STAT0032: Exercise Sheet #5

The exercises in this sheet focus on generalised linear models. As in the previous sheet, some of the questions are from James et al., "An Introduction to Statistical Learning" (ISLR).

1. (Exercise 1 of Chapter 4, ISLR.) Using a little bit of algebra, prove that (1) is equivalent to (2). In other words, the logistic function representation and logit representation for the logistic regression model are equivalent.

$$p(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} \tag{1}$$

$$\frac{p(X)}{1 - p(X)} = \exp(\beta_0 + \beta_1 X) \tag{2}$$

2. (Exercise 6 of Chapter 4, ISLR.) Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficients, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

   (a) Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.

   (b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

3. (Exercise 9 of Chapter 4, ISLR.) This problem has to do with odds.

   (a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

   (b) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default?

4. (Computer implementation) Exercise 13, items (a)-(d) and (i), of Chapter 4, ISLR. Read the corresponding explanation of confusion matrices from the book.

5. (Computer implementation) Exercise 14, items (a)-(c) and (f), of Chapter 4, ISLR.

6. (Computer implementation) (Adapted from Gelman and Hill, Chapter 6) The dataset `frisk_with_noise.dat` (open it in a text file first to better understand it) measures the number of stops of members of an ethnic group within a corresponding precint in New York City (3 ethnicities, 75 precints). It also includes the number of arrests for the same ethnicity × precint in the previous year (column `past.arrests`).

(a) Fit a Poisson regression model for the outcome "number of stops" with an offset of log of the population (pop). Use the intercept only, no further covariates. See the help file of glm for an explanation of the offset term. Interpret the outcome.

(b) Now add ethnicity indicators. Contrast this model with the one in item (a).

(c) Now add the precint variable. Report on the reduction of deviance compared to the expected value that we would get if precint was not an useful covariate. Provide further interpretation of the coefficients.

(d) Make a plot of the raw residuals against predicted values. Do the same for the standardized residuals that assume the Poisson output model. What can you say about overdispersion?

(e) Attempt a negative binomial fit, reporting the results.

(f) What if we use a different offset, like past arrests?

7. (Computer implementation) (Adapted from Gelman and Hill, Chapter 6) Dataset `risky.behavior` contains data from a randomized controlled trial targeting heterosexual patients at high risk of HIV infection. The intervention consistens of counseling sessions for practices that potentially reduce their likelihood of contracting HIV. Couples were randomized either to a control group (indicated by columns `couple` and `women_alone` being equal to zero), a group in which just the woman participated, or a group in which both members of the couple participated. One of the outcomes examined after three months was "number of unprotected sex acts" (`fupacts`). Each member of the couple is presented as a separate row, all the women first followed by the corresponding men. Other variables include the number of estimated acts prior to treatment (`bupacts`) and whether the corresponding participant was HIV positive (`bs_hiv`).

   The data is in the Moodle page. You will need to figure out how to load the data by yourself. Google is your friend.

   (a) Model the outcome as a function of treatment using Poisson regression. For some reason the outcome is not an integer, so round it up first. You might also want to transform some of the discrete variables into factors (see R documentation) so that the fitting will make sense. Does the model fit well? Is there evidence of overdispersion?

   (b) Extend the model to include pre-treatment variables. Does the model fit well? Is there evidence of overdispersion?

   (c) Fit an overdispersed Poisson model (look at the R documentation for that). What do you conclude regarding effectiveness of the intervention?

   (d) Given that we have data for both members of each couple, does this raise concerns how the data was used in your analysis? What would you do differently?

8. (Computer implementation) (Adapted from Gelman and Hill, Chapter 6) Using the individual-level survey data from the 2000 National Election study (file `nes5200_processed_voters_realideo.dta`) will contain many years besides 2000), predict party identification `partyid3` (which has 5 levels) using ideology (`ideo`) and demographics (I suggest restricting it to age and `income` to keep it simple) with an ordered multinomial logit model.

(a) Summarize the parameter estimates and explain the results of the fitted model.

(b) Now use the same covariates and logistic regression to predict the choice of supporting a mainstream candidate (Democrats or Republicans) against not voting/voting third party. Do this by first processing variable `presvote`. Interpret the model.

(c) Package ARM is a package for data analysis that includes a non-standard plot called a binned residual plot (`binnedplot`). Read the documentation and use a binned residual plot to assess the fit of the model in (b), commenting on the insights obtained. Pay attention to possible missing values in part (b)!

(d) Explain what the interpretation of the model in part (a) would be if we attempted to model the output as an ordinal variable, and which shortcomings you might expect.