

Regression and ANOVA

Stat 120

March 01 2023

Simple Linear Model

The population/true simple linear model is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

★ β_0 and β_1 are unknown parameters corresponding to the y-intercept and the slope, respectively

★ ε is the random error

★ Estimate with b_0 and b_1 from the least squares line

$$\hat{y} = b_0 + b_1 x$$

How accurate are the estimates?

Recall: Least Square Regression

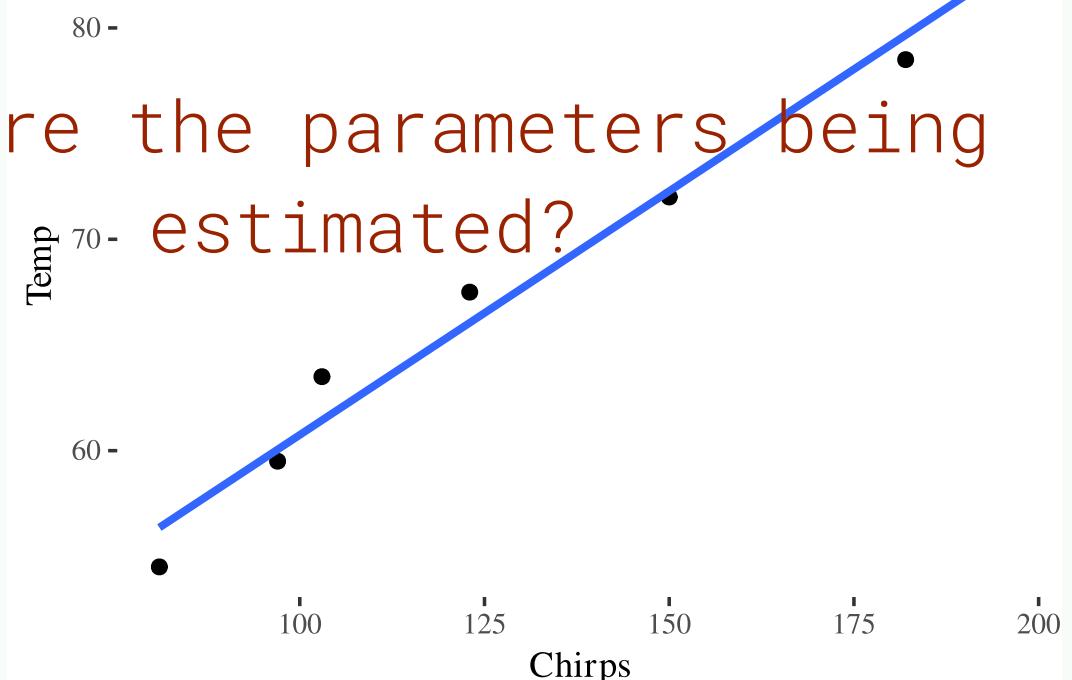
X = Cricket chirp rate

Y = Temperature

Chirps	Temp
81	54.5
97	59.5
103	63.5
123	67.5
150	72.0
182	78.5
195	83.0

$$\widehat{\text{Temp}} = 37.7 + 0.23 \text{ Chirps}$$

What are the parameters being estimated?



Inference for the Slope

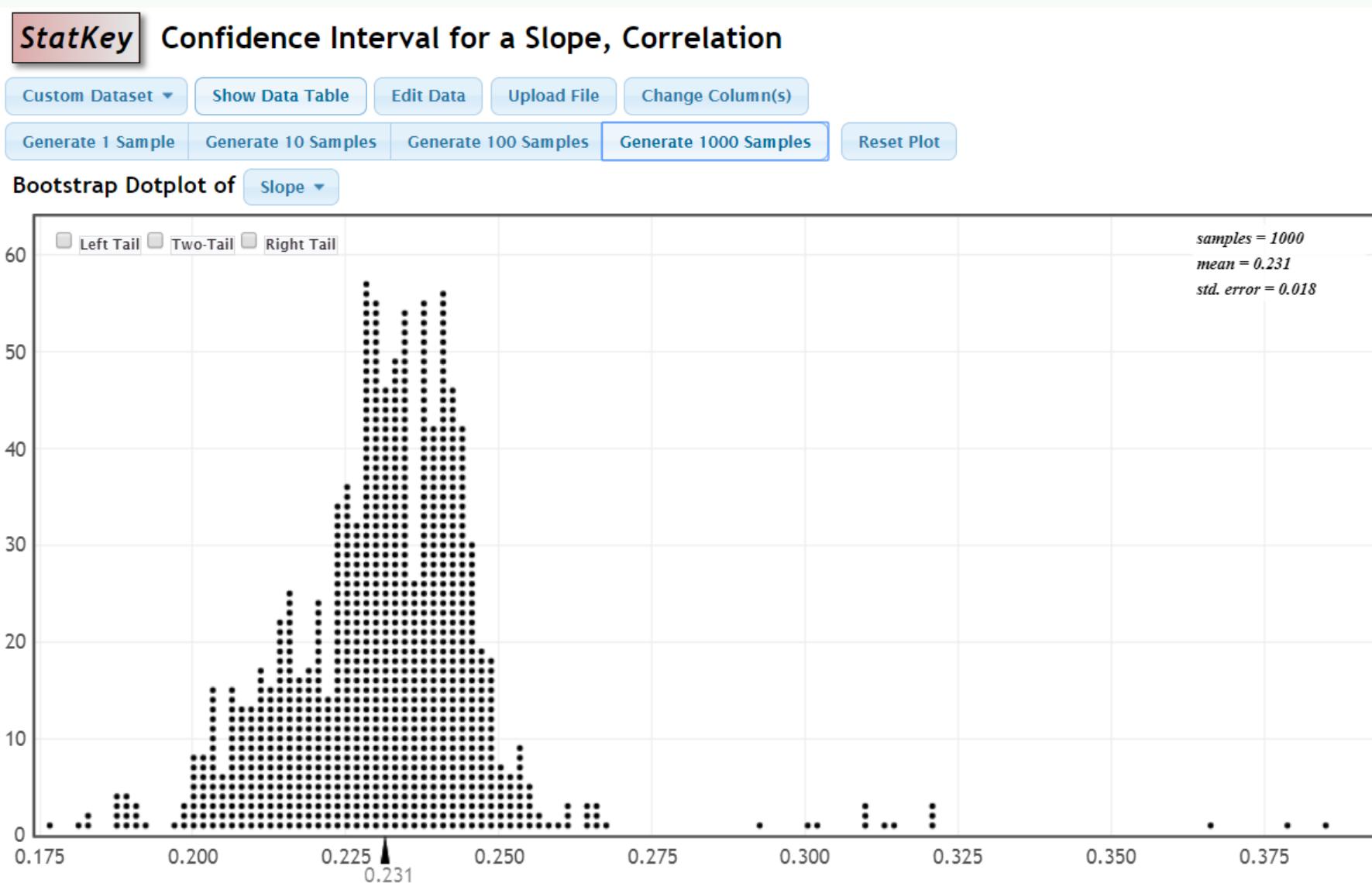
*Confidence intervals and hypothesis tests for the slope can be done using the familiar **formulas**:*

$$b_1 \pm t^* \cdot SE \quad t = \frac{b_1 - \text{null slope}}{SE}$$

But how do we estimate the standard error?

- ★ Bootstrap/Randomization distributions
- ★ Computer output

Statkey: Standard Error Using a Bootstrap Distribution



Technology Examples

Slope estimate and Standard Error

```
chirps.lm <- lm(Temp ~ Chirps, data = data)
summary(chirps.lm)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	37.67858	1.97817	19.05	7.35e-06 ***
Chirps	0.23067	0.01423	16.21	1.63e-05 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.528 on 5 degrees of freedom
Multiple R-squared: 0.9813

Confidence Interval for Slope

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	37.67858	1.97817	19.05	7.35e-06	***
Chirps	0.23067	0.01423	16.21	1.63e-05	***

We can use the values for b_1 and SE from the regression output to form a confidence interval in the usual way:

$$b_1 \pm t^* \cdot SE$$

Here, t^* uses $n - 2$ degrees of freedom, since we are estimating two parameters in the simple linear model.

Confidence Interval for Slope

Find a 95% confidence interval for the slope of the cricket temperature model.

Predictor	Coef	SE Coef	T	Pr(> t)
Constant	37.67858	1.97817	19.05	7.35e-06 ***
Chirps	0.23067	0.01423	16.21	1.63e-05 ***

$$b_1 \pm t^* \cdot SE$$

Hypothesis Test for Slope

Population Simple Linear Model: $y = \beta_0 + \beta_1 x + \varepsilon$

$H_0 : \beta_1 = 0$ \implies No linear relationship

$H_a : \beta_1 \neq 0$ \implies Some relationship

$$t = \frac{\text{statistic-null}}{SE} = \frac{b_1 - 0}{SE} = \frac{b_1}{SE}$$

Again, b_1 and SE come from R output.

We find the p-value by using a t distribution with $n - 2\text{df}$.

Hypothesis Test for Slope

*Confirm the **p-value** given by the regression output for testing the slope of the cricket chirp model.*

Predictor	Coef	SE Coef	T	Pr(> t)
Constant	37.67858	1.97817	19.05	7.35e-06 ***
Chirps	0.23067	0.01423	16.21	1.63e-05 ***

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

$$t = \frac{b_1}{SE}$$

Conditions for the test

- ★ *Linearity* → randomly scattered residuals around 0 in the residuals plot
- ★ *Nearly normally distributed residuals* → QQ-plot plot of residuals
- ★ *Constant variability of residuals (homoscedasticity)* → constant vertical spread of the residuals in residual plot, no patterns
- ★ *Independence of residuals (and hence observations)* → depends on data collection method, often violated for time-series data

If interested to learn more about this, take STAT 230: Applied Linear Regression next quarter!

Standard Error of the Slope

Although we generally rely on technology to obtain the SE for the slope, we can also obtain it as follows:

$$SE = \frac{s_{\varepsilon}}{s_x \sqrt{n - 1}}$$

where s_{ε} is the standard deviation of the error term and s_x is the standard deviation for the sample values of the predictor.

SE for Slope

The regression equation is

$$\widehat{\text{Temperature}} = 37.7 + 0.231 \text{ Chirps}$$

Predictor	Coef	SE Coef	T	Pr(> t)
Constant	37.67858	1.97817	19.05	7.35e-06 ***
Chirps	0.23067	0.01423	16.21	1.63e-05 ***
s = 1.52778 R-Sq = 98.1% R-Sq(adj) = 97.8%				

	Chirps
Mean	133.00000
Standard Dev.	43.84442
n	7.00000

$$SE = \frac{s_{\varepsilon}}{s_x \sqrt{n-1}} = \frac{\sqrt{MSE}}{s_x \sqrt{n-1}} = \frac{1.528}{43.84 \cdot \sqrt{7-1}} = 0.01423$$

Hypothesis Test for Correlation

How else can we measure the strength of association between two quantitative variables?

Recall: r = sample correlation, ρ = population correlation

$$H_0 : \rho = 0$$

$$H_a : \rho \neq 0$$

Find the p-value using a t-distribution with $n - 2$ df

$$\begin{aligned} t &= \frac{\text{statistic - null}}{SE} = \frac{r - 0}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \\ &= \left(\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \right) \end{aligned}$$

Hypothesis Test for Correlation

The correlation for the $n = 7$ cricket chirp data points is $r = 0.99062$. Compute the t-statistic for the test:

$$\begin{aligned}H_0 &: \rho = 0 \\H_a &: \rho \neq 0\end{aligned}$$

$$\begin{aligned}t &= \left(\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \right) \\&= \frac{0.99062\sqrt{7-2}}{\sqrt{1-0.99062^2}} = 16.21\end{aligned}$$

Coefficient of Determination, R^2

Recall that for correlation: $-1 \leq r \leq 1$

If we square the correlation, we get the **coefficient of determination**, which is a number between 0 and 1 that can be interpreted as a proportion or percentage.

$R^2 = \text{proportion of variability in the response variable, } Y, \text{ that is "explained" by the explanatory variable, } X.$

★ By convention we use a capital R^2 , although the value is just r^2 for a single explanatory variable.

Checking Condition

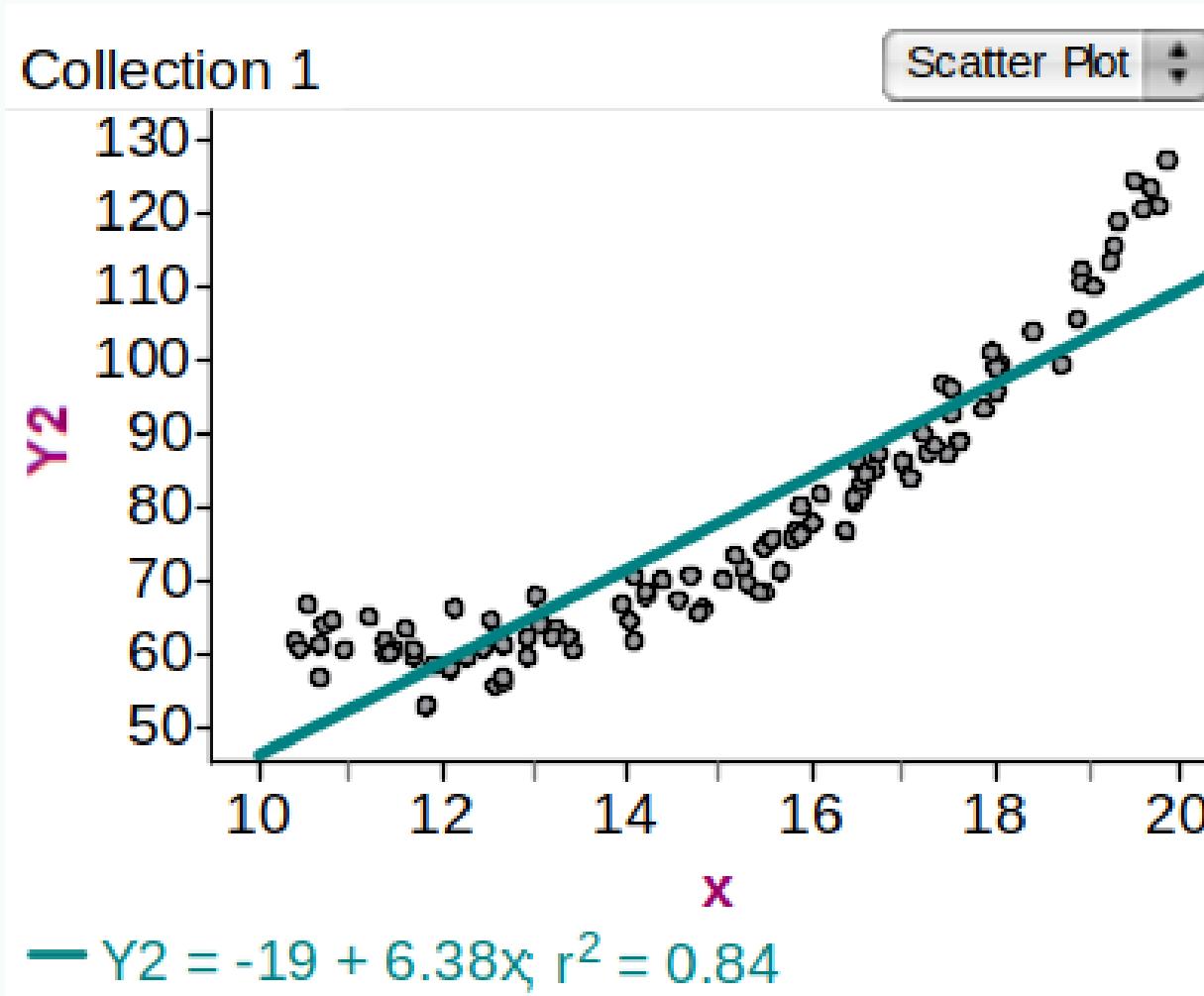
$$y = \beta_0 + \beta_1 x + \varepsilon$$

For a simple linear model, we assume the errors (ε) are randomly distributed above and below the line.

Quick check : Look at a scatterplot with regression line on it. Watch out for:

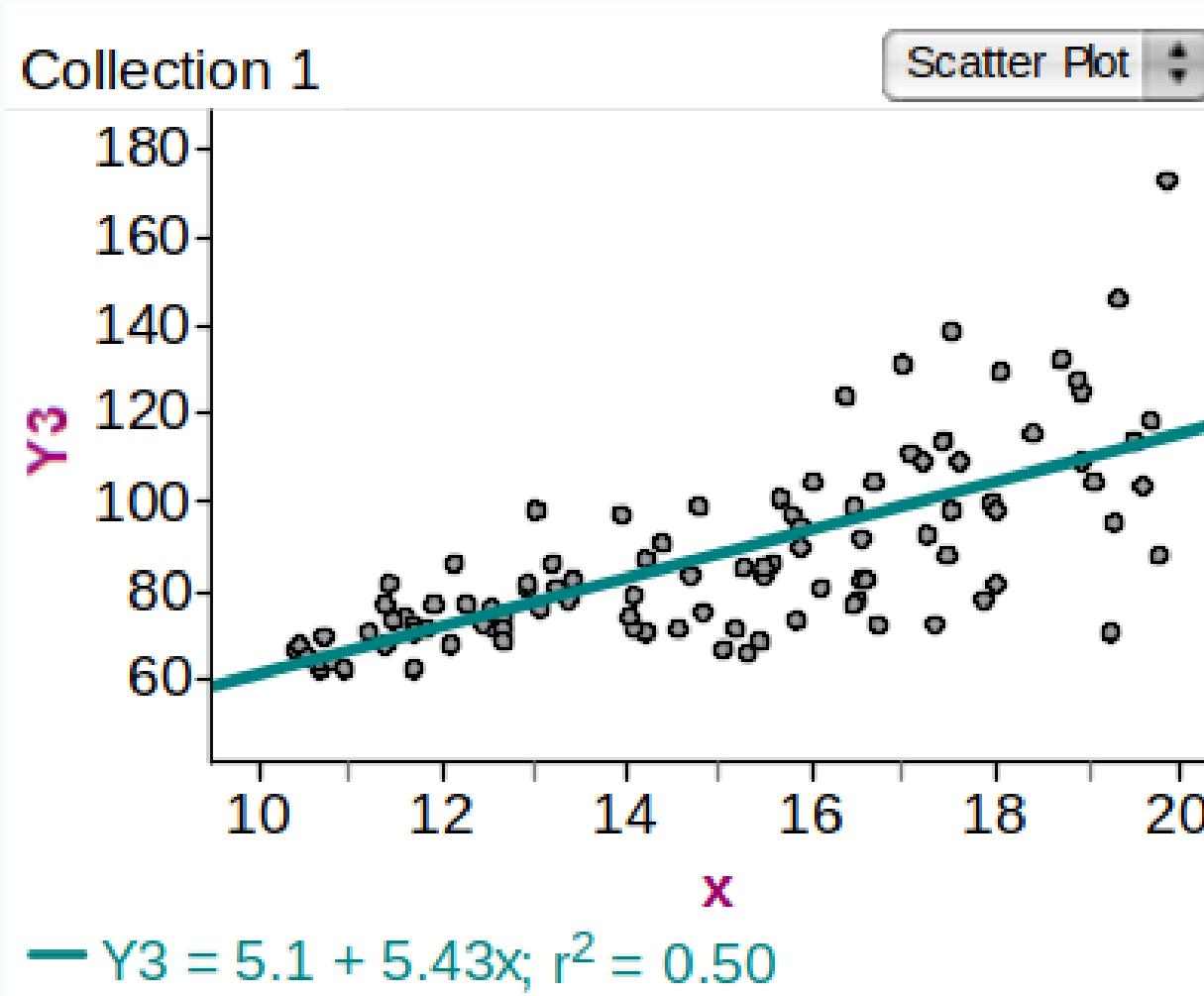
- ★ Curved (nonlinear) patterns in the data
- ★ Consistently changing variability
- ★ Outliers and influential points

Scatterplot: Check linearity and consistent variability



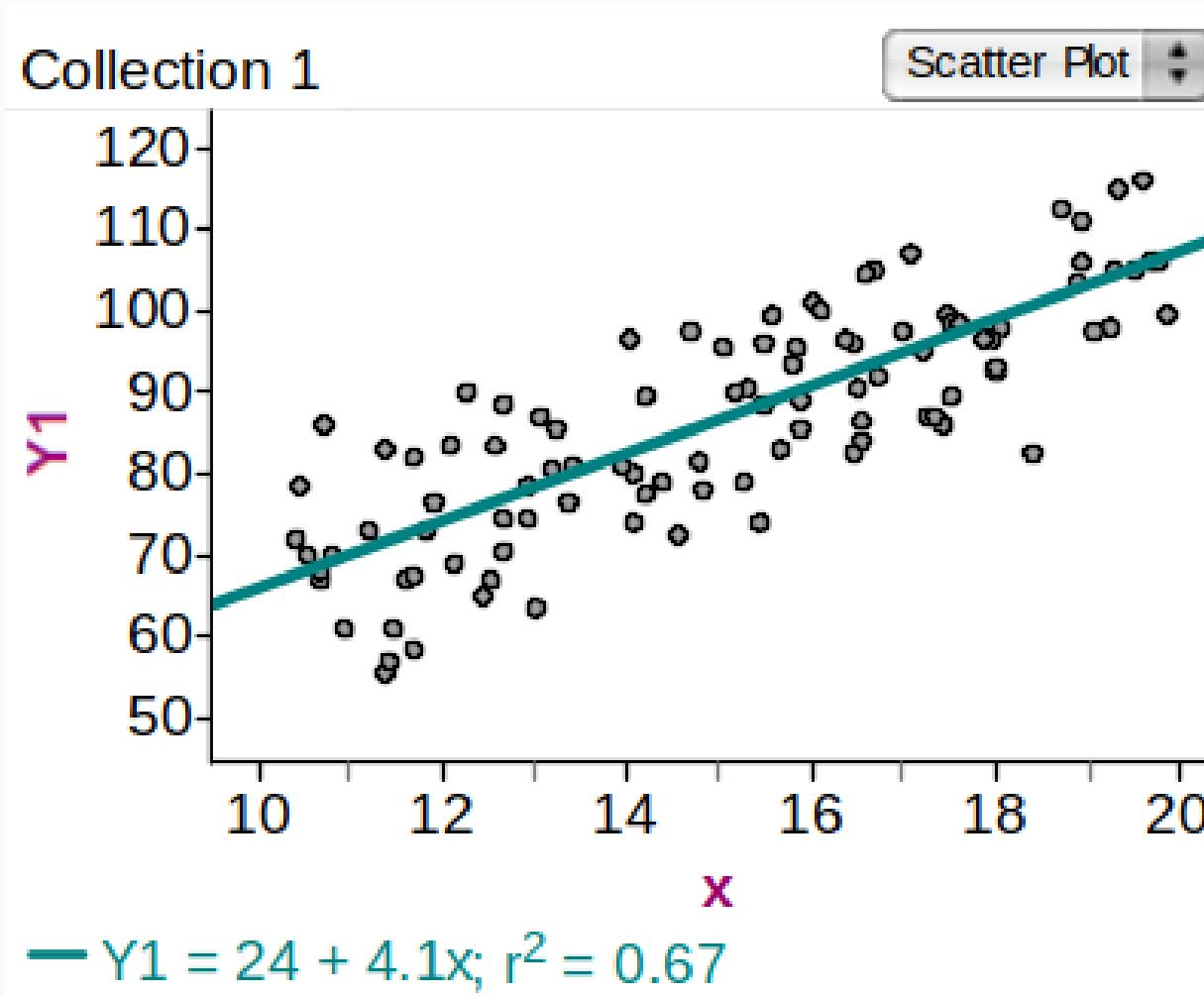
Problem

Scatterplot: Check linearity and consistent variability



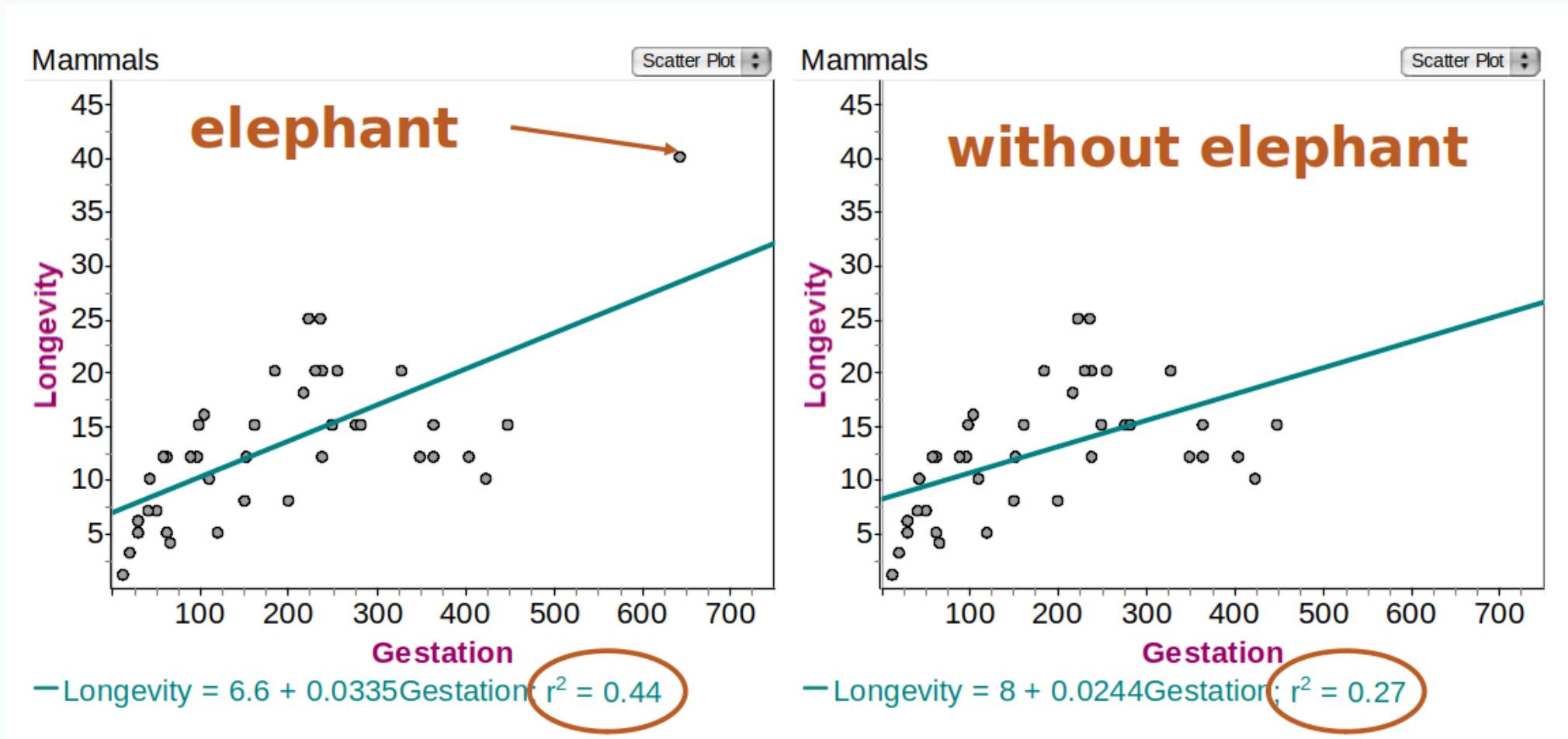
Problem

Scatterplot: Check linearity and consistent variability



Good

Scatterplot: Check for outliers or influential points



Partitioning Variability

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Data = Model + Error

Split the total variability in Y into two pieces, variability explained by the model + unexplained (residual error) variability

Total
Variability
in Y

=

Variability
Explained
by Model

+

Unexplained
Variability in
Error

Measuring Variability

$$\text{Total Variability in } Y = \text{Variability Explained by Model} + \text{Unexplained Variability in Error}$$

Total variability in Y : $SSTotal = \sum(y - \bar{y})^2$

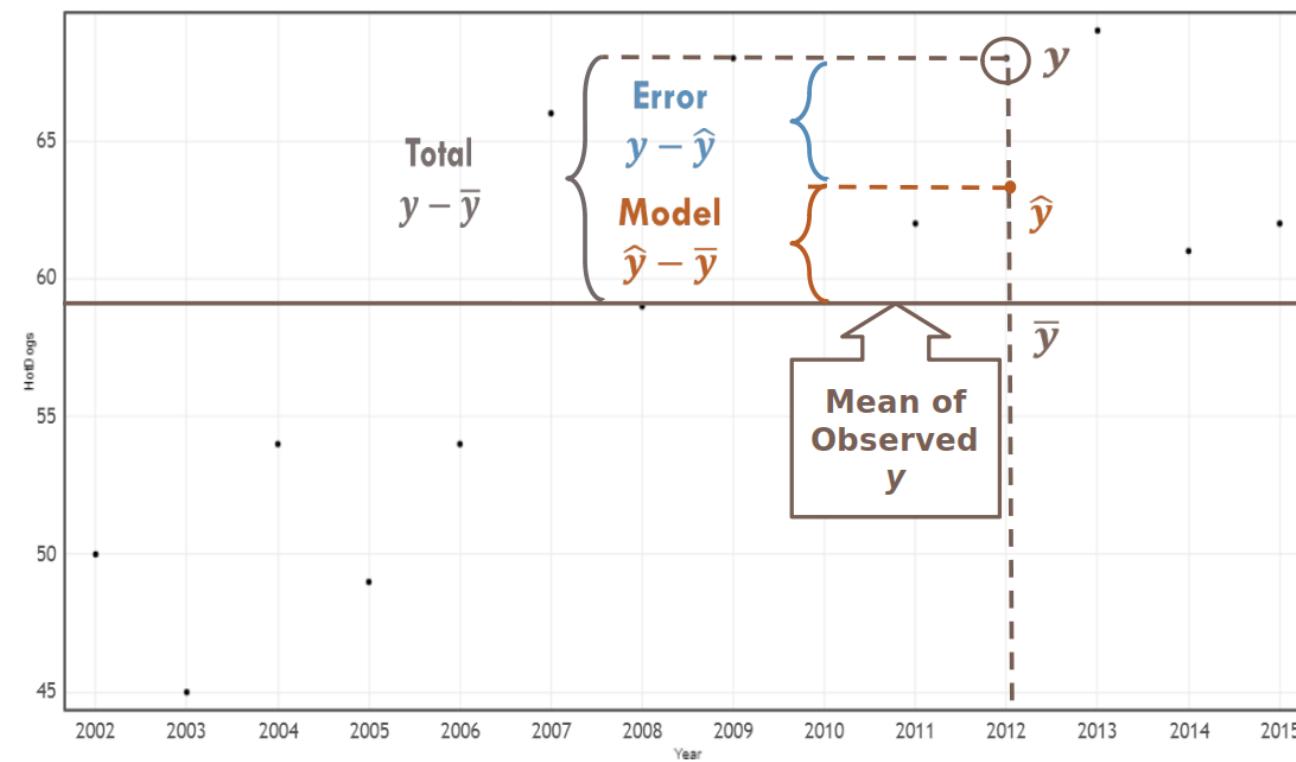
Explained variability: $SSModel = \sum(\hat{y} - \bar{y})^2$

Unexplained variability: $SSE = \sum(y - \hat{y})^2$

Graphically

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Data = Model + Error



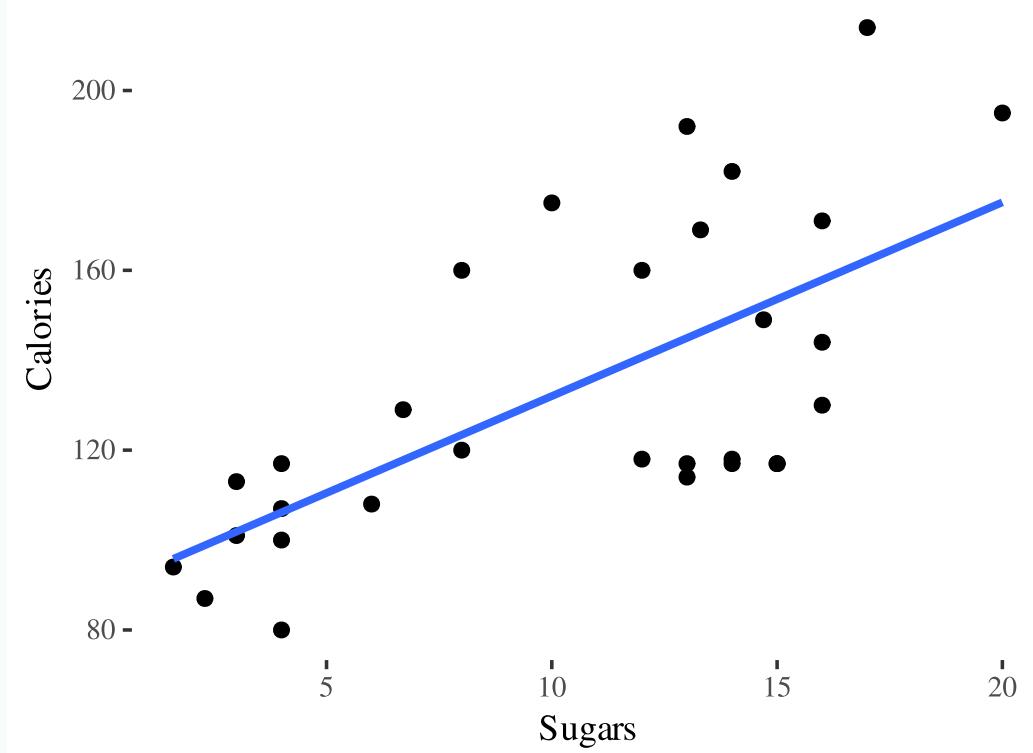
R: Calories Vs. Sugars

```
library(Lock5Data)
model <- lm(Calories~Sugars, data = Cereal)
anova(model)
```

Analysis of Variance Table

```
Response: Calories
  Df Sum Sq Mean Sq F value    Pr(>F)
Sugars     1 15316 15316.5  21.623 7.217e-05 ***
Residuals 28 19834   708.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' '
```

```
ggplot(Cereal, aes(x = Sugars, y = Calories)) +
  geom_point() +
  geom_smooth(method="lm", se = FALSE)
```



R^2 in Regression

```
library(Lock5Data)
model <- lm(Calories~Sugars, data = Cereal)
anova(model)
```

Analysis of Variance Table

Response: Calories

	Df	Sum Sq	Mean Sq	F value	Pr(>F)				
Sugars	1	15316	15316.5	21.623	7.217e-05 ***				
Residuals	28	19834	708.3						

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	.

$$R^2 = \frac{\text{Variability explained by Model}}{\text{Total variability in Y}}$$

$$R^2 = \frac{\text{SSModel}}{\text{SSTotal}}$$

$$R^2 = \frac{15316}{15316 + 19834}$$

$$R^2 = 0.4357$$

R: Calories Vs. Sugars

```
mod <- lm(Calories~Sugars, data = Cereal)
summary(mod)
```

Call:

```
lm(formula = Calories ~ Sugars, data = Cereal)
```

Residuals:

Min	1Q	Median	3Q	Max
-36.574	-25.282	-2.549	17.796	51.805

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	88.9204	10.8120	8.224	5.96e-09 ***
Sugars	4.3103	0.9269	4.650	7.22e-05 ***

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

Residual standard error: 26.61 on 28 degrees of freedom

Multiple R-squared: 0.4357, Adjusted R-squared: 0.4156

F-statistic: 21.62 on 1 and 28 DF, p-value: 7.217e-05

ANOVA for Regression

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

H_0 : The model is ineffective
 H_a : The model is effective

Source	df	Sum of Squares	Mean Square	F-statistic	p-value
Model	1	SSModel	$\frac{SSModel}{1}$	$F = \frac{MSModel}{MSE}$	$P(F_{1,n-2} > F)$
Error	$n - 2$	SSE	$\frac{SSE}{n-2}$		
Total	$n-1$	SSTotal			

P-value for Regression ANOVA

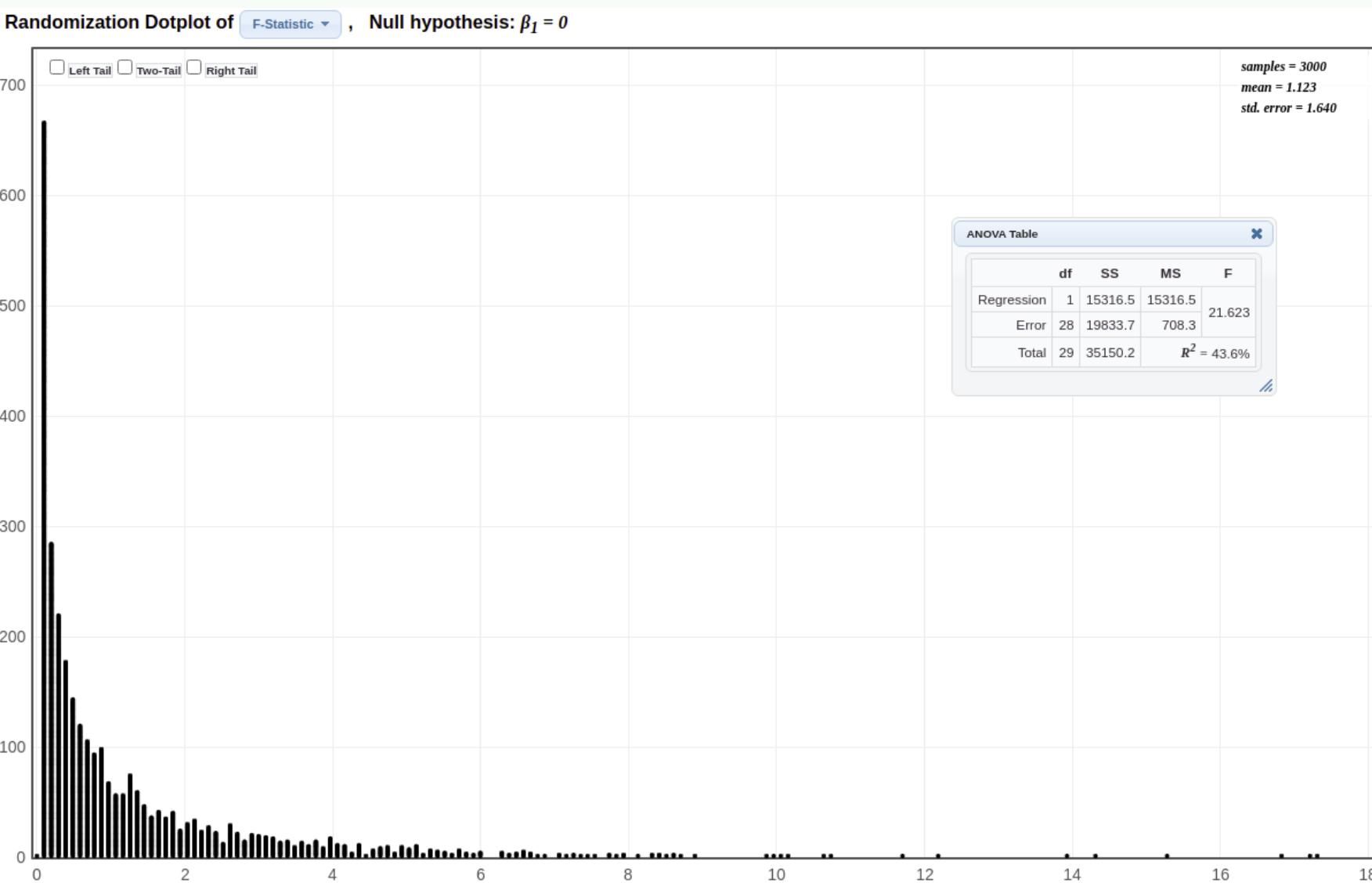
To find a p-value for the ANOVA F-statistic:

- ★ Create a randomization distribution, OR
- ★ Use a theoretical distribution

For a randomization distribution, we need to obtain samples where $H_0: \beta_1 = 0$ is true:

- ★ Randomly scramble the response (Y) values
- ★ Compute the ANOVA F-statistic for each sample

Randomization distribution



Theoretical Distribution: F-distribution

For testing an F-statistic (which is a ratio of variances)

★ Use an F-distribution, specifying the degrees of freedom for both the numerator and the denominator.

When performing an ANOVA for Regression with a single predictor:

- ★ 1 df (numerator)
- ★ $n - 2$ df (denominator)
- ★ Then use the upper tail beyond the F-statistic

Std. Dev. of Error in Regression

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Condition: The random errors (ε) have a common standard deviation, σ_ε

Estimation:

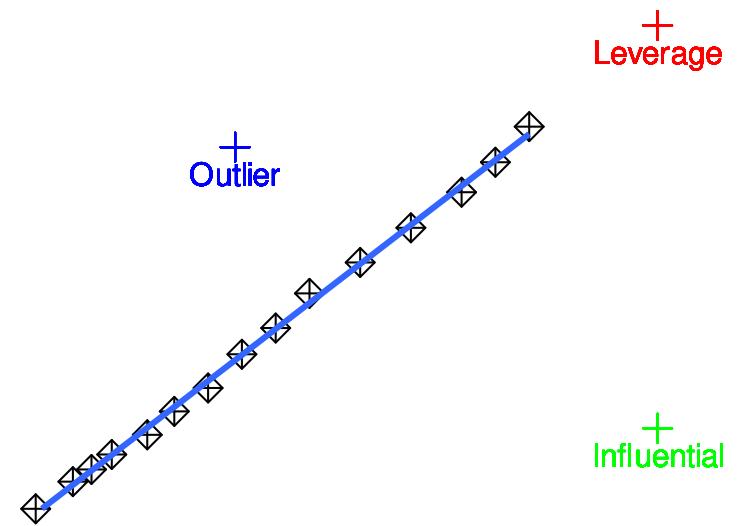
$$s_\varepsilon = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}} = \sqrt{\frac{SSE}{n - 2}} = \sqrt{MSE}$$

Identifying outliers in linear regression

Leverage points far away in the horizontal direction, do not change slope.

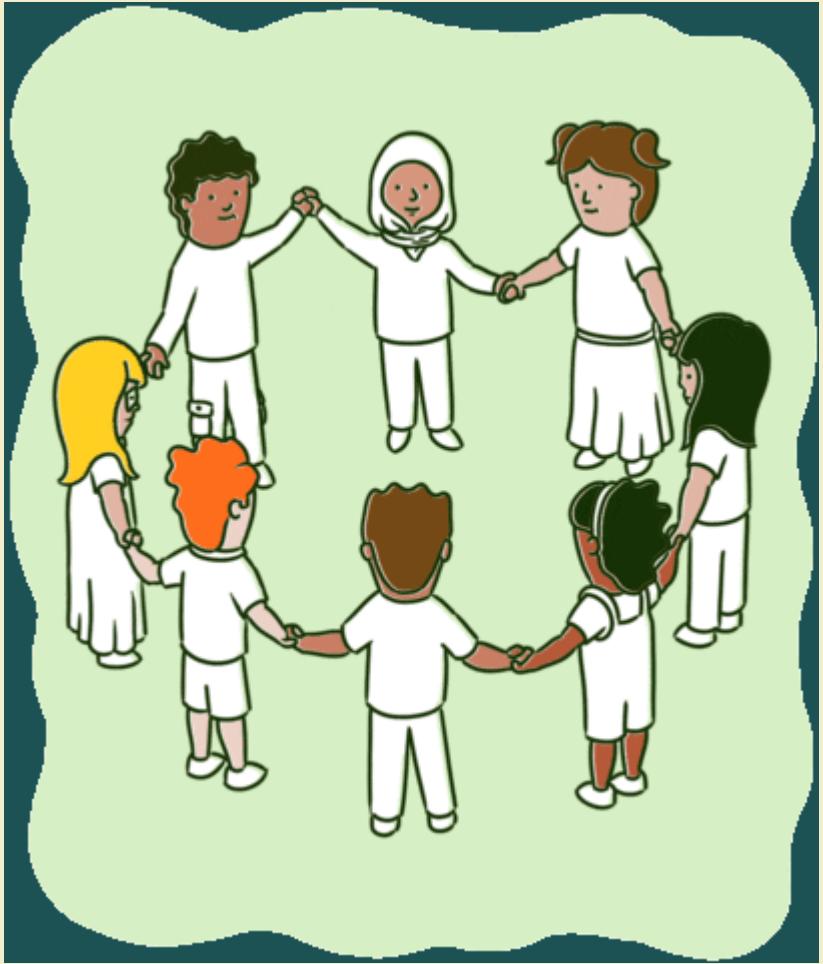
Influential points, slope changes with and without the influential point.

Outliers are points that are unusual and typically only affect the intercept of the data.



Your Turn 1

10:00



- ★ Go over to the in class activity file
- ★ Complete the remaining activity