36-617: Applied Linear Models
Fall 2020
HW07 – Due Mon Oct 25, 11:59pm

- Please turn the homework in online to gradescope using the link on the assignment page in canvas.
- You will need three data files for this assignment:
  - MissAmerica2008-binomial.txt
  - MissAmerica2008-bernoulli.txt
  - nes5200_processed_voters_realideo.dta

  All three are in the folder for this assignment on Canvas.
- You will also need to install three R packages for this assignment:
  - `marginalmodelplots`. We will use the `mmplots()` function from `library(marginalmodelplots)` rather than the `mmps()` function from `car`, since `mmps()` doesn't work right for glm's. To install it:
    1. download `marginalmodelplots_0.4.2.tar.gz` from canvas (a copy is in the directory for this hw assignment)
    2. Instal with these commands:
       ```
       install.packages("locfit")
       install.packages("marginalmodelplots_0.4.2.tar.gz", repos=NULL)
       ```
  - `DHARMa`. This provides an alternative set of residuals for all lm's and glm's based on a simulation method called the "parametric bootstrap". Install from `cran` as usual. Here is an example code that uses `DHARMa`:

    ```
    ## Simulate some data
    expit <- function(x) {exp(x)/(1+exp(x))}
    x <- rnorm(100)
    p <- expit(2 -3*x)
    y <- rbinom(100,1,p)
    ## fit a logistic glm
    mymodel <- glm(y ~ x, family=binomial)
    ## use DHARMa to check residuals
    simulationOutput <- simulateResiduals(fittedModel = mymodel, plot = T)
    ```

    This produces a qq plot on the left and a residuals vs fitted plot on the right.
    - If the model fits the data, DHARMa produces uniformly distributed residuals, instead of normally distributed residuals. You can use the qq plot to check for how close to uniform the residuals are, check for skewing, overdispersion, etc., just like the normal qq plot from the usual residual diagnostic plots in R.
    - The plot on the right is a residuals vs fitted plot that you can interpret much as for ordinary regression. The three guidelines help you assess nonconstant varianace. If the lines are horizontal, the variance is consistent with the model (e.g., constant variance for ordinary regression; variance proportional to $p_i(1-p_i)$ for logistic regression). Residuals that are outliers will be colored red in the plot.

    There is much more you can do with DHARMa; if you are curious, have a look at
    `https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html`
  - `arm`. This is the library that goes with the Gelman and Hill text, and provides the t binnedplot() function, which lets you look at local averages of residuals, as in ther lecture notes.

Exercises

1. Please do Sheather, Ch 8, pp 296ff, problem #2, using the data set `MissAmerica2008-binomial.txt`. Notes:

   - Because the data are grouped into binomial data (# of successes in the 9 years from 2000 to 2008) instead of bernoulli/binary data, instead of fitting models using code like this:

     ```
     glm.1 <- glm( y ~ x1 + x2 + x3 + x4 + x5, family=binomial)
     ```

     you will need to fit the models using code like this:

     ```
     glm.1 <- glm(cbind(y,9-y) ~ x1 + x2 + x3 + x4 + x5, family=binomial)
     ```

     You should use the data set `MissAmerica2008-binomial.txt` for this problem.
   - Sheather suggests (among other things) to use marginal model plots to check for missing variables, transformations, etc. Use the `mmplot()` function for this purpose.
   - Sheather also asks you to check for high leverage points, and the "bad" (i.e. influential points among the high leverage points). For this you will need to examine one of the plots from the usual residual diagnostic plots `plot(my.glm.fit)`. To examine the quality of the residuals, please do both of the following:
     - Examine binned residuals using `binnedplot()` from `arm`.
     - Examine simulated residuals using `simulateResiduals()` from `DHARMa`.

2. Now consider the data set `MissAmerica2008-bernoulli.txt`.

   (a) Examine the individual rows in the two data sets (`MissAmerica2008-binomial.txt` and `MissAmerica2008-bernoulli.txt`). What is the relationship between the two data sets?

   (b) Fit your final model from problem #1 to the `MissAmerica2008-bernoulli.txt`. This will require a modification of your response variable. Explain why you might expect equally good fit whether you are using `MissAmerica2008-binomial.txt` or `MissAmerica2008-bernoulli.txt`.

   (c) Compare AIC for the model fitted to `MissAmerica2008-binomial.txt` with AIC for the model fitted to `MissAmerica2008-bernoulli.txt`. Are they the same? Explain, as carefully as you can, why or why not.

3. 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?

4. [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 (4b), using this new data frame.

(d) Compare the results from parts (4b) and (4c). How are they consistent? Inconsistent? In this case, would you recommend converting all of the variables to numeric?

(e) NOT TO TURN IN. If you have time, it is interesting to repeat parts (4a) through (4d) 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. How are the results from 1976 and 2000 similar? How are they different?

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.