

Cross-Validation and Balance Checking

October 13, 2009

Cross-Validation

- How do we check the accuracy of a predictive model?
- Many predictive models tend to over-fit, so good practice is to choose a predictive model based on a training data set and then check its predictive accuracy on a separate validation dataset.
- Training datasets aren't available, as in our regression discontinuity case, but one useful technique for testing our model's predictive accuracy is known as **cross-validation**.

Cross-Validation

- We usually refer to prediction error as the expected squared difference between a future response and its prediction from the model:

$$\text{PE} = E\{(y - \hat{y})^2\}$$

- In cross validation, we use part of the data to fit the model, and a different part to test it.
- Suppose we split the data into K parts. Let $k(i)$ be the part containing observation i . Denote the by $\hat{y}_i^{-k(i)}$, the fitted value for observation i , computed with the $k(i)$ th part of the data removed. Then the cross-validation prediction error is:

$$\text{CV} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i^{-k(i)})^2$$

“Leave-One-Out”

- Often we choose $k = n$, resulting in **“leave-one-out”** cross-validation.
- For each observation i , we refit the model leaving that observation out of the data, and then compute the predicted value for the i th observation and compute the predicted value \hat{y}_i^{-i} . We do this for each observation and then compute the average cross-validation sum of squares
$$CV = \sum (y_i - \hat{y}_i^{-i})^2 / n$$

What are we predicting?

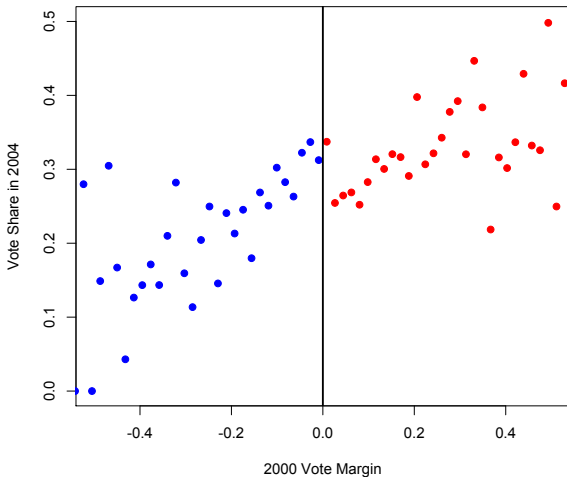
$$\min \sum_{i:c-h < X_i < c} (Y_i - \alpha_l - \beta_l \cdot (X_i - c))^2$$

and

$$\min \sum_{i:c \leq X_i < c+h} (Y_i - \alpha_r - \beta_r \cdot (X_i - c))^2$$

- The value of $\mu_l(c)$ is estimated as $\hat{\mu}_l(c) = \hat{\alpha}_l + \hat{\beta}_l \cdot (c - c) = \hat{\alpha}_l$ and $\hat{\mu}_r(c)$ is estimated as $\hat{\mu}_r(c) = \hat{\alpha}_r + \hat{\beta}_r \cdot (c - c) = \hat{\alpha}_r$.
- $\hat{\tau}_{RD} = \hat{\alpha}_r - \hat{\alpha}_l$.

Effect of Incumbency on Vote Share



Picking h

- We need to pick an h and cross-validation is a natural “hands-off” technique.
- Predict each y_i using x_i values within h . Note that we each treat each y_i as point at a boundary.
- To emulate the fact that RD estimates are based on regression estimates at the boundary, the regression is estimated using only observations with values of X on the left of X_i ($X_i - h \leq X < X_i$) for observations on the left of the cutpoint ($X_i < c$). For observations on the right of the cutoff point ($X_i \geq c$), the regression is estimated using only the observations with values of X on the right of X_i ($X_i < X \leq X_i + h$)

- Formally, let $\hat{Y}(X_i)$ be the predicted value of Y obtained using the regressions described above. The cross validation criterion is defined as

$$CV_Y(h) = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}(X_i))^2$$

with the corresponding cross-validation choice for the bandwidth

$$h_{CV}^{opt} = \arg \min_h CV_Y(h)$$

Results of Cross-Validation

