Evaluating Methods for Covariate Selection

Lauren Khoury

Kendra Wyant

Markus Brauer

John J. Curtin

2025-10-22

Abstract

Abstract of paper goes here and can span several lines.

# Introduction

Psychology researchers are often interested in experimental manipulations as a means of estabishing causal relationships between a variable of interest and a psychological outcome. For example, researchers may manipulate treatment conditions to determine whether a new psychotherapy intervention reduces anxiety symptoms.

The validity of causal interpretations depends on the prequisite of adequate statistical power (Cohen 1988, 1992). Statistical power is the probability of avoiding a Type II error, or failing to reject the null hypothesis when a true effect exists. The field has increasingly recognized the critical importance of adequately powered studies. Low-powered studies do not replicate well. They have a low probability of finding true effects and when an effect is detected, the magnitude of the effect is often inflated and the likelihood that the finding is a true positive is low (Button et al. 2013). These concerns have led to a replication crisis and have prompted increased caution and consideration around researcher decisions in statistical tests.

Adding covariates, variables measured at baseline prior to the manipulation, to statistical models has been shown to reduce Type II error by accounting for unexplained variance in the outcome that might otherwise be interpreted as noise. In the earlier example, where researchers test whether a new treatment reduces anxiety symptoms, they might consider including a measure of recent stressful life events as a covariate. Stressful events are expected to correlate with anxiety but, if measured after treatment assignment, would not be correlated with the manipulated variable thereby reducing variance in anxiety that is unrelated to the manipulation.

However, selecting which covariates to include is not straightforward. As a result, researchers have often iterated over their analyses, selectively adding covariates that improve the statistical significance (i.e., *p*-value) of their focal variable. This is a practice known as *p*-hacking (Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons 2014). It is now well-estabished that this is a statistically invalid method for covariate selection and leads to an increased Type I error rate (i.e., finding significant effects that do not truly exist; (Simmons, Nelson, and Simonsohn 2011; Ioannidis 2005)).

While this serves as an example of what not to do, there remains an important question on how covariates should be selected. In light of the serious consequences of *p*-hacking, as highlighted by Simmons, Nelson, and Simonsohn (2011) and others, one might conclude that perhaps no covariates should be used. Though, as noted earlier, this would increase the risk of Type II error. Another conclusion might be to use all available covariates that are reasonably believed to account for unexplained variance in the outcome. Each additional covariate, however, reduces degrees of freedom. In psychology, where potential covariates are numerous and often redundant, this can unnecessarily reduce degrees of freedom, again increasing the risk of Type II error.

It is clear, then, that covariate selection is essential for adequately powered analyses (i.e., those with low Type II error). However, criticially the method of covariate selection must not nominally inflate the Type I error rate, as occurs with *p*-hacking. This study aims to provide clear and accessible methods for researchers to select among a set of covariates. Specifically, we conducted 40,000 simulations of nine candidate methods across several research settings. These settings varied in population parameter, sample size, number of available covariates, proportion of good covariates, and the strength of relationships between good covariates and the outcome. We report Type I and Type II error rates for methods that use no covariates, all available covariates, a statistically invalid selection method (*p*-hacking), and three valid selection methods (Pearson correlation, full linear model, and least absolute shrinkage and selection operator [LASSO]) that were eached tested with and without controlling for the focal variable. These findings can help researchers determine the optimal covariate selection method for their specific research setting.

# Methods

## Covariate Selection Methods

We evaluated nine linear regression models with varying levels and methods of covariate selection. Two models did not use any covariate selection: a linear regression model that used no covariates and a linear regression model that used all available covariates. One model used a statistically invalid method of selecting covariates based on whether they lower the regression p-value (i.e., p-hacking). The remaining six models used three systematic covariate selection methods, selection based on the Pearson correlation coefficient (*r*), a full linear model, and LASSO, controlling for the focal manipulation (i.e., *X*) and without controlling for *X*. A summary of the nine models is presented in [Table 1](#tbl-methods).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: The nine linear regression models and their definition for covariate selection.   | Method | Definition | | --- | --- | | No covariates | Y is regressed on X without any covariates. | | All covariates | All available covariates are included in the regression model. | | P-hacking | Unsystematically adding covariates based on whether they lower the p-value of the main effect of X on Y. | | Bivariate correlation | Pearson correlation coefficient (r) of covariates on Y. Covariates are considered one at a time and included in the final model if they yield a significant effect on Y (p < .05). | | Partial correlation | Pearson correlation coefficient (r) of covariates on Y while controlling for X. Covariates are considered one at a time and included in the final model if they yield a significant effect on Y (p < .05). | | Full linear model | A full linear model that regresses Y on all available covariates. Covariates that have a statistically significant effect on Y (p < .05) are retained. | | Full linear model with X | A full linear model that regresses Y on all available covariates and X. Covariates that have a statistically significant effect on Y (p < .05) when controlling for X are retained. | | LASSO | A linear model that regresses Y on all available covariates and applies a penalty to shrink coefficients for less important covariates, potentially dropping them altogether (i.e., coefficient of 0). Covariates with non-zero coefficients were retained. | | LASSO with X | A linear model that regresses Y on all available covariates and applies a penalty to shrink coefficients for less important covariates. We assigned a 0 penalty to X to retain it in the model. Covariates with non-zero coefficients when controlling for X were retained. | |

Source: [Make Tables for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_tables-preview.html#cell-tbl-methods)

## Research Settings

We manipulated several variables designed to mimic varying research settings that psychology researchers might be working in. We crossed all levels of each variable to create a total of 540 unique research settings. A summary of these settings is presented in [Table 2](#tbl-dictionary). The variables include:

1. The true population parameter for X. We selected values that represent null (), medium () and large () effect sizes.
2. The sample size. We chose values for the number of observations that pertain to common sample sizes in experimental research: 50, 100, 150, 200, 300, and 400 observations.
3. The number of covariates available. We selected a wide range of possible scenarios: 4, 8, 12, 16, or 20 covariates.
4. The proportion of “good” covariates. We used varying proportions of good covariates across research settings (.25, .50, and .75) to represent a common reality when a researcher is faced with many covariates, but only some may be related to their outcome.
5. The strength of the relationship between the good covariates and Y. All good covariates were given a moderate and large relationship to *Y*, correlations of 0.3 and 0.5, respectively. Since these good covariates are all correlated with *Y*, it is likeley they are also correlated with each other. Therefore we assigned a moderate 0.3 correlation for all relationships among good covariates.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Research setting variables and values   | Research Setting Variable | Values | | --- | --- | | The population parameter for X | 0, 0.3, 0.5 | | The number of observations in the sample | 50, 100, 150, 200, 300, 400 | | The number of covariates | 4, 8, 12, 16, 20 | | The proportion of "good" covariates | 0.25, 0.50, 0.75 | | The correlation between Y and good covariates | 0.3, 0.5 | |

Source: [Make Tables for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_tables-preview.html#cell-tbl-dictionary)

## Data Analytic Plan

All data analyses were done in R (version 4.4.2). Simulations were run using high-throughput computing resources provided by the University of Wisconsin-Madison Center for High Throughput Computing (Center for High Throughput Computing 2006).

We ran 40,000 simulations for each research setting. Within each simulation we generated a unique datset that consisted of a dichotomous focal variable (*X*), varying numbers of quantitative covariates, where a subset are correlated with each other and with *Y* (see Research Settings section), and a quantitative outcome (*Y*) calculated by adding the *X* variable multiplied by the effect size (i.e., population parameter) to the *Y* generated from the correlation matrix with the covariates. We fit models from our nine methods on each simulated dataset. From each model we saved out the parameter estimate, standard error, and p-value for *X*. We also calculated true and false positive rates to evaluate the rate at which the model selected covariates correlated with *Y* (i.e., percentage of good covariates selected) and incorrectly selected covariates not correlated with *Y* (i.e., percentage of bad covariates selected).

For research settings where the population parameter for *X* is 0 (i.e., *X* has no effect on *Y*), we report the Type I error rate for each method both across and within research settings. For research settings where the population parameter for *X* is 0.3 or 0.5 (i.e., *X* has an effect on *Y*), we report the Type II error rate for each method across and within research settings. We also provide sampling distributions of the parameter estimate for *X* across research settings for each method, separately by true effect size (0, 0.3, and 0.5). Detailed tables of range and average error rates for each research setting and true and false positive rates for covariates are available in the supplement.

# Results

## Type I Error

Since this is simulation study, we were able to set the true population parameter for *X* to be zero (i.e., ). Therefore, any significant result found was a Type I error. [Figure 1](#fig-bar-1) shows the average Type I error rate across all research settings for each method. The p-hacking method for selecting covariates, unsurprisingly led to a highly inflated Type I error rate, consistent with the extant research published in the last several years on the connection between researcher degrees of freedom and false positives. Most other methods remained around the expected .05 threshold. There was some elevation in Type I errors for methods that controlled for *X* compared to those, however, it is not clear that these differences are substantial.

|  |
| --- |
| Figure 1: Type I Error by Method. Average type I error rate across all research settings is displayed for each method separately by color. Patterns indicate type of covariate selection: no selection (solid), p-hacking (crosshatch), selection without controlling for X (striped), selection controlling for X (dotted). |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-bar-1)

We assessed Type I error by individual research setting and found several patterns ([Figure 2](#fig-type1-panel)). First, small sample sizes are more susceptible to inflated Type I error rates when using covariate selection methods that control for *X*. As sample sizes get larger (i.e., *n* = 300), the methods become comparable. Second, as the number of available covariates to select from increases, Type I error rate increases for covariate selection methods that control for *X*, such as a full linear model with *X* and LASSO with *X*, and to a lesser degree partial correlation. Third, there appeared to be no definitive pattern in Type I error rate as the proportion of good covariates increased. Similarly we did not see changes in Type I error rate as function of the correlation strength between covariates and *Y*.

|  |
| --- |
| Figure 2: Type I error by method and research setting. Plots are paneled by aspects of the research setting (number of observations, number of covariates, proportioin of good covariates, and Y-covariate correlation strength). Methods are displayed separately by color. Methods with no covariate selection are depicted as solid lines. p-hacking is depicted as dashed line. Covariate selection methods that control for X are depicted as two dash lines. Covariate selection methods that do not control for X are depicted as long dash lines. |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-type1-panel)

## Type II Error

We simulated datasets with two possible population parameters for the effect of *X* on *Y* (, ). Any non-significant result found indicated a Type II error. Stronger statistical methods will have lower Type II error (implying greater statistical power). Since we demonstrated that p-hacking substantially inflates Type I error, making it a statistically invalid method for covariate selection, we do not evaluate Type II Error for this method. [Figure 3](#fig-bar-2) shows the average Type II error rate across all research settings for each method. Using no covariates results in the highest Type II error highlighting the importance of covariates for detecting true effects. Type II error rates trended lower for covariate selection methods that controlled for *X* compared to those that did not control for *X*.

|  |
| --- |
| Figure 3: Type II Error by Method. Average type II error rate across all research settings is displayed for each method separately by color. Patterns indicate type of covariate selection: no selection (solid), selection without controlling for X (striped), selection controlling for X (dotted). |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-bar-2)

We also assessed Type II error by individual research setting ([Figure 4](#fig-panel-2)). Across all research setting, using no covariates was associated with higher Type II error. Using all available covariates or selection methods that do not control for *X* performed had higher Type II error rates compared to the other methods when sample sizes were low. This pattern was especially notable for using all covariates and a full linear model without *X*. As sample size increased, the methods performed comparably with respect to Type II error. Using all covariates or a full linear model without *X* for selection produced higher Type II error rates compared to other methods when there was a larger number of covariates available. The full linear model without *X* also produced higher Type II error rates, compared to other methods when the proportion of good covariates was higher.

|  |
| --- |
| Figure 4: Type II error by method and research setting. Plots are paneled by aspects of the research setting (number of observations, number of covariates, proportioin of good covariates, Y-covariate correlation strength, and population parameter for X). Methods are displayed separately by color. Methods with no covariate selection are depicted as solid lines. Covariate selection methods that control for X are depicted as two dash lines. Covariate selection methods that do not control for X are depicted as long dash lines. |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-panel-2)

## Parameter Estimates

[Figure 5](#fig-distribution-bx) shows the sampling distribution for the parameter estimate of *X* for all nine methods separately by effect size (, , ). Distributions for parameter estimates for all methods are centered around the true population values. Using no covariates produces a wider distribution, indicating more variability in its parameter estimates. The top plot also highlights an unusual feature of p-hacking. When there is no true effect the distribution appears bimodal due to the artificial inflation or deflation of parameter estimates that occurs when selecting covariates to obtain a significant p-value.

|  |
| --- |
| Figure 5: Sampling Distribution for Population Parameter Estimates. |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-distribution-bx)

# Discussion

This present study sought to empirically compare methods for covariate selection in linear models. R scripts were used to generate data, fit linear models, and extract and visualize model results. The data generation process was based on full crossings of levels of variables which were chosen to represent common research settings in Psychology. Within each research setting, a unique dataset was simulated 40,000 times using high-throughput computing, with each dataset being used to fit models according to the nine methods discussed.

The first step in evaluating the nine methods, was to establish which methods are statistically valid (i.e., have a Type I error rate of 0.05). We replicated previous findings that p-hacking is not a statistically valid method as it yielded inflated Type I error rates and biased parameter estimates. The remaining eight methods were found to be statistically valid, although LASSO and full linear model show slight inflation.

With this established, the methods could be further compared by their Type II error rates. Lower Type II error can also be considered as higher statistical power. Including all covariates in a model shows a large reduction in Type II error compared to including no covariates in a model. Thus, researchers are encouraged to measure and use covariates. From there, performing a selection of covariates yields further reductions in Type II error compared to simply using all covariates. Overall, LASSO, partial correlation, and bivariate correlation resulted in the lowest Type II errors. However, there are more factors to take into consideration.

Depending on the research setting, researchers may prefer one method over another. As sample size increases (especially past ), the performance of methods tends to become indistinguishable. For the largest number of covariates tested (20), LASSO had the lowest Type II error, but for smaller numbers, it is comparable to partial correlation. LASSO showed more inflation of Type I error than the partial correlation approach which in turn showed more inflation than the bivariate correlation approach. Among these three, LASSO is the most computationally expensive and will come with a steep learning curve for new users. The bivariate correlation approach showed no inflation of Type I error, but underestimated the parameter estimates for the nonzero effects. The partial correlation approach did have slight inflation of Type I error, but correctly estimated the parameter estimates for the nonzero effects. The partial correlation approach also had lower Type II error than the bivariate correlation approach, overall and within different research contexts. With this information, researchers should consider what to prioritize depending on their goals.

# References

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013. “Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience.” *Nature Reviews Neuroscience* 14 (5): 365–76. <https://doi.org/10.1038/nrn3475>.

Center for High Throughput Computing. 2006. “Center for High Throughput Computing.” Center for High Throughput Computing. <https://doi.org/10.21231/GNT1-HW21>.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2 edition. Hillsdale, N.J: Routledge.

———. 1992. “A Power Primer.” *Psychological Bulletin* 112 (1): 155–59.

Ioannidis, John P. A. 2005. “Why Most Published Research Findings Are False.” *PLoS Medicine* 2 (8): e124. <https://doi.org/10.1371/journal.pmed.0020124>.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science* 22 (11): 1359–66. <https://doi.org/10.1177/0956797611417632>.

Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons. 2014. “P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results.” *Perspectives on Psychological Science* 9 (6): 666–81. <https://doi.org/10.1177/1745691614553988>.