

Assessing the U.S. Senate Vote on the Corporate Average Fuel Economy (CAFE) Standard

Abstract

The dataset presented here illustrates to students the utility of logistic regression. Its analysis results in a fit that explains much of how senators vote on a particular bill, and allows for quantification of the effects of ideology and money on the vote. A number of interesting quantitative interpretations follow from a good fit. A successful analysis makes use of a number of ideas discussed in applied courses: descriptive statistics, inferential methods, transformation of variables, and the handling of outliers and special cases. All these issues arise in the context of data on variables that require of students no specialized knowledge. Students have strong qualitative preconceptions about the relationships among the variables. The final results quantify, and nicely confirm, many of those conceptions.

1. Introduction

Logistic regression is core to intermediate and advanced courses in regression, and appears in syllabi for second semester survey courses in applied statistics. Some recent introductory applied textbooks (for example, Moore, McCabe, Duckworth and Sclove (2003), and Deveaux, Velleman and Bock (2005)) take up the topic. My goals for students include appreciation for model fitting and the descriptive aspects of logistic regression. In what follows I introduce a topical dataset, and the story about its analysis, that helps my students achieve these goals.

I use the CAFE dataset described below as the basis for an in-class activity in a second semester applied statistics course taught to mostly Business, Biology/Zoology, and Information Science majors. I am primarily interested in model fitting and descriptive statements about the fit. (The dataset does not lend itself to sample-based inferences about a population.) I have students study the analysis, and I revisit the exercise on homework and an exam, asking students to explain the decisions that lead to a good fit, and to provide interpretations of that fit.

In an upper-division regression course I have used the dataset in an assignment/take-home exam late in the course. Emphasis here is again on descriptive, rather than inferential, statistics. Students have worked with all the tools necessary to perform a good analysis, yet they have trouble wielding those tools together; consequently I require them to consult with me as they take actions in their analyses.

2. The Dataset

At the end of February 2002 the U.S. Senate considered comprehensive energy legislation. Senators John McCain and John Kerry proposed raising the Corporate Average Fuel Economy (CAFE) standard for cars and trucks. On March 13, 2002 the United States Senate voted on the Levin amendment (No. 2997), charging the National Highway Traffic Safety Administration with the development of a new standard and effectively shelving the McCain/Kerry proposal.

The dataset consists of information about each of the 100 U.S. Senators regarding their vote on the Levin amendment. A senator's vote (Vote) is the response variable. Provided explanatory variables include the state represented, political party affiliation (Party), and the lifetime total amount of contributions received from auto manufacturers (Amount). See Appendix 1 for a detailed description of the contents and format of the data file `cafe.dat`.

Motivation for studying the CAFE issue can be found in the media. An article in the May 2004 Consumer Reports describes how automakers classify cars as light trucks to “bend” the restrictions set by the standard. (Vehicles classified as light trucks fall under relaxed fuel efficiency standards. For instance, by having its Outback model reclassified as a light truck, Subaru was able to add weight to the vehicle without making expenditures compensating for the resulting reduction in fuel efficiency.) Paul Roberts' 2004 book, “The End of Oil,” provides a good summary of the issue in the context of energy concerns – see Appendix 2 for an excerpt. The New York Times (Hakim 2004b) has reported on the Bush administration's plan to propose changes to the national fuel economy regulations. John Kerry's 2004 presidential election campaign vow to pursue better fuel economy involved a stance on CAFE – an article on the Kerry campaign in the New York Times (Hakim 2004a) directly references the vote on the Levin amendment.

3. The Analysis

Prior to the data analysis, I ask students to search for information on the statement of the bill, the outcome of the senate vote, and the votes of their home-state senators. It is important to clarify the statement of the Levin amendment and the meaning of a YES vote. A YES vote is a vote to shelve the McCain/Kerry proposal and is consistent with most auto manufacturers wishes. We spend time in class discussing factors that might affect senators' votes. This leads us to identify Party (as a measure of ideology), and the Amount that auto manufacturers have contributed to the senator, as explanatory variables. Students are asked to make conjectures about the direction of the associations between these variables and the response variable Vote. I provide students with axes (probability of YES vote vs. Amount) and ask them to anticipate the curve for each party.

Students are led to the dataset on the internet and then import it into a spreadsheet. We first obtain univariate summaries for each of the variables. At this point the Jeffords (Independent) case provokes discussion. For the time being we omit this case.

A tally of Vote by Party results in a two-by-two table that can be put to a significance test (one that might have been performed during coverage of categorical data earlier in the term). Our conclusion is that the difference in vote between the two parties did not come about by randomly allocating the 62 YES votes among the 99 senators. We are now primed to include Amount as a second explanatory variable.

3.1 Preliminaries

A graphical and descriptive analysis of Amount by Party is informative and interesting.

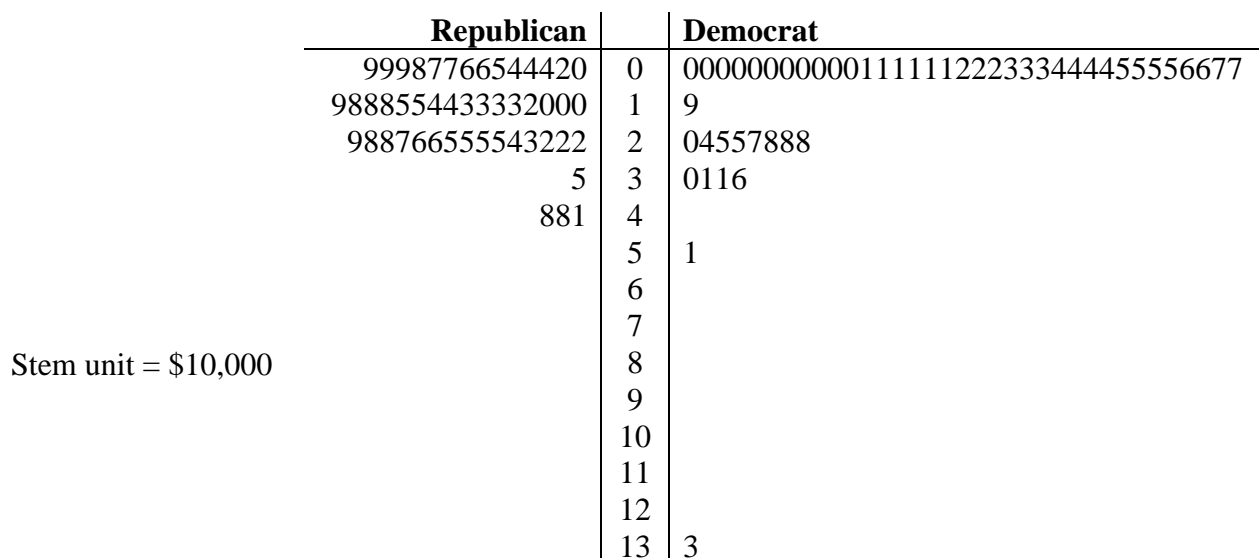


Figure 1. Back-to-back stemplots of contribution amounts for the two parties.

Table 1. Descriptive statistics of contribution amounts by party.

Variable	Party	Count	Mean	Minimum	Q1	Median	Q3	Maximum
Amount	D	50	10025	0	1000	4375	14250	133250
	R	49	17783	250	9250	15000	25000	48939

The distributions have skew, and outlying cases are identified. A transformation of Amount may be in order. I have settled on $\log_{10}(10 \times \text{Amount} + 1)$. (Using $\log_{10}(10 \times \text{Amount})$ expresses “how many ‘figures’ in the dollar amount,” (see “Dealing with Logarithms,” Deveaux, Velleman and Bock (2005), page 45), but leaves cases with Amount = 0 undefined. The proposed compromise allows students to quickly handle the transformation in the spreadsheet and results in essentially the same interpretation. And, to perform the transformation on a calculator, students merely affix the digit 1 to the end of Amount, then press the \log_{10} key.)

Table 2. Descriptive statistics of log-transformed contribution amounts by party.

Variable	Party	Count	Mean	Minimum	Q1	Median	Q3	Maximum
$\log(10A+1)$	D	50	4.124	0.000	4.000	4.641	5.154	6.125
	R	49	5.1318	3.3981	4.9660	5.1761	5.3979	5.6897

See the boxplots in Figure 2 and dotplots in Figure 3 for graphical displays of the transformed amounts. This transformation completely, and meaningfully, alters what constitutes an outlier.

Here again one may ask students to anticipate the fit. Supplied with dotplots of Amount by Party as in Figure 3, have them draw a curve for each party, summarizing the relative density of filled circles as Amount increases.

My students begin the logistic regression analysis (logit link function) fitting Vote by Party. (This amounts to estimating cell probabilities in a 2x2 table. Moore, McCabe, Duckworth and

Sclove (2003) use a 2x2 table as a first step in introducing logistic regression.) We confirm that the fit produces the observed probabilities. From this point forward we engage in an exercise fitting Vote as a function of Amount and Party.

3.2 The model

To fit the data I omit the Jeffords case and use the two predictors: An indicator for Political Affiliation, $PA = I(\text{Party} = \text{Republican})$; and the continuous variable for the approximate number of “figures” in the amount contributed, $F = \log_{10}(10 \times \text{Amount} + 1)$. With $p = \text{Pr}[\text{Vote} = \text{YES}]$, the full model is stated

$$\text{logit}(p) = \beta_0 + \beta_1 PA + \beta_2 F + \beta_3 PA \times F .$$

3.3 Fundamental analyses

Here is a discussion of the fundamental issues in fitting a logistic regression and interpreting the results.

3.3.1 Assessing interaction

A natural first issue centers on the necessity of the interaction term. To assess this, we have four available strategies.

A formal statistical test (the strategy I’d be least inclined to use with my students) reveals the interaction term to be insignificant (from Table 3, $P = 0.122$).

Table 3. Logistic regression results for the full model.

Predictor	Coef	SE Coef	Z	P	Odds	95% CI	
					Ratio	Lower	Upper
Constant	-3.20274	1.64536	-1.95	0.052			
Party							
R	-8.04599	6.46223	-1.25	0.213	0.00	0.00	101.49
log(10A+1)	0.618414	0.350380	1.76	0.078	1.86	0.93	3.69
Party*log(10A+1)							
R	2.01934	1.30583	1.55	0.122	7.53	0.58	97.39

Measures of Association:

(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures		
Concordant	1980	86.3	Somers' D		0.73
Discordant	301	13.1	Goodman-Kruskal	Gamma	0.74
Ties	13	0.6	Kendall's Tau-a		0.35
Total	2294	100.0			

Measures of association are commonly used to compare fits in regression. Summary measures from Table 3 and Table 4 show that the full model improves very little on the reduced (no interaction) model.

Table 4. Logistic regression results for the reduced model.

Predictor	Coef	SE Coef	Z	P	Odds	95% CI	
					Ratio	Lower	Upper
Constant	-4.51004	2.00965	-2.24	0.025			
Party							
R	1.91463	0.558377	3.43	0.001	6.78	2.27	20.27
log(10A+1)	0.898979	0.423062	2.12	0.034	2.46	1.07	5.63

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	1964	85.6	Somers' D	0.72
Discordant	318	13.9	Goodman-Kruskal Gamma	0.72
Ties	12	0.5	Kendall's Tau-a	0.34
Total	2294	100.0		

A third assessment involves a graphical comparison of the full and reduced models. On first inspection, the fits – more precisely the coefficients – are very different. A plot of the two fits shows them to differ substantially in the range of amounts from \$0 to \$1,000,000. On closer inspection it becomes clear that within the range of actual contribution amounts (the boxplots in Figure 2 outline these), the fits are quite similar.

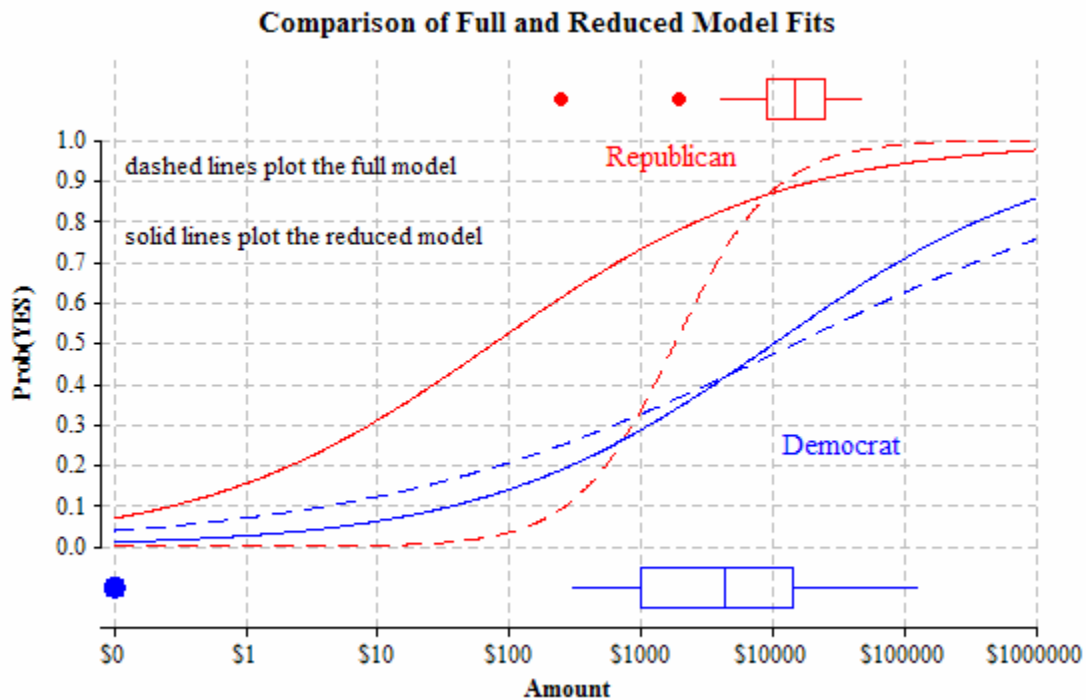


Figure 2. A graphical comparison of the competing models must take into account the range of explanatory values – here displayed with boxplots.

Finally, common sense prevails. A broader view of a plot of the fitted full model reveals that the curves intersect, and that below approximately \$1000 (\$964.76 to be exact), the full model puts the probability of a YES vote higher for a Democrat than a Republican. The result is counterintuitive and best avoided.

3.3.2 Understanding the fit

Effects of Campaign Contributions and Party on the CAFE Vote

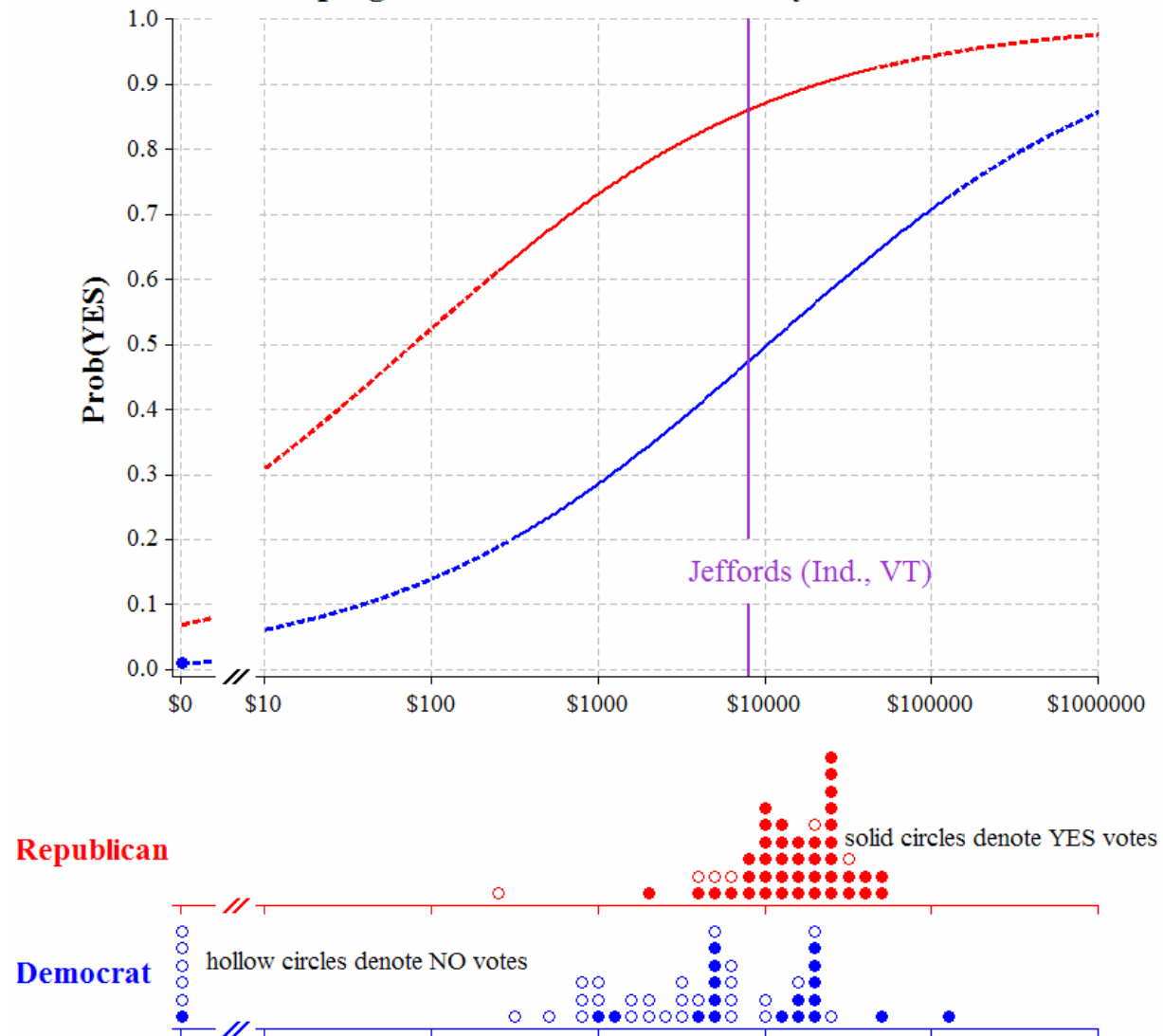


Figure 3. Dotplots of contribution amounts by party, and the fitted logistic regression of the probability of a YES on party and amount contributed. Dotted lines are used for extrapolated values for the regression (six Democrats received no contribution; they are indicated by a solid dot). The vertical line at the \$8000 contribution amount marks Vermont’s James Jeffords, the sole independent in the Senate.

Figure 3 shows a plot not only of the fit I obtain, but of the data. I like this plot because it serves also as a pedagogical tool: students can relate the logistic regression fit to the fill density of the dotplots. Students can obtain fairly accurate quantities from such a plot, helping them anticipate and verify some of the quantitative statements discussed below.

We settle on the reduced model. I want students to synthesize their understanding of the fit by a) identifying coefficients and odds-ratios displayed in Table 4, b) using the coefficients to produce the logit, odds, and probability of a YES vote for any combination of Amount and Party, c) obtaining odds-ratios from the coefficients, and d) interpreting odds ratios in the context of the CAFE setting.

For example, take a Republican who's received \$10,000. On a handheld calculator input 100001 , take the logarithm (the result $F = 5.00$ tells us \$10,000 is a 5-figure amount), multiply by 0.898979 , subtract 4.51004 , and then add 1.91463 (the Republican effect), to obtain a logit of 1.89949 . The odds then are $e^{1.89949} = 6.6825$ (to 1); the probability of a YES vote is $6.6825/(1+6.6825) = 0.8698$. For a Democrat, stop before adding 1.91463 , yielding a logit of -0.05114 . Exponentiate to obtain odds $e^{-0.05114} = 0.9850$ and then a probability of $0.9850/(1+0.9850) = 0.4962$. Students can now plot these estimated probabilities.

Since Amount = 0 is equivalent to $F = 0$, the coefficient of -4.51004 tells us that for a Democrat with Amount = 0, the log-odds of a YES vote are -4.51004 . The odds are $e^{-4.51004} = 0.0110$; the estimated probability of a YES vote is $0.0110/(1+0.0110) = 0.0109$.

The "Republican effect" is quantified by the coefficient 1.91463 . The odds ratio is $e^{1.91463} = 6.7844$. The odds for the Republican voting YES are 6.7844 times that for the Democrat. For example, suppose a Democrat has 0.2500 chance of voting YES. The odds then are 1:3. For a Republican with the same contribution amount, the odds are 6.7844:3, leading to a probability of 0.6934.

The lack of interaction implies that the effect of Amount on Vote is the same for Republicans as for Democrats. However, such a statement requires care, because of the transformation of the Amount variable. Suppose we increase Amount by a fixed amount (or percentage). What is the effect on the likelihood of a YES vote?

The coefficient of 0.898979 tells us that for each 1 unit increase in $\log_{10}(10 \times \text{Amount} + 1)$ there's a corresponding 0.898979 increase in the log-odds of a YES vote. Solved algebraically, increasing the amount 10-fold plus 90 cents results in a 0.898979 increase in the log-odds, which is equivalent to an increase in the odds of a YES vote by a multiplicative factor of 2.4571 . The 90 cents is insignificant – it's an artifact of adding a dime to each contribution amount before multiplying by ten and taking the logarithm. To avoid the algebra, appeal to the transformation of Amount in terms of "figures" and ignore the added dime: Increasing the amount 10-fold (or 900%) results in a 2.4571-fold increase in the odds. Consider the \$10,000 Republican from above – odds 6.6825. A \$100,000 Republican then has odds $2.4571 \times 6.6825 = 16.4196$, leading to a probability of 0.9426.

3.3.3 Some diagnostics

The six Amount = 0 cases – all Democrats – are, after the log transform, outlying cases in the predictor, and deserve attention. What effect do these cases have on the fit?

One of the six senators receiving no contribution voted YES. At issue here is whether the model adequately accounts for this. One can construct a rudimentary goodness of fit test to assess this. Taking 0.0109 as the base probability for a YES vote, the probability of at least one YES vote among six is $1 - 0.9891^6 = 0.0635$. This suggests some lack of fit for these cases. (The interaction term in the full model – which puts Democrats more likely to vote YES at low contribution amounts – is clearly influenced by these cases.)

The estimated 0.0109 probability of a YES for Democrats with Amount = 0 is somewhat inconsistent with the observed 0.1666. One alternative is to isolate these cases, and fit the remaining 93 with a logistic regression. Doing this presumes that a different model applies to Amount = 0 cases; one possible explanation (which students would need to check) is that among these cases are new senators who have had little opportunity to receive contributions from any source.

All the issues discussed in this paper can be addressed for the Amount = 0 cases deleted situation; naturally the values change. The full and reduced model coefficients are similar on inspection, and the interaction term plays less a role ($P = 0.690$ vs. $P = 0.122$ when the Amount = 0 cases are included). Still, an interaction implies that the curves cross: While they do so at a fairly unlikely contribution amount of \$9.22 when the Amount = 0 cases are omitted, this result remains unappealing.

3.4 Additional explorations

Here, in no particular order, are issues worth exploring. For some, computations using the fitted model may be supplemented or replaced by using the plot of the fitted model.

1. Compare the fit that includes the Amount \times Party interaction term to the two fits obtained by treating the Democrats and Republicans separately. (Contrast with least squares regression.)
2. Which senators' votes are most surprising?
3. What is the estimated probability of a YES vote for a senator of either party who received no contribution? (Why should we be careful with these answers?)
4. Sum the fitted probabilities by party. Compare these to the actual Vote by Party table.
5. How much additional money does it take to make a Democrat have the same probability of a YES vote as a Republican? (Or: For a Republican with Amount given, there's a probability p of voting YES. For a Democrat to have the same probability p , how much more must Amount be?)
6. What is the predicted aggregate vote if all senators receive the same Amount? (This depends on the amount. One might use the average contribution.)
7. For a Republican with Amount \$8000 what is the probability of a YES vote? For a non-Republican? Given Jeffords' \$8000 Amount and NO vote, what is the (posterior) probability

that he used to be a Republican? (Some students will know that the truly correct posterior probability is “Exactly 1!”).

8. What effect does including Jeffords as a Democrat (non-Republican) have on the fit?

See Appendix 3 for discussion of these issues.

4. Summary

Students’ chief mechanical difficulty with logistic regression involves the three step process required to work from predictors to logits to odds to probabilities. This is compounded by software without any easy menu commands for getting plots of the fitted regression function. Plots for the case of a single predictor (for example, Figure 27.17 on page 576 of Deveau, Velleman and Bock (2005)) are especially helpful, but are not quickly and easily obtained. (The exercise sets for logistic regression provided in Moore, McCabe, Duckworth and Sclove (2003) fail to call for a single plot.) For the CAFE data, as for any logistic regression data, my responsibilities include supplying students with appropriate graphical summaries. (The fit for the CAFE data is sufficiently complex to not expect that a plot be produced via a simple menu command.)

The CAFE dataset generates a good deal of class interest and discussion. It would be interesting and instructive to perform similar analyses for other issues voted on by lawmakers, and to compare results to those obtained for the CAFE data.

5. Getting the Data

The file `cafe.dat` contains the raw data. The file `cafe.txt` documents the data with a brief description. The dataset is online at <http://www.howdarethey.org/news/cafe/>.

Appendix 1 – Key to variables in `cafe.dat`

Columns

- 1-19 Name (Last, First)
- 20-22 Two-character abbreviation for state
- 23-24 Party affiliation (R = Republican, D = Democrat, I = Independent)
- 25-28 Vote on CAFE standard (YES, NO)
- 29-34 Lifetime contributions from auto manufacturers

Values are aligned and delimited by blanks. There are no missing values.

Appendix 2 – Summarizing the CAFE issue

The following is excerpted from Roberts (2004, pages 295-6).

...By any reasonable standard, the most important step the United States could take to simultaneously improve energy security, cut CO₂ emission, boost urban air quality, and deprive Middle Eastern terrorists of financing would be to raise fuel efficiency requirements. American cars and trucks burn two of every three barrels of oil used in the United States – and one of every seven barrels used worldwide – a figure that is hardly surprising, given that economy standards have been frozen since 1988. Today, American cars need to achieve an average fuel economy of just 27.5 miles per gallon, while “light trucks,” that hugely popular category that includes pickups and SUVs, need achieve only 20.5 miles per gallon. Even a modest improvement in fuel-economy standards – say thirty-two miles per gallon for cars and twenty-four per gallon for light trucks – would by 2010 be saving 2.7 million barrels per day – or nearly twice as much as could be pumped every day from the Arctic National Wildlife Refuge.

Yet so far, even that small change has proved to be a political impossibility. Although such efficiency improvements are already technically feasible – Ford’s Escape SUV, a gas-electric hybrid, gets 36 miles per gallon in the city [although Consumer Reports’ most recent test of the 2005 Escape, published in the August 2005 edition, found values of 22 (city) and 29 (highway)] – U.S. automakers and the big automotive unions have persuaded Congress not to raise fuel-efficiency standards since the late 1980s. Why? Among other reasons, because any regulations requiring greater fuel efficiency will initially favor Japanese and German automakers, whose fleets are already more fuel-efficient – thereby costing U.S. companies more of their market share and U.S. auto workers more of their jobs. And such losses are not inconsequential to American politicians. Since 1990, the U.S. transportation industry has made more the \$256 million in campaign contributions. Whereas nearly 70 percent has wound up with Republicans, Democrats haven’t been shy about asking for auto dollars, especially from the auto workers’ unions. No surprise the CAFE has never come close to being updated.

Appendix 3 – Discussion of items enumerated in Section 3.4

1. For logistic regression, the separate fits are equivalent to the fit that includes the interaction term. This is not the case for least squares, because fitting two groups via indicators assumes that error terms have the same variance across groups, whereas doing separate regressions estimates two variances.
2. McCain’s vote (perhaps not so surprising, given his reputation as a maverick). Senator McCain received a large contribution but voted No.
3. For Democrats, 0.0109: Section 3.3.3 is dedicated to discussion of this. For Republicans, 0.0694: Since no Republican received anywhere near no contribution, this should be treated with skepticism.
4. For Republicans the event probabilities sum to 42.6 (compare to 43 voting YES); for Democrats the sum is 19.00 (19 Democrats voted YES).
5. If A denotes the amount given (contributed to the Republican) then the Democrat must be given \$13.38 more than $134.83A$. The \$13.38 is due to the “extra dime,” and is negligible. One might recognize that the increase in number of figures needs to balance the party effect. Doing this requires the number of figures to increase by $1.91463/0.898979 = 2.1298$; $10^{2.1298} = 134.83$. In short: The Democrat requires almost 135 times the Republican amount.

6. The average amount contributed is \$13806.27 (this includes Jeffords and those senators who received nothing). For a Democrat receiving this amount the fitted probability is 0.5277. Of 50 Democrats, 26.4 are expected to vote YES. For a Republican receiving this amount the fitted probability is 0.8834. Of 49 Republicans, 43.3 are expected to vote YES. These sum to approximately 70 YES votes (exclusive of Jeffords). There were 62 YES votes.
 7. Republican: 0.8596. Democrat: 0.4745. Given a 50-50 split on party affiliation, the posterior probability is $0.8596/(0.8596+0.4745) = 0.6444$.
 8. Including Jeffords as a Democrat has negligible effect on the fit. Doing so results in a (reduced model) fit with coefficients that are, on inspection, very close to those displayed in Table 4. (None of the coefficients change by more than 0.1 of a standard error when Jeffords is included.)
-

References

- “Automobile and Light Truck Fuel Economy: The CAFE Standard” (2002), *Almanac of Policy Issues* [Online], September 25, 2002, http://www.policyalmanac.org/environment/archive/crs_cafe_standards.shtml.
- “CAFE Standard,” (2002) Public Campaign [Online], www.howdarethey.org/news/cafe/.
- “The fuel-economy shell game” (2004), *Consumer Reports* 69(5), 8, Yonkers, NY: Consumer’s Union.
- “Not easy being green” (2005), *Consumer Reports* 70(8), 50-57, Yonkers, NY: Consumer’s Union.
- DeVeaux, R. D., Velleman, P. F., and Bock, D. E. (2005), *Stats: Data and Models*, Boston: Pearson Education, Inc.
- Hakim, D. (2004, March 25), “Kerry Is Sticking With Plan to Raise Auto Fuel Efficiency,” *The New York Times* [Online].
- Hakim, D. (2004, May 5), “Average U.S. Car Is Tipping Scales at 4,000 Pounds,” *The New York Times* [Online].
- Hosmer, D. W., and Lemeshow, S. (2000), *Applied Logistic Regression* (2nd ed.), New York: John Wiley & Sons, Inc.
- Love, T. E. (1998), “A Project-Driven Second Course,” *Journal of Statistics Education* [Online], 6(1).
- Moore, D. S, McCabe, G. P., Duckworth, W. M., and Sclove S. L. (2003), *The Practice of Business Statistics*, New York: W. H. Freeman and Company.

Roback, P. J. (2003), "Teaching an Advanced Methods Course to a Mixed Audience," *Journal of Statistics Education* [Online], 11(2).

Roberts, Paul (2004), *The End of Oil: On the Edge of a Perilous New World*, New York: Mariner Books.

Simonoff, J. S. (1997), "The 'Unusual Episode' and a Second Statistics Course," *Journal of Statistics Education* [Online], 5(1).

U.S. Senate Roll Call Votes 107th Congress - 2nd Session [Online],
[www.senate.gov/legislative/LIS/roll_call_lists/roll_call_vote_cfm.cfm?congress=107&session=2
&vote=00047](http://www.senate.gov/legislative/LIS/roll_call_lists/roll_call_vote_cfm.cfm?congress=107&session=2&vote=00047).

Scott R. Preston
Department of Mathematics
SUNY Oswego
Oswego, NY 13126
srp@oswego.edu