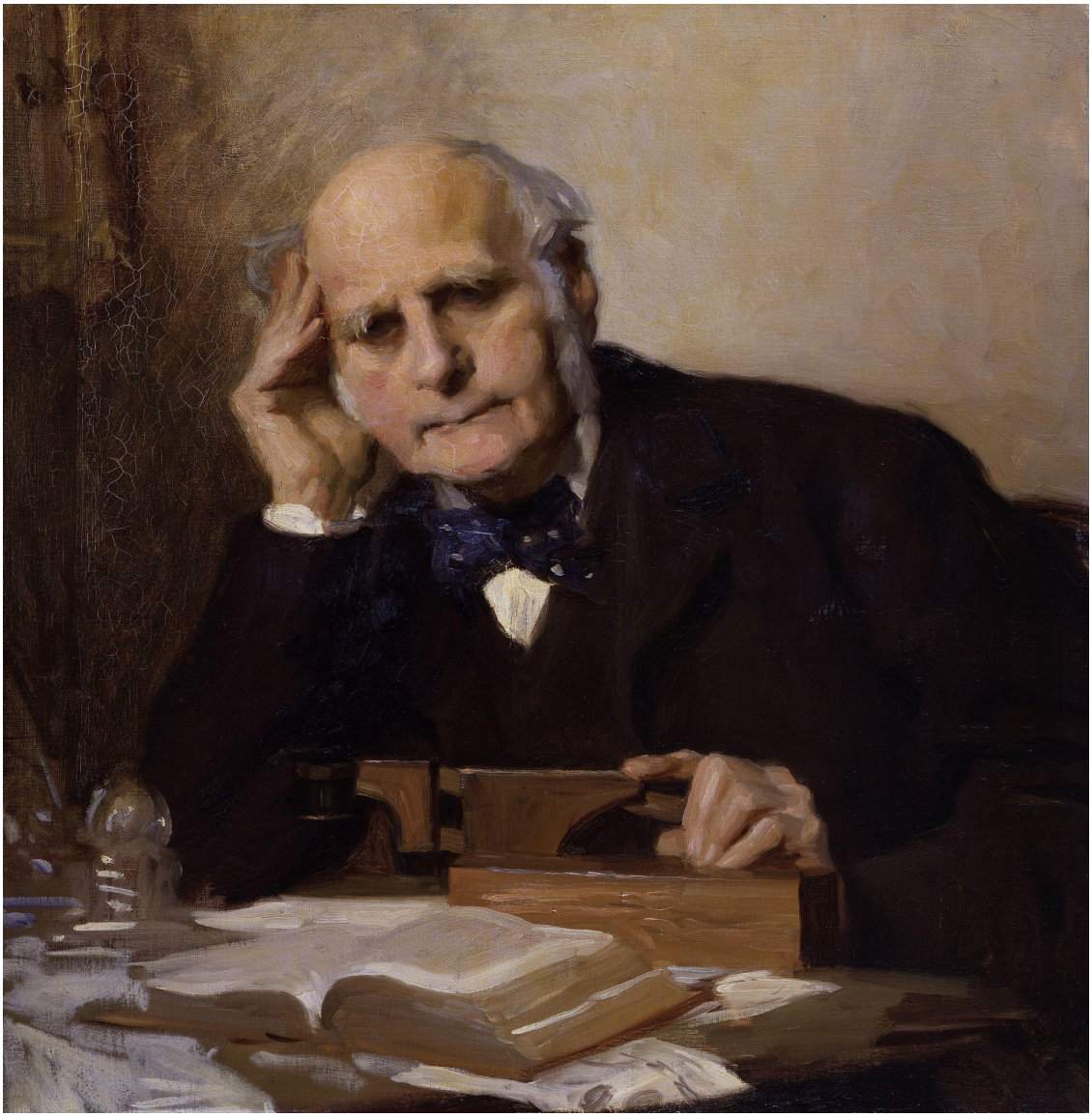
Linear Models



```
<IPython.core.display.HTML object>
```

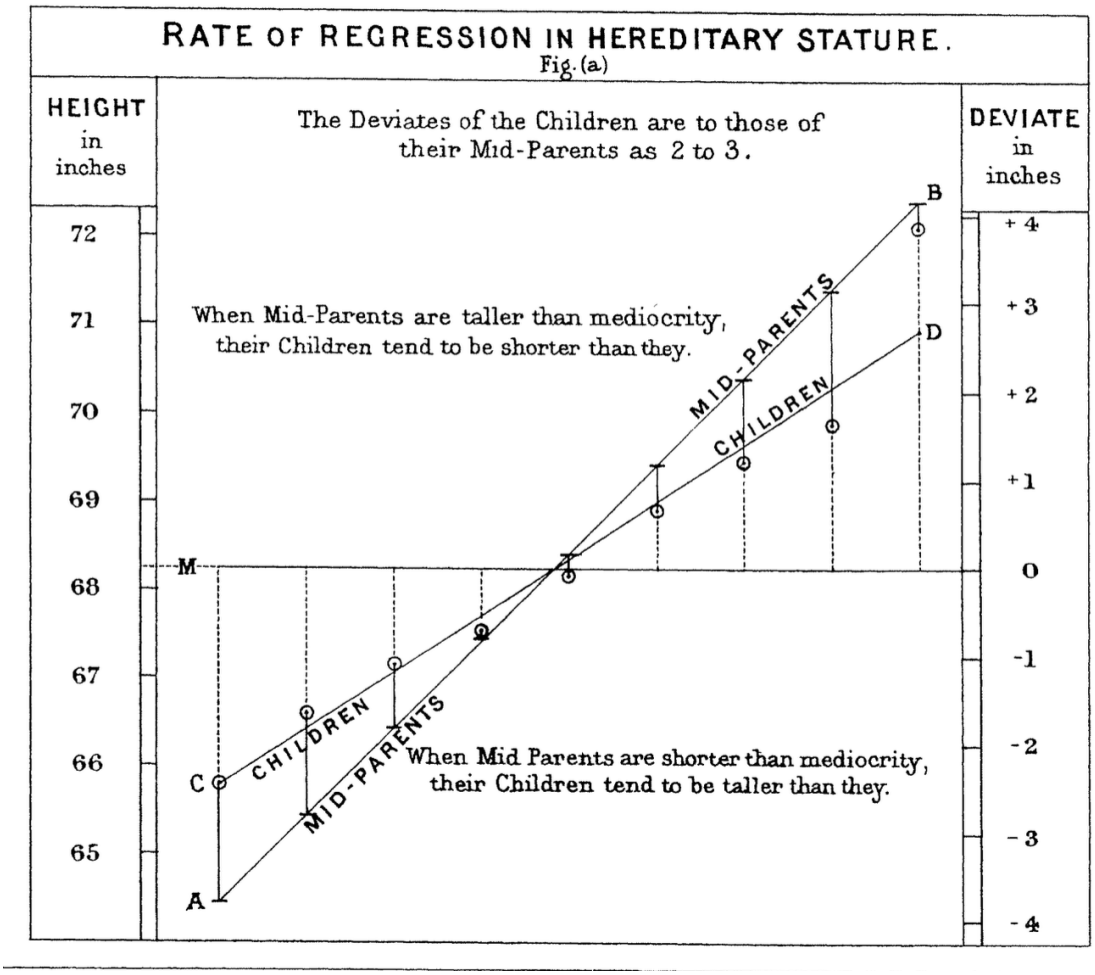
ANTHROPOLOGICAL MISCELLANEA.

Regression *towards* Mediocrity *in* Hereditary Stature.

By Francis Galton, F.R.S., &c.

In 1886 Francis Galton published his observations about how random factors affect outliers.

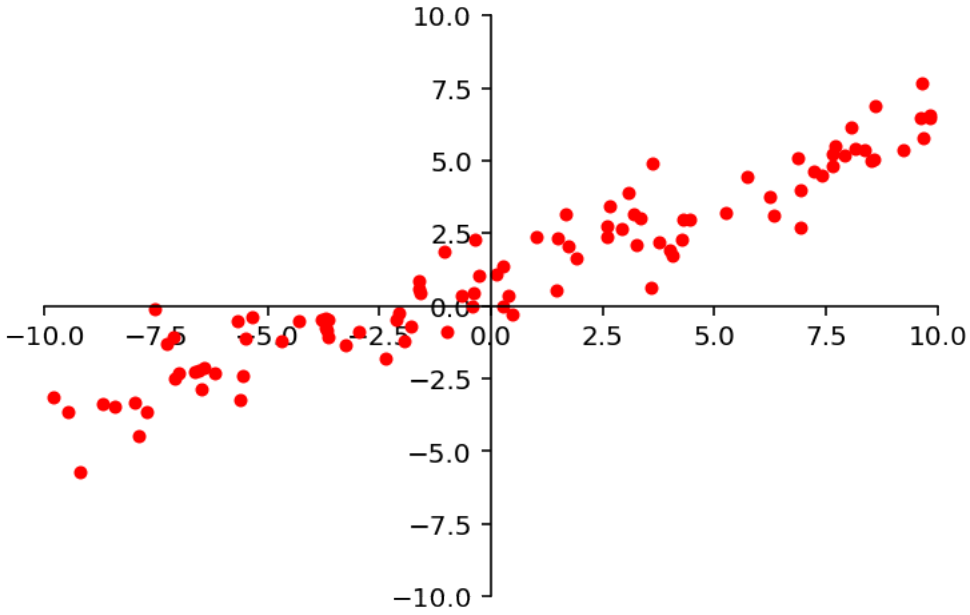
This notion has come to be called "regression to the mean" because unusually large or small phenomena, after the influence of random events, become closer to their mean values (less extreme).



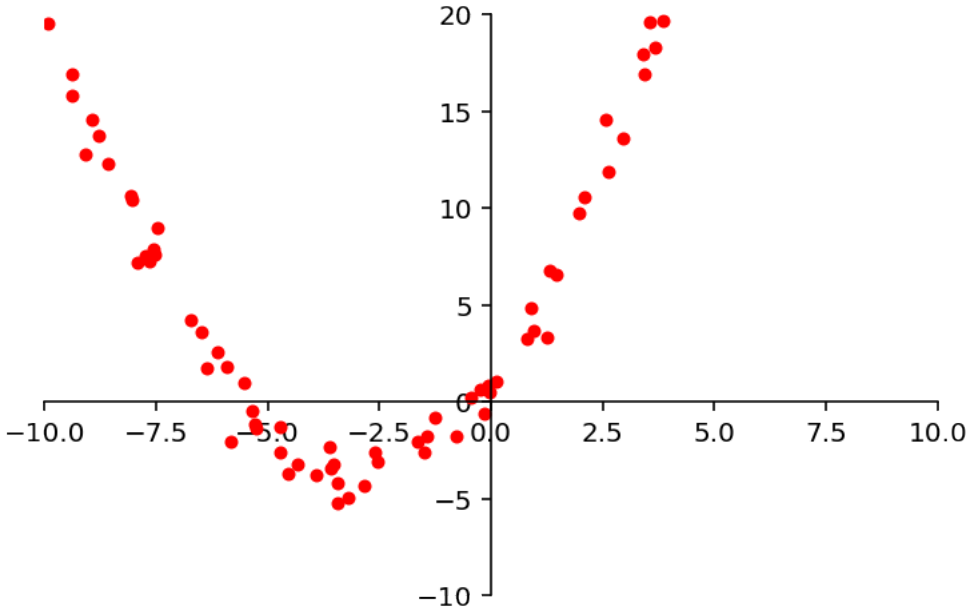
One of the most fundamental kinds of machine learning is the construction of a model that can be used to summarize a set of data.

The most common form of modeling is **regression**, which means constructing an equation that describes the relationships among variables.

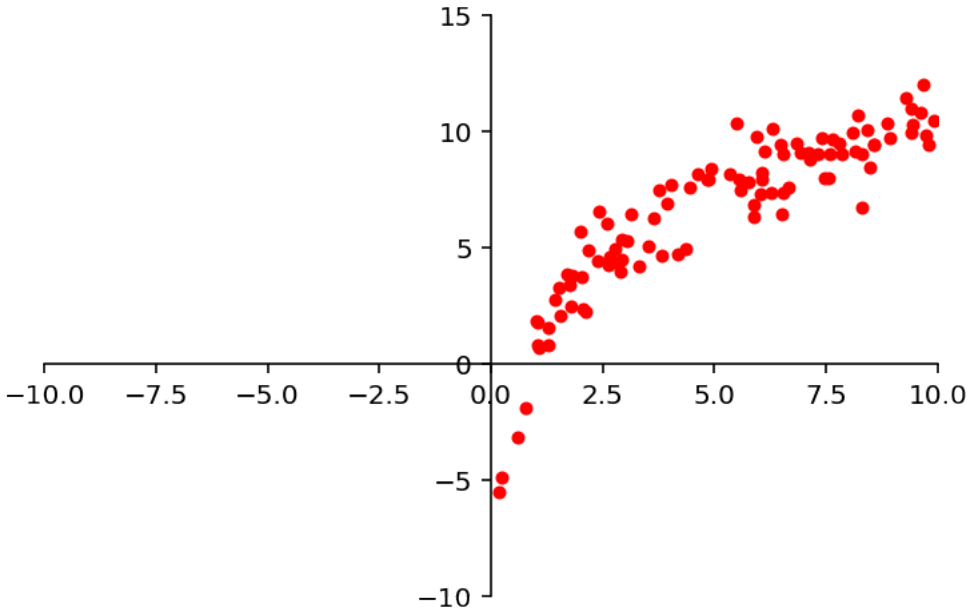
For example, we may look at these points and decide to model them using a line.



We may look at these points and decide to model them using a quadratic function.



And we may look at these points and decide to model them using a logarithmic function.

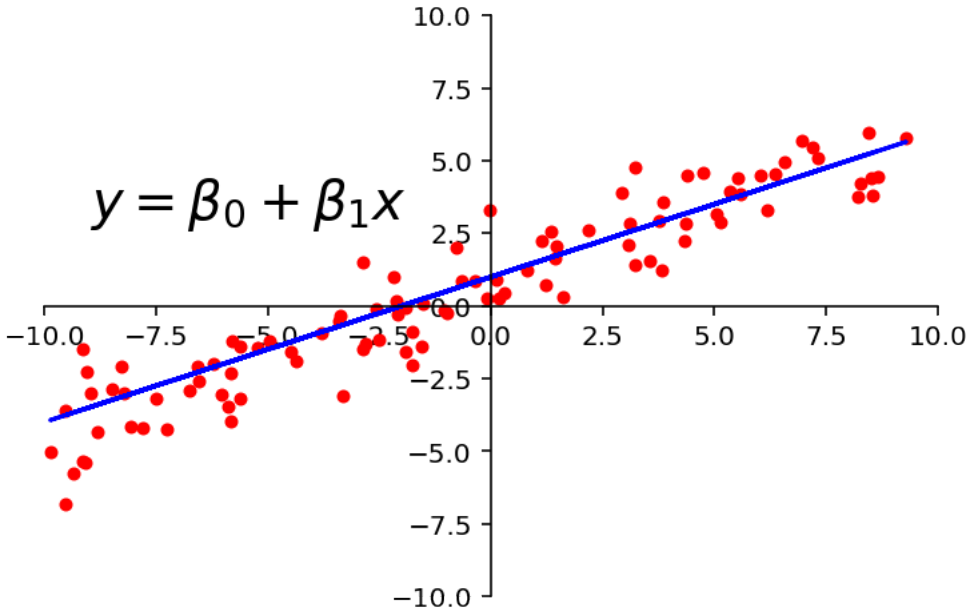


Clearly, none of these datasets agrees perfectly with the proposed model. So the question arises:
How do we find the **best** linear function (or quadratic function, or logarithmic function) given the data?
Framework.
This problem has been studied extensively in the field of statistics. Certain terminology is used:

- Some values are referred to as "independent," and
- Some values are referred to as "dependent."

The basic regression task is: * given a set of independent variables * and the associated dependent variables, * estimate the parameters of a model (such as a line, parabola, etc) that describes how the dependent variables are related to the independent variables.

The dependent variables are collected into a matrix X, which is called the **design matrix.**
The independent variables are collected into an **observation** vector **y**.
The parameters of the model (for any kind of model) are collected into a **parameter** vector β.



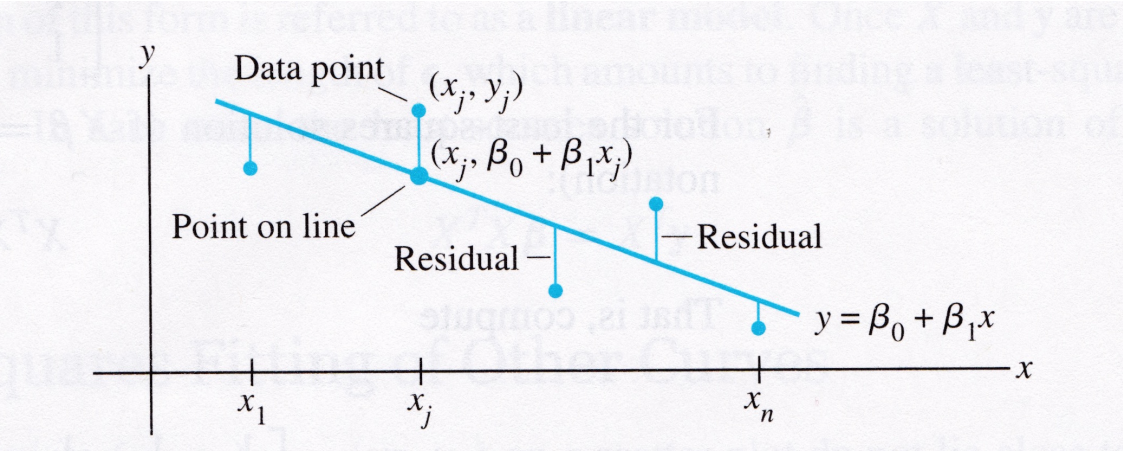
Least-Squares Lines

The first kind of model we'll study is a linear equation, $y = \beta_0 + \beta_1 x$.

Experimental data often produce points $(x_1, y_1), \ldots, (x_n, y_n)$ that seem to lie close to a line.

We want to determine the parameters β_0, β_1 that define a line that is as "close" to the points as possible.

Suppose we have a line $y = \beta_0 + \beta_1 x$. For each data point (x_j, y_j), there is a point $(x_j, \beta_0 + \beta_1 x_j)$ that is the point on the line with the same x-coordinate.



We call y_j the **observed** value of y and $\beta_0 + \beta_1 x_j$ the **predicted** y-value.

The difference between an observed y-value and a predicted y-value is called a **residual**.

There are several ways of measure how "close" the line is to the data.

The usual choice is to sum the squares of the residuals.

The **least-squares line** is the line $y = \beta_0 + \beta_1 x$ that minimizes the sum of squares of the residuals.

The coefficients β_0, β_1 of the line are called **regression coefficients.**

A least-squares problem.

If the data points were on the line, the parameters β_0 and β_1 would satisfy the equations

$$\begin{aligned}\beta_0 + \beta_1 x_1 &= y_1\\ \beta_0 + \beta_1 x_2 &= y_2\\ \beta_0 + \beta_1 x_3 &= y_3\\ &\vdots\\ \beta_0 + \beta_1 x_n &= y_n\end{aligned}$$

We can write this system as

$$X\mathbf{fi} = \mathbf{y}, \quad \text{where } X = \begin{bmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_n\end{bmatrix}, \ \mathbf{fi} = \begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}, \ \mathbf{y} = \begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}$$

Of course, if the data points don't actually lie exactly on a line,

... then there are no parameters β_0, β_1 for which the predicted y-values in $X\mathbf{fi}$ equal the observed y-values in $\mathbf{y}$,

... and $X\mathbf{fi} = \mathbf{y}$ has no solution.

Note that we are seeking the β that minimizes the sum of squared residuals, ie,

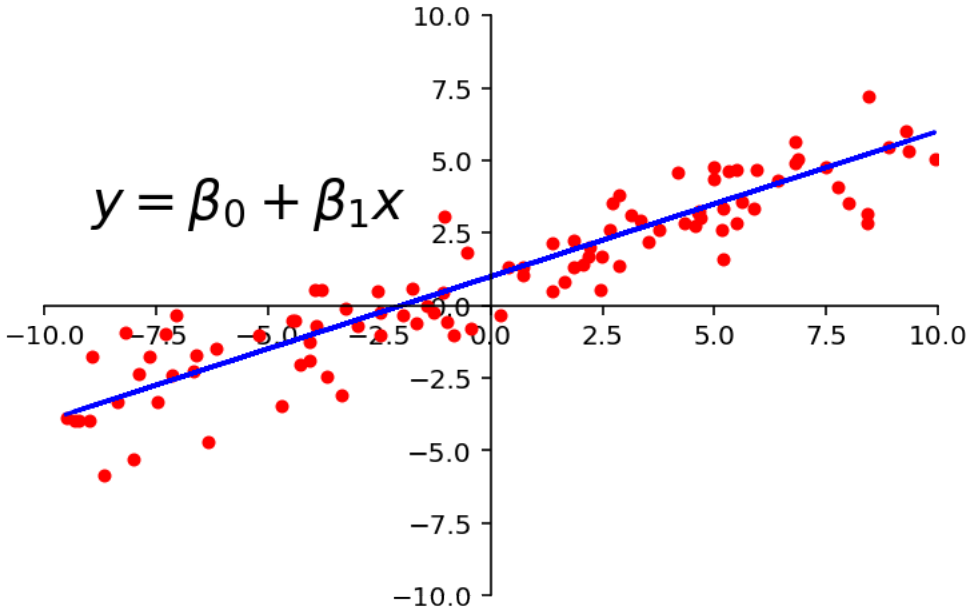
$$\sum_i (\beta_0 + \beta_1 x_i - y_i)^2$$

$$= \|X\beta - \mathbf{y}\|^2$$

This is key: **the sum of squares of the residuals** is **exactly** the **square of the distance between the vectors** $X\mathbf{fi}$ **and** $\mathbf{y}$.

This is a least-squares problem, $A\mathbf{x} = \mathbf{b}$, with different notation.

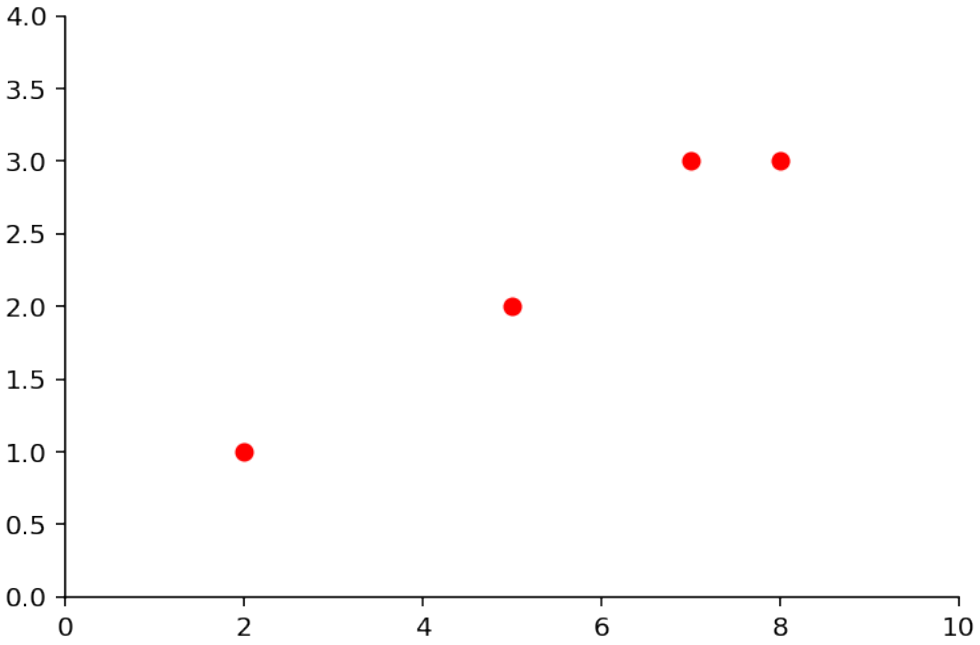
Computing the least-squares solution of $X\mathbf{fi} = \mathbf{y}$ *is equivalent to finding the* **fi** *that determines the least-squares line.*



Question Time! Q23.1

Example 1. Find the equation $y = \beta_0 + \beta_1 x$ of the least-squares line that best fits the data points (2,1), (5,2), (7,3), and (8,3).

```
[12]: ax = ut.plotSetup(0, 10, 0, 4)
      ut.centerAxes(ax)
      pts = np.array([[2,1], [5,2], [7,3], [8,3]]).T
      ax.plot(pts[0], pts[1], 'ro');
```



Solution. Use the x-coordinates of the data to build the design matrix X, and the y-coordinates to build the observation vector $\mathbf{y}$:

$$X = \begin{bmatrix} 1 & 2 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix}$$

Now, to obtain the least-squares line, find the least-squares solution to $X\mathbf{fi} = \mathbf{y}$.
We do this via the method we learned last lecture (just with new notation):

$$X^T X\mathbf{fi} = X^T\mathbf{y}$$

So, we compute:

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix}$$

$$X^T\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 9 \\ 57 \end{bmatrix}$$

So the normal equations are:

$$\begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 9 \\ 57 \end{bmatrix}$$

Solving, we get:

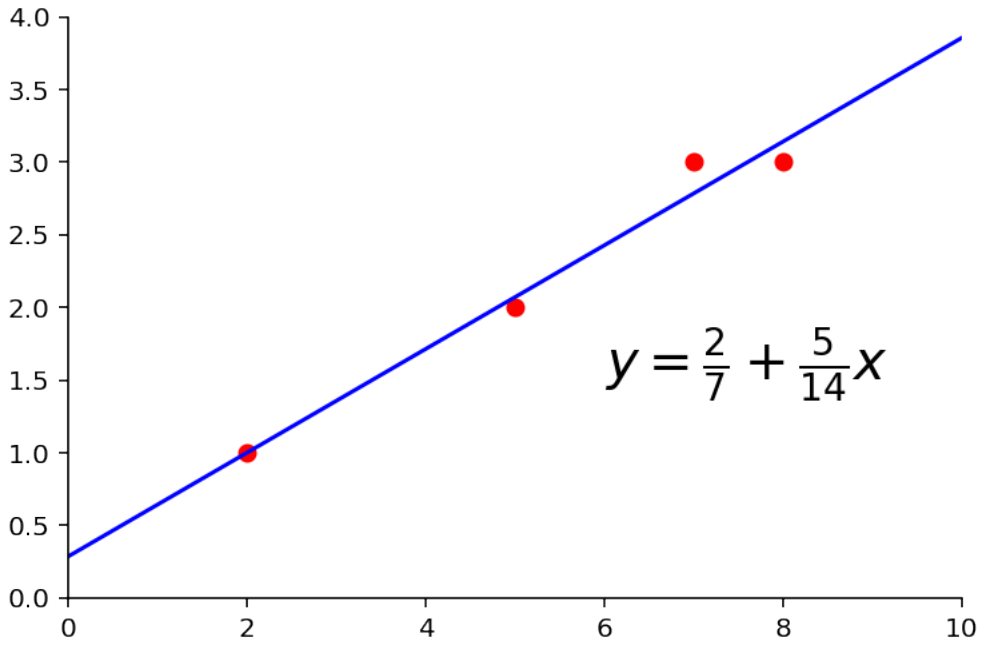
$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix}^{-1} \begin{bmatrix} 9 \\ 57 \end{bmatrix}$$

$$= \frac{1}{84} \begin{bmatrix} 142 & -22 \\ -22 & 4 \end{bmatrix} \begin{bmatrix} 9 \\ 57 \end{bmatrix}$$

$$= \begin{bmatrix} 2/7 \\ 5/14 \end{bmatrix}$$

So the least-squares line has the equation

$$y = \frac{2}{7} + \frac{5}{14}x.$$



The General Linear Model

Another way that the inconsistent linear system is often written is to collect all the residuals into a **residual vector.**

Then an exact equation is

$$y = X\mathbf{fi} + \mathbf{ffl}$$

Any equation of this form is referred to as a **linear model.**

In this formulation, the goal is to minimize the length of ϵ, ie, $\|\epsilon\|$.

In some cases, one would like to fit data points with something other than a straight line.

For example, think of Gauss trying to find the equation for the orbit of Ceres.

In cases like this, the matrix equation is still $X\mathbf{fi} = \mathbf{y}$, but the specific form of X changes from one problem to the next.

The least-squares solution $\hat{\mathbf{fi}}$ is a solution of the normal equations

$$X^T X\mathbf{fi} = X^T\mathbf{y}.$$

Least-Squares Fitting of Other Models

Most models have parameters, and the objection of **model fitting** is to to fix those parameters. Let's talk about model parameters.

In model fitting, the parameters are the unknown. A central question for us is whether the model is *linear* in its parameters.

For example, the model $y = \beta_0 e^{-\beta_1 x}$ is **not** linear in its parameters. The model $y = \beta_0 e^{-2x}$ **is** linear in its parameters.

For a model that is linear in its parameters, an observation is a linear combination of (arbitrary) known functions.

In other words, a model that is linear in its parameters is

$$y = \beta_0 f_0(x) + \beta_1 f_1(x) + \cdots + \beta_n f_n(x)$$

where $f_0, \ldots, f_n$ are known functions and $\beta_0, \ldots, \beta_k$ are parameters.

Example. Suppose data points $(x_1, y_1), \ldots, (x_n, y_n)$ appear to lie along some sort of parabola instead of a straight line. Suppose we wish to approximate the data by an equation of the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2.$$

Describe the linear model that produces a "least squares fit" of the data by the equation.

Solution. The ideal relationship is $y = \beta_0 + \beta_1 x + \beta_2 x^2$.

Suppose the actual values of the parameters are $\beta_0, \beta_1, \beta_2$. Then the coordinates of the first data point satisfy the equation

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon_1$$

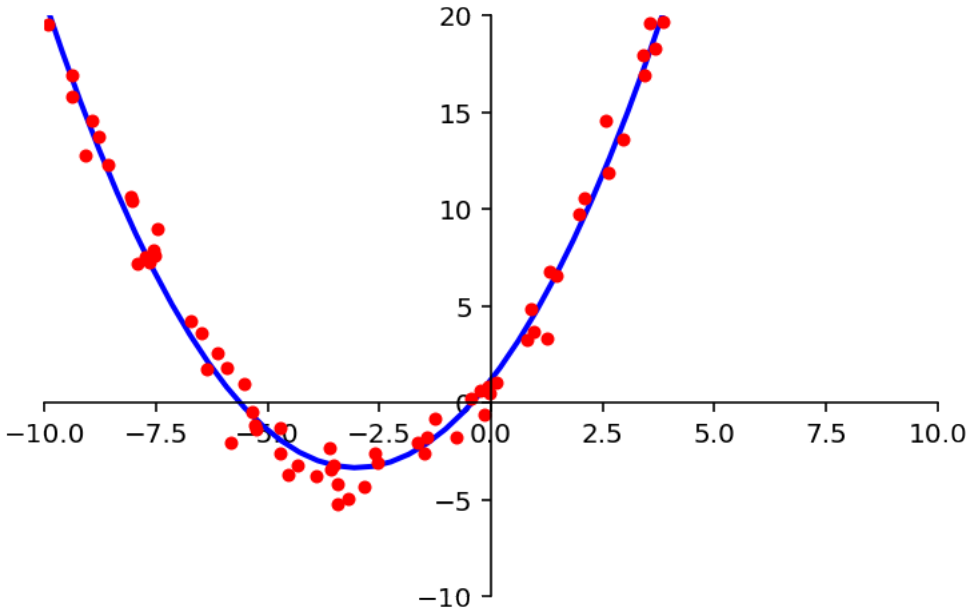
where ϵ_1 is the residual error between the observed value y_1 and the predicted y-value.

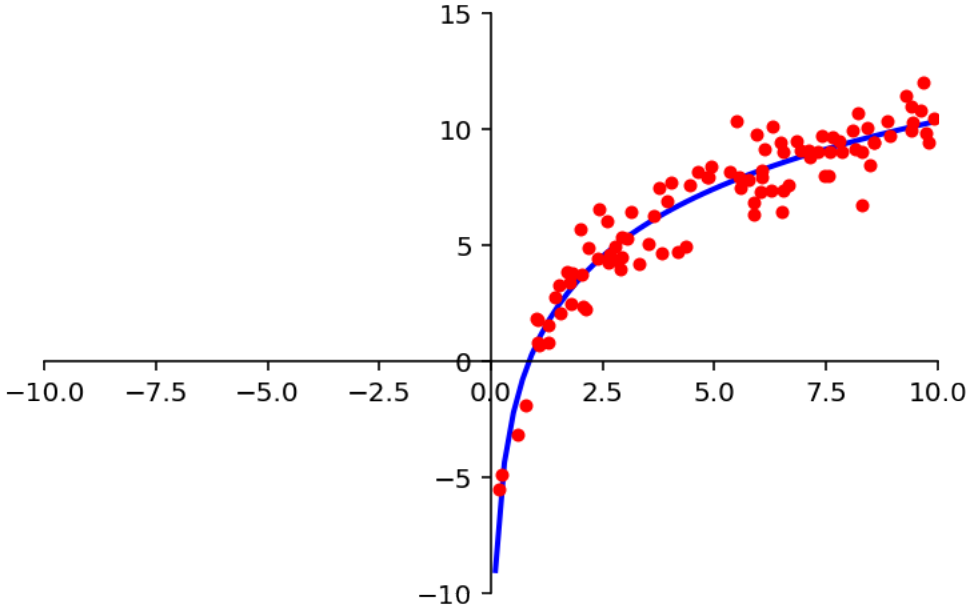
Each data point determines a similar equation:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \beta_2 x_n^2 + \epsilon_n \end{aligned}$$

Clearly, this system can be written as $\mathbf{y} = X\mathbf{fi} + \mathbf{ffl}$.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$





Question Time! Q23.2

Multiple Regression

Suppose an experiment involves two independent variables – say, u and v, – and one dependent variable, y. A simple equation for predicting y from u and v has the form

$$y = \beta_0 + \beta_1 u + \beta_2 v$$

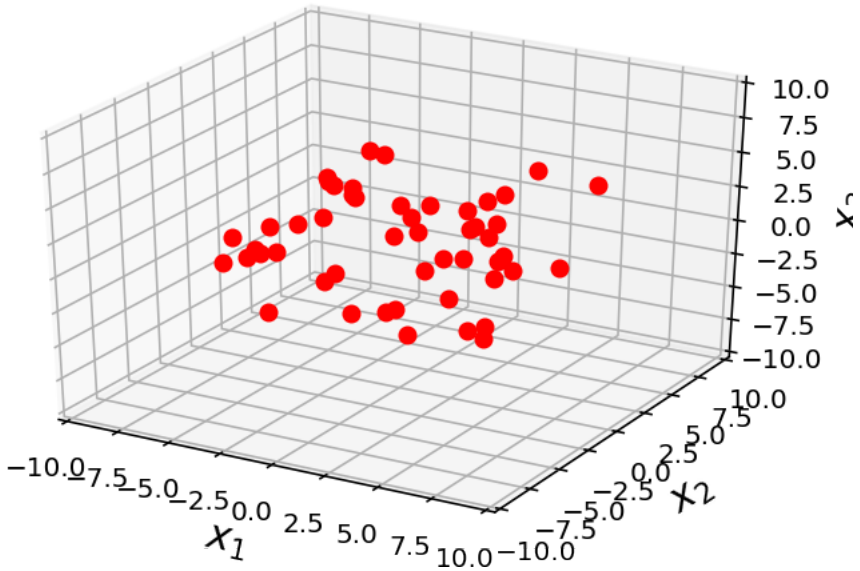
Since there is more than one independent variable, this is called **multiple regression.**
A more general prediction equation might have the form

$$y = \beta_0 + \beta_1 u + \beta_2 v + \beta_3 u^2 + \beta_4 uv + \beta_5 v^2$$

A least squares fit to equations like this is called a **trend surface.**
In general, a linear model will arise whenever y is to be predicted by an equation of the form

$$y = \beta_0 f_0(u, v) + \beta_1 f_1(u, v) + \cdots + \beta_k f_k(u, v)$$

with $f_0, \ldots, f_k$ any sort of known functions and $\beta_0, ..., \beta_k$ unknown weights.
Let's take an example. Here are a set of points in $\mathbb{R}^3$:



Example. In geography, local models of terrain are constructed from data $(u_1, v_1, y_1), \ldots, (u_n, v_n, y_n)$ where u_j, v_j, and y_j are latitude, longitude, and altitude, respectively.

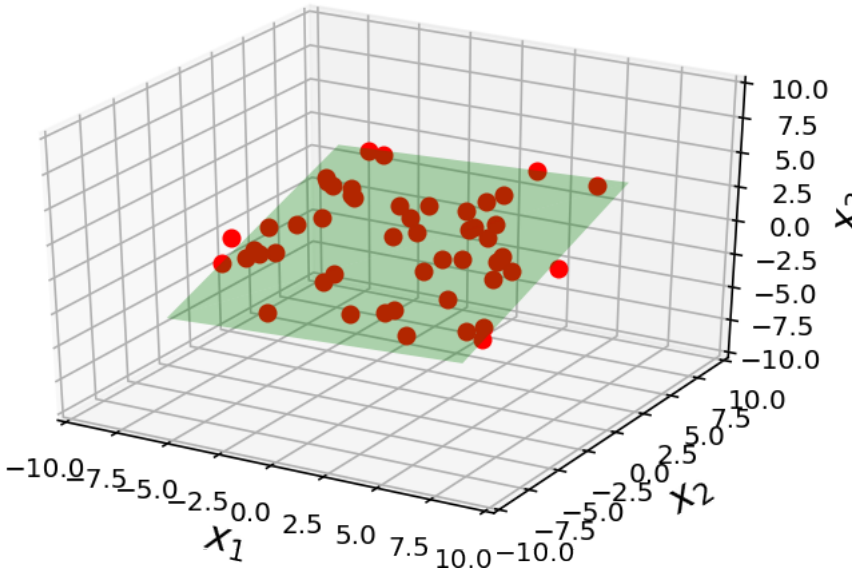
Let's describe the linear models that gives a least-squares fit to such data. The solution is called the least-squares *plane.*

Solution. We expect the data to satisfy these equations:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 u_1 + \beta_2 v_1 + \epsilon_1 \\ y_1 &= \beta_0 + \beta_1 u_2 + \beta_2 v_2 + \epsilon_2 \\ &\vdots \\ y_1 &= \beta_0 + \beta_1 u_n + \beta_2 v_n + \epsilon_n \end{aligned}$$

This system has the matrix for $\mathbf{y} = X\mathbf{fi} + \epsilon$, where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & u_1 & v_1 \\ 1 & u_2 & v_2 \\ \vdots & \vdots & \vdots \\ 1 & u_n & v_n \end{bmatrix}, \quad \mathbf{fi} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$



This example shows that the linear model for multiple regression has the same abstract form as the model for the simple regression in the earlier examples.

We can see that there the general principle is the same across all the different kinds of linear models.

Once X is defined properly, the normal equations for **fi** have the same matrix form, no matter how many variables are involved.

Thus, for any linear model where $X^T X$ is invertible, the least squares $\hat{\mathbf{fi}}$ is given by $(X^T X)^{-1} X^T \mathbf{y}$.