Sachin Date · Follow

May 11, 2020 · 9 min read ★

The Clifton suspension bridge (Copyright: sagesolar under CC BY 2.0)

Generalized Linear Models

What are they? Why do we need them?

Generalized Linear Models (GLMs) were born out of a desire to bring under one umbrella, a wide variety of regression models that span the spectrum from Classical Linear Regression Models for real valued data, to models for counts based data such as Logit, Probit and Poisson, to models for Survival analysis.

Models under the GLM umbrella

2. Analysis of Variance (ANOVA) models.

3. Models for ratios of counts. For e.g. models which predict the odds of winning, probability of machine failure etc. Some examples of this class are the Logit model (used in Logistic regression), Probit and Ordered Probit models, and the very powerful Binomial Regression model.

4. Models used for explaining (and predicting) event counts. For e.g. models that predict the number of footfalls at the supermarket, in a mall, in an emergency room. Examples of models of this class are the Poisson and Negative Binomial regression models, and the Hurdle model.

5. Models for predicting time to next failure of parts, machines (and human beings). Models for estimating lifespans of living (and non-living) things.

When each one of the above seemingly diverse set of regression models is expressed in the format of a Generalized Linear Model (and we'll get to explaining what that format is shortly), it gives the modeler the great benefit of applying a common training technique for all such models.

> With Generalized Linear Models, one uses a common training technique for a diverse set of regression models.

Furthermore, GLMs allow the modeller to express the relationship between the regression variables (a.k.a. covariates, a.k.a. influencing variables, a.k.a. explanatory variables) $\boldsymbol{X}$ and the response variable (a.k.a. dependent variable) $\boldsymbol{y}$, **in a linear and additive way** *even though the underlying relationships may be neither linear nor additive*.

> Generalized Linear Models let you express the relation between covariates X and response y in a linear, additive manner.

such as the following:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \cdots + \beta_p * x_p$$

Regression variables x_1, x_2, x_3,...x_p are additively related (Image by Author)

A CLR model is often the 'model of first choice': something that a complex model should be carefully compared with, before choosing the complex model for one's problem.

CLR models come with clear advantages:

Subject to certain conditions being met, they have a neat 'closed-form' solution, meaning, they can be fitted i.e. trained on the data by simply solving a linear algebraic equation.

It is also easy to interpret the trained model's coefficients. For e.g. if a trained CLR model is expressed by the following equation:

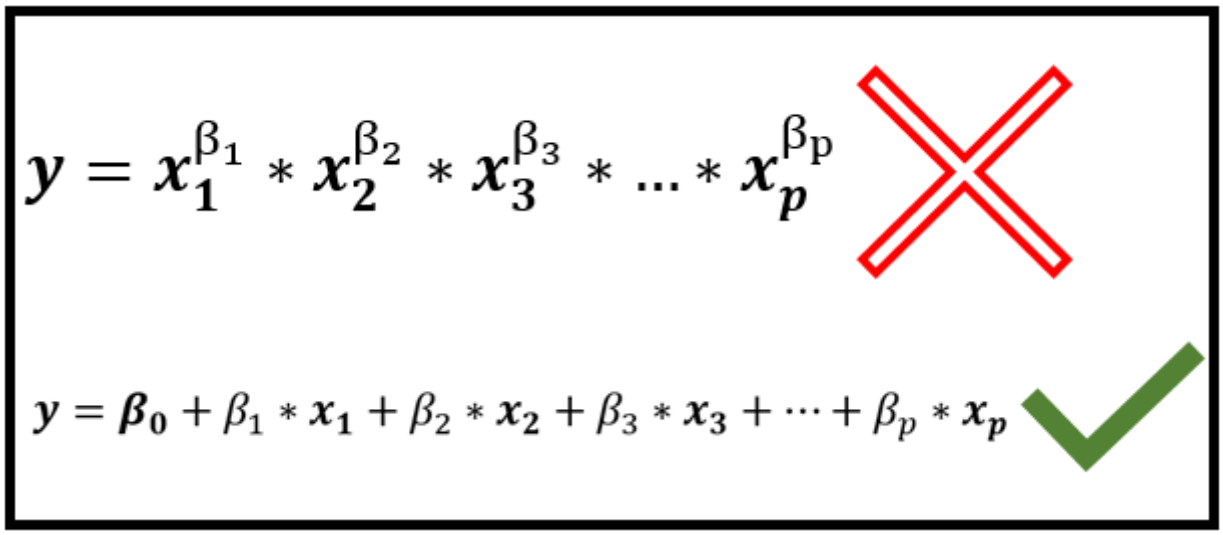
$$\begin{aligned} NumberOfFishCaught \\ = 0.79 * NumberOfCampers \\ - 1.08 * NumberOfChildren \\ - 2.49 \end{aligned}$$

A fitted linear regression model (Image by Author)

It is clear from this equation what the model has been able to find: that for each unit increase in the number of campers the number of fish caught increases by around 75%,

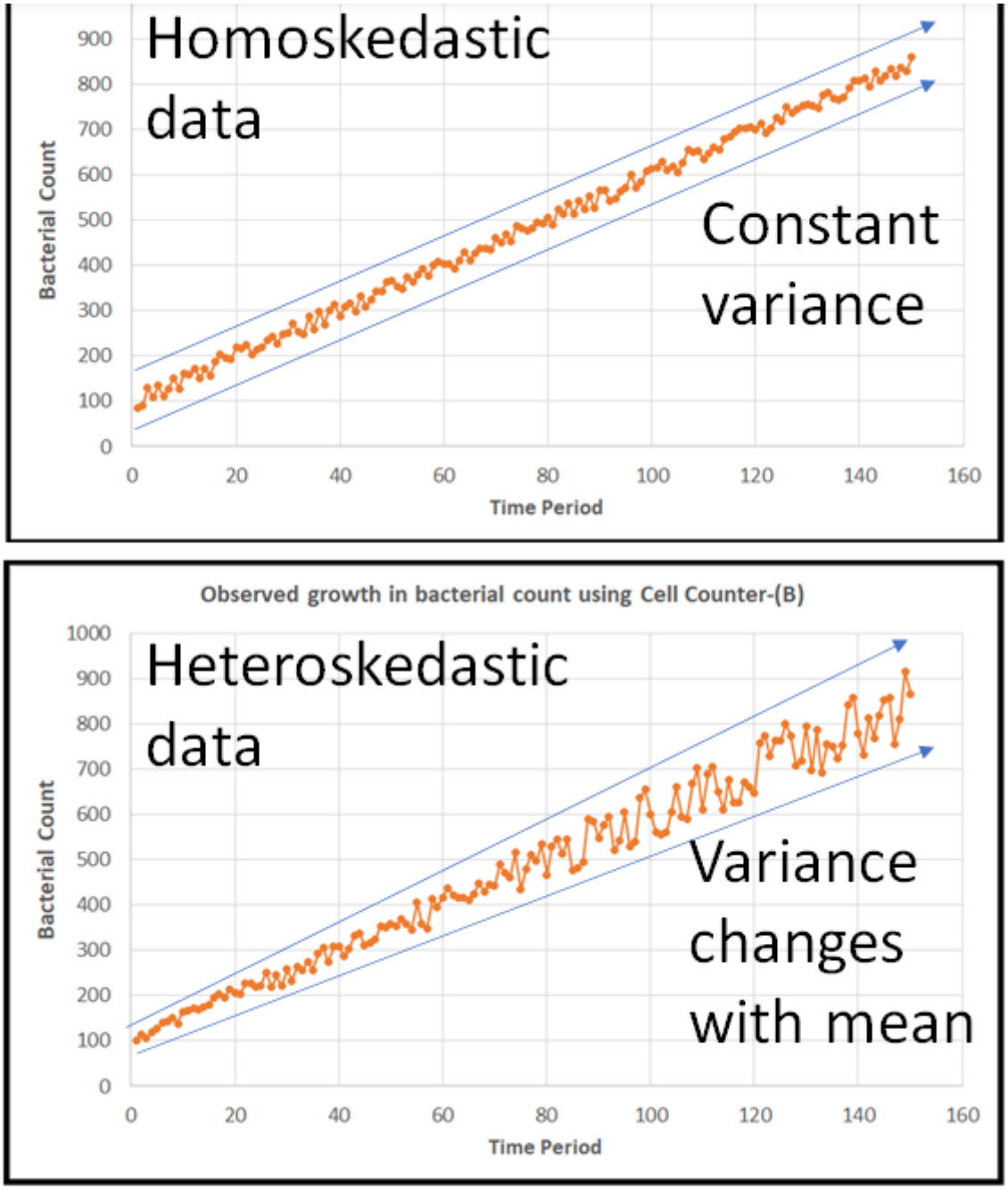
But Classical Linear Regression models also come with some strict requirements, namely:

- **Additive relationships:** Classical Linear models assume that the regression variables should have an additive relationship with each other.



Multiplicative and additive relationships (Image by Author)

- **Homoscedastic data:** Classical Linear models assume that the data should have constant variance i.e. the data should be homoscedastic. In real life, data is often *not* homoscedastic. The variance is not constant and sometimes it is a function of the mean. For e.g. the variance increases as the mean increases. This is common in monetary datasets.



Homoscedastic and heteroscedastic data (Image by Author)

- **Normally distributed errors:** Classical Linear models assume the errors of regression, also known as the residuals, are normally distributed with mean zero. This condition is also difficult to meet in real life.

- **Non-correlated variables:** Finally, the regression variables are assumed to be non-correlated with each other, and preferably independent of each other.

Therefore if your data set is non-linear, heteroscedastic and the residuals are not

The square root and the logarithm transformations are commonly used for achieving these effects as follows:

if:

$$y = x_1^{\beta_1} * x_2^{\beta_2} * x_3^{\beta_3} * \ldots * x_p^{\beta_p}$$

then:

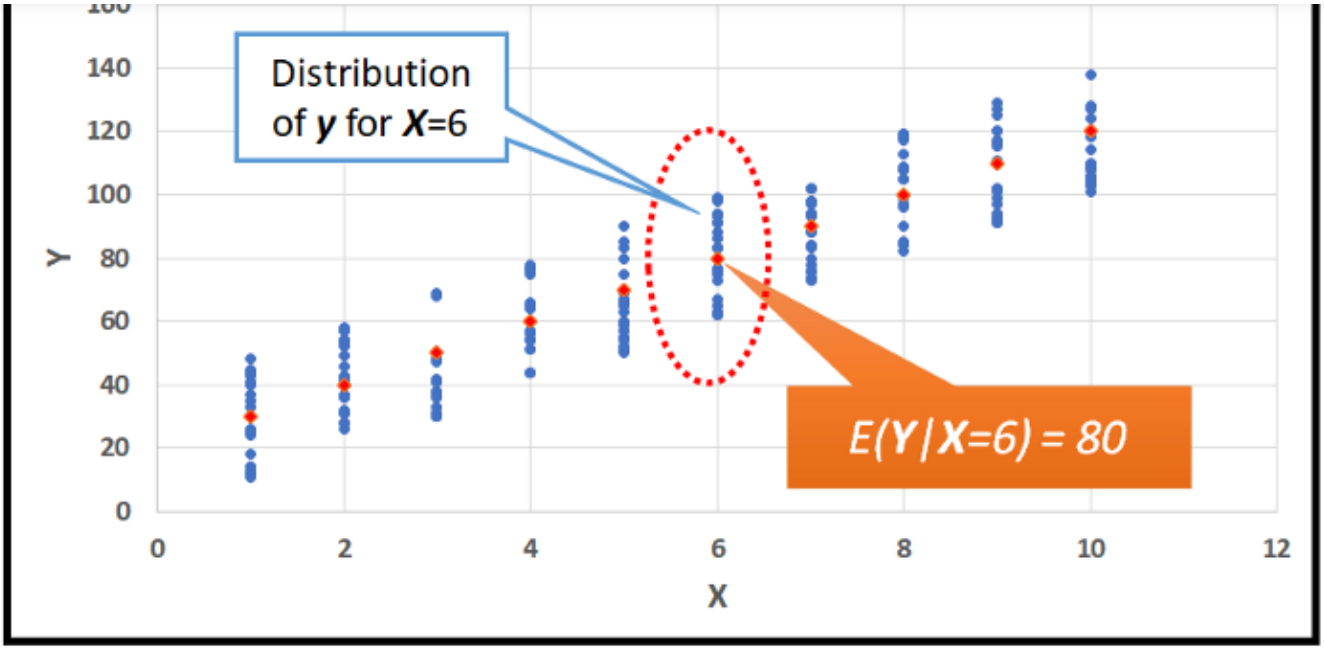
$$log(y) = \beta_1 * \log(x_1) + \beta_2 * \log(x_2) + \beta_3 * \log(x_3) + \cdots + \beta_p * \log(x_p)$$

A logarithm transforms a multiplicative relationship to an additive relationship (Image by Author)

Unfortunately, none of the available transforms are good at achieving all three effects at the same time, namely making the relation linear, minimizing heteroscedasticity and normalizing the error distribution.

There is another great problem with the transformation approach which is as follows:

Recollect that y is a random variable that follows some kind of a probability distribution. So for any given combination of x values in the data set, the real world is likely to present to you several random values of y and only some of these possible values will appear in your training samples. In the real world, these values of y will be randomly distributed around the conditional mean of y given the specific value of x. The conditional mean of y is denoted by $E(y|x)$. This situation can be illustrated as follows:



Conditional expectation E(**y|x**) as denoted by the red dot (Image by Author)

In the variable transformation approach, we make the unrealistically strong assumption that every single value of $\boldsymbol{y}$ i.e. each one of the blue dots in the above plot, after transformation using *log(), sqrt()* etc., will end up having a linear relationship with $\boldsymbol{X}$. This is obviously too much to expect.

What seems more realistic is that the *conditional mean (a.k.a. expectation) of* $\boldsymbol{y}$, i.e. $E(\boldsymbol{y}|\boldsymbol{x},)$ after a suitable transformation, ought to have a linear relationship with $\boldsymbol{X}$.

In other words:

if:

$$\boldsymbol{y} = \boldsymbol{x_1}^{\beta_1} * \boldsymbol{x_2}^{\beta_2} * \boldsymbol{x_3}^{\beta_3} * \ldots * \boldsymbol{x_p}^{\beta_p}$$

then:

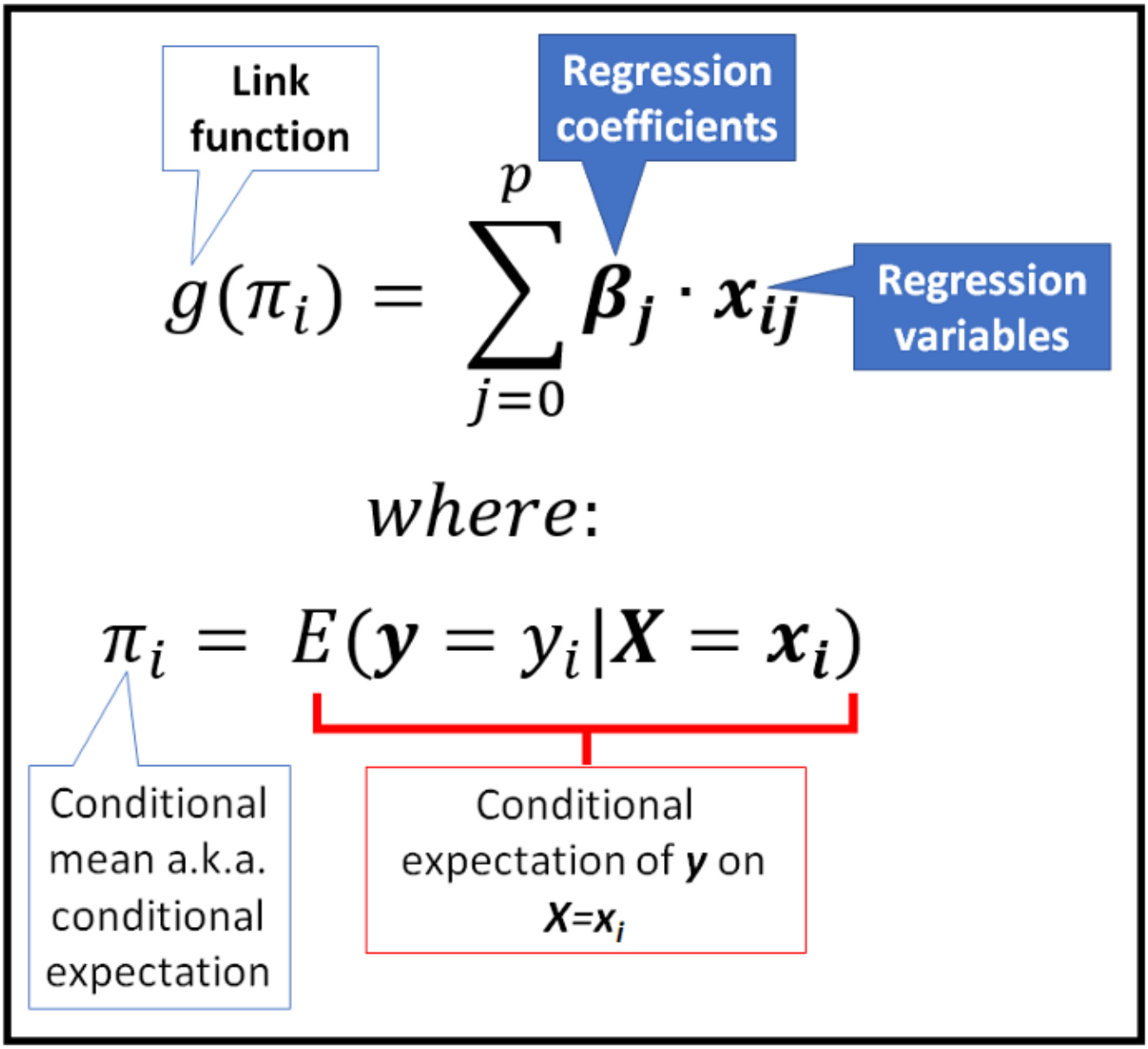
$$log(\boldsymbol{y}) = \beta_1 * \log(\boldsymbol{x_1}) + \beta_2 * \log(\boldsymbol{x_2}) + \beta_3 * \log(\boldsymbol{x_3}) + \cdots + \beta_p * \log(\boldsymbol{x_p})$$

$$log(\boldsymbol{E(y|X = x_i)}) = \beta_1 * \log(\boldsymbol{x_1}) + \beta_2 * \log(\boldsymbol{x_2}) + \beta_3 * \log(\boldsymbol{x_3}) + \cdots + \beta_p * \log(\boldsymbol{x_p})$$

variables $\boldsymbol{X}$.

The transformation function is called a **link function** of the GLM and is denoted by $g(.)$

We illustrate the action of $g()$ as follows:



The link function of Generalized Linear Models (Image by Author)

Thus, instead of transforming every single value of $\boldsymbol{y}$ for each $\boldsymbol{x}$, GLMs transform only the conditional expectation of $\boldsymbol{y}$ for each $\boldsymbol{x}$. So there is no need to assume that every single value of $\boldsymbol{y}$ is expressible as a linear combination of regression variables.

transformed conditional expectation of the dependent variable y as a linear combination of the regression variables X.

The link function $g(.)$ can take many forms and we get a different regression model based on what form $g(.)$ takes. Here are a few popular forms and the corresponding regression models that they lead to:

The Linear Regression Model

In Linear models, $g(.)$ is the following ***identity*** function:

$$g(\pi_i) = \pi_i$$

Identity function used by the Linear Regression Model (Image by Author)

The Logistic (and in general, Binomial) Regression Models

In the Logistic regression model, $g(.)$ is the following ***Logit*** function:

$$g(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$$

$$\therefore \pi_i = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$$

The Logit (log-odds) function used by the Logistic Regression model (Image by Author)

$$g(\pi_i) = \ln(\pi_i)$$

$$\therefore \pi_i = e^{x_i \beta}$$

The log-link function used by the Poisson regression model (Image by Author)

There are many other variants of $g(.)$ such as the Poisson-Gamma mixture leading to the Negative Binomial regression model and the inverse of the Cumulative Distribution Function of the Normal distribution, which leads to the probit model.

How to handle Heteroscedasticity in the data?

Finally, let's look at how GLMs handle heteroscedastic data i.e. data in which the variance is not constant, and how GLMs handle potentially non-normal residual errors.

GLMs account for the possibility of a non-constant variance by assuming that the variance is some function $V(\mu)$ of the mean μ, or more accurately the conditional mean $\mu|\boldsymbol{X}=\boldsymbol{x}$.

In each of the above mentioned models, we assume a suitable variance function $V(\mu|\boldsymbol{X}=\boldsymbol{x})$.

> In Generalized Linear Models, one expresses the variance in the data as a suitable function of the mean value.

In the **Linear regression model**, we assume $\boldsymbol{V(\mu) = some\ constant}$, i.e. variance is constant. Why? Because Linear models assume that $\boldsymbol{y}$ is Normally distributed and a

In the **Logistic and Binomial Regression models**, we assume, $V(\mu) = \mu - \mu^2/n$ for a data set size of n samples, as required by a Logit distributed y value.

Related Post: The Binomial Regression Model: Everything you need to know

In the **Poisson Regression model**, we assume $V(\mu) = \mu$. This is because, the Poisson regression model assumes that y has a Poisson distribution and in a Poisson distribution, *variance* = *mean*.

Related Post: An Illustrated Guide to the Poisson Regression Model

In the **Negative Binomial regression model**, we assume $V(\mu) = \mu + \alpha^*\mu^2$, where α is a dispersion parameter which allows us to deal with over-dispersed or under-dispersed data.

Related Post: Negative Binomial Regression Model: A Step by Step Guide

…and so on for other models.

In GLMs, it is possible to show that the model is not sensitive to the distributional form of the residual errors. In simple terms, the model doesn't care whether the model's errors are normally distributed or distributed any other way, as long as the mean-variance relationship that you assume, is actually satisfied by your data.

> Generalized Linear Models do not care if the residual errors are normally distributed as long as the

> the data.

This makes GLMs a practical choice for many real world data sets that are nonlinear and heteroscedastic and in which we cannot assume that the model's errors will always be normally distributed.

Finally, a word of caution: Similar to Classical Linear Regression models, GLMs also assume that the regression variables are uncorrelated with each other. Therefore GLMs cannot be used to model time series data which typically contain a lot of auto-correlated observations.

> Generalized Linear Models should not be used for modeling auto-correlated time series data.

Summary

Generalized Linear Models bring together under one estimation umbrella, a wide range of different regression models such as Classical Linear models, various models for data counts and survival models.

Here is a synopsis of things to remember about GLMs:

- GLMs deftly side-step several strong requirements of classical linear models such as additivity of effects, homoscedasticity of data and normality of residual errors.
- GLMs impose a common functional form on all models in the GLM family which consists of a link function $g(\mu|\boldsymbol{X}=\boldsymbol{x})$ that allows you to express the transformed conditional mean of the dependent variable $\boldsymbol{y}$ as a linear combination of the regression variables $\boldsymbol{X}$.
- GLMs require the specification of a suitable variance function $V(\mu|\boldsymbol{X}=\boldsymbol{x})$ for expressing the conditional variance in the data as function of the condition mean. What form $V(.)$ takes depends on the probability distribution that you assume for the dependent variable $\boldsymbol{y}$ in your data set.




- GLMs do assume that regression variables $\boldsymbol{X}$ are uncorrelated, thereby making GLMs unsuitable for modeling auto-correlated time series data.

Happy modeling!

Thanks for reading! I write about topics in data science, with a focus on regression and time series analysis and forecasting.

If you liked this article, please follow me at ***Sachin Date*** *to receive tips, how-tos and programming advice, on topics devoted to regression and time series analysis.*