Statistical Science
2002, Vol. 17, No. 4, 441–457

A Statistical Assessment of Buchanan's Vote in Palm Beach County

Richard L. Smith

Abstract. This article presents a statistical analysis of the results of the 2000 U.S. presidential election in the 67 counties of Florida, with particular attention to the result in Palm Beach county, where the Reform party candidate Pat Buchanan recorded an unexpectedly large 3,407 votes. It was alleged that the "butterfly ballot" had misled many voters into voting for Buchanan when they in fact intended to vote for Al Gore. We use multiple regression techniques, using votes for the other candidates and demographic variables as covariates, to obtain point and interval predictions for Buchanan's vote in Palm Beach based on the data in the other 66 counties of Florida. A typical result shows a point prediction of 371 and a 95% prediction interval of 219–534. Much of the discussion is concerned with technical aspects of applying multiple regression to this kind of data set, focussing on issues such as heteroskedasticity, overdispersion, data transformations and diagnostics. All the analyses point to Buchanan's actual vote as a clear and massive outlier.

Key words and phrases: Binary data, butterfly ballot, Florida election, heteroskedasticity, Monte Carlo tests, multiple regression, overdispersion, regression diagnostics, transformations.

1. INTRODUCTION

The 2000 U.S. presidential election was eventually settled when George Bush officially beat Al Gore by 537 votes, out of nearly 6 million cast, in the state of Florida. This gave Bush the 25 Electoral College votes he needed to become President under the U.S. Constitution. During the month that it took to resolve this election, many allegations of voting irregularities were made. One of the strongest arguments concerned Palm Beach county, the second largest of Florida's 67 counties, with over 400,000 votes cast for a total of 10 candidates. Here, it was alleged, the "butterfly ballot" design had misled many voters into voting for a different candidate from the one they intended. Specifically, it was claimed that many voters who had intended to vote for Al Gore in fact voted for Pat Buchanan, the candidate of the Reform party. The initial returns gave Buchanan 3,407 votes in Palm Beach, compared with about 1,300 that he might have expected based on his overall vote in the state. The excess of about 2,000 votes, had they indeed been credited to Gore, would have given Gore the state of Florida and hence the White House.

Within a few days of the election, numerous articles appeared on various websites, containing statistical analyses of the election results. Many of these articles focussed specifically on the Buchanan vote in Palm Beach county, using various political or demographic covariates to predict the number of votes he should have received. Most such analyses in fact put the predicted number of Buchanan votes at considerably less than 1,300. For example, Palm Beach overall is a Democratic county—Al Gore gained 62% of the total vote there—so one would have expected the right wing Buchanan to gain a smaller percentage of votes than he did over the whole state. Comparisons like this have generally been taken as strengthening the argument that Gore should have won, but they also raise questions over the correct application of regression methodology to the results of an election.

Richard Smith is Professor of Statistics in the Department of Statistics, University of North Carolina, Chapel Hill, North Carolina 27599-3260 (e-mail:rls@email.unc.edu).

The purpose of this article is to examine this hypothesis in detail by presenting a detailed regression analysis of the number of Buchanan votes across all 67 Florida counties, using both the votes for other candidates and demographic variables obtained from the Census as covariates. The analysis devotes particular attention to the problem of heteroskedasticity, an inevitable problem when there is such a large disparity in the populations of the various counties (from under 7,000 in Lafayette and Liberty counties, to over 2 million in Miami–Dade county). I consider various transformations of the data as remedies against this. Other questions considered include variable selection, overdispersion in a Poisson or binomial analysis of the data, and various forms of diagnostics and tests of fit. When applied to the prediction of Buchanan's Palm Beach vote based on the vote in the other 66 Florida counties, all the analyses result in point and interval predictions well short of his actual vote, and confirm that it was an enormous outlier.

The motivation for the present article is to draw attention to statistical rather than political issues. Even if it were proved that some voters were misled by the ballot design, there appears to be no legal or constitutional basis for overturning the result of an election on the basis that the results were inconsistent with some statistical prediction, and indeed the Gore team acknowledged this when they decided not to pursue a formal legal challenge based on the Palm Beach ballot design (instead they pursued a claim to have the votes in a number of Florida counties recounted by hand, the issue on which the U.S. Supreme Court eventually decided the election in Bush's favor). Another point is that the present analysis presents no evidence that the excess votes came from Gore's supporters rather than Bush's, though some other commentators have argued that point; see the review in Section 2.

On the other hand, the enormous public interest in the outcome of the Florida vote resulted in a number of statistical analyses of the results, many of which used regression analysis in some form. Most of these analyses did not go through a careful process of model building and model checking. No claim is made that the present analysis is in any sense definitive, but my hope is that it may serve as a positive example of how modern regression techniques may be applied to a question of real public interest.

An outline of the article is as follows. Section 2 is a review of other analyses that have appeared since the election. Section 3 describes the data used in this analysis. Section 4 goes through the various stages of building a regression model and discusses in detail such issues as homoskedasticity, normality of the errors, selection of variables, and outlier and influence diagnostics. Section 5 discusses the actual predictions, and Section 6 presents an alternative analysis that treats the votes cast as binary data. Finally, Section 7 presents a summary and conclusions.

Data and programs used in the present article are available from the World Wide Web address http://www.stat.unc.edu/faculty/rs/palmbeach.html.

2. REVIEW OF OTHER ANALYSES

Within a few days of the election, numerous statistical analyses appeared on various websites, and much has been published since then. It is not feasible to attempt a review of all of these, but this section provides an overview.

Adams (2001) provided a number of data plots to illustrate the point that Buchanan's vote in Palm Beach was highly anomalous. For example, he plotted the Buchanan votes in each of the 67 Florida counties against total votes cast, against the number of registered Reform party voters and against the 1996 votes for Buchanan in the Republican primary. All of these plots showed a clear outlier in Palm Beach. In an attempt to make the case that the butterfly ballot was specifically responsible for the anomaly, he also made a similar plot for the Socialist and Green party candidates—the Socialist candidate also achieved a much higher than expected vote in Palm Beach and since the Socialist candidate's name appeared immediately below Buchanan's on the ballot paper, it seems likely that voters were also misled into voting for the Socialist candidate by mistake. Adams' published work refrained from any detailed regression analysis, citing concerns such as the effect of different populations in the counties (an earlier World Wide Web version of the paper did include such analyses). Adams also did not consider any form of multiple regression analysis.

Other World Wide Web analyses attempted to take account of varying population size by various methods, for example, by assuming the variance to be inversely proportional to population size, or by some transformation, such as log transformation, or by considering proportions of votes for different candidates rather than total votes cast (such transformations do not remove the problems associated with heteroskedasticity, but they are better than simple linear regression applied to actual vote counts). Examples of such analyses include

Carroll (2000), Hansen (2000) and Monroe (2000). Although most of these analyses support the general contention that Buchanan's vote was an outlier, there are some analyses that express a contrary opinion (e.g., Farrow, 2000).

Agresti and Presnell (2001) summarized a number of the statistical issues in a paper written for lawyers. As examples, they made comparisons of Buchanan's vote in 2000 with the votes cast for Ross Perot in the 1996 presidential election (Perot was the Reform party's candidate in 1996), with the Buchanan vote in the 1996 Republican primary, with Bush and Gore votes in 2000, and with Reform party registration totals. They also drew attention to the Socialist Party's unusually large vote and discussed results from individual precincts within Palm Beach county (noting the positive correlation between Buchanan and Gore votes), from overvotes (ballots that were invalidated through voting for more than one candidate) and undervotes (ballots where no vote was made for President). All of these were argued as supporting the case that the anomalous vote for Buchanan was specifically a consequence of the butterfly ballot and that the excess votes came from Gore rather than Bush.

The most detailed and comprehensive analysis I have seen of the Palm Beach result is the article by Wand et al. (2001). Their abstract directly claimed that "the butterfly ballot ... caused more than 2,000 Democratic voters to vote by mistake for Reform candidate Pat Buchanan." They used county-level data to study the Buchanan vote not only in the whole of Florida, but also repeated the regression study in each of the other states (some of the smaller states were lumped together to create sufficiently large sample sizes). They argued that Palm Beach produced the most anomalous vote for Buchanan among all the 3,053 U.S. counties that they examined. To substantiate the claim that this anomaly was due to the butterfly ballot and that it took votes away from Gore rather than Bush, they used a number of other data sources. They looked at precinct-level data for Florida, they looked at election day versus absentee ballots (the absentee ballot in Palm Beach did not use the butterfly format and did not show any anomalous support for Buchanan) and they also looked at individual ballots in Palm Beach (a substantial number of the votes for Buchanan came from voters who voted Democratic for the other major offices on the ballot).

Concerning the actual regression analysis used by Wand et al., the main analyses used votes for Buchanan as the response variable together with a number of covariates including votes for other candidates in the 2000 presidential election, votes for the Republican and Reform candidates in the 1996 presidential election, and a set of demographic variables. Instead of considering the demographic variables individually, as the following analysis does, they performed a principal components analysis, and the leading principal component was treated as a single measure of demographic variability. For the analysis itself, the authors noted the twin problem of overdispersion (i.e., that the variance of the countywide Buchanan votes is much larger than could be explained by a simple binomial or Poisson model) and outliers, and proposed an apparently original method of "robust estimation of the overdispersed binomial regression model." Although their robust analysis indeed appears to reduce the influence of outliers, nowhere in their article did they attempt to validate in detail the assumptions of their model, based on the data.

The present analysis is restricted to the county-level data for Florida and examines Buchanan's vote as predicted by two types of covariates: (a) votes for the other candidates in the same election and (b) demographic variables. Variable selection methods are used to reduce the dimensionality of the covariates. The remainder of the analysis is concerned with detailed building of a regression model and the use of diagnostic techniques to examine the fit of the model, together with the role and possible influence of outliers.

3. DATA USED IN THIS STUDY

Two data sets have been assembled, one consisting of election returns from the Florida Division of Elections, and the other of demographic data compiled from the U.S. Census Bureau. Tables 1 and 2 give the votes for the 10 presidential candidates on the ballot in Florida, classified by county. These data are the provisional results issued immediately after the election, which may differ slightly from the final certified totals that were issued only after the first version of the present analysis had been completed. In addition to the votes for Bush, Gore and Buchanan, I also include in the analysis the votes for the Libertarian candidate Browne and the Green party candidate Nader. The other five candidates (Harris, Hagelin, McReynolds, Phillips and Moorehead) all achieved less than 0.1% of the vote and shall play no part in the analysis, although they may be of some interest for determining whether other minor-party candidates also showed unusual voting patterns, an aspect we shall not consider

TABLE 1
County voting data, part I

County	Bush	Gore	Brow	Nade	Har	Hag	Buc	Mc	Ph	Mo
Alachua	34,124	47,365	658	3226	6	42	263	4	20	21
Baker	5,610	2,392	17	53	0	3	73	0	3	3
Bay	38,637	18,850	171	828	5	18	248	3	18	27
Bradford	5,414	3,075	28	84	0	2	65	0	2	3
Brevard	115,185	97,318	643	4470	11	39	570	11	72	76
Broward	177,323	386,561	1212	7101	50	129	788	34	74	124
Calhoun	2,873	2,155	10	39	0	1	90	1	2	3
Charlotte	35,426	29,645	127	1462	6	15	182	3	18	12
Citrus	29,765	25,525	194	1379	5	16	270	0	18	28
Clay	41,736	14,632	204	562	1	14	186	3	6	9
Collier	60,433	29,918	185	1399	7	34	122	4	10	29
Columbia	10,964	7,047	127	258	1	7	89	2	8	5
Desoto	4,256	3,320	23	157	0	0	36	3	8	2
Dixie	2,697	1,826	32	75	0	2	29	0	3	2
Duval	152,098	107,864	952	2757	37	162	652	15	58	41
Escambia	73,017	40,943	296	1727	6	24	502	3	110	20
Flagler	12,613	13,897	60	435	1	4	83	3	3	12
Franklin	2,454	2,046	17	85	1	3	33	0	3	2
Gadsden	4,767	9,735	24	139	3	4	38	4	7	6
Gilchrist	3,300	1,910	52	97	0	1	29	0	2	4
Glades	1,841	1,442	12	56	0	3	9	1	0	1
Gulf	3,550	2,397	21	86	2	4	71	2	2	9
Hamilton	2,146	1,722	12	37	4	1	23	8	7	4
Hardee	3,765	2,339	17	75	0	2	30	0	2	3
Hendry	4,747	3,240	11	103	3	1	22	2	7	2
Hernando	30,646	32,644	116	1501	8	26	242	4	10	22
Highlands	20,206	14,167	64	545	6	16	127	3	7	8
Hillsborough	180,760	169,557	1138	7490	35	217	847	29	68	154
Holmes	5,011	2,177	18	94	1	7	76	3	6	2
Indian River	28,635	19,768	122	950	4	13	105	2	13	10
Jackson	9,138	6,868	40	138	0	2	102	1	4	7
Jefferson	2,478	3,041	14	76	2	1	29	1	0	0
Lafayette	1,670	789	6	26	2	0	10	1	1	0
Lake	50,010	36,571	204	1460	4	36	289	1	21	15
Lee	106,141	73,560	538	3587	30	81	305	5	34	96
Leon	39,053	61,425	330	1932	9	28	282	7	16	31
Levy	6,858	5,398	92	284	1	1	67	1	10	12
Liberty	1,317	1,017	12	19	0	3	39	0	1	2

here. The demographic data in Tables 3 and 4 include the following variables:

- Pop: County population in 1997.
- Whi: Percentage of whites in 1996.
- Bla: Percentage of blacks in 1996.
- Hisp: Percentage of Hispanics in 1996 (note that the percentages of whites, blacks and Hispanics sometimes add up to more than 100, because Hispanics include other races).
- ≥ 65: Percentage of the population aged 65 and over (actually calculated from the 1996 population aged 65 and over, divided by the 1997 total population).
- HS: Percentage of the population that graduated from high school (1990 census).
- Coll: Percentage of the population that graduated from college (1990 census).
- Inc: Mean personal income (1994).

As an initial analysis of the data, Figure 1 plots the Buchanan percentage vote against 12 covariates. In each case, Palm Beach county is marked with an $\times$. The plots show that the percentage of Buchanan votes is overall decreasing with total population size; not obviously dependent on the percentages of whites and

TABLE 2
County voting data, part II

County	Bush	Gore	Brow	Nade	Har	Hag	Buc	Mc	Ph	Mo
Madison	3,038	3,014	18	54	0	2	29	1	1	5
Manatee	57,952	49,177	242	2491	5	35	271	3	19	26
Marion	55,141	44,665	662	1809	13	26	563	6	22	49
Martin	33,970	26,620	109	1118	14	29	112	7	20	14
Miami–Dade	289,492	328,764	760	5352	87	119	560	35	69	124
Monroe	16,059	16,483	162	1090	1	26	47	0	3	7
Nassau	16,280	6,879	62	253	0	7	90	4	3	3
Okaloosa	52,093	16,948	313	985	4	15	267	2	33	20
Okeechobee	5,057	4,588	21	131	1	4	43	1	3	4
Orange	134,517	140,220	891	3879	13	65	446	7	41	46
Osceola	26,212	28,181	309	732	10	20	145	5	10	33
Palm Beach	152,846	268,945	743	5564	45	143	3407	302	188	103
Pasco	68,582	69,564	413	3393	19	83	570	14	16	77
Pinellas	184,823	200,629	1230	10022	41	442	1013	27	72	170
Polk	90,180	75,193	365	2062	8	59	532	5	46	36
Putnam	13,447	12,102	114	377	2	7	148	3	10	12
Santa Rosa	36,274	12,802	131	724	1	13	311	1	43	19
Sarasota	83,100	72,853	431	4069	11	94	305	5	15	59
Seminole	75,677	59,174	550	1946	6	38	194	5	18	26
St. Johns	39,546	19,502	210	1217	4	11	229	2	12	13
St. Lucie	34,705	41,559	165	1368	4	12	124	10	13	29
Sumter	12,127	9,637	53	306	2	2	114	0	3	17
Suwannee	8,006	4,075	52	180	2	4	108	0	9	5
Taylor	4,056	2,649	4	59	0	3	27	1	8	1
Union	2,332	1,407	15	33	1	0	37	0	1	0
Volusia	82,214	97,063	442	2903	8	36	496	5	20	69
Wakulla	4,512	3,838	30	149	2	3	46	1	0	6
Walton	12,182	5,642	68	265	3	11	120	2	7	18
Washington	4,994	2,798	32	93	0	2	88	0	9	5

blacks but strongly negatively correlated with Hispanics; decreasing with the percentage of population aged 65+; decreasing with mean high school graduation rate; decreasing with mean college graduation rate; and decreasing with mean personal income. There are also clear negative correlations with the proportions of Gore and Nader votes, and a positive correlation with the proportion of Bush votes. On several of these plots, Palm Beach county stands out as an outlier, in the sense that it is inconsistent with the pattern of the rest of the data points.

4. BUILDING A REGRESSION MODEL

4.1 Transformation of the Response Variable

Possibly the main technical difficulty with fitting a regression to these data is the enormous variation in county populations. If we define N_i to be the total number of votes cast and y_i to be the number of votes for Buchanan in county i, then the binomial distribution would suggest that the variance of y_i is approximately $N_i p_i(1 - p_i)$, where p_i is the proportion of voters who support Buchanan in county i. Since the overall value of p_i is about 0.003, we subsequently neglect $1 - p_i$ in this expression and write

$$\text{Var}(y_i) \approx N_i p_i. \tag{1}$$

Also, in view of the small p_i, the binomial distribution is very closely approximated by a Poisson distribution, for which (1) is exact. Thus, any direct attempt to perform least-squares regression analysis runs into the difficulty that the data are heteroskedastic, contradicting one of the standard assumptions of least-squares regression analysis.

In fact, as we shall see later, even (1) is a substantial underestimate of the true variance, because it turns out that there is substantial overdispersion in these data. However, the more fundamental issue is still the enormous variation of variance with N_i.

TABLE 3
County demographic data, part I

County	Pop	Whi	Bla	Hisp	≥ 65	HS	Coll	Inc
Alachua	198,326	74.4	21.8	4.7	9.4	82.7	34.6	19,412
Baker	20,761	82.4	16.8	1.5	7.7	64.1	5.7	14,859
Bay	146,223	84.2	12.4	2.4	11.9	74.7	15.7	17,838
Bradford	24,646	76.1	22.9	2.6	11.8	65.0	8.1	13,681
Brevard	460,977	88.3	9.2	4.1	16.5	82.3	20.4	19,567
Broward	1,470,758	80.3	17.5	10.9	20.3	76.8	18.8	24,706
Calhoun	12,337	81.6	16.9	1.6	14.3	55.9	8.2	12,570
Charlotte	133,681	94.3	4.4	3.4	33.4	75.7	13.4	18,977
Citrus	112,454	96.2	2.8	2.5	30.7	68.6	10.4	16,060
Clay	135,179	91.0	6.0	3.5	7.9	81.2	17.9	18,598
Collier	195,731	93.3	5.7	17.1	21.5	79.0	22.3	30,906
Columbia	52,856	78.3	20.5	1.9	12.3	69.0	11.0	15,349
Desoto	26,259	80.6	18.1	12.1	18.0	54.5	7.6	16,544
Dixie	12,563	89.8	9.5	1.2	14.4	57.7	6.2	12,035
Duval	732,622	69.4	27.5	3.4	10.7	76.9	18.4	20,686
Escambia	282,604	73.3	22.7	2.6	11.7	76.2	18.2	17,661
Flagler	46,128	88.5	9.8	5.9	23.0	78.7	17.3	15,613
Franklin	10,133	84.5	14.5	1.0	17.8	59.5	12.4	15,735
Gadsden	45,441	37.6	61.8	2.9	11.6	59.9	11.2	14,416
Gilchrist	13,367	90.0	9.3	2.1	13.0	63.0	7.4	12,865
Glades	9,698	79.6	13.7	10.1	15.3	57.4	7.1	14,789
Gulf	13,926	73.9	25.2	1.1	13.6	66.4	9.2	15,482
Hamilton	12,521	56.3	43.0	3.6	10.9	58.4	7.0	12,357
Hardee	22,113	93.1	5.9	28.4	13.3	54.8	8.6	16,812
Hendry	31,634	78.2	18.8	26.6	9.9	56.6	10.0	17,823
Hernando	125,537	94.4	4.6	4.0	29.6	70.5	9.7	16,062
Highlands	76,854	87.1	11.6	6.7	32.4	68.2	10.9	17,655
Hillsborough	909,444	82.8	14.9	16.0	12.3	75.6	20.2	20,167
Holmes	18,382	91.7	6.5	1.7	15.5	57.1	7.4	12,790
Indian River	99,215	89.2	9.9	3.9	26.6	76.5	19.1	28,977
Jackson	45,706	69.5	29.6	3.5	14.4	61.6	10.9	15,519
Jefferson	13,232	49.4	50.1	1.3	13.4	64.1	14.7	15,574
Lafayette	6,289	83.0	16.4	5.1	10.7	58.2	5.2	13,663
Lake	196,214	88.2	10.9	3.8	26.3	70.6	12.7	18,269
Lee	387,091	91.1	7.8	5.9	24.4	76.9	16.4	22,053
Leon	215,170	70.4	27.3	3.1	8.4	84.9	37.1	16,705
Levy	32,254	84.4	14.2	2.6	17.6	62.8	8.3	13,745
Liberty	6,703	78.1	20.9	3.1	10.7	56.7	7.3	14,896

One standard way to deal with this issue is to transform the response variable so that we are considering a model of the form

$$h(y_i) = \sum_j x_{ij}\beta_j + \varepsilon_i, \tag{2}$$

where h is the transformation function, $\{x_{ij}, 1 \le i \le n, 1 \le j \le p\}$ are the regressors and $\{\varepsilon_i\}$ are random errors. Indeed, it is well known that the transformation $h(y_i) = \sqrt{y_i}$ is approximately variance-stabilizing for a Poisson random variable, so this suggests a square root transformation for our analysis. On the other hand, a logarithmic transformation, suggested in several of the analyses reviewed in Section 2, is *not* variance-stabilizing. Alternatively, one could embed both transformations in the Box and Cox (1964) transformation family, usually written

$$h(y_i) = C_\lambda \frac{y_i^\lambda - 1}{\lambda}, \tag{3}$$

where C_λ is a scaling constant and λ may be any real number; the case $\lambda \to 0$ corresponds to a logarithmic transformation. The scaling constant C_λ is chosen to

TABLE 4
County demographic data, part II

County	Pop	Whi	Bla	Hisp	≥ 65	HS	Coll	Inc
Madison	17,558	53.9	45.6	1.9	13.8	56.5	9.7	13,002
Manatee	237,159	89.8	9.0	5.8	27.8	75.6	15.5	23,031
Marion	237,308	84.3	14.6	4.0	21.4	69.6	11.5	14,502
Martin	116,087	91.8	6.9	6.2	26.6	79.7	20.3	31,996
Miami–Dade	2,044,600	77.0	21.2	54.4	14.4	65.0	18.8	20,014
Monroe	81,919	92.3	6.2	15.8	15.9	79.7	20.3	25,160
Nassau	54,096	87.3	11.9	1.5	9.8	71.2	21.5	20,874
Okaloosa	167,580	85.3	10.3	4.2	9.1	83.8	21.0	18,959
Okeechobee	33,102	91.1	7.5	14.8	14.6	59.1	9.8	15,162
Orange	783,974	79.1	17.5	12.3	10.4	78.8	21.2	20,469
Osceola	142,128	90.7	6.6	15.3	13.2	73.7	11.2	16,256
Palm Beach	1,018,524	83.9	14.4	9.8	23.7	78.8	22.1	33,518
Pasco	320,253	96.5	2.3	4.4	32.0	66.9	9.1	16,924
Pinellas	871,766	89.1	9.0	3.1	26.6	78.1	18.5	24,796
Polk	448,646	83.3	15.4	5.3	18.2	68.0	12.9	17,824
Putnam	70,430	78.1	20.9	3.4	17.6	64.3	8.3	14,250
Santa Rosa	114,481	92.6	4.6	2.0	8.9	78.5	18.6	17,127
Sarasota	301,644	94.0	5.1	2.8	32.3	81.3	21.9	30,205
Seminole	344,729	87.4	9.8	8.4	10.1	84.6	26.3	21,815
St. Johns	112,707	88.7	10.1	3.0	15.6	79.9	23.6	25,637
St. Lucie	179,559	79.6	19.0	5.2	20.1	71.7	13.1	16,483
Sumter	39,428	81.0	18.1	3.1	20.3	64.3	7.8	14,606
Suwannee	33,077	82.2	16.9	2.0	15.8	63.8	8.2	14,773
Taylor	18,718	77.1	21.5	1.3	12.7	62.1	9.8	15,459
Union	12,359	71.0	27.8	4.8	7.0	67.7	7.9	10,783
Volusia	419,797	88.0	10.5	5.0	22.7	75.4	14.8	17,778
Wakulla	19,172	83.9	14.9	0.9	10.9	71.6	10.9	15,570
Walton	37,914	88.9	8.6	1.2	14.9	66.5	11.9	14,866
Washington	20,221	79.7	17.6	1.5	16.4	60.9	7.4	13,732

satisfy the relationship $\prod_i |h'(y_i)| = 1$, which for (3) leads to

$$C_\lambda = \dot{y}^{1-\lambda}, \tag{4}$$

where $\dot{y}$ is the geometric mean, that is, $(\prod_i y_i)^{1/n}$, where n is the number of observations. This choice of scaling constant makes the Jacobian of the transformation 1, and hence allows different values of λ to be compared in terms of the residual sum of squares of the regression model.

For the initial stages of the analysis, three response variables were used: y_i itself, $\sqrt{y_i}$ and $\log(y_i/N_i)$. According to (3) and (4), these should be multiplied by scaling constants which are, respectively, 1, $2\sqrt{\dot{y}}$ and $\dot{y}$. For this particular data set, $\dot{y}$, the geometric mean of Buchanan votes across all counties, is 127.2651.

Aside from transformation and scaling issues associated with the response variable, one must also consider the scaling issue with respect to the covariates. A number of authors on regression, such as Cook and Weisberg (1982) and Carroll and Ruppert (1988), have pointed out that there is often a need to transform both sides of a regression equation, and this appears to be a case in point. Consider income, for example, a significant covariate in all the models to be discussed. It does not make sense that the influence of income on the *total* Buchanan vote in Miami–Dade county (population 2,044,600) would be the same as that in, say, Lafayette county (population 6,289). Logically, if income affects votes at all, then the level of the effect is on the probability that an individual voter supports Buchanan, and if one wants to know the influence on the y_i, one must multiply by the total number of votes cast in the county. The same argument applies to all the covariates we are considering, so if the response variable is proportional to y_i^λ for some other value of λ, then the covariates should be scaled by N_i^λ. To be logically consistent, this rescaling should also include the intercept, that is, instead of implicitly defining a covariate 1 the coefficient

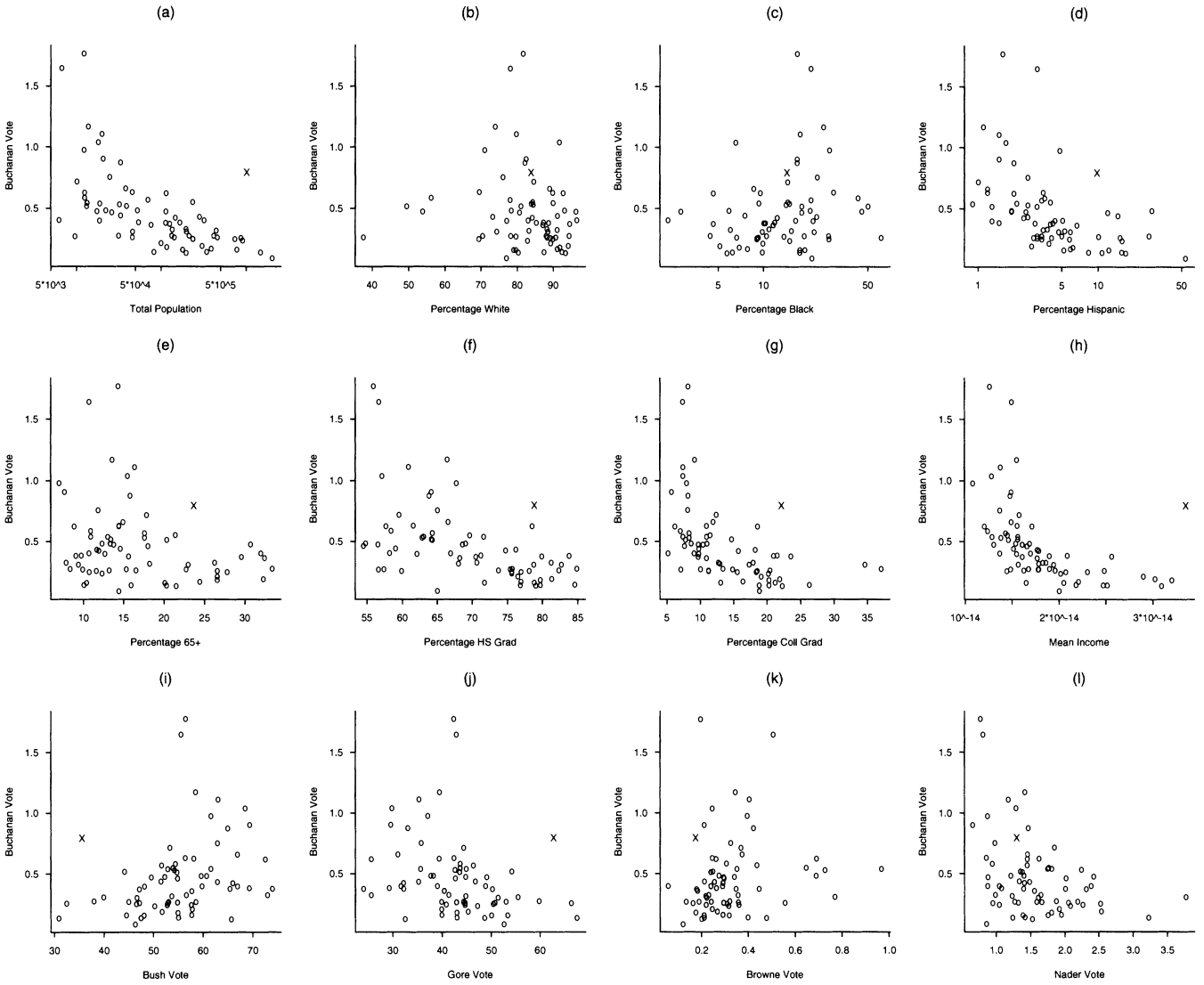

FIG. 1. *Percentage of Buchanan vote against* 12 *covariates. Palm Beach county is marked with an* $\times$.

of which is the intercept of the linear regression, we should replace this by a covariate N_i^λ, and then fit a regression without a separate intercept term. In its handling of the intercept term, the analysis presented in this article differs from an earlier version posted on my website. This rescaling of the covariates has been made in all the analyses presented below.

Specifically, when the response variable is y_i^λ for some $\lambda > 0$, the jth covariate x_{ij} is replaced by $x_{ij}^* = N_i^\lambda x_{ij}$, while the intercept term $x_{i1} \equiv 1$ is replaced by $x_{i1}^* = N_i^\lambda$. Then, ordinary least squares without an intercept is performed on y_i^λ against regressors x_{ij}^*, $j = 1, \ldots, p$. An alternative approach would be to define $(y_i/N_i)^\lambda$ to be the response, make no scaling of covariates and use a weighted linear regression, but the method that has been described is only slightly more complicated to implement and allows us to use the full range of diagnostic techniques associated with standard linear regression.

One minor difficulty with implementing the proposed analyses in SAS is that for many of the analyses, we would like to predict the Buchanan vote in Palm Beach county based on a regression analysis fitted to the other 66 counties. For this purpose, we would like to define the Buchanan vote in Palm Beach to be a missing value, but to get the predictions, the other covariates, including N_i, need to be defined for Palm Beach. To avoid this difficulty, the actual definition of N_i has been the sum of all votes cast *excluding* Buchanan. Given that the overall percentage of Buchanan's vote is so small, this should make no difference to the results. This also avoids a possible ob-

TABLE 5
List of covariates used in the analysis

Covariate	Definition
lpop	Log total population size
whit	Proportion of whites
lblac	Log proportion of blacks
lhisp	Log proportion of Hispanics
o65	Proportion of population aged 65 and over
hsed	Proportion graduated high school
coll	Proportion graduated college
inco	Mean personal income
pbush	Proportion voting for Bush
pbrow	Proportion voting for Browne
pnade	Proportion voting for Nader

jection that if N_i is being treated as one of the predictors for Buchanan's vote, the definition of N_i should not include the quantity we are trying to predict, although again, this should be a minor issue given the small overall percentage of votes for Buchanan.

4.2 Variable Selection

After both the covariates and the response variable were suitably transformed and rescaled, regression analyses were performed to determine which covariates were statistically significant. Noting from Figure 1 that some of the covariates give more meaningful responses when plotted on a logarithmic scale, a logarithmic transformation was applied in those cases. The full list of covariates considered is shown in Table 5. Note that the proportion of Gore votes has not been included in this analysis, but the reason for that is that it is so highly correlated with the proportion of Bush votes (in most counties, the two add up to very nearly 100%) that it makes no sense to consider both in the analysis—we would get very similar results to the following if we deleted the proportion of Bush voters from the analysis and replaced it with the proportion of Gore voters (with opposite signs in the corresponding regression coefficients).

Because of our strong prior suspicion, supported by Figure 1, that Palm Beach county was a very strong outlier, this has been omitted from all analyses used to determine variable selection.

Variable selection may be applied using any of the standard techniques used for multiple regression: the two criteria considered here are Mallows' C_p and backward selection. The variables selected by these two methods are shown in Table 6.

The models selected by backward selection and Mallows' C_p differ somewhat, for the models using y_i and $\log(y_i/N_i)$, but some comparisons of the models (not reported here) suggest that it makes very little difference to the eventual predictions which of the two model selection strategies is employed. However, the differences among the models fitted to the three response variables are far more significant, so we concentrate on that aspect in subsequent discussion. The residual sums of squares (RSS) for the three models selected by C_p are 127,298 (57 df), 72,712 (59 df) and 91,933 (60 df) for response variables y_i, $\sqrt{y_i}$ and $\log(y_i/N_i)$, respectively, after rescaling the response variable as mentioned earlier. This implies a clear preference for the square root transformation among the three models considered, but it also shows that a logarithmic transformation is much better than using y_i directly as a response.

However, after adjusting the residual sums of squares for the rescaling, the residual variance derived from the square root transformation model is 2.42, compared with an approximate variance of 0.25 for $\sqrt{y_i}$ if y_i is Poisson. This implies a very considerable amount of overdispersion compared with the Poisson distribution.

TABLE 6
Covariates selected by either Mallows's C_p or backward selection; all models include the rescaled intercept term

Response variable	Selection method	Variables selected
y_i	C_p	lpop, whit, lhisp, o65, hsed, coll, pbush, pnade
y_i	Backward	lpop, whit, lhisp, o65, hsed, pbush, pnade
$\sqrt{y_i}$	C_p	whit, lhisp, o65, hsed, inco, pbrow
$\sqrt{y_i}$	Backward	whit, lhisp, o65, hsed, inco, pbrow
$\log(y_i/N_i)$	C_p	lpop, lhisp, hsed, inco, pbush
$\log(y_i/N_i)$	Backward	lhisp, hsed, inco, pbush

4.3 Testing for Heteroskedasticity

After tentatively identifying a model, we can now make various tests to decide whether the model is appropriate. Given the strong emphasis we have placed so far on the homoskedasticity assumption, it seems natural to examine this first.

Suppose a model of the form of (2) has been fitted. After constructing estimates $\hat{\beta}_j$ of the parameters β_j, we form residuals

$$(5) \qquad e_i = h(y_i) - \sum_j x_{ij}\hat{\beta}_j, \quad 1 \le i \le n.$$

One way to examine the data for homoskedasticity is to plot the absolute values of the residuals against some other variable of interest—if the variability of the residuals increased or decreased over the range of the plot, that would indicate a problem with the model.

In Figure 2, the square root of absolute residuals from three regressions, (a) based on the original Buchanan votes as the response, (b) based on the square root of Buchanan votes and (c) based on the log proportion of Buchanan votes, is plotted against total vote (on a logarithmic scale) in each county, omitting Palm Beach. Also shown on each plot is a smoothed function through the scatterplot, calculated using the lowess function in S-PLUS. The covariates included in each regression were the ones selected as optimal by the C_p method in Table 6. The square root scale for the y axis was chosen after visually inspecting several plots for the one that gave the best visual representation of the scale variation of the residuals. In plot (a), it is clear that the residuals have a tendency to increase with the total votes in the county. In (b), the scale of the residuals seems approximately constant. In (c), the residuals appear to decrease sharply with increasing total vote over the left-hand half of the plot, then level off and even increase slightly. These three plots support the square root transformation as the one most consistent with an assumption of equal variance of the residuals. Other plots using different variables on the x axis supported the conclusion that the variance of residuals is approximately constant when the response variable in the regression is the square root of Buchanan votes.

These visual impressions may be backed up by formal tests of homoskedasticity. One such test is due to White (1980) and is implemented in SAS through the SPEC option with PROC REG. In this procedure, squared residuals (derived from the ordinary least-squares regression) are regressed on the squares and cross-products of all the regressors, and a test statistic is constructed based on the multiple correlation coefficient (nR^2) of this regression. The results are shown in Table 7. Given our heavy emphasis on the whole homoskedasticity issue so far, it is somewhat disappointing that in none of the five cases considered is the null hypothesis of homoskedasticity rejected at the 5% level of significance. This may be pointing to a lack of power of the test when the number of covariates is large and the number of observations is comparatively small.

Wetherill (1986) described a number of other tests for homoskedasticity. For example, a combination of ideas from Godfrey (1978) and Koenker (1981) led Wetherill to propose the test statistic

$$(6) \qquad \phi = \frac{n\{\sum_i (\hat{y}_i - \bar{y})e_i^2\}^2}{\sum_i (\hat{y}_i - \bar{y})^2 \sum_i (e_i^2 - \hat{\sigma}^2)^2},$$

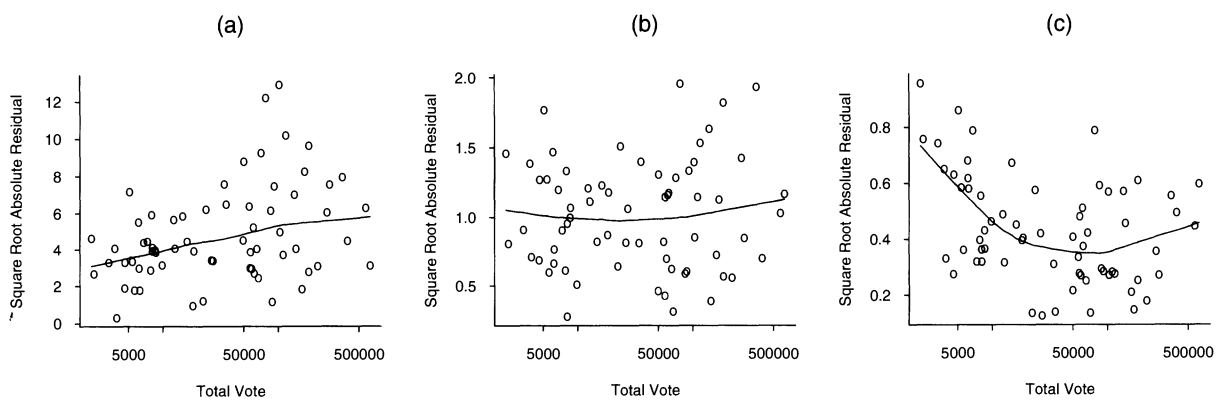

FIG. 2. (a) *Plot of the square root of absolute values of studentized residuals for the model with untransformed Buchanan votes as the response variable against total votes in the county, together with a smoothed curve from the lowess function in S-PLUS. Palm Beach county has been omitted from the plot.* (b) *The same thing based on regression using square root of Buchanan votes as the response.* (c) *The same thing based on regression using log proportion of Buchanan votes as the response.*

TABLE 7
Results of White's heteroskedasticity test applied to the five distinct models of Table 6

Response variable	Selection method	nR^2	DF	p value
y_i	C_p	43.52	45	0.53
y_i	Backward	47.81	36	0.09
$\sqrt{y_i}$	C_p or Backward	23.02	28	0.73
$\log(y_i/N_i)$	C_p	18.20	20	0.57
$\log(y_i/N_i)$	Backward	13.12	14	0.52

where $\hat{y}_i$ is the ith fitted value, $\bar{y}$ is the average of y_i or equivalently $\hat{y}_i$, e_i is the ith residual and $\hat{\sigma}^2 = \sum e_i^2/n$. Under the null hypothesis of homoskedasticity, ϕ has an approximate χ_1^2 distribution.

Another approach is to regress e_i^2 on selected covariates z_{ij}, which may or may not be the same as the covariates x_{ij}. This was the idea of Godfrey (1978) and appears to be particularly appropriate when we suspect the variance depends on a specific covariate or group of covariates (such as total population of the county in the present example). Godfrey defined Z to be a matrix of covariates including a column of 1s to represent an intercept term, and defined r as the vector with entries $r_1, \ldots, r_n$, where $r_i = e_i^2/\hat{\sigma}^2 - 1$. Then define

$$G = \tfrac{1}{2} r^T Z (Z^T Z)^{-1} Z^T r. \tag{7}$$

Godfrey claimed that under the null hypothesis of homoskedasticity, G has an asymptotic χ^2_{p-1} distribution, where p is the number of columns of Z.

For the present study, I have taken just a single variable z_i, and hence defined $G = (\sum r_i z_i)^2/(2\sum z_i^2)$, where $z_i = \log N_i$. (A corresponding analysis based on $z_i = N_i$ did not produce nearly such clear-cut results.) The p values associated with both ϕ and G have been assessed by simulation: for each regression model under study, the analysis was repeated 1,000 times using the same covariates but y_i generated as independent standard normal random variables, and ϕ and G computed for the simulated regression analysis. The quoted p values are the proportions of simulations for which the simulated ϕ or G value exceeded that calculated for the real data. For this analysis, which differs from Godfrey's in the omission of an intercept term in Z, no asymptotic χ^2 result appears applicable for the distribution of G.

For the Florida data, the test statistics ϕ and G are tabulated in Table 8, together with simulated p values. The test based on ϕ rejects all the models, except the one based on $\sqrt{y_i}$, at the 0.1 level of significance, and two of the p values are around 0.01–0.02. The G test is much more definitive, decisively rejecting the null hypothesis for all cases except the model based on square roots.

Thus our conclusion is that, for this data set, White's test apparently accepts the homoskedasticity hypothesis for all the models considered, but the other two tests are more definitive. In particular, the test based on the G statistic (7) leads to the clear-cut conclusion that the square root transformation is the only one for which homoskedasticity is valid.

4.4 Other Diagnostics of the Model Fit

We now consider a number of other diagnostics for the fit of the model. The studentized residuals are

$$d_i^* = \frac{h(y_i) - \sum_j x_{ij}\hat{\beta}_{j(i)}}{s_{(i)}}, \tag{8}$$

where $\hat{\beta}_{j(i)}$ denotes the estimate of β_j and $s_{(i)}$ denotes the estimate of residual standard deviation σ, based on all observations omitting the ith. It is a standard fact, used later, that if all the usual model assumptions of linear regression are satisfied, the distribution of d_i^* is exactly t_{n-p-1}.

There are also a number of measures used to define influence, of which one of the most popular is DFFITS

TABLE 8
Test statistics ϕ and G, with corresponding p values, applied to the five distinct models of Table 6

Response variable	Selection method	ϕ	p value	G	p value
y_i	C_p	2.95	0.04	0.269	0.00
y_i	Backward	4.49	0.01	0.328	0.00
$\sqrt{y_i}$	C_p or Backward	1.70	0.18	0.031	0.26
$\log(y_i/N_i)$	C_p	5.42	0.02	0.277	0.00
$\log(y_i/N_i)$	Backward	3.18	0.08	0.233	0.00

(see, e.g., Belsley, Kuh and Welsch, 1980; Cook and Weisberg, 1982; Atkinson, 1985; Neter, Kutner, Nachtsheim and Wasserman, 1996).

Atkinson (1985) proposed a plotting technique in which the absolute values of either the studentized residual or the DFFITS statistic are first arranged in increasing order and then plotted as a half-normal plot; that is, if there are n observations, a quantile–quantile plot is constructed as if they were the largest n observations from $2n + 1$ normal data values. As a measure of how much such a plot differs from what would be expected under normal-theory assumptions, Atkinson also proposed constructing confidence bands by simulation. The entire regression model fit, up to the calculation of Studentized residuals or DFFITS, is repeated but with response variables generated at random from a normal distribution. In this way, probability bands for the order statistics of Studentized residuals or DFFITS are constructed (separately for each order statistic). In Atkinson's own examples he used just 19 simulations to construct approximate 95% confidence bands, but for the experiments shown here, 1,000 simulations have been used and the 50th largest and smallest values (marked by dashes) correspond to approximate 5% tail probabilities in each tail.

Because our objective now is to examine how much Palm Beach county really is either an outlier or an influential observation, the regression model has been refitted including Palm Beach, for the model with $\sqrt{y_i}$ as the response and whit, lhisp, o65, hsed, inco and pbrow as regressors (the C_p-best model of Table 6). Ordered values of the studentized residuals and DFFITS are plotted in Figure 3, with simulated confidence bands as just described. In each case there is an enormous outlier, which corresponds to Palm Beach. The studentized residual associated with Palm Beach is 17.5, with a nominal t_{58} distribution and p value of 10^{-72} based on standard normality assumptions. This is the first clear evidence in this paper that Palm Beach is indeed extremely inconsistent with the rest of the data.

This exercise has been repeated using $\log(y_i/N_i)$ as the response variable, with results very similar to Figure 3. In this case, the studentized residual associated with Palm Beach is 6.36, with a nominal t_{59} distribution and a p value of 10^{-10}. Although the quoted p values should not be taken too literally (e.g., even in the absence of any anomaly associated with Palm Beach, it is unlikely that the errors in this kind of regression analysis are exactly normally distributed), they do serve to highlight the extreme nature of the Palm Beach outlier.

The question naturally arises whether Palm Beach is the only outlier. In Figure 3, it can be seen that some other large values of the Studentized residual or

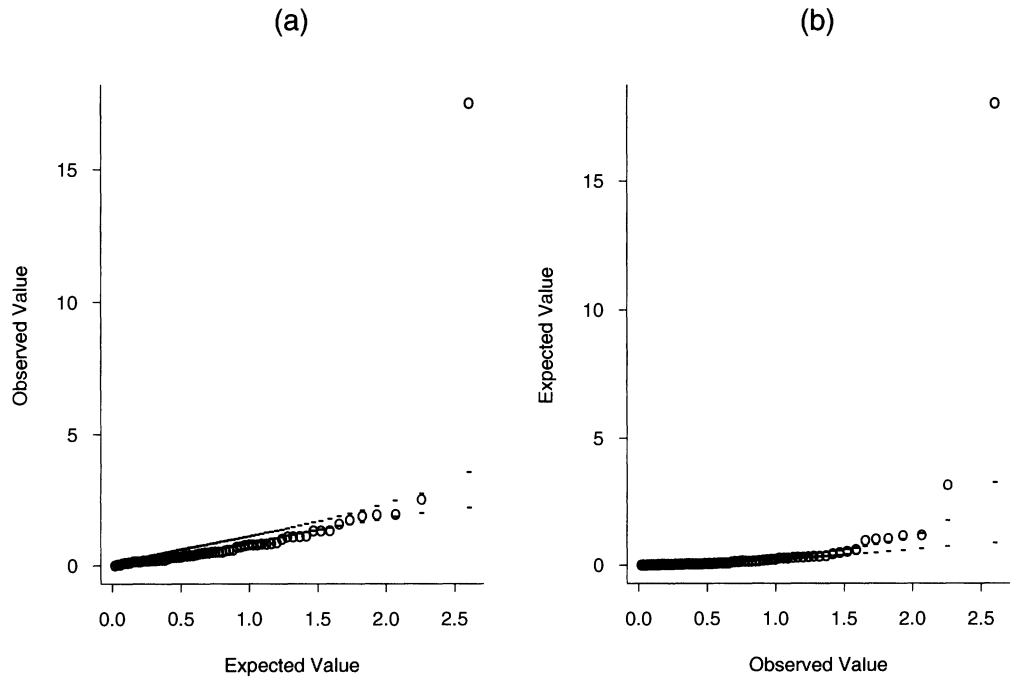

FIG. 3. (a) *Half-normal plot of ordered studentized residuals for the model based on square root Buchanan votes, with pointwise* 90% *simulation bounds.* (b) *Half-normal plot of ordered DFFITS for the model based on square root Buchanan votes, with pointwise* 90% *simulation bounds. Palm Beach county is the large outlier on both plots.*

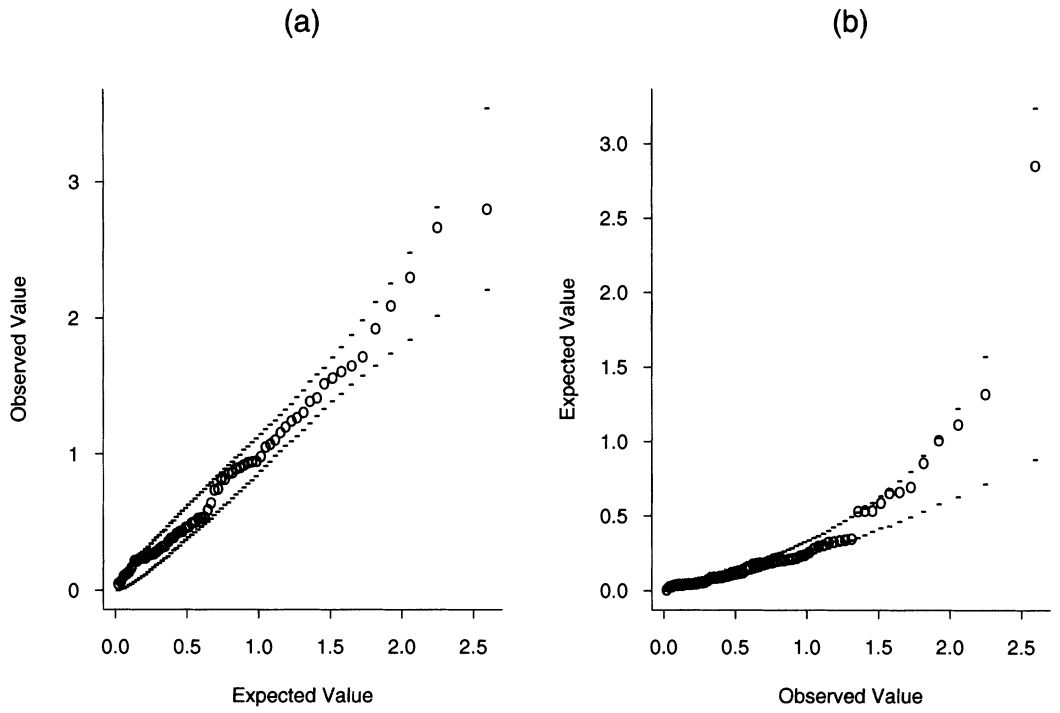

FIG. 4. *Same as Figure 3, but omitting Palm Beach altogether.*

DFFITS are outside the simulation bands, but it is not clear whether these represent other outliers, or whether this is simply part of the distorting effect of Palm Beach on the rest of the analysis. This question has been examined by repeating the whole analysis, including the calculations of Studentized residuals and DFFITS and the corresponding simulation bounds, for the data set in which Palm Beach has been deleted entirely. For the model with the square root of Buchanan vote as the response, Figure 4 shows the resulting plots. There is no sign that any other county, besides Palm Beach, is an outlier, but from Figure 4(b), it still appears that a number of the DFFITS statistics are close to the upper simulation envelope and therefore may represent influential observations.

One possible explanation for this is that, even in the absence of outliers, large cities are still influential. As an indicator of that, DFFITS has been plotted again in Figure 5, against N_i. From Figure 5(a), it does indeed appear that large counties are influential in the analysis using the square root transformation. However in Figure 5(b), which is the same plot using log proportion of Buchanan votes as the response variable in the regression, it appears to be the other way around, that is, in this case, small cities are more influential than large ones (except Miami–Dade).

4.5 Reexamination of the Transformation

So far, we have seen that both the square root and logarithmic transformations appear to work reasonably, but there is no clear preference between the two, and the theoretical argument in favor of a square root transformation is unclear because of the large overdispersion. Therefore, we are motivated to look further at the question of transformations.

One motivation behind the Box and Cox (1964) transformation (3) was the possibility of searching through different values of λ to obtain the best regression model. As noted earlier, if "best" is defined in terms of minimum residual sum of squares, then it is essential to adopt the scaling (4). In Figure 6, this has been done, plotting the RSS against λ, for $0 < \lambda < 1$, for the scaled transformation. The model used for this calculation included the same covariates as for the $\sqrt{y_i}$ regression earlier, although very similar results are obtained with other selections of covariates. For this calculation, Palm Beach has been omitted from the analysis—we get a very different curve from Figure 6 if we include Palm Beach, but we have already provided very strong evidence that Palm Beach is an outlier, so it is appropriate to omit it from this part of the analysis.

Figure 6 shows that the smallest RSS is obtained very near $\lambda = 0.4$, but the actual value of RSS at this

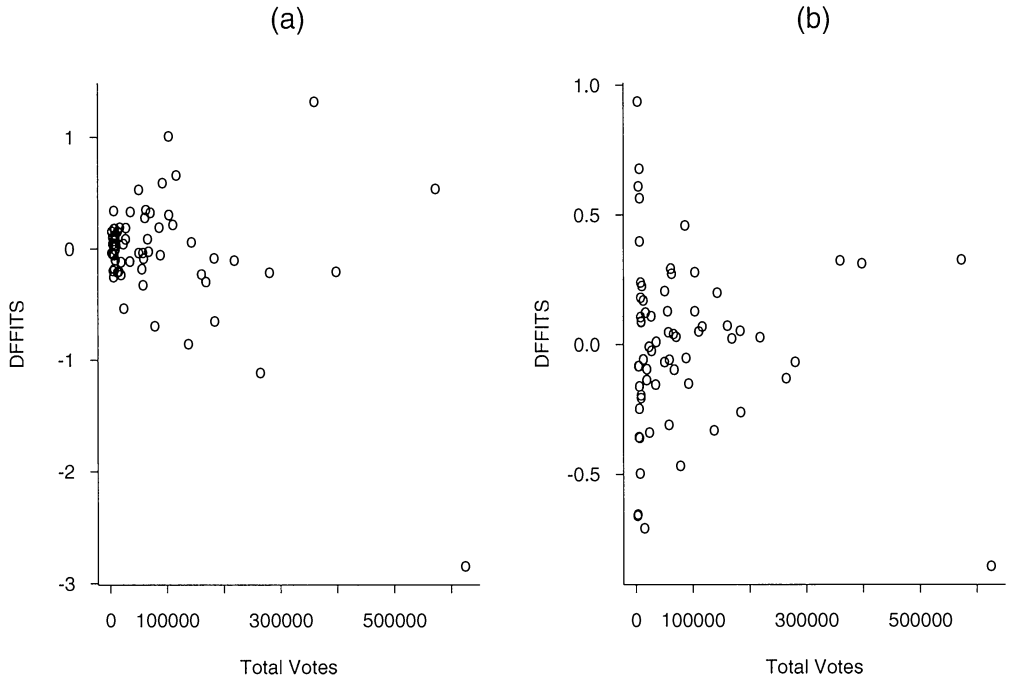

FIG. 5. *Influence diagnostics: Plotting the DFFITS statistic against total votes in the county.* (a) *Square root of Buchanan votes as the response variable.* (b) *Log percentage of Buchanan votes as the response variable.*

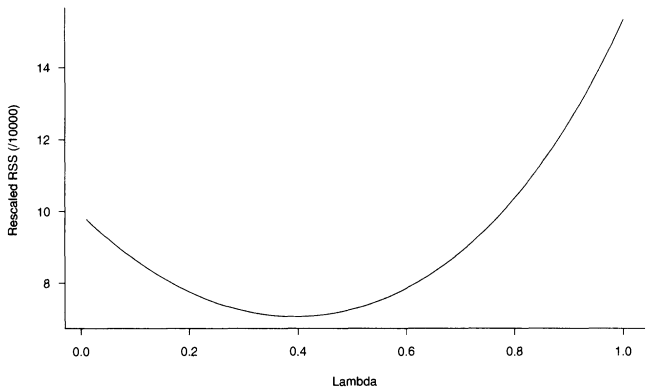

FIG. 6. *Selecting the transformation: Plotting the residual sum of squares against transformation parameter λ, for the normalized transformation.*

point (70,711) is not much reduced from the value at $\lambda = 0.5$ (72,712). As a test of this, we can form an F statistic,

$$F = \frac{72{,}712 - 70{,}711}{1} \cdot \frac{58}{70{,}711} = 1.64,$$

which p value (approximately 0.2) is not significant. This is not an exact F test because λ is not one of the linear regression parameters, but we are treating it as such for the purpose of the test.

Thus our conclusion is that although we could pursue other analyses based on different values of λ, it appears that $\lambda = 0.5$ is adequate.

4.6 Testing Normality

Another issue about whether any of these models fits the data is whether the normality assumption of errors is adequate.

Figure 3(a), in which a probability plot was drawn for the studentized residuals, from a regression analysis in which $\sqrt{y_i}$ was the response, provides clear evidence that the normality assumption is not appropriate for the data set that includes Palm Beach, because of the Palm Beach outlier itself. On the other hand Figure 4(a), which shows the same plot in which Palm Beach has been deleted from the analysis, appears consistent with the normality assumption. This still raises the question of a formal test, however.

As a further check on this, four test statistics of normality were computed: the statistic of Looney and Gulledge (1985), which is similar to the Shapiro–Wilk test, and three familiar goodness of fit statistics, the Kolmogorov–Smirnov statistic, the Cramér–von Mises statistic and the Anderson–Darling statistic (see, e.g., D'Agostino and Stephens, 1986). Critical values for all four statistics may be obtained by simulation—for this analysis, I preferred simulation to asymptotic or tabulated critical values because the simulations correctly

allow for the estimation of the regression parameters, which standard asymptotics or tables do not. The results reported here are based on the Studentized residuals, though similar results were obtained using the ordinary residuals without Studentization.

Using 1,000 simulations, the empirical p values for the four tests based on Figure 3(a) are 0, 0.031, 0.008 and 0, respectively—decisive rejection of the normality hypothesis in all four cases. However, based on Figure 4(a), the empirical p values are all greater than 0.3, indicating no problem with a normal distribution.

The corresponding results based on a $\log(y_i/N_i)$ response are essentially the same: rejection of the normal distribution hypothesis if the outlier is included and acceptance if it is excluded. However, if the same analyses are repeated using y_i itself as a response, the result is rejection of the normality hypothesis even if the outlier is excluded.

Acceptance of a null hypothesis does not prove that the hypothesis is correct, but these results strengthen the case that while regression based on y_i itself is not acceptable, either of the models based on $\sqrt{y_i}$ or $\log(y_i/N_i)$ gives an acceptable fit to a normal distribution, provided Palm Beach is excluded from the analysis. The earlier tests of homoskedastacity still lead us to favor the square root transformation overall.

5. PREDICTING THE PALM BEACH VOTE

We now return to the original purpose of the analysis: predicting the Palm Beach vote from a regression model in which Palm Beach itself has been omitted from the fit. In accordance with the previous results, we consider both $\sqrt{y_i}$ and $\log(y_i/N_i)$ as response variables, with corresponding model selection using either the C_p criterion or backward selection, as in Table 6.

Based on the model fitted to the other 66 counties, point estimates and 95% prediction intervals were calculated for the values of the covariates corresponding to Palm Beach county. The intervals calculated are prediction intervals rather than confidence intervals, that is, they are intended to reflect the result of a single vote rather than a long-run average over a series of votes. Predictions are calculated on the transformed scale, then transformed back to the original scale for the results reported here. It will be recalled that the actual Buchanan votes in Palm Beach county were 3,407.

Table 9 shows point predictions and prediction intervals for three versions of the model. The point predictions are remarkably close to one another, and even the interval estimates are sufficiently consistent to reinforce the point that Buchanan's actual vote was way out of line with predictions based on any reasonable regression model. In all three cases, the point prediction is under 400, and the widest bounds for the prediction interval are 180 at the lower end and 758 at the upper end.

TABLE 9
Point predictions and prediction intervals under three versions of linear model

Response variable	Variable selection	Point predictor	Prediction interval
$\sqrt{y_i}$	C_p or Backward	371	(219, 534)
$\log(y_i/N_i)$	C_p	363	(180, 735)
$\log(y_i/N_i)$	Backward	371	(182, 758)

6. BINARY DATA ANALYSIS

An alternative approach to the whole analysis is to take into account from the beginning that the data are counts and to use the binomial or Poisson distribution in conjunction with a logistic-linear or log-linear model to account for covariates. As already noted, however, this analysis is deficient in the present context because of a rather drastic overdispersion—Equation (1) appears to underestimate the true variance by a factor on the order of 10.

Suppose y_i is the ith count based on population N_i. As in the rest of this paper, y_i is the Buchanan vote in county i, N_i is the total vote excluding Buchanan in county i and we exclude Palm Beach for the purpose of model fitting and prediction. McCullagh and Nelder (1989, pages 124–128) described quasilikelihood approaches to overdispersion based on models of the form $\mathrm{Var}(y_i) = \phi N_i p_i(1 - p_i)$, where ϕ is estimated from either the Pearson or deviance statistics. An alternative approach due to Williams (1982) assumes that an individual voter votes for Buchanan with random probability P, where P is independent from voter to voter, and in county i has mean p_i and variance $\psi p_i(1 - p_i)$. As a result, y_i has mean $N_i p_i$ and variance $N_i p_i(1 - p_i)\{1 + (N_i - 1)\psi\}$. The case $\psi = 0$ corresponds to the case with no overdispersion. These methods are implemented in SAS as part of PROC LOGISTIC, using options SCALE=DEVIANCE, SCALE=PEARSON or SCALE=WILLIAMS.

For p_i, we assume as in standard logistic regression that $\log\{p_i/(1 - p_i)\} = \sum x_{ij}\beta_j$, where the x_{ij} are covariates, the same covariates as in Table 1. When this model is fitted using standard logistic regression

and backward selection of covariates, the only variable dropped from the regression is lblac. Using the resulting model to predict p_{50} (the value of p_i for $i = 50$, which is Palm Beach) leads to a point estimate 0.000884 and a 95% confidence interval (0.000814, 0.000960). With $N_{50} = 428{,}879$, this leads to a point estimate 379 and 95% confidence interval (349, 412) for the mean vote $N_{50}p_{50}$. Compared with the results in Table 9, the point estimate looks reasonable but the interval is too narrow.

The same variable selection (dropping only lblac) is made under the analyses with SCALE=DEVIANCE and SCALE=PEARSON, which also lead to the same point estimate for p_{50}, but the estimates of ϕ are 9.6932 for the deviance analysis and 9.7334 for the Pearson analysis, with both confidence intervals for p_{50} coming out to (0.000683, 0.001143). These lead to the interval (293, 491) for $N_{50}p_{50}$.

The results using SCALE=WILLIAMS are a little different: backward selection removes all variables except lhisp, hsed, inco and pbush; the estimated value of ψ is 0.000452. Estimating p_{50} from the other 66 counties, we get a point estimate 0.000804 with 95% confidence interval (0.000492, 0.001311). The resulting interval estimate of $N_{50}p_{50}$ is (211, 562) with a point prediction of 345. However, the variance formula for the Williams analysis seems less realistic, since it implies much greater overdispersion, compared with the binomial case, in large counties as compared with small counties. This would not be consistent with our earlier conclusions in which we showed that a squarE root transformation appears to lead to homoskedastic errors.

The intervals that have been quoted in this section are properly described as confidence intervals rather than prediction intervals because they do not take into account the variability of y_{50} given p_{50}. A precise resolution of this issue seems hard to achieve, since none of the quasilikelihood models specifies the full distribution of y_{50}. As an ad hoc solution to this problem, the following quasi-Bayesian argument has been adopted. For each of the logistic regression models, the confidence interval for $\log p_{50}$ (but not for p_{50} itself) is symmetric about the point estimate. Therefore, we assume $\log p_{50}$ has a normal posterior distribution, given the data in all counties other than Palm Beach, with a mean and standard deviation chosen to be consistent with the stated confidence interval. Conditionally on p_{50}, we assume y_{50} has a normal distribution with mean $N_{50}p_{50}$ and variance $\phi N_{50}p_{50}(1 - p_{50})$. The two sources of randomness are combined using a simple simulation to obtain a quasi-Bayesian 95% prediction interval for y_{50}.

TABLE 10
Point predictions, confidence and prediction intervals under four versions of logistic regression

Method	Point estimate	Confidence interval	Prediction interval
No overdispersion	379	(349, 412)	(330, 447)
Deviance	379	(293, 491)	(237, 606)
Pearson	379	(293, 491)	(237, 606)
Williams	345	(211, 562)	NA

This procedure leads to approximate prediction intervals for y_{50}, given the data in the other 66 counties, that are given in Table 10. For comparison, the confidence intervals for $N_{50}p_{50}$ are also summarized in the same table. The model with no overdispersion seems clearly wrong and leads to unrealistically narrow intervals. The deviance and Pearson results lead to identical estimates that are consistent with those of Table 9. Prediction intervals are not given for the Williams approach because this appears to lead to an unrealistically large variance for y_{50} and in fact led to a number of negative predictions of y_{50}.

In conclusion, the binary data analyses seem competitive with, but not superior to, the normal regression analyses based on a square root response. It is clearly necessary to account for overdispersion, and both the deviance and Pearson methods for estimating the overdispersion parameter lead to realistic results, whereas the Williams method probably overestimates the variance in those y_i for which N_i is large. On the other hand, the binary data analyses have not been subjected to such intensive diagnostic tests as those based on normal regression, and the approach that has been taken to the calculation of prediction intervals is ad hoc.

7. SUMMARY AND CONCLUSIONS

Previous analyses of the Florida election data have used either y_i itself or $\log(y_i/N_i)$ as a response variable, with several authors claiming that the latter is superior and some claiming that there is no evidence of a "Palm Beach effect" in this case. The present analysis confirms that the analysis based on y_i as a response fails on two counts: (a) lack of homoskedasticity and (b) failure to fit the normal distribution. A model with $\log(y_i/N_i)$ as a response is clearly superior to that based on y_i directly, but it still fails the homoskedasticity test, whereas the analysis based on $\sqrt{y_i}$ seems

satisfactory from all points of view. Binary data analyses are also satisfactory, provided one allows appropriately for overdispersion, but in many respects seem less clear-cut than the normal-theory regression analyses, for example, in the treatment of prediction intervals. In all cases, however, the outlier and influence diagnostics confirm that Palm Beach is a very significant outlier. Point predictions for Buchanan's vote in Palm Beach, based on the other 66 counties, are all under 400 in the analyses presented here, with upper bounds to a 95% prediction interval of under 800 even for analyses based on log proportions, for which some commentators have claimed Palm Beach was not an outlier.

The present analysis establishes conclusively that Buchanan's Palm Beach vote cannot be explained away as normal statistical variation and the likely distortion of the Florida vote was at least 2,500 votes. This article has not attempted to establish that the anomaly was specifically due to the butterfly ballot or was at the expense of Gore's votes rather than Bush's, but other authors, notably Wand et al. (2001), have argued these points persuasively. As far as the broader implications are concerned, from a legal point of view, the issues are probably already dead: Florida law has been changed to require electronic rather than punched card voting in future elections, and it seems safe to say that no future election in the Western world will use anything resembling the butterfly ballot that caused so much confusion in Florida. From a political or historical perspective, however, there remains much interest in the question, "Who really won the election?" Although there were many other contentious issues in this election—the failure of machines to record votes correctly, the alleged harrassment of racial minority voters, the disputed legality of some of the absentee ballots that were counted and the whole question of recounting the ballots by hand—most commentators agree that there is no proof that any of these factors were responsible for Bush winning the election. The most convincing evidence that Gore should have won the election is based on the outcome in Palm Beach county, and the present analysis may be viewed as strengthening the evidence in favor of that conclusion.

ACKNOWLEDGMENTS

I thank the Editor and the referees for numerous suggestions about an earlier version of the article and Alan Agresti, who suggested the binary data analysis of Section 6. The work was partially supported by NSF grants DMS-00-84375 and DMS-99-71980.

REFERENCES

ADAMS, G. (2001). Voting irregularities in Palm Beach, Florida. *Chance* **14**(1) 22–24.

AGRESTI, A. and PRESNELL, B. (2001). Statistical issues in the 2000 U.S. presidential election in Florida. *Journal of Law and Public Policy* **13** 117–133.

ATKINSON, A. C. (1985). *Plots, Transformations and Regression.* Oxford Univ. Press.

BELSLEY, D., KUH, E. and WELSCH, R. E. (1980). *Regression Diagnostics*. Wiley, New York.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. Ser. B* **26** 211–252.

CARROLL, C. D. (2000). How many Buchanan votes in Palm Beach County were erroneous? Available at http://www.econ.jhu.edu/people/ccarroll/carroll.html.

CARROLL, R. J. and RUPPERT, D. (1988). *Transformation and Weighting in Regression*. Chapman and Hall, London.

COOK, R. D. and WEISBERG, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, London.

D'AGOSTINO, R. B. and STEPHENS, M. A., eds. (1986). *Goodness-Of-Fit Techniques*. Dekker, New York.

FARROW, S. (2000). Untitled note in response to Adams (2001). Available at http://www.andrew.cmu.edu/user/sf08/Adams.pdf.

GODFREY, L. G. (1978). Testing for multiplicative heteroscedasticity. *J. Econometrics* **8** 227–236.

HANSEN, B. E. (2000). Who won Florida? Are the Palm Beach votes irregular? Available at http://www.ssc.wisc.edu/~bhansen/vote/vote.html.

KOENKER, R. (1981). A note on Studentizing a test for heteroscedasticity. *J. Econometrics* **17** 107–112.

LOONEY, S. W. and GULLEDGE, T. R. (1985). Use of the correlation coefficient with normal probability plots. *Amer. Statist.* **39** 75–79.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

MONROE, B. L. (2000). Did votes intended for Gore go to Buchanan? Available at http://www.indiana.edu/~playpol/pbmodel.pdf.

NETER, J., KUTNER, M. H., NACHTSHEIM, C. J. and WASSERMAN, W. (1996). *Applied Linear Statistical Models*, 4th ed. Irwin, Chicago.

WAND, J. N., SHOTTS, K. W., SEKHON, J. S., MEBANE, Jr., W. R., HERRON, M. C. and BRADY, H. E. (2001). The butterfly did it: The aberrant vote for Buchanan in Palm Beach County, Florida. *American Political Science Review* **95** 793–810.

WETHERILL, G. B. (1986). *Regression Analysis with Applications.* Chapman and Hall, London.

WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** 817–838.

WILLIAMS, D. A. (1982). Extra-binomial variation in logistic linear models. *Appl. Statist.* **31** 144–148.