Review for "Regression assumptions in clinical psychology research practice - A systematic review of common misconceptions"

Matt Williams

December 2016

1 General comments

This is a well-written article on an interesting and important research topic. I would certainly like to see this article published, although I do have some recommendations for how it can be improved. Many of my comments are minor and technical; for example, I have a particular interest in how the assumptions of regression can best be described narratively, so I've made a number of comments in this regard. However, the most substantial suggestion I have is to consider the content analysis of journal articles as an empirical study in its own right, and ensure that sufficient information is given about this content analysis.

I have provided detailed suggestions below. Please do be aware that most of these are not absolutely compulsory: The authors seem to have a good grasp of the relevant literature, and I wouldn't want to dictate their choices. I am happy to enter into a productive dialogue here - my interest is in helping the authors produce an excellent article, not in pushing a particular point of view. My suggestions are also quite voluminous, but please take that as a reflection of my personal style and of my strong interest in the specific area of research assumptions - it's not an indication that the article is of poor quality, which is by no means the case.

2 Basic reporting: Summary comments

- The article is generally written clearly.
- Relevant prior literature has been appropriately referenced.
- Figures are relevant.
- The submission seems like a self-contained appropriate unit of publication.
- Raw data has been made available. The raw data in this case are the ratings of the individual journal articles that were reviewed. This data

has been made available as online supplementary materials (as a very large pdf document). I do wonder whether providing this as a dataset more amenable to re-analysis (e.g., csv file or similar) might be helpful.

3 Experimental design: Summary comments

The method of the study seems largely appropriate and there are certainly no ethical problems.

4 Validity of the findings: Summary comments

The conclusions are appropriately stated and supported.

5 Specific suggestions

- Page 2 Line 11: When you say the journals were "representative", of which population are they representative?
- Page 3 Line 7: We do almost always assume that the X variables are measured without error, but don't necessarily have to assume they're fixed. See Montgomery et al. (2012).
- Page 3 Line 9: Do clarify the difference between errors and residuals for the reader and try to use the correct term for a given instance throughout the article - often you say "residual" when you may mean "error". Note:
  - An **error** is the difference between an observation on Y and the value predicted by the "true" regression model (for the whole population)
  - An **error term** is the distribution of errors across repeated sampling for each of the N sampling units. There are thus N error terms, each of which has a distribution, albeit that in OLS the N error terms are assumed to be identically distributed. The distributions of the error terms are what we make assumptions about).
  - A **residual** is the difference between an observation on Y and the value predicted by the model estimated using the sample of data at hand. We don't make assumptions about the residuals, but we use them to give us a rough idea of how the error terms behave.
- Page 3 Line 10: It is true that the errors are assumed uncorrelated with the values of X, but typically we assume that the error terms have conditional mean zero for any combination of values on the X variables, which is a stronger assumption and implies that the errors will be uncorrelated with the X variables. More on this later.

- Page 3 Line 14: Ideally here you would unpack the idea of "proper inference" - what do we require these assumptions for? (Unbiasedness, consistency, efficiency, parameters having normal sampling distributions, etc.) See my 2013 article for some info on this, though it's not a definitive source. A discussion of consistency, efficiency etc. would be a good time to really emphasise the fact that some of these assumptions are much more important than others, and explain which assumptions are required to achieve which desirable properties of the OLS estimator. The error-normality assumption, for example, is much less important than the others.

- Page 3 Line 17: I wouldn't really encourage people to use Tabachnick and Fidell - they're popular in the social sciences but the accuracy of their material is pretty dubious. A couple of somewhat more reliable sources on regression would be Montgomery et al. (2001) or Fox (2008). The Cohen et al. text is technically imperfect but reasonably ok. I don't know Miles and Shevlin.

- Page 3 Line 19: Clarify that these are the assumptions of linear regression models when estimated using *ordinary least squares*. Different assumptions apply to linear regression models using different estimation methods (e.g., GLS).

- Page 3 Lines 22-24: Two comments on your discussion of the linearity assumption:

  - Your statement of the assumption isn't wrong per se but might leave the reader with the impression that non-linear relationships can't be modelled using OLS. (Polynomial models can be fit using OLS, for example). In fact, the idea that you can only model linear relationships using OLS is a bit of a popular misconception that is worth combating in this piece. The OLS model is linear in the parameters and linear in the estimator, but not necessarily linear in the variables.
  - More broadly, the "real" assumption here is that the conditional mean of the errors is zero for any given combination of values of the predictor variables (and where this assumption is breached, the OLS estimator becomes biased). Unmodelled non-linearity is the most obvious reason why this assumption might be breached, but it isn't the only one. Measurement error that is correlated across the IVs and DV can have the same effect; non-random sampling might lead to a systematic sampling bias toward selecting subjects with more positive (or more negative) error for some levels of the predictors; etc. I wouldn't necessarily dictate that you break with tradition by titling this assumption as something other than "linearity", but do try to flesh it out a bit more.

- Page 4 Lines 5-6: I find this wording of the normality assumption a little confusing. It would be sufficient to say simply that the errors are assumed

to be normally distributed for any given combination of values on the predictor variables.

- Page 4 Lines 9-10: I think the statement that "the theoretical model of regression is constructed based on this assumption" is a bit too vague. Try to ensure that here or at some other point that you clarify what the normality assumption is (and isn't) required for - e.g., it is not required for the OLS estimator to be unbiased, consistent and efficient, but it is required for the true Type I error rate and confidence interval coverage to be the same as their nominal values (albeit that OLS is extremely robust to breaches of this assumption).

- Page 4 Line 13: I do find the idea of "subpopulations" to be a slightly odd way of phrasing assumptions about conditional distributions. Here you could just say something like "the variance of the errors is the same for any combination of values on the X variables".

- Page 4 Lines 20-21: It is not correct to say that we assume that the residuals should be independent of the observations. This couldn't really be true - the residuals and DV observations are necessarily correlated. The trick to communicating the independence assumption well is to ensure you've very clearly introduced the idea of an error term having a distribution, and then say simply that the error terms are assumed independent.

- Page 5 Line 1: "there seems to be some pattern in the model residuals" is maybe too vague a description. The specific issue is the apparent presence of autocorrelation.

- Page 5 Lines 4-5: There are indeed many misconceptions about the regression model! You certainly don't need to discuss all of them, but off the top of my head some others are:

  - That OLS regression can only be used to model linear relationships (as discussed above).
  - That OLS assumes the presence of zero measurement error whatsoever (in reality, random measurement error in the Y variable is fine)
  - That measurement error in the X variables can only cause relationships to be underestimated, and not overestimated.
  - That OLS regression assumes a lack of multicollinearity. (It doesn't - albeit that perfect multicollinearity will result in the model being inestimable).

- Page 5 Line 10: Osborne and Waters has now been viewed more like 535,000 times - see the hit counter on the PARE website: http://pareonline.net/genpare.asp?wh=0abt

- Page 6 Lines 4-14: I would suggest removing this paragraph - unfortunately none of the cited benefits for checking IV normality really stack up. Specifically:

  - I don't think it's true that IV normality "can enhance prediction through the enhancement of linearity". I can't see any reason why a normal IV would enhance the linearity of the relationship. In fact, the IV distribution that gives the best assurance of linearity is an IV that is dichotomous rather than normal (since only a linear relationship is then possible).
  - I'm also not sure that there's a good basis to say that in general IV normality reduces the problem of influential points - this could be the case sometimes, but not always (and ultimately if influential points are a concern, then the best thing is to check for that directly).
  - It is definitely not true that normality of the predictor variables will result in the estimates remaining unbiased in the presence of measurement error. See the R code appendix below for a simulation showing this.

- Page 7 Line 2: I would be wary of saying that p-values and confidence intervals will be "biased". The term *bias* has a very specific meaning in statistics - I suspect you might mean something more general like "untrustworthy"?

- Page 7 Lines 13-14: I am similarly a bit wary of the claim that "the ordinary least squares method of linear regression is a more powerful procedure than any of its non-parametric counterparts". Whether OLS or a rank-based alternative is more powerful depends on the distribution of the errors. This is a complex issue but there's a good discussion here in the context of a t test (special case of OLS regression) vs a Mann Whitney test: http://stats.stackexchange.com/questions/71302/power-of-a-mann-whitney-test-compared-to-a-t-test

- Page 8 Line 12: The statement that "the APA manual (American Psychological Association, 2010) barely provides guidelines on this" sounds a little harsh, which might be justified, but I'm a little unsure. Is the APA style manual really the right place to look for guidance on statistical issues, given how diverse statistical analyses are in psychology? If you think the guidelines aren't optimatl, could you perhaps give a bit more detail about how you think they might be improved?

- Page 9 Line 2: Could you (at some point in the article) explain why you chose to focus on clinical psychology journals?

- Page 9 Line 4: What do you mean by "all" clinical psychology journals? All those listed in SCImago 2014?

- Page 9 Lines 6-7: Can you explain why you used this particular selection procedure? (Three highest ranked journals within each quartile)

- Page 9 Line 20: Saying that you checked whether "papers adhered to the spirit of the guidelines" of the APA is maybe a bit too vague a description for my comfort.

- Page 10 Lines 4-17: I think in general there needs to be more information here about the content analysis performed.
  - How and why did you decide on this particular 12-category rubric?
  - Can you provide more information about each of the 12 rubric categories? I.e., you've provided brief definitions of each, but did you use expanded definitions or descriptions of each category when coding the data?
  - Can you provide excerpts from papers falling into some of the categories to serve as illustrations?
  - "6 Did not test all but some correct assumptions, included neither normality of variables nor residuals" - I find this rubric category confusing - it seems to be double-barelled? I.e., what about a paper that didn't list "all" the assumptions, but did test the normality of the residuals? And what exactly counts as "all" the assumptions? There are a few different ways to frame the assumptions, after all.
  - What kind of misconceptions fell into the "12 Other misconceptions about assumptions" category?
  - It could be worth including information about inter-rater reliability here in the Method section rather than in an Additional Information section.
- Pages 11 and 18-20 Results tables: Your tables seem to be set up in such a way as to show lots of information about how the different journals compared to one another. This makes it harder to effectively disply other information. For example, I'd really like to have seen a table or figure simply showing how many articles fell into each of the 12 individual rubric categories, rather than just the over-arching super-categories. Your conclusions and discussion certainly don't emphasise the comparisons between journals very heavily, so I do wonder if re-jigging your tables and figures to de-emphasise this aspect and instead directly communicate the percentages in various rubric categories (across all journals) might be worthwhile?
- Page 11 Lines 18-20: I'm not sure about the statement "Of all papers that employed regression, 92% did not mention anything at all about the assumptions". Doesn't the 92% figure relate to the articles falling into categories 5-7 on the rubric, as noted in the previous sentence? Only rubric category 7 is consistent with the description of "did not mention anything at all about the assumptions".
- Page 12 Lines 6-9: I'd suggest re-wording the first two sentences of this paragraph. They're grammatically shaky and maybe a little more pejorative than is really warranted.
- Page 12 Lines 18-21: Here you are in effect assuming that the 8% of papers that provided information about assumption checks is representative of the

92% that didn't provide information. That's a very strong assumption, so do ensure that you acknowledge the great uncertainty surrounding your conclusion here.

- Page 13 Lines 1-2: It could be useful here to be more specific about which pieces of information you think researchers should include. Which specific residual plots do you think they should include when running OLS, if any? Which particular diagnostic statistical tests, if any?
- Page 13 Lines 11-12: Are you referring to transparency of assumption checks or about provision of open data here?
- Page 13 Line 17: I'm not sure that statisticians would agree that they know how to "correctly check" assumptions. There isn't really a correct way to check assumptions per se - just better and worse ways, with quite a bit of debate about the usefulness or various methods.
- Page 13 Lines 19-20: Re. "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." - Might statisticians not have participated in a good number of the studies that you did discuss? I wouldn't be surprised if they did. I do wonder if you're being a little overconfident here in statisticians' ability and willingness to fix errors on the part of their applied collaborators.
- Page 14 Line 4: Please explain the term "meta-regression" for the reader.
- Page 14 Lines 10-14: It's great that you're covering limitations, but here you're really only covering limitations of *scope*. What were the limitations of the **methods** you used?
- Page 14 Line 15: I'm not sure that "are valid" is quite the right term to use here. Perhaps this should be "should be relevant to"?
- General remarks on Discussion section:
  - It might be nice to include some directions for future research? Perhaps survey studies checking researchers' understanding of regression assumptions, if that hasn't been covered sufficiently already?
  - I totally understand that this project was intended to focus on misconceptions about regression assumptions, but the clearest finding is obviously that clinical psychology researchers don't seem to provide much information about assumption checks in their articles. (So we don't really know whether they labour under misconceptions or not). I do wonder if it's worth trying to tease out the consequences of this lack of reporting a bit more. For example, a key consequence of a lack of information about assumption breaches is that reports published in clinical psychology journals are likely to underestimate the quantity of uncertainty surrounding their claims. I.e., a confidence

interval might be a valid depiction of the magnitude of uncertainty surrounding a parameter if we knew that the OLS assumptions to hold true... but the presence of uncertainty about whether the assumptions hold true implies the presence of additional uncertainty about the parameter that is *not* captured by the confidence interval/uncertainty interval.

- Page 15 Lines 3-5: Perfect reliability of classification seems surprising? Are you sure the check was fully independent? I.e., did Casper know how Anja had classified each paper when he performed his classification checks?

- Page 19 Category "wrong" - I think "incorrectly" might be a better term to use here.

- Page 21 Figure 1: I'm not super keen on this figure. It's a neat idea to use a single type of display to show all the assumptions. Sorry to say it, but I suspect you might need to use different types of figures to display the different types of violations. My reasons for saying this:
  - Using scatterplots of DV on IV measurements is only useful for checking assumptions in the case of simple regression. With more than one IV, you can't use this method.
  - The scatterplots work well for showing violations of linearity and homoscedasticity, but are much less suitable for visualising normality and independence problems.

- Page 22 Figure 2:
  - Explain to the reader what the red and blue (density?) lines represent
  - The actual scatter points aren't visible on my end
  - Line 5 "The distribution is normal" should read "the distribution of Y is normal"

6 Recommendation

I would like to see this manuscript published, though some revisions could improve it. I would like to see a revised version before final acceptance. Please feel free to contact me if you need clarification of any of my comments.

7 Appendix: R code showing biased estimates with measurement error and normal X

```
get_coef = function(N = 30){
  X_true =  rnorm(30, 0, 1) #normal true X
  X_error = rnorm(30, 0, 0.5) #normal error in X
  X_obs = X_true + X_error
  #Observed X a combination of true score and error
  #both of which are normally distributed
  #hence X_obs is normal too.
  Y_error = rnorm(30)
  Y = X_true*0.75 + Y_error #true slope is 0.75
  lm(Y ~ X_obs)
  as.numeric(lm(Y ~ X_obs)$coefficients[2])
}

sims = replicate(1000, get_coef())
mean(sims) #lower than true slope of 0.75
```

The simulation shows that the estimates of the slope are biased downward below the true value in this case due to the presence of measurement error in X, even though X is normally distributed.