Transformations for Regression Modeling

SDS 291 – Multiple Regression

March 30, 2020

What to Do When Regression Conditions Are Violated

- **Examples:**
- Lack of normality in residuals
- Patterns in residuals
- Heteroscedasticity (nonconstant variance)
- Outliers: influential points, large residuals

Data Transformations

- **Can be used to:**
- Address nonlinear patterns
- Stabilize variance
- Remove skewness from residual
- Minimize effects of outliers

Common Transformations

For either the response (*Y*) or predictor (*X*)...

Logarithm $Y \to \log(Y)$

Square root $Y \to \sqrt{Y}$

Exponentiation $Y \to e^{Y}$

Power function $Y \to Y^3$

Reciprocal $Y \to 1/Y$

Example: Planets

- *Year* = length of the “year” for planets
- *X* = distance from the Sun

```
library(tidyverse)

Planets<-
read.csv(url("https://sds291.netlify.com/15/Plane
ts.csv"))


qplot(x=Distance, y=Year, data=Planets)


Planets <- Planets %>%

         mutate(newyear=Year^(2/3))


qplot(x=Distance, y=newyear, data=Planets)
```

Example: Planets

- Planets.R
- Y = length of the "year" for planets
- X = distance from the Sun

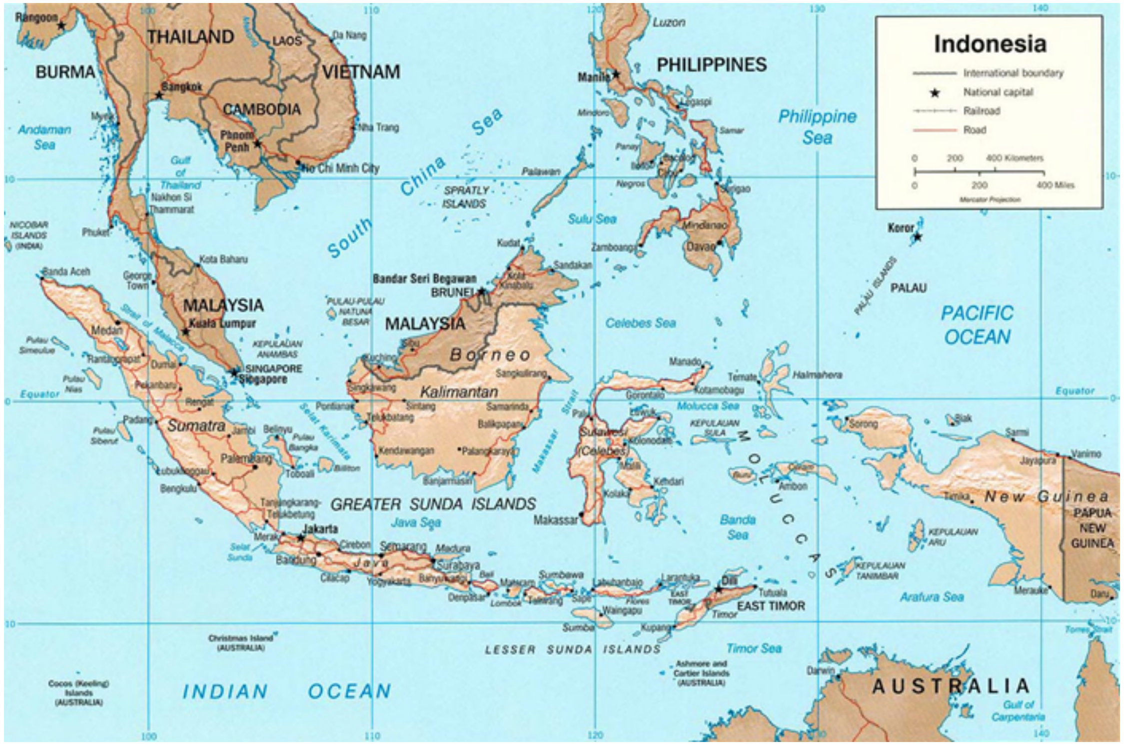
- **Try scatterplots and SLM with**
- Y vs. X
- $\log(Y)$ vs. X
- Y vs. $\log(X)$
- $\log(Y)$ vs. $\log(X)$

- Note: the default log in R is natural log (ln) or log base e.

Example: Mammal Species (1 of 2)

- Y = number of mammal species on an island
- X = area of the island

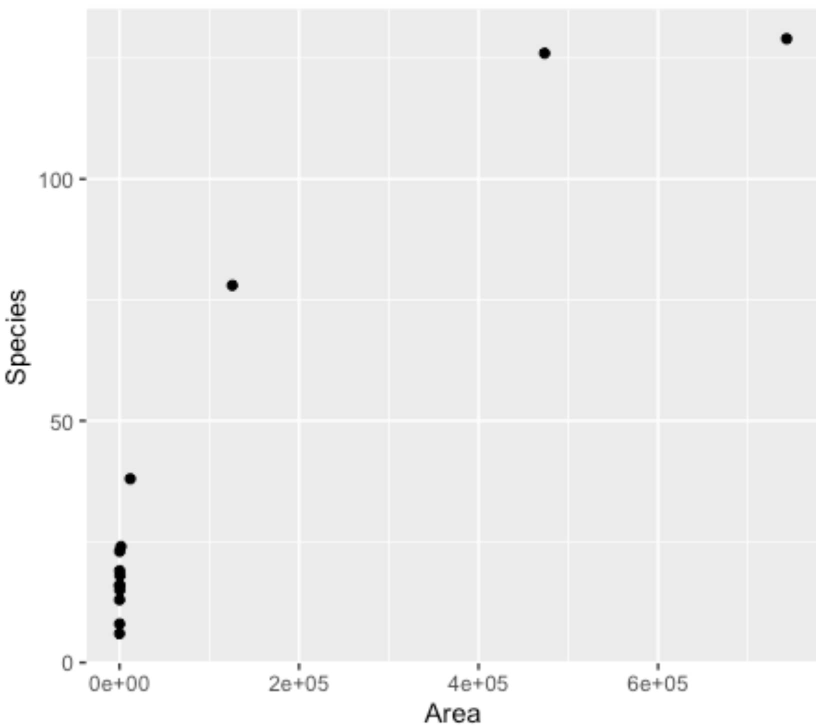
- Data on 14 islands in Southeast Asia are stored in `SpeciesArea`.


Indonesia
International boundary
National capital
Railroad
Road
400 Kilometers
400 Miles
Mercator Projection
THAILAND
LAOS
BURMA
VIETNAM
CAMBODIA
PHILIPPINES
Rangoon
Bangkok
Da Nang
Phnom Penh
Nha Trang
Ho Chi Minh City
Manila
Luzon
Legaspi
Mindoro
Samar
Panay
Negros
Palawan
Mindanao
Davao
Zamboanga
Philippine Sea
Andaman Sea
Gulf of Thailand
South China Sea
SPRATLY ISLANDS
Sulu Sea
Celebes Sea
NICOBAR ISLANDS (INDIA)
Phuket
Nakhon Si Thammarat
Kota Baharu
George Town
MALAYSIA
Kuala Lumpur
Bandar Seri Begawan
BRUNEI
Kudat
Sandakan
Kota Kinabalu
PULAU-PULAU NATUNA BESAR
KEPULAUAN ANAMBAS
SINGAPORE
Singapore
Banda Aceh
Medan
Pulau Simeulue
Pulau Nias
Pulau Siberut
Dumai
Pekanbaru
Rengat
Padang
Sumatra
Jambi
Palembang
Bengkulu
Pulau Bangka
Belitung
Borneo
Kalimantan
Kuching
Sibu
Singkawang
Pontianak
Sintang
Samarinda
Balikpapan
Palangkaraya
Kendawangan
Banjarmasin
Makassar Strait
Equator
Sulawesi (Celebes)
Manado
Gorontalo
Palu
Kolaka
Kendari
Makassar
Molucca Sea
KEPULAUAN SULA
Halmahera
Ternate
MOLUCCAS
Ambon
Banda Sea
Koror
PALAU
PALAU ISLANDS
PACIFIC OCEAN
Sorong
Biak
Sarmi
Jayapura
Vanimo
New Guinea
PAPUA NEW GUINEA
Timika
Merauke
Daru
KEPULAUAN ARU
KEPULAUAN TANIMBAR
Arafura Sea
Torres Strait
GREATER SUNDA ISLANDS
Java Sea
Jakarta
Merak
Bandung
Cirebon
Semarang
Madura
Java
Surabaya
Cilacap
Yogyakarta
Bali
Denpasar
Lombok
Mataram
Sumbawa
Flores
Sumba
Waingapu
Kupang
Timor
Dili
EAST TIMOR
LESSER SUNDA ISLANDS
Timor Sea
Christmas Island (AUSTRALIA)
Cocos (Keeling) Islands (AUSTRALIA)
Ashmore and Cartier Islands (AUSTRALIA)
INDIAN OCEAN
Darwin
AUSTRALIA
Gulf of Carpentaria

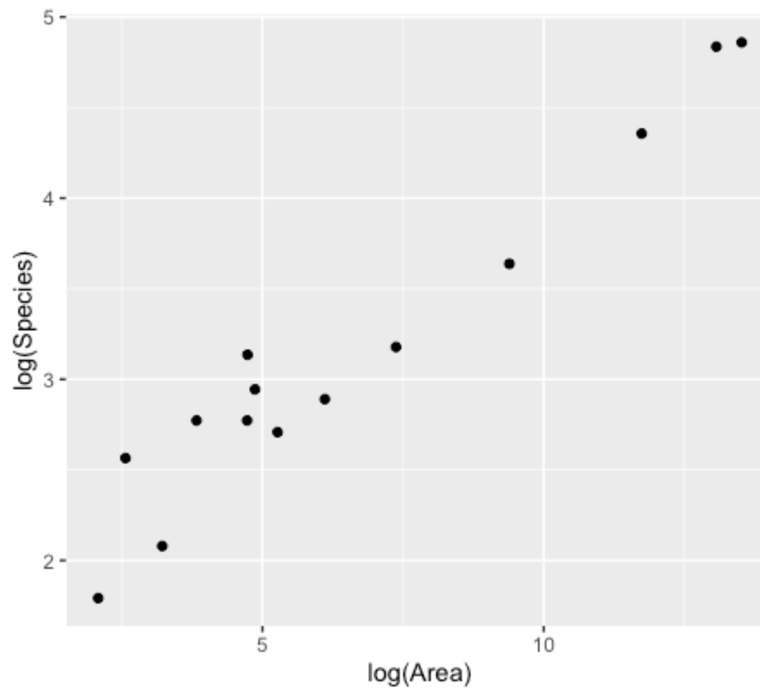
Example: Mammal Species (2 of 2)

Y = number of mammal species on an island & X = area of the island

Original Form



Logged Both X and Y



logSpecies= 1.625 + 0.235logArea

➔Species = 5.08 · Area$^{0.235}$

Code in [SpeciesArea.R]()

Why a Log Transformation?

Some relationships are *multiplicative* (not linear).

Example : Area of a circle

$A = \pi r^2$ (not linear)

$\log(A) = \log(\pi r^2) = \log(\pi) + 2\log(r)$

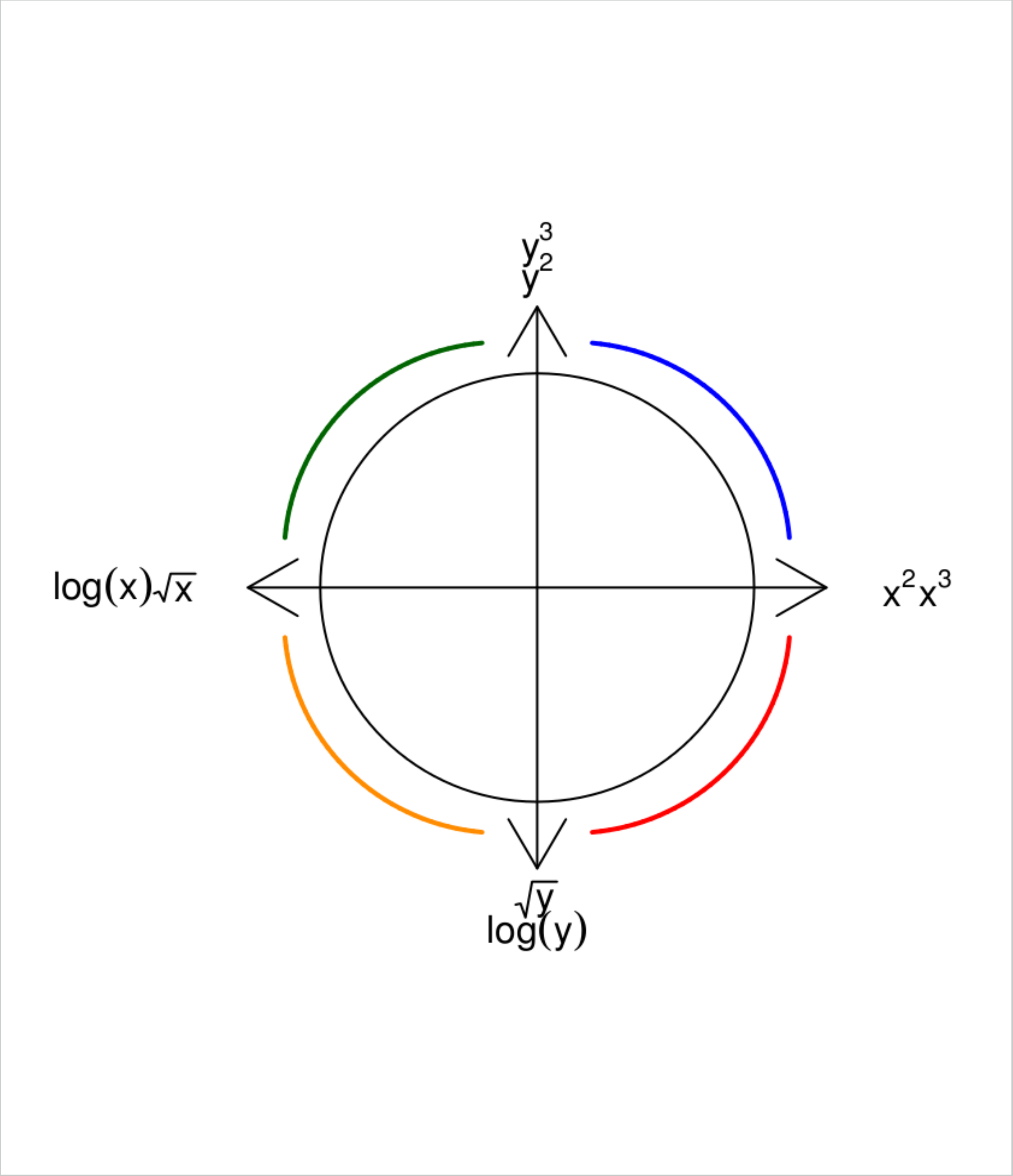
$\Rightarrow \log(A)$ is a linear function of $\log(r)$.

Look for:

- Strongly right-skewed distributions
- Curvature in scatterplot
- Increasing variability in scatterplot

What Kind of Transformation?

- Tukey's Ladder



Interaction

Recall:

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

Product allows for different Active/Rest slopes for females and males

Interaction: When the relationship between two variables changes depending on a third variable.

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

Example 3.11 in the Text: Fish Weights (1 of 3)

- Data: **`Perch`** (measurements for 56 fish)
- Predictors: **`Length, Width`** (in cm)
- Response: **`Weight`** (in gm)

- Fit a two-predictor model with an interaction.

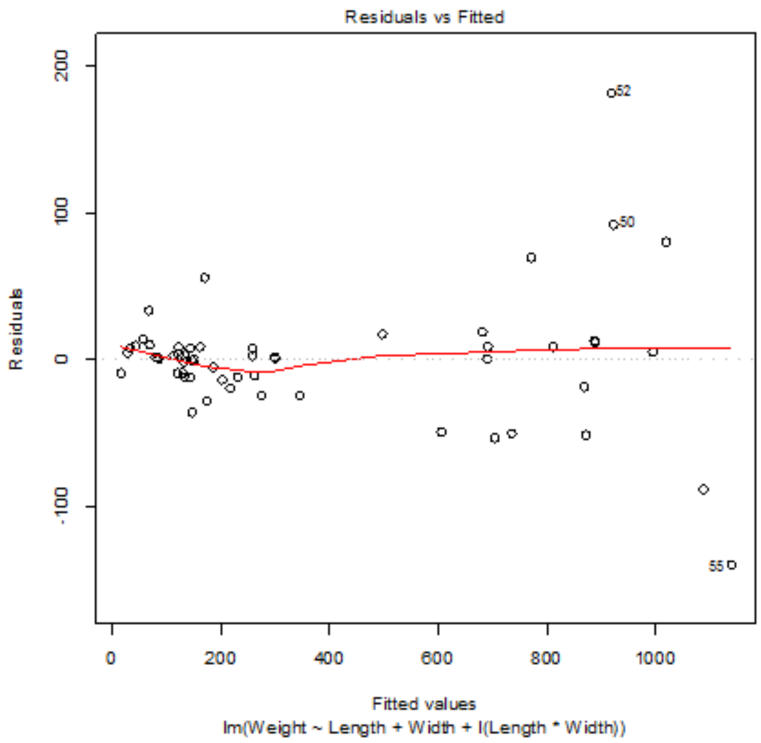
Example 3.11 in the Text: Fish Weights (2 of 3)

```
> Perchmodel=lm(Weight~Length+Width+I(Length*Width))
> summary(Perchmodel)
Call:
lm(formula = Weight ~ Length + Width + Length * Width)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      113.9349    58.7844   1.938    0.058 .
Length            -3.4827     3.1521  -1.105    0.274
Width            -94.6309    22.2954  -4.244  9.06e-05 ***
I(Length*Width)   5.2412     0.4131  12.687  < 2e-16 ***

Residual standard error: 44.24 on 52 degrees of freedom
Multiple R-squared: 0.9847,     Adjusted R-squared: 0.9838
F-statistic:  1115 on 3 and 52 DF,  p-value: < 2.2e-16
```

To avoid creating a new column, use **I( )** in the **lm( )**

Example 3.11 in the Text: Fish Weights (3 of 3)

- All three terms are significant. (But there is a pattern in the residual plot . . . might try **log(Weight)**.)




```
> anova(Perchmodel)
Analysis of Variance Table
Response: Weight
                Df  Sum Sq Mean Sq  F value    Pr(>F)
Length           1 6118739 6118739 3126.571 < 2.2e-16 ***
Width            1  110593  110593   56.511 7.416e-10 ***
I(Length*Width)1  314997  314997  160.958 < 2.2e-16 ***
Residuals       52  101765    1957
```

Example: State SAT Scores

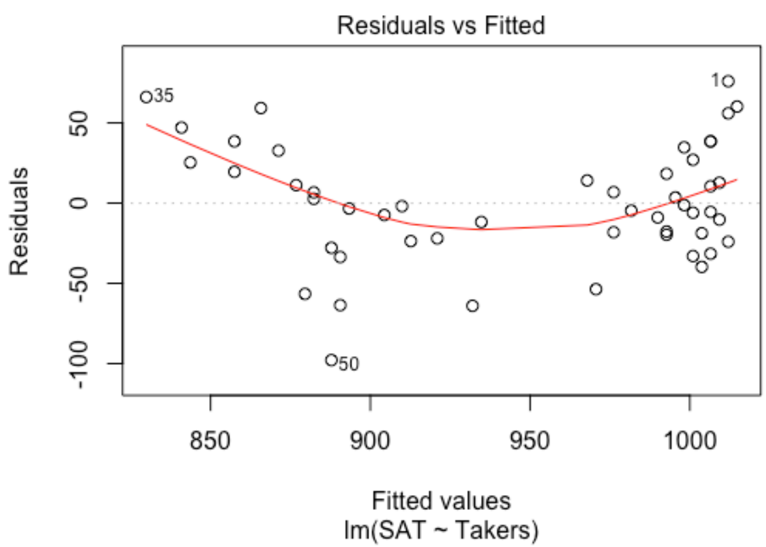
- **Response variable:**
- *Y* =average combined SAT score

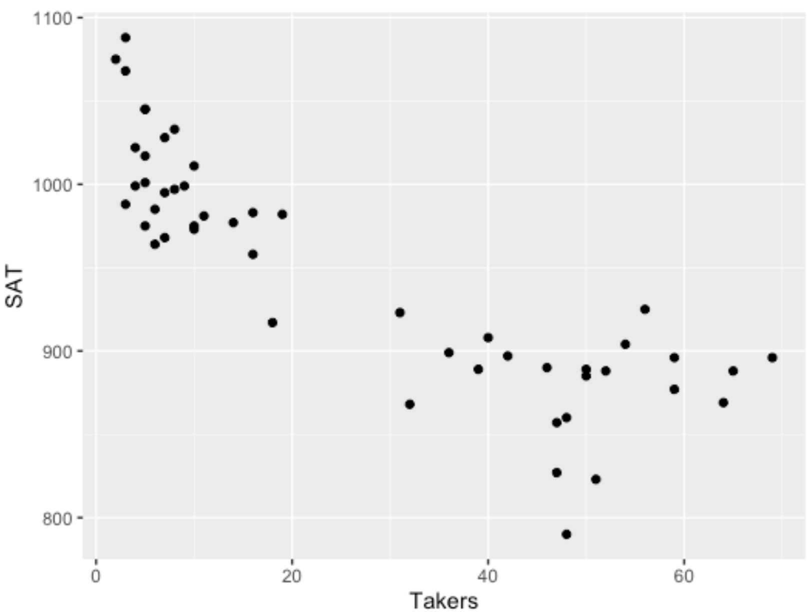
- **Potential predictors:**
- Takers = % taking the exam
- Expend = spend per student ($100's)

- Data file: StateSAT82

Example: State SAT

- Y = combined SAT
- X = % taking SAT





Polynomial Regression

For a single predictor X :

$$Y = \beta_o + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p + \varepsilon$$

$$Y = \beta_o + \beta_1 X + \varepsilon \text{ (linear)}$$

$$Y = \beta_o + \beta_1 X + \beta_2 X^2 + \varepsilon \text{ (quadratic)}$$

$$Y = \beta_o + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon \text{ (cubic)}$$

Polynomial Regression in R

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

```
library(tidyverse)

StateSAT82 <- StateSAT82 %>%

       mutate(Takers_Sq=Takers^2)

quadmod_method1 <- lm(SAT~Takers+Takers_Sq, data=StateSAT82)
```

To avoid creating a new column:

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

```
quadmod_method2 <- lm(SAT~Takers+I(Takers^2), data=StateSAT82)
```

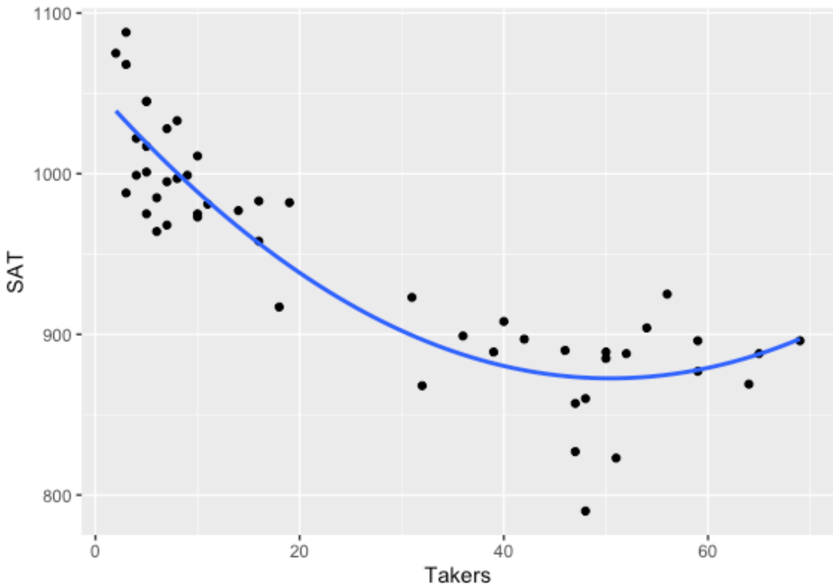
Code in [StateSAT.R]

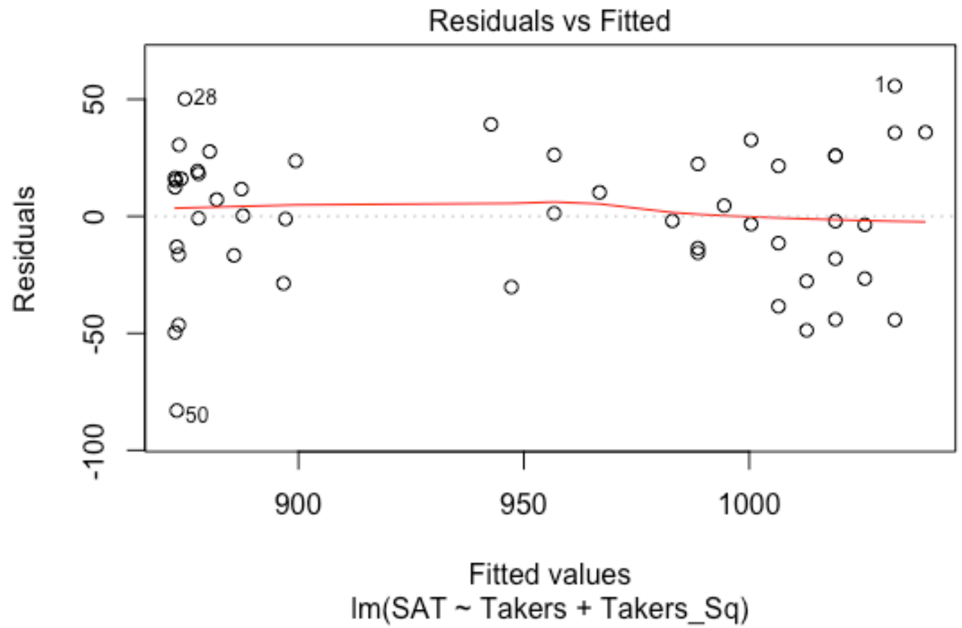
Quadratic Model for SAT

```
> quadmod_method2 <- lm(SAT~Takers+I(Takers^2), data=StateSAT82)
> summary(quadmod_method2)

Call:
lm(formula = SAT ~ Takers + I(Takers^2), data = StateSAT82)

Residuals:
    Min      1Q  Median      3Q     Max 
-83.015 -16.636   0.783  22.167  55.714 

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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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
```

Code in [StateSAT.R]()

Plot the Quadratic Fit





Code in StateSAT.R

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”

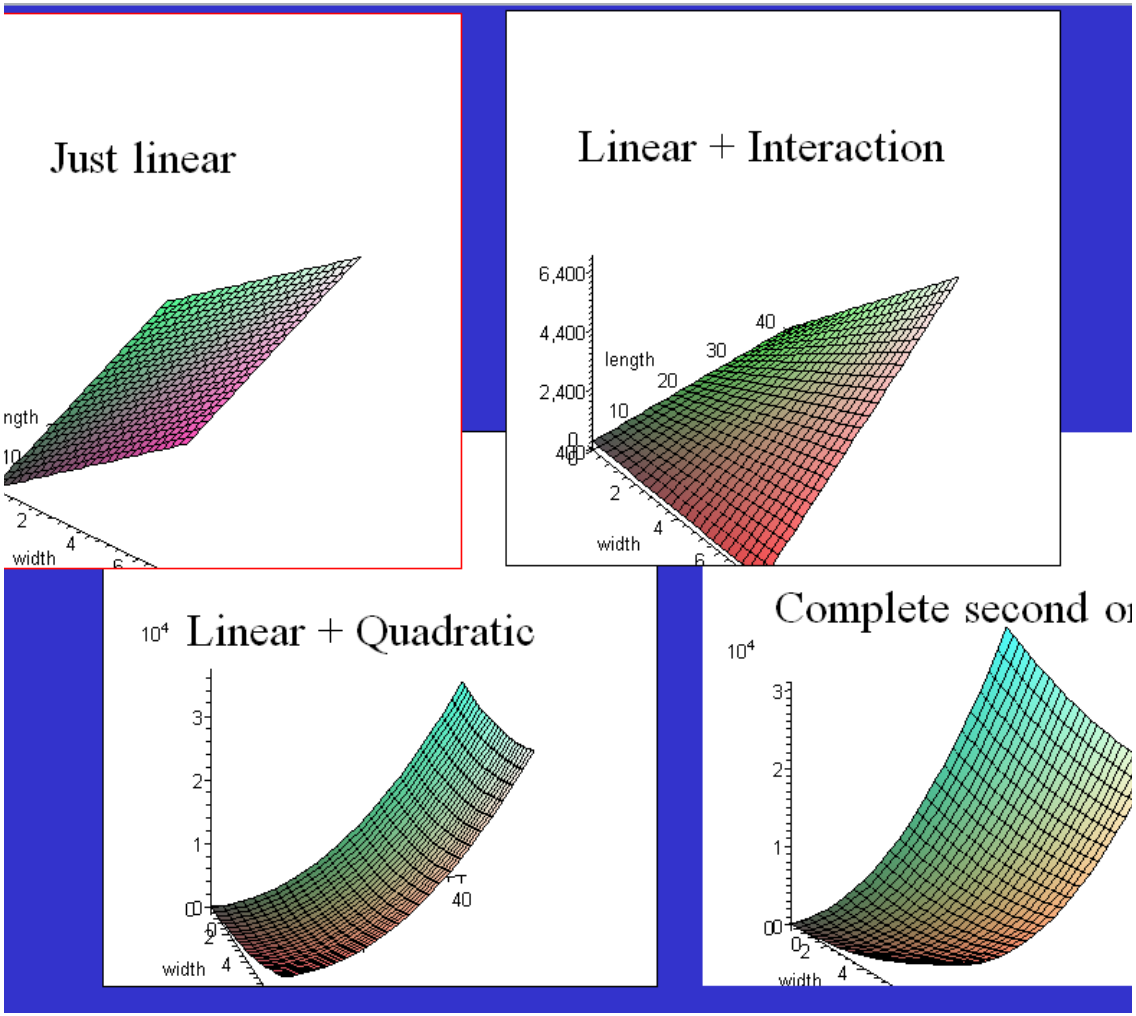
Complete Second-Order Models (1 of 2)

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

$$Y = \beta_o + \underbrace{\beta_1 X_1 + \beta_2 X_2}_{\text{first order}} + \underbrace{\beta_3 X_1^2 + \beta_4 X_2^2}_{\text{quadratic}} + \underbrace{\beta_5 X_1 X_2}_{\text{interaction}} + \varepsilon$$

Example: Try a full second-order model for

$Y = SAT$ using X_1 = Takers and X_2 = Expend



Complete Second-Order Models (2 of 2)

Second-Order Model for State SAT

```
modSAT5=lm(SAT~Takers+Expend+I(Takers^2)+I(Expend^2)+I(Takers*Expend)
,data=StateSAT)
summary(modSAT5)
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 ***
Expend              10.33333    2.49600   4.140 0.000155 ***
I(Takers^2)          0.07725    0.01328   5.816 6.28e-07 ***
I(Expend^2)         -0.11775    0.04426  -2.660 0.010851 *
I(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 need the interaction term? No

Do we need both quadratic terms?

Do we need the terms with *Expend*?

Nested *F* test (Section 3.6)