STAT 215
One-Way ANOVA to Model Multiple Means

Colin Reimer Dawson

Oberlin College

November 17, 2017

Outline

Quantitative Vs. Categorical Predictor and Response

		Response	
		Quantitative	**Categorical**
Predictor	**Quantitative**	Linear Reg.	Logistic Reg.
	Categorical	ANOVA	

Anxiety and Cognitive Functioning

Is there a relationship between anxiety levels and cognitive functioning? A collection of variables pertaining to cognitive and emotional status, sleep habits, and academic habits were collected from 253 college students. One of these, `AnxietyStatus` classifies students according to whether they have `Normal`, `Moderate`, or `Severe` anxiety. Another, `CognitionZscore`, measures (standardized) performance on a test of cognitive skills.

Quantitative Vs. Categorical Predictor and Response

		Response	
		Quantitative	**Categorical**
Predictor	**Quantitative**	Linear Reg.	Logistic Reg.
	Categorical	ANOVA	

ANOVA: Test vs. Model

- We have already seen "Analysis of Variance" (ANOVA) in the context of a test to compare models.
- Somewhat confusingly, ANOVA is used to refer both to this test, and to a means model for which the same test is used.
- Note: You have likely seen the test of this model in AP Stat, but likely not in expliictly model-based terms.

The One-Way (One-predictor) ANOVA (Means) Model

DATA $=$ PATTERN $+$ IDIOSYNCRACIES

The One-way ANOVA Population Model (X categorical)

$$Y = f(X) + \varepsilon$$
$$Y = \mu_k + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2_\varepsilon)$$

One μ_k for each level of X

Parameter Estimates

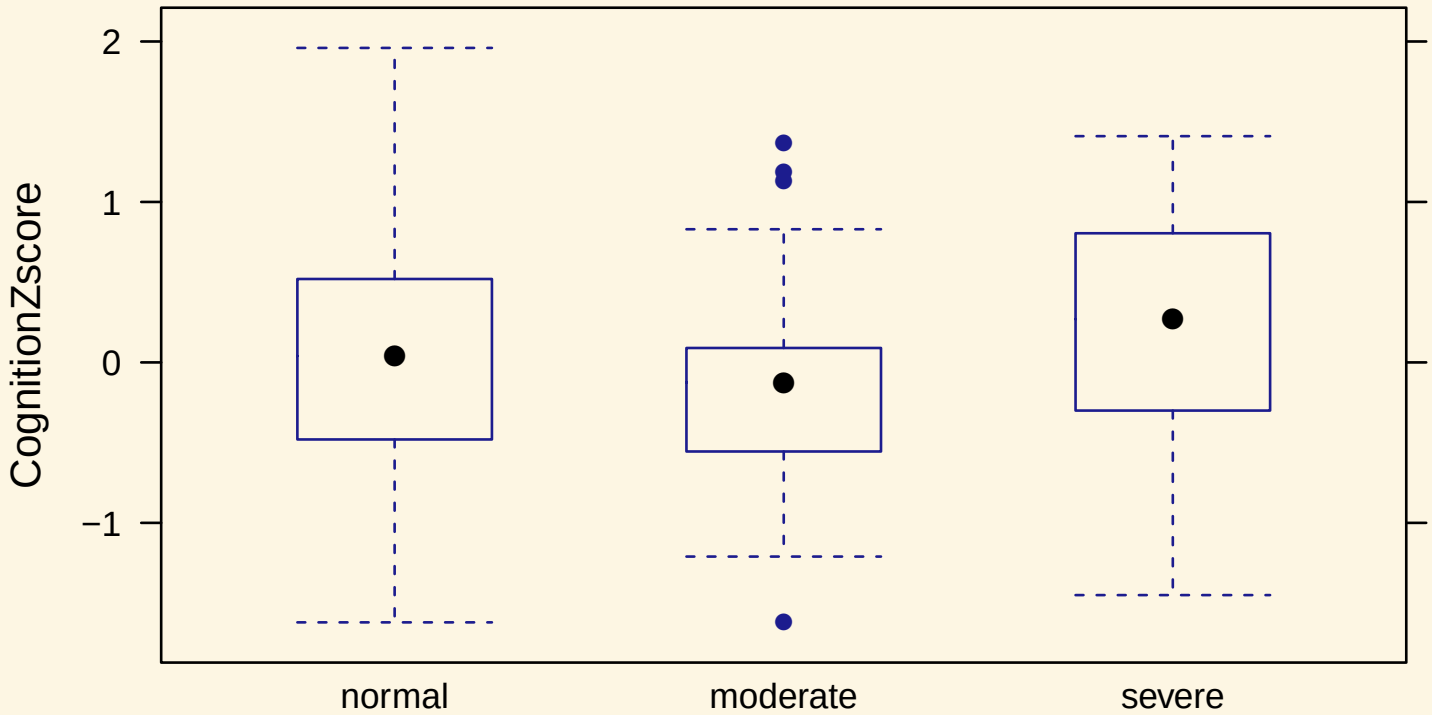
The One-way ANOVA Fitted Model (X categorical)

$$Y_i = \hat{f}(X_i) + \hat{\varepsilon}_i \qquad \hat{\varepsilon} \sim \mathcal{N}(0, \hat{\sigma}^2_\varepsilon)$$
$$Y_i = \bar{Y}_k + (Y_i - \bar{Y}_k)$$

One $\bar{Y}_k$ for each level of X

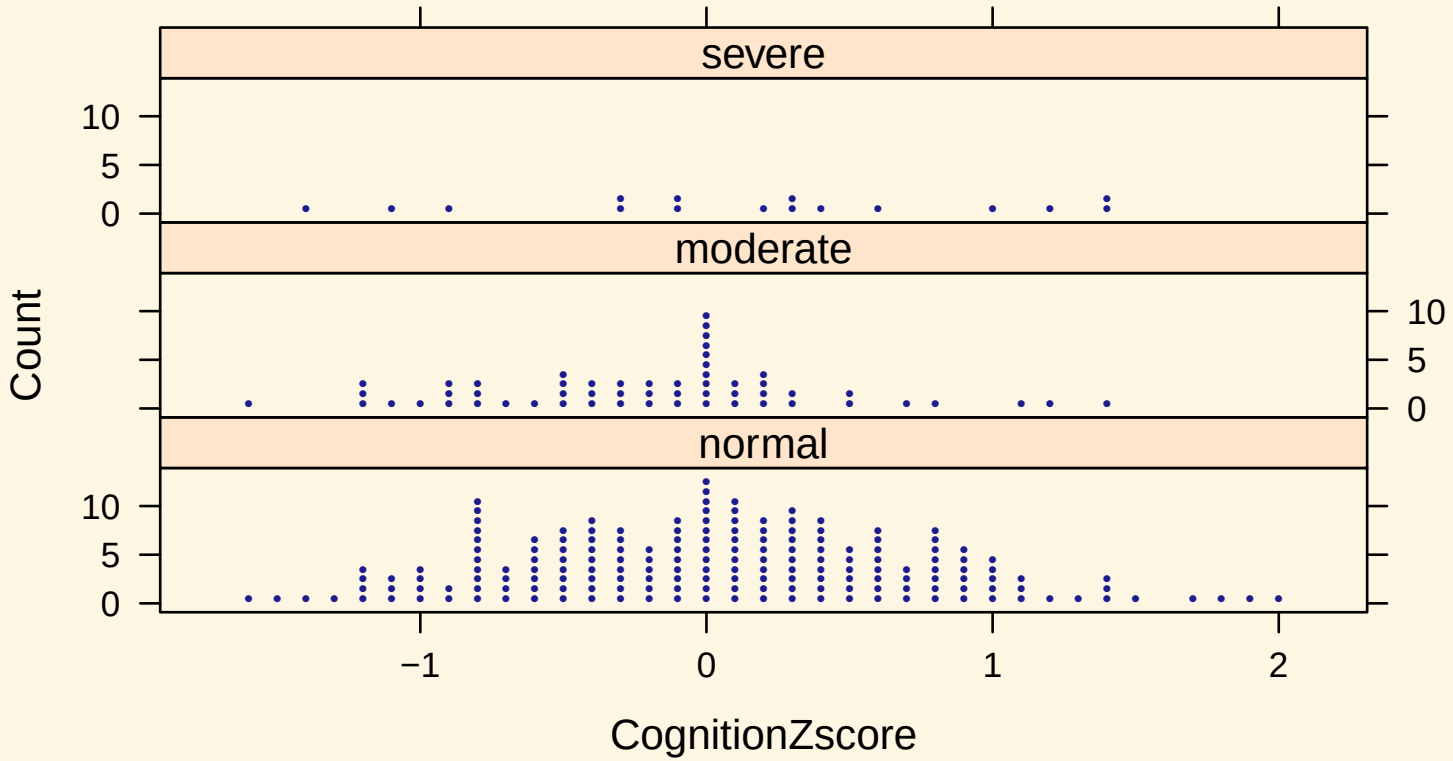
Anxiety and Cognitive Functioning

```
library("Lock5Data"); library("mosaic"); data("SleepStudy")
SleepStudy2 <-   # This line is just to reorder the categories for the plot
    mutate(SleepStudy,
           Anxiety = factor(AnxietyStatus,
                            levels = c("normal", "moderate", "severe")))
bwplot(CognitionZscore ~ Anxiety, data = SleepStudy2)
```



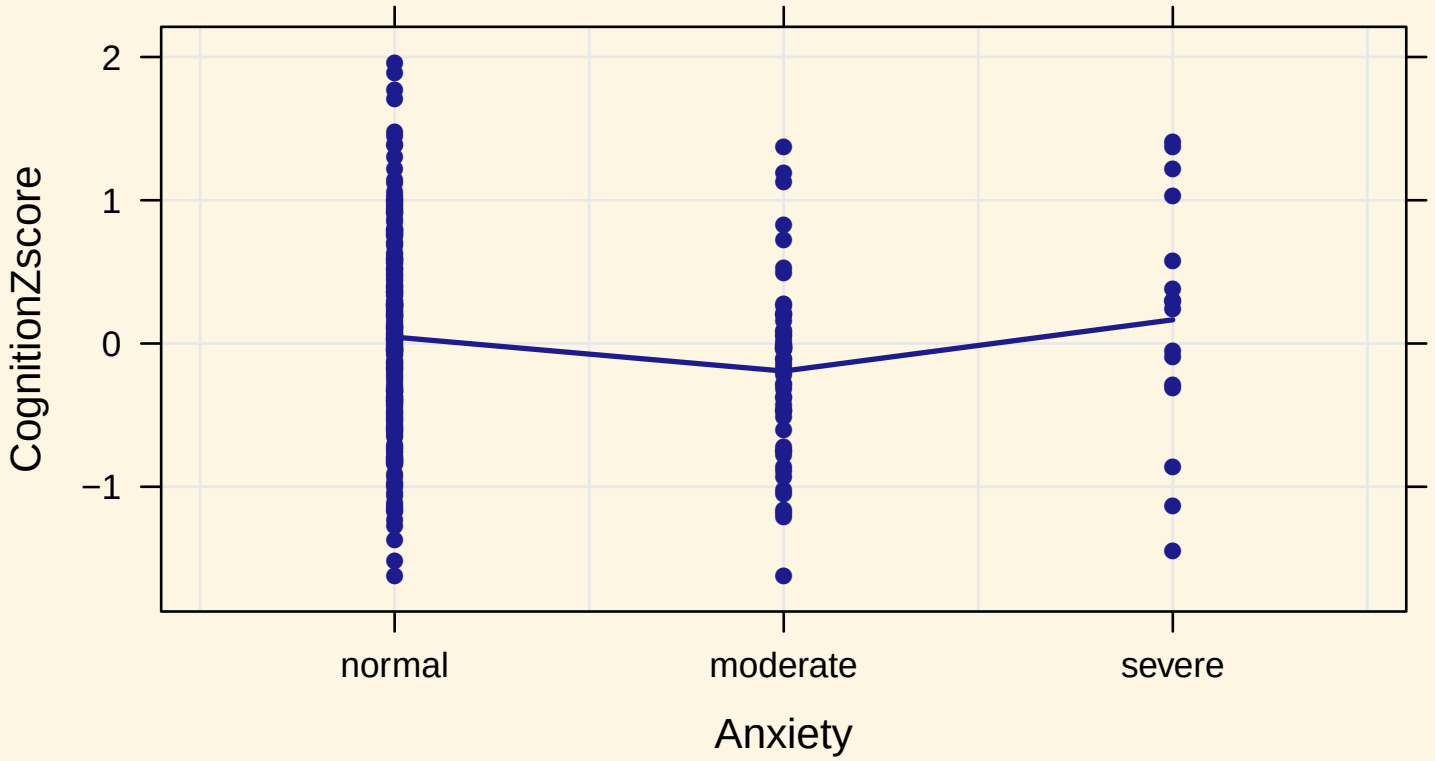
Anxiety and Cognitive Functioning

```
dotPlot(~CognitionZscore | Anxiety, data = SleepStudy2,
        width = 0.1, layout = c(1,3)) #controls bin width and arrangement
```



Anxiety and Cognitive Functioning

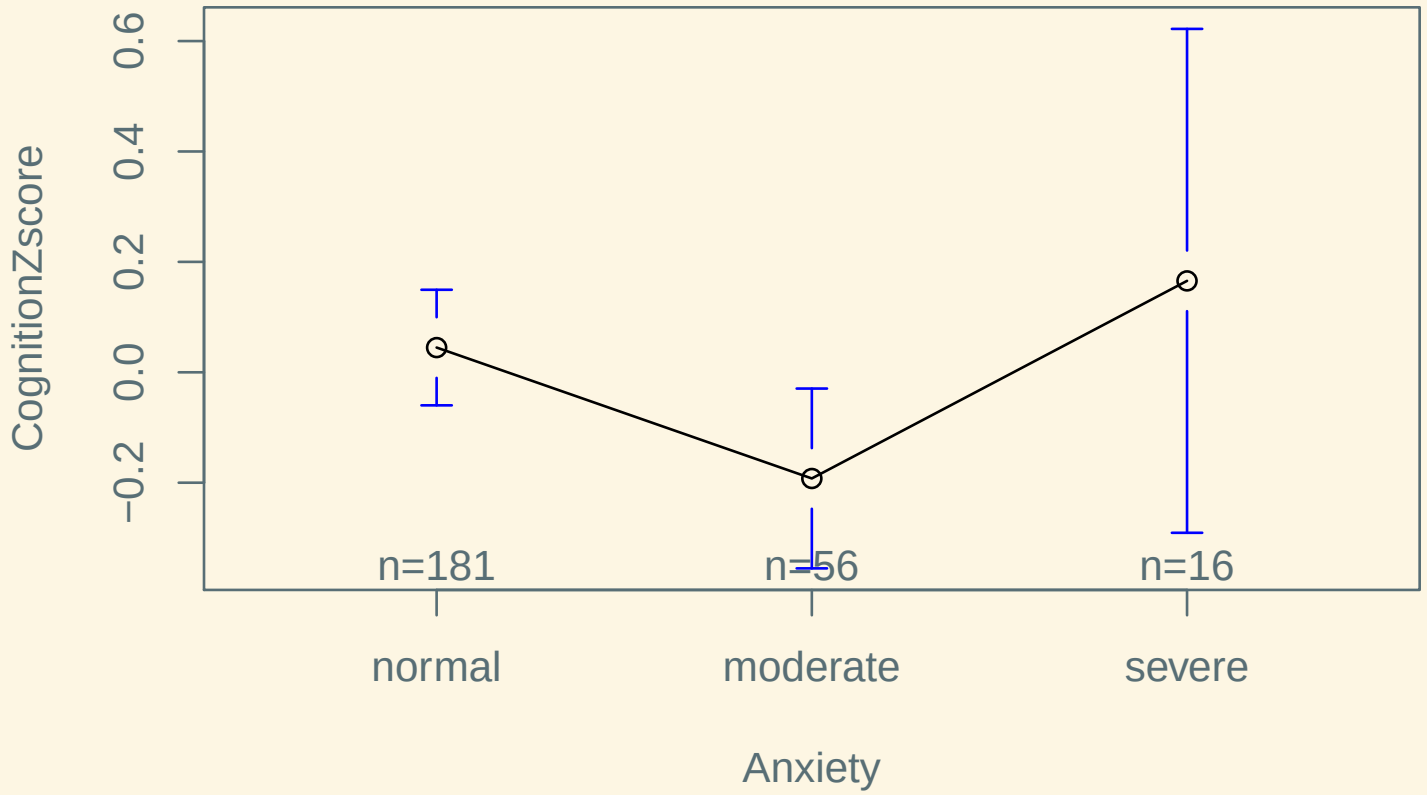
```
xyplot(CognitionZscore ~ Anxiety, data = SleepStudy2,
       type = c("p","a"), grid = TRUE)
```




```
## type = c("p", "a") plots the points and a line through the averages
## grid = TRUE draws a grid in the background
```

Anxiety and Cognitive Functioning

```
library("gplots")  # May need to install it first
plotmeans(CognitionZscore ~ Anxiety, data = SleepStudy2)
```




```
## Plots means and individual confidence intervals for the means
```

Anxiety and Cognitive Functioning

```
favstats(CognitionZscore ~ Anxiety, data = SleepStudy2)
```

	Anxiety	min	Q1	median	Q3	max	mean	sd	n	missing
1	normal	-1.62	-0.4800	0.040	0.5200	1.96	0.04480663	0.7136912	181	0
2	moderate	-1.62	-0.5325	-0.125	0.0900	1.37	-0.19232143	0.6081208	56	0
3	severe	-1.45	-0.2950	0.270	0.6925	1.41	0.16562500	0.8564887	16	0

The One-Way (One-predictor) ANOVA (Means) Model

The One-way ANOVA Population Model (X categorical)

$$Y = f(X) + \varepsilon$$
$$Y = \mu_k + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2_\varepsilon)$$

One μ_k for each level of X

Question: Is there evidence that the μ_k are different?

The One-Way ANOVA Model: Alternative Formulation

The One-way ANOVA Population Model (X categorical)

$$Y = f(X) + \varepsilon$$
$$Y = \mu + \alpha_k + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

One α_k for each level of X: group deviation from overall mean

Question: Is there evidence that the α_k are not all zero?

Parameter Estimates

The One-way ANOVA Fitted: (Alternative Formulation)

$$Y_i = \hat{f}(X_i) + \hat{\varepsilon}$$
$$Y_i = \bar{Y} + (\bar{Y}_k - \bar{Y}) + (Y_i - \bar{Y}_k)$$

One $\bar{Y}_k - \bar{Y}$ for each level of X: Group deviation from overall mean

Partitioning Variability

The One-way ANOVA Fitted: (Alternative Formulation)

$$Y_i = \bar{Y} + (\bar{Y}_k - \bar{Y}) + (Y_i - \bar{Y}_k)$$

One $\bar{Y}_k - \bar{Y}$ for each level of X: Group deviation from overall mean

Partitioning Variability

$$(Y_i - \bar{Y}) = (\bar{Y}_k - \bar{Y}) + (Y_i - \bar{Y}_k)$$

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\bar{Y}_k - \bar{Y})^2 + 0 + \sum_i (Y_i - \bar{Y}_k)^2$$

$$SS_{Total} = SS_{Model} + SS_{Error}$$

Anxiety and Cognitive Functioning

```
favstats(~CognitionZscore, data = SleepStudy2) %>%
    format(digits = 2, nsmall = 2)
```

```
     min    Q1 median   Q3  max   mean   sd   n missing
   -1.62 -0.48  -0.01 0.44 1.96 -4e-05 0.71 253       0
```

Overall mean, $\bar{Y} \approx 0$
Individual means $\bar{Y}_k = 0.045, -0.192, 0.166$

```
favstats(CognitionZscore ~ Anxiety, data = SleepStudy2) %>%
    format(digits = 2, nsmall = 2)
```

```
   Anxiety   min    Q1 median   Q3  max   mean   sd   n missing
1   normal -1.62 -0.48   0.04 0.52 1.96  0.045 0.71 181       0
2 moderate -1.62 -0.53  -0.12 0.09 1.37 -0.192 0.61  56       0
3   severe -1.45 -0.29   0.27 0.69 1.41  0.166 0.86  16       0
```

$$Y_i = \begin{cases} 0 + (0.045 - 0) + \hat{\varepsilon} & X_i = \text{normal} \\ 0 + (-0.192 - 0) + \hat{\varepsilon} & X_i = \text{moderate} \\ 0 + (0.166 - 0) + \hat{\varepsilon} & X_i = \text{severe} \end{cases}$$

Exercise: Find the SS components

$$Y_i = \begin{cases} 0 + (0.045 - 0) + \hat{\varepsilon} & X_i = \text{normal} \\ 0 + (-0.192 - 0) + \hat{\varepsilon} & X_i = \text{moderate} \\ 0 + (0.166 - 0) + \hat{\varepsilon} & X_i = \text{severe} \end{cases}$$

Anxiety	**CognitionZscore**	$(\bar{Y}_k - \bar{Y})^2$	$(Y_i - \bar{Y}_k)^2$
Normal	-0.26		
Normal	1.39		
Severe	0.38		
Severe	1.22		
Moderate	-0.04		
Moderate	0.72		
		$SS_{Model} =$	$SS_{Error} =$

How Much Variability is Explained?

- Still have R^2, even though there's no longer a correlation:

$$R^2 = \frac{SS_{Model}}{SS_{Total}}$$

Model Comparison (ASSESS)

Null Population Model:

$$Y_i = \mu + \varepsilon$$

Groups Population Model:

$$Y_i = \mu + \alpha_k + \varepsilon$$

H_0 : $\alpha_k \equiv 0$ for all $k \Rightarrow$ high SS_{Model} due to chance
H_1 : some $\alpha_k \neq 0 \Rightarrow$ high SS_{Model} due to grouping

Conditions for Test of ANOVA Model

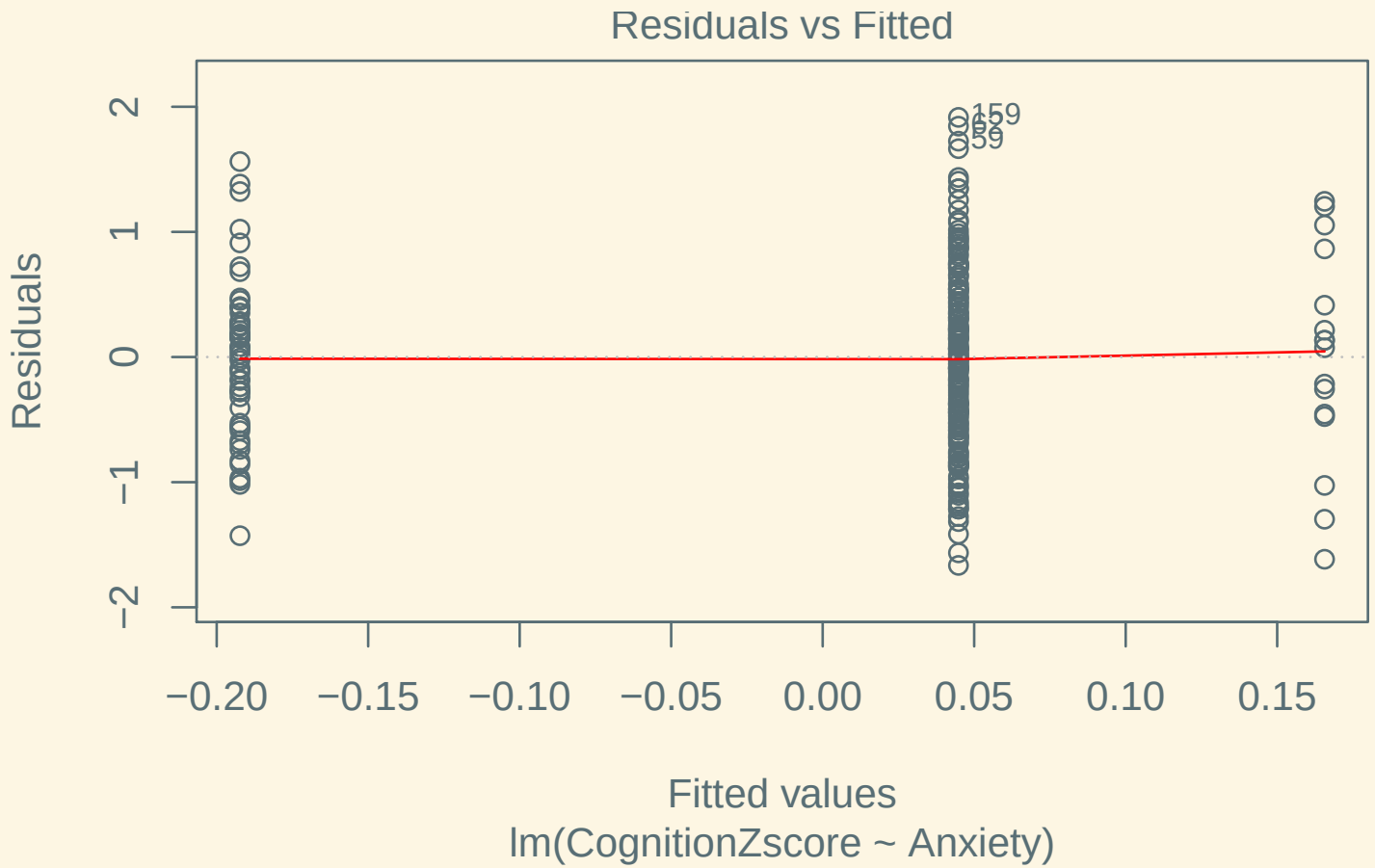
Can view these as regression models and compute P-value assuming the same conditions as for SLR (except linearity):

Conditions for Test of ANOVA Model

1. Zero mean: Residuals centered at 0
2. Constant variance: Same variability at all X (Homoskedasticity)
3. Independence: No relationship among errors
4. Normality (for standard form): At each X, Ys are Normally distributed

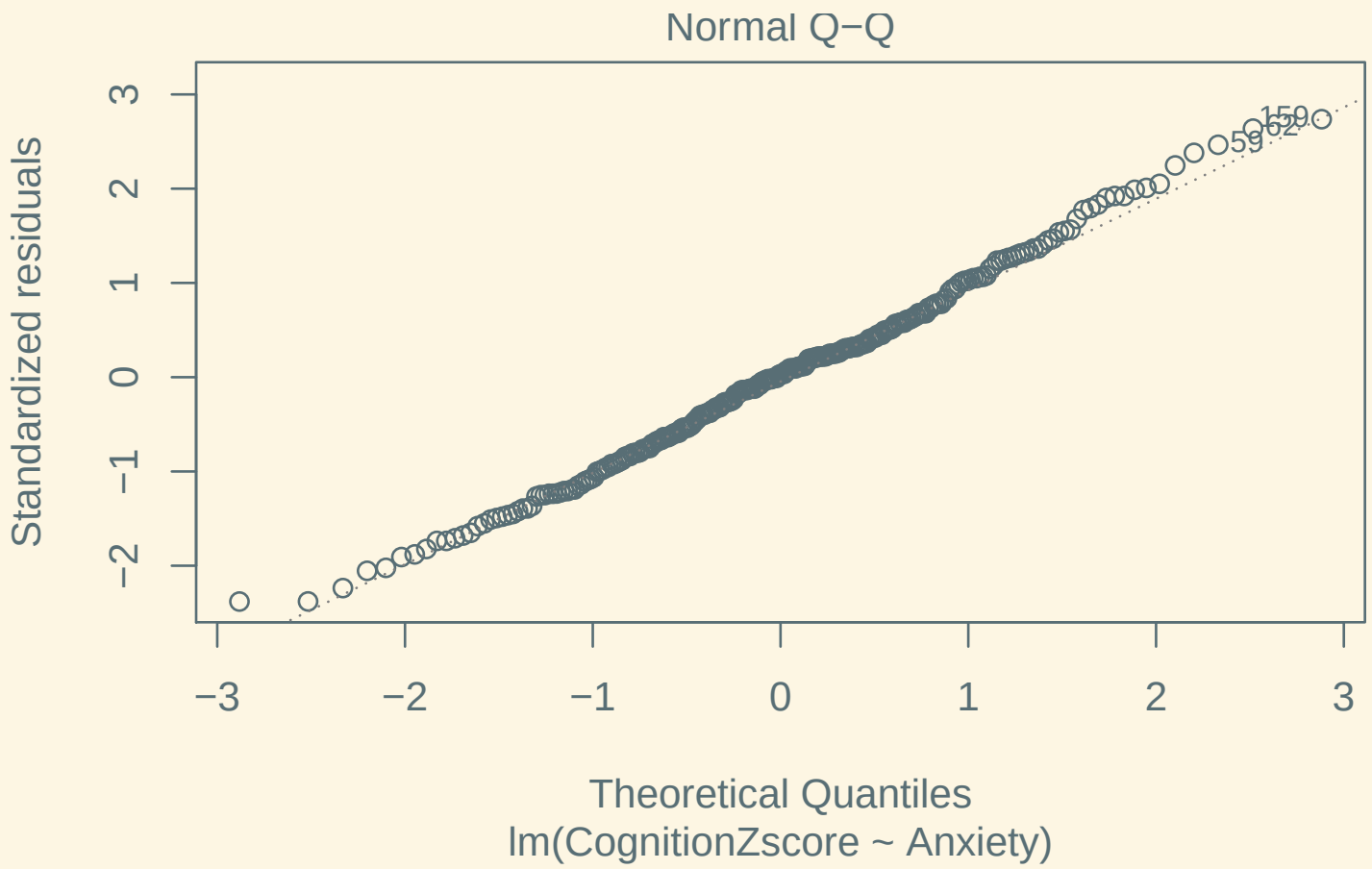
Checking Conditions: Residual Plots

```
anxiety.model <- lm(CognitionZscore ~ Anxiety, data = SleepStudy2)
plot(anxiety.model, which = 1)
```



Checking Conditions: Residual Plots

```
plot(anxiety.model, which = 2)
```



Checking Conditions: Homoskedasticity

Homoskedasticity: $=$ Equal variances

```
sd(CognitionZscore ~ Anxiety, data = SleepStudy2)
```

```
       normal  moderate    severe
    0.7136912 0.6081208 0.8564887
```

Rough rule: largest $s \leq 2 \cdot$smallest s

Conditions: A Caveat

If sample sizes are (nearly) equal within groups, the ANOVA test is fairly *robust* to violations of normality and homoskedasticity.

The ANOVA Table

```
anxiety.model <- lm(CognitionZscore ~ Anxiety, data = SleepStudy2)
anova(anxiety.model)
```

```
Analysis of Variance Table

Response: CognitionZscore
           Df  Sum Sq Mean Sq F value  Pr(>F)
Anxiety     2   2.874 1.43679  2.9197 0.05579 .
Residuals 250 123.027 0.49211
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

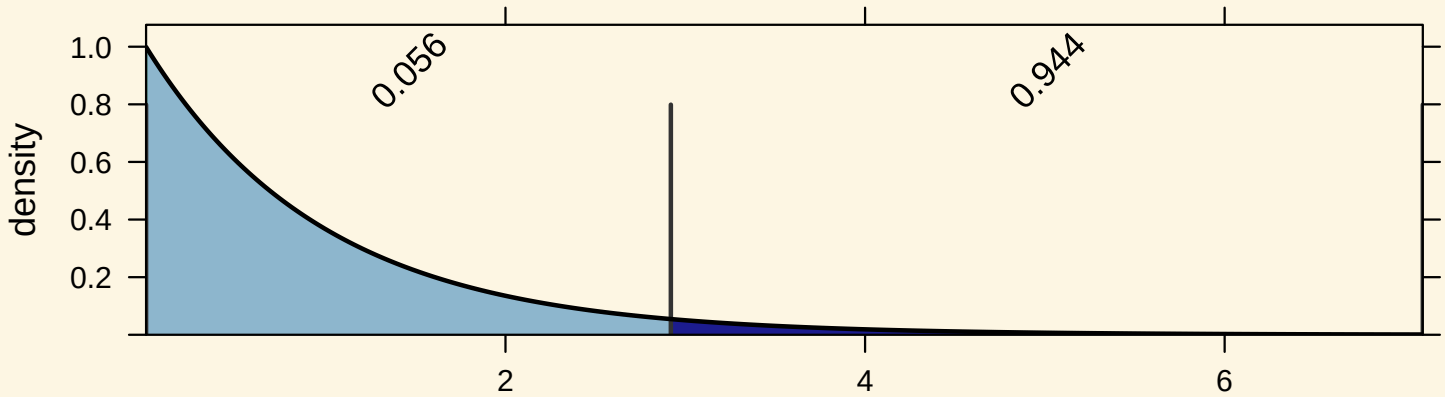
$$F = \frac{MS_{Model}}{MS_{Error}} = \frac{SS_{Model}/df_{Model}}{SS_{Error}/df_{Error}}$$

has an F distribution if H_0 is true and conditions are met.

P-value computation

For the record, we can compute the P-value from the F-statistic using an F-distribution:

```
xpf(2.9197, df1 = 2, df2 = 250, lower.tail = FALSE)
```




```
[1] 0.05579209
```

An Alternative: Randomization Test

- Recall: We can simulate H_0 for an SLR model by randomly associating X and Y values and recomputing the statistic of interest.
- Same logic applies here: randomly pair X and Y (i.e., scramble groups, but preserve group sizes), and compute F each time.
- The resulting **randomization distribution** can be used instead of the theoretical F distribution.
- `http://lock5stat.com/statkey`

Example: Stereotype Threat and Student Athletes

The term "stereotype threat" refers to a phenomenon whereby reminders of particular components of an individual's identity (race, gender, ethnicity) can result in the individual conforming to stereotypes about that group. For example, women perform worse on a math test after being reminded of their gender (Spencer et al., 1999). Some researchers (Steele, 1997) believe this is due to anxiety about the possibility of confirming negative stereotypes. Yopyk and Prentice (2005) administered a math test to student-athletes after either (A) reminding them of their athlete status, (B) reminding them of their student status, or (C) not reminding them of either component of their identity. The test scores had the following mean and standard deviations.

Summary Stats and ANOVA Table

	Athlete Prime	No Prime	Student Prime
n	12	13	12
$\bar{x}$	66.97	82.46	86.17
s	5.60	4.99	4.58

Source	df	SS	MS	F	P-value
Prime	?	2504.38	1252.19	48.68	1.05e-10
Residuals	?	874.5	25.72		

Conclusion: Some population mean is different from some other mean. But which one(s)?

Individual and Pairwise Inference

Items of Interest...

1. CIs for individual μ_ks
2. CIs for pairwise differences, $\mu_A - \mu_B$
3. t-tests for pairwise differences, $H_0 : \mu_A = \mu_B$, $H_1 : \mu_A \neq \mu_B$

Intervals and Tests to Compare Two Means

- Normally:

$$\text{CI for } \mu : \bar{Y} \pm t^* \cdot SE \text{ where } SE = \sqrt{\frac{\hat{\sigma}^2}{n}}$$

$$\text{CI for } \mu_1 - \mu_2 : \bar{Y} \pm t^* \cdot SE \text{ where } SE = \sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}$$

$$t_{obs} \text{ to test } H_0 : \mu_1 - \mu_2 = 0 \text{ is } t_{obs} = \frac{\bar{Y} - 0}{SE}$$

- For the ANOVA model, we assume, among other things, that there is one σ_ε^2 common to all groups, estimated by $\hat{\sigma}_\varepsilon^2 = MS_{Error}$.

So...

$$\text{CI for } \mu_k : \bar{Y} \pm t^* \cdot SE \text{ where } SE = \sqrt{\frac{MS_{Error}}{n_k}}$$

$$\text{CI for } \mu_A - \mu_B : \bar{Y} \pm t^* \cdot SE \text{ where } SE = \sqrt{\frac{MS_{Error}}{n_A} + \frac{MS_{Error}}{n_B}}$$

$$t_{obs} \text{ to test } H_0 : \mu_1 - \mu_2 = 0 \text{ is } t_{obs} = \frac{\bar{Y} - 0}{SE}$$

How many df for t^* and t_{obs}? Use df_{Error}, since this represents number of pieces of information about σ_ε^2

Example: Stereotype Threat and Student Athletes

	Athlete Prime	No Prime	Student Prime
n	12	13	12
$\bar{x}$	66.97	82.46	86.17
s	5.60	4.99	4.58

Source	df	SS	MS	F	P-value
Prime	2	2504.38	1252.19	48.68	1.05e-10
Residuals	34	874.5	25.72		

Let's compute a CI for $\mu_{Athlete} - \mu_{NoPrime}$.