3 STATISTICAL TRICKS EVERY HACKER SHOULD KNOW

adam hogan
@mittenchops on github
Presentation for hackNY Masters

New York, NY
September 28, 2013

OUR GOAL

- Learn 3 tricks that solve a lot of problems well enough to become more effective
- Enough to know what to look up when you get in over your head.
- Really important things we will not get into:
  - Model diagnostics
  - Back-testing, cross-validation
  - Theory behind stuff like Markov models

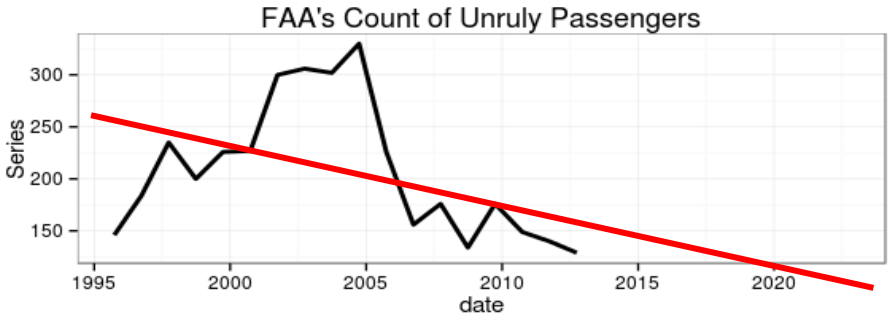
FORECASTING: HI, ARIMA MODELS

WHAT IT SOLVES

- I want to predict future values based on a bunch of values I know about the past.
  - To do it easily
  - To be generally right.
- Even if the data moves around a lot, this is safe when I think the underlying process that generates it stays the same over time.
- Best when the process doesn't rely on knowing a lot of past information, many periods back.

BASIC THEORY

- First thought is maybe just draw a "trend line"---or a regression.
- This neglects something very powerful though---you have the ORDER of items



- Y1, Y2, Y3 may depend on the change in the value before them, not the value itself.

not $Y \sim X$ but $Y_n \sim Y_{n-1}$

AUTOREGRESSIVE INTEGRATED MOVING AVERAGE MODELS

AR Integration MA

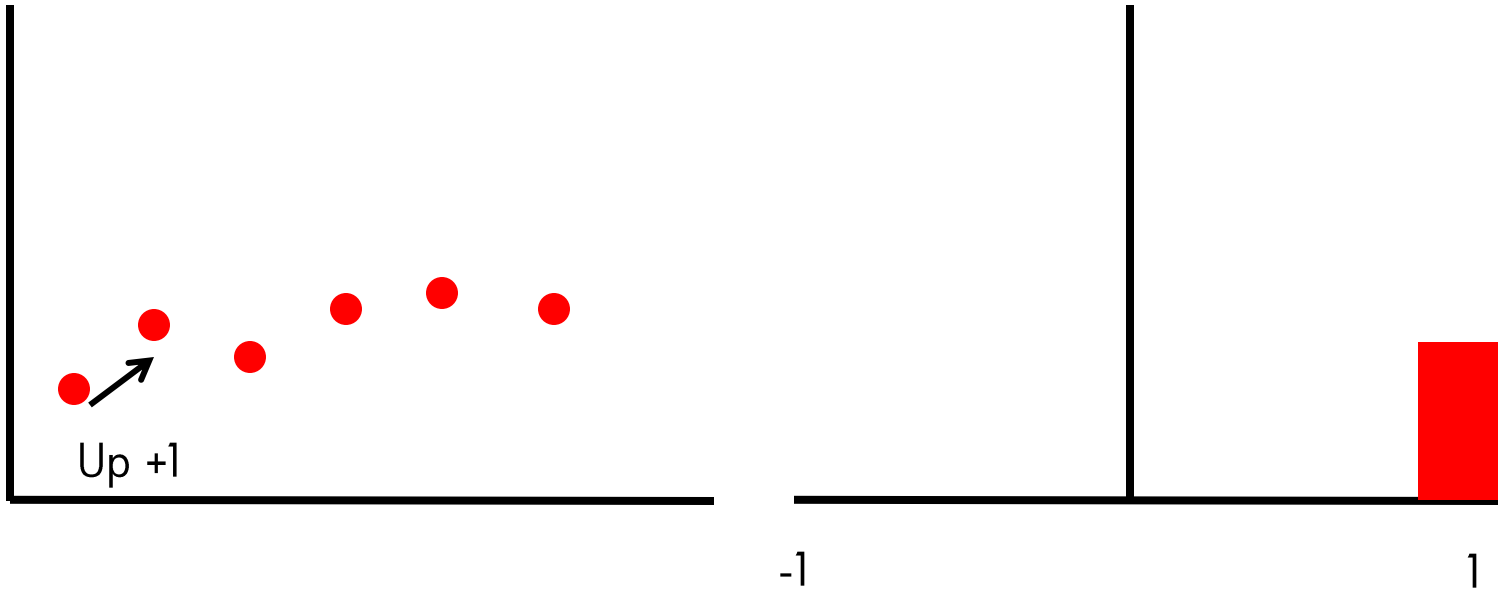
```
ARIMA[P,I,Q]
```

- AR = the trend component
- MA = the mean-reversion component

- The forecast is the dynamics between following the trend and going back to normal---like a spring.

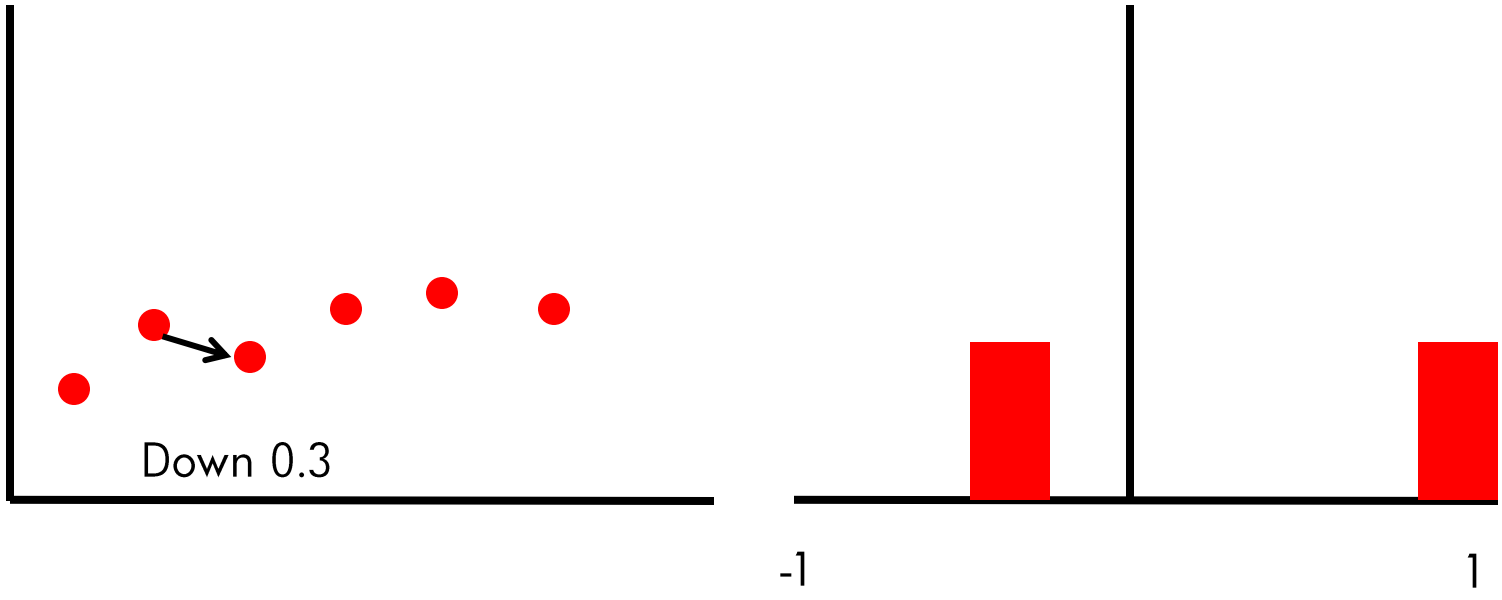
DIFFERENCING

- Often, you're not regressing on the levels, but on the relationship between Y_n and Y_{n-1}



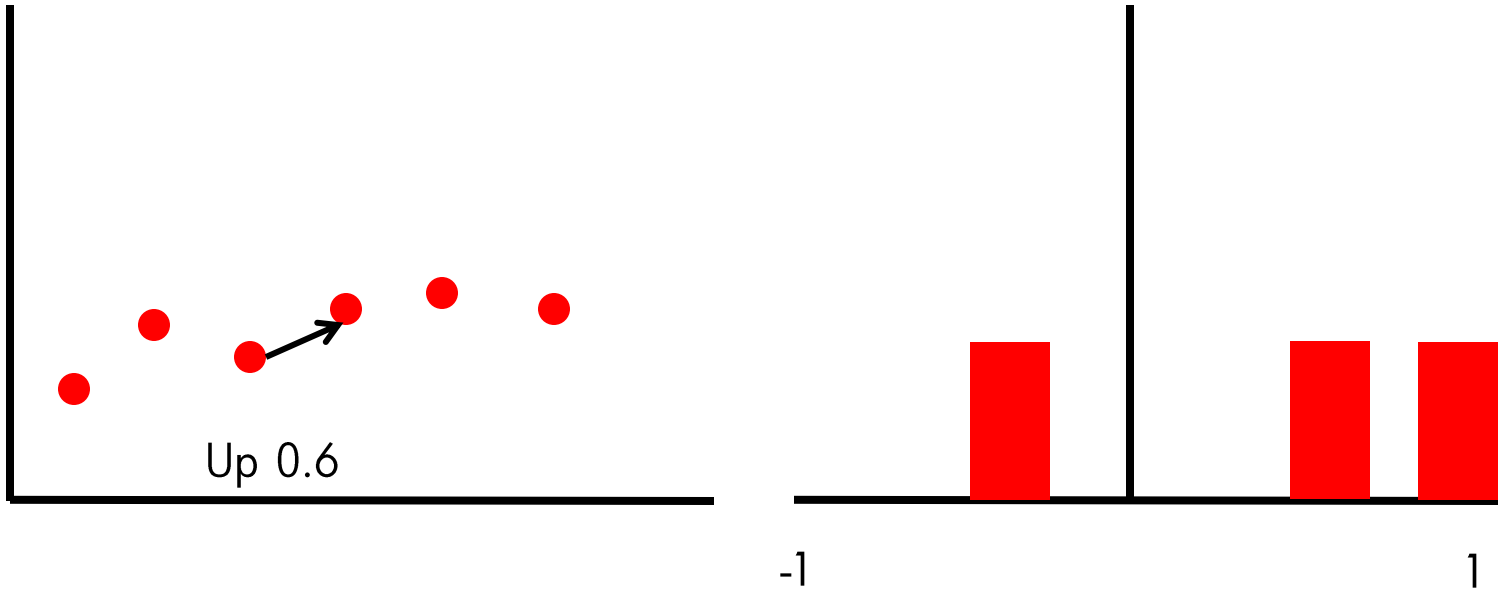
DIFFERENCING

- Often, you're not regressing on the levels, but on the relationship between Y_n and Y_{n-1}



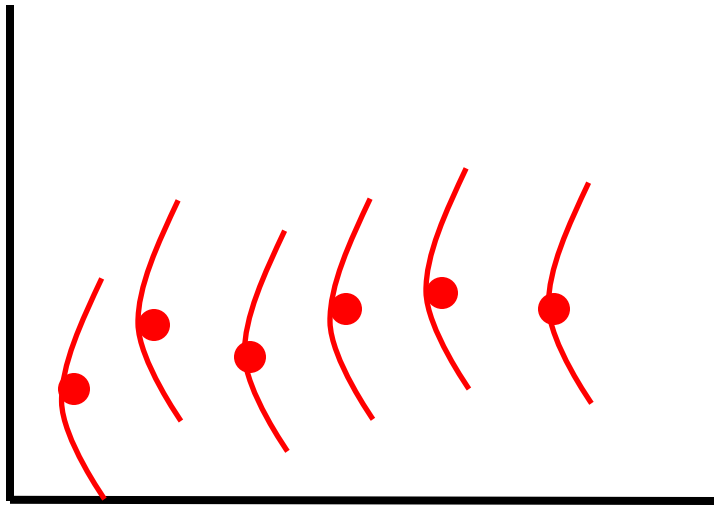
DIFFERENCING

- Often, you're not regressing on the levels, but on the relationship between Y_n and Y_{n-1}



THE I PART

- You want the relationship between entries to be consistent over time.



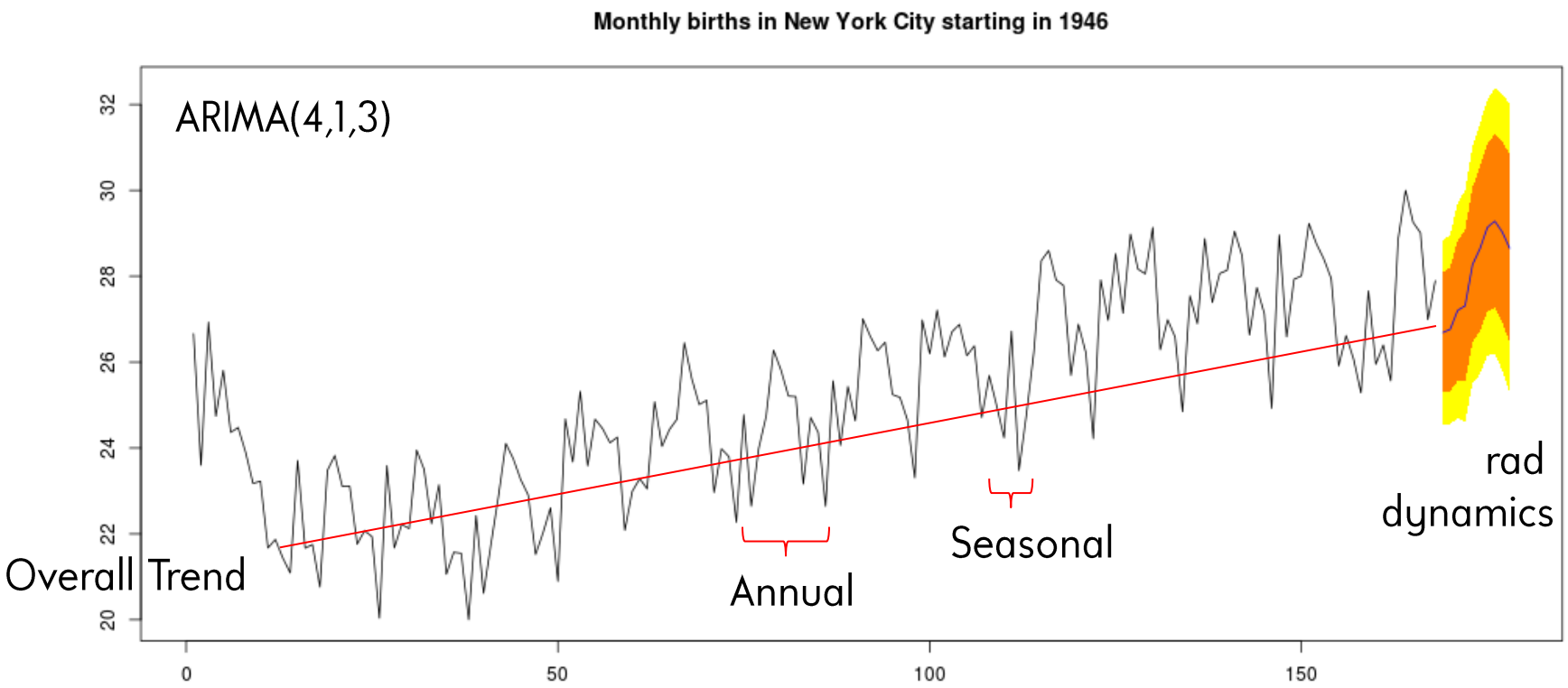
- Each point is telling you something about the distribution of the relationship between points
- Take differences until that's stable.

Model at the level where the randomness lives, not a function of the randomness

INTERPRETING P AND Q

- AR(p) and MA(q) deal with how many observations back continue to impact the present---this is how you deal with seasonality.
- The last 4 periods might be a financial quarter, the last 12 a year, maybe your series depends on a rhythm of time of day.
- A high order model, like ARIMA(12,2,5) can incorporate some really rad dynamics.

BIRTHS IN NYC OVER TIME

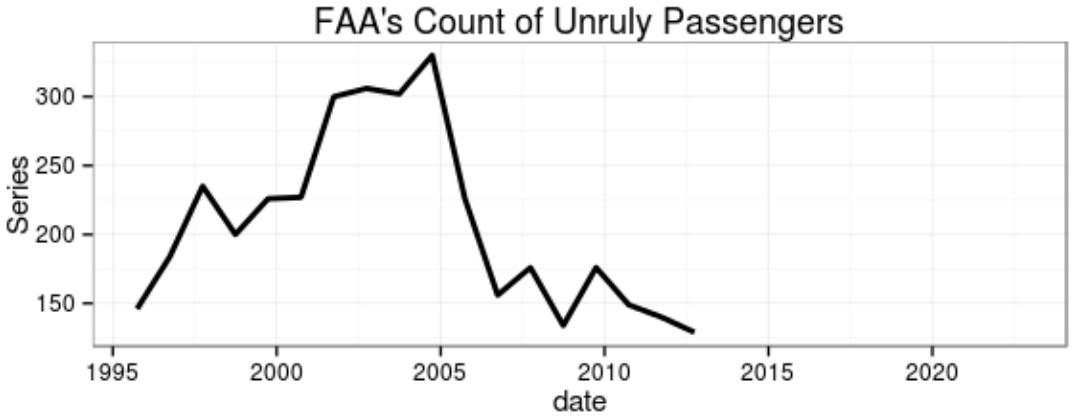


UNRULY PASSENGERS

- If we have a record of how many Americans were misbehaved on airplanes in the past, how many will there be next year?
  - Maybe I'm an airline, and I want to make security decisions
  - Maybe I'm a passenger, and I want to decide how ruly to be.

DATA

```
> library(XML)
> url2 <-
'http://www.faa.gov/data_research/passengers_cargo/unruly_passengers/'
> X <- readHTMLTable(url2, header=T, stringsAsFactors=FALSE)[[1]]
> X
   Year                    Total
1  1995                      146
2  1996                      184
3  1997                      235
4  1998                      200
5  1999                      226
6  2000                      227
7  2001                      300
8  2002                      306
9  2003                      302
10 2004                      330
11 2005                      226
12 2006                      156
13 2007                      176
14 2008                      134
15 2009                      176
16 2010                      149
17 2011                      140
18 2012                      129
```



CODE

- Get your data
- Convert to a date-indexed time series (xts is good)
- Apply `auto.arima(data)`
- Brute force searches for the best p,i,q values that optimize your model.

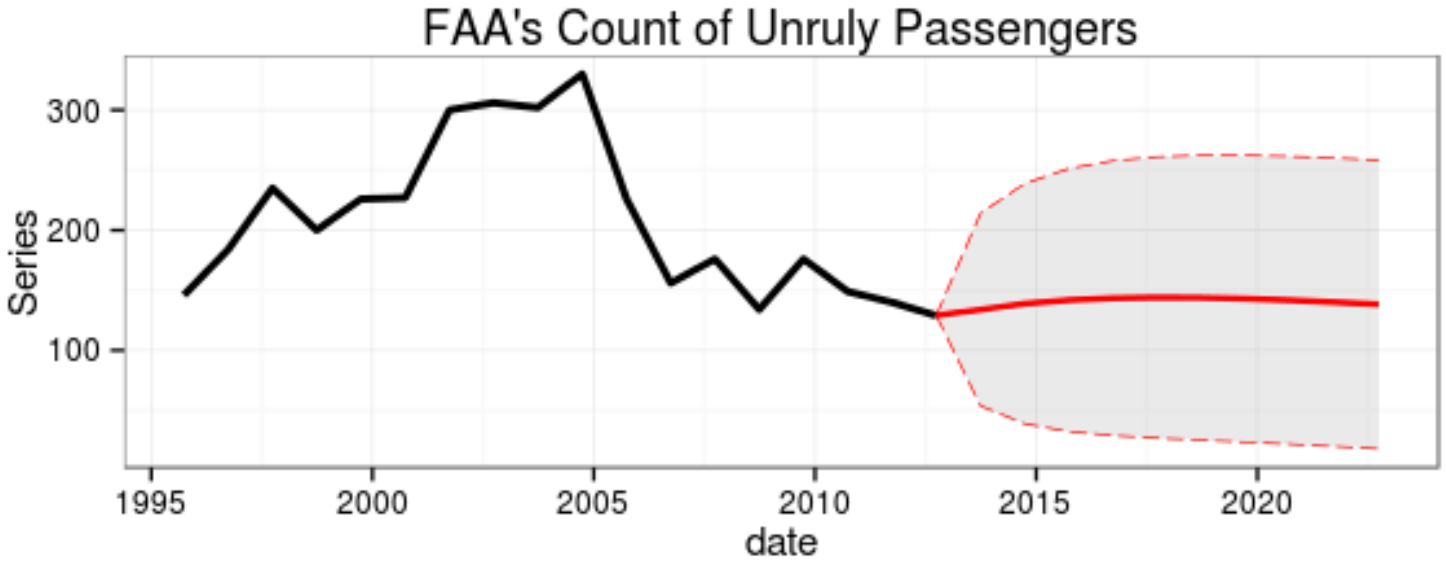
```
url2 <-
'http://www.faa.gov/data_research/passengers_cargo/unruly_passengers/'
X <- readHTMLTable(url2, header=T, stringsAsFactors=FALSE)[[1]]
unruliness <- as.xts(as.numeric(X[-19,2]), order.by = as.Date(X[-19,1], format="%Y"))

nobs <- length(unruliness[,1])
fit <-
auto.arima(as.vector(unruliness), xreg=1:nobs-1, ic="bic")
# (1,0,0)
fore <- predict(fit, 10, newxreg=(nobs+1):(nobs+10))
```

For X-validation, try K-folds: https://gist.github.com/mittenchops/5354856

RESULT

- This one returns an AR(1,0,0), but the coefficient is really weak, and the last values weren't strong, so we're going to predict a pretty flat future.
- Close to a random walk---written ARIMA(0,1,0)



TIME FOR A BETTER MODEL

- Highly seasonal
  - A different model, SARIMA, helps you separate seasonal effects better
  - ARFIMA is trendy right now.
- Other distribution
  - May need to do more transforms on the series first
  - GARCH() family of models---most of finance, econ
- Series changes over time ("regime-switching")
  - Lot of Econometric series do this.
- Diagnosing the terms is really hard to do well, but auto.arima should help you do it well enough for a lot of series.

THE SAME OR DIFFERENT?
HI, T-TEST

WHAT IT SOLVES

- There's a lot of randomness. If you start getting different numbers from what you've seen before---are they within the bounds of normal randomness, or do they suggest something new is happening?
- When people say "statistically different" this is usually what they mean.
- Pretty simple, amazingly useful, undiscovered until 1908.

BASIC THEORY

- This table is happening:

X	Y
33.4	24.4
33.3	24.3
33.1	24.2
33	24.1
32.9	24
32.8	23.9
…	…
AVG, 32.6	AVG, 23.7

- If X and Y are actually from the same distribution, the difference between their means will approach zero as we observe more temperatures.
- t-test asks how different that difference is from 0

IS IT REALLY COLDER IN CHICAGO?

- When I first moved here, New Yorkers would say, "Chicago is nice, but it's soooo cold!"
- Chicago is at 41.8°N, New York at 40.7°N; that's only 60 miles on sphere earth. That can't be that much colder.
- Is it actually colder there, or is that just rumor, myth, superstition, View of the World from 9th Avenue, etc.?

CODE

- Get your data
- This function is baked into R, python, even excel. Easy.

```
cparkjan <-
c(33.4,33.3,33.1,33,32.9,32.8
,32.7,32.6,32.5,32.4,32.4,32.
3,32.3,32.3,32.3,32.2,32.2,32
.3,32.3,32.3,32.3,32.4,32.4,3
2.5,32.5,32.6,32.7,32.8,32.9,
33,33.1) # central park in
january
ordinjan <-
c(24.4,24.3,24.2,24.1,24,23.9
,23.9,23.8,23.7,23.7,23.6,23.
6,23.5,23.5,23.5,23.5,23.4,23
.4,23.4,23.4,23.5,23.5,23.5,2
3.6,23.6,23.7,23.8,23.9,24,24
.1,24.2) # chicago in january


t.test(cparkjan,ordinjan,pair
ed=T)
```

```
        Paired t-test

data:  cparkjan and ordinjan
t = 644.8545, df = 30, p-value < 2.2e-16
alternative hypothesis: true difference
in means is not equal to 0
95 percent confidence interval:
 8.830011 8.886118
sample estimates:
mean of the differences
               8.858065
```

CODE

- Get your data
- This function is baked into R, python, even excel. Easy.

```
cparkjan <-
c(33.4,33.3,33.1,33,32.9,32.8
,32.7,32.6,32.5,32.4,32.4,32.
3,32.3,32.3,32.3,32.2,32.2,32
.3,32.3,32.3,32.3,32.4,32.4,3
2.5,32.5,32.6,32.7,32.8,32.9,
33,33.1) # central park in
january
ordinjan <-
c(24.4,24.3,24.2,24.1,24,23.9
,23.9,23.8,23.7,23.7,23.6,23.
6,23.5,23.5,23.5,23.5,23.4,23
.4,23.4,23.4,23.5,23.5,23.5,2
3.6,23.6,23.7,23.8,23.9,24,24
.1,24.2) # chicago in january


t.test(cparkjan,ordinjan,pair
ed=T)
```

```
        Paired t-test

data:  cparkjan and ordinjan
t = 644.8545, df = 30, p-value < 2.2e-16
alternative hypothesis: true difference
in means is not equal to 0
95 percent confidence interval:
 8.830011 8.886118
sample estimates:
mean of the differences
                8.858065
```

CHICAGO IS COLDER

WAIT, LET'S TRY SUMMER!

- Get your data
- This function is baked into R, python, even excel. Easy.

```
cparkjul <-
c(75.3,75.5,75.6,75.8,75.9
,76,76.1,76.2,76.3,76.4,76
.5,76.6,76.6,76.7,76.7,76.
7,76.8,76.8,76.8,76.8,76.8
,76.8,76.8,76.8,76.8,76.7,
76.7,76.7,76.7,76.6,76.6)
ordinjul <-
c(73.2,73.3,73.5,73.6,73.7
,73.8,73.9,74,74.1,74.2,74
.2,74.3,74.3,74.3,74.3,74.
3,74.3,74.3,74.3,74.2,74.2
,74.2,74.1,74.1,74,74,73.9
,73.9,73.8,73.8,73.7)


t.test(cparkjul,
ordinjul,paired=T)
```

```
        Paired t-test

data:  cparkjul and ordinjul
t = 53.4294, df = 30, p-value < 2.2e-16
alternative hypothesis: true difference
in means is not equal to 0
95 percent confidence interval:
 2.367211 2.555370
sample estimates:
mean of the differences
                2.46129
```

WAIT, LET'S TRY SUMMER!

- Get your data
- This function is baked into R, python, even excel. Easy.

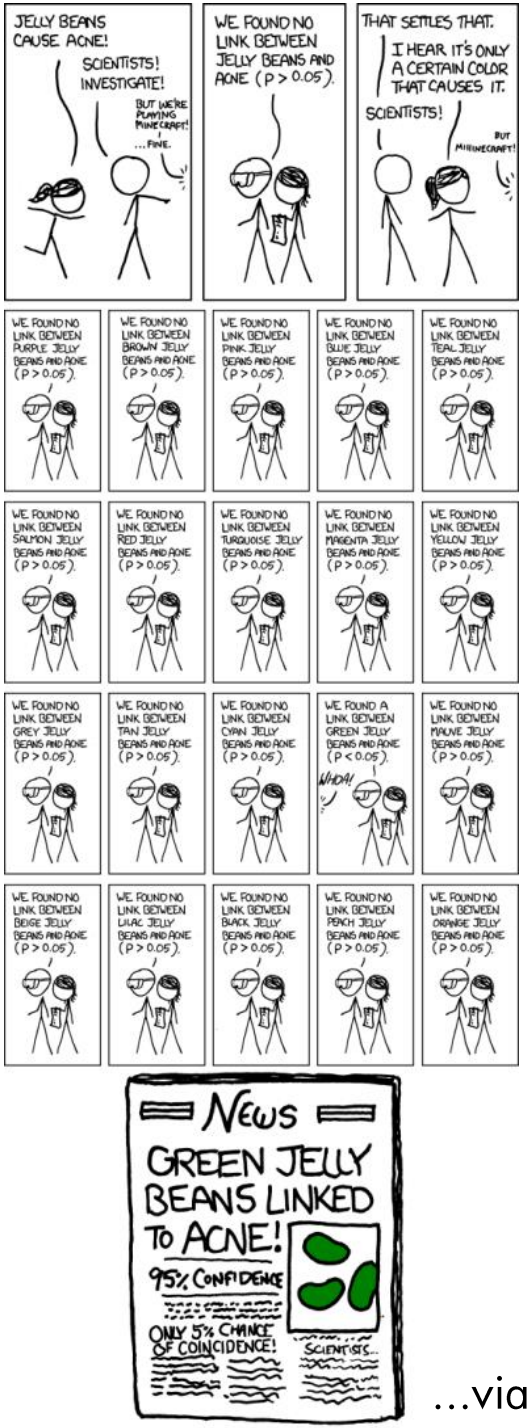
```
cparkjul <-
c(75.3,75.5,75.6,75.6,75.9
,76,76.1,76.2,76.3,76.4,76
.5,76.6,76.6,76.7,76.7,76.
7,76.8,76.8,76.8,76.8,76.8
,76.8,76.8,76.8,76.8,76.7,
76.7,76.7,76.7,76.6,76.6)
ordinjul <-
c(73.2,73.3,73.5,73.6,73.7
,73.8,73.9,74,74.1,74.2,74
.2,74.3,74.3,74.3,74.3,74.
3,74.3,74.3,74.3,74.2,74.2
,74.2,74.1,74.1,74,74,73.9
,73.9,73.8,73.8,73.7)


t.test(cparkjul,
ordinjul,paired=T)
```

```
        Paired t-test

data:  cparkjul and ordinjul
t = 53.4294, df = 30, p-value < 2.2e-16
alternative hypothesis: true difference
in means is not equal to 0
95 percent confidence interval:
 2.367211 2.555370
sample estimates:
mean of the differences
                2.46129
```

CHICAGO IS COLDER IN THE SUMMER TOO!



...via xkcd

WELL, THERE ARE 12 MONTHS

- Well, maybe this year was special---I need to compare 2 years ago Chicago against last year New York.
- This kind of thinking is where the lies and damn lies come in.

TIME FOR A BETTER MODEL

- When you suspect more than the mean matters
  - Kolmogorov-Smirnov test
- When you have more than 2 groups
  - ANOVA
- High non-normality, have a lot of dependent things
  - Paired, rank-tests, non-parametrics, lots and lots of other tests
- The important thing is testing! This simple question, are X and Y the same or not, is very powerful!

ODDS OF YES OR NO: HI, LOGISTIC REGRESSION

WHAT IT SOLVES

- Lots of stuff you can reduce to a yes/no question.
- (If you're creative enough, you can reduce *anything* to a yes/no question.)

BASIC THEORY

- Draw a boundary between these regions
- Machine learning people call this a classifier.

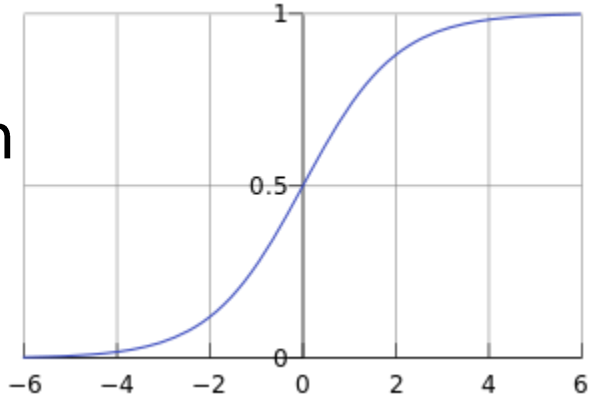
Actually the "logit" shhh

$$F(x) = \exp(\beta_0 + \beta_1 x)$$

Like a regular linear regression

But the logistic function part keeps it between 0 and 1

- The output is a number you can interpret as a probability.



EXAMPLE: WILL I FIND THIS CAT VIDEO INTERESTING?

- Let's say I want to make a robo cat-video watcher that will deliver only excellent cat videos into my inbox every day.
- How can I teach a robot what cat videos I would like?
- Let's take some cat-video parameters:
  - Is there music playing in the cat video?
  - How long is it?
  - How many views does it have?
  - How many cats are in the video?
  - Do the cats jump?
  - Are there humans in the video?

DATA

Y	url	music	length	numofcats	views	jumping	nopeople	title
1	…	1	6:28	1	3055989	1	1	I am maru
1	…	0	5:30	1	3541708	1	1	maru greatest hits IV
0	…	0	0:56	1	13000000	0	0	the original grumpy cat
0	…	1	2:29	0	14759405	0	0	cat friend versus dog friend
1	…	0	2:58	1	11126592	1	1	many too small boxes and maru
1	…	1	1:32	23	4908988	1	0	french ninja cats
1	…	1	2:30	1	2421475	0	1	henri, le chat
1	…	1	2:06	2	7895029	0	0	henri, paw de deux
0	…	1	2:40	15	3888232	1	0	Top 10 best cat videos of all time
1	…	0	0:25	2	309005	0	1	hana turns and maru watches
1	…	0	0:23	1	16848708	0	1	Cat gets caught barking by human and resumes
1	…	0	1:15	1	37264312	1	1	Brave kitten stands up to dog
0	…	0	0:30	1	10295	0	0	Hamilton the hipster cat
0	…	1	1:37	1	750745	0	0	Colonel Meow â€“ Longest cat fur on record
?	…	1	3:21	1	3078151	1	1	I am maru 5

CODE

- Bunch of omitted data-munging normalizing scores…
- Model step is `glm()`
- `Predict()` step asks, "what will he think about 'I am Maru 5?'"
- The model is outrageously sure I'll like it.

```
# INPUT
catdat <- read.csv(file='cats.csv',
header=T, stringsAsFactors=F)

# CUT OUT URL, TITLE VARS
cdata <- catdat[,-c(2,9)]
leftout <- cdata[last,]

# LOGISTIC REGRESSION
lfit <- glm(Y~., data=cdata[-last,],
family=binomial(link="logit"))

predict(lfit,leftout, type='response')
# 15
#  1

# BONUS REGRESSION TREE MODEL!
rfit <- rpart(Y ~ ., data=cdata[-last,])

predict(rfit, leftout[,-1])
#        15
# 0.6428571
```

Regression tree model is less sure.

CAUTION!

- Check your model! I actually don't have enough data to produce significant results, here. I would need to watch a lot more cat videos than 20.
- Use fewer parameters, test variations. ML-people call this research; social scientists call it data mining.
- Interpretation: Probability, odds ratios, log odds ratios--interpretation of the coefficients of logistic regression is more involved. Consult a statistical professional before making statements like "having music in the video makes you 10x more likely to want to see it!"

TIME FOR A BETTER MODEL

- You start categorizing things that answer the questions:
  - "how much" (linear regression);
  - "until when" (exponential); "how many by a certain time" (poisson);
- You want to put things in N categories, instead of 2.
- You have a complex multivariate dependence structure, or the model just plain doesn't work.
  - Fisher's LDA
  - Random forest

ALAS, THE SEARCH CONTINUES

- I leave it to you to make the perfect cat video regression tree discovery engine.
- But now you can…
  - Predict the future
  - Say when stuff is different and when it isn't.
  - Give odds on whether something will happen.
- You can learn more by…
  - Studying time series models (get obsessed with stocks?)
  - Take a regression class (hard to learn it otherwise)
  - Experimenting a lot

THANKS

- Data from anonymized and public sources.
  - Time Series Data Library (Births in NYC)
  - NOAA (temperatures)
  - FAA (unruly passengers)
- Code snippets from lots of people, sorry if I missed crediting anyone

Want to talk more about R or stats?

adam hogan

github: @mittenchops