Chapter 15: Correlation and regression - Relationships between measured values

Regression Analysis

TXCL7565/PHSC7565

What This Lecture Covers

- Simple statistical models
- Linear regression analysis
- Best fit line for the data
- Making predictions using the regression equation
- Reverse calculation

SIMPLE STATISTICAL MODEL

Simple Statistical Model

All statistical models can be represented by the equation:

$$response_i = model + error_i$$

This means that every response can be predicted from the model we choose to fit to the data plus some error.

Simple Statistical Models (cont.)

- The model is typically specified in terms of parameters, which are unknown constants that describe a population characteristic.
    - Parameters are usually represent by Greek letters such as: μ, σ, and β

- A statistic is an estimate of a parameter calculated from observed data.
    - Statistics are typically represented by phonetic letters (or Greek letter with a symbol) such as: $\bar{x}$, s, and $\hat{\beta}$

Ex. The mean as a statistical model

Suppose we would like to summarize the number of friends statistics instructors have using the model:

$$friends_i = \mu + error_i$$

where $\boldsymbol{\mu}$ is the mean number of friends for all statistics instructors.

Suppose we randomly selected 5 statistics instructors and measure the number of friends that they have. We observe 1, 2, 3, 3, 4. The mean number of friends is $\bar{x} = 2.6$

Our model is

Assessing the fit of a model

The fit of a model is often characterized using the deviance and sum of squares.

The **deviance** for observation ***i*** is the difference between the response for observation ***i*** and the estimated model.

- Deviance measures the amount of error in our predication of response ***i***.
- Using a formula, the deviance is

$$deviance_i = response_i - model$$

- Negative deviance indicates that our model overestimated the response
- Positive deviance indicates that our model underestimated the response

Assessing the fit of a model (cont.)

The sum of squared errors (SS) is the sum of the squared deviances and is a way to measure the fit of a model.

- The larger the SS is, the more poorly our model fits the data

Assessing the fit of a model

mean as a statistical model

Example: Friends/statistics instructors

- Our estimate of the population mean is 2.6 friends
- The deviances are given by the formula:

$$deviance_i = x_i - \bar{x}$$

where x_i is the response for observation i

Note: the sample variance is $s^2 = SS/(n-1)$

and the sample standard deviation is $s = \sqrt{SS/(n-1)} = \sqrt{s^2}$

LINEAR REGRESSION ANALYSIS

What is regression?

- Regression analysis provides a way of describing the distribution of a response (outcome/dependent) variable as a function of one or more explanatory (predictor/independent) variables.

- The **regression of the response variable on the explanatory variable** describes the mathematical relationship between the response variable and the explanatory variable.

Simple vs Multiple Linear Regression

- Simple linear regression - regression assuming that the relationship between a single explanatory variable and the response variable is linear.

- Multiple linear regression - regression using multiple explanatory variables that assumes the relationship can be written as a linear equation.

We will use the basic model

$$response_i = model + error_i$$

where model will be replaced by a regression model.

Simple linear regression model

The **simple linear regression model** postulates that the *mean* of the response variable $\boldsymbol{Y}$, as a function of a single explanatory variable $\boldsymbol{X}$ denoted by $\mu_{Y|X}$, is given by the linear relationship

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

This is called the **regression line**.

Relationship between Y_i and X_i

The relationship between the response for the ith observation (Y_i) and the value of the explanatory variable of the ith observation (X_i) is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Notice that

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ &= \mu_{Y|X} + \epsilon_i \\ &= model + \epsilon_i \end{aligned}$$

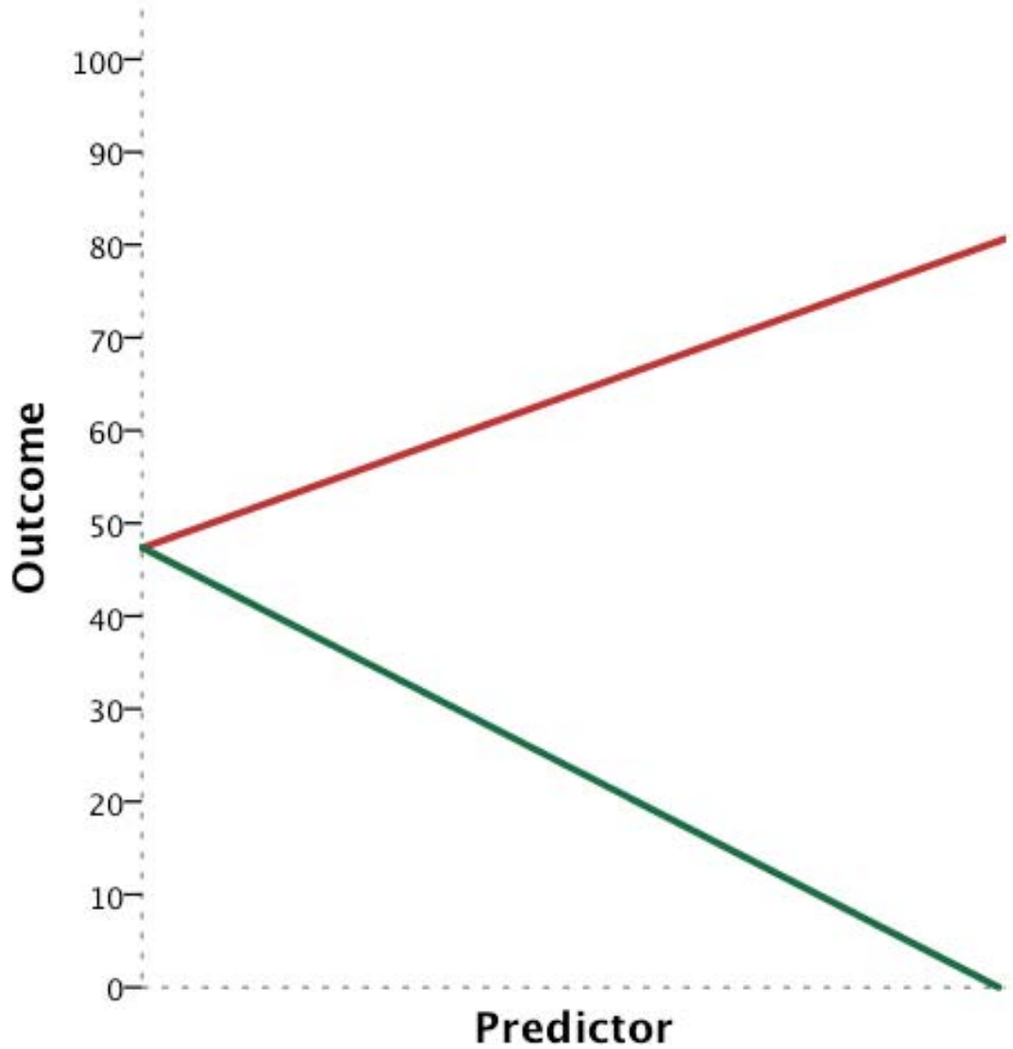
Thus, the value for the ith observation Y_i is the mean of Y when $X = X_i$ plus some random error.

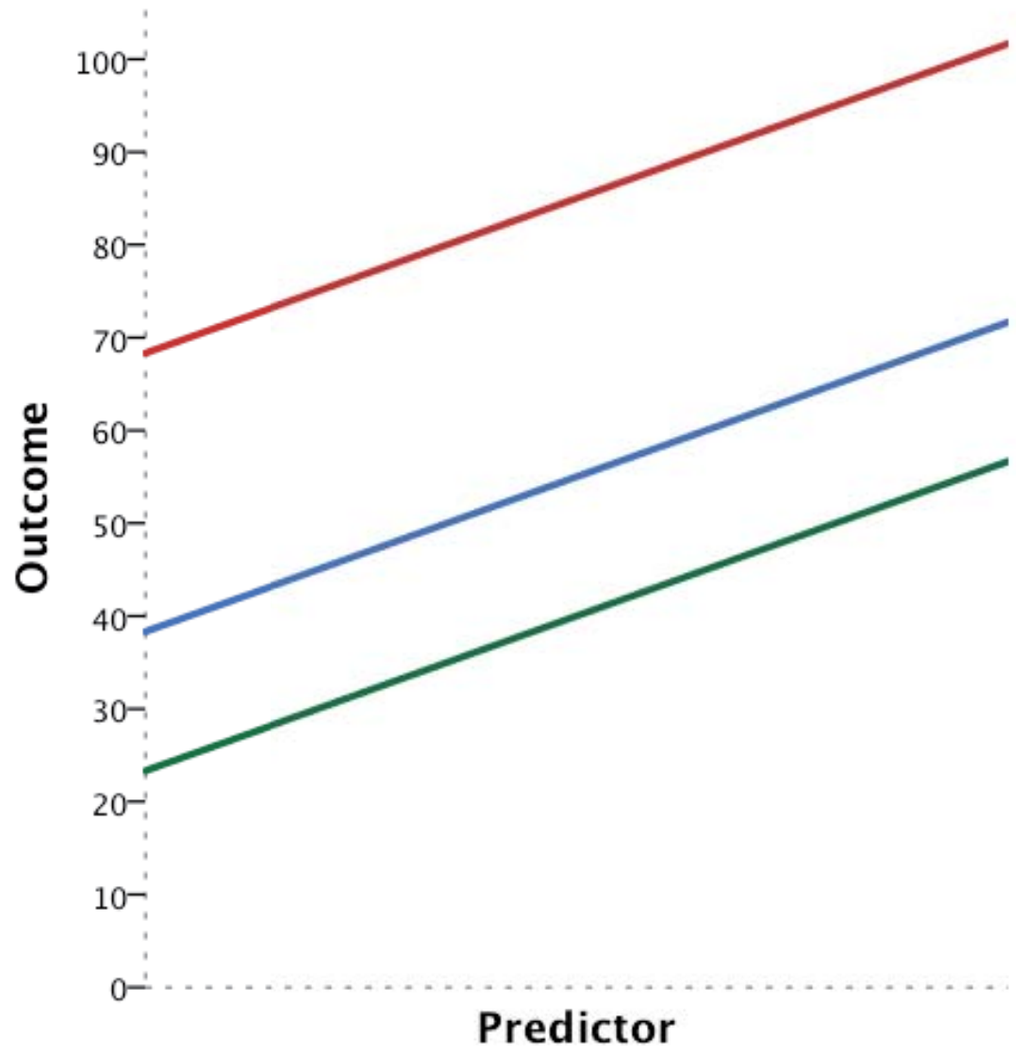
Terms in a simple linear regression model

- β_0 is the *intercept* of the model
  - It is the mean value for Y when $X = 0$.
  - It is where the regression line crosses the y-axis.
- β_1 is the *slope* of the model
  - It is the mean change in Y when the explanatory variable X increases by 1 unit.
  - Describes how the response variable is affected by the explanatory variable.

β_0 and β_1 are called **regression coefficients**.

Relationship between intercepts and slopes



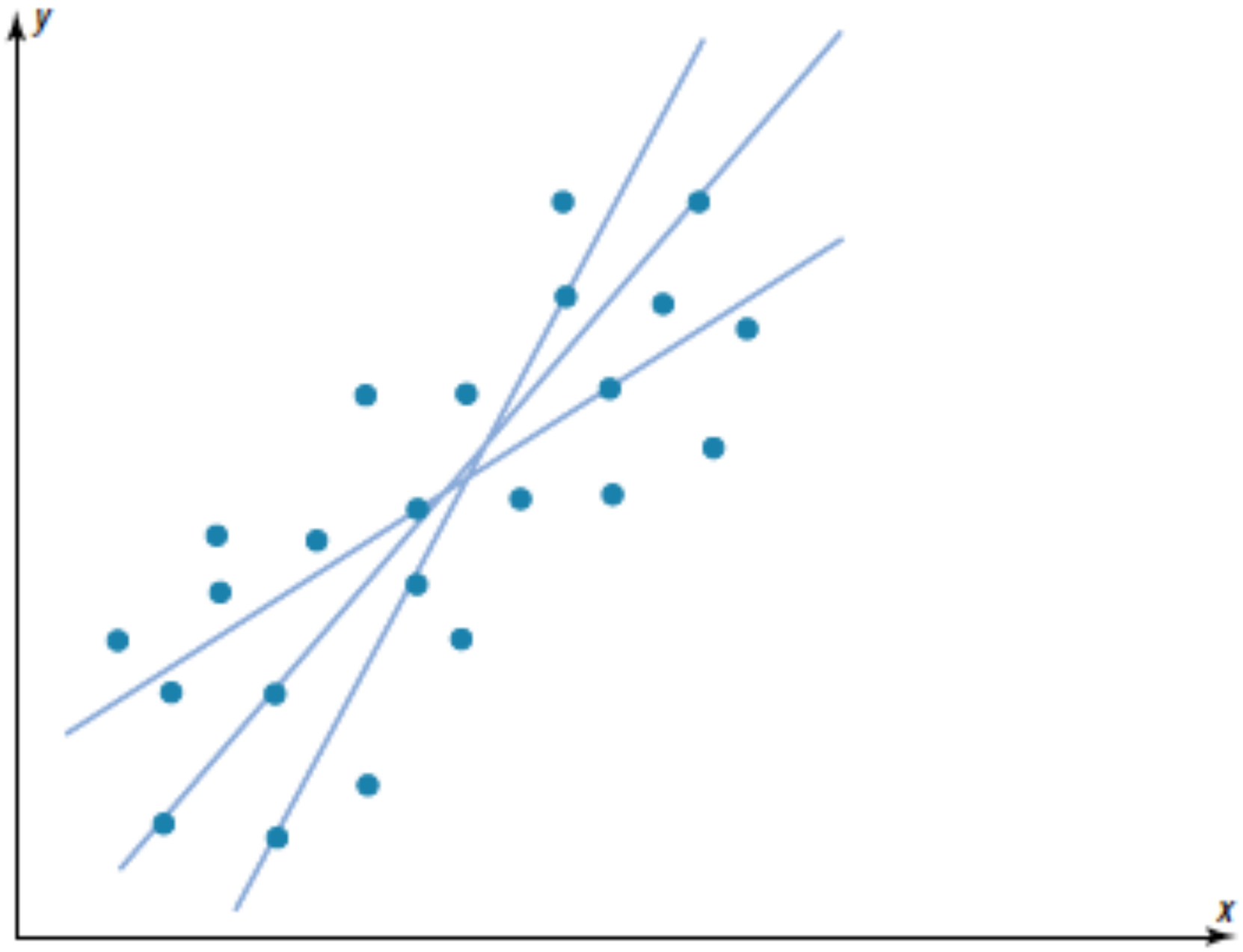


BEST FIT LINE FOR THE DATA

The method of least squares

We want to describe the relationship between the response variable and the explanatory variable with a regression line.

How do we decide which model is the “best” model since there are many possible straight lines?


y
x

Method of least squares

Method of least squares - a regression procedure that estimates the regression coefficients with the values that minimize the sum of the squared deviations (residuals).

The estimated mean response for Y_i is $\hat{Y}_i = \mu_{\hat{Y}|X} = \hat{\beta}_0 + \hat{\beta}_1 X_i$

where β_0 and β_1 are the estimated values of β_0 and β_1.

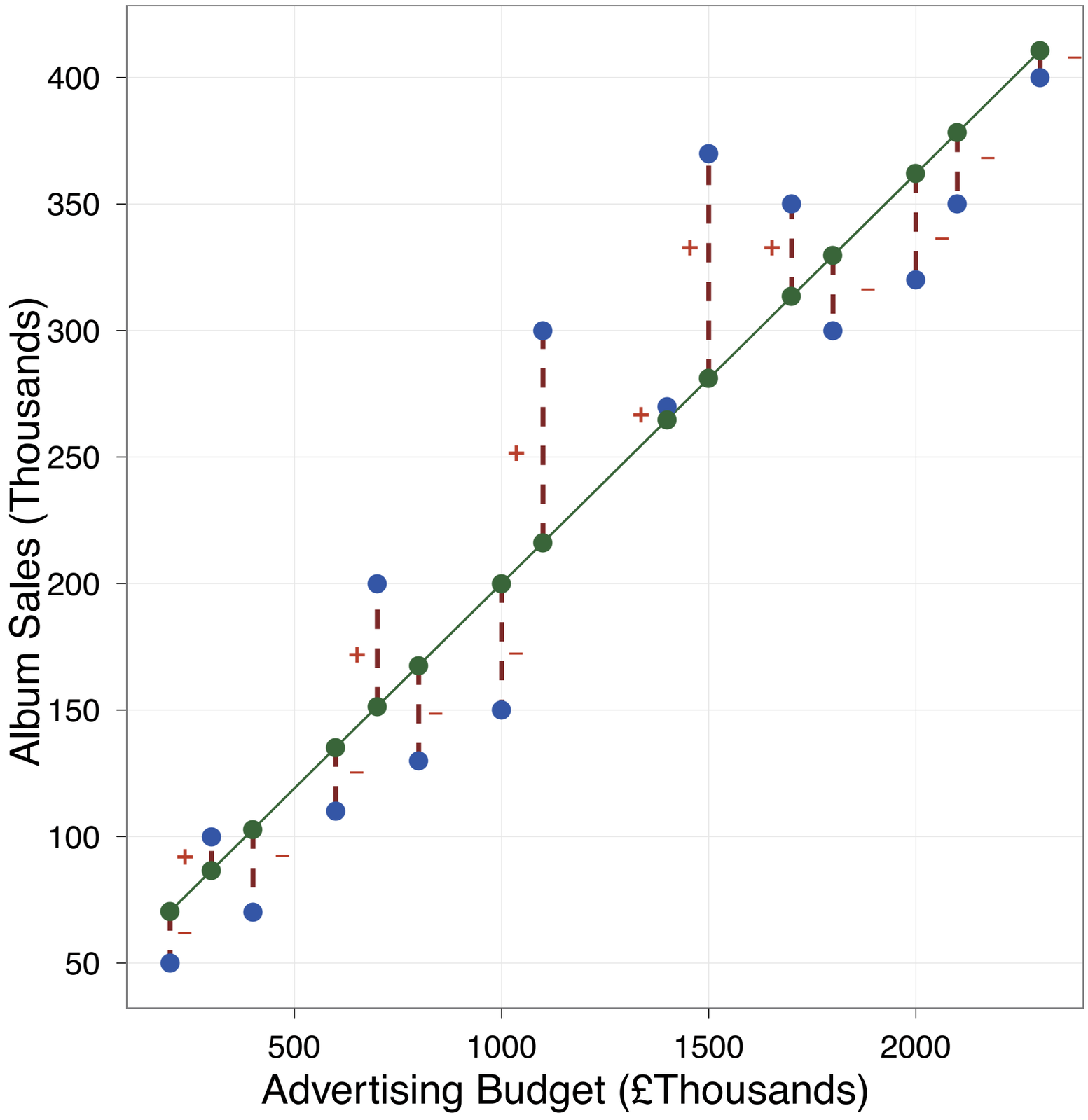
$\hat{Y}_i$ is known as the ***i*th fitted value**.

The *i*th residual (deviation) is $e_i = Y_i - \hat{Y}_i$

- This is the difference between the observed response and the estimated response for the *i*th observation.

The least squares method finds the $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} e_i^2$$


400
350
300
250
200
150
100
50
Album Sales (Thousands)
500
1000
1500
2000
Advertising Budget (£Thousands)

SIMPLE LINEAR REGRESSION - EXAMPLE

Fungal Toxin in Nuts

A drug precursor molecule is extracted from a type of nut. The nuts are commonly contaminated by a fungal toxin that is difficult to remove during the purification process. We suspect that the amount of fungus (and hence toxin) depends on the rainfall at the growing site. We would like to predict toxin concentration from rainfall in order to judge whether it would be worth paying additional rental charges for relatively drier sites. We analyze the toxin content in a series of batches of nuts and we know the rainfall at the growing sites during the four months when the nuts are forming.

Is this an observational study or a randomized experiment?

Can we make cause-and-effect conclusions from this study?

State the null and alternative hypothesis.

Table 15.3 Rainfall at the growing site and concentration of fungal toxin in nuts

Rainfall (cm.week^{-1})	Toxin (μg per 100 g)
1.30	18.1
2.28	28.6
1.11	15.9
0.74	19.2
1.32	19.3
0.51	14.8
1.56	21.7
1.32	16.5
2.05	23.8
1.37	19.0

Welcome to GraphPad Prism

GraphPad PRISM® 7.0a

New table & graph

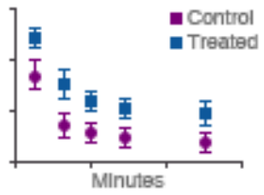
- XY
- Column
- Grouped
- Contingency
- Survival
- Parts of Whole

Existing file

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

XY tables: Each point is defined by an X and Y coordinate

		X	A			B		
		Minutes	Control			Treated		
		X	A:Y1	A:Y2	A:Y3	B:Y1	B:Y2	B:Y3
1	Title							
2	Title							
3	Title							



? Learn more

Enter/import data:

X:
- (•) Numbers
- ( ) Numbers with error values to plot horizontal error bars
- ( ) Dates
- ( ) Elapsed times

Y:
- (•) Enter and plot a single Y value for each point
- ( ) Enter 5 replicate values in side-by-side subcolumns
- ( ) Enter and plot error values already calculated elsewhere

Enter: Mean, SD, N

Use tutorial data:
- ( ) Linear regression - Compare slopes
- ( ) Nonlinear regression -- One phase exponential decay
- ( ) Dose-response - X is log(dose)
- ( ) Interpolate unknowns from a linear standard curve
- ( ) Correlation
- ( ) Entering dates into the X column
- ( ) Entering elapsed times into the X column
- ( ) More tutorial data...

Prism Tips

Cancel Create

fungalToxin

File | Sheet | Undo | Clipboard | Analysis | Change | Import | Draw | Write

New | Analyze

- Family
- Search results
- Data Tables
  - fungalToxin
- Info
  - Project info 1
- Results
- Graphs
  - fungalToxin
- Layouts

Table format: XY		X	Group A	Group B
		rainfall (cm.week^{-1})	Toxin (microgram per 100 grams)	Title
		X	Y	Y
1	Title	1.30	18.1	
2	Title	2.28	28.6	
3	Title	1.11	15.9	
4	Title	0.74	19.2	
5	Title	1.32	19.3	
6	Title	0.51	14.8	
7	Title	1.56	21.7	
8	Title	1.32	16.5	
9	Title	2.05	23.8	
10	Title	1.37	19.0	
11	Title			
12	Title			
13	Title			
14	Title			

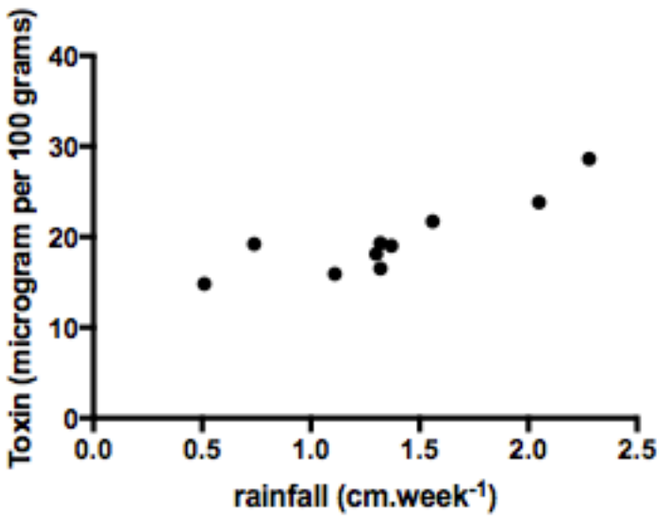
fungalToxin | Row 16, A: Toxin (micro...

fungalToxin — Edited

File Sheet Undo Clipboard Analysis Change Arrange Draw Write

New Analyze

- Family
- Search results
- Data Tables
  - fungalToxin
- Info
  - Project info 1
- Results
- Graphs
  - fungalToxin
- Layouts



fungalToxin

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 = 0.51

End regression line at:

◉ Auto

○ X = 2.28

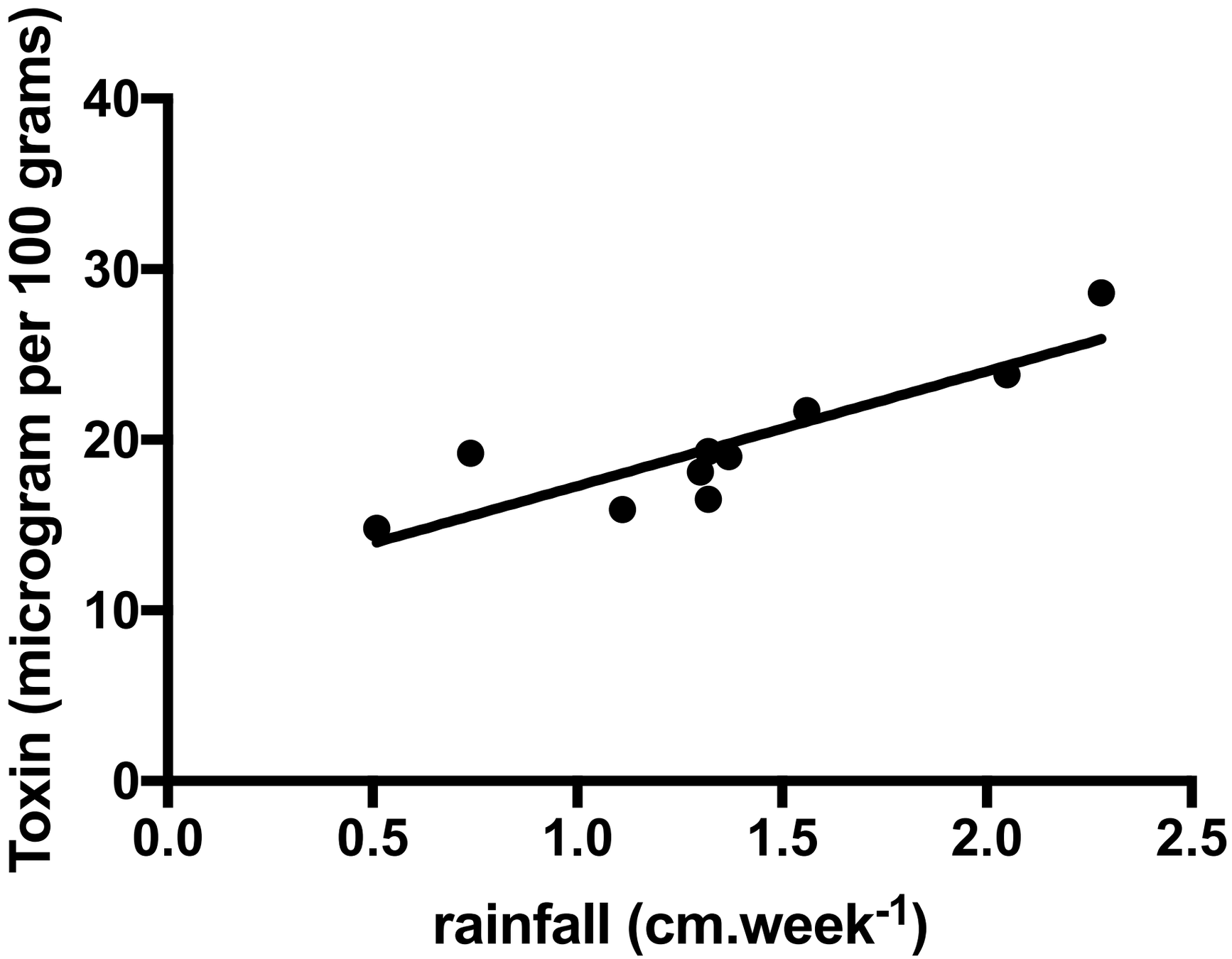
Output options

☐ Show table of XY coordinates

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

Show 4 significant digits.

? More choices... Cancel OK


Toxin (micrograms per 100 grams)
40
30
20
10
0
0.0
0.5
1.0
1.5
2.0
2.5
rainfall (cm.week-1)

fungalToxin — Edited

File Sheet Undo Clipboard Analysis Interpret Change Draw Write Text Export Print Send LA

New Analyze 14 Arial

- Family
- Search results
- Data Tables
  - fungalToxin
- Info
  - Project info 1
- Results
  - Linear reg. of fungalToxin
- Graphs
  - fungalToxin
- Layouts

Linear reg.

		A	B	C	D	E
		Toxin (microgram per 100 grams)	Title	Title	Title	Title
		Y	Y	Y	Y	Y
1	Best-fit values ± SE					
2	Slope	6.726 ± 1.356				
3	Y-intercept	10.57 ± 1.961				
4	X-intercept	-1.571				
5	1/slope	0.1487				
6						
7	95% Confidence Intervals					
8	Slope	3.6 to 9.852				
9						
10	Goodness of Fit					
11	R square	0.7547				
12	Sy.x	2.16				
13						
14	Is slope significantly non-zero?					
15	F	24.61				
16	DFn, DFd	1, 8				
17	P value	0.0011				
18	Deviation from zero?	Significant				
19						
20	Equation	Y = 6.726*X + 10.57				
21						
22	Data					
23	Number of X values	10				

Linear reg. of fungalToxin Tabular results

Linear reg.		A
		Toxin (microgram per 100 grams)
		Y
1	Best-fit values ± SE	
2	Slope	6.726 ± 1.356
3	Y-intercept	10.57 ± 1.961
4	X-intercept	-1.571
5	1/slope	0.1487
6		
7	95% Confidence Intervals	
8	Slope	3.6 to 9.852
9		
10	Goodness of Fit	
11	R square	0.7547
12	Sy.x	2.16
13		
14	Is slope significantly non-zero?	
15	F	24.61
16	DFn, DFd	1, 8
17	P value	0.0011
18	Deviation from zero?	Significant
19		
20	Equation	Y = 6.726*X + 10.57
21		
22	Data	
23	Number of X values	10

Intercept =
Slope =

Regression Model:

MAKING PREDICTIONS USING THE REGRESSION EQUATION

Having obtained the regression equation, we might now have the chance to rent two agricultural locations where we could grow a crop of nuts. We show that the weekly rainfall at Sites A and B during the fruiting season are 2.05 and 1.25 cm/week respectively. Therefore, we can predict the level of toxin in nuts grown at these two sites.

Site A

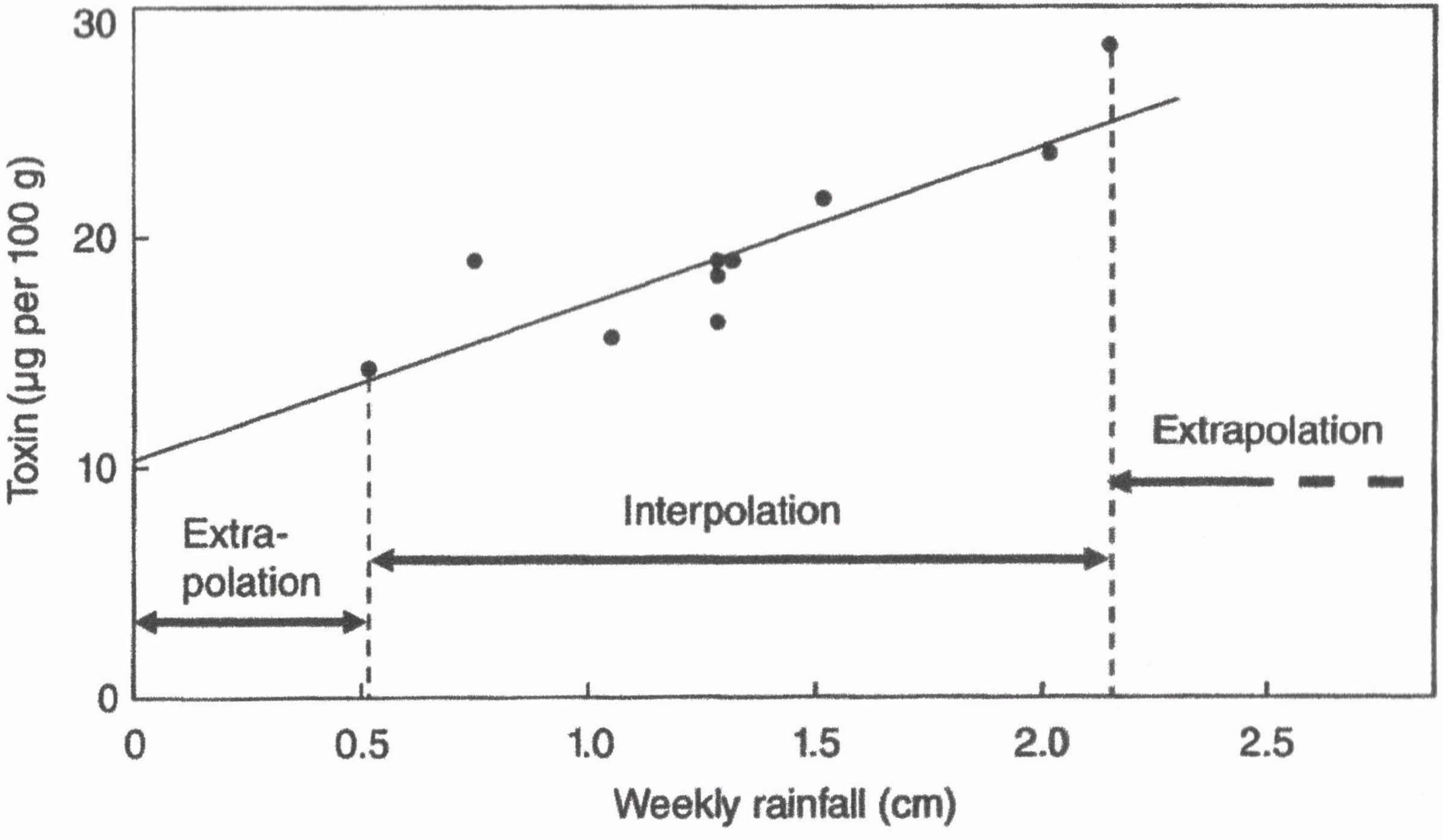
$$
\begin{aligned}
Toxin &= 10.6 + 6.73 \times Rainfall \\
&= 10.6 + 6.73 \times 2.05 \\
&= 10.6 + 13.8 \\
&= 24.4
\end{aligned}
$$

Site B

Interpolation and Extrapolation

- **Interpolation**: A prediction using a value of the independent variable that is within the observed range - uncontroversial.

- **Extrapolation**: A prediction using a value of the independent variable that lies outside the observed range. Extrapolation should be avoided unless there is sound reason to believe that the linear relationship extends beyond the observed range.


30
20
10
0
Toxin (µg per 100 g)
Extra-
polation
Interpolation
Extrapolation
0
0.5
1.0
1.5
2.0
2.5
Weekly rainfall (cm)

REVERSE CALCULATION

Reverse calculation

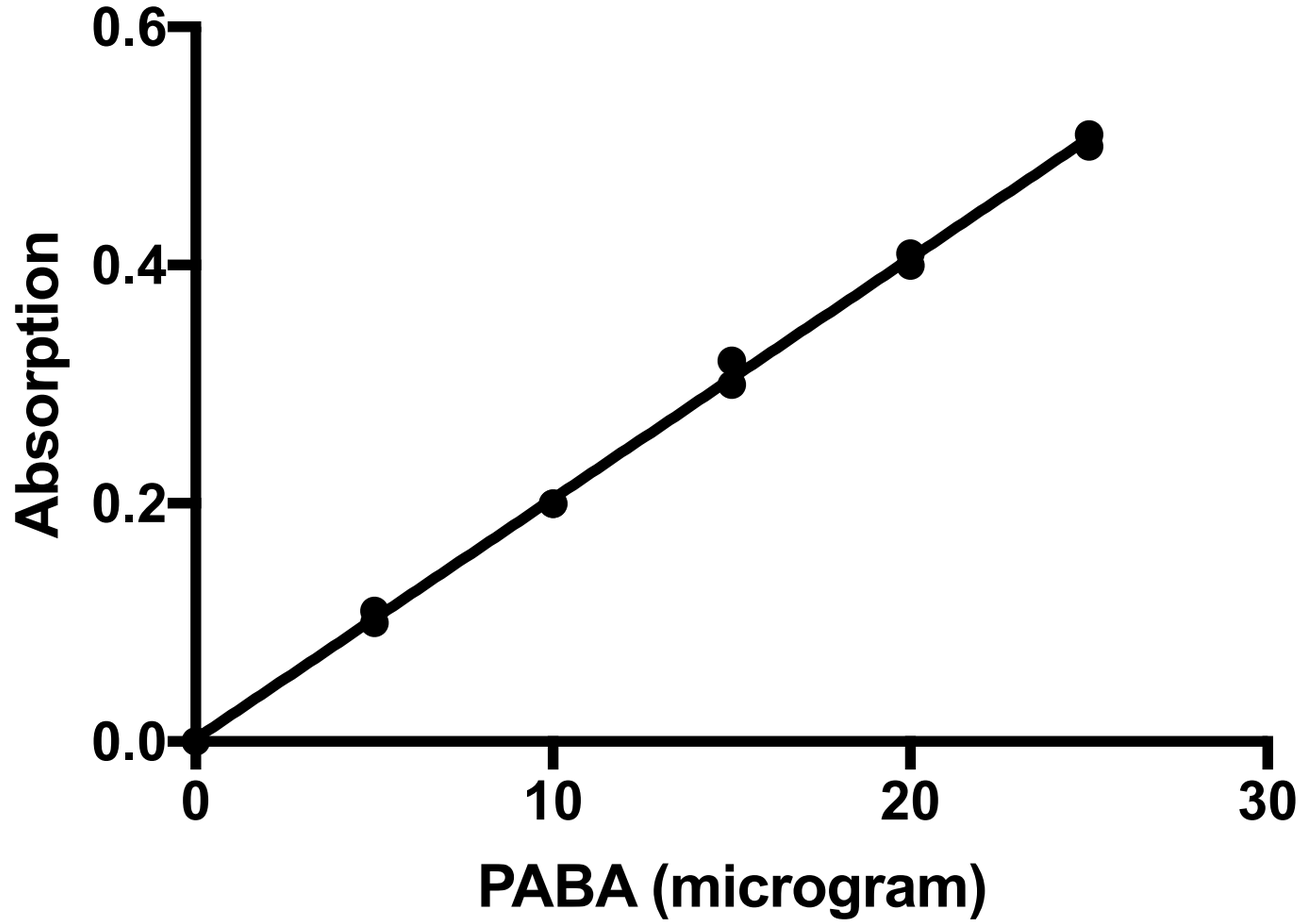
The normal purpose of regression is to obtain a value for the dependent variable from the independent variable.

However, there are times when we want to operate in the opposite direction, e.g., analytical methods that use a calibration curve.

Reverse calculation: To predict the value of the independent variable from the dependent variable, the 'normal' equation (predicting dependent from independent) is calculated initially and is then re-arranged.

Colorimetric example

PABA (µg) - STANDARD	Absorption
0	0.00
5	0.11
5	0.10
10	0.20
10	0.20
15	0.30
15	0.32
20	0.41
20	0.40
25	0.50
25	0.51



$$Absorption = 0.0023 + 0.0202 \times PABA(\mu g)$$

Reverse calculation

$$PABA(\mu g) = \frac{Absorption - 0.0023}{0.0202}$$

If we then have an unknown sample with an absorption of 0.41, its PABA content is estimated to be:

What did we learn?

- Correlation coefficients measures the magnitude and direction of the linear association between two factors.

- It is essential that you graph the data using a scatterplot prior to conducting a correlation analysis to check for a radical deviation from a linear relationship