

Why You Should smoothScatter() from The Dark Arts of Selecting and Diagnosing Models

Dale W.R. Rosenthal

11 Jun 2019

Q₃₆

Introduction

Me:

- Head partner, Q₃₆: quant finance firm.
- Stats PhD; ex-UofC; then UIC; lately UIUC; Organizer of R/Finance.

What I want to talk about:

- Why you need to look at your data.
- How to look at lots of data; no more excuses.

Q₃₆

Use The Force **YOUR FREAKING EYES**, Luke

- Modeling is a very difficult task: very likely NP-hard.
- There is no “cookbook” for building a good model.
- Thus we need to listen to the data, watch for errors.
- We should never ignore or torture the data...
- ... except that is exactly what many people do.
- The best way to do this is to use your eyes.
- This is also one of the *least-used* tools.

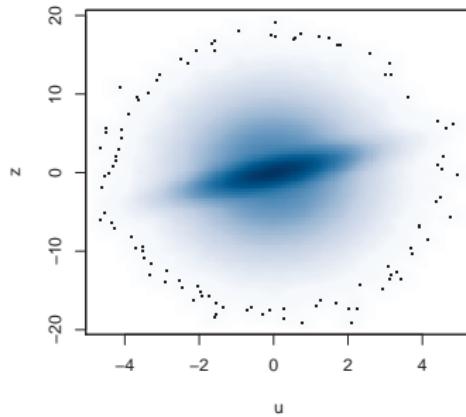
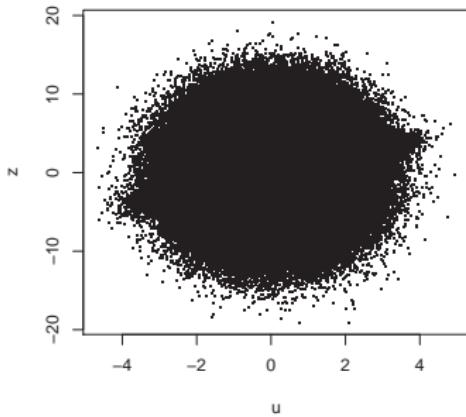
Q₃₆

Looking More Effectively

- Most people look for 10–30 seconds; cf. crossword puzzles.
 - How long to spend? 15 minutes to a few hours.
- Thankfully, *R* has some great tools for visualization.
- **⚠ Vorsicht!** Step #1 is *always* `summary()`, `hist()`, plots.
 - Look for data errors: Sensible range, median, mean, mode?
- If data OK: Time series plots; compare across time.
- ALSO: Create scatterplot matrix. Usually: `pairs()`

Looking At Big Data

- Looking at the data can be difficult sometimes.
- What if your data is big; not just studying a small slice?
 - Scatter contour plot: Scatter+data density contours, or
 - Scatter heat map: heat map+rarer points (`smoothScatter()`).
- Compare left and right plots; which is more useful?



Q₃₆

Going Further

- For more information on looking at data:
 - Weisberg: Applied Linear Regression
 - Cox and Snell: Applied Statistics
- Intrepidly quantitative? Interested in finance?
 - My book: *A Quantitative Primer on Investments with R*



Q
36