Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 18, Number 11, September 2013 ISSN 1531-7714

Assumptions of Multiple Regression: Correcting Two Misconceptions

Matt N. Williams, *Massey University, Auckland, New Zealand*
Carlos Alberto Gómez Grajales, *University of Veracruz, Poza Rica, Mexico*
Dason Kurkiewicz, *Iowa State University, Iowa.*

In 2002, an article entitled "Four assumptions of multiple regression that researchers should always test" by Osborne and Waters was published in PARE. This article has gone on to be viewed more than 275,000 times (as of August 2013), and it is one of the first results displayed in a Google search for "regression assumptions". While Osborne and Waters' efforts in raising awareness of the need to check assumptions when using regression are laudable, we note that the original article contained at least two fairly important misconceptions about the assumptions of multiple regression: Firstly, that multiple regression requires the assumption of normally distributed variables; and secondly, that measurement errors necessarily cause underestimation of simple regression coefficients. In this article, we clarify that multiple regression models estimated using ordinary least squares require the assumption of normally distributed errors in order for trustworthy inferences, at least in small samples, but not the assumption of normally distributed response or predictor variables. Secondly, we point out that regression coefficients in simple regression models will be biased (toward zero) estimates of the relationships between variables of interest when measurement error is uncorrelated across those variables, but that when correlated measurement error is present, regression coefficients may be either upwardly or downwardly biased. We conclude with a brief corrected summary of the assumptions of multiple regression when using ordinary least squares.

Testing of assumptions is an important task for the researcher utilizing multiple regression, or indeed any statistical technique. Serious assumption violations can result in biased estimates of relationships, over or under-confident estimates of the precision of regression coefficients (i.e., biased standard errors), and untrustworthy confidence intervals and significance tests (Chatterjee & Hadi, 2012; Cohen, Cohen, West, & Aiken, 2003). Unfortunately, the reporting of assumption checking in social science research articles is often relegated to a sentence or two, if that, in the method section. We might hope that most researchers nevertheless thoroughly and appropriately investigate the statistical assumptions of their analyses of choice, but we suspect that such a hope would be decidedly optimistic. In fact, a recent analysis of a sample of psychological researchers' data analysis practices found that assumptions were rarely checked, and the sample's knowledge about the assumptions of basic statistical tests was poor (Hoekstra, Kiers, & Johnson, 2012). Osborne and Waters' (2002) attempt to draw attention to the assumptions of multiple regression is therefore commendable, especially so in that it was published in an open-access journal (PARE). It is a testament to both the usefulness of clear writing on this topic and the success of PARE's open access model that Osborne and Waters' article has been viewed more than 275,000 times as at August 2013 (as per the hit counter on the html version of the article). This phenomenal number of page views achieves particular significance when we consider that Tenopir and King (2000) estimate that the average scientific article in the United States is read only 900 times. Osborne and Waters' article is also currently one of the first five results for a Google search for the search terms *regression assumptions*, no doubt contributing largely to its popularity. Its impact on the scientific literature has likewise been far from trivial, with Google Scholar listing 219 papers and books as citing the article as at June 2013.

It is the very popularity and ready accessibility of Osborne and Waters' article that prompts us to pen this response more than a decade after its original publication. Our concern is that Osborne and Waters' article contained two fairly substantial misconceptions about the assumptions of multiple regression. These misconceptions are that multiple regression requires the assumption of normally distributed *variables*; and that measurement error can only lead to under-estimation of bivariate relationships. Misconceptions about distributional assumptions can have serious consequences, including the expending of effort on checking unnecessary assumptions, the performing of problematic transformations and "corrections", and the neglect of the actual assumptions of the analysis being used. In this paper we correct the misconceptions contained in Osborne and Waters' article, making use of simple computer simulations to illustrate our points. We also provide a brief corrected summary of the assumptions of multiple regression. For simplicity, our examples are restricted to the bivariate or "simple" regression case—i.e., just one predictor and one response variable. Our statements nevertheless apply to both multiple and simple linear regression, and indeed can be generalized to other instances of general linear models with a single dependent variable such as between-subjects ANOVA and ANCOVA, and independent samples *t*-tests. Comments are restricted, however, to models in which the estimation method is ordinary least squares (OLS)—as is usually the case.

Desiderata for a statistical estimator

Before discussing the assumptions of multiple regression, it is important to discuss what we need to make these assumptions *for*. Remembering that a regression coefficient based on sample data is an estimate of a true regression parameter for the population the sample is drawn from, there are three particularly important properties for a statistical estimator (Dougherty, 2007). These three properties are true of regression coefficients calculated via ordinary least squares—provided that certain assumptions are met (Cohen et al., 2003).

Unbiased: An estimator is unbiased if its expected value (mean) is the same as the true parameter value in the population. In other words, an unbiased estimator has no systematic bias: It does not have a general tendency to over- or under-estimate the true parameter.

Consistent: An estimator of a parameter is consistent if the estimate converges to the true value of the parameter as the sample size increases. I.e. its accuracy tends to improve as the sample size grows larger.

Efficient: Efficiency refers to the accuracy of the estimates produced by the estimator. An estimator may be referred to as efficient if it is the most accurate (i.e., its variance is the smallest) of all unbiased estimators for the given parameter.

Aside from these three properties, it is also often desirable to assume a particular probability distribution for the sampling distribution of a given test statistic. A sampling distribution is the distribution of a particular statistic over repeated samplings from a population. For example, it is conventional to assume that the estimate of a regression coefficient will be normally distributed over repeated samplings, allowing researchers to make inferences about the value of the given regression parameter via confidence intervals and/or significance tests. The validity of this assumption, however, depends on the assumption of normally distributed model errors (at least when working with small samples), and this is the issue we turn to next.

The Normality Assumption: It's All About the Errors

In their summary of the assumptions of multiple regression, the first of four assumptions given focus by Osborne and Waters (2002) is the normality assumption. Osborne and Waters state: "Regression assumes that variables have normal distributions" (p. 1). They do not explicate which variables in particular they are referring to, but the implication seems to be that multiple regression requires that the predictor and/or response variables be normally distributed. In reality, only the assumption of normally distributed *errors* is relevant to multiple regression: Specifically, we may assume that errors are normally distributed for any combination of values on the predictor variables.

It is important to define at this point what we mean by errors, especially as the term is unfortunately

used to denote two different concepts that are relevant to a regression model. In a regression model, errors are the difference between subjects' observed values on the response variable and the values predicted by the true regression model for the population as a whole. This usage of the term *error* needs to be distinguished from the concept of *measurement* error, which will be defined and discussed later in this article.

The errors of a regression model cannot usually be directly observed, of course, since we rarely know the parameters of the true regression model. Instead, it is possible to investigate the properties of the errors by calculating the *residuals* of a regression model estimated using sample data (Weisberg, 2005). The residuals are defined as the differences between the observed response variable values and the values predicted by the estimated regression model. Another way of stating the normality assumption is that for any given combination of values on the predictor variables, we assume that the conditional distribution of the response variable is normal[1]—even though we do not assume that the marginal or "raw" distribution of the dependent variable is necessarily normal.

Osborne and Waters (2002) do mention briefly the assumption of normality of errors, but say that regression is robust to this assumption and do not give it any further discussion. The assumption of normally distributed errors is useful because when it holds true, we can make inferences about the regression parameters in the population that a sample was drawn from, even when the sample size is relatively small. Such inferences are usually made using significance tests and/or confidence intervals. However, when the sample is small, and errors are not normally distributed, these inferences will not be trustworthy. Normality violations can degrade estimator efficiency in at least a technical sense: When errors are normally distributed, OLS is the most efficient of all unbiased estimators (White & MacDonald, 1980), whereas in the presence of non-normal errors it is only the most efficient in the class of *linear* unbiased estimators (Wooldridge, 2009). More concretely, non-normal errors may also mean that coefficient *t* and *F* statistics may not actually follow *t* and *F* distributions.

On the other hand, the assumption of normally distributed errors is not required for multiple regression to provide regression coefficients that are unbiased and consistent, presuming that other assumptions are met. Further, as the sample size grows larger, inferences about coefficients will usually become more and more trustworthy, even when the distribution of errors is not normal. This is due to the central limit theorem which implies that, even if errors are not normally distributed, the sampling distribution of the coefficients will approach a normal distribution as sample size grows larger, assuming some reasonably minimal preconditions. This is why it is plausible to say that regression is relatively *robust* to the assumption of normally distributed errors.

The misconception that the normality assumption applies to the response and/or predictor variables is problematic in that there are certainly situations where the response and/or predictors are not normally distributed, but a normal distribution for the errors is still plausible. As one example, dichotomous predictors are often used in multiple regression; although such predictors are clearly not normally distributed, the errors of regression models using dichotomous predictors may still be normally distributed, allowing for trustworthy inferences. Furthermore, dichotomous variables that are particularly strong predictors of a response variable may induce a bimodality to the marginal distribution of the response variable, even if the errors are normally distributed. This is one situation in which neither predictor nor response variable has a normal distribution, despite the model errors being normally distributed.

Normality assumption simulation

The following simulation presents a situation where a dichotomous predictor *X,* that has a very strong effect on a response variable *Y,* results in the response variable not taking a normal distribution, despite the errors being normally distributed. The simulation was completed in R 2.15.2; the relevant code is attached in an appendix for readers interested in

[1] This alternative formulation of the normality assumption may be particularly helpful when considering generalized linear models, in which distributions other than the normal may be assumed for the conditional distribution of the response variable.

replicating our results. For this simulation, we define the true population model as the following:

$$Y_i = 5X_i + e_i$$

where $e \sim N(0, 1)$

In other words, in this scenario, the response variable *Y* is equal to the value of the predictor variable *X* multiplied by five, plus an error term. The error term is normally distributed with a mean of zero, and a standard deviation of one. In the simulation, *X* will be a fixed dichotomous variable, with an equal number of cases in each group (as might be the case for, say, a randomized controlled trial). For the purpose of illustration, this scenario is one where the effect size is very large: Approximately five standard deviation units. The reader may also note that given that we are analyzing the relationship between a single dichotomous predictor and a continuous response variable, it might be more conventional to use an independent samples *t*-test or perhaps an ANOVA. In fact, independent samples *t*-tests and between-subjects ANOVA are just special cases of regression, having the same assumptions and resulting in the same inferential statistics.

In the first step of the simulation, we simulate a single sample of 30 participants, with 15 participants in each of the two subsamples formed by the *X* variable. This allows us to offer a visual depiction of the distribution of the response variable *Y*. In Figure 1 a histogram shows that due to the strong influence of the dichotomous predictor variable *X*, the response variable *Y* is bimodal. Its non-normality is also clear in a normal q-q plot, where the quantiles of the distribution do not match those of a normal distribution, as indicated by the straight line. A Shapiro-Wilk normality test also provides evidence to reject a null hypothesis of a normal distribution for this variable, $W = 0.910$, $p = .015$. On the other hand, if we regress *Y* on *X* and then calculate the residuals, there is no evidence to reject a null hypothesis of normality for the marginal distribution of the errors, $W = 0.971$, $p = .562$. In sum, despite the presence of normally distributed errors, the response variable in this simulated example is clearly not normally distributed. The predictor variable is obviously likewise not normally distributed, being dichotomous.

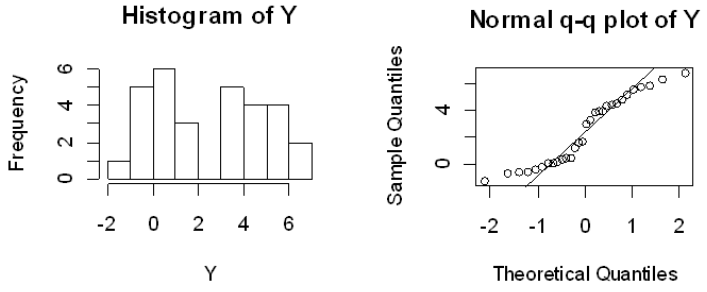

Figure 1. Histogram and Normal Q-Q Plots for the Simulated Response Variable Y

We can now proceed to check whether a regression model still produces trustworthy results in this scenario of non-normal predictor and response variables, but normal errors. Firstly, we will briefly check that regression via ordinary least squares provides coefficients that are *unbiased*. We will do this by running a simulation in which we generate a large number of samples (10,000), each sample having a total of 30 cases, or 15 cases in each subsample formed by the dichotomous predictor variable. A linear regression model is then fit in each sample, the coefficient for the effect of *X* on *Y* is estimated, and summary statistics are calculated for the coefficients. The results are displayed in Table 1.

Table 1: Results from simulation testing unbiasedness of coefficient estimates with non-normal X, non-normal Y, and normal errors

Statistic	Results
N of samples generated	10,000
Mean coefficient estimate	5.002
Minimum coefficient estimate	3.518
Maximum coefficient estimate	6.679
Standard deviation of coefficient estimates	0.362

The simulation demonstrates that the estimates of the regression coefficient for *X* are unbiased: The mean estimate nearly exactly equals the true parameter of 5, although of course the estimates vary around the true value. The unbiasedness of the estimates is

unremarkable, and in actuality the assumption of normally distributed errors is not required to achieve unbiasedness of coefficient estimates. We will omit evaluations of consistency and efficiency, but coefficients estimated in this situation would be both consistent and efficient.

Although not essential to achieve unbiasedness of regression coefficients, normally distributed errors are required to achieve trustworthy inferences (e.g., confidence intervals) in small samples (Weisberg, 2005). We can investigate the trustworthiness of confidence intervals calculated via OLS regression in this scenario by evaluating the *coverage* of confidence intervals. The coverage of a confidence interval is the proportion of intervals calculated using the given estimator that actually contain the true value of the parameter of interest, such as a regression parameter. If a 95% confidence interval has correct coverage, this means that if a large number of samples are drawn from the population of interest and a confidence interval calculated based on the data from each sample, 95% of these intervals will contain the true parameter value. In the simulation below, we generate 10,000 samples using the model discussed previously (dichotomous *X*, non-normal *Y*, normal errors). Once again, each sample contains 30 cases. We then investigate the coverage of 95% confidence intervals for the regression parameter for the predictor variable *X*. Results are presented in Table 2 **Error! Reference source not found.**.

Table 2: Results from simulation testing unbiasedness of coefficient confidence intervals with non-normal X, non-normal Y, and normal errors

Statistic	Results
N of samples generated	10,000
Number of intervals including the true parameter	9,518
Coverage	95.18%

This simulation demonstrates that the coverage of 95% confidence intervals is nearly exactly correct at 95.18%. The fact that only a finite number of samples can be generated explains the very slight difference of 0.18%. This result demonstrates the trustworthiness of small-sample inferences in this scenario in which neither the predictor nor the response variable is normally distributed (but errors are normally distributed). While we intentionally used a rather exaggerated case for illustrative purposes (a dichotomous predictor with a very strong effect on the response variable), scenarios substantively similar to this one may occur in real life. It is therefore important that researchers using multiple regression investigate how the residuals from their regression model behave, in order to determine how well they fit the assumption of normally distributed errors for the model under consideration. On the other hand, investigations of the distributions of the response and predictor variables may be useful for the sake of description, but have less bearing on whether the assumptions of multiple regression are actually met.

The Effects of Measurement Error on Regression Coefficients

We will now switch our focus to another kind of error: *Measurement* error. The formal definition of measurement error differs somewhat across different theories of measurement (e.g., classical test theory versus latent variable theory), but a loose conceptual definition is that measurement error is the difference between an observed score and either the subject's *true score* or the subject's actual level of the attribute of interest. Osborne and Waters (2002, p. 2) state that the absence of measurement error is an assumption of multiple regression, and claim: "In simple correlation and regression, unreliable measurement causes relationships to be *under-estimated*". (Simple regression involves only one predictor and one response; Osborne and Waters correctly note that in *multiple* regression, coefficients may be upwardly or downwardly biased by measurement error). They go on to provide formulae for correcting the attenuating effects of measurement error on zero-order and partial correlation coefficients.

The formulae provided by Osborne and Waters are closely related to classical test theory, and attempt to estimate the relationships between *true scores* on the measured variables. In classical test theory, true scores can be conceptualized as such: If we were able to administer a particular test to the same individual an extremely large number of times, with each

administration being independent of the other administrations, and the individual's level of the trait of interest remaining unchanged, then his or her true score would be the average score across all these administrations (Lord & Novick, 1968; Raykov & Marcoulides, 2011). In other words, an individual's true score on a test is his or her *expected* score on the test. The "correction for attenuation" formulae provided by Osborne and Waters allow for the estimation of zero-order and partial correlations between true scores on different variables, but only under the restrictive assumption that measurement error is uncorrelated across these variables. While classical test theorists have typically relied on this assumption, it is not guaranteed by the axioms of classical test theory, and may be false in real world situations (Zimmerman & Williams, 1977; Zimmerman, 1998).

Complicating the issue further is the fact that the true score relates to the reliability of scores, and not their validity. An individual's true score is defined as his or her expected score on the test itself, and is *not* necessarily the respondent's actual level of the particular attribute of interest, such as anxiety, intelligence, depression, and so forth. Attributes such as intelligence or anxiety, which are not directly observable, are commonly termed *latent* variables. It is the relationships between such latent variables that social scientists often wish to actually draw inferences about. If measurement error is correlated across the measured variables, regression coefficients may be downwardly or upwardly biased estimates of the actual relationships between the latent variables, depending partly on the magnitude and direction of the correlation between measurement error terms. Importantly, the "corrections" for measurement error suggested by Osborne and Waters simply are not designed to estimate relationships amongst latent variables: They are designed to estimate relationships between true scores. There is no particular basis to assume that they will improve estimates of relationships between latent variables. We illustrate this point with a simple simulation in which correlated measurement error results in regression coefficients *over*-estimating the relationship between two latent variables, where a "correction" for attenuation exacerbates rather than solves this issue.

This simulation explicitly takes a *latent variable* perspective on measurement. Other theories about the essential nature of measurement are possible, and include the classical concept of measurement (Michell, 1999), classical test theory, and representationalism (see Borsboom, 2005). However, given the pervasive use of latent variable models such as factor models, structural equation models, and item response theory models in the social sciences, and indeed recent research appearing on the pages of PARE (e.g., Baryla, Shelley, & Trainor, 2012; Han, 2012; Thompson & Weiss, 2011), we expect that the latent variable perspective on measurement will be the most familiar and relevant to readers.

Measurement error simulation

Imagine a scenario where a researcher is interested in the relationship between latent variable X and latent variable Y as presented in Figure 1. These variables might be particular cognitive abilities, personality traits, consumer satisfaction components, levels of psychopathology, or whatever example the reader prefers. The researcher cannot tap into direct error-free measurements of these latent variables, but instead has to obtain scores on particular tests: "Test score X" and "Test score Y". These tests are of course imperfect, and each subject to a degree of measurement error. Observed scores on each test are caused by a combination of an effect of measurement error, and an effect of the latent variable the test is meant to be measuring. The complicating factor we will include in this illustrative scenario is that measurement error is correlated across the two instruments. We will simulate a population of 10,000 observations, with a correlation of exactly 0.15 between the latent variables X and Y, and a correlation of 0.30 between measurement errors on the two variables. For simplicity, the regression coefficients for the effects of the latent variables and measurement error terms on the observed test scores are all set to 1, as are the variances of the latent variables; measurement error terms are set to a variance of 0.5. Using the simulated data, we can then calculate the correlation between observed scores on test X and test Y in this population. We will focus on the correlation coefficient at this point to facilitate an investigation of the effects of the correction for

measurement error suggested by Osborne and Waters (which is designed for correlation coefficients). The results would be otherwise similar were we to use an unstandardised regression coefficient to evaluate the relationship between test scores X and Y.

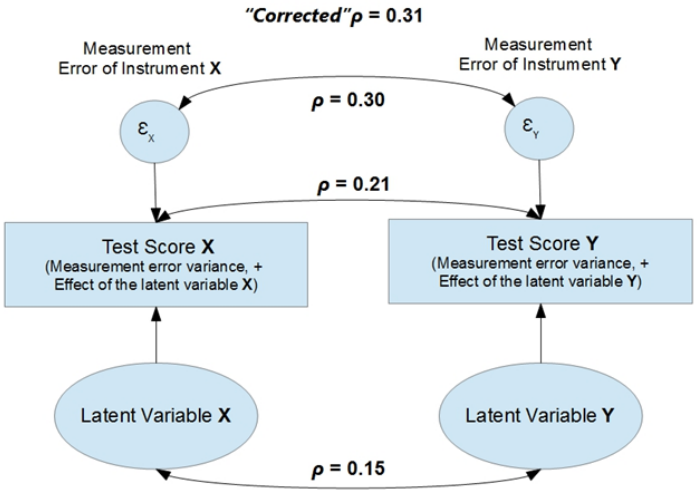

Figure 2. Path Diagram of the Measurement Error Simulation

The results are concerning. In this scenario, the population correlation between latent variables *X* and *Y* is just 0.15, but the correlation between observed scores on tests X and Y (ρ = 0.21) actually overestimates the correlation between the latent variables. This is due to the presence of correlated measurement error (ρ = 0.30). Of course, an actual researcher is unlikely to have access to data for a full population; the correlation between observed scores on tests X and Y has an expected value of 0.21 across samples from the population of 10,000 cases, but estimates based on sample data would fluctuate somewhat around this value. Estimates of the correlation between scores on test X and Y based on samples from the population of 10,000 cases will nevertheless be upwardly biased estimates of the population correlation between latent variables X and Y in this scenario.

Attempting to correct for the presence of measurement error using formula 1 provided by Osborne and Waters (the classic "correction for attenuation" formula) would not resolve this problem. Using the formulae in Raykov (2004) for calculating the reliability of measures in a latent variable model, we can obtain a population reliability value of 0.67 for scores on both Test X and Test Y. If we then apply Osborne and Waters formula 1 to obtain the "corrected" value for the population correlation between X and Y, the result is as below:

$$r_{12}^{*} = \frac{r_{12}}{\sqrt{r_{11}r_{12}}} = \frac{0.21}{\sqrt{0.67 * 0.67}} = 0.31$$

The "corrected" value for the population correlation between test scores X and Y of 0.31 thus even *further* overestimates the actual correlation of just 0.15 between the latent variables *X* and *Y*. The net result is that in this scenario, the presence of correlated measurement error means that the correlation between observed scores on tests X and Y overestimates the correlation between the latent variables of interest; and this problem is seriously *exacerbated* by the application of the correction for attenuation formula. Again, this simulation is conducted on the basis of a population dataset; estimates of the "corrected" correlation on the basis of samples from the population would fluctuate somewhat around the population value of 0.31.

The reader may object that this is an artificial example, and we have not provided a plausible case for why measurement error might be correlated across multiple measuring instruments or tests. In reality, a number of sources may produce correlated measurement error. For example, when the same method is used to measure multiple attributes, this may result in correlated measurement error across those variables (Andrews, 1984). Furthermore, situational variables such as variations in the health of participants and noise levels may cause correlations in the measurement error of attributes of participants measured at the same point in time (Zimmerman & Williams, 1977).

Measurement error, then, may certainly bias estimates of the relationships between particular constructs, but *not* necessarily in a predictably downwards fashion, even for simple bivariate regression (as Osborne and Waters seem to suggest). In turn, this means that researchers cannot comfortably assume that measurement error can only result in

making a given study's findings more conservative, in the sense of only reducing rather than inflating regression estimates.

Relatedly, we would caution against widespread use of the adjustments or "corrections" for measurement error that are suggested by Osborne and Waters: There is little basis to conclude that these adjustments will result in better estimates of relationships between the latent variables or constructs that researchers are interested in. Borsboom and Mellenbergh (2002) likewise argue against the use of the well-known correction for attenuation formula (formula 1 in Osborne and Waters), providing a more detailed examination of this particular question. Researchers who wish to account for the presence of measurement error when estimating the relationships between latent variables would be far better served by applications of modern latent variable modeling techniques such as structural equation modeling, which allow for the explicit modeling of (correlated and uncorrelated) measurement error.

So what are the statistical assumptions of multiple regression?

Having commented on two misconceptions about the assumptions of multiple regression, it is perhaps worthwhile closing with a brief (revised) summary of the assumptions of linear regression by ordinary least squares. The following assumptions apply regardless of whether simple (bivariate) or multiple linear regression is utilized, and also apply to other instances of general linear models with single dependent variables such as between-subjects ANOVA and ANCOVA, and independent samples *t*-tests. It is important to note in passing that we do not discuss assumptions about measurement levels in this article, restraining our focus to purely statistical assumptions. Some theoretical positions on measurement proscribe the use of parametric statistical procedures with data of certain measurement levels, such as the variety of representationalism advocated by S. S. Stevens (1946). However, this proscription does not necessarily apply to all theoretical positions on measurement (see Hand, 1996; Michell, 1986; Zand Scholten, 2011), and the *statistical* assumptions underlying parametric analyses do not include any assumptions about levels of measurement (Gaito, 1980). We refer the reader interested in the issue of measurement levels to the above-cited articles and omit further discussion of this issue here.

Assumption about the model: Linearity in the parameters

The model that relates the response Y to the predictors $X_1, X_2, X_3 \ldots X_p$ is assumed to be linear in the regression parameters (Chatterjee & Hadi, 2012). This means that the response variable is assumed to be a linear function of the parameters $(\beta_1, \beta_2, \beta_3 \ldots \beta_p)$, but not necessarily a linear function of the predictor variables $X_1, X_2, X_3 \ldots X_p$. Osborne and Waters (2002, p. 1) unfortunately repeat a common misconception in claiming that "Standard multiple regression can only accurately estimate the relationship between dependent and independent variables if the relationships are linear in nature". In reality, some types of non-linear relationships can be modeled within a linear regression framework. For example, a quadratic (U or reverse-U shaped) relationship between X and Y can be accommodated by including both X and X^2 as predictors, as in the equation:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

This regression equation is still a linear regression equation, because Y is modeled as a linear function of the parameters β_0, β_1 and β_2. On the other hand, a regression equation such as $Y_i = X_i^{\beta_1} + e_i$ is non-linear in the parameters, and cannot be modeled within a linear regression framework. As Osborne and Waters note, unmodeled non-linearity can be identified by plotting residuals against predicted values of Y. If the relationship between the response and predictor variables appears to take a form that is not linear in the regression parameters, non-linear regression models are available, although transformations may also be used to achieve a linear function in some cases (Chatterjee & Hadi, 2012). If the true model relating the predictors to the response variable is not of the form specified in the regression model (e.g., non-linear in the parameters, or simply of a different form to that specified), then the calculated coefficients will lead to erroneous conclusions about the strength and nature of the

relationships between the variables in the model. Furthermore, the important assumption that the errors have a conditional mean of zero will be breached in such a scenario; the formal consequences of this problem are reviewed below.

Assumptions about the model errors

It should perhaps go without saying that given that the following four assumptions apply to errors rather than the response and/or predictor variables, it is not possible to investigate these assumptions without estimating the actual regression model of interest itself. It is a common misconception that assumption checking can and should be fully completed prior to the running of substantive analyses; in reality, assumption checking should be an ongoing process throughout any data analysis.

1. Zero conditional mean of errors

The errors are assumed to have a mean of zero for any given value, or combination of values, on the predictor variables (Fox, 1997; Weisberg, 2005). When the conditional means of the errors are zero (and the other assumptions are also met), the desirable properties of OLS estimators discussed in this article apply regardless of whether the *X* values are fixed, as in an experiment, or random, as in sampled from a population (Berk, 2004; Snedecor & Cochran, 1980). On the other hand, if this assumption is violated, regression coefficients may be biased (Berk, 2004). Plausible reasons for a breach of this assumption include unmodeled non-linearity (e.g., if the model specifies a linear relationship between the predictor and the response and the true relationship is non-linear), or measurement error that is correlated across the response and predictor variable(s). As such, the sections above and below discussing assumptions with regard to linearity and measurement error provide advice that is also useful for identifying and responding to breaches of the assumption that the errors have conditional means of zero.

2. Independence of errors

The errors are assumed to be independent (Chatterjee & Hadi, 2012; Fox, 1997; Weisberg, 2005). Breach of this assumption leads to biased estimates of standard errors and significance, though the estimates of the regression coefficients remain unbiased, yet inefficient (Chatterjee & Hadi, 2012). Osborne and Waters (2002) state that independence of *observations* is required for linear regression, which is not entirely correct. Much in the same way that we assume that the errors (but not necessarily the raw variables) are normally distributed, we only need to assume independence of errors, not the observations themselves. In practice, many situations may produce dependent observations. For example, the observed values of data collected in the form of a time series may exhibit a form of independence breach in which observations are correlated with lagged values of the time series (i.e., current observations are autocorrelated with recent observations). However, a correctly specified time series model (e.g., perhaps including autoregressive terms) may result in independent errors and trustworthy results. The possibility of autocorrelated errors may be investigated by calculating an autocorrelation function (see Cryer & Chan, 2008), although other sources of error dependence may be identified using knowledge about the study design. For example, the use of cluster rather than random sampling can result in dependence of errors (Winship & Radbill, 1994). In general, the appropriate response to dependent errors depends on the source of this dependence. For example, the use of time series data may require the use of some form of time series analysis (see Cryer & Chan, 2008; Hamilton, 1994), while the analysis of nested data may require the use of a multilevel model (see Goldstein, 2011).

3. Homoscedasticity (constant variance) of errors

The model errors are generally assumed to have an unknown but finite variance that is constant across all levels of the predictor variables. This assumption is also known as the homogeneity of variance assumption. If the errors have a variance that is finite but not constant across different levels of the predictor/s (i.e., heteroscedasticity is present), ordinary least squares estimates will be unbiased and consistent as long as the errors are independent, but will not be efficient (Weisberg, 2005). The inference process will also be untrustworthy since conventionally computed confidence intervals and *t* and *F*-tests for OLS estimators can no longer be justified. As Osborne and Waters state, heteroscedasticity can be identified by

plotting standardized (or studentized) residuals against the predicted values of *Y*. When heteroscedasticity is encountered, several alternatives are available to the researcher. These alternatives include variance stabilizing transformations (Montgomery, Peck, & Vining, 2001; Weisberg, 2005), robust estimation methods for standard errors (e.g., Huber-White standard errors; White, 1980), bootstrap methods (Montgomery et al., 2001), estimation via Weighted Least Squares (Chatterjee & Hadi, 2012), or the specification of a Generalized Linear Model (Cohen et al., 2003; Montgomery et al., 2001).

4. Normal distribution of errors

This assumption has been discussed at length previously in this article. Normally distributed errors are not required for regression coefficients to be unbiased, consistent, and efficient (at least in the sense of being best linear unbiased estimates) but this assumption *is* required for trustworthy significance tests and confidence intervals in small samples (Cohen et al., 2003). The larger the sample, the lesser the importance of this assumption. This assumption formally applies to the distribution of the errors (or, equivalently, the conditional distribution of the response variable) for any given combination of values on the predictor variables. In some simple cases, such as for a single categorical predictor, it may be possible to investigate the distribution of residuals (or equivalently, the distribution of the response variable) at all values of the predictor variables. In many cases, however, there will be a very large number of possible values on the predictor variables. In this more general situation, it is only feasible to investigate the marginal distribution of the residuals, which may provide a reasonable guide to the accuracy of the normality assumption. A normal Q-Q plot may be useful for this purpose (see Cohen et al., 2003). Since the normality assumption is primarily of importance for small samples, non-normality of the errors may be addressed by increasing the sample size. When this is not possible, inference in small samples with non-normal errors can be achieved by using bootstrap methods (Efron & Tibshirani, 1986; Montgomery et al., 2001), or the specification of a Generalized Linear Model with an error distribution other than the normal (Cohen et al., 2003; Montgomery et al., 2001).

It is worth noting in passing that while the regression model requires only the normality of errors, the Pearson product moment *correlation* model requires that the two variables follow a bivariate normal distribution (Pedhazur, 1997). I.e., in the correlation model, both the marginal and conditional distribution of each variable is assumed to be normal.

Assumptions about measurement error

The predictor variables are assumed to be measured without error (Chatterjee & Hadi, 2012; Montgomery et al., 2001). Error in the *response* variable measurements (but not the predictors) will not harmfully affect inferences relating to unstandardized regression coefficients, provided this measurement error is not correlated with the predictor variable values. Aside from this special case, measurement error can result in either upwardly or downwardly biased coefficients, depending on whether measurement error is correlated or uncorrelated across the measured variables, and depending on the magnitude and direction of any correlations amongst error terms. Where measurement error exists for the predictors, or correlated measurement error exists for either the predictors or the response variable, analysis methods that allow measurement error to be explicitly modeled may be a better alternative to OLS regression. Structural equation modeling (see Kline, 2005) may allow for the detection (Raykov, 2004) and correction of both correlated and uncorrelated measurement error. For a more general introduction to psychometric theory and measurement, see Raykov and Marcoulides (2011).

Other potential problems

Although perhaps not best described as assumptions, since these are not theoretical constraints imposed in the definition of the General Linear Model, two important potential problems are often described in conjunction with discussions of the assumptions of linear regression: Multicollinearity and outliers.

1. Multicollinearity

The presence of correlations between the predictors is termed collinearity (for a relationship between two predictor variables) or multicollinearity (for relationships between more than two predictors).

In severe cases (such as a perfect correlation between two or more predictors), multicollinearity can mean that no unique least squares solution to a regression analysis can be computed (Belsley, Kuh, & Welsch, 1980; Slinker & Glantz, 1985). More commonly, less severe multicollinearity can lead to unstable estimates of the coefficients for individual predictors: That is, the standard errors and confidence intervals for the coefficient estimates will be inflated (Belsley et al., 1980). The extent to which multicollinearity is a concern depends somewhat on the aims of the analysis: If prediction is the main objective, multicollinearity is not a significant obstacle, as prediction of the response variable (including prediction intervals) will not be harmfully affected. If the aim is inference about population parameters, however, multicollinearity is more problematic. The variance inflation factor is one popular measure of multicollinearity, although several other diagnostics are available (Belsley et al., 1980; Cohen et al., 2003). Appropriate responses to multicollinearity may include the use of an alternative estimation method such as ridge regression (Montgomery et al., 2001), or principal components regression (Chatterjee & Hadi, 2012). Removing some of the highly correlated predictors may be considered too, but this solution is usually not ideal (Chatterjee & Hadi, 2012)

2. Outliers

In some cases, the results of a regression analysis may be strongly influenced by individual members of the sample that have highly unusual values on one or more variables under analysis, or a highly unusual combination of values. This is not necessarily a problem in itself, nor necessarily a justification for excluding such cases. However, if the outlying value(s) are a result of measurement or coding error such as a typographical mistake, or the result of the inclusion of a case that is not a member of the intended population, then the results of a regression analysis will obviously be deleteriously affected (Stevens, 1984). Several diagnostics are available to identify outliers (Belsley et al., 1980; Cohen et al., 2003), of which Cook's distance is perhaps the most widely used. Determining the correct course of action when outliers are detected may be a complex decision. It may not be justifiable to exclude an observation unless there is a valid substantive reason to consider it as an invalid observation (e.g., if there was an error in recording a data point, or if the case is not a member of the intended population; Montgomery et al., 2001). When outliers are excluded, it may be useful to present results both with and without outlier exclusions (Stevens, 1984). Alternatively, the use of "robust" regression methods may help to reduce the influence of outlying observations (Montgomery et al., 2001; Western, 1995).

Conclusions

Carefully considering the reasonableness of the assumptions of multiple regression in the context of a particular dataset and analysis is an important prerequisite to the drawing of trustworthy conclusions from data. It is our hope that this article will help everyday researchers to complete informed assessments of the actual assumptions of multiple regression and other general linear models when applying this important family of techniques. Thorough and well-informed assumption checks will help researchers to select appropriate analyses and, ultimately, to produce meaningful and robust conclusions.

References

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, *48*(2), 409–442. doi:10.1086/268840

Baryla, E., Shelley, G., & Trainor, W. (2012). Transforming rubrics using factor analysis. *Practical Assessment, Research & Evaluation*, *17*(4). Retrieved from http://www.pareonline.net/getvn.asp?v=17&n=4

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York, NY: John Wiley & Sons.

Berk, R. A. (2004). *Regression analysis: a constructive critique*. Thousand Oaks, CA: SAGE.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge, United Kingdom: Cambridge University Press.

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, *30*(6), 505–514. doi:10.1016/S0160-2896(02)00082-X

Chatterjee, S., & Hadi, A. S. (2012). *Regression analysis by example* (5th ed.). Hoboken, NJ: John Wiley & Sons.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Cryer, J. D., & Chan, K. (2008). *Time series analysis: with applications in R* (2nd ed.). New York, NY: Springer.

Dougherty, C. (2007). *Introduction to Econometrics* (3rd ed.). Oxford, United Kingdom: Oxford University Press.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*(1), 54–75. doi:10.1214/ss/1177013815

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: SAGE.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, *87*(3), 564–567. doi:10.1037/0033-2909.87.3.564

Goldstein, H. (2011). *Multilevel statistical models*. Chichester, United Kingdom: John Wiley & Sons.

Hamilton, J. D. (1994). *Time series analysis*. Cambridge, United Kingdom: Cambridge University Press.

Han, K. T. (2012). Fixing the c parameter in the three-parameter logistic model. *Practical Assessment, Research & Evaluation*, *17*(1). Retrieved from http://pareonline.net/pdf/v17n1.pdf

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *159*(3), 445–492. doi:10.2307/2983326

Hoekstra, R., Kiers, H. A. L., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, *3*. doi:10.3389/fpsyg.2012.00137

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: The Guilford Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, *100*(3), 398–407. doi:10.1037/0033-2909.100.3.398

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge, United Kingdom: Cambridge University Press.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2001). *Introduction to linear regression analysis* (3rd ed.). New York, NY: John Wiley & Sons.

Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*, *8*(2). Retrieved from http://pareonline.net/getvn.asp?v=8&n=2

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace College Publishers.

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy*, *35*(2), 299–331. doi:10.1016/S0005-7894(04)80041-8

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.

Slinker, B. K., & Glantz, S. A. (1985). Multiple regression for physiological data analysis: the problem of multicollinearity. *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology*, *249*(1), R1–R12.

Snedecor, G. W., & Cochran, W. G. (1980). *Statistical methods* (7th ed.). Ames, IA: Iowa State University Press.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin*, *95*(2), 334–344. doi:10.1037/0033-2909.95.2.334

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. doi:10.1126/science.103.2684.677

Tenopir, C., & King, D. W. (2000). *Towards electronic journals: Realities for scientists, librarians, and publishers*. Washington, DC: Special Libraries Association.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, *16*(1). Retrieved from http://www.pareonline.net/pdf/v16n1.pdf

Weisberg, S. (2005). *Applied linear regression*. Hoboken, NJ: John Wiley & Sons.

Western, B. (1995). Concepts and suggestions for robust regression analysis. *American Journal of Political Science*, *39*(3), 786–817. doi:10.2307/2111654

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, *48*(4), 817–838. doi:10.2307/1912934

White, H., & MacDonald, G. M. (1980). Some large-sample tests for nonnormality in the linear regression model. *Journal of the American Statistical Association*, *75*(369), 16–28. doi:10.2307/2287373

Winship, C., & Radbill, L. (1994). Sampling weights and regression analysis. *Sociological Methods & Research*, *23*(2), 230–257. doi:10.1177/0049124194023002004

Wooldridge, J. M. (2009). *Introductory econometrics: A modern approach*. Mason, OH: South-Western Cengage Learning.

Zand Scholten, A. (2011). *Admissible statistics from a latent variable perspective* (PhD dissertation). University of Amsterdam, Amsterdam, Netherlands. Retrieved from http://dare.uva.nl/record/366790

Zimmerman, D. W. (1998). How should classical test theory have defined validity? *Social indicators research*, *45*(1-3), 233–251. doi:10.1023/A:1006949915525

Zimmerman, D. W., & Williams, R. H. (1977). The theory of test validity and correlated errors of measurement. *Journal of Mathematical Psychology*, *16*(2), 135–152. doi:10.1016/0022-2496(77)90063-3

Appendix: R code for simulations

Generating a sample of data with dichotomous X, normal errors, and non-normal Y

```
set.seed(seed=123)
#setting seed for replicability
n.group <- 15
#subsample size of 15 cases (i.e., a small sample)
X = rep(0:1, each = n.group)
#create dichotomous X with equal samples sizes in each group
E = rnorm(2*n.group, 0, 1)
#Error is normally distributed with mean of zero, SD of 1
Y = X*5 + E
#Y is equal to 5*X plus normally distributed error
hist(Y) #histogram of Y
qqnorm(Y); qqline(Y) #q-q plot of Y
shapiro.test(Y) #Shapiro-Wilk test for response variable
fit<-lm(Y~X) #fit a linear model
shapiro.test(fit$res) #Shapiro-Wilk test for residuals
```

Testing unbiasedness of estimates

```
set.seed(seed=123)
fun1 <- function(n.group = 15){
   x <- rep(0:1, each = n.group)
   y <- 3+5*x + rnorm(n.group*2, 0, 1)
   o <- lm(y ~ x)
   return(coef(o))
}
out1 <- replicate(10000, fun1())
summary(out1[2,])
sd(out1[2,])
```

Testing coverage of confidence intervals

```
set.seed(seed=123)
CIfun <- function(n.group = 15){
   x <- rep(0:1, each = n.group)
   y <- 3+5*x + rnorm(n.group*2, 0, 1)
```

```
    o <- lm(y ~ x)
    xlimits <- c(confint(o)[2,1], confint(o)[2,2])
    return(xlimits)
}
out2 <- replicate(10000, CIfun())
#coverage
1-(sum(out2[1,]>5) + sum(out2[2,]<5))/ncol(out2)
```

Measurement error simulation

```
set.seed(123)
library(MASS)
sigmaL = matrix(c(1,0.15,0.15,1),2) #covariance matrix for latent variables
latents = mvrnorm(10000, mu = c(0,0), sigmaL, empirical = TRUE)
sigmaE = matrix(c(0.5,0.15,0.15,0.5),2) #covariance matrix for errors.
errors = mvrnorm(10000, mu = c(0,0), sigmaE, empirical = TRUE)
ObsX = 1*latents[,1] + 1*errors[,1]
ObsY = 1*latents[,2] + 1*errors[,2]
cor(ObsX, ObsY)
```

Acknowledgement

This article grew out of a discussion on the online statistics forum http://talkstats.com. A number of members of the forum made invaluable suggestions and comments at various stages during the drafting of this paper.

Citation:

Williams, Matt N., Grajales, Carlos Alberto Gómez, & Kurkiewicz, Dason (2013). Assumptions of Multiple Regression: Correcting Two Misconceptions. *Practical Assessment, Research & Evaluation*, 18(11). Available online: http://pareonline.net/getvn.asp?v=18&n=11

Corresponding Author:

Matt N. Williams
School of Psychology
Massey University
Private Bag 102904
North Shore, Auckland 0745
New Zealand
Mattnwilliams [at] gmail.com

Assumptions of multiple regression: Correcting two misconceptions

Williams, MN

2013-09-06