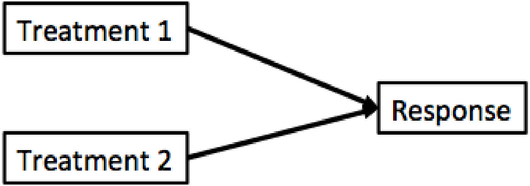
Multiple Predictor Variables: Regression & the General Linear Model

Contrasts for a Multiway ANOVA

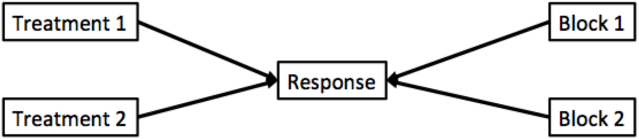
```
library(contrast)
contrast(zoop_lm,
         list(treatment="low", block=levels(zoop$block)),
         list(treatment="high", block=levels(zoop$block)),
         type="average")

# lm model parameter contrast
#
#   Contrast   S.E.     Lower Upper    t df Pr(>|t|)
# 1     0.62 0.2895 -0.04755 1.288 2.14  8   0.0646
```

Multiple Predictor Variables: Regression & the General Linear Model

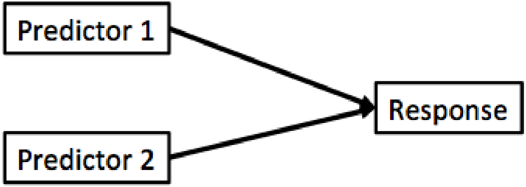
One-Way ANOVA Graphically



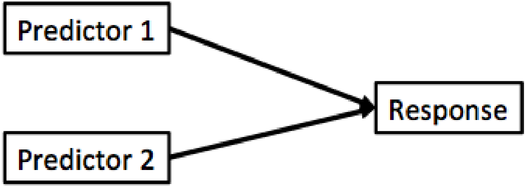
Two-Way ANOVA Graphically



Multiple Linear Regression?

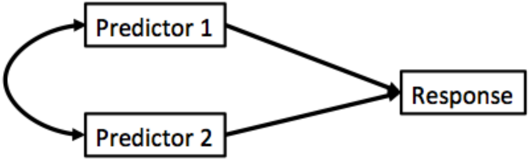


Multiple Linear Regression?



Note no connection between predictors, as in ANOVA. This is ONLY true if we have manipulated it so that there is no relationship between the two.

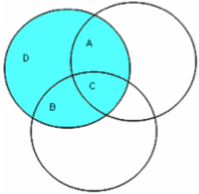
Multiple Linear Regression

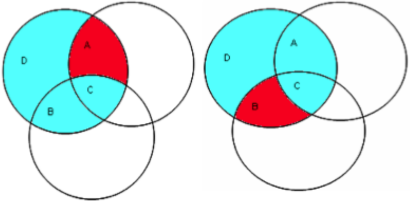


Curved double-headed arrow indicates COVARIANCE between predictors that we must account for.

Semi-Partial Correlation

- Semi-Partial correlation asks how much of the variation in a response is due to a predictor after the contribution of other predictors has been removed
- How much would R^2 change if a variable was removed?
- A / (A+B+C+D)
- $sr_{y1} = \frac{r_{y1} - r_{y2} y_{12}}{\sqrt{1 - r_{12}^2}}$





Calculating Multiple Regression Coefficients with OLS

$$\boldsymbol{Y} = \boldsymbol{b}\boldsymbol{X} + \boldsymbol{\epsilon}$$

Calculating Multiple Regression Coefficients with OLS

$$\boldsymbol{Y} = \boldsymbol{b}\boldsymbol{X} + \boldsymbol{\epsilon}$$

Remember in Simple Linear Regression $b = \frac{cov_{xy}}{var_x}$?

Calculating Multiple Regression Coefficients with OLS

$$\boldsymbol{Y = bX + \epsilon}$$

Remember in Simple Linear Regression $b = \frac{cov_{xy}}{var_x}$?

In Multiple Linear Regression $\boldsymbol{b = cov_{xy} S_x^{-1}}$

where $\boldsymbol{cov_{xy}}$ is the covariances of $\boldsymbol{x_i}$ with $\boldsymbol{y}$ and $\boldsymbol{S_x^{-1}}$ is the variance/covariance matrix of all *Independent variables*

Calculating Multiple Regression Coefficients with OLS

$$\boldsymbol{Y = bX + \epsilon}$$

Remember in Simple Linear Regression $b = \frac{cov_{xy}}{var_x}$?

In Multiple Linear Regression $\boldsymbol{b = cov_{xy} S_x^{-1}}$

where $\boldsymbol{cov_{xy}}$ is the covariances of $\boldsymbol{x_i}$ with $\boldsymbol{y}$ and $\boldsymbol{S_x^{-1}}$ is the variance/covariance matrix of all *Independent variables*

OR $bi = \frac{cov_{xy} - \sum cov_{x1xj} b_j}{var_(x)}$

Calculating Multiple Regression Coefficients with OLS

$$\boldsymbol{Y} = \boldsymbol{b}\boldsymbol{X} + \boldsymbol{\epsilon}$$

Calculating Multiple Regression Coefficients with OLS

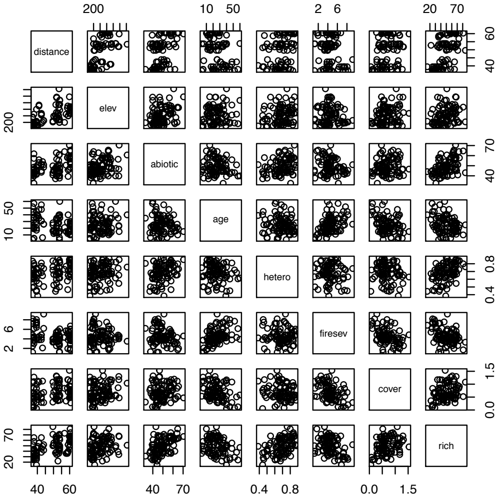
$$\boldsymbol{Y = bX + \epsilon}$$

Coefficient Estimates: $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{cov_{xy} S_x^{-1}}$

Coefficient Variance: $Var[\hat{\beta}_i] = \frac{\boldsymbol{\sigma^2}}{\boldsymbol{SXX_i}}$

Five year study of wildfires & recovery in Southern California shurblands in 1993. 90 plots (20 x 50m)
(data from Jon Keeley et al.)

Many Things may Influence Species Richness



Many Things may Influence Species Richness

```
klm <- lm(rich ~ cover + firesev + hetero, data=keeley)
```

Checking for Multicollinearity: Correlation Matrices

```
with(keeley, cor(cbind(cover, firesev, hetero)))

#             cover  firesev   hetero
# cover     1.0000 -0.43713 -0.16838
# firesev -0.4371  1.00000 -0.05236
# hetero  -0.1684 -0.05236  1.00000
```

Correlations over 0.4 can be problematic, but, they may be OK even as high as 0.8. Beyond this, are you getting unique information from each variable?

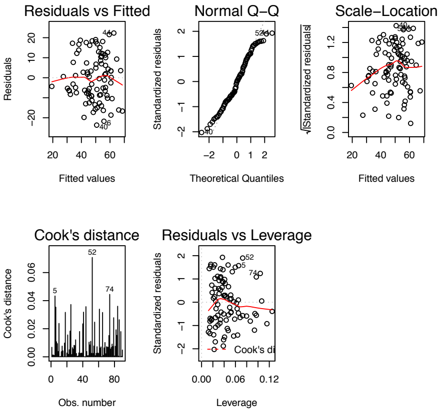
Checking for Multicollinearity: Variance Inflation Factor

$$VIF = \frac{1}{1 - R_j^2}$$

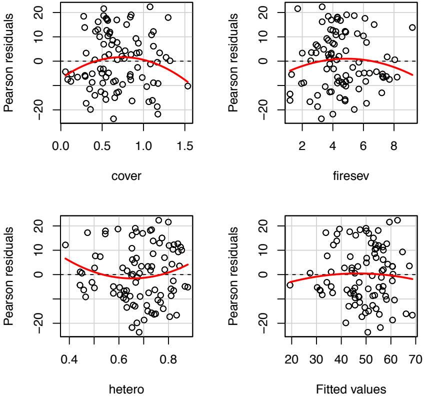
```
vif(klm)

#    cover firesev  hetero
#    1.295   1.262   1.050
```

VIF $>$ 5 or 10 can be problematic and indicate an unstable solution.

Other Diagnostics as Usual!



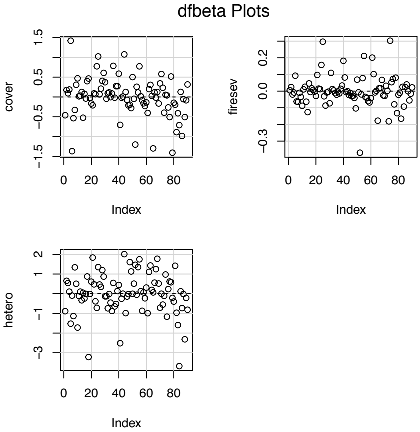
Other Diagnostics as Usual!




```
#              Test stat Pr(>|t|)
# cover           -1.602    0.113
# firesev         -1.087    0.280
```

New Diagnostic for Outliers: Leave One Out

```r
dfbetaPlots(klm)
```



Which Variables Explained Variation: Type II Marginal SS

```
Anova(klm)

# Anova Table (Type II tests)
#
# Response: rich
#           Sum Sq Df F value  Pr(>F)
# cover       1674  1   12.01 0.00083
# firesev      636  1    4.56 0.03554
# hetero      4865  1   34.91 6.8e-08
# Residuals  11985 86
```

If order of entry matters, can use type I. Remember, what models are you comparing?

The coefficients

```
summary(klm)$coef

#               Estimate Std. Error t value  Pr(>|t|)
# (Intercept)    1.679    10.6737  0.1573 8.754e-01
# cover         15.558     4.4886  3.4661 8.264e-04
# firesev       -1.817     0.8506 -2.1357 3.554e-02
# hetero        65.992    11.1694  5.9082 6.757e-08

cat(paste("R^2 = ", round(summary(klm)$r.squared, 2), sep=""))

# R^2 = 0.41
```

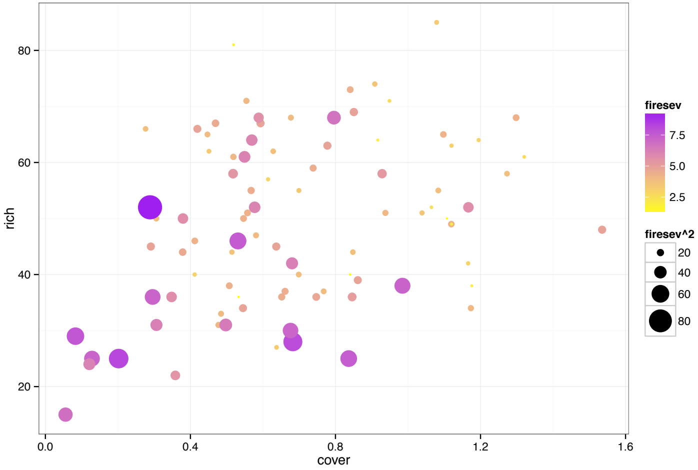
If order of entry matters, can use type I. Remember, what models are you comparing?

Comparing Coefficients on the Same Scale

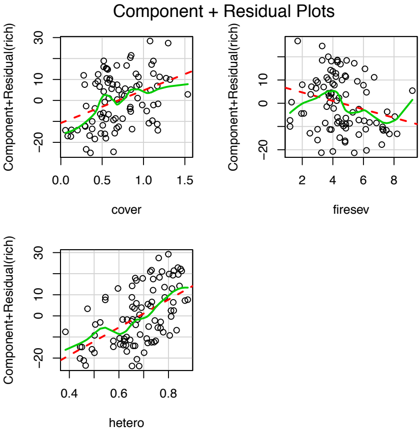
$$r_{xy} = b_{xy}\frac{sd_x}{sd_y}$$

```
library(QuantPsyc)
lm.beta(klm)

#    cover firesev   hetero
#   0.3267 -0.1987   0.5016
```

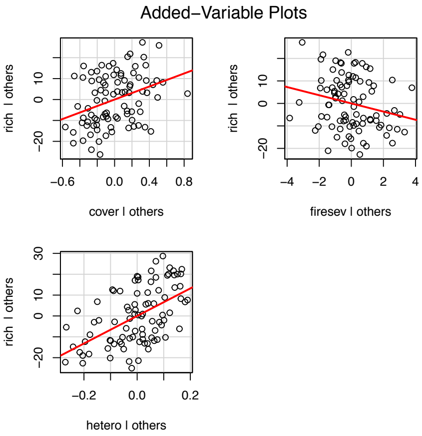
Visualization of Multivariate Models is Difficult



Component-Residual Plots Aid in Visualization



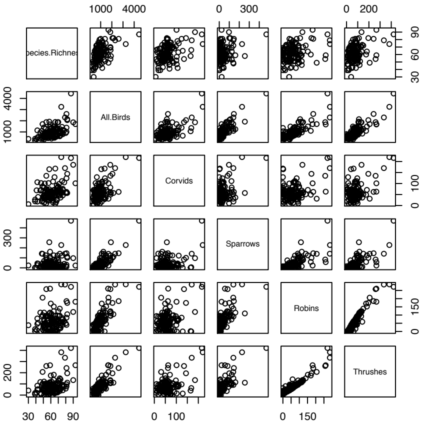
Added Variable Plots for Unique Contribution of a Variable



Analagous to the A part of the three-circle diagram from earlier.

Exercise: Bird Species Richness

- Which bird abundances influence Species Richness?
- Can we use every variable?
- Visualize Results



All of the Birds!

```
wnv_lm_vif <- lm(Species.Richness ~ Corvids +
                                    Sparrows +
                                    Robins +
                                    Thrushes , data=wnv)
```

Correlation Problems

```
cor(wnv[,c(3:8)])

#                 Species.Richness All.Birds Corvids
# Species.Richness          1.0000    0.5058  0.4326
# All.Birds                 0.5058    1.0000  0.5964
# Corvids                   0.4326    0.5964  1.0000
# Sparrows                  0.2406    0.8465  0.3846
# Robins                    0.2928    0.8075  0.4028
# Thrushes                  0.3859    0.8531  0.4960
#                 Sparrows Robins Thrushes
# Species.Richness   0.2406 0.2928   0.3859
# All.Birds          0.8465 0.8075   0.8531
# Corvids            0.3846 0.4028   0.4960
# Sparrows           1.0000 0.7083   0.7286
# Robins             0.7083 1.0000   0.9572
```

Multicollinearity Problems

```
vif(wnv_lm_vif)

#   Corvids Sparrows    Robins Thrushes
#     1.449    2.145    13.050   15.060
```

Odd Results from Robins and Sparrows

```
summary(wnv_lm_vif)

#
# Call:
# lm(formula = Species.Richness ~ Corvids + Sparrows + Robins +
#     Thrushes, data = wnv)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -24.997  -6.250  -0.093   6.827  22.074
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  53.3019     1.6681   31.95   <2e-16
# Corvids       0.0732     0.0262    2.79   0.0060
# Sparrows     -0.0150     0.0202   -0.74   0.4596
# Robins       -0.1235     0.0502   -2.46   0.0152
# Thrushes      0.1538     0.0471    3.27   0.0014
#
```

A New Model

```
wnv_lm <- lm(Species.Richness ~ Corvids +
                                Sparrows +
                                Robins, data=wnv)
```

No Multicollinearity Problem

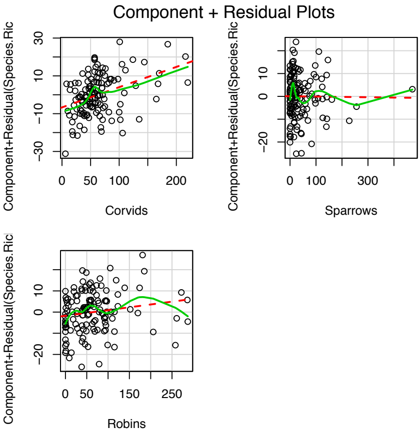
```
vif(wnv_lm)

#   Corvids Sparrows   Robins
#     1.223    2.055    2.091
```

A Corvid Story

```
Anova(wnv_lm)

# Anova Table (Type II tests)
#
# Response: Species.Richness
#            Sum Sq  Df F value   Pr(>F)
# Corvids      1793   1   18.36 3.6e-05
# Sparrows        1   1    0.01    0.94
# Robins        160   1    1.64    0.20
# Residuals  12306 126
```

A Corvid Story

```
crPlots(wnv_lm)
```



The General Linear Model

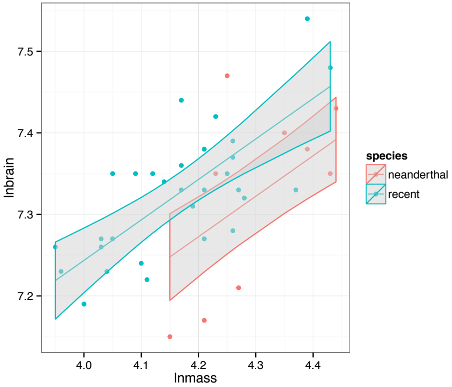
$$\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{\epsilon}$$

- This equation is huge. X can be anything - categorical, continuous, etc.
- One easy way to see this is if we want to control for the effect of a covariate - i.e., ANCOVA
- Type of SS matters, as 'covariate' is de facto 'unbalanced'

Neanderthals and the General Linear Model

How big was their brain?

Analysis of Covariance (control for a covariate)



ANCOVA: Evaluate a categorical effect(s), controlling for a *covariate* (parallel lines)
Groups modify the *intercept*.

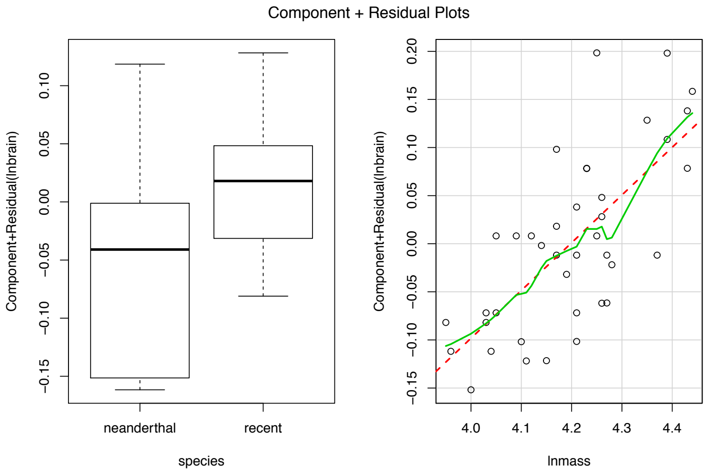
Exercise: Fit like a cave man

- Fit a model that will describe brain size from this data
- Does species matter? Compare type I and type II SS results
- Use Component-Residual plots to evaluate results

Type of SS Matters

```
# Analysis of Variance Table
#
# Response: lnbrain
#           Df Sum Sq Mean Sq F value  Pr(>F)
# species    1 0.0001  0.0001    0.01    0.91
# lnmass     1 0.1300  0.1300   29.28 4.3e-06
# Residuals 36 0.1599  0.0044

# Anova Table (Type II tests)
#
# Response: lnbrain
#           Sum Sq Df F value  Pr(>F)
# species   0.0276  1     6.2   0.017
# lnmass    0.1300  1    29.3 4.3e-06
# Residuals 0.1599 36
```

Species Effect



Species Effect

```
summary(neand_lm)$coefficients

#                Estimate Std. Error t value  Pr(>|t|)
# (Intercept)     5.18807    0.39526  13.126 2.736e-15
# speciesrecent   0.07028    0.02822   2.491 1.749e-02
# lnmass          0.49632    0.09173   5.411 4.262e-06

summary(neand_lm)$r.squared

# [1] 0.4486
```