Regression

Contents

1. Linear regression
2. Multiple regression
3. Logistic regression

1. Linear Regression

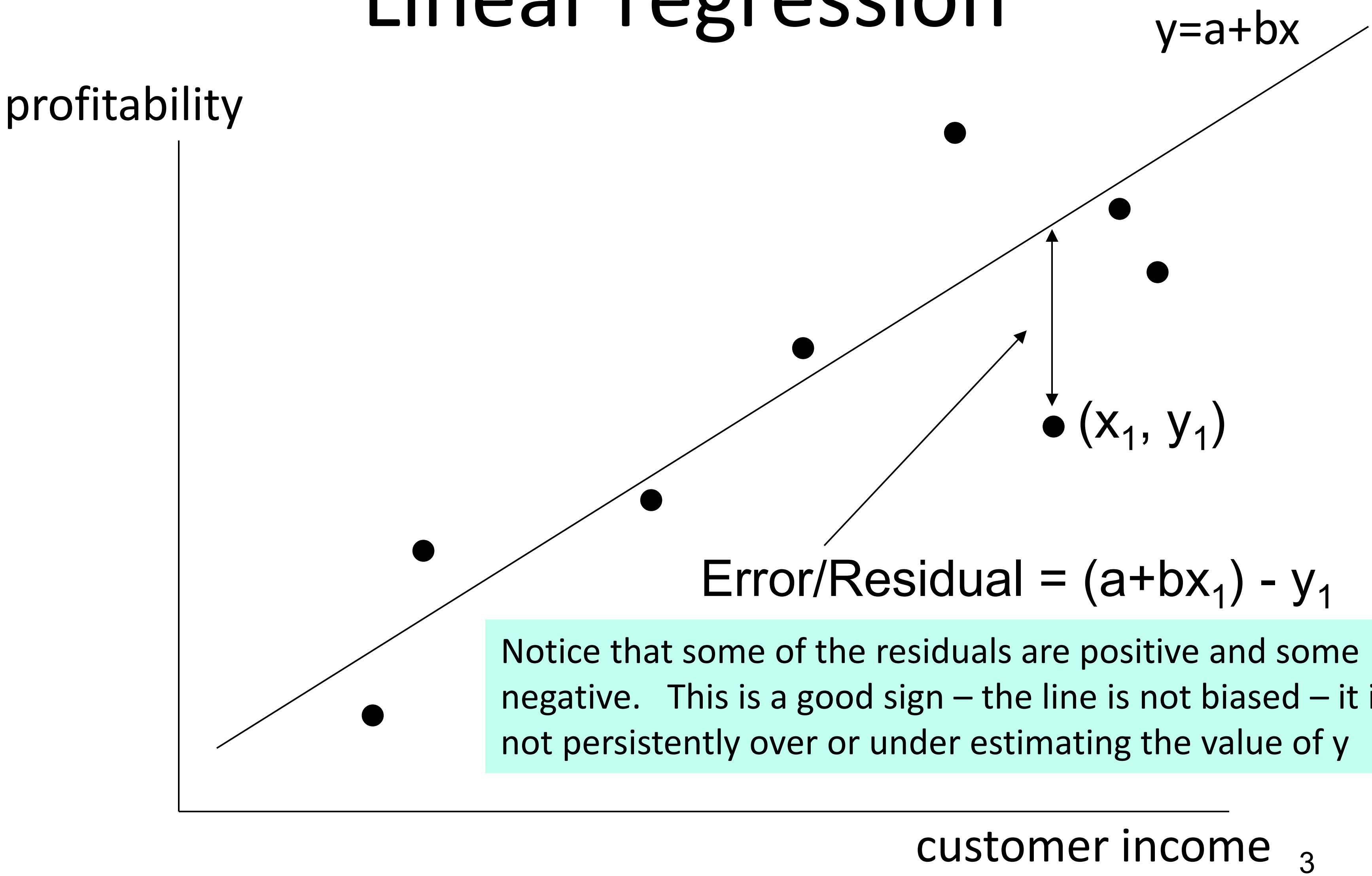
- Linear regression attempts to fit a straight line through a plot of *numeric* data
- Suppose that we have *known* numeric data points

$$(x_1, y_1), (x_2, y_2), \; . \; . \; . \;, (x_n, y_n)$$

These points are in effect our training set

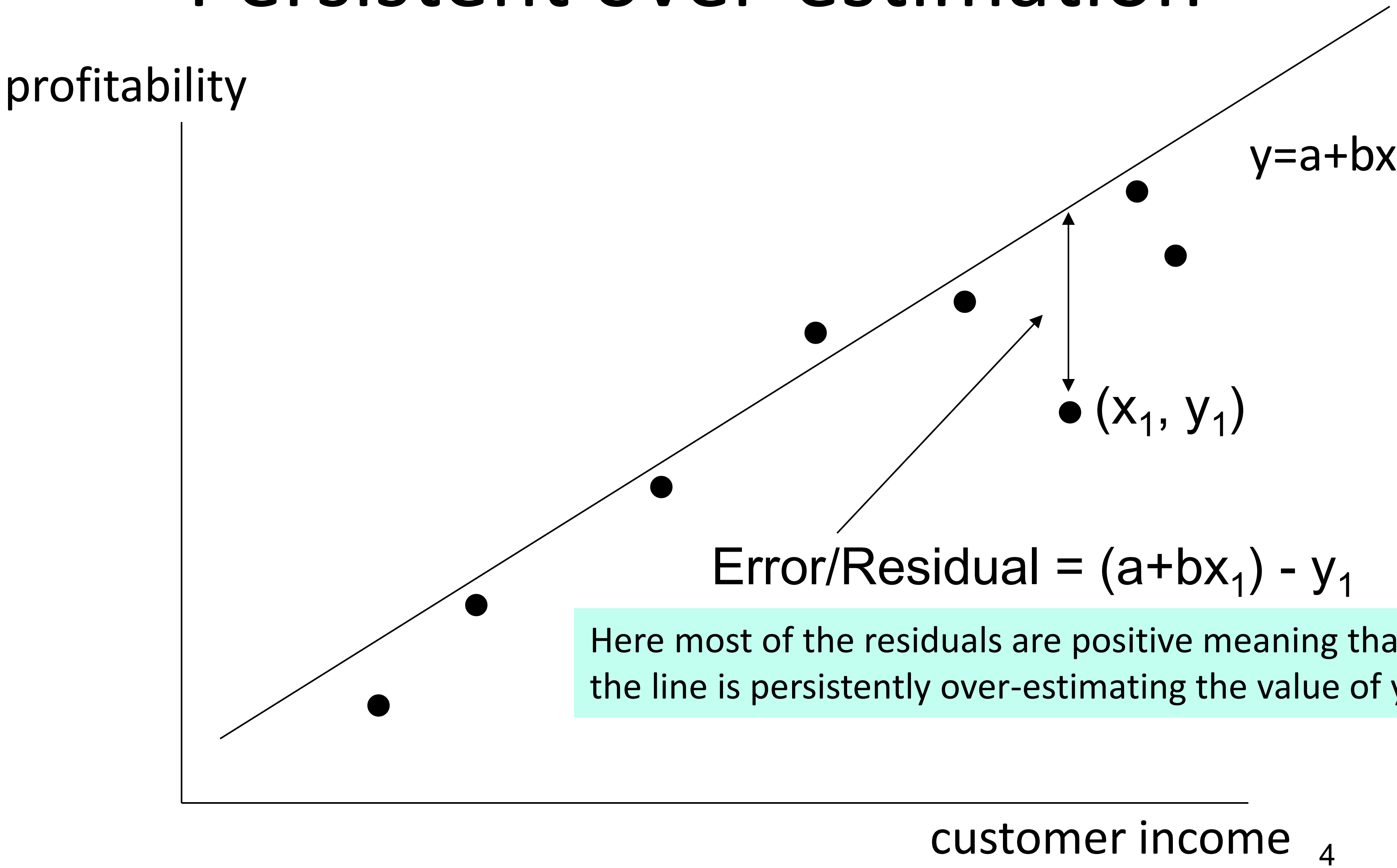
- A straight line has formula $y = a + bx$

Linear regression



Notice that some of the residuals are positive and some negative.   This is a good sign – the line is not biased – it is not persistently over or under estimating the value of y

Persistent over-estimation



Here most of the residuals are positive meaning that the line is persistently over-estimating the value of y

Residuals and the best fit

- In ordinary least squares regression the "best fit" is provided by the line that minimizes

$$R_1^2 + R_2^2 + \ldots + R_n^2$$

where R_i is the ith residual

$$R_i = (a+bx_i) - y_i$$

Calculating the best fit

A
S
I
D
E

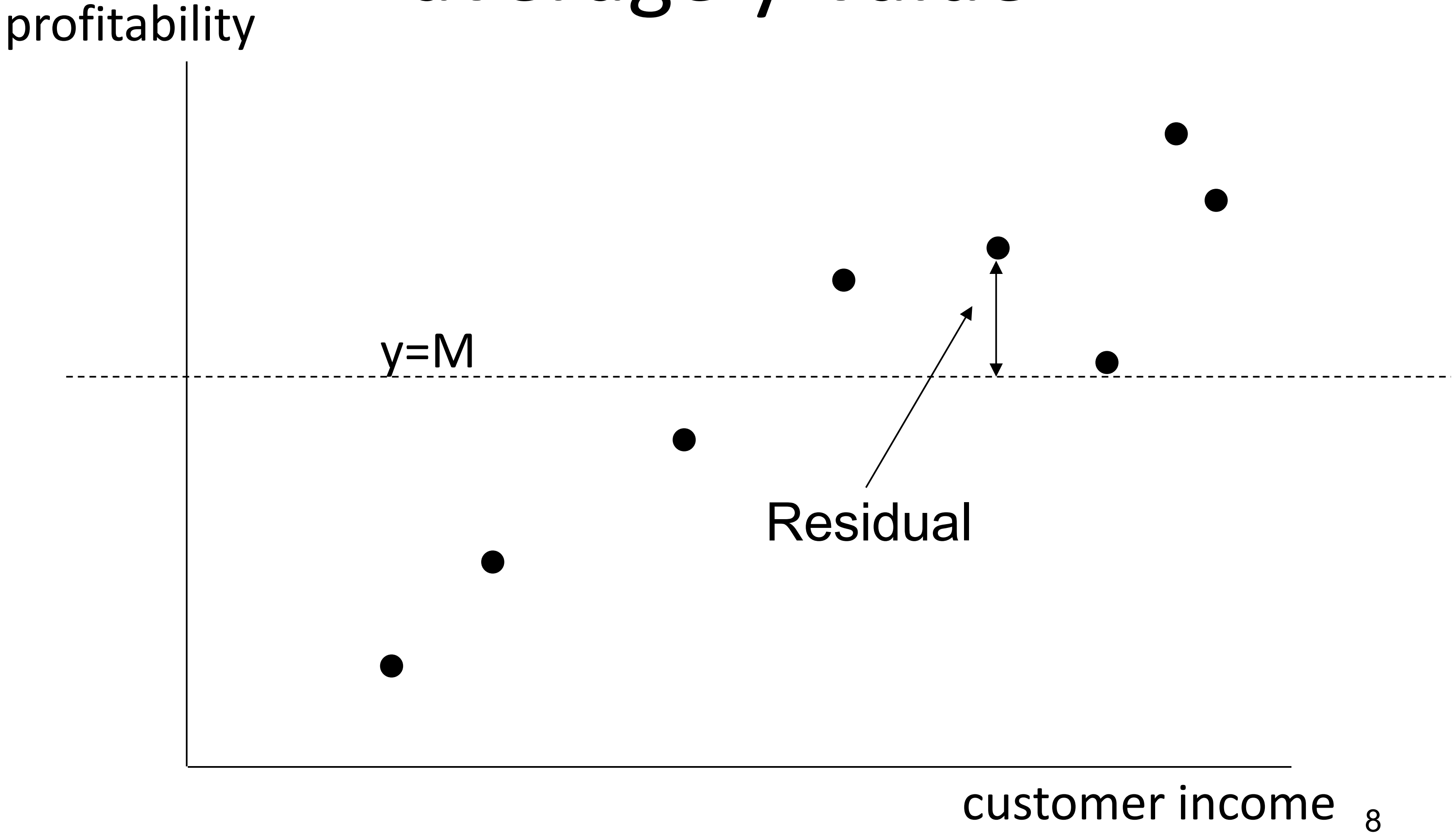
- Let $f = R_1^2 + R_2^2 + . . .\ + R_n^2$
- The expression f contains 2 unknowns (a and b): we want to find the values of a and b that minimize f
- When a curve finds its minimum value – its gradient is 0. This yields two equations in two unknowns

$$\partial f/\partial a = 0 \text{ and } \partial f/\partial b = 0$$

How good is the best fit?

- Let M be the average y-value of the data points
- Consider the most primitive prediction of y values, i.e., $y=M$
- Given that this prediction is so primitive, we'd expect its residuals to be *much* higher

A trivial prediction using the average y-value



Pearson's correlation coefficient

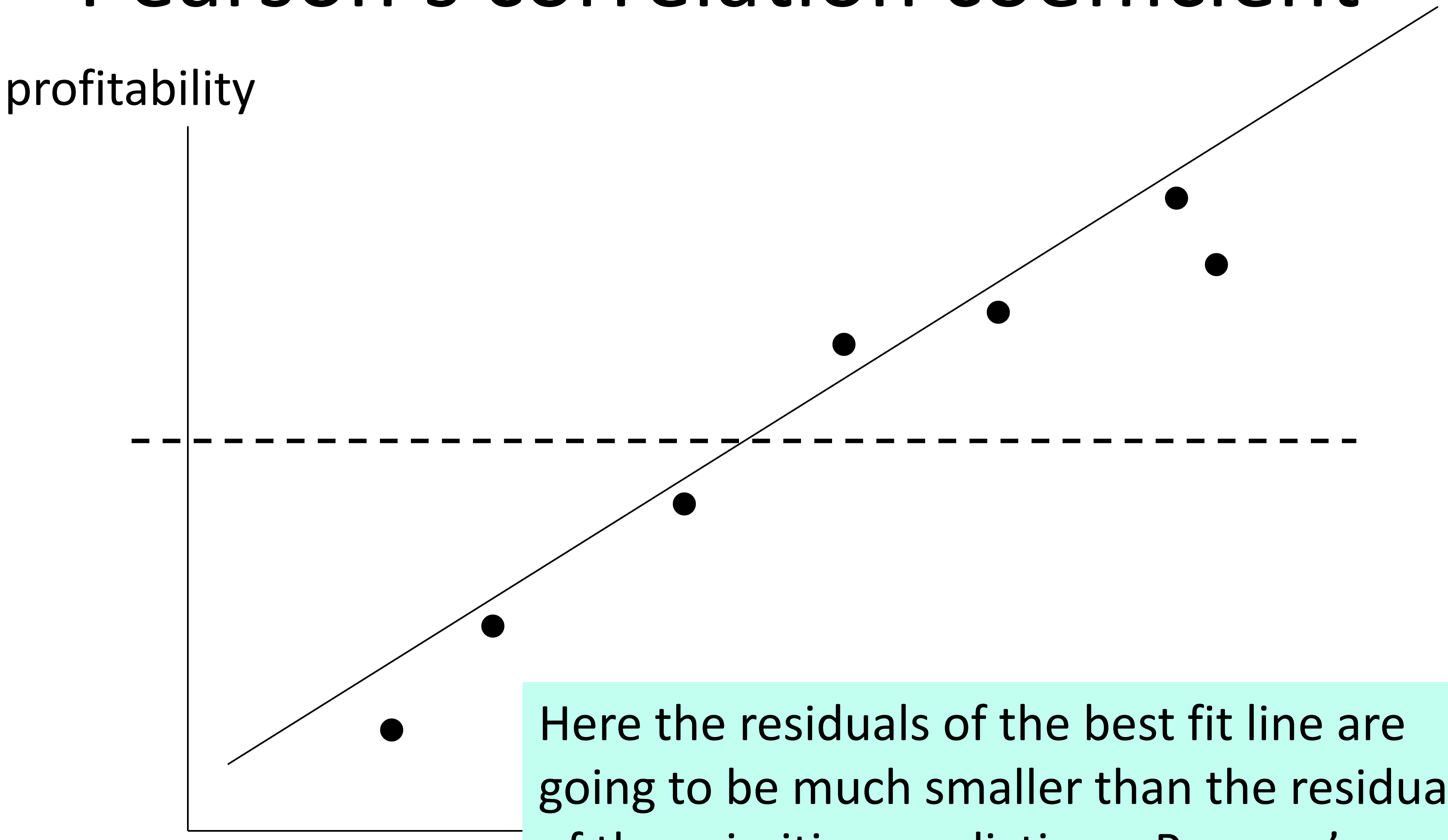
is given by

$$1 - \frac{R_1^2 + R_2^2 + \ldots + R_n^2}{P_1^2 + P_2^2 + \ldots + P_n^2}$$

where P_1, P_2, …, P_n are the residuals from the primitive prediction line $y=M$

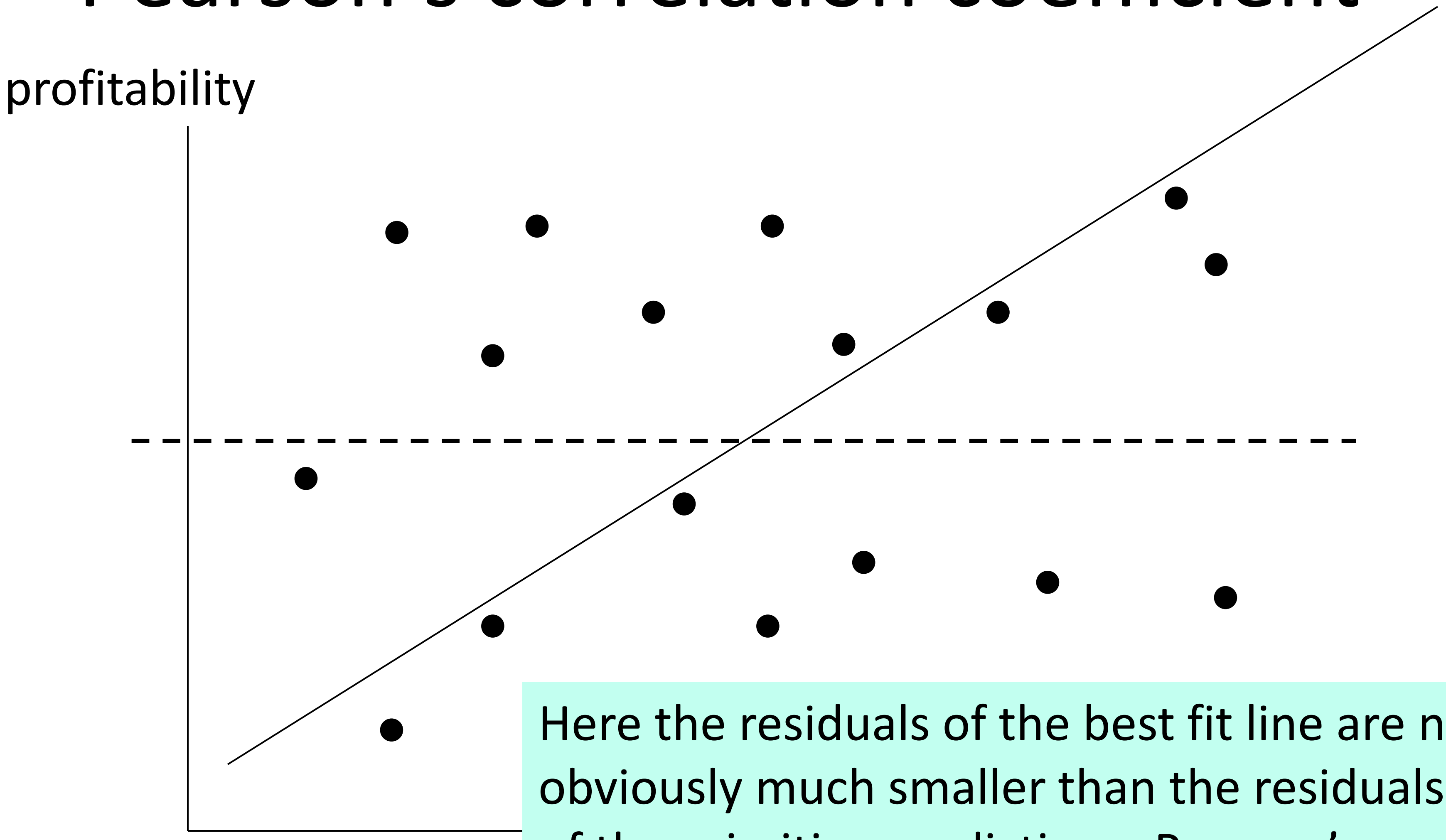
Actually a slight variant of the true Pearson coefficient

Pearson’s correlation coefficient



Here the residuals of the best fit line are going to be much smaller than the residuals of the primitive prediction. Pearson’s coefficient will therefore be close to 1.

Pearson's correlation coefficient

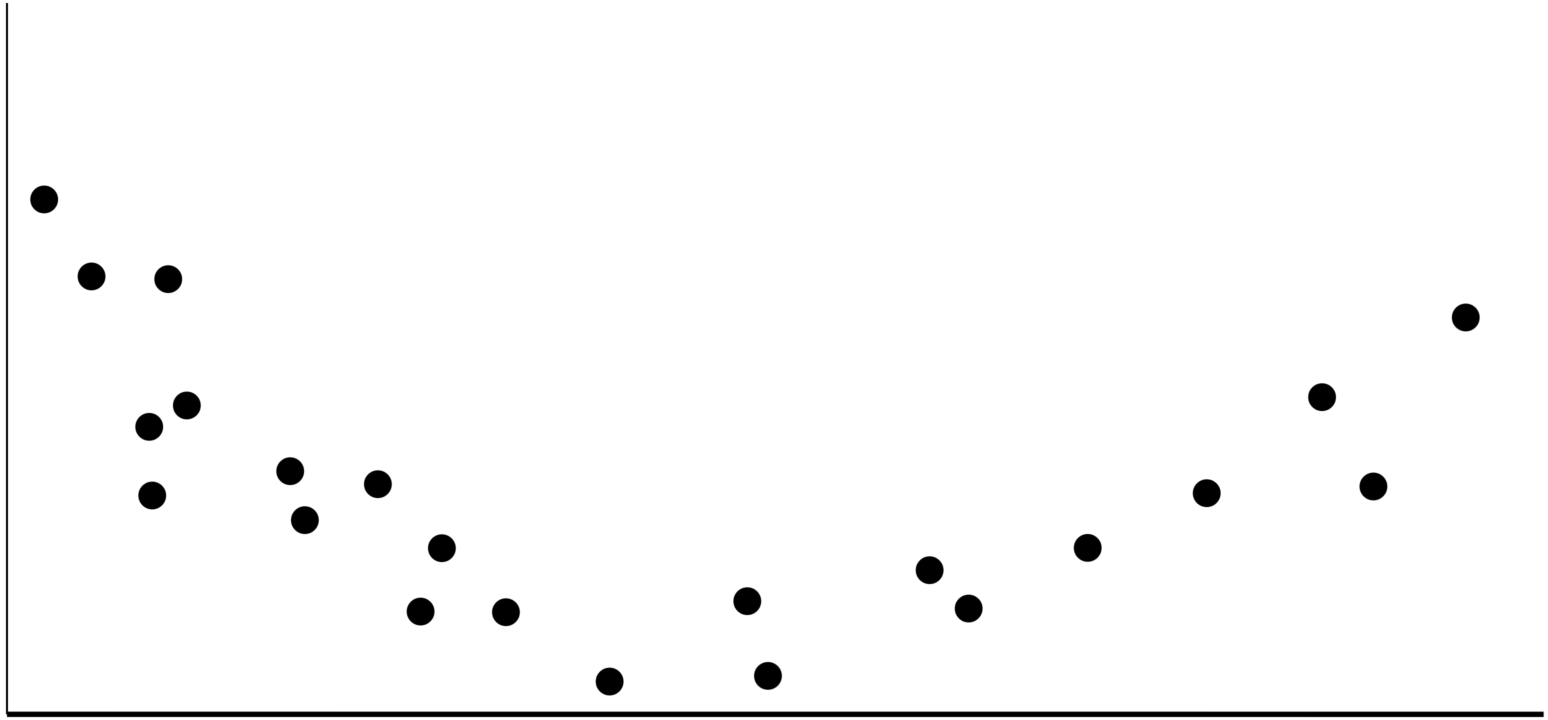


Here the residuals of the best fit line are not obviously much smaller than the residuals of the primitive prediction. Pearson's coefficient will therefore be close to 0.

Linear Regression

- Sensitive to the presence of outliers
- Only works well with linear data
  - Most data is not linear. Non-linear regression may help, but won't handle all cases
- Can produce the full range of y-values
  - This may not be appropriate
- Finds *global* patterns in the data
  - Not good for finding local patterns

The data is not inherently linear



Linear regression assumes that there is a single linear relationship that holds across the range of data values. But here for example there is not a single global pattern – two effects are at play. We could therefore create two models – *piecewise* regression.

The linear model might be inadequate

Here the residuals grow as customer income gets larger … the model is not capturing the full characteristics of the data

2. Multiple regression

- With several inputs (say x_1 and x_2) the equation for a linear model is

$$y = a + b_1x_1 + b_2x_2$$

- If say $b_1>0$, then an increase in x_1 causes an increase in y (& vice versa); the larger the value of $|b_1|$, the more sensitive y is to changes in x_1
  - provided x_1 and x_2 are independent

3. Logistic regression

- Linear regression is not applicable when the target variable takes on a limited subset of values
- Suppose for example we wish to predict a probability
  - Probabilities range from 0 to 1

Converting probabilities to odds

- The *odds* of something happening is given by the formula

  $$\text{odds} = p/(1-p)$$

  where p is its probability
- Whereas p ranges from 0 to 1, the odds range from 0 to ∞
- Taking ln (natural logarithms) yields values that can take any value (from -∞ to ∞)

probabilities vs odds

p	0	0.01	0.1	0.3	0.5	0.7	0.9	0.99	1
odds= p/(1-p)	0	1/99	1/9	3/7	1	7/3	9	99	∞
ln(p/(1-p))	-∞	-4.6	-2.2	-0.85	0	0.85	2.2	4.6	∞

Applying linear regression

We can apply linear regression

$$\ln(p/(1-p)) = a + bx$$

to find values for the constants a and b, and then rearranging yields

$$p = 1/(1+\exp(-a-bx))$$

Fitting the data

x =	0-20	20-40	40-60	60-80	80-100	100-120	120-140
Defaulters	2	5	10	16	23	25	31
Non-defaulters	25	21	18	8	7	4	1
Probability (p) of default	2/27	5/26	10/28	16/24	23/30	25/29	31/32
ln(p/(1-p))	-2.5	-1.4	-0.6	0.7	1.2	1.8	3.4

Fitting the data