Are Linear Regression Techniques Appropriate for Analysis When the Dependent (Outcome) Variable Is Not Normally Distributed?

Linear regression is a common technique used in the association study between the targeted outcome and some potential risk factors (e.g., age, sex). The violation of the normality assumption sometimes may be attributed by the skewed nature of the dependent variable, and may be a concern for naturally skewed outcome variables, such as best corrected visual acuity,1 refractive error,2 and Rasch score.3–6 The validation of normality sometimes can be ignored in the application of linear regression models.1,2,5,6

Normality violation will affect the estimates of the standard error (SE) and the confidence interval, and hence the significance of the risk factors. Nonparametric regression model or bootstrap techniques are suggested to be performed as they provide more robust estimates of SE.3,4 However, nonparametric techniques require large sample sizes to supply the model structure, and are very sensitive to the outliers.7 Thus, a key question is whether simple linear regression modelling still is valid if the “normality assumption” is violated.

First, we suggest there is a common misconception of the need to meet the “normality assumption” in linear regression techniques, and the validity of performing linear regression is compromised when this assumption is violated. Typically, the “normality assumption” often is checked from the histogram of the dependent variable. Statistically, however, it is more accurate to check that the errors of a linear regression model are distributed normally or the dependent variable has a conditional normal distribution (rather than if the dependent variable complies fully with a normal distribution) when evaluating whether the “normality assumption” is fulfilled for linear regression.

Second, by the law of large numbers and the central limit theorem,8 the ordinary least squares (OLS) estimators in linear regression technique still will be approximately normally distributed around the true parameter values, which implies the estimated parameters and their confidence interval estimates remain robust. Hence, in a large sample, the use of a linear regression technique, even if the dependent variable violates the “normality assumption” rule, remains valid.

We illustrate the concepts graphically. In Figure 1, we show that the outcome, Y, is non-normally distributed but is conditional normally distributed as error term is from normal distribution. Simulated non-normal or skewed error terms data in Figure 2 show trend of decreasing variations in estimates and standard errors with increasing sample size, indicating the accurateness and efficiency of linear regression estimates, although the normality assumption is violated.

In short, when a dependent variable is not distributed normally, linear regression remains a statistically sound technique in studies of large sample sizes. Figure 2 provides appropriate sample sizes (i.e., >3000) where linear regression techniques still can be used even if normality assumption is violated. Diagnostic checking in regression relationships nevertheless is important and, although linear regression still is appropriate in many situations, there are many other pitfalls.

Figure 1. Y is non-normally distributed but is conditional normally distributed.

Figure 2. Efficiency of estimation as sample size increases if normality assumption is violated.

3082

Copyright 2012 The Association for Research in Vision and Ophthalmology, Inc.
that may affect the quality of the interpretations and conclusions drawn from poorly fitted models.

Xiang Li¹,²
Wanling Wong¹,²
Ecosse L. Lamoureux¹,³
Tien Y. Wong¹,²,³

¹Singapore Eye Research Institute, Singapore National Eye Centre, Singapore; ²National University of Singapore, Singapore; and ³Centre for Eye Research Australia, University of Melbourne, Australia.
E-mail: ophwty@nus.edu.sg

References


Citation: Invest Ophtalmol Vis Sci. 2012;53:3082–3083. DOI:10.1167/iovs.12-9967