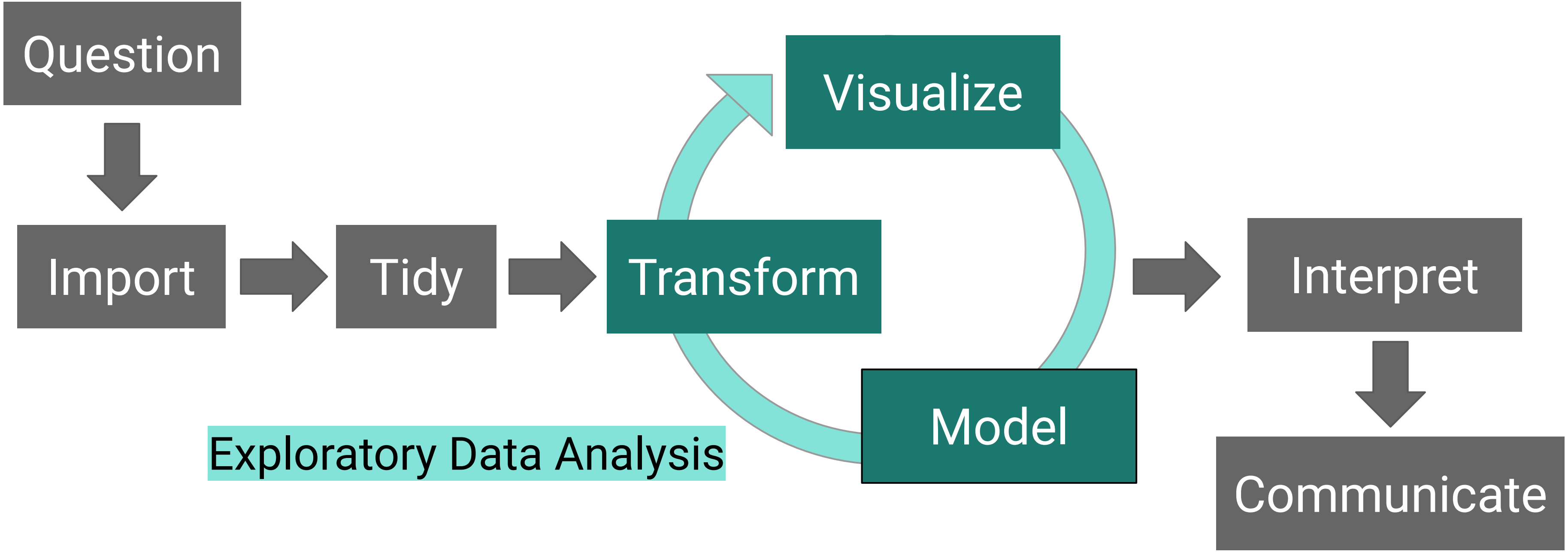
Statistical Modeling I

Lecture 11

Motivation



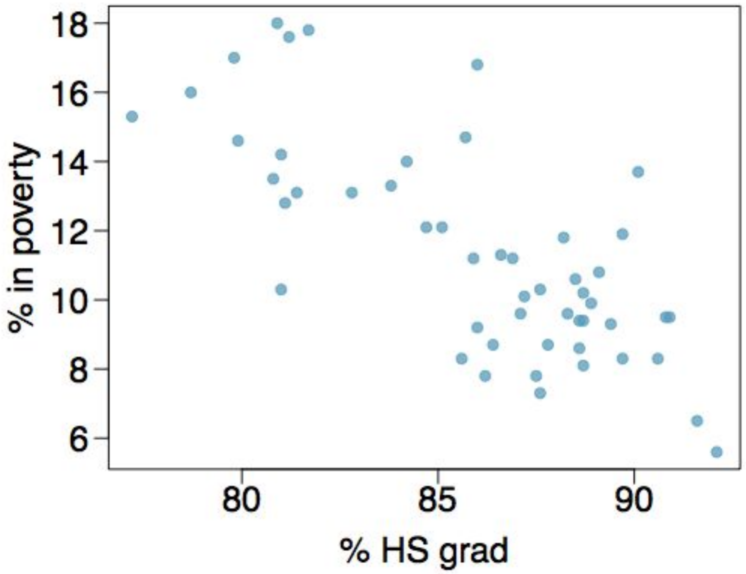
Modified from https://r4ds.had.co.nz/explore-intro.html#explore-intro

Linear Regression Models

- To quantify the relationship between two numerical variables
- To model numerical response or outcome variables using a numerical or categorical explanatory variable

Poverty vs HS graduate rate

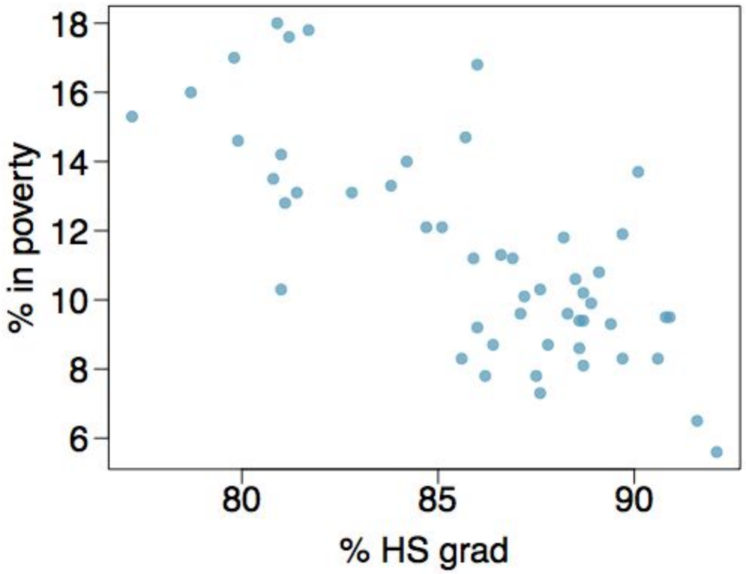
The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the percent of residents who live below the poverty line.



Response variable?

Poverty vs HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the percent of residents who live below the poverty line.

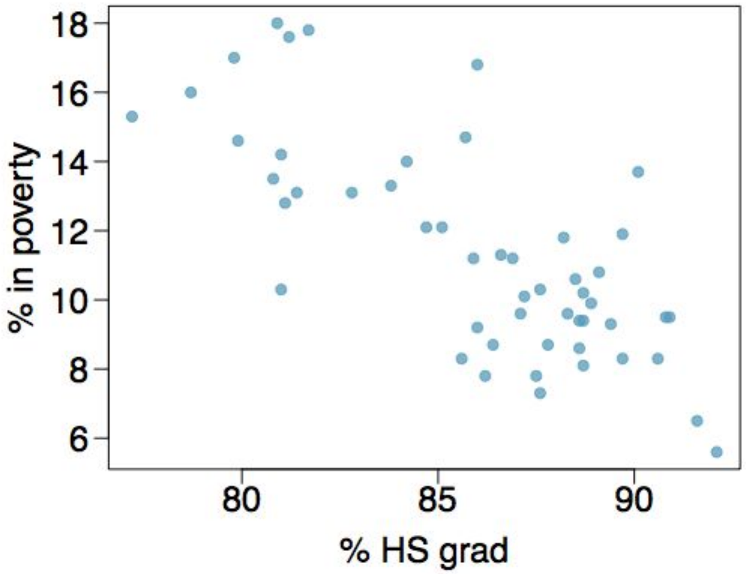


Response variable?

% in poverty

Poverty vs HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the percent of residents who live below the poverty line.



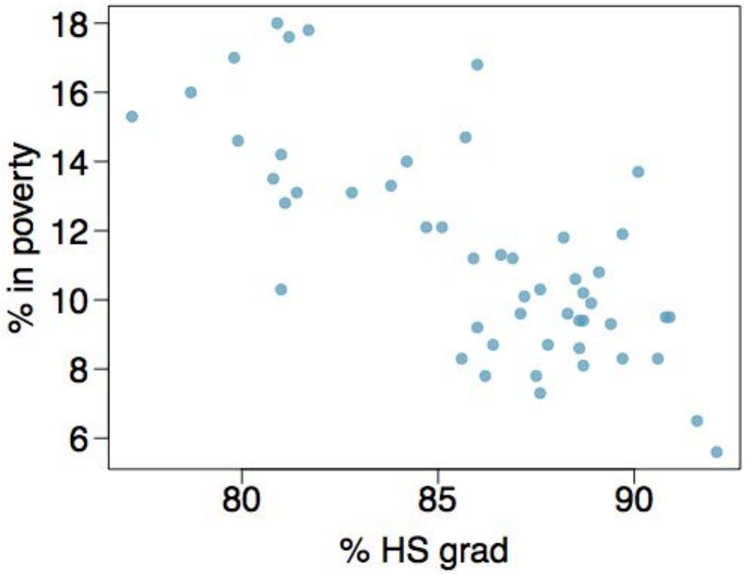
Response variable?

% in poverty

Explanatory variable?

Poverty vs HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the percent of residents who live below the poverty line.



Response variable?

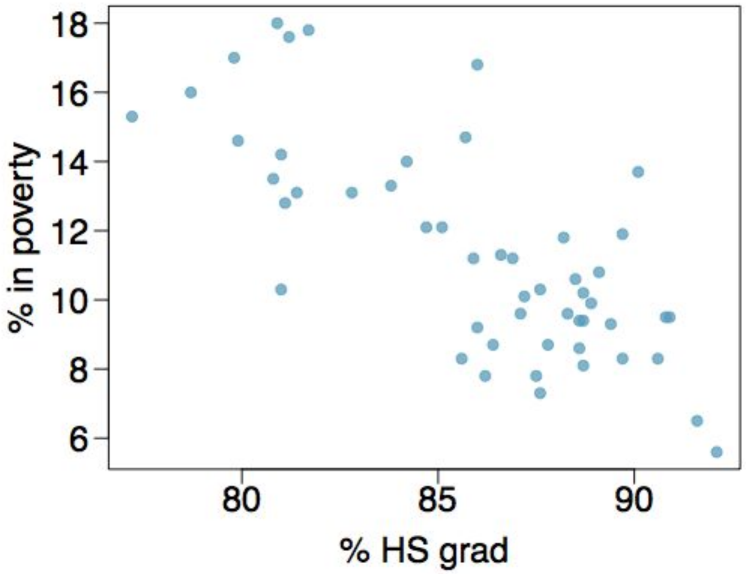
% in poverty

Explanatory variable?

% HS grad

Poverty vs HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the percent of residents who live below the poverty line.



Response variable?

% in poverty

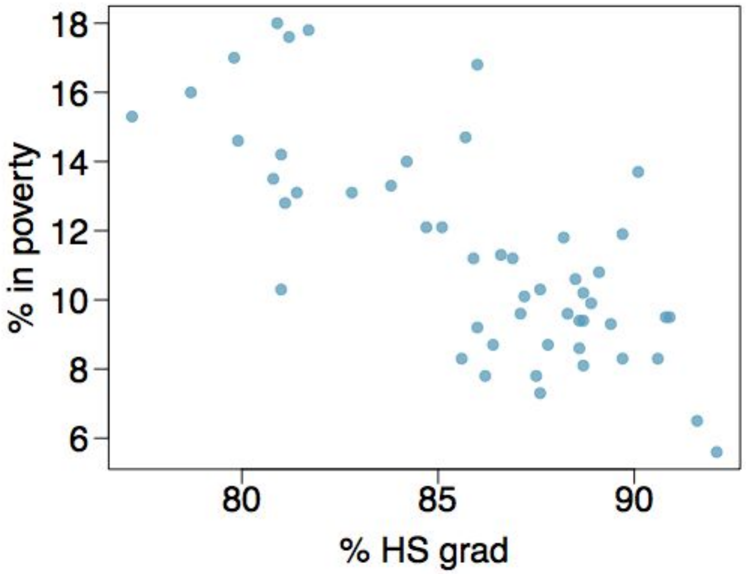
Explanatory variable?

% HS grad

Relationship?

Poverty vs HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the percent of residents who live below the poverty line.



Response variable?

% in poverty

Explanatory variable?

% HS grad

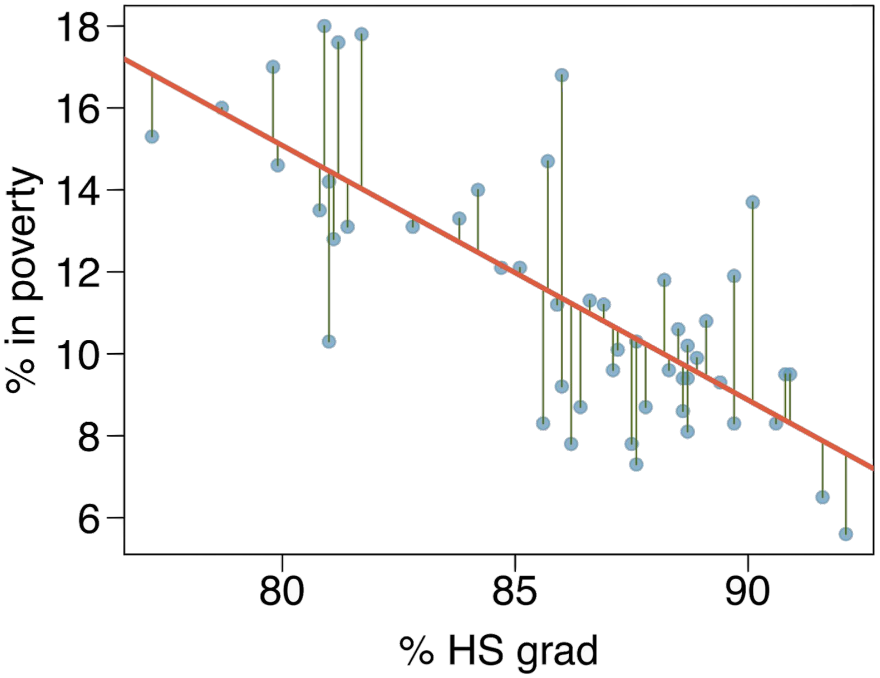
Relationship?

linear, negative, moderately strong

Fitting the curve

Residuals are the leftovers from the model fit:

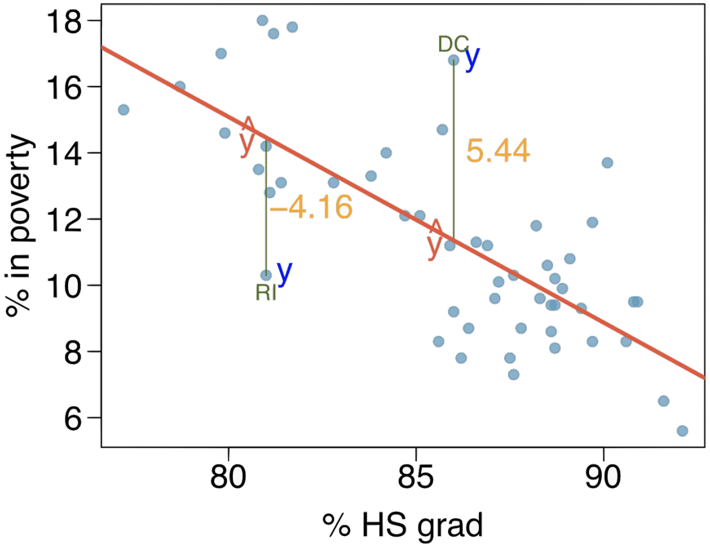
Data = Fit + Residual



Fitting the curve

Residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

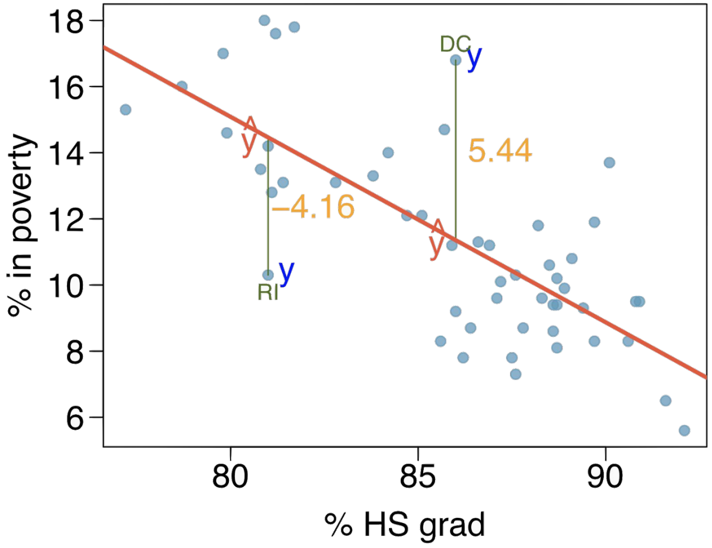
$$e_i = y_i - \hat{y}_i$$



Fitting the curve

Residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

$$e_i = y_i - \hat{y}_i$$

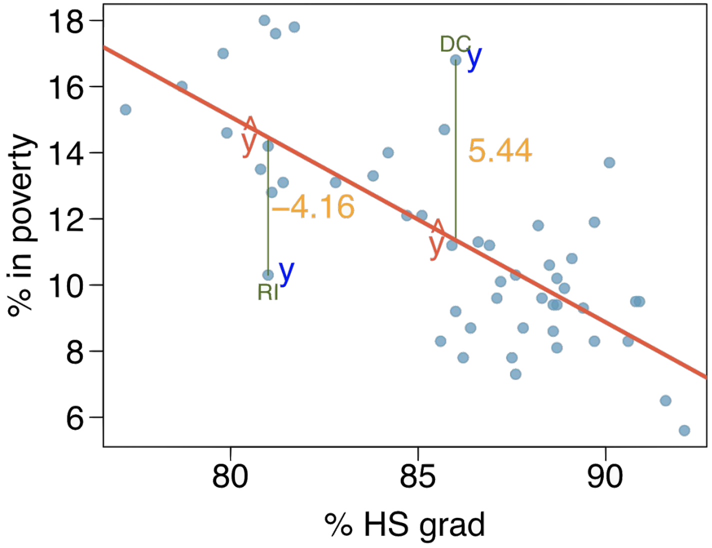


% living in poverty in DC is 5.44% more than predicted.

Fitting the curve

Residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

$$e_i = y_i - \hat{y}_i$$



% living in poverty in DC is 5.44% more than predicted.

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

Linear regression models

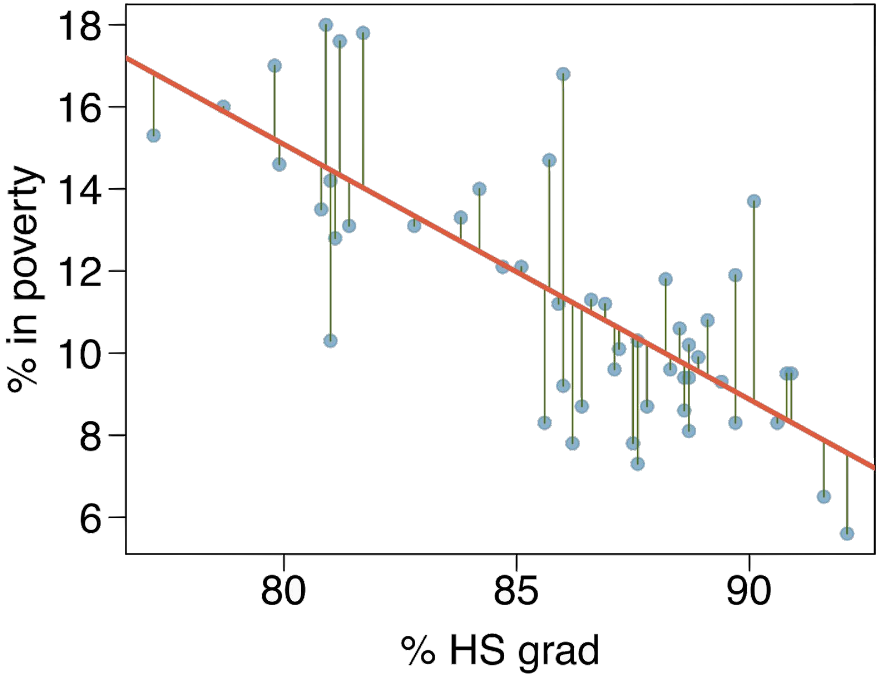
- linear models can be used for prediction or to evaluate whether there is a linear relationship between two numerical variables
- “fitting a straight line through data points”: use x (explanatory or predictor variable) to predict y (response)

$$y = \beta_o + \beta_1 x$$

- linear model parameters (β_o , β_1) are estimated using data

Poverty vs HS graduate rate

The linear model for predicting poverty from high school graduation rate in the US:



$$y = \beta_o + \beta_1 x$$

$$\hat{poverty} = 64.78 - 0.62 * HS_{grad}$$

Babies dataset in R

- The Child Health and Development Studies (USA) investigated the pregnancy in women in San Francisco, 1960-1967. Studied the relationship between mothers who smoked and weight of their babies.
- “babies” dataset available in “openintro” package in R
- 1,236 observations (rows) x 8 variables (columns)

variable	description
case	id number
bwt	birthweight, ounces
gestation	length of gestation, days
parity	binary indicator for a first pregnancy (0=first pregnancy)
age	mother’s age, years
height	mother’s height, inches
weight	mother’s weight, pounds
smoke	binary indicator whether the mother smoked, 1=smoker

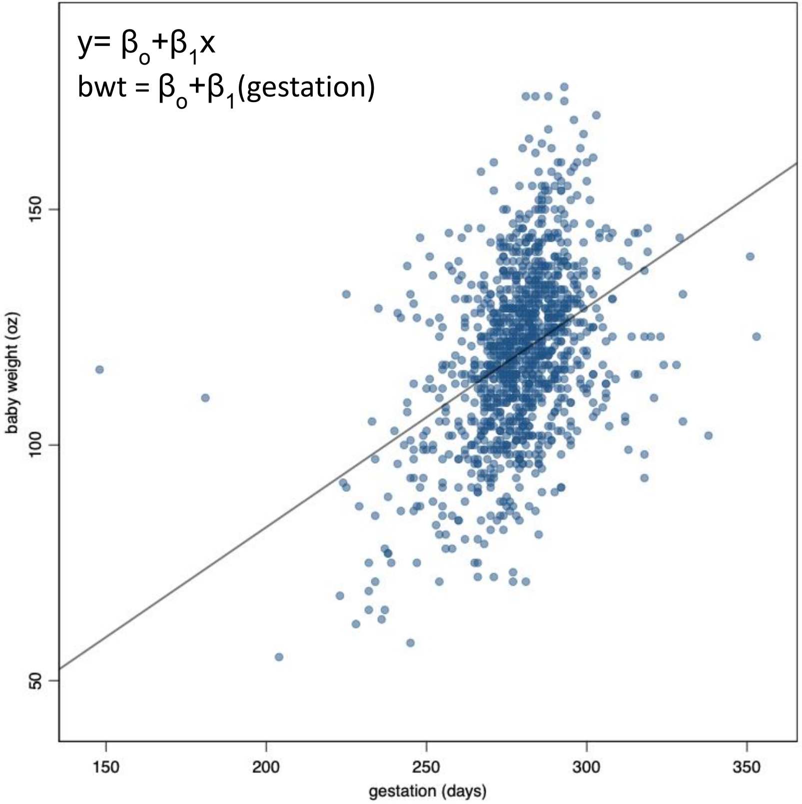
Linear regression model

$y = \beta_o + \beta_1 x$

lm(formula = y ~ x)

lm(formula = response ~ explanatory)

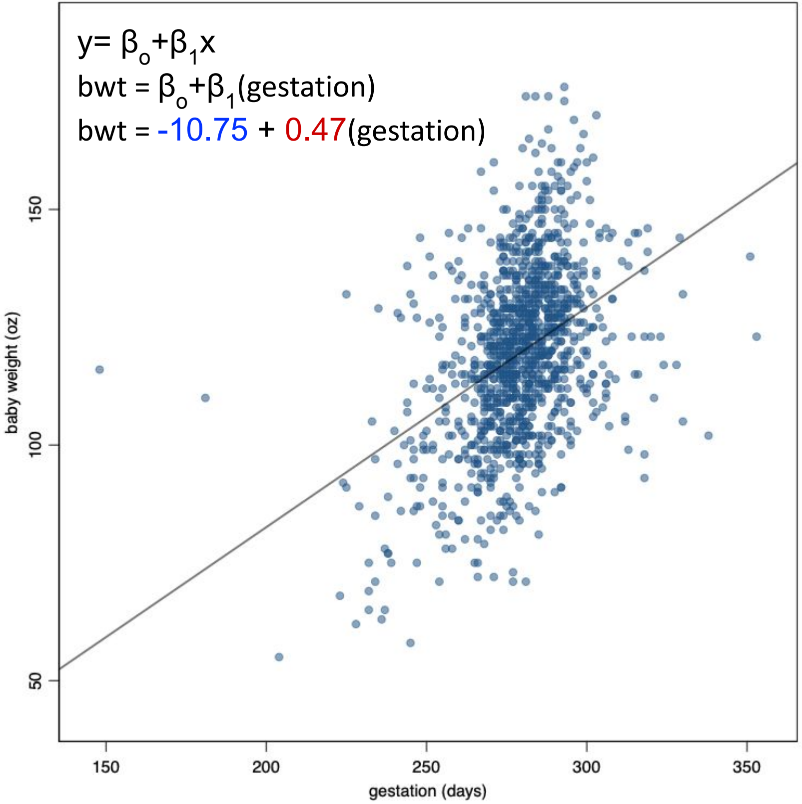
```
# Linear regression model on birthweight (oz) and length of gestation (days)
uni <- babies %>%
  lm(formula = bwt ~ gestation)
```



Linear regression model

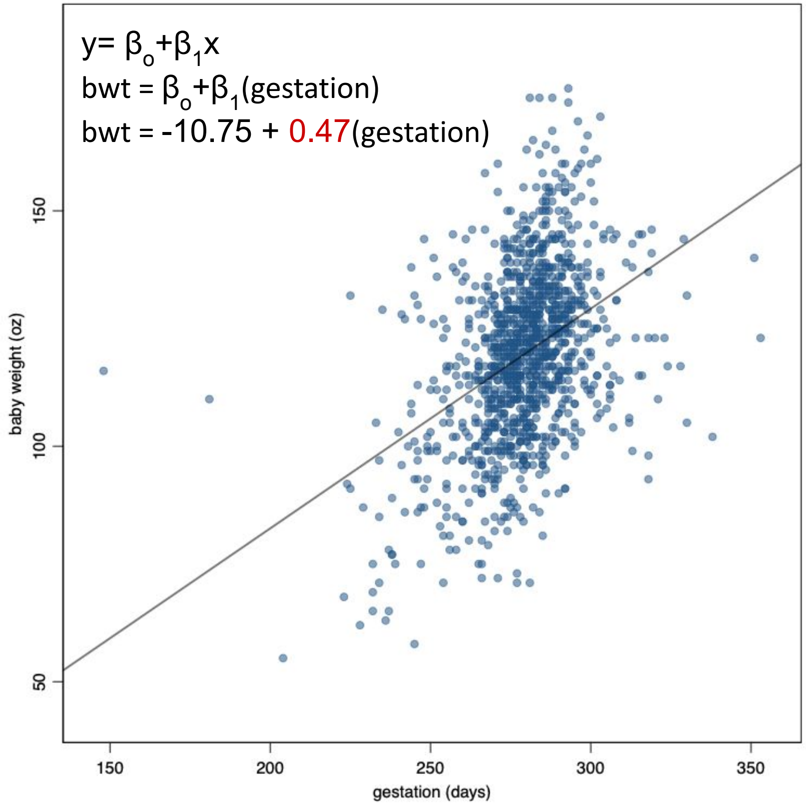
```
# Summarize linear model result
summary.lm(uni)
```

```
Call:
lm(formula = bwt ~ gestation, data = .)
Residuals:
   Min  1Q  Median     3Q      Max
-49.348 -11.065   0.218  10.101  57.704
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept) -10.75414    8.53693     -1.26         0.208
gestation        0.46656      0.03054     15.28      <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 16.74 on 1172 degrees of freedom
Multiple R-squared:  0.1661,  Adjusted R-squared:  0.1654
F-statistic: 233.4 on 1 and 1172 DF,  p-value: < 2.2e-16
```



Linear regression model

"The model predicts a 0.47 oz increase in average birth weight of newborns for each additional day of pregnancy."



Inference for regression

Is there a linear relationship between y and x?

- H_0: $\beta_1 = 0$

  The true linear model has a slope equal to zero.

- H_A: $\beta_1 \neq 0$

  The true linear model has a slope not equal to zero.

Inference for regression

Is there a linear relationship between birthweight and gestation?

bwt = $\beta_o + \beta_1$(gestation)

- H_0: $\beta_1 = 0$

  The true coefficient of gestation is zero.

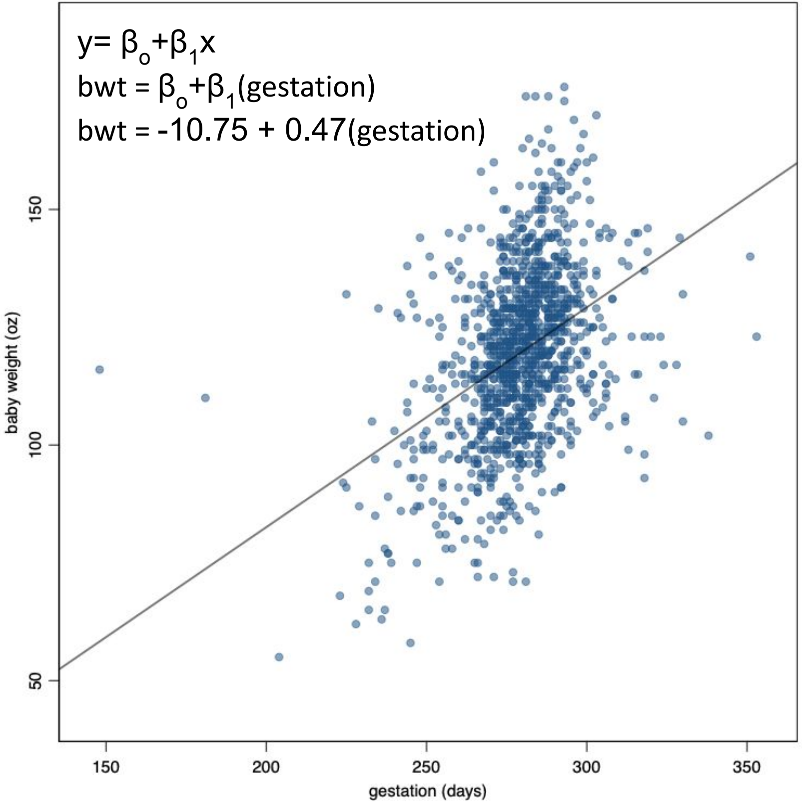
- H_A: $\beta_1 \neq 0$

  The true coefficient of gestation is not equal to zero.

Inference for regression

- p-value < 0.05, reject null and accept the alternative hypothesis
- the data provide strong evidence that the slope parameter (β_1) is not equal to zero.

```
Call:
lm(formula = bwt ~ gestation, data = .)
Residuals:
   Min  1Q  Median      3Q       Max
-49.348 -11.065   0.218  10.101  57.704
Coefficients:
                    Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   -10.75414    8.53693     -1.26       0.208
gestation         0.46656    0.03054      15.28     <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Coefficient of determination (R^2)

- coefficient of determination
- measure of how well the model is fitting the actual data (variance)
- R^2 = explained/total variation, [0-1]
- "About 16% of the variance in baby weight is explained by gestation."

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-10.75414	8.53693	-1.26	0.208
gestation	0.46656	0.03054	15.28	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

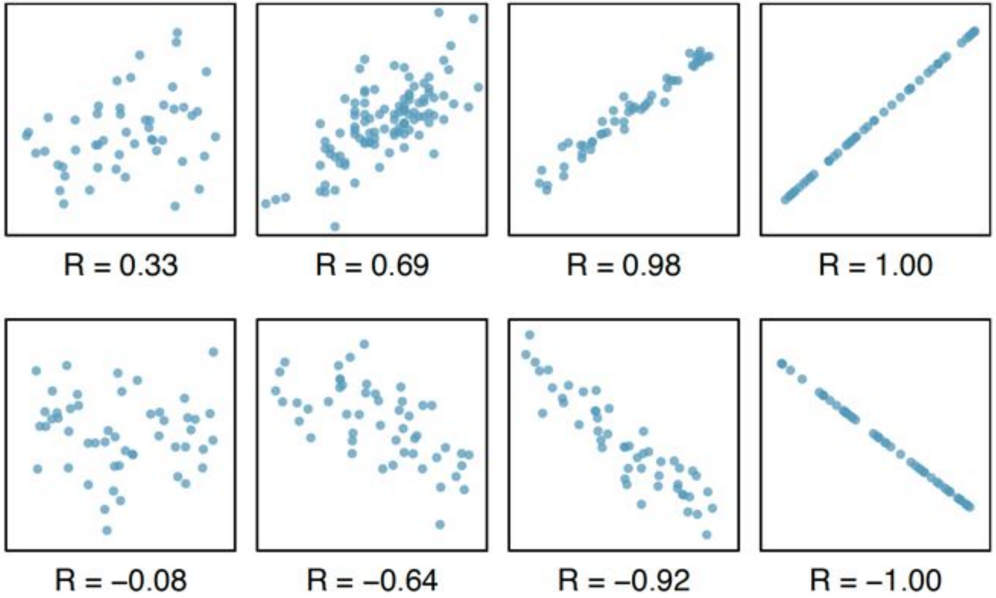
Residual standard error: 16.74 on 1172 degrees of freedom

Multiple R-squared: 0.1661, Adjusted R-squared: 0.1654

F-statistic: 233.4 on 1 and 1172 DF, p-value: < 2.2e-16

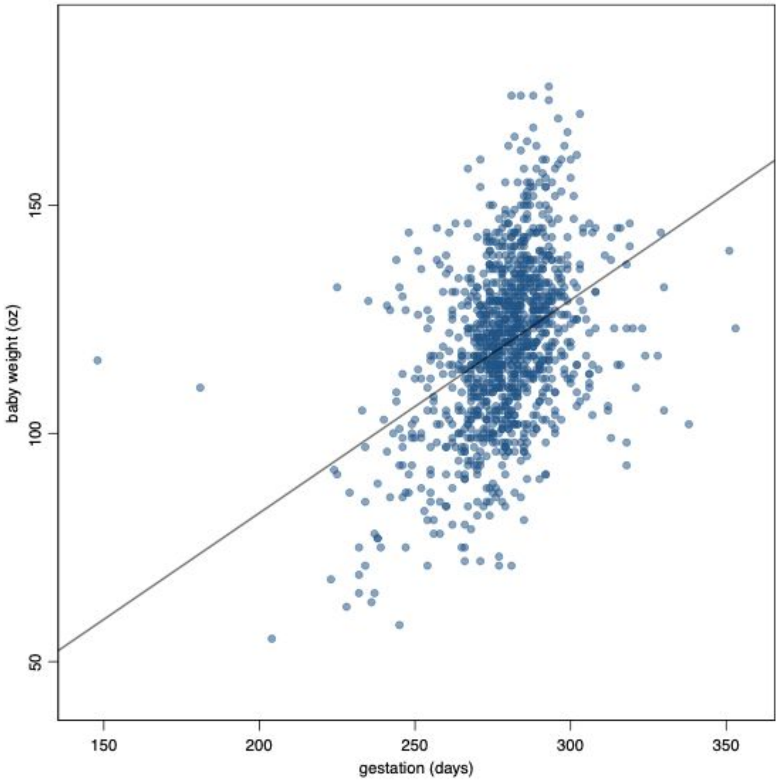
Correlation coefficient (R)

- correlation coefficient
- describes the strength of linear relationship between two variables
- R values [-1, 1]



Correlation coefficient (R)

- R = 0.41
- Birthweight and gestation are positively correlated



Correlation

- correlation is an assessment of the association between measured variable in a dataset
- caution 1: two variables covary does not necessarily follow that the changes in the two variables are causally connected
- *e.g.* significant correlation between nations chocolate consumption and likelihood of producing Nobel laureates (NEJM 2012; 367:1562-4)
- caution 2: wrong to assume that a lack of correlation demonstrates a lack of association
- *e.g*. bacteria growth over time does not exhibit linear correlation but exponential or log phase growth

Decision errors in hypothesis testing

- Hypothesis tests are not flawless.
- In the court system, innocent people are sometimes wrongly convicted, and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests as well.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics.

Decision errors in hypothesis testing

- There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		
	H_A true		

Decision errors in hypothesis testing

- There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		✓

Decision errors in hypothesis testing

- There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true		✓

- A **Type 1 Error** (α) is rejecting the null hypothesis when H_0 is true.

Decision errors in hypothesis testing

- There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- A **Type 2 Error (β)** is failing to reject the null hypothesis when H_A is true.

How to minimize Type II errors (false negative)?

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- Increase sample size
  - A larger sample size increases the chances to capture the differences in the statistical tests, as well as increasing the power of a test.

How to minimize Type I errors (false positive)?

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	Type 1 Error
	H_A true	Type 2 Error	✓

- Choose a lower value for significance level
  - For example, the significance level can be minimized to 1% (0.01). This indicates that there is a 1% probability of incorrectly rejecting the null hypothesis.

Power and sample size calculation

- critical for every study design and ethical considerations
- if sample size is too small, the study will risk of failing to detect an effect
- if sample size is too big, it is both economically and ethically unjustifiable
- to determine a number that will detect an effect as statistically significant
- 4 variables:
  - sample size, n
  - effect size, d (size difference among groups)
  - significance level = P(Type I error), probability of finding an effect that is not there
  - power = 1 – P(Type II error), probability of finding an effect that is there

New drug for hypertension

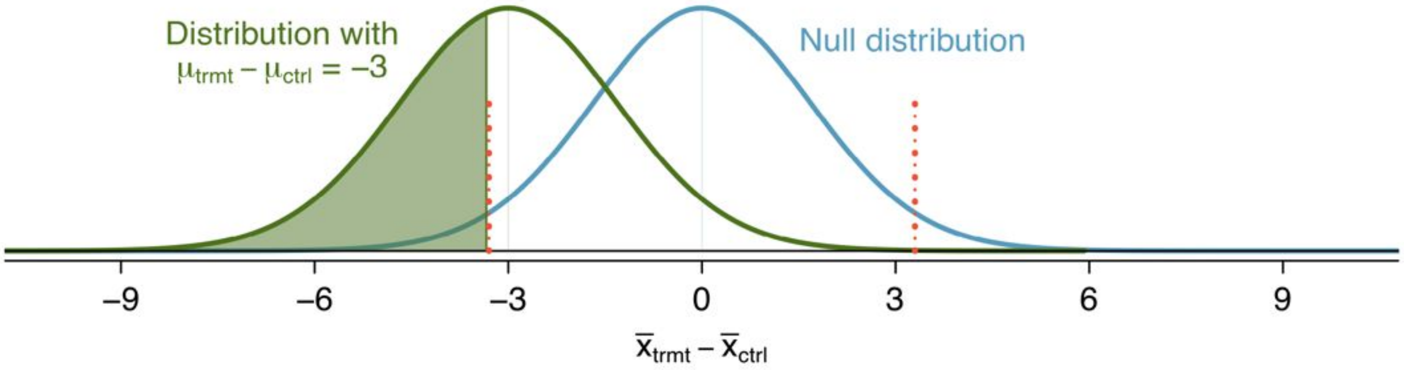
- Suppose a pharmaceutical company has developed a new drug for lowering blood pressure by 3 mmHg relative to standard medication. In preparation for a clinical trial to test the drug’s effectiveness, the company will try to recruit people who are taking a particular standard blood pressure medication. People in the control group will continue to take their current medication through generic-looking pills to ensure blinding.
- Previously published studies had shown that the standard deviation of patients’ blood pressures were about 12 mmHg and the distribution of patient blood pressures were approximately symmetric. If 100 patients will be recruitment per group, what is the probability that you can detect the decrease blood pressure by 3 mmHg?

Power calculation

- use “pwr” package in R
- n=100, α= 0.05
- effect size: d = $|\mu_1-\mu_2|$/sd = 3/12 = 0.25
- output: power = 42%

```
library(pwr)

pwr.t.test(n=100, d=0.25, sig.level = 0.05, type =
"two.sample")
```



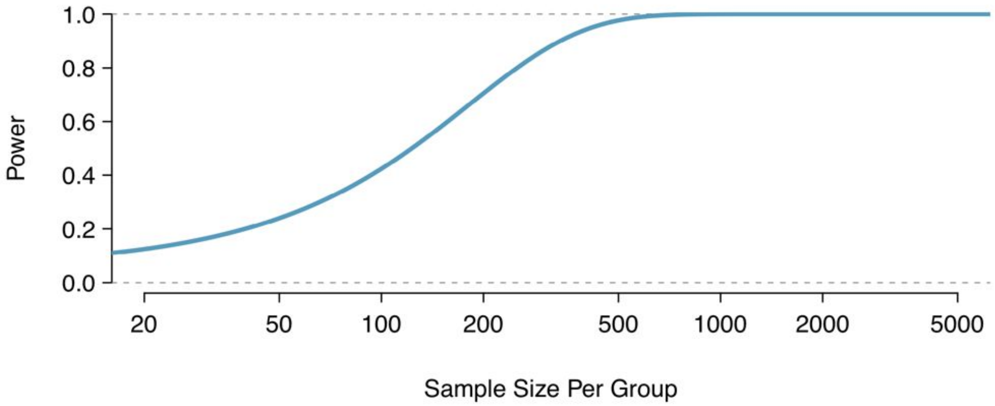
If we proceed with clinical trial with sample size of 100 in each group, we can only detect the drop of 3 mmHg with a probability of 0.42. Then, the data obtain may not support the hypothesis and fail to reject the null hypothesis. Then, waste millions of dollars.

Sample size calculation

- What sample size will lead to a power of 80%?
- power = 0.8; α= 0.05
- effect size: $d = |\mu_1-\mu_2|/sd = 0.25$
- output: n = 252 in each group

```
library(pwr)

pwr.t.test(power=0.8, d=0.25, sig.level = 0.05, type = "two.sample")
```



note that recruiting >300 subjects does not provide additional value in detecting an effect when α= 0.05

Take-away message

- Use linear regression to determine the relationship between two numerical variables.
- Power analysis is important in determining the minimum sample size that is suitable to detect the effect of a given test at the desired level of significance.