Babraham Institute

Analysis of Quantitative data
Linear and non-linear relationships

Anne Segonds-Pichon
v2020-08

Babraham Bioinformatics

Association between 2 continuous variables

Linear relationship

Signal-to-noise ratio

$$\frac{\text{Similarity}}{\text{Variability}} = \frac{\text{Signal}}{\text{Noise}}$$

$$\frac{\text{Signal}}{\text{Noise}} = \textbf{statistical significance}$$

$$\frac{\text{Signal}}{\text{Noise}} = \textbf{no statistical significance}$$

Signal-to-noise ratio and Correlation

$$\frac{\text{Similarity}}{\text{Variability}} = \frac{\text{Signal}}{\text{Noise}}$$

- Signal is **similarity** of behaviour between variable x and variable y.
- **Coefficient of correlation**: $r = \frac{\text{similarity}}{\text{variability}} = \frac{\text{Signal}}{\text{Noise}}$

covariance

$$r = \frac{\text{similarity}}{\text{variability}} = \frac{COV_{xy}}{SD_x SD_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1) SD_x SD_y}$$

Standard Deviation

Correlation

- Most widely-used correlation coefficient:
  - **Pearson product-moment correlation coefficient "r"**
    - The **magnitude** and the **direction** of the relation between 2 variables
    - It is designed to range in value between **-1 and +1**
    - **-0.6** < r > **+0.6** : exciting

Coefficient (+ve or –ve)	Strength of the relationship
0.0 to 0.2	Negligible
0.2 to 0.4	Weak
0.4 to 0.7	Moderate
0.7 to 0.9	Strong
0.9 to 1.0	Very strong

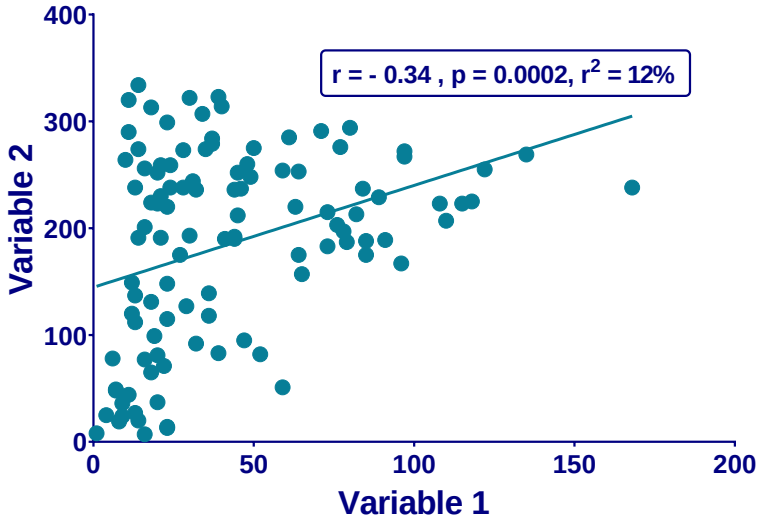
  - **Coefficient of determination "r^2"**
    - It gives the proportion of variance in Y that can be explained by X (in percentage).
      - It helps with the interpretation of r
      - It's basically the **effect size**

Correlation

$p = 0.0002$ 

$r = -0.34$ 

$r^2 = 0.12$ 

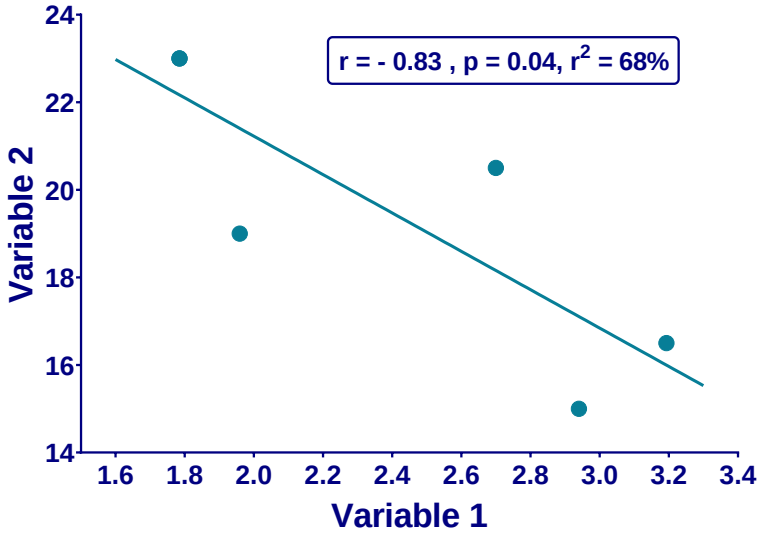


Power!!

$p = 0.04$ 

$r = -0.83$ 

$r^2 = 0.68$ 



Coefficient of determination

Unpaired t test

1	Table Analyzed	coyotes
2		
3	Column B	Males
4	vs.	vs.
5	Column A	Females
6		
7	**Unpaired t test**	
8	P value	0.1045
9	P value summary	ns
10	Significantly different (P < 0.05)?	No
11	One- or two-tailed P value?	Two-tailed
12	t, df	t=1.641, df=84
13		
14	**How big is the difference?**	
15	Mean of column A	89.71
16	Mean of column B	92.06
17	Difference between means (B - A) ± SEM	2.344 ± 1.428
18	95% confidence interval	-0.4964 to 5.185
19	R squared (eta squared)	0.03107

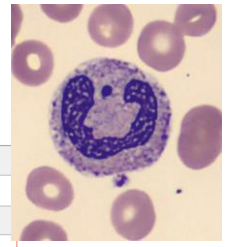
3%

5 cell lines on protein expression

Ordinary one-way ANOVA
ANOVA results

1	Table Analyzed	Transform of Protein expression
2	Data sets analyzed	A-E
3		
4	**ANOVA summary**	
5	F	8.127
6	P value	<0.0001
7	P value summary	****
8	Significant diff. among means (P < 0.05)?	Yes
9	R square	0.3081

31%



RM one-way ANOVA
ANOVA results

1	Table Analyzed	Neutrophils
2		
3	**Repeated measures ANOVA summary**	
4	Assume sphericity?	No
5	F	28.57
6	P value	0.0002
7	P value summary	***
8	Statistically significant (P < 0.05)?	Yes
9	Geisser-Greenhouse's epsilon	0.6916
10	R square	0.8772

88%

Beer goggles effect

2way ANOVA
ANOVA results

1	Table Analyzed	data for 2-way				
2						
3	**Two-way ANOVA**	Ordinary				
4	Alpha	0.05				
5						
6	**Source of Variation**	**% of total variation**	**P value**	**P value summary**	**Significant?**	
7	Interaction	22.06	<0.0001	****	Yes	
8	Alcohol Consumption	37.16	<0.0001	****	Yes	
9	Gender	1.882	0.1614	ns	No	
10						
11	**ANOVA table**	**SS**	**DF**	**MS**	**F (DFn, DFd)**	**P value**
12	Interaction	1978	2	989.1	F (2, 42) = 11.91	P<0.0001
13	Alcohol Consumption	3332	2	1666	F (2, 42) = 20.07	P<0.0001
14	Gender	168.8	1	168.8	F (1, 42) = 2.032	P=0.1614
15	Residual	3488	42	83.04		
16						

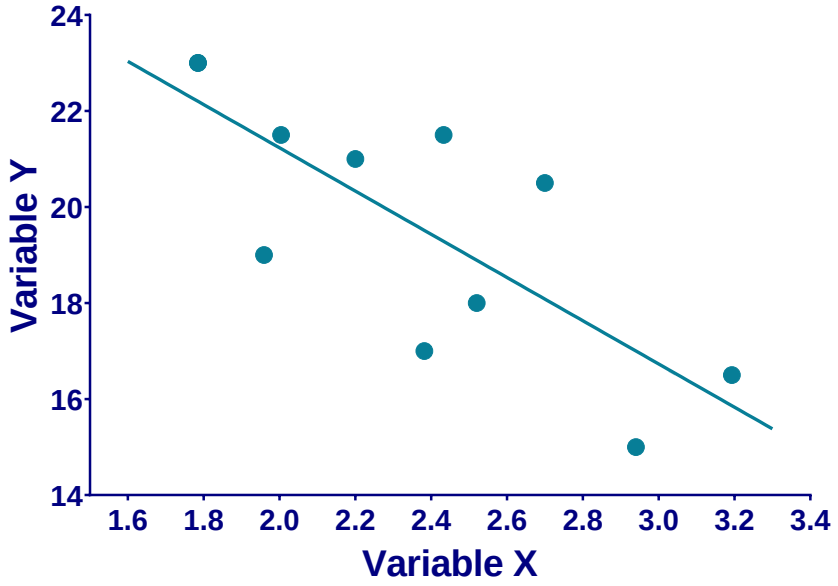
22%+37%+1.88% = 61%

Correlation: Two more things

Thing 1: Pearson correlation is a parametric test

First assumption for parametric test: Normality

Correlation: bivariate Gaussian distribution



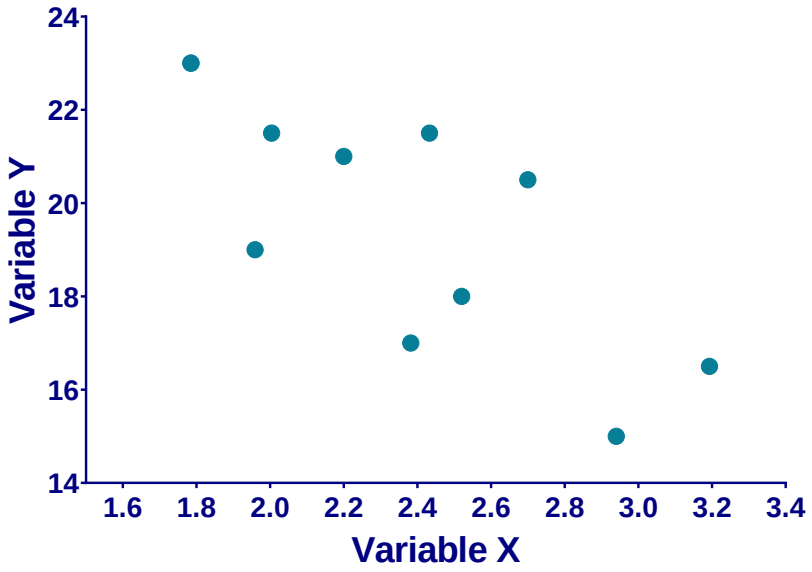
Symmetry-ish of the values on either side of the line of best fit.

Correlation: Two more things

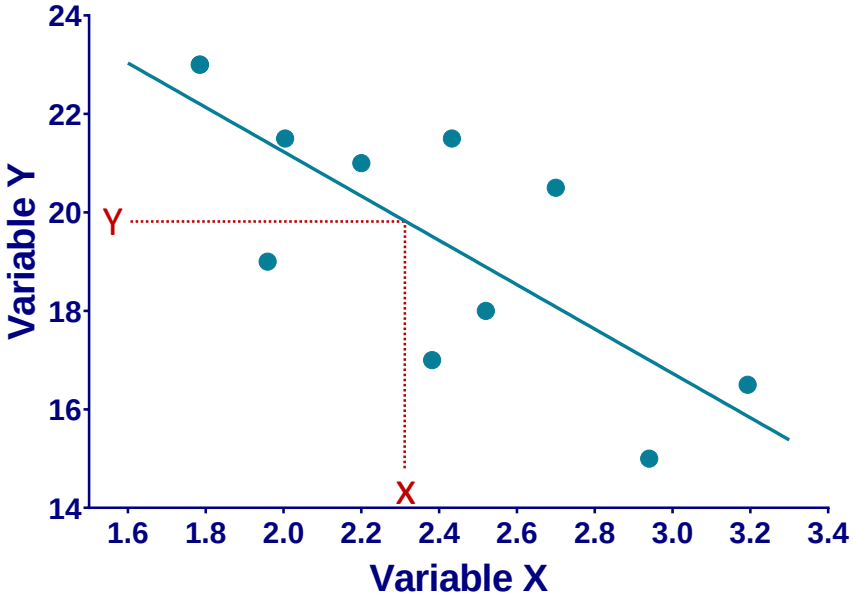
Thing 2: Line of best fit comes from a regression

Correlation: nature and strength of the association

Regression: nature and strength of the association and **prediction**



Correlation = Association



Regression = Prediction

Y = A*X + B

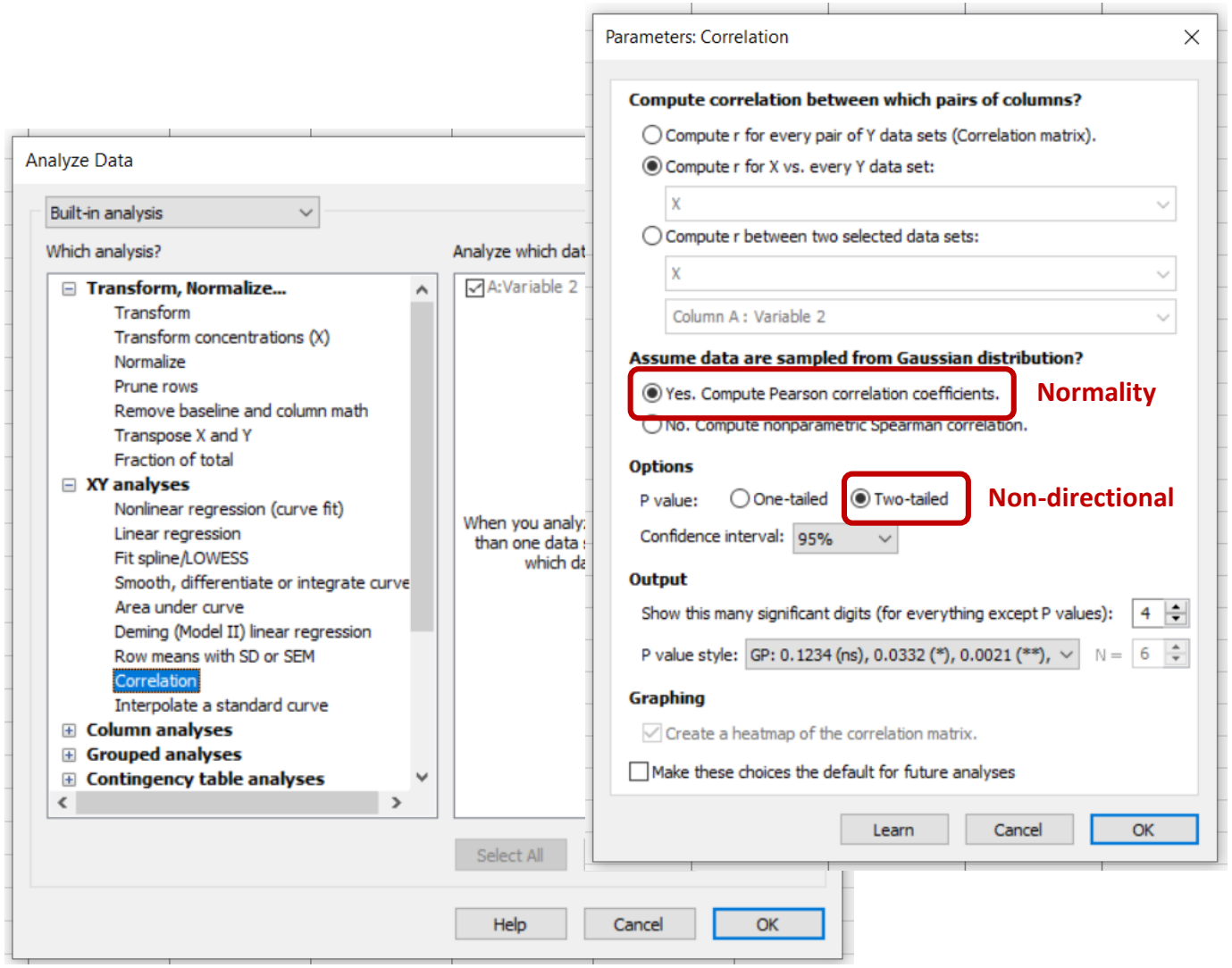
Correlation

treelight.xlsx

Amount of light in a tree

Have a go!

Correlation with Prism 8



Results

	Correlation	Depth vs. Light
1	**Pearson r**	
2	r	-0.8465
3	95% confidence interval	-0.9530 to -0.5537
4	R squared	0.7165
5		
6	**P value**	
7	P (two-tailed)	0.0003
8	P value summary	***
9	Significant? (alpha = 0.05)	Yes
10		
11	**Number of XY Pairs**	13

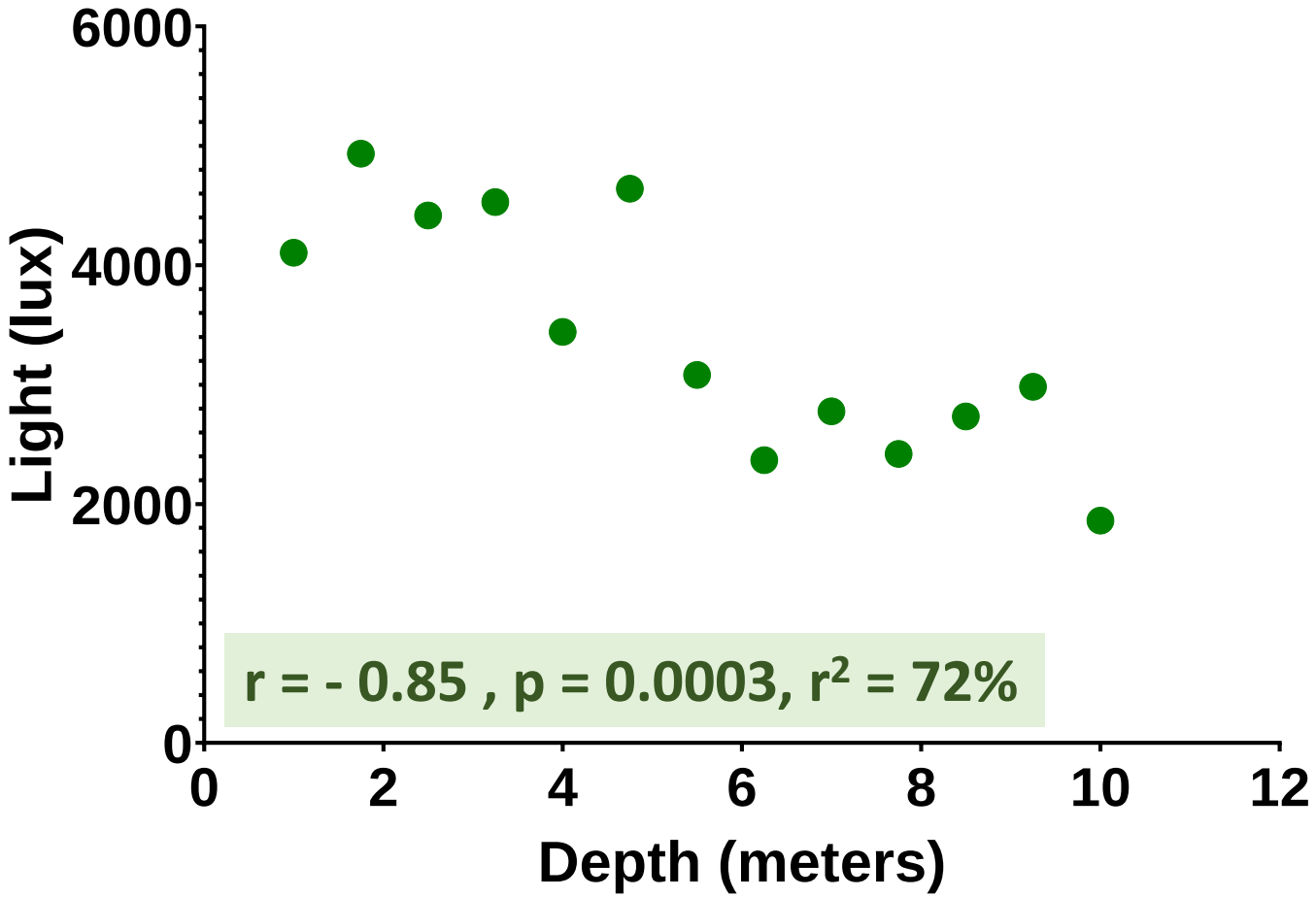
Association

Strong Negative

Significant

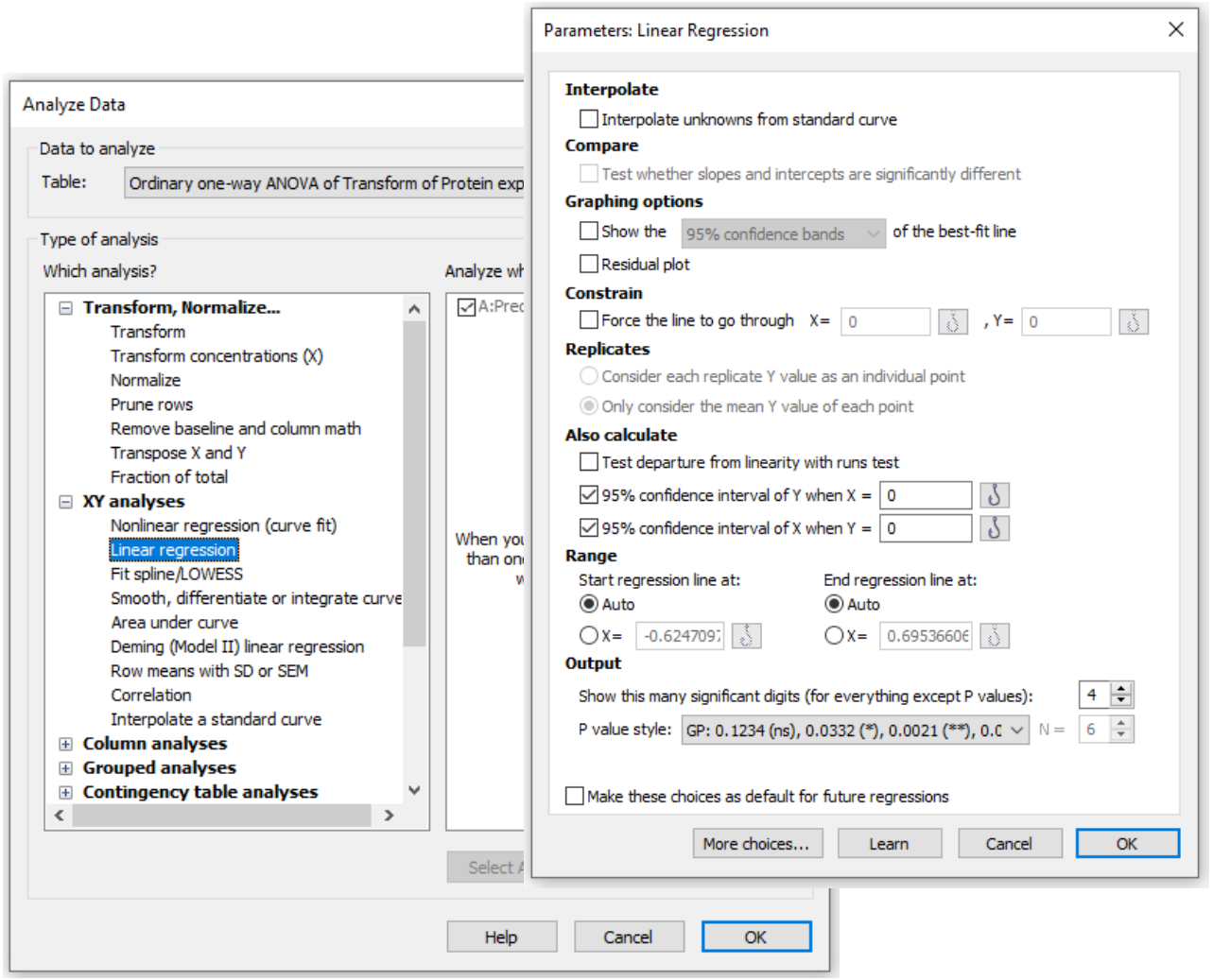
Correlation with Prism 8

Amount of light in a tree



Correlation with Prism 8

(if you also want a line of best fit)



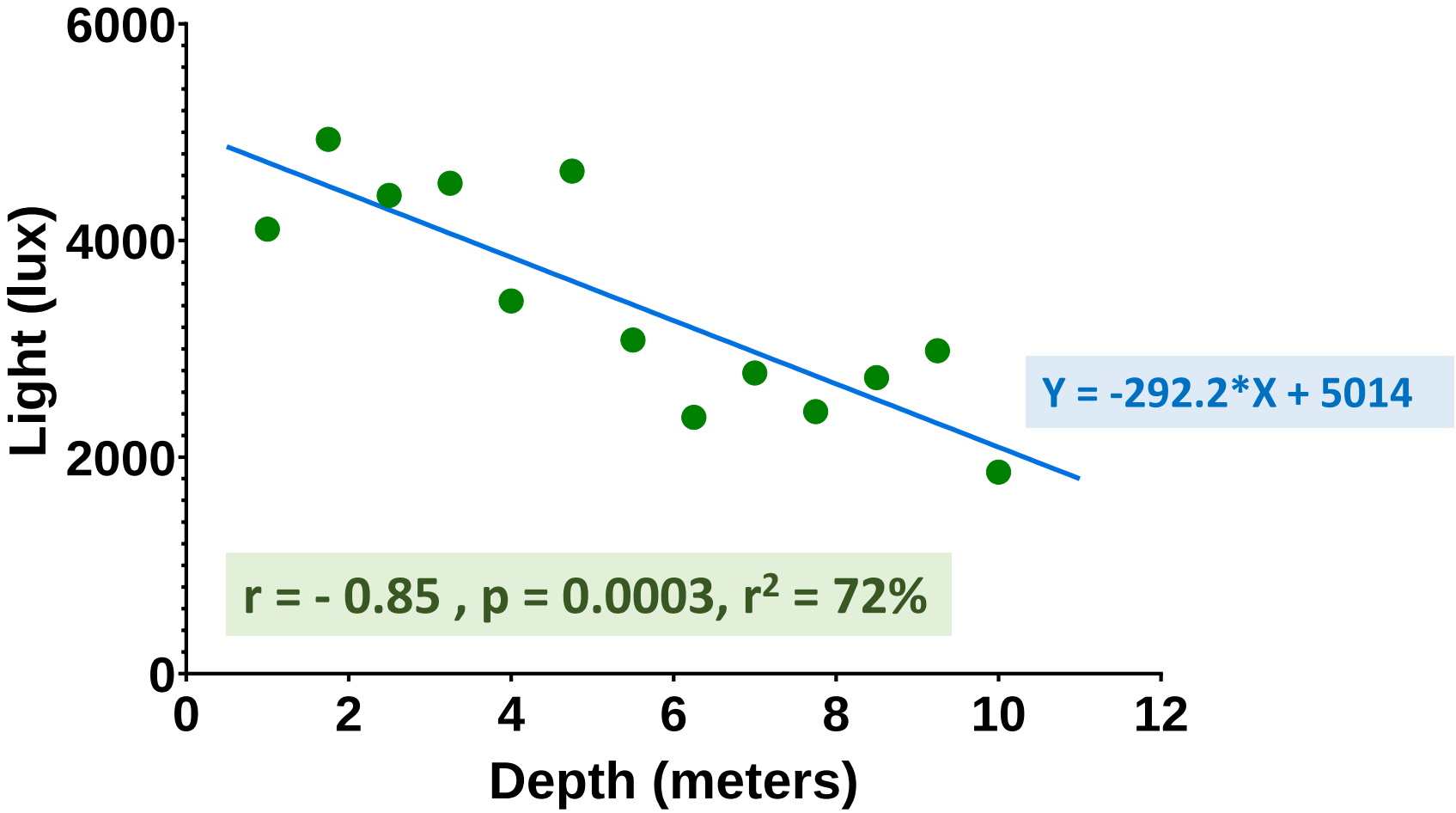
Results

	Linear reg. Tabular results	A Light
1	**Best-fit values**	
2	Slope	-292.2
3	Y-intercept	5014
4	X-intercept	17.16
5	1/slope	-0.003423
6		
7	**Std. Error**	
8	Slope	55.41
9	Y-intercept	342.2
10		
11	**95% Confidence Intervals**	
12	Slope	-414.1 to -170.2
13	Y-intercept	4261 to 5767
14	X-intercept	13.59 to 25.66
15		
16	**Goodness of Fit**	
17	R square	0.7165
18	Sy.x	560.7
19		
20	**Is slope significantly non-zero?**	
21	F	27.80
22	DFn, DFd	1, 11
23	P value	0.0003
24	Deviation from zero?	Significant
25		
26	**Equation**	Y = -292.2*X + 5014

Prediction

Correlation with Prism 8

Amount of light in a tree



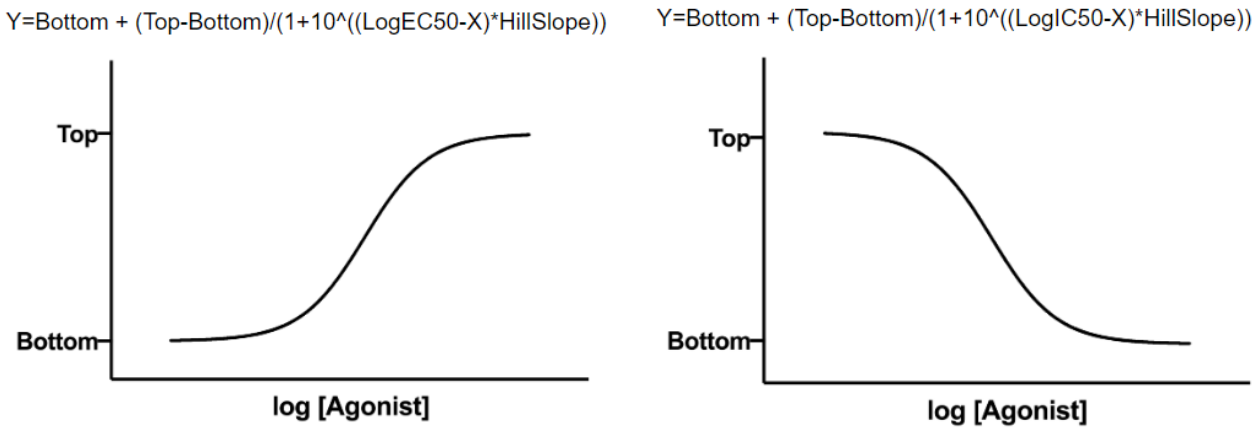
Association between 2 continuous variables

Non-linear relationship

Curve fitting

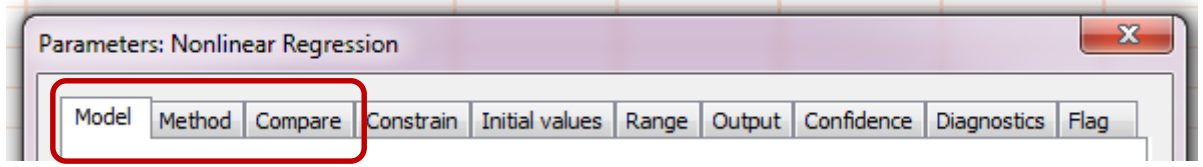
- **Dose-response curves**
  - Nonlinear regression
  - Dose-response experiments typically use around 5-10 doses of agonist, equally spaced on a logarithmic scale
  - Y values are responses

- The aim is often to determine the **IC50** or the **EC50**
  - **IC50 (I=Inhibition)**: concentration of an agonist that provokes a response half way between the maximal (Top) response and the maximally inhibited (Bottom) response.
  - **EC50 (E=Effective):** concentration that gives half-maximal response



Curve fitting

Example: Inhibition data.xlsx

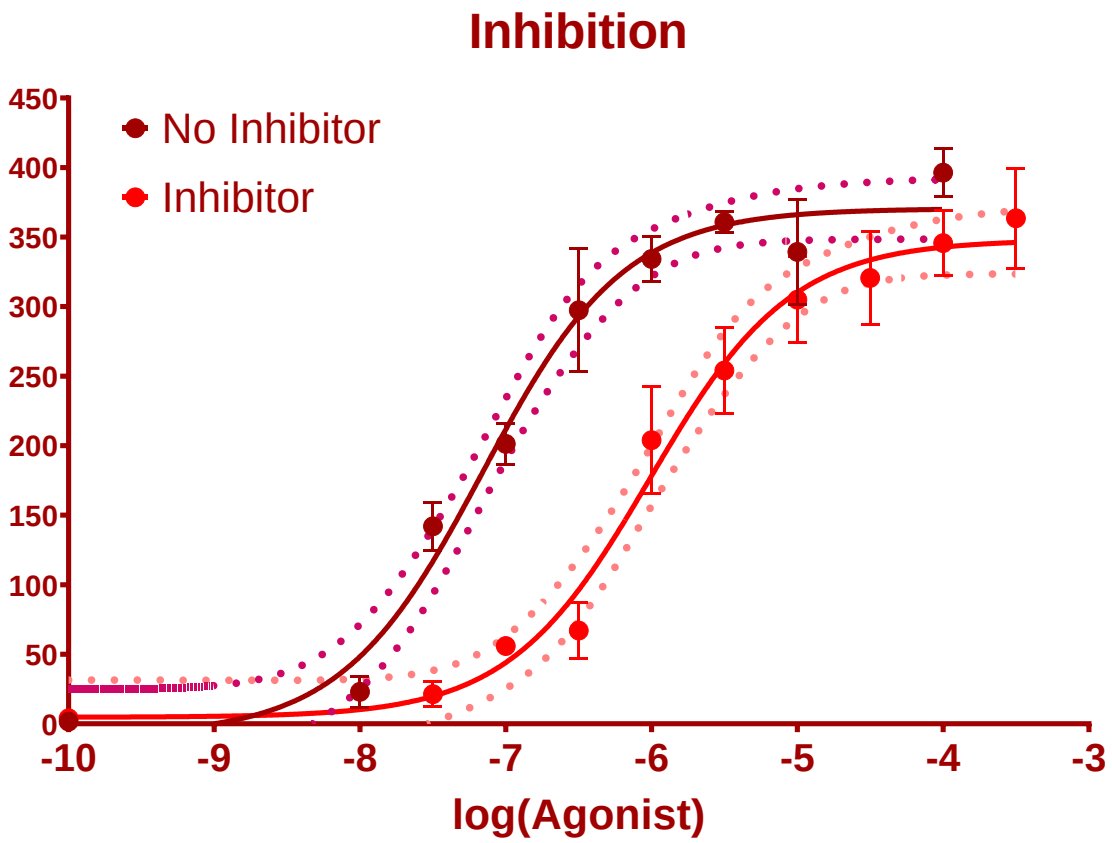


Step by step analysis and considerations:

1- Choose a **Model**:

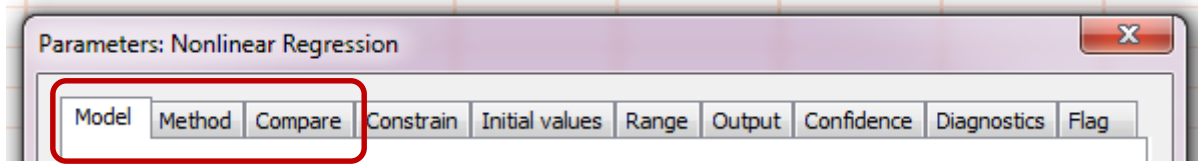
2- Choose a **Method**

3- **Compare** different conditions



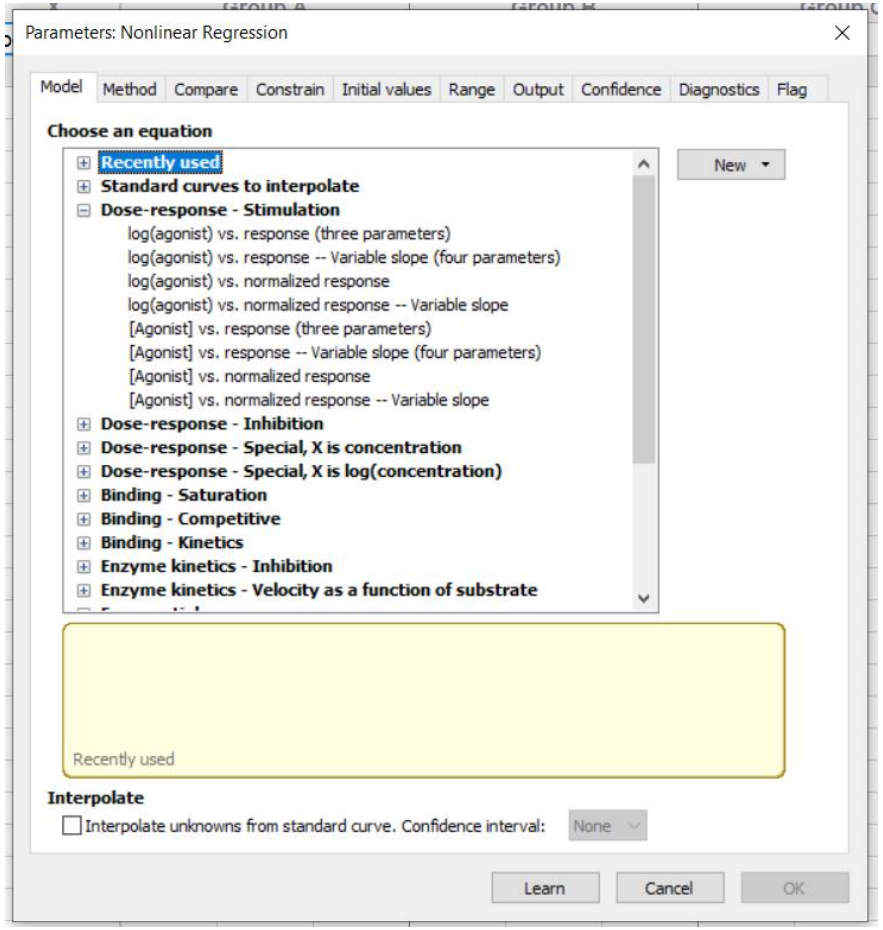
Curve fitting

Example: Inhibition data.xlsx



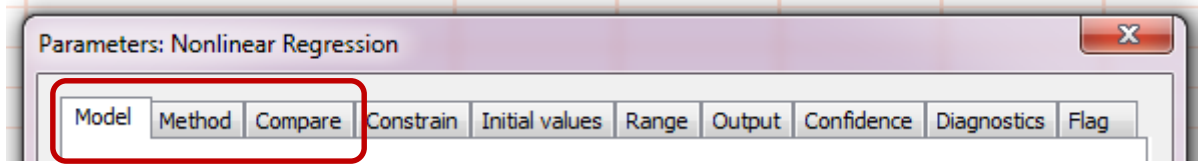
Step by step analysis and considerations:

1- Choose a **Model**



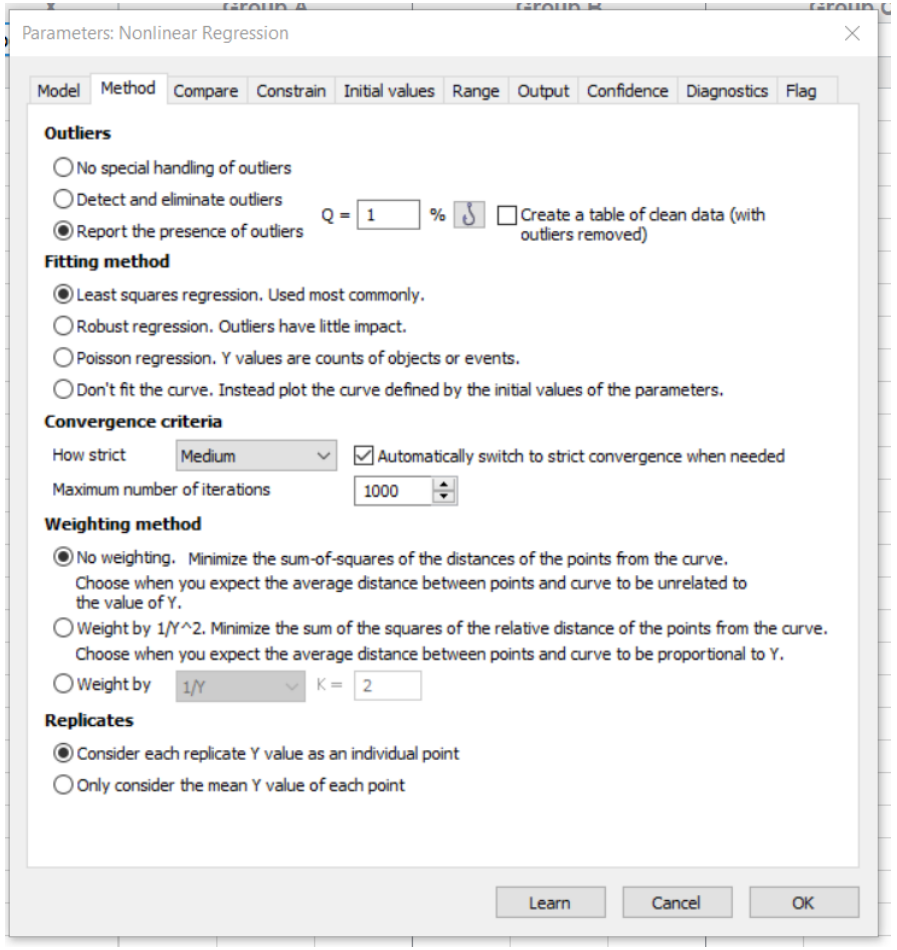
Curve fitting

Example: Inhibition data.xlsx



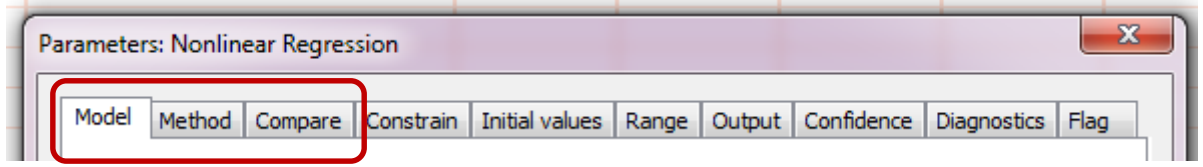
Step by step analysis and considerations:

2- Choose a **Method:**



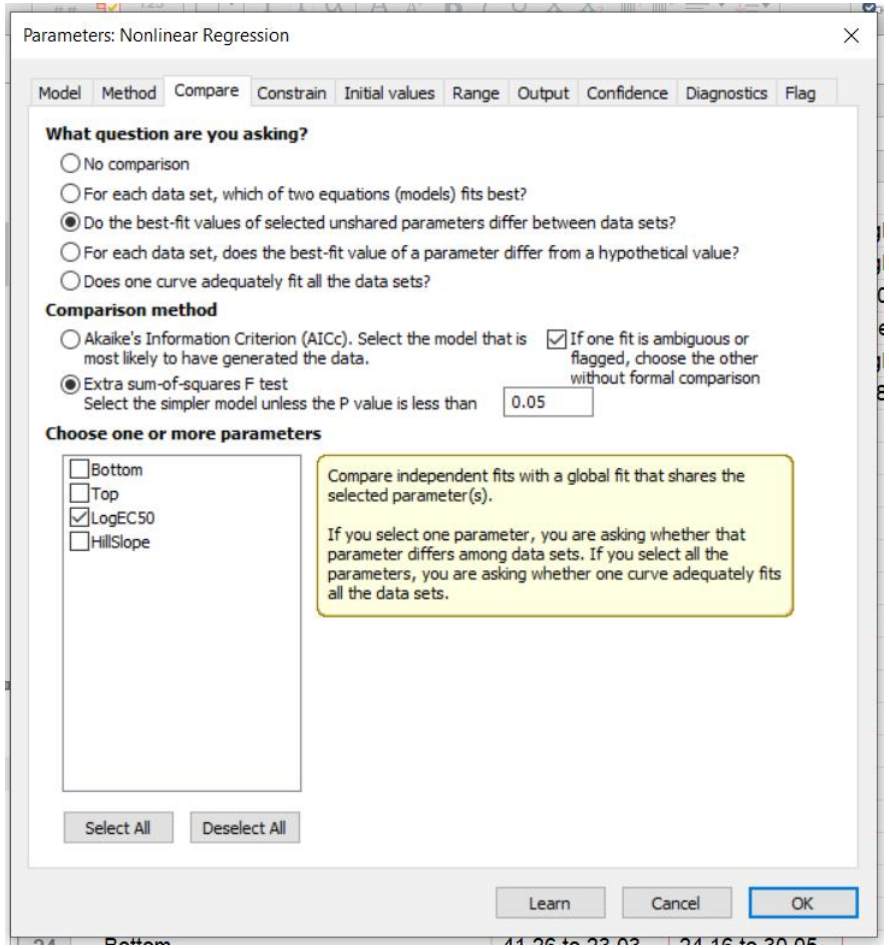
Curve fitting

Example: Inhibition data.xlsx



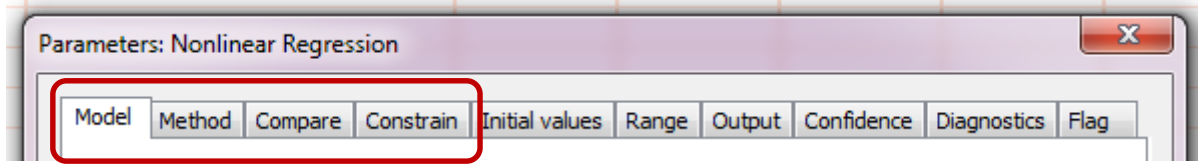
Step by step analysis and considerations:

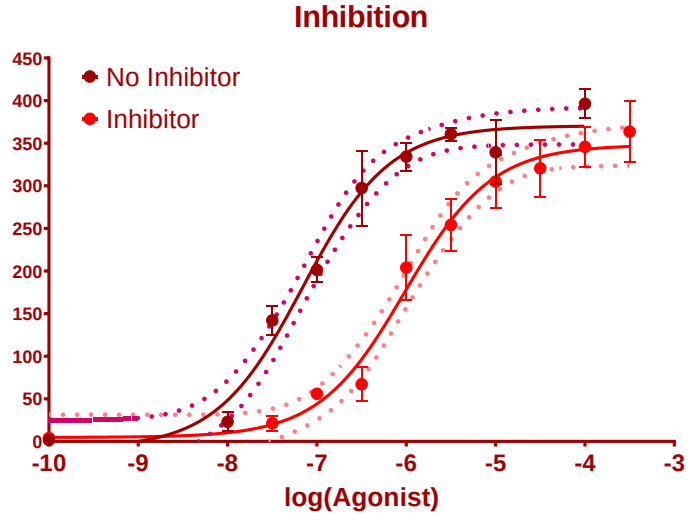
3- **Compare** different conditions:



Curve fitting

Example: Inhibition data.xlsx





Step by step analysis and considerations:

1- Choose a **Model**:

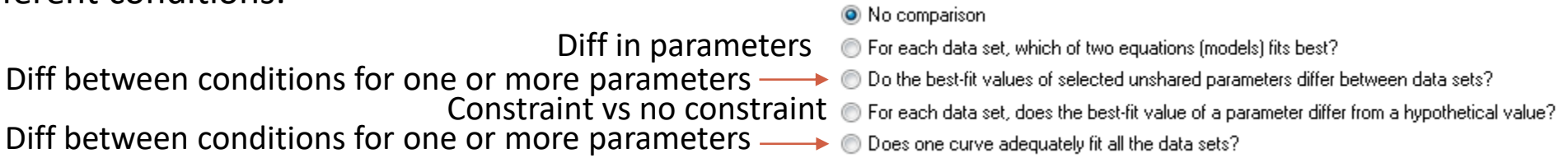
not necessary to normalise

should choose it when values defining 0 and 100 are precise

variable slope better if plenty of data points (variable slope or 4 parameters)

2- Choose a **Method:** outliers, fitting method, weighting method and replicates

3- **Compare** different conditions:



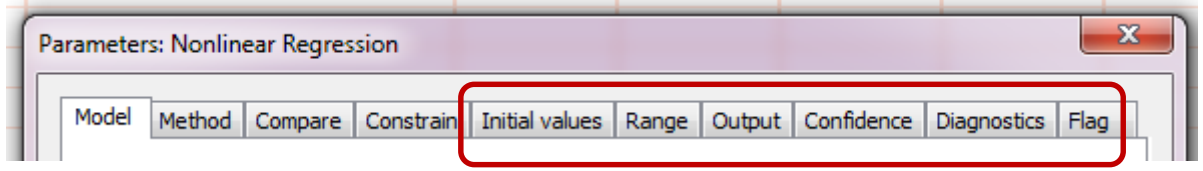
4- **Constrain**:

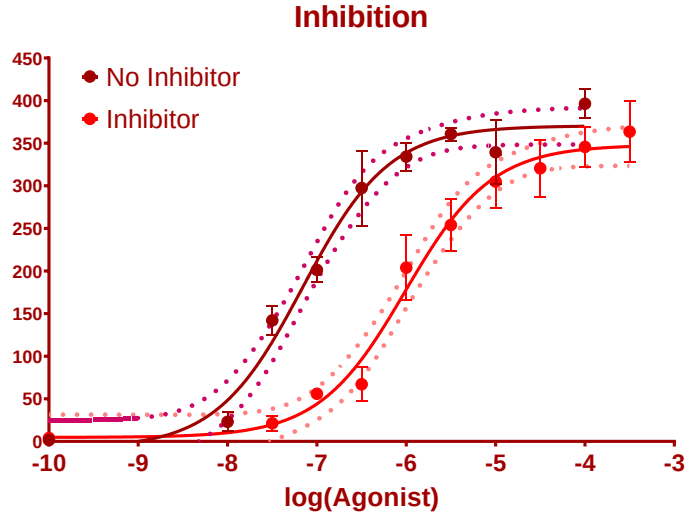
depends on your experiment

depends if your data don't define the top or the bottom of the curve

Curve fitting

Example: Inhibition data.xlsx





Step by step analysis and considerations:

5- **Initial values**:

defaults usually OK unless the fit looks funny

6- **Range**:

defaults usually OK unless you are not interested in the x-variable full range (ie time)

7- **Output**:

summary table presents same results in a … summarized way.

8 – **Confidence**: calculate and plot confidence intervals

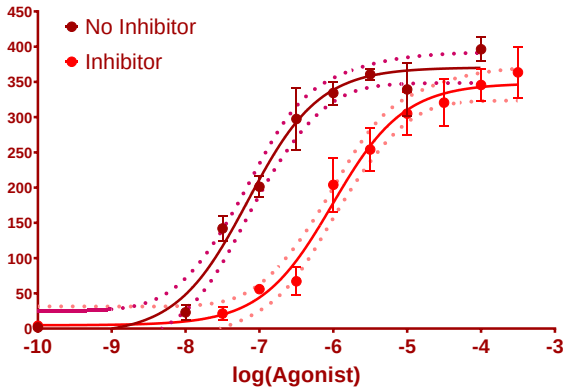
9- **Diagnostics**:

check for normality (weights) and outliers (but keep them in the analysis)
check Replicates test
residual plots

Curve fitting

Inhibition 4 parameters

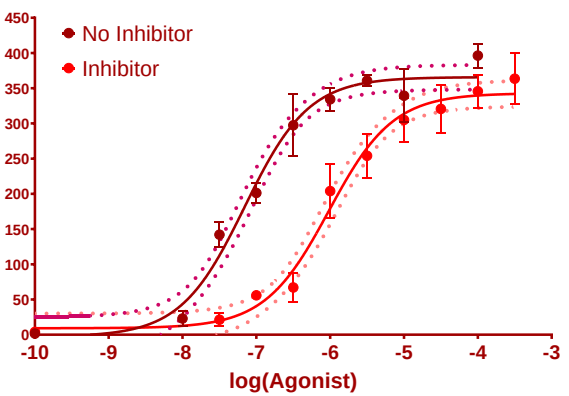
LogEC50 same for all data sets
LogEC50 different for each data set
< 0.0001
Reject null hypothesis
LogEC50 different for each data set
64.86 (1,48)



95% Confidence Intervals		
Bottom	-41.39 to 24.94	-22.15 to 31.56
Top	348.3 to 392.6	323.1 to 373.0
LogEC50	-7.324 to -6.991	-6.185 to -5.837
HillSlope	0.6347 to 1.159	0.6095 to 1.186
EC50	4.739e-008 to 1.020e-007	6.538e-007 to 1.455e-006
R square	0.9663	0.9653

Inhibition 3 parameters

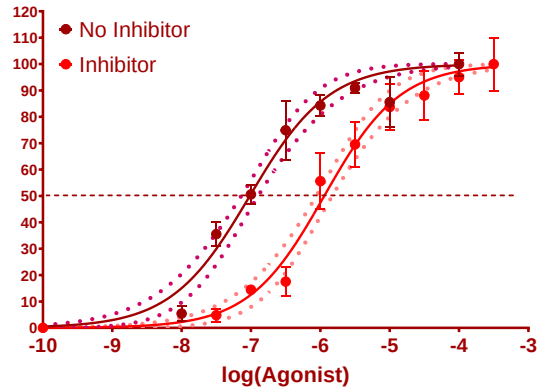
LogEC50 same for all data sets
LogEC50 different for each data set
< 0.0001
Reject null hypothesis
LogEC50 different for each data set
101.0 (1,50)



95% Confidence Intervals		
Bottom	-30.74 to 24.78	-11.65 to 30.07
Top	348.2 to 383.2	324.3 to 361.4
LogEC50	-7.312 to -7.006	-6.175 to -5.859
EC50	4.875e-008 to 9.858e-008	6.677e-007 to 1.385e-006
R square	0.9655	0.9648

Normalize 4 parameters

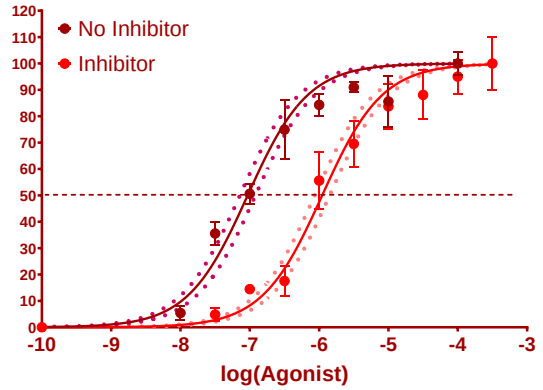
LogEC50 same for all data sets
LogEC50 different for each data set
< 0.0001
Reject null hypothesis
LogEC50 different for each data set
162.8 (1,52)



95% Confidence Intervals		
LogEC50	-7.137 to -6.897	-6.057 to -5.830
HillSlope	0.6094 to 0.9184	0.6467 to 0.9460
EC50	7.295e-008 to 1.268e-007	8.763e-007 to 1.481e-006
R square	0.9580	0.9635

Normalize 3 parameters

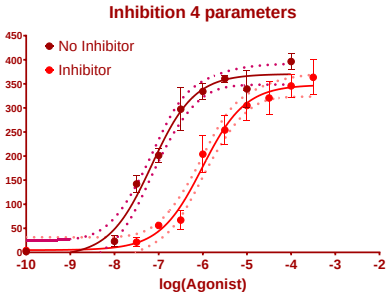
One curve for all data sets
Different curve for each data set
< 0.0001
Reject null hypothesis
Different curve for each data set
175.0 (1,54)



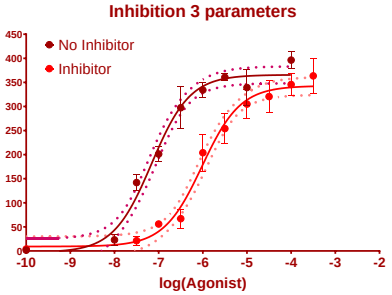
95% Confidence Intervals		
LogEC50	-7.144 to -6.917	-6.064 to -5.848
EC50	7.179e-008 to 1.209e-007	8.633e-007 to 1.420e-006
R square	0.9476	0.9568

Curve fitting

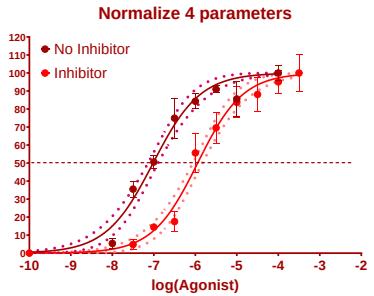
	No inhibitor	Inhibitor
Replicates test for lack of fit		
SD replicates	22.71	25.52
SD lack of fit	41.84	32.38
Discrepancy (F)	3.393	1.610
P value	0.0247	0.1989
Evidence of inadequate model?	Yes	No



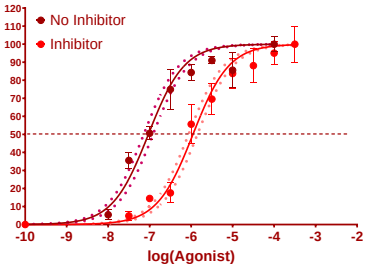
Replicates test for lack of fit		
SD replicates	22.71	25.52
SD lack of fit	39.22	30.61
Discrepancy (F)	2.982	1.438
P value	0.0334	0.2478
Evidence of inadequate model?	Yes	No



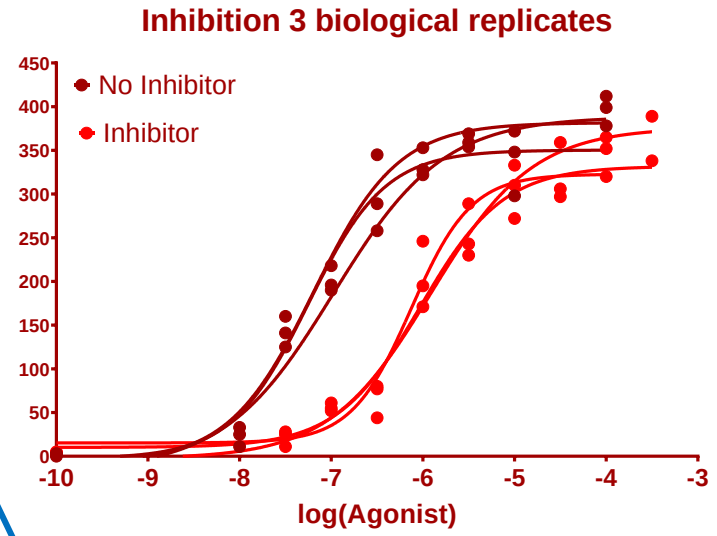
Replicates test for lack of fit		
SD replicates	5.755	7.100
SD lack of fit	11.00	8.379
Discrepancy (F)	3.656	1.393
P value	0.0125	0.2618
Evidence of inadequate model?	Yes	No

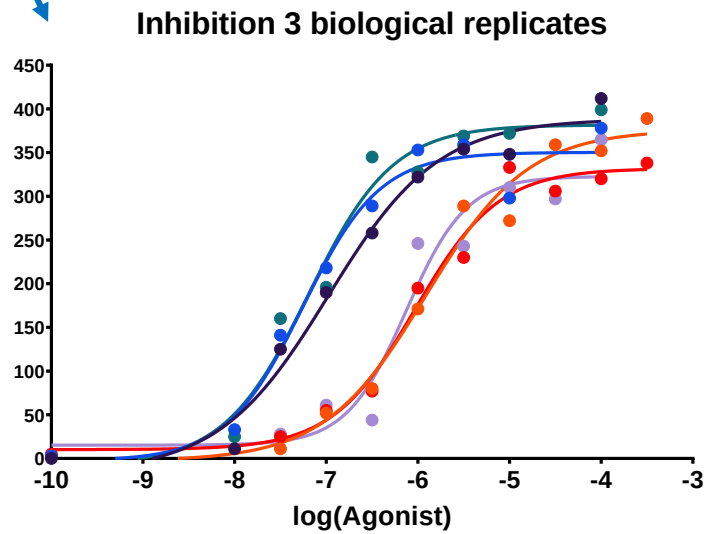


Replicates test for lack of fit		
SD replicates	5.755	7.100
SD lack of fit	12.28	9.649
Discrepancy (F)	4.553	1.847
	0.0036	0.1246
l?	Yes	No



Have a go!





Association between 2 continuous variables

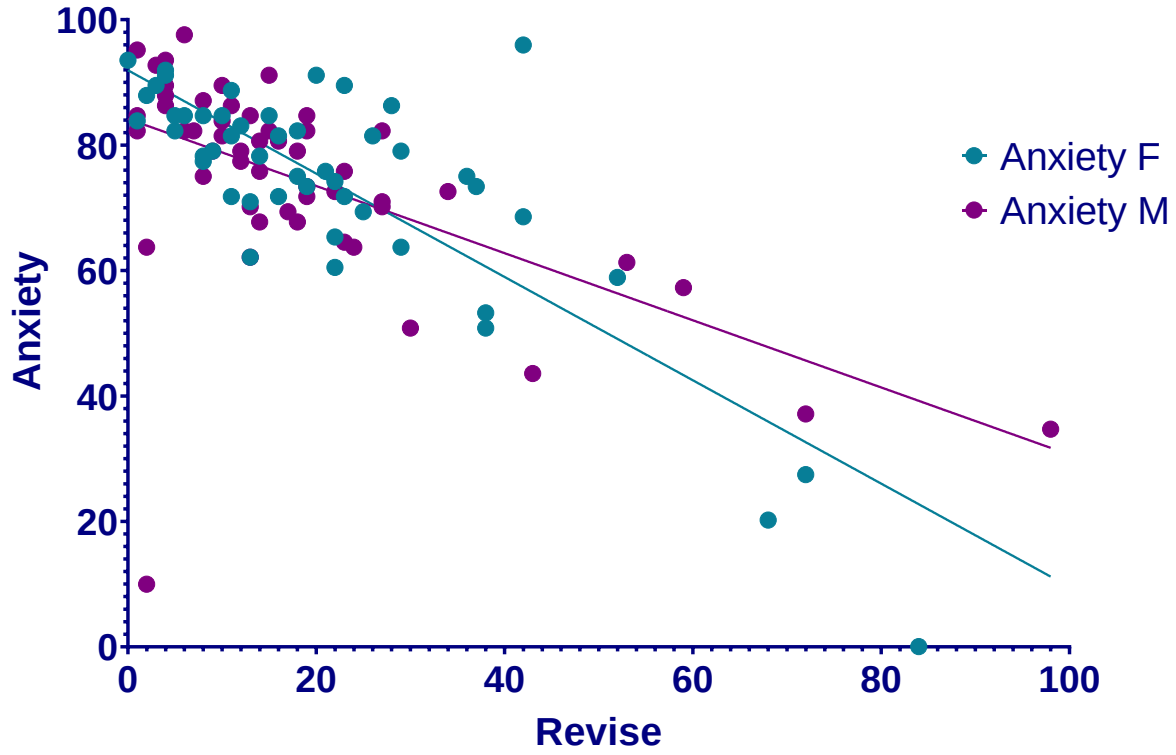
Linear relationship

Diagnostics: Goodness-of-fit

Regression: Goodness of fit

• **Question**: Is there a relationship between time spent revising (Revise) and exam anxiety (Anxiety)? And, if yes, are boys and girls different?

• **Focus**: how good is the model?



Regression: Goodness of fit

Set of 5 clues

Clue 1: Graphical exploration of the data: linearity

Clue 2: Identification of outliers

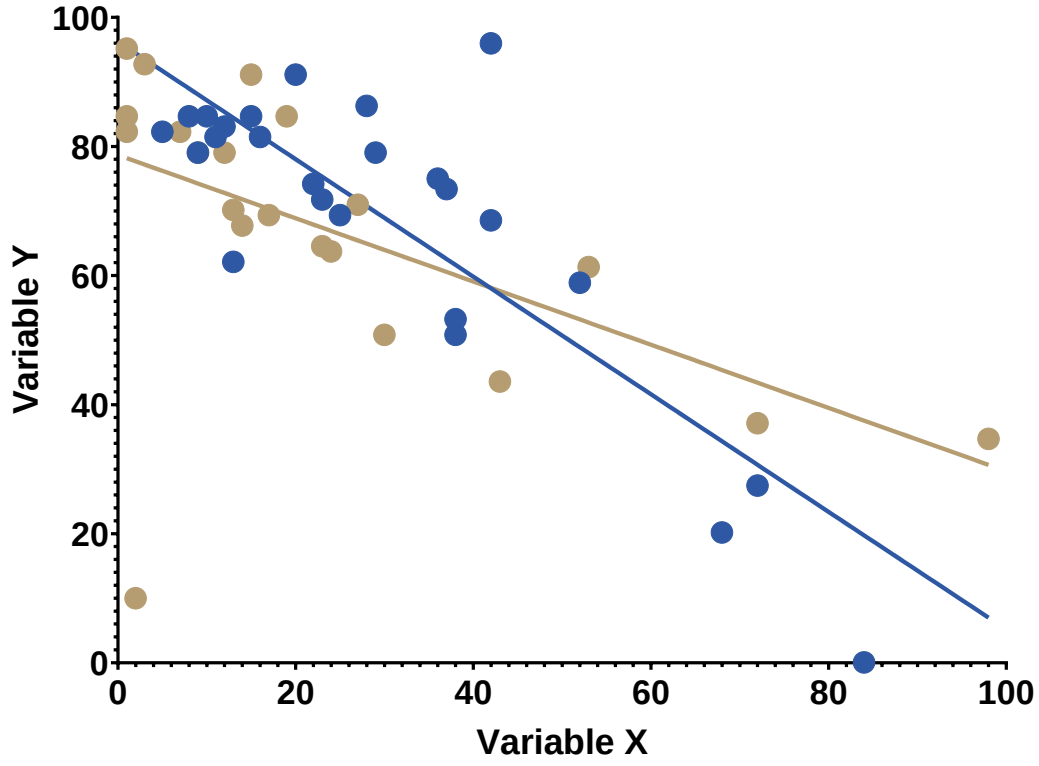
Clue 3: R^2, the Coefficient of determination

Clue 4: Distribution of residuals with statistical tests

Clue 5: Distribution of residuals with a QQ plot

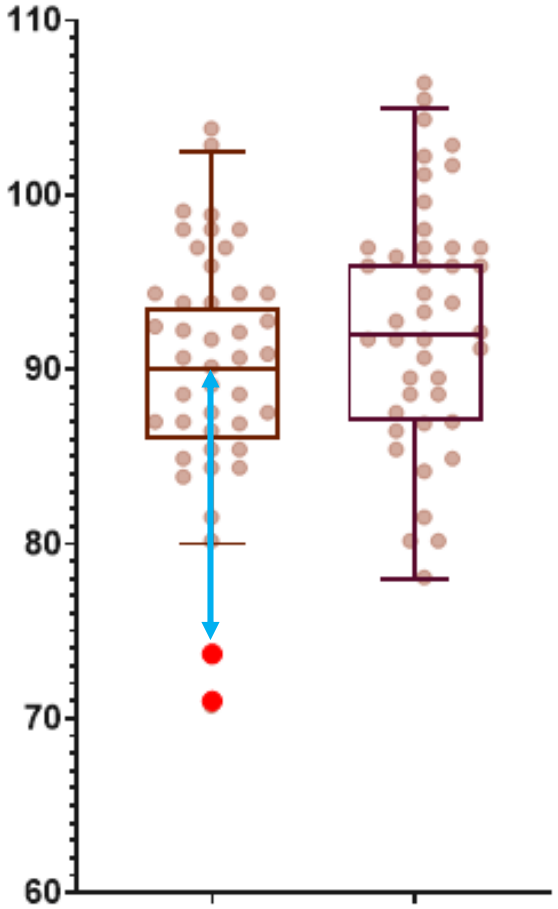
Regression: Goodness of fit

- **Clue 1**: Graphical exploration of the data

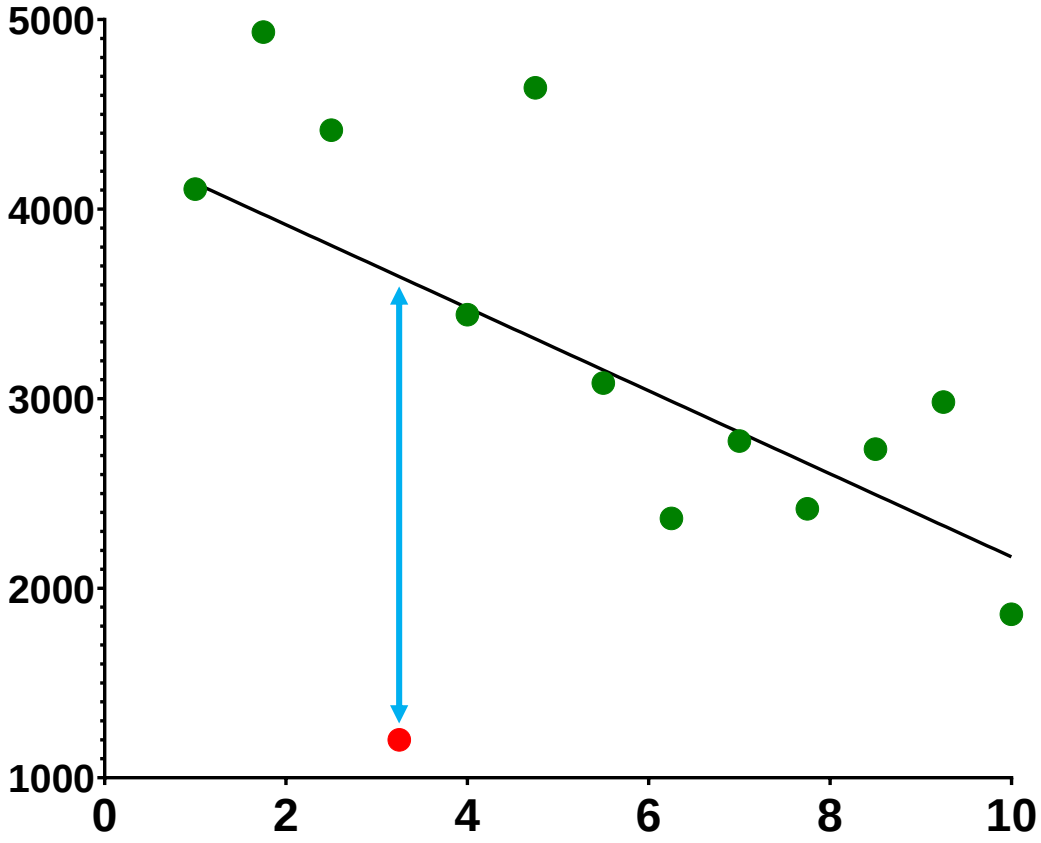


Regression: Goodness of fit

- **Clue 2**: Identification of outliers



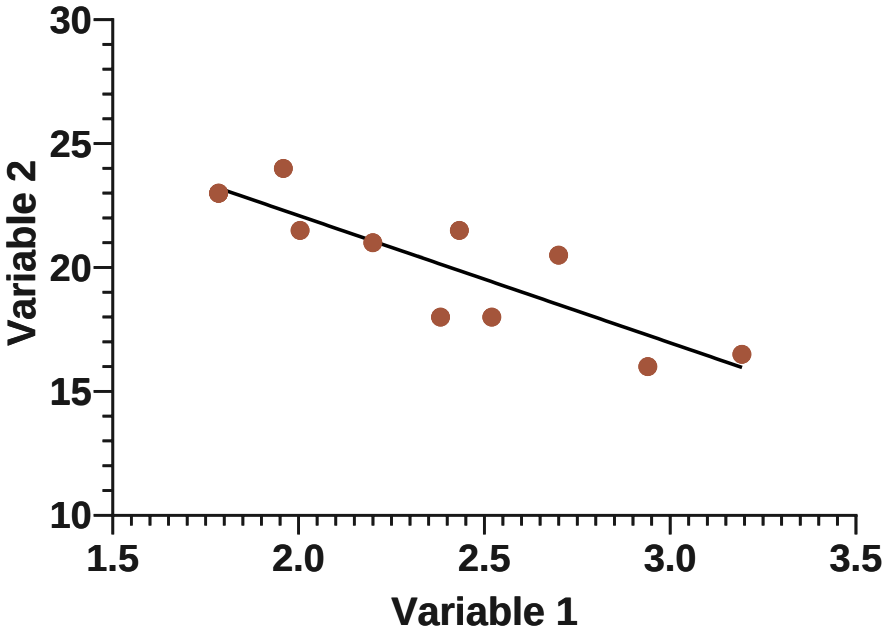
Categorical x and continuous y



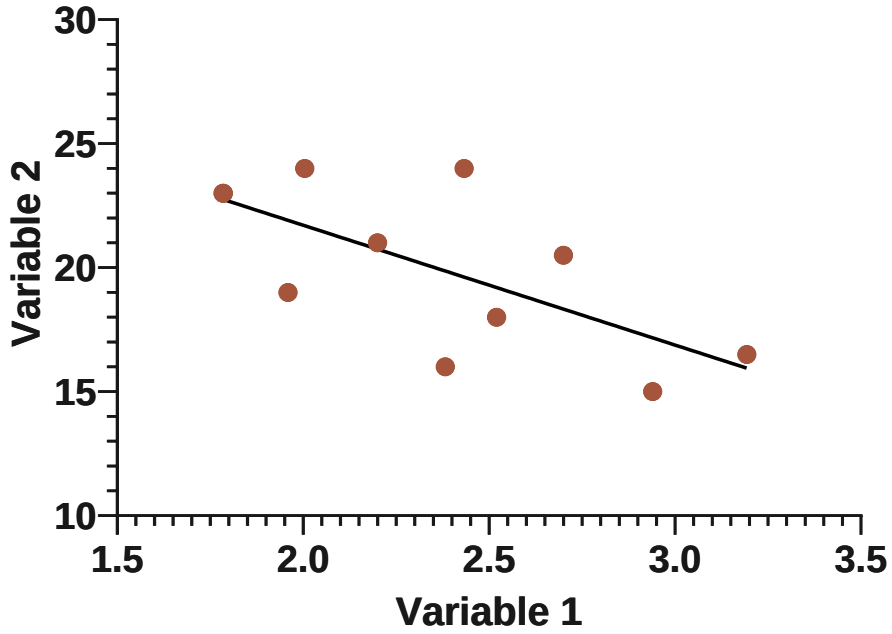
Continuous x and continuous y

Regression: Goodness of fit

- **Clue 3**: Coefficient of determination



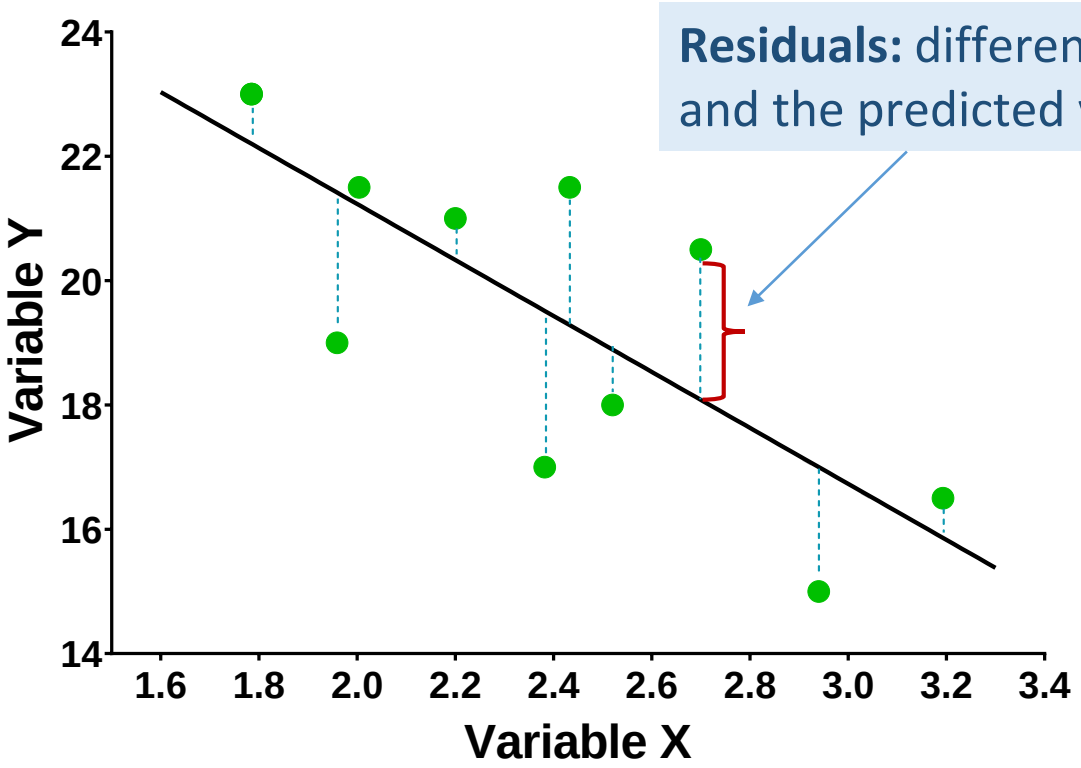
$R^2 = 76\%$



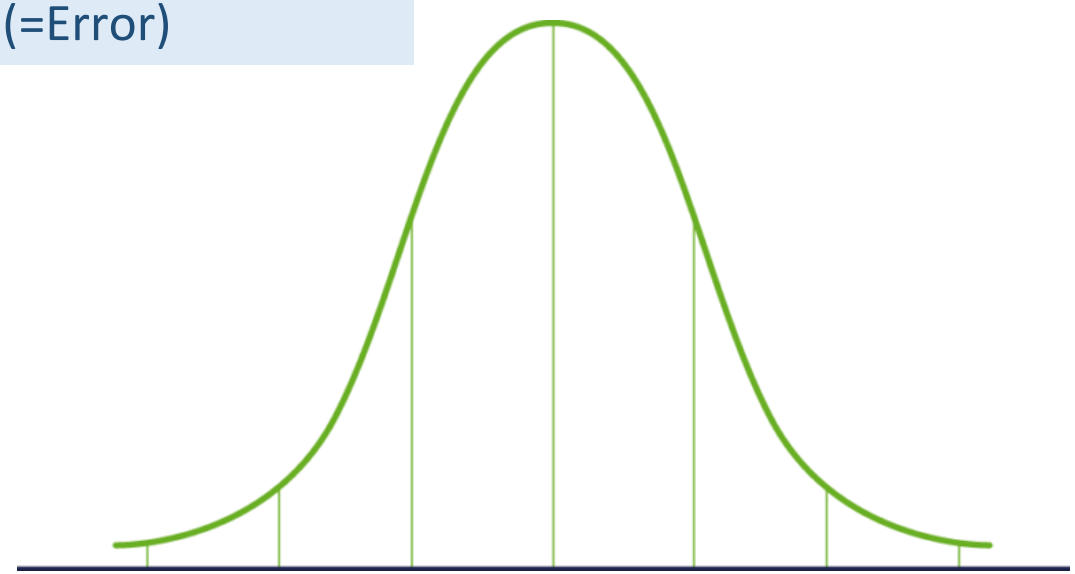
$R^2 = 46\%$

Regression: Goodness of fit

- **Distribution of the residuals**



Residuals

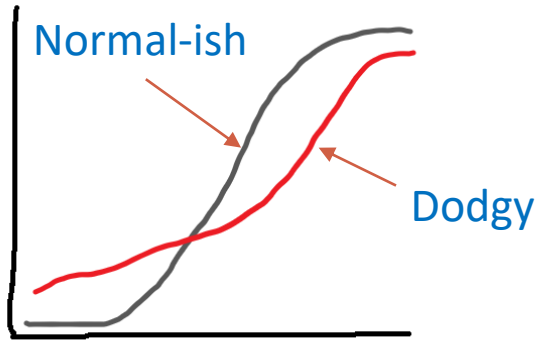


Distribution of Residuals

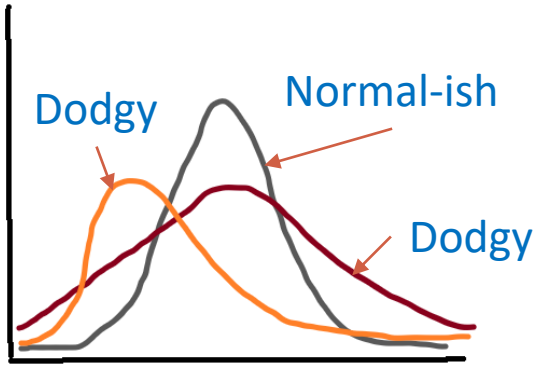
Regression: model goodness of fit

- **Clue 4**: Distribution of the residuals **with statistical tests**

- Statistical tests: **significant departure from normality ?**
  - **Anderson-Darling**:
    - cumulative distribution different from a normal one?



  - **D'Agostino-Pearson**:
    - asymmetry and shape different from normal distribution?

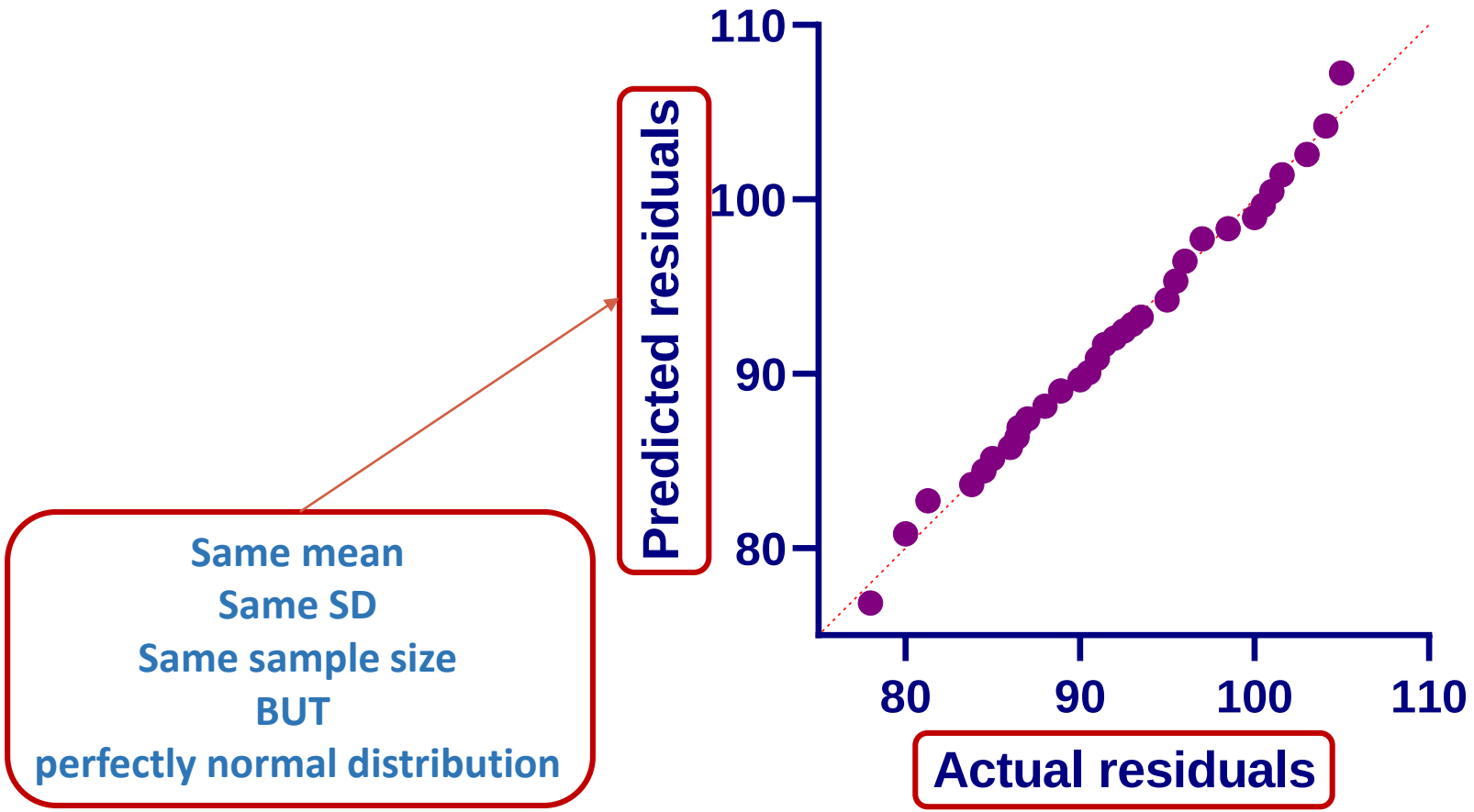


  - Other tests: **Shapiro-Wilk** and **Kolmogorov-Smirnov**

Regression: Goodness of fit

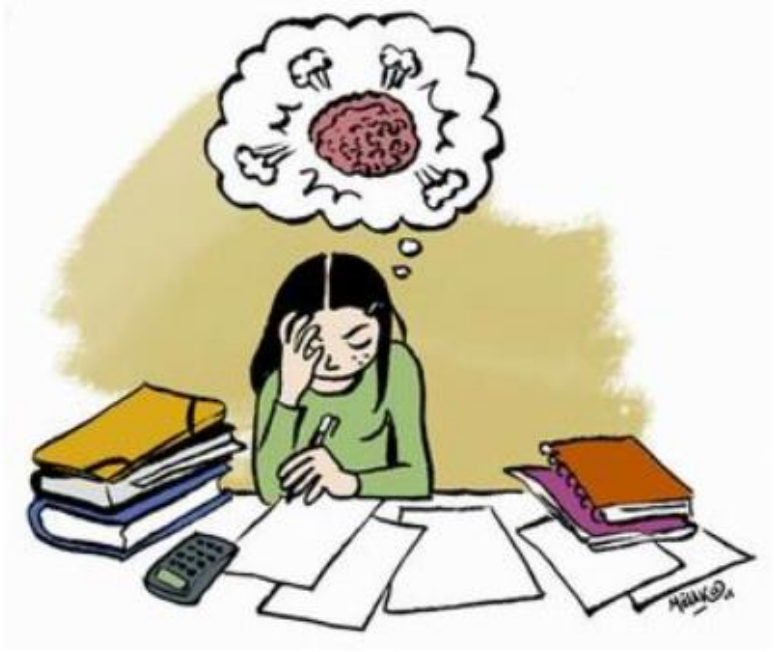
- **Clue 5**: Distribution of the residuals **with QQ plot**

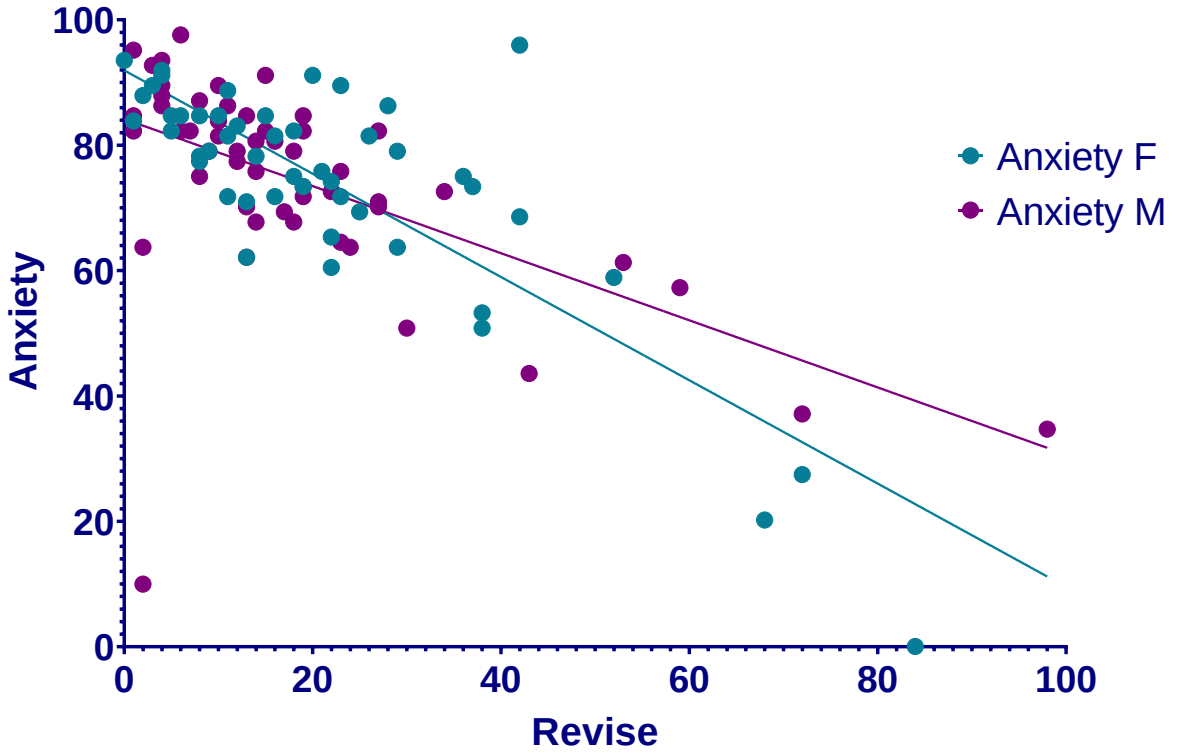
Normal QQ plot = Quantile – Quantile plot



Goodness of fit: let's do it

exam anxiety.xlsx (Don't need 'Exam')





Association between time spent revising and exam anxiety.

Are we getting it right?

Goodness of fit: let's do it

exam anxiety.xlsx (Don't need 'Exam')

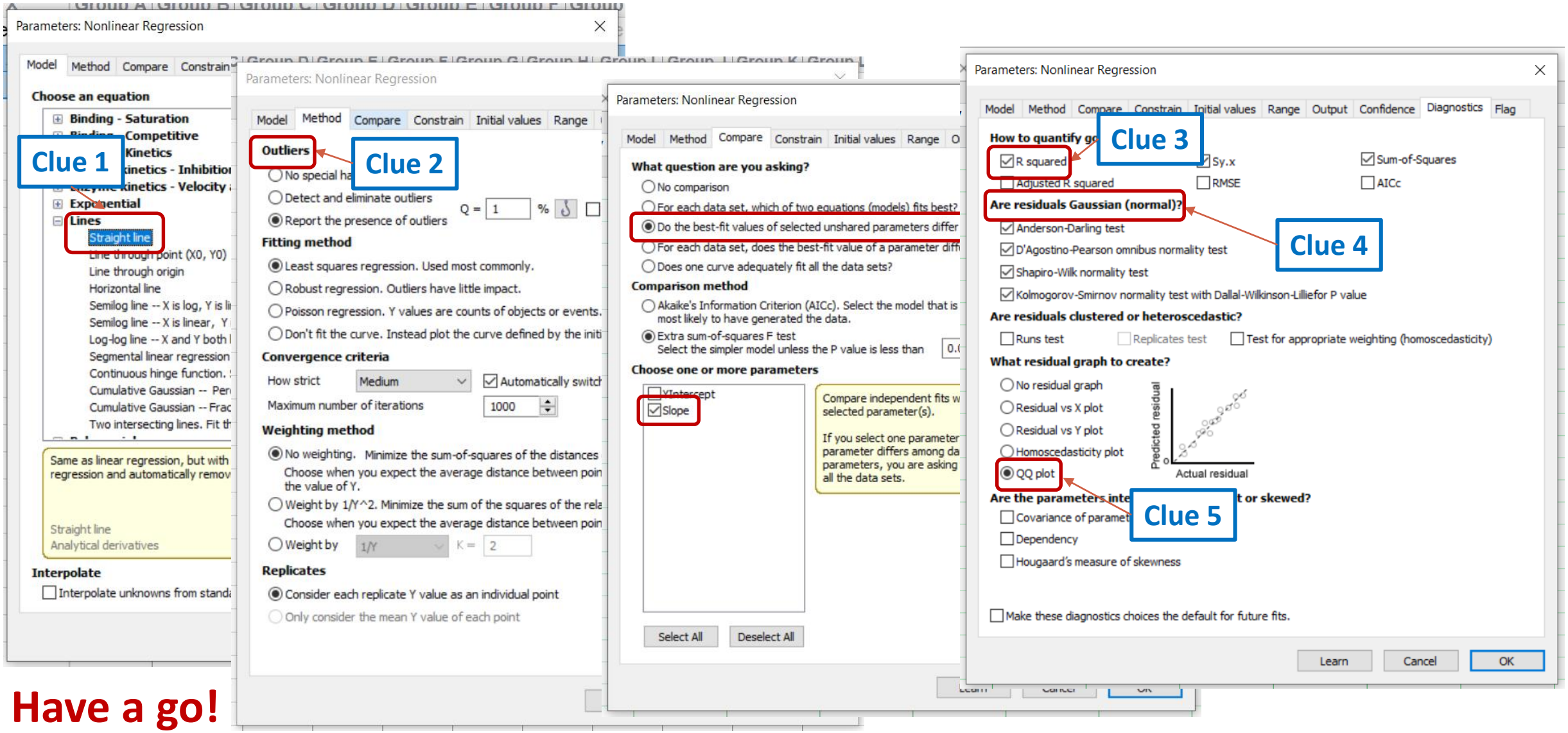
Excel File

	A	B	C	D	E
1	Code	Revise	Exam	Anxiety	Gender
2	2	11	65	88.716	Female
3	6	22	70	60.506	Female
4	7	16	20	81.462	Female
5	8	21	55	75.82	Female
6	9	25	50	69.372	Female
7	10	18	40	82.268	Female
8	14	18	50	75.014	Female
9	18	29	95	79.044	Female
10	19	4	50	91.134	Female
11	23	22	85	65.342	Female
12	24	84	90	0.056	Female
13	25	23	30	71.79	Female
14	26	26	60	81.462	Female
15	28	72	75	27.46	Female
16	29	37	27	73.402	Female
17	31	3	75	89.522	Female
18	32	36	90	75.014	Female
19	36	9	10	79.044	Female
20	39	12	5	83.074	Female
21	42	8	45	78.238	Female
22	44	22	70	74.208	Female
23	45	21	50	75.82	Female
24	50	19	50	73.402	Female
25	51	0	35	93.552	Female
26	52	52	80	58.894	Female
27	53	38	50	53.252	Female
28	55	23	75	89.522	Female
29	56	11	25	71.79	Female
30	60	42	70	68.566	Female

Prism File

Table format: XY	X	Group A	Group B
	Revise	Anxiety F	Anxiety M
	X	Y	Y
31	64	11	81.462	
32	67	4	91.940	
33	68	28	86.298	
34	70	29	63.730	
35	72	16	71.790	
36	74	10	84.686	
37	76	8	77.432	
38	77	5	82.268	
39	79	38	50.834	
40	82	6	84.686	
41	83	68	20.206	
42	85	1	83.880	
43	87	42	95.970	
44	88	13	62.118	
45	91	5	84.686	
46	92	12	83.074	
47	94	2	87.910	
48	97	15	84.686	
49	99	13	70.984	
50	100	14	78.238	
51	103	20	91.134	
52	1	4		86.298
53	3	27		70.178
54	4	53		61.312
55	5	4		89.522
56	11	18		79.044
57	12	16		80.656
58	13	13		70.178
59	15	98		34.714
60	16	1		95.164
61	17	14		75.820
62	20	23		64.536
63	21	14		80.656
64	22	12		77.432

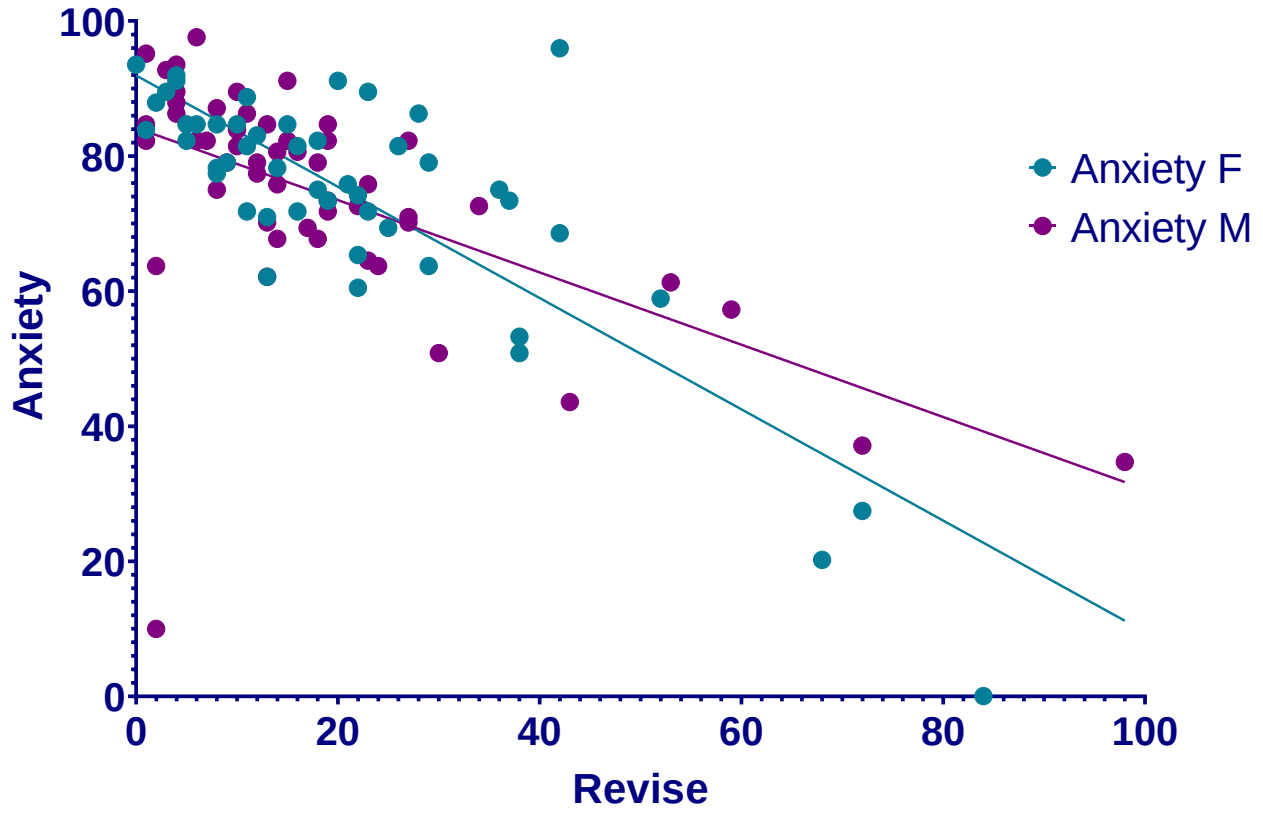
Goodness of fit with Prism 8



Have a go!

Regression: model goodness of fit

- **Clue 1**: Graphical exploration of the data: linear relationship



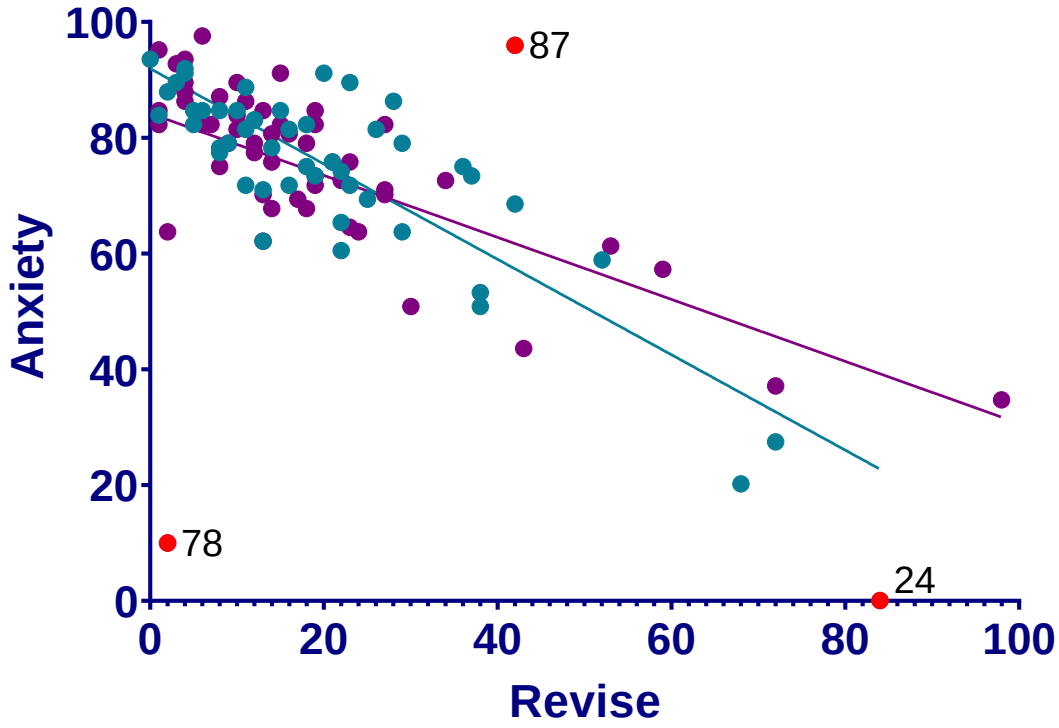
Goodness of fit with Prism 8

- **Clue 2**: Identification of outliers

Table of results | Outliers

		X	A	B
		Revise	Anxiety F	Anxiety M
		X		
1	24	84.000	0.056	
2	87	42.000	95.970	
3	78	2.000		10.000

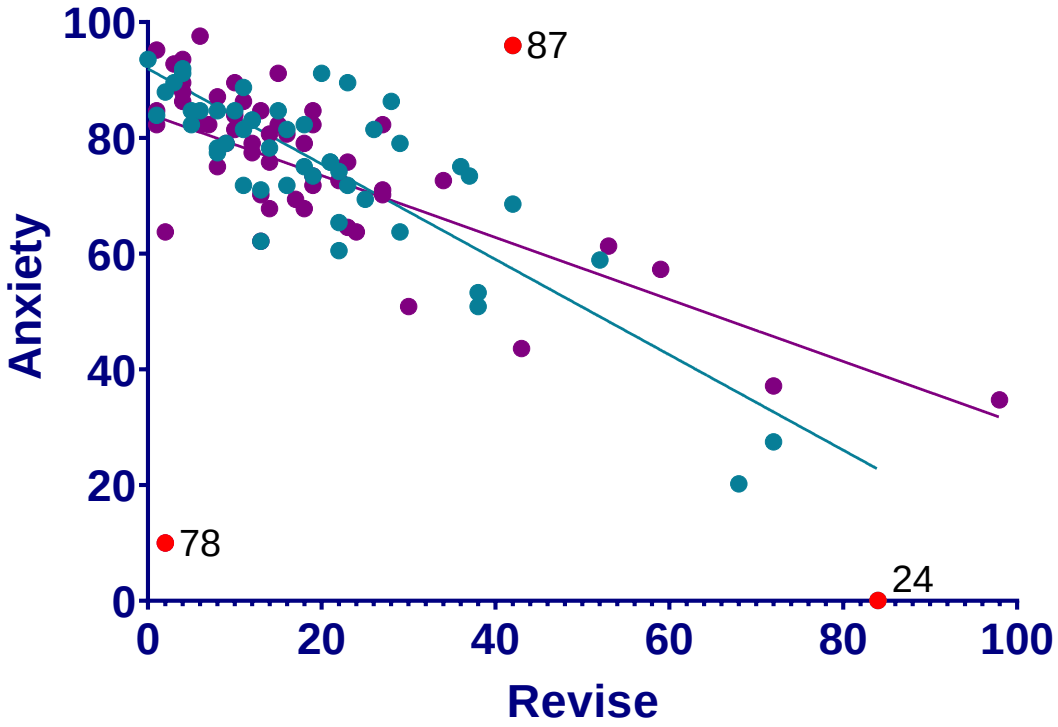
ID



Regression: model goodness of fit

- **Clue 3**: Coefficient of determination

Nonlin fit Table of results	A Anxiety F	B Anxiety M	C Global (shared)
Comparison of Fits			
Null hypothesis			Slope same for all data sets
Alternative hypothesis			Slope different for each data set
P value			0.0299
Conclusion (alpha = 0.05)			Reject null hypothesis
Preferred model			Slope different for each data set
F (DFn, DFd)			4.852 (1, 99)
Slope different for each data set			
Best-fit values			
YIntercept	91.94	84.19	
Slope	-0.8238	-0.5353	
95% CI (profile likelihood)			
YIntercept	87.36 to 96.52	78.93 to 89.46	
Slope	-0.9880 to -0.6596	-0.7394 to -0.3312	
Goodness of Fit			
Degrees of Freedom	49	50	
R squared	0.6746	0.3568	
Sum of Squares	5322	8845	
Sy.x	10.42	13.30	

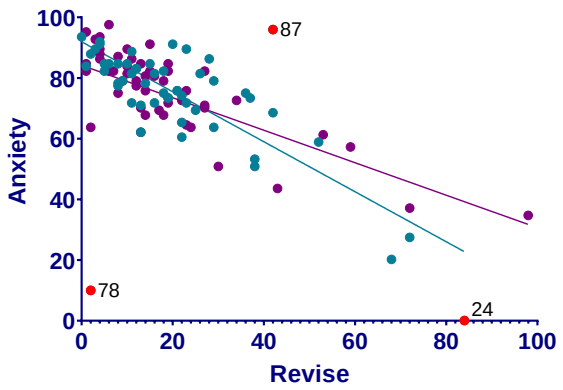


Regression: model goodness of fit

- **Clue 4**: Distribution of the residuals **with statistical tests**

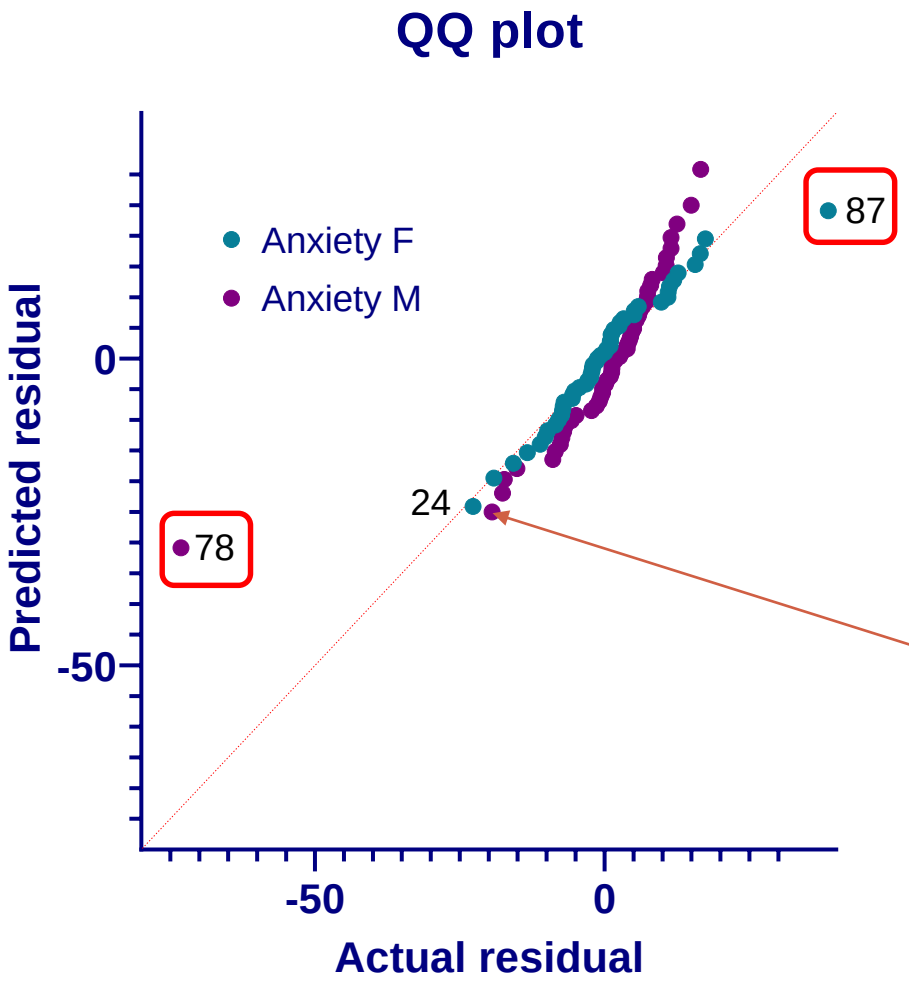
Significant departure from normality ☹

		Females	Males
1	**Normality of Residuals**		
2	**Anderson-Darling (A2*)**	0.7493	3.041
3	P value	0.0478	<0.0001
4	Passed normality test (alpha=0.05)?	No	No
5	P value summary	*	****
6	**D'Agostino-Pearson omnibus (K2)**	14.43	68.42
7	P value	0.0007	<0.0001
8	Passed normality test (alpha=0.05)?	No	No
9	P value summary	***	****
0	**Shapiro-Wilk (W)**	0.9429	0.6997
1	P value	0.0161	<0.0001
2	Passed normality test (alpha=0.05)?	No	No
3	P value summary	*	****
4	**Kolmogorov-Smirnov (distance)**	0.1243	0.1887
5	P value	0.0475	<0.0001
6	Passed normality test (alpha=0.05)?	No	No
7	P value summary	*	****
8			



Regression: model goodness of fit

- **Clue 5**: Distribution of the residuals **with QQ plot**



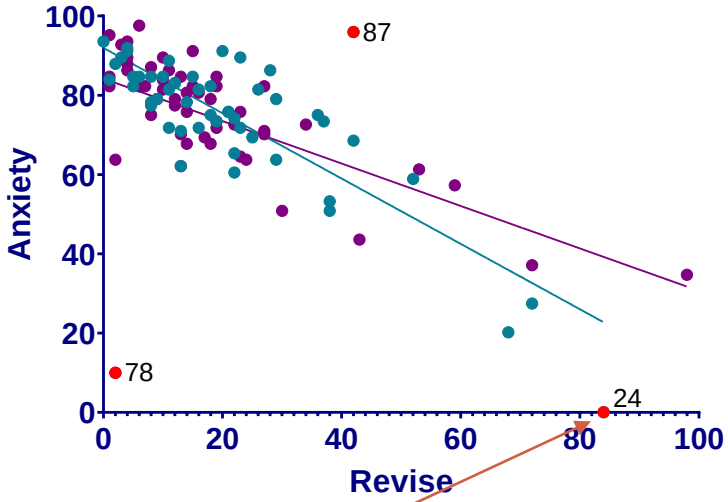
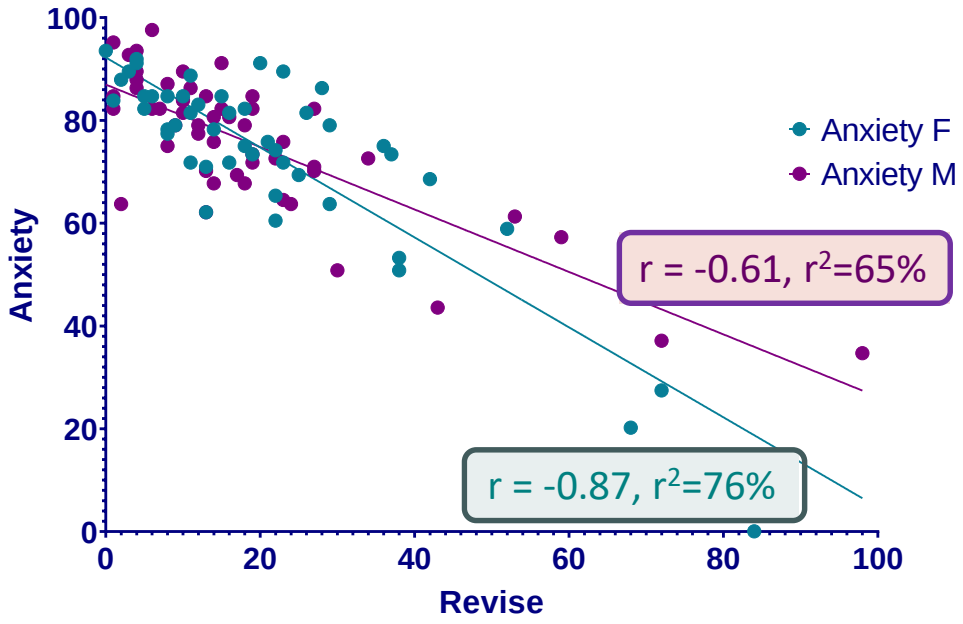


Table of results | Outliers

		X	A	B
		Revise	Anxiety F	Anxiety M
		X		
1	24	84.000	0.056	
2	87	42.000	95.970	
3	78	2.000		10.000

Goodness of fit with Prism 8

	Nonlin fit Table of results	A Anxiety F	B Anxiety M	C Global (shared)
1	**Comparison of Fits**			
2	Null hypothesis			Slope same for all data sets
3	Alternative hypothesis			Slope different for each data set
4	P value			0.0056
5	Conclusion (alpha = 0.05)			Reject null hypothesis
6	Preferred model			Slope different for each data set
7	F (DFn, DFd)			8.022 (1, 97)
8				
9	**Slope different for each data set**			
10	**Best-fit values**			
11	YIntercept	92.25	86.97	
12	Slope	-0.8750	-0.6075	
13	**95% CI (profile likelihood)**			
14	YIntercept	88.35 to 96.14	83.66 to 90.29	
15	Slope	-1.016 to -0.7336	-0.7347 to -0.4804	
16	**Goodness of Fit**			
17	Degrees of Freedom	48	49	
18	R squared	0.7633	0.6530	
19	Sum of Squares	3759	3306	
20	Sy.x	8.849	8.213	



Normality of Residuals			
Anderson-Darling (A2*)	0.6528	0.6557	
P value	0.0834	0.0821	
Passed normality test (alpha=0.05)?	Yes	Yes	
P value summary	ns	ns	
D'Agostino-Pearson omnibus (K2)	5.158	5.132	
P value	0.7727	0.0768	
Passed normality test (alpha=0.05)?	Yes	Yes	
P value summary	ns	ns	

Normality

There is a strong negative relationship between time spent revising and exam anxiety and that relationship is significantly stronger for girls than for boys (p=0.0056).