Regression Assumptions in Clinical Psychology Research Practice – A systematic review of common misconceptions

Anja F. Ernst and Casper J. Albers[1)]

University of Groningen

Heymans Institute for Psychological Research, Department of Psychometrics and Statistics, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands. [1)] Corresponding author: c.j.albers@rug.nl

Regression Assumptions in Clinical Psychology Research Practice – A systematic review of common misconceptions

Abstract

Misconceptions about the assumptions behind the standard linear regression model are widespread and dangerous. These lead to using linear regression when inappropriate, and to employing alternative procedures with less statistical power when unnecessary. Our systematic literature review investigated employment and reporting of assumption checks in twelve clinical psychology journals. ~~The selected journals were representative based on impact factor.~~ Findings indicate that normality of the variables themselves, rather than of the ~~residuals~~errors, was wrongfully held for a necessary assumption in 4% of papers that use regression. Furthermore, 92% of all papers using linear regression were unclear about their assumption checks, violating APA-recommendations. This paper appeals for a heightened awareness for and increased transparency in the reporting of statistical assumption checking.

Keywords: *Linear Regression, Statistical Assumptions, Literature Review, Misconceptions about Normality*

Regression Assumptions in Research Practice – A systematic review of common misconceptions

One of the most frequently employed models to express the influence of several predictors on a continuous outcome variable is the linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon_i.$$

This equation predicts the value of a case Y_i with values X_{ji} on the independent variables X_j (j = 1, …, p). The standard regression model takes X_j to be ~~fixed, i.e.~~ measured without error (cf. Montgomery, Peck & Vining, 2012, p.71). The various β_j slopes are a measure of association between the respective independent variable X_j and the dependent variable Y. The ~~residual, the~~ error ~~term~~ for the given Y_i, the difference between the observed value and value predicted by the population regression model, is denoted by ε_i and is supposed to be unrelated to the values of X_p. Here, β_0 denotes the intercept, the expected Y value when all predictors are equal to zero. The model includes p predictor variables. In case p = 1, the model is ~~called simple~~ denoted as the simple linear regression model.

The standard linear regression model is based on four assumptions. These postulate the properties that the variables should have in the population. The regression model only provides proper inference if the assumptions hold true (although the model is robust to mild violations of these assumptions). Many statistical textbooks (for instance, Miles & Shevlin, 2001; Cohen, Cohen, West & Aiken, 2003; Lomax & Hahs-Vaughn, 2012; ~~Tabachnick & Fidell, 2013~~Montgomery et al., 2012) provide more background on these assumptions as well as advice on what to do when these assumptions are violated.

Violations of these assumptions can lead to various types of problematic situations. First, estimates may become biased, that is~~i.e.~~ not estimating the true value on average. Second, estimators may become inconsistent, implying that ~~i.e.~~ convergence to the true value when the sample size increases is not guaranteed. Third, the ordinary least squares estimator~~s~~ may not be efficient anymore: ~~whilst not giving a 'wrong' estimate, other procedures are demonstrably better~~I.e., in the presence of assumption violations, OLS may provide less accurate parameter estimates than other available estimation procedures. Fourth and finally, NHST's and confidence intervals might become untrustworthy: *p*-values can be systematically too small or too large, and confidence intervals are too narrow or too wide. This can occur even if estimators are unbiased, ~~consistant~~consistent and efficient. For a more detailed description of these issues, see Williams et al. (2013). Please note that these assumptions are the assumptions when estimating using the Ordinary Least Squares (OLS) procedure, which is the default procedure in many software packages, including SPSS and *R*. Other type of estimation methods, such as GLS, apply other sets of assumptions.

Below, the four OLS~~-~~ assumptions will shortly be discussed. ~~For each assumption, Figure 1 displays what the scatterplot without violation of this assumption, and with mild or severe violation of this assumption can look like.~~

Linearity. The ~~relationship~~ conditional mean of the errors is assumed to be zero for any given combination of values of the predictor variables. This implies that, for standard multiple regression models, the relationship between every independent variable X_i and the population mean of the dependent variable Y, denoted by μ_Y, is assumed to be linear when the other variables are held constant. ~~This assumption is illustrated in the top row of Figure 1.~~ Furthermore, the relations between the various X_i and μ_Y are additive: thus, the relation of X_i with μ_Y is the same, regardless of the value of X_j ($j \neq i$). This relates to the issue of multicollinearity; a good model is expected to have as little overlap between predictors as possible. However, multicollinearity is not a model assumption but merely a necessity for a model to be parsimonious. Violation of this assumption can obviously occur when non-linear relations are unmodelled, but also in the case of measurement error (see Williams et al., 2013).

Normality. All ~~subpopulations defined by the values of the predictor variables are assumed to be normally distributed~~ around their mean, which implies that all ~~residuals~~ errors are normally distributed around zero. ~~Even though the linear regression model is quite robust to violations of this assumption (and the central limit theorem implies that for large samples this assumption is automatically, at least, approximately met) it is important to notice that the theoretical model of regression is constructed based on this assumption. Note that, unlike the linearity assumption, it is difficult to judge on basis of a scatterplot whether the assumption is violated, as can be seen in the second row of Figure 2. Alternative methods, such as QQ-plots are better suited for this.~~

Homoscedasticity. ~~***All***~~ The variance of the errors is the same for any combination of values of the independent variables. Thus, this variance~~subpopulations are expected to have an equal variance~~, which can then be denoted by a single symbol~~,~~ (e.g. σ^2). This assumption is also called the homoscedasticity assumption. Thus, the second and third regression assumptions combined specify that the ~~residuals~~ errors (ε_i) of the model should follow a normal distribution with a mean of zero and a (fixed) standard deviation σ. Heteroscedasticity often manifests itself through a larger spread of measurements around the regression line at one side of the scatterplot than at the other.~~, as is illustrated in the third row of Figure 1.~~

Independence. ~~***All***~~ The ~~residuals~~ error term~~s~~s ε_1, ε_2, …, ~~—~~ should be independent of one another: the pairwise covariances should be zero.~~ and all residuals should be independent of the observations. This implies that the observations should be independent of one another.~~ This assumption is not directly based on the distribution of the data but on the study design and it requires the sampling method to be truly random (see, for instance, Cohen, Cohen, West and Aiken, 2003). ~~Figure 1 (bottom row) displays violations of the independence assumption: there seems to be an some autocorrelated pattern in the model residuals.~~ As with the normality assumption, inspection

Comment [m1]: Did you mean to replace this text with something maybe?

Formatted: Font: Italic

of a scatterplot is not the best way to check for independence. A residual plot, or inspection of the autocorrelation of the residuals, is a better approach.

Comment [m2]: It's not quite clear here how you would use a residual plot / scatter plot to evaluate the independence assumption.

Common misconceptions about assumptions. There are many misconceptions about the regression model, most of which concern the ~~second and the third~~ assumptions of normality and homoscedasticity. Most commonly, researchers incorrectly assume that X_i, or both X_i and Y, should be normally distributed instead of the ~~residuals~~errors. Osborne and Waters (2002), a peer-reviewed article attempting to educate about regression assumptions, and with over ~~36~~540,000 online views times at the time of writing[1], make this mistake, demonstrating how widespread this misconception ~~illustrate how widespread this misconception~~ really is~~: this paper is a peer-reviewed article attempting to educate about regression assumptions~~, ~~yet it wrongly lists normality of the variables themselves as an assumption of linear regression instead of normality of residuals~~ (cf. Williams, Grajales & Kurkiewics, 2013). ~~The paper has been viewed online over 360,000 times.~~

Not assuming a normal distribution for X_i may seem counterintuitive at first~~,~~. ~~h~~However the indulgence of this assumption becomes more evident with an illustrative example. Take the standard Student's *t*-test which assesses if two distributions are statistically different from one another (e.g.~~:~~ ~~for~~ instance the *t*-test that compares the efficacy of a specific treatment compared to a placebo treatment). The population distributions in both conditions are assumed to be normally distributed with equal variances. This *t*-test can also be expressed as a regression model where the independent variable X dummy codes the group membership, ~~so~~ (i.e. if a participant is in the ~~control, (*X*~~control = 0~~)~~, or in the treatment condition ~~(,~~ $X = 1$~~)~~). This regression model and the *t*-test are mathematically equivalent and will thus lead to identical inference. Variable X will only attain two values, 0 and 1, as it is only used as label for group membership. The dependent variable Y will attain many different values: following a normal distribution for the treatment group and a (possibly other) normal

[1] Based on the journal's access counter, http://pareonline.net/genpare.asp?wh=0&abt=8

distribution for the control group. This resulting ‘condition membership’ distribution is nothing close to normal~~,~~. ~~h~~However, no assumption of the general linear model is violated because the *subpopulations* of *Y* for each of the *X* values follow a normal distribution with equal variances, as is visualised in Figure 1~~2~~. This example demonstrates that the assumptions of the *t*-test (standard normal distribution of the populations around the group mean and equal variances) coincide with the second regression assumption.

~~Although normality of the predictor variables is not a requirement of the model, it can be helpful for a range of reasons. It can enhance prediction through the enhancement of linearity between the independent variable and the dependent variable (Tabachnick & Fidell, 2013) and it reduces the problems corresponding to influential points (Miles & Shevlin, 2001). Most importantly, normality of variables is helpful when the predictor variables cannot be measured without error. In case of measurement error, parameter estimates can be biased (cf. Williams, Grajales & Kurkiewicz, 2013; Tabachnick & Fidell, 2013). When the predictor variables are normally distributed, however, the estimates will remain unbiased. As it is rarely the case in clinical practice that predictor variables are (all) measured without error, it is thus good practice to check for univariate and multivariate normality of observed scores. This, however, does not imply that one may neglect to check for normality of the residuals *as well*.~~

As a consequence of the second regression assumption, the distribution of the dependent variable conditional on some combination of values on the predictor variables, Y_i is actually normally distributed around μ_Y, the true conditional population mean. This becomes clear when remembering that the error of the regression estimation is normally distributed around mean zero and that Y_i is equal to $\mu_Y + \varepsilon_i$, that is, individual observations are the sum of the mean and a deviation from this mean. However, it is wrong to test the normality of the marginal distribution of the dependent variable *Y* because this would imply that all μ_Y values are the same which is, generally, not the case. (This situation occurs only when all regression slopes are zero and, thus, all predictor variables are linearly unrelated to *Y*.)

Regarding the linearity assumption, a common misconception is in thinking that only linear relationships can be modelled using the OLS framework. This is not the case: the linearity assumption deals with linearity in the parameters and the estimates, but not necessarily in the variables.

Consequences of violations of assumptions. Misconceptions like the ones outlined above ~~can~~ potentially ~~has~~ have severe effects on the ability to draw inferences from a data-analysis. First of all, the checking of wrong assumptions will most likely lead to the neglect of correct assumption checking. If the researcher will decide on a regression analysis without having tested the correct assumptions it is possible that some requirements of linear regression were not met~~. In that case *p*-values and confidence intervals will be biased.~~. However, in any case the neglect of correct assumption checking will always leave the reader or reviewer unable to trust the results because there is no way of knowing whether the model assumptions ~~could have been~~were actually met. Of course, the severity of this problem of non-transparency persists even when the researcher ensured the validity of all necessary assumptions and merely ~~missed~~ failed to report those findings. Not only does such non-transparency in data analysis lead to confusion ~~in~~ for researchers that are potentially interested in replicating or comparing the results, it also weakens the informational value of the research findings that are being interpreted.

A second problem that is caused by misconceptions about model assumptions occurs when a researcher decides against a linear regression analysis because of the violation of faulty assumptions that were unnecessary ~~to be met~~ in the first place. The difficulty of abandoning linear regression analysis for a non-parametric procedure is the fact that the ordinary least squares method of linear regression is a more powerful procedure than any of its non-parametric counterparts, if ~~the~~its assumptions are met. Hence, wrongfully deciding against the employment of linear regression in a data-analysis will lead to a decrease in power. ~~Especially because the regression model is quite robust to violations of the normality and homoscedasticity assumptions, one should only decide against the use of linear regression for valid reasons.~~ Thus, the understanding of the correct regression assumptions is crucial because it prevents the abandonment of the linear regression technique in cases in which it would be unjustified. Furthermore, the checking of assumptions has another advantage: it might help the researcher to think about conceptually alternative models. For instance, heteroscedasticity in the data could be a sign of an interaction between one or of the included independent variables ~~with~~ and an independent variable not (yet) included in the model.

Applying a linear regression model when assumptions are violated can lead to (severe) problems, but this does not have to be the case, depending on the type of violation. Violations of the linearity assumption and of the independence assumption can lead to biased, inconsistent and inefficient estimates (Chatterjee & Hadi, 2006; Williams et al., 2013). A proper check on these two assumptions ~~thus~~ is thus vital. ~~This is less the case~~The consequences of violations are less severe for the other two assumptions.

If normality of errors holds, the OLS method is the most efficient unbiased estimation procedure (White & MacDonald, 1980). If this~~ey~~ assumption does~~are~~ not hold (but the remaining assumptions do), OLS is only most efficient in the class of linear estimators (see Williams et al., 2013, for a detailed discussion). This implies that, as long as the other assumptions are met, estimates will still be unbiased and consistent in the presence of a normality violation, but the *p* --values might be biased. Furthermore, the central limit theorem implies that for large samples ~~this assumption is automatically, at least, approximately met~~the sampling distribution of the parameters will be at least approximately normal, even if the distribution of the errors is not. Hence, the regression model is robust with respect to violations of the normality assumption. Potential problems will, in practice, only occur in inferential problems (such as confidence intervals and testing) for small samples.

~~Also~~Similarly, violations of the homoscedasticity assumption are not necessarily problematic. Provided that the very mild assumption of finite variance holds, estimates will still be unbiased and consistent (Chatterjee & Hadi, 2006).

Best practices for checking of assumptions. There are many different ways to check the four assumptions of the regression model and there generally is no 'uniformly optimal' approach.

Generally, there are two classes of approaches: (i) formal tests (of the style 'H_0: the assumption is true' vs 'H_A: the assumption is violated') and (ii) graphical methods. For the normality assumption alone, there is an abundance of formal tests, such as the Shapiro-Wilk test, the Anderson-Darling test and the Kolmogorov-Smirnov test. Which approach is most powerful depends on the kind of violation from normality (Razali & Wah, 2011). However, the use of formal tests is discouraged (Albers, Boon & Kallenberg, 2000, 2001). Due to the nature of NHST, in α (usually 5%) of cases where the assumption actually is valid, the null hypothesis will still be rejected. Thus, applying a different approach in case of significant violations distorts the *p*-value distribution of the estimates of the regression model, even when no assumptions are violated.

Applying graphical methods is therefore a preferred approach. This is also suggested by the statistical guidelines for the APA set up by Wilkinson et al. (1999, p. 598): "Do not use distributional tests and statistical indices of shape (e.g. skewness, kurtosis) as a substitute for examining your residuals graphically". This advice builds upon the adagium by Chalmers et al. (1983) that "there is no single statistical tool that is as powerful as a well-chosen graph". A graph simply provides more information on an assumption than a single *p*-value ever can (see also Chatterjee & Hadi, 2006, Ch. 4).

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Comment [m3]: Your reasoning is a little unclear – maybe just re-word it a bit, or possibly highlight other reasons to not use statistical tests of normality. (To me, the main reason is simply that these tests only have good power when N is large – i.e., in exactly the scenario where the normality assumption doesn't actually matter).

Comment [m4]: Add page number for the quote

The linearity assumption can easily be checked using scatterplots or residual plots: plots of the residuals vs. either the predicted values of the dependent variable or against (one of) the independent variable(s). ~~(~~Note that residuals are the differences between the observed values and the values predicted by the *sample* regression model, whereas errors denote the difference with the values predicted by the *population* regression model. ~~)~~ Residual plots are also the best visual check for homoscedasticity. For the normality assumption, it is difficult to judge on the basis of a scatterplot whether the assumption is violated. A histogram of the residuals is also a poor visual check, as the 'shape' of the histogram heavily depends on the arbitrary choice of the bin width, especially in small samples. Normal probability plots, or QQ-plots, provide a much better way to check normality. Finally, a check on the independence assumption is done by studying the autocorrelation function of the residuals.

Comment [m5]: Could be worth explaining that method allow us to check for *temporal* dependence, but not other types of dependence structures (e.g., clustering of observations).

~~As with the normality assumption, inspection of a scatterplot is not the best way to check for independence. A residual plot, or inspection of the autocorrelation of the residuals, is a better approach.~~

Outline of this paper. Misconceptions about frequently employed statistical tools, like the *p*-value, are not rare, even amongst researchers (~~cf.~~see Bakker and Wicherts, 2011; Hoekstra, Morey, Rouder &~~and~~ Wagenmakers, 2014). Our paper aims to shed light onto potential misconceptions researchers and reviewers might hold about the linear regression model. Therefore, the documentary practices of psychological research papers with the linear regression model and its assumptions were investigated by means of a literature review. In this review, we investigate the proportion of papers where misconceptions around the assumptions of the statistical regression model occurred and which type of misconceptions occurred most often. This will provide important information, as the first step in solving flawed methodology in research is finding out where the flaws are and how predominant they are.

Although the consequences of incorrectly dealing with assumptions can be severe, the APA manual (American Psychological Association, 2010) barely provides guidelines on ~~this~~what to report and how to report ~~this~~. It does *recommend* being specific about "information concerning problems with statistical assumptions and/or data distributions that could affect the validity of findings" (p. 248) as part of the Journal Article Reporting Standards, but this is not obligatory. The APA Task Force on Statistical Inference (Wilkinson and Task Force on Statistical Inference, 1999) is more explicit in their recommendations: "You should take efforts to assure that the underlying assumptions required for the analysis are reasonable given the data. Examine the residuals carefully." (p. 598).

In this manuscript we present the findings of our literature review. Because the whole field of psychological science is too broad to study in a single paper, we ~~restirct~~restrict ourselves to ~~focus on~~ the field of ~~one~~ clinical psychology~~field of psychological research~~. We investigate how statistical assumptions were covered in various journals of clinical psychology and what types of misconceptions and mistakes are occurring most often. In the discussion section, possible explanations for the reported findings will be offered. The paper will conclude with several proposals of how potential shortcomings in the current practices with linear regression analysis could be overcome.

Method

Journals. The literature review restricted itself to articles that were published in clinical psychology journals in the year 2013. It is possible that problems with the checking of assumptions are less (or more) prominent in journals with a high impact, which is why we aimed for a selection of journals with varied impact factors. We employed the Scientific Journal Rankings (SJR) as reported on 16 December 2014 by the SCImago Journal and Country Rank ~~on the 16.12.2014~~ (SCImago, 2007) for clinical psychology journals of~~of~~ the year 2013 ~~(SCImago, 2014)~~ to divide all clinical psychology journals into four quartiles (Q1 – Q4), where Q1 contains the 25% of journals with the highest journal rank, etcetera. From every quartile the three highest ranked journals were selected to be included in the review. Hence, we obtained a balanced selection from all clinical psychology journals, as listed in Table 1. All articles published in the selected journals in 2013 were included, ~~including those that had already been published earlier as well~~including also papers that had potentially been published online earlier~~.~~. Letters, journal corrigenda, editorial board articles and book reviews were not included in the review. Basically, articles that were by design not containing a method section were not included in our lists of articles~~, also not in the section '*No Model of Interest*'~~. The focus of this review purely lies on published scientific articles.

Every article was retrieved directly from the official website of its respective journal (except for Q1.3 which was directly retrieved from its official database "PsycARTICLES"). All articles were in German (Q3.1), Spanish (part of Q3.3) or in English (all other). German articles were also included in the review; Spanish articles were excluded because of the authors' lack of proficiency in this language. Figure 2~~3~~ displays the Prisma workflow of the analysis. The conduction of our review adhered to the common meta-regression guidelines (Moher, Liberati, Tetzlaff, Altman, The PRISMA Group, 2009).

Procedure. It was evaluated whether and how papers ~~adhered to the spirit of the guidelines of the American Psychological Association (APA) which recommend a researcher~~ described ~~to~~ careful examination of e~~ion~~ ~~of~~ the data with regard to the underlying model assumptions whenever conducting statistical analysis (APA, 2010; Wilkinson et al., 1999). Papers were skimmed for the following criteria: if they had used linear regression, how they tested the regression assumptions or what kind of assumptions they indicated as being necessary, if they had transformed data on basis of correct or incorrect assumptions and if a paper had considered an ordinary least squares regression model but employed a different model on basis of either correct or incorrect assumptions. This resulted into a classification scheme of 12 different rubrics which are displayed in Table 2. This scheme is mutually exclusive and exhaustive; all studied papers are classified into exactly one rubric.

Papers that used linear regression were classified as follows. We assumed the most common misconception about linear regression to be the checking of the normality of the variables while failing to check the normality of the errors. Therefore, we created rubrics 8 to 11 to classify all papers that employed linear regression and checked or assumed the normality of *X* and/or *Y* but not of the errors. An example of a paper classified in rubric 8 ~~mentioned~~stated “Variable distributions were tested to ensure assumptions of normality, linearity, and variance equality were met, with no significant violations observed”. Often, when the normality assumption was mentioned it was unclear whether authors had checked the normality of errors or of the variables. Articles that were unclear in this regard were classified under rubric 5. For instance, one of the ~~an~~ articles classified in this ~~category would state~~ rubric stated “Preliminary analysis examined data for the presence of outliers and the appropriateness of assumptions of normality, linearity, and homoscedasticity~~.~~” with no more information provided on the assumption checks. Papers that indicated to have checked the most important assumptions (homoscedasticity and normality of the errors and linearity) were classified as “*Correct*” in rubric 4. Articles that mentioned at least a few correct assumptions, as opposed to giving no indication at all (rubric 7), were classified in rubric 6. Because all papers that checked or assumed the normality of *X* or *Y* but not of the errors were included in rubrics 8 to 11, we have named rubric 6 “*Did not test all but some correct assumptions, did not include normality of variables*”. After ~~having done~~performing the literature review it became apparent that none of the articles listed in this category had mentioned the normality of errors. Because we aimed to demonstrate how rare it is to read that researchers check the normality of the errors we have updated

Comment [m6]: I think you might as well cite the paper when giving these quotes – it’d be easy for a reader to find the paper anyway by googling the quote, so no point being polite by not saying who it was…

Comment [m7]: Bit confusing here – your earlier comments suggested that these are the two *less* important assumptions.

the name of the category into “*Did not test all but some correct assumptions, included neither normality of variables nor errors*”, even though the checking of the normality of errors was not employed as a criterion for inclusion in this category during the literature review.

Papers that did not fit into any of the eleven other ~~categories~~rubrics but included an aspect on linear regression assumptions that we found unsatisfactory were listed in the rubric “*Other misconceptions about assumptions*”. One example of a paper classified in this category claimed “All assumptions of multiple regression (linearity, multicollinearity, and homoscedasticity) were met” this paper was included in the category “*Other misconceptions*” because they did not only lack any mention whether normality of the residuals was checked (which would have resulted in a classification in rubric 6) but also claimed that a list not containing normality of residuals was complete. We found this claim unsatisfactory which was the reason we included this paper in rubric 12.

Whenever an article in our selection reported the results of a regression analysis of another paper or reviewed several linear regression articles, it was evaluated whether the paper reviewing all the previous regression analysis had made it a criterion of inclusion whether the assumptions have been met in the original articles. If a review article did not check or mention the assumptions of the papers that published the original analysis, the article was classified as ‘*Use of linear regression but no indication if any or which assumptions were tested*’. However, these sorts of papers constitute less than one percent of our selected articles. It should be noted that this only applies to papers which reported the data values of a linear regression or analysed regression results from other studies. A paper was not included if it only mentioned the direction of the outcomes of another paper’s regression model or stated that a relationship had been established by previous research findings.

Because the focus of this paper lies on the assumptions of linear regression, only linear regression model assumptions were examined in the literature review. Consequently, papers that analysed data by means of other types of regression, such as latent factor models, logistic regression, and proportional hazards models (Cox regression), were not inspected for assumption checking. ~~As long as~~When a paper used a ~~non-linear~~ regression model other than linear regression, and without mentioning that linear regression was ~~alternatively~~ considered for data analysis, it was classified as *'No Model of Interest'*.

Comment [m8]: We call regression via OLS "linear regression", but in reality models like logistic regression and many latent factor models are actually linear models too. (So I've reworded this so that you aren't referring to them as "non-linear").

Results

The results of the systematic literature review are displayed in Tables 3, 4 and 5 which display the number of occurrences of different classifications for the selected journals. In the online supplementary material we indicate for all of the 893 individual papers studied into which category they fall.

Table 3 shows the findings for all journals with the 12 different classification rubrics summarized into seven different columns. The three columns entitled ‘Dealing with assumptions’ list the number of different types of regression papers in a specific journal and shows the proportional amount of this type in relation to the complete number of regression articles in that journal. The two columns for ‘No regression’ list the number of papers which did not use a linear regression model and included in their method sections to have considered a linear regression analysis but decided against it on the basis of checking either correct or incorrect assumptions.

Table 4 specifies the details behind the articles which are listed in Table 3 under the column titled ‘incorrectly ~~wrong~~’. This table classifies the corresponding 10 papers into Rubrics 8 – 12 of Table 2. It may be noted that 4% of all articles that used linear regression checked normal distributions of some variables instead of normal distribution of ~~residuals~~errors.

Table 5 specifies the details behind the column ‘unclear’ in Table 2; that is it ~~i.e.~~ classifies the 159 corresponding papers into Rubrics 5 to 7 of Table 2. Of all papers that employed regression, 92% were unclear~~did not mention anything at all~~ about the assumptions of the linear regression model that were tested or were thought to be fulfilled.

Discussion

In our analysis, we studied 893 papers, representative for the work published in the field of clinical psychology, and classified the 172 papers (19.4%) which considered linear regression into three categories: those that dealt with the assumptions correctly, those that dealt with assumptions incorrectly, and those that did not specify how they dealt with assumptions.

Merely ~~a disappointing~~ 2% of these papers were both transparent and correct in their dealing with statistical assumptions. Furthermore, in ~~no less than~~ 6% of papers~~,~~ transparency was given but the dealing with assumptions was ~~plain wrong~~incorrect.~~; with the type of mistakes made being the type that is warned for in statistics textbooks aimed at undergraduate students in psychology.~~ Hoekstra, Kiers & Johnson (2012) might provide some insight into why researchers did not check assumptions. They list unfamiliarity with either the fact that the model rests on the assumption~~,~~ or with how to check the assumption as the top two reasons. As explained, incorrect dealing with the assumptions~~,~~ could lead to severe problems regarding the validity and power of the results. We hope that this manuscript creates new awareness of this issue with editors of clinical psychology journals and that this will assist~~s~~ in bringing down the number of publications with flawed statistical analyses.

A tremendous amount of papers that employed regression, 92% of those studied, were not clear on how they dealt with assumptions. It is not possible (not for us, nor for the reader) to judge from the text whether the analysis was performed correctly. ~~Given that, i~~In the group of transparent papers, the number of papers with fundamental mistakes in dealing with assumptions far outnumber the number of papers without mistakes. Thus, even though it is not possible to pinpoint an exact number to it,~~,~~ it would be naive to ~~-~~assume that only a small proportion of ~~is very reasonable to fear that a considerable proportion of~~ those 92% ~~is~~ also deal~~ing~~ with assumptions incorrectly.~~.~~

Comment [m9]: Do you mean something like “whether checks for assumption violations were performed correctly”?

We believe that most contemporary problems in the handling of regression methods could be counteracted by a more thorough coverage of the statistical assumption checks that were performed in order to determine the validity of the linear regression model. At the very least, transparency regarding how assumptions are approached, in line with the recommendations by Wilkinson et al. (1999), is essential. Thus, mentioning which assumptions were checked and what diagnostic~~is~~ tools were used to check them under what criteria, should be a minimum requirement. Preferably, the authors should ~~also~~ ~~show~~also show the results of these checks.

With transparency, the critical reader can distinguish correct approaches from incorrect ones, even if the author(s), editor(s) and referees fail to spot the flaws. These statistical checks can be given in the paper itself, but could also be provided in online supplementary material, a possibility most journals offer nowadays. Thus, increased length of the manuscript does not need to be an issue. Our aspiration for an increased transparency in statistical assumption checks is in line with recent developments in psychology such as open methods (obligatory in e.g. the APA-journal Archives of Scientific Psychology) and open data (either published as online supplementary material with a paper, or through special journals like Journal of Open Psychology Data)~~, which also encourage transparency~~. With open data, sceptical scientists can re-do the analyses and check ~~the~~ assumptions for themselves. Enforcing, or at least strongly encouraging, transparency can even have beneficial effects to the level of publications in the respective journal (Wicherts, Bakker and Molenaar, 2011). Even if publishing the data does not have a direct beneficial effect on the quality of work, it will be useful as it provides the sceptical reader with the required information to perform the assumption checks and thus the possibility to check the credibility of the published work.

Comment [m10]: Did you check supplementary materials for the articles you examined?

~~Another suggestion to improve the worrisome findings reported in this paper is to encourage authors to include a statistician or methodologist in the study more often. For statisticians, it is daily practice to correctly check assumptions (as well as dealing with all other challenges of data~~

~~analysis). It is our belief that many of the mistakes reported in this study could have been avoided if a statistician would have participated in the data analysis.~~

It is difficult to establish whether high ranking journals deal with assumptions more adequately than lower ranking journals. Even though the results in Table 5 indicate that higher ranked journals were more likely to test at least a few assumptions compared to lower ranked journals; the results do mainly show that there is great variability between journals regarding the number of papers with applied regression models they publish: two journals published no papers in 2013 that employed linear regression, and five journals published six or less of these papers. Because two of the three inspected Q1 journals are review journals they predominantly employed meta-regression, a special type of regression useful for conducting meta-analyses, and only rarely linear regression, it should be pointed out that of the 15 papers that used meta-regressions in our Q1.2 eleven tested at least some of the required assumptions (that is 73% of meta-regression papers were checked correctly for statistical assumptions). We believe that for these papers the percentage is much better than the overall percentage of 2% for applied regression papers~~,~~ because meta-analyses are usually carried out by a team of authors including at least one statistician or psychometrician.

We have limited our literature review to papers employing linear regression models, in order to keep the study feasible. We suspect that similar findings would arise when studying other classes of statistical models. Furthermore, we have also limited the review to papers published in the field of clinical ~~psychology,~~psychology; however we suspect that similar problems occur – albeit possibly in different proportions – in all areas of applied psychological research. Thus, our suggestions with respect to increased transparency and better evaluation of the employed methodology ~~are valid~~should be relevant for a wider range of papers than those studied here. Because our categorization of papers is reasonably straightforward, only one author conducted most of the review. While our rubrics allow objective classifications we cannot preclude a few single accidental ~~misclassification~~misclassifications. However, possible misclassification should be minimal at most and can therefore be expected to not have skewed the overall results that are based on a large number of papers. Thus, despite this limitation we are confident in the overall results. For future research, it would be interesting to do a similar literature review based on either alternative techniques or on another field of application. Furthermore, more research is needed in understanding the reasons that ~~underly~~underlie why researchers frequently do ~~not~~n't ~~check~~not check assumptions.

One of the consequences of the lack of reporting of assumption ~~checks,~~checks is that many published findings in clinical psychology are underestimating the uncertainty in their claims. For instance, reported confidence intervals in the literature describe the uncertainty surrounding the parameter, if the OLS-assumptions are met. The uncertainty of the validity of the assumptions should lead to wider confidence intervals, in general. For future research, it would be an interesting puzzle to assess the magnitude of this added uncertainty.

To summarise, in order to prevent the observed problems that were outlined above we suggest a more transparent methodological reporting. Research should cover which assumption checks were carried out. Furthermore, it should be mentioned if alternative statistical models have been considered and why they were not employed, if so. This will be a necessity for future research articles in order to be able to detect and prevent errors related to ~~the~~ widespread misconceptions but also to remove doubt from articles with an actual immaculate data analysis.

Additional information

A detailed breakdown of the systematic review, references to all websites employed to retrieve articles as well as a completed PRISMA checklist are provided as online supplementary material.

The search strategy has been carried out by Anja Ernst. Independently, Casper Albers checked and classified 10% of the manuscripts in the Q1-journals. No mismatch between both sets of classifications ~~occured~~occurred.

References

Albers, W., Boon, P. C., & Kallenberg, W. C. M. (2000). Size and power of pretest procedures. *Annals of Statistics*, 28,~~:~~ 195-214. Retrieved from: http://www.jstor.org/stable/2673986.

Albers, W., Boon, P. C., & Kallenberg, W. C. M. (2001). Power gain by pre-testing? *Statistics & Decisions*, 19(3), 254-276.

American Psychological Association (2010). *Publication Manual of the American Psychological Association* (Sixth Edition). Washington D.C.: American Psychological Association.

Bakker, M. & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, 43, 666-678, doi:10.3758/s13428-011-0089-5

Chambers, J. M., Cleveland, W. S., Kleiner, B., Tukey, P. A. (1983). *Graphical Methods for Data Analysis.* Pacific Grove, CA: Wadsworth & Brooks/Cole

Cohen, J., Cohen, P., West, S.G., Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (Third Edition). New York, NY: Routledge.

Chatterjee, S. & Hadi, A. S. (2006). *Regression Analysis by Example, Fourth Edition.* Hoboken, NJ: ~~-~~John Wiley & Sons.

Hoekstra, R., Kiers, H. A. L., & Johnson, A. L. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, ~~Vol.~~ 3, ~~, No.~~ 137, doi:10.3389/fpsych.2012.00137

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J~~.(~~. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21, 1157–1164. doi:10.3758/s13423-013-0572-3/21.12/2014

Lomax, R.G., & Hahs-Vaughn, D. L. (2012). Statistical Concepts: A Second Course. New York, NY: Routhledge.

Formatted: Font: Italic

Miles, J. & Shevlin, M. (2001). Applying Regression and Correlation: A Guide for Students and Researchers. London, UK: Sage.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. PLoS Medicine 6(6): e1000097. doi:10.1371/journal.pmed1000097/21.12/2014

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis (5th edition). New York: John Wiley & Sons.

~~(5th edition). New York: John Wiley & Sons.~~

Osborne, J. & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*, 8(2), 1-9. Retrieved from: http://PAREonline.net/getvn.asp?v=8&n=2/16.12/2014

Razali, N. & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1): 21-33.

SCImago. (2007). *SJR — SCImago Journal & Country Rank*. http://www.scimagojr.com. Retrieved on: 16.12.2014.

Tabachnick, B. G., & Fidell, L. S. (2013). Using multivariate statistics, 6th edition. London, UK: Pearson.

Wicherts, J. M., Bakker, M. & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLOS One*, 6(11), doi: 10.1371/journal.pone.0026828

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:10.2307/1912934.

Wilkinson, L. & Task Force on Statistical Inference (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist*, Vol. 54, No. 8, 594–604

Williams, M. N., Grajales, C. A . G. G., & Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research & Evaluation*, ~~Vol. ~~18(11)~~, No. 11~~, 1-14.

Tables

Label	*Journal*
Q1.1	Annual Review of Clinical Psychology
Q1.2	Clinical Psychology Review
Q1.3	Journal of Consulting and Clinical Psychology
Q2.1	International Psychogeriatrics
Q2.2	Journal of Attention Disorders
Q2.3	American Journal of Drug and Alcohol Abuse
Q3.1	Zeitschrift fur Klinische Psychologie und Psychotherapie
Q3.2	Journal of Obsessive-Compulsive and Related Disorders
Q3.3	International Journal of Psychology and Psychological Therapy
Q4.1	Internet Journal of Mental Health
Q4.2	Indian Journal of Psychological Medicine
Q4.3	Behaviour Change

Table 1: Selection of Clinical Psychology Journals. The first column gives the ranking of the journal, the first number denoting the quartile in which the journal falls, the second number the rank of the journal within that quartile.

Class.	Reason
Papers without a linear regression model:	
1	No Model of Interest
2	Rejection of linear regression on basis of correct assumptions
3	Rejection of linear regression on basis of not meeting incorrect assumptions
Papers with a linear regression model:	
4	Correct linear regression
5	Mentioned all correct assumptions but not if the 'normality assumption' was tested on the residuals or on *X* or *Y*
6	Did not test all but some correct assumptions, included neither normality of variables nor ~~residuals~~errors
7	Use of linear regression but no indication if any or which assumptions were tested
8	Assumed/tested normally distributed *X* but not the normality of the errors~~residuals~~
9	Assumed/tested normally distributed *Y* but not the normality of the errors~~residuals~~
10	Assumed/tested normally distributed *X* and *Y* but not the normality of the errors~~residuals~~
11	Assumed/tested normally distributed variables but did not indicate if *X* or *Y* or both and did not test the normality of the errors~~residuals~~
12	Other misconceptions about assumptions

Table 2: Classification of the reviewed regression papers. Rubrics 3 and 5 – 12 represent papers with imperfect handling of regression assumptions: in rubrics 5 – 7 it is unclear from whether assumptions are correctly dealt with; in rubrics 8 – 12 the dealing with assumptions was incorrect.

Journal	*Number of papers (~~col~~rub. 1–12)*	Number of papers with regression (rub~~col~~. 4–12)	Dealing with assumptions			No regression	
			Correctly (rub~~col~~. 4)	Unclear (rub~~col~~. 5–7)	Incorrectly~~Wrong~~ (rub~~col~~. 8–12)	Correct (violation of true assump-tion) (rub~~col~~. 2)	Incorrect (violation of false assump-tion) (rub~~col~~. 3)
Q1.1	33	0	0	0	0	0	0
Q1.2	86	6 (7%)	0	6 (100%)	0	0	0
Q1.3	98	26 (28%)	0	25 (100%)	0	3 (100%)	0
Q2.1	227	44 (19%)	3 (7%)	39 (89%)	2 (5%)	1 (100%)	0
Q2.2	199	52 (26%)	0	49 (94%)	3 (6%)	0	0
Q2.3	54	14 (26%)	0	14(100%)	0	0	0
Q3.1	23	5 (22%)	0	5 (100%)	0	1 (50%)	1 (50%)
Q3.2	59	21 (55%)	0	16 (71%)	5 (29%)	1 (100%)	0
Q3.3*	10*	2 (20%)*	0*	2 (100%)*	0*	0*	0*
Q4.1	2	1 (50%)	0	1 (100%)	0	0	0
Q4.2	82	0	0	0	0	0	0
Q4.3	20	2 (10%)	0	2 (100%)	0	0	0
Total	893	172 (19 %)	3 (2%)	159 (92%)	10 (6%)	6 (86%)	1 (14%)

Table 3: Proportion of various types of papers in our selected journals. Categorisations are mutually exclusive and exhaustive. Journals are referred by the labels assigned in Table 1. "~~Col~~Rub." refers to the ~~columns~~ rubrics in Table 2~~ that are included in the rubrics~~. The online supplementary material indicates which papers belong to each of the numbers in this table.

* Papers in Spanish excluded

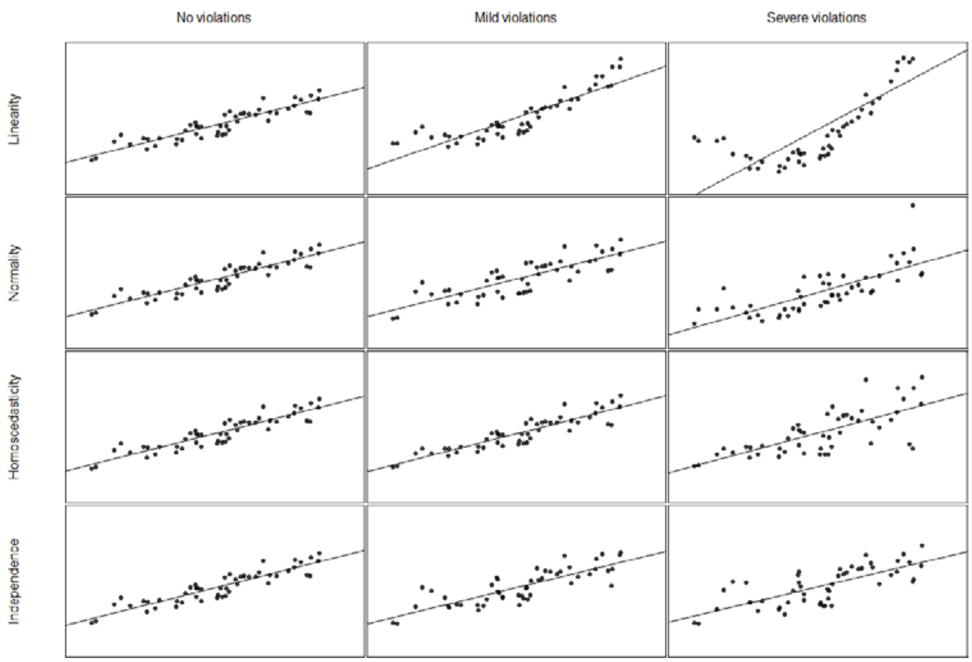
Journal	Articles with flawed linear regression model (rub. 8-12)	*Tested normality of X but not of residuals (rub. 8)*	*Tested normality of Y but not of residuals (rub. 9)*	*Assuming normally distributed variables but did not indicate if X or Y or both (rub.10)*	*Tested normality of X and of Y but not of residuals (rub. 11)*	Other misconceptions (rub. 12)
Q2.1	2	0	0		0	2 (100%)
Q2.2	3	2 (67%)	0	0	0	1 (33%)
Q3.2	5	4 (80%)	1 (20%)	0	0	0
Total	10	6 (60%)	1 (10%)	0	0	3 (30%)

Table 4: Breakdown of the types of mistakes that were observed. Only Journals with flawed models are listed. Categorizations are mutually exclusive and exhaustive. Journals are referred by the labels assigned in Table 1.

Journal	Papers in which handling of regression assumption was unclear *(rub. 5-7)*	Unclear		
		if the 'normality assumption' was tested on the residuals or on *X* or *Y* *(rub. 5)*	Did not test all but some correct assumptions *(rub. 6)*	no indication if any or which assumptions were tested *(rub. 7)*
Q1.2	6	0	2 (33%)	4 (67%)
Q1.3	26	0	0	25 (100%)
Q2.1	39	4 (10%)	5 (13%)	30 (77%)
Q2.2	49	1 (2%)	2 (4%)	46 (94%)
Q2.3	14	0	1 (7%)	13 (93%)
Q3.1	5	0	0	5 (100%)
Q3.2	16	0	0	16 (100%)
Q3.3	2	0	0	2 (100%)
Q4.1	1	0	0	1 (100%)
Q4.3	2	0	0	2(100%)
Total	159	5 (3%)	10 (6%)	144 (91%)

Table 5: Breakdown of the different types of 'Unclear' classifications. Only Journals with unclear models are listed. Categorizations are mutually exclusive and exhaustive. Journals are referred by the labels assigned in Table 1.

Figures



~~*Figure 1*. Visualisation of violations of the assumptions. From left to right, the columns indicate no violation (hence, the four figures in column 1 are the same), mild violation and severe violation of the assumptions listed in the rows. (As the scale of measurement is irrelevant for the visualisation, axis labels are omitted. The independent variable is plotted against the horizontal axis, the dependent variable against the vertical axis.)~~

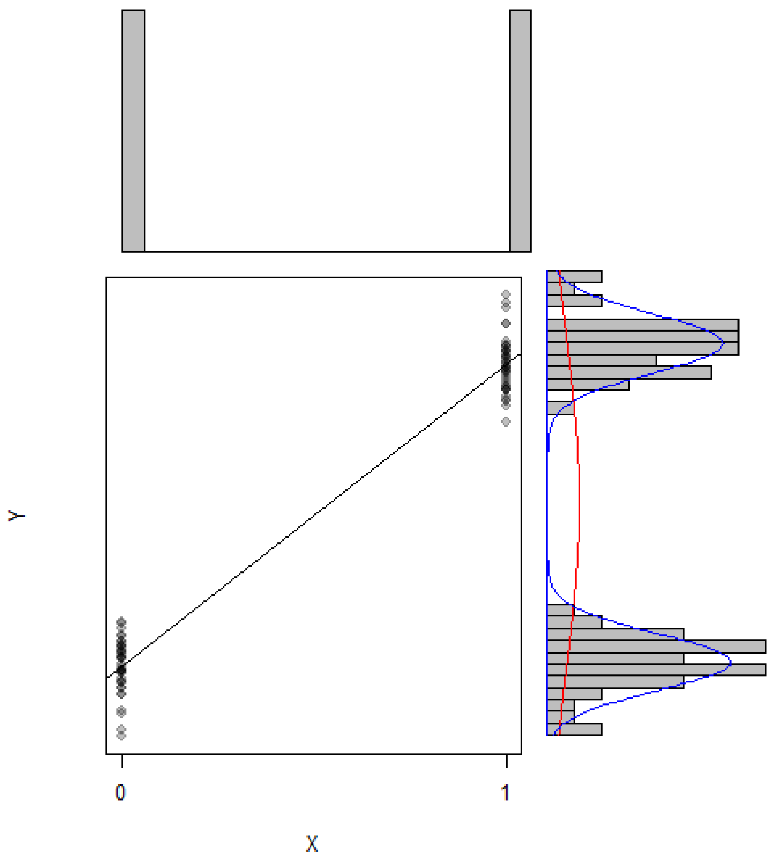
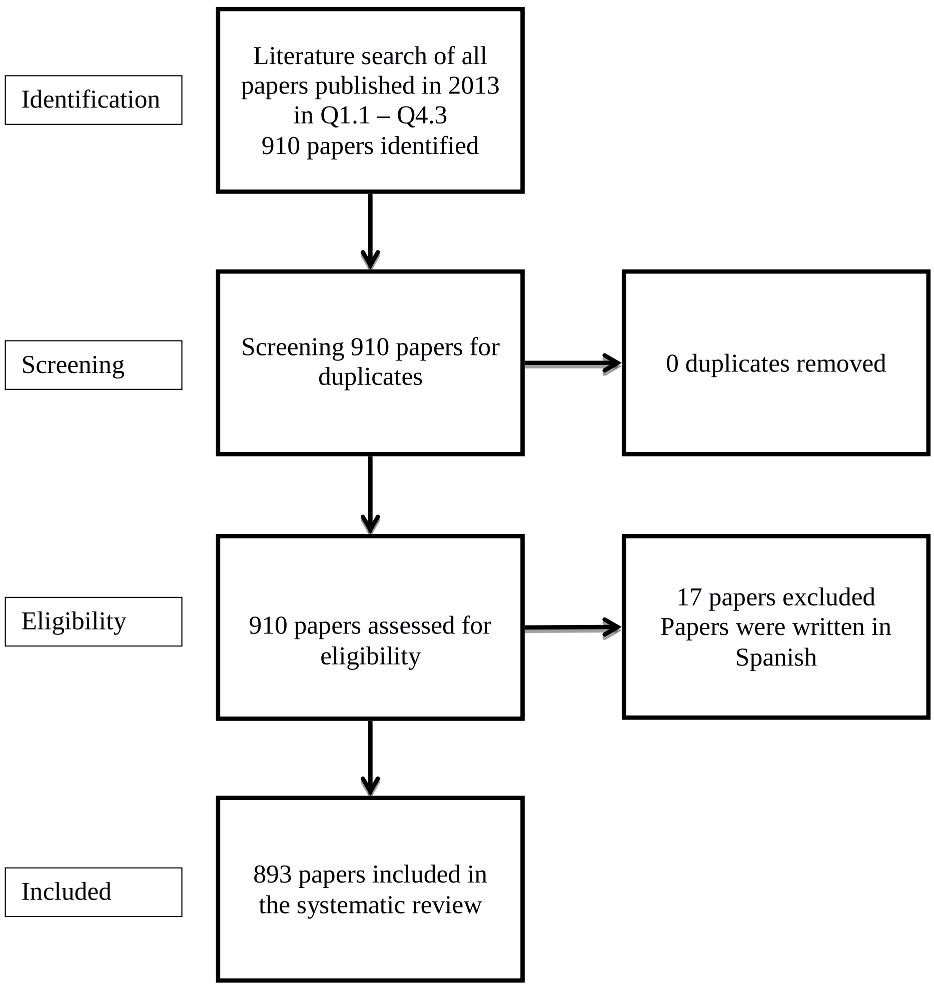


Figure 1~~2~~: Simulated example of a *t*-test based on $n = 40$ observations per group and no violations of the assumptions. The main panel shows a scatterplot of (X, Y)-scores. The red curve corresponds to the best-fitting normal distribution for Y, where the blue curves correspond to the best-fitting normal distribution for both subpopulations of Y. The histograms in the top and side panels clearly indicate non-normality for X and Y. However, within both subpopulations the distribution is normal ~~(blue curves).~~



Formatted: Line spacing: single, Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Figure 2~~3~~: Prisma flow diagram of included records