Quantitative Methods II
Assignment 5
September 25, 2011
Solutions

There is no way to determine which variable combination is *best* in an absolute sense, since we do not know the process that gave rise to the data. We can only select a variable combination that appears logical and has certain desirable properties. One of those desirable properties is that the model (these chosen variables) does not violate the assumptions of Ordinary Least Squares (OLS). Another is that the combination maximizes the adjusted R^2 measure.

We start with a data set and a goal of predicting the probability of a ballot measure passing in Washington. The model must contain one dichotomous variable and one continuous variable. They can enter additively or with an interaction term. As the dependent variable is bounded to the interval $(0, 1)$, we transform using the logit function.

As of now, the only method we have for comparing models is to compare adjusted R^2 values, with the model belonging to the larger of the two being preferred. Using this method will all of the possible combinations of variables nets us the research model

$$\text{logit}(\texttt{propFavor}) \sim \texttt{south} * \texttt{povertyRate}$$

Technically, the model with `povertyRank` provided a (marginally) better adjusted R^2. However, as it is almost always better to use rates than ranks, we use `povertyRate`. The adjusted R^2 for this model was 0.7247 — quite high for the social sciences.

With that as our model, we now must check the assumptions of OLS. Unfortunately, the independent variables are highly correlated ($\rho = 0.615, t = 4.347, \nu = 31, p = 0.0001$). This will affect the statistical significance of these two variables. However, as we are only concerned with the prediction, this will not be an issue.

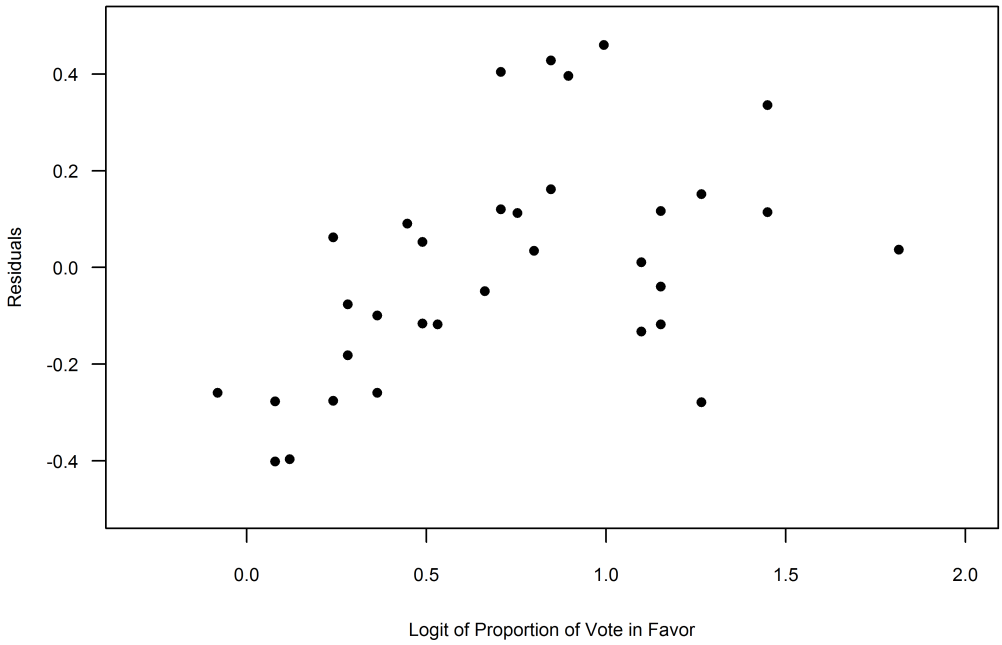

Figure 1. *Residual plot of the final model. The lack of an apparent funnel shape indicates that the heteroskedasticity is not too severe.*

The residuals have mean zero, are reasonably Normally distributed according to the Shapiro-Wilks test ($W = 0.9593, p = 0.2465$), and appear to have constant variance (there is no funnel shape in the residual plot, Figure 1.

Thus, we can reasonably conclude that the assumption of

$$e_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

is not severely violated in our model.

Thus, as the model does not violate our assumptions, we can proceed. For Washington (with poverty rate of 10.2% and not being in the South), we predict that 65% of the people will vote in favor of the ballot measure (with a 95% confidence interval of 42.4% to 82.5%). Using Monte Caarlo techniques, this means that the ballot measure will have a 90.6% chance of passing (see Figure 2).

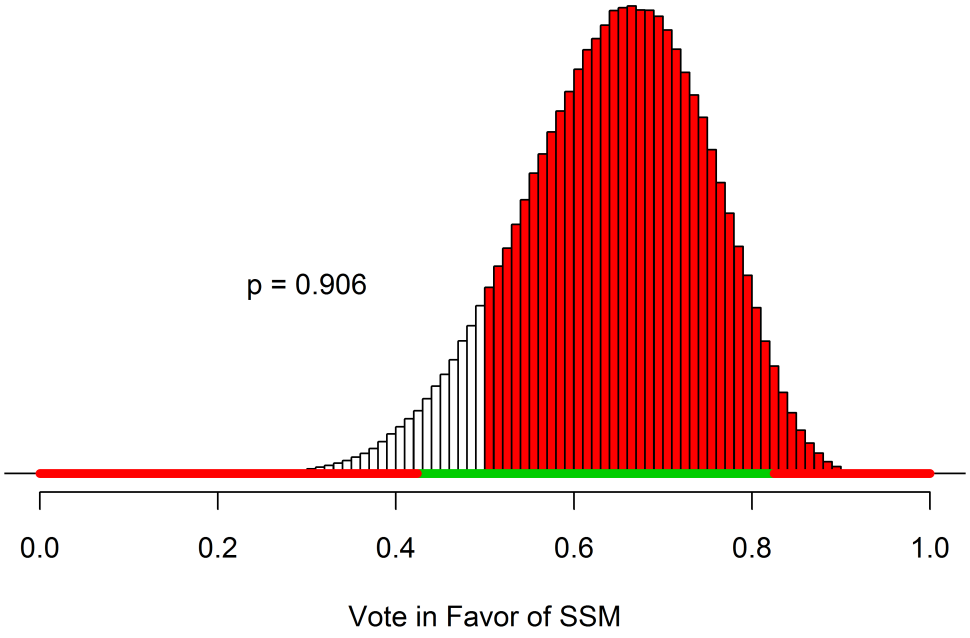

Figure 2. *Histogram of the predicted outcomes from the one million Monte Carlo trials. The red-shaded region indicates that the ballot measure passes. The underlines indicate the 95% confidence interval for our point estimate of a 65% vote in favor.*