

Introduction to Statistical Data Science

Ricardo Silva

ricardo@stats.ucl.ac.uk

Department of Statistical Science, UCL

STATISTICAL DATA SCIENCE

Data Science

- It will mean different things to different people.
- My minimalist definition: *the study of processes for extraction of information from data.*
- How is this different from Statistics?
- How is this different from Machine Learning, Data Mining, Predictive Analytics and so on?

Statistical Data Science

- Is the science and engineering behind a Google query also “data science”?
 - Yes, why not?
- In Statistics, though, we are concerned about **modelling uncertainty**.
- *Probability* will play an important role in either the **design of models**, or in understanding **properties of such models**.

Relation to Machine Learning

- Machine Learning has *Artificial Intelligence* (a.k.a., *autonomous systems*) as a key motivation.
- In principle, it tries to remove humans from the decision making loop.
- Down-to-earth applications: for instance, spam filtering, image recognition, advertising.
- Semi-autonomy by aiding decision making: for instance, detecting tumours in images.

This Module

- We will provide an overall view of the major statistical ideas.
- The intersection with machine learning is clear, but the emphasis is different.
- We will put less emphasis on prediction than in Machine Learning, more on modelling and analysing the properties of the models we develop. Prediction will still be important.
- Data Science is an iterative process, and the idea is to mimic this in class.

This Module

- Unlike the typical MSc Statistics module, this is designed to be more self-contained.
 - Including preparations to other modules.
- Emphasis is less on mathematics, more on learning by example.
- **Probability and Statistics will be intertwined.**
- **A degree of quantitative maturity is assumed, as well as some working knowledge of probability.**

Isn't This a Lot of Stuff?

- Yes, indeed.
- We will keep mathematics to a minimum, but as professional Data Scientists you need to be able to “talk Gaussian” sometimes.
- Office hours! **Make use of office hours!**

A Note on Examples

- Many examples will use R code.
- In *STATG003*, some of you will go into R programming in more detail.
- I don't need you to write R code in this module, but some exercises will be more interesting if you can (or MATLAB/Python/Julia/etc., if you prefer)
- Consider yourself highly encouraged to teach yourself the details of R coding used in my examples if not covered by *STATG003* yet.

OVERVIEW OF KEY CONCEPTS

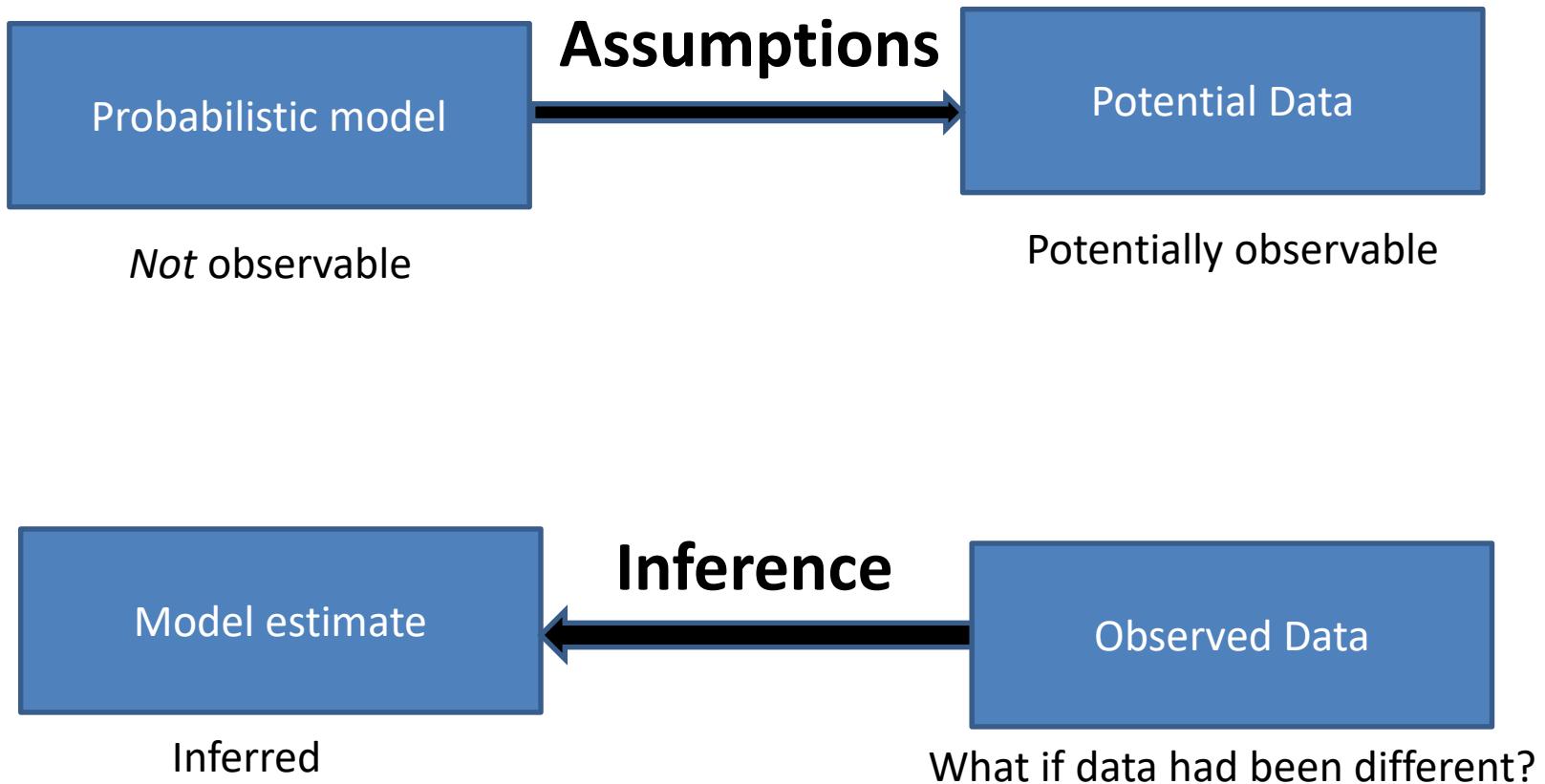
Outline

A first taste of statistical modelling:

From assumptions to data, from data to conclusions

1. Case study: human height and other measures
2. Tools: plots, simple probability models
3. A taster for statistical evaluation of models
4. A taster for computational aspects of model fitting
5. Conditional probability and estimation

A Sketch of Statistical Inference



A Sketch of Statistical Inference

- Without going into detail, in this module we will focus on **frequentist inference**.
 - Considering what would happen had the data been different, and the long run consequences.
- STATG004* teaches you about **Bayesian inference**.



A Working Example

- Third National Health and Nutrition Examination Survey
 - “*designed to provide national estimates of the health and nutritional status of the United States... population*”
 - http://www.cdc.gov/nchs/nhanes/nhanes3/data_files.htm
- We will consider measurements of sex, age, weight and height.

Activity

- Using chapter1.R from the Moodle page.
 - Needed also: nhanes.dat
- We will:
 - Examine heights
 - How they differ according to subgroups
 - How to model it
 - How to assess the variability of the model

Samples and Populations

- There is a group of people which we collected data from (**the sample**) and which we did not collect data from.
- All of these potential people form a **population**.
- **Inference: we want to characterize the population.**
 - What does it mean in a day and age where we can collect data much more cheaply than ever?

Samples and Populations

- Sometimes we can collect “all” the data. Is that the population? Is inference just descriptive in this case?
- **Populations can be infinite, never mind that you have “all” the data:**
 - A population can include future units: people to be born, new products to be recommended, new pesky spams to trash, political campaigns to be run, etc.

Samples and Populations

- As a matter of fact, a **statistic** is just a function of the data, that is, any summaries of your sample.
- It might sound dull, but it emphasizes an important distinction: statistics are based on quantities you directly observe.
 - If it depends on unknown quantities, then it is NOT a statistic.

Samples and Populations

- **Judicious assumptions are necessary to relate sample to population.**
- This includes why some units ended up being in the sample and why some did not.
 - You can learn more about the subtleties of this step in *STATG002*.

Samples and Populations

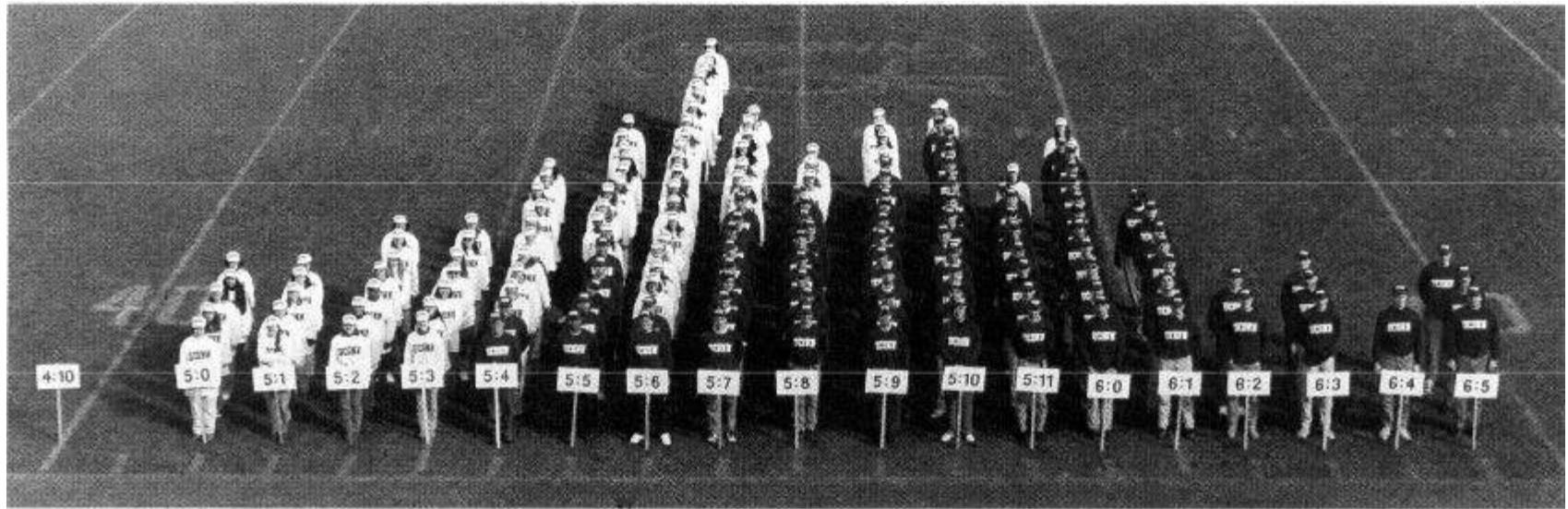


Figure 7. Living histogram of 143 student heights at University of Connecticut.

Schilling et al., (2002) "Is human height bimodal?".
The American Statistician Volume 56, Issue 3, 2002

“I take no prisoners, I make no assumptions”

- “Data speaks for itself”
 - If you believe that, this is an excellent time to leave.
- Data does not speak for itself because there is only so much information about the population that a sample can give you.
 - Do the math: for instance, if the population is infinite and your sample is finite, then...
 - **Mind the gap.** Assumptions matter, but this is **not** the same as blind belief.

Our First Link to Probability

- We assume that the variability in our data is generated probabilistically.
- Heights can be considered to be real, continuous numbers.
 - What can we possibly mean by “the probability of that person’s height being 1.90 meters is 0.002”?
 - Consider instead “the probability of that person’s height being between 1.89 and 1.91 meters”.
 - How to describe this more generally?

Randomness

- Our data can be interpreted as a collection of **random variables**.
- There is a formal mathematical definition of a random variable. It suffices to say it is a quantification of random **events**.
 - A series of random events gave you your current height, coming from some **sample space**. Your measurement correspond to the outcome of this process is a random variable.

Random Variables

- We can play with a random variable in an algebraic way, like ordinary variables.
- Say X is your weight (kg) and Y is your height (m), we can define a random body mass index:

$$Z = \frac{X}{Y^2}$$

- In our setup, Z is itself a random variable, because it is a function of random variables.

Notation

- We will use upper case letters to denote random variables (X, Y, Z , etc.).
- We will use lower case letters to denote values taken by the random variables (x, y, z , etc.)

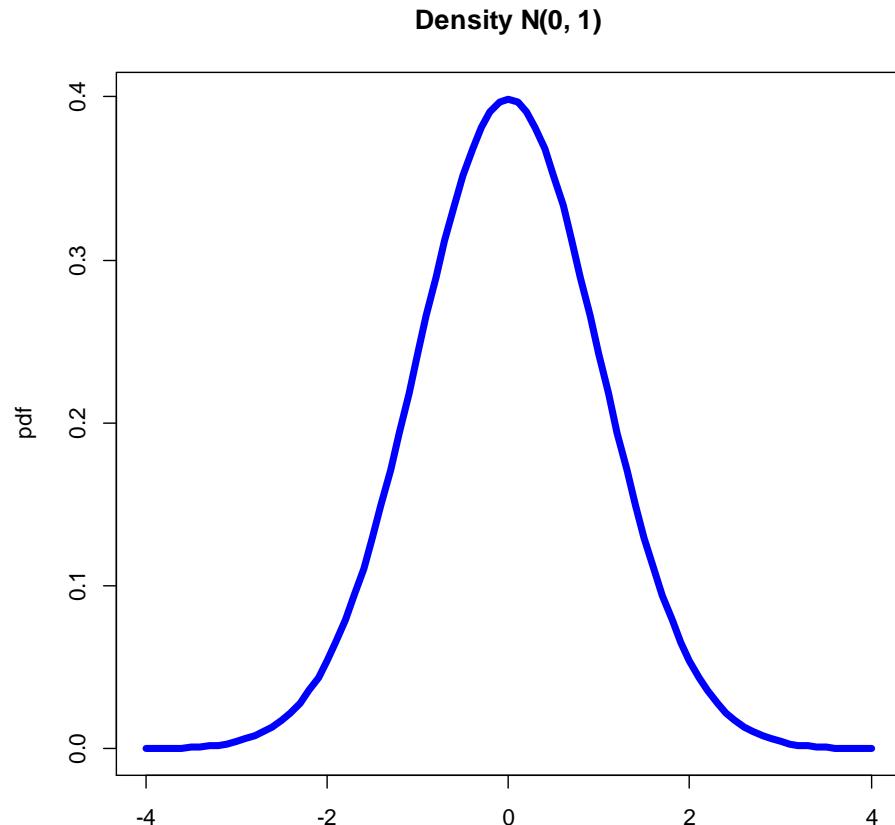
Scope

- We will typically divide random variables into two main classes: **discrete** and **continuous**.
- A discrete r.v. takes values from a “discrete” set.
 - Sounds obvious, but discrete sets can be infinite too.
 - Discrete random variables can represent categories as “male”, “female”, “heads”, “tails”, etc., but formally they are encoded as numbers (0, 1, etc.)

Example

- Say Y means height of a person (in meters).
- Let's simulate randomness by **subsampling**. I will generate multiple datasets of size 10000 from the original dataset.
- (R demo)

Probabilities, Densities and the Gaussian (or Normal) Distribution



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

We will explain
what these
mean soon

Density Functions

- What do we mean by **probability density function (pdf)**?
 - Recall we can't assign positive probabilities to outcomes in continuous spaces
 - We can change our thinking towards the events of finding outcomes in a particular interval.
- General idea: the **cumulative distribution function (cdf)**

$$F(x) \equiv P(X \leq x)$$

From cdfs to pdfs

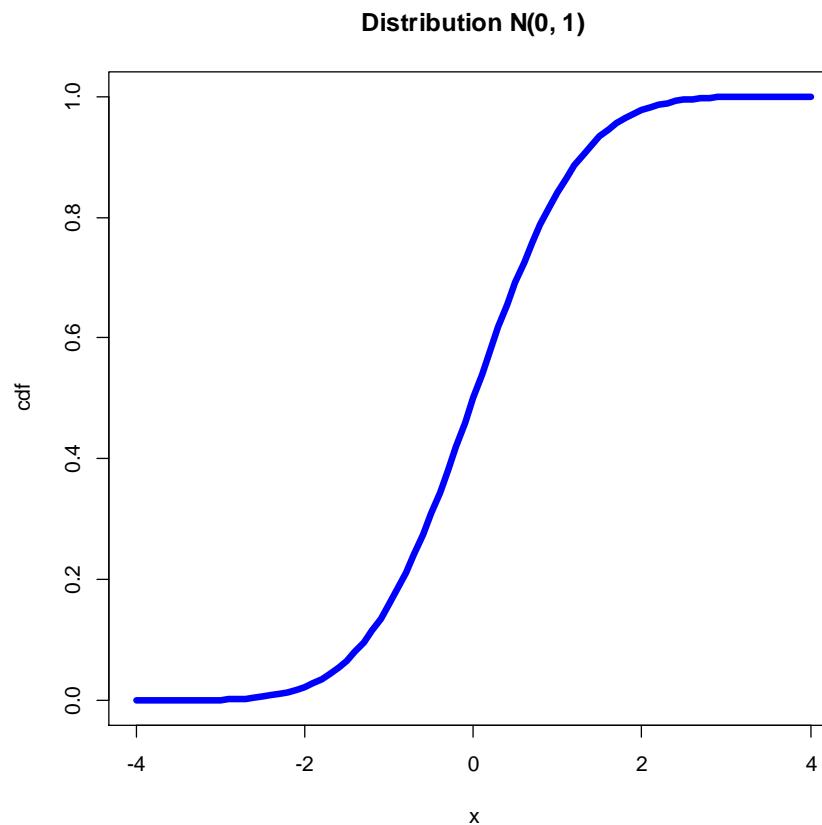
- For continuous data, the density function is just the derivative of the CDF.

$$p(x) \equiv \frac{dF(x)}{dx}$$

- In the majority of cases, it is easier to think with pdfs than cdfs.
- We will deal with discrete random variables later. In that context, we speak of **probability mass functions (pmfs)**, i.e. the probability of a discrete X taking a particular value.

For Illustration Purposes

- The Normal $(0, 1)$ cdf :



(Notice monotonicity, bounded between 0 and 1.)

The Normal a.k.a Gaussian a.k.a Bell Curve

- Physical motivation: under very general conditions, if we average many random variables, their distribution will converge to a Gaussian as we average more and more variables.
- This is known as **Central Limit Theorem**, and there are several variations of it (see Rice, Wasserman, etc.).

Demonstration

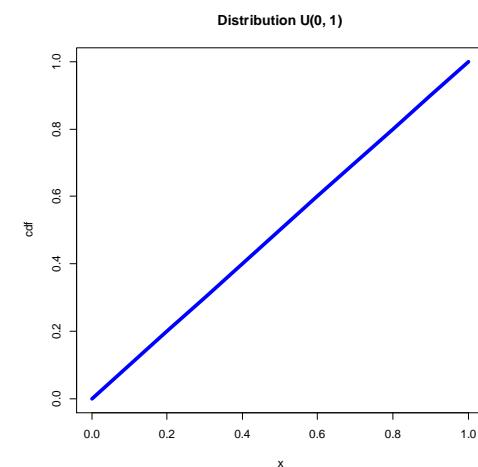
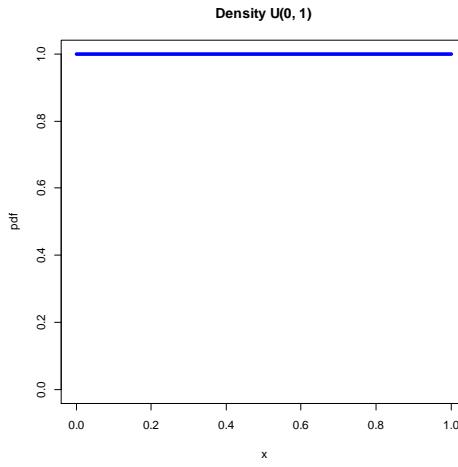
- Consider the **Uniform distribution** in $[0, 1]$.
We will generate a random variable X following it. Our notation:

$$X \sim U(0, 1)$$

$$p(x) = 1, F(x) = P(X \leq x) = x,$$

support

$$0 \leq x \leq 1$$



Demonstration

- Now, consider a **vector** of n independent and identically distributed random variables $X^{(i)}$, $i = 1, 2, \dots, n$:

$$X^{(i)} \sim U(0, 1)$$

Demonstration

- We will sometimes use bold faces to denote **random vectors**:

$$\mathbf{X} = \begin{bmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(n)} \end{bmatrix}$$

- Now, say you don't see \mathbf{X} . All you observe is the average

$$Y = \frac{1}{n} \sum_{i=1}^n X^{(i)}$$

Demonstration

- Finally, let's say you have measured Y_1, Y_2, \dots, Y_p , all of them following the same recipe with independent $X_j^{(i)}, j = 1, 2, \dots, p$.

$$Y_j \sim ?$$

- Let's see what happens as $n \rightarrow \infty$
 - R demo, $p = 1000$, $n = 1$ to 10.

Going Back to Our Example

- Let's assume that human height follows a normal/Gaussian distribution, **by whatever physical process, which we will just abstract away.**
 - Probabilities as a summary of ignorance.
- This is a model. **All models are approximations.** For instance, Gaussian random variables can be negative. Heights can't.

Which Gaussian?

- **Which Gaussian distribution?** The Gaussian has two **parameters**, a **mean** μ and a **variance** σ^2 (alternatively, **standard deviation** σ).

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

Parameters

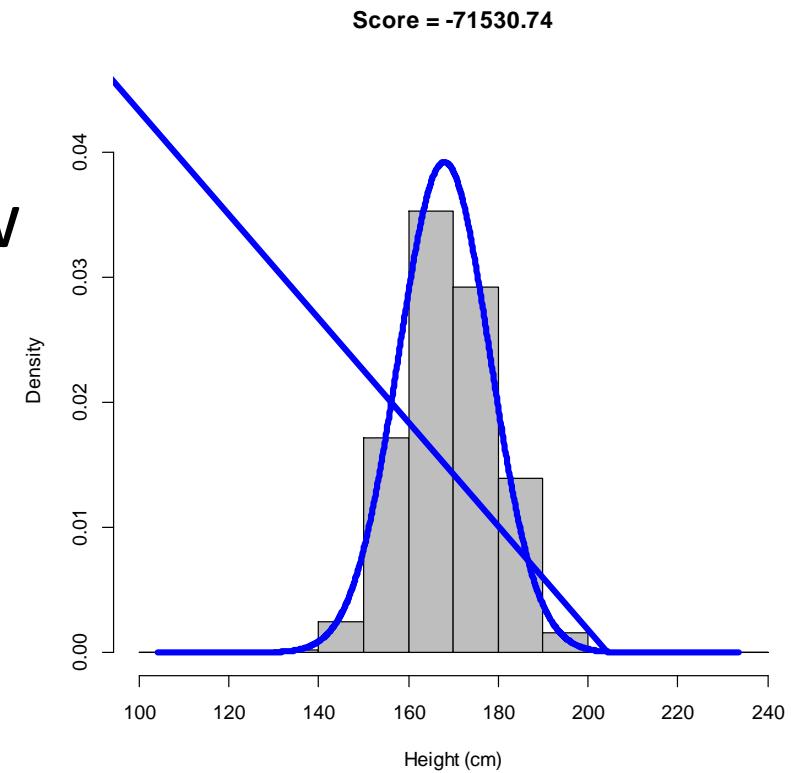
- The mean is a **location** parameter. It says where the “peak” of the density will be located.
- The variance is a **scale** parameter. It controls how the mass “spreads” around the mean.
- R demo.

Fitting

- What in Machine Learning is called “learning”.
- Adjust the free parameters so that the **implied model** “matches” the data somehow.
- There are different **scores** to quantify the quality of the matching, and different **algorithms** to perform the optimisation of the score. We will see some later on.
- R demo.

Fitting

- From this data we can say a good **estimate** of the model is $\hat{\mu} = 168$
- $\hat{\sigma} = 10.2$
- (We will explain later how to get those)



Verifying Fit

- Is a Gaussian model a good model?
- We can compare a “good” Gaussian cdf against the model **empirical cdf**.

$$F_n(x) \equiv \frac{1}{n} \sum_{i=1}^n I(x^{(i)} \leq x)$$

where $x^{(1)}, \dots, x^{(n)}$ is our observed data.

- R demo.

Verifying Fit

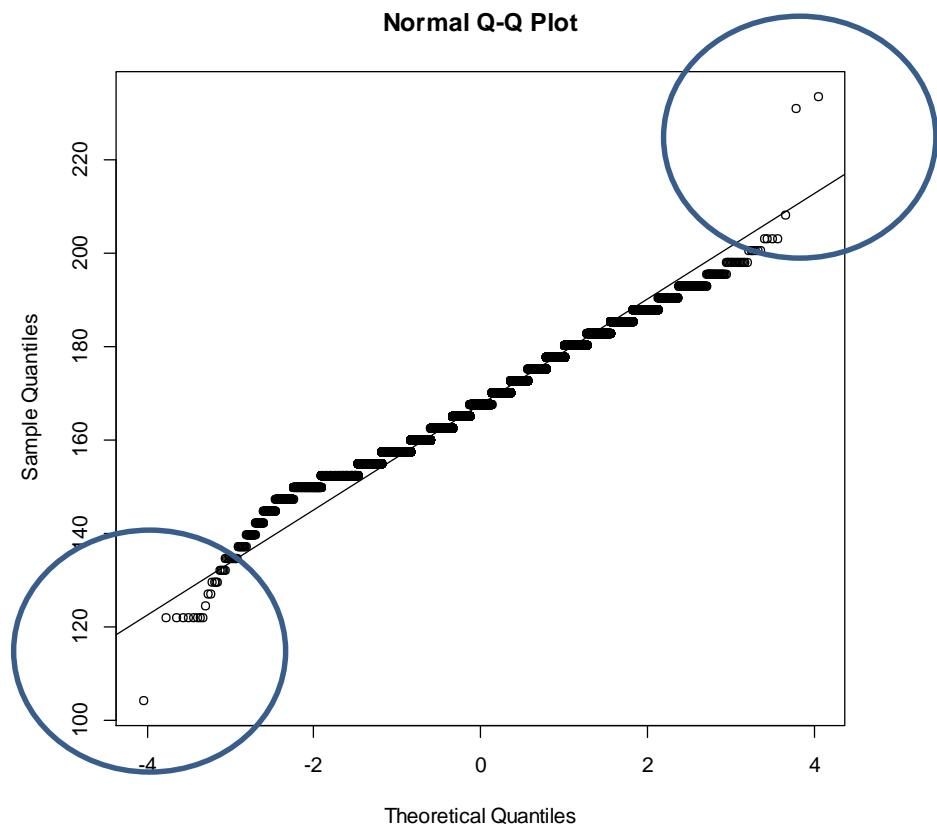
- The story doesn't end there. We might want to see whether this comparison is the result of chance or not.
- Remember: our data is assumed to have been generated by a probabilistic process. It could have been different.
- Part of the analysis is to quantify how likely is for the data to “look Gaussian” even if the population isn’t.
- A topic for another day.

The Quantile-quantile Plot

- Another way of (visually) assessing the fit is the quantile-quantile plot (qqplot).
- A **quantile** is just the inverse of the cdf. So e.g. the 0.9 quantile is the value of x such that $F(x) = 0.9$ and, as such, $F^{-1}(0.9) = x$.
- Some quantiles have special names. The **median** is the 0.5 quantile.
- The qqplot is an idea similar to comparing cdfs.

QQ Plot

- Clear outliers. The Gaussian assumption is a decent approximation, but fails at the **tails** of the distribution.



And Now What?

- We have learned something: an estimate of the (approximate) distribution of heights from a population.
- What can we do with it?
 - For instance, learn how to keep a stock of clothing items in a shop.
 - Snapshots vs trends: we haven't done that, but we can think of collecting data across time periods, assuming different populations so that we can compare them.

How to Compare?

- We can pick a summary of interest. For instance, the **expectation** of the distribution.

$$E[X] \equiv \int xp(x) \, dx$$

- Think of it as a weighted average, or “centre of mass” as it were. We also call this the **mean**.
- We might want to estimate the expectation of the distribution. So this becomes our **estimand**: a feature of the distribution that is a quantity of interest.

Side note

- We can have expectations of arbitrary functions of random variables, since they are also random variables:

$$E[f(X)] = \int f(x)p(x) \, dx$$

Are Expectations Good Summaries?

- It depends. For instance, if the data follows a Gaussian distribution

$$X \sim N(\mu, \sigma^2)$$

then

$$E[X] = \mu$$

- (You might want to check that yourself. Remember your calculus)

$$E[X] = \int x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} dx = \mu$$

When are Means Bad Summaries?

- Highly **skewed** (asymmetric) or **multi-modal** (multiple “bumps” or peaks) distributions.
 - Average life expectancy, Classical Rome: 20-30 years (https://en.wikipedia.org/wiki/Life_expectancy).
 - However, 47.5 *given* surviving past 10 years of age.
 - In *STATG015* and *STATG016*, for instance, we discuss particular ways of modelling life expectancy through models of *survival analysis*.

NHANES Data

- So, we had found already some
 $\hat{\mu} = 168$
- An estimated height expectation of 16.8cm.
But what if the data had been different?
- That is, we had here 19,219 individuals in this study. What if **different** 19,219 individuals had been sampled?
- We will study this later on, when we discuss **confidence intervals**.

The Dual Use of Probability

- That is, we will use probability not only to **express assumptions about the population.**
- We will also use it to analyse **sampling properties** of our **estimators.**

More About Summaries

- The Gaussian has a mean and variance parameters, which may confuse you because “mean” and “variance” also refer to general summaries of a distribution.
- For instance, what we mean in general as variance of a random variable X is:

$$Var(X) \equiv \int (x - E[X])^2 p(x) \, dx$$

This is just a way of quantifying dispersion (how “spread out” $p(x)$ is). There are others.

More About Summaries

- Without wanting to bore you, you may come across the notion of **moment** elsewhere. There are just summaries of the type $E[X^d]$.
- For example, the mean is “the first moment”, because we can get it for as the moment where $d = 1$.
- Variances can be written as a function of first and second moments:

$$Var(X) = E[X^2] - (E[X])^2$$

Multivariate Data

- We had already started by talking about two measurements: sex and height.
- These are of not **independent**, because we can see that the distribution of height is different in either case.
- In **multivariate models** we are interested in modelling such dependencies. It is vital in tasks such as **prediction**.

Prediction

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski^{a,1}, David Stillwell^a, and Thore Graepel^b

^aFree School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom; and ^bMicrosoft Research, Cambridge CB1 2FB, United Kingdom

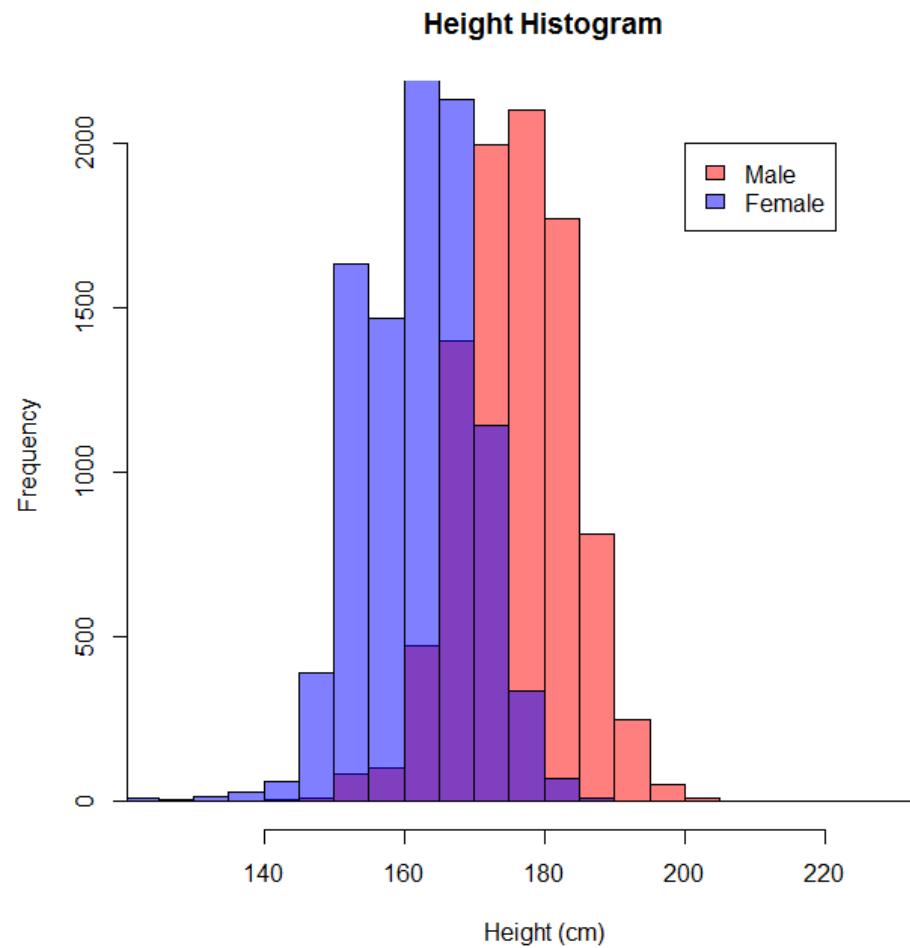
Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. The analysis presented is based on a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric tests. The proposed model uses dimensionality reduction for

browsing logs (11–15). Similarly, it has been shown that personality can be predicted based on the contents of personal Web sites (16), music collections (17), properties of Facebook or Twitter profiles such as the number of friends or the density of friendship networks (18–21), or language used by their users (22). Furthermore, location within a friendship network at Facebook was shown to be predictive of sexual orientation (23).

This study demonstrates the degree to which relatively basic digital records of human behavior can be used to automatically

Can We Predict Sex from Height?



Adding More Dimensions

- We can visualise how sex differences appear across two dimensions, height and weight.
- Notice that height and weight themselves are of course **associated**, and should not be counted as independent pieces of information.
- R demo.
- We will see how to formalize association and prediction problems more precisely in the next chapter.

Summary: Take-Home Messages

- Probability vs statistical inference:
models to data vs data to models.
- Two simple models so far: Gaussian
("normal") and the Uniform distributions.
- Notions of parameter estimation: explain the
data by matching model to data.
- Multivariate data, associations and prediction.
- Next topic: testing and confidence intervals.

Introduction to Statistical Data Science

Ricardo Silva

ricardo@stats.ucl.ac.uk

Department of Statistical Science, UCL

Statistical Assessment: Hypothesis Testing and Confidence Intervals

Challenging your Assumptions and Conclusions

- The phenomenon I'm studying seems to have an interesting property. **Is it real?**
- Data doesn't speak by itself. So we need assumptions. **Are they true?**
- Our assumptions implied conclusions. **How would these conclusions change if we had observed different data?** What can I say about the long-run behaviour of my conclusions?

Outline

- Hypothesis tests and p-values: the good, the bad and the ugly
- Confidence intervals
- Computational aspects: Central Limit Theorem and the Bootstrap

Hypothesis testing

A PRELIMINARY EXAMPLE

A Simple Initial Example

(The following is synthetic and created for the sake of illustration).

- Suppose you offer a training course. Is there gender imbalance among your participants?
- With a large sample and a large discrepancy, this might be easy to conclude. No “need” for statistical inference there.
- However, for example, what would you say if we observe 15 out of 40 participants are female? For simplicity, let’s assume that the true proportion of females is not greater than 0.5.

A Hypothesis Testing Approach

- What is it that we would like to test?
 - Is 0.5 the probability of the event of any given student being female?
 - Why 0.5?
- Technical term: **null hypothesis**.
- The idea is: assume the null is true. What would be the probability of observing the data we have?

Test Statistic

- Recall: a statistic is a summary of the data.
- A test statistic is a summary that can falsify the null, if it is indeed false.
- There are different ways of picking a test statistic, more on that later. In our example, intuitively, the number of female students provides such a summary.

Complementary Assumptions

- On top of the null hypothesis, we often need to make further assumptions to characterize the test statistic.
- In our example, we will assume that each student decides to enrol independently, so the sex of each student is independent of each other.
- Encoding the random variable: 0 for male, 1 for female. So we have **independent Bernoulli random variables** (“coin flips”).

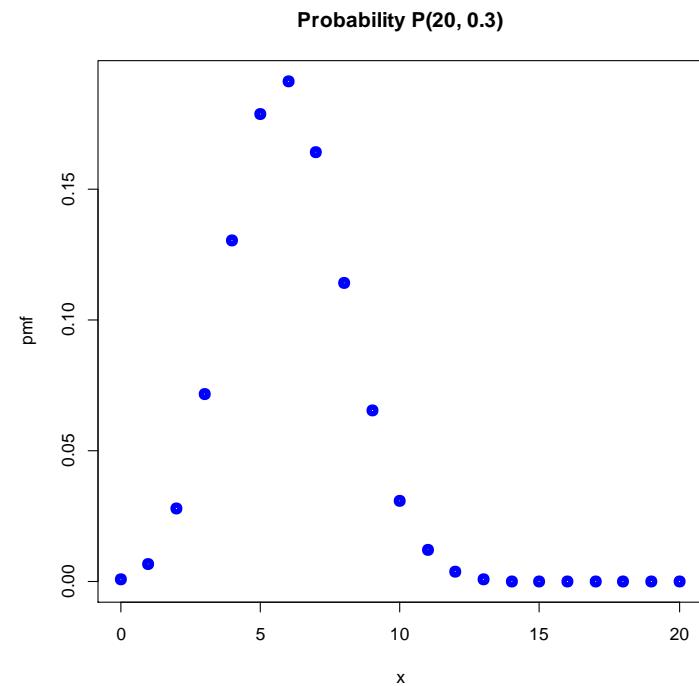
The Binomial Distribution

- If we have n independent Bernoulli trials, each with probability θ , then we have a **binomial distribution**.

$$X \sim Bin(n, \theta)$$

$$P(X = x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

(See R code to play with visualisations)



("pmf" here means **probability mass function**)

Back to Our Test Statistic

- In our example,

$$Y_1, \dots, Y_{40} \sim \text{Bernoulli}(0.5)$$

in an **independent, identically distributed (i.i.d.)** way.

- That is, we can show that

$$X \equiv \sum_{i=1}^{40} Y_i \sim \text{Bin}(40, 0.5)$$

and the more general statement says the sum is binomial (n, θ).

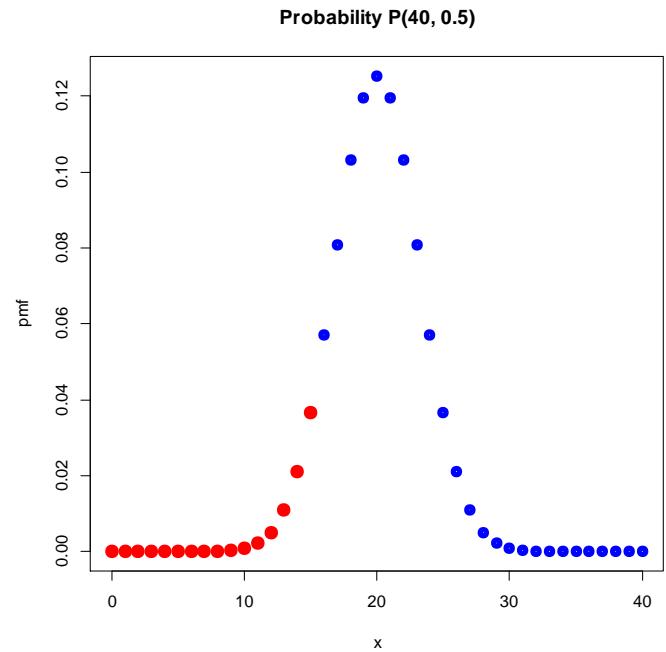
How is This Relevant?

- We can characterize whether the x we saw in our example (15) is likely under the null hypothesis H_0 that $\theta = 0.50$.
- As a matter of fact, let's characterize how probable values of 15 or smaller are.
 - Why? We will see.

Writing Down our Target: the p-value

$$p \equiv P(X \leq 15; H_0)$$

- That is, $p = \sum_{x=0}^{15} \binom{40}{x} 0.5^x (1 - 0.5)^{(40-x)}$
- p here is approximately 0.07.
What is your conclusion?
- Decision thresholds are used.



More on Interpretation

- The p-value is the probability of observing a test statistic X “at least as extreme as” the value x seen in the data, **assuming H_0 is true**.
- It is NOT the probability of H_0 being true!
- Let’s take a small detour to learn (or remind ourselves) of that.

$$P(H_0 \mid T = t) \neq P(T = t \mid H_0)$$

- More precisely

$$P(A, B) = P(A \mid B)P(B) \quad P(H_0 \mid T = t) = \frac{P(T = t \mid H_0)P(H_0)}{P(T = t)}$$

- This needs some $P(H_0)$ to be defined, which is not always easy.
- The inverse statement is discussed at length in *STATG004*, Bayesian Analysis.

A Crude Analogy

- In Logic, the implication

$$A \Rightarrow B$$

comes with the contrapositive

$$\neg B \Rightarrow \neg A$$

- The unwritten “logic” of hypothesis testing is that H_0 should imply “with high probability” values seen. If they don’t occur, this disproves H_0 by an informal contrapositive argument.

A Crude Analogy

- Although used in practice with a threshold of 0.05, this is an informal way of reasoning (“unlikely B implies unlikely A” is not really a contrapositive!) and can be easily criticized.
- We will see other interpretation based on long-run trade-offs between “false positives” and “false negatives”.
- Ultimately, we will also present a pragmatic guide on when/why use null hypothesis testing. This is not a tool without controversies.

A Crude Analogy

- One aspect of the contrapositive analogy is useful.

$$\neg B \Rightarrow \neg A$$

- This just means *something* in your assumptions is wary. It is easy to forget the problem might be with our other implicit assumptions.
- Can you point out what else might falsify the hypothesis in our example?

An Example to Remember

Journal of Personality and Social Psychology
2011, Vol. 100, No. 3, 407–425

© 2011 American Psychological Association
0022-3514/11/\$12.00 DOI: 10.1037/a0021524

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of *psi* are *precognition* (conscious cognitive awareness) and *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by "time-reversing" well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (d) in *psi* performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with *psi* performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about *psi*, issues of replication, and theories of *psi* are also discussed.

Keywords: *psi*, parapsychology, ESP, precognition, retrocausation

Hypothesis Testing

PROBABILITIES, POWER AND TYPES OF ERROR

**NOW YOU'RE
PLAYING
WITH POWER.**

- Statistical testing involves a trade-off between “false positives” (rejecting null hypotheses that are true) and “false negatives” (not rejecting those that are false).
 - “not rejecting” is not the same as “accepting”!
- The **power** of a test is the probability of avoiding a false negative.
 - But it needs an alternative hypothesis!

Going Back to our Example

- Let's define our rule "reject H_0 " as "reject if the event of observing an extreme count has probability 0.05 or less under the assumption H_0 is true" (which may not be).
- Two things to consider:
 - The p-value is actually a random variable!
 - How does the probability of the p-value being less than 0.05 change depending on "the way" H_0 is false?

The p-value as a Random Variable

- Take the view of p as a *black-box function of our data* (which you should recognize):

$$p_v(x) = \sum_{i=0}^x \binom{40}{x} 0.5^x (1 - 0.5)^{(40-x)} = F(x)$$

- This is for a fixed dataset that spits out summary statistic x (15, in our example).
- When we write $p_v(X)$, i.e. capital X , we are indicating that, *since the data generating process is random, so is the p-value*. Still with me?

The Distribution of p-values

- Assume continuous X for simplicity

$$P(F(X) \leq z) = P(F^{-1}(F(X)) \leq F^{-1}(z)) = P(X \leq F^{-1}(z)) = z$$

- What is the distribution where $P(Z \leq z) = z$, and where z is in $[0, 1]$? Uniform($0, 1$)
- p-values are uniformly distributed in $[0, 1]$, **under the null.**
 - Think of the intuitive explanation for it.
- Implications?

Error Control

- *If* the null hypothesis is true, *and* I reject it only when my test statistic is below the 0.05 quantile, *then* the probability of erroneously rejecting H_0 when it is true will be 0.05.
$$P(X \leq F^{-1}(0.05); H_0) = 0.05$$
- We say that $(-\infty, F^{-1}(0.05))$ is the **critical region** of this test, and that the **Type I error** is 0.05.

Frequentist Interpretation and Practical Motivation

- Of course we don't expect you to collect data for the same phenomenon over and over again. Error calibration is about using the procedure over a long range of problems.
- Of course this is an idealization. There will be approximations (the distribution of X is often not known exactly) and mistakes. But the idea is to be “less wrong”, if done appropriately.

Level

- In our example, 0.05 was the **level** of the test. The choice of level is problem-dependent. 0.05 is just an example, even if it the scientific literature appears to have a poorly motivated appetite for it.
- One way where the choice of level can be guided is by trading off **Type I** and **Type II** errors.

Type II Errors

- Erroneously failing to reject (“accept”) the null when it is false.
 - “But don’t we know it is always false?”
- The probability of avoiding this is the power of the test, as we introduced before.
- Unlike the level of the test, this will in general depend on what the true hypothesis is.

Type II Errors

- Power will vary with sample size (as it changes the distribution of the test statistic) and with the level of the test (as it changes the rejection region).
- When we speak of a trade-off, we mean level vs. power at a fixed sample size. Increasing sample size will increase power without changing the level.

(R demo)

Hypothesis Testing

THE STORY SO FAR

Recap

- We see the output of 40 “coin flips”. We observe 15 were “heads”.
 - For instance, 40 students in a class, where $Y^{(i)} = 0$ if male, $Y^{(i)} = 1$, if female. Summing all $Y^{(i)}$ gives us 15. Let’s call that X , so $X = 15$ in this example.
- The sum X follows a Binomial with $n= 40$, and “heads” probability θ , which we don’t know.
- I would like to check whether $\theta= 0.5$.

Recap

- What we will ***not*** do: treat θ as a random variable. If we are talking about a *specific* “coin”, then this is not a repeatable event. The following would not make sense in this scenario:

$$P(\theta = 0.5 \mid X = 15)$$

This will make sense in *STATG004*, but not without a “subjective” interpretation.

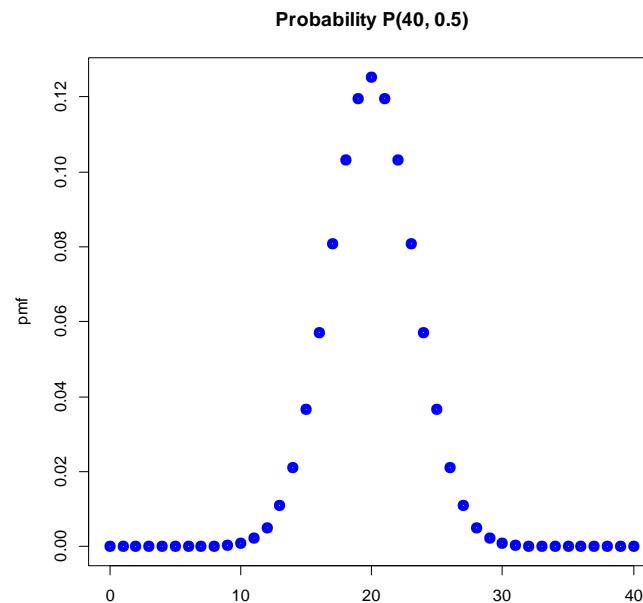
- Instead, we will focus on designing a **decision rule** that will have some **desirable behaviour** in terms of its *long-run frequency of applications*.

Recap

- Desirable behaviour?
 - If we apply your decision rule to many many problems, we will have some understanding of the frequency of mistakes we make in the long run.

Recap

- This is what the pmf of a $\text{Binomial}(40, 0.5)$ looks like.



- So we may *reject* the hypothesis of $\theta = 0.5$ if our observation is part of an unlikely event.

Recap

- How unlikely? Let's say, if the statistic falls in a region that corresponds to an event of probability 0.05. We call this **the level**.
- Which event? It pays off to look at “extreme values” of the statistic. Small number of “heads”, for instance, which would be “more likely” under *alternative hypotheses*.

Recap

- The two hypotheses, with alternative H_1 :

$$H_0 : \theta = 0.5$$

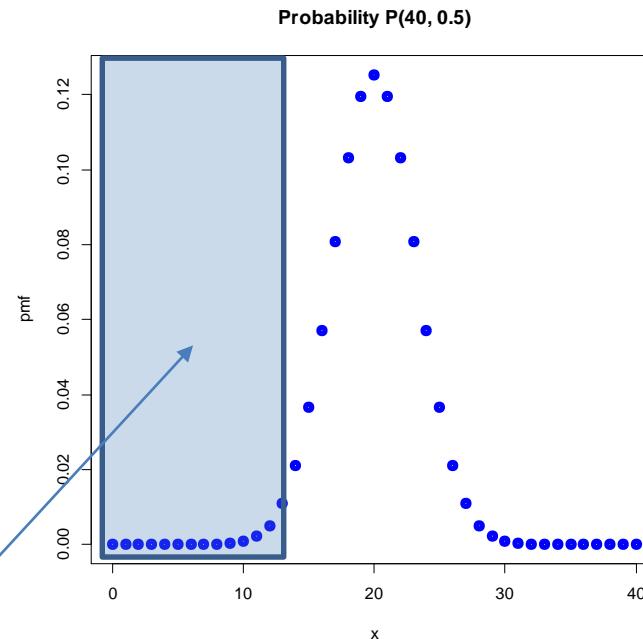
$$H_1 : \theta < 0.5$$



Yes, it could be $\theta \neq 0.5$.
This is my choice, assuming
it makes sense in the
real world.

Recap

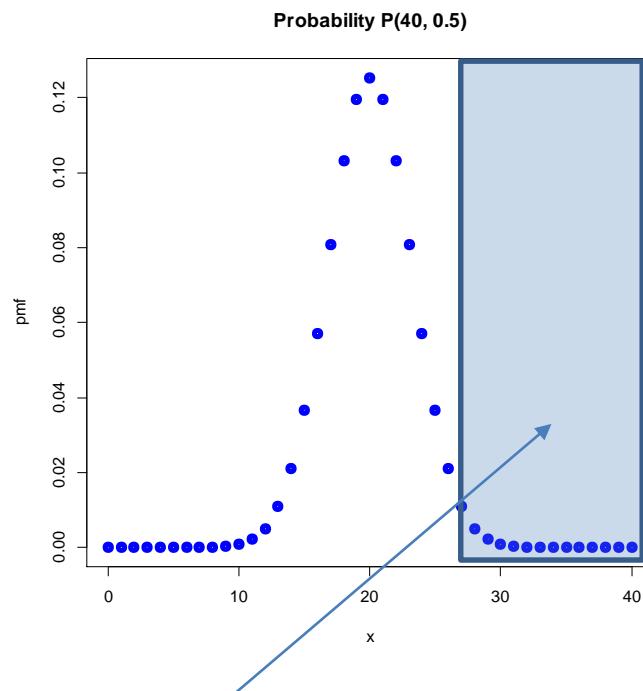
- So we could choose the event of “my count is among the smallest possible”.



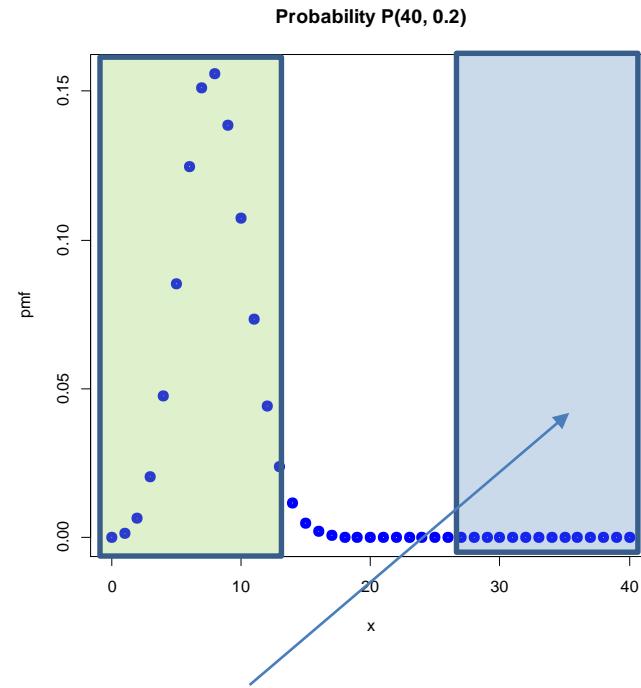
Zone of low probability (lots of points, but with very low mass)

Recap

- Why not this? It does not help to distinguish 0.5 from alternatives (remember H_1 here).



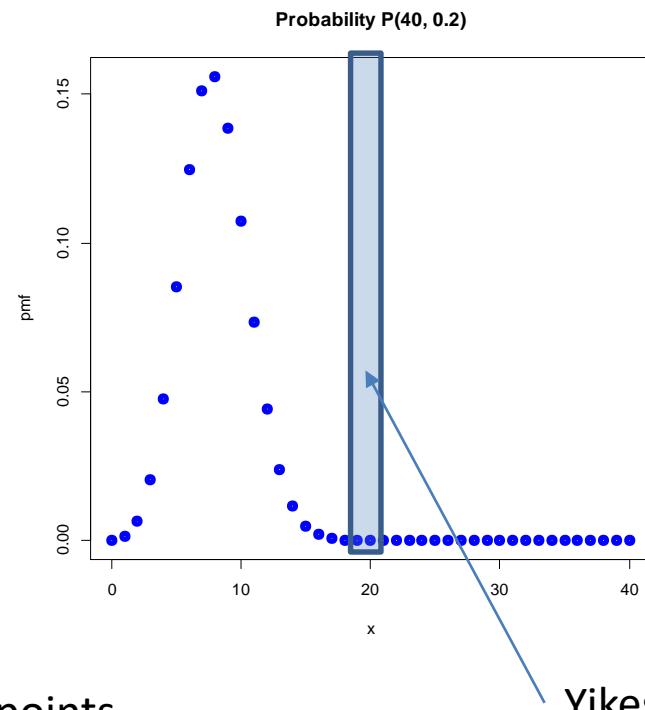
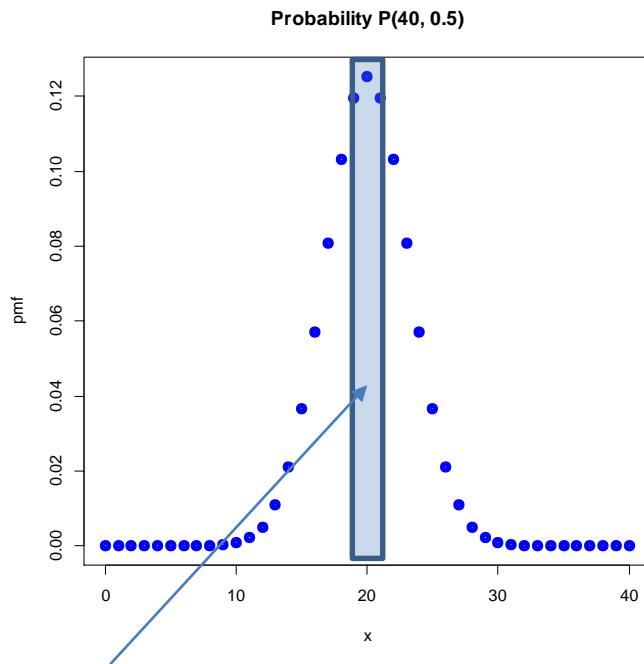
Zone of low probability



Zone of even lower probability

Recap

- Why not this? Same reasoning.



Zone of low probability, because there are very few points.

Yikes!

Recap

- So we agreed that, after looking at our statistic x , if our **p-value** $p_v(x) \equiv P(X \leq x)$ is smaller than 0.05, then we reject H_0 .
 - X here must follow a *Binomial(40, 0.5)* by definition!
- Rephrasing it in a equivalent way, reject if and only if x is smaller than the 0.05 quantile of a *Binomial(40, 0.5)* (**the critical region**).

Recap

- What is the probability of making a mistake by following this decision rule?
- Two types of errors:
 - **Type I**: rejecting H_0 when it is true
 - **Type II**: failing to reject (“accepting”) H_0 when it is false.

Recap

- What is the probability of a Type I error (wrongly rejecting H_0)?
 - It is the probability of our $p_v(X)$ being less than 0.05, a.k.a. the probability of our X being less than the 0.05 quantile of the *Binomial(40, 0.5)*
 - From both points of view ($p_v(X)$ is $\text{Uniform}(0, 1)$), we can see that this is exactly 0.05.

Recap

- And, it never hurts to emphasise, this “0.05” control is what we get in the long run by applying this decision rule to many many problems.

Recap

- Great, we got the understanding of one major property of our decision rule.



Type I error controlled

Recap

- What about power? It will depend on the actual state of nature θ .
- Recall our critical region. For a $Binomial(40, 0.5)$, the corresponding quantile is (approximately) 15.
- **Let's now fix the critical region " $X \leq 15$ " in stone, and compute the probability of that event under a variety of alternative θ s.**

Recap

- What if $\theta = 0.2$? $X \sim \text{Binomial}(40, 0.2)$, hence

$$P(X \leq 15; \theta = 0.2) \approx 1$$

- What if $\theta = 0.3$? $X \sim \text{Binomial}(40, 0.3)$, hence

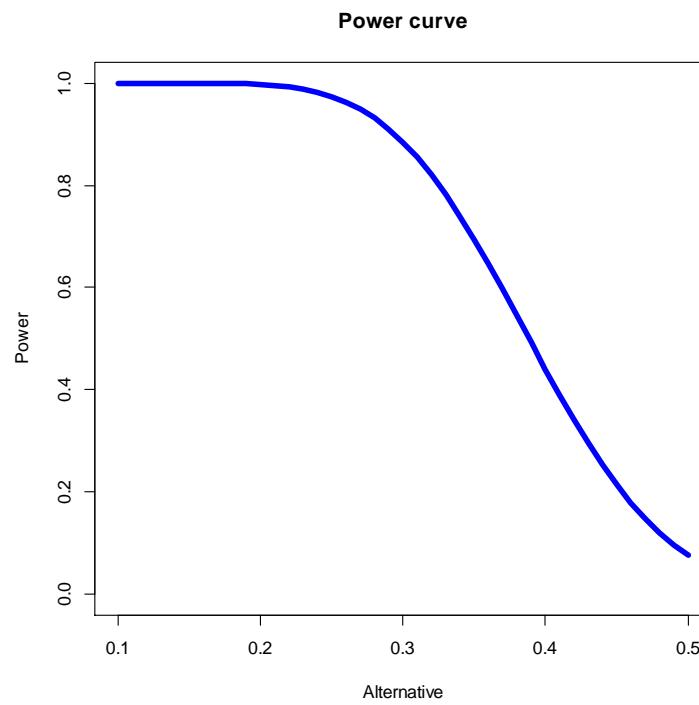
$$P(X \leq 15; \theta = 0.3) = 0.88$$

- What if $\theta = 0.45$? $X \sim \text{Binomial}(40, 0.45)$, hence

$$P(X \leq 15; \theta = 0.45) = 0.21$$

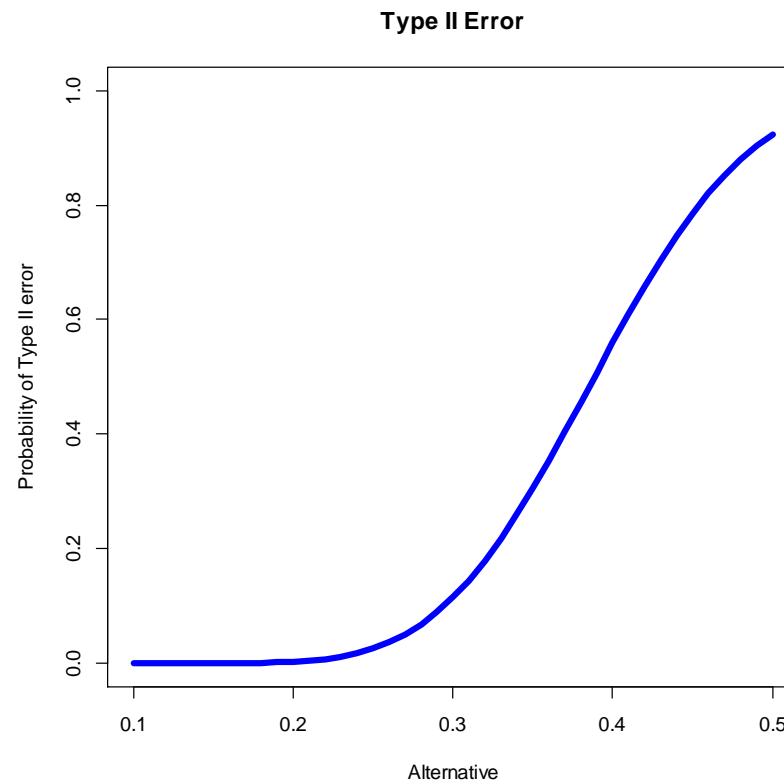
Recap

- We can calculate this to every θ in H_1 ($\theta < 0.5$), which gives us the **power function** for our test “reject if $X \leq 15$ ”:



Recap

- The probability of Type II error at each alternative θ is one minus the power:



Recap

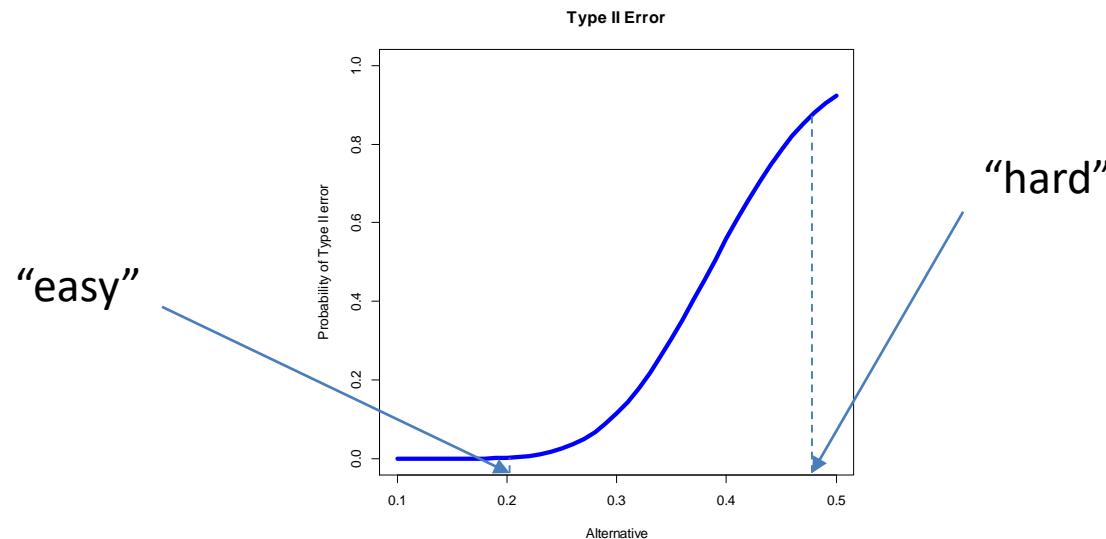
- We got the understanding of another major property of our decision rule.



How Type II error varies

Recap

- Can we control Type II? The power curve follows directly from our choice of “0.05”, H_1 , and test statistic X .
 - The curve just tells you how easily each alternative can be distinguished from H_0 .

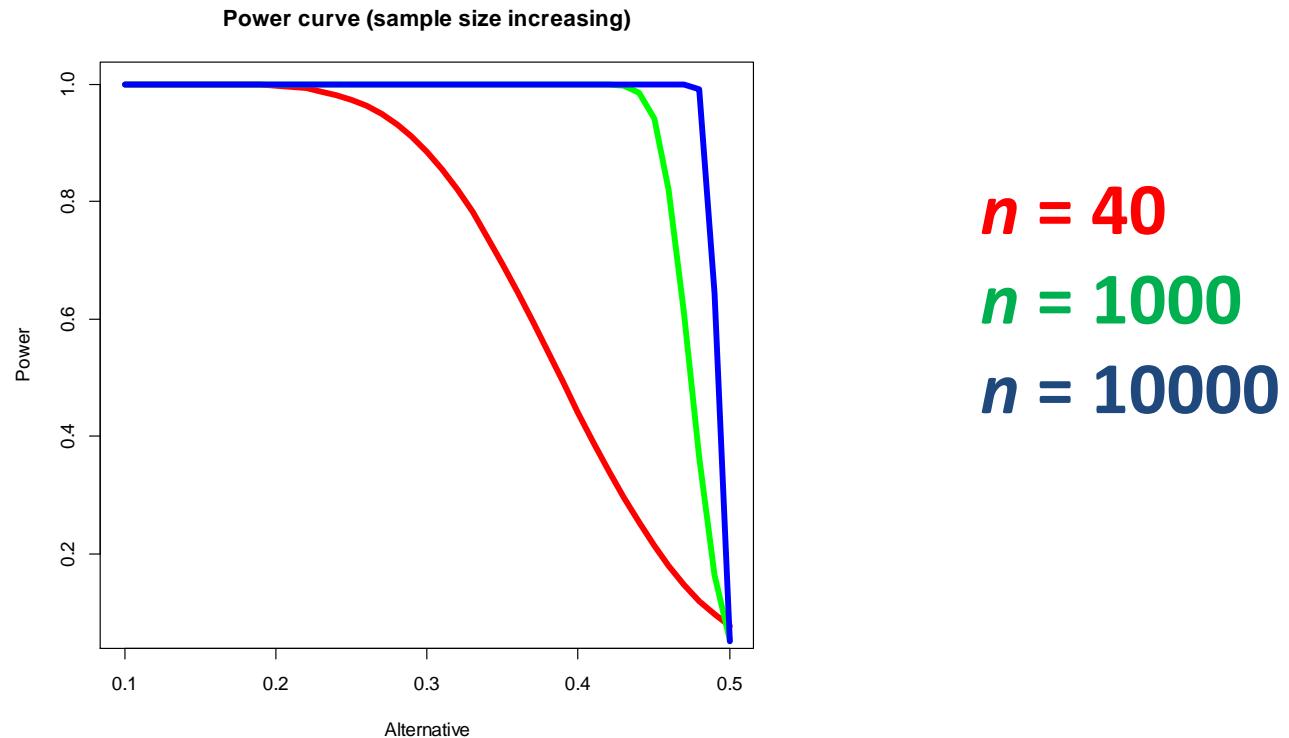


Recap

- How to increase power?
 - Get more data!
 - Allow for a higher Type I error
 - Look for a better test statistic
 - Make stronger assumptions

Recap

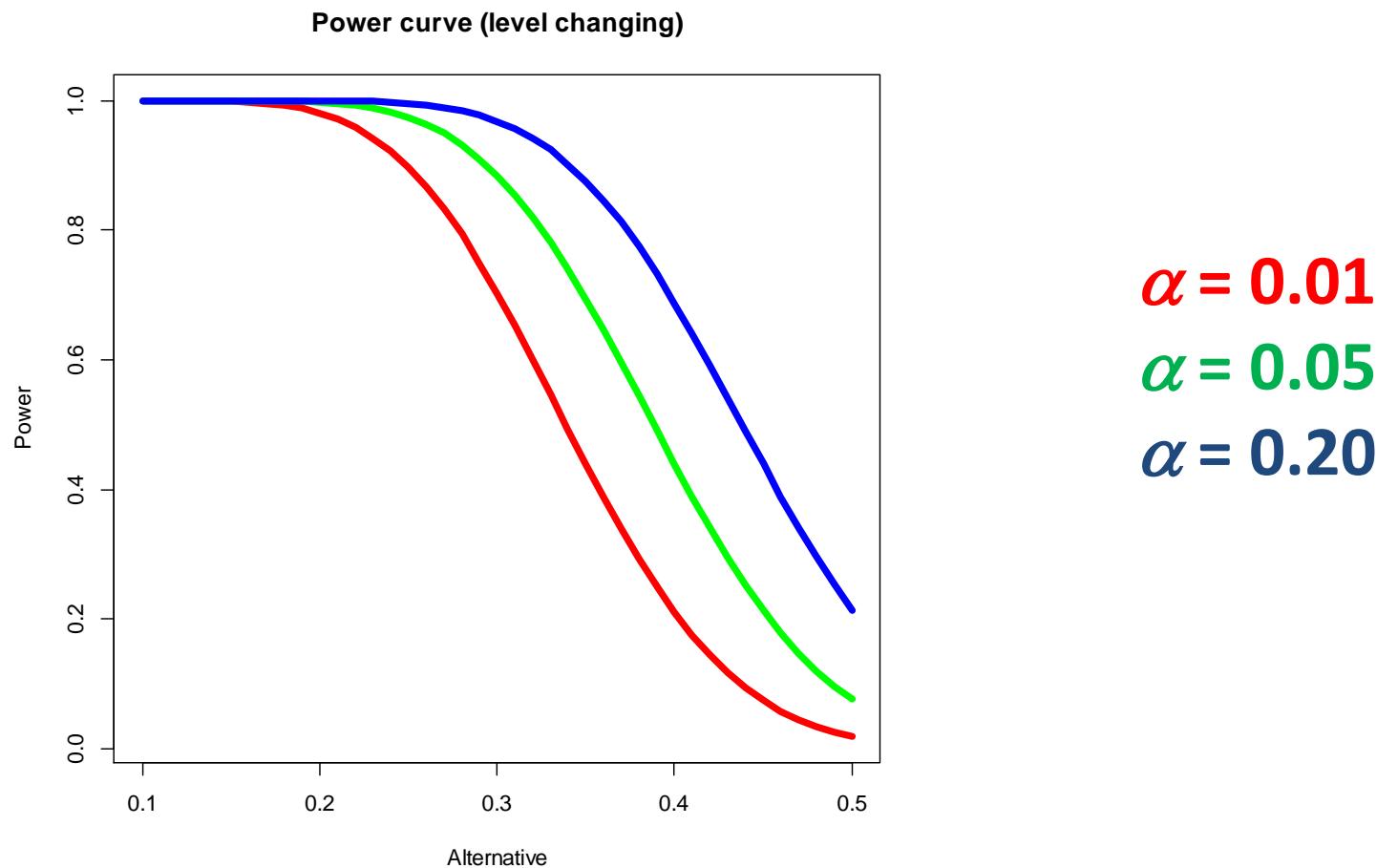
- “Get more data”: power curves allow you to design your study
 - That is, “how many samples should I collect”?



Recap

- Allow for a higher Type I error.
 - If you decrease/increase the critical region, you make it harder/easier to distinguish H_0 from alternatives.
- Level $\alpha = 0.01$: critical region is $\cong [0, 13]$
- Level $\alpha = 0.05$: critical region is $\cong [0, 15]$
- Level $\alpha = 0.20$: critical region is $\cong [0, 17]$

Recap



Recap

- To see how the choice of statistic matters, suppose we use instead X_{dumb} , where

$$X_{dumb} \equiv \sum_{i=1}^{20} Y^{(i)} \quad X_{dumb} \sim Binomial(20, 0.5) \text{ under } H_0$$

- What do you expect will happen?
- Looking for a better statistic to increase power may not be that easy.
 - There are such things as “uniformly most powerful tests”, we will not discuss that any further.
 - The example tests we will discuss are known to have good power.

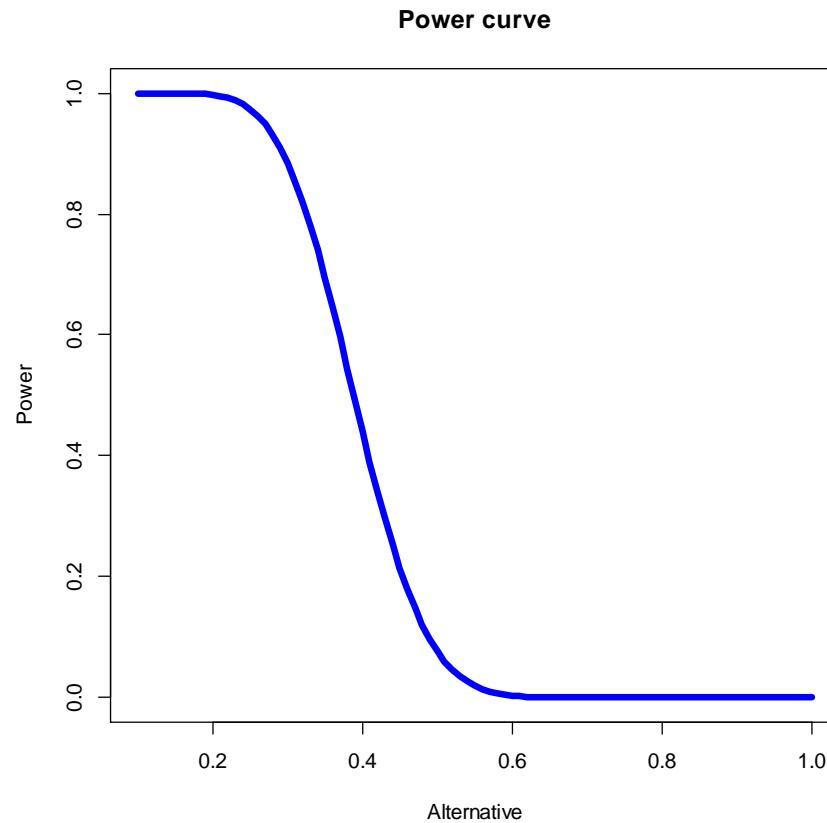
Recap

- Making strong assumptions: reducing your set of alternatives also increases power.
- To see it in the other way (expanding H_1 , observing loss of power): suppose now

$$\begin{aligned}H_0 : \quad \theta &= 0.5 \\H_1 : \quad \theta &\neq 0.5\end{aligned}$$

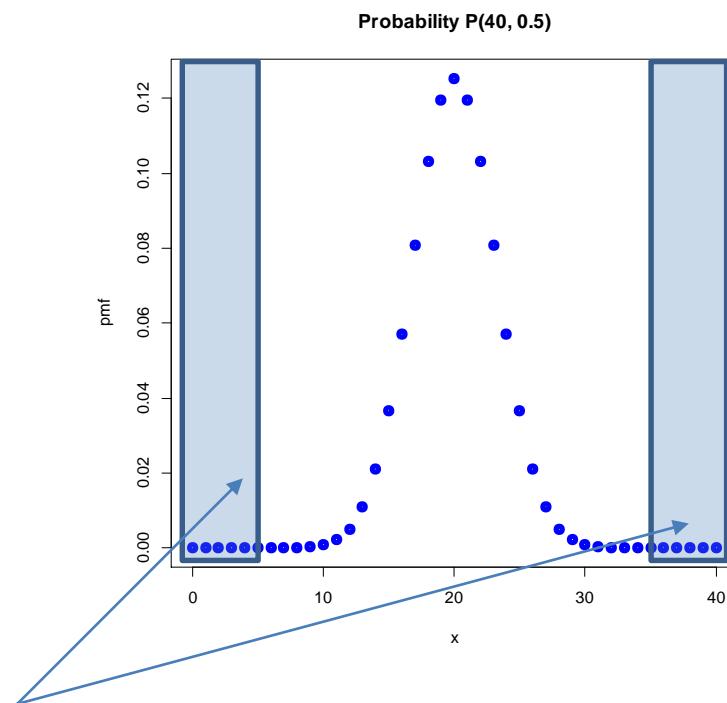
Recap

- Doesn't look good, does it?



Recap

- Rethinking the test statistic:



This will help us to distinguish H_0 from alternatives on “both sides” of H_0 .

Recap

- New critical region: the union of these two sets

$$\{X \leq c_{0.025}\} \cup \{X \geq c_{0.975}\}$$

where $c_{0.025}$ is the 0.025 quantile of the distribution of X under H_0 etc. In our example:

$$\{X \leq 14\} \cup \{X \geq 26\}$$

Recap

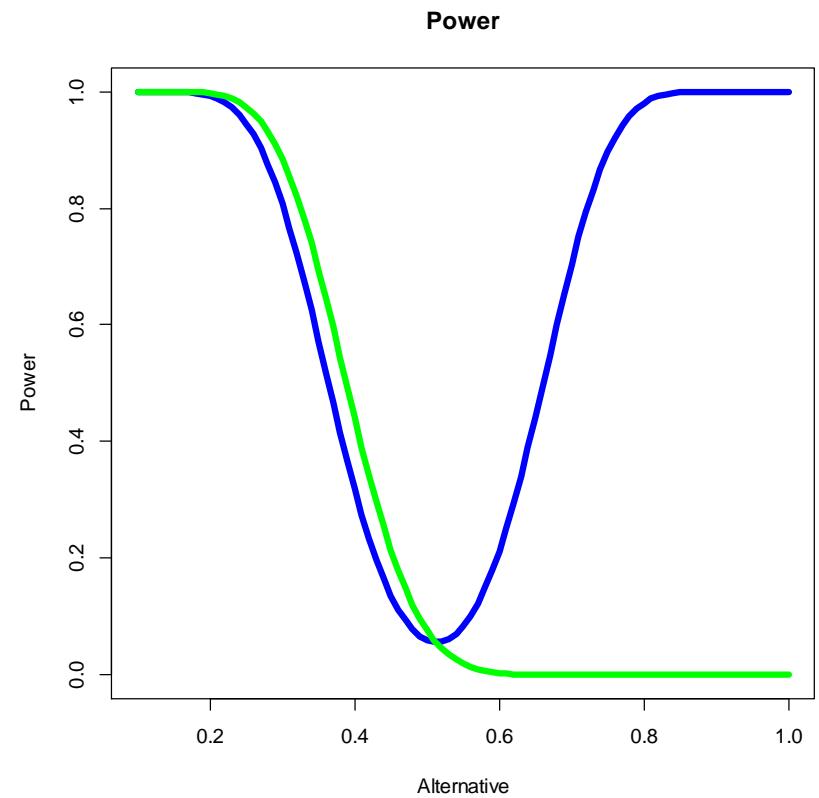
- This is called a **two-tailed test**, as opposed to the “one-tailed test” we saw before.

Recap

- This is now what the power curve is.
- Notice that the original test *is* more powerful for $\theta < 0.5$!

Rule: reject if $X \leq 15$

Rule: reject if $X \leq 14$ OR $X \geq 26$



Composite Hypotheses

- In principle, a null hypothesis can postulate more than one value for the target state of nature, as in this problem setup:

$$\begin{aligned} H_0 : \quad \theta &\geq 0.5 \\ H_1 : \quad \theta &< 0.5 \end{aligned}$$

- This is called a **composite hypothesis** (as opposed to a simple hypothesis).

Composite Hypotheses

- We will not have much to say about this.
- In our context, it suffices to say that we can look at the “hardest” value to falsify ($\theta = 0.5$) and proceed with that as a simple hypothesis.
- For other states of nature in H_0 (say $\theta = 0.6$), our Type I error will be of a smaller size (e.g., less than 0.5) than the designed level (e.g., 0.05). In the “worst-case scenario” ($\theta = 0.5$), we know we are still controlling Type I at 0.05.

Keep This in Mind!

- A large p-value might mean two different things:
 - H_0 is true
 - H_0 is false, but the power of the test is low.

Strategy

- For a given level, pick the test that maximizes power regardless of the true hypothesis.
- Easier said than done: only in some cases there are **uniformly most powerful tests** (that is, at least as good as anything else for any value of the true hypothesis).
- In what follows, I will avoid mathematical discussions and focus on some common tests and their applications.

A Historical Note

- This framework of controlling Type I and minimising Type II error was introduced by Neyman and Pearson in the early 20th century, which became known as the Neyman-Pearson framework.
- As a matter of fact, both Neyman and Pearson at some point worked here at UCL.

Hypothesis Testing

SOME USEFUL TESTS

Warning

- This will sound like an intense laundry list. The most important thing here is getting the logic. The rest comes with practice.
- There will be opportunities to get further details on these within *STATG002* and *STATG003*. See also later chapters of the *STAT1005* notes.

The t-test

- Original motivation: yields of barley. Let's illustrate it with a problem of **quality control** for the Guinness stout.
- Say you are measuring barley concentration in small beer samples. Your sample is assumed to follow some unknown i.i.d. Gaussian:

$$X^{(i)} \sim N(\mu, \sigma^2), i = 1, 2, \dots, n$$



The t-test

- We would like to know if we are correctly manufacturing it with target mean μ_0 . We can formulate it as a hypothesis test

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

- Notice that in most cases the **alternative hypothesis** is just the negation of H_0 .

The t-test

- Gosset (a.k.a. “Student”) derived the distribution of the following statistic

$$T = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{S_n} \sim \mathcal{T}(n - 1)$$

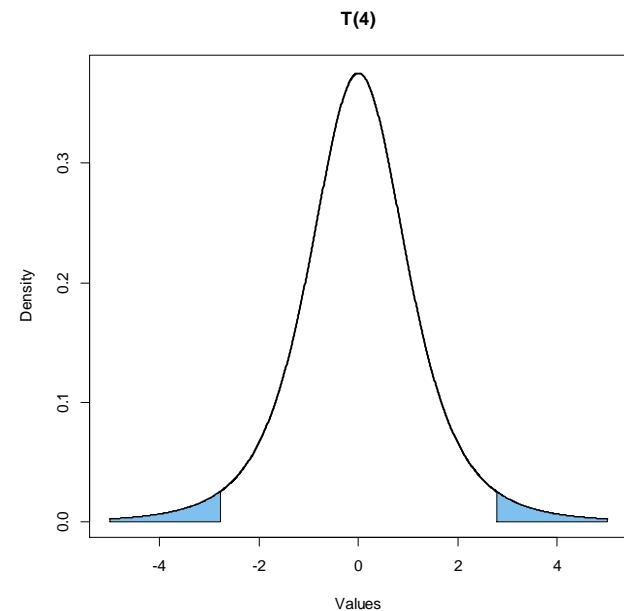
\bar{X}_n : sample mean

S_n : sample standard deviation

- This is called a t-distribution with $n - 1$ degrees of freedom. The formula is ugly, but for large n it is essentially a Normal (0, 1).

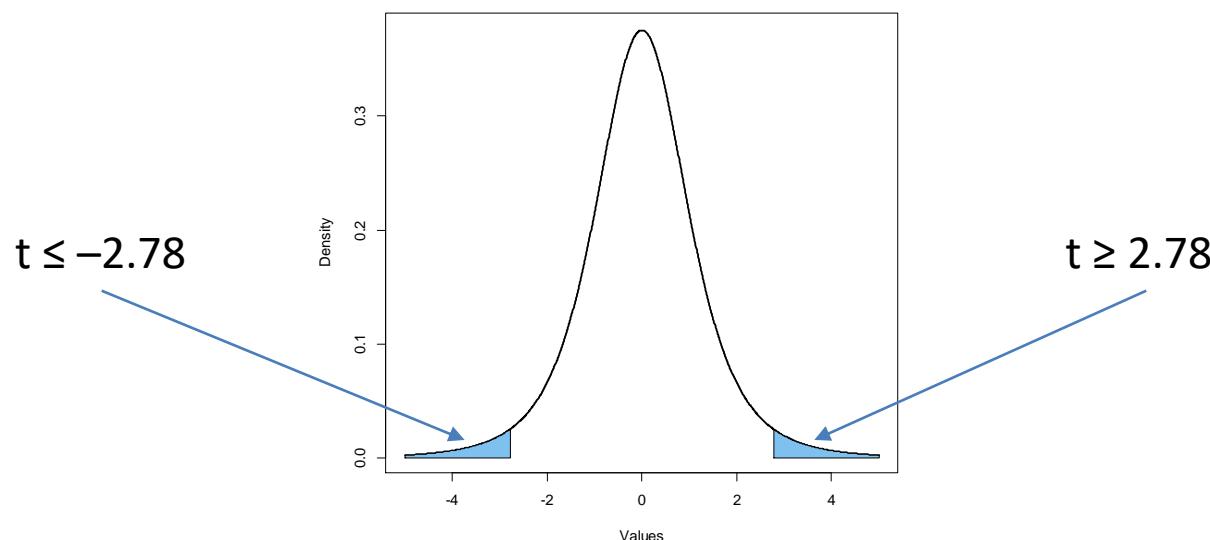
The t-test

- Now, we can look at the probability of “extreme values”.
- The true mean can differ from μ_0 by being smaller or larger. A two-tailed test is used in this example.
- The blue in the figure is the critical region.



The t-test

- We reject H_0 with level α only if t is smaller than the respective $\alpha / 2$ quantile, or larger than the $1 - \alpha / 2$ quantile.
- For instance, if $n = 5$, $\alpha = 0.05$, we have



The Wald Test

- Recall our friend, the Central Limit Theorem.
 - In a simplified way: averages of big samples look like Gaussian random variables.
- The t-test is motivated by small, Gaussian samples.
- The Wald test uses the same statistic (!) as the t-test. The interpretation is different.
 - Samples $X^{(i)}$ can be of “any” distribution.
 - Sample sizes are assumed to be “big enough” to that the CLT kicks in.
 - Hence the distribution of the statistic (let’s call it W , but it’s the same formula as T) is $N(0, 1)$ now.

Goodness-of-fit Tests

- We can think of testing more general assumptions. The t-test goes for a particular mean. What about other **constraints** in the distribution?

Goodness-of-fit Tests

- Like Michelangelo's “statue trapped in the stone”, your model is not what you add, but what you remove. *Hypothesis testing doesn't answer whether what you added to the model was valid, but whether what you subtracted didn't hurt you.*
- In modelling terms: subtraction = constraints.



Wikimedia Commons

Goodness-of-fit Tests

- Example: testing whether two discrete variables are independent.
- In probability terms, this means

$$P(X, Y) = P(X)P(Y)$$

– notice the implication: $P(Y | X) = P(Y)$

Goodness-of-fit Tests

- **Contingency table** of twin data
 - D_j = depression, sibling j
 - A_j = dependence on alcohol, sibling j

		$D_1 = 0$		$D_1 = 1$	
		$D_2 = 0$	$D_2 = 1$	$D_2 = 0$	$D_2 = 1$
$A_1 = 0$	$A_2 = 0$	288	80	92	51
	$A_2 = 1$	15	9	7	10
$A_1 = 1$	$A_2 = 0$	8	4	8	9
	$A_2 = 1$	3	2	4	7

- Is there an association between depression and alcohol dependency across different subjects?

Goodness-of-fit Tests

- The **chi-squared test** compares “expected” versus “observed” outcomes.
- In a nutshell: in this case, for every combination of values of the two variables, compare its frequency of co-occurrences against the product of its marginal frequencies. We can derive a test statistic using a particular way of aggregating these numbers.
- This statistic, Pearson’s χ^2 , has a so-called chi-squared distribution. In general, this test can be used to check whether a particular pmf explains some observed **multinomial data**.

Paired Tests

- As a final example, consider comparing measurements that are tied to a single unit (a person, a beer vat, and so on).
- This is typical when we apply two treatments to a same individual and contrast the results.
 - Further details? Yes, *STATG002*.



BEFORE



AFTER

Paired Tests

- A **s signed rank test** (Wilcoxon test) compares the differences of two $X^{(i)}$ and $Y^{(i)}$ measurements within a single individual i .
- The idea is to build data as if it came from the null. For that, build differences $X^{(i)} - Y^{(i)}$ for all possible pairs, look at the sign. Under the null, it is possible to find the distribution of a test statistic based on these rearrangements.
- The null here is whether $P(X^{(i)} > Y^{(i)}) = P(Y^{(i)} > X^{(i)})$.
 - Question: What would you do if we assume they are Gaussian distributed with the same variance?

Hypothesis testing

WORDS OF CAUTION AND PRAGMATIC ADVICE

Before We Conclude

- Let's remind ourselves why we are doing this!

Objections to Hypothesis Testing

- The null is “always false”, especially depending on the amount of precision used.
- Dichotomization of decisions may lead to inconsistencies. We can “accept” some null H_0^i , and “reject” some H_0^j , even if $H_0^i \Rightarrow H_0^j$!
- It is confusing and people often misuse it.
 - Well, don’t you disappoint me!

Why Do Hypothesis Testing

- A typical industry practice: A/B testing
 - Do two treatments. Say, offer product with two different variations (price, colour, user interface, etc.)
 - Does the distribution of outcomes (sales, consumer satisfaction, etc.) change in some way (mean, variance, maximum value, etc.)?
 - Set H_0 as the “no change” hypothesis.

Kohavi et al. (2009) "Controlled experiments on the web: survey and practical guide"

<http://dl.acm.org/citation.cfm?id=1485091>

Why Do Hypothesis Testing

- Granted, you may be/should be interested on the *size* of the effect (e.g. difference in sales means).
- However, would you have a large enough sample to distinguish it from zero?
- The machinery of hypothesis testing helps to tell you whether you are asking a ridiculous question to begin with.
 - that is, estimating effect size when the sample size you have can't really distinguish it from zero

Why Do Hypothesis Testing

- Models rely on assumptions that sometimes are “good enough”. For example, Gaussianity and independence.
- There is a more convincing story, instead of meaningless handwaving, if you actually show that your data cannot falsify these assumptions!

Why Do Hypothesis Testing

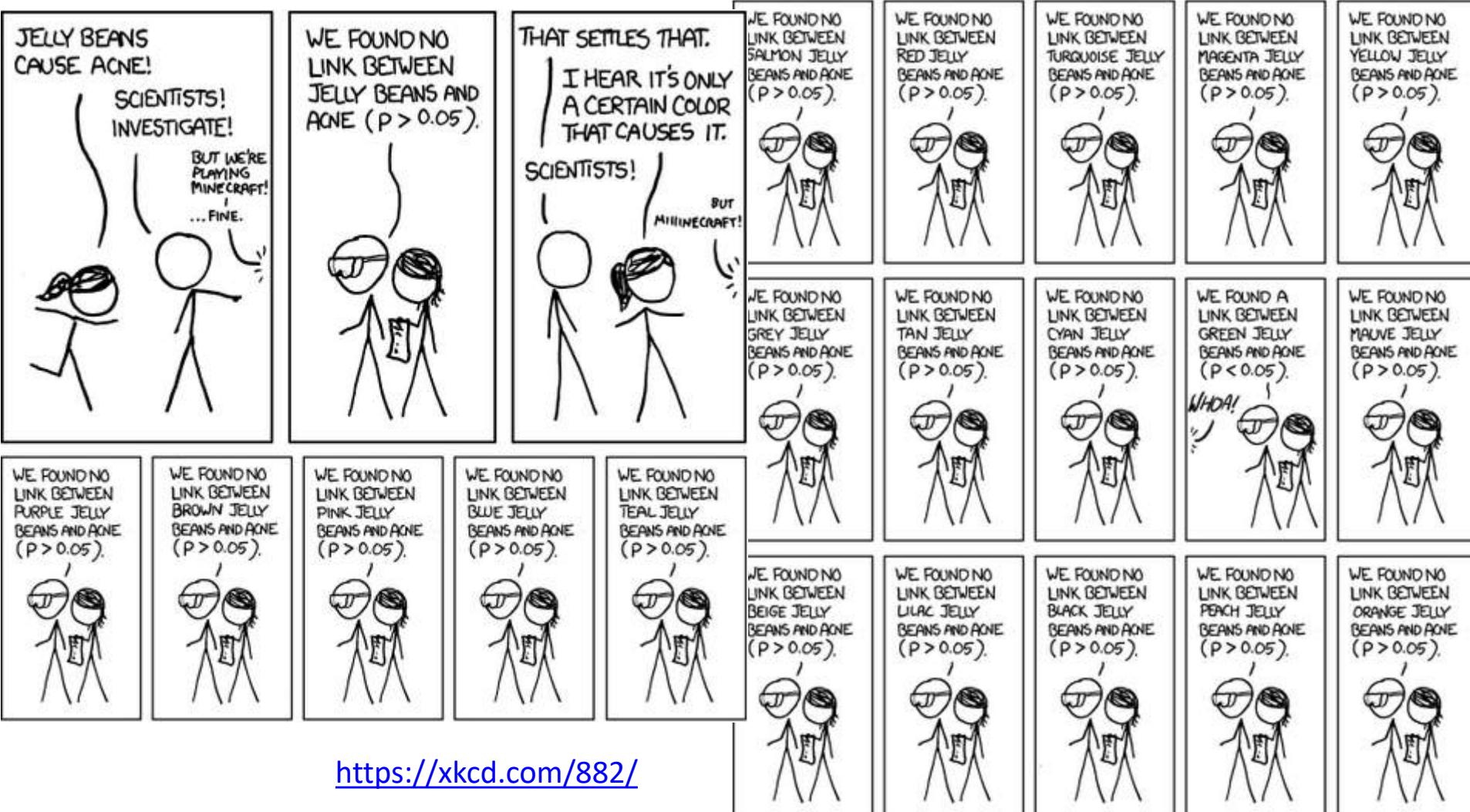
- Physical/social/psychological measurements have practical limits.
- A null hypothesis can be highly precise, and yet not falsifiable with the given technology.
- Moreover, background knowledge might tell us that the precision of the null is good enough. See also: Boson, Higgs.

Why Not Do It

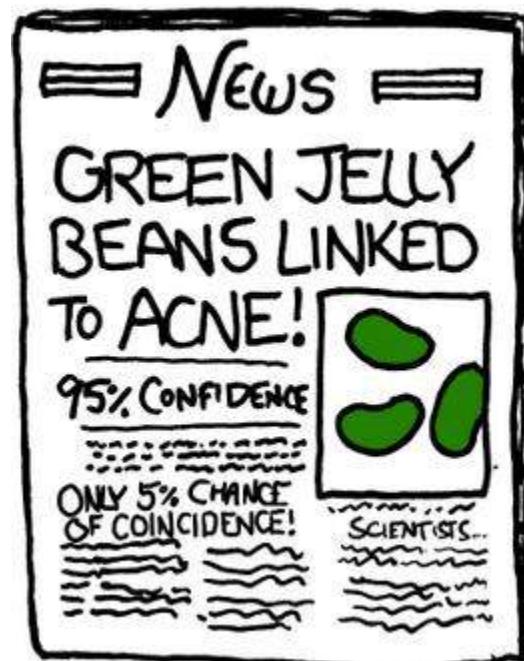
- Essentially, if you think you can get away without doing anything else.
- Hypothesis testing is hardly (or should be) convincing by itself. Effect size matters.
Validation of assumptions/sample size may be just the starting point of a solid analysis.

Statistical significance ≠ practical significance

The Sociology of Hypothesis Testing and p-hacking



The Sociology of Hypothesis Testing and p-hacking



The Sociology of Hypothesis Testing and p-hacking

- There are perverse incentives for “p-hacking”: making a selective reporting of p-values.
 - Multiple tests on a single data set are not independent. **And the minimum of a set is not going to be uniformly distributed.**
- Be responsible.

The Sociology of Hypothesis Testing and p-hacking

- Two different types of bad incentives:
 - “the null is bad”. If I’m proposing a new treatment and the null is a zero difference with respect to the old treatment. Down with the p-value!
 - “the null is good.” If I’m proposing a model and the null is “the model generated the data”. Up with the p-value!

Multiple Testing

- There is a considerable literature on multiple testing, which I will not cover.
- I will just mention one of the simplest techniques, the **Bonferroni correction**. It is motivated by $P(A \cup B) \leq P(A) + P(B)$.
- So if you have k hypotheses to test, think of As and Bs as the Type I error events.

Multiple Testing

- Without knowing the (complicated) joint distribution of these events, it suffices to control the level of the joint test by changing the level α of each individual test to α / k .
- It will control the level, but the probability of Type I error maybe much smaller than α . Bad power is likely to follow...

Take-Home Message

- Hypothesis testing: putting assumptions to test by deriving their consequences to the observed data.
- Despite its shortcomings, it is a common type of diagnostics. As long as we do not take them as the ultimate goal of an analysis (only in rare cases), they can be valuable.

Confidence Intervals

GENERAL CONCEPTS

Recall the NHANES Data

- Height data, where we found the estimated height expectation of 168 cm:

$$\hat{\mu} = 168$$

- We asked: what if **different** 19,219 individuals had been sampled?

Using Simulation to Understand Confidence Intervals

- Imagine the following experiment: let's generate **simulated data**, "God playing dice", like we have been doing in our R examples.
- This is called a **(Monte Carlo) simulation**, for which there are very old computer algorithms based on **pseudo-randomness**.

Using Simulation to Understand Confidence Intervals

- Let's say we generate datasets of size 50, based on a $N(\mu = 168, \sigma^2 = 103)$ distribution.
- Let's do it over and over again, say 20 times, calculate the sample average each time.
- R demo

Using Simulation to Understand Confidence Intervals

- The sample average

$$\bar{X} \equiv \frac{1}{n} \sum_{i=1}^n X_i$$

is of course a random variable, since it is a function of the data.

- Importantly, it is a function only of the data, not of the unknown parameters of the model.
 - Recall the name: statistic!
- But its *distribution* should be a function the data distribution.

Recall (from Chapter 1)

- If

$$X \sim N(\mu, \sigma^2)$$

then

$$E[X] = \mu$$

- (You might want to check that yourself. Remember your calculus)

$$E[X] = \int x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right\} dx = \mu$$

In Particular

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \mu$$

- (check this, if you are interested. It follows from the definition of $E[]$)
- That's why \bar{X} is a plausible estimator of μ . We typically denote estimators with hats, so in our case we chose $\hat{\mu} = \bar{X}$.

More Than This

- We can characterize the whole distribution of the sample average if each X_i is $N(\mu, \sigma^2)$:

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

- We can now use this to our advantage. Ask yourself: what is the distribution of

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} ?$$

(Use the fact that the $\text{Var}(Y / c) = \text{Var}(Y) / c^2$ for any random variable Y and constant c, and linear manipulations of a Gaussian are also Gaussian-distributed)

Bounding μ

- It is a $N(0, 1)$. So for instance we can make claims such as (the cutoff below is arbitrary),

$$P\left(\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leq 1.5\right) = P(Z \leq 1.5) \approx 0.93$$

where $Z \sim N(0, 1)$.

- So we can make the analogous claim

$$P\left(\mu \geq \bar{X} - 1.5\sigma/\sqrt{n}\right) \approx 0.93$$

Two IMPORTANT Observations

- In general, $\bar{X} - 1.5\sigma/\sqrt{n}$ is not a statistic!!
- Assume for now σ^2 is known, to simplify the argument (so the above IS a statistic in this case).
- Also: how to interpret this statement?

$$P(\mu \geq \bar{X} - 1.5\sigma/\sqrt{n}) \approx 0.93$$

The Key Point Concerning Confidence Intervals

- The randomness is not on μ !
- The randomness is in the data, which here is summarized by \bar{X} !
- What on Earth does it mean? After all, “I got my data already, it’s right here in front of me”.



Coverage

- Regardless of what μ is, if my sample size is n and my data follows a $N(\mu, \sigma^2)$, then in the limit of infinite repetitions of my dataset, the interval

$$[\bar{X} - 1.5\sigma/\sqrt{n}, +\infty)$$

will contain μ (approximately) 93% of the time.

- Another way of saying this: the **coverage** of this interval is 93%.

Lessons

- As I've mentioned before, nobody expects you to collect "infinitely many datasets" for a same problem.
- What this means is the following: if you provide a (say) 93% confidence interval for your quantity of interest in each problem you work on through your career, then in the long run the intervals you provided will contain the quantity of interest 93% of the time. You will earn the title "Mr/Ms/Mrs 93%".
- **You cannot know (without further data) for which intervals you got it right, just the long-run performance!**

In Practice

- There will be several assumptions in your model that will be violated, so coverage will not be exact.
- Despite being an idealization, reporting confidence intervals is of major importance in many applications.
 - **At the very least as a way of being more humble about which conclusions you can draw.**
- Just because some modern models have difficult-to-interpret parameters (e.g., neural nets), it doesn't mean confidence intervals will not be used at some point in your application (e.g., in the estimation the empirical performance of a neural net).

In Practice

- When we get to other topics (such as regression), we will talk again about confidence intervals in those contexts.

Another Interval

- It is much more common to report lower and upper bounds.
- Going back to our example. Say you want a 95% interval (the default in much software). This is typically done by finding the 2.5% and 97.5% quantiles of the distribution of your statistic. For instance,

$$[\bar{X} + z_{0.025}\sigma/\sqrt{n}, \bar{X} + z_{0.975}\sigma/\sqrt{n}]$$

2.5% quantile of $N(0, 1)$

97.5% quantile of $N(0, 1)$

A “Real” Statistic for the Gaussian Mean CI

- I owe you that. Start from:

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim \mathcal{T}(n - 1)$$

where S^2 is the sample variance:

$$S^2 \equiv \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Exercise: can you derive a 95% confidence interval using this?

A “Real” Statistic for the Gaussian Mean CI

- Notice: for large n , which we are assuming anyway, the following is used in practice
 - (the funny double-tilde is just an informal notation for “approximately distributed”)

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \approx N(0, 1)$$

Confidence Intervals

APPROXIMATIONS: CLT AND THE BOOTSTRAP

Those Bounds Again

- Say you want to trap a parameter of interest θ with coverage probability c . This is the general template of a confidence interval:

$$P(lower_c(\mathbf{X}) \leq \theta \leq upper_c(\mathbf{X})) = c$$

so depending on your choice of c , you will get (random) lower and upper bounds that depend on data \mathbf{X} .

- In general, it is not at all easy to find the actual distribution of $lower_c(\mathbf{X})$ and $upper_c(\mathbf{X})$!
 - It would be bonkers to assume Gaussianity in general. However, the good old Gaussian is useful in a different way.

Help me, Central Limit Theorem!

- [You may have seen this coming.]
- Many lower/upper functions depend on averages.
- We have seen what happens to averages for large n , correct?

(R demo)

Practical Advice

- Many confidence intervals in software packages rely on the Central Limit Theorem under the hood.
- It is not always obvious what a “large sample” is. The theory is about **asymptotics**. Sometimes checks can be done.

A Missing Ingredient

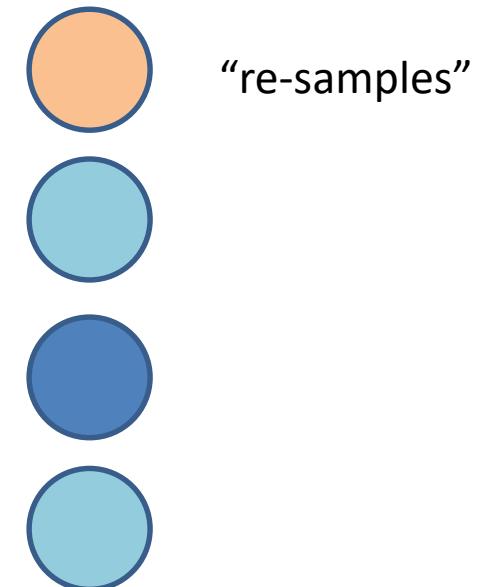
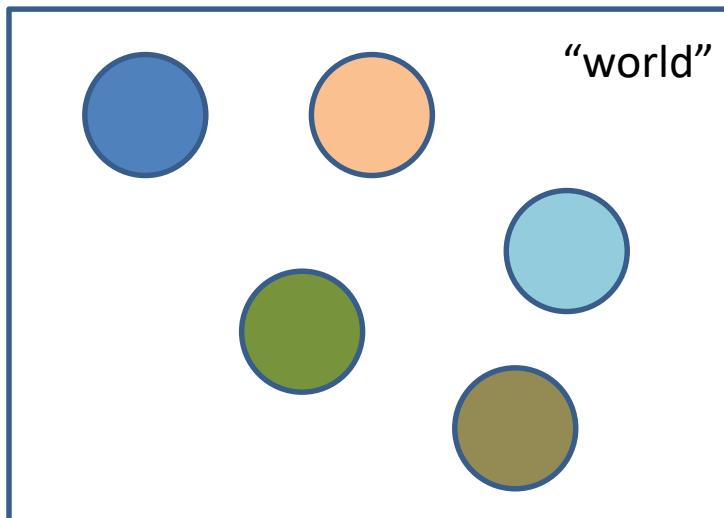
- Normal approximations for averages are fine in many cases. But what is the variance of your statistic??
- Sometimes this can be written down in a simple way. In many cases, not at all.
- When good old fashioned algebra fails, we should resort to computer-intensive alternatives. **Enter the bootstrap.**

The Bootstrap

- “The world” generated your data.
- **What if *your data was the world*? What does it mean by “your data generating data”?**

The Basic Idea

- Think of your sample as if it was the “world”, the **population**.
- A box which we can, now, *play with by sampling from it*.

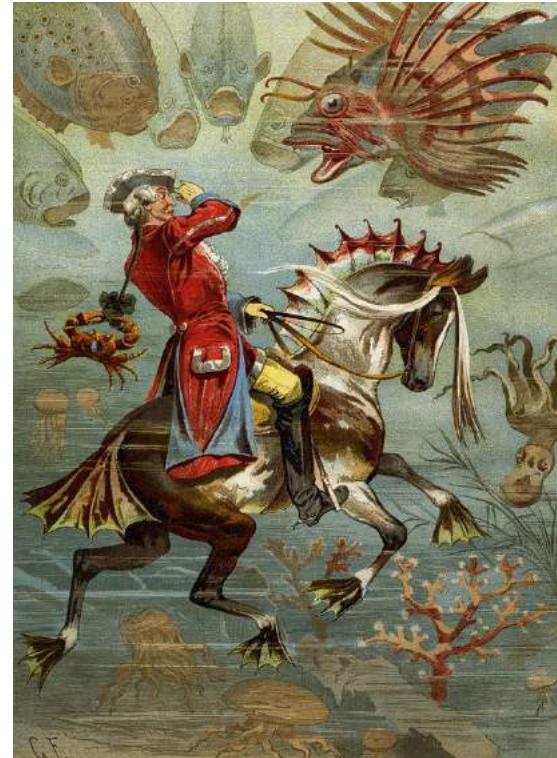


Sampling with Replacement, and the Size of the Sample

- This is **sampling with replacement**: choose a data point, add to the “re-sample”, but “put it back in the box”.
- We do this to generate a re-sample of the same size as the original sample. The idea is to mimic the same process that generated your data, down to the sample size.

The Bootstrap

- So, we are “using the data we have to generate data.” This will inform us about sampling variability.



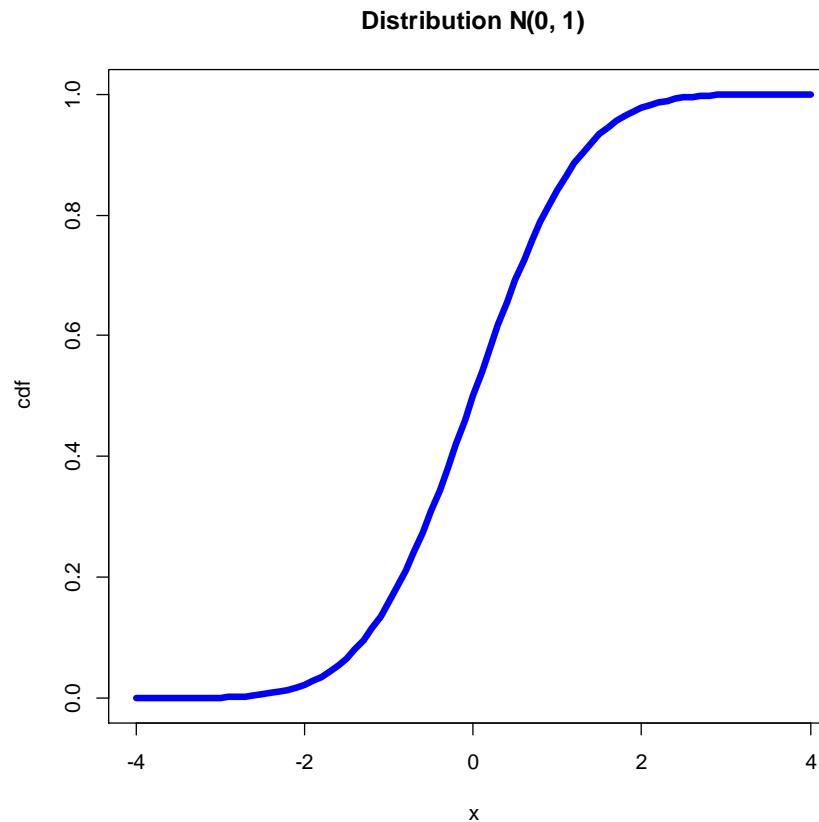
Baron Munchausen (Wikimedia Commons)

Why?

- Because we can now do it many many times to simulate in a computer the idea of “infinite replicates of a dataset”.
 - Ideally, we would do infinitely many replicates. In practice, we choose a large number and accept that we will get some approximation error.
- We can then see how our statistic varies across these synthetic replicates.

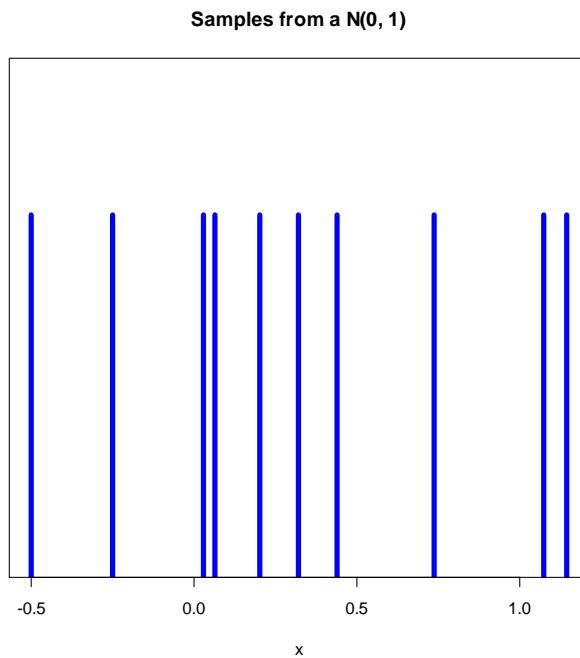
An Intuition of Why This Works

- Recall our friend, the cumulative distribution function $F(x) \equiv P(X \leq x)$. For instance,



The Empirical CDF

- Thought experiment: the data as our population. What would be the cdf when new data should be only at particular locations, with equal probability?



Each stick is placed at a particular draw from a $N(0, 1)$.

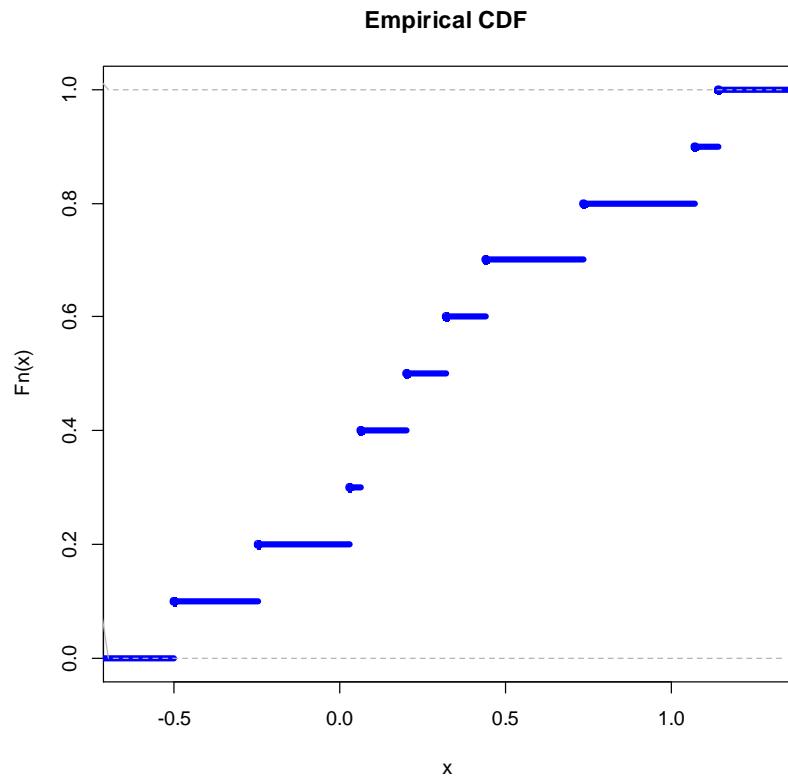
Now imagine you are generating x values only at the locations previously chosen by our draws.

The Empirical CDF

- For any particular level x , just count the frequency of points no larger than x .

$$\hat{F}_n(x) \equiv \frac{\#\text{data points no greater than } x}{\#\text{data points (i.e., } n)}$$

$$\hat{F}_n(x) \equiv \frac{1}{n} \sum_{i=1}^n I(x^{(i)} \leq x)$$



In the Limit

- Analogous to the **law of large numbers** (averages converge to means as sample size increases), empirical CDFs converge to the population cdfs (R demo).
- So, informally, we can say that the empirical distribution (what we get by re-sampling) carry some information about sampling according to the true cdf.

Using the Bootstrap

- The simplest way is to use it to calculate the variance of the statistic of interest along with the CLT approximation.
- Going back to our UK Gas consumption, let's create a 95% confidence interval for the mean.

standard error (“deviation”) obtained by bootstrap

$$[\bar{X} + z_{0.025} \hat{s}e_{boot}, \bar{X} + z_{0.975} \hat{s}e_{boot}]$$

The Algorithm

1. Draw $X^{(1)\star}, \dots, X^{(n)\star} \sim \hat{F}_n$
2. Compute \bar{X}_n^\star by averaging $X_1^\star, \dots, X_n^\star$
3. Repeat steps 1 and 2, B times, to get $\bar{X}_{n,1}^\star, \dots, \bar{X}_{n,B}^\star$
4. Let

$$s.e.boot \equiv \sqrt{\frac{1}{B} \sum_{b=1}^B \left(\bar{X}_{n,b}^\star - \frac{1}{B} \sum_{r=1}^B \bar{X}_{n,r}^\star \right)^2}$$

[Recall the empirical variance definition: $\frac{1}{n} \sum_{i=1}^n (X^{(i)} - \bar{X})^2$]

B ?

- Ideally, we would average over *every possible re-sample*, but this is not in general doable.
- So this is an approximation, which itself is justifiable by the law of the large numbers! Some packages might choose B for you.
- R demo

Wait

- The empirical mean is simple enough that we can find a decent approximation for its variance in the literature. There wasn't really a need for the bootstrap here.
- The bootstrap shines when this is not the case. In the UK energy example, we might be interested in a confidence interval for the *median* (can you guess why?).

Rephrasing It

- The idea is exactly the same. Instead of the sample average, our statistic is the sample median. Just think in terms of abstract T s.
 1. Draw $X^{(1)*}, \dots, X^{(n)*} \sim \hat{F}_n$
 2. Compute T_n^* from X_1^*, \dots, X_n^* according to its definition
 3. Repeat steps 1 and 2, B times, to get $T_{n,1}^*, \dots, T_{n,B}^*$
 4. Let

$$s.e.\text{-}boot \equiv \sqrt{\frac{1}{B} \sum_{b=1}^B \left(T_{n,b}^* - \frac{1}{B} \sum_{r=1}^B T_{n,r}^* \right)^2}$$

(R demo)

Rephrasing It

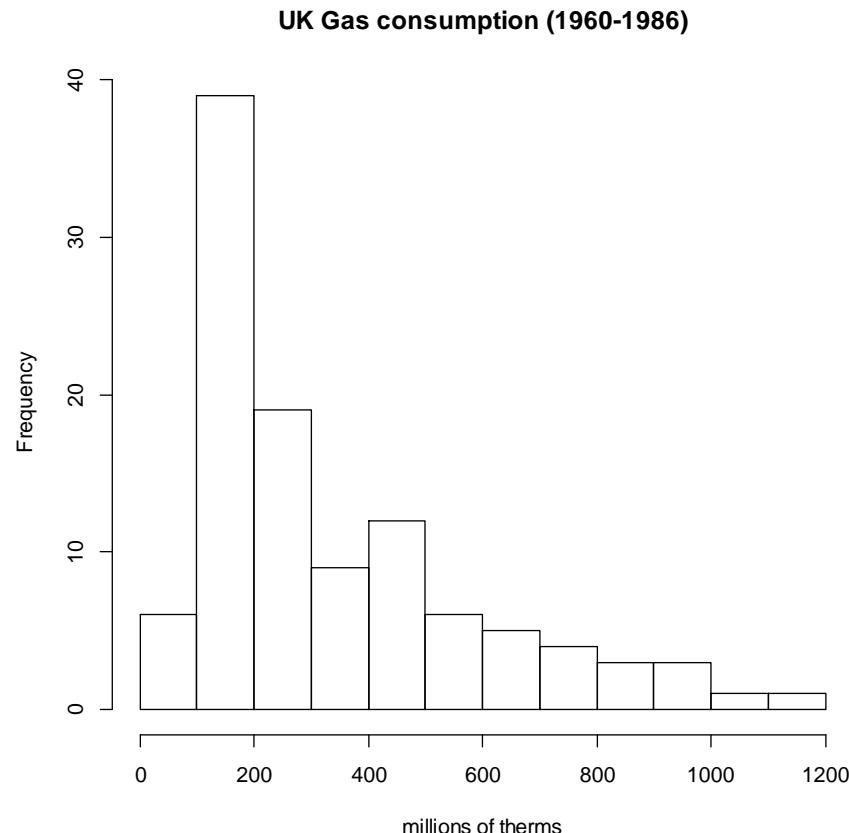
- But this only gives variances. What if we want a more precise way of building confidence intervals without using the Normal approximation?

Confidence Intervals

THE STORY SO FAR

Recap

- From data like this, we would like **to learn a property of the population**, like the mean or median.
- Preferably, **not only a single point**, but a set (an **interval**, more precisely) that will **contain the true value with high probability**.



Recap

- For instance, to learn about the mean μ , we can make use of the following claim

$$\bar{X} \approx N(\mu, S^2/n)$$

where n is the sample size and S^2 is the sample variance $S^2 \equiv \frac{1}{n} \sum_{i=1}^n (X^{(i)} - \bar{X})^2$.

- Why is this useful?

Recap

1. We can rewrite it like this
$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \approx N(0, 1)$$
2. Then find the (say) 0.025 and 0.975 quantiles of this known distribution to claim that
$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \leq 1.96\right) = 0.95,$$
3. And then re-express it in a way to emphasize μ :

$$P\left(\bar{X} - 1.96\sqrt{S^2/n} \leq \mu \leq \bar{X} + 1.96\sqrt{S^2/n}\right) = 0.95,$$

Note

- We actually have a name for quantities like this:

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}}$$

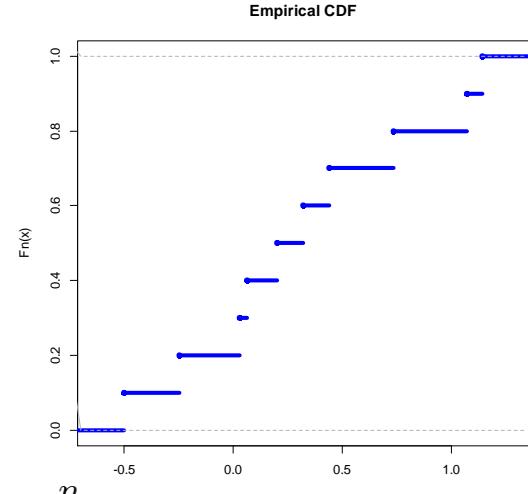
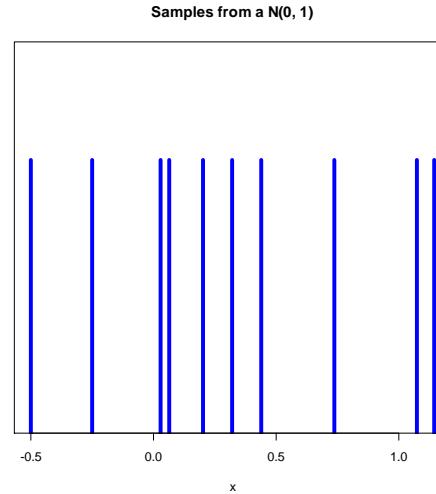
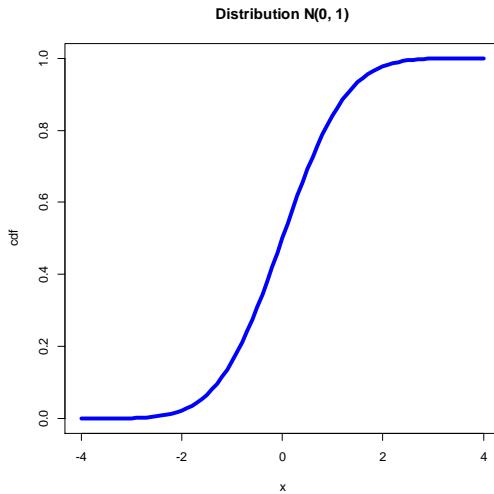
- This is called a **pivot**, a function of the parameter of interest which has a known distribution.
 - Notice a pivot is not a statistic by construction!
 - In general, hard to find!

Recap

- For instance: how to get a confidence interval for the median? Which pivot to use?

Recap

- Enter the bootstrap: a general trick that uses the **empirical distribution**.
 - We saw it to calculate variances. Now let's see how to build pivots with it.

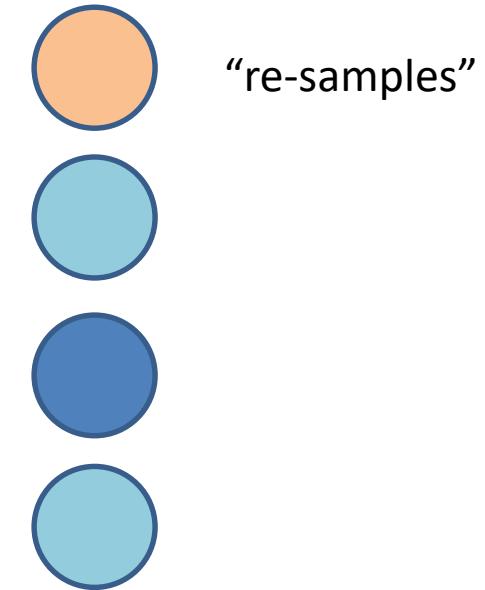
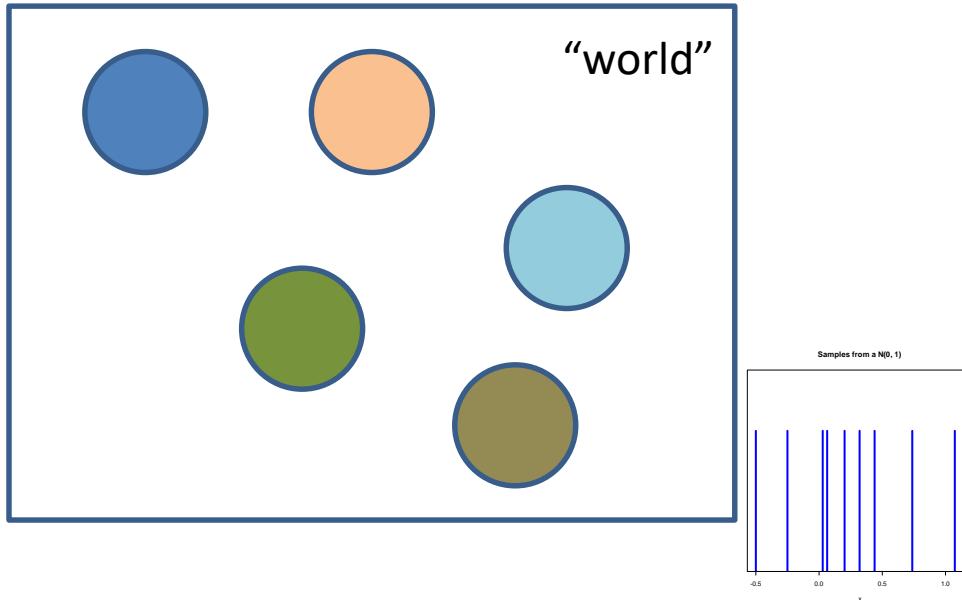


$$\hat{F}_n(x) \equiv \frac{\#\text{data points no greater than } x}{\#\text{data points (i.e., } n)}$$

$$\hat{F}_n(x) \equiv \frac{1}{n} \sum_{i=1}^n I(x^{(i)} \leq x)$$

Recap

- Basic idea: think of your sample as if it was the “world”, the **population**.
- A box which we can, now, *play with by sampling from it*.



The Bootstrap Pivotal Interval

- The Normal approximation may not be good, as in estimating the median.
- One of the most used bootstrap variants of confidence intervals is the **pivotal interval**.
- The problem: find confidence interval for parameter θ based on an estimator $\hat{\theta}_n$ built from a sample of size n .
 - Notice the explicit subscript n .

Alternative Use

- Essentially, use bootstrap to estimate the distribution of $\hat{\theta}_n - \theta$, **which we will be our pivot**. Then find its quantiles of interest.

Idea

- Let $H(r)$ be the cdf of the pivot, that is

$$H(r) \equiv P(\hat{\theta}_n - \theta \leq r)$$

- Define quantiles such that we get coverage $1 - \alpha$:

$$P(a(\hat{\theta}_n) \leq \theta \leq b(\hat{\theta}_n)) = 1 - \alpha$$

- I won't show to you, but the following satisfies the above:

$$a(\hat{\theta}_n) = \hat{\theta}_n - H^{-1}(1 - \alpha/2)$$

$$b(\hat{\theta}_n) = \hat{\theta}_n - H^{-1}(\alpha/2)$$

The Problem

- What are the distributions of these little monstrosities? Hard to find!

$$a(\hat{\theta}_n) = \hat{\theta}_n - H^{-1}(1 - \alpha/2)$$
$$b(\hat{\theta}_n) = \hat{\theta}_n - H^{-1}(\alpha/2)$$

- Solution: bootstrap 'em. First, generate B resampled datasets, with the respective **bootstrapped pivots** $R_{n,b}^*$, replacing $\hat{\theta}_n - \theta$!

This replaces $\hat{\theta}_n$

And this replaces θ !

$$R_{n,b}^* \equiv \theta_{n,b}^* - \hat{\theta}_n$$

Final Trick

- Use the distribution of the bootstrap samples to get an estimated $\hat{H}(r)$.

$$a(\hat{\theta}_n) = \hat{\theta}_n - \hat{H}^{-1}(1 - \alpha/2)$$
$$b(\hat{\theta}_n) = \hat{\theta}_n - \hat{H}^{-1}(\alpha/2)$$

- What is $\hat{H}^{-1}(\alpha)$? It's just the empirical quantile of the bootstrap distribution. Sort all of your B bootstrap pivots, pick the one in position $B\alpha$, or the closest integer.

That Is

- To get e.g. $b(\hat{\theta}_n) = \hat{\theta}_n - \hat{H}^{-1}(\alpha/2)$
 - Find $\hat{H}^{-1}(\alpha/2) = \theta_{n,B\alpha/2}^* - \hat{\theta}_n$
 - Set $b(\hat{\theta}_n) = 2\hat{\theta}_n - \theta_{n,B\alpha/2}^*$
 - Note: $B\alpha/2$ is rounded
- R demo

Take-Home Messages

- Estimation is important, but so is the assessment of your uncertainty.
 - In your career as Data Scientist, you will be pressed for easy answers. Don't fall for that.
- Confidence intervals provide coverage: an interval which traps the parameter of interest, regardless of its true value, with the advertised probability.

Take-Home Messages

- However, everything is predicted on given assumptions. Including the bootstrap. Don't ever forget that.
 - You should verify them as best as you can.
 - Again, **the game is about being “less wrong”, it is not about being infallible.**
- If the intervals look wide and uninformative, even with large sample sizes: tough luck. **Information doesn't come for free.** If you want more certainty, you need more assumptions (or more data).

Take-Home Messages

Hypothesis testing and confidence intervals are two core building blocks of statistical inference.

They will show up as again as needed, when we study regression and modelling during the rest of this course.

Introduction to Statistical Data Science

Ricardo Silva

ricardo@stats.ucl.ac.uk

Department of Statistical Science, UCL

Linear Regression

Outline

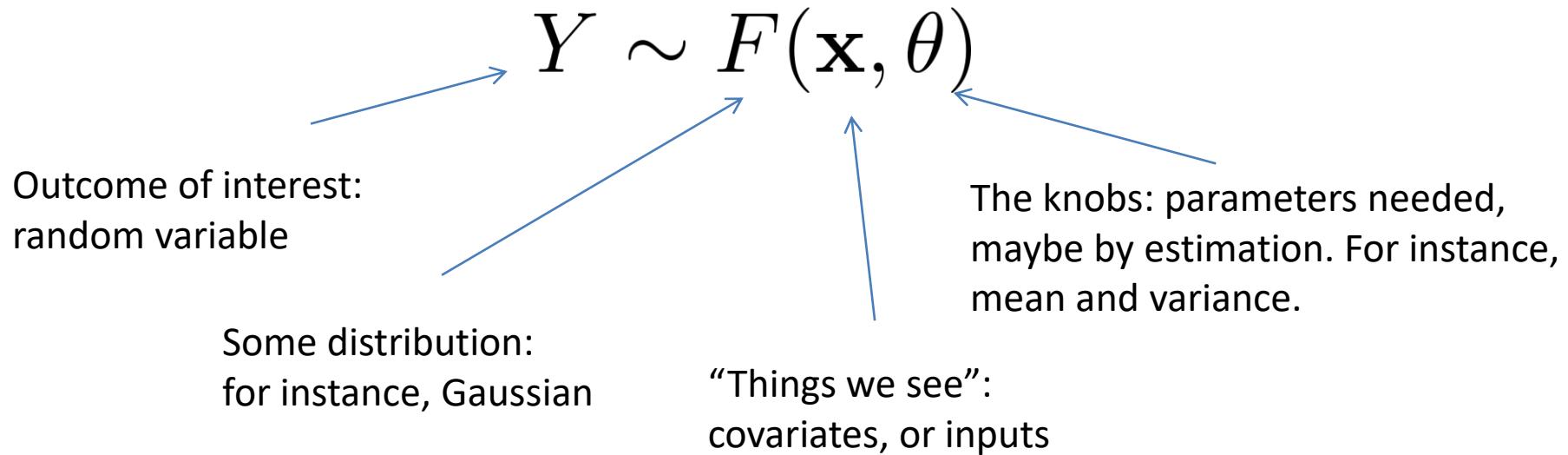
You have studied some of this in the Supervised Learning module. We here present an alternative view that emphasizes interpretation and statistical properties.

Outline

- Basic definitions
- Gaussian vs model-free points of view
- Model checks
- Hypothesis testing and confidence intervals
- Other practical issues

Learning a Relationship

- Our measurements are not independent.
- Often we want to characterize the distribution of an **outcome** Y given observable **covariates** \mathbf{x} :

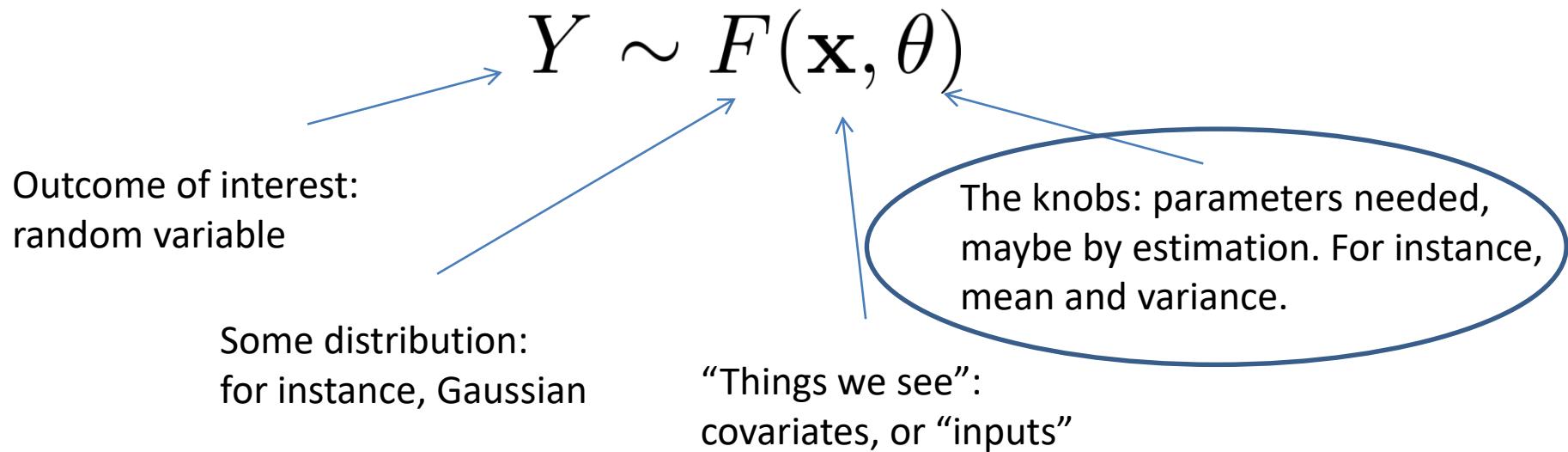


Many Names

- These covariates are sometime given many names:
 - Predictors
 - Inputs
 - Regressors
 - Independent variables (bad name!)
- The outcome is sometimes called:
 - Response
 - Output
 - Dependent variable (bad name!)

Learning a Relationship

- But how are parameters related to inputs? We need to specify how they interact to generate Y .



Example for This Section

- Advertising data (ISLR book).
- Goal: understanding how to improve sales of a particular product.
- Data: sales of that product in 200 markets
 - For each market, budgets spent on TV, radio and newspaper advertisement in thousands of dollars.
 - Outcome: sales, in thousands of units.
 - What is the relationship?

Problem Formulation

- In our problem, Y is sales volume. X , is the advertisement expenditure vector:
 - X_1 : TV
 - X_2 : Radio
 - X_3 : Newspaper
- Task: estimate how Y is related to X .
 - We can apply it to future campaigns, assuming **external validity**: that the relationship in the future remains the same. This can be a strong assumption!

In Matrix Notation

$$\begin{bmatrix} Y^{(1)} & X_1^{(1)} & X_2^{(1)} & X_3^{(1)} \\ Y^{(2)} & X_1^{(2)} & X_2^{(2)} & X_3^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ Y^{(200)} & X_1^{(200)} & X_2^{(200)} & X_3^{(200)} \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} Y^{(1)} \\ Y^{(2)} \\ \cdots \\ Y^{(200)} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} X_1^{(1)} & X_2^{(1)} & X_3^{(1)} \\ X_1^{(2)} & X_2^{(2)} & X_3^{(2)} \\ \cdots & \cdots & \cdots \\ X_1^{(200)} & X_2^{(200)} & X_3^{(200)} \end{bmatrix}$$

Regression

- Signal + noise:

$$Y^{(i)} = f_{\theta_1}(\mathbf{x}^{(i)}) + \epsilon^{(i)}$$

Signal: the **regression function**.
This is not random (\mathbf{x} is known)

Error, or “noise”.
This is random

$$\epsilon^{(i)} \sim F(\theta_2)$$

Distribution of error

In what follows, I will typically drop the superscript (i) to avoid complicating notation.

Linear Regression with Gaussian Noise

$$Y = \beta_0 + \beta^\top \mathbf{x} + \epsilon$$

- That is,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

and

$$\epsilon \sim N(0, \sigma_\epsilon^2)$$

- We have then four free parameters to fit, β_0 , β_1 , β_2 , β_3 and σ_ϵ^2 .

Why Linear?

- Because it is simple to understand.
- Computationally efficient.
- If you have many variables and not much data, might be as good as it gets (more on that later).
- Don't kid yourself, in most cases reality is not exactly linear, but:
 - **George E. P. Box's dictum, “All models are wrong but some are useful.”**

Simple Demo

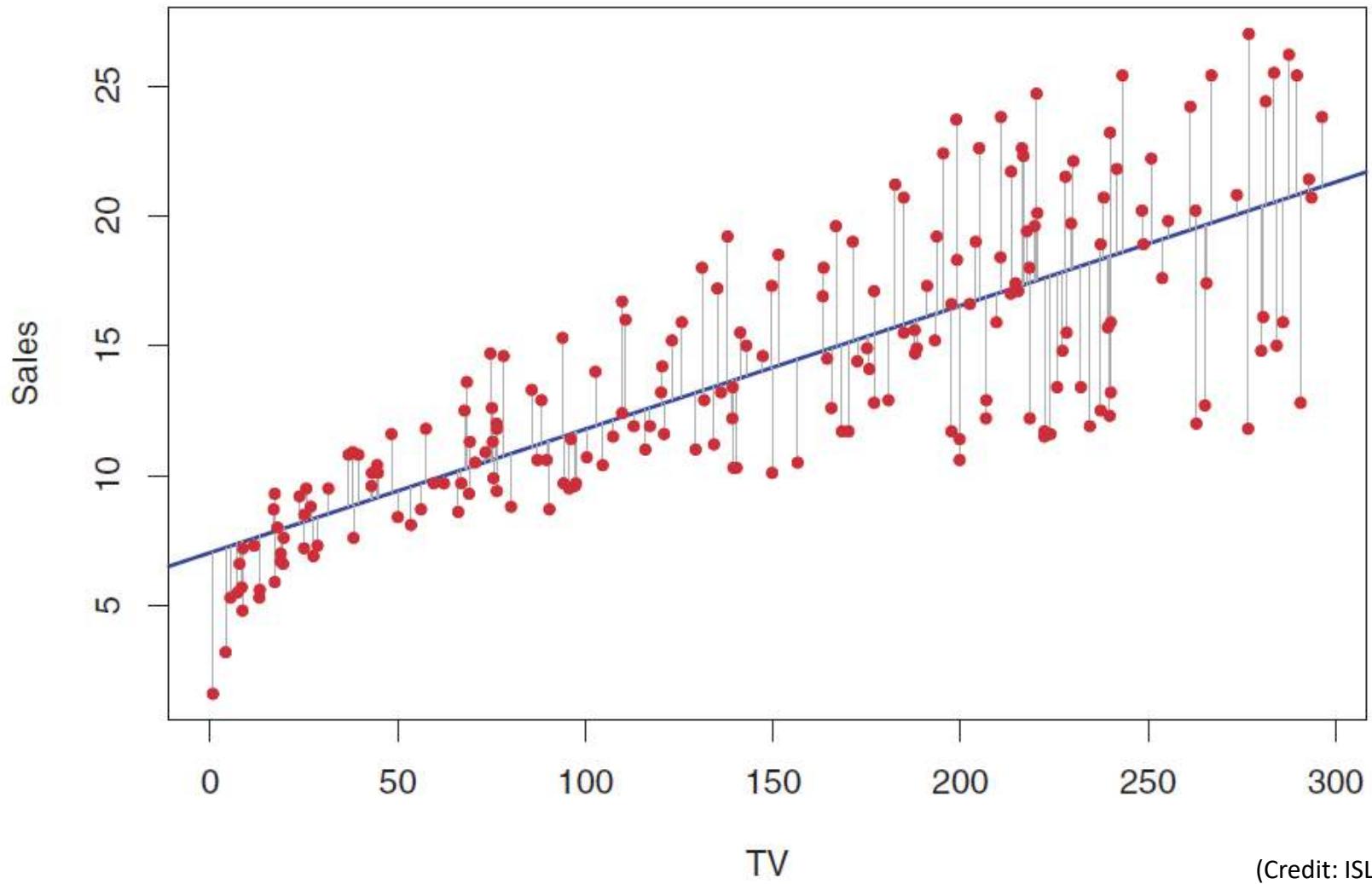
- One dimensional regression

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

where Y is sales, X_1 is TV budget and the error term here is not the same as in the previous slide (we use the same symbol as an abuse of notation).

- R demo.

The Fitted Model



Parameter Fitting: What Happened

- What is the model for the data? Recall that

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

- We assume each of our data points $Y^{(1)}, \dots, Y^{(200)}$ are independent given X .
- They are **not** identically distributed. What is their distribution?

The Distribution of $Y^{(i)}$

- Remember: $x^{(i)}$ here is fixed.
- Each $Y^{(i)}$ is a constant plus some Gaussian random variable $\varepsilon^{(i)}$.
- Without proof, we will claim that $Y^{(i)}$ is Gaussian itself. What are its mean and variance?
- We will use the notation $V = v \mid V' = v'$ to denote the **conditional distribution** of V given V' , even if V' is not a random variable.
 - Sometimes $V = v \mid v'$, when V' is obvious from context

Result

- For each data point,

$$Y^{(i)} \mid X_1^{(i)} = x_1^{(i)} \sim N(\beta_0 + \beta_1 x_1^{(i)}, \sigma_\epsilon^2)$$

- Why? Bear with me for a couple of slides.

Mean

- If Z is a random variable, what is $E[aZ + b]$ for two **constants** a and b ?

$$E[aZ + b] = \int (az + b)p(z)dz = a \int zp(z)dz + b \int p(z)dz = aE[Z] + b$$

- So if $Y = \beta_0 + \beta_1 x_1 + \epsilon$,

$$\begin{aligned} E[Y^{(i)} \mid X_1^{(i)} = x_1^{(i)}] &= \beta_0 + \beta_1 x_1^{(i)} + E[\epsilon^{(i)} \mid X_1^{(i)} = x_1^{(i)}] \\ &= \beta_0 + \beta_1 x_1^{(i)} \end{aligned}$$

Variance

- Variance is defined as

$$Var(Z) = E[(Z - E[Z])^2]$$

- That is, nothing but a quantification of how much Z differs (in expectation) from its mean by the squared Euclidean distance.
- The use of the name “variance” to describe the scale parameter of a Gaussian wasn’t a coincidence.
- We can show that $Var(aZ + b) = a^2Var(Z)$
- So

$$Var(Y^{(i)} \mid x_1^{(i)}) = 1^2Var(\epsilon^{(i)}) = \sigma_\epsilon^2$$

Now What?

- If I give you $(\beta_0, \beta_1, \sigma_\epsilon^2)$, you can tell me the probability (density) of each data point.
- We can “play with” the values of these parameters to **maximise the probability of the data occurring**
 - This is essentially the idea we sketched in the previous chapter.
- How to formalize it?

The Likelihood Function

- Probability of the data as a function of parameters:

$$L(\beta_0, \beta_1, \sigma_\epsilon^2) = \prod_{i=1}^{200} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp \left\{ -\frac{1}{2} \frac{(y^{(i)} - \beta_0 - \beta_1 x_1^{(i)})^2}{\sigma_\epsilon^2} \right\}$$

- It is easier to work on the log-scale, where we also drop constants:

$$\log L(\beta_0, \beta_1, \sigma_\epsilon^2) = -0.5 \sum_{i=1}^{200} \left(\log(\sigma_\epsilon^2) + \frac{(y^{(i)} - \beta_0 - \beta_1 x_1^{(i)})^2}{\sigma_\epsilon^2} \right)$$

The Algorithm

- Now it is a matter of computing the maximum of this likelihood function.
- This will be given the too-obvious name of **maximum likelihood estimator (MLE)**.
- Finding a MLE can be computationally hard in general (more about that in future chapters!), but here it can be done analytically.
 - That is, take derivatives, set them to zero, solve equations.

Result

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{200} (x_1^{(i)} - \bar{x}_1)(y^{(i)} - \bar{y})}{\sum_{i=1}^{200} (x_1^{(i)} - \bar{x}_1)^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1$$

$$\hat{\sigma}_\epsilon^2 = \frac{1}{200} \sum_{i=1}^{200} (y^{(i)} - \hat{\beta}_0 - \hat{\beta}_1 x_1^{(i)})^2$$

where

$$\bar{y} = \frac{1}{200} \sum_{i=1}^{200} y^{(i)} \quad \bar{x}_1 = \frac{1}{200} \sum_{i=1}^{200} x_1^{(i)}$$

Prediction

- Now for every point x_1 , we can provide a **prediction** for Y at that point.
- As in Chapter 1, we can think of the conditional expectation as an appropriate prediction.

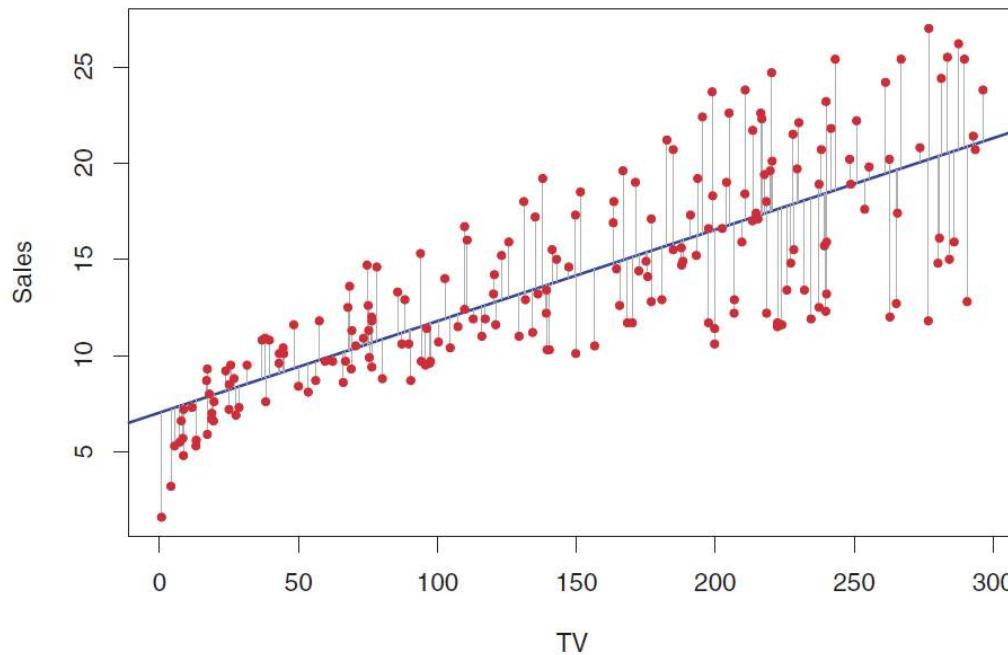
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

Prediction

- Using terminology from the Machine Learning literature, we call the data we used to fit the model the **training data**.
- It is common to reserve some data to evaluate how well we can perform **out-of-sample**, that is, with future unseen data. This data we reserved is called **test** (or testing) **data**.
 - There are more sophisticated ways of partitioning the data between training/test. See the Supervised Learning class (also, some in Chapter 5).

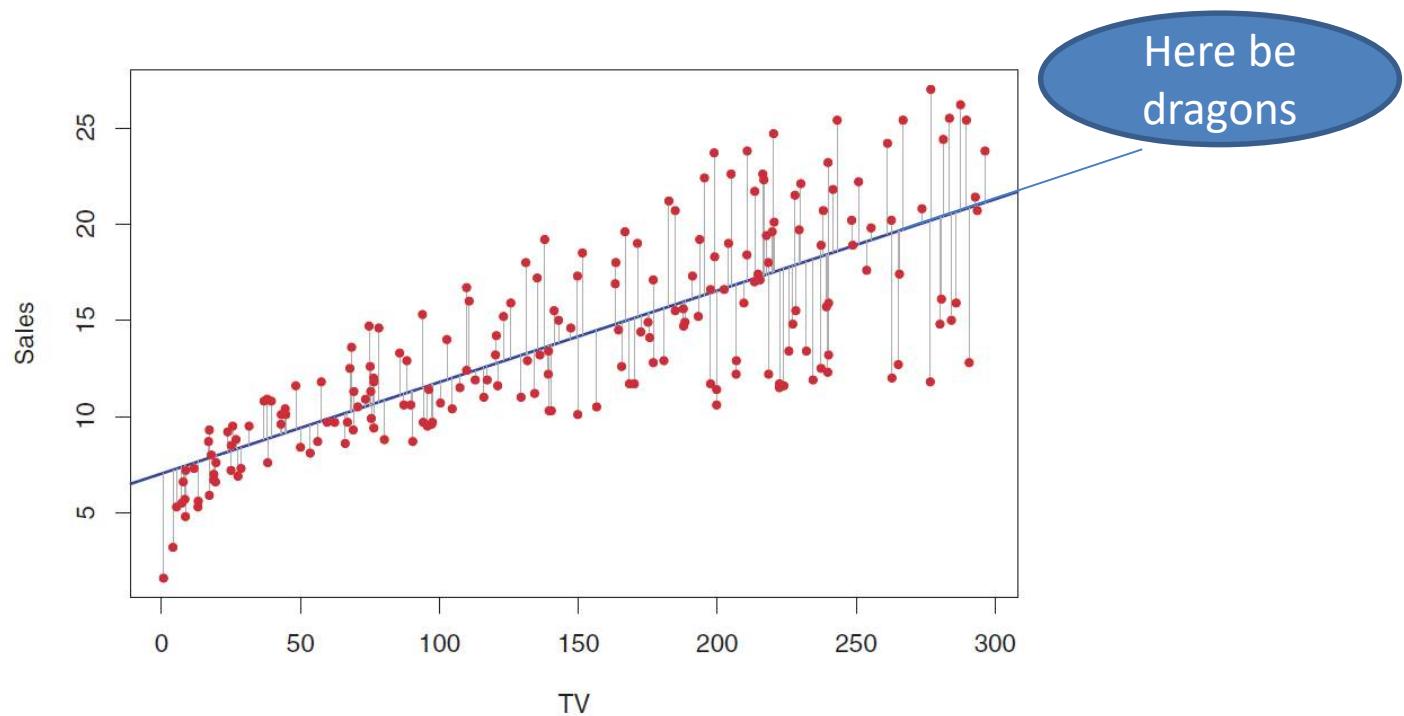
Smoothing

- We can think of **smoothing** as a way of “denoising” the **training data** you had. It provides estimates of the expectation **within-sample**.



Extrapolation

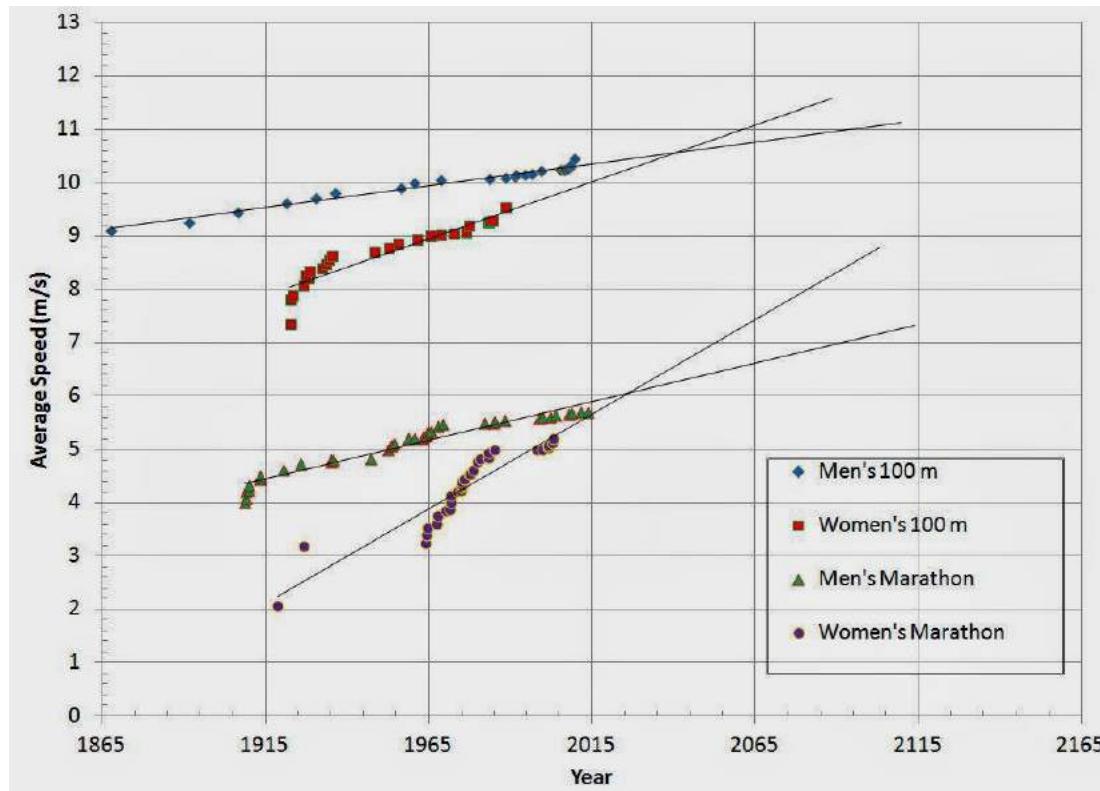
- Predictions “outside” the training data.
- Not always easy to define.
- Beware of unwarranted extrapolations!



Extrapolation

- Particularly tempting with linear models.

"This is not me talking, it's the data."



Extrapolation



If she loves you more each and every day,
by linear regression she hated you before you met.

Diagnostics

- Is this a good model? Bad? In which ways?
- Which kind of visual checks can we have as the number of inputs grows to 2, 3, ..., very many?
- Ways in which we can get things wrong:
 - Non-linearity!
 - Noise is not “homogenous” (heteroscedasticity)
 - Non-Gaussianity?
- In what follows, n will be used to denote sample size and p will denote number of inputs.

REGRESSION WITHOUT GAUSSIANITY ASSUMPTIONS

Non-Gaussianity

We will first discuss why it is not necessary to assume Gaussianity, and why and when we do/don't.

Residuals

- What you miss by your linear reconstruction.

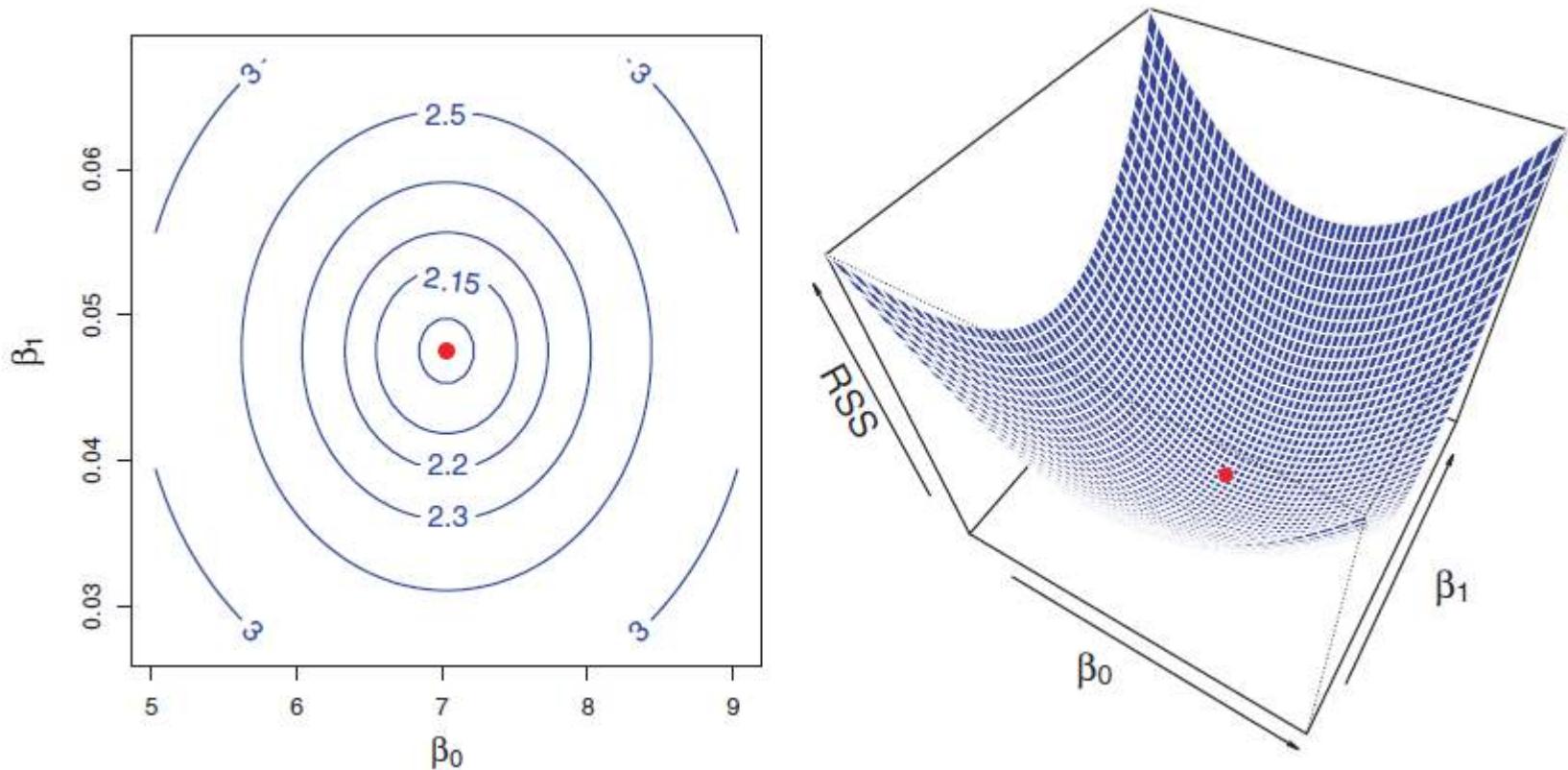
$$e^{(i)} \equiv y^{(i)} - \hat{\beta}_0 - \hat{\beta}_1 x_1^{(i)}$$

- Summary: **residual sum of squares (RSS)**

$$RSS \equiv \sum_{i=1}^n e^{(i)2}$$

- **What is the relation between that and the log-likelihood function?**

Least-Squares Interpretation



$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^n (y^{(i)} - \beta_0 - \beta_1 x_i^{(1)})^2$$

Least-Squares Estimator

This is identical to the Linear Gaussian MLE for the regression coefficients β .

What Does it Mean?

- Say you have a sample of independent, identically distributed (**i.i.d**) random variables

$$Y^{(i)} \sim F(\theta)$$

for $i = 1, 2, \dots, n$. Say you want to estimate the mean of this distribution by maximum likelihood.

- What would you do for Gaussians?

MLE for iid Gaussians

- If we maximise this

$$\log L(\mu, \sigma^2) = -0.5 \sum_{i=1}^n \log(\sigma_\epsilon^2) + \exp\{(y^{(i)} - \mu)^2 / \sigma^2\}$$

we get the **sample mean**

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y^{(i)}$$

- Notice: this is just a special case of Gaussian linear regression with an empty set \mathbf{X} .

What If We don't Want to Assume Gaussianity?

- Isn't this intuitive?

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y^{(i)}$$

- Of course it is, but how to justify it? Enter the **empirical cdf** again.

$$\hat{F}_n(x) \equiv \frac{1}{n} \sum_{i=1}^n I(x^{(i)} \leq x)$$

Empirical cdf vs Population cdf

- A central result of **nonparametric statistics** is that the empirical cdf converges (in a probabilistic sense I won't define) to the population cdf as n grows

$$\hat{F}_n(x) \rightarrow F(x)$$

- If you must know, this is called the Glivenko-Cantelli theorem.
- R demo.

Nonparametrics?

- I won't say much about this now, except that these are statistical models we can't describe with a finite number of parameters
 - More on that in Introduction to Supervised Learning, and later chapters.
- Suffices to say that unlike the Gaussian model that uses two parameters, the empirical cdf estimate uses n "parameters". It is not a fixed number.

How Do We Use This?

- Expectation, now with the “**empirical pdf**”!

$$E[Y] = \int y\hat{p}(y)dy = \sum_{i=1}^n y^{(i)} \frac{1}{n} = \bar{y}$$

- Implication? The sample mean is justified as a **consistent** estimator of means
 - It “converges” to the truth, as n increases
- This happens without assuming Gaussianity
 - even if it is exactly the same formula as in the Gaussian case

Implications to Regression

- Gaussianity assumption is not necessary to estimate the regression function, or even the error variance.
- However, we cannot say anymore that we can estimate the conditional distribution

$$Y \sim F(\mathbf{x}, \theta)$$

- So, if you need something other than mean/variance, you will need further assumptions such as Gaussianity.
- And if the conditional distribution is “far” from Gaussian, don’t fool yourself that least-squares will be reliable.

Concluding This Discussion

- We can do statistical inference without likelihood functions.
 - Bayesian inference requires likelihood function, see *STATG004*.
- There are important advantages in likelihood modelling
 - Full uncertainty modelling.
- However, more assumptions.
 - Keep in mind though that generality is not the same as reliability.

RESIDUAL ASSESSMENT AND MODEL CHECKS

Residual Assessment and Model Checks

Now we assess what residuals can tell us about accuracy, non-linearity and heteroscedasticity.

R^2 Statistic

- The RSS is not that straightforward to interpret because of its scale.
- R^2 is a proportion statistic. More exactly, the proportion of variance of Y explained by X . It always take values between 0 and 1.

$$R^2 \equiv \frac{TSS - RSS}{TSS} \quad \text{where}$$

$$RSS = \sum_{i=1}^n (y^{(i)} - \hat{y}(i))^2$$

$$TSS = \sum_{i=1}^n (y^{(i)} - \bar{y})^2$$

(Total sum of squares)

Interpretation

- $TSS - RSS$ measures “amount of variability in the outcome that is explained”.
- If we get 0, the linear model does not provide a good explanation for the data.
- Let’s check the R^2 of our advertisement example using R.
 - R also includes something called “adjusted R^2 ”.
The difference matters little for large sample sizes.

High $R^2 \neq$ Good Predictions

- High R^2 is good news, but may be not enough.
- Although in theory regression doesn't make assumptions about the distribution of covariates \mathbf{X} , its interpretation will require assumptions. This includes interpreting R^2 .
 - Recall the talk about extrapolation
- R demo with synthetic data.

High $R^2 \neq$ Good Predictions

$$R^2 = \frac{a^2 \text{Var}(X)}{a^2 \text{Var}(X) + \text{Var}(\epsilon)}$$

- This means $R^2 \rightarrow 0$ as $\text{Var}(X) \rightarrow 0$ and $R^2 \rightarrow 1$ as $\text{Var}(X) \rightarrow \infty$!
- Even with much non-linearity we can get high R^2 !
- In particular, bad predictions with high R^2 will follow if $\text{Var}(\epsilon)$ is high, but $\text{Var}(X)$ is much higher.
 - “bad” in the sense of absolute error, not necessarily in the sense of relative error with respect to not having seen x .

High $R^2 \neq$ Good Predictions

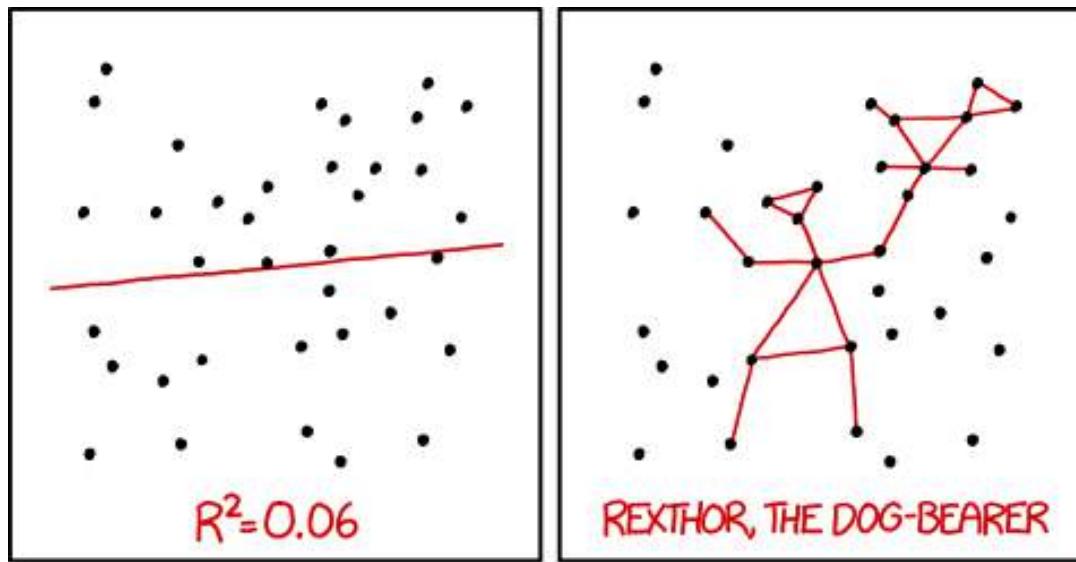
- Despite that, it is good practice to report R^2 , as a “necessary but not sufficient” diagnostic of how good your fit is.
- (Assuming fixed model) However, there is only so much we can achieve with a given \mathbf{x} :

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= [f(X) - \hat{f}(X)]^2 + \text{Var}(\epsilon) \end{aligned}$$

We can shrink this with better modelling

To shrink this, we may need to measure further variables

In Any Case...



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER
TO GUESS THE DIRECTION OF THE CORRELATION FROM THE
SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Residual Plots

- What should we expect to see in a good regression model?

$$e^{(i)} \equiv y^{(i)} - \hat{\beta}_0 - \hat{\beta}_1 x_1^{(i)}$$

- R demo: in what follows we will exemplify diagnostics by comparing the advertising model to the outcome of a well-behaved synthetic model.

Residual Plots

- R's *lm* plot 1: residuals vs fitted

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 \quad e^{(i)} \equiv y^{(i)} - \hat{\beta}_0 - \hat{\beta}_1 x_1^{(i)}$$

- What you should expect to see is lack of correlation between the two. In particular
 - The location (empirical average) and spread (empirical variance) of the residual axis stays similar across the value of the fitted outcomes.

Residual Plots

- With this plot, it might be possible to detect **outliers**, points “far from the curve” that may or may not indicate model failure.
 - Notice that the scale will depend on Y .
 - It could be the natural result of non-Gaussian error, for instance.
 - It could be the result of measurement error that *maybe* should be removed.
 - At this stage, we won’t be formal about outliers. One thing to keep in mind at this time, however, any outlier removal should be documented and justified.

Residual Plots

- R's *lm* plot 2: Normal Q-Q
 - As we have seen before, the assumption of normality is not necessary.
 - However, if there are highly skewed residuals, you might want to ask whether the mean of the outcome is a good estimand to target.
 - Violations of normality have other implications to model checking, to be discussed later.

Residual Plots

- R's *lm* plot 3: Scale-Location
 - Similar to plot 1, but transformed: square roots of absolute value of standardised residuals.
 - “standardised” = divided by empirical standard deviation
 - Rationale: horizontally, should show “no pattern” (flat red line, homogenous spread around it)
 - For Gaussian errors: approximately most points should be less than 2. But main point it to visualize homogeneity.
 - Square root is just to minimize the visual impact of more extreme points.

Residual Plots

- R's *lm* plot 4: Residuals vs. leverage
- Think of a concept that complements regression outliers: while outliers refer to point "off the y axis", it measures instead how points are "off the bulk of x values".
- This is straightforward to visualize in one dimension. For higher dimensions, we reduce it to a single number, **leverage**, which summarizes it.

Residual Plots

- R demo: let's compare two synthetic datasets that differ only by one data point.
- In one-dimension, the leverage statistic for data point x_i is

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^n (x_{i'} - \bar{x})^2}$$

Residual Plots

- The value of the leverage statistic is always between $1 / n$ and 1.
- If we have p inputs, the average leverage across inputs is $(p + 1) / n$.
- Values deviating “much” from this average can be flagged.

Residual Plots

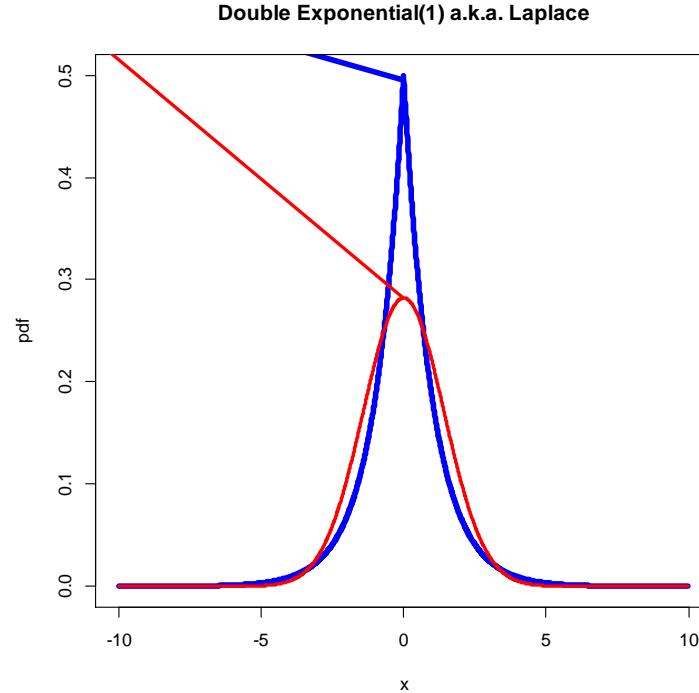
- In the residual vs. leverage plot, a point can be an **outlier** (large standardised residual) and/or a **high leverage** point.
- In R, the vertical axis is standardised, but the horizontal axis is relative. So the point of highest leverage may be inconsequential anyway.

R Demos

- Now let's walk though these plots again for an idealized examples contaminated with outliers, and one with high leverage points.

R Demos

- Now let's walk though these plots again for an idealized example with errors which follow a **double-exponential (a.k.a. Laplace)** distribution



In the figure, blue is the density of a $\text{Laplace}(1)$, red is $N(0, 2)$.

Both of these distributions have variance 2.

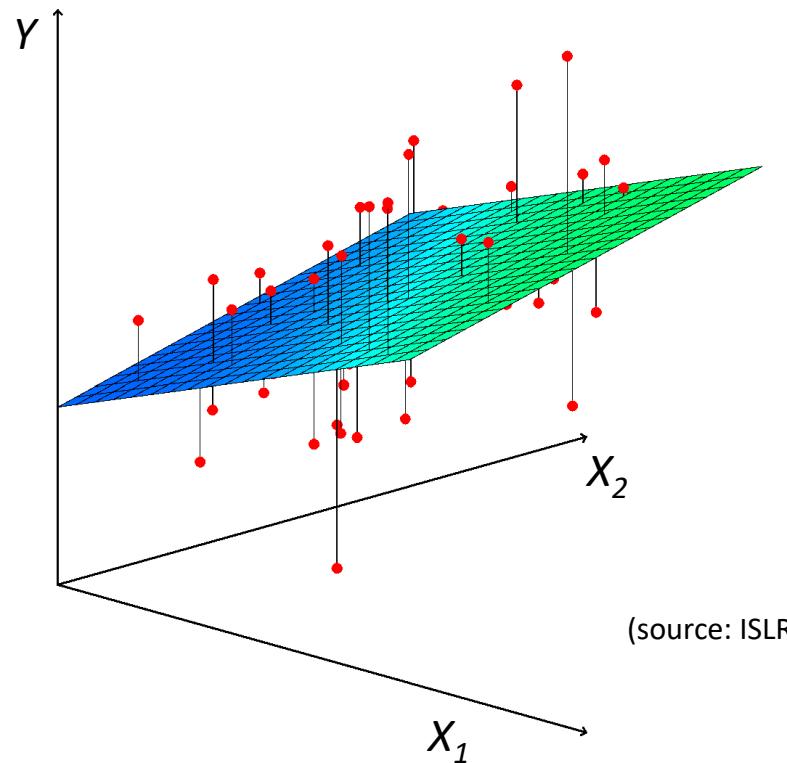
R Demos

- Now let's walk though the plots again for our advertisement data.

Finally: Multiple Regression

- Let's just fit this model, where X_1 , X_2 and X_3 are budgets for TV, radio and newspaper, respectively (R demo).

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$



(source: ISLR)

HYPOTHESIS TESTING AND CONFIDENCE INTERVALS

Two Basic Null Hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

- Rejecting it would mean β_i at least one coefficient is non-zero.
- That is, is there any association of any kind between input Y and X_i given the other inputs?
 - Notice the subtle “given the other inputs”. More on that in the future.

The All Zero H_0

- For the former hypothesis, consider the following statistic

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}$$

which refers to our old friends

$$RSS = \sum_{i=1}^n (y^{(i)} - \hat{y}(i))^2$$

$$TSS = \sum_{i=1}^n (y^{(i)} - \bar{y})^2$$

How Can F Falsify H_0 ?

- If you feel like doing some algebra, you will be able to show that

$$E \left[\frac{RSS}{n - p - 1} \right] = \sigma^2$$

where σ^2 is the variance of the error term.

- Under the null the following is also true:

$$E \left[\frac{TSS - RSS}{p} \right] = \sigma^2$$

- So, what would you suggest?

A Test of H_0

- The F statistic should be “close” to 1 under the null.
- We have a machinery to decide what closeness is:
 - Find the distribution of F
 - Assess the probability (with respect to the data distribution) that F is greater than 1
 - Technical note: $E [(TSS - RSS)/p] > \sigma^2$ if H_0 is false.
 - Reject H_0 if this probability is smaller than your agreed test level (sigh... “0.05” for the sake of illustration)

Implications

- If your p-value is low, H_0 is rubbish at explaining the data: reject it.
 - If you must know, the F statistic follows (approximately, in the non-Gaussian case) the unimaginatively named F distribution, which I won't explain.
- This is a test that is commonly reported, **but don't fool yourself that this is evidence of a good model.**
 - Your data is arguably very bad if this very strong H_0 is not rejected.

R Demo

- Recall that our linear model of sales volumes looks preposterous. But guess its p-value.

Testing Subsets

- There are analogous F statistics for the null $\beta_i = 0$ only (that is, the other coefficients are unconstrained). As a matter of fact, we can easily test whether any subset of coefficients is zero.
- Many software packages report the one-coefficient test automatically.

Implications

- If your p-value is high, there is evidence predictor X_i does not explain the variability of the outcome *given the other predictors*.
- This is not the same as input X_i not being important.
- R demo.

Implications

- If you do find evidence that predictors are important (e.g., tests give low p-values), again this does not mean the model is “good”.
- However, testing provides an useful indication of which variables are redundant or not quite useful, *given the sample size you have and the model assumptions.*
 - They *might* prove useful if you later collect larger sample sizes.

Beware of the Star-Chasing Complex

Call:

```
lm(formula = adv$Sales ~ adv$TV + adv$Radio + adv$Newspaper)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.8277	-0.8908	0.2418	1.1893	2.8292

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	Signif. codes:
(Intercept)	2.938889	0.311908	9.422	<2e-16	***
adv\$TV	0.045765	0.001395	32.809	<2e-16	***
adv\$Radio	0.188530	0.008611	21.893	<2e-16	***
adv\$Newspaper	-0.001037	0.005871	-0.177	0.86	

Signif. codes: 0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1

These might be
there even
if your model
has no predictive
value

Residual standard error: 1.686 on 196 degrees of freedom

Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956

F-statistic: 570.3 on 3 and 196 DF, p-value: < 2.2e-16

Beware of the Star-Chasing Complex

- In a later chapter we will discuss variable selection and what it means in practice.

Confidence Intervals

- This is very similar to the general idea. Find some pivot around which an interval can be built.
- As a technical aside, let us define an adjusted estimator for the error variance.

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n (Y_i - \beta_1 x_1 - \cdots - \beta_p x_p)^2$$

(from now on, we assume β_0 can be represented by $\beta_1 \times 1$, where X_1 is always 1.)

Confidence Intervals

- The following can be shown to be true when errors are Gaussian:

$$T_i \equiv \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma} \sqrt{v_{ii}}} \sim \mathcal{T}(n - p)$$

and v_{ii} is the i th entry of the diagonal of $(\mathbf{X}^T \mathbf{X})^{-1}$.

- Notice this requires $n > p$
 - As a matter of fact, least-squares is ill-defined if $n < p$.
- Exercise: write an expression for a confidence interval for β_1 of coverage $1 - \alpha$.
- Note: CLT applies and Gaussianity ends up again not being that important.

Predictive Intervals

- A different matter is **predictive intervals**.
- In Supervised Learning, you may see a lot about prediction. But going one step further, we might want to characterize uncertainty in the prediction. This takes into account uncertainty of the estimates.

What We mean by That

- If we assume a model like the Gaussian, then this implies uncertainty as a conditional distribution

$$Y = \beta_0 + \beta_1 x_1 + \epsilon \Leftrightarrow Y \mid X_1 = x_1 \sim N(\beta_0 + \beta_1 x_1, \sigma^2)$$

- First, let's look at the uncertainty of the **expected value of outcome** given inputs when all we have are parameter estimates.

Prediction

- Say a new data point x_1^* comes, and you want to predict the output as follows

$$\hat{Y}^* \equiv \hat{\beta}_0 + \hat{\beta}_1 x_1^*$$

- We know the coefficients themselves are random variables if we consider the training data to be random. What is the long-run variability of my prediction?

Answer

$$Var(\hat{Y}^*) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_1^*)$$

- Let's not worry about how to calculate this. The important message is the interpretation: **the randomness here is in the estimated coefficients**, and they come from the randomness in the training data.
- To get the variance of Y^* itself (notice the lack of a hat), we also use the variance of the error.

Predictive Variance

- In practice, we cheat a little bit: the estimated variance $\hat{\sigma}_\epsilon^2$ of the error term is given by the empirical variance of the residuals and treated *as if* it was known.
- Also, recall $\text{Var}(W_1 + W_2) = \text{Var}(W_1) + \text{Var}(W_2)$ for two arbitrary **independent** random variables W_1 and W_2 .
- So

$$\text{Var}(Y^*) = \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \epsilon) \approx \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_1^*) + \hat{\sigma}_\epsilon^2$$

OTHER PRACTICAL ISSUES AND DIAGNOSTICS

Interpretation of Regression Models

- We fit the model of sales volume against TV, radio and newspaper expenditures. We get this:

$$Y = 2.93 + 0.04x_1 + 0.19x_2 - 0.001x_3 + \epsilon$$

- What is its interpretation?
 - Recall first the units: sales volume is measured in thousands of units; each advertising budget is in thousands of dollars.

Interpretation of Regression Models

- The dangerous conclusion:

“If we increase the TV budget by one thousand then, other things being equal, I will sell 400 hundred more units of my product, in expectation.”

$$Y = 2.93 + 0.04x_1 + 0.19x_2 - 0.001x_3 + \epsilon$$

Careful!

- Let me tell you the following extra piece of information: this data came from an **observational study**.
- That is: there was no documented explanation on the causes leading to the level of TV expenses.
- Why this is relevant? Because there might be **common causes (confounding)** of both sales volume and TV expenses. **They may be hidden.**
 - For instance, TV budgets are bigger in markets that are stronger economically, where also people are more likely to buy your product anyway.

Careful!

- Under some strong assumptions, it might be possible to extract causal effects from observational studies. In other situations, you might have **randomized controlled trials**. Then your regression coefficients can be interpreted as causal effects.
 - A big topic in itself that I will leave entirely to *STATG002*
- Without these conditions, some people still refer to regression coefficients as “effects”. This is common, but I find it preposterous.

Interpretation of Regression Models

- A more sober conclusion:

“Other budgets being equal, an increase of TV budget by one thousand dollars will correspond to an increase of 400 hundred more units of my product, in expectation.”

$$Y = 2.93 + 0.04x_1 + 0.19x_2 - 0.001x_3 + \epsilon$$

- **Notice the major difference:** “increase” here means a increase “as in” the training set, whatever black-box mechanism that was.

Interpretation of Regression Models

- “Other budgets being equal”. **Your regression coefficients depend entirely on which other variables are included.**
- Notice the major difference!

$$Y = 2.93 + 0.04x_1 + 0.19x_2 - 0.001x_3 + \epsilon_{123}$$

$$Y = 12.35 + 0.05x_3 + \epsilon_3$$

I'm emphasizing which variables I'm using as inputs.

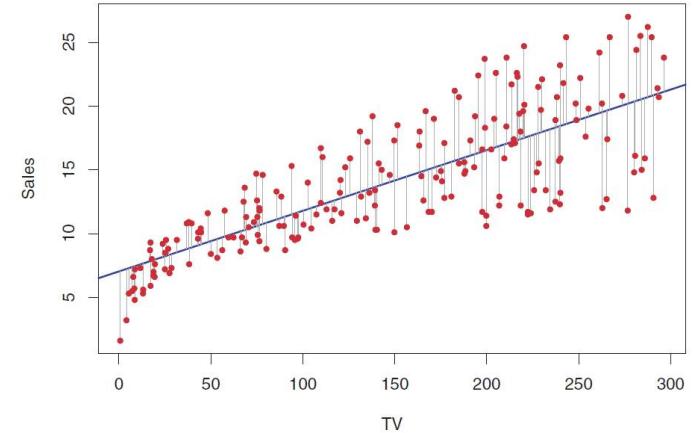
- Be careful to contextualize what you mean by a variable being “important”.

The Linear Elephant in the Room

- We know that for our advertising data, linearity is not particularly great.
- There are all sorts of great nonlinear black-box models
 - Introduction to Supervised Learning, and Chapter 5 of our course, will address very many of them.
- However, it sometimes pays off to improve the humble linear model with a change of representation.

Logarithm Transforms

- Heteroscedasticity looks strong in this problem.
- Sometimes it is the result of **multiplicative errors**.



$$Y = x\epsilon$$

- Logarithm transforms can be taken with non-negative data.

Logarithm Transforms

- In our advertising data, let's try taking
 - the logarithm of TV budget
 - sales volume
 - both
- R demo.

Well, That Wasn't Great Was it?

- But it illustrates the principle that we can stick to a linear model that builds a nonlinear mapping (logarithm, in this case).
 - A principle that is taken to the extreme with kernel methods, as discussed in Supervised Learning.
- Other transformations can be done, for instance using a quadratic polynomial.

$$Y = \beta_0 + \beta_1 x_1 + \beta_{12} x_1^2 + \epsilon$$

Interactions

- From the point of view of interpretability, one common use of the linear model is through quadratic or higher order polynomials, e.g.
- This is in part to the idea of interpreting **interactions**. For instance, with the advertising data, is there a “synergy” effect between media?
 - Spending more money on radio could “change the slope” for TV, if the extra exposure makes people pay more attention to TV adverts.

Interactions

- In a linear model, this is translated by constructing inputs derived from the product of more basic inputs.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

- R demo.
- Notice that pairwise interactions already substantially increase the number of inputs.

Notice

- By adding non-linear transformations as inputs to your linear model, this essentially gives you a test of whether the linear model in the original space was reasonable.
- Because if the hypotheses of zero-coefficient for the non-linear terms are rejected, then the original representation was not good enough.

Discrete Inputs and Interpretation

- In Supervised Learning, you may have seen already how to deal with discrete inputs.
 - In Statistics, sometimes we use the generic term **categorical variable** or **factor** to mean discrete variable. Discrete variables can also be **ordinal** if they have a meaningful ordering (think number of stars in a Netflix rating), and can, of course, be **counts**.

Discrete Inputs

- Binary variables (e.g., gender) can be typically represented as 0 or 1, as we have seen before.
- In the context of regression

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

translates to either

$$Y = \beta_0 + \beta_1 + \epsilon$$

or

$$Y = \beta_0 + \epsilon$$

Example

- The *Credit* dataset (from ISLR).
- “Balance” as output, “Gender” as input. We will treat (arbitrarily) level *Female* as 1, *Male* as 0.
Let’s do a R demo.
- Alternatively, we could code these levels as 1 and -1:

$$Y = \beta_0 + \beta_1 + \epsilon$$

$$Y = \beta_0 - \beta_1 + \epsilon$$

so β_0 can be interpreted as the “baseline credit balance”

Interpretation with More than Two Levels

- We can again create a “dummy” encoding, again with the idea of having one fewer variable than the number of values.
 - So, one dummy for binary variables, two for variables with three levels and so on.
- The reason for the “one fewer” rule is the lack of **identifiability** otherwise.
 - That is, there are infinitely many coefficients giving the same output.

Interpretation with More than Two Levels

- For instance, let's say we have X_1 as an indicator that someone is male ($x_1 = 0$ if person is not male, 1 otherwise). Let's have X_2 as an indicator that someone is female.
 - Clearly $x_1 + x_2 = 1$, so these two models are identical for any c :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$Y = (\beta_0 - c) + (\beta_1 + c)x_1 + (\beta_2 + c)x_2 + \epsilon$$

Interpretation with More than Two Levels

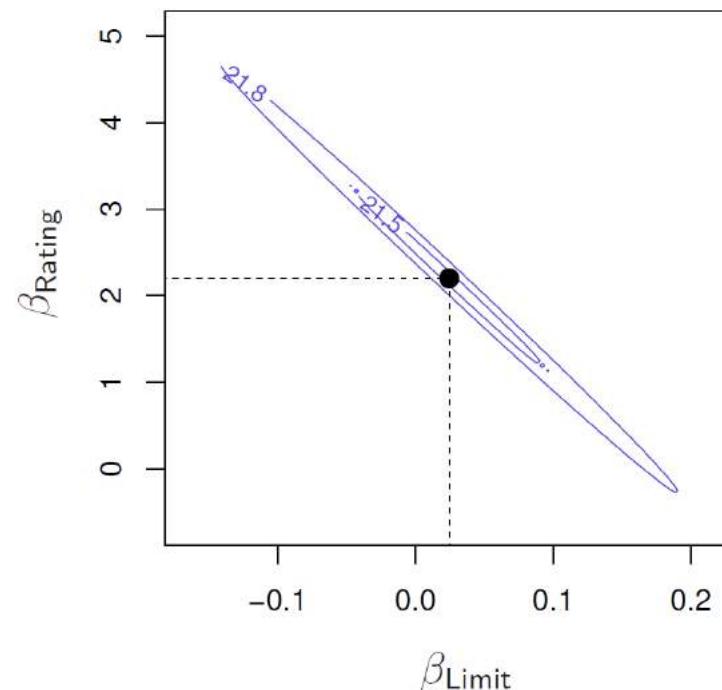
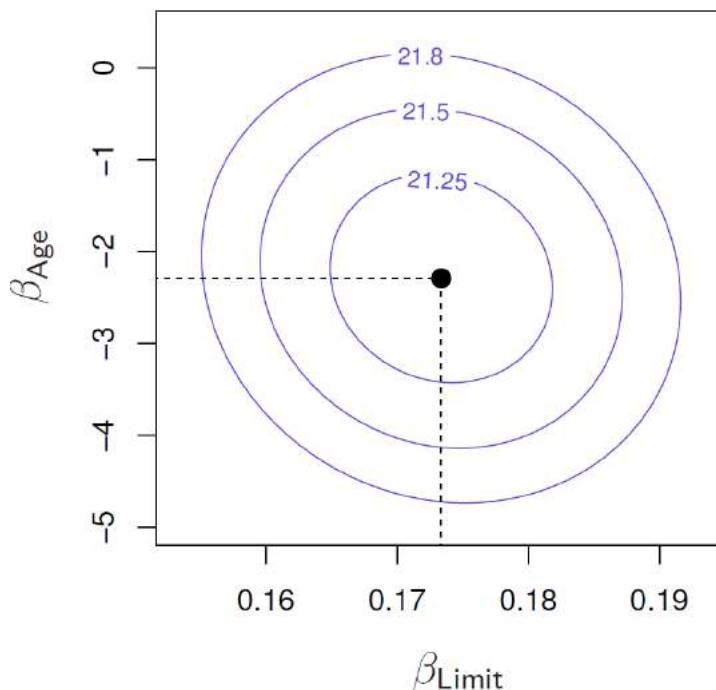
- This restricted encoding is not symmetric, but without it the linear model would break down.
 - R demo.
- The choice of “base level” is up to the practitioner.
- See also: **analysis of variance** in *STATG002*.

Final Comment: Collinearity

- A “softer” version of the problem of identifiability in linear models: variables which are almost linear combinations of others.
- What happens in the Credit dataset? For instance, the relation between credit and rating? (R demo)

Collinearity

- How does the RSS change with parameter values? Bivariate regression plots.



Collinearity

- Interpreting parameters of variables which are linearly related is not possible due to unidentifiability.
- Interpreting parameters of variables which are almost linearly related may be unreliable due to wide confidence intervals.

Collinearity

- Try to understand whether it makes sense to have nearly-collinear variables in your model. Remember that each coefficient describes the association between a given input and output holding the other inputs fixed.
 - This is “mutually assured destruction” if an input can basically be derived from the others.

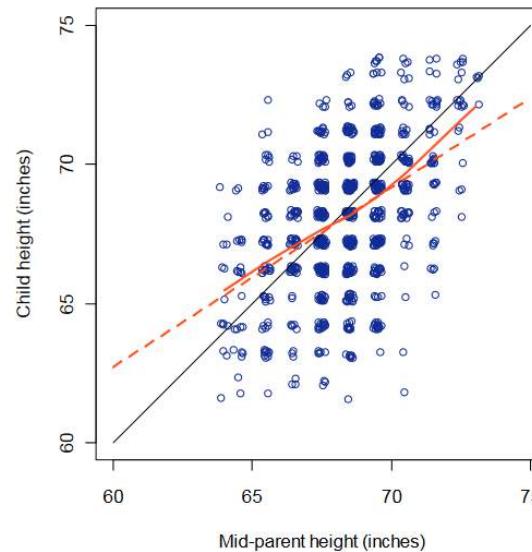
Take-Home Messages

- Linear regression is *the* workhorse of data analysis.
- Prediction is not everything: judicious interpretation of the model and where it fails to fit is also there to convey further messages.
- Confidence intervals help with that, while hypothesis testing provides some basic evidence of what your data can tell about the model components.

Next: model-based regression beyond Gaussianity.

A Historical Note

- The idea of least-squares dates back to at least Gauss and Legendre.
- The name “regression” itself comes from Francis Galton.



Yes, the very same Galton who names our lecture theatre. He was a mentor of Karl Pearson, who founded our department.

Introduction to Statistical Data Science

Ricardo Silva

ricardo@stats.ucl.ac.uk

Department of Statistical Science, UCL

Generalised Linear Models

Outline

We will have a more detailed look at regression models when the outcome is not Gaussian, but of a different probability family.

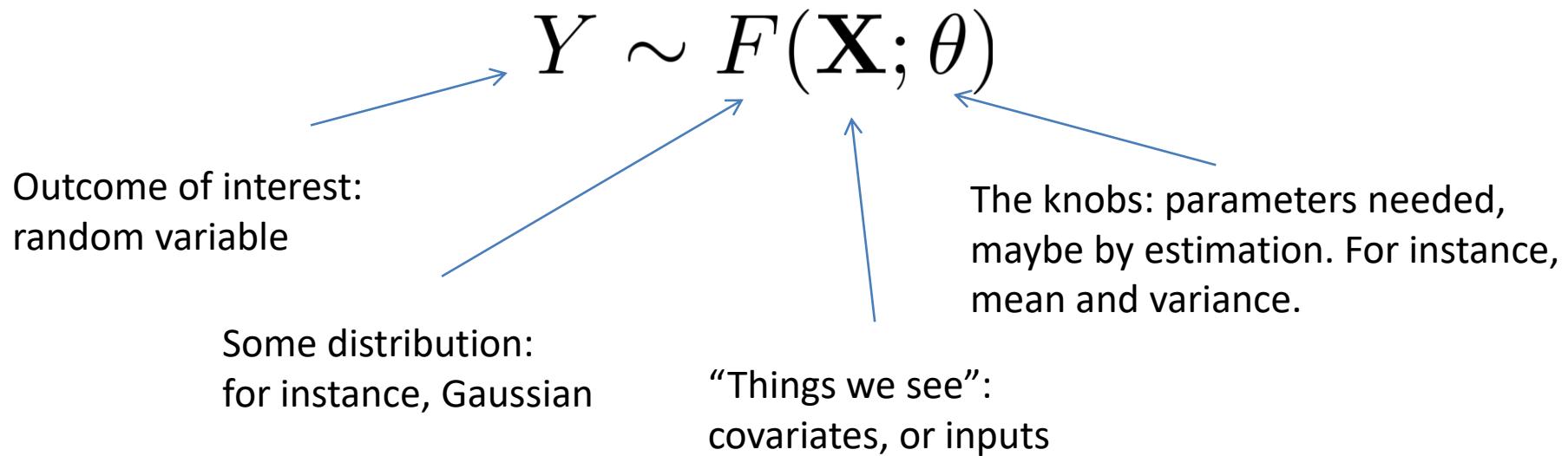
This will also be an opportunity to be exposed to more probabilistic modelling.

Outline

- Basic definitions
- Logistic regression
 - And **maximum likelihood**
 - And **deviance**
- Beyond logistic regression
 - Poisson, Negative Binomial and Ordered Logit models
- Brief mathematical comments
 - Putting the “general” in “generalised linear models”
 - Notions of numerical optimisation: the GLM example

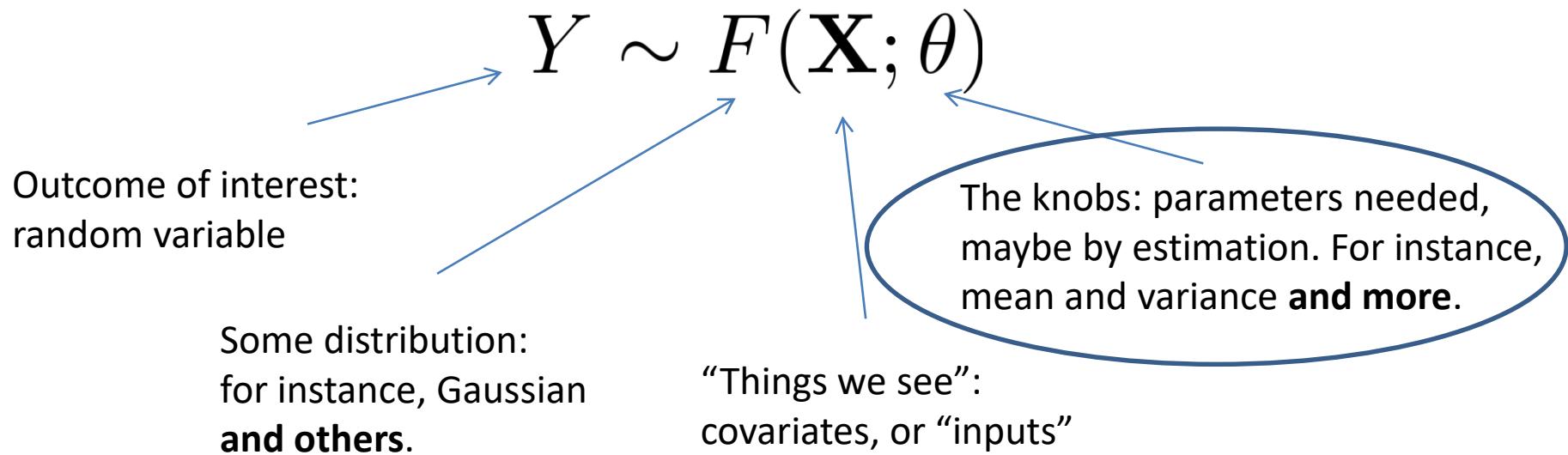
Recall: Learning a Relationship

- Our measurements are not independent.
- Often we want to characterize the distribution of an **outcome** Y given observable **covariates** \mathbf{X} :



Recall: Learning a Relationship

- But how are parameters related to inputs? We need to specify how they interact to generate Y .



Generalised Linear Models

- As before, we have inputs and coefficients β .
- Product $\mathbf{X}\beta$ is a real vector, each entry a real number. We will call it the **linear predictor**.

$$\mathbf{X} = \begin{bmatrix} X_1^{(1)} & X_2^{(1)} & \dots & X_p^{(1)} \\ X_1^{(2)} & X_2^{(2)} & \dots & X_p^{(2)} \\ \dots & \dots & \dots & \dots \\ X_1^{(n)} & X_2^{(n)} & \dots & X_p^{(n)} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

Generalised Linear Models

- Least-squares is intuitive as the expected value of a Gaussian. What if outcome is (say) binary?

$$Y \sim \text{Bernoulli}(\theta)$$

$$P(Y = y) = \theta^y(1 - \theta)^{1-y}, \quad y \in \{0, 1\}$$

Generalised Linear Models

- Least-squares is intuitive as the expected value of a Gaussian. What if outcome is (say) count data?

$$Y \sim Poisson(\theta)$$

$$P(Y = y) = \frac{\theta^y \exp^{-\theta}}{y!}, \quad y \in \{0, 1, 2, 3, \dots\}$$

Generalised Linear Models

- Idea: we do a “two-stage” model.
- Use linear predictor as a meta-parameter: transform it to get the parameter(s) of the target distribution.
- This transformation will be called a **link function**.

Generalised Linear Models

$$P(Y = y) = \theta^y(1 - \theta)^{1-y}, \quad y \in \{0, 1\}$$

- Say you have a single data point, linear predictor is

$$\eta^{(i)} \equiv \sum_{j=1}^p x_j^{(i)} \beta_j$$

- We want to model the probability of $Y^{(i)}$ taking value 1 according to the information in $x^{(i)}$.

$$\eta^{(i)} = g(\theta^{(i)})$$

Link function

LOGISTIC REGRESSION

Logistic Regression

- You may have seen some of this before in Supervised Learning: the **logit** link function. Its inverse is the **logistic** function:

$$\theta = g^{-1}(\eta) \equiv \frac{1}{1 + e^{-\eta}}$$

- This is convenient, as it maps a real number for the interval $[0, 1]$, which is precisely the range of allowable values for θ .

Interpretation

- This is equivalent to the following :

$$\log \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- We call the term in the left hand side the **log-odds** for Y (given X). Odds of 1 are equivalent to a probability of 0.5.
- The ratio of two odds is called... you guessed, an **odds ratio**.

Interpretation

- What is the relation between this and β_1 ?

$$\log \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- If X_1 increase by one unit, the log-odds for Y increases by β_1 .
- So $\exp(\beta_1)$ is the odds ratio for the odds of Y at $X_1 = x_1 + 1$ against odds at $X_1 = x_1$.
- In other words $\exp(\beta_1)$ tells you the **rate of change of the odds** per unit of X_1 , other things being equal.

Latent Data Interpretation

- Suppose that for each $Y^{(i)}$ there is some continuous, unobserved (or **hidden**, or **latent**) $Z^{(i)}$ such that

$$y^{(i)} = \begin{cases} 1, & \text{if } z^{(i)} > 0. \\ 0, & \text{otherwise.} \end{cases}$$

where

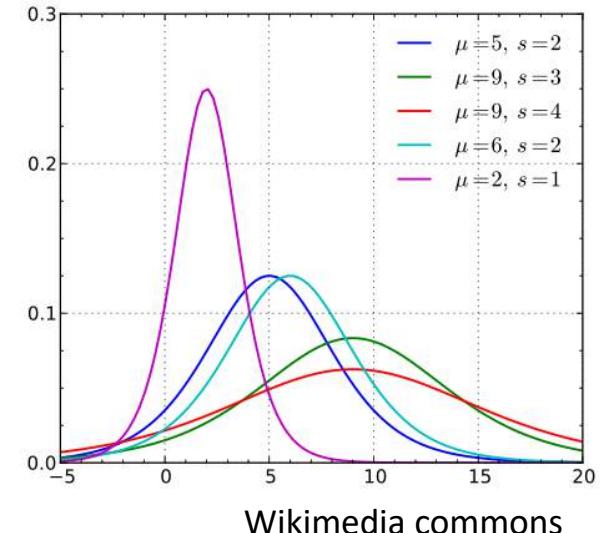
$$Z^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)} + \epsilon^{(i)}$$

and $\epsilon^{(i)}$ follows a Logistic (0, 1) distribution.

Sidenote

- Logistic distribution?

- It is not particularly important to know this distribution. It suffices to say it is very similar to a Gaussian distribution
 - Logistic (0, 1) being almost a Normal(0, 2.56)



- Notice that the scale of the logistic error is unimportant, because it is *unidentifiable*.

Latent Data Interpretation

$$y^{(i)} = \begin{cases} 1, & \text{if } z^{(i)} > 0. \\ 0, & \text{otherwise.} \end{cases}$$

$$Z^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)} + \epsilon^{(i)}$$

- This point of view requires an interpretation of the latent variable.
 - “Ability”, “propensity”, “utility” etc.
- Conditioned on this interpretation of the latent, logistic regression coefficients assume the same interpretation of linear regression coefficients *with respect to Z*.
- This interpretation is more convenient for **ordinal** variables, which we will see later.

Example

- “Default” data from ISLR: data on credit card defaults. Includes measurements such as balance (in dollars).
- Outcome variable: has the person defaulted or not? First, fit logistic regression to input “balance”.
 - I will be mostly interested on interpretation and fitting assessment. This data is clearly of relevance for **prediction**, but I’ll leave these matters mostly for the Supervised Learning course. (More on Chapter 5)
 - Notice: minimising least-squares error is theoretically **consistent** for separating binary outcomes. However, it might require much larger sample sizes than logistic regression (and change of representation).

Assessing Fit

- R demo.

$$\log \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)} = -10.65 + 0.005x$$

- Can you think of cases where interpretation will be relevant?
- Measures of fit: notice that R^2 is not reported. Instead we have something else called **deviance**.

Deviance

- This measure of fit is based on the likelihood function.
- Please bear with me for now: we will take a detour to give **more details on likelihoods and the maximum likelihood estimator**.
- These are very fundamental concepts that go far beyond generalised linear models.

Likelihood

- We have seen this concept before when we talked about linear regression with Gaussian errors. A blast from a recent past:

$$L(\beta_0, \beta_1, \sigma_\epsilon^2) = \prod_{i=1}^{200} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp \left\{ -\frac{1}{2} \frac{(y^{(i)} - \beta_0 - \beta_1 x_1^{(i)})^2}{\sigma_\epsilon^2} \right\}$$

- This is of course a probability (density) *with respect to the data*. However, we are interested on what happens to it *with respect to the parameters*.
 - Hence, the name “likelihood”, to distinguish it from probability (density). **THERE IS NO SUCH A THING AS “LIKELIHOOD OF THE DATA”. IT IS THE LIKELIHOOD OF PARAMETERS WE ARE TALKING ABOUT.**

Likelihood in the Logistic Case

$$P(Y^{(i)} = y \mid X^{(i)} = x) = \left(\frac{1}{1 + e^{-\beta_0 - \beta_1 x^{(i)}}} \right)^{y^{(i)}} \left(1 - \frac{1}{1 + e^{-\beta_0 - \beta_1 x^{(i)}}} \right)^{(1-y^{(i)})}$$

- As mentioned before, as we have many data points (10,000 in this case) it makes more sense to look at the log likelihood l :

$$l(\beta_0, \beta_1) = \sum_{i=1}^{10,000} -y^{(i)} \log(1 + e^{-\beta_0 - \beta_1 x^{(i)}}) + (1 - y^{(i)}) \log \left(\frac{e^{-\beta_0 - \beta_1 x^{(i)}}}{1 + e^{-\beta_0 - \beta_1 x^{(i)}}} \right)$$

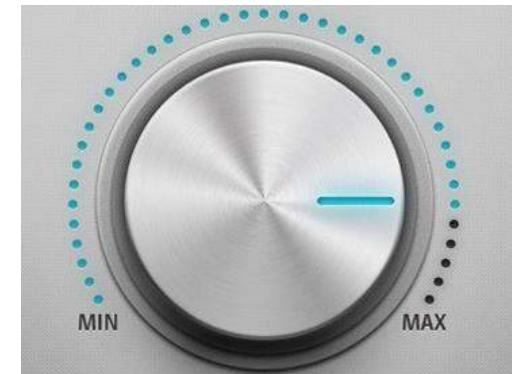
(Notice this can be further simplified)

Maximum Likelihood

- We *maximise* this function with respect to its arguments to obtain the “best fit”:
 - Intuition: making the data “as probable as possible”
- Unlike the Gaussian case, we cannot solve it analytically.
- In Supervised Learning, you may have seen some gradient-based methods. Many programming languages have black-box optimisers.
 - I will avoid repeating it, but I’ll have more to say about it at the end of this chapter.

Interpretation

- Let's ignore X for now. Say we want to fit a probability mass function for a Bernoulli random variable Y . All you see is a series of “coin flips”.
- You have only one dial to tune properly, parameter θ representing the probability of “heads”.

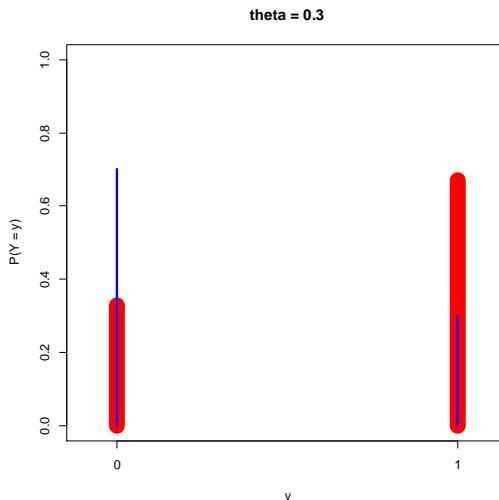


Matching the Empirical Distribution

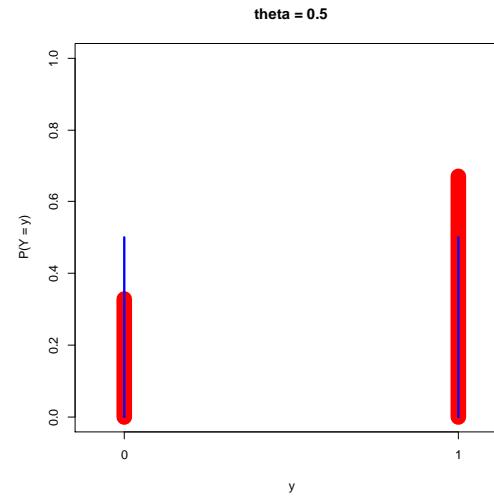
- For any given θ , there are different ways of quantifying how “far” we are from the data.
- Recall the *empirical pdf* from the previous chapter.
- For discrete data, the **empirical pmf** boils down to frequencies at the possible points, as found in the data.

Quantify This

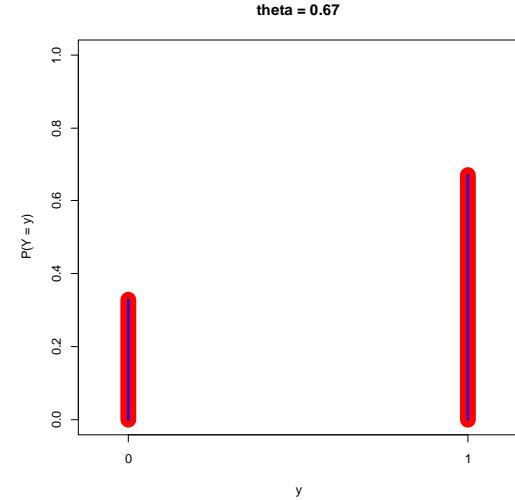
- Red lines: empirical pmf
- Blue lines: models proposed by turns of the dial.



“Poor” fit



“Mediocre” fit



“Perfect” fit

Quantify This

- Maybe we minimise the Euclidean difference between the heights of the bars?
 - Works in principle, but what does it mean when data is continuous?
 - Also, your “dials” may not be flexible enough. What does this mean when data is e.g. approximately Gaussian only?
 - **Statistical efficiency:** some minimisation criteria might be less stable (i.e., high variance) than others. Without getting in details, likelihoods are a good idea.

Maximum Likelihood as KL Divergence

- The Kullback-Leibler Divergence (**KL divergence**) between two pmfs “p” and “q” is defined as this (for y where $p(y) > 0$):

$$KL(p||q) \equiv \sum_y p(y) \log \frac{p(y)}{q(y)}$$

- There are ways of motivating this particular measure. It suffices to say it is never negative, and equals zero if and only if $p(y) = q(y)$ for all y .

Maximum Likelihood as KL Divergence

- We make $p(y)$ our empirical pmf,

$$\hat{p}_n(y) \equiv \frac{\#\text{number of data points equal to } y}{n}$$

- Make $q(x)$ our model, which we can tune with our dial θ . Then for the Bernoulli case we get

$$KL(p||q) = \text{constant} - \sum_y \hat{p}_n(y) \log q(y; \theta)$$

Something that does not depend on θ

Maximum Likelihood as KL Divergence

- So minimising KL with respect to θ is the same as maximising this:

$$\hat{p}_n(0) \log q(0; \theta) + \hat{p}_n(1) \log q(1; \theta)$$

=

$$\frac{1}{n} \sum_{i=1} (1 - y^{(i)}) \log(1 - \theta) + y^{(i)} \log(\theta)$$

- That is, maximum likelihood!

Maximum Likelihood as KL Divergence

- The Law of Large Numbers tells us

$$\hat{p}_n(y) \rightarrow p(y)$$

as n increases, which means maximum likelihood is consistent* (and consistency is good!), since maximising it will make $q(y) = p(y)$ in the limit!

- Although not directly pertinent to our logistic regression model, let's just end this detour with an example with continuous data. This will be relevant in general.

* Annoying technical conditions apply

Gaussian Example

- Recall the empirical pdf (now a bit more formally):

$$\hat{p}_n(y) = \begin{cases} 1/n, & \text{if } y \text{ is in the data.} \\ 0, & \text{otherwise.} \end{cases}$$

- KL divergence can be defined for pdfs too. Our dials will now be the mean and variance of a Gaussian model.

KL in the Gaussian Case

- Among two densities

$$KL(p||q) \equiv \int_{-\infty}^{+\infty} p(y) \log \frac{p(y)}{q(y)} dy$$

- Plug in empirical pdf $p(y) = 1 / n$, and model

$$\sum_{y \text{ in the data}} \frac{1}{n} \log \frac{1/n}{q(y; \mu, \sigma^2)}$$

KL in the Gaussian Case

- Get rid of constants, maximise

$$\sum_{i=1}^n \log q(y^{(i)}; \mu, \sigma^2)$$

which is by now our well-known expression

$$l(\mu, \sigma^2) = - \sum_{i=1}^n \left(\log(\sigma^2) + \frac{(y^{(i)} - \mu)^2}{\sigma^2} \right)$$

- Notice we got rid (again) of constant terms that do not depend on the parameters.

(R demo)

End of Detour

- All of the lessons here apply when there is a nesting of parameterizations.

$$\theta^{(i)} = g^{-1}(\eta^{(i)}) = g^{-1}(\beta, x^{(i)})$$

- Notice that in regression, output variables are not (conditionally) iid. They are typically *independent*, but NOT *identically distributed*.

Back to the Where We Left

- Deviance, for logistic regression.
 - More general formulation later.
- It is a function of the maximum likelihood of the model.
- It is called “deviance” because it is defined as a contrast of the desired model against the **saturated model**.

Saturated Model

- Imagine we could set $\theta^{(i)}$ to be *whatever we wanted* instead of a logistic regression model

$$L(\theta) = \prod_{i=1}^n (\theta^{(i)})^{y^{(i)}} (1 - \theta^{(i)})^{1-y^{(i)}}$$

- It doesn't take amazing calculus skills to realize that the MLE in this case is

$$\hat{\theta}_{sat}^{(i)} = y^{(i)} \quad \text{for all } i$$

Saturated Model

- This is of course a terrible model (in machine learning parlance, it badly overfits – it has zero generalization ability).
- But it allows us to derive a useful summary statistic with known asymptotic distribution. Below, the definition (using β as a common piece of notation):

$$D \equiv 2[l(\hat{\beta}_{sat}) - l(\hat{\beta})] \quad D \sim \chi^2_{n-p}$$

The Distribution

- That distribution is called a chi-squared with $n - p$ degrees of freedom. I will spare you of the details.
- What suffices to say of this statistic is that
 - (1) it quantifies what we lost by adding constraints (linearity of log-odds), and
 - (2) we know its approximate distribution.



Wikimedia Commons

General Model Comparison

- Moreover, say you have two **nested models**: model M_1 is just “bigger” than model M_2 , which excludes some of the inputs used by M_1 . M_2 is **nested** within M_1 .
- Another way of seeing that: both have the same inputs, but M_2 has some parameters β set to zero.
 - This is a constraint. It can be formulated as a null hypothesis.

Testing the Statistical Significance of Inputs

- H_0 is “ M_2 is correct”. We can think of M_1 as an **alternative hypothesis** H_1 instead of the saturated model.
- The ratio between the optimised likelihood for M_1 and the optimised likelihood for M_2 tells us some information about whether we lose by using M_2 .
- This type of testing is known as a **likelihood ratio test**.

Testing the Statistical Significance of Inputs

- With a difference of k parameters, the distribution of (twice) the log-likelihood ratio is

$$D \equiv 2[l(\hat{\beta}_{M_1}) - l(\hat{\beta}_{M_2})] \quad D \sim \chi_k^2$$

which is just the analogous idea of taking the difference of the number of inputs to calculate the parameter of the chi-squared.

Testing the Statistical Significance of Inputs

- Why do this? Shouldn't I get good evidence on whether M_1 predicts better than M_2 by cross-validation?
 - Yes, but that depends on your prediction measure. 0/1 classification for instance. This will not tell you about the probability estimates.
 - You could do cross-validation of predictive log-likelihood, and then test whether the difference is significant. This is silly when we have an option (the likelihood ratio test) that does not require cross-validation.
 - On the other hand, all this is saying is whether we should reject M_2 or not. *Again, statistical significance ≠ practical significance.* Maybe the gains of M_1 do not justify the possible cost of measuring extra inputs, for instance.

Informal Understanding of Deviance

$$D \equiv 2[l(\hat{\beta}_{sat}) - l(\hat{\beta})] \quad D \sim \chi^2_{n-p}$$

- Lower deviance means better fit. But:
 - Adding a useless predictor (independent of output) would on expectation decrease deviance by one unit
 - since p increases by 1, but fit doesn't get better.
 - When adding k informative predictors, we expect deviance to decrease by more than k on average.

Example

- R demo, back to the “Default” example.
 - Note: R doesn’t have a standard specialized plotting procedure for logistic regression
- Common summaries found in software packages:
 - **Null deviance**: deviance for the model without any inputs (empirical probability of each class)
 - **Residual deviance**: deviance for the model with the given inputs
 - **Coefficient p-values** are typically based on the approximate Gaussian distribution of MLE estimates, which we won’t discuss in detail – but the rationale is the same as always.

BEYOND LOGISTIC REGRESSION

Poisson Models

- The Poisson distribution is a distribution over natural numbers. The modelling of **count data** is one of its main applications. For $\theta > 0$,

$$Y \sim Poisson(\theta)$$

$$P(Y = y) = \frac{\theta^y \exp^{-\theta}}{y!}, \quad y \in \{0, 1, 2, 3, \dots\}$$

Facts about the Poisson

- If $Y \sim \text{Poisson}(\theta)$, then

$$E[Y] = \text{Var}(Y) = \theta$$

- Hence, a very constrained distribution. But a very important building block.
- Physical motivation:
 - think about partitioning a line segment of length θ very finely in n regions of equal length.
 - Now, do independent coin flips at each region with probability θ/n of “success”. In the limit $n \rightarrow \infty$, the total number of successful coin flips will follow a $\text{Poisson}(\theta)$.

Generalised Linear Models (Again)

- Idea: we do a “two-stage” model.
- Use linear predictor as a meta-parameter: transform it to get the parameter(s) of the target distribution.
- We will need a **link function**.

Generalized Linear Models for Poisson Regression

$$P(Y = y) = \frac{\theta^y \exp^{-\theta}}{y!}, \quad y \in \{0, 1, 2, 3, \dots\}$$

- Say you have a single data point, linear predictor is

$$\eta^{(i)} \equiv \sum_{j=1}^p x_j^{(i)} \beta_j$$

- We want to model the probability of $Y^{(i)}$ taking value 1 according to the information in $x^{(i)}$.

$$\eta^{(i)} = g(\theta^{(i)})$$

Link function

Generalized Linear Models for Poisson Regression

- We only need to satisfy $\theta > 0$.
- The most common link function (the default in many packages) is the logarithm function.

$$Y^{(i)} \sim Poisson(\theta^{(i)})$$

$$\theta^{(i)} \equiv \exp \left(\sum_{j=1}^p x_j^{(i)} \beta_j \right)$$

Interpretation

- The logarithm link function gives a mapping between the linear response and the log-conditional expected value of Y .

$$\log E[Y|X = x] = \log \theta = \sum_{j=1}^p x_j \beta_j$$

- So each coefficient β_j is the rate of change of the expected value of the outcome on log-scale per unit of X_j , fixing everything else.

Example

- Oyster card data
 - From <https://blog.tfl.gov.uk/2015/12/09/is-customer-flow-data-useful-to-developers/>
- Exit counts according to position
 - “Position” is measured as Euclidean distance to King’s Cross by latitude/longitude
- Non-linearity evident, we will use polynomial model with Poisson likelihood.

Analysing Fitness

- Is the Poisson mean/variance relationship valid?
- We can plot for each prediction its fitted interval
 - That is, plugging in parameter estimates and plotting (e.g.) 95% intervals centred around fitted value
- R demo

Overdispersion

- When many more points than expected lie outside the interval, it could be the result of a bad fit of the mean.
- But also, the result of a bad choice of likelihood function. It is common to find count data where variance is higher than the mean. This is called **overdispersion**.
 - Can you think of reasons for that in the Tube data?

Alternatives

- This is a good point to illustrate how possibly over-simplified distributions like the Poisson are powerful building blocks.
- Consider a **mixture** of Poissons: parameter θ being itself random, followed by the usual Poisson.
- The resulting distribution of the data can be a very flexible distribution over counts.

The Negative Binomial

- Consider this particular mixture:

$$\theta \sim \text{Gamma}(r, p/(1 - p))$$

$$Y \sim \text{Poisson}(\theta)$$

- One interpretation is that there are (infinitely many) hidden populations generating counts, and we only observe their aggregation.

The Negative Binomial

- Recall the following fact about probabilities:

$$\sum_b P(A = a \mid B = b)P(B = b) = P(A = a)$$

Since $P(A = a \mid B = b)P(B = b)$ is just another way of writing $P(A = a, B = b)$.

- The result of this

$$P(Y = y) = \int P(Y = y \mid \theta)p(\theta) d\theta$$

is a (r, p) **negative binomial** distribution.

$$Y \sim \text{NegativeBinomial}(r, p)$$

Negative Binomial

- There are other interpretations, but the mixture of Poissons point of view emphasizes this is a more flexible model for counts.
- In particular,

$$E[Y] = \frac{pr}{1-p} \quad Var(Y) = \frac{pr}{(1-p)^2}$$

- Notice we can write the model in terms of its mean and variance. Some packages offer both ways of writing a negative binomial distribution.

Negative Binomial Regression

- Notice the twist: there is more than one model parameter. How are linear responses used?
- This is not the first time we come across with this: what did we do in the Gaussian case?
 - Varying means, constant variances (homoscedastic model).
 - This is typically what we consider a generalized linear model: a “location” parameter being a transformation of the linear response, a “dispersion” parameter being independent of the inputs.
 - We can of course define a model where “dispersion” depends also on the inputs, but this would not be a GLM. A story for another course (for the sake of curiosity, in the machine learning literature, a classical model like that is the *mixture of experts*).
 - Strictly speaking, the negative binomial is not a “true” GLM as defined in the literature, for reasons we will very briefly explain later.

Negative Binomial Regression

- For an example, the following is based on the implementation in package MASS, from R:

$$\eta^{(i)} \equiv \sum_{j=1}^p x_j^{(i)} \beta_j \quad \mu^{(i)} = \exp(\eta^{(i)})$$

$$Y \sim NegativeBinomial_{alt}(\mu^{(i)}, v)$$

which is an alternative parameterization where

$$E[Y^{(i)}] = \mu^{(i)} \quad Var(Y^{(i)}) = \mu^{(i)} + \frac{\mu^{(i)2}}{v}$$

- Let's revisit our Tube data with a negative binomial regression model (R demo).

One Final Example: Ordinal Data

- Ordinal data is discrete and doesn't have numerical value, but it is ordered.
- Very common in surveys ("Strongly disagree", ... "Strongly agree").
 - Notice how magnitude of difference is not evident.
- Non-numerical ordering can be encoded in model.

Ordered Logit Model

- For outcome Y with arbitrary levels $1, 2, \dots, K$

$$\begin{aligned} P(Y^{(i)} > 1) &= \text{logistic}(\eta^{(i)}) \\ P(Y^{(i)} > 2) &= \text{logistic}(\eta^{(i)} - c_2) \\ P(Y^{(i)} > 3) &= \text{logistic}(\eta^{(i)} - c_3) \\ &\dots \\ P(Y^{(i)} > K-1) &= \text{logistic}(\eta^{(i)} - c_{K-1}) \end{aligned}$$

besides parameters in linear responses, we have also (non-increasing) **threshold parameters** c_2, \dots, c_{K-1} .

(R demo)

Ordered Logit Model

- Thresholds are forced to be monotone.
- Probabilities can be easily obtained out of the given expressions:

$$P(Y = k) = P(Y > k - 1) - P(Y > k)$$

where here $P(Y > 0) = 1$ and $P(Y > K) = 0$.

Latent Variable Interpretation

- Similar to the interpretation in binary logistic regression: $Z^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)} + \epsilon^{(i)}$

$$y^{(i)} = \begin{cases} 1, & \text{if } z^{(i)} < 0 \\ 2, & \text{if } z^{(i)} \in (0, c_2) \\ 3, & \text{if } z^{(i)} \in (c_2, c_3) \\ \dots \\ K-1, & \text{if } z^{(i)} \in (c_{K-2}, c_{K-1}) \\ K, & \text{if } z^{(i)} > c_{K-1} \end{cases}$$

This is a common interpretation, e.g. with the latent variable representing a “position” regarding the observable preference Y .

BRIEF MATHEMATICAL COMMENTS

NOT EXAMINABLE. FOR YOUR REFERENCE AND INFORMATION.

The Abstract Formulation

- A generic way of writing down GLMs.
- I mention this as it will show up in the documentation of software packages.
 - Also, other *STAT* modules, including *STATG003*

$$p(y; \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

The above is also known as a type of **exponential (dispersion) family**.

η ? ϕ ? a ? b ? c ?

- ϕ is a “dispersion parameter” that does not depend on inputs. We hinted at it before when we mentioned Gaussian regression as a GLM.
 - $\phi = 1$ in many models, like the logistic and Poisson.
- The rest are functions that will depend on the family.

Also

- *To be more in line with notation in other textbooks*, let's focus on the standard case where the linear predictor is used to model the *mean* of the distribution. This is the “textbook” GLM:

$$\eta^{(i)} \equiv \sum_{j=1}^p x_j^{(i)} \beta_j \quad \eta^{(i)} = g(\mu^{(i)})$$

$$P(Y = y) = \mu^y (1 - \mu)^{1-y}, \quad y \in \{0, 1\}$$

- So, instead of θ as we used before, we will use μ , and θ instead will refer to **the parameter in the actual exponential family representation**.

Where are We Going with This?

- At a more mature level as a Data Scientist, it pays off to recognize commonalities among models.
- This allows for distributions of statistics, and algorithms for parameter fitting, to be written in an unifying way.
- GLMs are a famous success story of this line of thought, and it is illuminating to understand this.

Example

- Rewriting the Gaussian according to the GLM template: find me θ, ϕ, a, b, c .

$$p(y) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(y - \mu)^2}{2\sigma^2} \right\}$$

$$p(y; \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

Example

- What about:

$$\theta = \mu, \phi = \sigma^2$$

$$c(y, \phi) = -(\log(2\pi) + y^2/\sigma^2)/2$$

$$a(\phi) = \sigma^2$$

$$b(\theta) = \theta^2/2$$

$$p(y) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(y - \mu)^2}{2\sigma^2} \right\}$$

$$p(y; \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

Another Example

- Rewriting Bernoulli regression according to the GLM template: find me θ, ϕ, a, b, c .

$$P(y) = \mu^y(1 - \mu)^{1-y}$$

$$p(y; \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

Another Example

- What about:

$$\eta = \log \left(\frac{\theta}{1 - \theta} \right), \phi = 1$$

$$c(y, \phi) = 0$$

$$a(\phi) = 1$$

$$b(\eta) = \log(1 - \theta) = -\log(1 + e^\eta)$$

$$P(y) = \mu^y (1 - \mu)^{1-y}$$

$$p(y; \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

Inference

- Previously we discussed the following result for logistic regression concerning deviance:

$$D \equiv 2[l(\hat{\beta}_{sat}) - l(\hat{\beta})] \quad D \sim \chi^2_{n-p}$$

- For the more general GLM, we need to take into account inference for ϕ .

Inference

- The definition of deviance is slightly different:

$$D \equiv 2[l(\hat{\beta}_{sat}) - l(\hat{\beta})]\phi$$

- Without going into details, the distribution is not chi-squared anymore if we estimate ϕ
 - We can use a ratio of deviances to make ϕ disappear. The ratio follows particular type of *F distribution*, if you must know.

Inference

- Confidence intervals can be obtained by CLT

$$\hat{\beta} \sim N(\beta, \mathcal{I}^{-1}(\beta))$$

which is a multivariate Gaussian distribution
(more on that in Chapter 6).

- Each marginal follows your well-known univariate Gaussian distribution.
- Matrix $\mathcal{I}(\beta)$ is sometimes called, for your information, *the Fisher Information Matrix*.

Inference

- The (j, k) entry of the Fisher Information Matrix is given by

$$\mathcal{I}_{jk}(\beta) = -E \left[\frac{\partial l}{\partial \beta_j} \frac{\partial l}{\partial \beta_k} \right]$$

- Notice that l is random since it depends on the data distribution of the outcome variable.
- **Why am I telling you this?** You might come across this in other modules, textbooks or even software package documentations. It is a general result in Statistics that is used beyond GLMs: CLT approximations to the sampling distributions of maximum likelihood parameter estimates.

Optimisation

- We will spend just a little more time now understanding how maximum likelihood works for GLMs.
- As a Data Scientist, it is helpful to understand at least one of these less straightforward case studies for parameter fitting (“learning”, in machine learning lingo).

Optimisation

- An optimisation problem has an **objective function** and **(decision) variables**
 - don't confuse these with random variables in a probabilistic model.
- The goal is to find variable assignments that maximise/minimise the objective function.
- In a GLM, the objective function is the log-likelihood function, the decision variables are the parameters of the linear responses, and perhaps the dispersion parameter.

Newton-Raphson

- To find the zeroes of a function
 - Useful in optimisation, as finding the zeroes of a derivative provide candidate optima
- One-dimensional case: iterate

$$x_{(i)} \leftarrow x_{(i-1)} - \frac{f(x_{(i-1)})}{f'(x_{(i-1)})}$$

where $f(x)$ is the first derivative of the function $g(x)$ we want to optimise, whatever x is.

- R demo:
 - Imagine original function is a cubic polynomial
 - So we will try to find zeroes of a quadratic polynomial

Multi-Dimensional Case

- Expressed as the original function, we need first and second derivatives. **Hessians** and **gradients** in the multi-dimensional case:

$$\hat{\beta}_{(i)} \leftarrow \hat{\beta}_{(i-1)} - \mathbf{H}(\hat{\beta}_{(i-1)})\mathbf{h}(\hat{\beta}_{(i-1)})$$

$$\mathbf{H}(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_p} \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_p} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_p} & \cdots & \frac{\partial^2 f(x)}{\partial x_p^2} \end{bmatrix} \quad \mathbf{h}(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_p} \end{bmatrix}$$

Expensive to compute in high-dimensions

A Statistician's “Hack”

- “ $g(x)$ ” in our case is the log-likelihood function of β , that is, $I(\beta)$.
- Statisticians love to substitute the Hessian by its expected value, as it is numerically more stable. Notice that

$$E[\mathbf{H}(\hat{\beta})] = -\mathcal{I}(\hat{\beta})$$

- So the expression for $\mathcal{I}(\hat{\beta})$ (which I will not show, but *STATG003* students will have fun with in Term 2) can be plugged in the Newton update equation.

Take-Home Messages

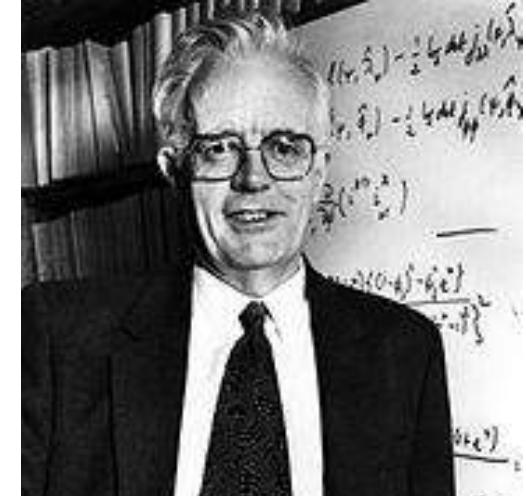
- We saw probabilistic modelling from a regression perspective.
 - Models with full likelihood functions. Contrast this to likelihood-free Statistics, or Empirical Risk Minimisation in Machine Learning.
- GLMs provide a good sandbox to define models, which by the end of the day have commonalities (same fitting algorithms, confidence intervals, etc.)

Take-Home Messages

- Shortcomings:
 - Still linear in some important sense...
 - We can't really calculate Hessians in high-dimensional problems...
 - Variance of estimates can be very high in high-dimensional problems...
- Solutions in the next Chapter:
nonparametrics, sparse models and more on model selection.

Historical Note

- Logistic regression was formalised by David Cox in the 1950s.
- He is still active these days, and you might still be able to sometimes spot him coming to give a talk at the London School of Hygiene and Tropical Medicine around the corner.



Wikimedia Commons

Journal of the Royal Statistical Society

SERIES B (METHODOLOGICAL)

Vol. XX, No. 2, 1958

THE REGRESSION ANALYSIS OF BINARY SEQUENCES

By D. R. COX

Birkbeck College, University of London

[Read before the RESEARCH SECTION of the ROYAL STATISTICAL SOCIETY, March 5th, 1958, Professor G. A. BARNARD in the Chair]

SUMMARY

A SEQUENCE of 0's and 1's is observed and it is suspected that the chance that a particular trial is a 1 depends on the value of one or more independent variables. Tests and estimates for such situations are considered, dealing first with problems in which the independent variable is preassigned and then with independent variables that are functions of the sequence. There is a considerable amount of earlier work, which is reviewed.

Introduction to Statistical Data Science

Ricardo Silva

ricardo@stats.ucl.ac.uk

Department of Statistical Science, UCL

Selected Topics in Modern Regression Practice

Outline

- This is a selection of regression methods which I believe is the core of practical, applied statistics beyond the widespread linear/GLM framework.
 - Not an exhaustive list, see e.g. *STATG011* (Forecasting)
- Like the exposure in linear/GLMs, this is from the point of view of not only providing black-boxes, but also understanding their properties.
- Little of what follows is cutting-edge research in Statistics itself. Nevertheless, these ideas are fundamental enough that still form the basis of much on-going methodological research.

Outline

- Model selection in regression
- High-dimensional regression
- Non-linear models and the generalised additive model (GAM)

Selected Topics

MODEL SELECTION IN REGRESSION

Model Selection, and Motivation

- Recall the linear model

$$Y = \beta_0 + \beta_1 x_1 + \dots \beta_p x_p + \epsilon$$

- We might be interested in
 - Improving **prediction**: are my estimates unreliable to the point that I could actually do better by removing some of the covariates?
 - Improving **interpretability**: does it make sense to include covariates for which I have no evidence that they contribute to explaining Y ?

Sparsity

- A model is sparse if many of its components are “inactive”. In our context, this will mean several regression coefficients are zero.
- We do not need to believe that Nature is sparse in order to agree that there are advantages on having sparse *estimates*.

Strategies

- **Combinatorial search** (a.k.a subset selection): search among possible subsets of inputs, fit final model based on the chosen subset.
- **Shrinkage** (a.k.a. regularization): change your fitting criterion to “incentivize” parameters to be zero, or near-zero.
 - We will exploit this more in the context of high-dimensional regression.
- **Dimensionality reduction**: transform your inputs to a smaller set of features, regress on them. More on that in Chapter 6 (see also, Supervised Learning/Neural Networks)

Combinatorial Search

- Core idea: given your original set of p predictors, propose several subsets of it.
 - Concern: how many subsets are out there?
- Score: for which subset, provide a quantification of how good/bad they are.
 - Concern: take the residual sums of squares (RSS). Take two sets S_1 and S_2 such that $S_1 \subset S_2$. Which one is best according to the RSS?

Combinatorial Search

- For the computation, we will need **heuristics**, ways of cutting down the search
 - No guarantees of optimality, unless restrictive assumptions are used.
- For the scoring, we will follow this general idea (written as functions to be *minimised*):

$$Score = -fitness + complexity \text{ penalty}$$

Increases by how much residuals
are minimised

Increases by how many parameters are used

The Best Subset Selection Algorithm

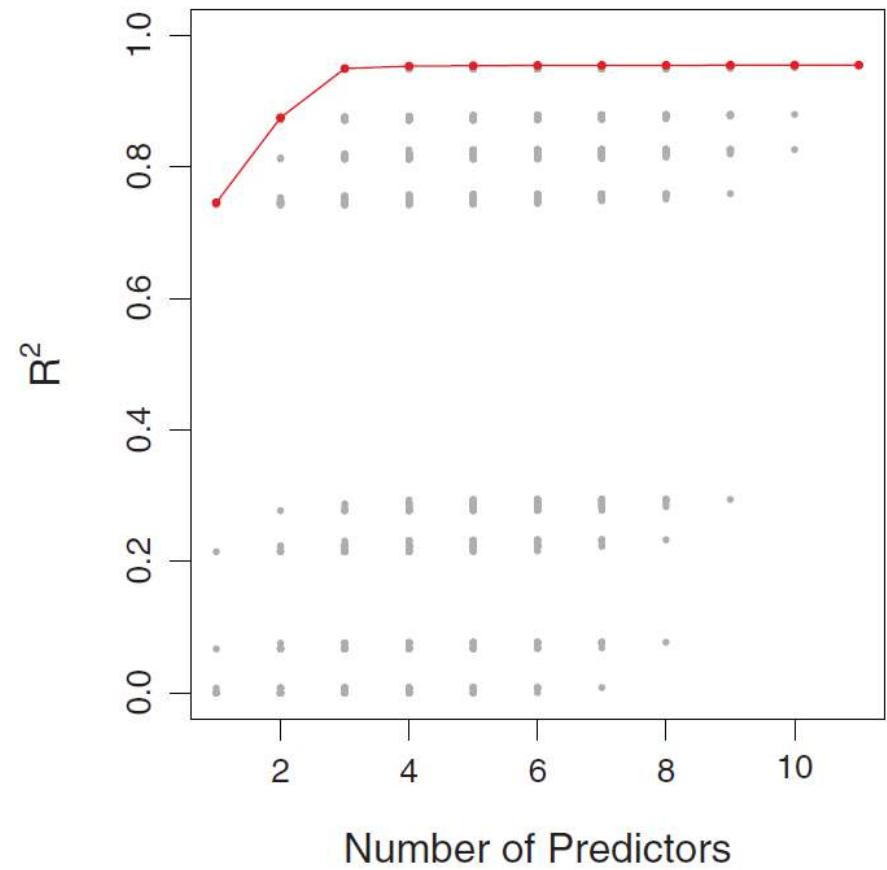
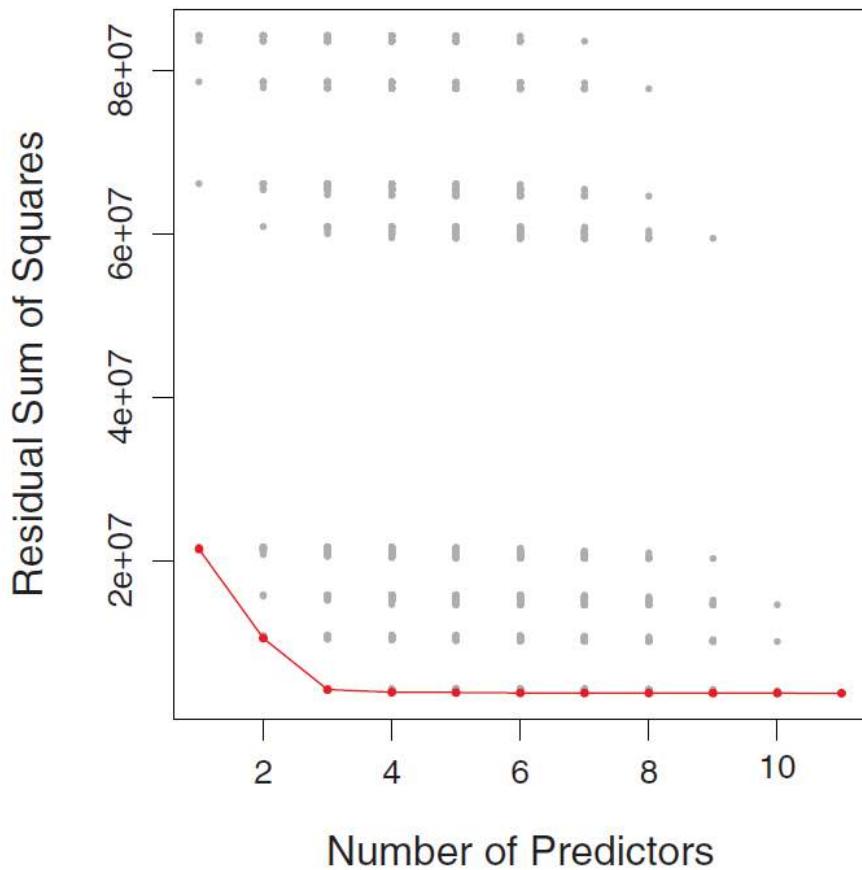
Algorithm 6.1 *Best subset selection*

1. Let \mathcal{M}_0 denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
 2. For $k = 1, 2, \dots, p$:
 - (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
 - (b) Pick the best among these $\binom{p}{k}$ models, and call it \mathcal{M}_k . Here *best* is defined as having the smallest RSS, or equivalently largest R^2 .
 3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .
-

Illustration

- *Credit* dataset: modelling balance.

(ISLR, Chapter 6)



The Different Criteria

- Cross-validation:
 - Split your data in K disjoint subsets
 - For every data subset D_k , where $k = 1, 2, \dots, K$, use the remaining data for fitting (training) the model, and assess the RSS on the “test set” D_k .
 - Report as a score the model average D_k across folds
- Notice that the penalty is implicit here

Cross-Validation

- Illustration of leave-one-out CV (LOOCV)



Mallow's C_p

Training RSS with input set S

Estimate of error variance using all covariates

$$C_p(S) \equiv \frac{1}{n} (RSS(S) + 2|S|\hat{\sigma}^2)$$

- Provides an estimate of test error.
 - In practice, for linear regression it provides a similar estimate as cross-validation, without the need to split the data
 - Unlike cross-validation, not easily adaptable outside linear regression.

AIC/BIC/*IC

- There is a whole industry of defining penalized fitness using likelihoods: “information criteria”.
 - In our context, we use Gaussianity assumptions for the error terms.
- Unlike cross-validation and C_p , they focus on model fitness (as given by the likelihood) instead of prediction as pre-specified by a cost function, like least-squares.
 - Sometimes these measures agree, but not always.
 - If the main goal is prediction, cross-validation is in general more appealing, but it is expensive and more unstable (as it depends on the data split)

AIC

- The AIC (Akaike Information Criterion):

$$AIC(S) \equiv -l(S) + |S|$$

 Log likelihood using all data

(Please notice that in some books, you might find different by equivalent definitions)

- It is not hard to show that, for linear Gaussian models, AIC and Mallow's C_p give exactly the same ranking of models.

BIC (Bayesian Information Criterion)

$$BIC(S) = -l(S) + \frac{|S|}{2} \log(n)$$

- Which to use? Both methods start from different assumptions, which are technical and won't be discussed.
- In practice, for linear models fit by least-squares or MLE, any of these will do reasonably OK. BIC adds a stronger penalty than AIC / C_p , so it will generate smaller models. AIC might still get better out-of-sample prediction errors.

Greedy Search

- In combinatorial optimisation, a common heuristic is **greedy search**:
 1. Start with a (possibly random) initial candidate model
 2. Look at its “neighbours”. If any of the neighbours is better than the current candidate, make it the current candidate. Else, stop
 3. Return to Step 2

Greedy Search

- A “starting model” can be the model with the intercept only, or a model with all covariates.
- In the context of model selection for regression models, a “neighbour” is a model that differs from the current one by having one more or one fewer covariate.
- Even simpler, depending on the starting point, we may decide only to add, or only to exclude, a covariate to/from the current candidate.
- This is commonly known as **stepwise selection**.

Forward Stepwise Selection

Algorithm 6.2 *Forward stepwise selection*

1. Let \mathcal{M}_0 denote the *null* model, which contains no predictors.
 2. For $k = 0, \dots, p - 1$:
 - (a) Consider all $p - k$ models that augment the predictors in \mathcal{M}_k with one additional predictor.
 - (b) Choose the *best* among these $p - k$ models, and call it \mathcal{M}_{k+1} . Here *best* is defined as having smallest RSS or highest R^2 .
 3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .
-

“Local” Minimum

- Just like as in standard (continuous) optimisation, it is possible to get trapped into a “local” minimum in a combinatorial optimisation problem.
 - “Local” here is defined with respect to the neighbourhood definition.
- Notice one aspect of the particular stepwise approach from the previous slide: its stopping criterion is “follow a path until no more variables can be added. Only then decide on the best.”

Computational Complexity

- Recall: original exhaustive search (also known as a **brute-force** approach) required the evaluation of 2^p models.
- The search space of stepwise selection here required $O(p^2)$ evaluations (why?).



(the “O” notation roughly means “something proportional to p^2 ”, up to multiplicative and additive constants)

Illustration

- *Credit data*
 - (also, R demo)

# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income, student, limit	rating, income, student, limit

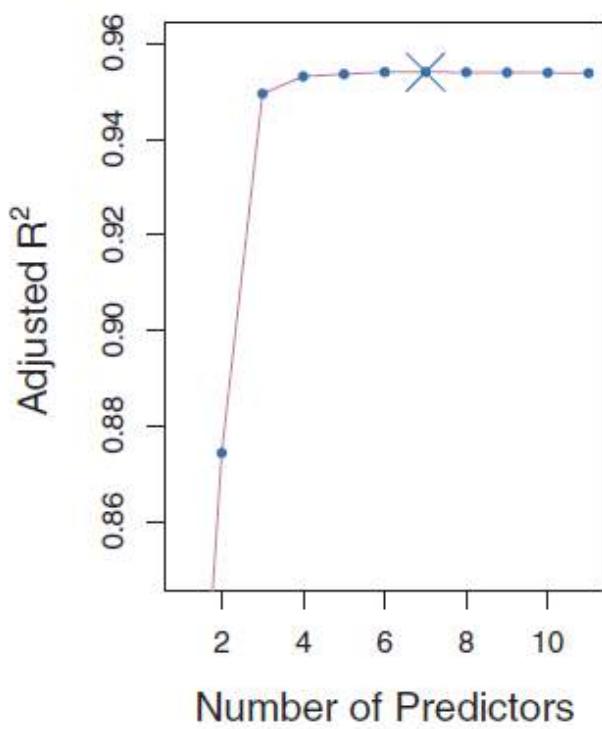
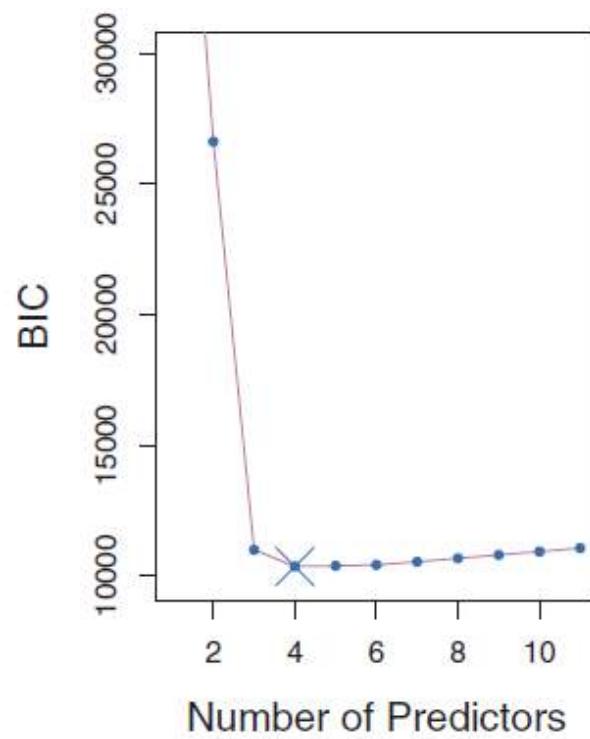
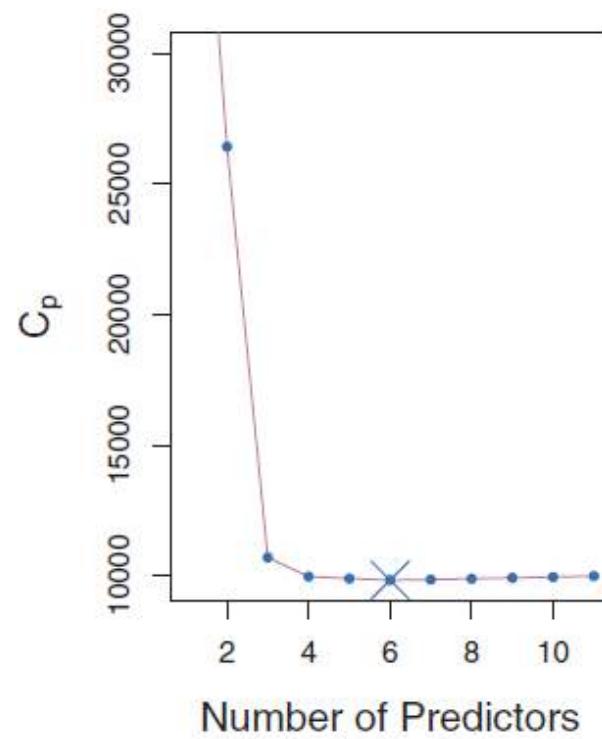
Backward Selection

Algorithm 6.3 Backward stepwise selection

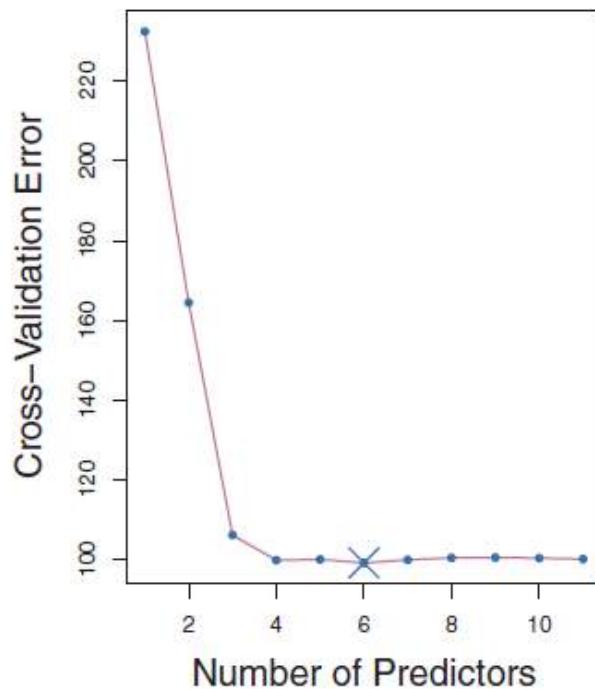
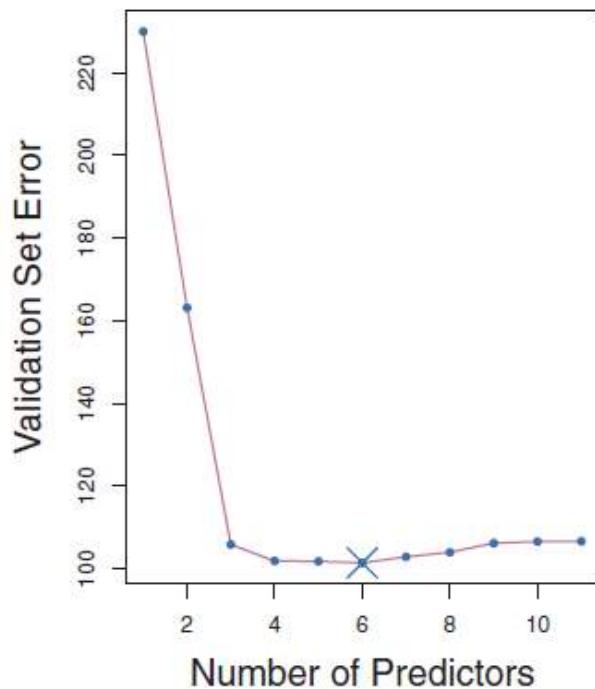
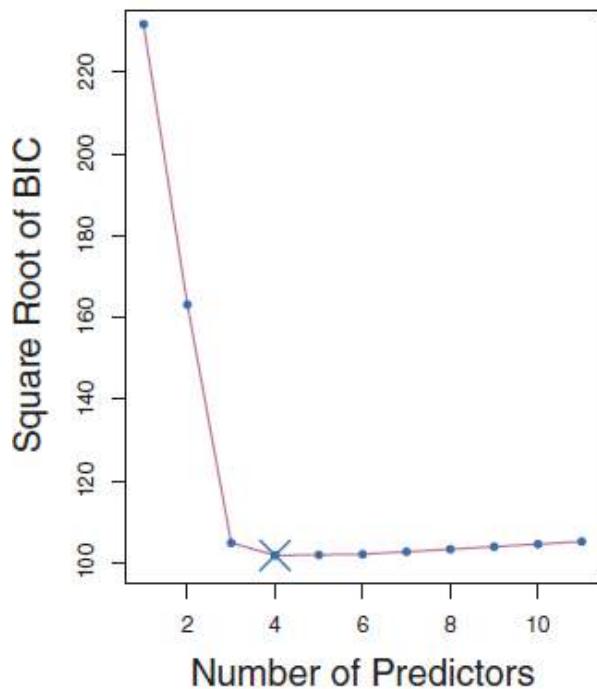
1. Let \mathcal{M}_p denote the *full* model, which contains all p predictors.
 2. For $k = p, p-1, \dots, 1$:
 - (a) Consider all k models that contain all but one of the predictors in \mathcal{M}_k , for a total of $k-1$ predictors.
 - (b) Choose the *best* among these k models, and call it \mathcal{M}_{k-1} . Here *best* is defined as having smallest RSS or highest R^2 .
 3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .
-

(R demo: different starting points)

Different Criteria



Held-out Data Criteria



Rules of Thumb

- Remember: **results in a test set are statistical estimates, not statements of fact about future out-of-sample behaviour.**
- The “one-standard-error rule” looks at the simplest model with an cross-validated error within one (estimated) standard deviation of the error given by the “optimal” model.
 - Like the “5 second rule for dropped food”, take it with some knowledge that it is partially folk knowledge, and not a hard guideline.

Questions

- Once I've selected the model, can I interpret the p-values of the coefficients as before?
 - No, for the same reason that training error is not a good estimate of out-of-sample performance.
 - You could evaluate p-values in a separate test data, although your sample size will be reduced, and as a consequence so will your power.
- Do I have theoretical guarantees of optimal variable selection (assuming "true sparsity" holds)?
 - Up to some point (conditions apply), if you were able to compute the optimal solution to the combinatorial search problem. But in general this is not possible. Sub-optimal selection can still be achieved.
- Do I have theoretical guarantees of improved test set error?
 - Again, up to some point and with fewer assumptions (more so with cross-validation than with other methods). It might not be optimal, but you can still get an improvement.

Final Note: The p-Value Route

- Why not just drop the covariates with a corresponding low p-value (maybe with some Bonferroni correction)?
- Recall the shortcoming: each coefficient β_i is what we learn about the contribution of x_i , **when we fix all other covariates**.
- When we drop *one* covariate, it is possible that previously non-significant coefficients now become significant.

Final Note: The p-Value Route

- As a heuristic, it might be a cheap alternative (R demo: see what happens with the *Credit* data), but it also has no guarantees of optimality (without further assumptions).
 - And it is even less clear what it would mean in terms of prediction quality.
- What about using it as the greedy search criterion?
 - Possible, but how to set the threshold of significance? The search path will destroy its interpretation as a Type I error control.

Final Note: The p-Value Route

- What about treating the threshold as a “free parameter”?
- Q. How to choose it?
 - A. What about we use cross-validation itself?
- Mmm... Details are unclear up to now, but this intriguing. What about ignoring p-values and think more generally about non-combinatorial ways of making models simpler?

Selected Topics

SHRINKAGE AND MODEL COMPLEXITY

Regression with Limited Data

- As the number p of covariates increases, less reliable greedy search becomes. As a matter of fact, even least-squares gets less reliable (meaning that confidence intervals get wider and wider).
- Also, in some applications we may even have situations where $p > n$. Least-squares is not even defined in that case.
- Let's tackle these issues using the sledgehammer of regularization, or **shrinkage**.

Complexity as “Parameter Size”

- Consider this modified objective function for least-squares regression: ***ridge regression***

$$\sum_{i=1}^n \left(y^{(i)} - \beta_0 - \sum_{j=1}^p \beta_j x_j^{(i)} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

Our old friend, the RSS:
measure of fit

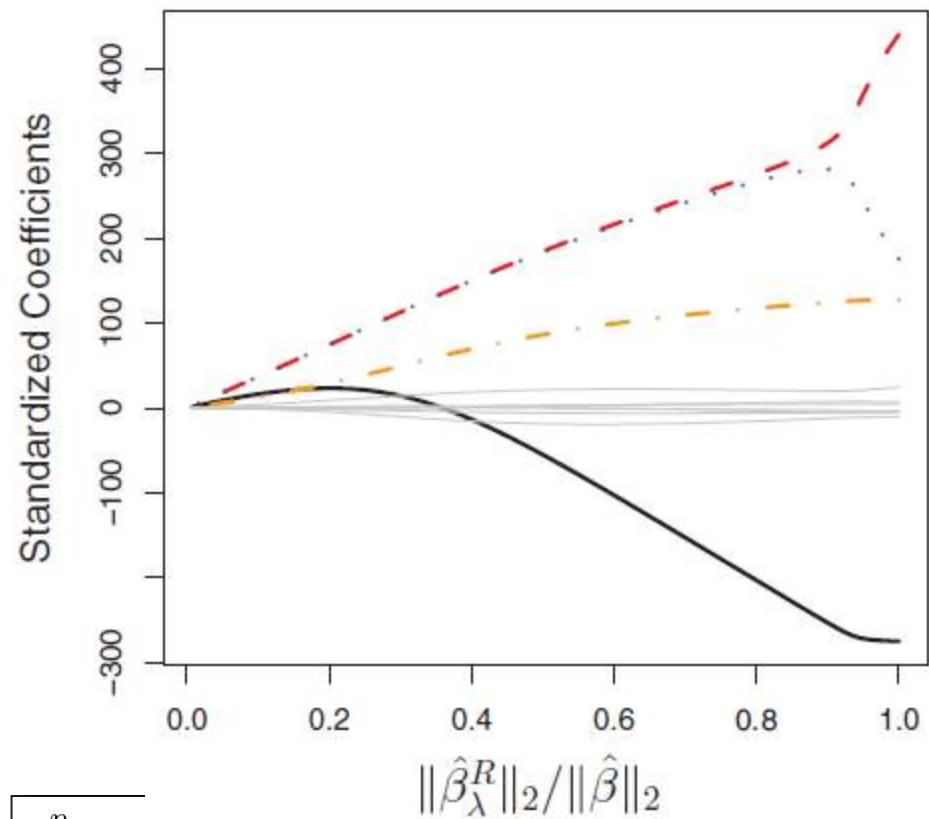
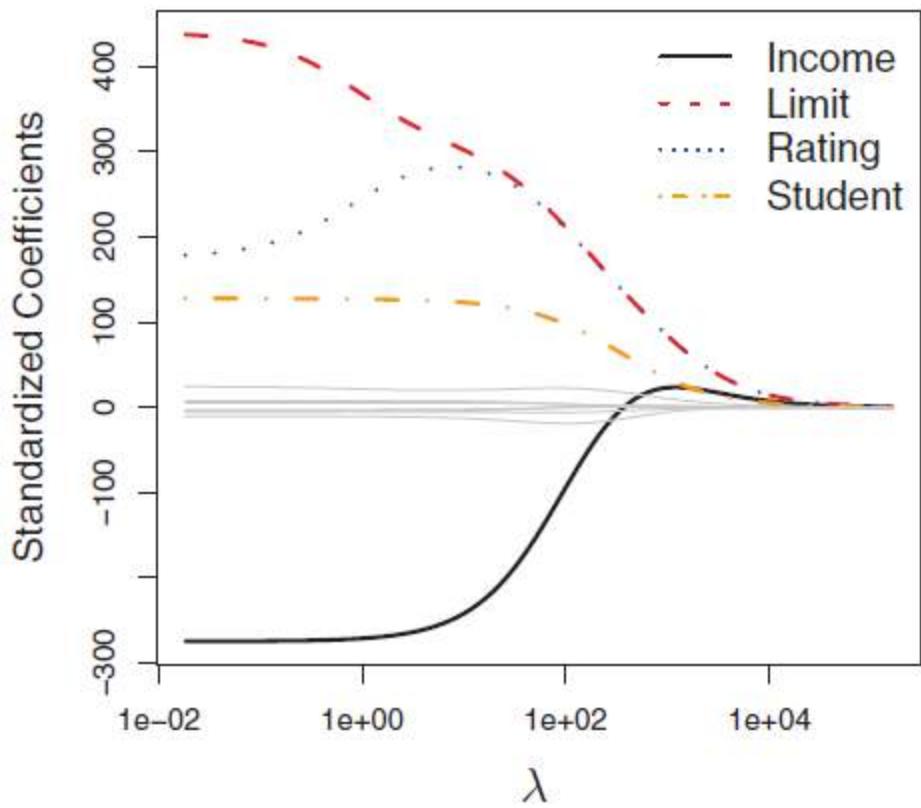
Increases with “size” of
parameters

Penalty term:
to be set by cross-validation

Notice: no penalty for the intercept

(R demo: size vs complexity)

Illustration: *Credit* dataset



$$\|\beta\|_2 \equiv \sqrt{\sum_{j=1}^p \beta_j^2}$$

IMPORTANT!

- Unlike unpenalized least-squares regression, regression with shrinkage is **scale-dependent**.
- Running penalized least-squares without considering the format of your data might give you suboptimal results.
- Consider standardization:

$$\tilde{x}_j^{(i)} \equiv \frac{x_j^{(i)}}{S_j}, \text{ where } S_j = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_j^{(i)} - \bar{x}_j)^2}$$

R demo: dangers of ignoring data scale in ridge regression.

More Intuition

- The **bias-variance trade-off**, mean-squared error style:

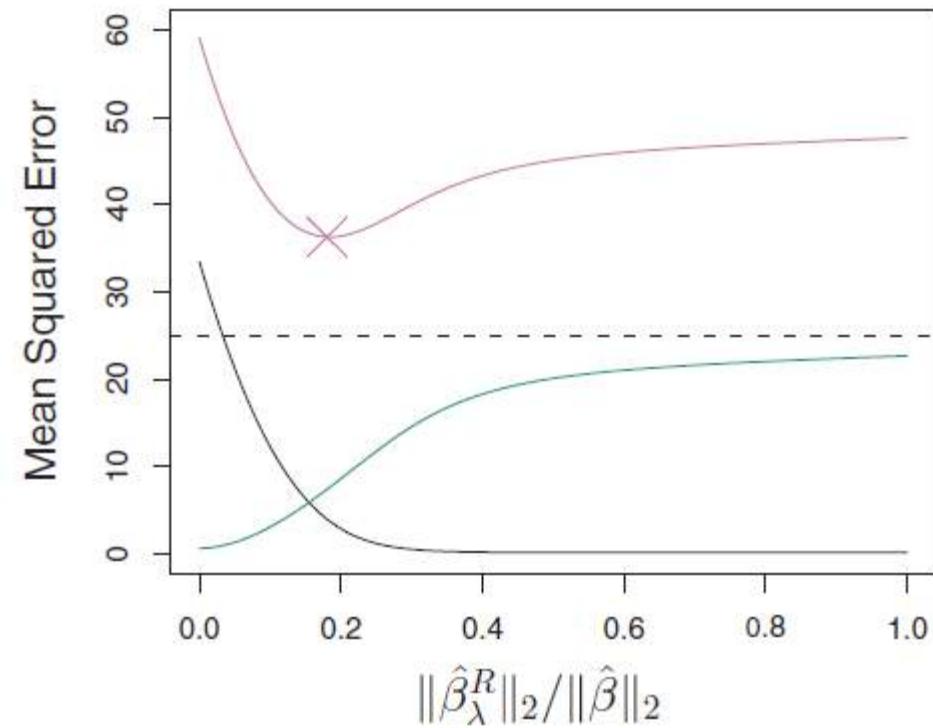
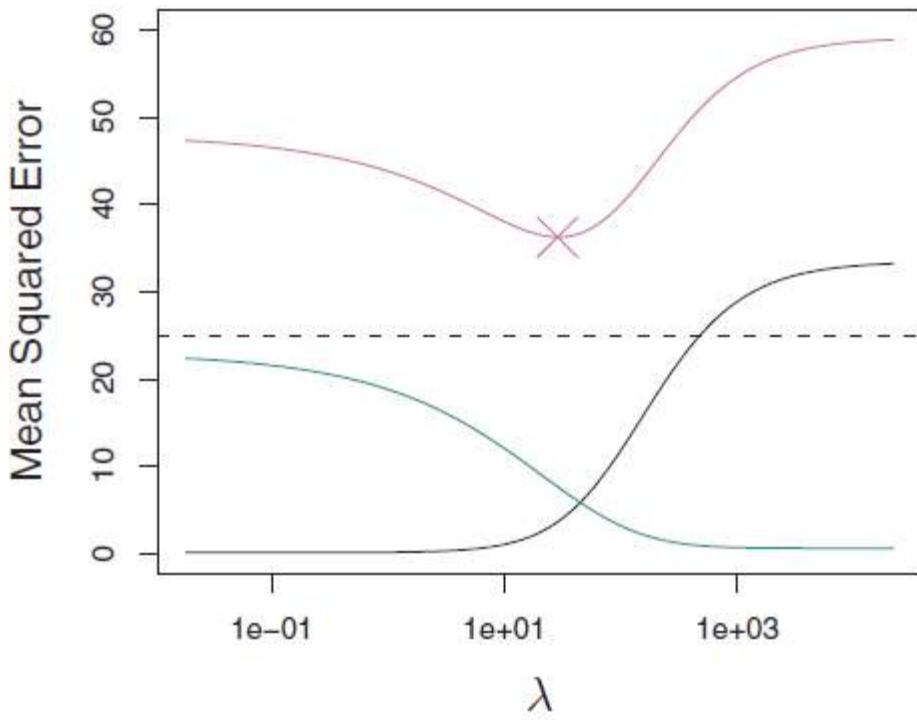
$$\begin{aligned}
 MSE &= E[(\theta - \hat{\theta})^2] = \\
 &= E[\theta^2] - 2E[\theta\hat{\theta}] + E[\hat{\theta}^2] \\
 &= \theta^2 - 2\theta E[\hat{\theta}] + E[\hat{\theta}^2] \\
 &= \theta^2 - 2\theta E[\hat{\theta}] + \textcolor{red}{E[\hat{\theta}]^2 - E[\hat{\theta}]^2} + E[\hat{\theta}^2] \\
 &= (\theta - E[\hat{\theta}])^2 + E(\hat{\theta} - E[\hat{\theta}])^2
 \end{aligned}$$

A blue arrow points from the term $(\theta - E[\hat{\theta}])^2$ to the label "bias, squared". Another blue arrow points from the term $E(\hat{\theta} - E[\hat{\theta}])^2$ to the label "variance".

- We can interpret the penalty term as a **bias**, pushing coefficients towards zero. The effect is a reduction on **variance**.

Example: *Credit* data

- Squared bias in black, variance in green, MSE by cross-validation.



Norms as Penalties

- The ridge regression penalty is known as the (square of) **L_2 norm** (“ell 2”) penalty.

$$\sum_{i=1}^n \left(y^{(i)} - \beta_0 - \sum_{j=1}^p \beta_j x_j^{(i)} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

- Why do we use it? Among other things, it is differentiable. So, computationally convenient.
 - If you know what convex optimisation is, the above is also convex. The upshot? Unlike subset selection, all local minima agree on the same global minima!

The l_0 penalty

- Subset selection can be seen as optimising this:

$$\sum_{i=1}^n \left(y^{(i)} - \beta_0 - \sum_{j=1}^p \beta_j x_j^{(i)} \right)^2 + \lambda \sum_{j=1}^p I(\beta_j = 0)$$

where $I(z) = 1$ if z is true, 0 otherwise. This is the same as counting non-zero parameters.

- l_0 is in one sense ideal, but computationally nasty, as we saw before. Definitely *not* differentiable, hence the combinatorial search.

The Lasso

- Motivated as a “relaxation” of the l_0 norm.

$$\sum_{i=1}^n \left(y^{(i)} - \beta_0 - \sum_{j=1}^p \beta_j x_j^{(i)} \right)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

- Although not differentiable at zero, this is still a convex function and a variety of algorithms (some gradient-based) can be used to optimise it.

Another Interpretation of Ridge Regression and the Lasso

- Minimise

$$\sum_{i=1}^n \left(y^{(i)} - \beta_0 - \sum_{j=1}^p \beta_j x_j^{(i)} \right)^2$$

subject to

$$\sum_{j=1}^p |\beta_j| \leq s \quad (\text{Lasso})$$

or

$$\sum_{j=1}^p \beta_j^2 \leq s \quad (\text{Ridge})$$

where s is related to λ .

- You may recognize this from the idea of Lagrange multipliers.

The Lasso’s “Feasible Region”

- Notice that the constraints for each coefficient in the lasso can also be written as

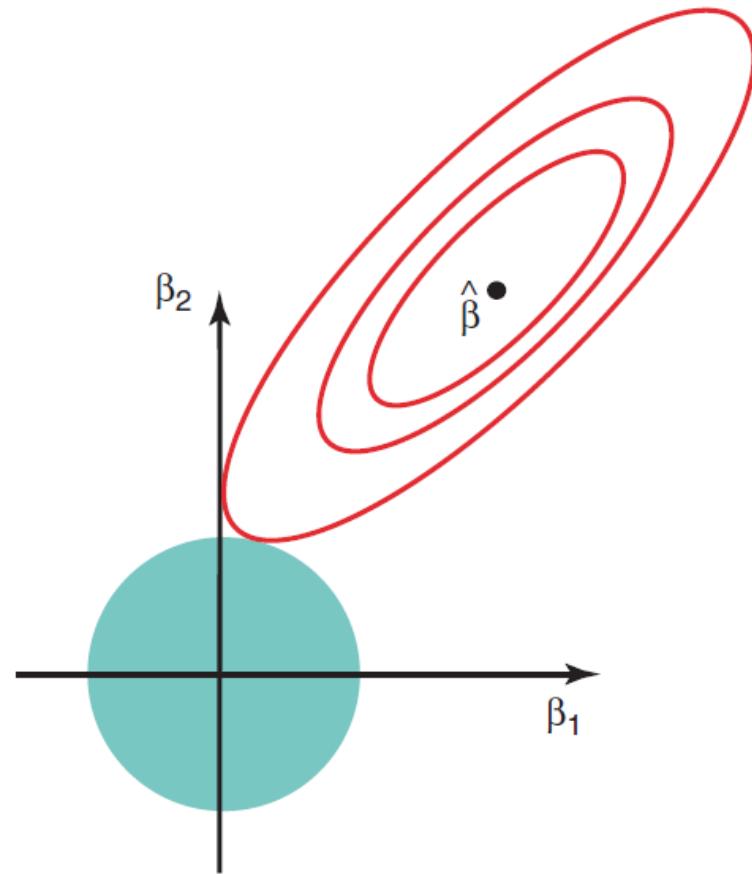
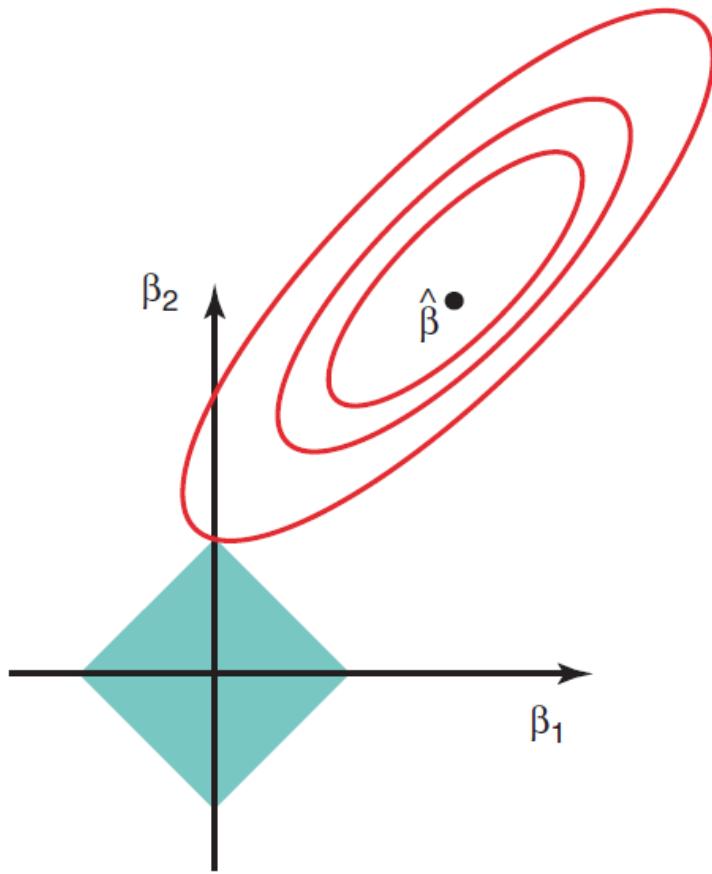
$$-(s + \text{constant}) \leq \beta_j \leq s + \text{constant}$$

if we fix the other coefficients.

- If we need to optimise a quadratic function within an interval, we can hit the boundary if the extreme of the quadratic function is outside of the feasible interval.

The Effect of the l_1 Constraints

- This has a nice consequence: **sparsity**.



Overview: l_0 vs l_1 vs l_2

- l_0 can give sparse solutions, but it is not computationally tractable. Regularisation boils down to zero/one penalties.
- l_2 can give regularised solutions, but in general we will never see a sparse solution.
 - Informally we may think of thresholding, but this is not an optimal solution.
- l_1 can regularise and “sparsify” solutions in a single pass, being computationally convenient.
 - But this conflation of magnitude and sparsity is not ideal as they are not synonyms. Recall that a signal can be “large” and statistically non-significant, and “small” but significant.

Notice

- We have only one penalty term (“**hyperparameter**”):

$$\sum_{i=1}^n \left(y^{(i)} - \beta_0 - \sum_{j=1}^p \beta_j x_j^{(i)} \right)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

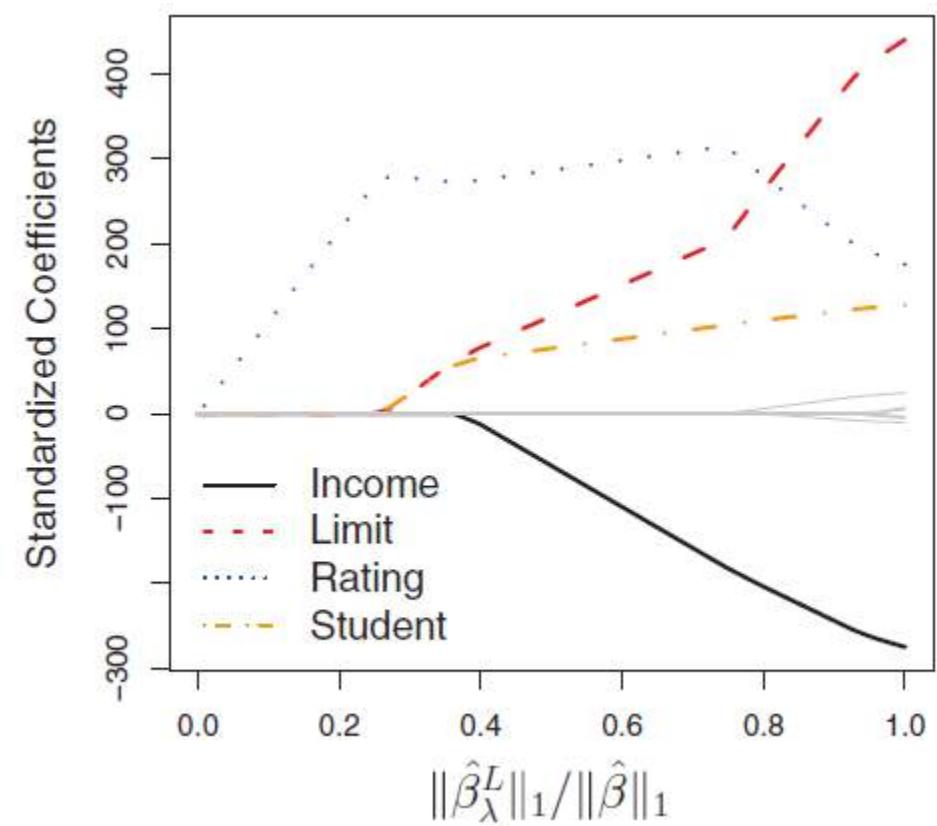
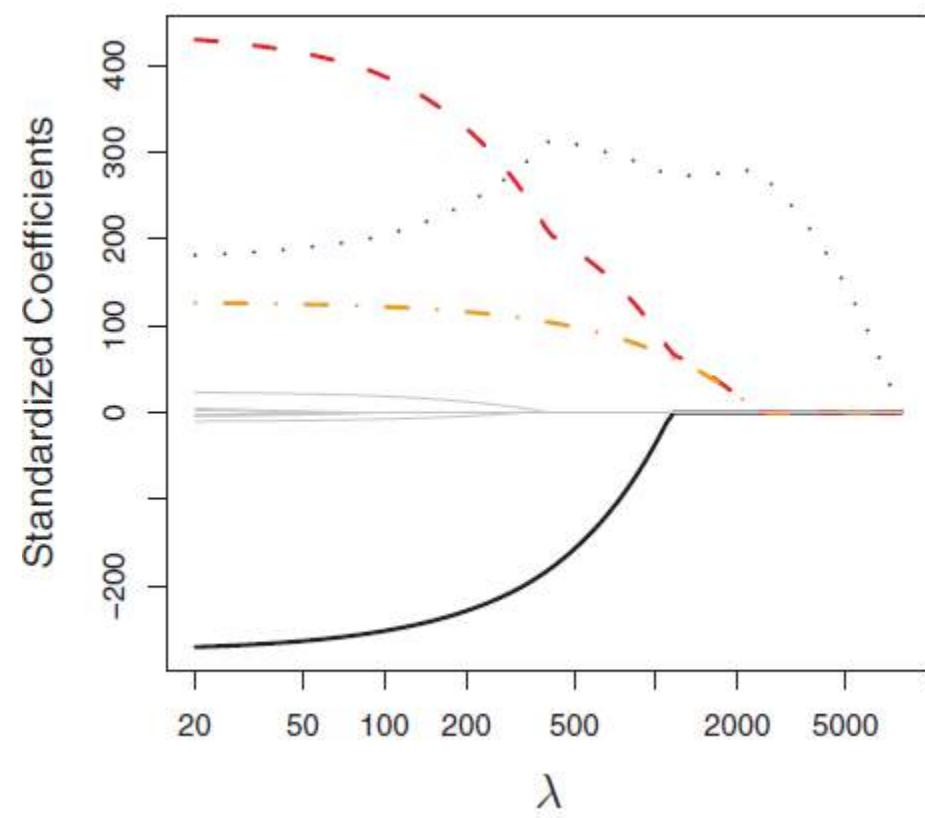
- This means that “large signals” (coefficients) will be “over-shrunk” if we push many other coefficients to zero.
 - There are some (computationally expensive) Bayesian solutions based on using separate penalties per coefficient, and “regularizing the regularizers”.
 - In practice, lasso is much more widely used, and a non-Bayesian alternative of doing cross-validation across many penalty terms is hardly ever done as the number of combinations would explode easily.

Notice

- Under some conditions, ℓ_0 and ℓ_1 agree on the same sparsity pattern.
- However, the usual conditions for that require that the dependency across the inputs is “weak”. So if you were able to solve ℓ_0 , we could expect to have a sparser solution.
- In particular, the problem would be “easy” if the inputs were all uncorrelated.
 - Exercise?

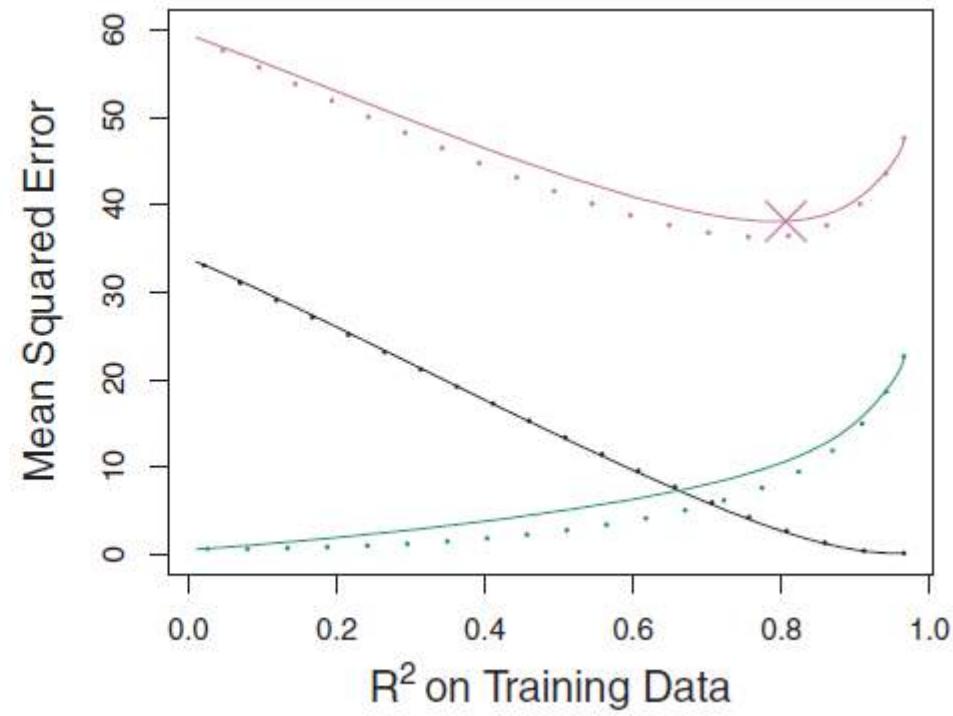
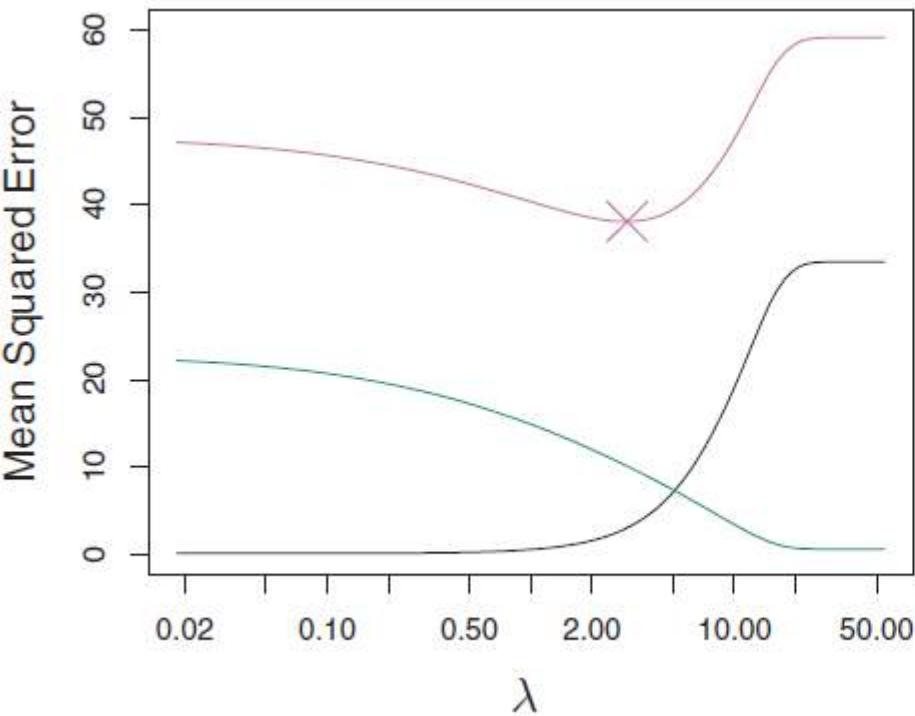
Example: Credit Data

- Notice: standardization is used.



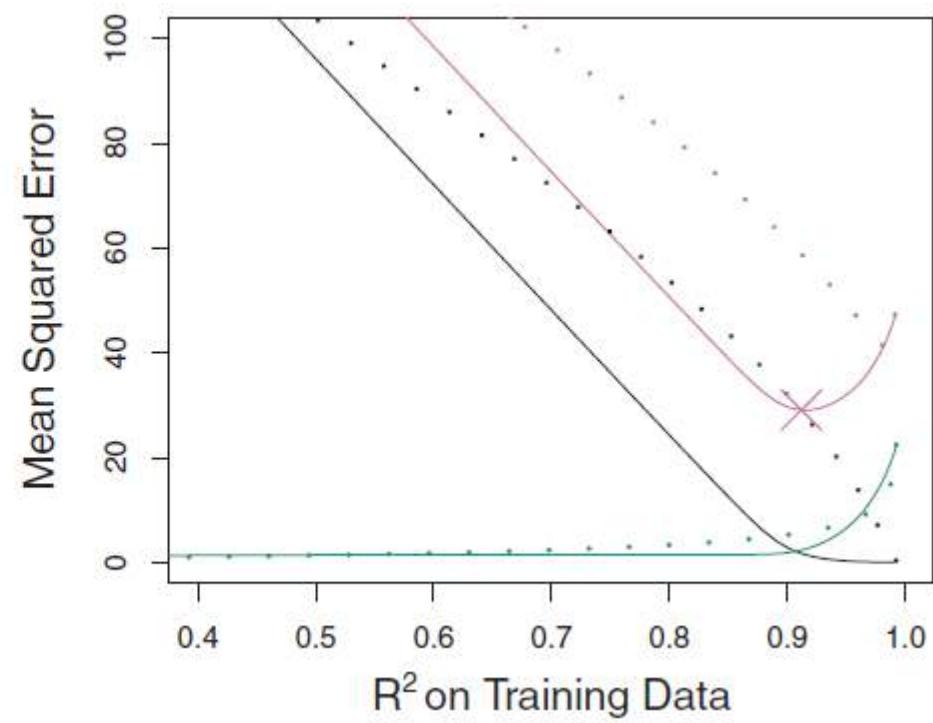
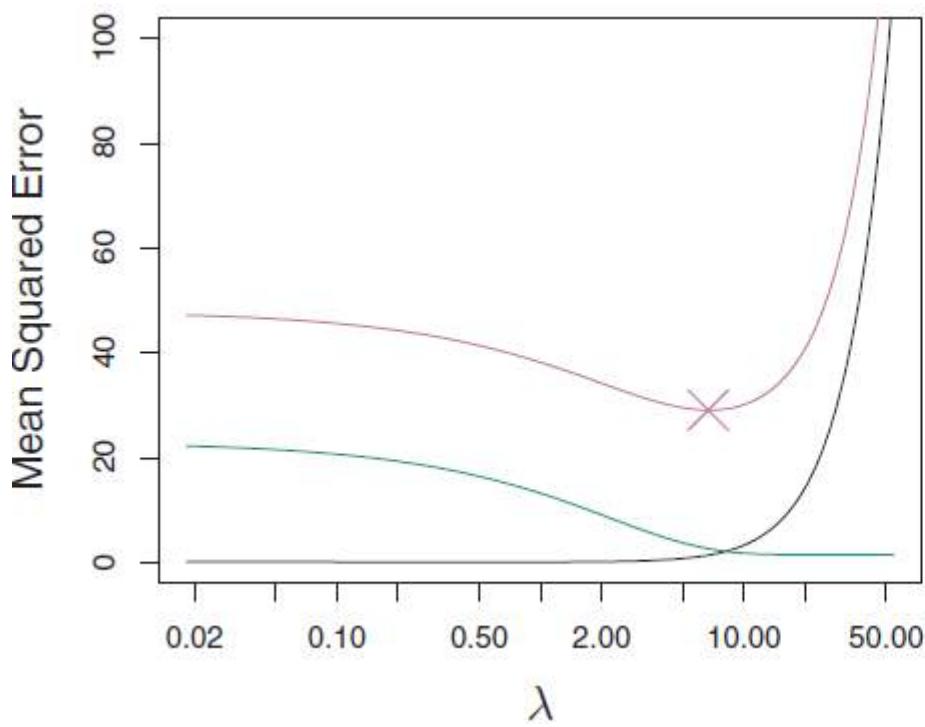
Ridge vs Lasso

- MSE with simulated data. On the right, lasso is solid, ridge is dotted.



Notice

- In the previous simulation, no true sparsity existed. Let's see what happens when only 2 out of 45 variables contribute to the outcome.



Walking Through a Simple Case

(For your reference. Not examinable.)

- We will go through a very simple artificial dataset to better highlight the differences.
- $n = p$, and covariate training matrix \mathbf{X} is diagonal.
- Assume we will not have an intercept, to make things simpler. In least-squares, we get to minimise

$$\sum_{j=1}^p (y^{(j)} - \beta_j)^2$$

Solution?

Solution

(For your reference. Not examinable.)

$$\hat{\beta}_j = y^{(j)}$$

- What if we do ridge regression?

$$\sum_{j=1}^p (y^{(j)} - \beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

We get

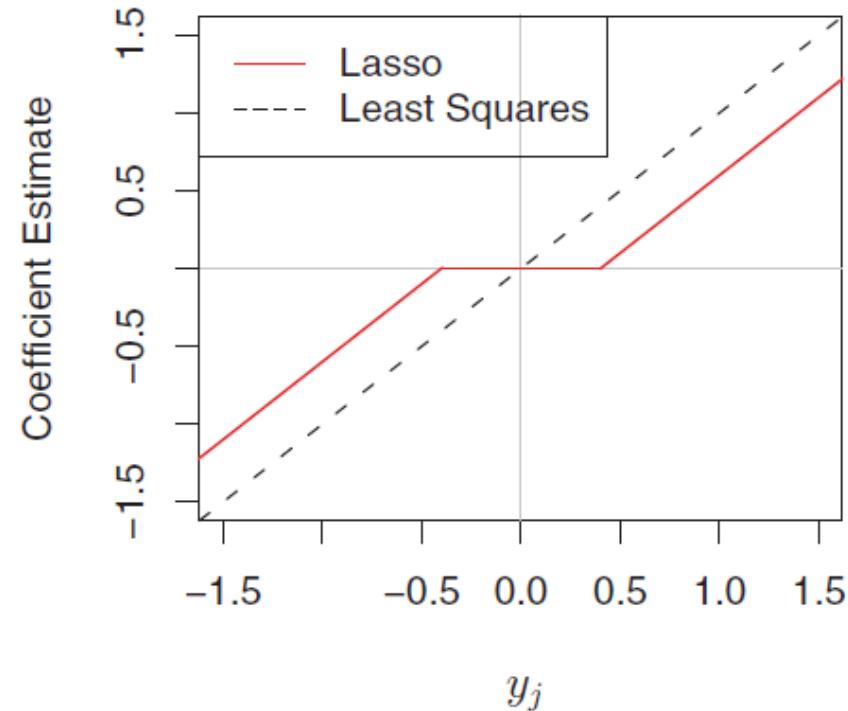
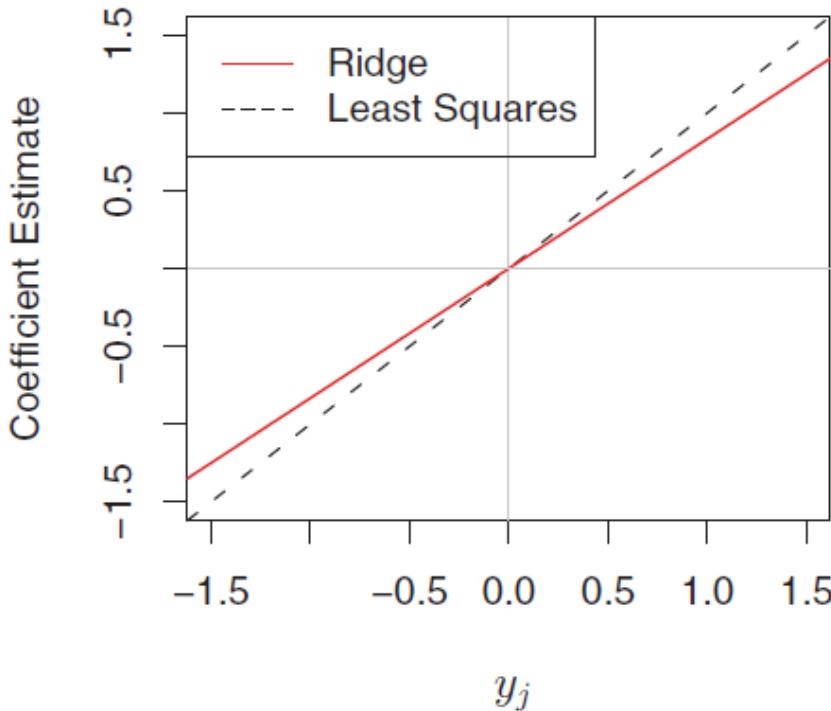
$$\hat{\beta}_j^R = y^{(j)} / (1 + \lambda)$$

The Lasso Solution

(For your reference. Not examinable.)

$$\hat{\beta}_j^L = \begin{cases} y^{(j)} - \lambda/2, & \text{if } y^{(j)} > \lambda/2; \\ y^{(j)} + \lambda/2, & \text{if } y^{(j)} < -\lambda/2; \\ 0, & \text{if } |y^{(j)}| \leq \lambda/2; \end{cases}$$

- That is,



“Soft Thresholding”

- Both ridge regression and lasso shrink the least-squares estimates (push them towards zero).
- After a particular threshold, lasso shrink them all the way to zero.
- In real problems, lasso/ridge will perform better/worse depending how useful soft thresholding might be.
 - Lasso still better in terms of interpretability.
- It is possible to combine both penalties.

A Note on Cross-Validation

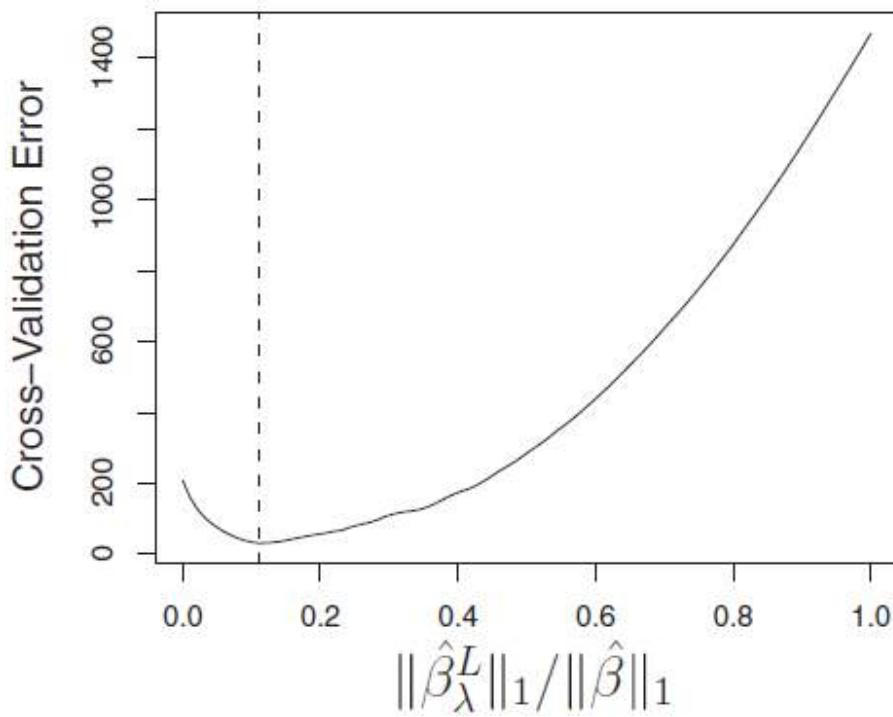
- If you implement cross-validation by yourself, you might be tempted to call a lasso optimiser repeatedly inside a loop over possible λ .
- Don't!
- There are algorithms for easily computing the “regularization path”, starting up lasso from the current value of λ to the next one in an efficient way.
 - R has great libraries for this, package **lars** is an example.
 - Also, manually setting up the range for λ may require some thought.

Interpreting Cross-Validation Plots

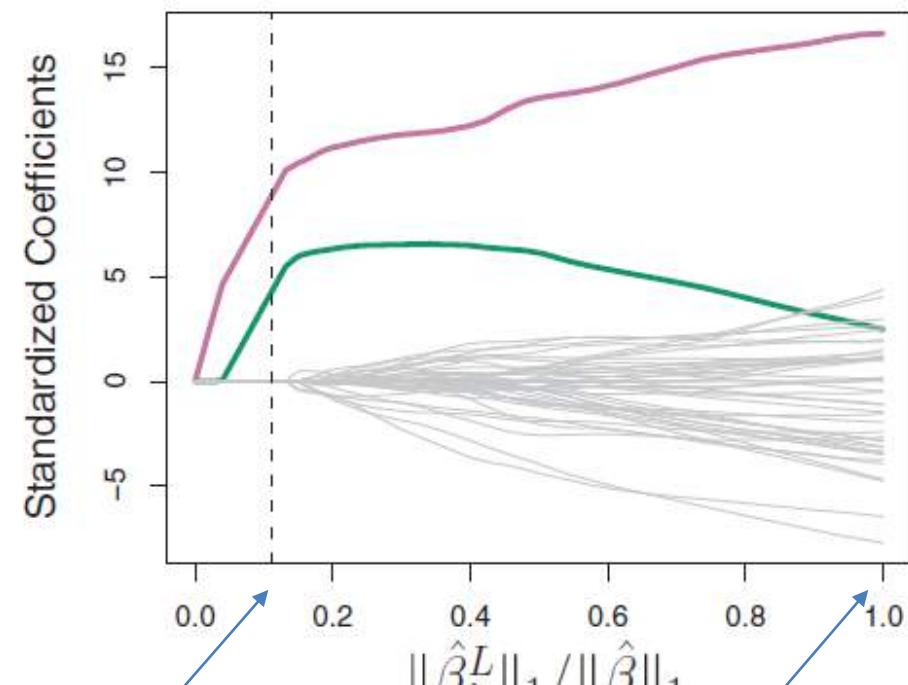
- Be careful when interpreting regularization plots.
- You may see plots of what happens as training progresses (e.g. in neural networks software).
- Instead, you may also see what happens when a penalty parameter changes (e.g., the plots given by lars).

Typical Regularization Path Plots

Simulation: 43 out of 45 variables irrelevant, $n = 50$



Lasso + CV solution



Least-squares solution!

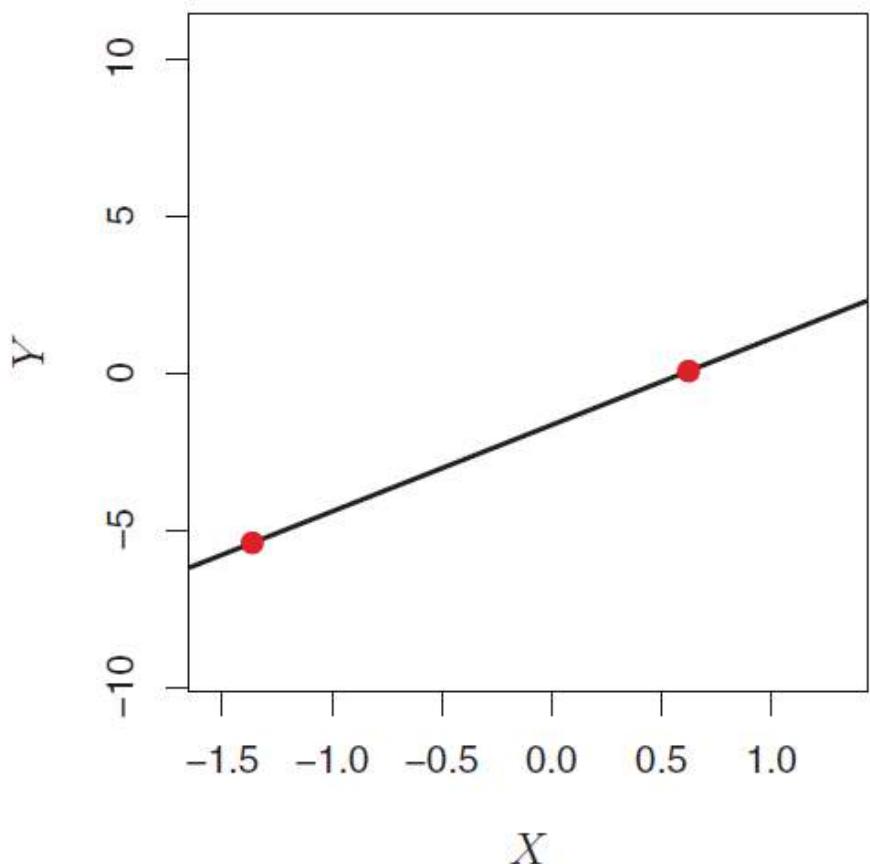
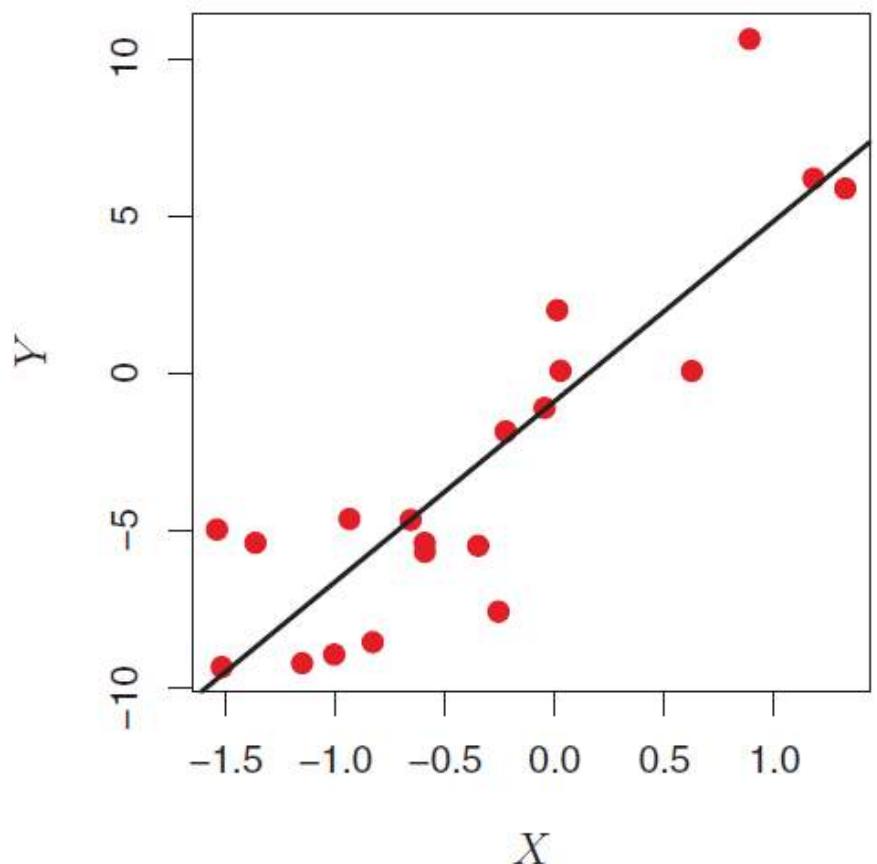
Selected Topics

NOTES ON HIGH-DIMENSIONAL REGRESSION

Technology as a Game Changer

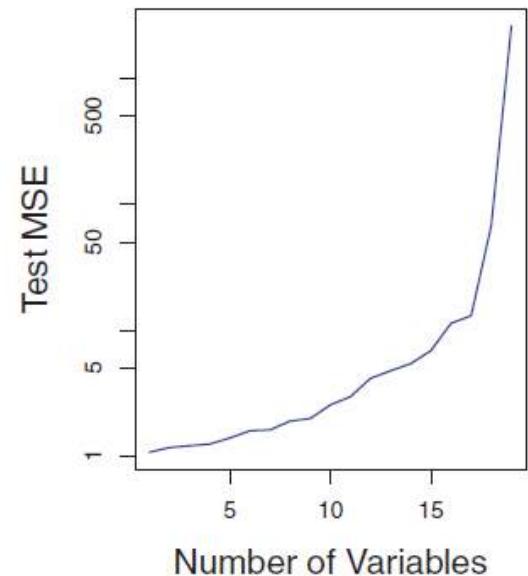
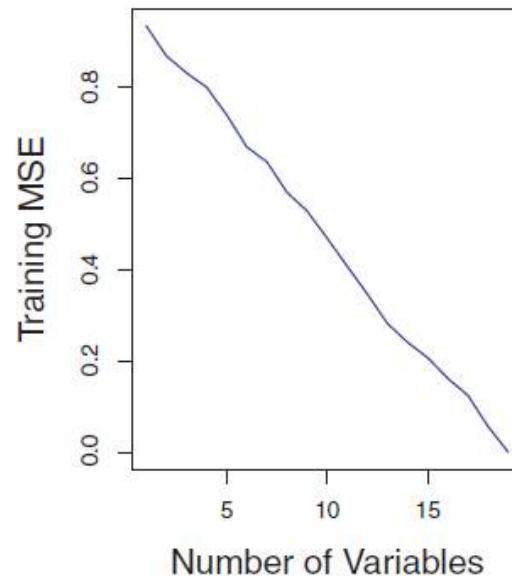
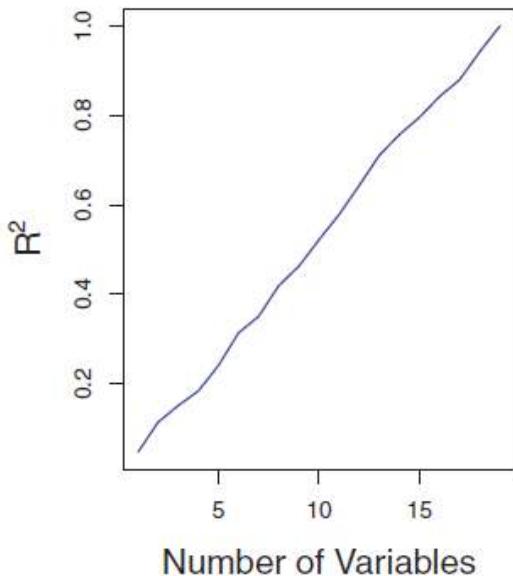
- It is now possible to measure a large number of covariates.
- Examples:
 - high-throughput biological data, at the order of hundreds of thousands genetic features of a cell.
 - text/image data, assuming words/pixels as individual covariates.
- When data is abundant and interpretation is not relevant (e.g., second example above), methods such as modern neural networks might be a good go-to choice.
- However, when data is (relatively) scarce and interpretation is desirable, we need to rethink what regression means.

$$n = p$$



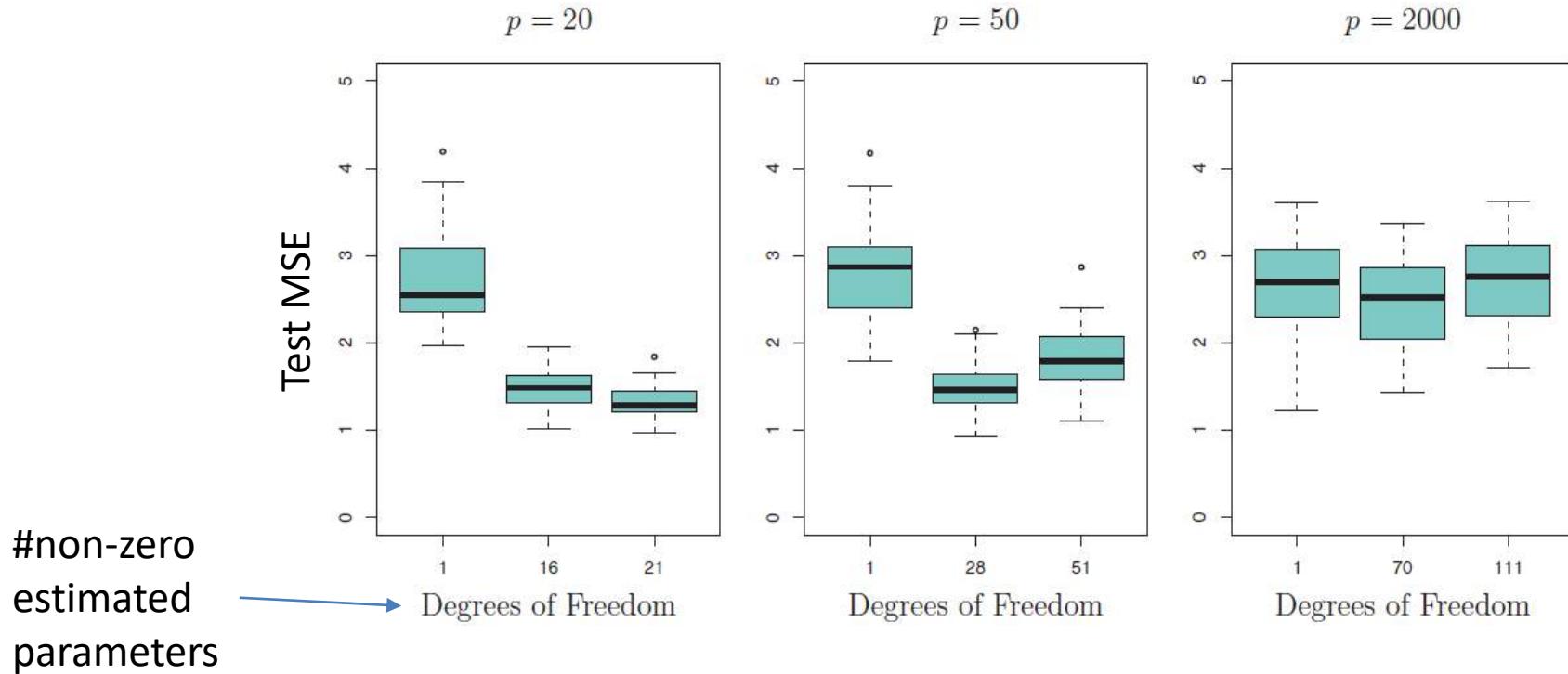
$n = 20, p$ from 1 to 20

- All 20 covariates are unrelated to the outcome
- Likelihood depends on empirical variance, which is even more unstable than the mean: C_p /AIC/BIC useless as p increases.



Simulation: Using Lasso, $n = 100$

- Only 20 features are related to the outcome.
- No cross-validation, just setting three levels of λ .



Interpretation of High-Dimensional Regression

- It is hopeless to think we can learn a “true” model when $p > n$ and there are many relevant covariates. This leaves us with two possibilities.
- First, learn a model “as good as it gets”.
 - We estimate a model that is the best in its class (linear model) and uses as many variables as it is feasible (n).
 - More variables **is** more data after all. For prediction, it should make things **easier** not harder. We just need to make sure we don’t bite more than what we can chew on.

Interpretation of High-Dimensional Regression

- Second, *if* the world is indeed sparse, that is, there are fewer than n relevant variables among those provided, *then* it is possible in theory to recover them.
 - Statistically, some technical assumptions about n / p are necessary.
 - Computationally, some technical assumptions about the correlation of the inputs are necessary.
- This property is sometimes called “*sparsistency*”.

In Practice

- What we recover is one of many possible models, one that can be statistically detected.
 - Training measures such as R^2 will be useless here, but cross-validated predictions are meaningful.
- Prediction-wise, it is silly to throw variables away without looking at the data and without a theoretical justification, so we should welcome large p .
- Concerning what we learn about the world, we will be selecting promising explanatory variables as allowed by our data resources.

Selected Topics

OTHER VARIATIONS OF PENALIZATION

(FOR YOUR REFERENCE. NOT EXAMINABLE.)

Motivation

- L_1 penalties is even more sensitive than least-squares to near-collinearity. As a way of example, if $X_{j'} = X_j$ and $\hat{\beta}_j > 0$ is the estimated lasso coefficient before $X_{j'}$ is introduced, then I'm allowed to set

$$\tilde{\beta}_j + \tilde{\beta}_{j'} = \hat{\beta}_j$$

in a way that keeps the RSS *and* the penalty invariant.

Motivation

- Interestingly, the l_2 penalty is more robust to that: it will just split $\hat{\beta}_j$ equally among the two coefficients (why?). So it is more stable.
- This was an extreme example, but in practice correlations among inputs increase the variance of the lasso estimates.
- l_2 is more stable in this sense, but it gives no sparsity, which might be desirable elsewhere.

Example

- From SLS, Chapter 4

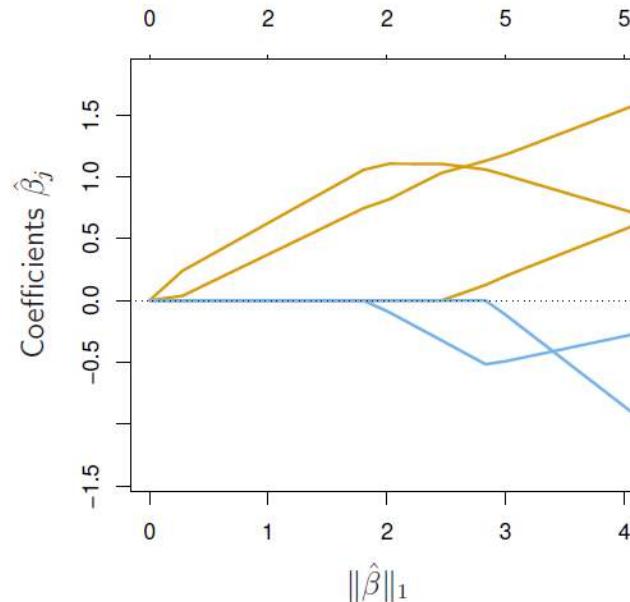
$$Z_1, Z_2 \sim N(0, 1)$$

$$Y = 3Z_1 - 1.5Z_2 + 2\epsilon, \text{ with } \epsilon \sim N(0, 1)$$

$$X_j = Z_1 + \xi_j/5, \text{ with } \xi_j \sim N(0, 1) \text{ for } j = 1, 2, 3, \text{ and}$$

$$X_j = Z_2 + \xi_j/5, \text{ with } \xi_j \sim N(0, 1) \text{ for } j = 4, 5, 6.$$

$$\alpha = 1.0$$

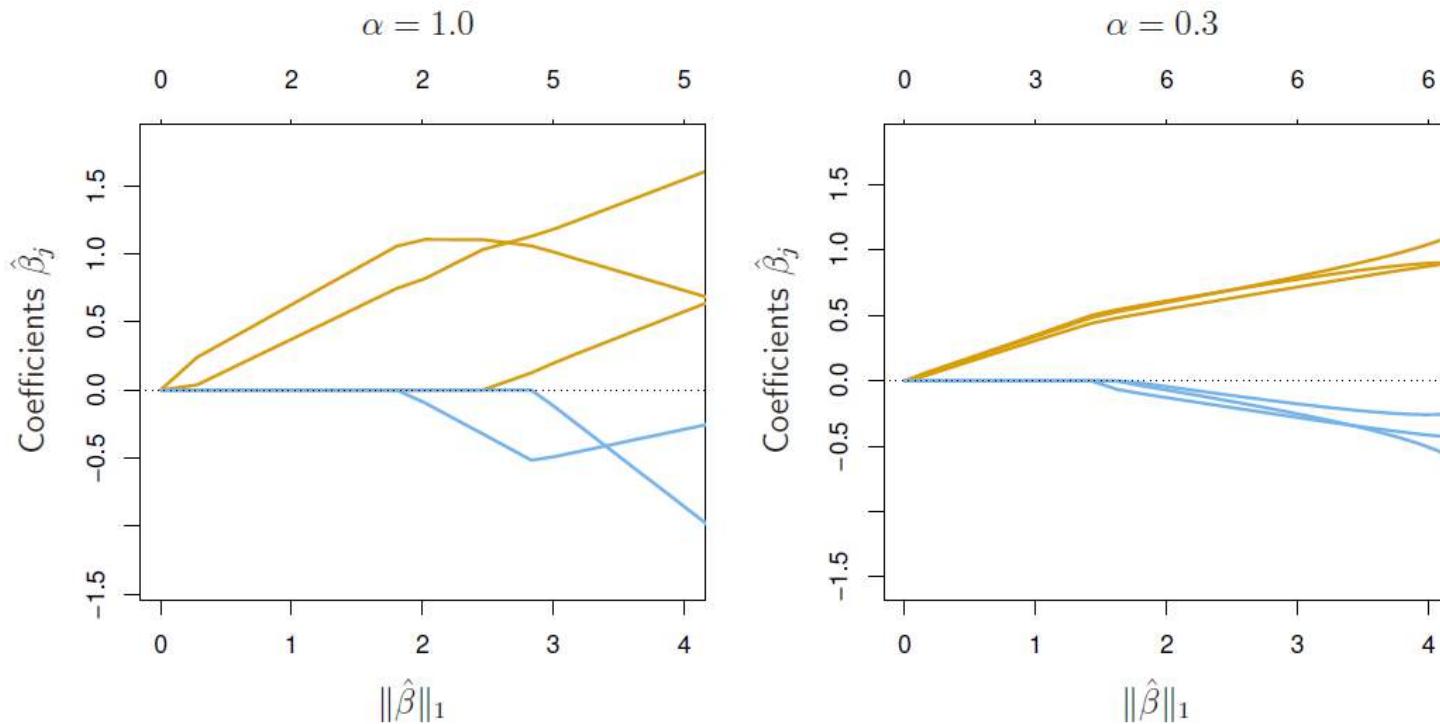


The Elastic Net

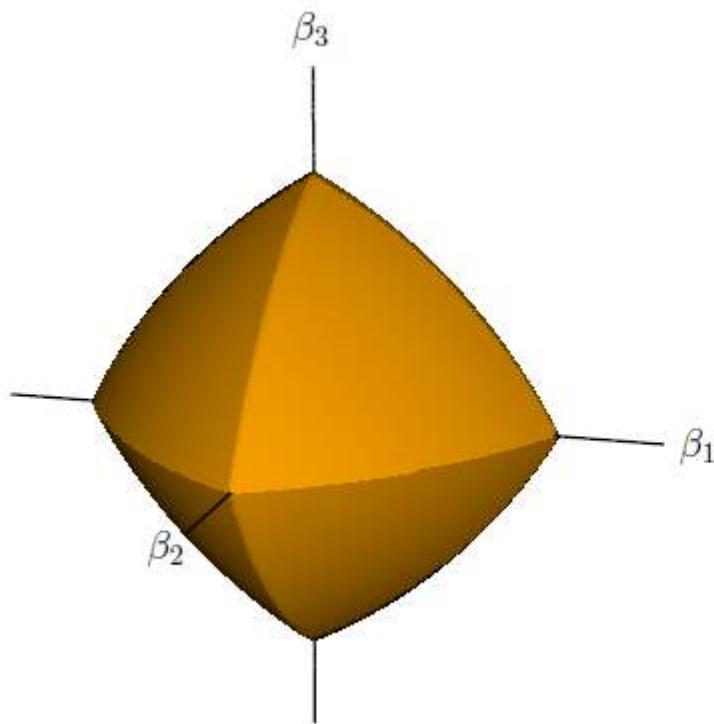
$$\frac{1}{2} \sum_{i=1}^n (y^{(i)} - \beta_0 - \beta^T \mathbf{x}^{(i)})^2 + \lambda \left[\frac{1}{2}(1-\alpha) \sum_{j=1}^p \beta_j^2 + \alpha \sum_{j=1}^p |\beta_j| \right]$$

- Two regularizers, λ and α . Cross-validation is more expensive as the grid of possibilities increases.
- When $\alpha = 1$, we recover the lasso. When $\alpha = 0$, ridge regression.
- The problem is still convex for $\alpha < 1$ and $\lambda > 0$.
- Let's see how it behaves in our synthetic case.

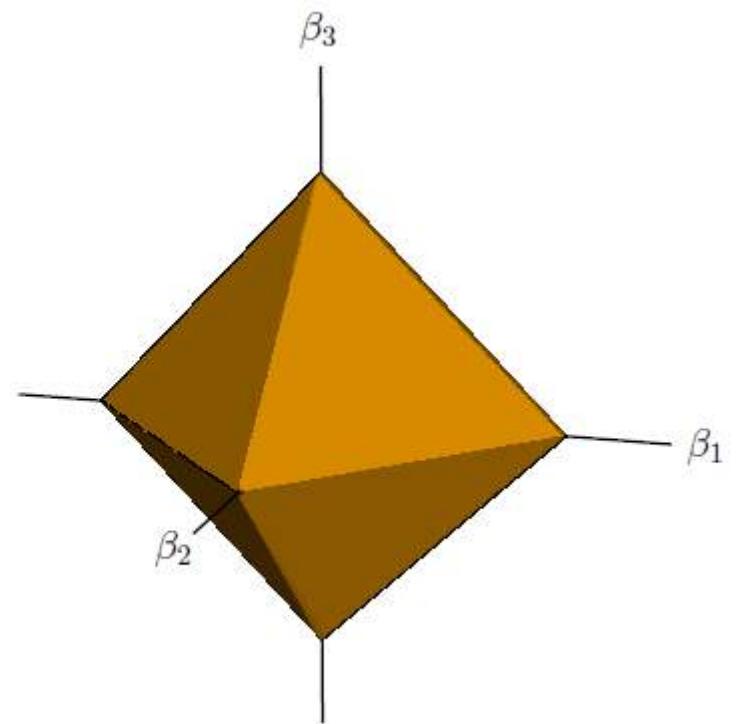
Elastic Net Results



Visualization of Feasible Region



Elastic net, $\alpha = 0.7$



Lasso

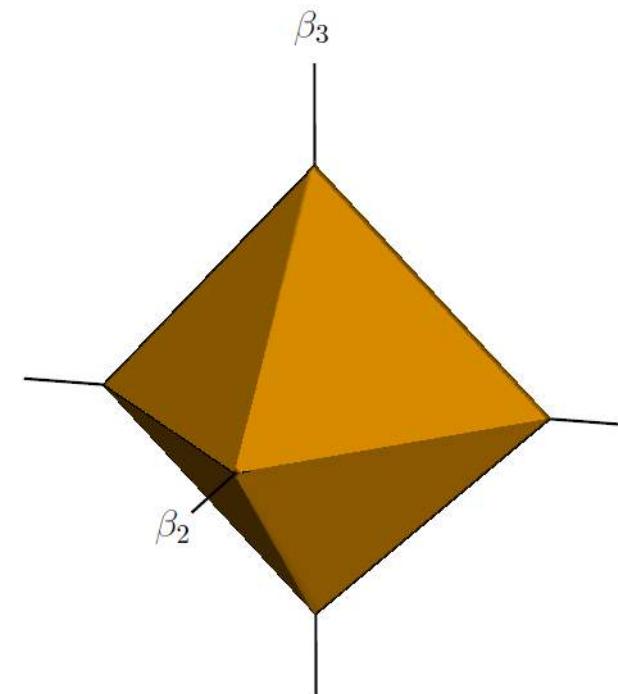
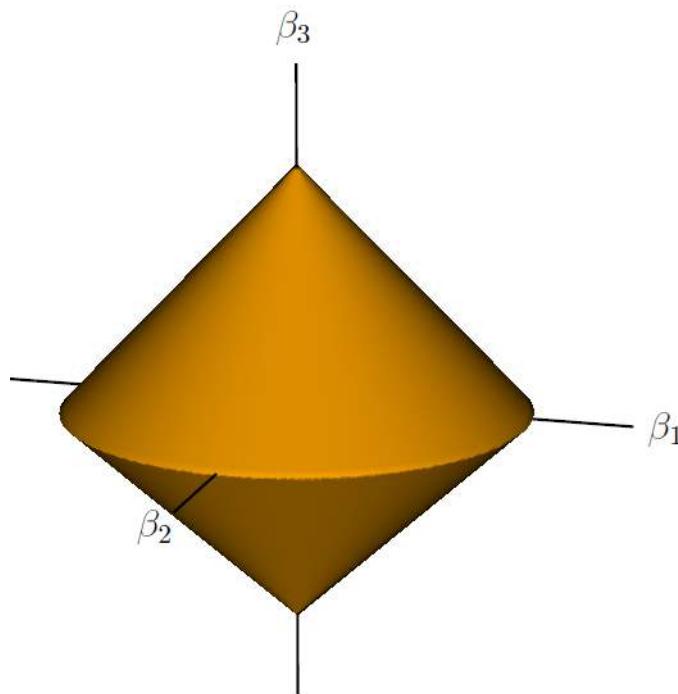
Another Variant: Group Lasso

- Sometimes covariates come in (known) groups
 - For instance, the one-of-K encoding of categorical variables.
 - It feels natural that either the whole group, or nobody in the group, should be selected.
- The group lasso for J groups:

$$\frac{1}{2} \sum_{i=1}^n (y^{(i)} - \beta_0 - \sum_{j=1}^J \beta_j^T \mathbf{x}_j^{(i)})^2 + \lambda \sum_{j=1}^J \sqrt{\beta_{j1}^2 + \dots + \beta_{jk_j}^2}$$

Visualization

Groups: $\{\beta_1, \beta_2\}$ and β_3



Application: Categorical Inputs

- Say you have a category with four levels (e.g., whether you live in England, Wales, Scotland or North Ireland).
- Let us represent it as a one-of-4 encoding (Z_1, Z_2, Z_3, Z_4), where England means $(1, 0, 0, 0)$, etc.
- Using group lasso (with other k_1 covariates \mathbf{x}_1), we can model it as

$$\frac{1}{2} \sum_{i=1}^n (y^{(i)} - \beta_0 - \beta_1^T \mathbf{x}_1^{(i)} - \beta_{21} z_1^{(i)} - \beta_{22} z_2^{(i)} - \beta_{23} z_3^{(i)} - \beta_{24} z_4^{(i)})^2 +$$

$$\lambda \left(\sqrt{\beta_{11}^2 + \dots + \beta_{1k_1}^2} + \sqrt{\beta_{21}^2 + \dots + \beta_{24}^2} \right)$$

Application: Categorical Inputs

- Without proof, I will state that the estimated coefficients in the group are either all zero or all non-zero.
- Notice that in this case we do not really worry about the fact that

$$z_1 + z_2 + z_3 + z_4 = 1$$

as the penalty will enforce this automatically that the corresponding coefficients add up to zero (see SLS Exercise 4.4 and Exercise Sheet #6).

Interactions and Hierarchies

- Group lasso also preserves hierarchies in interaction models. That is, if a product $x_j x_k$ is included, so will the “main effects” x_j and x_k .
- Example:

We can show (proof omitted) that if $\beta_{12} \neq 0$, then $\beta_1 \neq 0$ and $\beta_2 \neq 0$, also.

$$\frac{1}{2} \sum_{i=1}^n (y^{(i)} - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_{12} x_1 x_2)^2 + \lambda \left(\sqrt{\beta_1^2} + \sqrt{\beta_2^2} + \sqrt{\beta_{12}^2} \right)$$

Notice that in the one-dimensional case,
this is exactly the same as $|\beta_1|$ etc.

(See SLS, Example 4.3)

Finally

- It should go without saying that similar ideas apply to sparse GLMs.
See SLS for more.
- In R, look at the `glmnet` package in you are interested.

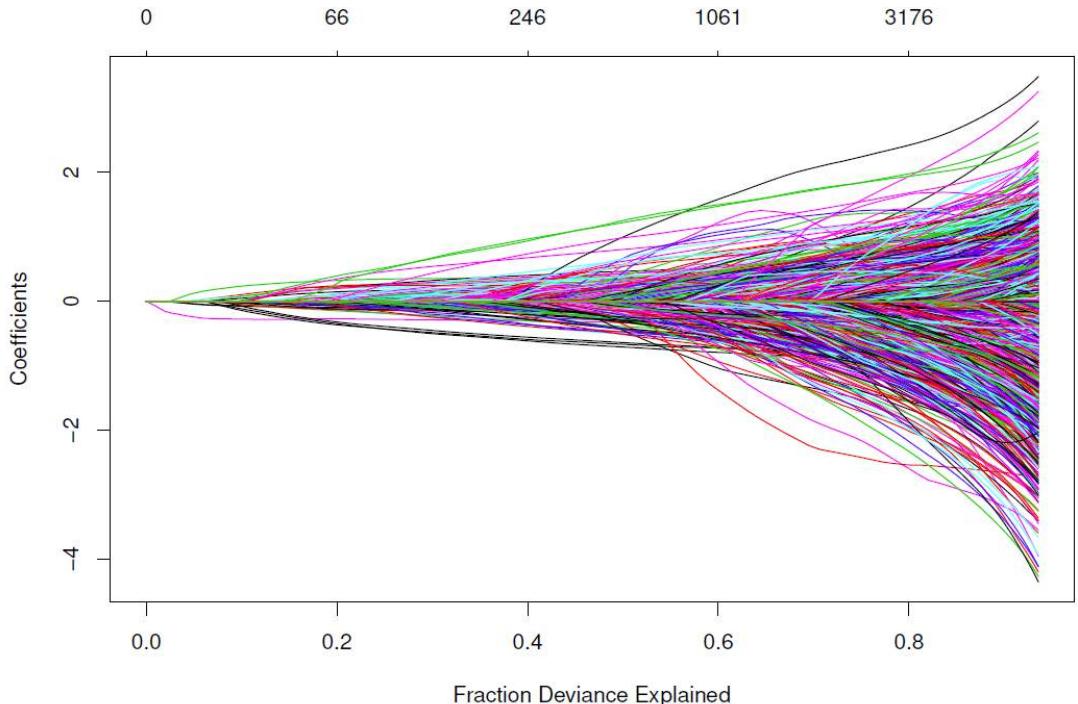


Figure 3.1 Coefficient paths for an ℓ_1 -regularized logistic regression for a document-classification task—the “NewsGroup” data. There are 11K documents roughly divided into two classes, and 0.78M features. Only 0.05% of the features are nonzero. The coefficients are plotted as a function of the fraction of null deviance explained.

Selected Topics

STATISTICAL INFERENCE FOR THE LASSO ESTIMATOR

Inference in Sparse Models

- We would like to have measures such as the standard deviation of lasso estimators, confidence intervals, or the frequentist probability of a coefficient estimate being zero.
- Unfortunately, this is far too hard to derive analytically even in the Gaussian likelihood case.
 - Who you gonna call? Bootstrap!
 - Side note: there is a on-going area of research on “post-selection” inference which I won’t mention, e.g., correcting confidence intervals after estimating the sparsity pattern with lasso/hypothesis testing etc. See Chapter 6 of SLS if you are curious.

Estimating Standard Errors

- Least-squares + l_1 penalty, an example using the bootstrap.
 - Adapted from Exercise 2.6 of SLS
 - **Data:** crime rate for $n = 50$ U.S. cities. Five predictors, including funding for police in dollars per resident, percentage of 16-19 not in high school, etc. Outcome is crime occurrences per million residents.
 - **We will focus here on finding standard errors,** but exactly the same idea applies to calculating confidence intervals using other bootstrap methods.

Estimating Standard Errors

Table 2.2 Results from analysis of the crime data. Left panel shows the least-squares estimates, standard errors, and their ratio (Z-score). Middle and right panels show the corresponding results for the lasso, and the least-squares estimates applied to the subset of predictors chosen by the lasso.

	LS coef	SE	Z	Lasso	SE	Z	LS	SE	Z
funding	10.98	3.08	3.6	8.84	3.55	2.5	11.29	2.90	3.9
hs	-6.09	6.54	-0.9	-1.41	3.73	-0.4	-4.76	4.53	-1.1
not-hs	5.48	10.05	0.5	3.12	5.05	0.6	3.44	7.83	0.4
college	0.38	4.42	0.1	0.0	-	-	0.0	-	-
college4	5.50	13.75	0.4	0.0	-	-	0.0	-	-

Algorithm

1. Estimate regularizer $\hat{\lambda}$ from full data using cross-validation
2. Draw $(Y^{(1)\star}, X^{(1)\star}), \dots, (Y^{(n)\star}, X^{(n)\star}) \sim \hat{F}_n$
3. Estimate β_n^\star from bootstrap data using regression with regularization $\hat{\lambda}$
4. Repeat steps 2 and 3, B times, to get $\beta_{n,1}^\star, \dots, \beta_{n,B}^\star$
5. Let

$$s.e.\text{-}boot \equiv \sqrt{\frac{1}{B} \sum_{b=1}^B \left(\beta_{n,b}^\star - \frac{1}{B} \sum_{r=1}^B \beta_{n,r}^\star \right)^2}$$

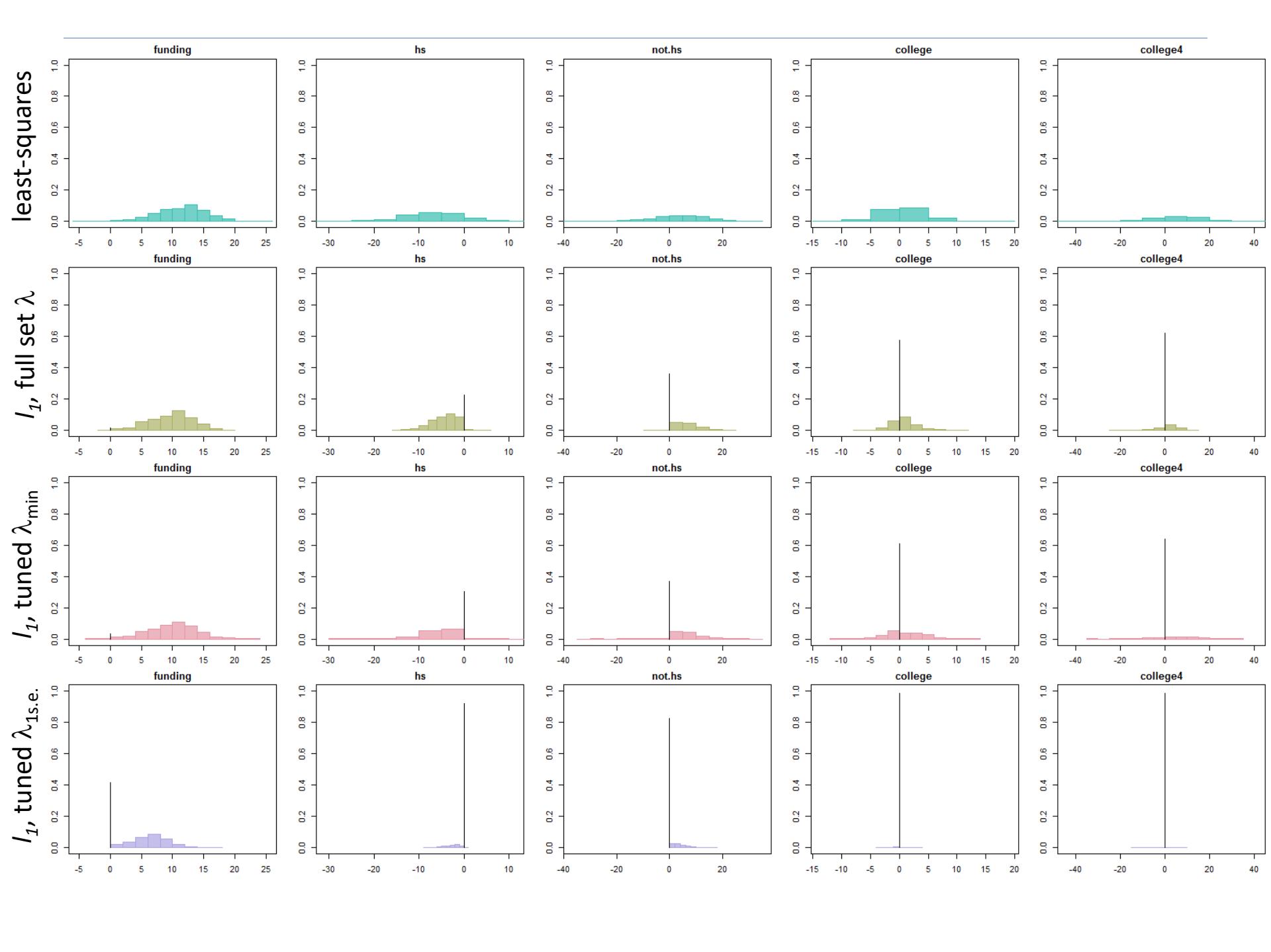
(Notice: β can be a vector)

Alternative Algorithm

1. Draw $(Y^{(1)\star}, X^{(1)\star}), \dots, (Y^{(n)\star}, X^{(n)\star}) \sim \hat{F}_n$
2. Estimate β_n^\star from bootstrap data using your regression with cross-validation to get regularizer λ_b^\star
3. Repeat steps 1 and 2, B times, to get $\beta_{n,1}^\star, \dots, \beta_{n,B}^\star$
4. Let

$$s.e.\text{-}boot \equiv \sqrt{\frac{1}{B} \sum_{b=1}^B \left(\beta_{n,b}^\star - \frac{1}{B} \sum_{r=1}^B \beta_{n,r}^\star \right)^2}$$

(Notice: β can be a vector)



Alternative Bootstrap: Resampling Residuals

1. Estimate regularizer $\hat{\lambda}$ from full data using cross-validation
2. Calculate residuals $\hat{\epsilon}^{(i)}$, define \hat{F}_n^ϵ as the empirical distribution of residuals
3. Draw $\epsilon^{(1)\star}, \dots, \epsilon^{(n)\star} \sim \hat{F}_n^\epsilon$
4. Generate $Y^{(i)\star}$ from $\hat{\beta}$, $X^{(i)}$ and $\epsilon^{(i)\star}$
5. Estimate β_n^* from bootstrap data using regression with regularization $\hat{\lambda}$
6. Repeat steps 3, 4 and 5 for B times, to get $\beta_{n,1}^*, \dots, \beta_{n,B}^*$
7. Let

$$s.e.\text{-}boot \equiv \sqrt{\frac{1}{B} \sum_{b=1}^B \left(\beta_{n,b}^* - \frac{1}{B} \sum_{r=1}^B \beta_{n,r}^* \right)^2}$$

Disclaimer

- The residual bootstrap for the lasso has some undesirable theoretical properties I will not get in detail.
- However, for ridge regression it remains a theoretically sound approach.

When to Use Which?

- The “residual bootstrap” tends to get narrower confidence intervals, as it forces the expectation of the outcome to follow a regression model.
- Which, remember, might not be true in reality, and it is not imposed by the “fully” nonparametric bootstrap.
- So this is another type of bias-variance trade-off, which is up to you as Data Scientist to accept or not.

Selected Topics

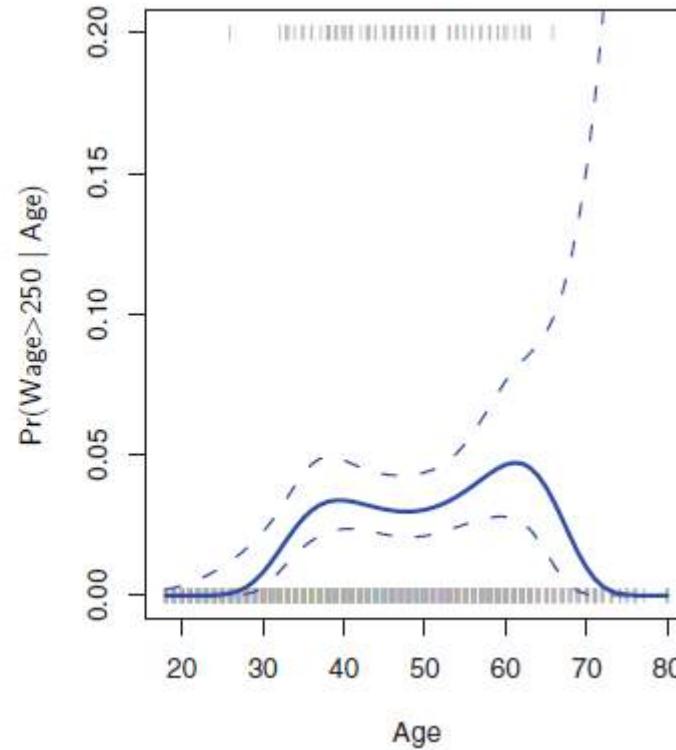
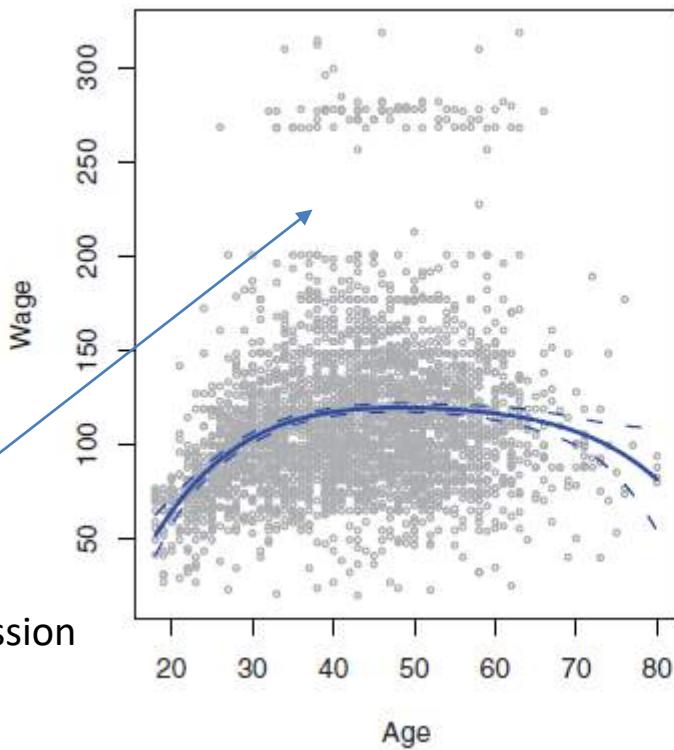
NONLINEAR MODELS: THE UNIVARIATE CASE

Polynomial Regression Revisited

- We can take a covariate x and expand it in terms of x, x^2, x^3 etc.
- What about confidence intervals in the original space? How do they behave?
 - Polynomials of degree > 3 are not particularly trustworthy. Avoid them for all practical purposes.
- We may want to look at other nonlinear transformations.

Example: 4th Order Logistic Regression

$$\log \frac{P(\text{Wage} > 250 \mid \text{Age} = x)}{P(\text{Wage} \leq 250 \mid \text{Age} = x)} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 x^3 + \hat{\beta}_4 x^4$$



Flexible Models from Local Basis Functions: the Step Case

- Any parameter in a polynomial regression problem has a global implication (why?).
- Let's think in terms of more general **basis functions**, nonlinear transformations of our inputs.
 - Sometimes, a collection of basis functions is called a **dictionary**.
 - If you have a signal processing background, you may be familiar with concepts such as Fourier or wavelet decomposition. Nonlinear regression modelling does share some of its “DNA” with signal processing.
- In particular, an alternative for flexible regression is the use of **local**, or **sparse**, basis functions: basis functions which are zero in most of the input space.

Binning

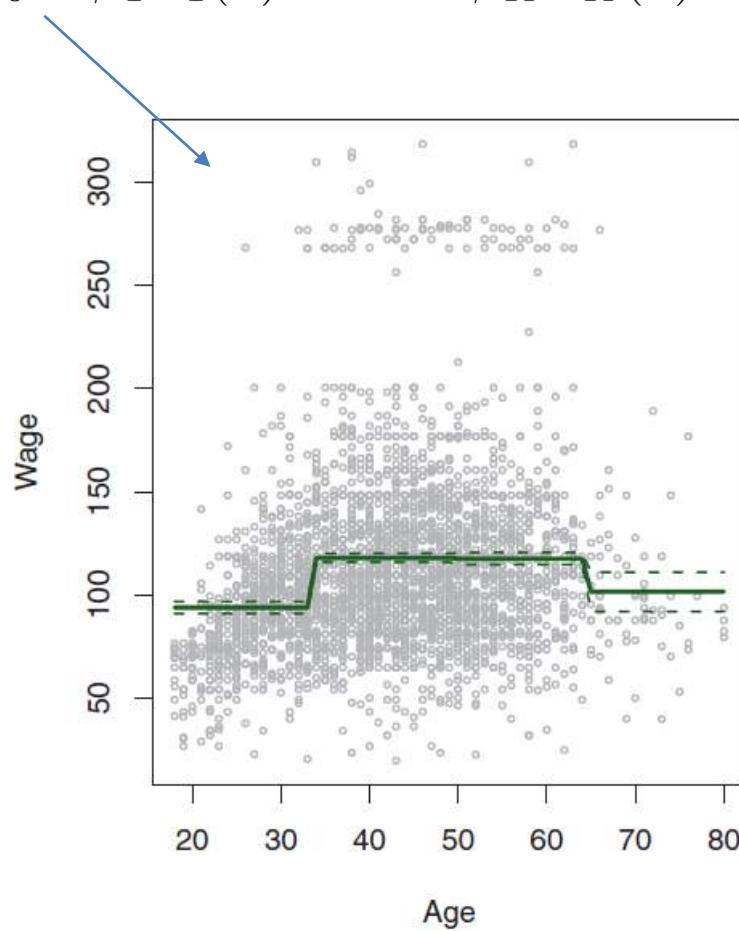
- Create cutpoints in the range of covariate X :

$$\begin{aligned}C_0(X) &= I(X < c_1) \\C_1(X) &= I(c_1 \leq X < c_2) \\C_2(X) &= I(c_2 \leq X < c_3) \\&\dots \\C_{K-1}(X) &= I(c_{K-1} \leq X < c_K) \\C_K(X) &= I(X \leq c_K)\end{aligned}$$

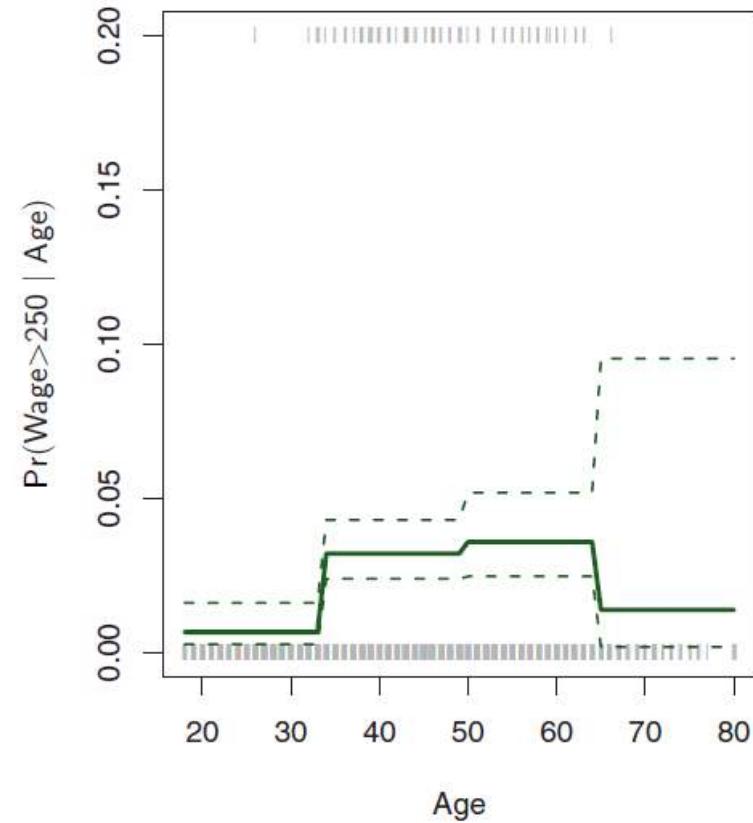
where once again remember that $I()$ is an indicator function (outputs 0 or 1 depending whether argument is false or true).

Example

$$Y = \beta_0 + \beta_1 C_1(x) + \cdots + \beta_K C_K(x) + \epsilon$$



- What is the interpretation of β_j ?
- Which shortcomings do you see?



Alternative: Regression Splines

- This combines ideas from the polynomial and step basis functions.
- Basic insight: fit **separate (low-degree) polynomials** over different regions of the input space. This is sometimes called a “piecewise” model.
- Most common building block: cubic polynomials.

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

Knots

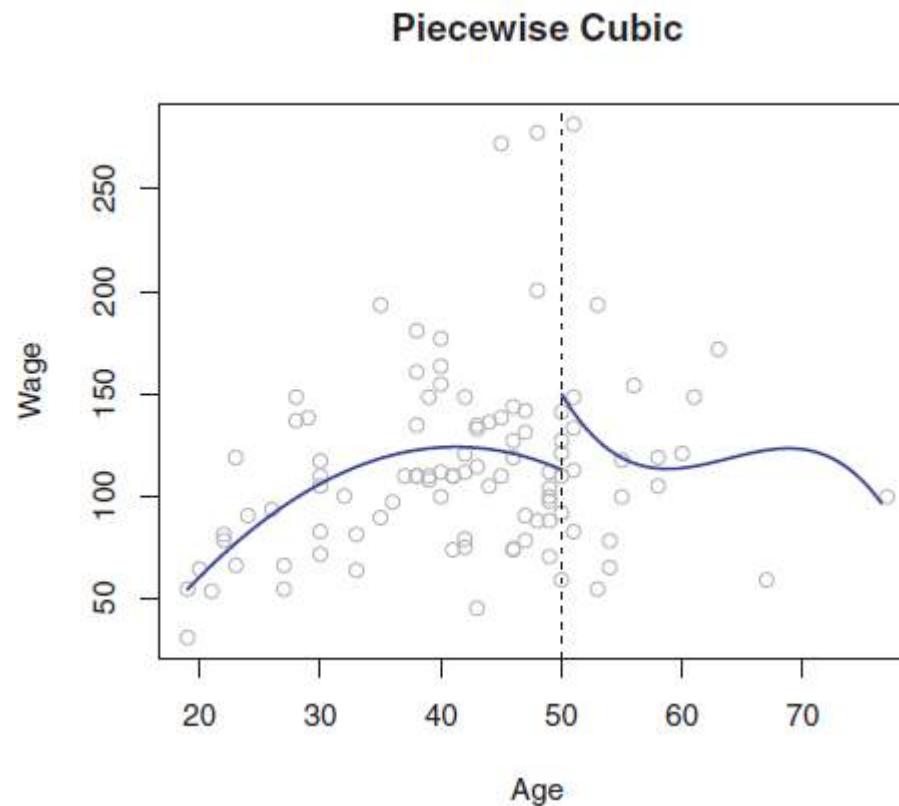
- The technical name given to points in the input space where coefficients change. For example, some point c such that:

$$Y = \begin{cases} \beta_{01} + \beta_{11}x + \beta_{21}x^2 + \beta_{31}x^3 + \epsilon, & \text{if } x < c; \\ \beta_{02} + \beta_{12}x + \beta_{22}x^2 + \beta_{32}x^3 + \epsilon, & \text{if } x \geq c. \end{cases}$$

- K knots will lead to $K + 1$ polynomials.
- Notice that the step function was a spline of degree 0 (“piecewise constant”).
- “Piecewise linear” is what we get by fitting linear models within each region.

First Attempt

- Let's get back to fitting a model for the wages data, splitting at age $c = 50$.



Needed

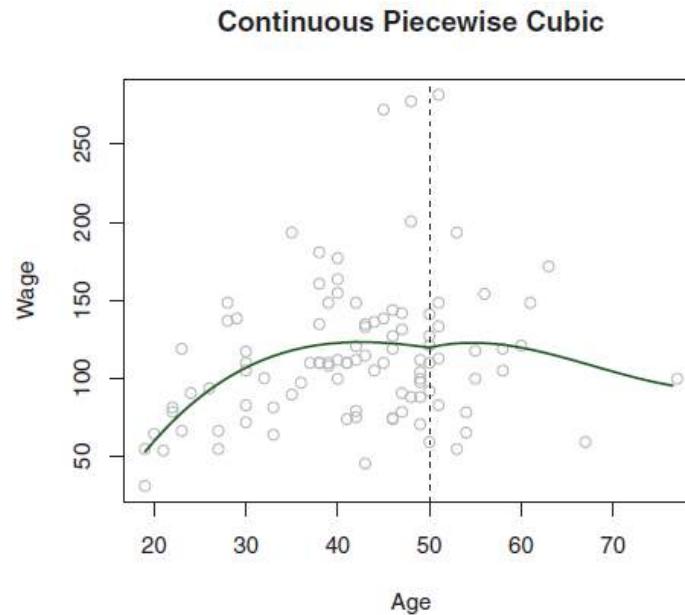
- A way of fitting parameters so that they agree at the knots.
- What does agreement mean? We can “tie the knot” by enforcing

$$\beta_{01} + \beta_{11}c + \beta_{21}c^2 + \beta_{31}c^3 = \beta_{02} + \beta_{12}c + \beta_{22}c^2 + \beta_{32}c^3$$

- How many free parameters do we get by fitting two cubic polynomials that agree at the knot?

One Tie, One Fewer Degree of Freedom

- We can express one intercept as a function of other parameters, so we have 7 free parameters (“degrees of freedom”).



- Still doesn't look right, does it?

More Smoothing

- Let's enforce not only agreement at the know, but also agreement of the *first derivatives*, and agreement of the *second derivatives*.

$$\beta_{01} + \beta_{11}c + \beta_{21}c^2 + \beta_{31}c^3 = \beta_{02} + \beta_{12}c + \beta_{22}c^2 + \beta_{32}c^3$$

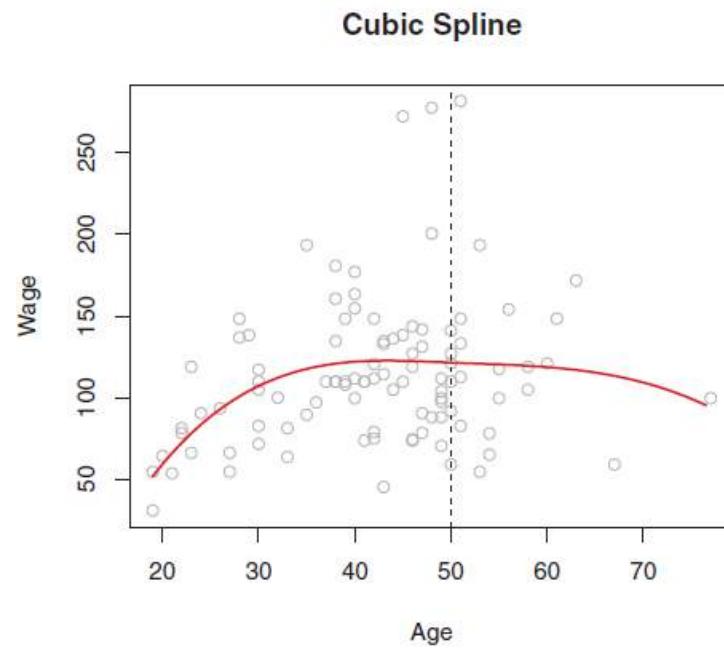
$$\beta_{11} + 2\beta_{21}c + 3\beta_{31}c^2 = \beta_{12} + 2\beta_{22}c + 3\beta_{32}c^2$$

$$2\beta_{21} + 6\beta_{31}c = 2\beta_{22} + 6\beta_{32}c$$

- We are left with 5 degrees of freedom.

Cubic Splines

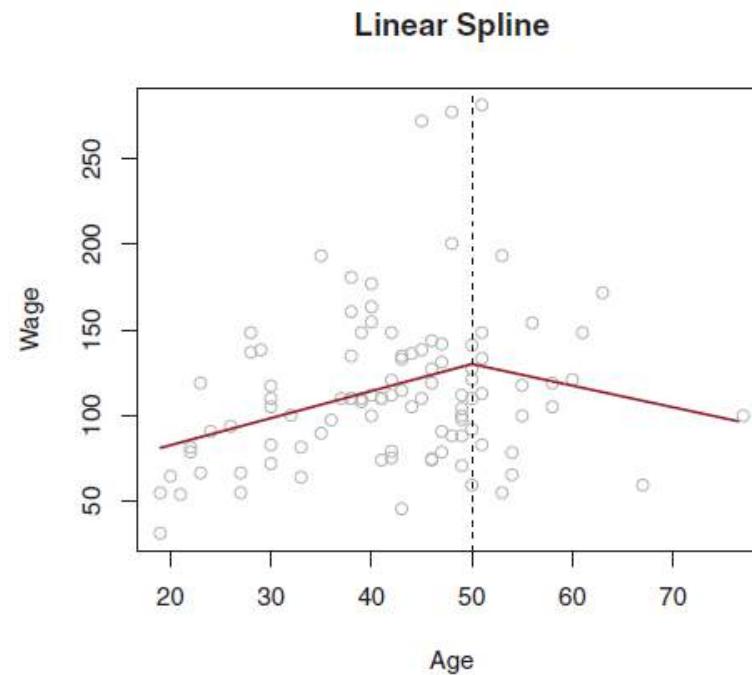
- The result is the unimaginatively named **cubic spline**.



- It is not hard to show that for K knots, we have $4 + K$ degrees of freedom.

Degree- d Spline

- We tie the derivatives up to degree $d - 1$. So we if had linear polynomials, “derivative 0” is just the original lines.



A More Suitable Formulation

- You may notice that optimising the parameters (say, by least squares) is a bit annoying because of the constraints.
- One idea is a **change of representation**, which in an abstract form relates to reparameterisation (recall GLMs!)
- What if we had $K + 3$ basis functions such that

$$Y = \beta_0 + \beta_1 b_1(x) + \beta_2 b_2(x) + \cdots + \beta_{K+3} b_{K+3}(x) + \epsilon$$

The Truncated Power Basis Representation

- For each knot ξ ,

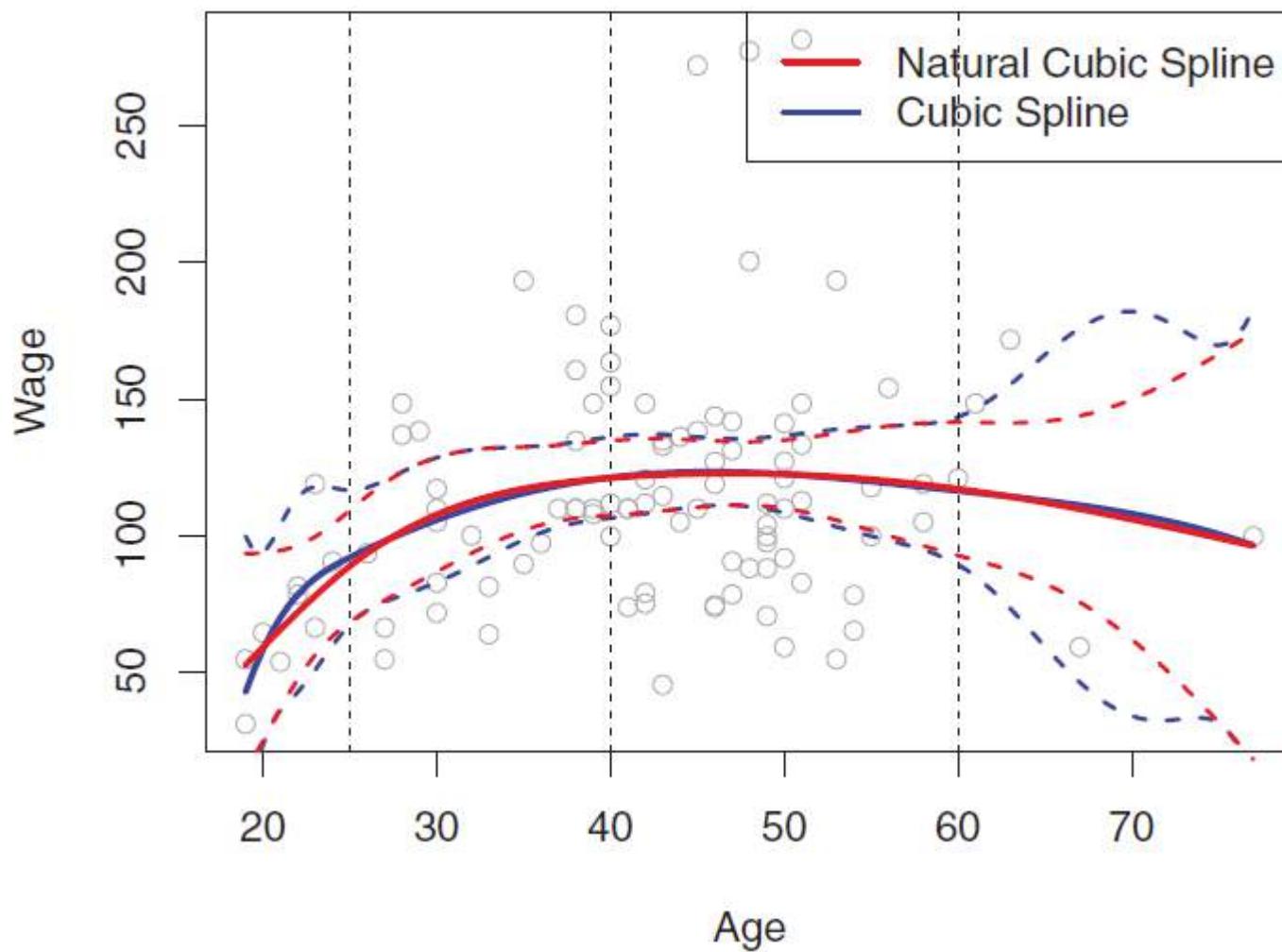
$$h(x, \xi) \equiv (x - \xi)_+^3 = \begin{cases} (x - \xi)^3 & \text{if } x > \xi \\ 0 & \text{otherwise} \end{cases}$$

- This function is discontinuous only at the third derivative (why?)
- The coefficients are now unconstrained!
- So, in order to fit this by least squares, calculate the basis functions (including also intercept, x , x^2 and x^3), and it is business as usual.

Further Corrections

- Points at the outer range of the training data (small/large values of x) are always the bane of regression. Splines are particularly sensitive.
- The smugly named **natural spline** “gives up” on nonlinearity at the “outer” knots: a linear function is used to fit points before the first knot/after the last knot.
 - If frees up 4 degrees of freedom, which can be cashed by spending them on 4 more knots.

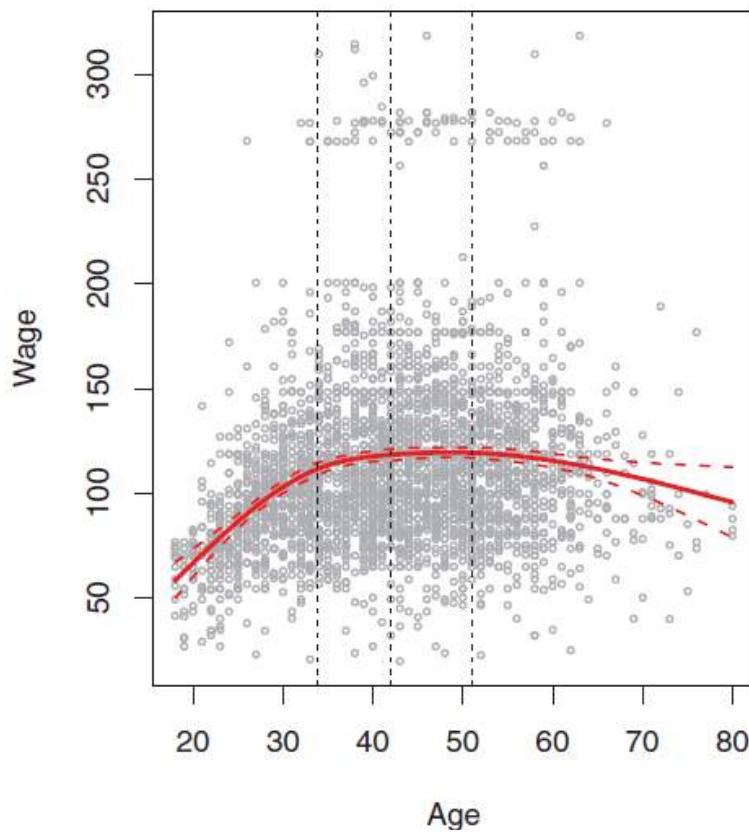
Example (3 Knots)



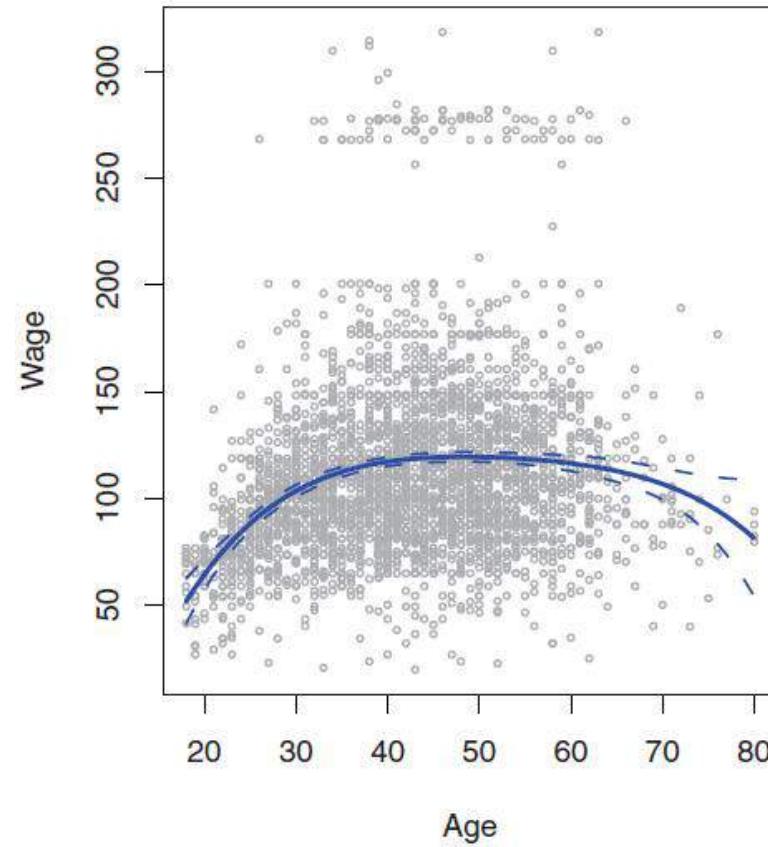
Choosing the Knots

- Intuition: putting more knots where the function is “wigglier”, fewer where it is particularly smooth.
 - But of course, we *don't know* the function. That's what we want to estimate!
- In practice, placing them uniformly in the range of x is one of the most common strategies.
 - Empirical quantiles of X are typically used.
- But how many knots?

Example: 3 Knots @ the 0.25, 0.50 (median) and 0.75 Empirical Quantiles



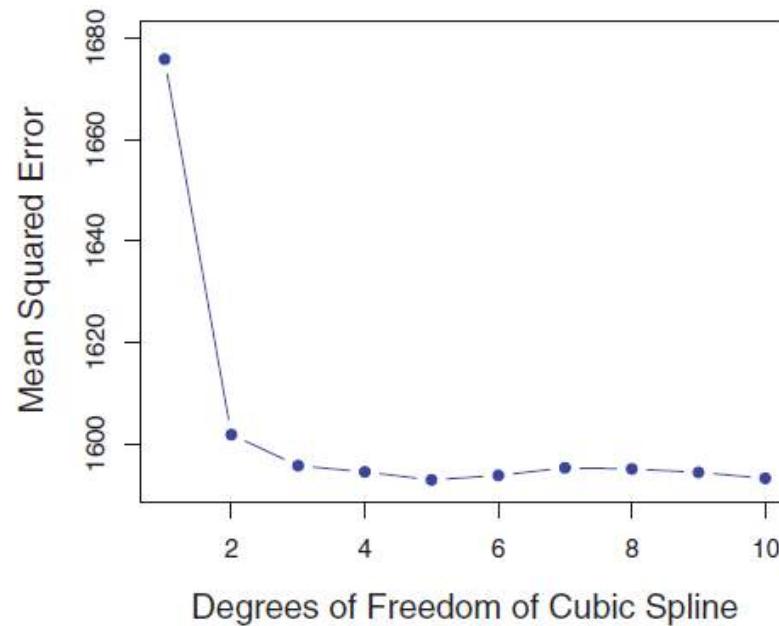
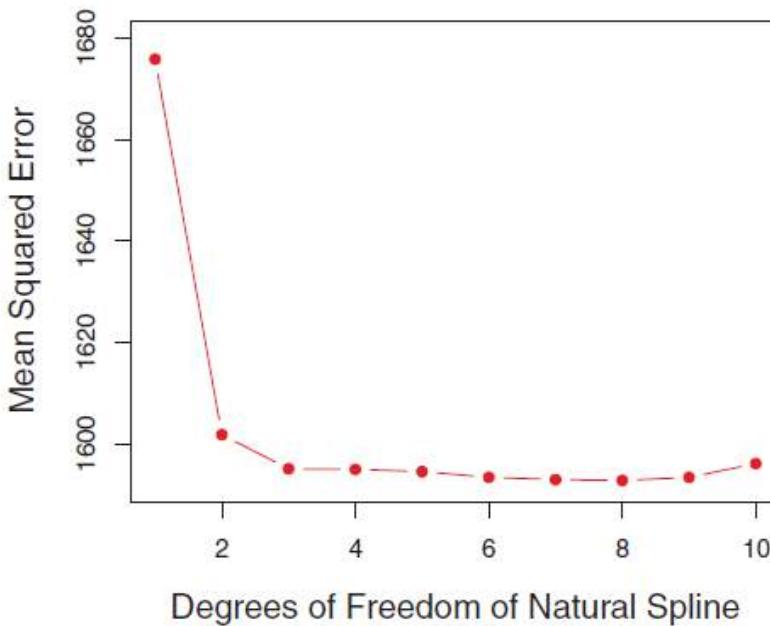
Natural Cubic Spline



Degree 4 Polynomial

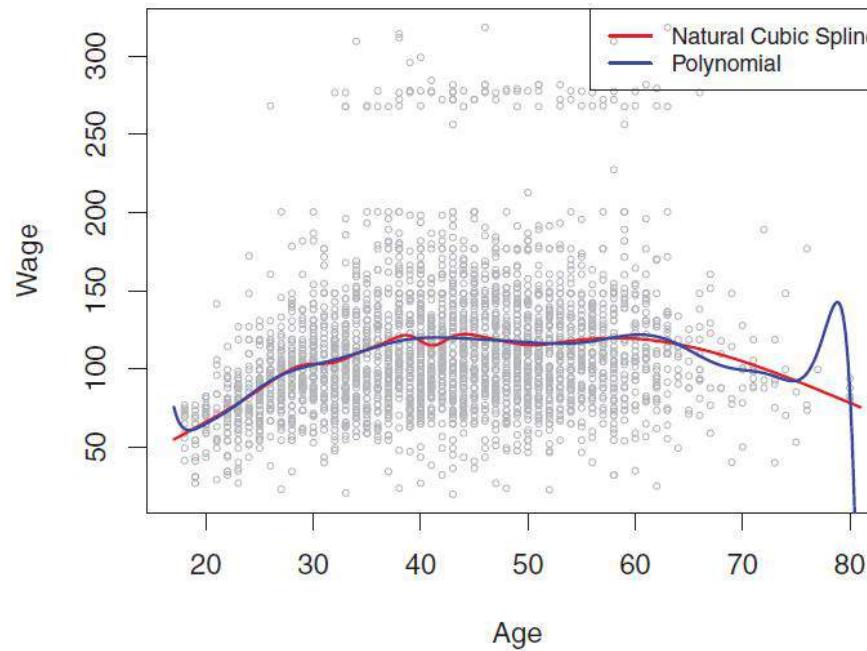
Example: Cross-Validation

- In practice, particularly when we have several inputs (details soon), it is not that uncommon to just fix the number of knots.



Comparison

- Spending 15 degrees of freedom on a natural cubic spline vs. polynomial regression.



- Representation matters!

Smoothing Splines

- Take no prisoners: *make every training point a knot.*
- How can we justify this? Cast least-squares in a very abstract sense

minimise $\sum_{i=1}^n (y^{(i)} - f(x^{(i)}))^2$ such that $f(\cdot)$ is a function

- Wait, what? I will just set $f(x^{(i)}) = y^{(i)}$, thank you very much. Not sure what will happen outside the training data.

Regularization

minimise $\sum_{i=1}^n (y^{(i)} - f(x^{(i)}))^2$ such that $f(\cdot)$ is a function that is not “too wiggly”

- Better phrased as

$$\text{minimise } \sum_{i=1}^n (y^{(i)} - f(x^{(i)}))^2 + \lambda \int f''(t)^2 dt$$

- The result of this “function optimisation” problem is a natural cubic spline with a knot on every training point!

Nonparametrics

- As hinted before, this is a genuine example of nonparametric statistics. The parameter space is **infinite**, even if the range of x was bounded.

Start with

$$\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \dots \\ \theta_K \end{bmatrix} = \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ \dots \\ f(K) \end{bmatrix}$$

But, then...

$$\begin{bmatrix} f(0) \\ f(1) \\ f(1.5) \\ f(2) \\ \dots \\ f(K) \end{bmatrix} ? \dots$$

$$\begin{bmatrix} f(0) \\ f(0.25) \\ f(0.5) \\ f(0.75) \\ f(1) \\ f(1.25) \\ f(1.5) \\ f(1.75) \\ f(2) \\ \dots \\ f(K) \end{bmatrix} ! \dots$$

Nonparametrics

- But even if the parameter is a weird infinite-dimensional “vector”, the **estimator** we get has a finite representation. It just happens to grow with the size of the training set...

$$Y = \beta_0 + \beta_1 b_1(x) + \beta_2 b_2(x) + \cdots + \beta_{K+3} b_{K+3}(x) + \epsilon$$

now $K = n$

Degrees of Freedom, Revisited

- When we introduce a regularizer, it is not a matter of parameter counting anymore.

$$\sum_{i=1}^n (y^{(i)} - f(x^{(i)}))^2 + \lambda \int f''(t)^2 dt$$

- $\lambda = 0$ corresponds to n degrees of freedom.
- $\lambda = \infty$ corresponds to 2 degrees of freedom.
- Why do we care?

Linear Smoothers

- Recall an exercise in Sheet #4 where we showed that in regression, the fitted values are weighted combinations of the training \mathbf{Y} .
- For natural cubic splines, the same applies:

$$\hat{\mathbf{f}}_\lambda = \mathbf{S}_\lambda \mathbf{y}$$

- Without doing the algebra, the diagonal entries of \mathbf{S}_λ decrease with λ . The “effective degrees of freedom” is the sum of the diagonal entries of \mathbf{S}_λ .

(See ESL, Section 5.4, if you want the gory details.)

Linear Smoothers

- What is my point? You should be aware that leave-one-out cross-validation can be computed very efficiently for linear smoothers:

$$\sum_{i=1}^n (y^{(i)} - \hat{f}^{(-i)}(x^{(i)}))^2 = \sum_{i=1}^n \left[\frac{y^{(i)} - \hat{f}_\lambda(x^{(i)})}{1 - \{\mathbf{S}_\lambda\}_{ii}} \right]^2$$

- Implications:
 - If for some reason you need to fit many nonlinear curves, a linear smoother like smoothing splines can be very handy!
 - I've once fitted 100,000s of spline models for a single research project in a humble desktop machine in a matter of hours.
 - If a piece of software is choking on delivering the fit of many linear smoothers, maybe this software is junk.
 - Also, many software packages report the “effective degrees of freedom” chosen by cross-validation instead of λ , because it is arguably more interpretable.

Local Regression

(For your reference. Not examinable.)

- Large-scale fitting is a problem with you have much data: smoothing splines cost* $O(n^3)$ computational steps.
- An alternative is to build a model around the prediction point you are interested.
 - Old-fashioned machine learning had wonderful names for that, *lazy learning* or *memory-based learning*.
- The main idea is to fit least-squares around a point by nonlinear reweighting the data around it.
 - It does require storing your training data at test time.

*There are less naïve ways of doing it, see ESL, Appendix to Chapter 5.

Algorithm

(For your reference. Not examinable.)

Algorithm 7.1 Local Regression At $X = x_0$

1. Gather the fraction $s = k/n$ of training points whose x_i are closest to x_0 .
2. Assign a weight $K_{i0} = K(x_i, x_0)$ to each point in this neighborhood, so that the point furthest from x_0 has weight zero, and the closest has the highest weight. All but these k nearest neighbors get weight zero.
3. Fit a *weighted least squares regression* of the y_i on the x_i using the aforementioned weights, by finding $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

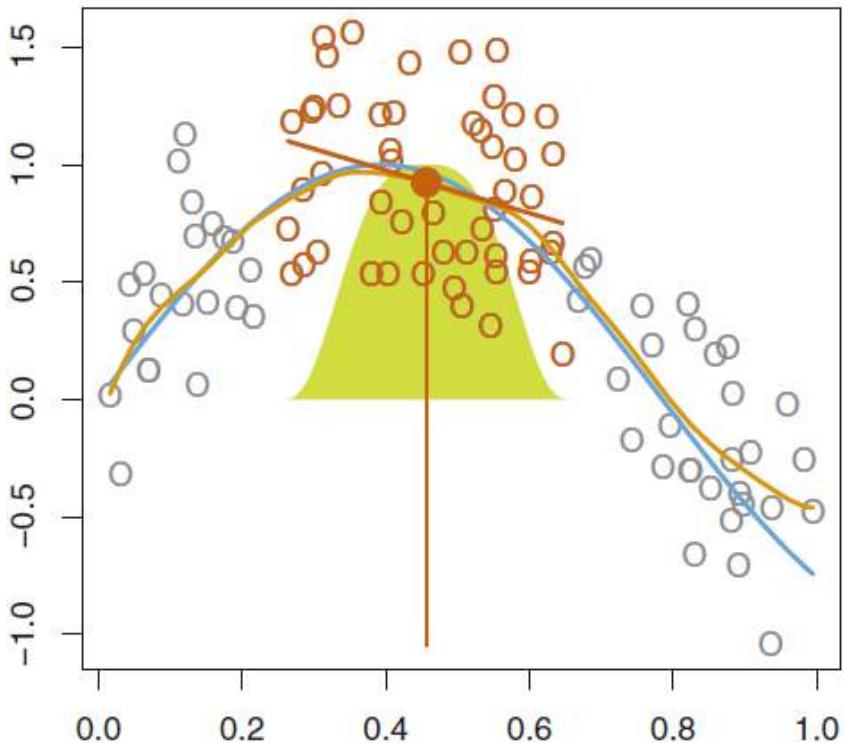
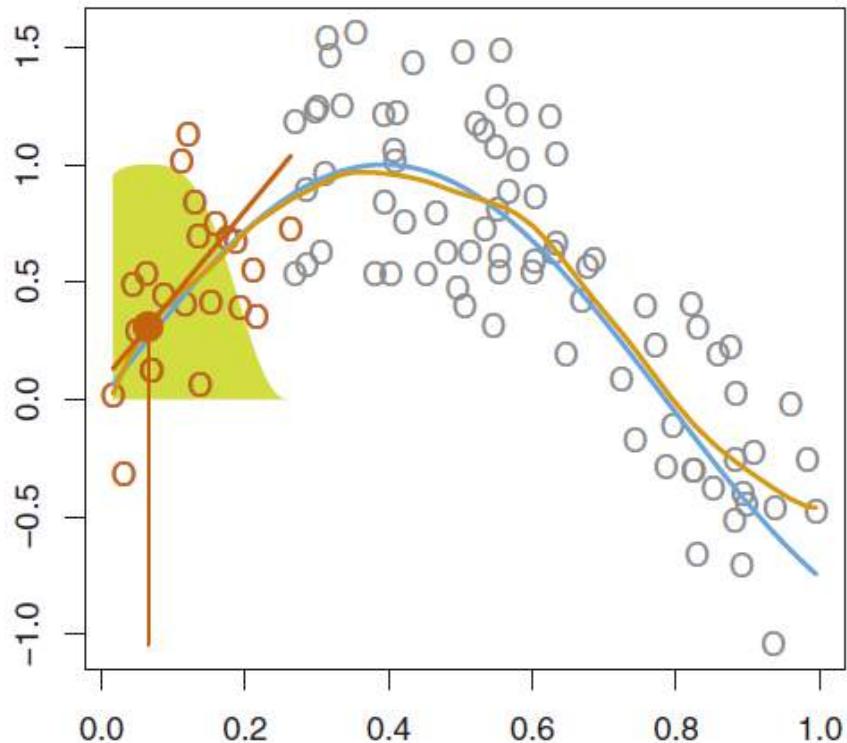
$$\sum_{i=1}^n K_{i0} (y_i - \beta_0 - \beta_1 x_i)^2. \quad (7.14)$$

4. The fitted value at x_0 is given by $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.
-

Illustration: Wage data again

(For your reference. Not examinable.)

Local Regression



Selected Topics

NONLINEAR MODELS: THE MULTIVARIATE CASE

Multidimensional Splines

- Imagine we have two covariates. We can define a **tensor product basis**:

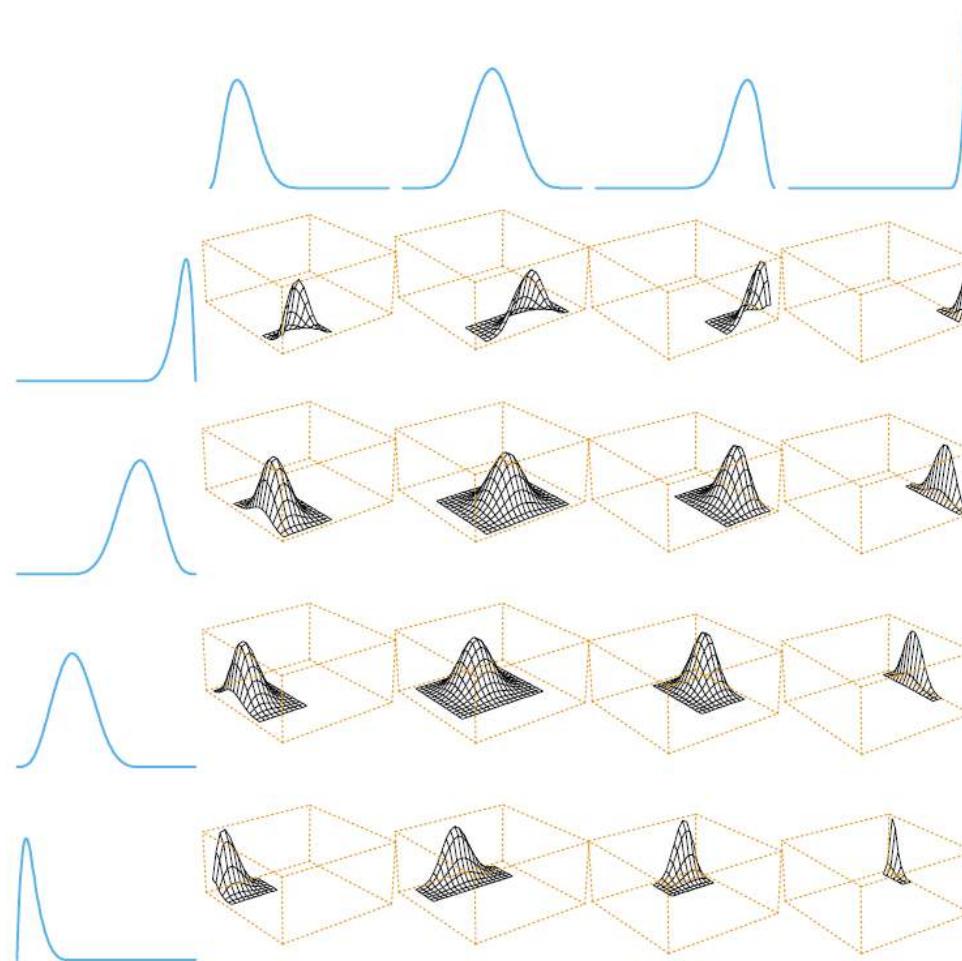
$$f_{jk}(x) = h_{1j}(x_1)h_{2k}(x_2)$$

for some $j = 1, \dots, M_1, k = 1, \dots, M_2$.

- This leads to the following parameterisation of the regression function

$$f(x) = \sum_{j=1}^{M_1} \sum_{k=1}^{M_2} \beta_{jk} g_{jk}(x)$$

Visualisation



(ESL, Chapter 5)

Problems

- The complexity of this tensor combination explodes as a function of dimensionality. Not really used even in few dimensions (e.g. 4+).
 - This is both computationally (takes too much time) and statistically (takes too much data) intractable.
 - There are alternatives: remember greedy search? It is used in the context of deciding which pieces of the tensor to be used, like in the wonderfully named MARS algorithm (ESL, Chapter 9)
- A useful compromise is **additive models**. Way fewer degrees of freedom, and also interpretable.

Generalised Additive Models

- Additivity here means on additivity on the covariates. On top of it, we can have additive errors (as in linear regression).

$$Y = \beta_0 + \sum_{j=1}^p f_j(x_j) + \epsilon$$

Typically, $\sum_{i=1}^n f_j(x_j^{(i)}) = 0$ is enforced.

Redundant?

- Algorithms for fitting these models can build upon existing methods for univariate regression.

Example: The Backfitting Algorithm

Algorithm 9.1 *The Backfitting Algorithm for Additive Models.*

1. Initialize: $\hat{\alpha} = \frac{1}{N} \sum_1^N y_i$, $\hat{f}_j \equiv 0, \forall i, j$.

2. Cycle: $j = 1, 2, \dots, p, \dots, 1, 2, \dots, p, \dots,$

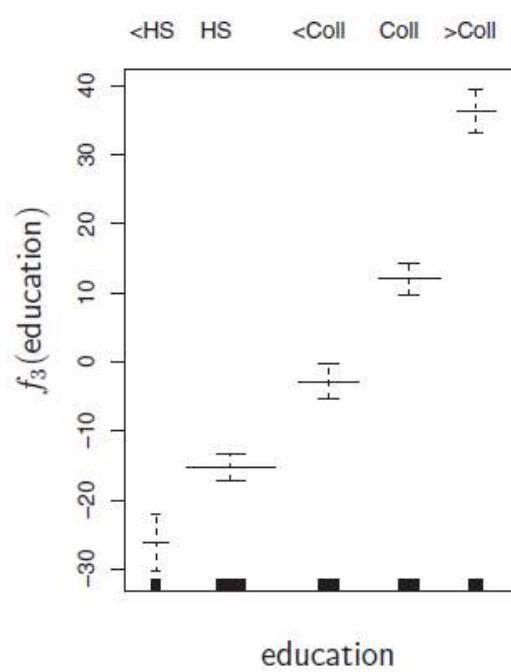
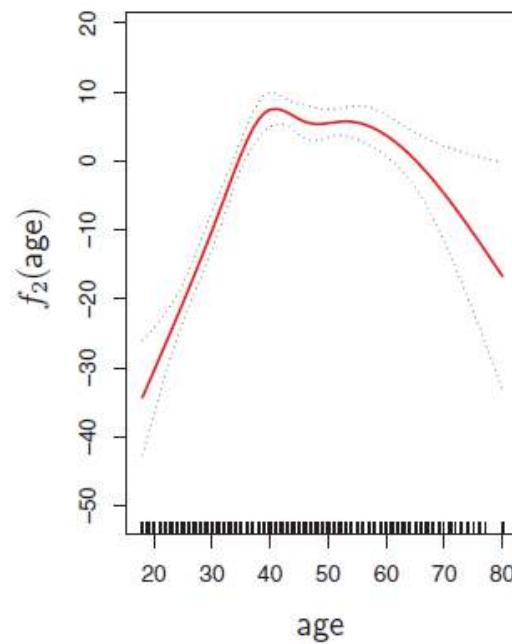
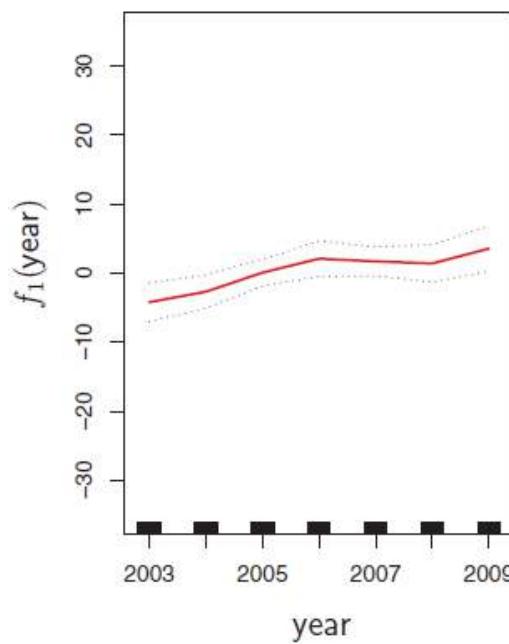
$$\hat{f}_j \leftarrow \mathcal{S}_j \left[\{y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik})\}_1^N \right],$$

Some smoother e.g.
natural smoothing splines

$$\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{N} \sum_{i=1}^N \hat{f}_j(x_{ij}).$$

until the functions \hat{f}_j change less than a prespecified threshold.

Example: Wage Data



- What is the interpretation?
- How does it differ with respect to linear models?

(ISLR, Chapter 7)

Additive Logistic Regression

- More of the same, in terms of definition.

$$\log \frac{P(Y = 1 \mid x)}{P(Y = 0 \mid x)} = \beta_0 + \sum_{j=1}^p f_j(x_j) + \epsilon$$

- Beware of interpretation and algorithm. The idea now is similar to Newton-Raphson, iterated per covariate.
 - Those taking *STATG003* will have a closer look into Newton-Raphson for GLMs in Lab 8.

Example: Backfitting for Additive Logistic Regression

Algorithm 9.2 Local Scoring Algorithm for the Additive Logistic Regression Model.

1. Compute starting values: $\hat{\alpha} = \log[\bar{y}/(1 - \bar{y})]$, where $\bar{y} = \text{ave}(y_i)$, the sample proportion of ones, and set $\hat{f}_j \equiv 0 \forall j$.
2. Define $\hat{\eta}_i = \hat{\alpha} + \sum_j \hat{f}_j(x_{ij})$ and $\hat{p}_i = 1/[1 + \exp(-\hat{\eta}_i)]$.

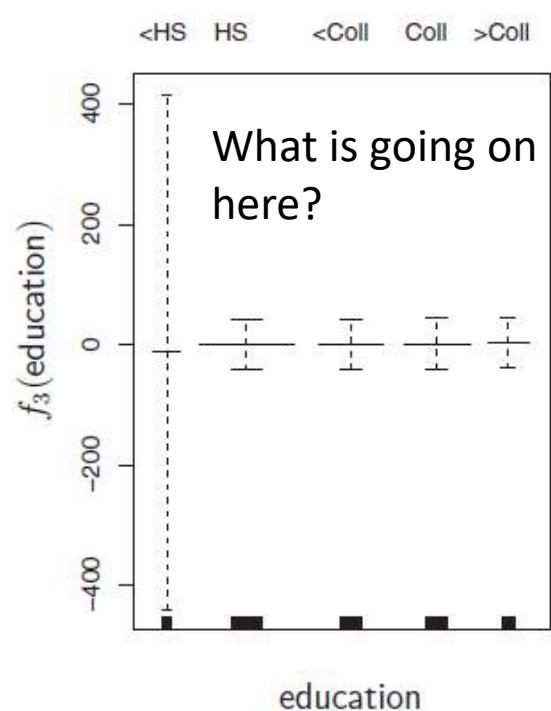
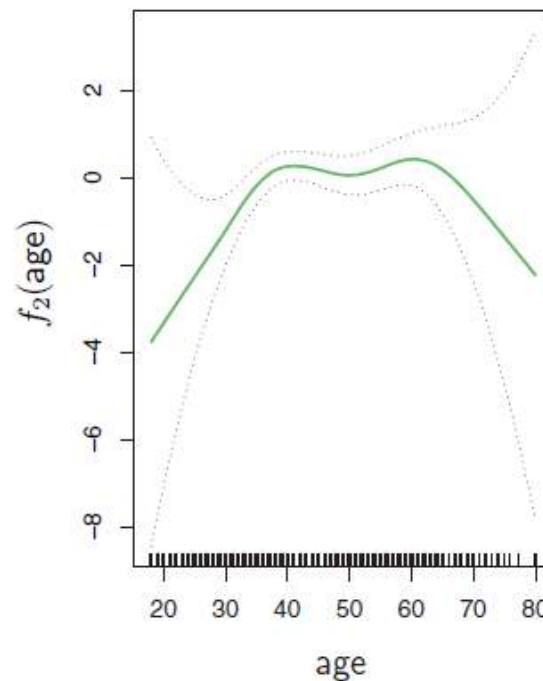
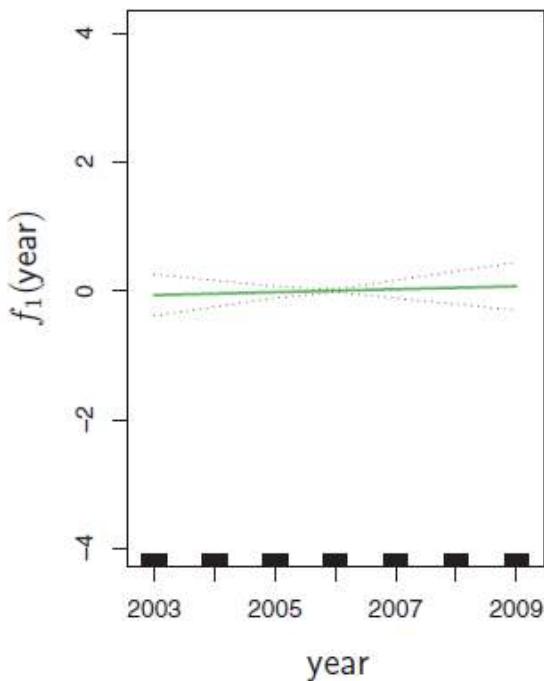
Iterate:

- (a) Construct the working target variable

$$z_i = \hat{\eta}_i + \frac{(y_i - \hat{p}_i)}{\hat{p}_i(1 - \hat{p}_i)}.$$

- (b) Construct weights $w_i = \hat{p}_i(1 - \hat{p}_i)$
 - (c) Fit an additive model to the targets z_i with weights w_i , using a weighted backfitting algorithm. This gives new estimates $\hat{\alpha}, \hat{f}_j, \forall j$
3. Continue step 2. until the change in the functions falls below a pre-specified threshold.

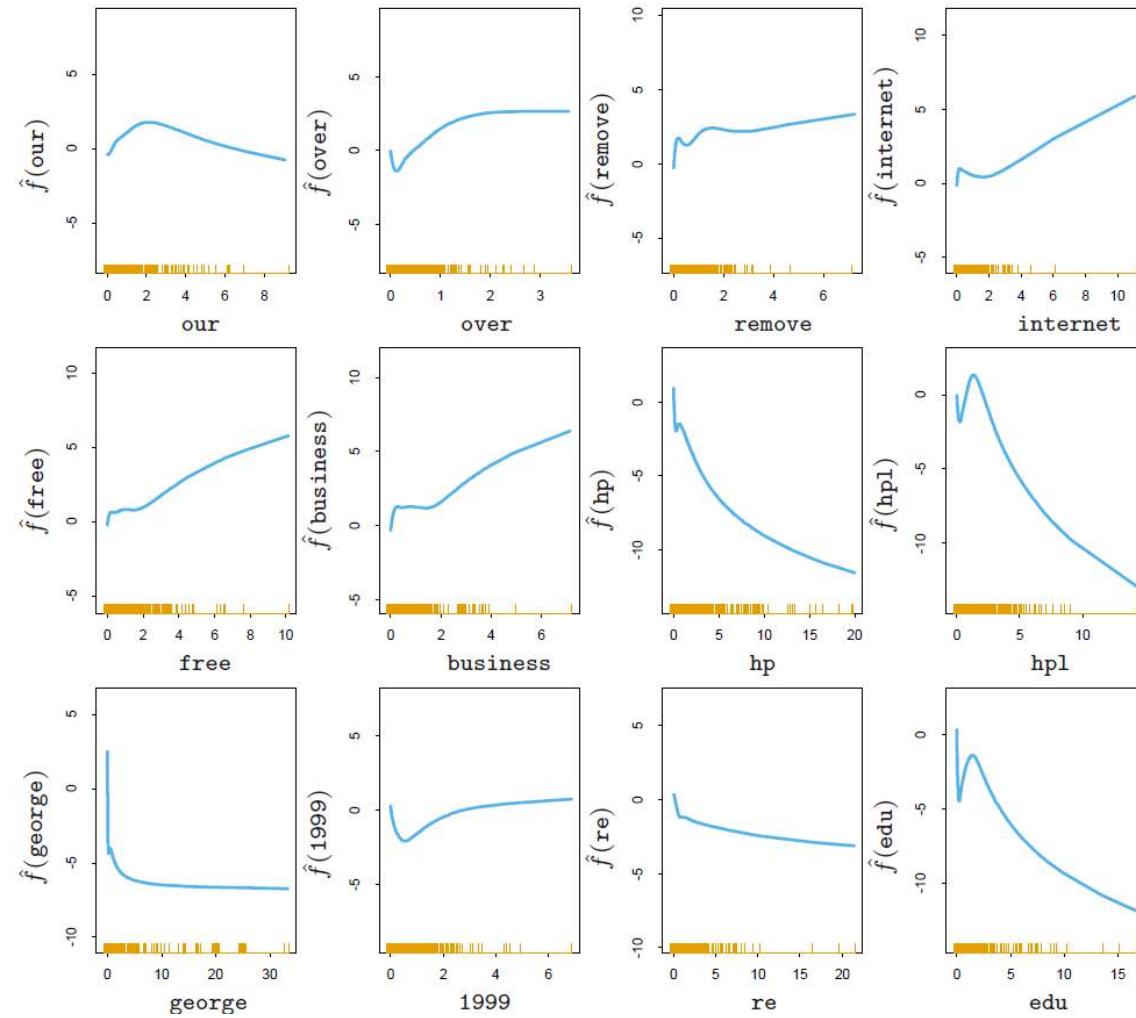
Example: Wage > 250?



- What is the interpretation?
- How does it differ with respect to linear models?

(ISLR, Chapter 7)

Example: Spam Filtering



Sparse Additive Models

- A “SpAM” of another type.
- Surely you saw this coming:

$$\sum_{i=1}^n (y^{(i)} - \sum_{j=1}^J f_j(x_j^{(i)}))^2 + \lambda \sum_{j=1}^J \sqrt{\sum_{i=1}^n f_j(x^{(i)})^2}$$

- Notice this is a type of group lasso: each evaluation of f_j , across the data points, is shrunk together.

Backfitting Revisited

- Before we had the “smooth the residuals step”:

$$\hat{f}_j = \mathcal{S}_j \left(\mathbf{y} - \sum_{k \neq j} \hat{\mathbf{f}}_k \right)$$

- Now (without getting in details), we have the soft threshold variation

$$\hat{f}_j = \left(1 - \frac{\lambda}{\sqrt{\sum_{i=1}^n \tilde{f}_j(x^{(i)})^2}} \right)_+ \tilde{f}_j, \text{ where } \tilde{f}_j = \mathcal{S}_j \left(\mathbf{y} - \sum_{k \neq j} \hat{\mathbf{f}}_k \right)$$

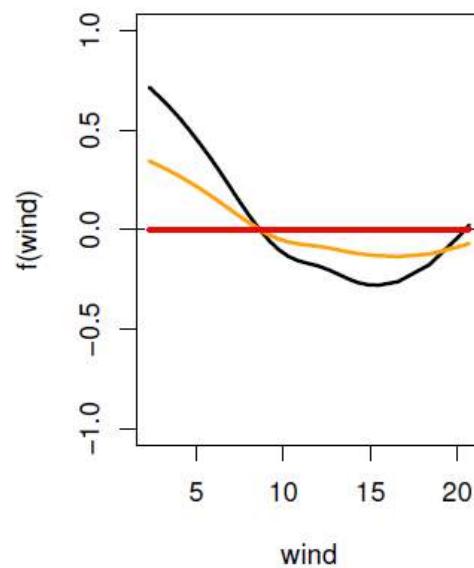
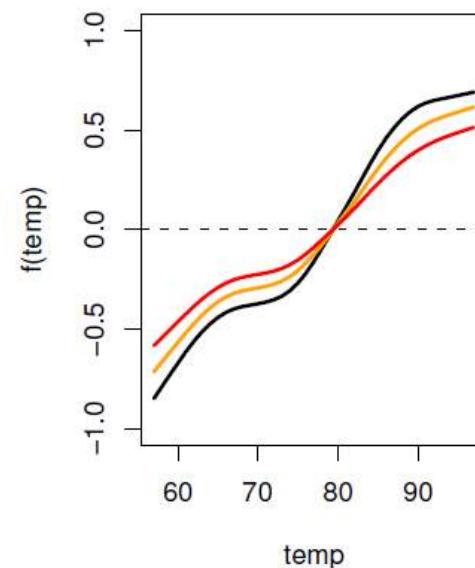
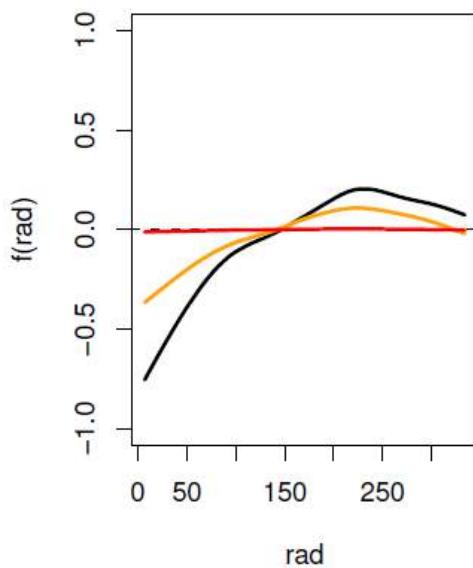
Example: Log-Ozone Concentration

$$\lambda = 0$$

$$\lambda = 2$$

$$\lambda = 4$$

$$\log(\text{ozone}) \sim s(\text{rad}) + s(\text{temp}) + s(\text{wind})$$



Take-Home Messages

- Model search: two views
 - Finding real structure in nature (even if in practice it might be approximate)
 - Shrinkage/regularization: just minimise the effect of overfitting
- Computational considerations: l_0 , l_1 , l_2 (and combinations) have their disadvantages and advantages.
 - Pragmatic advice: elastic net might be a good starting point.

Take-Home Messages

- Nonlinear regression/classification/etc. assumes many facets.
- The spline/additive approach is very popular in the Statistics community because of:
 - Computational simplicity
 - In many cases, confidence intervals are easy to calculate
 - Many approaches boil down to least-squares with data-independent regularization!
- Additivity is certainly not the greatest assumption, but it can be as good as it gets in problems with many variables and not “that much” data.

Take-Home Messages

- Much of this chapter also illustrates the development of statistical thinking: a mix-and-match combination of many fundamental building blocks.
- R packages such as **gam** and **glmnet** can go a long way in your practical applications, even if these are not particularly optimised pieces of code.
- Sparse/(Generalised) Additive Models complement nicely the Machine Learning toolbox of SVMs, neural nets etc. particularly for interpretation purposes and the selection of variables. But don't rely on additivity for additivity's sake.

Introduction to Statistical Data Science

Ricardo Silva

ricardo@stats.ucl.ac.uk

Department of Statistical Science, UCL

Unsupervised Learning: a Statistical View

Outline

- Unsupervised learning is often ill-defined, but it is basically the problem of inferring “interesting features” in the distribution of one variable or several variables.
- There is no outcome variable to be predicted.

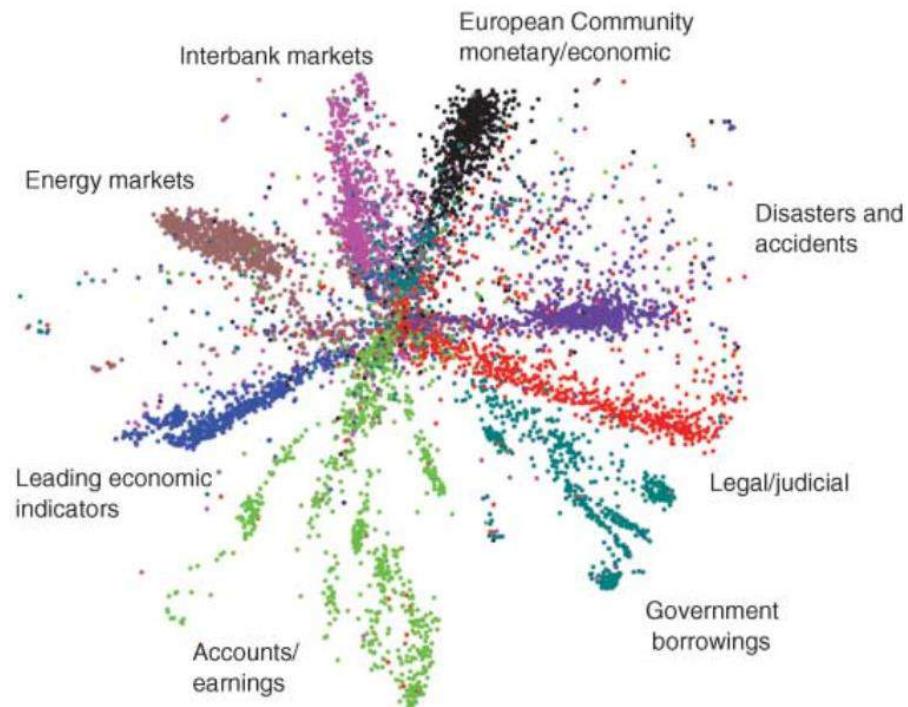
A Machine Learning Example

- Retrieval by content: find me stories similar to a given story
 - Story == news article == text document

Two-dimensional representation of a corpus of news articles, inferred from nothing but raw text. Labels shown were assigned by a human, but not used in the method!

“Similarity” is now given by Euclidean distance in this space.

Hinton and Salakhutdinov (2006). “Reducing the Dimensionality of Data with Neural Networks”. *Science*.



More Standard Statistical Applications

- Is this data point “atypical”? What is typical and what is not? I need a pdf for that.
- Which dependencies exist in my system? Can I infer which proteins in a cell “talk to each other” directly?
- I have too many variables. Can I “preserve information” in my data using fewer variables, regardless of a possible prediction problem?
- I may only see part of my data: I’m missing some data that I postulate to exist but I cannot measure.

Outline

- (Nonparametric) Density estimation
- Multivariate models, with a focus on the Gaussian
- Dimensionality reduction
- Latent variable and mixture models, with applications to clustering

Unsupervised Learning

NONPARAMETRIC DENSITY ESTIMATION

Models and Likelihoods

- The problem of **density estimation** is the problem of estimating the probability distribution of a population.
 - Sometimes the term “density estimation” is used even if the data is discrete (by machine learning people, mostly)
- We saw likelihood functions, mainly in the context of regression and some simple models. Like this:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

- We can use models to characterize observations as being “likely” or not.

Example: Outlier Detection

- Let's resurrect our old NHANES data!
 - An American “Health and Nutrition Examination Survey”.
- Is an individual of height 1.95m “too unusual” for the population we are studying?
 - For instance, if we were looking at kids of a particular age, a height that is too unusual might indicate problems with the data.

Gaussian Density Estimation

- We fit parameters by maximum likelihood.

$$\hat{\mu} \approx 168, \quad \hat{\sigma} \approx 10.2$$

- Then we calculate tail area probabilities:

$$P(Y > 190; \mu = 168, \sigma = 10.2) \approx 0.02$$

- Is this an unlikely event? This is a problem-dependent conclusion (just like p-values).
Regardless of it, you can (and must) explain how you got to that probabilistic statement.

Nonparametric Estimate

- Don't trust the Gaussian? What about using the empirical distribution?

$$P(Y > 190) \approx \frac{\#\text{people in sample with height greater than } 190}{n} \approx 0.0001$$

- The difference in estimate may or may not matter, depending on the application (see also, "value at risk").
- Avoiding the Gaussianity assumption is nice, but the empirical distribution itself has its shortcomings: it all boils down to the bias-variance trade-off.
 - Gaussian: (possibly) "high" bias, "low" variance
 - Empirical distribution: low bias, "high" variance

Alternative: the Histogram Estimate

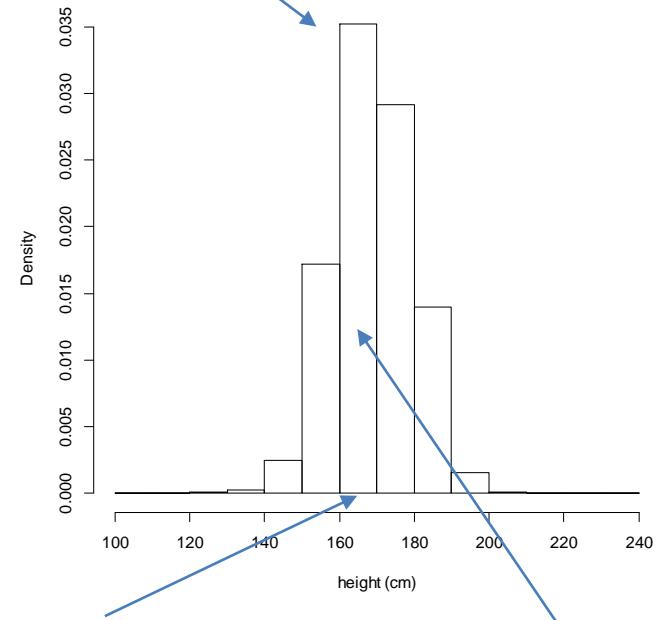
- Break the data in bins, calculate frequency of points falling in each bin. This will smooth the empirical distribution:
 - few bins = lot of smoothing (R demo)
- There is no likelihood function: divide the range of your distribution into m bins of length h .
 - For simplicity, assume we know the maximum and minimum of our space (say 0 and 2.5 in the NHANES data, so $h = 2.5 / m$)
 - For each bin B_k , compute the proportion of points which fall in B_k .

Example: NHANES Height

$$\hat{p}_k = \frac{\text{\#points falling in bin } B_k}{n}$$

$$\hat{p}(x) = \begin{cases} \hat{p}_1/h & x \in B_1 \\ \hat{p}_2/h & x \in B_2 \\ \dots \\ \hat{p}_m/h & x \in B_m \end{cases}$$

height is \hat{p}_k/h



width is $h = 10$

area is $\hat{p}_k/h \times h = \hat{p}_k$

Choosing m

- Cross-validation, duh!
- But what is the measure we are optimising?
There is no outcome variable, and no likelihood.
- Option: mean-squared on the density.

$$\begin{aligned} L(m) &= \int (\hat{p}(x) - p(x))^2 dx \\ &= \int \hat{p}^2(x) dx - 2 \int \hat{p}(x)p(x)dx + \int p^2(x)dx \end{aligned}$$

Does not depend on m

$$\approx \int \hat{p}^2(x)dx - 2 \times \frac{1}{n} \sum_{i=1}^n \hat{p}_{(-i)}(x^{(i)}) + \text{constant}$$

Easy to solve

This substitutes $p(x)$.

Histogram fit without point (i)

Confidence ~~Intervals~~ Bands

- We could look at each value x and build a confidence interval for $p(x)$, but more generally we would like to bound the entire function at one, regardless of x . That is, find some $l(x)$, $u(x)$ such that

$$P(l(x) \leq p(x) \leq u(x) \text{ for all } x) \geq 1 - \alpha$$

for a given α , where the probability is over the datasets that can be used to build $l(x)$ and $u(x)$.

Confidence Bands

- For a fixed number m of bins, we can get a confidence band for a “histogramized” version of $p(x)$, which we will call $\tilde{p}(x)$,

$$\tilde{p}(x) \equiv \frac{p_k}{h} \text{ for } x \in B_k, \text{ where } p_k \equiv \int_{B_k} p(x) dx$$

- Bizarre! What is this? I want my confidence band for $p(x)$!
 - Thought luck.
 - There are limits of what we can do without a likelihood.
 - The (informal) idea is that as the data grows, cross-validation will pick higher and higher values for m , so in principle we can get arbitrarily close to $p(x)$ if we feed in “enough” data.
 - The amount of uncertainty may be humbling (R demo).

(See Chapter 20 of Wasserman, if you want a formula and gory details.)

Kernel Density Estimation

- Histograms are simple, but discontinuous.
Kernel density estimation provides a smoother alternative.
- Function K is a kernel if:

$$K(x) \geq 0$$

$$\int K(x)dx = 1$$

Basically, a density function of zero mean and positive variance.

$$\int xK(x)dx = 0$$

$$\int x^2K(x)dx > 0$$

Estimate

- Given a kernel, the estimator requires a choice of scaling parameter (called **bandwidth**):

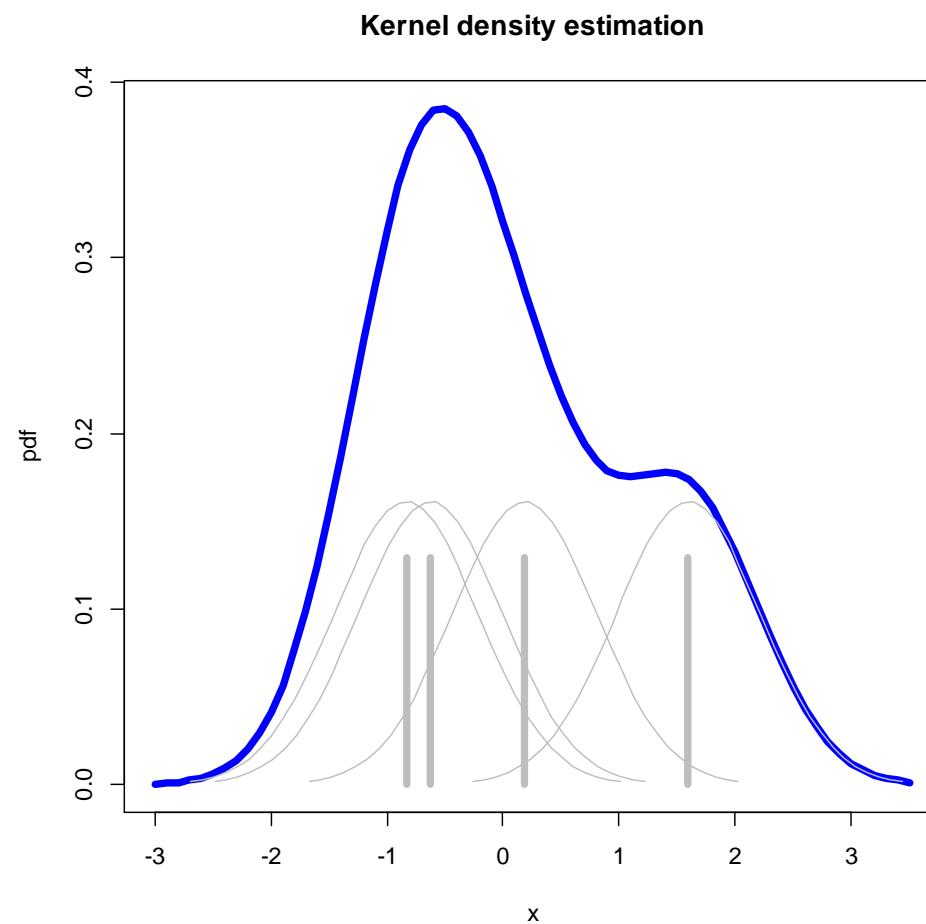
$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - x^{(i)}}{h}\right).$$

- One possible choice of kernel is the standard Gaussian:

$$K(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Example

- Gaussian kernel, selection of bandwidth by cross-validation.



R demo: effect of bandwidth selection

Confidence Bands

- Same issue as histograms: confidence for a given bandwidth h , meaning confidence over a smoothed version of the truth.

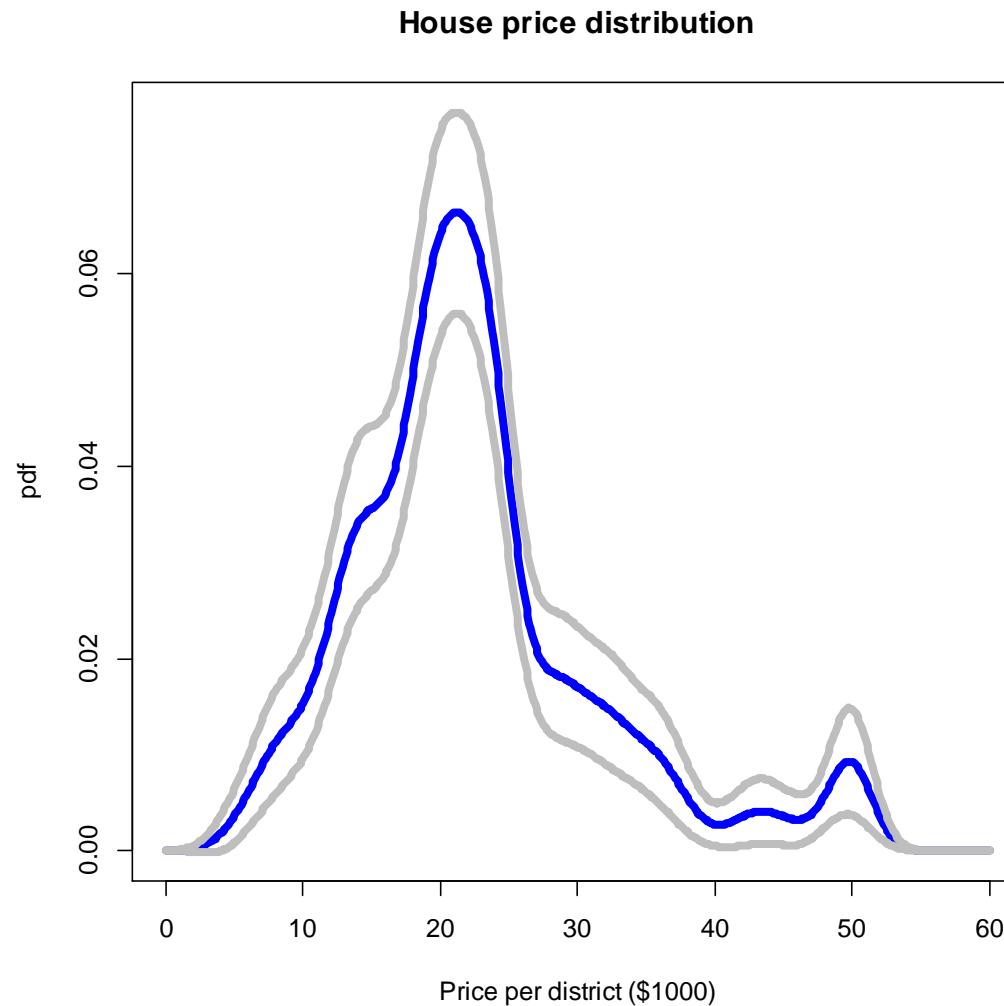
$$\tilde{p}(x) \equiv \int \frac{1}{h} K\left(\frac{x-u}{h}\right) p(u) du$$

Actual density

- As in the histogram case, the idea is that cross-validation gives $h \rightarrow 0$ as $n \rightarrow \infty$.

(Again, for those interested, Chapter 20 of Wasserman gives details.)

Example



Multivariate Density Estimation

- The kernel idea again applies to p -dimensional vectors \mathbf{x} :

$$\hat{p}(\mathbf{x}) \equiv \frac{1}{n} \sum_{i=1}^n K_h(\mathbf{x} - \mathbf{x}^{(i)})$$

$$K_h(\mathbf{x} - \mathbf{x}^{(i)}) = \frac{1}{nh_1 h_2 \dots h_p} \left\{ \prod_{j=1}^p K\left(\frac{x_j - x_j^{(i)}}{h_j}\right) \right\}$$

- Is there a catch? Of course there is a catch.

The Curse of Dimensionality

- On your right, the sample sizes required to ensure, as the dimension increases, a MSE less than 0.1 at **0**, when data comes from a multivariate Normal but you do not assume it.
- For this problem, this is like saying that having 800,000+ observations in a ten dimensional problem is as good as having 4 in one dimension!

Dimension	Sample Size
1	4
2	19
3	67
4	223
5	768
6	2790
7	10,700
8	43,700
9	187,000
10	842,000

Unsupervised Learning

BUILDING AND ESTIMATING MULTIVARIATE DISTRIBUTIONS

Modelling Joint Distributions

- You will hardly be able to learn some big distribution $P(X_1, X_2, \dots, X_p)$ with a purely nonparametric approach.
- A combination of domain knowledge and off-the-shelf ideas can go a long way.
- In what follows, we describe:
 - An example of a recipe for defining joint distributions.
 - The great classical example: the multivariate Gaussian in more detail.
 - Structured Gaussians: identifying independencies.
 - An example of how to exploit the multivariate Gaussian to build other distributions.

How to Build a Distribution over Multiple Variables?

- A blast from the past:

- (h) Suppose that the number of distinct uranium deposits in a given area is a Poisson random variable with parameter $\mu = 10$. If, in a fixed period of time, each deposit is independently discovered with probability $1/50$, find the probability that (i) exactly one, (ii) at least one and, (iii) at most one deposit is discovered during that time.

(Exercise sheet #1)

- We have D as the number of deposits and Y as the number of deposits discovered.

$$p(d, y) = p(d)p(y \mid d)$$

Maximum Likelihood

- If we have a dataset $(d^{(1)}, y^{(1)}), \dots (d^{(n)}, y^{(n)})$, how do we do maximum likelihood? Same old.

$$\log \prod_{i=1}^n p(d^{(i)}, y^{(i)}) = \log \prod_{i=1}^n p(d^{(i)}; \theta_1) p(y^{(i)} | d^{(i)}; \theta_2)$$

$$\sum_{i=1}^n \log p(d^{(i)}; \theta_1) + \sum_{i=1}^n \log p(y^{(i)} | d^{(i)}; \theta_2)$$

We can optimise these separately,
using the standard stuff

In some models, this could
be a GLM or a GAM, for instance

- This can be generalised to more variables.
 - See *COMP GI08* for the juicy details.
 - If you choose to go Bayesian next term, *STAT G004* will give you plenty of these too.

“Canonical” Models

- Where could we find the “natural” (“canonical”) generalisation of models we have seen?
 - Multivariate Binomial?
 - Multivariate Poisson?
 - Multivariate Gaussian?...
- Surprisingly few exist! No single agreeable way of defining a multivariate Poisson, for instance.

Multivariate Binomial

- Hinted at when we mentioned contingency tables (Chapter 2).
- Say you have a dataset of three variables
 - X_1 = did customer buy milk? Yes/No (0/1)
 - X_2 = did customer buy bread? Yes/No (0/1)
 - X_3 = did customer buy coffee? Yes/No (0/1)
- Which parameters?
 - Literally, for each $p(x_1, x_2, x_3)$, have a $\theta_{x_1 x_2 x_3}$.
 - Notice that

$$0 \leq \theta_{ijk} \leq 1, i \in \{0, 1\}, j \in \{0, 1\}, k \in \{0, 1\}$$

$$\sum_{i=0}^1 \sum_{j=0}^1 \sum_{k=0}^1 \theta_{ijk} = 1$$

Multivariate Gaussian

- Our main focus in this chapter (hinted at before in Chapter 1 and Exercise Sheet #4, we never got in details).
- The pdf of a p -dimensional Gaussian consists of a $p \times 1$ **mean vector** μ and a $p \times p$ **covariance matrix** Σ .

$$p(x_1, \dots, x_p) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})/2}$$

Example: Bivariate Gaussian with Zero Mean and Unit Variances

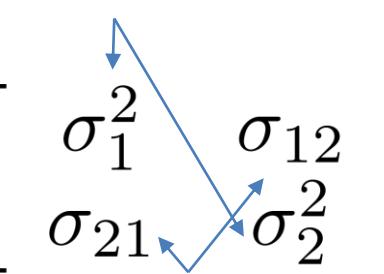
$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{Example: zero means}$$
$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{21} & 1 \end{bmatrix}$$

Variances

Covariances: they are always symmetric.

Explicit representation of symmetry.

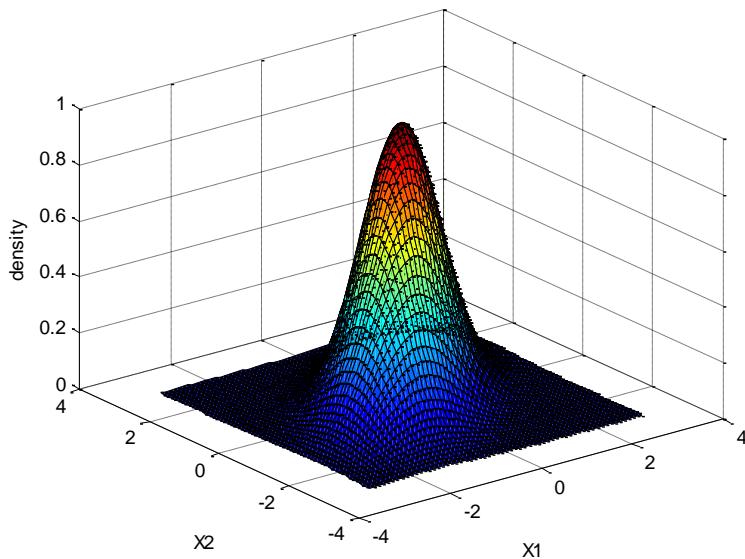
Example: unit variances



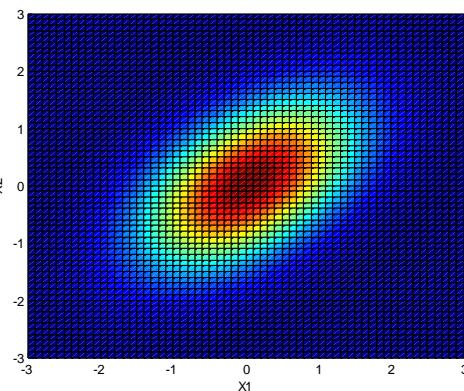
- Notice that Σ is a symmetric matrix:
 - $\sigma_{jk} = \sigma_{kj}$ for any two variables X_j and X_k .

Examples

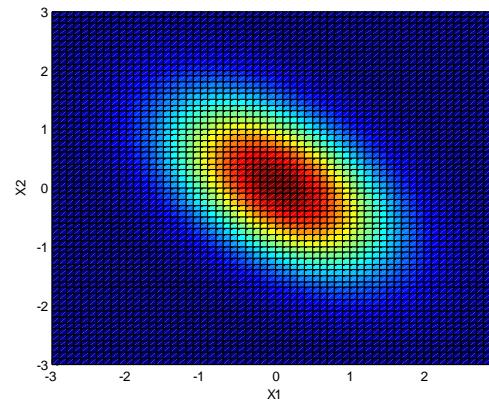
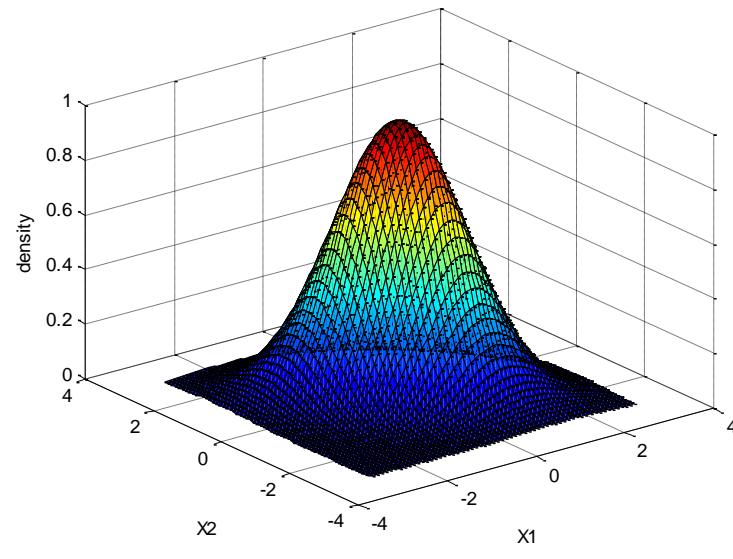
Zero mean, unit variance, $\sigma_{12} = 0.5$



Contour plot

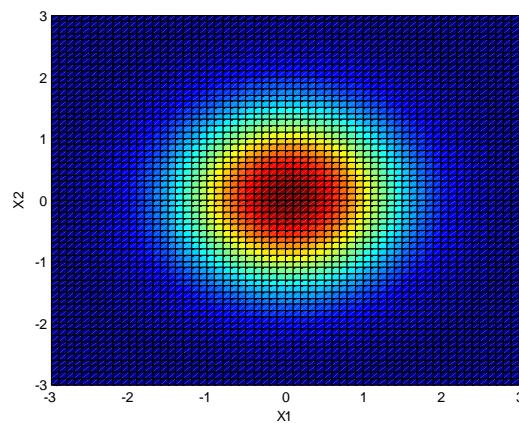
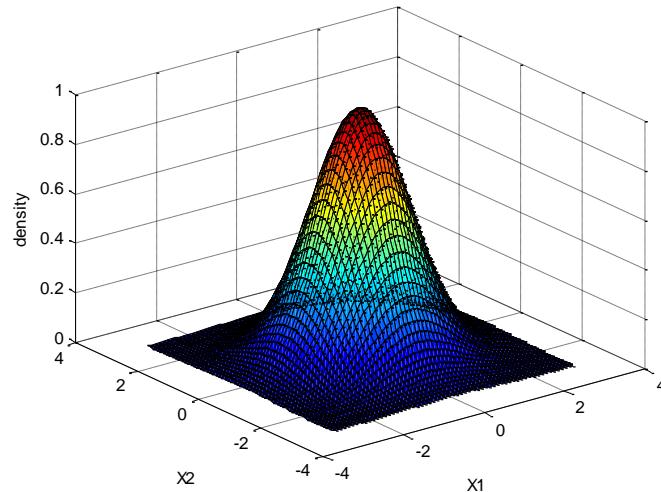


Zero mean, unit variance, $\sigma_{12} = -0.5$

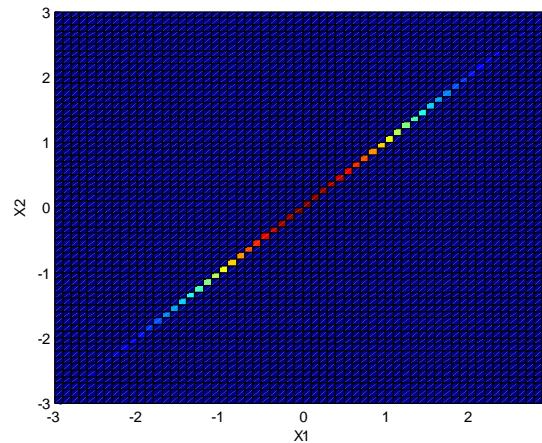
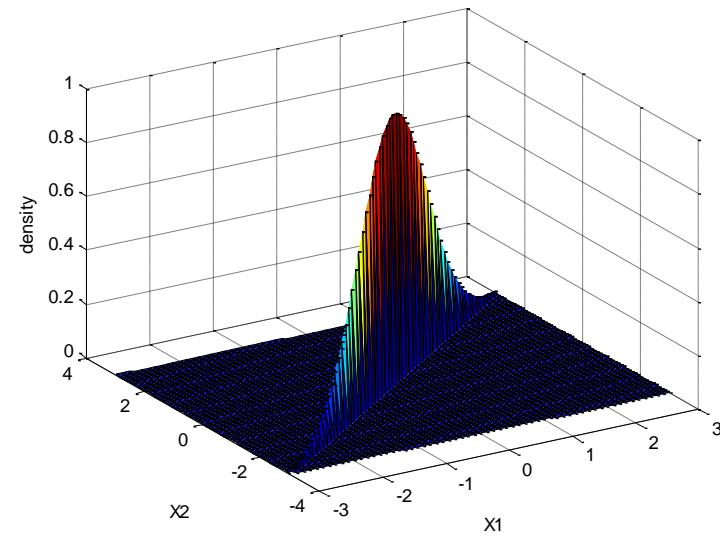


Examples

Zero mean, unit variance, $\sigma_{12} = 0$



Zero mean, unit variance, $\sigma_{12} \approx 1$



Properties

- **Marginalisation** gives back univariate Gaussians:

$$(X_1, X_2) \sim N(\mu, \Sigma) \Rightarrow X_1 \sim N(\mu_1, \sigma_1^2)$$

- As a matter of fact, any subset of a Gaussian vector \mathbf{X} will follow the corresponding Gaussian marginal:

$$(X_1, X_2, X_3) \sim N(\mu, \Sigma) \Rightarrow (X_1, X_3) \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{13} \\ \sigma_{31} & \sigma_3^2 \end{bmatrix}\right)$$

Properties

- **Conditioning** also gives back Gaussians,

$$(X_1, X_2) \sim N(\mu, \Sigma) \Rightarrow X_1 \mid X_2 \sim N(\mu_{1.2}, \sigma_{1.2}^2)$$

where $\mu_{1.2}$ $\sigma_{1.2}^2$ are the mean and variance of X_1 in the pdf given by

$$p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)}.$$

- Without getting in details, we get exactly what we would get by least-squares regression of X_1 on X_2 !
 - The idea is analogous when conditioning on multiple variables.

Properties

- **Linearly combining** a multivariate Gaussian vector gives back another Gaussian vector:

$$\mathbf{X} \sim N(\mu, \Sigma) \Rightarrow \mathbf{AX} \sim N(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^T)$$


 An arrow points from "Arbitrary $k \times p$ matrix" to \mathbf{A} . Another arrow points from "A k dimensional multivariate Gaussian" to $N(\cdot, \cdot)$.

Arbitrary $k \times p$ matrix A k dimensional multivariate Gaussian

- For example, let $(\hat{\beta}_0, \hat{\beta}_1) \sim N(\mu^\beta, \Sigma^\beta)$. Then,

$$\hat{Y} \equiv \hat{\beta}_0 + \hat{\beta}_1 x = [1 \ x] \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \Rightarrow \hat{Y} \sim N(\mu_0^\beta + \mu_1^\beta x, \sigma_0^{2\beta} + 2\sigma_{01}^\beta x + \sigma_1^{2\beta})$$

e Gaussian
constant
univariate Gaussian
“A” “X”

Interpretation and Estimation

- Just like variance, the name “covariance” applies both to the following particular summary of any distribution and parameters of a multivariate Gaussian:

$$\sigma_{ij} \equiv E[(X_i - \mu_i)(X_j - \mu_j)]$$

- Maximum likelihood for a multivariate Gaussian gives

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n x_j^{(i)} \quad \hat{\sigma}_{jk} = \frac{1}{n} \sum_{i=1}^n (x_j^{(i)} - \hat{\mu}_j)(x_k^{(i)} - \hat{\mu}_k)$$

Tests of Interest

- A common test of interest in a multivariate distribution is whether some variables are (conditionally) independent. Using the symbol $\perp\!\!\!\perp$, recall that:

$$X_j \perp\!\!\!\perp X_k \Rightarrow p(x_j \mid x_k) = p(x_j)$$

- For conditional independence:

$$\begin{aligned} X_j \perp\!\!\!\perp X_k \mid \{X_{s_1}, \dots, X_{s_v}\} &\Rightarrow \\ p(x_j \mid x_k, x_{s_1}, \dots, x_{s_v}) &= p(x_j \mid x_{s_1}, \dots, x_{s_v}) \end{aligned}$$

Independence in Gaussians

- It is true in general that

$$X_j \perp X_k \Rightarrow \sigma_{jk} = 0$$

- Moreover, if (X_j, X_k) is bivariate Gaussian then

$$\sigma_{jk} = 0 \Rightarrow X_j \perp X_k$$

- The intuition why this is not true in general:
 - Think of the contour plots of the Gaussian always showing “linear relationships”. For non-Gaussian variables, non-linear relationships can hold so that covariances might still be zero.

Testing for Zero Covariance (Independence in the Gaussian Case)

- We can test for zero covariance (= independence, for the Gaussian):

$$H_0 : \hat{\sigma}_{jk} = 0$$
$$H_1 : \hat{\sigma}_{jk} \neq 0$$

- Sometimes this is described as tests of **zero correlation** instead, which is equivalent to zero covariance. (Population) correlation and (sample) correlation are defined as:

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_j^2} \sqrt{\sigma_k^2}}$$

Always between -1 and 1

$$\hat{\rho}_{jk} = \frac{\hat{\sigma}_{jk}}{\sqrt{\hat{\sigma}_j^2} \sqrt{\hat{\sigma}_k^2}}$$

Recall this
from Exercise Sheet #4

Vanishing Partial Correlations

- To unclutter the following, define $\mathbf{X}_S \equiv \{X_{s_1}, \dots, X_{s_v}\}$.
- To test $X_j \perp\!\!\!\perp X_k \mid \mathbf{X}_S$, we can test for vanishing **partial correlations** $\rho_{jk.S}$ via

$$\begin{aligned} H_0 : \quad & \hat{\rho}_{jk.S} = 0 \\ H_1 : \quad & \hat{\rho}_{jk.S} \neq 0 \end{aligned}$$

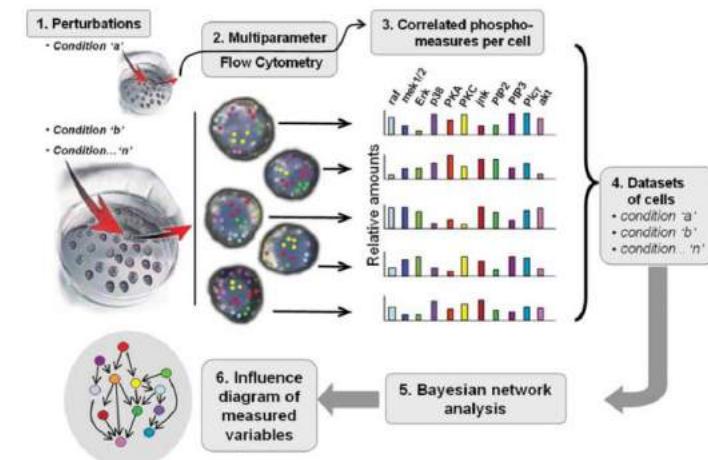
where $\rho_{jk.S}$ is the correlation between X_j and X_k in the distribution given by

$$p(x_j, x_k \mid x_{s_1}, \dots, x_{s_k}).$$

- There is a formula for $\rho_{jk.S}$, but it is ugly and not illuminating. Interestingly, it is closely related to the least-squares regression of X_j on X_k and X_S (or X_k on X_j and X_S)!

Application

- Single-cell data by Sachs et al. (2005).
- The goal is to infer some “dependency structure” among 11 molecular components in human immune system cells (proteins and lipids).
- This may lead to conclusions about the causal regulatory structure of this type of cell, under further assumptions.



Sachs et al. (2005). “Causal protein-signaling networks derived from multiparameter single-cell data”. *Science*.

Problem Setup

- We want to infer whether the concentrations of two molecules are independent, given everybody else.
- Independence will be assumed to be the same as zero partial correlation after transforming variables to be “Gaussian-like”.
 - This is an assumption. We can verify it up to some point.
- That is, for each pair we assess:

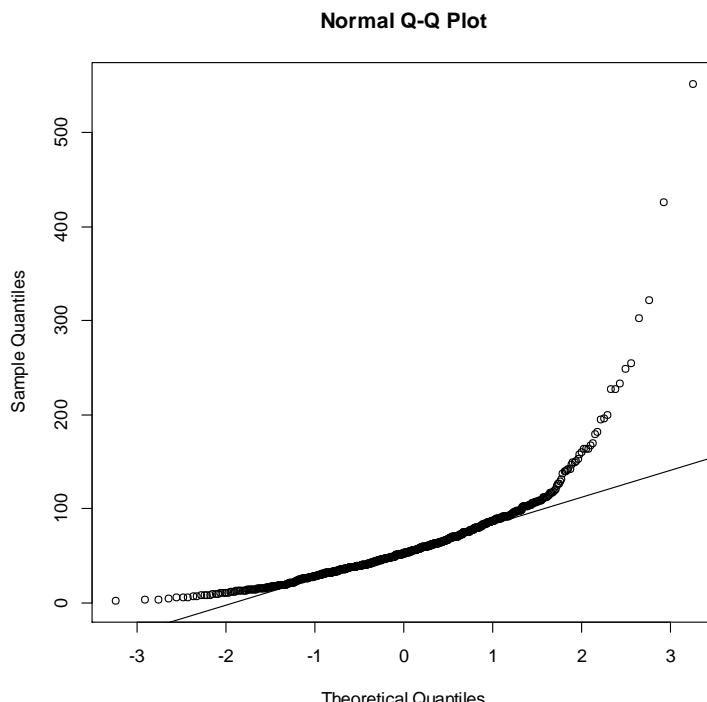
$$\begin{aligned} H_0 : \hat{\rho}_{jk.\backslash jk} &= 0 \\ H_1 : \hat{\rho}_{jk.\backslash jk} &\neq 0 \end{aligned}$$



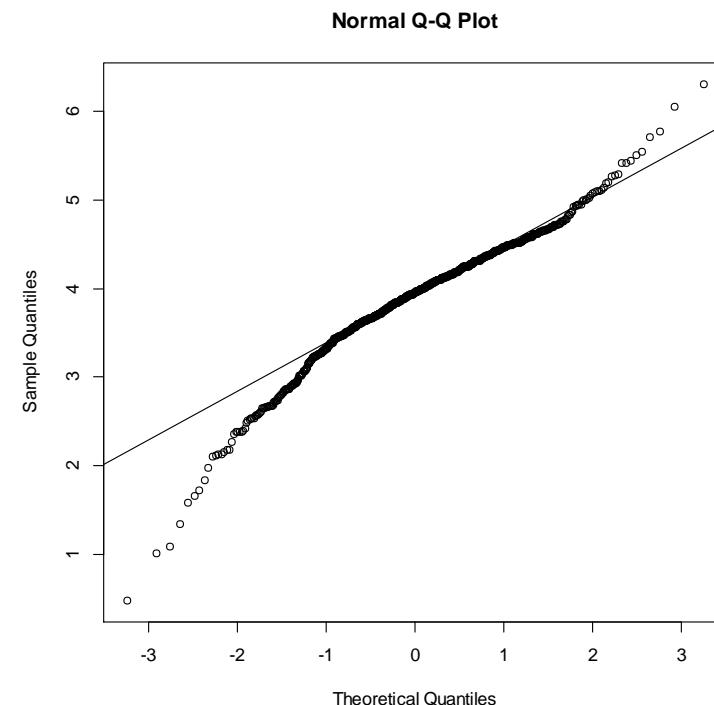
Partial correlation of
 X_j and X_k given all other
variables.

Data Modelling

- Example of plots of concentration of Raf protein:



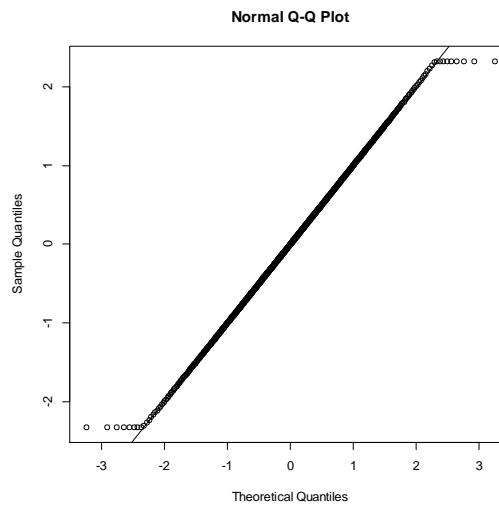
Original scale



Log scale (still preposterous)

Transformation

- For each measurement $x_j^{(i)}$, evaluate its corresponding empirical cdf $\hat{F}_n(x_j^{(i)})$ according to data in column j .
- Get data $z_j^{(i)} \leftarrow \Phi^{-1}(\hat{F}_n(x_j^{(i)}))$ that is, get the empirical approximation to a Gaussian.



In practice, we do some further adjustments to points close to the minimum/maximum. For example, see (the non-trivial) details in Lin, Lafferty and Wasserman (2009). “The Nonparanormal: semiparametric estimation of high dimensional undirected graphs”, Journal of Machine Learning Research (*this is advanced research-level, reading*).

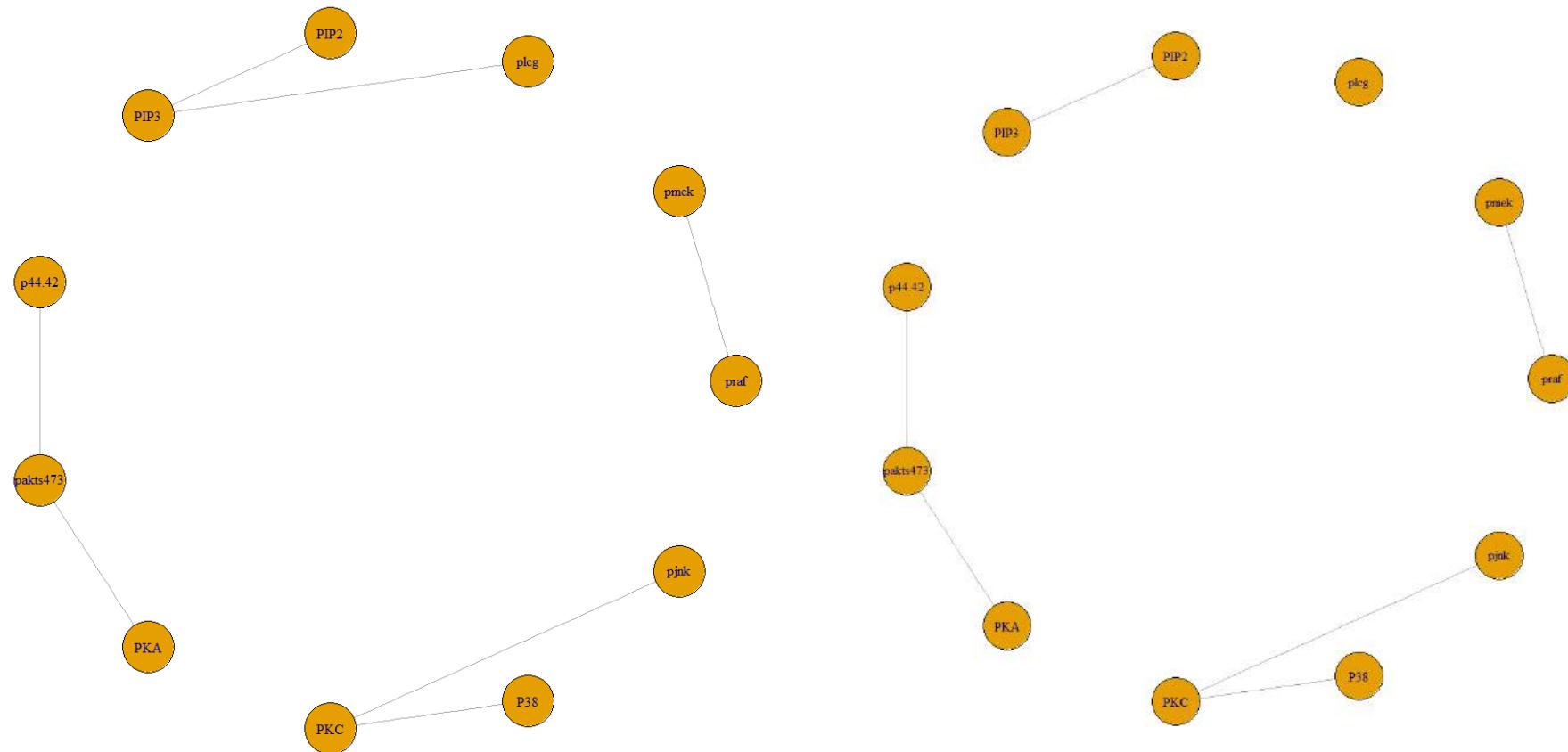
Interpretation

- If our variables have the “dependency structure” of a Gaussian (meaning that after the univariate Gaussian transformation, they look like multivariate Gaussian), then:
 - testing partial correlations of each pair (X_i, X_j) given everybody else recovers the corresponding conditional independencies.
 - It is common to interpret the resulting structure as an “undirected graph”: variables as nodes, edges removed for pairs which are conditionally independent given everybody else.
 - The hope is this tells us something about the physical process. That is, if two protein concentrations are conditionally independent, then these proteins do not cause changes to each other’s concentration except (possibly) via the other measured quantities.
 - This requires extra assumptions.
 - More? *COMPGL08*.

Interpretation

- If we test each pair at a level α (say, 0.05), then the chance of having at least one Type I error is way higher than α .
 - There are $p(p - 1) / 2$ different pairs, each one with an opportunity of Type I error.
- Corrections such as Bonferroni (which is very conservative, but alternatives exist) should be applied.

Example of Output



Without Bonferroni corrections, $\alpha = 0.05$

With Bonferroni corrections

Sidenote: Copulas

- This idea of transforming each variable to a Gaussian and treating the result as multivariate Gaussian is known as a “Gaussian copula”.
- This is a fairly decent compromise between a parametric (too biased) and nonparametric (too data hungry) strategy.
- It doesn't mean it should be applied without critical thinking...

FELIX SALMON BUSINESS 02.23.09 12:00 PM

RECIPE FOR DISASTER: THE FORMULA THAT KILLED WALL STREET



In the mid-'80s, Wall Street turned to the quants—brainy financial engineers—to invent new ways to boost profits. Their methods for minting money worked brilliantly... until one of them devastated the global economy. Photo: Jim Krantz/Gallery Stock

<https://www.wired.com/2009/02/wp-quant/>

(Disclaimer: filters against hysterical, oversimplifying journalism should be applied.)

Regularization

- Even the humble Gaussian distribution might require regularization.
 - 1000 variables? This means $1000 \times 999 / 2$ different covariances.
- One idea is to use l_1 penalisation in the entries of the inverse covariance matrix. This can also set several partial correlations to zero.

FYI: Graphical Lasso

- Maximise this, assuming zero mean. Here, $\theta = \Sigma^{-1}$, the inverse of the covariance matrix.

$$l(\theta) = \frac{n}{2} \log |\theta| - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}^{(i)})^T \theta (\mathbf{x}^{(i)}) - \lambda \sum_{j \neq k} |\theta_{jk}|$$

- Note: Without proof, let me claim that $\hat{\rho}_{jk} \setminus_{jk}$ is exactly what you find in the inverse of the empirical correlation matrix of your data.

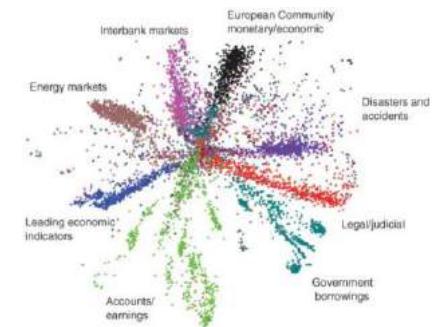
(If you are interested: for details, see SLS Chapter 9)

Unsupervised Learning

DIMENSIONALITY REDUCTION

Dimensionality Reduction

- “Represent the information” in your data with a smaller number of variables.
 - It is a type of data compression.
- Needed:
 - a way of quantifying the information preserved
 - a rule to establish a trade-off of compression vs information loss
 - a family of representations. For instance: compress p variables into 1 by a linear transformation.



$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

Principal Component Analysis (PCA)

- One of the earliest methods for dimensionality reduction based on linear transformations.
- For all that follows, let's assume our given variables have zero (empirical) mean.
- Idea: maximise sample variance of the transformation, subject to it being “normalized”.

$$\max_{\phi_{11}, \dots, \phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \phi_{j1} x_j^{(i)} \right)^2 \right\} \text{ subject to } \sum_{j=1}^p \phi_{j1}^2 = 1$$

Why?

Outcome

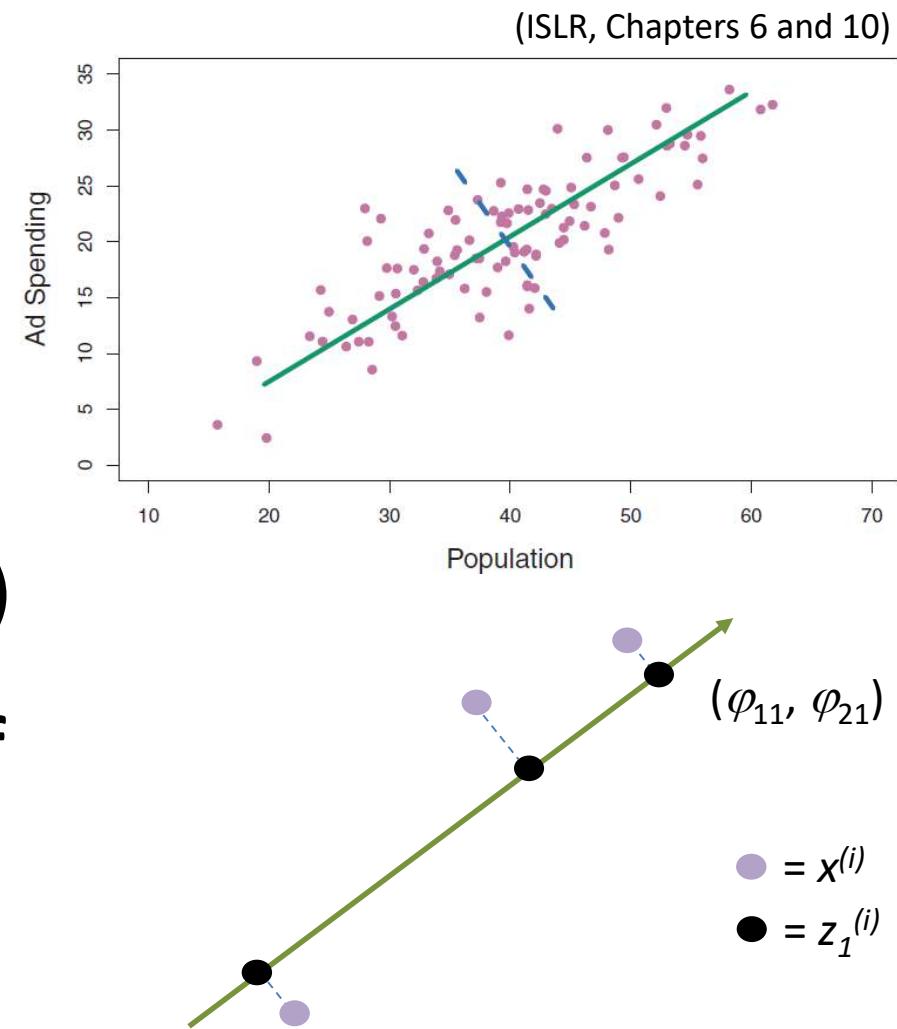
- So if we have a $n \times p$ dataset, we end up with a $n \times 1$ dataset instead.

$$\mathbf{X} = \begin{bmatrix} X_1^{(1)} & X_2^{(1)} & \dots & X_p^{(1)} \\ X_1^{(2)} & X_2^{(2)} & \dots & X_p^{(2)} \\ \dots & \dots & \dots & \dots \\ X_1^{(n)} & X_2^{(n)} & \dots & X_p^{(n)} \end{bmatrix} \xrightarrow{\text{PCA, single variable}} \mathbf{Z} = \begin{bmatrix} Z_1^{(1)} \\ Z_1^{(2)} \\ \dots \\ Z_1^{(n)} \end{bmatrix}$$

- Values $Z_1^{(i)}$ are also called the **scores**, or **projections**, of the data.
- **By construction**, this rule gives the single linear transformation of your data with maximum variance.

Interpretation

- Consider $p = 2$ with the *Advertising* data of ISLR. Let's reduce (*Population*, *Ad Spending*) to a single variable.
- Essentially, we find a direction (vector $\varphi_{11}, \varphi_{21}$) so that **the projection of the data into it will be of maximum variance**.



Adding More Components

- It will follow the same structure

$$Z_2 = \phi_{12}X_1 + \phi_{22}X_2 + \cdots + \phi_{p2}X_p$$

but we will need it to be “as different as possible” from Z_1 .

- PCA uses correlation as a measure of “difference”

$$\max_{\phi_{12}, \dots, \phi_{p2}} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \phi_{j2} x_j^{(i)} \right)^2 \right\} \text{ subject to } \sum_{j=1}^p \phi_{j2}^2 = 1 \text{ and } \sum_{j=1}^p \phi_{j1} \phi_{j2} = 0$$

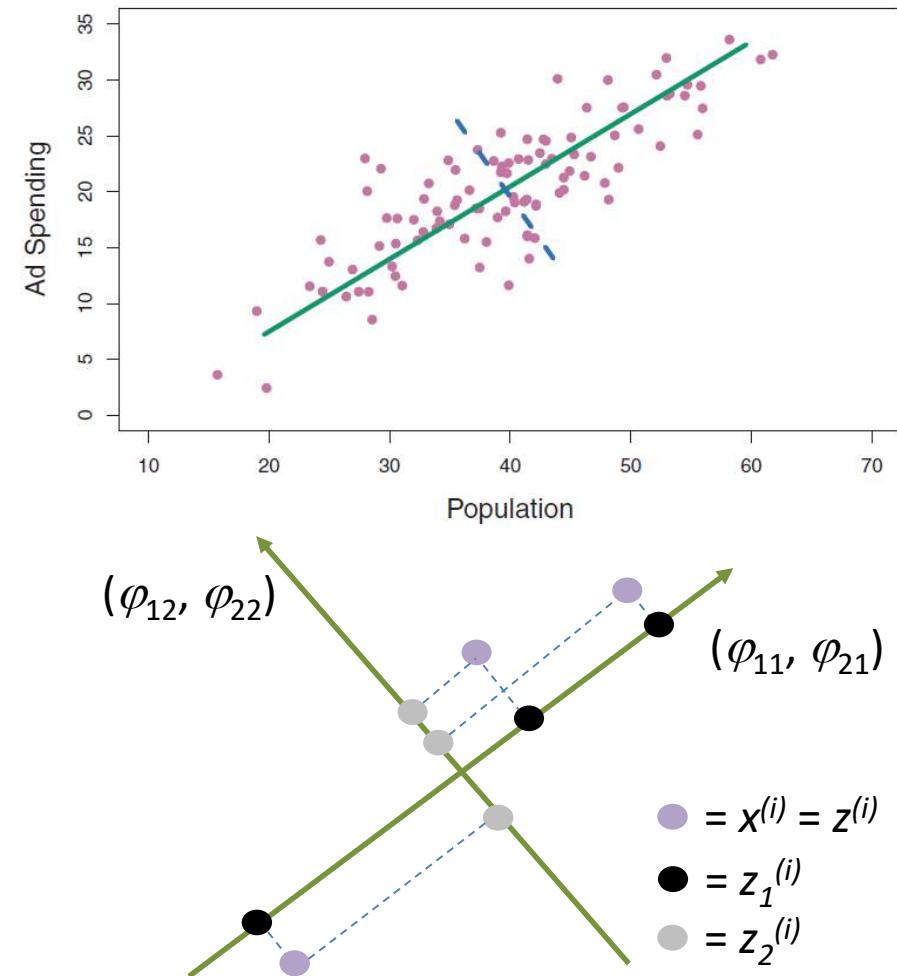
Vectors are perpendicular, this is enough for uncorrelated $z_1^{(i)}$ and $z_2^{(i)}$.

Interpretation

- This adds more dimensions to projection vector $\mathbf{z}^{(i)}$. With two dimensions, the new dataset is

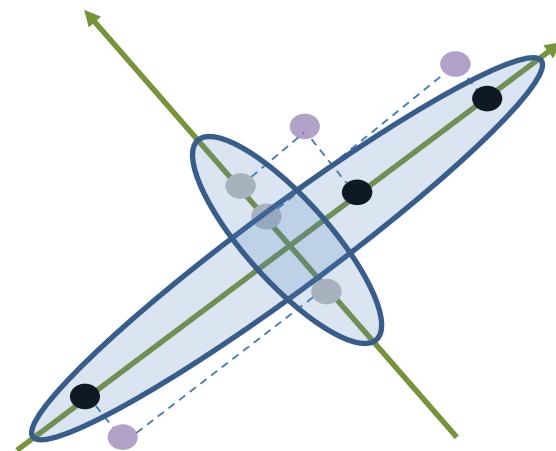
$$\mathbf{Z} = \begin{bmatrix} Z_1^{(1)} & Z_2^{(1)} \\ Z_1^{(2)} & Z_2^{(2)} \\ \dots & \dots \\ Z_1^{(n)} & Z_2^{(n)} \end{bmatrix}$$

- If the original data was two-dimensional, then $\mathbf{Z} = \mathbf{X}!$



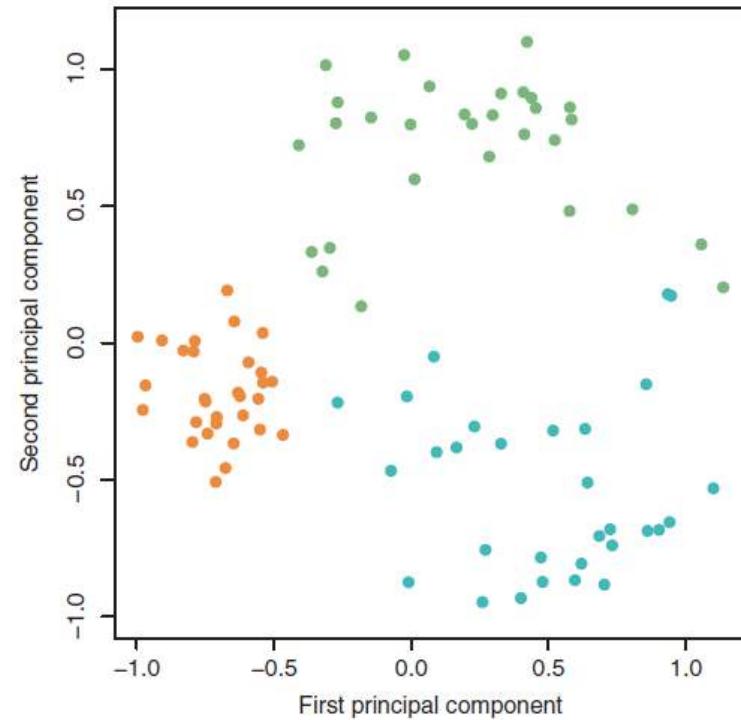
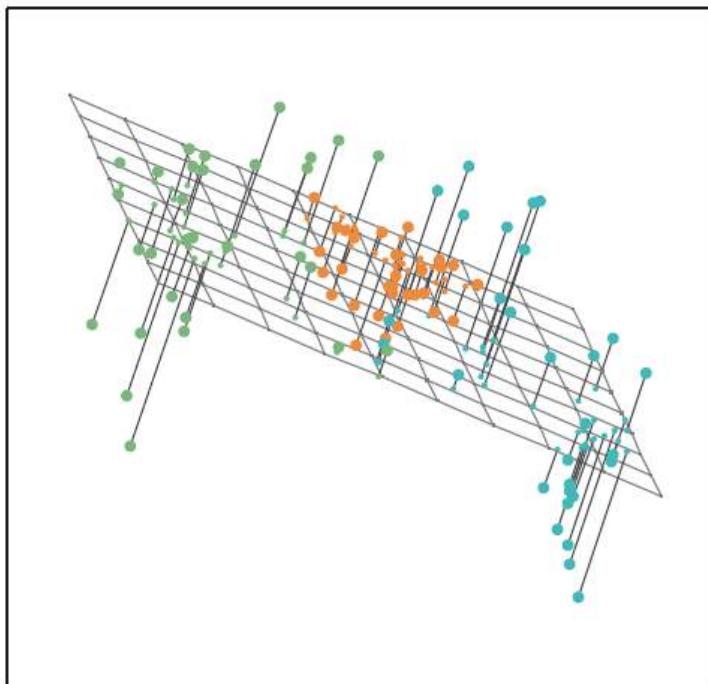
Interpretation

- Notice that the empirical variance around the points projected into the first principal component is greater than the one in the second.
- As it should be, by construction.
- With p variables, Z_3, Z_4, \dots, Z_p are defined in an analogous way: make each Z_m uncorrelated with all those preceding it.



$p = 3$ Example, Projection into 2 Components

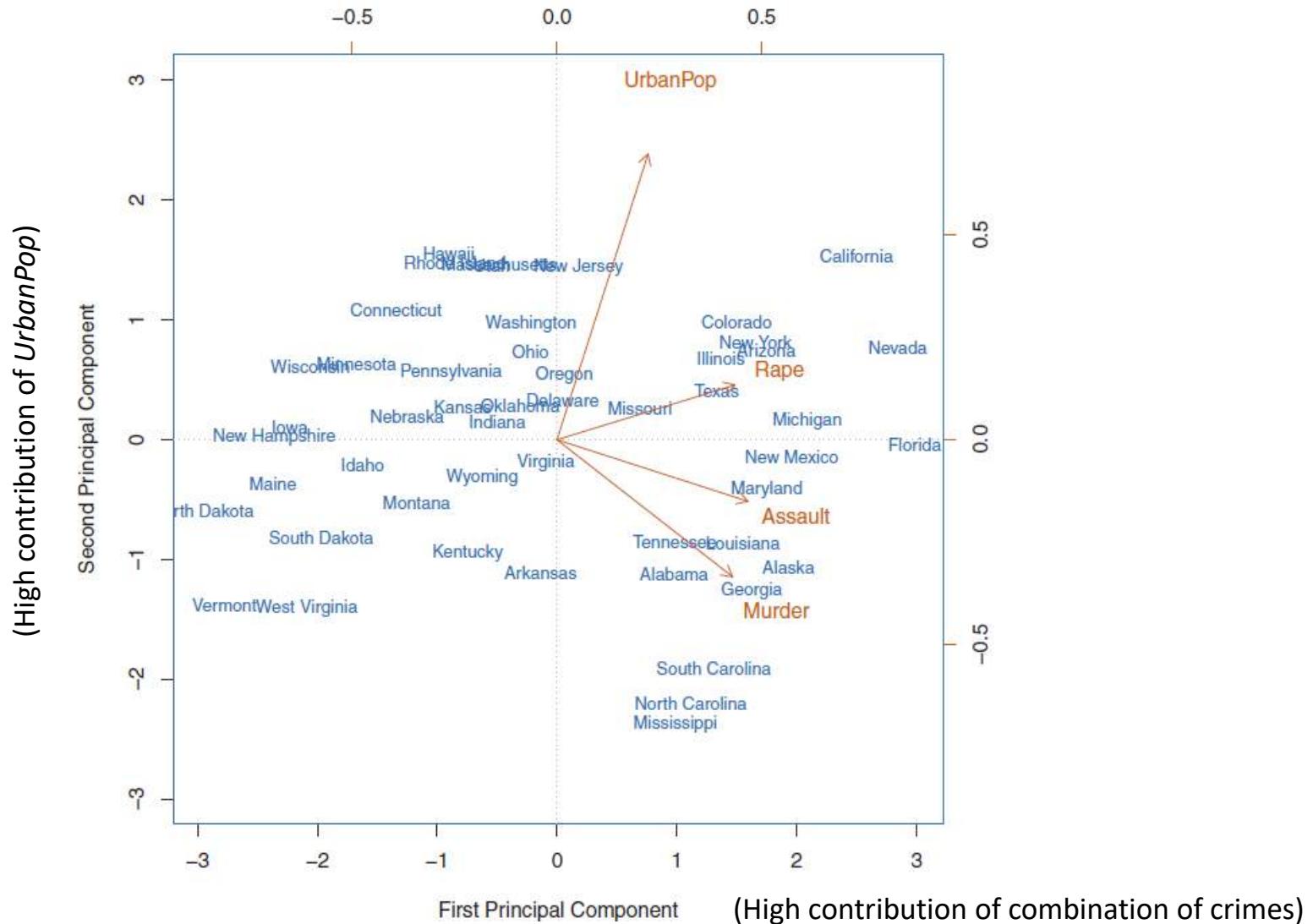
- Synthetic data (ISLR, Chapter 10)



Example: USA Arrests Data (ISLR)

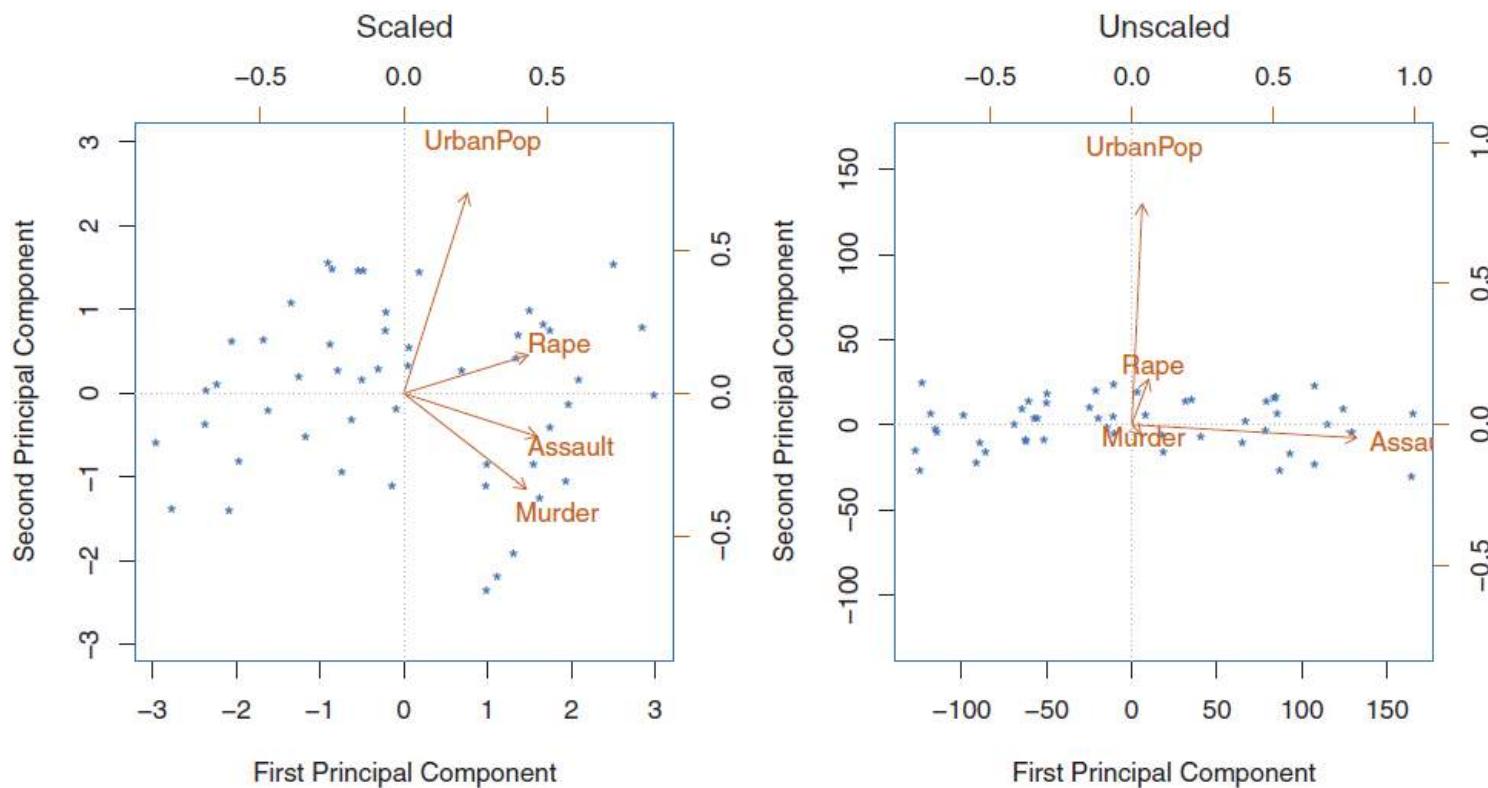
- For each of the $n = 50$ states in the US, this includes number of arrests per 100,000 residents for crimes of *Assault*, *Murder* or *Rape*.
- Also, percentage of population (*UrbanPop*) living in urban areas, per state.
- Goal: visualize how states differ.
 - scale the data to zero empirical mean/one standard deviation, do PCA with two components, get $n \times 2$ reduced data matrix Z , plot it.
- For each variable j , also plot the vector $(\varphi_{j1}, \varphi_{j2})$.

Result



Other Practical Issues

- **PCA results depend on how you scale your data.** In hindsight this is obvious, since the maximisation of variance means we will put more weight in variables of higher variance.



Other Practical Issues

- We may want to choose the number of components for the goal of using Z as input to some other statistical analysis.
 - For instance, regression for prediction purposes.
- How to select it?
 - Cross-validation is an option, but even then we might want to reduce the possibilities to a more manageable set.

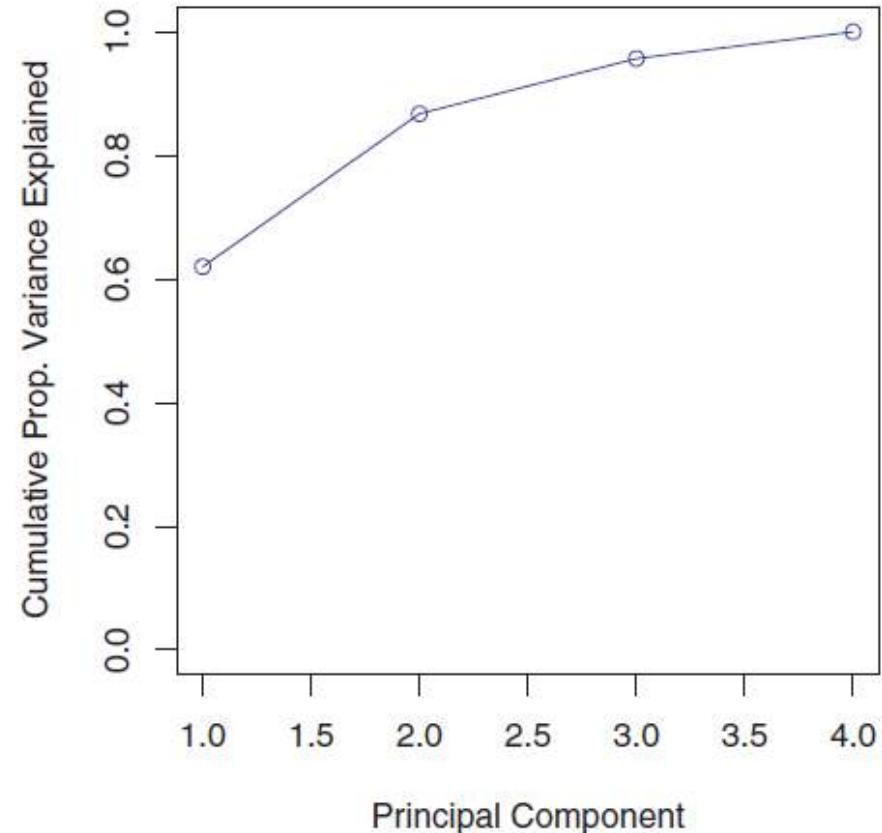
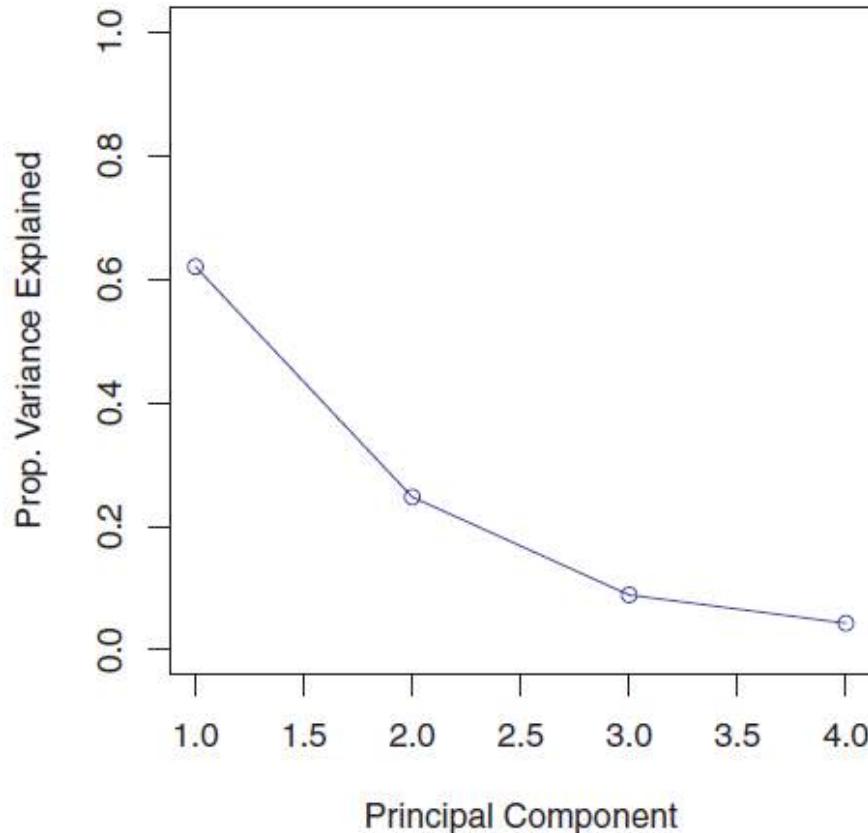
Other Practical Issues

- One heuristic is the proportion of variance explained (PVE) by m components, contrasted to total variance (TV). For zero-mean data,

$$TV \equiv \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n (x_j^{(i)})^2 \quad PVE(m) \equiv \frac{\frac{1}{n} \sum_{i=1}^n (z_m^{(i)})^2}{TV}$$

- We can plot $PVE(m)$ against m , visualize where it is not worthwhile to add more components.

Example: USA Arrests Data



- The plot on the left is sometimes called a **scree plot**.

Other Practical Issues

- PCA is particularly useful for multivariate Gaussian data. Sometimes unclear how good it is otherwise.
 - Alternatives: make data “Gaussian-like” by marginally transforming variables; use non-linear PCA (a story for another day), neural networks etc.
- It is possible to directly pipeline PCA with regression to optimise parameters in a different way.
 - See Principal Components Regression, Section 6.3 of ISLR.
 - In the non-linear case, this is basically what a multilayer perceptron is.
- Just like any statistical method, it is possible to have confidence intervals on PCA coefficients. They are less obvious to derive. Johnson and Wichern (see reading list) discuss this with more detail if you are interested.

Unsupervised Learning

LATENT VARIABLE MODELS, WITH APPLICATION TO CLUSTERING

Latent Variable Models

- We already discussed some concepts when discussing GLMs.
- A latent variable model is a model where some variables are not in the data (like the “latent propensities” in the ordinal regression models of Chapter 4).
- This is a vast area of modelling. The goal here is just to provide two canonical examples.
 - *COMPGL08* for more.

Latent Gaussian Models

- Consider a two-stage zero mean model,

$$X_k \sim N(0, 1), k = 1, 2, \dots, m$$

$$\begin{aligned} Y_j &= \beta_{j1}X_1 + \dots + \beta_{jm}X_m + \epsilon_j, \\ \epsilon_j &\sim N(0, \sigma^2), j = 1, 2, \dots, p \end{aligned}$$

- The previous line is equivalent to

$$Y_j \mid X_1, \dots, X_m \sim N(\beta_{j1}X_1 + \dots + \beta_{jm}X_m, \sigma^2)$$

Latent Gaussian Models

- Using properties of linear combinations of Gaussian variables (as seen earlier), we have the following **marginal likelihood** for $\beta \equiv (\beta_1, \dots, \beta_m)$ and σ^2 :

$$\prod_{i=1}^n p(\mathbf{y}^{(i)}) = \prod_{i=1}^n \frac{1}{(2\pi)^{p/2} |\Sigma(\beta)|^{1/2}} e^{-\mathbf{y}^{(i)T} \Sigma(\beta)^{-1} \mathbf{y}^{(i)} / 2}$$
$$\Sigma(\beta) \equiv \beta\beta^T + \sigma^2 \mathbf{I}$$

where \mathbf{I} is the $p \times p$ identity matrix.

Why Should We Care?

- Models like this postulate that there are hidden factors explaining the dependencies of our observations.
 - Interpretation: can we understand what these factors are?
 - Sparsity: notice that the number of parameters for the covariance matrix of \mathbf{Y} can be much smaller than the usual $p(p + 1) / 2$ in a full covariance matrix.

Interpretation

- Once we fit the parameters by maximum likelihood, we can provide an **estimate of the latent variables** such as

$$\hat{\mathbf{x}}^{(i)} \equiv E[\mathbf{X}^{(i)} \mid \mathbf{y}^{(i)}]$$

- Guess what? This is the same as the m **principal components** of the data!

Tipping and Bishop (1999). “Probabilistic PCA”.
Journal of the Royal Statistical Society.

Implications

- We can pick-and-choose from our knowledge of regression models to define new models. For instance:

$$Y_j \mid X_1, \dots, X_m \sim \text{Poisson}(\exp(\beta_{j1}X_1 + \dots + \beta_{jm}X_m))$$

- Fitting parameters can be difficult, but there are advanced off-the-shelf algorithms for that.
- Regularization, including l_1 , can be used to get sparse solutions! See SLS, Chapter 8, if you are curious.
- I must end the story here, but keep this in mind as a general lesson: statistical data science is much more flexible and richer than what we have learned so far.

A Simple Latent Variable Model: Mixture of Gaussians

- Let X be a scalar discrete variable taking values in $1, 2, \dots, K$ for a given K . Parameters:

$$\theta_k \equiv P(X = k)$$

- Let our observations (Y_1, \dots, Y_p) be given independently by

$$Y_j \mid X = k \sim N(\mu_k, \sigma^2), j = 1, 2, \dots, p$$

- This is a type of **Gaussian mixture model**.

Fitting

- The marginal likelihood is given by remembering that $p(\mathbf{y}) = \sum_{k=1}^K p(\mathbf{y} | X = k)P(X = k)$:

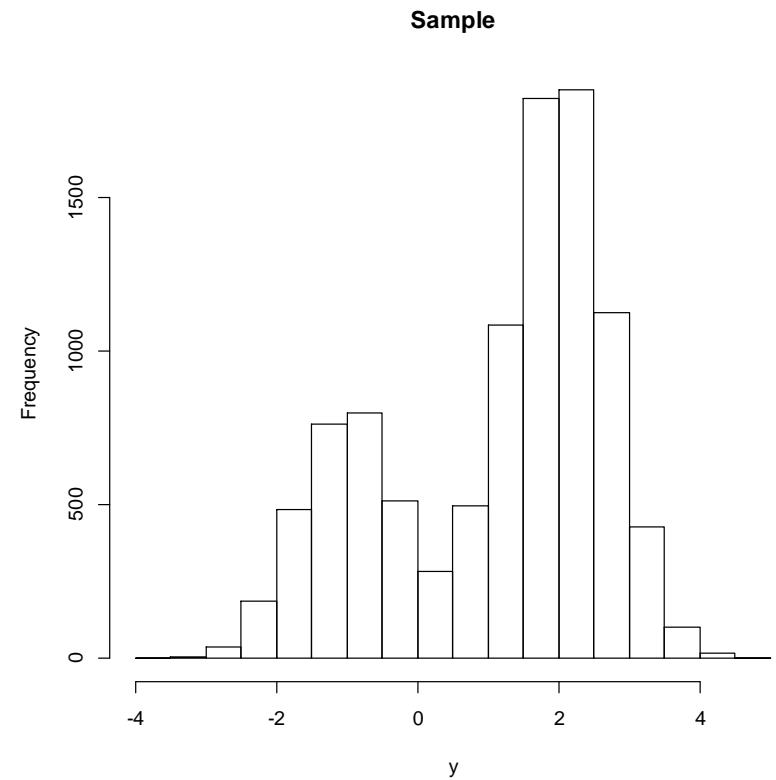
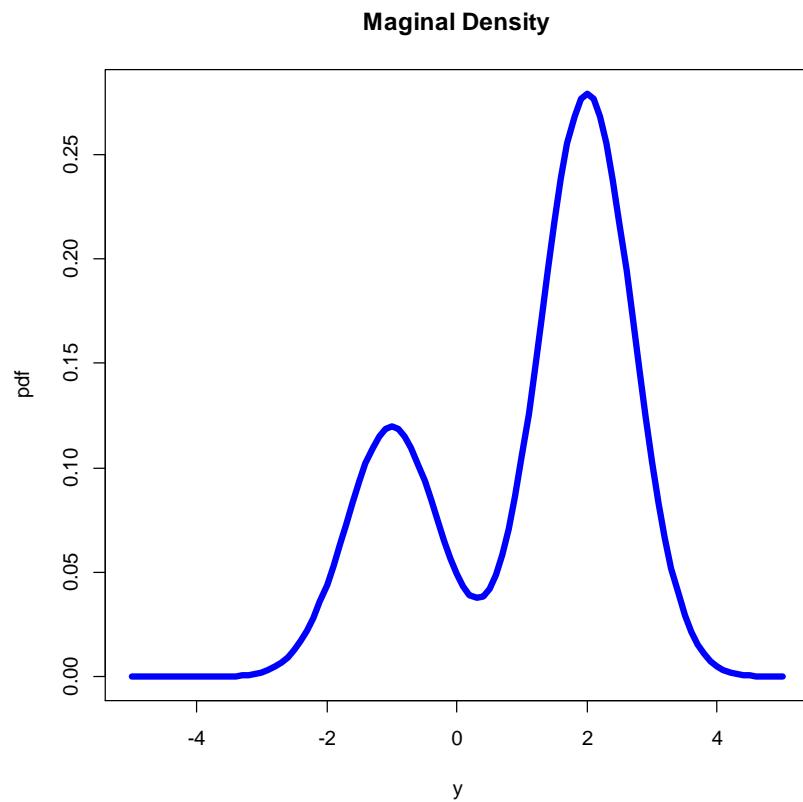
$$\prod_{i=1}^n p(\mathbf{y}^{(i)}) = \prod_{i=1}^n \left(\sum_{k=1}^K \theta_k \left\{ \prod_{j=1}^p \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_j^{(i)} - \mu_k)^2} \right\} \right)$$

$p(X^{(i)} = k)$ $p(\mathbf{y}^{(i)} | X^{(i)} = k)$

- We can then maximise this marginal likelihood to get estimates of $\{\mu_k\}, \sigma^2, \{\theta_k\}$.

Example

- $p = 1, K = 2.$



Clustering

- One interpretation of clustering (there are others) is: can we recover which mixture component generated each data point?
- More generally: find “natural” groups of data points.
 - Example: market segmentation, socio-economic stratification, etc.
- Clustering does not need a likelihood function, but latent variable models can motivate clustering.
 - Interpreting the outcome, by the end of the day, is interpreting latent variable assignments.

K-Means

- An algorithm originally motivated by grouping points by Euclidean distance. Find a partition (C_1, \dots, C_K) of $\{1, \dots, n\}$ to solve

$$\min_{C_1, \dots, C_K} \sum_{i \in C_k} \sum_{j=1}^p (x_j^{(i)} - \bar{x}_{jk})^2$$

Average of x_j among points in C_k

- This is a “kind” of maximum likelihood where θ_k is different for each i such that

$$\theta_k^{(i)} \in \{0, 1\} \text{ and } \sum_{k=1}^K \theta_k^{(i)} = 1$$

Solving K-Means

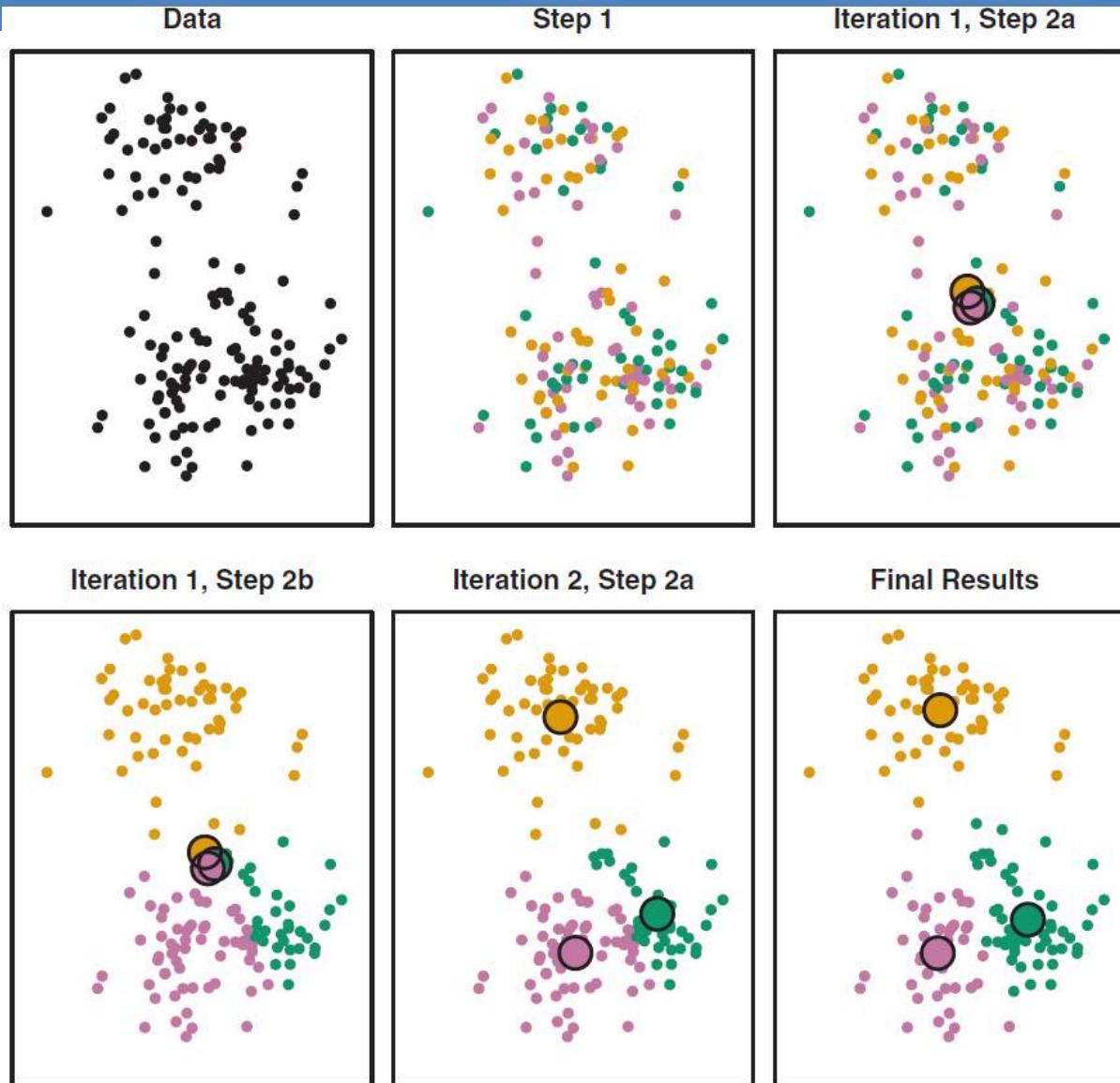
- Solving this exactly is pretty hard!
 - Exhaustive search: K possibilities for each of the n points, resulting on n^K possibilities
 - greater than number of particles in the observable universe for most problems.
- Practical solution:
 1. Allocate points to clusters randomly
 2. Find averages for each cluster
 3. Optimise cluster assignments for given averages
 4. Repeat 2-4 to convergence
 5. Repeat 1-4 a few times with different starting points, get best solution.

The Algorithm

Algorithm 10.1 *K*-Means Clustering

1. Randomly assign a number, from 1 to K , to each of the observations. These serve as initial cluster assignments for the observations.
 2. Iterate until the cluster assignments stop changing:
 - (a) For each of the K clusters, compute the cluster *centroid*. The k th cluster centroid is the vector of the p feature means for the observations in the k th cluster.
 - (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).
-

Illustration, $K = 3$



(ISLR, Chapter 10)

Practical Issues: Random Restarts



Practical Issues: Choice of K

- Many ways. Here are a few:
 - In principle, with an explicit likelihood, penalties like BIC could be applied.
 - Not necessarily a good thing. Do we believe in the mixture of Gaussians model?
 - “Compression” point of view:
 - Clustering is just extreme dimensionality reduction: one scalar per data point, the cluster assignment! Choose the largest K we feel happy with (e.g., how many targeted marketing campaigns are we willing to design?)
 - “Acceptable reconstruction error”: to avoid wasting resources with large K , allow for smaller K if the average/maximum distance of the points to the cluster mean is below some domain-specific error.

Practical Issues: Validation

- Like much of unsupervised learning, this is not easy to validate.
- All sorts of sensitivity analyses:
 - How does clustering change given a subset of the data? (a type of bootstrapping)
 - How does it change given a subset of the variables?
How much variance do the unused variables have within each cluster?
- I'm afraid there is no easy, domain-independent, answer here.

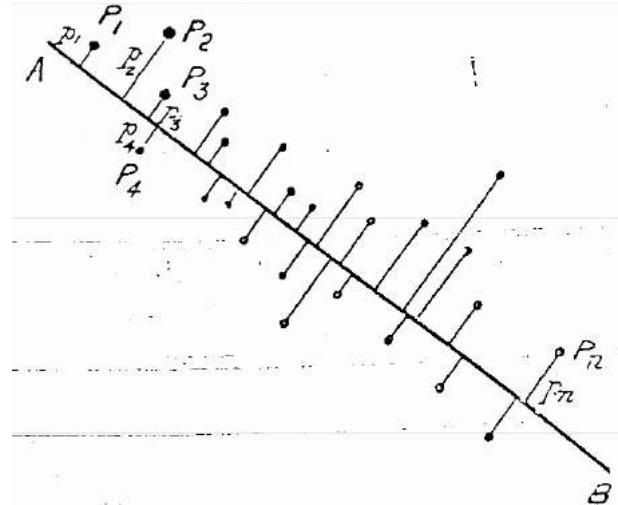
Take-Home Messages

- From a statistical perspective, we can see unsupervised learning as estimating joint distributions.
 - Sometimes merely as a tool to facilitate visualization or supervised learning.
- However, this is not the whole story, as we would like to characterize “features” of such distributions.
- Outliers, independencies and latent variables (including cluster assignments and PCA projections) are one way of describing what these features are, but they may come with strong assumptions too.

Historical Notes

PCA was invented by Karl Pearson in 1901. He was also responsible for the popularization of the correlation coefficient, the reason why it is also known today as “Pearson correlation”, distinguishing it from other less common measures of association.

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, F.R.S., University College, London *.



J. R. Statist. Soc. B (1979),
41, No. 1, pp. 1–31

One of the earliest formalizations of models of conditional independence was Dawid's 1979 paper. It was in it that the symbol || was introduced.

Conditional Independence in Statistical Theory

By A. P. DAWID†

University College London

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, October 18th, 1978, the Chairman Professor J. F. C. KINGMAN in the Chair]

...

Although some discussants have their doubts that conditional independence can live up to my inflated claims, I have been very gratified by the overall constructive and stimulating response to my paper. My appreciation and thanks go out to all the contributors. If my notation and general theory find application in the future work of others, I shall at least have justified to the Society's Printers my insistence that, somehow or other, they should find a way to print that awkward symbol ||.

END OF THE BEGINNING