36-401 Class Notes and Exercise for 10/9/01

Brian Junker and Kimberly Sellers

Interaction Effects

In the model

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \epsilon_i,\ i = 1, \ldots, n$$

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)) # '*' denotes interaction

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); e.g. see "Class Notes and Exercise for 9/20/01, Part II".

```
> wages_read.table("wages.dat", header=T)
> wages_cbind(wages, LogWage=log(wages$Wage))
> wage1.lm_lm(LogWage ~ Sex, data=wages)
> summary(wage1.lm)

Call: lm(formula = LogWage ~ Sex, data = wages)
Residuals:
    Min      1Q   Median     3Q   Max
 -2.165 -0.3759 0.006617 0.3685 1.861

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   2.1653   0.0303    71.4106   0.0000
        Sex  -0.2312   0.0448    -5.1658   0.0000

Residual standard error: 0.5155 on 532 degrees of freedom
Multiple R-Squared: 0.04776
F-statistic: 26.69 on 1 and 532 degrees of freedom, the p-value is
3.39e-07

Correlation of Coefficients:
    (Intercept)
Sex -0.6773
```

Note that the dummy variable regression performs essentially the same test as a two-sample t-test or one-way ANOVA (with one 2-category X variable) for the difference of two means. More generally, a oneway ANOVA with K-categories ($K > 2$) is equivalent to a regression on $K - 1$ dummy variables. We cannot estimate models with both a Y-intercept and coefficients on K dummies for a K-category variable, due to multicollinearity.

It is possible to maintain the intercept in the model and remove a covariate instead. However, if you leave the intercept in the model, you will not be able to directly interpret the coefficients. Thus, you are usually better off dropping the (overall) intercept.

```
> occ.prestige.lm_lm(occ.prestige ~ Prof + WC + BC + Unknown - 1,
data=newprestige)
> summary(occ.prestige.lm)

Call: lm(formula = occ.prestige ~ Prof + WC + BC + Unknown - 1, data =
newprestige)
Residuals:
    Min     1Q  Median    3Q   Max
 -19.93 -7.982 -0.3067 6.218 29.89

Coefficients:
          Value Std. Error t value Pr(>|t|)
   Prof 67.9067  1.8826    36.0701  0.0000
     WC 42.2435  2.1501    19.6471  0.0000
     BC 36.2067  1.5372    23.5543  0.0000
Unknown 34.7250  5.1558     6.7351  0.0000

Residual standard error: 10.31 on 98 degrees of freedom
Multiple R-Squared: 0.9589
```

```
F-statistic: 571.8 on 4 and 98 degrees of freedom, the p-value is 0

Correlation of Coefficients:
        Prof WC BC
     WC 0
     BC 0    0
Unknown 0    0  0
```

Slope Dummy Variables

Interactions of dummy variable covariates with measurement variables results in slope dummy variables. Consider the model:

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

where X_1 is a dummy variable and X_2 is a measurement variable. A t-test of $H_0 : \beta_3 = 0$ determines whether the two slopes (β_2 and $\beta_2 + \beta_3$) differ significantly. Different slopes imply that, for one of the two categories of X_1, the $Y - X_2$ relation is stronger or has a different sign.

If we suspect two categories might differ wrt. intercepts and slopes, then both the intercept and slope dummy variables could be included in the model.

An F-statistic can test the null hypothesis that neither intercept nor slope is different. Note that this technique works around multicollinearity.

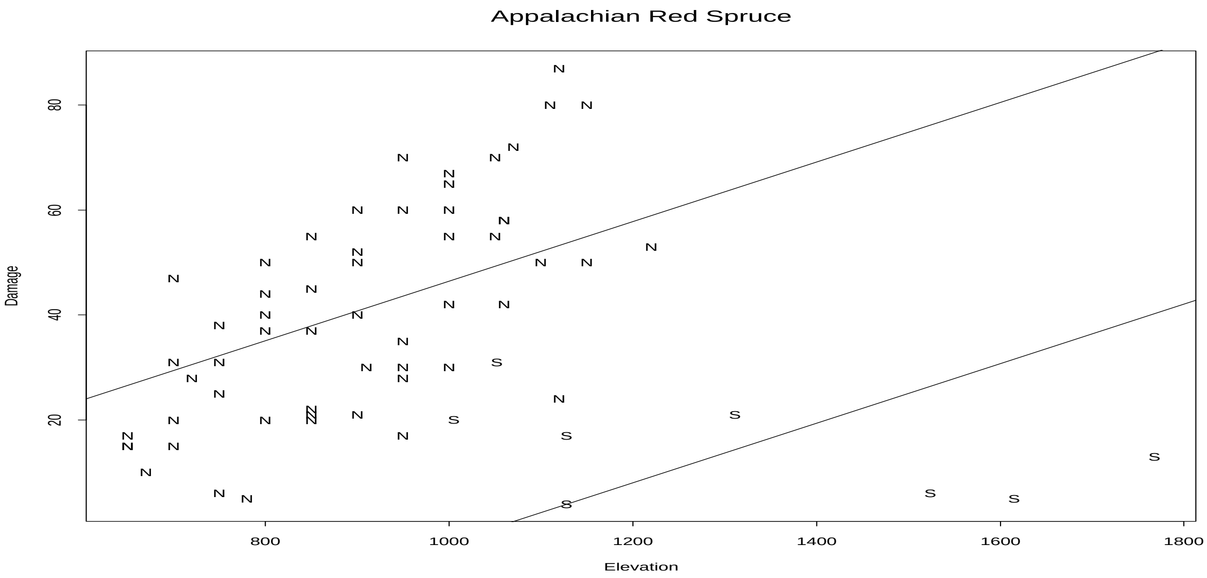
Breakout Session for 10/9/01

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 model on the top of this page tell you that these two separate regressions can't?

4. If you have time, answer the following question: In what way(s) are the regression assumptions unlikely to be met? Be as specific as possible.