Observational Studies 4 (2018) 250-259 Submitted 6/18; Published 8/18

Larry Brown: Remembrance and Connections of His Work to Observational Studies

Dylan S. Small **dsmall@wharton.upenn.edu**
Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, PA, U.S.A.

Statistics lost one of our giants when Larry Brown passed away in February, 2018 at the age of 77. Many lost a friend, a mentor and a teacher. I had the good fortune to be Larry's colleague for the past 16 years. While I wasn't as close to him as many others, Larry was my friend and I looked up to him.

Larry loved thinking. When Larry got interested in something and scratched his head, I could see his enjoyment. Steam seemed to come out as Larry was thinking and then he happily shared his thoughts.

Larry's classes and talks were full of insights. After joining the faculty at Wharton, I attended Larry's linear models first year PhD course even though I had seen the material before. I was glad I took the time to do so as I learned a lot, and seeing how Larry thought deeply through things from different perspectives (e.g., he often presented both a geometric and a statistical perspective) was memorable and inspiring. Larry often mentioned questions he had about methods or results, and directions of research he thought could be expanded upon, which I think motivated students to see statistics as a field full of open questions and research opportunities rather than a dead field.

Larry was generous with his time. Whenever I had a student for whom it was unclear which other faculty members had the expertise to serve on their dissertation committee, I suggested asking Larry because I knew he would be willing to spend time talking with the student, read the dissertation seriously and have something thoughtful to say. A few days before Larry's passing, when he knew his time was short, he was writing recommendation letters for students.

Larry spent much time on public service, and he encouraged me about its value to society even though one may not get recognition for it.

Larry worked hard. He was active in research, teaching and mentoring students and public service until his passing. The large number of Larry's former students who traveled to his funeral from places far away at short notice (Hong Kong even!) was a testament to Larry's impact on their lives. Larry also made good time for family and friends. Besides the much time spent together with his wife Linda and their family, Larry made the time for trips over the summer alone with his sons.

Larry treated people with respect and decency regardless of their status. At a time when I was the postdoctoral coordinator for our department, a PhD student from a little known university in India contacted Larry about a post doc opening in our department and

said he would be visiting Philadelphia and hoped to stop by. Larry e-mailed me that he would have met with the student but would be away in Washington for one of the many public service committees he was on, and asked me to meet with the student. Larry said that even though he doubted the student would qualify for the post doc position in our department, the student seemed earnest and that it was at least worth a little of our time to give the student advice and respect.

Larry was the dean of statistical decision theory. In Brown (1971), Larry connected whether an estimator of normal means in p dimensions is admissible to whether a stochastic process such as Brownian motion has probability one of returning to its starting point in p dimensions. What stroke of imagination and depth of understanding led Larry to think about the connection between these two phenomenona from different poles of statistics and probability? The math is beyond me, but when a math Ph.D. friend of mine claimed that statistics doesn't have any deep math, I showed him Larry's paper and my friend changed his mind.

Although Larry was most known for his work in mathematical statistics, he had broad interests in statistics and respect for all types of statisticians. Anirban Dasgupta's conversation with Larry (Dasgupta, 2005) gives a sense of Larry's wide ranging interests and contributions, including in statistical theory, statistical methods, the foundations of statistics and applied statistics. Larry mentions that although there was only one quarter course in statistics at his alma mater Cal Tech and it was taught by an algebraist, "The subject almost immediately appealed to me. I suppose that my interest in statistics was rooted in a desire to use formal mathematics in a pragmatic way." Larry fulfilled his desire well.

Larry didn't pretend to be an expert on things he wasn't, but he was willing to engage any topic in statistics and he wasn't shy to share his ideas when he thought he had something to say. When many brilliant statisticians stray from their areas of expertise, they poke holes which are hard to refute at the moment one is confronted with them, but on reflection seem to miss the heart of the matter. But for Larry, on topics I had thought a lot about, when he raised a question or concern, there was something serious to it even if I ultimately didn't fully agree with it. Observational studies was no exception. Larry recognized the importance of observational studies (as the chair of the National Research Council Committee on National Statistics, Larry supported forming panels on topics in observational studies and their applications), but he had a healthy skepticism. Larry's work on postselection inference that he was engaged in at the time of his passing was motivated in part by his concerns about the way observational studies are often conducted. Many practical observational studies are conducted by first regressing an outcome Y on some variables of interest $\mathbf{X}$, which include treatments of interest and potential confounders, selecting a model and then reporting confidence intervals for the coefficients of the selected model as if the model was chosen *a priori* rather than selected by using the data. Larry worried that some investigators fish for a model to support desired conclusions and sought a method of inference that protected against fishing. One of Larry's and his collaborators' solutions (Berk, Brown, Buja, Zhang and Zhao, 2013) is to form a confidence interval for a coefficient in the selected model that is valid regardless of the process used to select the model – they call this the Post-Selection Inference (POSI) confidence interval. For example, say there is a treatment of interest T and potential confounders $C_1, \ldots, C_p$ and one uses some model selection procedure to choose a subset of the confounders for the model and then linearly regresses the outcome

Y on T and the chosen confounders. Say the treatment and confounders are fixed but the Y is random, $Y_i \sim N(\alpha + \beta T_i + \tau_1 C_{i1} + \cdots + \tau_p C_{ip}, \sigma^2)$, $i = 1, \ldots, n$. In one sample of the data, one might choose confounders $C_2, C_3, \ldots, C_p$ (i.e, excluding C_1); denote the population regression coefficient on T when Y is regressed on $T, C_2, \ldots, C_p$ (i.e., the expected regression coefficient on T when Y is regressed on $T, C_2, \ldots, C_p$ in repeated samples) by $\beta_{T:C_2,\ldots,C_p}$, which equals $\beta + \tau_1 \theta_{T:C_1:C_2,\ldots,C_p}$ where $\theta_{T:C_1:C_2,\ldots,C_p}$ is the regression coefficient on T when C_1 is regressed on T,$C_2, \ldots, C_p$. In another sample of the data, one might choose confounders $C_1, C_3, C_4, \ldots, C_p$ (i.e., excluding C_2); denote the population regression coefficient on T when Y is regressed on $T, C_1, C_3, \ldots, C_p$ by $\beta_{T:C_1,C_3,\ldots,C_p}$, which equals $\beta + \tau_2 \theta_{T:C_2:C_1,C_3,\ldots,C_p}$ where $\theta_{T:C_2:C_1,C_3,\ldots,C_p}$ is the regression coefficient on T when C_2 is regressed on T,$C_1, C_3, \ldots, C_p$.. Let $\beta_{T_{sel}}$ be the random variable that is the population regression coefficient on T in the selected model; $\beta_{T_{sel}}$ is random because the selected model is random, e.g., $\beta_{T_{sel}} = \beta_{T:C_2,\ldots,C_p}$ when confounders $C_2, \ldots, C_p$ are selected and $\beta_{T_{sel}} = \beta_{T:C_1,C_3,\ldots,C_p}$ when confounders $C_1, C_3, \ldots, C_p$ are selected. Larry and his collaborators' 95% POSI confidence interval is an interval that has a $\geq 95\%$ chance of containing $\beta_{T_{sel}}$ regardless of how the model selection is done.

As an example of POSI in action, consider what is the effect of a man serving in the military (during the mid 1950s/early 1960s) on the man smoking, specifically pack years smoked by age 53 (the sum of average packs smoked per day for each year of a person's life up to age 53)? See Bedard and Deschênes (2006) for background on the effect of the military on smoking and its public health implications and see our supplementary materials for the data set we consider and a description of it. Potential confounders that are available in the data include father's and mother's education, family income, whether the man attended a Catholic school, high school class rank, teacher's evaluation of whether the man was an outstanding student and iq in ninth grade. If we first fit a linear regression model of pack years smoked on a military service dummy variable and the potential confounding variables, the significant ($p < 0.05$) confounders are class rank and iq. If we then fit a linear regression model of pack years smoked on the military service dummy, class rank and iq, then we estimate that military service increases pack years smoked by 3.4; the regression in R spits out

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     10.07478    3.82973   2.631 0.008572
veteran          3.41489    1.03281   3.306 0.000958
class.rank      -0.17631    0.02296  -7.678 2.28e-14
iq               0.14369    0.04171   3.445 0.000581
Residual standard error: 25.94 on 2589 degrees of freedom
Confidence Interval:
                      2.5 %     97.5 %
(Intercept)      2.56515066 17.5844187
veteran          1.38967285  5.4401145
class.rank      -0.22133338 -0.1312787
iq               0.06189378  0.2254852
```

The reported confidence interval for veteran of $(1.4, 5.4)$ is the usual linear regression confidence interval, estimate $\pm$ $t(.975, \text{degrees of freedom})$ (≈ 1.96) times the Std. Error. But this is not right – it pretends that we knew the model before looking at the data whereas we actually chose the model based on the data. The PoSI confidence interval, computed from the R package PoSI, uses a multiplier of 3.06 rather than 1.96 for the Std. Error, yielding a confidence interval of $(0.3, 6.6)$ for the effect of military service on pack years smoked. This is considerably wider than the usually reported confidence interval of $(1.4, 5.4)$ but still provides evidence that military service increased smoking. When model selection has taken place, the usually reported confidence interval is not valid in any sense because it ignores the effects of selection. The PoSI confidence interval is universally valid for $\beta_{T_{sel}}$ regardless of how the model was selected. A different approach is to seek a limitedly valid confidence interval that is valid if I followed exactly the model selection approach I *said* I took: first fit a linear regression model of pack years smoked on a military service dummy variable and the potential confounding variables, choose the $p < 0.05$ confounders and then fit a linear regression of pack years smoked on military service and the chosen confounders (work in the direction of constructing confidence intervals that are valid under a pre-specified model selection procedure includes bootstrap methods (e.g., Efron, 2014) and selective inference approaches (e.g., Taylor and Tibshirani, 2015)). Larry questioned whether such a limitedly valid approach would capture what really happened in the data analysis. I might have done some preliminary data analysis that I failed to report, perhaps because it wasn't tidy. Berk, Brown, Buja, Zhang and Zhao (2013) wrote, "There is a general question about the wisdom of proposing ever tighter confidence and retention intervals for practical use when in fact these intervals are valid only under tightly controlled conditions. It might be realistic to suppose that much applied work involves more data peeking than is reported in published articles." Such data peeking may be good statistical practice. Statisticians take pride in plotting the data before fitting a complicated model. However, Larry was concerned that the failure to account for such good exploratory data analysis leads to spurious findings. Buja and Brown (2014) wrote:

> Empirical research suffers from a systemic malady that is well reflected by Ioannidis' (2005) piece with the provocative yet realistic title "Why most published research findings are false." The culprit of first order is most likely publication bias, also called the "file drawer problem," that is, the fact that negative results tend not to see the light of publication. A culprit of second order we hypothesize to be the fact of unaccounted data analytic activity, ranging from meta-selection among variable selection methods to the use of informal EDA and diagnostics methods. It may just be the case that the most expert and thorough data analysts are also the ones who produce the most spurious findings in applied statistical work. This should not be construed as a call to apply less competence and abandon research into efficient statistical methods, but it should be motivation to create statistical inference that integrates ever more of the informal data analytic activities for which there is currently no accounting.

PoSI provides one way to integrate informal data analytic activities into inference. "With inference that is universally valid after any model selection procedure [PoSI], we have a way

to establish which rejections are safe, irrespective of unreported data peeking as part of selecting a model." (Berk et al., 2013).

From my experience working on collaborative projects, I share Larry's concern that data analysis is typically more interactive than described in a published paper and I worry about how my and my collaborators' preconceptions affect the results. Berk et al. described post hoc decisions that are often used in analyzing data like, "in retrospect only one of these two variables should be in the model." How are these post hoc decisions affected by our preconceptions about what the data should show? Dante wrote of St. Thomas Aquinas cautioning him, "opinion – hasty – often can incline to the wrong side, and then affection for one's own opinion binds, confines the mind." (Alighieri, trans. Mandelbaum). However, PoSI is allowing for not only cognitive biases but malice – it protects against an investigator who searches over all models to find the one that is most significant for the treatment of interest. I wonder what intermediate methods can be developed that assume an investigator has honest intentions but human cognitive flaws. Methods that allow the investigator to interact with the data before making inferences for the results of interest are one approach, for example constructing a matched observational study using only the covariates and treatment before looking at the outcomes to make inferences about the treatment effect (Rubin, 2007; Deshpande et al. 2017 provide an application of this approach).

Our discussion of the effect of military service on smoking example above has assumed fixed covariates, which was the setting Larry and collaborators assumed in Berk et al. (2013). In subsequent work, Larry and collaborators have made progress in extending PoSI to the random covariates setting, e.g., Kuchibhotla, Brown, Buja, George and Zhao (2018), work led by Larry's last Ph.D. student Arun Kumar Kuchibhotla. Larry argued that observational studies should assume random, not fixed, covariates. Typically we care not so much about the study subjects themselves, for their treatments have already been given, but future subjects on whom we might implement the results of the study. If the treatment effect is constant and we have a correctly specified regression model that includes all confounders, then even if we regard the covariates as random, the distribution of the covariates does not depend on the treatment effect and ancillarity of the covariates argues for making inferences conditional on the covariates, i.e., treating the covariates as fixed; see Savage (1976, p. 468) for a discussion of how both Bayesian and frequentist statisticians have come to this conclusion. Suppose instead the treatment effect is heterogeneous and we are interested in the average treatment effect, e.g., $E[Y^{(t)}|\mathbf{C}] = \boldsymbol{\delta}^T\mathbf{C} + \boldsymbol{\kappa}^T\mathbf{C}$ and $E[Y^{(c)}|\mathbf{C}] = \boldsymbol{\kappa}^T\mathbf{C}$ where $\mathbf{C}$ are all the confounders, $\boldsymbol{\delta}^T\mathbf{C}$ is the treatment effect for a subject with confounders $\mathbf{C}$ and $Y^{(t)}$, $Y^{(c)}$ are the potential outcomes under treatment and control respectively, so that the model for the observed data is

$$E[Y|T,\mathbf{C}] = T\boldsymbol{\delta}^T\mathbf{C} + \boldsymbol{\kappa}^T\mathbf{C} \tag{1}$$

and the average treatment effect is $E[\boldsymbol{\delta}^T\mathbf{C}]$. The average treatment effect now depends on the distribution of $\mathbf{C}$ and ancillarity breaks down. Buja, Berk, Brown, George, Pitkin, Traskin, Zhang and Zhao (2014, §4.1) discuss this breakdown of ancillarity from the perspective of misspecified regression models. For example, we might fit a linear regression model of the outcome on the treatment and the potential confounders $\mathbf{C}$, $\hat{Y} = \hat{\theta}T + \hat{\boldsymbol{\kappa}}^T\mathbf{C}$. The expected coefficient on the treatment when the true model is (1) is then a variance weighted av-

erage of the treatment effect at the different $\mathbf{C}$'s, $E[\hat{\theta}] = E_{\mathbf{C}}[Var(T|\mathbf{C})\boldsymbol{\delta}^T\mathbf{C}]/E_{\mathbf{C}}[Var(T|\mathbf{C}]$ (Angrist and Krueger, §2.3.1). The parameter $E[\hat{\theta}]$, which is estimated by $\hat{\theta}$ now depends on the distribution of $\mathbf{C}$ and ancillarity breaks down. Buja et al. (2014) advocate the use of sandwich standard errors or the x-y bootstrap (i.e., bootstrapping pairs of outcomes and treatment/covariates) which treat the covariates as random to obtain asymptotically correct standard errors.

In light of the breakdown of ancillarity with random covariates and heterogeneous treatment effects, a reviewer raised concerns about the strategy discussed in the last sentence two paragraphs above of constructing a matched observational study using only the covariates and treatment with the outcomes locked away and then making inferences conditional on the covariates and treatment. I agree with Larry that ideally observational studies should make inferences for future populations of subjects to whom the study results might be applied. But convincing observational studies often require studying specially chosen samples rather than random samples from natural populations (Rosenbaum, 1999). For example, consider the question, does malnutrition in a mother's womb reduce a person's cognitive abilities as an adult? This is hard to study in the whole population because mothers who are malnourished during pregnancy differ in many ways from those who are well nourished. A brutal natural experiment occurred in World War II. In September 1944, the Allies sought to force a bridgehead through the Rhine by linking paratroopers who landed in Arnhem, Holland with advancing troops. The attempt failed and in reprisal, the Nazis imposed a transportation embargo on Western Holland, causing a famine during the winter of 1944-1945 in affected areas in Western Holland but not other parts of Holland. When males in utero in Holland in 1944-1945 reached the age of 18, they were required to take psychological and medical examinations as part of induction into the military. Stein, Susser, Saenger and Marolla (1972) compared the results of these examinations for males who were in utero in a famished area in 1944-1945 to males who were in utero in a non-famished area in Holland in 1944-1945. Stein et al.'s study is only about Dutch men. We would ideally like to consider a population which is more representative of those who might experience malnutrition in the womb, including women and people from different racial and ethnic groups. But in most representative situations, malnutrition is tangled with other confounding factors. If we used the data from the Stein et al. study to report a confidence interval for the effect of malnutrition for a representative population that includes both men and women, it would be $(-\infty, \infty)$ because the study provides no information about women. However, I think reporting the confidence interval for the population we can study in the Stein et al. study, Dutch men, and being clear about what population is being studied, is useful.[1]

Larry and I had some lively discussions about the usefulness of the PoSI confidence interval for observational studies. I argued that what we care about for observational studies is β, the coefficient on the treatment T after controlling for all confounders $C_1, \ldots, C_p$ and not $\beta_{T_{sel}}$ which does not control for confounders which the model has excluded. Consider the following setting: there are 250 observations and 50 mutually independent covariates $C_1, \ldots, C_{50}$ each having a standard normal distribution, the binary treatment variable is $T = I(\frac{1}{\sqrt{50}}C_1 + \ldots + \frac{1}{\sqrt{50}}C_{50} \geq 0)$ and the true regression model is $E(Y|C_1, \ldots, C_{50}, T) = 0.08C_1 + 0.08C_2 + \cdots + 0.08C_{50}$ and $Var(Y|C_1, \ldots, C_{50}, T) = 1$ (this setting was suggested

1. The discussion of this example is drawn from Rosenbaum (2017).

to me by Dean Foster). The truth is there is no treatment effect. We use backward stepwise regression to select a model in the following way: (i) use backward stepwise model selection with the AIC criterion to choose a regression model for $E(Y|C_1,\ldots,C_{50})$; (ii) add the treatment variable to the regression model fit in Step (ii) and (iii) test whether there is evidence of a treatment effect at the 0.05 level. In 1000 simulations, this procedure resulted in an average treatment effect estimate of 0.25 and evidence of a treatment effect 36% of the time – variable selection has greatly exacerbated the probability of making a Type I error. About half of the confounders were excluded on average. Variable selection is causing "death by 1000 cuts" – exclusion of each confounding variable on its own does not cause much bias but the combined exclusion of many confounders causes considerable bias. For PoSI, the number of variables is too big to use the PoSI R package to obtain the exact minimum PoSI multiplier but we can use an upper bound, the Scheffé constant, $\sqrt{dF_{d,r,.95}}$ where d is the number of variables being considered, r is the corresponding degrees of freedom for the error when the maximum number of variables is considered and $F_{d,r,.95}$ is the .95 quantile of the F distribution with d, r degrees of freedom (Berk et al., 2013, §4.8). The Scheffé constant is ≈ 8.5 in our setting, meaning that the t-statistic on the treatment in the selected model would have to exceed 8.5 for PoSI to find evidence of a treatment effect, which it never did in 1000 simulations. PoSI correctly finds no evidence of a treatment effect, correcting for the overoptimism of naïve inference after variable selection. However, I find it a little unsettling that PoSI's target of inference $\beta_{T_{sel}}$ has mean 0.25 whereas the causal effect that I care about is $\beta = 0$.

The "death by 1000 cuts" example in the above paragraph is an instance of the general phenomenon that it is not possible to give "good" inference for a full model regression coefficient after doing model selection that has been illuminated by Leeb and Pötscher in a series of papers. For example, Leeb and Pötscher (2005) present results supporting that

1. Regardless of sample size, the sampling properties of post-model-selection estimators are typically significantly different from the nominal distributions that arise if a fixed model is supposed. Consequently, using standard t-intervals (as if the selected model had been given prior to the statistical analysis) can be highly misleading;

2. The finite-sample distributions of post-model-selection estimators [for the coefficients in the full model] are typically complicated and depend on unknown parameters. Estimation of these finite-sample distributions is "impossible" (even in large samples). No resampling scheme whatsoever can help to alleviate this situation.

Influenced by Leeb and Pötscher's results, in my collaborative work, I have discouraged doing model selection. In one study, a collaborator wanted to do model selection for estimating a treatment effect with about 50 potential confounders and about 2000 observations. I said, "This is not a good idea – we wouldn't gain much power by doing model selection and we will lose being able to make accurate inferences." In the setting in the above paragraph but with a treatment effect of 0.7, even with only 250 observations, not much power is lost by fitting the full model with 50 potential confounders – see Figure 1. But there is a limit to how much model selection can be avoided. Of course, if there are more variables p than observations n , model selection cannot be avoided and even when $p < n$, the power can start to drop severely as p comes close to n as Figure 1 shows. Even if there are only 50 potential

confounders, the potential confounders may have nonlinear effects so that if we consider powers and regression splines of the potential confounders, p will start to come close to or exceed n. So, avoiding model selection is no panacea. Larry and his collaborators were also influenced by Leeb and Pötscher's results and earlier results on problems associated with post-model-selection inference (see pg. 804 of Berk et al., 2013), but rather than seeking to avoid model selection, they have sought to make valid inferences for the coefficient on the treatment in the selected model, $\beta_{T_{sel}}$, rather than the coefficient on the treatment in the full model β which corresponds to the causal effect if the full model is correct. Part of Larry and his collaborators' motivation for shifting the focus from β to $\beta_{T_{sel}}$ is that they think the full model that is specified (e.g., in my work with my collaborator described above, a model that assumes the 50 potential confounders have linear effects) is often not believable. Buja, Berk, Brown, George, Kuchibhotla and Zhao (2016) cite Tukey, "The hallmark of good science is that it uses models and 'theory' but never believes them." (J.W. Tukey, cited by D. Brillinger, Toronto, 2016). But Tukey also said, "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." (Tukey, 1962). My view is that the coefficient on treatment in the full, correctly specified model with all confounders is the "right" question for observational studies that aim to make causal inferences about the effect of treatment.[2] Larry understood my argument (see Berk, Brown, Buja, Zhang and Zhao, 2017, Section 4.3) but thought my conception of observational studies was rather narrow. Berk, Brown, Buja, Zhang and Zhao (2013) wrote, "Generally, in any creative observational study involving novel predictors, it will be difficult a priori to exclude collinearities that might force a rethinking of the predictors."

I wish I could talk more with Larry about model selection in observational studies and other topics. I miss Larry and his love of statistics, kindness, intellect and decency. Larry's good influence on me and many in the statistical community will live on.

Acknowledgement: I thank the editor for this article, Ben Hansen, and the reviewers for insightful comments.

References

Alighieri, Dante (1320). Paradiso canto XIII: 118–20. Trans. Allen Mandelbaum (1995).

Angrist, J. D. and Krueger, A. B. (1999). Empirical strategies in labor economics. In *Handbook of labor economics* (Vol. 3, pp. 1277-1366). Elsevier.

Bedard, K. and Deschênes, O. (2006). The long-term impact of military service on health: Evidence from World War II and Korean War veterans. *American Economic Review*, 96, 176-194.

Berk, R., Brown, L., Buja, A., George, E. and Zhao, L. (2017). Working with misspecified regression models. *Journal of Quantitative Criminology*, DOI 10.1007/s10940-017-9348-7.

2. If there is treatment effect heterogeneity, then I would be interested in the coefficients of the treatment interacted with effect modifiers and/or the average treatment effect. For example, suppose $E[Y^{(t)}|\mathbf{C}] = \boldsymbol{\delta}^T\mathbf{C} + \boldsymbol{\kappa}^T\mathbf{C}$ and $E[Y^{(c)}|\mathbf{C}] = \boldsymbol{\kappa}^T\mathbf{C}$ where $\mathbf{C}$ are all the confounders and $Y^{(t)}$, $Y^{(c)}$ are the potential outcomes under treatment and control respectively, so that the model for the observed data is $E[Y|T,\mathbf{C}] = T\boldsymbol{\delta}^T\mathbf{C} + \boldsymbol{\kappa}^T\mathbf{C}$ where T is an indicator variable for treatment, then I would be interested in $\boldsymbol{\delta}$ or perhaps as a summary measure the average treatment effect over the sample $\sum_{i=1}^{n}\boldsymbol{\delta}^T\mathbf{C}_i$

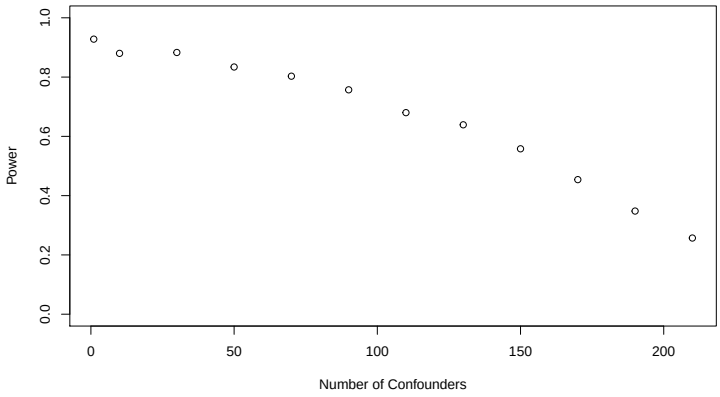

Figure 1: Power for testing hypothesis of no treatment effect when there are 250 observations and p mutually independent potential confounders $C_1, \ldots, C_{50}$ each having a standard normal distribution, the binary treatment variable is $T = I(\frac{1}{\sqrt{50}}C_1 + \ldots + \frac{1}{\sqrt{50}}C_{50} \geq 0)$ and the true regression model is $E(Y|C_1, \ldots, C_{50}, T) = 0.7T + 0.08C_1 + 0.08C_2 + \ldots + 0.08C_{50}$ and $Var(Y|C_1, \ldots, C_{50}, T) = 1$

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics*, 41, 802-837.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Annals of Statistics*, 42, 855-903.

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Traskin, M., Zhang, K. and Zhao, L., 2014. Models as approximations, Part I: A conspiracy of nonlinearity and random regressors in linear regression. arXiv preprint arXiv:1404.1578.

Buja, A., Berk, R., Brown, L., George, E., Kuchibhotla, A. K. and Zhao, L. (2016). Models as Approximations – Part II: A General Theory of Model-Robust Regression. arXiv preprint arXiv:1612.03257.

Buja, A. and Brown, L. D. (2014). Discussion: "A Significance test for the LASSO." *Annals of Statistics*, 42, 509-517.

Brown, L., and DasGupta, A. (2005). A Conversation with Larry Brown. *Statistical Science*, 20, 193-203.

Deshpande, S.K., Hasegawa, R.B., Rabinowitz, A.R., Whyte, J., Roan, C.L., Tabatabaei, A., Baiocchi, M., Karlawish, J.H., Master, C.L. and Small, D.S. (2017). "Association of playing high school football with cognition and mental health later in life." *JAMA Neurology*, 74, 909-918.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109, 991-1007.

Ioannidis, J.P.A. (2005). Why most published research findings are false. *Chance*, 18, 4047.

Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I., and Zhao, L. (2018). Valid post-selection inference in assumption-lean linear regression. arXiv preprint arXiv:1806.04119.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21, 21-59.

Rosenbaum, P. R. (1999). Choice as an alternative to control in observational studies. *Statistical Science*, 14, 259-278.

Rosenbaum, P.R. (2017). *Observation & Experiment: An Introduction to Causal Inference*. Harvard University Press.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26, 20-36.

Savage, L. J. (1976). On rereading R.A. Fisher. *Annals of Statistics*, 4, 441-500.

Stein, Z., Susser, M., Saenger, G. and Marolla, F. (1972). Nutrition and mental performance. *Science*, 178, 708-713.

Taylor, J. and Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112, 7629-7634.

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, 33, 1-67.