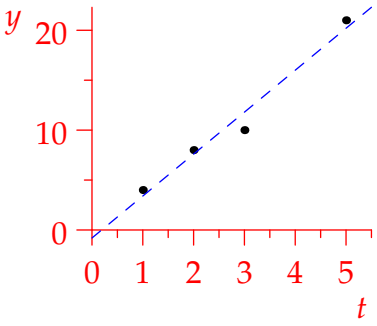
3 Regression and Best-fitting Polynomials

The simplicity of polynomials makes them well-suited to modelling problems. Straight lines in particular, are often used to approximate experimental data and infer relationships between variables.

Example 3.1. Suppose at time t hours in the afternoon, a hiker's GPS locator says that they've travelled y miles along a hiking trail;

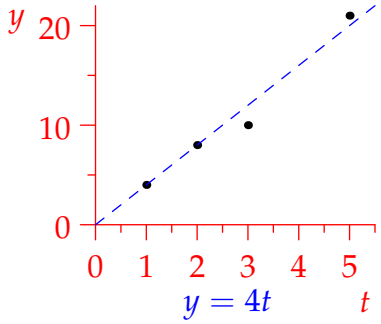
t_i	1	2	3	5
y_i	4	8	10	21

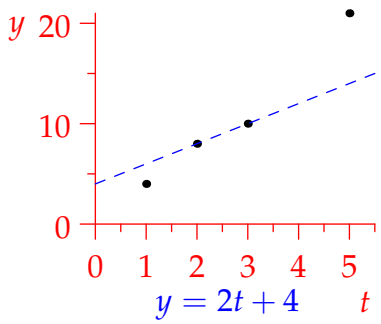


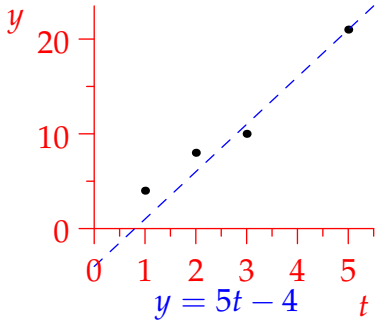
How far has the hiker traveled after 4 hours?

By plotting the points, the relationship looks to be approximately[1] linear: $y \approx mt + c$. What is the *best* choice of line, and how should we find the coefficients m, c?

Exercise Discuss what good criteria would be when choosing our line. What do we mean by *best*? Intuitively, we want the points to be as close to the line as possible, but what should *close* mean here? What would we like to use the approximating line to do? Here are three lines plotted with the data set: of the choices, which seems best and why?







In our problem, we want to use the line to *predict* the hiker's location: for a given t we want to predict $\hat{y} = mt + c$. It seems reasonable that we'd want our line to minimize *vertical* errors $\hat{y}_i - y_i$, since these are a measure of how bad our prediction is. Returning to the above exercise, we can compute these errors: since a positive error is as bad as a negative, we make all the errors positive

	t_i	1	2	3	5
	y_i	4	8	10	21
$y = 4t$	$\lvert\hat{y}_i - y_i\rvert$	0	0	2	1
$y = 2t + 4$	$\lvert\hat{y}_i - y_i\rvert$	2	0	0	7
$y = 5t - 4$	$\lvert\hat{y}_i - y_i\rvert$	3	2	1	0

It seems reasonable to claim that the first of these lines is better than the others. But can we do better; is there a *best* line? We might define *best* as meaning that the sum of the absolute errors is minimal $\sum |\hat{y}_i - y_i|$. This is certainly a reasonable approach. However, for reasons of computational simplicity, statistical interpretation, and to particularly discourage *large* errors, the standard approach is to minimize the sum of squares.

[1] There are at least *two* reasons why the data for the distance traveled by the hiker might not be perfectly linear; why?

Definition 3.2. Suppose we have a finite data set $\{(t_i, y_i)\}$. If $\hat{y} = mx + c$ is a linear predictor for y given x, then

- The i^{th} error is the difference $e_i = \hat{y}_i - y_i = mt_i + c - y_i$.
- The *regression line* or *best-fitting least-squares line* for the data is the function $\hat{y} = mt + c$ which minimizes the sum $\sum e_i^2$ of the *squares* of the errors.

To return to our example, suppose the predictor was $y = mt + c$. We expand the table

t_i	1	2	3	5
y_i	4	8	10	21
$\hat{y}_i$	$m + c$	$2m + c$	$3m + c$	$5m + c$
e_i	$m + c - 4$	$2m + c - 8$	$3m + c - 10$	$5m + c - 21$

Our goal is therefore to minimize the 'sum of squares' function

$$S(m, c) = \sum e_i^2 = (m + c - 4)^2 + (2m + c - 8)^2 + (3m + c - 10)^2 + (5m + c - 21)^2$$

This might look horrible, but it is fairly easy to deal with if we use a little calculus. At a minimum (m, c), moving either m or c should result in $S(m, c)$ getting larger; if we treat first c and then m as a constant, this says that the derivatives of S with respect to both m, c must be zero:[a]

- c constant, differentiate with respect to m;

$$\begin{aligned} S_m &= 2(m + c - 4) + 4(2m + c - 8) + 6(3m + c - 10) + 10(5m + c - 21) \\ &= 2\Big[(1 + 2^2 + 3^2 + 5^2)m + (1 + 2 + 3 + 5)c - 4 - 2 \cdot 8 - 3 \cdot 10 - 5 \cdot 21\Big] \\ &= 2\Big[39m + 11c - 155\Big] = 2\big[\big(\sum t_i^2\big)\, m + \big(\sum t_i\big)\, c - \sum t_i y_i\big] \end{aligned}$$

- m constant, differentiate with respect to c;

$$\begin{aligned} S_c &= (m + c - 4) + (2m + c - 8) + (3m + c - 10) + (5m + c - 21) \\ &= (1 + 2 + 3 + 5)m + (1 + 1 + 1 + 1)c - 4 - 8 - 10 - 21 \\ &= 11m + 4c - 43 = \big(\sum t_i\big)\, m + nc - \sum y_i \end{aligned}$$

Since both of these should be zero, we have a pair of simultaneous equations which can be solved in the usual manner

$$\begin{cases} 39m + 11c = 155 \\ 11m + 4c = 43 \end{cases} \implies m = \frac{21}{5}, c = -\frac{4}{5} \implies \hat{y} = \frac{1}{5}(21t - 4)$$

This is the line graphed in the original problem! We can also answer the original question, we expect the hiker to have traveled approximately $\hat{y} = \frac{1}{5}(21 \cdot 4 - 4) = 16$ miles after 4 hours.

The sum of the squared errors for our regression line is $\sum(\hat{y}_i - y_i)^2 = 4.4$; this compares to 5, 53 and 14 for our three options above.[b]

[a] This is really partial differentiation $\frac{\partial S}{\partial m} = 0 = \frac{\partial S}{\partial c}$. Fear not if you've never seen this, a formula is coming!

[b] Interestingly, the sum of the absolute errors $\sum |\hat{y}_i - y_i|$ is better ($3 < 3\frac{3}{5}$) for the line $y = 4t$. We are making a choice here as to what *best-fitting* means!

In general we have the following result.

Theorem 3.3 (Linear Regression). *Given a set of data pairs $\{(t_i, y_i) : i = 1, \ldots, n\}$ the best-fitting least-squares line has equation $\hat{y} = mt + c$ where m, c satisfy*

$$\begin{cases} \left(\sum t_i^2\right) m + \left(\sum t_i\right) c = \sum t_i y_i \\ \left(\sum t_i\right) m + nc = \sum y_i \end{cases} \iff \begin{cases} \overline{t^2} m + \bar{t} c = \overline{ty} \\ \bar{t} m + c = \bar{y} \end{cases}$$

This is a pair of simultaneous equations for the coefficients m, c, which can be solved either as a matrix problem or by substitution:

$$c = \bar{y} - m\bar{t}, \qquad m = \frac{\overline{ty} - \bar{t}\bar{y}}{\overline{t^2} - \bar{t}^2} = \frac{n\sum t_i y_i - \sum t_i \sum y_i}{n\sum t_i^2 - \left(\sum t_i\right)^2} = \frac{\sum (t_i - \bar{t})(y_i - \bar{y})}{\sum (t_i - \bar{t})^2}$$

where $\bar{t}, \bar{y}$ are the average *values of t_i, y_i, and $\overline{ty} = \frac{1}{n}\sum t_i y_i$, $\overline{t^2} = \frac{1}{n}\sum t_i^2$, etc.*

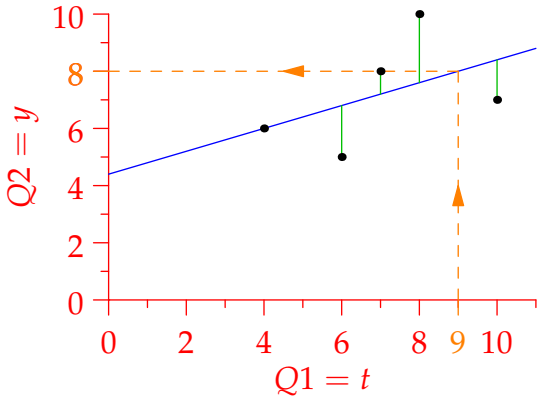
Example 3.4. Five students' scores on two quizzes were as follows:

Quiz 1	8	10	6	7	4
Quiz 2	10	7	5	8	6

1. If a student scores 9/10 on the first quiz, what might we expect them to score on the second?
2. If a student scores 8/10 on the second quiz, predict their score on the first.

We'll answers these two problems using two different approaches. For part 1, we compute using linear equations; the t data is Quiz 1 and y is Quiz 2, whence

$$\sum_{i=1}^{5} t_i^2 = 64 + 100 + 36 + 49 + 16 = 265$$
$$\sum_{i=1}^{5} t_i = 8 + 10 + 6 + 7 + 4 = 35$$
$$\sum_{i=1}^{5} y_i = 10 + 7 + 5 + 8 + 6 = 36$$
$$\sum_{i=1}^{5} t_i y_i = 80 + 70 + 30 + 56 + 24 = 260$$



$$\begin{cases} 265m + 35c = 260 \\ 35m + 5c = 36 \end{cases} \iff \begin{cases} 20m = 8 \\ c = \frac{36}{5} - 7m \end{cases} \iff (m, c) = \left(\frac{2}{5}, \frac{22}{5}\right)$$

Our regression line is therefore $\hat{y}(t) = \frac{2}{5}(t + 11)$: this is the line which minimizes the sum of the squares of the vertical deviations. The prediction given a Quiz 1 score of 9/10 is that the student will score $\hat{y}(9) = \frac{2}{5} \cdot 20 = 8$. Notice that the average score $\bar{y} = 7.2$ on Quiz 2 is *higher* than $\bar{t} = 7$ on Quiz 1, and yet the hypothetical student's score is expected to go down!

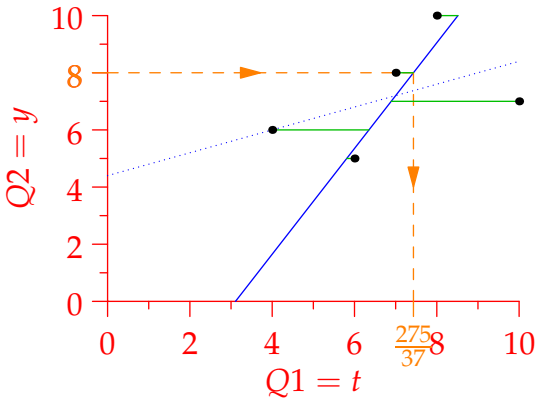
To solve the second problem, we need to swap the roles of our variables. This time we use the explicit formulæ (note that $y \leftrightarrow t$!):

$$\bar{t} = \frac{35}{5} = 7, \qquad \bar{y} = \frac{36}{5} = 7.2,$$

$t_i - \bar{t}$	1	3	−1	0	−3
$y_i - \bar{y}$	2.8	−0.2	−2.2	0.8	−1.2
$(y_i - \bar{y})^2$	7.84	0.04	4.84	0.64	1.44
$(y_i - \bar{y})(t_i - \bar{t})$	2.8	−0.6	2.2	0	3.6

$$m = \frac{\sum (y_i - \bar{y})(t_i - \bar{t})}{\sum (y_i - \bar{y})^2} = \frac{8}{14.8} = \frac{20}{37} \approx 0.541$$

$$c = \bar{t} - m\bar{y} = 7 - \frac{20 \cdot 36}{5 \cdot 37} = \frac{115}{37} \approx 3.108$$



The line for predicting Quiz 1 given Quiz 2 is $\hat{t} = \frac{1}{37}(20y + 115)$, whence $\hat{t}(8) = \frac{275}{37} \approx 7.432$. Notice how this minimizes the sum of the squares of the horizontal deviations and is therefore *different* to the line in the previous question (dotted in the picture).

Interpretation The regression line passes through the *center of mass* $(\bar{t}, \bar{y})$ of the data. If you've studied statistics, you might have seen the formula for m written in terms of the *variance* and *covariance*

$$\operatorname{Var} t = \frac{1}{n}\sum_{i=1}^{n}(t_i - \bar{t})^2, \qquad \operatorname{Cov}(t, y) = \frac{1}{n}\sum_{i=1}^{n}(t_i - \bar{t})(y_i - \bar{y})$$

Variance quantifies how much the values t_i deviate[2] from their mean $\bar{t}$. Covariance essentially measures to what extent t and y deviate from their means *in the same/opposite directions*: $\operatorname{Cov}(t, y) > 0$ means that when $t > \bar{t}$ we expect $y > \bar{y}$; in such a case the regression line has positive slope.

A further piece of interpretation comes from consideration of the *coefficient of determination*[3]

$$R^2 := 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = m^2 \frac{\sum (t_i - \bar{t})^2}{\sum (y_i - \bar{y})^2} = m^2 \frac{\overline{t^2} - \bar{t}^2}{\overline{y^2} - \bar{y}^2} = m^2 \frac{\operatorname{Var} t}{\operatorname{Var} y} = \frac{\operatorname{Cov}(y, t)}{\sqrt{\operatorname{Var} y \operatorname{Var} t}}$$

The coefficient of determination measures the fraction of the total variation in the output y which is explained by the linear prediction $\hat{y} = mt + c$. By the last formula, R^2 is *symmetric*; it does not matter whether y predicts t or vice versa. If R^2 is close to 1 it means that the linear model makes for a good predictor; the sum of the squared errors is small.

To return to our examples:

- In Example 3.1, $\bar{y} = \frac{4+8+10+21}{4} = \frac{43}{4} = 10\frac{3}{4}$, and

  $$\sum (y_i - \bar{y})^2 = \frac{27^2 + 11^2 + 3^2 + 41^2}{4^2} = \frac{635}{16}, \qquad \sum e_i^2 = 4.4 \implies R^2 \approx 0.9723$$

  The interpretation here is that the data is very close to being linear; the output y_i is very closely approximated by the linear model $\hat{y}_i = mt_i + c$.

[2] The *standard deviation* is the square root $\sigma_t = \sqrt{\operatorname{Var} t}$. This has the advantage of having the same units as t.

[3] The equivalence of these expressions is a messy exercise; use whichever formula you like! The point is to obtain interpretations, not to verify all this formally.

- In Example 3.4, $\sum(y_i - \overline{y})^2 = 14.8$ and

$$\sum(\hat{y}_i - y_i)^2 = \sum\left(\frac{2}{5}t_i + \frac{22}{5} - y_i\right)^2 = 3.2 \implies R^2 = \frac{3.2}{14.8} \approx 0.2162$$

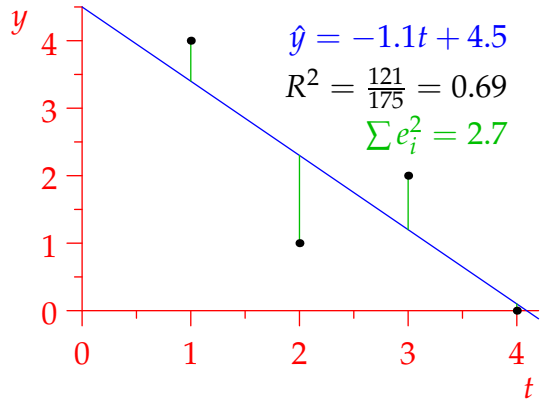
  In this case the coefficient of determination is small, which indicates that the model does not explaining much of the variation in the output.

Example 3.5. We do one more easy example with simple data

$$\{(t_i, y_i)\} = \{(1,4),(2,1),(3,2),(4,0)\}$$

this time using the formulae involving averages. You should feel confident doing this by hand, entirely without the assistance of even a calculator!

	data				average
t_i	1	2	3	4	$\overline{t} = \frac{10}{4}$
y_i	4	1	2	0	$\overline{y} = \frac{7}{4}$
t_i^2	1	4	9	16	$\overline{t^2} = \frac{15}{2}$
y_i^2	16	1	4	0	$\overline{y^2} = \frac{21}{4}$
$t_i y_i$	4	2	6	0	$\overline{ty} = 3$



$$m = \frac{\overline{ty} - \overline{t}\,\overline{y}}{\overline{t^2} - \overline{t}^2} = \frac{3 - \frac{70}{4^2}}{\frac{15}{2} - \frac{100}{4^2}} = -\frac{11}{10} = -1.1$$

$$c = \overline{y} - m\overline{t} = \frac{7}{4} + \frac{11 \cdot 10}{10 \cdot 4} = \frac{9}{2} = 4.5$$

The line for predicting y given t is $\hat{y} = -\frac{11}{10}t + \frac{9}{2} = -1.1t + 4.5$. Moreover, the coefficient of determination is

$$R^2 = m^2\frac{\overline{t^2} - \overline{t}^2}{\overline{y^2} - \overline{y}^2} = \frac{121}{100} \cdot \frac{\frac{15}{2} - \frac{100}{4^2}}{\frac{21}{4} - \frac{49}{4^2}} = \frac{121}{100} \cdot \frac{20}{35} = \frac{121}{175} \approx 0.691$$

Finally, the minimized square error is also easily computed:

$$\sum e_i^2 = \sum(\hat{y}_i - y_i)^2 = (3.4 - 4)^2 + (2.3 - 1)^2 + (1.2 - 2)^2 + (0.1 - 0)^2 = 2.7$$

While linear regression is undoubtedly useful it has obvious weaknesses, for instance:

- Outliers massively influence the regression line. Dealing with this problem is complicated; there are other approaches and definitions for a best-fitting straight line. For a simple approach with very large data sets, simply throw out the top and bottom 10% of data values! It is important to remember that any approach to modelling requires some *subjective choice*.

- If the data is 'not very linear' then the regression model will not be of much use. There are several ways around this. Sometimes data looks more linear after some manipulation, particularly by exponential or logarithmic functions. We'll think about this a bit later. For the present we consider how to find polynomial models for data, though to do this without losing sanity points requires a little detour into matrix multiplication.

Matrix Multiplication and Linear Regression

This is a very quick primer on how to multiply matrices. It is very easy if you are comfortable with the dot product! As an application, we obtain a useful way to think about the system of equations involved in linear regression which extends easily to polynomial regression.

- An $m \times n$ matrix is an array of mn numbers with m rows and n columns. Here, for example, is a 3×4 matrix,

$$\begin{pmatrix} 3 & 2 & 0 & 0 \\ -2 & \frac{1}{2} & -1 & 0 \\ 0 & 4 & 1 & -1 \end{pmatrix}$$

- Suppose A is $m \times n$ and B is $n \times r$. If a_{ij} is the entry in the i^{th} row, j^{th} column of A (and similarly for B), then the product AB is the $m \times r$ matrix with ik^{th} entry

$$(AB)_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk} = a_{i1}b_{1k} + a_{i2}b_{2k} + \cdots + a_{in}b_{nk}$$

  Equivalently, if we write $A = \begin{pmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_n^T \end{pmatrix}$ in terms of its n *rows*, and $B = \begin{pmatrix} \mathbf{b}_1 & \dots & \mathbf{b}_n \end{pmatrix}$ in terms of its n *columns*, then the i^{th} row j^{th} column of AB is the dot product $\mathbf{a}_i \cdot \mathbf{b}_j$. This is often said as "multiply along the row and down the column." For example

$$\begin{pmatrix} 3 & 2 & 0 \\ -2 & 1 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 & 2 & 0 \\ -2 & -1 & 3 & 1 \\ 0 & 5 & -1 & 1 \end{pmatrix} = \begin{pmatrix} -1 & 4 & 12 & 2 \\ -4 & 12 & -4 & 4 \end{pmatrix}$$

  where the 1^{st} row 2^{nd} column of the product was computed via

$$\begin{pmatrix} 3 & 2 & 0 \end{pmatrix} \begin{pmatrix} 2 \\ -1 \\ 5 \end{pmatrix} = 3 \cdot 2 + 2(-1) + 0 \cdot 5 = 4$$

- The $m \times m$ *identity* matrix has all of its entries zero, *except* down the main diagonal, all of whose entries are 1. For instance, the 3×3 identity matrix is

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

  The identity behaves very like the number 1: for any matrix A, we have $IA = AI = A$.

- A square matrix A has an *inverse* A^{-1} if $AA^{-1} = A^{-1}A = I$. For 2×2 matrices there is an explicit formula, provided $ad - bc \neq 0$,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

  Explicit formulas can be found for larger matrices, though they aren't practically useful. We won't explicitly invert anything beyond 2×2. Just know that computers are expert at this!

How does this relate to linear regression? The system of equations in Theorem 3.3 can be written in as a 2×2 matrix problem! For a data set with n pairs, the coefficients m, c satisfy

$$\begin{pmatrix} \sum t_i^2 & \sum t_i \\ \sum t_i & n \end{pmatrix} \begin{pmatrix} m \\ c \end{pmatrix} = \begin{pmatrix} \sum t_i y_i \\ \sum y_i \end{pmatrix}$$

This is nice because we can decompose the square matrix on the left as the product of a simple matrix and its transpose (switch the rows and columns);

$$\begin{pmatrix} \sum t_i^2 & \sum t_i \\ \sum t_i & n \end{pmatrix} = \begin{pmatrix} t_1 & t_2 & \cdots & t_n \\ 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} t_1 & 1 \\ t_2 & 1 \\ \vdots & \vdots \\ t_n & 1 \end{pmatrix} = P^T P$$

We can also view the right side as the product of P^T and the column vector of output values y_i:

$$\begin{pmatrix} \sum t_i y_i \\ \sum y_i \end{pmatrix} = \begin{pmatrix} t_1 & t_2 & \cdots & t_n \\ 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = P^T \mathbf{y}$$

A little theory tells us that if *at least two* of the t_i are distinct, then the matrix $P^T P$ is invertible;[4] there is a *unique* regression line whose coefficients may be found by taking the matrix inverse

$$\begin{pmatrix} m \\ c \end{pmatrix} = (P^T P)^{-1} P^T \mathbf{y} \implies \hat{y} = mt + c = (t \ \ 1) \begin{pmatrix} m \\ c \end{pmatrix} = (t \ \ 1)(P^T P)^{-1} P^T \mathbf{y}$$

We can also easily compute the prediction vector of values $\hat{y}_i$ given t_i:

$$\hat{y} = \begin{pmatrix} t_1 & t_2 & \cdots & t_n \\ 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} m \\ c \end{pmatrix} = P(P^T P)^{-1} P^T \mathbf{y}$$

and therefore the squared error $\sum e_i^2 = \sum |\hat{y}_i - y_i|^2 = ||\hat{\mathbf{y}} - \mathbf{y}||^2$, which leads to an alternative expression for the coefficient of determination

$$R^2 = \frac{||\hat{\mathbf{y}}||^2 - n\bar{y}^2}{||\mathbf{y}||^2 - n\bar{y}^2}$$

where $||\mathbf{y}||$ is the *length* of a vector.

[4] For those who've studied linear algebra, P and $P^T P$ have the same null space and thus rank, since

$$P\mathbf{x} = \mathbf{0} \implies P^T P\mathbf{x} = \mathbf{0} \quad \text{and} \quad P^T P\mathbf{x} = \mathbf{0} \implies \mathbf{x}^T P^T P\mathbf{x} = \mathbf{0} \implies |P\mathbf{x}| = \mathbf{0} \implies P\mathbf{x} = \mathbf{0}$$

For linear regression, having two distinct t_i values means P has rank (2); so does $P^T P$ which is therefore invertible.

Examples 3.6. 1. We revisit the previous example in this language

$$P = \begin{pmatrix} t_1 & 1 \\ t_2 & 1 \\ \vdots & \vdots \\ t_n & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{pmatrix} \implies P^T P = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{pmatrix} = \begin{pmatrix} 30 & 10 \\ 10 & 4 \end{pmatrix}$$

from which

$$\begin{pmatrix} m \\ c \end{pmatrix} = (P^T P)^{-1} P^T \mathbf{y} = \begin{pmatrix} 30 & 10 \\ 10 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 4 \\ 1 \\ 2 \\ 0 \end{pmatrix}$$

$$= \frac{1}{30 \cdot 4 - 10^2} \begin{pmatrix} 4 & -10 \\ -10 & 30 \end{pmatrix} \begin{pmatrix} 12 \\ 7 \end{pmatrix} = \frac{1}{20} \begin{pmatrix} 48 - 70 \\ -120 + 210 \end{pmatrix} = \frac{1}{10} \begin{pmatrix} -11 \\ 45 \end{pmatrix}$$

The prediction vector given inputs t_i is therefore

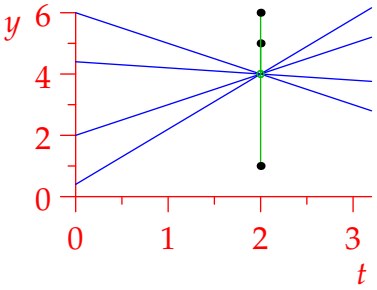
$$\hat{\mathbf{y}} = P \begin{pmatrix} m \\ c \end{pmatrix} = \frac{1}{10} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} -11 \\ 45 \end{pmatrix} = \frac{1}{10} \begin{pmatrix} 34 \\ 23 \\ 12 \\ 1 \end{pmatrix}$$

from which the coefficient of determination is, as before

$$R^2 = \frac{||\hat{\mathbf{y}}||^2 - 4\overline{y}^2}{||\mathbf{y}||^2 - 4\overline{y}^2} = \frac{\frac{1}{100}(34^2 + 23^2 + 12^2 + 1^2) - 4 \cdot \frac{7^2}{4^2}}{(4^2 + 1^1 + 2^2 + 0^2) - 4 \cdot \frac{7^2}{4^2}} = \frac{121}{175}$$

2. Given the data set $\{(3,1),(3,5),(3,6)\}$, we have $P = \begin{pmatrix} 3 & 1 \\ 3 & 1 \\ 3 & 1 \end{pmatrix}$ and $P^T P = \begin{pmatrix} 27 & 9 \\ 9 & 3 \end{pmatrix}$ which isn't invertible: $27 \cdot 3 - 9 \cdot 9 = 0$. The linear regression method doesn't work!

   It is easy to understand this from the picture. Since the three data points are vertically aligned, any line minimizing the sum of the squared errors must pass through the average $(3,4)$. It could however, have *any* slope!



It is unnecessary to use the matrix approach; particularly if you have to compute a small example you should use whichever approach you feel most comfortable with. There are, however, further advantages to the matrix approach.

- Computers store and manipulate data in essentially matrix format, so this approach is computer-ready; essential for *large* data set which would be prohibitive to work with manually.
- Suppose you repeat an experiment several times, taking measurements y_i at times times t_i. Since it depends only on the t-data, you need only compute the matrix $(P^T P)^{-1} P^T$ *once*; this makes it very efficient to compute the regression line for each experiment run.
- The method is easily generalizable to higher-degree polynomial regression...

Polynomial Regression

The pattern here works in almost the same way as for linear regression, you just need more terms. We work through the approach for a quadratic approximation.

Suppose we have a data set $\{(t_i, y_i) : 1 \leq i \leq n\}$ and that we desire a best-fitting quadratic polynomial $\hat{y} = at^2 + bt + c$ which minimizes the sum of the vertical errors

$$S = \sum e_i^2 = \sum (at_i^2 + bt_i + c - y_i)^2$$

This looks terrifying, but can be attacked as before using differentiation: to minimize the sum of the squared errors, we need all three of the derivatives of S with respect to the *variables* a, b, c to be zero.

$$\begin{aligned}
\frac{\partial S}{\partial a} &= 2\sum_{i=1}^{n} t_i^2(at_i^2 + bt_i + c - y_i) = 2\sum_{i=1}^{n} at_i^4 + bt_i^3 + ct_i^2 - t_i^2 y_i = 0 \\
\frac{\partial S}{\partial b} &= 2\sum_{i=1}^{n} t_i(at_i^2 + bt_i + c - y_i) = 2\sum_{i=1}^{n} at_i^3 + bt_i^2 + ct_i - t_i y_i = 0 \\
\frac{\partial S}{\partial c} &= 2\sum_{i=1}^{n} at_i^2 + bt_i + c - y_i = 0
\end{aligned}$$

which is equivalent to a system of equations for a, b, c:

$$\begin{cases}
a\sum t_i^4 + b\sum t_i^3 + c\sum t_i^2 = \sum t_i^2 y_i \\
a\sum t_i^3 + b\sum t_i^2 + c\sum t_i = \sum t_i y_i \\
a\sum t_i^2 + b\sum t_i + cn = \sum y_i
\end{cases}$$

This looks fairly nasty, and could be rephrased in terms of averages as before (e..g. $\overline{t^4} = \frac{1}{n}\sum t_i^4$). Instead, observe that we have the same matrix problem as previously, just with an $n \times 3$ matrix P:

$$P^T P \begin{pmatrix} a \\ b \\ c \end{pmatrix} = P^T \mathbf{y} \quad \text{where} \quad P = \begin{pmatrix} t_1^2 & t_1 & 1 \\ \vdots & \vdots & \vdots \\ t_n^2 & t_n & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

It can be checked that, provided at least three of the t_i are distinct, then the matrix $P^T P$ is invertible[5] and there is a unique least-squares quadratic minimizer

$$\hat{y} = at^2 + bt + c = \begin{pmatrix} t^2 & t & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

The predictions $\hat{y}_i = \hat{y}(t_i)$ therefore form a vector $\hat{\mathbf{y}} = P\begin{pmatrix} a \\ b \\ c \end{pmatrix} = P(P^T P)^{-1} P^T \mathbf{y}$, and the coefficient of determination may be computed as previously.

[5]Remember that the goal of this class is *not* to practice inverting 3×3 matrices. This is what computers are for! The point is to know that it can easily be done.

Example 3.7. Suppose we have the data

$$\{(t_i, y_i)\} = \{(1,2),(2,5),(3,7),(4,4)\}$$

We compute both the linear an quadratic regression curves.

1. To find the best-fitting least-squares line, we use the same P (and thus P^TP) from the previous example:

$$\begin{pmatrix} m \\ c \end{pmatrix} = (P^TP)^{-1}P^T\mathbf{y} = \begin{pmatrix} 30 & 10 \\ 10 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 5 \\ 7 \\ 4 \end{pmatrix}$$

$$= \frac{1}{10}\begin{pmatrix} 2 & -5 \\ -5 & 15 \end{pmatrix}\begin{pmatrix} 49 \\ 18 \end{pmatrix} = \begin{pmatrix} 0.8 \\ 2.5 \end{pmatrix}$$

   which yields $\hat{y} = 0.8t + 2.5$. The predicted values and the coefficient of determination are easily found:

$$\hat{\mathbf{y}} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 0.8 \\ 2.5 \end{pmatrix} = \begin{pmatrix} 3.3 \\ 4.1 \\ 4.9 \\ 5.7 \end{pmatrix}, \qquad R^2 = \frac{84.2 - 81}{94 - 81} \approx 0.2462$$

   The linear model is not very accurate.

2. Now compute the quadratic model:

$$P = \begin{pmatrix} 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \\ 16 & 4 & 1 \end{pmatrix} \implies P^TP = \begin{pmatrix} 1 & 4 & 9 & 16 \\ 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \\ 16 & 4 & 1 \end{pmatrix} = \begin{pmatrix} 354 & 100 & 30 \\ 100 & 30 & 10 \\ 30 & 10 & 4 \end{pmatrix}$$

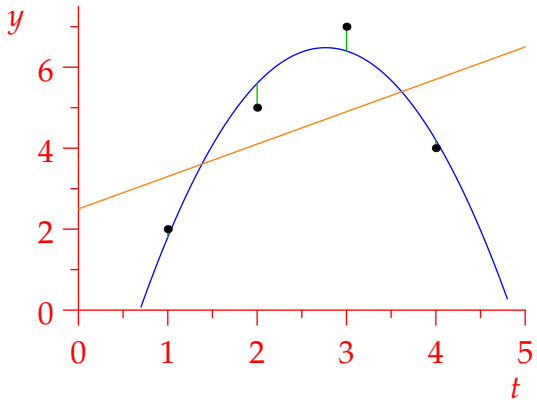
$$\implies \begin{pmatrix} a \\ b \\ c \end{pmatrix} = (P^TP)^{-1}P^T\begin{pmatrix} 2 \\ 5 \\ 7 \\ 4 \end{pmatrix} = \begin{pmatrix} 354 & 100 & 30 \\ 100 & 30 & 10 \\ 30 & 10 & 4 \end{pmatrix}^{-1}\begin{pmatrix} 149 \\ 49 \\ 18 \end{pmatrix} = \begin{pmatrix} -1.5 \\ 8.3 \\ -5 \end{pmatrix}$$

   so we have the model $\hat{y} = -1.5t^2 + 8.3ty - 5$. To quantify its accuracy, compute

$$\hat{\mathbf{y}} = P\begin{pmatrix} -1.5 \\ 8.3 \\ -5 \end{pmatrix} = \begin{pmatrix} 1.8 \\ 5.6 \\ 6.4 \\ 4.2 \end{pmatrix}$$

$$R^2 = \frac{||\hat{\mathbf{y}}||^2 - 4\bar{y}^2}{||\mathbf{y}||^2 - 4\bar{y}^2} = \frac{93.2 - 81}{94 - 81} \approx 0.9385$$

   The quadratic model is far superior to the linear.



That a quadratic model would provide a significantly better fit should have been obvious simply by plotting the data points. If this were real world data, there might have been other clues; perhaps the data is related to the trajectory of a object, which we know to follow a parabola!

We can repeat the approach for cubic, and higher-order, polynomials. Everything stays the same *except* for the matrix P, which gets extra columns: for a degree d polynomial fitting n data points, P will be an $n \times (d+1)$ matrix whose first column contains t_i^d. This is rarely useful in practice, and might even be counter-productive.

Returning to the example, there is in fact a unique cubic polynomial passing through the four data points[6]

$$\hat{y} = \frac{1}{6}(-4t^3 + 21t^2 - 17t + 12)$$

The best-fitting cubic polynomial to the data therefore has *no error.* This is likely a bad interpretation of a real-world situation, where the values y_i are likely *inaccurate* measurements from an experiment. You shouldn't expect observed data to be perfect; the 'perfect' cubic model not only takes much longer to compute, but it is likely only amplifying whatever noise was present in the original measurements y_i.

We'll return to regression later once we've thought about exponential functions.

[6]Compare with a unique line through two points, and a unique quadratic through three.