Chapter Abstracts for Handbook for Regression Modeling in People Analytics

Keith McNulty

15/02/2021

## Chapter 1

**Book Title:** Handbook of Regression Modeling in People Analytics: With Examples in R and Python  
**Chapter Author**: Keith McNulty (ORCiD: 0000-0002-2332-1654)  
**Chapter Number & Title**: Chapter 1, The Importance of Regression in People Analytics

Regression models are families of inferential statistical techniques that play an extremely important role in people analytics and in many other disciplines, because they are particularly well suited to explaining outcomes of interest such as performance, promotion or employee engagement. In this chapter, the role and importance of regression in people analytics is described through differentiating between inferential models and predictive models and describing the aims of each, outlining the situations in which it is more appropriate to use inferential models such as regression. The chapter provides a formal definition for the concept of inferential modeling and describes and discusses the different types of models available for common use cases. Some elementary concepts related to statistical inference are also introduced in this chapter, such as sampling of populations and random error.

## Chapter 2

**Book Title:** Handbook of Regression Modeling in People Analytics: With Examples in R and Python  
**Chapter Author**: Keith McNulty (ORCiD: 0000-0002-2332-1654)  
**Chapter Number & Title**: Chapter 2, The Basics of the R Programming Language

R is a powerful programming language for any form of analytics and contains an extensive functionality to support both elementary and advanced statistical modeling. Many of the walkthrough examples in this book use R as the primary analysis tool. In this chapter, fundamental aspects of R programming are introduced to provide a foundation for the methods introduced in later chapters. The chapter progressively covers the installation and use of R, data and data types, data structures, loading, cleaning and managing data in dataframes, using functions and packages, plotting and graphing methods and documenting work using R Markdown. The chapter serves as a basic programming foundation for the remainder of the book, and as an elementary introduction to R for those new to the language.

## Chapter 3

**Book Title:** Handbook of Regression Modeling in People Analytics: With Examples in R and Python  
**Chapter Author**: Keith McNulty (ORCiD: 0000-0002-2332-1654)  
**Chapter Number & Title**: Chapter 3, Statistics Foundations

Regression models depend on some foundational concepts of inferential statistics and it is challenging to design, run or interpret models without a good knowledge of these concepts. This chapter reviews some common measures in descriptive statistics such as mean, variance, standard deviation, covariance and correlation, with example data sets used as practical examples. The chapter then discusses the statistical properties of random variables, the central limit theorem and the idea of statistical confidence. The theory of hypothesis testing is introduced and then illustrated through three specific types of hypothesis tests: Welch’s t-test of difference in means, correlation testing and Chi-square tests for different distributions between categories in a data set. The chapter uses example data sets to illustrate the steps involved in hypothesis testing and how to interpret results before illustrating the convenient functions in R and Python which perform these tests.

## Chapter 4

**Book Title:** Handbook of Regression Modeling in People Analytics: With Examples in R and Python  
**Chapter Author**: Keith McNulty (ORCiD: 0000-0002-2332-1654)  
**Chapter Number & Title**: Chapter 4, Linear Regression for Continuous Outcomes

Linear Regression, or Ordinary Least Squares (OLS) Regression, is one of the most widely used and established regression methods available and is a frequent choice for modelers because of its underlying simplicity and ease of interpretation. This chapter contains a step by step walkthrough of the ordinary least squares method using an example related to student test scores. Many foundational regression concepts are introduced such as residual errors, fit, and goodness of fit. The interpretation of linear models is discussed in detail, including coefficients, confidence intervals and coefficient significance. Model diagnostics are reviewed as a way to ensure that a linear regression model is an appropriate choice for the data. Simple extensions of linear regression to include interaction terms and polynomial terms are discussed as ways to improve model fit. Other important concepts related to model design are introduced, such as input variable selection, dummy variables, collinearity and multicollinearity.

## Chapter 5

**Book Title:** Handbook of Regression Modeling in People Analytics: With Examples in R and Python  
**Chapter Author**: Keith McNulty (ORCiD: 0000-0002-2332-1654)  
**Chapter Number & Title**: Chapter 5, Binomial Logistic Regression for Binary Outcomes

Binary outcomes are commonplace in people analytics and many other fields. Promotion and attrition are examples of binary outcomes in the workplace, and often other forms of outcomes are converted to binary through the use of cutoffs. This chapter deals with modeling the likelihood of a binary event using binomial logistic regression. It introduces the logistic function as an approximation of a probability distribution, and proceeds to demonstrate how this gives rise to a highly interpretable model based on the concept of odds. Using an example related to the promotion of salespeople, the chapter goes through the steps of fitting a binomial logistic regression model, measuring its fit with the data, interpreting the coefficients as log odds and converting these to odds ratios. The concepts of model parsimony and simplification are introduced. More complex topics such as determining goodness of fit and performing diagnostics on these models are also touched upon in this chapter, which also sets a foundation for studying other log-likelihood models covered in later chapters.

## Chapter 6

**Book Title:** Handbook of Regression Modeling in People Analytics: With Examples in R and Python  
**Chapter Author**: Keith McNulty (ORCiD: 0000-0002-2332-1654)  
**Chapter Number & Title**: Chapter 6, Multinomial Logistic Regression for Nominal Category Outcomes

Frequently outcomes of interest will take the form of several different events that have no natural order or ranking to them. For example, the choice of a political party in an election or the choice of career following graduation are all nominal categorical outcomes. This chapter covers common options for how to model and explain such outcomes. Using an example of the choice of a health insurance product by company employees, the chapter starts with the concept of stratified binomial models which model each category as a binary outcome versus the other categories. The chapter then proceeds to describe the multinomial approach, where the likelihood or relative risk of each category is determined against a common reference category. The chapter covers the interpretation of the coefficients of a multinomial model relative to its reference, how to change the reference and how to simplify the model without risking the loss of important variables.

## Chapter 7

**Book Title:** Handbook of Regression Modeling in People Analytics: With Examples in R and Python  
**Chapter Author**: Keith McNulty (ORCiD: 0000-0002-2332-1654)  
**Chapter Number & Title**: Chapter 7, Proportional Odds Logistic Regression for Ordered Category Outcomes

Ordered category or ordinal outcomes are arguably the most common outcomes in people analytics. Most measurement scales have an ordinal form, such as performance ratings, survey responses, academic grades and ranked preference voting systems. This chapter covers the most common regression technique for modeling and explaining ordinal outcomes: proportional odds logistic regression. A simple, easily interpretable model is derived and then applied to an example related to the levels of discipline of players in soccer games. As in previous chapters, time is spent on the design, simplification and interpretation of this model and on assessing its fit and goodness-of-fit. Importantly, the chapter also covers the key assumption that underlies the proportional odds model and shows ways to determine whether or not that assumption is violated for any given data set. The chapter also briefly reviews alternative approaches to modeling ordinal outcomes when the proportional odds assumption is violated.

## Chapter 8

**Book Title:** Handbook of Regression Modeling in People Analytics: With Examples in R and Python  
**Chapter Author**: Keith McNulty (ORCiD: 0000-0002-2332-1654)  
**Chapter Number & Title**: Chapter 8, Modeling Explicit or Latent Hierarchy in Data

It is not uncommon for an analyst to have to consider hierarchy in data. Sometimes this hierarchy can be explicit and visible in the data and other times there is a belief that the data represents some latent, invisible factors at play. The first part of this chapter covers modeling situations where the data has an explicit hierarchy. This can happen when observations are of individuals who are members of groups, or where multiple observations are taken of the same individual over time. The concept of mixed (hierarchical or multilevel) models that have fixed effects and random effects are introduced. Data from a speed dating experiment is used to illustrate a mixed binomial logistic regression model and time is spent illustrating the difference when interpreting a mixed model versus a standard fixed effects model. The second part covers a common situation where large, often disorganized, survey instruments need to be modeled. Using the example of a survey on voting preferences, structural equation modeling is introduced as a method used to confirm a smaller set of unmeasured latent variables which explain the responses to the measured survey items (confirmatory factor analysis), and then explain the outcome of interest against the latent variables (structural modeling).

## Chapter 9

**Book Title:** Handbook of Regression Modeling in People Analytics: With Examples in R and Python  
**Chapter Author**: Keith McNulty (ORCiD: 0000-0002-2332-1654)  
**Chapter Number & Title**: Chapter 9, Survival Analysis for Modeling Singular Events Over Time

In standard regression techniques, there is no time component. However, many outcomes in people analytics and in other fields can happen at different times over a period of study. Promotion or attrition can occur at different times for different individuals. Subscription sign-up or cancellation can occur at differing points for different customers. This chapter covers the common elements of survival analysis, which considers the time to event as well as the event itself in explaining the outcome of interest. Using an example of employee attrition, the chapter covers Kaplan-Meier estimates of survival rates and survival curves as a way of establishing *prima facie* relevance of a variable to the outcome. Cox proportional hazard models are studied as a way of performing multivariate regression on a time to event outcome. The underlying assumptions behind Cox models are reviewed and methods for checking these assumptions are presented. Frailty models are then briefly studied as a method for conducting regression on a population where there are differing background risks of an event occurring.

## Chapter 10

**Book Title:** Handbook of Regression Modeling in People Analytics: With Examples in R and Python  
**Chapter Author**: Keith McNulty (ORCiD: 0000-0002-2332-1654)  
**Chapter Number & Title**: Chapter 10, Alternative Technical Approaches in R and Python

Analysts will sometimes need to work with models that have more predictable outputs or will need to run models efficiently over many subgroups of nested data. Additionally, many analysts are more familiar with Python as their primary analytics tool and do not work frequently in R. This chapter brings together the various methods studied in the book and illustrates different ways of running these methods in both R and Python. In R, the focus is on producing tidier, more predictable output from models, integrating models into ‘tidyverse’ frameworks to allow efficient modeling on grouped or nested data sets and creating abstracted and generalizable model frameworks for frequent use. The chapter then covers how to implement the various models in the book using Python, starting with linear regression and binomial regression and moving through multinomial regression, structural equation modeling and survival analysis.

## Chapter 11

**Book Title:** Handbook of Regression Modeling in People Analytics: With Examples in R and Python  
**Chapter Author**: Keith McNulty (ORCiD: 0000-0002-2332-1654)  
**Chapter Number & Title**: Chapter 11, Power Analysis for Estimating Required Sample Sizes for Modeling

When evaluating potentially costly or labor-intensive experiment proposals involving regression analysis, it is important to know what conditions are required for the analytic methods to have a chance of showing significant results. This chapter covers the topic of statistical power in hypothesis testing and how it relates to sample sizes, effect sizes and other criteria. Examples such as concurrent validity studies of selection instruments are used to illustrate how to calculate the required minimum sample size to achieve a target statistical power. The chapter moves through power analysis for basic hypothesis testing and then on to linear and logistic regression. Different effect size measurements sich as Cohen’s and are introduced. Example code is provided in both R and Python.

## Chapter 12

**Book Title:** Handbook of Regression Modeling in People Analytics: With Examples in R and Python  
**Chapter Author**: Keith McNulty (ORCiD: 0000-0002-2332-1654)  
**Chapter Number & Title**: Chapter 12, Further Exercises for Practice

This chapter introduces a series of five typical problems in people analytics with data sets provided, with the aim of providing practice in choosing and applying the methods covered in the book. Examples relate to compensation, recruiting, performance, promotion and learning. Context is provided for each example along with a description and link to the data set. A series of discussion questions are presented to encourage up front consideration of the method and approach for each problem. Then a series of data exercises are listed to gently guide in the analytical approach. The methods intended to be used on these exercises range from simple statistical hypothesis testing, to elementary regression methods to more advanced methods such as mixed models and Cox proportional hazard models. The material in this chapter is particularly suitable as project work or homework for appropriate courses in statistics, regression or people analytics.