Knowledge for Analytics & Big Data

What's the role for statistical significance?

Bob Stine
Department of Statistics
The Wharton School
University of Pennsylvania

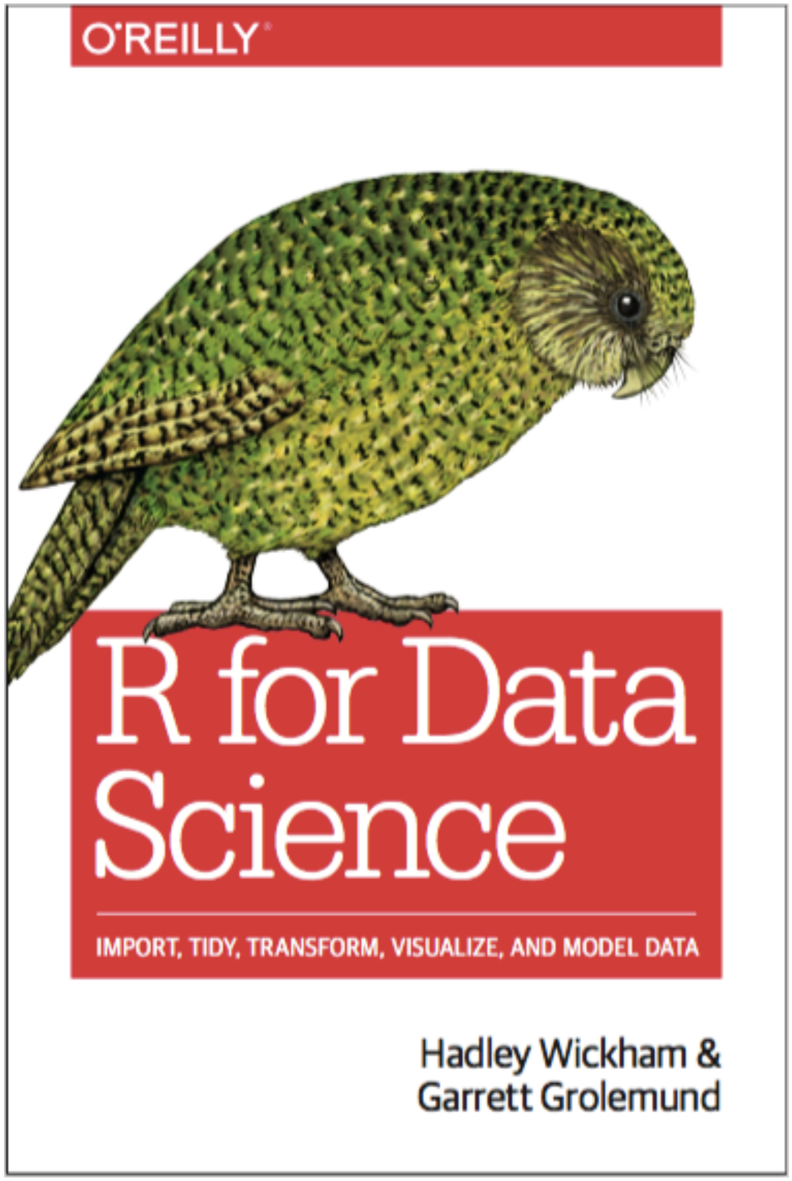
Wharton
Department of Statistics

An R Shout Out

Interested in how to use R for analytics?

Check out this book…

- different style of R
- ggplot
- rplyr
- and many others



Perspective

Motivation

- Let's not screw up this wave of interest in statistics (aka, data science)
- Unless we teach students to think carefully about significance with big data, they will think all we told them was wrong and forget us.

Standard error and significance are THE major concepts we bring to the table

- We need to make sure we convey these well.

Three concerns …

- Told through a sequence of examples

First Example

Question

- Do assets that perform well in one year also perform well the following year?
- That is, can we use performance this year to anticipate performance next year?

Not unique to finance and investing

Analogous situations

- Forecasting sales at Amazon
- Performance of retail market segments

Statistical Significance?

Question

Do assets that perform well in one year also perform well the following year?

That is, can we use performance this year to anticipate performance next year?

Data analysis

Simple regression

Regress of stock return of companies this year on stock return last year

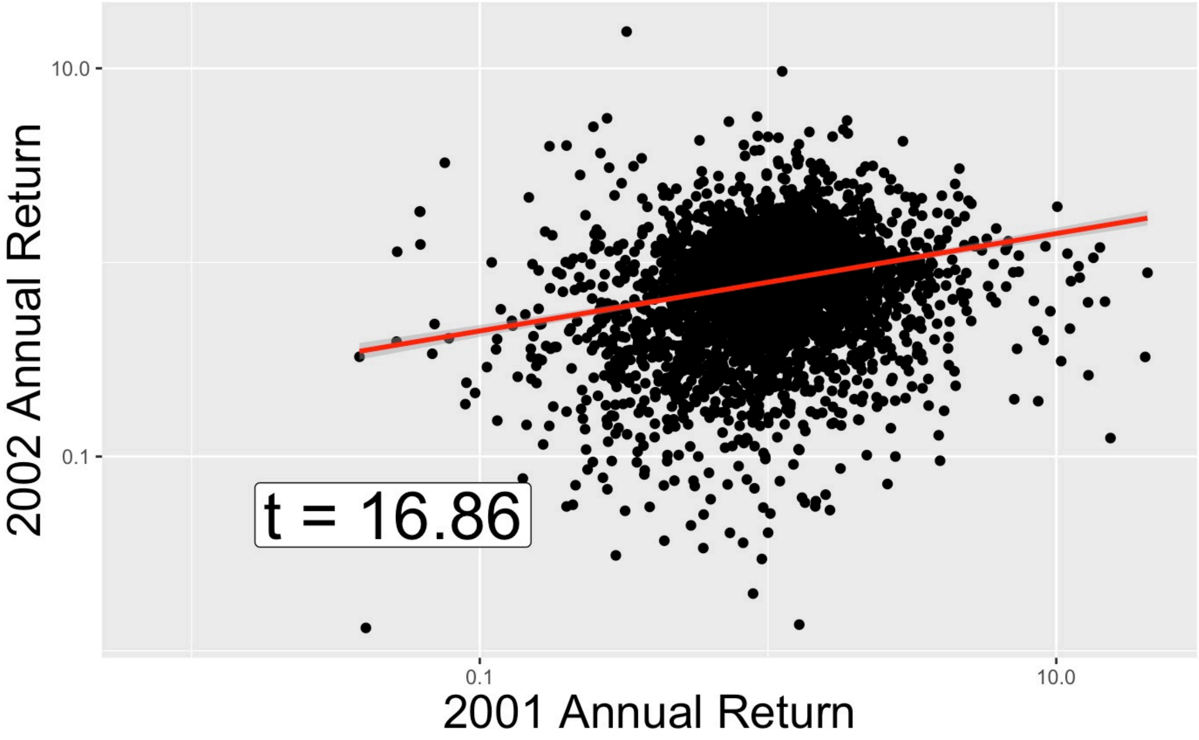
Lots of data: 3,500 assets in typical year.

Statistical Significance?

Data analysis

- Regress of stock return this year on stock return last year
- Significantly positive

2002
on
2001



Wharton
Department of Statistics

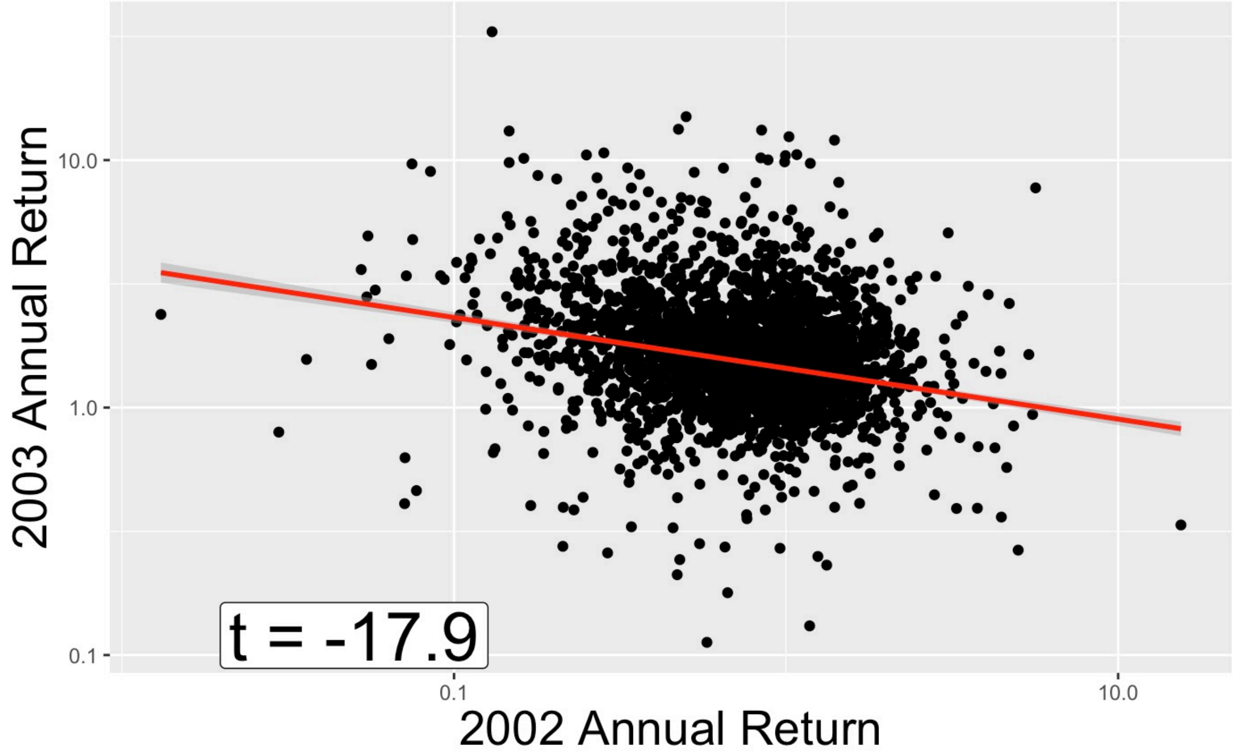
Statistical Significance?

Data analysis

Regress of stock return this year on stock return last year

Significantly negative!

2003
on
2002

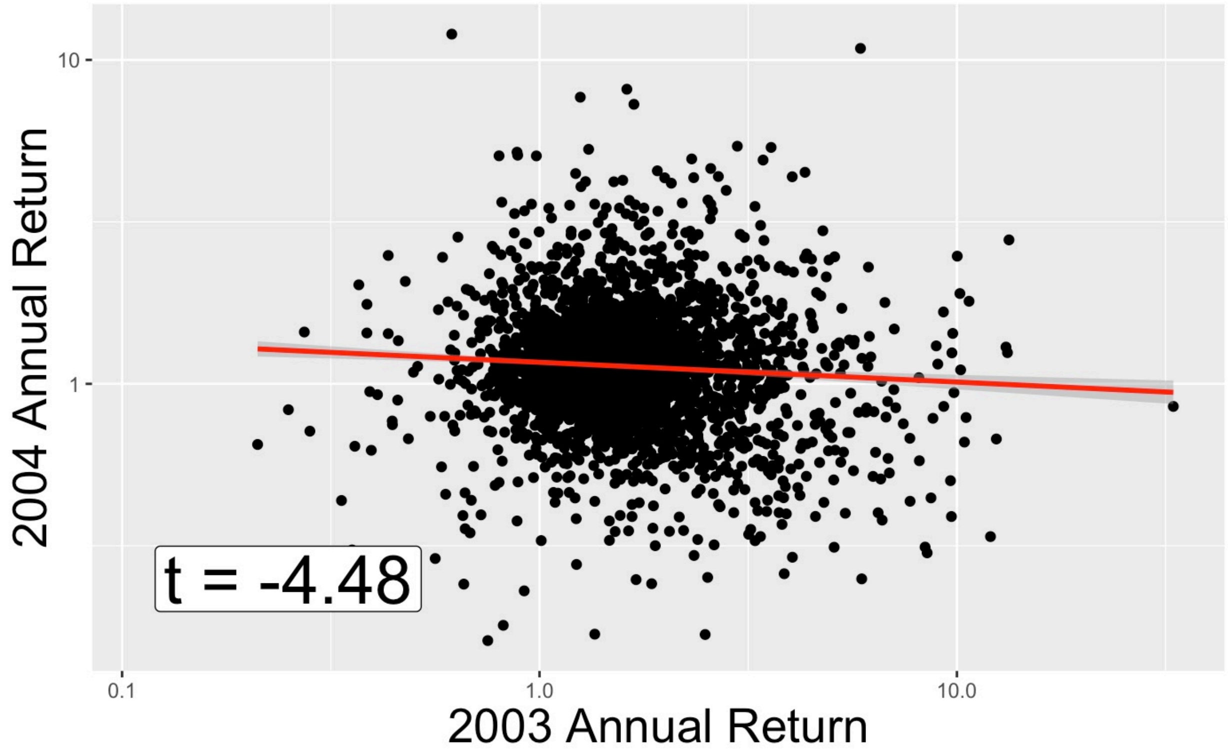


Statistical Significance?

Data analysis

- Regression of stock return this year on stock return last year
- Significantly negative

2004 on 2003



Wharton
Department of Statistics

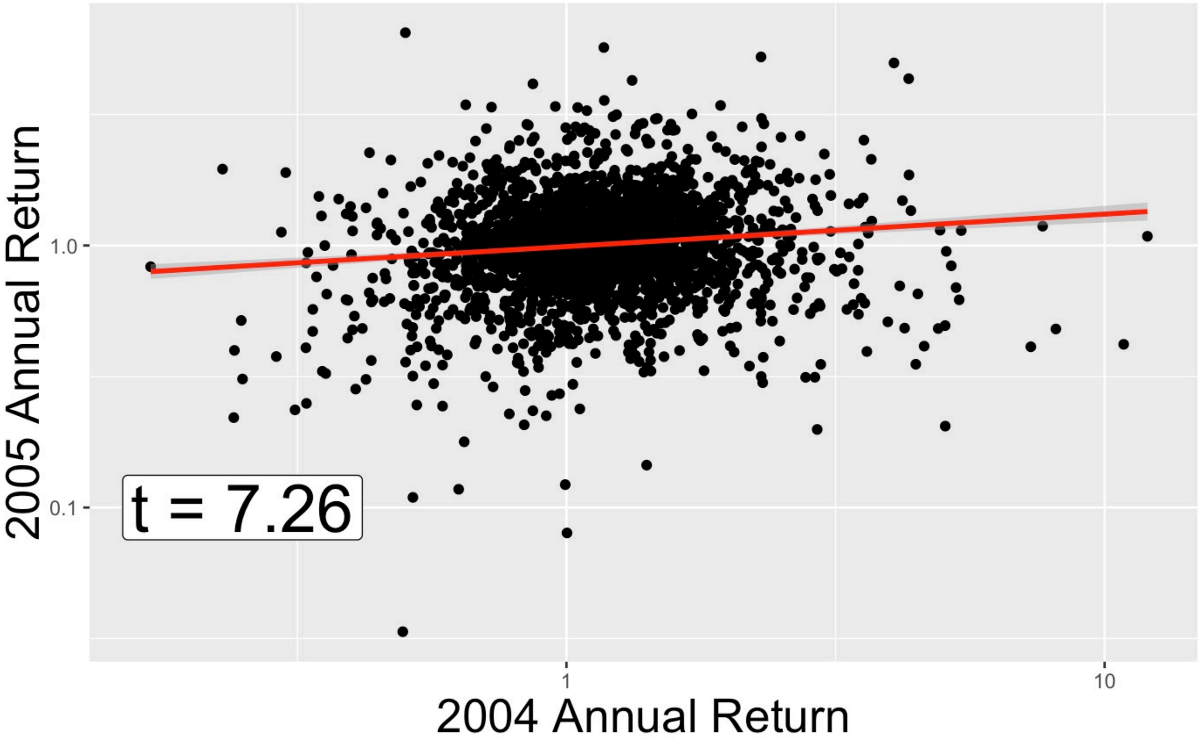
Statistical Significance?

Data analysis

Regression of stock return this year on stock return last year

Significantly positive

2005
on
2004



Wharton
Department of Statistics

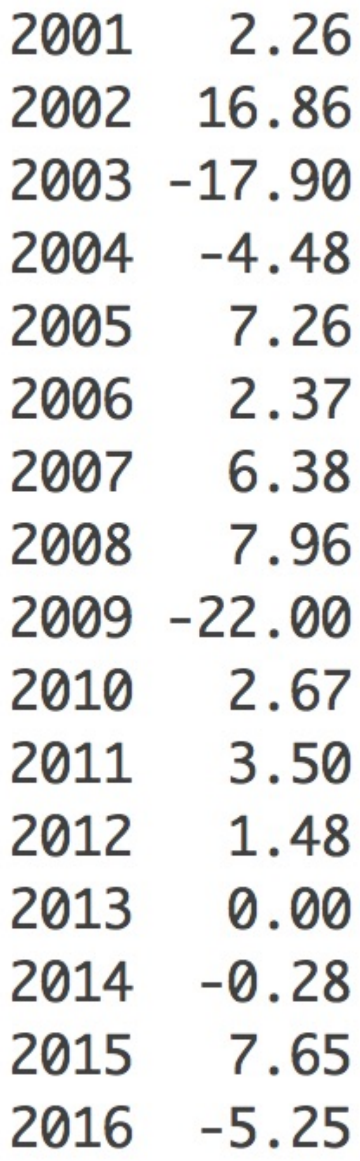
Statistical Significance?

Data analysis

t-statistics from regression of return this year on return last year

Question

What does it mean to find "significantly" positive one year, but "significantly" negative the next?



```
2001   2.26
2002  16.86
2003 -17.90
2004  -4.48
2005   7.26
2006   2.37
2007   6.38
2008   7.96
2009 -22.00
2010   2.67
2011   3.50
2012   1.48
2013   0.00
2014  -0.28
2015   7.65
2016  -5.25
```

Concern #1

Heuristic

Claims for hurricane insurance are very different from claims for car insurance

Explanation

Significance determined by effect size and sample size

Sample size = count of independent cases

Stocks not independent observations

All move in a correlated fashion

Lesson

Many rows in data table ≠ many degrees of freedom

Inference for years, not individual companies

See: hierarchical models, repeated measures, latent variables

Second Example

Question

- Do technical rules predict the movement of the overall stock market?

Again, not unique to finance

Analogous problems

- “Wide” data with more explanatory features than available cases.
- Deciding the location for a new retail outlet
  - Lots of possible features
  - Zip code, census, social media
- Genetics

Wharton
Department of Statistics

Second Example

Question

Do technical trading rules predict the direction and movement of overall stock market?

Results

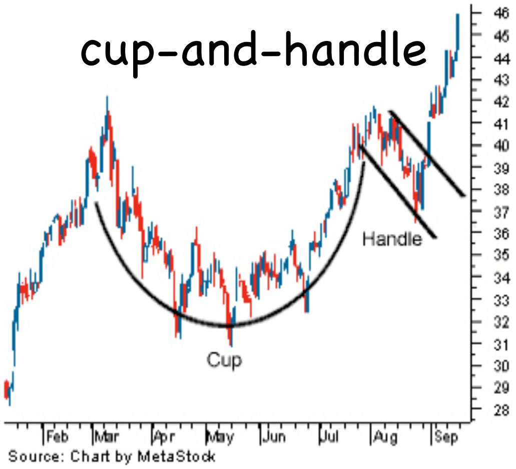
Regress daily returns (% change) on the S&P 500 stock market index in 2014

Predictors are technical trading rules based on observed properties of the market

Designed to be easy to extrapolate

Include combinations of these rules

cup-and-handle

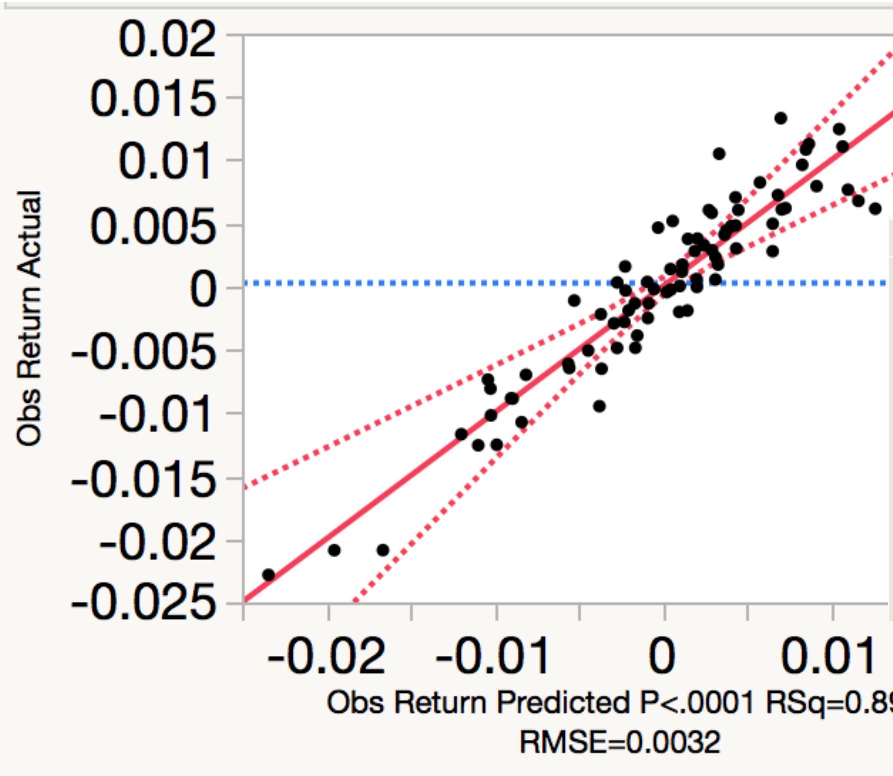


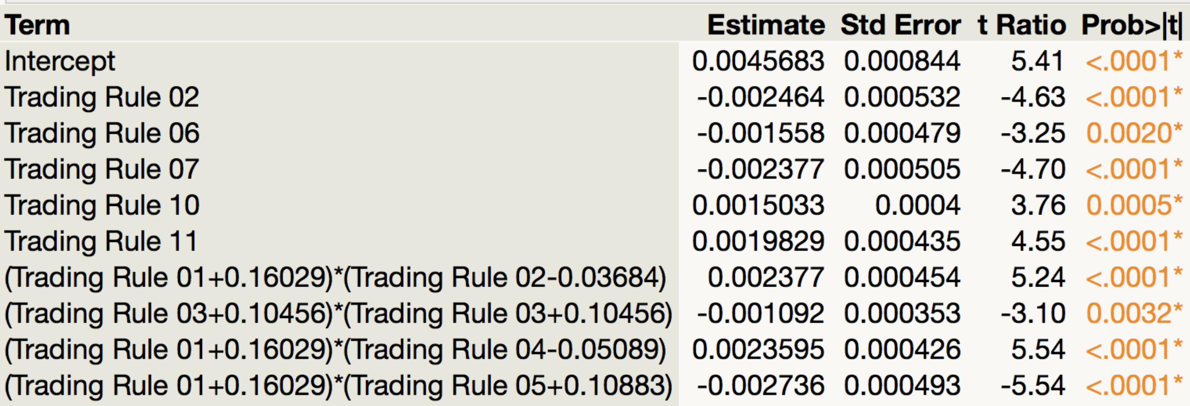
Source: Chart by MetaStock

Model Summary

Model has numerous features but is very predictive and highly stat significant

- Identify using AIC
- Most p-values exceed Bonferroni standard





Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0045683	0.000844	5.41	<.0001*
Trading Rule 02	-0.002464	0.000532	-4.63	<.0001*
Trading Rule 06	-0.001558	0.000479	-3.25	0.0020*
Trading Rule 07	-0.002377	0.000505	-4.70	<.0001*
Trading Rule 10	0.0015033	0.0004	3.76	0.0005*
Trading Rule 11	0.0019829	0.000435	4.55	<.0001*
(Trading Rule 01+0.16029)*(Trading Rule 02-0.03684)	0.002377	0.000454	5.24	<.0001*
(Trading Rule 03+0.10456)*(Trading Rule 03+0.10456)	-0.001092	0.000353	-3.10	0.0032*
(Trading Rule 01+0.16029)*(Trading Rule 04-0.05089)	0.0023595	0.000426	5.54	<.0001*
(Trading Rule 01+0.16029)*(Trading Rule 05+0.10883)	-0.002736	0.000493	-5.54	<.0001*

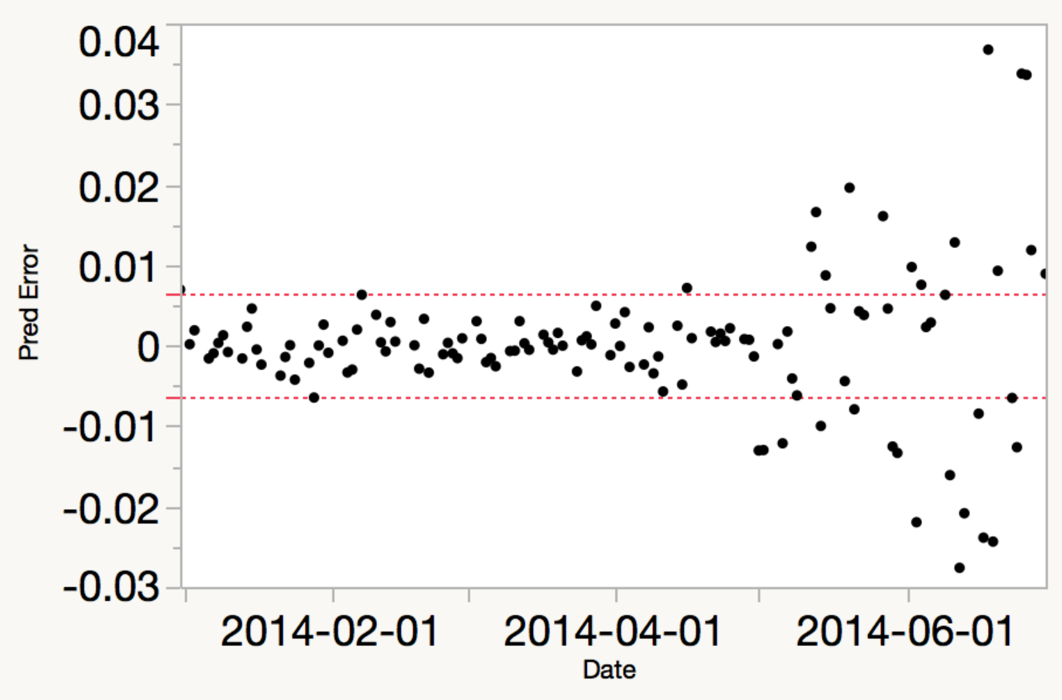
Predicts Future?

Compare claimed to actual performance

R^2 = 89% with RMSE = 0.0032

How well does it predict future?

SD of prediction errors larger than claimed



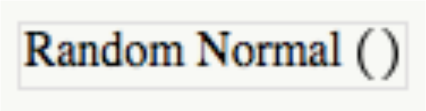
How were we so deceived?

Wharton
Department of Statistics

What went wrong?

Overfitting, multiplicity

- “Statistics rewards persistence”
- Trading rules in the model are random noise
- X_j = random normal values



Model selection process flawed

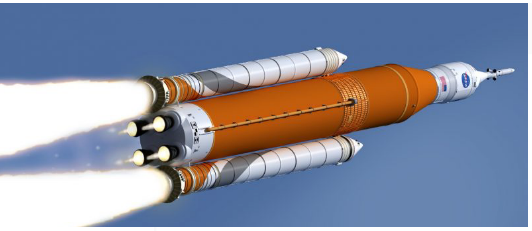
- More features than cases, so can't estimate σ^2
- Resulting bias from selection procedure ruins usual estimates of standard error.

Lesson

- To appreciate significance, must validate the procedure used to choose the model

Corollary

Model selection and multiplicity arise without fitting regression models…

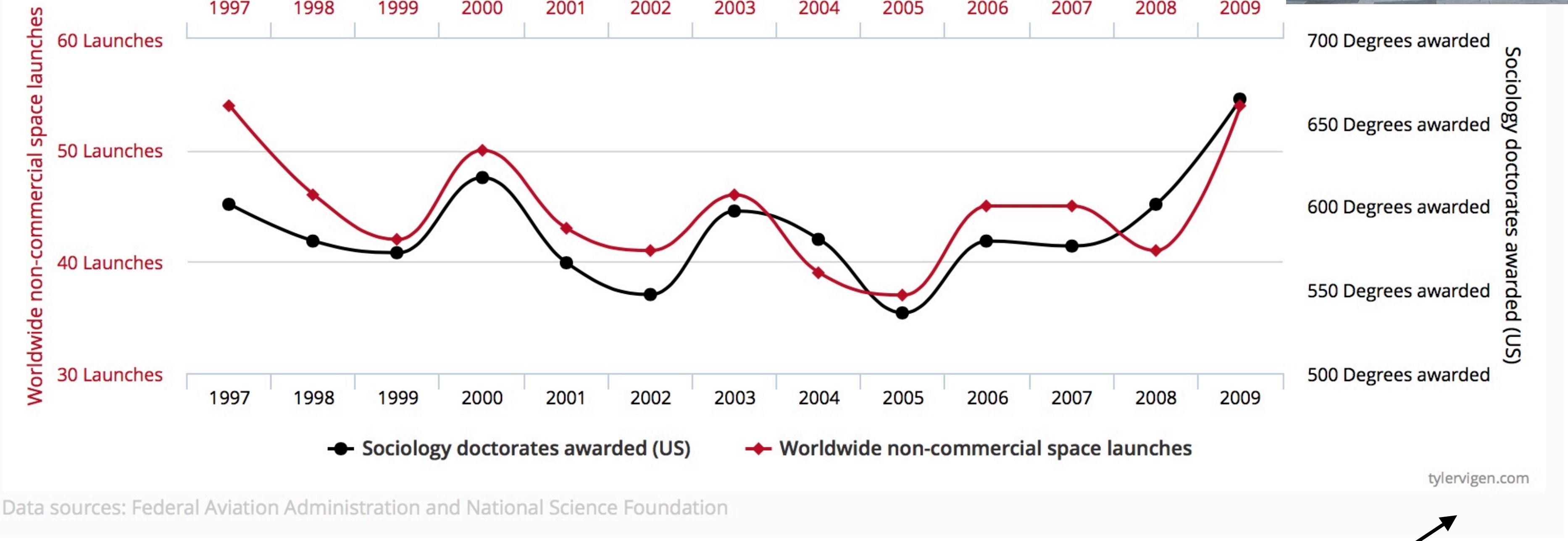


Worldwide non-commercial space launches

correlates with

Sociology doctorates awarded (US)

Correlation: 78.92% (r=0.78915)



Data sources: Federal Aviation Administration and National Science Foundation

tylervigen.com

Corollary

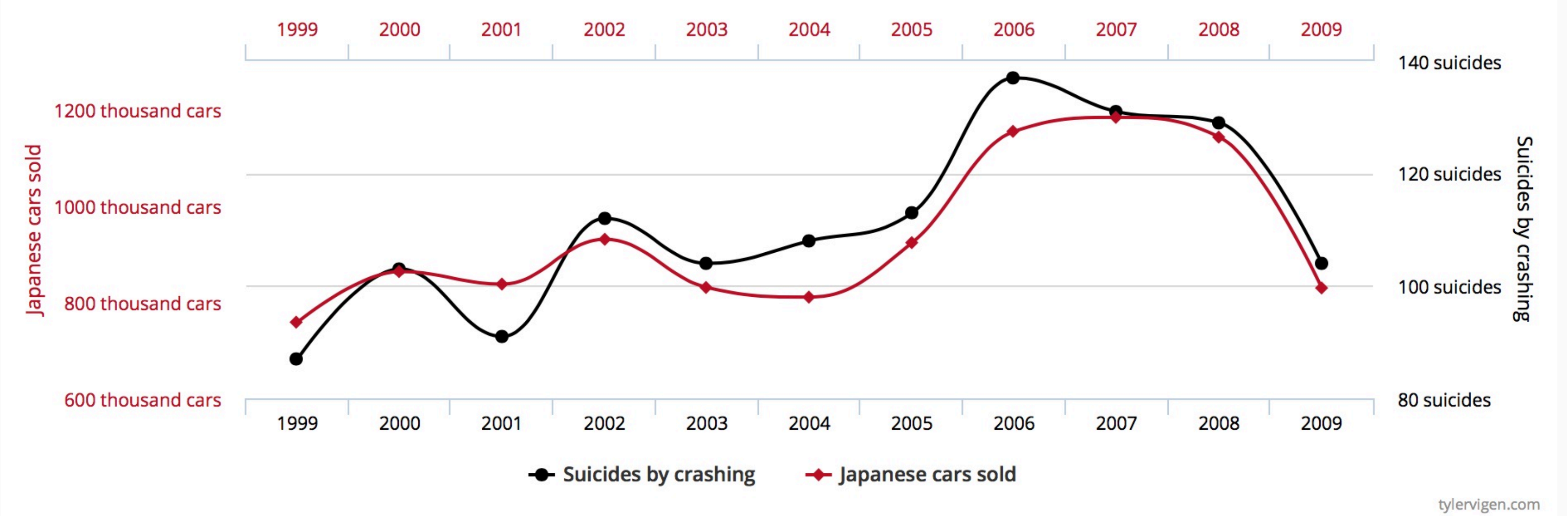
Model selection and multiplicity arise without fitting regression models…

Japanese passenger cars sold in the US

correlates with

Suicides by crashing of motor vehicle

Correlation: 93.57% (r=0.935701)



Data sources: U.S. Bureau of Transportation Statistics and Centers for Disease Control & Prevention

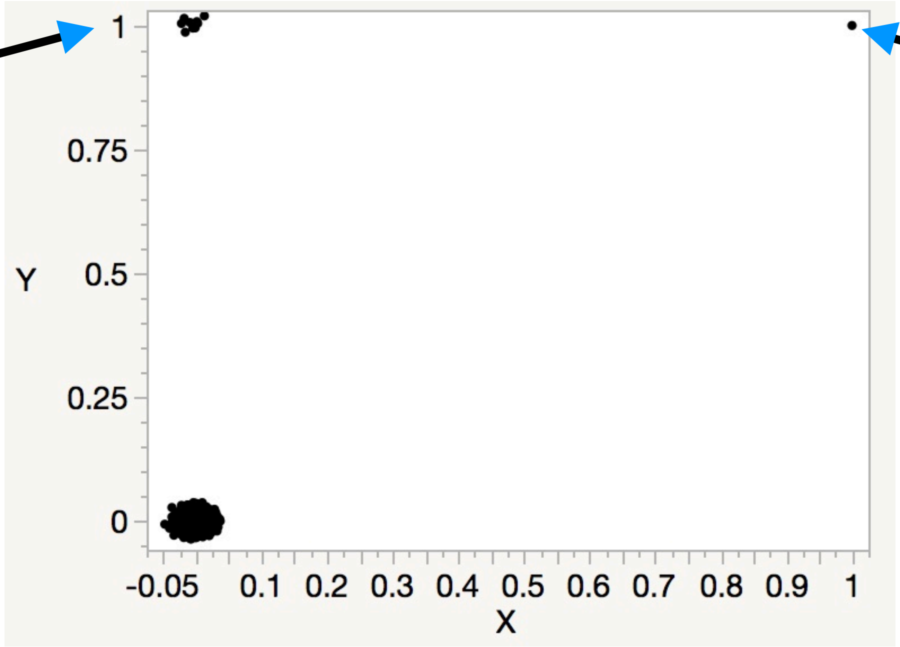
Example #3

Question

Is this sparse feature an important risk factor?

Context

Sparse variables, rare events common in big data



10 cases

1 case
x=1

n=10,000
x ≈ 0

Wharton
Department of Statistics

Statistical Significance?

Question

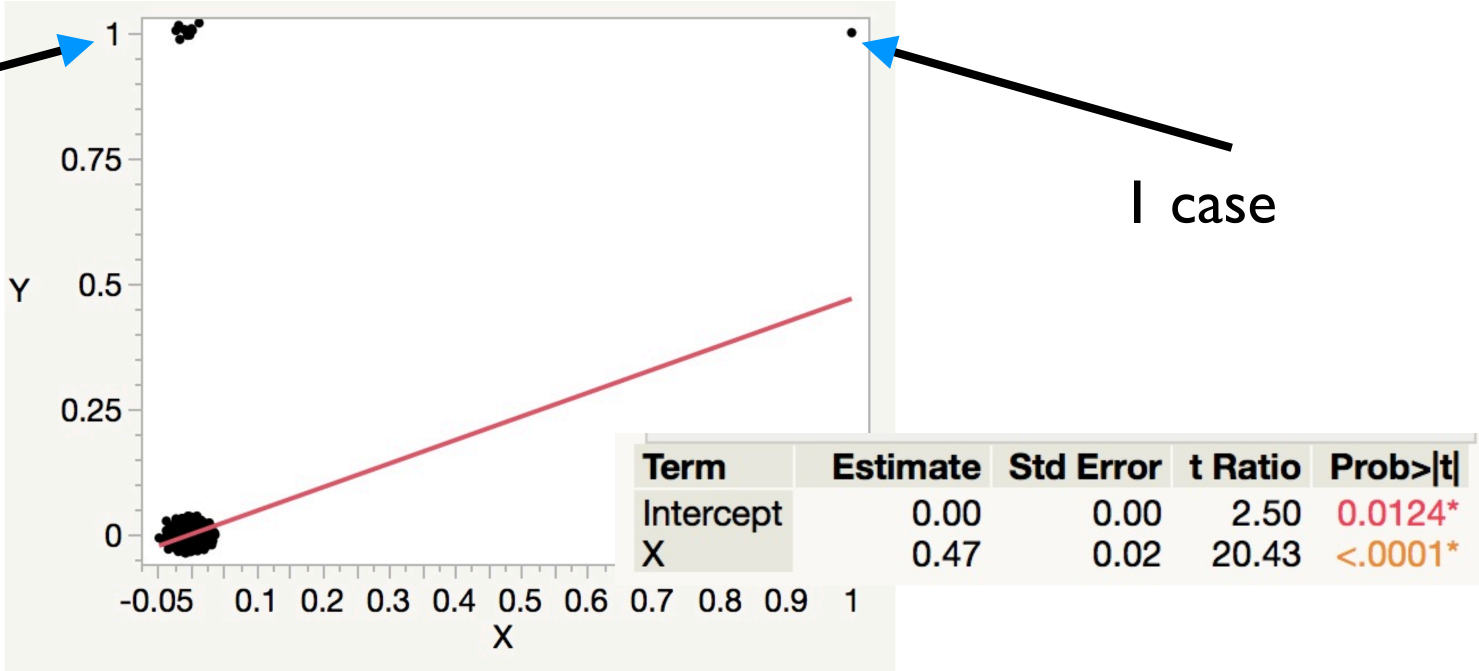
Is this variable an important risk factor?

Statistics

What's a common sense p-value for this feature?

10 cases

1 case



Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.00	0.00	2.50	0.0124*
X	0.47	0.02	20.43	<.0001*

n=10,000
$x \approx 0$

Concern #3

Explanation

- Assumptions of simple regression are not met
- Not even close to normal distribution with equal variance

Lesson

- Large n ≠ normality of estimates
- Plots remain relevant

  You might have numerous cases and many variables but plots remain important to judge results

Other Neglected Topics

- Data isn’t free
  - So, you want to run an A/B experiment?
  - Can you access all of that data quickly?
- Missing values are everywhere
  - Except in introductory stat textbooks!
- Heterogeneity of big data
  - By time homogeneous, often quite small!
- Most business data is transactional, not sampled
  - Relational data is so different.
  - Combining SQL tables

Summary

Let's not screw up this wave of interest in statistics (aka, data science)

Key learning objectives

- Students recognize dependence and distinguish number of relevant independent observations from count of the rows in a data table.
- Students realize importance of process: significance can be abused by searching over many "theories"
- Students appreciate the role of assumptions and recognize value of plots