Feb 8: Dichotomous indicator variables

- What is the relationship between percentage of damaged trees and elevation for the North and the South?
- Consider 3 models; conclude *non-parallel lines model*
- Conclusions:
  - Rates of change in percentage damaged as a linear function of elevation differ in the North and South. In the North, the percentage increases with increased elevation. In the South, the percentage decreases as elevation increases.
- Critique:
  - Differing North and South sample sizes.
  - Weak linear relationship in South; elevation may not be the adequate predictor of percentage damaged.

Feb. 13, 15 Overview

- Ch. 10: Inferential tools in Multiple Regression
- Inference for single coefficients and for linear combinations of coefficients
- Indicator variables with >2 levels: bat data
- More on polynomial regression
- R^2 in multiple regression, R^2 adjusted
  - Ratio of explained variance to total variance
  - "adjusted": penalty for unnecessary *X's*

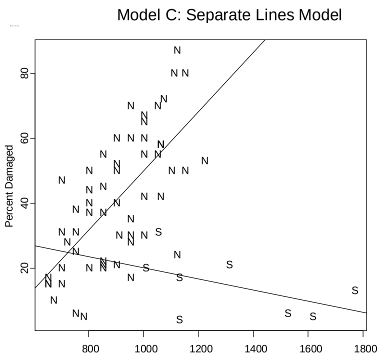
Longer range view of the course

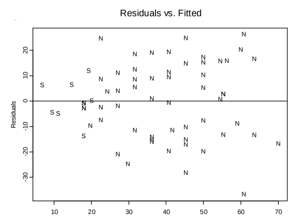
- Model checking and refinement
  - identifying and quantifying influence of observations
- Strategies for variable selection
  - dealing with multicollinearity
  - sequential variable selection procedures
  - model-fitting criteria
- Multiple regression for binary response variables and for Poisson counts

Further issues in the spruce data

- *Extra SS F-test*: Comparison of equal lines model with separate lines model.
- What is the best estimate of σ^2 ? Comparison of separate lines model with individual North/South models.
- Eliminate "elevation" term from final model?
- Confidence interval for slope terms in the North and South? *In North, SE of slope is a function of two parameter estimates.*

Model C





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.

Model C: Splus Output

```
Call: lm(formula = Perc.Damage ~ Elevation + Location +
Elevation:Location)

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 df, p-value is 2.825e-010

Correlation of Coefficients:
                   (Intercept) Location Elevation
         Location -0.9074
        Elevation -0.9799      0.8892
Location:Elevation  0.8098     -0.9683  -0.8264
```

Eliminate "elevation" term from final model?

- The elevation variable is not significant in the final separate lines model.
- Indicates that the slope for the South may not be very different from zero.
- Problem with deleting "elevation" variable: logical inconsistency to say that the effect of elevation depends on location (L:E term) but there is no elevation effect.
- Better to leave elevation term in the final model, but report and discuss a confidence interval for slope in the South (will include zero).

Separate Lines Model: CI's for slopes?

South: $\hat{\mu}\{D|E,L\}=\hat{\beta}_0+\hat{\beta}_1 E$

$\hat{\beta}_1 \pm t_{0.975;n-p} SE(\hat{\beta}_1)$ (-0.0558, 0.0214)

North: $\hat{\mu}\{D|E,L\}=(\hat{\beta}_0+\hat{\beta}_2)+(\hat{\beta}_1+\hat{\beta}_3) E$

$(\hat{\beta}_1+\hat{\beta}_3) \pm t_{0.975;n-p} SE(\hat{\beta}_1+\hat{\beta}_3)$

Variance, Covariance, Correlation

For random variables X and Y, and a constant

$Var(aX)=a^2 Var(X)$

$Var(a+X)=Var(X)$

$Var(X+Y)=Var(X)+Var(Y)+2Cov(X,Y)$

$Cov(X,Y)$ is the mean of $(X-\mu_X)(Y-\mu_Y)$

If $Cov(X,Y)=0$, X and Y are independent

$$\rho_{XY}=Cor(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Estimating the variance of a linear combination of random variables

- Linear combination of population parameters

$$\gamma=C_0\beta_0+C_1\beta_1+C_2\beta_2+\ldots+C_p\beta_p$$

- Estimate γ by

$$g=\hat{\gamma}=C_0\hat{\beta}_0+C_1\hat{\beta}_1+C_2\hat{\beta}_2+\ldots+C_p\hat{\beta}_p$$

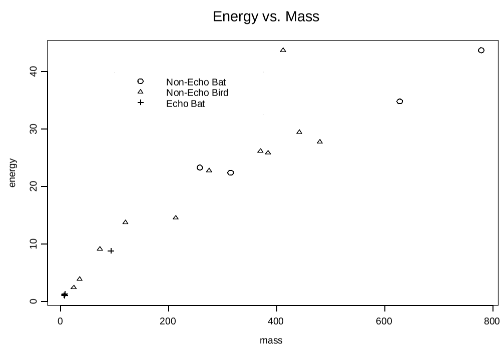
- Consider the case of $p=2$.

$$Var(g)=C_0^2 SE(\hat{\beta}_0)^2+C_1^2 SE(\hat{\beta}_1)^2+C_2^2 SE(\hat{\beta}_2)^2+2C_0C_1 Cov(\hat{\beta}_0,\hat{\beta}_1)$$
$$+2C_0C_2 Cov(\hat{\beta}_0,\hat{\beta}_2)+2C_1C_2 Cov(\hat{\beta}_1,\hat{\beta}_2)$$

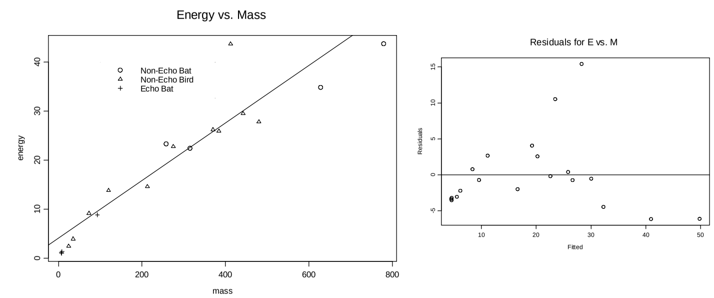
- $SE(g)=\sqrt{Var(g)}\,(d.f.=n-p)$

Bat Echolocation Dataset

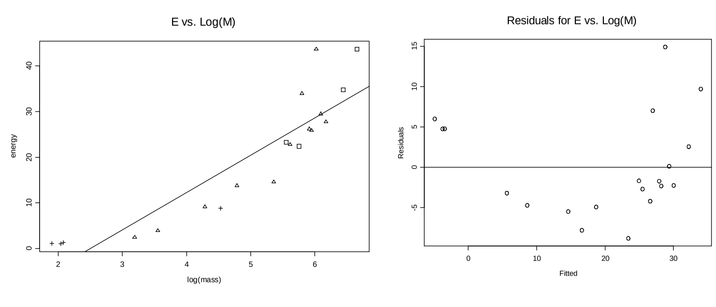
Is the energy expenditure the same for echolocating bats as for non-echolocating bats after accounting for body mass?



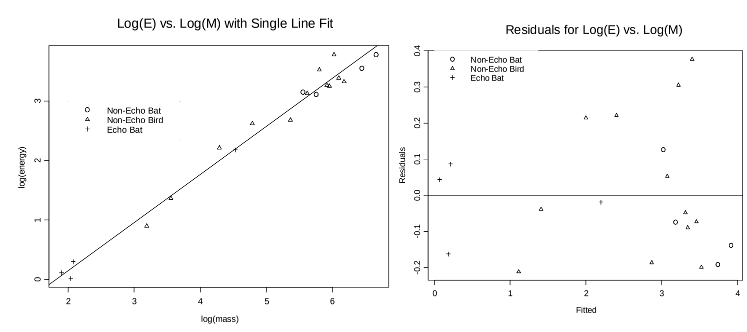
Linear Model, Untransformed Data



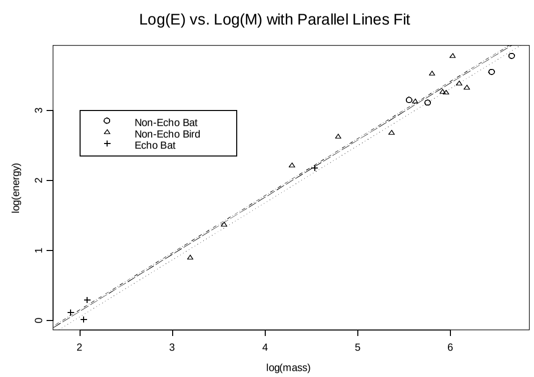
Log-transform Body Mass



Log(energy)~Log(mass)



Log(energy)~Log(mass)+type



Model A: Equal Lines

```
Coefficients:
                 Value Std. Error  t value Pr(>|t|)
(Intercept)  -1.4683   0.1372   -10.7046   0.0000
  log(mass)   0.8086   0.0268    30.1270   0.0000

Residual standard error: 0.18 on 18 degrees of freedom
Multiple R-Squared: 0.9806
F-statistic: 907.6 on 1 and 18 degrees of freedom, the
p-value is 1.11e-016

Correlation of Coefficients:
          (Intercept)
log(mass)  0.956
```

Model B: Parallel Lines Model

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  -1.5764   0.2872    -5.4880   0.0000
  log(mass)   0.8150   0.0445    18.2966   0.0000
      type2   0.1023   0.1142     0.8956   0.3837
      type3   0.0787   0.2027     0.3881   0.7030

Residual standard error: 0.186 on 16 degrees of freedom
Multiple R-Squared: 0.9815
F-statistic: 283.6 on 3 and 16 degrees of freedom,
               the p-value is 4.463e-014

Correlation of Coefficients:
          (Intercept) log(mass)   type2
log(mass) -0.9462
    type2 -0.5856      0.3403
    type3 -0.8685      0.7610    0.6326
```

ANOVA Tables

Equal Lines

```
          Df Sum of Sq  Mean Sq  F Value         Pr(F)
log(mass)  1  29.39191 29.39191 907.6384 1.110223e-016
Residuals 18   0.58289  0.03238
```

Parallel Lines

```
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
log(mass)  1  29.39191 29.39191 849.9108 0.0000000
     type  2   0.02957  0.01479   0.4276 0.6593195
Residuals 16   0.55332  0.03458
```

Centering Trick: CI, PI for NE Bats at Mass=100g

```
Call: lm(formula = log(energy) ~ I(log(mass) - log(100)) +
as.factor(type))

Coefficients:
                         Value Std. Error t value Pr(>|t|)
            (Intercept)  2.1767  0.1144    19.0264  0.0000
I(log(mass) - log(100))  0.8150  0.0445    18.2966  0.0000
                  type2  0.1023  0.1142     0.8956  0.3837
                  type3  0.0787  0.2027     0.3881  0.7030

Residual standard error: 0.186 on 16 degrees of freedom
Multiple R-Squared: 0.9815
F-statistic: 283.6 on 3 and 16 df, p-value is 4.463e-014
```

Note: Six separate model fits are required to create Display 10.9.

For Birds, need to recode data to make Birds the "reference level".
That is, let Birds variable have "Type"=1 and repeat the procedure.