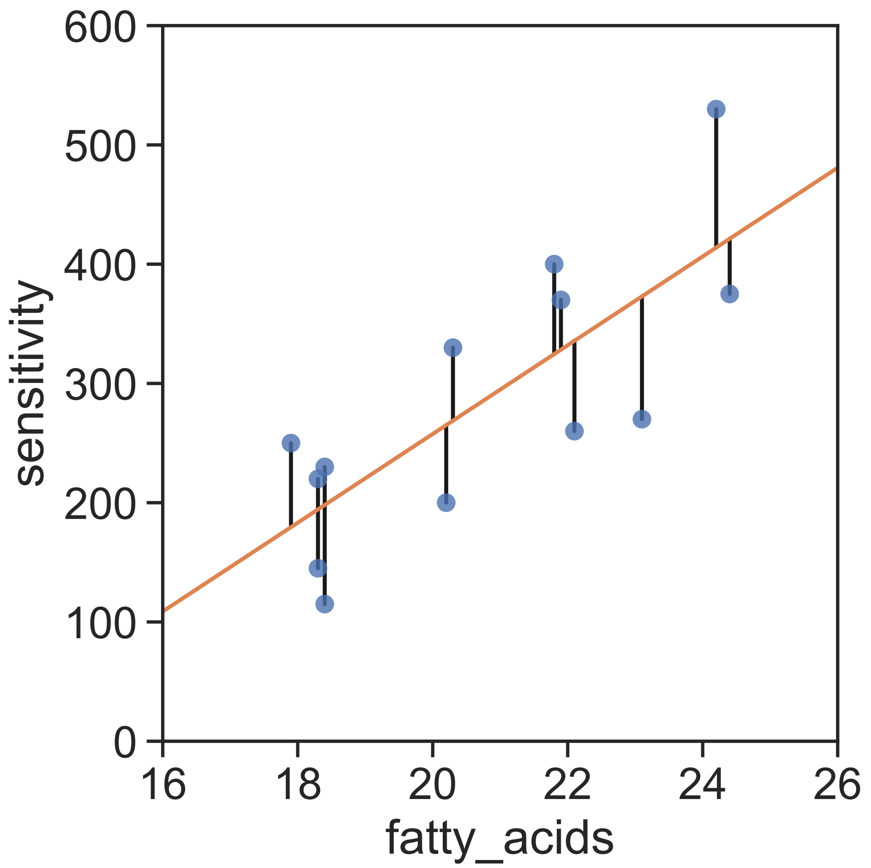
Linear regression
Nonlinear regression
Survival analysis

Biostatistics Course 2024
Lecture 5
Friday, 12 July 2024
2:00pm - 4:00pm

Linear regression

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



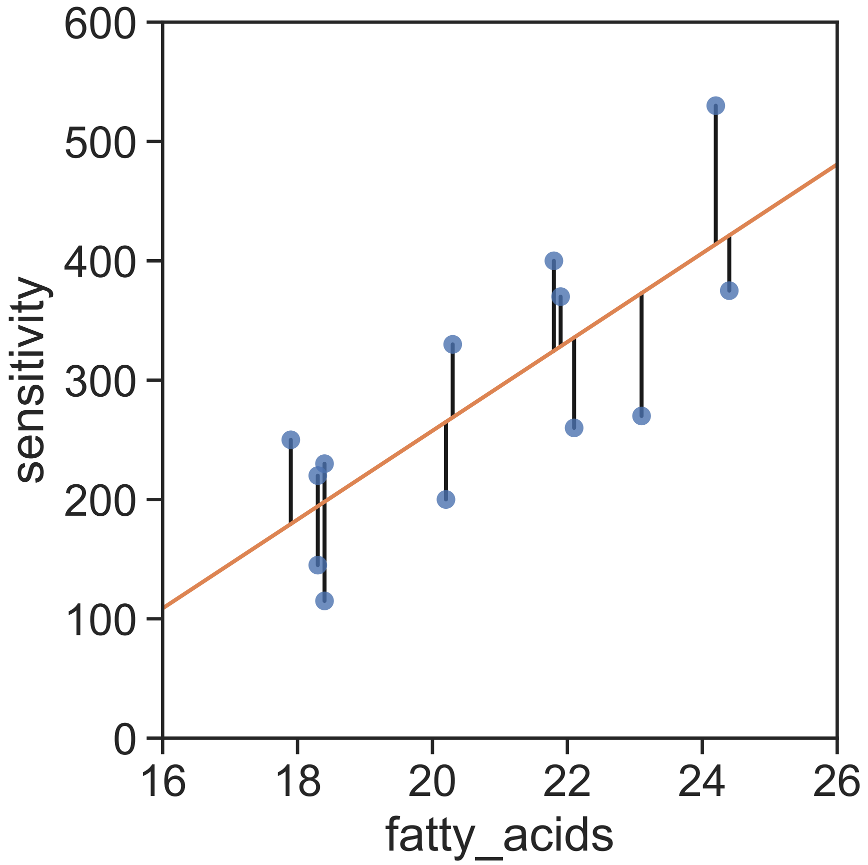
$$y_i = a + bx_i + \epsilon_i$$

a: y-intercept

b: slope

ϵ_i: the “residuals”

Parameters are chosen to minimize the sum of squared deviations



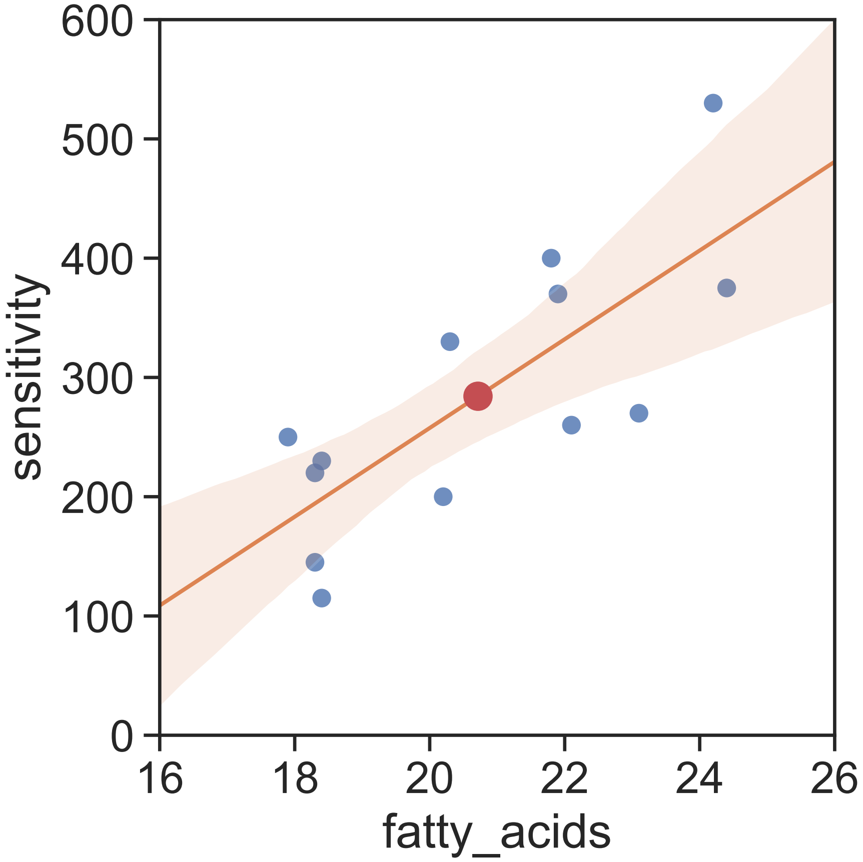
$$y_i = a + bx_i + \epsilon_i$$

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

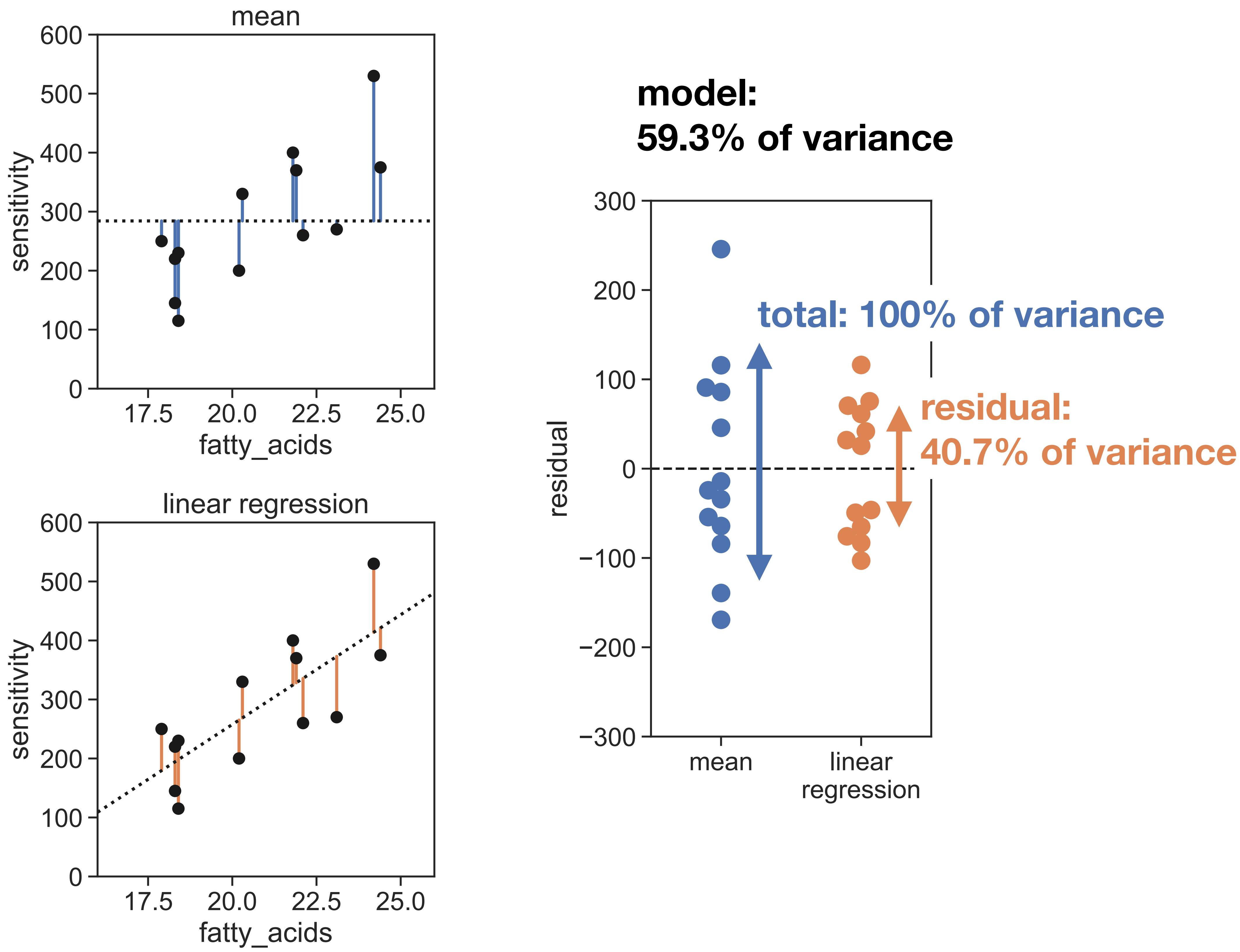


The center of mass point of the data, $(\hat{\mu}_x, \hat{\mu}_y)$, lies on the regression line.

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

total: 100% of variance

residual: 40.7% of variance

Linear regression explains a fraction of the variance

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

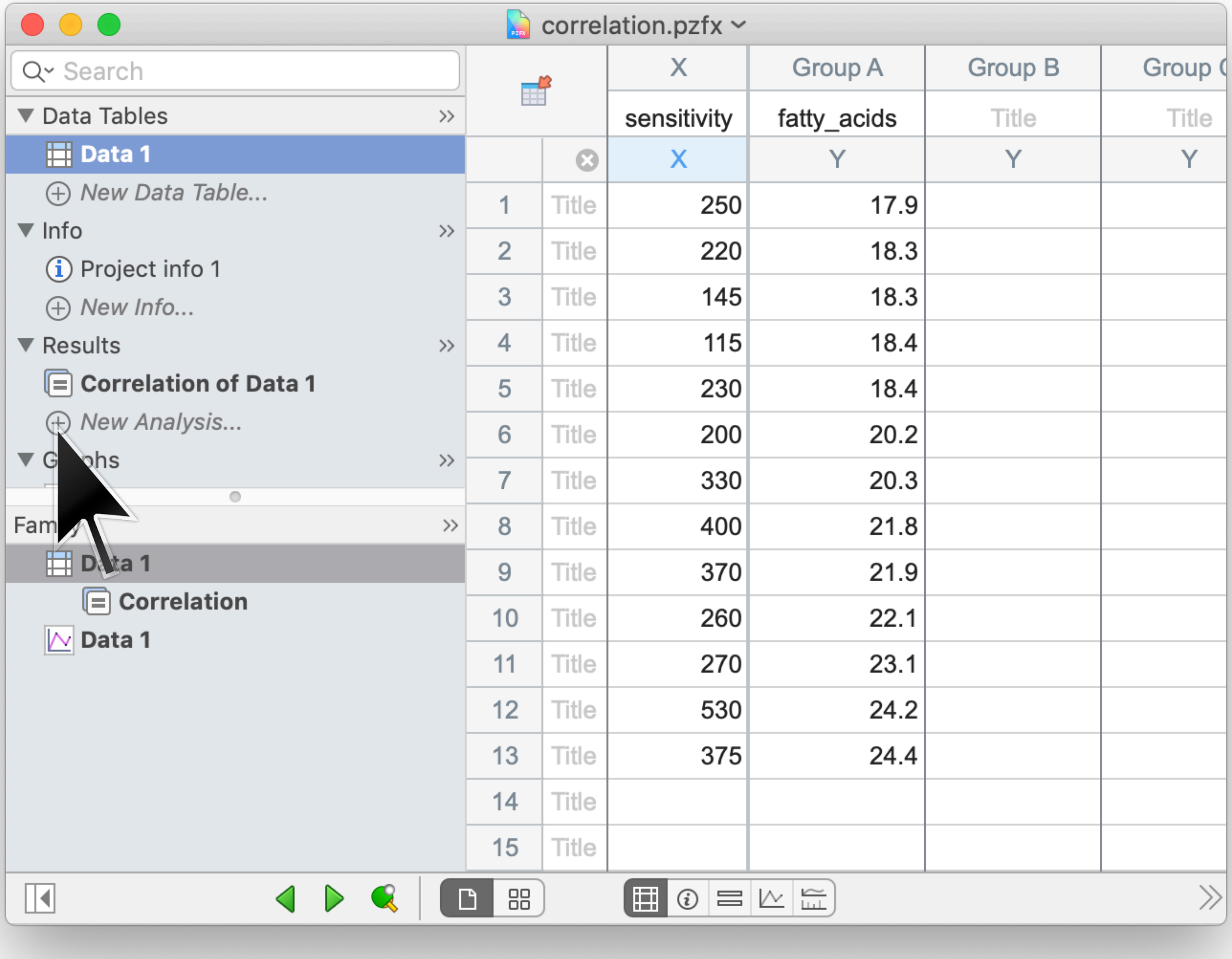
$(n-1) \times$ variance:

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

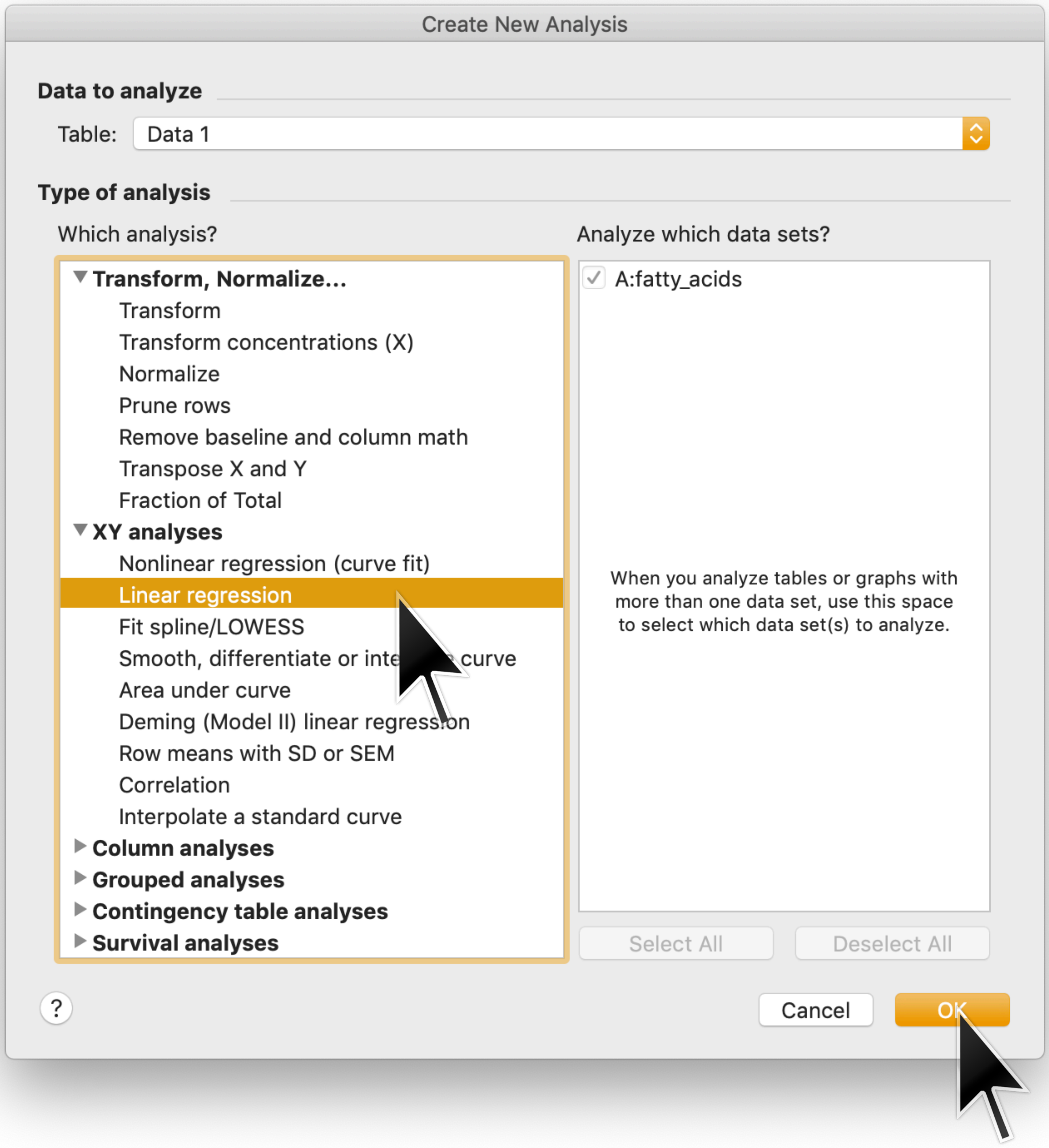
total: 100% **residual: 40.7%** **model: 59.3%**

r^2 is the fraction of variance explained:

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



		X	Group A	Group B	Group
		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

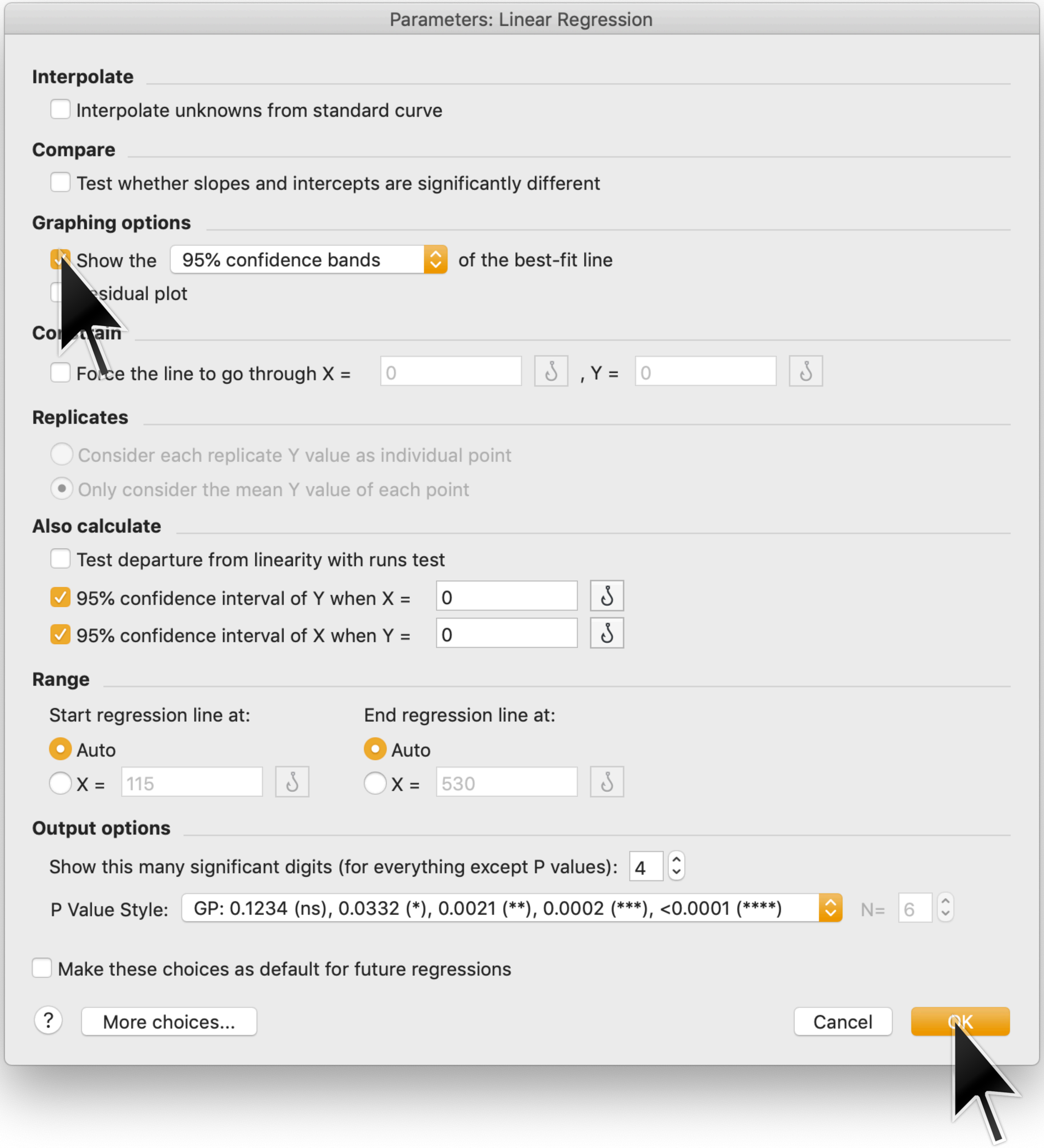
Type of analysis

Which analysis?

Analyze which data sets?

☑ 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.



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

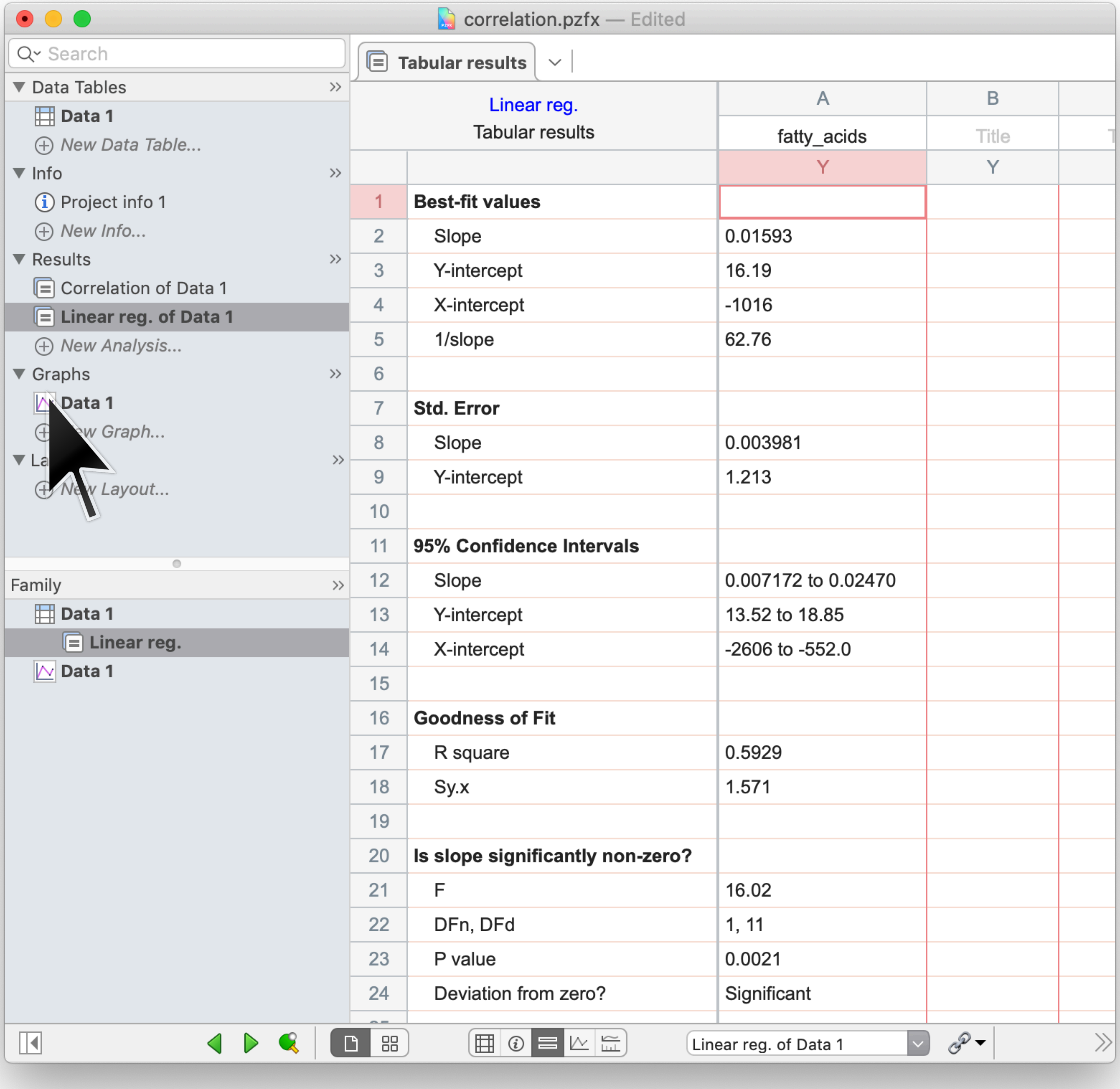
○ X = 530

Output options

Show this many significant digits (for everything except P values): 4

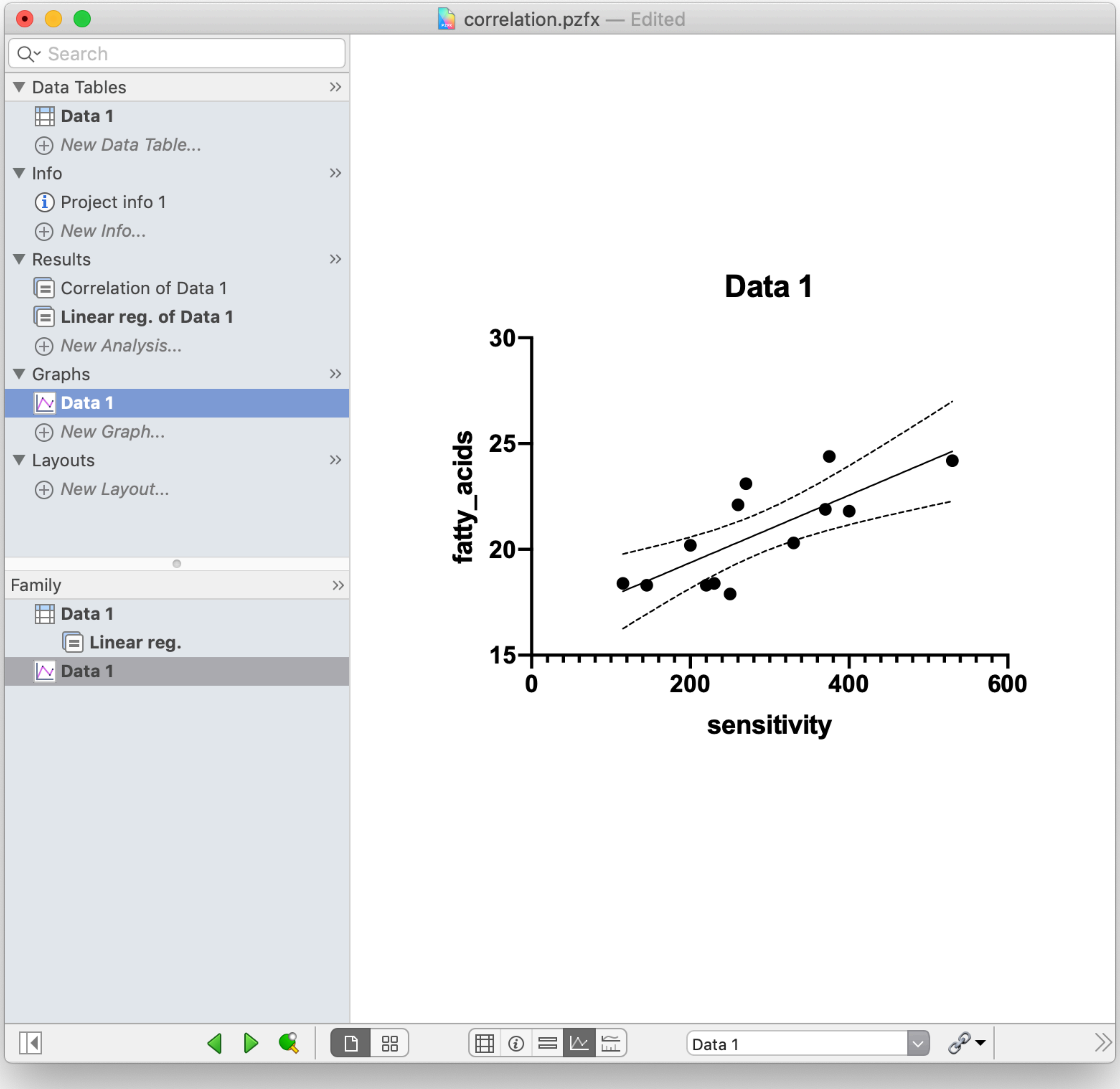
P Value Style: GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.0002 (***), <0.0001 (****) N= 6

□ Make these choices as default for future regressions



Linear reg.
Tabular results

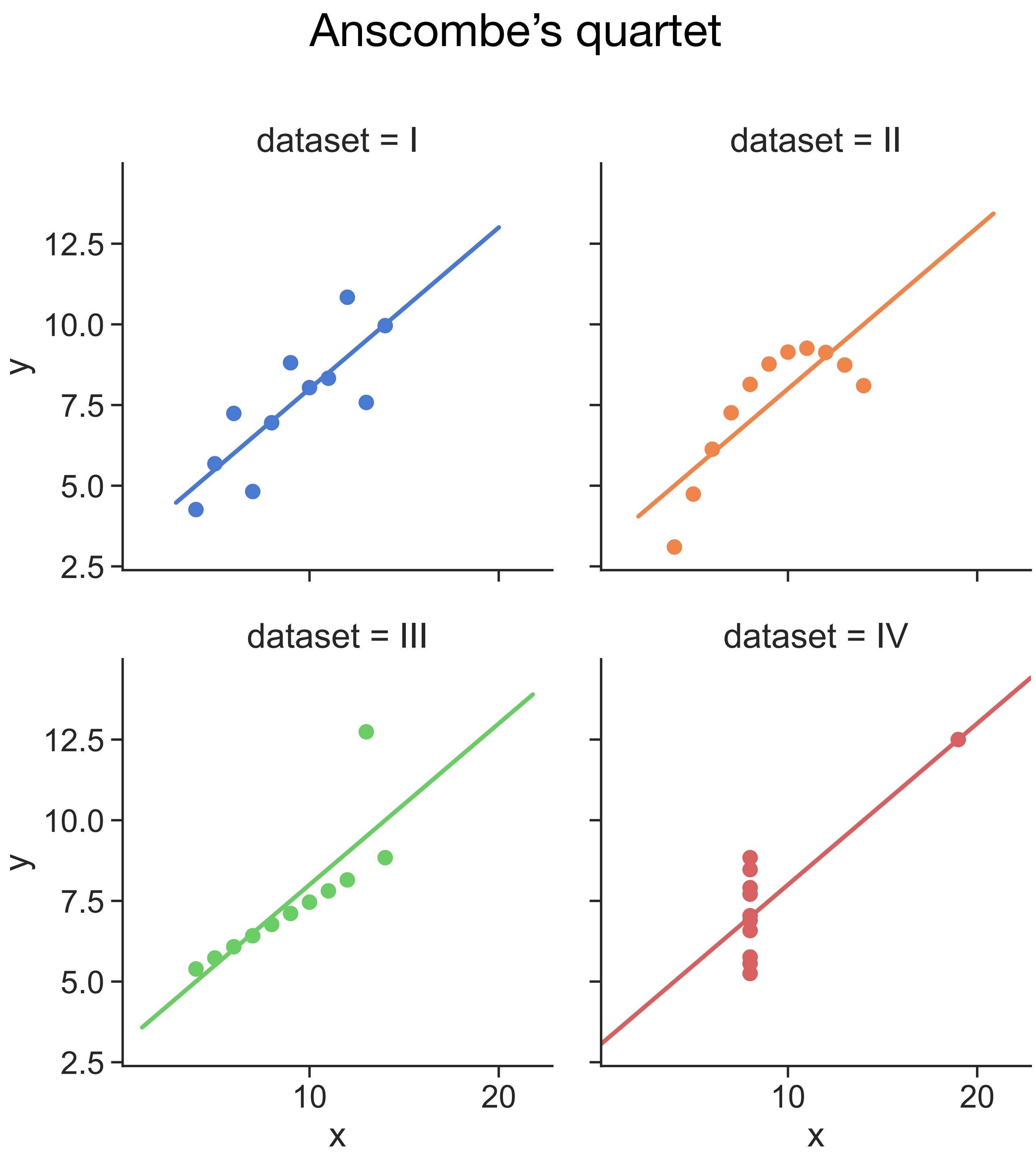
		A	B
		fatty_acids	Title
		Y	Y
1	**Best-fit values**		
2	Slope	0.01593	
3	Y-intercept	16.19	
4	X-intercept	-1016	
5	1/slope	62.76	
6			
7	**Std. Error**		
8	Slope	0.003981	
9	Y-intercept	1.213	
10			
11	**95% Confidence Intervals**		
12	Slope	0.007172 to 0.02470	
13	Y-intercept	13.52 to 18.85	
14	X-intercept	-2606 to -552.0	
15			
16	**Goodness of Fit**		
17	R square	0.5929	
18	Sy.x	1.571	
19			
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	


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
30
25
20
15
fatty_acids
0
200
400
600
sensitivity
Data 1

Linear regression assumptions

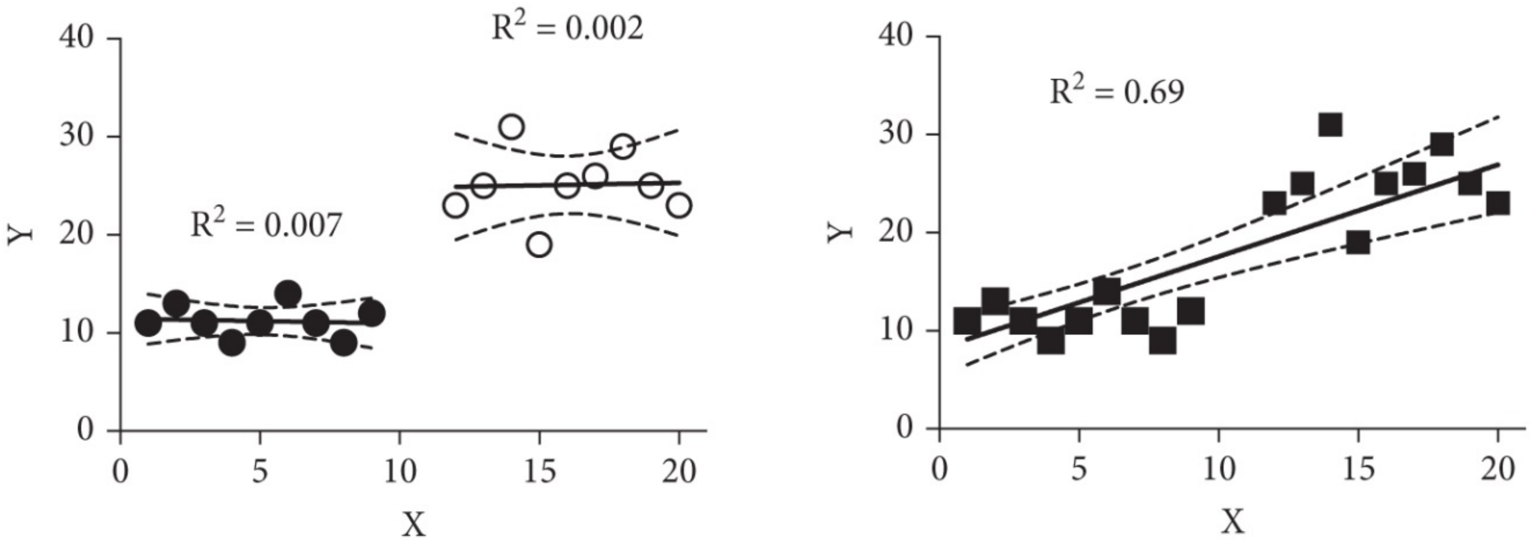
- The model is correct, i.e. the expected value for y is indeed a linear function of x for some correct choice of parameters.
- The noise (i.e. the residuals) is Gaussian and has mean zero.
- The residual for each data point is statistically independent
- The magnitude of the noise (i.e. variance of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

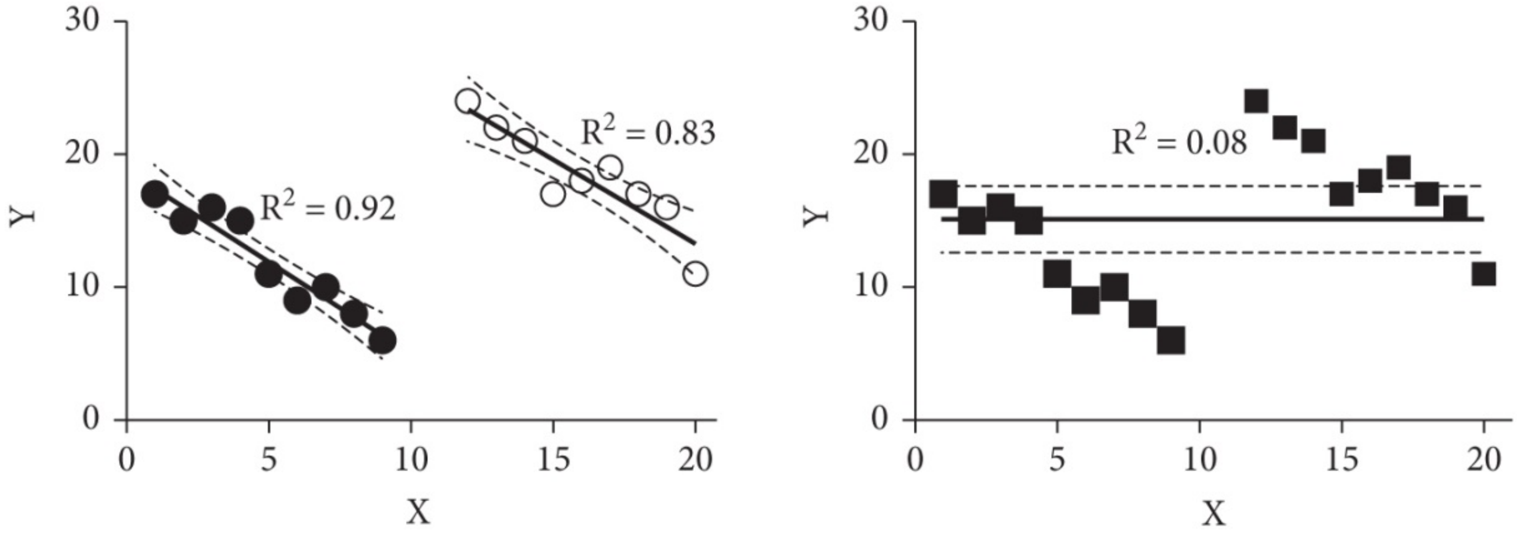


Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21.

Beware of combining distinct groups into one



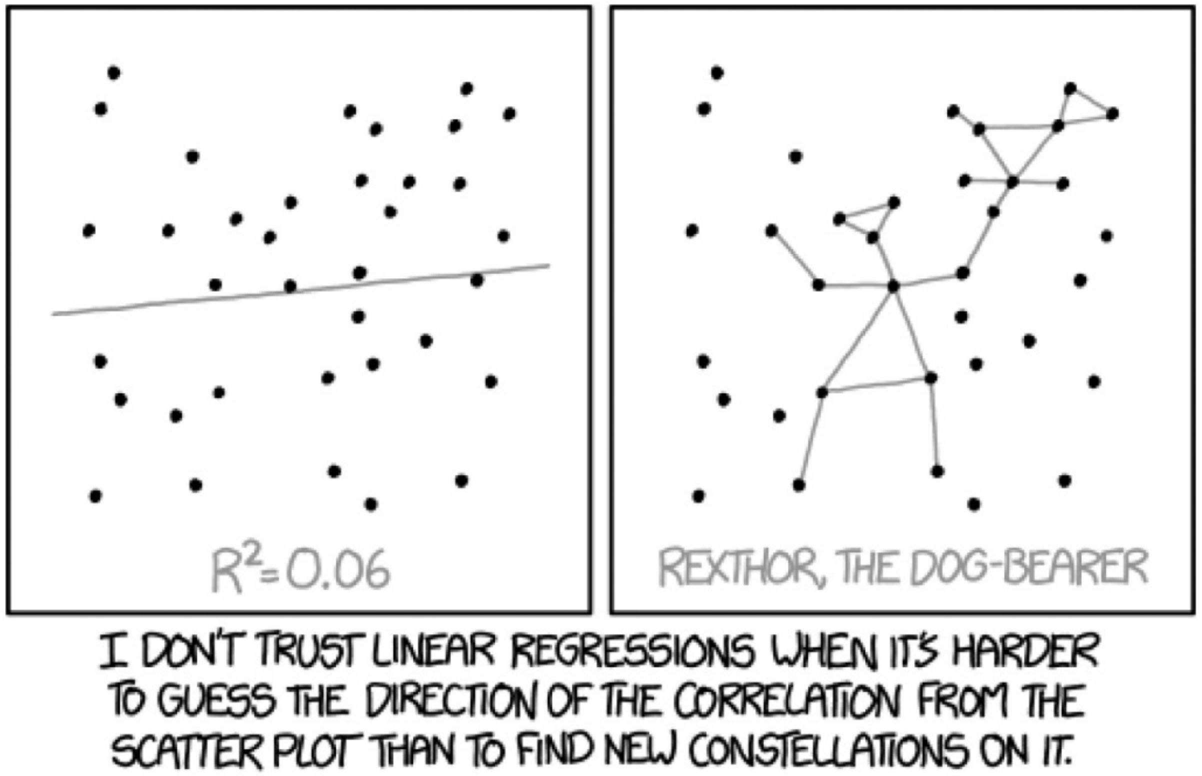
Combining two groups into one regression can mislead by creating a strong linear relationship.



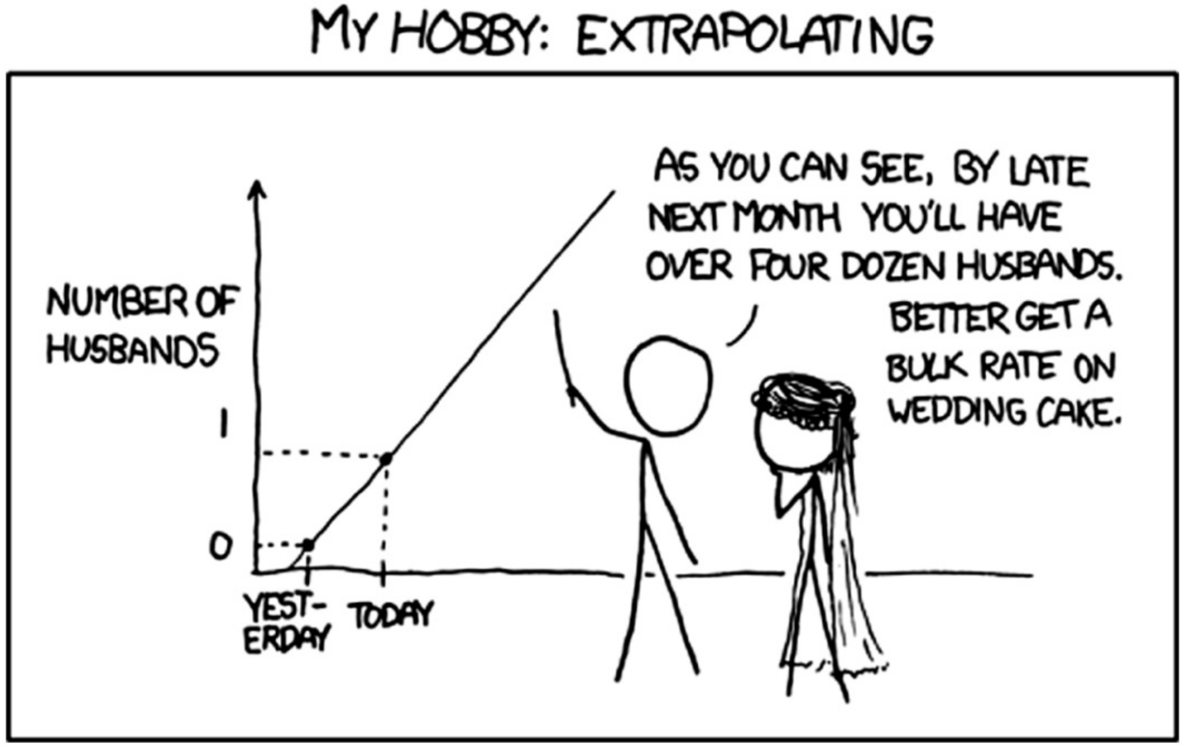
Combining two groups into one regression can mislead by hiding a trend.

Beware of reading too much into a regression result

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


R²=0.06
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Don't over-extrapolate


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS. BETTER GET A BULK RATE ON WEDDING CAKE.

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.

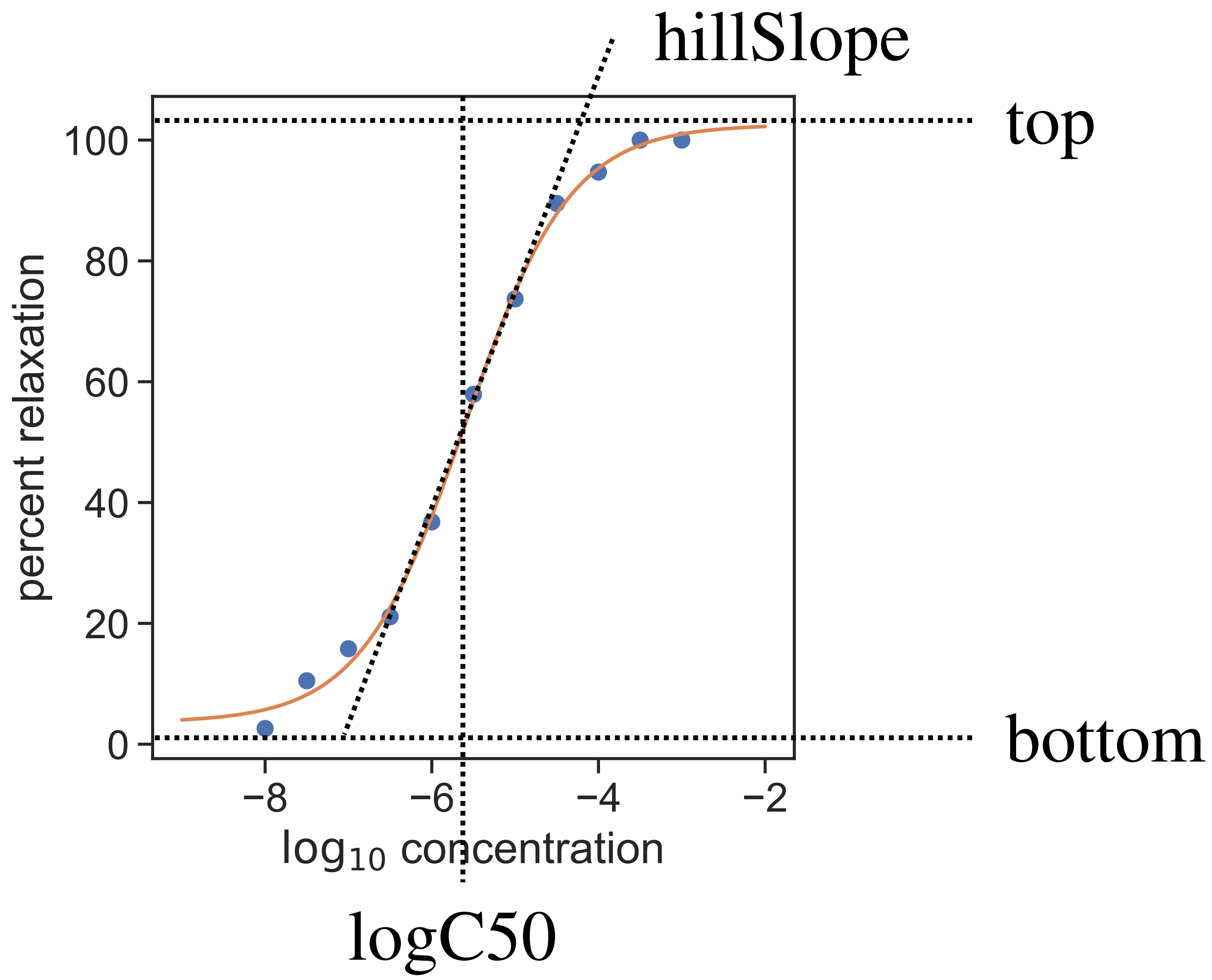
The data from each rat was analyzed to determine the maximum relaxation and the concentration of norepinephrine that relaxes the muscle half that much (C50)

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



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

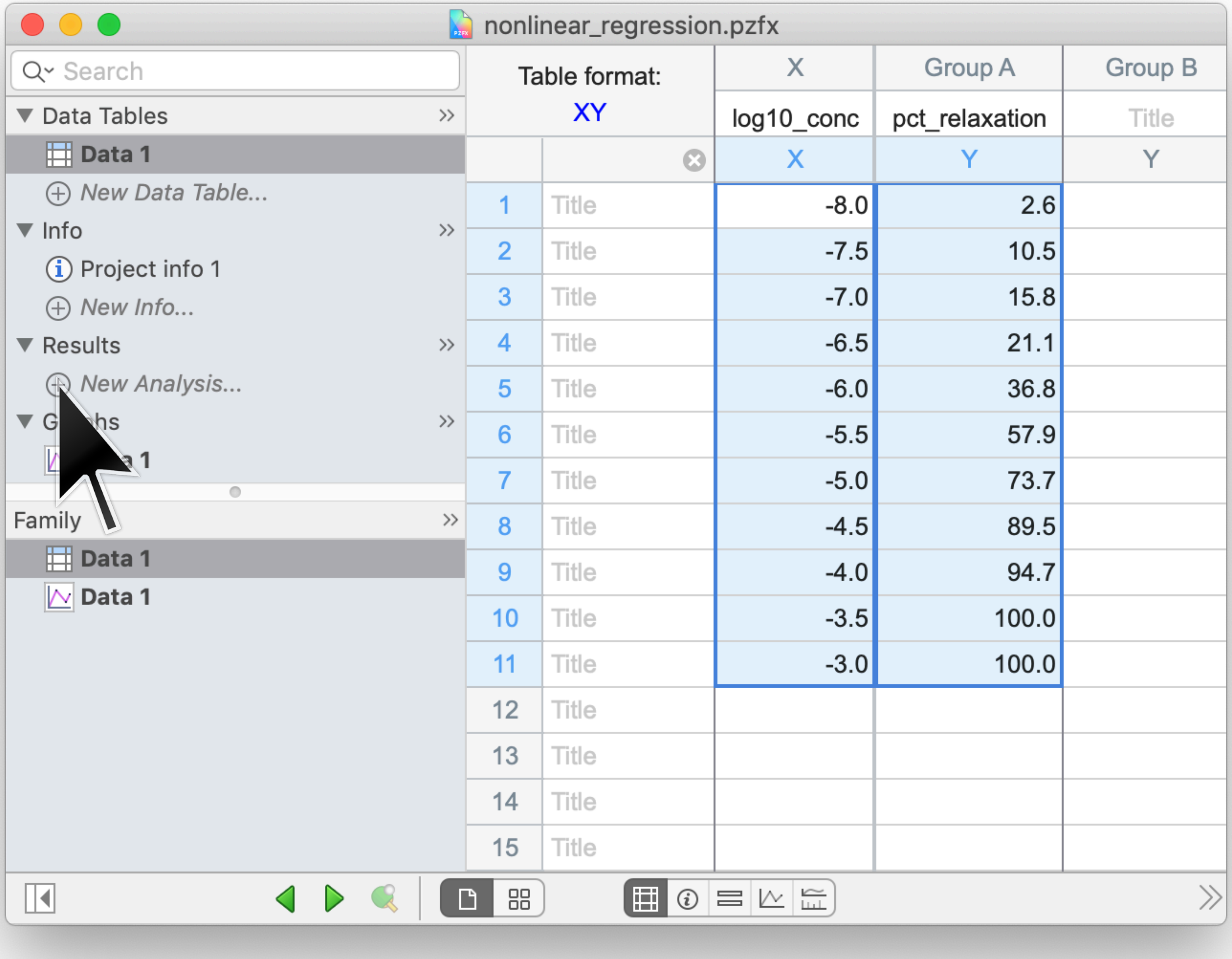
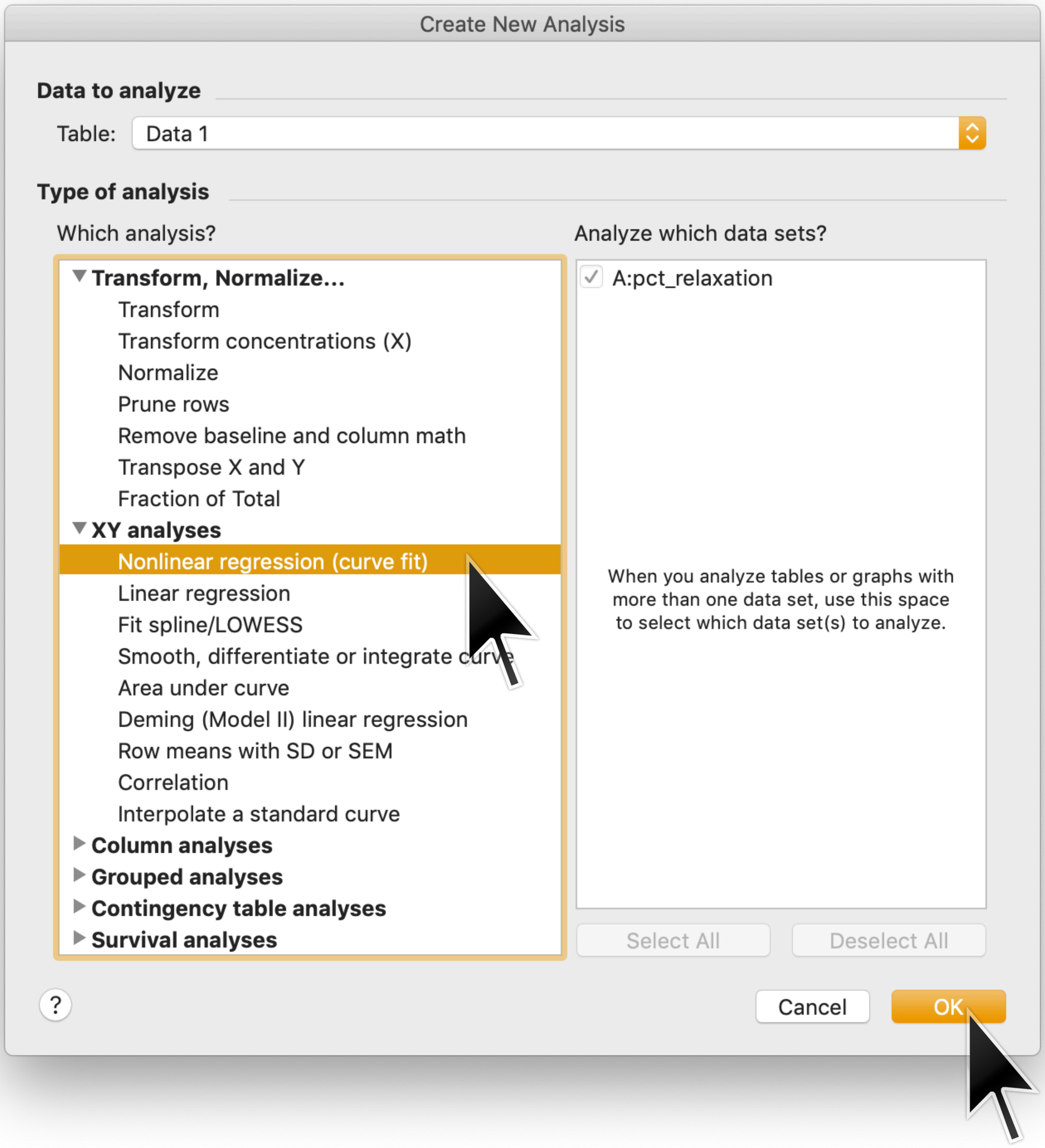


Table format: XY

		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			



Data to analyze

Table: Data 1

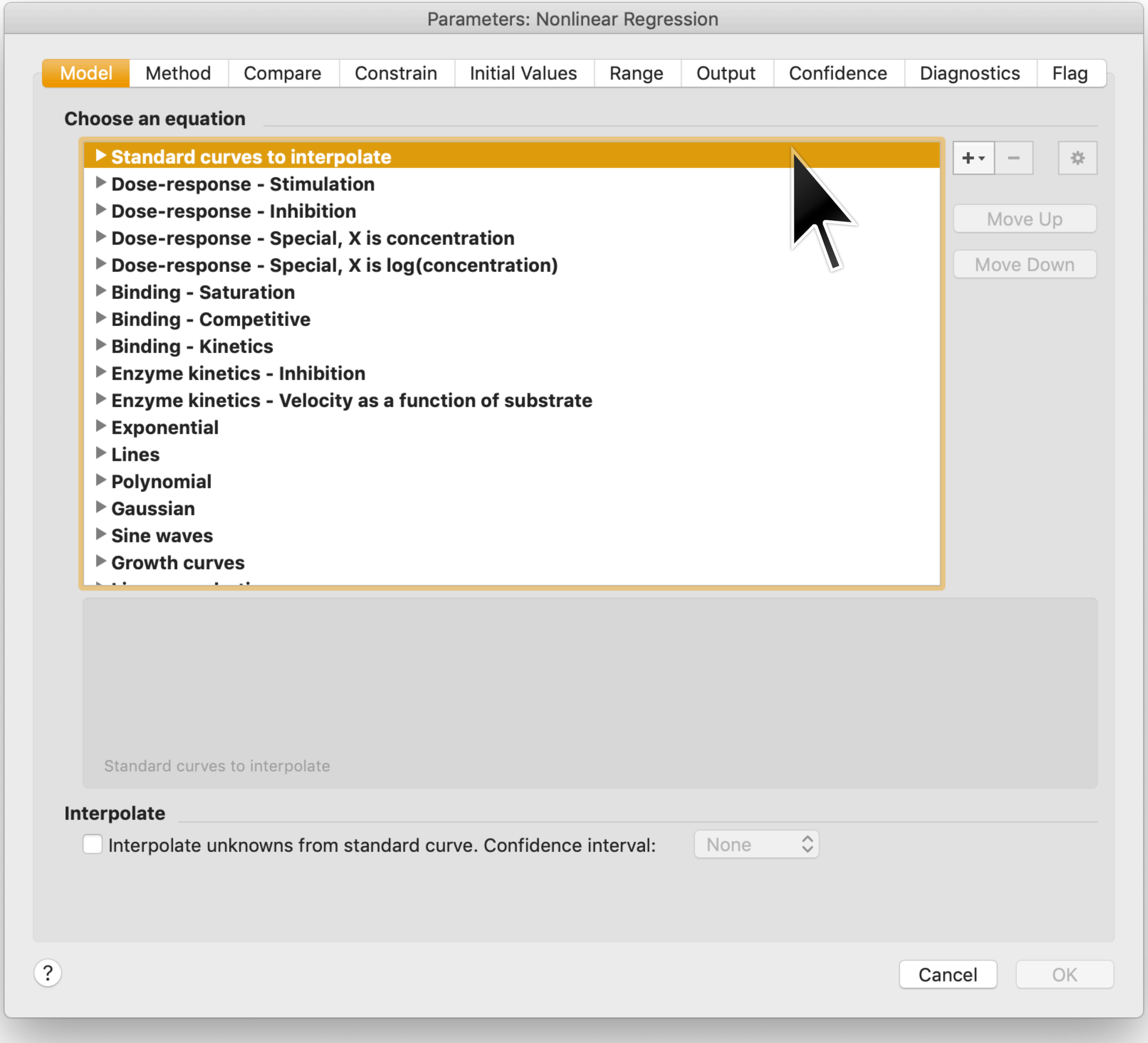
Type of analysis

Which analysis?

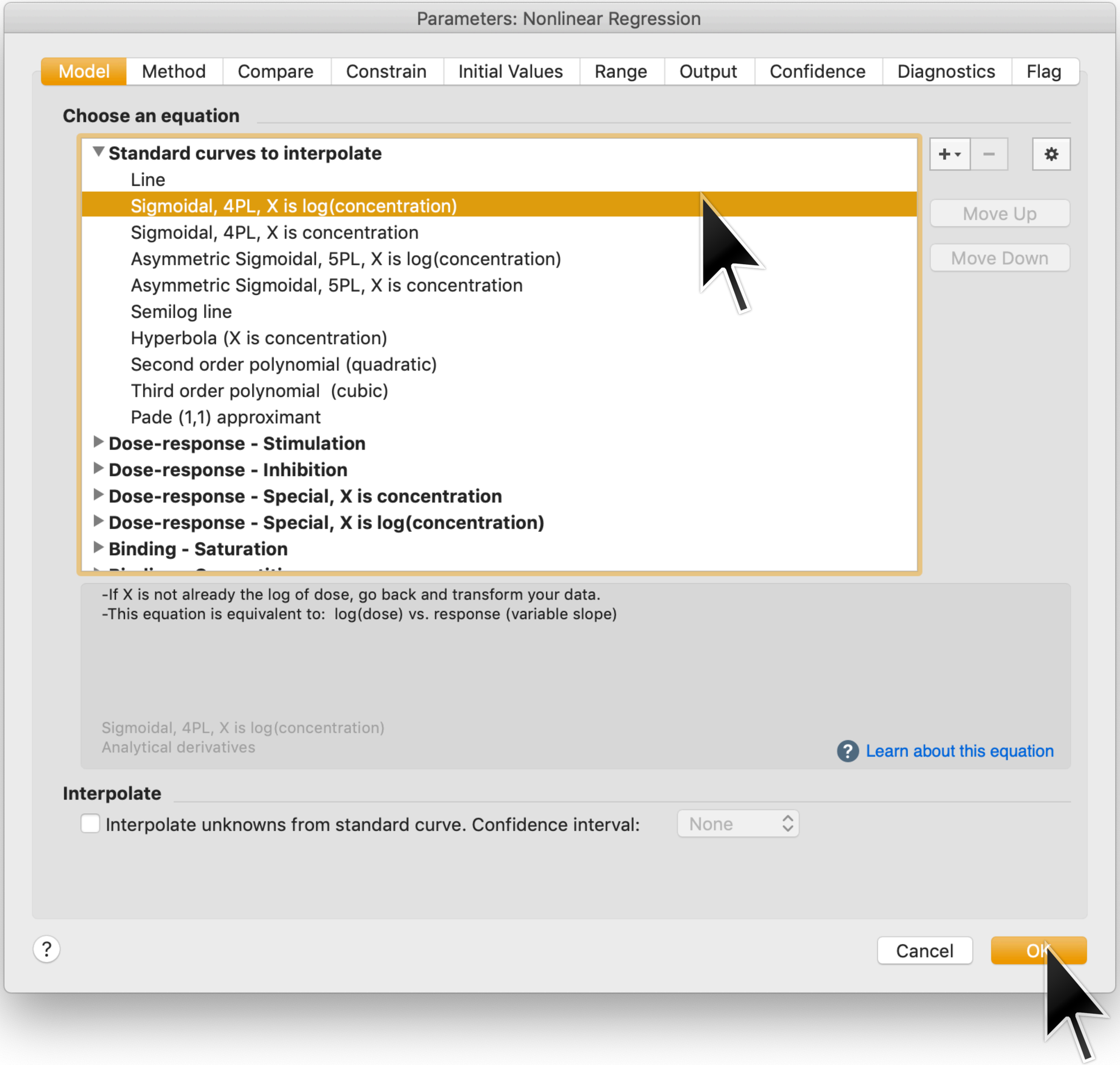
Analyze which data sets?

☑ 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.


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
?
Cancel
OK

□ Interpolate unknowns from standard curve. Confidence interval: None



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

-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)

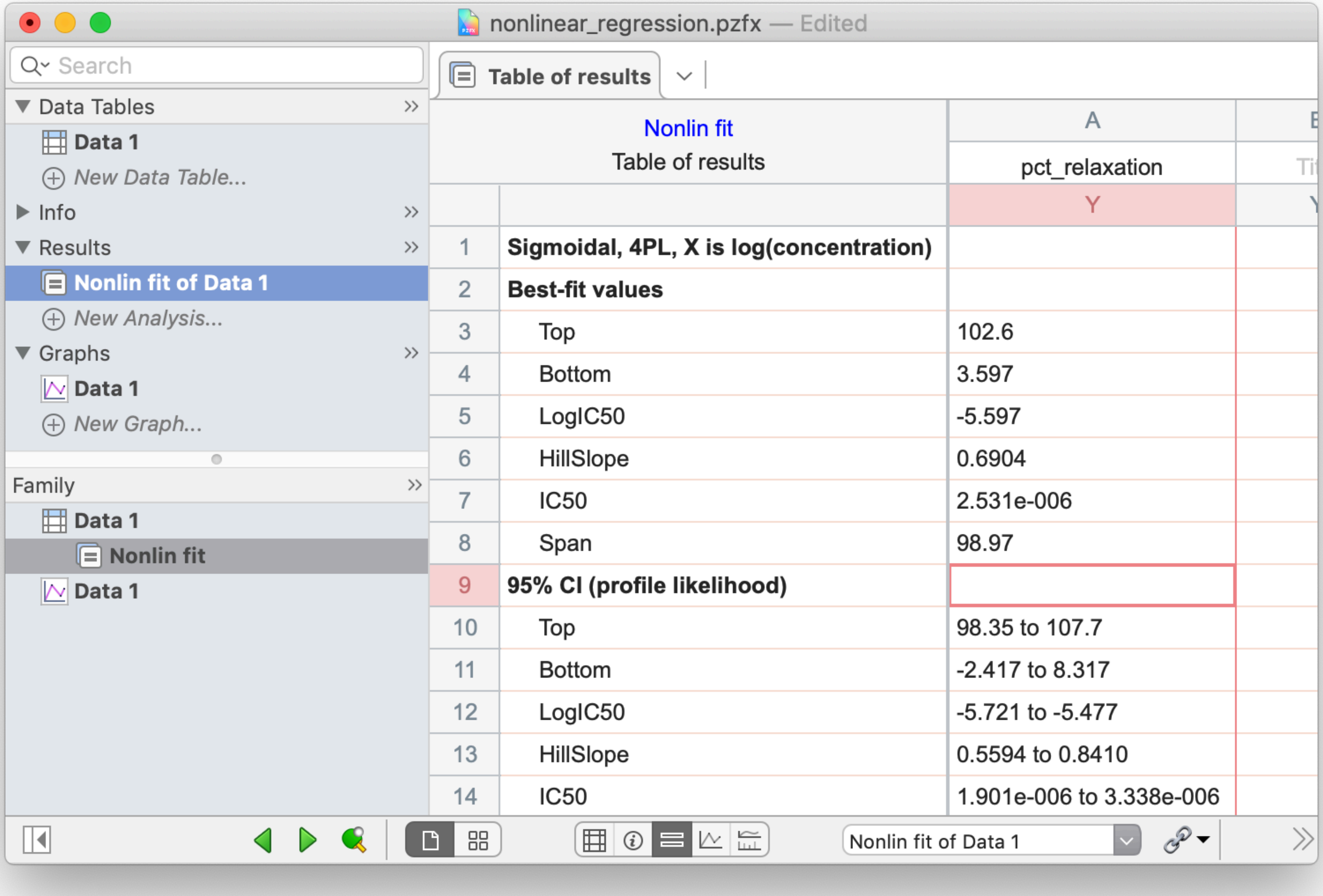
Sigmoidal, 4PL, X is log(concentration)
Analytical derivatives

Interpolate

□ Interpolate unknowns from standard curve. Confidence interval: None

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

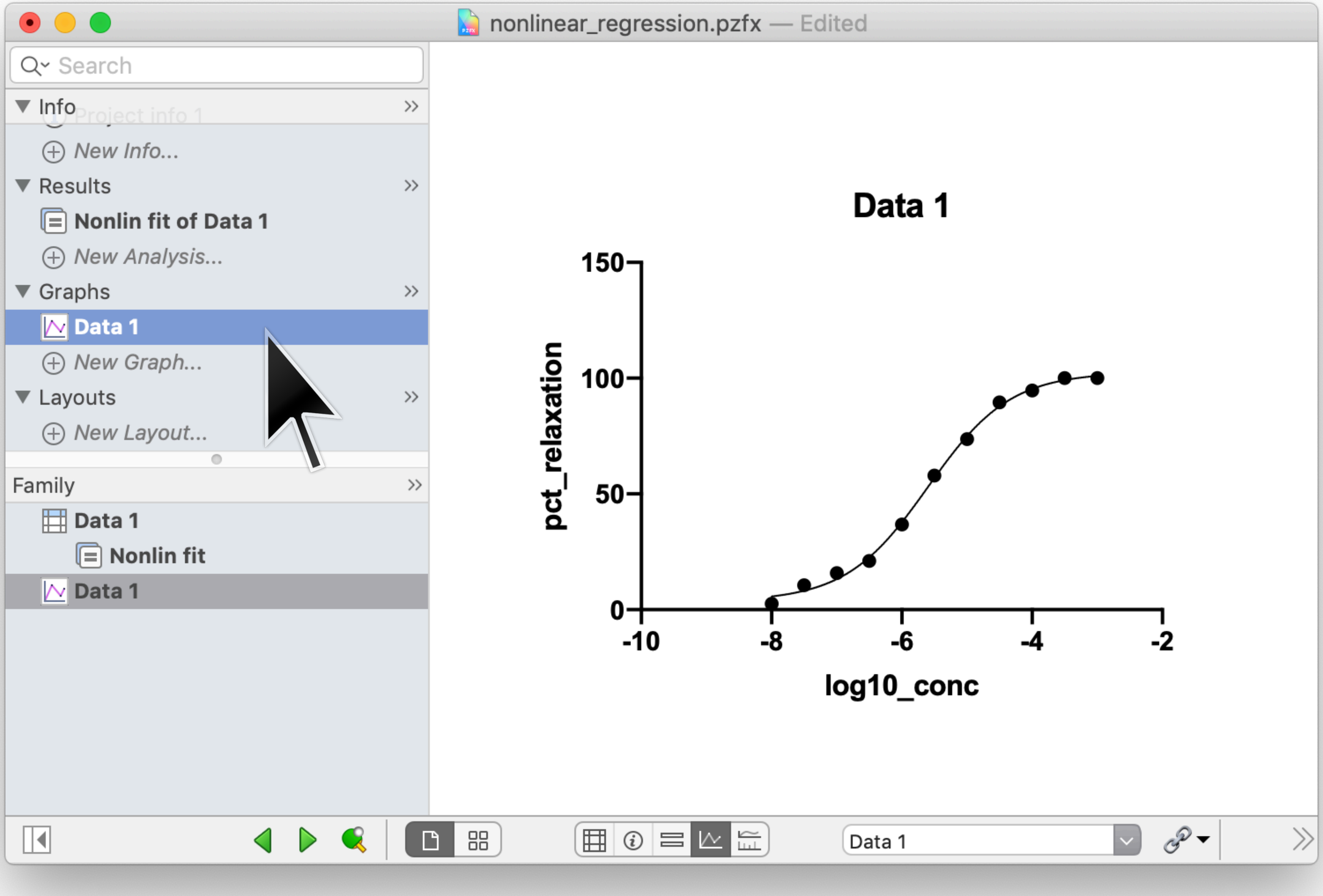
4 parameters: Bottom, Top, LogIC50, HillSlope



	Nonlin fit Table of results	A pct_relaxation Y
1	**Sigmoidal, 4PL, X is log(concentration)**	
2	**Best-fit values**	
3	Top	102.6
4	Bottom	3.597
5	LogIC50	-5.597
6	HillSlope	0.6904
7	IC50	2.531e-006
8	Span	98.97
9	**95% CI (profile likelihood)**	
10	Top	98.35 to 107.7
11	Bottom	-2.417 to 8.317
12	LogIC50	-5.721 to -5.477
13	HillSlope	0.5594 to 0.8410
14	IC50	1.901e-006 to 3.338e-006

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

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

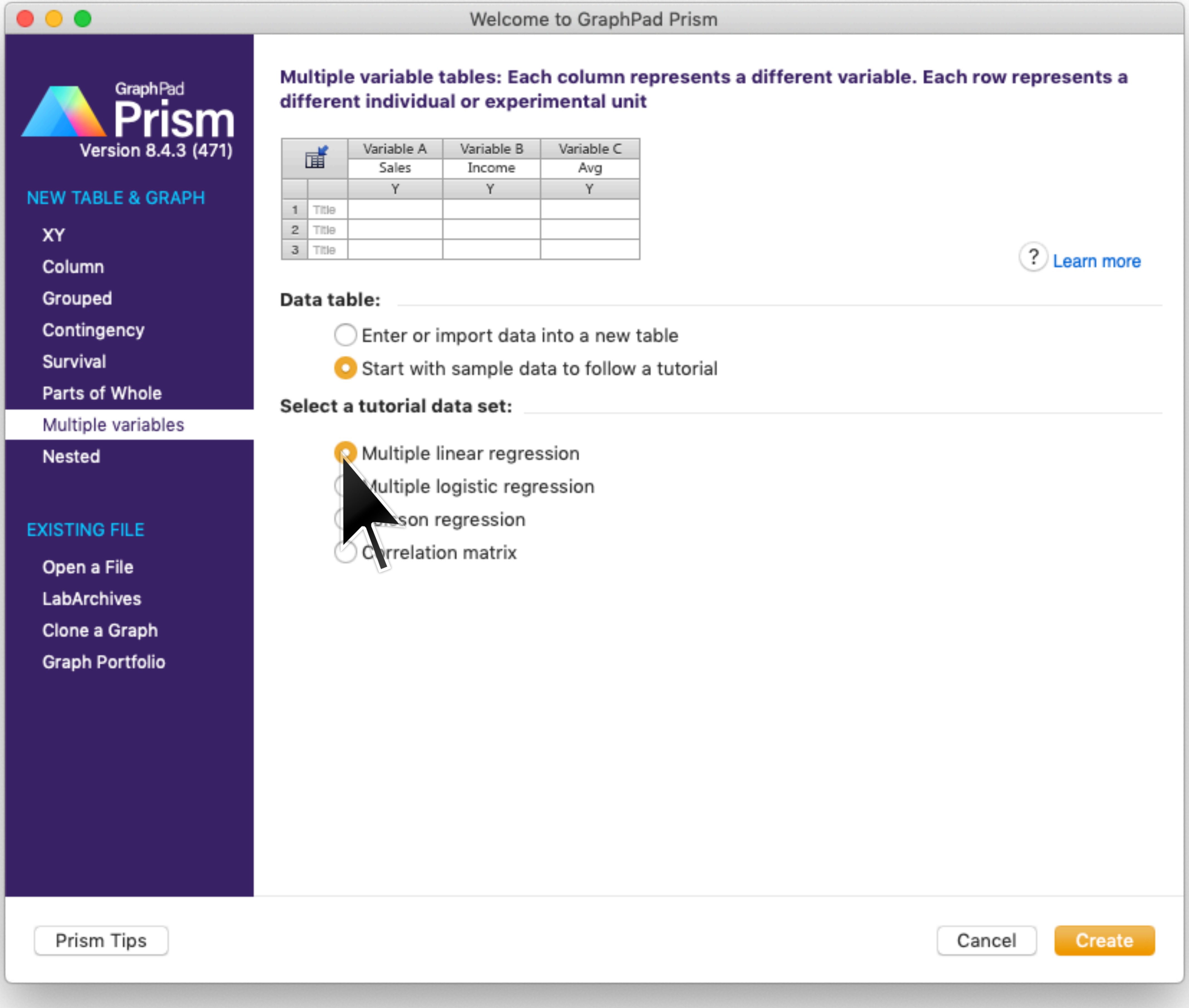
Logistic regression is used to model data where each data point $(\vec{x}_i, y_i)$ consists of a vector $\vec{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$ that represents D covariants, and one dependent variable y_i that is **binary**.

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

$$\log \mathrm{Odds_i} = \log \left[\frac{p(y_i = 1 \,|\, \vec{x}_i)}{p(y_i = 0 \,|\, \vec{x}_i)} \right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_D x_{iD}$$

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.



Multiple variable tables: Each column represents a different variable. Each row represents a different individual or experimental unit

		Variable A	Variable B	Variable C
		Sales	Income	Avg
		Y	Y	Y
1	Title			
2	Title			
3	Title			

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

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 .

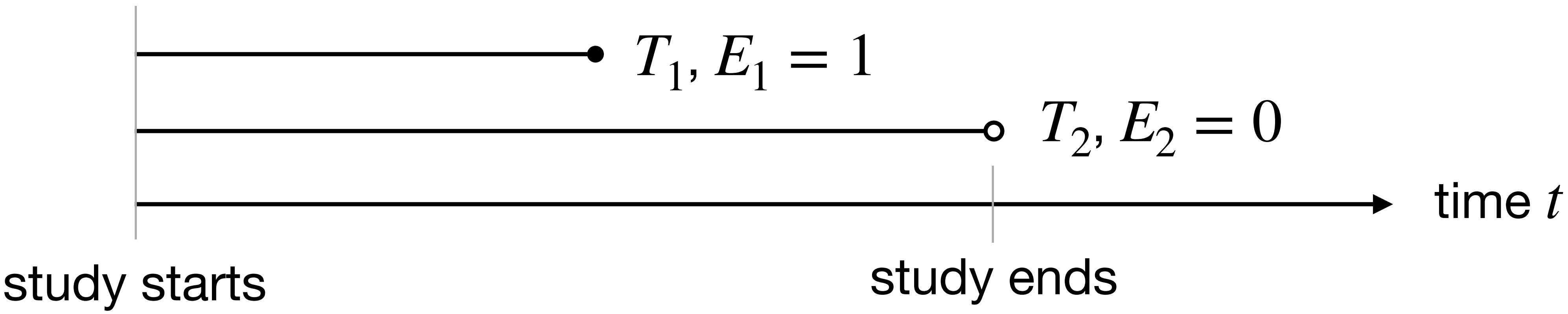
$$\hat{S}(t) = \frac{n(t)}{n(0)}$$

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:

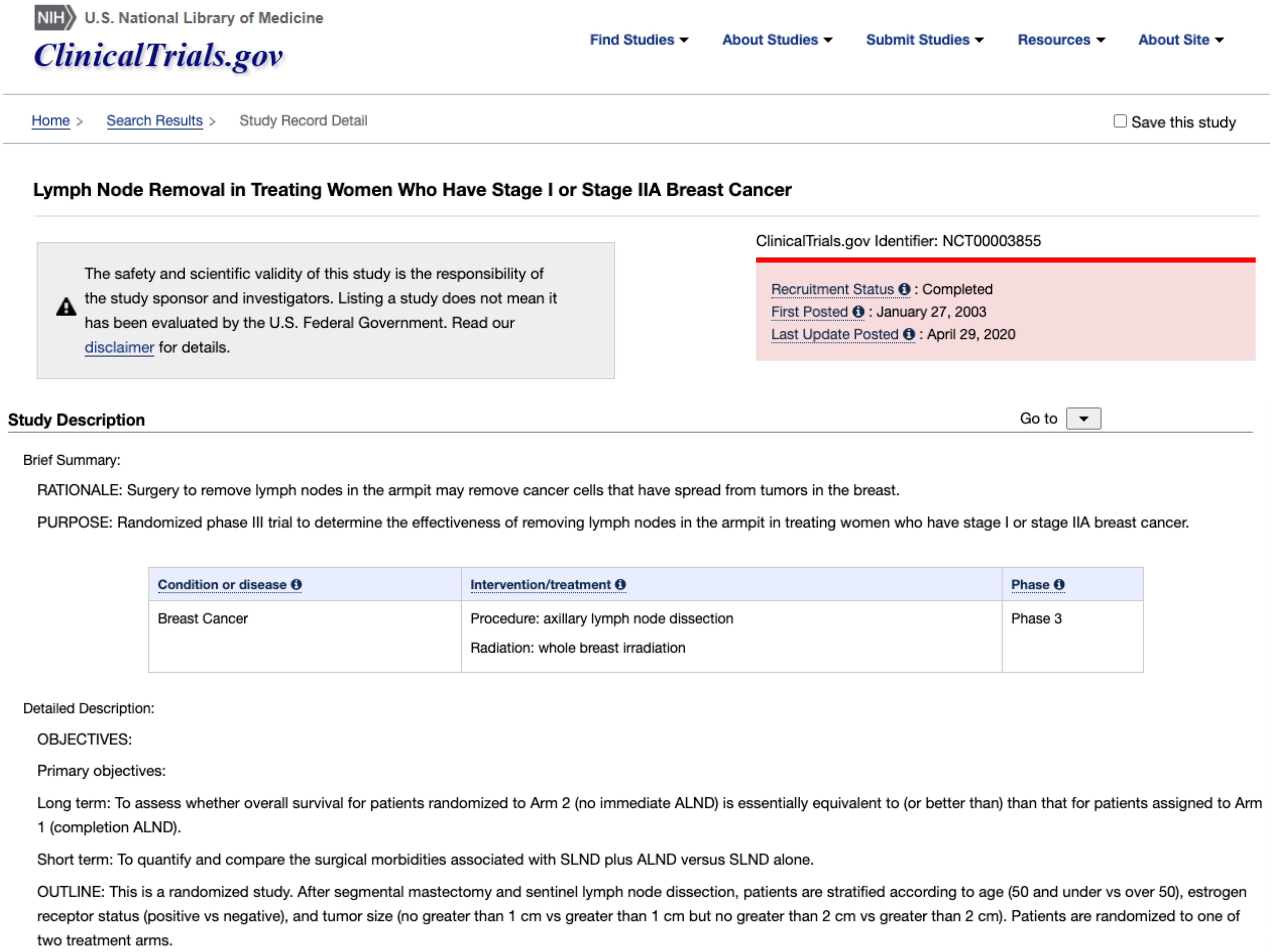
$$\hat{S}(t) = \prod_{i\,:\,T_i<t} \frac{n_i - d_i}{n_i}.$$

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

⚠ The safety and scientific validity of this study is the responsibility of the study sponsor and investigators. Listing a study does not mean it has been evaluated by the U.S. Federal Government. Read our disclaimer for details.

ClinicalTrials.gov Identifier: NCT00003855

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

Study Description

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

Detailed Description:

OBJECTIVES:

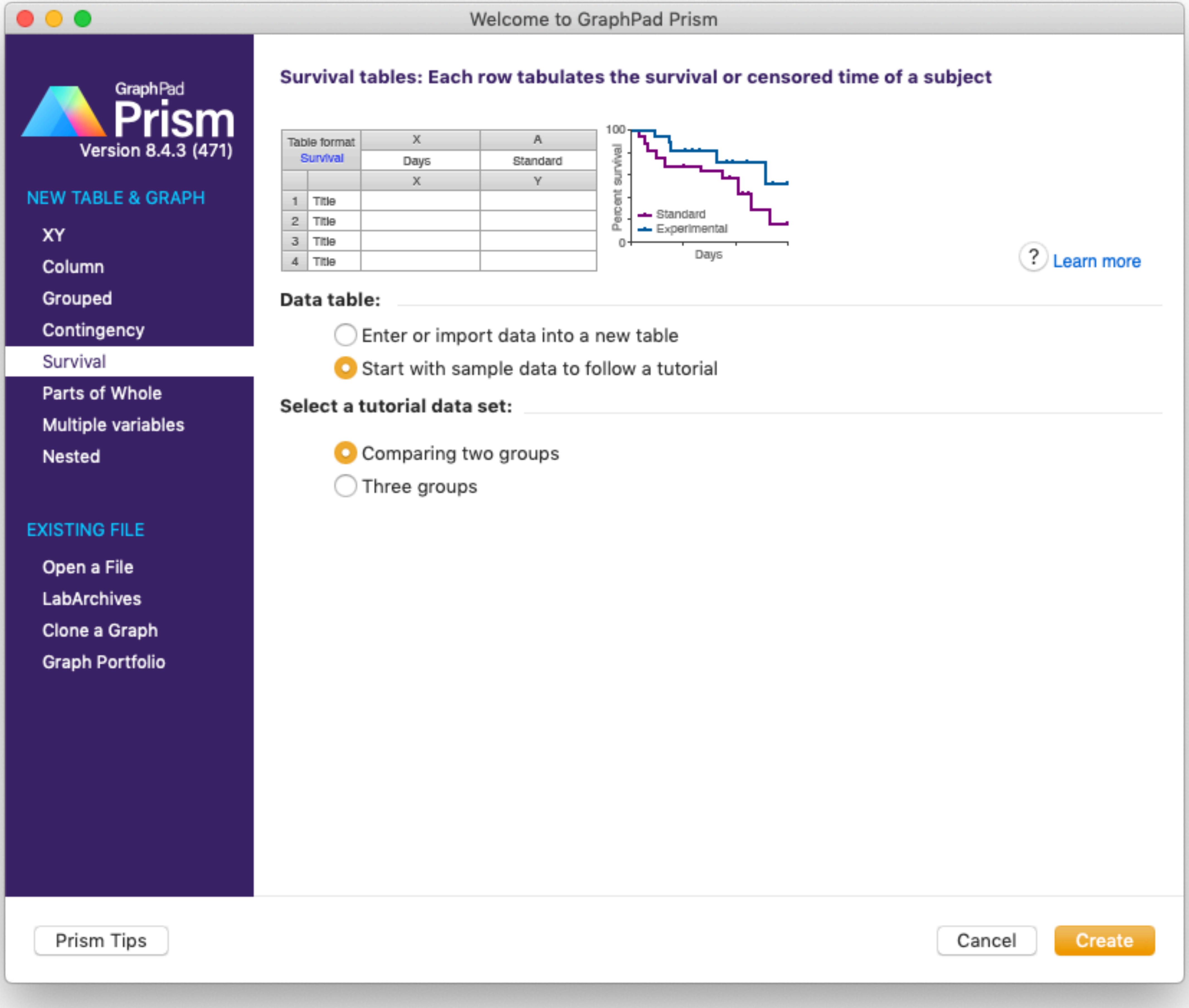
Primary objectives:

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.

https://clinicaltrials.gov/ct2/show/NCT00003855



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		

Data table:

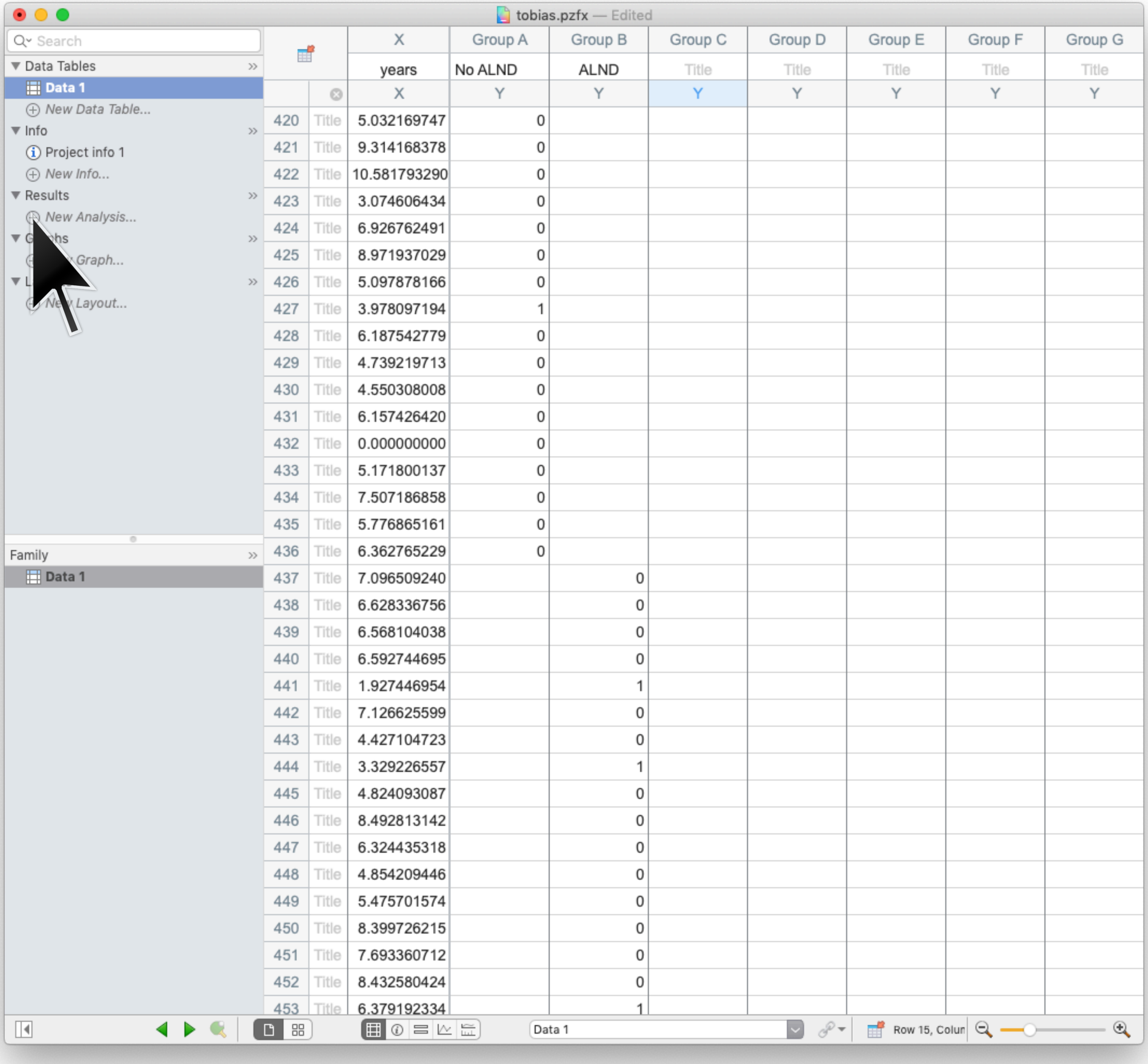
○ Enter or import data into a new table

● Start with sample data to follow a tutorial

Select a tutorial data set:

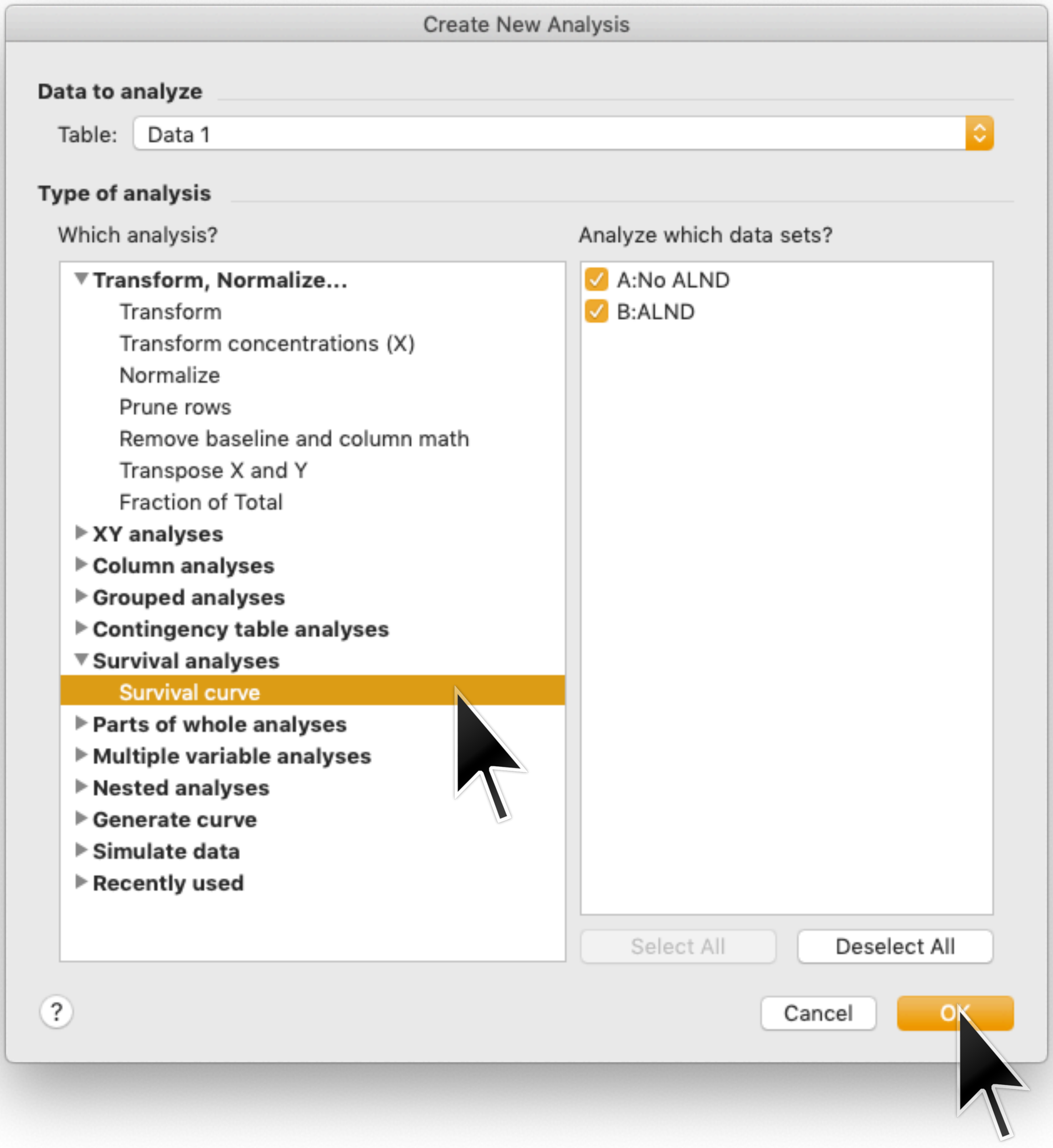
● Comparing two groups

○ Three groups



		X	Group A	Group B	Group C	Group D	Group E	Group F	Group G
		years	No ALND	ALND	Title	Title	Title	Title	Title
		X	Y	Y	Y	Y	Y	Y	Y
420	Title	5.032169747	0						
421	Title	9.314168378	0						
422	Title	10.581793290	0						
423	Title	3.074606434	0						
424	Title	6.926762491	0						
425	Title	8.971937029	0						
426	Title	5.097878166	0						
427	Title	3.978097194	1						
428	Title	6.187542779	0						
429	Title	4.739219713	0						
430	Title	4.550308008	0						
431	Title	6.157426420	0						
432	Title	0.000000000	0						
433	Title	5.171800137	0						
434	Title	7.507186858	0						
435	Title	5.776865161	0						
436	Title	6.362765229	0						
437	Title	7.096509240		0					
438	Title	6.628336756		0					
439	Title	6.568104038		0					
440	Title	6.592744695		0					
441	Title	1.927446954		1					
442	Title	7.126625599		0					
443	Title	4.427104723		0					
444	Title	3.329226557		1					
445	Title	4.824093087		0					
446	Title	8.492813142		0					
447	Title	6.324435318		0					
448	Title	4.854209446		0					
449	Title	5.475701574		0					
450	Title	8.399726215		0					
451	Title	7.693360712		0					
452	Title	8.432580424		0					
453	Title	6.379192334		1					

(data curtsey of Tobias Janowitz)



Create New Analysis

Data to analyze

Table: Data 1

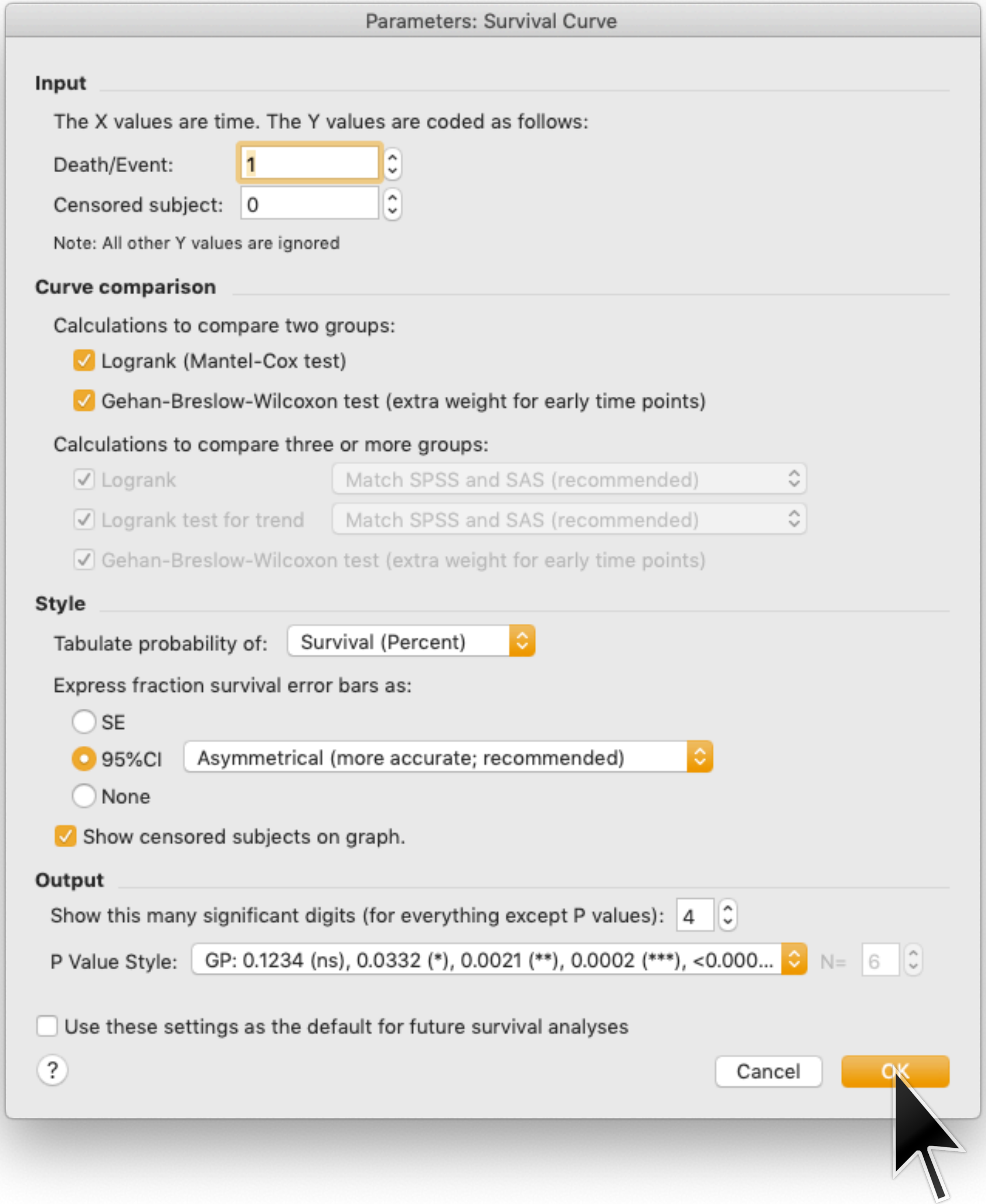
Type of analysis

Which analysis? Survival curve

Analyze which data sets?

☑ A:No ALND

☑ B:ALND



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

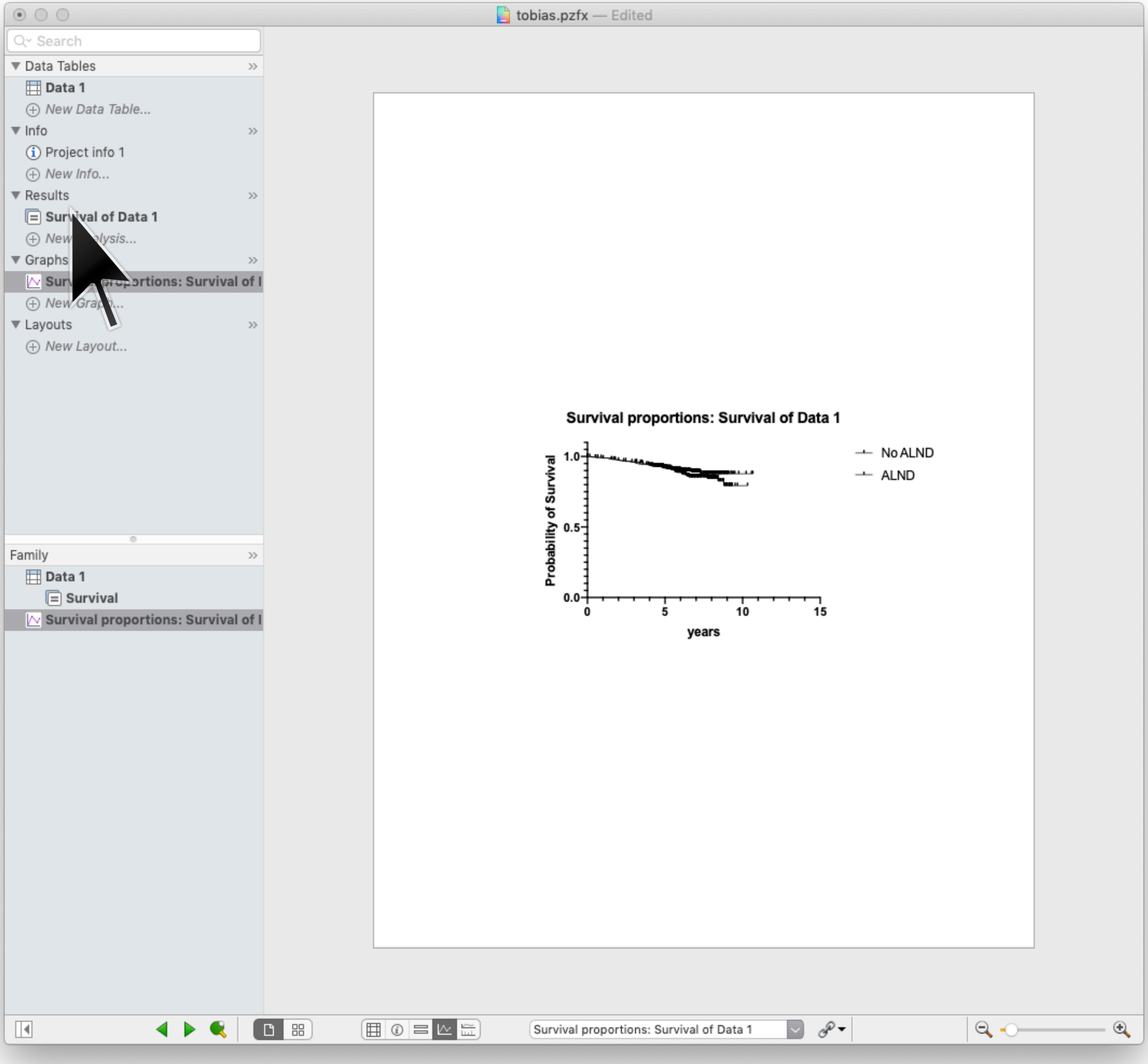
☑ Show censored subjects on graph.

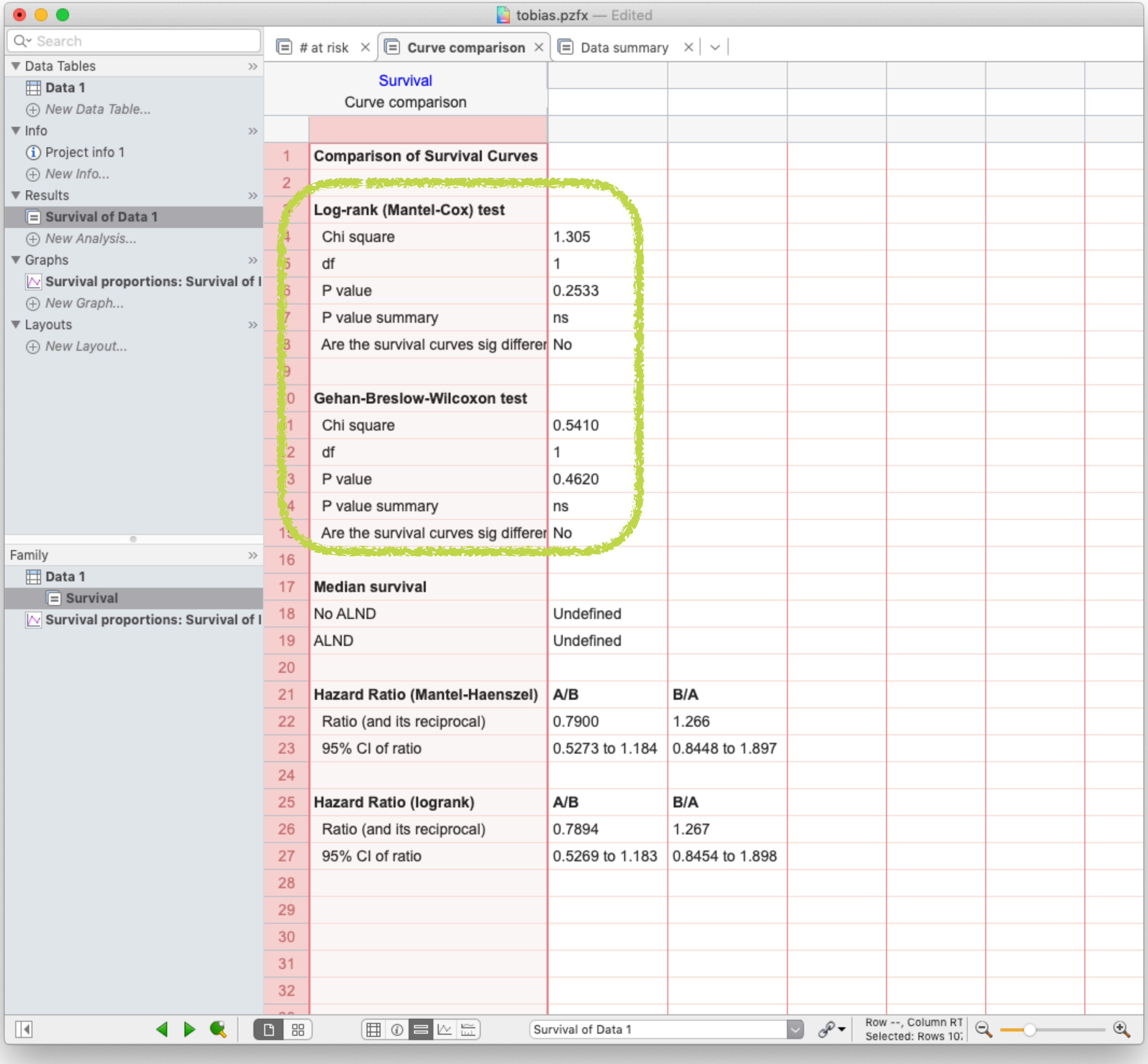
Output

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


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
Survival proportions: Survival of Data 1
Probability of Survival
1.0
0.5
0.0
0
5
10
15
years
No ALND
ALND
Survival proportions: Survival of Data 1



Survival
Curve comparison

1	**Comparison of Survival Curves**		
2			
3	**Log-rank (Mantel-Cox) test**		
4	Chi square	1.305	
5	df	1	
6	P value	0.2533	
7	P value summary	ns	
8	Are the survival curves sig differer	No	
9			
10	**Gehan-Breslow-Wilcoxon test**		
11	Chi square	0.5410	
12	df	1	
13	P value	0.4620	
14	P value summary	ns	
15	Are the survival curves sig differer	No	
16			
17	**Median survival**		
18	No ALND	Undefined	
19	ALND	Undefined	
20			
21	**Hazard Ratio (Mantel-Haenszel)**	**A/B**	**B/A**
22	Ratio (and its reciprocal)	0.7900	1.266
23	95% CI of ratio	0.5273 to 1.184	0.8448 to 1.897
24			
25	**Hazard Ratio (logrank)**	**A/B**	**B/A**
26	Ratio (and its reciprocal)	0.7894	1.267
27	95% CI of ratio	0.5269 to 1.183	0.8454 to 1.898
28			
29			
30			
31			
32			

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

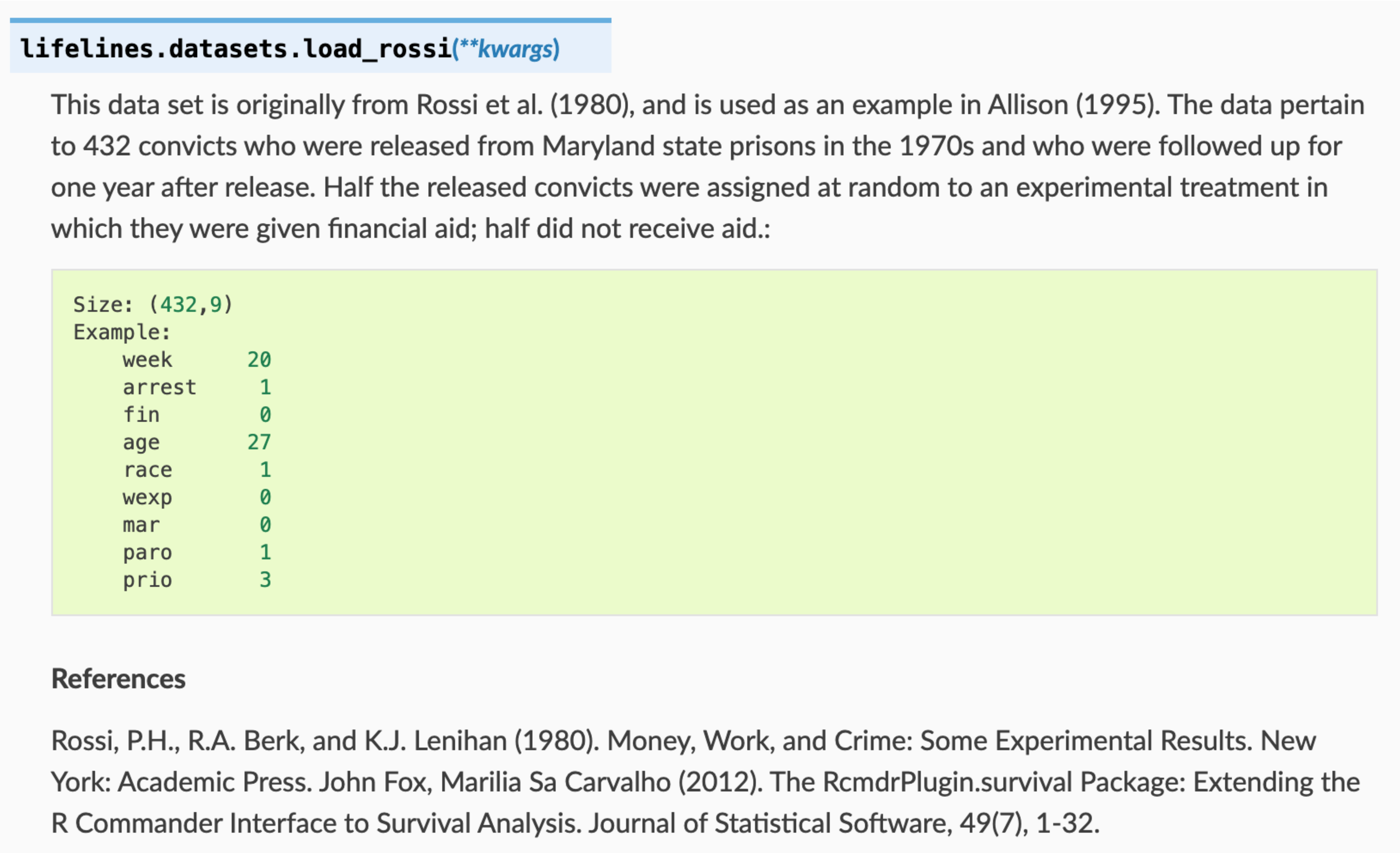
Suppose that each individual i has, in addition to an event time t_i and event flag, has a set of D covariants $x_{i1}, x_{i2}, \ldots, x_{iD}$, which can be either real numbers or binary.

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.:

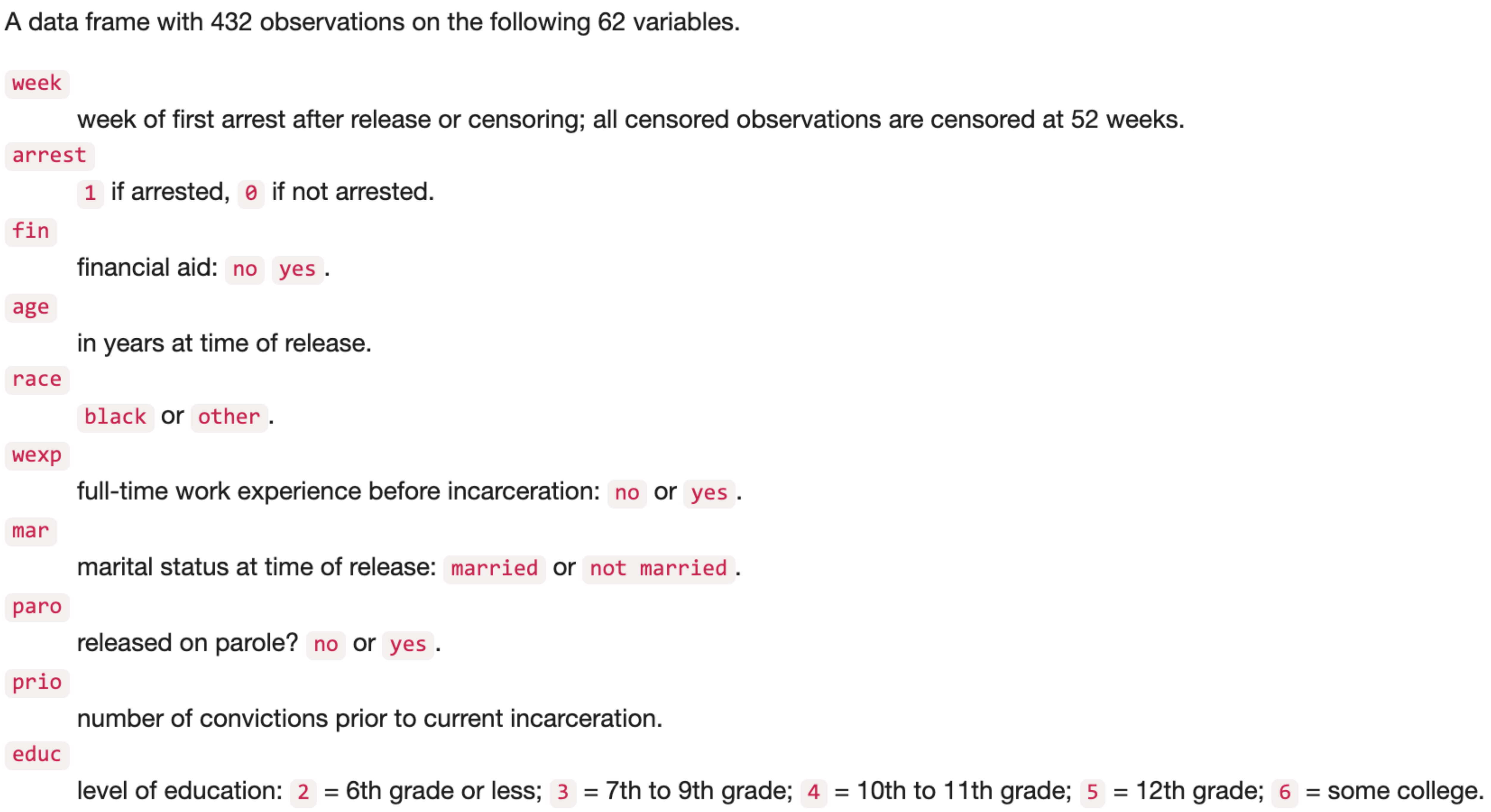
```
Size: (432,9)
Example:
    week      20
    arrest     1
    fin        0
    age       27
    race       1
    wexp       0
    mar        0
    paro       1
    prio       3
```

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`

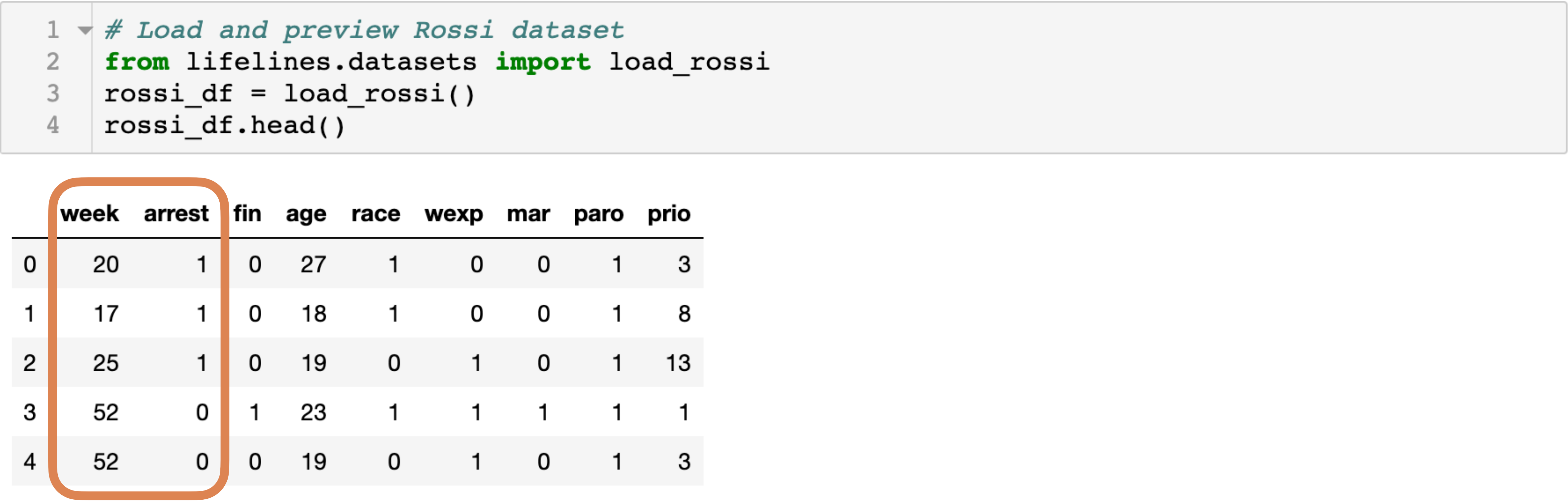
number of convictions prior to current incarceration.

`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



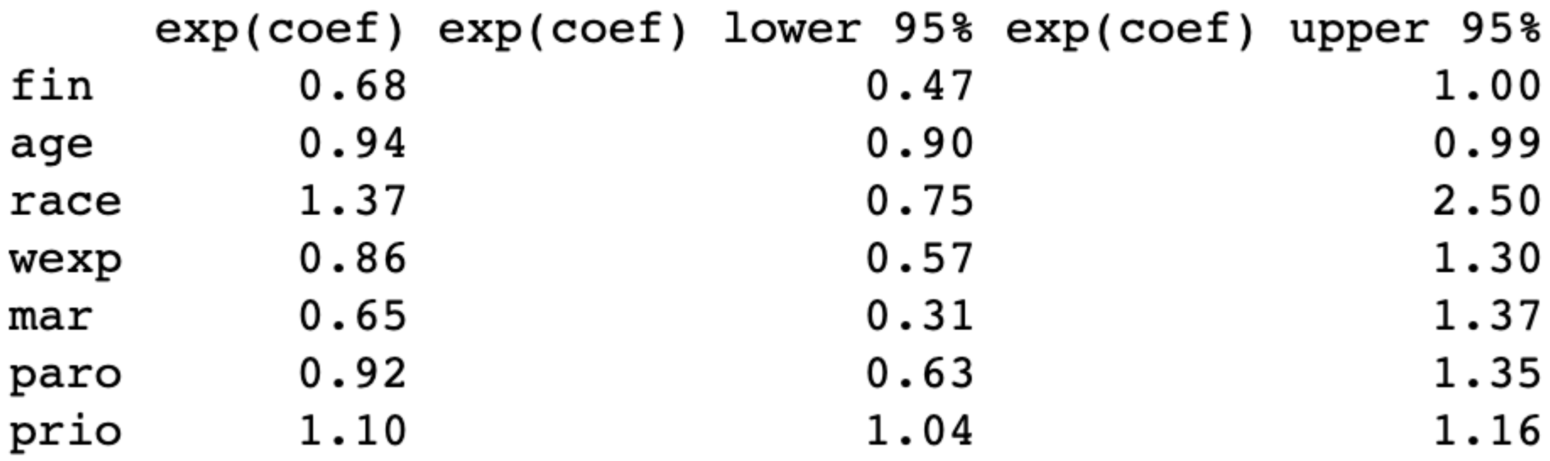
```python
# Load and preview Rossi dataset
from lifelines.datasets import load_rossi
rossi_df = load_rossi()
rossi_df.head()
```

	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)

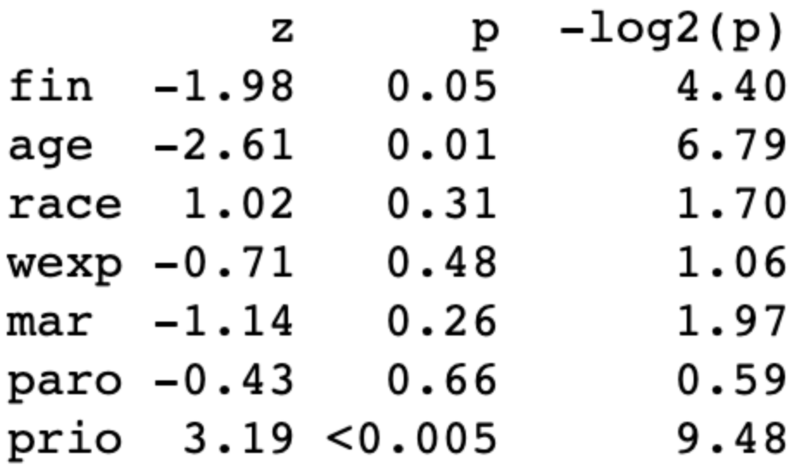
The results of Cox Regression is a statement about the effect size and significance of each variable

effect size (hazard)



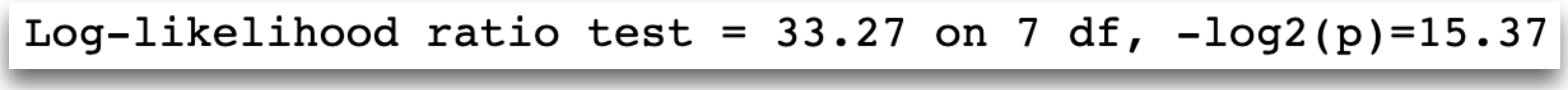
```
     exp(coef) exp(coef) lower 95% exp(coef) upper 95%
fin       0.68                0.47                1.00
age       0.94                0.90                0.99
race      1.37                0.75                2.50
wexp      0.86                0.57                1.30
mar       0.65                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.

10:00a - 12:00p. Finished right on time, though rather rushed at the end.