Evaluating Methods for Covariate Selection

Lauren Khoury

Kendra Wyant

John J. Curtin

2025-10-16

Abstract

Abstract of paper goes here and can span several lines.

# Introduction

## 1

We are often interested in experimental manipulations in Psychology. - Namely, to test if a focal variable has a significant effect on an outcome variable of interest. (give simple example)

## 2

We all know that power is important (frame in terms of type 2 error). - low powered studies don’t replicate well (Cohen, replication crisis) - Statistical power is the ability to find a significant effect when one truly does exist. Studies with low power do not replicate well, which led to a replication crisis and an increased awareness of false-positive results. - Type 2 error rate is positively correlated with the “likelihood that a nominally statistically significant finding actually reflects a true effect” - “problems with low power when all other research practices are ideal: low probability of finding true effects, low ppv when an effect is claimed, and exaggerated estimate of the magnitude of the effect when a true effect is discovered (Effect inflation is worst for small, low-powered studies, which can only detect effects that happen to be large.)” (PPV is the probability that a ‘positive’ research finding reflects a true effect (that is, the finding is a true positive)). - statistical power definition - “The statistical power of a test is the probability that it will correctly reject the null hypothesis when the null hypothesis is false (that is, the probability of not committing a type II error or making a false negative decision). The probability of committing a type II error is referred to as the false negative rate (β), and power is equal to 1 – β.”

statistical power citations - (Cohen 1988, 1992)

(Button et al. 2013) - shows the average statistical power of studies in neuroscience is very low, causing overestimates of effect sizes and low reproducibility of results. They also raise ethical concern that unreliable research is inefficient and wasteful.

## 3

Covariates (measured at baseline before manipulation) to increase power (classic cites?) - keep type 2 error rate down. - Covariates are variables that do not have focal hypotheses, but are added to increase statistical power and model precision of the focal parameter estimate. - They explain some of the variance that might otherwise show up as noise. - Including covariates increases detection of true effects of the manipulated variable of focus. For example, if researchers were testing the effects of a new treatment technique on depression, variables such as age, sex, gender identity, socioeconomic status, and education could be measured at baseline and included in the model as covariates. (continue earlier example with covariates added that intuitively might be related to outcome) This would increase power to find if there is a true significant effect of the new treatment on depression. This increase in power comes from a reduction in the error of the model as covariates account for more unexplained variance in the outcome. This reduces the standard error of the effect which yields a more precise parameter estimate.

## 4

Evidence that p-hacking is problematic (Simons) - Previously, it was common practice to test multiple statistical analyses until statistical significance was reached with only the significant result being reported (Simmons, Nelson, and Simonsohn 2011). This repeated testing and selective reporting of significant results came to be known as p-hacking (Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons 2014). This led to an increase in “false-positive psychology” where researchers were finding significant effects that did not truly exist. This was motivated by two factors: an inherent publication bias of journals to only publish significant results, leading to researchers striving to reach this result and an ambiguity in decision-making (Simmons, Nelson, and Simonsohn 2011). Researchers have various decisions such as sample size of the study, handling of outliers, and including covariates in linear models. There are not always clear guidelines on how to make these decisions.

Ioannidis, J.P. Why most published research findings are false. PLoS Med. 2, e124 (2005). - This study demonstrates that many (and possibly most) of the conclusions drawn from biomedical research are probably false. The reasons for this include using flexible study designs and flexible statistical analyses and running small studies with low statistical power.

Simmons, J.P., Nelson, L.D. & Simonsohn,U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22, 1359–1366 (2011). - This article empirically illustrates that flexible study designs and data analysis dramatically increase the possibility of obtaining a nominally significant result. However, conclusions drawn from these results are almost certainly false.

## 5

Perhaps shouldn’t use covariates. Though as we mentioned this would come with cost to power/increased type 2 error. Another conclusion might be to use all covs (all variables reasonably believed to account for variance in outcome) - reduces degrees of freedom. - The prevalence of possible covariates to measure in the field of Psychology is high, making it difficult for researchers to know how many and which ones to add to a model. - Some measured covariates may be redundant as they explain overlapping variance in the outcome or they may not be related to the outcome at all. - Ideally, only the “good” covariates should be included in statistical models. “Good” covariates are uncorrelated with the focal independent variable and explain variance in the outcome. In experimental work, covariates are measured at baseline before the manipulation and are, therefore, not correlated with the manipulated variable, so the first part is satisfied. The second part is less trivial and needs further exploration.

## 6

We need a method for selecting covariates that doesn’t nominally change type 1 error rate. - introduce proposed methods - least absolute shrinkage and selection operator (LASSO)

Our research aims to provide clear and accessible methods for researchers to use to select among a set of covariates by simulating … across several research settings.

We replicate that p-hacking is not a statistically valid method and propose eight other methods for selecting covariates. The methods are tested and compared in systematically varied research settings accounting for different effect sizes, sample sizes, amounts of available covariates, proportions of good covariates, and correlations. The methods are compared by their Type I error rates (the probability of rejecting the null hypothesis when no effect exists) and Type II error rates (the probability of failing to reject the null hypothesis when an effect does exist).

# 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. **?@fig-typeI-bar** 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). The horizontal dotted black line indicates expected Type 1 Error rate for an alpha = .05. |

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

We assessed Type I error by individual research setting and found several patterns emerged ([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. The horizontal dotted black line depicts expected type I error rate for an alpha of 0.5. |

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. **?@fig-typeII-bar** shows the average type II error rate across all research settings for each method.

|  |
| --- |
| 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-typeII-bar)

## Parameter Estimates

As mentioned, this section considers a population parameter of *X* set to zero. Figure [Figure 4](#fig-distribution-bx) shows the sampling distribution for the parameter estimate of *X* for all nine methods. For further information about the average estimate, see the table in the Supplemental Material section. This table also displays the standard deviation of the estimates from the linear model, the average standard error of the model, and their difference. The standard error of the sampling distribution is expected to match the standard error of the parameter estimates. The table shows a large difference in these values for the p-hacked method and slight differences for LASSO, full linear model, and partial correlation approaches.

This distribution highlights another negative consequence of p-hacking. While its distribution is still centered around zero, it is not normally distributed around zero, emphasizing a biasing of the parameter estimates. The remaining distributions are centered around zero. The model fit with no covariates has the widest distribution, indicating greater variability in its estimate for the effect.

As mentioned, this section considers two population parameters of set to 0.3 and 0.5. Figure show the sampling distributions for the parameter estimate of for all eight methods. See the tables in the Supplemental Material section for more information about the average estimates, the standard deviation of the estimates, and the average standard error of the model for both 0.3 and 0.5 effect sizes. The tables show that the bivariate correlation, full linear model without , and LASSO without approaches slightly underestimate the parameter estimate.

Both distributions show parameter estimates centered around the population values. In both effect size cases, the method including no covariates has a wider distribution, and therefore, more variability in its parameter estimates.

|  |
| --- |
| Figure 4: 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.

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>.