

Regression Models

By the end of this chapter you should:

- Be able to perform linear regression ~~or analysis of variance~~ using MATLAB and correctly interpret the output.
- Be able to perform appropriate diagnostic checks.
- Be able to perform appropriate inference for model parameters.

variables taking numerical values

The simplest form of association to analyse between two quantitative variables is a linear relationship.

Recall: A variable y is a linear function of x if

$$y = mx + c$$

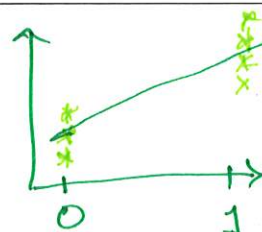
gradient/slope (pointing to m) *y-intercept.* (pointing to c)

We want to model the relationship between a single random variable Y called the response variable and an explanatory (as called *predictor*) variable x . Our inferences will be based on sample data $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$, where the y_1, \dots, y_n are realisations of independent random variables Y_1, \dots, Y_n however, they do not all have the same distribution. The distribution of Y_i will depend on the explanatory variable x_i .

Connection to two sample inference for means:

$(y_1, 0), (y_2, 0), \dots, (y_m, 0)$ $(y_{m+1}, 1), (y_{m+2}, 1), \dots, (y_{m+n}, 1)$

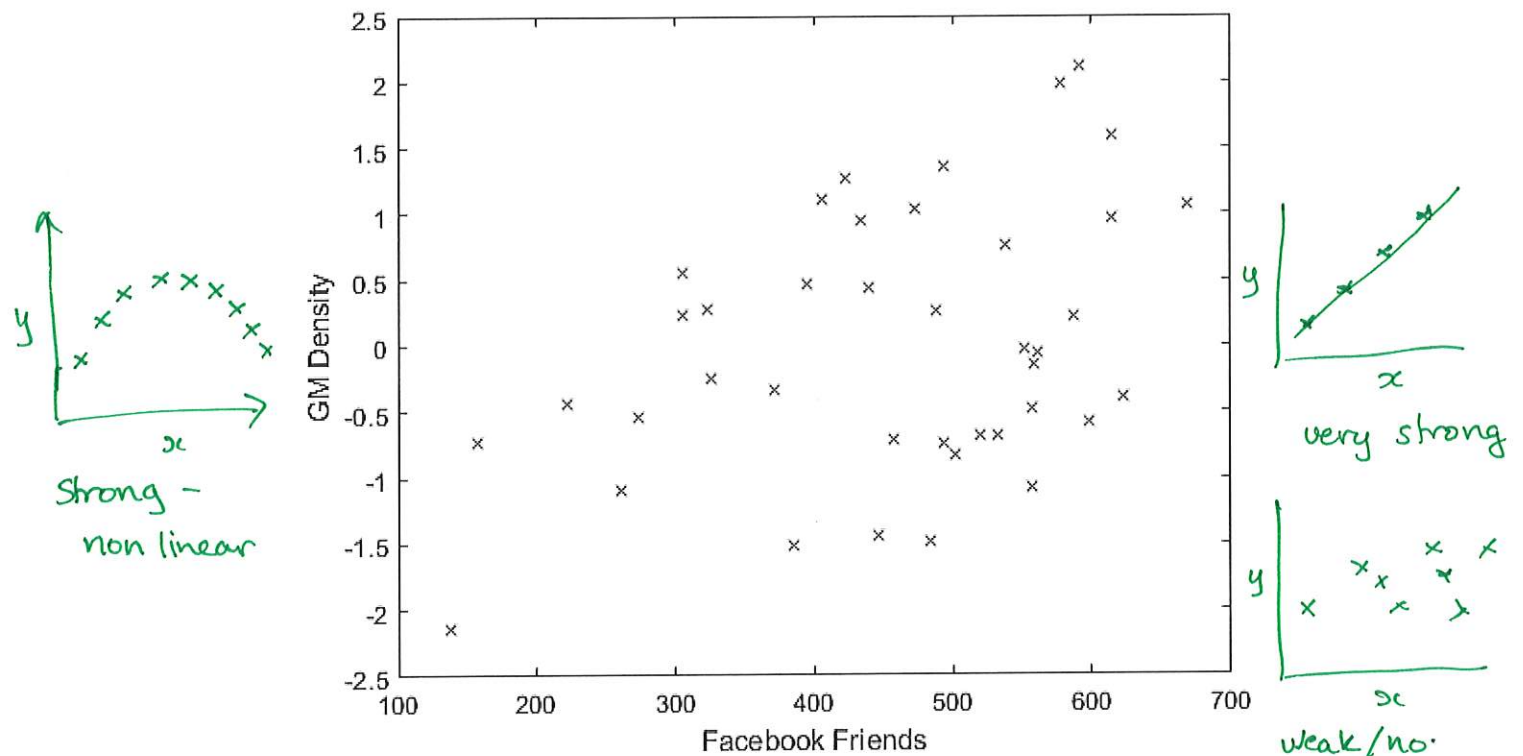
sample size m from population 1 *sample of size n from population 2*



Example. Researchers at University College London took a sample of 40 student volunteers and used MRI to measure grey matter (GM) density within three small volumes of the brain. They then looked at the association between these densities and the number of Facebook friends.

The results for one of the volumes, the left middle temporal gyrus, are shown in the following scatterplot. The GM densities are given in standard units.

```
1 facebook = readtable('facebook.xlsx');
2 plot (facebook.Facebook, facebook.GMDensity, 'x')
3 xlabel('Facebook Friends')
4 ylabel('GM Density')
```



In this chapter we model the relationship between two variables as a trend in the mean of the response variable plus variability about that trend.

How would you describe the relationship between grey matter density and the number of Facebook friends?

- Direction **positive** (more friends \Rightarrow higher GM density)
- Linearity **no obvious departure from linearity**
- Strength **weak relationship? considerable variation about overall trend**

In the above examples there appears to be a linear trend in the mean of the response variable. In the case of the Hubble data, this is what we expect from the physics. In the case of the Facebook data, there is not theoretical reason why there should be a linear relationship between the number of facebook friends a person has and there grey matter density, but it does appear to describe the data well.

There are usual two main objectives to regression analysis:

- Prediction — Given a new value of the explanatory variable, we would like to predict the value of the response variable with the smallest possible error.
- Explanation — We would like to describe the relationship between the response variable and the explanatory variable.

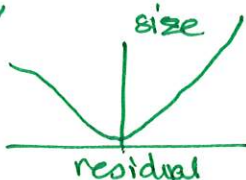
Supposing we have determined that a straight-line relationship is appropriate, how can we fit a *line of best fit* to the data? In other words, given a line, $b_0 + b_1x$, how can we judge how well it fits the data?

residuals = observation - (value of line)

minimise the "size of all residuals"

(1) absolute value of residual

* (2) squared residuals



A widely used measure of fit, and the one we will use in this course, is the sum of squared errors (or sum of squared deviations):

$$\text{residual}_i = y_i - (b_0 + b_1 x_i)$$

mathcal{S}

measure of fit

$$\rightarrow S(b) = \sum_{i=1}^n (y_i - [b_0 + b_1 x_i])^2$$

The line that minimises this is called the *least-squares line*.

We can use calculus to find the values of b_0 and b_1 which minimises $S(b)$. First, we differentiate $S(b)$ with respect to b_0 and b_1

$$\begin{aligned} \frac{\partial S(b)}{\partial b_0} &= \frac{\partial}{\partial b_0} \sum_{i=1}^n (y_i - [b_0 + b_1 x_i])^2 = \sum_{i=1}^n \frac{\partial}{\partial b_0} (y_i - [b_0 + b_1 x_i])^2 \\ &= -2 \sum_{i=1}^n (y_i - [b_0 + b_1 x_i]) = -2 \sum_{i=1}^n y_i + 2nb_0 + 2b_1 \sum_{i=1}^n x_i \\ \frac{\partial S(b)}{\partial b_1} &= \frac{\partial}{\partial b_1} \sum_{i=1}^n (y_i - [b_0 + b_1 x_i])^2 = \sum_{i=1}^n \frac{\partial}{\partial b_1} (y_i - [b_0 + b_1 x_i])^2 \\ &= -2 \sum_{i=1}^n x_i (y_i - [b_0 + b_1 x_i]) \\ &= -2 \sum_{i=1}^n x_i y_i + 2b_0 \sum_{i=1}^n x_i + 2b_1 \sum_{i=1}^n x_i^2 \end{aligned}$$

Setting these derivatives equal to zero leads to the following system of equations

$$\begin{aligned} nb_0 + \left(\sum_{i=1}^n x_i \right) b_1 &= \sum_{i=1}^n y_i \\ \left(\sum_{i=1}^n x_i \right) b_0 + \left(\sum_{i=1}^n x_i^2 \right) b_1 &= \sum_{i=1}^n x_i y_i \end{aligned}$$

Provided not all the x_i are the same, this system of equations has a unique solution given by

$$\begin{aligned} b_0 &= \frac{(\sum_{i=1}^n y_i)(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n x_i y_i)}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2} \\ b_1 &= \frac{n(\sum_{i=1}^n x_i y_i) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2} \end{aligned}$$

We (almost) never find the least squares line by hand using these equations. The equations determining the least-squares line are more easily presented in matrix form. Define

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

The matrix \mathbf{X} is called the *design matrix*. Then

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \quad \text{and} \quad \mathbf{X}^T \mathbf{y} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

We can now write our system of equations in matrix form;

$$\mathbf{X}^T \mathbf{X} \mathbf{b} = \mathbf{X}^T \mathbf{y} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

non-singular matrix

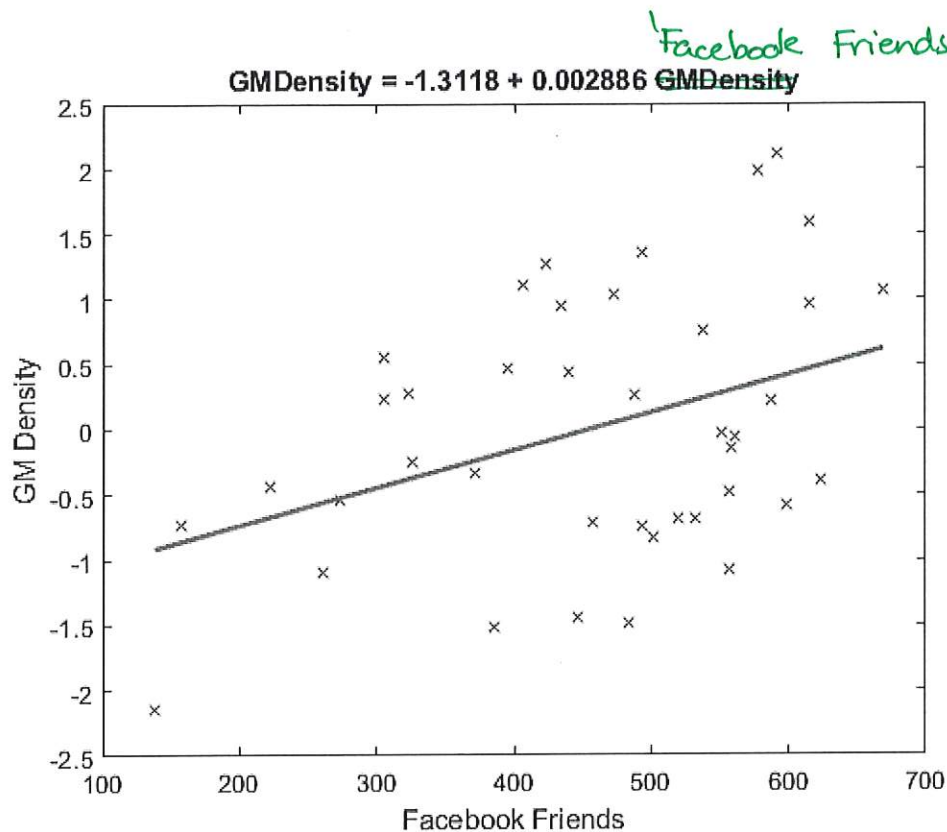
whose solution is

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

Example. Returning to the facebook example, we can use the `fitlm` function in MATLAB to determine the least-squares line. The least-squares line is plotted with the data in the figure below.

```
1 fitlm(facebook, 'Facebook~GMDensity')
2 facebooklm.Coefficients.Estimate
3
4 ans =
5
6     -1.3118
7         0.0029
```

`facebooklm = fitlm (facebook, 'GMDensity ~ Facebook')`



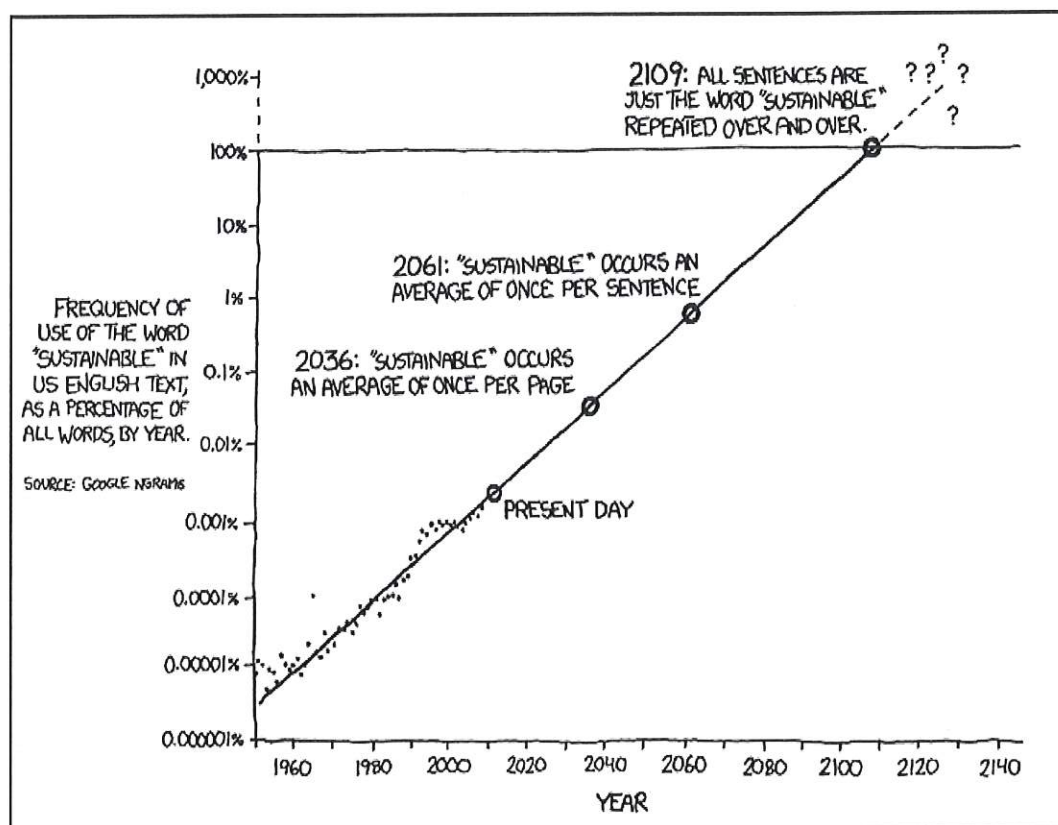
The least-squares line is

$$-1.311785 + 0.002886 \times \text{Facebook Friends}.$$

How do we interpret the slope and intercept of this line?

- Slope: One additional facebook friend is associated with an increase of 0.0029 (unit) in GM density.
- Intercept: for some with ~~0~~ no facebook friends we expect a GM density of -1.31 (units).

Sensible interpretation of the intercept depends on the context. Extrapolating beyond the range of the data is generally to be avoided.



THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

xkcd.com/1007/

This procedure allows us to fit a straight-line to our data. However, we know that our data is just the realisation of some random process and so we can think of our estimated line as being the realisation of some random process as well. So, in the facebook example, if we collected data from a different group of 40 students, would we have obtained exactly the same least-squares line?

We need to understand the variation in our fitted straight-line in order to be able to answer questions like:

- How much variability is there in the estimate of the slope?
- Is the evidence that the slope is non-zero?
- Construct confidence interval for the mean response at a given value of the explanatory variable.

To answer these questions we need a model for the process generating our data.