

Predictive Analytics

STUART KLUGMAN

Senior Staff Fellow

June 7, 2018



Predictive Analytics

- What is it?
- Why does the SOA care about it?
- What is the SOA doing about it?
 - Pre-qualification education
 - Continuing education
 - Certificate program
- What do you want to know?

What is Predictive Analytics?

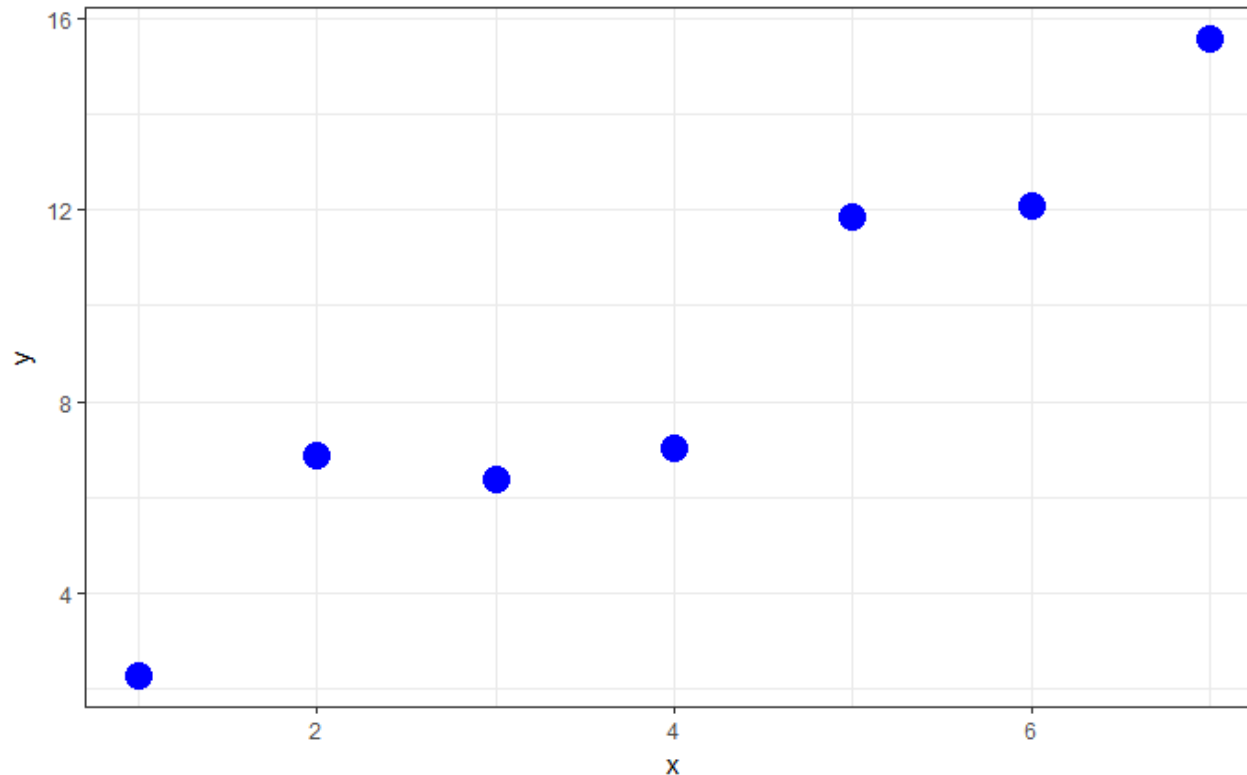


What is Predictive Analytics

- Also known as
 - Data Science
 - Big Data
 - Statistics, but with better marketing
- How does it differ?
 - Actuaries have had data (and put “science” in the name of the discipline)
 - Actuaries have had big data (e.g., mortality studies)
 - Didn't I learn statistics 30 years ago?

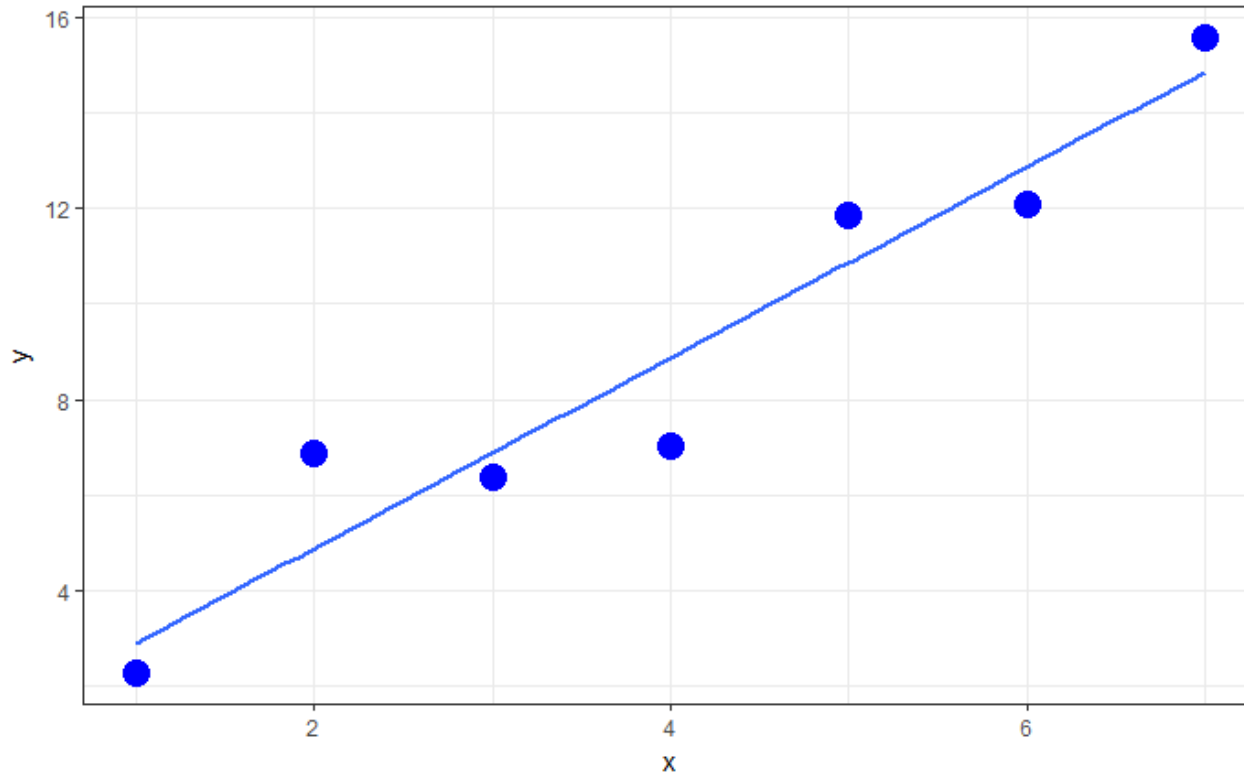
What is new - generic

- A small data example
- What is the pattern in these seven data points?



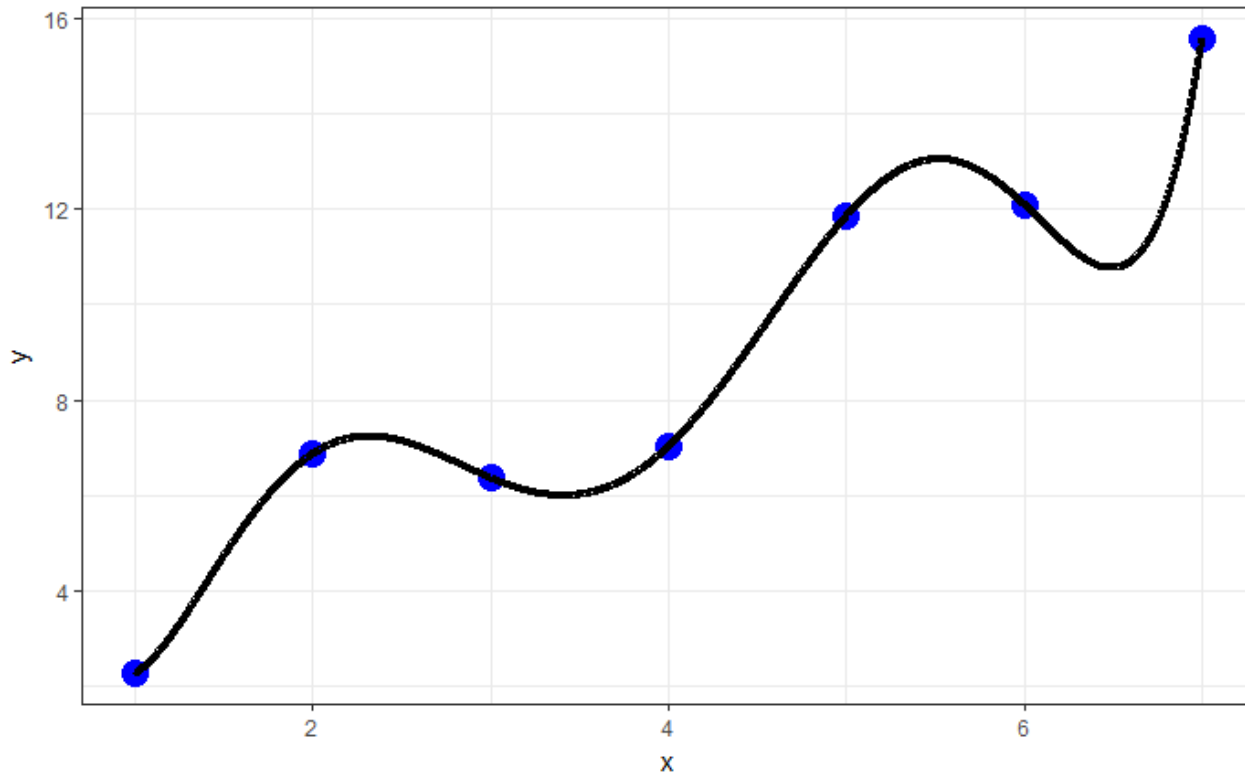
Is it a straight line?

- Low variance (a new dataset will yield a similar line)
- High bias (reality is not as simple as this)



Is it a sixth degree polynomial?

- High variance (new data will give a different curve)
- Low bias (complexity may match reality)



Big data

- Suppose we have 1,000,000 observations on 150 variables
- There will be patterns everywhere.
- Even if all 150 variables were insignificant, about 7 would be declared significant with a t-test at 5%.
- Standard statistical methods from the small data era will tend to have high variance (a new data set will select a different 7 variables) and low bias (have added complexity with little predictive power).

Solutions

- Validation/cross validation
 - Train your model on one set of data
 - Test your model on a new set of data
 - Only models that test well are kept
- Regularization
 - An alternative way to select regression variables by extracting a penalty for many variables or large coefficients.
 - Can be combined with validation methods to tune the regularization parameter.

Models

- Can use more complex models (because we can manage the bias/variance tradeoff)
- Can combine models (bagging, blending, stacking, etc.) due to easy computing access
- Speaking of computing
 - Anything new comes along, a free computing package becomes available

Why do We Care?



We have the data

- Application data
- Tracking data
- Social media data
- Other financial data

We have questions

- Who is likely to lapse?
- Who is likely to have a particular type of claim?
- What factors drive (you name it)

We have software

- Free
- Open source
- Specialized
 - actuar
 - ChainLadder
 - cplm

Our students are learning it

- Becoming a standard part of programs in
 - Business
 - Mathematics/Statistics
 - Actuarial science
- A bit of history – over my teaching career, the most common minor/second major changed from
 - Accounting to
 - Finance to
 - Data Science

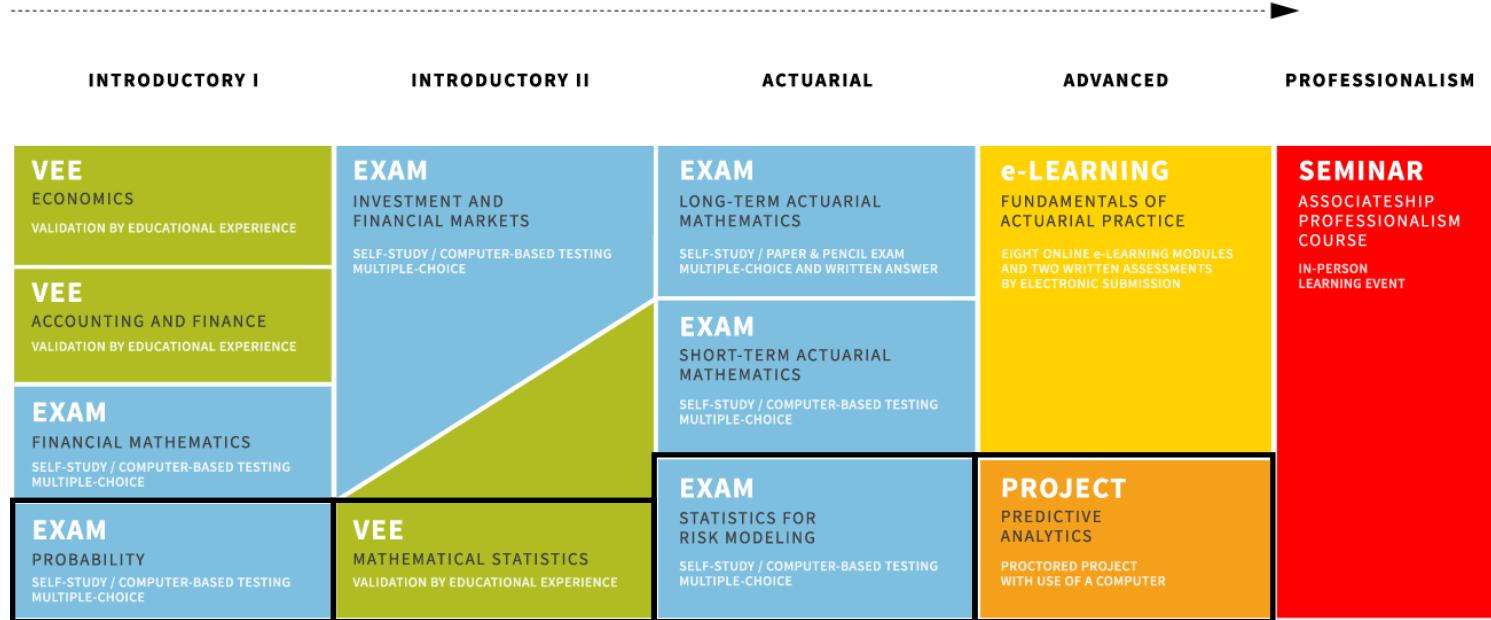
Challenges

- How to educate candidates and members
 - Read a book?
 - Have directed practice?
 - Attend a seminar?
- Assessing candidates and members
 - More than recalling facts
 - Need to verify they can conduct an analysis from start to finish

Society of Actuaries Activities



Candidates – New ASA pathway



YOU MUST COMPLETE 2 EXAMS BEFORE APPLYING TO VALIDATE YOUR EDUCATIONAL EXPERIENCE (EDUCATIONAL EXPERIENCES MAY HAVE OCCURRED PRIOR TO COMPLETING EXAMS).

Analytics exams

- P – Probability
- VEE Mathematical Statistics
 - Fundamentals of estimation and hypothesis testing
- SRM – Statistics for Risk Modeling
 - Multiple choice via Computer-Based Testing
 - Generalized Linear Models, Time Series, Principal Components, Decision Trees, Clustering
 - Cross-validation and regularization

Analytics exams

- PA – Predictive Analytics
 - Instruction via e-Learning modules
 - Data preparation, understanding, display
 - Communication
- Proctored Project
 - Five hours
 - R/RStudio, Excel, Word
 - Dataset plus business problem
 - Held at Prometric test center
 - Graded using fellowship exam protocols

Analytics Texts

- SRM Exam
 - *Regression Modeling with Actuarial and Financial Applications* by E. W. Frees (Linear models and Time series)
 - *An Introduction to Statistical Learning with Applications in R* by G. James, et. al. (All models and methods except time series, free download)
- PA Exam
 - Above plus *R for Everyone* by J. P. Lander and *Data Visualization* by K. Healy (free online version, for now)

Members – Certificate Program

- Pilot in 2017
- Fully operational in 2018
 - Open to any credentialed actuary
 - Two cohorts of 50, first one underway
 - Extensive e-Learning program
 - Two-day seminar + project-based assessment
 - Discussion forum
 - Webinar

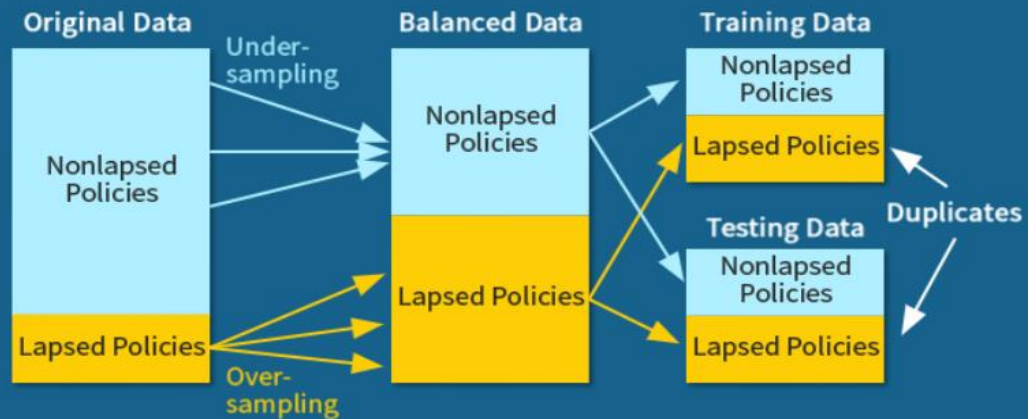
Sample content

Unbalanced Data

Oversampling and Validation

Before we move onto tuning our parameters, we need to make an important point about oversampling. You may have noticed that we only performed oversampling on the training data, that is, we didn't duplicate the lapse cases in the full data before we split into test/training. This is crucial, especially as we move onto cross validation. Oversampling before we split into training and test sets increases the chances of duplicate records in our training and test data. This is, in effect, cheating, by training the model on the same data that is in the testing dataset.

Whenever we sample data with replacement (as we do in oversampling), we need to make sure that we perform it after the data have been split into training and holdout samples (i.e., testing or validation). This prevents identical observations from leaking into the holdout sample and artificially reducing our holdout prediction error.



Members - Symposium

- First held in 2017, 200+ attendees
- Set for September 20-21, 2018 in Minneapolis
 - Registration opens soon
 - Tracks for beginners, intermediate/advanced, and managers

Members – Other resources

- Seminars
 - Advanced Business Analytics
 - Health Analytics
 - Short courses attached to other meetings
- Sessions at major meetings
- Webinars
- Predictive Analytics and Futurism Section newsletter



**SOCIETY OF
ACTUARIES®**