# Phd Program in Transportation

## Transport Demand Modeling

Filipe Moura

# Generalized Linear Models – Part 2

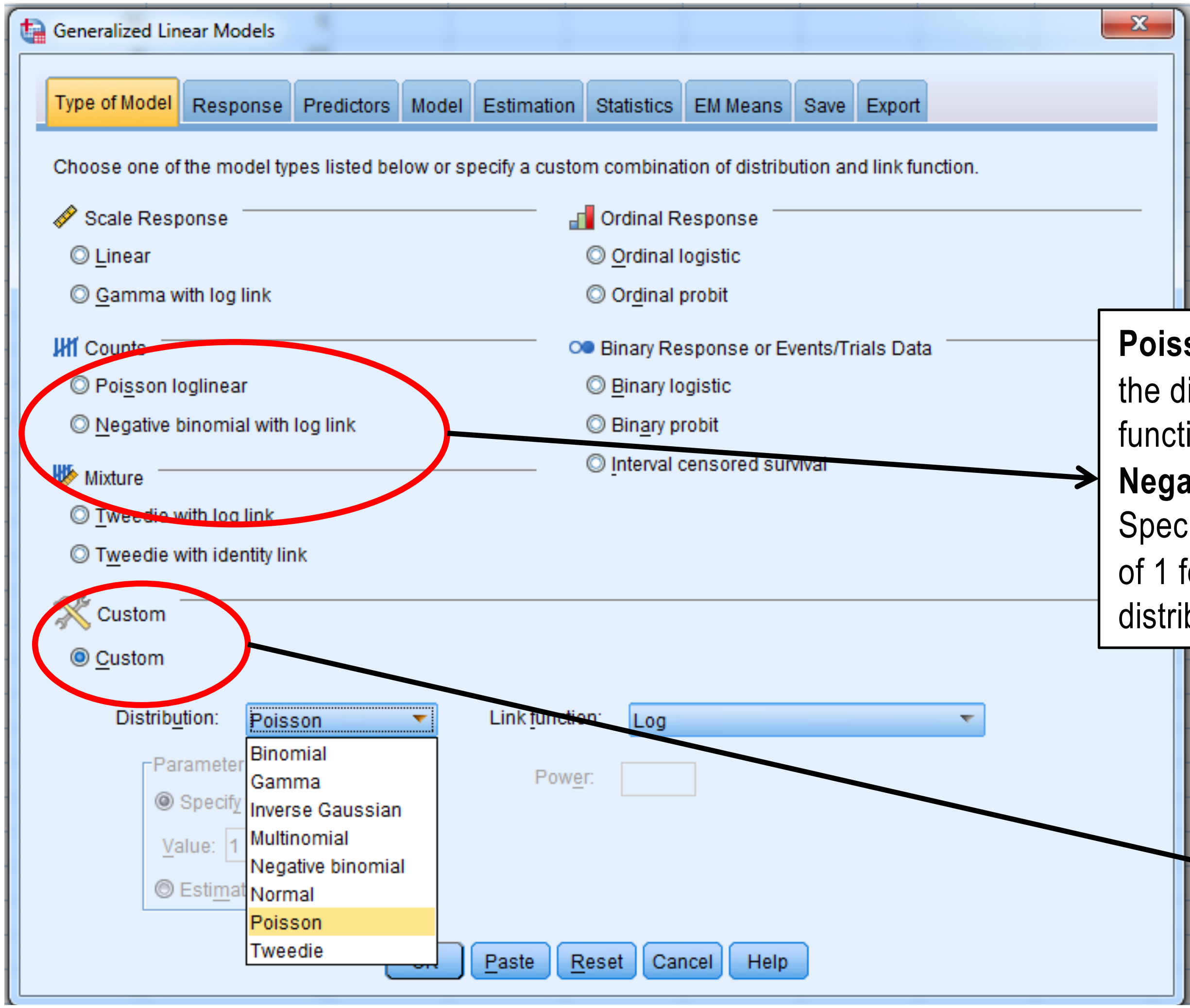
- Analysis steps
  - Model Formulation
  - Model Adjusment
  - Model Selection and Validation

- Model Formulation

  - **Random Component** - Dependent Variable (distributed as a Poisson or Negative Binomial)

  - **Systematic Component** - Independent variables (explaining the dependent variable)

  - **Link or connection function** (logarithmic)

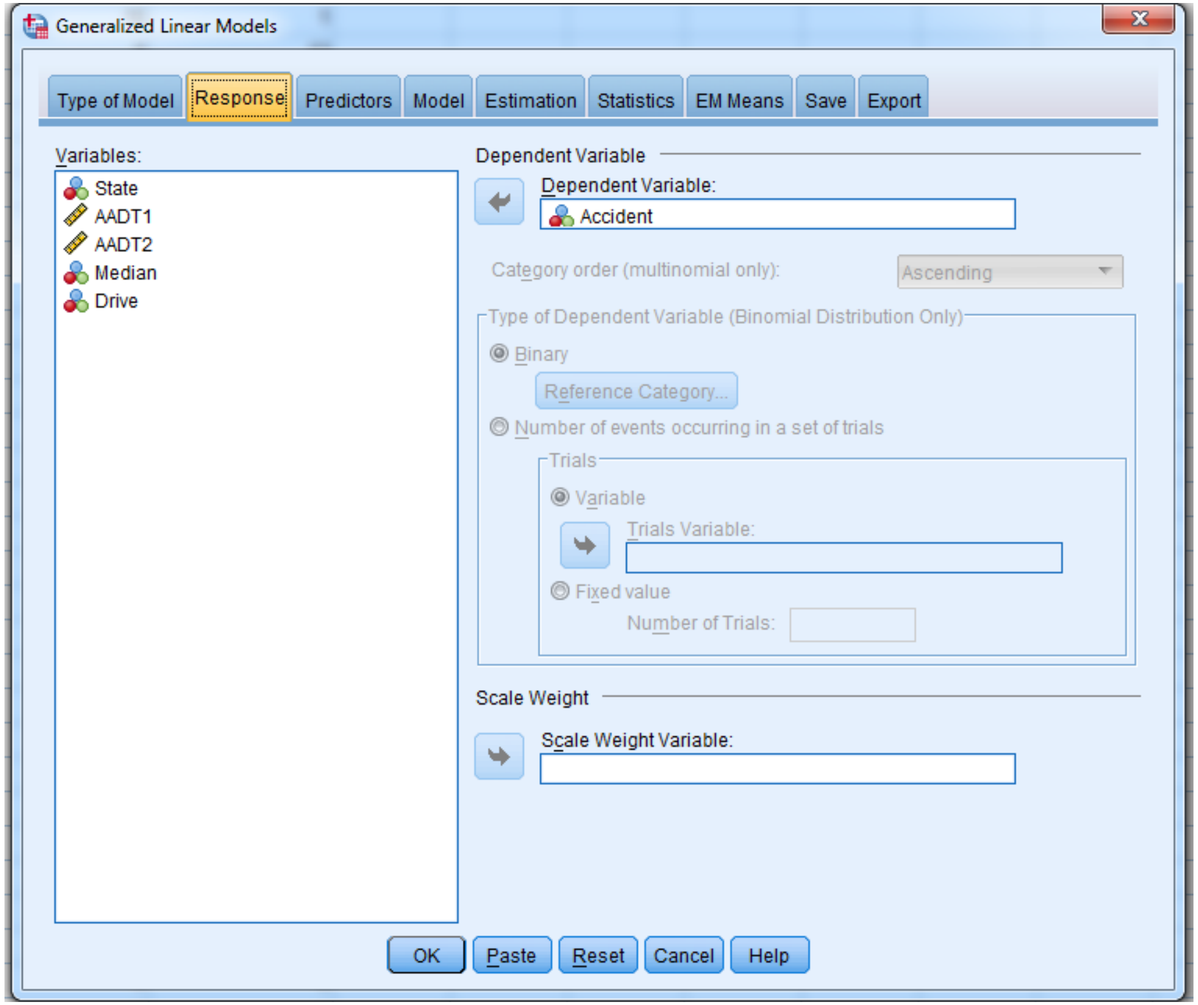



The type of model could be selected among a series of model types

**Poisson loglinear**-Specifies Poisson as the distribution and Log as the link function.
**Negative binomial with log link**. Specifies Negative binomial (with a value of 1 for the ancillary parameter) as the distribution and Log as the link function.

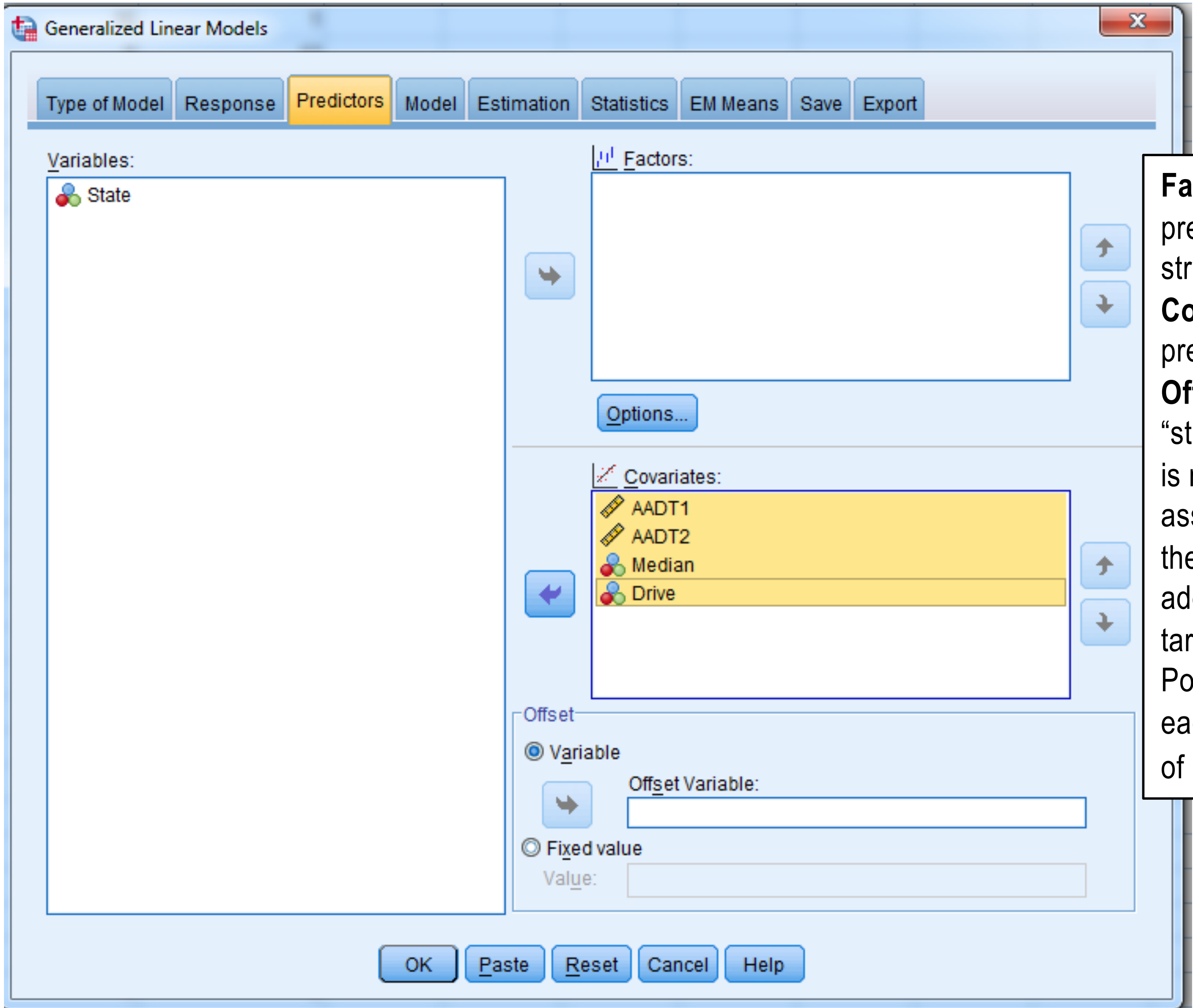
The custom tool allows the selection of specific models (specific distribution) together with a specific link function

# GZLM
# Count Data




The dependent variable is defined here

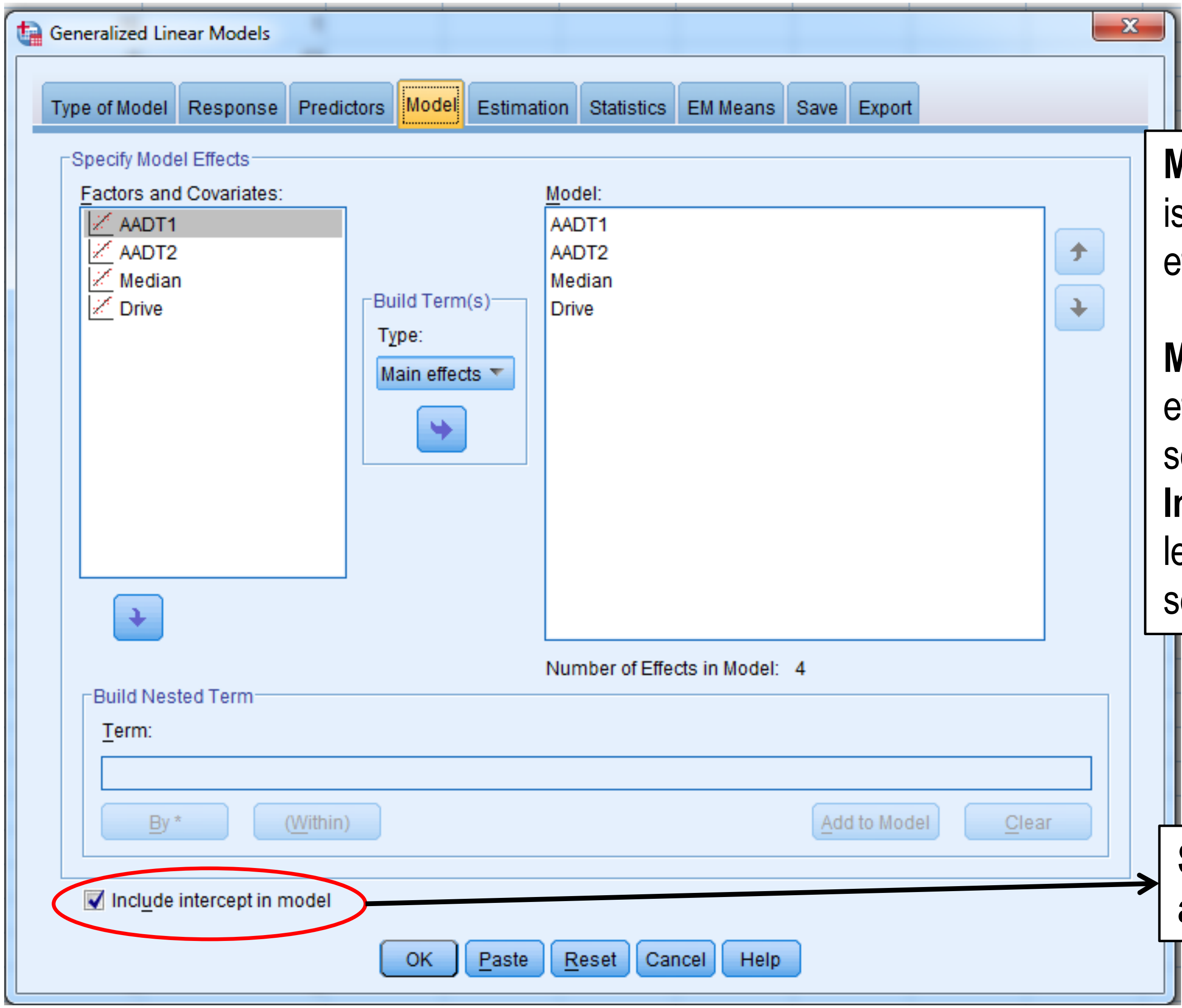
# GZLM
# Count Data




**Factors** - Factors are categorical predictors; they can be numeric or string.
**Covariates** - Covariates are scale predictors; they must be numeric
**Offset** - The offset term is a "structural" predictor. Its coefficient is not estimated by the model but is assumed to have the value 1; thus, the values of the offset are simply added to the linear predictor of the target. This is especially useful in Poisson regression models, where each case may have different levels of exposure to the event of interest.

When modeling accident rates for individual drivers, there is an important difference between a driver who has been at fault in one accident in three years of experience and a driver who has been at fault in one accident in 25 years! The number of accidents can be modeled as a Poisson or negative binomial response with a log link if the natural log of the experience of the driver is included as an offset term.

# GZLM
# Count Data




**Model Effects** - The default model is intercept-only, the other model effects must be explicitly specified

**Main effects** - Creates a main-effects term for each variable selected.

**Interaction** - Creates the highest-level interaction term for all selected variables.

**Selecting a model with an intercept term**

## ❒ Model Adjustment

- **Maximum likelihood method (to estimate variables' coefficients and dispersion parameter φ)**
  - Interactive computational estimation method:
    1. For the exponential family

$$f(y_i \mid \theta_i, \varphi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\varphi)} + c(y_i, \varphi)\right\}, y_i \in \Re$$

    1. The Log of Maximum likelihood estimation is given by

$$L(\vec{\theta}, \varphi; y) = \sum_{i=1}^{N} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\varphi)} + c(y_i, \varphi)\right\}$$

- Model Adjustment – Variable Coefficients
  - **Maximum likelihood method** maximizes the likelihood function $Y_i$ in relation to $\beta_j$, and therefore it allows to determine the absolute maximum (since the logarithmic function is monotonic and growing) . We must then solve the system of equations $S(\theta_i)=0$, for coefficient.

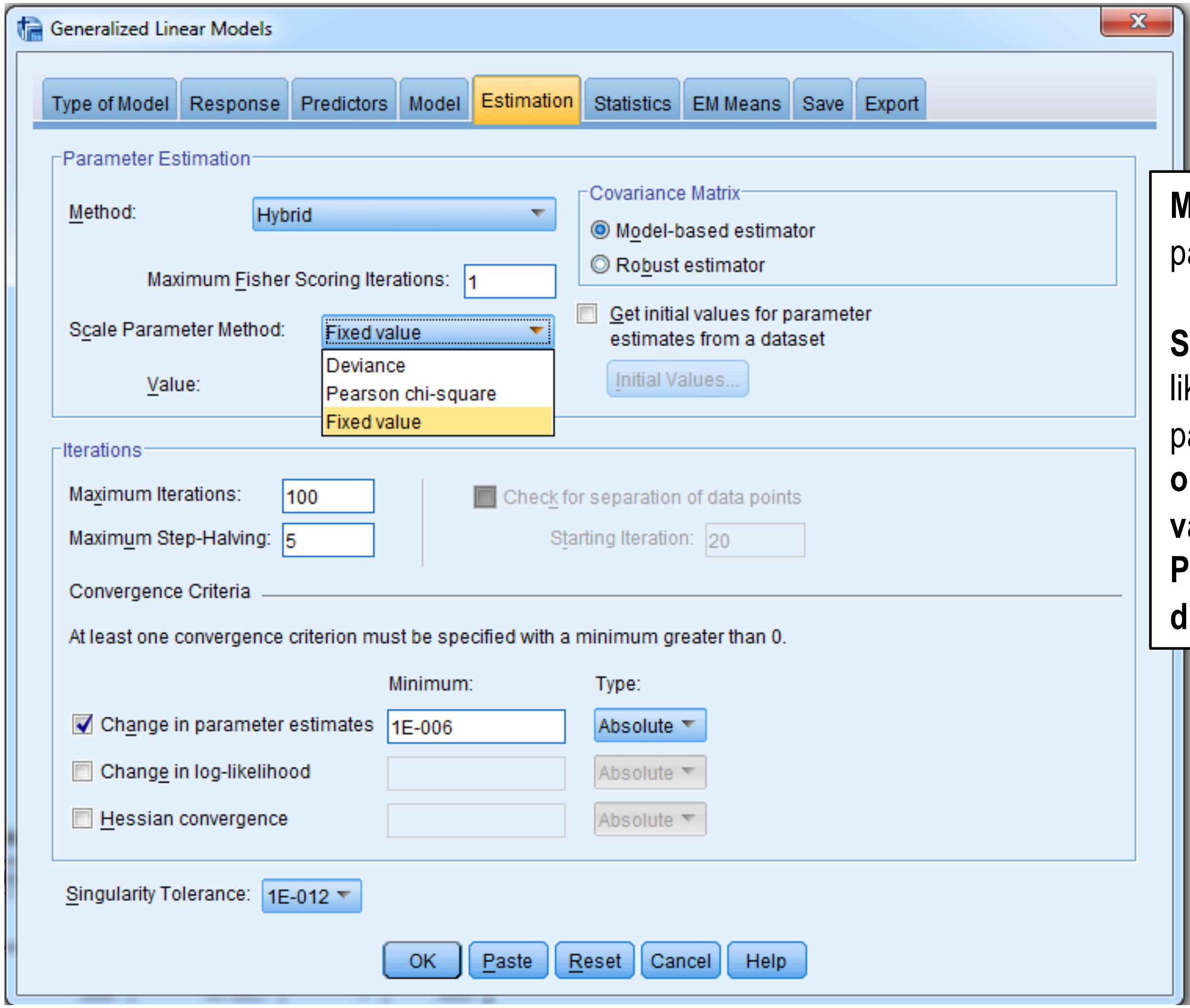
$$S(\theta_i) = \frac{\partial L(\vec{\theta}, \varphi; y)}{\partial \beta_j}$$

    - Since it is a system of non linear equations it must be estimated iteratively. The methods are:
      - Newton-Raphson
      - Fisher-Scoring
      - Hybrid (Fisher on a set of initial iterations and than changed to Newton)

## ❒ Model Adjustment – Scale Parameter φ

- The **scale parameter φ** has a different nature then vector β
  - β has a direct influence on the $\lambda_i$ – expected value of variable $Y_i$ – and the parameter **φ** reveals the data dispersion of the data
  - **On some exponential families such as Poisson, the parameter φ is fixed and not estimated**
  - On other distributions **φ** must be estimated through maximum likelihood log for the $Y_i$ vector, by a derivate in order to φ and being equal to zero.

# GZLM
# Count Data




**Method** - Estimation methods for the parameters could be selected here

**Scale parameter method** - Maximum-likelihood jointly estimates the scale parameter with the model effects. **This option is not valid if the response variable has a negative binomial, Poisson, binomial, or multinomial distribution**.

## ❐ Model Adjustment - Scale Parameter φ

❖ Estimated through '**Deviance**' $\boldsymbol{\varphi_D}$

$$\varphi_D = \frac{D}{N-p} = \frac{2(L^c - L^m)}{N-p}$$

*where*

- $L^c$ is the maximum likelihood log of the complete model (with all the variables)
- $L^m$ is the maximum likelihood log of the model under analysis
- If *Deviance* is higher than *N-p,* the model is 'over-dispersed'
- *N* observations (e.g., road segments) and *p* variables
- *D* is the *Deviance*

## ❒ Model Adjustment - Scale Parameter φ

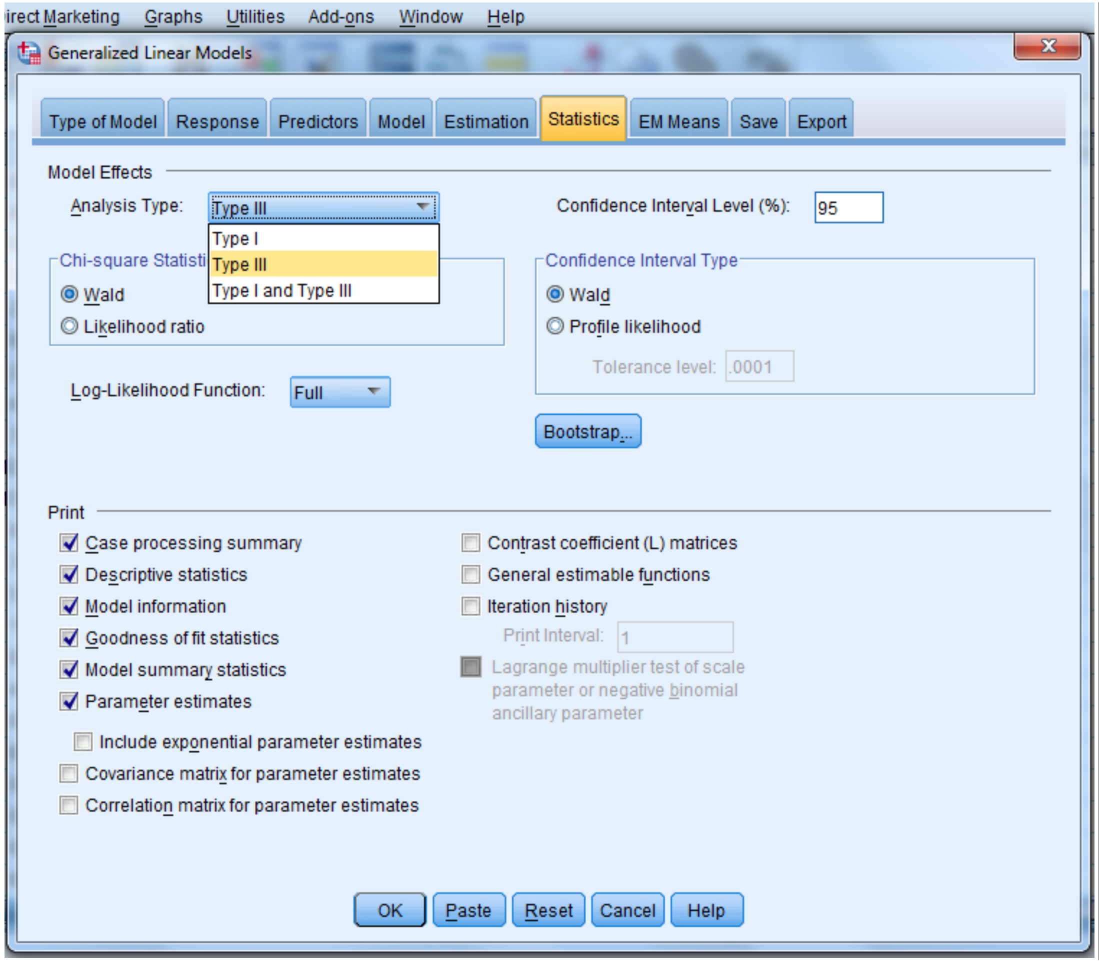
- Or through the statistic '**$\chi^2$ of Pearson' ($\varphi_{\chi 2}$)**

$$\varphi_{\chi 2} = \frac{\chi^2}{N-p} = \frac{1}{N-p}\sum_{i=1}^{N}\frac{(y_i - \hat{y}_i)^2}{\mathrm{var}(\hat{y}_i)}$$

where

- $\chi^2$ is the statistic of Pearson
- If $\chi^2$ is superior to *N-p* the model is 'over-dispersed'
- *N* observations (e.g., road segments) and *p* variables

➢ **Both should be close to 1 in order to use Poisson Regression**

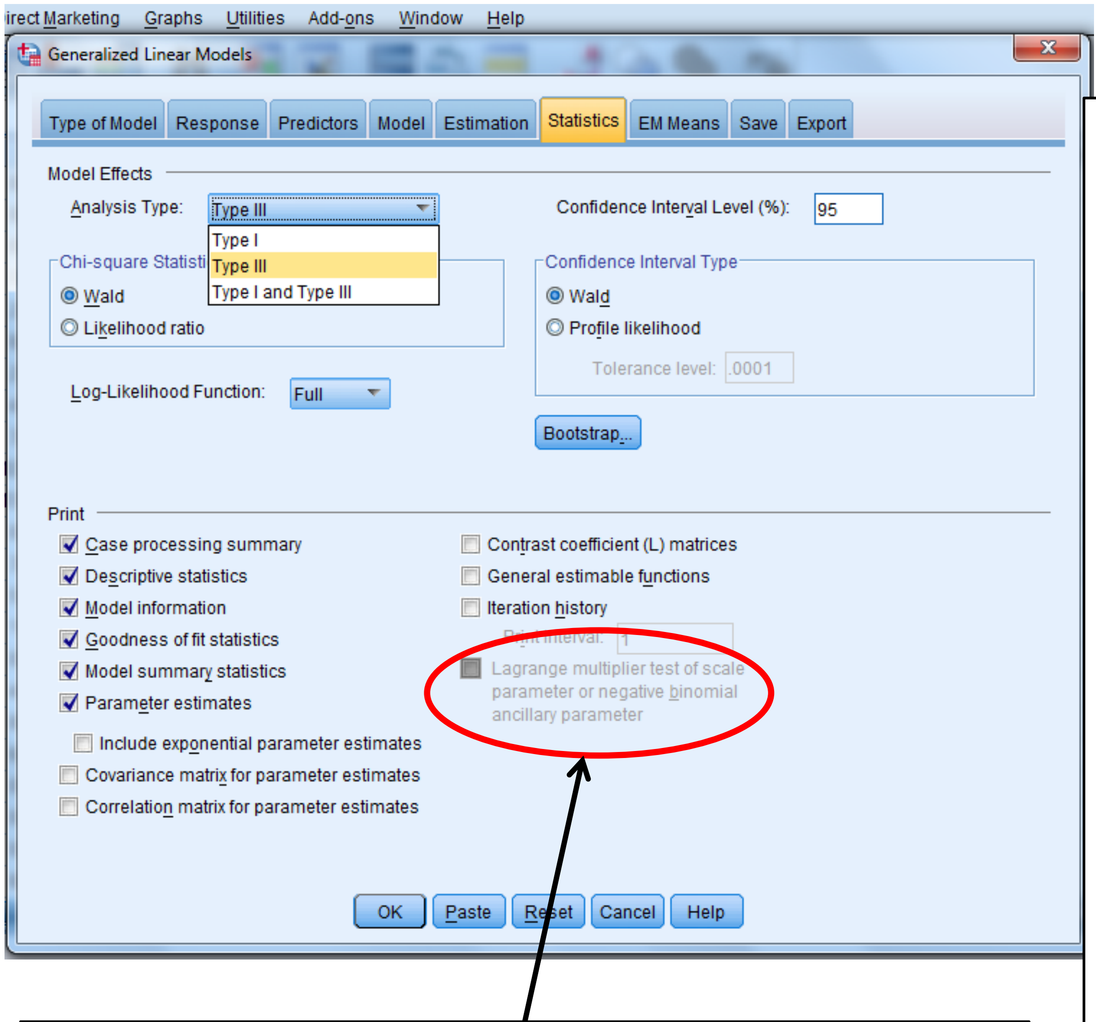



**Analysis type** - Type I analysis is generally appropriate when there are a priori reasons for ordering predictors in the model. Type III is more generally applicable. The chi squared statistics could be estimated either using Wald or likelihood-ratio.

**Confidence intervals** - Wald intervals are based on the assumption that parameters have an asymptotic normal distribution; profile likelihood intervals are more accurate but can be computationally expensive. The tolerance level is the criteria used to stop the iterative algorithm used to compute the intervals.

**Log-likelihood function** -This controls the display format of the log-likelihood function. The full function includes an additional term that is constant with respect to the parameter estimates; it has no effect on parameter estimation.

# GZLM
# Count Data




**Lagrange multiplier test** - Lagrange multiplier test statistics for assessing the validity of a scale parameter that is computed using the deviance or Pearson chi-square. For the negative binomial distribution, this tests the fixed ancillary parameter.

## Print

**Case processing summary** - number and percentage of cases included and excluded from the analysis and the Correlated Data Summary table.

**Descriptive statistics** - descriptive statistics and summary information about the dependent variable, covariates, and factors.

**Model information** - dataset name, dependent variable or events and trials variables, offset variable, scale weight variable, probability distribution, and link function.

**Goodness of fit statistics** - Deviance and scaled deviance, Pearson chi-square and scaled Pearson chi-square, log-likelihood, Akaike's information criterion (AIC), finite sample corrected AIC (AICC), Bayesian information criterion (BIC), and consistent AIC (CAIC).

**Model summary statistics** - likelihood-ratio statistics for the model fit omnibus test and statistics for the Type I or III contrasts for each effect.

**Parameter estimates** - Displays parameter estimates and corresponding test statistics and confidence intervals. In addiction it can optionally display exponentiated parameter estimates.

- Model Selection and Validation
  - Over-dispersion of data should be the first analysis to be perform in order to evolve over Poisson distribution
    - Maximum Likelihood Ratio and Lagrange Tests
  - Statistical significance of the parameters should be verified
    - Wald test and p-values
  - The predictive capacity should be analysed
    - Omnibus test (for improvement of the restricted model); Pseudo $R^2$
  - Comparison between models with different specifications or different distributions of the Yi
    - Improvement of the log maximum likelihood together with AIC/AICC/BIC/CAIC

- Model Selection and Validation - **Maximum likelihood ratio**
  - This test analyses the equality between the mean and the variance through Poisson Regression Standard against the alternative of the variance exceeding the mean (Negative Binomial)
  - The corresponding hypothesis test can be formulated as the over dispersion parameter K (sometimes $\alpha$ in the literature and software):
    - H0:K=0
    - H1:K≥0
  - The test is performed by calculating the corresponding $X^2$ statistic with

$$X^2 \sim -2[L(P) - L(NB)]$$

where $X^2$ follows a $\chi^2$ distribution

  - If *p* value is below 0.05 than the null hypothesis is rejected and over-dispersion is than identified (mean ≠ variance), recommending for the negative binomial
  - Note: Overdispersed Poisson regression can also be tested where a scale parameter is admissible

- Model Selection and Validation - **Lagrange tests**
  - Likewise, Lagrange test on K detects the over-dispersion of data around the mean
  - Again, the hypothesis test can be formulated as:
    - H0:K=0
    - H1:K $\geq$ 0
      - If the $\chi 2$ statistic is non-significant (i.e., p<0.05) then there is over-dispersion and the Negative Binomial is more adequate
      - If it is significant (i.e., p>0.05) then there is no over-dispersion, the mean is equal to the variance and the Poisson distribution is recommended
      - Note: Overdispersed Poisson regression can also be tested where a scale parameter in admissible
  - It is often the case that over-dispersion is related with excess of zeros:
    - The solution is opting for **Zero Inflated Poisson**
      - Note: not possible without the presence of zero accidents segments

- Model Selection and Validation
  - Testing for the statistical signifcance of each coeficient β
  - Assimptotical test or **Wald Test**

$$WS = \frac{(\beta_j)^2}{\mathrm{var}(\beta_j v)}$$

where the hypothesis test is:

$$H_0: \hat{\beta}_j = 0$$
$$H_a: \hat{\beta}_j \neq 0$$

  - For low *p* values (i.e., below 0,05), the null hypothesis is rejected and the variable is influent in the model

- Model Selection and Validation
  - Omnibus test calculated with the statistic

$$X^2 = -2[LL(\beta_R) - LL(\beta_U)]$$

where $X^2$ follows a $\chi^2$ distribution

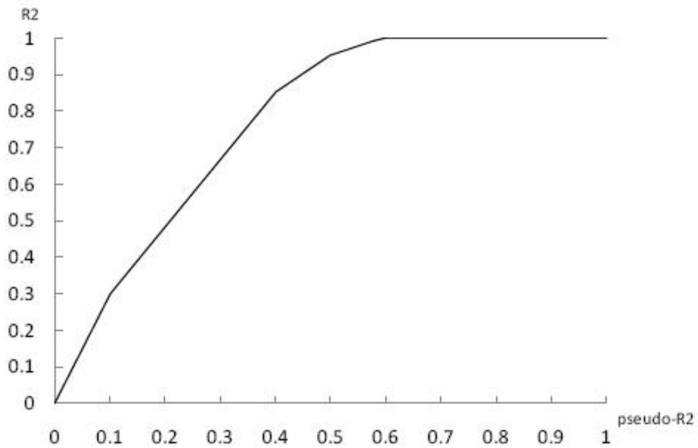
  - If significant (i.e., p-value <0,05), then the estimated model is better than the null model (i.e., model with constant only)
    - $LL(\beta_U)$ is the log likelihood of the unrestricted model
    - $LL(\beta_R)$ is the log likelihood of the restricted (or null) model (without independent variables)
    - Note: degrees of freedmon are equal to the diference between the number of parameters in the **restricted** and **unrestricted** model

## ❒ Model Selection and Validation

- With the the values obtained with the previous testes, the *LL(β$_U$)* and *LL(β$_R$)* of the unrestricted and restricted model, respectively, it possible to calculate the pseudo r-square (rho-square) comparable to the linear model's r-square
- Pseudo r-square is calculated as follows:

$$\rho^2 = 1 - \frac{LL(\beta_U)}{LL(\beta_R)}$$

- The value of the Pseudo $R^2$ can be compared with the linear models $R^2$ through the empirical relation set by *Domencich and Macfaden (1975)*



## ❒ Model Selection and Validation

- Other information criteria to compare models:
  - AIC: $AIC = -2L(\hat{\beta}) + 2p^*$
  - AICC (for finite samples): $AICC = -2L(\hat{\beta}) + \frac{2p^* \times N}{N - p^* - 1}$
  - BIC: $BIC = -2L(\hat{\beta}) + p^* \times \ln(N)$
  - CAIC: $CAIC = -2L(\hat{\beta}) + p^* \times (\ln(N) + 1)$

# AIC – Akaike Information Criteria

- It is is an estimator of **the relative quality of statistical models** for a given set of data.
- Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, **AIC provides a means for model selection**.
- AIC estimates the **relative information lost** by a given model: the less information a model loses, the higher the quality of that model.
  - AIC deals with the **trade-off between the goodness of fit** of the model and the **simplicity of the model**
  - Given a set of candidate models for the data, the **preferred model is the one with the minimum AIC value**.
  - AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The **penalty discourages overfitting**, because increasing the number of parameters in the model almost always improves the goodness of the fit.

FEUP

# BIC – Bayesian Information Criteria

- It is **similar to the formula for AIC**, but with a different penalty for the number of parameters.
  - With AIC the penalty is 2k, whereas with BIC the penalty **is ln(n) k**.
- It is interpreted in the same way, i.e. the **minimum BIC value indicates the preferred model**.
- Comparing AIC with BIC:
  - Different opinions on which to chose and when
  - Some authors argue that BIC is best at indicating “the true model” (that, ultimately, never exists) and is better for **forecasting models**
  - AIC would be preferred for **explanatory models**

# GZLM
# Count Data

**Goodness of Fit[a]**

| | Value | df | Value/df |
|---|---|---|---|
| Deviance | 176.540 | 79 | 2.235 |
| Scaled Deviance | 176.540 | 79 | |
| Pearson Chi-Square | 186.482 | 79 | 2.361 |
| Scaled Pearson Chi-Square | 186.482 | 79 | |
| Log Likelihood[b] | -169.260 | | |
| Akaike's Information Criterion (AIC) | 348.519 | | |
| Finite Sample Corrected AIC (AICC) | 349.288 | | |
| Bayesian Information Criterion (BIC) | 360.673 | | |
| Consistent AIC (CAIC) | 365.673 | | |

Dependent Variable: Accident
Model: (Intercept), AADT1, AADT2, Median, Drive

a. Information criteria are in small-is-better form.

b. The full log likelihood function is displayed and used in computing information criteria.

**Omnibus Test[a]**

| Likelihood Ratio Chi-Square | df | Sig. |
|---|---|---|
| 153.851 | 4 | .000 |

Dependent Variable: Accident
Model: (Intercept), AADT1, AADT2, Median, Drive

a. Compares the fitted model against the intercept-only model.

$$X^2 = -2[LL(\beta_R) - LL(\beta_U)]$$

The Omnibus test verifies if the explained variance is significantly greater than the unexplained variance

Deviance compares the given model with the full model (the full model has one parameter for each observation, therefore has a perfect fit). The deviance in a perfect fit model is 0.
The deviance could be used to have information about over dispersion or not (testing if H0: K=0).
In the present case, we reject that hypothesis since the deviance value is higher than the $X^2_{critical}$, therefore the p-value is 0,00. When the Value/df >1, there is a sign of over dispersion

**Tests of Model Effects**

| Source | Type III Wald Chi-Square | df | Sig. |
|---|---|---|---|
| (Intercept) | 12.756 | 1 | .000 |
| AADT1 | 47.602 | 1 | .000 |
| AADT2 | 54.560 | 1 | .000 |
| Median | 7.450 | 1 | .006 |
| Drive | 20.639 | 1 | .000 |

Dependent Variable: Accident
Model: (Intercept), AADT1, AADT2, Median, Drive

Type III tests examine the significance of each partial effect, that is, the significance of an effect with all the other effects in the model. The chi-squared is a likelihood ratio for testing the significance of the effect added to the model containing all of the other effects

**Parameter Estimates**

| Parameter | B | Std. Error | 95% Wald Confidence Interval | | Hypothesis Test | | |
|---|---|---|---|---|---|---|---|
| | | | Lower | Upper | Wald Chi-Square | df | Sig. |
| (Intercept) | -.826 | .2312 | -1.279 | -.373 | 12.756 | 1 | .000 |
| AADT1 | 8.122E-005 | 1.1771E-005 | 5.814E-005 | .000 | 47.602 | 1 | .000 |
| AADT2 | .001 | 7.4400E-005 | .000 | .001 | 54.560 | 1 | .000 |
| Median | -.060 | .0220 | -.103 | -.017 | 7.450 | 1 | .006 |
| Drive | .075 | .0165 | .043 | .107 | 20.639 | 1 | .000 |
| (Scale) | 1[a] | | | | | | |

Dependent Variable: Accident
Model: (Intercept), AADT1, AADT2, Median, Drive

a. Fixed at the displayed value.

Wald test for statistical inference of β coefficients for the independent variables

# Goodness of fit

Goodness of Fit[a]

| | Value | df | Value/df |
|---|---|---|---|
| Deviance | 176.540 | 79 | 2.235 |
| Scaled Deviance | 176.540 | 79 | |
| Pearson Chi-Square | 186.482 | 79 | 2.361 |
| Scaled Pearson Chi-Square | 186.482 | 79 | |
| Log Likelihood[b] | -169.260 | | |
| Akaike's Information Criterion (AIC) | 348.519 | | |
| Finite Sample Corrected AIC (AICC) | 349.288 | | |
| Bayesian Information Criterion (BIC) | 360.673 | | |
| Consistent AIC (CAIC) | 365.673 | | |

Dependent Variable: Accident
Model: (Intercept), AADT1, AADT2, Median, Drive

a. Information criteria are in small-is-better form.

b. The full log likelihood function is displayed and used in computing information criteria.

Goodness of Fit[a]

| | Value | df | Value/df |
|---|---|---|---|
| Deviance | 330,391 | 83 | 3,981 |
| Scaled Deviance | 330,391 | 83 | |
| Pearson Chi-Square | 358,073 | 83 | 4,314 |
| Scaled Pearson Chi-Square | 358,073 | 83 | |
| Log Likelihood[b] | -246,185 | | |
| Akaike's Information Criterion (AIC) | 494,370 | | |
| Finite Sample Corrected AIC (AICC) | 494,418 | | |
| Bayesian Information Criterion (BIC) | 496,800 | | |
| Consistent AIC (CAIC) | 497,800 | | |

Dependent Variable: Accident
Model: (Intercept)
a. Information criteria are in small-is-better form.
b. The full log likelihood function is displayed and used in computing information criteria.

The Omnibus test could be used to estimate the pseudo r-square:

$$\rho^2 = 1 - \frac{LL(\beta_u)}{LL(\beta_r)} = 1 - \frac{-169{,}260}{-246{,}185} = 0{,}312$$

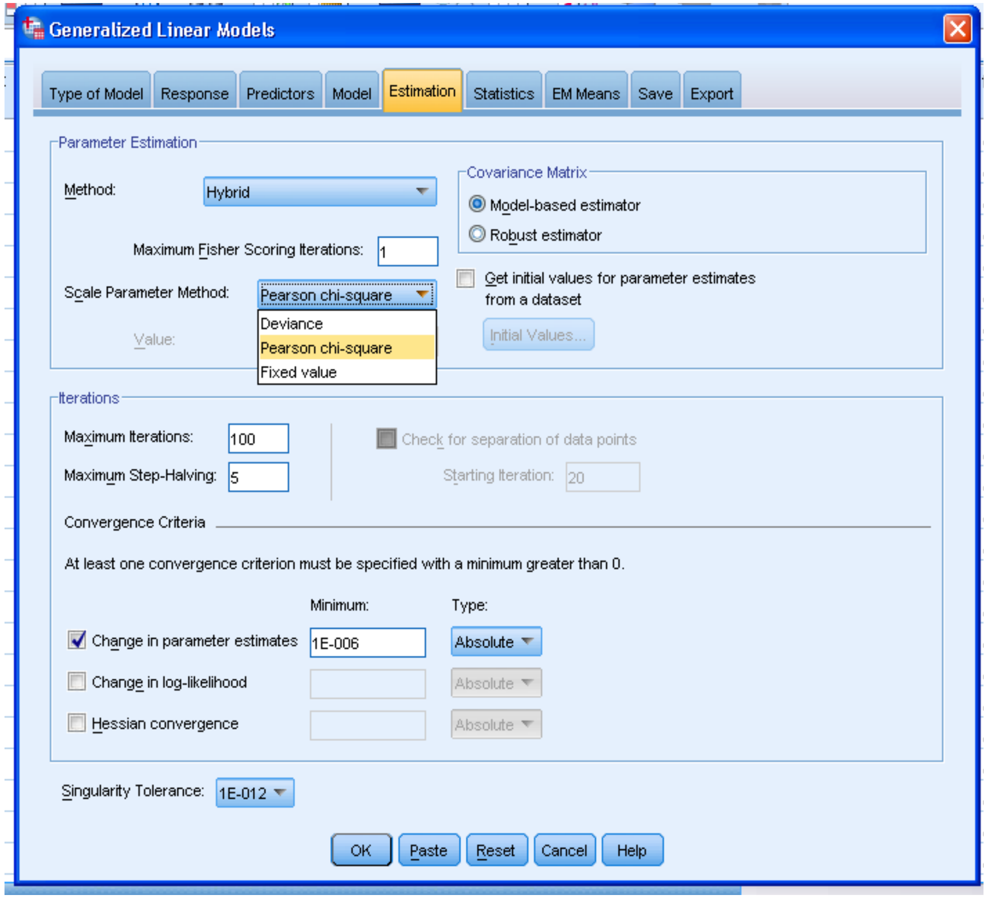
- It is possible to estimate the $LL(\beta_r)$ of the restricted model (with only the constant), by running a new model retrieveing the covariates and calculating the intercept only.

# Over dispersed Poisson

- Since there is an indication for **overdispersion**, two other models must be tested
  - **Overdispersed Poisson regression** (where a scale parameter in admissible)
  - **Negative Binomial**

# Over dispersed Poisson






**The main difference with the Poisson Regression Model is that the scale parameter is estimated and not fixed.**
**The Pearson Chi-squared method is used to estimate the Scale Parameter**

**The scale parameter has a different nature then vector β of coefficients**
**β has a direct influence on the expected value of variable Yi, and the parameter reveals the data dispersion**

# Over dispersed Poisson

**Parameter Estimates**

| Parameter | B | Std. Error | 95% Wald Confidence Interval | | Hypothesis Test | | |
|---|---|---|---|---|---|---|---|
| | | | Lower | Upper | Wald Chi-Square | df | Sig. |
| (Intercept) | -,826 | ,3553 | -1,522 | ,130 | 5,404 | 1 | ,020 |
| AADT1 | 8,122E-005 | 1,8086E-005 | 4,577E-005 | ,000 | 20,166 | 1 | ,000 |
| AADT2 | ,001 | ,0001 | ,000 | ,001 | 23,114 | 1 | ,000 |
| Median | -,060 | ,0338 | -,126 | ,006 | 3,156 | 1 | ,076 |
| Drive | ,075 | ,0253 | ,025 | ,124 | 8,743 | 1 | ,003 |
| (Scale) | 2,361[a] | | | | | | |

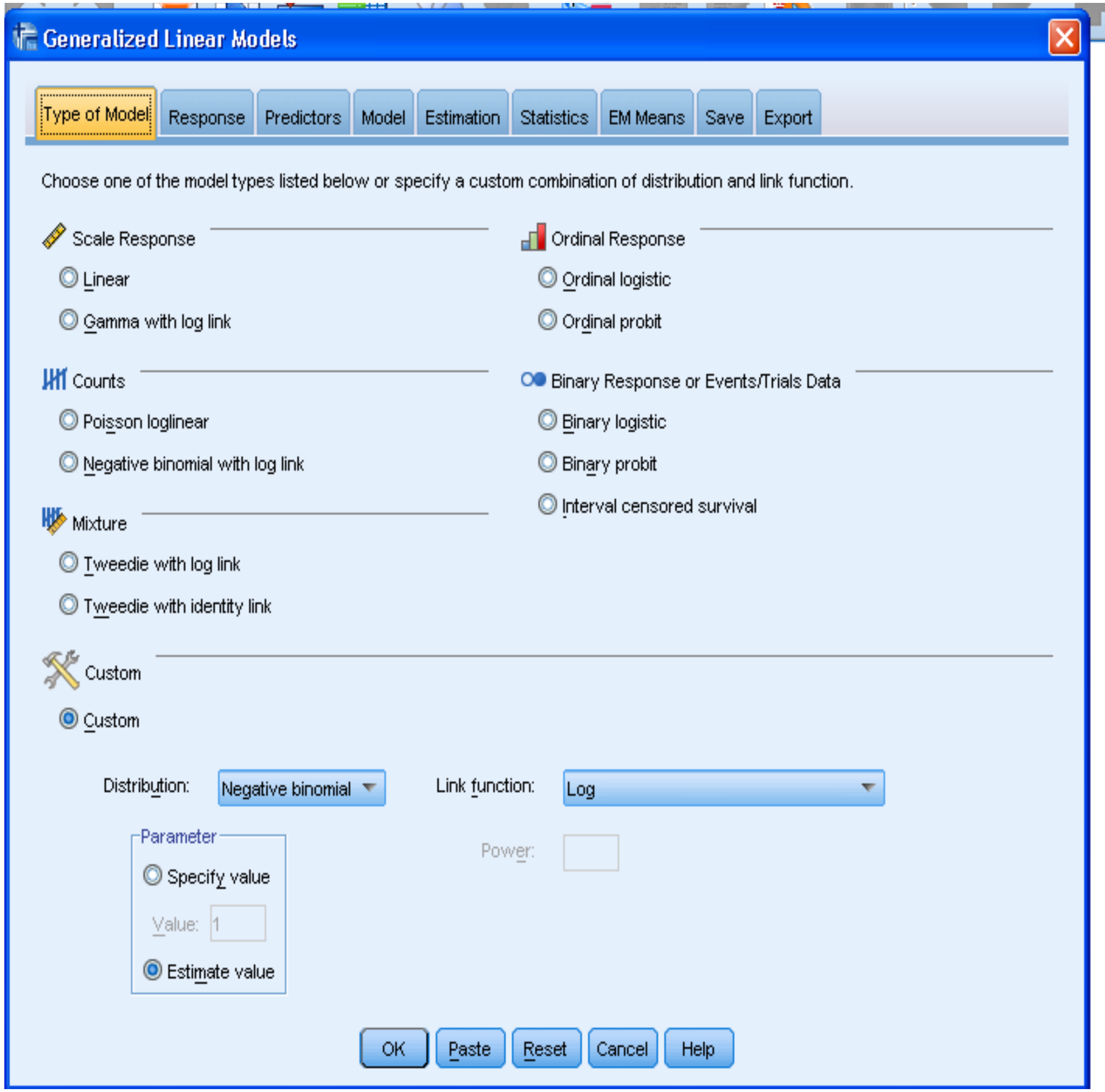
Dependent Variable: Accident

Model: (Intercept), AADT1, AADT2, Median, Drive

a. Computed based on the Pearson chi-square.

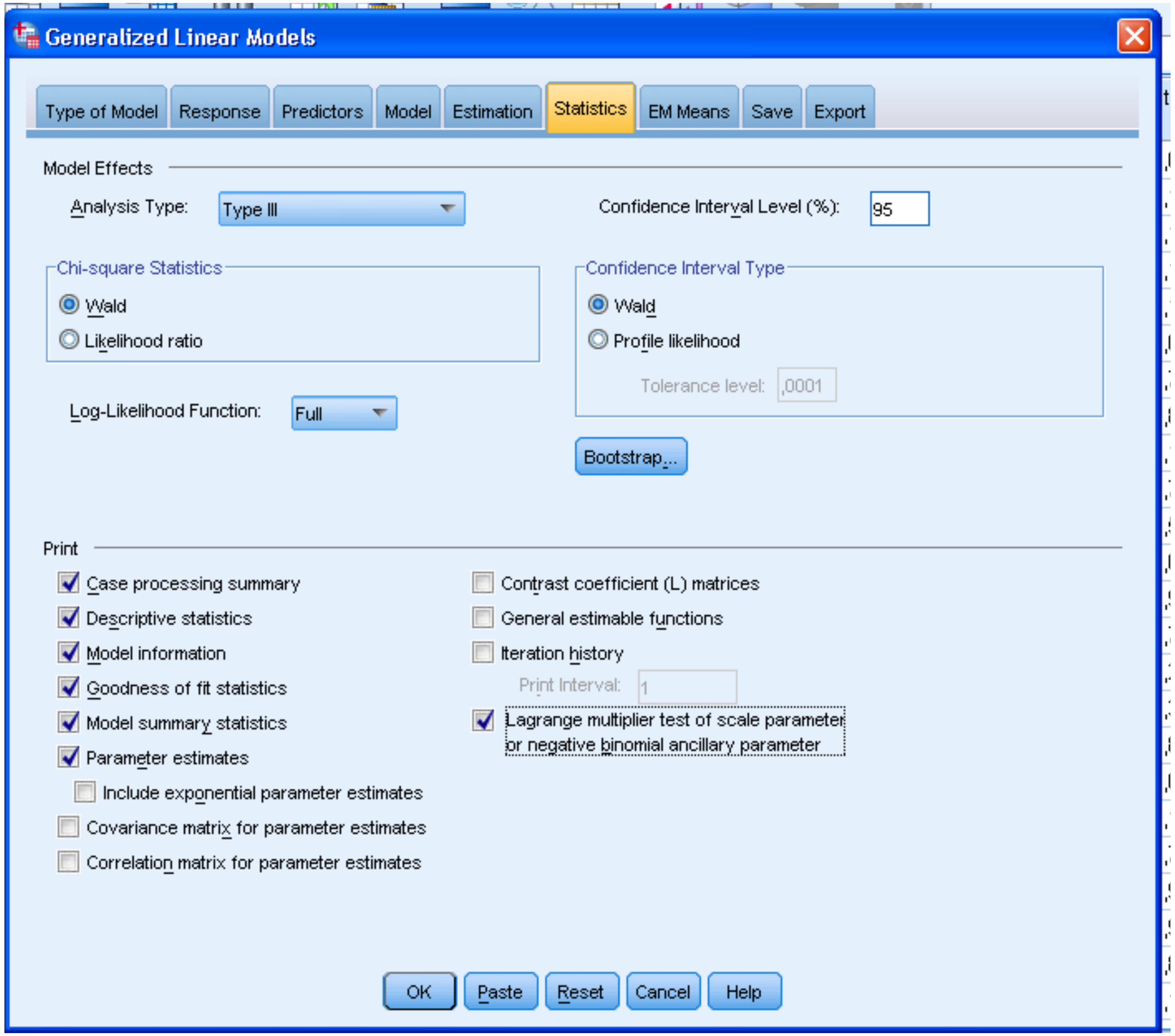
- The coefficient estimates are similar to the ones obtained with the Poisson model.
- Still, the standard errors are bigger, because they are adjusted by the scale parameter
  - When there is over dispersion, the variance of the parameters is also larger
  - As such, the standard errors of the parameters become inflated

# Negative Binomial




To estimate the Negative Binomial, and estimate the scale parameter using maximum likelihood

# Negative Binomial




**The Lagrange Multiplier test This test could only be performed if the scale parameter is fixed**

# Negative Binomial



FEUP

**Goodness of Fit**[a]

| | Value | df | Value/df |
|---|---|---|---|
| Deviance | 88,200 | 78 | 1,131 |
| Scaled Deviance | 88,200 | 78 | |
| Pearson Chi-Square | 88,922 | 78 | 1,140 |
| Scaled Pearson Chi-Square | 88,922 | 78 | |
| Log Likelihood[b] | -153,284 | | |
| Akaike's Information Criterion (AIC) | 318,567 | | |
| Finite Sample Corrected AIC (AICC) | 319,658 | | |
| Bayesian Information Criterion (BIC) | 333,152 | | |
| Consistent AIC (CAIC) | 339,152 | | |

Dependent Variable: Accident
Model: (Intercept), AADT1, AADT2, Median, Drive

a. Information criteria are in small-is-better form.

b. The full log likelihood function is displayed and used in computing information criteria.

**Omnibus Test**[a]

| Likelihood Ratio Chi-Square | df | Sig. |
|---|---|---|
| 48,526 | 4 | ,000 |

Dependent Variable: Accident
Model: (Intercept), AADT1, AADT2, Median, Drive

a. Compares the fitted model against the intercept-only model.

# Negative Binomial

**Tests of Model Effects**

| Source | Type III | | |
|---|---|---|---|
| | Wald Chi-Square | df | Sig. |
| (Intercept) | 7,621 | 1 | ,006 |
| AADT1 | 21,910 | 1 | ,000 |
| AADT2 | 15,723 | 1 | ,000 |
| Median | 4,480 | 1 | ,034 |
| Drive | 4,767 | 1 | ,029 |

Dependent Variable: Accident
Model: (Intercept), AADT1, AADT2, Median, Drive

**Parameter Estimates**

| Parameter | B | Std. Error | 95% Wald Confidence Interval | | Hypothesis Test | | |
|---|---|---|---|---|---|---|---|
| | | | Lower | Upper | Wald Chi-Square | df | Sig. |
| (Intercept) | -,931 | ,3372 | -1,592 | -,270 | 7,621 | 1 | ,006 |
| AADT1 | 8,962E-005 | 1,9146E-005 | 5,209E-005 | ,000 | 21,910 | 1 | ,000 |
| AADT2 | ,001 | ,0002 | ,000 | ,001 | 15,723 | 1 | ,000 |
| Median | -,067 | ,0317 | -,129 | -,005 | 4,480 | 1 | ,034 |
| Drive | ,063 | ,0290 | ,006 | ,120 | 4,767 | 1 | ,029 |
| (Scale) | 1[a] | | | | | | |
| (Negative binomial) | ,516 | ,1718 | ,269 | ,991 | | | |

Dependent Variable: Accident
Model: (Intercept), AADT1, AADT2, Median, Drive

a. Fixed at the displayed value.

## Negative Binomial

**Lagrange Multiplier Test**

| | Chi-Square | df | Sig. |
|---|---|---|---|
| Ancillary Parameter[a] | 4,064 | 1 | ,044 |

a. Tests the null hypothesis that the negative binomial distribution ancillary parameter equals 1

- The negative binomial model is the same as the Poisson model when the binomial model's ancillary (dispersion) parameter, $\alpha$, equals 0.
- The Lagrange multiplier test is a test of the null hypothesis that $\alpha = 1$.
- A significant Lagrange test coefficient indicates that $\alpha$ can be assumed to be different from *0*, and hence there is over-dispersion in the data.
  - A negative binomial model would be preferred over a Poisson model.
- Yet, if LL(p) is substantially smaller than LL(NB), then, the use of a Negative Binomial might not improve the model results (even with over dispersion).

- Poisson example – Accidents at intersections
  - *Washington, Simon P., Karlaftis, Mathew G. e Mannering (2003) Statistical and econometric Methods for Transportation Data Analysis, CRC*

**TABLE 10.1**

Summary of Variables in California and Michigan Accident Data

| Variable Abbreviation | Variable Description | Maximum/ Minimum Values | Mean of Observations | Standard Deviation of Observations |
|---|---|---|---|---|
| *STATE* | Indicator variable for state: 0 = California; 1 = Michigan | 1/0 | 0.29 | 0.45 |
| *ACCIDENT* | Count of injury accidents over observation period | 13/0 | 2.62 | 3.36 |
| *AADT1* | Average annual daily traffic on major road | 33058/2367 | 12870 | 6798 |
| *AADT2* | Average annual daily traffic on minor road | 3001/15 | 596 | 679 |
| *MEDIAN* | Median width on major road in feet | 36/0 | 3.74 | 6.06 |
| *DRIVE* | Number of driveways within 250 ft of intersection center | 15/0 | 3.10 | 3.90 |

## Example 1

- Poisson example – Accidents at intersections
  - *Washington, Simon P., Karlaftis, Mathew G. e Mannering (2003) Statistical and econometric Methods for Transportation Data Analysis, CRC*

**TABLE 10.2**

Poisson Regression of Injury Accident Data

| Independent Variable | Estimated Parameter | t Statistic |
|---|---|---|
| Constant | –0.826 | –3.57 |
| Average annual daily traffic on major road | 0.0000812 | 6.90 |
| Average annual daily traffic on minor road | 0.000550 | 7.38 |
| Median width in feet | – 0.0600 | – 2.73 |
| Number of driveways within 250 ft of intersection | 0.0748 | 4.54 |
| Number of observations | 84 | |
| Restricted log likelihood (constant term only) | –246.18 | |
| Log likelihood at convergence | –169.25 | |
| Chi-squared (and associated *p*-value) | 153.85 (<0.0000001) | |

- Negative Binomial – Accidents at intersections **Example 1**
  - *Washington, Simon P., Karlaftis, Mathew G. e Mannering (2003) Statistical and econometric Methods for Transportation Data Analysis, CRC*

**TABLE 10.4**

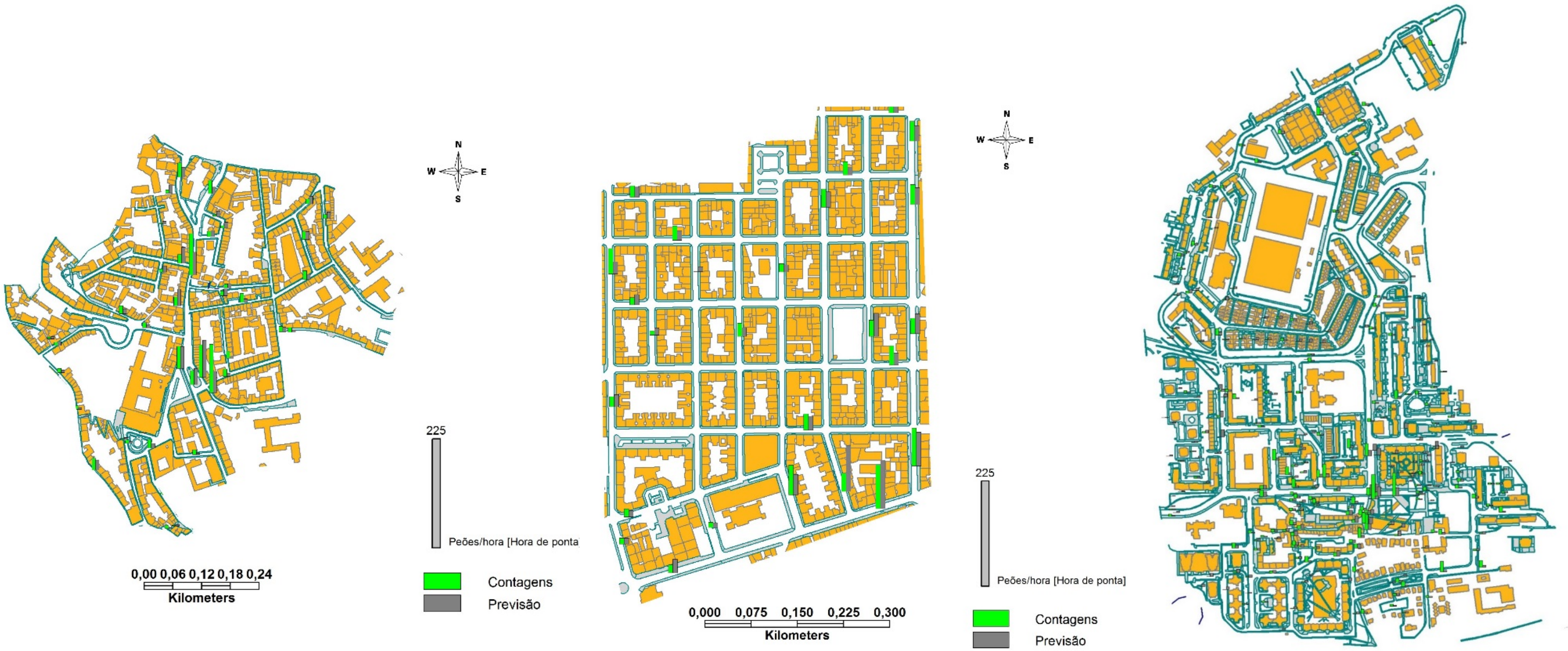
Negative Binomial Regression of Injury Accident Data

| Independent Variable | Estimated Parameter | $t$ Statistic |
|---|---|---|
| Constant | –0.931 | –2.37 |
| Average annual daily traffic on major road | 0.0000900 | 3.47 |
| Average annual daily traffic on minor road | 0.000610 | 3.09 |
| Median width in feet | – 0.0670 | –1.99 |
| Number of driveways within 250 ft of intersection | 0.0632 | 2.24 |
| Overdispersion parameter, α | 0.516 | 3.09 |
| Number of observations | 84 | |
| Restricted log likelihood (constant term only) | –169.25 | |
| Log likelihood at convergence | –153.28 | |
| Chi-squared (and associated $p$-value) | 31.95 (<0.0000001) | |

- Overdispersed Poisson – Pedestrian countings
  - *Barros, A.P., Martinez, L.M., Viegas, J.M., Silva, P.C., Holanda, F. (2013) Análise da mobilidade de pedestres sob o prisma de três configurações urbanas distintas – Estudo de caso em Lisboa, ANPET.*

| Variáveis | Coef. | Coef. Pad. | Erro pad. | Wald Chi$^2$ | Sig. |
|---|---|---|---|---|---|
| (Termo independente) | 3.926 | 3.926 | 0.398 | 97.196 | 0.000 |
| Índice de integração (HH) | 0.685 | 0.394 | 0.232 | 8.748 | 0.003 |
| Conectividade | -0.242 | -1.352 | 0.060 | 16.034 | 0.000 |
| Compacidade viária | -0.071 | -0.476 | 0.033 | 4.637 | 0.031 |
| Calçadas estreitas | -0.360 | -0.051 | 0.197 | 3.340 | 0.068 |
| Presença de escadas | -0.771 | -0.019 | 0.289 | 7.143 | 0.008 |
| Presença de árvores | 0.285 | 0.112 | 0.122 | 5.464 | 0.019 |
| Declive elevado | -0.566 | -0.043 | 0.276 | 4.192 | 0.041 |
| Área de Comércio | 0.179 | 0.177 | 0.041 | 18.970 | 0.000 |
| Área de Educação | 0.209 | 0.043 | 0.084 | 6.131 | 0.013 |
| Alimentação e lazer | 0.116 | 0.046 | 0.101 | 1.311 | 0.252 |
| Entropia | 0.387 | 0.279 | 0.162 | 5.688 | 0.017 |
| Número de Portas | 0.035 | 0.384 | 0.006 | 37.086 | 0.000 |
| Proximidade ônibus | 0.306 | 0.052 | 0.144 | 4.494 | 0.034 |
| Proximidade metrô | 1.534 | 34.279 | 0.375 | 16.756 | 0.000 |
| Linhas de ônibus | 0.200 | 0.108 | 0.050 | 16.349 | 0.000 |
| (Parâmetro de sobredispersão) | 48.140 | | | | |

- Overdispersed Poisson – Pedestrian countings **Example 2**
  - *Barros, A.P., Martinez, L.M., Viegas, J.M., Silva, P.C., Holanda, F. (2013) Análise da mobilidade de pedestres sob o prisma de três configurações urbanas distintas – Estudo de caso em Lisboa, ANPET.*



# Your Home assignment

- **Objective**
  - To evaluate the importance/impact of the **International friction index – IFI of the pavements** on the level of accidents

- **You should use the same methodology:**
  - Compare 3 Generalized Linear Models (SPSS), for which you should perform, and explain in your report, the following major steps:
    1. Model Formulation
    2. Model Adjustment
    3. Model Validation

# References

## ❒ References


- ❖ Hardin, J. W., & Hilbe, J. M. (2007). Generalized linear models and extensions (2nd ed.). College Station, TX: StataCorp LP.
- ❖ Hilbe, Joseph M. (2007). Negative binomial regression. New York: Cambridge University Press.
- ❖ **Hoffman, J. P. (2004). Generalized linear models: An applied approach. Boston: Allyn & Bacon. An accessible extended introduction**.
- ❖ McCullagh, P. & J.A. Nelder (1989). Generalized linear models. Second Edition. Boca Raton: Chapman and Hall/CRC. ISBN 0-412-31760-5.
- ❖ **Nelder, J. A. & Wedderburn, R. W. N. (1972). Generalized linear models. Journal of the Royal Statistical Society 135: 370-384. The seminal article for GZLM.**


FEUP

# References

## ❒ References


- J. B. S. Haldane, "On a Method of Estimating Frequencies", Biometrika, Vol. 33, No. 3 (Nov., 1945), pp. 222–225. JSTOR 2332299
- **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
- Domencich and Mcfadden (1975) Urban Travel Demand: A Behavioral Analysis. North-Holland Publishing Co., 1975. Reprinted 1996.


FEUP