Week 2, HW4: Logit and LASSO

Instructor: Jacob LaRiviere

Email: jlariv@microsoft.com

1) Go to github and read in the SAT dataset. You'll see there is a .txt file too which is a readme. It has descriptions of all the variables.
    a. Lets try to predict who is going to take the SAT like the online Rmd output.
    b. First use dplyr to summarize the data. What percent of the sample took the SATs?
        i. What is the average score rank of high school students who took the SAT versus those that didn't?
    c. Use everything as a RHS variable except `pict`, `lgsc`, and `mosaic` [since we don't know what they are and `sat` since it isn't a direct outcome
    Note: I think normalized scores are normalized SAT scores rather than normalized grades, which is what I thought initially so you might leave those out too.
    d. Consider the following logit object commands. What would you use to compare OLS coefficients to logit coefficients (hint: it's the third one!)? Does this comparison make sense? Estimate the OLS model and compare visually.
```
summary(fit) # display results
confint(fit) # 95% CI for the coefficients
exp(coef(fit)) # exponentiated coefficients
exp(confint(fit)) # 95% CI for exponentiated coefficients
predict(fit, type="response") # predicted values
residuals(fit, type="deviance") # residuals
```
    e. If you believed this model, what would you do from a policy perspective to try to encourage more students to take the SAT?
2) Let's return to the orange juice dataset and investigate how store demographics are related to demand.
    a. Run a LASSO model for the same model cross validated OLS that gave you the lowest MSE.
    b. What are the coefficients that are selected from the LASSO technique? Is it all of them?
        i. What are the point estimates of those features?
    c. Now set *alpha* in the glmnet function to .5. Does the predictive power of the model increase? Why or why not do you think?
    d. Now use all of this same code but start by withholding 10% of the OJ data for the training of LASSO.
        i. Use this holdout data to perform out of sample testing of the LASSO model with the lambda with the min out of sample MSE.
        Here is the code for using an existing model on new data:

```
predict(cvfit, newx = x[1:5,], s = "lambda.min")
```

        See: https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html
        NOTE: you can get around the feature engineering by doing the following:

```
X <- model.matrix(formula, dataframe) # where formula is a general formula like you would run for OLS.
```

https://www.rdocumentation.org/packages/MatrixModels/versions/0.4-1/topics/model.Matrix