Journal of Memory and Language 72 (2014) 37–48

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

# What residualizing predictors in regression analyses does (and what it does *not* do) ☆

Lee H. Wurm [a,*], Sebastiano A. Fisicaro [b]

[a] Department of Psychology and Program in Linguistics, Wayne State University, Detroit, MI 48202, USA
[b] Department of Psychology, Wayne State University, Detroit, MI 48202, USA

ARTICLE INFO

*Article history:*
Received 24 May 2013
revision received 12 December 2013
Available online 23 January 2014

*Keywords:*
Regression analysis
Mixed-effects modeling
Statistical analysis
Collinearity
Residualization
Orthogonalization

ABSTRACT

Psycholinguists are making increasing use of regression analyses and mixed-effects modeling. In an attempt to deal with concerns about collinearity, a number of researchers orthogonalize predictor variables by residualizing (i.e., by regressing one predictor onto another, and using the residuals as a stand-in for the original predictor). In the current study, the effects of residualizing predictor variables are demonstrated and discussed using ordinary least-squares regression and mixed-effects models. Some of these effects are almost certainly not what the researcher intended and are probably highly undesirable. Most importantly, what residualizing does *not* do is change the result for the residualized variable, which many researchers probably will find surprising. Further, some analyses with residualized variables cannot be meaningfully interpreted. Hence, residualizing is not a useful remedy for collinearity.



## Introduction

In psycholinguistics there has been a move toward regression studies, which offer several advantages over traditional factorial designs. Baayen, Wurm, and Aycock (2007), for example, used mixed-effects modeling[1] to examine auditory and visual lexical decision and naming times. They found a number of curvilinear effects that are difficult to detect with factorial designs. Even more interesting, the authors found sequential dependencies in the response times, such that response latency on a given trial could be predicted by latencies on the previous four trials. This sequential dependency, which cannot be assessed in a factorial design, ultimately exhibited more explanatory power than nearly all of the other predictors that were examined.

A second advantage of regression designs is pragmatic. With the increased complexity of many theoretical models, it becomes impractical to isolate a difference on one predictor while adequately equating stimulus materials on the growing number of other variables known to affect psycholinguistic processing. Baayen et al. (2007) examined 18 predictor variables. The influential megastudy of Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) examined 19. A factorial design matching on all but one or two of the variables in situations like these is virtually inconceivable, and so a large number of potentially interesting studies simply could not be done. The Balota et al. (2004) study is interesting for the additional reason that they included as stimuli virtually all single-syllable monomorphemic words in English. An exhaustive study such as this cannot be done in a factorial manner, because the words in the language are naturally correlated on a number of variables of theoretical interest.

☆ Portions of this study were presented at the 54th annual meeting of the Psychonomic Society in Toronto, Ontario (November 14–17, 2013).

* Corresponding author. Address: Department of Psychology and Program in Linguistics, Wayne State University, 5057 Woodward, 7th floor, Detroit, MI 48202, USA. Fax: +1 313 577 7636.
*E-mail address:* lee.wurm@wayne.edu (L.H. Wurm).

[1] The distinction between ordinary least-squares regression and the kind of mixed-effects model described by Baayen, Davidson, and Bates (2008) is unimportant for the current study. Both techniques are used here, and the same findings and conclusions hold.

0749-596X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jml.2013.12.003

Many researchers express concern about the extent to which these natural correlations between predictors might lead to collinearity and computational problems. For example, Tabachnick and Fidell (2007) assert that, with predictor intercorrelations of .90 and above, there are statistical difficulties in the precision of estimation of regression coefficients (citing Fox, 1991). Further, Cohen, Cohen, West, and Aiken (2003) state that the estimates of the coefficients will be "very unreliable" and "of little or no use" (p. 390). In addition, Darlington (1990) emphasizes the loss of statistical power of tests on the individual regression slopes.

However, Friedman and Wall (2005) assert and demonstrate that improvements in algorithms and computer accuracy have eliminated the computational difficulties. The current study lends additional support to their claim. Further, Friedman and Wall (2005), along with others, also note that collinearity *per se* is not necessarily bad. For example, if a researcher's goal is simply to maximize explained variance, collinearity can be ignored (Darlington, 1990; Tabachnick & Fidell, 2007). The goal of most psycholinguistic applications of regression, though, is to evaluate the effects of several of the individual predictor variables. The potential interpretational problems caused by collinearity here can be thorny, even if the computational problems are not.

Because of concerns like this, some researchers have attempted to deal with collinearity by residualizing one of the correlated predictor variables. To do this, one runs a preliminary regression analysis using one of the predictor variables to predict the other (e.g., using $X_2$ to predict $X_1$). The residuals from this analysis constitute a new predictor variable, $X_{1\text{resid}}$, that is used in subsequent analyses in lieu of $X_1$. $X_{1\text{resid}}$ is guaranteed to be uncorrelated with $X_2$, providing an apparent solution to the problem of collinearity. Thus, residualizing seems like a useful and appropriate technique.

Psycholinguists have offered several justifications for residualizing. A review of some of those justifications is instructive, as it illustrates a considerable range of beliefs, some erroneous, about what residualizing accomplishes[2]:

> "To avoid problems with increased multicollinearity, we included the residuals…in our mixed-effects model…These residuals are thus corrected for the influence of all variables correlated with the original familiarity and meaningfulness measures" (Lemhöfer et al., 2008, p. 23)

> To dissociate the effect of one predictor from another and demonstrate that the effect of one predictor does not explain the effect of the other (Green, Kraemer, Fugelsang, Gray, & Dunbar, 2012, pp. 267–268)

> To help rule out the possibility that the effect of one predictor masks the effect of another (Kuperman, Bertram, & Baayen, 2010, p. 89)

> "…to assess the effect of " [a predictor] (Jaeger, 2010, p. 33)

> "…to ensure a true effect of" [a predictor] (Cohen-Goldberg, 2012, pp. 191–192)

> "…to allow for assessment of the respective contributions of each predictor" (Ambridge, Pine, & Rowland, 2012, p. 267)

> "…to determine the unique contribution of" [a predictor] (Cohen-Goldberg, 2012, p. 188)

> To provide "…a reliable estimate of the unique variance explained by each" [predictor] (Ambridge et al., 2012, p. 268)

> To pit predictors against one another and determine whether one explains variance that the other cannot (Ambridge et al., 2012, p. 268)

> "…to reliably assess effect *directions* for collinear predictors" and to be able to simultaneously assess "…the independent effects of multiple hypothesized mechanisms" (Jaeger, 2010, p. 30; emphasis in original)

> to test the effect of one predictor beyond the properties of two other predictors (Jaeger, 2010, p. 33)

> "Orthogonalisation of such variables is crucial for the accuracy of predictions of multiple regression models. Teasing collinear variables apart is also advisable for analytical clarity, as it affords better assessment of the independent contributions of predictors to the model's estimate of the dependent variable" (Kuperman, Bertram, & Baayen, 2008, p. 1098).

Most researchers do not specify precisely what would trigger the strategy. Cohen-Goldberg (2012) said it was done when a predictor "…was collinear with one or more control variables…" (p. 188). Jaeger and Snider (2013) did it "since the two predictor variables were correlated" (p. 63). Kahn and Arnold (2012) residualized "Because of high correlations" between the predictor variables (p. 317). This last case is interesting for the additional fact that the residualization was restricted to variables that were included only for purposes of statistical control. The individual effects of these variables were not of interest – the goal was simply to be able to assure readers that the analysis had controlled for them. Below, we show that residualizing accomplishes literally nothing in this case. Further, examination of the cut-off values that are reported reveals a lack of consensus about when one should residualize: Kuperman et al. (2008) residualized whenever a zero-order correlation between predictors exceeded 0.50, whereas Bürki and Gaskell (2012) used 0.30 as a cut-off.

Use of this strategy in psycholinguistics is a relatively recent phenomenon. The earliest example we have identified is Baayen, Feldman, and Schreuder (2006). The scope of what Baayen et al. (2006) did was restricted, and the reasons for it were principled and clearly articulated. They wanted to determine if a subjectively-rated version of word frequency offered anything beyond various objective measures. They partialed the objective measures from the subjective measure, and added the residuals to a model they had already specified as more or less complete. They did mention collinearity in this context, but it was not their

[2] One reviewer wondered if perhaps researchers were guilty of imprecise writing, rather than misunderstanding residualization. Evidence is presented later that there is genuine misunderstanding in at least some of these cases.

primary motivation. Indeed, in this study, they handled collinearity among their primary predictors in other ways.

Examples of residualizing can be found in at least a dozen papers published in three of the top journals in the field in 2012 (*Cognition; Journal of Experimental Psychology: Learning, Memory, and Cognition; Journal of Memory and Language*). Judging by the descriptions found in these studies, some of which were included above, there seems to have been significant drift in researchers' implementation of the strategy. Concerned that enthusiasm for the technique might be outpacing understanding of what it does, we decided to examine more closely exactly what is (and what is *not*) achieved by residualization of predictor variables. Our ultimate goal is to clear up some misconceptions and improve statistical practice in psycholinguistics.

## Study 1: Reanalysis of data from Lorch and Myers (1990)

Lorch and Myers (1990) presented a data set to illustrate a recommended way to analyze repeated-measures regression data. The DV was time to read a sentence. The predictor variables of theoretical interest were the number of words in the sentence (WORDS) and the number of new arguments in the sentence (NEWARGS). They also included an index of the serial position of each sentence in the experimental list. To make certain points clearer, we exclude this variable from analysis. The data set included seven sentences, each of which was read by 10 participants. Here, we reanalyze those data with and without residualization, showing that different results obtain depending on which variable is residualized.

### *Method*

We analyzed the data using a linear mixed-effects model with participant and sentence included as crossed random effects (Baayen et al., 2008). The DV in all analyses was reading time in seconds. Fixed effects included WORDS (either the original variable or residualized from NEWARGS) and/or NEWARGS (either the original variable or residualized from WORDS). We used version 2.11.1 of R (R Development Core Team, 2010) and version 1.0 of the *languageR* library (Baayen, 2010).

### *Results and discussion*

We begin by presenting the results of two linear mixed-effects analyses, one for each of the predictor variables in its original form, with no other predictors in the model. As can be seen in Table 1, each predictor has a significant effect when it is the sole predictor of reading times.

Both of these effects make sense. Sentences with more words require longer to read, and sentences with more new linguistic arguments also require longer to read. Note, however, that because these analyses include only one predictor each, the effect in each case reflects nothing more than that predictor's zero-order correlation with the DV.

A researcher might reasonably wonder whether the effect of NEWARGS holds when accounting for the number of words in the sentence, and whether the effect of WORDS holds when accounting for the number of new arguments in the sentence. The typical way to answer this type of question is to include both predictors in one simultaneous analysis. Table 2 presents the results of such an analysis. The statistical tests on the regression coefficients indicates that the effect of WORDS is statistically different from 0 and the effect of NEWARGS is not.

**Table 1**
Results of two linear mixed-effects analyses of reading time. In one, the predictor variable is the number of words. In the other it is the number of new arguments (data from Lorch & Myers, 1990). Neither analysis involved residualization.

| | *b* | *SE b* | *t* |
|---|---|---|---|
| Analysis 1: Predictor = WORDS | 0.437 | 0.090 | 4.857* |
| Analysis 2: Predictor = NEWARGS | 1.512 | 0.477 | 3.169* |

* $p < .05$.

Fig. 1 illustrates this situation with a Venn diagram. WORDS is assigned the variance represented by section a and NEWARGS is assigned the variance represented by section b. Neither predictor is given credit for the variance represented by section c because it is not uniquely attributable to either one. Contrast this with Fig. 2, which shows the situation when NEWARGS was the only predictor in the model. In this analysis, NEWARGS is assigned the variance represented by both sections b and c.

In Table 2, the standard error for NEWARGS has decreased from .477 to .424. Examination of the equation

**Table 2**
Results of linear mixed-effects analysis of reading time as a function of number of words and new arguments (data from Lorch & Myers, 1990). Neither predictor was residualized.

| Predictor | *b* | *SE b* | *t* |
|---|---|---|---|
| WORDS | 0.317 | 0.110 | 2.875* |
| NEWARGS | 0.664 | 0.424 | 1.566 |

* $p < .05$.

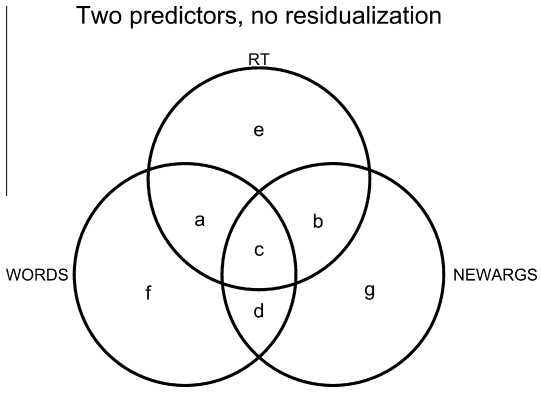


**Fig. 1.** Venn diagram illustrating the effects of WORDS and NEWARGS in a simultaneous analysis with no residualization. Section a represents the variance assigned to WORDS and section b represents the variance assigned to NEWARGS. The variance represented by section c is included in the $R^2$ calculations but is not assigned to either predictor variable because it is not uniquely attributable to either one.

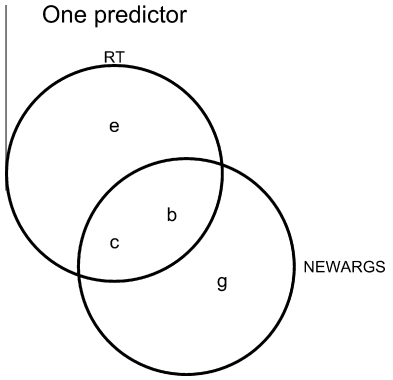


**Fig. 2.** Venn diagram illustrating the effect of NEWARGS when it is the only predictor in the model. Sections b and c represent the variance assigned to NEWARGS.

for the standard error shows why such a reduction is usually to be expected when moving from a one-predictor to a two-predictor model. When there is only one predictor, the standard error for an unstandardized regression coefficient is calculated as

$$SE_b = \sqrt{\frac{1-R^2_{YX}}{(N-2)}}\left(\frac{S_Y}{S_X}\right) = \sqrt{\frac{1-.6676}{(7-2)}}\left(\frac{1.906}{1.030}\right) = .477 \quad (1)$$

where $S$ stands for the standard deviation. When there are two predictors, the equation is

$$SE_{b_1} = \sqrt{\frac{1-R^2_{Y.12}}{(1-R^2_{1.2})(N-3)}}\left(\frac{S_Y}{S_1}\right) = \sqrt{\frac{1-.8916}{(1-.4838)(7-3)}}\left(\frac{1.906}{1.030}\right) = .424 \quad (2)$$

As is to be expected because linguistic arguments require words, these two predictor variables are positively correlated ($r = .696$). Psycholinguists seem to be increasingly expected to demonstrate that such intercorrelations do not contaminate their key results and render them unstable or somehow "incorrect" (e.g., Roland, Mauner, O'Meara, & Yun, 2012, p. 496). In the current example, we might decorrelate the predictors by residualizing the WORDS variable. We would regress WORDS on NEWARGS and save the unexplained portion of the variance in this analysis (the residuals) as a new predictor variable, $WORDS_{resid}$. Balling and Baayen (2012) say that a residualized variable created in this way "...can be straightforwardly understood as [the residualized predictor] in so far as this cannot be predicted from [the other predictor]" (p. 87).

Though it might seem like a minor quibble, we believe that this language can invite the interpretation that $WORDS_{resid}$ is somehow a purer or improved (or "corrected" – see Lemhöfer et al., 2008) measure of the number of words in a sentence. Lemhöfer et al. (2008) illustrate this danger when they change from calling a variable "residual meaningfulness" to "meaningfulness" in the same sentence: "Both residual familiarity and residual meaningfulness had significant facilitatory effects on RTs: familiarity ($\beta = -.0002$), $t(38{,}800) = -1.99$, $p < .05$; meaningfulness ($\beta = -.0001$), $t(38{,}800) = -2.38$, $p < .02$" (p. 23). This is perhaps understandable as shorthand presentation of a statistical result, but the next sentence from the main text says "Meaningfulness also interacted with participant group, with a stronger facilitatory effect of..." (p. 23). What they are calling meaningfulness is not meaningfulness. Similarly, in the current context, $WORDS_{resid}$ is not an improved, purified, or corrected version of WORDS; it is simply the errors of prediction with which one is left when predicting the number of words in a sentence from the number of new arguments in the sentence.

Table 3 shows the results of the analysis with $WORDS_{resid}$ and NEWARGS as predictors. Notice that the result for $WORDS_{resid}$ is identical to the result for the original WORDS variable in Table 2. This includes not only the coefficient, but the standard error as well (and thus the $t$ value and the significance level).

The fact that residualizing does not affect any aspect of the outcome for the residualized variable may come as a surprise to some researchers, as illustrated by much of the language reviewed in the Introduction. The following quote from Cohen-Goldberg (2012) provides a very typical example of the underlying logic while at the same time illustrating the result we have just shown:

> A significant inhibitory effect of similarity was found ($\beta = .02$; s.e. = .005; $t = 4.7$). Since onset–onset similarity was strongly correlated with sonority ($r = .78$), initial segment voicing ($r = .74$) and moderately correlated with letter similarity ($r = .41$), additional tests were performed to ensure a true effect of similarity... a significant inhibitory effect remained when similarity was residualized against sonority, initial segment voicing, and letter similarity ($\beta = .02$; s.e. = .005; $t = 4.7$) (pp. 191–192).

The statistical result for similarity is identical before and after it is residualized.

Ambridge et al. (2012) seem similarly unaware of this consequence of residualizing, and in fact misinterpret the outcome for one of their key theoretical variables – Pre-emption. Across three different age groups, the $\beta$s for Pre-emption changed from −.27, −.28, and .00 in single-predictor models to +.28, +.17, and +.60 in two-predictor models (their Table 5). Such sign changes are what many researchers hope to avoid by residualizing (e.g., Jaeger, 2010), but Ambridge et al. conclude that residualizing in fact *caused* them:

> ...the residualized pre-emption predictor is working in the opposite direction to that predicted...This most likely reflects a statistical quirk arising from the residu-

**Table 3**
Results of linear mixed-effects analysis of reading time as a function of number of words and new arguments (data from Lorch & Myers, 1990). WORDS was residualized.

| Predictor | *b* | *SE b* | *t* |
|---|---|---|---|
| $WORDS_{resid}$ | 0.317 | 0.110 | 2.875* |
| NEWARGS | 1.512 | 0.305 | 4.964* |

* $p < .05$.

alization process... Supporting this interpretation, note that for the older children and adults, the pre-emption predictor is in the predicted negative direction in the Pre-emption-only model, but the residualized version flips (though is not significant) in the Entrenchment + Pre-emption model (p. 271).

Their conclusion that residualizing is an appropriate strategy in light of these results is, to say the least, puzzling; but, in any event, they are incorrect about what caused the results. As just demonstrated, residualizing has no effect on the result for the residualized variable. The positive $\beta$s are what would have been observed in two-variable models even without residualization. What caused the changes in sign is not residualization, but moving from one-variable to two-variable statistical models. Study 2 (below, including the Extensions section) illuminates this issue further.

We turn next to the result for NEWARGS. Note that the coefficient (1.512) is the same as that observed in the one-variable model (Table 1). Fig. 3 illustrates why: Residualizing WORDS assigns section c to NEWARGS. Note also that the standard error (.305) is smaller than in Table 1. A reduction was expected based on Eqs. (1) and (2) above, but what is interesting is that the standard error is also considerably less than that observed in the two-variable model of Table 2. Eq. (3) shows why this happens. When the two predictors are uncorrelated, as they must be with residualization, Eq. (2) simplifies to

$$SE_{b_1} = \sqrt{\frac{1 - R^2_{Y.12}}{(N-3)}}\left(\frac{S_Y}{S_1}\right) = \sqrt{\frac{1-.8916}{(7-3)}}\left(\frac{1.906}{1.030}\right) = .305 \quad (3)$$

It is crucial to understand that this result for NEWARGS does not control for WORDS; it controls for $\text{WORDS}_{\text{resid}}$. That is, it controls for that part of WORDS which is independent of NEWARGS, which is to say that it controls for nothing.

An additional point is worth noting here: The outcome for $\text{WORDS}_{\text{resid}}$ is not only exactly the same as the result from the analysis using the original unresidualized WORDS variable (Table 2), but it is also exactly the same as would be obtained for the original unresidualized WORDS variable in a hierarchical (rather than simultaneous) analysis. In Table 4, we show the results of such an analysis. NEWARGS was entered at Step 1 and WORDS was entered at Step 2.

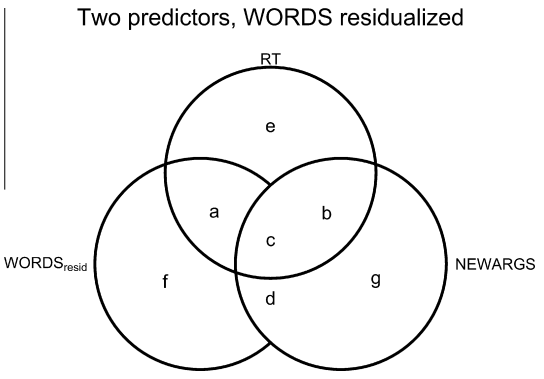


**Fig. 3.** Venn diagram illustrating the effects of $\text{WORDS}_{\text{resid}}$ and NEWARGS. Residualizing WORDS assigns section c to NEWARGS.

**Table 4**
Results of hierarchical linear mixed-effects analysis of reading time as a function of number of words and new arguments (data from Lorch & Myers, 1990). Neither predictor was residualized, but they were entered in two discrete steps.

| Predictor | *b* | *SE b* | *t* |
|---|---|---|---|
| Step 1 | | | |
| NEWARGS | 1.512 | 0.477 | 3.169* |
| Step 2 | | | |
| WORDS | 0.317 | 0.110 | 2.875* |

* $p < .05$.

**Table 5**
Results of linear mixed-effects analysis of reading time as a function of number of words and new arguments (data from Lorch & Myers, 1990). NEWARGS was residualized.

| Predictor | *b* | *SE b* | *t* |
|---|---|---|---|
| WORDS | 0.437 | 0.079 | 5.518* |
| $\text{NEWARGS}_{\text{resid}}$ | 0.664 | 0.424 | 1.566 |

* $p < .05$.

The result for NEWARGS is identical to that shown in Table 1, as it must be because Step 1 of the hierarchical analysis is identical to the one-predictor model. The WORDS result is identical to that seen in Table 2, and to the result for $\text{WORDS}_{\text{resid}}$ in Table 3. Thus, a simultaneous analysis with $\text{WORDS}_{\text{resid}}$ and NEWARGS produces the same coefficients as a hierarchical analysis with NEWARGS entered a step prior to WORDS. The difference is that the statistical significance of NEWARGS is exaggerated in the analysis using $\text{WORDS}_{\text{resid}}$ because of an artificially small standard error.

For the sake of completeness, Table 5 shows the results of a (simultaneous) analysis that includes the original version of WORDS and a residualized version of NEWARGS. The results are predictable from the foregoing demonstrations. The result for $\text{NEWARGS}_{\text{resid}}$ is identical to that involving the original, unresidualized NEWARGS in Table 2. The coefficient for WORDS (.437) is the same value observed in Table 1, when WORDS was the only predictor in the model, but its statistical importance is exaggerated here.

It is worth noting that the models in Tables 2–5 explain exactly the same proportion of variance and have identical values on all five of the model fit indices produced by the software (version 2.11.1 of R; R Development Core Team, 2010). Thus, the differences between the models concern only the individual coefficients and their associated statistical tests. We have seen that the results vary depending on whether (and on which predictor) residualization is done.

We conclude that the potential for misinterpretation of results involving residualized predictors is extremely high. Indeed, as our review of researchers' language in the Introduction suggested, some researchers using this approach either misunderstand what it does or do an inadequate job of describing their results. As another example of the

high potential for misunderstanding, one reviewer proposed residualizing both predictors against each other and using the two residualized variables in an analysis together in place of the original variables. Given a particular specific circumstance this might seem to be a reasonable course of action. However, the correlation between two variables residualized against each other will always have the same magnitude as that between the original variables, with the opposite sign. For example, in the Lorch and Myers (1990) data set, the correlation between WORDS and NEWARGS is .696, and the correlation between $WORDS_{resid}$ and $NEWARGS_{resid}$ is −.696. Such an analysis thus does not reduce collinearity, and worse, contains neither of the original predictors of interest.

What, then, should we consider the "true" effect of WORDS or of NEWARGS (as Cohen-Goldberg, 2012, put it)? It depends on what is meant by "true," but we argue that the simultaneous analysis with the original predictors (see Table 2) comes closest to what researchers generally want. According to Darlington (1990), one would say that the effect of WORDS, holding NEWARGS constant, is .317. The effect of NEWARGS, holding WORDS constant, is .664. In the language of Lorch and Myers (1990), the unique effect of WORDS adjusted for NEWARGS is .317 and the unique effect of NEWARGS adjusted for WORDS is .664. Baayen, del Prado, and Martín (2005) would say that the independent contribution of WORDS, while NEWARGS is held constant, is .317 and the independent contribution of NEWARGS, while WORDS is held constant, is .664. Cohen et al. (2003) would say that, for any given value of NEWARGS, the effect of WORDS is .317 and, for any given value of WORDS, the effect of NEWARGS is .664. Trying to apply this language to any table other than Table 2 quickly leads to tortured text and logic, and dramatically increases the likelihood that a researcher will produce an inappropriate or inaccurate description of the results. We are not advocating needlessly technical language, but rather, careful and clear descriptions. If one's text and logic are tortured, it is possible that a different analysis might have been more appropriate.

## Study 2: Simulated data

Friedman and Wall (2005) demonstrated that $r_{12}$ is merely one of the three pieces of information that determine the presence and extent of problems relating to collinearity. We now simulate data for additional analyses, incorporating all three pieces of information, with the goal of achieving a more systematic revelation and understanding of the underlying issues.

In a typical behavioral study of word recognition, a researcher might have each of 50 participants respond to each of 100 items. A traditional way of analyzing such data would be to calculate a mean reaction time (RT) for each of the items by averaging over participants, and using item-specific values on some variables to try to predict those RTs. There are more sophisticated ways to analyze such data that take their repeated-measures nature into account, such as the multi-level models used in Study 1. However, at present there is no agreed-upon method for calculating $R^2$ for those models, so for Study 2 we opt for the traditional method of analysis in order to demonstrate certain points more clearly.

For the simulations, five different values of predictor intercorrelation were examined: $\rho_{12} = -.50$, 0, .35, .75, and .95. The case where $\rho_{12} = 0$ illustrates the idealized situation in which interpretation of the individual coefficients is least ambiguous. Because we built a medium and a small effect into the data (details are provided below), each of the other four values of $\rho_{12}$ resides in a distinct region discussed in Friedman and Wall's (2005) very useful computational framework.

### Method

Data were simulated with the *mvrnorm* module of the MASS package (version 7.3-6; Venables & Ripley, 2002) and version 2.11.1 of R (R Development Core Team, 2010). Each call to *mvrnorm* produces a sample of a desired size (100 in this case, to simulate data for 100 imaginary items) from a specified multivariate normal distribution. As our starting point, we specified a covariance matrix in which $\rho_{12} = 0$, $\rho_{Y1} = .32$, and $\rho_{Y2} = .22$. That is, the predictor variables were uncorrelated, $X_1$ correlated .32 with $Y$, and $X_2$ correlated .22 with $Y$. On average these values produce a medium effect for $X_1$ (roughly 10% explained variance) and a small effect for $X_2$ (roughly 5% explained variance). After each of 10,000 calls to *mvrnorm*, the simulated data were analyzed with a linear model. $Y$ was the DV and $X_1$ and $X_2$ were predictors. All three variables were $z$-scores, so the resulting regression coefficients are standardized ($\beta$s) rather than unstandardized ($b$s). Several statistics were recorded from each analysis: $r_{12}$, $r_{Y1}$, $r_{Y2}$, $\beta_1$, $\beta_2$, $R^2$, and the $p$-values associated with the test on each $\beta$. An additional 40,000 calls were made to *mvrnorm*, 10,000 at each of the other values of $\rho_{12}$, and the data were analyzed in the same way.

### Results and discussion

Table 6 presents the results. As can be seen from the first column of Table 6, when predictor variables are uncorrelated, the regression analysis reflects the underlying correlation between the DV and each predictor. To two decimal places, mean $\beta_1$ = mean $r_{Y1}$ and mean $\beta_2$ = mean $r_{Y2}$. Mean adjusted $R^2 = .15$, which is the sum of the medium effect ($.32^2$) and the small effect ($.22^2$) that were built into the data. Power is quite acceptable for the medium effect and a bit less than the desired .80 for the small effect.

The second column shows the results when $\rho_{12} = -.50$. Each predictor has a positive relationship with the DV, but the predictors have a negative relationship with each other. This is in a region Friedman and Wall (2005) call Region I, Enhancement. Both $\beta_1$ and $\beta_2$ are greater than in the uncorrelated case, and $R^2 > r_{Y1}^2 + r_{Y2}^2$. This is one example of the kind of situation Hamilton (1987) had in mind (see also Region IV below) when cautioning against the conclusion that correlated variables are necessarily redundant. For both effects, the likelihood of statistical significance is nearly 1.

**Table 6**
Results of simulations ($N$ = 10,000 for each value of $\rho_{12}$). Neither $X_1$ nor $X_2$ was residualized.

| | $\rho_{12}$ | | | | |
|---|---|---|---|---|---|
| | 0 | −.50 | .35 | .75 | .95 |
| Mean $r_{12}$ | −0.001 | −0.498 | 0.347 | 0.747 | 0.950 |
| Mean $r_{Y1}$ | 0.317 | 0.319 | 0.319 | 0.316 | 0.319 |
| Mean $r_{Y2}$ | 0.220 | 0.220 | 0.218 | 0.217 | 0.219 |
| Mean $\beta_1$ | 0.318 | 0.572 | 0.277 | 0.351 | 1.135 |
| Mean $\beta_2$ | 0.220 | 0.506 | 0.122 | −0.046 | −0.860 |
| Mean $R^2$ | 0.165 | 0.306 | 0.132 | 0.118 | 0.189 |
| Mean adjusted $R^2$ | 0.148 | 0.291 | 0.114 | 0.100 | 0.173 |
| Power for effect of | | | | | |
| $X_1$ | 0.913 | >0.999 | 0.777 | 0.673 | 0.964 |
| $X_2$ | 0.646 | 0.998 | 0.222 | 0.060 | 0.818 |

The third column shows the results when $\rho_{12}$ = .35. This is in a region Friedman and Wall (2005) call Region II, Redundancy. Both $\beta_1$ and $\beta_2$ are smaller than in the uncorrelated case, and $R^2 < r_{Y1}^2 + r_{Y2}^2$. The likelihood of either effect reaching statistical significance is lower, substantially so in the case of the smaller effect. The Lorch and Myers (1990) data analyzed in Study 1 provide an example of Redundancy, and this would seem to be the region in which researchers most often find themselves. We would generally expect the correlation between predictors with the same kind of effect on the DV to be positive. The interpretation here is familiar and sensible: We would say that the effect of $X_2$ no longer holds when one takes $X_1$ into account, as we concluded in connection with Table 2 above.

The fourth column shows the results when $\rho_{12}$ = .75. This is in a region Friedman and Wall (2005) call Region III, Suppression. The main characteristics of this region are that the $\beta$ for the smaller effect has changed sign, and the $\beta$ for the larger effect is greater than in the ideal uncorrelated case. $R^2$ in this region is increasing from a minimum value but remains less than $r_{Y1}^2 + r_{Y2}^2$.[3] The likelihood of either effect reaching statistical significance is low (extremely so for the smaller effect). The likely conclusion here is the same as in Region II: The effect of $X_2$ no longer holds when one takes $X_1$ into account.

The last column shows the results when $\rho_{12}$ = .95. Friedman and Wall (2005) call this Region IV, Enhancement, as $R^2 > r_{Y1}^2 + r_{Y2}^2$. It is distinguished from Region I, also called Enhancement, by the changed sign of $\beta_2$. Both $\beta$s are becoming more extreme, and the power values indicate a strong likelihood that both effects, including the one with the changed sign, will be statistically significant.

The general boundaries of the regions are shown in Table 7. The lower bound of Region I and the upper bound of Region IV are the theoretical minimum and maximum values of $r_{12}$, respectively. Fig. 4 illustrates the behavior of the $\beta$s over the range of possible values of $r_{12}$ in the specific context of the effect sizes used here ($-.854 < r_{12} < .995$). The behavior of the $\beta$s in the figure is the straightforward result of the regression equation arriving at the optimal least-squares solution to the analytical problem.

A major part of Jaeger's (2010) motivation in residualizing appeared to relate to effects changing sign (i.e., Regions III and IV). One might wonder how common this is. For a data set like the Lorch and Myers (1990) example used in Study 1, $r_{12}$ needs to be above .903 for entry into Region III and above .995 for entry into Region IV. For a data set with perhaps more realistic effect sizes, like the one simulated here, $r_{12}$ needs to be above .688 for entry into Region III and above .934 for entry into Region IV. We doubt that researchers will find themselves in this territory all that often, unless they are using regression to adjudicate between two highly-correlated predictor variables. We assert that regression is not well-suited to this task. We return to this point below, under "Recommendations and conclusion."

However, there is a more important point. The regression model is not influenced by whether the sign of a coefficient makes sense given the researcher's theoretical model. A changed sign may be an indication that the variable under consideration is of lesser theoretical importance. The variable that changes sign will always have the smaller of the two correlations with the DV. It does not relate to the DV in the way theorized, but operates "as a measure of the sources of error" in the other predictor (Darlington, 1990, p. 155), whose effect is stronger. Put another way, the predictor whose sign has changed accounts for (or suppresses) a portion of the variance in the other predictor that is unrelated to the DV (Pandey & Elliott, 2010).

*What happens when $X_1$ is residualized?*

The simulated data were reanalyzed, using the original version of predictor $X_2$ but a residualized version of $X_1$. The results of these analyses are shown in Table 8.

A row-wise comparison of Tables 6 and 8 shows that residualization has affected some aspects of the results while leaving others unchanged. $r_{1\text{resid}2}$ is of course now 0 for all values of $\rho_{12}$, as it must be. Thus, we can now conceptualize the analytic situation as being at the $X = 0$ location on figures like Fig. 4, but we must remember that Fig. 4 is no longer the correct representation because residualizing has changed the multivariate correlational structure of the data. Not surprisingly, residualization of $X_1$ had no effect on $r_{Y2}$, but the zero-order correlation between the other predictor and the DV has changed. That is, $r_{Y1\text{resid}}$ does not equal $r_{Y1}$, because $X_{1\text{resid}}$ is not $X_1$ unless $r_{12} = 0$. Therefore, although we may find it comforting that this analysis will produce "true" $\beta$s because the predictor intercorrelation has been set to 0, it must be remembered that those $\beta$s will be for variables that are not the original variables.

It is worth noting that residualizing $X_1$ had no effect whatsoever on the $\beta$ or on the likelihood of detecting an effect of what was originally $X_1$. This was to be expected given Study 1 but, given some of the language reviewed in the introduction, this outcome might surprise some authors. Residualizing $X_1$ also had no effect on any of the

[3] Discussion of the different definitions of suppression is beyond the scope of this paper. Here we use the capitalized word Suppression in referring specifically to Region III as defined by Friedman and Wall (2005). We use suppression (lower case) in the generic sense of a predictor's regression coefficient having a different sign than that predictor's zero-order correlation with the DV, as it is such sign changes that seem to call for either an explanation or countermeasures in psycholinguistic analyses (e.g., Ambridge et al., 2012; Jaeger, 2010).

**Table 7**
$r_{12}$ boundaries of the four regions discussed in Friedman and Wall (2005), assuming $r_{Y1} > r_{Y2} > 0$.

| Region | Lower bound | Upper bound |
|---|---|---|
| I, Enhancement | $r_{Y1}r_{Y2} - \sqrt{(1-r_{Y1}^2)(1-r_{Y2}^2)}$ | 0 |
| II, Redundancy | 0 | $\frac{r_{Y2}}{r_{Y1}}$ |
| III, Suppression | $\frac{r_{Y2}}{r_{Y1}}$ | $\frac{2r_{Y1}r_{Y2}}{r_{Y1}^2+r_{Y2}^2}$ |
| IV, Enhancement | $\frac{2r_{Y1}r_{Y2}}{r_{Y1}^2+r_{Y2}^2}$ | $r_{Y1}r_{Y2} + \sqrt{(1-r_{Y1}^2)(1-r_{Y2}^2)}$ |

**Table 8**
Reanalysis of data from Table 6 with $X_1$ residualized.

| | $\rho_{12}$ | | | | |
|---|---|---|---|---|---|
| | 0 | −.50 | .35 | .75 | .95 |
| Mean $r_{1resid2}$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Mean $r_{Y1resid}$ | 0.316 | 0.493 | 0.258 | 0.232 | 0.353 |
| Mean $r_{Y2}$ | 0.220 | 0.220 | 0.218 | 0.217 | 0.219 |
| Mean $\beta_{1resid}$ | 0.318 | 0.572 | 0.277 | 0.351 | 1.135 |
| Mean $\beta_2$ | 0.220 | 0.220 | 0.218 | 0.217 | 0.219 |
| Mean $R^2$ | 0.165 | 0.306 | 0.132 | 0.118 | 0.189 |
| Mean adjusted $R^2$ | 0.148 | 0.291 | 0.114 | 0.100 | 0.173 |
| Power for effect of | | | | | |
| $X_{1resid}$ | 0.913 | >0.999 | 0.777 | 0.673 | 0.964 |
| $X_2$ | 0.638 | 0.701 | 0.625 | 0.616 | 0.651 |

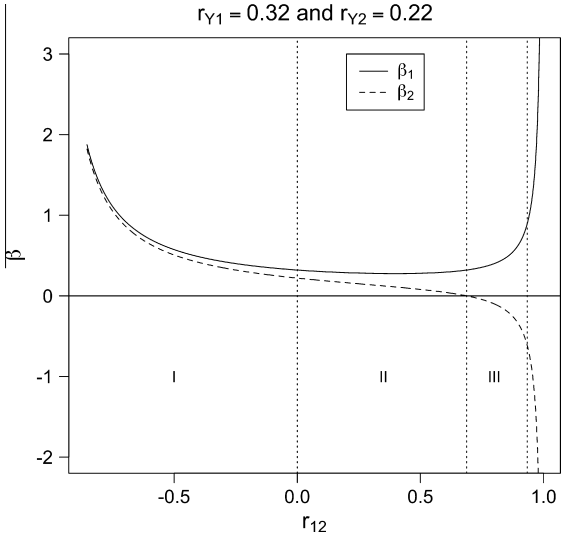


**Fig. 4.** $\beta$ values for $X_1$ and $X_2$ as a function of the correlation between the predictors ($r_{12}$), following Friedman and Wall (2005). $r_{Y1}$ is the effect built into the data for $X_1$. $r_{Y2}$ is the effect built into the data for $X_2$. Region I = Enhancement, as both $\beta_1$ and $\beta_2$ are greater than in the uncorrelated case, and $R^2 > r_{Y1}^2 + r_{Y2}^2$. Region II = Redundancy, as both $\beta_1$ and $\beta_2$ are smaller than in the uncorrelated case, and $R^2 < r_{Y1}^2 + r_{Y2}^2$. Region III = Suppression, as the $\beta$ for the smaller effect has changed sign, and the $\beta$ for the larger effect is greater than in the ideal uncorrelated case. $R^2$ remains less than $r_{Y1}^2 + r_{Y2}^2$. Region IV = Enhancement, as $R^2 > r_{Y1}^2 + r_{Y2}^2$. It is distinguished from Region I, also called Enhancement, by the changed sign of $\beta_2$. It begins at a value of $r_{12} = .934$, and is too small to label.

$R^2$ values. What differs, then, is how the total pool of explained variance is being assigned. As argued in Study 1, this is a function of what we have artificially done by giving $X_2$ first access to the variance. This point can also be seen by examination of the $\beta_2$ values. In all cases, they have the same value as $r_{Y2}$.

To our knowledge, no researcher has ever discussed the effect of residualizing $X_1$ on the likelihood of finding an effect of $X_2$, even though it can be fairly dramatic. In Regions I and IV (i.e., $\rho_{12} = -.50$ and .95), residualizing $X_1$ has *decreased* the probability of finding an effect of $X_2$ by .297 and .167, respectively. In Regions II and III residualizing $X_1$ has *increased* the probability of finding an effect of $X_2$ by .403 and .556, respectively. This is an unappealing state of affairs given the logic usually invoked for residualizing: A researcher wanting to know the "true" or "incremental" effect of $X_1$, over and above the effect of $X_2$, residualizes $X_1$. The result for $X_{1resid}$ is identical to what it would have been for $X_1$, but the procedure has had a dramatic effect on the likelihood of finding an effect of the *other* predictor, $X_2$.

*What happens when $X_2$ is residualized?*

The data from Table 6 were reanalyzed using the original version of predictor $X_1$ but a residualized version of $X_2$. The results of these analyses are shown in Table 9. In general, they are quite predictable from the foregoing analyses. Compared to the original analysis in Table 6, we can see that residualizing $X_2$ has had no effect on any of the $R^2$ values, or on $r_{Y1}$, or on the $\beta$, or on the likelihood of finding an effect of what was originally $X_2$. Where it did have an effect is in the zero-order correlation between the DV and what was originally $X_2$, and on $\beta_1$, and on the likelihood of finding an effect of $X_1$ (which now exceeds .90 in all cases).

The reappearance of the −.046 in the fourth column of the table is interesting. This is of course exactly the same value that was observed in Table 6 but, in the current analyses, it is not suppression. Residualization in fact changed $r_{Y2resid}$ to negative, so this $\beta$ does not represent a change in sign. We believe most researchers will consider this an unexpected and unwelcome outcome given that the effect built into these data was positive.

This last point underscores how much it matters which variable is residualized. If it is the one with the larger effect, the result will be two coefficients with no apparent suppression. If it is the one with the smaller effect, the result can be apparent suppression that is not really suppression. We believe that this situation captures what was happening in the data of Ambridge et al. (2012), which they dismissed as a "quirk arising from the residualization process" (p. 271).

**Table 9**
Reanalysis of data from Table 6 with $X_2$ residualized.

| | $\rho_{12}$ | | | | |
|---|---|---|---|---|---|
| | 0 | −.50 | .35 | .75 | .95 |
| Mean $r_{12\text{resid}}$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Mean $r_{Y1}$ | 0.317 | 0.319 | 0.319 | 0.316 | 0.319 |
| Mean $r_{Y2\text{resid}}$ | 0.219 | 0.436 | 0.114 | −0.031 | −0.267 |
| Mean $\beta_1$ | 0.317 | 0.319 | 0.319 | 0.316 | 0.319 |
| Mean $\beta_{2\text{resid}}$ | 0.220 | 0.506 | 0.122 | −0.046 | −0.860 |
| Mean $R^2$ | 0.165 | 0.306 | 0.132 | 0.118 | 0.189 |
| Mean adjusted $R^2$ | 0.148 | 0.291 | 0.114 | 0.100 | 0.173 |
| Power for effect of | | | | | |
| $X_1$ | 0.914 | 0.941 | 0.910 | 0.905 | 0.918 |
| $X_{2\text{resid}}$ | 0.646 | 0.998 | 0.222 | 0.060 | 0.818 |

Power for the non-residualized variable has been affected, as in the analyses in which $X_1$ was residualized. In Regions I and IV (i.e., $\rho_{12} = -.50$ and .95), residualizing $X_2$ has *decreased* the probability of finding an effect of $X_1$ by .058 and .046, respectively. In Regions II and III, residualizing $X_2$ has *increased* the probability of finding an effect of $X_1$ by .133 and .232, respectively. These are smaller than the corresponding numbers from the previous section because $X_2$ is the smaller of the two effects built into the data.

## Extensions

One might wonder whether our results scale up to more realistic data sets. We present three different pieces of evidence showing that they do. First, we mentioned above an analysis in which Cohen-Goldberg (2012) found an exactly identical result for a predictor before and after residualizing it. That analysis included more than 10,000 responses and had 19 predictor variables. The predictor in question had been residualized against three other predictors.

Second, we took an actual data set consisting of lexical decision times for 106 items, each of which was responded to by 88 subjects. Response errors were deleted, making the data array ragged (i.e., not all subjects are represented an equal number of times, and not all items are represented an equal number of times). We chose 10 numeric predictor variables, all measured on different scales. Their distributions deviated by varying degrees from normal; in some cases this deviation was substantial. Predictor intercorrelations ranged from −.66 to +.67. We added three more predictors consistent with a real-world analysis: Trial number, and two two-level factors (subject gender and voicing of the onset phoneme).

We used mixed-effects analyses with subjects and items as crossed random factors, the same type of analysis used in Study 1. Adopting a strategy cited in the Introduction, we residualized whenever a predictor correlation exceeded .50 in absolute magnitude (i.e., regardless of sign). There were five such correlations, involving seven of the 10 numeric predictors. $X_1$ was residualized against $X_6$. $X_9$ was residualized against $X_5$. $X_3$ was residualized against $X_5$, $X_7$, and $X_{12}$. We believe that the difficulties in interpretation noted above only get worse with each additional residualizing variable, but this is sometimes done (e.g., Kahn & Arnold, 2012). We make no defense of such a statistical model, but present it as an example of realistic complexity with real data.

We first ran the analysis with all of the original variables. Then, we substituted the three residualized versions for their original counterparts and re-ran the analysis. The results for the three residualized variables matched the results of the first analysis, in line with the findings presented above. The results for the residualizing variables ($X_5$, $X_6$, $X_7$, and $X_{12}$) changed, again in line with the findings above. Finally, the results for the remaining variables, uninvolved in any residualizations, remained unchanged.

Our third kind of evidence is conceptual. Consider the result for $X_3$ vs. $X_{3\text{resid}}$, for example. In the original analysis, $X_3$ was assigned the variance it could explain that no other predictor could. This exact same variance was assigned to $X_{3\text{resid}}$ in the second analysis: Variance that used to be explainable by $X_3$ and/or $X_5$ and/or $X_7$ and/or $X_{12}$ was expressly taken away from $X_3$, but $X_3$ was never given credit for that variance anyway. Thus, our findings do scale up as expected.

## General discussion

The current study has shown several of the effects of residualizing a predictor variable (assume $X_1$ here) in regression analyses. First and foremost, it produces an intercorrelation between predictors of 0, which was of course its desired effect. It is important to note that it does this by substituting a new predictor for one of the originals (e.g., $X_{1\text{resid}}$ for $X_1$). This has the concomitant effect of substituting $r_{Y1\text{resid}}$ for $r_{Y1}$. The difference between these two correlations depends on $r_{12}$ and can be dramatic.

Residualizing also gives the non-residualized predictor first access to the shared variance, which (conceptually) could be desirable and appropriate. Refusing to give a new predictor the same access to variance as more established predictors would seem to be a conservative approach. However, this creates an analysis that is neither simultaneous nor hierarchical in terms of the original variables, but which blends aspects of both. Specifically, residualizing exaggerates the statistical importance of the non-residualized predictor in a region of Redundancy or Suppression, and underestimates it in a region of Enhancement (as defined by Friedman & Wall, 2005). Finally, residualizing replaces the problem of collinearity (to the extent that it is a problem) with one that is less obvious and less well-understood. For these reasons, residualizing sometimes creates conceptual difficulty and leaves the researcher unable to draw any firm conclusions.

The current study has also demonstrated several things that residualizing does *not* do. Probably the most important is that it does not change the result for the predictor that was residualized. Further, residualizing (a) does not create an improved, purified, or corrected version of the original predictor, (b) does not change the overall explanatory power of the model, and (c) does not change any of the indices of model fit (AIC, BIC, log likelihood, etc.).

## Additional interpretational issues

Psycholinguists using regression have been concerned about statistical undercontrol (i.e., failing to take some important variable into account), but aside from collinearity concerns, they seem to have placed far less emphasis on the issue of statistical overcontrol (i.e., including too many predictor variables in a model). Meehl (1970) framed the conceptual consequences of this in terms of investigators interpreting counterfactual situations (e.g., a world in which written word frequency is uncorrelated with spoken word frequency) after having created a "virtual or idealized sample" (p. 401) that is given fictional values. He asserts that "When a social scientist of methodological bent tries to get clear about the meaning, proof, and truth of those counterfactuals that interpret statistical formalisms purporting to 'control the influence' of nuisance variables, he is disappointed to discover that the logicians are still in disagreement about just how to analyze counterfactuals" (p. 385; see also Campbell, Converse, & Rodgers, 1976). Anderson (1963) was succinct in making a similar point: "...one may well wonder exactly what it means to ask what the data would be like if they weren't what they are" (p. 170).

Breaugh (2006) presents an illustrative example built around the hypothesis that taller basketball players get more rebounds. Regression analyses using data downloaded from the website of the National Basketball Association suggest that the hypothesis is true only if players' weights are not controlled for. However, Breaugh questions whether a height variable from which weight has been residualized is even interpretable as anything. In the real world, these quantities are strongly correlated, so how are we to conceptualize this new variable? Breaugh (2006) says that "...making subjunctive statements based upon a residual variable is inappropriate. Simply stated, there is no basis to assume that, if in reality height and weight were uncorrelated, height would not be related to rebounds. Given they are correlated, and highly so, we simply have no way of knowing" (p. 439).

## Recommendations and conclusion

Some researchers consider mean-centering to be a viable alternative to the residualizing of predictors because they contend that it reduces collinearity (Kromrey & Foster-Johnson, 1998). However, the strategy is misguided because it ignores the crucial distinction between essential and non-essential collinearity (e.g., Dalal & Zicker, 2012). Mean-centering reduces non-essential collinearity, which is due to the way in which variables are scaled, but not essential collinearity, which is due to the underlying relationships between variables. As Pedhazur (1997) notes: "...centering $X$ in the case of essential collinearity does not reduce it, though it may mask it by affecting some of the indices used to diagnose it" (p. 306; see also Belsley, 1984). Mean-centering can sometimes facilitate interpretation of regression coefficients (but cf. Cohen, 1978; Kromrey & Foster-Johnson, 1998); but it does not reduce essential collinearity.

It might be worthwhile to investigate the effectiveness and appropriateness of techniques that have not been widely used in psycholinguistics but that have been developed to solve similar analytic issues. One such technique, random forests (Strobl, Malley, & Tutz, 2009) uses permutations of predictor variables to rank the importance of predictors. Many models are computed, with variables in their original and permuted forms. If the permuted versions of a variable lead to substantially worse models, that variable is assigned a relatively higher conditional importance. One disadvantage is that even with modern computers and algorithms, computation of the models can take many hours (Tagliamonte & Baayen, 2012).

Another alternative is ridge regression (Hoerl, 1962). In the context of high collinearity, ridge regression produces parameter estimates with less error variance than that seen with ordinary least-squares regression. The price a researcher pays is that the parameter estimate has a slight conservative bias. A far more important shortcoming for most psycholinguistic applications is that there is no way to use this in a repeated-measures design. One would need to collapse across items or participants.

A method of model comparison exists that formally tests whether a more complex model is statistically justified (i.e., whether the additional explained variance is worth increasing the complexity of a model). Cohen-Goldberg (2012) employs this method, as does Jaeger (2010), who notes that it is "robust against collinearity" (p. 37) and thus does not suffer from the reduced power associated with the $t$-test on a regression coefficient. Jaeger (2010) also points out the shortcoming of the method: The direction of the effect cannot be assessed.

Whatever statistical technique is chosen, researchers must be clear about what they wish to test. Language about "true" or "accurate" or "reliable" effects is probably meaningless without further elaboration. A review of the literature (see examples in the "Introduction") suggests that researchers' hypotheses in these situations are usually about the unique explanatory power of a predictor, beyond that of other predictors. When this is true, simultaneous multiple regression with the original predictors is the way to proceed (Breaugh, 2006; Lorch & Myers, 1990; Pedhazur, 1997) because this provides the basis for the appropriate interpretation of the resulting coefficients. As we have seen, these resulting coefficients need not reflect the zero-order correlation between any given independent variable and the dependent variable, because that is a different statistical question which is not addressed by multiple regression.

A hierarchical analysis might sometimes be preferable, insofar as it makes explicit the researcher's desire to know if a single additional predictor explains variance beyond that of an already-established model. We repeat, though, that the result of this analysis *as regards the last added predictor* is identical to the result of the simultaneous analysis (e.g., the result for WORDS in Tables 2 and 4).

Tabachnick and Fidell (2007) list several options for researchers concerned about collinearity: Ignore it (if the goal is simply to maximize $R^2$); eliminate one or more of the variables; make a composite variable (for example, by making ratios of different frequency measures, as was

done in Baayen et al., 2006 and Wurm, 2007); or subject the variables to a principle components analysis (as was done, for example, in Baayen et al., 2006). However, none of these is satisfactory if the researcher's goal is to evaluate the effects of one or more of the individual predictors. Models that differ by just one predictor can suggest dramatically different effects. For example, omitting a suppressor variable produces underestimates of the effect of $X$ on $Y$ (Cohen et al., 2003), and we have shown above that residualizing such a variable has differing effects depending on whether the variable resides in Region III or IV of Friedman and Wall's (2005) framework. Making a composite variable destroys any possibility of choosing between two similar variables for theory-building purposes, as does a principle component analysis.

Worries about suppression might be overblown, though. Darlington (1990) says "Suppression rarely occurs in real data" (p. 155), and Cohen et al. (2003) say that it is more likely to be seen in fields like economics where variables or actions often have simultaneous equilibrium-promoting effects. The computational framework of Friedman and Wall (2005) provides an easy way to see whether an analysis will produce a sign change (or any of the other possibilities discussed above): Assuming $r_{Y1} > r_{Y2} > 0$, the sign of the coefficient for $X_2$ will change if $r_{12}$ exceeds $r_{Y2}/r_{Y1}$. Above, we showed that these cut-off values were .903 for the Lorch and Myers (1990) data set, and .688 for the parameters used in Study 2. As reviewed in the Introduction, some researchers residualize at values of $r_{12}$ considerably smaller than this.[4] The larger point, and one of our main conclusions, is that to the extent that collinearity is a problem, residualizing does not solve it.

Researchers should understand that suppression does not indicate computational problems or model instability. Thus, one of the reasons given for residualizing, namely instability of computational estimates in the context of high collinearity, appears not to be valid. Friedman and Wall (2005) write (and demonstrate) that, because of advances in computational algorithms and accuracy, "multicollinearity does not affect standard errors of regression coefficients in ways previously taught" (p. 127).

Researchers might mean something different by "instability," though. It is true that in the presence of high collinearity relatively minor changes in the structure of a data set, even small differences due to random error in a replication study, can potentially reverse the order of importance of $X_1$ and $X_2$. The current study has shown that under some circumstances this could lead to opposite conclusions about whether a predictor's effect is facilitative or inhibitory. Pedhazur (1997) discusses these issues, but crucially, concludes that "none of the proposed methods of dealing with collinearity constitutes a cure. High collinearity is symptomatic of insufficient, or deficient, information, which no amount of data manipulation can rectify" (p. 318).

When psycholinguists encounter high collinearity, they should examine closely the reason(s) why. In such situations they may have to come to terms with the possibility that multiple regression is simply ill-suited to some of the purposes for which they would like to use it. Darlington (1990), amplifying earlier remarks (Darlington, 1968), wrote that it is a "misconception about collinearity. . .that more advanced statistical methods might someday eliminate the problem. But the problem is essentially that when two variables are highly correlated, it is harder to disentangle their effects than when the variables are independent. This is simply an unalterable fact of life" (p. 131; see also Breaugh, 2006; Meehl, 1970; Pedhazur, 1997). We are not as comfortable as Darlington in predicting what might be possible with future statistical techniques, but we have shown in the current study that residualization of predictor variables is not the hoped-for panacea.

[4] One reviewer noted that there are no such simple rules of thumb when one gets beyond two predictors. Deriving such cross-over points with several predictors is indeed complex, but it can be done if one wants them (e.g., Peters & Van Voorhis, 1935; Peters & Wykes, 1931a, 1931b). If it is simply the presence of a sign change that is of interest, it would be far easier to simply check whether the zero-order correlation between a predictor and the DV has a different sign than that predictor's regression coefficient.

## References


Ambridge, B., Pine, J. M., & Rowland, C. F. (2012). Semantics versus statistics in the retreat from locative overgeneralization errors. *Cognition, 123*, 260–279.

Anderson, N. H. (1963). Comparison of different populations: Resistance to extinction and transfer. *Psychological Review, 70*, 162–179.

Baayen, R. H. (2010). *languageR: Data sets and functions with Analyzing linguistic data: A practical introduction to statistics*. R package version 1.0. <http://CRAN.R-project.org/package=languageR>.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language, 55*, 290–313.

Baayen, R. H., del Prado, Moscoso., & Martín, F. (2005). Semantic density and past-tense formation in three Germanic languages. *Language, 81*, 666–698.

Baayen, R. H., Wurm, L. H., & Aycock, J. (2007). Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *The Mental Lexicon, 2*, 419–463.

Balling, L. W., & Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition, 125*, 80–106.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133*, 283–316.

Belsley, D. A. (1984). Demeaning conditioning diagnostics through centering. *The American Statistician, 38*, 73–77.

Breaugh, J. A. (2006). Rethinking the control of nuisance variables in theory testing. *Journal of Business and Psychology, 20*, 429–443.

Bürki, A., & Gaskell, M. G. (2012). Lexical representation of schwa words: Two mackerels, but only one salami. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 617–631.

Campbell, A., Converse, P. E., & Rodgers, W. L. (1976). *The quality of American life: Perceptions, evaluations, and satisfactions*. New York: Russell Sage Foundation.

Cohen, J. (1978). Partialed products are interactions; partialed powers are curve components. *Psychological Bulletin, 85*, 858–866.

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

Cohen-Goldberg, A. M. (2012). Phonological competition within the word: Evidence from the phoneme similarity effect in spoken production. *Journal of Memory and Language, 67*, 184–198.

Dalal, D. K., & Zicker, M. J. (2012). Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organizational Research Methods, 15*, 339–362.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin, 69*, 161–182.

Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill Publishing Company.

Friedman, L., & Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician, 59*, 127–136.

Fox, J. (1991). *Regression diagnostics*. Newbury Park, CA: Sage Publications.

Green, A. E., Kraemer, D. J. M., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2012). Neural correlates of creativity in analogical reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 264–272.

Hamilton, D. (1987). Sometimes $R^2 > r^2_{yx1} + r^2_{yx2}$: Correlated variables are not always redundant. *The American Statistician, 41*, 129–132.

Hoerl, A. E. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress, 58*, 54–59.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology, 61*, 23–62.

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition, 127*, 57–83.

Kahn, J. M., & Arnold, J. E. (2012). A processing-centered look at the contribution of givenness to durational reduction. *Journal of Memory and Language, 67*, 311–325.

Kromrey, J. D., & Foster-Johnson, L. (1998). Mean centering in moderated multiple regression: Much ado about nothing. *Educational and Psychological Measurement, 58*, 42–67.

Kuperman, V., Bertram, R., & Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes, 23*, 1089–1132.

Kuperman, V., Bertram, R., & Baayen, R. H. (2010). Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language, 62*, 83–97.

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 12–31.

Lorch, R. F., Jr., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 149–157.

Meehl, P. E. (1970). Nuisance variables and the ex post facto design. In M. Radner & S. Winokur (Eds.), *Analyses of theories and methods of physics and psychology* (pp. 373–402). Minneapolis, MN: University of Minnesota Press.

Pandey, S., & Elliott, W. (2010). Supressor variables in social work research: Ways to identify in multiple regression models. *Journal of the Society for Social Work and Research, 1*, 28–40.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research*. Forth Worth, TX: Harcourt Brace & Co.

Peters, C. C., & Van Voorhis, W. R. (1935). *Statistical procedures and their mathematical bases*. State College, PA: The Pennsylvania State College.

Peters, C. C., & Wykes, E. C. (1931a). Simplified methods for computing regression coefficients and multiple and partial correlations. *Journal of Educational Research, 24*, 44–52.

Peters, C. C., & Wykes, E. C. (1931b). Simplified methods for computing regression coefficients and partial and multiple correlations. *Journal of Educational Research, 23*, 383–393.

R Development Core Team (2010). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. <http://www.R-project.org>.

Roland, D., Mauner, G., O'Meara, C., & Yun, H. (2012). Discourse expectations and relative clause processing. *Journal of Memory and Language, 66*, 479–508.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*, 323–348.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson Education Inc.

Tagliamonte, S., & Baayen, R. H. (2012). Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change, 24*, 135–178.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.

Wurm, L. H. (2007). Danger and usefulness: An alternative framework for understanding rapid evaluation effects in perception? *Psychonomic Bulletin & Review, 14*, 1218–1225.