2.286 12.5 28.71

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  93.7143   4.2233    22.1896   0.0000
         DH   9.5714   7.3151     1.3085   0.2031
         DM  -4.8810   7.7107    -0.6330   0.5327

Residual standard error: 15.8 on 24 degrees of freedom
Multiple R-Squared: 0.1088
F-statistic: 1.465 on 2 and 24 degrees of freedom, the p-value is 0.2511

Correlation of Coefficients:
   (Intercept)      DH
DH -0.5774
DM -0.5477      0.3162
```

Here it is not at all clear that SES matters anymore. Still we can compute the mean IQ in each SES group (I am "lazy" so I do this by computing the regression with no intercept):

```
> summary(lm(fos.IQ ~ DH + DM + DL -1))

Call: lm(formula = fos.IQ ~ DH + DM + DL - 1)
Residuals:
    Min     1Q Median   3Q   Max
 -30.71 -12.27  2.286 12.5 28.71

Coefficients:
      Value Std. Error  t value Pr(>|t|)
DH 103.2857   5.9727    17.2929   0.0000
DM  88.8333   6.4513    13.7699   0.0000
DL  93.7143   4.2233    22.1896   0.0000
```

and one can see that crude 95% CI's for the three group means ($\hat{\beta} \pm 2$ SE's) overlap quite a bit.

Finally, it is useful to examine the standard diagnostic plots of the two regressions:

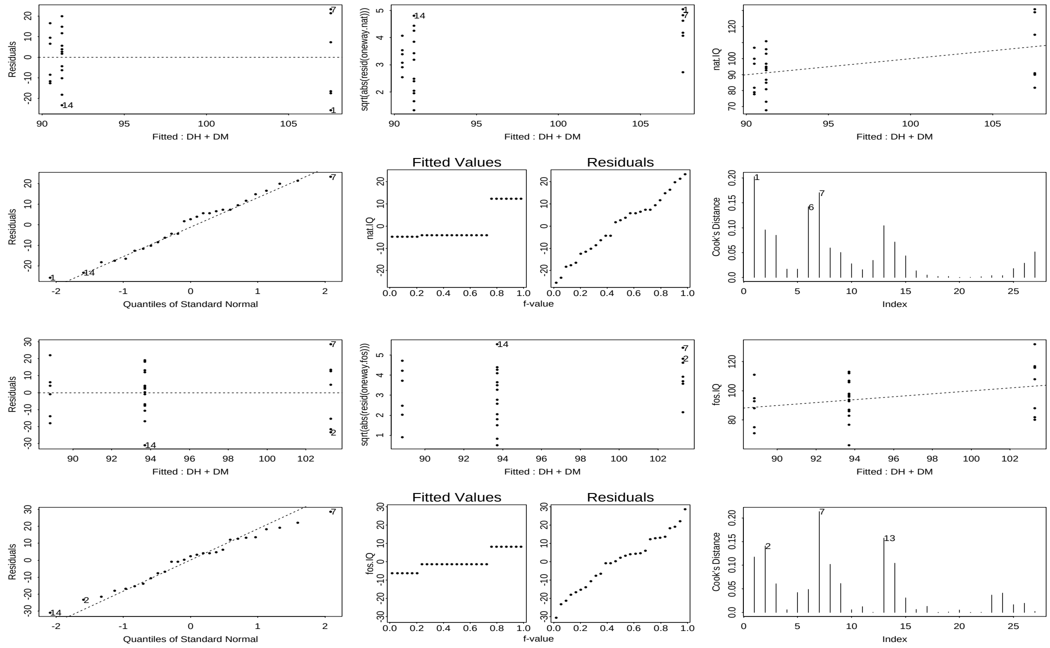

Figure 3: Standard diagnostic plots for regressing natural IQ on SES (upper two rows of plots) and foster IQ on SES (lower two rows of plots).

```
> par(mfrow=c(4,3))
> plot(oneway.nat)
> plot(oneway.fos)
```

These are shown in Figure 3. No severe violations of the statistical assumptions of linear regression can be found from the plots. Since the basic residual plots here plot residuals vs fitted values, one can see indirectly from the residual plots how the fitted means for each SES group are spread out. It seems that SES matters for IQ of twins raised by their natural parents since the high-SES mean is so much different from the other two, which seem almost indistinguishable from each other. On the other hand, the group means for the twins raised by foster parents are spread more uniformly across the IQ range, and this may be making it difficult to distinguish them using the given data.

3.4 How Does Splus Do It?

In making our first figure, Figure 1, we had to work around the fact that Splus converts categorical, or discrete, data to an internal representation it calls a "factor" when it constructs a data frame using `read.table()`. Splus constructs a factor from a discrete variable by precoding a set of dummies for the categories of the variable. So for example we could have learned whether SES matters for IQ in the two groups of twins by simply running the regressions

```
> summary(lm(nat.IQ ~ SES))

Call: lm(formula = nat.IQ ~ SES)
Residuals:
    Min      1Q Median    3Q   Max
 -25.57 -10.86  2.786 8.464 23.43
```

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  96.4286   2.9730    32.4347   0.0000
       SES1  -8.1786   3.3446    -2.4453   0.0222
       SES2  -2.9643   2.2605    -1.3113   0.2022

Residual standard error: 14.45 on 24 degrees of freedom
Multiple R-Squared: 0.2215
F-statistic: 3.415 on 2 and 24 degrees of freedom, the p-value is 0.04954

Correlation of Coefficients:
     (Intercept)    SES1
SES1 -0.2500
SES2  0.2055      0.1644


> summary(lm(fos.IQ ~ SES))

Call: lm(formula = fos.IQ ~ SES)
Residuals:
    Min     1Q Median   3Q   Max
 -30.71 -12.27  2.286 12.5 28.71

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  95.2778   3.2511    29.3060   0.0000
       SES1  -4.7857   3.6575    -1.3085   0.2031
       SES2  -3.2222   2.4720    -1.3035   0.2048

Residual standard error: 15.8 on 24 degrees of freedom
Multiple R-Squared: 0.1088
F-statistic: 1.465 on 2 and 24 degrees of freedom, the p-value is 0.2511

Correlation of Coefficients:
     (Intercept)    SES1
SES1 -0.2500
SES2  0.2055      0.1644
>
```

and not coding our own dummy variables, worrying about imposing constraints, and so forth. There are two things to note about this approach:

- The number of regressors, R^2's and omnibus F-tests are the same. These aspects of a regression are invariant to the parametrizatio you choose for the regression, as long as it is "the same model". Another invariant is the predictions ($\hat{y}$'s) that you would make from the fitted regression.

- The particular coefficients estimated don't look anything like the coefficients we estimated when we constructed the dummies by hand. Unless we know how Splus coded its dummy variables, and what constraints it imposed, we will not be able to interpret Splus's coefficient estimates (as we did our own coefficient estimates above).

There are some statisticians who would say that it is never safe to interpret coefficients, and that all one should focus on are the aspects of fit and predictions, that are invariant to the parametrization anyway (see first bullet). In more complicated models, interpreting parameter estimates can be very tricky and non-intuitive, so they protect themselves against making errors of interpretation by simply not making the interpretation in the first place. If your view is like this, then you may always proceed using Splus's internal dummy codings for factor variables without thinking too much about what they mean. However, we will later learn something about Splus's internal codings, since it is useful to know what the software is doing, and it really is occasionally useful to be able to interpret parameter estimates (as long as you are careful!).

4 A Preview of Two-Way Analysis of Variance

Two-way analysis of variance involves regressing a y variable on two categorical predictors. We might ask, for the present data set, what is the joint effect of SES and being raised in a natural vs foster environment on childrens' IQ's?

Our data set is not set up to do the two-way ANOVA but we can easily construct a data set that does the trick. The idea is to "stack" the natural IQ data on top of the foster IQ data, replicating the SES variable, and including a new variable indicating natural vs foster home status.

```
> dim(twins)
[1] 27 3
> newtwins _ data.frame(IQ = c(fos.IQ,nat.IQ),SES=rep(SES,2),
+ Home=c(rep("F",27),rep("N",27)))
```

Now we construct indicators for the two discrete variables, SES and Home, and consider some regression models.

```
> sesH _ ifelse(newtwins$SES=="H",1,0)
> sesM _ ifelse(newtwins$SES=="M",1,0)
> sesL _ ifelse(newtwins$SES=="L",1,0)
> HomeF _ ifelse(newtwins$Home=="F",1,0)
> HomeN _ ifelse(newtwins$Home=="N",1,0)
```

First we fit an "additive" model that says the contributions of SES and Home are functionally independent of one another and combine additively to predict IQ.

```
> additive _ lm(IQ ~ HomeF + sesH + sesM,data=newtwins)
> summary(additive)

Call: lm(formula = IQ ~ HomeF + sesH + sesM, data = newtwins)
Residuals:
    Min      1Q Median     3Q   Max
 -29.37 -11.36  2.036 10.39 26.66

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  92.5569   3.4730    26.6501   0.0000
      HomeF  -0.1852   4.0589    -0.0456   0.9638
       sesH  12.9643   4.8816     2.6558   0.0106
       sesM  -2.7976   5.1456    -0.5437   0.5891

Residual standard error: 14.91 on 50 degrees of freedom
Multiple R-Squared: 0.1551
F-statistic: 3.061 on 3 and 50 degrees of freedom, the p-value is 0.03653
```

```
Correlation of Coefficients:
      (Intercept)   HomeF    sesH
HomeF -0.5843
 sesH -0.4685      0.0000
 sesM -0.4445      0.0000  0.3162


> anova(additive)
Analysis of Variance Table

Response: IQ

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
HomeF      1      0.46    0.463 0.002082 0.9637916
sesH       1   1975.96 1975.956 8.884216 0.0044332
sesM       1     65.74   65.744 0.295596 0.5890728
Residuals 50  11120.60  222.412
```

Note that the parametrization I've chosen makes "Low-SES, raised by natural parents" the baseline category, and I've dropped one indicator variable for each discrete predictor (why did I drop two, instead of one as I did in the previous example?). From the `summary()` output it appears that the coefficient on HomeF is really not different from zero; but remember that this t-statistic is for the HomeF variable added last after all the other variables. So I've also included the `anova()` output, which for this model gives an F-test for adding HomeF first in the model. Clearly, "it don't make no nevermind"! Also, consistent with what we have seen before, sesH seems to matter a lot, and sesM less so, as with our previous analyses. The standard diagnostic plots do not reveal severe problems here.

Another model to consider is one in which the two factors, SES and HomeF, interact in some way, so that—to give a concrete example—SES might matter more for children in foster homes than children in their natural parents' homes. This can be fitted as indicated in the next Splus display.

Note that in specifying the model I have only indicated terms like `HomeF*sesH` and Splus has expanded these into terms like `HomeF + sesH + HomeF:SesH`. Thus, Splus naturally tries to obey the *principle of marginality* in constructing models:

> *Principle of marginality (a.k.a. "hierarchy principle"):*
> Whenever an interaction is fitted, all of the lower-order interactions and simple terms should also be fitted.

```
> interactive _ lm(IQ ~ HomeF*sesH + HomeF*sesM,data=newtwins)
> summary(interactive)

Call: lm(formula = IQ ~ HomeF * sesH + HomeF * sesM, data = newtwins)
Residuals:
    Min    1Q Median    3Q   Max
 -30.71 -11.3  2.536 11.21 28.71

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  91.2143   4.0467    22.5402   0.0000
      HomeF   2.5000   5.7229     0.4368   0.6642
       sesH  16.3571   7.0091     2.3337   0.0238
       sesM  -0.7143   7.3883    -0.0967   0.9234
 HomeF:sesH  -6.7857   9.9124    -0.6846   0.4969
 HomeF:sesM  -4.1667  10.4486    -0.3988   0.6918


Residual standard error: 15.14 on 48 degrees of freedom
Multiple R-Squared: 0.164
F-statistic: 1.883 on 5 and 48 degrees of freedom, the p-value is 0.1149

Correlation of Coefficients:
           (Intercept)   HomeF    sesH    sesM HomeF:sesH
     HomeF -0.7071
      sesH -0.5774      0.4082
      sesM -0.5477      0.3873  0.3162
HomeF:sesH  0.4082     -0.5774 -0.7071 -0.2236
HomeF:sesM  0.3873     -0.5477 -0.2236 -0.7071  0.3162



> anova(interactive)
Analysis of Variance Table

Response: IQ

Terms added sequentially (first to last)
           Df Sum of Sq  Mean Sq  F Value     Pr(F)
HomeF       1      0.46    0.463 0.002019 0.9643439
sesH        1   1975.96 1975.956 8.618677 0.0050944
sesM        1     65.74   65.744 0.286761 0.5947764
HomeF:sesH  1     79.45   79.448 0.346533 0.5588422
HomeF:sesM  1     36.46   36.458 0.159023 0.6918261
Residuals  48  11004.69  229.264
```

In this case it does not look as if the interaction terms need to be in the model. We can have Splus compute the incremental F test for the interaction terms directly as follows:

```
> anova(interactive,additive)
Analysis of Variance Table

Response: IQ

                        Terms Resid. Df      RSS                   Test Df
1 HomeF * sesH + HomeF * sesM        48 11004.69
2         HomeF + sesH + sesM        50 11120.60 -HomeF:sesH-HomeF:sesM -2
  Sum of Sq   F Value     Pr(F)
1
2 -115.9061 0.2527782 0.7776674
```

but of course the story is the same.

Once again we could also use Splus's internal coding of discrete data into "factors" to do the analysis more efficiently. Once again the thing to note is that the parameter estimates have changed (because Splus's parametrization isn't the same as our "coded-by-hand" parametrization) but, because the models are the same, the F-tests, R^2, etc. (which are all invariant under parametrization) are the same.

```
> add _ lm(IQ ~ Home + SES,data=newtwins)
> summary(add)

Call: lm(formula = IQ ~ Home + SES, data = newtwins)
Residuals:
    Min     1Q Median    3Q   Max
 -29.37 -11.36  2.036 10.39 26.66

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  95.8532   2.1696    44.1803   0.0000
       Home   0.0926   2.0295     0.0456   0.9638
       SES1  -6.4821   2.4408    -2.6558   0.0106
       SES2  -3.0933   1.6496    -1.8751   0.0666

Residual standard error: 14.91 on 50 degrees of freedom
Multiple R-Squared: 0.1551
F-statistic: 3.061 on 3 and 50 degrees of freedom, the p-value is 0.03653

> anova(add)
Analysis of Variance Table

Response: IQ

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
Home       1      0.46    0.463 0.002082 0.9637916
SES        2   2041.70 1020.850 4.589906 0.0147880
Residuals 50  11120.60  222.412
```

```
> int _ lm(IQ ~ Home*SES,data=newtwins)
> summary(int)

Call: lm(formula = IQ ~ Home * SES, data = newtwins)
Residuals:
    Min    1Q Median    3Q   Max
 -30.71 -11.3  2.536 11.21 28.71

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  95.8532   2.2028    43.5150   0.0000
       Home   0.5754   2.2028     0.2612   0.7950
       SES1  -6.4821   2.4781    -2.6158   0.0119
       SES2  -3.0933   1.6749    -1.8469   0.0709
   HomeSES1  -1.6964   2.4781    -0.6846   0.4969
   HomeSES2   0.1290   1.6749     0.0770   0.9389

Residual standard error: 15.14 on 48 degrees of freedom
Multiple R-Squared: 0.164
F-statistic: 1.883 on 5 and 48 degrees of freedom, the p-value is 0.1149

> anova(int)
Analysis of Variance Table

Response: IQ

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
Home       1      0.46    0.463 0.002019 0.9643439
SES        2   2041.70 1020.850 4.452719 0.0168306
Home:SES   2    115.91   57.953 0.252778 0.7776674
Residuals 48  11004.69  229.264

> anova(add,int)
Analysis of Variance Table

Response: IQ

       Terms Resid. Df      RSS      Test Df Sum of Sq   F Value     Pr(F)
1 Home + SES        50 11120.60
2 Home * SES        48 11004.69 +Home:SES  2  115.9061 0.2527782 0.7776674
```

A reasonable approach in general might be to use Splus's "factor" codings to get a good initial idea of what is important or not in a model fitting exercise, and then go back and code dummy variables whose coefficients you understand well enough to interpret directly when you do a more detailed analysis based on the models you like.

5 Analysis of Covariance: Interactions between discrete and continuous preductors

Let us return to the original `twins` data set now to answer a different question. Since the data are on IQ's of twins raised apart, one in the natural parents' home and one in a foster home, we can consider the "natural-raised" twin to be a kind of control for the "foster-raised" twin. From Figure 1 (page 3) it appears that nat.IQ is a very good predictor of fos.IQ. After accounting for the variability in fos.IQ explained by nat.IQ, is there additional variability that can be explained by SES? Answering this question would in some way contribute to the "nature vs. nurture" question of intelligence.

The simplest (most restrictive) model is just simple linear regression predicting fos.IQ from nat.IQ. Here we go:

```
> simple _ lm(fos.IQ ~ nat.IQ)
> summary(simple)

Call: lm(formula = fos.IQ ~ nat.IQ)
Residuals:
    Min     1Q  Median    3Q   Max
 -11.35 -5.731 0.05742 4.324 16.35

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 9.2076 9.2999     0.9901  0.3316
     nat.IQ 0.9014 0.0963     9.3575  0.0000

Residual standard error: 7.729 on 25 degrees of freedom
Multiple R-Squared: 0.7779
F-statistic: 87.56 on 1 and 25 degrees of freedom, the p-value is 1.204e-09

Correlation of Coefficients:
       (Intercept)
nat.IQ -0.9871
```

The standard diagnostic plots don't reveal much, except that observations 23 and 24 appear to be mildly outlying.

Now let's examine the simple regression model for variations among the three SES groups. There are 4 models that are commonly considered. We will list them in order from most general to most restrictive. They are:

1. *MOST GENERAL:* Each of the three SES groups has its own slope and intercept parameters: think of three randomly-dropped toothpicks.

2. *PARALLEL REGRESSIONS:* Each of the three SES groups has the same slope, but different intercepts. Think of three parallel toothpicks.

3. *CONCURRENT REGRESSIONS:* The three SES groups have different slopes, but they intersect at a common point. Usually this point is on the y-axis (same intercept), but by subtracting (or adding) a constant from the X-values the common point could be moved to some other location. Think of three toothpicks connected at one end and emanating in three directions.

4. *COINCIDENT REGRESSIONS:* The three regression lines are the same for all three SES groups. This is the simple regression we fit above. (Now the three toothpicks occupy the same space.)

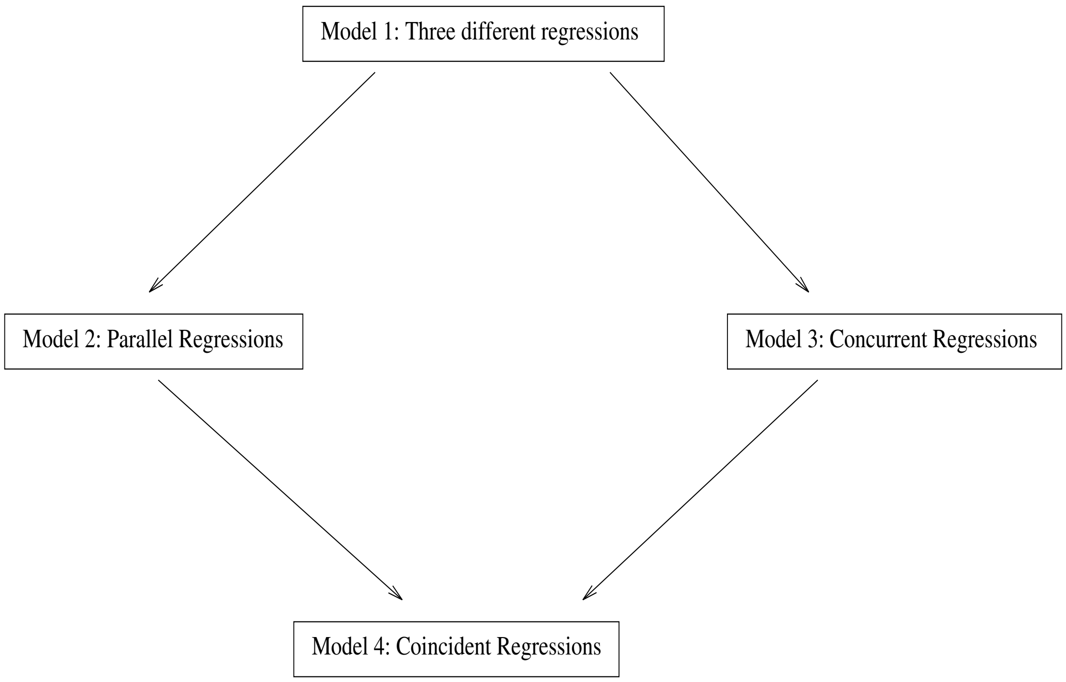

Figure 4: Nesting of the models for the twins data. Each (smaller) model at the end of an arrow can be obtained from the (larger) model at the beginning of the arrow by setting some β's $= 0$ in the larger model.

These models are nested within each other according to Figure 4: each "lower" model in the figure can be obtained from a "higher" model to which it is connected by an arrow, by imposing one or more linear constraints (setting some β's $= 0$) on the "higher" model. There are two kinds of comparisons that would be common here.

- compare 2, 3, and 4 with 1
- compare 4 against 2 (this is a classic ANCOVA [analysis of covariance])

Because of the nature of the F-test for comparing models,

$$\frac{(RegSS_{big} - RegSS_{small})/q}{RSS_{big}/(n-k-1)} = \frac{(RSS_{small} - RSS_{big})/q}{RSS_{big}/(n-k-1)}$$

it only makes sense to compare models which are nested, for these are the models where it is guaranteed that $RegSS_{big} - RegSS_{small} = RSS_{small} - RSS_{big} \geq 0$.

Now let's fit the four models. I am going to choose a parametrization without an overall intercept, to make interpreting the parameters in terms of the individual regressions as straightforward as possible.

```
> model1 _ lm(fos.IQ ~ (DH + DM + DL)*nat.IQ - 1 - nat.IQ)
> summary(model1)

Call: lm(formula = fos.IQ ~ (DH + DM + DL) * nat.IQ - 1 - nat.IQ)
Residuals:
    Min      1Q Median     3Q  Max
 -14.48 -5.248 -0.155 4.582 13.8
```

```
Coefficients:
             Value Std. Error  t value Pr(>|t|)
       DH  -1.8720  17.8083    -0.1051   0.9173
       DM   0.8160  26.1092     0.0313   0.9754
       DL   7.2046  16.7513     0.4301   0.6715
DH:nat.IQ   0.9776   0.1632     5.9902   0.0000
DM:nat.IQ   0.9726   0.2863     3.3973   0.0027
DL:nat.IQ   0.9484   0.1822     5.2061   0.0000

Residual standard error: 7.921 on 21 degrees of freedom
Multiple R-Squared: 0.9948
F-statistic: 663.2 on 6 and 21 degrees of freedom, the p-value is 0

Correlation of Coefficients:
               DH      DM      DL DH:nat.IQ DM:nat.IQ
       DM  0.0000
       DL  0.0000  0.0000
DH:nat.IQ -0.9858  0.0000  0.0000
DM:nat.IQ  0.0000 -0.9923  0.0000  0.0000
DL:nat.IQ  0.0000  0.0000 -0.9920  0.0000    0.0000
> plot(model1) # reveals nothing untoward
> model2 _ lm(fos.IQ ~ (DH + DM + DL) + nat.IQ - 1)
> summary(model2)

Call: lm(formula = fos.IQ ~ (DH + DM + DL) + nat.IQ - 1)
Residuals:
    Min     1Q  Median    3Q  Max
 -14.82 -5.237 -0.1111 4.476 13.7

Coefficients:
          Value Std. Error  t value Pr(>|t|)
    DH  -0.6076  11.8551    -0.0513   0.9596
    DM   1.4277  10.1604     0.1405   0.8895
    DL   5.6188   9.9628     0.5640   0.5782
nat.IQ   0.9658   0.1069     9.0306   0.0000

Residual standard error: 7.571 on 23 degrees of freedom
Multiple R-Squared: 0.9947
F-statistic: 1089 on 4 and 23 degrees of freedom, the p-value is 0


Correlation of Coefficients:
            DH      DM      DL
    DM  0.9244
    DL  0.9502  0.9328
nat.IQ -0.9704 -0.9526 -0.9792
> plot(model2) # reveals nothing much more or less than model1

> model3 _ lm(fos.IQ ~ 1 + (DH + DM + DL):nat.IQ - nat.IQ)
> summary(model3)

Call: lm(formula = fos.IQ ~ 1 + (DH + DM + DL):nat.IQ - nat.IQ)
Residuals:
```

```
   Min     1Q   Median    3Q   Max
 -15.4 -5.188 -0.05831 4.596 13.58

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  2.5623 10.5984     0.2418  0.8111
  DH:nat.IQ  0.9375  0.0993     9.4423  0.0000
  DM:nat.IQ  0.9536  0.1202     7.9319  0.0000
  DL:nat.IQ  0.9985  0.1164     8.5747  0.0000

Residual standard error: 7.594 on 23 degrees of freedom
Multiple R-Squared: 0.8027
F-statistic: 31.2 on 3 and 23 degrees of freedom, the p-value is 2.791e-08

Correlation of Coefficients:
          (Intercept) DH:nat.IQ DM:nat.IQ
DH:nat.IQ -0.9643
DM:nat.IQ -0.9592      0.9249
DL:nat.IQ -0.9819      0.9468    0.9418
> plot(model3) # similar to previous diagnostic plots

> model4 _ lm(fos.IQ ~ nat.IQ)
> summary(model4)

Call: lm(formula = fos.IQ ~ nat.IQ)
Residuals:
    Min     1Q  Median    3Q   Max
 -11.35 -5.731 0.05742 4.324 16.35

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 9.2076 9.2999     0.9901  0.3316
     nat.IQ 0.9014 0.0963     9.3575  0.0000

Residual standard error: 7.729 on 25 degrees of freedom
Multiple R-Squared: 0.7779
F-statistic: 87.56 on 1 and 25 degrees of freedom, the p-value is 1.204e-09

Correlation of Coefficients:
       (Intercept)
nat.IQ -0.9871
> plot(model4) # observations 23 and 24 appear to be mildly outlying
```

Now let's do the first set of comparisons (models 2, 3, and 4 vs. model 1).

- `2 (PARALLEL) vs 1 (THREE REGRESSIONS):`

  By looking at `anova(model1)` and `anova(model2)` you should verify that the F statistic is $[(1318.4 - 1317.5)/2]/[1317.5/(27 - 6)] = 0.007172676$, clearly we don't have to refer to an F table to know that the smaller model is acceptable here.

- `3 (CONCURRENT [SAME INTERCEPT]) vs 1 (THREE REGRESSIONS):`

  By looking at `anova(model1)` and `anova(model3)` you should verify that the F statistic is $[(1326.46 - 1317.5)/2]/[1317.5/(27 - 6)] = 0.071407963$, again no checking of the F distribution is needed to prefer the smaller model.

- `4 (IDENTICAL REGRESSIONS) vs 1 (THREE REGRESSIONS):`

  By looking at `anova(model1)` and `anova(model4)` you should verify that the F statistic is $[(1493.533 - 1317.5)/(6 - 2)]/[1317.5/(27 - 6)] = 0.7014597$, and again it is not necessary to check the F distribution to prefer the smaller model.

If anything, there's evidence of overfit here! Remember, each F statistic is usually bigger than 1; here all the F's are less than 1...

Now, let's look at the second comparison: model 4 vs model 2. This is exactly like our 1-way ANOVA from before but now the SES means (under H1) and the overal means (under H0) are corrected for the predictive quality of the "covariate" '`nat.IQ`'. This is usually called the "analysis of covariance" or ANCOVA...

To be explicit, the two hypotheses are:

$$
\begin{aligned}
H_0: &\quad y = \alpha + \beta_4 nat.IQ + \varepsilon \\
H_1: &\quad y = \beta_1 DH + \beta_2 DM + \beta_3 DL + \beta_4 nat.IQ + \varepsilon
\end{aligned}
$$

(NOTE: Since the simple ANOVA failed to detect anything, it'd be somewhat surprising if the ANCOVA detected anything. Usually the effect of including the covariate is to decrese differences between the cell means in an ANOVA)

You should verify from the `anova(model4)` and `anova(model2)` output that the F statistic is $[(1493.5 - 1318.4)/2]/[1318.4/(27 - 4)] = 1.52734$, and using $pf()$ in Splus that $P[F(2, 23) >= 1.52734] = 0.24$, so again we don't reject the simpler (more restrictive) model...

Another way to do this is by using the `anova()` function to compare the two models directly

```
> anova(model2,model4)
Analysis of Variance Table

Response: fos.IQ

                        Terms Resid. Df      RSS       Test Df Sum of Sq
1 (DH + DM + DL) + nat.IQ - 1        23 1318.403
2                      nat.IQ        25 1493.533 -DH-DM-DL -2 -175.1306
   F Value     Pr(F)
1
2 1.527607 0.2382767
```

And finally, we get rid of the data set we've been playing with...

```
> detach()
```