Feb 1 lecture

- ANOVA table for simple linear regression
  - computed relative to a specific larger model and a smaller model obtained from the full model by setting some parameters to zero.
  - In simple linear regression, test hypothesis:
    - $H_o: \mu\{Y|X\} = \beta_o$
    - $H_{A:} \mu\{Y|X\} = \beta_o + \beta_1 X$
  - Equivalent to a test of $H_o: \beta = 0; H_{A:} \beta \neq 0$
- Composite ANOVA table for lack-of-fit test
  - for data with replicates at each X (can estimate pure error) assess adequacy of linear model.

Linear Models

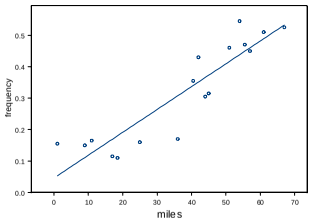
- A model that is linear in the coefficients (β's)
- Polynomial regression
- Transformed variables
- Interaction effects
  - qualitative variables (indicator or dummy variables)
    - Additive effect? Interaction effect? No effect?
  - multiple level categorical variables

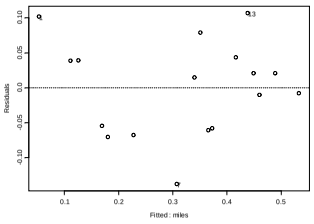
Polynomial Regression

- $\mu\{Y|X\} = \beta_o + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$
- Often of interest when only one independent variable
- For most biological work, terms of X no higher than cubic are used
- Note that for each higher order term added, we lose a degree of freedom for estimating $\sigma = \sigma\{Y|X\}$
- Difficult to read structural meaning into the terms X^2, X^3
- Usually forward selection approach

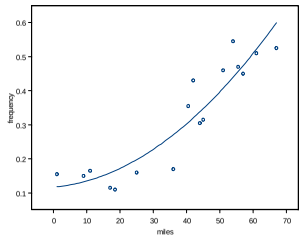
Gene Frequency and Distance

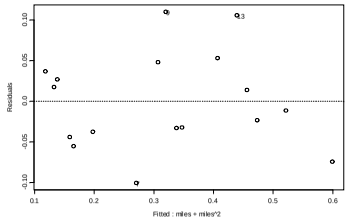
- We model frequency of an allele (*Lap*^{94}) in the blue mussel as a function of distance (miles east of Southport, Connecticut).





Add a quadratic term





Splus Output

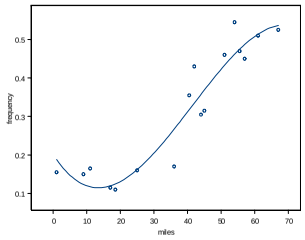
$$\mu\{Frequency|Distance\} = \beta_0 + \beta_1 Distance$$

```
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 0.0458 0.0364     1.2604  0.2268
      miles 0.0073 0.0009     8.4204  0.0000

Residual standard error: 0.06955 on 15 degrees of freedom
Multiple R-Squared: 0.8254
F-statistic: 70.9 on 1 and 15 degrees of freedom, the p-value is 4.557e-007
```

$$\mu\{Frequency|Distance\} = \beta_0 + \beta_1 Distance + \beta_2 Distance^2$$

```
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 0.1173 0.0480     2.4409  0.0285
      miles 0.0008 0.0032     0.2618  0.7973
 I(miles^2) 0.0001 0.0000     2.0470  0.0599

Residual standard error: 0.06316 on 14 degrees of freedom
Multiple R-Squared: 0.8656
F-statistic: 45.09 on 2 and 14 degrees of freedom, the p-value is 7.919e-007
```

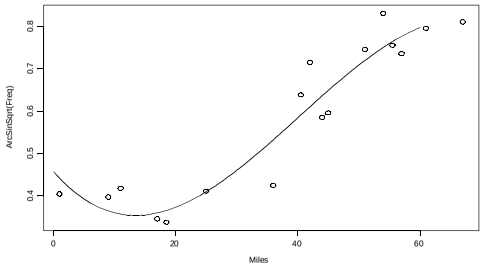
Cubic Term



$$\mu\{Frequency|Distance\} = \beta_0 + \beta_1 Distance + \beta_2 Distance^2 + \beta_3 Distance^3$$

```
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept)  0.2010  0.0547     3.6761  0.0028
      miles -0.0137  0.0068    -2.0277  0.0636
 I(miles^2)  0.0006  0.0002     2.7554  0.0164
 I(miles^3)  0.0000  0.0000    -2.3662  0.0342
```

Model Suggested by Koehn




```
(Intercept)       miles   I(miles^2)     I(miles^3)
  0.4576821 \0.01647738 0.0007355729 -6.112638e-006
```

Koehn, 1978, "Physiology and biochemistry of enzyme variation: The interface of ecology and population genetics." In: *Ecological Genetics: The Interface.*

Model Suggested by Koehn

```
Call: lm(formula = asin(sqrt(freq)) ~ miles + miles^2 +
miles^3)
Residuals:
     Min       1Q   Median       3Q      Max
 -0.1076 -0.03728 -0.00682 0.03319 0.1049

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept)  0.4577  0.0596     7.6840  0.0000
      miles -0.0165  0.0074    -2.2371  0.0434
 I(miles^2)  0.0007  0.0002     3.0540  0.0092
 I(miles^3)  0.0000  0.0000    -2.6969  0.0183

Residual standard error: 0.0597 on 13 degrees of freedom
Multiple R-Squared: 0.9112
F-statistic: 44.46 on 3 and 13 degrees of freedom, the
p-value is 4.268e-007
```

Interaction Effects

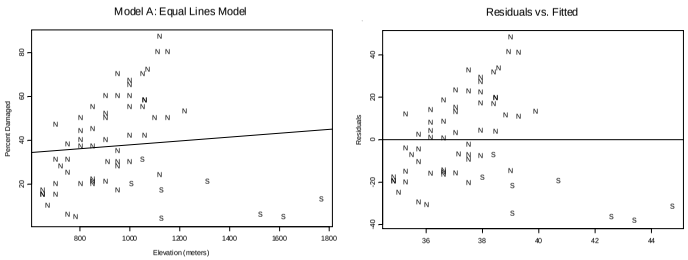
- Red spruce forests in the Appalachian Mountains show signs of decline, with many dead or dying trees.
- Deposition of airborne pollutants such as metals or acids tends to be heavier at higher elevations, where red spuce predominate.
- Linear model relating the percentage of dead or badly damaged trees to elevation (meters) and region (North or South).
- 64 sites. Eight of the sites are in southern states (West Virginia, Virginia and North Carolina); the remainder are northern (New Hampshire, Vermont and New York).
- Cite: Johnson & Siccama, Committee on Monitoring and Assessment of Trends in Acid Deposition, 1986.

Three models to describe region effect

- *Model A: Equal Lines.* There is a linear relationship between Damage and Elevation; there is no difference in this relationship by region.
- *Model B: Parallel Lines.* There is a linear relationship between Damage and Elevation; the rates of change are the same for the regions, but the mean percent damaged at each elevation differs.
- *Model C: Separate Lines.* There is a linear relationship between Damage and Elevation, but it is different for each region (separate slope, intercept for each).

See Display 9.8 of *Sleuth*, page 239.

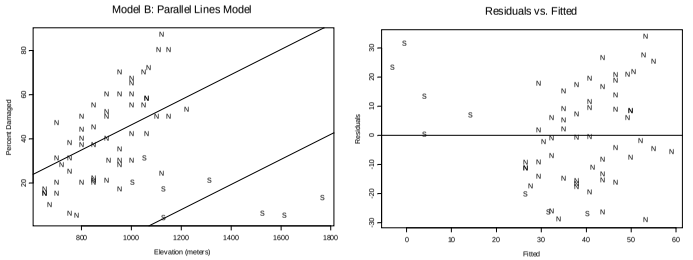
Model A



Model A: Splus Output

```
Call: lm(formula = Perc.Damage ~ Elevation)
Residuals:
    Min     1Q Median    3Q   Max
 -38.39 -16.29 -1.064 15.35 47.99

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 29.1034 11.9125     2.4431  0.0174
  Elevation  0.0088  0.0122     0.7256  0.4708

Residual standard error: 21.18 on 62 degrees of
freedom
Multiple R-Squared: 0.008421
F-statistic: 0.5265 on 1 and 62 degrees of
freedom, the p-value is 0.4708
```

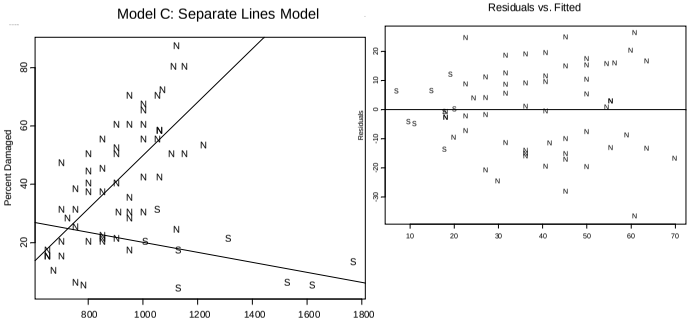
Model B



Model B: Splus Output

```
Call: lm(formula = Perc.Damage ~ Elevation +
Location)
Residuals:
    Min     1Q  Median   3Q   Max
 -29.25 -13.79 -0.8327 13.2 33.75

Coefficients:
             Value Std. Error  t value Pr(>|t|)
(Intercept) -60.1523  17.5731    -3.4230   0.0011
  Elevation   0.0568   0.0126     4.5245   0.0000
   Location  49.7802   8.2448     6.0378   0.0000

Residual standard error: 16.89 on 61 degrees of
freedom
Multiple R-Squared: 0.3793
F-statistic: 18.64 on 2 and 61 degrees of freedom,
the p-value is 4.809e-007
```

Model C



Model C: Splus Output

```
Call: lm(formula = Perc.Damage ~ Elevation + Location +
Elevation:Location)
Residuals:
    Min     1Q Median    3Q   Max
 -36.78 -11.61 0.3082 11.04 26.22

Coefficients:
                    Value Std. Error  t value Pr(>|t|)
      (Intercept)  37.2836  25.8967     1.4397   0.1551
        Elevation  -0.0172   0.0193    -0.8929   0.3755
         Location -78.6220  28.5397    -2.7548   0.0078
Elevation:Location   0.1084   0.0233     4.6464   0.0000

Residual standard error: 14.61 on 60 degrees of freedom
Multiple R-Squared: 0.5436
F-statistic: 23.82 on 3 and 60 degrees of freedom, the
p-value is 2.825e-010
```

Modify Model C to eliminate "Elevation"?

```
Call: lm(formula = Perc.Damage ~Location+Elevation:Location)
Residuals:
    Min     1Q Median    3Q   Max
 -36.78 -10.87 0.7189 11.04 26.22

Coefficients:
                    Value Std. Error  t value Pr(>|t|)
      (Intercept)  14.6250   5.1556     2.8367   0.0062
         Location -55.9634  13.0375    -4.2925   0.0001
Elevation:Location   0.0912   0.0131     6.9522   0.0000

Residual standard error: 14.58 on 61 degrees of freedom
Multiple R-Squared: 0.5375
F-statistic: 35.45 on 2 and 61 degrees of freedom, the
p-value is 6.106e-011
```

Extra Sum of Squares F-test (10.3, *Sleuth)*

$$H_O: \mu\{P|E,L\} = \beta_0 + \beta_1 E$$
$$H_A: \mu\{P|E,L\} = \beta_0 + \beta_1 E + \beta_2 L + \beta_3 E \times L$$

REDUCED:

```
                Df Sum of Sq  Mean Sq   F Value     Pr(F)
Elevation  1    236.17 236.1704 0.5265261 0.4708012
Residuals 62  27809.77 448.5446
```

FULL:

```
Terms added sequentially (first to last)
                    Df Sum of Sq  Mean Sq  F Value     Pr(F)
Elevation  1    236.17   236.17  1.10696 0.2969622
Location  1  10402.81 10402.81 48.75935 0.0000000
Elevation:Location  1   4605.95  4605.95 21.58872 0.0000190
Residuals 60  12801.00   213.35
```

F-statistic:

$$F = \frac{(27809.77 - 12801.00)/(62-60)}{213.35} = 35.17$$

Reject H_o if $F > F_{(.95;2,60)} = 3.15$

Individual North/South Models

South Only

```
Coefficients:
                          Value Std. Error  t value Pr(>|t|)
            (Intercept)  37.2836  15.4525     2.4128   0.0524
Elevation[Location == 0]  -0.0172   0.0115    -1.4964   0.1852

Residual standard error: 8.716 on 6 degrees of freedom
Multiple R-Squared: 0.2718
F-statistic: 2.239 on 1 and 6 degrees of freedom, the p-value is 0.1852
```

North Only

```
Coefficients:
                          Value Std. Error  t value Pr(>|t|)
            (Intercept) -41.3384  12.4165    -3.3293   0.0016
Elevation[Location == 1]   0.0912   0.0136     6.7049   0.0000

Residual standard error: 15.12 on 54 degrees of freedom
Multiple R-Squared: 0.4543
F-statistic: 44.96 on 1 and 54 degrees of freedom, the p-value is 1.241e-008
```

Compare these results to Model C results.