2021 ■ Vol. 17 ■ no. 1

How prevalent is overfitting of regression models? A survey of recent articles in three psychology journals

Lauren Dalicandro[a], Jane A. Harder[a], Dwight Mazmanian[a] & Bruce Weaver[a,b]

[a]Lakehead University
[b]Northern Ontario School of Medicine

Abstract ■ Since 2011, there has been much discussion and concern about a "replication crisis" in psychology. An inability to reproduce findings in new samples can undermine even basic tenets of psychology. Much attention has been paid to the following practices, which Bishop (2019) described as "the four horsemen of the reproducibility apocalypse": Publication bias, low statistical power, p-hacking (Simmons et al., 2011) and HARKing (i.e., hypothesizing after the results are known; Kerr, 1998). Another practice that has received less attention is overfitting of regression models. Babyak (2004) described overfitting as "capitalizing on the idiosyncratic characteristics of the sample at hand", and argued that it results in findings that "don't really exist in the population and hence will not replicate." The following common data-analytic practices increase the likelihood of model overfitting: Having too few observations (or events) per explanatory variable (OPV/EPV); automated algorithmic selection of variables; univariable pretesting of candidate predictor variables; categorization of quantitative variables; and sequential testing of multiple confounders. We reviewed 170 recent articles from three major psychology journals and found that 96 of them included at least one of the types of regression models Babyak (2004) discussed. We reviewed more fully those 96 articles and found that they reported 286 regression models. Regarding OPV/EPV, Babyak recommended 10 -15 as the minimum number needed to reduce the likelihood of overfitting. When we used the 10 OPV/EPV cut-off, 97 of the 286 models (33.9%) used at least one practice that leads to overfitting; and when we used 15 OPV/EPV as the cut-off, that number rose to 109 models (38.1%). The most frequently occurring practice that yields overfitted models was univariable pretesting of candidate predictor variables: It was found in 61 of the 286 models (21.3%). These findings suggest that overfitting of regression models remains a problem in psychology research, and that we must increase our efforts to educate researchers and students about this important issue.

Keywords ■ overfitting, replication crisis, regression, statistical best practices.

✉ ldalican@lakeheadu.ca

10.20982/tqmp.17.1.p001

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer.

Introduction

The so-called replication crisis in psychology started to gain attention in 2011. Arguably, a key factor was the publication of Bem's (2011) article in which he apparently demonstrated evidence of extrasensory perception (ESP) using common research practices. That claim prompted some serious scrutiny of common research practices in psychology. Some of the key papers that followed included those by Simmons et al.'s (2011), the Open Science Collaboration (2015), and Gelman and Loken (2014 – the "garden of forking paths" paper). In summarizing the main issues raised in those articles (and others), Bishop (2019) referred to the following problems as "the four horsemen of the reproducibility apocalypse": publication bias, low statistical power, p-hacking (Simmons et al., 2011) and HARKing (i.e., hypothesizing after the results are known; Kerr, 1998).

Another issue that has received less explicit attention in the psychological literature on replication is overfitting of (regression) models.[1] One exception is Babyak's (2004) article, "What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models". The following excerpt from Babyak (2004, p. 411) makes clear both his definition of overfitting and how it relates to the replication crisis:[2]

> In the present article, I will discuss a relatively narrow but important concept that has been considerably illuminated by simulation studies: the problem of capitalizing on the idiosyncratic characteristics of the sample at hand, also known as overfitting, in regression-type models. Overfitting yields overly optimistic model results: "findings" that appear in an overfitted model don't really exist in the population and hence will not replicate.

Babyak (2004) described several common data-analytic practices that increase the likelihood of overfitting :

- Having too few observations (or events) per variable
- Automated algorithmic selection of variables (e.g., stepwise regression)
- Univariable pretesting (or screening) of candidate predictor variables
- Categorization of quantitative variables (e.g., dichotomization)
- Sequential testing of multiple confounders

The best evidence that these practices produce overfitted models comes from simulation studies. Babyak (2004) explained the important role of simulation studies as follows.

> A statistical simulation study of modeling begins with a computer-generated population for which, much like knowing the correct diagnosis, the correct model is already known. The computer algorithm then simulates the activity of drawing a sample from the known population and conducting a regression model on the data from the sample. Because it is all performed on a computer, however, this act is repeated many thousands of times in a few seconds or minutes, each time using a newly drawn sample from the population (simulation studies often use 10,000 or more samples). The results from the many thousands of models are tallied and compared with the true population model. Most importantly, we can systematically manipulate various aspects of sampling and analytic activity.(p. 412)

He also noted that many simulation studies are "designed like a factorial experiment, systematically manipulating various aspects of sampling and analysis, such as the sample size, the shape of the distribution of x or y, or the presence of missing data or noise variables," and so on. Babyak (2004) himself conducted a small simulation study (see his Figure 2), but many other larger scale simulation studies have been done (e.g., Steyerberg, Eijkemans, & Habbema, 1999; Steyerberg, Eijkemans, Jr, E., & Habbema, 2001; Subramanian & Simon, 2013).

Regarding the first point above, Babyak (2004) cautioned that for ordinary least squares (OLS) models, one must have a minimum of 10-15 observations per variable (OPV) in the model if one wishes to reduce the likelihood of overfitting. For binary logistic regression models and survival models, on the other hand, the corresponding recommendation is 10-15 events-per-variable (EPV), where an event is the outcome variable category with the lower frequency of occurrence.

Before going on, we must emphasize that Babyak's (2004) guidelines recommending 10-15 OPV or EPV (depending on the type of model) are not about ensuring sufficient power to detect some minimally important effect size. Rather, they are only concerned with reducing the likelihood of model overfitting. Unless the minimally important effect size one wishes to detect is quite large, we suspect that the sample sizes determined via a priori power analysis will generally be greater than the minimal sample sizes required by Babyak's guidelines.

Using the lower limit of Babyak's (2004) 10-15 OPV guideline for OLS models yields a rule of thumb that is well known: One should have at least 10 observations per variable when fitting a linear regression model. This is sometimes described as "the rule of 10" (e.g., Norman & Streiner, 2014). Unfortunately, many research workers apply that same rule of thumb when estimating binary logistic regression or survival models. They appear to be unaware that for those types of models, it is the number of events, not observations, per variable that is important. For example, one of the authors remembers attending a thesis defense in which the main statistical model was a binary logistic regression model with 15 explanatory variables. There were 223 observations, but only 30 events. With 2 EPV, overfitting was virtually guaranteed. Neither the student nor the supervisory committee expressed any concern about overfitting, presumably because everyone

[1] In data science, machine learning, artificial intelligence, and related fields, there has been comparatively more discussion of how overfitting makes replication difficult. Interested readers can find many blog posts on this topic by doing a search on *X overfitting replication crisis*, replacing the X with the name of a particular field of research.

[2] Overfitting may be somewhat less problematic in purely predictive models that use regularization and shrinkage, but the majority of analyses in psychological research are explanatory in nature (Yarkoni & Westfall, 2017).

was applying the rule of 10 with respect to observations, not events, and there were nearly 15 observations-per-variable ($223/15 = 14.87$).

Unfortunately, such bad practices are not limited to thesis defenses: They can also be found in published articles that have undergone peer review. For example, Freedland, Reese, and Steinmeyer (2009) reviewed 60 randomly selected articles published in 2005 (in psychosomatic or behavioural medicine journals) and found that approximately 30% of them reported a limiting sample size that violated the rule of 10. They also found that 25% of models used univariable prescreening, 8% of models used automated algorithmic selection of variables (e.g., stepwise regression), and nearly half of the studies categorized quantitative variables prior to analysis. Only 1% of studies employed some kind of cross-validation.

Building from these results, the purpose of our study was to estimate the prevalence of overfitted regression models in recent issues of three high quality psychology journals. Of particular interest was the proportion of articles reporting regression-type models that show evidence of one or more of the bad practices Babyak (2004) described. We focused on those types of models because they are the types for which Babyak provided clear guidelines regarding OPV or EPV needed to reduce the likelihood of overfitting.

Method

Recent issues from three scientific, peer-reviewed journals (Personality and Individual Differences (PAID), the Journal of Personality and Social Psychology (JPSP), and Psychosomatic Medicine (PM) were selected for the present review. The former two journals were chosen based on their prominence in the field of psychology as well as their high scientific standards. Psychosomatic Medicine was also included, as it was the journal in which Babyak's (2004) article on overfitting was published. The three most recently published issues at the start of data collection (September 2019) from each of these journals were included in the review, for a total of nine issues. We included Issues 2-4 of Volume 117 for JPSP, Volumes 147-149 for PAID , and Issues 6-8 of Volume 81 for PM.

Reviewers began by recording each type of analysis or statistical model that was included in an article, with one row per analysis. A "continue" flag was set and more detailed information was collected only if the analysis used one of the following types of models:

- OLS regression (including between-subjects ANOVA and ANCOVA)
- Binary logistic regression
- Count regression models (e.g., Poisson or negative binomial regression)
- Survival models

As mentioned earlier, we focused on these types of models because they are the ones for which Babyak (2004) provided clear rules of thumb regarding OPV or EPV needed to reduce the likelihood of overfitting. For analyses that used one of the model types listed above, further information was recorded, as follows:

- The number of OPV or EPV (depending on type of model)
- Use of automated algorithmic variable selection
- Univariable pretesting candidate predictor variables
- Categorization of quantitative variables (e.g., dichotomization)
- Sequential testing of multiple confounders
- Whether overfitting was mentioned in the article
- Whether any type of cross-validation was used

For the first point above, reviewers recorded the actual number of OPV or EPV. For the remaining items, they entered 1 for Yes or 0 for No.

Results

As shown in Table 1, we reviewed 170 articles from three journals, and those articles reported 782 statistical analyses in total. We also show in Table 1 the number (and percentage) of articles reporting specific types of common analyses. Overall, the four most frequent types of analyses were OLS regression (40.6%), some kind of t-test (25.9%), ANOVA (25.3%), and mediation analysis (19.4%). Given the increasing popularity of multilevel modeling over recent years, we were somewhat surprised to see that only 9 articles (5.3%) reported using it.

The final row in Table 1 shows that we included 96 articles for more detailed analysis. The first row in Table 2 shows that we examined 286 models reported in those 96 articles (because they were one of the types for which Babyak provided clear guidelines). When Babyak's (2004) list of practices that produce overfitting are examined individually, bad practices do not appear as prevalent as we had anticipated: The percentage of models with specific bad practices range from 0.0% (for automated variable selection) to 21.3% (for univariable pretesting of candidate predictors). However, as the final two rows in Table 2 show, at least one bad practice was used in 33.9% of the models when using 10 OPV or EPV as the rule of thumb, and in 38.1% of articles when using 15 OPV or EPV.

When we categorize by model type rather than by journal, 32.5% of OLS models (89 of 274) and 66.7% of other models (8 of 12) included at least one bad practice using the 10 OPV or EPV as the rule of thumb (see Table 3). Similarly, 36.5% of OLS models (100 of 274), and 75.0% of other models (9 of 12) included at least one bad practice using the 15 OPV or EPV as the rule of thumb (see Table 3).

Table 1 ■ Number of articles plus numbers and types of analyses or models by journal

	Journal			
	JPSP	PAID	PM	Total
Total number of articles	30 (17.6%)	109 (64.1%)	31 (18.2%)	170 (100%)
Total number of analyses/models	265 (33.9%)	441 (56.4%)	76 (9.7%)	782 (100%)
Number of articles reporting:				
t-test (any type)	10 (33.3%)	32 (29.4%)	2 (6.5%)	44 (25.9%)
Chi-square test (any type)	2 (6.7%)	8 (7.3%)	4 (12.9%)	14 (8.2%)
Rank-based test/procedure (any type)	1 (3.3%)	4 (3.7%)	2 (6.5%)	7 (4.1%)
ANOVA	16 (53.3%)	23 (21.1%)	4 (12.9%)	43 (25.3%)
ANCOVA	0 (0.0%)	2 (1.8%)	2 (6.5%)	4 (2.4%)
MANOVA or MANCOVA	2 (6.7%)	7 (6.4%)	0 (0.0%)	9 (5.3%)
OLS regression (OLS)	10 (33.3%)	49 (45.9%)	10 (32.3%)	69 (40.6%)
Mediation analysis	9 (30.0%)	23 (21.1%)	1 (3.2%)	33 (19.4%)
Multilevel modeling	5 (16.7%)	2 (1.8%)	2 (6.5%)	9 (5.3%)
Binary logistic regression (BLR)	4 (13.3%)	5 (4.6%)	2 (6.5%)	11 (6.5%)
Multinomial or ordinal logistic regression	0 (0.0%)	2 (1.8%)	0 (0.0%)	2 (1.2%)
Count regression model	1 (3.3%)	2 (1.8%)	2 (6.5%)	5 (2.9%)
Survival model (SM)	0 (0.0%)	1 (0.9%)	5 (16.1%)	6 (3.5%)
Articles included for further analysis*	21 (70%)	59 (54.1%)	16 (51.6%)	96 (56.5%)

Note. *: Because they included at least one type of model for which Babyak (2004) provided guidlines.

Finally, as shown in Table 4, very few articles mentioned overfitting explicitly; and only a few more included some kind of cross-validation analysis. We intended to also report in Table 1 the number (and %) of articles that cited Babyak (2004). However, we found that none of the articles included for further analysis cited Babyak.

Discussion

Babyak (2004) described five common practices that lead to overfitting of regression-type models: 1) Having too few OPV or EPV; 2) Using some type of automated, algorithmic variable selection procedure; 3) Univariable pre-testing of predictor variables to include in a multivariable model; 4) Categorization of quantitative variables; and 5) Sequential testing of multiple confounders.

When we examined these practices individually, univariable pre-testing of candidate predictor variables was the most frequent bad practice and was found in 21.3% of the models we checked. This suggests that many researchers are still unaware of the "phantom degrees of freedom" problem Babyak (2004) described—namely, that each univariable model uses up one degree of freedom that is not necessarily apparent in the final model.

The next most frequent bad practice was having fewer than 15 OPV or EPV (depending on the type of model). This was observed for 10.5% of the models we assessed. When the slightly more lenient rule of thumb requiring 10 OPV or EPV was used, that percentage dropped to 5.6% of models. In our view, 10 OPV/EPV rule of thumb should be considered a bare minimum. We also remind readers (again) that these rules of thumb are about having sufficient sample size to reduce the likelihood of overfitting, not to guarantee sufficient power to detect the smallest effect size of interest (SESOI; Lakens, 2014). Sample size estimation is required for the latter.

Perhaps the most surprising result was that none of the models we examined were developed using automated variable selection (e.g., stepwise selection). It may be that the drawbacks to automated variable selection are finally becoming widely known in the research community in psychology. However, we must also consider the possibility that some authors may have used an algorithmic approach to variable selection but failed to report it. In hindsight, while reviewing the models, we should have also flagged cases where all variables in the final model were statistically significant. This excerpt from Frank Harrell's Author Checklist (discourse.datamethods.org/t/author-checklist/3407) explains why that is generally a sign that non-significant variables have been filtered out by some means.

> Unless the sample size is huge, this is usually the result of the authors using a stepwise variable selection or some other approach for filtering out "insignificant" variables. Hence the presence of a table of variables in which every variable is significant is usually the sign of a

Table 2 ■ Number of analyses, by journal, that used practices that tend to produce overfitted models

	Journal			
	JPSP	PAID	PM	Total
Number of models we assessed [1]	119	139	28	286
<10 OPV (or EPV) [2]	3 (2.5%)	8 (5.8%)	5 (17.9%)	16 (5.6%)
<15 OPV (or EPV)	6 (5.0%)	17 (12.2%)	7 (25%)	30 (10.5%)
Automated variable selection	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Univariable pretesting of predictors	3 (2.5%)	45 (32.4%)	13 (46.4%)	61 (21.3%)
Categorization of quantitative variable(s)	8 (6.7%)	8 (5.8%)	9 (32.1%)	25 (8.7%)
Multiple testing of confounders	0 (0.0%)	15 (10.8%)	7 (25.0%)	22 (7.7%)
At least one bad practice (using 10 OPV/EPV)	14 (11.8%)	62 (44.6%)	21 (75.0%)	97 (33.9%)
At least one bad practice (using 15 OPV/EPV)	17 (14.3%)	69 (49.6%)	23 (82.1%)	109 (38.1%)

Note. [1] We included only models for which Babyak (2004) provided recommendations concerning observations (or events) per variable. [2] OPV = observations-per-variable; EPV = events-per-variable. Percentages are (column) percentages of all models that were assessed.

Table 3 ■ Number of analyses, by model type, that used practices that tend to produce overfitted models.

	Type of Model		
	OLS [1]	Other [2]	Total
Number of models we assessed	274	12	286
<10 OPV (or EPV)	14 (5.1%)	2 (16.7%)	16 (5.6%)
<15 OPV (or EPV)	27 (9.9%)	3 (25%)	30 (10.5%)
Automated variable selection	0 (0.0%)	0 (0.0%)	0 (0.0%)
Univariable pretesting of predictors	55 (20.1%)	6 (50.0%)	61 (21.3%)
Categorization of quantitative variable(s)	19 (6.9%)	6 (50.0%)	25 (8.7%)
Multiple testing of confounders	22 (8.0%)	0 (0.0%)	22 (7.7%)
At least one bad practice (using 10 OPV/EPV)	89 (32.5%)	8 (66.7%)	97 (33.9%)
At least one bad practice (using 15 OPV/EPV)	100 (36.5%)	9 (75.0%)	109 (38.1%)

Note. [1] All models using OLS, including ANOVA, including ANOVA, ANCOVA, and linear regression. [2] Binary logistic regression models, count regression models, and survival models. Percentages are (column) percentages of all models that were assessed.

serious problem.

Finally, we note that none of the 96 articles included for more detailed review cited Babyak (2004), that only 4 of them (4.2%) mentioned the issue of overfitting, and that only 20 of them (20.8%) used some kind of cross-validation.

Our findings suggest that psychology researchers have made some strides in avoiding overfitting of their models, perhaps most notably by reducing their use of algorithmic variable selection methods such as stepwise selection. But given that approximately 35% of the models we examined included at least one bad practice that leads to overfitting, there is still work to do. We must increase our efforts to better educate researchers and students about the nature of model overfitting, how it relates to the replication crisis, and how to avoid it. Ten years ago Freedland et al. (2009) argued that researchers, statisticians, editors, and reviewers should increase their awareness of these issues and be attentive to the potential problems as they strive for "methodological excellence" (p. 213). Given the increased use of regression models in many areas of psychology, and the introduction of more advanced analytic modelling techniques, such calls for improvement are perhaps even more relevant than they ever were. Undergraduate psychology students taking statistics courses are routinely exposed to the perils of inflated Type I error rates associated with multiple comparisons or "fishing expeditions". Perhaps we should also strive to instil an awareness of the types of problems Babyak identified.

Authors' note

Co-first authors, LD and JH are presented in alphabetical order.

Table 4 ■ Number of articles mentioning overfitting, or using cross-validation

	Journal			
	JPSP	PAID	PM	Total
Articles mentioning overfitting	2 (9.5%)	0 (0.0%)	2 (12.5%)	4 (4.2%)
Articles using some kind of cross-validation	4 (19.0%)	14 (23.7%)	2 (12.5%)	20 (20.8%)

References

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*(3), 411–421. doi:10.1097/00006842-200405000-00021

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–437. doi:10.1037/a0021524

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature*, *568*(7753), 435–435. doi:10.1038/d41586-019-01307-2

Freedland, K. E., Reese, R. L., & Steinmeyer, B. C. (2009). Multivariable models in biobehavioral research. *Psychosomatic Medicine*, *71*(2), 205–216.

Gelman, A., & Loken, E. (2014). The statistical crisis in science: Data-dependent analysis–a" garden of forking paths"–explains why many statistically significant comparisons don't hold up. *American Scientist*, *102*(6), 460–466.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. doi:10.1207/s15327957pspr0203_4

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. doi:10.1002/ejsp.2023

Norman, G. R., & Streiner, D. L. (2014). *Biostatistics: The bare essentials (4th ed.)* Shelton, Connecticut: People's Medical Publishing House.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 344–344. doi:10.1126/science.aac4716

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. doi:10.1037/e519702015-014

Steyerberg, E. W., Eijkemans, M. J. C., & Habbema, J. D. F. (1999). Stepwise selection in small data sets: A simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology*, *52*(10), 935–942.

Steyerberg, E. W., Eijkemans, M. J. C., Jr, H., E., F., & Habbema, J. D. F. (2001). Prognostic modeling with logistic regression analysis: In search of a sensible strategy in small data sets. *Medical Decision Making*, *21*(1), 45–56.

Subramanian, J., & Simon, R. (2013). Overfitting in prediction models– is it a problem only in high dimensions? *Contemporary Clinical Trials*, *36*(2), 636–641.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Citation

Dalicandro, L., Harder, J. A., Mazmanian, D., & Weaver, B. (2021). How prevalent is overfitting of regression models? A survey of recent articles in three psychology journals. *The Quantitative Methods for Psychology*, *17*(1), 1–6. doi:10.20982/tqmp.17.1.p001

Received: 18/10/2020 ∼ Accepted: 18/01/2021