Homework 3: Binary and Count Regressions

your name

Due: February 17 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 [10 points]: Manually write your own Fisher scoring algorithm which maximizes the logistic regression log likelihood for the CCSO 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.

Problem 2 [10 points]: Manually write your own Fisher scoring 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.

Problem 3 [5 points]: Derive the log-odds ratio of $x + 1$ to x when $Y = 1$, and observe that the log-odds ratio does not depend on x. Comment on this finding.

Problem 4 [10 points]: Complete the following parts:

(a) 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.

(b) 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 5 [10 points]: Derive expressions and compute standard errors $\text{se}(\hat{\mu})$ in the logistic regression CCSO example without using `predict.glm`. Then construct Wald based confidence intervals for the estimated mean value parameters. Also construct confidence intervals $(g(\hat{\beta} - z_{\alpha/2}\text{se}(\hat{\beta})), g(\hat{\beta} + z_{\alpha/2}\text{se}(\hat{\beta})))$. Comment on the differences between these two confidence intervals for $\hat{\mu}$.

Problem 6 [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 7 [15 points]: Use the `dvisits` data in the `faraway` package to answer the follow parts:

(a) Make plots which show the relationship between the response variable, doctorco, and the potential predictors, age and illness.

(b) 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.

(c) 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?

(d) 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.

(e) 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?

(f) 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.

(g) 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.

(h) Fit a comparable (Gaussian) linear model and graphically compare the fits. Describe how they differ.

Problem 8 [20 points]: Analyze the CCSO data set with Days in Jail as the response variable. You are allowed to dichotomize the response into a binary variable. Restrict attention to other traffic offenses as done in class. Your analysis needs to consider the variables considered in class as well as repeat offenders, multiple offenses, released reason, and agency. The determination of repeat offenders and multiple offenses can be done via the jacket number variable. Report interesting significant and null findings and determine that your final model is appropriate. You are allowed to use other inferential techniques than GLM. If you do so, then you need to justify your choices.