Class 31

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-11-20

Review: 2 Numeric Variables

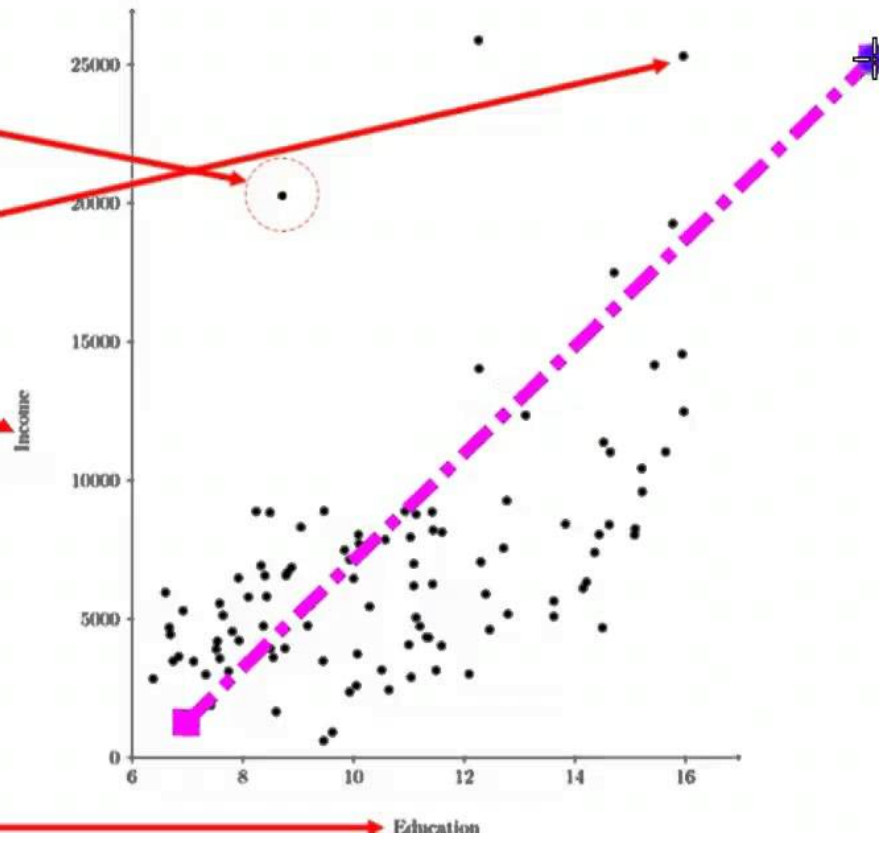
Understand how to read a scatter plot.

Outlier: a point that is far away from the pattern of the other data points

Influential point: a point that is far away from the other data points

Response variable / dependent variable: the variable that reacts to the explanatory variable; it is plotted on the y-axis

Explanatory variable / independent variable: the variable that causes the other variable to change; it is plotted on the x-axis



Review: Describing Associations

- ***Independence***: an increase in X is not associated with a change in Y

Review: Describing Associations

- ***Independence***: an increase in X is not associated with a change in Y
- ***Positive association***: an increase in X is associated with an *increase* in Y

Review: Describing Associations

- ***Independence***: an increase in X is not associated with a change in Y
- ***Positive association***: an increase in X is associated with an *increase* in Y
- ***Negative association***: an increase in X is associated with a *decrease* in Y

Review: Describing Associations

- ***Independence***: an increase in X is not associated with a change in Y
- ***Positive association***: an increase in X is associated with an *increase* in Y
- ***Negative association***: an increase in X is associated with a *decrease* in Y
- ***Weak association***: data points are very far apart from each other

Review: Describing Associations

- ***Independence***: an increase in X is not associated with a change in Y
- ***Positive association***: an increase in X is associated with an *increase* in Y
- ***Negative association***: an increase in X is associated with a *decrease* in Y
- ***Weak association***: data points are very far apart from each other
- ***Strong association***: data points are tightly clustered

Pratice

Which image shows a ***positive*** relationship between the explanatory and response variables?

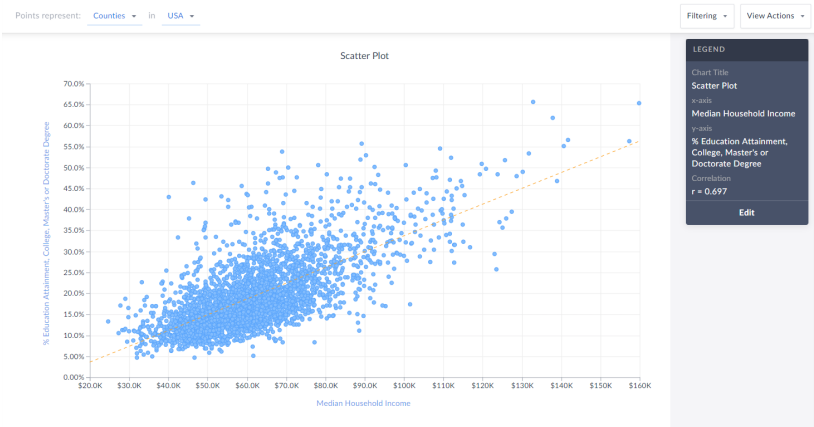


Figure 1: Income vs Education

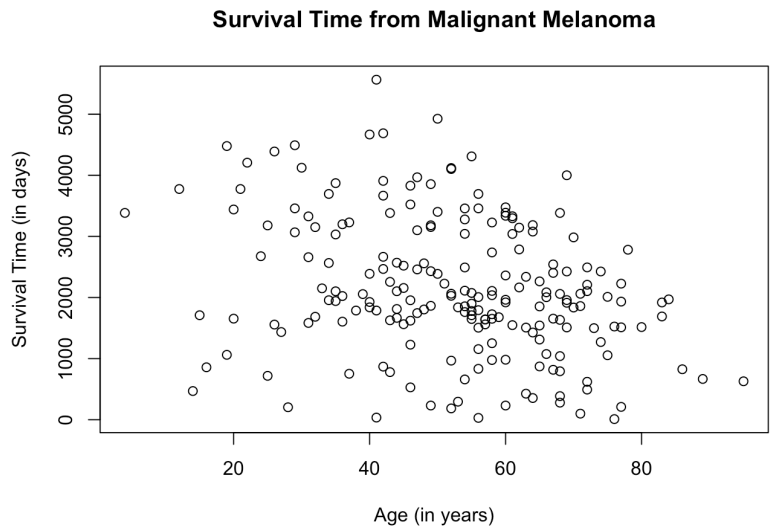


Figure 2: Age vs Survival

Practice

Which image shows a ***weak*** relationship between the explanatory and response variables?

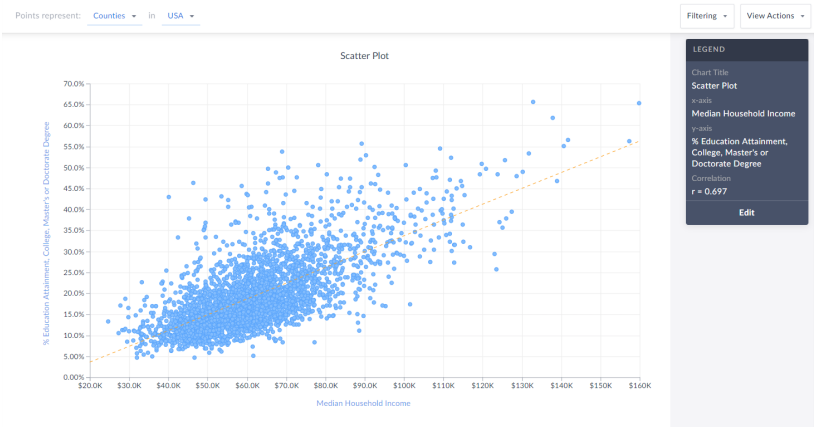


Figure 3: Income vs Education

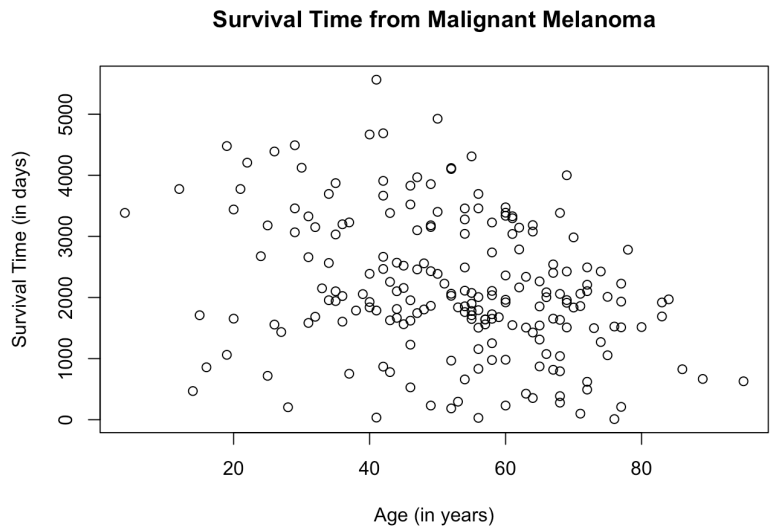


Figure 4: Age vs Survival

Correlation

- Describes the direction and strength of the association between 2 numeric variables

Correlation

- Describes the direction and strength of the association between 2 numeric variables
- A correlation ranges from -1 to 1
  - A perfect negative correlation equals -1
  - A perfect positive correlation equals 1

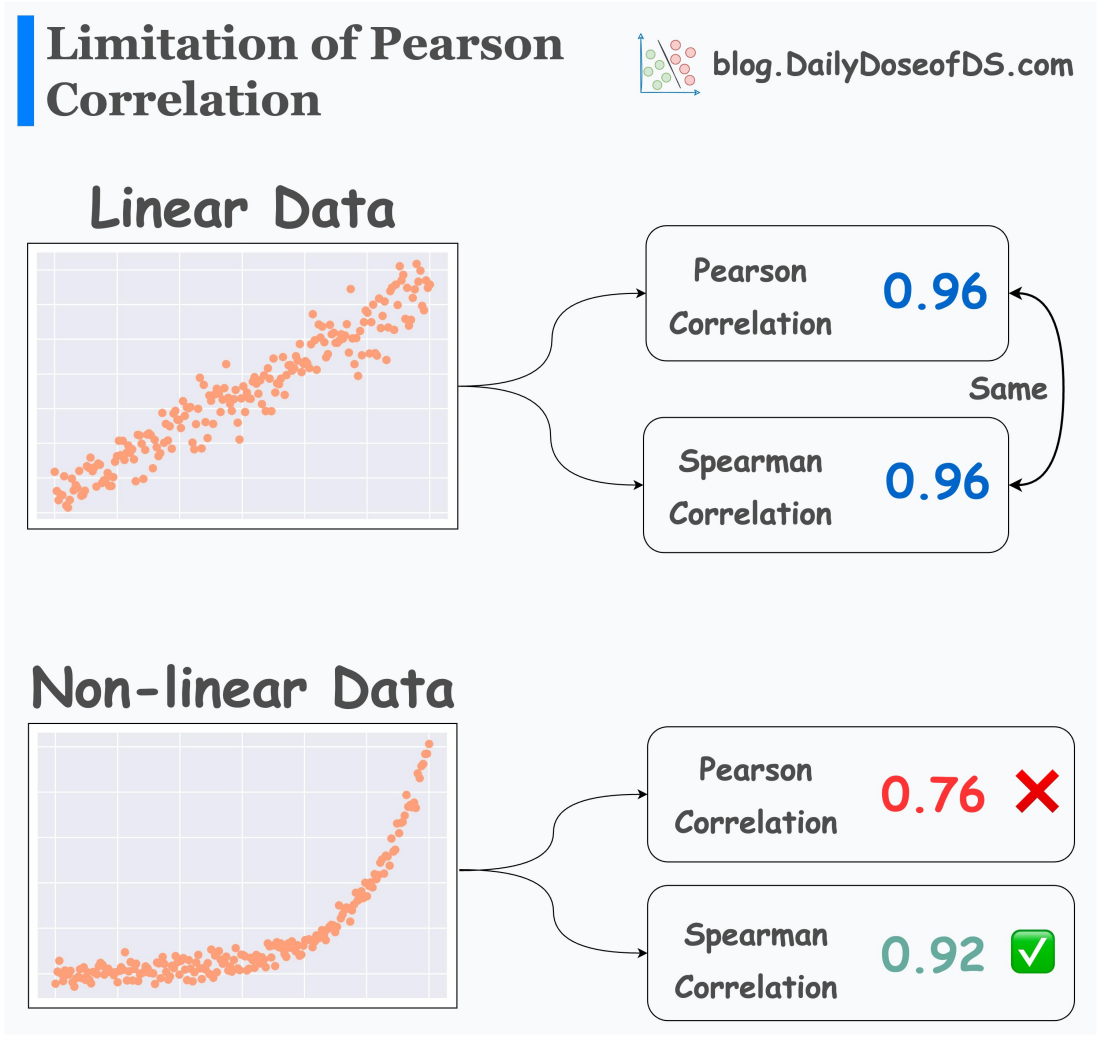
Correlation

- Describes the direction and strength of the association between 2 numeric variables
- A correlation ranges from -1 to 1
  - A perfect negative correlation equals -1
  - A perfect positive correlation equals 1
- A correlation of 0 indicates the two variables are independent (no relationship)

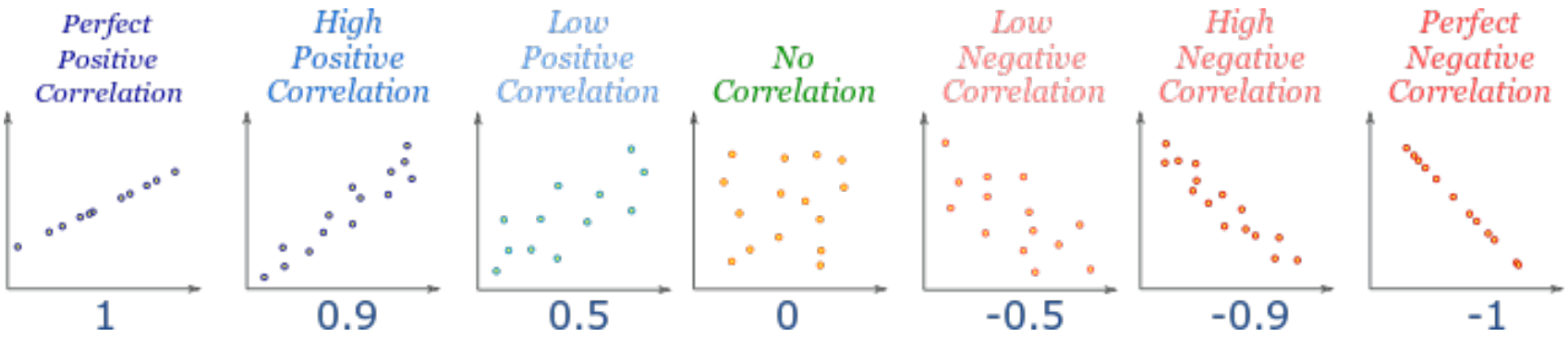
Correlation

- Describes the direction and strength of the association between 2 numeric variables
- A correlation ranges from -1 to 1
  - A perfect negative correlation equals -1
  - A perfect positive correlation equals 1
- A correlation of 0 indicates the two variables are independent (no relationship)
- We use the Pearson correlation for linear relationships

Linear vs Non-Linear



Interpreting Correlations



Example: Poverty vs Graduation Rate

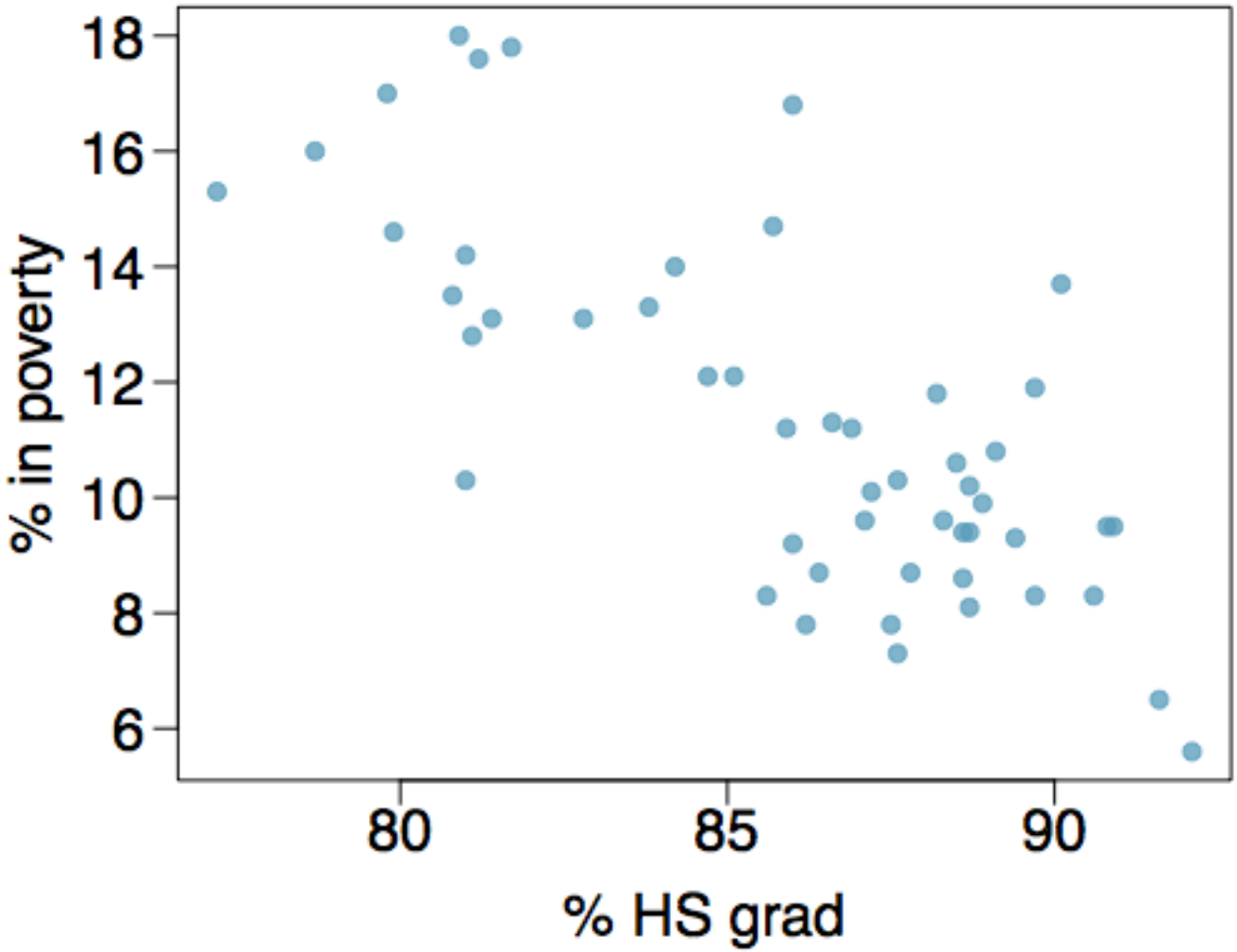


Figure 5: What's the response variable?

Example: Poverty vs Graduation Rate

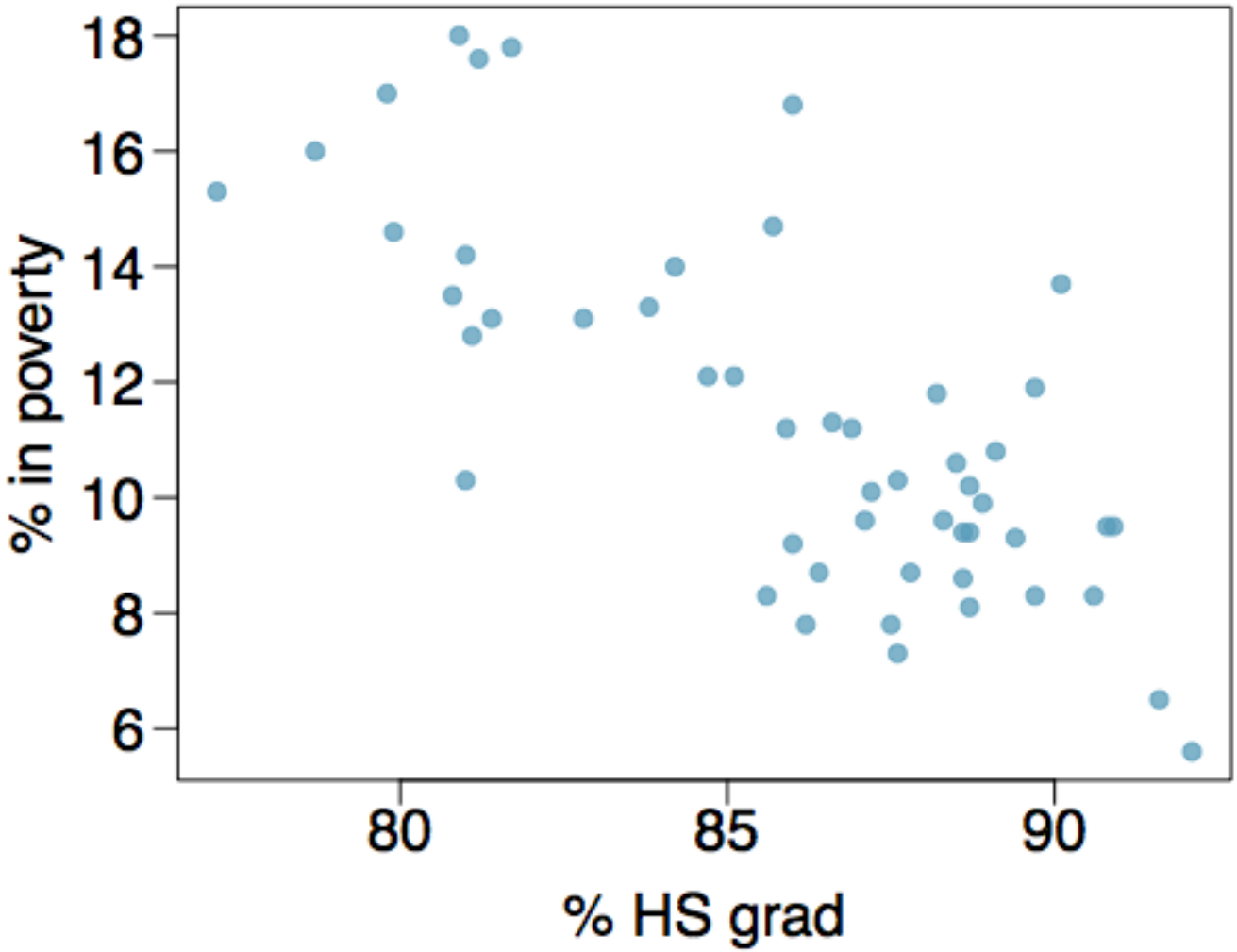


Figure 5: What's the response variable?

Response Variable: Percent of people in poverty

Example: Poverty vs Graduation Rate

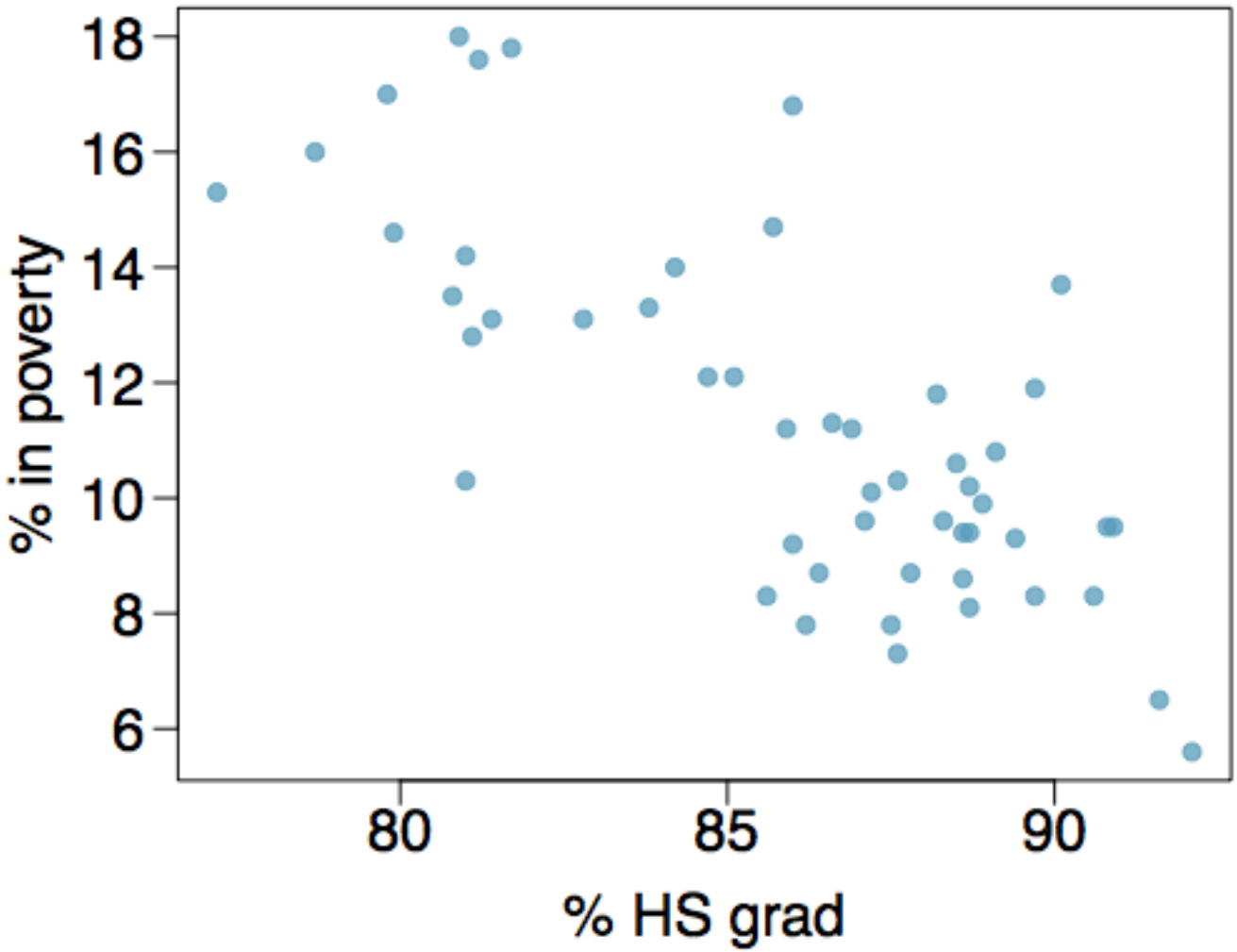


Figure 6: What's the explanatory variable?

Example: Poverty vs Graduation Rate

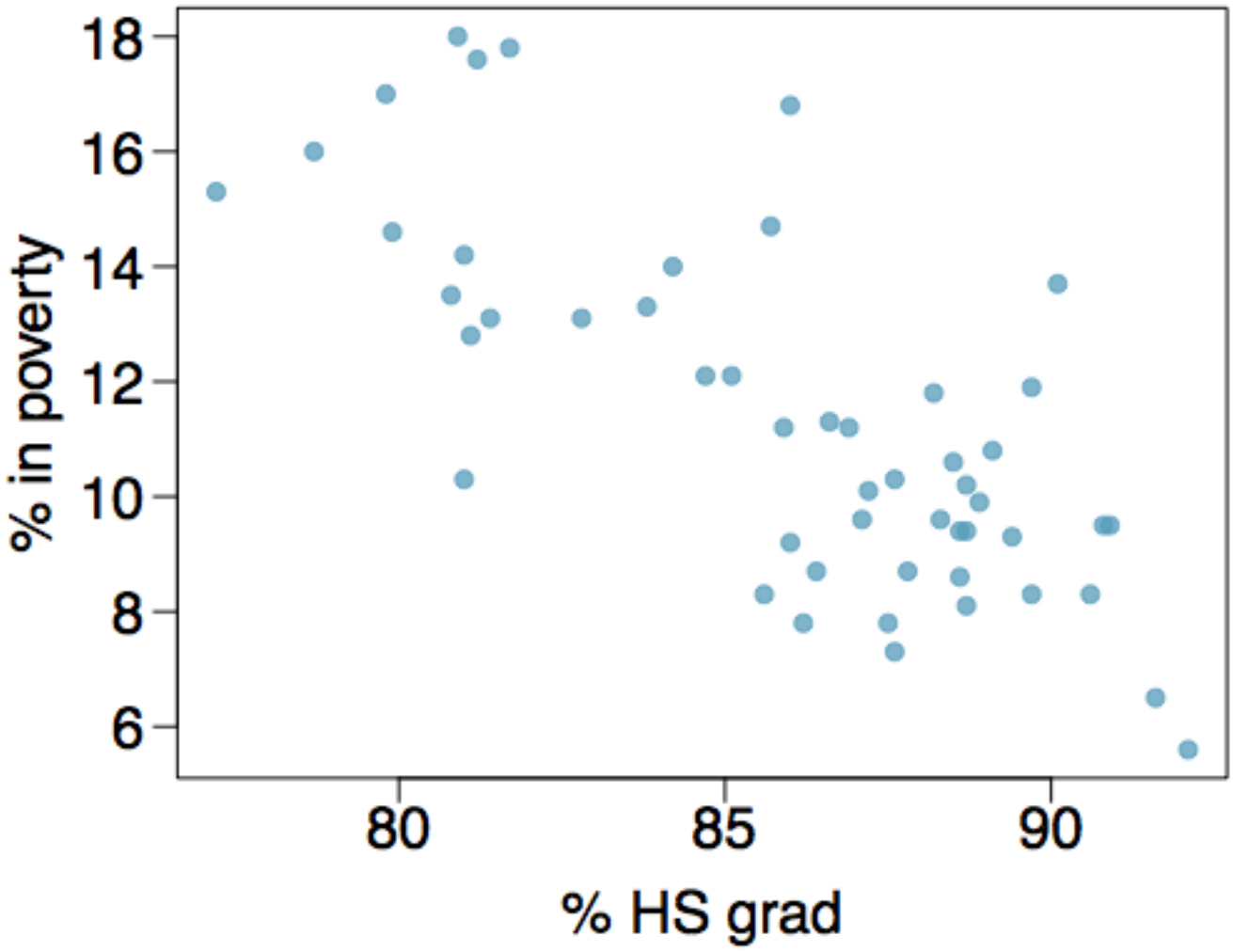


Figure 6: What's the explanatory variable?

Explanatory variable: Percent of people who graduated high school

Example: Poverty vs Gradution Rate

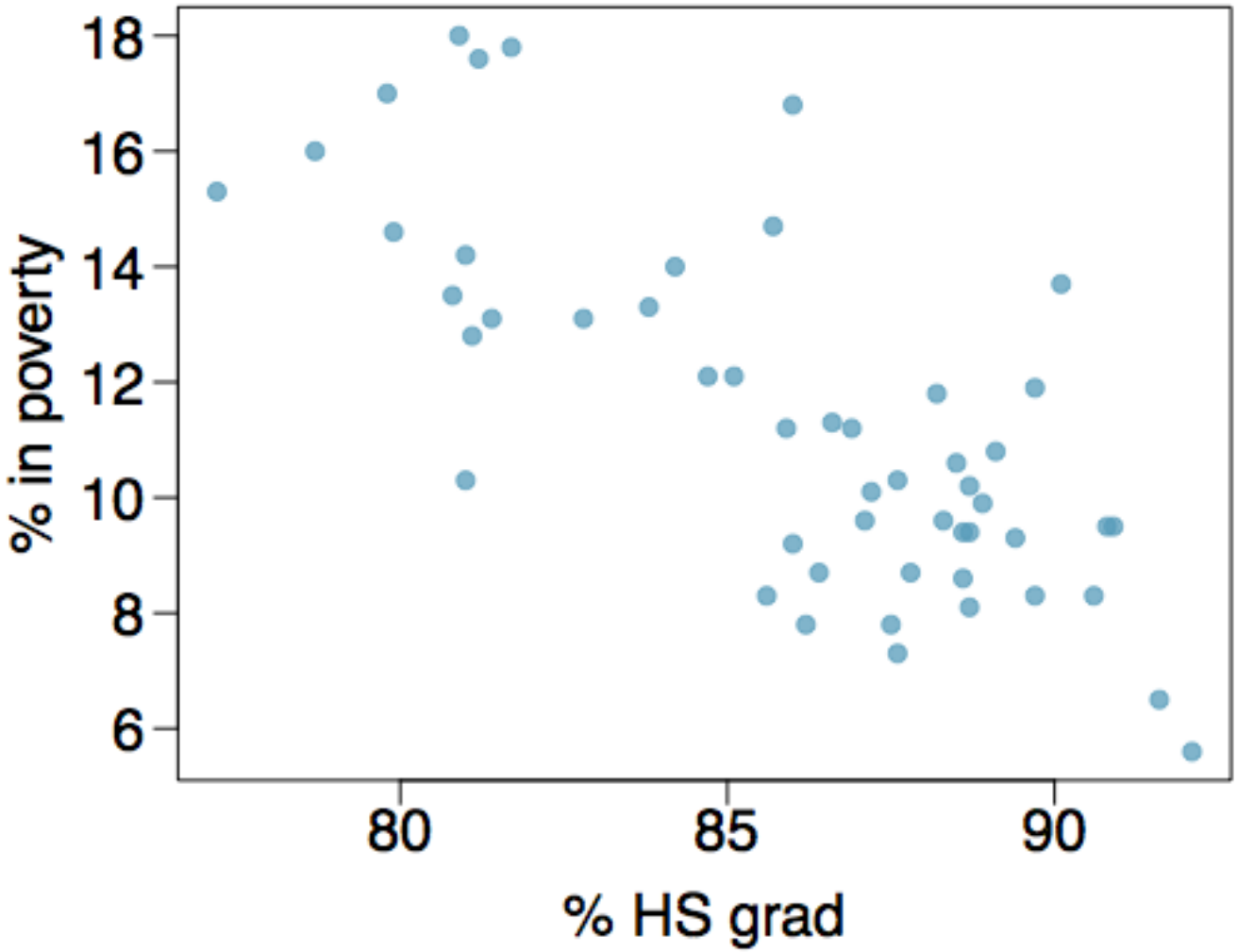


Figure 7: Describe the relationship between these 2 variables.

Example: Poverty vs Gradution Rate

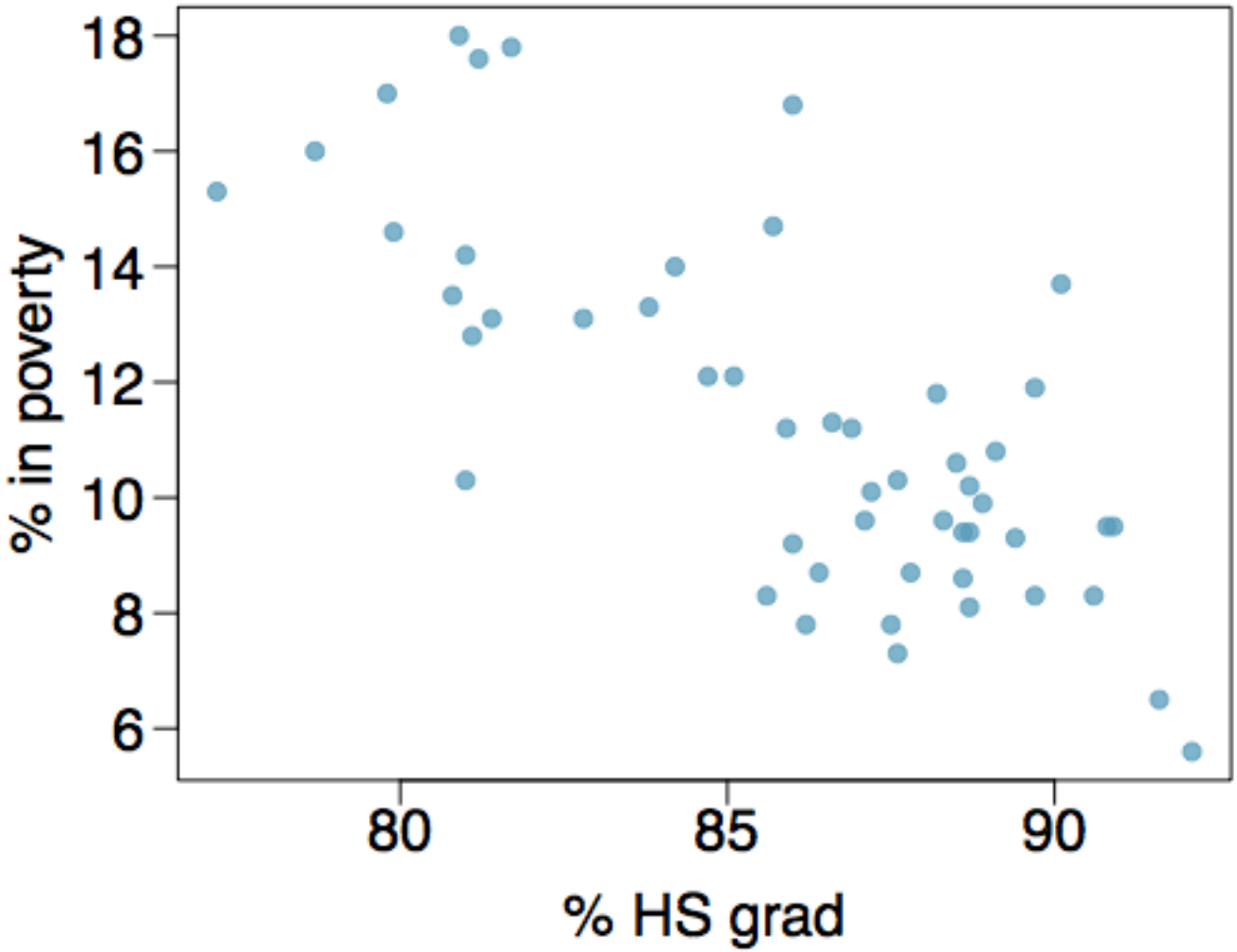


Figure 7: Describe the relationship between these 2 variables.

Relationship: linear, negative, moderate to strong

Example: Poverty vs Graduation Rate

Which of the following is the most likely correlation? A) 0.60 B) -0.25 C) -0.75 D) 0.35

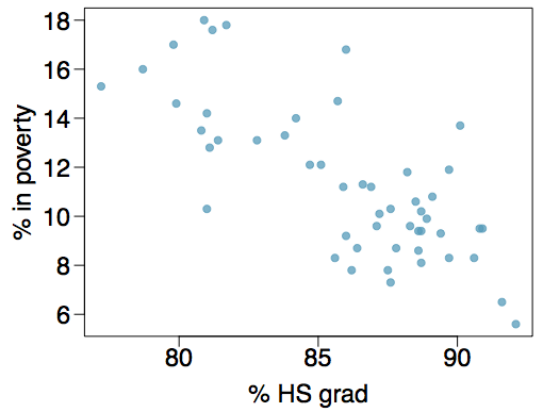


Figure 8: Describe the relationship between these 2 variables.

Example: Poverty vs Graduation Rate

Which of the following is the most likely correlation? ***C.*** ***-0.75***

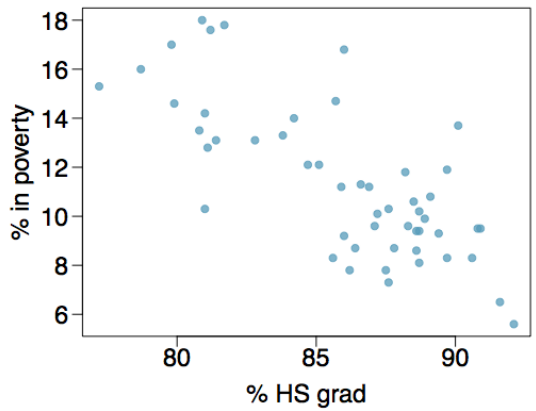


Figure 9: Describe the relationship between these 2 variables.

Testing a Correlation

- Null Hypothesis: The two variables are independent (correlation $= 0$)

$$H_0 : \rho = 0$$

Testing a Correlation

- Null Hypothesis: The two variables are independent (correlation $= 0$)

$$H_0 : \rho = 0$$

- Alternate Hypothesis: the two variables are dependent

$$H_A : \rho > 0$$
$$\rho < 0$$
$$\rho \neq 0$$

Test Statistic

The test statistic t for the population Pearson correlation ρ (Greek letter rho) is estimated using the observed correlation r.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Getting a p-value

Use the Student's t distribution with degrees of freedom $\text{df} = n - 2$ to find a p-value for the observed correlation r in a sample of size n under the null hypothesis $H_0 : \rho = 0$.

```
# specify the test statistic and degrees of freedom
pt(test_statistic,
   df = n-2,
   lower.tail = F) # optional parameter
```

Eyeballing a Line

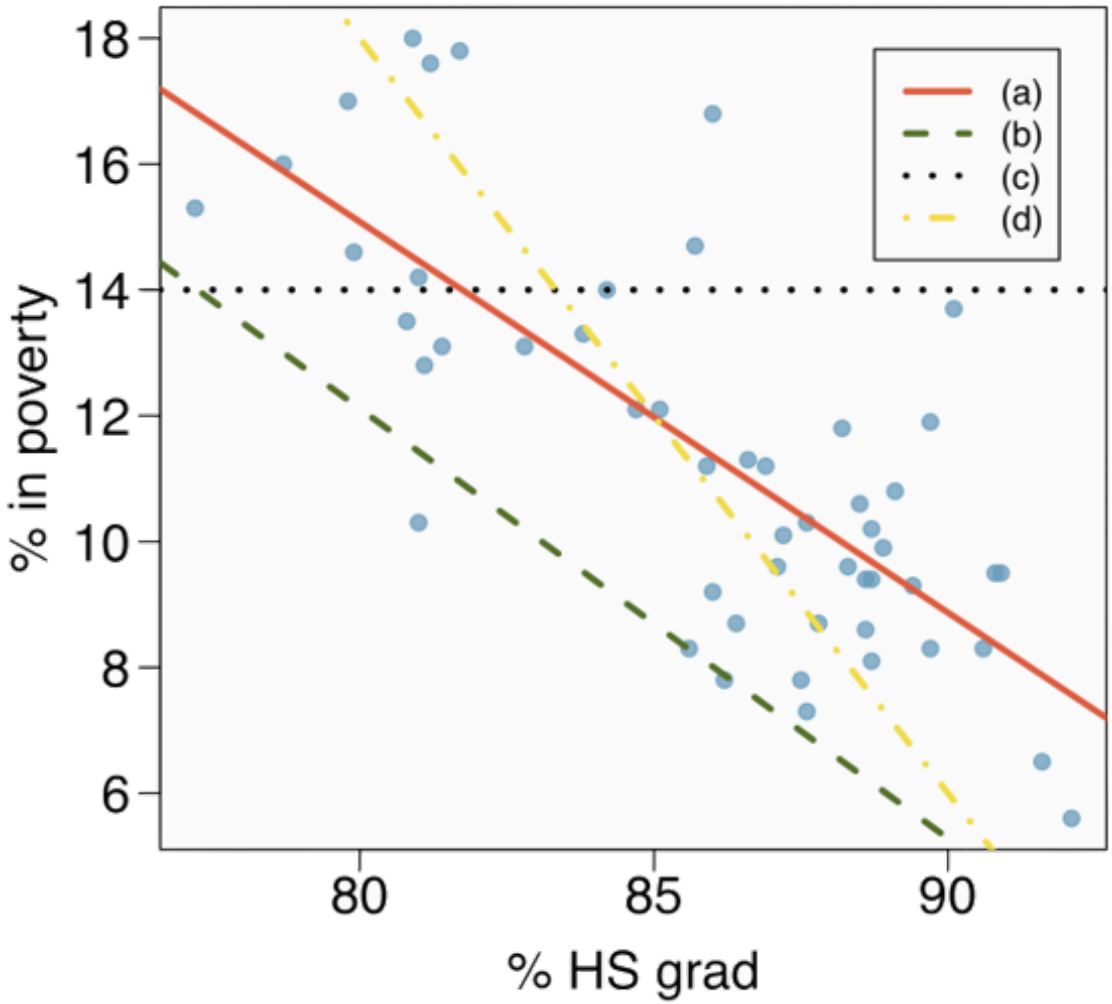
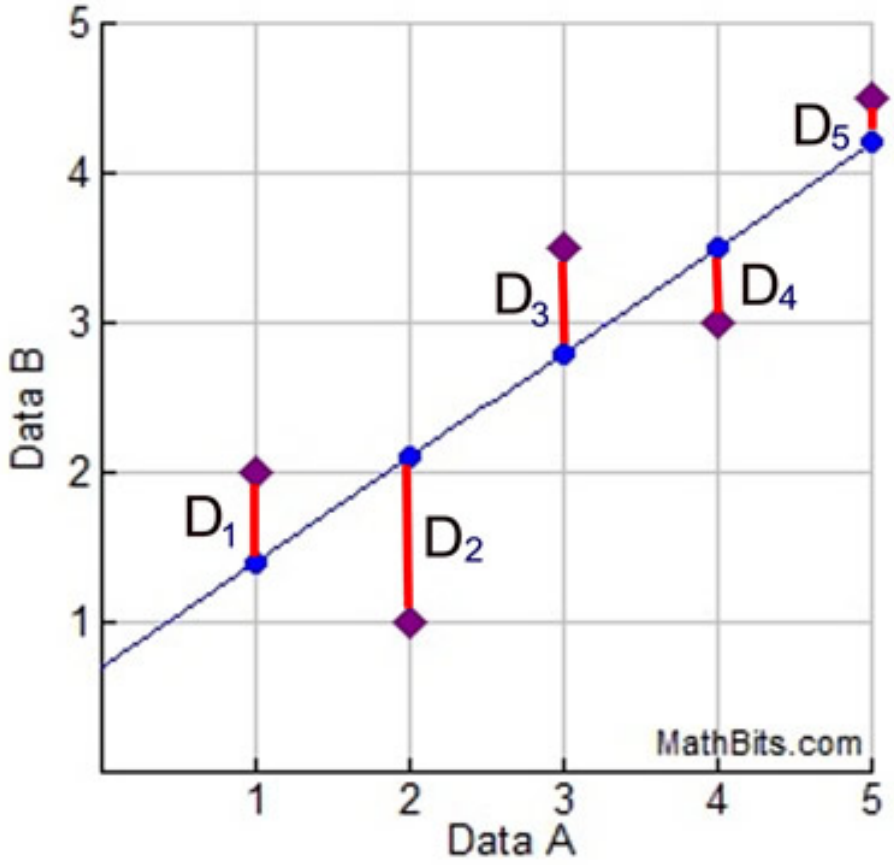


Figure 10: How do we find the best line to draw through variables that appear to have a linear relationship?

Quantifying Error: Residuals



◆ Scatter Plot Points:

{(1,2), (2,1), (3,3½), (4,3), (5,4)}

● Regression Points

{(1,1.4), (2,2.1), (3,2.8), (4,3.5), (5,4.2)}

The Red Line Segments:

The red line segments represent the distances between the y-values of the actual scatter plot points, and the y-values of the regression equation at those points.

The lengths of the red line segments are called RESIDUALS.

Figure 11: Residuals are the difference between the observed values and the predicted values.

Special Topic: Correlation vs Causation

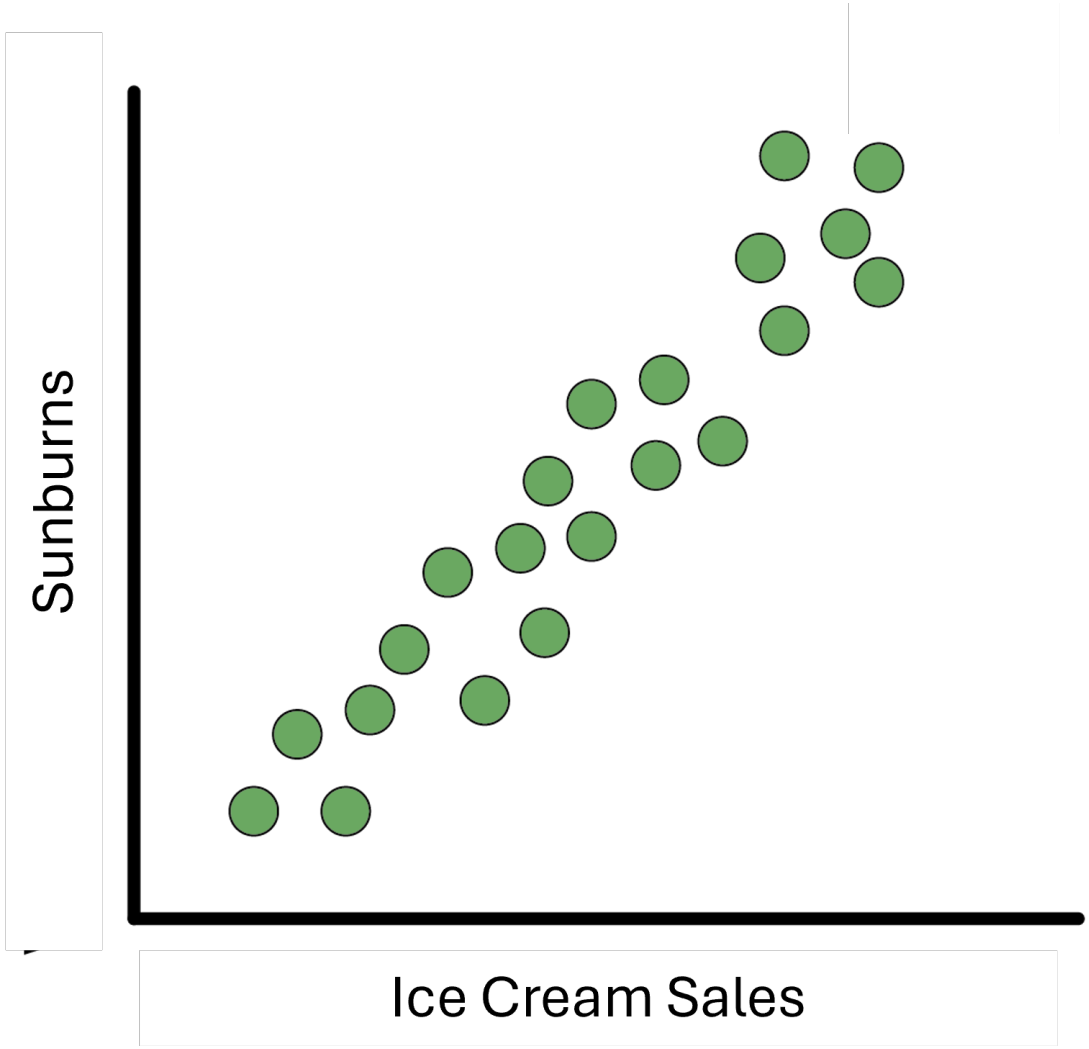
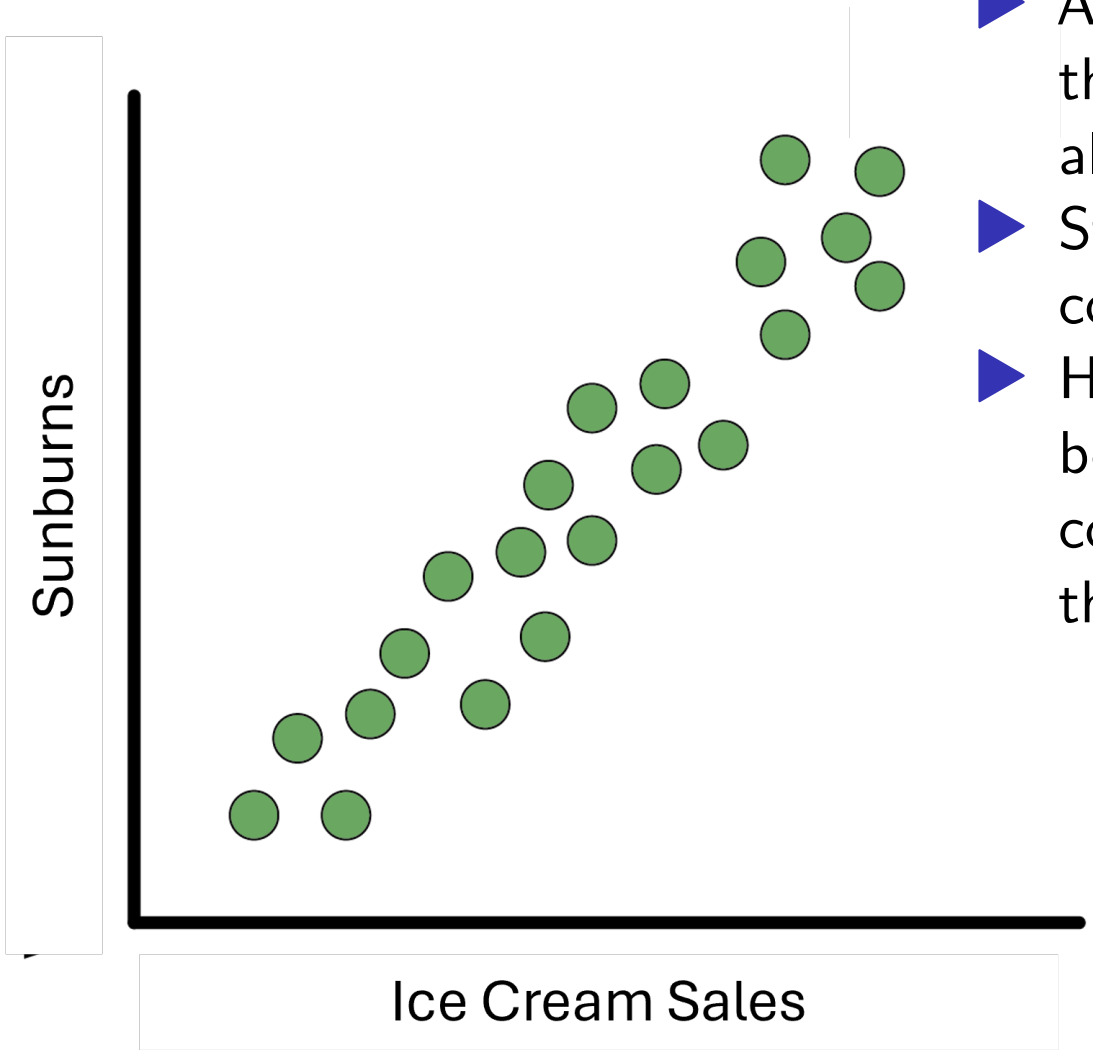


Figure 12: Research Question: Do ice cream sales cause sunburns?

Special Topic: Correlation vs Causation



- As ice cream sales increase, the number of sunburns also increases
- Strong, positive, linear correlation
- High temperatures affect both ice cream consumption and behaviors that lead to sunburn

Confounding Variables

When you have a confounding variable, you might find dependence between two unrelated variables that are only connected by the confounder.