Regression and ANOVA

Stat 120

May 26 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?

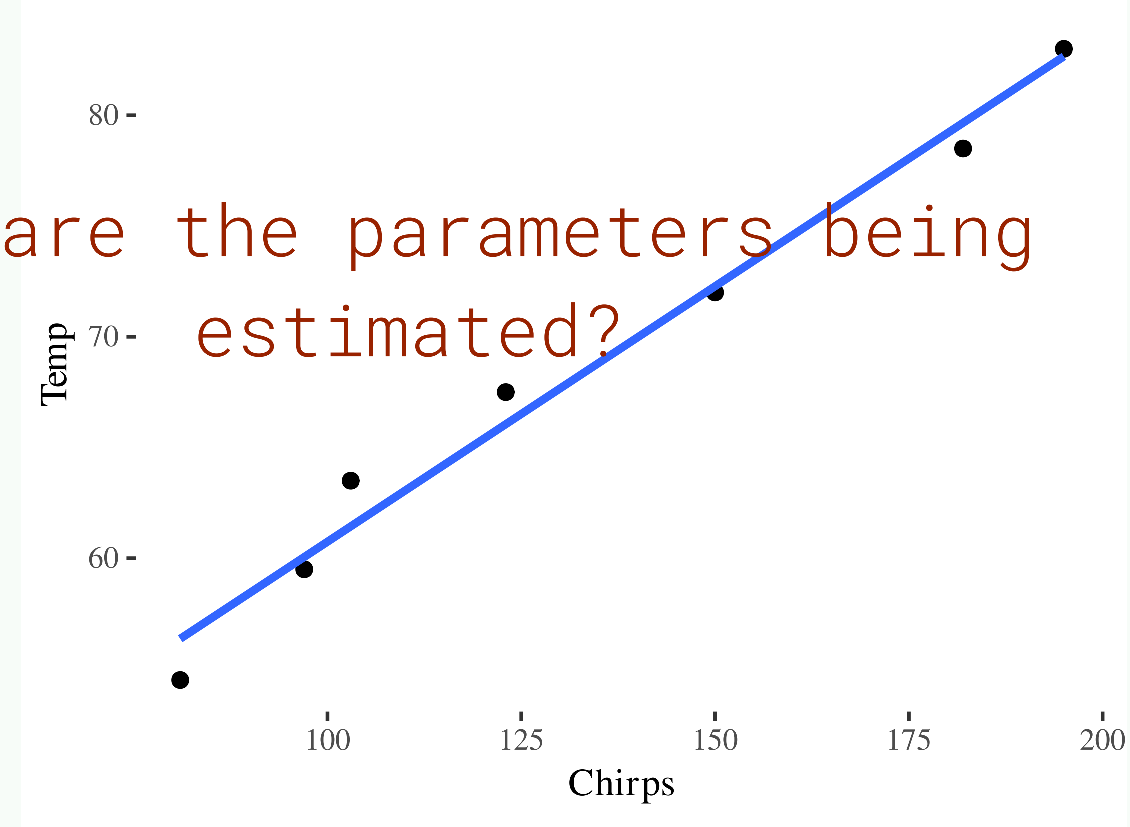
Recall: Least Square Regression

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:

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

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.

```
Temperature  =  37.7 + 0.231 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 ***
```

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

Hypothesis Test for Slope

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

$$\begin{aligned} H_0 : \beta_1 = 0 \quad &\Longrightarrow \text{No linear relationship} \\ H_a : \beta_1 \neq 0 \quad &\Longrightarrow \text{Some relationship} \end{aligned}$$

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

- **Again, b_1 and SE come from R output.**
- **We find the p-value by using a t distribution with $n-2$ df**

Hypothesis Test for Slope

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

```
Temperature  =  37.7 + 0.231 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 ***
```

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

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

Hypothesis Test for Correlation

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

Recall: r = sample correlation, ρ = population correlation

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

$$t = \frac{\text{statistic - null}}{SE} = \frac{r - 0}{\sqrt{\frac{1-r^2}{n-2}}} == \frac{r}{\frac{\sqrt{1-r^2}}{\sqrt{n-2}}}$$

$$= \left(\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right)$$

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

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 = proportion of variability in the response variable, Y, 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***

Partitioning Variability

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

$$\text{Data} = \text{Model} + \text{Error}$$

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

Total Variability in Y = **Variability Explained by Model** + **Unexplained Variability in Error**

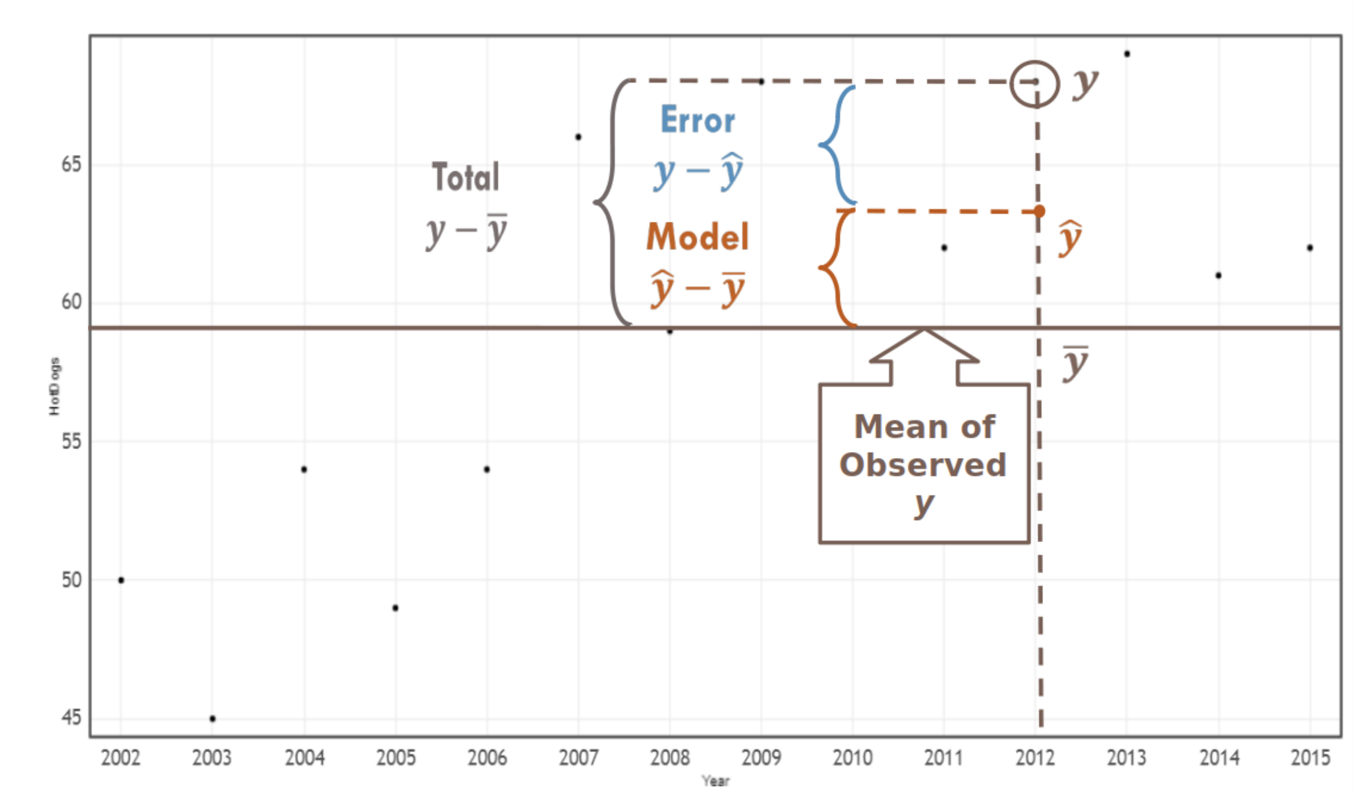
Total variability in Y: $SSTotal = \sum(y - \bar{y})^2$

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

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

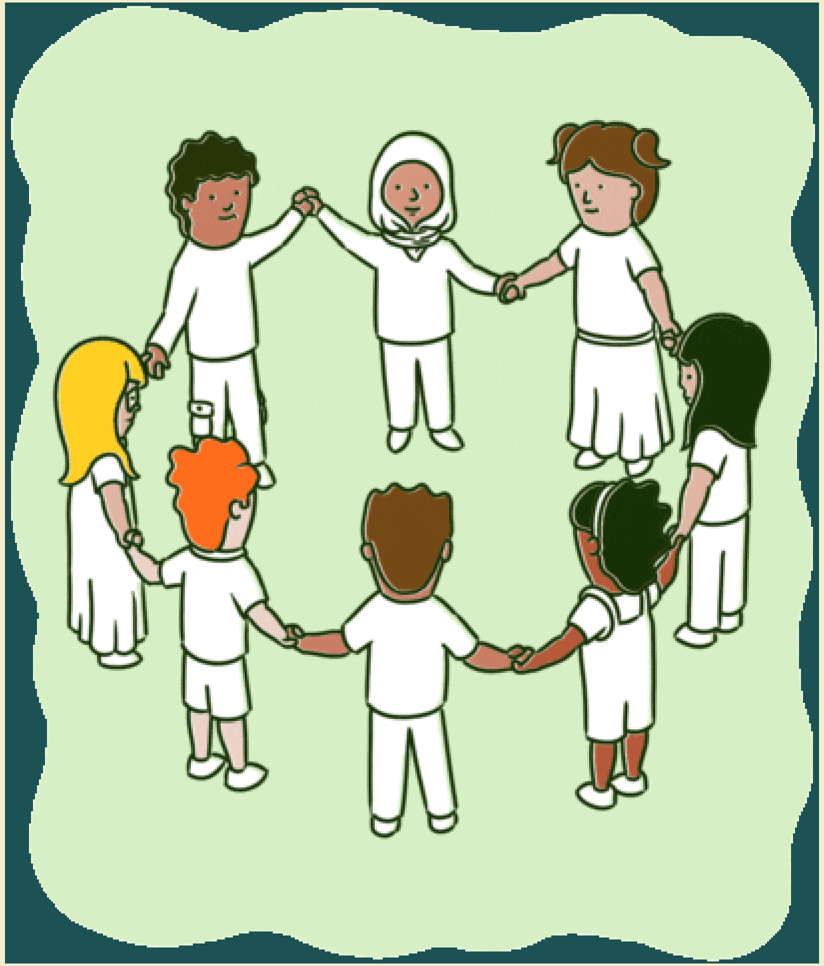
Graphically

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \text{Data} = \text{Model} + \text{Error}$$



YOUR TURN 1

05:00



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