A Regression Analysis on the Senate Items for Passage from 2004 to 2013

William F. Lamberti

Introduction:

The goal of this report is to find a model to best describe the number of votes an item from the Senate will receive once it has reached the procedural point of passage. It is also desirable to say which variables are the most important in predicting the number of votes. The data was collected from www.GovTrack.us and www.Senate.gov. The sample was collected through Neyman Allocation, a special procedure of a stratified random sample. Neyman Allocation was applied so that any difference in the variation might be kept within each year.

The independent variables initially were the number of Republicans in the Senate, the number of Democrats in the Senate, the number of Independents in the Senate, if the President was Republican or Democrat, if the sponsor was Republican, Democrat, or Independent, the number of African American Senators, and the number of women Senators. The general approach utilized a partial least squares multiple regression fit of the data.

Analysis on the voting behaviors of the Senate is a heavily researched topic. For instance, Riddlesperger and King conducted an analysis on the energy votes in the United States Senate from 1973 to 1980.[1] They concluded that political party and region is the clearest and most important variable for voting patterns. They created three regression models that were applied to the 93rd to the 96th Senate session. For the sake of brevity, model three produced the highest R^2 with values of .610, .691, .577, and .534 for each session respectively.[2] Additionally, studies on the Senate voting on Supreme Court Nominees has also been discussed by Cameron, Cover, and Segal.[3] They found that "Senators routinely vote to confirm nominees who are

[1] Riddlesperger, James W., and James D. King. "Energy Votes in The U.S. Senate, 1973–1980." *The Journal of Politics*, 1982, 838. Accessed December 5, 2014.
[2] Ibid.
[3] Cameron, Charles M., Albert D. Cover, and Jeffrey A. Segal. "Senate Voting on Supreme Court Nominees: A Neoinstitutional Model." *The American Political Science Review*, 1990, 525. Accessed December 5, 2014.

perceived as well qualified and ideologically proximate to Senators' constituents."[4] They performed a spatial analysis and resulted in a model with an R^2 of .746.[5]

The analysis performed within this report differs however from others. It only analyzed those bills and other resolutions that reached the point of a final vote. This means that they were not limited to a specific issue. Additionally, an analysis of the voting behavior of Supreme Court nominees was not included.

Methods:

The general approach entailed a partial least squares multiple regression fit to the sample data. Dependent variable correlation and variance inflation was checked for to ensure that no multicollinearity existed. This led to the removal of the number of Republicans. None of the other variables needed to be removed as the variance inflation factors were all below 10. With the remaining regressors, multiple regression was performed using SAS Studio 3.2. Analysis included following a careful procedure. The data was collected from www.GovTrack.us and www.Senate.gov. The specific sample was made through Neyman Allocation in order to keep the variation within each year. The final number of observation was 43 and edited to check for multicollinearity. Next, a multiple regression first order model was proposed. This model had 8 regressors, but it was expected that this will be refined upon further analysis. Next, parameters were estimated throughout our analysis through partial least squares regression while following the Guass-Markov theorem. Next, the errors were estimated as normally distributed to allow for the approximation of the variance and standard deviations. Following that, the F value and t values were considered for further analysis. Next, the normality of the error distribution was established by residual visualization. This was conducted for each other model analyzed.

[4] Ibid.
[5] Ibid.

Lastly, for the best model, the overall regression model and confidence intervals on parameter estimates were derived using the assumption of normally distributed errors.

The first order model was refined by considering interactions. While all possible interactions were first considered, the number of interactions exceeded the number of observations. Therefore, only those interactions with the number of democrats were considered. Backwards, forwards, and stepwise methods were used to try to find an unbiased and useful model. However, all those models produced were biased. However, retaining all the interactions and first order terms resulted in a model that was unbiased. Unfortunately, the number of regressors and interactions was too high for such a small sample. To avoid over fitting the data, three interactions and one first order term were removed based on the high p values that were clearly not significant. The final model resulted in 6 first order terms and 3 interactions. Analysis of residuals and confidence intervals provided additional support for the final model. Since there were no missing observations, no analysis on missing observations was required.

Results:

Neyman allocation was first applied to determine the size of the sample. With a bound on the error of 5, the total sample size was determined to be 40.50. However, in the process of determining each year's size, the total final sample size chosen was 43. This was done to be conservative in the analysis. The calculation of sample sizes is provided in Table 1. Each years votes were ordered as last bill up for a vote as 1 and the first bill up for vote as the last number. The samples were chosen from each year using a random number generator on a TI-84 Silver Plus Edition calculator. Those selected from the population are listed in Table 2.

Table 1: The results from the Neyman Allocation sample size determination

Year	Calculated Sample Size	Final Sample Size
2013	2.005678863	2
2012	3.160178703	4
2011	6.082803092	6
2010	2.093164643	2
2009	4.493341097	5
2008	2.826762561	3
2007	5.794799482	6
2006	3.69780858	4
2005	6.077152988	6
2004	4.76830999	5
	Final Total	n=43

Table 2: The table includes the observations used in the sample, the determined sample size for each year, and the total number of items within each year

Year	Observation	Final Sample Size	Number of Items
2013	9, 4	2	20
2012	15, 1, 22, 20	4	22
2011	30, 31, 2, 37, 24, 13	6	37
2010	7, 15	2	17
2009	25, 3, 37, 34, 17	5	43
2008	6, 21, 18	3	26
2007	27, 24, 10, 13, 34, 30	6	48
2006	12, 2, 10, 19	4	29
2005	31, 36, 45, 17, 7, 10	6	48
2004	14, 8, 24, 21, 27	5	34

Once the data was collected, the analysis could begin. The variance inflation factor was compared to check for multiculinearity. The first VIF analysis was shown in Table 3. Since 2 variables have VIFs above 190, the highest one, the number of republicans in the Senate, was removed, and the VIF analysis was performed once again. After, all the VIFs were below 10. The results were provided in Table 4. Since all of the variables are less than 10, the rest of the variables can be retained. Therefore, with the remaining regressor variables, the following resulting first order model was

Table 3: Initial VIF Analysis

Variable	Variance Inflation
Intercept	0
num_rep	231.51253
num_dem	209.60860
num_ind	6.38988
pres_r	5.50063
spon_r	8.63316
spon_d	5.28759
senate	1.61582
num_afr_am	2.08595
num_wom	8.49284

Table 4: VIF Analysis after removal

Variable	Variance Inflation
Intercept	0
num_dem	4.27641
num_ind	5.93007
pres_r	5.28798
spon_r	8.61138
spon_d	5.27419
senate	1.56224
num_afr_am	1.89744
num_wom	7.76752

$$Model\ I = \widehat{num\_yea}$$
$$= 112.18 + 0.198(num\_dem) + 13.78(num\_ind) + 9.53(pres\_r) + 8.05(spon\_rep) + 8.47(spon\_dem) + 0.70(senate) - 12.82(num\_afr\_am) - 4.72(num\_wom)$$

However, the F value for this model is 0.88 with a corresponding p value of 0.54. Additionally, none of the t values or p values were initially significant. This model had a R^2 value of 0.1722 and a R_a^2 of -0.0226. While this is concerning and may suggest that a different kind of analysis would be more appropriate, the analysis continued.

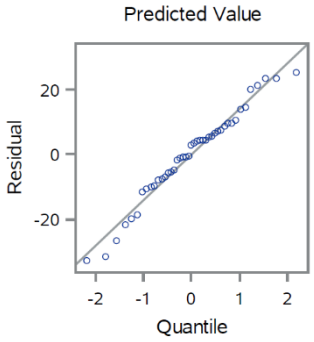

Figure 1: QQ Plot before transformation

The QQ Plot was assessed to see if the residuals have uniform scedasicity. As Figure 1 indicates, this was not as linear as one would desire. From this conclusion, transformations were explored. Eventually, performing a log transformation on the response variable made the QQ Plot more linear as shown in Figure 2. Therefore, the transformation was appropriate. Next, possible interactions were explored. When all possible interactions were considered, many of them were being reported as biased. Therefore, those were removed. Moreover, analyzing the residual plots suggested that the number of democrats had pinching. Therefore, all possible interactions were explored with this variable. Even still, the model continued to remain biased with these interactions. However, it was discovered that by removing the number of Independents, the model became unbiased. Using forward, backward and stepwise techniques with the remaining regressors and interactions, all methods failed to produce adequate models. The forward and backward technique failed to have all the needed first order terms for its interactions in their final models. The backward technique had an interaction

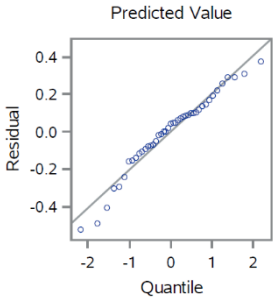

Figure 2: QQ Plot after log transformation

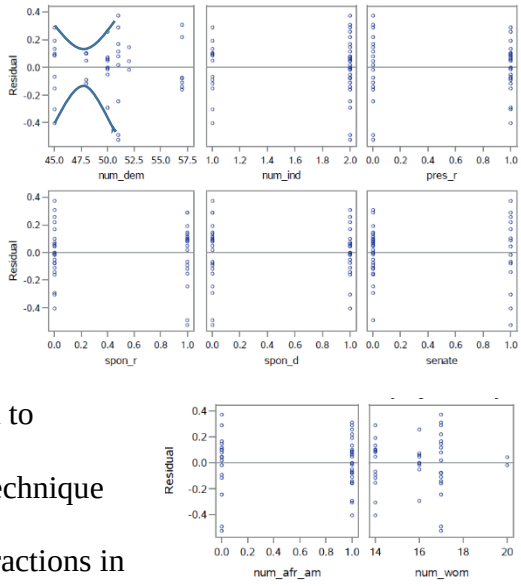

Figure 3: Residual Plots for first order model. Pinching can be seen in first plot for the number of democrats.

term between the number of democrats and the number of African Americans, but it did not include the number of African Americans first order term. The backward technique had an interaction term between the number of democrats and the number of women, but it did not include the number of women first order term. Additionally, models created previous to the final one had Cps not high enough and were biased. The stepwise technique would not accept any variables whatsoever. The threshold p values for the techniques were not changed from what SAS had as the standard threshold value. However, when all interactions and regressors were retained, the Cp was high enough with 14. Thusly, this model with a total of 13 terms was not biased. The resulting model was

$$\begin{aligned} Model\ II = \ & Log(\widehat{num\_yea}) \\ & = -2.00 + 2.10(num\_dem) + 1.64(pres\_r) + 1.90(spon\_r) + 1.49(spon\_d) \\ & - 0.62(senate) + 1.59(num\_afr\_am) + 1.92(num\_wom) - 1.62(d\_presr) \\ & - 1.91(d\_sponr) - 1.47(d\_presd) - 0.58(d\_senate) - 1.66(d\_af) - 1.99(d\_wom) \end{aligned}$$

This model was significant by with an F value of 2.17 and a corresponding p value 0.0405. This model had a R^2 value of 0.4874 and a R_a^2 of 0.2824. A brief description of the model is provided in Table 5. Upon further investigation though, the model had 13 degrees of freedom. Since the data only had 43 observations, it was concluded that some of the interactions and first order variables must be removed to ensure that model was not over fitting to the sample. This was done by comparing p values and one by one, removing those with the highest values. They were in order, the interaction term between the number of Democrats and if the item originated in the Senate, the interaction between the number of Democrats and if the sponsor was a Democrat, the regressor term indicating if the bill originated from the Senate, and lastly, the interaction term between the number of Democrats and the political party of the president. Therefore, 3 interactions were

Table 5: Model summary with interactions with Number of Democrats

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	1.02244	0.07865	2.17	0.0405
Error	29	1.05046	0.03622		
Corrected Total	42	2.07289			

removed as well as one first order term. The first order term could be removed as none of the remaining interactions contained it within the model. The final model chosen had 9 parameters and 3 interactions. The model was

$$Model\ III = Log(\widehat{num\_yea}\ ) = -43.02 + 1.01(num\_dem) + 0.32(pres\_r) + 7.59(spon\_r) + 0.32(spon\_d) + 12.19(num\_afr\_am) + 1.88(num\_wom) - 0.15(d\_sponr) - 0.25(d\_af) - 0.04(d\_wom)$$

This was much better number of interactions and regressors considering the low number of observations. This model still had an F value of 2.82 and a corresponding p value of 0.0143. This model had a R^2 value of 0.4344 and a R_a^2 of 0.2802. Additionally, it has a Cp of 10 and is not biased. It is briefly described in Table 6.

Table 6: Brief description of Model III

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	0.90050	0.10006	2.82	0.0143
Error	33	1.17239	0.03553		
Corrected Total	42	2.07289			

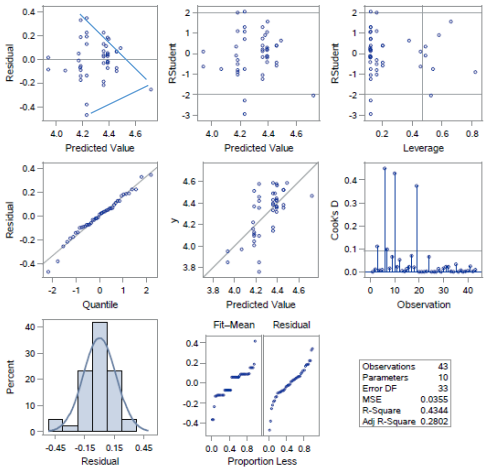

Figure 4: Important plots and graphs for checking GM assumptions and the 4th assumption of the normality of the errors.

A residual analysis was performed on the final model. The distribution of the errors appeared to be normally distributed and centered at 0. There is minimal kurtosis present. There are some observations which are close to 3 standard deviations away, however, this is not of major concern as most of the observations are clearly within 2 standard deviations. The QQ Plot is fairly linear. This satisfies the 3 parts of the Guass-Markov and the 4th assumption of the normal distribution of the errors. However, the predicted value vs the transformed response variables does not perfectly predict the observations. This makes sense as the R^2 is less than .50. This can be seen in Figure 4. However, despite the transformation on the model, the residual

plot indicated that a different transformation might be more appropriate. Perhaps doing a weighted least squares analysis would be a better way to determine this.

The model clearly had some observations pulling the model towards itself as seen in the Cook's D Plot in Figure 5. However, these observations cannot be removed due to the nature of the sample and the matter in which it was collected. Since Neyman Allocation was utilized in creating the sample, each year was made proportional in comparison to the total number of items in each year and the total amount. Some years only had 2 observations allocated because the sample size was only 43 observations. Since so many years had low numbers of observations, removing even 1 observation would destroy the proportions that were so carefully constructed. If the data set was larger, perhaps 200 observations instead of 40, then removing 1 observation would most likely not ruin the proportions. Additionally, Neyman Allocation is should be used as each year the political climate can change drastically. In short, the problem was that the sample size collected was simply too small. Therefore, the final model and its important figures were present and analyzed in the Discussion section.

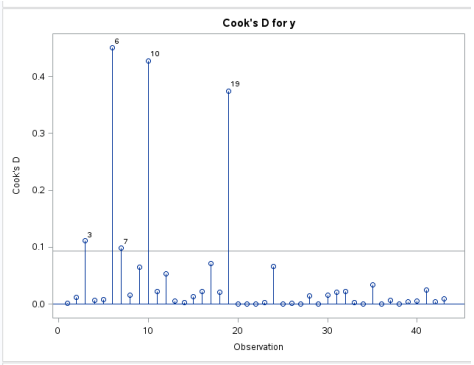

Figure 5: The Cook's D Plot shows that some observation are pulling the model towards themselves strongly.

Discussion:

The data created through Neyman Allocation on the Senate Items for Passage from 2004 to 2013 had been analyzed. The final model, Model III includes 6 regressor variables and 3 interactions. As stated previously, the final model is

$$Model\ III = Log(\widehat{num\_yea}\ ) = -43.02 + 1.01(num\_dem) + 0.32(pres\_r) + 7.59(spon\_r) + 0.32(spon\_d) + 12.19(num\_afr\_am) + 1.88(num\_wom) - 0.15(d\_sponr) - 0.25(d\_af) - 0.04(d\_wom)$$

where num_yea indicates the number of votes an item will receive in the senate.

Model III had captured about 43.4% of the variation within the model as indicated by the R^2 of 0.4344. With a coefficient of variation of about 4.38%, this means that the model will usually lead to more precise predictions. Additionally, since this is lower than the rule of thumb of about 10%, this is a positive sign for the accuracy of the predictions. Furthermore, with a Root MSE of about 0.188, this means that about 95% of the observations will fall within 2(0.188), or .376, of the predicted value. Even though there were no missing observations, quickly checking the 95% confidence interval for a single item's value and the mean value were somewhat narrow. By looking at the confidence intervals for the regressor terms and their interactions, only one contains zero. Therefore, this helps to show that the components of the model are helpful and are meaningful. Additionally, the term that had the biggest impact on the R^2 was the interaction of the number of democrats and the number of women. This was deduced because it had the highest Type I SS Error. These figures have been summarized in Tables 7 and 8.

Table 7: Parameter Estimates

Variable	Parameter Estimate	Standard Error	Type I SS	95% Confidence Limits	
Intercept	-43.01915	12.07022	N/A	-67.57619	-18.46210
num_dem	1.01005	0.25982	0.06631909	0.48144	1.53866
pres_r	0.32251	0.20543	0.01862965	-0.09544	0.74047
spon_r	7.59429	1.85065	0.02833699	3.82911	11.35946
spon_d	0.32420	0.14277	0.01788529	0.03374	0.61466
num_afr_am	12.19327	5.79183	0.12656190	0.40970	23.97684
num_wom	1.88954	0.59394	0.00349618	0.68116	3.09791
d_sponr	-0.15216	0.03761	0.10854338	-0.22868	-0.07563
d_af	-0.25490	0.11899	0.10651376	-0.49698	-0.01281
d_wom	-0.04178	0.01209	0.42421593	-0.06637	-0.01718

Table 8: Important Figures

Root MSE	0.18849
Dependent Mean	4.30276
Coeff Var	4.38058
R-Square	0.4344
Adj R-Sq	0.2802

The results in this analysis had been somewhat comparable to Riddlesperger's and King's study on energy votes in the Senate. They were able to present models with more predictive power. They included the region of the country that senators represent.[6] This is an important variable that was not considered previous to this analysis. This aspect of rationality

[6] Riddlesperger, Ibid.

should be included for other investigations. However, because of the time frame in which they analyzed, they were not able to look into the effects of gender because women simply were not members of the Senate. This is a distinguishing factor in their analysis and the analysis performed in this report. Additionally, they build specific models for each congressional session. They were able to look at every single energy related item as well.[7] Due to the nature of this analysis, this could not have been done. However, future studies might be able to include more observations as to be able to test more interactions and other variables.

The learning process was an insightful because it illustrated many basic principles of basic research. The first is that perseverance is extremely important. For many hours, the analysis was not able to produce a meaningful result. This may be due to in part that a regression analysis may not have been the best method. However, eventually a significant model was produced, but only after continually trying different methods of analysis.

Another important lesson was that a big sample size is vital for models with many variables. By not having enough observations, analysts quickly become limited in the number of interactions and other variables that can be included in the model. Therefore, much larger sample sizes are helpful to ensure that more variables and interactions can be tested for more complicated problems.

[7] Ibid.

Works Cited

Cameron, Charles M., Albert D. Cover, and Jeffrey A. Segal. "Senate Voting on Supreme Court Nominees: A Neoinstitutional Model." *The American Political Science Review*, 1990, 525. Accessed December 5, 2014.

Riddlesperger, James W., and James D. King. "Energy Votes in The U.S. Senate, 1973–1980." *The Journal of Politics*, 1982, 838. Accessed December 5, 2014.