STAT 215: SUPPLEMENTARY LOGISTIC REGRESSION PROBLEMS

COLIN REIMER DAWSON

You will need R to do the following problems. You are encouraged, but not required, to write up your answers using RMarkdown (you may want to use one Markdown document for this and the textbook problems together).

(1) The `CAFE` dataset (in the `Stat2Data`), also described in examples throughout Chapter 10 of the text, includes information about how various U.S. senators voted on an amendment that would hamper the proposed Corporate Average Fuel Economy (CAFE) bill. The bill would have tightened regulations on fuel economy standards, and so a Yes vote on the amendment acts in opposition to tightened regulations. Examine the documentation on the dataset for more information.

The quantitative variable `LogContr` records how much money each senator received from the auto industry, on a log scale. The `Dem` indicator is 1 if the senator caucused with Democrats, 0 otherwise. Fit and compare a set of logistic regression models to address the following questions. For each question, identify the model(s) you used to address the question, and interpret each coefficient.

(a) Does the probability of a `Yes` vote increase with (log) campaign contributions?

(b) Does the probability of a `Yes` vote differ between parties?

(c) Is the relationship between log campaign contribution and Vote different for those that caucus with Democrats?

(d) Does knowing whether a senator caucuses with Democrats improve predictive ability *after controlling for* campaign contributions?

Date: Last Revised November 15, 2017.

(2) The dataset `Leukemia` records treatment outcomes for 51 leukemia patients (in the binary variable `Resp`, where 1 means the patient responded to treatment). Pretreatment covariates (predictors) that might be relevant are recorded in `Age`, `Smear`, `Infil`, `Index`, `Blasts`, and `Temp`.

(a) Fit a logistic model to predict `Resp` from the other six variables. Interpret the relationship between `Age` and `Resp`, and between `Temp` and `Resp`.

(b) Consider performing model selection (you don't need to actually perform it) to choose a subset of the predictors, so that physicians know what information is valuable to record when deciding whether to treat. If a predictor is nonsignificant in the full model, is it possible that it will end up in the final model? Explain.

(c) Use a nested likelihood ratio test (same code as nested F-test but adding `test = "LRT"` to see whether the full model fits any better than a model that simply retains all of the "significant" predictors.

(d) In the reduced model above, how if at all have standard errors changed for the coefficients of the remaining predictors, compared to the full model? If there is a substantial change, explain why that might be.

(e) Perform stepwise selection to identify a set of predictors to keep. Does the resulting model significantly differ from the full model, based on a nested likelihood ratio test?

Key differences in R code between linear and logistic regression (all caps indicates a placeholder).

```
## To fit a model
model <- glm(FORMULA, family = "binomial", data = DATA)
## To do an overall likelihood ratio test
anova(model, test = "LRT")
## To do a nested likelihood ratio test
anova(REDUCED, FULL)
## To do stepwise selection based on AIC
step(NULLMODEL, scope = list(upper = FULLMODEL))
```