5 Regression Models

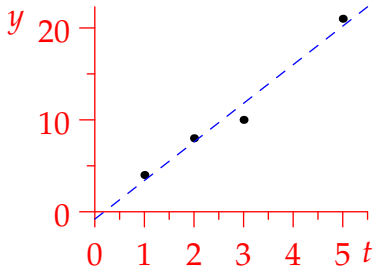
We've studied several types of function and seen how to spot whether a given data set might suit a particular model. To get further with this analysis, we need a method for comparing how bad a particular model is for given data.

5.1 Best-fitting Lines and Linear Regression

We start with an example of some data which appears reasonably linear.

Example 5.1. At t p.m., a trail-runner's GPS locator says that they've travelled y miles along a trail;

t_i	1	2	3	5
y_i	4	8	10	21

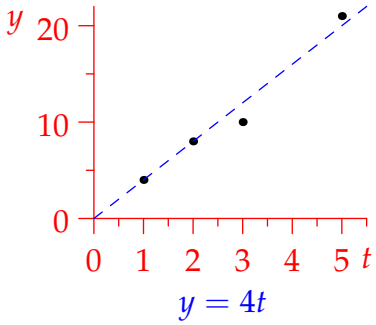


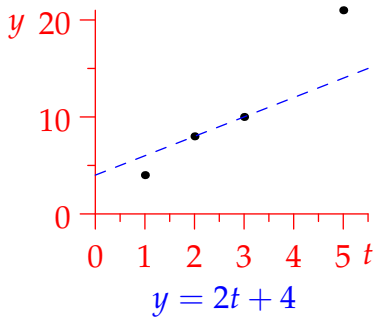
We'd like a simple model for how far the runner has travelled as a function of t. We might use this to predict where they would be at a given time; say at 6 p.m., or at 2 p.m. if they were to attempt the trail on another day.

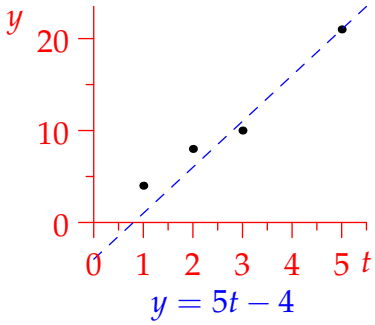
By plotting the points, the relationship looks to be approximately[12] linear: $y \approx mt + c$. What is the *best* choice of line, and how should we find the coefficients m, c?

What might be good criteria for choosing our line? What should we mean by *best*? Plainly, we want the points to be close to the line, but measured how? What use do we want to make of the approximating line?

Here are three candidate lines plotted with the data set: of the choices, which seems best and why?







Since we want our model to *predict* the hiker's location $y \approx \hat{y} = mt + c$ at a given time t, we'd like our model to minimize *vertical* errors $\hat{y}_i - y_i$. We've computed these in the table; since a positive error is as bad as a negative, we make all the errors positive. It therefore seems reasonable to claim that the first line is the best choice of the three.

But can we do better?

	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

[12] Why should we not expect the distance traveled by the hiker to be perfectly linear?

We need a sensible definition of *best-fitting line* for a given data set. One possibility is to minimize the sum of the vertical errors:

$$\sum_{i=1}^{n} |\hat{y}_i - y_i|$$

For reasons of computational simplicity, uniqueness, statistical interpretation, and to discourage large individual errors, we *don't* do this! The standard approach is instead to minimize the *sum of the squared errors.*

Definition 5.2. Let (t_i, y_i) be data points with at least two distinct t-values. Let $\hat{y} = mt + c$ be a linear predictor (model) for y given t.

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

Having at least two distinct t-values (some $t_i \neq t_j$) is necessary for the regression line to be unique.

Example (5.1, cont). Suppose the predictor was $\hat{y} = mt + c$. 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 to minimize the 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 is easy to deal with if we invoke some calculus. If (m, c) minimizes $S(m, c)$, then the first derivative tests says that the (partial) derivatives of S must be zero.

- Keep c constant and differentiate with respect to m:

$$\begin{aligned}\frac{\partial S}{\partial m} &= 2(m + c - 4) + 4(2m + c - 8) + 6(3m + c - 10) + 10(5m + c - 21) \\ &= 2\Big[39m + 11c - 155\Big]\end{aligned}$$

- Keep m constant and differentiate with respect to c:

$$\begin{aligned}\frac{\partial S}{\partial c} &= (m + c - 4) + (2m + c - 8) + (3m + c - 10) + (5m + c - 21) \\ &= 11m + 4c - 43\end{aligned}$$

The regression line is found by solving a pair of simultaneous equations

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

By 6 p.m., we predict that the runner would have covered 24.4 miles. The sum of the squared errors for our regression line is $\sum e_i^2 = \sum |\hat{y}_i - y_i|^2 = 4.4$, compared to 5, 53 and 14 for our earlier options.

To obtain the general result for n data points, we return to our computations of the partial derivatives:

$$\frac{\partial S}{\partial m} = \sum \frac{\partial}{\partial m}(mt_i + c - y_i)^2 = 2\sum t_i(mt_i + c - y_i) = 2\Big[\big(\sum t_i^2\big)\, m + \big(\sum t_i\big)\, c - \sum t_i y_i\Big]$$

$$\frac{\partial S}{\partial c} = \sum \frac{\partial}{\partial c}(mt_i + c - y_i)^2 = 2\sum (mt_i + c - y_i) = 2\Big[\big(\sum t_i\big)\, m + nc - \sum y_i\Big]$$

These sums are often written using a short-hand notation for *average*:

$$\overline{t} = \frac{1}{n}\sum_{i=1}^{n} t_i, \qquad \overline{t^2} = \frac{1}{n}\sum_{i=1}^{n} t_i^2, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i^2, \qquad \overline{ty} = \frac{1}{n}\sum_{i=1}^{n} t_i y_i$$

Theorem 5.3 (Linear Regression). *Given n data points (t_i, y_i) with at least two distinct t-values, the best-fitting least-squares line has equation $\hat{y} = mt + c$, where m, c satisfy*

$$\begin{cases} \big(\sum t_i^2\big)\, m + \big(\sum t_i\big)\, c = \sum t_i y_i \\ \big(\sum t_i\big)\, m + nc = \sum y_i \end{cases} \quad \leftrightsquigarrow \quad \begin{cases} \overline{t^2} m + \overline{t} c = \overline{ty} \\ \overline{t} m + c = \overline{y} \end{cases}$$

This is a pair of simultaneous equations for the coefficients m, c, with solution

$$m = \frac{\overline{ty} - \overline{t}\,\overline{y}}{\overline{t^2} - \overline{t}^2}, \qquad c = \overline{y} - m\overline{t}$$

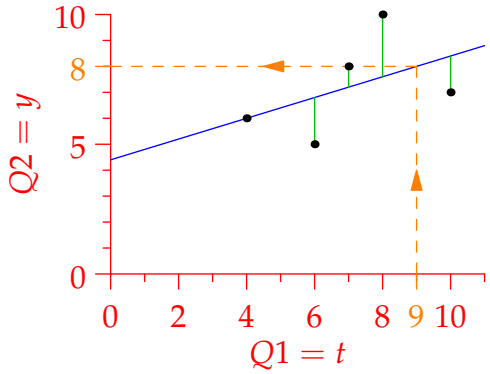
As the next section shows, having two distinct t-values guarantees a non-zero denominator $\overline{t^2} - \overline{t}^2$. The expression for c shows that the regression line passes through the data's *center of mass* $(\overline{t}, \overline{y})$.

Example 5.4. Five students' scores on two quizzes are given. If a student scores 9/10 on the first quiz, what might we expect them to score on the second?

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

To put the question in standard form, suppose Quiz 1 is the t-data and Quiz 2 the y-data. It is helpful to rewrite the data and add lines to the table so that we may more easily compute everything.

Data						$\sum$	Average
t_i	8	10	6	7	4	35	7
y_i	10	7	5	8	6	36	7.2
t_i^2	64	100	36	49	16	265	53
$t_i y_i$	80	70	30	56	24	260	52



$$m = \frac{52 - 7 \times 7.2}{53 - 7^2} = \frac{1.6}{4} = 0.4, \quad c = 7.2 - 0.4 \times 7 = 4.4$$

$$\implies \hat{y}(t) = \frac{2}{5}(t + 11)$$

This line which minimizes the sum of the squares of the vertical deviations. The prediction is that the hypothetical student scores $\hat{y}(9) = \frac{2}{5} \cdot 20 = 8$ on Quiz 2. Note that the predictor isn't symmetric: if we reverse the roles of t, y we don't get the same line!

Exercises 5.1. 1. Compute the sum of the absolute errors $\sum |\hat{y}_i - y_i|$ for the regression line and compare it to the sum of the absolute errors for $\hat{y} = 4t$: what do you notice?

2. Let $\hat{y} = mt + c$ be a linear predictor for the given data.

t_i	0	1	2	3
y_i	1	2	2	3

   (a) Compute the sum of squared-errors $S(m,c) = \sum e_i^2 = \sum |\hat{y}_i - y_i|^2$ as a function of m and c.

   (b) Compute the partial derivatives $\frac{\partial S}{\partial m}$ and $\frac{\partial S}{\partial c}$.

   (c) Find m and c by setting both partial derivatives to zero; hence find the equation of the regression line for these data.

   (d) Compare the sum of square errors S for the regression line with the errors if we use the simple predictor $y(t) = 1 + \frac{2}{3}t$ which passes through the first an last data points.

3. Consider Example 5.4.

   (a) Compute the sum of square-errors $S = \sum e_i^2 = \sum |\hat{y}_i - y_i|^2$ for the regression line.

   (b) Suppose a student was expected to score *exactly* the same on both quizzes; the predictor would be $\hat{y} = t$. What would the sum of squared-errors be in this case?

   (c) If a student scores 8/10 on *Quiz 2*, use linear regression to predict their score on Quiz 1. (*Warning: the answer is NOT* $\frac{5}{2} \cdot 8 - 11 = 9\ldots$)

4. Ten children had their heights (inches) measured on their first and second birthdays. The data was as follows.

1^{st} birthday	28	28	29	29	29	30	30	32	32	33
2^{nd} birthday	30	32	31	34	35	33	36	37	35	37

   Given this data, find a regression model and use it to predict the height at 2 years of a child who measures 32 inches at age 1.

   (*It is acceptable—and encouraged!—to use a spreadsheet to find the necessary ingredients. You can do this by hand if you like, but the numbers are large; it is easier with some formulæ from the next section.*)

5. (a) Let a, b be given. Find the value of y which minimizes the sum of squares

$$(y-a)^2 + (y-b)^2$$

   (b) For the data set $\{(t,y)\} = \{(1,1),(2,1),(2,3)\}$, find the unique least-squares linear model for predicting y given t.
   (*Hint: think about part (a) if you don't want to compute*)

   (c) Show that there are *infinitely many* lines $\hat{y} = mt + c$ which minimize the sum of the absolute errors $\sum_{i=1}^{3} |\hat{y}_i - y_i|$.

5.2 The Coefficient of Determination

In the sense that it minimizes the sum of the squared errors $S = \sum e_i^2$, the linear regression model is as good as it can be—but *how* good? We could use S as a *quantitative* measure of the model's accuracy, but it doesn't do a good job at comparing the accuracy of models for *different* data sets. The standard approach to this problem relies the concept of variance.

Definition 5.5. The *variance* of data sequence $(y_1, \ldots, y_n)$ is the average of the squared deviations from their mean $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$,

$$\operatorname{Var} y := \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$$

The *standard deviation* is $\sigma_y := \sqrt{\operatorname{Var} y}$.

Variance and standard-deviation are measures of how data deviates from being constant.

Example 5.6. Suppose $(y_i) = (1, 2, 5, 4)$. Then

$$\overline{y} = \frac{1}{4}(1+2+5+4) = 3 \qquad \operatorname{Var} y = \frac{1}{4}\Big((-2)^2 + (-1)^2 + 2^2 + 1^2\Big) = \frac{5}{2} \qquad \sigma_y = \frac{\sqrt{10}}{2}$$

The square-root means that σ_y has the same units as y. Loosely speaking, a typical data value is expected to lie approximately $\sigma_y = \frac{1}{2}\sqrt{10} \approx 1.58$ from the mean $\overline{y} = 3$.

To obtain a measure for how well a regression line fits given data (t_i, y_i), we ask what *fraction* of the variance in y is explained by the model.

Definition 5.7. The *coefficient of determination* of a model $\hat{y} = mt + c$ is the ratio

$$R^2 := \frac{\operatorname{Var}\hat{y}}{\operatorname{Var} y}$$

Examples 5.8. We start by considering two extreme examples.

1. If the data were perfectly linear, then $y_i = mt_i + c$ for all i. The regression line is therefore $\hat{y} = mt + c$ and the coefficient of determination is precisely $R^2 = \frac{\operatorname{Var} y}{\operatorname{Var} y} = 1$. All the variance in the output y is explained by the model's transfer of the variance in the input t.

2. By contrast, consider the data in the table where we work out all necessary details to find the regression line:

	data				average
t_i	0	0	2	2	$\overline{t} = 1$
y_i	1	3	1	3	$\overline{y} = 2$
t_i^2	0	0	4	4	$\overline{t^2} = 2$
$t_i y_i$	0	0	2	6	$\overline{ty} = 2$

$$m = \frac{\overline{ty} - \overline{t}\,\overline{y}}{\overline{t^2} - \overline{t}^2} = 0, \quad c = \overline{y} - m\overline{t} = 2$$

The regression line is the *constant* $\hat{y} \equiv 2$, whence $\hat{y}$ has *no variance* and the coefficient of determination is $R^2 = 0$.
In this example, the regression model doesn't help explain the y-data in any way: the t-values have no obvious impact on the y-values.

In fact, the coefficient of determination always lies somewhere between these extremes $0 \le R^2 \le 1$: Exercise 6 demonstrates this and that the extreme situations are essentially those just encountered; in practice, therefore, $0 < R^2 < 1$. Before we revisit our examples from the previous section, observe that the average of the model's outputs $\hat{y}_i$ is the same as that of the original data:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(mt_i + c) = m\bar{t} + c = \bar{y}$$

This makes computing the variance of $\hat{y}$ a breeze!

Example 5.1. Recall that $\hat{y} = \frac{1}{5}(21t - 4)$. Everything necessary is in the table

	data				average
t_i	1	2	3	5	$\bar{t} = 2.75$
y_i	4	8	10	21	$\bar{y} = 10.75$
$\hat{y}_i$	3.4	7.6	11.8	20.2	$\bar{\hat{y}} = 10.75$

$$\operatorname{Var} y = \frac{6.75^2 + 2.75^2 + 0.75^2 + 10.25^2}{4} = 39.6875$$
$$\operatorname{Var} \hat{y} = \frac{7.35^2 + 3.15^2 + 1.05^2 + 9.45^2}{4} = 38.5875$$

from which $R^2 = \frac{\operatorname{Var}\hat{y}}{\operatorname{Var} y} = \frac{3087}{3175} \approx 97.23\%$. The interpretation here is that the data is very close to being linear; the output y_i is very closely approximated by the regression model with approximately 97% of its variance explained by the model.

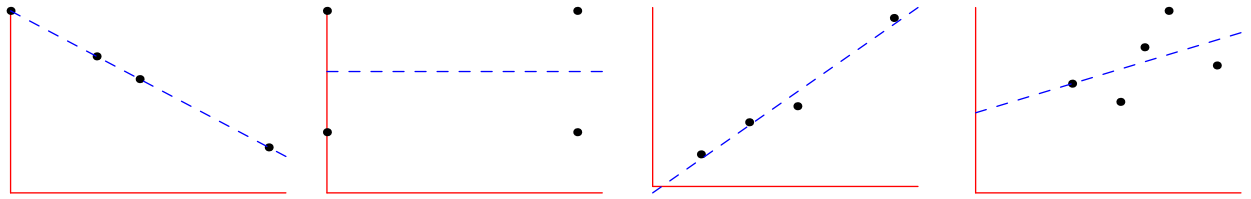
Example 5.4. This time $\hat{y} = \frac{2}{5}(t + 11)$.

	data					average
t_i	8	10	6	7	4	$\bar{t} = 7$
y_i	10	7	5	8	6	$\bar{y} = 7.2$
$\hat{y}_i$	7.6	8.4	6.8	7.2	6	$\bar{\hat{y}} = 7.2$

$$\operatorname{Var} y = \frac{2.8^2 + 0.2^2 + 2.2^2 + 0.8^2 + 1.2^2}{5} = 2.96$$
$$\operatorname{Var} \hat{y} = \frac{0.4^2 + 1.2^2 + 0.4^2 + 0^2 + 1.2^2}{5} = 0.64$$

from which $R^2 = \frac{\operatorname{Var}\hat{y}}{\operatorname{Var} y} = \frac{8}{37} \approx 21.62\%$. In this case the coefficient of determination is small, which indicates that the model does not explain much of the variation in the output.

The four examples are plotted below for easy visual comparison between the R^2-values.



Perfect model $R^2 = 1$ Useless model $R^2 = 0$ Good model $R^2 = 0.97$ Poor model $R^2 = 0.22$

Efficient computation of R^2 If you want to compute by hand, our current process is lengthy and awkward. To obtain a more efficient alternative we first consider an alternative expression for the variance of any collection of data:

$$\operatorname{Var} x = \frac{1}{n}\sum(x_i - \bar{x})^2 = \frac{1}{n}\sum x_i^2 - \frac{2\bar{x}}{n}\sum x_i + \frac{\bar{x}}{n}\sum x_i = \overline{x^2} - \bar{x}^2$$

Plainly $\operatorname{Var} x \ge 0$ with equality if and only if all data values x_i are equal. The alternative expression $\overline{x^2} - \bar{x}^2$ justifies the uniqueness of the regression line in Definition 5.2 and Theorem 5.3.

Now expand the variance of the predicted outputs:

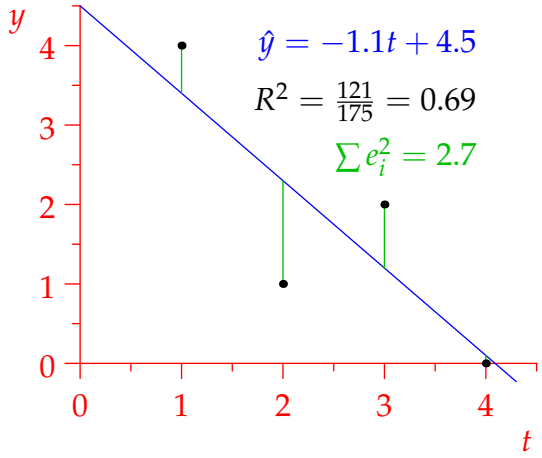
$$\operatorname{Var}\hat{y} = \frac{1}{n}\sum(\hat{y}_i - \overline{y})^2 = \frac{1}{n}\sum\big(mt_i + c - (m\overline{t} + c)\big)^2 = \frac{m^2}{n}\sum(t_i - \overline{t})^2 = m^2 \operatorname{Var} t$$

Putting these together, we obtain several equivalent expressions for the coefficient of determination:

$$R^2 = \frac{\operatorname{Var}\hat{y}}{\operatorname{Var} y} = m^2\frac{\operatorname{Var} t}{\operatorname{Var} y} = m^2\frac{\overline{t^2} - \overline{t}^2}{\overline{y^2} - \overline{y}^2} = \frac{(\overline{ty} - \overline{t}\,\overline{y})^2}{(\overline{t^2} - \overline{t}^2)(\overline{y^2} - \overline{y}^2)} \qquad (*)$$

Example 5.9. We do one more easy example with simple data $(t_i, y_i) : (1,4), (2,1), (3,2), (4,0)$.

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 regression line is $\hat{y} = -\frac{11}{10}t + \frac{9}{2} = -1.1t + 4.5$, and the coefficient of determination is

$$R^2 = m^2\frac{\overline{t^2} - \overline{t}^2}{\overline{y^2} - \overline{y}^2} = \frac{121}{100}\cdot\frac{\frac{15}{2} - \frac{100}{4^2}}{\frac{21}{4} - \frac{49}{4^2}} = \frac{121}{100}\cdot\frac{20}{35} = \frac{121}{175} = 69.1\%$$

The minimized square error is also easily computed:

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

Reversion to the Mean & Correlation By $(*)$, the regression model may be re-written in terms of the standard-deviation and R^2:

$$\hat{y}(t) = mt + c = \overline{y} + m(t - \overline{t}) = \overline{y} + \sqrt{R^2}\frac{\sigma_y}{\sigma_t}(t - \overline{t}) \implies \hat{y}(\overline{t} + \lambda\sigma_t) = \overline{y} + \lambda\sqrt{R^2}\sigma_y$$

Definition 5.10. The *correlation coefficient* is the value $r := \pm\sqrt{R^2}$ (sign equal to that of m).

An input λ standard-deviations above the mean ($t = \overline{t} + \lambda\sigma_t$) results in a prediction λr standard-deviations above the mean ($\hat{y} = \overline{y} + \lambda r\sigma_y$). Unless the data is perfectly linear, we have $R^2 < 1$; relative to the 'neutral' measure given by the standard-deviation a prediction $\hat{y}(t)$ is closer to the mean than the input t

$$\frac{|\hat{y}(t) - \overline{y}|}{\sigma_y} = r\frac{|\hat{t} - \overline{t}|}{\sigma_t} < \frac{|\hat{t} - \overline{t}|}{\sigma_t}$$

Example (5.9, cont). We compute the details. The correlation coefficient is $r = -\sqrt{R^2} \approx -0.832$; we say that the data is *negatively correlated,* since the output y seems to *decrease* as t increases. The standard deviations may be read off from the table:

$$\sigma_t = \sqrt{\operatorname{Var} t} = \sqrt{\overline{t^2} - \bar{t}^2} = \frac{\sqrt{5}}{2} \approx 1.118, \qquad \sigma_y = \sqrt{\operatorname{Var} y} = \sqrt{\overline{y^2} - \bar{y}^2} = \frac{\sqrt{35}}{4} \approx 1.479$$

The predictor may therefore be written (approximately)

$$\hat{y}(\bar{t} + \lambda\sigma_t) = \hat{y}(2.5 + 1.12\lambda) = \bar{y} + \lambda r \sigma_y = 1.75 - 1.23\lambda$$

As a sanity check,

$$\hat{y}(2.5 + 1.12) = \hat{y}(3.62) = -1.1 \times -3.98 + 4.5 = 0.52 = 1.75 - 1.23$$

Weaknesses of Linear Regression There are two obvious issues:

- Outliers massively influence the regression line. Dealing with this problem is complicated and there are a variety of approaches that can be used. It is important to remember that any approach to modelling, including our regression model, requires some *subjective choice.*
- If the data is not very linear then the regression model will produce a weak predictor. There are several ways around this as we'll see in the remaining sections: higher-degree polynomial regression can be performed, and data sometimes becomes more linear after some manipulation, say by an exponential or logarithmic function.

Exercises 5.2. 1. Suppose $(z_i) = (2, 4, 10, 8)$ is *double* the data set in Example 5.6. Find $\bar{z}$, $\operatorname{Var} z$ and σ_z. Why are you not surprised?

2. Use a spreadsheet to find R^2 for the predictor in Exercise 5.1.4. How confident do you feel in your prediction?
3. Find the standard deviations and correlation coefficients for the data in Examples 5.1 and 5.4.
4. The adult heights of men and women in a given population satisfy the following:

   Men: average 69.5 in, $\sigma = 3.2$ in. Women: average 63.7 in, $\sigma = 2.5$ in.

   The height of a father and his adult daughter have correlation coefficient 0.35. If a father's height is 72 in (mother's height unknown), how tall do you expect their daughter to be?
5. Suppose R^2 is the coefficient of determination for a linear regression model $\hat{y} = mt + c$. Use one of the alternative expressions for R^2 (page 59) to find the coefficient of determination for the reversed predictor $\hat{t}(y)$? Are you surprised?
6. Suppose that a data set $\{(t_i, y_i)\}_{1 \le i \le n}$ has at least two distinct t- and y-values (some $t_i \neq t_j$, etc.), that it has regression line $\hat{y} = mt + c$ and coefficient of determination R^2.
   (a) Show that $R^2 = 0 \iff m = 0$.
   (b) (Hard) Prove that the sum of squared errors equals $S = \sum_{i=1}^{n} e_i^2 = n(\operatorname{Var} y - \operatorname{Var} \hat{y})$.
   (c) Obtain the alternative expression $R^2 = 1 - \frac{S}{n \operatorname{Var} y}$. Hence conclude that $R^2 \le 1$, with equality if and only if the original data set is perfectly linear.

5.3 Matrix Multiplication & Polynomial Regression

In this section we consider how to find a best-fitting least-squares polynomial for given data. To see how to do this, it helps to rephrase the linear approach using matrices.[13]

We start by observing that the system of equations in Theorem 5.3 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 $2 \times n$ 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 2×2 matrix $P^T P$ is invertible;[14] 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 vector of predicted values $\hat{y}_i = \hat{y}(t_i)$:

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

and 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

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

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

---

[13] Matrix computations are non-examinable. The purpose of this section is to be see how the regression may easily be automated and generalized by computer and to understand a little of how a spreadsheet calculates best-fitting curves of different types.

[14] 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 at least two distinct t_i values means rank $P = 2$, whence $P^T P$ is invertible.

Examples 5.11. 1. We revisit the Example 5.9 in this language.

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

from which

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

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

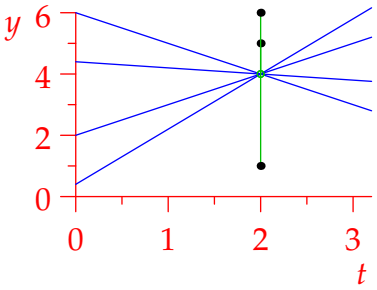
The prediction vector given inputs t_i is therefore

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

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 $P^TP = \begin{pmatrix} 27 & 9 \\ 9 & 3 \end{pmatrix}$ 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 $(3,4)$, though it could have *any* slope!
This illustrates our fundamental assumption: linear regression requires at least two distinct t-values.

It is unnecessary ever to use the matrix approach for linear regression, though the method has significant advantages.

- Computers store and manipulate data in matrix format, so this method is computer-ready.
- Suppose you repeat an experiment several times, taking measurements y_i at times t_i. Since P depends only on the t-data, you need only compute the matrix $(P^TP)^{-1}P^T$ *once*, making computation of the regression line for repeat experiments very efficient.
- The method generalizes (easily for computers!) to polynomial regression...

Polynomial Regression

The pattern is almost identical when we use matrices; you just need to make the matrix P a little larger…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 quadratic polynomial predictor $\hat{y} = at^2 + bt + c$ which minimizes the sum of the squared vertical errors

$$S(a,b,c) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (at_i^2 + bt_i + c - y_i)^2$$

This might look terrifying, but can be attacked exactly as before using differentiation: to minimize S, we need the derivatives of S with respect to the coefficients a, b, c to be zero.

$$\begin{cases} \dfrac{\partial S}{\partial a} = 2\sum at_i^4 + bt_i^3 + ct_i^2 - t_i^2 y_i = 0 \\ \dfrac{\partial S}{\partial b} = 2\sum at_i^3 + bt_i^2 + ct_i - t_i y_i = 0 \\ \dfrac{\partial S}{\partial c} = 2\sum at_i^2 + bt_i + c - y_i = 0 \end{cases} \iff \begin{cases} a\sum t_i^4 + b\sum t_i^3 + c\sum t_i^2 = \sum t_i^2 y_i \\ a\sum t_i^3 + b\sum t_i^2 + c\sum t_i = \sum t_i y_i \\ a\sum t_i^2 + b\sum t_i + cn = \sum y_i \end{cases}$$

As a system of equations for a, b, c this looks fairly nasty, but by rephrasing in terms of matrices, we see that it is exactly the same problem as before!

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

corresponds to

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

The only change is that P is now an $n \times 3$ matrix so that $P^T P$ is 3×3. Analogous to the linear situation, provided at least three of the t_i are distinct, the matrix $P^T P$ is invertible and there is a unique least-squares quadratic minimizer

$$\hat{y} = at^2 + bt + c = \begin{pmatrix} t^2 & t & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} t^2 & t & 1 \end{pmatrix} (P^T P)^{-1} P^T \mathbf{y}$$

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 before.

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

The method generalizes in the obvious way: if you want a cubic minimizer, give P an extra column of *cubed* t_i-terms! This would be hard work by hand, but is standard fodder for computers: this isn't a linear algebra class, so don't try to invert a 3×3 matrix!

Example 5.12. We are given data $\{(t_i, y_i)\} = \{(1,2),(2,5),(3,7),(4,4)\}$.

1. For the best-fitting linear model, 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 $\hat{y}(t) = 0.8t + 2.5$. The predicted values and coefficient of determination are then

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

   The linear model predicts only 24.6% of the variance in the output; not very accurate.

2. For a quadratic model; all that changes is the matrix P

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

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

   from which $\hat{y} = -1.5t^2 + 8.3t - 5$. To quantify its accuracy, compute the vector of predicted values $\hat{y}_i = \hat{y}(t_i)$ and the coefficient of determination:

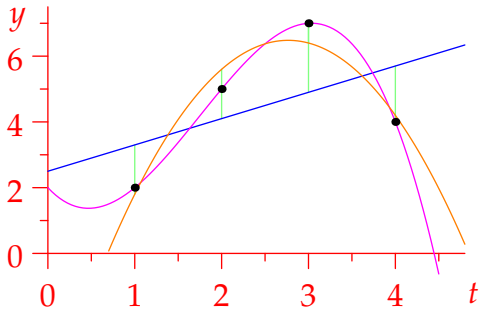
$$\hat{\mathbf{y}} = P\begin{pmatrix} -1.5 \\ 8.3 \\ -5 \end{pmatrix} = \begin{pmatrix} 1.8 \\ 5.6 \\ 6.4 \\ 4.2 \end{pmatrix} \qquad R^2 = \frac{||\hat{\mathbf{y}}||^2 - 4\overline{y}^2}{||\mathbf{y}||^2 - 4\overline{y}^2} = \frac{93.2 - 81}{94 - 81} \approx 0.9385$$

   The quadratic model is far superior to the linear, explaining 94% of the observed variance.

3. We can even find a cubic model (P is a 4×4 matrix!)

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

   The cubic passes through all four data points, there is *no error* and $R^2 = 1$.



For real-world data this is possibly *less useful* than the quadratic model—it certainly takes longer to find! More importantly, likely experimental error in the y-data has a strong effect on the 'perfect' model—we are, in effect, modelling *noise.* Do you expect $y(5)$ to be closer to -1 or -8?

Exercises 5.3. 1. Recall Example 4.2, with the following *almost* linear data set.

x	0	2	4	6	8	10
y	3	23	41	59	77	93

Find the best-fitting straight line for the data, then use a spreadsheet to find the best-fitting quadratic. Is the extra effort worth it?

2. You are given the following data consisting of measurements from an experiment recorded at times t_i seconds.

t_i	1	2	3	4	5	6	7	8	9	10
y_i	7	5	3	2	3	5	6	9	8	12

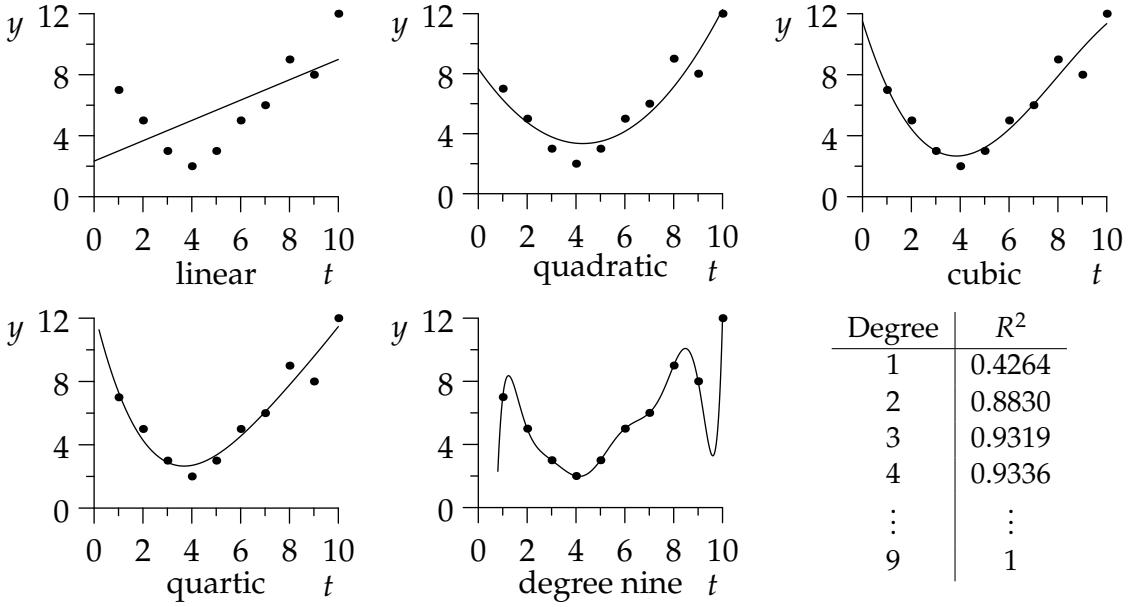
(a) Given the values

$$\sum t_i = 55, \quad \sum t_i^2 = 385, \quad \sum y_i = 60, \quad \sum t_i y_i = 385$$

find the best-fitting least-squares linear model for this data, and use it to predict $\hat{y}(13)$.

(b) Find the best-fitting quadratic model for the data: feel free to use a spreadsheet!

(c) The graphs below show the best-fitting least-squares linear, quadratic, cubic, quartic, and ninth-degree models and their coefficients of determination.



Degree	R^2
1	0.4264
2	0.8830
3	0.9319
4	0.9336
⋮	⋮
9	1

Which of these models would *you* choose for this data and why? What considerations would you take into account?

5.4 Exponential & Power Regression Models

If you suspect that your data would be better modelled by a non-polynomial function, there are several things you can try.

Minimizing the sum of squared-errors might be very difficult for non-polynomial functions because there is likely no simple tie-in with linear equations/algebra. Attempting this is likely to result in a horrible *non-linear* system for your coefficients which is difficult to analyze either theoretically or using a computer.[15]

Log Plots The most common approach when trying to fit an exponential model $\hat{y} = e^{mt+c}$ to data is to use a log plot: taking logarithms of both sides results in

$$\ln \hat{y} = mt + c$$

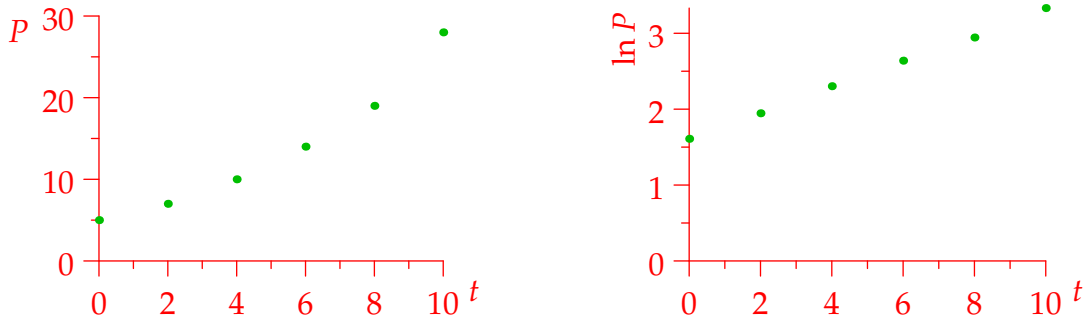
If we take $\hat{Y} := \ln \hat{y}$ as a new variable, the model is now a *straight-line*! The idea is then to use linear regression to find the coefficients $m, \ln a$.

Example (4.4, cont). Recall our earlier rabbit-population $P(t)$, repeated in the table below. We previously considered modelling this with an exponential function for two reasons:

1. We were told it was population data!
2. The *t-differences* are constant (2), while the *P-ratios* are approximately so (≈ 1.41).

t_i	0	2	4	6	8	10
P_i	5	7	10	14	19	28
$\ln P_i$	1.61	1.95	2.30	2.64	2.94	3.33

After constructing a log-plot, the relationship is much clearer:



Since the relationship between t and $\ln P$ appears linear, we perform a linear regression calculation to find the best-fitting least-squares line for the $(t_i, \ln P_i)$ data.

[15] As an example of how horrific this is, suppose you want to minimize the sum of square-errors for data (t_i, y_i) using an exponential model $\hat{y}(t) = ae^{kt}$. The coefficients of our model, a, k should minimize

$$S(a,k) = \sum_{i=1}^{n} \left(ae^{kt_i} - y_i\right)^2$$

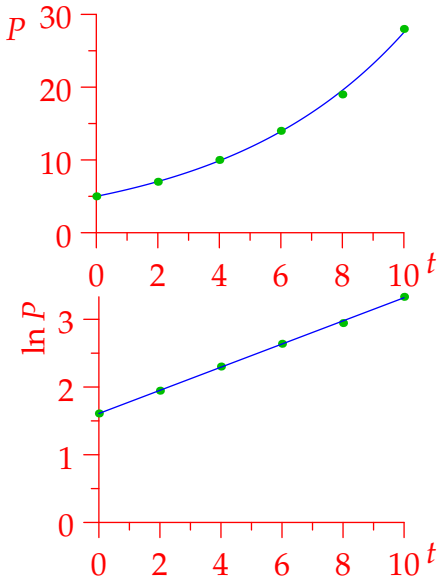
Differentiating this with respect to a, k and setting equal to zero results in

$$\begin{cases} \frac{\partial S}{\partial a} = 2\sum e^{kt_i}\left(ae^{kt_i} - y_i\right) = 0 \\ \frac{\partial S}{\partial k} = 2a\sum t_i e^{kt_i}\left(ae^{kt_i} - y_i\right) = 0 \end{cases} \implies \left(\sum y_i e^{kt_i}\right)\left(\sum t_i e^{2kt_i}\right) = \left(\sum e^{2kt_i}\right)\left(\sum t_i y_i e^{kt_i}\right)$$

where we substituted for a to obtain the last equation. Remember that this is an equation for k; if you think you can solve this easily, think again!

Everything necessary comes from extending the table.

	Data						average
t_i	0	2	4	6	8	10	5
P_i	5	7	10	14	19	28	13.83
$\ln P_i$	1.61	1.95	2.30	2.64	2.94	3.33	2.46
t_i^2	0	4	16	36	64	100	36.67
$t_i \ln P_i$	0	3.89	9.21	15.83	23.56	33.32	14.30



$$m = \frac{\overline{t \ln P} - \bar{t} \cdot \overline{\ln P}}{\overline{t^2} - \bar{t}^2} = \frac{14.30 - 5 \cdot 2.46}{36.67 - 5^2} = 0.171$$

$$c = \overline{\ln P} - m\bar{t} = 3.46 - 0.171 \cdot 5 = 1.609$$

which yields the exponential model

$$\hat{P}(t) = e^{0.171t + 1.609} = 4.998(1.186)^t$$

This is very close to the model $(5(1.188)^t)$ we obtained previously by pure guesswork. The approximate doubling time T for the population satisfies

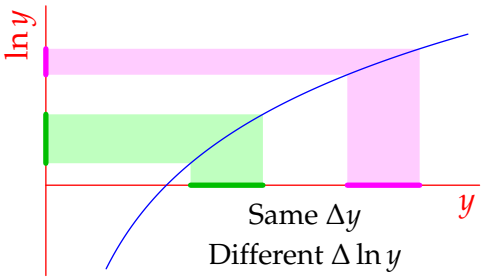
$$e^{mT} = 2 \implies T = \frac{\ln 2}{m} = 4.06 \text{ months}$$

When using the log plot method, interpreting errors and the goodness of fit of a model is a little more difficult. Typically one computes the coefficient of determination R^2 of the *underlying linear model*: in our example,[16]

$$R^2 = m^2 \frac{\operatorname{Var} t}{\operatorname{Var} \ln P} = 99.3\%$$

It is important to appreciate that the log plot method does not treat all errors equally: taking logarithms tends to reduce error by a greater amount when the output y is large. This should be clear from the picture, and more formally by the mean value theorem: if $y_1 < y_2$, then there is some $\xi \in (y_1, y_2)$ for which

$$\ln y_1 - \ln y_2 < \frac{1}{\xi}(y_1 - y_2)$$



The log plot approach therefore places a higher emphasis on accurately matching data when the output y is *small*. This isn't such a bad thing, since our intuitive view of error depends on the size of the data. For instance, you might be very annoyed to discover that you've misplaced a $100 bill, but if you've just bought a house for $1 million, a $100 mistake in escrow is unlikely to concern you very much! Exponential data can more easily vary over large orders of magnitude than linear or quadratic data.

[16] This needs more decimal places of accuracy for the log-values than what's in our table!

Log-Log Plots If you suspect a *power function model* $\hat{y} = at^m$, then taking logarithms

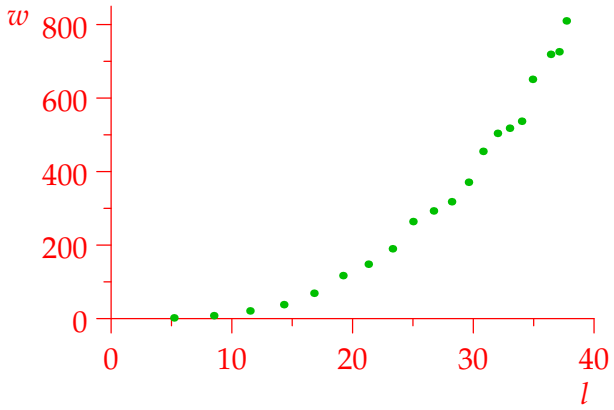
$$\ln \hat{y} = m \ln t + \ln a$$

results in a linear relationship between $\ln y$ and $\ln t$. As before, we can apply a linear regression approach to find a mode; the goodness of fit is again described by the coefficient of determination of the underlying model.

Exercises 5.4. 1. You suspect a logarithmic model for a data set. Describe how you would approach finding a model in the context of this section.

2. The table shows the average weight and length of a fish species measured at different ages.

Age (years)	Length(cm)	Weight (g)
1	5.2	2
2	8.5	8
3	11.5	21
4	14.3	38
5	16.8	69
6	19.2	117
7	21.3	148
8	23.3	190
9	25.0	264
10	26.7	293
11	28.2	318
12	29.6	371
13	30.8	455
14	32.0	504
15	33.0	518
16	34.0	537
17	34.9	651
18	36.4	719
18	37.1	726
20	37.7	810



(a) Do you think an exponential model is a good fit for this data? Take logarithms of the weight values and use a spreadsheet to obtain a model $\hat{w}(\ell) = ae^{m\ell}$ where w, ℓ are the weight and length respectively.

(b) What happens if you try a log-log plot? Given what we're measuring, why do you expect a power model to be mode accurate?

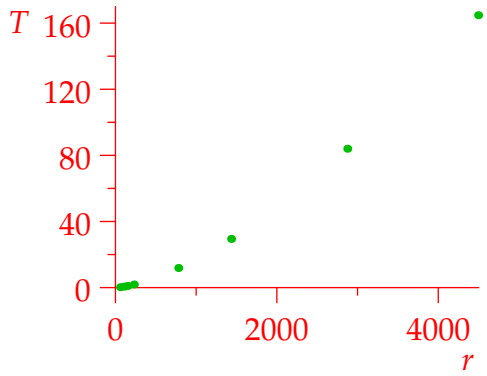
3. Population data for Long Beach CA is given.

Using a spreadsheet or otherwise, find linear, quadratic, exponential and logarithmic regression models for this data.
Which of these models seems to fit the data best, and which would you trust to best predict the population in 2020?
Look up the population of Long Beach in 2020; does it confirm your suspicions? What do you think is going on?

Year	Years since 1900	Population
1900	0	2,252
1910	10	17,809
1920	20	55,593
1930	30	142,032
1940	40	164,271
1950	50	250,767
1960	60	334,168
1970	70	358,879
1980	80	361,498
1990	90	429,433
2000	100	461,522
2010	110	462,257

4. In the early 1600s, Johannes Kepler used observational data to derive his *laws of planetary motion*, the third of which relates the orbital period T of a planet (how long it takes to go round the sun) to its (approximate) distance r from the sun.

Planet	T (years)	r (millions km)
Mercury	0.24	58
Venus	0.61	110
Earth	1	150
Mars	1.88	230
Jupiter	11.9	780
Saturn	29.5	1400
Uranus	84	2900
Neptune	165	4500



The table shows the data for all the planets. Use a spreadsheet to analyze this data and find a model relating T to r.

Kepler did not known about Uranus and Neptune and only had relative distances for the planets. Research the correct statement of Kepler's third law and compare it with your findings.