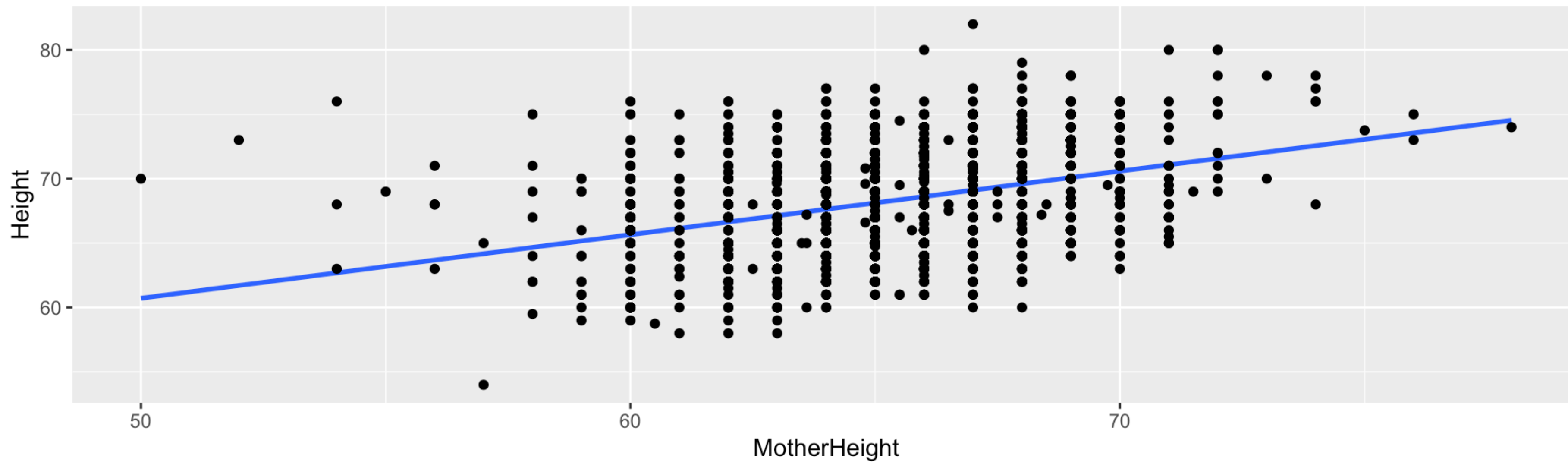
Simple Linear Regression - Inference

Research Objective

Research Question: Is the adult height of a student determined by the height of the mother? In other words, what is the relationship between a student's height and mother's height for all BYU students?

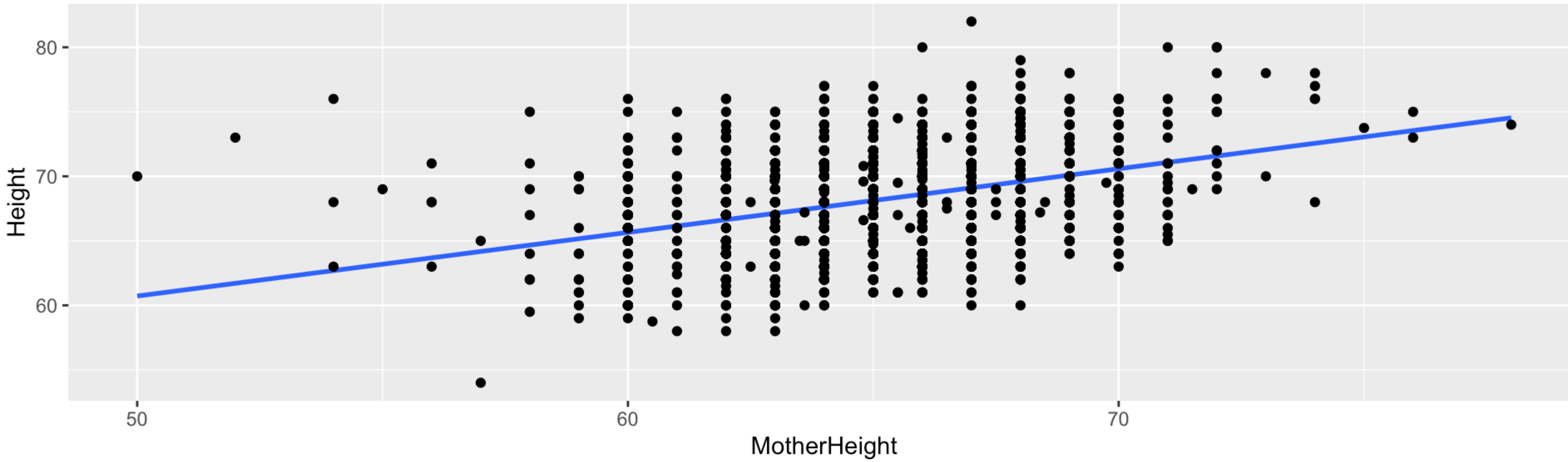


Our model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

Considering the research question, What would it mean if $\beta_1 = 0$?

- There is no relationship between mother's height (x) and student's height (y).

Population vs. Sample Slope



Our model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

Our fitted model: $\hat{y} = 36.059 + 0.493 \times x$

So, doesn't this mean that $\beta_1 \neq 0$ because $\hat{\beta}_1 = 0.493$?

- Not necessarily! $\beta_1 \neq \hat{\beta}_1$
- We need to do a test for β_1

Hypothesis Testing for β_1

Research Question: Does mother's height impact a child's height?

Steps of hypothesis testing:

1. Formulate null and alternative hypotheses.
2. Gather the data and see if our sample data matches (or doesn't match) the null hypothesis.
3. Draw a conclusion about H_0.

Hypothesis Testing for β_1 - Step 2

Step 2 - Compare our data result with what we expect to see if the null hypothesis is true.

From our sample, we have $\hat{\beta}_1 = 0.493$ is this “different enough” from 0 to conclude that $H_a : \beta_1 \neq 0$?

Hypothesis Testing for β_1 - Step 2

Step 2 - Compare our data result with what we expect to see if the null hypothesis is true.

From our sample, we have $\hat{\beta}_1 = 0.493$ is this "different enough" from 0 to conclude that $H_a : \beta_1 \neq 0$?

First, standardize using the formula (or let the computer do this for you):

$$t = \frac{\hat{\beta}_1 - \overbrace{\beta_1}^{0}}{\frac{\hat{\sigma}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 14.744$$

Interpret t as the number of standard errors our $\hat{\beta}_1$ is from the hypothesized β_1.

Hypothesis Testing for β_1 - Step 2

Theorem. Sampling Distribution of beta_1

If the LINE assumptions of the regression model are appropriate, then

$$t = \frac{\hat{\beta}_1 - \overbrace{\beta_1}^{0}}{\frac{\hat{\sigma}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

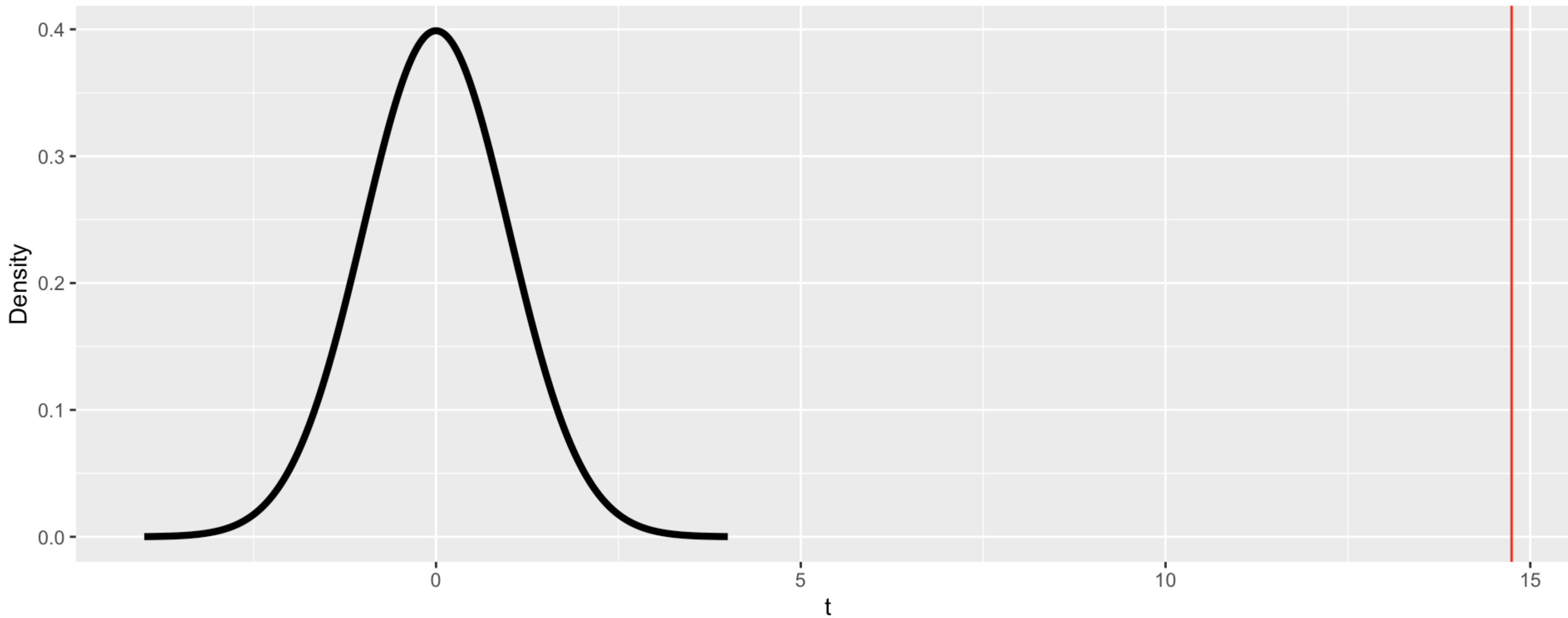
is a standardized statistic and follows t distribution with center 0 and spread 1 and degrees of freedom $n - 2$.

Note, above we would set $\beta_1 = 0$ because we assume H_0 is true unless proven otherwise.

- So…what does this mean?

Hypothesis Testing for β_1 - Step 2

IF the LINE assumptions holds, the talues of t that are consistent with the claim $H_0 : \beta_1 = 0$ are given by the distribution (curve):



- But we are getting ahead of ourselves because the LINE assumptions have to be true for the above picture to be correct.

Checking LINE Assumptions

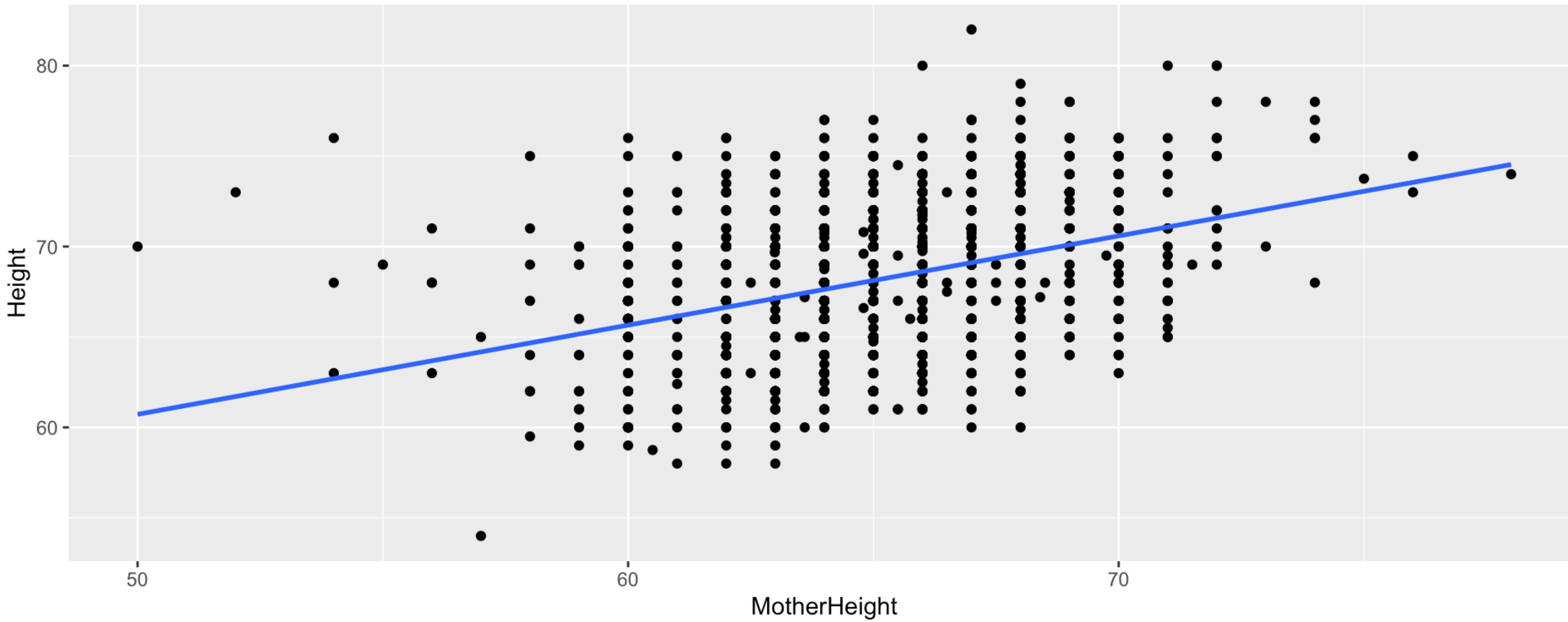
Reminder, the LINE assumptions are:

- L - Linear relationship between x and y
- I - Independence (one obs. doesn't impact the other)
- N - Normal residuals (distance from line is normal)
- E - Equal variance of residuals (spread about the line is constant)

How would we see if there is a linear relationship between x and y?

- Scatterplot!

Checking LINE Assumptions



Is this (approximately) linear for the bulk of the data?

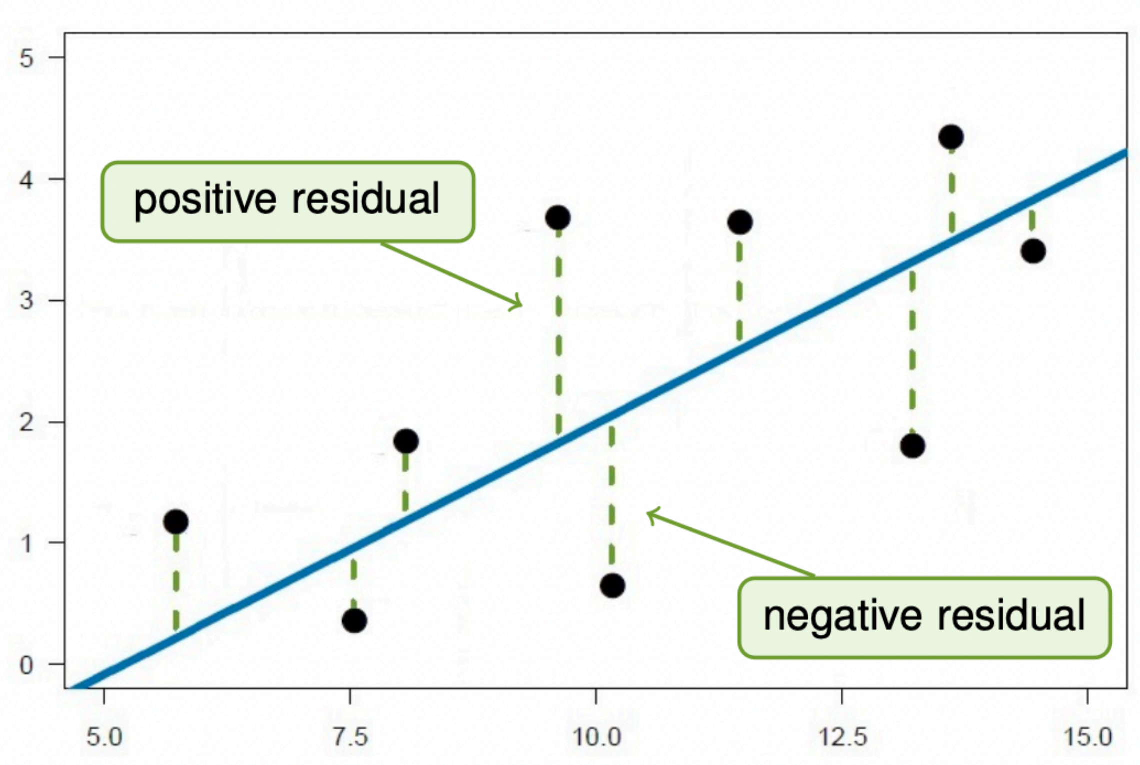
Checking LINE Assumptions

How would we see if there is independence? In other words, how can we “check” if one observation doesn’t influence another?

- Critical Thinking!
- Does it make sense that one student’s height would determine another student’s height?
- Likely a minimal influence.

Checking LINE Assumptions

How would we see if the residuals are normal?

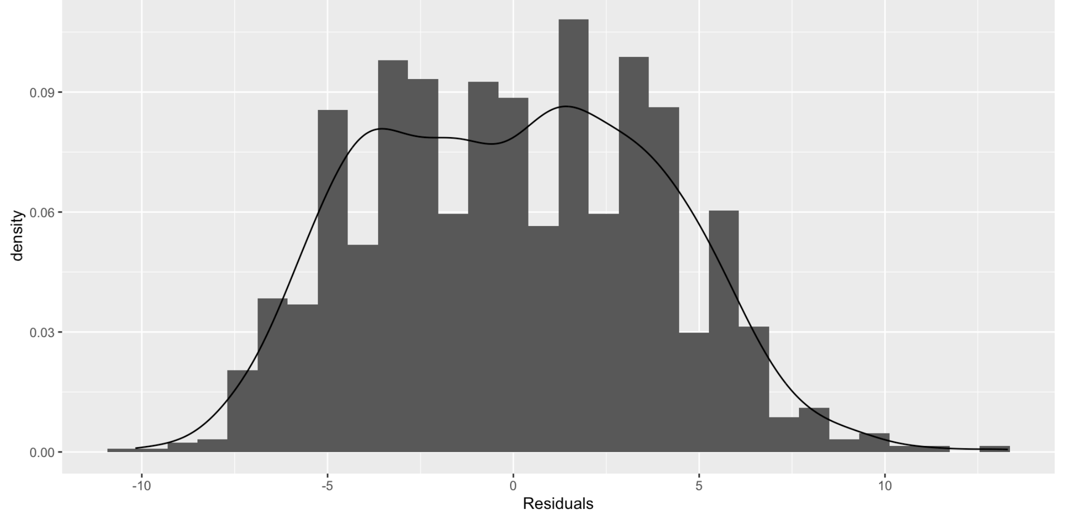


1. Calculate the residuals as $\epsilon_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ (don't worry - the computer will do this for you)
2. Draw a histogram (or density plot) of residuals

Checking LINE Assumptions

How would we see if the residuals are normal?

1. Calculate the residuals as $\epsilon_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ (don't worry - the computer will do this for you)
2. Draw a histogram (or density plot) of residuals



Is this approximately normal?

- Close enough. Skew = 0.1195708

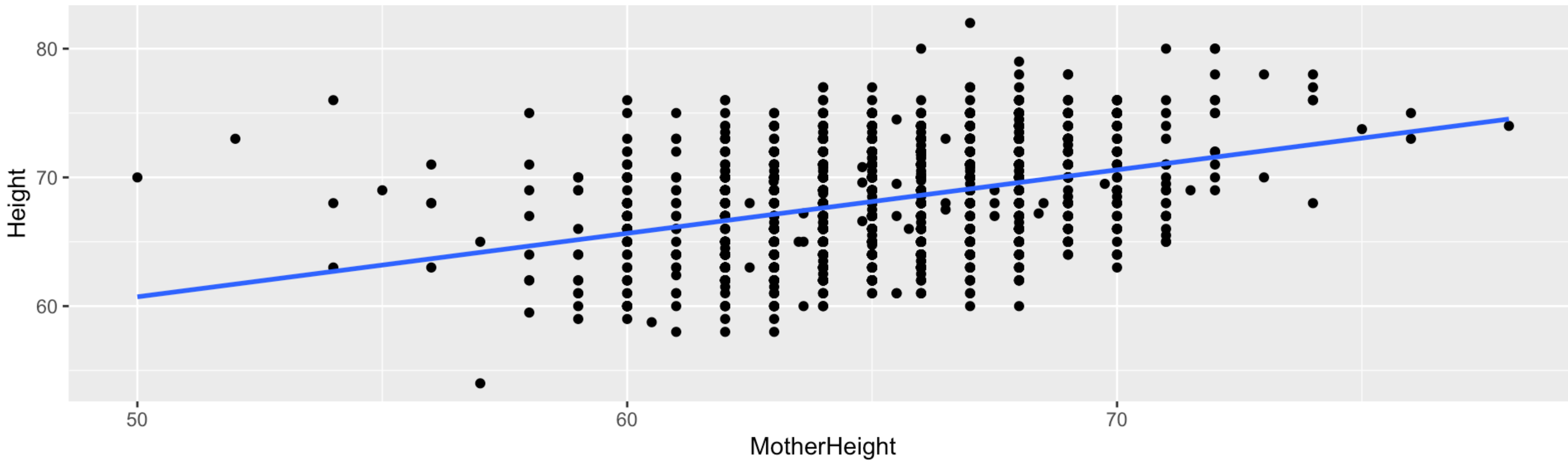
Checking LINE Assumptions

How would we see if there is “equal spread” of the residuals about the fitted line?

Checking LINE Assumptions

How would we see if there is “equal spread” of the residuals about the fitted line?

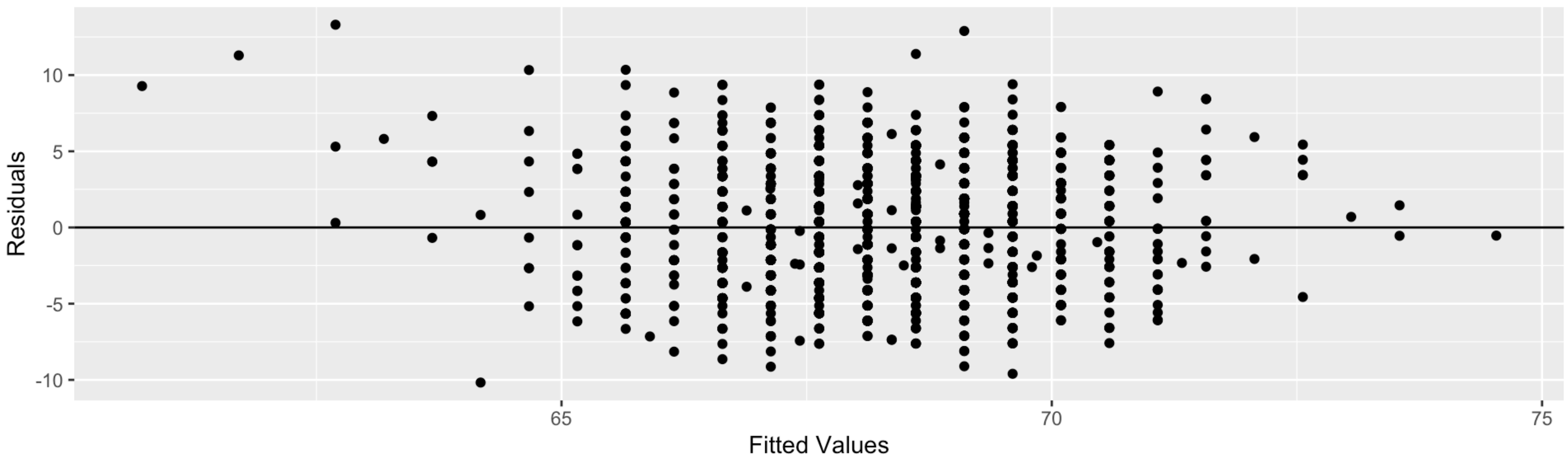
- Option 1: Scatterplot with fitted line



Checking LINE Assumptions

How would we see if there is “equal spread” of the residuals about the fitted line?

- Option 2: Fitted values vs. residuals plot (just like a scatterplot with fitted line but made to be easier to see visually)

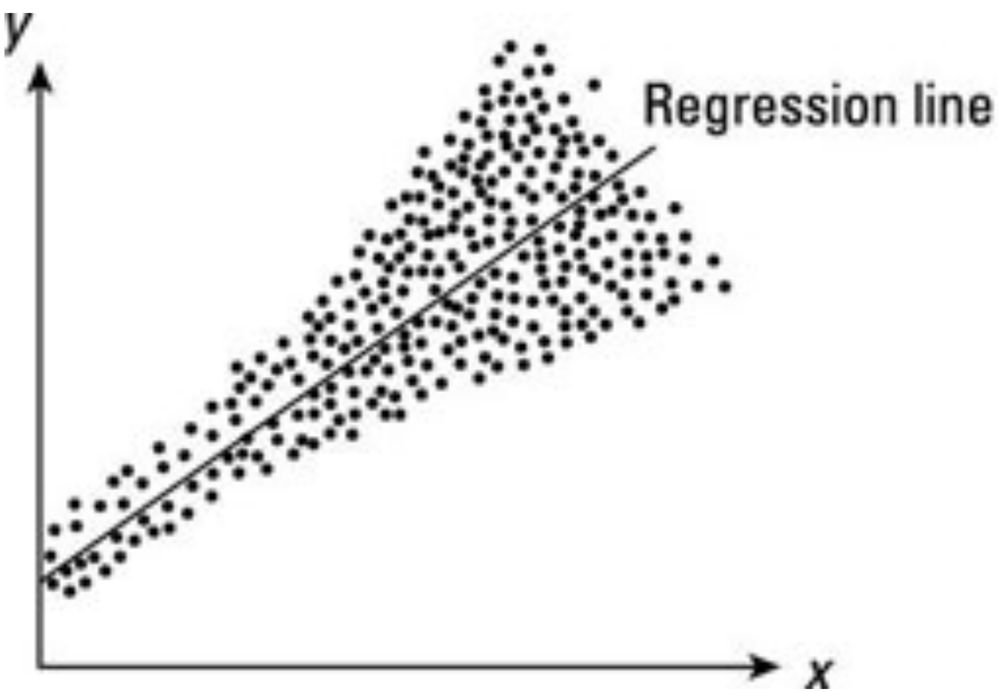


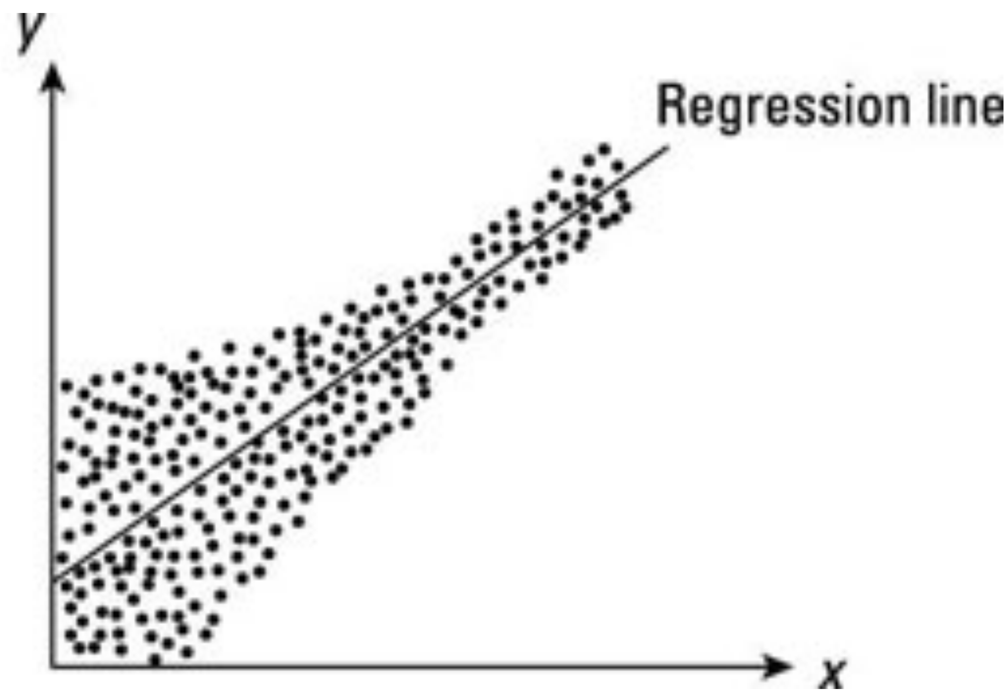
Is this roughly “equal spread”?

- Close enough except for 1 or 2 outliers

Checking LINE Assumptions

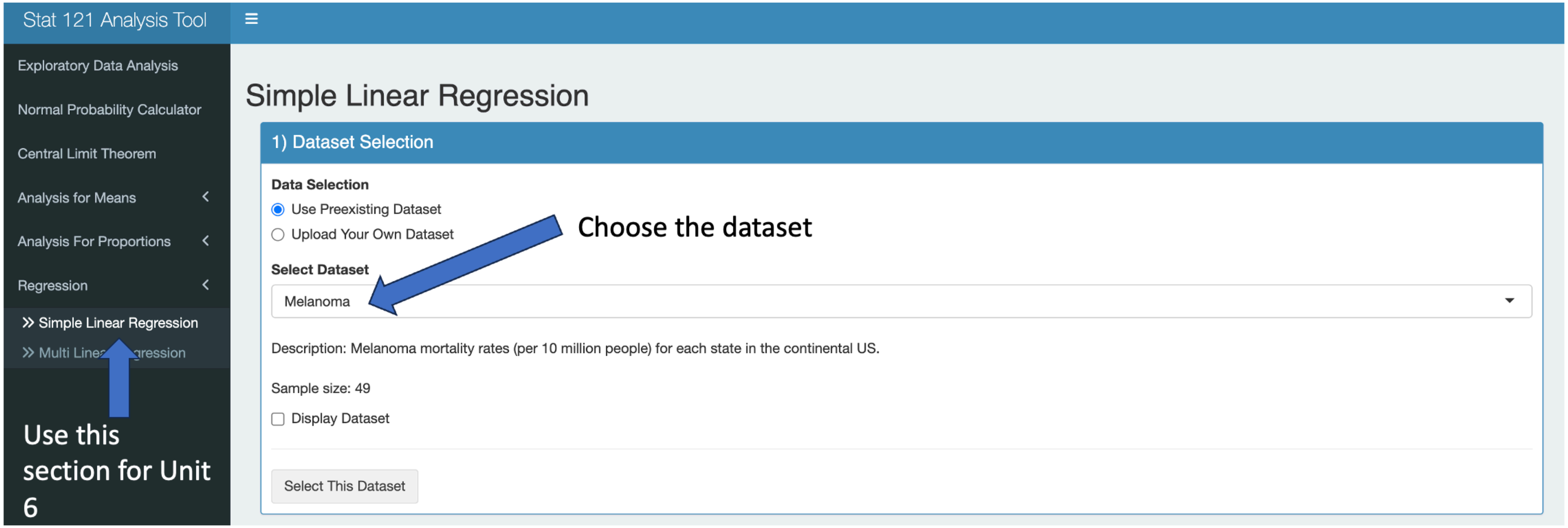
Examples of NOT equal spread



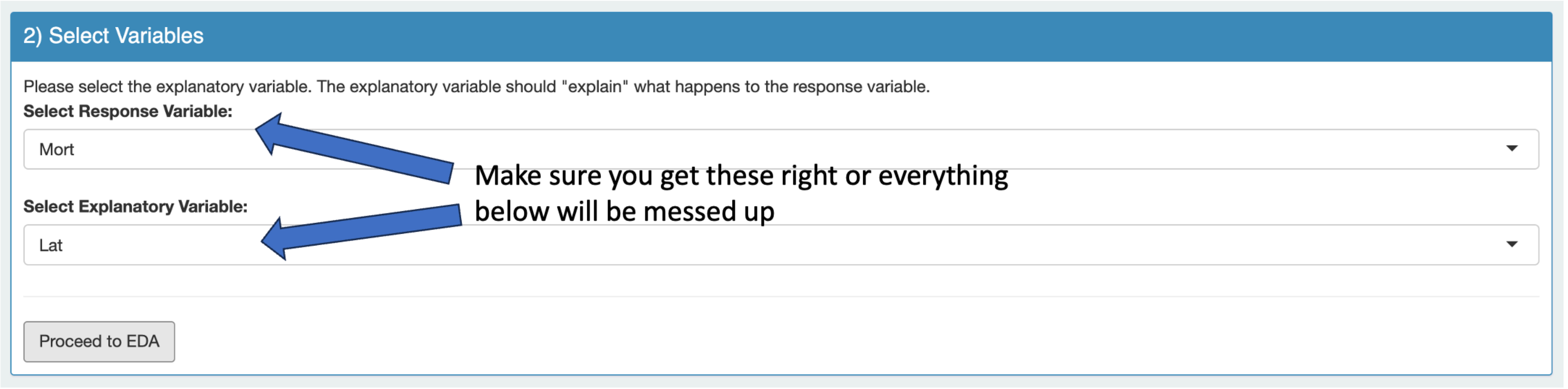


Using the Analysis Tool

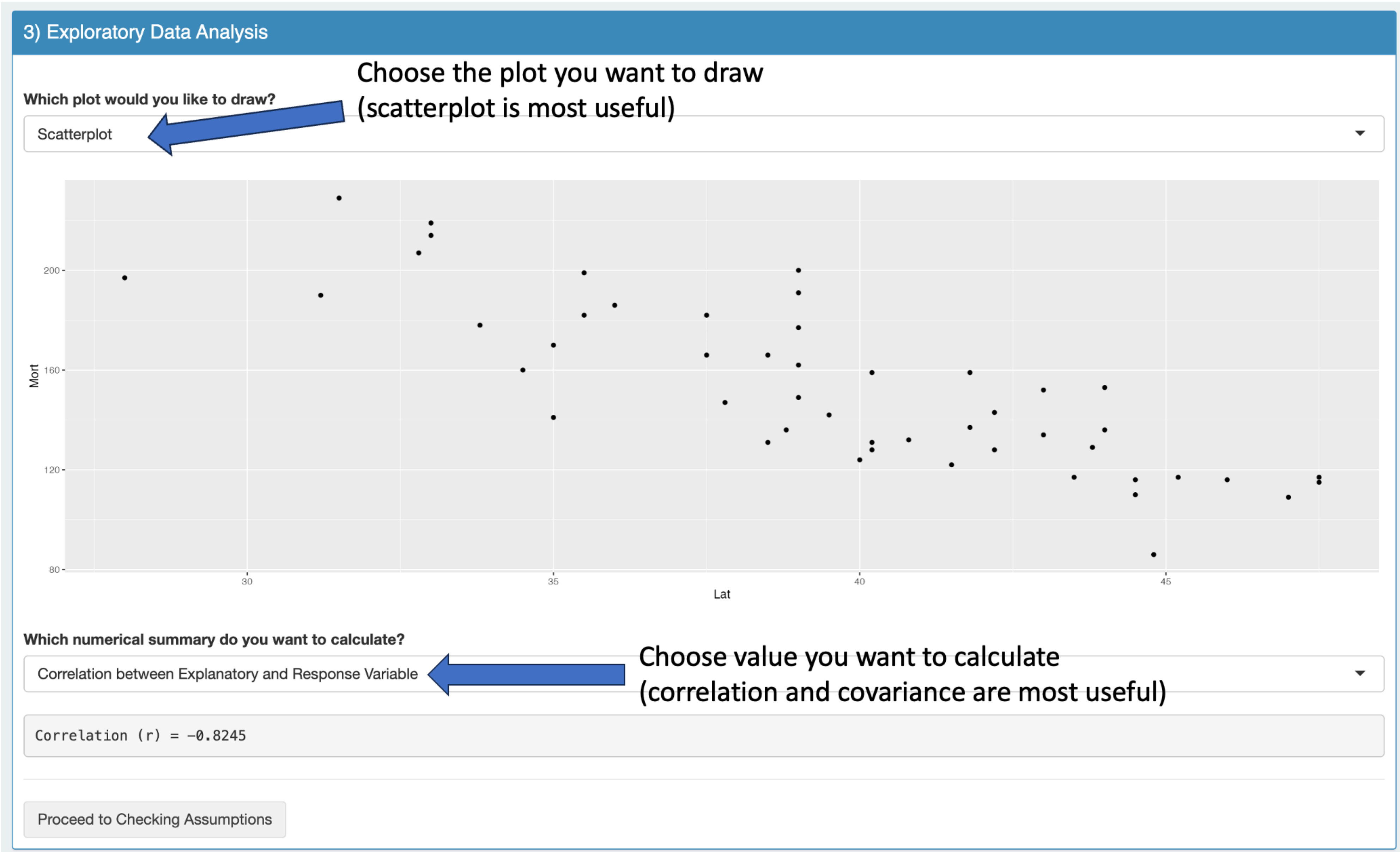
Melanoma is highly related to sun exposure. Hence, areas with greater sun have a greater risk of melanoma.



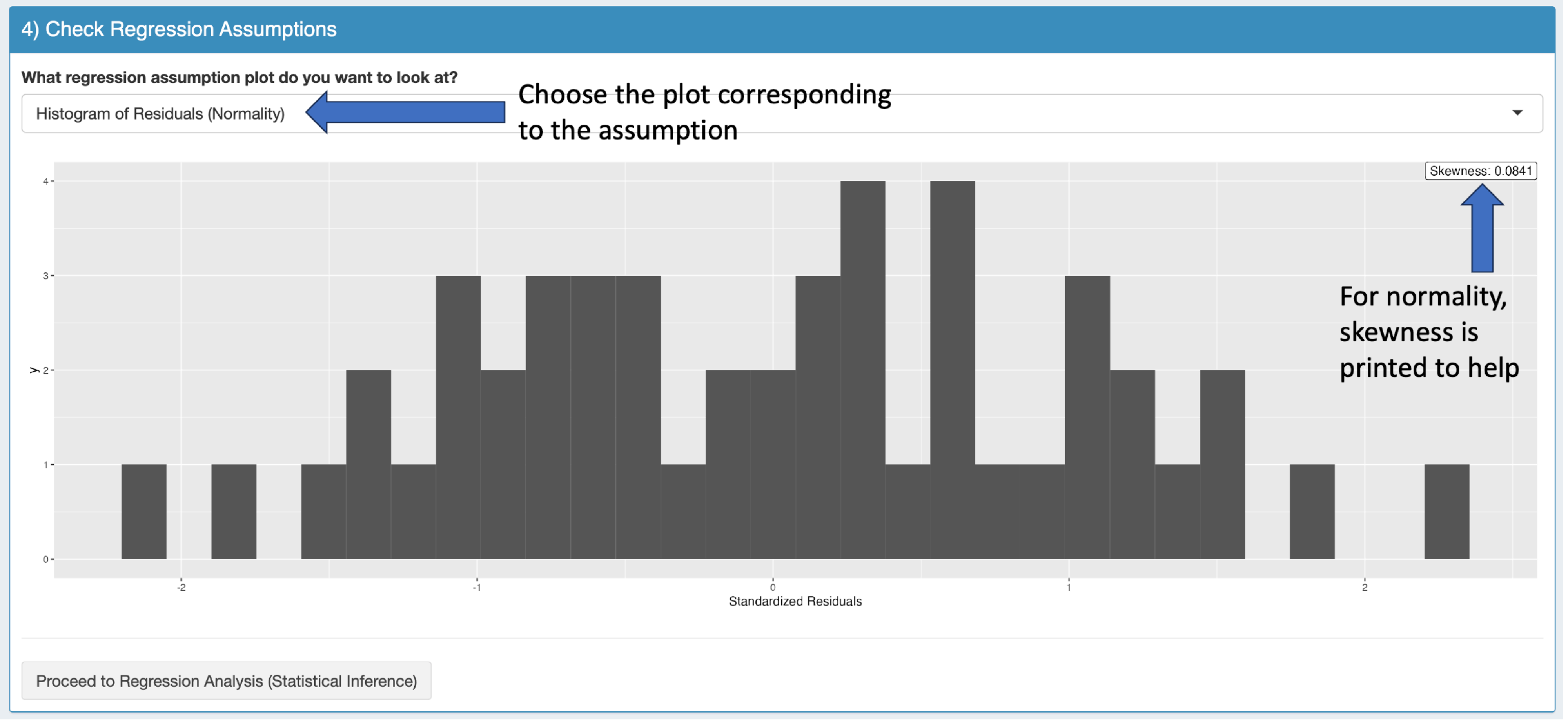
Using the Analysis Tool



Using the Analysis Tool



Using the Analysis Tool



Hypothesis Testing for β_1

Back to Step 2 - - gather the data and see if our sample data matches (or doesn't match) the null hypothesis (note: do this only if LINE assumptions are valid)

Measuring if our data is consistent with the null hypothesis:

1. **Standardized test statistic**: the number of standard errors away from the hypothesized value our data is. In our rent example $t =$ 14.7438605.
2. p-**value**: probability of observing our sample result or "more extreme" (as stated by H_a) if the null hypothesis is true. Our p-value is 0.

Step 3: Draw a conclusions about $H_0 : \beta_1 = 0$. Using $\alpha = 0.05$, what do we conclude about β_1?

- Our data is NOT consistent with the null hypothesis so we conclude that the mother's height does have an effect on the student's height.

Using the Analysis Tool

5) Regression Analysis

Confidence Level for Slope and Intercept:

0.5 — 0.95 — 0.99

0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 0.99

```
Regression Analysis of Mort (Y) explained by Lat (X)
Coefficient Table:
```

Show 5 **entries**

Focus on this line because it gives information about the slope

	Estimate	t value	p-value	CI Lower Bound	CI Upper Bound
(Intercept)	389.1894	16.344	0	341.2852	437.0936
Lat	-5.9776	-9.9898	0	-7.1814	-4.7739

Slope ($\hat{\beta}_1$)

t statistic for the slope

p-value for the slope

Showing 1 to 2 of 2 entries

Previous 1 Next

```
R-squared:  0.6798
sigma:  19.115
```

☐ Show Fitted Regression Line

Proceed to Predictions

Vagueness of Hypothesis Tests

If we reject $H_0 : \beta_1 = 0$ and conclude $H_A : \beta_1 \neq 0$ then we really haven't concluded anything other than there is an effect.

- Use a confidence interval for more informative answers.

Confidence Intervals for β_1

Using the same ideas for building a confidence interval as before, a C% confidence interval for β_1 is:

$$\hat{\beta}_1 \pm t^{\star} \frac{\hat{\sigma}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

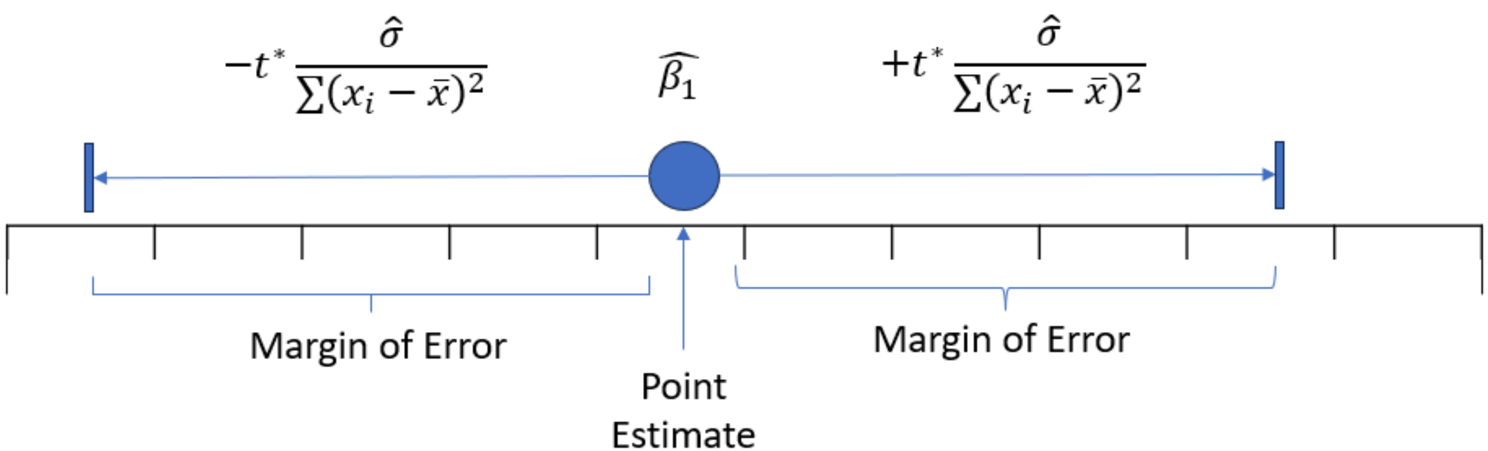
- Don't worry about the formula, the computer will calculate it for you.

Confidence Intervals for β_1

Research Question: As the mother's height increases, what happens to the child's height?

Answer:

- A 95% confidence interval for β_1 is calculated as (0.428,0.559).
- How do we interpret this interval?
  - We are 95% confident that as the mother's height goes up by 1 inch, we **expect** the student's height to go up between (0.428,0.559) inches.
  - Notice, that the interpretation says **expect** NOT will.



Using CIs to do Tests

Research Question: If the mother's height goes up by 1 inch, can we expect the student's height to change by 1in?

Answer:

- A 95% confidence interval for β_1 is calculated as (0.428,0.559).
- No because 1 is not in the interval at the 0.05 significance level.
- Principle: You can use CIs to do 2-sided hypothesis tests (i.e. alternative hypothesis with "$\neq$")

Using the Analysis Tool

5) Regression Analysis

Confidence Level for Slope and Intercept:

0.5 — 0.95 — 0.99

0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 0.99

Set the confidence level

```
Regression Analysis of Mort (Y) explained by Lat (X)
Coefficient Table:
```

Show 5 **entries**

Focus on this line because it gives information about the slope

	Estimate	t value	p-value	CI Lower Bound	CI Upper Bound
(Intercept)	389.1894	16.344	0	341.2852	437.0936
Lat	-5.9776	-9.9898	0	-7.1814	-4.7739

Showing 1 to 2 of 2 entries

Previous 1 Next

Interval bounds

```
R-squared:  0.6798
sigma:  19.115
```

☐ Show Fitted Regression Line

Proceed to Predictions

Nuances of Inference for β_1

What do we do if the LINE assumptions aren't quite appropriate?

- Throw out outliers (not recommended)
- Ignore them and do inference anyway (but acknowledge that your inferences could be very wrong - not recommended)
- Use more explanatory variables. For example, use father's height AND shoe size to explain height (we'll learn this next unit).
- Consult a statistician (or better yet - take more stats classes and we'll teach you)

Additional Practice:

Measuring possum head size can be difficult. However, measuring total possum length is easier. What is the relationship between possum length and head size? Use a simple linear regression model (and the course app) to answer the following questions:

1. Do the LINE assumptions all hold for this example?
   - Yes
2. Does total length have a linear effect on head length?
   - Yes because the test on the slope rejects at the $\alpha = 0.05$ level.
3. What would a Type 1 Error be for the hypothesis test in #1?
   - Saying there is a relationship between total and head length when there isn't.
4. If the total length goes up by 1, how much do we expect the head length to change?
   - We are 90% confident that head length will go up between (0.6904, 0.977)

Key Terminology

- LINE Assumptions
- Hypothesis tests for β_1
- Confidence intervals for β_1
- Checking LINE assumptions