Generelized Linear Model

1 Introduction/Intuition

So, uptill now we were dealing with linear model where we are assuming that response variable follows Gaussian's distribution, thus also it should be always continuous but what if it's discrete in nature ? How will you deal with such kind of models ? This is where we generalize linear model for the inclusion of these situations. Let's see.

Ordinary linear regression predicts the expected value of a given unknown quantity (the response variable, a random variable) as a linear combination of a set of observed values (predictors). This implies that a constant change in a predictor leads to a constant change in the response variable (i.e. a linear-response model). This is appropriate when the response variable can vary, to a good approximation, indefinitely in either direction, or more generally for any quantity that only varies by a relatively small amount compared to the variation in the predictive variables, e.g. human heights.

However, these assumptions are inappropriate for some types of response variables. For example, in cases where the response variable is expected to be always positive and varying over a wide range, constant input changes lead to geometrically (i.e. exponentially) varying, rather than constantly varying, output changes.

2 Theory

So, in linear model we have $\mathbf{X}\vec{\beta} = E(\mathbf{Y}/\mathbf{X})$ i.e., Expected value of response variable is a linear combination of given independent variables.
But now for Generalized linear models we will assume a function *'g'* which is known as **Link function**. So $g(\mathbf{X}\vec{\beta}) = E(\mathbf{Y}/\mathbf{X})$ i.e., A function of expected value of response variable is equal to the linear combination of independent variables.
Also, we are assuming that this response variable is coming from a distribution which is essentially exponential family of distribution. Let's say it as function *'F'*.
Thus,

$$\mathbf{Y} \sim F(g(\mathbf{X}\vec{\beta}))$$

There is always a well-defined canonical link function which is derived from the exponential of the response's density function. However, in some cases it makes sense to try to match the domain of the link function to the range of the distribution function's mean.

3 Model Components

- Response variable i.e., $\mathbf{Y}$.
- A particular distribution for modeling $\mathbf{Y}$ from among those which are considered exponential families of probability distributions i.e., *'F'*.
- A linear predictor $\mathbf{X}\vec{\beta}$ such that $(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-1}$ exists.
- A link function *'g'* such that $g(\mathbf{X}\vec{\beta}) = E(\mathbf{Y}/\mathbf{X})$.

4 Table for GLM

<table>
<tr><th colspan="4">Common distributions and canonical link functions</th></tr>
<tr><th>Distribution
'F'</th><th>Support of distribution</th><th>Link function
'g'</th><th>Link name</th></tr>
<tr><td>Normal</td><td>$(-\infty, +\infty)$</td><td>$\mathbf{X}\vec{\beta} = \mu$</td><td>Identity</td></tr>
<tr><td>Exponential</td><td rowspan="2">$(0, +\infty)$</td><td rowspan="2">$\mathbf{X}\vec{\beta} = -\mu^{-1}$</td><td rowspan="2">Negative Inverse</td></tr>
<tr><td>Gamma</td></tr>
<tr><td>Inverse Gaussian</td><td>$(0, +\infty)$</td><td>$\mathbf{X}\vec{\beta} = -\mu^{-2}$</td><td>Inverse Squared</td></tr>
<tr><td>Poisson</td><td>$\mathbf{W} : 0, 1, 2, 3, ...$</td><td>$\mathbf{X}\vec{\beta} = ln(\mu)$</td><td>Logarithm</td></tr>
<tr><td>Bernoulli</td><td>$\{0, 1\}$</td><td rowspan="3">$\mathbf{X}\vec{\beta} = ln(\frac{\mu}{1-\mu})$</td><td rowspan="4">Logit</td></tr>
<tr><td>Categorical</td><td>integer in $[0, N)$ K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1</td></tr>
<tr><td>Multinomial</td><td>K-vector integer:$[0, N]$</td></tr>
<tr><td>Binomial</td><td>all integer in : $[0, N]$</td><td>$\mathbf{X}\vec{\beta} = ln(\frac{\mu}{n-\mu})$</td></tr>
</table>

For example,
We can consider death of rats using a particular dose of the chemical, we want check what is the relation between death of rats with this chemical. So, here it's clear that death will follow Bernoulli's distribution and hence doses of that chemical will be connected with death of rats using link function i.e. logit. So, we can use logistic regression for predicting death of rats under the influence of concentration of dose of the chemical.