# Math 4392 HW2

Washington University, SP 2024

**Instruction**: Please type up your assignment neatly including R code, relevant console output, graphs and written portion in a single document. You are encouraged to use Rmarkdown, but other format is acceptable as long as the report demonstrates your work clearly and coherently.

If possible, submit the resulting file in `pdf` format due to limitation of the grading comment feature on Canvas. Raw `.Rmd` or `.R` files are not accepted in case of trouble compiling.

Please note - there is always some freedom in deciding which methods to use, in what order to apply them, and how to interpret the results. So there may not be one clear right answer and good analysts may come up with different models.

This assignment covers approximately chapter 2 and 3 of the Faraway textbook

**Due Date**: Sunday, Feb 25, 11:59 pm on Canvas

1.[4 pts] A study was conducted on children who had corrective spinal surgery. We are interested in factors that might result in kyphosis (a kind of deformation) after surgery. The data can be loaded by `data(kyphosis,package="rpart")`

a) Make plots of the response as it relates to each of the three predictors. You may find a jittered scatterplot more effective than the interleaved histogram for a dataset of this size. Comment on how the predictors appear to be related to the response.

b) Fit a GLM with the kyphosis indicator as the response and the other three variables as predictors. Plot the deviance residuals against the fitted values. What can be concluded from this plot?

c) Produce a binned residual plot as described in the text. You will need to select an appropriate amount of binning. Comment on the plot.

d) Plot the residuals against the Start predictor, using binning as appropriate. Comment on the plot.

e) Produce a normal QQ plot for the residuals. Interpret the plot.

f) Make a plot of the leverages. Interpret the plot.

g) Check the goodness of fit for this model. Create a plot like Figure 2.9. Compute the Hosmer-Lemeshow statistic and associated p-value. What do you conclude?

h) Use the model to classify the subjects into predicted outcomes using a 0.5 cutoff. Produce cross-tabulation of these predicted outcomes with the actual outcomes. When kyphosis is actually present, what is the probability that this model would predict a present outcome? What is the name for this characteristic of the test?

2.[3 pts] A biologist analyzed an experiment to determine the effect of moisture content on seed germination. Eight boxes of 100 seeds each were treated with the same moisture level. Four boxes were covered and four left uncovered. The process was repeated at six different moisture levels.

a) Plot the germination percentage against the moisture level on two side-by-side plots according to the coverage of the box. What relationship do you see?

b) Create a new factor describing the box (the data are ordered in blocks of 6 observations per box). Add lines to your previous plot that connect observations from the same box. Is there an indication of a box effect?

c) Fit a binomial response model including the coverage, box and moisture predictors. Use the plots to determine an appropriate choice of model.

d) Test for the significance of a box effect in your model.Repeat the same test but using the Pearson's Chi-squared statistic instead of the deviance.

e) At what value of moisture does the predicted maximum germination occur for noncovered boxes? For covered boxes?

f) Produce a plot of the residuals against the fitted values and interpret.

3.[3 pts] This problem concerns the modeling of the quantitative structure-activity relationships (QSAR) of the inhibition of dihydrofolate reductase (DHFR) by pyrimidines. We want to relate the physicochemical and/or structural properties as exhibited by the 26 predictors in pyrimidines with an activity level. We have structural information on 74 2,4-diamino- 5-(substituted benzyl) pyrimidines used as inhibitors of DHFR in E. coli. All the variables lie in [0,1].

a) Plot the activity(response) against the first three predictors. Are any outlier sin the response apparent? Remove any such cases.

b) Fit a Gaussian linear model for the response with all 26 predictors. How well does this model fit the data in terms of R2? Plot the residuals against the fitted values. Is there any evidence of a violation of the standard assumptions?

c) Fit a quasi-binomial model for the activity response. Compare the predicted values for this model to those for the Gaussian linear model. Take care to compute the predicted values in the appropriate scale. Compare the fitted coefficients between the two models. Are there any substantial differences?

d) Fit a Gaussian linear model with the logit transformation applied to the response. Compare the coefficients of this model with the quasi-binomial model.

e) Fit a Beta regression model. Compare the coefficients of this model with that of logit response regression model.

f) What property of the response leads to the similarity of the models considered thus far in this question?