

## POL 681

Spring 2004

### Problem Set 2: From Output to Interpretation ANSWER KEY

1. Regression of Buchanan Vote on Bush Vote

```
. reg Buchanan Bush
```

Source	SS	df	MS	Number of obs =	67
Model	5218965.12	1	5218965.12	F( 1, 65) =	41.67
Residual	8141533.66	65	125254.364	Prob > F =	0.0000
				R-squared =	0.3906
				Adj R-squared =	0.3813
				Root MSE =	353.91
Total	13360498.8	66	202431.80		

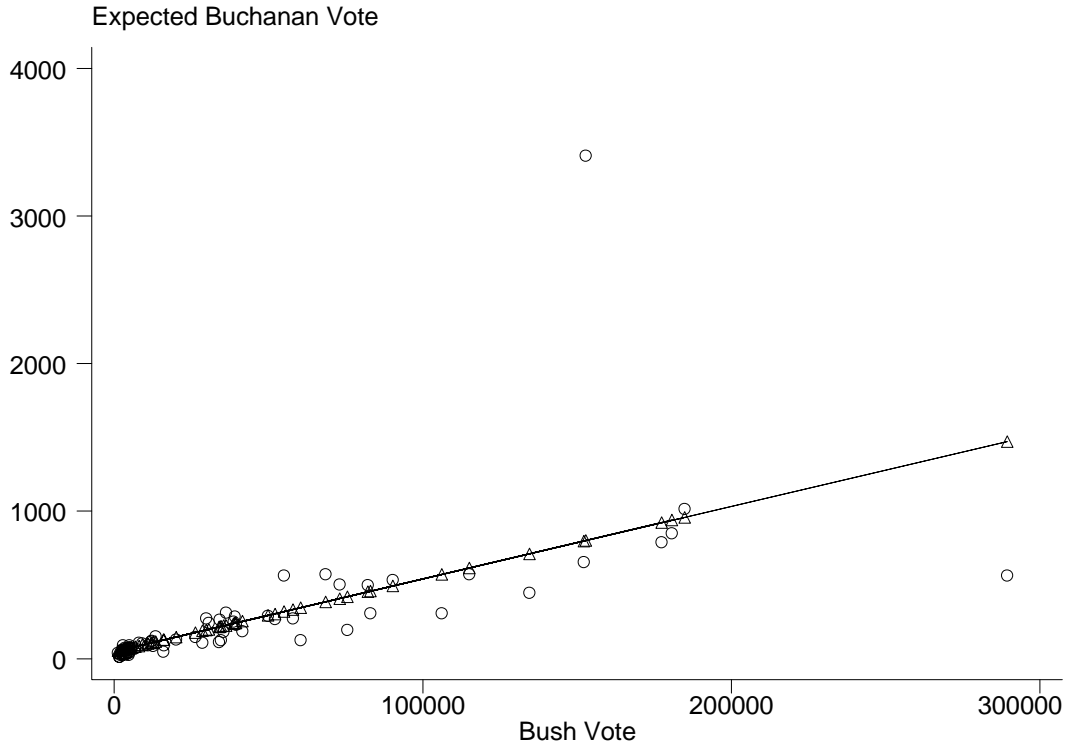
  

Buchanan	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Bush	.0049201	.0007622	6.45	0.000	.0033978 .0064423
_cons	46.97282	54.45616	0.86	0.392	-61.78365 155.7293

Substantively, we see a positive relationship between the Buchanan vote and the Bush vote. Specifically, for each 1 vote increase in the Bush vote, the expected Buchanan vote increases by about .005 votes. Thus, for a 1,000 vote increase in the Bush vote, the expected Buchanan vote increases by about 5 votes. The constant term only makes substantive sense when the Bush vote is 0. Under this condition (which does not occur in the observed data), the expected Buchanan vote would be about 47 votes.

2. The standard error of the regression estimate is 353.91. This indicates that the average residual (which is scaled in terms of the Buchanan vote) is about 354 votes. The standard error of the estimate is the RMSE, which is the square root of the MSE. In turn, the  $MSE = (SS_{Residual} / (n - k - 1))$ . A large RMSE will be indicative of a poorly fitting model. The poor fit could be due to a lack of a relationship between x and y or because of outlying observations (i.e. those with large residuals).

3. Below is the graph of the expected Buchanan vote by the Bush vote.



The central features of this graph are the following: apart from a couple of counties, the regression function closely fits the observed data. Two counties seem to be outliers (i.e. have large residuals). The regression function is positively sloping, indicating a positive relationship between the Bush vote and the expected Buchanan vote.

4. Here I regress the Buchanan vote on the Bush vote, but omit Palm Beach County.

```
. reg Buchanan Bush if County~="PALM BEACH"
```

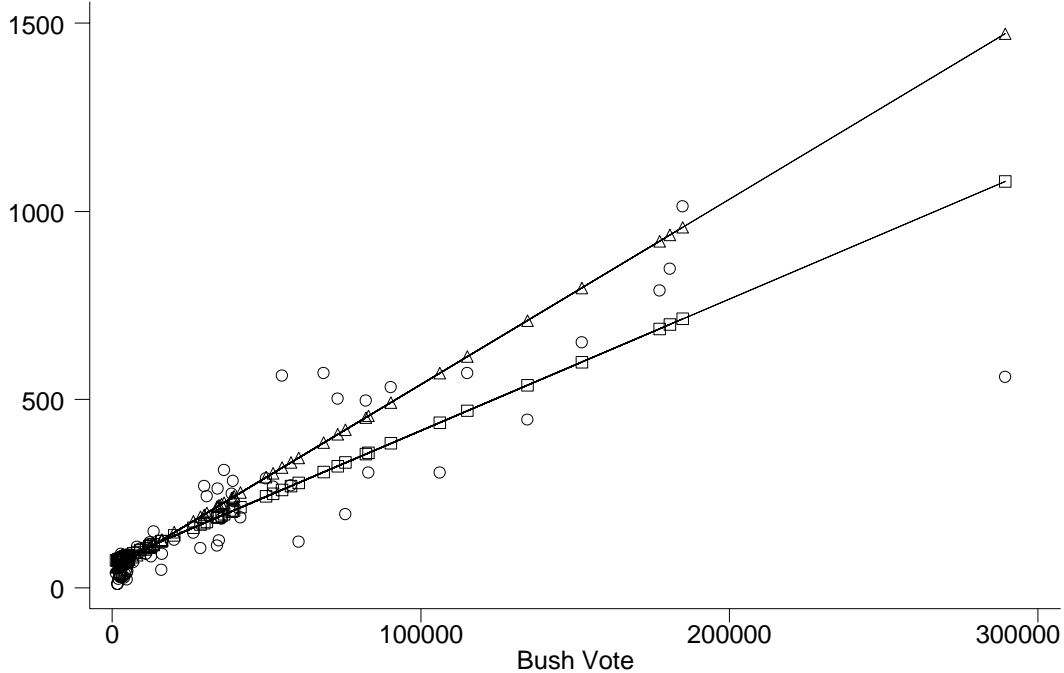
Source	SS	df	MS	Number of obs =	66
Model	2486824.13	1	2486824.13	F( 1, 64) =	193.08
Residual	824301.869	64	12879.7167	Prob > F =	0.0000
Total	3311126.00	65	50940.40	R-squared =	0.7511
				Adj R-squared =	0.7472
				Root MSE =	113.49

Buchanan	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Bush	.0034962	.0002516	13.90	0.000	.0029936 .0039989
_cons	66.94032	17.48248	3.83	0.000	32.01506 101.8656

Substantively, I find the following. There remains a positive relationship between the Buchanan vote and the Bush vote, though after omitting PBC, the coefficient drops in value. Specifically, for each 1 vote increase in the Bush vote, the expected Buchanan vote increases by about .003 votes. Thus, for a 1,000 vote increase in the Bush vote, the expected Buchanan vote increases by about 3 votes. The constant term only makes substantive sense when the Bush vote is 0. Under this condition (which does not occur in the observed data), the expected Buchanan vote would be about 67 votes. Thus, we see that by omitting PBC, the slope "flattens" out and the intercept increases. Compared to the previous model, this relationship is less "steep." We can see this below:

```
. gr Buchanan xb xb2 Bush, ylab xlab c(.11.) b2("Bush Vote") t1("Expected Buchanan Vote") , if Count~="PALM BEACH"
```

### Expected Buchanan Vote



The bottom line represents the model just estimated (note that Palm Beach County is omitted from this figure). Because PBC had, in the original model, the effect of "pulling" the regression line towards its data point, the slope coefficient was heavily "influenced" by the PBC data. Omitting this data point removes the influential observation and the slope "flattens" out a little bit.

5. Noteworthy, the RMSE drops precipitously after omitting PBC from the analysis. This makes sense. Since the RMSE measures the average residual, the large residual due to PBC has a huge impact on this measure. Omitting this one case cuts the RMSE by about 70 percent.

### Multiple Regression

1. Solve the model "by hand." (Could be done in numerous ways. Will evaluate on case-by-case basis.)
2. The regression model is shown below.

```
. reg Buchanan Bush Gore
```

Source	SS	df	MS			
Model	6369332.24	2	3184666.12	Number of obs =	67	
Residual	6991166.53	64	109236.977	F( 2, 64) =	29.15	
				Prob > F =	0.0000	
				R-squared =	0.4767	
				Adj R-squared =	0.4604	
Total	13360498.8	66	202431.80	Root MSE =	330.51	

Buchanan	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
Bush	-.0002427	.0017429	-0.14	0.890	-.0037245	.0032392
Gore	.004313	.0013291	3.25	0.002	.0016579	.0069681
_cons	83.94324	52.11566	1.61	0.112	-20.16977	188.0562

There are several features of the model to note. First, we see that after including the Gore variable, the Bush variable is no longer statistically significant. Its coefficient is no different from 0. On the other hand, we see that the Gore vote is strongly and positively related to the Buchanan vote. On the face of it, this seems counterintuitive. Specifically, we find that for

every 1 vote increase in the Gore vote, the expected Buchanan vote increases by about .004 votes. In contrast, for each 1 vote increase in the Bush vote, the expected Buchanan vote decreases by about .0002 votes. The constant tells us that in the impossible case of Gore and Bush each getting 0 votes, the model predicts Buchanan would receive about 84 votes. The real interpretive problem, however, lies in the fact that the Bush and Gore variables are likely to be highly correlated. In fact, the correlation between the two is about .91. Although multicollinearity is often a tolerable problem for the OLS estimator, in this case, we see that the estimator is having a "harder" time identifying an unambiguous effect of the Bush vote (as well as the Gore vote) on the Buchanan vote. Given the high correlation between the two variables, a substantively natural interpretation is difficult. Additionally, because votes are recorded in as actual votes (not transformed, for example, as logs), then it must be the case that large population counties will have *both* a large Gore vote and a large Bush vote *as well* as a large Buchanan vote. This most likely explains the counterintuitive result that the Gore vote is positively related to the Buchanan vote. Moreover, given the problem in the 2000 election with Palm Beach County, the significant and positive association between the Gore vote and Buchanan vote could be due to this data point.