Week 3:
Checks of Data Integrity

The Steps of Data Analysis

6 Week Webinar Class

Presented By

William Bannon, PhD

StatsWhisperer

New York, New York

Checks of Data Integrity

The process of making **Checks of Data Integrity** is concerned with one central question: **Are the Data Appropriate for Data Analysis?** This is perhaps the most important question of any study, even more so than the research question(s) and hypothesis(es), as if the data are not appropriate for data analysis, the research question(s) and hypothesis(es) are likely to be answered incorrectly.

Here (pages 85-182 & 248-254, for sample studies one and two, respectively) we review four areas that are critical to consider when assuring data integrity, which are statistical power, test assumptions, missing data, and measurement tools.

Statistical Power

In order to perform an effective quantitative study, we must be sure we have an appropriate level of statistical power, which includes having a sufficient number of study participants in our sample (please see pages 86-89 & 249-252 regarding sample studies one and two, respectively, for details). The power analysis is the primary tool in statistics used to determine how many study participants you need for your study.

Power Analysis

The power analysis indicates the number of study participants needed to generate enough statistical power to examine the relationships between study variables, such as the independent and dependent variables. A power analysis is perhaps best conceptualized as a dish composed of four ingredients, which are settings for power (pages 86-88 & 249-252), alpha (pages 87-88 & 249-252), effect size (pages 87-88 & 249-252), and sample size (pages 87-89 & 249-252). Each of these topics are expanded in the text. In a power analysis, the sample size required for the study is most often calculated based upon the other three settings (the power, alpha, and effect size).

Sample Power Analyses

Please see pages 86-89 for a sample power analysis for a test of Multiple Linear Regression, i.e., for a study with a continuous dependent variable measured at one timepoint.

Please also see pages 249-252 for a sample power analysis for a test of Binary Logistic Regression, i.e., for a study with a dichotomous (a categorical variable with two categories) dependent variable measured at one timepoint.

Why Only One Power Analysis? Page 89 mentions, that typically, within a grant proposal, manuscript, research report, etc., one study power analysis is presented, which is based upon the final multivariate model.

Test assumptions

All quantitative models have certain assumptions that must be reflected within the data before a statistical procedure can be applied. Violations of these assumptions can lead to statistical findings that misrepresent the study data. Please see pages 89-182 & 248-253 for a description of the test assumptions for sample studies one and two, respectively.

Test Assumptions: Multiple Linear Regression

Pages 89-182 describe that there are five principal assumptions (along with several smaller assumptions) that must be met regarding the multiple linear regression procedure, which is the final multivariate model to be used in the first sample study.

Specifically, this section presents six subsections that describe essential test assumptions that must be addressed when performing multiple linear regression, including a normal distribution (pages 90-119), multicollinearity (pages 119-126), homoscedasticity (pages 126-133), linearity (pages 133-138), no undue influence of outlier scores (pages 138-142), and other test assumptions (pages 142-144). Within the text, each section discusses what the assumption is, how to test to see if the assumption has been met, and what to do if testing indicates that the assumption has not been met.

Test Assumptions: Binary Logistic Regression

Page 253 describes the principal assumptions made when conducting a study employing a binary logistic regression as the final multivariate procedure, as is the case in the second sample study. We see far fewer assumptions regarding the data need to be examined (i.e., only checks for *multicollinearity*), relative to the multiple linear regression procedure used in the first sample study.

Missing Data

Missing data generally refers to the presence of a missing data value within a study variable(s) for a particular case(s). Within every study, a researcher must assess the level of missing data, as well as formulate a solution to account for the missing data (please see pages 144-171 for greater detail). In general, the greater the challenge

presented by the missing data, the more sophisticated the method needed to address the challenge**.**

Assessing the level of challenge presented by the missing data, as well as formulating a solution for the missing data involves defining missing data in a study (see pages 145-148), assessing the amount of missing data (see pages 148-158), assessing the pattern of the missing data (see pages 158-166), and make a selection regarding the treatment of missing data (see pages 166-171).

Within this section of the text (pages 144-171) this process is applied to the data used in samples studies one and two as an illustration of how these topics may be addressed. There is also mention that if possible and appropriate: **The best way to handle missing data is not to have any!**

Measurement Tools

The study measurement tools play a significant role in determining if data are appropriate for analysis (see pages 171-182 for greater detail). Specifically, it is necessary to assess the **reliability** and **validity** associated with each tool, toward assessing if measurement tools and the data are appropriate for statistical analysis.

Pages 172-179 of the text present issues related to the **reliability** of the study measurement tools. A primary focus of this section in one of the most commonly expressed dimension of reliability in quantitative research, which is internal consistency measured via the Cronbach's alpha. Several dimensions of study measurement tool **validity** (pages 180-182) are also presented.

In Summary

The checks of data integrity mentioned in this section are usually the most time consuming and complex tasks within a quantitative study. In fact, the actual inferential statistical analysis is often far less complex relative to these checks of data integrity. However, these checks are the foundation of the study. In short, if these checks are not made correctly, the entire study is unstable. Therefore, it is both necessary to perform these checks, as well as report how these checks were made when publishing a manuscript, research report, etc. If these checks are not reported, it is difficult to be confident that these procedures were completed to the reader's satisfaction or done at all. This is a prime reason for stating that performing and reporting these checks is a true make of sophistication within a quantitative study.