The Thin Reed: Accommodating Weak Evidence for Critical Parameters in Cost-Benefit Analysis

David L. Weimer (University of Wisconsin–Madison)

Working Paper prepared for:

Methods for Research Synthesis:
A Cross-Disciplinary Workshop

Harvard Center for Risk Analysis

October 3, 2013

www.hcra.harvard.edu

Corresponding author: weimer@lafollette.wisc.edu

Disclaimer: The findings and conclusions of this paper are those of the author and do not imply endorsement by any component of Harvard University or other sponsors of this workshop. Comments should be directed to the author.

Acknowledgements: I thank Marc Ratkovic for advice and assistance on an earlier version of this paper. I also thank Dana Mukamel for providing helpful comments on this draft. Any errors and all interpretations are my own.

The Thin Reed: Accommodating Weak Evidence for Critical Parameters in Cost-Benefit Analysis

David L. Weimer

Abstract

Policy analysis often demands quantitative prediction. This is especially the case in of cost-benefit analysis, which requires the comprehensive quantification and monetization of all valued policy impacts. For all but the simplest of policies, achieving comprehensiveness requires analysts to take parameter values and shadow prices from statistical analyses done by others. These parameter estimates are uncertain and usually come with estimates of their precisions. Using this information to assume distributions for all parameters and shadow prices, Monte Carlo simulation provides an appropriate way to create a distribution of net benefits that conveys the level of certainty about the fundamental question of interest: Are net benefits positive? However, there is considerable controversy about how to move from empirical estimates to the distributions of parameters needed for the cost-benefit analysis. Unfortunately, most social science researchers frame their work in terms of hypotheses about the particular parameters, uncritically privileging Type I over Type II error. The inappropriate focus on hypothesis testing rather than prediction sometimes leads analysts to treat statistically insignificant coefficients as if they, and their standard errors, are zero. One alternative method is to use all estimates and their standard errors whether or not the estimates are statistically significant. Another alternative method is to use all estimates but to shrink them and standard errors toward zero in an effort to guard against regression to the mean. Comparing the three methods (only use statistically significant estimates and their standard errors, use all estimates and their standard errors, use shrunk estimates and shrunk standard errors) in Monte Carlo simulation suggests that treating statistically insignificant coefficients as zero rarely minimizes the mean squared error of prediction. Using shrunk estimates appears to provide a more robust minimization of the mean squared error of prediction. These results suggest the heuristic that, when confronted with a necessary but statistically insignificant estimate of a necessary parameter, shrink it and use it! Indeed, the simulations presented here suggest that routinely shrinking estimates is a robust approach in the face of relatively uninformed priors about the true values.

Key words: Shrinkage estimators, Monte Carlo simulation, cost-benefit analysis

1.0 Introduction

Demand for cost-benefit analysis (CBA) to assess the efficiency of public policies has grown in recent years. Whereas it was once employed almost exclusively to assess infrastructure projects, it now sees routine use in environmental policy, and increasingly in health and social policy (Weimer and Vining, 2009). President Reagan's Executive Order 12291 and President Clinton's Executive Order 12866 expanded the use of CBA by federal agencies. Although not yet widely used by state governments, the value of its sophisticated application to social policy has been demonstrated by the Washington State Institute for Public Policy, which routinely conducts CBAs at the request of the state legislature. As the validity of CBA as an assessment of efficiency depends on its comprehensive accounting of valued impacts, analysts seeking to apply it to any but the simplest policies must usually glean effect sizes and shadow prices from empirical work done by others or expediently by themselves. These empirical estimates are never certain. How should analysts make use of these estimates to predict costs and benefits? How should analysts appropriately convey the uncertainty that the use of these estimates as a basis for prediction creates in net benefits?

The answer to the second question is obvious, but not uniformly followed: every CBA should employ Monte Carlo simulation to produce distributions of net benefits that take account of uncertainty in effect sizes and shadow prices. (For illustrations, see Nicol, 2001; Weimer and Sager, 2009). The distribution of net benefits is the basis for "testing" hypotheses about net benefits. It provides a correct estimate of expected net benefits, which may differ from that resulting from calculations based on point-estimate values of parameters when uncertain effects are multiplied by uncertain shadow prices. Most importantly, it conveys useful information not only about expected net benefits, but also about how likely it is that positive net benefits will be realized.

The focus of this essay is on answering the first question, which has received virtually no attention by policy analysts. Moving from empirical estimates to appropriate predictions would seem to be one of the fundamental tasks for not just CBA but for any policy analysis with quantitative components such as risk analysis or cost-effectiveness analysis. Yet, common practice tends to make two types of errors. The first is to apply naively concepts of hypothesis testing in drawing on empirical evidence. As social scientists, our norms require us to give great deference to the convention of assuming parameters are zero unless we can claim that the

probabilities of their Type I errors are no more 5 percent (or 10 percent for the more accepting). As policy analysts, however, we want the best prediction, which is often closer to the estimated but not-statistically-significant parameter than it is to zero. The second error is to confuse efficient estimation with accurate prediction. An unbiased and efficient estimator does not necessarily give the most appropriate prediction. Indeed, in circumstances in which the null hypothesis reflects beliefs that the true parameter may actually be zero, parameter estimates from ordinary least squares or logistic regressions tend to make predictions that are too large from the perspective of minimizing mean squared error and therefore may appropriately be "shrunk" toward zero.

Recent years have witnessed increased attention to the synthesis of data relevant to particular predictions through systematic techniques such as meta-analysis. Here I am addressing a different sort of synthesis: drawing together estimates of sets of parameters necessary to complete an informative policy analysis. Especially for policies with multiple effects, such analysis often requires synthesizing across many sources, including sometimes individual studies with imprecise results. My focus is on this latter sort of synthesis, specifically when a parameter is based on the thin reed of a single study.

2.0 Problems with Current Evidence-Based Approaches

Imagine that an analyst wanted to do a CBA to answer the question of whether the replication of an experimentally evaluated program would produce positive net benefits. Also assume that the experiment was perfectly implemented and produced estimates of a dozen impacts relevant to the CBA. For example, the impacts might include changes in school achievement, participation in juvenile delinquency, and high school completion. Now suppose that eight of the estimated impacts would be starred to show statistical significance at the 5 percent level based on their individual t-tests. However, following the *What Works Clearinghouse Procedures and Standards Handbook* (Institute of Education Sciences, 2008), the researchers employ the Benjamini and Hochberg (1995) method for adjusting critical values to ensure that the Type I error for the collection of estimated impacts remains at 5 percent (actually, the probability of rejecting any one hypothesis when all the nulls are true), finding that only six of the impacts remain statistically significant. (They are nonetheless delighted with this method

because using the less powerful but commonly employed Bonferroni correction would have only resulted in four statistically significant impacts!)

The researchers might follow one of the following three approaches to moving from their experimental results to a prediction of net benefits.

First, they could embrace what Stephen Ziliak and Deirdre McCloskey (2008) refer to as the cult of statistical significance and only use the estimates and standard errors for the eight estimates of impacts that were individually statistically significant in their CBA. As a consequence, they would be assuming that the other four impacts were *exactly* zero. That is, because there was more than a 5 percent chance that the true value of these statistically insignificant impacts is zero, they ignore their estimated magnitudes and standard errors. These assumptions of zero impact very likely bias their estimate of net benefits toward zero. Further, by assuming their standard errors are also zero, their Monte Carlo analysis will produce a distribution of predicted net benefits that does not convey the true level of uncertainty.

Second, they could be even more diligent followers of the cult and base their CBA only on the six statistically significant impacts after taking account of multiple comparisons. The result would be even more bias and a greater underestimation of uncertainty in their prediction of net benefits for the replication.

Third, realizing that the hypothesis of interest is *whether replicating the program would produce positive net benefits*, they decide to use their estimates of all twelve impacts, both those that are statistically significant and those that are not. To predict the distribution of net benefits, they use all the estimates and their standard errors in their Monte Carlo analysis. They have escaped the cult of statistical significance! However, by treating estimates as predictions, their approach tends to bias their predictions of net benefits away from zero, especially when the true effects are small.

3.0 Inference and Prediction

Under the canonical assumptions (non-random regressors and identically and independently distributed errors with zero mean and constant variance), ordinary least squares (OLS) produces the unbiased estimates with the smallest possible standard errors. The correction we will consider introduces a useful form of bias. Published estimates are generally unbiased in finite samples (least squares) or asymptotically (maximum likelihood methods, i.e.

logit or probit). Unbiasedness is desirable in questions of inference, especially where the focus is on whether or not to reject a null of no effect. The issues in proper inference are daunting (see, e.g., Anderson, 2008; Perneger, 1998), and are central to retrospective program evaluation. CBA asks a different question. Rather than assessing a past policy, asking what costs and benefits appear to have been actually realized, CBA attempts to *predict* the costs and benefits that a policy would produce if it were to be adopted.

This changes the problem from one of inference to one of prediction. If the goal is optimal prediction, then some level of bias may be desirable. We are not concerned with an unbiased estimate of what happened in the data set analyzed by a study, but instead we wish to estimate the benefits in some future replication of the policy. The basic theoretical insight arises from the use of "shrinkage" in reducing the mean squared error of prediction. Shrinkage methods introduce some bias towards zero in coefficient effect sizes, which, through the bias-variance trade off, results in better predictions. Basically, we adjust effect size in order to account for regression to the mean.

4.0 Using Shrinkage to Reduce Predictive Error

The basic insight for reducing predictive error dates back to Francis Galton, who plotted the mid-heights of fathers and mothers versus offspring (Galton 1886). He observed that particularly tall parents, on average, had children who were taller than average but not as extremely so as the parents, while particularly short parents had children who were still short, but less extremely so than the parents. Galton termed this phenomenon of heights "Regression toward Mediocrity in Hereditary Stature."

Translated into modern parlance, a given observation or effect size has two components: a systematic and a random component. When predicting the value of the effect size for a new draw of data, the best unbiased predictor (BUP), in the sense of minimizing mean squared predictive loss, will shrink the unbiased estimate towards the overall mean of all observations. In statistics, this is referred to as Stein's Paradox, which states that the unbiased estimate (i.e., a sample mean or regression coefficient) can always be transformed into a better predictor by accounting for the random component that generated the data. In Galton's example, the grand mean of all parental midpoints carried information about the height of any given child, because

these heights shrink towards the mean, even though the parents are independent observations. Combining means across parents can lead to a better estimator than any given parents' mean.

Shrinkage estimates work by balancing some data-specific value and some grand mean, with weights determined by the sample variances of each. A large literature discusses the properties of this approach (Thompson 1968; Efron and Morris 1971, 1975; Stein 1981; Copas 1983, 1987). This essay briefly assesses the literature on shrinkage estimates and how they should be used in bringing thin evidence into CBA and other types of quantitative analyses. It employs simulations to illustrate the implications of the following approaches to dealing with tenuous links, especially statistically insignificant effects:

1. Treating statistically significant effects and their standard errors as zero.
2. Using all estimates and their standard errors.
3. Using shrunk estimates that would be available in secondary analysis.

5.0 Simple Introduction to the Theory of Shrinkage

The basic idea behind shrinkage is that by accepting some bias, the mean squared error of prediction can be reduced. Consider an unbiased and efficient estimator, such as an OLS coefficient, $\hat{\theta}$, under the canonical assumptions. Although $\hat{\theta}$ has zero bias and the smallest possible variance among the class of linear unbiased estimators, it may not have the smallest mean square error, $\text{MSE} \equiv \text{E}[(\hat{\theta} - \theta)^2]$, where θ is the true value of the parameter. Following Thompson (1968), consider a "natural origin" for θ of θ_0. One can think of this as a Bayesian prior—in our applications we assume that it is zero. The shrinkage estimator is derived by finding the value of c that minimizes the following expression:

$$\text{E}[(c(\hat{\theta} - \theta_0) - (\theta - \theta_0))^2]$$

Applying *c* to the estimator will yield a smaller MSE around the natural origin, but a larger MSE far away from it.

Taking the derivative with respect to c and solving yields the following equation for *c*:

$$c = \frac{(\theta - \theta_0)^2}{(\theta - \theta_0)^2 + Var(\hat{\theta})}$$

We can estimate *c* as

$$\hat{c} = \frac{(\hat{\theta} - \theta_0)^2}{(\theta - \theta_0)^2 + \hat{V}ar(\hat{\theta})}$$

Letting θ_0=0, Thompson gives the shrinkage formula for estimating the mean, μ, from a random sample as:

$$\hat{\mu} = \frac{\bar{x}^2}{\bar{x}^2 + s^2/n}\bar{x}$$

where $\bar{x}$ is the sample mean, *n* is the number of observations, and s^2 is the estimate of the population variance. A similar formula was derived by Copas (1997) for OLS regression and MLE logistic regressions. We take a similar approach to derive a shrinkage formula for a parameter estimate from an ordinary least squares regression.

For our purposes, we consider shrinkage of the OLS estimator based on its Student's t statistic, which can be calculated from the estimate and its standard error, which are routinely reported by researchers. Seeking to minimize error when the true parameter is zero, we want to choose *k* to minimize:

$$MSE_k = E[(k\hat{\beta}_{OLS} - \beta)^2]$$

Taking the partial derivative with respect to *k* and solving the first order condition for an extreme value yields:

$$k = \frac{\beta^2}{Var(\hat{\beta}_{OLS}) + \beta^2}$$

Substituting $\hat{B}_{OLS}$ for β and rearranging gives:

$$\hat{k} = \frac{\hat{\beta}_{OLS}^2}{Var(\hat{\beta}) + \hat{\beta}_{OLS}^2}$$

Dividing through by $Var(\hat{\beta})$ yields:

$$\hat{k} = \frac{t^2}{t^2 + 1}$$

Therefore, the shrunken value is given by:

$$\hat{\beta} = \left[\frac{t^2}{t^2 + 1}\right]\hat{\beta}_{OLS}$$

where *t* is the Student's *t* statistic for the test of the hypothesis that the coefficient of the independent variable is zero. Note that the shrinkage declines as *t* increases. Using the Delta method,[1] we can approximate the variance of $\hat{\beta}$ as:

$$Var(\hat{\beta}) = \left[\frac{(t^8 + 6t^6 + 9t^4}{(t^2 + 1)^4} \right] Var(\hat{\beta}_{OLS})$$

The simulations that follow assume that coefficients from OLS regressions are needed to complete a comprehensive CBA.

6.0 Structure of the Simulations

Figure 1 provides an overview of the simulation process. A simulation trial begins by drawing "true values" of three parameters that sum to total net benefits (Step 1). These true values and random errors are used to create a data set relating the true parameters to observed outcomes (Step 2). OLS regressions are then used to estimate the three parameters and their standard errors (Step 3). Next, the three decision rules are applied: (1) assume statistically insignificant coefficients and their standard errors are zero; (2) use all coefficients and their standard errors; and (3) use shrunken coefficients and their shrunken standard errors (Step 4). Finally, the coefficients and standard errors resulting from each rule are applied in a Monte Carlo simulation of the sort that would be done by analysts to estimate a distribution of predicted net benefits (Step 5). Two summary statistics for each Monte Carlo simulation are recorded: first, the squared deviation of the true value of net benefits and the mean predicted value; and second, whether or not the true value of net benefits is smaller than the 5 percent point of the cumulative distribution of predicted net benefits or is larger than the 95 percent point; in other words, whether the true value falls outside the derived 90 percent confidence interval.

These five steps constitute one "trial." The investigation involves executing a large number of trials.

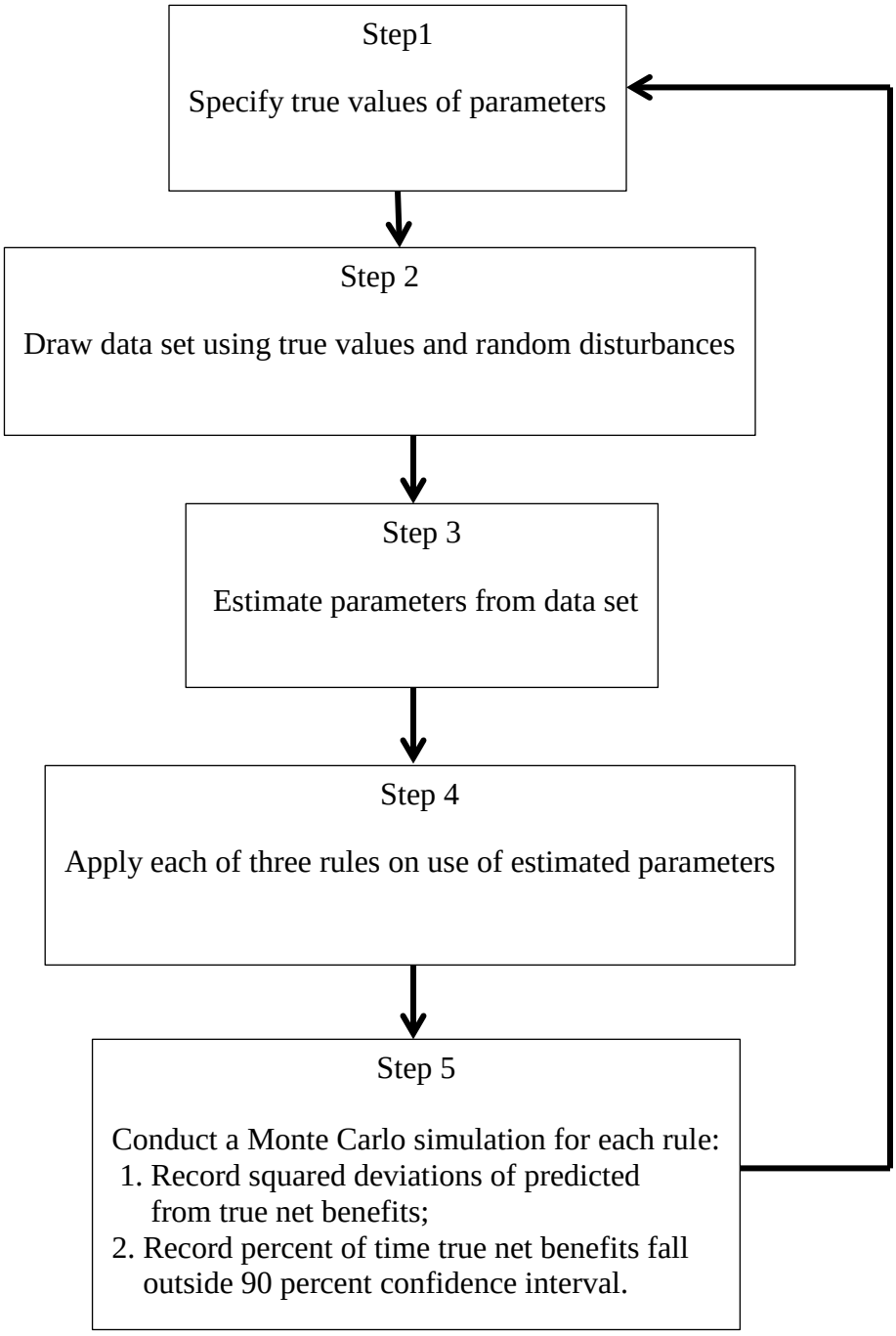
In the trials reported on here, the true but unobserved parameters are represented as three impacts measured in standardized units that correspond, say, to millions of dollars. Impact 1 is assumed to have a 20 percent chance of taking the value 0 and an 80 percent chance of a value

---

[1] $g(\hat{\beta}) = \left[\frac{\hat{\beta}_{OLS}^3 / Var(\hat{\beta}_{OLS})}{\hat{\beta}_{OLS}^2 / Var(\hat{\beta}_{OLS}) + 1} \right]$ and $Var(\hat{\beta}) = \left[dg(\hat{\beta}) / d\hat{\beta} \right]^2 Var(\hat{\beta}_{OLS})$.

drawn from a uniform distribution from 0 to 1. Impact 2 is assumed to have a similar distribution, but with a 40 percent chance of taking a value of 0 and a 60 percent chance of a

Figure 1: Overview of Simulation Procedure



value drawn from a uniform distribution from 0 to 1. Impact 3 is assumed to be uniformly distributed over the range -0.4 to 1. The true net benefit is assumed to be the sum of these three impacts, so that it ranges from -0.4 to 3, with an expected value of 1.

We generate three data sets each with 100 observations split equally between the treatment and control cases.[2] Each dependent variable equals the error drawn from a normal distribution for control cases and a draw from a normal distribution plus the impact for treatment cases. One can think of the dependent variables as the impact measured with error so that each regression provides an estimate of one of the impacts and a constant, where the true value of the constant is zero.

The regression analysis applies a two-sided t-test to the estimated coefficients. Those that are statistically insignificant at the 5 percent level are set equal to zero (and their standard errors are set equal to zero) under rule 1. Rule 2 simply uses the full set of estimated coefficients and their standard errors. Rule 3 also uses all coefficients and their standard errors, but shrinks each of them using its t-statistic in the shrinkage equations previously displayed.

The Monte Carlo simulation creates a distribution of predicted net benefits for each decision rule by drawing values from Student's t distributions with means equal to the values of the coefficients and standard errors equal to the values of the standard errors employed under that rule. Each Monte Carlo simulation is based on 1,000 draws.

The overall simulation has sets of 10,000 trials for each of five different assumed standard deviations of the normally distributed disturbance in the regressions: 0.6, 0.8, 1.0, 1.2, 1.4. The middle value of these standard deviations corresponds to the mean value of true net benefits across the trials. To estimate the mean squared error of prediction, the average of the squared deviations of the difference between the true value of net benefits and the mean of the distribution of predicted net benefits is taken across trials. Also, the fraction of trials for which the true value of net benefits falls outside of the 90 percent confidence interval implied by the Monte Carlo distribution are calculated.

---

[2] One could also structure the estimation as a single regression with a set of indicators for the three impacts. If one does this with, say, 30 mutually exclusive indicators for the three impacts and 10 observations with all indicators set to zero, then not taking account of covariances among the estimators leads to a predicted variance of the sum of impacts approximately half of the actual value. Thus, treating statistically insignificant coefficients as zero would further underestimate the true variance if the covariances involving the insignificant coefficients were also set equal to zero.

7.0 Illustrative Results

Imagine that one held prior beliefs about the distributions of the true values of the impacts that correspond to those used in the first step of these simulations. That is, one thought the true values had the specified distributions. It then makes sense to ask: Which method of using coefficients would on average yield the greatest accuracy in the sense of minimizing the average of the squared deviations of the mean values of the Monte Carlo simulations, the value usually reported by analysts as the point estimate of net benefits, and the true value of net benefits? Table 1 shows these averages by the assumed values of the standard deviations of the errors in the regressions.

Table 1

Average Squared Deviations of True and Mean Predicted Net Benefits by Method

	Standard Deviation of Regression Error				
	0.6	0.8	1.0	1.2	1.4
Method 1: Significant Coefficients Only	.094	.090	.148	.218	.310
Method 2: All Coefficients	.077	.078	.121	.168	.239
Method 3: Shrunk Coefficients	.075	.074	.115	.157	.216

Note that either using all coefficients or using shrunk coefficients gives on average more accurate predictions than using only significant coefficients across the range of standard deviations. Further, using shrunk coefficients consistently gives the most accurate predictions, though using all coefficients does almost as well.

It is also interesting to ask which method does better over particular ranges of true net benefits. Table 2 displays the simulation results for average squared deviations between the true value of net benefits and the mean of the Monte Carlo distribution conditional on the range of true values of net benefits. First, note that using only statistically significant coefficients (method 1) gives a better prediction than the other approaches only when the true value of net benefits is non-positive and the standard deviation of the regression error is relatively large. In these cases, it gives a much more accurate prediction on average than using all the coefficients (method 2) and it does moderately better than using shrunk coefficients (method 3). For true

values of net benefits between 0 and 1, the shrunk coefficients give more accurate predictions than using only statistically significant coefficients for all the standard deviations, and they give more accurate predictions than using all the coefficients except in the case of the smallest standard deviation, where the two methods do equally well. For larger values of true net benefits, those between 1 and 2, using shrunk coefficients and all the coefficients provide comparable accuracy for smaller standard deviations, while using all the coefficients does better

Table 2

MSE for Each Method by True Net Benefit Intervals

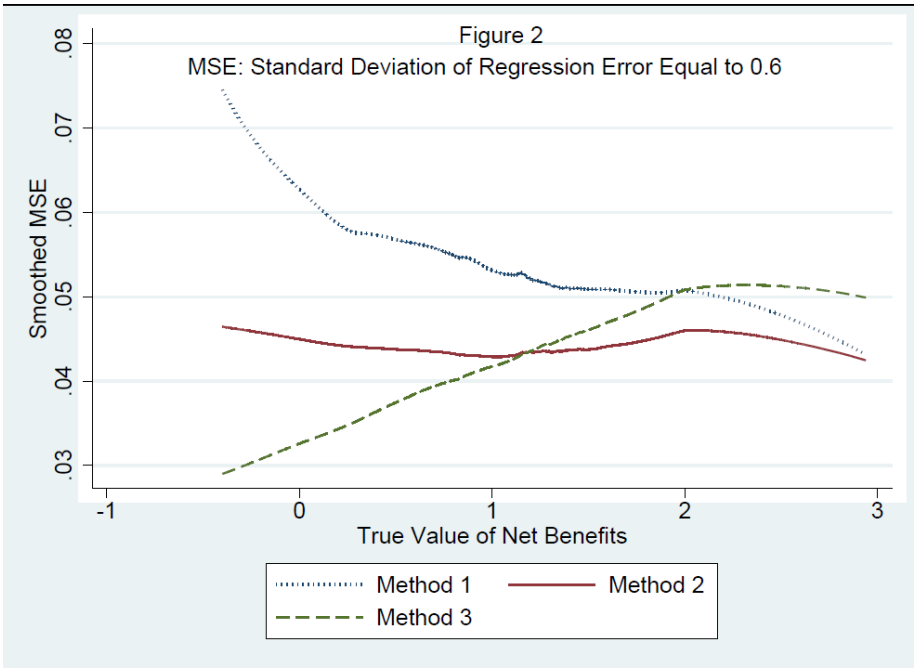
[Significant Coefficients Only
All Coefficients
Shrunk Coefficients
(Number of Trials)]

		Interval of True Net Benefits			
		-0.4 to 0	0 to 1	1 to 2	2 to 3
Standard Deviation of Regression Error	0.6	.07 .04 **.03** (496)	.06 **.04** **.04** (4,624)	.05 **.04** **.04** (4,309)	.05 **.04** .05 (571)
	0.8	.07 .08 **.05** (524)	.09 .08 **.06** (4,604)	.10 **.08** **.08** (4,291)	.11 **.09** .11 (581)
	1.0	**.07** .12 **.07** (497)	.13 .12 **.10** (4,622)	.17 **.12** .13 (4,309)	.18 **.13** .16 (572)
	1.2	**.09** .18 .11 (508)	.17 .17 **.13** (4,646)	.27 **.16** .18 (4,273)	.29 **.16** .22 (573)
	1.4	**.09** .23 .13 (509)	.22 .24 **.18** (4,654)	.41 **.24** .25 (4,279)	.53 **.23** .34 (558)

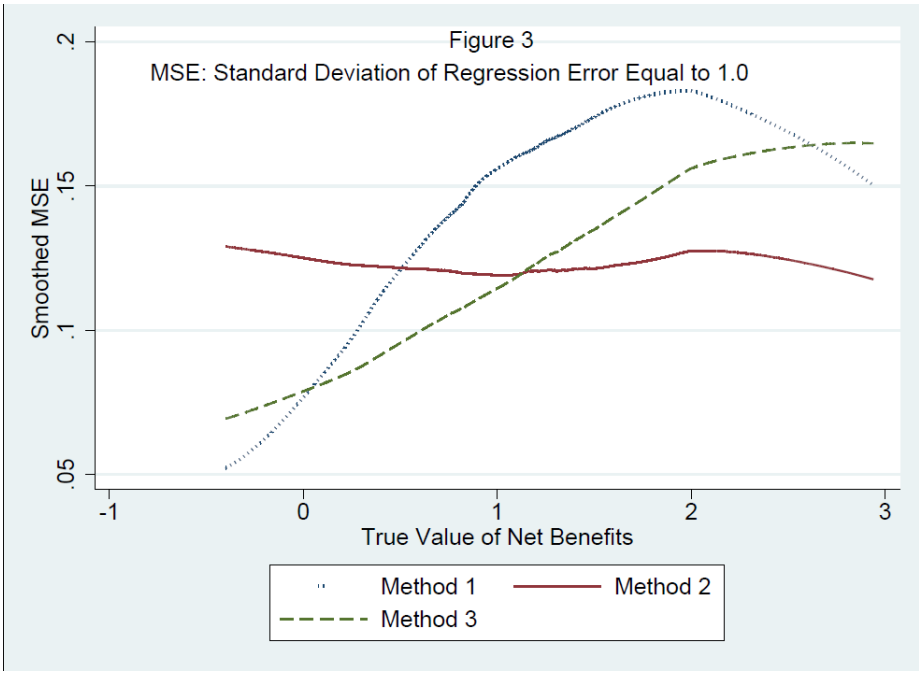
Minimum values in red**.**

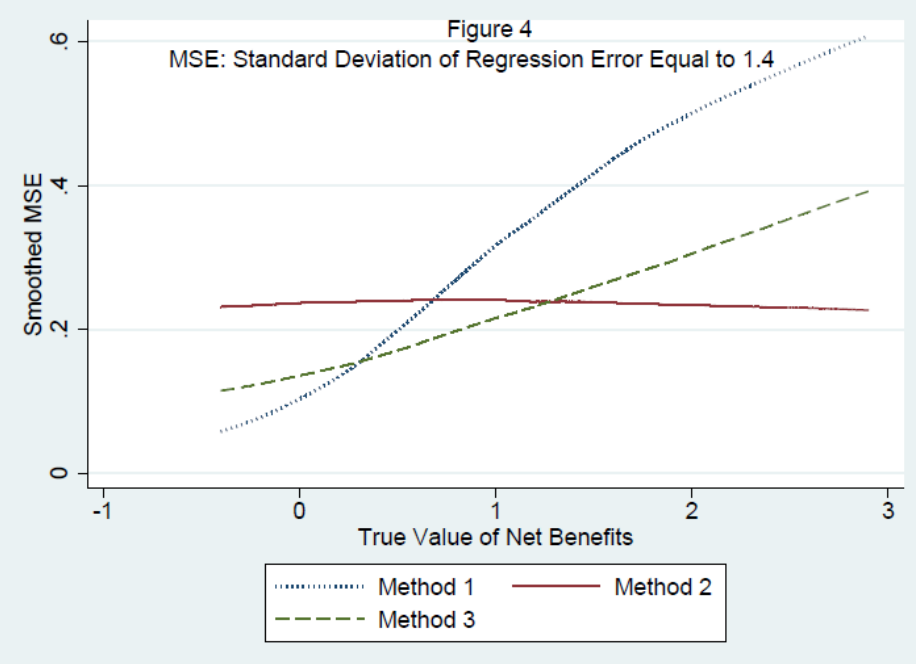
for the larger variances. When the true value of net benefits takes its largest values, between 2 and 3, using all the coefficients gives the most accurate predictions. Its advantage over the other two methods is small for the smallest standard deviations but grows large for the larger ones. For the larger standard deviations, shrunk coefficients also yield more accurate predictions than using only statistically significant coefficients.[3]

Figures 2, 3, and 4 show results for standard deviations of 0.6, 1.0, and 1.4, respectively. Each figure plots smoothed values of mean squared error for each of the three methods: dots for treating insignificant coefficients as zero; solid lines for using all coefficients; and dashes for shrunk estimates. Figure 2, the simulation with the least noise, shows shrunk estimates giving the smallest mean squared errors until true net benefits reach about 1, and using all coefficients doing so for larger values. Figure 3, the simulation with moderate noise, shows treating insignificant coefficients as zero does best until true net benefits reach about zero. Shrunk estimates do best until true net benefits go slightly beyond 1, and then using all coefficients does best. Finally, Figure 4, the simulation with the greatest noise, shows a similar pattern, but one showing treating statistically insignificant coefficients as zero more favorably for small values of true net benefits. However, it does much worse than either of the other methods beyond 1, the mean value of true net benefits.



[3] Just shrinking statistically insignificant coefficients does not outperform shrinking all in terms of mean squared error.





Which of these methods would yield distributions of predicted net benefits that best conveyed the true level of uncertainty in the prediction of net benefits? In other words, for which method would the confidence interval reported from the Monte Carlo simulation be most valid?

Table 3

Fraction of Times True Value Falls Outside of Monte Carlo 90 Percent Confidence Interval

[Significant Coefficients Only
All Coefficients
Shrunk Coefficients
(Number of Trials)]

		Interval of True Net Benefits			
		-0.4 to 0	0 to 1	1 to 2	2 to 3
Standard Deviation of Regression Error	0.6	1.0 .10 .08 (496)	.42 .10 .11 (4,624)	.21 .10 .11 (4,309)	.14 .12 .11 (571)
	0.8	1.0 .08 .09 (524)	.44 .10 .11 (4,604)	.26 .10 .12 (4,291)	.17 .11 .14 (581)
	1.0	1.0 .10 .08 (497)	.47 .10 .12 (4,622)	.30 .10 .12 (4,309)	.22 .12 .13 (572)
	1.2	1.0 .12 .11 (508)	.51 .09 .11 (4,646)	.35 .09 .13 (4,273)	.24 .11 .13 (573)
	1.4	1.0 .09 .08 (509)	.56 .10 .11 (4,654)	.39 .11 .14 (4,279)	.31 .10 .13 (558)

Table 3 shows how often the true value falls outside the derived 90 percent confidence interval for predicted net benefits. First, note that in most of the cells, the 90 percent confidence intervals do cover the true value about 90 percent of the time when using all the coefficients or shrunk coefficients. However, the confidence intervals that result from using only statistically significant coefficients and their standard errors are much too short. Indeed, for the smallest values of true net benefits, they never were wide enough to cover the true value. More generally, looking across cells in Table 3, the most valid confidence intervals using only significant coefficients have error rates of 40 percent (14 percent versus the expected 10 percent), which is

as large as the least valid confidence intervals using shrunk coefficients. The confidence intervals based on only significant coefficients do so poorly because they ignore the standard errors of the excluded coefficients in generating the distribution of predicted net benefits. This could be corrected by including the standard errors of statistically insignificant coefficients, but that would beg the question of why they are set to zero rather than to their estimated values.

The shrunk coefficients result in confidence intervals that are two wide when the true value of net benefits in this simulation are negative but too narrow for values greater than zero. One possible reason for this is that the variance estimate for the shrunk coefficients is based only on the first term of the Delta method expansion. Another reason is that in the derivation of the weights, the regression estimates of the impacts are substituted for their true values.

8.0 Shrinkage in Perspective

The purpose of this analysis is to assess the use of shrunk estimates to improve the predictive value of ordinary least squares estimators taken from single studies when the true parameters are near zero. Therefore, it uses only information that would be routinely available to analysts in reported results——coefficients and their standard errors. The approach considered minimized mean squared error and therefore falls squarely within classical statistics. Nonetheless, it can be interpreted in Bayesian terms and it is embedded in hierarchical linear modeling.

If analysts have available the original data upon which a regression of interest has been estimated, then they could consider adopting an informative Bayesian prior distribution centered at zero. For commonly used convenient conjugate priors, the resulting parameter estimate would be a weighted average of the prior value and the ordinary least squares estimate, with the weights depending on the prior and ordinary least squares variances. This would shrink the ordinary least squares estimate toward zero, but less so the more precise the ordinary least squares estimates. Thus, it would operate in a similar way to the shrinkage factor we employed.

Shrinkage also occurs in hierarchical linear models that include random effects for groups. Taking account of differences in sample sizes, and hence standard errors, the random effect models employing empirical Bayesian estimation shrink group means toward the overall average across groups to improve predictive accuracy. (As the shrinkage is larger for larger numbers of groups, it also obviates the need for multiple comparison adjustments if one is

interested in the hypothesis that all the nulls are true.) Thus, if one uses hierarchical linear models, one has already embraced a form of shrinkage.[4]

9.0 Discussion and Conclusion

The simulations suggest two observations in using empirical evidence in CBA or other predictive analysis. First, treating statistically insignificant coefficients as if they and their standard errors are exactly zero will only provide the most accurate predictions if the true values are actually very close to zero. On the surface, this seems like a conservative approach. However, CBA often requires estimating parameters related to costs, or negative benefits—— consider the negative region for the third impact in the simulation. Treating these parameters as zero when their estimates are statistically insignificant would bias CBA toward project acceptance. Even if one accepts these risks, also ignoring the standard errors of the insignificant coefficients results in Monte Carlo distributions of predicted net benefits that are much too tight.

Obviously, the problems that arise in treating statistically insignificant coefficients as if they are zero are exacerbated by multiple comparison corrections to ensure that the overall frequency of false positives within an analysis satisfy some critical level overall. The correction for multiple comparisons initially arose out of situations where repeated samples are taken from, say, a production line to try to determine if a product is defective (Perneger 1998). In such cases, the overall null hypothesis that none of the samples exceed specifications is appropriate and adjustment for multiple comparisons is indeed appropriate. However, in situations where we are interested in impacts on subgroups within an experimental sample, it is unlikely the case that we care about the overall null hypothesis. Indeed, it is unlikely that we care at all about any hypotheses relevant to the particular sample; rather, we want to use the estimates to test a policy-relevant hypothesis such as net benefits are positive. If so, then in the blind pursuit of a "correct" Type I error not only ignores the reality of Type II error, statistical insignificance when the null is false, because of an underpowered test, but also the more fundamental error of asking the wrong question. Consequently, when we conduct a field experiment to assess some policy intervention, we should report all the plausibly relevant subgroup effects, with their standard

[4] On the advantages and disadvantages of the use of shrunk estimators in report cards on health care providers see Mukamel et al. (2010).

errors, and not hide or throw away policy relevant information in the inappropriate devotion to a correct Type I error for an irrelevant hypothesis.

A possible caveat to this harsh assessment of treating statistically insignificant coefficients as zero is the problem of publication bias. In industries like pharmacology there is a strong financial incentive not to publish the results of studies with insignificant findings of positive effects. In the social sciences, our reluctance as referees to recommend publication of even well-designed studies that fail to find statistically significant results creates an incentive for researchers to abuse their data in ways that make the reported coefficients, standard errors, and significance levels suspect (Weimer, 1986; Humphreys et al., 2013). In fields where this sort of selection bias is operating, the sorts of simulations used in this analysis would not necessarily be valid because some of the draws yielding statistically insignificant coefficients would be disregarded. One shocking non-intuitive (and, yes, perhaps bogus) implication is that if one can only find one relevant empirical study when the nature of the question and the likely availability of data suggest multiple studies should be available, one might put more confidence if the one found is a working paper posted on the internet than if it is published in a peer reviewed journal! Of course the former does not have the benefit of peer review of the methods and data, so one would bear greater burden of assessment. In effect, one would be taking on this burden in an effort to avoid the selection bias.

Second, shrinking estimates and their standard errors provides better predictions than either treating statistically significant coefficients as zero or using the raw estimates for relatively small values of true parameters. However, if the true parameter values are very large and noisy, then the raw estimates will provide substantially more accurate predictions than the shrunk estimates.

These observations together I think lead to the following heuristic: Rather than assuming statistically insignificant coefficients and their standard errors are zero, use shrunk coefficients and shrunk standard errors!

References

Anderson, Michael L. (2008) Multiple Inference and Gender Differences in the Effects of EarlyIntervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association* 103(484), 1481–1495.

Benjamini, Yoav and Yosef Hochberg (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series A* 57(1), 289–300.

Copas, J.B. (1997) Using Multiple Regression Models for Prediction: Shrinkage and Regression to the Mean. *Statistical Methods in Medical Research* 6(2), 167–183.

Copas, J. B . (1983) Regression, Prediction and Shrinkage. *Journal of the Royal Statistical Society (Series B)* 45(3), 311–254.

Efron, Bradley and Carl Morris (1977) Stein=s Paradox in Statistics. *Scientific American* 236, 119–127.

Galton, Francis (1886) Regression Towards Mediocrity in Hereditary Stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt (2013) Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Regulation. Political Analysis 21(1), 1–20.

Institute of Education Sciences (2008) *What Works Clearinghouse Procedures and Standards Handbook Version 2.0*. ed.gov/ncee/wwc/.

Mukamel, Dana B., Laurent G. Glance, Andrew W. Dick, and Turner M. Osler (2010) Measuring Quality for Public Reporting of Health Provider Quality: Making It Meaningful to Patients. *American Journal of Public Health* 100(2), 264–269.

Nicol, Kristen L. (2001) Cost-Benefit Analysis of a Strategy to Vaccinate Healthy Working Adults against Influenza. *Archives of Internal Medicine* 163(5), 749–759.

Perneger, Thomas V. (1998) What=s Wrong with Bonferroni Adjustments. *British Medical Journal* 316(7139), 1236–1238.

Stein, Charles M. (1981) Estimation of the Mean of a Multivariate Normal Distribution. *Annals of Statistics* 9(6), 1135–1151.

Thompson, James (1968) Shrinkage Techniques for Estimating the Mean. *Journal of the American Statistical Association* 63(321), 113–122.

Weimer, David L. (1986) Collective Delusion in the Social Sciences. *Review of Policy Research* 5(4), 705–708.

Weimer, David L. and Mark A. Sager (2009) Early Identification and Treatment of Alzheimer's Disease: Social and Fiscal Outcomes. *Alzheimer's & Dementia* 5(3), 215–226.

Weimer, David L. and Aidan R. Vining (2009) *Investing in the Disadvantaged: Assessing the Benefits and Costs of Social Policies* (Washington, DC: Georgetown University Press).

Ziliak, Stephen T. and Deirdre N. McCloskey (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives* (Ann Arbor: University of Michigan Press).