6: Generalized Linear Models

```
$ echo "Data Science Institute"
```

What happens if we are faced with a situation where the response Y is neither quantitative or qualitative?

Motivation

We have learned about linear and logistic regression which are generalized linear models (GLM). But exactly is GLM? Let's explore!

What is Generalized Linear Model?

The linear predictor η is a linear combination of the predictors X_1, X_2 ... X_p:

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots . \beta_p X_p$$

where $\beta_0, \beta_1 \ldots \beta_p$ are the coefficients to be estimated.

The link function $g(.)$ connects the mean of the response variable μ (which is $E(Y)$) to the linear predictor η

$$g(\mu) = \eta$$

This can be written as

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots . \beta_p X_p$$

Examples of GLM

Different choices of the link function g(•) and the distribution of Y give rise to different models. Here are three examples of GLM:

- Linear Regression with identity link, i.e $g(\mu) = \mu$
- Logistic Regression with link $g(\mu) = log(\frac{\mu}{1-\mu})$
- Poisson Regression with link $g(\mu) = log(\mu)$

`Bikeshare` dataset overview

- The response is `bikers`, the number of hourly users of a bike sharing program in Washington, DC.
- This response value is neither qualitative nor quantitative: it is *counts*. We will consider predicting `bikers` using the predictors `mnth` (month of the year), `hr` (hour of the day, from 0 to 23), `workingday` (an indicator variable that equals 1 if it is neither a weekend nor a holiday), `temp` (the normalized temperature in Celsius), and `weathersit` (a qualitative variable that takes on one of four possible values: clear; misty or cloudy; light rain or light snow; or heavy rain or heavy snow.)

Poisson Distribution

Suppose that a random variable Y takes on nonnegative integer values, i.e. Y=0,1,2,... If Y follows the Poisson distribution then

$$Pr(Y=k)=\frac{e^{-\lambda}\lambda^k}{k!}\text{for k}=0,1,2\ldots$$

Here $\lambda > 0$ is the expected value of Y, i.e $E(Y)$. It turns out that λ also equals the variance of Y. This means that if Y follows the Poisson distribution, then the larger the mean of Y , the larger its variance.

Note: $k! = k*(k-1)*(k-2)\ldots*3*2*1$

Poisson Regression

$$log(\lambda(X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

Note: Taking the log ensures that λ can only be non-negative.

This is equivalent to representing the mean λ as follows:
$\lambda = \mathrm{E}(Y) = \lambda(X_1, \ldots, X_p) = e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}$

Interpreting the Coefficients

Each coefficient β_i can be interpreted as the expected change in the *log count* for a one-unit change in the predictor X_i, holding all other variables constant.

Breakout Room

What are some advantages of Poisson Regression over Linear Regression?

Exercise: Linear and Poisson Regression on Bikeshare data

References

Chapter 4 of the ISLP book:

James, Gareth, et al. "Classification." An Introduction to Statistical Learning: with Applications in Python, Springer, 2023.