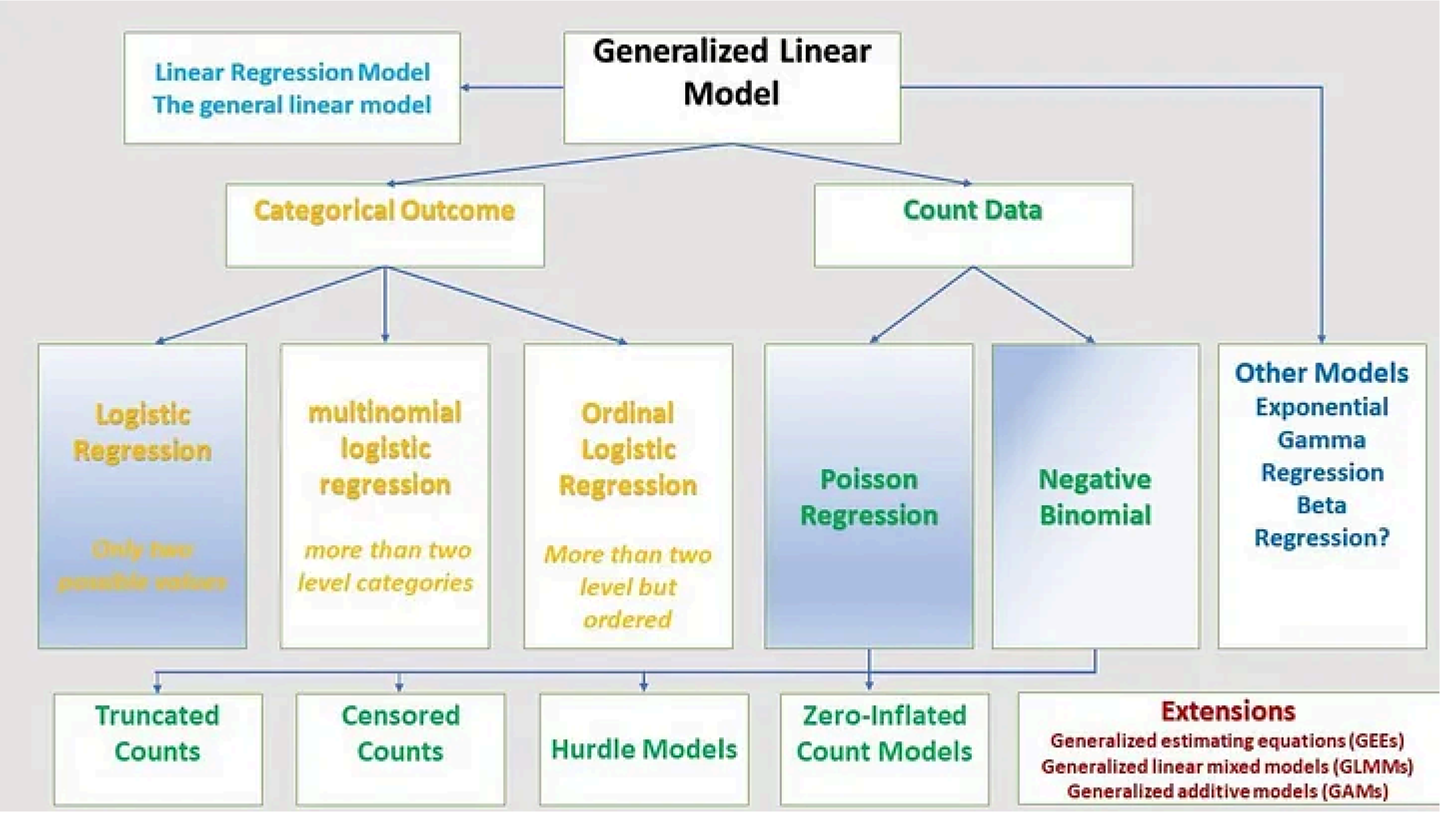
Generalized Linear Models



- This is a diagram **GLMs** Generalized Linear Models which we can notice that all the statistical learning methods we covered so far are are derived from **Linear Regression** by extending it to accommodate various types of response Y variables .

Recall that Linear Regression model rely on these two assumptions :

- Relationship between X and Y is **Linear**
- Normality of Residuals $Y - \hat{Y}$
- Independence of Error
- constant variance of error terms

GLMs try to expand the Linear Regression and make it more flexible by :

- Allow for the response to follow various distributions $\rightarrow$ **Binomial** , **Poisson** , **gamma**
- Allow non-linear relationships within the linear framework
- More flexible on the assumptions and limitations
- Allow for different error structures accommodating hetroscedasticity

Scope of Application

- Linear Regression Models are best suited for **Continuous data** that fits Gaussian distribution
- **GLMs** can handle border scope of data types $\rightarrow$ binary, count and continuous data

Math Behind GLM:

GLMs are based on the assumption of the response Y variable distribution comes from the **exponential family** that includes : Gaussian, Binomial, Poisson and gamma

$$f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

- y is the response outcome variable
- θ is the parameter as β the coefficient in linear regression
- ϕ Dispersion parameter which is the scales variance
- $a(\phi), b(\theta), c(y, \phi)$ Known functions that define the distribution

Exponential Family of Distributions

Distribution	Domain	Parameter θ	$b(\theta)$	$Var(Y)$	Used For
Gaussian	$(-\infty, \infty)$	μ	$\frac{\theta^2}{2}$	σ^2	Linear Regresion
Binomial	$(0, 1, \ldots, n)$	$\log\left(\frac{p}{1-1}\right)$ Logit	$\log(1+e^{\theta})$	$np(1-p)$	Logistic Regresison
Multinomial		$\log\left(\frac{e^{X_i}}{\sum e^{X_j}}\right)$	$\log\left(1+\sum_{k=1}^{K-1} e_k^{\theta}\right)$		Multi-class Logistic Regression
Poisson	$(0, 1, 2\ldots)$	$\log(\lambda)$	e^{θ}	λ	Poisson Regression
Exponential	$(0, \infty)$	$-\lambda$	$-\log(-\theta)$	$\frac{1}{\lambda^2}$	Survival Analysis
Gamma	$(0, \infty)$	$-\frac{\nu}{\mu}$	$-\log(-\theta)$	$-\frac{\mu^2}{\nu}$	Positive Continuous Data

GLM's Components

These are the core components that define the Generalized Linear Models structure and functionality

1. Error distribution
2. Linear predictor
3. Link function

Error Distribution

This referred to the probability distribution of the response variables which determines the form of the likelihood function used for the parameters estimations, the selection of the distribution is guided by the nature of the response variables (**continuous, binary, count,...**) The **GLM's** include :

- **Gaussian Distribution** Used for continuous response Y and where the residuals are assumed to be normally distributed
- **Binomial Distribution** Fits the Binary response variables as in Logistic Regression where the outcome is in the form **Two classes**
- **Poisson Distribution** Fits when the response variables represent an event that occur in a fixed interval or space, as in the **Poisson Regression example** discuss later on
- **Gamma Distribution** Used for continuous positive data, often used when the response Y variables represent time until event occurs

Linear Predictor

The linear predictor as heavily discuss in the Linear Regression chapter where it a **Linear combination** of these variables each multiplied by a Coefficient that quantifies the relationship between the predictor and the response variables Y :

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Which can be written in the matrix form of :

$$\eta = X\vec{\beta}$$

Link Function

The link function connects the linear predictor to the mean μ , It transforms the expected value of the response Y variable to the scale of the linear predictor, to make sure it stay within the range for example **Logistic Regression** logit function that transforms the probability to an unbounded

Distribution	Function	Purpose	Formula
Gaussian	Identity	Directly relates the linear predictor to the response	$\eta = \mu$
Binomial	Logit	Transforms the probability of the success to an unbounded scale	$\eta = \log\left(\frac{\mu}{1-\mu}\right)$
Multinomial	Softmax	Generalizes the logit function for multi class outcome	$\eta_i = \log\left(\frac{e^{X_i}}{\sum e_i^X}\right)$
Poisson	Log	Relates the log of the meant count to the linear predictors, Used for count data	$\eta = \log(\mu)$

- μ is he response variable's distribution

Poisson Regression

This section covers the **Poisson Regression** and its use case when the **Linear Regression** doesn't fit the problem at hand

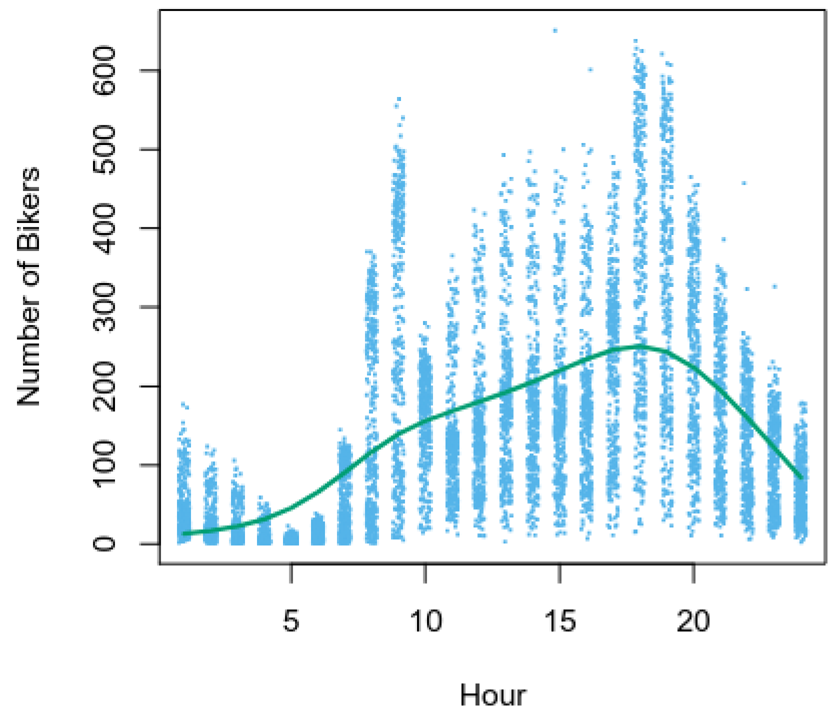
The example at hand is **Bike Share** data where the response Y is **bikers** the number of hourly **bikers** users, this response isn't qualitative nor qualitative it's more of a count

Linear Regression

Now trying to fit the **Linear Regression** model to this data set, which after fitting results in this following table

	Coefficient	Std. error	t-statistic	p-value
Intercept	73.60	5.13	14.34	0.00
workingday	1.27	1.78	0.71	0.48
temp	157.21	10.26	15.32	0.00
weathersit[cloudy/misty]	-12.89	1.96	-6.56	0.00
weathersit[light rain/snow]	-66.49	2.97	-22.43	0.00
weathersit[heavy rain/snow]	-109.75	76.67	-1.43	0.15

- 9.6% of the the fitted values are negative which predict negative users which is plane wrong
- Which makes us question the accuracy and the meaningfulness of the predictions



- This graph clearly shows a violation of the constant variance σ^2
- The linear regression results in a continuous response Y which isn't for the kind of response we want fro this data

Sometimes transforming the response for example :

$$\log(Y) = \sum_{j=1}^{p} X_j\beta_j + \varepsilon$$

where the $\log$ avoids negative values for the predictions and reduce the variance, but its not a perfect solution since transforming the response will make it challenging to interpret

Poisson Regression

To overcome this limitation that the linear regression present there is another approach called **Poisson regression** which fits the data and response at hand since it measures where :

- Its a Discrete distribution

- Takes only positive values
- Used to model counts

The Poisson distribution follows this :

$$\Pr(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

for $k = 0, 1, 2 \ldots$

- Where $\lambda > 0$ is the expected value of the mean of the response Y , Which is also the variance $\lambda = E(Y) = Var(Y)$
- Which means the larger the expected value the larger the variance which fits the non-constant our data got
- In practical examples the expected value λ is a function that changes depending on various predictors and observations $\lambda(X_1, \ldots, X_p)$

$$\log \lambda(X_1, \ldots, X_p) = X\beta$$

$$\lambda(X_1, \ldots, X_p) = e^{X\beta}$$

And same as the other linear models to estimate the values of coefficients $\beta_0, \ldots \beta_p$ we maximize the likelihood of the coefficients

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \frac{e^{-\lambda_{x_i}}\lambda(x_i)^{y_i}}{y_i!}$$

- Where $\lambda(x_i) = e^{x\beta}$
- Here we maximize the likelihood of the coefficients that makes observed data as likely as possible

Now after fitting the **Poisson Regression** model resulting in these values :

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	4.12	0.01	683.96	0.00
`workingday`	0.01	0.00	7.5	0.00
`temp`	0.79	0.01	68.43	0.00
`weathersit[cloudy/misty]`	-0.08	0.00	-34.53	0.00
`weathersit[light rain/snow]`	-0.58	0.00	-141.91	0.00
`weathersit[heavy rain/snow]`	-0.93	0.17	-5.55	0.00

Linear Regression Vs. Poisson Regression

Interpretation : in the Poisson regression one unite change in X_i is associated with change in the expected value $E(Y) = \lambda$ by a factor of e^{β_i} for example a change by $e^{-0.08} = 0.923$ which means only 92.3% people will use it when its cloudy relative to when its clear

Mean-variance relationship :The Poisson mode $\lambda = E(Y) = Var(Y)$ which means that given each hour k it got a variance to it unlike the linear regression which always takes a constant value, in the used data each hour got a variance to it which Poisson is able to capture and model

non-negative fitted values : There is no negative predictions in the Poisson regression model, uses it only allows for non negative values by the Poisson formula distribution