Do econometric models provide more accurate forecasts when they are more conservative?

A test of political economy models for forecasting elections

Andreas Graefe[a], Kesten C. Green[b], J. Scott Armstrong[c]

[a] *Department of Communication Studies and Media Research,*
LMU Munich, Germany. a.graefe@lmu.de
[b] *University of South Australia Business School, and the Ehrenberg-Bass Institute,*
University of South Australia, Australia. kesten.green@unisa.edu.au
[c] *Wharton School, University of Pennsylvania, USA, and the Ehrenberg-Bass Institute,*
University of South Australia. armstrong@wharton.upenn.edu

August 11, 2015; GREF-68.docx

Revised title on March 11

Abstract. The assumptions of multiple regression analysis are not met in many practical forecasting situations and, as a result, regression models are insufficiently conservative. We tested the effect on forecast accuracy of applying three evidence-based forecasting guidelines to 18 political economy models for forecasting elections in nine countries, all of which were originally estimated using multiple regression analysis. The guidelines direct modelers to account for uncertainty of econometric model forecasts by (1) modifying estimates of the strength of variable effects, (2) combining forecasts from diverse models, and (3) taking account of all variables that are known to be important. Out-of-sample forecast accuracy was compared with the accuracy of forecasts from the originally published econometric models representing typical practice. While damping the estimated variable weights did not improve accuracy, equalizing them reduced error compared the original model forecasts by 10%. Combining forecasts from models for US (N=8) and Australian (N=2) elections reduced error by 25% on average. Including more causal knowledge, by using all unique variables from the different models in equal-weights index models, reduced error on average 26%.

Keywords: combining forecasts**,** damping, equalizing, elections, golden rule of forecasting, index method, shrinkage.

Acknowledgments. We received helpful suggestions when presenting an early version of this paper at the 2014 APSA Annual Meeting in Washington, DC. Lynn Selhat edited the paper.

Introduction

The development of causal models for forecasting elections has become an important sub-discipline of political science. These models rely on theories of voting behavior and use a set of structural variables to predict election outcomes. The dominant theory underlying most models is the idea of retrospective voting, which views an election as a referendum on the incumbent government's performance or, more narrowly defined, its ability to handle the economy. Retrospective voting theory thus assumes that voters reward the government for good performance and punish the incumbent party otherwise. Causal models typically build on this idea by incorporating one or more economic variables such as GDP growth, unemployment, or inflation to measure the government's economic performance. In addition, such models often also include poll-based measures such as such as popularity, which is commonly seen as a proxy variable for measuring the incumbent's overall performance and is thus assumed to also include information about the electorate's satisfaction with the government handling non-economic issues. Finally, many models include information about a country's specific electoral system, such as the time the incumbent party has held power. For example, first-time elected leaders often enjoy a honeymoon phase in their first term but the chance of reelection decreases the longer they hold office due to the electorate's increasing desire for change. Due to their reliance on economic and political variables, such models are commonly known as political economy models.

In the US, political economy models have been an established by-product of presidential elections since the late 1970s (Fair 1978). For the past six elections since 1992, political scientists and economists have published their models and forecasts prior to the election in special symposia of scientific journals such as *Political Methodologist* 5(2), *American Politics Research* 24(4) and *PS: Political Science and Politics* 34(1), 37(4), 41(4), and 45(4). This work also spearheaded the development of election forecasting models in other countries, as shown with two special issues of the *International Journal of Forecasting* 26(1) and 28(4). In particular, researchers have developed models for major European countries (e.g., France, Germany, and the UK) as well as for neglected democracies such as Japan, Portugal, Spain, and Turkey. These models are used to test theories of voting and to estimate the relative effects of specific variables on the aggregate vote. Most importantly, however, the models provide ex ante forecasts of future election outcomes, usually months in advance. The goal of the present paper is to test whether evidence-based principles derived from the forecasting literature can help to improve the models' predictive accuracy.

The dominant method for estimating political economy models is multiple regression analysis. Multiple regression analysis estimates variable weights that provide the best possible solution for a given sample. The resulting "optimal" (in terms of least squares) variable weights are then used to predict new (out-of-sample) data. When estimating variable weights, multiple regression analysis accounts for error in the measurement of variables and for uncertainty about the strength of the relationships. However, regression analysis does not account for other sources of uncertainty such as bias in the data, use of proxy variables to represent the true causal variables, omission of important variables, inclusion of irrelevant variables, lack of variation in variable values, or error in predicting the causal variables. As a result, regression models are insufficiently conservative, which harms their accuracy when predicting new data.

In an attempt to make forecasts more conservative and thus to increase their accuracy, Armstrong, Green and Graefe (2015), hereafter referred to as AGG, proposed the Golden Rule of Forecasting, which is "to be conservative" (p. 1718). Conservative forecasting requires forecasters to adhere closely to cumulative prior knowledge about the situation and about relevant evidence-based forecasting methods. From their Rule, AGG developed 28 conservative forecasting guidelines.

These guidelines were derived by logical deduction to describe how conservatism should be applied to the different aspects of a forecasting problem, and to different kinds of forecasting problems. Specifically, the guidelines address how to formulate a forecasting problem, how to forecast with judgmental, extrapolative, and causal methods, how to combine forecasts from different methods, and how to adjust forecasts. AGG then assessed the effects of each of the guidelines on out-of-sample forecast accuracy by reviewing published studies that compared the accuracy of forecasts from conservative and non-conservative forecasting methods. Of the 105 studies they identified, 102 supported the guidelines. On average, ignoring a guideline increased forecast error by more than 40%.

The present paper tests the value of applying the conservative guidelines to 18 political economy models for forecasting elections. We first describe the models that are the subject of our empirical validation tests. Second, we describe three conservative guidelines for dealing with uncertainty in the estimation of causal models. Third, we describe the methods and the data that we used. Fourth, we present our results. Finally we discuss the results and draw conclusions regarding the future of econometric forecasting.

Political economy models for forecasting elections

We used five criteria for including a model in our analysis. In particular, a model needed to be (1) published in an academic journal, (2) predict national elections, and (3) estimated with multiple linear ordinary least squares (OLS) regression analysis. Furthermore, (4) we deliberately excluded models with a fit (R^2) higher than 0.95 (e.g., X. The reason is that these models were developed shortly before an election and are highly fitted to given data, so our method of cross-validation is unlikely to reveal the underlying uncertainty. Finally, (5) the model data needed to be available. The latter criteria led to the exclusion of a number of models, since some forecasters did not publish their data and did not respond to our request to use their data (e.g., X, Y, Z).

Table 1: Key features of the 18 models analyzed in the present study

				No. of variables		
Country / Election / Model	**Dependent variable**	**Model fit (R^2)**	**No. of elections**	**Total**	**Economic**	**Political**
Australia (general)						
Cameron & Crosby (2000)*	Incumbent vote	0,24	40	5	4	1
Jackman (1995)*	Incumbent vote	0,19	22	3	2	1
Canada (general)						
Bélanger & Godbout (2010)	Incumbent vote	0,78	19	4	1	3
Nadeau & Blais (1993)	Liberal vote	0,54	13	4	3	1
Italy (national, european, and local)						
Bellucci (2010)	Incumbent vote	0,73	9	3	0	3
Japan (general)						
Lewis-Beck & Tien (2012a)	LDP (percent seats)	0,77	17	3	1	2
Portugal (general)						
Magalhães & Aguiar-Conraria (2009)	Incumbent vote	0,94	11	3	1	2
Spain (general)						
Magalhães, Aguiar-Conraria & Lewis-Beck (2012)	Liberal vote	0,62	14	4	2	2
Turkey (general)						
Toros (2011)	Incumbent vote change	0,91	11	3	1	2
UK (general)						
Lewis-Beck, Nadeau & Bélanger (2004)	Incumbent vote	0,88	12	3	1	2
US (presidential)						
Fair (2009)	Incumbent vote	0,90	25	7	4	3
Cuzan (2012)	Incumbent vote	0,92	25	5	3	2

Abramowitz (2012)	Incumbent vote	0,91	17	3	1	2
Campbell (2012)	Incumbent vote	0,84	17	2	1	1
Lewis-Beck & Tien (2012b)	Incumbent vote	0,90	16	4	2	2
Holbrook (2012)	Incumbent vote	0,83	16	3	0	3
Erikson & Wleizen (2012)	Incumbent vote	0,76	16	2	1	1
Lockerbie (2012)	Incumbent vote	0,77	15	2	1	1
	Median across all 18 models	**0,80**	**16**	**3**	**1**	**2**

* Data obtained from Leigh and Wolfers (2006), who updated both models up to the 2004 election.

Ultimately, we ended up with 18 models from nine countries. While these models are of course not exhaustive of the election forecasting literature, we believe that they provide a representative sample of the work done in different countries. Table 1 provides an overview of the 18 models' key features such as the model fit, the number of elections (observations in the sample), and the number of variables, split up by economic and political variables. Given the attention that election forecasting attracts in the US, it is not surprising that models for forecasting US presidential election form the largest group, with a total of eight models. Two models were available for Australian and Canadian general elections, whereas only one model each was available for Italy, Japan, Portugal, Spain, Turkey and the UK. Across all 18 models, the median model fit was quite high, with an R^2 of 0.80. The median ratio of observations to variables was 16 to 3, which means that about five observations per independent variable were available to estimate a model. The general specification for each of these political economy models can be written as:

$$V = a + \sum_{i=1}^{k} b_i x_i$$

where *V* is the incumbent party's share of the national two-party popular vote, *a* is the intercept and the b_i's are the coefficients—all estimated from historical data—of the *k* predictor variables, x_i to x_k.

Conservative guidelines for causal models

AGG provide four conservative guidelines for causal models. We test three: (1) modify effect estimates to reflect uncertainty, (2) combine forecasts from dissimilar models, and (3) include all variables that are important in the model. We hypothesized that following these guidelines would result in forecasts that were more accurate than those from models estimated using regression analysis.

The three guidelines address well-known weaknesses in forecasting models developed using regression analysis. Multiple regression analysis estimates the sizes of causal variables' effects from a given data set, a procedure that will fail to produce good forecasting models in many situations. As summarized by Graefe (2013), the accuracy of forecasts of election results from regression models is adversely affected by the use of non-experimental data, small sample sizes, the exclusion of many important variables, measurement errors, variables that correlate with one another, omitted variables, and important variables do not vary much in the estimation sample. Under such conditions, which are common for social science problems, regression is best confined to estimating models for situations in which there are only a few causal variables that are important (Armstrong, 2012).

We now describe the three guidelines that can help address weakness in regression analysis as a method for developing forecasting models, and describe their application to political economy models for forecasting election results.

Modify effect estimates to reflect uncertainty

Regression reduces the estimated effect of a variable in response to unexplained variation in the historical data. However, it does not compensate for all sources of uncertainty. In order to compensate for the unaccounted for uncertainty, damping causal variable coefficient estimates is sometimes proposed as an appropriate conservative strategy.

Damping coefficients

Damping reduces the size of a variable's coefficient toward having no effect. For causal models, it assumes that the actual causal effects are weaker than the effects that are estimated from the data by regression analysis, thus forecasts will stay closer to the regression model constant. Damping has been shown to be an effective guideline for extrapolation models, where, on average over xx studies, it reduced error by 12% (AGG). However, unlike regression analysis, extrapolation does not by its nature adjust for uncertainty. While users of regression analysis have often been advised to use "shrinkage" (refs), which is analogous to damping, we are unaware of any research that tests the conditions under which shrinkage improves upon the ex ante forecasting accuracy of the original regressions.

Damping is a conditional guideline. In articular, it is not expected to work if the estimated coefficient is lower than that provided by prior knowledge, for example where many previous studies have estimated a higher price elasticity of demand for a good one might expect better forecasts by increasing the size of the coefficient toward the prior estimates (ref??). If, on the other hand, the forecaster is uncertain over whether causal variables will take on values or combinations that are more extreme than those in the estimation data, the case for damping would seem stronger, the greater the uncertainty the greater should be the damping. The question of whether and when damping is useful for causal model forecasting remains.

A simple strategy for damping is to multiply the estimated weights with a factor *d*. The "damped" version of the original regression model can be written as:

$$V = a + (1 - d)\sum_{i=1}^{k} b_i x_i$$

The factor *d* can range from 0 to 1. For *d*=0, the original regression model would remain unchanged, which means no damping. For *d*=1, on the other hand, the model coefficients are in effect zero and the model forecast is simply the value of the intercept *a*, which is the incumbent's vote share that would be obtained if the predictor variables were equal to their historical mean. The bigger the factor *d*, the greater is the shrinking toward the historical average incumbent vote share.

Equalizing coefficients

Equalizing provides a conservative response to uncertainty caused by the possibility of multicollinearity, the problem of estimating weights when causal variables are correlated with other causal variables, whether included in the model or not. In other words, equalizing is useful if there is uncertainty about the relative importance of the predictor variables; the greater the uncertainty, the more one should adjust the coefficients towards equality. When relative effect sizes are highly uncertain, one should consider the most extreme case of equalizing and assign equal weights to all variables expressed as differences from their mean divided by their standard deviation (i.e., standardized).

To equalize, standardized the variables, estimate the model using regression analysis, and adjust the estimated coefficients toward equality. The adjusted vote equation can be written as:

$$V = a + (1 - e)\sum_{i=1}^{k} b_i x_i + \frac{e}{k}\sum_{i=1}^{k} b_i \sum_{i=1}^{k} x_i$$

where *e* is the equalizing factor, which can range from 0 to 1. The greater the equalizing factor *e*, the greater the amount of equalizing. An equalizing factor *e*=0 yields the original regression model, and thus amounts to zero equalizing. For *e*=1, all model coefficients are assigned equal weights, which is the most extreme case of equalizing.

Graefe (2015) reviewed comparative studies on equal weights published since the 1970s in a variety of areas, and concluded that equal weights models often provide ex ante forecasts that are more accurate than those from regression models. For example, Dana and Dawes (2004) analyzed the relative predictive accuracy of forecasts from regression and equal weights models by making out-of-sample predictions using five real non-experimental social science datasets and a large number of synthetic datasets. Regression weights were inferior to equal weights where there were fewer than 100 observations per predictor variable available for estimating the model. This is the case for many practical problems, and for election forecasting.

In the domain of election forecasting, Cuzán and Bundrick (2009) found that equal-weights versions of the Fair (2009) model and the fiscal model (Cuzán, 2012) provided out-of-sample election forecasts that were at least as accurate as those from the original regression models. In addition, Graefe (2015) showed that equal-weights versions of six of nine established regression models for election forecasting yielded more accurate forecasts than the original models. On average across the 10 elections from 1976 to 2012, the equal-weights models reduced the original regression models' ex ante forecast errors by 5%.

Combine forecasts from dissimilar models

The second of the Golden Rule guidelines for causal methods that we consider in this paper recommends combining forecasts from models that incorporate different data and information. Hundreds of studies have shown that combining forecasts is an effective method for using additional knowledge and to thereby improve forecast accuracy (see Graefe, et al. 2014 for a review).

Combining has also been applied to election forecasting. *The Economist* published the first "poll of polls" in 1992. Nowadays, combining polls from different pollsters (e.g., RealClearPolitics's Poll Average, and Huffington Post Pollster) is common practice (Blumenthal, 2014). As demonstrated in the PollyVote project, combining is effective in reducing error when forecasts from different methods are combined as an equally weighted average. Since 2004, PollyVote.com has provided consistently accurate forecasts of the popular vote in U.S. presidential elections by averaging forecasts within and across four methods: polls, prediction markets, expert judgment, and causal models. The PollyVote forecast has been more accurate than the forecasts from each component method and reduced error relative to the error of forecasts from the best single method in that combination—prediction markets—by 16% on average (Graefe, Armstrong, Jones Jr. and Cuzán, 2014).

An early review of more than 200 papers on combining forecasts failed to find evidence that complex combining schemes can consistently provide forecasts that are more accurate than simple averages (Clemen, 1989). Graefe, Küchenhoff, Stierle, and Riedl (2015) reviewed studies since then and found that the results still hold, though there are exceptions such as when one puts heavier weight on methods that have been found to be most suitable for the situation. In addition, that paper provides further evidence for U.S. presidential election forecasting: across the elections from 1976 to 2012, the error of simple unweighted averages of forecasts from six election-forecasting models was 25% lower than the corresponding error of the forecasts from the much more complex "Ensemble Bayesian Model Averaging." In light of the evidence, we calculated simple unweighted averages of the forecasts from all eight models to generate a combined model forecast for this study**.**

Try to use all important variables

The third guideline recommends trying to include all variables that are known to be important in a model. This guideline is difficult to implement with regression analyses. As is well known, the practical limit of regression is a handful of variables (Armstrong, 2012). Researchers typically deal with this problem by using only some of the variables that are known to be important.

One way to avoid the practical limits that regression places on the number of variables in a model is to use prior knowledge—*not* statistical methods—to select causal variables and to determine the direction and size of their effects. This calls for a review of the cumulative knowledge from prior research. This approach was well-received in the mid-1900s Refer to conditional regression analysis (Wold an Jureen etc . . . .(use material from Long-range Forecasting 1985, pages 214-217 and 232-235)

Another approach can be traced back to at least as early as Benjamin Franklin, who suggested a method for solving a binary choice problem, which he called "Moral Algebra, or Method of Deciding Doubtful Matters" (Sparks, 1844, p. 20). In particular, Franklin recommended to first, identify all important variables, second, code each variable according to its directional impact on the outcome (e.g., positive: 1; negative: -1), third, weight each variable by importance, and, fourth, add up the variable scores to see which outcome is favored.

While Franklin originally suggested differential weighting, forecasters often lack adequate prior knowledge about the variables' relative importance. In such situations, equalizing the variables' weights as described, as described above, provides a good starting point. First, as the number of variables in a model increase, the magnitudes of effects become less important. Wilks (1938) showed mathematically that for models with a large number of intercorrelated variables, the variable weighting has virtually no effect on the prediction. {mention Schmidt as first empirical test?} Second, equalizing prevents forecasters from weighting variables on the basis of unsubstantiated preconceptions or in ways that suit their biases. [First application to economic problems is??

Franklin's approach is geared towards solving binary choice problems. However, the sum of the variable scores can also be used to make numerical forecasts, by using the sum as the independent variable in a regression analysis. The major advantage of this approach, which is known as the index method, is that variable are included on the basis of prior knowledge about their importance and effect size and direction, not from the given data. Thus an unlimited number of variables can be included in a model.

Commented [KG1]: WE WORKED OUT A TERMINOLOGY IN AN EARLIER PAPER… WE SHOULD BE CONSISTENT WITH IT, BUT CAN'T REMEMBER OF THE TOP OF MY HEAD

Graefe (2015) tested the benefits of the index method by assigning equal weights to all 27 (standardized) variables that were included in nine established models for forecasting U.S. presidential elections. The resulting model was used to generate *ex ante* forecasts of the ten elections from 1976 to 2012 with an average error of 1.3 percentage points. That error was 48% smaller than the typical model's error and 29% smaller than the most accurate model's error.

The present study uses a similar approach and sums up the standardized values of all variables that are used in different models that predict the same target variable in order to calculate an index variable. The resulting vote equation is:

$$V = a + b\sum_{i=1}^{25} x_i$$

where the x_i's are the standardized values of all unique variables used in different models.

Method and data

For each of the 18 models, we standardized the original data and transformed some of the variables to ensure that all predictor variables correlated positively with the dependent variable. We analyzed the accuracy of forecasts across all observations available for each model. All forecasts were out-of-sample using an N-1 cross-validation procedure, an approach that is also known as jackknifing. The method allows for a more powerful test of predictive validity because it expands the number of ex ante forecasts. Specifically, to forecast an election outcome, we estimated models using the data on the other elections.

We report the relative absolute error (RAE) of the forecasts that result from the application of each guideline (see Armstrong and Collopy, 1992). The RAE is calculated as the mean absolute error (MAE) of forecasts from a model that follows the guideline, divided by the corresponding MAE of the original model . Values of RAE greater than 1 mean that following a guideline yielded forecasts that were *less* accurate than those from the original model, whereas values less than 1 mean that following the guidelines yielded more accurate forecasts. All data and calculations will be made available at the Harvard Dataserve upon publication.

Accuracy gains from applying the guidelines

Modifying effect estimates

Damping the estimated coefficients

Table 2 shows the relative absolute errors for each model with various levels of damping. Across all 18 models, damping of 10% and 20% slightly reduced forecast errors, with a median error reduction of 3% (=1 - 0.97). For example, damping of 10% provided error reductions for 12 and had no effect for 3 of the 18 models; for the remaining 3 models, 10% damping harmed accuracy. In general, there was little gain from damping, and heavier damping harmed accuracy.

Table 2: Relative absolute error (RAE) of forecasts from damping compared to forecasts from the original regression models

	MAE	**RAE depending on damping factor (in %)**									
Model	**0%**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
Abramowitz (2012)	1,69	1,00	*1,05*	*1,16*	*1,34*	*1,53*	*1,74*	*1,94*	*2,15*	*2,36*	*2,57*
Campbell (2012)	1,97	*1,03*	*1,06*	*1,09*	*1,12*	*1,27*	*1,45*	*1,64*	*1,82*	*2,01*	*2,20*
Cuzán (2012)	2,05	1,00	*1,06*	*1,17*	*1,34*	*1,54*	*1,73*	*1,93*	*2,12*	*2,32*	*2,52*
Bellucci (2010)	2,21	0,97	0,95	0,93	0,92	0,90	0,88	0,86	0,86	0,87	0,89
Lewis-Beck & Tien (2012b)	2,25	0,98	0,98	*1,05*	*1,14*	*1,26*	*1,42*	*1,57*	*1,73*	*1,89*	*2,05*
Lewis-Beck, Nadeau & Bélanger (2004)	2,33	*1,03*	*1,08*	*1,12*	*1,18*	*1,24*	*1,30*	*1,36*	*1,46*	*1,58*	*1,71*
Erikson & Wlezien (2012)	2,47	0,97	0,99	*1,03*	*1,10*	*1,23*	*1,36*	*1,49*	*1,62*	*1,75*	*1,88*
Jackman (1995)	2,54	0,98	0,97	0,96	0,94	0,93	0,92	0,91	0,91	0,91	0,91
Holbrook (2012)	2,59	0,96	0,91	0,95	*1,07*	*1,19*	*1,31*	*1,43*	*1,55*	*1,67*	*1,79*
Lockerbie (2012)	2,73	1,00	*1,01*	*1,05*	*1,08*	*1,15*	*1,25*	*1,36*	*1,46*	*1,57*	*1,67*
Fair (2009)	2,80	0,97	0,96	0,99	*1,08*	*1,21*	*1,34*	*1,46*	*1,59*	*1,72*	*1,85*

Magalhães & Aguiar-Conraria (2009)	3,17	0,97	1,00	*1,11*	*1,24*	*1,44*	*1,64*	*1,84*	*2,04*	*2,25*	*2,47*
Cameron & Crosby (2000)	4,22	0,97	0,94	0,91	0,90	0,88	0,87	0,86	0,86	0,85	0,86
Bélanger & Godbout (2010)	4,59	0,96	0,93	0,90	0,89	0,89	0,92	0,95	*1,02*	*1,12*	*1,24*
Lewis-Beck & Tien (2012a)	4,79	0,96	0,96	0,98	1,00	*1,02*	*1,06*	*1,09*	*1,13*	*1,25*	*1,38*
Nadeau & Blais (1993)	7,01	0,96	0,92	0,88	0,85	0,82	0,78	0,75	0,73	0,70	0,67
Magalhães, Aguiar-Conraria & Lewis-Beck (2012)	9,86	0,95	0,91	0,86	0,82	0,79	0,76	0,73	0,70	0,67	0,67
Toros (2011)	12,44	*1,02*	*1,08*	*1,18*	*1,27*	*1,36*	*1,46*	*1,55*	*1,68*	*1,82*	*1,96*
Median	**2,66**	**0,97**	**0,97**	***1,01***	***1,08***	***1,20***	***1,30***	***1,39***	***1,51***	***1,62***	***1,75***

Notes:

- Models are ordered from most to least accurate by mean absolute error (MAE) of the original individual regression models' out-of-sample error (determined by N-1 cross validation) across all available observations.
- The MAEs are the percentage point error that one would achieve without any (i.e., 0%) damping.
- Underlined: Damped model was more accurate than original model.
- *Italics*: Damped model was less accurate than original model**.**

Equalizing the estimated coefficients

Table 3 shows the relative absolute errors for each model with various levels of equalizing. Equalizing reduced forecast error for 169 (or 94%) out of the 180 combinations (i.e., 18 models x 10 equalizing levels). Forecast accuracy tended to increase with higher levels of equalizing for all models except Holbrook (2012) and Magalhães, Aguiar-Conraria & Lewis-Beck (2012). At the most extreme case of equalizing, in which all predictor variables are assigned equal weights by setting *e*=1, the median RAE was 0.90. That is, equal weights reduced forecast error by 10%.

Table 3: Relative absolute error (RAE) of forecasts from equalizing compared to forecasts from the original regression models

	MAE	**RAE depending on equalizing factor (in %)**									
Model	**0%**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
Abramowitz (2012)	1,69	0,99	0,98	0,98	0,97	0,96	0,95	0,95	0,94	0,93	0,92
Campbell (2012)	1,97	0,99	0,99	0,98	0,98	0,97	0,96	0,96	0,95	0,95	0,94
Cuzán (2012)	2,05	0,98	0,96	0,95	0,93	0,91	0,90	0,89	0,89	0,89	0,90
Bellucci (2010)	2,21	0,97	0,94	0,91	0,88	0,85	0,81	0,78	0,75	0,72	0,69
Lewis-Beck & Tien (2012b)	2,25	1,00	0,99	0,99	0,98	0,98	0,98	0,97	0,98	0,99	1,00

Lewis-Beck, Nadeau & Bélanger (2004)	2,33	0,98	0,95	0,93	0,91	0,89	0,86	0,84	0,82	0,80	0,77
Erikson & Wlezien (2012)	2,47	0,88	0,99	0,98	0,98	0,97	0,96	0,96	0,95	0,94	0,94
Jackman (1995)	2,54	0,99	0,98	0,97	0,96	0,96	0,95	0,95	0,95	0,96	0,96
Holbrook (2012)	2,59	0,99	0,98	0,96	0,95	0,95	0,95	0,97	0,99	*1,01*	*1,03*
Lockerbie (2012)	2,73	0,99	0,99	0,98	0,97	0,97	0,96	0,96	0,95	0,94	0,94
Fair (2009)	2,80	0,95	0,90	0,86	0,83	0,80	0,77	0,77	0,78	0,79	0,80
Magalhães & Aguiar-Conraria (2009)	3,17	0,95	0,94	*1,02*	*1,10*	*1,18*	*1,27*	*1,40*	*1,54*	*1,68*	*1,82*
Cameron & Crosby (2000)	4,22	0,97	0,94	0,92	0,89	0,87	0,86	0,85	0,85	0,84	0,84
Bélanger & Godbout (2010)	4,59	0,98	0,96	0,95	0,93	0,91	0,89	0,88	0,86	0,84	0,82
Lewis-Beck & Tien (2012a)	4,79	0,99	0,97	0,96	0,94	0,93	0,92	0,90	0,89	0,88	0,87
Nadeau & Blais (1993)	7,01	0,97	0,94	0,91	0,88	0,86	0,83	0,81	0,79	0,77	0,76
Magalhães, Aguiar-Conraria & Lewis-Beck (2012)	9,86	0,96	0,93	0,89	0,86	0,82	0,79	0,75	0,72	0,69	0,66
Toros (2011)	12,44	0,97	0,94	0,91	0,87	0,86	0,87	0,88	0,88	0,89	0,90
Median	**2,66**	**0,98**	**0,96**	**0,95**	**0,94**	**0,92**	**0,91**	**0,89**	**0,89**	**0,89**	**0,90**

Notes:

- Models are ordered from most to least accurate by mean absolute error (MAE) of the original individual regression models' out-of-sample error (determined by N-1 cross validation) across all available observations.
- The MAEs are the percentage point error that one would achieve without any (i.e., 0%) equalizing.
- Underlined: Equalized model was more accurate than original model.
- *Italics*: Equalized model was less accurate than original model**.**

Combining forecasts

The benefits of combining forecasts can be tested for elections for which (a) more than one model is available and (b) the models predict the same dependent variable. This was the case for the eight models that forecast US presidential elections and the two models that forecast Australian general elections. Note that although two models were available for predicting Canadian federal elections, these models predict a different outcome (i.e., incumbent vs. liberal vote) and thus their forecasts could not be combined. Table 4 shows the results.

Table 4: Relative absolute errors from combining forecasts compared to the original regression models

	MAE (original)	**RAE (combining)**
Australia, 22 elections from 1951 to 2004, MAE of combined forecast: 2.26		

Cameron & Crosby (2000)	2,68	0,84
Jackman (1995)	2,54	0,89
Mean	***2,61***	***0,86***
US, 15 elections from 1956 to 2012, MAE of combined forecast: 1.48		
Abramowitz (2012)	1,76	0,84
Campbell (2012)	1,99	0,74
Cuzán (2012)	2,07	0,72
Erikson & Wlezien (2012)	2,54	0,58
Fair (2009)	2,49	0,60
Holbrook (2012)	2,55	0,58
Lewis-Beck & Tien (2012b)	2,29	0,65
Lockerbie (2012)	2,73	0,54
Mean	***2,30***	***0,64***

For Australian elections, model forecasts were combined across the 22 elections from 1951 to 2004 for which forecasts from both models were available. The MAE of the combined forecast was 2.30 percentage points, which was again more accurate than each individual model. Compared to the typical model forecast, which obtained an error of 2.61 percentage points, combining reduced error by 14%.

For US elections, model forecasts were combined across the 15 elections from 1956 to 2004 for which forecasts from all eight models were available. The MAE of the combined forecast was 1.48 percentage points and was thus smaller than the average errors of each of the eight individual models, which ranged from 1.76 to 2.73 percentage points. Compared to the error of the typical model, which was 2.30 percentage points, combining reduced error by 36% (Table 2). Compared to the error of forecasts from Abramowitz's model, the RAE of the combined forecast was 0.84, which means that forecast combining reduced error by 16% compared to the single model that performed best *ex post*.

Using all important variables with equal effect sizes

Similar to the tests of combining forecasts, the benefits from using all important variables could be tested only for US and Australian elections. Table 5 shows the error reductions achieved by using all available variables in an index model that weights the variables equally.

In the Australian case, the index model included a total of six variables: the five variables used by Cameron & Crosby (2000), plus one additional variable (i.e., a different measure of unemployment) used by Jackman (1994). The other two variables in the Jackman (1993) model, inflation and honeymoon, are already included in Cameron and Crosby. Across the 22 elections, the index model achieved an error of 2.35 percentage points, which is lower than the error of each individual model. Compared to the typical model, the index model reduced error by 10%.

In the US case, the index model consisted of 24 variables. While the total number of variables used in the eight models is 28, four variables were excluded. The reason is that the models by Fair (2009) and Cuzán (2012) use three identical variables, which are only included once. In addition, Fair (2009) uses a dummy variable to account for the World War II, which is unnecessary in our case, since we only model the 15 elections from 1956, for which data from all eight models are available. Across the 15 elections, the MAE of the index model forecasts was 1.32 percentage points, which is lower than the error of each individual models. Compared to forecasts from the typical model, the index model reduced error by 43%. Compared to forecasts from the best individual model, the index model reduced forecast error by 25%.

Table 5: Relative absolute errors from using all variables in an equal-weights index model, compared to the original regression models

	MAE (original)	RAE (index model)
Australia, 22 elections from 1951 to 2004, MAE of index model forecast: 2.35		
Cameron & Crosby (2000)	2,68	0,88
Jackman (1995)	2,54	0,92
Mean	***2,61***	***0,90***

US, 15 elections from 1956 to 2004, MAE of index model forecast: 1.32		
Abramowitz (2012)	1,76	0,75
Campbell (2012)	1,99	0,66
Cuzán (2012)	2,07	0,64
Erikson & Wlezien (2012)	2,54	0,52
Fair (2009)	2,49	0,53
Holbrook (2012)	2,55	0,52
Lewis-Beck & Tien (2012b)	2,29	0,58
Lockerbie (2012)	2,73	0,48
Mean	***2,30***	***0,57***

Discussion

We applied three conservative forecasting guidelines to the problem of forecasting election results: (1) modify effect estimates to reflect uncertainty, (2) combine forecasts from dissimilar models, and (3) include all variables that are important in the model.

For the first guideline, modifying effect estimates, we tested two approaches: damping and equalizing. Damping failed to provide substantive gains substantially harmed accuracy at high levels of damping. This suggests that regression is sufficiently conservative with respect to "regressing to the mean." We expect that these findings would apply to shrinkage but such testing was beyond the scope of this study. In contrast, equalizing the regression coefficients almost always improved forecast accuracy and reduced ex ante forecast error by 10% in comparison to the typical original model forecasts.

Applying the second guideline, combining forecasts, to eight US models and two Australian models produces more accurate forecasts than the most accurate model in each case. Compared to the typical model forecast, error was reduced by 36% in the US case and 14% in the Australian case, or 25% on average. The result are thus in line with the average of 22% error reduction for five comparative studies from different areas (including forecasts of economic variables) that examined combining across dissimilar causal models (Armstrong, Green, and Graefe 2015). The results are consistent with the guideline that forecasters should aim to include all important information in the forecast, rather than seeking to estimate statistically optimal effect sizes for a small set of selected variables from historical data. The "combine forecasts from dissimilar models" guideline is an established strategy for incorporating more information.

The third guideline, use all important variables, recommends an alternative approach to incorporating more information into a forecast model. Two index models provided forecasts that were more accurate than even the best individual model. Compared to the typical forecast, an index model that assigned equal weights to all unique variables available reduced forecast error by 10% in the Australian case and 43% in the US case, or 26% on average.

The gains from combining forecasts and using all important variables were achieved for election forecasting models that, for the most part, used similar variables. We expect that further gains in accuracy could be achieved by incorporating information from other important variables, such as biographical information (Armstrong and Graefe, 2011) about the candidates or perceptions of their issue-handling competence and leadership skills (Graefe, 2013).

The results demonstrate the importance of *a priori* analyses for model specification decisions. Unfortunately, this practice is on the decline, since (a) an *a priori* analysis requires the researcher to find and understand the relevant research, which is time consuming, expensive, and difficult, and (b) there are inexpensive alternatives. In particular, over the past half century, an ever-increasing abundance of data, computing power, and sophisticated statistical software have led researchers to increasingly adopt statistical methods (e.g., stepwise regression and data mining) to conduct specification searches. Ziliak and McCloskey (2004) provide evidence for this trend with their analysis of papers that were published in the *American Economic Review* in the 1980s and then in the 1990s. While 32 percent chose variables solely on the basis of statistical significance in the 1980s, 74 percent did so in the 1990s. While the resulting models best fit given data, they tend to be overfitted and thus perform poorly when predicting new data. In particular, the models tend to predict too much change. A review of recent research suggests that econometric forecasts from econometric forecasts are not sufficiently conservative.

The assumptions of regression analysis are seldom met in practice. The question about what is the best choice of method for developing a forecasting model cannot be answered by asserting the superior statistical properties of an optimal regression model. Our results support the contention that the implementation of *two* of the conservative forecasting guidelines will provide forecasts that are more accurate than those from regression models in practical forecasting situations. Specifically, trying to include (1) all important variables in a model and—in the absence of strong theory or evidence to the contrary—one can start with the assumption that they are (2) equally important. If one has good prior knowledge about the relative importance of variables, differential weights obtained from a priori analysis might improve the predictive validity of an index model. For example, weighting the importance of issues based on information from issue-salience polls reduced error by 28% relative to the error of forecasts from an equal-weights issue-index model (Graefe, 2013).

The gains in accuracy we report in this paper were achieved for election forecasting, a forecasting problem that involves little uncertainty and only modest complexity. Larger gains in forecast accuracy are likely from following the Golden Rule of Forecasting guidelines for econometric models involving complex problems that involve much uncertainty. Such problems include forecasting election outcomes in more volatile political jurisdictions, but also less-structured problems, such as forecasting the onset of political conflicts, the costs and benefits of government policies, and the long-term economic growth of nations. Comparative studies on the forecast validity of index models versus various types of regression models would help to assess the conditions under which index models can contribute to the forecast accuracy.

Conclusions

Two of the evidence-based conservative guidelines substantially improved on the accuracy of forecasts from established models. In out-of sample tests of forecast accuracy of election forecasting models, equalizing the coefficients of eight established econometric models reduced forecast error on average by 5%. Combining the eight original models' forecasts—i.e., the model outputs—reduced error by 25%. Two index models that use all unique variables that are included in two US and two Australian models reduced error on average by 26%.

The results suggest that the index method provides a substantial improvement to combining forecasts as a way to include all information in a forecasting procedure. Further research is necessary to validate whether these large gains in accuracy also hold for other forecasting problems.

Acknowledgements

{Show prior peer review.] ????????? Ray Fair??, Robin Hogarth??, John R. Lott??, Jason Dana?

Include the statement that each of the papers was read by at least one person (after making sure this is so.

References

Abramowitz, A. I. (2012). Forecasting in a polarized era: The time for change model and the 2012 presidential election. *PS: Political Science & Politics, 45*(4), 618–619.

Armstrong, J. S. (2012). Illusions in regression analysis. *International Journal of Forecasting, 28*(3), 689–694.

Armstrong, J. S., and Graefe, A. (2011). Predicting elections from biographical information about candidates: A test of the index method. *Journal of Business Research*, *64(7),* 699-706.

Armstrong, J. S., and Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting, 8(1)*, 69–80.

Armstrong, J. S., Green, K. C., and Graefe, A. (2015). Golden Rule of Forecasting: Be conservative. *Journal of Business Research, 68(8),* 1717-1731.

Bélanger, É., and Godbout, J.-F. (2010). Forecasting Canadian federal elections. *PS: Political Science & Politics, 43*(4), 691-699.

Bellucci, P. (2010). Election cycles and electoral forecasting in Italy, 1994–2008. *International Journal of Forecasting, 26*(1), 54-67.

Blumenthal, M. (2014). Polls, forecasts, and aggregators. *PS: Political Science & Politics*, 47(2), 297–300.

Cameron, L., and Crosby, M. (2000). It's the economy stupid: Macroeconomics and federal elections in Australia. *Economic Record, 76*(235), 354-364.

Campbell, J. E. (2012). Forecasting the presidential and congressional elections of 2012: The trial-heat and the seats-in-trouble models. *PS: Political Science & Politics, 45*(4), 630–634.

Campbell, A., Converse, P.E., Miller, W.E., and Stokes, D. E. (1960). *The American Voter*. New York: John Wiley and Sons.

Clarke, H. D., Kornberg, A., Scotto, T. J., Reifler, J., Sanders, D., Stewart, M. C., and Whiteley, P. (2011). Yes we can! Valence politics and electoral choice in America, 2008. *Electoral Studies*, 30(3), 450–461.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting,* 5(4), 559–583.

Cuzán, A. G. (2012). Forecasting the 2012 presidential election with the fiscal model. *PS: Political Science & Politics, 45*(4), 648–650.

Cuzán, A. G., and Bundrick, C. M. (2009). Predicting presidential elections with equally weighted regressors in Fair's equation and the fiscal model. *Political Analysis, 17*(3), 333–340.

Dana, J., and Dawes, R. M. (2004). The superiority of simple alternatives to regression for social science predictions. *Journal of Educational and Behavioral Statistics, 29*(3), 317–331.

Erikson, R. S., and Wlezien, C. (2012). The objective and subjective economy and the presidential vote. *PS: Political Science & Politics, 45*(4), 620–624.

Fair, R. C. (2009). Presidential and congressional vote-share equations. *American Journal of Political Science, 53*(1), 55–72.

Fair, R. C. (1978). The effect of economic events on votes for president. *The Review of Economics and Statistics, 60(2),* 159-173.

Graefe, A. (2015). Improving forecasts using equally weighted predictors. *Journal of Business Research*, 68(8), 1792-1799.

Graefe, A. (2013). Issue and leader voting in U.S. presidential elections. *Electoral Studies, 32*(4), 44–57.

Graefe, A., Armstrong, J. S., Jones, R. J. J., and Cuzán, A. G. (2014). Combining forecasts: An application to elections. *International Journal of Forecasting, 30*(1), 43–54.

Graefe, A., Küchenhoff, H., Stierle, V., and Riedl, B. (2015). Limitations of ensemble Bayesian model averaging for forecasting social science problems. *International Journal of Forecasting*, 31(3), 943-951.

Holbrook, T. M. (2012). Incumbency, national conditions, and the 2012 presidential election. *PS: Political Science & Politics, 45*(4), 640–643.

Jackman, S. (1995). Some more of all that: A reply to Charnock. *Australian Journal of Political Science, 30*(2), 347-355.

Lewis-Beck, M. S., and Tien, C. (2012a). Japanese election forecasting: Classic tests of a hard case. *International Journal of Forecasting, 28*(4), 797-803.

Lewis-Beck, M. S., and Tien, C. (2012b). Election forecasting for turbulent times. *PS: Political Science & Politics, 45*(4), 625–629.

Lewis-Beck, M. S., Nadeau, R., and Bélanger, E. (2004). General election forecasts in the United Kingdom: a political economy model. *Electoral Studies, 23*(2), 279-290.

Lockerbie, B. (2012). Economic expectations and election outcomes: The Presidency and the House in 2012. *PS: Political Science & Politics, 45*(4), 644–647.

Magalhães, P. C., and Aguiar-Conraria, L. (2009). Growth, centrism and semi-presidentialism: Forecasting the Portuguese general elections. *Electoral Studies, 28*(2), 314-321.

Magalhães, P. C., Aguiar-Conraria, L., and Lewis-Beck, M. S. (2012). Forecasting Spanish elections. *International Journal of Forecasting, 28*(4), 769-776.

Mayer, W. G. (2014). What, if anything, have we learned from presidential election forecasting? *PS: Political Science & Politics, 47*(2), 329–331.

Nadeau, R., and Blais, A. (1993). Explaining election outcomes in Canada: Economy and politics. *Canadian Journal of Political Science/Revue canadienne de science politique, 26*(4), 775-790.

Toros, E. (2011). Forecasting elections in Turkey. *International Journal of Forecasting, 27*(4), 1248-1258.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, 3(1), 23-40.

Ziliak, S. T. and McCloskey, D. N. (2004). Size matters: the standard error of regressions in the American Economic Review. *The Journal of Socio-Economics*, 33, 527–546.