

## Stat 201 - Generalized Linear Models

### Mandatory Assignment - due on November 20th 2021

- ~~1~~ The Chinese Longitudinal Health and Longevity Survey (CLHLS) includes health-related information of a sample of Chinese subjects collected during an interview: age at the time of the interview (in months); gender (1= female); type of residence (1=rural or 0=urban); whether the subject is sedentary or active (active = 1); limits in activities of daily living (ADL; six activities including bathing, dressing, eating, indoor, transferring, toileting and continence,;  $adl = 0$  if no adl limits, 1 if one adl limit, 2 if two adl limits or more); the number of correct answers in a 23-item MMSE (Mini Mental State Examination) questionnaire.

The data are stored in the file `mmse.txt`: for this exercise, use `mmse` as the response variable and ignore the variables `duration` and `status`.

- ~~(a)~~ Assuming that all the MMSE items have the same probability of being correctly answered, fit a logistic regression model (without interactions) to estimate the influence of the available covariates on the probability of a correct answer. Display and interpret the output.
- ~~(b)~~ Using the model computed above, provide a picture that shows the probability of answering an MMSE item as a function of age varying between 80 and 100 years, for a male with urban residence, sedentary lifestyle and no ADL limits (note: age is in months in `mmse.txt`)
- ~~(c)~~ By comparing two appropriate models through a deviance statistic, test whether the influence of age on the probability of a correct item varies between males and females. Do females perform better or worse than males as age increases?

- ~~2~~ The data stored in the file `penalty.txt` include a cross-sectional study of 326 defendants in homicide indictments in Florida, during the period 1976-1977. Data are clustered according to the final verdict (death penalty or no death penalty), and the race of both the victim and the defendant.

- ~~(a)~~ Test the independence of the three variables, by fitting an appropriate log-linear model.
- ~~(b)~~ Fit a battery of log-linear models, by including one pairwise interaction at the time. What is the best model? Interpret the output of the proposed model.
- ~~(c)~~ Consider now all the possible unsaturated models. Can you find a model that improves the model found in the previous item?

- ~~3~~ Consider the `mmse.txt` again, but now focus on `duration` as the response variable that describes the exit time (in months) of the subjects *after the interview*, while `status` indicates whether the exit is due to death ( $status = 1$ ) or the subject is still alive ( $status = 0$ ).

- ~~(a)~~ Compute the Kaplan-Meier estimate of the survival function of the males with urban residence, sedentary lifestyle and no ADL limits
- ~~(b)~~ Using the K-M estimate of the previous item, obtain the median survival time
- ~~(c)~~ Fit a Gompertz proportional hazards model and interpret the results (hint: do not include age as a covariate)
- ~~(d)~~ Fit a AFT Weibull model and interpret the results (hint: do not include age as a covariate)
- ~~(e)~~ Using the model proposed at the previous item, compute the acceleration factor of a male with urban residence, sedentary lifestyle and no ADL limits, scoring an MMSE of 20