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:
    • $H_0: \mu(Y|X) = \beta_0$
    • $H_A: \mu(Y|X) = \beta_0 + \beta_1 X$
  - Equivalent to a test of $H_0: \beta = 0; H_A: \beta \neq 0$
• 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 ($\beta$'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

• $\mu(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$
• 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 $\sigma^2 = \sigma^2(Y|X)$
• 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 ($Lap^{nd}$) in the
  blue mussel as a function of distance (miles east
  of Southport, Connecticut).
Add a quadratic term

\[ \mu(\text{Frequency}|\text{Distance}) = \beta_0 + \beta_1 \text{Distance} + \beta_2 \text{Distance}^2 \]

Coefficients:

| Coefficient | Value | Std. Error | t value | Pr(>|t|) |
|-------------|-------|------------|---------|----------|
| (Intercept) | 0.201 | 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   |

Splus Output

\[ \mu(\text{Frequency}|\text{Distance}) = \beta_0 + \beta_1 \text{Distance} + \beta_2 \text{Distance}^2 \]

Coefficients:

| Coefficient | Value | Std. Error | t value | Pr(>|t|) |
|-------------|-------|------------|---------|----------|
| (Intercept) | 0.457682 | 0.01647738 | 28.3652 | 0.0000   |
| miles       | 0.0008 | 0.00032   | 2.618  | 0.0007   |
| I(miles^2)  | 0.0000 | 0.0000    | -2.3662 | 0.0342   |

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 spruce 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 Slewth, page 239.
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 B: Parallel Lines Model
Residuals vs Fitted

Residuals vs Fitted

Model C: Separate Lines Model
Residuals vs Fitted

Residuals vs Fitted
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-10

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-11

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

\[ H_0: \mu[P|E,L] = \beta_0 + \beta_1 E \]
\[ H_1: \mu[P|E,L] = \beta_0 + \beta_1 E + \beta_2 L + \beta_3 E \times L \]

Reduced:

- Terms added sequentially (first to last)
  - Elevation

Full:

- Terms added sequentially (first to last)
  - Location

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

Reject \( H_0 \) if \( F > F_{(0.05,2,60)} = 3.15 \)

Individual North/South Models

South Only

Coefficients:
  Value Std. Error t value Pr(>|t|)
(Intercept) -41.3384 12.4165 -3.3239 0.0016
Elevation[Location = 0] -0.0172 0.0115 -1.4964 0.1852

Residual standard error: 8.716 on 6 degrees of freedom
Multiple R-Squared: 0.2719
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 -3.3239 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: 0.4563
F-statistic: 44.96 on 1 and 54 degrees of freedom, the p-value is 2.41e-05

Compare these results to Model C results.