Actuarial Analysis Using Open Source Software and Open Data

Brian A. Fannin, ACAS

November 29, 2013

Contents

Preface

There was a time not too long ago, when computers were expensive and heavy; they stayed on a desk and didn't like to travel. Business trips involved careful consideration of what data and analysis needed to be rendered onto paper and stuffed into a briefcase. The fortunate ones could borrow a laptop from the company pool. This was a great help once you had copied all of the data, analysis, exhibits and software you needed. Inside your company, information was available from one of two sources. There was a LAN which was in frequent danger of being overrun with information. The LAN was organized according to the whim of the hundreds or thousands of users who created, changed or deleted files there at will. Your second source of information was transactional data from business processing systerms. This was tightly secured in vague databases whose format and operating systems sounded like something from the cold war. Database administrators ensured that production reports were readily available for middle managers. Actuaries were often the last in line for information. When they were given information at all, it was received at a high level of aggregation at scheduled intervals. Ad hoc data requests needed strong justification to justify the consumption of some other department's scarce capacity. Information from outside your company was made available to you by copying it to a CD and physically delivering it to your office. This data had traversed the same IT gauntlet that safeguards information in your own company. You were given only what you had asked for and you had been cautioned to ask for only that which you could expect to receive. There were scant opportunities to make subsequent requests for more detail or to correct inaccuracies.

All of this data, the raw material that feeds the analytic process, was received by eager actuaries who would dutifully process the information in a spreadsheet. This processing would use whatever formulae happened to ship with the software. This meant that one could calculate means and standard deviations, perform a univariate linear regression and visualize the results in a scatter plot or a bar graph. Combining data from more than one source meant a complicated sequence of lookup functions, whose integrity could be destroyed by a column which had been moved. The data in the spreadsheet was copied in once, twice, as many times as were necessary to keep it consistent with data reported by anyone else. Functions for limited expected value, algorithms to estimate the parameters of a loss distribution, methods to determine capital consumptions and return, must all be coded by hand. Other departments and other compa-

nies all had their own favorite and propietary tools which did much the same thing (although just a little different). Software costs are not trivial. A large company must meet the demands of human resources, your legal department, underwriting, marketing, claims, risk management and IT. This massive cost meant that your desktop office software was typically three to five years out of date. From time to time, you would read about an innovative new statistical technique and wonder whether you would ever have time, data or technology to study it more. You would then return to the rote task of trying to reconcile the latest inconsistency in the data supporting the loss reserve study.

Things have changed.

As this book is being written, data analysts are living in a golden age. Computing hardware is cheap and plentiful and what you can't compute on your laptop, you can compute in the cloud. People much younger than your author have never seen a floppy disk, much less understand what one is and find CDs a quaint notion for data storage. Flash drives with enough memory to hold your company's history are given away for free at professional events, proudly embossed with a sponsor's name. More than a century of weather, employment, demographics, stock trades, disease, natural disaster, election results, box office receipts and sports results are available on the internet. Not more than fifteen years ago, if you wanted to know the name of the actor who played Tackleberry in the Police Academy movies and how many of those movies were made, you'd need to rely on your memory. Now those answers are a few mouse clicks away[1].

And then there's the software. It would be difficult to find anyone who has paid money for an internet browser in the past ten years. Similarly, spreadsheets and other office software need not cost money. Google supports a web-based suite of office products. Want something on your machine? Sun supports the Open Office suite of products which implement a set of open standards used to store office documents. It's the same standard used by Microsoft in their Office suite.

Predictive modeling is not only ubiquitous and available, it's sexy.Today, you can tell people at a cocktail party what you do for a living and regale them with stories about finding hidden meaning in massive piles of data.

Or can you? The aim of this book is to ensure that you can. The foremost objective is to ensure that anyone who reads this book will become a better actuary. By that, I mean an actuary who spends less time reconciling and transporting data and instead is an actuary who can effectively support sound business decisions. To do this, you will learn about

This is not a theoretical book. I'm not capable of writing it and I don't need to. There are piles of books and articles about the theory of data storage, software construction, statistical analysis and actuarial practice. I love to read them and hope that some of that insight finds its way into this book. But there will be no new ground broken here. Instead, what I hope to achieve is something written by and for the practicing actuary. These are not abstract problems and impractical solutions. This is the result of twenty years in the industry trying

[1]David Graf and seven, with one planned for 2014. There was also a TV show.

to make sense out of scarce data, high expectations and a limited toolkit. The open source revolution has changed the way that I approach solving problems. I spend less time worrying about the boundaries of my software and more about understanding the statistical processes which take place in the real world.

This is an exciting time to be an actuary. We have the ability to be masters of our own information, we have the opportunity to employ leading edge technologies to the questions we ask and the tools to render the science intelligible to our stakeholders. At last, we have the power to know and the power to explain.

Part I

Preliminaries

Chapter 1

What is open source?

There is a difference between

Chapter 2

The data and analysis process

This will form the structure for the rest of the book. I look at the process[1] as follows:

- Data structure
- Data maintenence and audit
- Reporting
- Visualization
- Analysis
- Publication

The first two elements relate primarily to data. For this, I use the metaphor of the construction of a building. The first step is to develop a rational blueprint. This is a plan for how the building will be constructed, with consideration for its purpose, its occupants, physical and legal limitations and the like.

We'll look at each of these individually.

2.1 Data structure

This is the most critical element for any enterprise and one over which actuaries have the least influence. This is because data structure in an insurance company is tightly linked with transactional data.

[1] It's not a cycle, by the way. Analytics is meant to have an end, which supports a decision or behavior change. You can repeat the analysis as data changes, or analyze something new, but it's not a cycle.

2.2 Data maintenence and audit

I will treat this category broadly and include the subject of Extract, Transform and Load: ETL. Data maintenence is the practice of housing

Chapter 3

The software you will use

You do not need to install all of this software all at once. It will be useful to refer back to this section as needed. I will describe the most significant components. Other ancillary tools will follow at the end. Consistent with the focus on data first, we will talk about data storage and retrieval technologies before delving into analytics tools.

3.1 PostgresSQL

PostgreSQL is among the preeminent open source relational databases. Its history stretches back to the 1990s and blah, blah, blah.

3.2 MongoDB

3.3 Knime

Knime sits at the boundary between data maintenence and analytics. It is useful as a structured ETL process, but also has some highly sophisticated exploratory analysis and more advanced data mining tools. It may be downloaded from SOME WEBSITE. Note that the tool itself doesn't install

3.4 R

The explosion of popularity of R is blah, blah, blah. Here are the libraries that you will need:

3.5 Python

Python has been around since the early '90s, but only recently has it gotten much attention from the analytic community. The publication of the Python libraries numpy and pandas have had a great deal to do with that.

3.6 TeX

TeX is a markup language which will aid in writing professional summaries of your analytic work. This book was written in TeX.

3.7 Others

RapidMiner, Weka, GitHub

Chapter 4

The SampleData

To facilitate the learning process and in keeping with this book's spirit as a practical guide, we will work with an actual set of data. This is not a trivial amount of information. It's meant to reflect the volume and scope with which an actuary could expect to work. It will be constructed automatically and randomly and you may reproduce it on your computer. To do so, you will need to have installled R and all of the required packages. Information on doing so may be found in INSERT REFERENCE HERE.

Most demographic information may be obtained from the US Census. These are huge datasets and a bit of patience is required to download them.

For each county, the number of insureds is distributed as a Poisson where N is equal to lambda times exposure. Exposures is equal to the county's population.

Part II

Data

Chapter 5

Data Structure

We'll talk about tabular, document and graph databases.

Part III

Analysis

Part IV

Publication