A Partial Critique of Hout, Mangels, Carlson and Best's "The Effect of Electronic Voting Machines on Change in Support for Bush in the 2004 Florida Elections"

by
B. D. McCullough
Department of Decisions Sciences and Department of Economics
Drexel University
and
Florenz Plassmann
Department of Economics
SUNY Binghamton

2 December 2004

> "It is commonly believed that anyone who tabulates numbers is a statistician. This is like believing that anyone who owns a scalpel is a surgeon. A statistician is one who has learned how to get valid evidence from statistics and how (usually) to avoid being misled by irrelevant facts. It's too bad that we apply the same name to this kind of person that we use for those who only tabulate. It's as if we had the same name for barbers and brain surgeons because they both work on the head."
>
> Robert Hooke, *How to Tell the Liars from the Statisticians*
> New York, NY: Marcel-Dekker, Inc., 1983, page 1

MOTIVATION

The recent working paper by Hout, Mangels, Carlson and Best ("HMCB")[1] concludes that "electronic voting raised President Bush's advantage from the tiny edge he held in 2000 to a clearer margin of victory in 2004." In particular, the paper's authors allege that Bush received somewhere between 130,000 and 260,000 "excess votes" that were not actually cast. We are suspicious of persons who call press conferences to release "results" that have not been peer-reviewed. As professors who teach statistics and econometrics to undergraduate and graduate students, we are always on the lookout for good examples of "what not to do" so that we may better instruct our students in the responsible use of statistics. Therefore we have examined the HMCB study with a critical eye. We conclude that the study is entirely without merit and its "results" are meaningless.

[1] The paper and data are available for inspection at http://ucdata.berkeley.edu/new_web/VOTE2004/

An easy way to show that there is something seriously wrong with a statistical study is to use the same data and the same approach to reach the opposite conclusion. HMCB apparently never bothered to check this aspect of their model, else they'd have easily found it: we show that HMCB's modelling approach also supports the contention that electronic voting favored Kerry.

Before we can do this, however, we must present a basic modelling error made by HMCB. HMCB claim to estimate two models, one for the change in support for Bush between 2000 and 2004, and another that estimates Bush's percentage of the vote in 2004. The former is critical to their calculation of "excess votes". In fact, they only estimate the latter. Therefore, their calculation of "excess votes" is illusory.

HMCB'S MODELLING ERROR

HMCB claim to have two models, one that predicts the difference between Bush's proportions of the vote in 2004 and 2000 (we call this "Model A"), and another that predicts Bush's proportion of the vote in 2004 ("Model B").[2] In the former case, the dependent variable is $p_{B04} - p_{B00}$ and in the latter case it is p_{B04}.

Model A

$$p_{B04} - p_{B00} = c + b_1 p_{B00} + b_2 p_{B00}^2 + b_3 \text{size} + b_4 \text{EV} + b_5 p_{B00} * \text{EV} + b_6 p_{B00}^2 * \text{EV}$$

Model B

$$p_{B04} = c + b_1 p_{B00} + b_2 p_{B00}^2 + b_3 \text{size} + b_4 \text{EV} + b_5 p_{B00} * \text{EV} + b_6 p_{B00}^2 * \text{EV}$$

where

p_{B04}	proportion voting for Bush in 04
p_{B00}	proportion voting for Bush in 00
size	total votes for Bush and Kerry
EV	electronic voting, 1 if the county has it, 0 otherwise

[2] HMCB's Table 2 presents two versions, short and long, of Model A; Model B is discussed in their text and shown in their Table 6.

The only difference between the two models is the "$-p_{B00}$" on the left hand side of the equals sign in Model A. As any student of linear models or econometrics knows, subtracting a right hand side variable from the left hand side does not alter the model (though it does change the estimated coefficient for that particular variable by one unit – it is trivial to show this). In other words, Model A and Model B both describe the same analysis of p_{B04}: HMCB *do not* have two models, one for the change in Bush's support and another for Bush's support; all they have is a model for Bush's support, albeit two times.

All this is obvious to any discerning, statistically-trained reader of HMCB's paper. For the benefit of those who are not statistically-trained, our Table 1 presents the results of two estimations, one with the change in Bush's support ($p_{B04} - p_{B00}$) as the dependent variable, and the other with Bush's proportion of the vote in 2004 (p_{B04}) as the dependent variable. We did not have to run these regressions, as these results are in HMCB's paper. The first column of our Table 1 comes from the first column of their Table 2. The second column of our Table 1 comes from the first column of their Table 6.

As can be seen, the coefficients are exactly the same, except for the coefficient on p_{B00}, for which the second estimation has a value one unit greater than the first estimation, 2.102 instead of 1.202. As we mentioned earlier, it is trivial to show that this is the effect of subtracting a right-hand side variable from the left-hand side.

Aside and apart from the modelling error, HMCB's two models, taken together, make no intuitive sense; if they had interpreted each model, they would have realized this. HMCB claim that these are two different models explaining two different phenomena. Yet according to HMCB, the variables p^2_{B00}, size, EV, $p_{B00}*$EV, and $p^2_{B00}*$EV have the same effect (*i.e.,* have the same coefficients) on both the change in Bush's support and Bush's support. That the coefficients are the same simply makes no sense – for example, if 'size' has an effect of some magnitude on the *proportion* of the vote that Bush got in 2004, it must have an effect of smaller magnitude on the

	Model A (from HMCB's Table 2)	Model B (from HMCB's Table 6)
	dependent variable	
	$p_{B04} - p_{B00}$	p_{B04}
p_{B00}	1.102	2.102
p_{B00}^2	-0.849	-0.849
size	-9.13E-8	-9.13E-8
EV	0.494	0.494
p_{B00}*EV	-1.478	-1.478
p_{B00}^2*EV	1.026	1.026
constant	-0.299	-0.299

Table 1: Coefficients from HMCB's Models

difference between his 2000 and 2004 proportions (because the proportion is a relatively large number and the difference between two proportions is a relatively small number). Intuitively, the effect of 'size' should not be the same on both Bush's support and the *change in* Bush's support.[3] Obviously, neither model estimates the percentage change in support for Bush, though *both* models (purport to) estimate the percentage of the votes for Bush in 2004.

HMCB and their reviewers were all but aware of this. Their Appendix A notes: "Other reviewers asked if the results would be different if we used the simpler percent for Bush instead of percent change in Bush support as the dependent variable. Since we used percentage of votes for Bush in 2000 as an independent variable, it essentially appears twice in our regression equations." HMCB ran the two regressions, ostensibly to compare them, and either did not notice that five of the six coefficients were identical and the sixth was off by exactly 1.0 or, if they did notice, did not recognize this as something that needed to be explained. We believe that any undergraduate student of linear models or econometrics asked to compare the two models would have caught this – certainly, our students would be severely marked down for such an elementary error.

HMCB's concept of "excess votes" is based on their model of the change in Bush's support. But since HMCB did not actually estimate such a model, their entire concept of "excess votes"

[3]Unless, of course, 'size' had absolutely no effect on the 2000 proportion. But how can 'size' affect the 2004 proportion without affecting the 2000 proportion?

is empirically vacuous.

In the sequel, we do not pretend to model the difference between proportions in 2004 and 2000, and instead focus only on the proportion in 2004. But we focus on Kerry's vote, not Bush's.

A MODEL FOR KERRY'S PROPORTION OF THE VOTE

We estimate for Kerry a model similar in spirit to that which HMCB estimated for Bush[4]:

Our Model

$$p_{K04} = c + b_1 p_{G00} + b_2 \text{EV} + b_3 \text{size} + b_4 p_{D96} + b_5 p_{G00} * \text{EV} + b_6 p_{G00} * \text{size}$$

$$+ b_7 \text{EV} * \text{size} + b_8 \text{EV} * \text{size} * p_{G00}$$

where p_{K04} is the proportion of the vote that went to Kerry in 2004, p_{G00} is the proportion of the vote that went to Gore in 2000, and p_{D96} is the proportion of the vote that went to Dole in 1996. The results for this model are given in Table 2: all coefficients are statistically significant at conventional levels. In HMCB's model, the cofficient on electronic voting is positive and significant and electronic voting helps Bush. In the model we present, the cofficient on electronic voting is also positive and significant and electronic voting helps Kerry, not Bush. Of course, we could add other variables to make the electronic voting coefficient negative or zero. The point is that, theoretically, our model is at least as valid as HMCB's, and ours has a higher goodness of fit[5] than theirs.

Both models can't be correct: electronic voting cannot help both Bush and Kerry. We could go through the motions of calculating "excess votes" for Kerry in the same way that HMCB

[4]We acknowledge and appreciate an objection raised by Professor Hout to a different nonsense model that we presented in an earlier draft.

[5]For the more technically inclined, here we refer to R^2, and we are well-aware that it is not the end-all-be-all of model specifi cation. We are also aware that our model strains the limits of credulity, but we wish to drive home the point that even nonsense can be statistically signifi cant.

variable	coefficient	t-statistic
p_{G00}	1.498	16.54
EV	0.154	2.98
size	8.5E-7	3.18
p_{D96}	0.282	3.05
p_{G00}*EV	-0.363	-3.06
EV*size	-6.6E-7	-2.16
p_{G00}*size	-1.8E-6	-3.12
p_{G00}*EV*size	1.6E-6	2.49
constant	-0.400	-4.94
	R^2= 0.9670	

Table 2: Results for the Kerry Model

calculate "excess votes" for Bush, but there is no point. The notion of interpreting coefficient estimates and calculating "excess votes" requires that there first be a reliable statistical model.

However, there is no reliable statistical model so far, and we do not even pretend that our model is correct; it is fatally flawed for many reasons, and so is HMCB's model. We shall spare the reader a discussion of technical details. We will be more than happy to take the issue up in a journal, however, if HMCB wish to engage us in a debate.

ADDITIONAL DOUBT ABOUT HMCB'S MODEL

Since, as we have shown, HMCB do not model the change in Bush's support, we now focus on their model for support for Bush in 2004, about which we have grave doubts. We have no idea how HMCB selected their variables. It appears to us that they neglected to consider many relevant variables. The press has written much of how Christian Evangelicals who stayed home in 2000 came out to vote in 2004, yet we see no attempt to incorporate this into their model. Republican "get-out-the-vote" was much larger and, apparently, more successful in 2004 than in 2000; again, HMCB omit this important variable from their model. Either of these variables could explain much of the ephemeral "excess votes" that HMCB claims Bush received.

Also, exploratory regressions suggest the possibility of a nonlinear relationship. For example,

replacing the variable 'size' by its constituent components ('number of 2004 votes for Bush' and 'number of 2004 votes for Kerry') reduces the size of the coefficient of the electronic voting variable by 2/3 and renders it statistically insignificant. At the very least, HMCB should justify their use of a linear regression when a nonlinear regression seems more appropriate.

We are generally concerned about the formal method – or lack thereof – whereby HMCB settled on their models. It appears that they just ran regressions until they got the answer they wanted. On what basis did they include/exclude variables from his models? Comparing their Model 1 and Model 2 (their Table 2), they added the variable "proportion voting for Dole in 96" and found its coefficient to be insignificant. Yet, just adding "number of votes for Dole" instead of "proportion voting for Dole" has a very profound effect on Model 2: not only is the number of votes for Dole statistically significant, but the goodness of fit increases from 0.538 to 0.602 (an appreciable increase) and the coefficient on the electronic voting variable switches from positive and significant to negative and insignificant! One can only wonder why HMCB included "proportion for Dole" instead of "votes for Dole". HMCB need to explain why they chose to report the worse fitting model with the positive coefficient instead of the better fitting model with the negative coefficient.

Even if their model had none of the above flaws, there would still be grounds for questioning their results. HMCB have confused the alleged effect of electronic voting with a county-specific effect. If their claim about electronic voting is true, it should be true for all counties. They were all but aware of this, They write (page 1), "[The impact of electronic voting] was especially large in Broward, Palm Beach, and Miami-Dade." Actually, it appears that those were the only counties in which there was an "effect of electronic voting." If we delete these three counties from the dataset, then the coefficient on electronic voting becomes zero: there is no effect of electronic voting in the other counties, *according to HMCB's own model*!

Much more plausible than some problem with the electronic voting machines in those three

counties is the notion that something else happened in those counties. Perhaps this is where Evangelical Christians reside. The media have also reported that Bush got a higher proportion of the Jewish vote in 2004 than in 2000; perhaps these three counties have a large Jewish population. Perhaps the GOP concentrated its resources in these three counties, since they are contiguous. There is any number of explanations more plausible than voting machine fraud, yet HMCB made no attempt to investigate these possibilities.

SUMMARY

HMCB's concept of "excess votes" depends critically on a model for the change in Bush's support, and we have shown that HMCB estimated no such model; they actually just subtracted a right hand side variable from the left hand side of their model for Bush's support in 2004. As to their model for Bush's support in 2004, it has been discredited. Therefore, all the claims that HMCB made are without any foundation. We specifically rebut every single point that they made in their summary document (which was separate from their paper):

- Since HMCB's concept of "excess votes" is empirically meaningless, they did not uncover statistical irregularities associated with the electronic voting machines used in Florida.
- We have shown that HMCB were not modelling increases for support for President Bush between 2000 and 2004, but instead were only modelling support for Bush in 2004. Therefore, there can be no evidence that counties with electronic voting machines were significantly likely to show increases in support for President Bush between 2000 and 2004.
- There is no credible, statistical basis for asserting the existence of "excess votes", hence President Bush cannot be said to have received 72,000 of them in Broward County.
- In their summary, HMCB claim that "We can be 99.9% sure that these effects are not attributable to chance." We respectfully disagree: we are 99.999% sure that HMCB's paper constitutes no evidence that there was anything amiss with the electronic voting in Florida.