Title:

Name:

Institutional Affiliation:

Course:

**Generalized Linear Regression**

It is an advanced statistical modeling technique formulated by John Nelder and Robert Wedderburn in 1972. It encompasses many other models, which allows the response variable y to have an error distribution other than a normal distribution. The models include Linear Regression, Logistic Regression, and Poisson Regression. In a Linear Regression Model, the dependant variable ‘y’ is expressed as a linear function of all the independent variables 'x .'The underlying relationship between the response and the predictors is linear; we can visualize the relationship in a straight line.

Generalized linear regression models allow us to build a linear relationship between the response and predictors, even though their underlying relationship is not linear. We can use a link function, which links the response variable to a linear model. Unlike Linear Regression models, the error distribution of the response variable need not be normally distributed. The errors in the response variable are assumed to follow an exponential family of distribution (i.e., normal, binomial, Poisson, or gamma distributions) since we are trying to generalize a linear regression model that can also be applied in these cases, the name Generalized Linear Models.

General Linear Regression Model may be considered because there are circumstances where the linear regression model may not be suitable. Such cases are like when the relationship between X and y is not linear. There exists some non-linear relationship between them. For example, y increases exponentially as X increases. Also, when the variance of errors in y (commonly called Homoscedasticity in Linear Regression) is not constant and varies with X.

**Components of Generalized Linear Regression Model**

A GLM consists of three components:

* A random component,
* A systematic component, and
* A link function.

**Systematic Component/Linear Predictor:**

It is just the linear combination of the Predictors and the regression coefficients.

β0+β1X1+β2X2

**Link Function:**

Represented as η or g(μ), it specifies the link between a random and systematic components. It indicates how the expected/predicted value of the response relates to the linear combination of predictor variables.

**Random Component:**

It refers to the probability distribution, from the family of distributions, of the response variable.

The family of distributions, called an exponential family, includes normal distribution, binomial distribution, or poisson distribution. Normal distribution falls under the identity function, binomial function falls under logit or sigmoid function. Poisson distribution falls under log function.

**Assumptions of Generalized Linear Regression Model**

There are some basic assumptions for Generalized Linear Models as well. Most of the premises are similar to Linear Regression models, while some of the beliefs of Linear Regression are modified. The premises are as follows;

* The original response variable need not have a linear relationship with the independent variables, but the transformed response variable (through the link function) is linearly dependent on the independent variables.
* Homoscedasticity (that is, constant variance) need not be satisfied. Response variable Error variance can increase or decrease with the independent variables.
* The response variable y does not need to be normally distributed, but the distribution is from an exponential family (example; binomial, Poisson, multinomial, normal)
* Feature engineering on the Independent variable can be applied, i.e., instead of taking the original raw independent variables, variable transformation can be done, and the transformed independent variables, such as taking a log transformation, squaring the variables, reciprocal of the variables, can also be used to build the GLM model.

1. **Normally Distributed Data**

For normally distributed data, the response variable y does not need to be normally distributed, but the distribution is from an exponential family (e.g., binomial, Poisson, multinomial, normal)

The assumption of normality claims that the sampling distribution of the mean is normal or that the distribution of means across samples is normal. This should not be confused with the presumption that the values within a given piece are typically distributed or that the values within the population from which the sample was taken are average. Instead, the core element of this assumption is that the distribution of sample means (across independent samples) is usually distributed.

The Shapiro-Wilk test is a statistical test used to check if a continuous variable follows a normal distribution. The null hypothesis (H0) states that the variable is normally distributed, and the alternative theory (H1) states that the variable is NOT normally distributed.

So after running this test:

* If p ≤ 0.05, the null hypothesis can be rejected (i.e., the variable is NOT normally distributed).
* If p > 0.05, the null hypothesis cannot be rejected (i.e., the variable MAY BE customarily distributed).

Errors are independent but need not be normally distributed.

1. **Homoscedasticity, Constant Variance**

It is assumed that HomoscedasticityHomoscedasticity (i.e., constant variance) need not be satisfied. Response variable Error variance can increase or decrease with the independent variables.

Homoscedasticity refers to whether these residuals are equally distributed or whether they tend to bunch together at some values and other values, spread far apart. In the context of t-tests and ANOVAs, you may hear this same concept referred to as equality of variances or homogeneity of variances. Data is homoscedastic if it looks like a shotgun blast of randomly distributed data. The opposite of HomoscedasticityHomoscedasticity is heteroscedasticity, where you might find a cone or fan shape in your data. You check this assumption by plotting the predicted values and residuals on a scatterplot.

Residual scatter plots provide a visual examination of the assumption homoscedasticity between the predicted dependent variable scores and the prediction errors. The primary benefit is that the assumption can be viewed and analyzed with one glance; therefore, any violation can be determined quickly and easily.

A residual scatter plot is a figure that shows one axis for predicted scores and one axis for errors of prediction. Initial visual examination can isolate any outliers in the data set, otherwise known as extreme scores. Tabachnick and Fidell (2007) explain the residuals (the difference between the obtained DV and the predicted DV scores), and the variance of the residuals should be the same for all predicted scores (HomoscedasticityHomoscedasticity). If this is true, the assumption is met. The scatter plot takes the (approximate) shape of a rectangular; scores will be concentrated in the center (about the 0 points) and distributed in a rectangular pattern. More simply, scores will be randomly scattered about a horizontal line. In contrast, any systematic design or clustering of scores is considered a violation.

Heteroscedasticity is any set of data that isn't homoscedastic. More technically, it refers to data with unequal variability (scatter) across a set of second predictor variables. Heteroscedastic data tends to follow a cone shape on a scatter graph. It happens when the standard deviations of a predicted variable, monitored over different values of an independent variable or as related toperiodsperiods, are non-constant.

1. **Spatial auto co-relation among regression residuals**

It describes the presence of systematic spatial variation in a variable. Positive. Positive spatial autocorrelation, which is most often encountered in practical situations, tends to close areas or sites together to have similar values. It is assumed that feature engineering on the Independent variable can be applied, i.e., instead of taking the original raw independent variables, variable transformation can be done. The transformed independent variables, such as taking a log transformation, squaring the variables, reciprocal of the variables, can also build the GLM model. Autocorrelation occurs when the residuals are not independent of each other.

**References**

Anderson, D.; Feldblum, S; Modlin, C; Schirmacher, D.; Schirmacher, E.; and Thandi, N., “A Practitioner’s Guide to Generalized Linear Models” (Second Edition), CAS Study Note, May 2005.

Indskopf, D. (2012). Generalized linear models. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), APA handbook of research methods in psychology, Vol. 3. Data analysis and research publication (pp. 191–206)

Fox, J., & Fox, J. (2008). Applied regression analysis and generalized linear models. Los Angeles: Sage.