

**Working Paper: The lack of effect of electronic voting machines on change in support for Bush in the 2004 Florida elections**

**Alex Strashny**  
**Institute for Mathematical Behavioral Sciences**  
**University of California, Irvine**  
**[alex@strashny.com]**

**November 21, 2004**

***Motivation***

Because of the controversial Bush win in Florida in 2000, even before the 2004 Presidential elections, some people were concerned about voting irregularities in Florida that would favor Bush. The 2004 Presidential election was also the first election in which electronic voting machines were used in many states, including Florida. It is natural that the suspicion of voting irregularities was connected by some to this new voting method.

On November 18, 2004, several researchers from University of California - Berkley posted a working paper (Hout et. al.) which alleges, based on statistical analysis, that in Florida, the use of electronic voting machines increased the number of votes cast for Bush beyond the number that Bush "should have" received had the new voting machines not been used. The finding was covered in Wired News later that day (Zetter).

This present paper is a critique of Hout et. al.. A major error of the paper is that it does not account for the demographic shift that had occurred in Florida between 2000 and 2004. Once I account for this shift, I find that the use of electronic voting machines *did not* have any impact on the votes cast for Bush.

***Data***

The observations are made on the 67 of Florida's counties. The following data was collected by Hout et. al. and is available on the authors' website [[http://ucdata.berkeley.edu/new\\_web/VOTE2004/index.html](http://ucdata.berkeley.edu/new_web/VOTE2004/index.html)]: number of people voting for Bush and Gore in 2000, number of people voting for Bush and Kerry in 2004, whether the county used electronic voting machines in 2004.

Voter registration statistics for 2000 and 2004 are available from Florida Department of State [<http://election.dos.state.fl.us/voterreg/index.shtml>]. Note that the 2004 voter registration data is dated October 4, 2004, which was before the Election Day, November 2; the collection of the data did not involve any electronic voting machines.

From these raw data, I construct the following variables:

- %Bush Y: proportion of votes cast for Bush in year Y, counting only the votes cast for Bush and the Democratic candidate (Gore or Kerry)
- $\Delta\%Bush$ : %Bush 2004 - %Bush 2000
- EV: 1 if the county uses electronic voting machines, 0 otherwise
- Size: votes cast for Bush plus votes cast for Kerry in 2004
- P Y: the number of voters who registered for party P in year Y
- %P Y: the number of voters who registered for party P in year Y divided by the total number of registered voters
- $\Delta P$ : number of new registered voters in party P, relative to 2000:  $(P\ 2004 - P\ 2000) / P\ 2000$
- RepShift: shift of voters toward the Republican party:  $\Delta\%Republican - \Delta\%Democratic$

The idea behind the last two variables is that the new voters who register for a party will probably vote for that party's candidate.

The following table shows descriptive statistics for some of these variables.

	Mean	Std Dev	Min	Max
$\Delta\%Bush$	0.037	0.0289	-0.0296	0.1071
%Bush 2000	0.5629	0.0934	0.3145	0.7545
EV	0.2239	0.42	0	1
Size	111140	158960	3000	714360
% Republican 2004	0.3398	0.1264	0.0785	0.5718
% Democratic 2004	0.5083	0.1795	0.2436	0.8827
$\Delta\%Republican$	0.2935	0.2362	-0.0431	0.9699
$\Delta\%Democratic$	0.049	0.1272	-0.1882	0.4002
RepShift	0.2445	0.2998	-0.1391	0.9864

Note that the table shows a strong shift toward the Republican party in voter registrations: between 2000 and 2004, the number of registered Republicans increased by an average of 29%, whereas the number of registered Democrats increased by an average of only 5%. In three counties (Baker, Gilchrist, and Liberty), the number of registered Republicans increased by more than 90%.

### ***Replication of the Hout et. al. result***

The main model ("Model 1") in Hout et. al. uses the Ordinary Least Squares (OLS) model to regress  $\Delta\%Bush$  on  $\%Bush$  2000, [ $\%Bush$  2000 squared], Size, EV, [ $\%Bush$  2000 \* EV], and [ $\%Bush$  2000 squared \* EV]. The following table replicates the result:

<b><math>\Delta\%Bush</math></b>	<b>Coefficient</b>	<b>t-stat</b>
Constant	-0.2994	-3.3759
$\%Bush$ 2000	1.1024	3.5001
$\%Bush$ 2000 squared	-0.8492	-3.0643
Size	-9.10E-08	-3.5496
EV	0.4941	3.2563
$\%Bush$ 2000 * EV	-1.4777	-2.6046
$\%Bush$ 2000 squared * EV	1.0259	1.9321

$$R^2 = 44.9\%$$

$$SC = -7.2578$$

I also report the Schwarz Criterion (SC), which Hout et. al. does not report. I discuss its use further below.

Because the coefficient on EV is positive and statistically significant, the regression seems to suggest that there is a relationship between the use of electronic voting machines and a boost in the votes received by Bush. Because the coefficient on [ $\%Bush$  2000 \* EV] is negative and statistically significant, the regression seems to suggest that most of the boost related to the electronic voting machines comes from "Democratic counties" (that is, counties where Bush received a lower percentage of votes in 2000). In fact, this is the interpretation made in Hout et. al..

## **Model construction and model selection**

Contrary to popular belief, data does not "speak for itself". In statistics, data is always seen through the prism of a model. What we see from the data depends on the data itself and on the model through which it is analyzed.

Where do models come from? Ideally, a model comes from an established scientific theory. When such a theory does not exist, the model must come from logical considerations. A model *does not* properly come from trying to use many different independent variables to see which ones maximize "fit".

If there are several plausible models, a "best" model can be selected. One popular model selection method uses the Schwarz Criterion (SC) (see Schwarz 1978). Among several models, the model with the lowest SC is the best.

Note that using an incorrect model can lead to incorrect interpretation of the data, even if all the estimates are "statistically significant" and there is a high "fit" as measured by  $R^2$ . The model with a higher  $R^2$  is not necessarily better.

## **A critique of logic**

What is the theory or logic behind the Hout et. al. model? It is never explained in the paper. The choice of independent variables seems rather arbitrary.

Why is Size used as one of the independent variables? Why is the population of a county, an *absolute* value, related to the change in the *proportion* of votes cast for Bush?

Why are the two variables containing [%Bush 2000 squared] in the model?

Why is [%Bush 2000] in the model? Why is it reasonable to think that a large proportion of Bush voters in 2000 is associated with an increase (or a decrease) in the proportion of Bush voters? After all, [%Bush 2000] is a *level*, whereas the dependent variable,  $\Delta\%$ Bush, is a *change*.

Most importantly, why doesn't the model use any independent variables that contain 2004 data? Four years is a long time. Perhaps there was a demographic shift between 2000 and 2004? The model should account for it.

## Another model

In my view, explaining the change in votes cast for Bush by new Republican and Democratic voters more logical.

% $\Delta$ Republican measures the number of new registered voters who are Republican, relative to 2000. Likewise,

% $\Delta$ Democratic measures the number of new registered voters who are Democrats.

My idea is that, especially in the 2004 election, when both parties made a strong appeal to voters to come out and vote, the change in Bush support was driven by the new voters. Because population increases with time, it's possible that there is both an increase in the number of people who support Republicans and in the number of people who support the Democrats. That is, it's possible that a county has both a positive % $\Delta$ Republican and a positive % $\Delta$ Democratic. I thus calculate RepShift, the shift toward the Republican party, as the difference between these two variables.

In Model A, I regress  $\Delta\%$ Bush on RepShift; in Model B, I regress  $\Delta\%$ Bush on RepShift and EV; finally, in Model C, I regress  $\Delta\%$ Bush on RepShift and [EV \* %Democratic 2004]. The table below summarizes the results.

Model	A		B		C	
$\Delta\%$ Bush	Coeff	t-stat	Coeff	t-stat	Coeff	t-stat
Constant	0.0237	6.2252	0.024	5.045	0.0233	4.9528
RepShift	0.0545	5.5132	0.054	4.9577	0.0551	5.1022
EV or EV * %Dem 2004			-0.0009	-0.1124	0.0033	0.1562
R <sup>2</sup>	31.86%		31.88%		31.89%	
SC	-7.3584		-7.2958		-7.296	

Note the following:

- The sign of the coefficient on RepShift is positive, as logic dictates. The coefficient is highly statistically significant.
- In Models B and C, the added EV term is statistically insignificant.
- The Schwarz Criterion for all three models is slightly less than the Schwarz Criterion for the Hout et. al. model.

Model A is better than the Hout et. al. model, both logically, and on the basis of Schwarz Criterion, a standard model selection method. Conditional on RepShift, the electronic voting variables are not correlated with the change in the percent of votes cast for Bush. This means that the use of electronic voting machines did not boost the Bush vote (Model B), and that this boost did not occur in Democratic counties (Model C).

### ***More errors that impact the conclusion***

There are several other problems with Hout et. al.. For example, it uses data mining. I will write more on these problems later...

### ***Conclusion***

A correct analysis of the data indicates that the use of electronic voting machines had no impact on the Bush vote in Florida.

Hout, M., Mangels, L., Carlson, J., and Best, R. (2004)  
Working Paper: The Effect of Electronic Voting  
Machines on Change in Support for Bush in the 2004  
Florida Elections. (Working paper)  
[[http://ucdata.berkeley.edu/new\\_web/VOTE2004/index.html](http://ucdata.berkeley.edu/new_web/VOTE2004/index.html)].

Schwarz, G. (1978) Estimating the Dimension of a Model.  
*Annals of Statistics*, 6, 461-464.

Zetter, K. (November 18, 2004) Researchers: Florida Vote  
Fishy. *Wired News*.  
[<http://www.wired.com/news/evote/0,2645,65757,00.html>]