Assumptions in Multiple Regression

Paul F. Tremblay

September, 2013

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

X is fixed (or measured without error). As Fox (2008) points out, X values are often sampled (i.e., observational research) rather than fixed by design. In this case, an imposed assumption is that X is measured without error (i.e., there is no measurement error), and errors are uncorrelated with X (see next). 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.

Errors (residuals) are uncorrelated with X. 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. 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).

Assumptions Regarding Errors/Residuals

Mean = 0.

Independence of residuals. The observations are samples independently (e.g., no clustering effects, observations not temporally linked)

Homoscedasticity. The variance of the errors (residuals) remains the same at different values of the predictor (X)

Normality of the errors. The errors (residuals) are normally distributed.

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.