Stat 201: Statistics I
Chapter 10

Metropolitan State University

date

Chapter 10
Correlation and Regression

Comparing samples from two populations

Example

Consider the following data:

X	83	85	66	89	96	78
Y	90	90	66	86	99	85

How should this data be analyzed? It depends on the context of the data or what the data represents.

Comparing samples from two populations

Example

Consider the following data:

X	83	85	66	89	96	78
Y	90	90	66	86	99	85

How should this data be analyzed? It depends on the context of the data or what the data represents.

- Suppose the data is test scores from two different statistics classes.

Comparing samples from two populations

Example

Consider the following data:

X	83	85	66	89	96	78
Y	90	90	66	86	99	85

How should this data be analyzed? It depends on the context of the data or what the data represents.

- Suppose the data is test scores from two different statistics classes.
    - One possible analysis would be a two-sample t-test comparing the mean test scores from each class.

Comparing samples from two populations

Example

Consider the following data:

X	83	85	66	89	96	78
Y	90	90	66	86	99	85

How should this data be analyzed? It depends on the context of the data or what the data represents.

- Suppose the data is test scores from two different statistics classes.
  - One possible analysis would be a two-sample t-test comparing the mean test scores from each class.
- Suppose the data are scores from the midterm and the final for one set of students.

Comparing samples from two populations

Example

Consider the following data:

X	83	85	66	89	96	78
Y	90	90	66	86	99	85

How should this data be analyzed? It depends on the context of the data or what the data represents.

- Suppose the data is test scores from two different statistics classes.
  - One possible analysis would be a two-sample t-test comparing the mean test scores from each class.
- Suppose the data are scores from the midterm and the final for one set of students.
  - One possible analysis would be a matched pairs t-test comparing the mean difference between the midterm and the final for each student.

Comparing samples from two populations, cont.

Example

X	83	85	66	89	96	78
Y	90	90	66	86	99	85

- Suppose X is a students score on the statistics final and Y is that students yearly salary, in thousands of dollars, a year later.

Comparing samples from two populations, cont.

Example

X	83	85	66	89	96	78
Y	90	90	66	86	99	85

- Suppose X is a students score on the statistics final and Y is that students yearly salary, in thousands of dollars, a year later.
  - It doesn't really make sense to compare means. The two samples represent entirely different kinds of data. There is no meaningful way to compare means.

Comparing samples from two populations, cont.

Example

X	83	85	66	89	96	78
Y	90	90	66	86	99	85

- Suppose X is a students score on the statistics final and Y is that students yearly salary, in thousands of dollars, a year later.
  - It doesn't really make sense to compare means. The two samples represent entirely different kinds of data. There is no meaningful way to compare means.
  - It is useful to examine the association between the data. Is a higher test score associated with a higher salary? Or are the samples independent, values of one having no effect on values of the other?

Section 10.1
Correlation

Correlation

Correlation is the measure of how strongly associated values of paired data are with each other.

Correlation

Correlation is the measure of how strongly associated values of paired data are with each other.

Linear correlation is the measure of how strong of a linear relationship (along a straight line) values of paired data have with each other.

Correlation

Correlation is the measure of how strongly associated values of paired data are with each other.

Linear correlation is the measure of how strong of a linear relationship (along a straight line) values of paired data have with each other.

- A positive correlation indicates that as one value in a pair increases the other will tend to increase.

Correlation

Correlation is the measure of how strongly associated values of paired data are with each other.

Linear correlation is the measure of how strong of a linear relationship (along a straight line) values of paired data have with each other.

- A positive correlation indicates that as one value in a pair increases the other will tend to increase.
- A negative correlation indicates that as one value in a pair increases the other will tend to decrease.

Correlation

Correlation is the measure of how strongly associated values of paired data are with each other.

Linear correlation is the measure of how strong of a linear relationship (along a straight line) values of paired data have with each other.

- A positive correlation indicates that as one value in a pair increases the other will tend to increase.
- A negative correlation indicates that as one value in a pair increases the other will tend to decrease.
- No correlation indicates that the two values of a pair have no relationship with each other. They are independent.

Correlation coefficient

The sample **correlation coefficient**, denoted as r, is the numeric value for the strength of linear correlation between values of paired data.

Correlation coefficient

The sample **correlation coefficient**, denoted as r, is the numeric value for the strength of linear correlation between values of paired data.

- r is always between -1 and 1

Correlation coefficient

The sample **correlation coefficient**, denoted as r, is the numeric value for the strength of linear correlation between values of paired data.

- r is always between -1 and 1
- When $r = 1$, the samples have perfect positive correlation

Correlation coefficient

The sample **correlation coefficient**, denoted as r, is the numeric value for the strength of linear correlation between values of paired data.

- r is always between -1 and 1
- When $r = 1$, the samples have perfect positive correlation
- When $r = -1$, the samples have perfect negative correlation

Correlation coefficient

The sample **correlation coefficient**, denoted as r, is the numeric value for the strength of linear correlation between values of paired data.

- r is always between -1 and 1
- When $r = 1$, the samples have perfect positive correlation
- When $r = -1$, the samples have perfect negative correlation
- When $r = 0$, the samples are perfectly independent

Correlation coefficient

The sample **correlation coefficient**, denoted as r, is the numeric value for the strength of linear correlation between values of paired data.

- r is always between -1 and 1
- When $r = 1$, the samples have perfect positive correlation
- When $r = -1$, the samples have perfect negative correlation
- When $r = 0$, the samples are perfectly independent
- Most samples will have some other value for r.

Correlation coefficient

The sample **correlation coefficient**, denoted as r, is the numeric value for the strength of linear correlation between values of paired data.

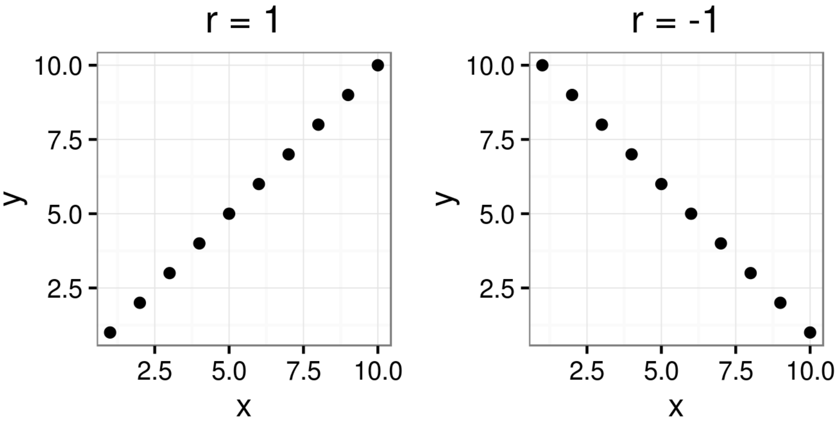
- r is always between -1 and 1
- When $r = 1$, the samples have perfect positive correlation
- When $r = -1$, the samples have perfect negative correlation
- When $r = 0$, the samples are perfectly independent
- Most samples will have some other value for r.
- The order of the values, i.e. (x, y) vs. (y, x), have no effect of the value of r.

Correlation coefficient

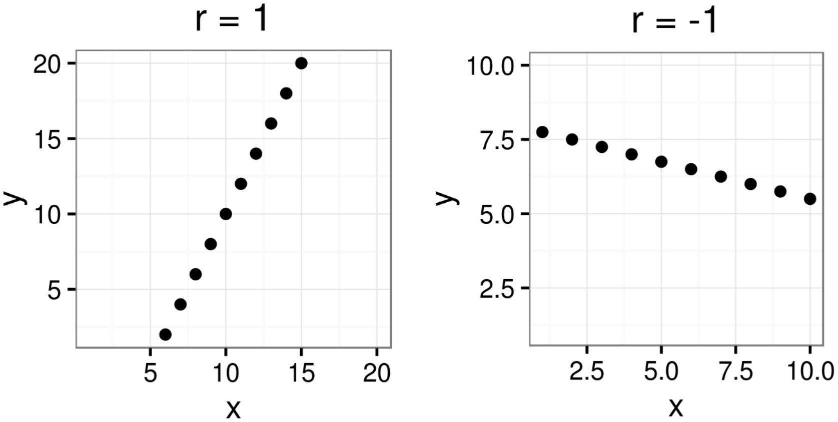
The sample **correlation coefficient**, denoted as r, is the numeric value for the strength of linear correlation between values of paired data.

- r is always between -1 and 1
- When $r = 1$, the samples have perfect positive correlation
- When $r = -1$, the samples have perfect negative correlation
- When $r = 0$, the samples are perfectly independent
- Most samples will have some other value for r.
- The order of the values, i.e. (x, y) vs. (y, x), have no effect of the value of r.
- The units of the values also have no effect on r. A correlation on height will be the same whether it is measured in inches, feet or meters, for example.

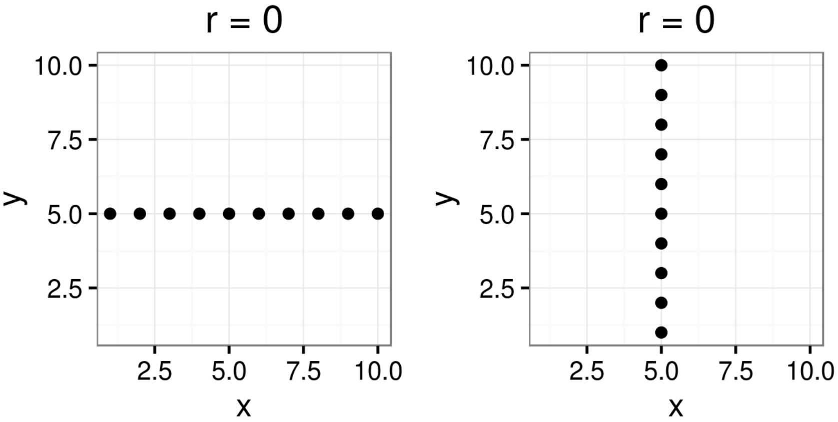
Perfect correlation



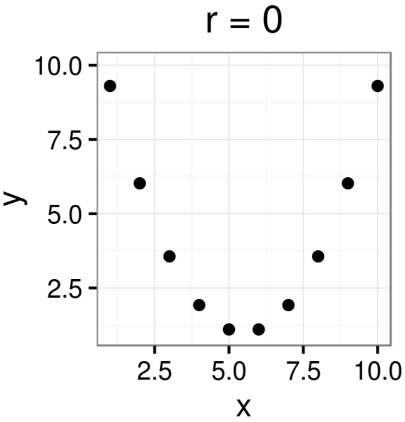
Perfect correlation, cont.



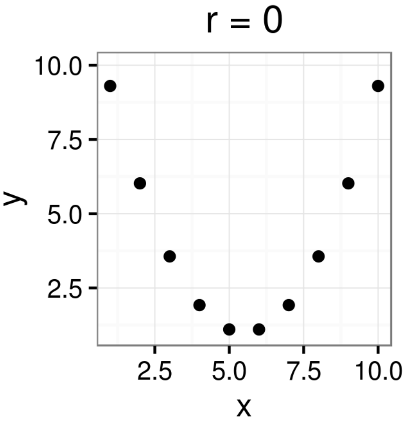
Zero correlation

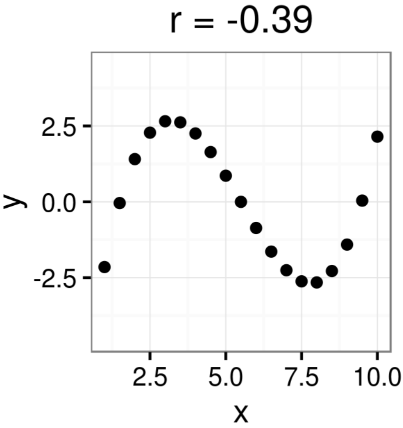


Non-linear correlation

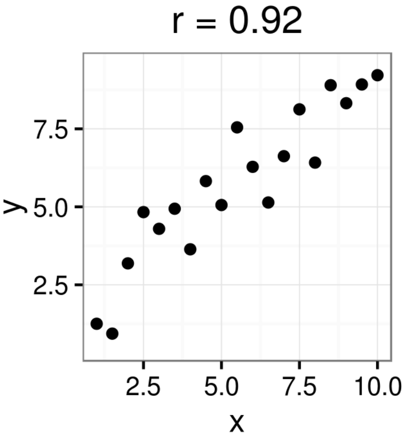


Non-linear correlation

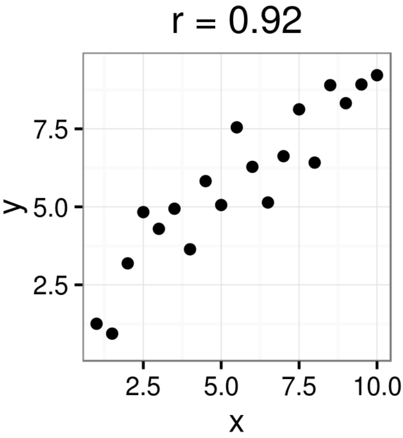


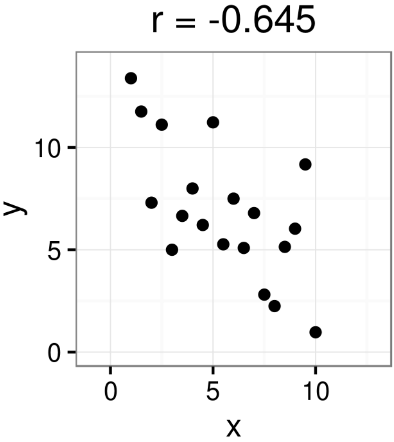


"Real world" correlation

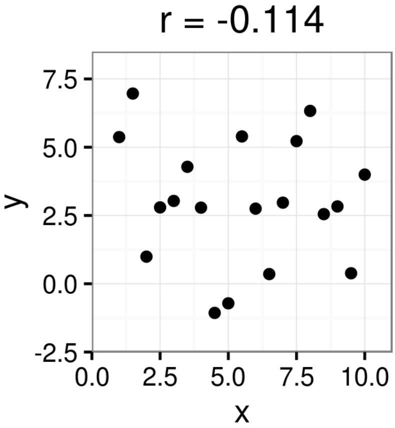


"Real world" correlation



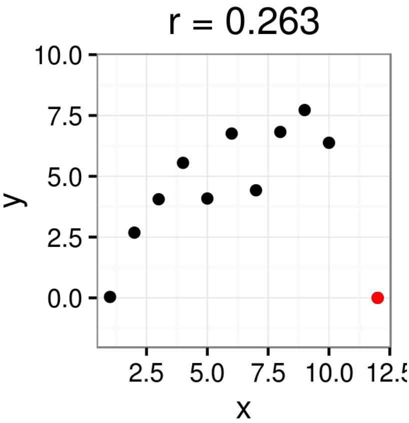


"Real world" correlation, cont.



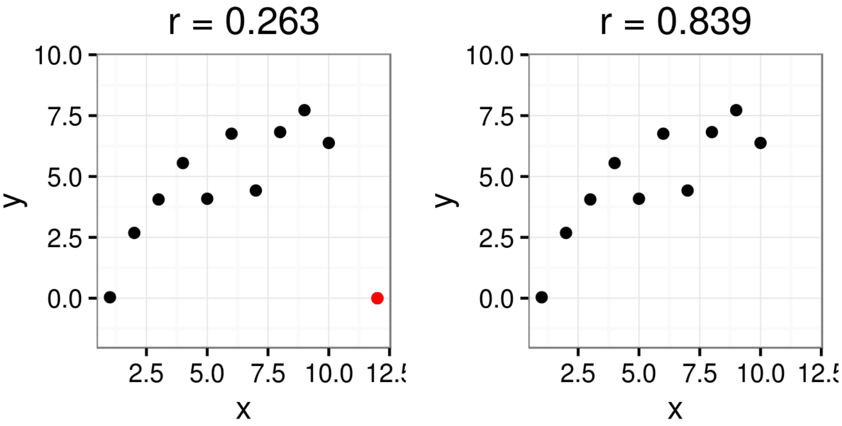
Correlation and outliers

The correlation coefficient r is sensitive to outliers (very unusual values).



Correlation and outliers

The correlation coefficient r is sensitive to outliers (very unusual values).



Calculating correlation coefficient

The book uses the following formula for calculating r:

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

Calculating correlation coefficient

The book uses the following formula for calculating r:

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

- While this formula will give you a value for r (after much work), it doesn't help much in understanding how the correlation coefficient works.

Calculating correlation coefficient

The book uses the following formula for calculating r:

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

- While this formula will give you a value for r (after much work), it doesn't help much in understanding how the correlation coefficient works.
- Since technology should be used to calculate the correlation calculation anyway, it would be useful to look at r defined with a more informative formula.

Calculating correlation coefficient, cont.

Consider: ($\propto$ means "is proportional to")

$$r \propto \sum (x - \bar{x})(y - \bar{y})$$

Calculating correlation coefficient, cont.

Consider: ($\propto$ means "is proportional to")

$$r \propto \sum (x - \bar{x})(y - \bar{y})$$

From this second formula is can be seen that:

Calculating correlation coefficient, cont.

Consider: ($\propto$ means "is proportional to")

$$r \propto \sum (x - \bar{x})(y - \bar{y})$$

From this second formula is can be seen that:

- If when x values are big (greater than $\bar{x}$) then y values are big (greater than $\bar{y}$), and vice versa ($x < \bar{x} \rightarrow y < \bar{y}$), then r is proportional to the sum of positive numbers, i.e. $r > 0$.

Calculating correlation coefficient, cont.

Consider: ($\propto$ means "is proportional to")

$$r \propto \sum (x - \bar{x})(y - \bar{y})$$

From this second formula is can be seen that:

- If when x values are big (greater than $\bar{x}$) then y values are big (greater than $\bar{y}$), and vice versa ($x < \bar{x} \rightarrow y < \bar{y}$), then r is proportional to the sum of positive numbers, i.e. $r > 0$.
- If when x values are big (greater than $\bar{x}$) then y values are small (less than $\bar{y}$), and vice versa ($x < \bar{x} \rightarrow y > \bar{y}$), then r is proportional to the sum of negative numbers, i.e. $r < 0$.

Calculating correlation coefficient, cont.

Consider: ($\propto$ means "is proportional to")

$$r \propto \sum (x - \bar{x})(y - \bar{y})$$

From this second formula is can be seen that:

- If when x values are big (greater than $\bar{x}$) then y values are big (greater than $\bar{y}$), and vice versa ($x < \bar{x} \rightarrow y < \bar{y}$), then r is proportional to the sum of positive numbers, i.e. $r > 0$.
- If when x values are big (greater than $\bar{x}$) then y values are small (less than $\bar{y}$), and vice versa ($x < \bar{x} \rightarrow y > \bar{y}$), then r is proportional to the sum of negative numbers, i.e. $r < 0$.
- If when x values are far away from $\bar{x}$ then y values are also far away from $\bar{y}$ (whether positive or negative), then r is proportional to the sum of larger numbers and will be closer to 1 or -1.

Calculating correlation coefficient, example

Example

X	2	4	6	8	$\bar{x} = 5$
Y	1	6	7	12	$\bar{y} = 6.5$

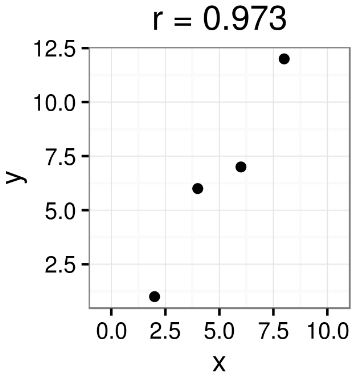
U	2	4	6	8	$\bar{u} = 5$
W	7	12	1	6	$\bar{w} = 6.5$

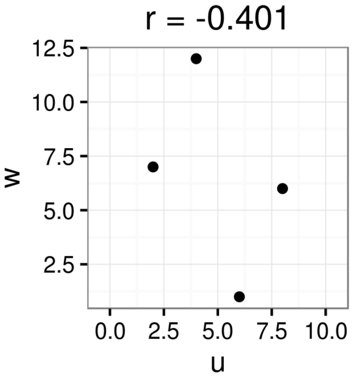
Calculating correlation coefficient, example

Example

X	2	4	6	8	$\bar{x} = 5$
Y	1	6	7	12	$\bar{y} = 6.5$

U	2	4	6	8	$\bar{u} = 5$
W	7	12	1	6	$\bar{w} = 6.5$





Calculating correlation coefficient, example

Example

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
2	1	-3	-5.5	16.5
4	6	-1	-0.5	0.5
6	7	1	0.5	0.5
8	12	3	5.5	16.5
				$r \propto 34$

Calculating correlation coefficient, example

Example

u	w	$(u-\bar{u})$	$(w-\bar{w})$	$(u-\bar{u})(w-\bar{w})$
2	7	-3	0.5	-1.5
4	12	-1	5.5	-5.5
6	1	1	-5.5	-5.5
8	6	3	-0.5	-1.5
				$r \propto -14$

Correlation hypothesis tests

The sample correlation coefficient r is an unbiased estimator for the population correlation coefficient ρ (rho).

Like population proportions and means, hypothesis tests can be conducted of the parameter ρ to determine if a population of paired data is correlated.

Correlation hypothesis tests

The sample correlation coefficient r is an unbiased estimator for the population correlation coefficient ρ (rho).

Like population proportions and means, hypothesis tests can be conducted of the parameter ρ to determine if a population of paired data is correlated.

- $H_0 : \rho = 0$

Correlation hypothesis tests

The sample correlation coefficient r is an unbiased estimator for the population correlation coefficient ρ (rho).

Like population proportions and means, hypothesis tests can be conducted of the parameter ρ to determine if a population of paired data is correlated.

- $H_0 : \rho = 0$
- $H_a : \rho \neq 0$
  (It is uncommon to use the other forms of the alternative hypothesis)

Correlation hypothesis tests

The sample correlation coefficient r is an unbiased estimator for the population correlation coefficient ρ (rho).

Like population proportions and means, hypothesis tests can be conducted of the parameter ρ to determine if a population of paired data is correlated.

- $H_0 : \rho = 0$
- $H_a : \rho \neq 0$
  (It is uncommon to use the other forms of the alternative hypothesis)
- Test statistic: $t = \dfrac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ $\qquad df = n - 2$

Correlation hypothesis tests

The sample correlation coefficient r is an unbiased estimator for the population correlation coefficient ρ (rho).

Like population proportions and means, hypothesis tests can be conducted of the parameter ρ to determine if a population of paired data is correlated.

- $H_0 : \rho = 0$
- $H_a : \rho \neq 0$
  (It is uncommon to use the other forms of the alternative hypothesis)
- Test statistic: $t = \dfrac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ $\qquad df = n - 2$
- Make a decision, two ways:
  - P-value: If $p < \alpha$, reject null hypothesis
  - Critical value (from table A-6): If $|t| > t_{crit}$, reject null hypothesis

Correlation coefficient and p-value in StatCrunch

- Stat → Summary Stats → Correlation
- Select the columns which contain the data
- Check "Two-sided P-value"
- Click "Compute!"
- The correlation coefficient r is given, the p-value is in parentheses

Note: The test statistic for the hypothesis test can be found in StatCrunch by doing a regression, discussed in the next section.

Strength of correlation

Even with statistically significant correlation, it might be useful to specify the strength of the correlation. While, like many things in statistics, the description of strength can be subjective and dependent of the context of the data, the following rule of thumb may be used:

Strength of correlation

Even with statistically significant correlation, it might be useful to specify the strength of the correlation. While, like many things in statistics, the description of strength can be subjective and dependent of the context of the data, the following rule of thumb may be used:

- $|r| \geq 0.7$ is a strong correlation

Strength of correlation

Even with statistically significant correlation, it might be useful to specify the strength of the correlation. While, like many things in statistics, the description of strength can be subjective and dependent of the context of the data, the following rule of thumb may be used:

- $|r| \geq 0.7$ is a strong correlation
- $0.4 \leq |r| < 0.7$ is a moderate correlation

Strength of correlation

Even with statistically significant correlation, it might be useful to specify the strength of the correlation. While, like many things in statistics, the description of strength can be subjective and dependent of the context of the data, the following rule of thumb may be used:

- $|r| \geq 0.7$ is a strong correlation
- $0.4 \leq |r| < 0.7$ is a moderate correlation
- $|r| < 0.4$ is a weak correlation

Strength of correlation

Even with statistically significant correlation, it might be useful to specify the strength of the correlation. While, like many things in statistics, the description of strength can be subjective and dependent of the context of the data, the following rule of thumb may be used:

- $|r| \geq 0.7$ is a strong correlation
- $0.4 \leq |r| < 0.7$ is a moderate correlation
- $|r| < 0.4$ is a weak correlation

Thus, $r = 0.5$ indicates a moderate positive correlation and $r = -0.8$ indicates a strong negative correlation.

Correlation, example

Example

In 1886, Sir Francis Galton, a British sociologist, published the paper "Regression towards Mediocrity in Hereditary Statures", in which he examined the heights of parents and their adult children. The core of modern uses of correlation and regression come from this paper (he also invented standard deviation). His data for fathers and sons is in the file "Galton-father-son.csv" on D2L.

Is there correlation between heights of fathers and their adult sons? Test at $\alpha = 0.05$ level of significance.

Correlation, example

Example

In 1886, Sir Francis Galton, a British sociologist, published the paper "Regression towards Mediocrity in Hereditary Statures", in which he examined the heights of parents and their adult children. The core of modern uses of correlation and regression come from this paper (he also invented standard deviation). His data for fathers and sons is in the file "Galton-father-son.csv" on D2L.

Is there correlation between heights of fathers and their adult sons? Test at $\alpha = 0.05$ level of significance.

- $H_0 : \rho = 0$
  $H_a : \rho \neq 0$

Correlation, example

Example

In 1886, Sir Francis Galton, a British sociologist, published the paper "Regression towards Mediocrity in Hereditary Statures", in which he examined the heights of parents and their adult children. The core of modern uses of correlation and regression come from this paper (he also invented standard deviation). His data for fathers and sons is in the file "Galton-father-son.csv" on D2L.

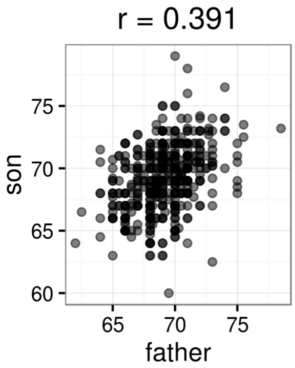
Is there correlation between heights of fathers and their adult sons? Test at $\alpha = 0.05$ level of significance.

- $H_0 : \rho = 0$
  $H_a : \rho \neq 0$
- $r = 0.39131736$
  $p < 0.0001$

Correlation, example

Example

- $p < 0.0001 < \alpha = 0.5$. Reject null hypothesis
- There is evidence that there is a correlation between the heights of fathers and their adult sons. However, $r < 0.4$ indicates weak correlation.



Coefficient of determination

The **coefficient of determination**, designated by R^2, represents the proportion of variation in one variable that is explained by the association with the other variable.

- The coefficient of determination is the correlation coefficient squared, $R^2 = r^2$

Coefficient of determination

The **coefficient of determination**, designated by R^2, represents the proportion of variation in one variable that is explained by the association with the other variable.

- The coefficient of determination is the correlation coefficient squared, $R^2 = r^2$

Example

In the previous example, heights of fathers and sons had a correlation coefficient of $r = 0.391$. Thus, $R^2 = (0.391)^2 = 0.153$.

About 15% of the variation of the heights of adult men can be explained by the association with their fathers heights.

Cautions

If a sample of paired data has a correlation coefficient that is zero or very low, that does not necessarily mean that there is not a association between the variables, only that there is not a *linear* association.

Cautions

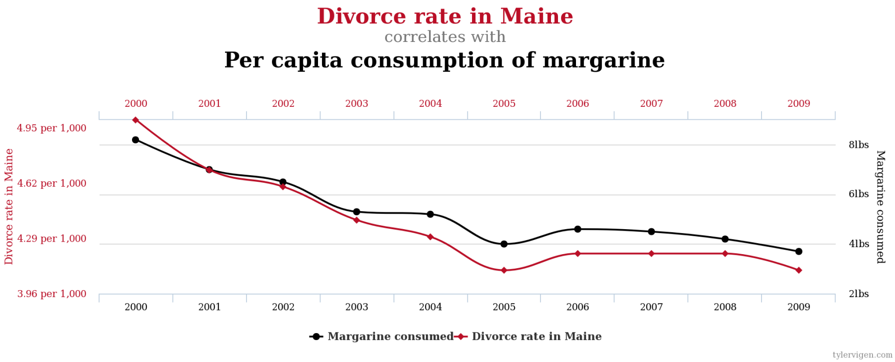
If a sample of paired data has a correlation coefficient that is zero or very low, that does not necessarily mean that there is not a association between the variables, only that there is not a *linear* association.

!!!

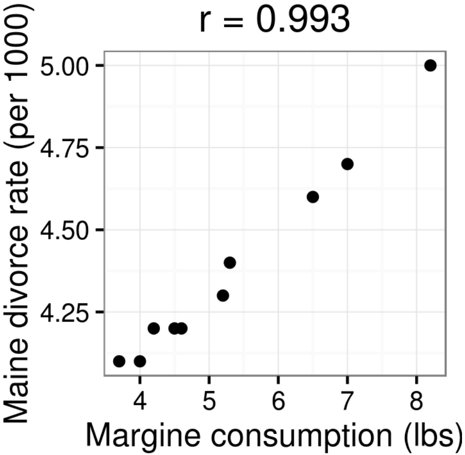
Correlation does not imply causation.

Spurious correlations

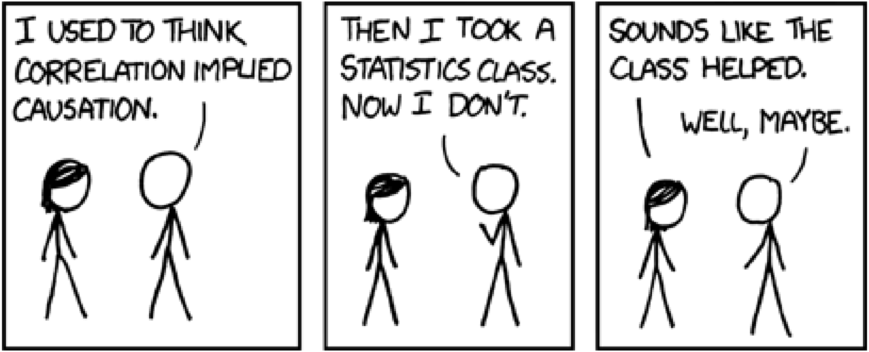
From the *Spurious Correlations* website:
`http://www.tylervigen.com/spurious-correlations`



Spurious correlations, cont.



Correlation does not imply causation



Group work

- For all the questions, complete parts (a) and (b).

Section 10.2
Regression

Regression

Once it has been determined that two variables have a linear relationship, that is they have an association described by a straight line, the next logical question is what is that line.

Regression

Once it has been determined that two variables have a linear relationship, that is they have an association described by a straight line, the next logical question is what is that line.

Regression is the statistical technique for finding the line that best describes a linear relationship between two paired variables.

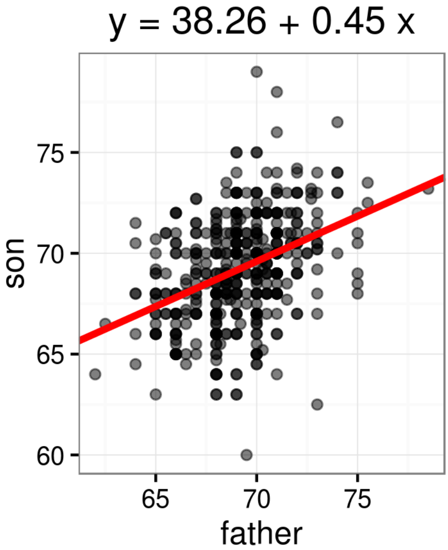
Regression

Once it has been determined that two variables have a linear relationship, that is they have an association described by a straight line, the next logical question is what is that line.

Regression is the statistical technique for finding the line that best describes a linear relationship between two paired variables.

The line found is known as the **regression line** or the line of best fit.

Regression, example



Algebra review, lines

The equation for a line generally has the following form:

$$y = b + mx$$

- b is the y-intercept, or where the line crosses the y-axis ($x = 0$).
- m is the slope of the line. It is the amount the y value increases as the x value increase by one.

Regression population models

A linear relationship between populations of variables X and Y can be described by:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Regression population models

A linear relationship between populations of variables X and Y can be described by:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- X is known as the predictor variable, or the explanatory variable, or the independent variable.

Regression population models

A linear relationship between populations of variables X and Y can be described by:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- X is known as the predictor variable, or the explanatory variable, or the independent variable.
- Y is the response variable, or the dependent variable.

Regression population models

A linear relationship between populations of variables X and Y can be described by:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- X is known as the predictor variable, or the explanatory variable, or the independent variable.
- Y is the response variable, or the dependent variable.
- β_0 and β_1 are the intercept and slope of a line describing the association of X and Y.

Regression population models

A linear relationship between populations of variables X and Y can be described by:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- X is known as the predictor variable, or the explanatory variable, or the independent variable.
- Y is the response variable, or the dependent variable.
- β_0 and β_1 are the intercept and slope of a line describing the association of X and Y.
- Like other population parameters, β_0 and β_1 are thought of as fixed, but unknown

Regression population models

A linear relationship between populations of variables X and Y can be described by:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- X is known as the predictor variable, or the explanatory variable, or the independent variable.
- Y is the response variable, or the dependent variable.
- β_0 and β_1 are the intercept and slope of a line describing the association of X and Y.
- Like other population parameters, β_0 and β_1 are thought of as fixed, but unknown
- ϵ (epsilon) is a random error term. It is usually assumed that $\epsilon \sim N(0, \sigma^2)$.

Regression lines

The regression line for a sample paired data (x, y) describes the linear relationship between x and y. It is given by the equation:

$$\hat{y} = b_0 + b_1 x$$

Regression lines

The regression line for a sample paired data (x, y) describes the linear relationship between x and y. It is given by the equation:

$$\hat{y} = b_0 + b_1 x$$

- $\hat{y}$ is the estimated y value for a given x, or the predicted response.

Regression lines

The regression line for a sample paired data (x, y) describes the linear relationship between x and y. It is given by the equation:

$$\hat{y} = b_0 + b_1 x$$

- $\hat{y}$ is the estimated y value for a given x, or the predicted response.
- b_1, the slope of the regression line, is the amount of change in the predicted response for each unit increase in x

Regression lines

The regression line for a sample paired data (x, y) describes the linear relationship between x and y. It is given by the equation:

$$\hat{y} = b_0 + b_1 x$$

- $\hat{y}$ is the estimated y value for a given x, or the predicted response.
- b_1, the slope of the regression line, is the amount of change in the predicted response for each unit increase in x
- b_1 can be calculated from the correlation coefficient, $b_1 = r\dfrac{s_y}{s_x}$

Regression lines

The regression line for a sample paired data (x, y) describes the linear relationship between x and y. It is given by the equation:

$$\hat{y} = b_0 + b_1 x$$

- $\hat{y}$ is the estimated y value for a given x, or the predicted response.
- b_1, the slope of the regression line, is the amount of change in the predicted response for each unit increase in x
- b_1 can be calculated from the correlation coefficient, $b_1 = r\dfrac{s_y}{s_x}$
- b_0, the y-intercept of the regression line, is usually not of much interest.

Predictions

If there is no correlation between values of paired data (if failed to reject null hypothesis for correlation test), then the best predictor for the response variable is simply $\bar{y}$.

Predictions

If there is no correlation between values of paired data (if failed to reject null hypothesis for correlation test), then the best predictor for the response variable is simply $\bar{y}$.

However, if there is a valid regression equation, it can be used to make predictions of the response variable for given values of the predictor variable. Replace x in the equation with the given predictor value and calculate the predicted response $\hat{y}$.

Predictions, example

Example

The regression line equation from the Galton data, for fathers height as predictor x and sons height as response y, is

$$\hat{y} = 38.26 + 0.45x$$

What is the predicted adult height of a son whose father is 68 inches tall?

Predictions, example

Example

The regression line equation from the Galton data, for fathers height as predictor x and sons height as response y, is

$$\hat{y} = 38.26 + 0.45x$$

What is the predicted adult height of a son whose father is 68 inches tall?

- $\hat{y} = 38.26 + 0.45 \times 68 = 68.86$ inches

Extrapolating

Predictions for predictor values outside the range of x values used to find the regression line are highly suspect. This is known as **extrapolating** and should be avoided.

Extrapolating

Predictions for predictor values outside the range of x values used to find the regression line are highly suspect. This is known as **extrapolating** and should be avoided.

Example

For the 2016-2017 season, NBA teams had field goal percentages between 43.5% and 49.5%. A regression line of the relationship between FG% (x) and numbers of games won (y) is

$$\hat{y} = -208.69 + 5.46x$$

The GM for the Timberwolves wants to know how many games the team would win if they could get their FG% up to 60%, based on this data.

Extrapolating

Predictions for predictor values outside the range of x values used to find the regression line are highly suspect. This is known as **extrapolating** and should be avoided.

Example

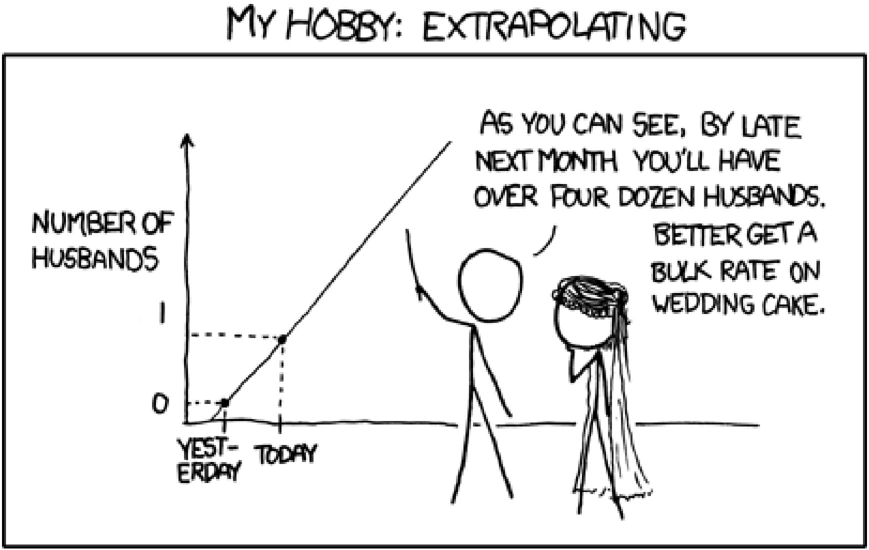
For the 2016-2017 season, NBA teams had field goal percentages between 43.5% and 49.5%. A regression line of the relationship between FG% (x) and numbers of games won (y) is

$$\hat{y} = -208.69 + 5.46x$$

The GM for the Timberwolves wants to know how many games the team would win if they could get their FG% up to 60%, based on this data.

- $\hat{y} = -208.69 + 5.46 \times 60 = 118.91$ games won

Extrapolating, example



Regression in StatCrunch

- Stat → Regression → Simple Linear
- Select columns for X and Y variables
- Select "Hypothesis tests" (default values are fine)
- If desired, enter "X value(s)" for "Prediction of Y"
- If desired, select graphs to generate (the default "Fitted line plot" is usually best)
- Click "Compute!"

Interpreting regression results in StatCrunch

- "Simple linear regression results" section contains:
  - The regression equation in the form of "YVAR = intercept + slope XVAR"
  - The correlation coefficient r as "R (correlation coefficient)"
  - The coefficient of determination R^2 as "R-sq"
- The "Parameter estimates" table contains:
  - The estimates for intercept (b_0) and slope (b_1)
  - The t statistic and p-value for the slope are the same as for a correlation hypothesis test with this data
- Ignore "Analysis of variance table for regression model"
- If predicted values were asked for, they will be in the "Predicted values" table
- Click on the right arrow at the bottom of the results window to get to any graphs that were selected

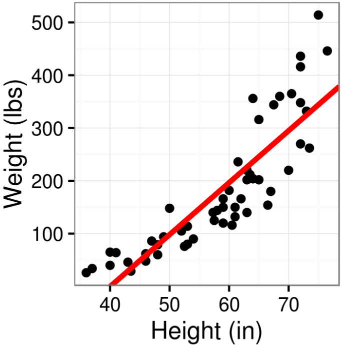
Regression, example

Example

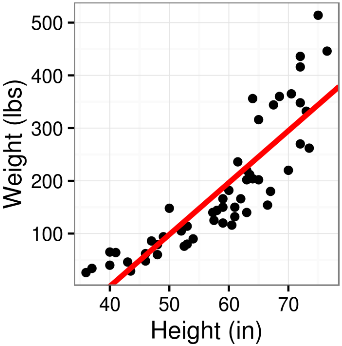
The Department of Natural Resources wishes to track the weight of bears in the wild. While it is very difficult to weigh a bear, it is fairly easy to estimate the length of bear using photos. The data set "bears.csv" on D2L contains measurements made from anesthetized wild bears.

- Find a relationship, if any, between the length and weight of bears using the data.
- What is the best predicted weight of a bear thought to be 71 inches long?
- Would it be appropriate to predict the weight of a bear 39 inches long? 89 inches?

Regression, example



Regression, example



Example

From the StatCrunch results:

- $r = 0.864,\ R^2 = 0.747.$
- About 75% of the variation in bear weight is explained by the association with bear height.

Regression, example

Example

- Regression line equation: $\hat{y} = -393.84 + 9.84x$

Regression, example

Example

- Regression line equation: $\hat{y} = -393.84 + 9.84x$
- For every inch long a bear is, its weight will increase by about 10 lbs.

Regression, example

Example

- Regression line equation: $\hat{y} = -393.84 + 9.84x$
- For every inch long a bear is, its weight will increase by about 10 lbs.
- A bear that is 71 inches long should weigh: $\hat{y} = -393.84 + 9.84 \times 71 = 304.7282$ lbs.

Regression, example

Example

- Regression line equation: $\hat{y} = -393.84 + 9.84x$
- For every inch long a bear is, its weight will increase by about 10 lbs.
- A bear that is 71 inches long should weigh: $\hat{y} = -393.84 + 9.84 \times 71 = 304.7282$ lbs.
- It would be appropriate to predict the weight of a bear that is 39 inches long, because 39 is within the range of lengths used to find the regression line (36, 76.5).

Regression, example

Example

- Regression line equation: $\hat{y} = -393.84 + 9.84x$
- For every inch long a bear is, its weight will increase by about 10 lbs.
- A bear that is 71 inches long should weigh: $\hat{y} = -393.84 + 9.84 \times 71 = 304.7282$ lbs.
- It would be appropriate to predict the weight of a bear that is 39 inches long, because 39 is within the range of lengths used to find the regression line (36, 76.5).
- 89 inches is not in that range, so it would not be appropriate to try to predict the weight of a 89 inch long bear.

Group work

- For all the questions, complete parts (c) and (d).