# APSTA 2022 Homework 5 GLM

This homework is due April 24, 5pm. Please submit the homework in pdf format online. A latex pdf document works the best to incorporate figures and formulas. If you work with Word, please convert the .doc file into a .pdf file. If you need to write the math formulas by hand, you can scan the work and save it into a pdf file. Any hand-writings must be legible, or they will not be graded.

1. Show that the Poisson distribution belongs to the Exponential family. For $Y \sim \text{Poisson}(\lambda)$, identify $\theta$, $b(\theta)$ and $c(Y_i, \phi)$. Derive mean and variance of $Y$ in terms of $\theta$, and further specify the variance as a function of the mean parameter and $\phi$. [1pt]

2. Derive the Maximum Likelihood Estimator of a Poisson Regression: $Y_i \sim \text{Poisson}(\lambda_i)$, $\log(\lambda_i) = X_i^\top \beta$, $i = 1, \ldots, n$, assuming common offset value for each observation. [2pts]

    - Find sufficient statistic for $\beta$.
    - Show that $\log(\lambda)$ is a canonical link.
    - Derive the score vector and the Hessian matrix, and express them in terms of $X$, $y$ and the $W$ matrix.
    - Derive the Fisher Scoring Algorithm to obtain $\hat{\beta}_{\text{MLE}}$.

3. In this part, you will use a dataset from Long(1990) on the number of publications produced by Ph.D. biochemists. There are following variables in the dataset: [4pts]

    - art: articles in last three years of Ph.D.
    - fem: female=1, male=0
    - mar: coded one if married
    - kid5: number of children under age six

- phd: prestige score of Ph.D. program
- ment: articles by mentor in three years

The focus of this question is on how to fit a Poisson model and interpret the coefficients (rather than selecting the best model):

- First fit a Poisson regression model that include all the available covariates (additive effects only) in `R` using the `glm` command, with `log` link function. Report the results of the model, and interpret the effects of the covariates in terms of Risk Ratio. Report 95% confident interval for the Risk Ratio and comment on their statistical significance at level 5%.
- Fit the model using the Fisher Scoring algorithm that you developed in the previous question in `R`.
  - You can generate the design matrix $X$ using `model.matrix` function.
  - The initial values can be found running OLS of $Y$ on $X$.
- Estimate the observed information matrix, and explain briefly that under Poisson regression, this is also the Information matrix.
- Estimate the standard errors for $\hat{\beta}_{\text{MLE}}$.
- Reproduce other statistics (such as Deviance) from the `glm` output.

4. In this part, you are asked to improve the Logistic regression model on children's attendance in religious school that we studied in class. Use dataset religion1.csv. [3pts+1bonus pt]

   - The covariates are: sex (1=female, 0=male); edu (1=college or above; 0=below college); age, agesq(age squared); income level (in $10,000 brackets);attend (frequency of attending religious services); race(1=white,0=nonwhite); Married (Yes and No)

   Pay particular attention to the following issues: [3pts]

   - The candidate models include all additive models and two-way interactions between any two pair of variables.
     - Use the `formula` command and write a loop to generate all the possible additive models. (See Lab 10 example)

  - For variables ~~education level and~~ income in $10,000 brackets, think carefully how you can incorporate those variables respecting the parsimony principle. Hint: You can compare the models between treating income as factors and treating income as linear predictors under an additive model with all possible covariates. You can also try other ways of dealing with income in the model
  - Based on the additive model(s) you select, build in two-way interactions.

- For each model, record the log-Likelihood, deviance, df, AIC, BIC, pseudo-$R^2$, classification rate, sensitivity and specificity.
- Choose the best model and report the results, make sure you interpret the effects of the covariates in terms of odds ratio. (The benchmark model has AIC=369.16, BIC=395.41, to get full credit, your best model should be comparable to this performance)
- For the model of your choice, conduct an ROC analysis and plot the ROC curve, estimate the area under the curve using a stepwise function approximation.
- **Bonus** Best model competition: Try to build a model that outperforms the benchmark model. The student who builds the best model gets a extra point. For any other models that outperform the benchmark model, you will get bonus point that is proportional to the improvement you accomplish relative to the best model.