

# Lecture 11

Review Section 3.5 from last Monday (on board)

Overview of today's example (on board)

Section 3.6, Continued: Nested F tests, review on board first

Section 3.4:

Interaction for quantitative variables (on board)

Polynomial Regression

Especially quadratic

Second-order models (including interaction)

# Example: 1982 State SAT Scores (First year state by state data available)

Unit = A state in the United States

Response Variable:

$Y$  = Average combined SAT Score

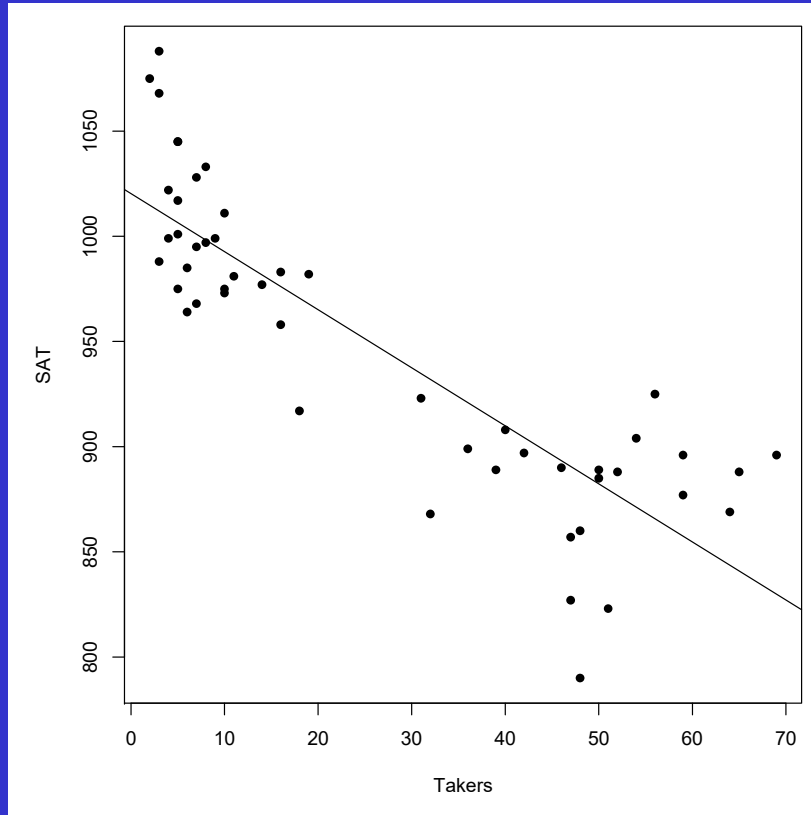
Potential Predictors:

$X_1$  = Takers = % taking the exam out of all eligible students in that state

$X_2$  = Expend = amount spent by the state for public secondary schools, per student (\$100's)

Is  $Y$  related to one or both of these  $X$  variables?

# Example: State SAT with $X_1$ only

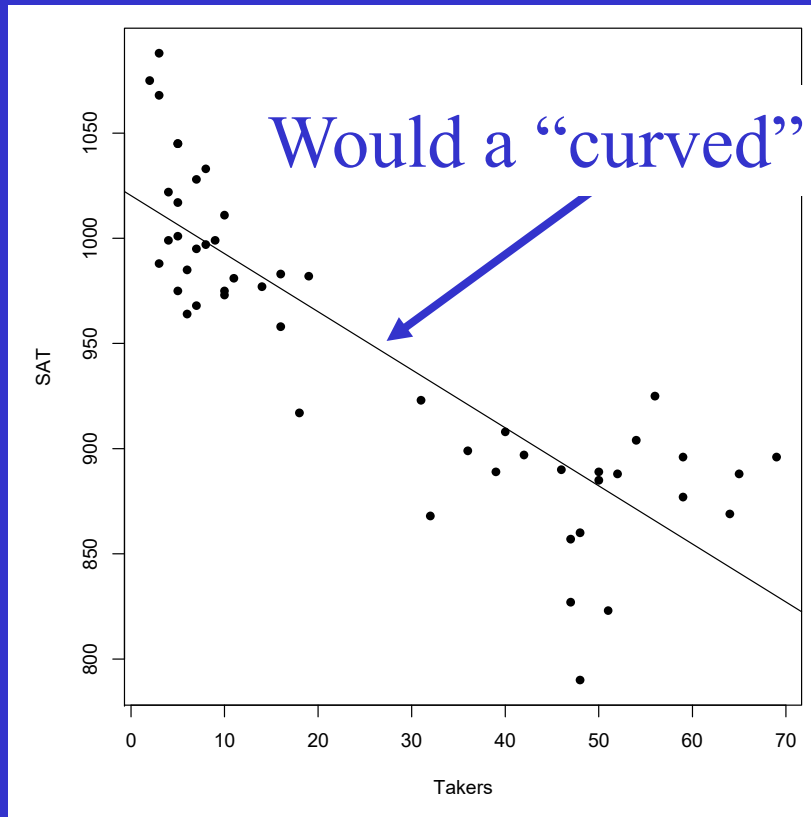


$Y =$  Combined SAT  
 $X =$  % Taking SAT

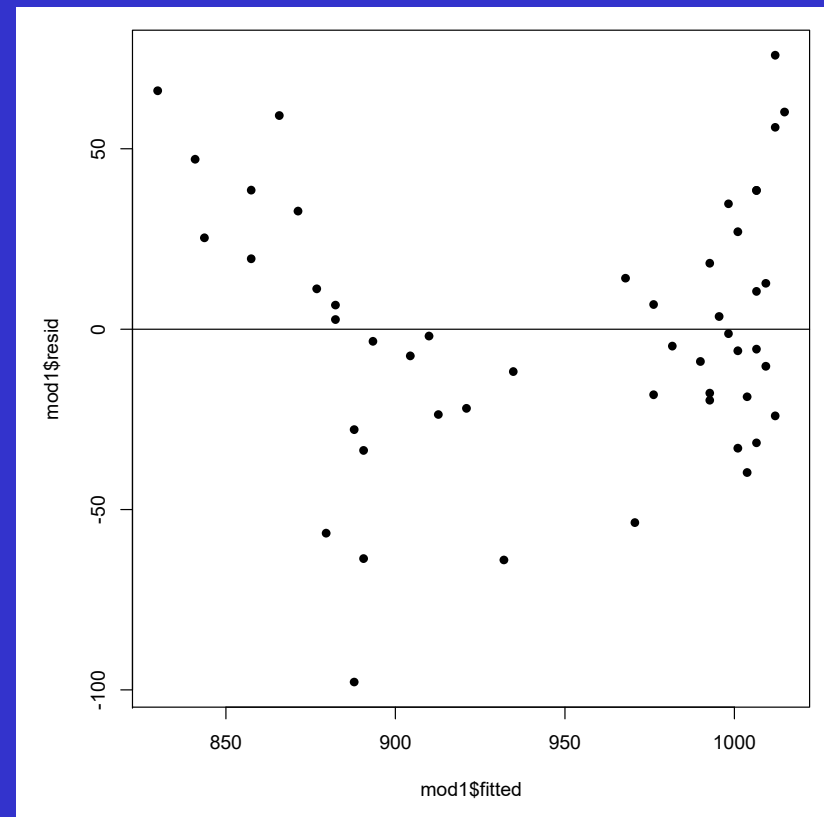
## Things to notice:

- Two clusters in the  $X$  range. Why?
- Possible curved relationship
- As %Takers goes up, average SAT goes *down*.

# Example: State SAT with $X_1$ only



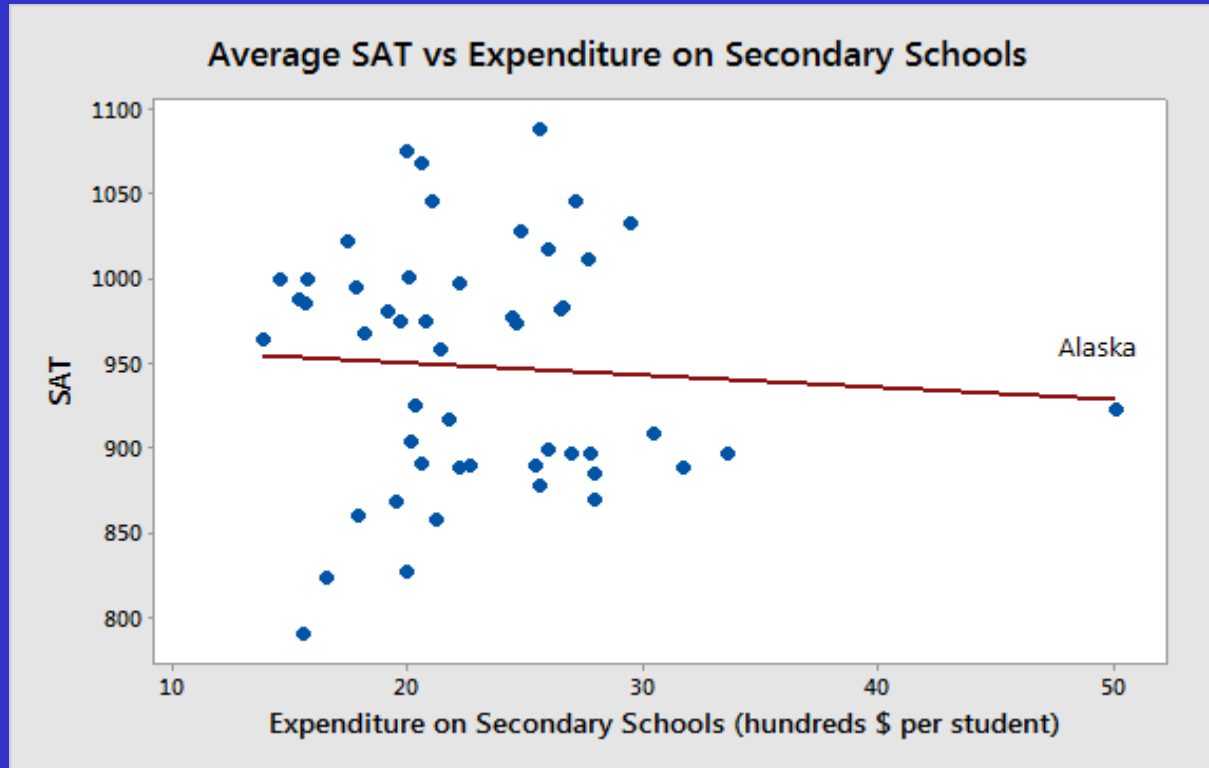
Would a “curved” line work better?



$Y =$  Combined SAT  
 $X =$  % Taking SAT

Residuals vs fitted values

# Example: State SAT with $X_2$ only



$Y =$  Combined SAT  
 $X =$  Expenditure

Not clear what pattern is; Alaska is very influential.

# Polynomial Regression

For a single predictor  $X$ :

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (\text{Linear})$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \quad (\text{Quadratic; curve})$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon \quad (\text{Cubic})$$

## Polynomial Regression in R

Method #1: Create new columns with powers of the predictor.

To avoid creating a new column...

Method #2: Use `I( )` in the `lm( )`

```
quadmod=lm(SAT~Takers+I(Takers^2))
```

Method #3: Use `poly`

```
quadmod=lm(SAT~poly(Takers, degree=2, raw=TRUE))
```

# Quadratic Model

```
> Quad<-lm(sat~takers+I(takers^2), data=StateSAT)
```

```
> summary(Quad)
```

Call:

```
lm(formula = sat ~ takers + I(takers^2), data = StateSAT)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	1053.13112	9.27372	113.561	< 2e-16	***
takers	-7.16159	0.89220	-8.027	2.32e-10	***
I(takers^2)	0.07102	0.01405	5.055	6.99e-06	***

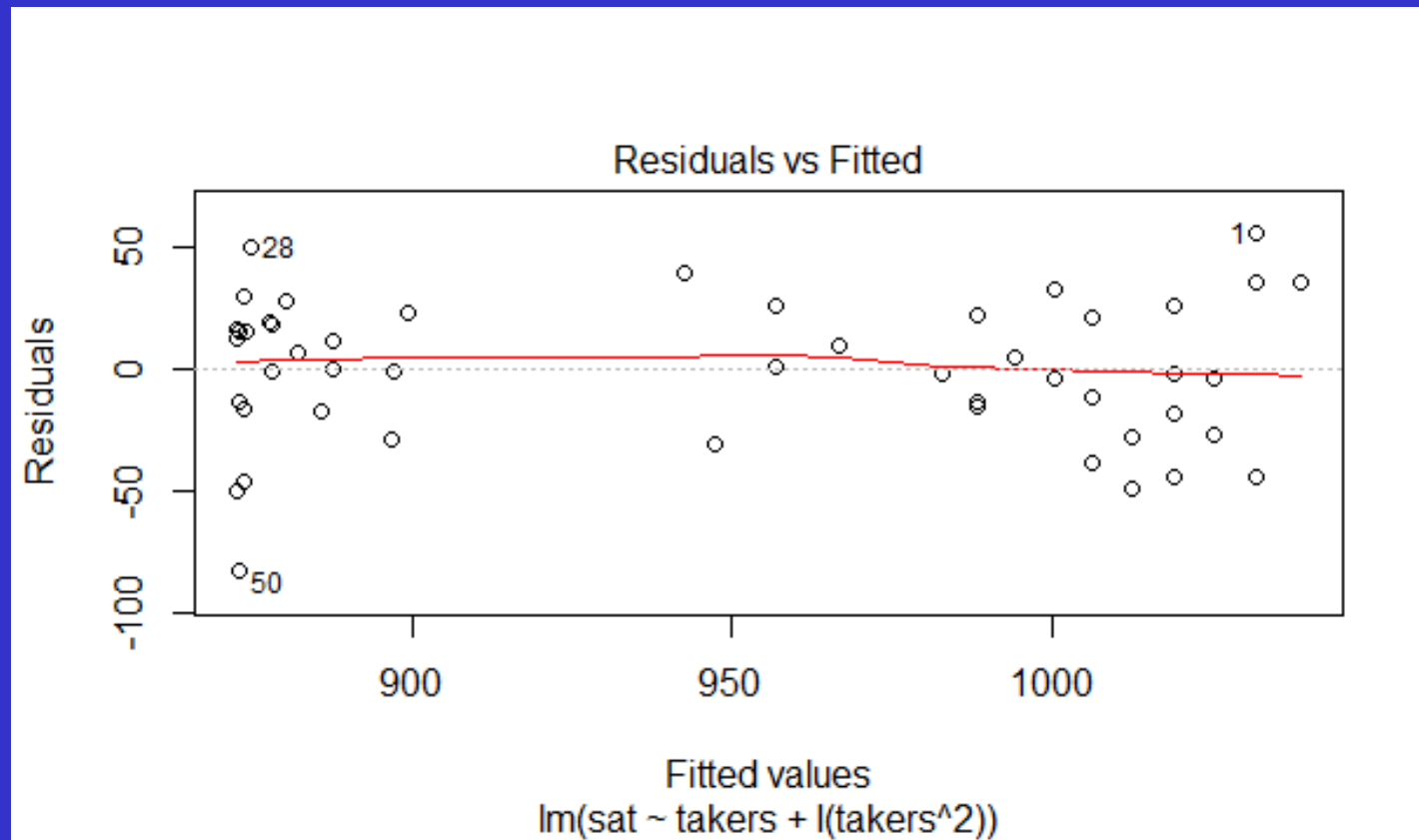
Residual standard error: 29.93 on 47 degrees of freedom

Multiple R-squared: 0.8289, Adjusted R-squared: 0.8216

F-statistic: 113.8 on 2 and 47 DF, p-value: < 2.2e-16



# Residual Plot Looks Good (Two clusters still obvious)



# How to Choose the Polynomial Degree?

- Use the minimum degree needed to capture the structure of the data.
- Check the t-test for the highest power.
- (Generally) keep lower powers—even if not “significant.”

# Interaction

Recall:

$$Active = \beta_0 + \beta_1 Rest + \beta_2 Gender + \beta_3 Rest * Gender + \varepsilon$$


Product allows for different Active/Rest slopes for different genders

**In General:** Interaction is present if the relationship between two variables (e.g.  $Y$  and  $X_1$ ) changes depending on a third variable (e.g.  $X_2$ ).

**Modeling tip:** Include a product term to account for interaction.

# Complete Second-order Models

**Definition:** A complete second-order model for two predictors would be:

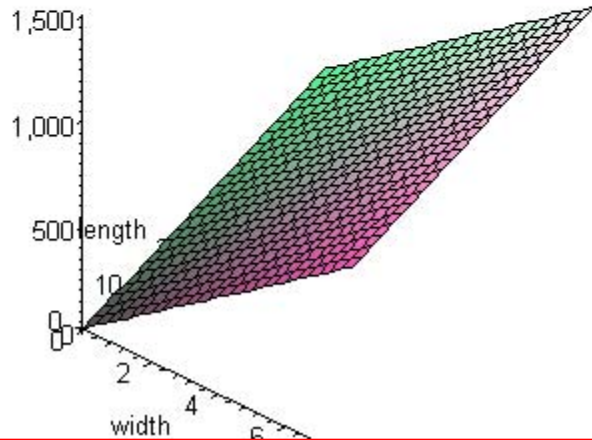
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon$$

First order

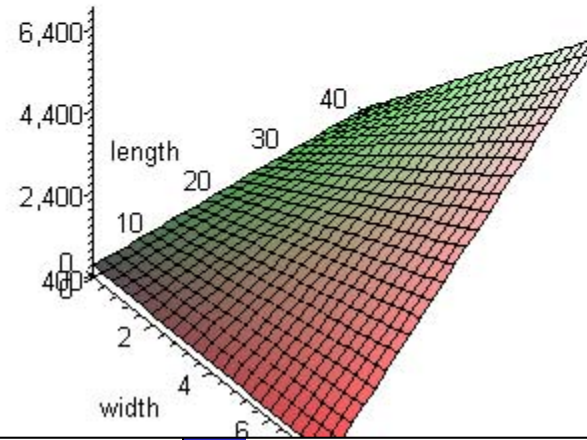
Quadratic

Interaction

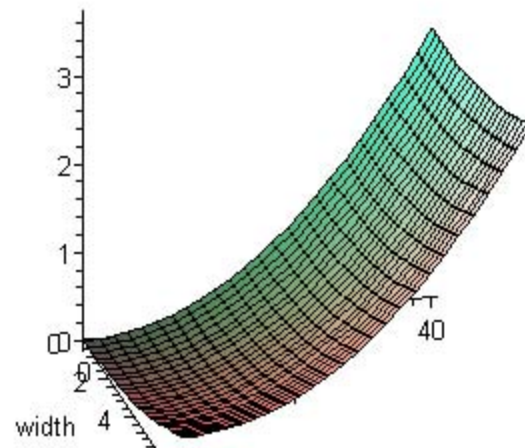
Just linear for both



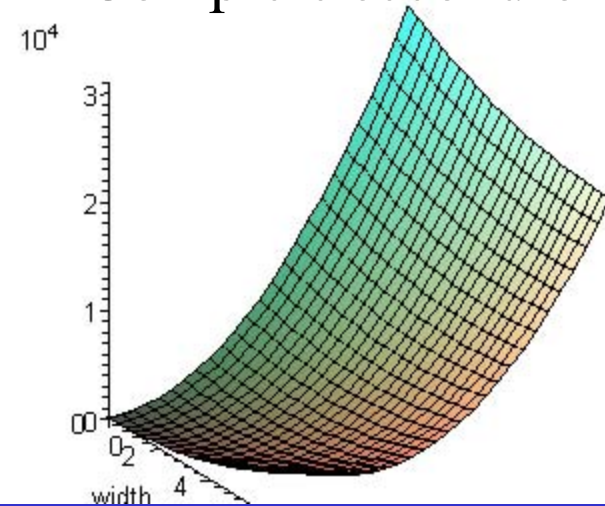
Linear +  $X_1 X_2$  Interaction



$10^4$  Linear + Quadratic



Complete second-order



# Second-order Model for State SAT

Example: Try a full second-order model for  $Y = \text{SAT}$  using  $X_1 = \text{Takers}$  and  $X_2 = \text{Expend}$ .

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon$$

```
secondorder=lm(SAT~Takers+I(Takers^2)  
+Expend+I(Expend^2)+Takers:Expend,  
data=StateSAT)
```

# Second-order Model for State SAT

## summary(secondorder)

```
lm(formula = sat ~ takers + I(takers^2) + expend + I(expend^2) +  
    takers:expend, data = StatesAT)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	893.66283	36.14094	24.727	< 2e-16	***
takers	-7.05561	0.83740	-8.426	9.96e-11	***
I(takers^2)	0.07725	0.01328	5.816	6.28e-07	***
expend	10.33333	2.49600	4.140	0.000155	***
I(expend^2)	-0.11775	0.04426	-2.660	0.010851	*
takers:expend	-0.03344	0.03716	-0.900	0.373087	

Residual standard error: 23.68 on 44 degrees of freedom

Multiple R-squared: 0.8997, Adjusted R-squared: 0.8883

F-statistic: 78.96 on 5 and 44 DF, p-value: < 2.2e-16

Do we really need the quadratic terms?

Nested  
F-test

## anova(secondorder) [FULL MODEL]

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
takers	1	181024	181024	322.8794	< 2.2e-16	***
I(takers^2)	1	22886	22886	40.8198	9.035e-08	***
expend	1	11700	11700	20.8678	3.956e-05	***
I(expend^2)	1	5278	5278	9.4148	0.003677	**
takers:expend	1	454	454	0.8098	0.373087	
Residuals	44	24669	561			

## firstorder=lm(SAT~Takers\*Expend) [REDUCED = NO QUADRATIC]

### anova(firstorder)

Response: SAT

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Takers	1	181024	181024	152.6279	3.245e-16 ***
Expend	1	8709	8709	7.3428	0.009429 **
Takers:Expend	1	1720	1720	1.4499	0.234710
Residuals	46	54558	1186		

### anova(firstorder,secondorder) [COMPARE]

Model 1: SAT ~ Takers \* Expend

Model 2: SAT ~ Takers + I(Takers^2) + Expend + I(Expend^2) + Takers:Expend

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	46	54558				
2	44	24669	2	29889	26.656	2.608e-08 ***

The quadratic terms are significant as a pair (as well as individually).



# Do we really need the terms with Expend? Nested F-test

Simultaneously test all three terms involving “Expend” in the second order model with “Takers” to predict SAT scores.

```
anova(secondordermodel)
```

Response: SAT					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Takers	1	181024	181024	322.8794	< 2.2e-16 ***
I(Takers^2)	1	22886	22886	40.8198	9.035e-08 ***
Expend	1	11700	11700	20.8678	3.956e-05 ***
I(Expend^2)	1	5278	5278	9.4148	0.003677 **
Takers:Expend	1	454	454	0.8098	0.373087
Residuals	44	24669	561		

# Do we really need the terms with Expend? Nested F-test

Simultaneously test all three terms involving “Expend” in the second order model with “Takers” to predict SAT scores.

```
anova(secondordermodel)
```

Response: SAT	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Takers	1	181024	181024	322.8794	< 2.2e-16 ***
I(Takers^2)	1	22886	22886	40.8198	9.035e-08 ***
Expend	1	11700	11700	21.1111	1.111e-05 ***
I(Expend^2)	1	5278	5278	9.5556	0.003316 **
Takers:Expend	1	454	454	0.8222	0.369118
Residuals	44	24669	560.659		

17432 explained by adding the three predictors

$$t.s. = \frac{17432 / 3}{24669 / 44} = 10.36$$

Compare to F(3,44)

P-value=0.000028

(next slide)

## anova(secondorder) [FULL MODEL]

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
takers	1	181024	181024	322.8794	< 2.2e-16	***
I(takers^2)	1	22886	22886	40.8198	9.035e-08	***
expend	1	11700	11700	20.8678	3.956e-05	***
I(expend^2)	1	5278	5278	9.4148	0.003677	**
takers:expend	1	454	454	0.8098	0.373087	
Residuals	44	24669	561			

## Takersmodel=lm(SAT~Takers+I(Takers^2)) [REDUCED MODEL]

## anova(Takersmodel)

Response: SAT

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Takers	1	181024	181024	202.089	< 2.2e-16 ***
I(Takers^2)	1	22886	22886	25.549	6.992e-06 ***
Residuals	47	42101	896		

## anova(Takersmodel,secondorder) [COMPARE]

Model 1: SAT ~ Takers + I(Takers^2)

Model 2: SAT ~ Takers + I(Takers^2) + Expend + I(Expend^2) + Takers:Expend

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	47	42101				
2	44	24669	3	17432	10.364	2.787e-05 ***

Three “new” predictors reduce the SSE by 17432, a sig. amount.

# Second-order Model for State SAT

## summary(secondorder)

```
lm(formula = sat ~ takers + I(takers^2) + expend + I(expend^2) +  
    takers:expend, data = StatesAT)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	893.66283	36.14094	24.727	< 2e-16	***
takers	-7.05561	0.83740	-8.426	9.96e-11	***
I(takers^2)	0.07725	0.01328	5.816	6.28e-07	***
expend	10.33333	2.49600	4.140	0.000155	***
I(expend^2)	-0.11775	0.04426	-2.660	0.010851	*
takers:expend	-0.03344	0.03716	-0.900	0.373087	

Residual standard error: 23.68 on 44 degrees of freedom

Multiple R-squared: 0.8997, Adjusted R-squared: 0.8883

F-statistic: 78.96 on 5 and 44 DF, p-value: < 2.2e-16

Do we really need the interaction? T-test for  
takers:expend

# SUMMARY

- Full second-order model is better than the model with no quadratic terms
- Full second-order model is better than the quadratic model with “Takers” only
- Model with no interaction seems acceptable
- Comparing Adjusted R-squared:
  - Full model: 88.83%; No interaction: 88.88%
  - No quadratic terms: 76.38%
  - Takers only, quadratic: 82.16%
  - Expend only, quadratic: -3.59%, partly because of the extreme outlier for Alaska!