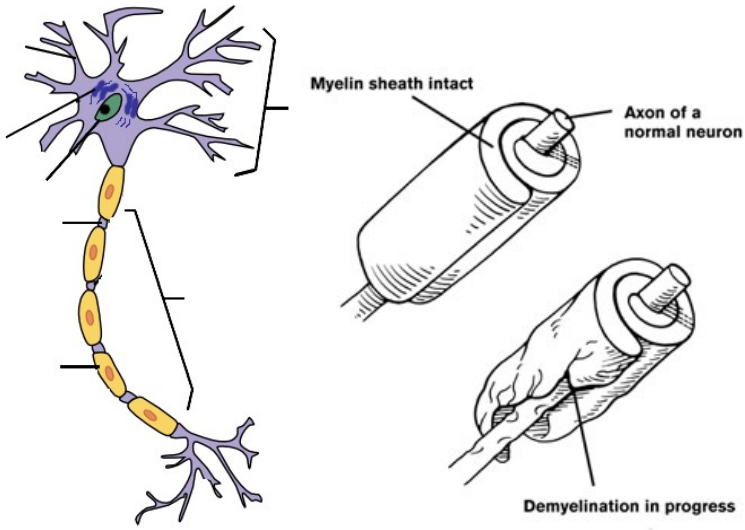
Handling Categorical Predictors: ANOVA

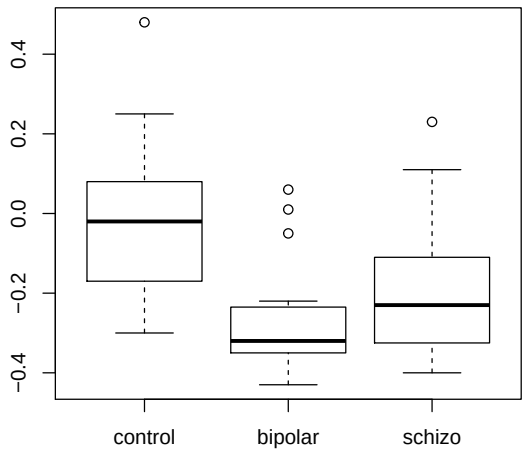
I Hate Lines!

- When we think of experiments, we think of manipulating categories
- Control, Treatment 1, Treatment 2
- Models with Categorical Predictors still reflect an underlying data generating process and error generating process
- In many ways, it's like having many processed generating data
- Linear (or Generalized Linear) Models still underlie ANOVA or other analyses with categorical predictors

Categorical Predictors: Gene Expression and Mental Disorders



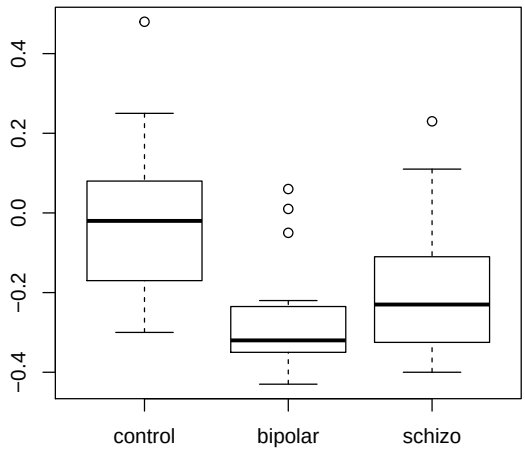
Categorical Predictors



How do we determine the importance of categorical predictors?

Aside: Reordering Factors

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



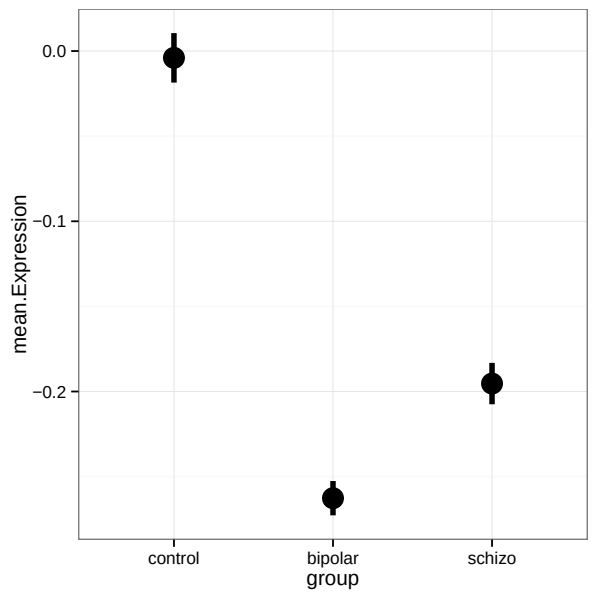
How do we determine the importance of categorical predictors?

Categorical Predictors Ubiquitous

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

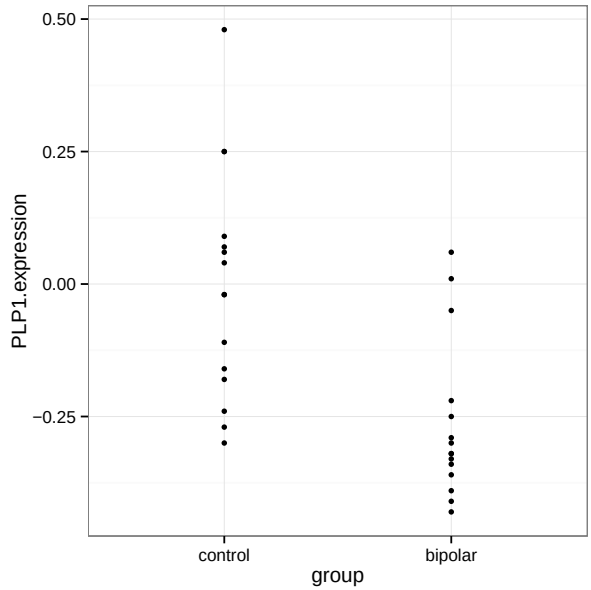
Modeling Categorical Predictors

Traditional Way to Think About Categories

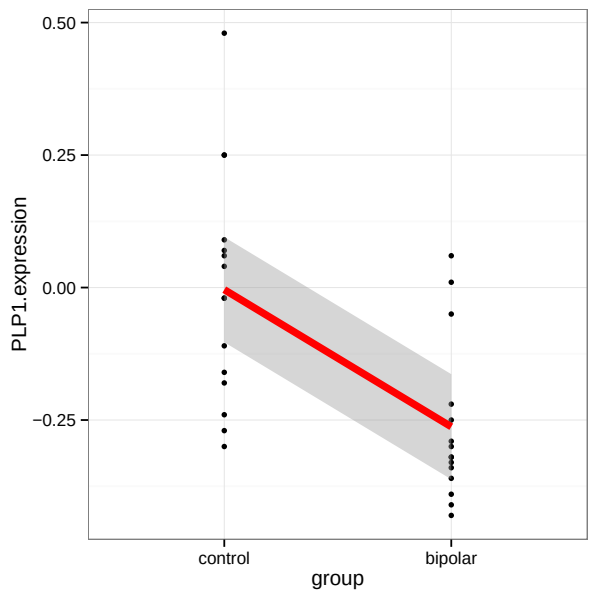


What is the variance between groups v. within groups?

But How is the Model Fit?

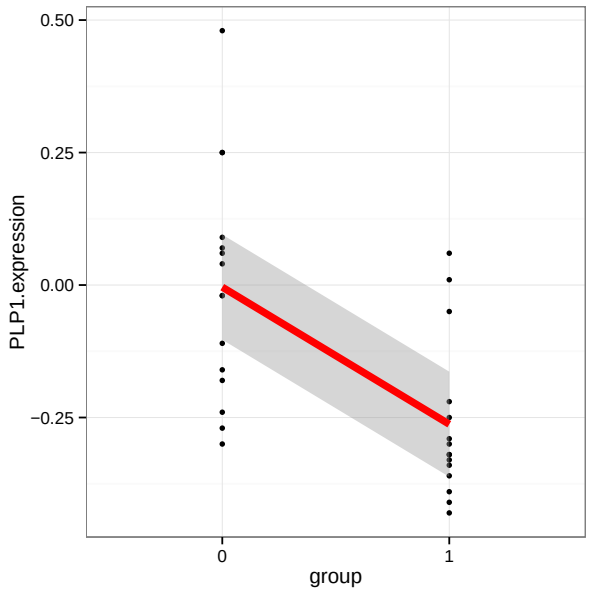


But How is the Model Fit?



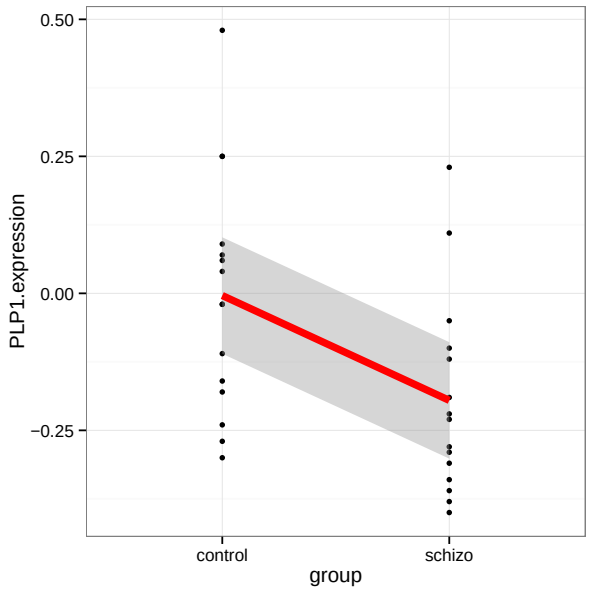
Underlying linear model with control = intercept, dummy variable for bipolar

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

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

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

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

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)
```

Evaluating Categorical Predictors

Hypothesis Testing with a Categorical Model: ANOVA

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

OR

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

F-Test to Compare

Regression: $SS_{Total} = SS_{Model} + SS_{Error}$
$SS_{Total} = SS_{Between} + SS_{Within}$

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

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

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

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

(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 0.54025 0.270127  7.8231 0.001294
# Residuals 42 1.45023 0.034529
```

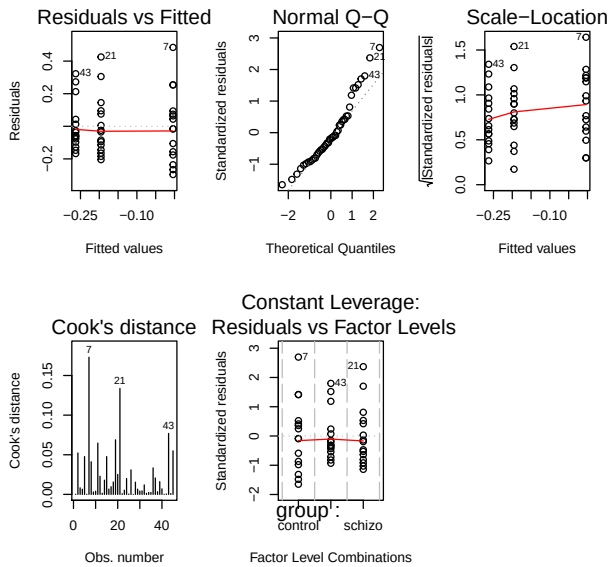

Is using ANOVA valid?

Assumptions of Ordinary Least Squares Regression

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

Inspecting Assumptions

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



Levene's Test of Homogeneity of Variance

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

# Levene's Test for Homogeneity of Variance (center = median)
#       Df F value Pr(>F)
# group  2  1.0067 0.3741
#       42
```

Levene's test robust to departures from normality

What do I do if I Violate Assumptions?

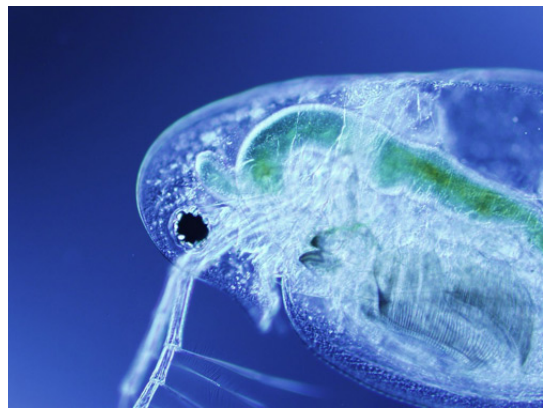
- Nonparametric Kruskal-Wallace (uses ranks)
- Transform?
- GLM with ANODEV

Kruskal Wallace Test

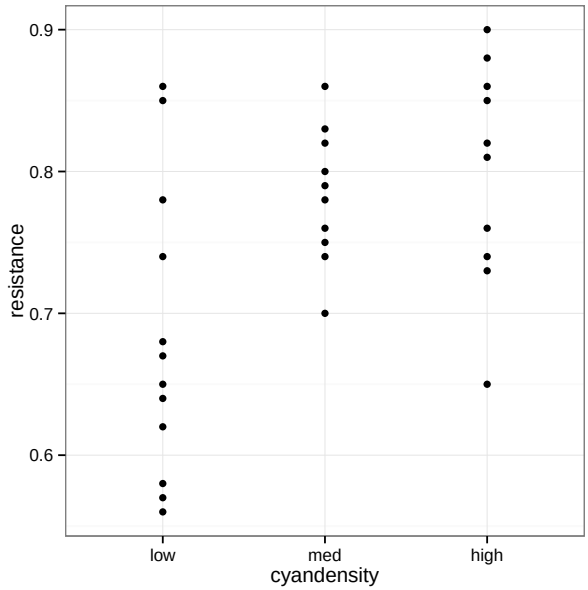
```
kruskal.test(PLP1.expression ~ group, data=brainGene)

#
#  Kruskal-Wallis rank sum test
#
# data:  PLP1.expression by group
# Kruskal-Wallis chi-squared = 13.1985, df = 2, p-value
# = 0.001361
```


Exercise: Daphnia Resistance

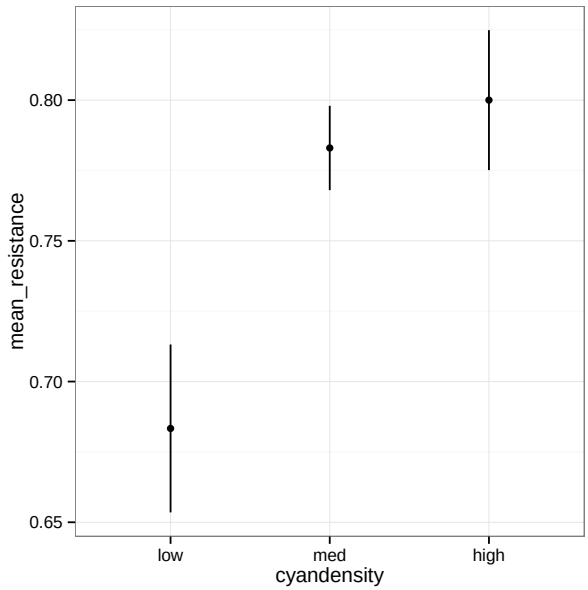
- Evaluate whether the data is appropriate for ANOVA
- Fit an ANOVA and check diagnostics
- Evaluate results & compare to Kruskal-Wallace and a glm with a Gamma distribution



Daphnia Data



Daphnia Means



How about HOV?

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

# Levene's Test for Homogeneity of Variance (center = median)
#       Df F value Pr(>F)
# group  2  2.0019 0.1533
#       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 0.089195 0.044598  6.6916 0.004078
# Residuals   29 0.193277 0.006665
```

KW shows an Effect

```
#
#  Kruskal-Wallis rank sum test
#
# data:  resistance by cyandensity
# Kruskal-Wallis chi-squared = 8.1996, 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    0.52908
# cyandensity  2  0.16216        29    0.36692
```

Diagnostics Also Good