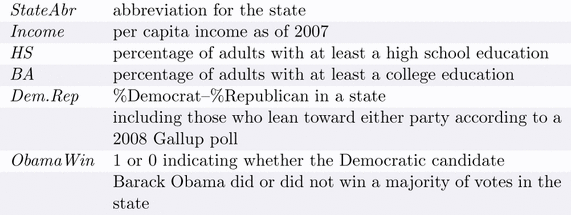
Can we use state-level variables to predict whether a state votes for the Democratic versus the Republican presidential nominee? The file **Election08** contains data from 50 states plus the District of Columbia. The variables recorded are:



See the attached R output to help answer the questions below.

1. Consider the separate logistic regression models that predict *ObamaWin* using each of the predictors *Income, HS, BA*, and *Dem.Rep* (model1, model2, model3, and model4). Which of these variables does the most effective job of predicting this response? Which is the least effective? Explain the criteria you use to make these decisions.

With only this information available to us (see #2), we must use p-values to choose the best (and worst) predictor. *Income* and *Dem.Rep* have p-values that are essentially equal, and both are highly significant. *HS* has the largest p-value, and is thus the least effective. This is not surprising when we look at the descriptive stats and boxplots, since the difference in *HS* for the two groups (state voted for Obama or not) is very small.

1. What additional information do you need to determine which of the four variables is the most appropriate predictor of *ObamaWin*?

Is the condition of linearity met? We need empirical logit plots (log(odds of ObamaWin) vs. x) for each predictor. We are looking for linearity in those plots.

1. Using model1, use the estimated slope from the logistic regression to compute an estimated odds ratio and write a sentence that interprets this value in the context of this problem.

Slope = 0.0003494 🡪 OR = e^(0.0003494) = 1.000349

A state with $1 higher per capita income has .035% higher odds of voting for Obama compared to the state with $1 lower per capita income.

1. Interpret the 95% confidence interval for the odds ratio in #3.

CI for slope = (0.000144, 0.000555) 🡪 CI for OR = (1.000144, 1.000555)

We are 95% confident that a state with $1 higher per capita income has between 100.0144% and 100.0555% the odds of voting for Obama compared to the state with $1 lower per capita income.

1. The units of the *Income* variable are dollars, with values ranging from $28,845 (Mississippi) to $61,092 (District of Columbia). The odds ratio and interval in #3 and #4 are awkward to interpret since they deal with the change in the odds when state income changes by $1, a very trivial amount! To get an odds ratio that may be more meaningful, we created a new variable (call it *IncomeTh*) using *Income*/1000 to express the state per capita incomes in $1000s. The logistic regression using *IncomeTh* as the predictor of *ObamaWin* is model5. How does the fitted prediction equation change?

The significance of the predictor (p-value) is the same. The coefficient is 0.3494, which is 1000 times the coefficient in model1 (not surprisingly!).

1. Using model5, use the estimated slope from the logistic regression to compute an estimated odds ratio and write a sentence that interprets this value in the context of this problem.

Slope = 0.3494 🡪 OR = e^(0.3494) = 1.418216

A state with $1000 higher per capita income has 142% the odds of voting for Obama compared to the state with $1000 lower per capita income.

1. Using model5, calculate the probability of Obama winning a state whose per capita income is $35000.

Odds = e^(-12.4251+0.3494\*35) = 0.82193

Prob = odds/(1+odds) = 0.45113

1. The logistic model with *ObamaWin* as the response and *Dem.Rep, HS, BA*, and *Income* as the predictors is model6. Which predictor has the strongest relationship with the response in this model? Which predictors (if any) are not significantly related to *ObamaWin* in that model?

In this model, *Dem.Rep* is the only predictor that has a significant relationship with *ObamaWin*, as seen in the Wald tests for the individual predictors (all other p-values>0.1).

1. The insignificance of certain variables in model6 seems to conflict with the significance of the variables in models 1,2,3,4 in #1. Explain this apparent contradiction.

The Wald tests for each predictor are conducted assuming the other predictors are all in the model. So even though *Income* was highly significant by itself (in model1), when *HS*, *BA*, and *Dem.Rep* are all in the model, *Income* is no longer significant. (Similarly for *HS* and *BA*.) *Dem.Rep* is a significant predictor of *ObamaWin* even when the other 3 variables are in the model.

In addition, notice that *Income* and *BA* have a relatively strong correlation (r = 0.83). So even though each variable is highly significant by itself, when the other variable is already included in the model, that term becomes insignificant.

1. Perform a nested drop-in-deviance test to compare model6 to model4.
   1. Write down the hypotheses for this test.

Ho: model4 (Dem.Rep only) is suitable for predicting ObamaWin

Ha: model6 (all 4 variables) is a significant improvement over model4

* 1. Write down the test statistic, and how you would find the p-value (you may write the R code needed to find the p-value; or explain in words, but you must be explicit!).

Test stat = G = 27.167 – 9.7252 = 17.4418

This should follow a chi-square distribution with 3 degrees of freedom. To find the p-value, the R code would be:

anova(model4, model6, test=”Chisq”)

OR

1 – pchisq(17.4418, df=3)

* 1. The p-value from part (b) is 0.000573. Make a conclusion in context.

The addition of the 3 variables significantly improves the prediction compared to the single-variable model of *Dem.Rep*; at least one of these variables must be a significant predictor of *ObamaWin*.

1. You want to find the best 1-, 2-, 3-, or 4-variable model to predict *ObamaWin*. If you were working with this data, what would your next step be?

There are two logical directions:

1. Start with the 4-variable model (model6) and start deleting insignificant terms one at a time. *HS* would go first; re-run the model with the other 3 terms and see what happens.
2. Since *Income* and *BA* have a pretty strong correlation, we probably only want one of them in the model. *HS* is clearly not significant when the others are included, so delete that as an option. Thus two potential models would be *Dem.Rep* and *Income*, or *Dem.Rep* and *BA*. Fit both of these models and compare based on p-values or misclassification rates.
3. We talked about the relationship between chi-squared tests and logistic regression. Could a chi-square test be performed to tell us something about the relationship between *ObamaWin* and any of these 4 variables? If your answer is No, explain why not. If your answer is Yes, explain what the hypotheses the chi-square test would test.

No, a chi-sq test would not be useful here because all of our potential predictors are numerical. Chi-square tests are only relevant when we have two categorical (or binary) variables to compare. Here, we have our binary response variable, but no categorical/binary predictor.