STAT 215
Comparison of Individual Means

Colin Reimer Dawson

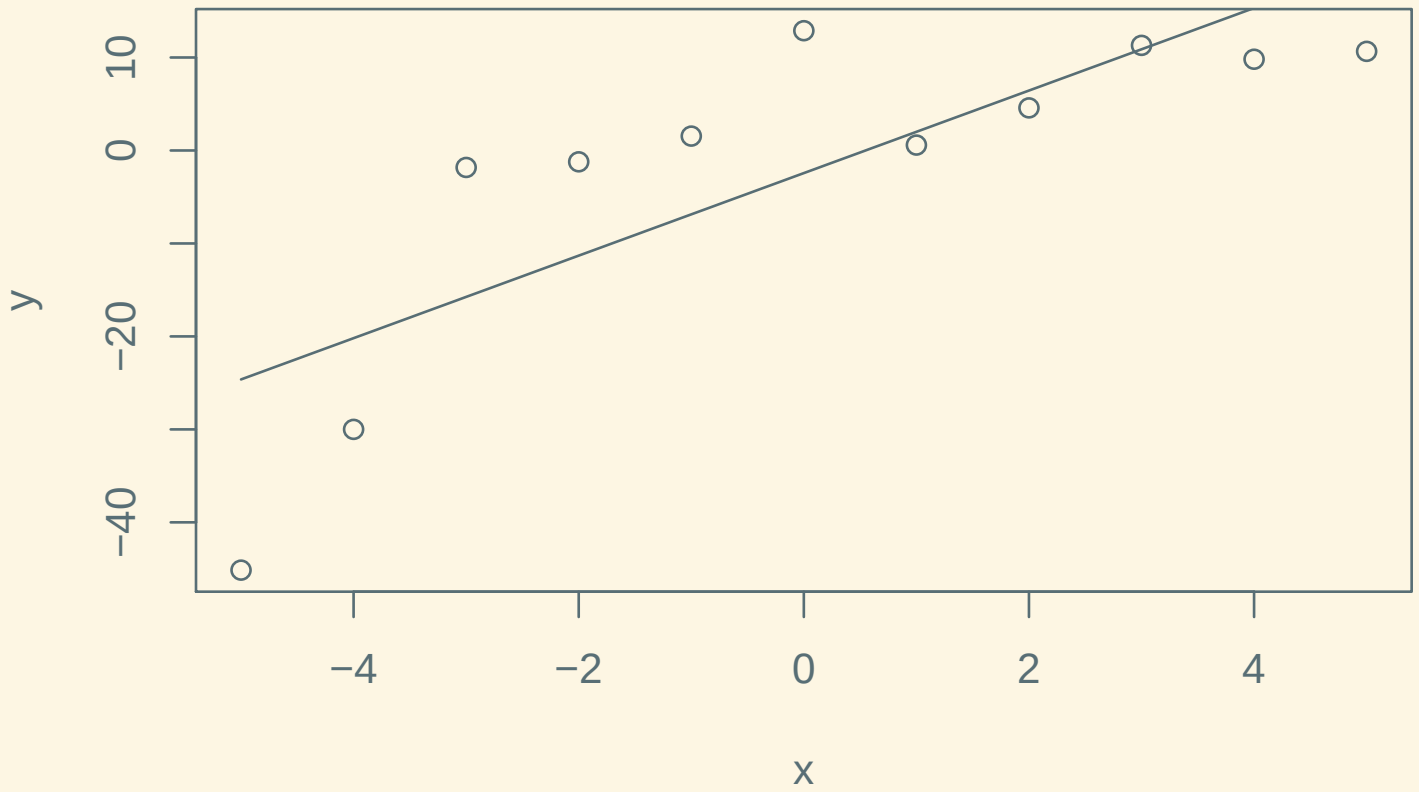
Oberlin College

November 21, 2017

Outline

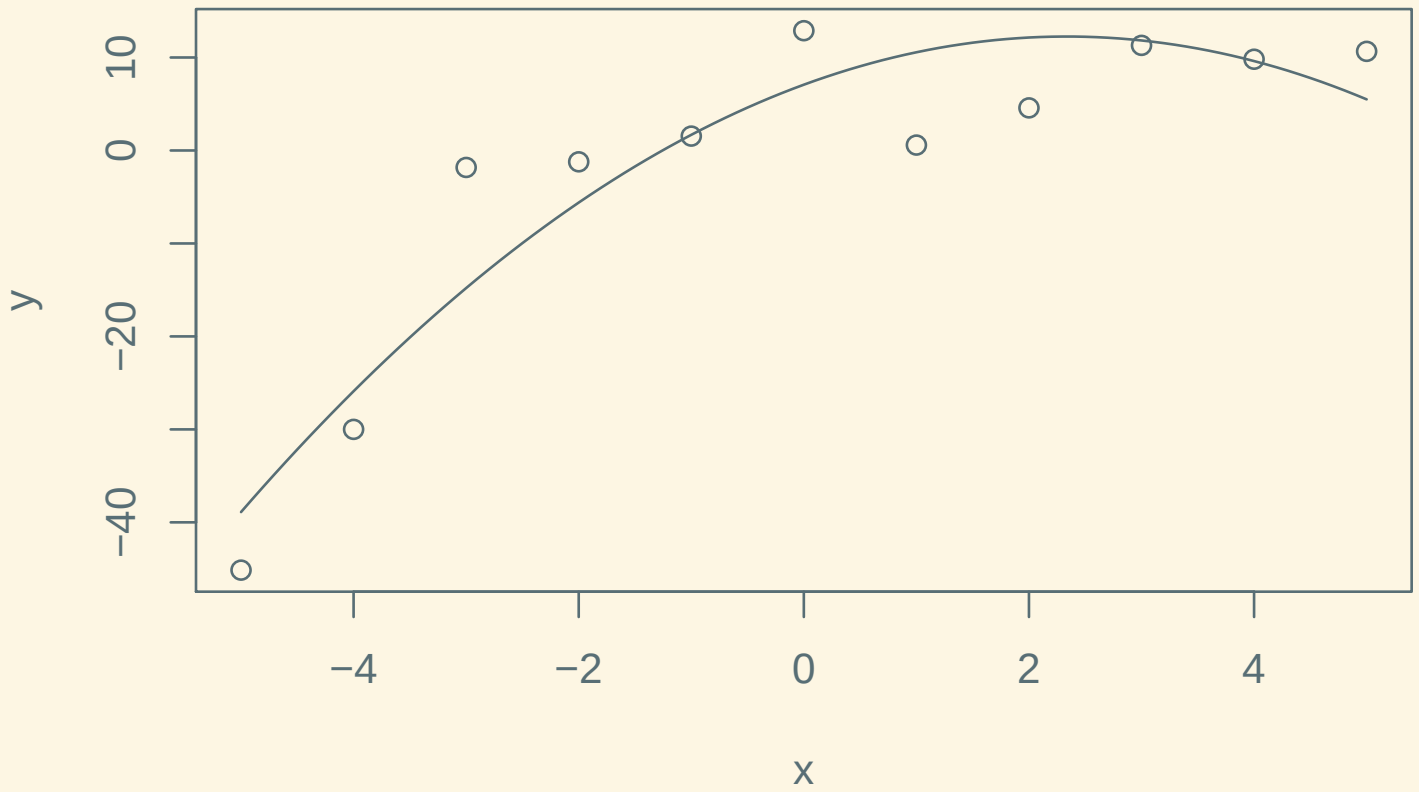
Data and Models

```
model1 <- lm(y ~ poly(x, degree = 1, raw = TRUE), data = TheData)
f.hat <- makeFun(model1)
plot(y ~ x, data = TheData)
curve(f.hat(x), add = TRUE)
```



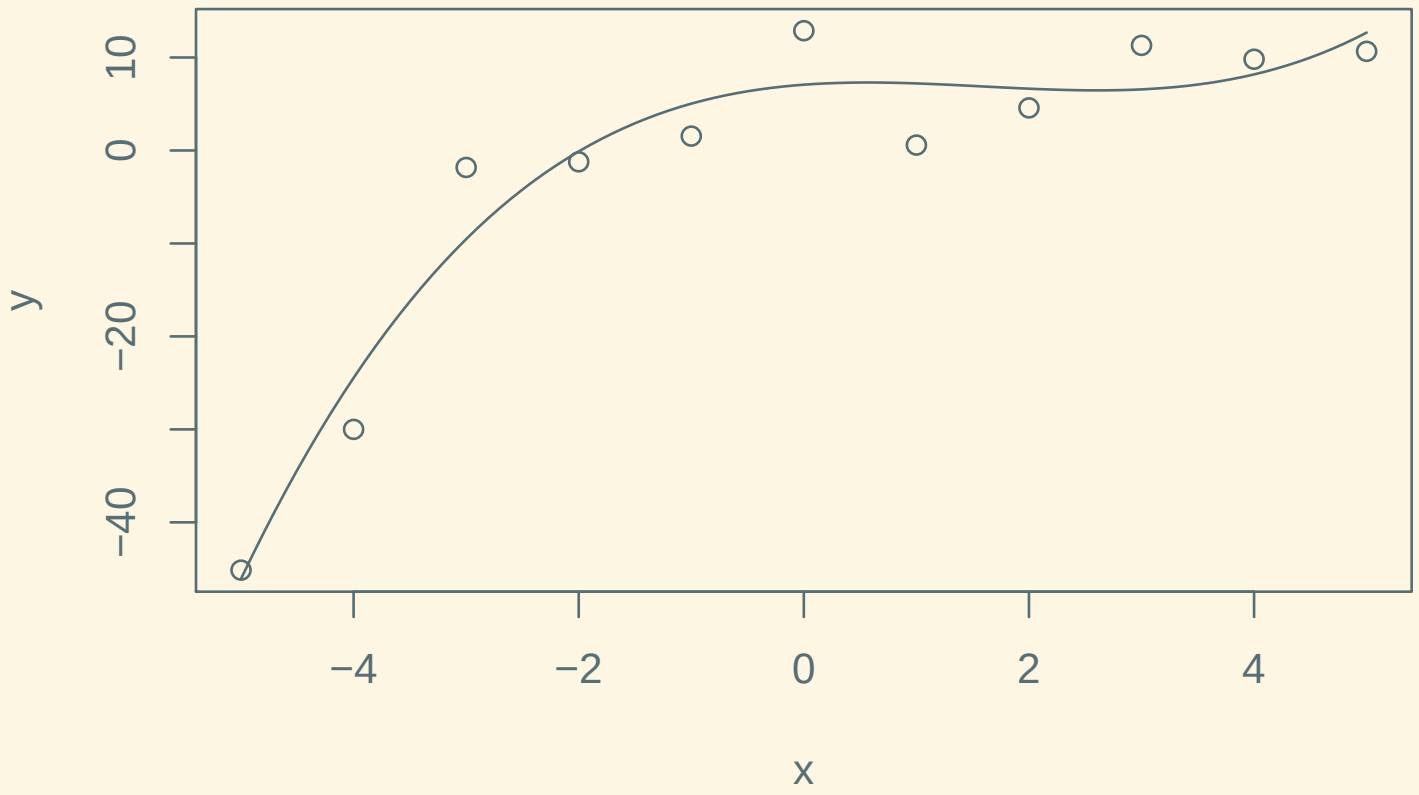
Data and Models

```
model2 <- lm(y ~ poly(x, degree = 2, raw = TRUE), data = TheData)
f.hat <- makeFun(model2)
plot(y ~ x, data = TheData)
curve(f.hat(x), add = TRUE)
```



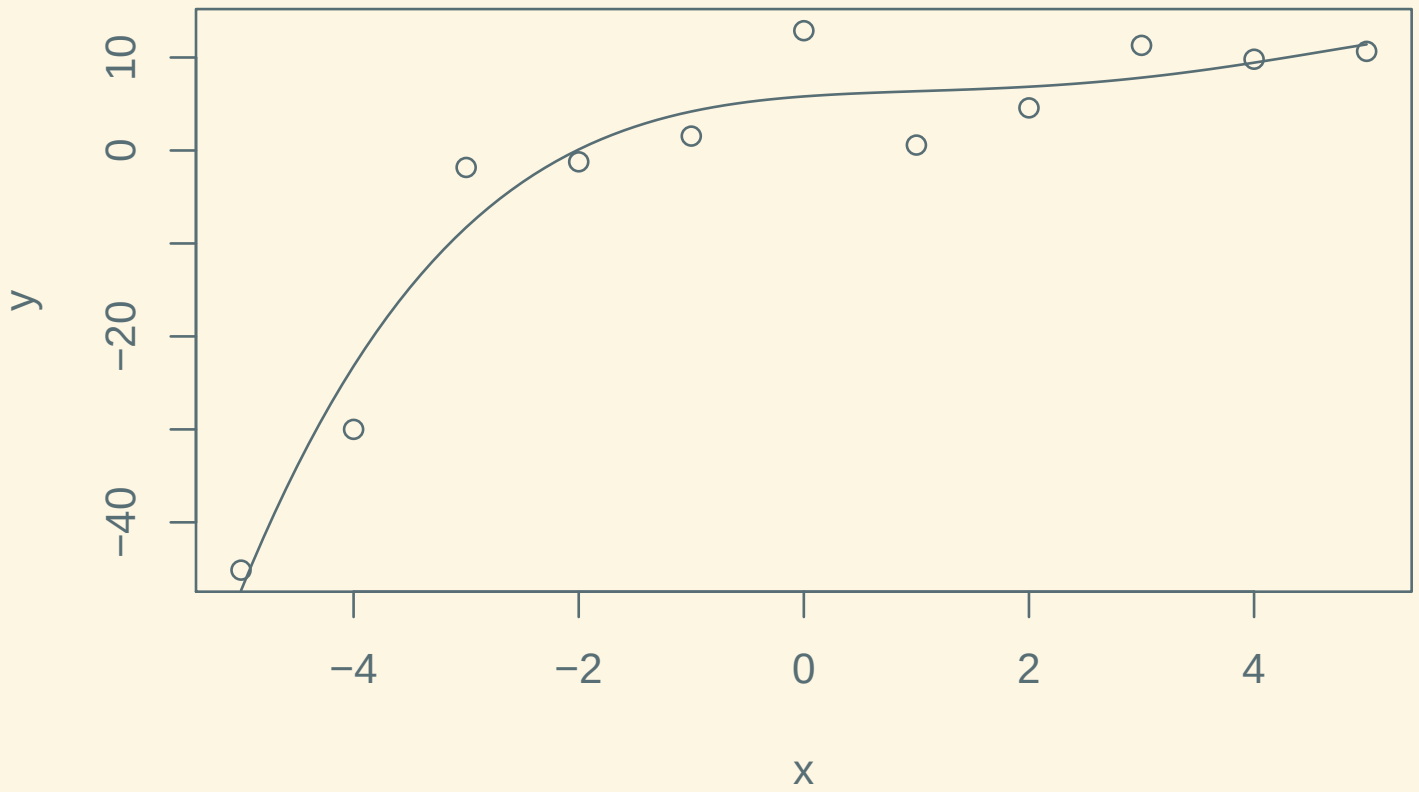
Data and Models

```
model3 <- lm(y ~ poly(x, degree = 3, raw = TRUE), data = TheData)
f.hat <- makeFun(model3)
plot(y ~ x, data = TheData)
curve(f.hat(x), add = TRUE)
```



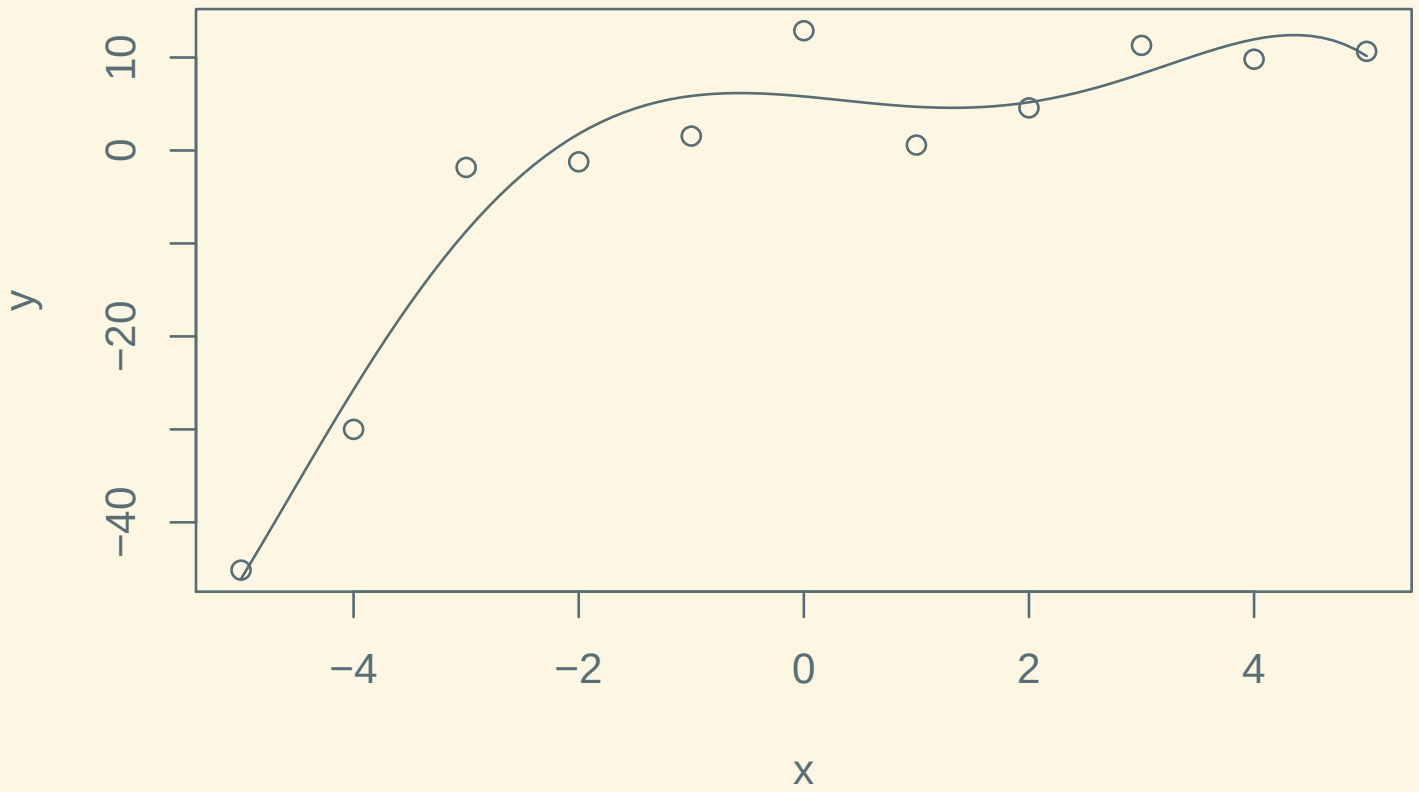
Data and Models

```
model4 <- lm(y ~ poly(x, degree = 4, raw = TRUE), data = TheData)
f.hat <- makeFun(model4)
plot(y ~ x, data = TheData)
curve(f.hat(x), add = TRUE)
```



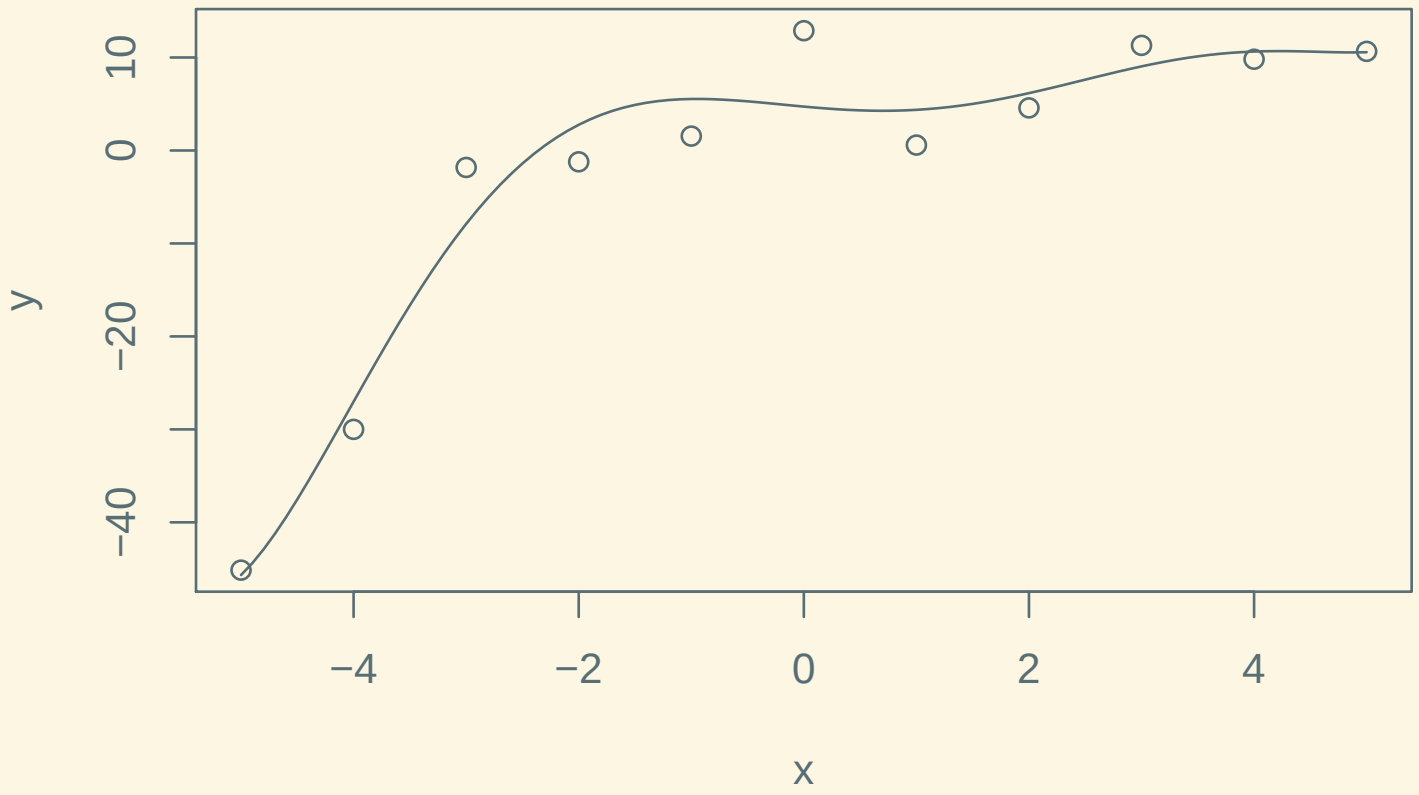
Data and Models

```
model5 <- lm(y ~ poly(x, degree = 5, raw = TRUE), data = TheData)
f.hat <- makeFun(model5)
plot(y ~ x, data = TheData)
curve(f.hat(x), add = TRUE)
```



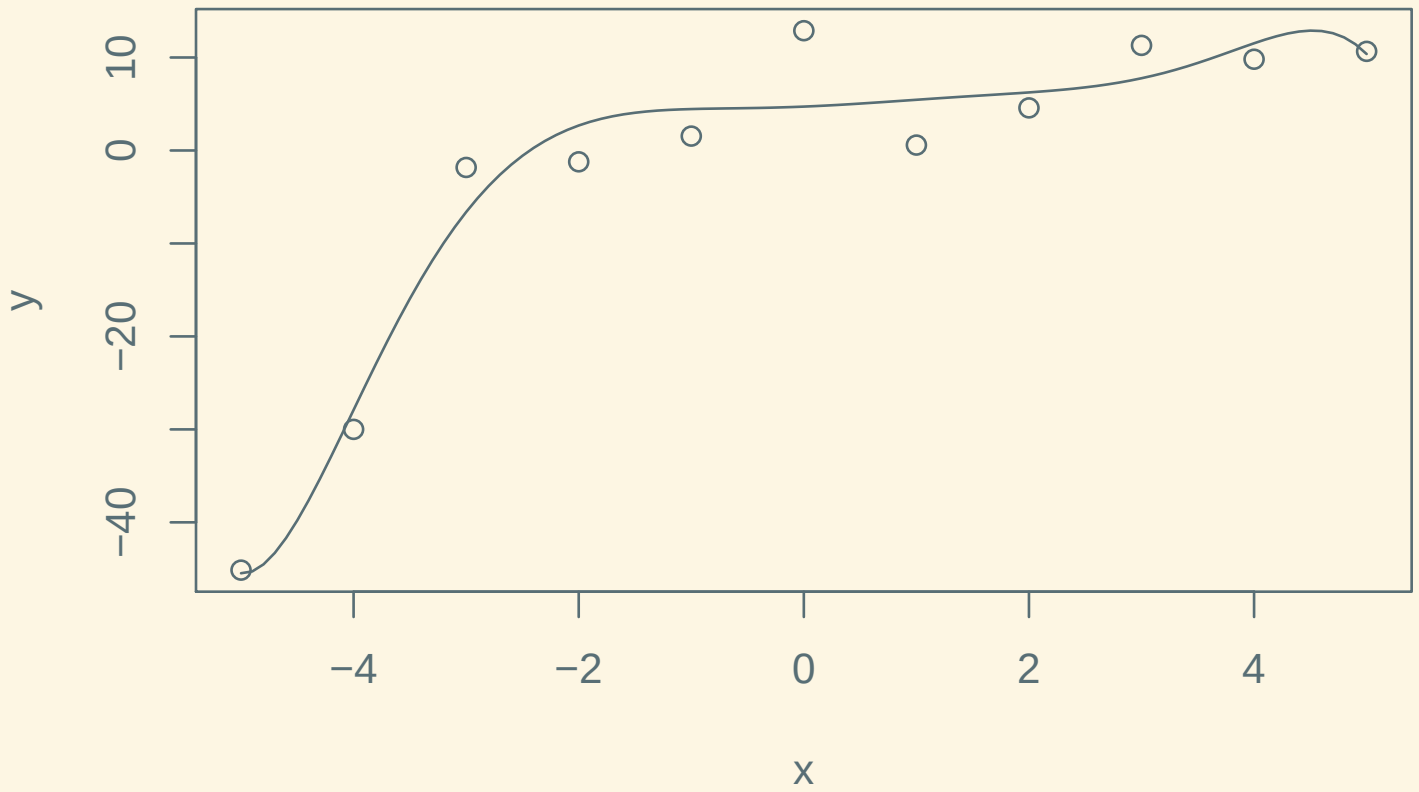
Data and Models

```
model6 <- lm(y ~ poly(x, degree = 6, raw = TRUE), data = TheData)
f.hat <- makeFun(model6)
plot(y ~ x, data = TheData)
curve(f.hat(x), add = TRUE)
```



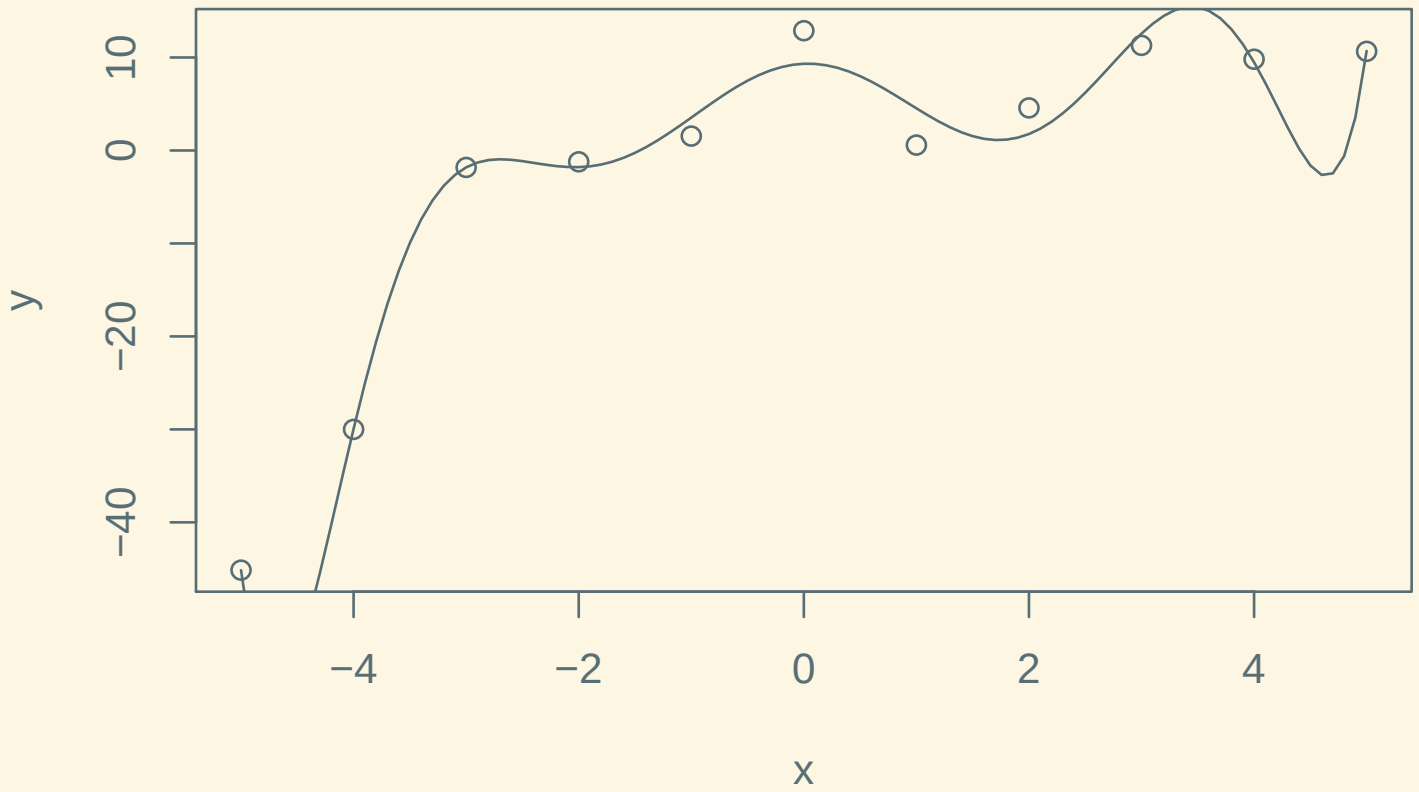
Data and Models

```
model7 <- lm(y ~ poly(x, degree = 7, raw = TRUE), data = TheData)
f.hat <- makeFun(model7)
plot(y ~ x, data = TheData)
curve(f.hat(x), add = TRUE)
```



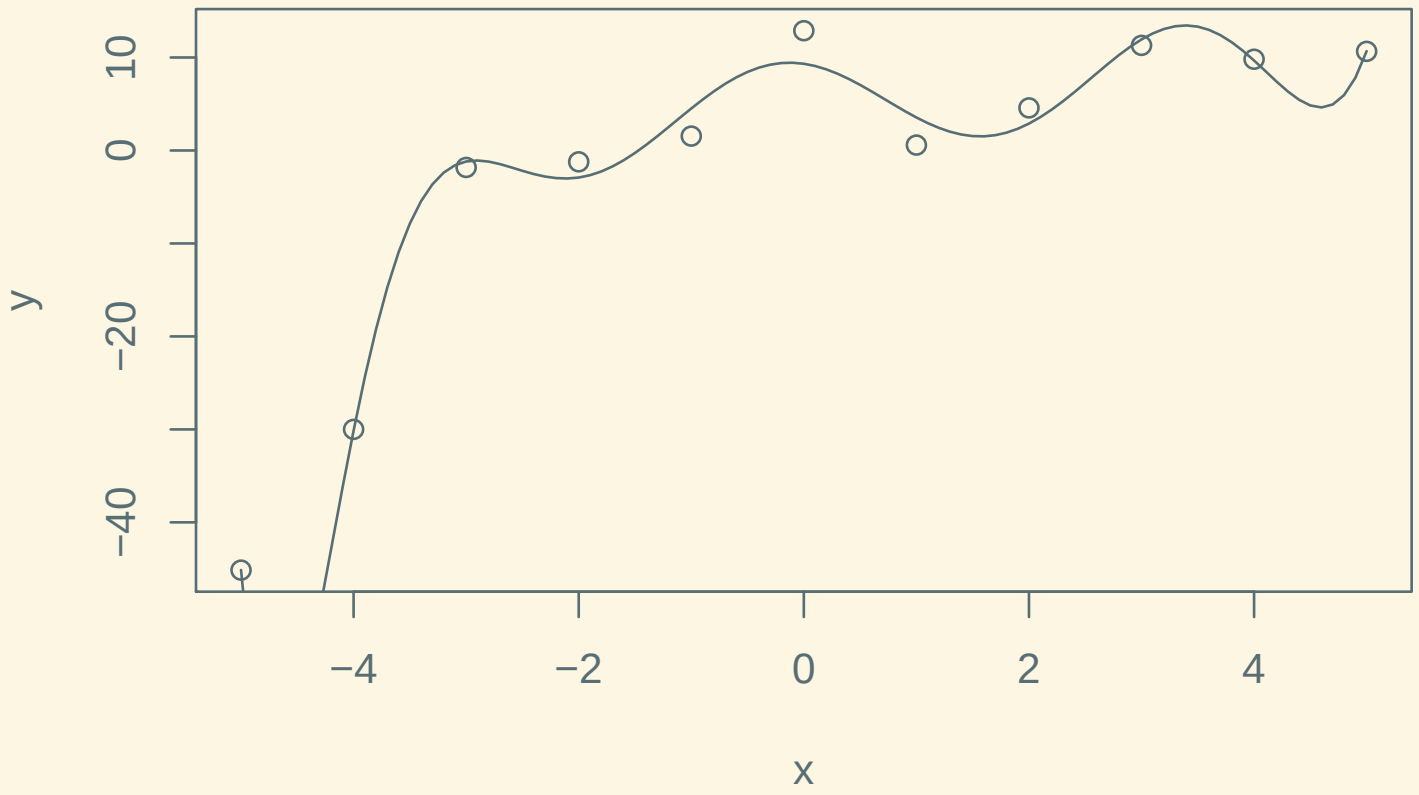
Data and Models

```
model8 <- lm(y ~ poly(x, degree = 8, raw = TRUE), data = TheData)
f.hat <- makeFun(model8)
plot(y ~ x, data = TheData)
curve(f.hat(x), add = TRUE)
```



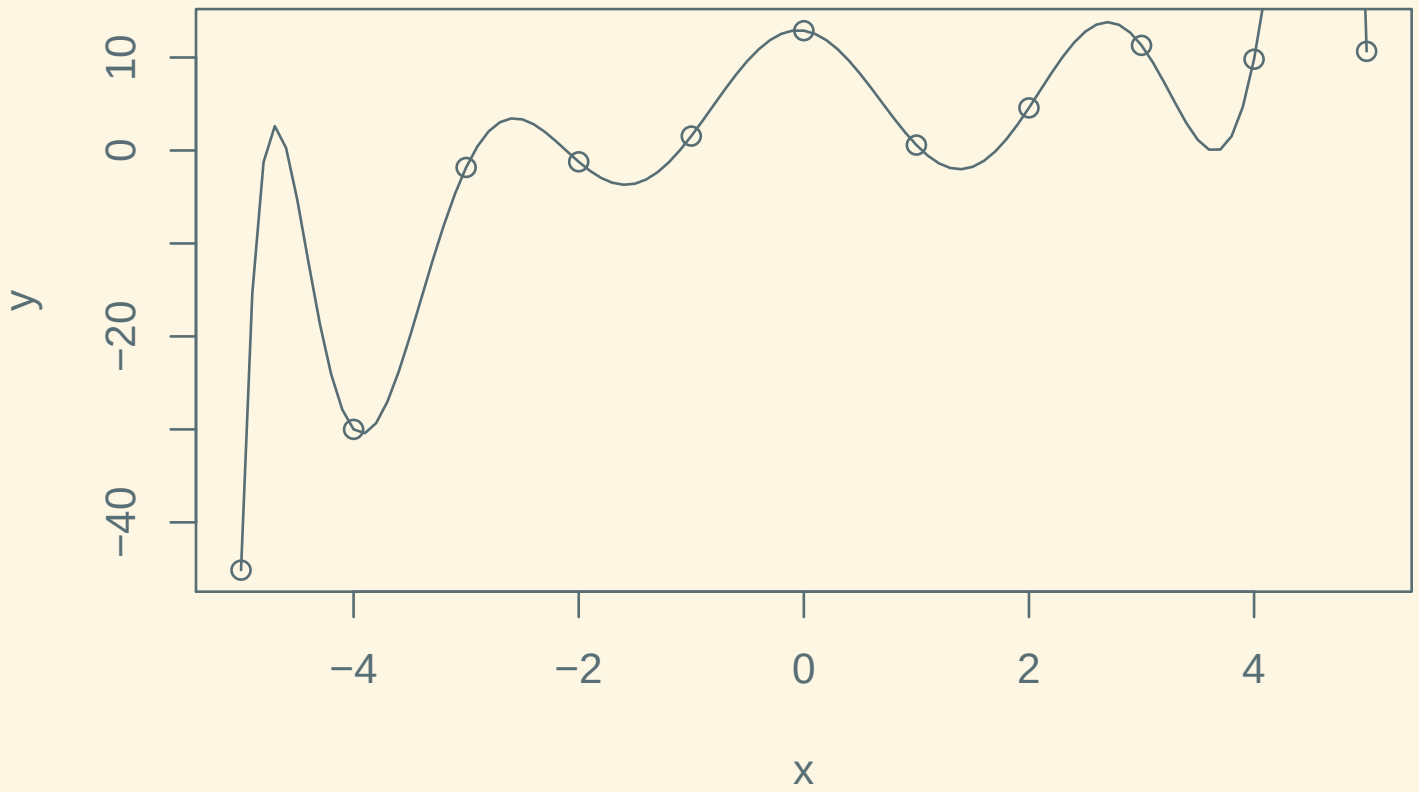
Data and Models

```
model9 <- lm(y ~ poly(x, degree = 9, raw = TRUE), data = TheData)
f.hat <- makeFun(model9)
plot(y ~ x, data = TheData)
curve(f.hat(x), add = TRUE)
```

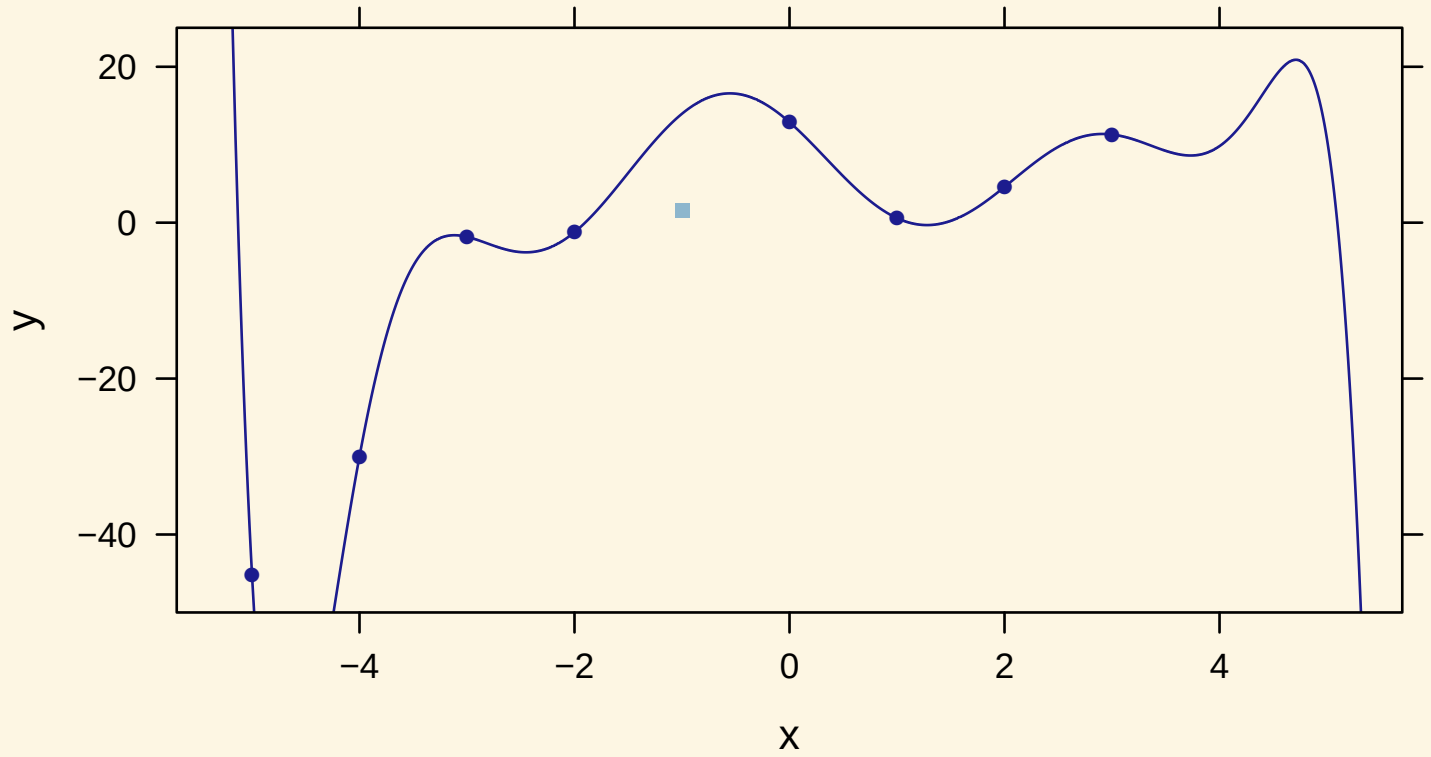


Data and Models

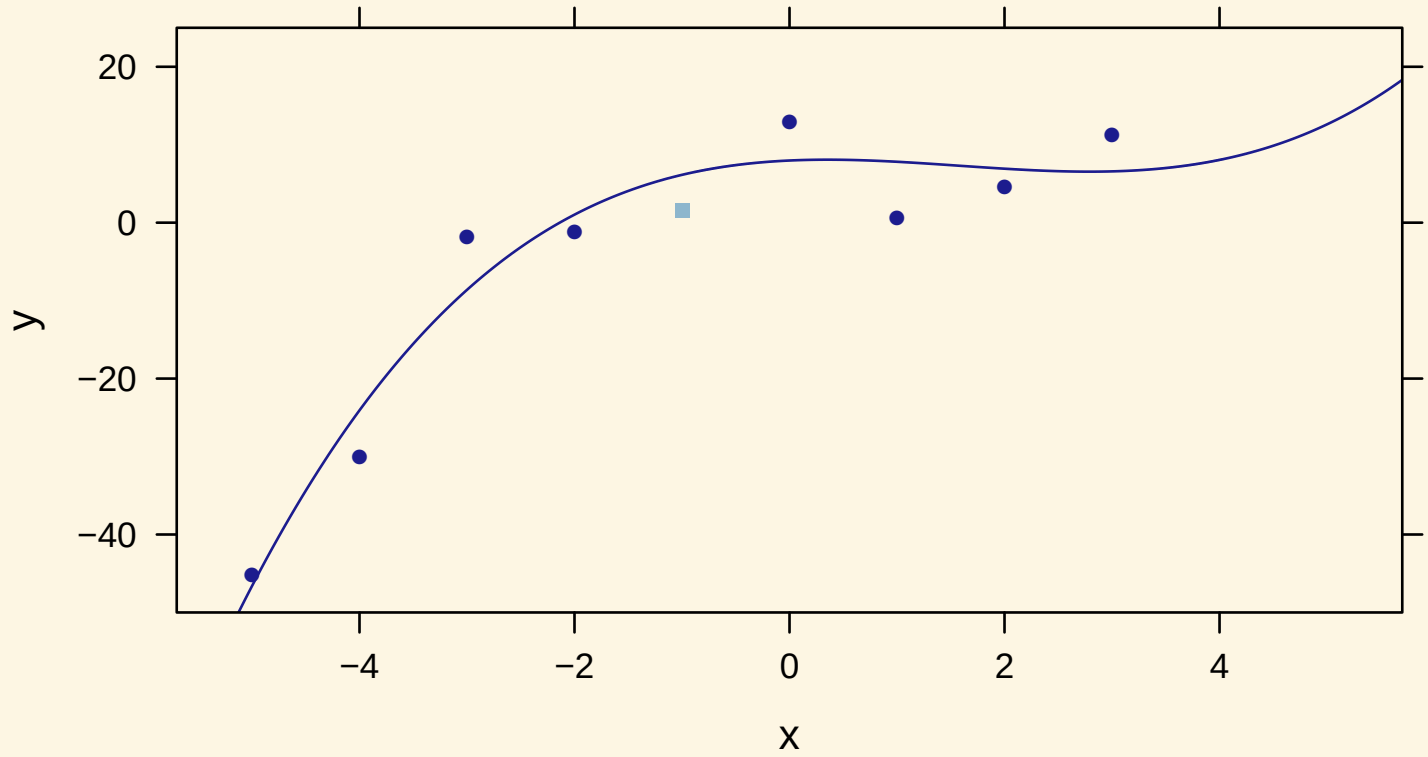
```
model10 <- lm(y ~ poly(x, degree = 10, raw = TRUE), data = TheData)
f.hat <- makeFun(model10)
plot(y ~ x, data = TheData)
curve(f.hat(x), add = TRUE)
```



Out-of-Sample Prediction



Out-of-Sample Prediction



Quantitative Vs. Categorical Predictor and Response

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

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)

$$Y = \hat{f}(X) + \hat{\varepsilon}$$
$$Y = \bar{Y}_k + \hat{\varepsilon}, \qquad \hat{\varepsilon} \sim \mathcal{N}(0, s_\varepsilon^2)$$

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

Testing the 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

Hypothesis Test: Is there evidence that the μ_k are not all equal?

The One-Way ANOVA Model: Alternative Formulation

The One-way ANOVA Population Model (X categorical)

$$Y = f(X) + \varepsilon$$

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

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

Hypothesis Test: 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

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

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

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

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 H_1 : some $\alpha_k \neq 0$

Conditions for ANOVA Test

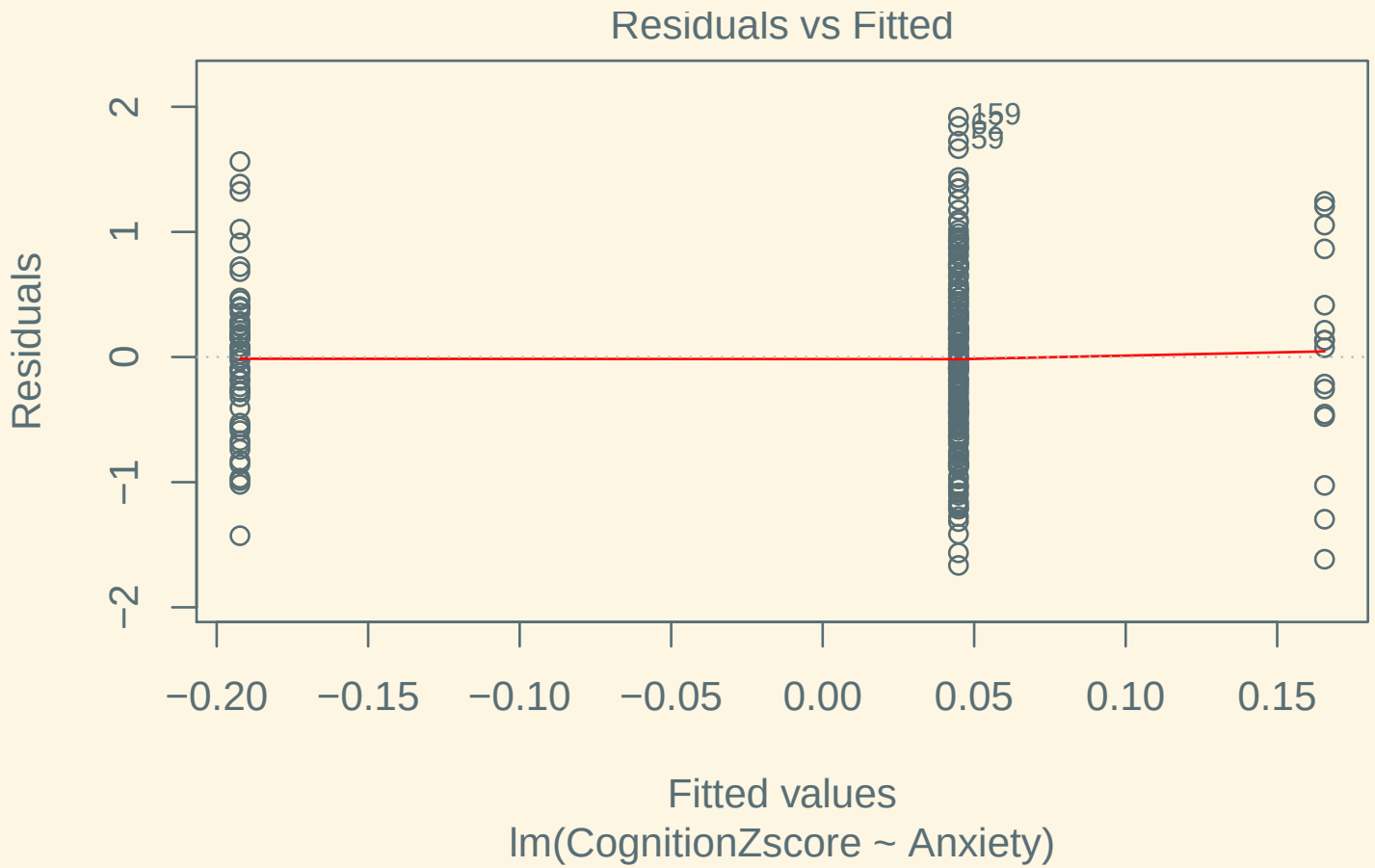
Can view these as regression models and compute P-value assuming the same conditions as for SLR (except linearity):

Conditions for ANOVA Test

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

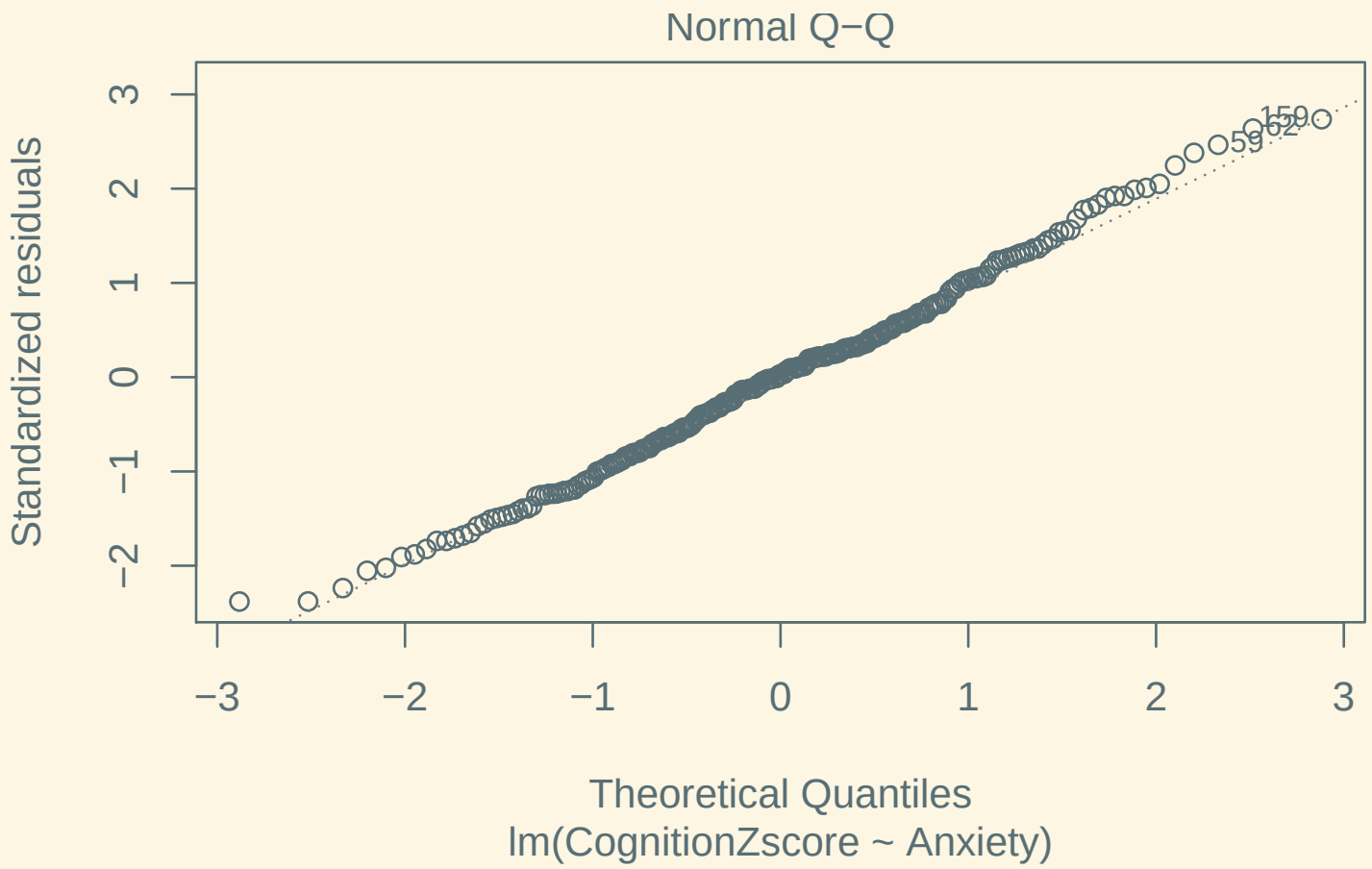
Checking Conditions: Residual Plots

```
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.

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^2_ε

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 $\mu_{Athlete} - \mu_{NoPrime}$.