Covariate Selection in Small Randomized Studies

David R. Judkins
Abt Associates, 6130 Executive Blvd., Rockville, MD 20852

Abstract
Considerable progress has been made in recent years in the asymptotic properties of procedures for covariate control in randomized studies. This paper focuses on the special challenges in small randomized studies. Simulation studies demonstrate that a cross-validated LASSO is an excellent choice. No selection at all results in power loss compared to no covariate adjustment at all. Stepwise procedures yield lower variances but also result in underestimated variances and excess type I error.

Key Words: LASSO, Koch’s method, Backward selection

1. Background

It is well known that covariance adjustment is optional for the analysis of randomized studies. One can use non-parametric tests due to Fisher for the strong null hypothesis of no effect of treatment at all, or a simple t-test for the difference in means for the weak null hypothesis of no mean shift due to treatment (Deaton and Cartwright, 2018). Despite this superfluity, covariance adjustment is the norm in the analysis of randomized evaluations of social interventions. Motivations vary among researchers. Some use it in the hope that it will boost power. Other use it in the hope that it will enhance the face validity of results by “controlling” for random imbalances in baseline covariates. It is also well known that while covariance adjustment has the potential to increase power (sometimes substantially), it can also thwart the first of these hopes by increasing variances on estimated effects rather than decreasing them (Freedman, 2008a, 2008b) and that the second goal is a hopeless endeavor given the infinite number of ways that a randomized sample can be “out of balance” (Tukey, 1991). Moreover, regression adjustment can compromise true validity by causing underestimation of variances on estimated effects (Freedman, 2008a, 2008b) and enabling researcher misbehaviour in which effects are manufactured by open-blind decisions about covariate selections, where researchers hunt (consciously or subconsciously) for the set of covariates that best supports the impacts expected or desired by the researcher. Lastly, the burden of explaining complex adjustment techniques may actually reduce face validity by clouding the transparency of methods. Nonetheless, I am unaware of any important randomized studies of social interventions where regression adjustment was not employed.

There is a rich literature on asymptotically optimal procedures for regression adjustment, whereby variances are both reduced and consistently estimated. Lin (2013) rehabilitated the regression adjustment given infinite sample sizes and a fixed set of regressors. Tsiatis et al. (2008) proved the asymptotic equivalence of several popular methods of choosing regressor and suggested a novel method that has the potential to reduce the risk of tainted results due to researcher misconduct. However, there is still a lack of guidance for practitioners in the analysis of small randomized studies. This is a gap that this paper aims to fill. It examines the properties of several popular methods for regression adjustment in the context of small studies with many useless covariates.

In section 2, I describe the alternate methods. Section 3 sets up a simulation study. Section 4 presents results from it, and section 5 closes with discussion.

2. Alternative Methods

This paper considers four methods for covariate selection. The first alternative is to make no selection at all, but to use every available covariate as in (2.1), where Y_i is the outcome, T_i is a 0/1 dummy variable indicating treatment group membership, X_i is a row vector of baseline covariates, β is the vector of parameters indicating the influence of each covariate on the outcome, δ is the effect of treatment, and e_i is an error term.

$$Y_i = X_i\beta + T_i\delta + e_i\,, \tag{2.1}$$

Obviously, there are settings where this alternative will not be feasible, but if *p*, the rank of the design matrix, is say less than *n*/5 for interval-valued outcomes (or perhaps *n*/30 for rare binary outcomes) and *n*, the sample size, is large, this method should produce consistent variance estimates, but is unlikely to achieve the lower bound on variance. The second method is to use backward selection with a p-value to retain of 0.20. It will also run into problems as *p* approaches *n*, but it is feasible for many problems.

The third method is a modification I suggested to Koch’s method. Koch, et al. (1998) referred to their method as nonparametric ANCOVA, but since then, most authors have referred to it as Koch’s estimator. With Koch’s method, a working model (2.2) is fit for the outcome on interest in terms of the full set of covariates but omitting the treatment indicator, and then the effect of treatment is estimated as the mean difference in residuals across treatment and control, as is equations (2.3) and (2.4). My suggested modification to Koch’s method is to fit 2.2 just on the control sample instead of on the full sample as originally suggested by Koch and co-authors.

My motivation for this suggestion was the demonstration by Lesaffre and Senn (2003) that Koch’s estimator can produce overly-liberal significance tests (i.e., tests with type I error rates higher than the claimed nominal rate), because of negative bias in the variance estimator shown in equation 2.5. My hope was that only using half the sample to estimate residuals would lessen this negative bias in the variance estimates. An additional thought that 2.4 would be easier to explain to non-technical audiences when the null hypothesis is rejected. With my modification to Koch’s estimator, the point estimate is the difference between the population-wide average outcome under universal treatment and the population-wide average outcome under the status quo. The point estimator with the original procedure does not admit such a simple explanation.

$$Y_i = X_i\beta + e_i\,, \tag{2.2}$$

$$\hat{r}_i = Y_i - X_i\hat{\beta}, \tag{2.3}$$

$$\hat{\delta} = \hat{\mu}_T - \hat{\mu}_C = \frac{\sum_i T_i\hat{r}_i}{\sum_i T_i} - \frac{\sum_i (1-T_i)\hat{r}_i}{\sum_i (1-T_i)}, \tag{2.4}$$

$$\mathrm{var}(\hat{\delta}) = \frac{\sum_i T_i(\hat{r}_i - \hat{\mu}_T)^2}{\sum_i T_i - 1} + \frac{\sum_i (1-T_i)(\hat{r}_i - \hat{\mu}_C)^2}{\sum_i (1-T_i) - 1}, \tag{2.5}$$

The fourth method is the 10-fold cross-validated LASSO.[1] A step-by-step description of the procedure is given below. Briefly with the ordinary LASSO, the sum of absolute values of the estimated regression coefficients in Equation A.2 is constrained to be less than a tuning parameter, λ. If the value for λ is small enough, many coefficients in Equation A.2 will be forced to zero in order to fit within the cap on the sum of absolute coefficient values and thus can be removed from the list of baseline covariates. The 10-fold cross-validation is used to optimize the value of λ, rather than just relying on an arbitrary choice.

Details of the procedure are as follows:

1. With 10-fold cross-validation, the sample (both treatment and control group members) is divided into 10 equal and mutually exclusive random subsamples.
2. For each of a range of candidate values of λ, the LASSO procedure is run to select covariates on a sample in which one of the 10 subsamples has been dropped.
3. The model in Equation A.2 is fit on the same sample using just the variables selected in the second step for each of the candidate values of λ.
4. The model is used to create out-of-sample predictions of the outcome for everyone in the excluded piece of the sample, and the prediction error $\hat{Y}_i - Y_i$ is measured for each of the candidate values of λ.
5. Steps 2 through 4 are repeated 10 times for each candidate value of λ. On each iteration, a different one of the 10 subsamples is dropped. In this manner, out-of-sample prediction errors are obtained for the entire sample.
6. Mean squared prediction errors across all 10 replicates are then calculated for each of the candidate values of λ.
7. The value of λ that minimizes this cross-validated mean squared prediction error and thus captures most of the variation reduction possible with the available covariates is selected as the optimal constraint.[2] Whichever variables have nonzero coefficients in the model for that optimal constraint are used as covariates in the impact regressions. All other baseline characteristics are discarded. All of this is done automatically in SAS®/GLMSELECT with the "CHOOSE" parameter set to CVPRESS.

[1] "Least absolute shrinkage and selection operator." See Bühlmann and van de Geer (2011) for a full explanation.

[2] One could simply use the LASSO to select covariates with a pre-specified value of the constraint, but the 10-fold cross-validation provides a principled method for selecting the constraint.

With all methods except the modified Koch method, whichever covariates are selected are then used in equation 2.1 to estimates the treatment effect. One could use variance estimates that correct for heteroscedasticity, but for this paper, I used standard OLS procedures to estimate the effect of treatment and of the associated variance. See Judkins and Porter (2016) for a discussion of why it is unnecessary to use a logistic regression to analyze the effect of treatment on a binary outcome.

3. Simulation Study Design

Equation 2.6 shows the basic structure of the superpopulation model used for the simulations. The outcome in this superpopulation is binary, with a logit propensity that is a linear function of a single standard normal covariate, *x*, and binary treatment, *T*. Different values of α lead the outcome to be more or less rare. The rather odd looking value for δ ensures that the simulated experiment has decent power to reject the null hypothesis when it is false. The two values of β lead to mild or strong value in using *x* as a regressor. In addition to *x*, I simulated either 34 or 100 additional useless covariates, and sample sizes of 500, 750, 1000, and 2000 (all sizes that were commonly encountered in evaluations of labor force interventions). In total, this paper is based on the performance of the alternative covariate selection procedures across 128 scenarios, 64 for the null and 64 for the alternative. For each scenario, there are 2000 Monte Carlo replications of the superpopulation, each of which were analysed with each of the four methods for covariate selection.

$$
\begin{aligned}
&\log\left(\frac{\Pr\{y=1\}}{1-\Pr\{y=1\}}\right)=\alpha+\beta x+\delta T \\
&\alpha=-2.94,-2.20,-1.10,\text{ or }0 \\
&\beta=1.5\text{ or }2.5 \\
&\mu=\exp(\alpha)/(1+\exp(\alpha)) \qquad (2.6)\\
&\delta\text{=0 or }1.8\sqrt{4\mu(1-\mu)/n} \\
&x\sim\text{N}(0,1) \\
&T\sim\text{B}(1,0.5)
\end{aligned}
$$

4. Simulation Results

I measured type I error rates, bias in the estimated effect of treatment, and precision gains from regression adjustment. However, there was no difference in the methods in terms of bias for the effect of treatment. They were all unbiased under the null and very slightly biased under the alternative hypothesis. As a result, I only tabulate Type 1 error rates and precision.

Table 1 shows the error-control properties of the four methods. For reference, it also includes what one would obtain if one somehow knew which *x* really belonged in the model. The findings imply that both backward selection and the modified Koch method are invalid procedures. They fail to control type I error rates. In contrast, no selection and the 10-fold cross-validated LASSO control the Type I error rates just as well as if one were able to use the single *x* that matters. Also note that the type I error rate varies more across scenarios with backward selection and the modified Koch method. Type I error rates are worse (not shown) for the smaller sample sizes and larger numbers of useless covariates.

Table 2 shows the precision gain due to use of regression adjustment compared to a method that just compares means. The cross-validated LASSO clearly gives the strongest variance reduction. In fact, for reasons unclear to me, it slightly edges out knowing the single correct covariate. Even so, there are some scenarios for which regression adjustment does worsen precision. Also, the modified Koch method was an unfortunate idea. It consistently gives less precision improvement than any of the other methods. Backward selection is lightly better than no selection, but given the lack of control of Type I error rates, one should still not consider using it.

Table 1. Type I Error Control

Method	Type I Error Rate	p-value (for error rate>0.05)	Standard Deviation Across 64 Scenarios
No selection	0.0502	0.345	0.004
Backward with p-value of 0.2 to retain	0.0601	<0.001	0.010
Modified-Koch	0.0721	<0.001	0.018
10-Fold Cross-validated LASSO	0.0507	0.109	0.005
Correct model	0.0508	0.104	0.005

Note: Across 64 scenarios, with varying sample sizes, numbers of useless covariates, explanatory power of the single useful covariate, and event rarity. Simulation error on estimated type I error rate is ±0.0012.

Table 2. Precision Gain due to Regression Adjustment

	Ratio of Standard Error to Standard Error of Unadjusted Effect of Treatment (Across 128 Scenarios)			
Method	Mean	Min	Max	Standard Deviation Across 128 Scenarios
No selection	0.921	0.770	1.184	0.087
Backward with p-value of 0.2 to retain	0.907	0.765	1.136	0.082
Modified-Koch	0.972	0.779	1.376	0.115
10-Fold Cross-validated LASSO	0.880	0.751	1.062	0.077
Correct model	0.881	0.752	1.063	0.077

Note: Smaller is better here. A value of 0.921 means that the standard error is 7.9 percent smaller than would be obtained without covariance adjustment.

5. Discussion

The 10-fold cross-validated LASSO appears to be a safe and useful tool for covariate selection in studies where *p* is not more than *n*/5. It may be safe for higher values of *p* as well, but I did not study that range. Note that to realize the precision gains estimated here in an evaluation, the covariate selection must be customized for every outcome. This appears to run against the instincts and customs of some professional evaluators. At least some in my acquaintances prefer having a single set of covariates for every outcome in an evaluation study. Having a single set of covariates improves method transparency and eases the documentation burden. Taking the union of covariates that are selected for any outcome would be a possible approach, but if the set of outcomes is large, this may end up performing more like no selection.

The other issue in application will be whether to add some covariates because they show strong imbalance at baseline. This will also tend to make the system perform more like no selection. Adding covariates that are out of balance but unrelated to the outcome is certain to increase variances on estimated effects.

Acknowledgments

Thanks to Austin Nichols for suggesting the simulation design and to Jacob Klerman for helpful comments on the draft. This research was funded by a contract with the Office of Program Evaluation in the Administration for Children and Family to evaluate PACE (http://www.career-pathways.org/acf-sponsored-studies/pace/). However, the views expressed are those of the author and not necessarily those of Abt Associates or of OPRE.

References

Bühlmann, P., and S. van de Geer. 2011. *Statistics for High-Dimensional Data.* Berlin, Heidelberg, Germany: Springer.

Deaton, A. and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*. 210, 2-21.

Freedman, D. A. (2008a). On regression adjustments to experimental data. *Adv. in Appl. Math.* **40** 180–193.

Freedman, D. A. (2008b). On regression adjustments in experiments with several treatments. *Ann. Appl. Stat.* **2** 176–196.

Judkins, D. R. and Porter, K.E. (2016). Robustness of ordinary least squares in randomized clinical trials. *Statistics in Medicine,* 35(11), 1763-1773. *doi: 10.1002/sim.6839*.

Koch, G. G., Tangen, C. M., Jung, J.-W., & Amara, I. A. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine,* **17**, 1863-1892.

Lesaffre, E., & Senn, S. (2003). A note on non-parametric ANCOVA for covariate adjustment in randomized clinical trials. *Statistics in Medicine,* **22**, 3586-3596.

Lin, W. (2013). “Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman’s Critique.” Annals of Applied Statistics **7**, 295–318.

Tsiatis AA, Davidian M, Zhang M, Lu X. (2008). Covariate adjustment for two-sample treatment comparison in randomized clinical trials; A principled yet flexible approach. *Statistics in Medicine*, **27**, 4658-4677.

Tukey, J.W. (1991). Use of many covariates in clinical trials. *International Statistical Review* **59**(2), 123-137.