36-617: Applied Linear Models
Fall 2020
HW07 – Due Mon Oct 26, 11:59pm Pgh Time

- Please turn the homework in online to gradescope using the link on the assignment page in canvas.
- There is no quiz this week!
- You will need two data files for this assignment:
  - MissAmerica2008-binomial.txt
  - MissAmerica2008-bernoulli.txt

  Both 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. Notes:

   - Because the data are grouped into binomial data (# of successes in the 9 years from 2000 to 2008) instead of bernoilli/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 probem.

   - Sheather suggests (among other things) to use marginal model plots to check for missing variables, transformations, etc. Use the `mmplots()` 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.