36-401 Class Notes and Exercise for 10/3/00

Howard Seltman

Interaction Effects

In the model

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \epsilon_i,$$

the change in Y due to a unit change in X_1 holding X_2 constant is always β_1. If we define the *interaction term* $X_{i3} = X_{i1}X_{i2}$, then for the model

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + \epsilon_i,$$

the change in Y due to a unit change in X_1 holding X_2 constant is $\beta_1 + X_{i2}\beta_3$.

This means that our (linear) regression model is fitting a curved surface to the data, not a plane. (The other way we have fit a curve is with transformations of the X or Y variables; in those cases the relationship on the transformed scale is linear, but on the original scale it is curved.)

Example:

```
> summary(lm(bio~sal+pH+sal*pH,grass))

Coefficients:     Value Std. Error    t value   Pr(>|t|)
(Intercept) -5476.9326  2397.6754    -2.2843     0.0276
        sal   146.6476    76.3007     1.9220     0.0616
         pH  1470.4926   503.4351     2.9209     0.0057
     sal:pH   -34.0018    16.0340    -2.1206     0.0401

Residual standard error: 408.8 on 41 degrees of freedom
Multiple R-Squared: 0.6426
```

At salinity=25, an increase in pH of 1 unit is associated with a biomass increase of 1470 - (34.0*25) = 620, while at salinity=35, the increase is only 1470 - (34.0*35) = 280.

Conversely, at pH=4, a salinity increase of 1 is associated with a biomass increase of 146.6 - (34.0*4) = 10.6, while at pH=6, the change is 146.6 - (34.0*6)= -57.4, a big drop.

Interaction of two (or more) continuous variables represents complex relationships, such as synergy or antagonism.

Indicator (dummy) variables

A two-level categorical independent variable can be coded as, e.g. 0 and 1. We have already learned that these "dummy" variables represent different intercepts in situations where the relationship of the outcome with the other covariates is otherwise identical (parallel).

For more than $j > 2$ levels, we must use $j - 1$ dummy variables so that each category can have its own intercept. Use of e.g. (0, 1, 2) in a single variable for (blue, red, green) codes the usually mistaken assumption that the fitted lines (planes) are not just parallel, but also equally spaced.

In Splus, "factor" objects are automatically expanded to $j-1$ columns as needed. Character variables are automatically converted to factors by read.table(). Numeric dummy variables must be converted with a statement such as

```
mydtf$mydummy_factor(mydtf$mydummy).
```

A nice way to check for factors is `sapply(mydtf,is.factor)`. You can get the original values back with `unclass(mydtf$mydummy)`. For two-level variables making a factor is unnecessary.

Splus uses its own parameterization for factors, which you don't need to know. If you leave the intercept in the model, you will not be able to directly interpret the coefficients, but to find the intercepts for a level of a factor, you can use predict as shown below. Or you can drop the (overall) intercept to get a more interpretable result:

```
> summary(lm(damage~loc+elev, spruce))$coef
                  Value  Std. Error   t value     Pr(>|t|)
(Intercept) -60.15228618 17.57305598 -3.422984 1.111818e-03
        loc  49.78022529  8.24476064  6.037801 1.009653e-07
       elev   0.05680546  0.01255495  4.524548 2.855655e-05

> spruce$locfac_factor(spruce$loc)
> fac.lm_lm(damage~locfac+elev, spruce)
> summary(fac.lm)$coef
                  Value  Std. Error   t value     Pr(>|t|)
(Intercept) -35.26217354 14.28024915 -2.469297 1.635295e-02
     locfac  24.89011264  4.12238032  6.037801 1.009653e-07
       elev   0.05680546  0.01255495  4.524548 2.855655e-05
> predict(fac.lm, data.frame(elev=0, locfac=factor("0", levels=levels(spruce$locfac))))
 -60.15229
> fac.lm$contrasts   # Peek at S-plus's parameterization
 $locfac:
  0   -1
  1    1

> summary(lm(damage~locfac+elev-1, spruce))$coef
              Value  Std. Error    t value     Pr(>|t|)
locfac0 -60.15228618 17.57305598 -3.4229838 1.111818e-03
locfac1 -10.37206089 11.53372334 -0.8992812 3.720392e-01
   elev   0.05680546  0.01255495  4.5245482 2.855655e-05
```

Slope Dummy Variables

Interactions of dummy variable covariates with continuous covariates results in slope dummy variables (see below).

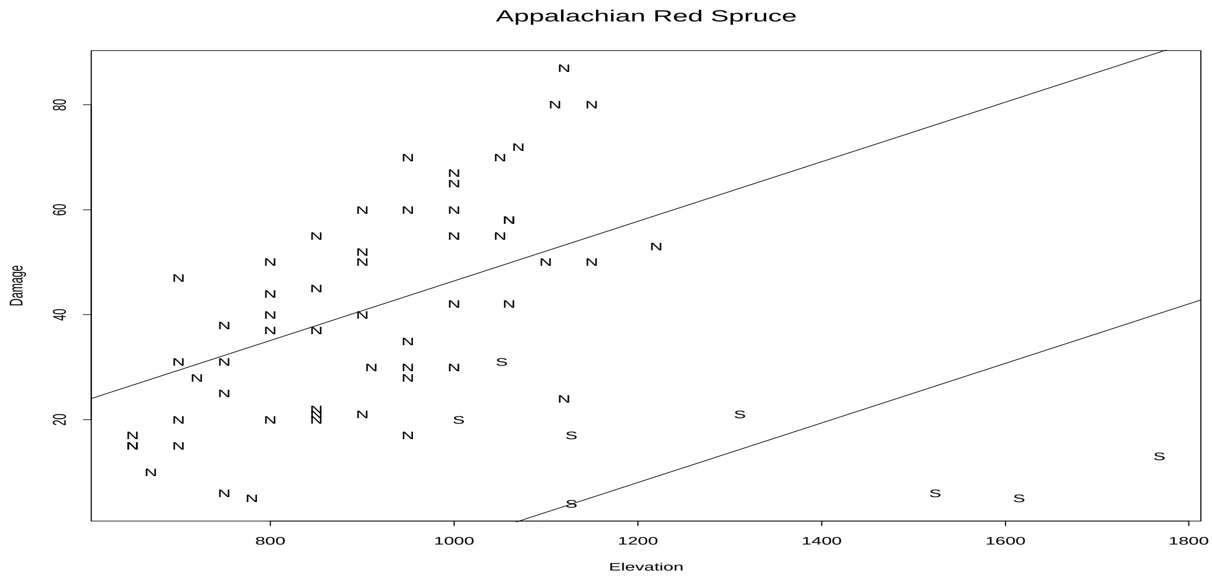
Breakout Session for 10/3/00

This study is from RwG, chapter 3, exercises 1-11. The data are from 64 wooded sites in the Appalachian Mountains. Ten are from southern states (loc=0), and the remainder from northern ones (loc=1). The elevation in feet and the percent of damaged trees are recorded at each site.

```
> spruce_read.table("spruce.dat",T)
> dim(spruce)   # 64  4
> spruce[1:3,]
  site loc elev damage
1    1   0 1615      5
2    2   0 1768     13
3    3   0 1524      6

> splm_lm(damage~loc+elev, spruce)
> summary(splm)
Coefficients:  Value Std. Error  t value Pr(>|t|)
(Intercept) -60.1523  17.5731    -3.4230   0.0011
        loc  49.7802   8.2448     6.0378   0.0000
       elev   0.0568   0.0126     4.5245   0.0000
Multiple R-Squared: 0.3793

plot(spruce$elev, spruce$damage, xlab="Elevation", ylab="Damage",
  main="Appalachian Red Spruce", type="n")  # "n" means no points plotted
STH_spruce$loc==0
NTH_spruce$loc==1
points(spruce$elev[STH], spruce$damage[STH], pch="S")
points(spruce$elev[NTH], spruce$damage[NTH], pch="N")
tmp_splm$coef
abline(tmp[1], tmp[3])          # South (loc=0)
abline(tmp[1]+tmp[2], tmp[3])   # North (loc=1)
```



1. In what way(s) is the model inadequate? How can we improve it?

Consider a different model:

```
> splmi_lm(damage~loc+elev+loc*elev, spruce)
> summary(splmi)
Coefficients:
                 Value Std. Error  t value Pr(>|t|)
  (Intercept)  37.2836  25.8967     1.4397   0.1551
          loc -78.6220  28.5397    -2.7548   0.0078
         elev  -0.0172   0.0193    -0.8929   0.3755
     elev:loc   0.1084   0.0233     4.6464   0.0000

Residual standard error: 14.61 on 60 degrees of freedom
Multiple R-Squared: 0.5436
```

2. What are the new fit equations? Plot the new regression fit on the graph.

3. The new fit equations can also be obtained by

```
lm(damage~elev, spruce, subset=STH)
lm(damage~elev, spruce, subset=NTH)
```

What does the new model tell you that two separate regression can't?

4. “Extra credit” In what way(s) are the regression assumptions unlikely to be met? Be as specific as possible.