STAT 213: Project III (100 pts)

Colin Reimer Dawson

Due by the end of classes (Fri. 5/6)

The data for this project concerns U.S. House members' votes for against the Affordable Care Act ("Obamacare") in March 2010.

Data URL: `http://colinreimerdawson.com/data/obamacare.csv`

The variables are:

`Vote`	(0 = no or 1 = yes)
`Private`	% with private health insurance in district
`Public`	% with public health insurance in district
`Uninsured`	% without health insurance
`State`	state
`DWNom`	a measure of conservative (positive) to liberal (negative) voting
`ObamaMargin`	Obama% - McCain% in the district in 2008

The response variable here is `Vote`. The first goal is to determine which factors predict `Vote`, i.e., to construct a model that describes the relationship between `Vote` and other variables.

The second goal is to predict the Vote results of each of the 216 representatives in the "holdout" sample. I'll use your model to make predictions of $\pi_{new} = P(\texttt{Vote} = 1)$ for each case in the holdout sample. If the true value of `Vote` is 1, then the error for that case is $1 - \hat{\pi}_{new}$. If the true value is `Vote = 0`, then the error is $\hat{\pi}$. In other words, the closer the predicted probability is to the true value, the lower the error. Part of your grade is determined by the score your model achieves on the held out data.

In your writeup:

1. Describe the process you used to select a model. Include any relevant plots, etc. However, although you will produce your writeup using Markdown, unlike in previous projects, you should *not* have the code or raw R output appear in the final pdf. Instead, use natural language and, where relevant, include numbers, to make clear what you did without having to see the code or raw output. To control this, add the line

```
opts_chunk$set(
    echo = FALSE,
    message = FALSE,
    results = 'hide'
)
```

   in the first code chunk before you turn in your final result (you may find it useful to set `echo = TRUE` and `results = 'markdown'` while you are writing).

2. Give two regression equations for your final model: one in the form:

$$\hat{\pi} = \frac{exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}{1 + exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}$$

   the other in the form

$$\text{logit}(\hat{\pi}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

   and interpret each coefficient.

3. Report any relevant fit measures that you used to decide among models.

The project grade is determined by the following components:

(a) Completeness and statistical correctness (40%).

(b) Clarity / presentation (30%).

(c) Predictive ability of your final model (30%). The error measure statistic can range from 0 to 216, but it is not reasonable to expect to achieve a perfect zero; as such, the grade scale will be determined relative not to perfect prediction, but relative to the performance of the best possible model using the six available predictors.