A Vote Equation and the 2004 Election

Ray C. Fair*

November 22, 2004

1 Introduction

My presidential vote equation is a great teaching example for introductory econometrics.[1] The theory is straightforward, the data are easy to understand, the estimates are easily duplicated, and the equation, while fitting well, suffers from many potential econometric problems. It reveals both the strengths and weaknesses of econometric methodology. This paper discusses what we might (or might not) have learned from the equation about the 2004 election. In particular, should the Democrats be concerned that they are losing their grip?

*Cowles Foundation and International Center for Finance, Yale University, New Haven, CT 06520-8281. Voice: 203-432-3715; Fax: 203-432-6167; e-mail: ray.fair@yale.edu; website: *fairmodel.econ.yale.edu*. The original version of this paper used 51.6 percent as the actual value of vote. The final value was 51.2 percent, and this version of the paper uses the 51.2 number. Nothing else has been changed, and the date has been left the same.

[1] See Fair (1978) for the original equation and Fair (1996) for a significant update. The latest work is on the website: *http://fairmodel.econ.yale.edu/vote2004/index2.htm*. A non technical discussion is in Fair (2002). This paper only briefly discusses the equation; the reader is referred to the website for more details.

2 The Equation

The variable explained is the incumbent share of the two-party popular vote. Data on elections back to 1916 are used in the estimation. The theory is that the vote share is affected by the incumbency situation and the economy. The variables used are the following:

- $VOTE$ = Incumbent share of the two-party presidential vote.
- $PARTY$ = 1 if there is a Democratic incumbent at the time of the election and -1 if there is a Republican incumbent.
- $PERSON$ = 1 if the incumbent is running for election and 0 otherwise.
- $DURATION$ = 0 if the incumbent party has been in power for one term, 1 if the incumbent party has been in power for two consecutive terms, 1.25 if the incumbent party has been in power for three consecutive terms, 1.50 for four consecutive terms, and so on.
- WAR = 1 for the elections of 1920, 1944, and 1948 and 0 otherwise.
- $GROWTH$ = growth rate of real per capita GDP in the first three quarters of the election year (annual rate).
- $INFLATION$ = absolute value of the growth rate of the GDP deflator in the first 15 quarters of the administration (annual rate) except for 1920, 1944, and 1948, where the values are zero.
- $GOODNEWS$ = number of quarters in the first 15 quarters of the administration in which the growth rate of real per capita GDP is greater than 3.2 percent at an annual rate except for 1920, 1944, and 1948, where the values are zero.

The equation has $VOTE$ on the left hand side and the other variables plus a constant term on the right hand side. It is linear in coefficients and is estimated by ordinary least squares. The first set of estimates in Table 1 uses the 1916-2000 sample period (22 observations). (Ignore for now the second set of estimates.)

Table 1
Estimates of the Vote Equation

	Sample	
	1916–2000	1916–1960
$GROWTH$	.691	.805
	(6.72)	(7.94)
$INFLATION$	-.775	-.477
	(-2.71)	(-1.34)
$GOODNEWS$	.837	.701
	(3.12)	(2.90)
$PERSON$	3.25	5.21
	(2.50)	(4.26)
$DURATION$	-3.63	-2.08
	(-3.04)	(-2.34)
$PARTY$	-2.71	-3.58
	(-4.65)	(-6.23)
WAR	3.85	3.88
	(1.46)	(1.72)
INTERCEPT	49.61	47.36
	(18.08)	(21.88)
SE	0.0237	0.0147
R^2	0.923	0.987
No. obs.	22	12

The growth rate and the number of good news quarters have a positive effect on the vote share, and inflation has a negative effect. Regarding the incumbency variables, the vote share is positively affected if the President is running again and if the party in power is Republican. It is negatively affected if the party has been in power for two or more consecutive times. There is no theory as to why, other things being equal, the Republican party has a bias in its favor. This is just what the estimates show. The WAR variable is not really a war variable. Because $INFLATION$ and $GOODNEWS$ were zeroed out for the 1920, 1944, and 1948 elections, the constant term in the equation is different for these three elections, which is what WAR is picking up. It has nothing to do with wars like Korea,

Table 1 (continued)
Actual and Predicted Values of VOTE

		1916–2000		1916–1960	
	Actual	Predicted		Predicted	
Election	$VOTE$	$VOTE$	Error	$VOTE$	Error
1916	51.7	50.9	-0.8	50.9	-0.8
1920	36.1	39.2	3.1	36.4	0.2
1924	58.2	57.3	-1.0	57.6	-0.7
1928	58.8	57.6	-1.2	57.4	-1.4
1932	40.8	38.8	-2.1	41.2	0.4
1936	62.5	63.8	1.4	63.5	1.1
1940	55.0	55.7	0.7	55.4	0.4
1944	53.8	52.5	-1.2	53.8	0.1
1948	52.4	50.5	-1.8	52.1	-0.3
1952	44.6	44.4	-0.2	43.9	-0.7
1956	57.8	57.3	-0.5	57.6	-0.2
1960	49.9	51.6	1.7	51.8	1.9
1964	61.3	61.1	-0.3	59.5	-1.8
1968	49.6	50.2	0.6	49.2	-0.4
1972	61.8	59.4	-2.4	61.6	-0.2
1976	48.9	48.9	0.0	51.3	2.3
1980	44.7	45.7	1.0	45.8	1.1
1984	59.2	62.0	2.9	63.7	4.6
1988	53.9	51.3	-2.6	52.1	-1.8
1992	46.5	51.7	5.1	55.2	8.6
1996	54.7	53.7	-1.0	53.0	-1.8
2000	50.3	48.9	-1.3	47.1	-3.2

Vietnam, and Iraq.

The equation fits the data fairly well. This can be seen at the bottom of Table 1, where the actual and predicted values are presented. The predicted values use the actual values of the economic variables. They are ex post, within-sample predictions. The largest error occurred in the 1992 election. President Bush I got 46.5 percent of the two-party vote and was predicted to get 51.7 percent, which is an error of 5.1 percentage points. Otherwise, the errors are fairly small. The estimated standard error of the equation is 2.37 percentage points.

3 Potential Econometric Problems

It should be clear from the choice of the explanatory variables that a serious potential problem with this econometric work is the possibility of data mining. Much searching over many years has been done in arriving at the final version. The equation in Table 1 differs from the original version in Fair (1978): the specification of the equation has been changed over time to try to improve the fit as new observations have become available. The three variables that have been around in one form or another from the beginning are the growth rate, the inflation rate, and the person-running-again. $GOODNEWS$, $DURATION$, and $PARTY$ were added later. So an important question, which is interesting to pose to students, is how much weight, if any, should be placed on these results? Are they completely spurious?

If an equation fits the data well simply because it has been continually changed (by searching) to fit each new observation well, it is not likely to do well when estimated over shorter sample periods and these estimates used to predict the future. In other words, it should not do well in outside-sample tests. To test this in the present context, the vote equation was estimated only through 1960. These estimates are presented in Table 1. They are based on only 12 observations and 4 degrees of freedom. Given the small sample size, the equation appears to hold up fairly well. Even the good news variable, which was not chosen until after the 1992 election, has a coefficient estimate fairly close to the estimate in the equation estimated through 2000. The predictions of this equation are also fairly good beyond the end of the estimation period. Remember that the prediction for 2000,

for example, is outside sample by 40 years! The error for 1992 is the largest at 8.6 percentage points. The average of the absolute values of the 10 errors from 1964 through 2000 is 2.58 percentage points, which compares to 1.72 percentage points for the equation estimated through 2000. The outside-sample errors are thus on average only moderately larger than the within-sample ones.

The results using the shorter estimation period thus suggest that data mining might not be as serious a problem as one might otherwise have thought. Note that it is not poor science that the specification of an equation is changed over time if one is getting closer to the truth with each change. For example, the $GOODNEWS$ variable was not chosen (discovered?) until after the 1992 election, but this doesn't mean it was not important before. It's just that it had not yet been discovered. The estimates using the shorter sample period in Table 1 show that it should have been included from the beginning had it been known, since it is significant.

The specification of the vote equation has not been changed since the changes following the 1992 election. All that was done after the 1996 and 2000 elections was to reestimate the equation using the latest data. There have thus been no potential data mining issues since the 1992 changes.

A final econometric point that is useful to make to students is to note the difference between the vote equation here and vote equations that have survey variables among the explanatory variables. Survey variables include variables like presidential popularity and voting intentions. Equations that include survey variables may be useful for forecasting purposes, but they are not structural in the econometric sense. With survey variables one is essentially sampling ahead of time what people are likely to do when they vote. The variables are not "causal." The

same is true even for stock-market variables, where the same "fundamental" forces may be acting on both investors' views and voters' views. For the vote equation here the theory is that the state of the economy as measured by output growth and inflation influence voting behavior in a causal sense.

4 Real-Time Predictions for 2004

Prior to the 2004 election the equation was used to predict the vote share. Given the incumbency situation and forecasts of the economic variables, a vote prediction can be made. (Note that a vote prediction can be made long in advance of the election as long as forecasts of output growth and inflation are available.) For the 2004 election $PARTY$ is -1, $PERSON$ is 1, and $DURATION$ is 0. This is the best possible incumbency situation that an incumbent can have: a Republican running again with no duration effect. According to the equation, it takes a very poor economy to defeat an incumbent in this situation.

Table 2 presents the predictions that I made of the 2004 election. The first prediction is from Box 4-2, page 65, in Fair (2002). This prediction used the equation estimated only through 1996, not the estimates in column 1 of Table 1 above. The other predictions used the equation in column 1. The economic values for the last prediction are actual values as of October 29, 2004.

One noticeable feature of the vote prediction is that it did not change much over the three year period. As far back as November 2001 it was predicting that because of his good incumbency situation President Bush would be hard to beat if the economy were moderately good or better. Compare this stability to the large

Table 2
Real-Time Predictions of the 2004 Election

Date	Economic Forecasts Used *GROWTH*	*INFLATION*	*GOODNEWS*	Predicted *VOTE*	Actual *VOTE*	Error
November 2001	1.5	3.0	3	56.9	51.2	5.7
November 1, 2002	2.0	1.9	1	56.3	51.2	5.1
January 30, 2003	2.0	1.9	1	56.3	51.2	5.1
April 25, 2003	2.0	1.9	1	56.3	51.2	5.1
July 31, 2003	2.4	1.8	1	56.7	51.2	5.5
October 31, 2003	2.4	1.9	3	58.3	51.2	7.1
February 5, 2004	3.0	1.9	3	58.7	51.2	7.5
April 29, 2004	3.2	2.0	3	58.7	51.2	7.5
July 31, 2004	2.7	2.1	2	57.5	51.2	6.3
October 29, 2004	2.9	2.0	2	57.7	51.2	6.5

fluctuations in the polls even in the last few months before the election!

The other noticeable feature of the vote prediction is that it substantially overestimated Bush's actual vote share. The final error was 6.5 percentage points. This (outside-sample) error is slightly larger than the within-sample error of 5.1 percentage points for 1992. The vote equation says that with Bush's incumbency advantages and with the moderately good economic variables, he should have done much better than he did.[2]

This is where the econometrics gets interesting for students. What are we to make of the 6.5 percentage point error? If you are not convinced by the outside-sample results above and believe that the equation is spurious, then the error has no information content. (If you believe this, you have probably not read this far anyway.) Otherwise, if you think there may be something to the equation, the error does have information content. It is 2.7 times larger than the estimated standard

[2]Using the last (actual) economic values and the equation above estimated only through 1960, the predicted vote share is 58.9, which is an error of 7.7 percentage points. This error is slightly smaller than the error of 8.6 percentage points for 1992 for the same equation.

error of 2.37 percentage points, and it is 2.5 times larger than the average absolute error of 2.58 percentage points for the 10 outside-sample errors in Table 1 above. The error term in the equation reflects all the factors that affect the vote share that are not captured by the incumbency and economic variables. The large error in 2004 means that some other factors were important. Since there are many possible factors, it is not possible to test that one particular factor is responsible for most of the error in an election. There are many stories and one observation. My personal view is that were it not for Iraq and with the economy as it was, Bush would have come close to the equation's prediction, but again this cannot be tested.

What one can say, however, is that conditional on the vote equation being a good approximation, the Democrats did well. Bush should have won by more than he did, and so the Democrats need not be wringing their hands about the demise of the party. See Nordhaus (2004) for an interesting analysis of this.

5 Predictions for 2008

Another reason the Democrats should not be too depressed is that according to the equation the incumbency situation is much less favorable to the Republicans in 2008 than it was in 2004. There will be no person-running-again effect ($PERSON$ = 0), and there will be a negative duration effect ($DURATION$ = 1). If, say, $GROWTH$ is 3.0, $INFLATION$ is 3.0, and $GOODNEWS$ is 2, which is a moderately good economy, then using the first set of estimates in Table 1 above the vote prediction for the Republicans is 50.1 percent, a dead heat. So the main message for 2008 is that the election will be close if the economy is moderately

good. It would take a quite strong economy for the equation to predict a comfortable Republican win, and it would take a quite weak economy for the equation to predict a comfortable Democratic win. The Democrats clearly have a much better shot in 2008 than they had in 2004 according to the equation.

References

[1] Fair, Ray C., 1978, "The Effect of Economic Events on Votes for President," *Review of Economics and Statistics*, 60, 159-173.

[2] Fair, Ray C., 1996, "The Effect of Economic Events on Votes for President: 1992 Update," *Political Behavior*, 18, 119-139.

[3] Fair, Ray C., 2002, *Predicting Presidential Elections and Other Things*, Stanford: Stanford University Press.

[4] Nordhaus, William, 2004, "Profile of an Election: The Story of a Non-Mandate," November.

[5] website: http://fairmodel.econ.yale.edu/vote2004/index2.htm.