Simple Linear Regression - EDA

Research Objective

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

Population: All BYU students.

Parameter of Interest:

- Some number measuring the "relationship" between students height and the mother's height.

Sample: A convenience sample of 1575 BYU students who are in Stat 121.

Are there any issues with this study setup?

More Problem Definitions

Response Variable (y): The height of the student.

- This is a **continuous quantitative variable** meaning it can be any number (including decimals)

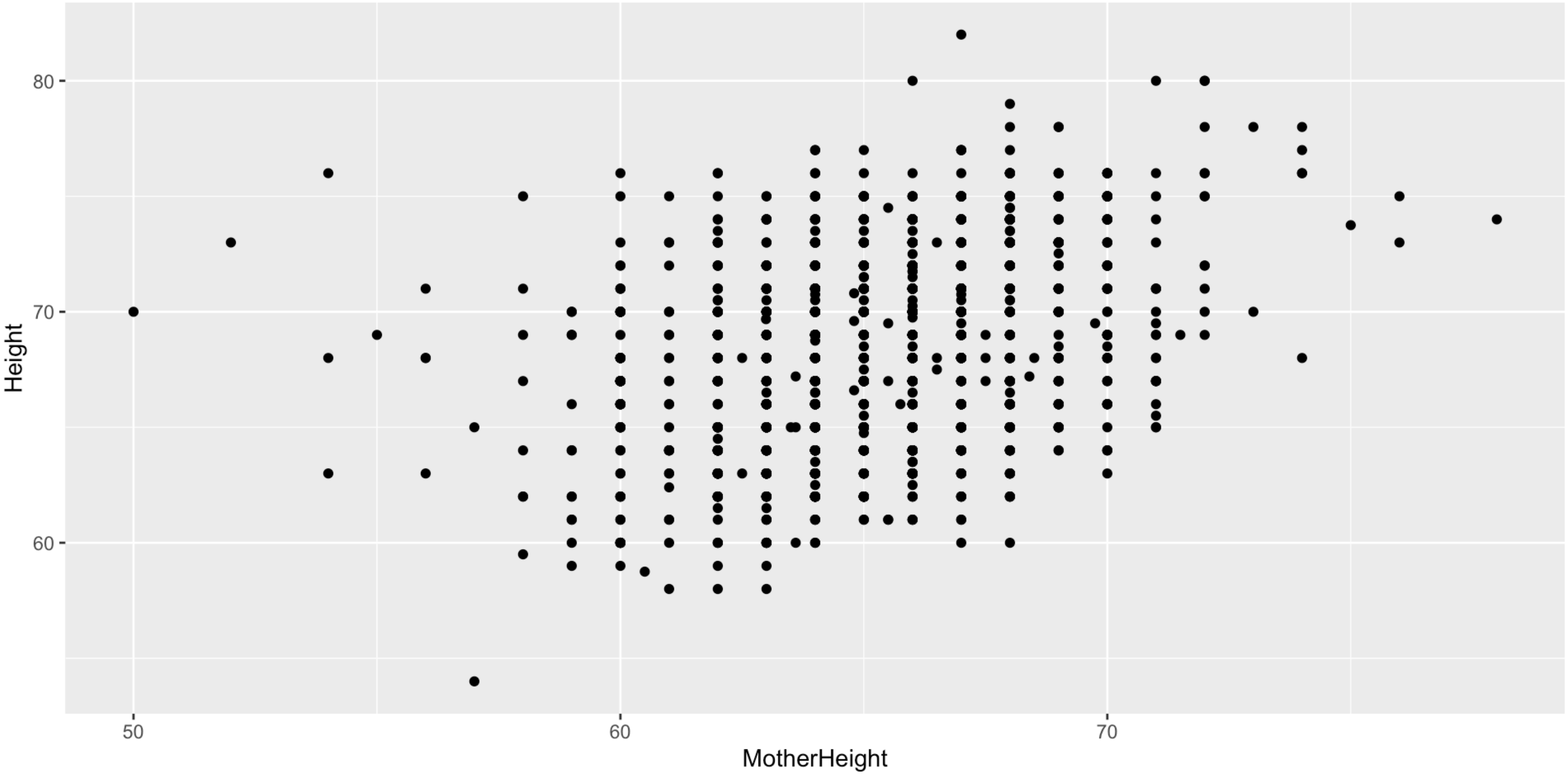
Explanatory Variable (x): The height of the mother.

- This is also **continuous quantitative variable**.

Exploratory Data Analysis (EDA)

Main goal: Investigate the relationship between student’s height and mother’s height.

Tool #1 - Scatterplots

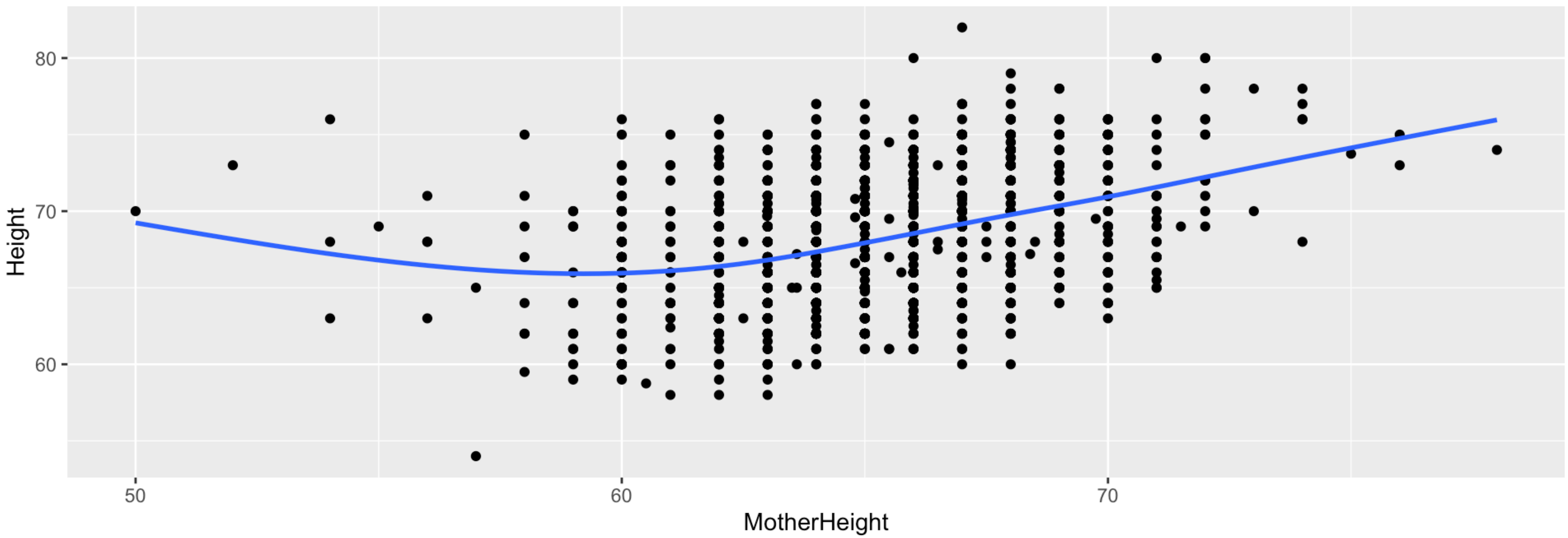


Tool #1 - Scatterplots

Things to look for in a scatterplot:

- Form: linear, non-linear or nothing
- Direction: positive or negative
- Strength: amount of “scatter” about the trend-line
- Outliers (data points out by themselves)

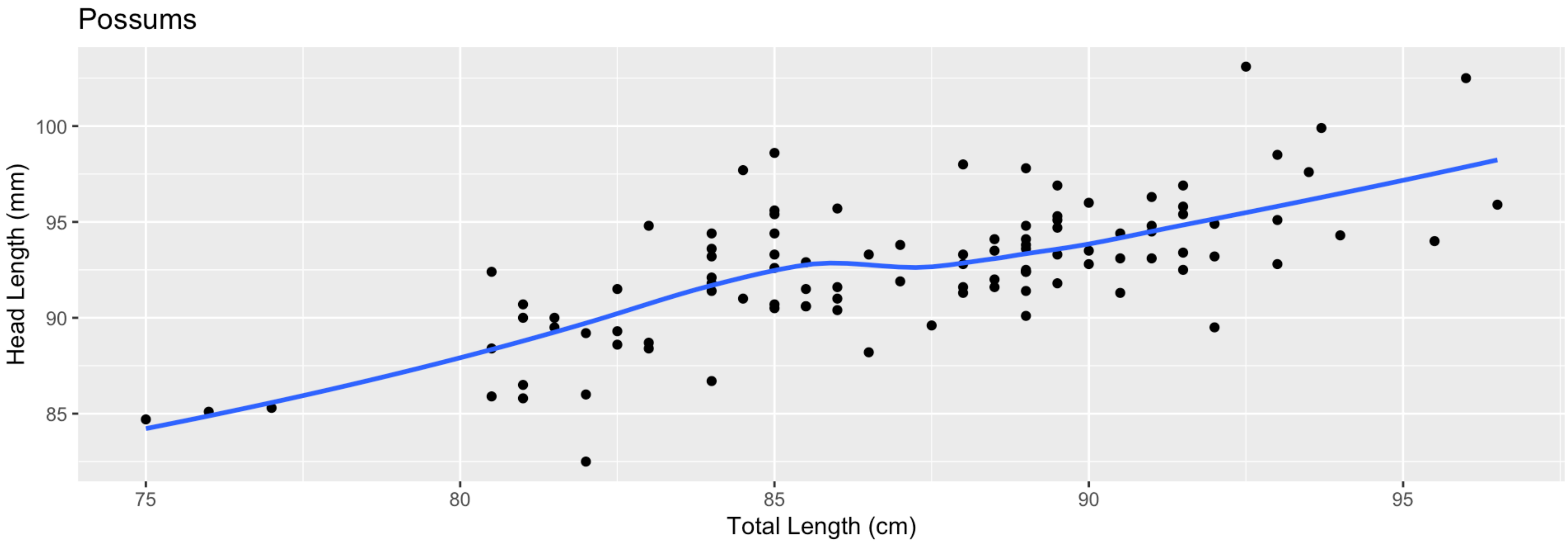
Tool #1 - Scatterplots w/trend line



Form? Direction? Strength? Outliers?

Tool #1 - Scatterplot Practice

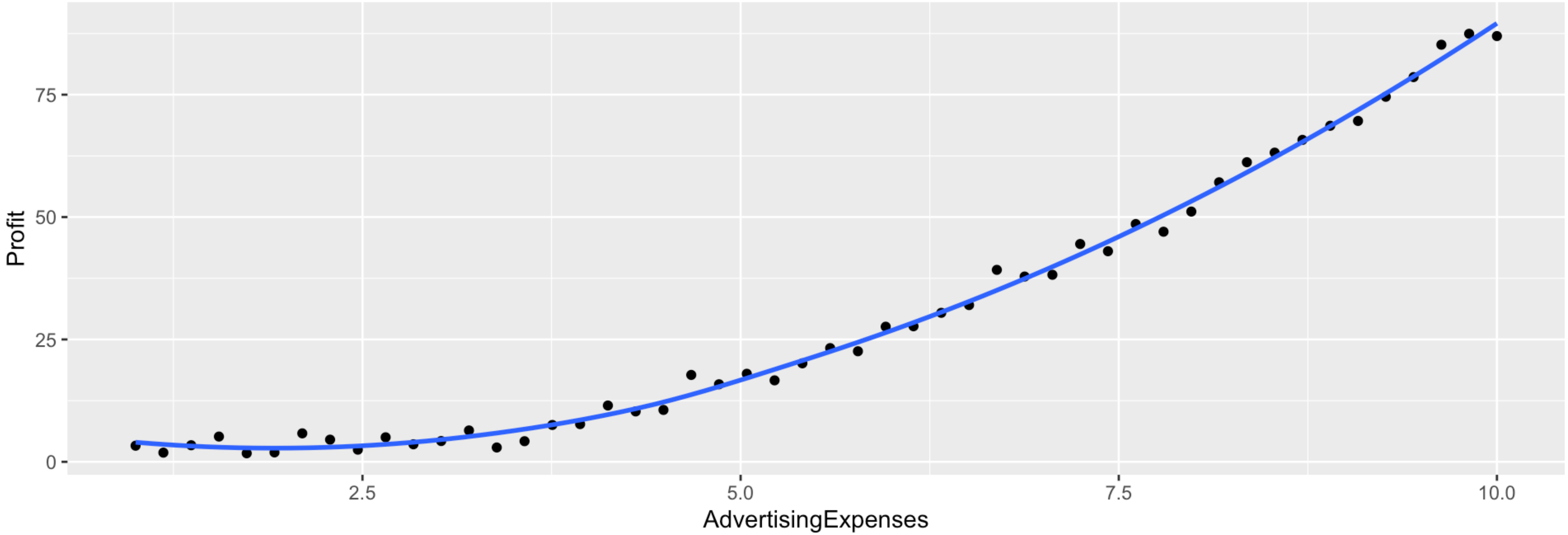
Ecology example: Is possum length related to head length?



Form? Direction? Strength? Outliers?

Tool #1 - Scatterplot Practice

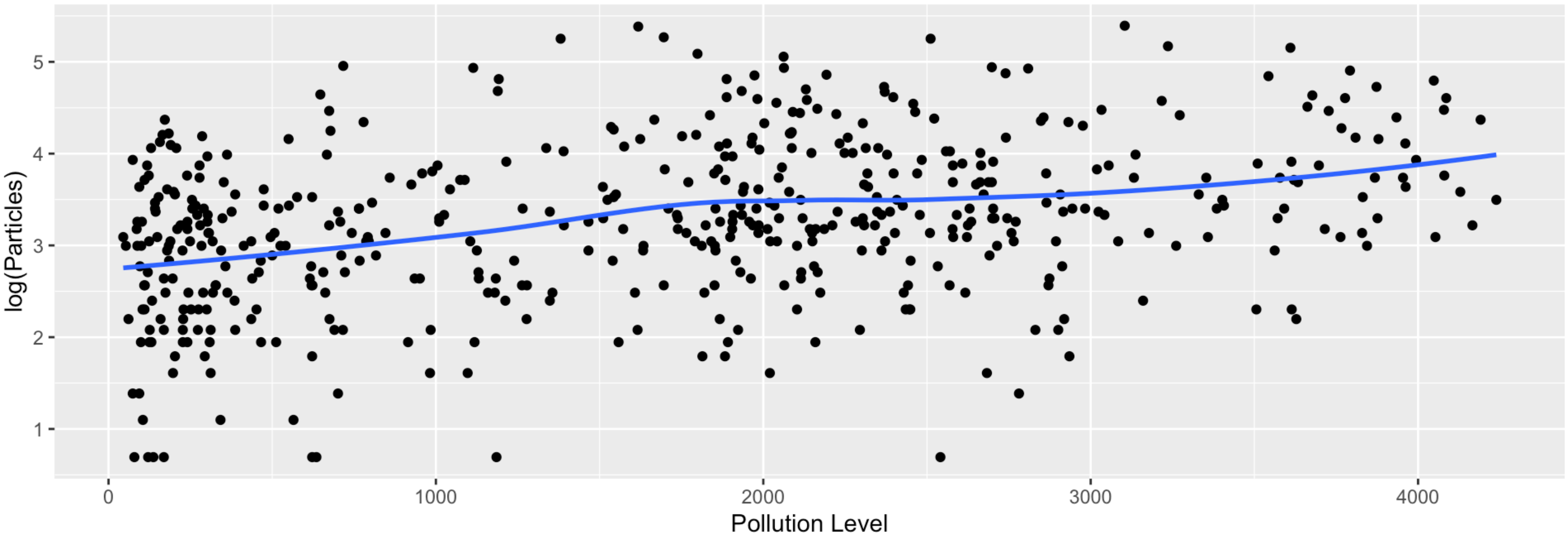
Marketing example: Is advertising expenses related to profit?



Form? Direction? Strength? Outliers?

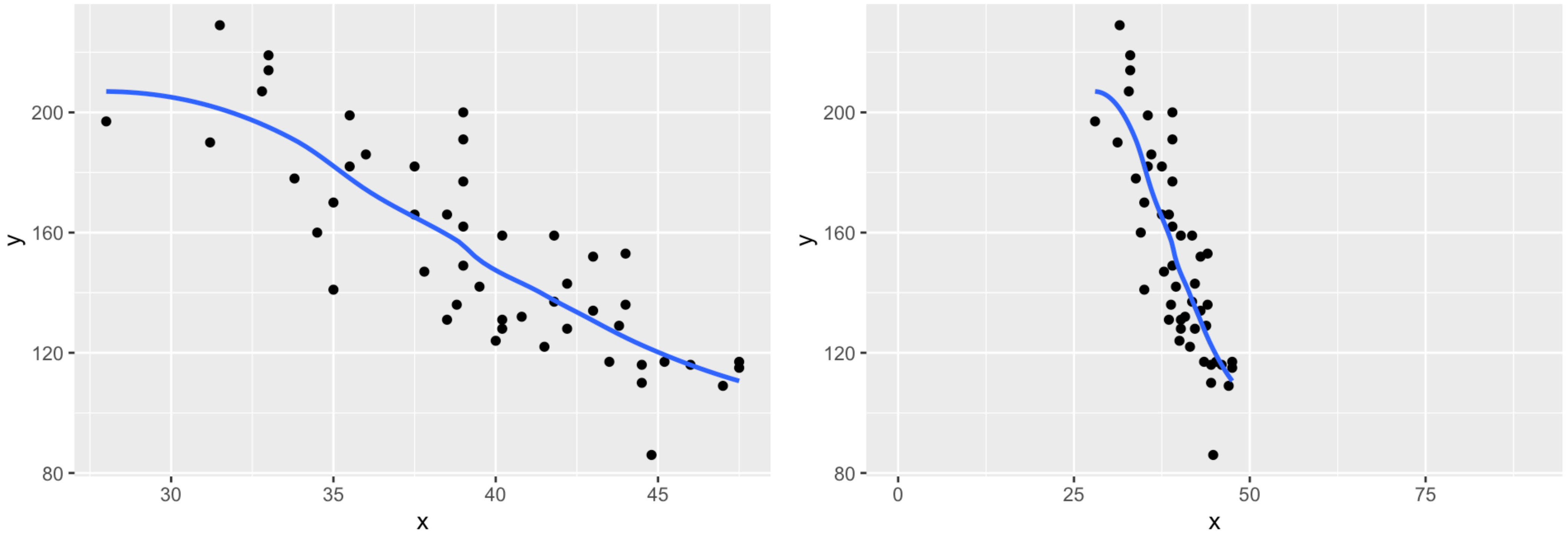
Tool #1 - Scatterplot Practice

Environment example: How much pollution do cars produce?



Form? Direction? Strength? Outliers?

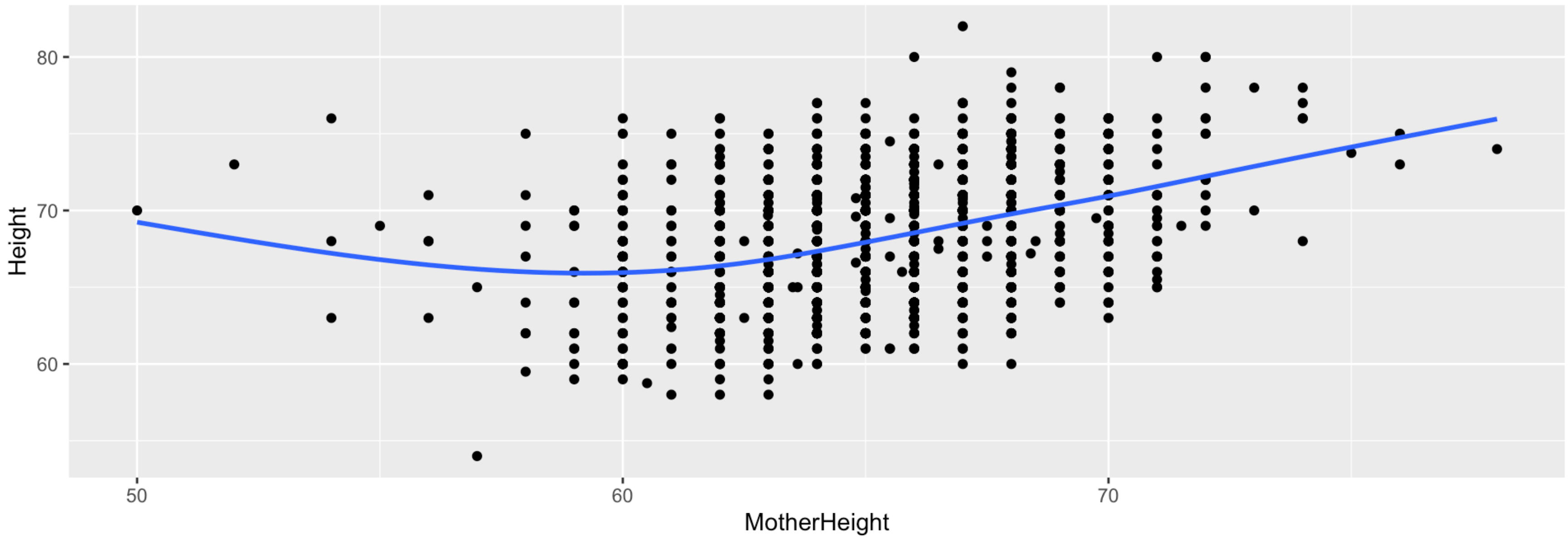
Tool #1 - Scatterplot Practice



Which graph has a stronger relationship?

- Trick question- they are the same data!
- We need a numeric (objective) measure of strength.

Tool #2 - Covariance



Covariance: a measure of the linear relationship between y and x (how much y changes as x changes), but with units that are difficult to interpret.

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\ &= 4.251\end{aligned}$$

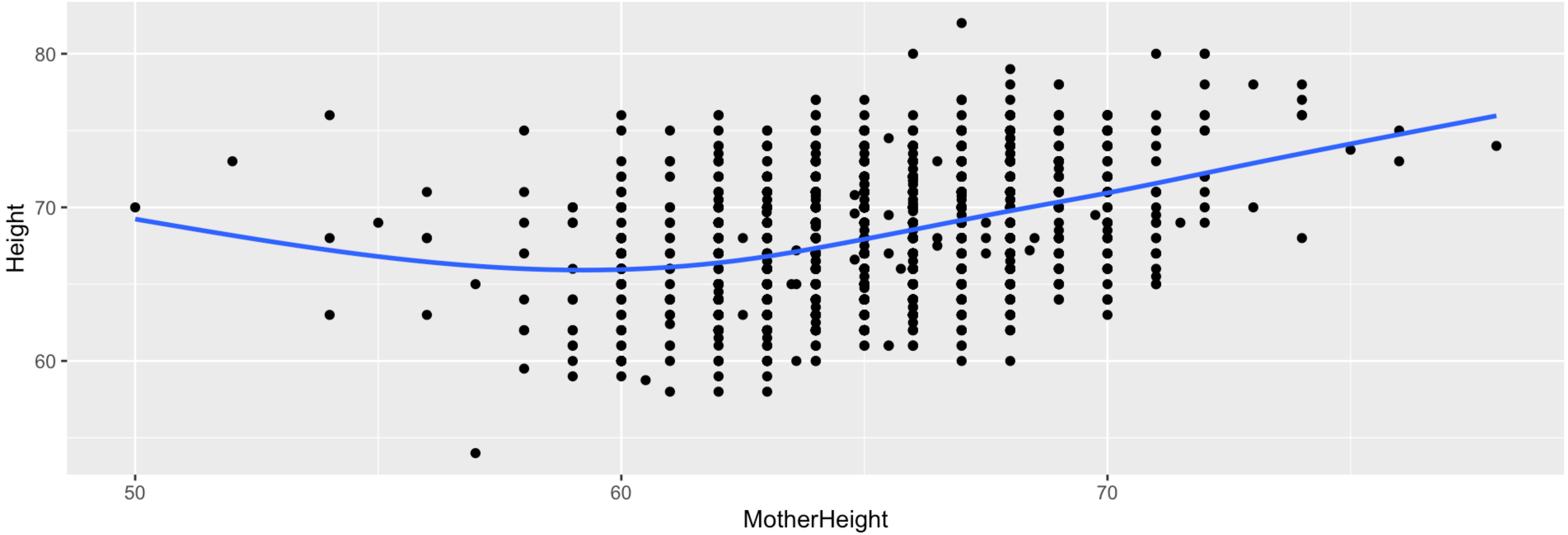
Tool #2 - Covariance

Properties of Covariance:

- If $\mathrm{Cov}(X, Y) < 0 \Rightarrow$ negative linear relationship
- If $\mathrm{Cov}(X, Y) > 0 \Rightarrow$ positive linear relationship
- Highly impacted by the unit of measurements for X and Y.
- Highly impacted by outliers

- What we really want is a standardized measure of strength

Tool #3 - Correlation



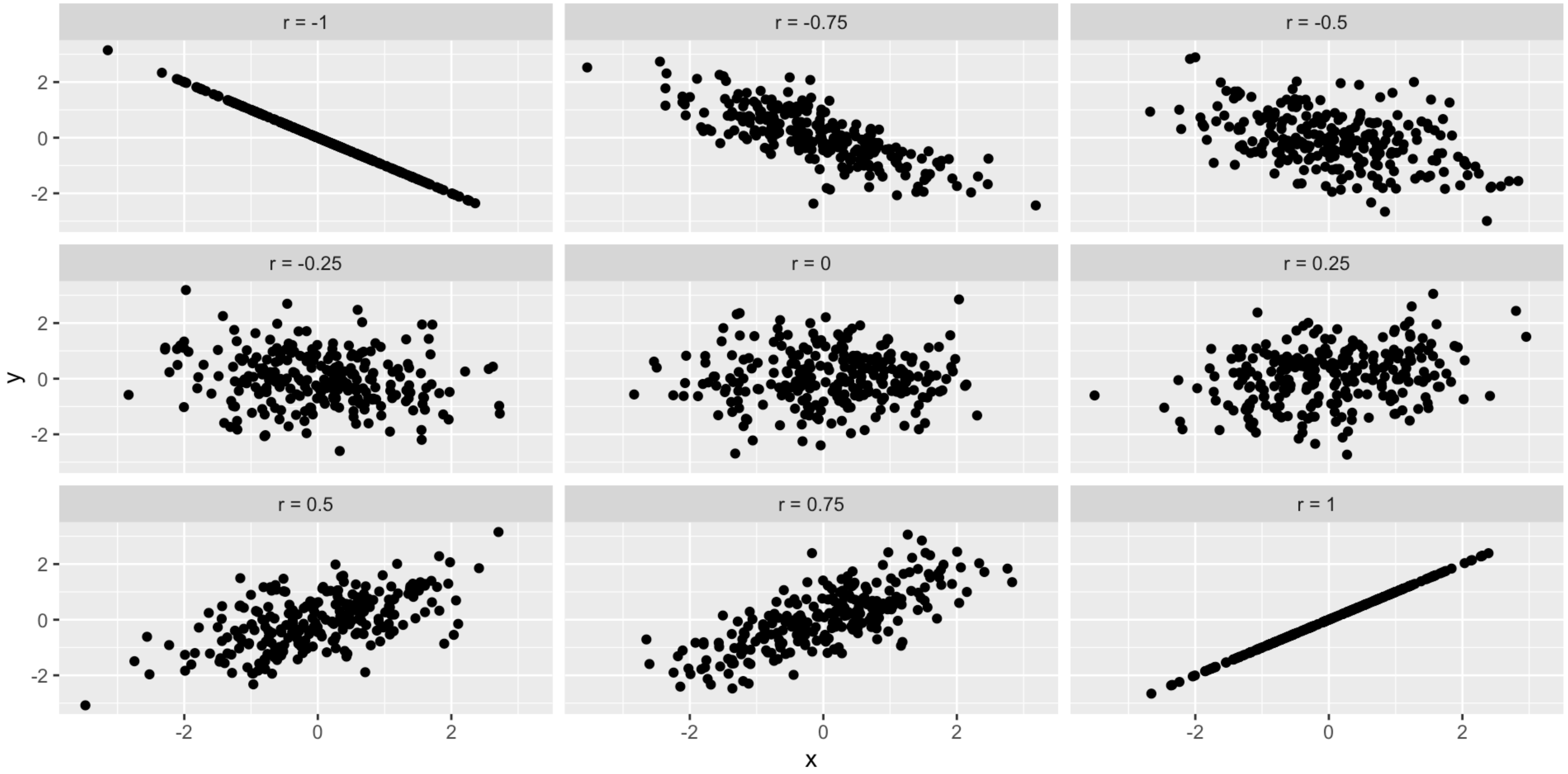
Correlation: A standardized measure of strength between -1 and 1:

$$\begin{aligned}\mathrm{Corr}(X,Y) = r &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)\\ &= 0.348\end{aligned}$$

Tool #3 - Correlation

Properties of Correlation (r):

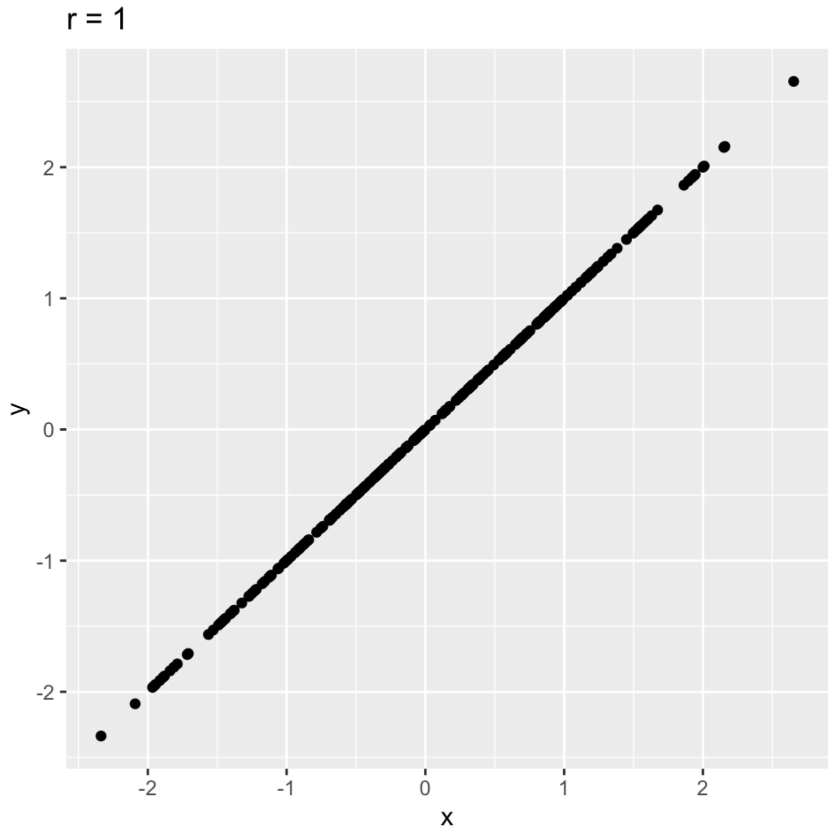
- $-1 < r < 1$

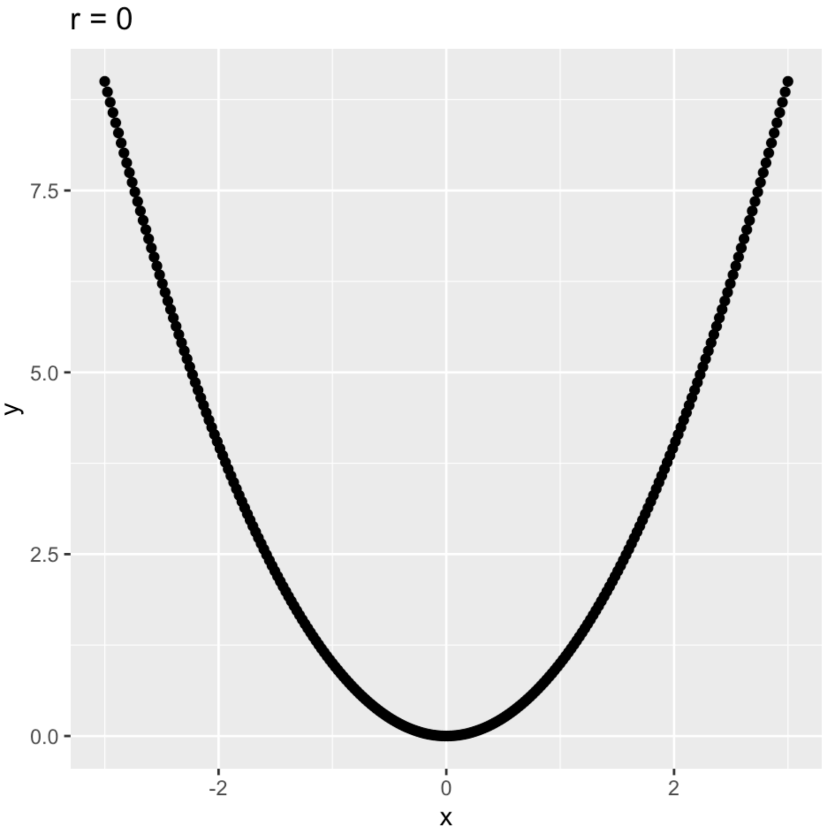


Tool #3 - Correlation

Properties of Correlation (r):

- $-1 < r < 1$
- Only appropriate for LINEAR relationships


r = 1


r = 0

Tool #3 - Correlation

Properties of Correlation (r):

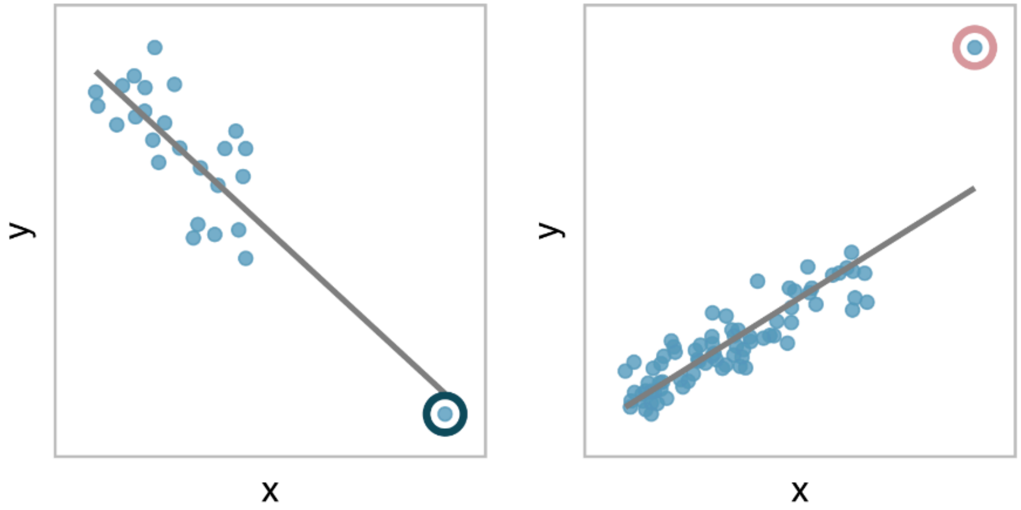
- $-1 < r < 1$
- Only appropriate for LINEAR relationships
- NOT impacted by scale of data (scale invariant). For example:

$$\begin{aligned}&\text{Cor(Height in inches, Weight in pounds)} = \\ &\qquad \text{Cor(Height in meters, Weight in kg)}\end{aligned}$$

Tool #3 - Correlation

Properties of Correlation (r):

- $-1 < r < 1$
- Only appropriate for LINEAR relationships
- NOT impacted by scale of data (scale invariant). For example:
- Highly impacted by outliers



In one case the outlier made r go up, in the other r goes down.

Tool #3 - Correlation

Properties of Correlation (r):

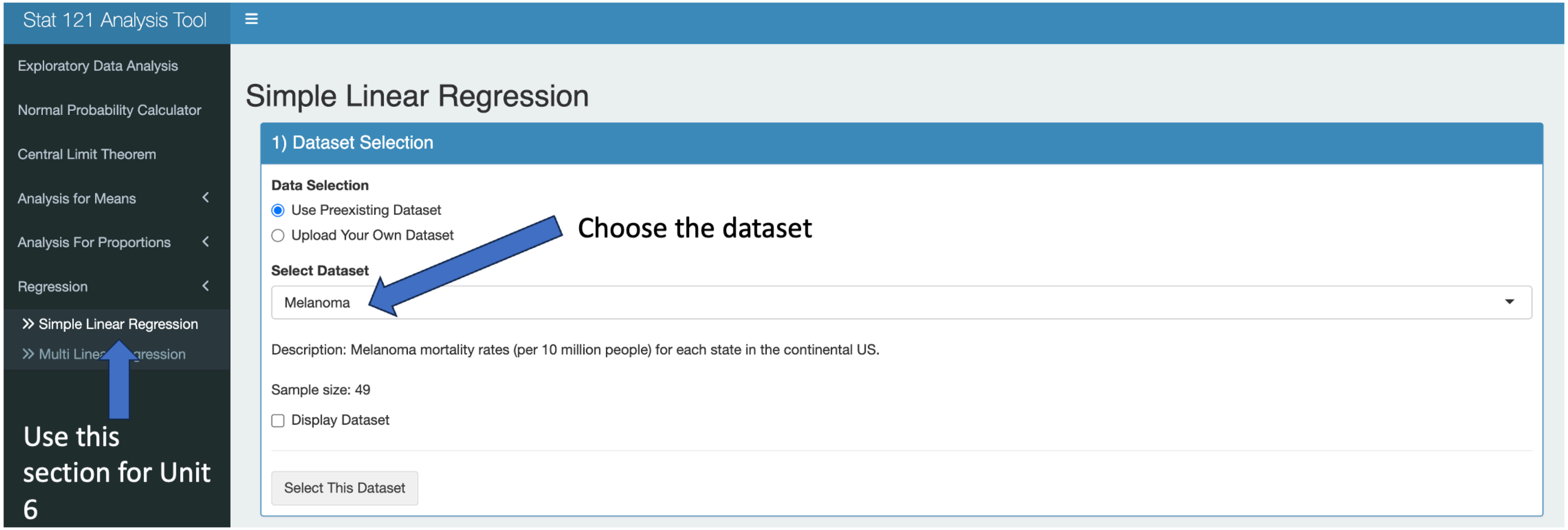
- $-1 < r < 1$
- Only appropriate for LINEAR relationships
- NOT impacted by scale of data (scale invariant). For example:
- Highly impacted by outliers
- Only for 2 quantitative variables. For example, correlation between state and income doesn’t make sense.

Tool #3 - Correlation

Properties of Correlation (r):

- $-1 < r < 1$
- Only appropriate for LINEAR relationships
- NOT impacted by scale of data (scale invariant). For example:
- Highly impacted by outliers
- Only for 2 quantitative variables. For example, correlation between state and income doesn’t make sense.
- $\text{Cor}(X, Y) = \text{Cor}(Y, X)$

Using the Analysis Tool



Stat 121 Analysis Tool
≡
Exploratory Data Analysis
Normal Probability Calculator
Central Limit Theorem
Analysis for Means
Analysis For Proportions
Regression
» Simple Linear Regression
» Multi Linear Regression

Simple Linear Regression

1) Dataset Selection

Data Selection

● Use Preexisting Dataset

○ Upload Your Own Dataset

Select Dataset: Melanoma

Description: Melanoma mortality rates (per 10 million people) for each state in the continental US.

Sample size: 49

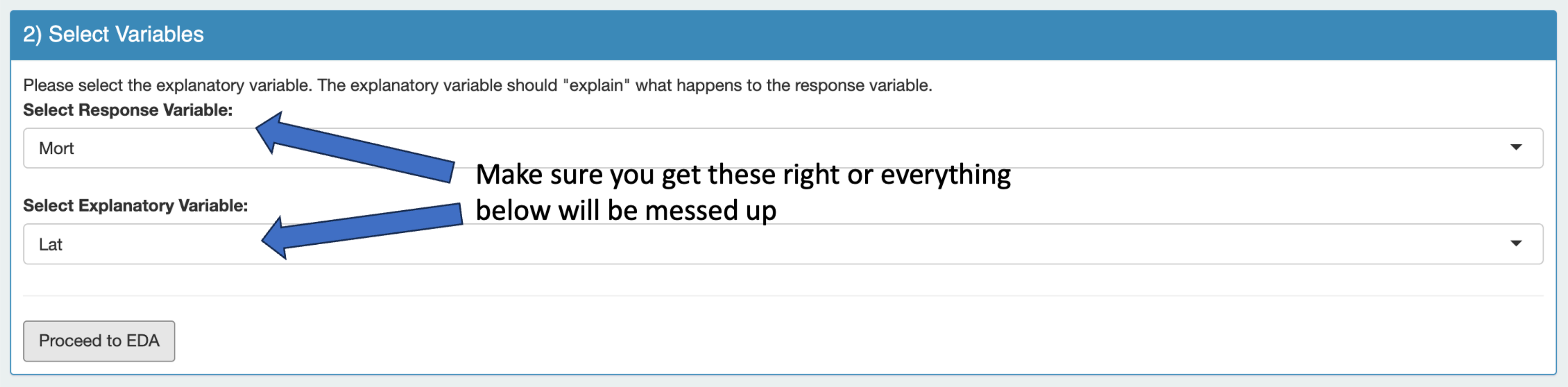
□ Display Dataset

Select This Dataset

Choose the dataset

Use this section for Unit 6

Using the Analysis Tool



2) Select Variables

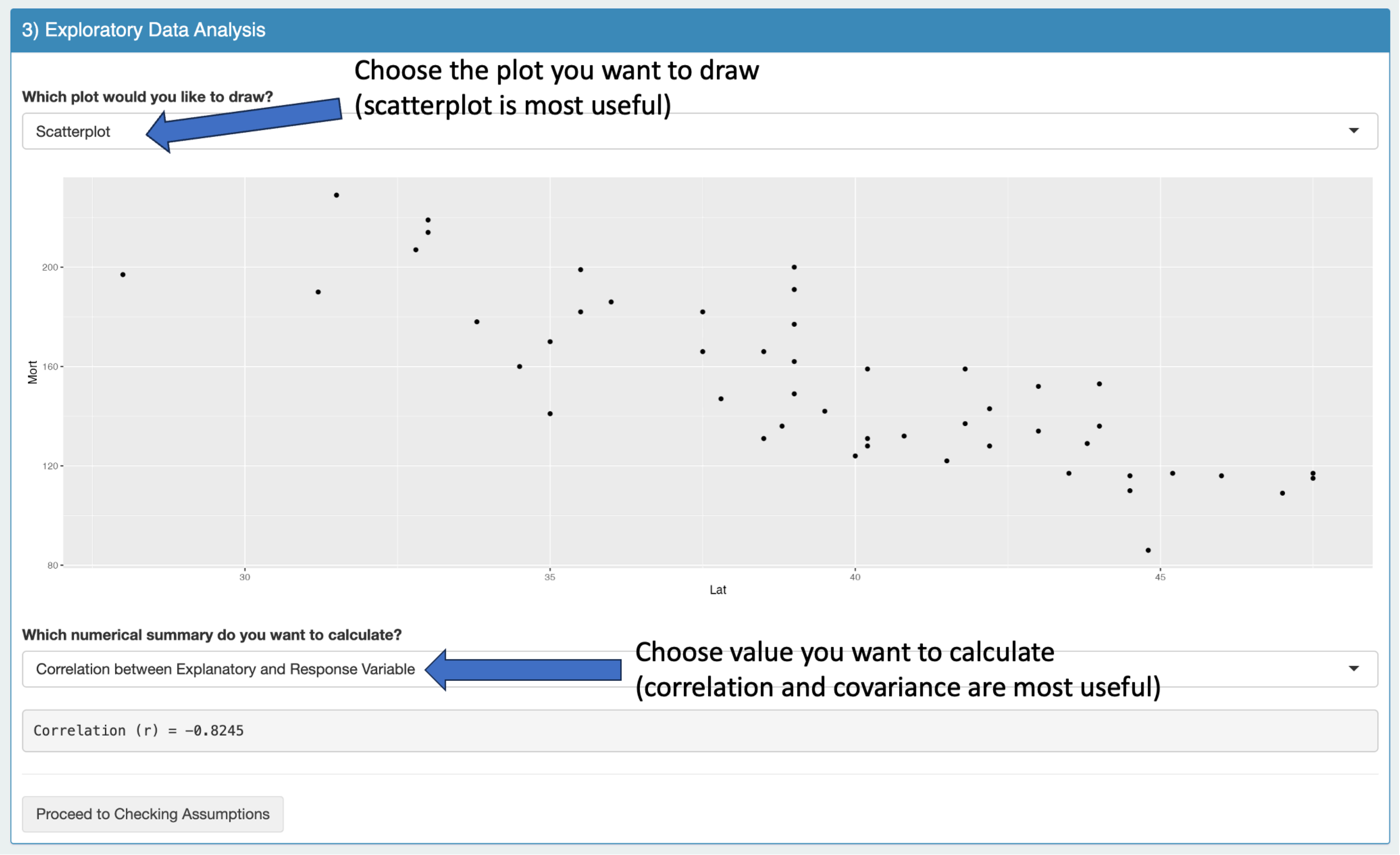
Please select the explanatory variable. The explanatory variable should "explain" what happens to the response variable.

Select Response Variable: Mort

Select Explanatory Variable: Lat

Make sure you get these right or everything below will be messed up

Using the Analysis Tool



3) Exploratory Data Analysis

Choose the plot you want to draw
(scatterplot is most useful)

Which plot would you like to draw?

Scatterplot

Which numerical summary do you want to calculate?

Correlation between Explanatory and Response Variable

Choose value you want to calculate
(correlation and covariance are most useful)

```
Correlation (r) = −0.8245
```

Correlation is not causation

Just because two variables are correlated, does not mean that one causes the other. For example (examples taken from spurious correlations):

1. The correlation between the number of movies made by Nicolas Cage and the number of drowning deaths is 0.66. Does this mean that Nicolas Cage movies cause drownings?
2. The correlation between the number of global shark attacks and ice cream sales in 0.81. Does this mean that shark attacks cause people to buy ice cream?
3. The correlation between the per capita consumption of margarine and the divorce rate in Maine is 0.99. Does this mean that eating more margarine causes divorce?

- Causal relationships can only be established if we have done an experiment and tried to control for as many lurking variables as possible.

Homework Choices for Unit 6

1. Rate my professor - what matters in determining a rate my professor score?
2. Supervisor - what makes people like their manager?
3. Body Fat - what body measurements are predictive of your BMI?
4. Basketball Salary - what skills lead to a higher salary?

Key Terminology

- Scatterplot
- Form
- Direction
- Strength
- Outliers
- Correlation and Properties
- Covariance and Properties