Phd Program in Transportation

Transport Demand Modeling

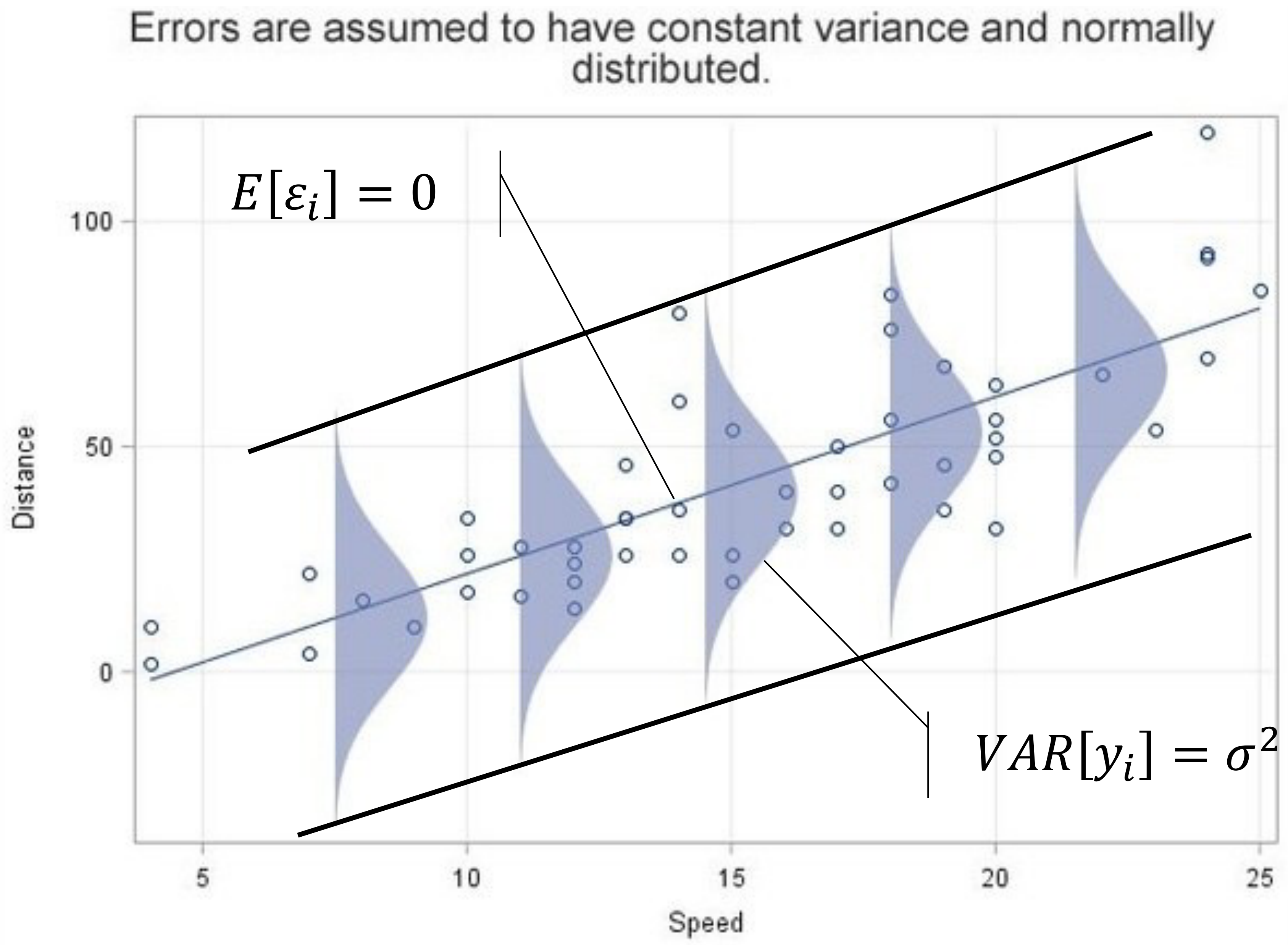
Filipe Moura

Generalized Linear Models

Why Generalized Linear Models?

- **Why using GZLM?**
  - We shall see that these models extend the linear modelling framework to models where:
    - The dependent variable may not be continuous
    - The effect of independent variables may not be linear
    - The expected value of the errors terms might not be 0.

  - GZLM unify all non linear models, used to explain the situation were the linear normal regression was not able to explain the relation under analysis
  - GZLMs are most commonly used to model binary or count data, so we will focus on models for these types of data.

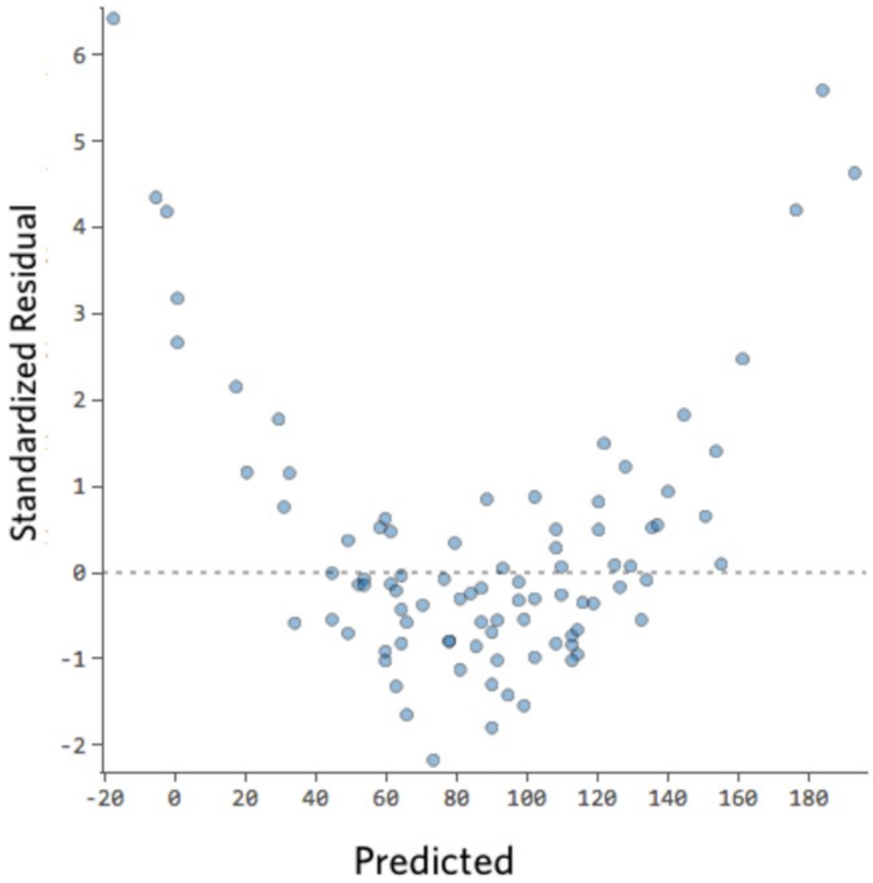

Errors are assumed to have constant variance and normally distributed.
$E[\varepsilon_i] = 0$
$VAR[y_i] = \sigma^2$
Distance
100
50
0
5
10
15
20
25
Speed

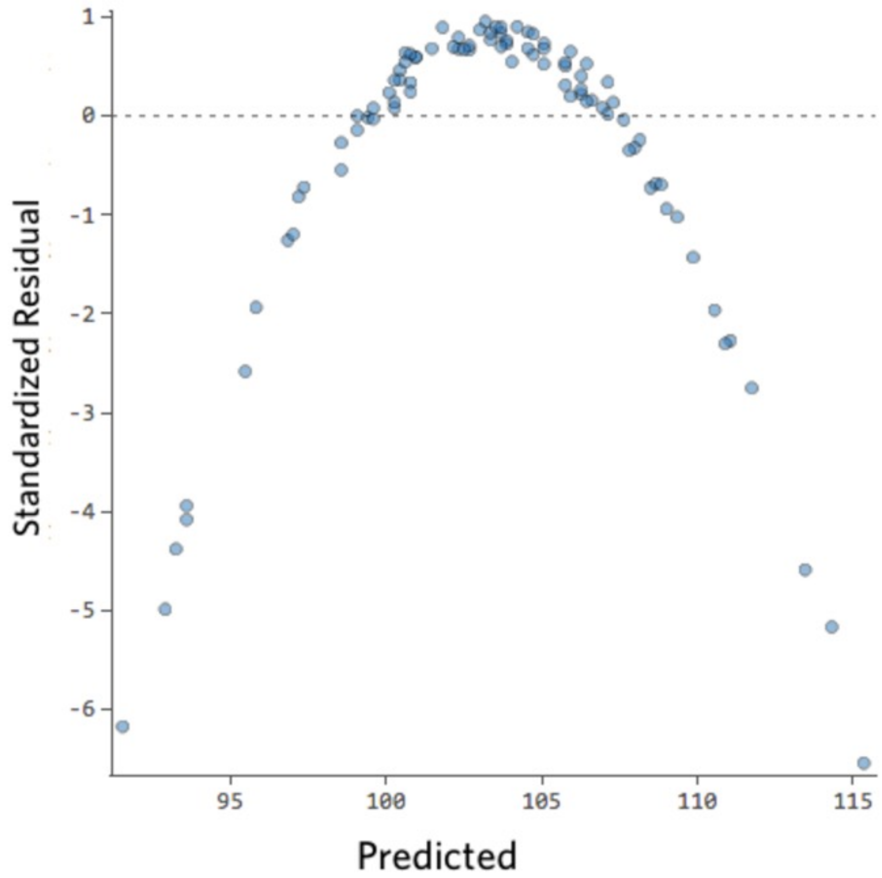
When do GZLM come into play?

- With MLR:
  - $Y_i = \mathrm{B}X + \varepsilon$, where X is a vector of predictors and B is a vector of coefficients β
  - $E[Y_i] = \widehat{\mathrm{B}}X$ because $E[\varepsilon_i] = 0$ and $VAR[y_i] = \sigma^2$
  - Etc.
- When such conditions are not met,
  - When $E[\varepsilon_i] \neq 0$ or $VAR[Y_i] \neq \sigma^2$
- you use GZLM where…
  - The variation (probabilistic distribution) in the response variable Yi can be explained in terms of the values of X
  - We want to find some link function *g(.)*, that mediates the response variable (Yi) and the regressors Xi, such that

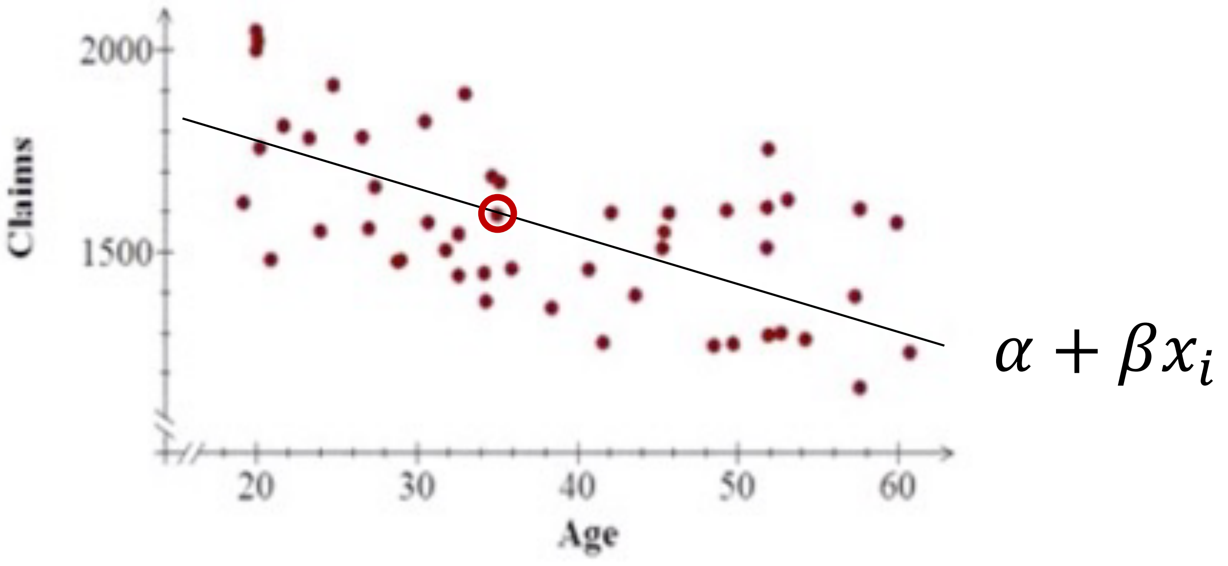
    $E[g(Y_i)] = \widehat{\mathrm{B}}X$, where *g(.)* is the link function.

Nonlinear


Standardized Residual
6
5
4
3
2
1
0
-1
-2
-20 0 20 40 60 80 100 120 140 160 180
Predicted


Standardized Residual
1
0
-1
-2
-3
-4
-5
-6
95 100 105 110 115
Predicted

Recap on linear models and generalizing



- Distributions of the Y_i's: $Y_i \sim N(\mu_i, \sigma^2)$
- Function of the explanatory variable, x_i's: $\alpha + \beta x_i$
- Connection between explanatory variable and the distribution of Y_i:

$$\mu_i = E[Y_i] = \alpha + \beta x_i$$

- We will now generalize the distribution of the Yi variables according to different distributions, besides the normal distribution

Components of the GZLM

A generalized linear model is made up of a **linear predictor**:

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

And two functions

- a **link function** that describes how the mean, $E(Y_i) = \mu_i$, depends on the linear predictor:

$$g(\mu_i) = \eta_i$$

- a **variance function** that describes how the variance, var(Y_i) depends on the mean

$$var(Y_i) = \emptyset V(\mu_i)$$

where Ø is a dispersion parameter and is constant across the observations *i*

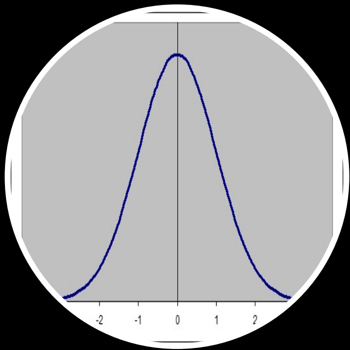
Components of the GZLM

- **Distribution of the Y_i's**
  - Linear models: $Y_i \sim N(\mu_i, \sigma^2)$
  - GZLM: $Y_i \sim exponential\ family$
- **Linear predictor =** function of the covariates (explanatory variables)
  - Linear models: $\eta_i = \alpha + \beta x_i$
  - GZLM: e.g. $\eta_i = \alpha + \beta x_i + \gamma z_i$

    $\eta_i = \alpha + \beta x_i + \gamma x_i^2$
- **Link function** = connection between the linear predictor and $E[Y_i] = \mu_i$
  - Linear models: $\eta_i = \mu_i \Rightarrow \mu_i = \alpha + \beta x_i$
  - GZLM: e.g. $\eta_i = ln(\mu_i) \Rightarrow \mu_i = e^{(\alpha + \beta x_i + \gamma x_i^2)}$

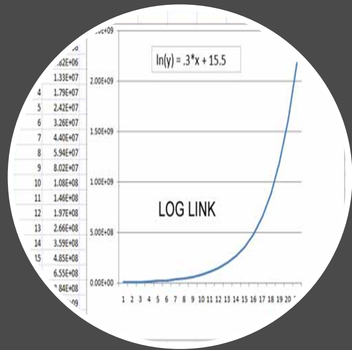
Inverse function

Structure of Generalized Linear Models



Random Component

$Y_i \sim exp.family$



Link Function

$g(\mu_i) = \eta_i$

Systematic Component

$\eta_i = \beta_0 + \sum_{pi} \beta_p x_{pi}$

Fundamental condition for using GZLM

- Response variable distribution must be a member of the

Exponential Family

- It corresponds to a $\underline{u}$ function that belongs to the exponential family with a single parameter θ and a probability distribution function (pdf) such as

$$f(u,\theta) = s(u).t(\theta).\exp\{a(u).b(\theta)\},$$ where s, t, a, b are known functions

or

$$f(u,\theta) = \exp\{a(u).b(\theta) + d(u) + c(\theta)\}$$

where $d(u) = \ln\big(s(u)\big)$ and $c(\theta) = \ln\big(t(\theta)\big)$

- When $\boldsymbol{a(u) = u}$, the distribution is said to be in **canonical form**.
- $\boldsymbol{b(\theta)}$ is called the **natural parameter** of the distribution function.
- For each function of the Exponential Family, one **parameter is of interest**. The remaining are said **nuisance parameters**.

Members of the Exponential Family: Normal Distribution

FEUP

- Normal distribution: N(μ, σ²)

$$f(u,\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp^{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2}$$, with $-\infty \leq \mu \leq \infty$

$f(u,\mu)$: u – Random variable; μ – Interest parameter

$$f(u,\mu) = exp\left\{u.\frac{\mu}{\sigma^2} + \left[\frac{-\mu^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2)\right] - \left(\frac{-u^2}{2\sigma^2}\right)\right\}$$

$a(u)$ = u

$b(\mu)$ = $\frac{\mu}{\sigma^2}$ ← Natural parameter

$c(\mu)$ = $\frac{-\mu^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2)$

$d(u)$ = $\left(\frac{-u^2}{2\sigma^2}\right)$

Members of the Exponential Family: Binomial Distribution

- Binomial distribution: Bin(n, p)

$$f(u,p) = \binom{n}{u}.p^u.(1-p)^{n-u}$$, n=0,1,2,....,n trials

$$= \binom{n}{u}.\left(\frac{p}{1-p}\right)^u.(1-p)^n$$

Random variable: u

Interest parameter: p

$$f(u,\mu) = exp\left\{u.ln\left(\frac{p}{1-p}\right) + nln(1-p) - ln\binom{n}{u}\right\}$$

$a(u)$ = u

$b(p)$ = $ln\left(\frac{p}{1-p}\right)$ — Natural parameter

$c(p)$ = $nln(1-p)$

$d(u)$ = $ln\binom{n}{u}$

Members of the Exponential Family: Poisson Distribution

- Poisson distribution: P(λ)

$$f(u,\lambda) = \frac{e^{-\lambda}.\lambda^{u}}{u!}$$, u=0,1,2,....,n observations

Random variable: u

Interest parameter: λ

$$f(u,\mu) = exp\{u.ln(\lambda) - \lambda - ln(u!)\}$$

$a(u)$: u

$b(\lambda)$: $ln(\lambda)$ — Natural parameter

$c(\lambda)$: λ

$d(u)$: $ln(u!)$

Calibration of GZLM

- Suppose we have a set of independent observations, where
  - Y_i, X_i, for i=1,2,…n observations
    and X_i is a vector of regressors = $X_1, X_2, \ldots. X_p$
    and Y_i is a response variable (dependent variable) we want to estimate and that belongs to some Exponential Family distribution where *a(Y)=Y*
  - The joint pdf can be written as:

$$f(Y_1, Y_2, \ldots, Y_n, \theta, \phi) = \prod_{i=1}^{n} exp\{Y_i . b(\theta_i) + c(\theta_i) + d(Y_i)\}$$

$$= exp\left\{\sum_{i=1}^{n} Y_i . \sum_{i=1}^{n} b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(Y_i)\right\}$$

  - The variation of Yi (**through the link function**) can be explained in terms of the regressors Xi, based on the calibrated coefficients β
  - **NOT THE VALUES OF Yi THEMSELVES!!!!**

Calibration of GZLM

- For X_i, a vector of regressors = X_1, X_2, ... X_p
  - We hope to find a set of parameters $\beta=(\beta_1, \beta_2,\ldots, \beta_p)$ that fits the regressor values to a link function $g(\mu_i)$ that transforms the response variable (Yi) such that, in the case of the normal distribution – $N(\mu, \sigma^2)$,

$$E[\mu_i] = \mu_i = g(\mu_i) = \mathrm{B}X$$

Link function

  - When the response variable Yi follows a normal distribution, the GZLM is equal to the MLR

$$E[Y_i] = \mu_{yi} = \mathrm{B}X + E[\epsilon] = \mathrm{B}X, \text{ as } E[\epsilon]=0$$

Calibrating regressions
Maximum Likelihood Estimation (MLE) - Recap

- **Obtain the likelihood**:

$$L(\mu) = f(y_1).f(y_2)\dots f(y_n)$$

- **Log it** – to make it easier to differentiate

$$ln[L(\mu)]$$

- **Differentiate** and **set** the derivative equal **to zero:**

$$\frac{\delta}{\delta\mu}(ln[L(\mu)]) = 0 \Rightarrow \hat{\mu} = \cdots$$

- **Check its maximum**:

$$\frac{\delta^2}{\delta\mu^2}(ln[L(\mu)]) < 0 \Rightarrow max$$

Calibrating GLZM
Maximum Likelihood Estimation (MLE)

- **Obtain the likelihood**:

$$L(\mu_1, \mu_2, \ldots, \mu_n) = f(y_1).f(y_2)\ldots f(y_n)$$

- **Log it** – to make it easier to differentiate

$$ln[L(\mu_1, \mu_2, \ldots, \mu_n)]$$

- Use the link function to replace the $\mu_i's$:

$$ln[L(\beta_1, \beta_2, \beta_3, \ldots)]$$

- **Differentiate** and **set** the derivative equal **to zero:**

$$\frac{\delta}{\delta\beta_1}(ln[L(\beta_1, \beta_2, \beta_3, \ldots)]) = 0 \Rightarrow \widehat{\beta_1} = \cdots$$

$$\frac{\delta}{\delta\beta_2}(ln[L(\beta_1, \beta_2, \beta_3, \ldots)]) = 0 \Rightarrow \widehat{\beta_2} = \cdots$$

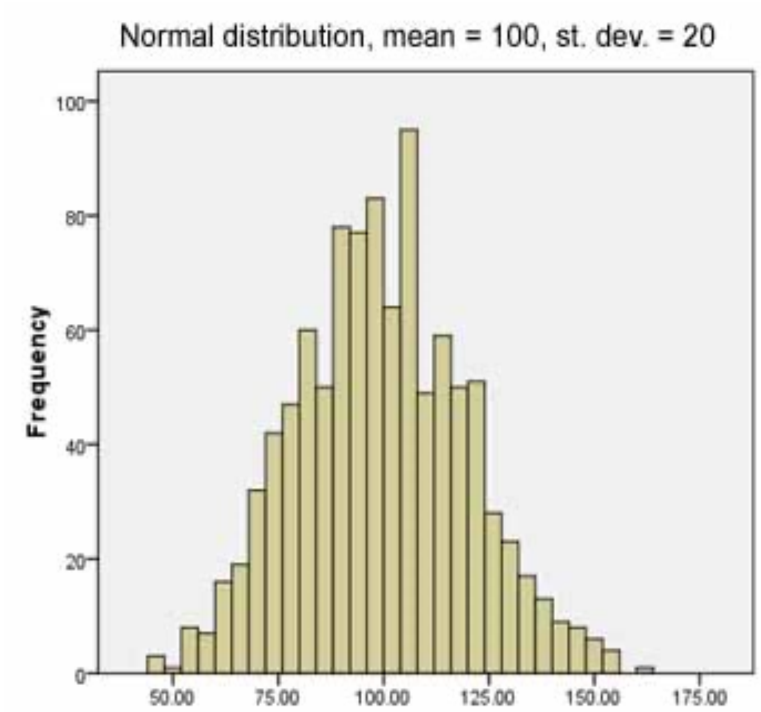
- **Obtain the vector of $\boldsymbol{\beta_i}$ that fit the model estimates to the link function** $g(\mu_i)$

Link functions and Inverse functions

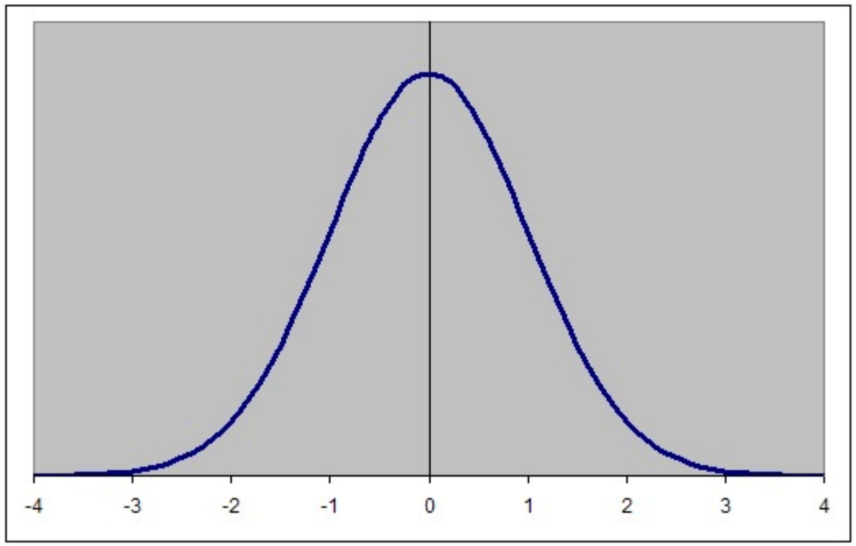
- If the random component of *Yi* follows a **normal distribution**
  - The corresponding **natural parameter** is: $b(\mu_i) = \mu_i$
  - The **link function** is: $g(\mu_i) = \mu_i$ (**Identity link**)
  - Then: $g(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots = \mathrm{B}X$
  - And inversely: $\mu_i = \mathrm{B}X$
- If the random component of *Yi* follows a **binomial distribution**
  - The corresponding **natural parameter** is: $b(\mu_i) = ln\left(\frac{p_i}{1-p_i}\right)$
  - The **link function** is: $g(\mu_i) = ln\left(\frac{p_i}{1-p_i}\right)$ (**logit link** – or logistic)
  - Then: $g(p_i) = ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots = \mathrm{B}X$
  - And inversely: $\mathrm{E}[Y_i] = p_i = \frac{exp^{(\mathrm{B}X)}}{1+exp^{(\mathrm{B}X)}}$

- The distribution of the dependent variable has the form of the bell-shaped symmetrical curve centered in the mean.
- This implies the dependent variable is continuous.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$



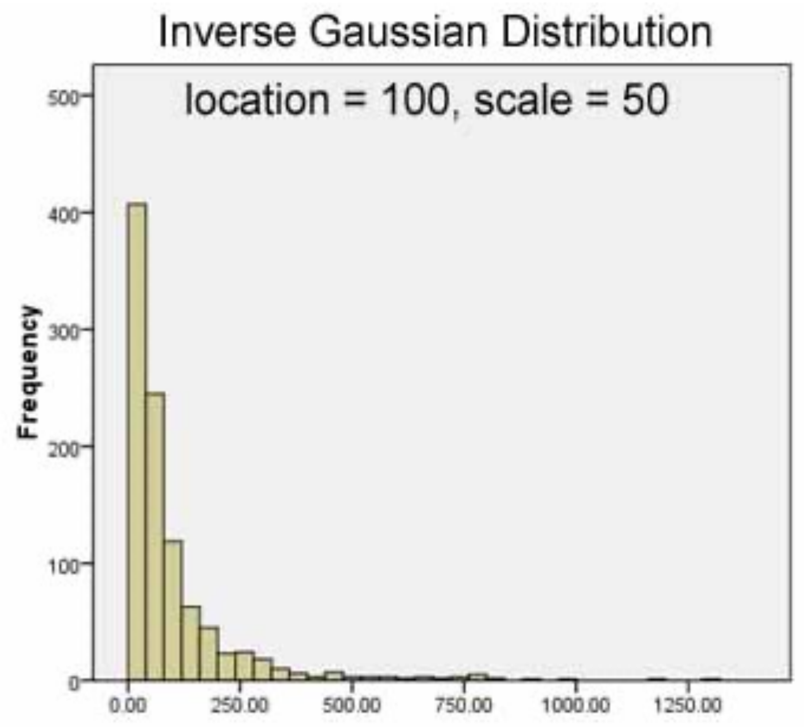
Normal distribution (standardized)



Generalized Linear Models

Inverse Gaussian (Wald Distribution)

- It is used for dependent variables that are **positively skewed** and have values always greater than 0.
- Values must be greater than 0 or are dropped. It has been used to model **diffusion processes, insurance claims, etc**



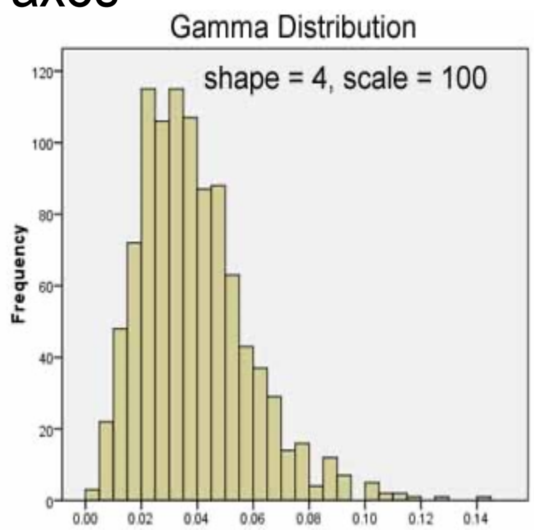
If λ tends to infinity, the distribution becomes like a Normal distribution

$$f(x;\mu,\lambda) = \left[\frac{\lambda}{2\pi x^3}\right]^{1/2} \exp \frac{-\lambda(x-\mu)^2}{2\mu^2 x}$$

Gamma

- This is an **alternative for positively skewed** dependent variables. It is highly sensitive to the shape parameter.
  - When the shape parameter is greater than 1, the gamma distribution is bell-shaped but positively skewed as shown in the figure below.
  - When the shape parameter is 1, the gamma distribution is exponentially declining.
  - When the shape parameter is less than 1, the gamma distribution is also exponentially declining and asymptotic to the axes
- The gamma distribution has been used in **survival analysis** and modeling **duration-of-event data**.

$$f(x; k, \theta) = x^{k-1} \frac{e^{-x/\theta}}{\theta^k \, \Gamma(k)} \text{ for } x \geq 0 \text{ and } k, \theta > 0.$$



Generalized Linear Models
Multinomial

- This distribution is used when the dependent variable has a **finite number of categories**, such as text string values, or is **ordinal**.
- The distribution among categories, not shown, is arbitrary.

$$f(x_1, \ldots, x_k; n, p_1, \ldots, p_k) = \Pr(X_1 = x_1 \text{ and } \ldots \text{ and } X_k = x_k)$$

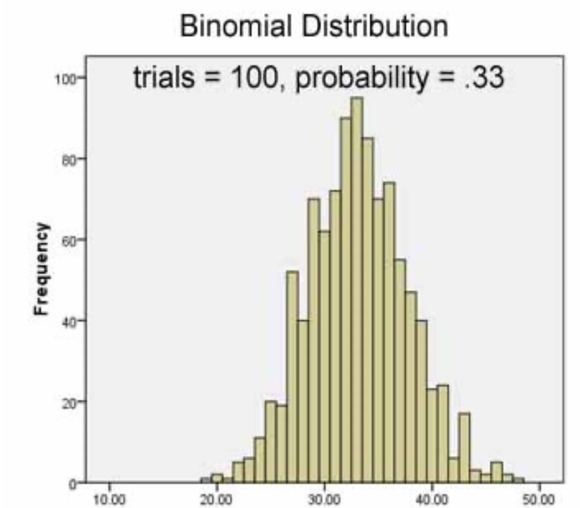
$$= \begin{cases} \dfrac{n!}{x_1! \cdots x_k!} p_1^{x1} \cdots p_k^{xk}, & \text{when } \sum_{i=1}^{k} x_i = n \\ 0 & \text{otherwise,} \end{cases}$$

Generalized Linear Models
Binomial

- Used when the **dependent variable is binary**.
- The count of events in a fixed number of trials also has a binary distribution. Examples of binomial data are attributes "present/not present", "innovation adopted/not adopted", or "success/failure" data.
- It is assumed that the two values have a fixed rather than changing probability of occurrence (as in coin-flipping), even if that probability is not known

$$f(k;n,p) = \Pr(K=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

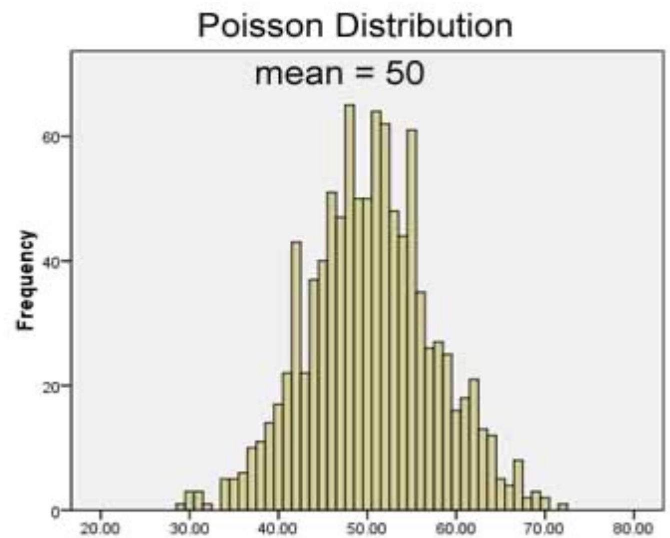
$$F(x;n,p) = \Pr(X \le x) = \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$$



Generalized Linear Models
Poisson (COUNT DATA)

- The Poisson distribution is also used for count data and is preferred **when events are rare, as in modeling accidents**, wars, or epidemics
- The binomial distribution is used when the dependent variable corresponds to **data counts of successes per given number of trials**
- **The Poisson distribution is used to count successes per given number of time units**

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$



Generalized Linear Models
Poisson (COUNT DATA)

- A rule of thumb is to use a Poisson rather than binomial distribution when n is $\geq$ 100, the probability of each event is below 0.05
- The Poisson distribution is also used when "events" can be counted but **non-occurrence of events cannot be counted (unreported)**.

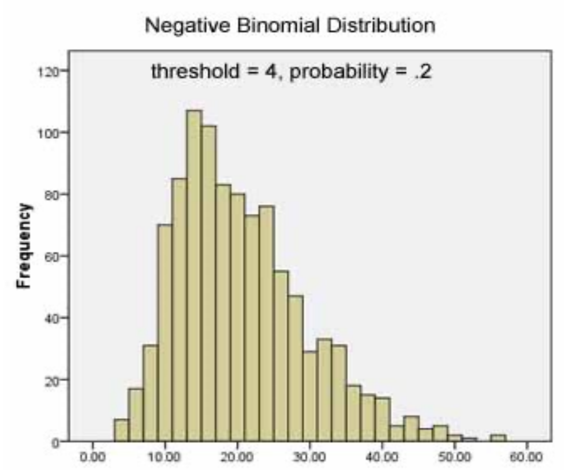
- In Poisson distributions, the **mean equals the variance**
  - Presence of homoscedasticity since variance doesn't change over data and λ is assumed constant
  - As such, there is **no over-dispersion of data**.
- All values are non-negative integers
  - Thus, count data, which cannot be negative, are better represented by Poisson than normal distributions

FEUP

Negative binomial

- It is like the Poisson distribution, also used for count data, but it is used when the variance is larger than the mean => **over-dispersion of data**.
- Typically, this is characterized by "**there being too many 0's**."
  - As such, not all cases have an equal probability of experiencing the rare event, but instead, events may be clustered.
  - The negative binomial model is therefore sometimes called the "**over dispersed Poisson model**" . Values must still be non-negative integers.
- The negative binomial is specified by an **ancillary /dispersion parameter k** (sometimes referred to as α or ψ).
  - When k=0, the negative binomial is equal to the Poisson distribution.

$$f(k) \equiv \Pr(X = k) = \binom{k+r-1}{k}(1-p)^r p^k \quad \text{for } k = 0, 1, 2, \ldots$$



❒ The linear predictor

- Quantity that incorporates the information about the independent variables into the model
- For a matrix of $\underline{n}$ observations and of $\underline{p}$ variables, the linear predictor $\underline{\eta}$ can be expressed as:

$$\eta_i = \sum_{j=1}^{p} x_{ij} \beta_j$$

- Where
  - ❖ each x_{ij} is the value of the j[st] IV for the i[st] observation
  - ❖ β_j belong to a vector of unknown parameters to be estimated

Generalized Linear Models
Link function

- It provides the **relationship between the linear predictor and the mean of the distribution function**
  - The way the two previous components relate to each other
  - ***In fact, the link function is a transformation of the response variable***
- It is a monotonous and differentiable function $g(\mu_i)$ that transforms

  μ_i in η_i where $g(\mu_i) = \eta_i$
- Inversely, μ_i can be obtained with (inverse function)

$$E[Y_i] = \mu_i = g^{-1}(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}) + e_i$$

where μ_i is the expected value of Y_i

and X_{ij} are the predictors or explanatory variables

Generalized Linear Models
Link function

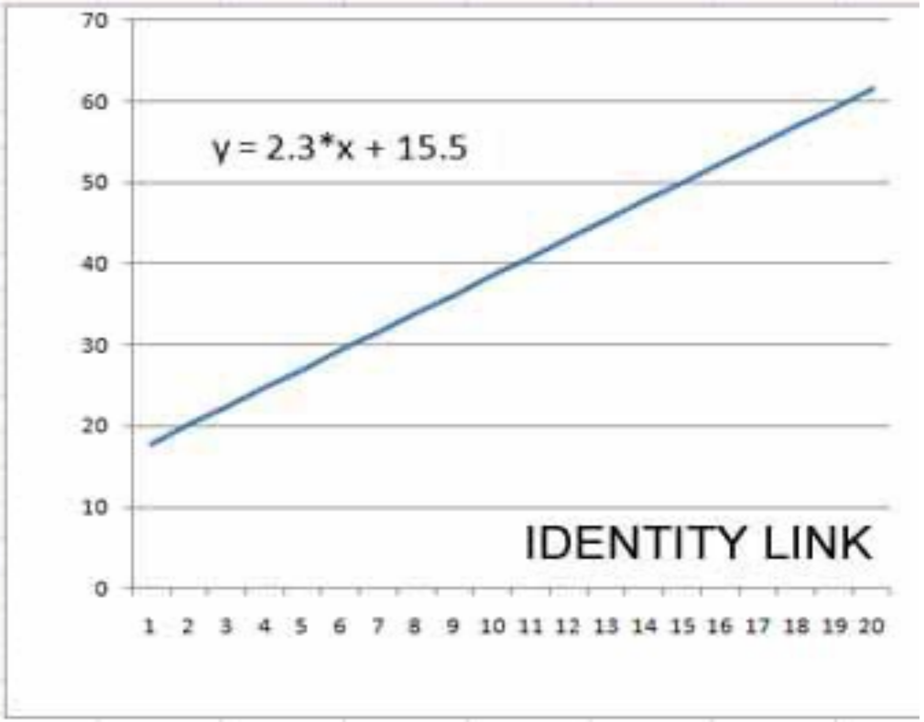
- It is used to **maintain a linear relationship** between the coefficients and predictors on the right-hand side of the model equation and the dependent variable transformed by the link function on the left-hand side of the equation

- **The choices of the link function depend on the natural parameter of the original distribution of the dependent variable** Yi

Generalized Linear Models
Link function

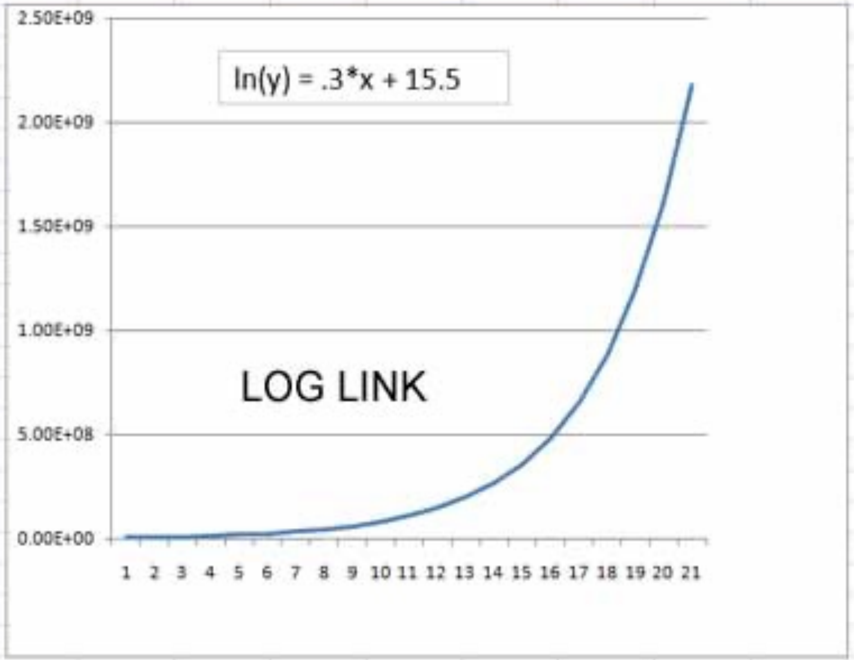
- Normal distribution: Identity function

X	Y
1	17.8
2	20.1
3	22.4
4	24.7
5	27
6	29.3
7	31.6
8	33.9
9	36.2
10	38.5
11	40.8
12	43.1
13	45.4
14	47.7
15	50
16	52.3
17	54.6
18	56.9
19	59.2
20	61.5



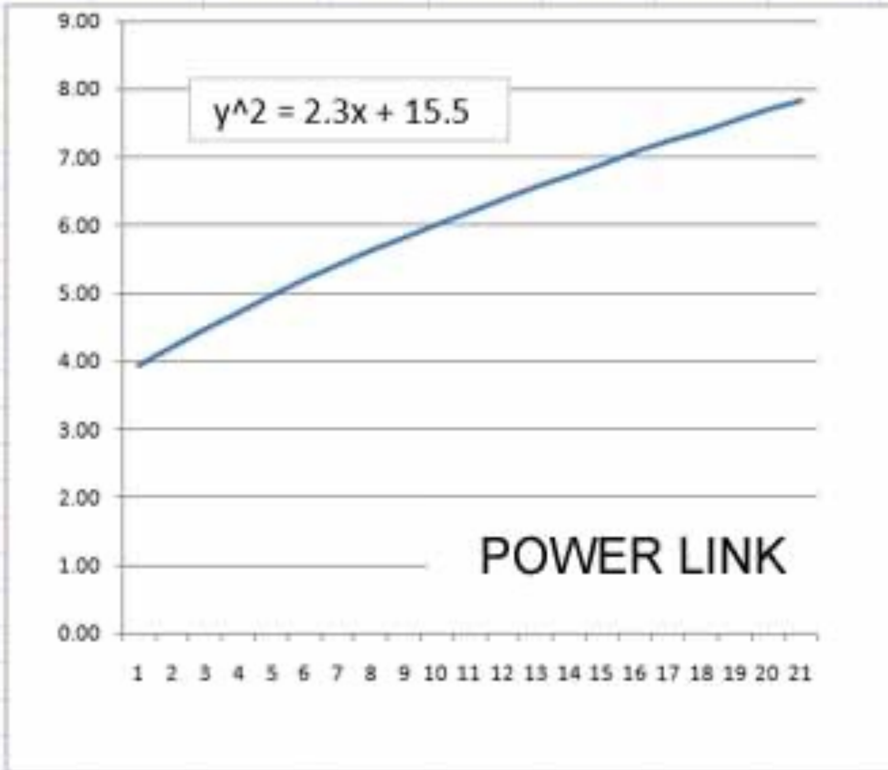
- **Poisson distribution: Log function**
  - **Loglinear models**: assume a Poisson distribution and use a log link function

X	Y
0	5.39E+06
1	7.28E+06
2	9.82E+06
3	1.33E+07
4	1.79E+07
5	2.42E+07
6	3.26E+07
7	4.40E+07
8	5.94E+07
9	8.02E+07
10	1.08E+08
11	1.46E+08
12	1.97E+08
13	2.66E+08
14	3.59E+08
15	4.85E+08
16	6.55E+08
17	8.84E+08
18	1.19E+09
19	1.61E+09
20	2.17E+09



- Many other distributions: Power functions

X	Y
0	3.94
1	4.22
2	4.48
3	4.73
4	4.97
5	5.20
6	5.41
7	5.62
8	5.82
9	6.02
10	6.20
11	6.39
12	6.57
13	6.74
14	6.91
15	7.07
16	7.23
17	7.39
18	7.54
19	7.69
20	7.84



Generalized Linear Models
Link function

FEUP

Common relations between distributions and link functions

Canonical Link Functions			
Distribution	**Name**	**Link Function – η_i 0 $g(\mu_i)$**	**μ_i - Mean (Inverse) Function**
Normal	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential	Inverse	$\mathbf{X}\boldsymbol{\beta} = \mu^{-1}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1}$
Gamma			
Inverse Gaussian	Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Binomial	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Multinomial			

GZLM for count of events
Poisson Regression Model

- **The Poisson Distribution** is commonly used to describe the count of events occurring at random in time or space
- The Poisson condition is that $E[Y_i] = VAR[Y_i] = \lambda_i$ or that
- Examples:
  - The number of cars passing through an intersection during a certain hour
  - The number of calls for emergency ambulance service during a tour of duty
  - The number of fires arising in a neighborhood
  - Number of vehicles waiting in a queue
  - Auto breakdowns in an express way in rush hour
  - Number of heart attack deaths per week in a county
  - Number of homes destroyed by a fire during the summer
  - Number of accidents in a road section or intersection

FEUP

GZLM for count of events
Poisson Regression Model

- The most common relationship between the explanatory and the Poisson parameter is the log-linear model (because the logarithm of this function produces the linear combination of explanatory variables)

$$\lambda_i = e^{(\beta X i)}$$

or

$$Ln(\lambda_i) = \beta X i$$

Inverse function

$$E[Yi] = \lambda_i = \exp\left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta\right)$$

- The expected number of accidents per period is given by

$$E[y_i] = \lambda_i = e^{(\beta X i)}$$

GZLM for count of events
Poisson Regression Model

- Poisson model
  - For the case of count data (e.g., accidents), a variable Q_i is added and corresponds to the unit of exposure (e.g., vehicles per year)
  - It is also referred to as the ***offset* value**

$$E[Y_i] = Q_i \times exp\left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right)$$

GZLM for count of events
Poisson Regression Model - MLE

- Estimation by standard maximum likelihood methods, with the likelihood function given as

$$L(\beta) = \prod \frac{EXP[EXP(\beta X_i)][EXP(\beta X_i)]^{y_i}}{y_i!}$$

or

$$LL(\beta) = \sum_{i=1}^{n} [-EXP(\beta X_i) + y_i \beta X_i - LN(y_i!)]$$

- Maximum likelihood estimates produce Poisson parameters witch are consistent, asymptotically normal, and asymptotically efficient

GZLM for count of events
Poisson Regression Model

- Elasticity is computed on the parameter estimation to evaluate the marginal effects of the independent variables
  - Effect of a 1% change in the variable on the expected frequency λ_i
  - Computed for each observation and then a single average is reported
  - **Continuous variables**

$$E_{x_{ik}}^{\lambda_i} = \frac{\delta\lambda_i}{\lambda_i} \times \frac{x_{ik}}{\delta x_{ik}} = \beta_k x_{ik}$$

  - Count data

$$E_{xik}^{\lambda i} = \frac{EXP(\beta_k) - 1}{EXP(\beta_k)}$$

GZLM for count of events
Negative Binomial Regression

FEUP

- When the Poisson condition is violated (i.e., $E[Y_i] \neq VAR[Y_i]$), two situations can occur:
  - $E[Y_i] > VAR[Y_i]$ (dispersed)
  - $E[Y_i] < VAR[Y_i]$ (over dispersed)
- As such, the link function is rewritten
  - from $\lambda_i = EXP(\beta Xi)$ for each observation , with

$$\lambda_i = EXP(\beta Xi + \varepsilon i)$$

➢ where ε_i is the dispersion term

GZLM for count of events
Negative Binomial Regression

- Therefore the variance differ from the mean through the addition of a quadratic term to the variance that represents over dispersion

$$\text{var}(Y_i) = \lambda_i + K(\lambda_i)^2$$

- The Poisson model is regarded as a limited model of the negative binomial as K approaches 0.
- This K parameter is called the **over dispersion** parameter.

GZLM for count of events

Negative Binomial Regression- MLE

- Negative Binomial pdf

$$P(y_i) = \frac{\Gamma\left(y_i + \frac{1}{K}\right)}{y_i!\,\Gamma\frac{1}{K}} \left(\frac{K\lambda_i}{(1+K\lambda_i}\right)^{y_i} \left(\frac{1}{1+K\lambda_i}\right)^{\frac{1}{K}}$$

  - where Γ is a Gamma Function and K is an estimated parameter representative of dispersion

- The corresponding likelihood function is:

$$L(\lambda_i) = \prod \frac{\Gamma\left(y_i + \frac{1}{K}\right)}{y_i!\,\Gamma\frac{1}{K}} \left(\frac{K\lambda_i}{(1+K\lambda_i}\right)^{y_i} \left(\frac{1}{1+K\lambda_i}\right)^{\frac{1}{K}}$$

GZLM for count of events
Negative Binomial Regression

❒ Negative Binomial with log link

- Specifies a negative binomial distribution (**with the ancillary K parameter = 1**) with a log link function
  - It is used to modeling count data that violates the Poisson assumption of equality of mean and variance.
  - Also, negative binomial regression is thought to be more stable than Poisson regression for **small datasets**.
  - An error term ξ of Gamma distribution and variance K_2 is added to the Poisson Regression

$$\log(\lambda_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \xi_i$$

GZLM for count of events
Negative Binomial Regression

- Tests on over dispersion - Lagrange Multiplier test (**in SPSS**)
  - The Lagrange multiplier test may be used to **test if a negative binomial model is significantly different from a Poisson model**
  - Since the negative binomial model is the same as the Poisson model when the binomial model's ancillary (dispersion) parameter, K=0, **the Lagrange multiplier test analyses the null hypothesis that K = 0**

- A significant Lagrange test coefficient (i.e., ***p-value > 0,05***) indicates that ***K* cannot be assumed to be different from *0***, and hence a Poisson model would be preferred over a negative binomial model (negative binomial models have one more parameter, *k*)

❐ References

- **McCullagh, Peter; Nelder, John (1989). Generalized Linear Models, Second Edition. Boca Raton: Chapman and Hall/CRC. ISBN 0-412-31760-5.**
- J. B. S. Haldane, "On a Method of Estimating Frequencies", Biometrika, Vol. 33, No. 3 (Nov., 1945), pp. 222–225. JSTOR 2332299
- Hilbe, Joseph M., Negative Binomial Regression, Cambridge, UK: Cambridge University Press (2007) Negative Binomial Regression - Cambridge University Press
- **Washington, Simon P., Karlaftis, Mathew G. e Mannering (2003) Statistical and Econometric Methods for Transportation Data Analysis, CRC**
- Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-Gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention , pp. 35-46.
- Fernandes, A. (2010) Programas de manutenção de características da superfície de pavimentos associados a critérios de segurança rodoviária. Tese de Doutoramento em Engenharia Civil. Instituto Superior Técnico, Universidade Técnica de Lisboa.

FEUP