Ch. 10: Two Quantitative Variables

Navigation

By Section

- 10.1: start - end
- 10.2: start - end
- 10.3: start - end
- 10.4: start - end
- 10.5: start - end

10.1: Two Quantitative Variables

Scatterplots and Correlation

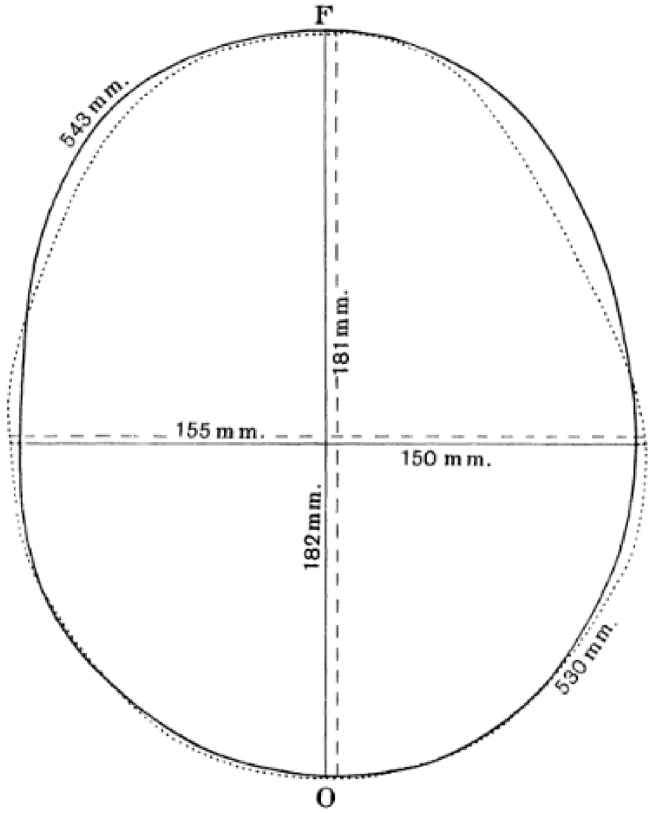
Graphical Summaries of Quantitative Variables

R.J. Gladstone (1905). "A Study of the Relations of the Brain to the Size of the Head", Biometrika, Vol. 4, p 105-123.

Data collected during 237 autopsies at Middlesex Hospital in London, excluding cases "in which the brain showed a distinctly pathological condition which would have obviously affected its weight"

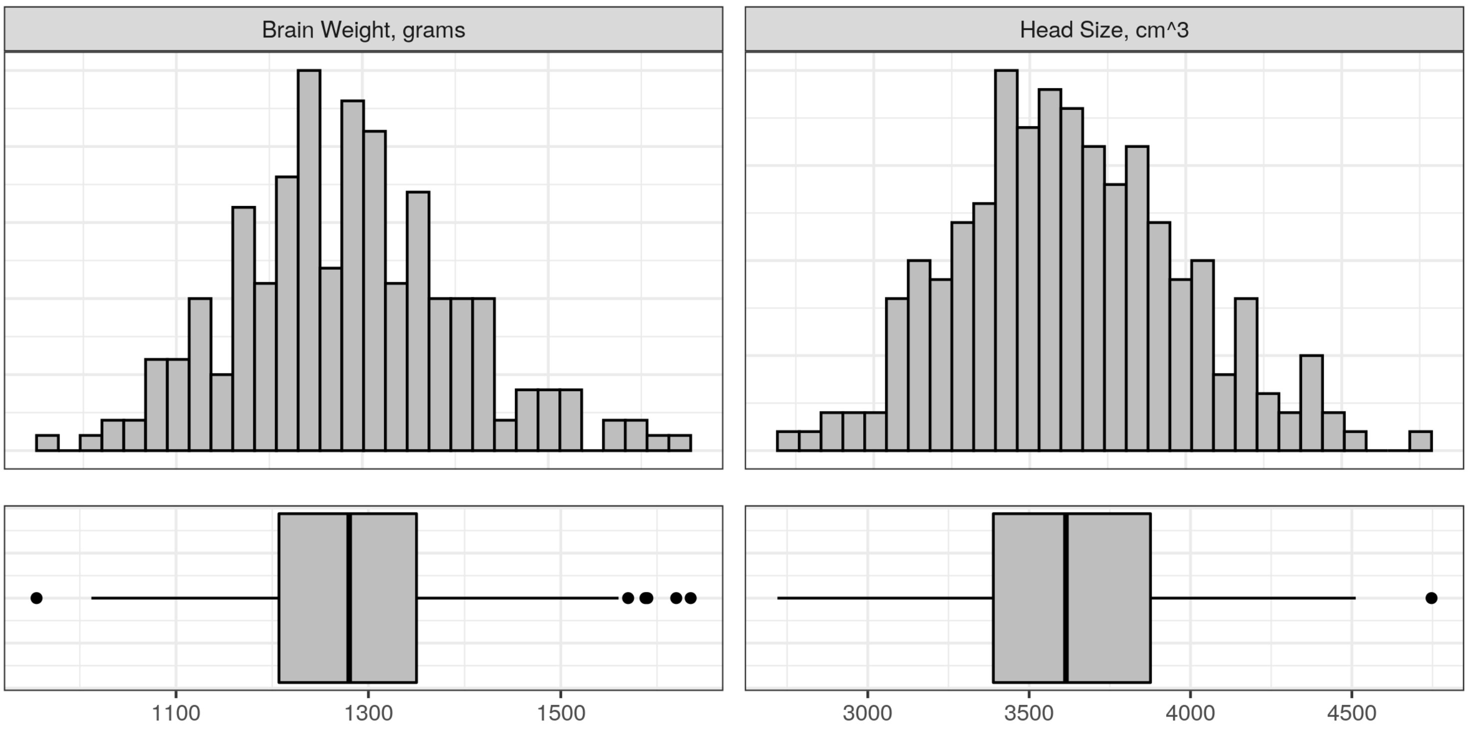
Variables:

- Gender
- Age (20 - 45 or 46+)
- Brain Weight (g)
- Head Size (cubic cm) the smallest rectangular block which could contain the head



Graphical Summaries of Quantitative Variables

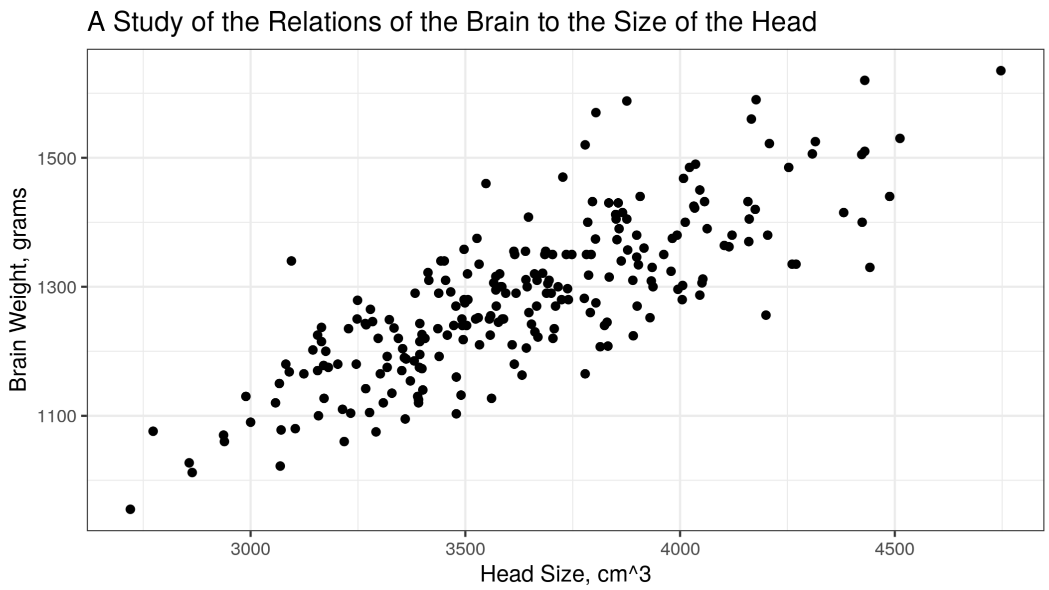
A single quantitative variable can be summarized visually using a histogram or a bar chart:



Graphical Summaries of Quantitative Variables

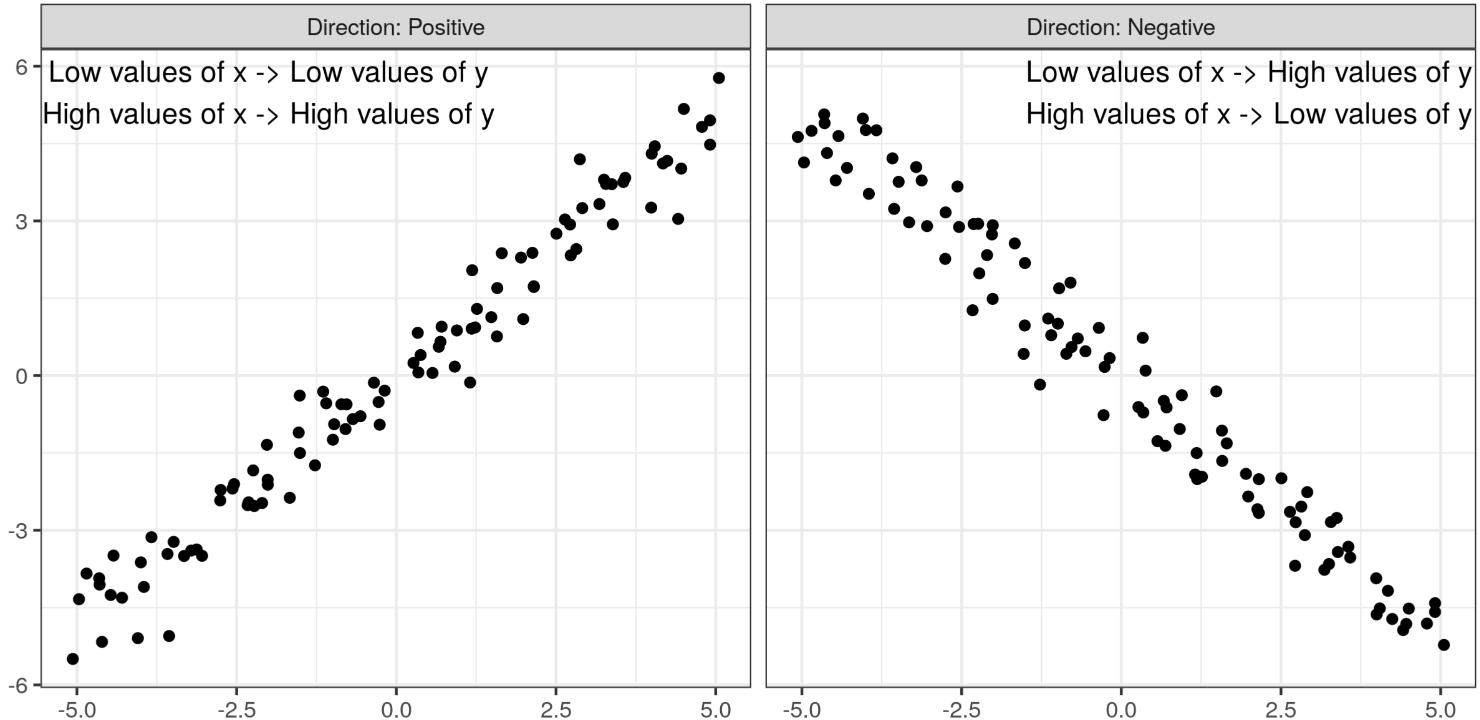
But, summarizing each variable separately doesn't tell us how the two variables might be related.

Is there a relationship between brain weight and head size? How do you know?



A **scatterplot** is a plot with the explanatory variable on the x-axis, and the response variable on the y-axis. Observations are shown as points corresponding to a set of quantitative measurements.

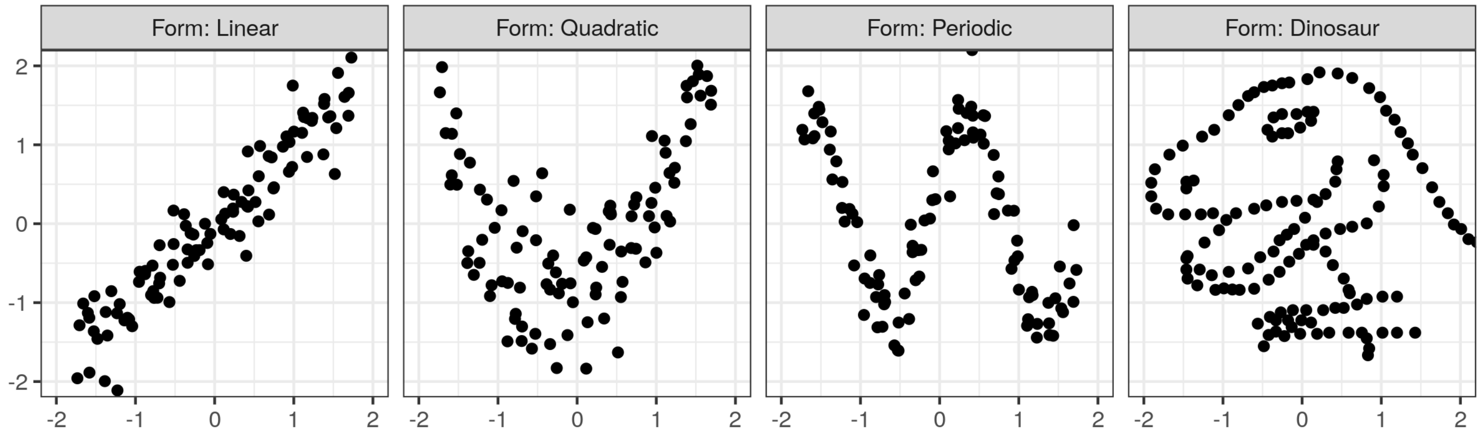
Describing Variable Relationships: Direction



Positive slope: as x increases, y increases too.

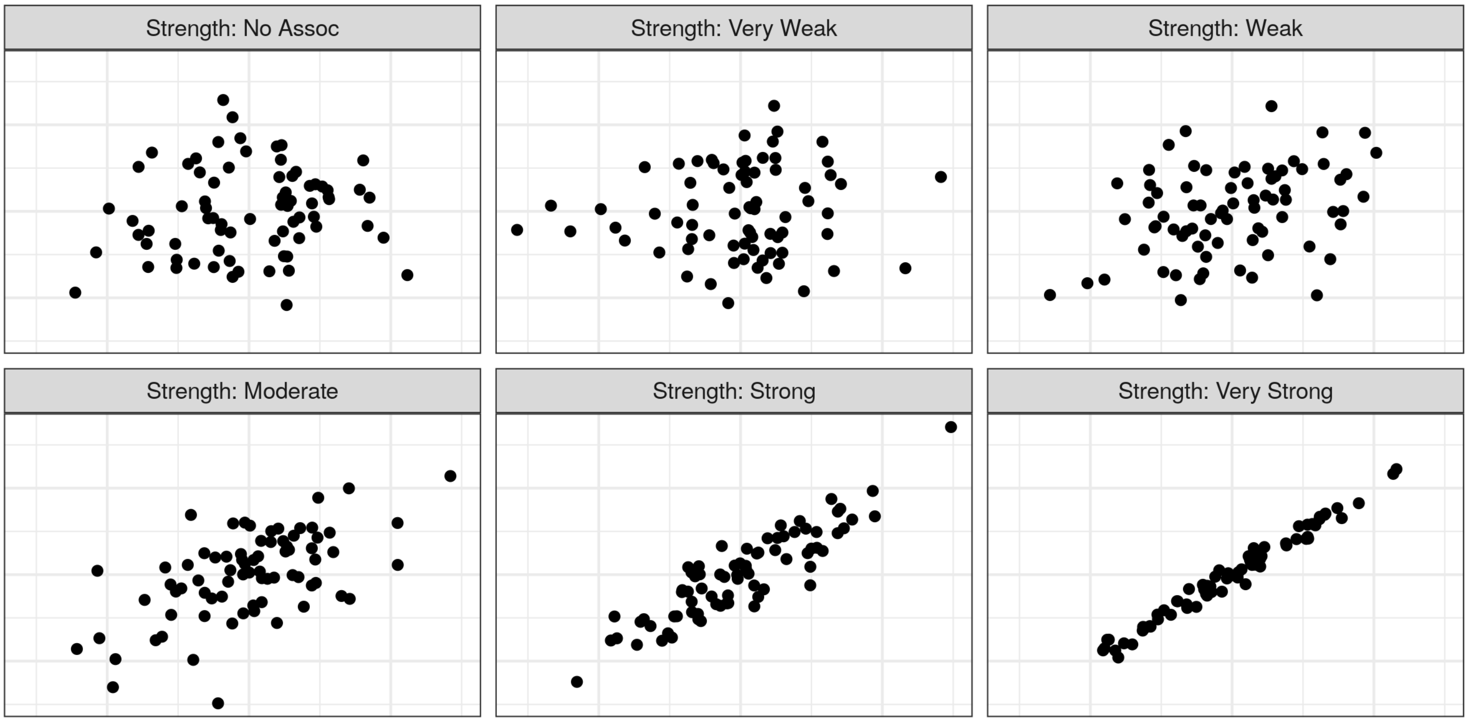
Describing Variable Relationships: Form

The **form** of an association is whether it follows a linear pattern, or some sort of more complicated pattern - periodic, polynomial (quadratic, cubic, etc.)



Describing Variable Relationships: Strength

The **strength** of an association indicates how well the value of one variable can be predicted if you know the value of the other variable.

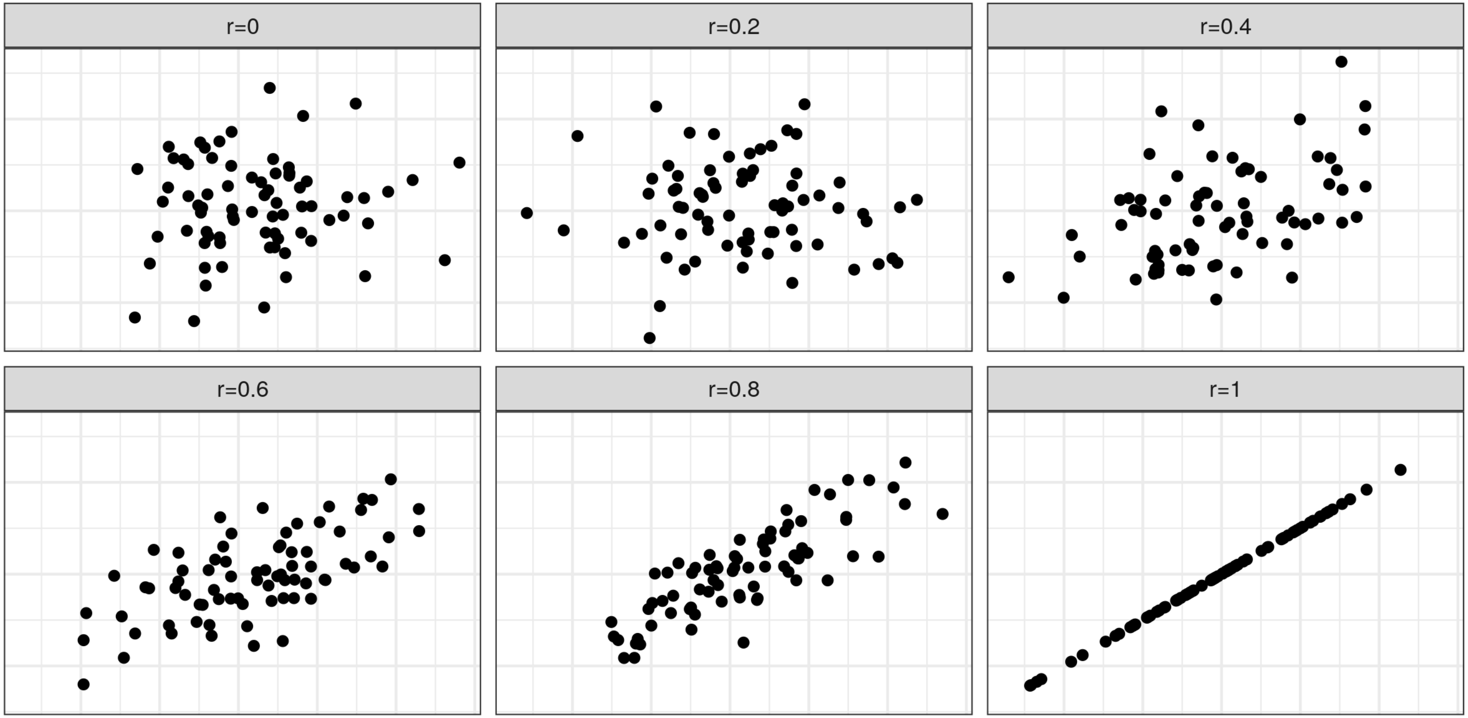


We can describe the strength and direction of a *linear* relationship using the **correlation coefficient**

Correlation Coefficient

The **correlation coefficient**, r, is always between -1 and 1.

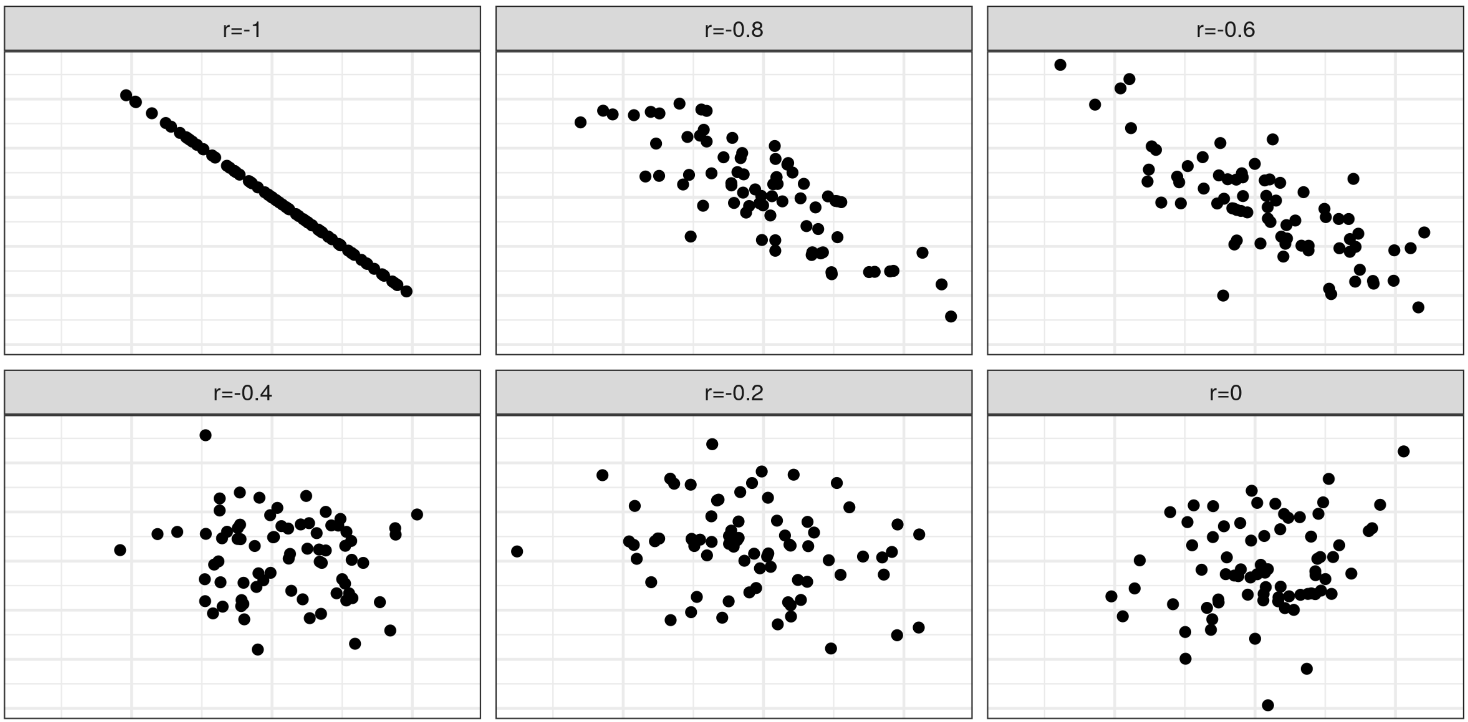
-1	-0.5	0	0.5	1
perfect negative correlation	moderate negative correlation	no correlation	moderate positive correlation	perfect positive correlation



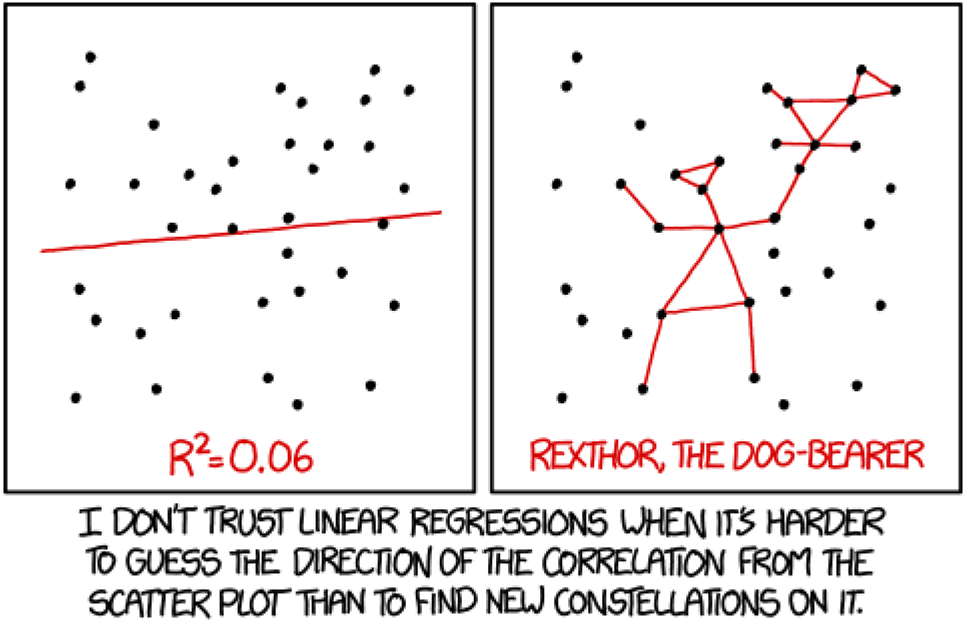
Correlation Coefficient

The **correlation coefficient**, r, is always between -1 and 1.

-1	-0.5	0	0.5	1
perfect negative correlation	moderate negative correlation	no correlation	moderate positive correlation	perfect positive correlation

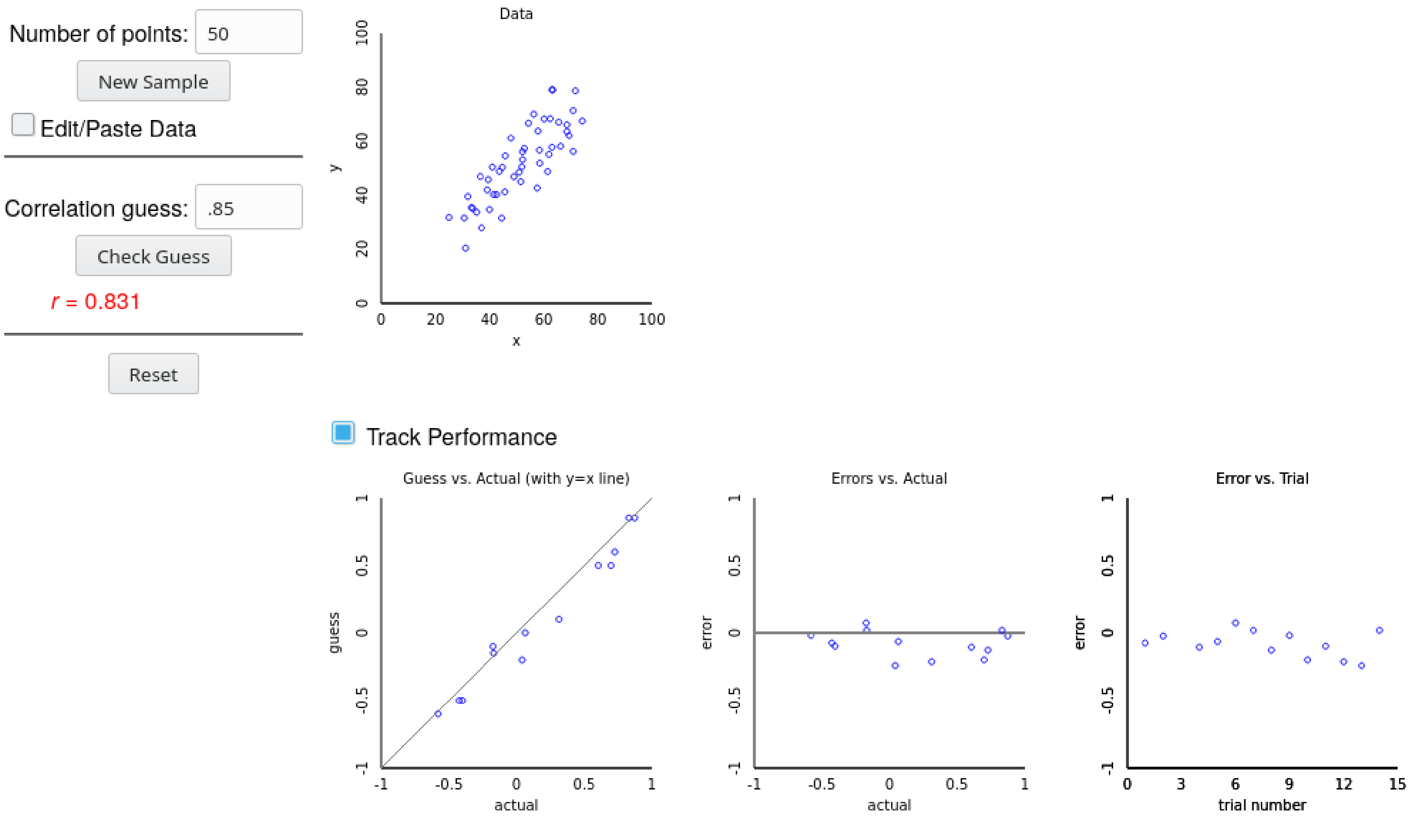


Correlation Coefficient



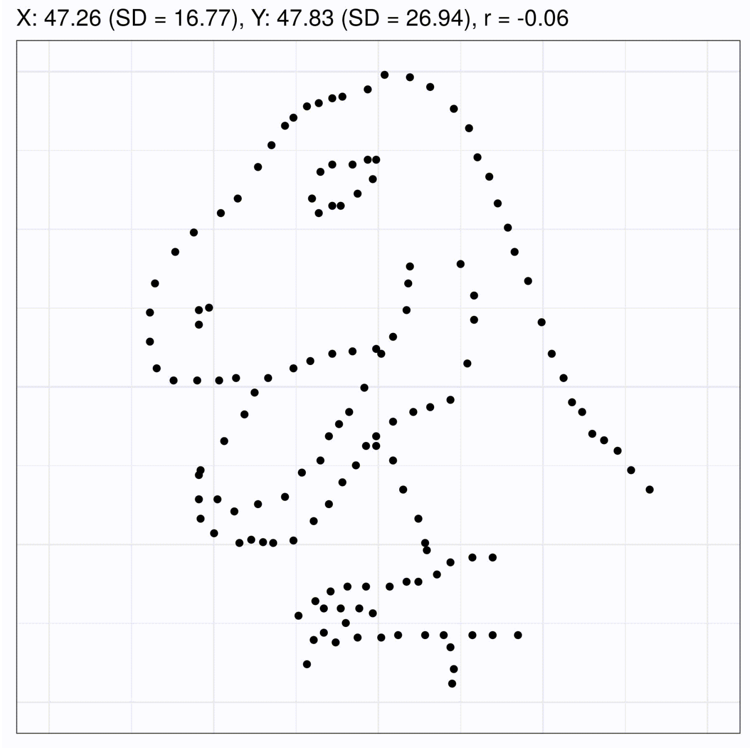
Correlation Coefficient

Get a feel for it by playing the correlation guessing game!



Correlation Coefficient

The correlation coefficient is only useful for showing the strength of linear relationships.



All of these plots have essentially the same correlation coefficient, but in some cases there are very clear associations between x and y

Correlation Coefficient

(From Calculation Details in the Appendix)

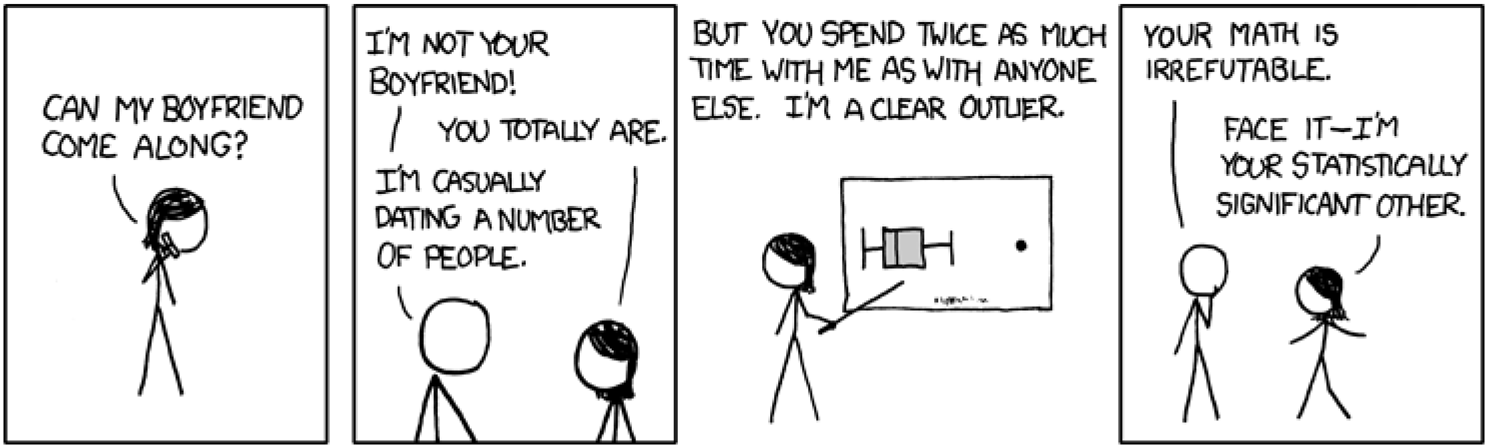
$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

Correlation is not Causation

Just in case you haven't heard this chant yet, "correlation is not causation". Say it a few times.

It's important to remember that correlation is a measure of association, but that doesn't mean there's any causal factors involved. In some cases, the choice of explanatory and response variables are arbitrary.

Outliers and Influential Observations

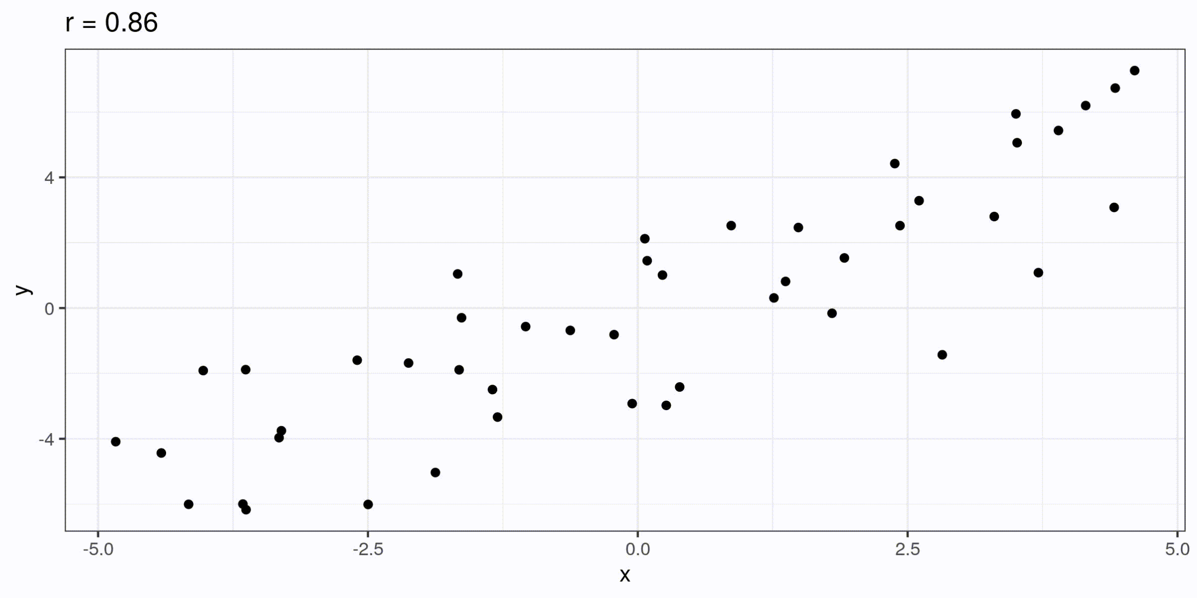


With one variable, outliers are fairly easy to spot

When there are two variables, we don't just have to worry about outliers in one dimension; we also have to worry about **influential observations**

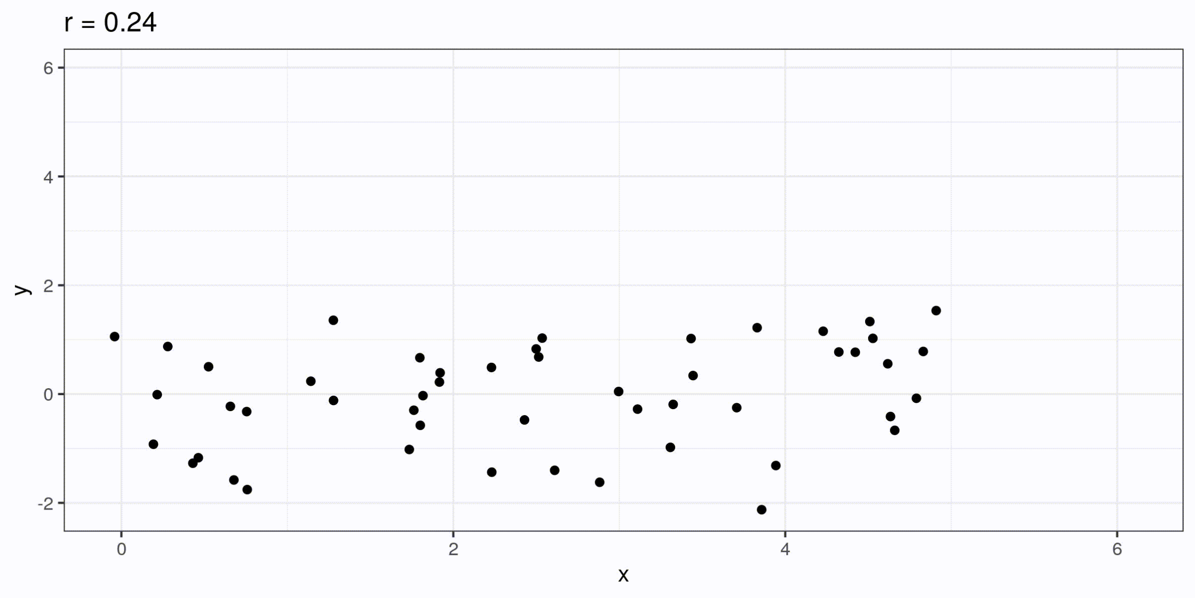
Outliers and Influential Observations

Influential observations are observations which, if included, change our understanding of the relationship between two variables.



Outliers and Influential Observations

Influential observations are observations which, if included, change our understanding of the relationship between two variables.



Exploration 10.1

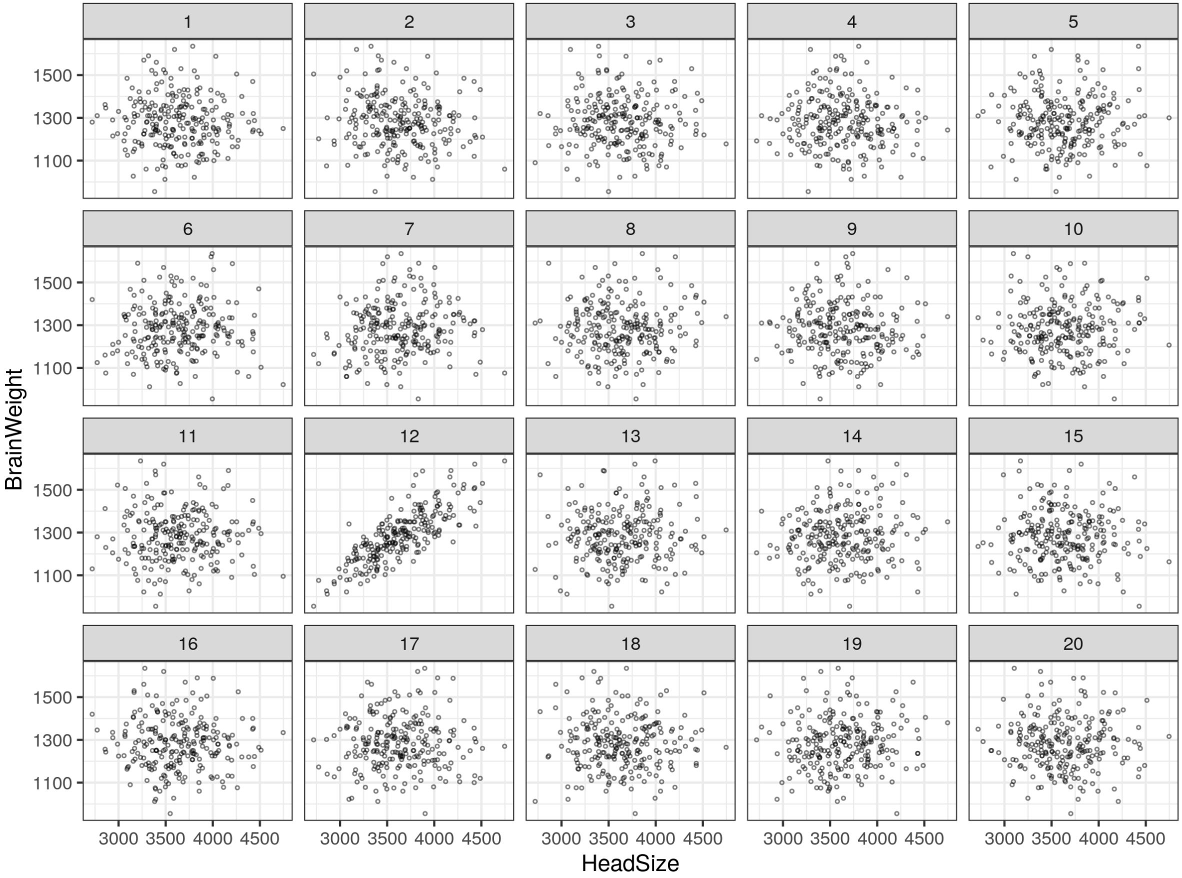
Work through Exploration 10.1 to get a chance to put the material in this section into practice. You can turn it in for 10 points of extra credit in the "Assignment" category.

10.2: Inference for the Correlation Coefficient

Simulation Based Approach

One of these things is not like the others

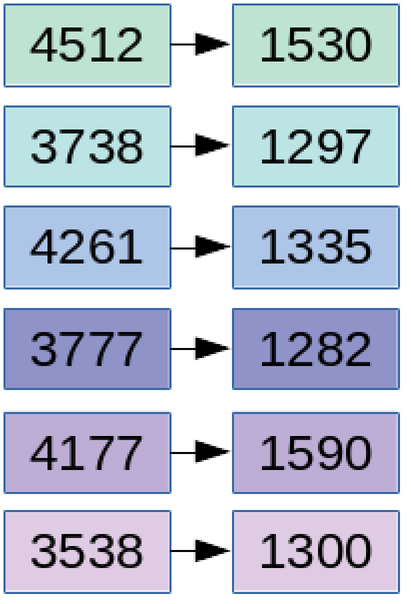
Which one of these things is not like the others?



Simulation-based Inference for Correlation Coefficient

Our null hypothesis is H_0 : No relationship between x and y

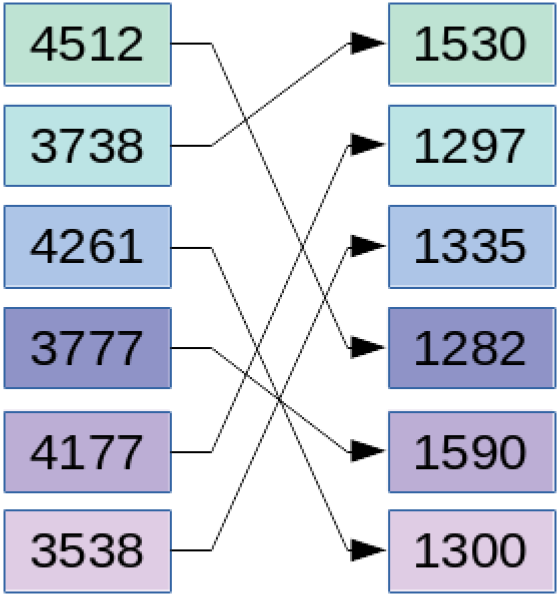
How can we simulate this?



Simulation-based Inference for Correlation Coefficient

Our null hypothesis is H_0 : No relationship between x and y

How can we simulate this?

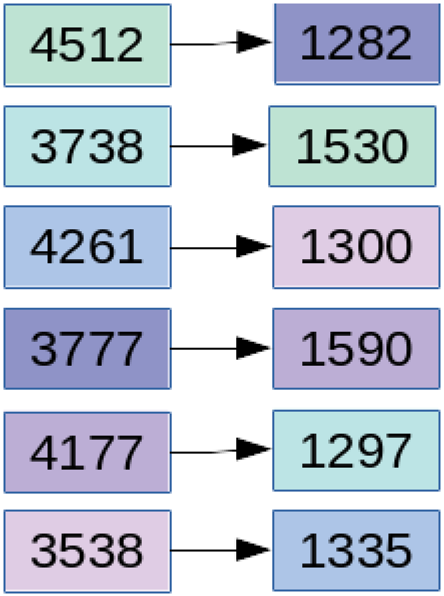


If there's no relationship between x and y, then it doesn't really matter what x value is paired with a given y value… so we can just change which values are paired together.

Simulation-based Inference for Correlation Coefficient

Our null hypothesis is H_0 : No relationship between x and y

How can we simulate this?

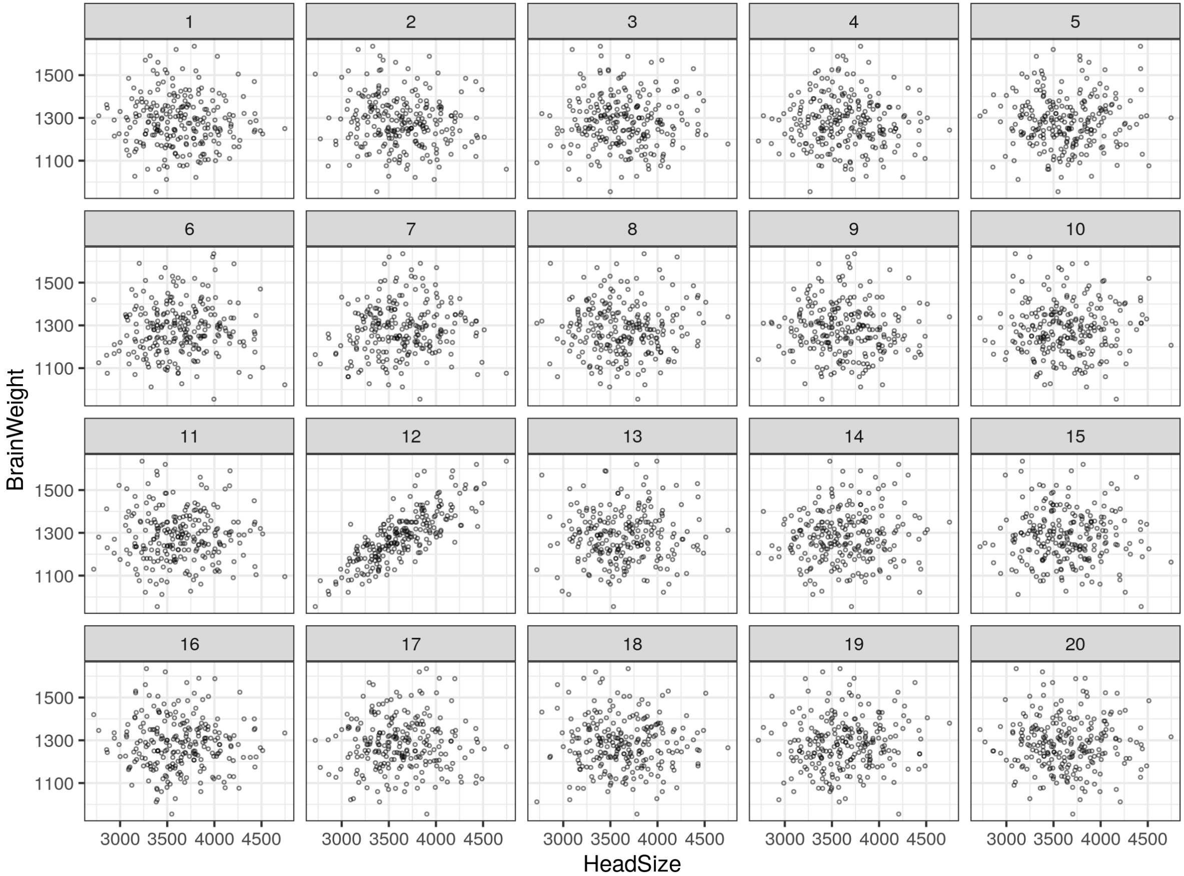


This is equivalent to shuffling the order of y and creating a new regression

Summary of Simulation Model

- Null hypothesis: No association between x and y variables
  The population symbol for r is ρ, so $H_0 : \rho = 0$
- One repetition: Re randomizing the response outcomes to the explanatory variable values (randomize y values)
- Statistic: Correlation coefficient, r

Which one of these things is not like the others?



Here, our y values have been shuffled for each sub-plot that isn't plot 12. Plot 12 contains the original data.

Hypotheses for Correlation Coefficient

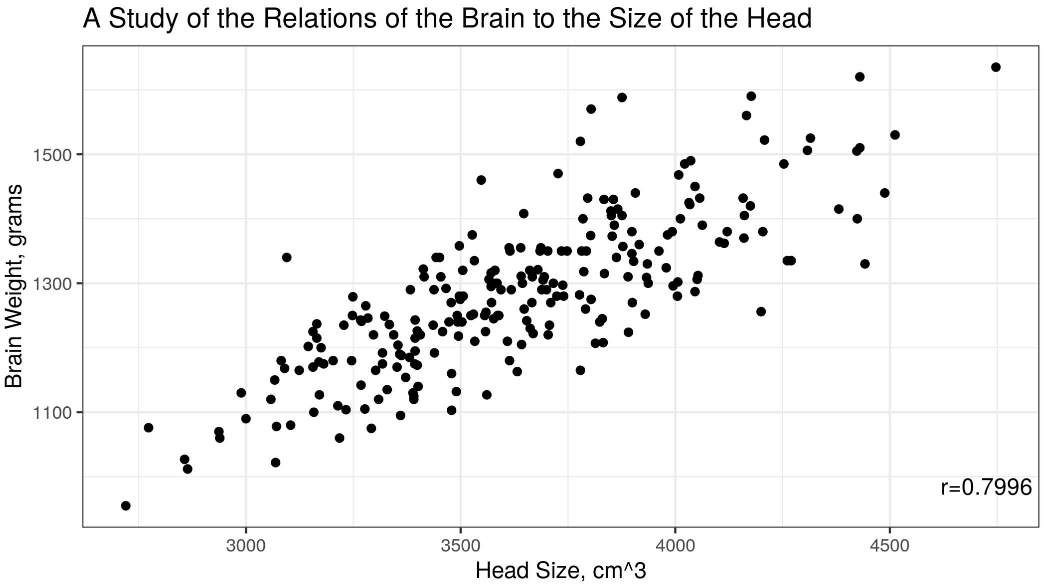
H_0 : There is no relationship between x and y $(\rho = 0)$

H_A : There is a relationship between x and y $(\rho \neq 0)$

ρ (pronounced "row", spelled "rho") is the population version of r

Caution ρ is not p and is NOT a p-value.

Example: Head Size and Brain Weight



H_0 : No linear relationship between head size and brain weight $(\rho = 0)$

H_A : There is a linear relationship between head size and brain weight $(\rho \neq 0)$

Example: Head Size and Brain Weight

1. **Statistic** - correlation coefficient from the sample
2. **Simulate**
   - Assume there is no relationship between head size (x) and brain weight (y)
   - Shuffle the values of y
   - Calculate the correlation coefficient r^* from the simulated data
3. **Strength of evidence** - in how many simulated samples did we get a correlation coefficient r^* with magnitude greater than our sample correlation coefficient r?

Example: Head Size and Brain Weight

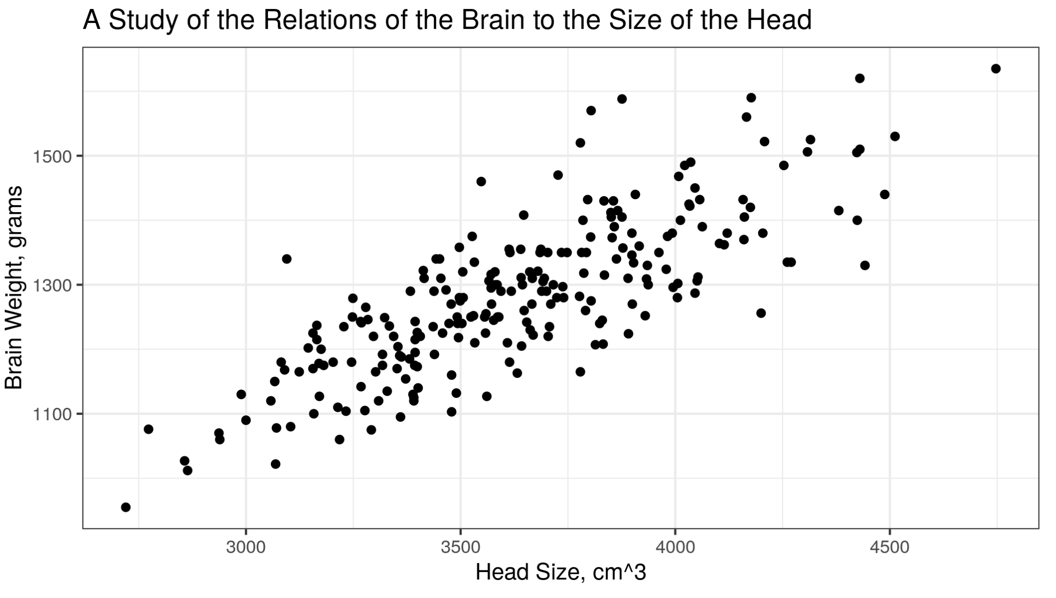
Conclusion:

I reject H_0 that there is no linear relationship between head size and brain weight with $p < 0.001$. There is very strong evidence that the linear relationship observed between the variables did not occur due to random chance, thus, we must conclude that there is a linear relationship between head size and brain weight, that is, that $\rho \neq 0$.

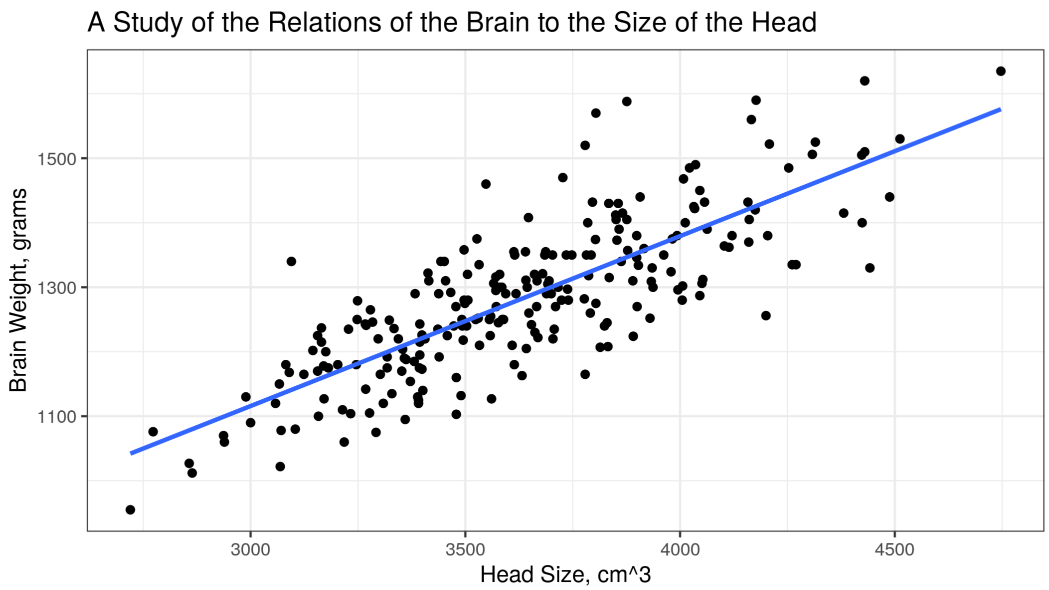
10.3: Least Squares Regression

Motivation

In the last two sections, we've talked about the Brain weight vs. Head size data:



Regression Line



General equation: $\hat{y} = a + bx$ where

- x is the explanatory variable
- $\hat{y}$ is the response variable
- a is the **y-intercept** (the predicted value when $x = 0$)
- b is the **slope**

Regression Line

The equation of the regression line in the picture on the last slide is:

$$\hat{y} = 325.573 + 0.263x$$

What does that mean?

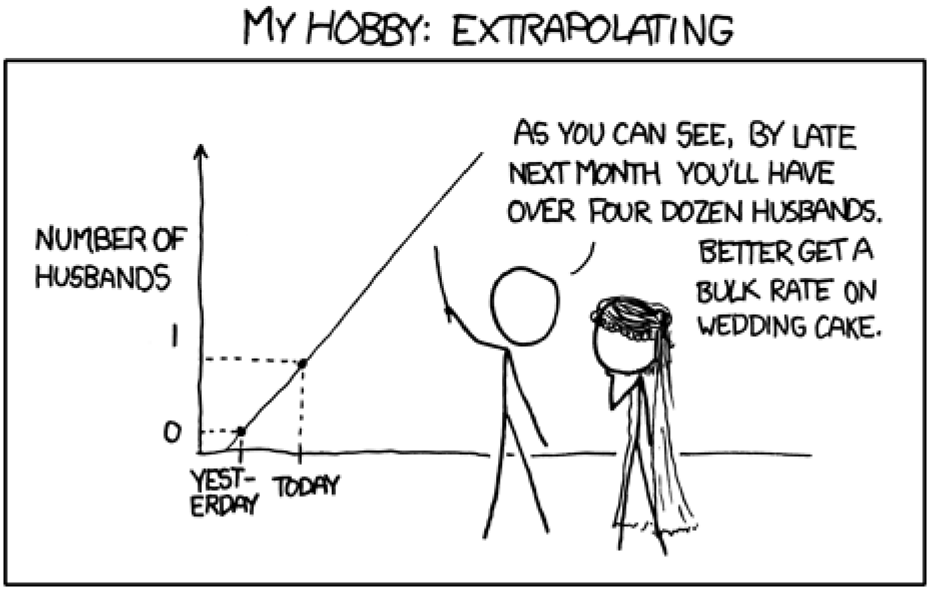
- With a head size of 0 cm^3, we would expect a brain weight of …
  325.57 g (on average)

  Sometimes, the interpretation of the intercept doesn't make a lot of sense in context.

- An increase in head size of 1 cm^3 leads to a predicted change in average brain weight of 0.2634 grams

Note that the sign of the slope is the same as the sign of the correlation coefficient, because both indicate the direction of the association.

Extrapolation



Predicting values for the response variable for explanatory variable values outside the range of the original data is known as **extrapolation** and can lead to very misleading predictions.

Extrapolation

Predicting values for the response variable for explanatory variable values outside the range of the original data is known as **extrapolation** and can lead to very misleading predictions.

For instance, if we try to predict the brain weight of a 2.5 month old child included in the original brain weight publication (but not in our dataset)...

$x = 1212$, so $\hat{y} = ?$

$325.573 + 0.263(1212) = 644.850$

Our model would predict that the child's brain would weigh about 645 grams, on average. In fact, it weighted 490 grams.

Our prediction had an error of $490 - 644.850 = -154.850$

This error is called a **residual**

Extrapolation

Extrapolation is sometimes reasonable, but often is ill-advised.

- I have data on the foot length and height of 58 people, ranging from 4'10 to 6'1. I want to predict the foot length of someone who is
  - 4'2 :
  - 6'3 :
  - 7':
- I have data on the number of ice cream cones sold at a shop and the temperature for January - April. I want to predict how many cones will be sold at a shop per day during
  - July :
  - early May :
  - April 30, in a similarly sized midwestern town :
  - in Saudi Arabia during May :

Extrapolation

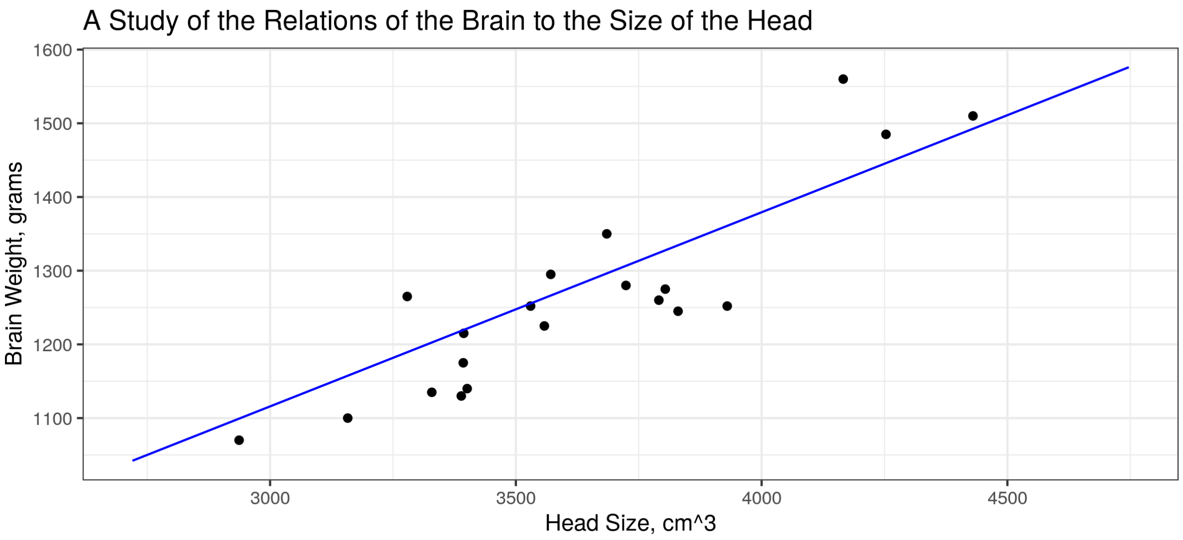
Extrapolation is sometimes reasonable, but often is ill-advised.

- I have data on the foot length and height of 58 people, ranging from 4'10 to 6'1. I want to predict the foot length of someone who is
  - 4'2 : not a good idea - most adults are taller, the relationship may not hold for children
  - 6'3 : probably fine - we're extrapolating by 2", and someone who is 6'1 is probably not so different from someone who is 6'3
  - 7': not a good idea, only a few people have gotten to be that tall, and the general population may not represent them well

- I have data on the number of ice cream cones sold at a shop and the temperature for January - April. I want to predict how many cones will be sold at a shop per day during
  - July : not a good idea (much higher avg. temp)
  - early May : probably fine (avg temp approximately same as late April)
  - April 30, in a similarly sized midwestern town : may be reasonable, but use caution
  - in Saudi Arabia during May : not a good idea (much, much higher avg. temp, different location that is not similar)

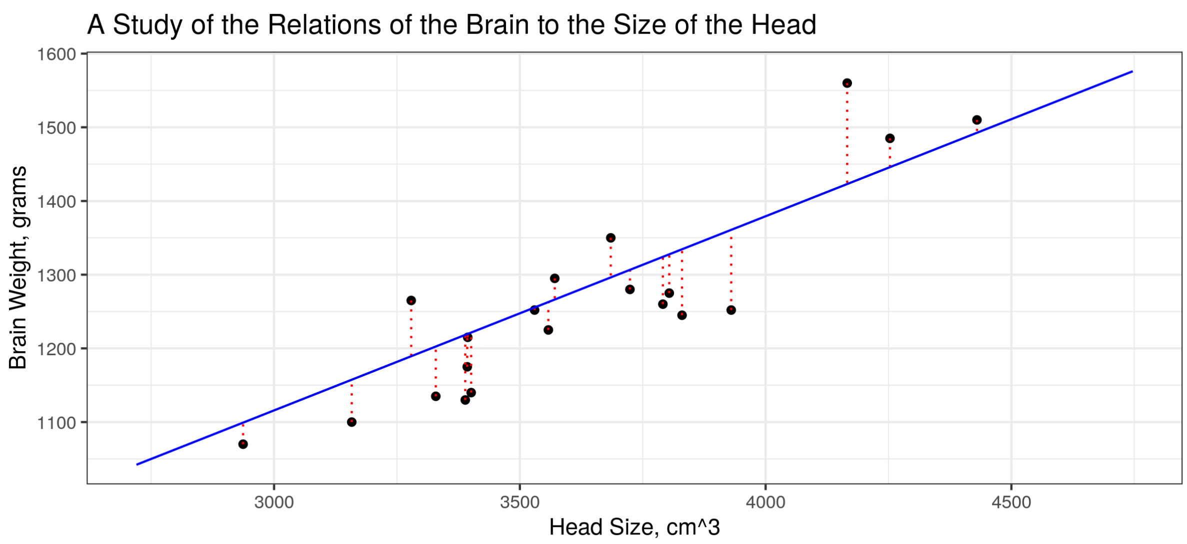
Least Squares Regression

How do we get our regression line?



Least Squares Regression

How do we get our regression line?



The least-squares regression line minimizes the sum of the squared residuals from the line - the sum of $(y - \hat{y})^2$

You can get a feel for what this means here - play with the first two charts.
(this page has regression with multiple variables as well, but ignore that part unless you're interested - this class only does single-variable regression.)

Least Squares Regression

How do we get our regression line?

From the computational details section of the appendix:

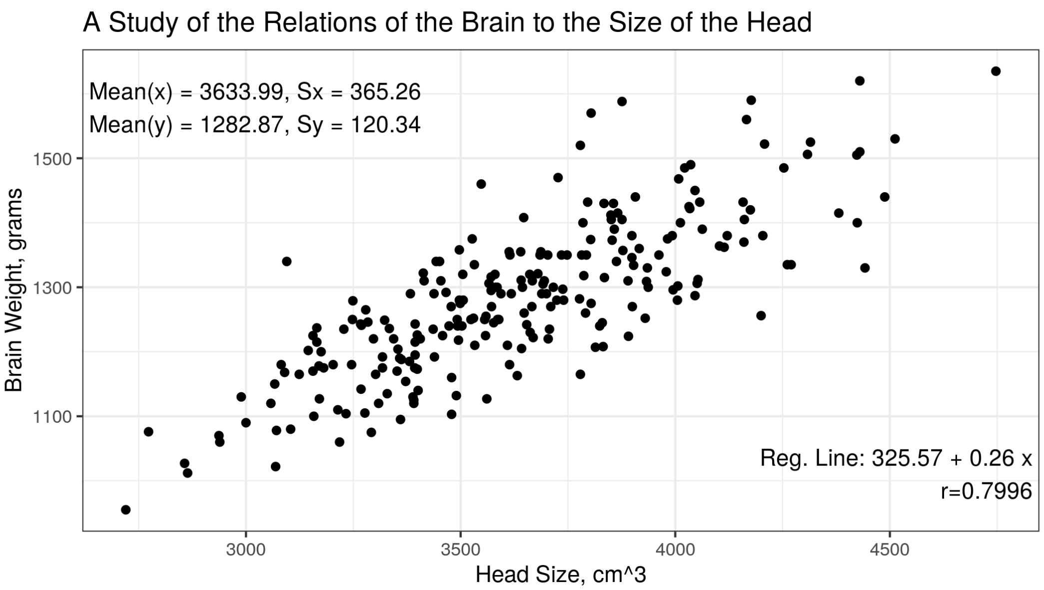
$$b = r\frac{s_y}{s_x}$$

where r is the correlation coefficient.

Then,

$$a = \overline{y} - b\overline{x}$$

Example: Head Size and Brain Weight



We can compute b and a from the other information given on the graph:

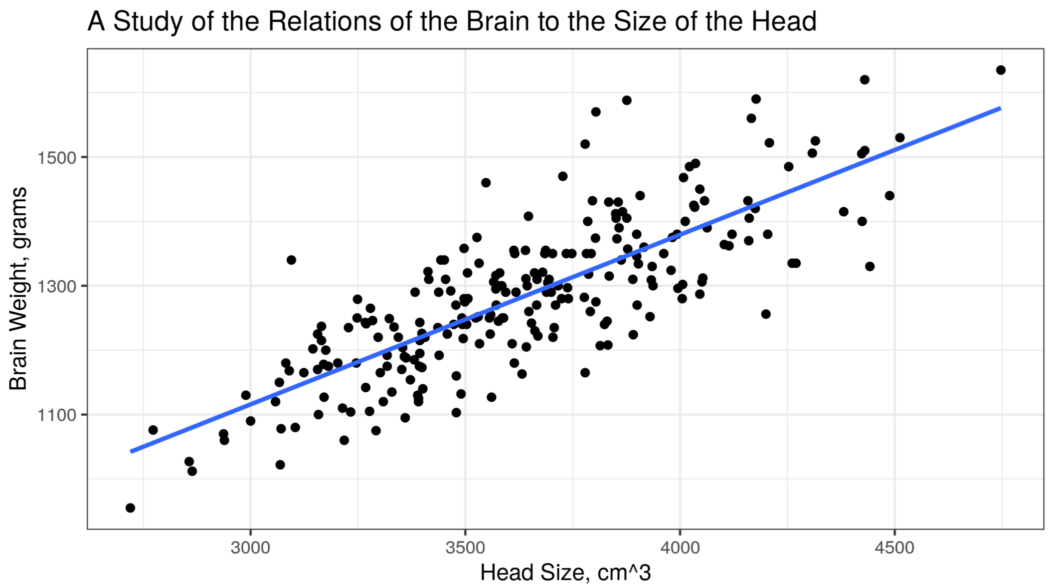
$$b = r\frac{s_x}{s_y} = 0.8 \times \frac{120.34}{365.261} = 0.263$$

$$a = \overline{y} - b\overline{x} = 1282.873 - 0.263(3633.992) = 325.573$$

Coefficient of Determination (R^2)

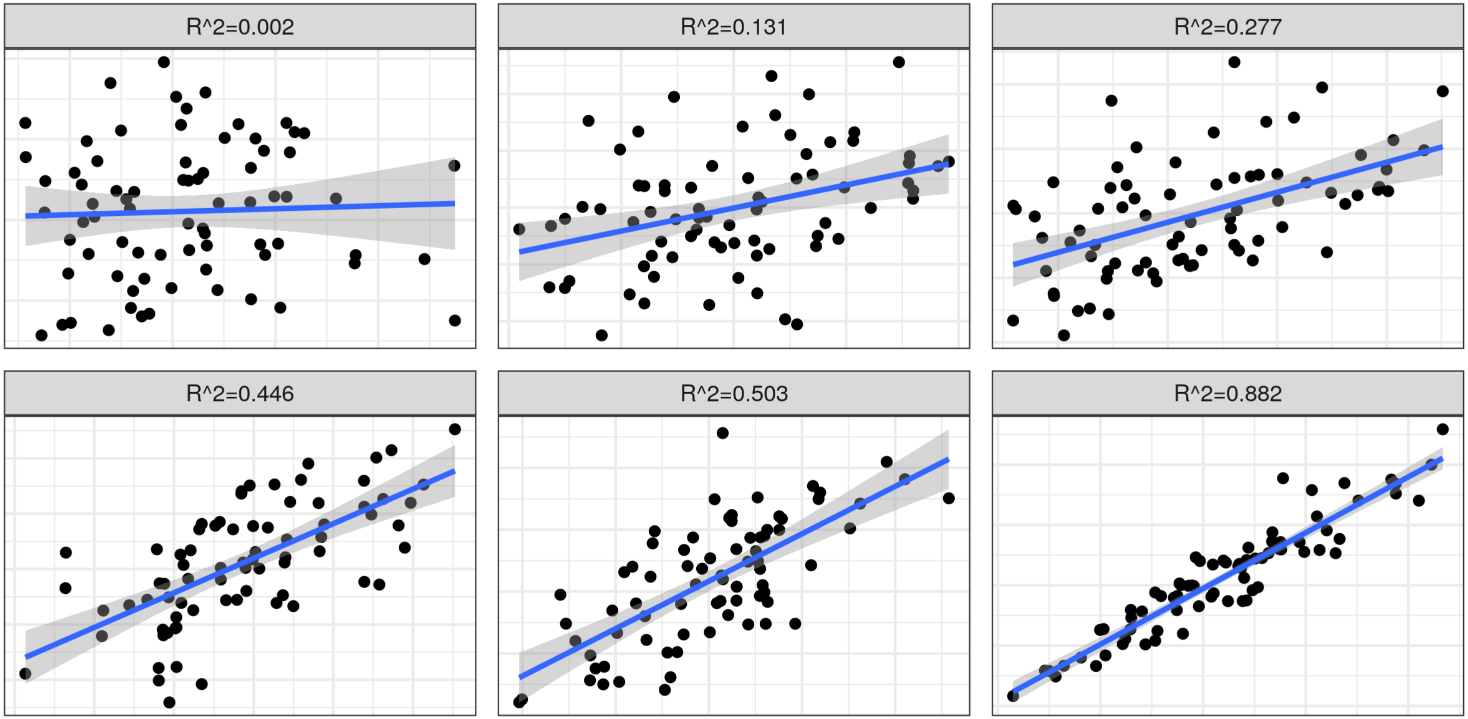
We can measure how well our data fit our regression line with R^2 (which is literally the correlation value, squared).

R^2 is between 0 and 1, where 1 = perfect predictions and 0 = no relationship at all.



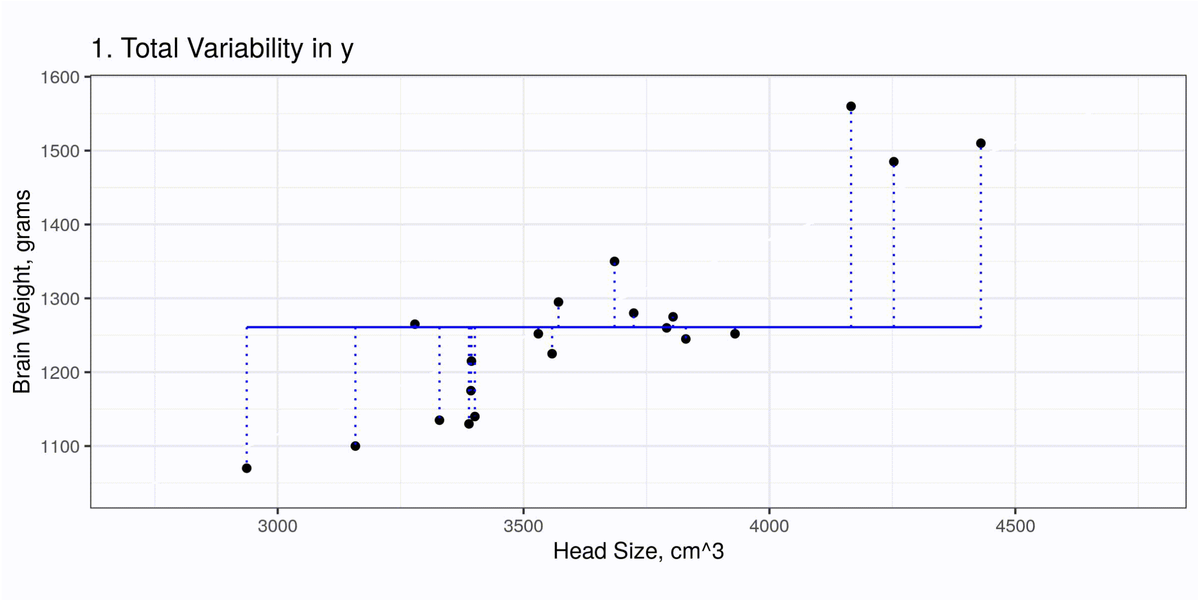
The R^2 for this plot is 0.639

Coefficient of Determination (R^2)



Coefficient of Determination (R^2)

R^2 can be interpreted as the percentage of variability in y which is explained by the regression line.



Coefficient of Determination (R^2)

In math, the last series of graphs can be shown like this:

$$\underbrace{SSE(y-\overline{y})}_{\text{total error}} = \underbrace{SSE(y-\hat{y})}_{\text{residual error}} + \underbrace{SSE(\hat{y}-\overline{y})}_{\text{explained by regression}}$$

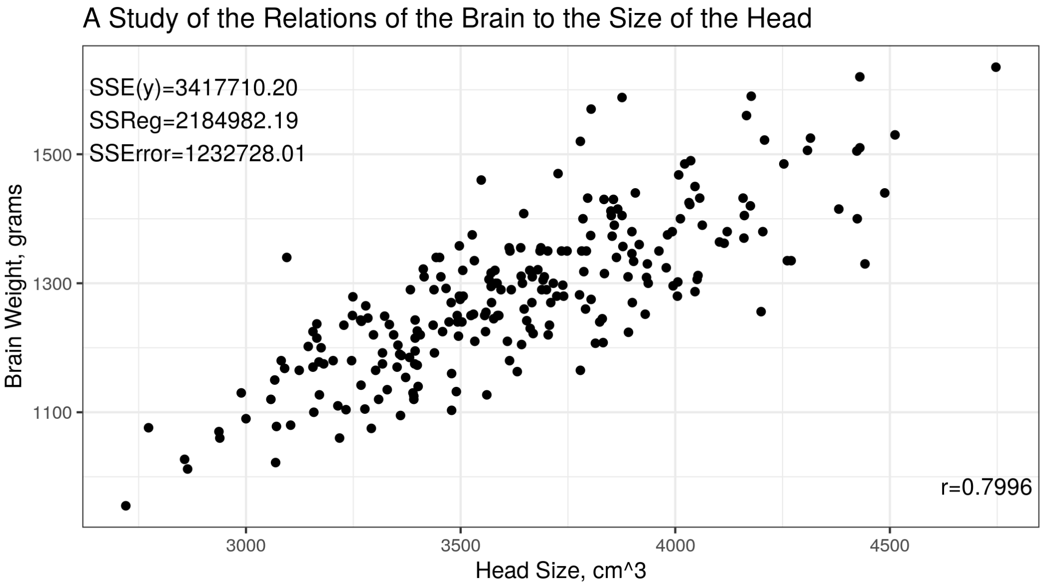
This is slightly different than the book's notation, where $SSE(y-\overline{y}) = SSE(\overline{y})$ and $SSE(\hat{y}-\overline{y})$ is written as SSE(regression line)

In math, then, we can write R^2 as

$$R^2 = 100 \times \left(1 - \frac{SSE(y-\hat{y})}{SSE(y-\overline{y})}\right) = 100 \times \left(1 - \frac{\text{residual std error}}{\text{total std error}}\right)$$

Notice that this equation works out to equal the proportion of the total error in y explained by the regression.

Example: Head Size and Brain Weight



Let's compute R^2 using the SSE formula and check that it's the same as the value of r, squared.

Example: Head Size and Brain Weight

$$R^2 = \frac{SS(\hat{y} - \overline{y})}{SS(y - \overline{y})} = \frac{\text{SSReg}}{SSE(y)} = \frac{2.185 \times 10^6}{3.418 \times 10^6} = 0.639$$

$$R^2 = r^2 = (.7996)^2 = 0.639$$

Exploration 10.3

Work through Exploration 10.3 (ignoring the parts which say compare with classmates) to review the concepts covered in this lecture.

Upload your answers to Canvas for extra credit.

10.4: Inference for the Regression Slope:

Simulation-Based Approach

Review

Regression equation: $\hat{y} = a + bx$

- $\hat{y}$ is the
- a is the
- b is the
- x is the

Review

Regression equation: $\hat{y} = a + bx$

- $\hat{y}$ is the response variable
- a is the intercept
- b is the slope
- $\hat{x}$ is the explanatory variable

Inference for Regression Slope

To test whether two variables are linearly associated, we usually want to know if $b = 0$.

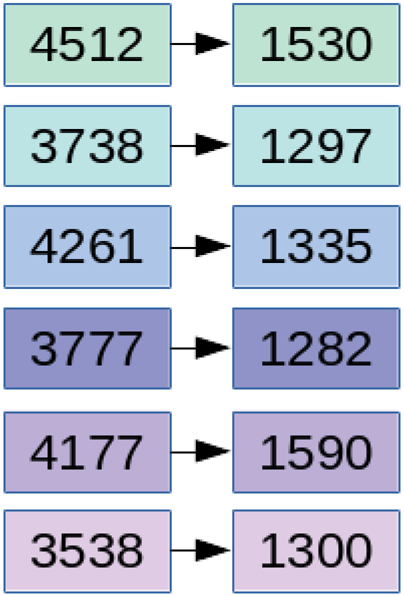
If $b = 0$ then $\hat{y} = a + 0x = a$... That is, x doesn't affect y at all.

What happens if we estimate a $b \neq 0$ by chance?

Simulation-based Inference for the Slope

Our null hypothesis is H_0 : No relationship between x and y, that is, $\beta = 0$
β is the greek letter corresponding to b - the "true/population slope"

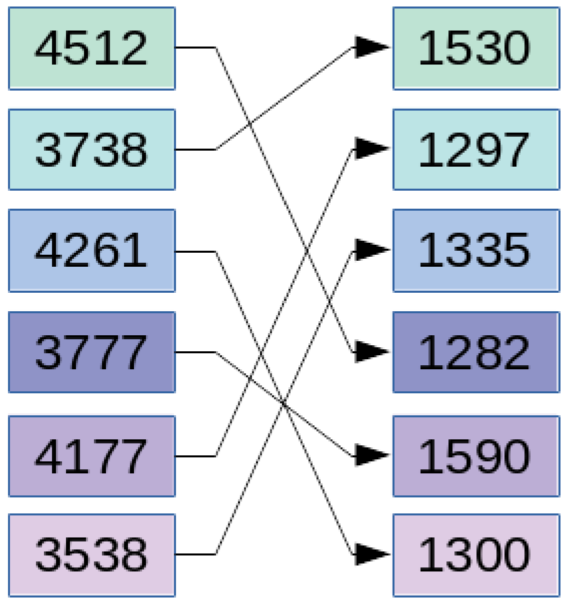
How can we simulate this?



Simulation-based Inference for the Slope

Our null hypothesis is H_0 : No relationship between x and y, that is, $\beta = 0$
β is the greek letter corresponding to b - the "true/population slope"

How can we simulate this?

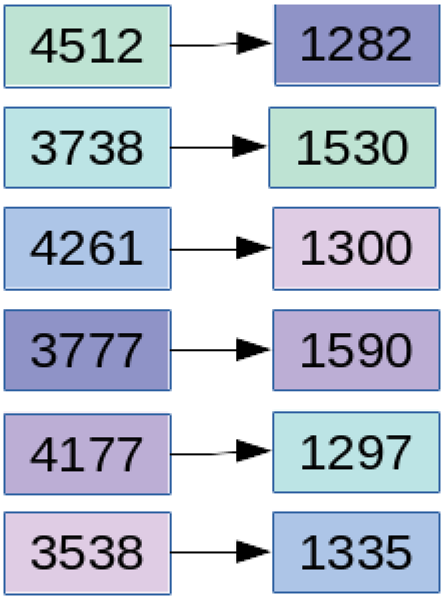


If there's no relationship between x and y, then it doesn't really matter what x value is paired with a given y value... so we can just change which values are paired together.

Simulation-based Inference for the Slope

Our null hypothesis is H_0 : No relationship between x and y, that is, $\beta = 0$
β is the greek letter corresponding to b - the "true/population slope"

How can we simulate this?

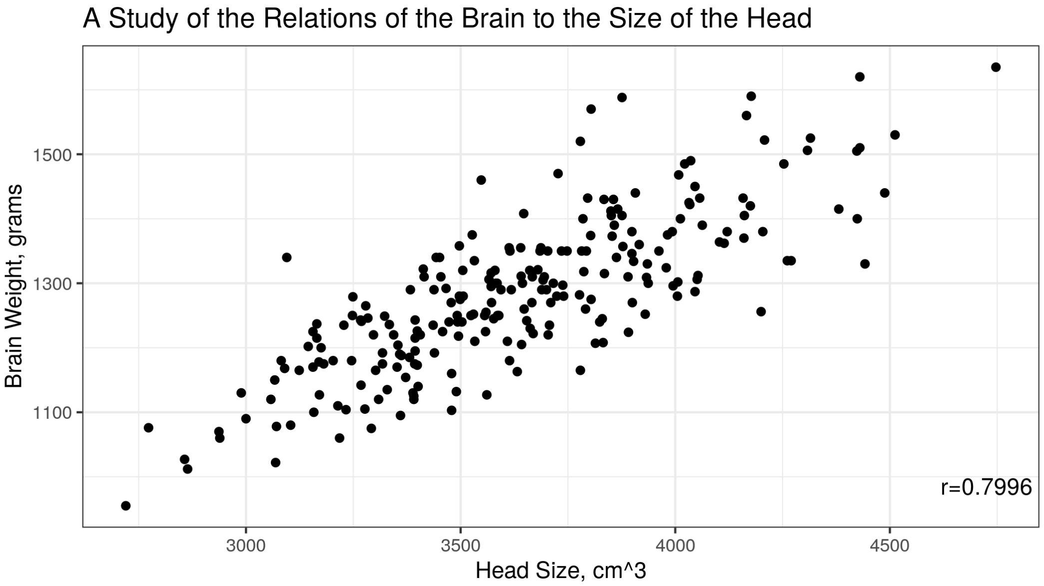


This is equivalent to shuffling the order of y and creating a new regression

Summary of Simulation Model

- Null hypothesis: No association between x and y variables ($\beta = 0$)
- One repetition: Re randomizing the response outcomes to the explanatory variable values (randomize y values)
- Statistic: Slope, b

Example: Head Size and Brain Weight

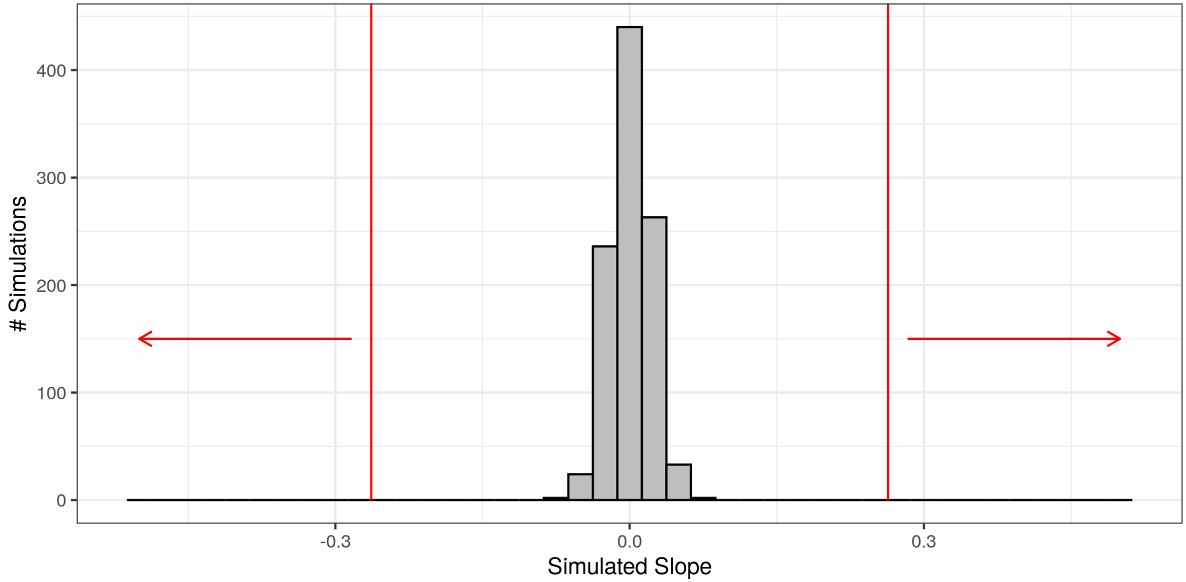


H_0 : No linear relationship between head size and brain weight $(\beta = 0)$

H_A : There is a linear relationship between head size and brain weight $(\beta \neq 0)$

Example: Head Size and Brain Weight

1. **Statistic** - slope computed from the sample
2. **Simulate**
3. **Strength of evidence** - in how many simulated samples did we get a slope b^* with magnitude greater than our sample slope b?



In 0 samples out of 1000, $|b^*| > 0.263$, so $p < 0.001$

Example: Head Size and Brain Weight

Conclusion:

I reject H_0 that there is no linear relationship between head size and brain weight with $p < 0.001$. There is very strong evidence that the linear relationship observed between the variables did not occur due to random chance, thus, we must conclude that there is a linear relationship between head size and brain weight, that is, that $\beta \neq 0$.

Note that this is the same interpretation (word for word) as we used in Section 10.2 (with the parameter changed)

10.5: Inference for the Regression Slope:

Theory-Based Approach

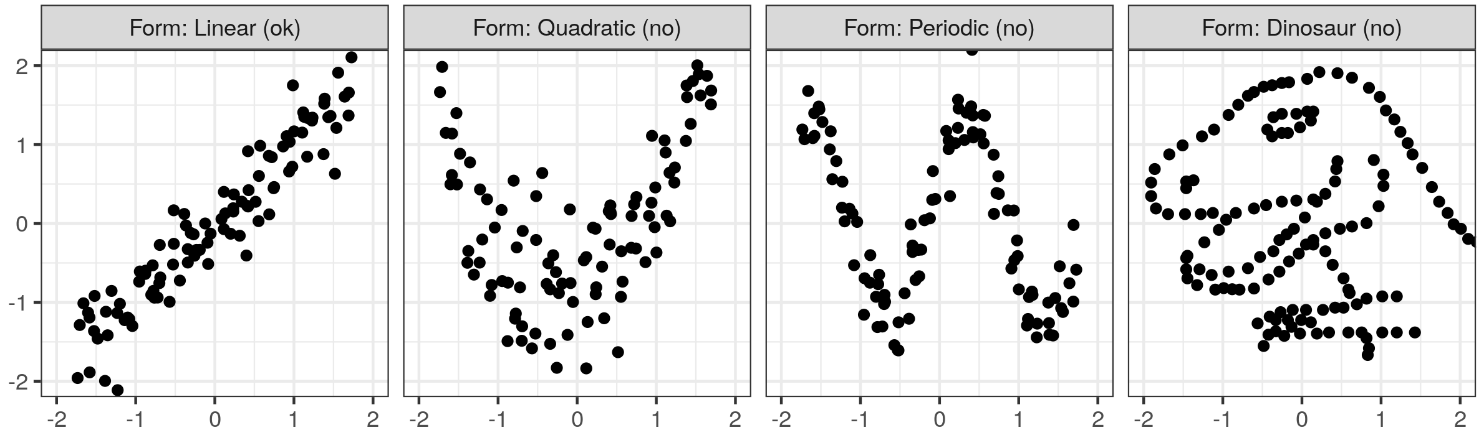
Requirements for Theory-Based Inference

Validity conditions:

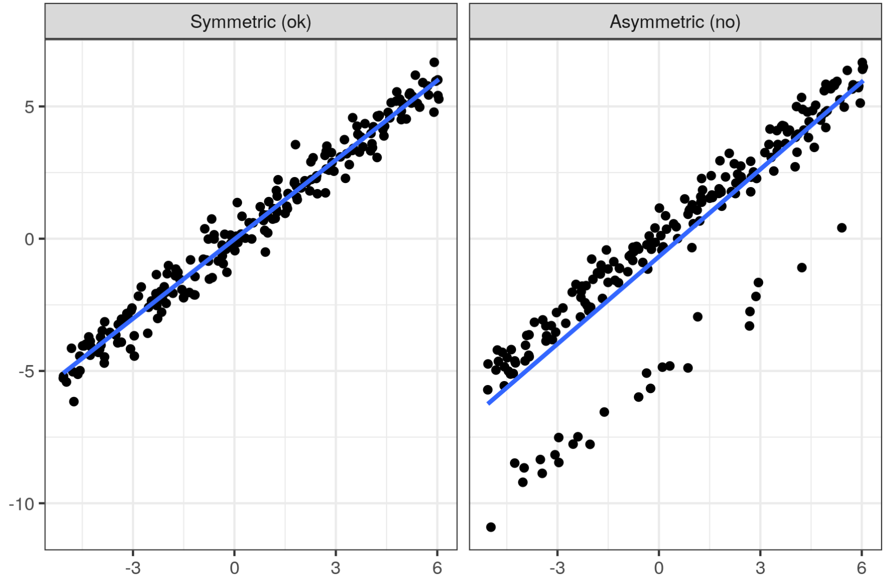
- general pattern of the points should follow a linear trend (no curved or nonlinear patterns)
- about the same number of points above and below the regression line (symmetry)
- variability of points around the regression line should be similar regardless of the value of x (equal variance)

Requirements for Theory-Based Inference

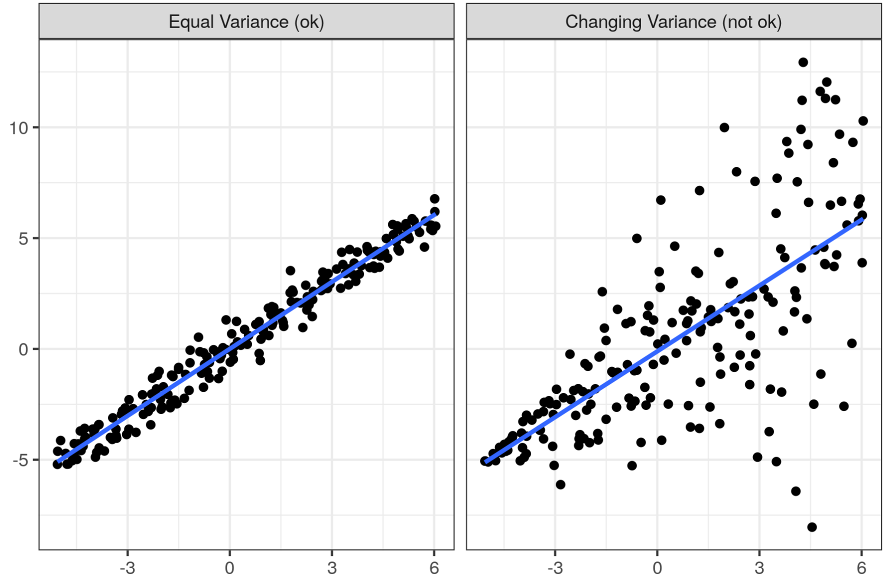
- general pattern of the points should follow a linear trend (no curved or nonlinear patterns)



Requirements for Theory-Based Inference

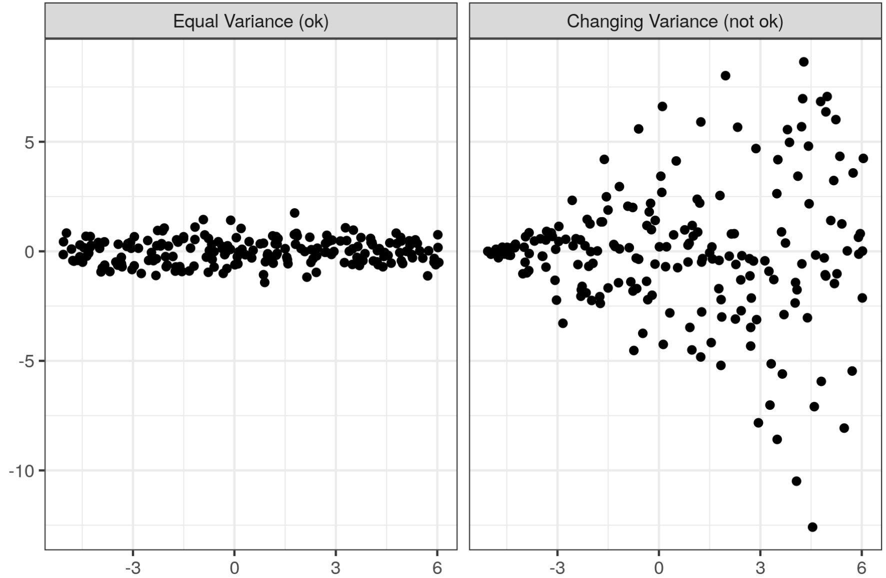


Requirements for Theory-Based Inference



Requirements for Theory-Based Inference

Looking at the residuals from your regression can be more useful:



Theory-Based Inference

We also need a formula for the standard error of our statistic, r

$$SE(r) = \sqrt{\frac{1 - r^2}{n - 2}}$$

Then,

$$t = \frac{r}{SE(r)}$$

Because t can be calculated from either the slope or the correlation coefficient, the test results will be identical.

We could also make a theory-based confidence interval for ρ using the standard error.

Theory-Based Inference

If we work with b, we will be given $SE(b)$ by statistical software:

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  325.573    47.1409    6.91 4.61e-11
## HeadSize       0.263     0.0129   20.41 5.96e-54
```

Then,

$$t = \frac{b - 0}{SE(b)}$$

Or, just get t (and the corresponding p-value) from the table…

Because t can be calculated from either the slope or the correlation coefficient, the tests are identical.

We could also make a theory-based confidence interval for β using the standard error.

Interpretations - Slope

Hypothesis test interpretation:

With t = …, I have (strength) evidence to suggest that there is a linear relationship between (explanatory variable) and (response variable). I (accept/reject) H_0 and conclude that (conclusion in context of the problem)

Confidence interval interpretation:

I am 95% confident that as explanatory variable increases by 1 unit, the predicted population average value of response variable will change by between (lower bound of b interval) and (upper bound of b interval)

Example: Head Size and Brain Weight

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  325.573    47.1409    6.91 4.61e-11
## HeadSize       0.263     0.0129   20.41 5.96e-54
```

We can see that the t value is 20.409, which corresponds to extremely strong evidence that there is a relationship between head size and brain weight.

Example: Head Size and Brain Weight

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  325.573    47.1409    6.91 4.61e-11
## HeadSize       0.263     0.0129   20.41 5.96e-54
```

If we wanted to make a 95% CI for the slope of the line (e.g. for the increase in brain weight when head size increases by 1 cm^3), we could use the 2*SE method:

$$0.263 \pm 2 * 0.0129 = (0.237, 0.289)$$

I am 95% confident that the population average brain weight will increase by between 0.237 and 0.289 grams for each additional cubic centimeter of head size.

Interpretations - Correlation Coefficient

Hypothesis test interpretation:

With t = …, I have (strength) evidence to suggest that there is a linear relationship between (explanatory variable) and (response variable). I (accept/reject) H_0 and conclude that (conclusion in context of the problem)

Confidence interval interpretation:

I am 95% confident that the population correlation between explanatory variable and response variable is between lower bound and upper bound