Simple Linear Regression

Michael Noonan

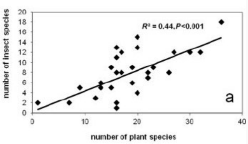
Biol 520C: Statistical modelling for biological data

Table of contents

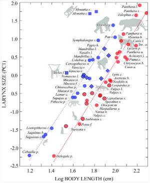
Simple Linear Regression

Simple Linear Regression

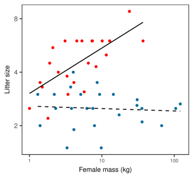
When observing biological systems, one of the first questions we often ask ourselves is: "Is there a relationship between X and Y?".



Chmura et al. 2013



Bowling et al. 2020



Johnson et al. 2017

Our verbal hypothesis in this case is 'X is proportional to Y'. But looking at the data isn't enough. So how do we approach the problem statistically?

Simple Linear Regression Cont.

Fitting a straight line to data is the root of all modern modelling.

The ‘simple’ in simple linear regression refers to the fact that there is only one parameter affecting the relationship between x and y.

The method itself isn’t simple and there’s a lot going on under the hood.

Regression Data

Data is of the form:

$$d = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

X	Y
x_1	y_1
x_2	y_2
...	...
x_n	y_n

Verbal description of the hypothesis:

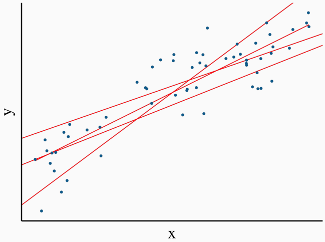
y increases with x or... y is proportional to x

Regression Model

More formally, a straight line is described by an intercept (β_0) and a slope (β_1): $y_i = \beta_0 + \beta_1 x_i$

With data $d = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, the question is what values of β_0 and β_1 best describe the relationship between x and y

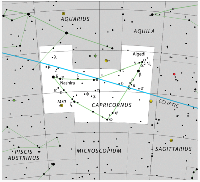
(i.e., what line do you draw through the data?)



Least Squares Fitting

Least Squares Fitting

The methods for fitting lines/shapes/curves date back thousands of years and are rooted in astronomy and geodesy. These original approaches served humanity well for thousands of years, but the challenges of navigating the Earth's oceans during the 'Age of Exploration' required more precise methods and there were a flurry of activity during the course of the eighteenth century.



Source: www.constellation-guide.com

In 1805, Legendre published an algebraic procedure for fitting linear equations to data. His 'least squares' approach assumed each observation y_i is accompanied by some amount of noise ε_i. If you further constrain the problem such that the sum of the squared errors needs to be minimized, only one line fits the data.

Source: Wikipedia

Least Squares Fitting Cont.

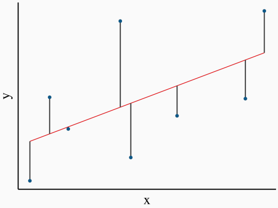
For a given observation, a line predicts y_i to be $\beta_0 + x_i\beta_1$.

This implies that the error for y_i is $\varepsilon_i = y_i - (\beta_0 + x_i\beta_1)$

i.e., observed $-$ expected

...and the sum of the squared errors is $\sum_{i=1}^{n} \varepsilon_i^2$ or $\sum_{i=1}^{n}(y_i - (\beta_0 + x_i\beta_1))^2$.

We want to find the value for β_0 and β_1 that minimizes this quantity.



Parameter estimation

So how we estimate the parameters β_0 & β_1?

One solution is to calculate $\sum_{i=1}^{n}(y_i - (\beta_0 + x_i\beta_1))^2$ for all values of β_0 and β_1 between $-\infty$ and ∞.

But who wants to do that?

Matrix Algebra Review

Matrices

Matrices are rectangular collections of numbers, generally denoted via bold capital letters.

$$\mathrm{A} = \begin{pmatrix} 2 & 7 & -3 & 4 \\ -7 & 1 & 1 & 8 \\ -9 & 4 & 5 & -1 \end{pmatrix}$$

The **dimension** of a matrix is expressed as number of rows × number of columns. So, A is a 3 × 4 matrix.

It is common to refer to elements in a matrix by subscripts, with the row first and the column second:

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \end{pmatrix}$$

So here, $a_{3,2} = 4$, and $a_{1,3} = -3$.

Vectors

Vectors are special matrices with only one row (called a row vector) or only one column (called a column vector).

$$B = \begin{pmatrix} 2 & 7 & -3 & 5 \end{pmatrix}$$

B is a 4 dimensional row vector (or a 1×4 matrix)

$$C = \begin{pmatrix} 2 \\ 9 \\ -3 \end{pmatrix}$$

C is a 3 dimensional column vector (or a 3×1 matrix)

‘Ordinary’ numbers can be thought of as a 1 × 1 matrices, or scalars (e.g., $D = 7$).

Matrix Addition/Subtraction

To perform matrix addition/subtraction, two matrices must have the **same** number of rows and columns (i.e., dimensions). In that case simply add/subtract each of the individual components:

$$A + B = \begin{pmatrix} 1 & -5 & 4 \\ 2 & 5 & 3 \end{pmatrix} + \begin{pmatrix} 8 & -3 & -4 \\ 4 & -2 & 9 \end{pmatrix} =$$

$$\begin{pmatrix} 1+8 & -5-3 & 4-4 \\ 2+4 & 5-2 & 3+9 \end{pmatrix} =$$

$$\begin{pmatrix} 9 & -8 & 0 \\ 6 & 3 & 12 \end{pmatrix}$$

Matrix addition has many of the same properties as normal addition.

$$A + B = B + A$$

$$A + (B + C) = (A + B) + C$$

Matrix Transposition

To take the transpose of a matrix, simply switch the rows and columns around. The transpose of A can be denoted as A' or A^T.

$$A = \begin{pmatrix} 1 & -5 & 4 \\ 2 & 5 & 3 \end{pmatrix} \qquad A' = A^T = \begin{pmatrix} 1 & 2 \\ -5 & 5 \\ 4 & 3 \end{pmatrix}$$

If a matrix is its own transpose, then that matrix is said to be symmetric, e.g.:

$$A = \begin{pmatrix} 1 & -5 & 4 \\ -5 & 7 & 3 \\ 4 & 3 & 3 \end{pmatrix} = A' = A^T$$

Matrix Multiplication

To multiply a matrix by a scalar, also known as scalar multiplication, multiply every element in the matrix by the scalar.

$$6 \times A = 6 \times \begin{pmatrix} 1 & -5 & 4 \\ 2 & 5 & 3 \end{pmatrix} = \begin{pmatrix} 6 \times 1 & 6 \times -5 & 6 \times 4 \\ 6 \times 2 & 6 \times 5 & 6 \times 3 \end{pmatrix} = \begin{pmatrix} 6 & -30 & 24 \\ 12 & 30 & 18 \end{pmatrix}$$

To multiply two vectors with the same length together, multiply every entry in the two vectors together and then add all the products up (called dot product).

$$x \cdot y = \begin{pmatrix} 1 & -5 & 4 \end{pmatrix} \times \begin{pmatrix} 4 & -2 & 5 \end{pmatrix} = (1 \times 4) + (-5 \times -2) + (4 \times 5) = 34$$

Matrix Multiplication Cont.

To perform matrix multiplication, the first matrix must have the same number of columns as the second matrix has rows. The dimensions of the resulting matrix equals the number of rows of the first matrix, and the number of columns of the second matrix (e.g., a 3×5 matrix × a 5×7 matrix = a 3×7 matrix). To perform the multiplication, you take the dot product of the corresponding row of the first matrix and the corresponding column of the second matrix.

$$C \times D = \begin{pmatrix} 3 & -9 & -8 \\ 2 & 4 & 3 \end{pmatrix} \times \begin{pmatrix} 7 & -3 \\ -2 & 3 \\ 6 & 2 \end{pmatrix} =$$

$$\begin{pmatrix} (3 \times 7) + (-9 \times -2) + (-8 \times 6) & (3 \times -3) + (-9 \times 3) + (-8 \times 2) \\ (2 \times 7) + (4 \times -2) + (3 \times 6) & (2 \times -3) + (4 \times 3) + (3 \times 2) \end{pmatrix} =$$

$$\begin{pmatrix} 21 + 18 - 48 & -9 - 27 - 16 \\ 14 - 8 + 18 & -6 + 12 + 6 \end{pmatrix} = \begin{pmatrix} -9 & -52 \\ 24 & 12 \end{pmatrix}$$

Matrix Properties

An identity matrix is a square matrix where every diagonal entry is 1 and all the other entries are 0

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

The trace of a $n \times n$ matrix is the sum of all the diagonal entries. In other words, for $n \times n$ matrix $trace(A) = tr(A) = \sum_{i=1}^{n} a_{i,i}$

$$tr(I) = tr \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = 1 + 1 + 1 = 3$$

Matrix Inversion

The inverse of a matrix is a special matrix that, when multiplied with its inverse, turn any matrix into an Identify matrix.

e.g., the matrix B is the inverse of matrix A if $AB = BA = I$.

The inverse of matrix is denoted as $B = A^{-1}$, so $AA^{-1} = I$

Inverting matrices requires a complicated algorithm, so we usually rely on computers to perform the calculations (e.g. the `solve()` function in `R`).

Linear regression and matrix notation

Linear regression and matrix notation

Given our dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ we can re-write x, y, and our regression parameters as matrices:

The observations of the response variable y are grouped into a single column, $n \times 1$, matrix

$$\mathrm{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

The regression coefficients β_0 and β_1 are grouped into a 2×1 matrix

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

The observations of the predictor are grouped into a two column, $n \times 2$ matrix.

$$\mathrm{x} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

Linear regression and matrix notation

Why the column of 1s in **x**? When we multiply **x** by β we get:

$$\mathrm{x}\beta = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad = \begin{pmatrix} 1 \times \beta_0 + x_1 \times \beta_1 \\ 1 \times \beta_0 + x_2 \times \beta_1 \\ \vdots \\ 1 \times \beta_0 + x_n \times \beta_1 \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{pmatrix}$$

Mean Squared Error

At each data point, our model results in some amount of error in the prediction, so we have n errors. These form a vector:

$$\varepsilon = y - \mathrm{x}\beta \quad = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{pmatrix} = \begin{pmatrix} y_1 - (\beta_0 + \beta_1 x_1) \\ y_2 - (\beta_0 + \beta_1 x_2) \\ \vdots \\ y_n - (\beta_0 + \beta_1 x_n) \end{pmatrix}$$

So our original regression problem in matrix notation is:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Linear regression in matrix notation

So how does this help us estimate our regression parameters?

We can also rewrite our sum of squares equation in matrix form

$$\sum_{i=1}^{n} \varepsilon_i^2 \rightarrow \sum_{i=1}^{n} \varepsilon^T \varepsilon$$

After some derivations we won't go over, we obtain a formula for the least squares estimates of the parameters:

$$\beta = (\mathrm{x}^{\mathrm{T}}\mathrm{x})^{-1}\mathrm{x}^{\mathrm{T}}\mathrm{y}$$

So instead of plugging in all of the possible values of β_0 and β_1 between $-\infty$ and ∞ to obtain our parameter estimates, all we have to do is a matrix calculation.

Assumptions of linear regression

A note on assumptions

As we just saw, translating a conceptual, verbal hypothesis into something that can actually be estimated with data requires the use of mathematical formulae.

In order to work out these formulae, we often rely on making assumptions/approximations to make the math more manageable.

Some assumptions don't have large impacts on outcomes, while others can be critically important.

Just because a specific estimator makes assumptions that aren't met by real data, this doesn't mean that the relationship doesn't exists or that the estimator is useless, but it does tell you that your estimator can be improved.

Linear regression assumptions

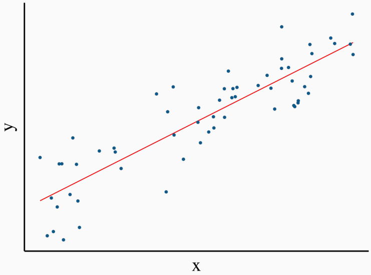
Applying linear regression to a problem relies on satisfying 5 assumptions:

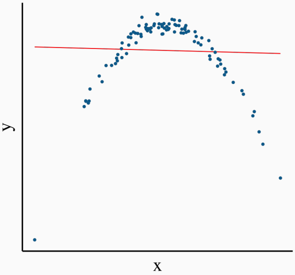
- Correct model specification
- Normality of the residuals
- Homogeneity
- Fixed x
- Independence

i) Correct model specification

In model based inference we need to apply some sort of model to our data.

One of the first things you need to ask yourself before fitting a simple linear model to a dataset is: “Is this really a good model for my data?”

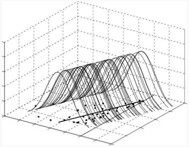




ii) Normality of the residuals

The least squares derivations assume the errors, ε_i, are normally distributed.

This doesn't mean the data need to be normally distributed (why?). It means that the **residuals** at each x value should be normally distributed.

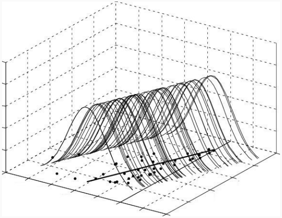


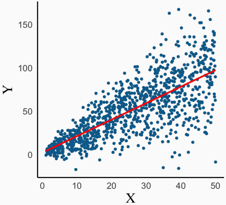
Source: Zuur et al. (2009)

In practice, we usually don't have many repeat measures of a specific x value, so checking for this usually means pooling all of the residuals and checking for normality. Normality of pooled residuals is reassuring, but does not necessarily mean the population is normally distributed.

iii) Heterogeneity

Heterogeneity is related to the assumption of normality. We just saw that the residuals at each x value should be normally distributed, but they also need to be drawn from the same distribution.





What will heterogeneity do to your estimates?

iv) Fixed x

This assumption means you are assuming there is no stochasticity around your x values (i.e., x is known exactly and entirely deterministic).

If you have defined the exact values at which x and y are measured, and there is no measurement error, this assumption is perfectly fine.

Situations where x is accompanied by a meaningful amount measurement error can break this assumption.

v) Independence

The assumption of independence is perhaps the most important assumption made by simple linear regression. Serial dependence can enter into your data in a number of ways, but the impact is typically the same: you over estimate the amount of information contained in a dataset.

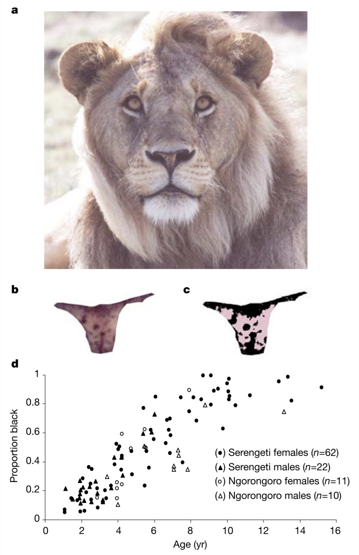
For example if I'm sitting in my back yard counting the number of birds, and I see a crow at 8:30:31, and then again at 8:30:32, and then again at 8:30:33, do I really have three unique pieces of information?

The standard deviation is given by: $\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$, what effect does breaking the assumption of independence have?

Linear regression example

Example: Lion noses

The Problem: In lion populations, the sustainable application of trophy hunting is often used as a way of maintaining stable populations while generating valuable funds to support conservation efforts. If you hunt older lions that are past their reproductive prime, the impact on the population is negligible, but if you hunt lions that are too young, there is a risk of the population destabilising. Whitman et al. (2004) looked at whether there was a relationship between how black a male lion's nose was and its age.



Lion noses: the data

proportionBlack	Age
0.21	1.1
0.14	1.5
0.11	1.9
0.13	2.2
0.12	2.6
0.13	3.2
0.12	3.2
0.18	2.9
0.23	2.4
0.22	2.1
0.2	1.9
0.17	1.9
0.15	1.9
0.27	1.9
0.26	2.8
0.21	3.6
0.3	4.3
0.42	3.8
0.43	4.2
0.59	5.4
0.6	5.8
0.72	6
0.29	3.4
0.1	4
0.48	7.3
0.44	7.3
0.34	7.8
0.37	7.1
0.34	7.1
0.74	13.1
0.79	8.8
0.51	5.4

The regression problem in matrix notation is:

$$\begin{pmatrix} 0.21 \\ 0.14 \\ \vdots \\ 0.51 \end{pmatrix} = \begin{pmatrix} 1 & 1.1 \\ 1 & 1.5 \\ \vdots & \vdots \\ 1 & 5.4 \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Lion noses: Estimating the parameters

Calculating our parameters using $\beta = (\mathrm{x}^{\mathsf{T}}\mathrm{x})^{-1}\mathrm{x}^{\mathsf{T}}\mathrm{y}$ can easily done in R:

```
data <- read.csv("LionNoses.csv")

x <- matrix(c(rep(1, nrow(data)),data$ageInYears),
            nrow = nrow(data), ncol = 2)
y <- matrix(data$proportionBlack,
            nrow = nrow(data), ncol = 1)

xtx <- t(x) %*% x
xtx.inv <- solve(xtx)
xty <- t(x) %*% y

beta <- xtx.inv %*% xty
```

which gives us $\beta_0 = 0.06969626$ and $\beta_1 = 0.05859115$

Lion noses: Estimating the parameters

We can also do this the easy way by using the `lm()` function:

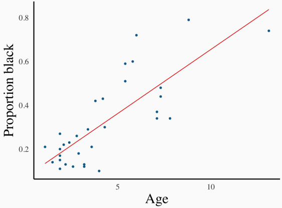
```
lm(proportionBlack ~ ageInYears, data = data)
```

```
Call:
lm(formula = proportionBlack ~ ageInYears, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.20406 -0.07758 -0.01750  0.07913  0.29876

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept) 0.069696   0.041956   1.661    0.107
ageInYears  0.058591   0.008307   7.053 7.68e-08 ***
---

Residual standard error: 0.1238 on 30 degrees of freedom
Multiple R-squared:  0.6238,  Adjusted R-squared:  0.6113
F-statistic: 49.75 on 1 and 30 DF,  p-value: 7.677e-08
```



Least Squares and Stochasticity

The least squares method provides a path for parametrising a model's deterministic component, but without any statement about the stochasticity of the system.

To solve this issue, we need to approach the problem as probalists and assume that each error term ε_i comes from some distribution ϕ.

We'll continue along this train of thought next lecture.