Announcements

- HW3 due Monday
- Lab 3 write-up due Monday as well
- Next lab is next Wednesday
- Read Ch 5.3 for Monday!
- Physics Tea today at 3:00pm
- Physics Seminar today on quantum communication and cryptography (3:30pm, Collins 320)

Warm Up Question

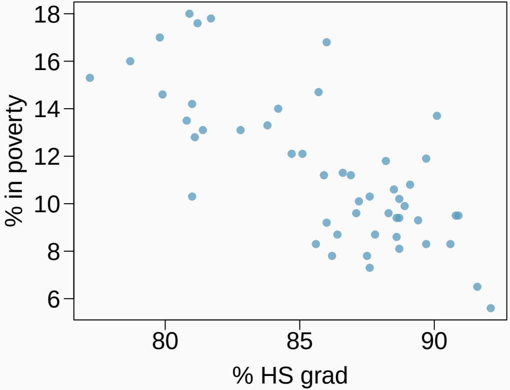
Which of the following is the best guess for the correlation between % in poverty and % HS graduates?

A) 0.6

B) -0.75

C) -0.1

D) -1.5



Warm Up Question

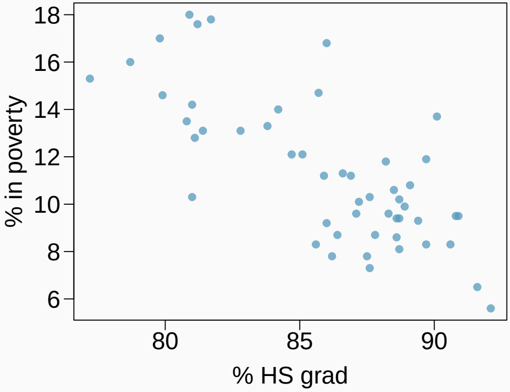
Which of the following is the best guess for the correlation between % in poverty and % HS graduates?

A) 0.6

B) -0.75

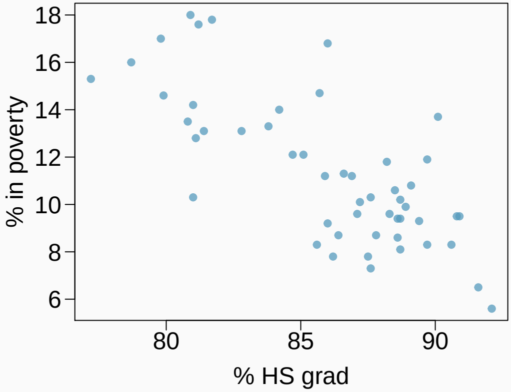
C) -0.1

D) -1.5



Our Starting Point

Least square linear fits have certain properties that we can take advantage of. Given:



	% HS grad (x)	% in poverty (y)
mean	$\bar{x} = 86.01$	$\bar{y} = 11.35$
sd	$s_x = 3.73$	$s_y = 3.1$
	correlation	$R = -0.75$

The Slope!

The Slope

The ***slope*** of the regression can be calculated as

$$b_1 = \frac{s_y}{s_x} R$$

The Slope!

The Slope

The ***slope*** of the regression can be calculated as

$$b_1 = \frac{s_y}{s_x} R$$

Example

In this case:

$$b_1 = \frac{3.1}{3.73}(-0.75) = -0.62$$

The Slope!

The Slope

The ***slope*** of the regression can be calculated as

$$b_1 = \frac{s_y}{s_x} R$$

Example

In this case:

$$b_1 = \frac{3.1}{3.73}(-0.75) = -0.62$$

Interpretation:

For each additional % point in HS graduation rate, we would expect the % living in poverty to be lower on average by 0.62% points.

An Interception!

The ***intercept*** is where the regression line intersects the y-axis.

- Conveniently, it will always pass through the point $(\bar{x}, \bar{y})$

As such, we can define b_0 in terms of b_1 and the means.

The Intercept

The intercept can be written as:

$$b_0 = \bar{y} - b_1 \bar{x}$$

An Interception!

The ***intercept*** is where the regression line intersects the y-axis.

- Conveniently, it will always pass through the point $(\bar{x}, \bar{y})$

As such, we can define b_0 in terms of b_1 and the means.

The Intercept

The intercept can be written as:

$$b_0 = \bar{y} - b_1\bar{x}$$

Example

$$b_0 = 11.35 - (-0.62) \times 86.01 = 64.676$$

Understanding Check

Which of the following is the correct interpretation of the intercept?

A) For each % point increase in HS graduate rate, % living in poverty is expected to increase on average by 64.68%

B) For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%

C) Having no HS graduates leads to 64.68% of residents living below the poverty line.

D) States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.

Understanding Check

Which of the following is the correct interpretation of the intercept?

A) For each % point increase in HS graduate rate, % living in poverty is expected to increase on average by 64.68%

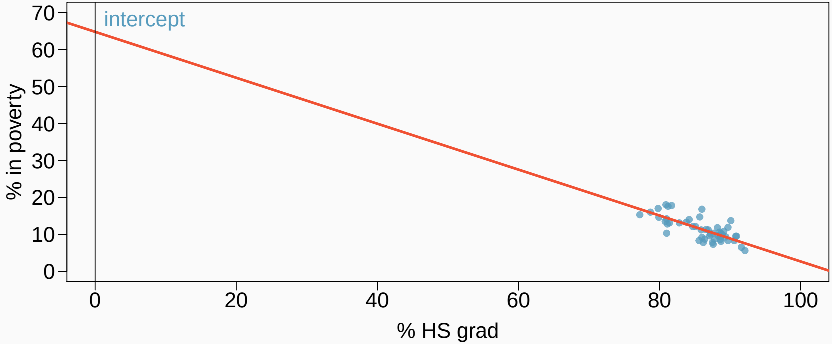
B) For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%

C) Having no HS graduates leads to 64.68% of residents living below the poverty line.

D) States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.

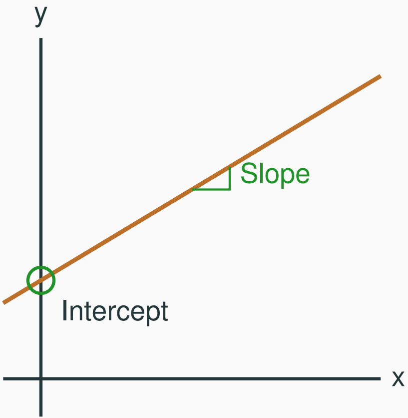
Final Intercept Thoughts

Since there are no states in the dataset with no HS graduates, the intercept is of no interest, not very useful, and not reliable since the predicted value of the intercept is so far from the bulk of the data.



Interpretations

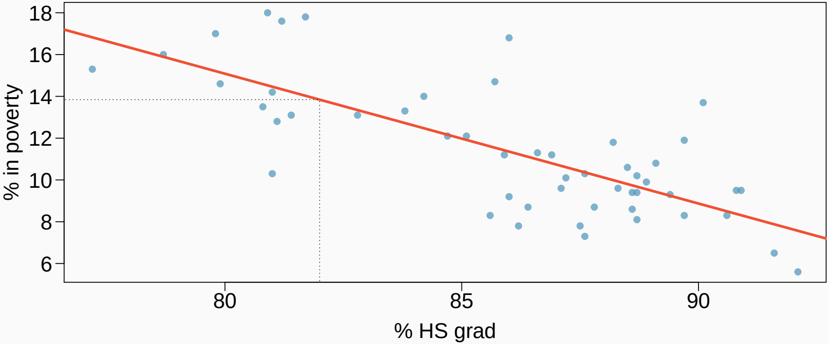
- ***Intercept:*** When $x = 0$, y is expected to equal the intercept
- ***Slope:*** For each unit x, y is expected to increase or decrease on average by the slope



Note: These statements are <u>not</u> causal, unless the study is a randomized and controlled experiment.

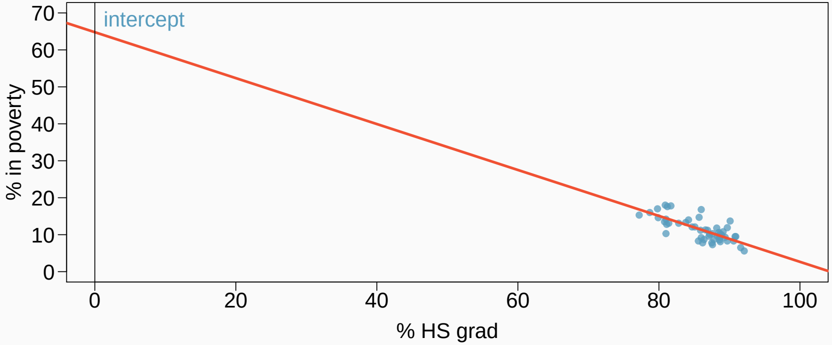
Powers of Prediction

- Using the linear model to estimate the value of the response variable for a given value of the explanatory variable is called ***prediction***.
- There will always still be some uncertainty associated with the predicted value

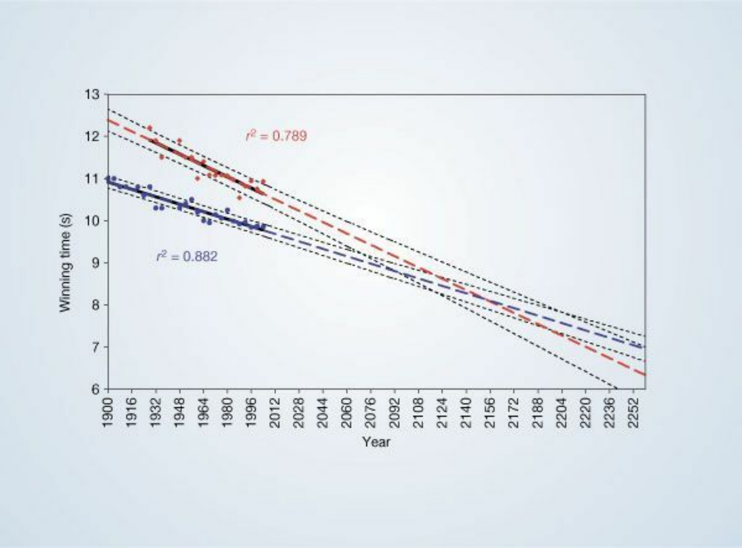


Extrapolation

- Applying a model estimate to values outside the realm of the original data is called ***extrapolation***.
- In some cases (like earlier) the intercept might be an extrapolation.



Extrapolating Examples...



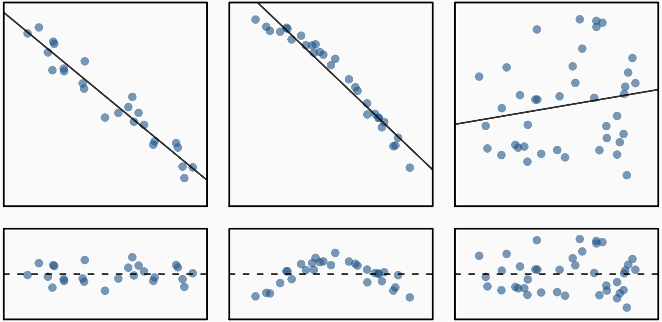
Important Conditions

We need certain things to be true to be safe in applying a least-squares line:

1. Linearity
2. Symmetric, single peaked residuals distribution
3. Constant variability

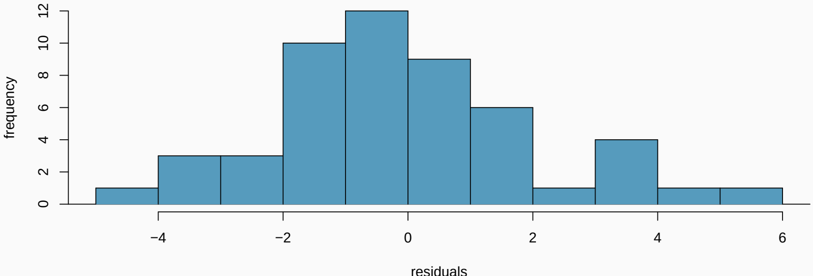
Condition 1: Linearity

- Relationship between explanatory and response variable must be linear
- Non-linear methods exist, but are outside of what we'll talk about in this class
- Check using a scatterplot or a ***residuals plot***.



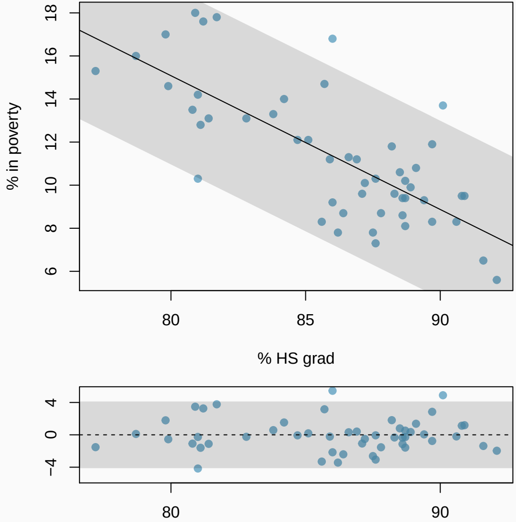
Condition 2: Symmetric (Normal) Residuals

- The residuals distribution should have a single peak and a symmetric falloff on either side (what we'll come to call "normal")
- Condition may fail when there are unusual observations that don't follow the trend of the data
- Check with a histogram



Condition 3: Constant Variability

- The variability of points around the least-squares line should be roughly constant
- Also means the variability of residuals around the 0 line should be roughly constant
- Check with a histogram or plot of residuals



Checking Conditions

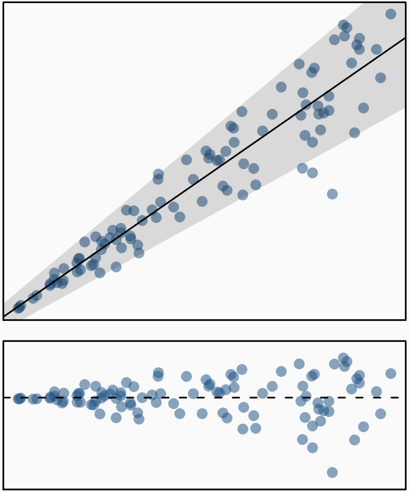
What condition is this linear model obviously violating?

A) Constant variability

B) Linear relationship

C) Symmetric residuals

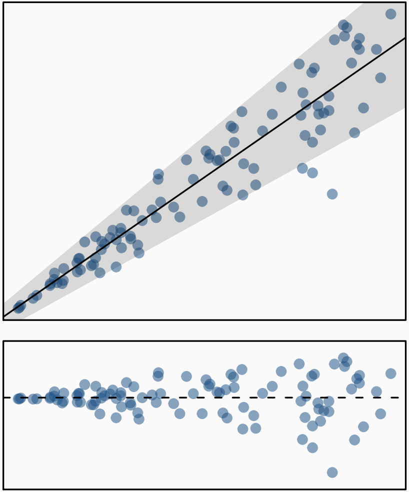
D) No extreme outliers



Checking Conditions

What condition is this linear model obviously violating?

A) Constant variability
B) Linear relationship
C) Symmetric residuals
D) No extreme outliers



R^2

- The strength of the fit of a linear model is most commonly evaluated using R^2.
- R^2 is calculated as the square of the correlation coefficient.
- Tells us what percent of variability in the response variable is explained by the model.
- The remainder variability is explained by variables not in the model or by inherent randomness

Example

In the model we've been looking at

$$R = -0.75$$

and so

$$R^2 = (-0.75)^2 = 0.56$$