Knowledge for Analytics & Big Data

What’s the role for statistical significance?

Bob Stine
Department of Statistics
The Wharton School
University of Pennsylvania
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 $2001$

$t = 16.86$
Statistical Significance?

Data analysis

Regress of stock return this year on stock return last year

Significantly negative!

2003 on 2002

\[ t = -17.9 \]
Statistical Significance?

Data analysis

Regression of stock return this year on stock return last year

Significantly negative

2004 on 2003

t = -4.48
Statistical Significance?

Data analysis

Regression of stock return this year on stock return last year

Significantly positive

$t = 7.26$
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?

<table>
<thead>
<tr>
<th>Year</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>2.26</td>
</tr>
<tr>
<td>2002</td>
<td>16.86</td>
</tr>
<tr>
<td>2003</td>
<td>-17.90</td>
</tr>
<tr>
<td>2004</td>
<td>-4.48</td>
</tr>
<tr>
<td>2005</td>
<td>7.26</td>
</tr>
<tr>
<td>2006</td>
<td>2.37</td>
</tr>
<tr>
<td>2007</td>
<td>6.38</td>
</tr>
<tr>
<td>2008</td>
<td>7.96</td>
</tr>
<tr>
<td>2009</td>
<td>-22.00</td>
</tr>
<tr>
<td>2010</td>
<td>2.67</td>
</tr>
<tr>
<td>2011</td>
<td>3.50</td>
</tr>
<tr>
<td>2012</td>
<td>1.48</td>
</tr>
<tr>
<td>2013</td>
<td>0.00</td>
</tr>
<tr>
<td>2014</td>
<td>-0.28</td>
</tr>
<tr>
<td>2015</td>
<td>7.65</td>
</tr>
<tr>
<td>2016</td>
<td>-5.25</td>
</tr>
</tbody>
</table>
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
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
Model Summary

Model has numerous features but is very predictive and highly statistically significant.

Identify using AIC

Most p-values exceed Bonferroni standard
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?
What went wrong?

Overfitting, multiplicity

“Statistics rewards persistence”
Trading rules in the model are random noise

\[ X_j = \text{random normal values} \]

Model selection process flawed

More features than cases, so can't estimate \( \sigma^2 \)
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
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

$n=10,000$
$x \approx 0$

10 cases

1 case
$x=1$
Statistical Significance?

Question
Is this variable an important risk factor?

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

n=10,000
x ≈ 0

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* |
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