Evaluating Methods for Covariate Selection

Lauren Khoury

Kendra Wyant

John J. Curtin

2025-08-27

Abstract

Abstract of paper goes here and can span several lines.

# Introduction

Linear regression models are a commonly used tool in the field of Psychology for researchers to test hypotheses. Namely, to test if a focal variable has a significant effect on an outcome variable of interest.

While a simple model could be fit regressing the dependent variable solely on the focal variable, model performance can be improved by adding covariates. Covariates are variables that do not have focal hypotheses, but are added to increase statistical power and model precision of the focal parameter estimate.

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. 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. 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.

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. 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.

Collecting data on many covariates can be simple, but selecting which covariates to include in linear models is less so. 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. How can researchers decide which covariates to add to their models?

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 (**urisimonsohnPCurveEffectSize2014?**). 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.

Our research aims to provide clear and accessible methods for researchers to use to select among a set of collected covariates to include in a linear model in order to increase their ability to find a true significant effect. 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).

# Method

All computations were conducted in R, including scripts to generate data based on crossings of research setting variables, fit linear models based on different methods for covariate selection, and extract and visualize the results from these models. The process was repeated across multiple simulations via high-throughput computing.

## Data Generation Process

For the data generation process, we manipulated variables as shown in [Table 1](#tbl-dictionary) with levels chosen that are commonly found in Psychology researh. The population parameters for were chosen to cover both zero and nonzero effects, with the nonzero effects covering a medium and larger effect size. We chose values for the number of observations that pertain to common sample sizes in experimental research.

The four variables handling the covariates were chosen with the plan of crossing all levels. Hence, the numbers of covariates had to be divisible by four to cross with the chosen levels of percentages. The values cover a range of amounts of covariates that researchers might have available to them. The percentages of “good” covariates represent the reality that while many might be available, there are varying strengths to which these covariates relate to the outcome. The number of total available covariates by the percentages of good covariates gives us the number of good covariates. This is factored into the data generating process in the covariance matrix, where there is a nonzero entry in the matrix intersection between and each good covariate. The value for this is given by the -covariate correlation (r\_ycov). Since these good covariates are correlated with , they must be correlated with each other. This gives a nonzero entries in the matrix intersections between the good covariates. This value is given by the covariate correlation (r\_cov). The values for the two correlation variables (between and the covariates and between the covariates themselves) were selected with what we feel is expected in social science contexts. These values also had to satisfy a positive-definite covariance matrix as and the covariates were generated from a multivariate normal distribution. The variable was generated as a dichotomous variable representing an experimental manipulation (e.g., condition versus treatment). The final values for were then calculated by adding the generated with the covariates to the variable multiplied by the given population parameter for .

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Data Dictionary   | Variable | Description | Values | | --- | --- | --- | | b\_x | the population parameter for X | 0, 0.3, 0.5 | | n\_obs | number of observations in sample | 50, 100, 150, 200, 300, 400 | | n\_covs | number of covariates | 4, 8, 12, 16, 20 | | r\_ycov | correlation between Y and covariates | 0.3, 0.5 | | p\_good\_covs | percentage of "good" covariates | 0.25, 0.50, 0.75 | | r\_cov | correlation between "good" covariates | 0.3 | |

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

## Covariate Selection

There are nine methods for selecting covariates to include in linear models that we evaluated. The methods can best be broken down into two categories: those that do not involve systematically selecting covariates and those that do. For the former category, we explored inlcuding (1) no covariates in the linear models and including (2) all available covariates. Here, we consider all available covariates to mean all covariates that were measured prior to manipulation.

For the latter category, we first employed the method of (3) p-hacking to replicate that this is not a statistically valid method to use for covariate selection. To accomplish this method, we include covariates that lower the p-value for when comparing it to the p-value from the model regressing only on .

This leaves six remaining methods that we predicted to be viable and successful methods for covariate selection. We consider the Pearson correlation coefficient () for both bivariate and partial correlations. In the (4) bivariate correlation model, we consider one covariate at a time and include it in the final model if it has a significant effect on . Note, that for this paper a significance level of is used. In the (5) partial correlation model, we consider one covariate at a time and include it in the final model if it has a significant effect on while also controlling for . We then fit full linear models in cases of including and excluding . We fit a (6) full linear model regressing on all covariates, and we included all covariates in the final model that are significant on when controlling for . Similarly, we fit a (7) full linear model without , where we include all covariates significant on without controlling for . Finally, we considered a more complex approach of covariate selection by employing the least absolute shrinkage and selection operator (LASSO, also known as L1 regularization). We tuned the penalty value across 100 bootstraps, fit the best model with the penalty that yielded the lowest RMSE, and selected all covariates with nonzero coefficients for inclusion in the final linear model. This process was the same for a (8) LASSO model that had a zero penalty factor applied to (i.e., was retained in the model) and for a (9) LASSO model that did not contain .

To summarize, these are the nine models compared in this paper used for covariate selection:

1. No covariates
2. All covariates
3. P-hacking
4. Bivariate correlation
5. Partial correlation
6. Full linear model
7. Full linear model without
8. LASSO
9. LASSO without

## Simulations

The crossing of all levels of each variable resulted in 540 unique research settings. We ran 40,000 simulations for each unique setting using the Center for High Throughput Computing (CHTC) at the University of Wisconsin, Madison (**chtc?**). A seed was set for each simulation for the purpose of reproducibility. Within each simulation, the scripts first generated a unique dataset based on the variables and process discussed. Then, a linear model was fit according to each of the nine aforementioned methods for covariate selection. Finally, the results from these models were saved to a dataframe.

These results include the corresponding parameter estimate, standard error, and p-value for from the linear model. The numerator and denominator degrees of freedom were also extracted from the model. We calculated true positive rates and false positive rates to identify the rates at which the different methods correctly selected covariates that did relate to and incorrectly selected covariates that did not relate to , respectively.

## Data Analysis

The nine methods will be compared in two separate conditions - zero and nonzero effects. In the former condition, the Type I error will be compared and in the latter condition, the Type II error will be compared. For both error rates, the overall error of each method will be considered in addition to examining the error at different levels of the research context variables. The sampling distributions of the parameter estimates for will also be presented.

# Results

The product of the sizes of the unique research settings (540), simulations (40,000), and methods (9) yielded a total of 194,400,000 observations. The data contained 16 columns, six of which were the variables indicating the research setting (see [Table 1](#tbl-dictionary)) and one indicating the selection method applied. Five columns were values extracted from the linear model output: the parameter estimate, standard error, and p-value for and the numerator and denominator degrees of freedom of the model. The true and false positive rates were also calculated. The remaining two columns held information about the simulations we were running including the job number and identification number of the simulation; these two columns were not part of the data analysis.

The results will be considered first by the zero effect followed by the nonzero effects. The Type I and Type II errors will be compared across methods and across levels of research setting variables. Figures will be used to visualize the error rates, but tables providing the minimum, maximum, and average error can be found in the Supplemental Material section. Note that for the line plots throughout this report, a solid line will indicate a method that does not involve performing selection of covariates (i.e., using no or all covariates). A dashed line will indicate a method that does involve performing a non-trivial selection of covariates. A dotted black line will indicate the expected value (if applicable). For example, there will be a dotted line at , the signifance level considered here and hence, the expected Type I error rate.

## Type I Error

In the case of a zero effect, we set the population parameter for to be zero (i.e., ), so that any significant result found is a Type I error. The first comparison will look at selection method overall, aggregated across all research settings. In Figure (**typeI-bar?**), the average proportion of significant effects – the Type I error – is calculated and displayed. Refer to the Supplemental Material section to see the distribution of Type I error of each method.

|  |
| --- |
| Figure 1: Type I Error by Method |

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

Previous findings are replicated that using p-hacking for selecting covariates leads to a highly inflated Type I error rate (0.198). The p-hacked model consistently shows high Type I error, as is evident in the remaining figures in this section. Most methods are at the expected 0.050 mark, while the partial correlation approach shows slight inflation (0.052), with both LASSO and full linear model showing further inflation (0.057 and 0.060, respectively). These eight methods (excluding p-hacking) can be considered statistically valid as they show little to no inflation of Type I error rates.

### By Research Settings

This section will consider the Type I error rates across the established values of the research context variables.

Figure **?@fig-typeI-nobs** displays the change in Type I error as the number of observations in a sample increases for all nine methods.

|  |
| --- |
| Figure 2: Type I Error by Number of Observations |

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

At the smallest sample size (), the discrepancy between methods is greatest, with partial correlation, LASSO, and full linear model deviating from the expected value of 0.05. However, as sample size increases, all methods become comparable, especially with or larger.

Figure **?@fig-typeI-ncovs** displays the change in Type I error as the number of covariares increases for all nine methods. Recall that the number of covariates refers to the number of available covariates, not necessarily the number included in the final model.

|  |
| --- |
| Figure 3: Type I Error by Number of Covariates |

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

There is a slight increase in Type I error for LASSO and full linear model as the number of covariates increases, while the other methods stay at or around 0.05.

Figure **?@fig-typeI-pgoodcovs** displays the change in Type I error as the proportion of good covariares increases for all nine methods.

|  |
| --- |
| Figure 4: Type I Error by Proportion of Good Covariates |

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

Most of the methods remain stagnant in Type I error as the proportions increase, except the error of LASSO decreases slightly and that of full linear model increases.

Figure **?@fig-typeI-rycov** shows the change in Type I error as the correlation between and the good covariates increases for all nine methods. The correlation among the good covariates did not vary, so only the -covariate correlation will be compared.

|  |
| --- |
| Figure 5: Type I Error by Y-Covariate Correlation |

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

All of the valid methods did not see a great change in Type I error as the correlation increased.

## Parameter Estimates

As mentioned, this section considers a population parameter of set to zero. Figure [Figure 6](#fig-distribution-bx-0) shows the sampling distribution for the parameter estimate of 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.

|  |
| --- |
| Figure 6: Sampling Distribution for b\_x=0 Estimate |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-distribution-bx-0)

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.

## Type II Error

In the case of a nonzero effect, we tested two values for the population parameter for , 0.3 and 0.5 (i.e., ), so that any non-significant result found is a Type II error. There is no benchmark of expected Type II error as there was for Type I, so the stronger methods will be those with lower Type II error (implying greater statistical power). As the p-hacked method was established as a method that is not statistically valid, it will not be considered in comparisons of Type II error.

The first comparison will look at selection method overall, aggregated across all research settings, for both effect sizes. In Figure (**typeII-bar-03?**) and Figure (**typeII-bar-05?**), the proportion of non-significant effects – the Type II error – is calculated and displayed. See the Supplemental Material section for the distributions of Type II error of each valid method.

|  |
| --- |
| Figure 7: Type II Error by Method, b\_x=0.3 |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-typeII-bar-03)

|  |
| --- |
| Figure 8: Type II Error by Method, b\_x=0.5 |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-typeII-bar-05)

For both effect sizes, similar trends are shown. Fitting a model with no covariates results in the highest Type II error. When further examining the error rates within research settings, the no covariates approach will continue to yield the highest Type II error. There is a large reduction in error when all covariates are included in the model instead of no covariates. Additionally, performing selection of these covariates further reduces Type II error from simply including all covariates.

### By Research Settings

Figure **?@fig-typeII-nobs-03** and Figure **?@fig-typeII-nobs-05** display the change in Type II error as the number of observations in a sample increases for all eight methods.

|  |
| --- |
| Figure 9: Type II Error by Number of Observations, b\_x=0.3 |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-typeII-nobs-03)

|  |
| --- |
| Figure 10: Type II Error by Number of Observations, b\_x=0.5 |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-typeII-nobs-05)

At the smallest sample size, the methods differ in Type II error, with LASSO and partial correlation approaches yielding the lowest in both effect size cases. As sample size increases, the seven methods (excluding the no covariates method) tend to become indistinguishable.

Figure **?@fig-typeII-ncovs-03** and Figure **?@fig-typeII-ncovs-05** display the change in Type II error as the number of covariates increases for all eight methods.

|  |
| --- |
| Figure 11: Type II Error by Number of Covariates, b\_x=0.3 |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-typeII-ncovs-03)

|  |
| --- |
| Figure 12: Type II Error by Number of Covariates, b\_x=0.5 |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-typeII-ncovs-05)

For the smallest number of covariates, the methods are comparable in Type II error, but for larger numbers of covariates, there is a larger gap in performance. As the number of covariates increases, LASSO begins to have the lowest Type II error, followed by the partial correlation and bivariate correlation approaches. Additionally, the method of including all covariates becomes out-performed by the selection methods.

Figure **?@fig-typeII-pgoodcovs-03** and Figure **?@fig-typeII-pgoodcovs-05** display the change in Type II error as the proportion of good covariates increases for all eight methods.

|  |
| --- |
| Figure 13: Type II Error by Proportion of Good Covariates, b\_x=0.3 |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-typeII-pgoodcovs-03)

|  |
| --- |
| Figure 14: Type II Error by Proportion of Good Covariates, b\_x=0.5 |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-typeII-pgoodcovs-05)

For higher proportions of good covariates, LASSO, partial correlation, and bivariate correlation have the lowest Type II error.

Figure **?@fig-typeII-rycov-03** and Figure **?@fig-typeII-rycov-05** display the change in Type II error as the correlation between and the good covariates increases for all eight methods.

|  |
| --- |
| Figure 15: Type II Error by Y-Covariate Correlation, b\_x=0.3 |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-typeII-rycov-03)

|  |
| --- |
| Figure 16: Type II Error by Y-Covariate Correlation, b\_x=0.5 |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-typeII-rycov-05)

All methods (excluding the no covariates approach) show similar decreasing trends in Type II error as the correlation between and the good covariates increases. The differences between these methods are minimal.

## Parameter Estimates

As mentioned, this section considers two population parameters of set to 0.3 and 0.5 Figure [Figure 17](#fig-distribution-bx-03) and Figure [Figure 18](#fig-distribution-bx-05) 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.

|  |
| --- |
| Figure 17: Sampling Distribution for b\_x=0.3 Estimate |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-distribution-bx-03)

|  |
| --- |
| Figure 18: Sampling Distribution for b\_x=0.5 Estimate |

Source: [Make Figures for Main Manuscript](https://jjcurtin.github.io/study_cov/notebooks\mak_figures-preview.html#cell-fig-distribution-bx-05)

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.

# 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

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>.