Review Exercise

Psychology 319 – Multilevel Regression Modeling

Instructions. These questions constitute a broad review of course material and key principles. Show your R/HLM code, your input, and your output. Feel free to email me for hints if you get stumped.

1. *Logistic Regression – Theory.* In a class of 100 students, a logistic regression is performed on the course outcome (pass or fail) with midterm exam score as a predictor. The midterm is a continuous variable with a long-run normal distribution with a mean of 60 and a standard deviation of 15. Imagine that the actual model is $\Pr(\text{Pass}) = \text{logit}^{-1}(-24 + 0.4x)$.

   (a) Create artificial *integer* data on x ($N = 100$) that fit this model. (Hint. For the exam scores, you could sample from x using `rnorm()`, then round to 0 decimals.)

   (b) Create artificial data for the outcome. (Hint. Compute the individual's probability of passing from the model, then use `rbinom` to create the outcome.)

   (c) Suppose that, in the population, you transformed the exam scores to have a mean of 0 and a standard deviation of 1. What would the model for Pr(Pass) change to?

   (d) Fit the model to your artificial data, using the logistic regression capabilities of R. When you get the coefficient estimates, plot the estimated curve of the form $\Pr(\text{Pass}) = \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x)$. (Hint: you might consider loading the R library `arm` and using the `invlogit()` function.) Add to this plot, in a different color, the actual model that generated the data, using the `curve()` function.

   (e) Suppose there was an additional predictor of passing the course. Call it `goodlooks`. Imagine that this predictor has a normal distribution in the population with the same metric as the exam scores, but is totally independent of academic performance in any aspect of the course. Create 100 simulated observations on this variable. Now simulate the actions of an analyst who things `goodlooks` might predict the probability of passing the course, and add it to your logistic regression predictor list and refit the

model. What does the deviance change to? Use the `anova` option to test the significance of the new predictor.

2. *Regression Modeling in Practice.* The file *baseball.txt* has data from the last two seasons of play in the National League. Included are the abbreviation for the team name, the average runs for, the average runs against, "winning percentage" (actually the proportion of wins), wins, losses, and games played. It is well known to baseball statistics enthusiasts that, over the course of a season, the percentage of wins can be predicted quite well from the the average number of runs scored for and average number of runs scored against. A question of interest is whether, and by how much, reducing these two variables to one hurts the ability to predict the number of wins.

   Here is what I want you to do. Load in the file *baseball.txt*. Create a new variable called `pct <- wins/games`, and a new variable called `diff <- runs.for - runs.against`. Now, do the following:

   (a) Predict `pct` from `diff` using ordinary linear regression. Save the fit as `lm.fit.1`.

   (b) Predict `pct` from `diff` using logistic regression via the `glm` module. Since your outcome variable is a proportion, you will need to input the number of binomial trials (i.e., `games`) as `weights` as an option to the `glm()` function. Save the fit object as `logistic.fit.1`.

   (c) Predict `wins` from `diff` using Poisson regression. Be sure to input `offset = log(games)` since the number of games varies slightly across teams. Save the fit object as `poisson.fit.1`.

   (d) Use the `predict` function to get predictions of `pct` for the linear and logistic regression models. Call them `pct.hat.linear` and `pct.hat.logistic`. (Hint. *Remember* that logistic and Poisson regression do not directly predict the outcome variable you give them. They each produce predicted values that need to be reverse transformed to give you the predicted value of the outcome variable.) Once you have transformed the logistic output to the percentage metric, multiply the predicted percentage by `games` (which, remember, will vary slightly by team) to get a predicted number of wins. Call these `wins.hat.linear` and `wins.hat.logistic`. Use `predict()` to obtain `wins.hat.poisson` directly (after the appropriate exponential transformation) from the Poisson fit. Now you have predicted wins from 3 different models.

(e) There is a famous formula in baseball statistics. It is called the *Pythagorean Formula* and is used to predict the number of wins from runs scored and runs allowed. The formula is

$$\text{pct} = \frac{\text{runs.for}^{1.82}}{\text{runs.for}^{1.82} + \text{runs.against}^{1.82}} \tag{1}$$

Use the Pythagorean formula to predict the number of wins for the teams. Call this `wins.hat.pythagorean`. Compute errors for each of the 4 models, and display plots of the predicted versus error values. Which of the 4 models has the smallest sum of squared errors? Which has the largest? Do the models all work well? Does that surprise you? Why, or why not?

(f) *Cross-Validation.* The file *baseball2.txt* contains data for the 1954, 1998, and 2001 American League seasons. Use your 4 models (and the estimated coefficients) obtained from the previous fitting exercise to predict winning percentage for these 3 seasons. (Remember, do not obtain new constants by fitting the new data. Use the coefficients from the old data!) Which models work best? Where are the biggest errors?

(g) According to some other very careful analysis results, it is believed that, by adding Joe Mauer to your team to replace your current catcher Kris Kathan, you will increase your team's runs-scored by an average of .32 runs a game from the current 4.77, while improving your defense (reducing runs allowed) by .12 runs per game from the current 4.55. Your team won 84 and lost 78 games this past year. How many more games do you estimate you win with Joe?

3. *Poisson Regression.* The file *risky.behavior.dta* is a Stata file containing data on a randomized trial targeting couples at high risk of HIV infection. This file was provided with the Gelman text, and, unfortunately, little information was provided about it, so some of my characterization of it is based on guesswork. The participants were 217 heterosexual couples. The female members of each couple are the first 217 observations, and the male members of each couple are observations 218–434. We will use only the data for the females. Included in the file are the following variables:

   - `sex` of the participant

- `couples` An indicator variable for the experimental condition in which both partners received training
- `women_alone` An indicator variable for the experimental condition in which only the female partner received training. A 0 on both `couples` and `women_alone` indicates that the participant was in the control condition
- `bs_hiv` The HIV status of the participant prior to the study
- `bupacts` A measure of risky sexual behaviors of the participant prior to the study
- `fupacts` A measure of risky sexual behaviors *after* the treatment—This is the outcome variable

Load in the file, using the `read.dta` command. *Select only the female data for analysis.*

(a) Model the outcome as a function of the treatment variable(s), using Poisson regression. Does the model fit well? Is there evidence of overdispersion?

(b) Now add the pretreatment measures. Does the model fit well? Is there evidence of overdispersion?

(c) Fit an overdispersed Poisson model. Does the treatment intervention appear to be effective? How effective?

4. *Continuous Probability Simulation.* The logarithm of weights in pounds of men in the US are approximately normally distributed with mean 5.13 and standard deviation 0.17. For women the numbers are 4.96 and 0.20. Suppose 10 adults selected at random from a population that is 51.38% women step on an elevator with a capacity of 1750 pounds. What is the probability that the elevator cable breaks?

5. *Causal Inference.* Do exercise 4, p. 195 of Gelman and Hill. Remember, treat the data in the table as the entire population, not as a sample. This means you can calculate conditional distributions exactly.

6. *Propensity Scores.* Do exercise 1, p. 231–232 of Gelman and Hill. The Lalonde(1986) and Dehejia and Wahba(1999) papers are available on JSTOR if you want to look at them. The `lalonde` data file is available in the online data repository at Gelman's website. I have placed the zip file containing all the data for all the Gelman on the course website as well, in case something goes wrong at one of the sites.

7. *Regression Discontinuity Analysis.* Do exercise 2, p. 232 of Gelman and Hill. The data are available in the Gelman zip file.

8. *Instrumental Variables.* Do exercise 3, p. 233 of Gelman and Hill.

9. *2-Level Multilevel Model.* In class, we discussed 2-Level multilevel modeling employing both HLM and R's `lmer` module. Attached are 16 pages from a recent article by J.K. Holt. The "wide format" data for this example are in a file called *reading.all.txt*. You can also find the level 1 and level 2 files as reading.level.1.txt and reading.level.2.txt.

   Data are from the kindergarten cohort of the Early Childhood Longitudinal Survey (ECLS-K) and include repeated observations of students from 1998 through 2002 collected in the fall and spring of kindergarten, fall and spring of 1st grade, and the spring of 3rd grade (National Center for Education Statistics, 2004).

   Level 1 File:

   Variables

   CHILDID - Child identification number

   READING - Reading IRT scale score

   SCHCHG - Whether student changed schools between waves

   0 = did not change schools

   1 = did change schools

   TERM_CTR - Term centered at the fall of 1st grade

   TERM_CTR_SQ - Term centered at the fall of 1st grade squared

   Level 2 File:

   Variables

   CHILDID - Child identification number

   GENDER

   0 = Female

   1 = Male

   Table 4.1 shows 4 models fit by Holt to the data. Work through the text, in which she describes her models and accompanying rationale. Set up models 1,2,3,4 and compare your results with Holt's. You will not get the same results, but you should come close. Construct a table

in the same format as hers, giving your results. Don't bother with the asterisks.

(Hint: You might want to check out your analyses with HLM, but if you only have the student version, the 16400+ observations will present a problem. However, you might still wish to use HLM to help generate the mixed model, and there is a simple way – "pare down" the data file by taking a subset of the data at level 2, then grabbing the level 1 observations that match those IDs, and performing your analysis on those data.)