PHAR2821

Drug Discovery and Design B – Statistics

GARTH TARR

SEMESTER 2, 2013

THE UNIVERSITY OF SYDNEY

Housekeeping

Contact details

- Email: garth.tarr@sydney.edu.au
- Room: 806 Carslaw Building
- Consultation: by appointment (email to arrange a time)

Tutorials

- Weeks 2 and 3 (check your timetable for details)

Quiz

- Week 4: Thursday 22nd August at 2:00

Online resources

- sydney.edu.au/science/maths/u/gartht/PHAR2821

Calculator

You need to bring a (non-programmable) calculator with you to all lectures, tutorials and the quiz!

Outline

1. Correlation coefficient
   - calculation
   - properties
   - interpretation
2. Simple linear regression:
   - Least squares method and the assumptions used
   - Residual plots and interpretation
   - Using Excel to perform linear regression analysis
   - Hypothesis testing for the slope parameter
   - Interpretation of the slope parameter
3. Linearising transformations:
   - Allometric
   - Exponential

Outline

Bivariate data

- Observations are sometimes observed in pairs, i.e. two measurements may be observed on each person in a study.
- We can write n observations as:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

- Often data will be stored in tables:

Observation (i)	1	2	3	...	n
Variable 1 (x_i)	x_1	x_2	x_3	...	x_n
Variable 2 (y_i)	y_1	y_2	y_3	...	y_n

Example (Bivariate data)

A researcher is exploring whether there is a relationship between blood pressure and weight. For each person in the study, she records the person's weight and their blood pressure.

Bivariate data

Definition (Independent and dependent variables)

An independent variable is a variable that can be controlled to determine the value of a dependent variable.

Lots of words that mean the same thing:

Independent variable	**Dependent variable**
explanatory variable	outcome variable
predictor variable	response variable
controlled variable	measured variable
regressor	regressand
manipulated variable	observed variable
input variable	output variable
x	y

Independent and dependent variables

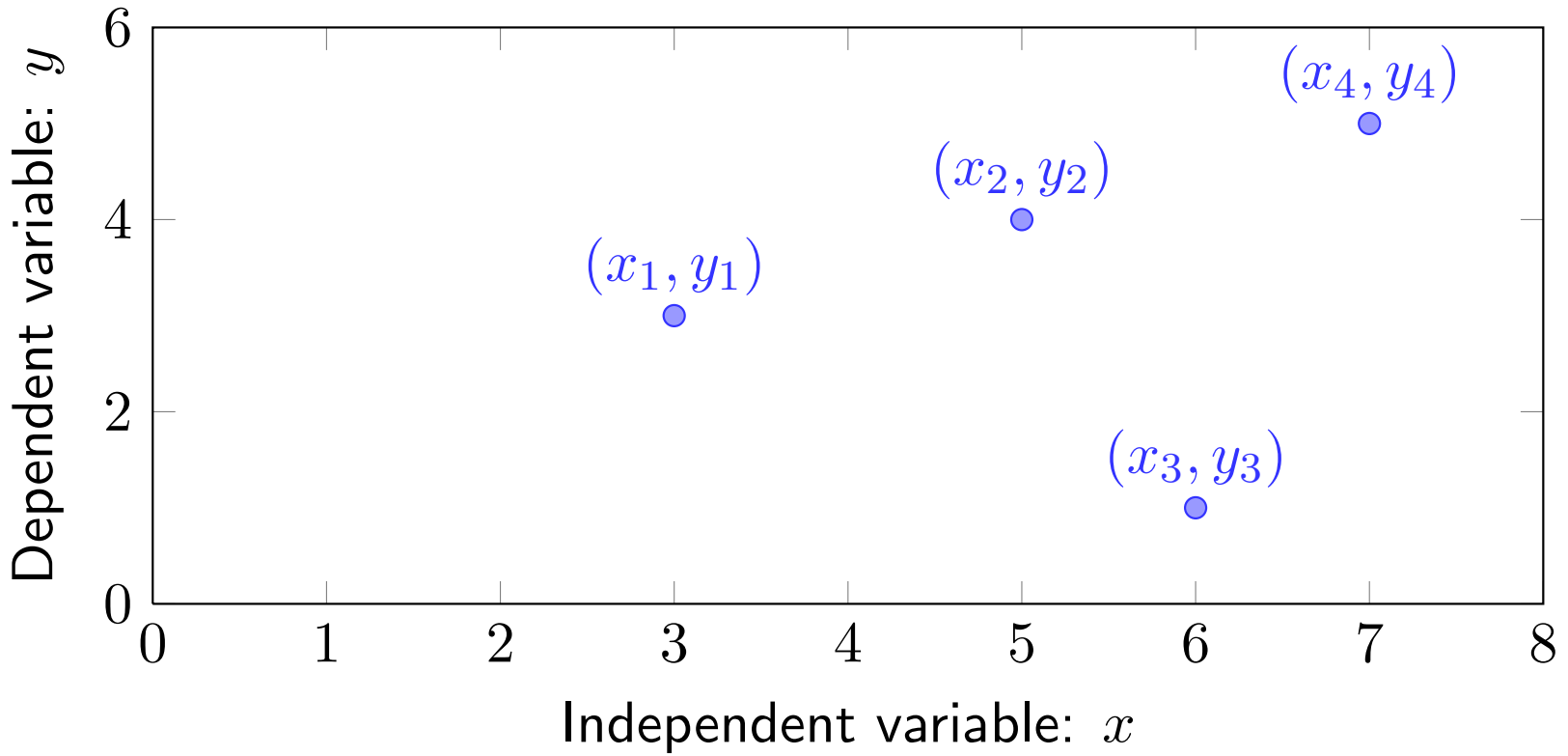
- Using the labels "independent" and "dependent" is only important when we want to estimate one measurement on the basis of the value of the other.
- The convention is to use the label y for the variable to be estimated (the dependent variable) and to use x for the independent variable.

Your turn: how would you label these variables?

1. drug concentration and blood pressure
2. height and weight
3. income and education

Relationship between x and y?

- The point of this course is to enable you to determine if there is a significant linear relationship between two variables.
- The first step in looking for some structure in the data, is to draw a scatterplot.



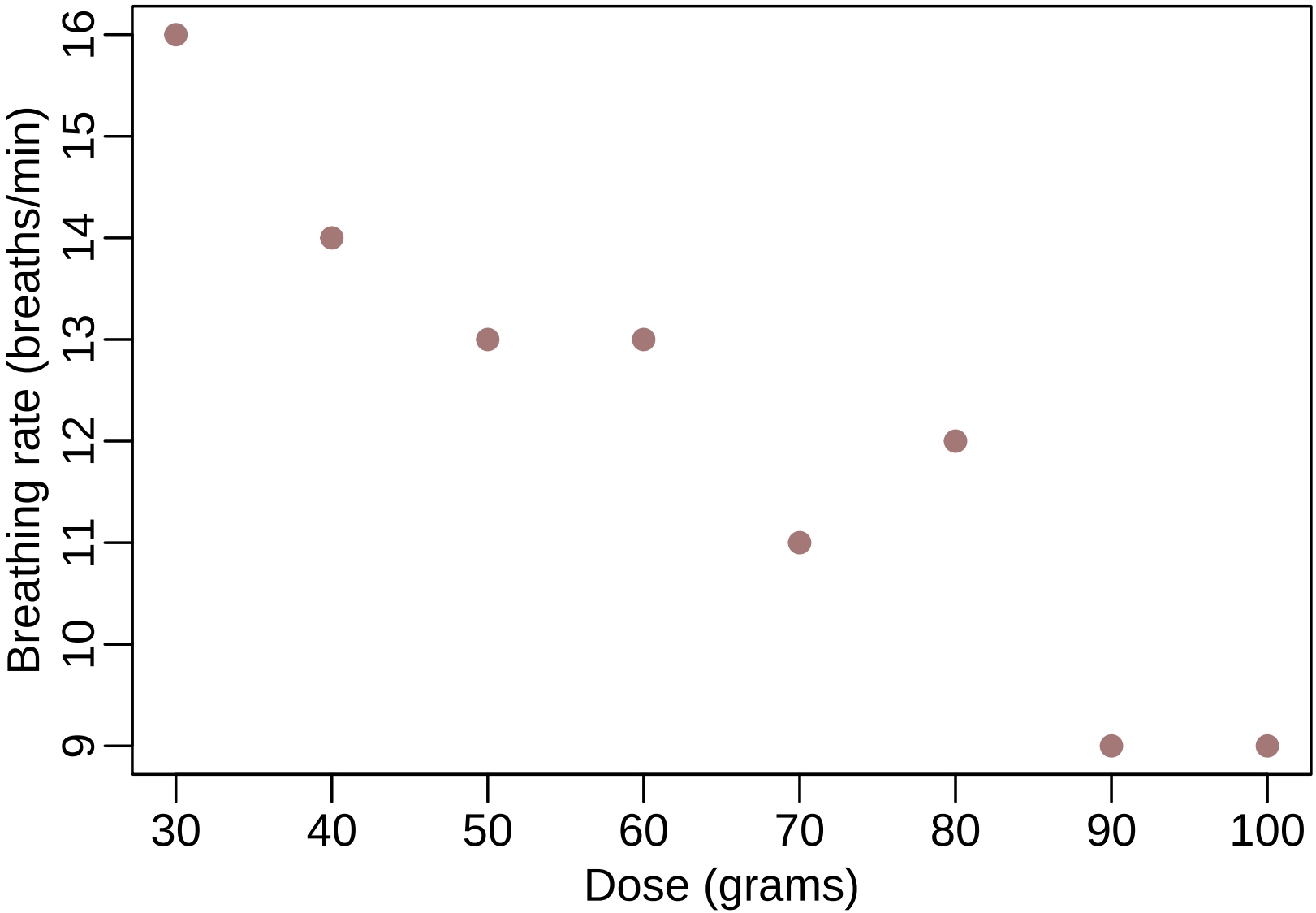
Haldane effect

- John Burdon Sanderson Haldane (1892–1964) was a physiologist famous for intrepid self-experimenting which led to many important discoveries about the human body and the nature of gases.
- One of Haldane's specialties was the physiology of gas absorption and binding in humans.

- To assess carbon dioxide regulation of blood pH he ingested large quantities of sodium bicarbonate ($NaHCO_3$) to make his blood basic.
- Then he measured his breathing rate (in breaths per minute).

Dose (grams):	30	40	50	60	70	80	90	100
Breathing Rate:	16	14	13	13	11	12	9	9

Haldane effect ($n = 8$)



Source: J.B.S. Haldane (circa 1920)

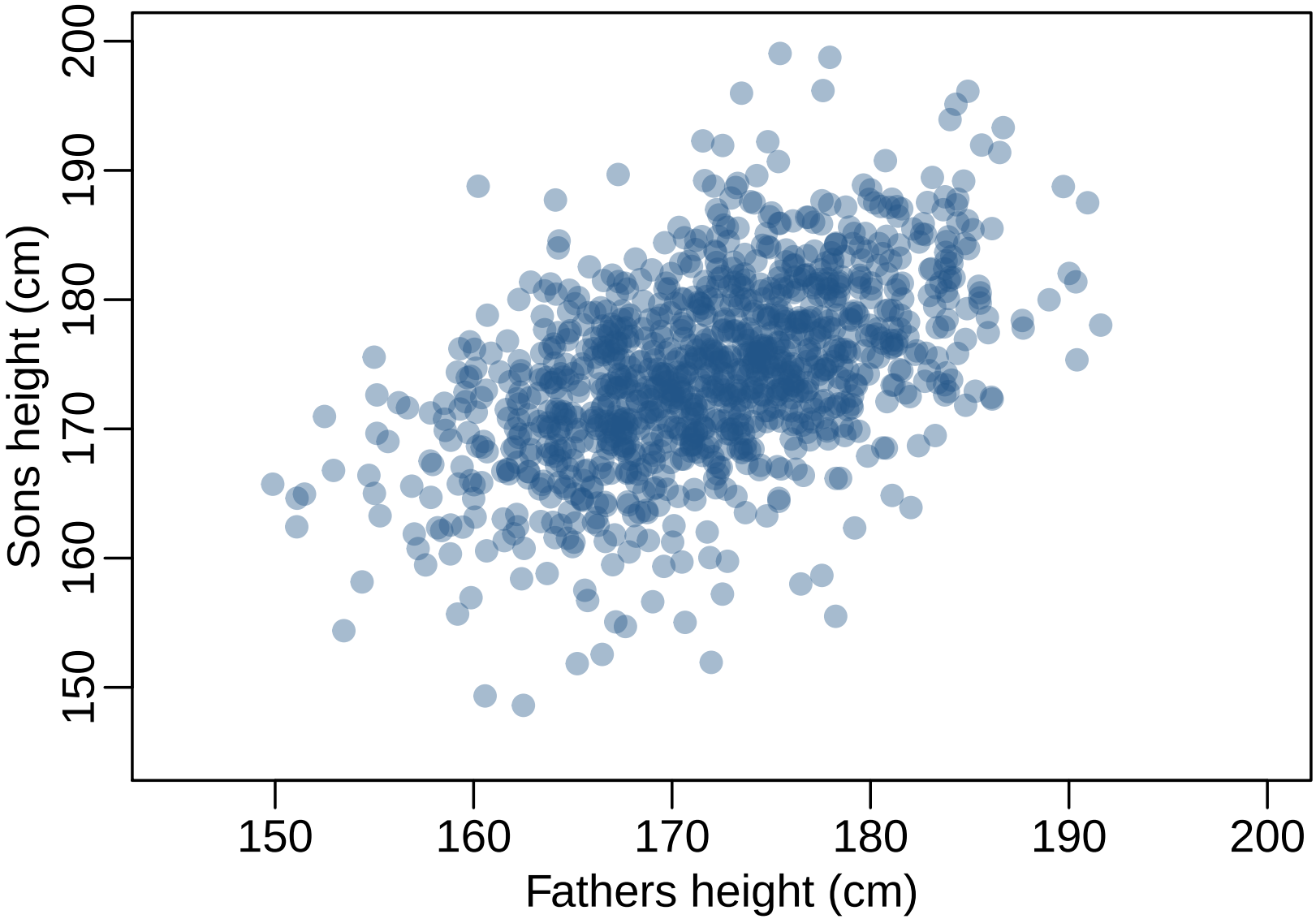
Father and son heights

- Statisticians in Victorian England were fascinated by the idea of quantifying hereditary influences.
- They gathered huge amounts of data in pursuit of this goal.
- Karl Pearson (1857–1936) studied 1078 pairs of fathers and their grown-up sons.
- He looked for a relationship between the heights of fathers and their sons.

Your turn...

Which is the independent variable and which is the dependent variable?

Father and son heights ($n = 1078$)



Source: Pearson, K. and Lee, A. (1903). *Biometrika*, 2(4):357–462.

Outline

Correlation coefficient

Definition (Correlation coefficient)

A measure of linear association for a bivariate data set is the correlation coefficient,

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

- $S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}.$
- $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2.$
- $S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2.$
- Here $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the sample means of x and y respectively.

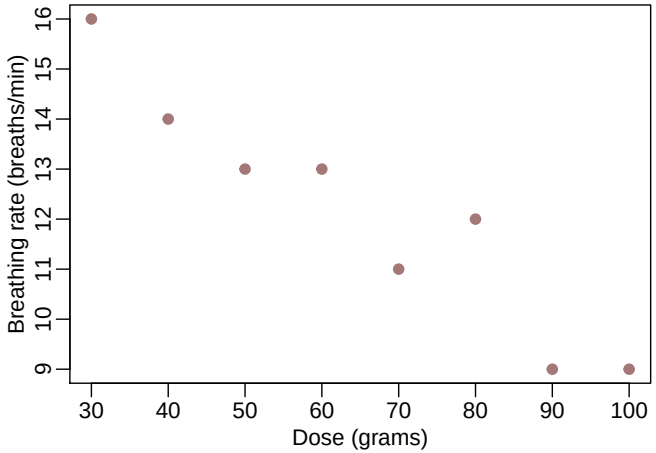
Correlation coefficient

- S_{xx} is proportional to the sample variance of x and similarly, S_{yy} is proportional to the sample variance of y:

$$\text{Estimated variance of } x = s_x^2 = \frac{1}{n-1} S_{xx}.$$

- S_{xx} and S_{yy} are both guaranteed to be positive. (You have made a mistake if they are not!)
- S_{xy} could be positive or negative.
- r is bounded: $-1 \leq r \leq 1$.
- r is not affected by a change of scale or origin. In other words, r is scale invariant and location invariant.
- r is symmetric in x and y: the correlation between x and y is the same as the correlation between y and x.

Correlation of the Haldane data



Need to know:

- $n = 8$
- $\sum_{i=1}^{8} x_i = 520$
- $\sum_{i=1}^{8} x_i^2 = 38000$
- $\sum_{i=1}^{8} y_i = 97$
- $\sum_{i=1}^{8} y_i^2 = 1217$
- $\sum_{i=1}^{8} x_i y_i = 5910$

Your turn...

- $\overline{x} = \frac{1}{8} \sum_{i=1}^{8} x_i =$
- $\overline{y} = \frac{1}{8} \sum_{i=1}^{8} y_i =$
- $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 =$
- $S_{yy} = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 =$
- $S_{xy} = \sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y} =$
- $r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} =$

Scale invariance with Haldane data

Consider the $n = 8$ observations from the Haldane data set, but now we measure dose in decagrams instead of grams:

Dose (in decagrams) x:	3	4	5	6	7	8	9	10
Breathing rate y:	16	14	13	13	11	12	9	9

- $\sum_{i=1}^{8} x_i = 52$
- $\sum_{i=1}^{8} x_i^2 = 380$
- $\sum_{i=1}^{8} x_i y_i = 591$
- $\sum_{i=1}^{8} y_i = 97$
- $\sum_{i=1}^{8} y_i^2 = 1217$
- $\overline{y} = 12.125$
- $S_{yy} = 40.875$

Your turn...

- $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i =$
- $S_{xy} = \sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y} =$
- $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 =$
- $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} =$

Calculating the correlation in Excel

Demonstration with the fathers' and sons' height data.

Interpretation of the correlation coefficient, r

- The sign of r reflects the trend of the points. It is positive if y increases with x and negative if y decreases as x increases.
- If $r = 1$, the points lie on a straight line of positive slope.
- If $r = -1$, the points lie on a straight line of negative slope.
- If $r = 0$, there is no linearity in the points even if there is some other relationship.
- It is important **not to interpret a high value of $|r|$ as a cause/effect relationship**.

Important!

A value of r close to ± 1 is not necessarily an indication of causality!

Spurious correlation

Example

As ice cream sales increase, the rate of drowning deaths increases sharply. Therefore,

(a) ice cream consumption causes drowning; or

(b) more ice cream is sold in summer months than during colder times, and it is during summer that people are more likely to go swimming and are therefore more likely to drown.

Example

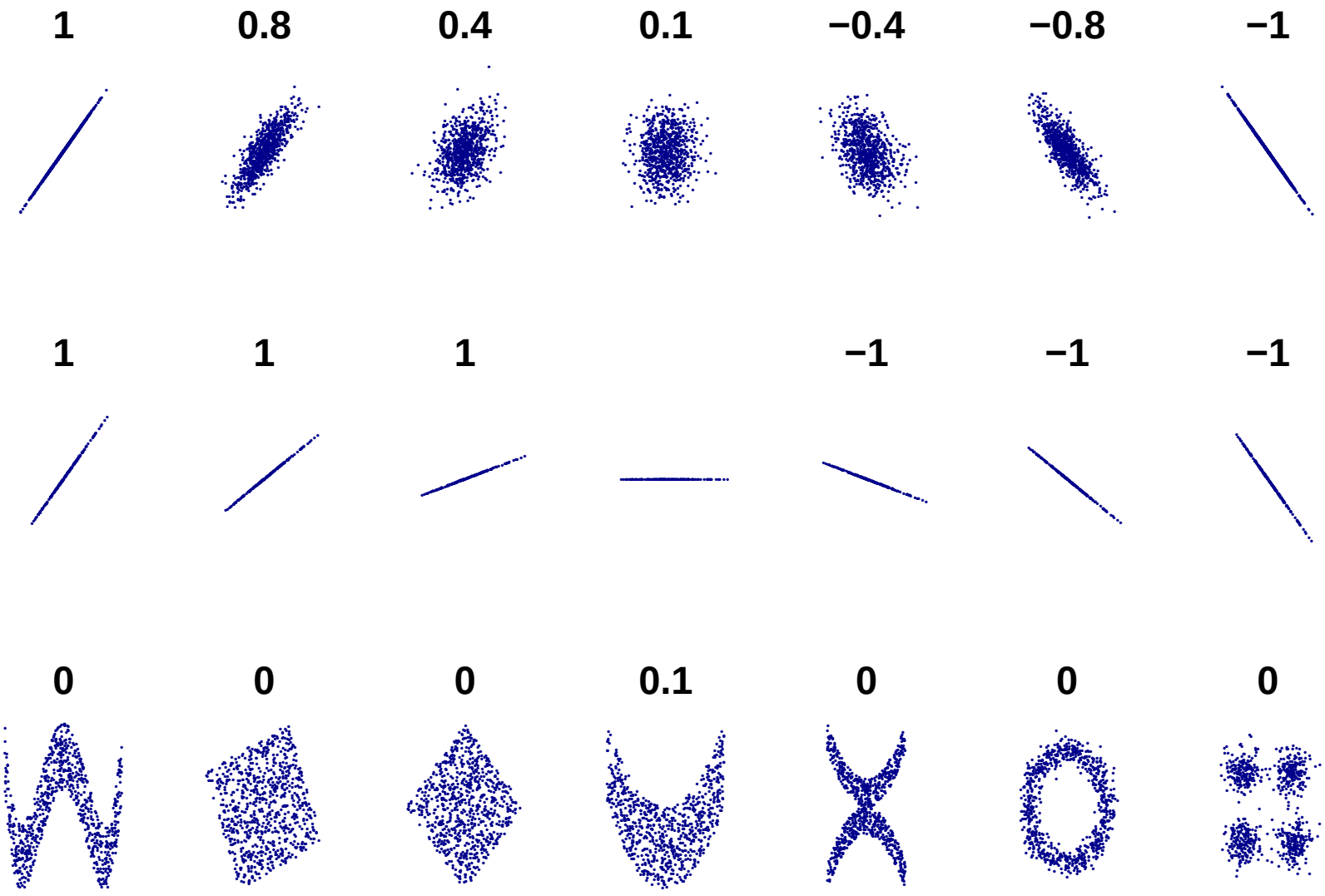
Young children who sleep with the light on are much more likely to develop myopia in later life. Therefore,

(a) sleeping with the light on causes myopia; or

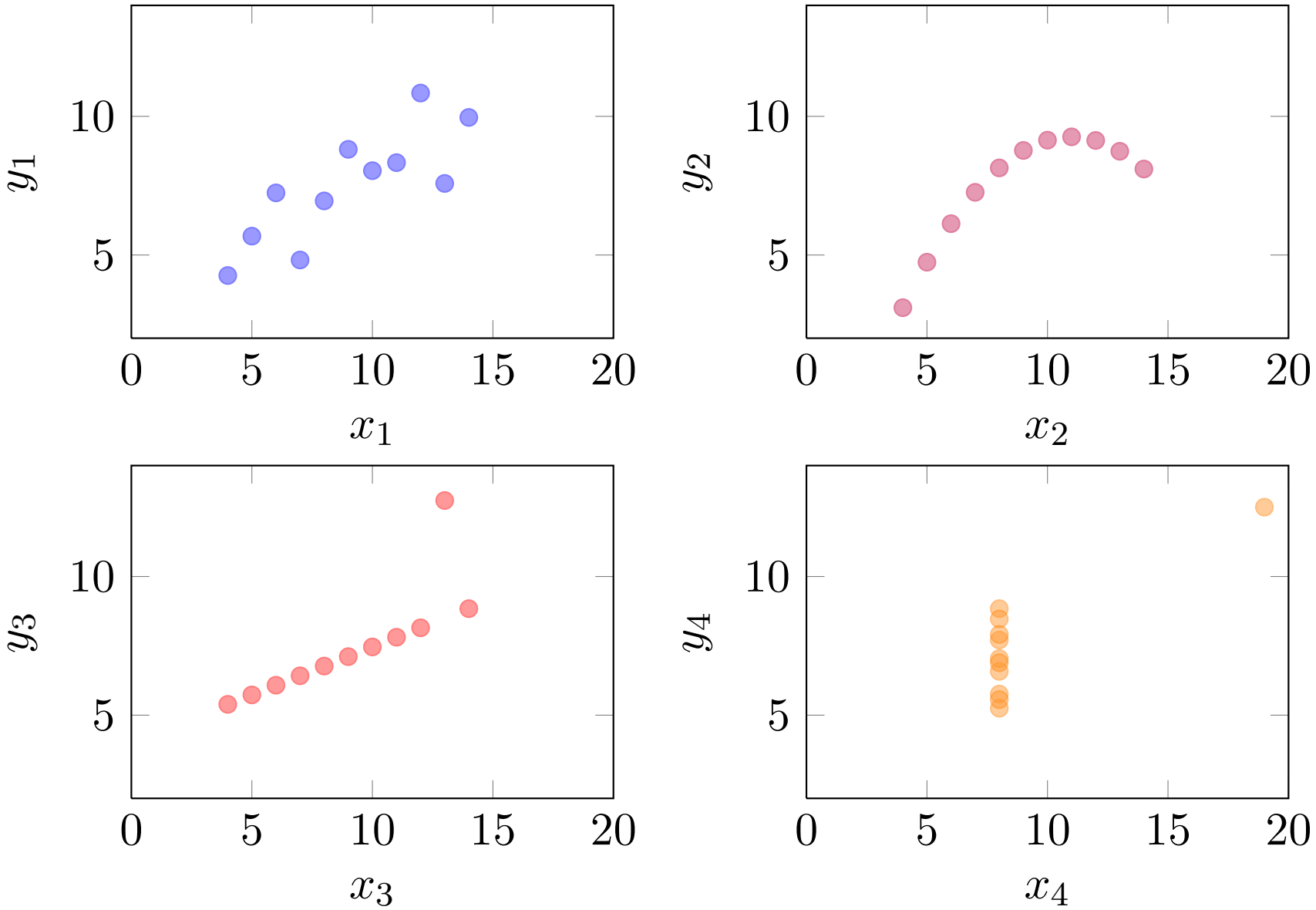
(b) myopia is a genetic trait and adults with myopia are more likely to leave the light on in their children's bedroom.

▸ More

Scatter plots and their correlation



Anscombe's quartet



Source: Francis Anscombe (1973)

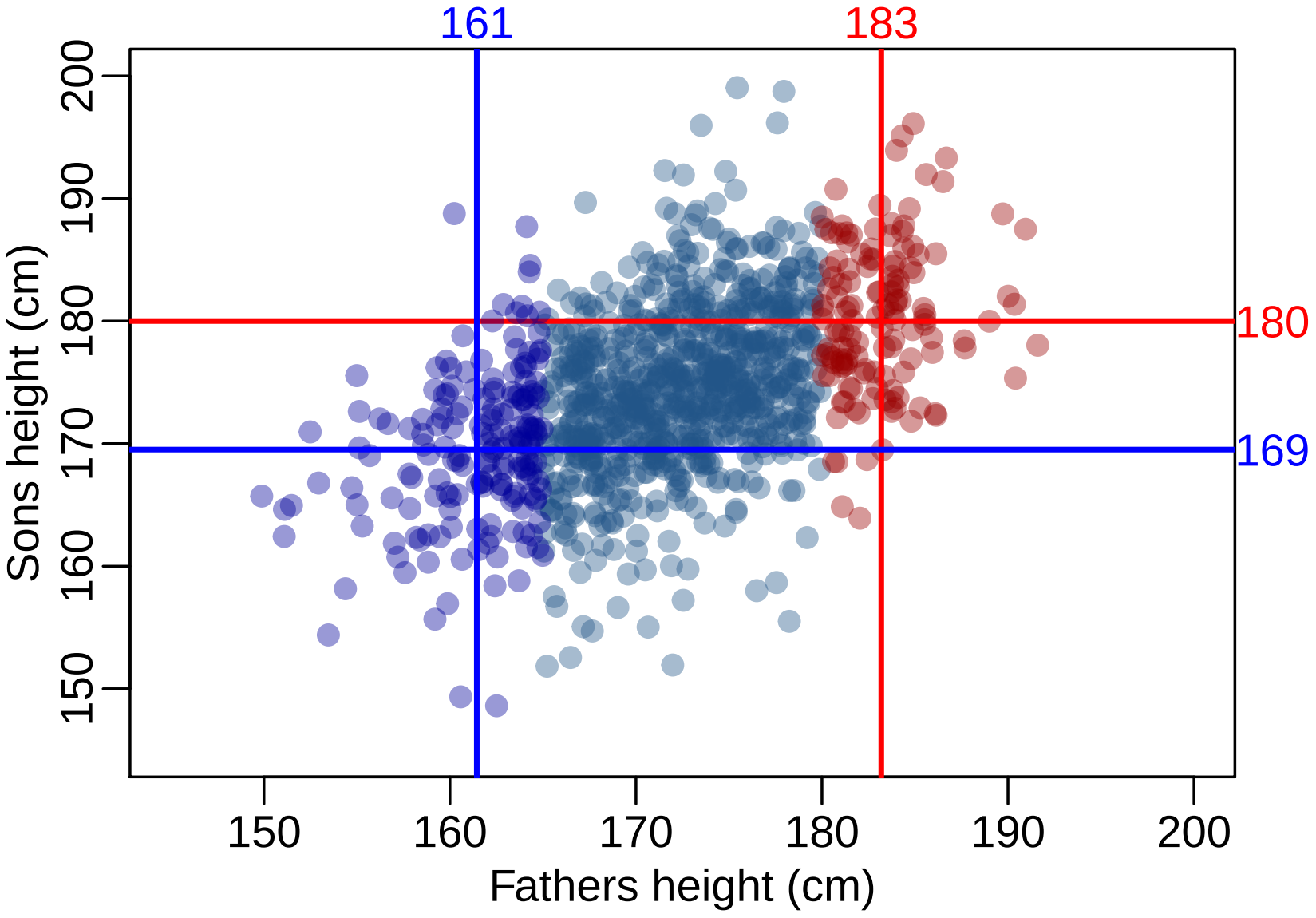
Outline

Regression to the mean

"Taller parents have shorter children, on average."

- Sir Francis Galton (1822–1911) was the first to note that tall parents have shorter children, on average.
- His protege and colleague Karl Pearson studied 1078 father-and-son pairs.
- He found that the tall fathers had sons that were one inch shorter, on average.
- On the other hand, on average, short fathers had sons that were three inches taller.
- Galton termed this phenomenon regression to mediocrity.
- Ever since, the method of studying how one variable relates to another variable has been called regression analysis.

Father and son heights



The aim of regression analysis

- **Aim:** to estimate the relationship between two variables,

$$y = f(x).$$

- In this course we will focus on simple linear regression. So, $f(x)$ is a linear function, $f(x) = \alpha + \beta x$, i.e.,

$$y = \alpha + \beta x.$$

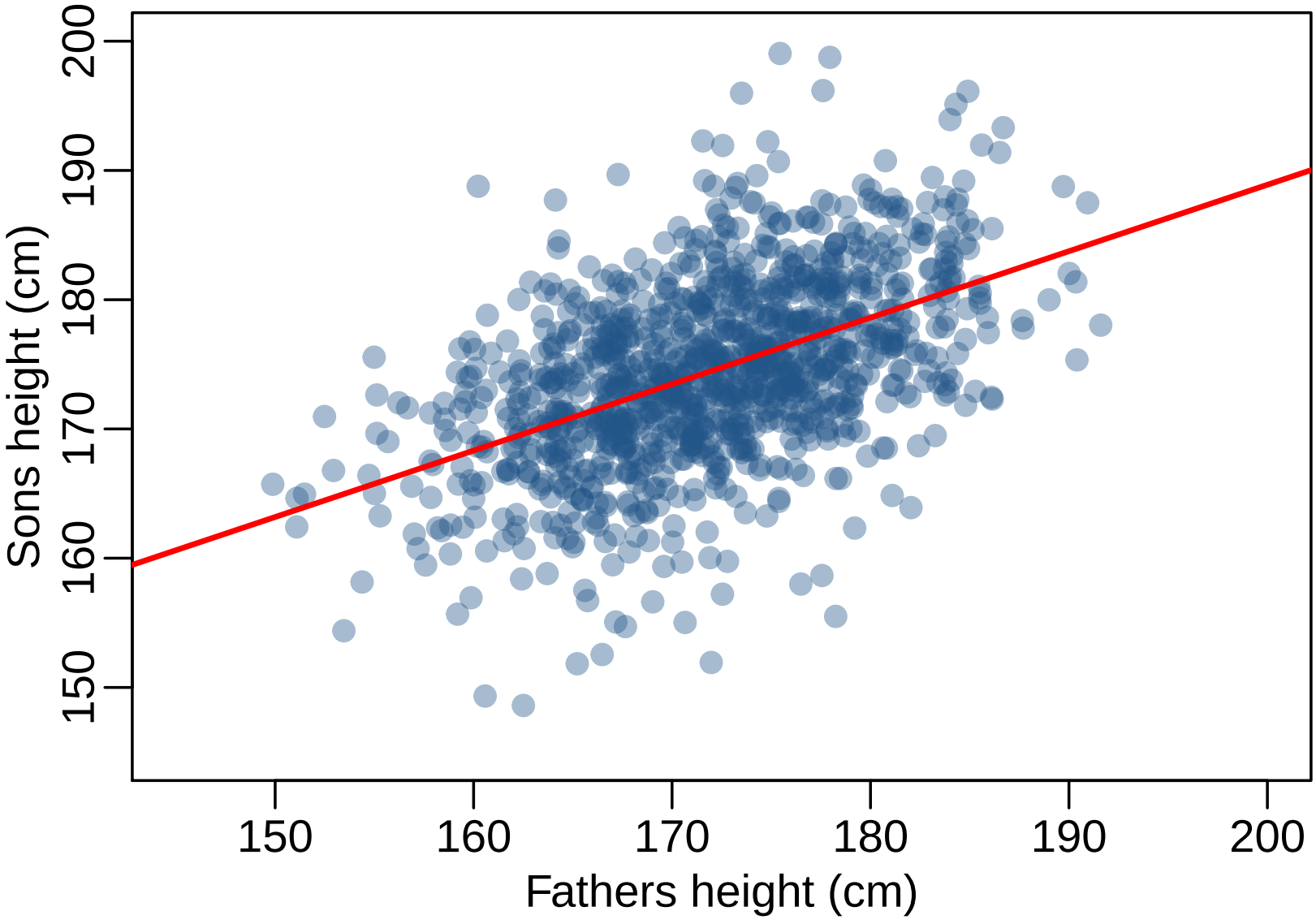
- But our observations do not lie on a perfectly straight line, there's an error component:

$$y = \alpha + \beta x + \varepsilon.$$

- The error, ε, is the difference between the observed y value and the value predicted by the line:

$$\varepsilon = y - (\alpha + \beta x).$$

The aim of regression analysis



Estimating the regression line

- The "true" relationship between y and x is given by the population regression function:

$$y = \alpha + \beta x + \varepsilon.$$

- Given a sample of data, we need to estimate the intercept α and the slope β, with the estimated regression function:

$$\widehat{y} = a + bx.$$

- I.e. a and b estimate α and β in the same way that the sample mean $\bar{x}$ is an estimate of the population mean, μ.

Problem: how to find a and b?

To answer this question, we need to consider the estimate of the error, ε, known as the residual, e.

Residuals

- Suppose our estimate of the line $y = \alpha + \beta x$ is $\widehat{y} = a + bx$.

Definition (Predicted or fitted value)

If we substitute $x = x_i$ into the estimated line, this gives us

$$\widehat{y}_i = a + bx_i$$

as our estimate of the y at some observed x_i. The estimate, $\widehat{y}_i$, is called the fitted value or predicted value of y at $x = x_i$.

- However, at x_i we actually observed y_i.

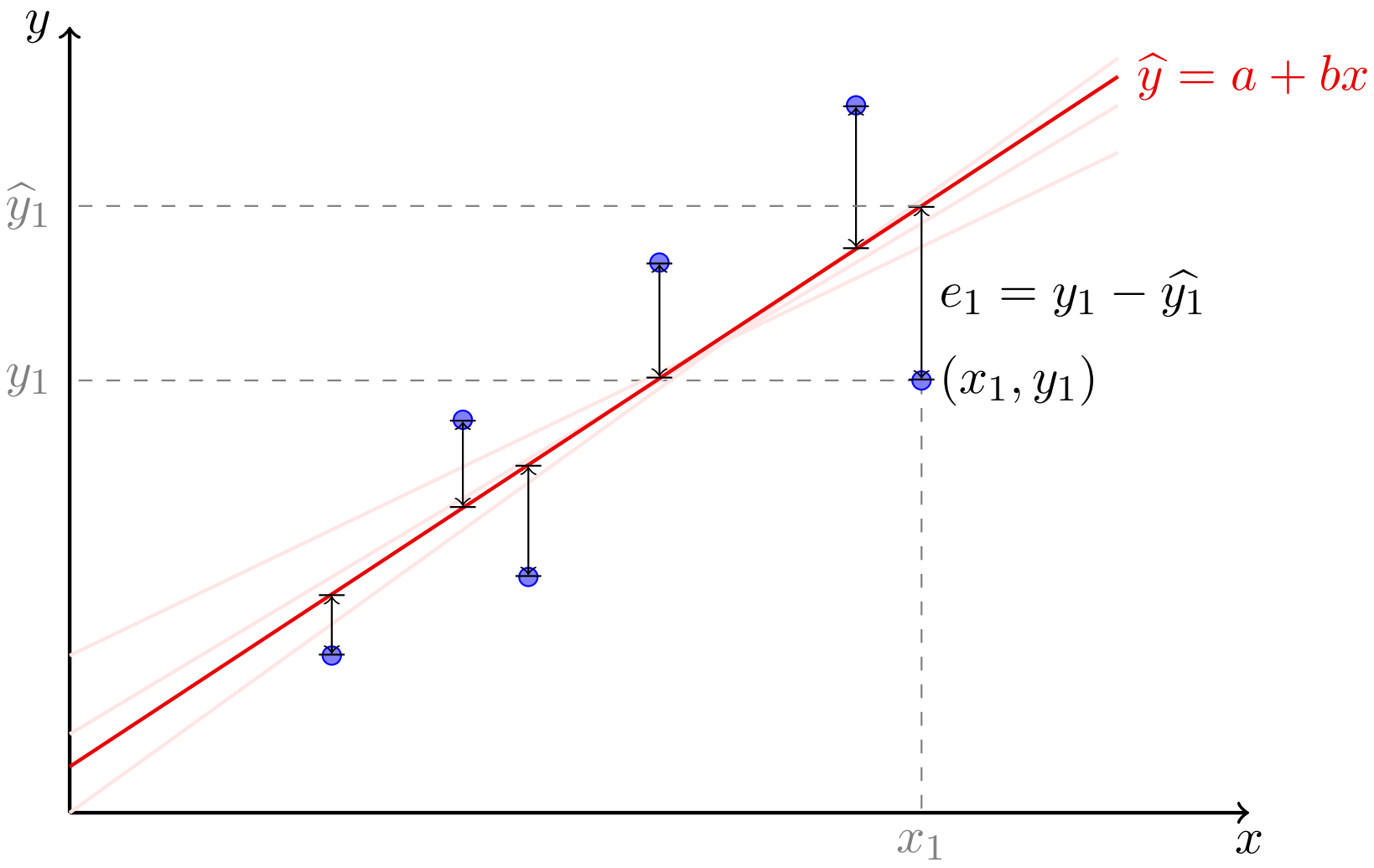
Definition (Residuals)

The residuals, e_i, are the difference between the observed values of y_i and the predicted values $\widehat{y}_i$:

$$e_i = y_i - \widehat{y}_i \quad \text{for } i = 1, 2, \ldots, n.$$

Which line is best?

We want our residuals to be small in absolute size, otherwise our line would have been bad at predicting the points we already have.



How to estimate the regression line: $\widehat{y} = a + bx$

The method of least squares

Find the values of a and b that minimise the sum of squared residuals (SSR):

$$\text{SSR} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 ,$$

subject to the constraint that the overall mean residual is zero:

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n} [y_i - (a + bx_i)] = 0.$$

Result:

$$a = \bar{y} - b\bar{x} \quad \text{and} \quad b = \frac{S_{xy}}{S_{xx}}$$

where $S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$, $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

The method of least squares: proof

From the constraint that the overall mean error is zero, we know

$$\frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} [y_i - (a + bx_i)] = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} [y_i - a - bx_i] = 0$$

$$\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) - a - b\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = 0$$

$$\overline{y} - a - b\overline{x} = 0$$

$$a = \overline{y} - b\overline{x}.$$

Noting that if a is a constant then $\frac{1}{n}\sum_{i=1}^{n} a = \frac{1}{n}na = a$.

The method of least squares: proof

Substituting $a = \overline{y} - b\overline{x}$ in the formula for the residuals gives,

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - (a + bx_i) \\ &= y_i - (\overline{y} - b\overline{x} + bx_i) \\ &= y_i - \overline{y} + b\overline{x} - bx_i \\ &= (y_i - \overline{y}) - b(x_i - \overline{x}). \end{aligned}$$

We can write the sum of squared residuals as,

$$\begin{aligned} \text{SSR} &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} \left[(y_i - \overline{y})^2 - 2b(y_i - \overline{y})(x_i - \overline{x}) + b^2(x_i - \overline{x})^2\right] \\ &= S_{yy} - 2bS_{xy} + b^2 S_{xx} \end{aligned}$$

The method of least squares: proof

We want to minimise the sum of squared residuals, so we take the derivative with respect to b:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}b}\text{SSR} &= \frac{\mathrm{d}}{\mathrm{d}b}\left[S_{yy} - 2bS_{xy} + b^2 S_{xx}\right] \\ &= -2S_{xy} + 2bS_{xx}.\end{aligned}$$

To find the value of b that produces a minimum SSR, we set the derivative equal to zero and solve for b:

$$\begin{aligned}-2S_{xy} + 2bS_{xx} &= 0 \\ b &= \frac{S_{xy}}{S_{xx}}.\end{aligned}$$

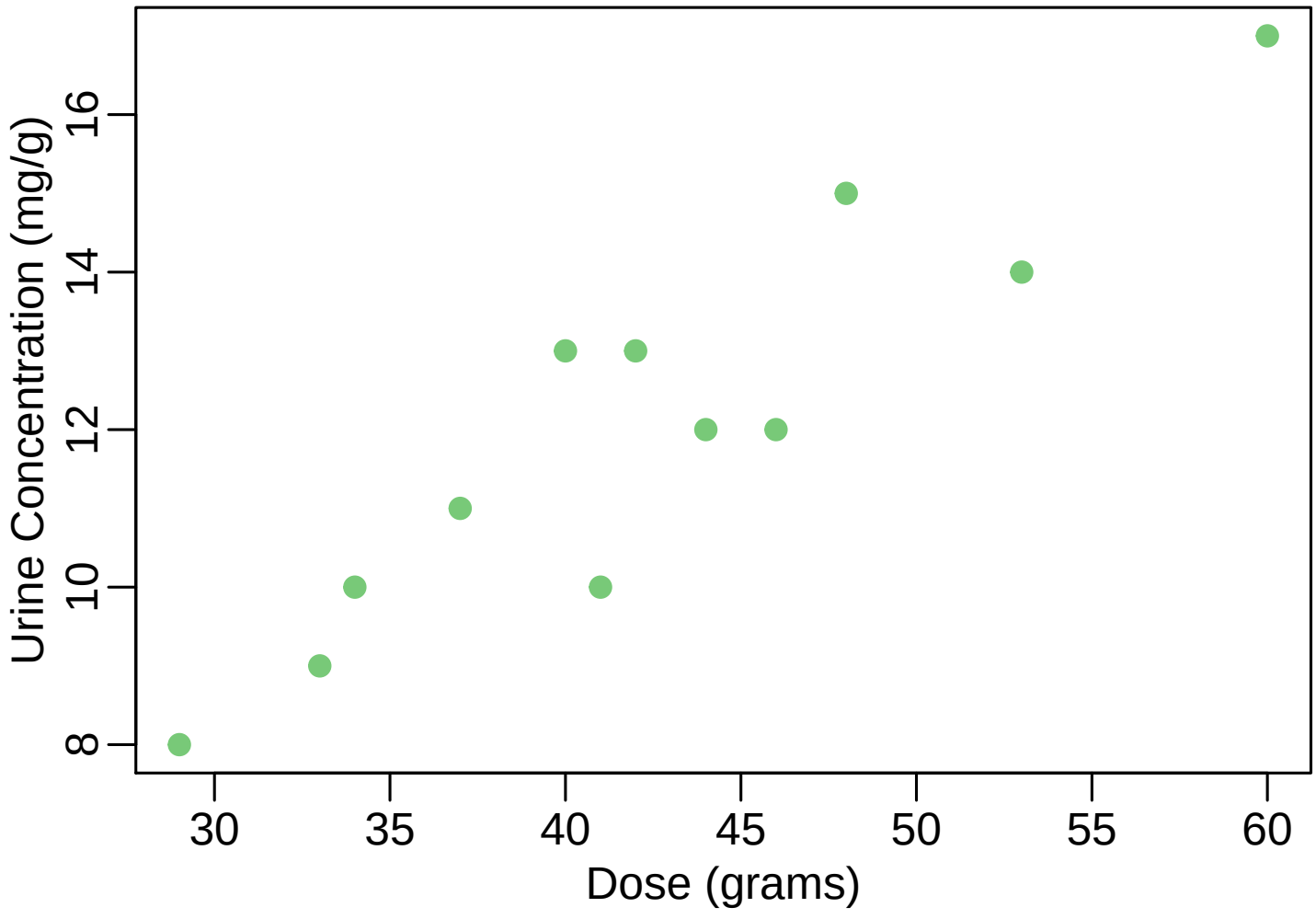
Result! $\widehat{y} = a + bx$

$$a = \bar{y} - b\bar{x} \quad \text{and} \quad b = \frac{S_{xy}}{S_{xx}}$$

Linear regression – example

In a study on the absorption of a drug, the dose x (in grams) and concentration in the urine y (in mg/g) were recorded as:

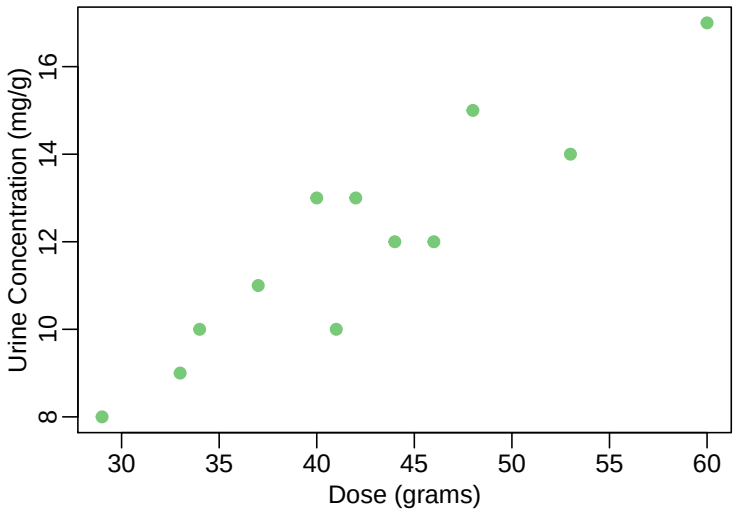
x	y
46	12
53	14
37	11
42	13
34	10
29	8
60	17
44	12
41	10
48	15
33	9
40	13



Linear regression – example

i	x	y	x^2	y^2	xy
1	46	12	$46^2 = 2116$	$12^2 = 144$	$46 \times 12 = 552$
2	53	14	2809	196	742
3	37	11	1369	121	407
4	42	13	1764	169	546
5	34	10	1156	100	340
6	29	8	841	64	232
7	60	17	3600	289	1020
8	44	12	1936	144	528
9	41	10	1681	100	410
10	48	15	2304	225	720
11	33	9	1089	81	297
12	40	13	1600	169	520
	507	144	22265	1802	6314
	$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum y_i^2$	$\sum x_i y_i$

Linear regression – example



- $n = 12$
- $\sum_{i=1}^{12} x_i^2 = 22265$
- $\sum_{i=1}^{12} y_i^2 = 1802$
- $\sum_{i=1}^{12} x_i y_i = 6314$
- $\bar{x} = 507/12 = 42.25$
- $\bar{y} = 144/12 = 12$

Your turn. . .

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$
$$=$$
$$=$$

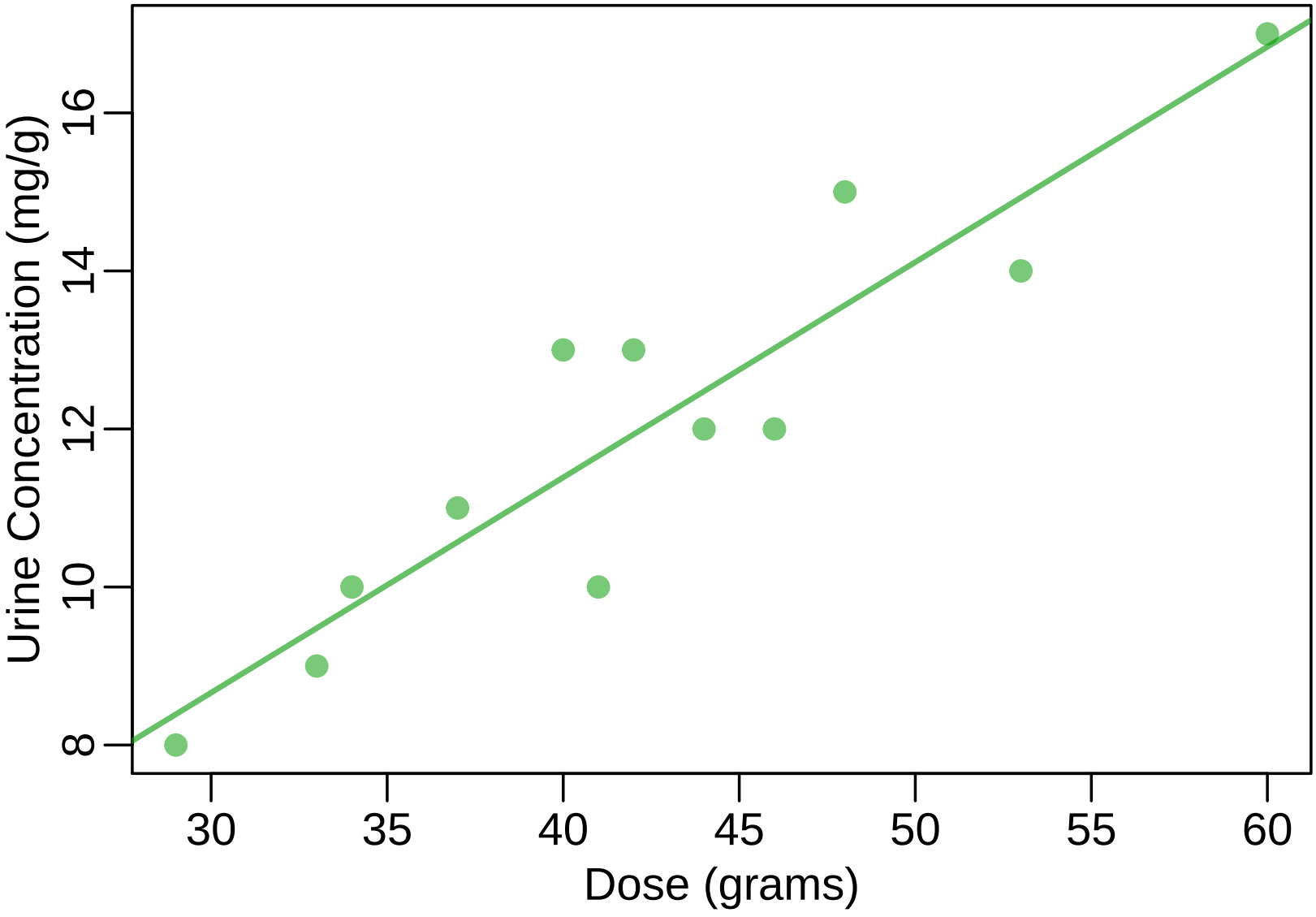
$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$
$$=$$
$$=$$

Hence,

$$b = \frac{S_{xy}}{S_{xx}} =$$

$$a = \overline{y} - b\overline{x}$$
$$=$$
$$=$$

Linear regression – example: $\widehat{y} = 0.49 + 0.27x$



Linear regression – example: Microsoft Excel

This is the relevant portion from Excel's output for this example.

Regression Statistics		
Multiple R	0.92	Absolute value of correlation coefficient, $\lvert r \rvert$
R Square	0.85	Square of correlation coefficient, r^2
Adjusted R2	0.83	Not discussed in this course
Standard Error	1.06	Estimated standard deviation of residuals, $\widehat{\sigma}$
Observations	12	Sample size, n

	Coefficients	Standard Error	t Stat	P-value
Intercept (a)	**0.49**	1.58	0.31	0.76
Dose (b)	**0.27**	0.04	7.43	0.00

Interpretation of a and b in $\widehat{y} = a + bx$

Slope coefficient, b

The average change in y for a one unit change in x:

On average, a one unit increase in x will cause a b unit change in y.

Example (Dose and urine concentration: $\widehat{y} = 0.49 + 0.27x$)

On average, a one gram increase in dose results in a 0.27 mg/g increase in urine concentration.

Intercept, a

The intercept is the value of y predicted by the model when $x = 0$. The intercept is often outside the range of observed values and sometimes makes no physical sense. Often it is not an important component to interpret.

Prediction

- How do we predict y at $x = x_i$ using the regression line?

  Answer: Find a and b, and substitute $x = x_i$ into $\widehat{y} = a + bx$, taking care to keep sufficient decimal places to estimate with the required precision.

Your turn...

Given our model,

$$\widehat{y} = 0.49 + 0.27x$$

predict the urine concentration of a person who has been given a dose of 55 grams:

$$\widehat{y} =$$

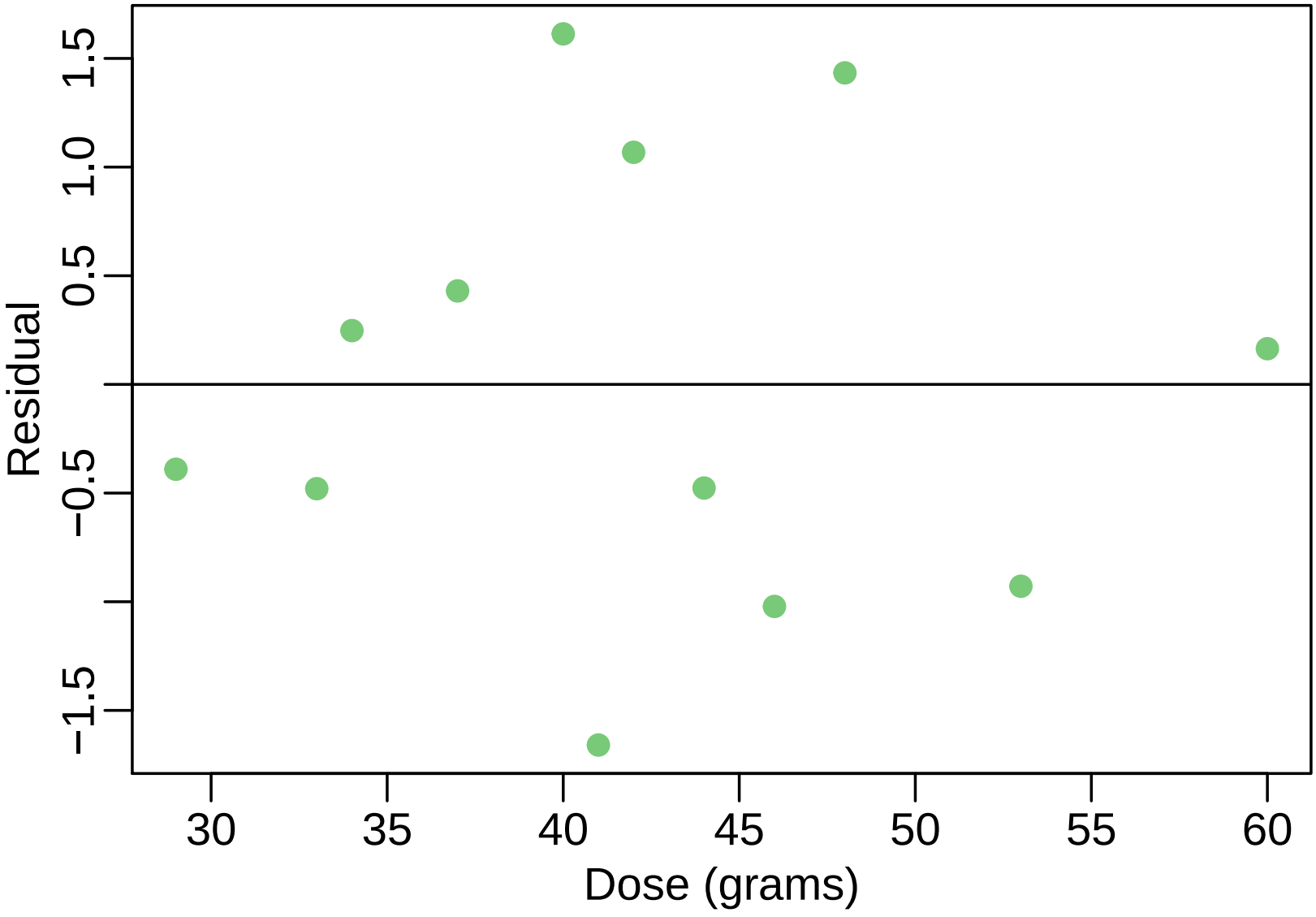
Residual plots

- The residual at a typical point (x_i, y_i) is given by $e_i = y_i - \widehat{y}_i$ where $\widehat{y} = 0.49 + 0.27x$.

i	x	y	$\widehat{y}$	e
1	46	12	$0.49+0.27\times46 = 12.91$	$12-12.91 =$ -0.91
2	53	14	14.80	-0.80
3	37	11	10.48	0.52
4	42	13	11.83	1.17
5	34	10	9.67	0.33
6	29	8	8.32	-0.32
7	60	17	16.69	0.31
8	44	12	12.37	-0.37
9	41	10	11.56	-1.56
10	48	15	13.45	1.55
11	33	9	9.40	-0.40
12	40	13	11.29	1.71

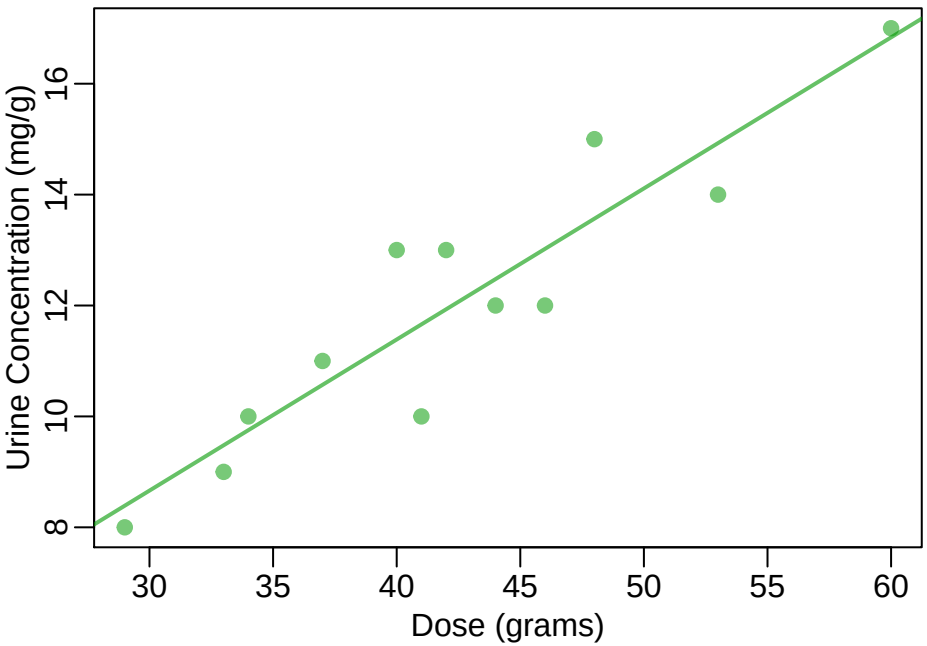
- The residual plot is a plot of all of the e_is vs all of the x_is.

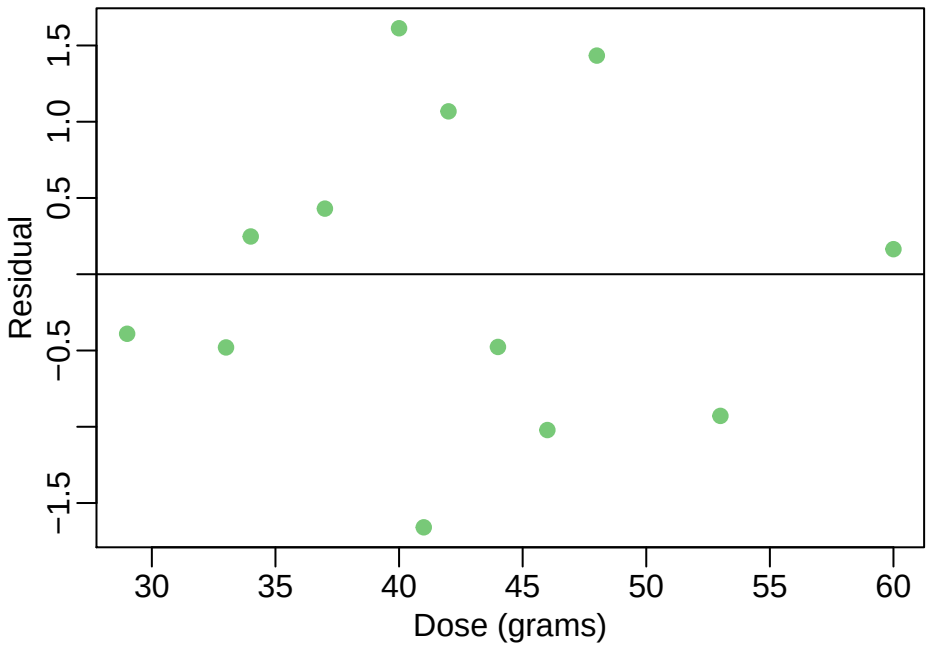
Residual plots



Residual plots – interpretation

- The residual plot indicates that there is no pattern in the residuals, just random scatter about the horizontal line through zero.
- Together with the rough linear scatter in the scatterplot, this tells us that the LSR line is a reasonable model.





Least squares assumptions

Recall that the "true" or population model for the data is:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Assumptions:

1. **Linearity** – we assume that the model is linear.
2. **Independence** – all the observations are obtained independently of one another.
3. **Homoskedasticity** – the errors have constant variance: $\text{var}(\varepsilon_i) = \sigma^2$ (an unknown constant) for all $i = 1, 2, \ldots, n$.
4. **Normality** – the errors are normally distributed: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ for all $i = 1, 2, \ldots, n$.

The last three can be written succinctly as:

$$\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

This reads: the errors are independently and identically distributed as a normal random variable with mean 0 and variance σ^2.

Assumption 1: Linearity

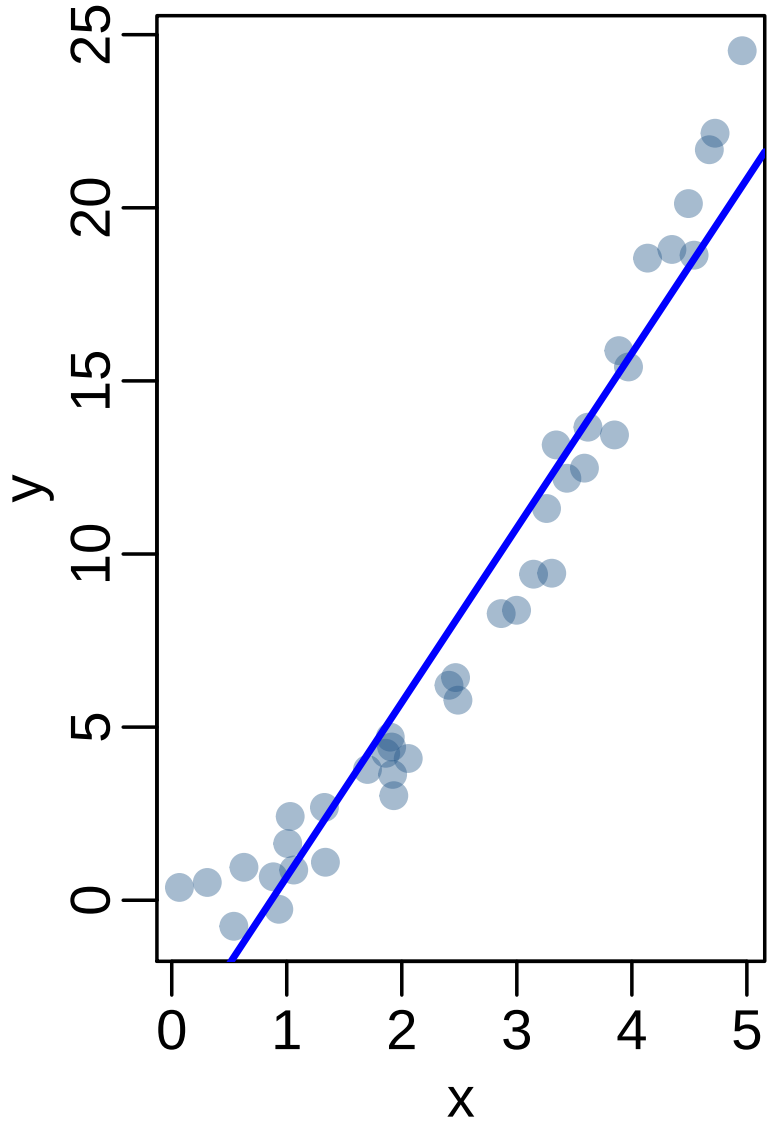
- Often the linearity of the relationship between y and x can be assessed after you've gathered the data but before you run the regression.
- Plot y against x and look to see if the relationship is approximately linear.
- Violations to linearity are quite serious, it means your predictions are likely to be wrong, particularly when extrapolating outside the range of observed values.
- After you have run the regression you can check for linearity using a residual plot: plot e_i against x_i.
- The points should be symmetrically distributed around a horizontal line.
- Look carefully for evidence of a "curved" pattern which indicates that in parts the model regularly overestimates y and in parts the model regularly underestimates y.

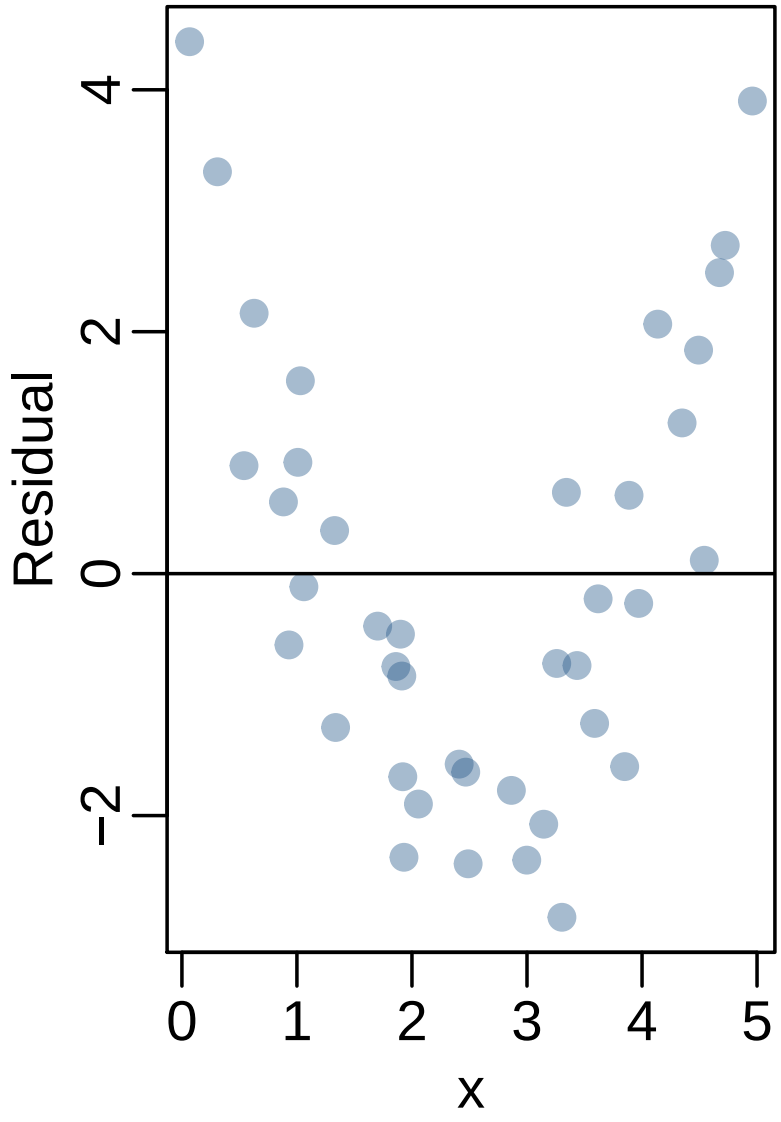
Assumption 1: Linearity

To understand what a pattern in the residual plot tells us, try this:

1. Draw a rough scatterplot of points which follow a slight convex quadratic pattern.
2. Draw the best straight line you can through the points.
3. Notice that the points are initially below the line (negative residuals, e_i) then there is a run of points above the line (positive residuals, e_i) and then points below again.
4. A plot of e_i vs x_i will show an obvious quadratic pattern.

Assumption 1: Linearity





Assumption 2: Independence

- The assumption of independence between the errors is usually dealt with in the experimental design phase – before data collection.
- You aim to design the experiment so that the observations are not related to one another.
- For example, you randomly sample your participants rather than just using members of your own family.
- If you don't have a random sample, your estimates a and b may be biased.
- Violations of independence usually arise in time series data where observations are measured on the same subject through time and therefore may be related to one another. This is beyond the scope of PHAR2821.

The other two assumptions are more important for inference and will be discussed in the third lecture.

Measuring model performance

Definition (Coefficient of determination)

The value r^2, the square of the correlation, can be interpreted as the proportion of the variation in the values of y that are explained by a linear fit of the data. It is a measure of goodness of fit. The closer r^2 is to 1, the better the fit.

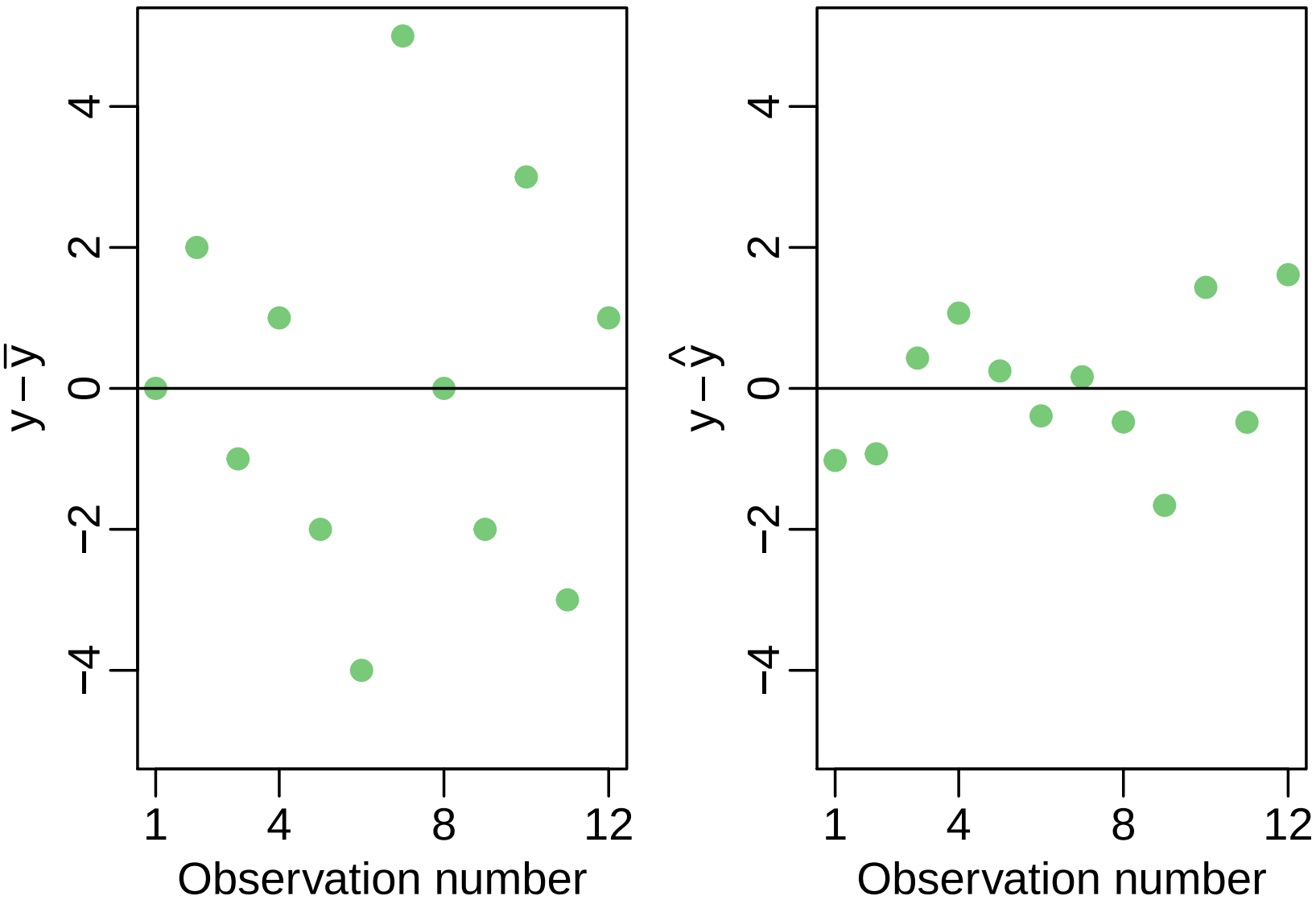
Example

In the example of urine concentration against dose, the correlation coefficient is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.92$$

so the coefficient of determination is $r^2 = 0.92^2 = 0.85$. Hence, 85% of the variability in urine concentration can be explained by dose.

Reduction in variation for urine concentration example

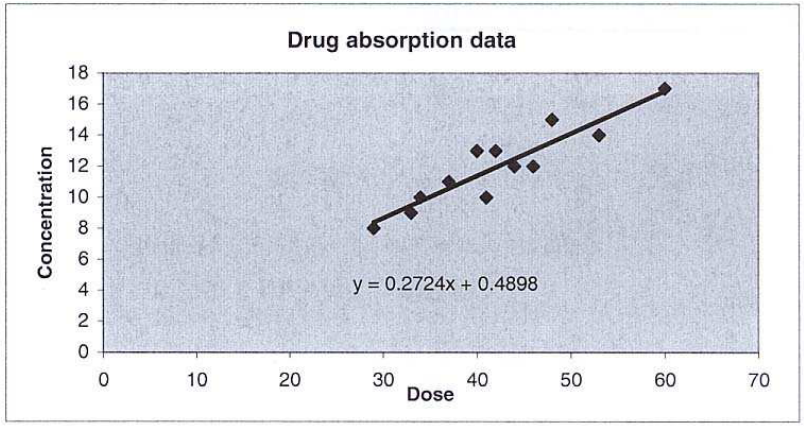


Excel demonstration

Excel output for regression

46	12
53	14
37	11
42	13
34	10
29	8
60	17
44	12
41	10
48	15
33	9
40	13

SUMMARY OUTPUT

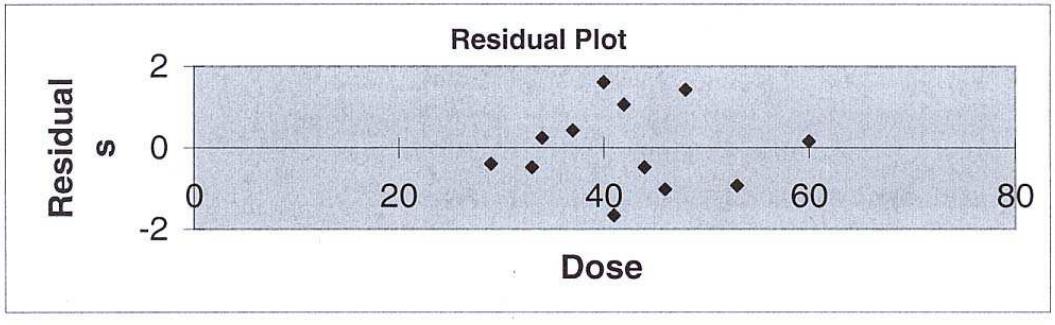


Regression Statistics	
Multiple R	0.920188
R Square	0.846745
Adjusted R	0.83142
Standard E	1.064934
Observatio	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	1	62.65916	62.65916	55.25092694	2.2281E-05
Residual	10	11.34084	1.134084		
Total	11	74			P-value = 0

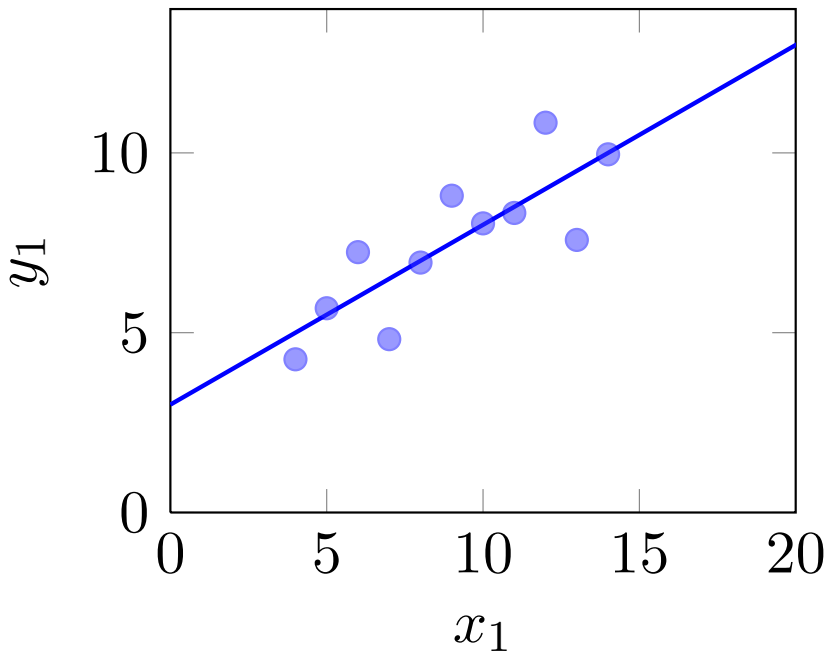
	Coefficients	*standard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.489784	1.578729	0.310239	0.76274823	-3.027845	4.007413	-3.02784519	4.007412854
X Variable	0.272431	0.036651	7.433097	2.2281E-05	0.190767	0.354095	0.190767404	0.3540949

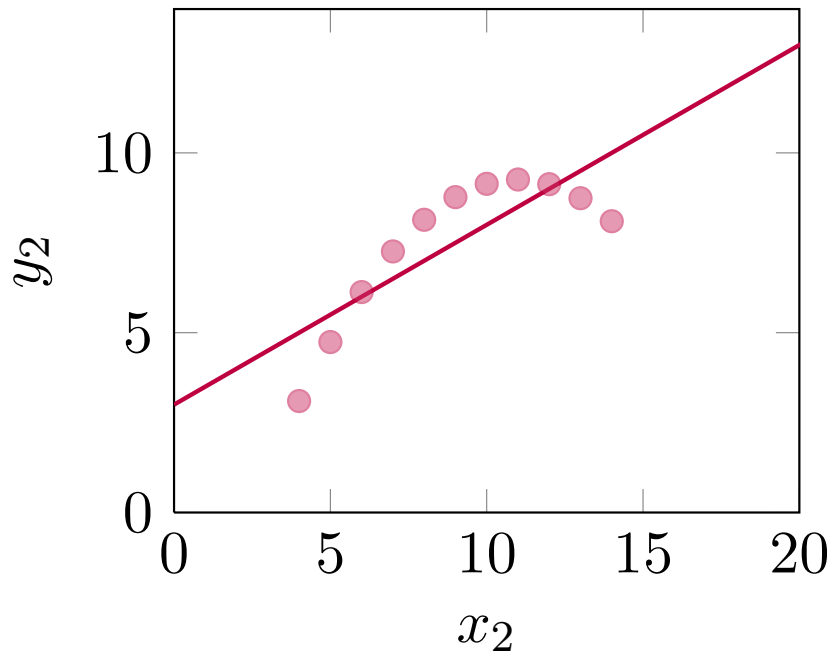


Caution with regression analysis

Before using bivariate data (x_i, y_i), $i = 1, \ldots, n$, to estimate y from a value of x, the first thing you must do is **draw a scatterplot** and ask yourself the following questions.

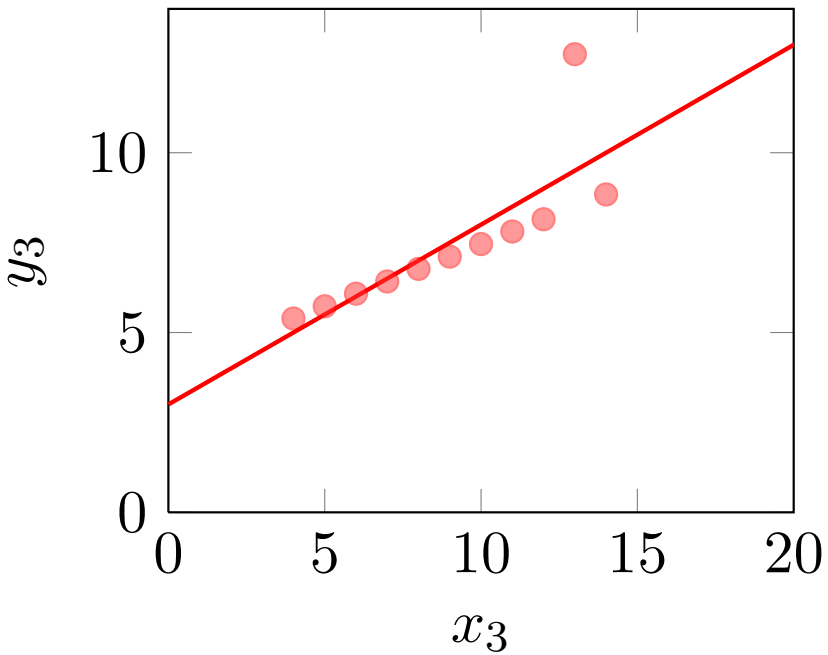
1. Are the data scattered about a rough **linear trend**?
   If not, there is no point in using the regression line, $y = \alpha + \beta x$, for estimation.

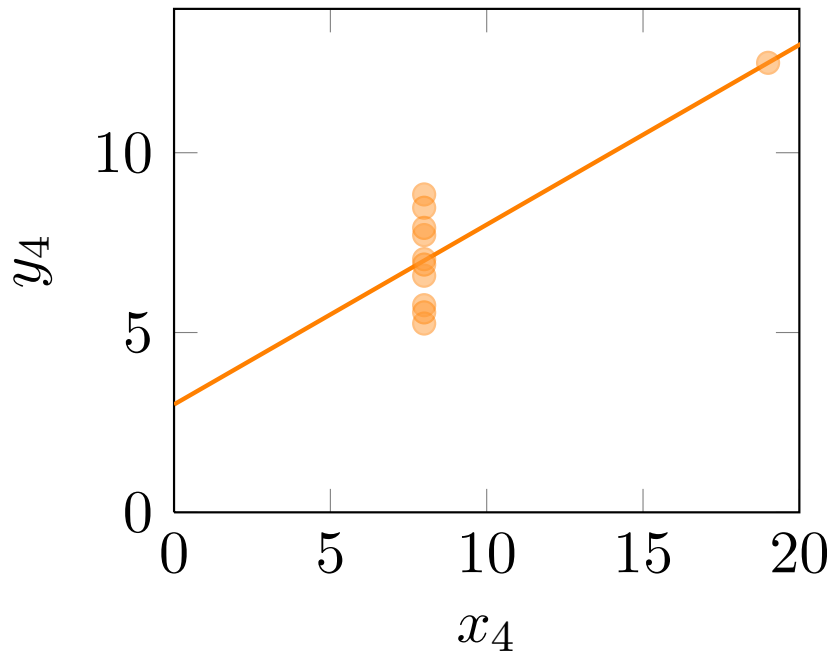




Caution with regression analysis

2. Look for **outliers**. They may simply be data entry errors, but they may have the effect of producing an artificially high value for r. They will also distort any estimated regression line.





Caution!

These are the same sorts of considerations that you need to be aware of before calculating the correlation coefficient.

Link between correlation and regression

- The correlation coefficient measures the strength of the linear association between x and y and is defined as

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

- The regression slope coefficient measures the change in y for a unit increase in x and is defined as

$$b = \frac{S_{xy}}{S_{xx}}.$$

- As S_{xx} and S_{yy} are always positive, r and b have the same sign as S_{xy}.
- If r is positive (negative), then b is also positive (negative).

Outline

Regression vs experts

Yale law professor Ian Ayres on regression:

> *William Grove, completed a meta-analysis of 136 human versus machine studies. In only 8 out of 136 studies was expert opinion found to be appreciably more accurate than statistical prediction. . . Indeed, regression equations are so much better than humans... that even very crude regressions with just a few variables have been found to outpredict humans.*

Cognitive psychologists Richard Nisbett and Lee Ross on regression:

> *Human judges are not merely worse than optimal regression equations; they are worse than almost any regression equation.*

But how do we know if what we've found is statistically significant?

Statistical inference on regression parameters

- Due to the fact that a and b are estimates of α and β we might want to show how confident we are about these estimates. To do this, we can construct confidence intervals.
- Another important decision we might wish to make is to provide statistical evidence for whether the slope coefficient is significantly different to zero. I.e. is there a relationship between y and x. This is equivalent to testing the null hypothesis, $H_0 : \beta = 0$.
- Recall the population regression function,

$$y = \alpha + \beta x + \varepsilon.$$

If $\beta = 0$ this implies that,

$$y = \alpha + \varepsilon.$$

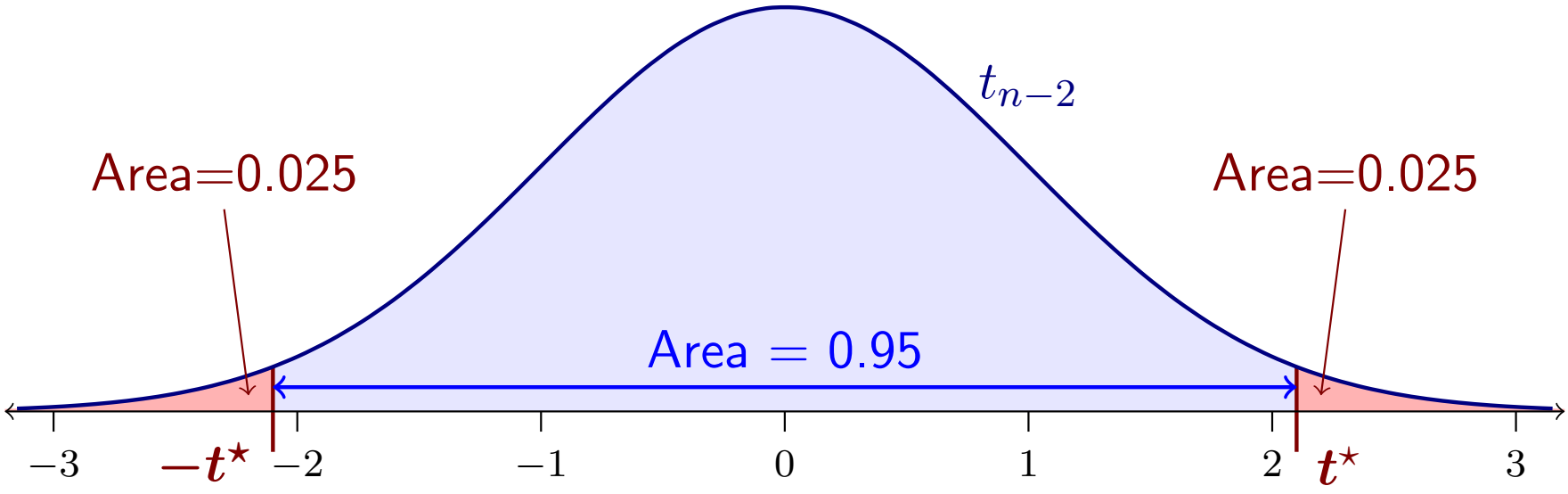
I.e. y does not depend on x: changing the value of x does not affect the value of y.

Confidence intervals for regression parameters

Approximate 95% confidence intervals for α and β are

$$a \pm t^\star \times \text{SE}(a) \quad \text{and} \quad b \pm t^\star \times \text{SE}(b)$$

where $t^\star$ is the appropriate quantile from a t distribution with $n-2$ degrees of freedom: $P(|t_{n-2}| > t^\star) = 0.05$. In Excel this can be found using `=TINV(0.05,n-2)` where `n` is the sample size.



Standard errors

- The standard error of a parameter is the standard deviation of a parameter – it measures the uncertainty in the estimate.
- The standard error of a is given by,

$$\text{SE}(a) = \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}.$$

- The standard error of b is given by,

$$\text{SE}(b) = \frac{\widehat{\sigma}}{\sqrt{S_{xx}}}.$$

- These can be found in the Excel output:

	Coefficients	Standard Error	...
Intercept	a	SE(a)	...
X variable	b	SE(b)	...

Estimating the error variance

An estimate of the variance of the residuals:

$$\widehat{\sigma}^2 = \frac{\text{RSS}}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2}.$$

The standard deviation of the residuals is known as the standard error of the regression. It can be found in the regression output in Excel:

Regression Statistics		
Multiple R	$\lvert r \rvert$	Absolute value of correlation coefficient
R Square	r^2	Square of correlation coefficient
Standard Error	$\widehat{\sigma}$	Estimated standard deviation of residuals
Observations	n	Sample size

Confidence intervals in Excel

Confidence intervals are given explicitly in the standard Excel regression output:

	Coeff.	Std. Error	...	Lower 95%	Upper 95%
Intercept	a	SE(a)	...	$a - t^\star \times$SE(a)	$a + t^\star \times$SE(a)
X variable	b	SE(b)	...	$b - t^\star \times$SE(b)	$b + t^\star \times$SE(b)

Recall that $t^\star$ can be found in Excel using `=TINV(0.05,n-2)` where `n` is the sample size.

Interpreting confidence intervals

- At the most basic level a confidence interval is:

  A range of plausible values for the parameter.

- Technically, we interpret confidence intervals as:

  If we run an experiment a large number of times, and each time we construct a 95% confidence interval for our parameter then, on average, we can expect 95% of those confidence intervals to contain the true population parameter.

- That is, we should really think about confidence intervals in the context of a process: a series of experiments.
- It is not correct to say that we are 95% sure that a particular confidence interval contains the true parameter.

Confidence intervals for regression parameters

- In general, if zero is included in the 95% confidence interval for β, then this provides evidence against x having a linear effect on y.
- For example, if a 95% confidence interval for β was,

$$(-0.65, 0.90),$$

  we would argue that x does not have a linear effect on y.
- This does not exclude the possibility that x could have a nonlinear effect on y.
- In general, if zero is not included in the 95% confidence interval for β, then this provides evidence for x having a linear effect on y.
- For example, if the 95% CI for β was,

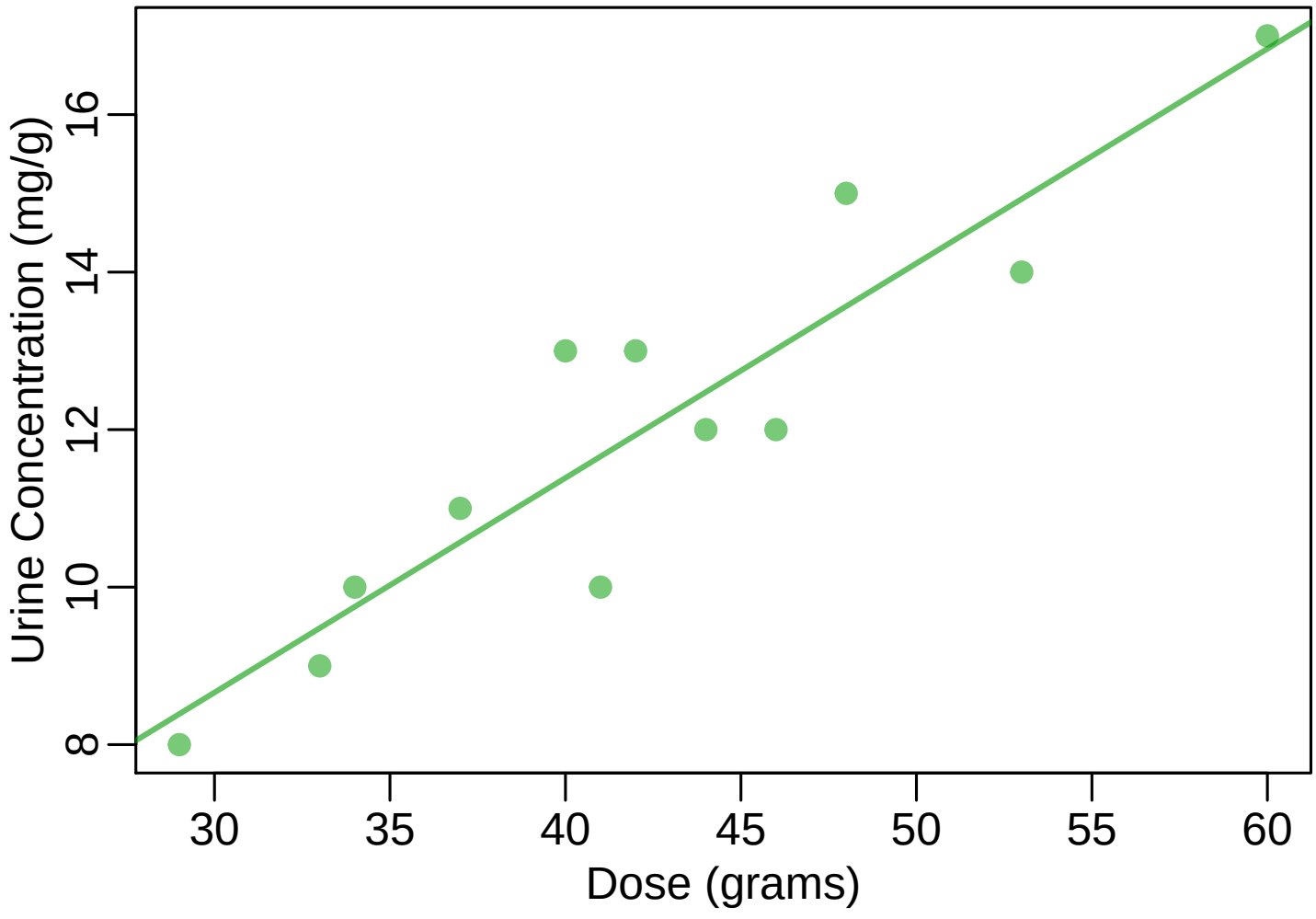
$$(0.65, 0.90),$$

  then we would argue that x does have a linear effect on y.

Confidence intervals for regression parameters – example

Recall the data from a study on the absorption of a drug, the dose x (in grams) and concentration in the urine y (in mg/g):

x	y
46	12
53	14
37	11
42	13
34	10
29	8
60	17
44	12
41	10
48	15
33	9
40	13



Confidence intervals for regression parameters – example

- $n = 12$
- $\sum_{i=1}^{12} x_i = 507$
- $\sum_{i=1}^{12} y_i = 144$
- $\sum_{i=1}^{12} x_i^2 = 22265$
- $\sum_{i=1}^{12} y_i^2 = 1802$
- $\sum_{i=1}^{12} x_i y_i = 6314$
- $\bar{x} = 42.25$
- $\bar{y} = 12$
- $S_{xy} = 230$
- $S_{xx} = 844.25$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{230}{844.25} = 0.27 \text{ (to 2 d.p)}$$

$$a = \bar{y} - b\bar{x} = 0.49 \text{ (to 2 d.p)}$$

Your turn. . .

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$
$$=$$
$$=$$

$$\widehat{\sigma}^2 = \frac{S_{yy} - bS_{xy}}{n-2}$$

$$=$$

$$=$$

Confidence intervals for regression parameters – example

Your turn...

Using the values $S_{xx} = 844.25$, $n = 12$, $a = 0.49$, $b = 0.27$ and $\widehat{\sigma}^2 = 1.13$ and noting from Excel that `=TINV(0.05,12-2)` is 2.228, the 95% confidence intervals for α and β are given by,

$$a \pm t^\star \times \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}} =$$

$$=$$

$$=$$

$$b \pm t^\star \times \frac{\widehat{\sigma}}{\sqrt{S_{xx}}} =$$

$$=$$

$$=$$

Confidence intervals for regression parameters – example

Excel does all this for us!

	Coeff.	Std. Error	...	Lower 95%	Upper 95%
Intercept	0.49	1.58	...	–3.03	4.01
Dose	0.27	0.04	...	0.19	0.35

Note: any differences can be put down to rounding errors in our calculations.

Revision: t tests

Recall the general set up for a one sample t test.

- If we have sample of size n and we want to test,

$$H_0 : \mu = 5 \quad \text{against} \quad H_1 : \mu \neq 5.$$

- The observed test statistic is:

$$t_{\text{obs}} = \frac{\bar{x} - \mu}{\text{SE}(\bar{x})} = \frac{\bar{x} - 5}{s/\sqrt{n}} \sim t_{n-1}.$$

- The p-value is the likelihood of getting the observed test statistic or something more extreme if the null hypothesis is true. In this case it is:

$$2P(t_{n-1} > |t_{\text{obs}}|).$$

- We reject the null hypothesis if the p-value is less than 0.05 and do not reject otherwise.

Hypothesis testing on regression slope parameter

- Recall our model, $y = \alpha + \beta x + \varepsilon$, of particular importance is determining whether x actually has a linear effect on y.
- This is equivalent to testing the hypothesis

$$H_0 : \beta = 0 \qquad \text{versus} \qquad H_1 : \beta \neq 0.$$

- The appropriate test statistic is $t_{\text{obs}} = \dfrac{b - \beta}{\text{SE}(b)} = \dfrac{b}{\sqrt{\widehat{\sigma}^2 / S_{xx}}}$.
- Under the null hypothesis, H_0, the value t_{obs} follows a t distribution with $n - 2$ degrees of freedom.
- The p-value is given by

$$2P(t_{n-2} \geq |t_{\text{obs}}|) = 2 * \texttt{TDIST}(\texttt{ABS}(t_{\text{obs}}), \texttt{n} - 2, 1) \text{ in Excel.}$$

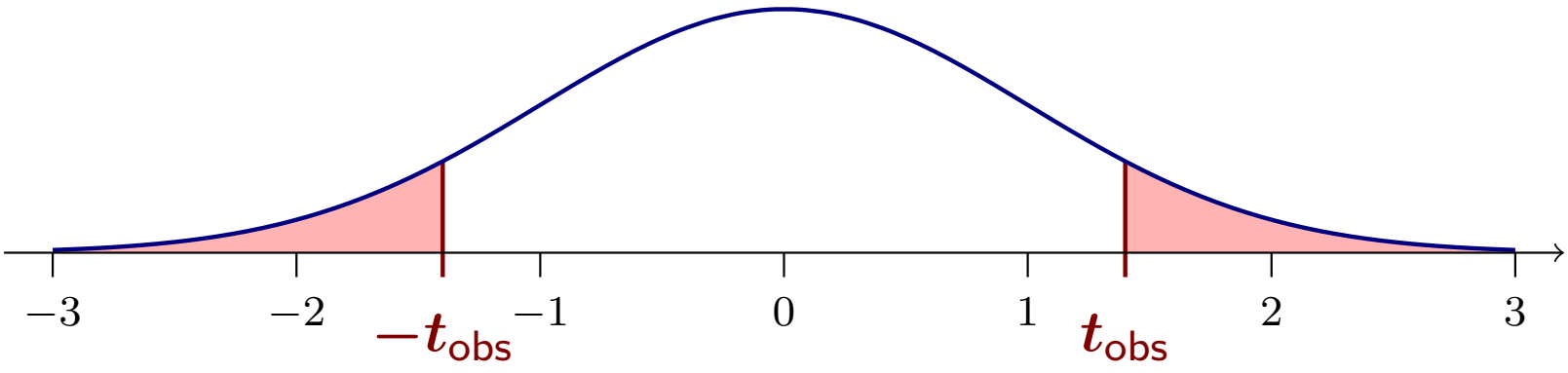
- If the p-value is bigger than the significance level (e.g. 0.05), we do not reject H_0.

Hypothesis testing on regression slope parameter in Excel

	Coeff.	Std. Error	t-Stat	P-value
Intercept	a	$\mathsf{SE}(a)$	$t_0 = \dfrac{a}{\mathsf{SE}(a)}$	$2P(t_{n-2} > \lvert t_0 \rvert)$
X variable	b	$\mathsf{SE}(b)$	$t_1 = \dfrac{b}{\mathsf{SE}(b)}$	$2P(t_{n-2} > \lvert t_1 \rvert)$

$$t_{\mathsf{obs}} = \frac{b}{\mathsf{SE}(b)}; \quad \mathsf{SE}(b) = \frac{b}{t_{\mathsf{obs}}}; \quad b = t_{\mathsf{obs}} \times \mathsf{SE}(b)$$

- Excel tests the null $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ so the p-value is for a two tail test:



Hypothesis testing on regression slope parameter – example

In our dose and urine concentration example:

$$b = 0.27, \quad \widehat{\sigma}^2 = 1.13, \quad S_{xx} = 844.25 \quad \text{and} \quad n = 12.$$

We are testing,

$$H_0 : \beta = 0 \quad \text{against} \quad H_1 : \beta \neq 0.$$

The observed test statistic is,

$$t_{\text{obs}} = \frac{b}{\text{SE}(b)} = \frac{b}{\sqrt{\widehat{\sigma}^2 / S_{xx}}} = \frac{0.27}{\sqrt{1.13/844.25}} = 7.43$$

and the p-value is given by

$$2P(t_{n-2} \geq |t_{\text{obs}}|) = 2 * \texttt{TDIST}(\texttt{ABS}(7.43), 10, 1)$$

$$=$$

Since the p-value is less than 0.05 we reject the null hypothesis, i.e. we reject $H_0 : \beta = 0$. Hence, there is strong evidence to suggest that x has a linear effect on y.

Hypothesis testing on regression slope parameter – Excel

	Coeff	Std Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0.49	1.58	0.31	0.76	-3.03	4.01
Dose	0.272	0.0367	7.41	0.00	0.19	0.35

- Excel automatically tests $H_0 : \beta = 0$ against $H_0 : \beta \neq 0$.
- The t statistic Excel calculates is:

$$t = \frac{b}{\mathsf{SE}(b)} = \frac{0.272}{0.0367} = 7.41.$$

- The corresponding p-value is:

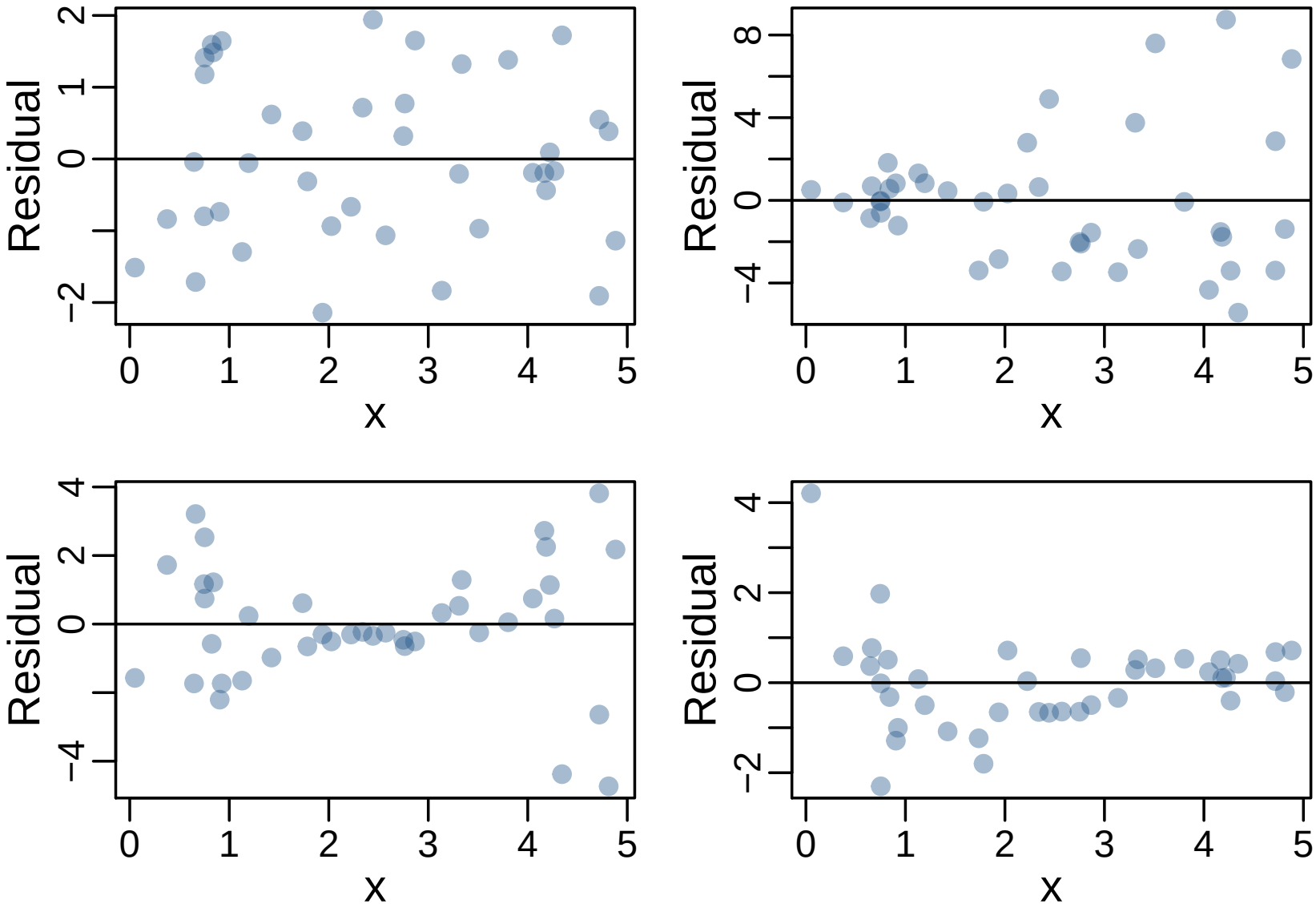
$$2P(t_{n-2} \geq |t_{\mathsf{obs}}|) = 2P(t_{10} \geq 7.41) < 0.001.$$

- Note the agreement between the p-value and the confidence interval – we conclude that the slope coefficient is significant (significantly different to zero).

Assumption 3: Homoskedasticity (constant error variance)

- Homoskedasticity (homo: same, skedasticity: spread) constant variance is good.
- Violations of homoskedasticity, called heteroskedasticity, make it difficult to estimate the "true" standard deviation of the errors, usually resulting in confidence intervals that are too wide or too narrow.
- Heteroskedasticity may also have the effect of giving too much weight to small subset of the data (namely the subset where the error variance was largest) when estimating coefficients.
- Heteroskedasticity appears in plots of residuals versus x. Look for evidence of residuals that are getting larger (more spread-out).

Assumption 3: Homoskedasticity



Assumption 4: Normality

- Violations of normality of the errors can compromise our inferences. The calculation of confidence intervals may be too wide or narrow and our conclusions from our hypothesis tests may be incorrect.
- You can use a boxplot or histogram to check for normality.
- In some cases, the problem may be due to one or two outliers. Such values should be scrutinised closely: are they genuine, are they explainable, are similar events likely to occur again in the future.
- Sometimes the extreme values in the data provide the most useful information.

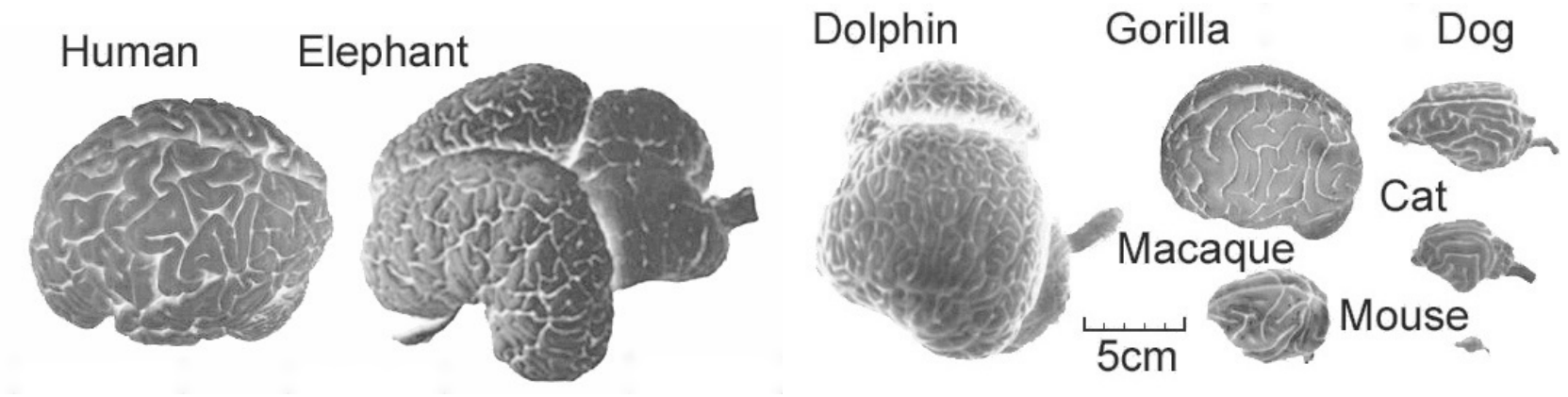
Outline

Alternative models

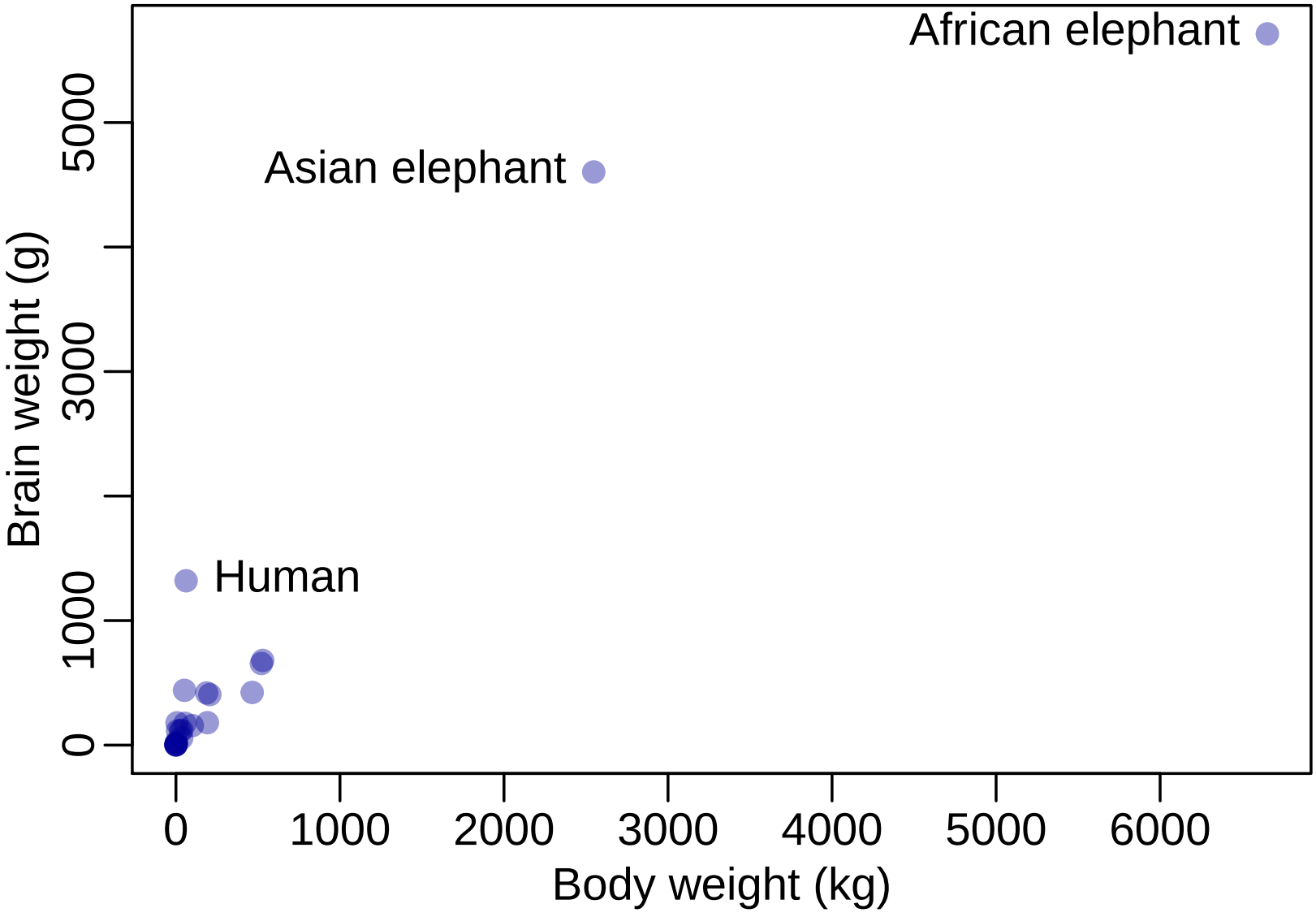
- A scatterplot can be used to give an initial indication as to whether a linear model is appropriate or not.
- What can we do if it is clear that a linear relationship between y and x will not be useful for estimating y from x?

Example (Brain to body mass ratio)

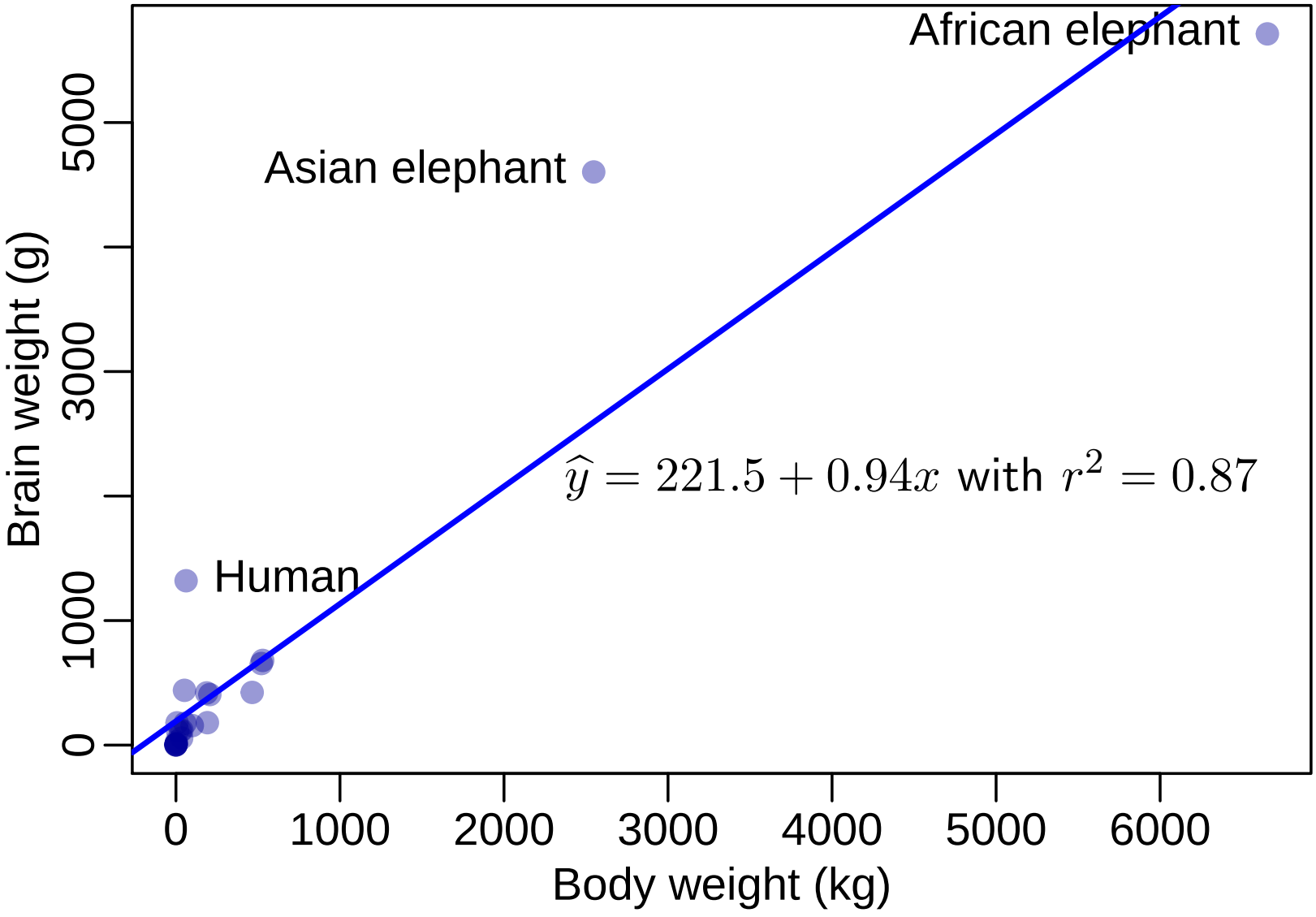
- Brain size usually increases with body size in animals.
- The relationship is not linear. Generally, small mammals have relatively larger brains than big ones.



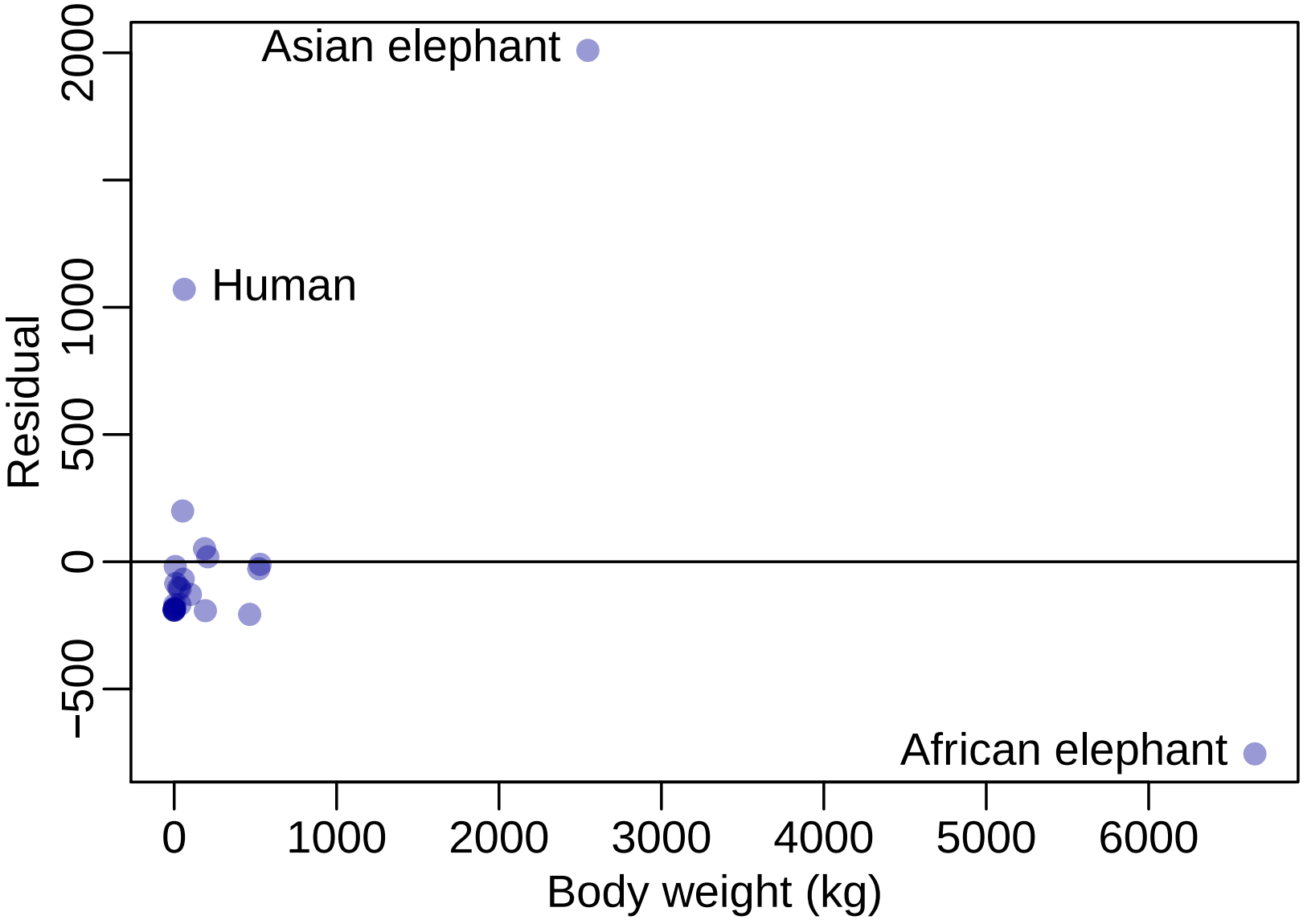
Brain to body mass ratio



Brain to body mass ratio – linear regression



Brain to body mass ratio – residual plot



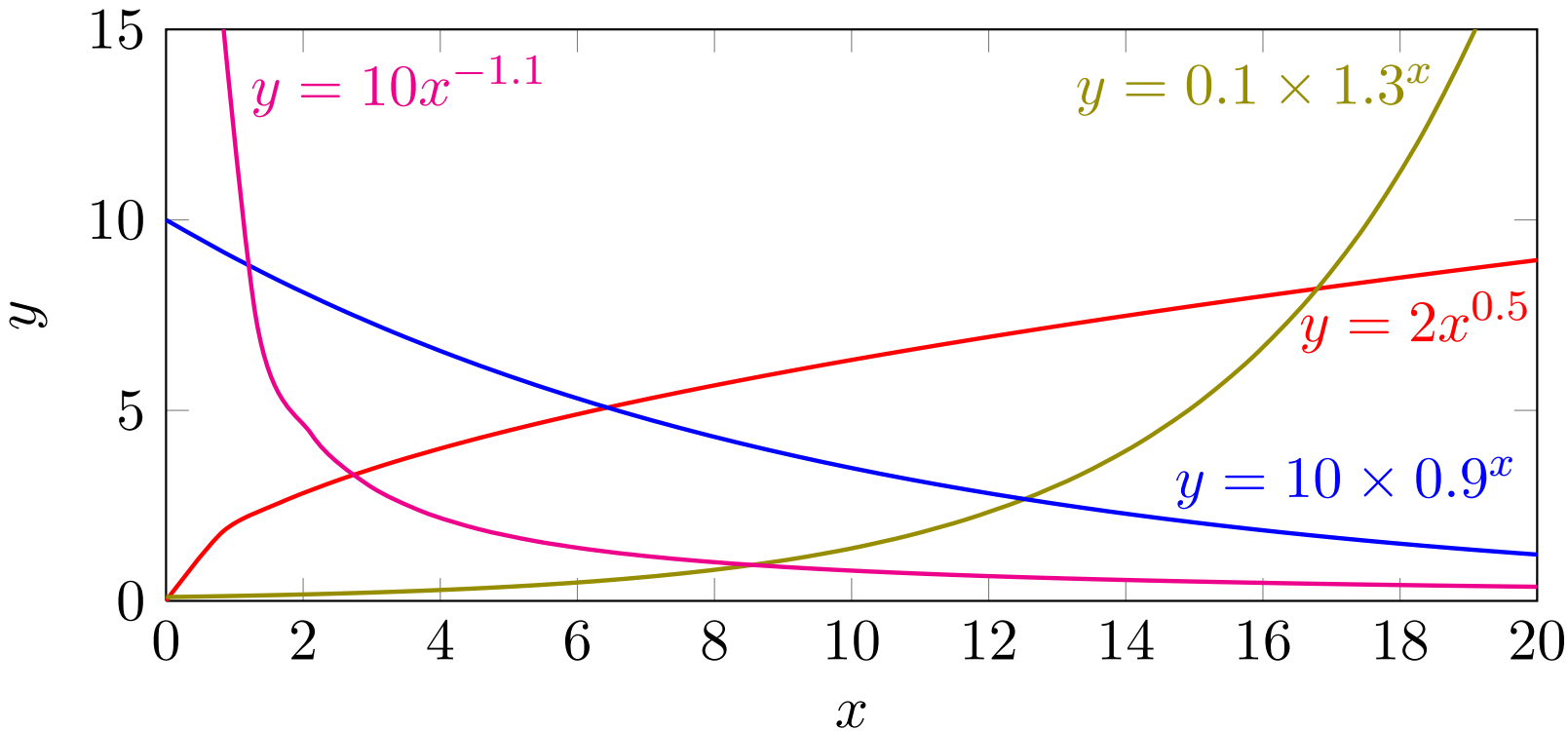
Transformations

- If the points don't follow a linear pattern, it is not be sensible to use a linear model.
- Apart from the usual scatter plot, residual plot might also help to indicate which alternative model may be appropriate.
- The general idea is:
  1. apply a function to your data to linearise the points – we will introduce variables Y and X (functions of the original y and x) which will result in a new set of points (X_i, Y_i) which have a linear trend.
  2. estimate a straight line, checking any necessary assumptions.
  3. transform estimates back to the original model.
- It's not always obvious which linearising transformation is appropriate – often scientific theory can give us guidance.
- The important question is how to define the linearising transformation.

Linearising transformations

There is no general rule, but there are two frequently occurring classes of relationships for which standard linearising transformations are available:

- allometric $y = Ax^B$; and
- exponential $y = AB^x$ relationships.



Exponential transformation

- The transformation is so called because the independent variable occurs as an exponent. It uses

$$y = AB^x \quad \text{or equivalently} \quad y = Ae^{Cx},$$

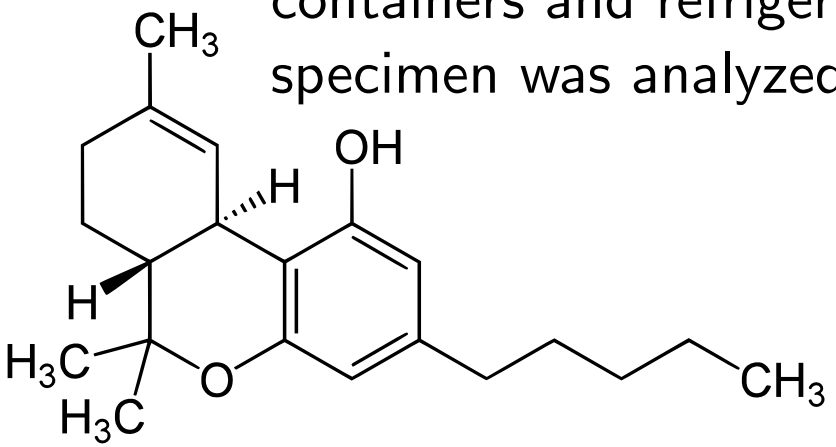
  where $C = \log(B)$.
- Here A and B (or C) are parameters which need to be estimated.

Example

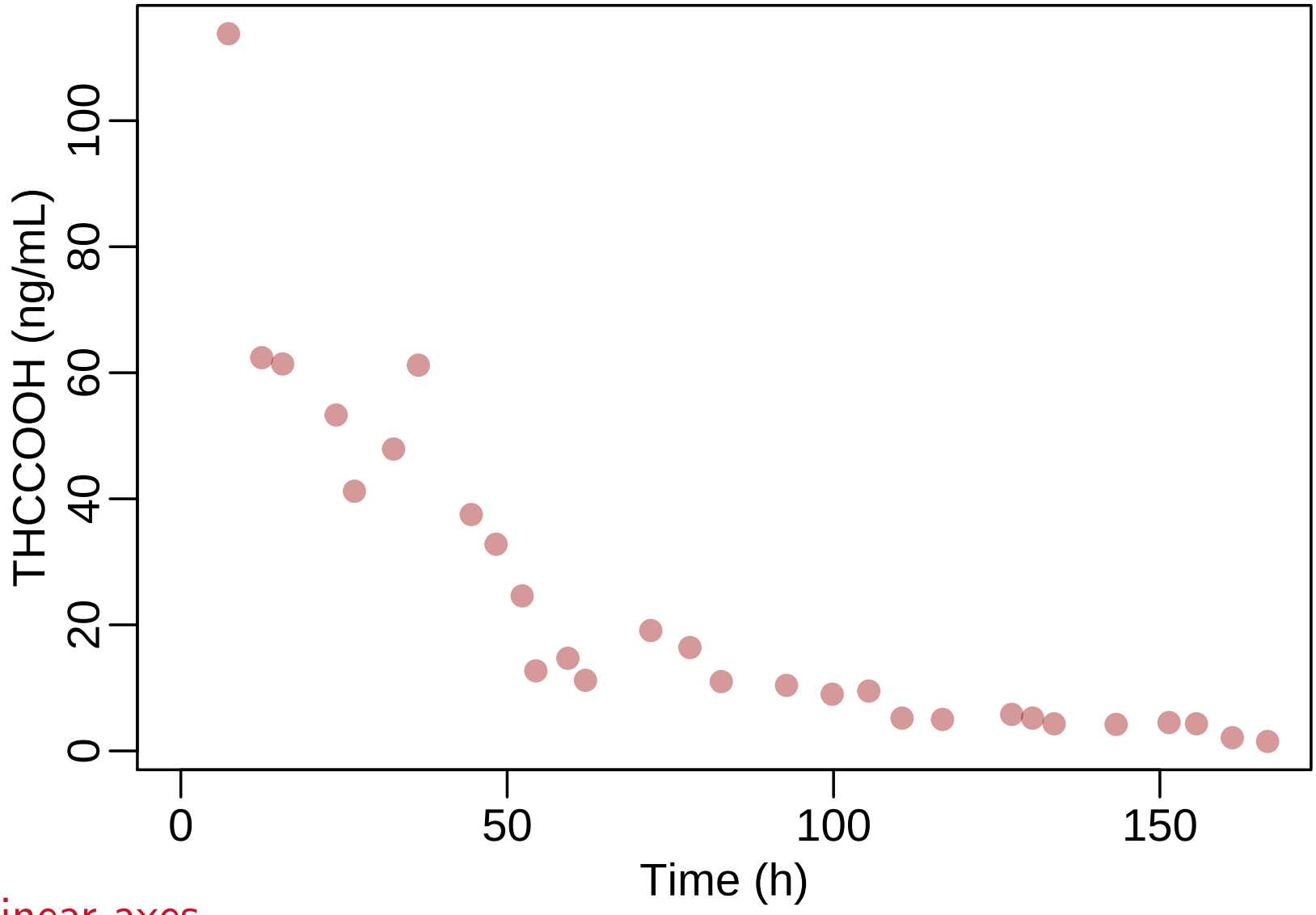
- Radioactive decay: $N(t) = N_0 e^{-\lambda t}$.
- Bacterial growth: $N(t) = N_0 2^{\nu t}$ where $N(t)$ is the number of bacteria at time t, N_0 is the initial number of bacteria and ν is the doubling rate (number of divisions per unit time).

Exponential example – THC excretion

- Tetrahydrocannabinol (Δ^9-THC), is the principal psychoactive constituent of the cannabis plant.
- More than 55% of THC is excreted in the feces and 20% in the urine. The main metabolite in urine is THCCOOH.
- Huestis et. al. (1996) characterises the urinary excretion profiles of THCCOOH in a healthy male subject after single, short-term, smoked dose of marijuana.
- The concentration (y) over time (x) can be modelled using an exponential model: $y = AB^x$.
- Specimens ($n = 29$) were collected in polypropylene containers and refrigerated immediately after urination. Each specimen was analyzed for THCCOOH by GC-MS.

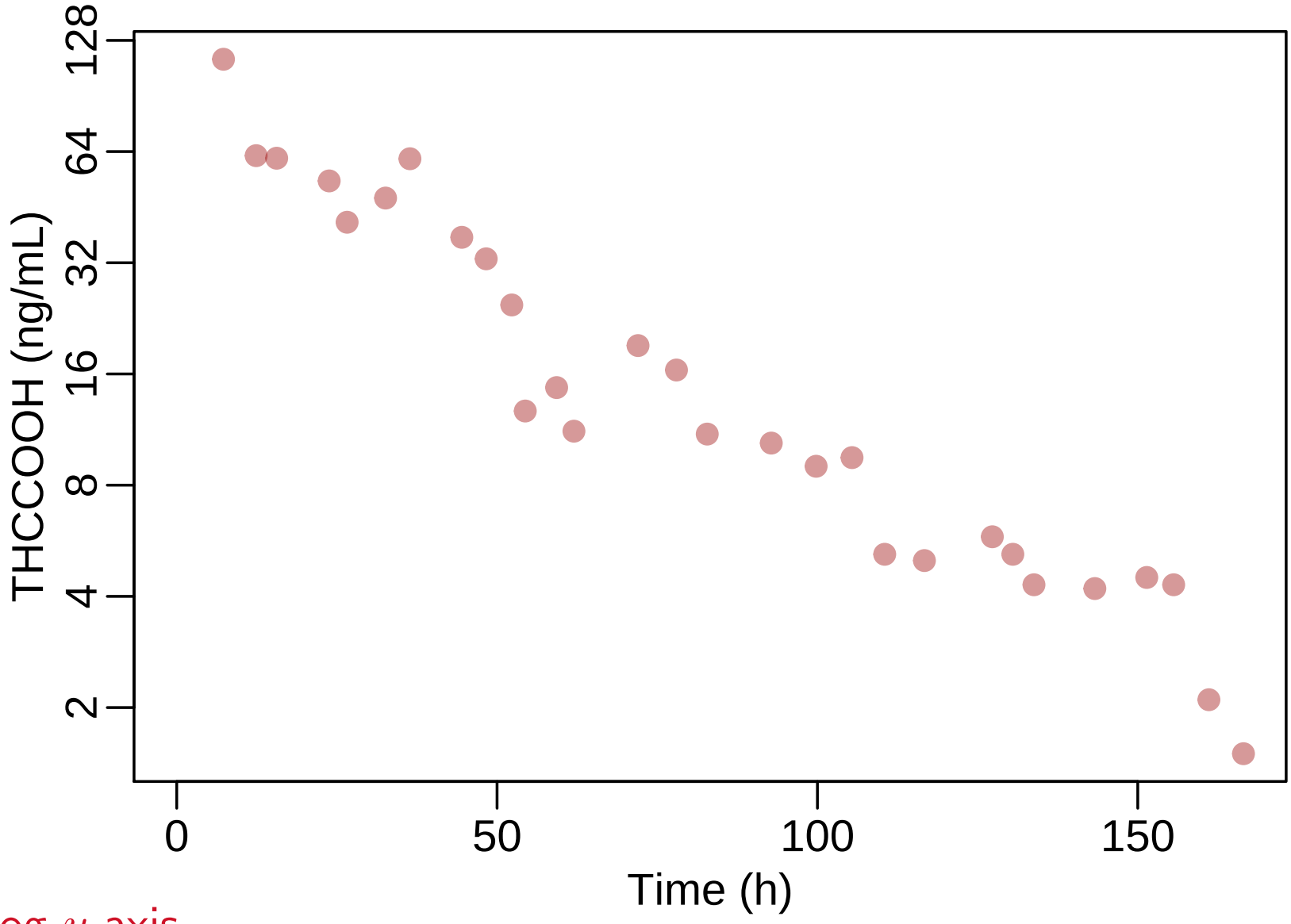


Exponential example – THC excretion



Linear axes

Exponential example – THC excretion



Log y axis

Allometric transformation

- An allometric (allo = different, metric = measure) relationship is where one variable scales at a different rate to the other:

$$y = Ax^B$$

- Here A and B are parameters which need to be estimated.

Example

- Kleiber's law (1932) relating metabolic rate and body mass:

$$\text{Metabolic Rate} = 70(\text{Body Mass})^{0.75}.$$

- Prediction of human pharmacokinetic (PK) parameters [e.g., clearance, volume of distribution V_d, elimination half life $t_{1/2}$] based on body weight (W):

$$\text{PK} = a\text{W}^b.$$

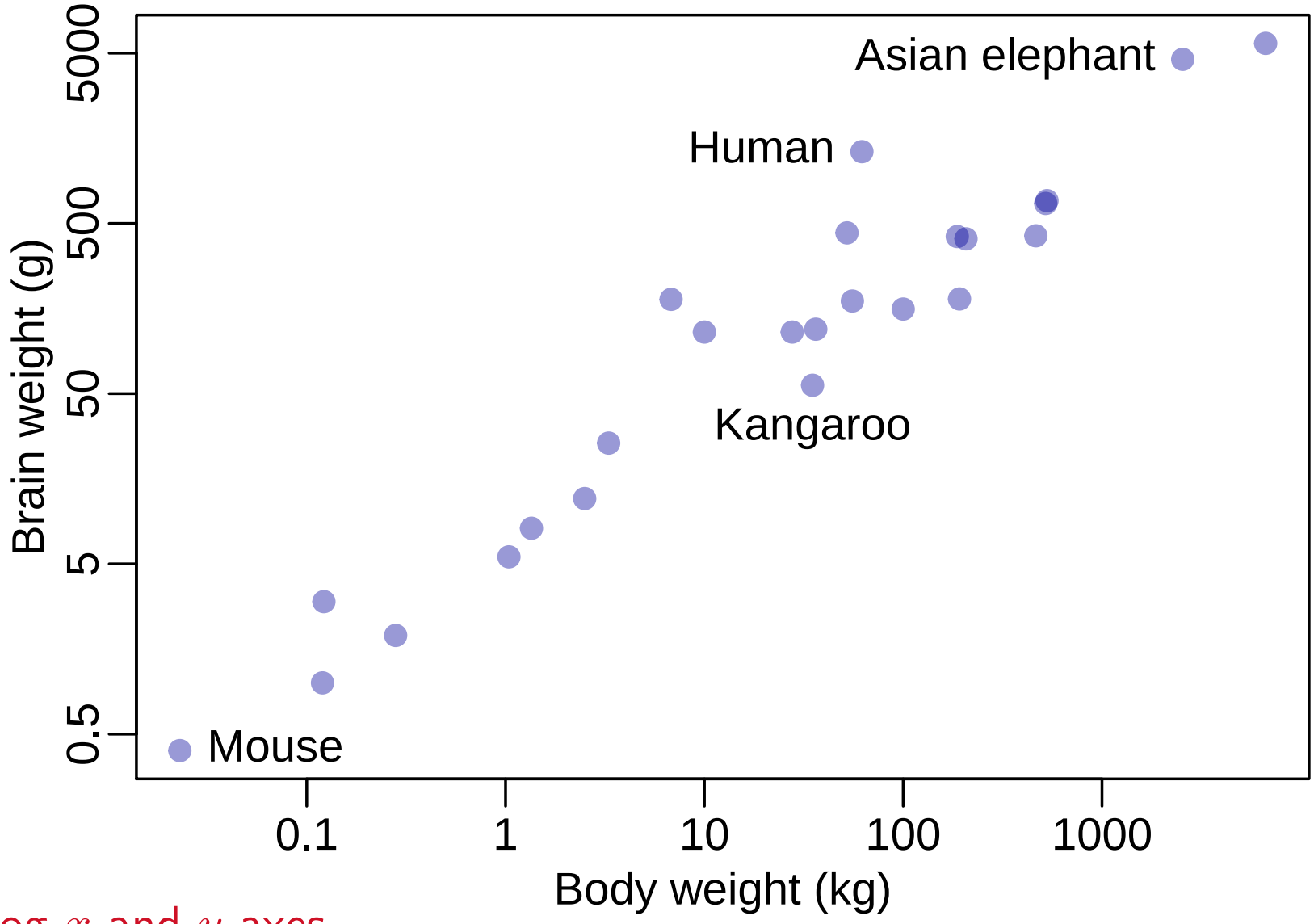
Allometric example – brain to body mass ratio

Body weight and brain size can be modelled using an allometric relationship $y = Cx^B$, where y and x are body and brain weights and C is known as the cephalisation factor.

Consider the following data on $n = 24$ mammals:

Animal	Body (kg)	Brain (g)	Animal	Body (kg)	Brain (g)
Mouse	0.023	0.4	Chimpanzee	52.16	440
Golden hamster	0.12	1	Sheep	55.5	175
Mole	0.122	3	Human	62	1320
Guinea pig	1.04	5.5	Jaguar	100	157
Mountain beaver	1.35	465	Donkey	187.1	419
Rabbit	2.5	12.1	Pig	192	180
Cat	3.3	25.6	Gorilla	207	406
Rhesus monkey	6.8	179	Cow	465	423
Potar monkey	10	115	Horse	521	655
Goat	27.66	115	Giraffe	529	680
Kangaroo	35	56	Asian elephant	2547	4603
Grey wolf	36.33	119.5	African elephant	6654	5712

Allometric example – brain to body mass ratio



Log x and y axes

How do we apply these linearising transformations

What are the linearising transformations in these two standard cases? Let's take the log of both sides:

Allometric, $y = Ax^B$

$$\begin{aligned}\log(y) &= \log(Ax^B) \\ &= \log(A) + \log(x^B) \\ &= \log(A) + B\log(x)\end{aligned}$$

This linear relationship has the log of y as the dependent variable and the log of x as the explanatory variable.

Double log or Log-log model

Exponential, $y = AB^x$

$$\begin{aligned}\log(y) &= \log(AB^x) \\ &= \log(A) + \log(B^x) \\ &= \log(A) + \log(B)x\end{aligned}$$

This linear relationship has the log of y as the dependent variable but just x as the explanatory variable.

Semi-log or Log-linear model

Semi-log transformation (exponential relationship)

- Say an **exponential** trend of the type $y = AB^x$ is expected.
- Take (natural) logs of both sides to obtain

$$\log(y) = \log(A) + \log(B)x$$
$$Y = \alpha + \beta X$$

and so if we put $Y = \log(y)$, $X = x$, $\alpha = \log(A)$ and $\beta = \log(B)$ the line we now want to estimate is $Y = \alpha + \beta X$.

Procedure:

1. Perform a semi-log transform: for each of the observations, $i = 1, 2, \ldots, n$, set $X_i = x_i$ and $Y_i = \log(y_i)$.
2. Find a least squares regression line for $Y = \alpha + \beta X$ in the usual way.
3. Transform back to obtain the fitted curve $y = AB^x$ using $A = e^{\alpha}$ and $B = e^{\beta}$.

Exponential example – THC excretion – formulation

- Our model for THC excretion can be written as

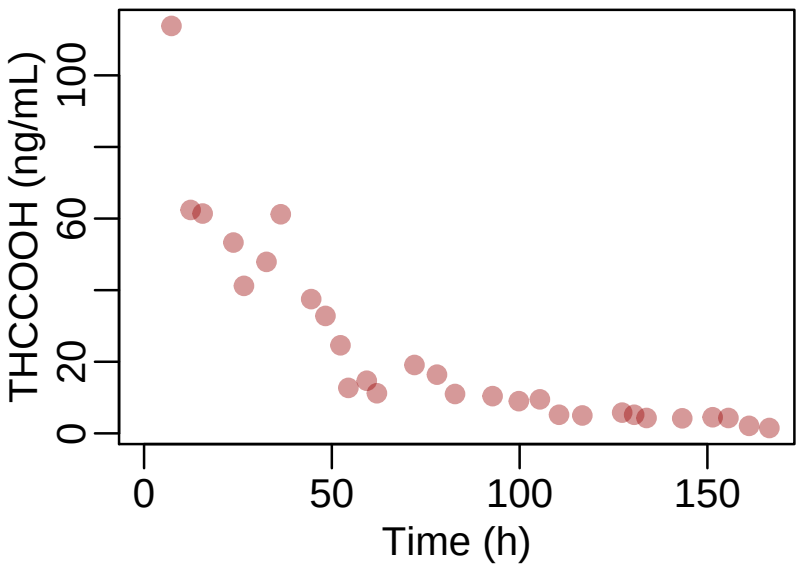
$$y = AB^x$$

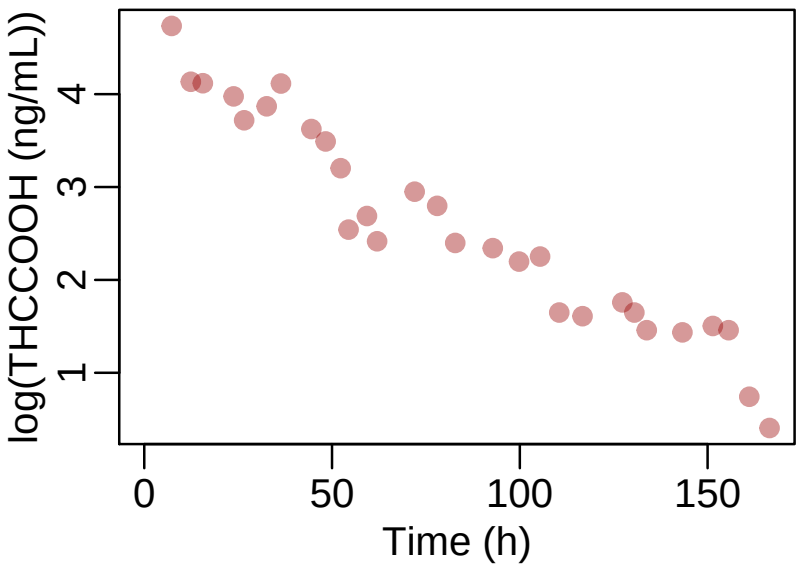
where y is the concentration of THCCOOHL (in ng/mL) in the urine and x is the time since inhalation.

- To estimate A and B we use the semi-log transformation:

$$Y = \alpha + \beta X$$

where $Y = \log(y)$, $X = x$, $\alpha = \log(A)$ and $\beta = \log(B)$.





Exponential example – THC excretion – calculation

x	y	$\log(y)$	x	y	$\log(y)$
7.3	113.8	4.73	82.8	11	2.4
12.4	62.4	4.13	92.8	10.4	2.34
15.6	61.4	4.12	99.8	9	2.2
23.8	53.3	3.98	105.4	9.5	2.25
26.6	41.2	3.72	110.5	5.2	1.65
32.6	47.9	3.87	116.7	5	1.61
36.4	61.2	4.11	127.3	5.8	1.76
44.5	37.5	3.62	130.5	5.2	1.65
48.3	32.8	3.49	133.8	4.3	1.46
52.3	24.6	3.2	143.3	4.2	1.44
54.4	12.7	2.54	151.4	4.5	1.5
59.3	14.7	2.69	155.6	4.3	1.46
62	11.2	2.42	161.1	2.1	0.74
72	19.1	2.95	166.5	1.5	0.41
78	16.4	2.8			

Recall: $X = x$ and $Y = \log(y)$

$$\sum X_i = \sum x_i = 2403$$

$$\sum X_i^2 = \sum x_i^2 = 267596.08$$

$$\sum Y_i = \sum \log(y_i) = 75.24$$

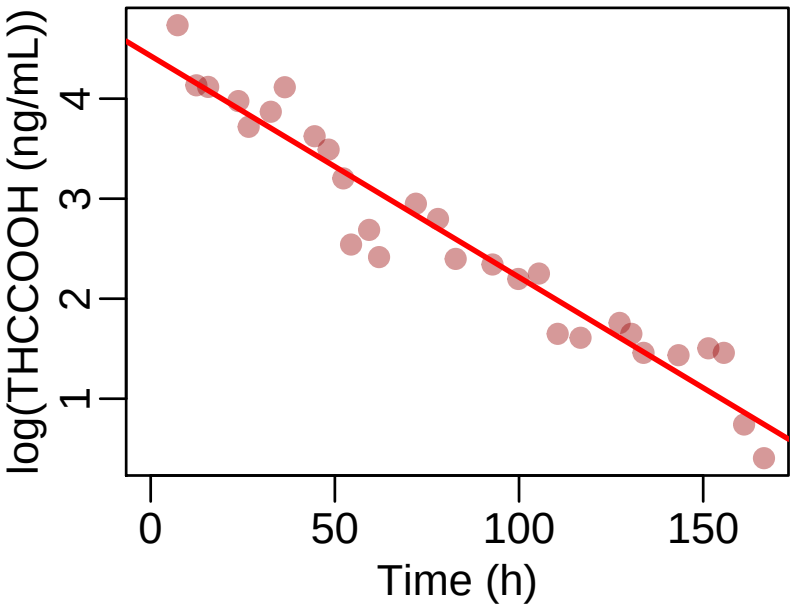
$$\sum Y_i^2 = \sum \log(y_i)^2 = 231.1384$$

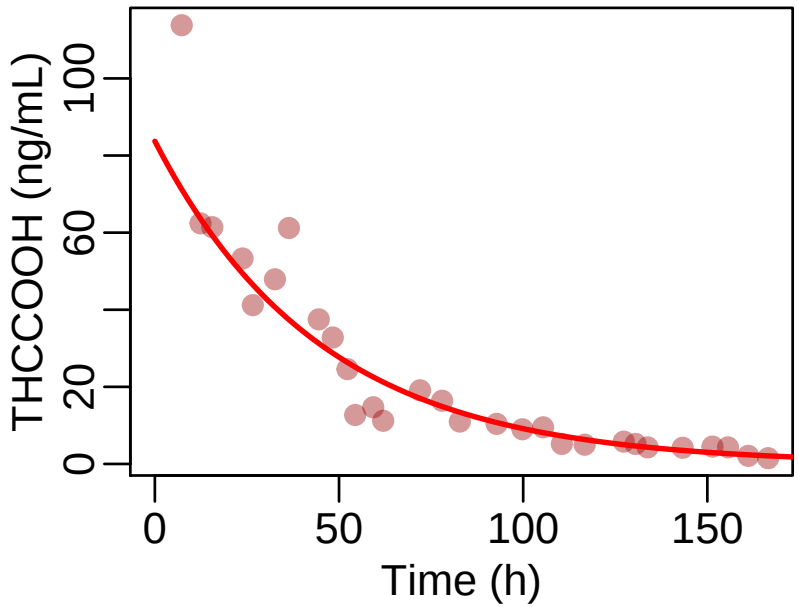
$$\sum X_i Y_i = \sum x_i \log(y_i) = 4720.627$$

$$\overline{X} = \overline{x} = 82.862$$

$$\overline{Y} = \overline{\log(y)} = 2.594$$

Exponential example – THC excretion – results





Our fitted least squares regression line is:

$$\widehat{\log(y)} = 4.427 - 0.022x.$$

We can transform back to the original model:

$$\widehat{y} = 83.7 \times 0.978^x,$$

as $e^{4.427} = 83.7$ and $e^{-0.022} = 0.978$.

Exponential example – THC excretion – interpretation

Definition (Semi-log model interpretation)

If the estimated model is

$$\widehat{\log(y)} = a + bx,$$

we interpret this as follows:

> *On average, a one unit change in x will result in a $b \times 100\%$ change in y.*

E.g. if the estimated coefficient is $b = 0.05$ that means that a one unit increase in x will generate a 5% increase in y.

Example (THC model: $\widehat{\log(y)} = 4.427 - 0.022x$)

On average, for each hour of time that passes, the concentration of THC in the body will decrease by 2.2%.

Double log transformation (allometric relationship)

- Say an **allometric** trend of the type $y = Ax^B$ is expected.
- Take (natural) logs of both sides to obtain

$$\log(y) = \log(A) + B\log(x)$$
$$Y = \alpha + \beta X$$

and so if we put $Y = \log(y)$, $X = \log(x)$, $\alpha = \log(A)$ and $\beta = B$ the line we now want to estimate is $Y = \alpha + \beta X$.

Procedure:

1. Perform a double log transform: for each of the observations, $i = 1, 2, \ldots, n$, set $X_i = \log(x_i)$ and $Y_i = \log(y_i)$.
2. Find a least squares regression line for $Y = \alpha + \beta X$ in the usual way.
3. Transform back to obtain the fitted curve $y = Ax^B$ using $A = e^{\alpha}$ and $B = \beta$.

Allometric example – body and brain mass – formulation

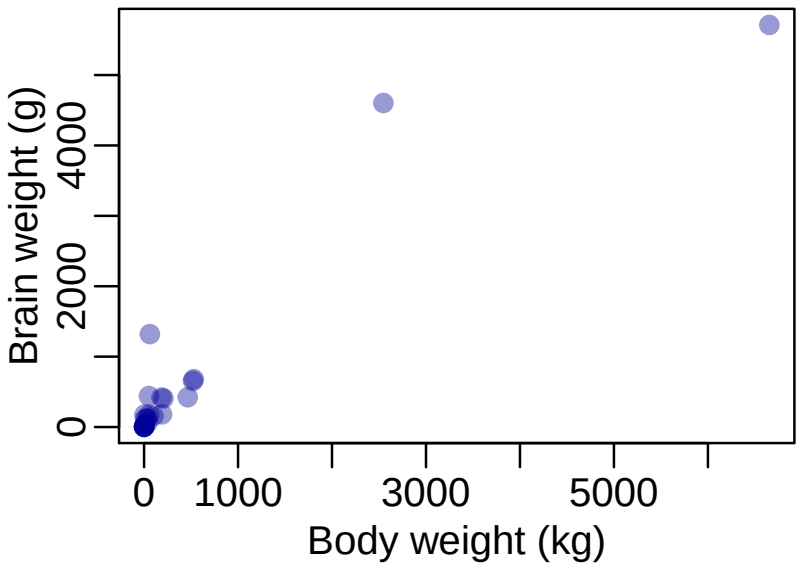
- The large variation in scale for both the y and x axes in this example indicate that an allometric model may be appropriate:

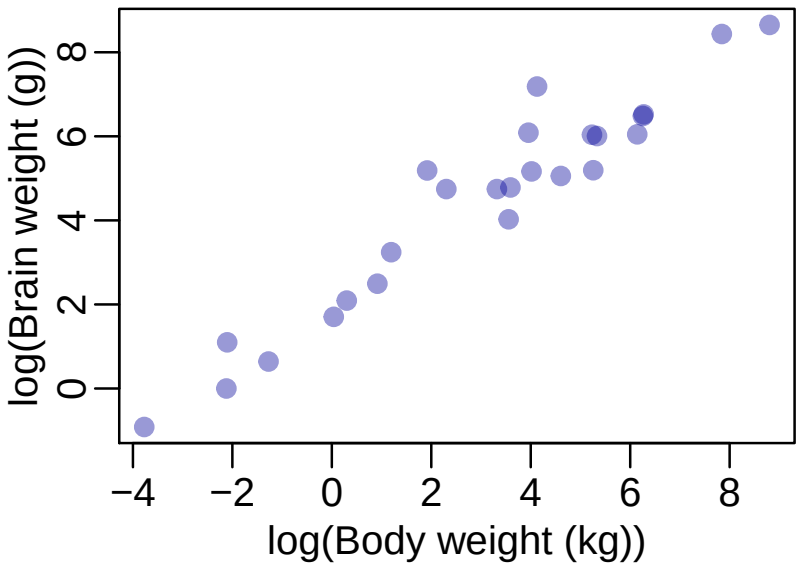
$$y = Ax^B.$$

  where y is the brain mass and x is the body mass.
- To estimate A and B we use the double log transformation:

$$Y = \alpha + \beta X$$

  where $Y = \log(y)$, $X = \log(x)$, $\alpha = \log(A)$ and $\beta = B$.





Allometric example – body and brain mass – calculation

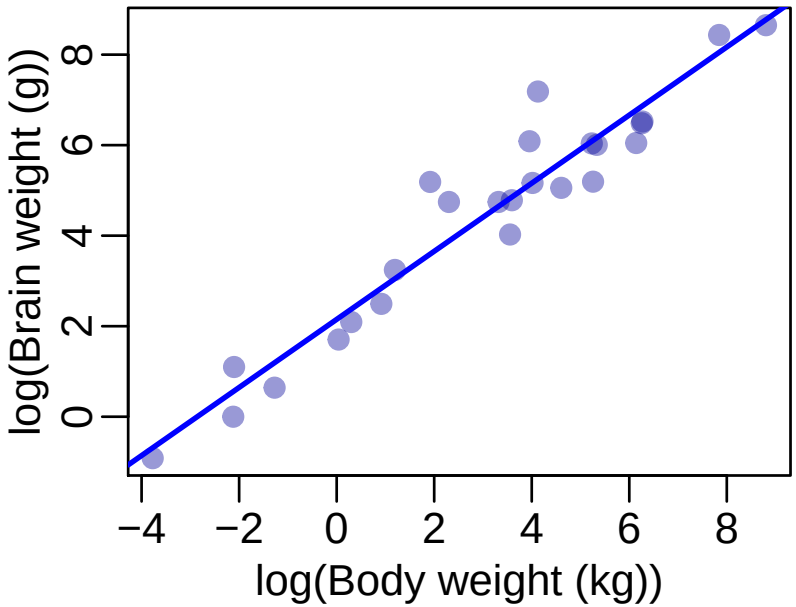
In the excel data file, `BodyBrainData.xlsx` create 2 new columns:

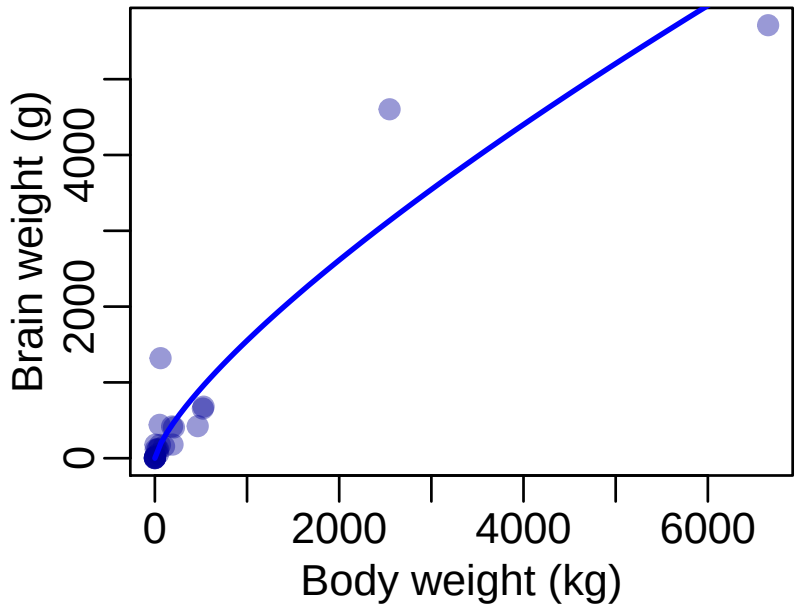
	Body (kg)	Brain (g)	log(body)	log(brain)
Mouse	0.023	0.4	`=ln(B2)`	`=ln(C2)`
Golden hamster	0.12	1	-2.10	0.00
Mole	0.122	3	-2.10	1.10
⋮	⋮	⋮	⋮	⋮

Then use these two columns to perform a simple linear regression:

	Coeff	Std Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.15	0.20	10.72	0.00	1.75	2.57
log(body)	0.75	0.05	16.45	0.00	0.66	0.85

Allometric example – body and brain mass – results





Our fitted least squares regression line is:

$$\widehat{\log(y)} = 2.15 + 0.75 \log(x).$$

We can transform back to the original model:

$$\widehat{y} = 8.58 \times x^{0.75},$$

as $e^{2.15} = 8.58$.

Allometric example – body and brain mass – interpretation

Definition (Double log model interpretation)

If the estimated model is

$$\widehat{\log(y)} = a + b\log(x),$$

we interpret this as follows:

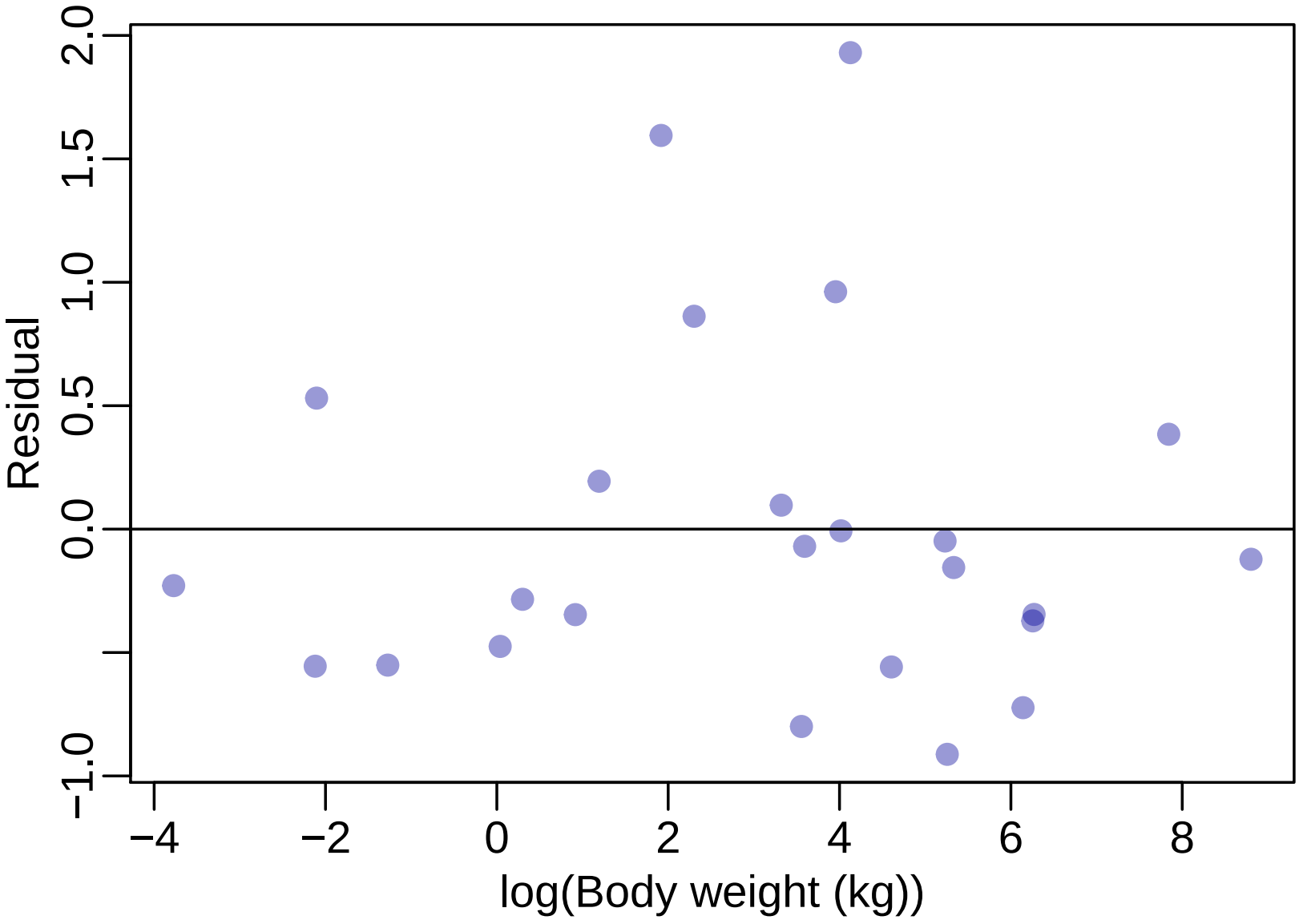
> *On average, a one percent change in x will result in a b% change in y.*

E.g. if the estimated coefficient is $b = -2$ that means that a one percent increase in x will generate a 2% decrease in y.

Example (Body and brain mass: $\widehat{\log(y)} = 2.15 + 0.75\log(x)$)

For mammals, on average, a one percent increase in body mass leads to a 0.75% increase in brain mass.

Allometric example – body and brain mass – residuals



Allometric example – body and brain mass – testing

- Jerison (1983) suggests that if brain size is "driven" by body surface area then the relationship should be of the form:

$$y = Ax^{2/3} \quad \text{i.e.} \quad \log(y) = \log(A) + \frac{2}{3}\log(x).$$

- To test the null hypothesis, $H_0 : \beta = 2/3$, consider the Excel output:

	Coeff	Std Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.15	0.20	10.72	0.00	1.75	2.57
log(body)	0.75	0.05	16.45	0.00	0.66	0.85

- The hypothesised value of $2/3$ is on the lower edge of the CI – this is consistent with other studies which have found that $3/4$ is a more likely value, suggesting that there's more to the relationship than just than body surface area.

▸ More

Allometric example – body and brain mass – caution

- As this formula is based on data from mammals, it should be applied to other animals with caution.