Article from:

Health Watch

January 2014 – Issue 74

Advanced Business Analytics: Applied Statistics, Programming and Actuaries

By Sheamus Parkes

Sheamus Parkes, FSA, MAAA, is a health actuary at Milliman, Inc. in Indianapolis, Ind. He can be reached at *shea.parkes@milliman.com.*

I'm enjoying the Society of Actuaries (SOA) recent emphasis on advanced business analytics. The SOA's classic education and examination programs produce professionals with robust subject matter expertise; however, once actuaries are credentialed, the education opportunities available to them are still focused on the same topics as the exams. It can be hard to learn what you don't realize you need to learn as there are a lot of unknown unknowns in a recent fellow's knowledge space. Almost any pertinent topic has a rich discussion going on in both academia and business if you are just willing and able to find and join it.

Many of those knowledge gaps exist in advanced business analytics. The SOA's Advanced Business Analytics initiative closely overlaps with the larger data science movement. Actuaries come into the data science realm with a healthy amount of domain knowledge, but they are weak on statistics and programming skills. By focusing on these extra skills, actuaries have a large opportunity to break out of their regulation-enabled roles and compete in the broader business world on pure analytic strengths.

The three sections below shine light on some of those opportunities. The first is an interesting case study that shows how closely related classic actuarial knowledge can be to applied statistics. The last two sections illuminate what there is to learn about applied statistics and programming in general.

Case study: ACOs

Accountable care organizations (ACOs) have a fresh incentive to find savings in the health care system. Some clients have found the required willpower to take on provider profiling, which is a political and statistical minefield. Success takes a mix of tact and sophistication. This case study will focus mostly on the statistical point of view.

The outcome studied is often a rate (utilization/cost per service/episode/member), but the sample size of exposure by provider is often akin to a Pareto distribution. Actuaries are trained to understand that the majority of providers will not have credible observed rates, but their toolset for addressing this issue is often inadequate. They throw out low-sample providers or use rough partial credibility blending. A more useful approach involves mixed modeling (also known as hierarchical modeling) (Gelman and Hill 2007). Properly applied mixed modeling is statistically identical to least-squares credibility, and can be easily extended.

This case study investigated the cost of one of three planned procedures at various facilities. Much energy was invested in using a medical episode grouper and pruning/combining the results into some semblance of homogeneity. Business rules had to be defined to exclude episodes that could possibly have been emergency procedures or those done in an outpatient setting. Expert clinicians were consulted and presented with summaries and statistical visualizations; as part of this, the analysts grew their own subject matter expertise. After that manual exercise, a series of robust statistical procedures were developed to prevent any particular observations from having unlimited influence in this or any future refresh of the analysis. Robust statistics are very useful tools to learn and apply to messy real world data (Maronna, Martin and Yohai 2006).

The primary modeling exercise then began and involved important subjective decisions such as:

- Choice of conditional gamma distribution
- Inclusion of eligibility status covariates to provide implicit case mix adjustment
- Inclusion of region covariates to provide room for region-appropriate full credibility targets

The effects of interest were the estimated costs of the three chosen procedures by facility. The modeling space was expanded such that a given facility's relative performance on different procedures could be correlated; e.g., it could be that if a hip replacement costs more at a certain facility than at the average facility, knee replacement will likely cost more there as well.

Uncertainty around these estimates was calculated, but limited to the estimated uncertainty given the model was actually true. We carefully communicate these uncertainty estimates as being most useful for comparing relative credibility between facilities. The observational nature of the data is also limit-

FIGURE 1

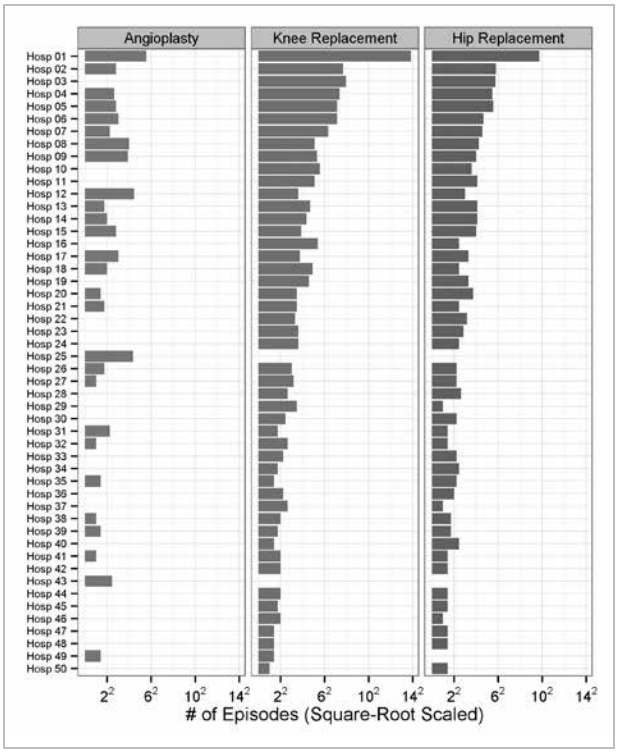
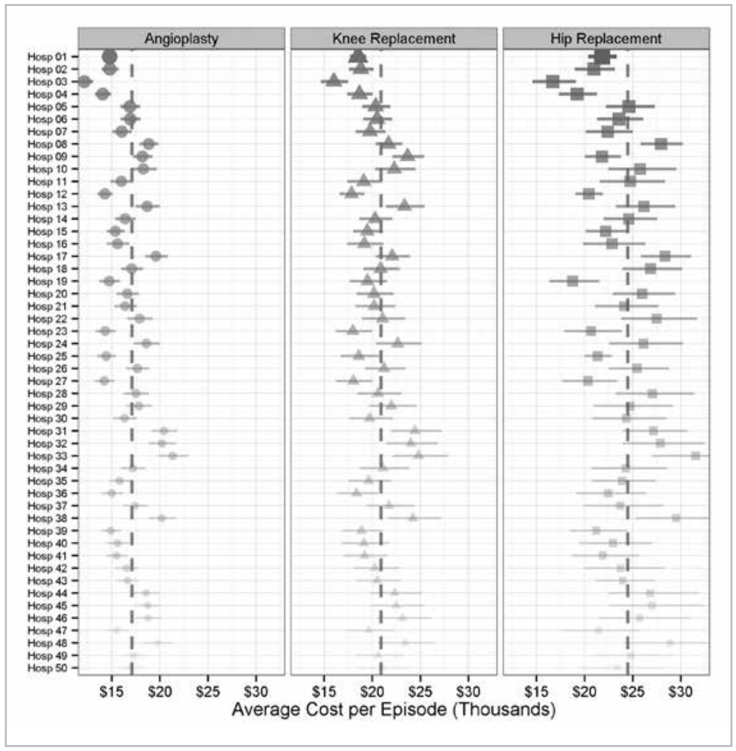


FIGURE 2



ing when planning action steps from the results; it is a leap of faith that a facility would continue to have similar costs per episode if more episodes were performed there. All of the results must be consumed with common sense and a healthy dose of skepticism.

A series of visualizations give insight into the case study. These particular visualizations were designed for the audience of this article; the complexity would be scaled back for a less analytic audience. **Figure 1** shows the frequency of episodes by facility and episode trigger (angioplasty, hip replacement or knee replacement). Actuaries would intuitively understand that the majority of the facilities would have weakly credible costs per episode. Seeing the gaps of zero episodes, especially for angioplasty, would suggest some facilities just don't provide certain services and any estimates of their costs for doing so should be disregarded.

Figure 2 shows the estimated average costs per episode for each facility and episode trigger. The facilities are still sorted as in Figure 1 (descending number of episodes). The points represent the best estimates of the model and the horizontal bars are one standard error wide; the vertical dashed lines represent the overall average cost. Size and transparency of the shapes were used to emphasize the more credible results. The generally tighter bars on the angioplasty results suggest facilities are more consistent in their costs for angioplasty. However, that tightness can somewhat be an artifact of the multiplicative nature of the model and the lower overall cost of angioplasty.

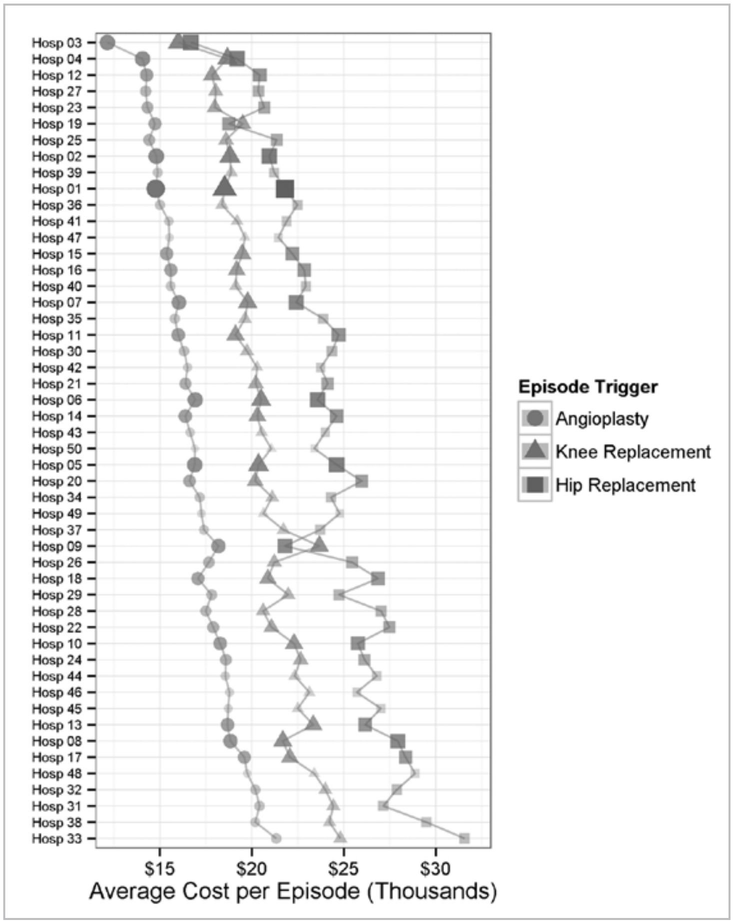
The effect of penalization/actuarial credibility/shrinkage can be seen in that the facilities with low support are usually constrained to be close to the average. The shrinkage would be more obvious if the unadjusted cost estimates were presented for comparison. It is possible to observe some of the correlation between angioplasty, knee replacement and hip replacement in Figure 2, but Figure 3 makes it more obvious.

Figure 3 on page 10 presents the same information, but the facilities are now sorted by their average estimated costs across the three episode triggers. This visual goes closest to the heart of the business need: Which facilities would best minimize costs? This ordering is much more useful than a raw observed cost ordering; the estimated cheapest facilities are all of credible size, and yet not all credible facilities are cheap. The correlation assumption appears front and center; the episode trigger lines are almost painfully parallel. That strong correlation assumption enabled borrowing strength between episode triggers to reach stronger conclusions about facilities with moderate support.

More simplistic approaches could still have been useful. It would be reasonable to just limit reporting

CONTINUED ON **PAGE 10**

FIGURE 3



to facilities with a minimum number of episodes per trigger and sort by their observed averages. The generalized linear mixed model approach can provide some competitive advantage by keeping those facilities with moderate to weak support in consideration. A facility that showed consistently excellent performance across a small sample of different procedures could be compelling evidence when the modeled correlation pools strength between the procedures. It also alleviates the need of choosing the full credibility threshold of a simpler method.

The next two sections highlight the subject areas that would support an analysis such as one shown in this case study.

Applied statistics

"Statistics: a subject which most statisticians find difficult but in which nearly all physicians are expert," wrote Stephen Senn (2008). This could also apply to many young actuaries who believe the exams did a thorough job covering applied statistics. They are taught the basics of frequentist statistics, but seldom with a true understanding. They won't realize how many pitfalls are looming when trying to bend a statistics paradigm designed for random control trials to a common business problem domain. They will brush aside issues of observational data. They will present significant p-values when they have enough data for any effect measured in any form to provide a p-value less than 0.001. They might present some uncertainty in their parameter estimates, but they are unlikely to think of the uncertainty in their modeling or data choices.

Bayesian statistics provide a bit more palatable rationale in the business world, but they are not a silver bullet. There are no silver bullets. Machine learning is a great lens to view predictive modeling through, but actuaries are often focused on inferences and quantifying uncertainty as well as accuracy (Hastie, Tibshirani and Friedman 2009). In addition to learning about the modern fusion of statistics and computers, actuaries should be reading the classic works of enlightened practitioners such as George Box (Box, Hunter and Hunter 2005) and John Tukey (1977). It takes broad knowledge and experience to produce, communicate and defend useful results.

Programming

Business analytics can't be considered advanced until they are reproducible and reusable. Point-and-click interfaces are wholly inadequate. Spreadsheets are a land mine of horrible practices; even well thought out and strictly enforced formatting guidelines might enable an almost sane separation of data and analysis. A similarly colossal effort will keep analysis flowing along a single path, at least for a while.

Real analysts write code. Programming provides a clean separation of data and analysis. It takes only an achievable level of effort to ensure smooth flow of logic through a code base. Abstracting repetitive tasks into reusable routines is a cinch, and any useful language likely already has an appropriate routine if you just look for it.

Writing good code can be difficult, however; skill differences among any sample of programmers commonly vary by many orders of magnitude (McConnell 2009). It is possible for an individual to improve, but it takes effort and practice. In addition to knowledge of specific languages, an analyst needs to have strong programming fundamentals. Intelligent programming is all about managing complexity; duplication is evil and modularity is bliss (Hunt and Thomas 1999).

There are many useful programming languages available, and different business problems will fit into different languages more easily. Learning additional languages will let your mind expand to see problems from different angles. Time and willpower should be the only limitations on learning; many of the best languages are available as open source (free to use even in a commercial setting). My personal favorite is R ("a language and environment for statistical computing and graphics") (R Core Team 2013). R is a domain-specific language for applied statistics and those are the problems I am most often solving. Python is a general purpose language with many packages to extend its applicability to statistics (or any other problem space). Commercial software such as SAS and SPSS are excellent choices, especially if they are already in use in an organization (Littell, et al. 2006). Every analyst should be comfortable in some variants of SQL; countless commercial and open source options are available.

Additional References

This article was written for a technical audience, but the communicated skepticism can be retained even when presenting to a more business-oriented audience. As George Box said, "All models are wrong, but some are useful" (Box and Draper 1987). To learn about the particular methods used in the case study (and advanced business analytics in general), I recommend the books *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Gelman and Hill 2007) and *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis* (Harrell 2001). This specific analysis and visualization was completed in the R programming language (R Core Team 2013) utilizing the lme4 (Bates, et al. 2013) and ggplot2 (Wickham 2009) packages respectively. For information about learning R, I recommend *The Art of R Programming: A Tour of Statistical Software Design (Matloff 2011)* and *R for Dummies* (Meys and de Vries 2012).

END NOTES

Bates, Douglas, Ben Bolker, Martin Maechler, and Steven Walker. *lme4: Linear Mixed-Effects Models Using Eigen and S4*. R package version 1.0-5, 2013.

Box, George E. P., and N. R. Draper. *Empirical Model-Building and Response Surfaces*. New York, NY: John Wiley & Sons, 1987.

Box, George E. P., J. Stuart Hunter, and William G. Hunter. *Statistics for Experimenters: Design, Innovation, and Discovery*. 2nd ed. New York, NY: Wiley-Interscience, 2005.

Gelman, Andrew, and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, MA: Cambridge University Press, 2007.

Harrell, Frank E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY: Springer, 2001.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer, 2009.

Hunt, Andrew, and David Thomas. *The Pragmatic Programmer: From Journeyman to Master*. Addison-Wesley Professional, 1999.

Littell, Ramon C., George A. Milliken, Walter W. Stroup, Russell D. Wolfinger, and Oliver Schabenberger. *SAS for Mixed Models*. 2nd ed. Cary, NC: SAS Institute, 2006.

Maronna, Ricardo A., Douglas R. Martin, and J. Victor Yohai. *Robust Statistics: Theory and Methods*. West Sussex, England: Wiley, 2006.

Matloff, Norman S. *The Art of R Programming: A Tour of Statistical Software Design*. San Francisco, CA: No Starch Press, 2011.

McConnell, Steve. *Code Complete: A Practical Handbook of Software Construction*. 2nd ed. Sebastopol, CA: O'Reilly Media Inc., 2009.

Meys, Joris, and Andrie de Vries. *R For Dummies*. West Sussex, England: John Wiley & Sons, 2012.

R Core Team. R: *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2013.

Senn, Stephen S. *Statistical Issues in Drug Development*. West Sussex, England: John Wiley & Sons, 2008.

Tukey, John. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.

Wickham, Hadley. *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer, 2009.