36-617: Applied Linear Models
Fall 2020
HW08 – Due Mon Nov 2, 11:59pm

- Please turn the homework in to Gradescope using the appropriate link in our course webspace at canvas.cmu.edu, under Assignments.
- Next week we will discuss causal inference (causal-inference.pdf) and hopefully begin discussing generalized least squares (Sheather Ch 9). There will be a quiz on Monday, but it will only cover material on causal inference.

Exercises

1. The `warpbreaks` data set included in R gives the results of an experiment to determine the the effect of wool type (A or B) and tension (low, medium or high) on the number of warp breaks per loom. Data was collected for nine looms for each combination of settings. You can get more information on this data set from `help(warpbreaks)`. Use `View(warpbreaks)` and `xtabs(breaks ~ wool + tension, data=warpbreaks)` to familiarize yourself with the data. (You don't have to turn in any of this initial exploration.)

   (a) Fit the two Poisson regression models `breaks ~ wool + tension` and `breaks ~ wool * tension` (dont't forget `family=poisson`!). Consider `summary()`'s, `plot()`'s of the fitted glm's, `mmplot()`'s and also `binnedplot()`'s (from `library(arm)`) and/or plots from `DHARMa`, and any other plots you think are useful. Comment on the fits of each model, using the graphical evidence you have obtained *(if you think some plot(s) are* **not** *useful for this, please tell which ones & why)*.

   (b) We would like to know whether the interaction between `wool` and `tension` is needed.

      i. Compare AIC's for the two models, and then compare BIC's for the two models. Do AIC and BIC choose the same model?

      ii. Conduct a likelihood ratio test for including the interaction. Does the LR test choose the same model as either AIC or BIC? *(Hint/recipe: For the likelihood ratio test, use the* `logLik()` *function to obtain the log-likelihood of the two models, compute* -2 *times the difference between the null and alternative model log-liklihoods, and obtain a p-value from the tail of an approporiate* χ^2 *distribution.)*

   (c) For the best model from part (b), we would like to know if overdispersion is a problem. Use the Pearson residuals to conduct a test of overdispersion (hint: follow the recipe from lecture #17.1). Does this agree with the conclusion of `testDispersion()` from `library(DHARMa)`?

   (d) Refit the model from part (c) using `family=quasipoisson`, and note any differences in overdispersion parameter, coefficient estimates, SE's, etc. Are there any differences in which predictors are significant?

[More exercises on the next page]

2. [Based on some exercises in Gelman & Hill.] The data set `nes5200_processed_voters_realideo.dta` has data from the National Election Survey in every congressional election year in the US from 1948 to 2002. You can read it into R using the function `read.dta()` from `library(foreign)`. We are interested in learning whether some of the survey data is useful in predicting respondents' vote for president.

   (a) Extract from the full NES data set above a data frame with just data from the presidential election year 1976, and the following variables: `ideo7, black, female, rep_pres_intent, presapprov, age, educ1, urban, income, union` and `perfin1`. The variable `rep_pres_intent` is 1 if the respondent intended to vote for the Republican candidate (Gerald Ford, who assumed the presidency when Richard Nixon resigned amid an impeachment investigation), and 0 if the respondent intended to vote for the Democratic candidate (Jimmy Carter). Familiarize yourself with all of the variables by using `summary()` on your 1976 data frame[1] to look at the categories. You don't have to turn anything in for this, but you should do it.

   (b) Fit a logistic regression predicting `rep_pres_intent` from the other variables, for 1976. Assess the fit using appropriate graphical methods. Interpret the results: which variables seem to matter for predicting voters' preference for president? How do they affect the odds of voting for the Republican candidate?

   (c) Make a new data frame by converting all of the variables in the 1976 data frame to numeric variables, using `as.numeric()`. Repeat part (2b), using this new data frame.

   (d) Compare the results from parts (2b) and (2c). How are they consistent? Inconsistent? In this case, would you recommend converting all of the variables to numeric?

   (e) Repeat parts (2a) through (2d) for the presidential election year 2000. In this year, the democratic candidate was Al Gore, who served as vice-president to Bill Clinton, who was impeached but not removed from office. The Republican candidate was George Bush.

   (f) Finally, compare your results from 1976 and 2000. Interpret any similarities or differences in terms of voter preferences.

   Of course there are very many other variables, as well as transformations and interactions, to explore in this data set, and the models you fitted above are not "optimal" in any sense. However I hope you see that you can still get interesting signal out of the data, even with non-optimal models.

[1] A command like `attr(nes_data,"var.labels")` on the full data set will give you brief definitions of all of the variables.