Study cov

Lauren Khoury

Name 2

Name 3

John J. Curtin

2025-08-05

Abstract

Abstract of paper goes here and can span several lines.

# Introduction

This is how you cite a paper using the bibtex key in Zotero (Simmons, Nelson, and Simonsohn 2011).

In Psychology research, we often use linear regression models to test if a focal variable has a significant effect on an outcome variable of interest. While we could fit a model regressing our dependent variable only on our independent variable, it would benefit us to include covariates. Covariates are variables that are not focal independent variables, but are added to a model to increase power and precision. Statistical power is the ability to find a significant effect when one truly does exist. Studies with low power do not replicate well (replication crisis).

For the purposes of this paper, we will consider the case where our focal variable () is an experimental manipulation. In this context, covariates are baseline measures taken before the manipulation and are therefore uncorrelated with the manipulation. They increase power to detect effects of the manipulated variable of focus. For example, if researchers were interested in 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. This would increase power to find if there is a true significant effect of the new treatment. The increase in power comes from a reduction in the error of the model. Covariates reduce 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. Ideally, only the “good” covariates should be included in statistical models. By “good” we mean a couple of things: we want variables that are uncorrelated with the focal IV and variables that explain variance in the outcome. Again, in experimental work, variables measured at baseline are not correlated with the manipulated variable.

Collecting data on many covariates is the simple part, but then selecting which covariates to include in models is less simple. Some measured covariates may be redundant (i.e., explain overlapping variance in the outcome) or are not related to the outcome. (p-hacking history).

# Method

## Data Generation Process

In order to evaluate methods for covariate selection, we wrote R scripts to generate data, fit linear models, and extract results from these linear models. For the data generation process, we manipulated variables as shown in **?@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 .

| 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. A seed was set for each simulation ran 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. 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

# 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. A glimpse/head of the data is shown below:

*figure*

We will examine the results by zero and nonzero effects. The Type I and Type II errors will be compared across methods and across levels of the research setting variables. 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 expected Type I error rate.

## Zero X Effect

We begin by considering the condition with a zero effect. In this case, 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, across all research settings. In this figure (**typeI-bar?**), the proportion of significant effects – the Type I error – is calculated and displayed.

|  |
| --- |
| Figure 1 |

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

From this bar plot, we see that the p-hacking method for selecting covariates leads to highly inflated Type I error rates. In the remaining figures in this section, we will continue to see this inflation from p-hacking. Most of the approaches are at the expected 0.05 mark, while the partial correlation approach shows slight inflation, with the LASSO and full linear models showing further inflation of Type I error.

We will now present the Type I error rates of each method for the different levels of the number of observations in a sample.

|  |
| --- |
| Figure 2 |

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

For smaller sample sizes, we see that partial correlation, LASSO, and full lm perform worse than the other methods, but they all become comparable as sample size increases.

When comparing across the number of covariates, this is not necessarily the number of covariates included in the final model, but the amount of available ones.

|  |
| --- |
| Figure 3 |

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 for full lm and LASSO as the number of covariates increases, while the other methods stay around 0.05.

|  |
| --- |
| Figure 4 |

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 across the different percentages, except LASSO decreases slightly and full lm increases as the percentage of good covariates increases.

The correlation among the good covariates did not vary, so only the -covariate correlation will be compared.

|  |
| --- |
| Figure 5 |

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

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

|  |
| --- |
| Figure 6 |

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. In addition to inflation of Type I error rates, it biases the parameter estimates. While its distribution is centered around zero, it has a bimodel distribution, emphasizing how it biases the estimates. The remaining distributions are centered around zero. The no covariates approach has the widest distribution, showing that is has more variability in its estimate.

## Nonzero X Effect

We continue on to consider conditions with a nonzero effect. Here, we tested two nonzero values for the population parameter for , 0.3 and 0.5 (i.e., and ). In this case, any non-significant result found is a Type II error.

|  |
| --- |
| Figure 7 |

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

|  |
| --- |
| Figure 8 |

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

|  |
| --- |
| Figure 9 |

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

|  |
| --- |
| Figure 10 |

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

|  |
| --- |
| Figure 11 |

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

|  |
| --- |
| Figure 12 |

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

|  |
| --- |
| Figure 13 |

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

|  |
| --- |
| Figure 14 |

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

|  |
| --- |
| Figure 15 |

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

|  |
| --- |
| Figure 16 |

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

|  |
| --- |
| Figure 17 |

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

|  |
| --- |
| Figure 18 |

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

# Discussion

# 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>.