Homework 3: Binary and Count Regressions

your name

Due: February 23rd at 11:59 PM

This homework set will cover problems concerning binary and count regression models. Point totals for specific problems are given, 10 points will be reserved for correct submission of the homework assignment.

Problem 1 [15 points]: This problem concerns manual creation of summary tables from nothing more than the observed data and the assumed model.

- **part a** [5 points]: Manually write your own iteratively reweighted least squares algorithm which maximizes the logistic regression log likelihood for the CCSO example in the notes. Report the estimated submodel canonical parameter vector $\hat{\beta}$ and reproduce the summary table (up to convergence tolerance differences) without using the `glm` or `summary` commands. You can ignore deviance residuals.
- **part b** [5 points]: Manually write your own iteratively reweighted least squares algorithm which maximizes the Poisson regression log likelihood for the Galapagos example in the notes. Report $\hat{\beta}$ and reproduce the summary table (up to convergence tolerance differences) without using the `glm` or `summary` commands. You can ignore deviance residuals.
- **part c** [5 points]: Manually write your own Fisher scoring algorithm for one of the parts above, and compare estimates of β from the Fisher scoring algorithm and the iteratively rewighted least squares algorithm.

Problem 2 [10 points]: Complete the following parts:

- **part a** [5 points]: Explain important findings and model information from the summary table produced by a call to summary(m1) in the CCSO example in the logistic regression notes. Keep in mind that we restricted attention to "other traffic offenses" in the CCSO example, and that this data is observational.
- **part b** [5 points]: Explain important findings and model information from the summary table produced by a call to summary(m1) in the Galapagos islands example in the count regression notes.

Problem 3 [10 points]: This problem concerns MLEs and inferences of modeling parameters using the CCSO example in class. Do the following:

- **part a** [5 points]: Compute MLEs and estimated standard errors for the saturated model parameter μ from the logistic regression with race, sex, and arrestAge as predictors for atleastone fit to the CCSO data restricted to "other traffic offenses". Compare with `predict.glm`.
- **part b** [5 points]: Then construct Wald based confidence intervals for the estimated mean value parameters. Also construct confidence intervals

$$(g(\hat{\beta} - z_{\alpha/2}se(\hat{\beta})), g(\hat{\beta} + z_{\alpha/2}se(\hat{\beta}))).$$

Comment on any noticeable differences between these two confidence intervals for $\hat{\mu}$.

Problem 4 [10 points]: Construct a nonparametric bootstrap procedure that estimates the uncertainty associated with both estimates of the average treatment effect (ATE) of online learning in the logistic regression notes. Do the conclusions change when we factor in the uncertainty obtained from the nonparametric bootstrap procedure? Explain.

Problem 5 [15 points]: Use the `dvisits` data in the `faraway` package to answer the follow parts:

- **part a** [1 points]: Make plots which show the relationship between the response variable, doctorco, and the potential predictors, age and illness.
- **part b** [2 points]: Combine the predictors chcond1 and chcond2 into a single three-level factor. Make an appropriate plot showing the relationship between this factor and the response. Comment.
- **part c** [2 points]: Build a Poisson regression model with doctorco as the response and sex, age, agesq, income, levyplus, freepoor, freerepa, illness, actdays, hscore and the three-level condition factor as possible predictor variables. Considering the deviance of this model, does this model fit the data?
- **part d** [2 points]: Plot the residuals and the fitted values — why are there lines of observations on the plot? Make a QQ plot of the residuals and comment.
- **part e** [2 points]: Use a stepwise AIC-based model selection method. What sort of person would be predicted to visit the doctor the most under your selected model?
- **part f** [2 points]: For the last person in the dataset, compute the predicted probability distribution for their visits to the doctor, i.e., give the probability they visit 0, 1, 2, etc. times.
- **part g** [2 points]: Tabulate the frequencies of the number of doctor visits. Compute the expected frequencies of doctor visits under your most recent model. Compare the observed with the expected frequencies and comment on whether it is worth fitting a zero-inflated count model.
- **part h** [2 points]: Fit a comparable (Gaussian) linear model and graphically compare the fits. Describe how they differ.

Problem 6 [30 points]: This problem will expand on analyses of the CCSO data done in the course notes. The goal is to investigate racial biases in the propensity of people who spend at least one day in jail. We will restrict attention to other traffic offenses as done in class. Do the following:

- **part a** [15 points]: Analyze the CCSO data restricted to other traffic offenses using a binary response regression models. Your analysis must consider the following variable:
  - race
  - sex
  - arrestAge
  - employmentStatus
  - releaseReason
  - repeatOffenders: a variable that you will have to create to indicate whether or not an arrested individual was previously arrested.
  - multipleOffenses: a variable that you will have to create to indicate whether the arrested individual has committed multiple offenses upon a single arrest.

Note that "consider" does not mean that a variable has to be included in a regression model in this context; it means that your analysis needs to be informed by the above variables. You may want to combine factor levels in these variables, or you may want to throw out individuals belonging to a factor level that may be hard to interpret or is sparse. You are encouraged, but not required, to consider other variables. Report your final regression model, and justify your choice for your final model. Report interesting findings.

- **part b** [5 points]: Report observed propensities of spending at least one day in jail broken up by race and employment status after restricting attention to people who were released because of a bond payment. Comment on racial discrepancies.
- **part c** [5 points]: Pretend you are an expert witness in a court case where the Champaign County Sheriff's Office is being sued for racial bias in sentencing. Suppose you are hired by the side prosecuting the CCSO. Present an argument for racial bias in sentencing based on your analysis above. You can add further analyses if you think they are needed.
- **part d** [5 points]: Pretend you are an expert witness in a court case where the CCSO is being sued for racial bias in sentencing. Suppose you are hired by the CCSO. Present an argument that there is no racial bias in sentencing based on your analysis above and any additional analyses if you think they are needed. You are allowed to criticize the model you presented in part a.