Re-engaging with Regression

Keith McNulty

15/02/2021

At the end of last year I released my new *Handbook of Regression Modeling in People Analytics* as an open book online. The book is intended as a guide to statistical inference and common regression methodologies for those who work in people-related disciplines.

I wrote this book and made it available online for a number of reasons. First, as a people analytics leader and practitioner, I find that what people want from us the most is not a prediction of when something will happen, but rather an explanation of *why* things are happening, something that regression is particularly well-positioned to support. Second, I was getting worried about the impact of the explosion in predictive analytics on our knowledge of—and appetite for—techniques which explain things rather than predict things. At a time when we need inferential methods more than ever, what will we do if data scientists and quantitative analysts are too busy running neural network or random forest learning algorithms to know how to run and interpret an inferential statistical model? Finally, in the past couple of years I am seeing a growing appetite for advanced inference methods outside clinical and academic settings, with increasing transfer into public and private enterprise settings. Methodologies like survival analysis, mixed models and latent variable analysis are now no longer the preserve of the clinician or the academic statistician or psychometrician and many businesses are now looking for skill sets that can set up, run and interpret these methods.

## Popularity of machine learning can create a methodological ‘blind spot’

With the intense focus on machine learning algorithms in data science education in recent years, there is a growing tendency for practitioners to dive straight into predictive methods when faced with an analytic objective, and in many cases this is not consistent with the broader evidence-based goal of the research. A typical situation I have observed on numerous occasions is where stakeholders or leaders ask whether a certain phenomenon (for example, performance, promotion or attrition) can be *predicted*. The data scientist immediately builds a set of learning algorithms and calculates their predictive accuracy, only to find out later that the stakeholders don’t care about the predictive accuracy because they never intended to use this work to make predictions in the first place. What they actually wanted was help to *explain* the phenomenon. Suddenly , the data scientist finds that their methods were not optimal for answering this new question, because predictive methods are not built with inferential interpretation as their primary objective.

As my colleague and fellow people analytics leader Alexis Fink points out in the foreword to my book, regression is the best ‘swiss army knife’ we have for answering the question of ‘why?’ in our field. Most parametric regression techniques allow us to do three things which are critical to understanding what may drive an outcome of interest. First, we can make statistical inferences on which variables are actually significant in explaining the outcome, allowing us to confirm or rule out certain beliefs and hypotheses, dispel long-held myths and simplify the universe of possible causal elements of a problem. Second, we can quantify and articulate the strength of the effect of each significant variable on the outcome, something I have personally found is of great interest to stakeholders. Finally we can estimate the overall ‘explainability’ of the outcome. In people disciplines we often find that our outcomes of interest have limited overall explainability, but our clients and stakeholders can be so excited about specific inferences that they frequently overstate their importance. This can be managed more effectively with good quantification of overall explainability through statistics like or pseudo- measures.

## Competent regression modeling requires skill and judgment

The advent of machine learning workflows such as those available in packages like scikit-learn in Python encourages a ‘software development’ approach to analytics. While there are many reasons why this approach is useful, efficient and valuable in many contexts, it is also a dangerous development because it can discourage the exercising of important elements of judgment and interpretation that are often critical in skilled inferential modeling. In many parts of my book, I deep dive on these important elements.

Early in the book, I propose a theoretical definition of inferential modeling as a process which uses the statistical properties of a sample to *infer and describe* relationships in a larger population to a high degree of statistical certainty. As the book proceeds, I highlight what that definition means for the management and selection of data for a model, for the choice of an appropriate model, for the setup and execution of a model, for the translation of results from a model and for the diagnostics and checking of important assumptions behind a model.

All of this is done through example problems and downloadable data sets which allow the reader to engage directly with the problem themselves, with all necessary code in R and Python provided. The problems are set in a variety of disciplines on topics such as academic scoring, employee engagement and performance, sports player discipline and political voting behavior and preferences. In end-of-chapter exercises, readers are encouraged to put their coding, judgment and interpretation skills into practice with fresh data sets and learning exercises.

## Growing demand for more advanced methods outside academic and clinical settings

The content of the book reflects a range of commonly used regression methodologies such as ordinary least squares linear regression, binomial and multinomial logistic regression and proportional odds regression. The book then extends to methods that might previously have been considered highly technical and academic in nature, but which I have observed are increasing in use in enterprise settings in recent years.

Desire for more nuanced analysis of populations and the commonalities and differences within them has led to an increasing interest in *mixed models*, and the book provides an interesting and fun application of this in the context of a speed dating experiment. *Latent variable modeling* is reviewed in the context of surveys, where there is growing desire to generate simpler and more intuitive insights through reducing dimensionality in these often complex and disorganized instruments. *Survival analysis* is illustrated as a way of modeling events that occur over time, now increasingly used for the modeling of both negative events such as employee attrition and positive events such as promotion.

*Handbook of Regression Modeling in People Analytics is available online now at* [*http://peopleanalytics-regression-book.org*](http://peopleanalytics-regression-book.org) *and is soon to be released in print/ebook form by Chapman and Hall/CRC Press with all the author’s proceeds donated to the R-Ladies Global foundation to promote diversity in statistical programming. The print/ebook version will contain additional exclusive content including a chapter of additional data sets and learning exercises.*

**Caption for image**: Proportional odds regression is illustrated in Chapter 7 of the book as a method to model ordered category outcomes, such as performance ratings and survey responses. In this 3D visualization taken from the book, the different colored surfaces represent probabilities of increasing disciplinary action against players in soccer games, based on their disciplinary record in their prior 25 games. Blue represents no disciplinary action, yellow represents a yellow card (warning), red represents a red card (dismissal).