Model Card - GLM

Model Details

- GLM - Generalized Linear Model
- Developed by John Nelder and R.W.M. Wedderbrun in 1972
- A flexible generalization of ordinary linear regression

Form

$$y_i \sim N(x_i^T \beta, \sigma^2)$$

- y_i (Response Variable) is assumed to follow exponential family distribution with mean $\mu_i = x_i^T \beta$
- x_i contains known covariates
- β contains the coefficients to be estimated
- Fit by least squares and weighted least squares
  - Using SAS's GLM procedure or R's $lm()$ function

Intended Use

GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function. No linear relationship is assumed between the response variable and explanatory variables.

- Binary Logistic Regression (Target is binary)
  - $logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$
- Poisson Regression (For modelling events whose outcomes are counts)
  - $\log(\lambda_i) = \beta_0 + \beta x_i$
- Advantages over traditional OLS regression
  - No need to transform the response to have a normal distribution
  - Choice of link function is separate from the choice of random component (more flexibility in modeling)
  - Models fitted via MLE, likelihood functions and parameter estimate benefit from asymptotic normal and chi-square distributions

Factors

- Random Component
  - Specifies the probability distribution of the response variable
    - Normal Distribution in ordinary linear regression model
    - Binomial Distribution in binary logistic regression model
- Systematic Component
  - Specifies the explanatory variables $(x_1, x_2, \ldots, x_k)$ in the model
- Link Function η or $g(\mu)$
  - Specifies the link between the random and the systematic components
  - Indicates how the expected value of the response variable relates to the linear combination of explanatory variables
  - $\eta = g\big(E(Y_i)\big) = E(Y_i)$ for classical regression
  - $\eta = \log\left(\frac{\pi}{1-\pi}\right) = logit(\pi)$ for logistic regression

Caveats and Recommendations

- The random component (response variable) doesn't not have a separate error term
- Dependent variable Y_i does not need to be normally distributed, typically assumes a distribution from an exponential family (e.g. Binomial, Poisson, Multinomial, Normal, etc.)
- Assumes data Y_i are independently distributed
- Does not assume a linear relationship between the response variable and explanatory variables
- Assume a Linear relationship between the transformed expected response (in terms of the link function) and the explanatory variables (e.g. binary logistic regression $logit(\pi) = \beta_0 + \beta_1 x$)
- Errors are independent (not normally distributed)
- Parameter estimation with MLE