Assumptions in Multiple Linear Regression

Paul F. Tremblay

January 2019

The first important point to note is that most of the assumptions in bivariate or multiple linear regression involve the residuals. Note that the residuals (i.e., the Y – Y' values) refer to the residualized or conditioned values of the outcome variable Y. It is beyond the scope of this paper to show why the assumptions involve the residualized Y values and not Y values. Suffice it to say that it is violation of the assumptions at the residual level that will be associated with incorrect values and inferences about the regression coefficient population values (parameters).

Before focusing on the residuals, let's consider the predictor(s). There are no distributional assumptions involving the predictors (Xs). The only issue is whether the predictors are measured without error and this relates to whether these variables are fixed or random as discussed below.

X is fixed (or measured without error). As Fox (2008) points out, X values are often sampled (e.g., observational or survey research) rather than fixed (e.g., three drug doses selected apriori in an experimental design). In other words, on a questionnaire scale that varies from 15 to 75, we don't know ahead of time, how many cases will fall in each increment. Furthermore, when a particular case has an observed value of 37, we don't know if the true score is indeed 37 or whether there is some measurement error. An assumption is that X is measured without error (i.e., there is no measurement error). When X is not error-free, the regression coefficient will be attenuated (lower than the parameter value in the population). It is not uncommon to use a correction for unreliability in the estimation of correlations in two measures. Note however that a measure can have different sources of error (e.g., lack of internal consistency, temporal stability, or inter-rater agreement). The use of latent variables in structural equation modeling separates True from Error variance in constructs of interest and therefore provides better estimates of regression coefficients. In a regression model, any measurement error in the outcome variable Y is absorbed in the residual, and the regression coefficient will not be biased. However, the standardized regression coefficient and the proportion of variance explained by the predictor will be attenuated.

X is uncorrelated with the residuals. The residual variance is the proportion that is not explained by X and therefore can include omitted causes of Y as well as random error. The assumption of independence would be satisfied as long as the omitted cause is unrelated to X. Otherwise there would be a correlation between X and e (residual). When this assumption is not satisfied, we have made an error in specification (Fox, 2008; Kline, 2011). The consequence is that the regression coefficients will be biased. The important point here is to strive for a model that includes all important predictors especially when these predictors overlap with each other.

Example. Let's say I use Number of drinks as a predictor of Aggression, the residual would include unknown sources of variation. I know from previous research that one of these unknown sources would be Sex (i.e., men get into more physical fights than do women) and Sex correlates with the Number of drinks (men drink more). So here the residual correlates with the predictor because there is an important omitted variable (Sex) that has not been brought into the model. The impact is that the regression coefficient associated with Number of drinks will be biased (lower or higher than the population parameter) when Sex is not included in the model.

Linear relationship between X and Y. Non-linear associations can be modeled in different ways (e.g., adding a quadratic component). Modeling a non-linear relation without taking into account the non-linear component would lead to inaccurate results.

Assumptions Regarding Errors/Residuals

Mean of 0. The residuals at each level of the predictor X in a bivariate regression or at each combination of the predictors (Xs) in a multiple regression should have a mean of 0. Departure from a mean of 0 could be due to a nonlinear relation between the predictor(s) and the outcome (Williams et al., 2013).

Normality of the errors. The errors (residuals) are normally distributed. Violation of this assumption is problematic only in small samples (Williams et al., 2013).

Independence of residuals. Again this assumption involves the residuals and not the actual Y values. In multilevel or clustered data, where the observations at the lower level are often non-independent (i.e., they are correlated), disregarding this non-independence will probably result in non-independence of the residuals. Multilevel modeling addresses this problem by separating the variance in the outcome attributable to the individual units at the lowest level and the clusters at the higher level(s). Violation of this assumption can cause significant inaccuracies in the standard error and confidence intervals of the regression coefficient estimates.

Homoscedasticity. The variance of the errors (residuals) remains the same at different values of the predictor (X) in a bivariate regression or combinations of the predictors in a multiple regression. Violation of this assumption can lead to inaccurate inferential tests. One approach to address this violation is to use robust estimation or bootstrapping methods (Williams et al., 2013).

References

Fox, J. (2008). *Applied regression analysis and generalized linear models. Second edition*. Thousand Oaks, CA: Sage Publications.

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling. Third Edition*. New York: Guilford Press.

Williams, M. N., Grajales, C. A. G., Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research & Evaluation*, *18*, 1 - 14. https://pareonline.net/getvn.asp?v=18&n=11