An introduction to Psychometrics using R

William Revelle
Northwestern University

Abstract
A brief syllabus of a one week course on psychometrics and R. Readings will be taken from Revelle (in prep). Many examples will use the psych package Revelle (2013). Students will be expected to do a number of applied problem sets using various R packages.

Contents

Day 1: Introduction to psychometrics and to R 1

Day 2: Basics of correlation and covariance, multiple correlation and regression 3

Day 3: Constructs, components, and factor models 3

Day 4: Classical Test Theory and Item Response Theory models of reliability 4

Day 5: Putting it all together: Structural Equation Models and Validity 5

Day 1: Introduction to psychometrics and to R

Psychometrics is that area of psychology that specializes in how to measure what we talk and think about. It is how to assign numbers to observations in a way that best allows us to summarize our observations in order to advance our knowledge. Although in particular it is the study of how to measure psychological constructs, the techniques of psychometrics are applicable to most problems in measurement. The measurement of

contact: William Revelle revelle@northwestern.edu
Draft version of February 6, 2014
intelligence, extraversion, severity of crimes, or even batting averages in baseball are all grist for the psychometric mill. Any set of observations that are not perfect exemplars of the construct of interest is open to questions of reliability and validity and to psychometric analysis.

Although it is possible to make the study of psychometrics seem dauntingly difficult, in fact the basic concepts are straightforward. This course is an attempt to introduce the fundamental concepts in psychometric theory so that the reader will be able to understand how to apply them to real data sets of interest. It is not meant to make one an expert, but merely to instill confidence and an understanding of the fundamentals of measurement so that the student can better understand and contribute to the research enterprise.

At first glance, it would seem that we have an infinite way to collect data. Measuring the diameter of the earth by finding the distance to the horizon, measuring the height of waves produced by a nuclear blast by nailing (empty) beer cans to a palm tree, or finding Avogadro’s number by dropping oil into water are techniques that do not require great sophistication in the theory of measurement. In psychology we can use self report, peer ratings, reaction times, psychophysiological measures such as the Electric Encephelagram (EEG), the basal level of Skin Conductance (SC), or the Galvanic Skin Response (GSR). We can measure the number of voxels showing activation greater than some threshold in a functional Magnetic Resonance Image (fMRI), or we can measure life time risk of cancer, length of life, risk of mortality, etc. Indeed, the basic forms of data we can collect probably are unlimited. But in fact, it is possible to organize these disparate forms of data in terms of an abstract organization in terms of what is being measured and in comparison to what.

The challenge of psychometrics is assign numbers to observations in a way that best summarizes the underlying constructs. The ways to collect observations are multiple and can be based upon comparisons of order or of proximity. But given a set of observations, how best to describe them? This is a problem not just for observational but also for experimental psychologists for both approaches are attempting to make inferences about latent variables in terms of statistics based upon observed variables.

For the experimentalist, the problem becomes interpreting the effect of an experimental manipulation upon some outcome variable in terms of the effect of manipulation on the latent outcome variable and the relationship between the latent and observed outcome variables. For the observationalist, the observed correlation between the observed Person Variable and Outcome variable is interpreted as a function of the relationship between the latent person trait variable and the observed trait variable, the latent outcome variable and the observed outcome variable and most importantly for inference, the relationship between the two latent variables.

(Revelle, in prep, Chapters 1, 2, and 3)
Day 2: Basics of correlation and covariance, multiple correlation and regression

A fundamental question in science is how to measure the relationship between two variables. The answer, developed in the late 19th century, in the the form of the correlation coefficient is arguably the most important contribution to psychological theory and methodology in the past two centuries. Whether we are examining the effect of education upon later income, of parental height upon the height of offspring, or the likelihood of graduating from college as a function of SAT score, the question remains the same: what is the strength of the relationship? We examine measures of relationship between two variables. Generalizations to the problem of how to measure the relationships between sets of variables (multiple correlation and multiple regression) are also discussed.

The general solution of how to determine the relationship between two variables and how to predict one from the other is to consider the ratio of the covariance between two variables to the variance of the predictor variable (regression) or the ratio of the covariance to the square root of the product the variances (correlation). This solution may be generalized to the problem of how to predict a single variable from the weighted linear sum of multiple variables (multiple regression) or to measure the strength of this relationship (multiple correlation). As part of the problem of finding the weights, the concepts of partial covariance and partial correlation will be introduced. To do all of this will require finding the variance of a composite score, and the covariance of this composite with another score, which might itself be a composite. Much of psychometric theory is merely an extension, an elaboration, or a generalization of these concepts. Almost all tests are composites of items or subtests. An understanding how to decompose test variance into its component parts, and conversely, an understanding how to analyze tests as composites of items, allows us to analyze the meaning of tests. But tests are not merely composites of items. Tests relate to other tests. A deep appreciation of the basic Pearson correlation coefficient facilitates an understanding of its generalizations.

(Revelle, in prep, Chapters 4 and 5)

Day 3: Constructs, components, and factor models

Parsimony of description has been a goal of science since at least the famous dictum commonly attributed to William of Ockham to not multiply entities beyond necessity. The goal for parsimony is seen in psychometrics as an attempt either to describe (components) or to explain (factors) the relationships between many observed variables in terms of a more limited set of components or latent factors.

The typical data matrix represents multiple items or scales usually thought to reflect fewer underlying constructs. At the most simple, a set of items can be be thought of representing random samples from one underlying domain or perhaps a small set of domains. The question for the psychometrician is how many domains are represented and how well does each item represent the domains. Solutions to this problem are examples of factor
analysis (FA), principal components analysis (PCA), and cluster analysis (CA). All of these procedures aim to reduce the complexity of the observed data. In the case of FA, the goal is to identify fewer underlying constructs to explain the observed data. In the case of PCA, the goal can be mere data reduction, but the interpretation of components is frequently done in terms similar to those used when describing the latent variables estimated by FA. Cluster analytic techniques, although usually used to partition the subject space rather than the variable space, can also be used to group variables to reduce the complexity of the data by forming fewer and more homogeneous sets of tests or items.

(Revelle, in prep, Chapters 6)

Day 4: Classical Test Theory and Item Response Theory models of reliability

Whether discussing ability, affect, or climate change, as scientists we are interested in the relationships between our theoretical constructs. We recognize, however, that our measurements are not perfect and that any particular observation has some unknown amount of error associated with that measurement for all measurement is befuddled by error (McNemar, 1946, p 294). When estimating central tendencies, the confidence interval of the mean may be estimated by the standard error of the sample observations, and may be calculated from the observed standard deviation and the number of observations. This is, of course, the basic concept of Gossett and Fisher.

Error may be both random as well systematic. Random error reflects trial by trial variability due to unknown sources while systematic error may reflect situational or individual effects that may be specified. Perhaps the classic example of systematic error (known as the personal equation) is the analysis of individual differences in reaction time in making astronomical observations. “The personal equation of an observer is the interval of time which habitually intervenes between the actual and the observed transit of a star...” (Rogers, 1869). Before systematic individual differences were analyzed, the British astronomer Maskelyn fired his assistant Kennebrook for making measurements that did not agree with his own (Stigler, 1986). Subsequent investigations of the systematic bias (Safford, 1898; Sanford, 1889) showed consistent individual differences as well as the effect of situational manipulations such as hunger and sleep deprivation (Rogers, 1869).

Classical test theory is concerned with the reliability of a test and assumes that the items within the test are sampled at random from a domain of relevant items. Reliability is seen as a characteristic of the test and of the variance of the trait it measures. Items are treated as random replicates of each other and their characteristics, if examined at all, are expressed as correlations with total test score or as factor loadings on the putative latent variable(s) of interest. Characteristics of their properties are not analyzed in detail. This led Mellenbergh (1996) to the distinction between theories of tests (Lord and Novick, 1968) and a theories of items (Lord, 1952; Rasch, 1960). The so-called “New Psychometrics” (Embretson and Hershberger, 1999; Embretson and Reise, 2000; Van der Linden
and Hambleton, 1997) is a theory of how people respond to items and is known as Item Response Theory or IRT. Over the past twenty years there has been explosive growth in programs that can do IRT, and within R there are at least four very powerful packages: eRm (Mair and Hatzinger, 2007), ltm Rizopoulos (2006), lme4 (Doran et al., 2007) and MiscPsycho, (Doran, 2010). Additional packages include mokken (van der Ark, 2010) to do non-metric IRT and plink (Weeks, 2010) to link multiple groups together. More IRT packages are being added all of the time.

(Revelle, in prep, Chapters 7 and 8)

Day 5: Putting it all together: Structural Equation Models and Validity

Structural Equation Modeling combines the power of regression with the use of latent variables. Three powerful sem packages in R are lavaan Rosseel (2012), sem Fox et al. (2012) and OpenMx Boker et al. (2011). Although easy to use, the interpretation of results require some caution in that seemingly directional hypotheses can be misinterpreted.

At its most basic, validity addresses the question of does a test measure what we think it measures.

(Revelle, in prep, Chapters 9 and 10)

References


