Curve fitting

Business Mathematics

VU UNIVERSITY AMSTERDAM
Faculty of Economics and Business Administration

CONTENTS

The estimation problem
A criterion for estimation
OLS regression
Old exam question

THE ESTIMATION PROBLEM

Suppose we think there is a relation between two variables

- floor area (x) and price (y) of the housing stock

Suppose we even have a postulated function

- $y = f(x) = ax + b$
- where x is the floor area (m^2), y the price (€)
- and where a and b are coefficients, to be determined

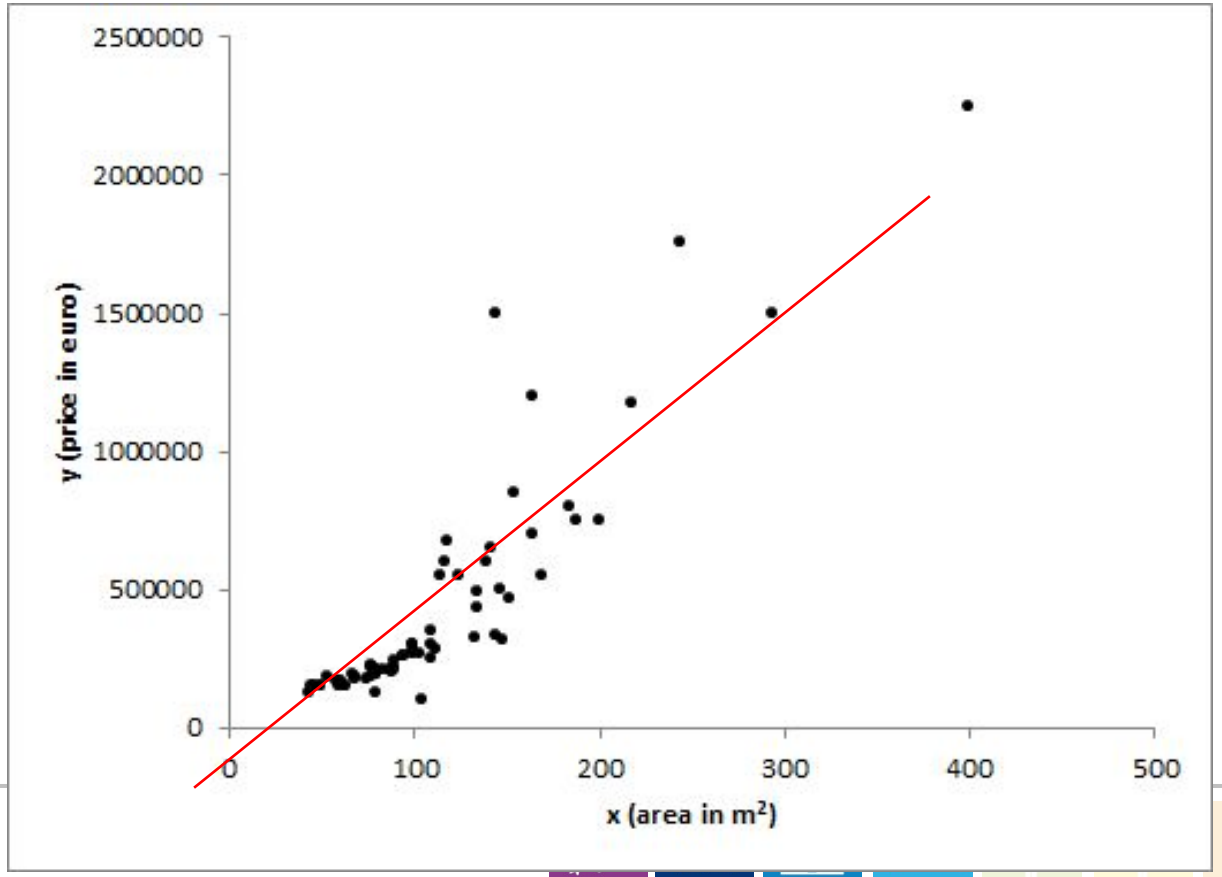
Also suppose we have two data vectors, with floor areas $\mathbf{x}$ and with prices $\mathbf{y}$

Can we estimate the coefficients that achieve the best fit?

THE ESTIMATION PROBLEM

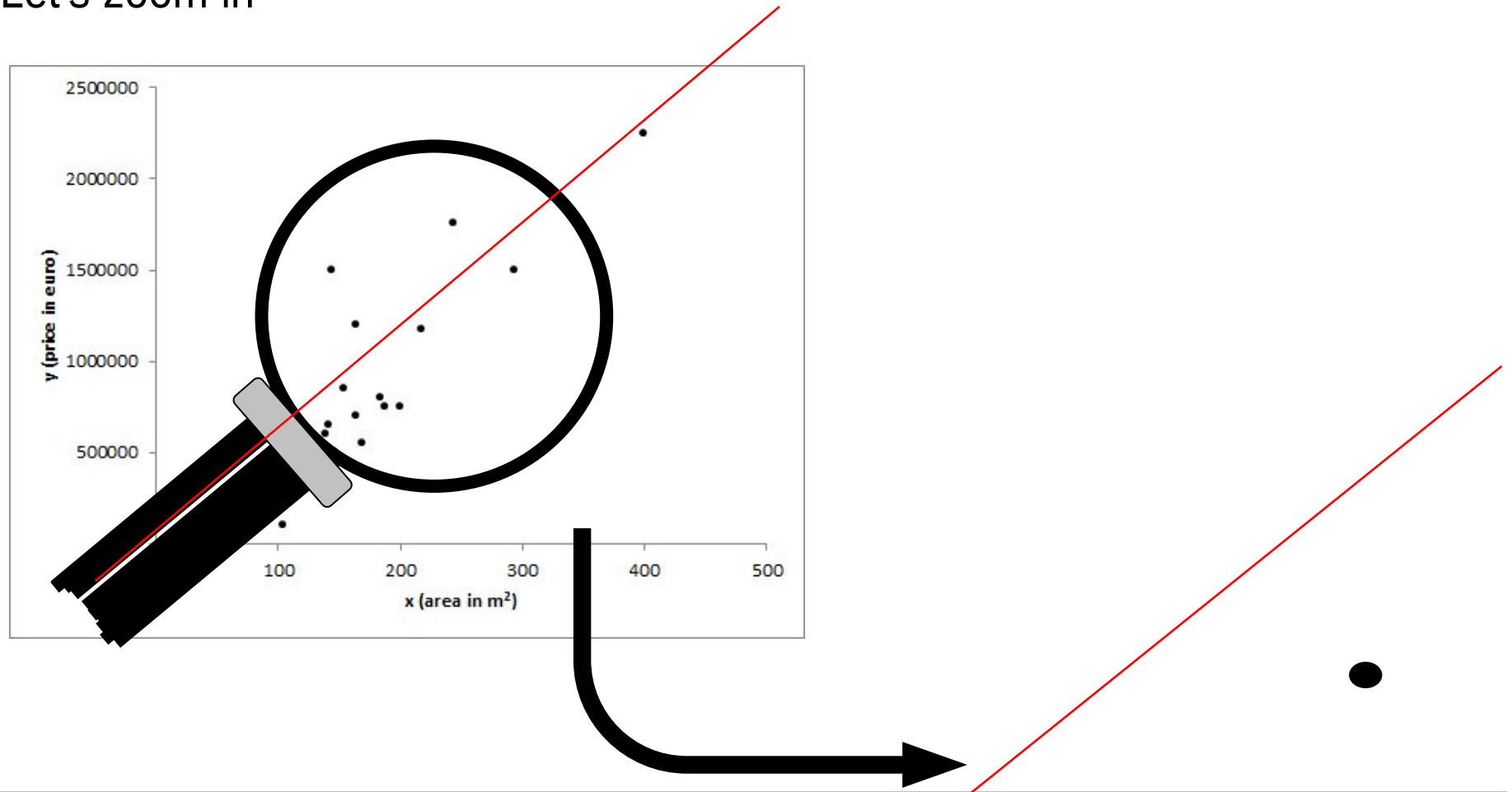
Sample of $n = 71$ house prices (y) and floor areas (x)

Assumed relation: $y = ax + b$ (red line)



THE ESTIMATION PROBLEM

Let's zoom in



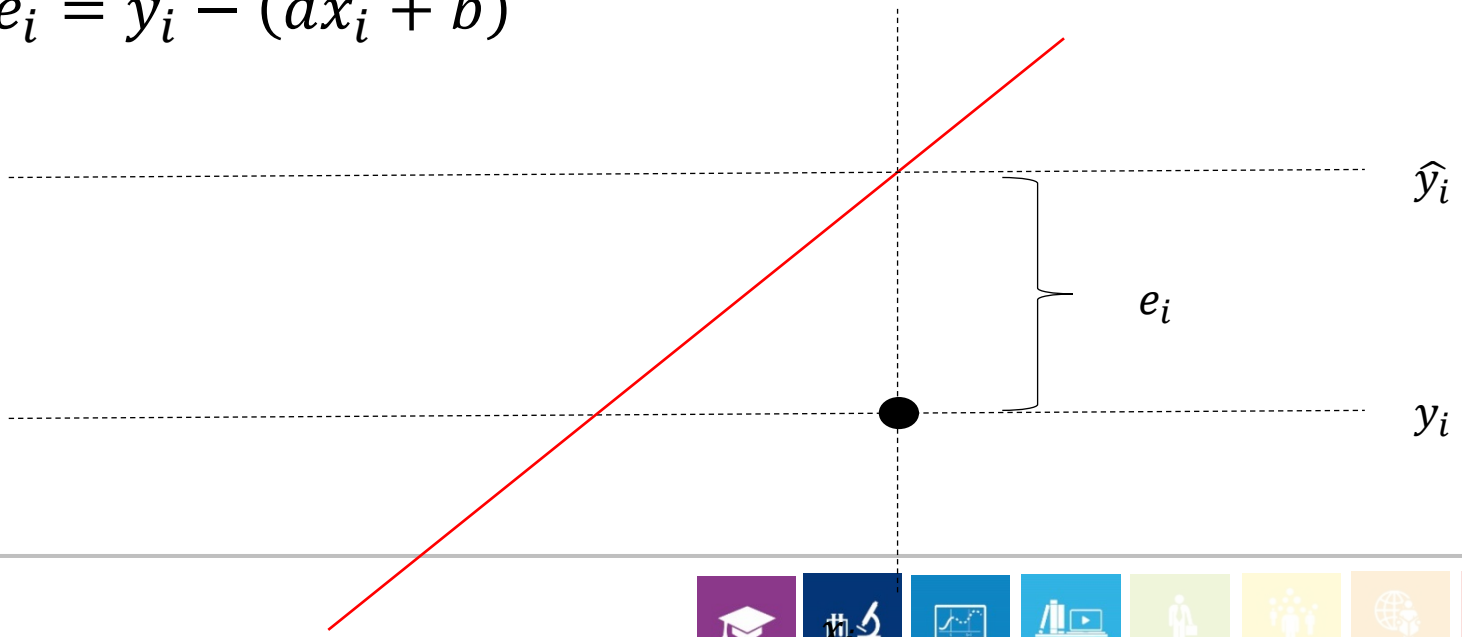
THE ESTIMATION PROBLEM

Each x_i has

- an observed value y_i
- but also a model given value $\hat{y}_i$
- obviously, $\hat{y}_i = ax_i + b$

The "misfit" (error) for observation i is $e_i = y_i - \hat{y}_i$

- so $e_i = y_i - (ax_i + b)$



A CRITRERION FOR ESTIMATION

So the error for one observation i is e_i

But we want to minimize the "total" error

We could wish to minimize the sum of errors: $\sum_{i=1}^{n} e_i$

- that is wrong, because "too low" $(+)$ is then compensated by "too high" $(-)$

So we minimize the sum of square errors: $\sum_{i=1}^{n} (e_i)^2$

This defines a total "misfit" (error) function

$$\mathcal{E} = f(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

actual (observed) value

model (estimated) value

A CRITRERION FOR ESTIMATION

So our problem is:

- find (a, b) such that $\mathcal{E} = f(a, b)$ is a minimum
- so, find the extreme points of $\mathcal{E} = f(a, b)$
- and check which extreme point is the global minimum

Stationary points require:

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

and simultaneously

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

Solving straightforward

- although a bit cumbersome (see extra theory)

Notice the strategy of reversal: normally a and b are fixed coefficients and x and y are variable, but now x and y are given, and a and b are the variables to be decided on.

SOLVING THE ESTIMATION PROBLEM

Take, as an example, the second one

- $\frac{\partial f(a,b)}{\partial b} = \frac{\partial \sum_{i=1}^{n}(y_i-(ax_i+b))^2}{\partial b} = 0$

Can be rewritten as (recall the product rule)

- $\sum_{i=1}^{n} \frac{\partial (y_i-(ax_i+b))^2}{\partial b} = 0$

SOLVING THE ESTIMATION PROBLEM

So (recall the chain rule)

- $\sum_{i=1}^{n} 2(y_i - (ax_i + b)) \times -1 = 0$

So (skip factor -2)

- $\sum_{i=1}^{n} (y_i - (ax_i + b)) = 0$

So (split terms inside $\sum$)

- $\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} ax_i - \sum_{i=1}^{n} b = 0$

So (bring terms without index i in front)

- $\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - nb = 0$

SOLVING THE ESTIMATION PROBLEM

So far, the second requirement for a stationary point

- $\frac{\partial f(a,b)}{\partial b} = 0$

leads to

- $\sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i - nb = 0$

Likewise, the first requirement for a stationary point

- $\frac{\partial f(a,b)}{\partial a} = 0$

leading to

- $\sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} (x_i)^2 - b \sum_{i=1}^{n} x_i = 0$

These are two linear equations in two unknowns

- so, we can easily solve

SOLVING THE ESTIMATION PROBLEM

Solution:

- $a = \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - n \sum_{i=1}^{n} x_i y_i}{\left(\sum_{i=1}^{n} x_i\right)^2 - n \sum_{i=1}^{n} (x_i)^2}$

and

- $b = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i\right)$

This is the only stationary point

It is obviously a minimum point

- why?

SOLVING THE ESTIMATION PROBLEM

What to remember of this?

- not the formulas for a and b
- not how to derive or prove them

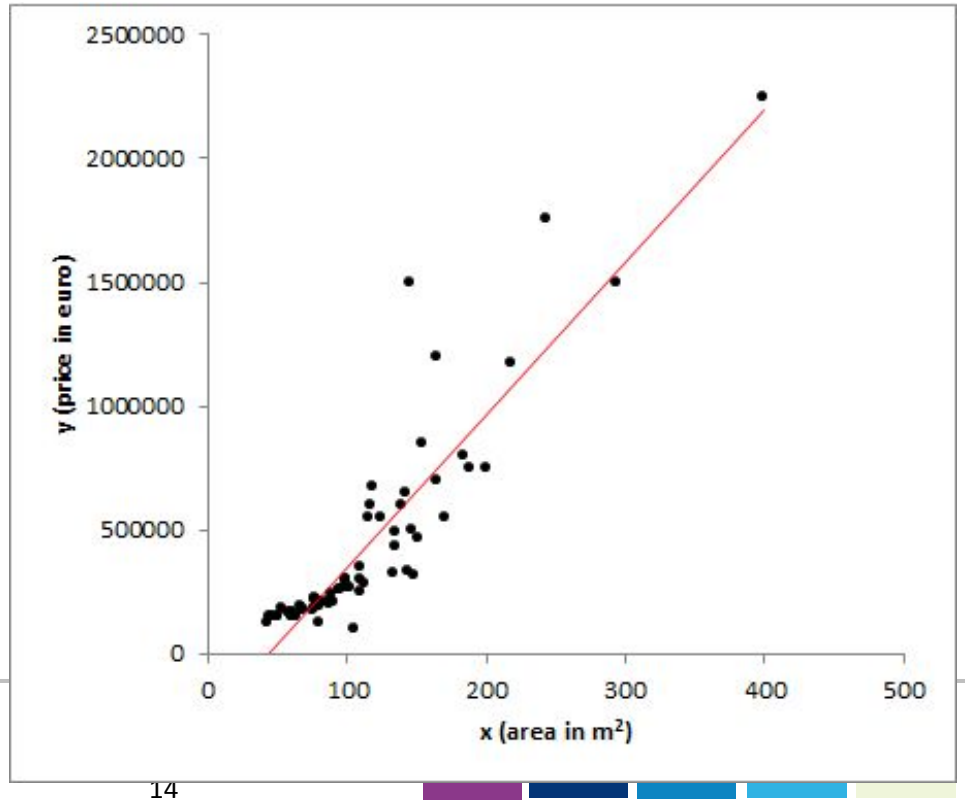
But rather

- that it is possible to derive them
- what the principle is
- what the assumptions are
- how you can combine your skills in summation, differentiation, and solving a systems of equations

OLS REGRESSION

Procedure is known as "(ordinary) least squares" regression (OLS)

- very important in economics and business analysis
- can be generalized to multi-variable and non-linear cases
- will be discussed further in the courses on statistics



OLD EXAM QUESTION

22 October 2014, Q1i

A data set $(\mathbf{x}, \mathbf{y})$, where the vector $\mathbf{x}$ denotes advertisement expenses (in euros) and the vector $\mathbf{y}$ denotes sales (in euros), is modelled by a regression equation $y = ax + b$. Which of the coefficients a and b (or none, or both) will change when we translate the currency from euros into Swedish crowns? (text)




OLD EXAM QUESTION

27 March 2015, Q1k

In a regression problem, a business analyst finds that $\ln y = 38.1 + 4.6 \ln x$. He rewrites this equation as $y = ax^b$. Specify the value of coefficient b. (1 decimal)