Lecture 10: Multilinear Regression

Dr. Roman V Belavkin

BIS3226

Contents

1 Multivariate Data and Models

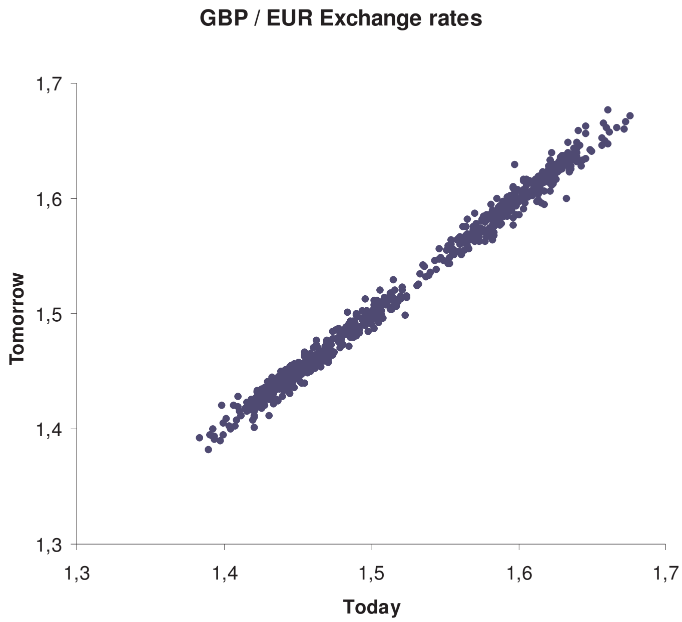
Data-Driven Models

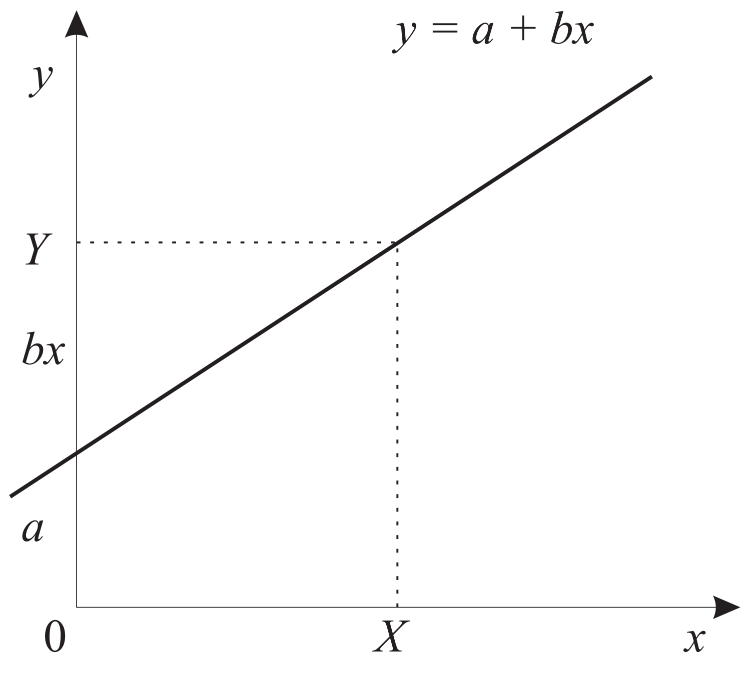
If there are just two variables x (e.g. 'Today') and y (e.g. 'Tomorrow'), then we can use a function $f(x)$ of one variable to model y:

$$\text{Tomorrow} \approx f(\text{Today})$$

Table 1: GBP/EUR rates 4–8 Jan, 2010

Date	Today	Tomorrow
2010/01/04	0.89513	0.89966
2010/01/05	0.89966	0.89934
2010/01/06	0.89934	0.89963
2010/01/07	0.89963	0.89771
2010/01/08	0.89771	?





For example, we can use linear model with parameters a (intercept) and b (slope):

$$y \approx f(x) = a + b\,x$$

Multivariate Data and Models

Case:	Age	Gender	M. Income (£ K)	M. Expenses (£ K)	Home owner	Credit score
1	21	0	2	1	0	3
2	18	1	1	2	0	1
3	50	1	6	2	1	5
4	23	0	3	1	1	4
5	40	1	3	2	0	2

- Data is a 'footprint' of reality.
- Does the credit score depend on a person's income?
- Can we find a function $f(\cdot)$ such that

$$\text{Credit score} = f(\text{Income, Expenses, Age, Gender}, \dots)$$

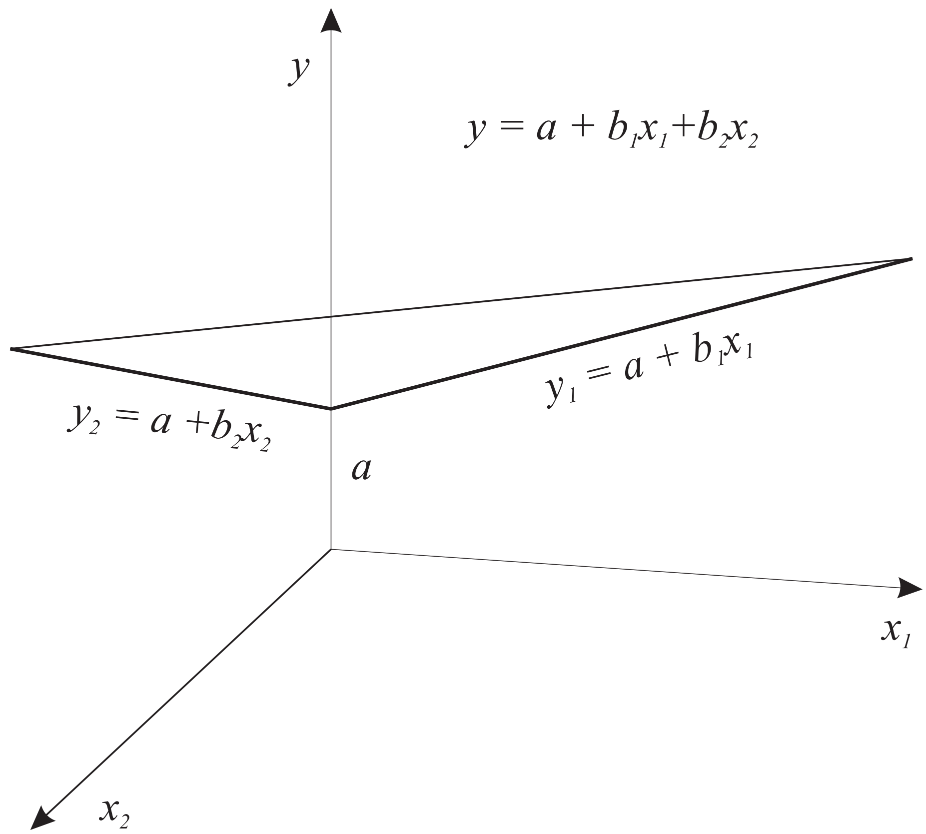
- Data-driven modelling is a search for such functions that represent the dependencies between different variables.

2 Linear Functions of Multiple Variable

Planes and Hyperplanes

- Variable y may depend on several variables x_1, x_2,...
- A linear function $y = f(x_1, x_2)$ of two variables describes a plane, which we can plot on an x, y, z (or x_1, x_2, y) chart.

$$f(x_1, x_2) = a + b_1\, x_1 + b_2\, x_2$$



- A linear function of m variables $y = f(x_1, \ldots, x_m)$ defines a *hyperplane* in an $m+1$ dimensional space.
- It has $m+1$ parameters: one intercept and m 'slopes' called *regression coefficients*.

Multiple Linear Regression

x_1	x_2	$\cdots$	x_m	y
X_{11}	X_{12}	$\cdots$	X_{1m}	Y_1
X_{21}	X_{22}	$\cdots$	X_{2m}	Y_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
X_{n1}	X_{n2}	$\cdots$	X_{nm}	Y_n

- Here, y depends not on one, but on several variables

$$y \approx f(x_1, \ldots, x_m) = a + b_1\, x_1 + \cdots + b_m\, x_m$$

- Thus, we need to find one intercept a and m **regression coefficients** b_1, b_2, …, b_m ('slopes')

3 Example: Credit Score Model

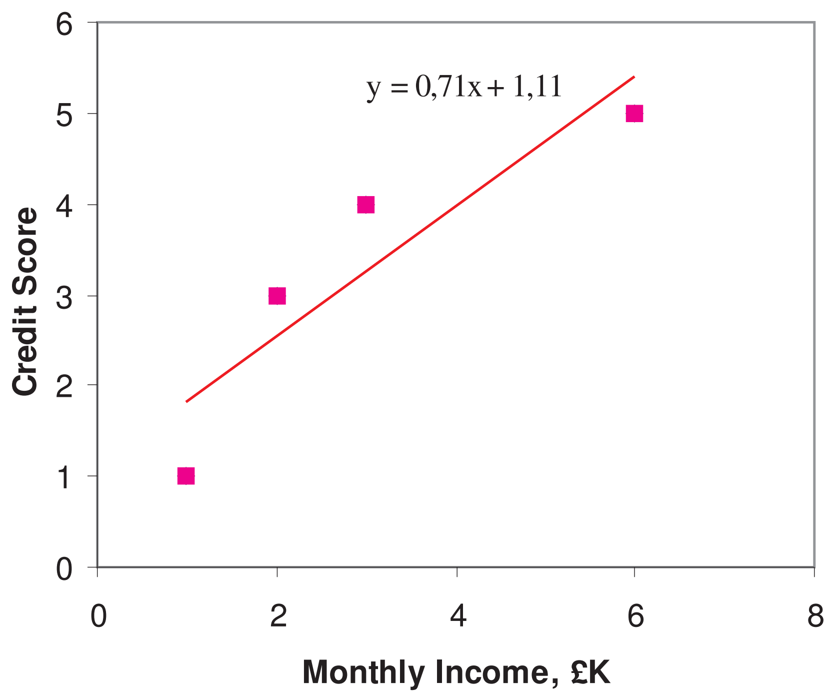
A Simple Model for Credit Score

Monthly Income (£ K)	Credit Score
2	3
1	1
6	5
3	4

- Denote by x the income and by y the credit score.
- Construct a linear model $y \approx a + b\,x$
- We need to find slope (b) and intercept (a) from the data

Solution

$$\begin{aligned}
E\{x\} &= (2+1+6+3)/4 = 3 \\
E\{y\} &= (3+1+5+4)/4 = 3,25 \\
Cov(x,y) &= [(2-3)(3-3,25) + \cdots + (3-3)(4-3,25)]/4 = 2,5 \\
Var(x) &= [(2-3)^2 + \cdots + (3-3)^2]/4 = 3,5
\end{aligned}$$



$$b = \frac{Cov(x,y)}{Var(x)} = 0,71$$

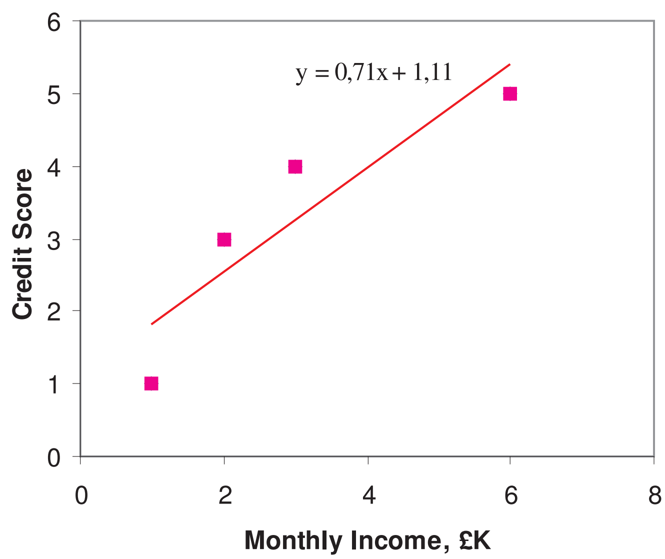
$$a = E\{y\} - b\,E\{x\} = 1,11$$

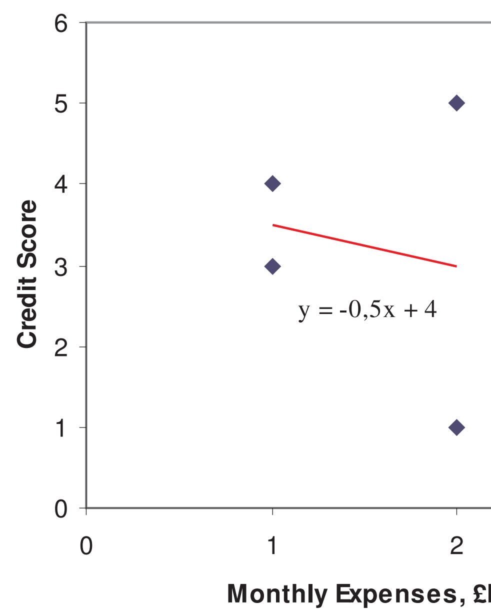
A More Complex Credit Score Model

Monthly Income (£ K)	Monthly Expenses (£ K)	Credit Score
2	1	3
1	2	1
6	2	5
3	1	4

- Denote by x_1 the income, by x_2 expenses and by y the credit score.
- Construct a linear model $y \approx a + b_1\,x_1 + b_2\,x_2$
- We need to find two slopes (b_1, b_2) and one intercept (a)

Approximate Solution





$$b_1 = \frac{Cov(x_1, y)}{Var(x_1)} = 0,71 \qquad b_2 = \frac{Cov(x_2, y)}{Var(x_2)} = -0,5$$

$$a = E\{y\} - b_1\,E\{x_1\} - b_2\,E\{x_2\} = 1,86$$

$$f(x_1, x_2) = 1,86 + 0,71\,x_1 - 0,5\,x_2$$

4 Conclusions

Slope, Correlation and Dependency

- Recall that correlation is

$$Corr(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)Var(y)}}$$

- Thus, we can compute the slope as

$$b = \frac{Cov(x, y)}{Var(x)} = Corr(x, y)\sqrt{\frac{Var(y)}{Var(x)}}$$

- Positive correlation means positive slope $b > 0$
- Negative correlation means negative slope $b < 0$ (anticorrelated)
- Zero correlation means zero slope $b = 0$ (uncorrelated)

Remark 1. *In multiple linear regression, the regression coefficients* $b_1, \ldots, b_m$ *represent linear dependency between multiple variables, and they are related to multiple correlations.*

Advantages of Linear Models

- Given data, they are easy to implement
- Multiple linear mean-square regression is a standard feature of many analytical tools
- If there is a strong linearity in the data, then the mean-square regression can always find the optimal model
- Such a model can be used to explain and understand the dependencies in data (i.e. using slopes or correlations)
- The model can be used for prediction and 'what-if' analysis.

Limitations of Linear Models

- There can be no significant linear dependency
- Linear models cannot account for nonlinear effects
- Mean-square error (quadratic cost) is very sensitive to *outliers* (unusual cases)
- It is much more difficult to find linear models optimising non-quadratic cost functions (e.g. an absolute error $|y - f(x)|$)

Summary

- Models are simplified representations of reality
- The unexplained part of reality results in an error of the model
- Linear functions, defining lines, planes and hyperplanes, can be used to construct the simplest data-driven models
- Linear mean-square regression is a standard method of computing such models
- Linear models can reveal linear dependencies in data