CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

MODULE 2: REGRESSION AND CLASSIFICATION

DAT405/DIT405, 2020-2021, STUDY PERIOD 2

Lecture 3

Regression

Module 2 - Learning objectives

- give examples of how machine learning (ML) are applied in data science and AI
- use appropriate programming libraries and techniques to implement basic transformations, visualizations and analyses of example data
- identify appropriate types of analysis problems for some concrete data science applications
- discuss advantages and drawbacks of different types of approaches and models within data science and AI.
- reflect on inherent limitations of data science methods and how the misuse of statistical techniques can lead to dubious conclusions
- show a reflective attitude in all learning

Arrange these tasks into groups:

A. Predict whether a manuscript will be a bestseller novel
B. Find texts in a corpus that probably have the same author
C. Predict what will be a company’s share price tomorrow
D. Predict which companies’ shares will go up tomorrow
E. Find evolutionary relationships among a set of species
F. Determine whether a news article is fake
G. Find ”communities” of users of a music service who have similar tastes
H. Predict the population of Gothenburg in 2030
I. Identify sets of genes that are “switched on” in similar conditions
J. Predict a patient’s blood pressure one hour from now
K. Diagnose a genetic disorder based on facial shape
L. Identify whether a picture is of a cat or a dog
M. Predict how long your journey home will take today
N. Arrange a set of data science tasks into groups

A. Predict whether a manuscript will be a bestseller novel

Regression	Predicting a numerical quantity
Classification	Assigning a label from a discrete set of possibilities
Clustering	Grouping items by similarity

B. Find texts in a corpus that probably have the same author

Regression	Predicting a numerical quantity
Classification	Assigning a label from a discrete set of possibilities A
Clustering	Grouping items by similarity

C. Predict what will be a company's share price tomorrow

Regression	Predicting a numerical quantity
Classification	Assigning a label from a discrete set of possibilities A
Clustering	Grouping items by similarity B

D. Predict which companies' shares will go up tomorrow

Regression	Predicting a numerical quantity C
Classification	Assigning a label from a discrete set of possibilities A
Clustering	Grouping items by similarity B

E. Find evolutionary relationships among a set of species

Regression	Predicting a numerical quantity C
Classification	Assigning a label from a discrete set of possibilities A D
Clustering	Grouping items by similarity B

F. Determine whether a news article is fake

Regression Predicting a numerical quantity

C

Classification Assigning a label from a discrete set of possibilities

A D

Clustering Grouping items by similarity

B E

G. Find ”communities” of users of a music service who have similar tastes

Regression Predicting a numerical quantity

C

Classification Assigning a label from a discrete set of possibilities

A D F

Clustering Grouping items by similarity

B E

H. Predict the population of Gothenburg in 2030

Regression Predicting a numerical quantity

C

Classification Assigning a label from a discrete set of possibilities

A D F

Clustering Grouping items by similarity

B E G

I. Identify sets of genes that are "switched on" in similar conditions

Regression Predicting a numerical quantity

C H

Classification Assigning a label from a discrete set of possibilities

A D F

Clustering Grouping items by similarity

B E G

J. Predict a patient's blood pressure one hour from now

Regression	Predicting a numerical quantity C H
Classification	Assigning a label from a discrete set of possibilities A D F
Clustering	Grouping items by similarity B E G I

K. Diagnose a genetic disorder based on facial shape

Regression

Predicting a numerical quantity

C H J

Classification

Assigning a label from a discrete set of possibilities

A D F

Clustering

Grouping items by similarity

B E G I

L. Identify whether a picture is of a cat or a dog

Regression	Predicting a numerical quantity C H J
Classification	Assigning a label from a discrete set of possibilities A D F K
Clustering	Grouping items by similarity B E G I

M. Predict how long your journey home will take today

Regression	Predicting a numerical quantity C H J
Classification	Assigning a label from a discrete set of possibilities A D F K L
Clustering	Grouping items by similarity B E G I

N. Arrange a set of data science tasks into groups

Regression

Predicting a numerical quantity

C H J M

Classification

Assigning a label from a discrete set of possibilities

A D F K L

Clustering

Grouping items by similarity

B E G I

Core data science tasks

- Regression
  - Predicting a numerical quantity
- Classification
  - Assigning a label from a finite set of possibilities
- Clustering
  - Grouping items by similarity

Topics

- Linear regression
- Residuals
- Covariance
- Correlation
- Multidimensional regression
- Regularization
- Applications of linear regression
- Using linear regression

Linear regression

Luis Serrano video

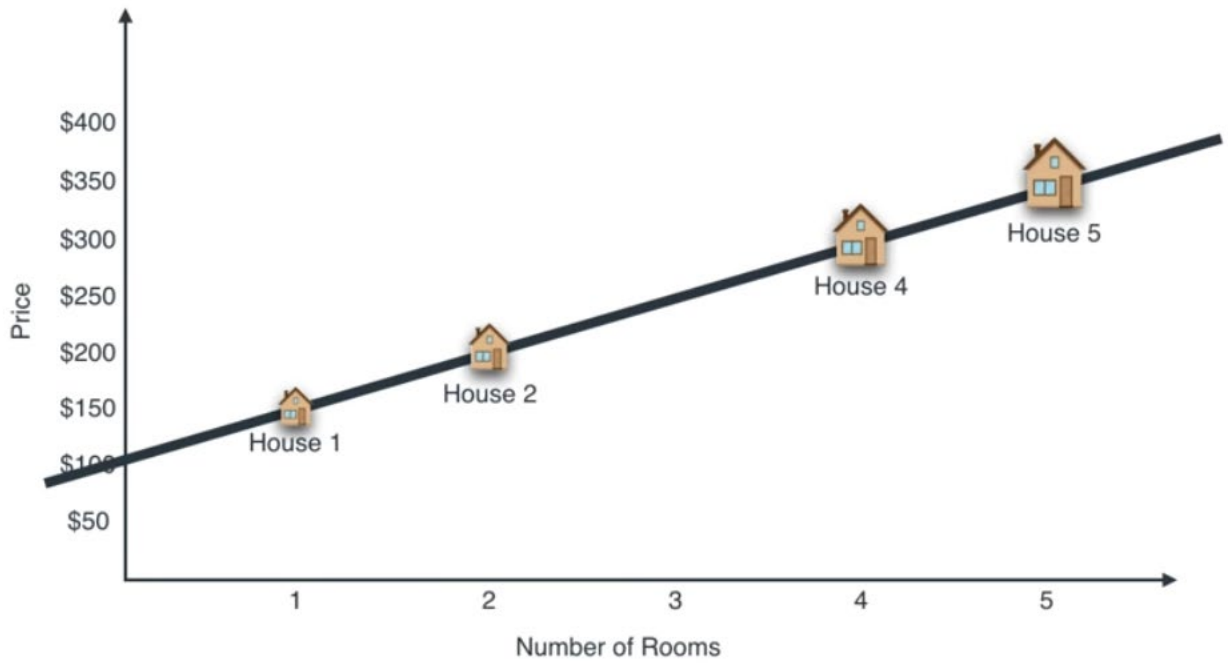
Estimating house prices

What price is reasonable for a house with 3 rooms (interpolation) or 6 rooms (extrapolation)?

Estimating house prices

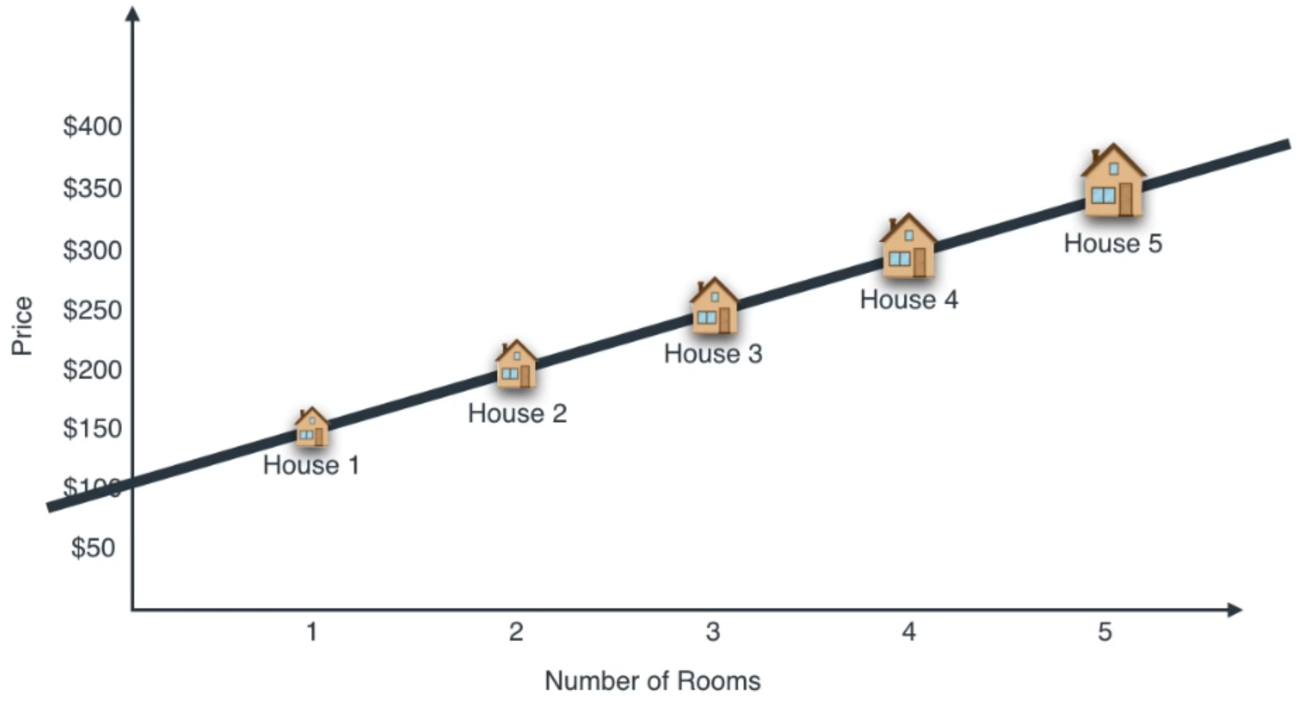
Put the houses in a diagram.

Estimating house prices



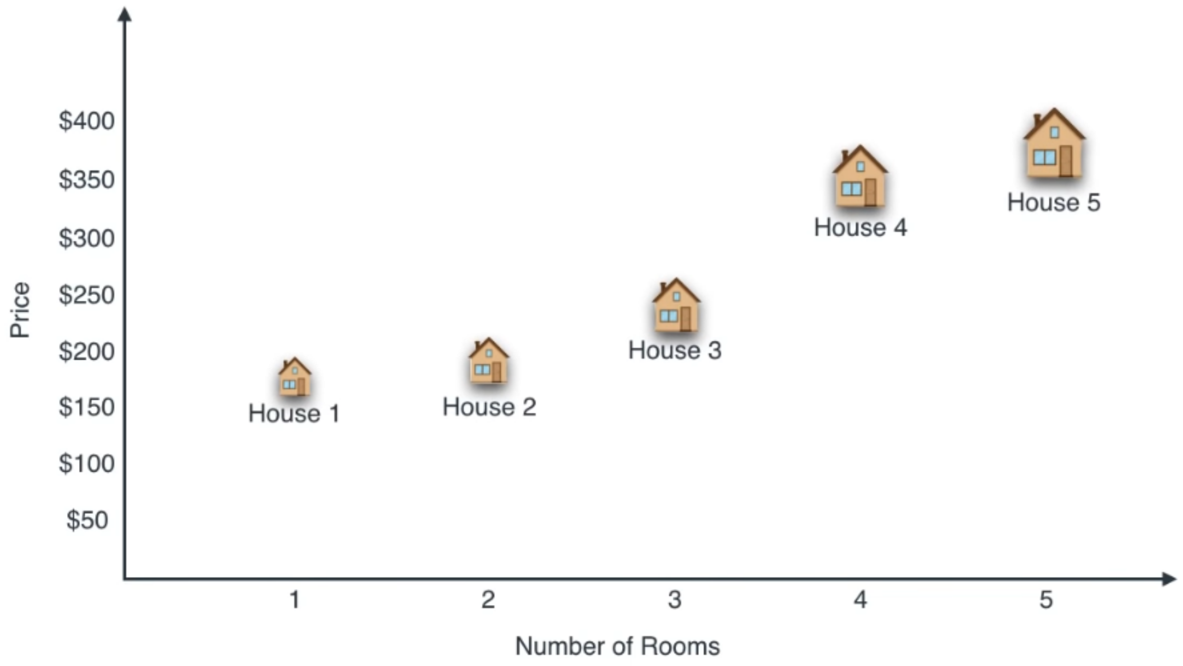
Draw a line through the points (possible in this case).

Estimating house prices



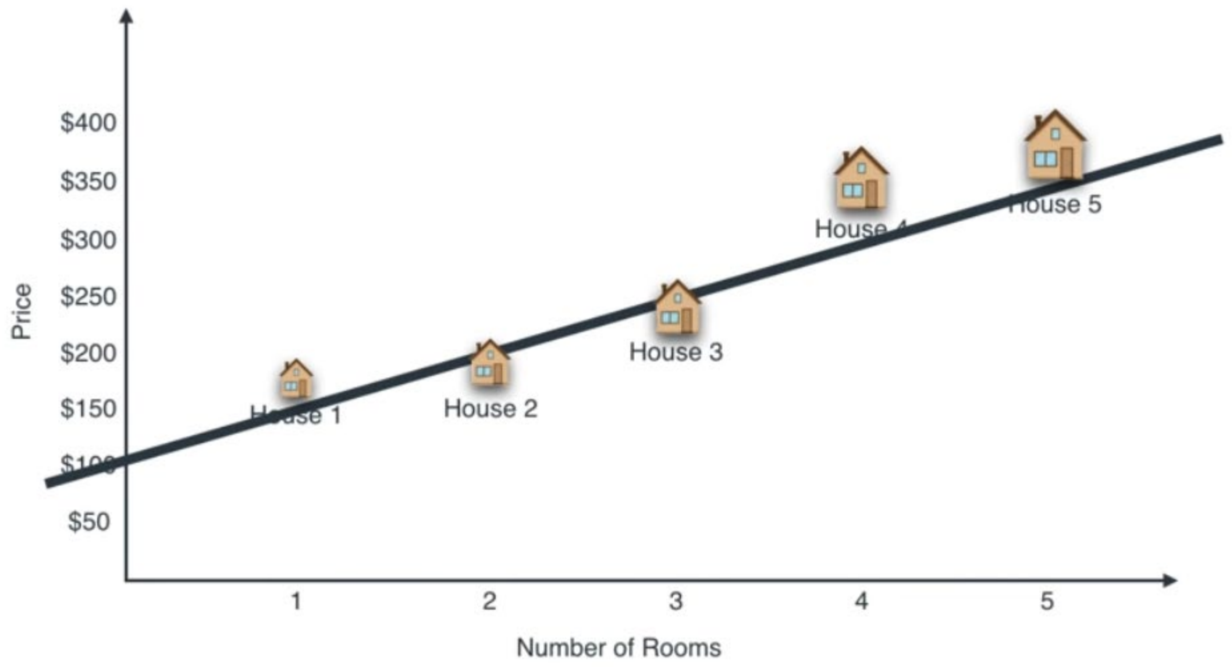
Put house 3 over the line. Read off the price.

Estimating house prices



A more realistic situation.

Estimating house prices

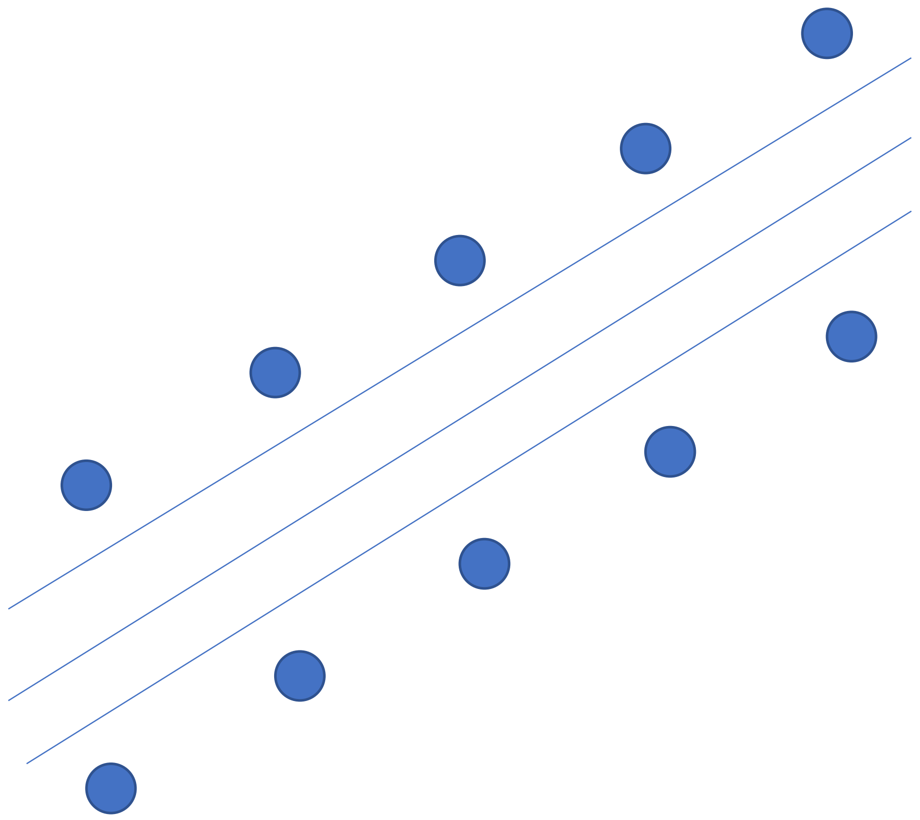


Let us try to fit a line to our data.

Goal

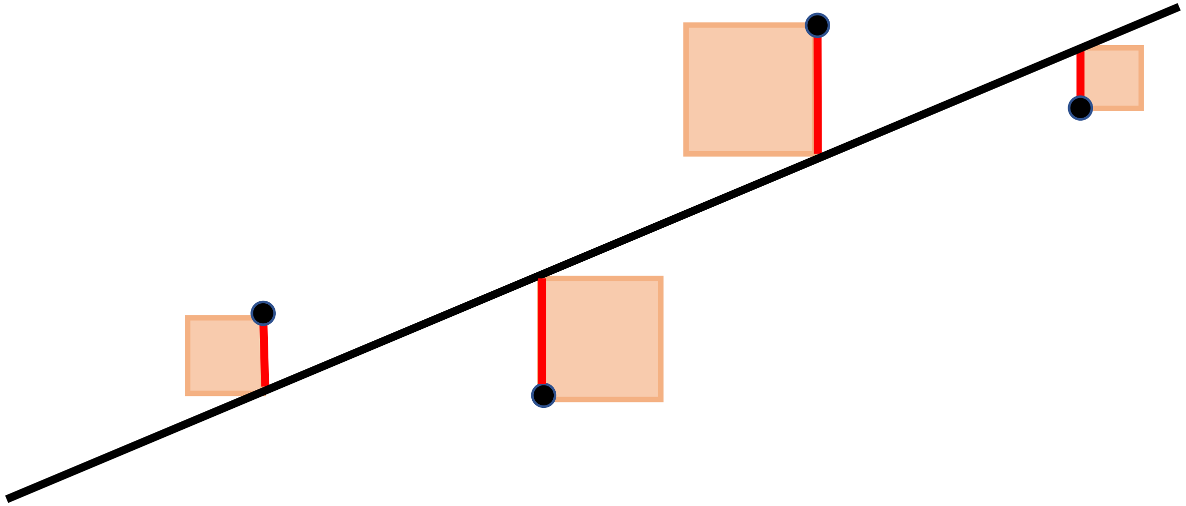
- We want to fit a line $f(x) = k \cdot x + m$ to our data
- We want to select k and m so that the total error is minimal
- But how do we define the error?

How do we measure the error?



These three lines are equally good if our error is the sum of the absolute errors. But we prefer the line in the middle…

How do we measure the error?



Instead we will measure error of the line as the sum of the squared errors of the datapoints. This is a somewhat arbitrary choice that is easy to work with. We want our line to minimize this sum.

Computing the error

correct value

predicted value

$$Error = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$= \sum_{i=1}^{n} (y_i - (k \cdot x_i + m))^2$$

A function with two variables: k and m. The x_i and y_i are constants!

$$= \sum_{i=1}^{n} {y_i}^2 - 2 \cdot y_i(k \cdot x_i + m) + (k \cdot x_i + m)^2$$

$$= a \cdot k^2 + b \cdot k \cdot m + c \cdot m^2 + d$$

Here a, b, c, d are constants! So the error function is a quadratic function in the variables k and m.

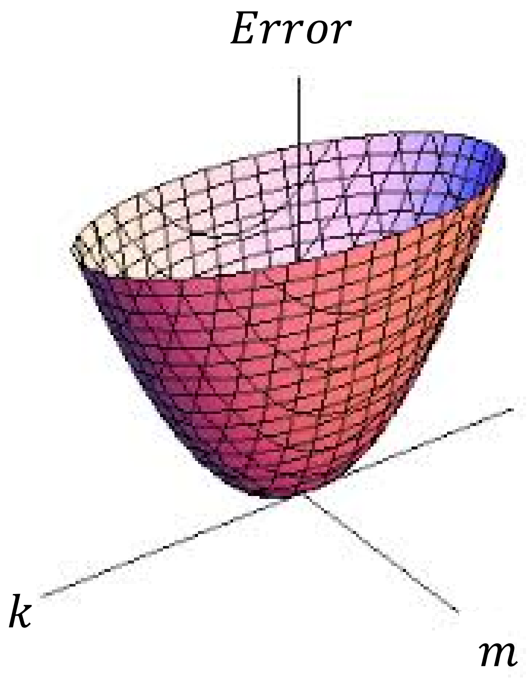
How to minimize the error

Two variables so we can plot it in 3D

$$Error = a \cdot k^2 + b \cdot k \cdot m + c \cdot m^2 + d$$

Now we can find the min of Error: the two partial derivatives are easy to compute and both should be 0.

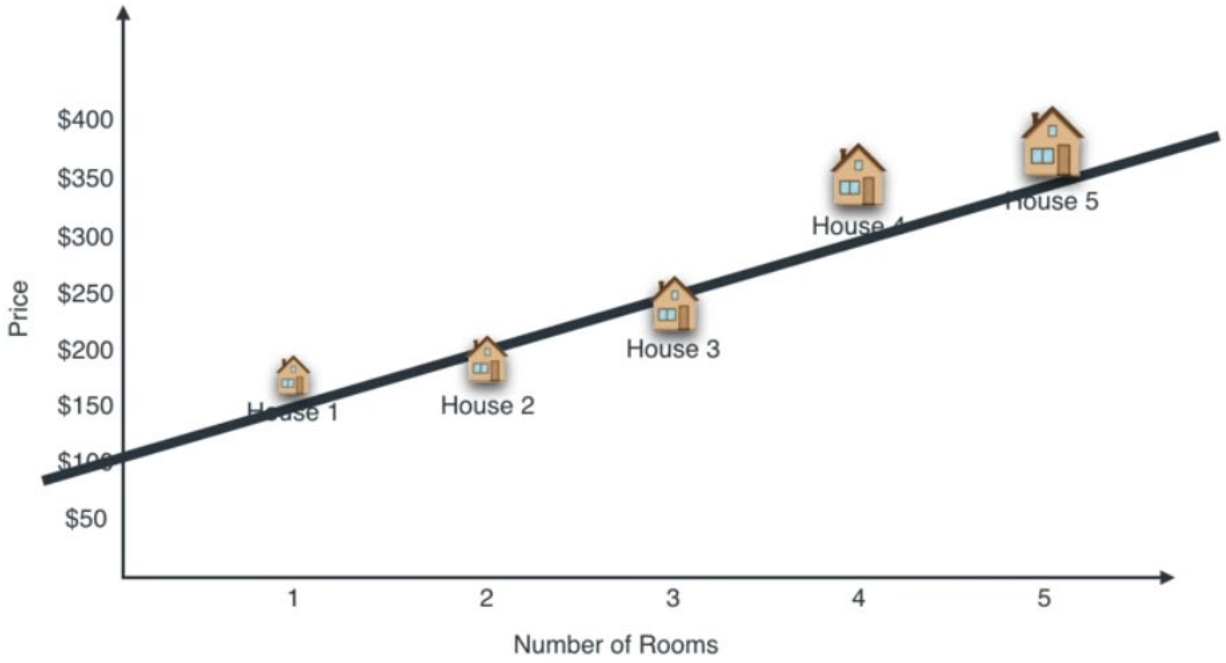
$$\begin{cases} 2ka + bm = 0 \\ 2mc + bk = 0 \end{cases}$$



This is a 2x2 linear equation. Solve it to get k and m and hence the line f(x)=kx+m!

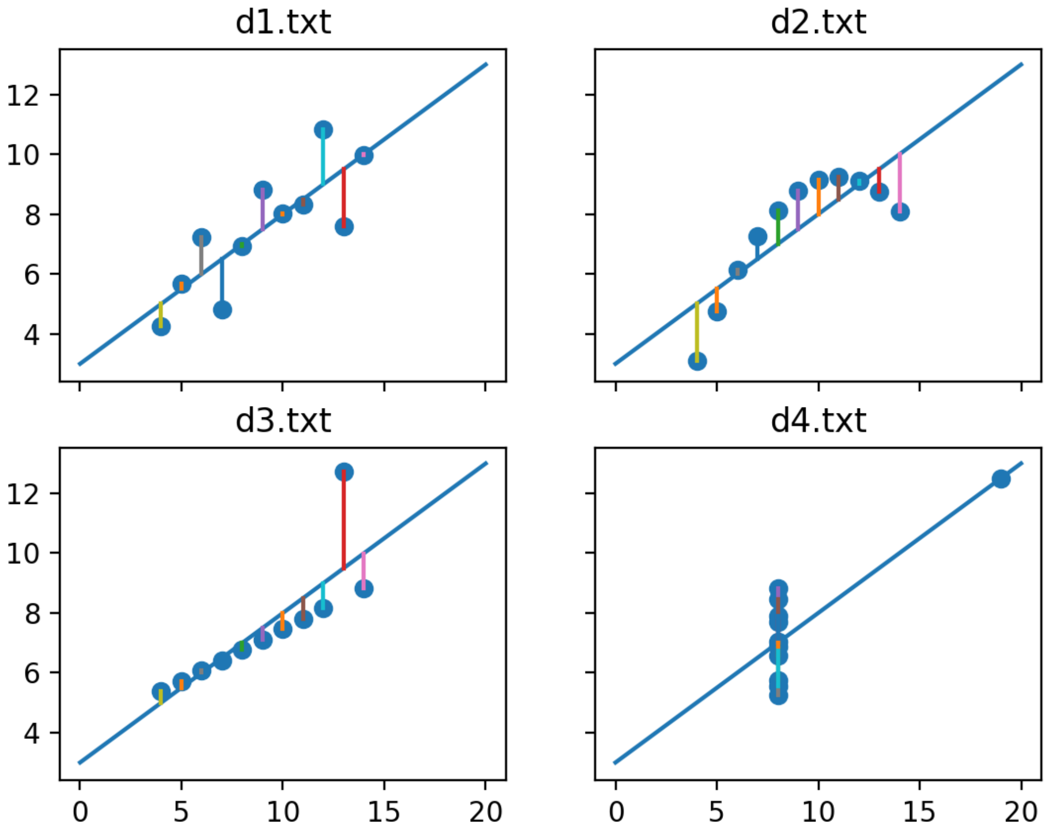
This Error function has no maximum, so what we find here will be a minimum (or check the second derivatives).

Estimating house prices



So now we have a method called *linear regression* for fitting a line to a set of data.

Examples of linear regression



These four data sets all have the same regression line.

Arbitrary base functions

- Now we assumed the form $f(x) = kx + m$.
- But linear regression actually works with arbitrary base functions!
- We may put, e.g., $f(x) = w_3 x^2 + w_2 \cos(x) + w_1 e^x + w_0$.
- Here there are four variables w_0, w_1, w_2, w_3. The datapoint coordinates x_i will be constants like before. Hence x_i^2, $\cos(x_i)$, and e^{x_i} will also be constants.
- The Error function will be analogous to what we had before. Error will be minimal when all four partial derivatives are 0. So we can find the values of the four variables w_0, w_1, w_2, w_3 as before!
- Thus we can capture a non-linear relationship with a linear model!

Example: Polynomial Regression

This is a linear model, but the curve is quadratic rather than a line:

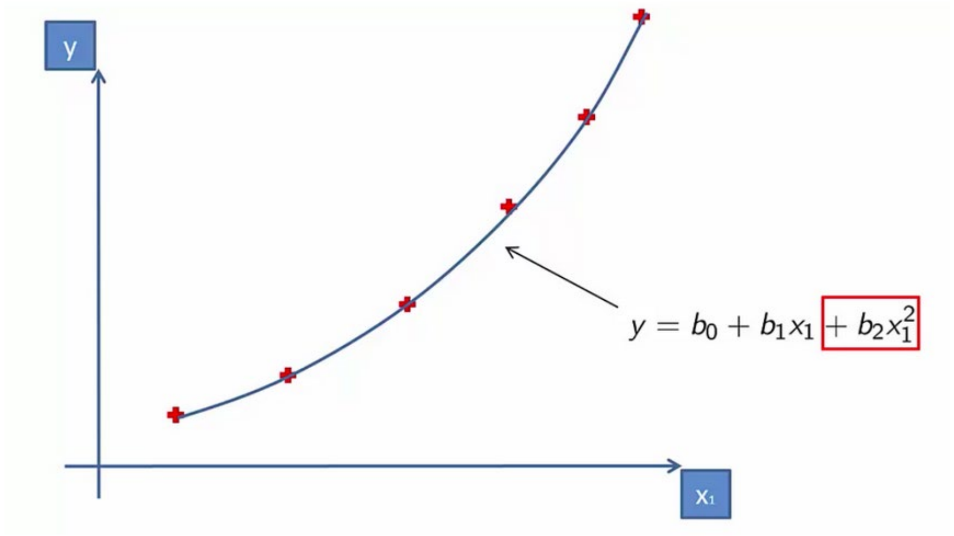


Image source

Residuals

Brushtail possums (n=104)

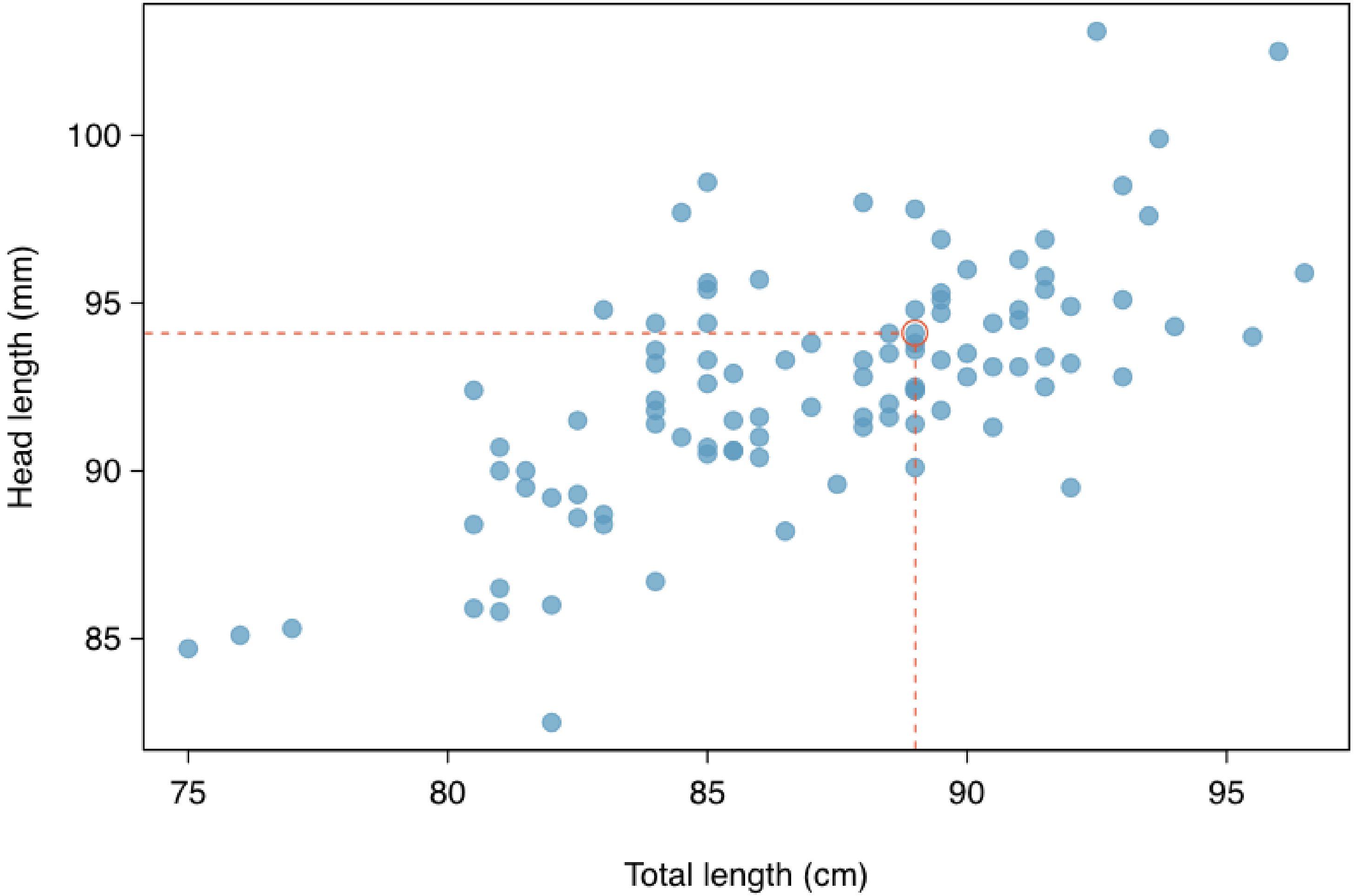
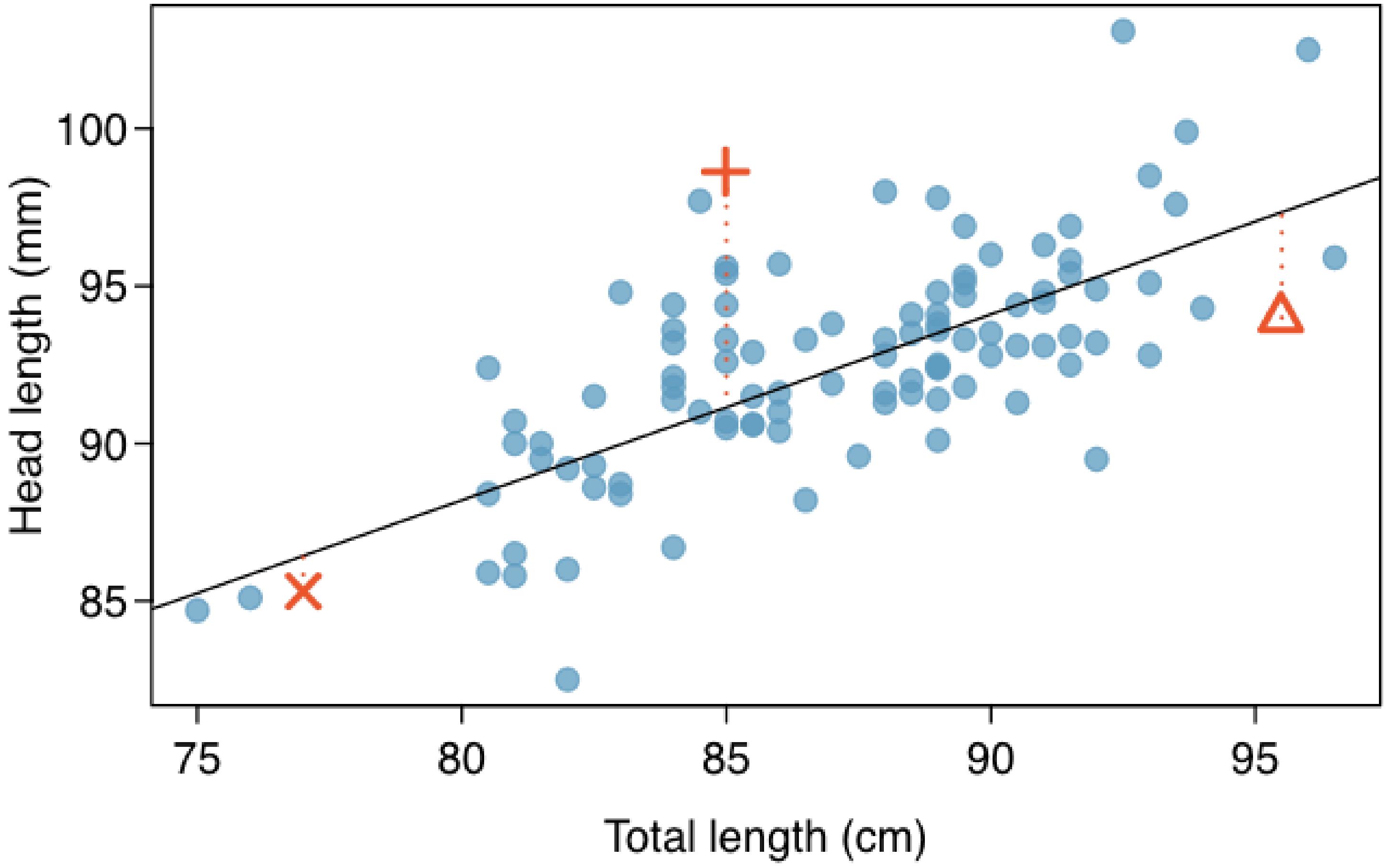


Image by JJ Harrison, https://commons.wikimedia.org/wiki/File:Trichosurus_vulpecula_1.jpg

Goal:
Express one variable as a function of other(s)

https://courses.lumenlearning.com/odessa-introstats1-1/chapter/line-fitting-residuals-and-correlation/

A linear model



Then we can use linear regression to construct a line.

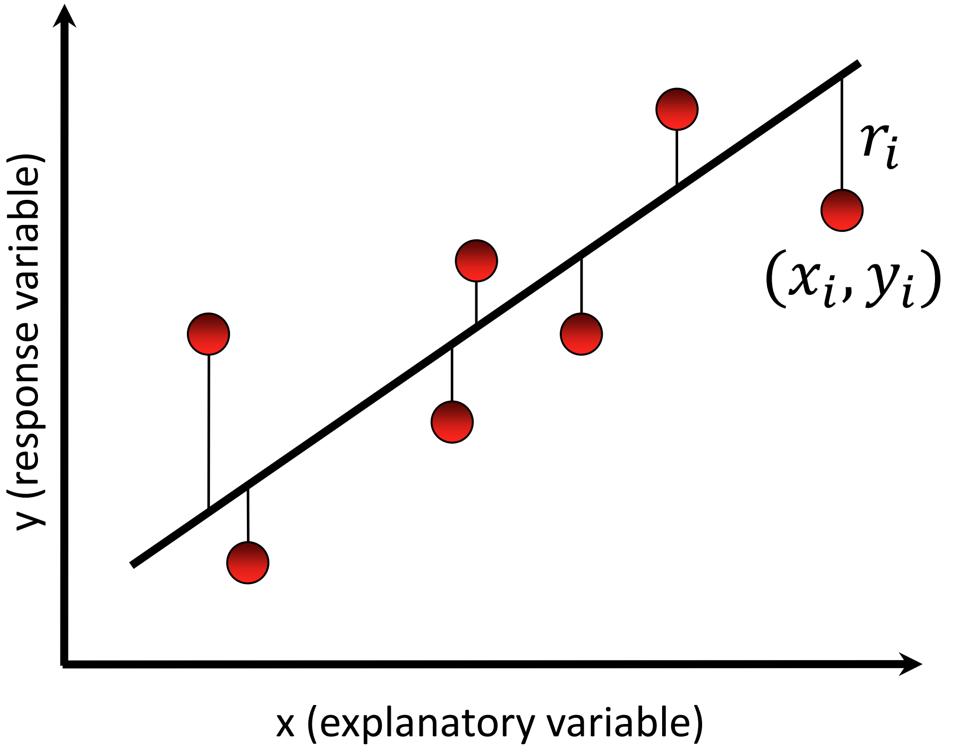
How good is this model?

https://courses.lumenlearning.com/odessa-introstats1-1/chapter/line-fitting-residuals-and-correlation/

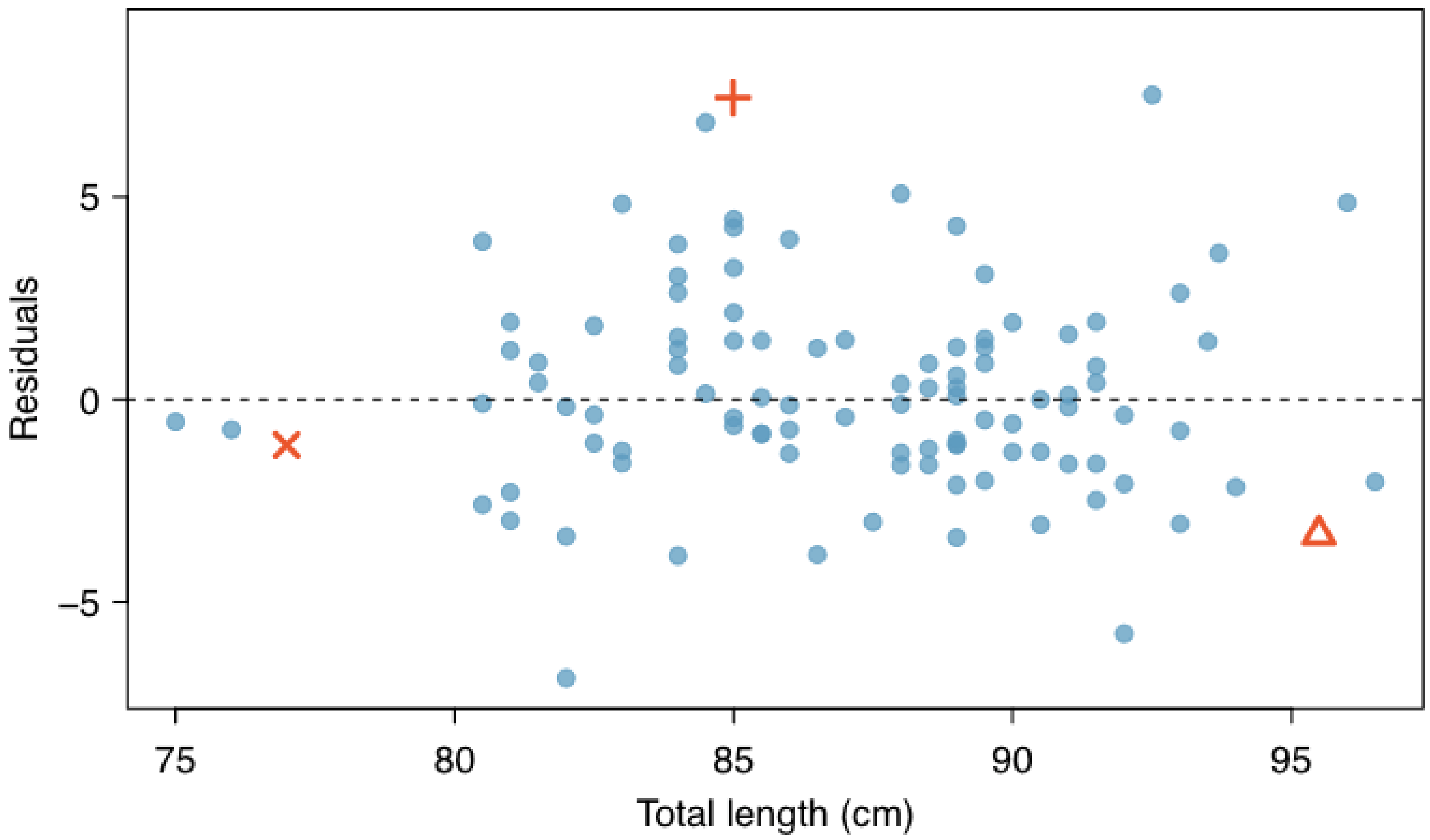
Residuals

Given any model $f(x)$, each data point (x_i, y_i) will have a *residual* r_i :

$$y_i = f(x_i) + r_i$$

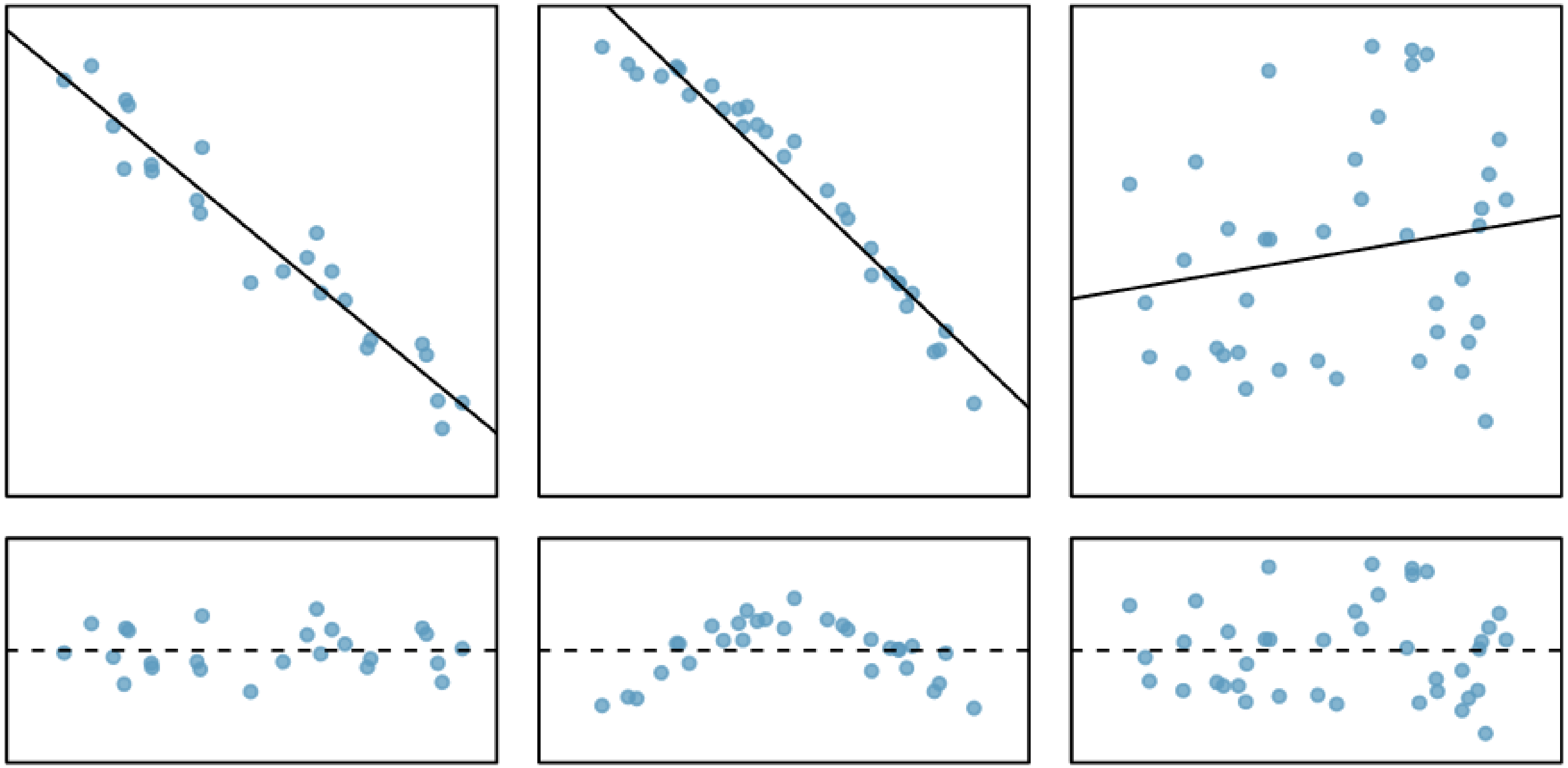


Residual plot



https://courses.lumenlearning.com/odessa-introstats1-1/chapter/line-fitting-residuals-and-correlation/

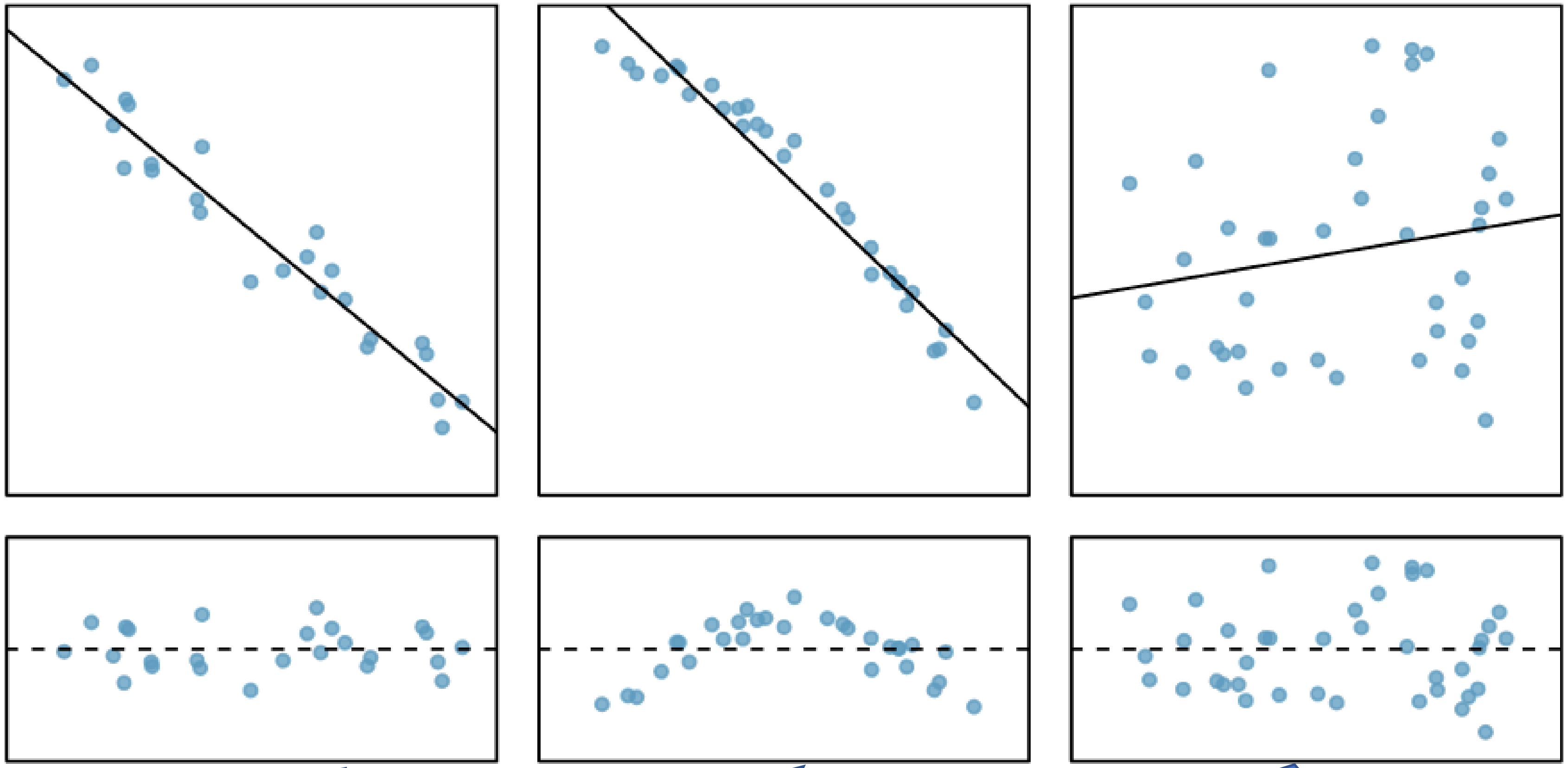
Scatter plots and residual plots



https://courses.lumenlearning.com/odessa-introstats1-1/chapter/line-fitting-residuals-and-correlation/

Scatter plots and residual plots

Are the residuals balanced?



https://courses.lumenlearning.com/odessa-introstats1-1/chapter/line-fitting-residuals-and-correlation/

Variance

- Consider the numbers $x_1, \dots, x_n$
- Then the *mean* is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- The *population variance* is $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$
- The *sample variance* is $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Average squared deviation from the mean

Almost the same as the population variance (for large n). Sometimes practical to have n-1 instead of n in the denominator

Standard deviation

Sample standard deviation: $\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$

This is the square root of the sample variance

It has the same unit as the x_i

Correlation

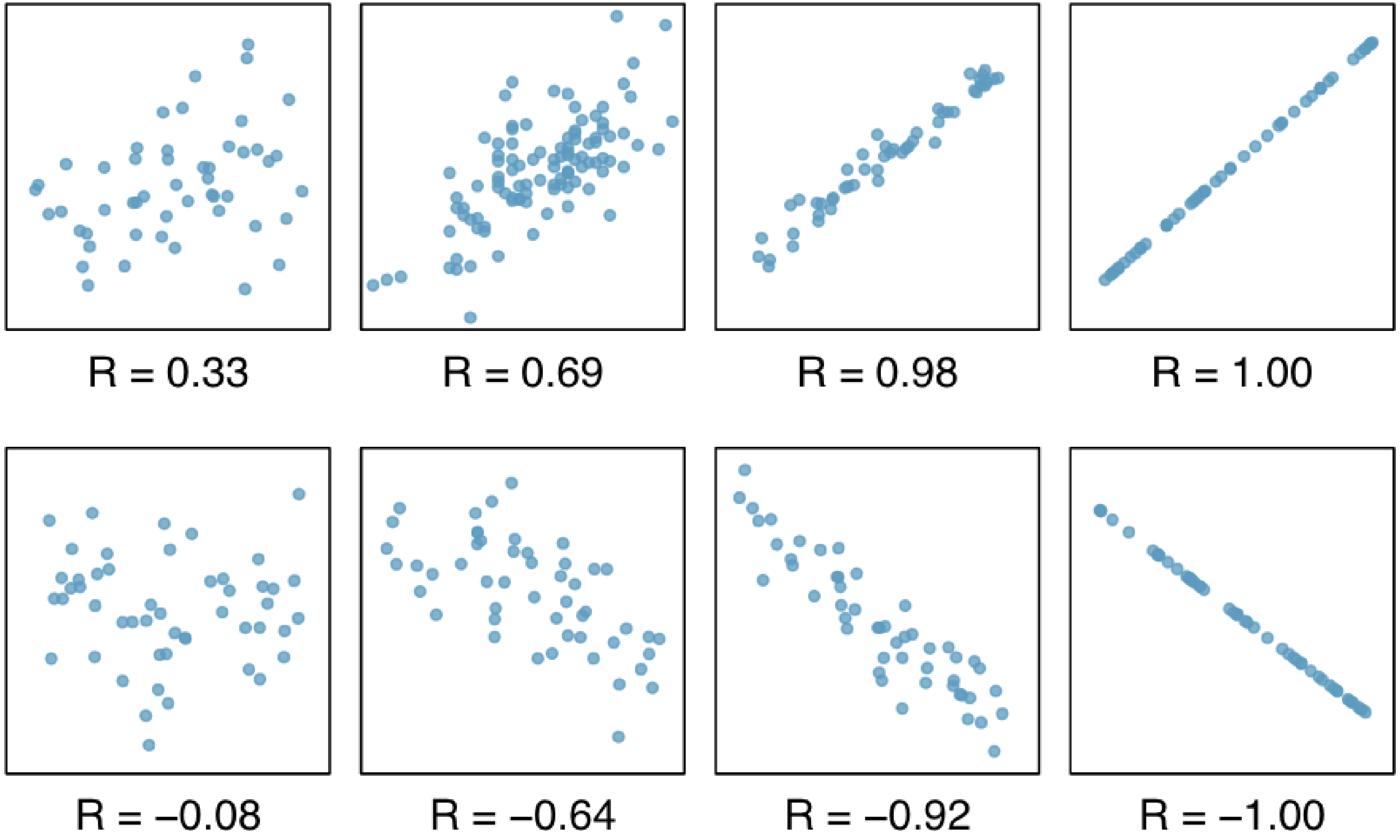
mean x and y

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

standard deviation of x and y

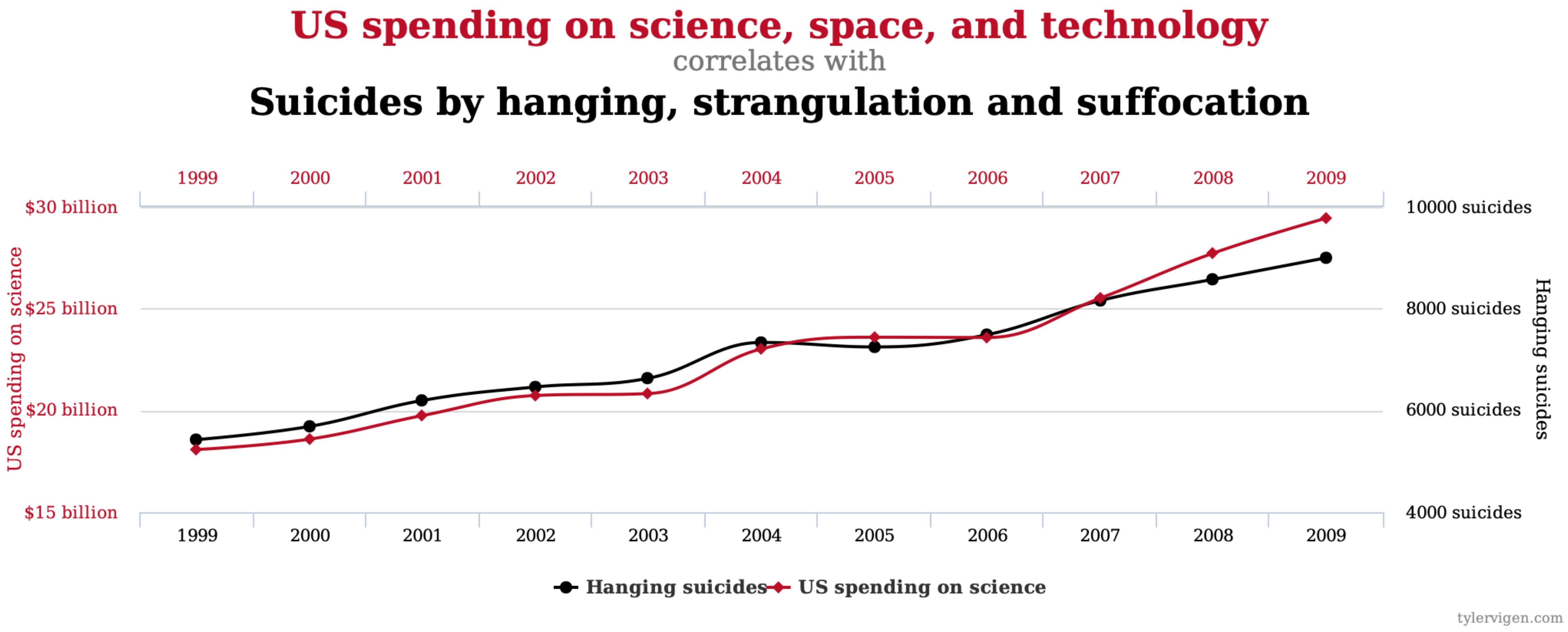
Quantifies the strength of a linear trend

Examples of correlations

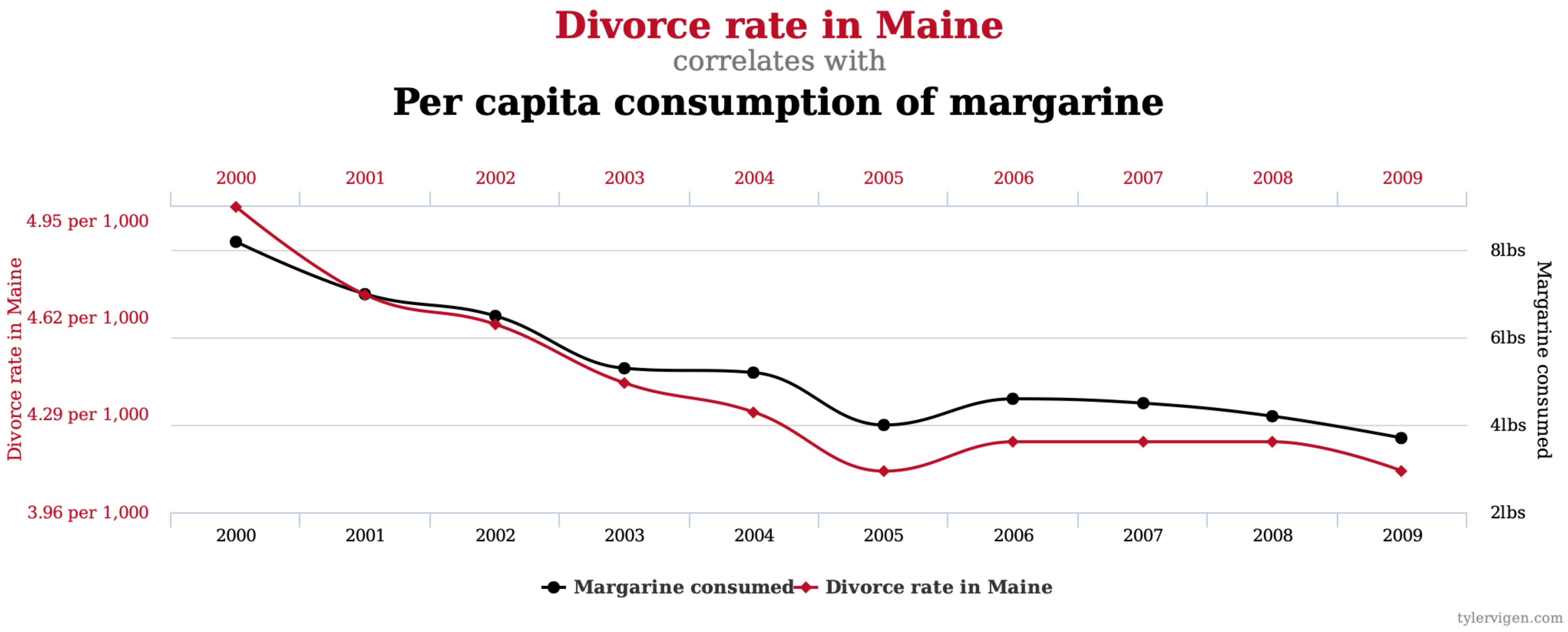


R measures to what extent the datapoints fall on a slanted line.

So R measures to what extent a set is suitable for linear regression



http://tylervigen.com/spurious-correlations

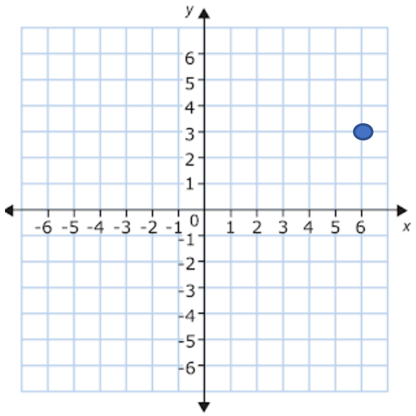


http://tylervigen.com/spurious-correlations

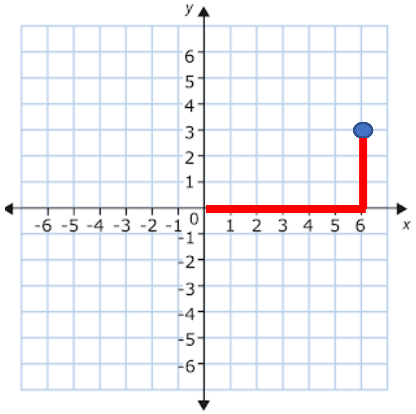
Multidimensional regression

Norms in two dimensions

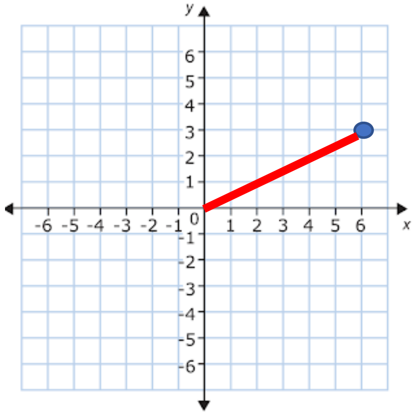
- A *norm* is a way of measuring the length of a vector.



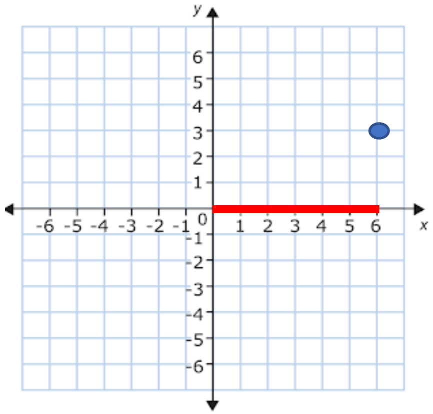
How do we measure the distance from (0,0) to the blue dot?



L1 norm = Manhattan distance = 6+3



L2 norm = Euclidean distance = $\sqrt{6^2 + 3^2}$



L-infinity norm = max coordinate length = 6

Norms in several dimensions

Length of a vector in n dimensions

We will use norms for measuring error vectors generated by n data points

- L1-norm: $||(v_1, \ldots, v_n)||_1 = \sum_{i=1}^{n} |v_i|$

- L2-norm: $||(v_1, \ldots, v_n)||_2 = \sqrt{\sum_{i=1}^{n} v_i^2}$

- L-infinity norm: $||(v_1, \ldots, v_n)||_\infty = \max_{1 \le i \le n} |v_i|$

Multidimensional regression

- Given a dataset $(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_n}, y_n)$

Here the $\boldsymbol{x_i}$ are N-dimensional vectors of real numbers and the y_i are real numbers.

- We want to find a prediction function $f(\boldsymbol{x})$ of a certain form (e.g. a polynomial of degree m) such that the error vector norm

$$\|y_1 - f(\boldsymbol{x_1}), \ldots, y_n - f(\boldsymbol{x_n})\|_? \text{ is minimized.}$$

The error at $\boldsymbol{x_1}$

The error at $\boldsymbol{x_n}$

Thus we aggregate all prediction errors into a single number

Could be 1, 2, or ∞

An optimization problem!

The point: if we have such a prediction function f, then we can use it to make a prediction $f(\boldsymbol{x})$ for any $\boldsymbol{x}$.

Multidimensional regression

- The n datapoints are of the form $\left(x_{\mathrm{i},1}, \ldots, x_{\mathrm{i},\mathrm{N}}, \mathrm{y}_{\mathrm{i}}\right)$.

- The prediction function that we are looking for is of the form

$$f(x_1, \ldots, x_{\mathrm{N}}) = w_0 + w_1 x_1 + \cdots + w_N\, x_N$$

- Such an f is called a *hyperplane* in N dimensions. When N=2, f is a plane.

Multidimensional regression

- We would ideally like to find values of the variables w_j so that the following equations hold:

$$w_N x_{N,1} + \ldots + w_1 x_{1,1} + w_0 = y_1$$
$$w_N x_{N,2} + \ldots + w_1 x_{1,2} + w_0 = y_2$$
$$\ldots$$
$$w_N x_{N,n} + \ldots + w_1 x_{1,n} + w_0 = y_n$$

Datapoint 1 correctly predicted

Datapoint n correctly predicted

- We can write this more compactly using matrices:

$$\boldsymbol{X}\boldsymbol{w} = \boldsymbol{y}$$

Multidimensional regression

- Since we typically have $n > N$, this equation is usually not solvable:

$$\boldsymbol{X}\boldsymbol{w} = \boldsymbol{y}$$

- But we can always try to minimize Error = $\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}$ (a vector of dim n):

$$\min_{\boldsymbol{w}} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2 = \min_{\boldsymbol{w}} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

With square root

Squared so that the square root disappears

This reduces to the sum of the squared errors. To find the minimum we just set the partial derivatives to 0 and solve the equation system like before!

So now we can do linear regression also in the multidimensional case!

Regularization

Occam’s Razor

"The simplest solution is most likely the right one.”

William of Occam (c. 1287–1347)

Occam's razor is used for “shaving” off unnecessary assumptions!

Regularization

Regularization is a tool for keeping models simple (in the spirit of Occam)

- Ordinary error:

$$\min_{\boldsymbol{w}} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2$$

It encourages small weights (by penalizing large weights)

- With *ridge regularization*:

$$\min_{\boldsymbol{w}} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2 + \alpha||\boldsymbol{w}||_2^2$$

Avoid large weights

- With *lasso regularization*:

$$\min_{\boldsymbol{w}} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2 + \alpha||\boldsymbol{w}||_1$$

Encourage many 0 weights
$Small^2 \ll Small$

Applications of linear regression

Linear regression

- “Linear regression is a bread-and-butter modeling technique that should serve as your baseline approach to building data-driven models.”
- “These models are typically easy to build, straightforward to interpret, and often do quite well in practice.”
- “With enough skill and toil, more advanced machine learning techniques might yield better performance, but the possible payoff is often not worth the effort.”
- “Build your linear regression models first, then decide whether it is worth working harder to achieve better results.”

S.S. Skiena, “The Data Science Design Manual”, 2017

Applications of linear regression

- House price based on size
- Tip received based on bill
- Sales forecast
- Price of a stock
- Spread of a disease

Many trends follow more complex curves, e.g. exponential or logistic. Still, linear regression can be very useful for making rough estimates!

Using linear regression

Open linear_regression_intro (on Canvas)