Checking Regression Assumptions When Using Multiple Imputation for Missing Data

Arndt Regorz, M.Sc.

Abstract

Multiple imputation is currently the most sophisticated technique for dealing with missing data in a regression analysis. However, this approach does not change the need to check the regression assumptions before interpreting the results of the analysis. The key question is what data should be used to assess whether the regression assumptions are met. In this brief tutorial, recommendations for this choice are presented depending on the missing data mechanism present in the data (missing completely at random, missing at random).

KEYWORDS: REGRESSION ASSUMPTIONS, MULTIPLE IMPUTATION, MISSING DATA

1. Introduction

If you want to conduct a simple or multiple regression analysis with incomplete data, one of the most sophisticated techniques currently available to deal with missing data is *multiple imputation* (van Ginkel et al., 2020). Provided sufficient imputation samples are used, the efficiency of this method can be quite high (Sinharay et al., 2001). However, regression analyses have assumptions, and these assumptions are as relevant with missing data as with complete data; the use of multiple imputation does not change that (Rubin, 1996). This raises the following questions: with what data should you test the different regression assumptions when using multiple imputation for a linear regression analysis? Currently, there is little practical information on this. This short tutorial is intended to fill this gap.

2. Different Missing Data Mechanisms

The literature distinguishes between three different missing data mechanisms (Baraldi & Enders, 2010; Rubin 1976):

With *missing completely at random* (MCAR), the probability of missing data for a particular variable is not related to other measured variables in the sample or to the values of that particular variable. With *missing at random* (MAR), the presence of missing data for a particular variable is related to other measured variables, but not to the values of that particular variable. With *missing not at random* (MNAR), the missing values for a particular variable are systematically related to the (hypothetical) missing values.

How you should handle regression assumptions depends on the missing data mechanism you assume for the data.

3. MCAR

If the missing data in your sample are MCAR, the subsample of complete cases is a random sample of the hypothetical complete data set (Baraldi & Enders, 2010). In this case, if the sample is random, the set of complete cases is representative of the distribution in the population. Therefore, you can evaluate the regression assumptions by looking only at the complete cases.

In practice you run a complete case analysis (listwise exclusion) and check the regression assumptions there by, for example, running the regression with listwise exclusion and analyzing the residuals (normality, homoskedasticity, outliers).

4. MAR

If the missing data are MAR, the complete cases are not necessarily representative of a full data set. It is possible that some elements are overrepresented in the complete data, while others are underrepresented. That could lead to incorrect conclusions about the regression assumptions if you were to base your decisions on a complete case analysis. Therefore, in this case, you need to look at the imputed data to check the regression assumptions. But how many imputation samples do you have to look at? Is one sufficient? Or do you need to look at all imputation samples when checking the regression assumptions?

When multiple imputation was first introduced, a recommendation for the number of imputation samples was five, as simulation analysis showed that this number of samples generally produced unbiased parameter estimates (Schafer, 1999). The number of five remains as a default in several software packages for multiple imputation (IBM, 2021; van Buuren, 2021). However, the current view is that one should use a higher number of imputation samples, e.g., between 20 and 100 (Austin et al., 2021). The reason for this, and it is important for our question, is not that a lower number leads to biased parameter estimates. A higher number of samples is used because it can reduce the standard error of the pooled estimates and thus increase the test power (Austin et al., 2021).

Therefore, I recommend checking the regression assumptions with five imputation samples drawn randomly from all imputation samples, since for the regression assumptions we are not interested in the standard errors, but in the estimates. One way to do this is to decide, before running the imputation, which samples you will use to check the assumptions. For example, if you have 20 imputation samples, you could commit in advance to analyze the imputation samples #4, 8, 12, 16, and 20 for possible violations of the regression assumptions. If the result is unanimous (an assumption is met or is not met for all five samples), then you have your result. However, if you get conflicting data on an assumption, you have a choice: you could increase the number of imputation samples to assess the regression assumptions or, which I recommend, play it safe and assume that the assumption is indeed violated.

5. MNAR

If the missing data mechanism is MNAR, then the issue of checking regression assumptions with multiple imputation is not longer relevant, because in this case one should not use multiple imputation in the first place (or at least should not trust its results).

6. Summary

When analyzing MCAR data, you can check regression assumptions by looking at the complete cases. For MAR data you should look at imputed data sets; this paper recommends checking the regression assumptions with *five* randomly drawn imputed data sets. And when the data is MNAR you should not use multiple imputation at all.

References

Austin, P. C., White, I. R., Lee, D. S., & van Buuren, S. (2021). Missing data in clinical research: a tutorial on multiple imputation. *Canadian Journal of Cardiology, 37*(9), 1322-1331. https://doi.org/10.1016/j.cjca.2020.11.010

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5-37. https://doi.org/10.1016/j.jsp.2009.10.001

IBM. (2021). *IBM SPSS missing values 28.* https://www.ibm.com/docs/en/SSLVMB_28.0.0/pdf/IBM_SPSS_Missing_Values.pdf

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*(434), 473-489. https://doi.org/10.1080/01621459.1996.10476908

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8*(1), 3-15. https://doi.org/10.1177/096228029900800102

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, *6*(4), 317-329. https://doi.org/10.1037/1082-989X.6.4.317

van Buuren, S. (2021). *Package "mice".* The Comprehensive R Archive Network. https://cran.r-project.org/web/packages/mice/mice.pdf

van Ginkel, J. R., Linting, M., Rippe, R. C., & van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment, 102*(3), 297-308. https://doi.org/10.1080/00223891.2018.1530680

Citation

Regorz, A. (2022). C*hecking regression assumptions when using multiple imputation for missing data* (Issues in Applied Statistics 1/22). Regorz Statistik. http://www.regorz-statistik.de/en/regression_assumptions_multiple_imputation.pdf

Arndt Regorz
Alemannenstrasse 6
D 44793 Bochum
Germany
www.regorz-statistik.de/en
mail@regorz-statistik.de