COM9417 – Week 1

Regression (1)

COMP9417 Machine Learning and Data Mining

Term 2, 2022

Acknowledgements

Material derived from slides for the book
"Elements of Statistical Learning (2nd Ed.)" by T. Hastie,
R. Tibshirani & J. Friedman. Springer (2009)
`http://statweb.stanford.edu/~tibs/ElemStatLearn/`

Material derived from slides for the book
"Machine Learning: A Probabilistic Perspective" by P. Murphy
MIT Press (2012)
`http://www.cs.ubc.ca/~murphyk/MLbook`

Material derived from slides for the book
"Machine Learning" by P. Flach
Cambridge University Press (2012)
`http://cs.bris.ac.uk/~flach/mlbook`

Material derived from slides for the book
"Bayesian Reasoning and Machine Learning" by D. Barber
Cambridge University Press (2012)
`http://www.cs.ucl.ac.uk/staff/d.barber/brml`

Material derived from slides for the book
"Machine Learning" by T. Mitchell
McGraw-Hill (1997)
`http://www-2.cs.cmu.edu/~tom/mlbook.html`

Material derived from slides for the course
"Machine Learning" by A. Srinivasan
BITS Pilani, Goa, India (2016)

Aims

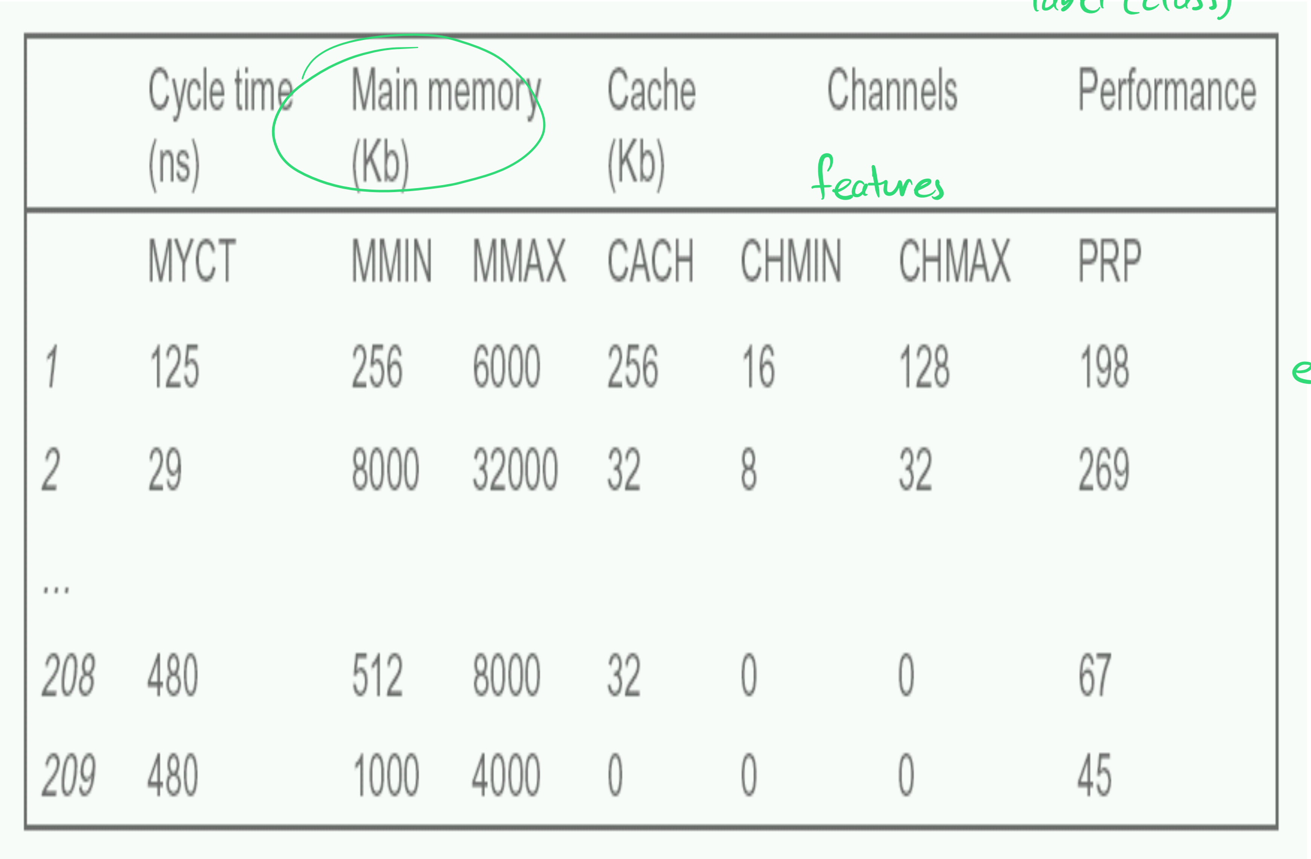
This lecture will introduce you to machine learning approaches to the problem of numerical prediction. Following it you should be able to reproduce theoretical results, outline algorithmic techniques and describe practical applications for the topics:

- the supervised learning task of numeric prediction ✓✓
- how linear regression solves the problem of numeric prediction ✓✓
- fitting linear regression by least squares error criterion ✓✓
- non-linear regression via linear-in-the-parameters models ✓✓
- gradient descent to estimate parameters for regression ✓✓

Introduction to Regression

Task[1] is to learn a model to predict CPU performance from a dataset of examples of 209 different computer configurations:

"label" (class)



	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
...							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

features

example

[1] Available from: https://archive.ics.uci.edu/ml/datasets/Computer+Hardware

One possible model is a linear model, e.g.,

```
PRP =
  - 56.1
    + 0.049 MYCT
    + 0.015 MMIN
    + 0.006 MMAX
    + 0.630 CACH
    - 0.270 CHMIN
    + 1.46 CHMAX
```

recall MATH2831

values of input variables

Coefficients or parameters

Regression

Features = Variables = Attribute

Covariate

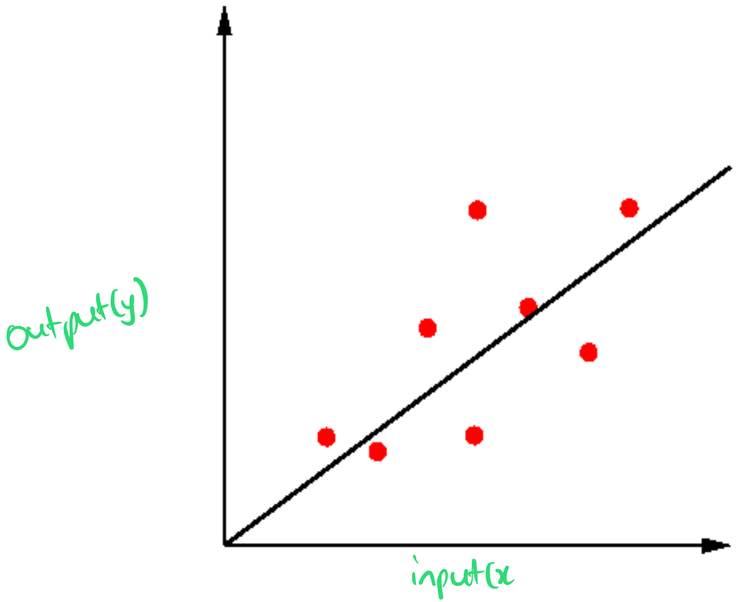
We will look at the simplest model for numerical prediction:
a *regression equation*

Outcome will be a linear sum of feature values with appropriate weights.

Note: the term *regression* is overloaded – it can refer to:

- the process of determining the weights for the regression equation, or ✓✓
- the regression equation itself. ✓✓

Linear Regression



features	label
inputs	outputs
$x1 = 1$	$y1 = 1$
$x2 = 3$	$y2 = 2.2$
$x3 = 2$	$y3 = 2$
$x4 = 1.5$	$y4 = 1.9$
$x5 = 4$	$y5 = 3.1$

Assumes: expected value of the output given an input, $E[y|x]$, is linear.
Simplest case: $\text{Out}(x) = bx$ for some unknown b.
Learning problem: given the data, estimate b (i.e., $\hat{b}$).

Linear Models

- Data has p numeric features, and we need numeric prediction $\Rightarrow$ regression
- Linear models, i.e., outcome is *linear* combination of attributes

$$\boxed{y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p}$$

$$x^{(1)} = \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_p^{(1)} \end{bmatrix}$$

- **Predicted** value for first training instance $\mathbf{x}^{(1)}$ is:

$$\boxed{b_0 x_0^{(1)} + b_1 x_1^{(1)} + b_2 x_2^{(1)} + \ldots + b_p x_p^{(1)} = \sum_{i=0}^{p} b_i x_i^{(1)}}$$

- $p + 1$ weights or coefficients must be learned on training data
- $x_0^{(1)} = 1$

Minimizing Mean Squared Error

Difference between *predicted* and *actual* values is the error !

$p + 1$ coefficients are chosen so that mean of sum of squared errors on all instances in training data is minimized.

$\hat{y}$ is the predicted value

Mean Squared Error (MSE):

$$\boxed{\frac{1}{n}\sum_{j=1}^{n}\left(\overset{\text{Actual}}{y^{(j)}} - \overset{\text{Predicted}}{\sum_{i=0}^{p} b_i x_i^{(j)}}\right)^2}$$

$$\frac{1}{n}\sum (y-\hat{y})^2$$

- Actual values are given

Coefficients b_i can be derived using calculus. (derivative)

Can be done if there are more instances than attributes (roughly speaking).

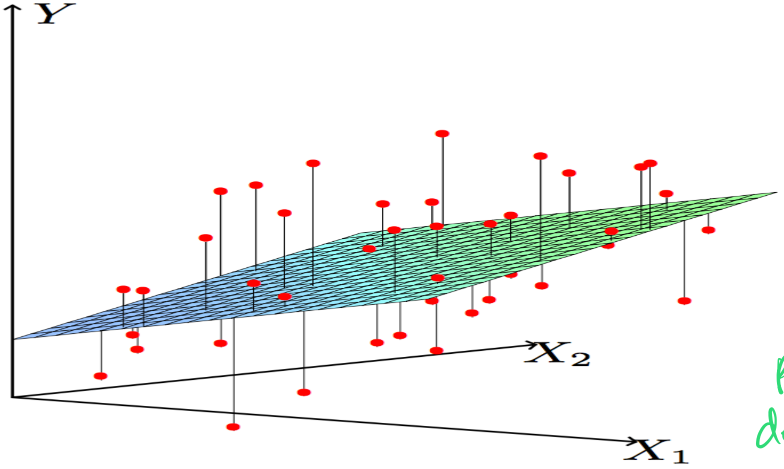
Known as "Ordinary Least Squares" (OLS) regression – minimizing the mean of the sum of squared distances of data points to the estimated regression line.

same thing as MSE, except we take absolute value.

Not the only approach — could use absolute error, etc. but this the most widely used . . .

Multivariate (Multiple) Regression

Given 2 real-valued variables X_1, X_2, labelled with a real-valued variable Y, find "plane of best fit" that captures the dependency of Y on X_1, X_2.



Minimising MSE involves taking difference from each point, and then determining the average.

Learning here is by minimizing MSE, i.e., average of squared vertical distances of actual values of Y from the learned function $\hat{Y} = \hat{f}(\mathbf{X})$.

Step back: Statistical Techniques for Data Analysis

Probability vs Statistics: The Difference

- **Probability** versus **Statistics**
- Probability: reasons from populations to samples
  - This is deductive reasoning, and is usually *sound* (in the logical sense of the word)
- Statistics: reasons from samples to populations
  - This is inductive reasoning, and is usually *unsound* (in the logical sense of the word)

Statistical Analyses

- Statistical analyses usually involve one of 3 things:
  1. The study of populations;
  2. The study of variation; and
  3. Techniques for data abstraction and data reduction
- Statistical analysis is more than statistical computation:
  1. What is the question to be answered?
  2. Can it be quantitative (i.e., can we make measurements about it)?
  3. How do we collect data?
  4. What can the data tell us?

Sampling

Where do the Data come from? (Sampling)

- For groups (populations) that are fairly homogeneous, we do not need to collect a lot of data. (We do not need to sip a cup of tea several times to decide that it is too hot.)
- For populations which have irregularities, we will need to either take measurements of the entire group, or find some way of get a good idea of the population without having to do so
- *Sampling* is a way to draw conclusions about the population without having to measure all of the population. The conclusions need not be completely accurate
- All this is possible if the sample closely resembles the population about which we are trying to draw some conclusions

What We Want From a Sampling Method

- No systematic bias, or at least no bias that we cannot account for in our calculations
- The chance of obtaining an unrepresentative sample can be calculated. (So, if this chance is high, we can choose not to draw any conclusions.)
- The chance of obtaining an unrepresentative sample decreases with the size of the sample

For the class of *numeric* representations, machine learning is viewed as:

"searching" a space of **functions** . . .

represented as mathematical models (linear equations, neural nets, . . .).

Which model is best for some sample(s) of data ?

Which model is best for generalising to the population ?

Methods to predict a numeric output from statistics and machine learning:

- linear regression (statistics) determining the "line of best fit" using the least squares criterion
- linear models (machine learning) learning a predictive model from data under the assumption of a linear relationship between predictor and target variables

Very widely-used, many applications

Ideas that are generalised in Artificial Neural Networks (Deep Learning) and other types of learning ...

Regression as a term occurs in many areas of machine learning:

- linear regression the classic
- non-linear regression by adding non-linear basis functions
- multi-layer neural networks (machine learning) learning non-linear predictors via hidden nodes between input and output
- regression trees (statistics / machine learning) tree where each leaf predicts a numeric quantity
- local (nearest-neighbour) regression

The inductive learning hypothesis

"machine learning in a nutshell"

Any estimate[2] found to approximate the target function well over a sufficiently large set of training examples will also approximate the target function well over other unobserved examples[3].

[2] Estimate forming part of a model, such as regression coefficients, or other model parameters.

[3] After Mitchell (1997)

Estimation

Estimation from a Sample

- Estimating some aspect of the population using a sample is a common task. Along with the estimate, we also want to have some idea of the accuracy of the estimate (usually expressed in terms of *confidence limits*)
- Some measures calculated from the sample are very good estimates of corresponding population values. For example, the sample mean m is a very good estimate of the population mean μ. But this is not always the case. For example, the range of a sample usually under-estimates the range of the population
- We will have to clarify what is meant by a "good estimate". One meaning is that an estimator is correct on average. For example, on average, the mean of a sample is a good estimator of the mean of the population

Estimation from a Sample

- For example, when a number of samples are drawn and the mean of each is found, then average of these means is equal to the population mean
- Such an estimator is said to be *statistically unbiased*

Sample Estimates of the Mean and the Spread I

Mean. This is calculated as follows.

- Find the total T of n observations. Estimate the (arithmetic) mean by $m = T/n$.
- This works very well when the data follow a symmetric bell-shaped frequency distribution (of the kind modelled by “normal” distribution)
- A simple mathematical expression of this is $m = \frac{1}{n}\sum_i x_i$, where the observations are x_1, $x_2 \ldots x_n$
- If we can group the data so that the observation x_1 occurs f_1 times, x_2 occurs f_2 times and so on, then the mean is calculated as $m = \frac{1}{\sum_i f_i}\sum_i x_i f_i$

Sample Estimates of the Mean and the Spread II

- If, instead of frequencies, you had relative frequencies (i.e. instead of f_i you had $p_i = f_i/n$), then the mean is simply the observations weighted by relative frequency. That is, $m = \sum_i x_i p_i$
- We want to connect this up to computing the mean value of observations modelled by some theoretical probability distribution function. That is, we want to a similar counting method for calculating the mean of random variables modelled using some known distribution

Sample Estimates of the Mean and the Spread III

- Correctly, this is the mean value of the *values of the random variable function*. But this is a bit cumbersome, so we will just say the "mean value of the r.v." For discrete r.v.'s this is:

$$\boxed{\mathbb{E}(X) \;=\; \sum_i x_i p(X = x_i)}$$

Variance. This is calculated as follows:

- Calculate the total T and the sum of squares of n observations. The estimate of the standard deviation is $s = \sqrt{\frac{1}{n-1}\sum_i (x_i - m)^2}$
- Again, this is a very good estimate when the data are modelled by a normal distribution

why is it n-1 and not just n?

Sample Estimates of the Mean and the Spread IV

- For grouped data, this is modified to $s = \sqrt{\frac{1}{n-1}\sum_i (x_i - m)^2 f_i}$
- Again, we have a similar formula in terms of expected values, for the scatter (spread) of values of a r.v. X around a mean value $\mathbb{E}(X)$:

$$\boxed{\begin{aligned} Var(X) &= \mathbb{E}((X - \mathbb{E}(X))^2) \\ &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \end{aligned}}$$

- You can remember this as “the mean of the squares minus the square of the mean”

Covariance and Correlation

Correlation

X	Y
42	691
37	713
---	---

- The *correlation coefficient* is a number between -1 and +1 that indicates whether a pair of variables X and Y are associated or not, and whether the scatter in the association is high or low
  - High values of X are associated with high values of Y *and* low values of X are associated with low values of Y, and scatter is low +1 pos.
  - A value near 0 indicates that there is no particular association and that there is a large scatter associated with the values
  - A value close to -1 suggests an inverse association between X and Y neg.
- Only appropriate when X and Y are roughly linearly associated (doesn't work well when the association is curved)

Correlation

- Correlation between X and Y as a population quantity is:

$$r = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}$$

← "association"

← "scatter"

- Sample variance for X:

$$S_{XX} = \sum (x_i - \overline{x})^2$$

n examples:

(x_1, y_1)

(x_2, y_2)

...

- Sample variance for Y:

$$S_{YY} = \sum (y_i - \overline{y})^2$$

- Sample covariance for X, Y:

$$S_{XY} = \sum (x_i - \overline{x})(y_i - \overline{y})$$

- The formula for computing correlation from a sample of data on X and Y is:

$$\cancel{\hat{r} = \frac{S_{XY}}{S_{XX}} S_{YY}}$$

$$\boxed{\hat{r} = \frac{S_{xy}}{S_{xx} S_{yy}}}$$

Correlation

- What does "covariance" intuitively mean ? Consider
  1. Case 1: $x_i > \overline{x}$, $y_i > \overline{y}$
  2. Case 2: $x_i < \overline{x}$, $y_i < \overline{y}$
  3. Case 3: $x_i < \overline{x}$, $y_i > \overline{y}$
  4. Case 4: $x_i > \overline{x}$, $y_i < \overline{y}$

  In the first two cases, x_i and y_i vary together, both being high or low relative to their means. In the other two cases, they vary in different directions
- If the positive products dominate in the calculation of S_{XY}, then the value of r will be positive. If the negative products dominate, then r will be negative. If 0 products dominate, then r will be close to 0.

What Does Correlation Mean? I

- r is a quick way of checking whether there is some linear association between X and Y
- The sign of the value tells you the direction of the association
- All that the numerical value tells you is about the scatter in the data
- The correlation coefficient does not model any relationship. That is, given a particular X you cannot use the r value to calculate a Y value
    - It is possible for two datasets to have the same correlation, but different relationships
    - It is possible for two datasets to have different correlations but the same relationship

What Does Correlation Mean? I

- MORAL: Do not use correlations to compare datasets. All you can derive is whether there is a positive or negative relationship between X and Y
- ANOTHER MORAL: Do not use correlation to imply X causes Y or the other way around (see HW0)

Regression

Regression

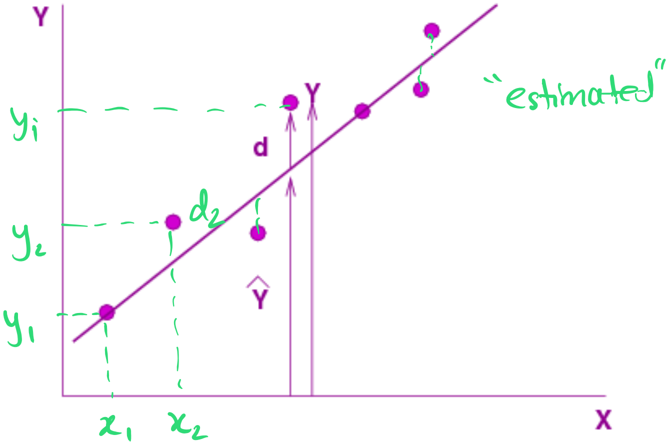
- Given a set of data points x_i, y_i, what is the relationship between them? (We can generalise this to the “multivariate” case later)
- One kind of question is to ask: are these linearly related in some manner? That is, can we draw a straight line that describes reasonably well the relationship between X and Y
- Remember, the correlation coefficient can tell us if there is a case for such a relationship
- In real life, even if such a relationship held, it will be unreasonable to expect all pairs x_i, y_i to lie precisely on a straight line. Instead, we can probably draw some reasonably well-fitting line. But which one?

Univariate linear regression

Linear Relationship Between 2 Variables

$Y = a + bX$

$\hat{y} = a + bx$



X	Y
x_1	y_1
x_2	y_2
...	...

n example

Program Regression

input: x

output: y

$\boxed{\;---\;}$

- GOAL: fit a line whose equation is of the form $\hat{y} = \underline{a} + \underline{b}x$
- HOW: minimise $\sum_i d_i^2 = \sum_i (y_i - \hat{y_i})^2$ (the "least squares estimator")

Linear Relationship Between 2 Variables

- The calculation for b is given by:

$$b = \frac{S_{XY}}{S_{XX}}$$

- Then we have $a = \overline{y} - b\overline{x}$

Meaning of the Coefficients a and b

unit change in x = add 1 unit to x

- b: change in Y that accompanies a unit change in X
- If the values of X were assigned at random, then b estimates the unit change in Y *caused* by a unit change in X Econometrics, Epidemiology
- If the values of X were not assigned at random (for example, they were data somebody observed), then the change in Y will include the change in X and any other confounding variables that may have changed as a result of changing X by 1 unit. So, you cannot say for example, that a change of X by 1 unit causes b units of change in Y

ML

- $b = 0$ means there is no linear relationship between X and Y, and then best we can do is simply say is $\hat{Y} = a = \overline{Y}$. Estimating the sample mean is therefore a special case of the MSE criterion

Finding the parameters for univariate linear regression

Train

Program LearnRegression

Input: $\{(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)\}$

Output: $\hat{a},\hat{b}$ (estimated) (actual)

$$\boxed{?}$$

Program PredictRegression

Input: x

Output: $\hat{y}$

$$\boxed{\hat{y}=a+bx}$$

Univariate linear regression

Training Dataset :

cm h	kg w
180	75
181	78.5
170	69.3
---	---

Example:

Suppose we want to investigate the relationship between people's height and weight[4].

We collect n height and weight measurements $(h_i, w_i), 1 \le i \le n$.

Univariate linear regression assumes a linear equation $w = a + bh$, with parameters a and b chosen such that the sum of squared residuals $\sum_{i=1}^{n}(\underset{\substack{\uparrow \\ \text{actual weight}}}{w_i} - \underset{\substack{\uparrow \\ \text{predicted weight}}}{(a + bh_i)})^2$ is minimised.

"loss function"

"cost function"

"objective function"

[4] This example from Flach (2012).

Univariate linear regression

In order to find the parameters we take partial derivatives, set the partial derivatives to 0 and solve for a and b:

$$\frac{\partial}{\partial a}\sum_{i=1}^{n}(w_i-(a+bh_i))^2 = -2\sum_{i=1}^{n}(w_i-(a+bh_i)) = 0$$

$$\Rightarrow \hat{a} = \overline{w} - \hat{b}\overline{h}$$

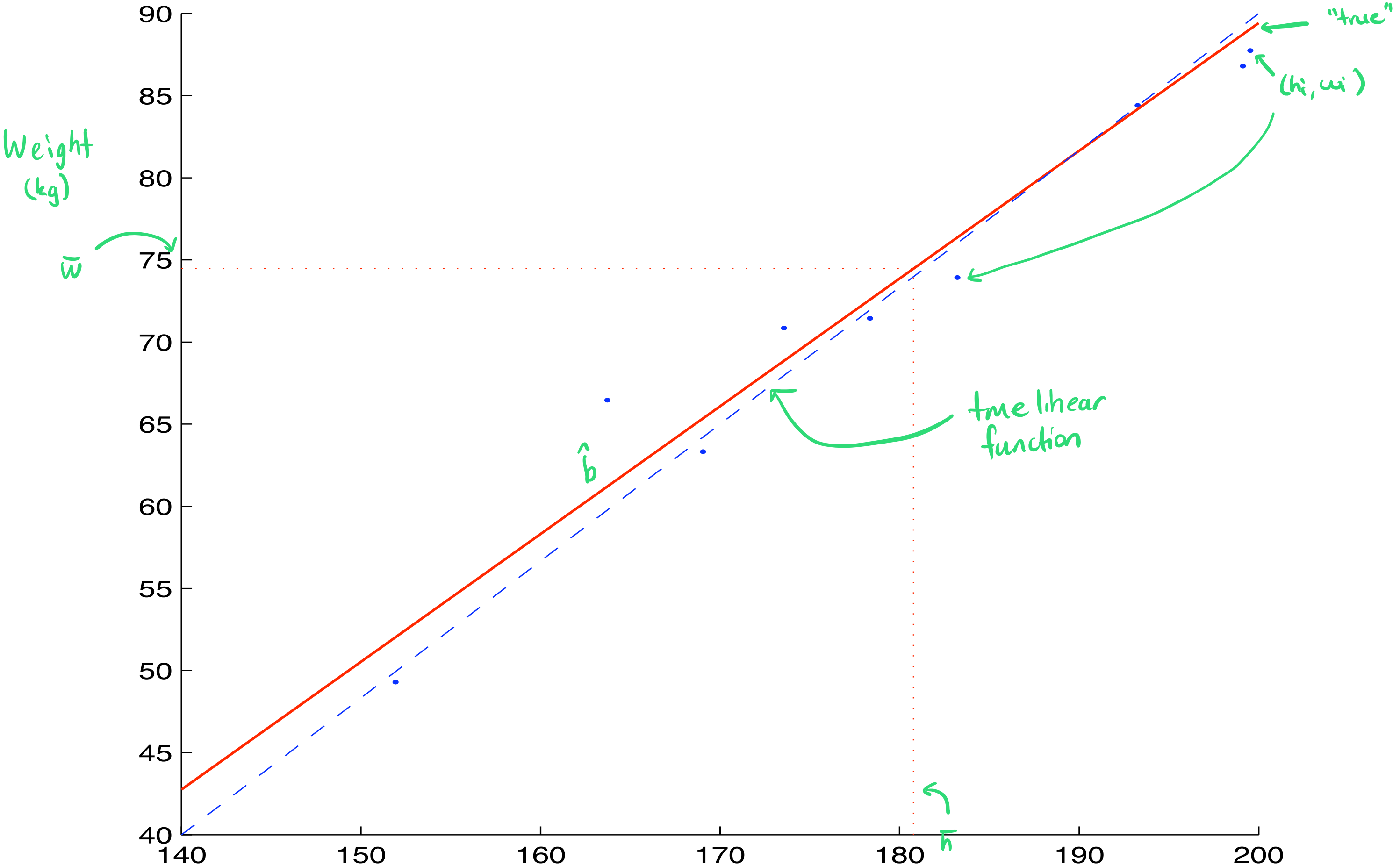
$$\frac{\partial}{\partial b}\sum_{i=1}^{n}(\underset{\substack{\uparrow\\ \text{data}}}{w_i}-(a+b\underset{\substack{\uparrow\\ \text{values}}}{h_i}))^2 = -2\sum_{i=1}^{n}(w_i-(a+bh_i))h_i = 0$$

$$\Rightarrow \hat{b} = \frac{\sum_{i=1}^{n}(h_i-\overline{h})(w_i-\overline{w})}{\sum_{i=1}^{n}(h_i-\overline{h})^2} = \frac{S_{hw}}{S_{hh}}$$

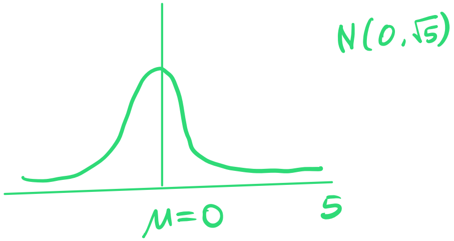
Linear Regression Model

So the solution found by linear regression is $\boxed{\overset{\substack{y\\ \downarrow}}{w} = \underset{\substack{\uparrow\\ \text{parameters}}}{\hat{a}} + \underset{\uparrow}{\hat{b}}\overset{\substack{x\\ \downarrow}}{h} = \overline{w} + \hat{b}(h-\overline{h}).}$

Univariate linear regression



Univariate linear regression



Shown on previous slide:

The red solid line indicates the result of applying linear regression to 10 measurements of body weight (on the y-axis, in kilograms) against body height (on the x-axis, in centimetres). The orange dotted lines indicate the average height $\overline{h} = 181$ and the average weight $\overline{w} = 74.5$; the regression coefficient $\hat{b} = 0.78$. The measurements were simulated by adding normally distributed noise with mean 0 and variance 5 to the true model indicated by the blue dashed line ($b = 0.83$).

true model: $\boxed{w = 40 + 0.83h}$

Linear regression: intuitions

For a feature X and a target variable Y, the regression coefficient is the covariance between X and Y in proportion to the variance of X:

$$\hat{b} = \frac{S_{XY}}{S_{XX}}$$

This can be understood by noting that the covariance is measured in units of X times units of y (e.g., metres times kilograms above) and the variance in units of X squared (e.g., metres squared), so their quotient is measured in units of y per unit of X (e.g., kilograms per metre).

Linear regression: intuitions

The *intercept* $\hat{a}$ is such that the regression line goes through $(\overline{X}, \overline{Y})$.

Adding a constant to all X-values (a translation) will affect only the intercept but not the regression coefficient (since it is defined in terms of deviations from the mean, which are unaffected by a translation).

So we could *zero-centre* the X-values by subtracting $\overline{X}$, in which case the intercept is equal to $\overline{Y}$.

We could even subtract $\overline{Y}$ from all Y-values to achieve a zero intercept, without changing the problem in an essential way.

Linear regression: intuitions

Another important point to note is that the sum of the residuals of the least-squares solution is zero:

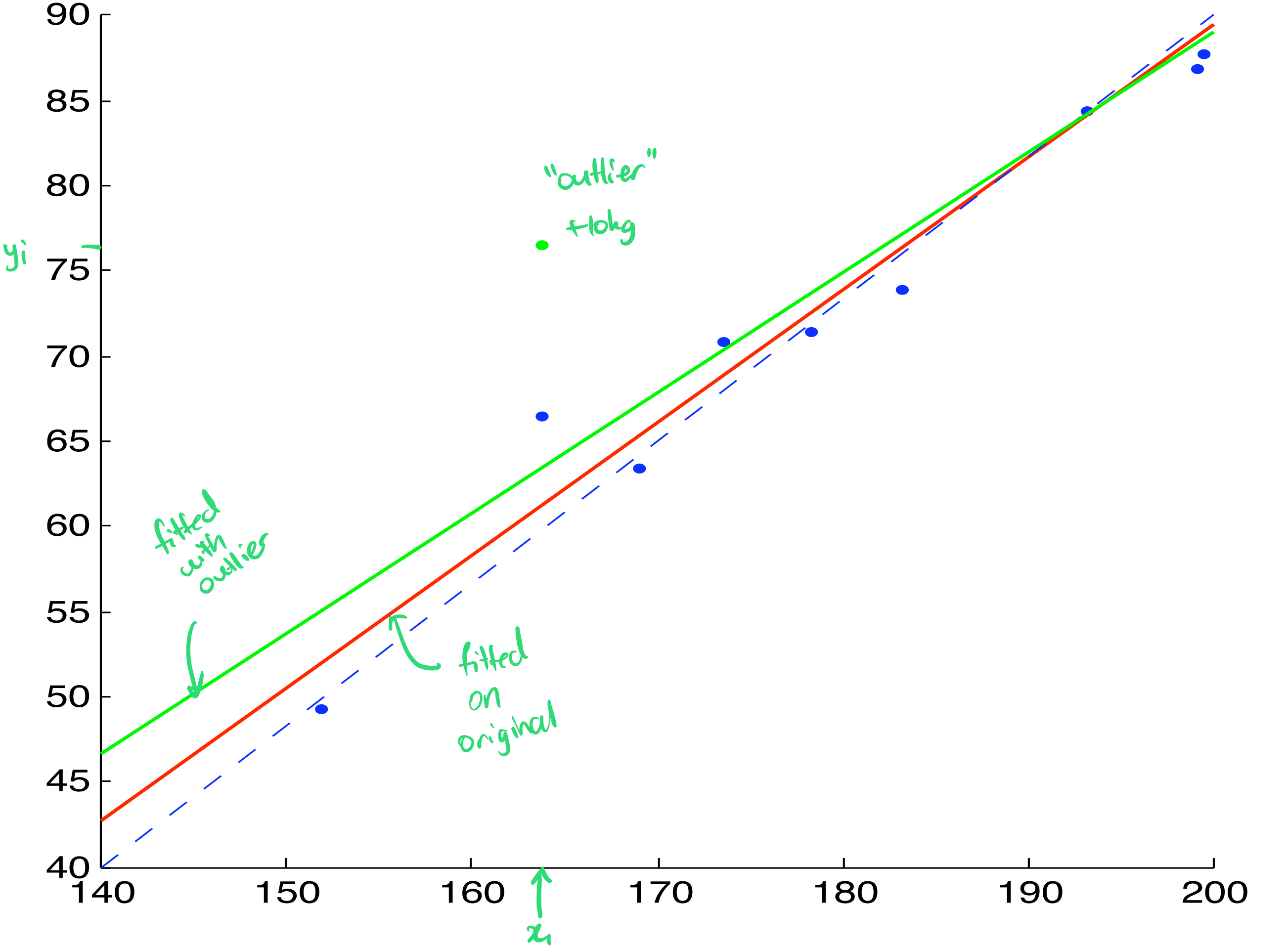
$$\sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i)) = n(\overline{y} - \hat{a} - \hat{b}\overline{x}) = 0$$

$\hat{a} = \overline{y} - \hat{b}\overline{x}$

The result follows because ~~$\hat{a} = \overline{Y} - \hat{b}\overline{X}$~~, as derived above.

While this property is intuitively appealing, it is worth keeping in mind that it also makes linear regression susceptible to *outliers*: points that are far removed from the regression line, often because of measurement errors.

The effect of outliers



The effect of outliers

Shown on previous slide:

Suppose that, as the result of a transcription error, one of the weight values from the previous example of univariate regression is increased by 10 kg. The diagram shows that this has a considerable effect on the least-squares regression line.

Specifically, we see that one of the blue points got moved up 10 units to the green point, changing the red regression line to the green line.

Flach, P. (2012). *Machine Learning*. Cambridge University Press.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill, New York.

End of Lecture 1 Questions (31/05/22):

Q1) When we calculate the sample variance, why is it 1/(n-1) instead of 1/n?

$$\mathrm{Var} = \frac{1}{n-1}\sum_i (x_i - m)^2$$

You need to understand the concept of bias of an estimator. Suppose theta_hat is an estimator of theta. Then bias(theta_hat) = E(theta_hat) - theta. When we estimate population variance with sample variance using n - 1 in the denominator, the bias turns out to be 0 whereas with n, the bias is -sigma^2/n.

Q2) What's the difference between r and r-squared?

R is the correlation coefficient, it's a number between -1 and 1 and tells us about the linear relationship between two variables. The R^2 metric is used in linear regression, in which we try to learn something about a response Y from a set of features X_1,..., X_p. The R^2 value tells us how much of the information (variance) in Y is explained by the features X_1,..., X_p. A good model should in theory have a high R^2.