How are *two continuous* variables related? Correlation and Simple Linear Regression

Kathleen Torkko

November 11, 2019

Objectives

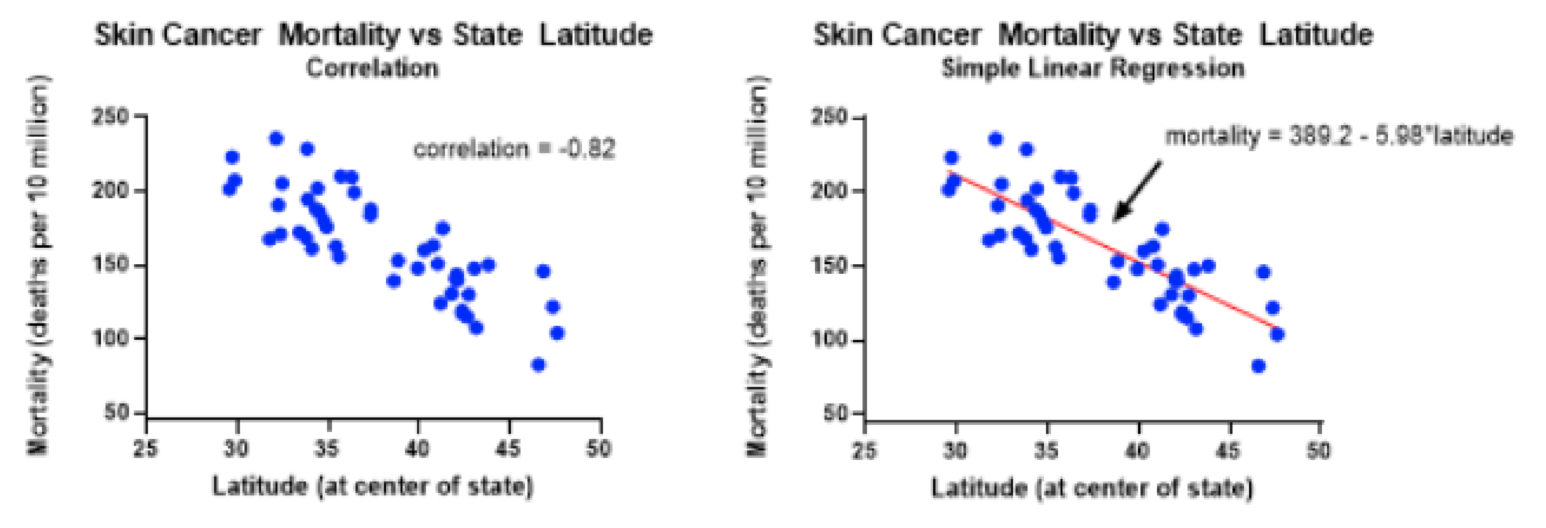
- Learn the difference between correlation and simple linear regression
- Understand when to use Pearson or Spearman correlation
- Learn the difference between causation and correlation
- Understand the basics of simple linear regression
- Learn how to do correlation and SLR in Prism

How are X and Y related?

Correlation and simple linear regression both quantify the direction and strength of the relationship between two continuous variables

Correlation uses a single variable, the correlation coefficient or r, and ranges between -1.0 and 1.0.

Simple linear regression relates X to Y using an equation for a line Y = a + βX.



How are correlation and regression different?

<u>Correlation</u>

Correlation quantifies the amount to which two variables covary

- Correlation does not fit a line through the data

Does not imply causation

Correlation coefficient is measure of effect, *i.e.*, the direction and strength of the association

<u>Simple Linear* Regression (only two variables**)</u>

Uses a linear model to quantify how well the X variable predicts the Y variable

- Fits the "best line" through the data

Implies that one variable, X, influences (causes?) the other, Y, to change

Slope of the line is the measure of effect

* Can be non-linear, too

**>2 variables is multiple linear regression

Correlation

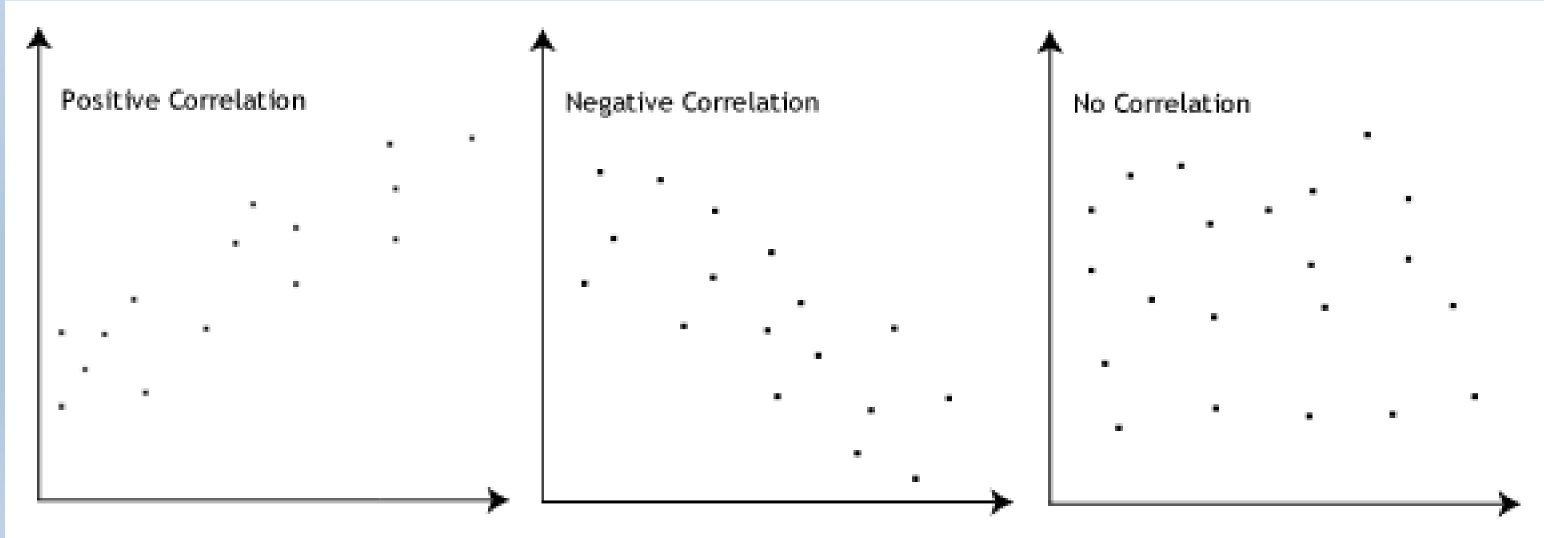
How closely do the two variables covary?

- Pearson correlation – must covary in a linear relationship
- Spearman correlation – must covary in a monotonic relationship

When X increases, what does Y tend to do?

- If Y tends to increase along with X, there's a positive relationship.
- If Y decreases as X increases, that's a negative (inverse) relationship

Does not imply causation (*repeat after me*)



Simple Linear Regression

How well does a linear model explain the relationship of the two variables?

$$Y = a + \beta X$$

a = Y intercept, β = slope

Implies that one variable influences (causes?) the other to change

Does an independent variable (X, AKA predictor variable) cause the dependent variable (Y, AKA outcome variable) to change?

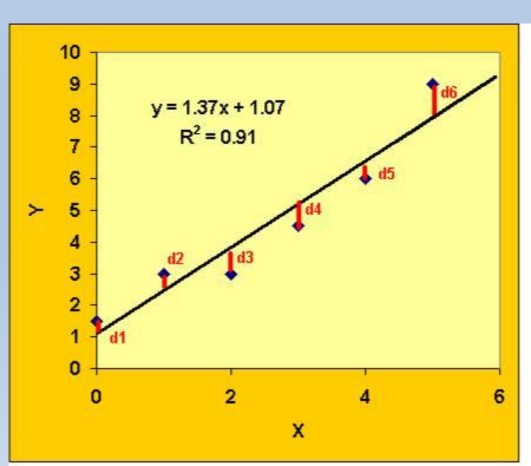
Pearson correlation coefficient, r

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

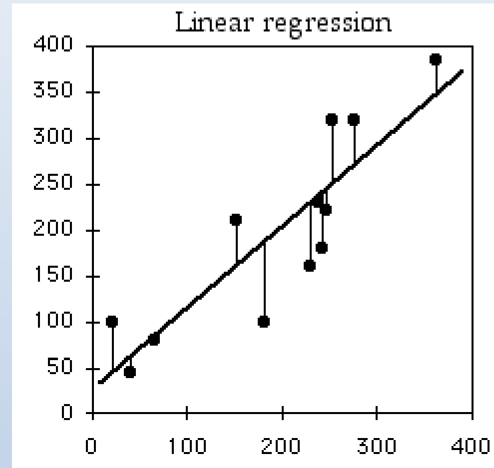
Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

The sample covariance (a measure of the joint variability of two random variables) of the variables divided by the product of their sample standard deviations



Simple linear regression, equation for a line



Statisticians typically use the least squares method to arrive at the equation for the best fit regression line.

The regression line is the line that minimizes the sum of the squared vertical distances between the data points and the line.

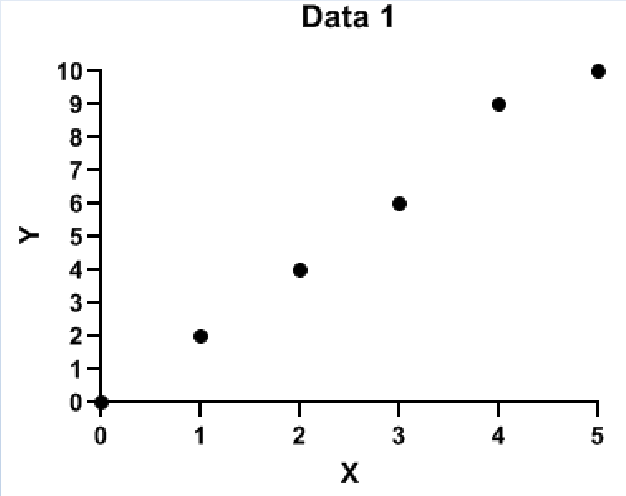
Calculating a Pearson Correlation Coefficient by Hand

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

	X	Y	(X)(Y)	x^2	y^2
	0	0	0	0	0
	1	2	2	1	4
	2	4	8	4	16
	3	6	18	9	36
	4	9	36	16	81
	5	10	50	25	100
SUM	15	31	114	55	237
SUM square	225	961			
SUMxSUM	465	=15x31			
N	6				
sum of xy	114				
(sum of x)(sum of y)	465				
sum of x^2	55				
sum of y^2	237				
(sum of x)^2	225				
(sum of y)^2	961				
numerator	219				
denominator	220.01	105	461	48405	
r	0.995403				



	Correlation	A X vs. Y
1	Pearson r	
2	r	0.9954

Correlation Coefficient r (rho, ρ)

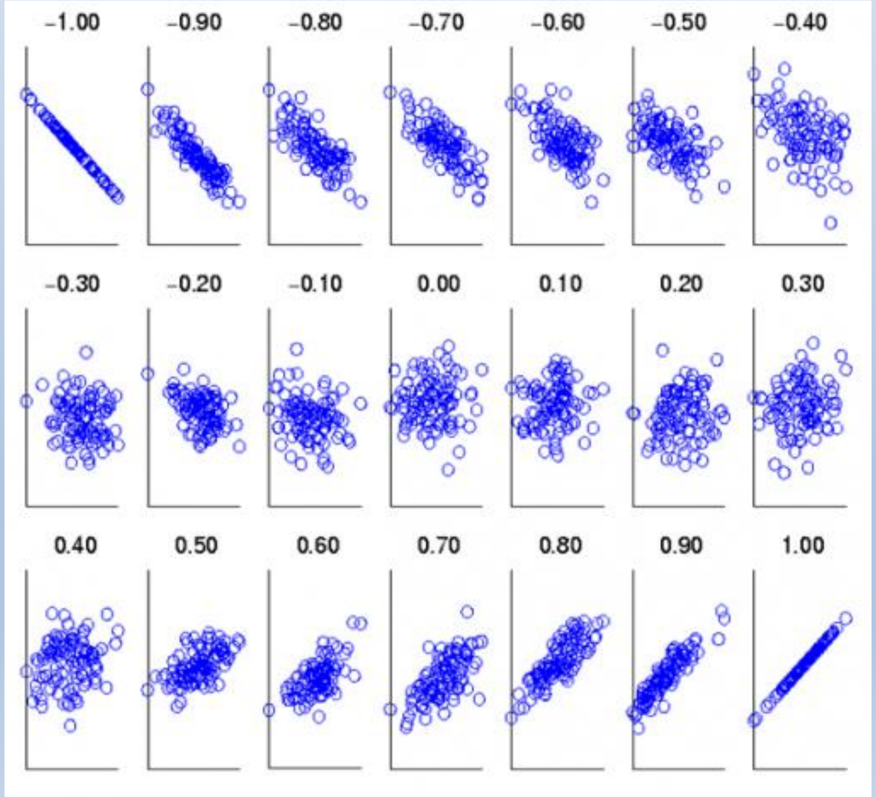
rho measures strength of a linear relationship between 2 continuous variables

$-1 \leq r \leq 1$

r=0 mean no linear relationship

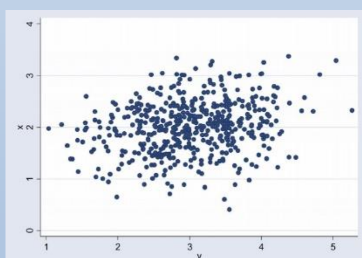
no distinction between x and y variables

does not represent the slope of the line of best fit.

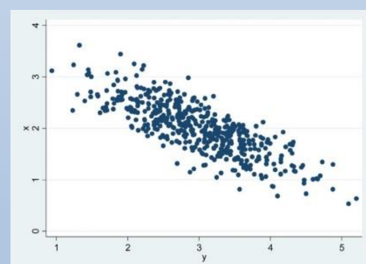


Rule of Thumb for Interpreting a Correlation Coefficient

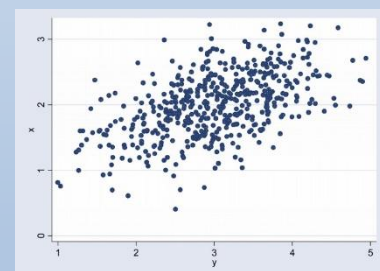
Size of Correlation, r	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very strong correlation
.70 to .90 (−.70 to −.90)	Strong correlation
.50 to .70 (−.50 to −.70)	Moderate correlation
.30 to .50 (−.30 to −.50)	Weak correlation
.00 to .30 (.00 to −.30)	Negligible correlation



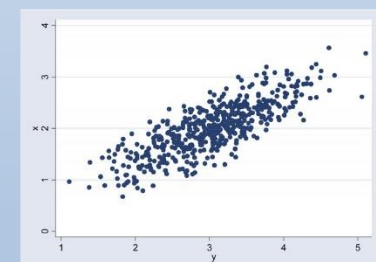
r=0.20



r= - 0.80



r=0.50



r=0.80

Malawi Medical Journal; 24(3): 69-71 September 2012

Two main types of correlation coefficients
Pearson's and Spearman's

Pearson correlation coefficient (parametric)
(AKA Pearson product-moment correlation coefficient)

a measure of the strength and direction of the *linear* relationship between two continuous variables

H_o: $\rho = 0$ (no correlation, random scatter)
H_A: $\rho \neq 0$

Tests the null hypothesis that the population correlation $\rho = 0$
NOT that there is a strong relationship

Assumptions: Pearson correlation coefficient

The two variables must be continuous

The two variables must be approximately normally distributed

No outliers

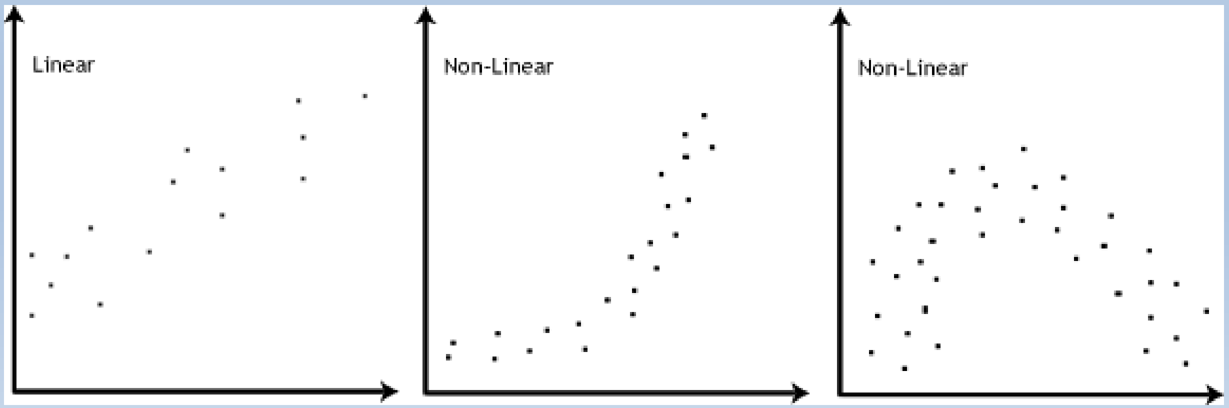
Homoscedasticity along the range of X values

There is a linear relationship between the two variables

Every data point must be in pairs

- Every independent X variable observation must have a corresponding observation for the dependent Y variable

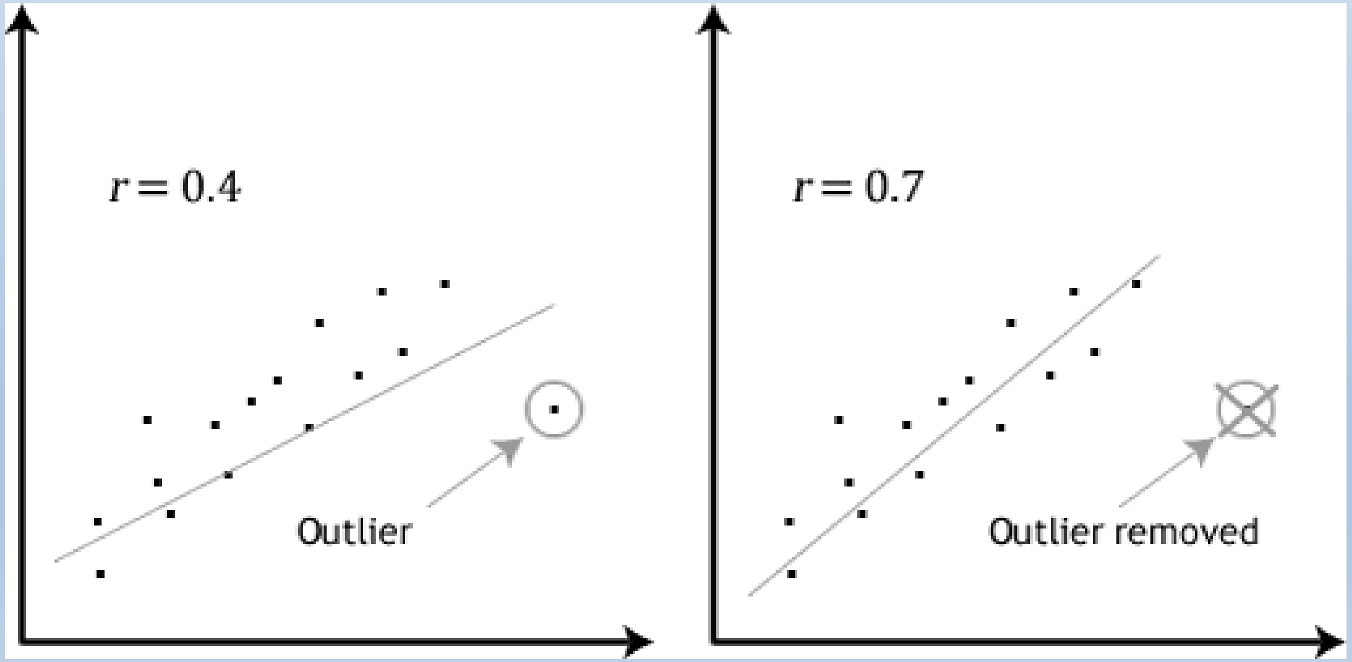
Always graph data using a scatter plot



Pearson correlation coefficient: The effect of outliers

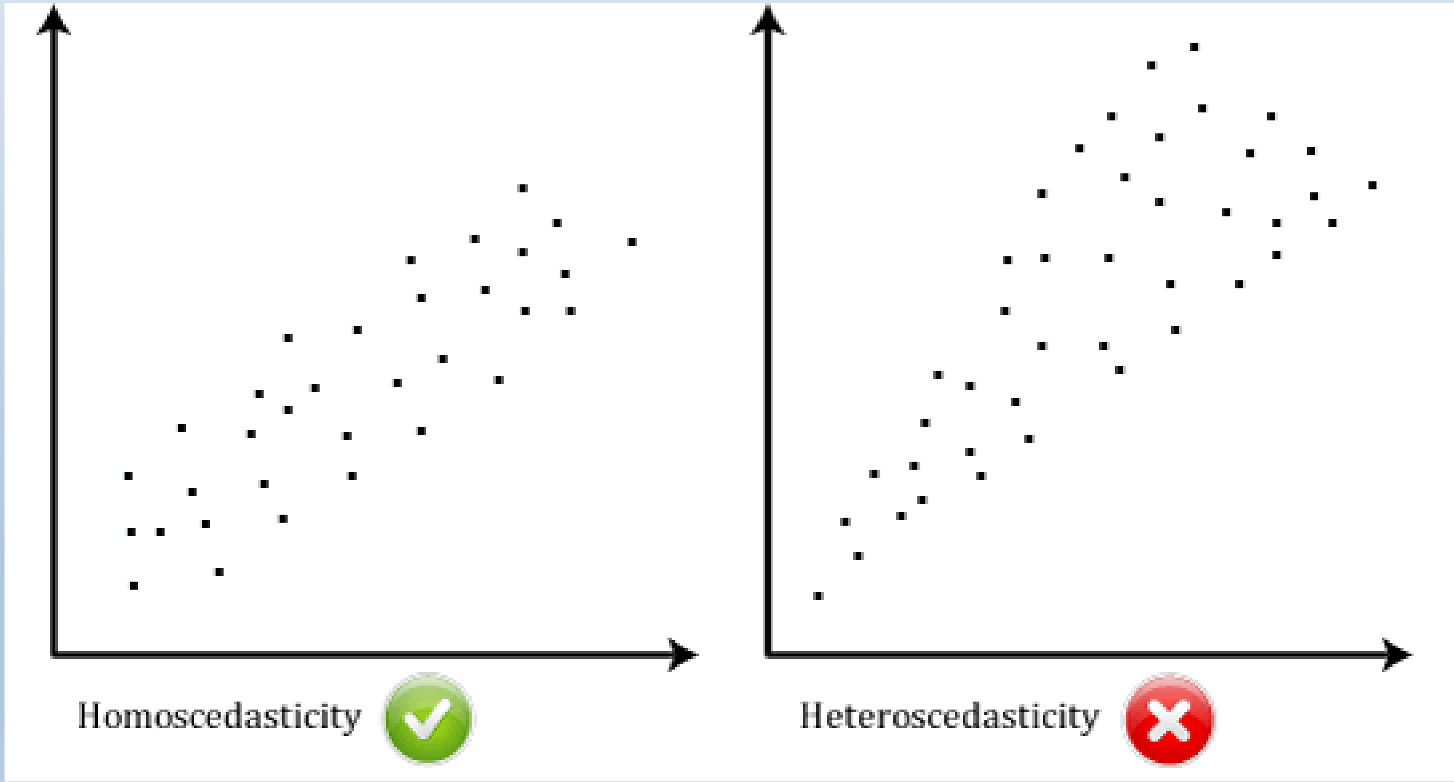
Outliers can have a very large effect on the line of best fit and the Pearson correlation coefficient, which can lead to very different conclusions from the data

(ignore the lines in the graphs…)



Homoscedasticity

The variances along the range of x values remain similar



Two main types of correlation coefficients
Pearson's and Spearman's

Spearman's coefficient (non-parametric)

Rank correlation, uses relationships between ranks of the variables

Spearman's correlation determines the strength and direction of the relationship between the ranks of the two continuous* variables as long as it is monotonic

$$r_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

H_o: $\rho_{ranks} = 0$ (no correlation)

H_A: $\rho_{ranks} \neq 0$

n=number of pairs

D=the difference of pairs of ranking

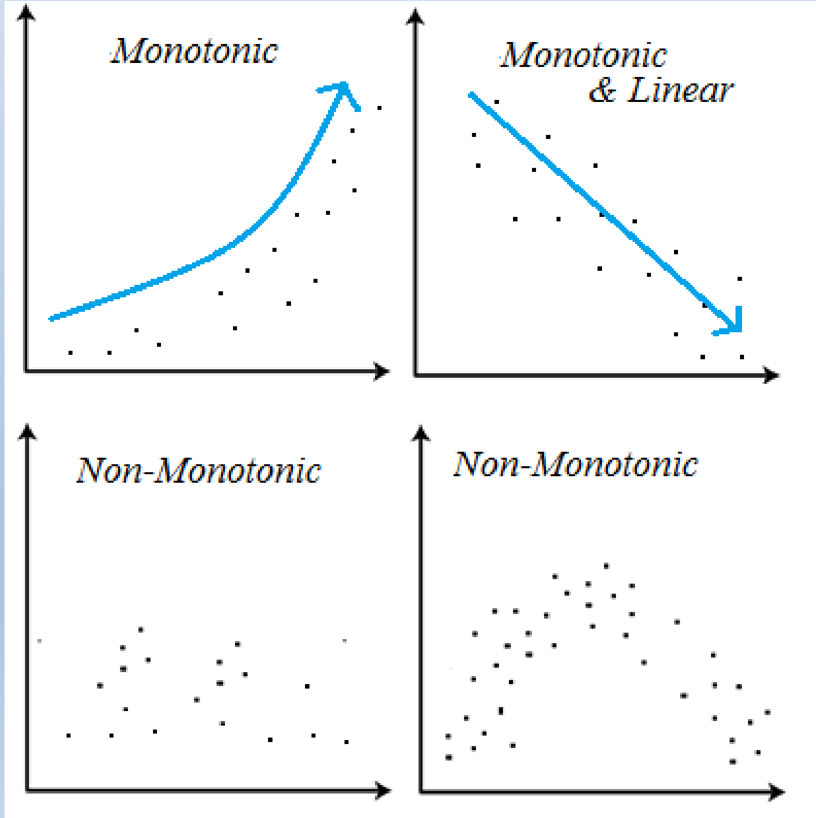
* The variables can also be ordinal

Spearman Correlation Coefficient Assumptions

The two variables should be continuous or ordinal

The variables have a monotonic relationship

Monotonic = not necessarily linear but increasing (or decreasing) together (and not increasing **and** decreasing)

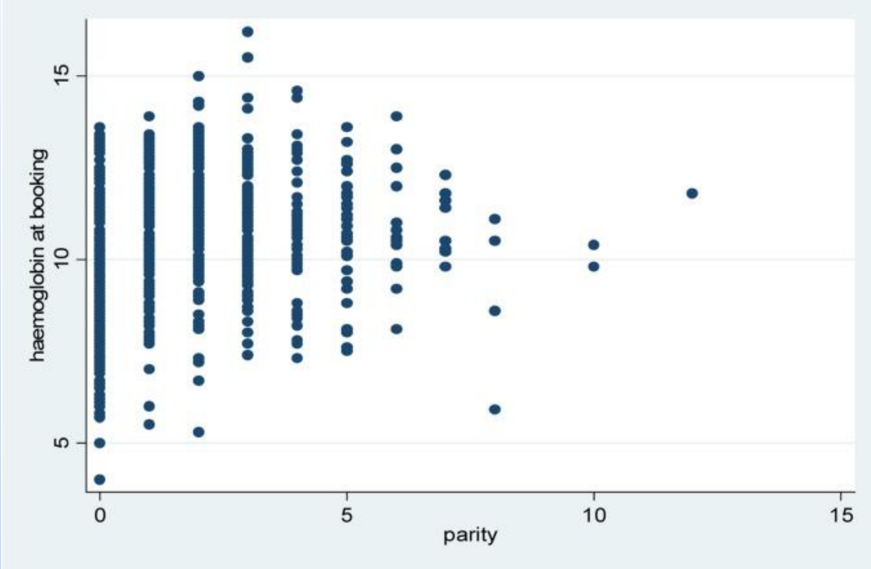


Rho interpreted similarly to the Pearson rho

If the scatterplot shows a linear relationship with no outliers, there is only a small difference numerically between the Pearson and Spearman correlation coefficients

Effect of outliers

Charles Spearman



Spearman's and Pearson's Correlation coefficients

Statistic	Extreme values included		7 Extreme values removed	
	n	r	n	r
Spearman's	783	0.3	776	0.3
Pearson's	783	0.2	776	0.3

If there are outliers, use Spearman

Malawi Med J. 2012 Sep; 24(3): 69–71.

Let's look at correlation between x and y for four different datasets

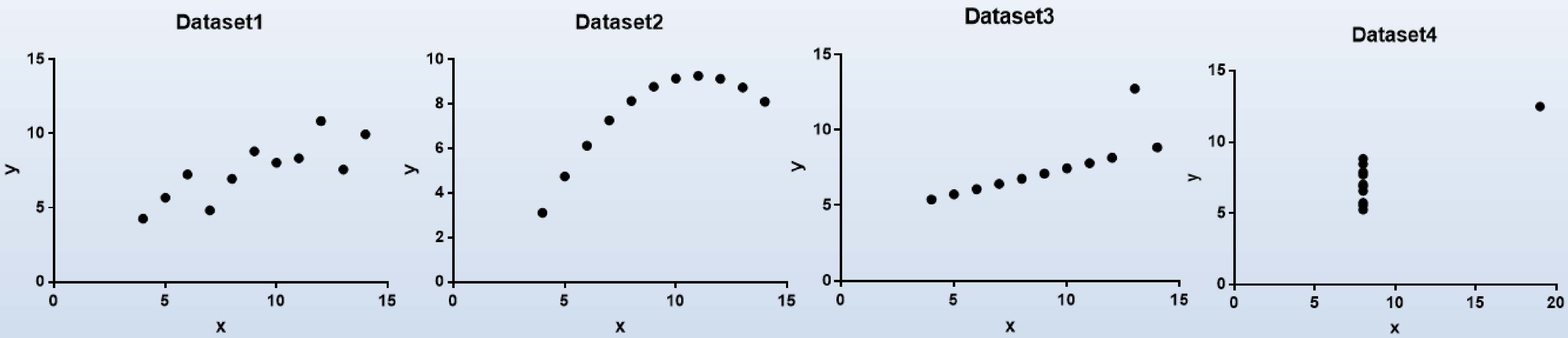
1		2		3		4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Dataset	Pearson correlation coefficient	p-value
1	0.816	0.002
2	0.816	0.002
3	0.816	0.002
4	0.816	0.002

1		2		3		4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Dataset	Pearson correlation coefficient	p	Spearman correlation coefficient	p
1	0.816	0.002	0.818	0.003
2	0.816	0.002	0.691	0.023
3	0.816	0.002	0.991	<0.0001
4	0.816	0.002	0.500	0.182

Anscombe’s quartet



Dataset	Pearson correlation coefficient	p	Spearman correlation coefficient	p
1	0.816	0.002	0.818	0.003
2	0.816	0.002	0.691	0.023
3	0.816	0.002	0.991	<0.0001
4	0.816	0.002	0.500	0.182

Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. **27** (1): 17–21.

Correlation Example

Special Article

NEJM 1999, 340:1881-7

THE RELATION BETWEEN FUNDING BY THE NATIONAL INSTITUTES OF HEALTH AND THE BURDEN OF DISEASE

CARY P. GROSS, M.D., GERARD F. ANDERSON, PH.D., AND NEIL R. POWE, M.D., M.P.H., M.B.A.

<u>Measure of Burden</u> (among several)

DALY (disability-adjusted life-years)
one DALY= loss of one year of healthy life to disease

Used 1990 DALY data

CONDITION OR DISEASE	NIH RESEARCH FUNDS
	thousands of dollars (% of total) 1996
AIDS	1,410,925 (28.7)
Breast cancer	381,880 (7.8)
Dementia	304,411 (6.2)
Diabetes mellitus	298,920 (6.1)
Ischemic heart disease	269,100 (5.5)
Alcohol abuse	256,600 (5.2)
Injuries	198,700 (4.0)
Dental and oral disorders	187,100 (3.8)
Cirrhosis	169,800 (3.4)
Depression	143,800 (2.9)
Lung cancer	127,796 (2.6)
Stroke	120,280 (2.4)
Schizophrenia	111,479 (2.3)
Colorectal cancer	105,525 (2.1)
Sexually transmitted diseases	102,583 (2.1)
Prostate cancer	92,661 (1.9)
Multiple sclerosis	82,800 (1.7)
Asthma	81,600 (1.7)
Parkinson's disease	77,158 (1.6)
Tuberculosis	64,125 (1.3)
Chronic obstructive pulmonary disease	62,400 (1.3)
Pneumonia	61,900 (1.3)
Cervical cancer	60,180 (1.2)
Epilepsy	55,100 (1.1)
Ovarian cancer	42,168 (0.8)
Perinatal conditions	26,400 (0.5)
Uterine cancer	13,956 (0.3)
Otitis media	9,100 (0.2)
Peptic ulcer	6,000 (0.1)

Example: Distribution of NIH Funds and DALY

Research Question

Are DALYs associated with NIH funding levels?

Two ways to look at this question

1. Are DALYs and NIH funding level correlated?
   - Pearson or Spearman correlation coefficients
   - Do DALYs and funding increase or decrease together, or does one increase as the other decreases
   - No causation is inferred

2. Do DALYs “predict” levels of NIH funding
   - Linear regression
   - Higher levels of DALY “cause” NIH funding to increase or decrease
   - Implies a causal relationship – but doesn’t prove one

Correlation Example: Distribution of NIH Funds and DALY

You decide to test correlation

You are in a hurry and you don't want to take the time to graph the data or test the assumptions. You just want a quick answer

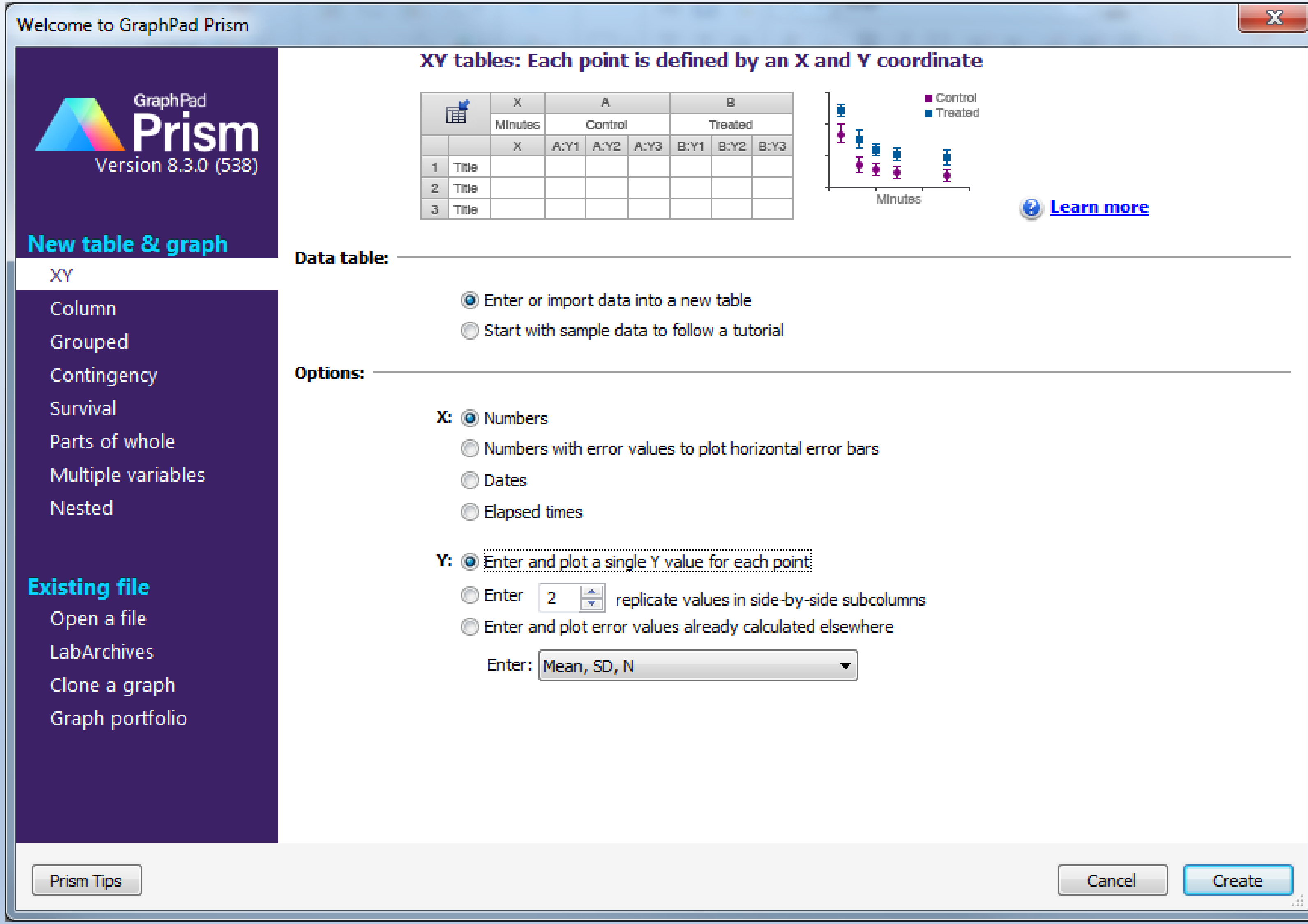
You put the data into Prism, accepting all the defaults (Pearson correlation), and you get

$r=0.12$, $p=0.54$

You conclude there is no correlation (you fail to reject the null hypothesis of no correlation": $\rho = 0$)

But are you right?

To Start a Correlation Analysis



Correlation: Prism Data Structure

Table format: XY		X	Group A
		DALY	NIHFunds
		X	Y
1	Title	8	9.10
2	Title	118	64.13
3	Title	185	13.96
4	Title	192	60.18
5	Title	236	82.80
6	Title	239	6.00
7	Title	375	42.17
8	Title	404	102.58
9	Title	447	77.16
10	Title	505	55.10
11	Title	574	92.66
12	Title	870	187.10
13	Title	1236	81.60
14	Title	1263	61.90
15	Title	1267	1410.93
16	Title	1421	381.88
17	Title	1584	169.80
18	Title	1626	105.53
19	Title	1767	26.40
20	Title	2249	111.48
21	Title	2284	62.40
22	Title	2357	298.92
23	Title	2866	304.41
24	Title	2987	127.80
25	Title	4690	256.60
26	Title	4977	120.28
27	Title	8393	143.80
28	Title	8608	198.70
29	Title	8876	269.10

Or if you use "Column" data/graph structure

	Group A	Group B
	DALY	NIHFunds
1	8	9.10
2	118	64.13
3	185	13.96
4	192	60.18
5	236	82.80
6	239	6.00
7	375	42.17
8	404	102.58
9	447	77.16
10	505	55.10
11	574	92.66
12	870	187.10
13	1236	81.60
14	1263	61.90
15	1267	1410.93
16	1421	381.88
17	1584	169.80
18	1626	105.53
19	1767	26.40
20	2249	111.48
21	2284	62.40
22	2357	298.92
23	2866	304.41
24	2987	127.80
25	4690	256.60
26	4977	120.28
27	8393	143.80
28	8608	198.70
29	8876	269.10

Assumptions Pearson Correlation

The variables must be continuous - yes

The variables must be approximately normally distributed

No outliers

Homoscedasticity along the range of X values

There is a linear relationship between the two variables

Every data point must be in pairs - yes

	Group A	Group B
	DALY	NIHFunds
1	8	9.10
2	118	64.13
3	185	13.96
4	192	60.18
5	236	82.80
6	239	6.00
7	375	42.17
8	404	102.58
9	447	77.16
10	505	55.10
11	574	92.66
12	870	187.10
13	1236	81.60
14	1263	61.90
15	1267	1410.93
16	1421	381.88
17	1584	169.80
18	1626	105.53
19	1767	26.40
20	2249	111.48
21	2284	62.40
22	2357	298.92
23	2866	304.41
24	2987	127.80
25	4690	256.60
26	4977	120.28
27	8393	143.80
28	8608	198.70
29	8876	269.10

The variables must be continuous - Yes

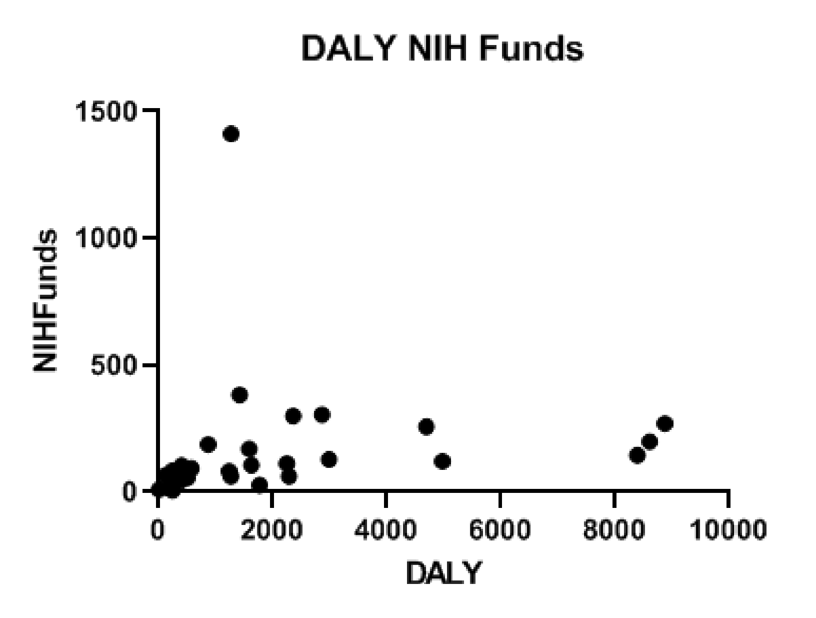
The variables must be approximately normally distributed

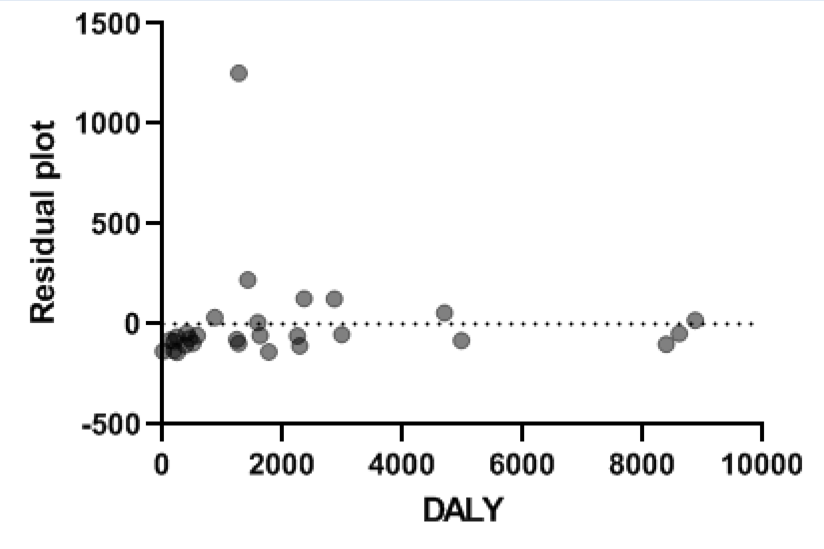
No outliers – there is an outlier

Homoscedasticity along the range of x values - No

There is a linear relationship between the two variables – Yes?

Every data point must be in pairs - Yes





The variables must be continuous - Yes

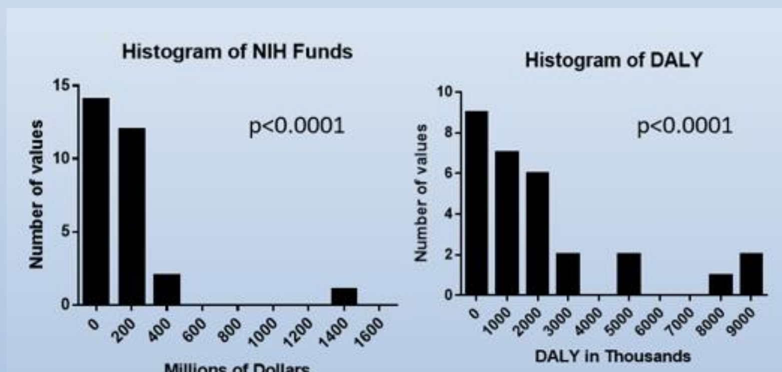
The variables must be approximately normally distributed - No

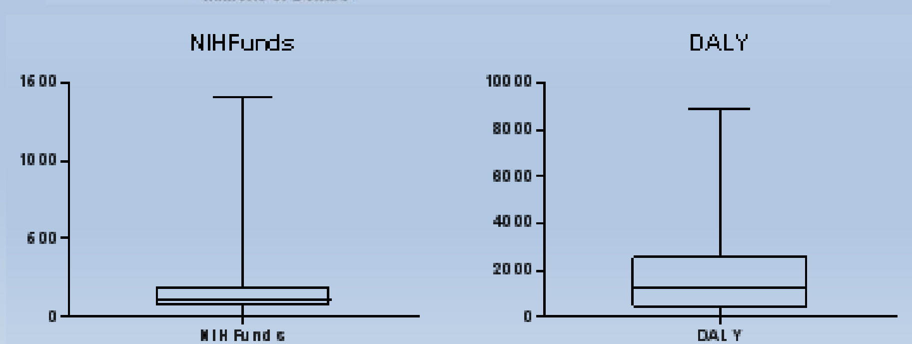
No outliers - there is an outlier

Homoscedasticity along the range of x values - No

There is a linear relationship between the two variables – Yes?

Every data point must be in pairs - Yes





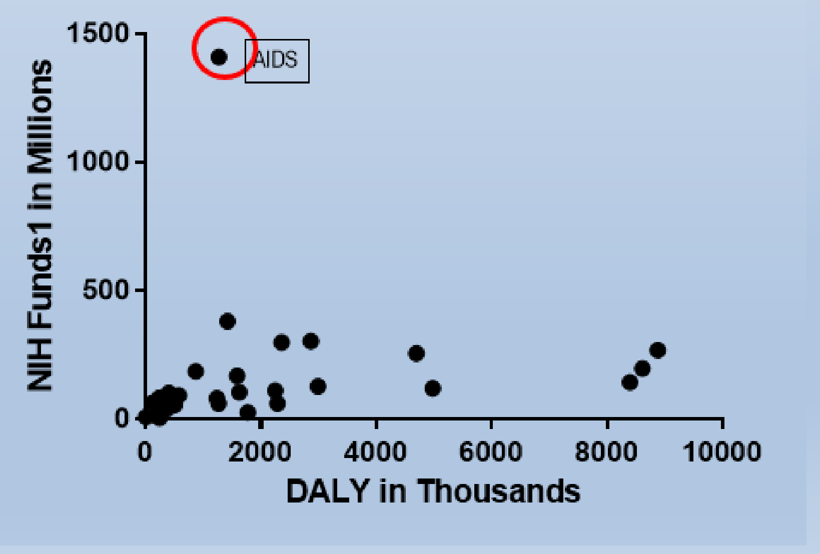
		A	B
		DALY	NIHFunds
1	Number of values	29	29
2			
3	Minimum	8.000	6.000
4	25% Percentile	389.5	61.04
5	Median	1267	102.6
6	75% Percentile	2612	192.9
7	Maximum	8876	1411
8	Range	8868	1405
9			
10	Mean	2159	169.8
11	Std. Deviation	2570	257.7
12	Std. Error of Mean	477.3	47.85
13			
14	Skewness	1.752	4.261

Correlation Example Distribution of NIH Funds and DALY Decision Time

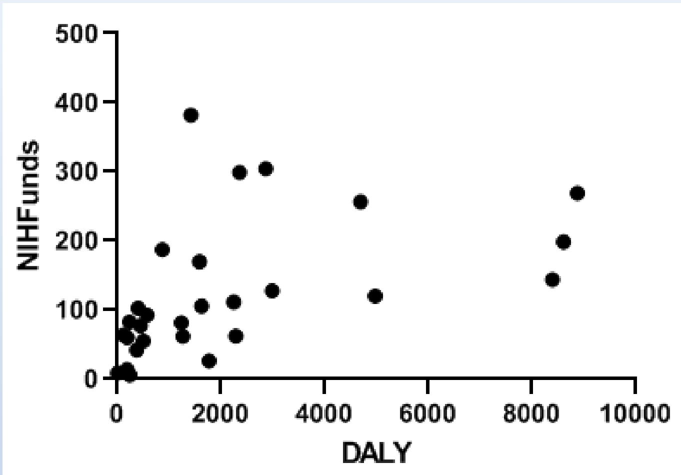
The data do not meet the assumptions of a normal distribution and homoscedasticity, and there is one obvious outlier

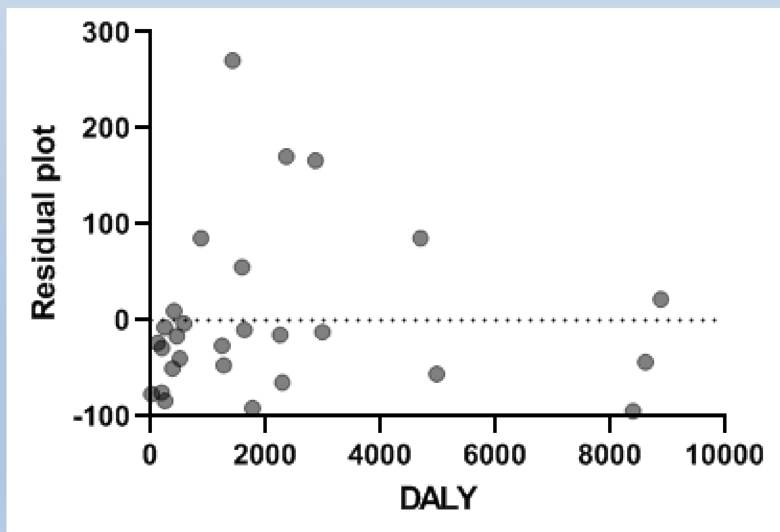
1. Do a Spearman rather than a Pearson correlation
   Some say not to do this because of reduced power

2. Transform the data or eliminate the outlier
   Is outlier real data? Yes, but…

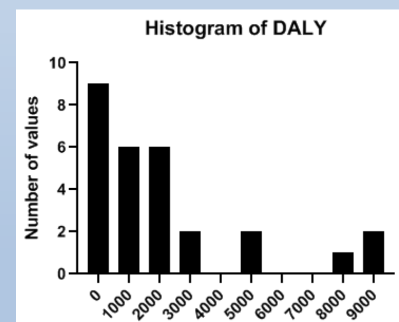


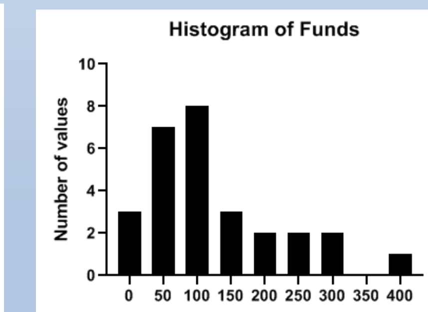
Without outlier





		A	B
		DALY	NIHFunds
1	Number of values	28	28
2			
3	Minimum	8.000	6.000
4	25% Percentile	382.3	60.61
5	Median	1342	97.62
6	75% Percentile	2739	182.8
7	Maximum	8876	381.9
8	Range	8868	375.9
9			
10	Mean	2191	125.5
11	Std. Deviation	2612	98.78
12	Std. Error of Mean	493.5	18.67
13			
14	Skewness	1.699	1.068





The variables must be continuous - Yes

The variables must be approximately normally distributed - No

No outliers – Yes – “outlier” eliminated

Homoscedasticity along the range of x values - No

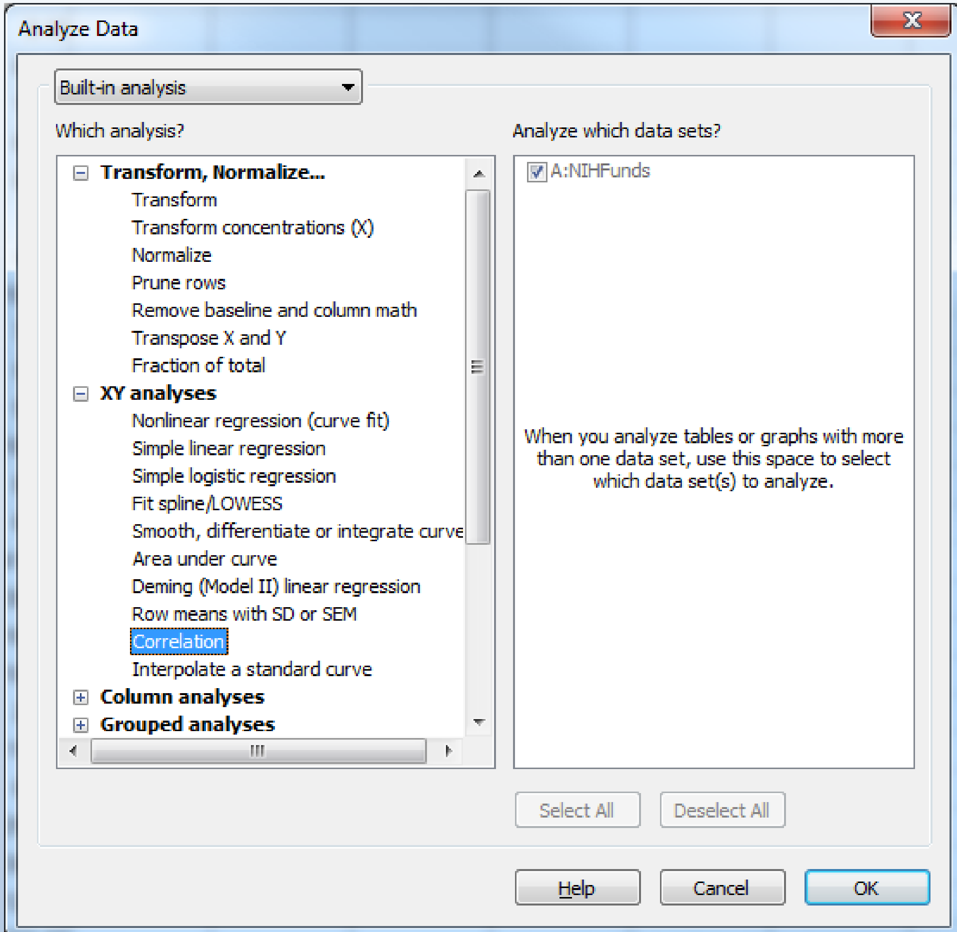
There is a linear relationship between the two variables – Yes?

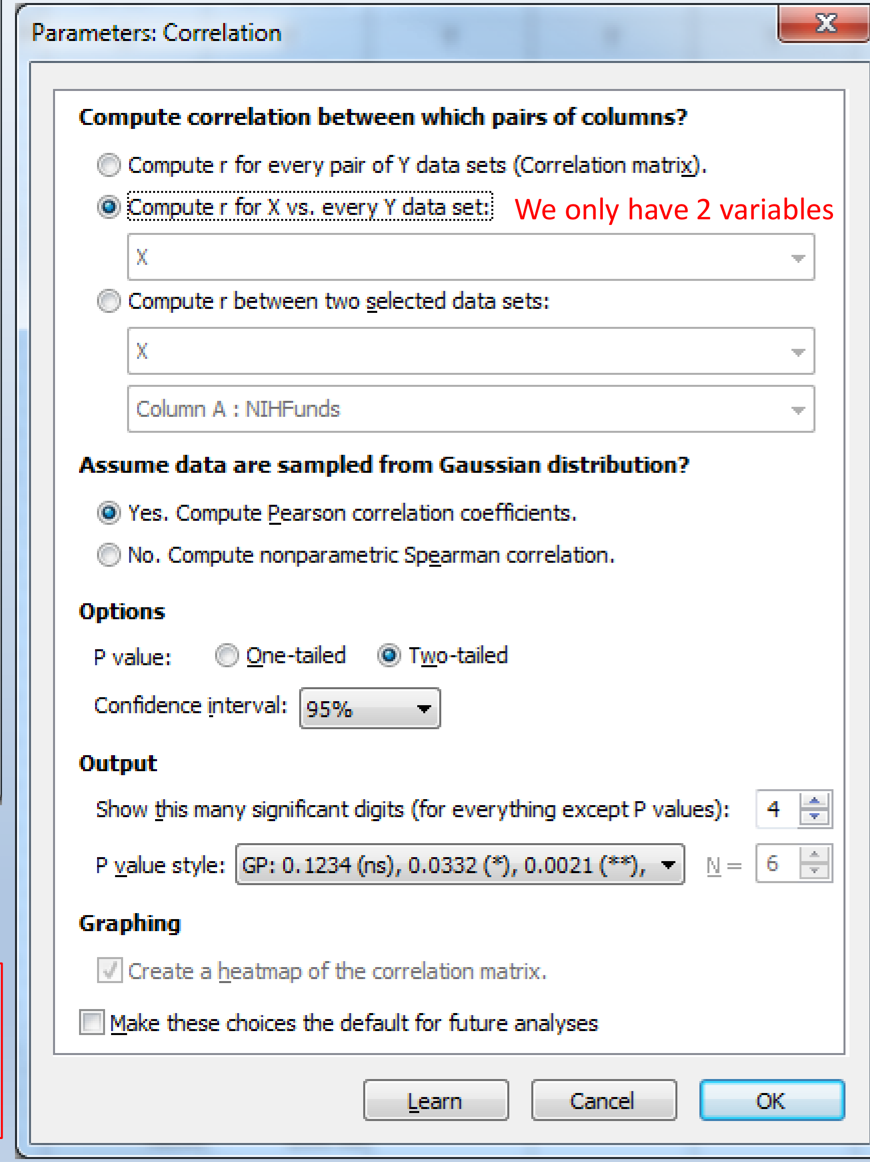
Every data point must be in pairs - Yes

What to do?
I would either transform data or use Spearman Correlation

For the purposes of this class, we will do Pearson and Spearman on datasets with and without outlier and try data transformation.

Pearson Correlation





We only have 2 variables

Note for later: to do a Spearman test, check the option “no” under “Assume data are sampled from Gaussian distribution.”

	Correlation	A DALY vs. NIHFunds
1	**Pearson r**	
2	r	0.1189
3	95% confidence interval	-0.2589 to 0.4651
4	R squared	0.01413
5		
6	**P value**	
7	P (two-tailed)	0.5391
8	P value summary	ns
9	Significant? (alpha = 0.05)	No
10		
11	**Number of XY Pairs**	29

Results on data with outlier:

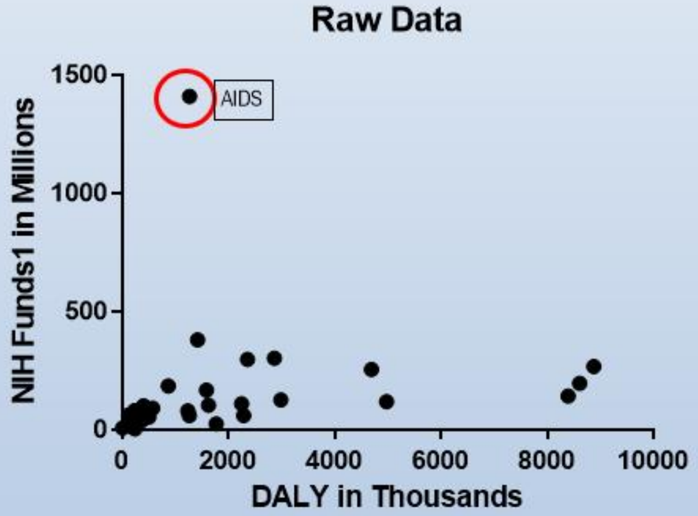
Pearson r=0.12
P=0.54

You fail to reject the null hypothesis of no linear relationship and conclude there is no correlation

Report the *r* value and the corresponding *p*-value. e.g., DALY was not correlated with NIH funding (Pearson *r* = 0.12, p=0.54).

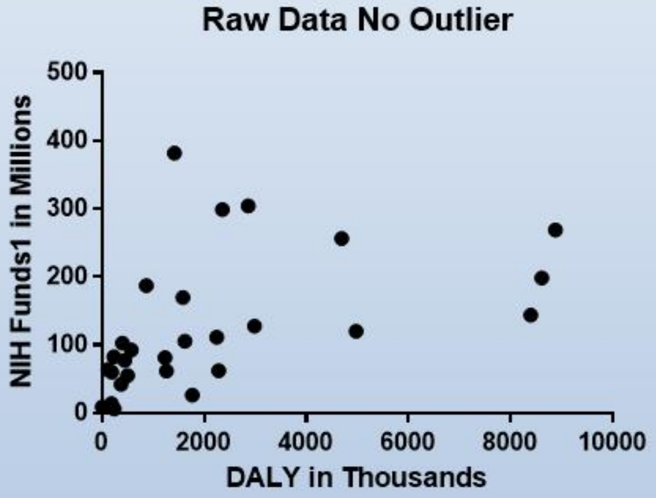
Plot of NIH Funding and DALY, Raw Data, With and Without AIDS Outlier

With Outlier



Pearson r=0.12, p=0.54
Spearman r=0.67, p=0.0001

Without Outlier

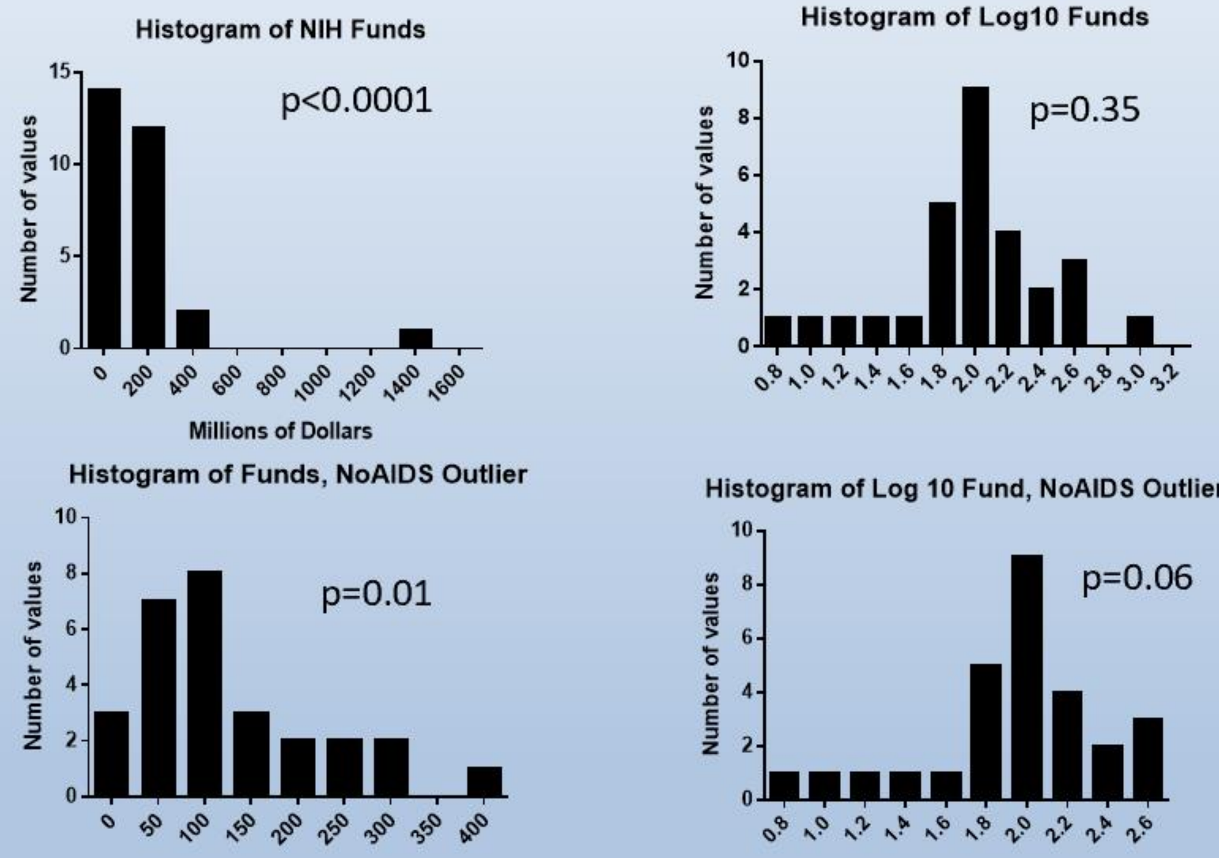


Pearson r=0.48, p=0.01
Spearman r=0.71, p<0.0001

Distribution of NIH Funding Raw and Transformed Data, With and Without AIDS Outlier

With outlier
n=29

Minus outlier
n=28

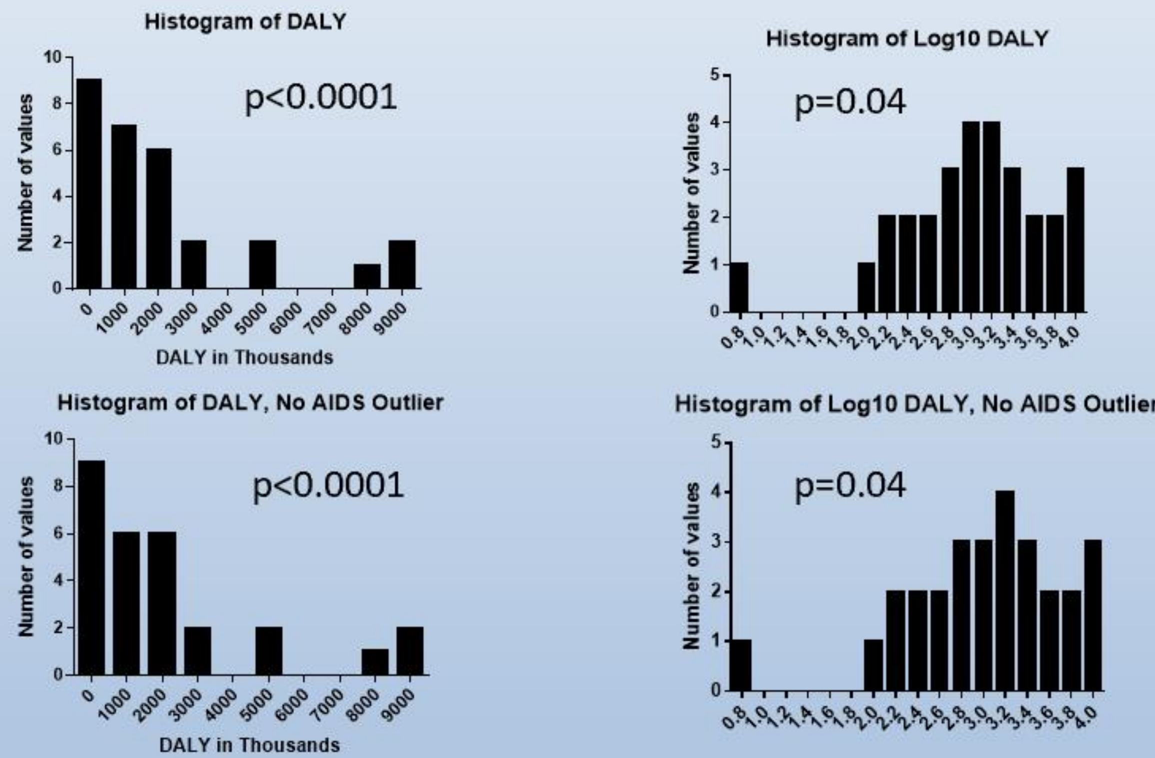


p-values are for the Shapiro-Wilk test for normal distribution. If p>0.05, reject the null hypothesis that data are not normally distributed.

Distribution of DALY Raw and Transformed Data, With and Without AIDS Outlier

With outlier n=29

Minus outlier n=28



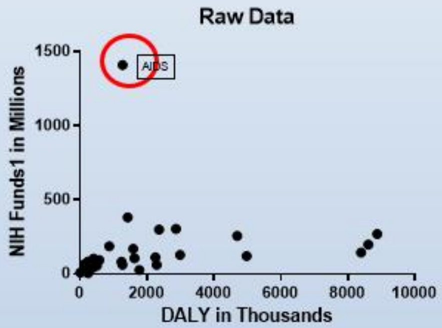
p-values are for the Shapiro-Wilk test for normal distribution.

Plots of NIH Funding and DALY, Raw and Transformed Data, With and Without AIDS Outlier

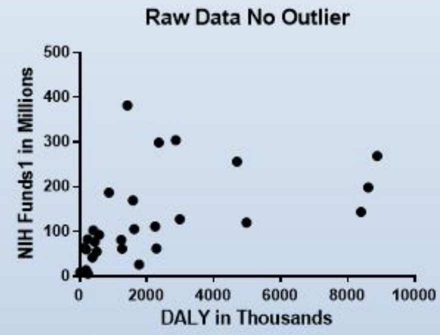
With Outlier

Without Outlier

Raw Data

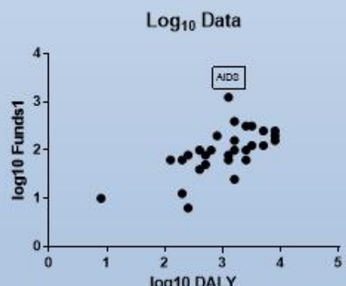


Pearson r=0.12, p=0.54
Spearman r=0.67, p=0.0001

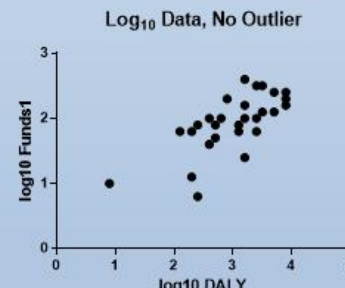


Pearson r=0.48, p=0.01
Spearman r=0.71, p<0.0001

Log_{10} Data



Pearson r=0.64, p=0.0002
Spearman r=0.67, p<0.0001



Pearson r=0.70, p<0.0001
Spearman r=0.72, p<0.0001

Lessons Learned

Always graph your data

Look for linear or monotonic relationship and outliers

Make sure you meet test assumptions

Otherwise you could come to the wrong conclusions!

Cautions

A correlation between two variables doesn't mean causality

r is highly influenced by sample size

e.g., sample size of 150 could find $r = 0.16$ and $p \leq 0.05$

Correlation vs. Possible Relationships Between Variables

Direct cause and effect

X causes Y, i.e., overexpression of protein X causes tumors to grow

Both cause and effect

Coffee consumption causes nervousness **and** nervous people drink more coffee

Relationship caused by a third variable

Drinking alcohol and lung cancer. Both are related to cigarette smoking

Smoking is a *confounder* in the relationship of alcohol to lung cancer

Coincidental relationship

Correlation occurs at random

Causation or just random correlation?
Does chocolate make you clever or crazy?

- A paper in the New England Journal of Medicine claimed a relationship between chocolate and Nobel Prize winners

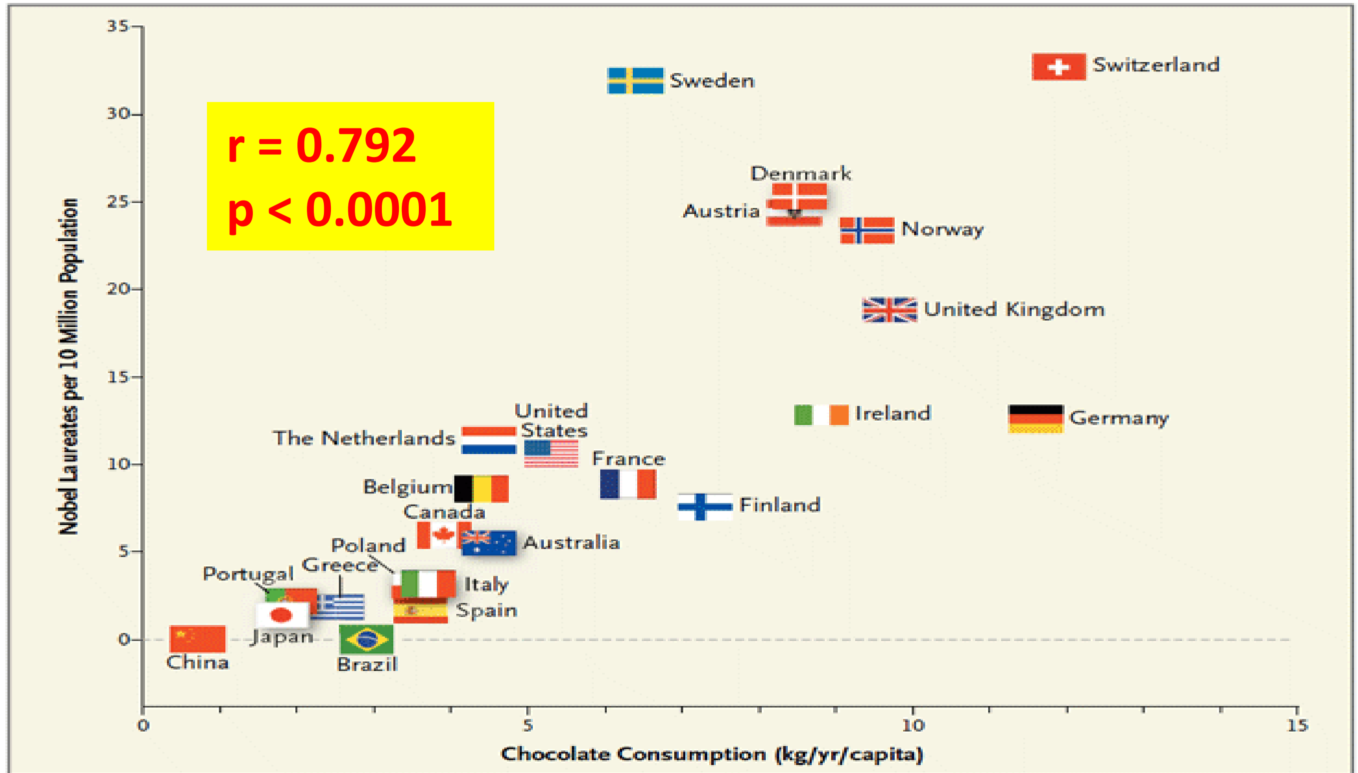
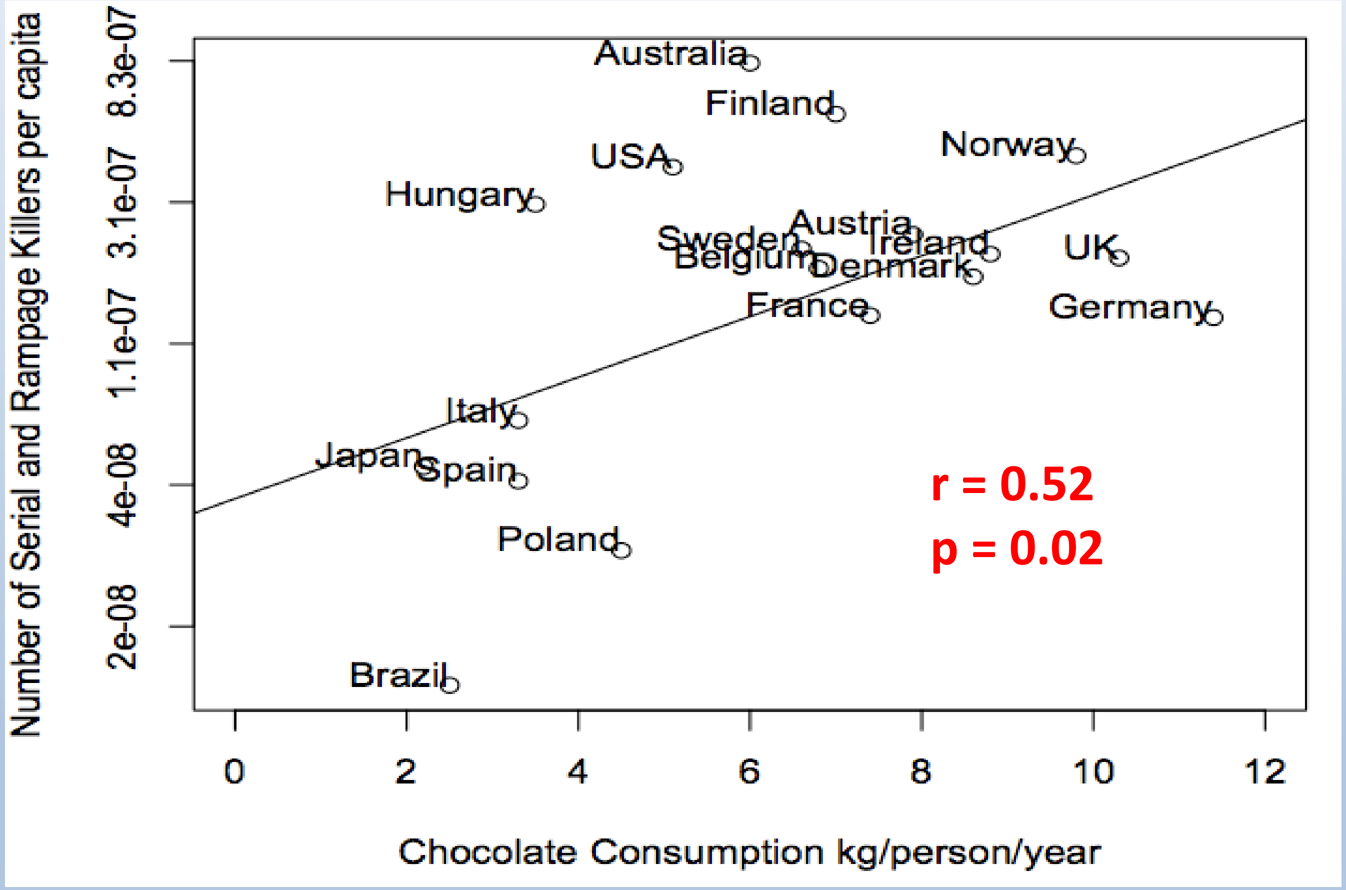


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Messerli FH N Engl J Med. 2012 Oct 18;367(16):1562-4.

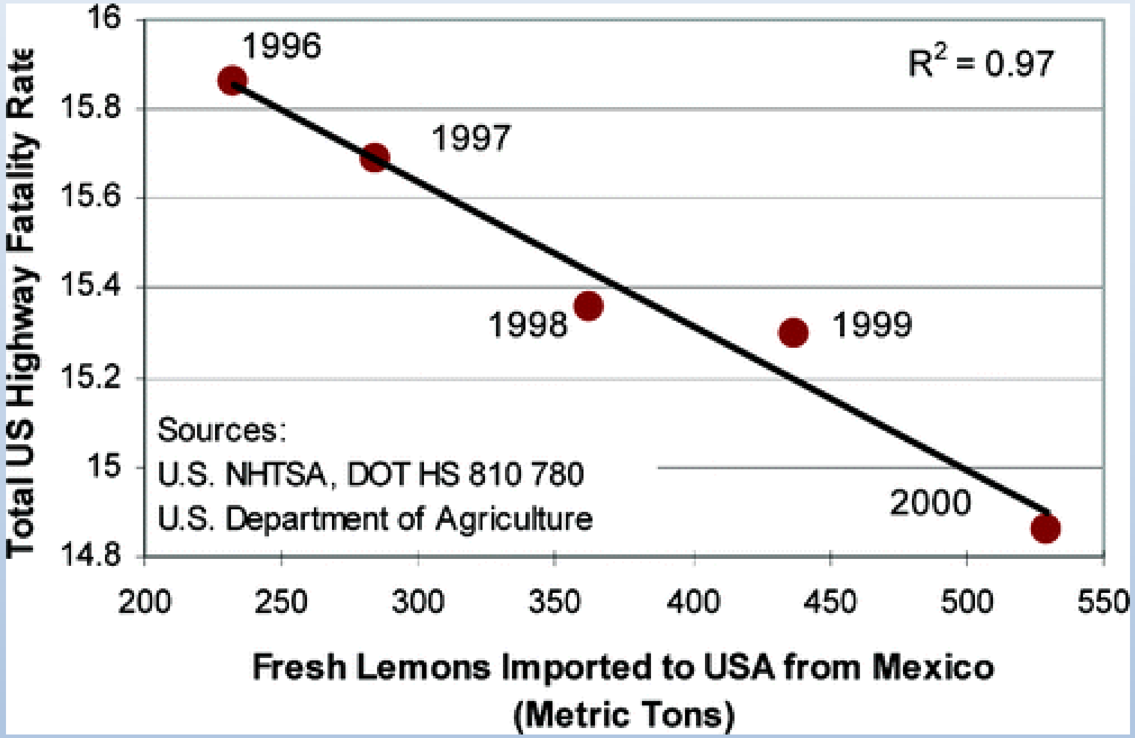
Chocolate and serial killers

- What else is related to chocolate consumption?

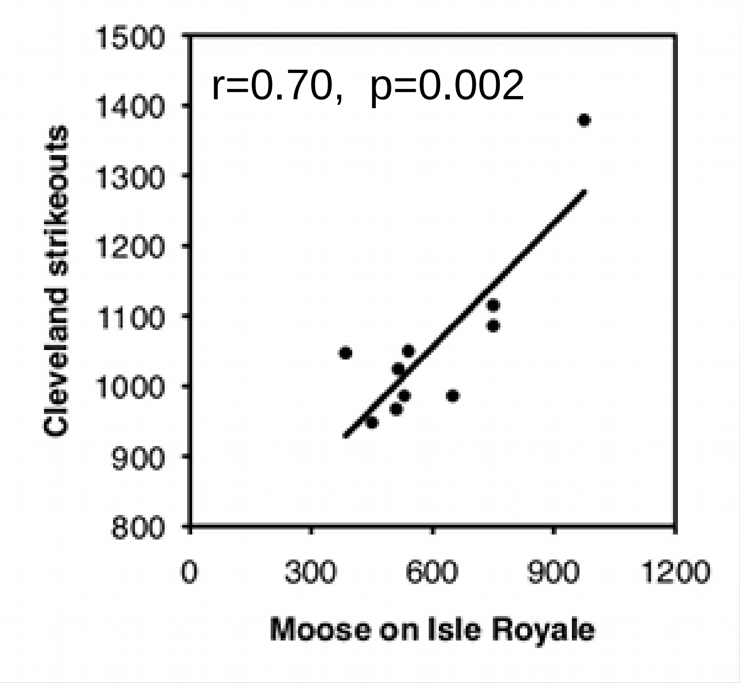


http://www.replicatedtypo.com/chocolate-consumption-traffic-accidents-and-serial-killers/5718.html

Correlation Does Not Prove Causation

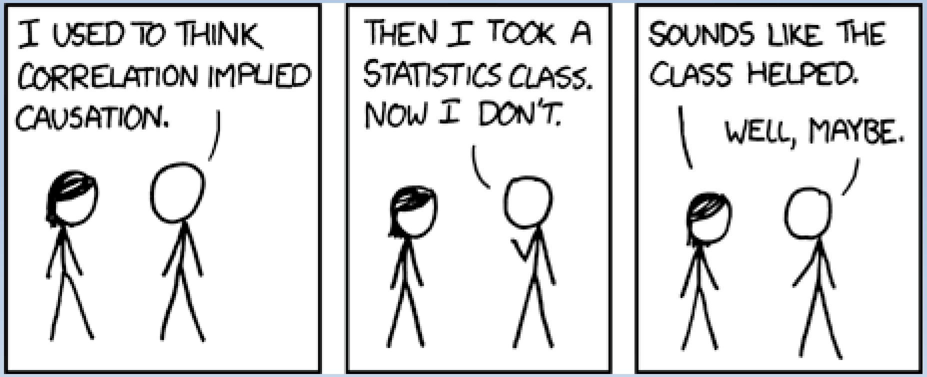


J. Chem. Inf. Model., 2008, 48 (1), pp 25–26



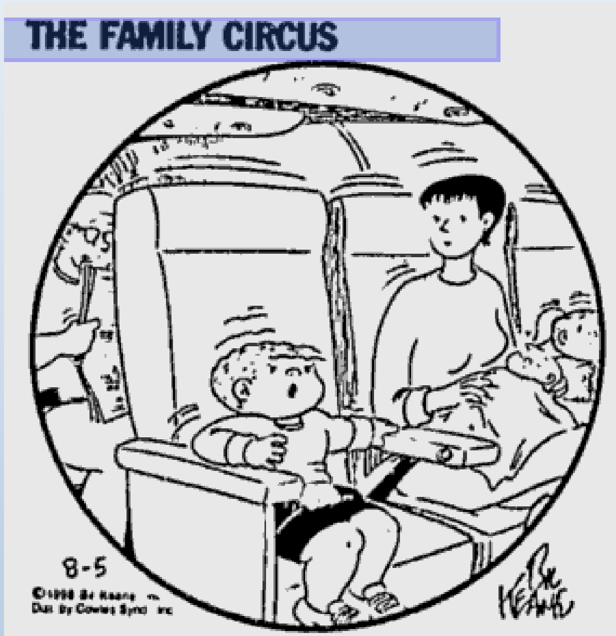
Correlation or Causation?

Number of moose on Isle Royale and strikeouts by the Cleveland baseball team, showing how easy it is to get an impressive-looking correlation from two unrelated data sets.

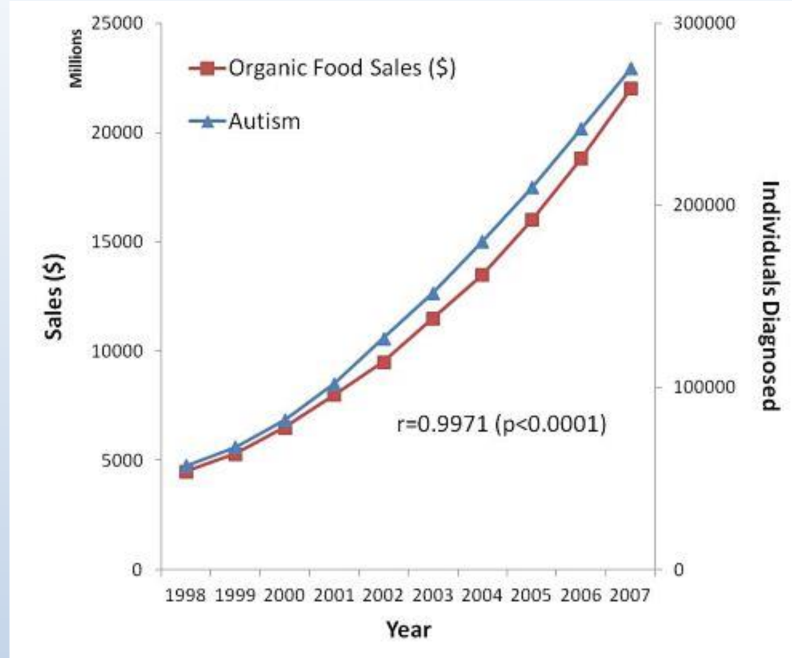


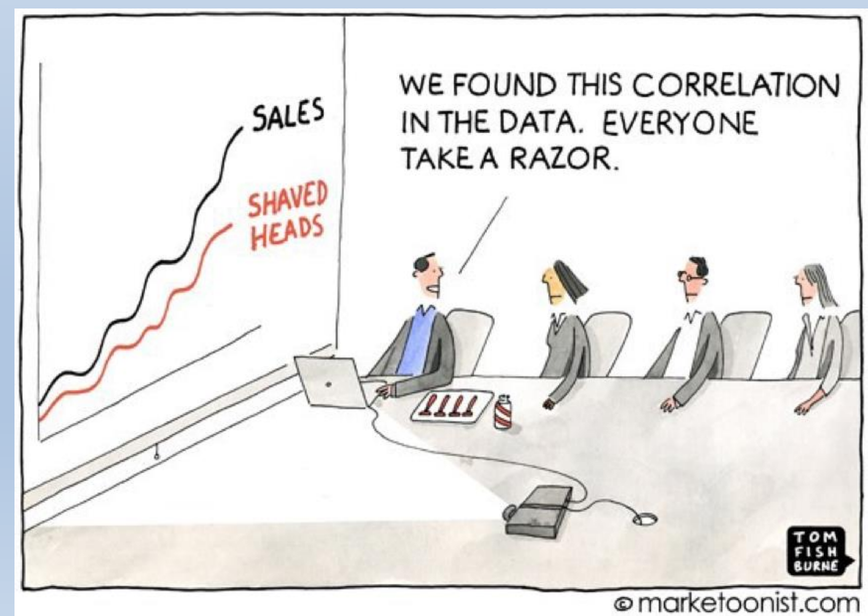
http://www.biostathandbook.com/linearregression.html

http://stats.stackexchange.com/questions/423/what-is-your-favorite-data-analysis-cartoon



"I wish they didn't turn on that seatbelt sign so much! Every time they do, it gets bumpy."





The danger of mixing up causality and correlation:
Ionica Smeets at TEDxDelft

https://www.youtube.com/watch?v=8B271L3NtAw

SIMPLE LINEAR REGRESSION

Simple Linear Regression:
Association between two continuous variables

Prediction of one variable from knowledge another variable

Called "simple" because there are only two variables in the model

How good is a linear model (*y=a+βx,* equation for a straight line) to explain the relationship of two variables?

If there is such a relationship, we can 'predict' the value y for a given x.

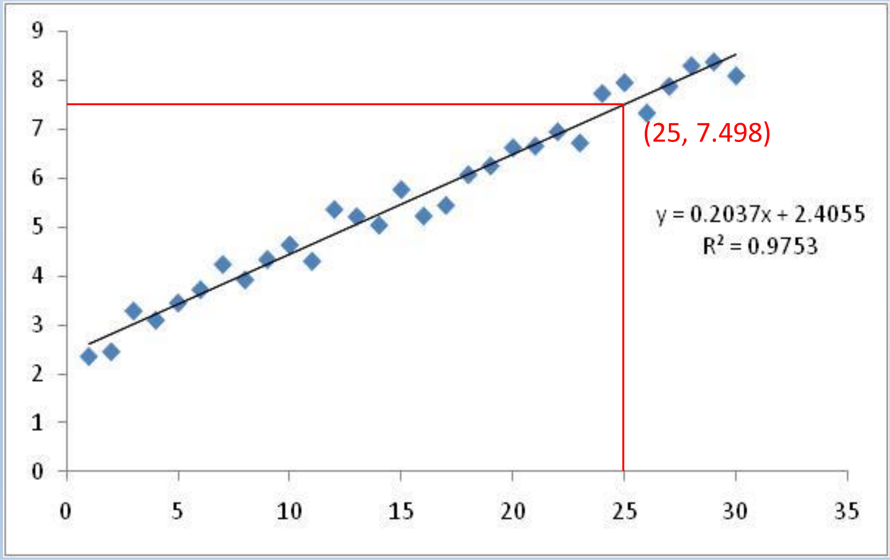
Dependent variable

Independent variable

$$y = a + bx$$

Intercept

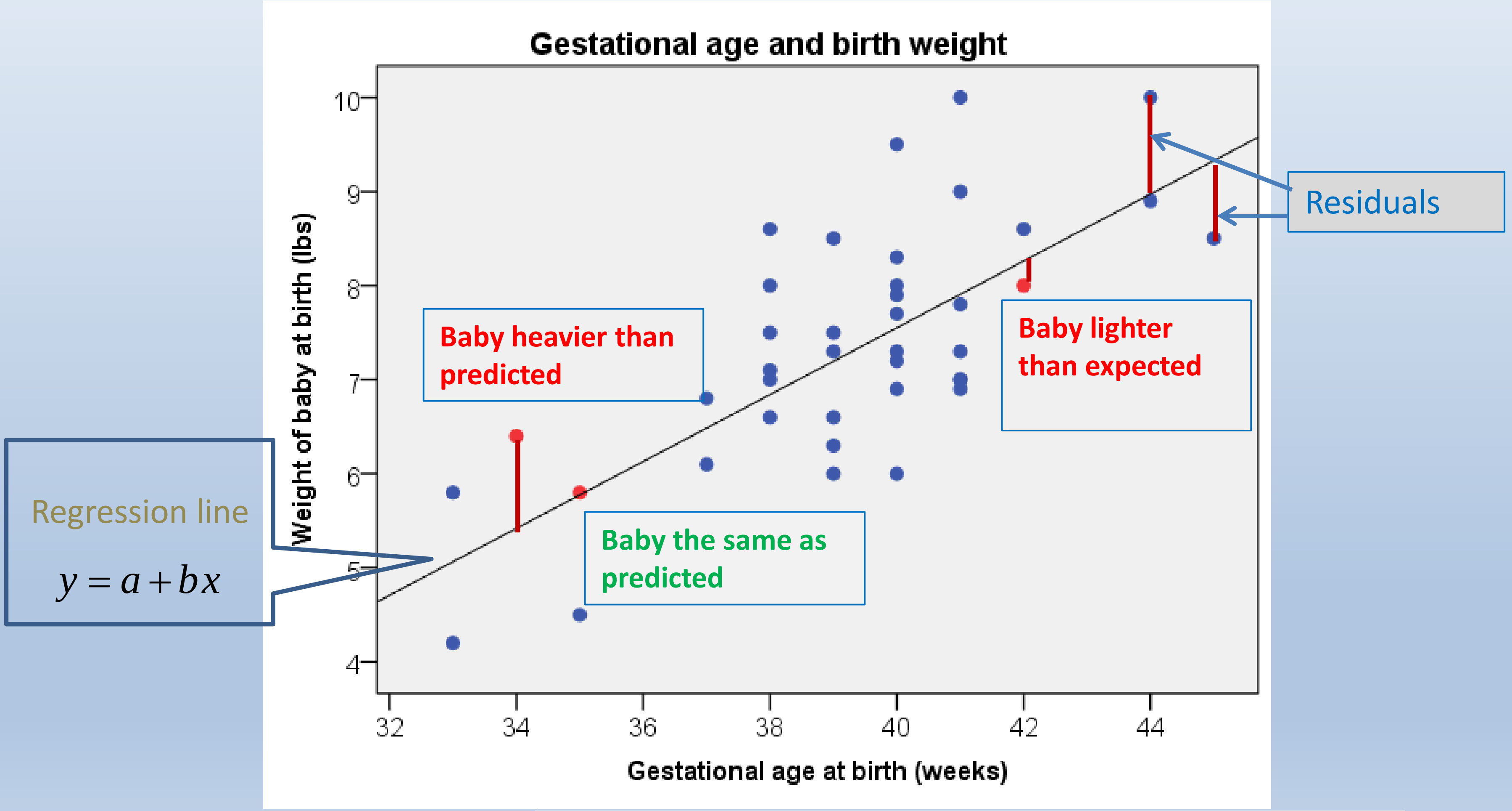
Slope



It involves estimating the line of best fit through the data which minimises the sum of the squared residuals

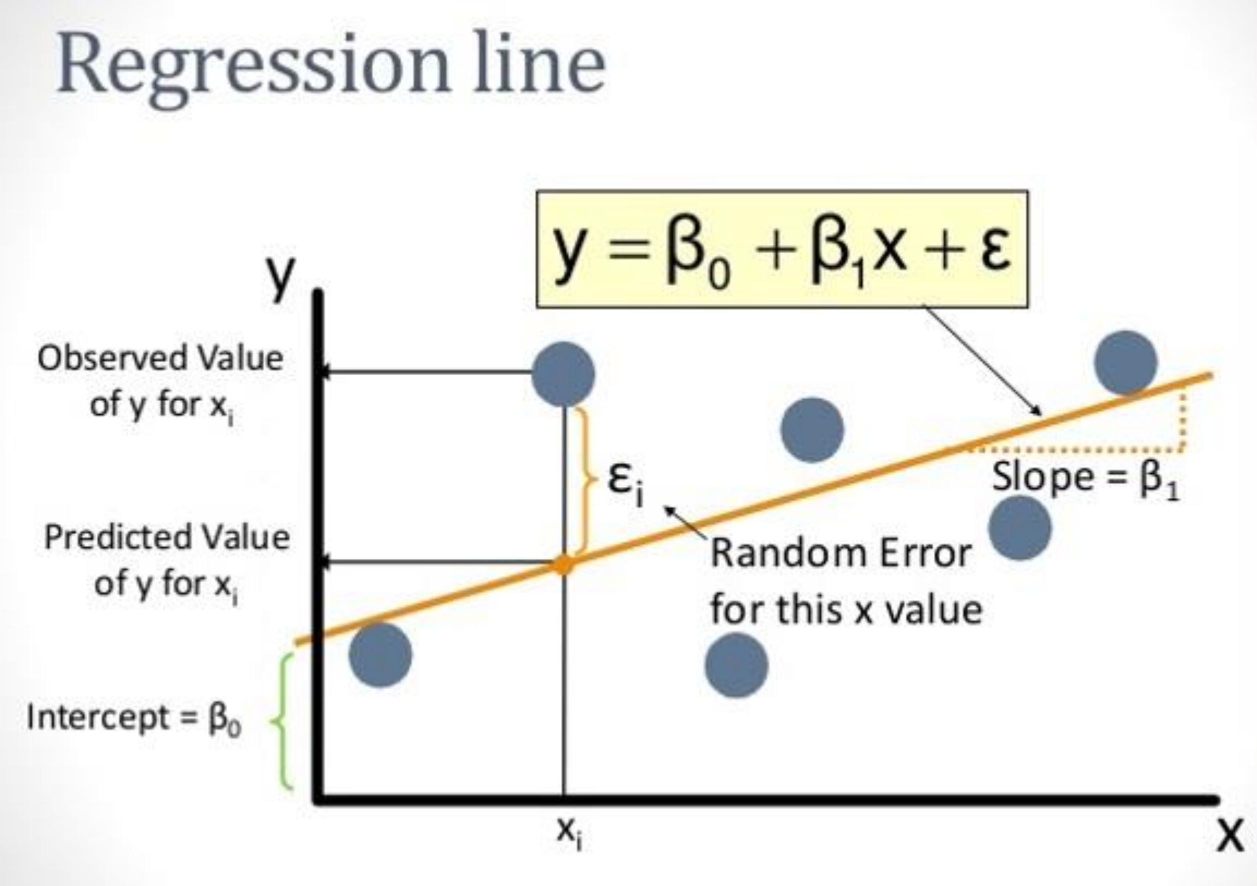
Residuals

Residuals are the differences between the observed dependent variables and the predicted value from the regression equation. These residuals are squared and added together.



X Predictor / explanatory variable (independent variable)

Different way to name the same equation



Hypothesis testing

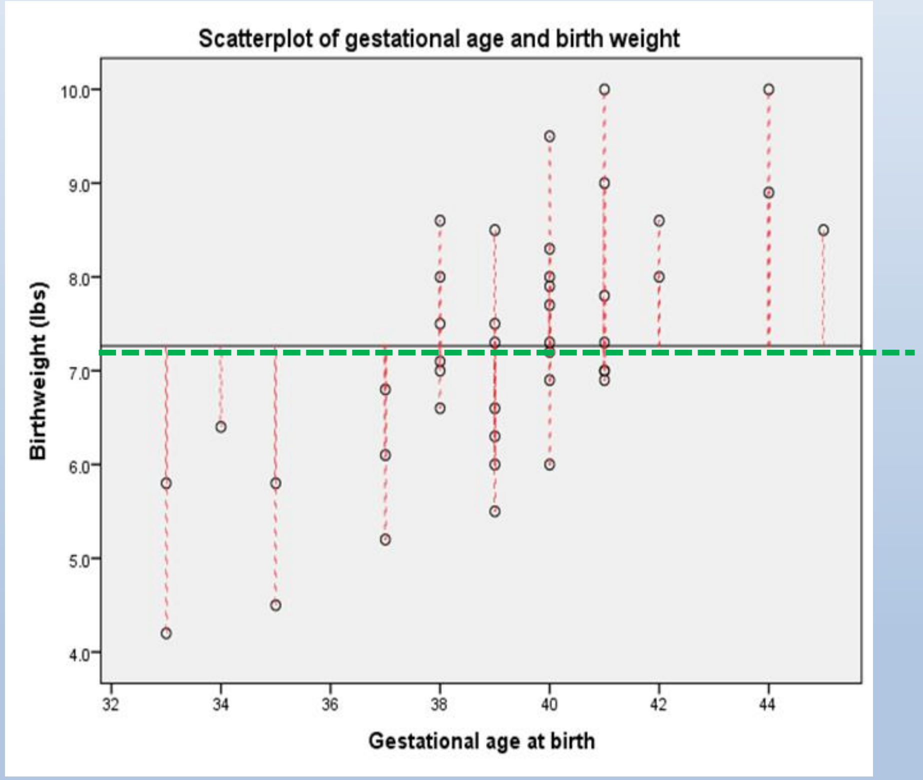
If there is no relationship between X and Y then the value of β (the slope) will be zero.

Green line represents a slope of zero.

No matter how X changes, Y remains the same

$$H_0 : \beta = 0$$

i.e., the slope of the line in the population is 0



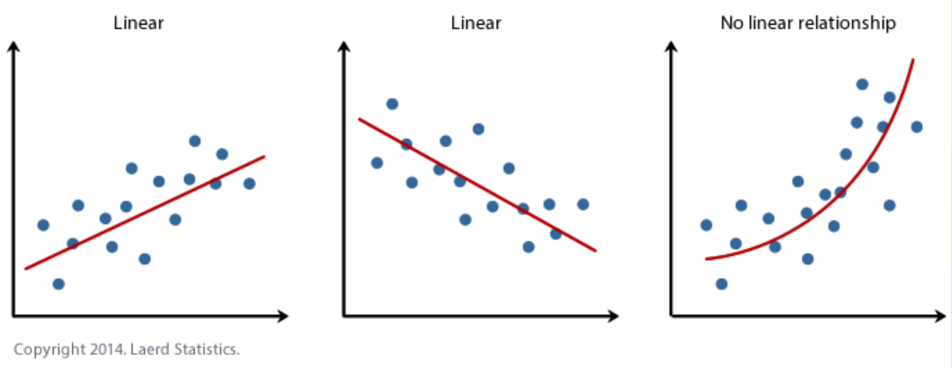
Assumptions of Simple Linear Regression Models

Normally distributed continuous dependent (Y) variable

Most important for small samples; large samples are quite robust against this assumption.

Independent/Predictor variable (X) has a linear relationship with Y

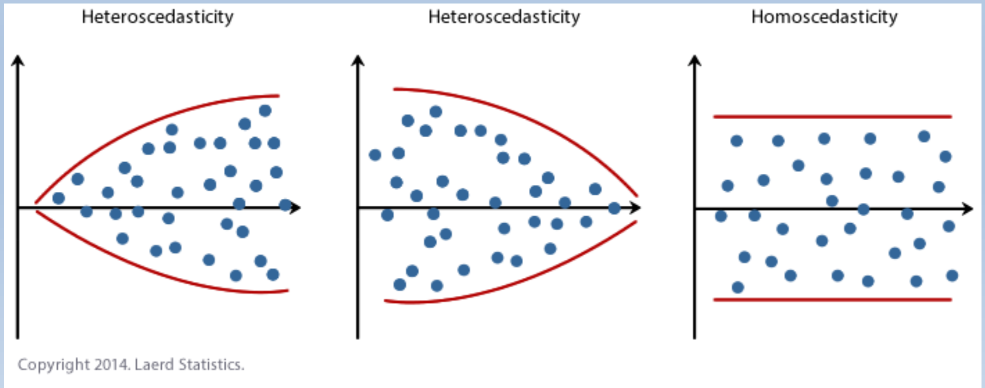
Graphing the data can help evaluate this



Independence

Residuals are normally distributed

The variance of Y at every value of X is the same (homogeneity of variances)

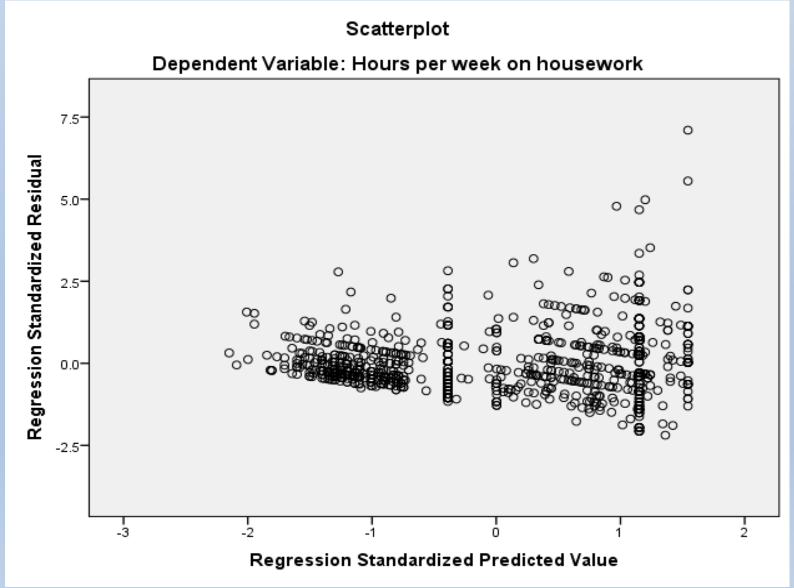


What if assumptions are not met?

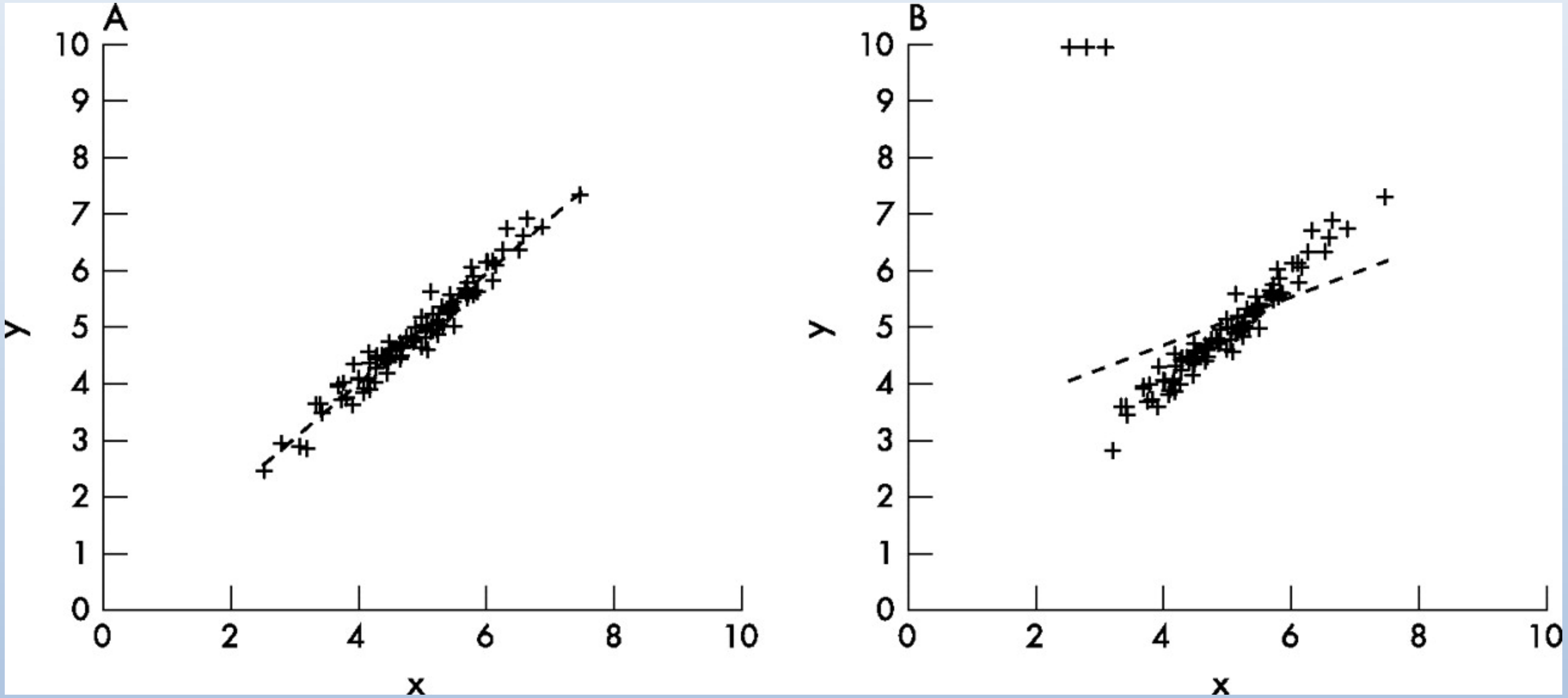
If the residuals are heavily skewed or the residuals show different variances as predicted values increase, the data needs to be transformed

Try taking the natural log (ln) or log10 of the dependent Y variable. Then repeat the analysis and check the assumptions. If necessary, also transform the X variable

Heteroscedasticity

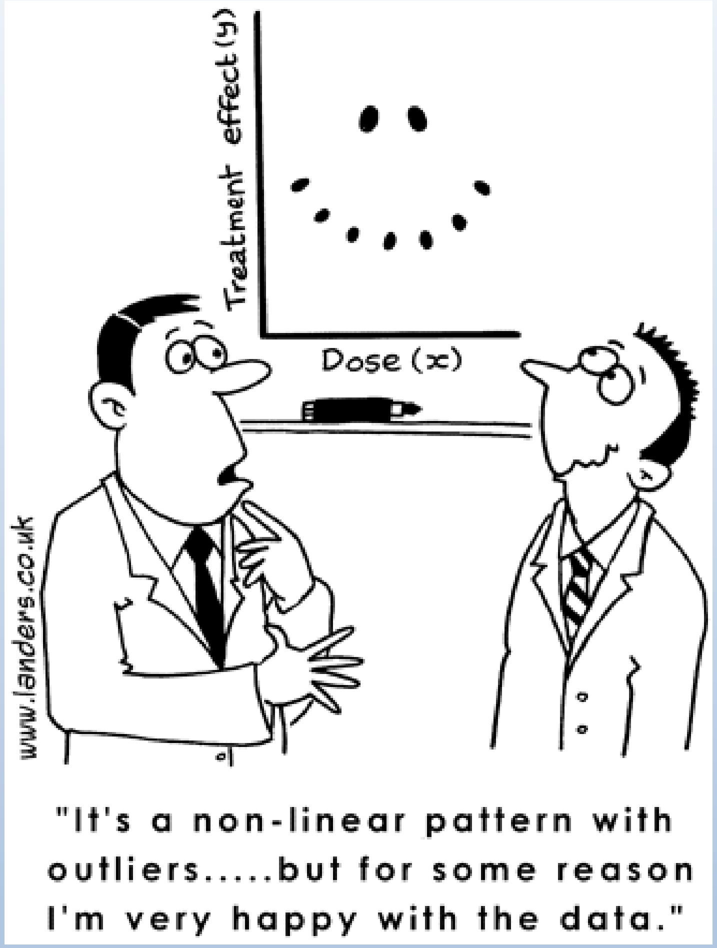


Effect of Outliers



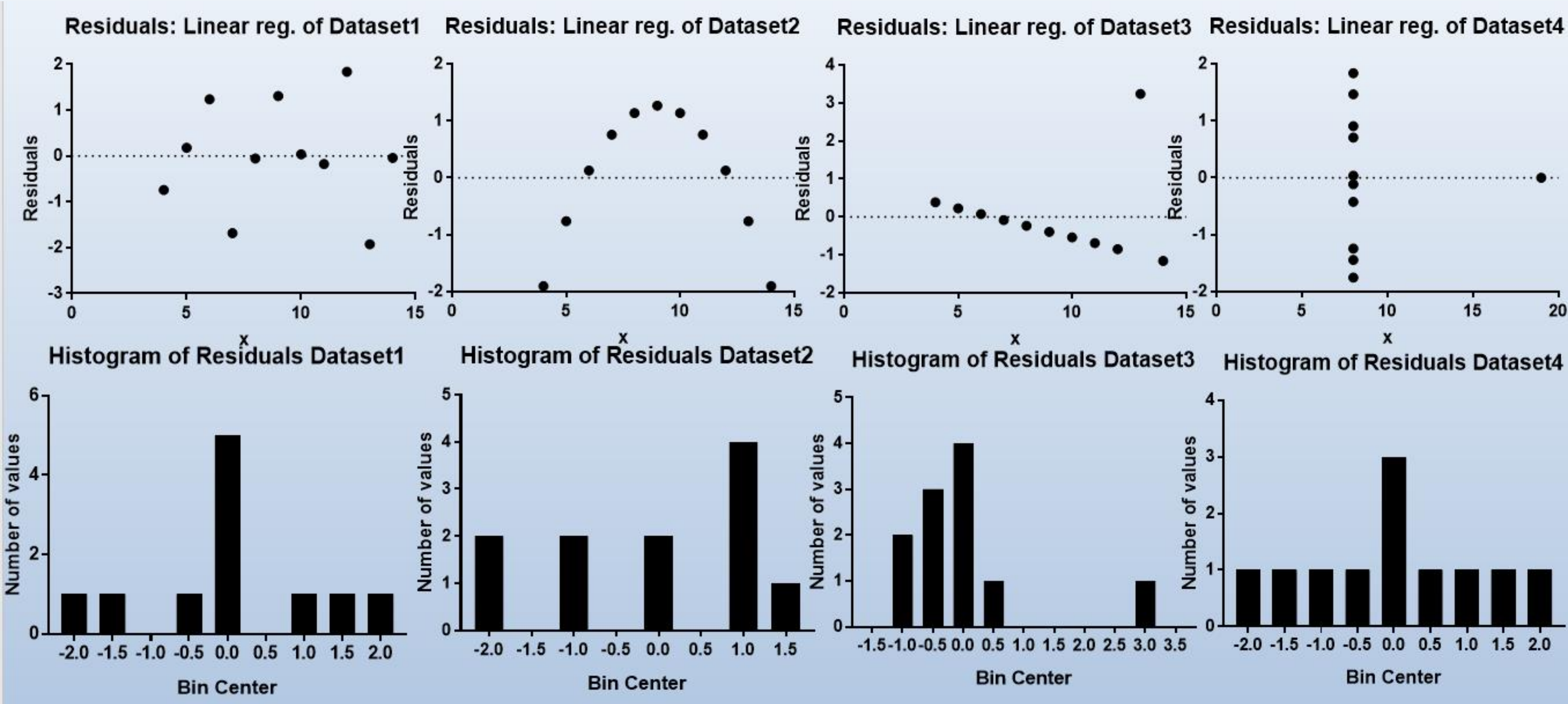
Two scatter plots showing a regression line fitted to data with (A) no outliers and (B) three outliers that have a profound effect on the estimated regression line

Steff Lewis Pract Neurol 2007;7:259-264



Steff Lewis Pract Neurol 2007;7:259-264

Anscombe's quartet: Testing Linear and normality assumptions



Dataset	equation	p
1	y=3.0+0.50x	0.002
2	y=3.0+0.50x	0.002
3	y=3.0+0.50x	0.002
4	y=3.0+0.50x	0.002

Simple Linear Regression Example

Special Article

NEJM 1999, 340:1881-7

THE RELATION BETWEEN FUNDING BY THE NATIONAL INSTITUTES OF HEALTH AND THE BURDEN OF DISEASE

CARY P. GROSS, M.D., GERARD F. ANDERSON, PH.D., AND NEIL R. POWE, M.D., M.P.H., M.B.A.

<u>Measure of Burden</u> (among several)

DALY (disability-adjusted life-years)
one DALY= loss of one year of healthy life to disease

Used 1990 DALY data

CONDITION OR DISEASE	NIH RESEARCH FUNDS 1996
	thousands of dollars (% of total)
AIDS	1,410,925 (28.7)
Breast cancer	381,880 (7.8)
Dementia	304,411 (6.2)
Diabetes mellitus	298,920 (6.1)
Ischemic heart disease	269,100 (5.5)
Alcohol abuse	256,600 (5.2)
Injuries	198,700 (4.0)
Dental and oral disorders	187,100 (3.8)
Cirrhosis	169,800 (3.4)
Depression	143,800 (2.9)
Lung cancer	127,796 (2.6)
Stroke	120,280 (2.4)
Schizophrenia	111,479 (2.3)
Colorectal cancer	105,525 (2.1)
Sexually transmitted diseases	102,583 (2.1)
Prostate cancer	92,661 (1.9)
Multiple sclerosis	82,800 (1.7)
Asthma	81,600 (1.7)
Parkinson's disease	77,158 (1.6)
Tuberculosis	64,125 (1.3)
Chronic obstructive pulmonary disease	62,400 (1.3)
Pneumonia	61,900 (1.3)
Cervical cancer	60,180 (1.2)
Epilepsy	55,100 (1.1)
Ovarian cancer	42,168 (0.8)
Perinatal conditions	26,400 (0.5)
Uterine cancer	13,956 (0.3)
Otitis media	9,100 (0.2)
Peptic ulcer	6,000 (0.1)

Simple Linear Regression Example: Distribution of NIH Funds and DALY

Research Question
Are DALYs associated with NIH funding levels?

Two ways to look at this question

1. Are DALYs and NIH funding level correlated?
   - Pearson or Spearman correlation coefficients
   - Do DALYs and funding increase or decrease together, or does one increase as the other decreases
   - No causation is inferred

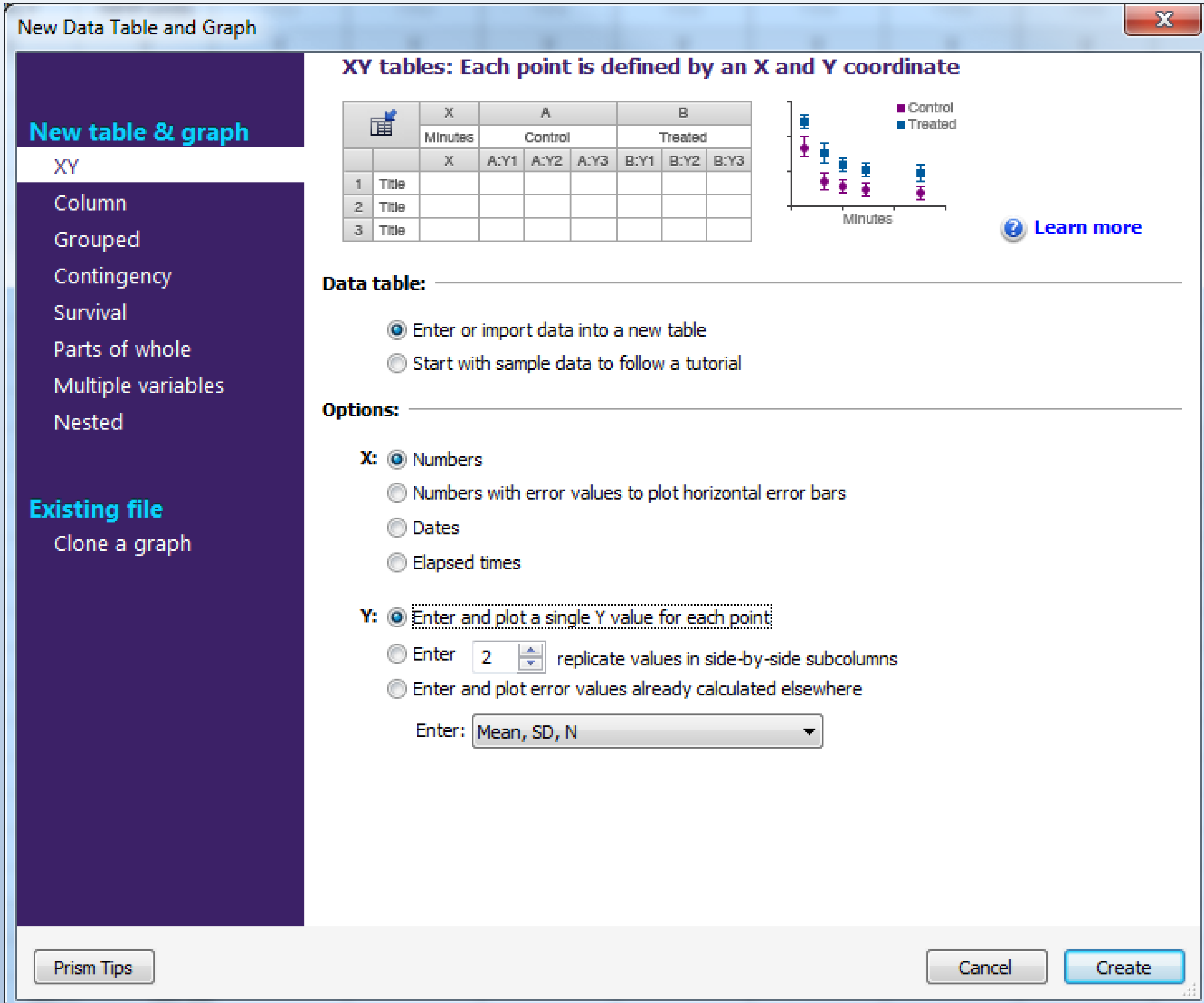
2. Do DALYs "predict" levels of NIH funding
   - Linear regression
   - Higher levels of DALY "cause" NIH funding to increase or decrease
   - Implies a causal relationship – but doesn't prove one

Graph Data, Look for Outliers
Check Normal Distribution

We did this for correlation and determined we needed to $\log_{10}$ transform the data to get a normal distribution

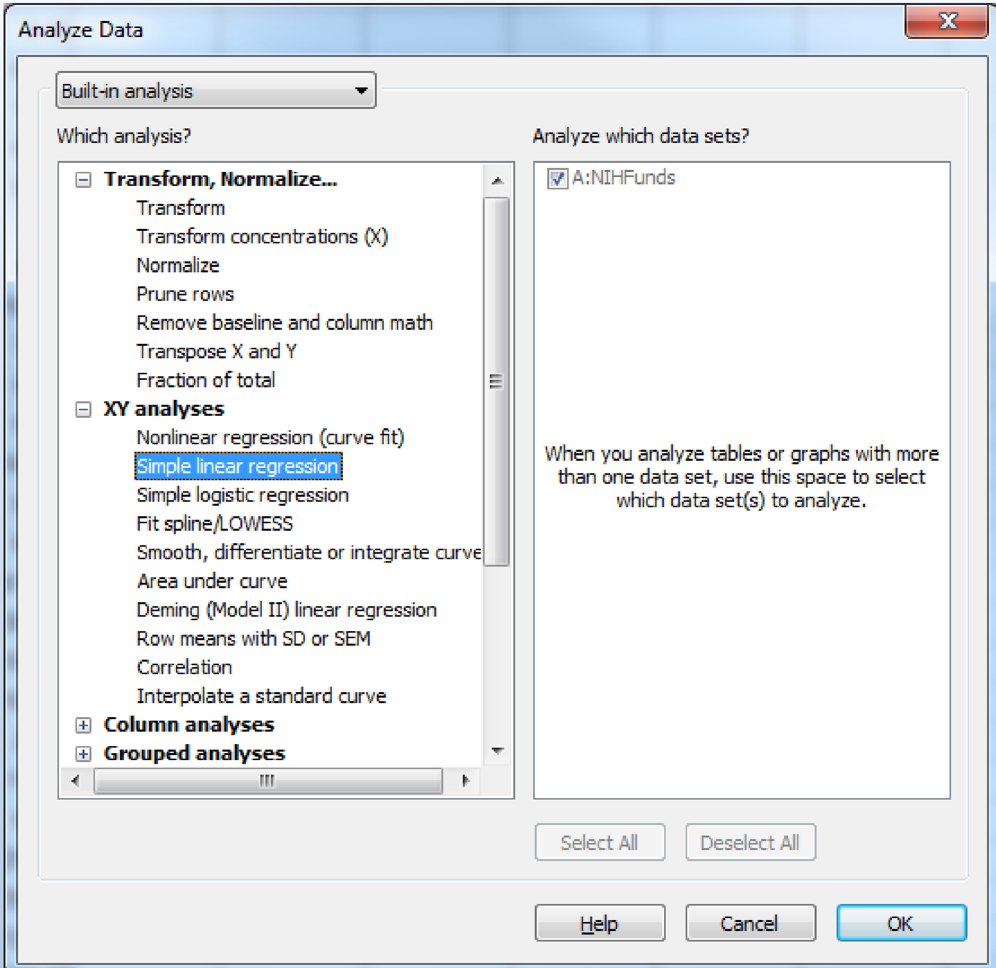
We also identified AIDS as an outlier so will not include in the final analysis.

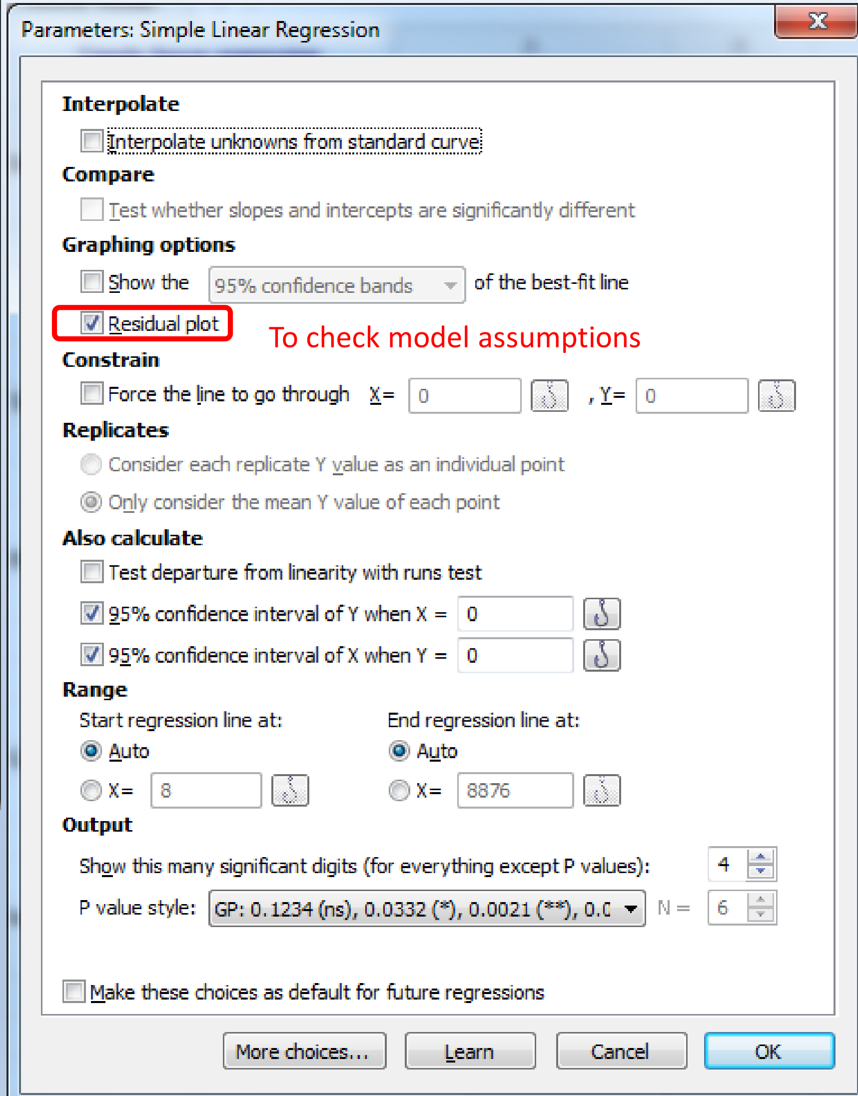
However, we will look at the effects of non-normality and the AIDS outlier

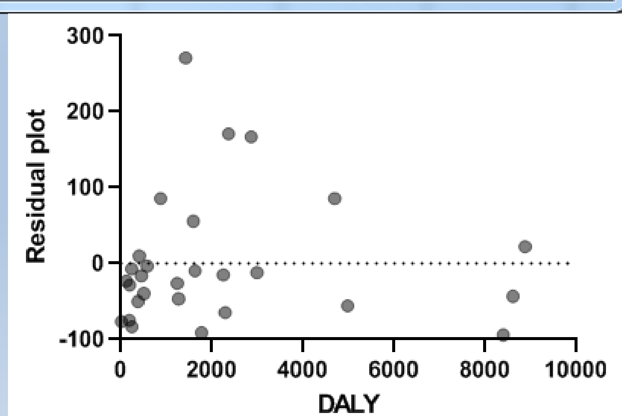


	X	Group A	Group B
	daly	NIHfunds	Title
	X	Y	Y
1	8	9.10	
2	118	64.13	
3	185	13.96	
4	192	60.18	
5	236	82.80	
6	239	6.00	
7	375	42.17	
8	404	102.58	
9	447	77.16	
10	505	55.10	
11	574	92.66	
12	870	187.10	
13	1236	81.60	
14	1263	61.90	
15	1267	1410.93	
16	1421	381.88	
17	1584	169.80	
18	1626	105.53	
19	1767	26.40	
20	2249	111.48	
21	2284	62.40	
22	2357	298.92	
23	2866	304.41	
24	2987	127.80	
25	4690	256.60	
26	4977	120.28	
27	8393	143.80	
28	8608	198.70	
29	8876	269.10	
30			

Without AIDS outlier







Simple linear regression Tabular results	A NIHFunds
Best-fit values	
Slope	0.01819
Y-intercept	85.63
X-intercept	-4707
1/slope	54.97
Std. Error	
Slope	0.006503
Y-intercept	21.93
95% Confidence Intervals	
Slope	0.004826 to 0.03156
Y-intercept	40.54 to 130.7
X-intercept	-24157 to -1440
Goodness of Fit	
R squared	0.2314
Sy.x	88.25
Is slope significantly non-zero?	
F	7.826
DFn, DFd	1, 26
P value	0.0096
Deviation from zero?	Significant
Equation	Y = 0.01819*X + 85.63
Data	
Number of X values	28

Slope 95%CI do not include 0 (the null hypothesis value) so p will be <0.05

Using a simple linear regression model, NIH funding was predicted by DALY (two sided test, $F(1,26)=7.83$, $p=0.01$, $R^2=23.1\%$, $\alpha=0.05$).

For every unit increase in X there is a 0.02 unit increase in Y

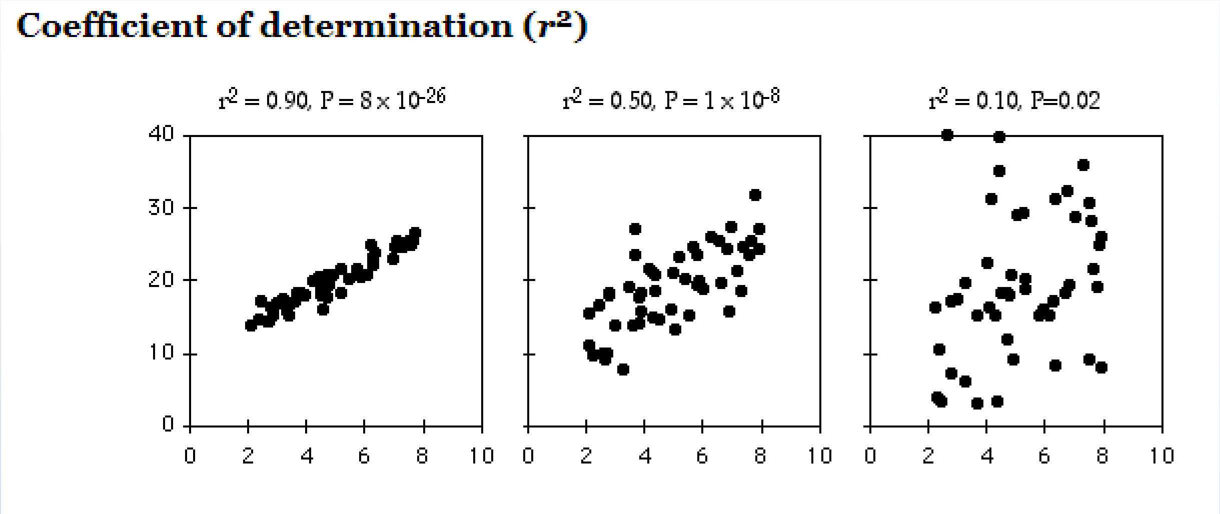
Simple linear regression Tabular results		A NIHFunds
1	**Best-fit values**	
2	Slope	0.01819
3	Y-intercept	85.63
4	X-intercept	-4707
5	1/slope	54.97
6		
7	**Std. Error**	
8	Slope	0.006503
9	Y-intercept	21.93
10		
11	**95% Confidence Intervals**	
12	Slope	0.004826 to 0.03156
13	Y-intercept	40.54 to 130.7
14	X-intercept	-24157 to -1440
15		
16	**Goodness of Fit**	
17	R squared	0.2314
18	Sy.x	88.25
19		
20	**Is slope significantly non-zero?**	
21	F	7.826
22	DFn, DFd	1, 26
23	P value	0.0096
24	Deviation from zero?	Significant
25		
26	**Equation**	Y = 0.01819*X + 85.63
27		
28	**Data**	
29	Number of X values	28

R squared (R^2) is a measure of how well the model fits the data (AKA coefficient of determination).

For simple linear regression, R^2 is the Pearson correlation coefficient squared (r^2).

It therefore takes values between 0 and 1.

Turning it into a percentage makes it easier to explain. Here 23.1% of the variation in NIH funding is explained by DALY in the model.



Three relationships with the same slope, same intercept, and different amounts of scatter around the best-fit line.

r^2 (correlation coefficient squared) is the proportion of the variation in the *Y* variable that is "explained" by the variation in the *X* variable.

values near 1 mean the *Y* values fall almost right on the regression line, while values near 0 mean there is very little relationship between *X* and *Y*.

regressions can have a small r^2 and not look like there's any relationship, yet they still might have a slope that's significantly different from zero.

Checking Assumptions

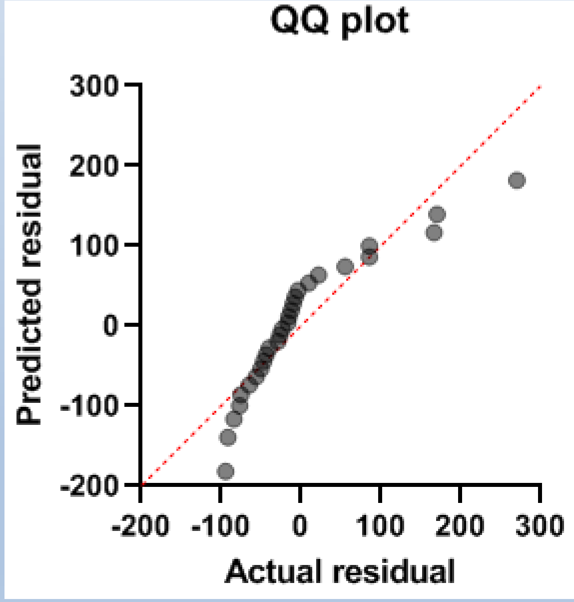
Normally distributed continuous dependent (Y) variable - No

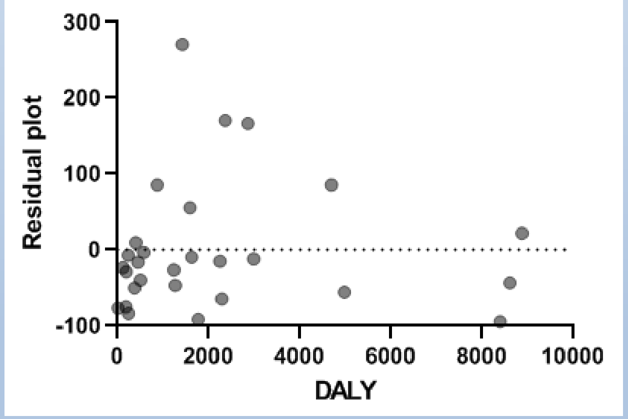
Independent/Predictor variable (X) has a linear relationship with the outcome – Yes?

Independence - Yes

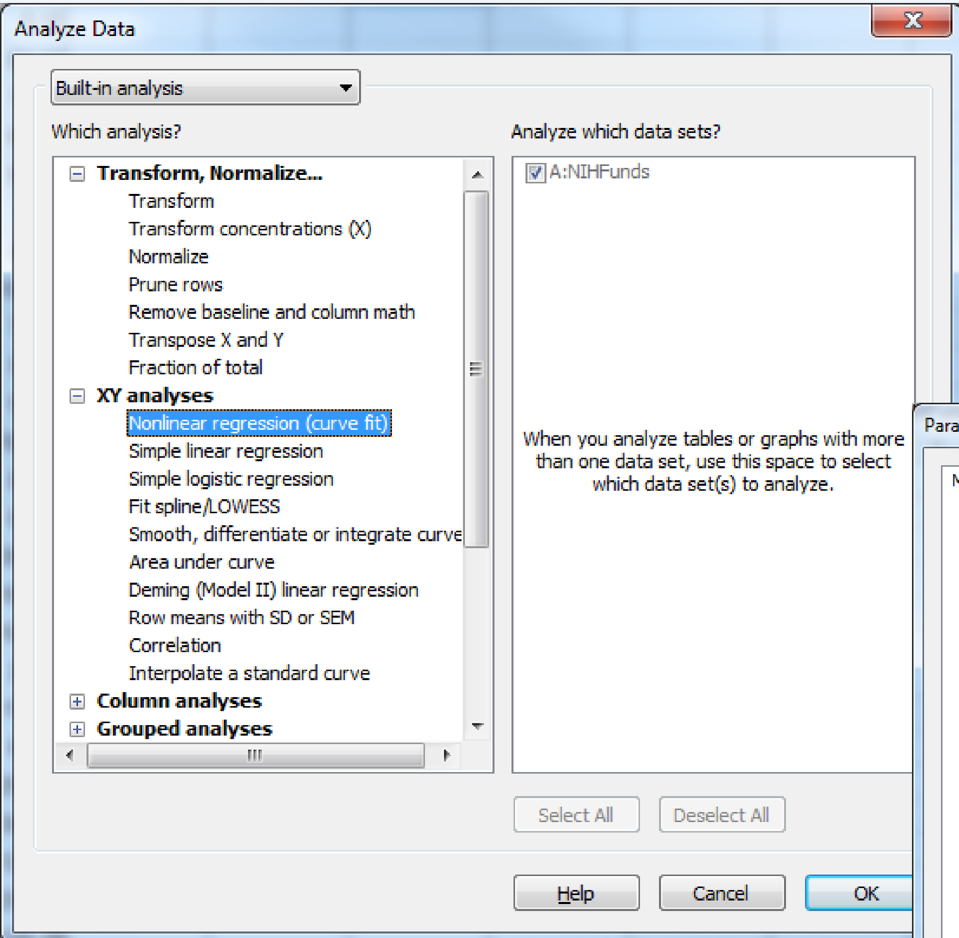
Residuals are normally distributed - No

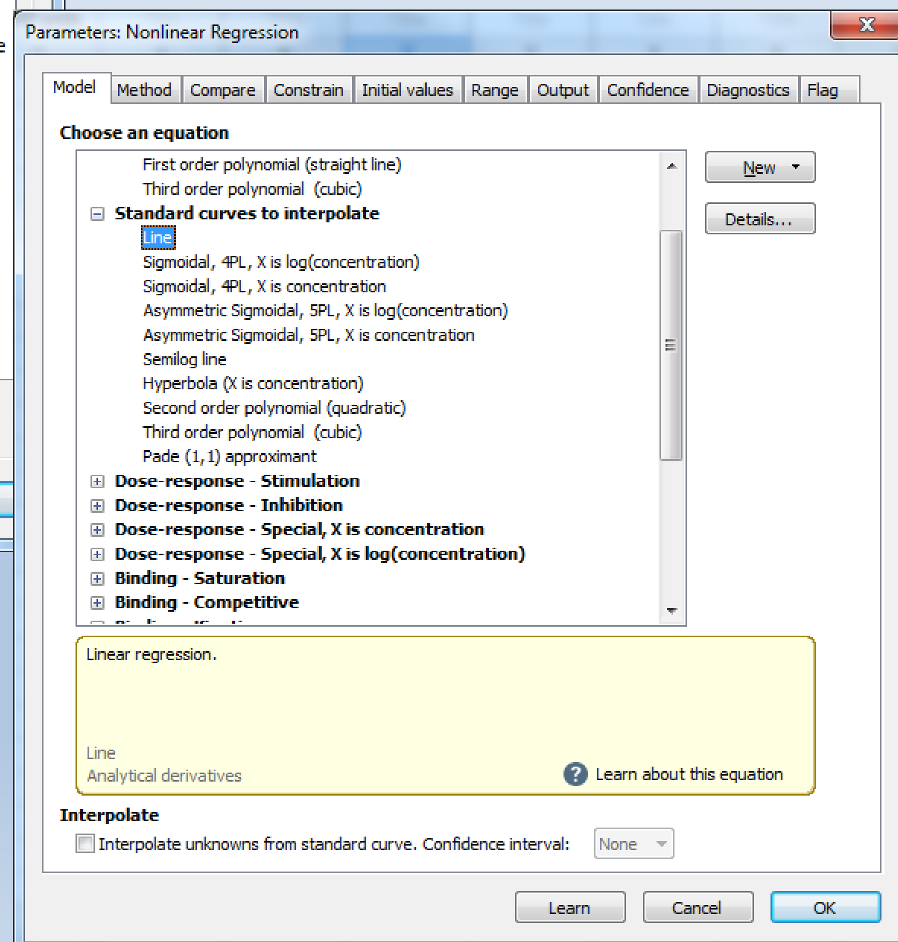
The variance of residuals at every value of X is the same - No

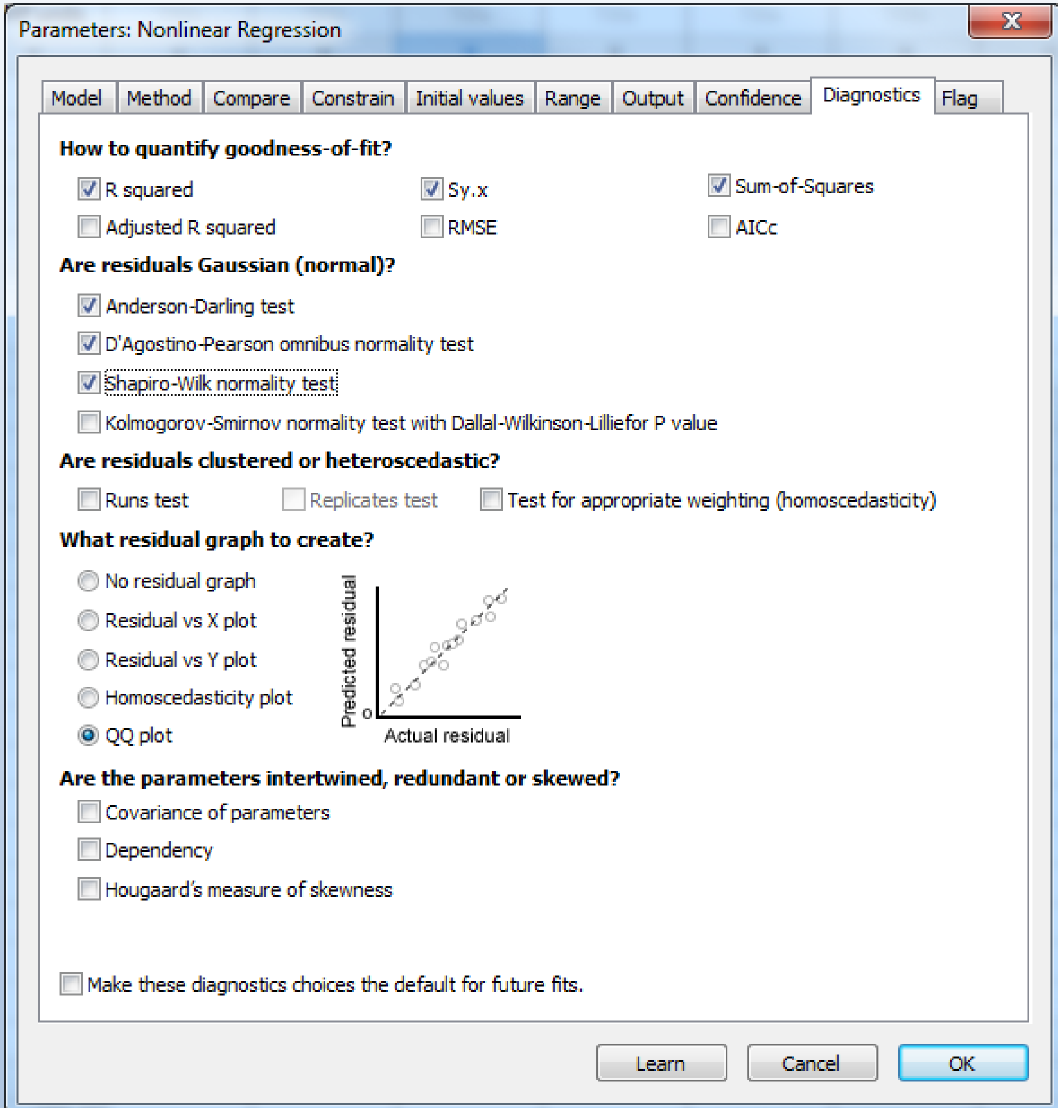


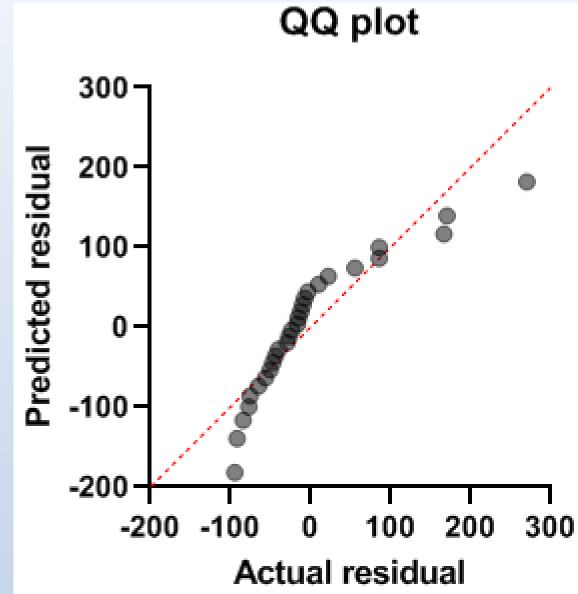


Use straight line or line regression options in nonlinear regression to get more options including Q-Q plot to check normality of residuals and homoscedasticity plot



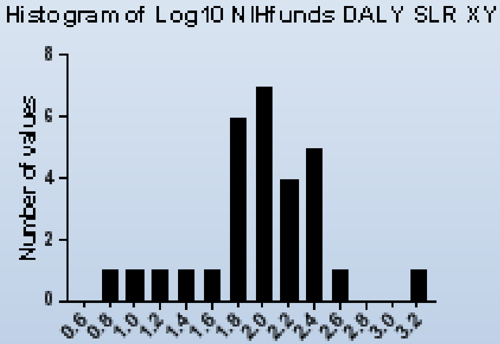


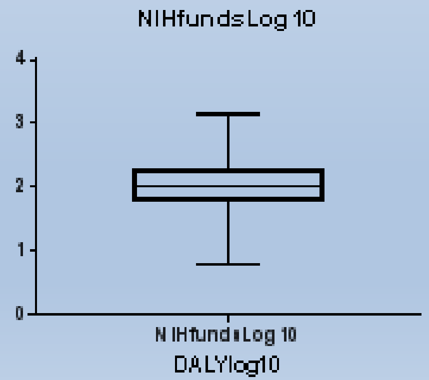


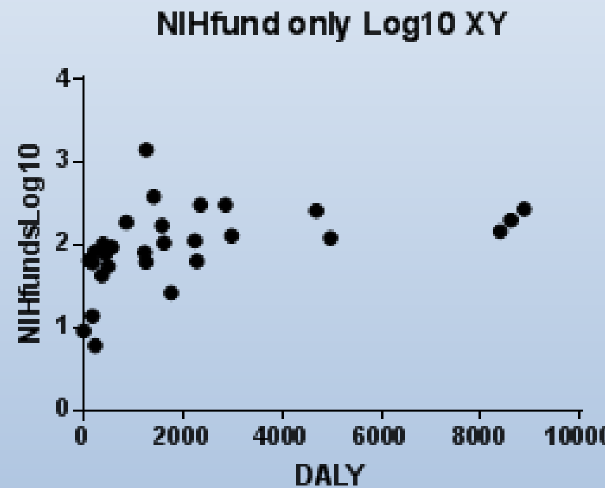


Normality of Residuals	
Anderson-Darling (A2*)	1.608
P value	0.0003
Passed normality test (alpha=0.05)?	No
P value summary	***
D'Agostino-Pearson omnibus (K2)	15.56
P value	0.0004
Passed normality test (alpha=0.05)?	No
P value summary	***
Shapiro-Wilk (W)	0.8325
P value	0.0004
Passed normality test (alpha=0.05)?	No
P value summary	***

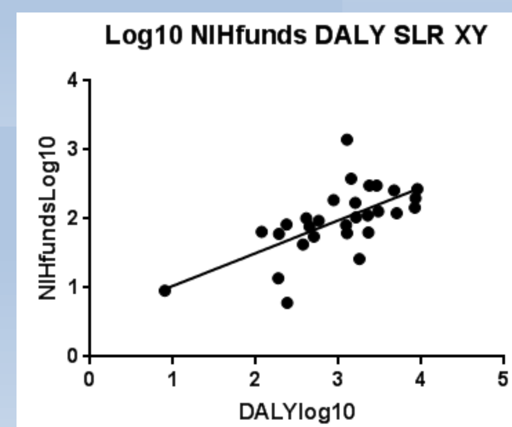
To meet the test assumptions, the Y variable needs to be normal (after log transformation) and in a linear relationship with the X variable



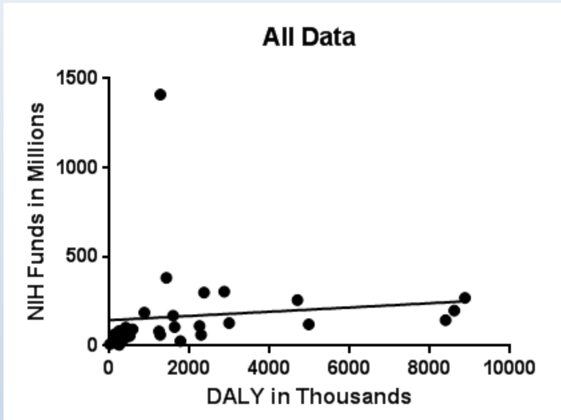




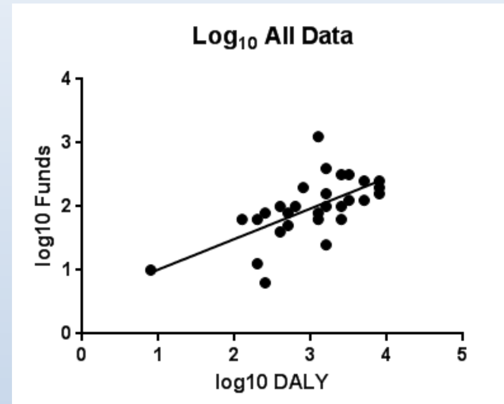
Not linear so let’s transform the X variable, too



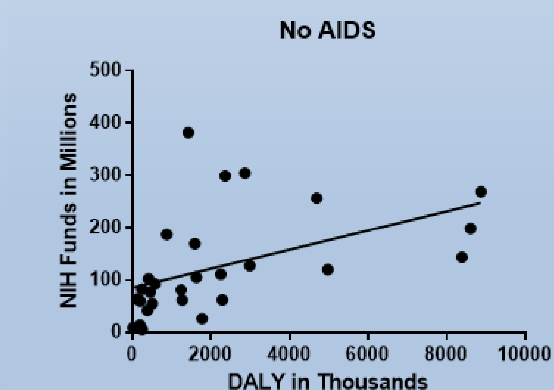
Simple Linear Regression Example:
NIH Funds and DALY



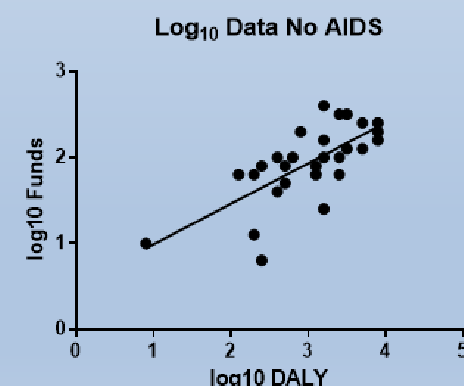
β = 0.01, p=0.54
Y = 0.01*X + 144.10



β = 0.48, p=0.0002
Y = 0.48*X + 0.53



β = 0.01, p=0.01
Y = 0.01*X + 85.63



β = 0.47, p<0.0001
Y = 0.47*X + 0.52

Simple linear regression Tabular results	A NIHfundsLog10
1 **Best-fit values**	
2 Slope	0.4657
3 Y-intercept	0.5421
4 X-intercept	-1.164
5 1/slope	2.147
6	
7 **Std. Error**	
8 Slope	0.09148
9 Y-intercept	0.2801
10	
11 **95% Confidence Intervals**	
12 Slope	0.2776 to 0.6537
13 Y-intercept	-0.03363 to 1.118
14 X-intercept	-3.998 to 0.05180
15	
16 **Goodness of Fit**	
17 R squared	0.4991
18 Sy.x	0.3207
19	
20 **Is slope significantly non-zero?**	
21 F	25.91
22 DFn, DFd	1, 26
23 P value	<0.0001
24 Deviation from zero?	Significant
25	
26 **Equation**	Y = 0.4657*X + 0.5421
27	
28 **Data**	
29 Number of X values	28

Transformed data

By log transforming the data we met the assumptions a simple linear regression model. NIH funding can be predicted by DALY (two-sided test, $F(1,26)=25.9$, $p<0.0004$, $R^2=49.9\%$, $\alpha=0.05$). The regression equation is

NIHfundsLog10 = 0.47 x DALYLog10 + 0.54.

For every 1% increase in X there is a 0.47% increase in Y

Calculating Y Values from X Values

Say we wanted to know what Y (NIH funds) would be if X (DALY) =1000. The model we chose as most accurately representing the relationship of DALY to NIH Funds was the $\log_{10}$ data.

$$Log10(Y) = 0.47*Log10(X) + 0.54$$

Start by taking the log10 of 1000 so we can plug the appropriate number into the equation.

$\log10(1000)=3.0$

$Y = (0.47)(3.0) + 0.54 = 1.95$

To transform back to a natural number $Y = 10^{1.95}$ = **89.12 (=$89,120,000)**

Let's check this value against what would be predicted for the model with the raw data.

$$Y=0.01*X + 85.63$$

$Y = (0.01)(1000) + 85.63$ = **95.63 (=$95,630,000)**

Why the difference? The non-transformed model did not meet all the assumptions so the equation represents a biased estimate of the relationship