Curve fitting and least-squares regression

Business Mathematics (BK/IBA)
Quantitative Research Methods I (EBE)

October 2015

1 Introduction

Suppose we have a series of paired data points for two variables. For example, we have a sample of data on house characteristics in terms of floor area (in square metre) in relation to their price (in thousands of euro). Thus, for one specific house, with floor area 142 m^2 and price 649 thousands of euro, we could indicate this as a point in the floor area-price surface with coordinates $(142, 649000)$. For another house, we could have different data, e.g., $(80, 125000)$. From now on, we will abbreviate the variables floor area and price with the symbols x (in m^2) and y (in euro) respectively. Thus, the first house has $x_1 = 142$ and $y_1 = 649000$. Table 1 shows a fragment of a sample of $n = 71$ houses in Amstelveen in May 2014.

The data are visualised as a scatterplot in Figure 1.

As we see and expect, in general, house prices are higher for larger houses. It even seems that a linear relationship would be approximately appropriate. A linear formula for this has the general form

$$\boxed{y = ax + b} \tag{1}$$

house (i)	floor area in m^2 (x_i)	price in euro (y_i)
1	142	649000
2	80	125000
...	...	...
71	89	237500

Table 1: Fragment of the data set of 71 house prices and floor areas. Source: www.funda.nl.

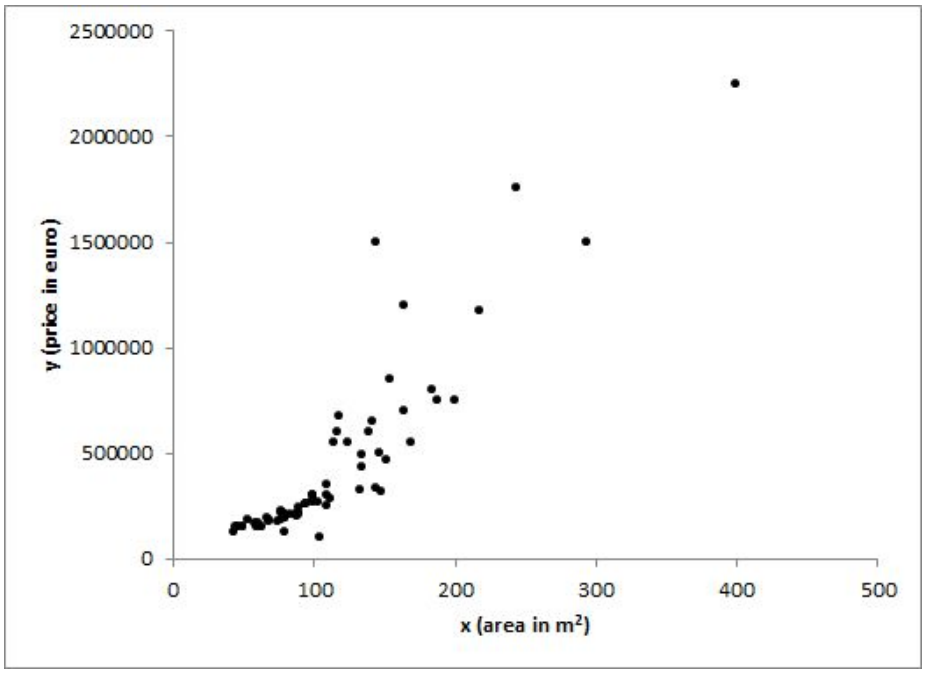

Figure 1: Scatterplot of the data of Table 1.

The question is now: what is the value of the coefficients a and b such that we have the best possible line ?

To answer this question, we first need to acknowledge that the data do not follow the linear relationship in a precise way. That is, we will expect that the linear formula is approximately correct, but that for many (or for some, or for all) data points, the points will not exactly all on the line. Thus, we need to look for a straight line that somehow is the best fitting line to the data.

In this document, we will explore how to find the straight line that "best" fits the data. This line is often referred to as a "regression line". The learning goals of this document are multiple:

- to discuss the important concept of a regression line;
- to show the strategy how to derive a theory of finding a regression line;
- to demonstrate how the summation operator, the theory of stationary points ad the concept of partial derivatives come together in an important application with business relevance.

We will put emphasize on the strategy and the proofs. We do not require you to be able to reproduce these proofs at the exam, but we do expect you to grasp the main line of argument.

2 Mathematical treatment

2.1 Formulating the problem

We first introduce notation. We have already defined the two variables as x and y, with x the so-called independent variable (floor area), and y the dependent variable (price). With this, we imply that y depends on x rather than that x depends on y. House prices depend on floor area, rather than the other way around. Mathematically, dependence is a mutual phenomenon: if y depends on x through $y = ax + b$, x depends on y through $x = \frac{1}{a}(y - b)$, which may be written as $x = cy + d$. But logically, the dependence is unidirectional.

So, our model equation is $y = ax + b$. We are now given a sample of observations. In observation 1, the value of independent variable was x_1, while the value of dependent variable was y_1. For the second observation, we have the values x_2 and y_2. Altogether, we have n observations, so the last one can be indicated as having the values x_n and y_n.

In general, we will speak of observation i with observed values x_i and y_i. Here, the index variable i can assume all integer values between 1 and n. Notice that the number of observed x-values is equal to the number of observed y-values: n. In a problem with paired data, we always have that $n_x = n_y = n$.

Now, for every observation i, we expect that the observed value for y (so y_i) is approximately given by the observed value for x (so x_i) inserted in the formula. Thus y_i should be approximately equal to $a + bx_i$. In symbols:

$$y_i \approx ax_i + b \tag{2}$$

There is an important alternative way of writing this, namely as

$$y_i = ax_i + b + e_i \tag{3}$$

Here, e_i is the error term for observation i. It measures how "wrong" the linear formula is for this particular observation i.

Our straight line, determined by the coefficients a and b, should be such that the error terms are as small as possible. Obviously, for one observation (namely i), we can always choose the coefficients in such a way that the error is zero. In that case the line passes through the data point (x_i, y_i). We can even do this for a second data point. But in general, we cannot do that for a third, fourth or for all of the remaining points. Whenever it passes through observation 1 and 2, it will not pass through observation 3, 4, etc, and whenever it passes through the observation 1 and 3, it will not pass through observation 2, 4, etc. Which points should we favour?

The general rule is that we will not favour any point, but rather will make sure that the total mismatch is as small as possible. Thus, we will minimize the errors collectively. To do this, we must define what we mean by total error.

An obvious initial choice seems to be the sum of the error terms for all observations, so

$$\sum_{i=1}^{n} e_i \tag{4}$$

But look carefully: the error of observation i is given by

$$e_i = y_i - (ac_i + b) \tag{5}$$

and this error will be positive when the point lies above the straight line, and negative when the point lies below it. So, overshoots of one point and undershoots of another point would cancel in this total error formula. This means we must develop a better formula for the total error, one that somehow manages to get rid of the minus sign in the case of an overshoot. One way to do this is by taking the error without the sign:

$$\sum_{i=1}^{n} |e_i| \tag{6}$$

However, this form turns out to yield cumbersome calculations. A third attempt is by squaring the error term, so by considering $(e_i)^2$ instead of e_i:

$$\sum_{i=1}^{n} (e_i)^2 \tag{7}$$

This form turns out to have desirable properties, and it is the form usually taken.

Now, with this modification in mind, we come to the following equation:

$$\boxed{\varepsilon = \sum_{i=1}^{n} (e_i)^2} \tag{8}$$

where ε is the sum of squared error terms, which we have defined as the measure of the total error.

We will introduce one simplification at this point: we will skip the $i = 1$ and n in order to keep the text more easy to read. So we will write

$$\varepsilon = \sum (e_i)^2 \tag{9}$$

where we implicitly understand that the summation runs over all data points i.

In summary, we wish to choose the coefficients a and b such that the sum of squares of errors ε is minimized. Now, you will understand why the title of this document is about the "least squares method": it is a method that determines the coefficients a and b such that the sum of squared error terms is least. The least squares principle goes back to the mathematician Gauss in the early eighteenth century.

2.2 Solving the problem

So far, this is about the strategy: we have defined a conceptual idea of a line that is supposed to satisfy a criterion of minimizing the sum of squares. Now, we enter the second stage of the mathematical treatment: the tactics. We are supposed to minimize something, and we know a method to do that: look for stationary points.

In order to do that, we must define ε as a function, and subject that function to the well-known minimization procedure. But if ε is a function, of what is it a function? Of x_i and/or y_i? No, because with given (i.e., fixed) x_i and y_i, we will be changing the coefficients a and b such that ε is minimized. In other words, ε is a function of a and b:

$$\varepsilon = f(a, b) \tag{10}$$

The sum of squared errors can be written as

$$\boxed{\varepsilon = f(a,b) = \sum (e_i)^2 = \sum \left(y_i - (ax_i + b)\right)^2} \tag{11}$$

We will determine the stationary point(s) of $f(a,b)$. The stationary points are those points for which the derivative of f is zero. In this case, f is a function of two variables (a and b), so we must use partial derivatives for this. The result is two equations:

$$\boxed{\frac{\partial f(a,b)}{\partial a} = 0} \tag{12}$$

and

$$\boxed{\frac{\partial f(a,b)}{\partial b} = 0} \tag{13}$$

We have two equations, one for the partial derivative with respect to a and one for the partial derivative with respect to b. The only thing is to work out the derivatives and to solve the system of two equations.

Doing the derivatives is a good exercise in applying the rules of differentiation (sum rule, product rule, chain rule) and the rules of working with the summation operator. Let's go and first give the partial derivative with respect to a a try.

We start by

$$\frac{\partial f(a,b)}{\partial a} = \frac{\partial \left(\sum \left(y_i - (ax_i + b)\right)^2\right)}{\partial a} = 0 \tag{14}$$

At the righthand-side, we have a derivative of a sum. The sum rule tells us that this is equal to the sum of the derivatives. In other words, we may swap differentiation and summation. This yields

$$\sum \frac{\partial \left(\left(y_i - (ax_i + b)\right)^2\right)}{\partial a} = 0 \tag{15}$$

The second step employs the chain rule: we differentiate the term with the square:

$$\sum \left(2\left(y_i - (ax_i + b)\right) \times -x_i\right) = 0 \tag{16}$$

This can be worked out as

$$\sum \left(-2\left(x_i y_i - a(x_i)^2 - bx_i\right)\right) = 0 \tag{17}$$

We can divide the factor -2 out (why?):

$$\sum \left(x_i y_i - a(x_i)^2 - bx_i\right) = 0 \tag{18}$$

Look carefully: we see a summation operator with three linear terms inside. This may be expanded as:

$$\sum (x_i y_i) - \sum \left(a(x_i)^2\right) - \sum (bx_i) = 0 \tag{19}$$

The coefficients a and b do not depend on i, so we may bring them in front of the summation operators:

$$\boxed{\sum x_i y_i - a \sum x_i^2 - b \sum x_i = 0} \tag{20}$$

In this equation three sums over the observations show up: $\sum x$, $\sum x^2$, and $\sum xy$. For a given set of observations, these three sums are just numbers. So. basically, this equation can just be something like

$$23 - a \times 21 - b \times -5 = 0 \tag{21}$$

Yet, this is one equation in two unknowns, so we cannot solve a and b. This changes when we include the second derivative, the one respect to b:

$$\frac{\partial f(a,b)}{\partial b} = \frac{\partial \left(\sum \left(y_i - (a x_i + b) \right)^2 \right)}{\partial b} = 0 \tag{22}$$

Now, we basically do the same tricks. We swap differentiation and summation to get

$$\sum \frac{\partial \left(\left(y_i - (a x_i + b) \right)^2 \right)}{\partial b} = 0 \tag{23}$$

and apply next the chain rule to find

$$\sum \left(2 \left(y_i - (a x_i + b) \right) \times -1 \right) = 0 \tag{24}$$

which can be worked out as

$$\sum \left(-2 \left(y_i - (a x_i + b) \right) \right) = 0 \tag{25}$$

Skipping the factor -2 again, we obtain

$$\sum \left(\left(y_i - (a x_i + b) \right) \right) = 0 \tag{26}$$

or

$$\sum y_i - \sum \left(a x_i \right) - \sum b = 0 \tag{27}$$

The special thing here is the term $\sum b$. This is $n \times b$. So, we obtain

$$\boxed{\sum y_i - a \sum x_i - nb = 0} \tag{28}$$

This is another equation with two unknowns a and b and otherwise just numbers that are defined by the observed data.

Summarizing, we now have the system of linear equations

$$\sum x_i y_i - a \sum x_i^2 - b \sum x_i = 0 \tag{29}$$

$$\sum y_i - a \sum x_i - nb = 0 \tag{30}$$

These are two linear equations with two unknowns (a and b). Solving this should be straightforward. We multiply the second equation with $\sum x$ to find

$$\sum x_i \sum y_i - a \sum x_i \sum x_i - nb \sum x_i = 0 \tag{31}$$

We also multiply the first equation with n:

$$n\sum x_i y_i - na\sum x_i^2 - nb\sum x_i = 0 \tag{32}$$

Now, the third terms of the two equations coincide, so we can subtract the two to get rid of this middle term:

$$\sum x_i \sum y_i - a\sum x_i \sum x_i - n\sum x_i y_i + na\sum x_i^2 = 0 \tag{33}$$

Keeping the terms with a on the lefthand-side gives

$$-a\sum x_i \sum x_i + na\sum x_i^2 = -\sum x_i \sum y_i + n\sum x_i y_i \tag{34}$$

From this, a can be isolated as follows:

$$a = \frac{-\sum x_i \sum y_i + n\sum x_i y_i}{-\sum x_i \sum x_i + n\sum x_i^2} \tag{35}$$

Once a is known, finding b is easy. We just insert the solution for a in one of the two equations. Take the second one, and rewrite it such that b is isolated:

$$b = \frac{1}{n}\left(\sum y_i - a\sum x_i\right) \tag{36}$$

Often, these results are written in a slightly different way:

$$\boxed{a = \frac{\sum x \sum y/n - \sum xy}{\sum x \sum x/n - \sum x^2}} \tag{37}$$

and

$$\boxed{b = \sum y/n - a\sum x/n} \tag{38}$$

Yet a different notation is adopted once we realize that

$$a = \frac{cov(x,y)}{var(x)} \tag{39}$$

and

$$b = \bar{y} - a\bar{x} \tag{40}$$

2.3 From stationary point to minimum

So, now we have found expressions for the coefficients a and b such that $\varepsilon = f(a,b)$ is a stationary point. Moreover, it is the only stationary point. The only thing that is left to be done is to find out the nature of the stationary point. After all, we're not looking for a maximum or a saddle point, but for a minimum. Theoretically, we could explore this using the second-order partial derivatives. But we will not do so, and just use our intuition to argue that the line determined by $y = ax + b$ is the best fitting line, not the worst fitting line nor a saddle point.

2.4 The results

We can summarize our results in a concise mathematical statement.

Let two samples of n paired observations be given by $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$. The straight line of best fit is defined to be the line $y = ax + b$ that minimizes the sum of squares:

$$\varepsilon = \sum (y_i - (ax_i + b))^2 \tag{41}$$

The coefficients of this straight line are given by

$$a = \frac{\frac{1}{n}\sum x_i \sum y_i - \sum x_i y_i}{\frac{1}{n}\sum x_i \sum x_i - \sum x_i^2} \tag{42}$$

and

$$b = \frac{1}{n}\sum y_i - \frac{a}{n}\sum x_i \tag{43}$$

where all sums run from $i = 1$ to n.

2.5 Vector notation

We will briefly explore how the least-squares problem can be formulated in vector notation. After all, this course advocates the use of vectors and matrices to concisely work with indexed variables.

The observed values for x and y are two series of values: $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$. We can conveniently write these as two vectors $\mathbf{x}$ and $\mathbf{y}$ and formulate the situation as

$$\mathbf{y} \approx a\mathbf{x} + b \tag{44}$$

or equivalently

$$\mathbf{e} = \mathbf{y} - (a\mathbf{x} + b) \tag{45}$$

where $\mathbf{e}$ is the vector of differences between observed and modelled value.

How can we cast the expression $\sum_{i=1}^{n}(e_i)^2$ in matrix form? Notice that

$$\sum_{i=1}^{n}(e_i)^2 = (e_1)^2 + (e_2)^2 + \cdots + (e_n)^2 = e_1 \times e_1 + e_2 \times e_2 + \cdots + e_n \times e_n \tag{46}$$

This looks like the inner product of the vector $\mathbf{e}$ with itself:

$$\sum_{i=1}^{n}(e_i)^2 = \mathbf{e} \cdot \mathbf{e} = \mathbf{e}'\mathbf{e} \tag{47}$$

Now, we have to merge the two coefficients a and b in one vector, say $\mathbf{b}$, by defining

$$\mathbf{b} = \begin{pmatrix} b \\ a \end{pmatrix} \tag{48}$$

The observed values for y can be arranged in a vector of observations of length n:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} \tag{49}$$

Finally, for the observed values for x we need to take care of the constant. We can do so by using a matrix $\mathbf{X}$ with the first column consisting of 1s only:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{x} \end{pmatrix} \tag{50}$$

Using this notation, we can write the model equation as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{51}$$

This leads to the expression of the error as

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b} \tag{52}$$

and for the sum of squared errors as

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})' (\mathbf{y} - \mathbf{X}\mathbf{b}) \tag{53}$$

We will not elaborate the details on how to find the stationary points using vector notation. Suffice to say that it can be done, and that the vector of coefficients $\mathbf{b}$ can be expressed as

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \tag{54}$$

assuming that $\mathbf{X}'\mathbf{X}$ is invertible.

2.6 Link to statistics

The regression model will reappear in the course on statistics. Above, we just had a set of data points, and wished to find the best straight line. When we're doing statistics, we will consider the data as a sample of data, and use a slightly different notation, not a and b for the coefficients, but rather α and β for the theoretical model based on the unknown population, and $\hat{a}$ and $\hat{b}$ as the estimated coefficients from the sample taken.

We will also assess the quality of the regression result using statistical theory. Except for pathological cases, the regression theory will always produce the best-fitting coefficients for a straight line, even if there is no evidence at all for a linear relationship. Statistical theory allows us to consider the goodness of fit, and to test whether the model is doing a better job than just a constant.

3 Back to the example

We return to the example on the housing market, mentioned in the introduction.

As we see, there is a clear pattern that bigger houses are in general more expensive than smaller ones. However, it is also clear that the pattern is not completely straightforward. There are obvious outliers. Perhaps some houses have a very big garden, or are in very bad state of repair. Moreover, it is not even clear if a straight line is the right relationship. Perhaps a quadratic one would fit better. Such considerations will not be made here. Rather we will

Parameter	Value
n	71
$\bar{x}$	110
$\bar{y}$	414×10^3
$var(x)$	3673
$var(y)$	1.696×10^{11}
$cov(x, y)$	2.260×10^7
a	6.152×10^3
b	-2.647×10^5

Table 2: Summary statistics of the data of Table 1.

now concentrate on finding the optimal straight line, i.e. the straight line that best fits the data in a least squares sense.

Table 2 gives the key information derived from the dataset to calculate the regression coefficients. In this example, we will define the regression line as $y = ax + b$, with a the slope of the regression line, and b its intercept with the vertical axis.

Altogether, the least squares method applied to the data yields the regression coefficient $p_0 = -264700$ and $p_a = 6152$. This implies that the equation of the regression line is given by $p = -264700 + 6152 \times a$.

The slope 6152 refers to the increase of price when the floor area increases by 1 m^2. The intercept -264700 indicates the price of a house of 0 m^2. This is a bit problematic, of course, it merely reflects that the model is probably not useful for extremely small houses.

Figure 2 shows the data as in Figure 1, but now with the regression line added.

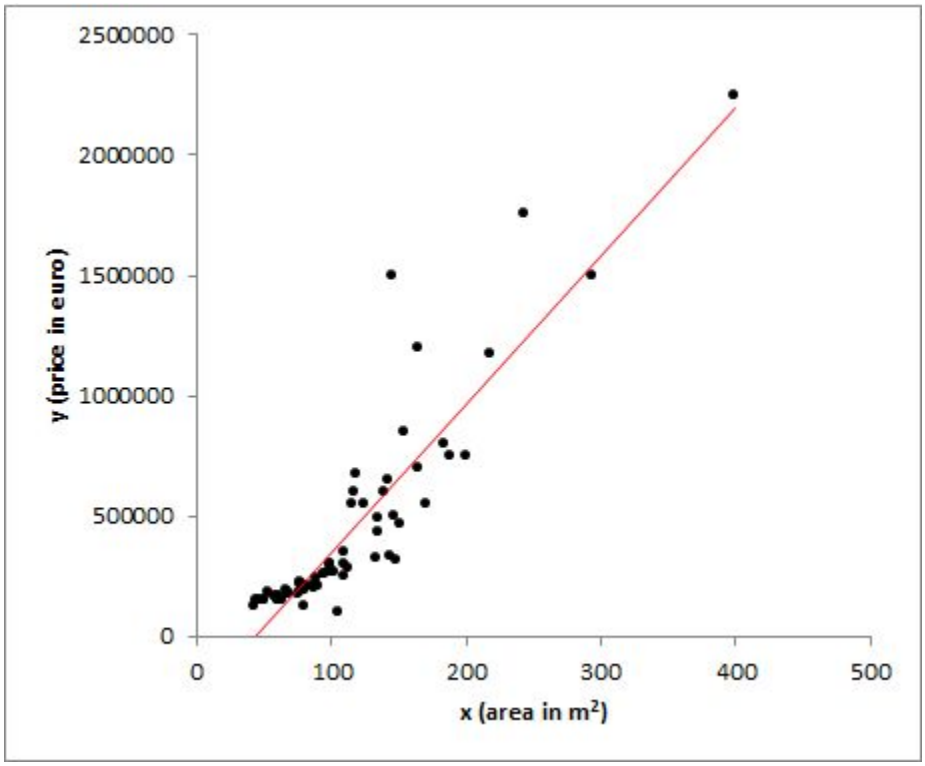

Figure 2: Similar to Figure 1, but with the regression line added.