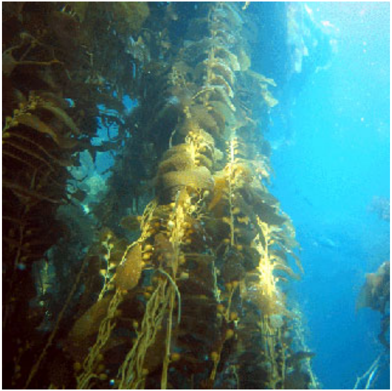
Handling Categorical Predictors: plyr, ANOVA, and more

Group Properties: Kelp

- Kelp sampled at multiple sites annually
- At each transect, holdfast diameter and # of fronds counted



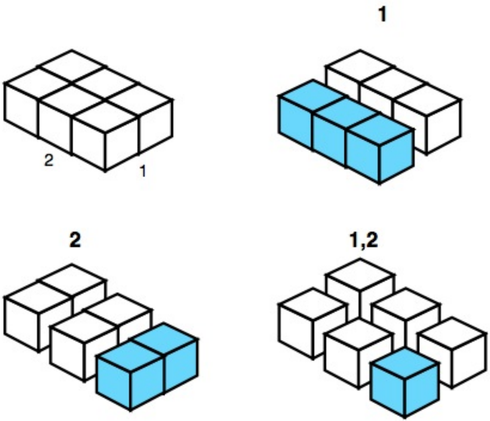
How can we get quick summaries by site?, year, or both?

```
#    YEAR MONTH       DATE SITE TRANSECT QUAD SIDE FRONDS
# 2  2000     9 2000-09-28 BULL        1   20            4
# 8  2000     9 2000-09-28 BULL        2   20           11
# 9  2000     9 2000-09-28 BULL        2   20           16
# 10 2000     9 2000-09-28 BULL        2   20           34
# 16 2000     9 2000-09-28 BULL        3   20           27
# 17 2000     9 2000-09-28 BULL        3   20           38
#    HLD_DIAM
# 2         7
# 8        65
# 9        55
# 10       55
# 16       65
# 17       60
```

For loops for Summarization by Site

```r
# number of groups
k <- length(levels(kelp$SITE))
#blank means vector
means <- rep(NA, k)
#the loop
for(i in 1:k) {
  #split the data first
  subdata <- subset(kelp, kelp$SITE == levels(kelp$SITE)[i])

  #apply the means function,
  #combine with previous means
  means[i] <- mean(subdata$FRONDS, na.rm=T)
}
```

The Split, Apply, Combine Strategy



Wickham 2011

ddply from Hadley Wickham's plyr library

```r
library(plyr)
#
kelpMeans <- ddply(kelp, .(SITE), summarize,
                   mean.FRONDS = mean(FRONDS, na.rm=T))
```

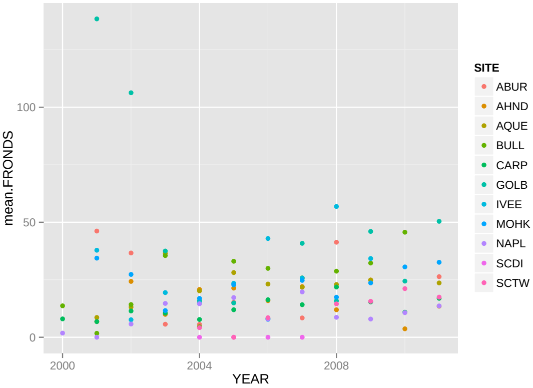
ddply from Hadley Wickham's plyr library

```
kelpMeans

#    SITE mean.FRONDS
# 1  ABUR       29.26
# 2  AHND       17.63
# 3  AQUE       21.04
# 4  BULL       27.30
# 5  CARP       13.11
# 6  GOLB       42.16
# 7  IVEE       25.81
# 8  MOHK       20.04
# 9  NAPL       13.16
# 10 SCDI        0.00
# 11 SCTW       14.73
```

Multiple Groups & ddply

```
kelpMeans2 <- ddply(kelp, .(YEAR, SITE), summarize,
                    mean.FRONDS = mean(FRONDS, na.rm=T))
```

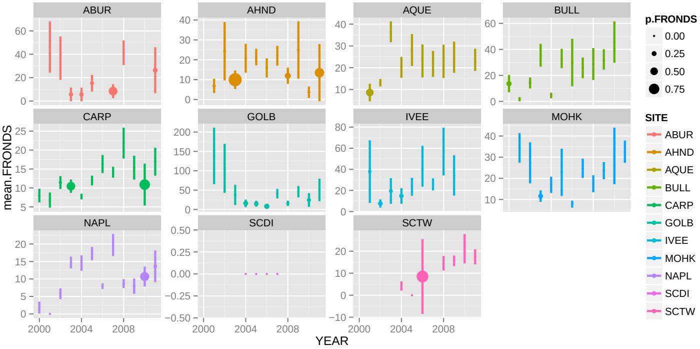
Multiple Groups & ddply



Complex Functions & ddply

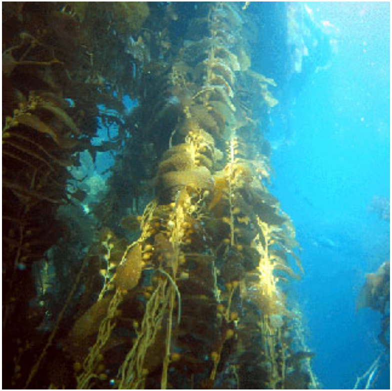
```
kelpMeans3 <- ddply(kelp, .(YEAR, SITE), function(aFrame){
  #calculate metrics for a 1-sample T test comparison against
  #grand mean of 10 fronds/m^2
  m <- mean(aFrame$FRONDS, na.rm=T)
  n<-length(na.omit(aFrame$FRONDS))
  se <- sd(aFrame$FRONDS, na.rm=T)/sqrt(n)
  t <- (m-10)/se
  p <- 2*pt(abs(t), df=n-1, lower.tail=F)

  # return everything
  return(c(mean.FRONDS=m, n.FRONDS=n,
           se.FRONDS=se, t.FRONDS=t,
           p.FRONDS = p))
})
```

Complex Functions & ddply



Exercise: Correlation!

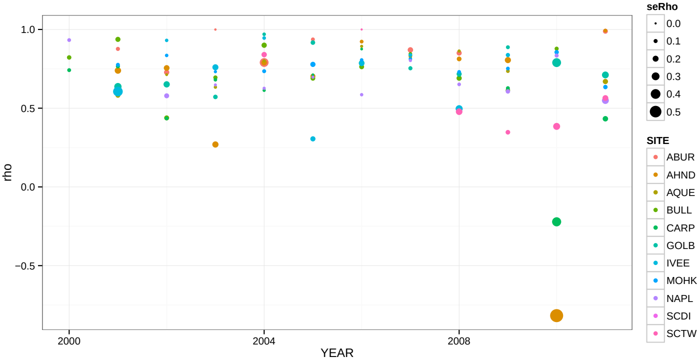
- Evaluate the correlation between fronds and holdfasts by site and year
- Plot it
- Extra: include the SE of the correlation visually



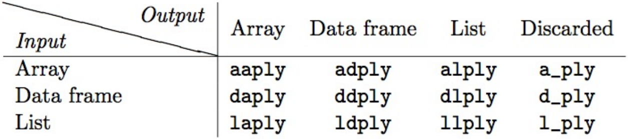
Exercise: Correlation!

```r
kelpCor <- ddply(kelp, .(YEAR, SITE), function(adf){
  #first get the correlation
  cors <- cor(adf$FROND, adf$HLD_DIAM)

  #use this to calculate it's SE
  seCor <- sqrt((1-cors^2) / (nrow(adf)-2))

  #return both
  return(c(rho = cors, seRho = seCor))

})
```

Exercise: Correlation!



Many plyr Functions

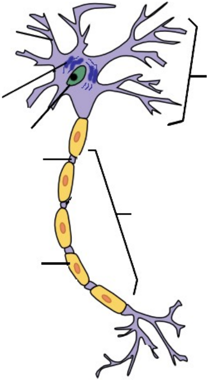


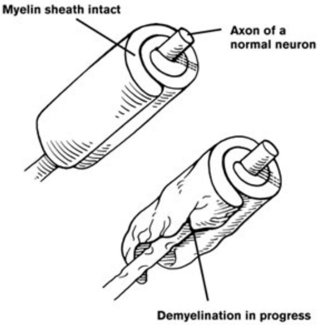
Output *Input*	Array	Data frame	List	Discarded
Array	`aaply`	`adply`	`alply`	`a_ply`
Data frame	`daply`	`ddply`	`dlply`	`d_ply`
List	`laply`	`ldply`	`llply`	`l_ply`

Also r*ply to replicate an action and return an object. Great for simulation.

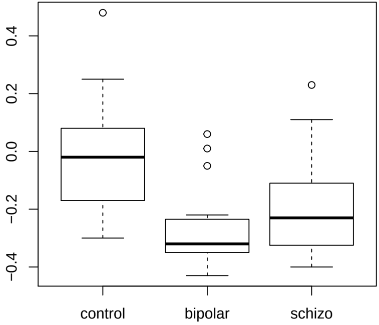
See also `colwise` and `each` for everyday use!

Categorical Predictors: Gene Expression and Mental Disorders





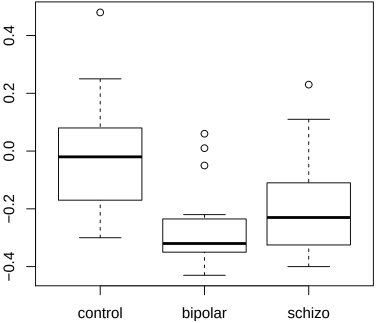
Categorical Predictors



How do we determine the importance of categorical predictors?

Aside: Reordering Factors

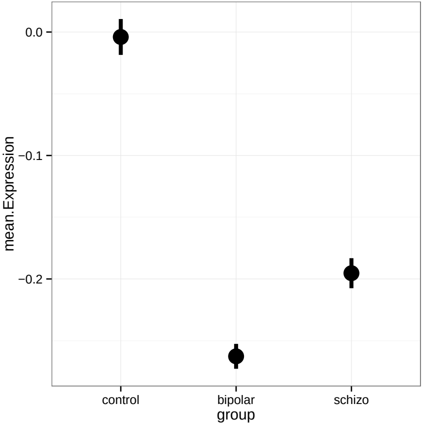
```r
brainGene$group <- factor(brainGene$group,
                          levels=c("control", "bipolar", "schizo"))
```



Categorical Predictors Ubiquitous

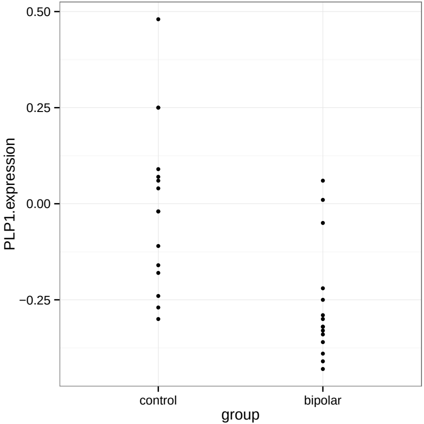
- Treatments in an Experiment
- Spatial groups - plots, Sites, States, etc.
- Individual sampling units
- Temporal groups - years, seasons, months

Traditional Way to Think About Categories

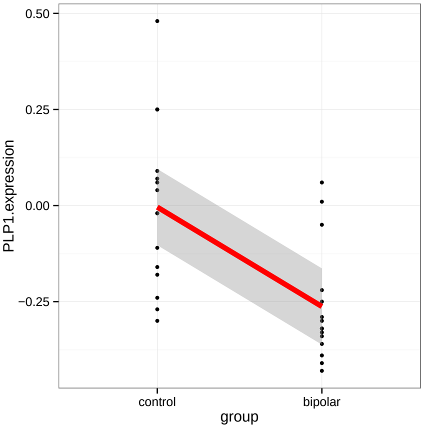


What is the variance between groups v. within groups?

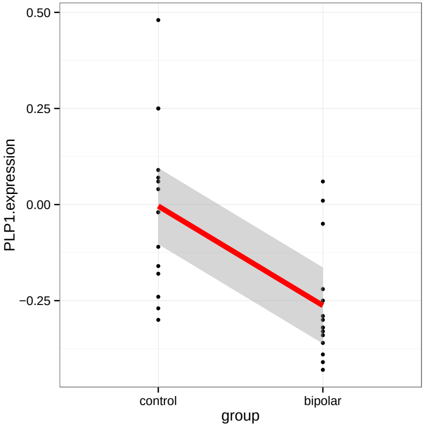
But How is the Model Fit?



But How is the Model Fit?

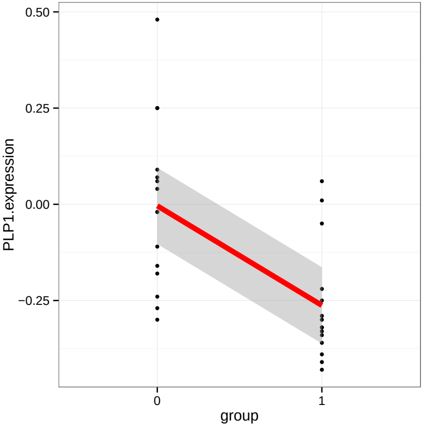


But How is the Model Fit?



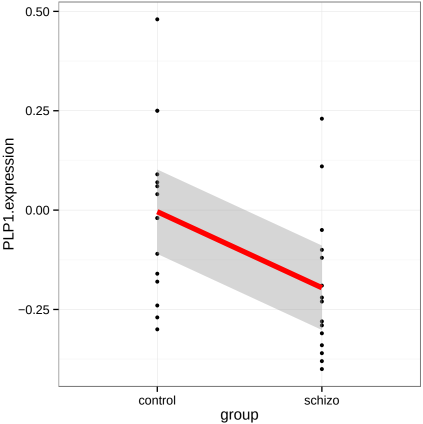
Underlying linear model with control = intercept, dummy variable for bipolar

But How is the Model Fit?



Underlying linear model with control = intercept, dummy variable for bipolar

But How is the Model Fit?



Underlying linear model with control = intercept, dummy variable for schizo

Different Ways to Write a Categorical Model

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

Different Ways to Write a Categorical Model

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

Different Ways to Write a Categorical Model

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

$$y_j = \beta_0 + \sum \beta_i x_i + \epsilon_j, \qquad x_i = 0, 1$$

Different Ways to Write a Categorical Model

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

$$y_j = \beta_0 + \sum \beta_i x_i + \epsilon_j, \qquad x_i = 0, 1$$

x_i inidicates presence/abscence of a category

Different Ways to Write a Categorical Model

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

$$y_j = \beta_0 + \sum \beta_i x_i + \epsilon_j, \qquad x_i = 0, 1$$

x_i inidicates presence/abscence of a category

Traditional ANOVA special case where all x_i are orthogonal

Different Ways to Write a Categorical Model

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

$$y_j = \beta_0 + \sum \beta_i x_i + \epsilon_j, \qquad x_i = 0, 1$$

x_i inidicates presence/abscence of a category

Traditional ANOVA special case where all x_i are orthogonal

Often one category set to β_0 for ease of fitting

This is a Linear Model

```
bg.sub.lm <- lm(PLP1.expression ~ group, data=brainGene)
```

Hypothesis Testing with a Categorical Model: ANOVA

$$H_0 = \mu_1 = \mu 2 = \mu 3 = ...$$

Hypothesis Testing with a Categorical Model: ANOVA

$$H_0 = \mu_1 = \mu 2 = \mu 3 = ...$$

OR

$$\beta_0 = \mu, \qquad \beta_i = 0$$

Assumptions of Ordinary Least Squares Regression

- Independence of data points
- Normality within groups
- Homoscedasticity (homogeneity of variance)

F-Test to Compare

$$SS_{Total} = SS_{Between} + SS_{Within}$$

$SS_{Between} = \sum_i \sum_j (\bar{Y_i} - \bar{Y})^2$, df=k-1

$SS_{Within} = \sum_i \sum_j (Y_{ij} - \bar{Y_i})^2$, df=n-k

F-Test to Compare

$$SS_{Total} = SS_{Between} + SS_{Within}$$

$$SS_{Between} = \sum_i \sum_j (\bar{Y_i} - \bar{Y})^2, \text{ df=k-1}$$

$$SS_{Within} = \sum_i \sum_j (Y_{ij} - \bar{Y_i})^2, \text{ df=n-k}$$

To compare them, we need to correct for different DF. This is the Mean Square.

MS = SS/DF, e.g, $MS_W = \frac{SS_W}{n-k}$

F-Test to Compare

$F = \frac{MS_B}{MS_W}$ with DF=k-1,n-k

F-Test to Compare

$F = \frac{MS_B}{MS_W}$ with DF=k-1,n-k

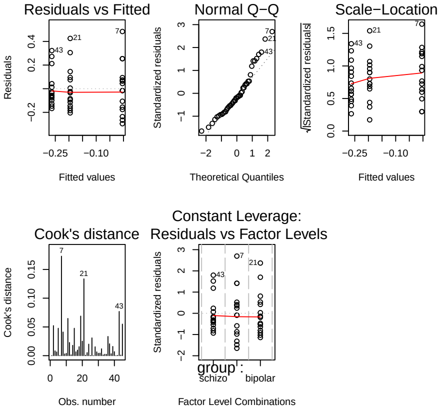
(note similarities to SS_R and SS_E notation of regression)

ANOVA

```
anova(bg.sub.lm)

# Analysis of Variance Table
#
# Response: PLP1.expression
#           Df Sum Sq Mean Sq F value Pr(>F)
# group      2   0.54  0.2701    7.82 0.0013
# Residuals 42   1.45  0.0345
```

Inspecting Assumptions

```r
par(mfrow=c(2,3))
plot(bg.sub.lm, which=1:5 )
```



Levene's Test of Homogeneity of Variance

```
library(car)

# Loading required package:  MASS
# Loading required package:  nnet

leveneTest(PLP1.expression ~ group, data=brainGene)

# Levene's Test for Homogeneity of Variance (center = median)
#       Df F value Pr(>F)
# group  2    1.01   0.37
#       42
```

Levene's test robust to departures from normality

What do I do if I Violate Assumptions?

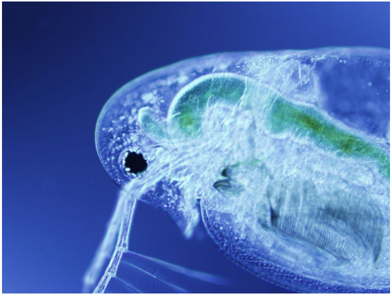
- Nonparametric Kruskal-Wallace (uses ranks)
- Transform?
- GLM with ANODEV

Kruskal Wallace Test

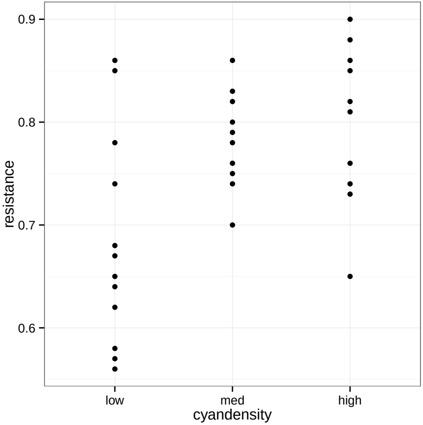
```
kruskal.test(PLP1.expression ~ group, data=brainGene)

#
#  Kruskal-Wallis rank sum test
#
# data:  PLP1.expression by group
# Kruskal-Wallis chi-squared = 13.2, df = 2, p-value =
# 0.001361
```

Exercise: Daphnia Resistance

- Plot the mean and SE of the data by group
- Evaluate whether the data is appropriate for ANOVA
- Fit an ANOVA and check diagnostics
- Evaluate results & compare to Kruskal-Wallace and a glm with a Gamma distribution



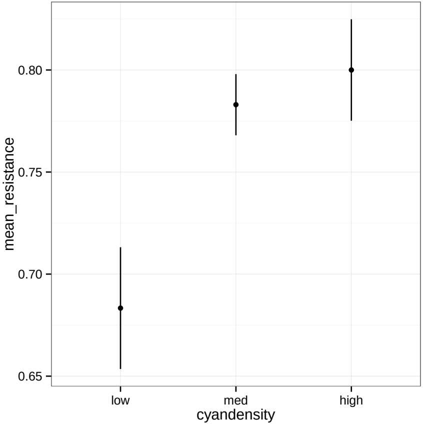
Daphnia Data



Daphnia Means

```r
#first use plyr to get means and SE
dsummary <- ddply(daphnia, .(cyandensity), summarize,
                  mean_resistance = mean(resistance),
                  se = sd(resistance) / sqrt(length(resistance)))
#
ggplot(dsummary, aes(x=cyandensity, y=mean_resistance,
                     ymin=mean_resistance-se,
                     ymax=mean_resistance+se)) +
  geom_pointrange() + theme_bw()
```

Daphnia Means



How about HOV?

```
leveneTest(resistance ~ cyandensity, data=daphnia)

# Levene's Test for Homogeneity of Variance (center = median)
#       Df F value Pr(>F)
# group  2       2   0.15
#       29
```

ANOVA shows an Effect

```
daphniaLM <- lm(resistance ~ cyandensity, data=daphnia)
anova(daphniaLM)

# Analysis of Variance Table
#
# Response: resistance
#             Df Sum Sq Mean Sq F value Pr(>F)
# cyandensity  2 0.0892  0.0446    6.69 0.0041
# Residuals   29 0.1933  0.0067
```

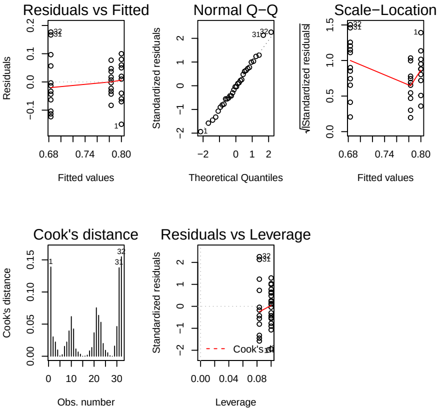
KW shows an Effect

```
#
#  Kruskal-Wallis rank sum test
#
# data:  resistance by cyandensity
# Kruskal-Wallis chi-squared = 8.2, df = 2, p-value =
# 0.01658
```

Bad GLM Does Not

```
# Analysis of Deviance Table
#
# Model: Gamma, link: identity
#
# Response: resistance
#
# Terms added sequentially (first to last)
#
#
#             Df Deviance Resid. Df Resid. Dev
# NULL                           31      0.529
# cyandensity  2    0.162        29      0.367
```

Diagnostics Also Good


Residuals vs Fitted
Residuals
Fitted values
0.68
0.74
0.80
−0.1
0.0
0.1
0.2
32
31
1
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−2
0
1
2
−2
−1
0
1
2
31
32
1
Scale−Location
√|Standardized residuals|
Fitted values
0.68
0.74
0.80
0.0
0.5
1.0
1.5
32
31
1
Cook's distance
Cook's distance
Obs. number
0
10
20
30
0.00
0.05
0.10
0.15
1
32
31
Residuals vs Leverage
Standardized residuals
Leverage
0.00
0.04
0.08
−2
−1
0
1
2
32
31
1
Cook's d