N-hacking revisited
Linear regression
Nonlinear regression
Analysis of variance (ANOVA)
Survival analysis

Practical Statistics for Experimental Biologists
Lecture 5
Friday, 31 July 2020
10:00am - 12:00pm

N-hacking

Multiple comparisons arise in many many contexts

multiple subgroups:
You perform tests on multiple subgroups of your data.

multiple ways to dichotomize:
You do pairwise comparisons between different combinations of subgroups.

multiple sample sizes:
You keep collecting data until you find $P < 0.05$.

DO NOT DO THIS.

multiple ways to preprocess the data:
You analyze data preprocessed in multiple different ways.

multiple statistical tests:
You use different statistical tests on the same data before finding $P < 0.05$.

Motulsky, Ch. 23

N-hacking might actually be good in principle!

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.12.868489. this version posted December 16, 2019.

Is N-Hacking Ever OK? A simulation-based study

Pamela Reinagel

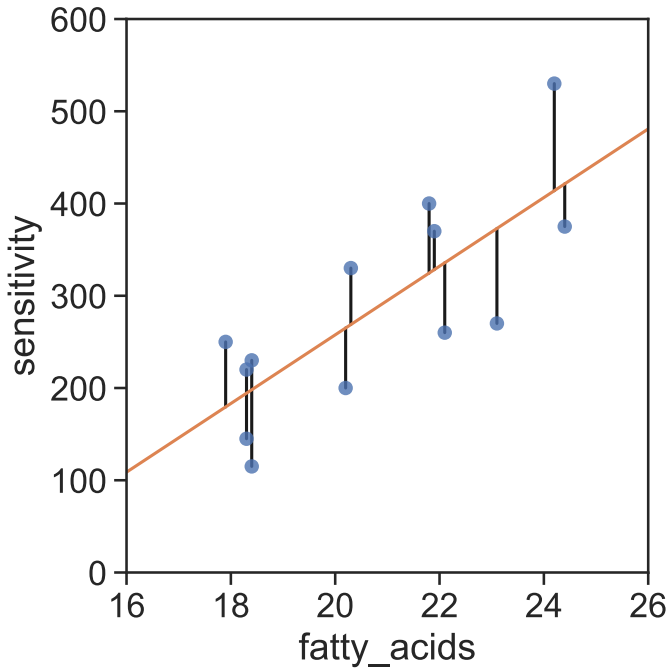
Section of Neurobiology, Division of Biological Science, University of California, San Diego.

Scientists in exactly this situation are currently being told (by teachers, advisors, reviewers, editors, and even staff biostatisticians) that if they have obtained a non-significant finding with a P value just above α, they cannot validly add more samples to their data set to improve statistical power; they must either run a completely independent replication, or accept the null hypothesis. The results shown here imply that this is bad advice. It is true that adding samples after the test violates the basic premise of null hypothesis significance testing (NHST). But that is not the same as being *invalid*. Adding more samples *with disclosure* is never invalid, and there are methods for rigorous correction of the P value within the NHST framework. Moreover, these simulations show that there are statistical benefits of incremental sampling that are often overlooked.

If you find yourself in this situation, do not throw away your data.
Rather, talk to someone in QB about how to proceed.

Linear regression

Linear regression seeks to explain y as a linear function of x plus Gaussian noise



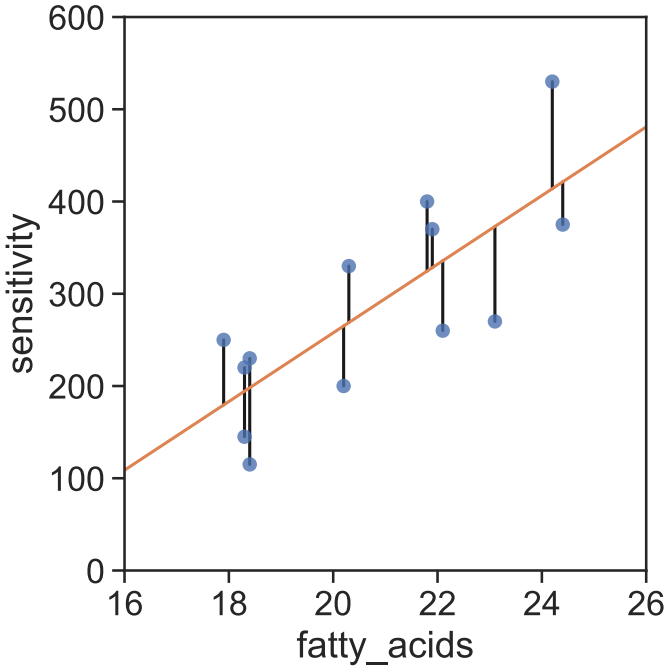
$$y_i = a + bx_i + \epsilon_i$$

a: y-intercept

b: slope

ϵ_i: the “residuals”

Parameters are chosen to minimize the sum of squared deviations



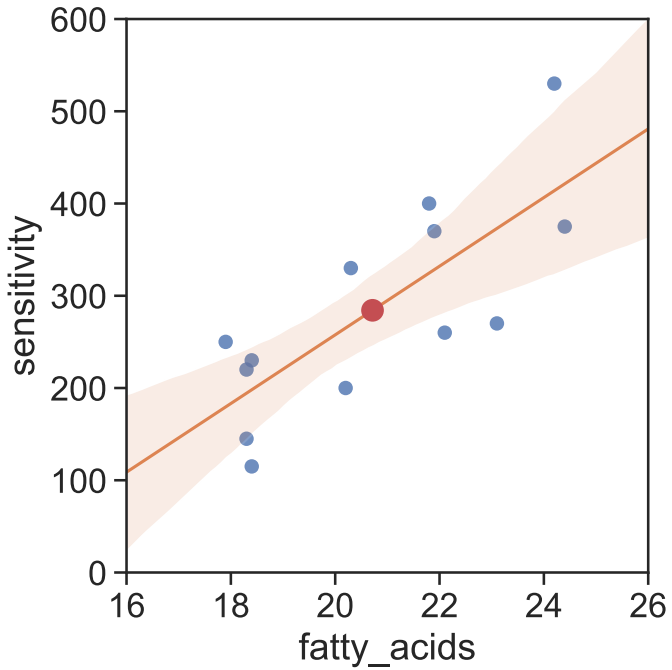
$$y_i = a + bx_i + \epsilon_i$$

The model "parameters", a and b, are chosen to minimize this quantity: $\sum_i \epsilon_i^2$.

This can be done mathematically, and one finds that,

$$b = r\frac{\hat{\sigma}_y}{\hat{\sigma}_x} \quad \text{and} \quad a = \hat{\mu}_y - b\hat{\mu}_x$$

Some properties of linear regression

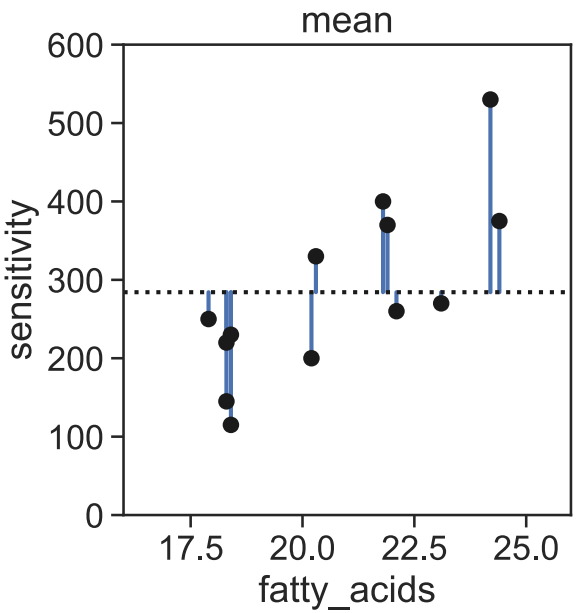


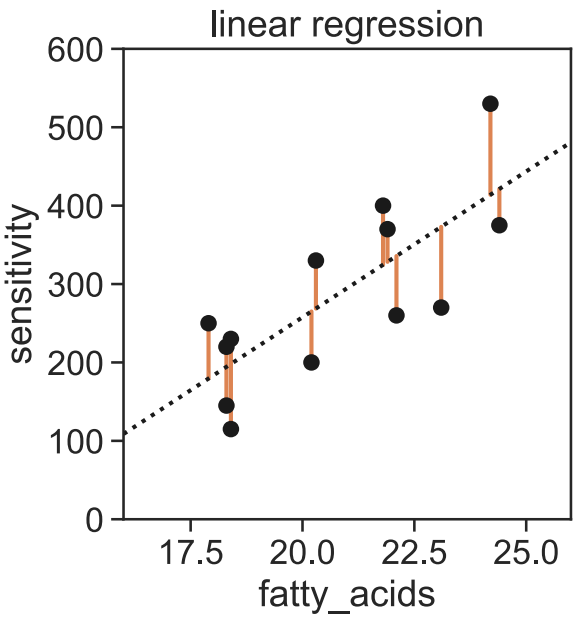
The center of mass point of the data, $(\hat{\mu}_x, \hat{\mu}_y)$, lies on the regression line.

Confidence intervals (shaded region) are curved because of uncertainty in both a and b.

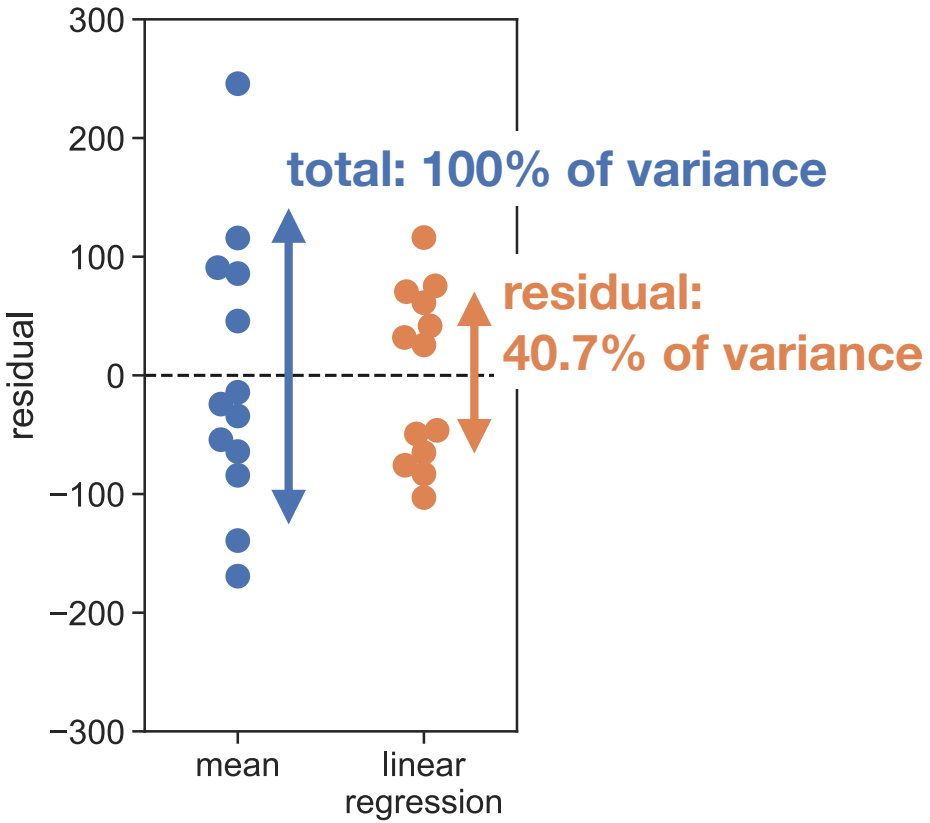
Any reported P-values correspond to the null hypothesis that $b = 0$.

Linear regression explains a fraction of the variance





model:
59.3% of variance



Linear regression explains a fraction of the variance

model: $\hat{y}_i = a + bx_i$

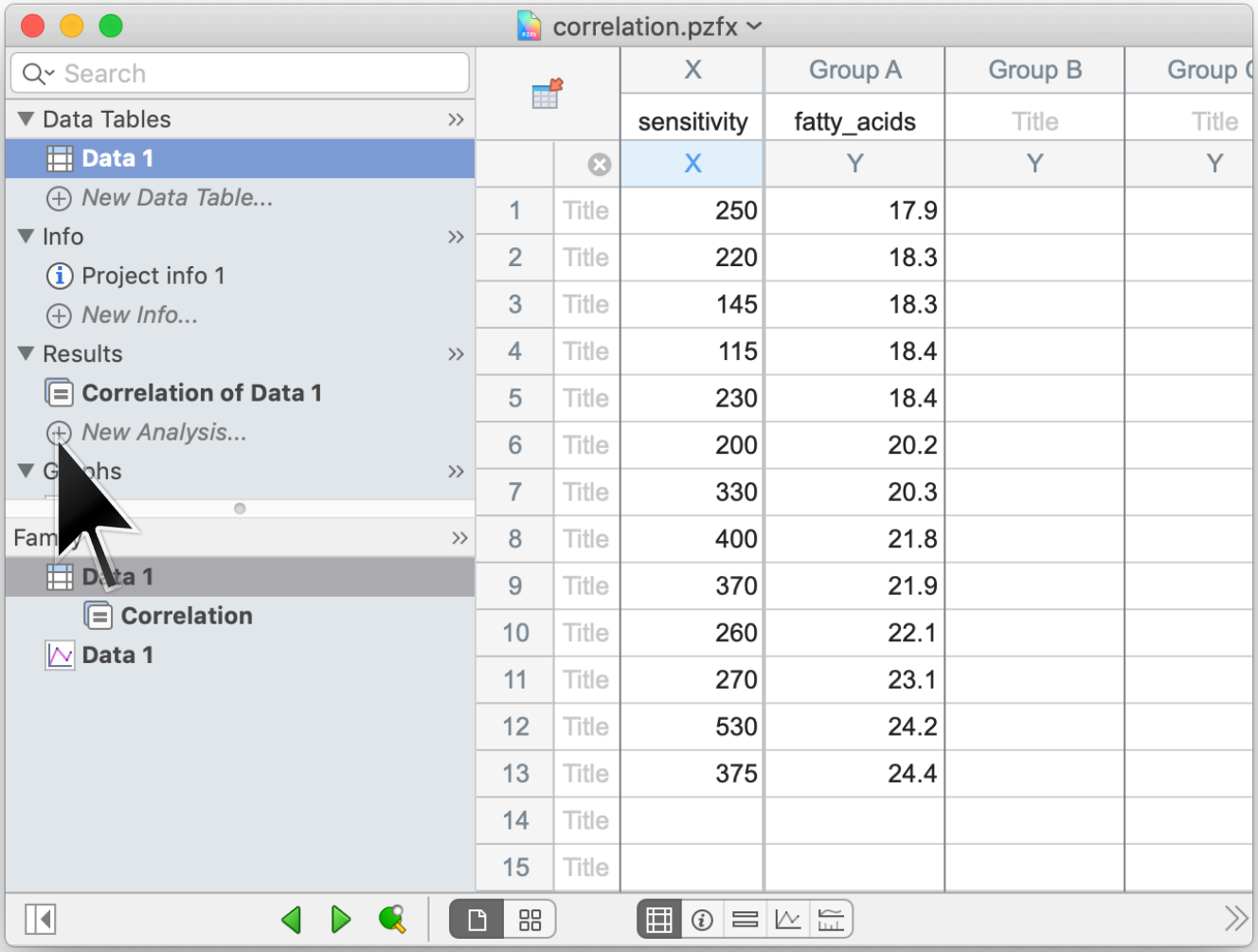
$(n-1) \times$ variance:

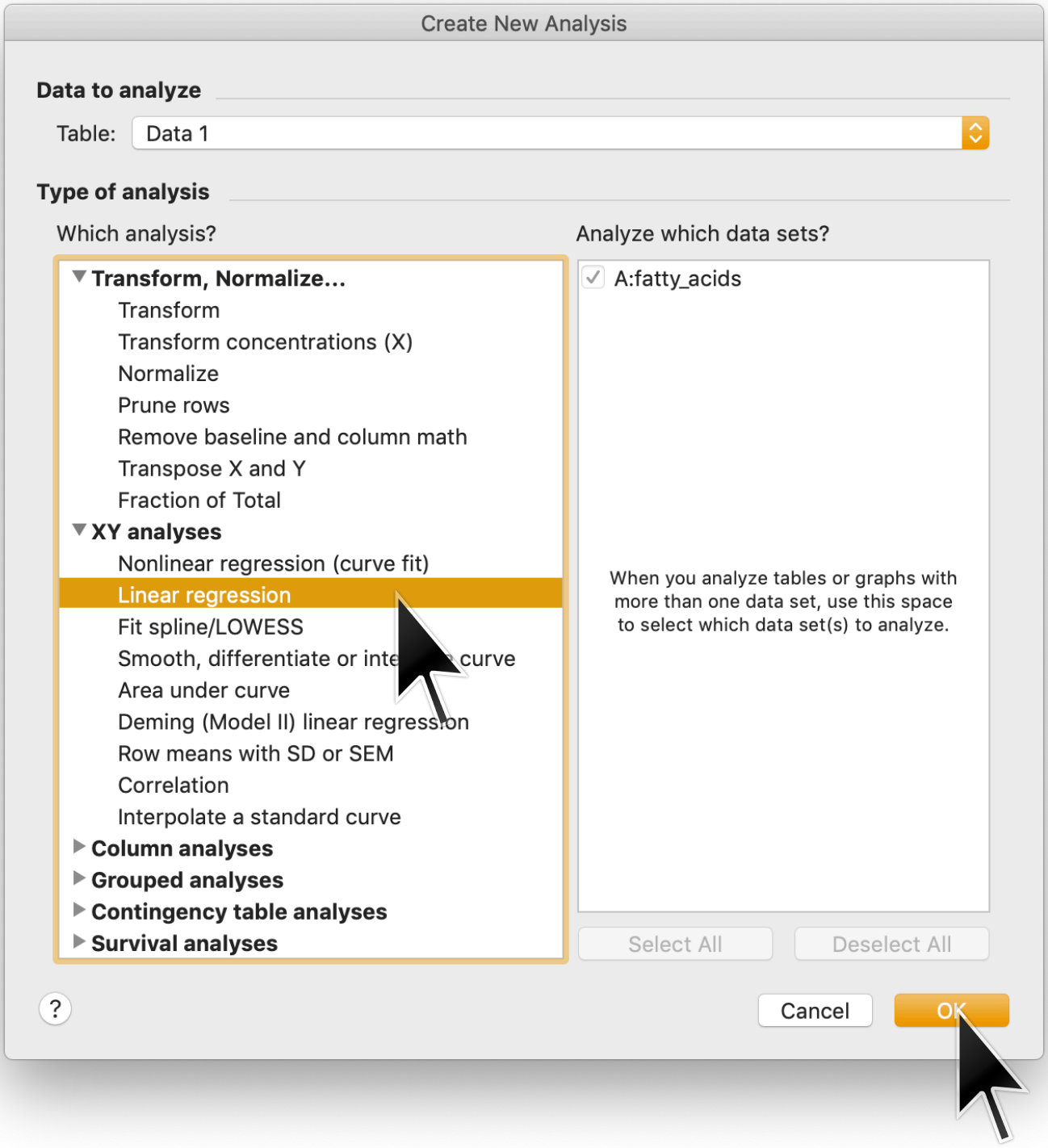
$$\sum_i (y_i - \hat{\mu}_y)^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \hat{\mu}_y)^2$$

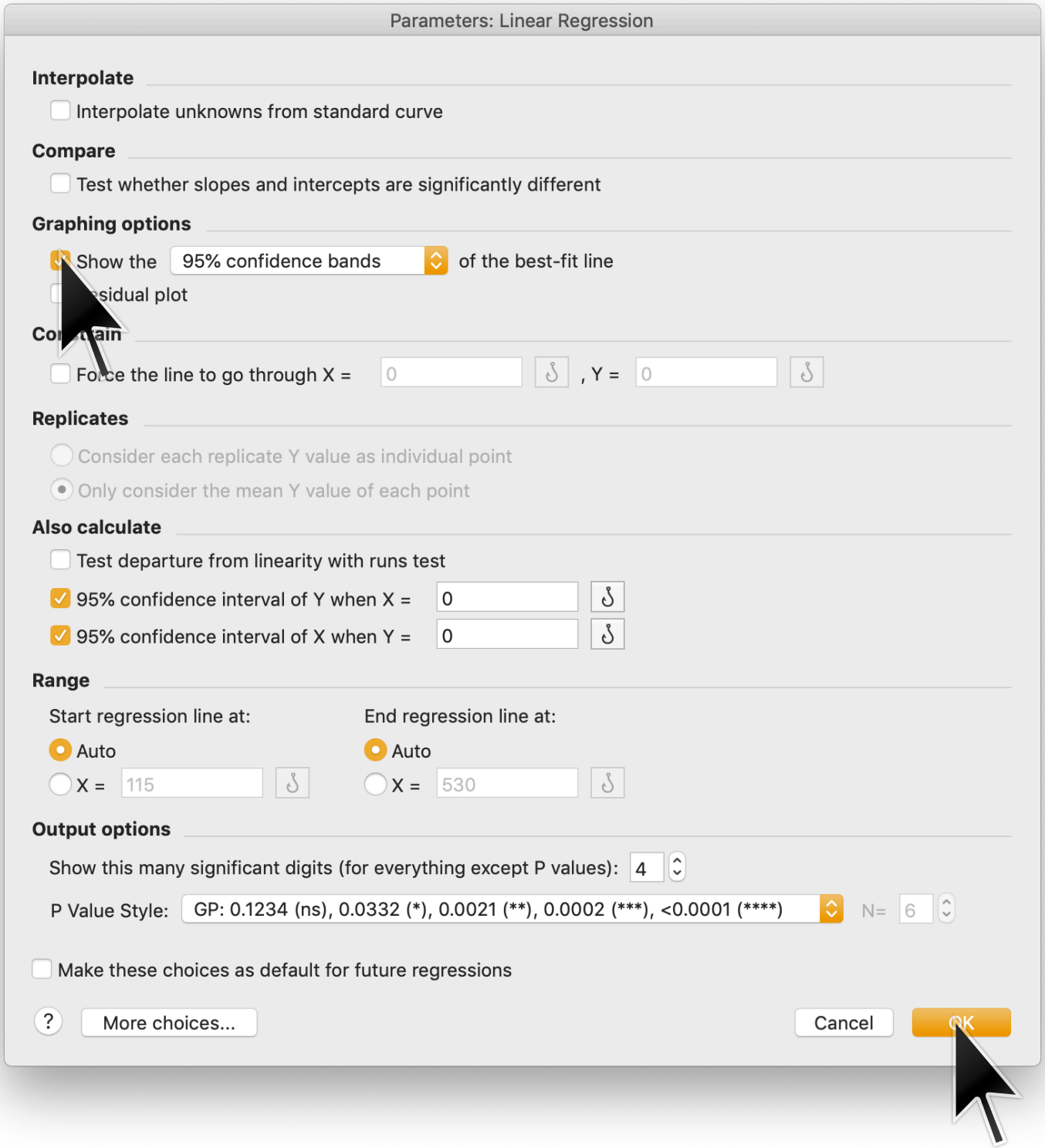
total:	**residual:**	**model:**
100%	**40.7%**	**59.3%**

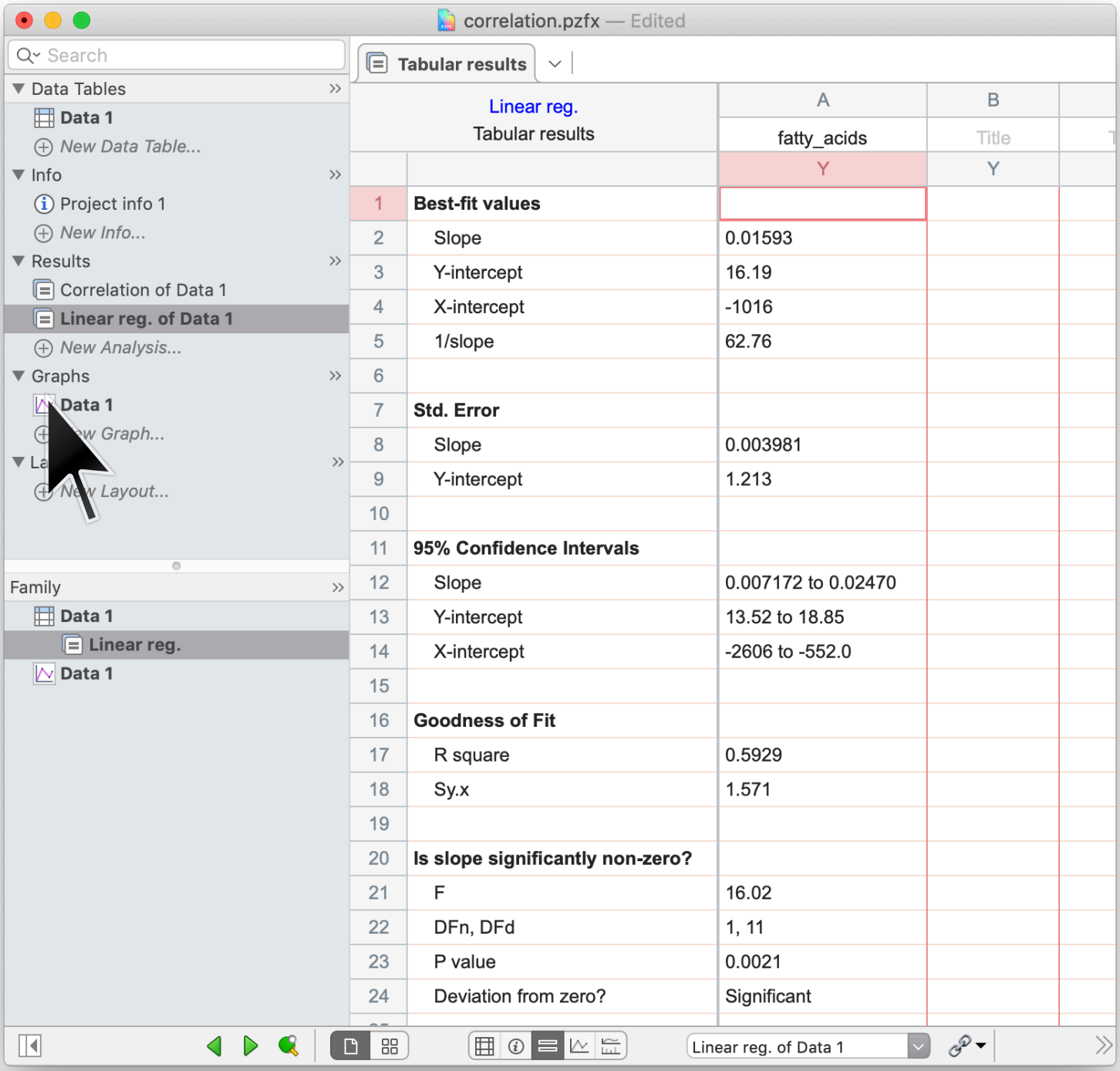
r^2 is the fraction of variance explained:

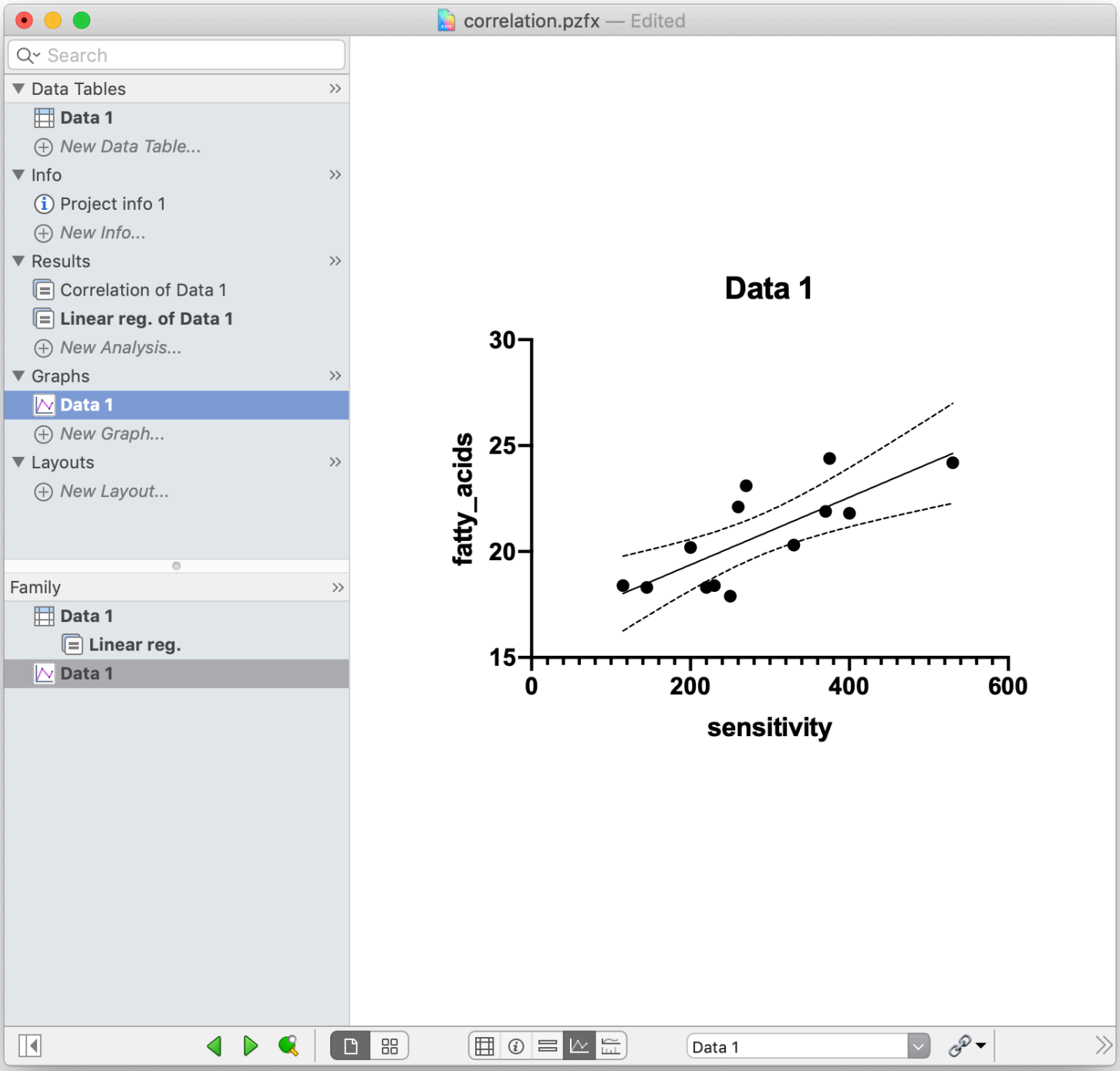
$$r^2 = \frac{\sum_i (\hat{y}_i - \hat{\mu}_y)^2}{\sum_i (y_i - \hat{\mu}_y)^2} = \mathbf{0.593}$$


correlation.pzfx
Search
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
Correlation of Data 1
New Analysis...
Data 1
Correlation
Data 1
X
Group A
Group B
sensitivity
fatty_acids
Title
Title
X
Y
Y
Y
1 Title 250 17.9
2 Title 220 18.3
3 Title 145 18.3
4 Title 115 18.4
5 Title 230 18.4
6 Title 200 20.2
7 Title 330 20.3
8 Title 400 21.8
9 Title 370 21.9
10 Title 260 22.1
11 Title 270 23.1
12 Title 530 24.2
13 Title 375 24.4
14 Title
15 Title


Create New Analysis
Data to analyze
Table:
Data 1
Type of analysis
Which analysis?
Analyze which data sets?
Transform, Normalize...
Transform
Transform concentrations (X)
Normalize
Prune rows
Remove baseline and column math
Transpose X and Y
Fraction of Total
XY analyses
Nonlinear regression (curve fit)
Linear regression
Fit spline/LOWESS
Smooth, differentiate or inte curve
Area under curve
Deming (Model II) linear regression
Row means with SD or SEM
Correlation
Interpolate a standard curve
Column analyses
Grouped analyses
Contingency table analyses
Survival analyses
A:fatty_acids
When you analyze tables or graphs with more than one data set, use this space to select which data set(s) to analyze.
Select All
Deselect All
?
Cancel
OK


Parameters: Linear Regression
Interpolate
Interpolate unknowns from standard curve
Compare
Test whether slopes and intercepts are significantly different
Graphing options
Show the
95% confidence bands
of the best-fit line
Residual plot
Constrain
Force the line to go through X =
0
, Y =
0
Replicates
Consider each replicate Y value as individual point
Only consider the mean Y value of each point
Also calculate
Test departure from linearity with runs test
95% confidence interval of Y when X =
0
95% confidence interval of X when Y =
0
Range
Start regression line at:
Auto
X =
115
End regression line at:
Auto
X =
530
Output options
Show this many significant digits (for everything except P values):
4
P Value Style:
GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.0002 (***), <0.0001 (****)
N=
6
Make these choices as default for future regressions
?
More choices...
Cancel
OK

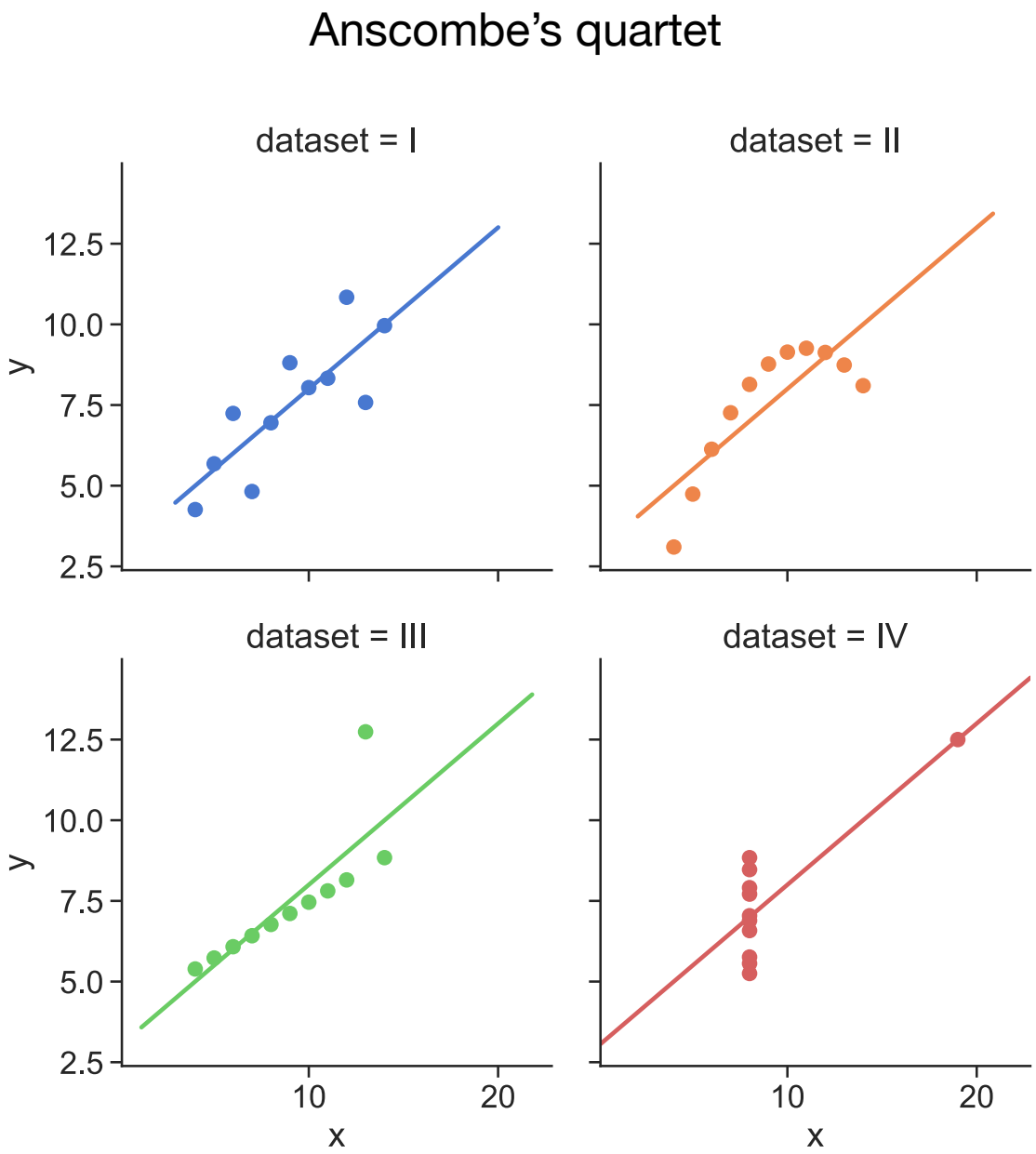

correlation.pzfx — Edited
Search
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
Correlation of Data 1
Linear reg. of Data 1
New Analysis...
Graphs
Data 1
New Graph...
New Layout...
Family
Data 1
Linear reg.
Data 1
Tabular results
Linear reg.
Tabular results
A
fatty_acids
B
Title
Y
Y
1 Best-fit values
2 Slope 0.01593
3 Y-intercept 16.19
4 X-intercept -1016
5 1/slope 62.76
7 Std. Error
8 Slope 0.003981
9 Y-intercept 1.213
11 95% Confidence Intervals
12 Slope 0.007172 to 0.02470
13 Y-intercept 13.52 to 18.85
14 X-intercept -2606 to -552.0
16 Goodness of Fit
17 R square 0.5929
18 Sy.x 1.571
20 Is slope significantly non-zero?
21 F 16.02
22 DFn, DFd 1, 11
23 P value 0.0021
24 Deviation from zero? Significant
Linear reg. of Data 1


correlation.pzfx — Edited
Search
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
Correlation of Data 1
Linear reg. of Data 1
New Analysis...
Graphs
Data 1
New Graph...
Layouts
New Layout...
Family
Data 1
Linear reg.
Data 1
Data 1
fatty_acids
30
25
20
15
0
200
400
600
sensitivity
Data 1

Linear regression assumptions

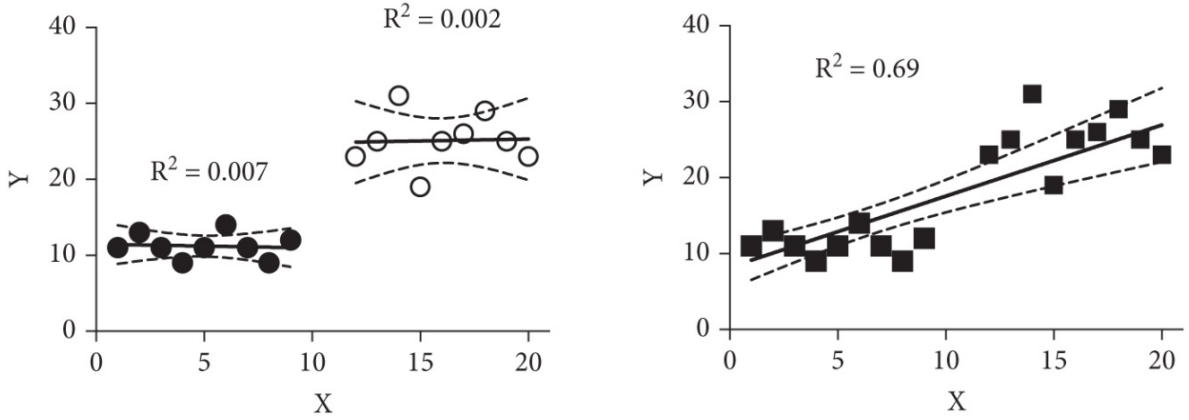
- The model is correct, i.e. the expected value for y is indeed a linear function of x for some correct choice of parameters.
- The noise (i.e. the residuals) is Gaussian and has mean zero.
- The residual for each data point is statistically independent
- The magnitude of the noise (i.e. variance of the Gaussian) is the same at all x values.
- Each x_i is known exactly.

As with correlation, many different-looking datasets can have exactly the same regression line

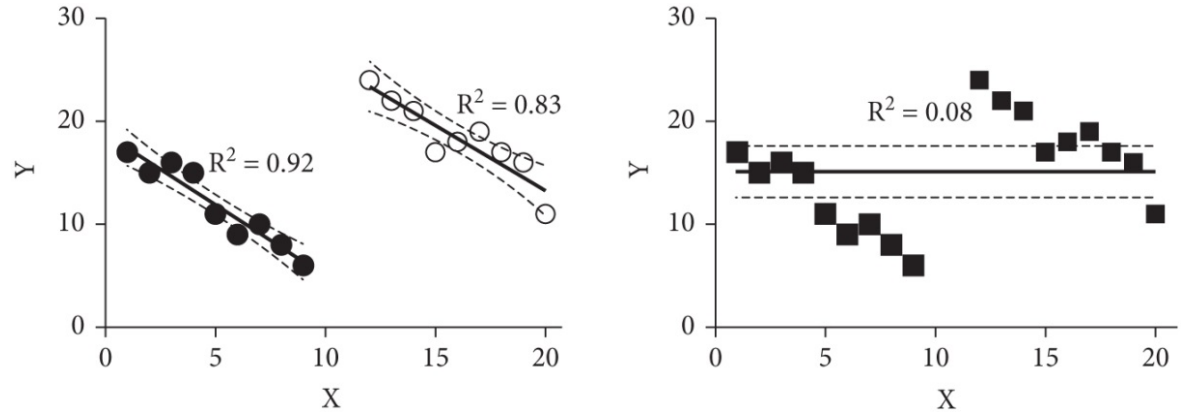


Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21.

Beware of combining distinct groups into one



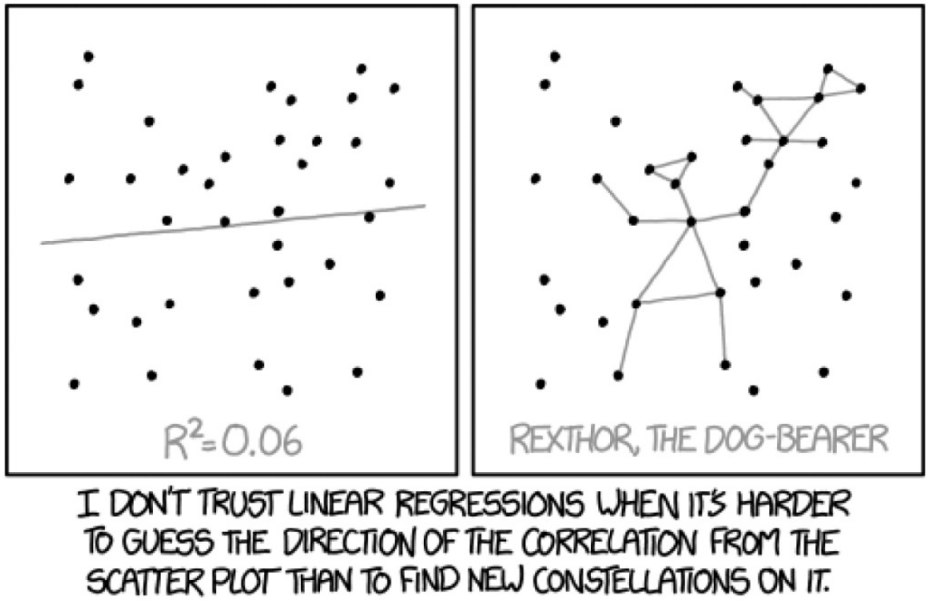
Combining two groups into one regression can mislead by creating a strong linear relationship.



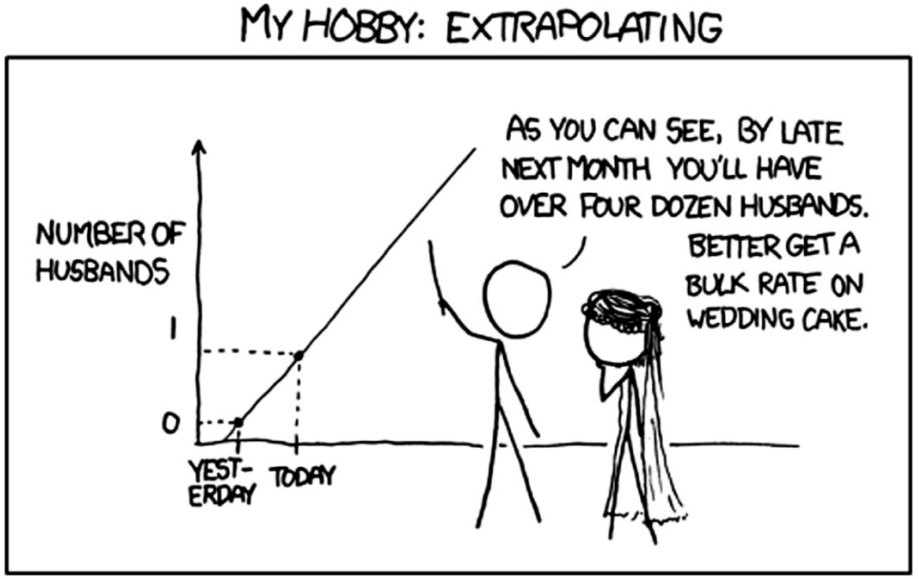
Combining two groups into one regression can mislead by hiding a trend.

Beware of reading too much into a regression result

Don’t trust regression results that you can’t verify by eye



Don’t over-extrapolate



Nonlinear regression

Example: effect of norepinephrine on muscle relaxation

log10_conc	pct_relaxation
-8.0	2.6
-7.5	10.5
-7.0	15.8
-6.5	21.1
-6.0	36.8
-5.5	57.9
-5.0	73.7
-4.5	89.5
-4.0	94.7
-3.5	100.0
-3.0	100.0

Frazier et al (2006) measured the degree to which the neurotransmitter norepinephrine relaxes bladder muscle in rats.

Strips of bladder muscle were exposed to various concentrations of norepinephrine, and percent muscle relaxation was measured.

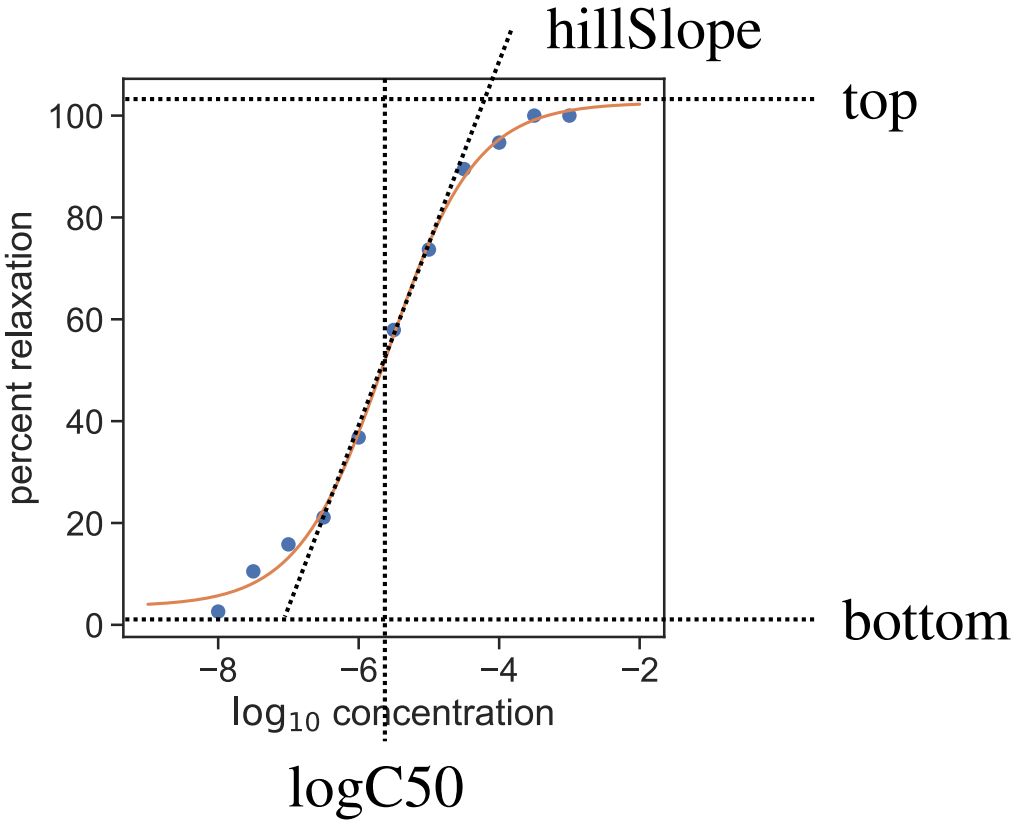
The data from each rat was analyzed to determine the maximum relaxation and the concentration of norepinephrine that relaxes the muscle half that much (C50)

Example: effect of norepinephrine on muscle relaxation

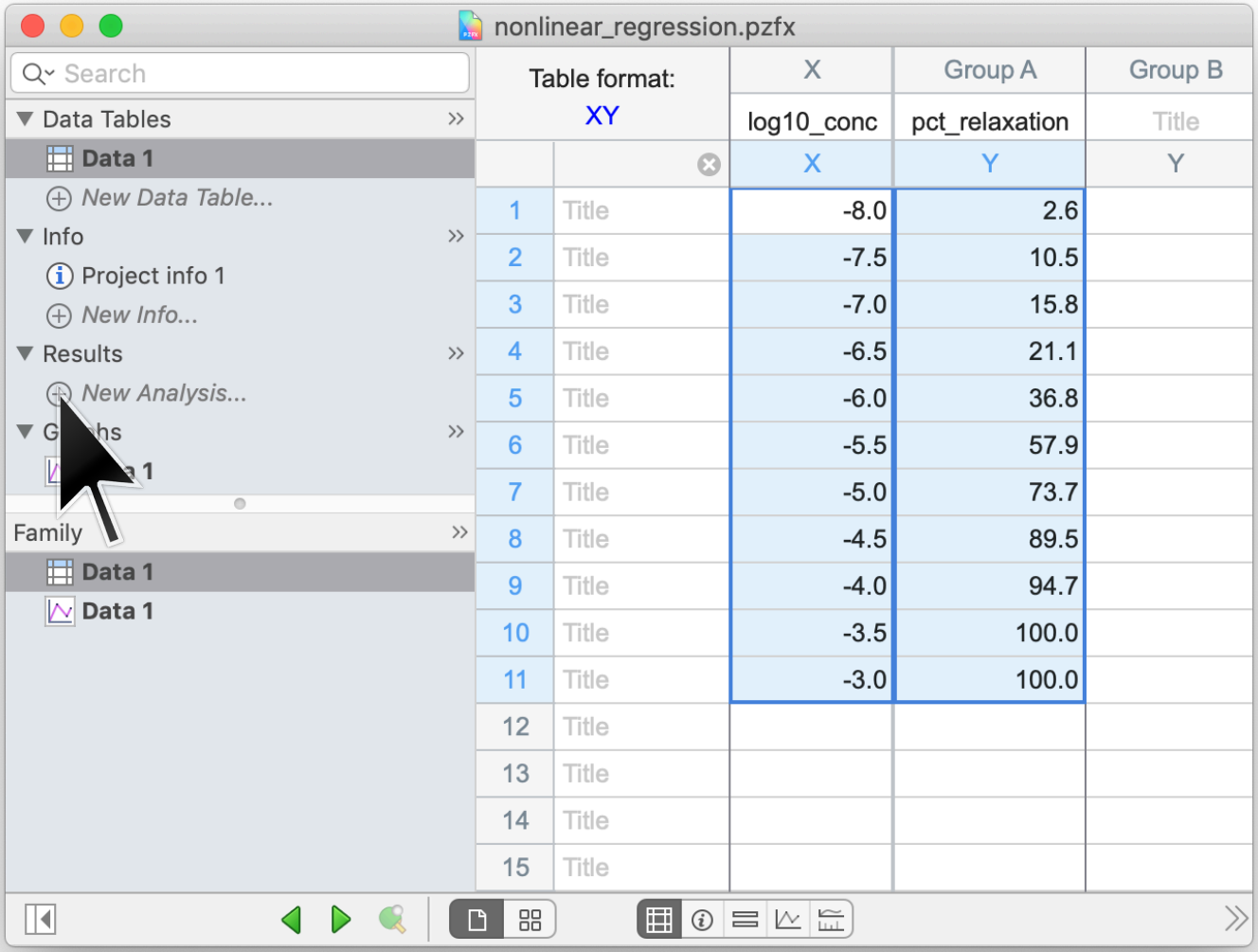
log10_conc	pct_relaxation
-8.0	2.6
-7.5	10.5
-7.0	15.8
-6.5	21.1
-6.0	36.8
-5.5	57.9
-5.0	73.7
-4.5	89.5
-4.0	94.7
-3.5	100.0
-3.0	100.0

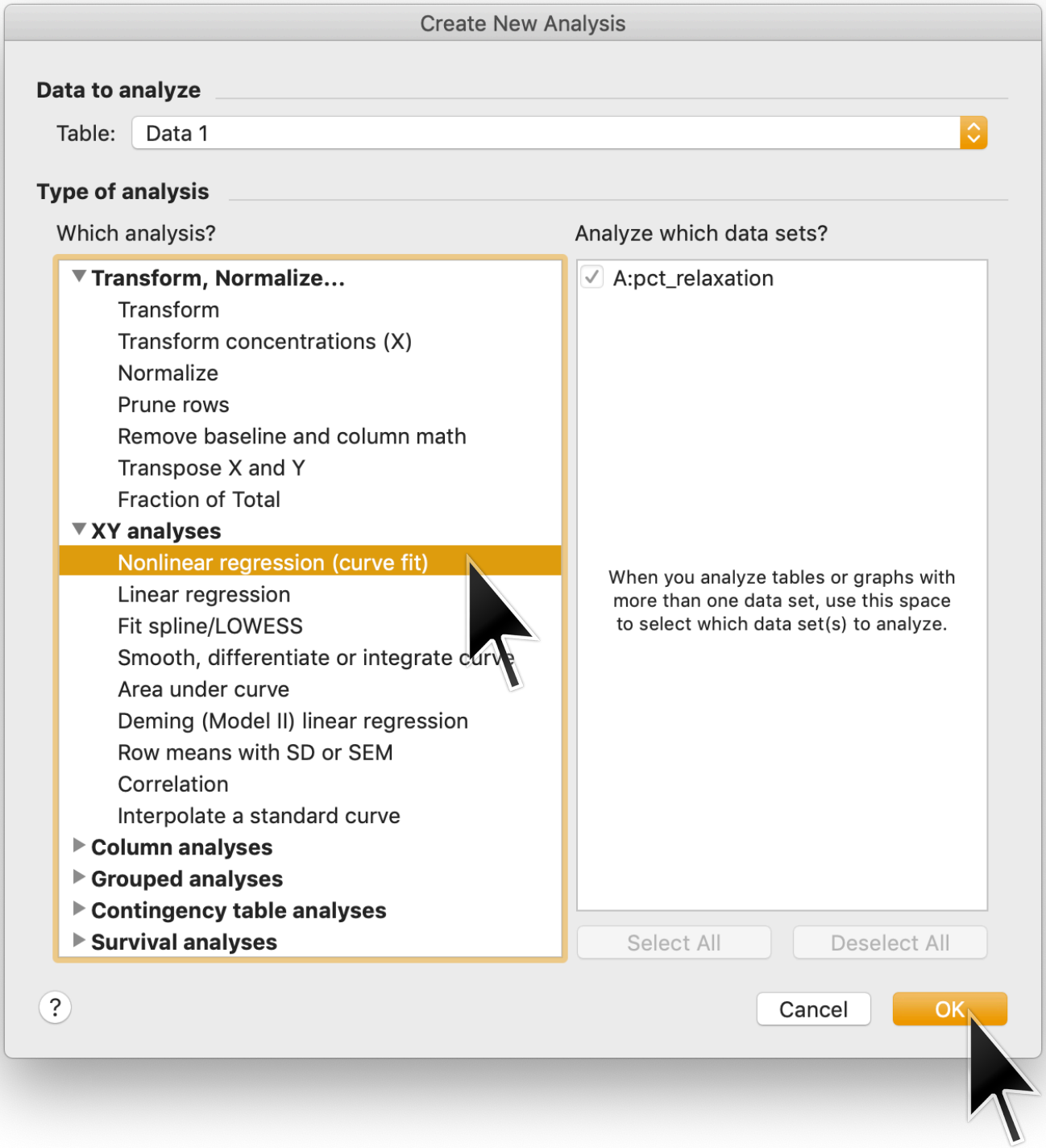
x: $\log_{10}$ concentration (in M)

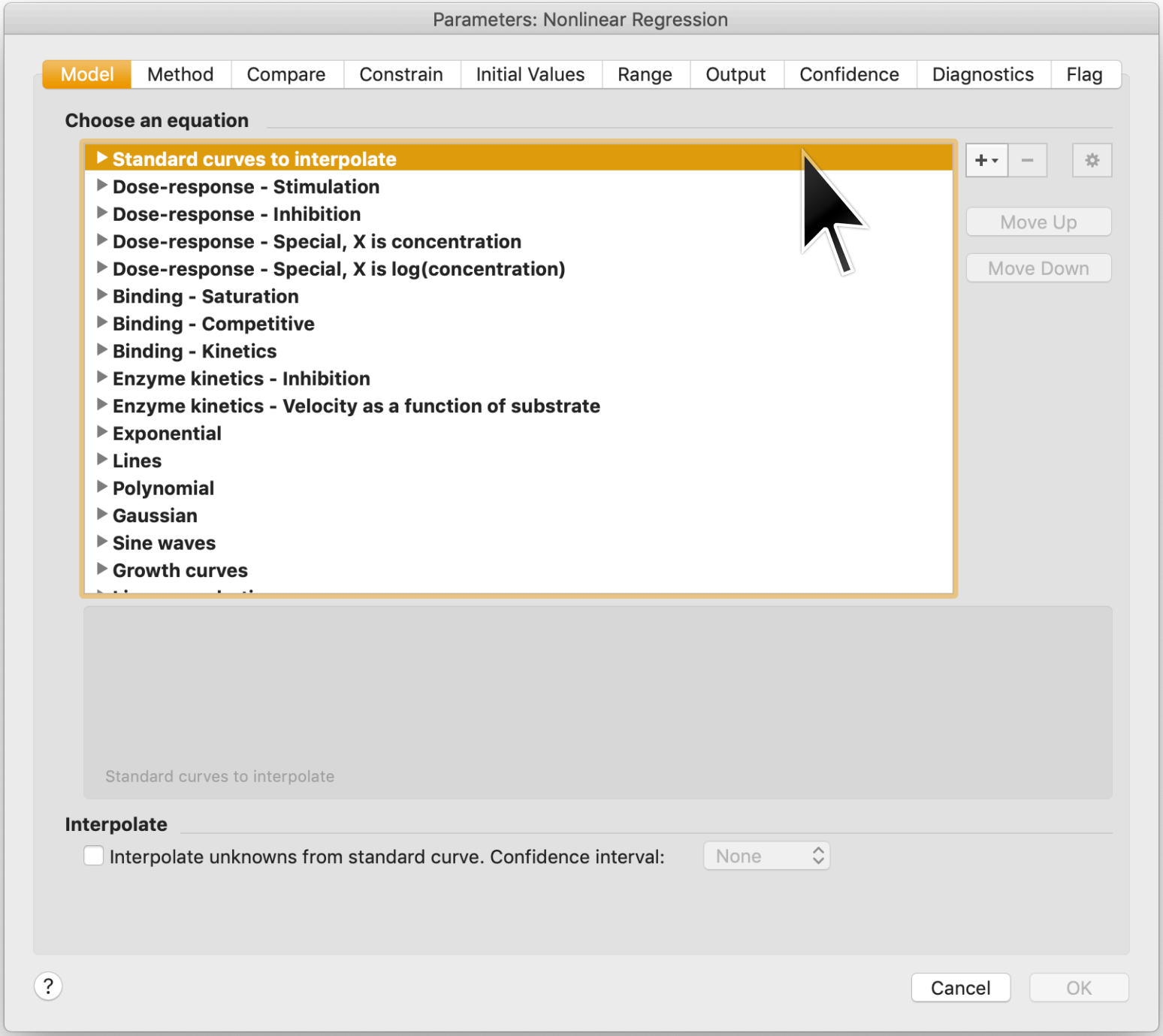
y: percent muscle relaxation

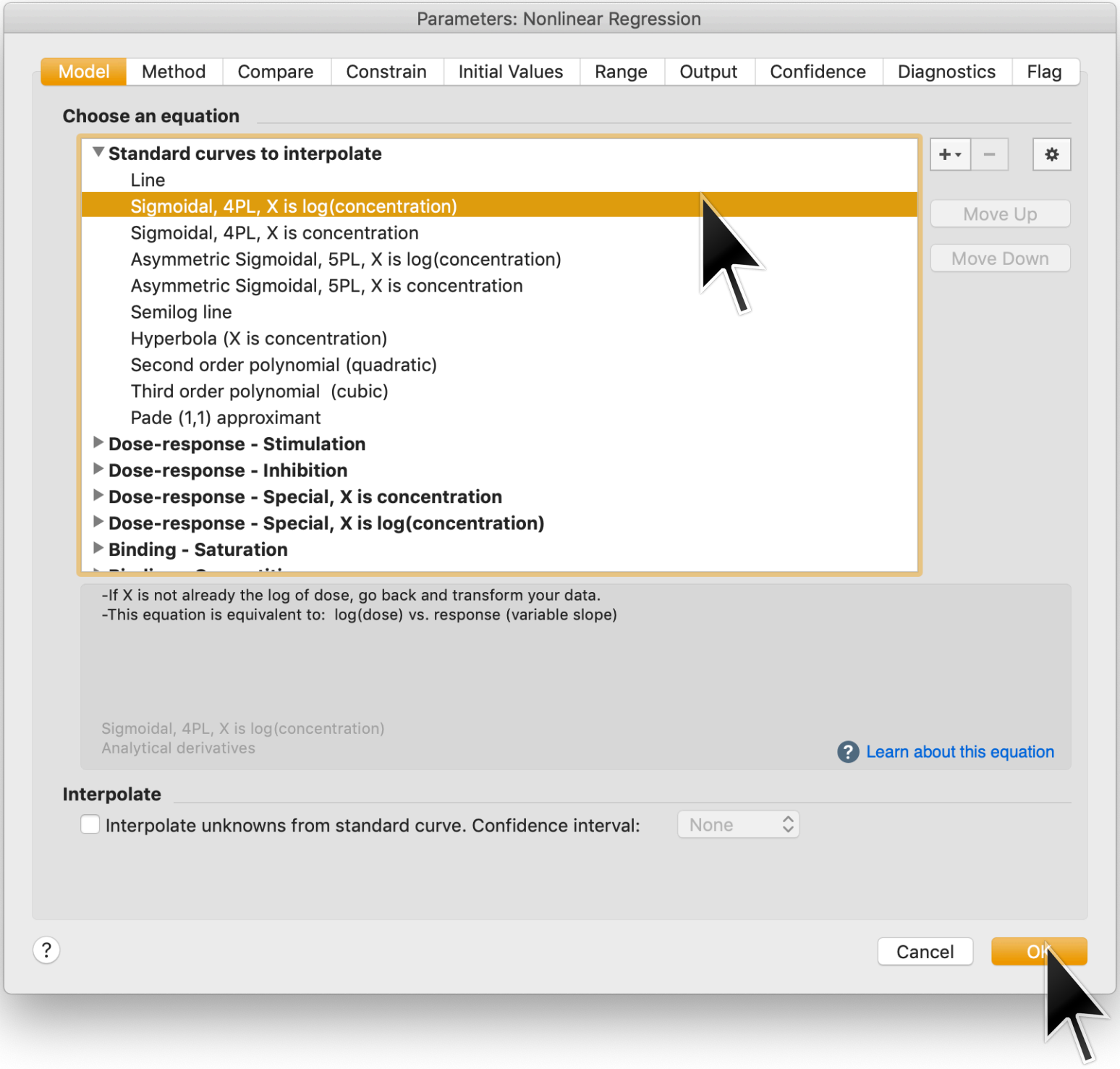


$$f(x) = \text{bottom} + \frac{\text{top} - \text{bottom}}{1 + 10^{(\text{logC50} - x)\cdot \text{hillSlope}}}$$


nonlinear_regression.pzfx
Search
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
New Analysis...
Family
Data 1
Data 1
Table format:
XY
X
Group A
Group B
log10_conc
pct_relaxation
Title
X
Y
Y
1 Title -8.0 2.6
2 Title -7.5 10.5
3 Title -7.0 15.8
4 Title -6.5 21.1
5 Title -6.0 36.8
6 Title -5.5 57.9
7 Title -5.0 73.7
8 Title -4.5 89.5
9 Title -4.0 94.7
10 Title -3.5 100.0
11 Title -3.0 100.0
12 Title
13 Title
14 Title
15 Title

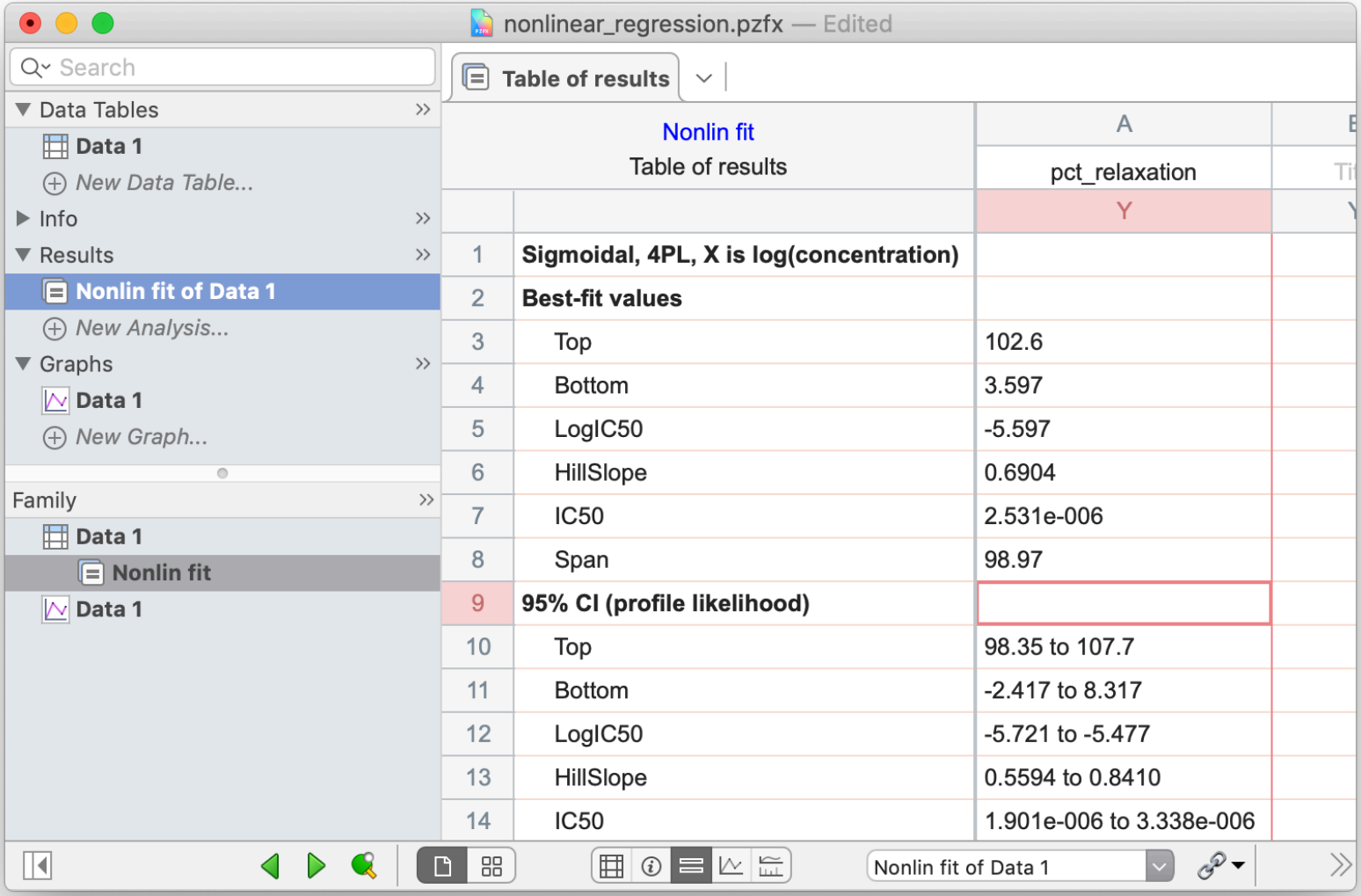

Create New Analysis
Data to analyze
Table:
Data 1
Type of analysis
Which analysis?
Analyze which data sets?
Transform, Normalize...
Transform
Transform concentrations (X)
Normalize
Prune rows
Remove baseline and column math
Transpose X and Y
Fraction of Total
XY analyses
Nonlinear regression (curve fit)
Linear regression
Fit spline/LOWESS
Smooth, differentiate or integrate curve
Area under curve
Deming (Model II) linear regression
Row means with SD or SEM
Correlation
Interpolate a standard curve
Column analyses
Grouped analyses
Contingency table analyses
Survival analyses
A:pct_relaxation
When you analyze tables or graphs with more than one data set, use this space to select which data set(s) to analyze.
Select All
Deselect All
?
Cancel
OK


Parameters: Nonlinear Regression
Model
Method
Compare
Constrain
Initial Values
Range
Output
Confidence
Diagnostics
Flag
Choose an equation
Standard curves to interpolate
Dose-response - Stimulation
Dose-response - Inhibition
Dose-response - Special, X is concentration
Dose-response - Special, X is log(concentration)
Binding - Saturation
Binding - Competitive
Binding - Kinetics
Enzyme kinetics - Inhibition
Enzyme kinetics - Velocity as a function of substrate
Exponential
Lines
Polynomial
Gaussian
Sine waves
Growth curves
Move Up
Move Down
Standard curves to interpolate
Interpolate
Interpolate unknowns from standard curve. Confidence interval:
None
?
Cancel
OK


Parameters: Nonlinear Regression
Model
Method
Compare
Constrain
Initial Values
Range
Output
Confidence
Diagnostics
Flag
Choose an equation
Standard curves to interpolate
Line
Sigmoidal, 4PL, X is log(concentration)
Sigmoidal, 4PL, X is concentration
Asymmetric Sigmoidal, 5PL, X is log(concentration)
Asymmetric Sigmoidal, 5PL, X is concentration
Semilog line
Hyperbola (X is concentration)
Second order polynomial (quadratic)
Third order polynomial (cubic)
Pade (1,1) approximant
Dose-response - Stimulation
Dose-response - Inhibition
Dose-response - Special, X is concentration
Dose-response - Special, X is log(concentration)
Binding - Saturation
Move Up
Move Down
-If X is not already the log of dose, go back and transform your data.
-This equation is equivalent to: log(dose) vs. response (variable slope)
Sigmoidal, 4PL, X is log(concentration)
Analytical derivatives
Learn about this equation
Interpolate
Interpolate unknowns from standard curve. Confidence interval:
None
Cancel

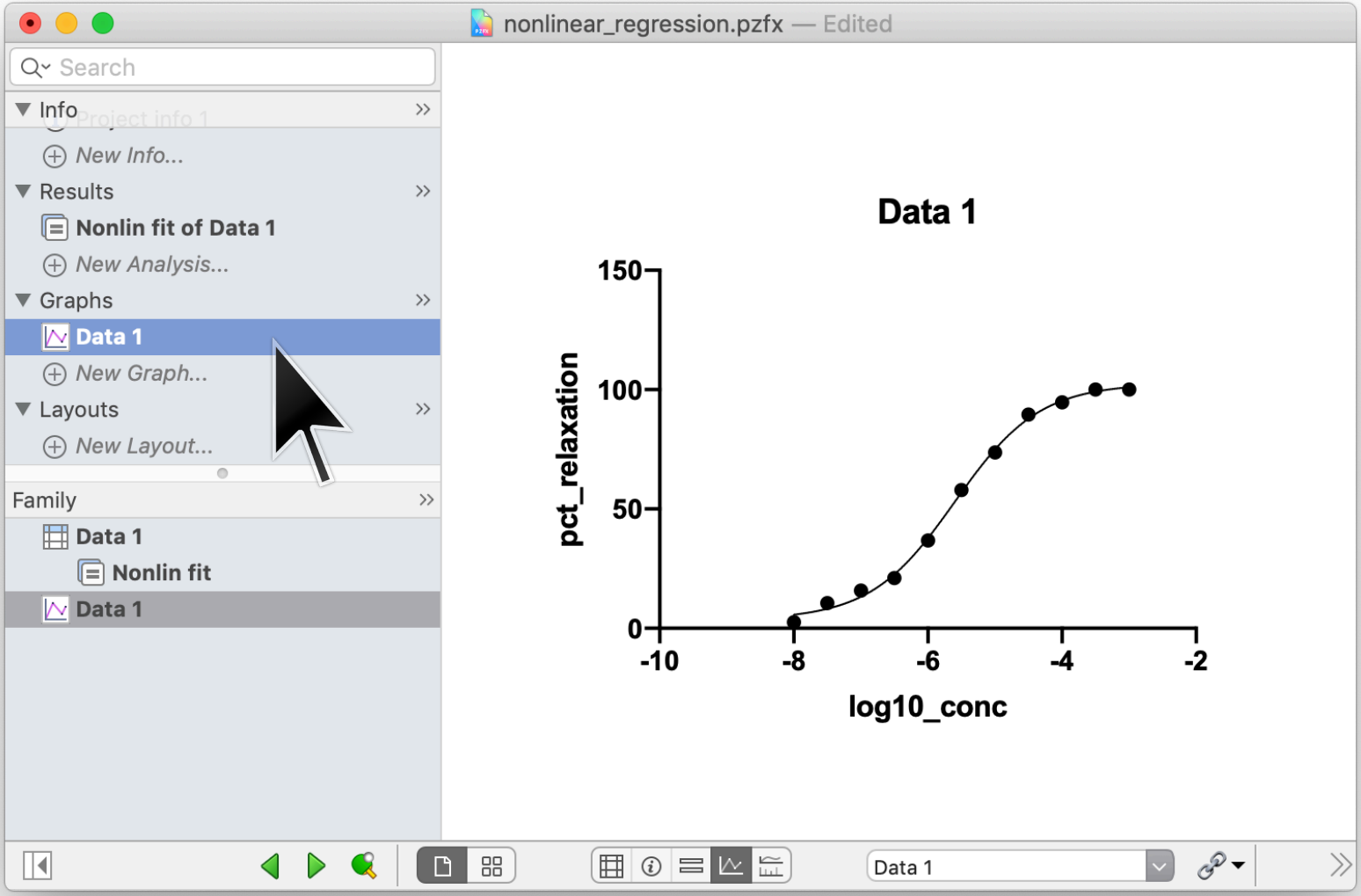
$$y = \text{Bottom} + \frac{\text{Top} - \text{Bottom}}{1 + 10^{(\text{LogIC50} - x)\cdot \text{HillSlope}}}$$

4 parameters: Bottom, Top, LogIC50, HillSlope



$$y = \text{Bottom} + \frac{\text{Top} - \text{Bottom}}{1 + 10^{(\text{LogIC50} - x)\cdot \text{HillSlope}}}$$

4 parameters: Bottom, Top, LogIC50, HillSlope



Multiple linear regression and logistic regression

Multiple linear regression is used to model a continuous number that depends on multiple covariates

Multiple linear regression (often just called "linear regression") is used to model data where each data point $(\vec{x}_i, y_i)$ consist of an independent variable $\vec{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, which is a D-dimensional vector, and a dependent variable y_i, which is a single number. Often the entries of the vector $\vec{x}_i$ are called "covariates".

The key assumption is that each dependent variable y_i is related to the corresponding independent variables via

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_D x_{iD} + \epsilon_i$$

where the residual ϵ_i is due to random Gaussian noise.

The covariants that define $\vec{x}$ are often a mixture of continuous and binary variables.

Logistic regression is used to model probabilities that depend on multiple covariates

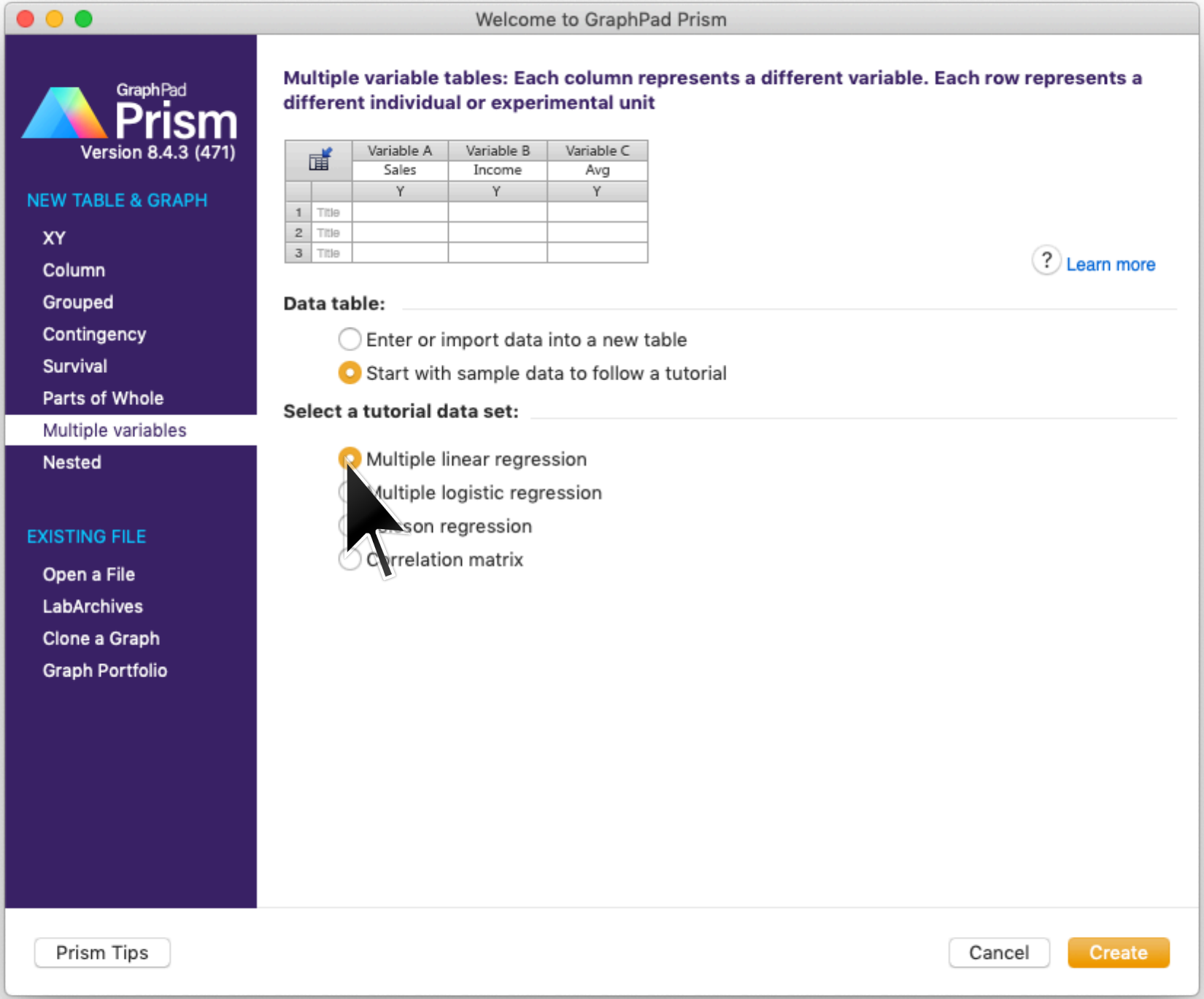
Logistic regression is used to model data where each data point $(\vec{x}_i, y_i)$ consists of a vector $\vec{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$ that represents D covariants, and one dependent variable y_i that is **binary**.

The key assumption is that the log odds of y_i is a linear function of $\vec{x}_i$:

$$\log \text{Odds}_\text{i} = \log \left[\frac{p(y_i = 1 \mid \vec{x}_i)}{p(y_i = 0 \mid \vec{x}_i)} \right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_D x_{iD}$$

Note that there is no need for a “residual” contribution since the model is inherently probabilistic.

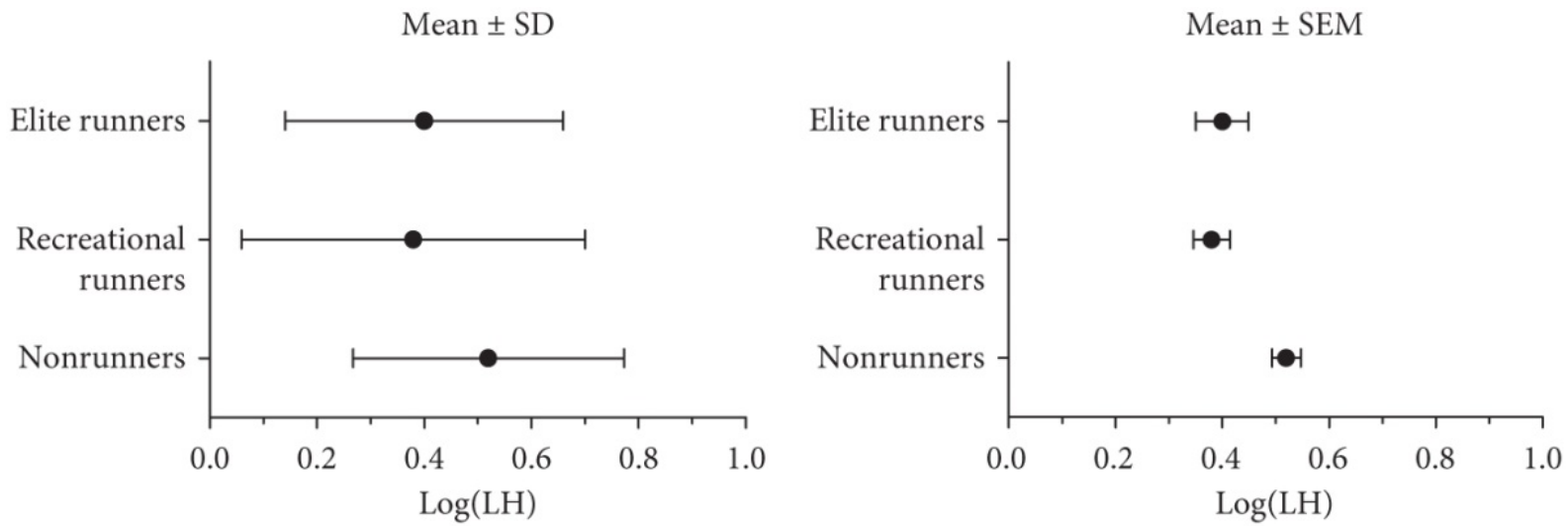
Again, the covariants that define $\vec{x}$ are often a mixture of continuous and binary variables.


Welcome to GraphPad Prism
GraphPad Prism
Version 8.4.3 (471)
NEW TABLE & GRAPH
XY
Column
Grouped
Contingency
Survival
Parts of Whole
Multiple variables
Nested
EXISTING FILE
Open a File
LabArchives
Clone a Graph
Graph Portfolio
Multiple variable tables: Each column represents a different variable. Each row represents a different individual or experimental unit
Variable A
Sales
Y
Variable B
Income
Y
Variable C
Avg
Y
1 Title
2 Title
3 Title
Learn more
Data table:
Enter or import data into a new table
Start with sample data to follow a tutorial
Select a tutorial data set:
Multiple linear regression
Multiple logistic regression
Poisson regression
Correlation matrix
Prism Tips
Cancel
Create

Analysis of variance (ANOVA)

One-way ANOVA example: hormone levels in runners

Hetland et al. (1993) investigated the level of luteinizing hormone (LH) in runners. Runners were classified into three groups: elite runners, recreational runners, and nonrunners.



GROUP	LOG(LH)	SD	SEM	N
nonrunners	0.52	0.25	0.027	88
recreational runners	0.38	0.32	0.034	89
elite runners	0.40	0.26	0.049	28

One-way ANOVA analyzes whether group means are significantly different

Null hypothesis: different groups have identical means

Alternative hypothesis: different groups have different means

SS = sum of squares

$$\underbrace{\sum_i (y_i - \hat{\mu})^2}_{\text{SS}_{\text{total}}} = \underbrace{\sum_i (y_i - \hat{\mu}_{g_i})^2}_{\text{SS}_{\text{within}}} + \underbrace{\sum_i (\hat{\mu}_{g_i} - \hat{\mu})^2}_{\text{SS}_{\text{between}}}$$

grand mean:

$$\hat{\mu} = \frac{1}{N} \sum_i y_i$$

group means:

$$\hat{\mu}_g = \frac{1}{N_g} \sum_{i|g} y_i$$

fraction variance explained:

$$\eta^2 = R^2 = \frac{\text{SS}_{\text{between}}}{\text{SS}_{\text{total}}}$$

One-way ANOVA analyzes whether group means are significantly different

$$\underbrace{\sum_i (y_i - \hat{\mu})^2}_{\text{SS}_{\text{total}}} = \underbrace{\sum_i (y_i - \hat{\mu}_{g_i})^2}_{\text{SS}_{\text{within}}} + \underbrace{\sum_i (\hat{\mu}_{g_i} - \hat{\mu})^2}_{\text{SS}_{\text{between}}}$$

DF = degree of freedom

$$\text{DF}_{\text{within}} = N - G, \quad \text{MS}_{\text{within}} = \frac{\text{SS}_{\text{within}}}{\text{DF}_{\text{within}}}$$

MS = mean square

similar if null is true

$$\text{DF}_{\text{between}} = G - 1, \quad \text{MS}_{\text{between}} = \frac{\text{SS}_{\text{between}}}{\text{DF}_{\text{between}}}$$

The corresponding F statistic is: $F = \dfrac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}}$

$F \approx 1$ if null is true

The null hypothesis, implies that: $F \sim \text{FDist}(\text{DF}_{\text{between}}, \text{DF}_{\text{within}})$

Alternatively, ANOVA can be thought of as a form of linear regression

$y_i = \log(\text{LH})$

$x_{i1} =$ elite runner? (1 or 0)

$x_{i2} =$ recreational runner? (1 or 0)

Null model: $y = \beta_0 + \epsilon_i$

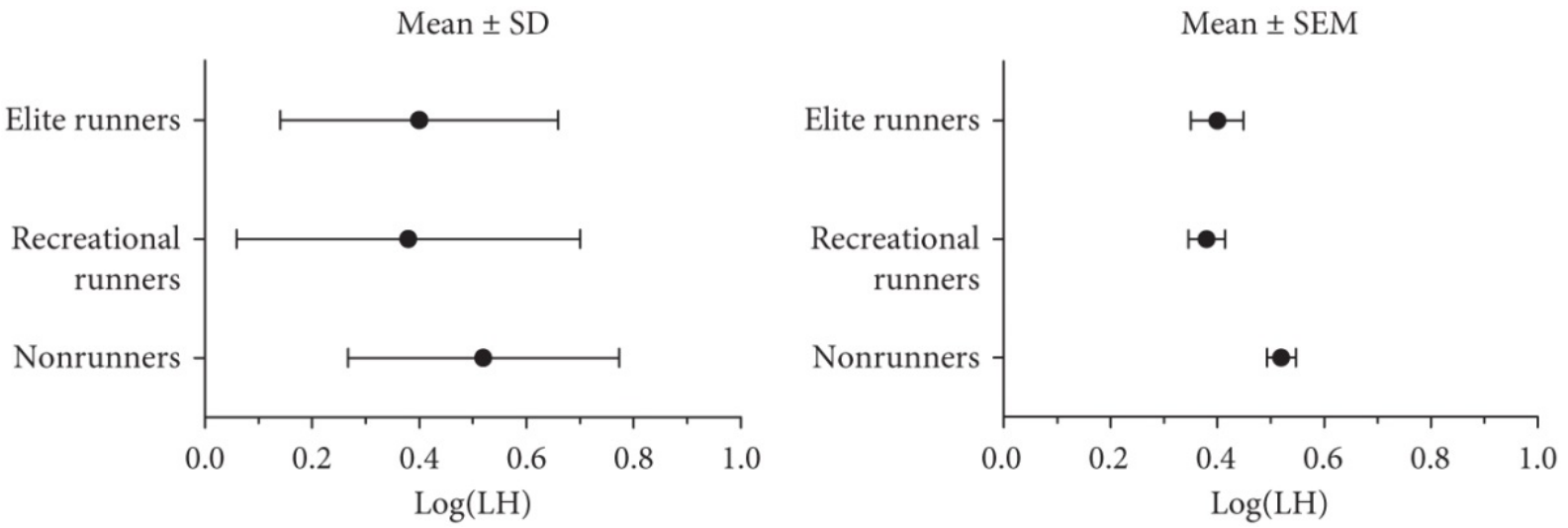
Alternative model: $y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

parameter correspondence: $\beta_0 = \mu_{\text{non}}, \ \beta_1 = \mu_{\text{elite}} - \mu_{\text{non}}, \ \ \beta_2 = \mu_{\text{rec}} - \mu_{\text{non}}$

The alternative model will always fit the data better. But how much better?

The F test tests whether the extra parameters, β_1 and β_2 are worth it.

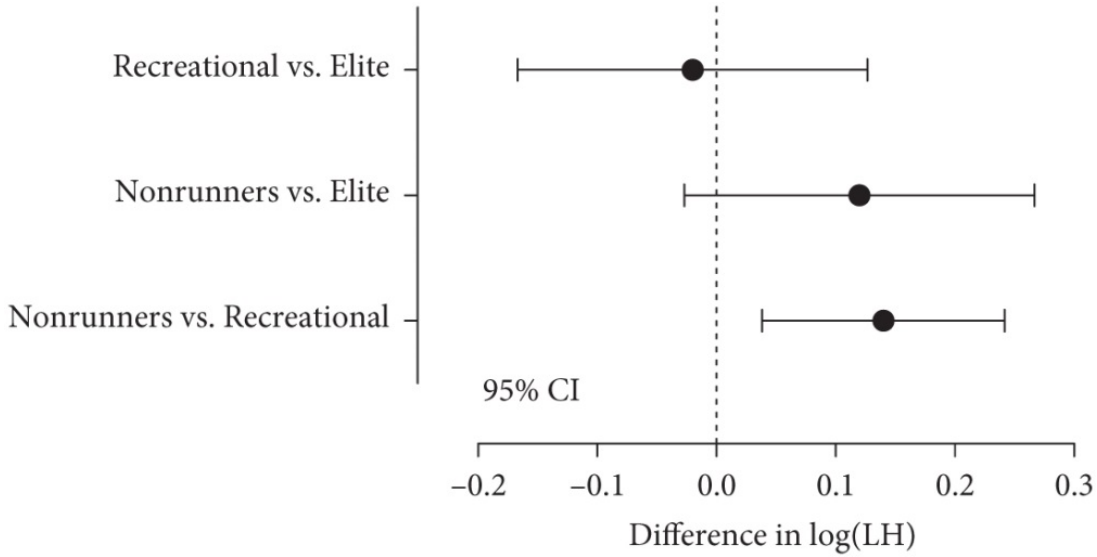
One-way ANOVA analyzes whether group means are significantly different



	SOURCE OF VARIATION	SUM OF SQUARES	DF	MS	F RATIO	P VALUE
	Between groups	0.93	2	0.46	5.69	0.0039
-	Within groups (resid.)	16.45	202	0.081		
=	Total	17.38	204			

This shows that the three groups have significantly different means. It does NOT, however, say which pairs of means (if any) are significantly different.

Tukey's test analyzes which pairwise comparisons in a one-way ANOVA, if any, are significant.



Tukey's test automatically incorporates the necessary multiple hypothesis correction into the test of significance.

There are other ANOVA post-hoc tests as well.

Welcome to GraphPad Prism

GraphPad Prism
Version 8.4.3 (471)

NEW TABLE & GRAPH

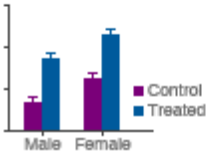
- XY
- Column
- Grouped
- Contingency
- Survival
- Parts of Whole
- Multiple variables
- Nested

EXISTING FILE

- Open a File
- LabArchives
- Clone a Graph
- Graph Portfolio

Grouped tables have two grouping variables, one defined by columns and the other defined by rows

Table format Grouped		A			B		
		Control			Treated		
		A:Y1	A:Y2	A:Y3	B:Y1	B:Y2	B:Y3
1	Male						
2	Femal						



Learn more

Data table:

- (•) Enter or import data into a new table
- ( ) Start with sample data to follow a tutorial

Options:

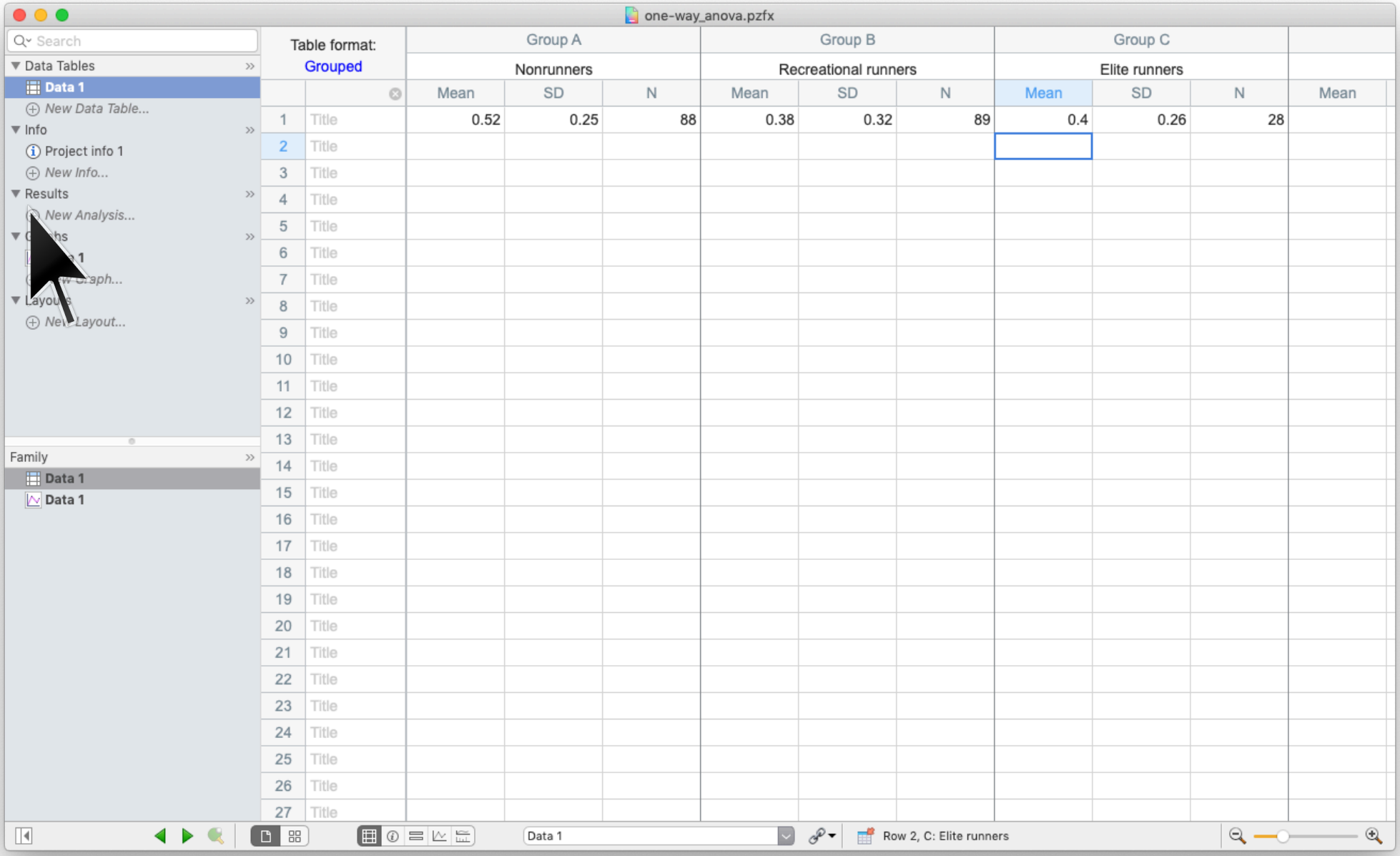
- ( ) Enter and plot a single Y value for each point
- ( ) Enter 3 replicate values in side-by-side subcolumns
- (•) Enter and plot error values already calculated elsewhere

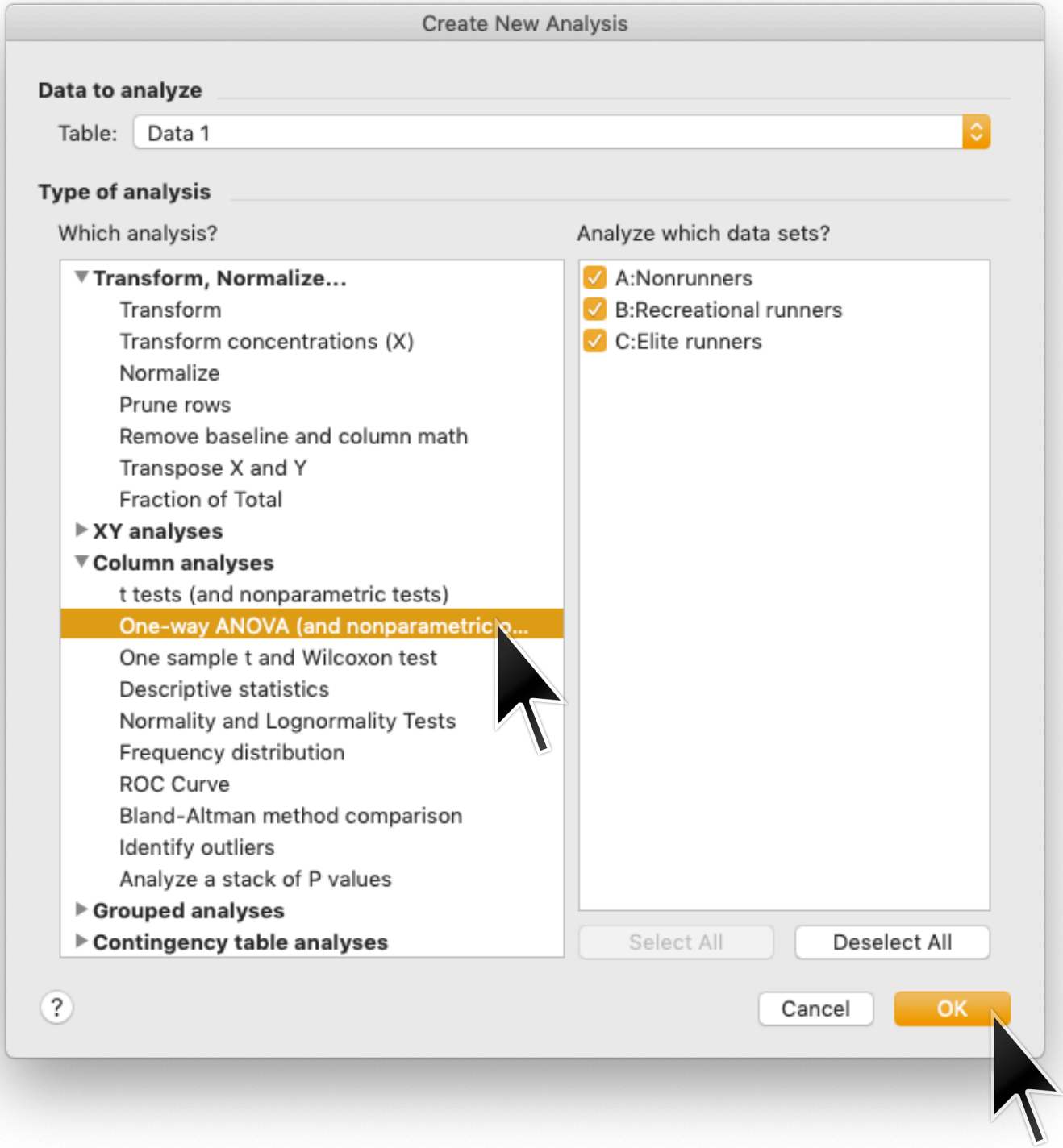
nter: Mean, SD, N

Prism Tips

Cancel

Create


one-way_anova.pzfx
Search
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
New Analysis...
New Layout...
Family
Data 1
Data 1
Table format:
Grouped
Group A
Nonrunners
Group B
Recreational runners
Group C
Elite runners
Mean
SD
N
Mean
SD
N
Mean
SD
N
Mean
1 Title 0.52 0.25 88 0.38 0.32 89 0.4 0.26 28
2 Title
3 Title
4 Title
5 Title
6 Title
7 Title
8 Title
9 Title
10 Title
11 Title
12 Title
13 Title
14 Title
15 Title
16 Title
17 Title
18 Title
19 Title
20 Title
21 Title
22 Title
23 Title
24 Title
25 Title
26 Title
27 Title
Data 1
Row 2, C: Elite runners

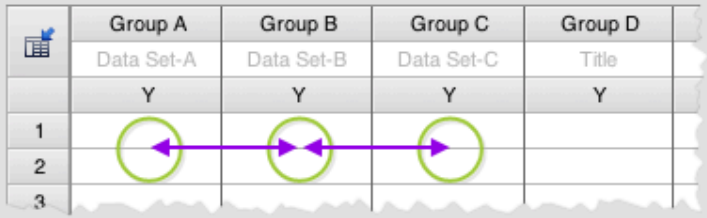

Create New Analysis
Data to analyze
Table:
Data 1
Type of analysis
Which analysis?
Analyze which data sets?
Transform, Normalize...
Transform
Transform concentrations (X)
Normalize
Prune rows
Remove baseline and column math
Transpose X and Y
Fraction of Total
XY analyses
Column analyses
t tests (and nonparametric tests)
One-way ANOVA (and nonparametric...
One sample t and Wilcoxon test
Descriptive statistics
Normality and Lognormality Tests
Frequency distribution
ROC Curve
Bland-Altman method comparison
Identify outliers
Analyze a stack of P values
Grouped analyses
Contingency table analyses
A:Nonrunners
B:Recreational runners
C:Elite runners
Select All
Deselect All
?
Cancel
OK

Parameters: One-Way ANOVA (and Nonparametric or Mixed)

Experimental Design | Repeated Measures | Multiple Comparisons | Options | Residuals

Experimental design

- (•) No matching or pairing
- ( ) Each row represents matched, or repeated measures, data



Assume Gaussian distribution?

- (•) Yes. Use ANOVA.
- ( ) No. Use nonparametric test.

Assume equal SDs?

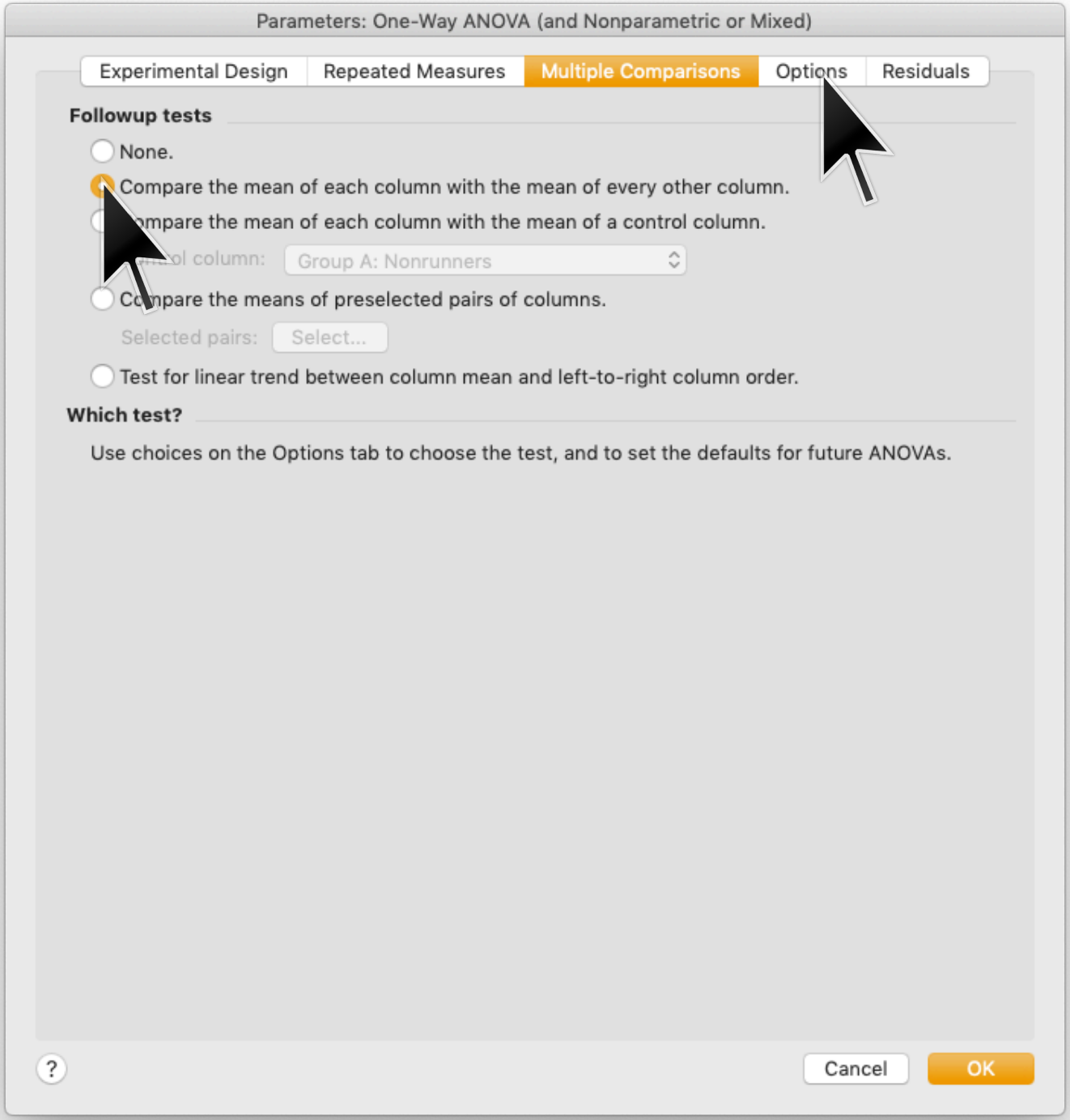
- (•) Yes. Use ordinary ANOVA test.
- ( ) No. Use Brown-Forsythe and Welch ANOVA tests.

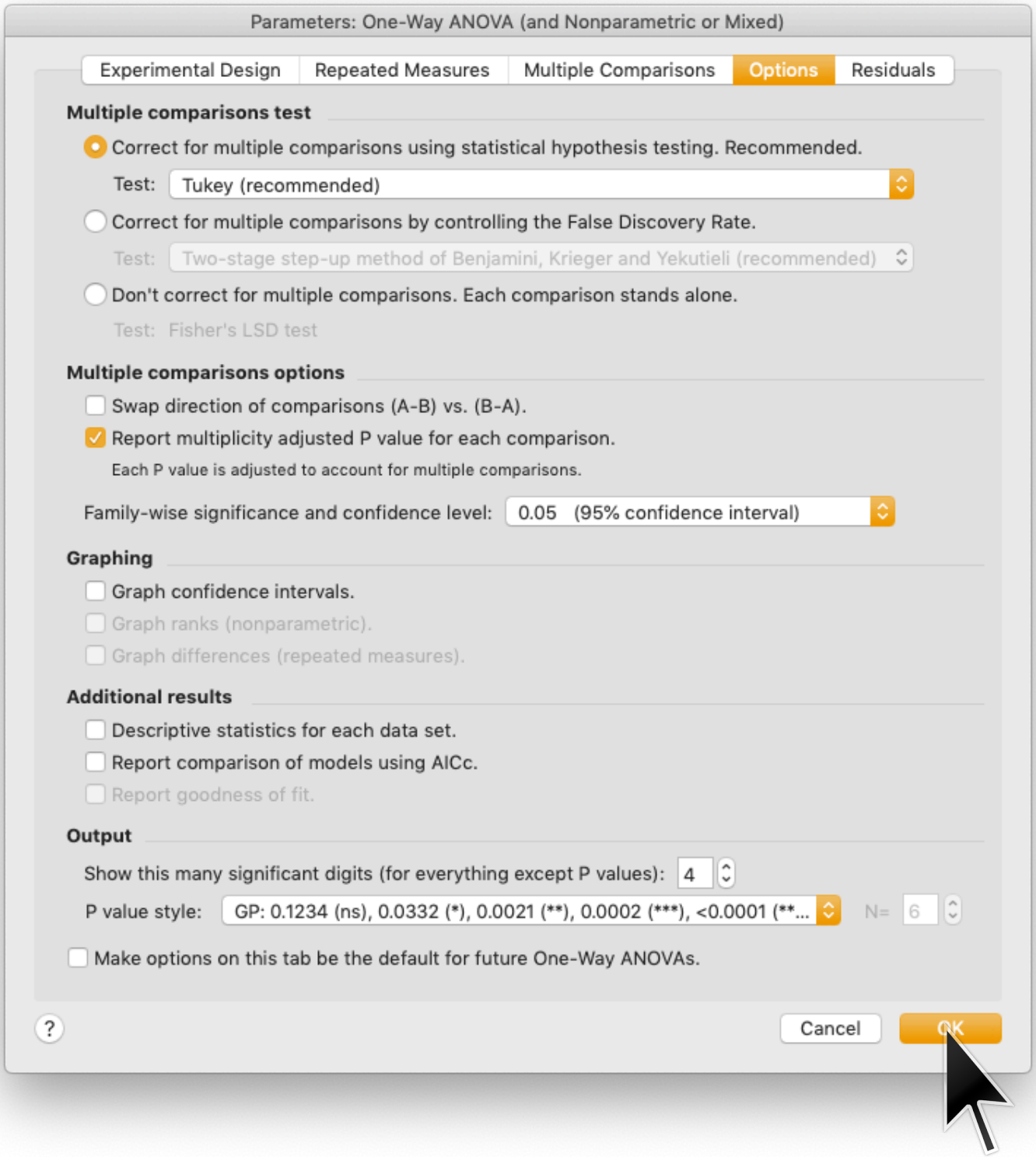
Based on your choices (on all tabs), Prism will perform:

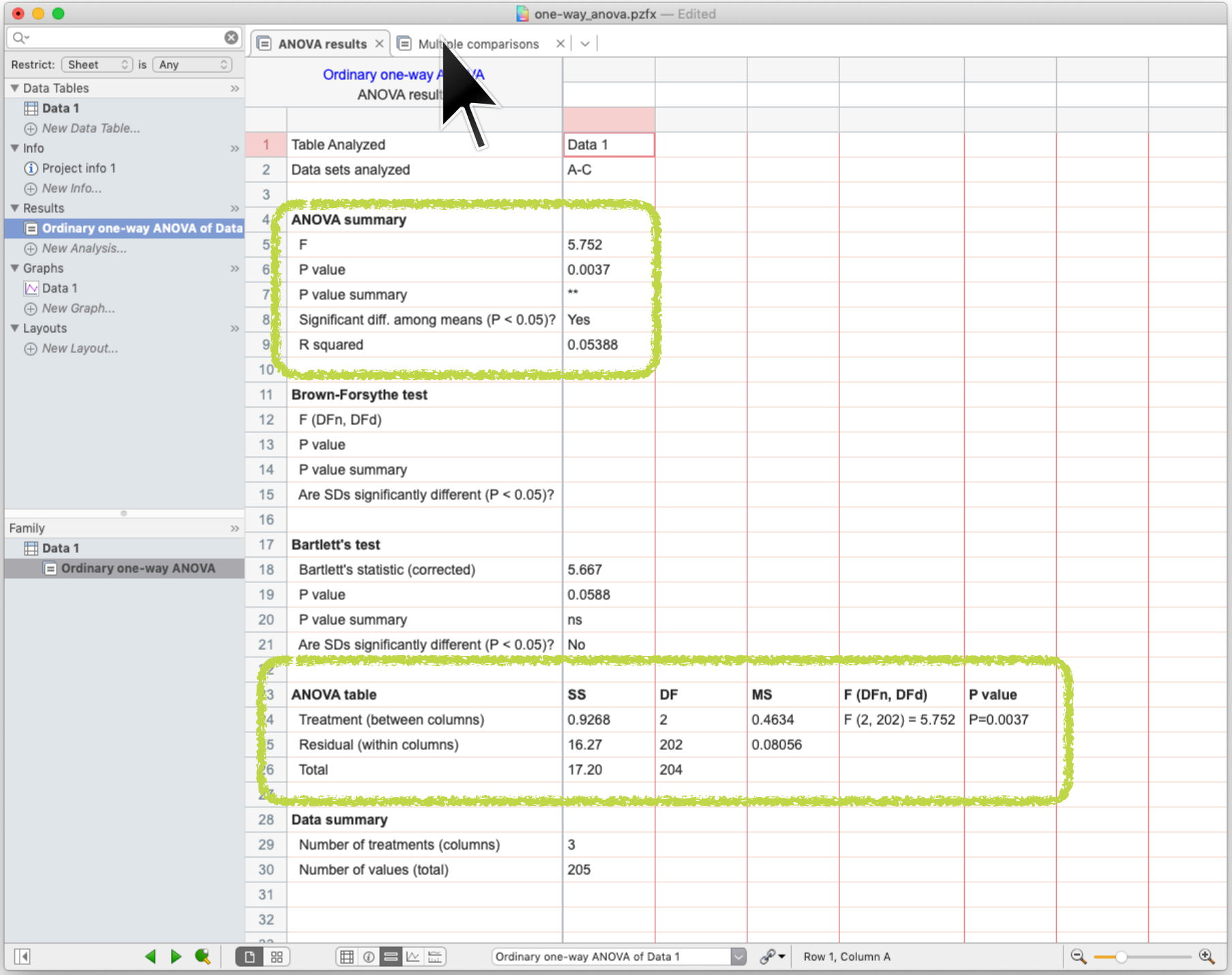
- Ordinary one-way ANOVA.

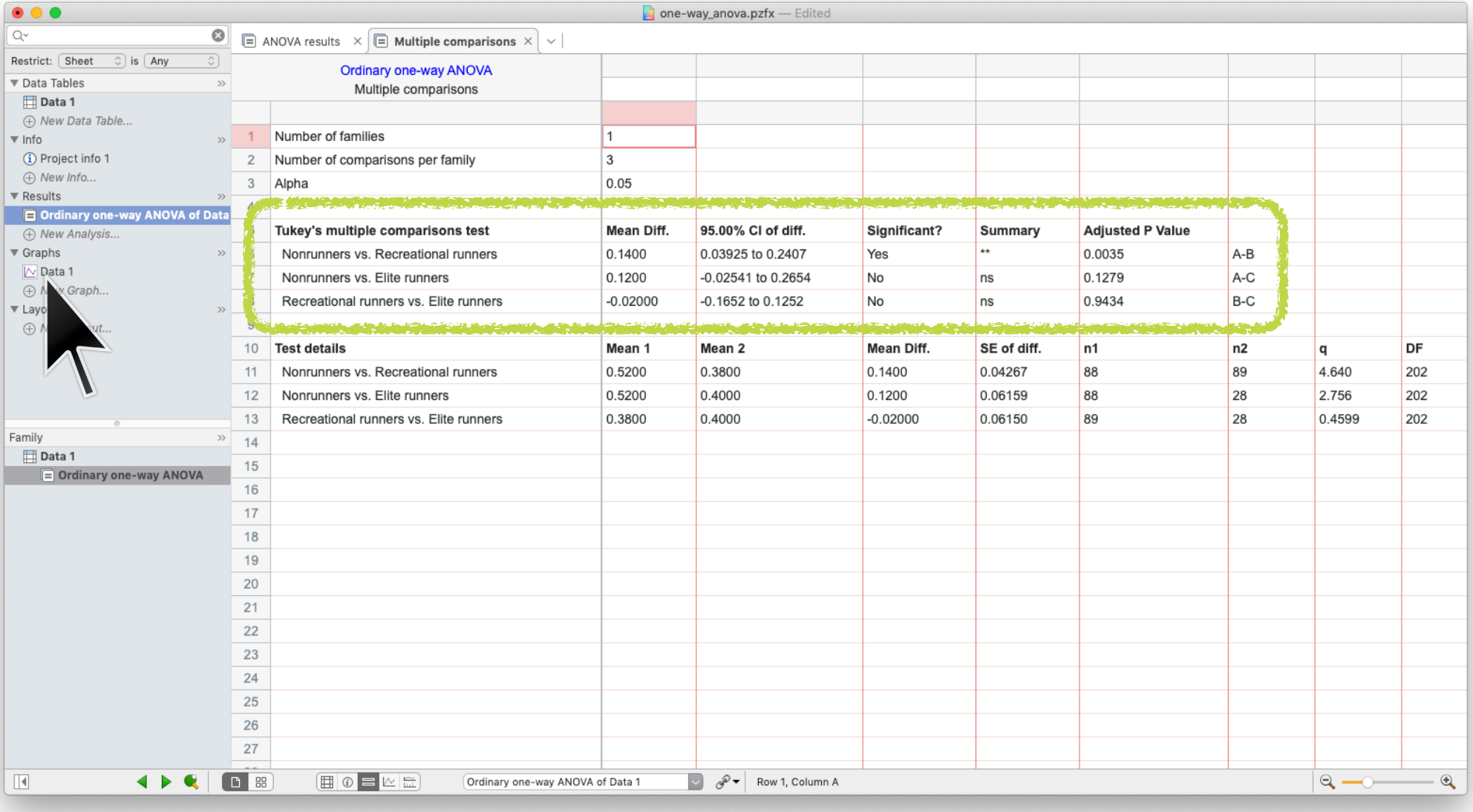
?

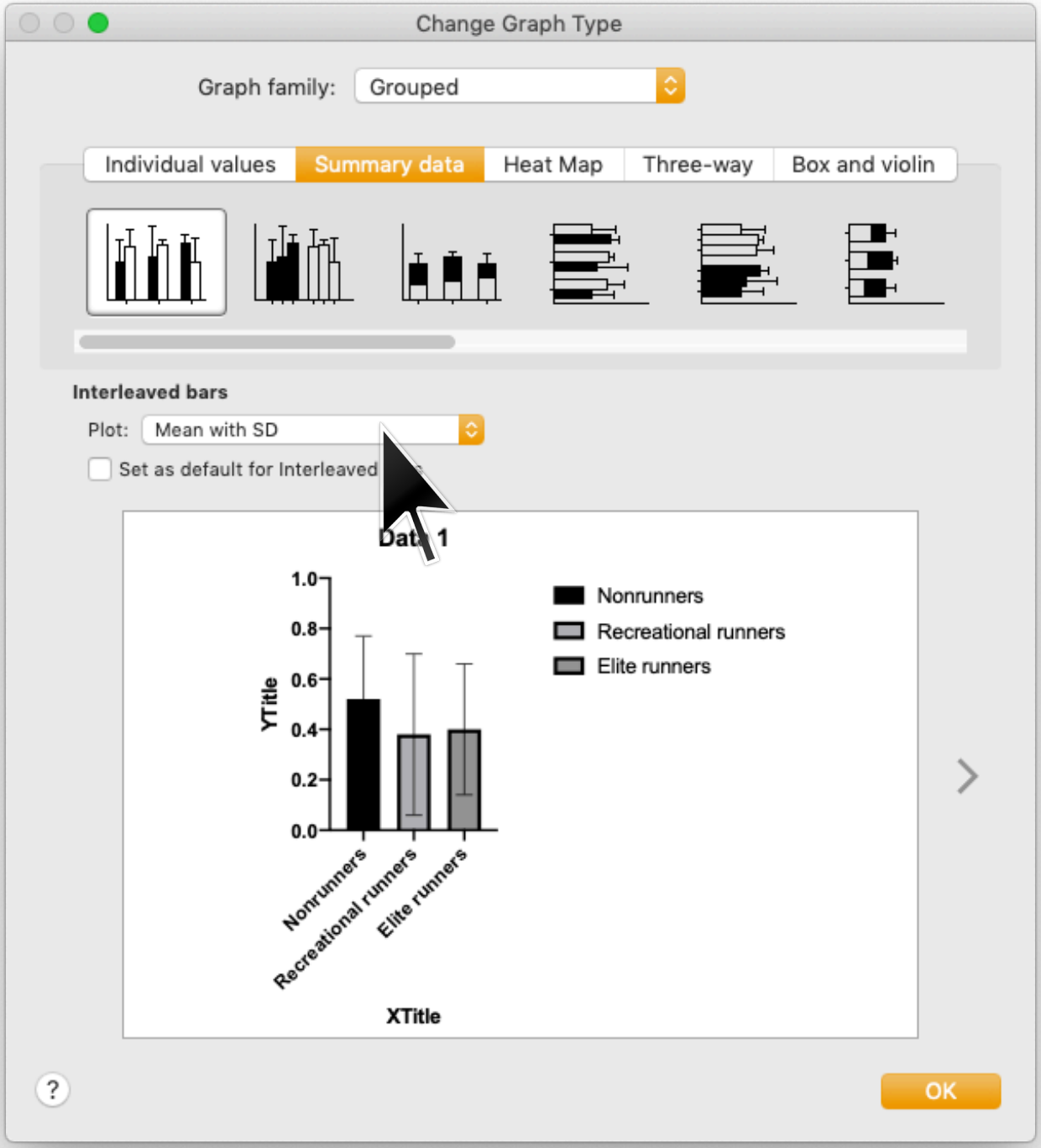
Cancel | OK

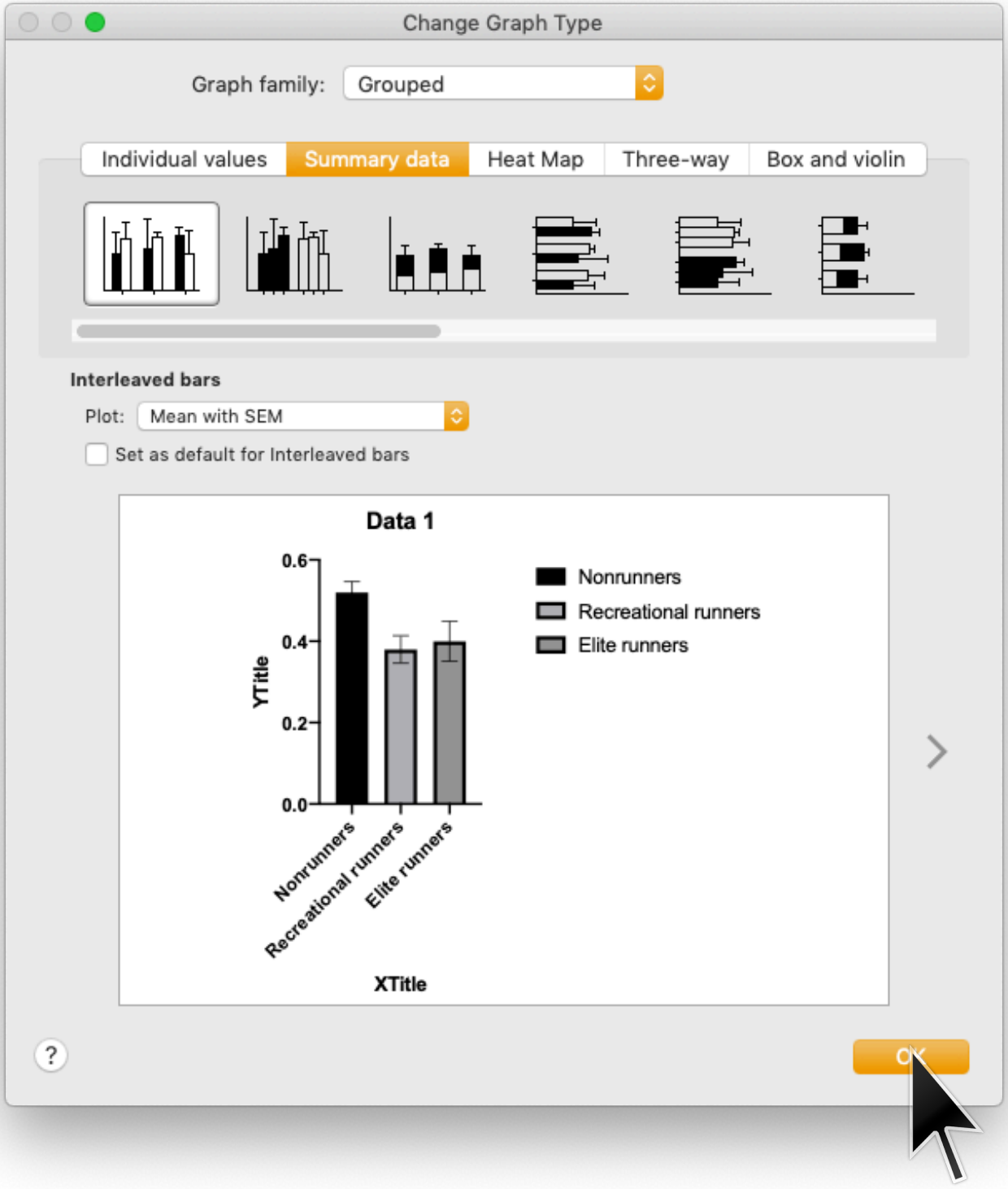

Parameters: One-Way ANOVA (and Nonparametric or Mixed)
Experimental Design
Repeated Measures
Multiple Comparisons
Options
Residuals
Followup tests
None.
Compare the mean of each column with the mean of every other column.
mpare the mean of each column with the mean of a control column.
ol column:
Group A: Nonrunners
Compare the means of preselected pairs of columns.
Selected pairs:
Select...
Test for linear trend between column mean and left-to-right column order.
Which test?
Use choices on the Options tab to choose the test, and to set the defaults for future ANOVAs.
?
Cancel
OK


Parameters: One-Way ANOVA (and Nonparametric or Mixed)
Experimental Design
Repeated Measures
Multiple Comparisons
Options
Residuals
Multiple comparisons test
Correct for multiple comparisons using statistical hypothesis testing. Recommended.
Test: Tukey (recommended)
Correct for multiple comparisons by controlling the False Discovery Rate.
Test: Two-stage step-up method of Benjamini, Krieger and Yekutieli (recommended)
Don't correct for multiple comparisons. Each comparison stands alone.
Test: Fisher's LSD test
Multiple comparisons options
Swap direction of comparisons (A-B) vs. (B-A).
Report multiplicity adjusted P value for each comparison.
Each P value is adjusted to account for multiple comparisons.
Family-wise significance and confidence level: 0.05 (95% confidence interval)
Graphing
Graph confidence intervals.
Graph ranks (nonparametric).
Graph differences (repeated measures).
Additional results
Descriptive statistics for each data set.
Report comparison of models using AICc.
Report goodness of fit.
Output
Show this many significant digits (for everything except P values): 4
P value style: GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.0002 (***), <0.0001 (**...
N= 6
Make options on this tab be the default for future One-Way ANOVAs.
?
Cancel
OK


one-way_anova.pzfx — Edited
ANOVA results
Multiple comparisons
Ordinary one-way ANOVA
ANOVA result
Restrict: Sheet is Any
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
Ordinary one-way ANOVA of Data
New Analysis...
Graphs
Data 1
New Graph...
Layouts
New Layout...
Family
Data 1
Ordinary one-way ANOVA
1 Table Analyzed Data 1
2 Data sets analyzed A-C
4 ANOVA summary
5 F 5.752
6 P value 0.0037
7 P value summary **
8 Significant diff. among means (P < 0.05)? Yes
9 R squared 0.05388
11 Brown-Forsythe test
12 F (DFn, DFd)
13 P value
14 P value summary
15 Are SDs significantly different (P < 0.05)?
17 Bartlett's test
18 Bartlett's statistic (corrected) 5.667
19 P value 0.0588
20 P value summary ns
21 Are SDs significantly different (P < 0.05)? No
ANOVA table SS DF MS F (DFn, DFd) P value
Treatment (between columns) 0.9268 2 0.4634 F (2, 202) = 5.752 P=0.0037
Residual (within columns) 16.27 202 0.08056
Total 17.20 204
28 Data summary
29 Number of treatments (columns) 3
30 Number of values (total) 205
Ordinary one-way ANOVA of Data 1
Row 1, Column A


one-way_anova.pzfx — Edited
ANOVA results
Multiple comparisons
Restrict: Sheet is Any
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
Ordinary one-way ANOVA of Data
New Analysis...
Graphs
Data 1
Family
Data 1
Ordinary one-way ANOVA
Ordinary one-way ANOVA
Multiple comparisons
1 Number of families 1
2 Number of comparisons per family 3
3 Alpha 0.05
Tukey's multiple comparisons test | Mean Diff. | 95.00% CI of diff. | Significant? | Summary | Adjusted P Value
Nonrunners vs. Recreational runners | 0.1400 | 0.03925 to 0.2407 | Yes | ** | 0.0035 | A-B
Nonrunners vs. Elite runners | 0.1200 | -0.02541 to 0.2654 | No | ns | 0.1279 | A-C
Recreational runners vs. Elite runners | -0.02000 | -0.1652 to 0.1252 | No | ns | 0.9434 | B-C
10 Test details | Mean 1 | Mean 2 | Mean Diff. | SE of diff. | n1 | n2 | q | DF
11 Nonrunners vs. Recreational runners | 0.5200 | 0.3800 | 0.1400 | 0.04267 | 88 | 89 | 4.640 | 202
12 Nonrunners vs. Elite runners | 0.5200 | 0.4000 | 0.1200 | 0.06159 | 88 | 28 | 2.756 | 202
13 Recreational runners vs. Elite runners | 0.3800 | 0.4000 | -0.02000 | 0.06150 | 89 | 28 | 0.4599 | 202
Ordinary one-way ANOVA of Data 1
Row 1, Column A


Change Graph Type
Graph family:
Grouped
Individual values
Summary data
Heat Map
Three-way
Box and violin
Interleaved bars
Plot:
Mean with SD
Set as default for Interleaved
Data 1
1.0
0.8
0.6
0.4
0.2
0.0
YTitle
Nonrunners
Recreational runners
Elite runners
XTitle
Nonrunners
Recreational runners
Elite runners
?
OK

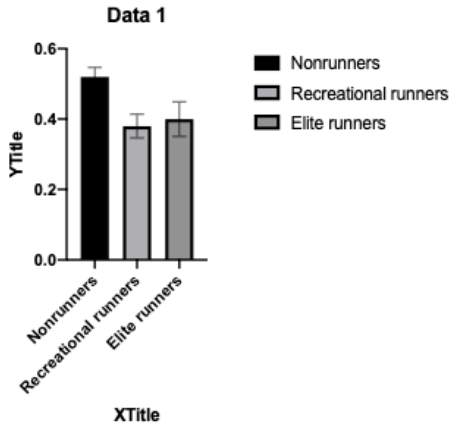

Change Graph Type
Graph family:
Grouped
Individual values
Summary data
Heat Map
Three-way
Box and violin
Interleaved bars
Plot:
Mean with SEM
Set as default for Interleaved bars
Data 1
0.6
0.4
0.2
0.0
YTitle
Nonrunners
Recreational runners
Elite runners
XTitle
Nonrunners
Recreational runners
Elite runners
?
OK

one-way_anova.pzfx — Edited

Restrict: Sheet is Any

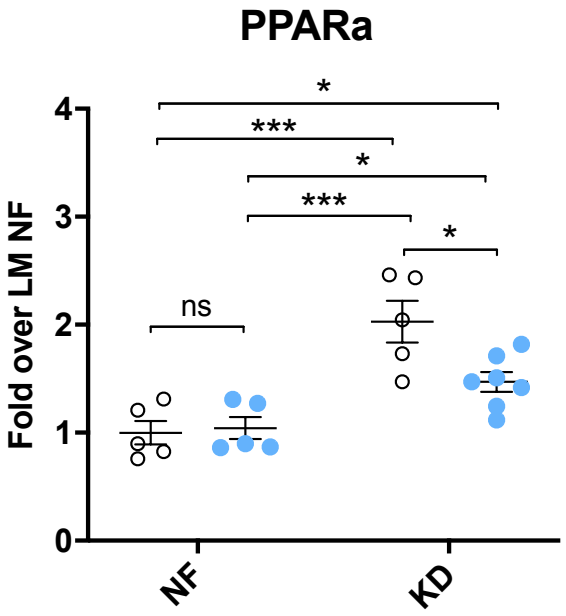
- Data Tables
  - Data 1
  - *New Data Table...*
- Info
  - Project info 1
  - *New Info...*
- Results
  - Ordinary one-way ANOVA of Data 1
  - *New Analysis...*
- Graphs
  - Data 1
  - *New Graph...*
- Layouts
  - *New Layout...*

Family
- Data 1
- Data 1



Data 1

Two-way ANOVA tests whether to see if there is an interaction between groups



LM
C26

y_i = PPARa mRNA expression

x_{i1} = cancer presence (C26=tumor, LM=litter mate)

x_{i2} = food (NF=normal, KD=ketogenic)

(data curtsey of Tobias Janowitz)

Null model: $y_i = \beta_0 + \epsilon_i$

Alternative model #1: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

Alternative model #2: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \epsilon_i$

interaction term

Welcome to GraphPad Prism

GraphPad Prism
Version 8.4.3 (471)

NEW TABLE & GRAPH

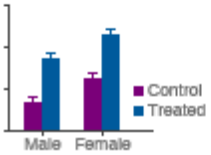
- XY
- Column
- Grouped
- Contingency
- Survival
- Parts of Whole
- Multiple variables
- Nested

EXISTING FILE

- Open a File
- LabArchives
- Clone a Graph
- Graph Portfolio

Grouped tables have two grouping variables, one defined by columns and the other defined by rows

Table format Grouped		A			B		
		Control			Treated		
		A:Y1	A:Y2	A:Y3	B:Y1	B:Y2	B:Y3
1	Male						
2	Femal						



Learn more

Data table:

- (•) Enter or import data into a new table
- ( ) Start with sample data to follow a tutorial

Options:

- ( ) Enter and plot a single Y value for each point
- (•) Enter 7 replicate values in side-by-side subcolumns
- ( ) Enter and plot error values already calculated elsewhere

Enter: Mean, SD, N

Prism Tips

Cancel Create

two-way_anova.pzfx

Search

- Data Tables
  - PPARa
  - *New Data Table...*
- Info
  - *New Info...*
- Results
  - 2way ANOVA of PPARa
  - *New Analysis...*
- Graphs
  - PPARa
  - *New Graph...*
- Layouts
  - *New Layout...*

Family

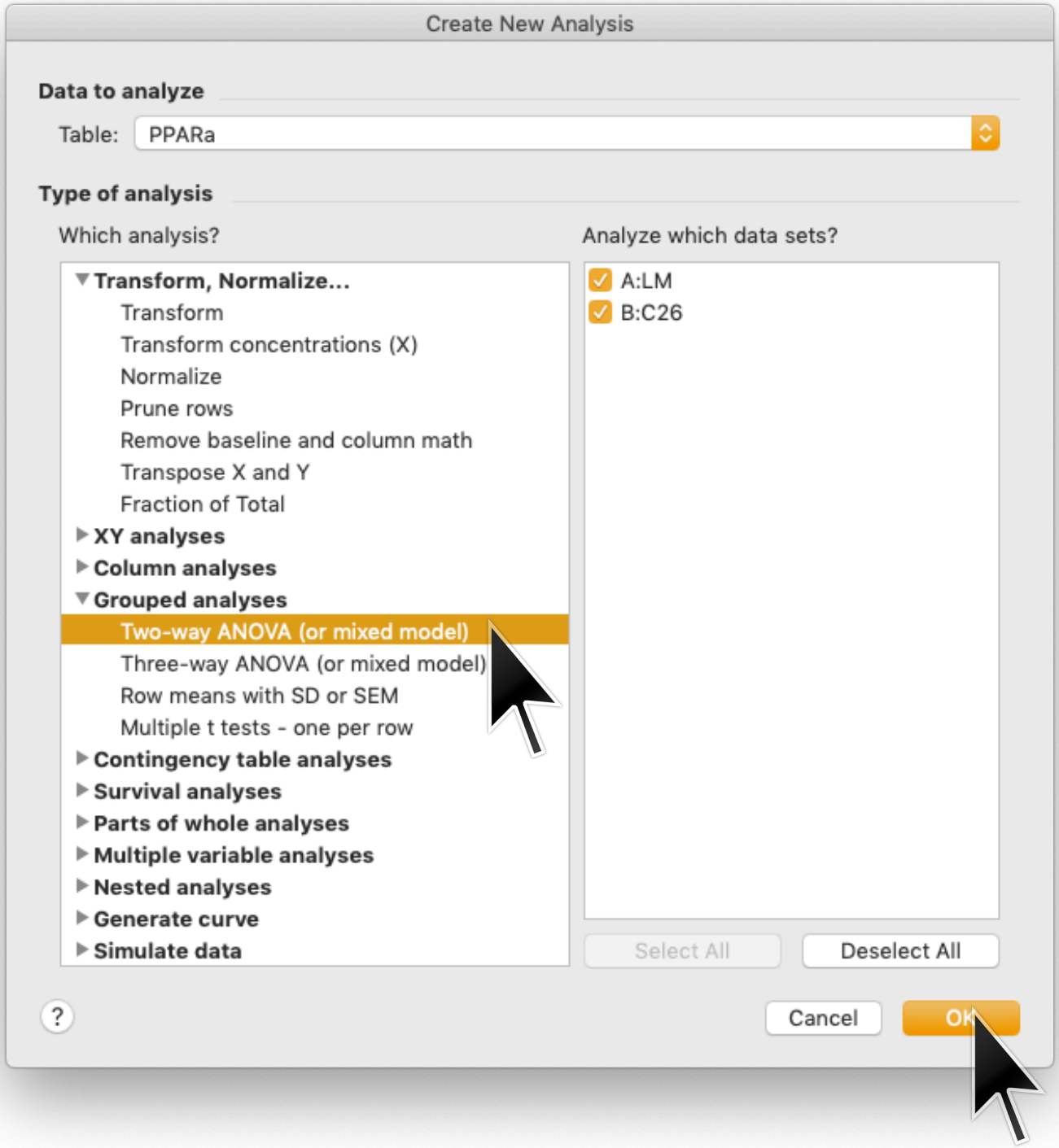
- PPARa
  - 2way ANOVA
- PPARa

Table format: Grouped

		Group A							Group B									
		LM							C26									
		A:Y1	A:Y2	A:Y3	A:Y4	A:Y5	A:Y6	A:Y7	B:Y1	B:Y2	B:Y3	B:Y4	B:Y5	B:Y6	B:Y7	C:Y1	C:Y2	C:
1	NF	0.896878	0.757779	1.209183	0.824336	1.311823			0.861320	0.868462	0.899686	1.307027	1.272415					
2	KD	2.435268	2.045139	2.460515	1.472005	1.732875			1.119927	1.244879	1.509778	1.416881	1.819409	1.710366	1.470989			
3	Title																	
4	Title																	
5	Title																	
6	Title																	
7	Title																	
8	Title																	
9	Title																	
10	Title																	
11	Title																	
12	Title																	
13	Title																	
14	Title																	
15	Title																	
16	Title																	
17	Title																	
18	Title																	
19	Title																	
20	Title																	
21	Title																	
22	Title																	
23	Title																	
24	Title																	
25	Title																	
26	Title																	
27	Title																	
28	Title																	
29	Title																	
30	Title																	
31	Title																	

PPARa

Row 2, B: C26

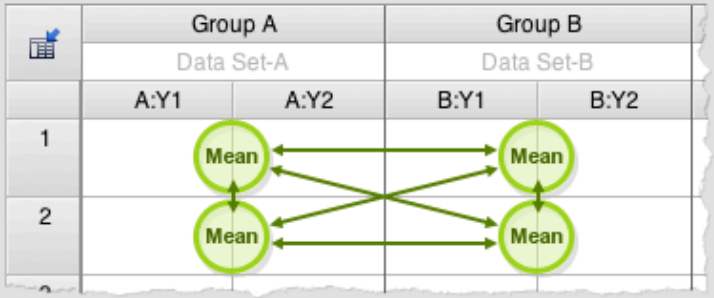

Create New Analysis
Data to analyze
Table:
PPARa
Type of analysis
Which analysis?
Analyze which data sets?
Transform, Normalize...
Transform
Transform concentrations (X)
Normalize
Prune rows
Remove baseline and column math
Transpose X and Y
Fraction of Total
XY analyses
Column analyses
Grouped analyses
Two-way ANOVA (or mixed model)
Three-way ANOVA (or mixed model)
Row means with SD or SEM
Multiple t tests - one per row
Contingency table analyses
Survival analyses
Parts of whole analyses
Multiple variable analyses
Nested analyses
Generate curve
Simulate data
A:LM
B:C26
Select All
Deselect All
?
Cancel
OK

Parameters: Two-Way ANOVA (or Mixed Model)

RM Design | RM Analysis | Factor Names | Multiple Comparisons | Options | Residuals

What kind of comparison?

Compare cell means regardless of rows and columns



How many comparisons?

- (•) Compare each cell mean with every other cell mean.
- ( ) Compare each cell mean with the control (upper-left) cell mean.

Control cell: LM : NF

How many families?

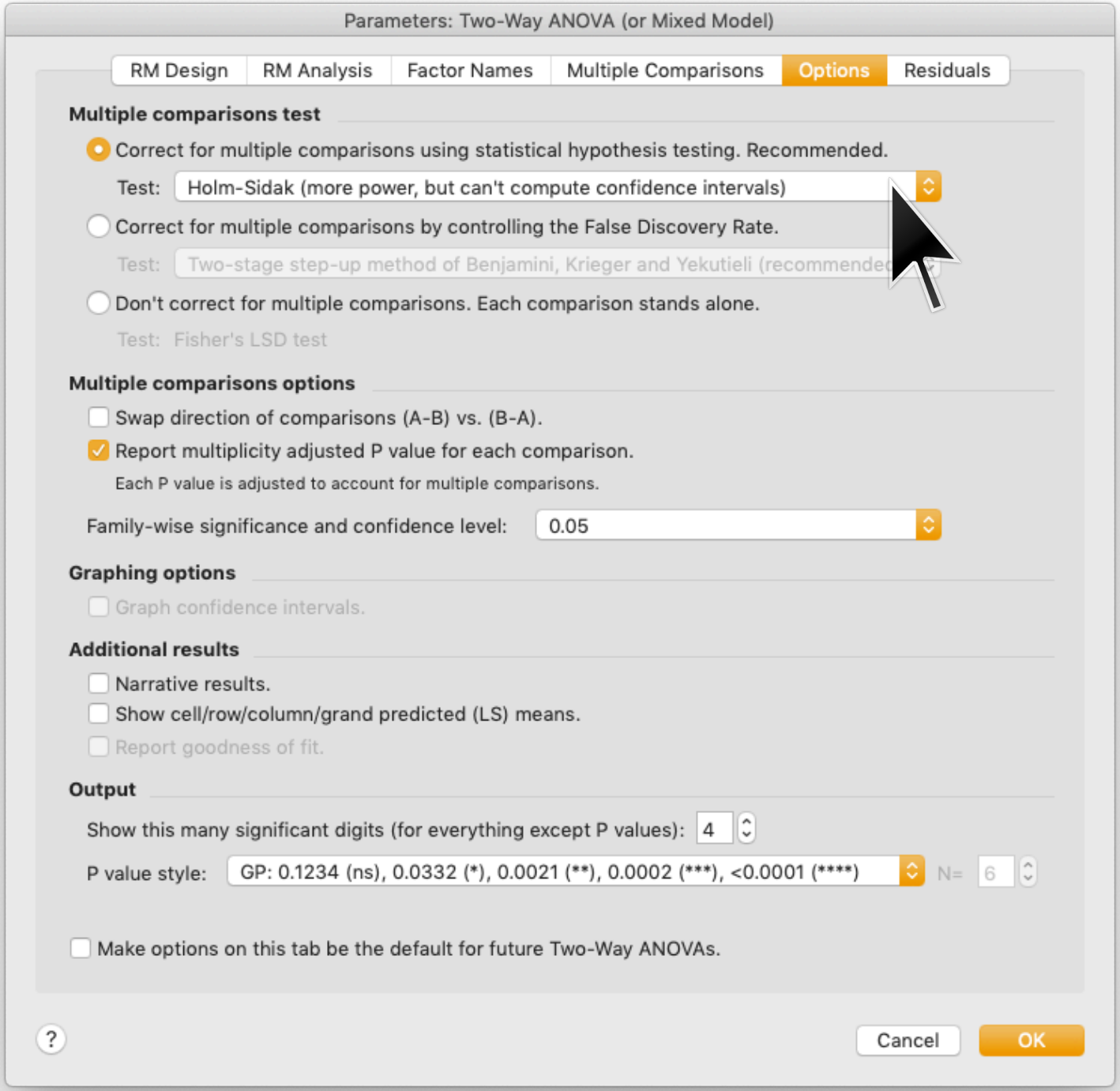
One family for all the comparisons

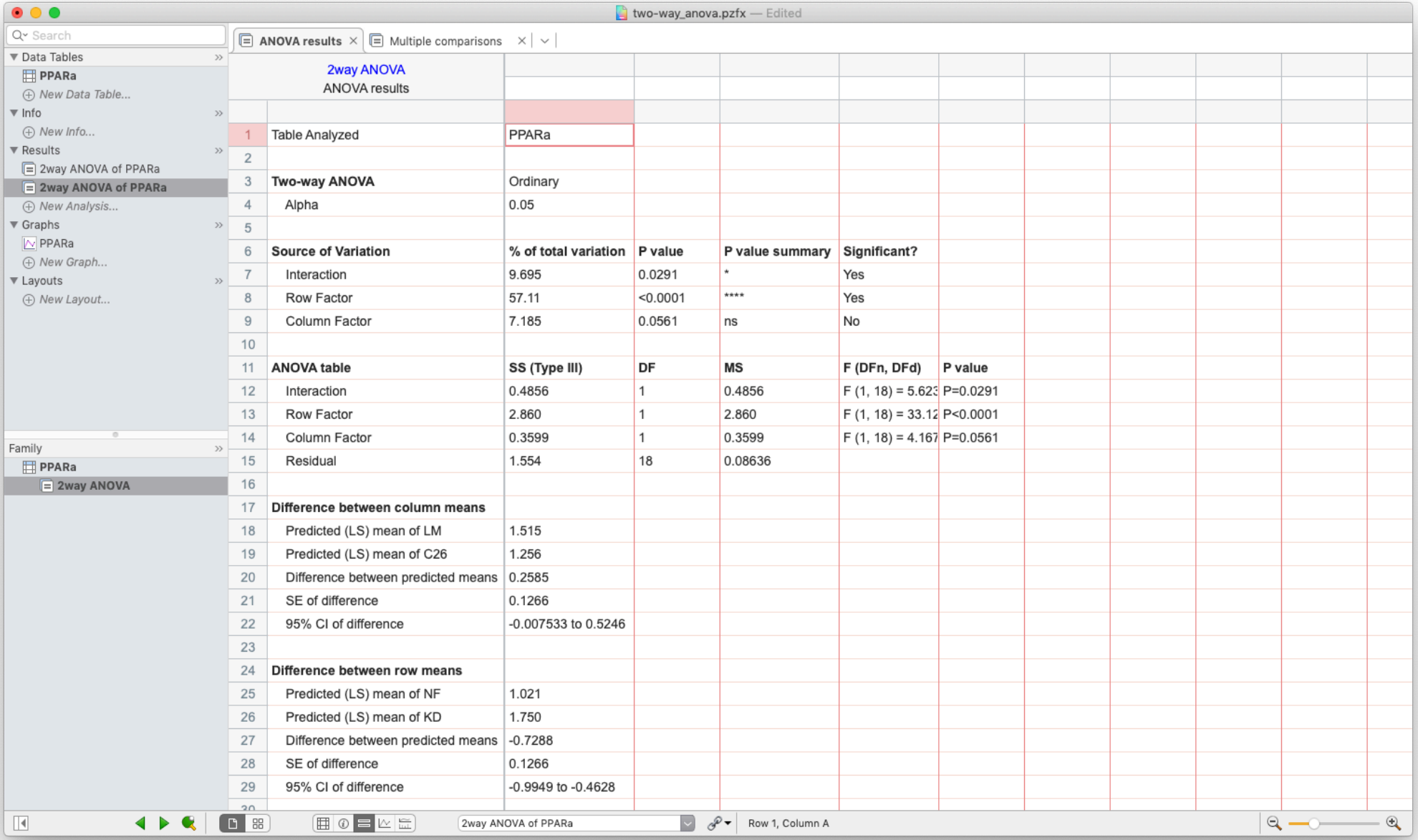
Which test?

Use choices on the Options tab to choose the test, and to set the defaults for future ANOVAs.

?

Cancel | OK


Parameters: Two-Way ANOVA (or Mixed Model)
RM Design
RM Analysis
Factor Names
Multiple Comparisons
Options
Residuals
Multiple comparisons test
Correct for multiple comparisons using statistical hypothesis testing. Recommended.
Test: Holm-Sidak (more power, but can't compute confidence intervals)
Correct for multiple comparisons by controlling the False Discovery Rate.
Test: Two-stage step-up method of Benjamini, Krieger and Yekutieli (recommended)
Don't correct for multiple comparisons. Each comparison stands alone.
Test: Fisher's LSD test
Multiple comparisons options
Swap direction of comparisons (A-B) vs. (B-A).
Report multiplicity adjusted P value for each comparison.
Each P value is adjusted to account for multiple comparisons.
Family-wise significance and confidence level: 0.05
Graphing options
Graph confidence intervals.
Additional results
Narrative results.
Show cell/row/column/grand predicted (LS) means.
Report goodness of fit.
Output
Show this many significant digits (for everything except P values): 4
P value style: GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.0002 (***), <0.0001 (****)
N= 6
Make options on this tab be the default for future Two-Way ANOVAs.
?
Cancel
OK



1	Table Analyzed	PPARa				
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	9.695	0.0291	*	Yes	
8	Row Factor	57.11	<0.0001	****	Yes	
9	Column Factor	7.185	0.0561	ns	No	
10						
11	**ANOVA table**	**SS (Type III)**	**DF**	**MS**	**F (DFn, DFd)**	**P value**
12	Interaction	0.4856	1	0.4856	F (1, 18) = 5.623	P=0.0291
13	Row Factor	2.860	1	2.860	F (1, 18) = 33.12	P<0.0001
14	Column Factor	0.3599	1	0.3599	F (1, 18) = 4.167	P=0.0561
15	Residual	1.554	18	0.08636		
16						
17	**Difference between column means**					
18	Predicted (LS) mean of LM	1.515				
19	Predicted (LS) mean of C26	1.256				
20	Difference between predicted means	0.2585				
21	SE of difference	0.1266				
22	95% CI of difference	-0.007533 to 0.5246				
23						
24	**Difference between row means**					
25	Predicted (LS) mean of NF	1.021				
26	Predicted (LS) mean of KD	1.750				
27	Difference between predicted means	-0.7288				
28	SE of difference	0.1266				
29	95% CI of difference	-0.9949 to -0.4628				

two-way_anova.pzfx — Edited

Search

Data Tables
- PPARa
- New Data Table...

Info
- New Info...

Results
- 2way ANOVA of PPARa
- New Analysis...

Graphs
- PPARa
- New Graph...

Layouts
- New Layout...

Family
- PPARa
- 2way ANOVA

ANOVA results | Multiple comparisons

2way ANOVA
Multiple comparisons

1	Compare cell means regardless of rows and columns										
2											
3	Number of families	1									
4	Number of comparisons per family	6									
5	Alpha	0.05									
6											
7	**Holm-Sidak's multiple comparisons test**	**Predicted (LS) mean diff.**	**Significant?**	**Summary**	**Adjusted P Value**						
8											
9	NF:LM vs. NF:C26	-0.04178	No	ns	0.8247						
10	NF:LM vs. KD:LM	-1.029	Yes	***	0.0002						
11	NF:LM vs. KD:C26	-0.4703	Yes	*	0.0404						
12	NF:C26 vs. KD:LM	-0.9874	Yes	***	0.0002						
13	NF:C26 vs. KD:C26	-0.4285	Yes	*	0.0450						
14	KD:LM vs. KD:C26	0.5588	Yes	*	0.0178						
15											
16											
17	**Test details**	**Predicted (LS) mean 1**	**Predicted (LS) mean 2**	**Predicted (LS) mean diff.**	**SE of diff.**	**N1**	**N2**	**t**	**DF**		
18											
19	NF:LM vs. NF:C26	1.000	1.042	-0.04178	0.1859	5	5	0.2248	18.00		
20	NF:LM vs. KD:LM	1.000	2.029	-1.029	0.1859	5	5	5.537	18.00		
21	NF:LM vs. KD:C26	1.000	1.470	-0.4703	0.1721	5	7	2.733	18.00		
22	NF:C26 vs. KD:LM	1.042	2.029	-0.9874	0.1859	5	5	5.313	18.00		
23	NF:C26 vs. KD:C26	1.042	1.470	-0.4285	0.1721	5	7	2.490	18.00		
24	KD:LM vs. KD:C26	2.029	1.470	0.5588	0.1721	5	7	3.248	18.00		
25											
26											
27											
28											
29											

2way ANOVA of PPARa

Row 1, Column A

two-way_anova.pzfx — Edited

Search

Data Tables
- PPARa
- New Data Table...

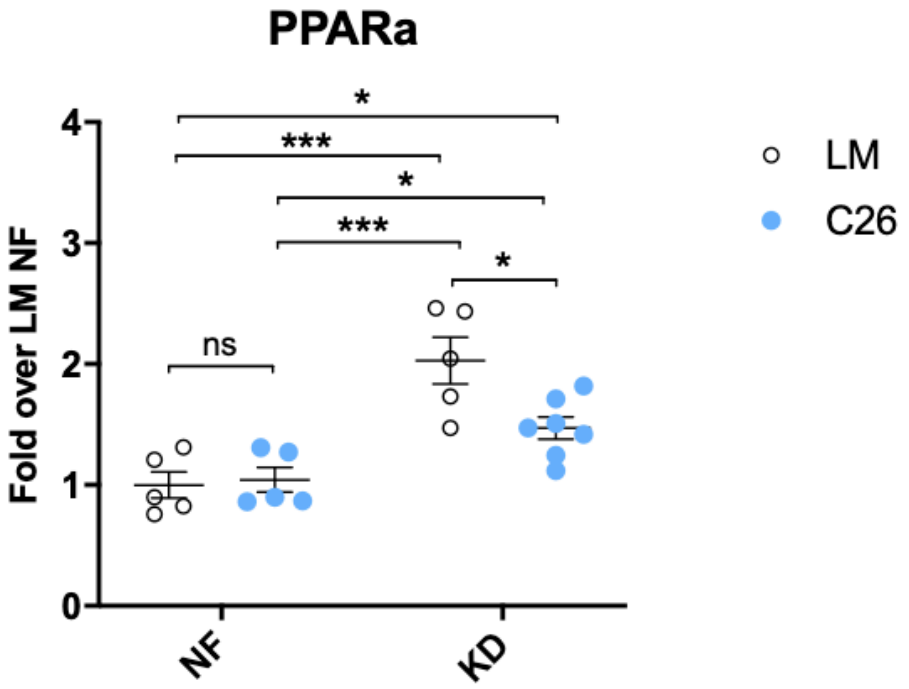
Info
- New Info...

Results
- 2way ANOVA of PPARa
- New Analysis...

Graphs
- PPARa
- New Graph...

Layouts
- New Layout...

Family
- PPARa
- PPARa



PPARa

Survival analysis

The Survival function $S(t)$

Uppercase T indicates the time of an individual's death. This is a random variable that changes from individual to individual. Alternatively, T can be the time of some other event an individual can experience once and only once. Not all individuals under study need to experience this event.

Lowercase t denotes a time value that we wish to inquire about; it is not specific to any individual.

The survival function $S(t)$ is the probability of survival to time t, i.e.

$$S(t) = p(T > t)$$

Here are some properties of the survival function:

1. $S(0) = 1$ (by convention)
2. $0 \leq S(t) \leq 1$ at all times t
3. $S(t)$ is a non-increasing function of t

The hazard function $h(t)$

The hazard function $h(t)$ is the probability of death per unit time (i.e. death rate) at time t, given that a subject has already survived up until time t.

The hazard function and the survival function are related to each other via

$$S(t) = \exp\left(-\int_0^t dt' h(t')\right) \quad \text{and} \quad h(t) = -\frac{d}{dt}\log S(t).$$

The cumulative hazard function $H(t)$ is the integral of the hazard function:

$$H(t) = \int_0^t dt' \; h(t'),$$

which is related to the survival function via $S(t) = e^{-H(t)}$.

Estimating the survival function: no censoring

The survival function is usually the primary thing we are interested in estimating from data. Suppose we have n individuals who are all alive at time $t = 0$. Further assume that we observe all death events that do occur. We can then estimate $S(t)$ quite simply as the fraction of these individuals who remain alive at time t .

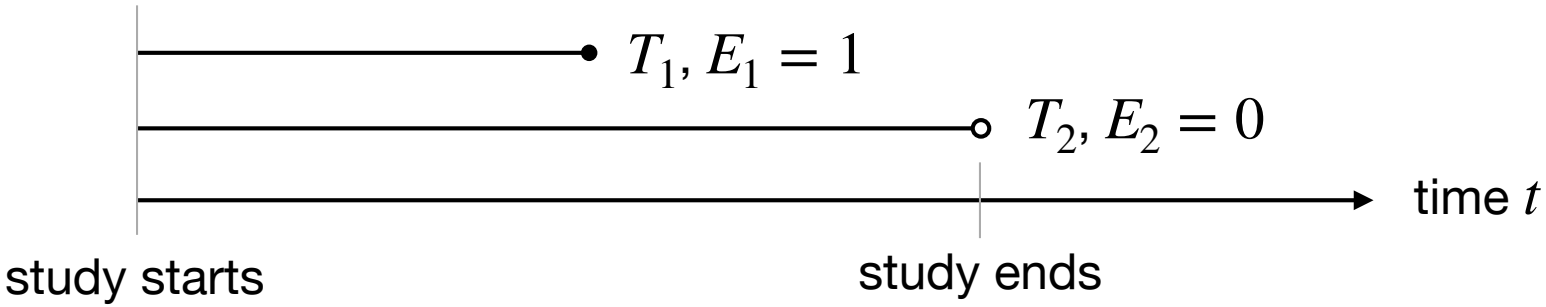
$$\hat{S}(t) = \frac{n(t)}{n(0)}$$

where $n(t)$ is the number of subjects alive at time t.

Right censoring

Survival data is “right-censored” when we know that an individual i survived up to time T_i, but after that we loose track of that individual.

Censoring is usually indicated by an event flag E_i that is 1 if the event is observed or 0 if the event is censored.



Censoring occurs for many different reasons

Censoring can occur for many different reasons.

1. Subjects enroll in a clinical trial on a rolling basis, and survival time is computed from the date of enrollment. When the trial ends, the subjects how still survive will have survived for different periods of time.

2. Subjects in a clinical trial leave because they don't want to participate anymore, they require protocol-breaking treatment, or they are lost to follow-up.

3. In an animal study, animals become available for experimentation at different times.

4. An animal in a study is subject to some unexpected mishap (lost, etc.)

Do not throw away censored data! This will invalidate your entire analysis.

The Kaplan-Meier estimator is the standard way to estimate survival curves

Let $T_1, T_2, \ldots, T_K$. , be the times, in increasing order at which individuals either die or are censored. We allow for multiple individuals dying and/or being censored at the same time.

Let n_i denote the number of individuals at risk at time T_i.

Let d_i denote the number of individuals that actually die at time T_i.

The Kaplan-Meier estimate $\hat{S}(t)$ for the survival curve is given by:

$$\hat{S}(t) = \prod_{i\,:\,T_i<t} \frac{n_i - d_i}{n_i}.$$

Use the log-rank test to compare two survival curves

The log-rank test is (also called the Mantel-Cox test) is the standard test used to compare survival curves for two distinct groups

Null hypothesis: the two populations are governed by the same survival curve and hazard rate

How it works: computes a summary statistic that quantifies how evenly distributed deaths are across the populations in question. Under the null hypothesis, this statistic approximately follows a χ^2 distribution with 1 degree of freedom.

Lymph Node Removal in Treating Women Who Have Stage I or Stage IIA Breast Cancer

ClinicalTrials.gov Identifier: NCT00003855

Recruitment Status : Completed
First Posted : January 27, 2003
Last Update Posted : April 29, 2020

Study Description

Go to

Brief Summary:

RATIONALE: Surgery to remove lymph nodes in the armpit may remove cancer cells that have spread from tumors in the breast.

PURPOSE: Randomized phase III trial to determine the effectiveness of removing lymph nodes in the armpit in treating women who have stage I or stage IIA breast cancer.

Condition or disease	Intervention/treatment	Phase
Breast Cancer	Procedure: axillary lymph node dissection Radiation: whole breast irradiation	Phase 3

Detailed Description:

OBJECTIVES:

Primary objectives:

Long term: To assess whether overall survival for patients randomized to Arm 2 (no immediate ALND) is essentially equivalent to (or better than) than that for patients assigned to Arm 1 (completion ALND).

Short term: To quantify and compare the surgical morbidities associated with SLND plus ALND versus SLND alone.

OUTLINE: This is a randomized study. After segmental mastectomy and sentinel lymph node dissection, patients are stratified according to age (50 and under vs over 50), estrogen receptor status (positive vs negative), and tumor size (no greater than 1 cm vs greater than 1 cm but no greater than 2 cm vs greater than 2 cm). Patients are randomized to one of two treatment arms.

Welcome to GraphPad Prism

GraphPad Prism
Version 8.4.3 (471)

NEW TABLE & GRAPH

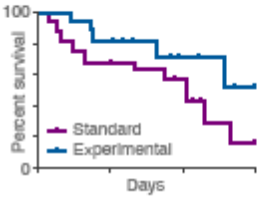
- XY
- Column
- Grouped
- Contingency
- Survival
- Parts of Whole
- Multiple variables
- Nested

EXISTING FILE

- Open a File
- LabArchives
- Clone a Graph
- Graph Portfolio

Survival tables: Each row tabulates the survival or censored time of a subject

Table format: Survival		X	A
		Days	Standard
		X	Y
1	Title		
2	Title		
3	Title		
4	Title		



? Learn more

Data table:

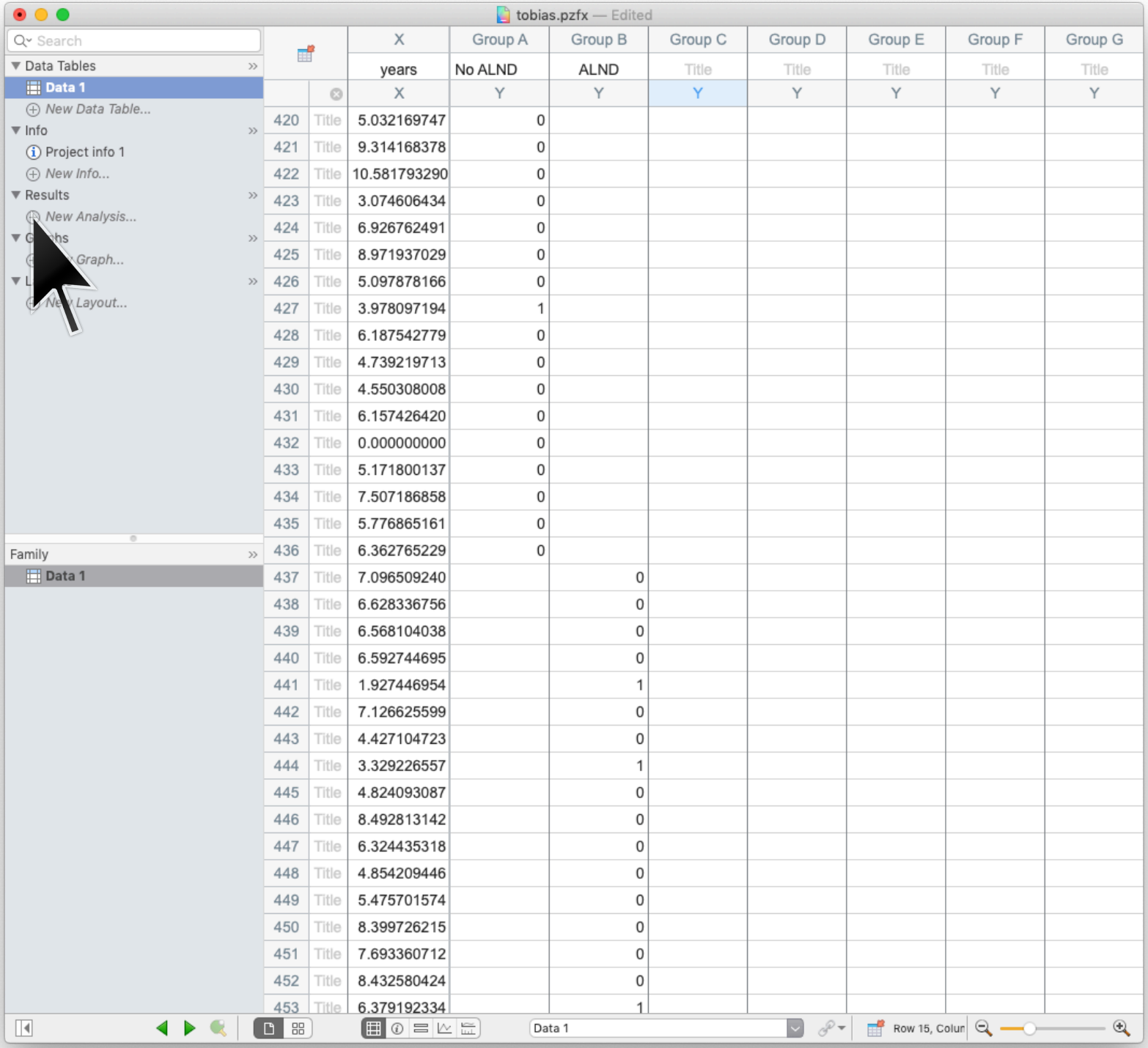
- ○ Enter or import data into a new table
- ◉ Start with sample data to follow a tutorial

Select a tutorial data set:

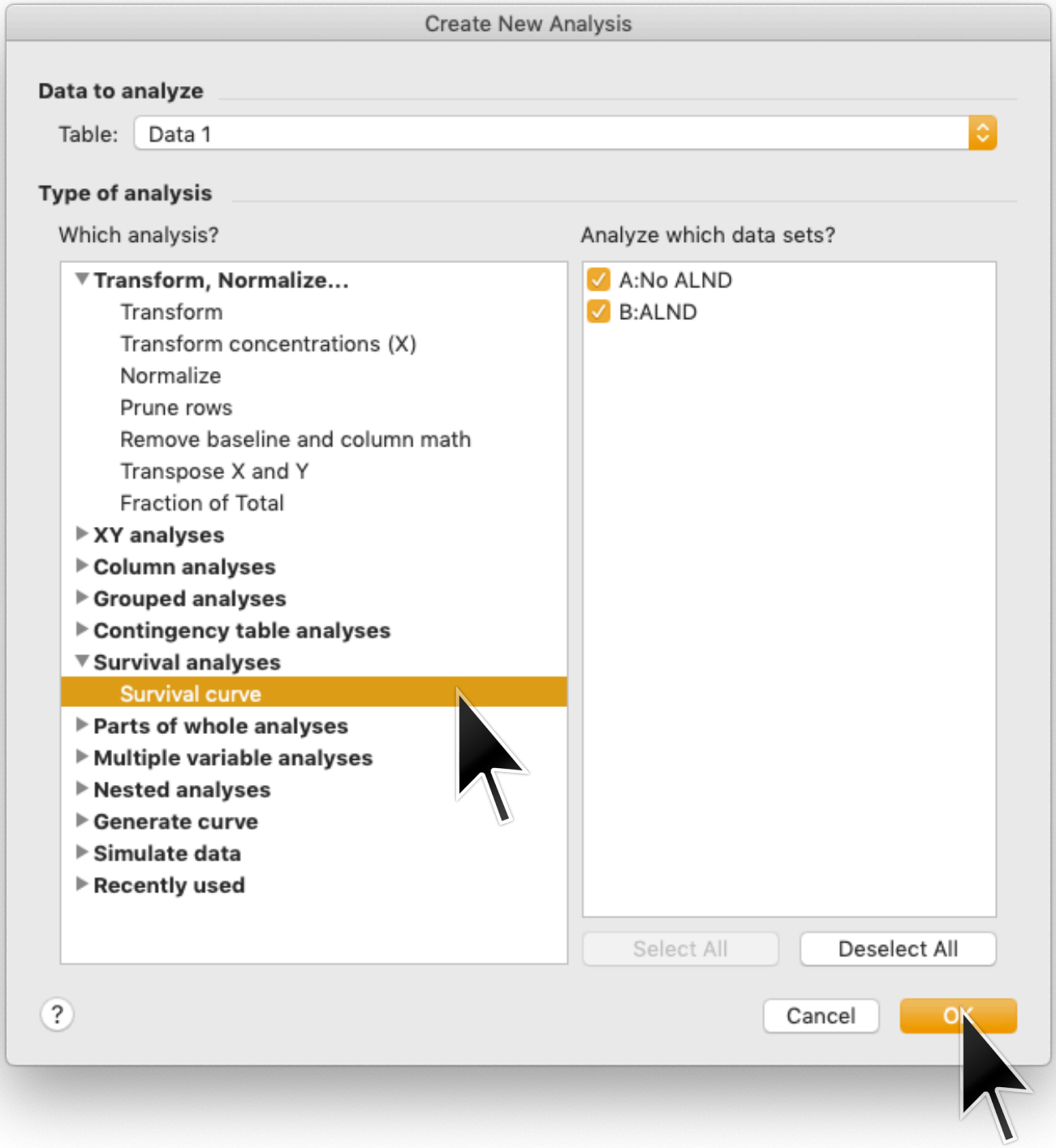
- ◉ Comparing two groups
- ○ Three groups

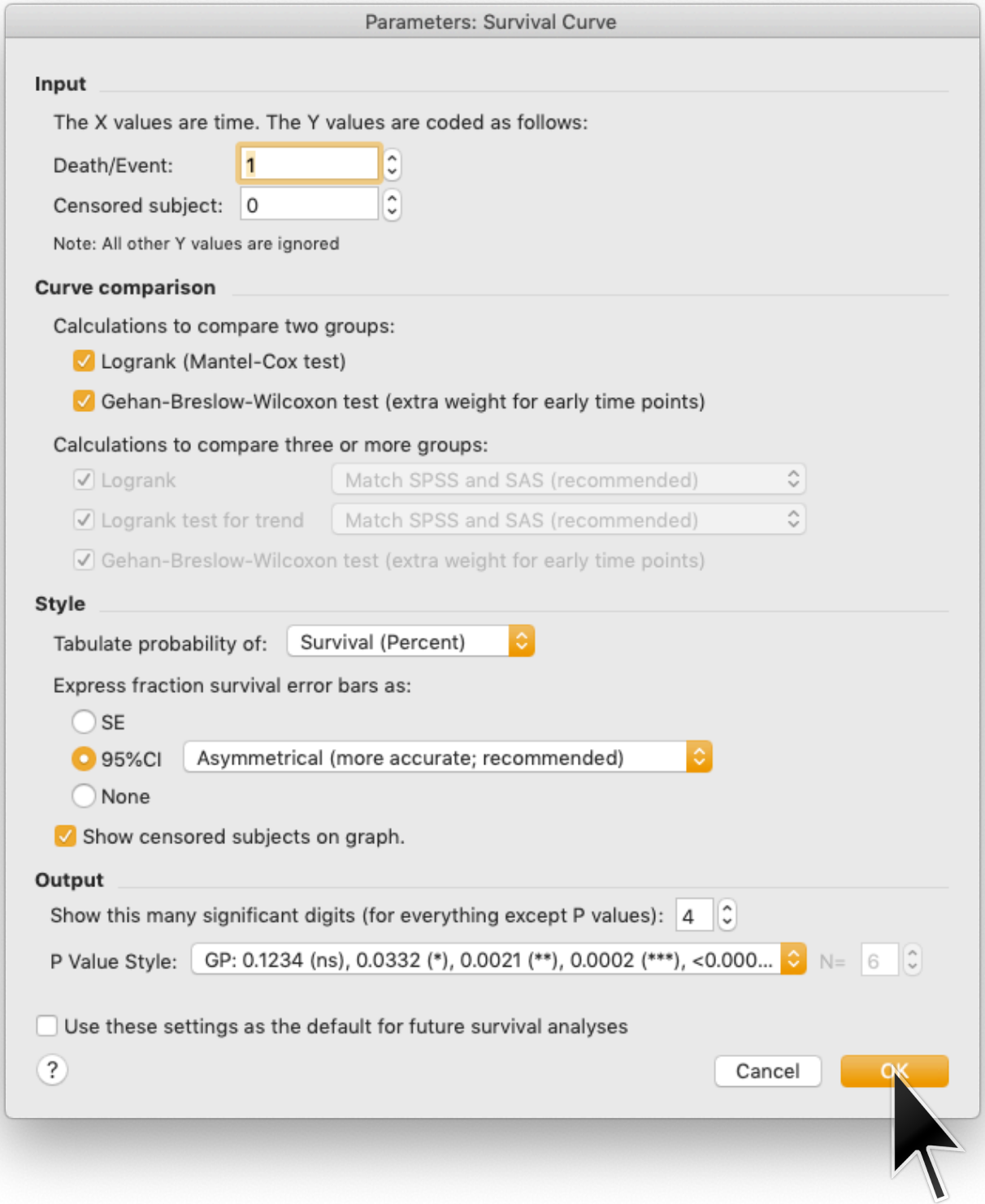
Prism Tips

Cancel | Create



(data curtsey of Tobias Janowitz)


Create New Analysis
Data to analyze
Table:
Data 1
Type of analysis
Which analysis?
Analyze which data sets?
Transform, Normalize...
Transform
Transform concentrations (X)
Normalize
Prune rows
Remove baseline and column math
Transpose X and Y
Fraction of Total
XY analyses
Column analyses
Grouped analyses
Contingency table analyses
Survival analyses
Survival curve
Parts of whole analyses
Multiple variable analyses
Nested analyses
Generate curve
Simulate data
Recently used
A:No ALND
B:ALND
Select All
Deselect All
?
Cancel
OK


Parameters: Survival Curve
Input
The X values are time. The Y values are coded as follows:
Death/Event:
1
Censored subject:
0
Note: All other Y values are ignored
Curve comparison
Calculations to compare two groups:
Logrank (Mantel-Cox test)
Gehan-Breslow-Wilcoxon test (extra weight for early time points)
Calculations to compare three or more groups:
Logrank
Match SPSS and SAS (recommended)
Logrank test for trend
Match SPSS and SAS (recommended)
Gehan-Breslow-Wilcoxon test (extra weight for early time points)
Style
Tabulate probability of:
Survival (Percent)
Express fraction survival error bars as:
SE
95%CI
Asymmetrical (more accurate; recommended)
None
Show censored subjects on graph.
Output
Show this many significant digits (for everything except P values):
4
P Value Style:
GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.0002 (***), <0.000...
N=
6
Use these settings as the default for future survival analyses
?
Cancel
OK

tobias.pzfx — Edited

Search

Data Tables
- Data 1
- New Data Table...

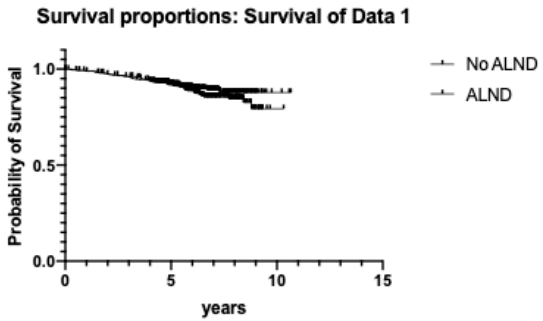
Info
- Project info 1
- New Info...

Results
- Survival of Data 1
- New Analysis...

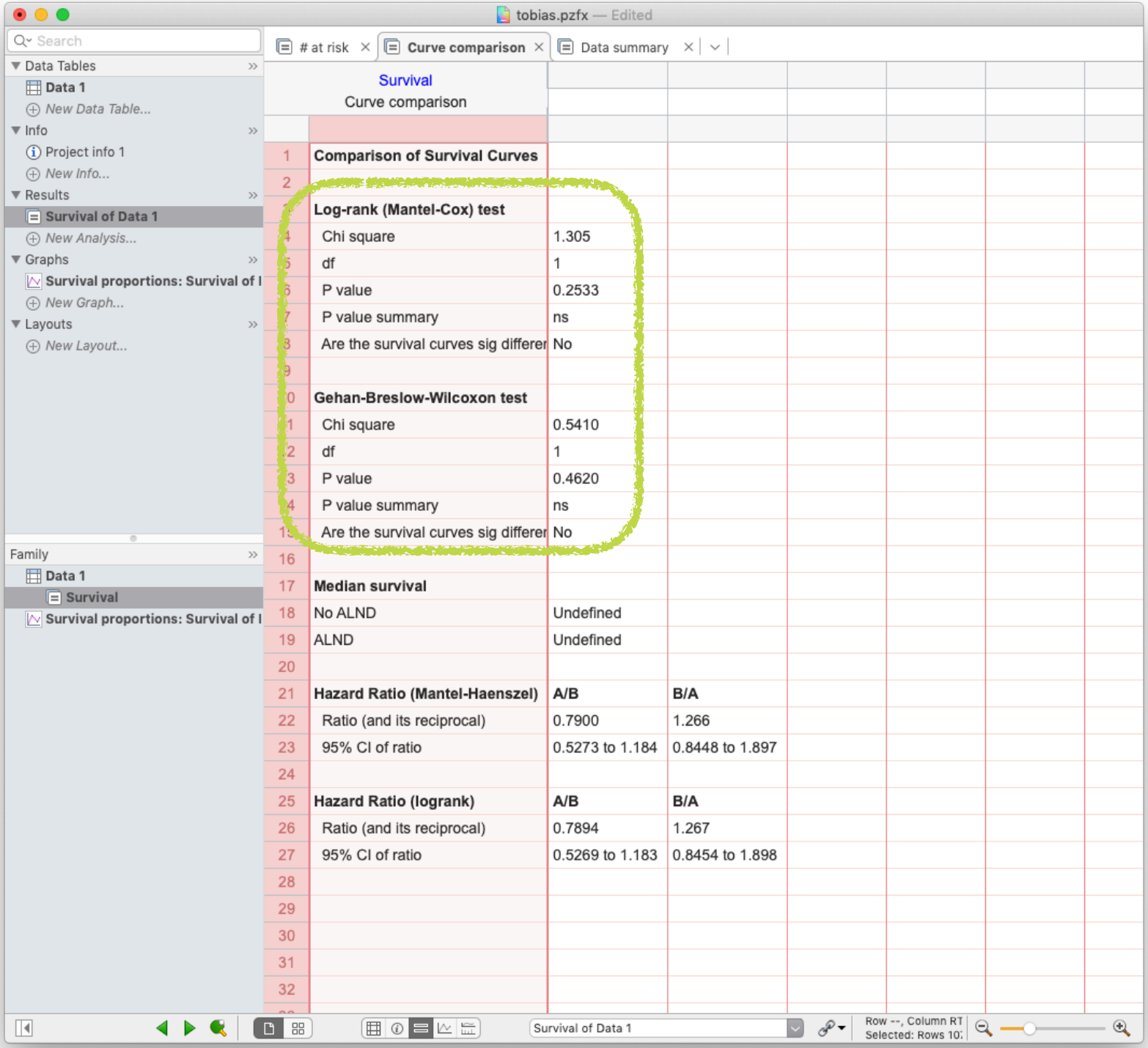
Graphs
- Survival proportions: Survival of Data 1
- New Graph...

Layouts
- New Layout...

Family
- Data 1
  - Survival
- Survival proportions: Survival of Data 1



Survival proportions: Survival of Data 1


tobias.pzfx — Edited
Search
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
Survival of Data 1
New Analysis...
Graphs
Survival proportions: Survival of I
New Graph...
Layouts
New Layout...
Family
Data 1
Survival
Survival proportions: Survival of I
at risk
Curve comparison
Data summary
Survival
Curve comparison
1 Comparison of Survival Curves
Log-rank (Mantel-Cox) test
Chi square 1.305
df 1
P value 0.2533
P value summary ns
Are the survival curves sig differer No
Gehan-Breslow-Wilcoxon test
Chi square 0.5410
df 1
P value 0.4620
P value summary ns
Are the survival curves sig differer No
Median survival
No ALND Undefined
ALND Undefined
Hazard Ratio (Mantel-Haenszel) A/B B/A
Ratio (and its reciprocal) 0.7900 1.266
95% CI of ratio 0.5273 to 1.184 0.8448 to 1.897
Hazard Ratio (logrank) A/B B/A
Ratio (and its reciprocal) 0.7894 1.267
95% CI of ratio 0.5269 to 1.183 0.8454 to 1.898
Survival of Data 1
Row --, Column RT
Selected: Rows 10:

The Cox proportional hazards model is the most common way to analyze how different variables influence survival

Suppose that each individual i has, in addition to an event time t_i and event flag, has a set of D covariants $x_{i1}, x_{i2}, \ldots, x_{iD}$, which can be either real numbers or binary.

The Cox proportional hazards model assumes that subjects are governed by a hazards function that has the following form.

$$h_i(t) = h_0(t) \times \exp\left[\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_D x_{iD}\right]$$

Each coefficient β_j is the "effect size" for the corresponding covariate $x_{\cdot j}$. If the value for β_j is significantly different than 0, it means that the covariate $x_{\cdot j}$ effects survival.

Example: Rossi recidivism dataset

lifelines.datasets.load_rossi(***kwargs*)

This data set is originally from Rossi et al. (1980), and is used as an example in Allison (1995). The data pertain to 432 convicts who were released from Maryland state prisons in the 1970s and who were followed up for one year after release. Half the released convicts were assigned at random to an experimental treatment in which they were given financial aid; half did not receive aid.:

```
Size: (432,9)
Example:
    week      20
    arrest     1
    fin        0
    age       27
    race       1
    wexp       0
    mar        0
    paro       1
    prio       3
```

References

Rossi, P.H., R.A. Berk, and K.J. Lenihan (1980). Money, Work, and Crime: Some Experimental Results. New York: Academic Press. John Fox, Marilia Sa Carvalho (2012). The RcmdrPlugin.survival Package: Extending the R Commander Interface to Survival Analysis. Journal of Statistical Software, 49(7), 1-32.

https://lifelines.readthedocs.io/en/latest/lifelines.datasets.html#lifelines.datasets.load_rossi

Example: Rossi recidivism dataset

A data frame with 432 observations on the following 62 variables.

`week`
: week of first arrest after release or censoring; all censored observations are censored at 52 weeks.

`arrest`
: `1` if arrested, `0` if not arrested.

`fin`
: financial aid: `no` `yes`.

`age`
: in years at time of release.

`race`
: `black` or `other`.

`wexp`
: full-time work experience before incarceration: `no` or `yes`.

`mar`
: marital status at time of release: `married` or `not married`.

`paro`
: released on parole? `no` or `yes`.

`prio`
: number of convictions prior to current incarceration.

`educ`
: level of education: `2` = 6th grade or less; `3` = 7th to 9th grade; `4` = 10th to 11th grade; `5` = 12th grade; `6` = some college.

https://rdrr.io/cran/RcmdrPlugin.survival/man/Rossi.html

Example: Rossi recidivism dataset

```
# Load and preview Rossi dataset
from lifelines.datasets import load_rossi
rossi_df = load_rossi()
rossi_df.head()
```

	week	arrest	fin	age	race	wexp	mar	paro	prio
0	20	1	0	27	1	0	0	1	3
1	17	1	0	18	1	0	0	1	8
2	25	1	0	19	0	1	0	1	13
3	52	0	1	23	1	1	1	1	1
4	52	0	0	19	0	1	0	1	3

week: survival time
arrest: 1 if arrested (event), 0 if not arrested (censored)

The results of Cox Regression is a statement about the effect size and significance of each variable

effect size (hazard)

	exp(coef)	exp(coef) lower 95%	exp(coef) upper 95%
fin	0.68	0.47	1.00
age	0.94	0.90	0.99
race	1.37	0.75	2.50
wexp	0.86	0.57	1.30
mar	0.65	0.31	1.37
paro	0.92	0.63	1.35
prio	1.10	1.04	1.16

statistical significance

	z	p	-log2(p)
fin	-1.98	0.05	4.40
age	-2.61	0.01	6.79
race	1.02	0.31	1.70
wexp	-0.71	0.48	1.06
mar	-1.14	0.26	1.97
paro	-0.43	0.66	0.59
prio	3.19	<0.005	9.48

Likelihood ratio test

```
Log-likelihood ratio test = 33.27 on 7 df, -log2(p)=15.37
```

The likelihood ratio test is an extremely general way of comparing two models. It is an approximate test, though, valid only in the large data regime.

Likelihood ratio test uses a statistic given by:

$$\chi^2 = 2\log\left(\frac{\text{Likelihood}_{\text{alt}}}{\text{Likelihood}_{\text{null}}}\right)$$

Under the null hypothesis, χ^2 follows a chi square distribution where the number of degrees of freedom is:

$$\text{DOF} = (\#\ \text{alt model parameters}) - (\#\ \text{null model parameters})$$

It tests the necessity of all parameters; it does not say whether individual parameters are required.