Introduction to Big Data

Lecture 1

Connor Dowd

March 30th 2021

A Plea

You are not required to keep your cameras on.

Please do. It helps me immensely to be able to see how people respond as I say things.

If 15% of you are lost, I want to know about it, but with 20 cameras on, I have to notice one of only three faces being lost.

More generally though, please, jump in with questions when you have them. Use the chat if you prefer.

I'm here to teach you.

- If you have questions,
- If you have trouble,
- If you have feedback for me,
- If you have ideas for improving the course,

Please let me know.

Today's Class

1. Introduction
   - Goals
   - Material
   - Syllabus
2. What is Big Data?
3. Computation: R, Big Data
4. Data Viz, Statistics, Dimension Reduction
5. Testing: False Discovery

Big Data

Introduction

Big data **is not** an end in itself

We are doing Inference at LARGE scale!

The goal here is to learn what we can trust, how it can be used, and how to learn more.

We want to help you make good decisions with lots of messy data.

By necessity then, this class will be a mix of theory and practice.

Theory and Practice

You need a solid foundation in statistical principles.

- We don’t want to shoot ourselves in the foot

You also need a hefty dose of ‘rules of thumb’.

- Inefficient procedures can be wildly inefficient
- **BUT** we don’t want to reinvent the wheel.

This is hands on work.

What is Big Data?

There are a lot of names out there for a large cluster of very similar disciplines:

- Econometrics
- Data Science
- Big Data
- Statistics
- Datamining
- Machine Learning

There are differences in focus and style here. But the similarities are larger than the distinctions.

What is Big Data?

Big Data as a name, originates with computer scientists working with data too large for any single computer

“Big Data is one GB larger than my RAM”.

But often it is associated with administrative data, where statistical notions of sampling error may start to fall apart.

More generally, it is in a nexus where a fairly tight connection between inference and prediction can start to break down.

What is Big Data?

Big Data is focused on extracting useful truth from large datasets.

- Infer patterns in high dimensional data
- Simple and scalable algorithms
- Honest and humble model selection
- Manage conflict between “useful” and “true”
- Make some **decision**

What is Big Data?

Big in number of observations (size n).

Big in number of variables (dimension p).

In these settings you *cannot*:

- Look at each variable and make a decision (t-tests).
- Choose from a small set of nested models (F-tests).
- Plot every variable to look for interactions and transformations.

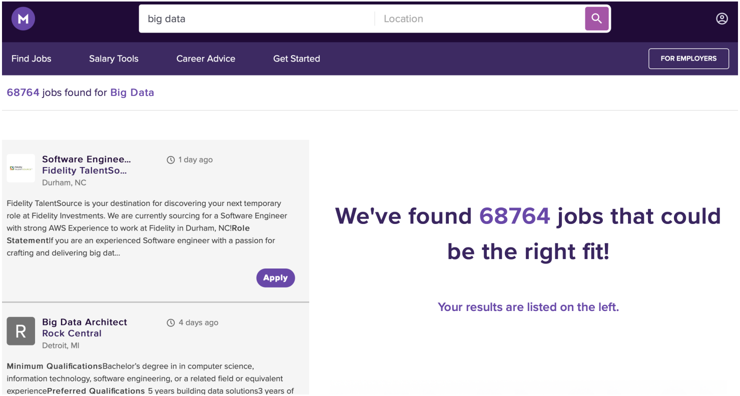
Some of our tools are straight out of Stats 101 (regression, confidence intervals), some are close relatives of Stats 101 (PCA, MSE), some are new beasts (trees, Bagging).

Why should you care?

- Extremely employable.

Why should you care?

- Extremely employable.



68764 jobs found for Big Data

Software Enginee...
Fidelity TalentSo...
Durham, NC
1 day ago

Fidelity TalentSource is your destination for discovering your next temporary role at Fidelity Investments. We are currently sourcing for a Software Engineer with strong AWS Experience to work at Fidelity in Durham, NC!**Role Statement**If you are an experienced Software engineer with a passion for crafting and delivering big dat...

Apply

Big Data Architect
Rock Central
Detroit, MI
4 days ago

Minimum Qualifications Bachelor's degree in in computer science, information technology, software engineering, or a related field or equivalent experience**Preferred Qualifications** 5 years building data solutions3 years of

We've found 68764 jobs that could be the right fit!

Your results are listed on the left.

-

Why should you care?

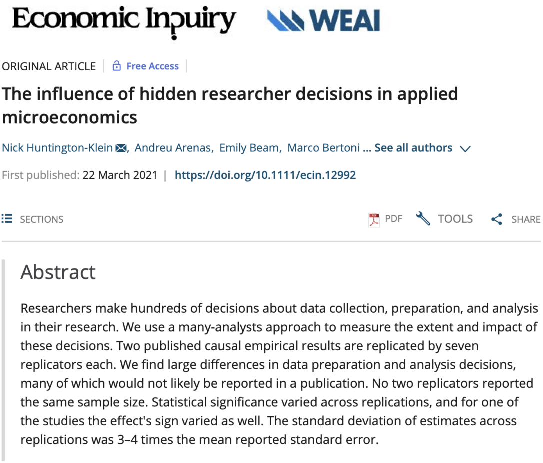
- Extremely employable
- Avoid being hoodwinked by nutjobs with numbers.
- Avoid hoodwinking yourself with numbers. *They want to lie to you*.

Why should you care?

- Extremely employable
- Avoid being hoodwinked by nutjobs with numbers.

- Avoid hoodwinking yourself with numbers. *They want to lie to you*.
- Its fun. Genuinely.

Does data really want to lie to you? | YES



ORIGINAL ARTICLE | Free Access |

The influence of hidden researcher decisions in applied microeconomics

Nick Huntington-Klein, Andreu Arenas, Emily Beam, Marco Bertoni ... See all authors

First published: 22 March 2021 | https://doi.org/10.1111/ecin.12992

Abstract

Researchers make hundreds of decisions about data collection, preparation, and analysis in their research. We use a many-analysts approach to measure the extent and impact of these decisions. Two published causal empirical results are replicated by seven replicators each. We find large differences in data preparation and analysis decisions, many of which would not likely be reported in a publication. No two replicators reported the same sample size. Statistical significance varied across replications, and for one of the studies the effect's sign varied as well. The standard deviation of estimates across replications was 3–4 times the mean reported standard error.

Figure 1: Influence of Researcher Decisions can reverse results

Admin

Course Schedule

subject to change. . .

1. Data: Computing, plotting, principles. False Discovery. Loss.
2. Regression: Everything in two days. Linear and Logistic.
3. Model Selection: Penalties, information criteria, cross-validation
4. Treatment Effects: High Dimensional Controls, AB testing, Bootstrap.
5. Classificiation: Multinomials, KNN, sensitivity/specificity
6. Trees: CART, Random Forests, Ensembles
7. Networks?: Co-occurence, directed graphs
8. Clustering?: k-means, mixture models, association rules
9. Factors?: Latent variables, PCA, PLS

Big Data Team

TA: Yuxiao Li – Booth PhD. She will primarily be grading your work.

Me: Booth PhD. First (and last) time teaching this course.

Office hours: Happy to meet with you. For now, email me if you want to meet. If that proves overwhelming for me, I may set formal office hours.

Structure: Today I'll see how much I get through, and then Thursday will be what we don't get to, as well as discussion of homework 1 (group assignment due next week), and more.

Course Design

This is also a good time to note that this course is based closely off the course by the same name taught by Professor Veronika Rockova.

But some things will be different.

First time for this course with undergrads. So we don't know what pace makes sense. Feedback is welcome.

Textbooks

The primary text for this course is “Elements of Statistical Learning” by Hastie, et al.

It is recommended as an excellent reference going forwards. PDFs are (legally) available for free at https://web.stanford.edu/~hastie/Papers/ESLII.pdf and on my website.

I will also try to recommend other textbooks as we go along for specific topics. If there is a topic you’re interested in that I haven’t recommended a textbook for, email me about it.

Homeworks

Homeworks will be in small groups of 3. You may not have groups smaller than 3.

There will be ∼1 homework a week, graded from 1-5.

1. Oops
2. Eh
3. Fine
4. Good. – Typical score for homework meeting my expectations. (Nothing wrong with it)
5. Nailed it. – Somehow you impressed the TA.

I don't expect there to be many 5s. Literally nothing is wrong with a 4.

There will also be a prediction competition.

Other Grade Components

1. Participation NEW – I care a lot about asking and answering questions on the canvas discussion board.
2. Midterm – possible, curious how you all feel about a 20 question multiple choice/numeric answer midterm.
3. Take home final

Analysis

We will be working with real data. The essence of big data is an excellent ability to make predictions using data.

To that end, we will also run a low key prediction competition. For now, there is one thing to predict.

If there is interest in the prediction competition, we will have more questions.

First prediction assignment: US covid vaccination predictions.

Details

“How many people in the US will have had at least one dose by end of day on April 30th?”

- Predictions due before class on Thursday.
- Full details in Assignment on website
  - Along with 3 CDC datasets over time.
- You only need to provide 3 numbers:
  - A prediction
  - A 90% CI lower bound
  - A 90% CI upper bound
- Enter on canvas for full credit
- Enter on google form to compete against me
  - And register interest in more questions.

Break

Computation

R

All of our analysis will be in R.

This is the premier data analysis platform. Other platforms are better at some things, none is better at everything.

Its free, open source, cross platform, very capable.

R

Used by academics (in statistics, marketing, finance, genetics, etc), companies (EBAY, Google, Microsoft, Boeing, Citadel, IBM, etc), governments (RAND, DOE Labs, US Navy, etc).

R's big strength is that it is open source and has a large community contributing packages. E.g. it is very easy to install a package to do LASSO or other fancy regression tools.

But it can be unpolished. The non-standard packages are as varied as their contributors. Some are super specific, some are super general. Many are useless.

R is not flawless. But it has a large community that will help you overcome flaws.

I **strongly** recommend using the RStudio IDE. (This lecture, your homeworks, much else in this course has been made with it)

R

The big barrier to entry is the command line interface.

- You type a command, R follows the command.

This can mean quite the learning curve. But it is well worth it.

- Once you write code for something, it becomes easy to repeat that operation.
- This is a critical skill for using modern statistical tools.

I will post all code online. I strongly recommend you look at my solutions to HWs, etc if you run into trouble.

R – Introduction

Open R (or RStudio). Just like any other program.

In some sense, R is a jumped up calculator. (e.g. *,/,+,-,^,%%)

If you want to save a number, or anything else, give it a name.

```
A = 2
```

If you want to see what a variable is, just type the name:

```
A
```

```
## [1] 2
```

R – Vectors

We can build vectors with `c()` or “`:`”

```r
1:5
```

```
## [1] 1 2 3 4 5
```

```r
B = c(1,5,3)
B
```

```
## [1] 1 5 3
```

R – Vectors

We can grab subsets of vectors (and other groupings) with `[ ]`.

```
B[2]
```

```
## [1] 5
```

And then do math with them. (“;” will end a line)

```
D = A+B[2]; D
```

```
## [1] 7
```

Sometimes, its not clear what you want, and R takes a guess:

```
E = A+B
```

R – Errors

```
E
```

```
## [1] 3 7 5
```

And sometimes you ask it to do something it can’t do. So it tells you it can’t do that, and tries to explain itself.

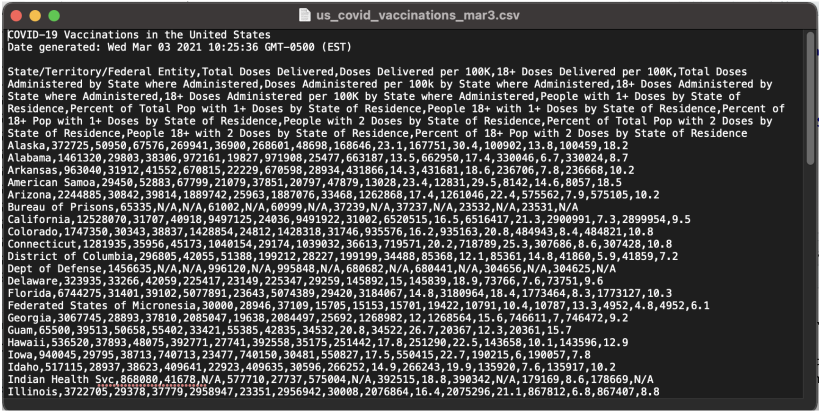
```
G = A+b[2]
```

```
## Error in eval(expr, envir, enclos): object ’b’ not found
```

If something isn’t working, googling often helps. (e.g. “R error object not found”)

R – CSVs

We are usually going to work with data in .csv files. Incredibly human-readable files, and easy to move between Excel and other programs.



```
COVID-19 Vaccinations in the United States
Date generated: Wed Mar 03 2021 10:25:36 GMT-0500 (EST)

State/Territory/Federal Entity,Total Doses Delivered,Doses Delivered per 100K,18+ Doses Delivered per 100K,Total Doses
Administered by State where Administered,Doses Administered per 100k by State where Administered,18+ Doses Administered by
State where Administered,18+ Doses Administered per 100K by State where Administered,People with 1+ Doses by State of
Residence,Percent of Total Pop with 1+ Doses by State of Residence,People 18+ with 1+ Doses by State of Residence,Percent of
18+ Pop with 1+ Doses by State of Residence,People with 2 Doses by State of Residence,Percent of Total Pop with 2 Doses by
State of Residence,People 18+ with 2 Doses by State of Residence,Percent of 18+ Pop with 2 Doses by State of Residence
Alaska,372725,50950,67576,269941,36900,268601,48698,168646,23.1,167751,30.4,100902,13.8,100459,18.2
Alabama,1461320,29803,38306,972161,19827,971908,25477,663187,13.5,662950,17.4,330046,6.7,330024,8.7
Arkansas,963040,31912,41552,670815,22229,670598,28934,431866,14.3,431681,18.6,236706,7.8,236668,10.2
American Samoa,29450,52883,67799,21079,37851,20797,47879,13028,23.4,12831,29.5,8142,14.6,8057,18.5
Arizona,2244885,30842,39814,1889742,25963,1887076,33468,1262868,17.4,1261046,22.4,575562,7.9,575105,10.2
Bureau of Prisons,65335,N/A,N/A,61002,N/A,60999,N/A,37239,N/A,37237,N/A,23532,N/A,23531,N/A
California,12528070,31707,40918,9497125,24036,9491922,31002,6520515,16.5,6516417,21.3,2900991,7.3,2899954,9.5
Colorado,1747350,30343,38837,1428854,24812,1428318,31746,935576,16.2,935163,20.8,484943,8.4,484821,10.8
Connecticut,1281935,35956,45173,1040154,29174,1039032,36613,719571,20.2,718789,25.3,307686,8.6,307428,10.8
District of Columbia,296805,42055,51388,199212,28227,199199,34488,85368,12.1,85361,14.8,41860,5.9,41859,7.2
Dept of Defense,1456635,N/A,N/A,996120,N/A,995848,N/A,680682,N/A,680441,N/A,304656,N/A,304625,N/A
Delaware,323935,33266,42059,225417,23149,225347,29259,145892,15,145839,18.9,73766,7.6,73751,9.6
Florida,6744275,31401,39102,5077891,23643,5074389,29420,3184067,14.8,3180964,18.4,1773464,8.3,1773127,10.3
Federated States of Micronesia,30000,28946,37109,15705,15153,15701,19422,10791,10.4,10787,13.3,4952,4.8,4952,6.1
Georgia,3067745,28893,37810,2085047,19638,2084497,25692,1268982,12,1268564,15.6,746611,7,746472,9.2
Guam,65500,39513,50658,55402,33421,55385,42835,34532,20.8,34522,26.7,20367,12.3,20361,15.7
Hawaii,536520,37893,48075,392771,27741,392558,35175,251442,17.8,251290,22.5,143658,10.1,143596,12.9
Iowa,940045,29795,38713,740713,23477,740150,30481,550827,17.5,550415,22.7,190215,6,190057,7.8
Idaho,517115,28937,38623,409641,22923,409635,30596,266252,14.9,266243,19.9,135920,7.6,135917,10.2
Indian Health Svc,868080,41678,N/A,577710,27737,575004,N/A,392515,18.8,390342,N/A,179169,8.6,178669,N/A
Illinois,3722705,29378,37779,2958947,23351,2956942,30008,2076864,16.4,2075296,21.1,867812,6.8,867407,8.8
```

Figure 2: CSV examples

R – Loading Data

We can load the data into R using the `read.csv` command. It is good practice to set a working directory first.

```r
setwd("~/Github/bigdata/predictions/")
mar3 = read.csv("us_covid_vaccinations_mar3.csv",skip=3)
```

Now we can play with (some of) the data.

R – Dataframes

mar3 is a dataframe. We can index columns with names as well as their position. Dataframes also improve on matrices by allowing different variable types in each column.

```r
mar3[1,] # The first row
mar3[1:10,] # 10 rows
mar3[,2] # The second column
mar3$Total.Doses.Delivered # The second column
mar3[mar3$Total.Doses.Delivered > 1000,] #Rows w/ >1k doses
nrow(mar3)
ncol(mar3)
dim(mar3)
colnames(mar3)
```

R – Variable types

R stores different types of data differently.

- Numeric data are just numbers.
  - Though they can take several types (integer, float, etc)
- Factor data take a small number of values.
  - Think vaccine types, or car models.
- logical/boolean values are either `TRUE` or `FALSE`
- Characters and strings are words or letters.
  - "hello". "Alaska". "aagta". "September 1st". "19".

We have many tools for converting between types and identifying types. `is.numeric`, `factor()`, `levels()`.

R – Misc

Functions look like `f(arg1,arg2,...)`
And are incredibly useful. e.g. Create new variables:

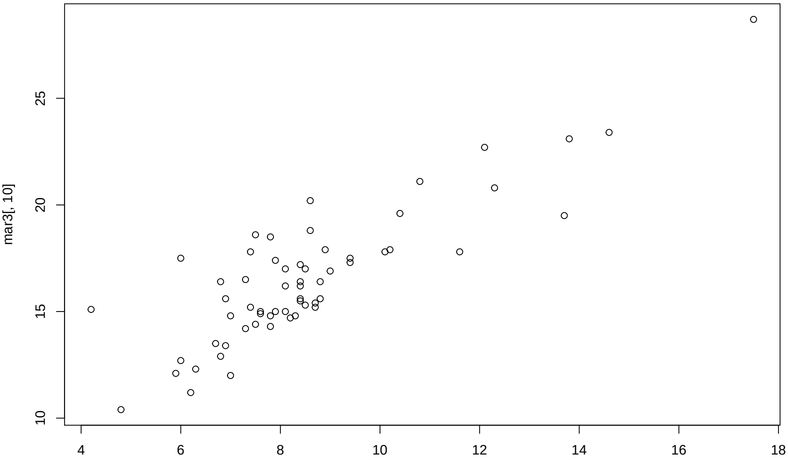
```
mar3$log.doses = log(mar3$Total.Doses.Delivered)
```

Finding out how a function works is simple `?log` will take you to the help page.

R – Plotting

Is straightforward for simple plots, but rapidly gets complicated.

```r
plot(mar3[,10]~mar3[,14])
```



R – Plotting

Histograms.

```
hist(mar3$Doses.Delivered.per.100K)
```

```
## Error in hist.default(mar3$Doses.Delivered.per.100K): ’x
```

```
typeof(mar3$Doses.Delivered.per.100K)
```

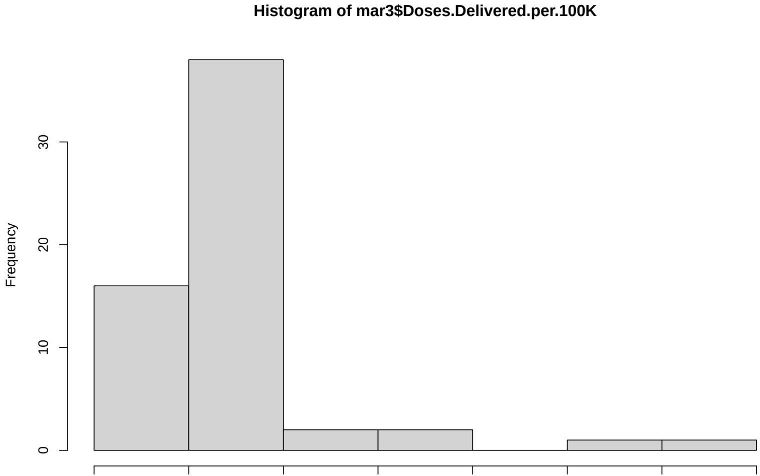
```
## [1] "character"
```

```
head(mar3$Doses.Delivered.per.100K)
```

```
## [1] "50950" "29803" "31912" "52883" "30842" "N/A"
```

R – Plotting

```
mar3$Doses.Delivered.per.100K =
  as.numeric(mar3$Doses.Delivered.per.100K)
hist(mar3$Doses.Delivered.per.100K)
```



R – Terrible Colnames

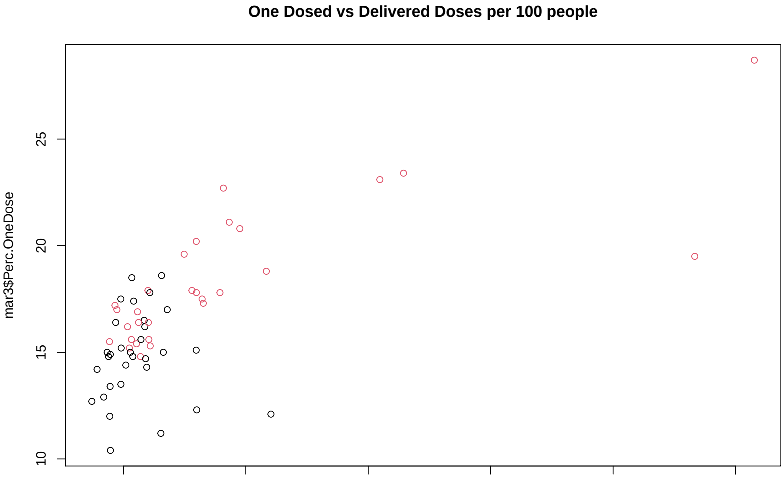
```r
colnames(mar3)[10:15]
```

```
## [1] "Percent.of.Total.Pop.with.1..Doses.by.State.of.Res
## [2] "People.18..with.1..Doses.by.State.of.Residence"
## [3] "Percent.of.18..Pop.with.1..Doses.by.State.of.Reside
## [4] "People.with.2.Doses.by.State.of.Residence"
## [5] "Percent.of.Total.Pop.with.2.Doses.by.State.of.Resid
## [6] "People.18..with.2.Doses.by.State.of.Residence"
```

```r
colnames(mar3) = c("State","Delivered","Delivered.100k",
   "X18.Delivered.100k","Administered","Administered.100k",
   "X18.Administered","X18.Administered.100k","OneDose",
   "Perc.OneDose","X18.OneDose","X18.Perc.OneDose",
   "TwoDose","Perc.TwoDose","X18.TwoDose","X18.Perc.TwoDose
```

R – Plotting - In color?

```r
plot(mar3$Delivered.100k/1000,mar3$Perc.OneDose,
     col=(mar3$Perc.TwoDose > 8.2)+1,
     main = "One Dosed vs Delivered Doses per 100 people")
```



R – Scatterplots

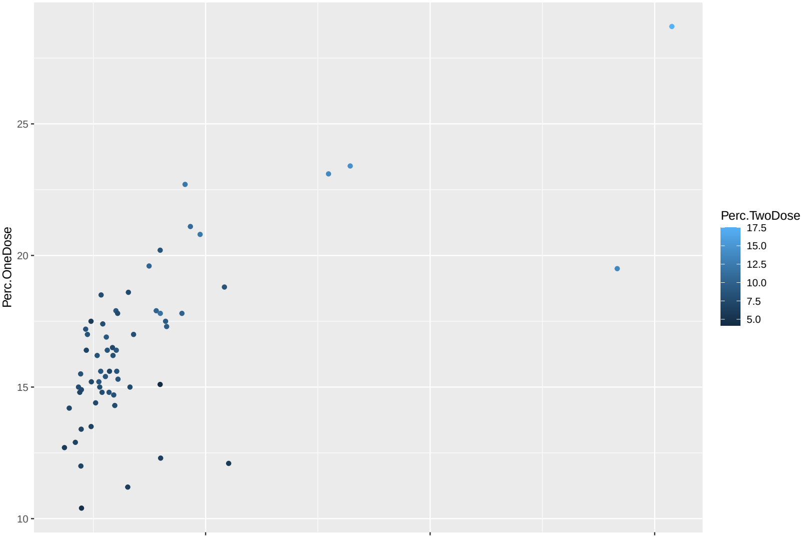
Scatterplots are fundamental to statistics. If you can show a compelling scatterplot, you can convince people of anything.

In part, this is because they are able to contain nearly as many dimensions of information as we can interpret. We tend to max out at 3-4 dimensions.

Critical to reduce dimensions down to interpretable level for good visualization.

R – Tidyverse

```
library(tidyverse)
ggplot(mar3,aes(x=Delivered.100k/1000,y = Perc.OneDose,col=
```



Statistics

R – Regression

```r
fit = lm(Perc.OneDose~Delivered.100k+Perc.TwoDose, data=mar
summary(fit)
```

```
##
## Call:
## lm(formula = Perc.OneDose ~ Delivered.100k + Perc.TwoDos
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3031 -1.1608 -0.3072  0.9775  3.9143
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.469e+00  8.593e-01   7.528 4.20e-10 **
## Delivered.100k 9.691e-06  3.525e-05   0.275    0.784
## Perc.TwoDose   1.138e+00  1.468e-01   7.752 1.77e-10 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
```

Regression

The basic model is as follows:
$E[Perc.OneDose] = \beta_0 + \beta_1 Delivered.100k +$
$\beta_2 Perc.TwoDose$

Or, similarly:

$Perc.OneDose = \beta_0 + \beta_1 Delivered.100k + \beta_2 Perc.TwoDose + \epsilon$
Where $E[\epsilon] = 0$.

Hypothesis Testing

Is “Delivered.100k” important?

H0: $\beta_1 = 0$.

```
summary(fit)
```

```
##
## Call:
## lm(formula = Perc.OneDose ~ Delivered.100k + Perc.TwoDos
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3031 -1.1608 -0.3072  0.9775  3.9143
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.469e+00  8.593e-01   7.528 4.20e-10 **
## Delivered.100k 9.691e-06  3.525e-05   0.275    0.784
## Perc.TwoDose   1.138e+00  1.468e-01   7.752 1.77e-10 **
##
```

A Single Test

The “p-value” is the probability of observing an event at least as extreme as the observed event, assuming that the null hypothesis is true.

Testing Procedure Choose a cutoff α. Conclude that there is a real effect if $p < \alpha$.

This procedure tries to give only α probability of a false positive.

$$P(p < \alpha | H0) = \alpha$$

Challenges

Big Data situations pose two different problems for p-values.

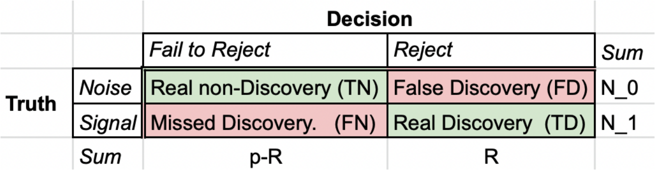
1. With a big enough sample (i.e. large n), we wind up rejecting the null even for effects that are of no practical significance, but which have a small positive β.
2. With many parameters (i.e. large p), we are performing many tests, and need to account for the number of tests we perform.

Large Scale Testing

We wish to test p simultaneous null hypothesis:

$$H0_1, H0_2, ..., H0_p$$

Out of the p null hypothesis, N_0 are true nulls and $N_1 = p - N_0$ are false – i.e. there is an effect.



		Decision		
		Fail to Reject	*Reject*	*Sum*
Truth	*Noise*	Real non-Discovery (TN)	False Discovery (FD)	N_0
	Signal	Missed Discovery. (FN)	Real Discovery (TD)	N_1
	Sum	p-R	R	

Multiplicity

α is typically for a single test. If we repeat p tests, we expect $p\alpha$ significant even when all the null hypotheses are true.

Suppose we have 100 regression coefficents, of which 5 are true effects. Now suppose we test all 100 coefficients, all 5 true effects come up (if we're lucky), as well as $95\alpha \approx 5$ more. So of the significant variables, 50% are fake.

This is the "false discovery proportion".

It depends closely on α, the underlying truth, and how good our methods are. But it can be really big.

False Discovery Rate

Big data is about making choices.
Instead of single tests, we'll consider:

FD Proportion = False positives / Significant = $\frac{FD}{R}$

FDP is a property of the fitted model and the truth. We don't know it.

We can control its expectation though: False Discovery Rate, $FDR = E[FDP]$.

This is the multivariate analogue of α.

If all tests are tested at α level, we have $\alpha = E[FD/N_0]$, whereas $FDR = E[FD/R]$

FDR Control

Suppose we want to know that $FDR \leq q = 0.1$.

Benjamini + Hochberg Algorithm

1. Rank your p-values smallest to largest.
2. Set p-value cutoff as $\alpha^* = max\{p_{(k)} : p_{(k)} \leq q\frac{k}{p}\}$

Then $FDR \leq q$ – assuming approximate independence between tests.

Wrap up

Things to do

Before Thursday:

- Group up on canvas
- Submit predictions
- Play with R, tidyverse, ggplot2

Rehash

- I'm Connor Dowd
- Big Data brings new problems
- R is important
- Graphs are important
- False Discovery Rates can be controlled

Bye!