Introduction to Linear Regression

Phase 3

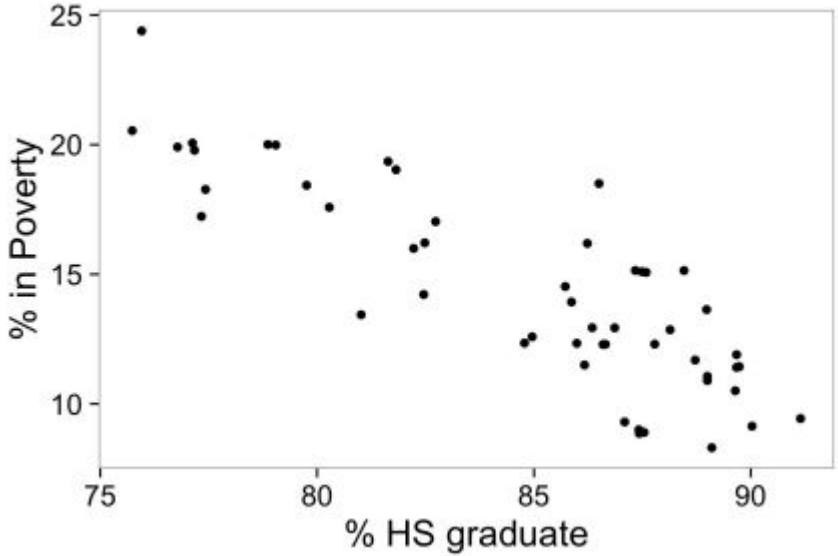
Key ideas

1. In previous lectures, we have learned how to **statistically** assess whether **two** different **variables** have any **relationship** or not.
2. In this lecture, we are going to generalize this by building a (parametric) model that represents a **linear** relationship between two variables.
3. Such a model is called **linear regression**.
4. This model defines a line that tries to **minimize** the **distance** between the **predicted** and **observed** values.

Today example

Relationship between high school graduation rate and the percent of residents who live below the poverty line in all 50 US states + DC in 2012. (income below $23,050 for a family of 4).

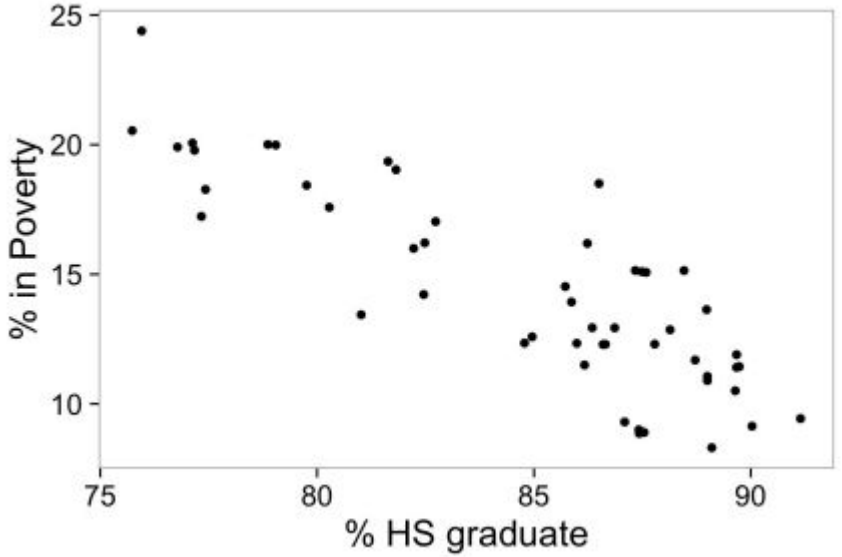
How would you describe this relationship?



Guess the correlation

Which of these is your best guess for the correlation between poverty and high school graduation?

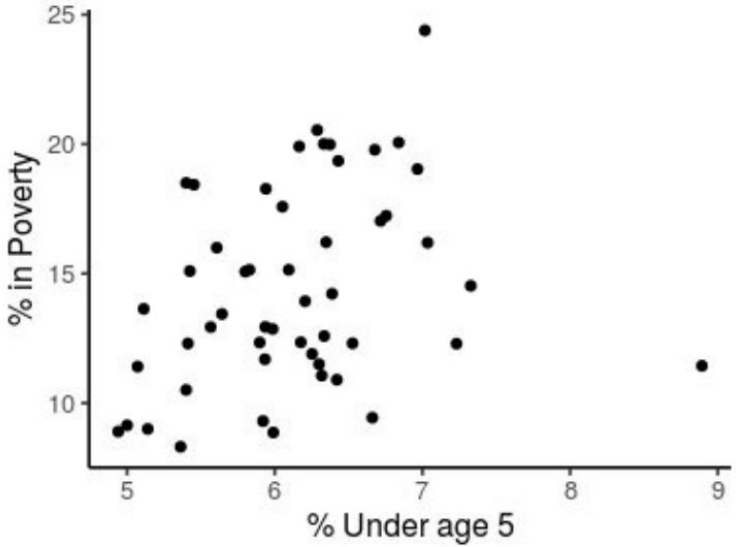
a) 0.6
b) -0.85
c) 0.02
d) 1.5



Guess the correlation

Which of these is your best guess for the correlation between poverty and and the proportion of the population under 5 years of age?

a) -0.3
b) 0.01
c) 1.1
d) 0.3



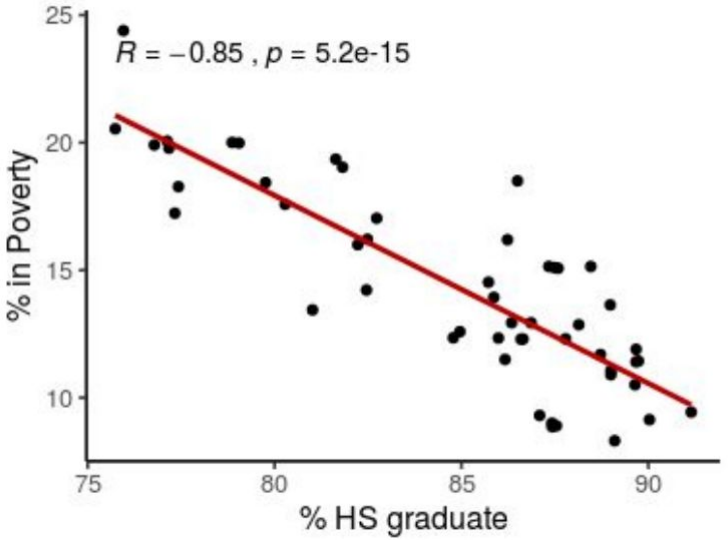
Relationship between two variables

- In all these previous examples, whenever we think about the relationship between two variables, the first thing that comes to our mind is a line.

- This line tries to be as close as possible to our points (Fit), although there is always some error (Residual).

  In other words:

  DATA = FIT + RESIDUAL



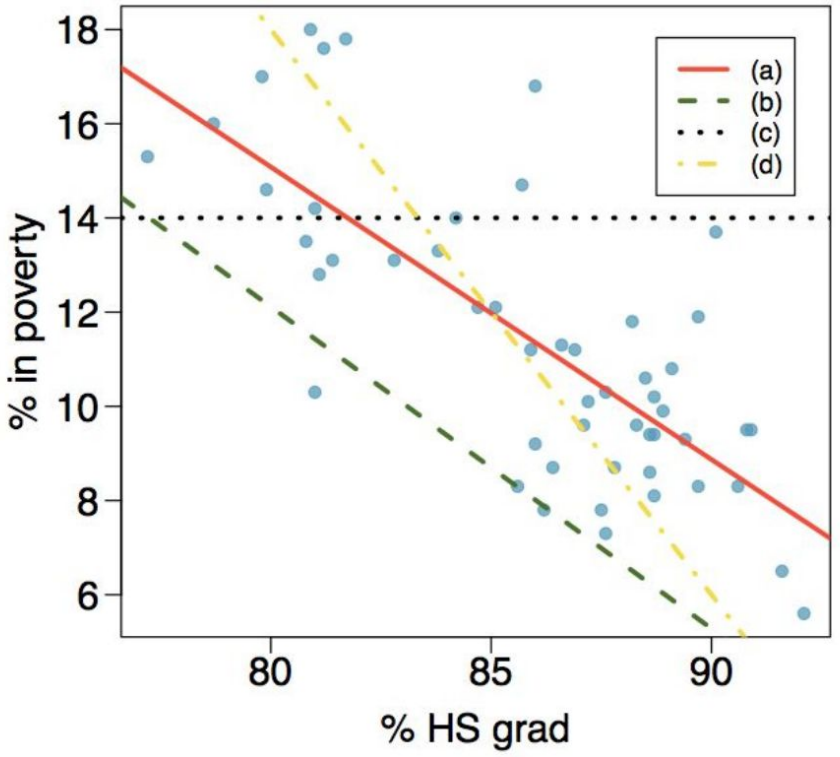
Which of these lines best represents the trend?

a)

b)

c)

d)

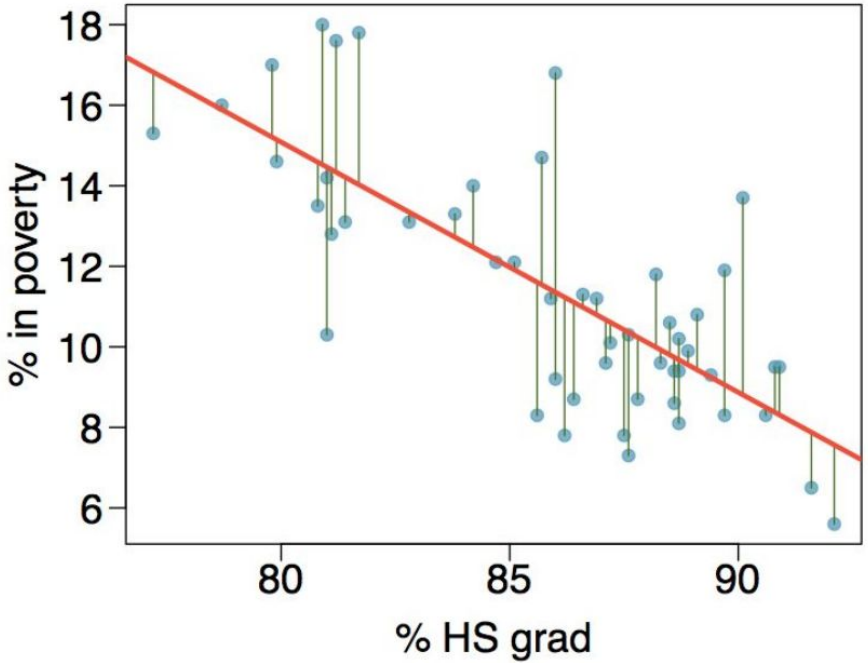


Linear regression model

- We want to find the **line** that **minimizes** the **residuals**: the distances between each point and the line.
- This is basically a **regression model,** which is composed of two things:

  (1) A best-fit line, and

  (2) the residuals between each point and the line.



Linear regression model

A **linear regression model** is just the line that tries to address the (linear) relationship between two variables.

$$Y_i = \alpha + \beta_i \cdot X_i + \epsilon_i$$

Residuals

Outcome variable
dependent variable
response variable

Intercept

Slope

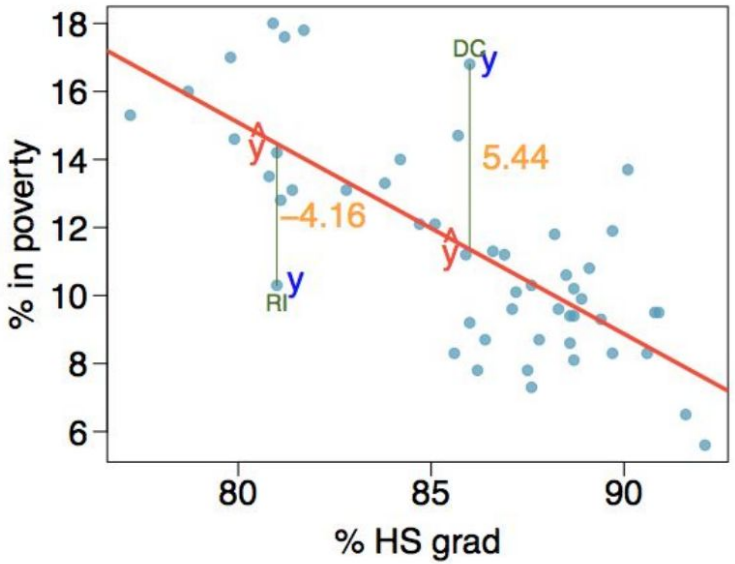
Input variable
independent variable
predictor variable

What is a residual?

A **residual** is the difference between the observed and the estimated regression line.

Percent living in poverty in DC is 5.44% more than predicted based on HS grad % alone.

Percent living in poverty in RI is 4.16% less than predicted.



Estimating the linear regression model

- Given some input data, *X*, a regression model allows you to generate prediction $\hat{Y}$.

Estimating the linear regression model

- Given some input data, *X*, a regression model allows you to generate prediction $\hat{Y}$.
- We want to find $\hat{\alpha}$ and $\hat{\beta}$ that give the **smallest residuals**.

$$\sum_i \epsilon_i^2 = \sum_i (Y - \hat{Y}_i)^2$$

Estimating the linear regression model

- Given some input data, *X*, a regression model allows you to generate prediction $\hat{Y}$.
- We want to find $\hat{\alpha}$ and $\hat{\beta}$ that give the **smallest residuals**.

$$\sum_i \epsilon_i^2 = \sum_i (Y - \hat{Y}_i)^2$$

Ordinary Least Squares

$$\hat{\alpha} = < Y > - \hat{\beta} < X >$$

$$\hat{\beta} = r \frac{\sigma_y}{\sigma_x}$$

Understanding the linear regression model

- Once we have estimated the regression model, we can make **predictions** on a new X**.**

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} \cdot X_i$$

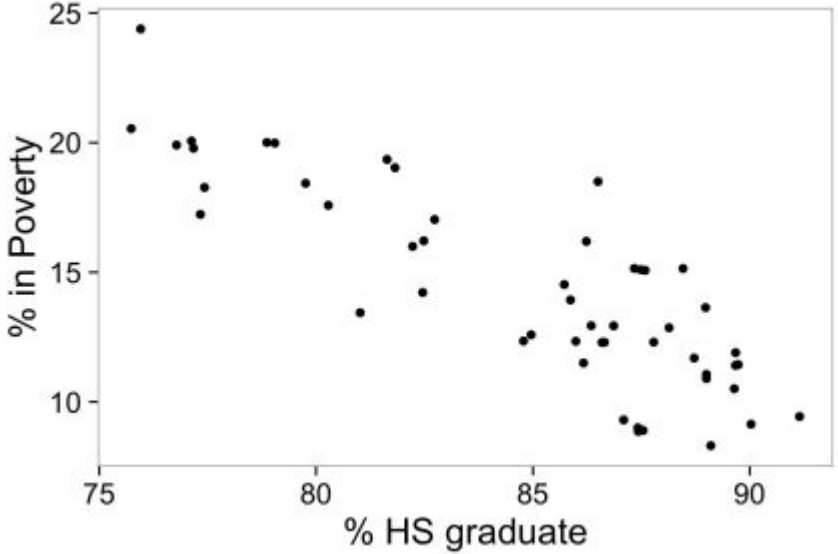
- Slope**:** For each unit increase in X, Y is expected to be **higher/lower** on average by the **slope**.

$$\hat{\beta} = r \frac{\sigma_y}{\sigma_x}$$

- Intercept: When X=0 , the intercept is the expected value of Y.

$$\hat{\alpha} = < Y > - \hat{\beta} < X >$$

Interpreting the linear regression model



	% 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	$R = -0.75$

Interpreting the linear regression model: slope

The slope of the regression: $\hat{\beta} = r\frac{\sigma_y}{\sigma_x}$

For this problem:

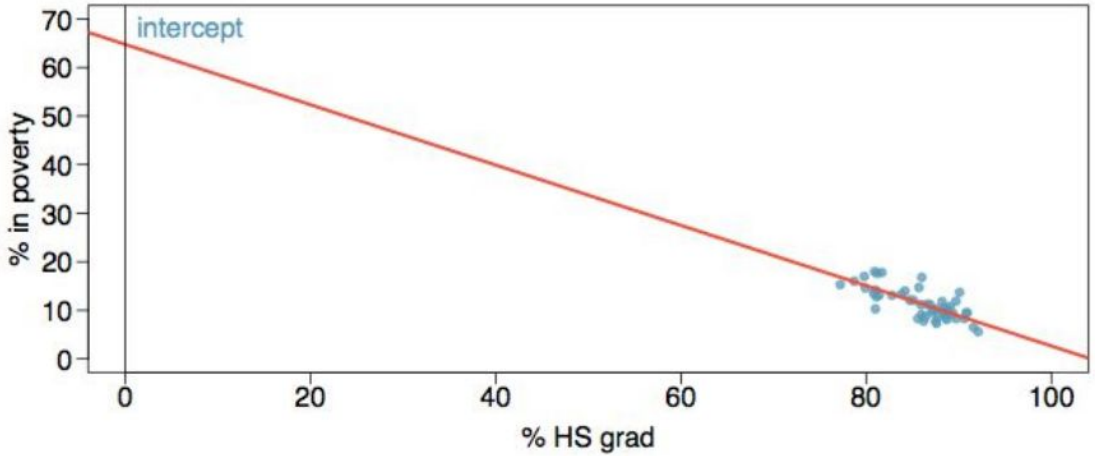
$$\hat{\beta} = \frac{3.1}{3.73} \times -0.75 = -0.62$$

	% 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	$R = -0.75$

Interpretation: For each additional % increase in HS graduate rate, we would expect the % living in poverty to be lower on average by 0.62.

Interpreting the linear regression model: intercept

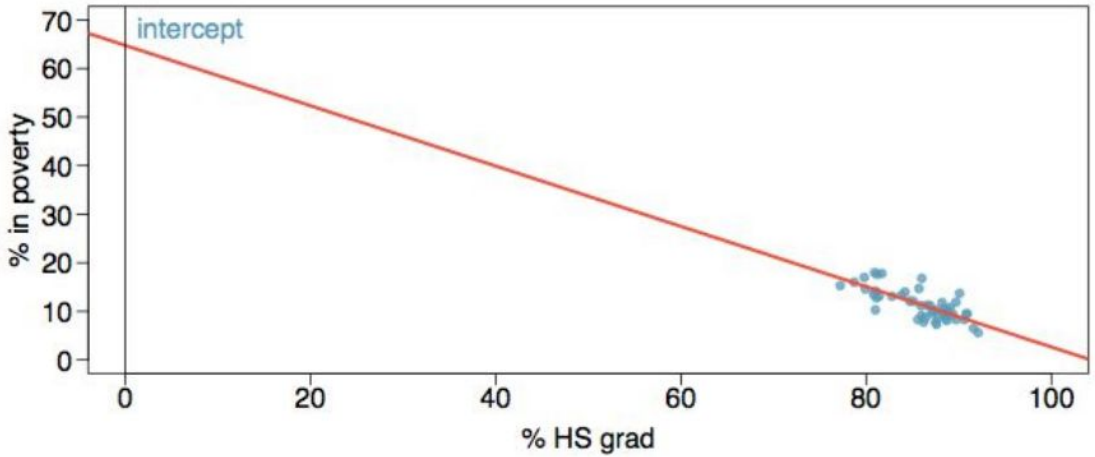
The intercept is where the line crosses the y-axis: $\hat{\alpha} = < Y > - \hat{\beta} < X >$



	% 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	$R = -0.75$

Interpreting the linear regression model: intercept

The intercept is where the line crosses the y-axis: $\hat{\alpha} = < Y > - \hat{\beta} < X >$



	% 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	$R = -0.75$

$$\hat{\alpha} = 11.35 - (0.62) \times 86.01 = 64.68$$

Practice question 1

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.

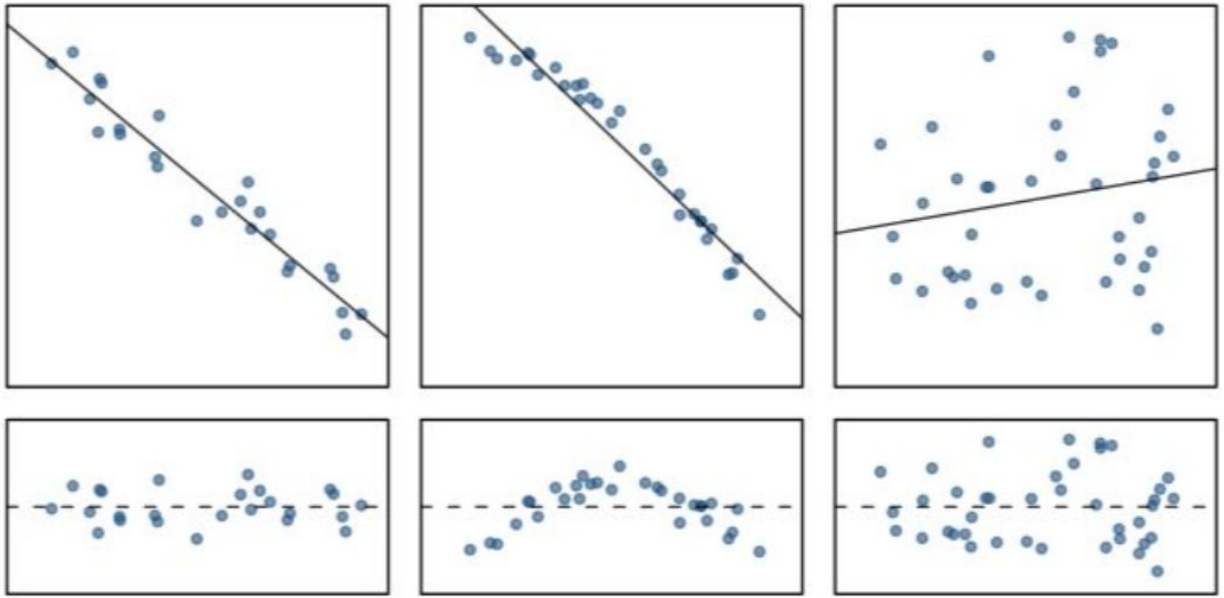
(d) States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.

Assumptions

- **Linearity**. The relationship between X and Y should actually be linear!
- **Normality of residuals**: (It's actually okay if the X and Y are non-normal, as long as the residuals are normal).
- **Homogeneity of variance**. Residuals are generated from a normal distribution with mean 0, and with a standard deviation that is the same for every single residual. Practically, this can be done by checking that the standard deviation of the residual is the same for all values of $\hat{Y}$.
- **No extreme outliers:** Data points very far away from the rest can exert undue influence on the model parameters.

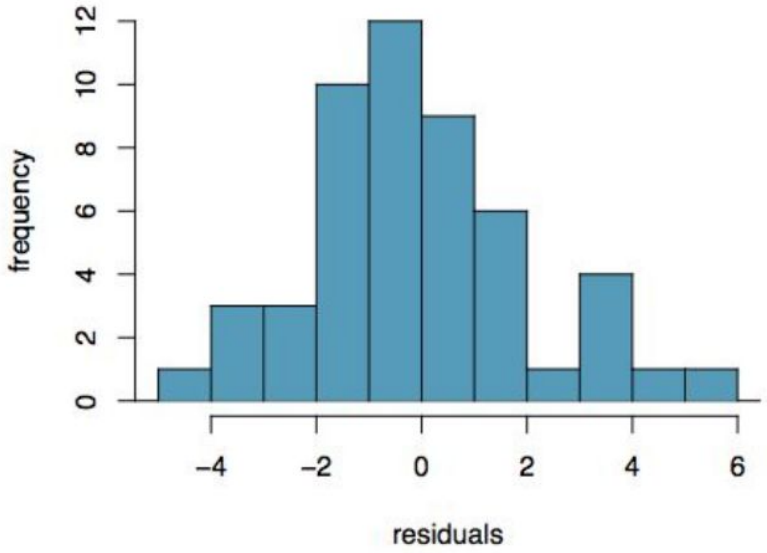
Assumption 1: Linearity

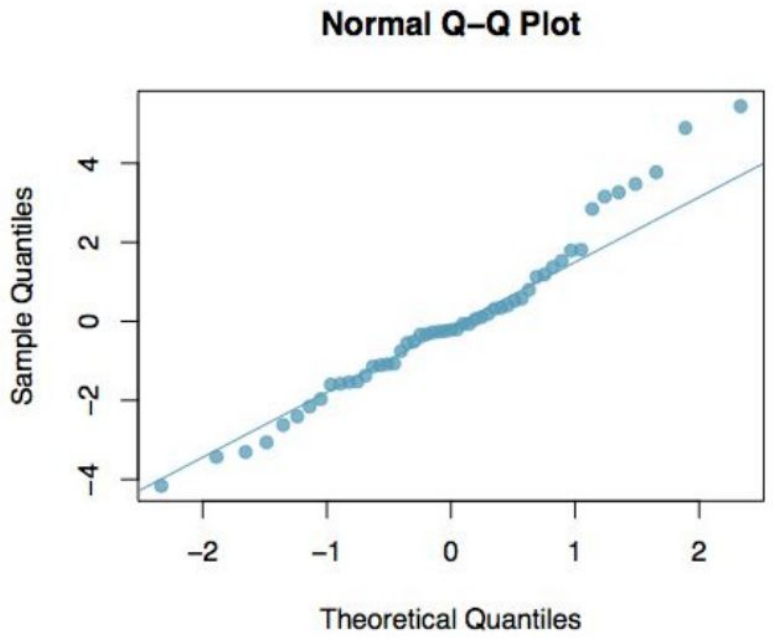
Check using scatterplots, or **residuals** plots (We'll see how to generate these in the tutorials).



Assumption 2: Normality

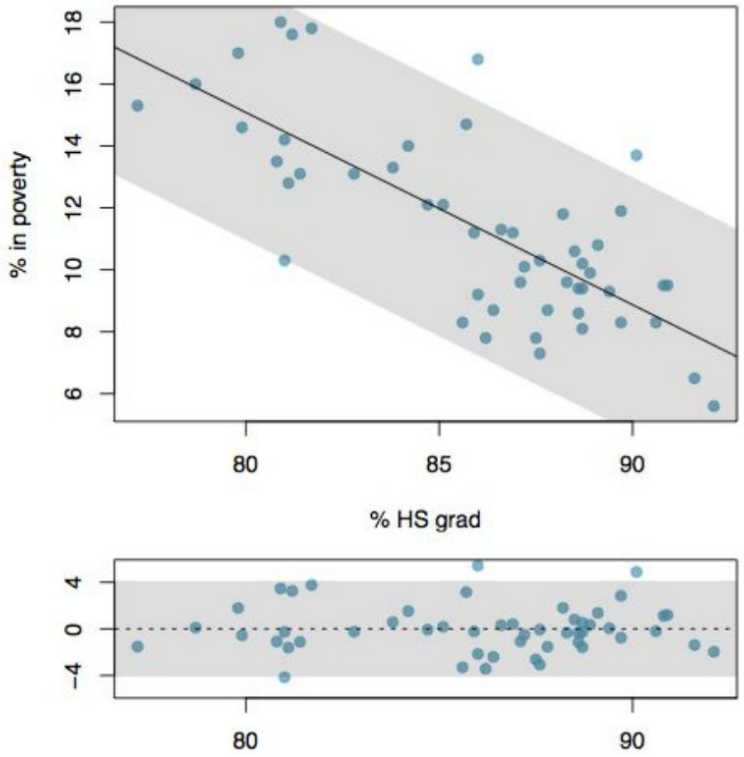
Check using a histogram or normal probability plot of **residuals**.





Assumption 3: Homogeneity of Variance

- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.
- Also called **homoscedasticity**.
- Check using a histogram or normal probability plot of residuals.



Practice question 2

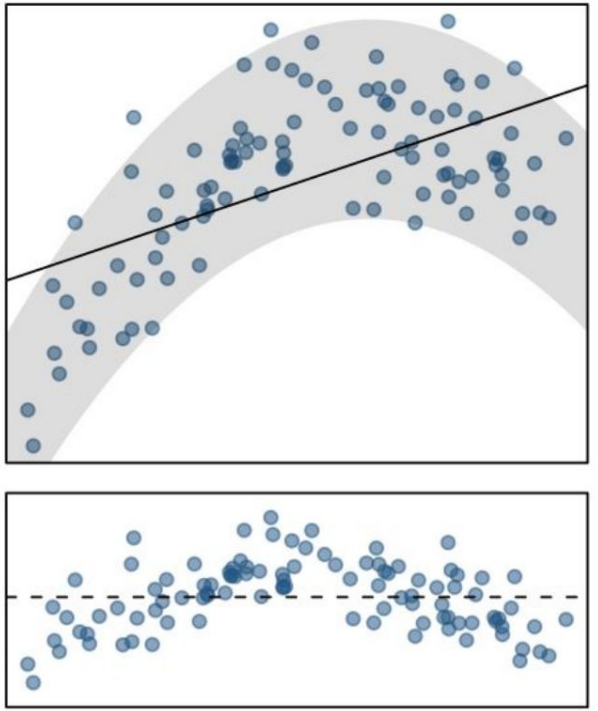
Which condition is this model violating?

(a) Constant variability

(b) Linear relationship

(c) Normal residuals

(d) No extreme outliers



Practice question 3

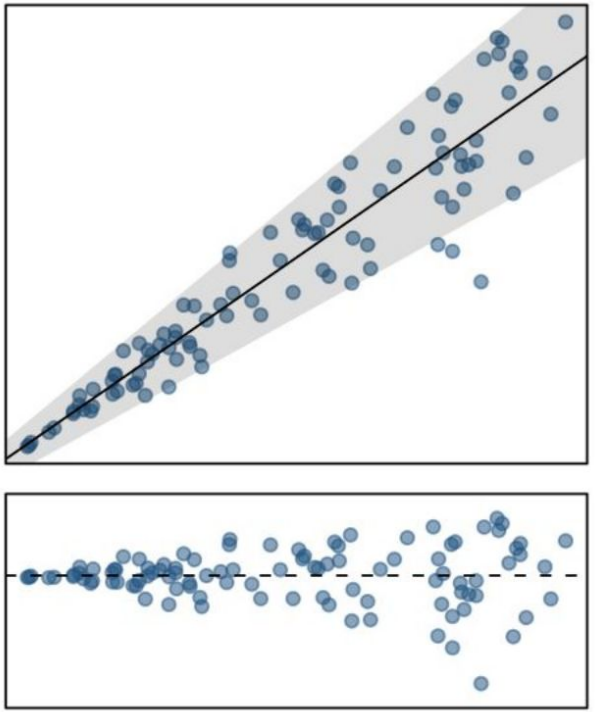
Which condition is this model violating?

(a) Constant variability

(b) Linear relationship

(c) Normal residuals

(d) No extreme outliers



Coefficient of determination

- The performance of the fit of a regression model is most commonly evaluated using the coefficient of determination, R^2.
- R^2 is calculated as the square of the correlation coefficient: it tells us the percent of variability in the response variable explained by the model.
- The remainder of the variability is explained by variables not included in the model or by inherent randomness in the data.
- For the model we've been working with, $R^2 = (-0.75)^2 = 0.56$.

Coefficient of determination

- The performance of the fit of a regression model is most commonly evaluated using the coefficient of determination, R^2.
- R^2 is calculated as the square of the correlation coefficient: it tells us the percent of variability in the response variable explained by the model.
- The remainder of the variability is explained by variables not included in the model or by inherent randomness in the data.
- For the model we've been working with, $R^2 = (-0.75)^2 = 0.56$.

How to account for the rest of variability? → Multiple linear regression (next lecture)

Practice question 4: Interpreting R^2

Which of the following is the correct interpretation of R^2 =0.56?

(a) 56% of the variability in the % of HS graduates among the 51 states+DC is explained by the model.

(b) 56% of the variability in the % of residents living in poverty among the 51 states+DC is explained by the model.

(c) 56% of the time % HS graduates predict % living in poverty correctly.

(d) 75% of the variability in the % of residents living in poverty among the 51 states+DC is explained by the model.

Recap

- We can use **linear regression** to estimate the relationship between **two variables**.
- A **regression line** is the line that **minimizes** the **residuals** between each point and the line.
- We can use the **slope** and **intercept** of a regression line to make **predictions.**
- Like other statistical tools explored so far, linear regression models are **appropriate** only when some **conditions** are **met**.