GLM

Divya Sharma (ds655)

Overview

GLMs (Generalized Linear Models) are a flexible extension of linear regression models. They allow the response variable to have a distribution other than a normal distribution, such as binomial, Poisson, etc.

Like linear regression, GLMs relate the response variable to explanatory variables using a linear function. But they also include a *link function* that transforms the response variable so its distribution matches the assumptions.

The key components of a GLM are:

- The response variable Y and its distribution (e.g. binomial, Poisson)
- The linear predictor - a linear function of the explanatory variables (like in linear regression)
- The link function g() that transforms the expected value of Y to match the linear predictor

The link function connects the linear predictor to the expected response, allowing a non-normal distribution for Y.

For example, a GLM could model student test scores based on hours studied and previous GPA. The response variable test score may follow a Poisson distribution. The linear predictor combines the explanatory variables hours studied and GPA. The log link function transforms the expected test score to match this linear combination.

A multinomial GLM handles a categorical response variable with more than 2 categories. The purpose is to model the probability of different outcomes. For example, a multinomial GLM could predict political party affiliation based on age, income, gender, etc. The response is party, a categorical variable with multiple outcomes. The linear predictor combines the explanatory variables. The logit link connects this to the probability of each party.

Some potential research questions for a GLM model can be:

- How does income level correlate with Republican vs Democrat affiliation?
- Do gender and age interact in predicting party?
- What variables most strongly predict party affiliation?

In summary, a multinomial GLM can model a categorical response to explain the factors influencing different outcomes. The link function handles the non-normal distribution.

Probability Distribution

Briefly describe the probability distribution that is assumed for the outcome. What is the support? What are the parameters and what values can they take?

Model

Write out the general form of your GLM. What is the link function and why is it appropriate for that type of outcome? What are the model assumptions?

Data Example

- Introduce the dataset. Provide a few summary statistics and/or plots. Include the fact that this is a simulated dataset.
- Fit the model. Include all relevant code, including `library()` for packages if needed.
- Explain how to interpret coefficient estimates for the predictors.
- Show and describe a plot that illustrates the results of the model.
- Describe how to assess the model and include the code to do so. Include how to assess any assumptions that are unique to that model (e.g., proportional odds, overdispersion). *Note: For the ordinal model, assess the assumption using the method shown in the class exercise, not the hypothesis test shown in the videos.*