Correlation and LSR

Donna Dietz

American University

dietz@american.edu

STAT 202 - Spring 2020

Correlation

We can calculate (usually using a computer) a value r which we call *correlation* on a set of data points.

Correlation facts:

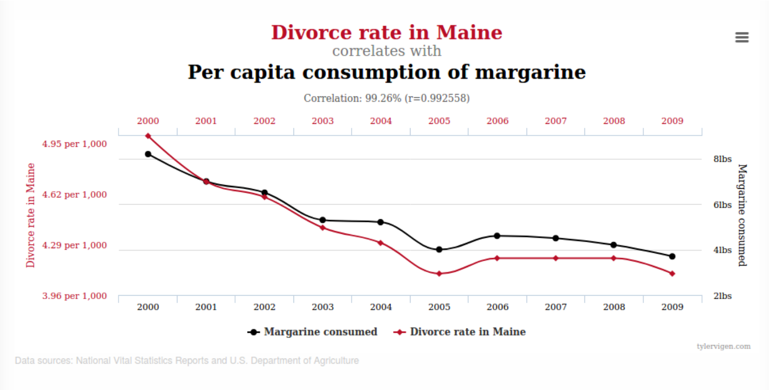
- Correlations can range from -1 to 1, inclusive
- For correlations of 1 or -1, all the points lie on a line
- A correlation of 0 means no linear association of the variables
- We usually compare data against a line unless we say otherwise
- Low correlations don't necessarily mean no association

Correlation does not imply causation

Warning:

If two variables are correlated, this does not imply that one causes the other. They may be influenced by a third *lurking* variable, or the correlation may also be *spurious*. (There are enough data sets floating around that some will appear to have an uncanny similarity for no real reason.)

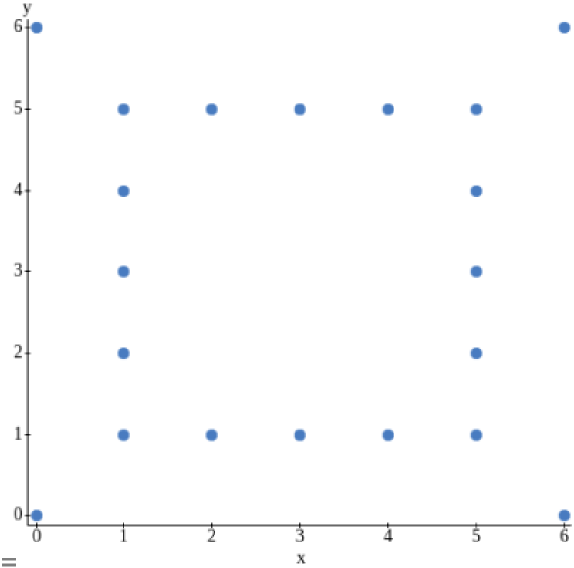
Spurious correlations



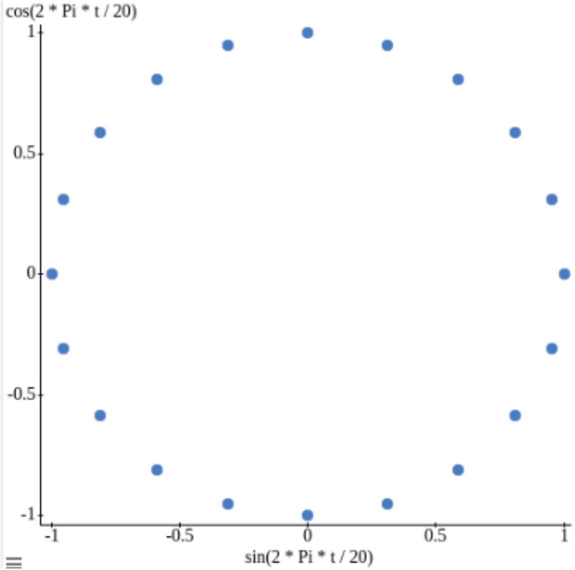
Lurking variables

Lurking variables can cause two data sets to really be connected, and we may think they are spurious. For example, white cars get more speeding tickets than other colors (according to cnet.com), followed by red cars. However, the people who choose those colors may be more likely to drive fast. Or, perhaps police really are more likely to stop white cars. Who knows!? But something like this is more likely to be caused by a lurking variable of some sort.

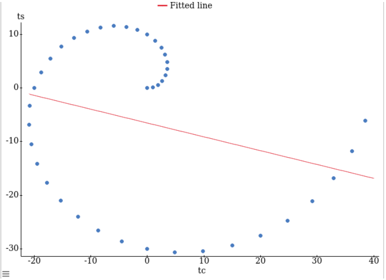
Examples of zero correlation



Examples of zero correlation



Example

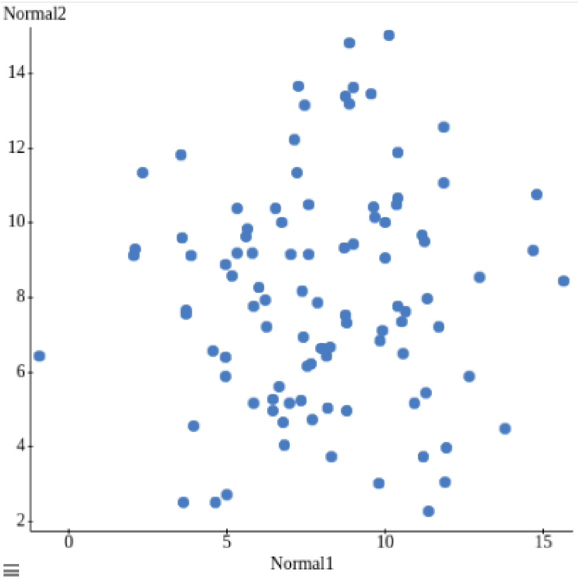


This technically has a correlation of $r = -.28$ but what does this mean? The correct model would fit perfectly. If you extended this spiral for several more turns, the linear correlation would drop even more. Here, low correlation means you have the wrong model. There is a relationship between these variables! But a line does not capture the relationship!

More about correlation

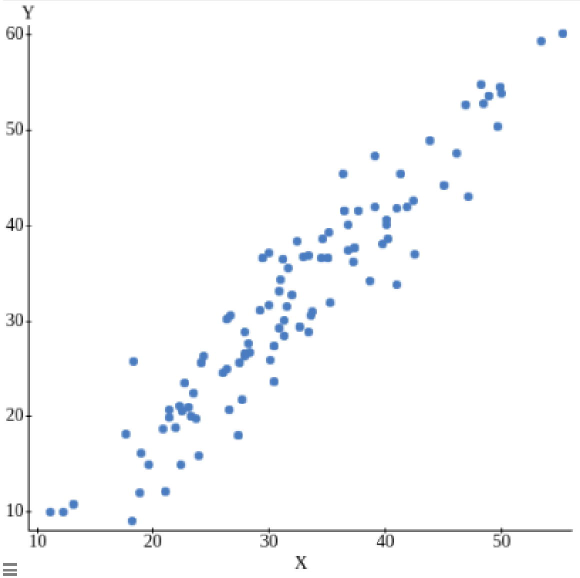
- For small data sets: BE CAREFUL
- The more significant the correlation, the farther from zero r is
- Positive correlation gives best-fit lines with positive slope
- Negative correlation gives best-fit lines with negative slope

Examples

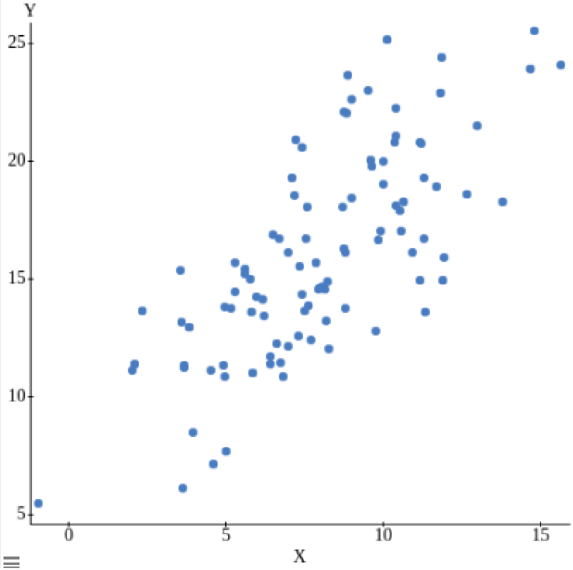


r=0.05

Examples

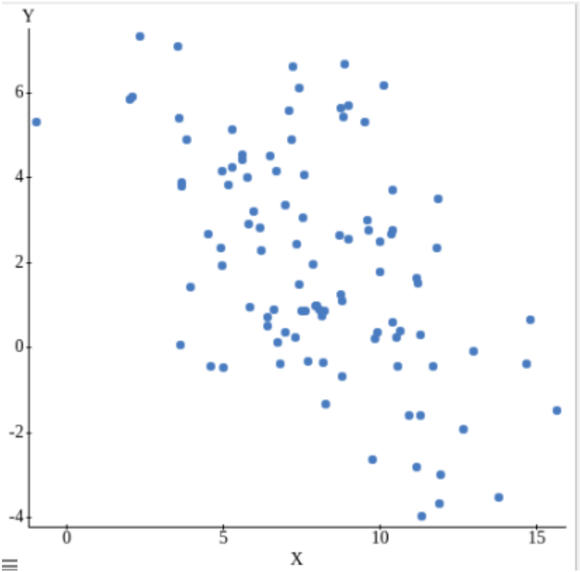


r=0.95



r=0.73

Examples

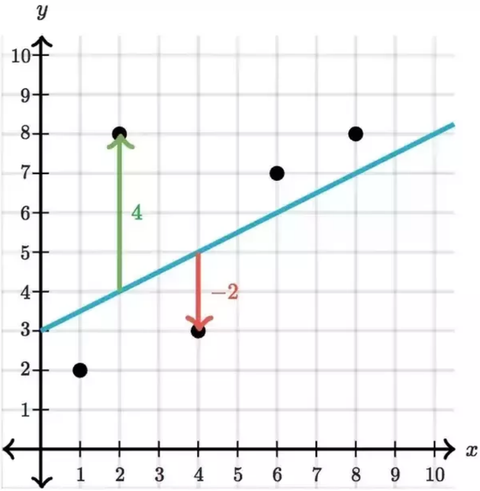


r= -0.54

Best-Fit Lines

Best-fit lines or LSR (Least Squares Regression) means that we find a line that decreases the squares of the residuals. The arithmetic for calculating residuals and squaring them is very much the same process we used to calculate variance and standard deviations.

Residuals



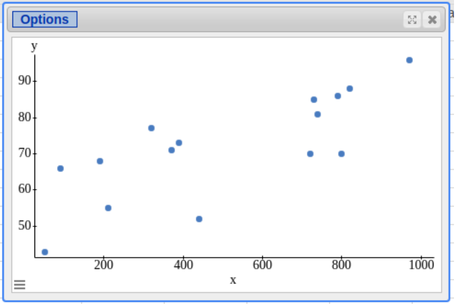
quora.com

Eyeballing it

You can often find a reasonable best-fit line by eyeballing it, and to be honest, your eyeball-fit line may be every bit as good at estimating the relationship as your super-duper computer-generated one. However, this is such a common task, we need to be able to pass it off to a computer, and to get repeatable results.

LSR by eyeball

x	y
970	96
740	81
90	66
210	55
820	88
190	68
320	77
390	73
50	43
370	71
790	86
800	70
440	52
720	70
730	85



Where do you think the line should go?

LSR by computer

Options (1 of 2)

Simple linear regression results:
Dependent Variable: Y
Independent Variable: X
Y = 5.7239042 - 0.46392592 X
Sample size: 100
R (correlation coefficient) = -0.53622318
R-sq = 0.2875353
Estimate of error standard deviation: 2.2229383

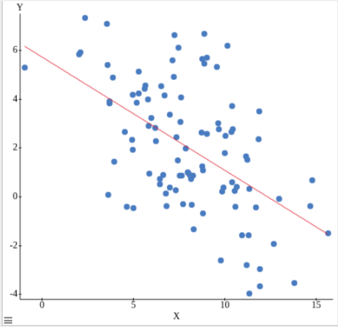
Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	5.7239042	0.63201147	≠ 0	98	9.056646	<0.0001
Slope	-0.46392592	0.073768625	≠ 0	98	-6.2889326	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	195.43786	195.43786	39.550674	<0.0001
Error	98	484.26256	4.9414547		
Total	99	679.70043			

Options



Interpolation and Extrapolation

Estimations based on models come in two broad categories:
Interpolation: Estimates within the domain of the given data
Extrapolation: Estimates outside the domain of the given data

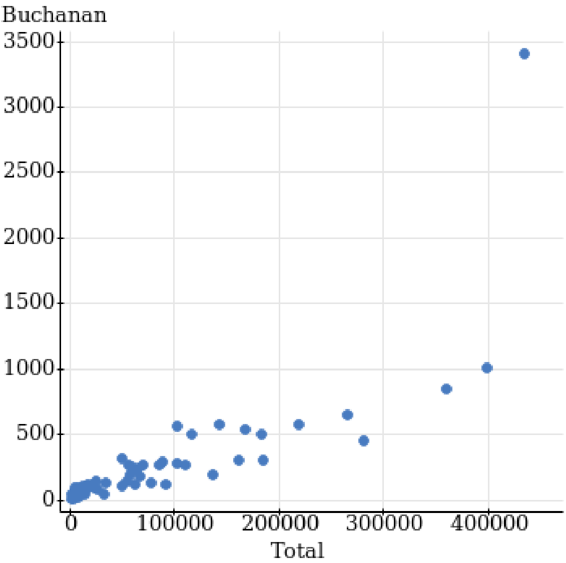
Which is better?

If you have good solid data, interpolations are often relatively safe. However, the farther away from the data you get, often the less accurate your predictions will become. Think of the weather predictions.

Exceptions to this are really awesome

There are, of course, amazing exceptions. We know with great precision when Halley's comet will pass by the earth, based on computations taken a long time past! It passes by only about once every 75 years! But you should understand that data from the past typically do not accurately predict the future!

Influential outliers



By county in Florida: Total votes vs. Buchanan votes

Florida

In the presidential election of 2000, a confusing ballot caused a third party candidate to receive many more votes than he should have. This graphic demonstrates evidence for this claim. Those votes lost due to this error were part of what caused one party to lose the presidency that year.

If you drew an LSR line on that scatterplot with and without the point from Palm Beach County, you would get a different line!

This is a real-world example of an influential outlier!

Reminders about lines

If you sketch a line by hand, you will want to estimate the slope and intercept.

To estimate the intercept, draw the line until it hits the y-axis.

To estimate the slope, pick two convenient points on your sketch:

$$slope = \frac{y_2 - y_1}{x_2 - x_1}$$

$$y = intercept + slope \cdot x$$

No more $y = mx + b$

For whatever reason, statisticians insist on writing lines this way:

$$responseVariable = intercept + slope \cdot explanatoryVariable$$

This is simplified as $y = a + bx$ and often confuses students, because suddenly the letter b is being used as the slope rather than the intercept, which they are used to. So avoid using the letter b altogether and just "B" careful!

Predictions

If you have a line $y = P + Q \cdot x$ and you have an x value you wish to find a y for, just push the x through the formula as usual.

Example: $y = 1 + 3x$ could be your LSR line and you wish to predict what y will go with $x = 4$.

$$y = 1 + 3 \cdot 4 = 13$$

"reverse" predictions

If you have a line $y = P + Q \cdot x$ and you have an y value you wish to find an x for, you need to solve for that x value.

Example: $y = 1 + 3x$ could be your LSR line and you wish to predict what x will go with $y = 7$.

$$7 = 1 + 3x$$

$$6 = 3x$$

$$2 = x$$

Significance

This fits together with correlation very nicely. When you create an LSR line on StatCrunch, it gives you all the information together: A LSR line, correlation (which it calls R instead of r), and R^2, as well as p-values.

We will look at the p-values for the Intercept and Slope to see how significant these are.

LSR Significant Intercept

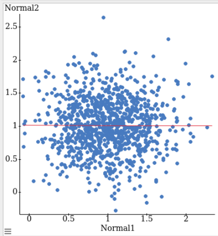
Simple linear regression results:
Dependent Variable: Normal2
Independent Variable: Normal1
Normal2 = 1.0165233 - 0.0029221711 Normal1
Sample size: 1000
R (correlation coefficient) = -0.00288482
R-sq = 0.0000083221865
Estimate of error standard deviation: 0.40640613

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	1.0165233	0.034500844	≠ 0	998	29.463723	<0.0001
Slope	-0.0029221711	0.032064158	≠ 0	998	-0.091135127	0.9274

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	0.0013718041	0.0013718041	0.0083056113	0.9274
Error	998	164.83561	0.16516594		
Total	999	164.83698			



Your Intercept may be very significant.
The p-value is very low for the Intercept but not the slope.

LSR Significant Slope

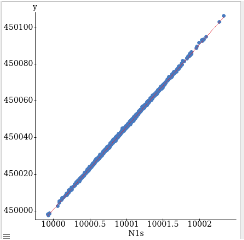
Simple linear regression results:
Dependent Variable: y
Independent Variable: N1s
y = 30.136889 + 44.997088 N1s
Sample size: 1000
R (correlation coefficient) = 0.99974624
R-sq = 0.99949254
Estimate of error standard deviation: 0.40679245

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	30.136889	320.97832	≠ 0	998	0.093890729	0.9252
Slope	44.997088	0.032094627	≠ 0	998	1402.0131	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	325274.41	325274.41	1965640.7	<0.0001
Error	998	165.14914	0.1654801		
Total	999	325439.56			



Your Slope may be very significant.
The p-value is very low for the Slope but not the intercept.
Note: This is a common issue when not recentering data!

LSR Significant Slope and Intercept

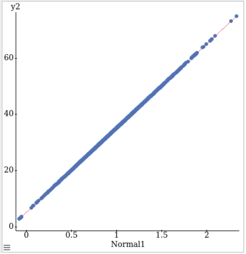
Simple linear regression results:
Dependent Variable: y2
Independent Variable: Normal1
y2 = 5.1016523 + 29.999708 Normal1
Sample size: 1000
R (correlation coefficient) = 0.9999943
R-sq = 0.9999886
Estimate of error standard deviation: 0.040640613

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	5.1016523	0.0034500844	≠ 0	998	1478.7036	<0.0001
Slope	29.999708	0.0032064158	≠ 0	998	9356.1502	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	144582.21	144582.21	87537547	<0.0001
Error	998	1.6483561	0.0016516594		
Total	999	144583.86			



Your Slope and Intercept may be very significant.
The p-value is very low for the Slope and the Intercept.

LSR for noise

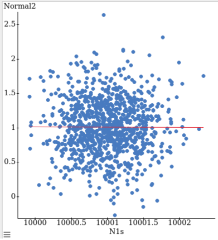
Simple linear regression results:
Dependent Variable: Normal2
Independent Variable: N1s
Normal2 = 30.238216 - 0.0029221693 N1s
Sample size: 1000
R (correlation coefficient) = -0.0028848191
R-sq = 0.0000083221814
Estimate of error standard deviation: 0.40640613

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	30.238216	320.6735	≠ 0	998	0.094295962	0.9249
Slope	-0.0029221693	0.032064148	≠ 0	998	-0.091135098	0.9274

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	0.0013718033	0.0013718033	0.0083056061	0.9274
Error	998	164.83561	0.16516594		
Total	999	164.83698			



Perhaps neither the Intercept nor Slope is significant.

What about R^2

R^2

The square of the correlation, or R^2, tells us how much of the variation in the data can be attributed to the model.

In the above slides, where we had high correlation, the points were mostly along their best-fit line, and R^2 was around 0.999 or very close to 1 or 100%. This means that most of the variation was accounted for by the line model. For the low-correlation slides, R^2 was near zero, meaning that none of the variation was captured by the model.

Example

If the correlation for a LSR model is $r = .5$, then only .25 or 25% of the variation is explained by the model.

However, if the correlation is $r = .9$ then 81% of the variation is explained by the model!

Residual Plots

We often plot residuals against the line $y = 0$ to emphasize the variation which remains after the model (in this case just a line) has been removed.

Original LSR analysis

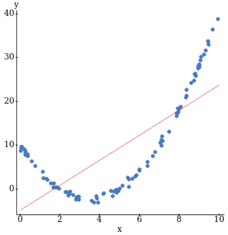
Simple linear regression results:
Dependent Variable: y
Independent Variable: x
y = -4.9034101 + 2.8607158 x
Sample size: 100
R (correlation coefficient) = 0.74844034
R-sq = 0.56016294
Estimate of error standard deviation: 7.8514158

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	-4.9034101	1.5054156	≠ 0	98	-3.2571804	0.0015
Slope	2.8607158	0.25606511	≠ 0	98	11.17183	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	7693.8653	7693.8653	124.80978	<0.0001
Error	98	6041.1836	61.644731		
Total	99	13735.049			



Polynomial Regression Results:
Dependent Variable: y
Independent Variable: x
y = 10.054939 + -7.0536831 X + 1.0060917 X^2

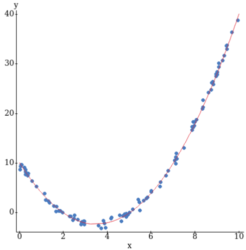
Parameter estimates

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	10.054939	0.1350627	≠ 0	97	74.446456	<0.0001
X	-7.0536831	0.065514217	≠ 0	97	-107.66645	<0.0001
X^2	1.0060917	0.0064418232	≠ 0	97	156.1812	<0.0001

Analysis of variance table for polynomial regression model:

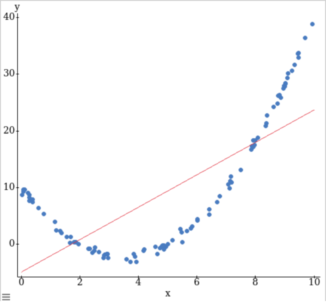
Source	DF	SS	MS	F-stat	P-value
Model	2	13711.121	6855.5603	27790.864	<0.0001
Error	97	23.928344	0.24668396		
Total	99	13735.049			

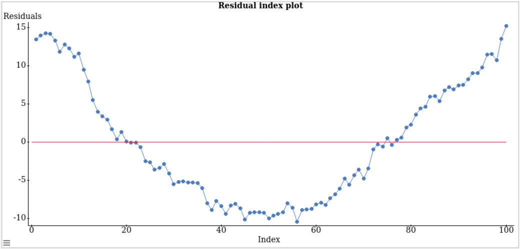
Summary of fit:
Root MSE: 0.49667289
R-squared: 0.9983
R-squared (adjusted): 0.9982



We do not do polynomial interpolation in STAT202.
But, it's just a click away!
In real life, don't hesitate to use it if you want to!

LSR with line model and Residuals Plot





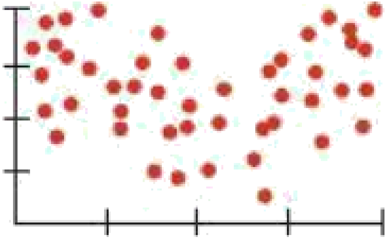
Residuals Plots

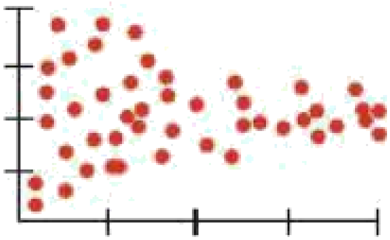
- Will not reveal anything the scatterplot and model won't reveal
- Does often make it easier to see
- Average value must be zero

Typical insights from residual plots

If the residuals cross over the line $y = 0$ a few times in quite distinct places, you probably need a new model.

When your residual plot hugs the line $y = 0$ more closely for part of the data, you can see where your model does a better job.





Small Warning

If you switch your two variables (x and y) for all your data and calculate an LSR line, this is not the same as just flipping the line. The distances (residuals) are computed on the independent variable only, so if you switch the roles of your variables, you'll change the model entierly.

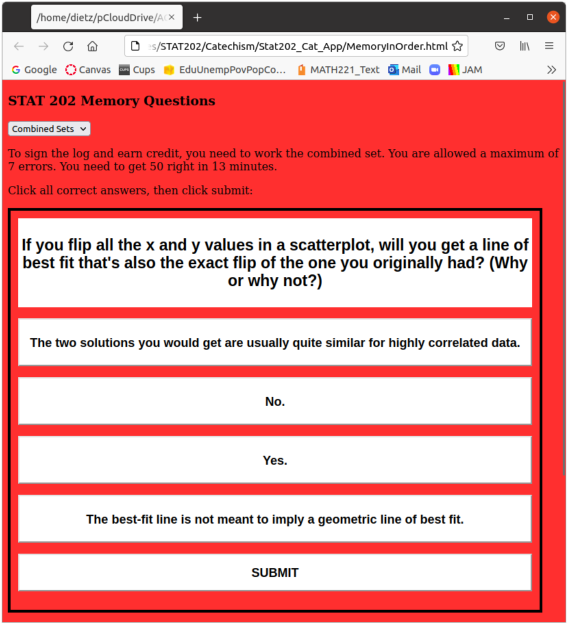
However, if your data do follow a best fit line fairly well, and you flip everything, you will probably end up with two very good LSR fits! They should be in roughly the same place.

Worksheet hint:

Remember that p-values greater than 0.05 are generally “boring”.

MEMORY QUESTIONS

7 today!


/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If you flip all the x and y values in a scatterplot, will you get a line of best fit that's also the exact flip of the one you originally had? (Why or why not?)
The two solutions you would get are usually quite similar for highly correlated data.
No.
Yes.
The best-fit line is not meant to imply a geometric line of best fit.
SUBMIT

/home/dietz/pCloudDrive/A

s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html

Google Canvas Cups EduUnempPovPopCo... MATH221_Text Mail JAM

STAT 202 Memory Questions

Combined Sets

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

If you flip all the x and y values in a scatterplot, will you get a line of best fit that's also the exact flip of the one you originally had? (Why or why not?)

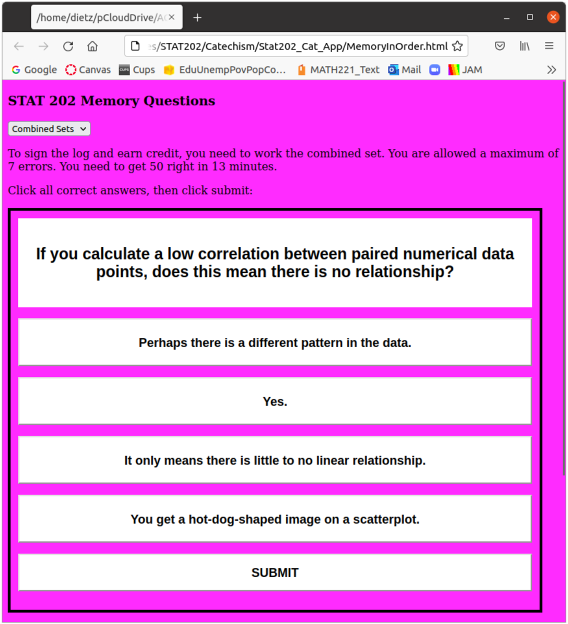
The two solutions you would get are usually quite similar for highly correlated data.

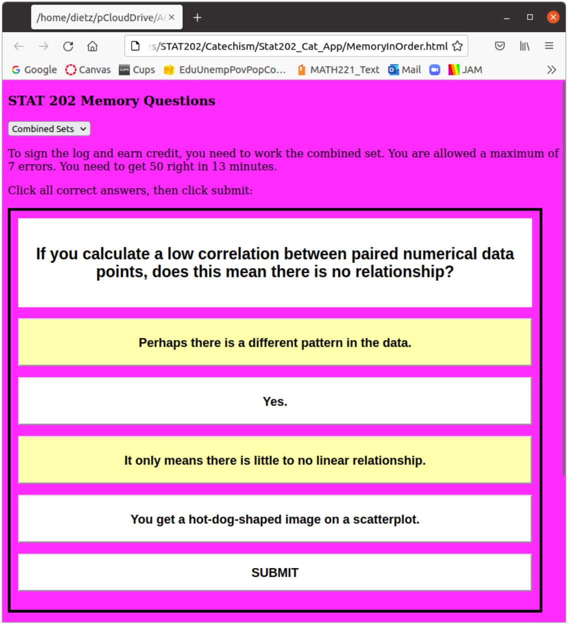
No.

Yes.

The best-fit line is not meant to imply a geometric line of best fit.

SUBMIT


/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If you calculate a low correlation between paired numerical data points, does this mean there is no relationship?
Perhaps there is a different pattern in the data.
Yes.
It only means there is little to no linear relationship.
You get a hot-dog-shaped image on a scatterplot.
SUBMIT


/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google Canvas Cups EduUnempPovPopCo... MATH221_Text Mail JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If you calculate a low correlation between paired numerical data points, does this mean there is no relationship?
Perhaps there is a different pattern in the data.
Yes.
It only means there is little to no linear relationship.
You get a hot-dog-shaped image on a scatterplot.
SUBMIT

/home/dietz/pCloudDrive/A

s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html

Google Canvas Cups EduUnempPovPopCo... MATH221_Text Mail JAM

STAT 202 Memory Questions

Combined Sets

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

If there is no relationship between two paired numerical sets of data, what do you expect the correlation to be?

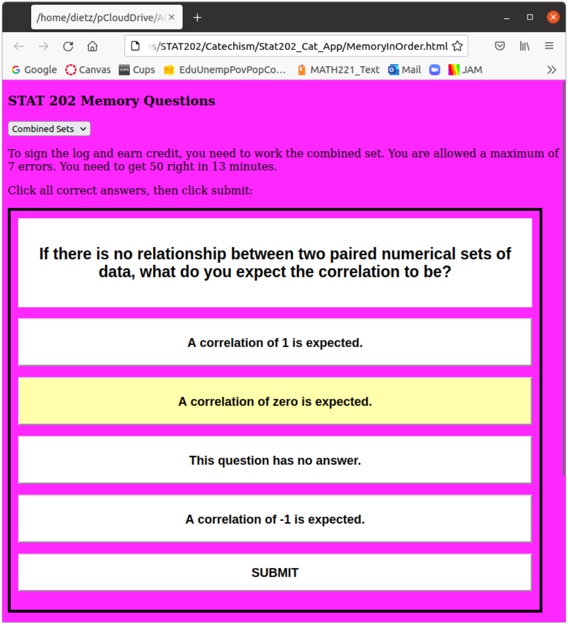
A correlation of 1 is expected.

A correlation of zero is expected.

This question has no answer.

A correlation of -1 is expected.

SUBMIT


/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If there is no relationship between two paired numerical sets of data, what do you expect the correlation to be?
A correlation of 1 is expected.
A correlation of zero is expected.
This question has no answer.
A correlation of -1 is expected.
SUBMIT

STAT 202 Memory Questions

Combined Sets

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

If someone thinks 0.8 is a high correlation, what would that same person say about a correlation of -0.85?

They should say it's lower.

Correlations can't be negative.

They should say it's the same.

They should say it's higher.

SUBMIT

STAT 202 Memory Questions

Combined Sets

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

If someone thinks 0.8 is a high correlation, what would that same person say about a correlation of -0.85?

They should say it's lower.

Correlations can't be negative.

They should say it's the same.

They should say it's higher.

SUBMIT

/home/dietz/pCloudDrive/A

s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html

Google Canvas Cups EduUnempPovPopCo... MATH221_Text Mail JAM

STAT 202 Memory Questions

Combined Sets

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

If I create a random set of x-y data, do I expect the correlation to be zero?

No.

Yes, if it's truly random.

I expect it to be negative.

I expect it to be positive.

SUBMIT

STAT 202 Memory Questions

Combined Sets

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

If I create a random set of x-y data, do I expect the correlation to be zero?

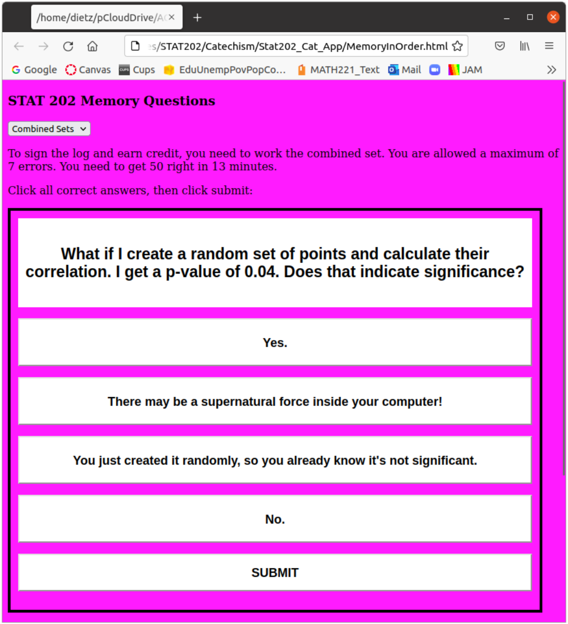
No.

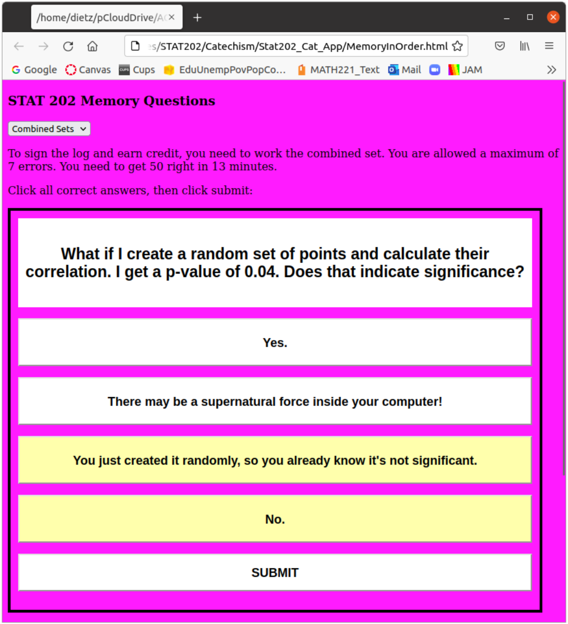
Yes, if it's truly random.

I expect it to be negative.

I expect it to be positive.

SUBMIT


/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
What if I create a random set of points and calculate their correlation. I get a p-value of 0.04. Does that indicate significance?
Yes.
There may be a supernatural force inside your computer!
You just created it randomly, so you already know it's not significant.
No.
SUBMIT


/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
What if I create a random set of points and calculate their correlation. I get a p-value of 0.04. Does that indicate significance?
Yes.
There may be a supernatural force inside your computer!
You just created it randomly, so you already know it's not significant.
No.
SUBMIT

/home/dietz/pCloudDrive/A ×

s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html

Google Canvas Cups EduUnempPovPopCo... MATH221_Text Mail JAM

STAT 202 Memory Questions

Combined Sets

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

Does correlation imply causation?

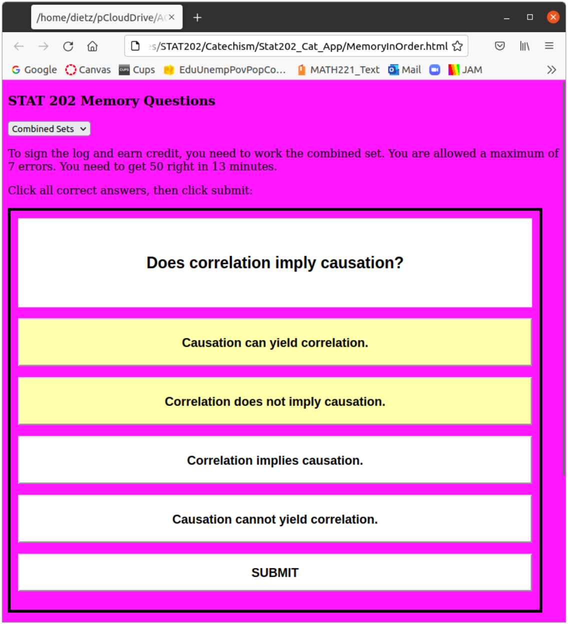
Causation can yield correlation.

Correlation does not imply causation.

Correlation implies causation.

Causation cannot yield correlation.

SUBMIT


/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Does correlation imply causation?
Causation can yield correlation.
Correlation does not imply causation.
Correlation implies causation.
Causation cannot yield correlation.
SUBMIT