**9. Introduction to Generalized Linear Models I**

A Generalized Linear Model (GLZ), defined below, extends General Linear (GLM) Models in three ways:



1. The distribution of the observations can come from the family of exponential distributions; no assumption of normality is required. This includes distributions such as the normal, gamma, Poisson, binomial, and negative binomial distributions.
2. The variance of the response variable may be specified as a function of its mean. (In the case of the normal distribution, the relationship may be expressed as . See below for details.)
3. The link function *g* is introduced to fit the linear model. This link function must be monotonic, but does not have to be the identity function, as is the case for general linear models.

The general form for a density or probability function of the exponential family can be expressed as , where *θ* is the location (canonical) parameter, *φ* is the scale (dispersion) parameter and *a*(.), *b*(.), *c*(.) are known functions.

For this family of distributions, the variance of *y* can be expressed as a function of the mean of *y* as:



For the normal distribution, for example, , that shows  since () and .

Note that as the link function is monotonic, it can be inverted to obtain . In practice, any link function could be used but the most useful and mathematically convenient ones are *canonical* link functions that relate to the canonical parameter .

**Canonical Link Function**

An alternative parameterisation for the general form for a density or probability mass function for an exponential family distribution is. When written in this form, the *canonical link* can be identified as *W(θ)*.

**Example 9.1**

To identify the canonical link for binary data, start with the pdf of a Bernoulli (binary) random variable and rearrange to this format (shown previously):



**Example 9.2**

The probability distribution function for Poisson distribution is given by:

*f* (y | λ)

In this expression, *y* is a non-negative integer value and *λ* is the expected value of *y*. It can be shown that *Var*(*y*) = *λ.* and the scale parameter is 1.

To identify the canonical link, rearrange it to the format:



The last term indicates that the natural log is the canonical link, where the location parameter is λ and the scale parameter is 1.

**Poisson Regression for Count Data**

The Poisson regression is often used to analyse count data, where the response variable has nonnegative integer values. Poisson regression can also be used to model the rate or incidence of an event. This type of outcome is widely seen in the medical sciences, biological sciences, social sciences, agriculture, engineering, and business.

The assumption in Poisson regression is that the conditional distribution of the response variable follows a Poisson distribution. Although ordinary least squares regression can be used to analyse count data, Poisson regression has the advantage of being precisely tailored to the discrete, often skewed distribution of count data.

In addition to being skewed, the sample distribution should have a fairly small mean if Poisson regression is the method of choice. The mean should certainly be below 10, preferably below 5, and ideally in the neighbourhood of 1.

**Note:** The gamma distribution or the lognormal distribution might be more appropriate for highly skewed data with large mean values.

Recall that the Poisson distribution for the random variable *X* is defined as:



It is fully defined by one parameter, the mean rate of occurrence of events, *λ*. An unusual property of the Poisson distribution is that the mean and variance are equal. This can be a serious limitation because count observations often exhibit variability exceeding that predicted by the Poisson distribution. This leads to *over-dispersion*.



For rare events, the Poisson distribution is different from the normal distribution. As the mean increases however, the Poisson distribution approximates the normal distribution.

Note that the Poisson distribution is a discrete distribution, which might be more conventionally represented by a bar chart. The graph shown above represents a continuous approximation to the bar chart for the Poisson distribution.

In Poisson regression the dependent variable is an observed count which follows the Poisson distribution. So *λ* is now determined as and thus for observation *i*:





The method of maximum likelihood is used to estimate the parameters of Poisson regression models.

Like the logistic models, the parameter estimates represent the expected change in the log scale. That is, if we calculate , we obtain the percentage change in the expected number of events with each one-unit increase in the predictor variable. For example, if , then a one-unit increase in ** yields a 20% increase in the estimated mean and if, then a one-unit increase in ** yields a 20% decrease in the estimated mean.

**Example 9.3**

A survey was undertaken to examine which factors are related to ear infections among swimmers. The response variable is the number of self-diagnosed ear infections reported by the participant. The variables in the data set are:

**Infections** number of self-diagnosed ear infections

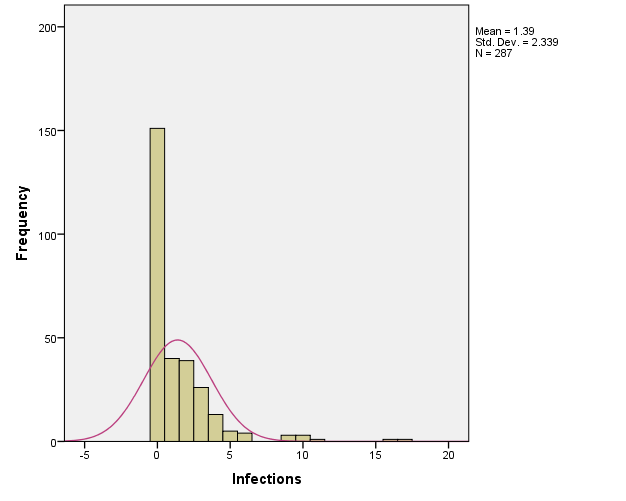
**Swimmer** swimmer’s perception of whether he or she is a frequent ocean swimmer or occasional ocean swimmer

**Location** swimmer’s typical swimming location (NonBeach = not a beach swimmer, Beach = usually a beach swimmer)

**Age** age in years

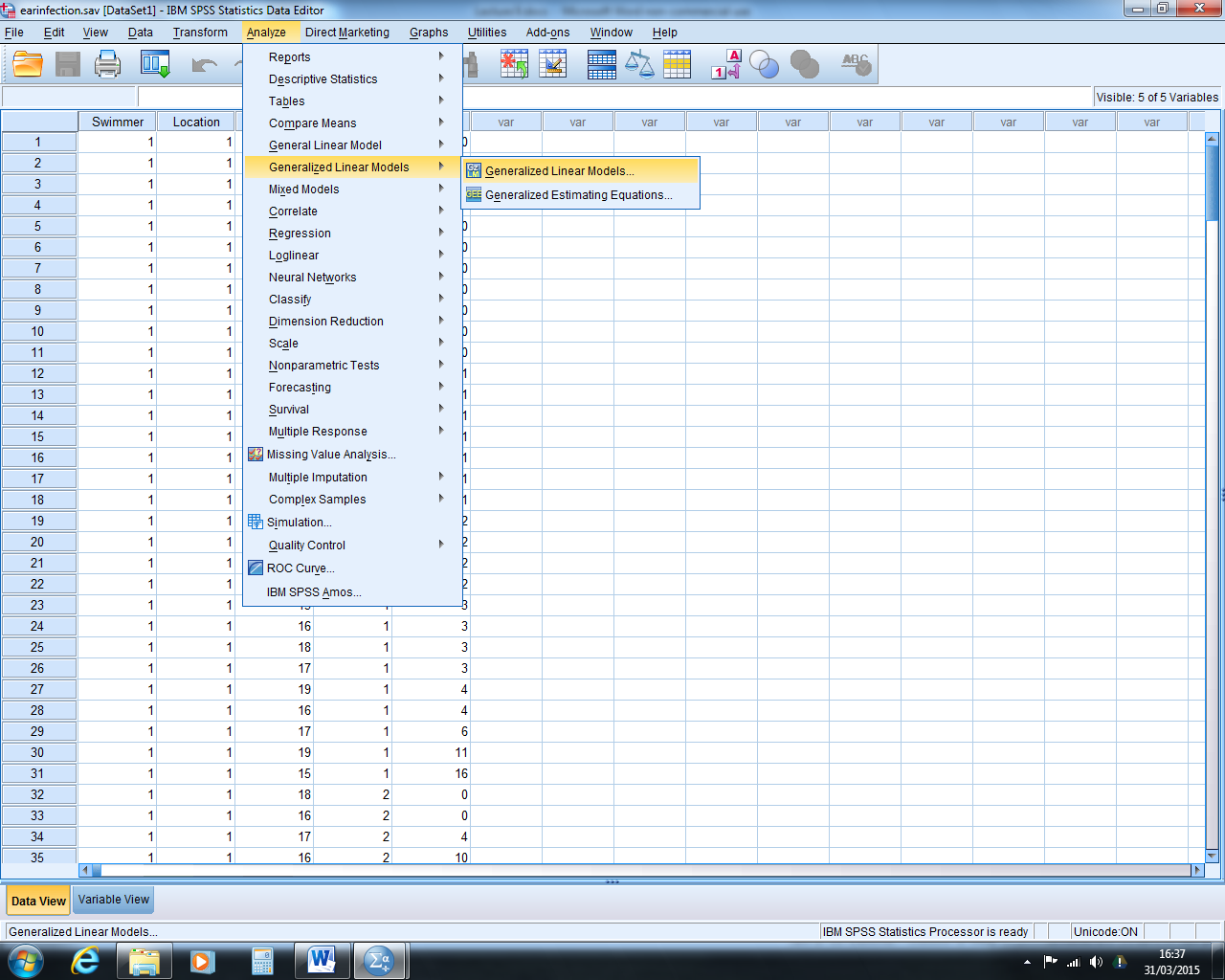
**Sex** sex of swimmer.

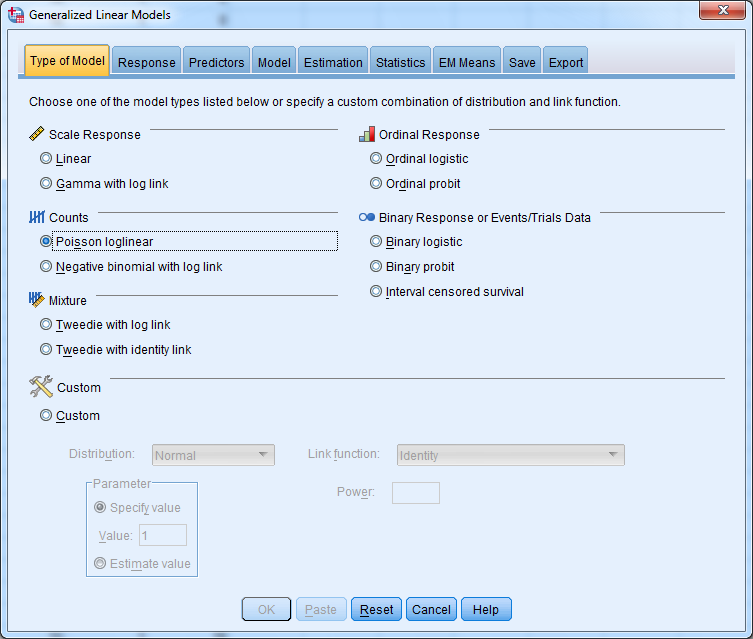
We will examine the dependent variable and fit a Poisson regression model to the data.

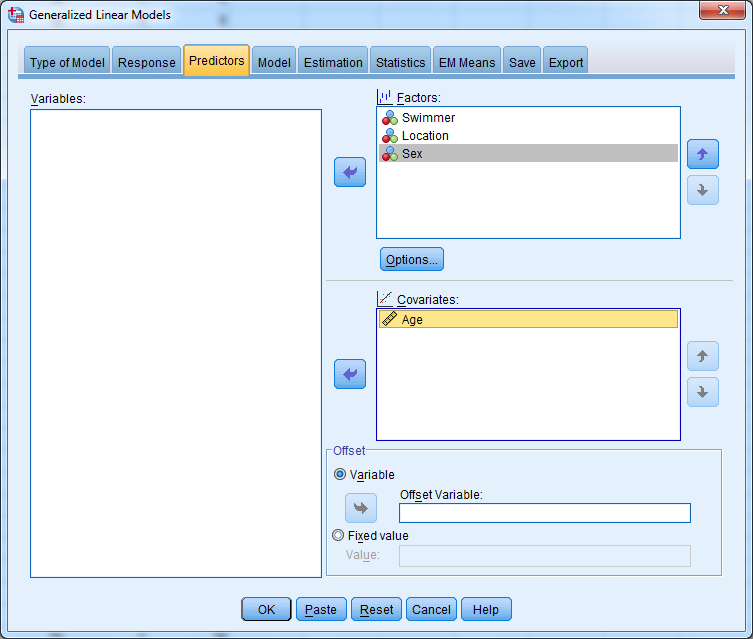


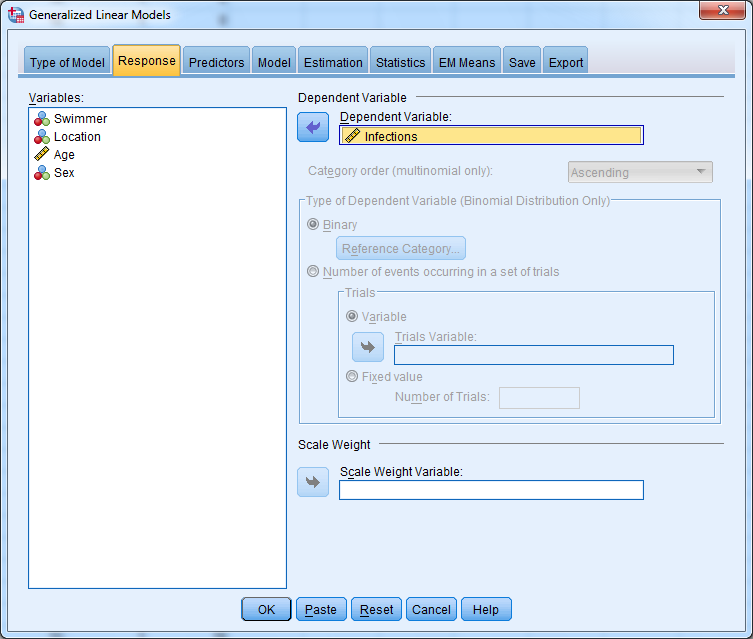
It is obvious that the response distribution is far away from normal and no appropriate transformation can resolve this problem.

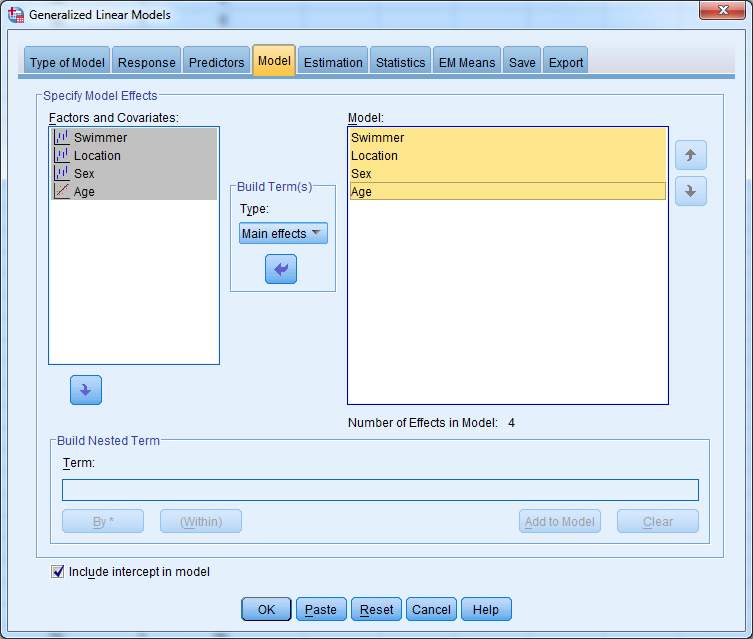
To analyse the data by a Poisson regression, we start by **Analyze** 🡆 **Generalized Linear Model** 🡆 **Generalized Linear Model**.



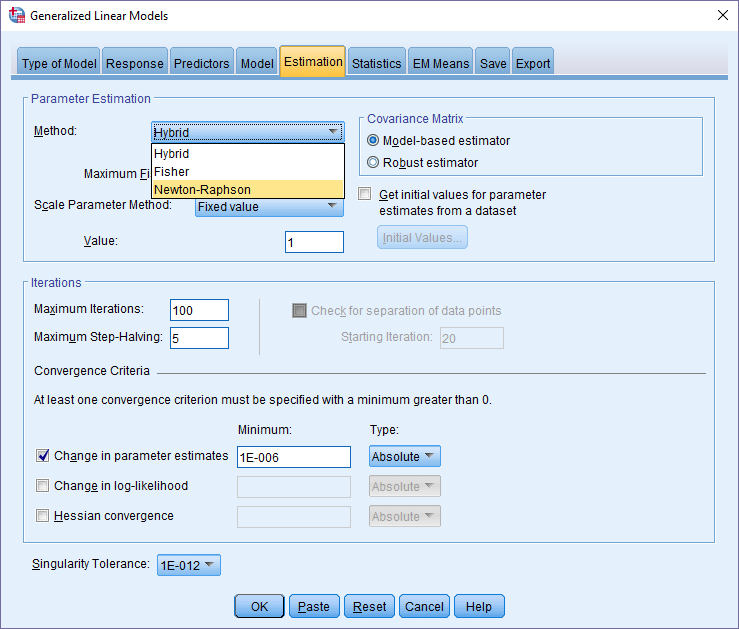


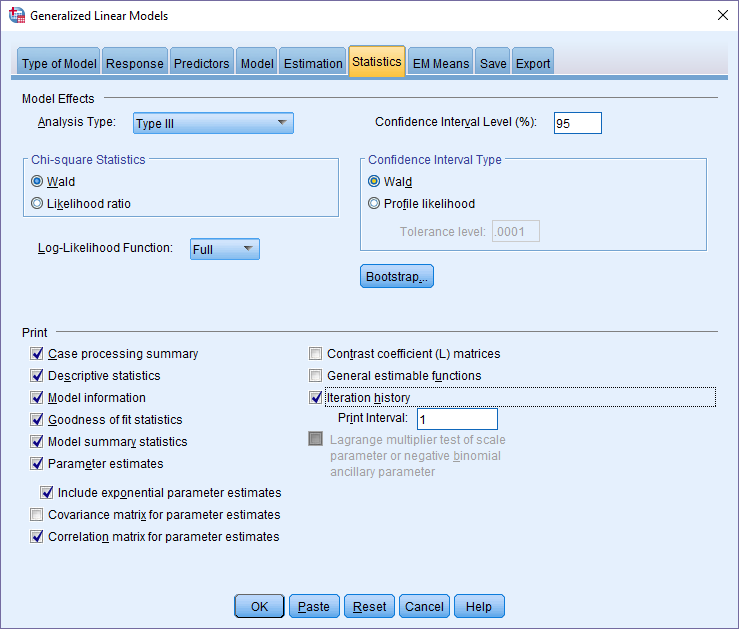




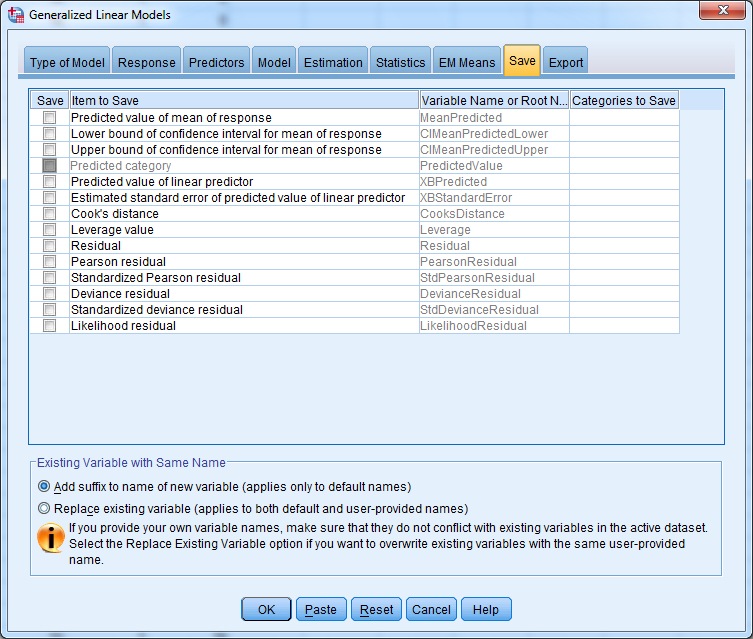


In the last dialogue box you can also build interaction terms.





Check “Include exponential parameter estimates”, Correlation matrix for parameter estimates and “Iteration history”.



These measures can be saved and used for different plots if required.

|  |  |
| --- | --- |
| **Model Information** | |
| Dependent Variable | Infections |
| Probability Distribution | Poisson |
| Link Function | Log |

|  |  |  |
| --- | --- | --- |
| **Case Processing Summary** | | |
|  | N | Percent |
| Included | 287 | 100.0% |
| Excluded | 0 | 0.0% |
| Total | 287 | 100.0% |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Categorical Variable Information** | | | | |
|  | | | N | Percent |
| Factor | Swimmer | Occasional | 144 | 50.2% |
| Frequent | 143 | 49.8% |
| Total | 287 | 100.0% |
| Location | NonBeach | 140 | 48.8% |
| Beach | 147 | 51.2% |
| Total | 287 | 100.0% |
| Sex | Male | 188 | 65.5% |
| Female | 99 | 34.5% |
| Total | 287 | 100.0% |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Continuous Variable Information** | | | | | | |
|  | | N | Minimum | Maximum | Mean | Std. Deviation |
| Dependent Variable | Infections | 287 | 0 | 17 | 1.39 | 2.339 |
| Covariate | Age | 287 | 15 | 29 | 20.77 | 4.319 |

Summary information on model, estimation methods and variables.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Iteration History** | | | | | | | | | |
| Iteration | Update Type | Number of Step-halvings | Log Likelihoodb | Parameter | | | | | |
| (Intercept) | [Swimmer=1] | [Location=1] | [Sex=1] | Age | (Scale) |
| 0 | Initial | 0 | -843.373 | 1.088657 | .548341 | .121471 | .025896 | -.008564 | 1 |
| 1 | Scoring | 0 | -597.727 | .521775 | .571417 | .257232 | .007587 | -.014829 | 1 |
| 2 | Newton | 0 | -567.243 | .289070 | .597768 | .416961 | -.016369 | -.022437 | 1 |
| 3 | Newton | 0 | -566.203 | .260363 | .608034 | .484809 | -.028349 | -.025812 | 1 |
| 4 | Newton | 0 | -566.200 | .260366 | .608627 | .489589 | -.029380 | -.026063 | 1 |
| 5 | Newton | 0 | -566.200 | .260370 | .608628 | .489603 | -.029384 | -.026064 | 1 |
| 6 | Newtona | 3 | -566.200 | .260370 | .608628 | .489603 | -.029384 | -.026064 | 1 |
| Redundant parameters are not displayed. Their values are always zero in all iterations.  Dependent Variable: Infections  Model: (Intercept), Swimmer, Location, Sex, Age | | | | | | | | | |
| a. All convergence criteria are satisfied. | | | | | | | | | |
| b. The full log likelihood function is displayed. | | | | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Gradient Vector and Hessian Matrix** | | | | | | |
|  | | Parameter | | | | |
| (Intercept) | [Swimmer=2] | [Location=2] | [Sex=2] | Age |
| Gradient Vector | | .000 | .000 | .000 | .000 | .000 |
| Hessian Matrix | (Intercept) | -398.000 | -140.000 | -155.000 | -131.000 | -8034.000 |
| [Swimmer=2] | -140.000 | -140.000 | -53.378 | -44.141 | -2838.889 |
| [Location=2] | -155.000 | -53.378 | -155.000 | -67.412 | -3209.724 |
| [Sex=2] | -131.000 | -44.141 | -67.412 | -131.000 | -2744.139 |
| Age | -8034.000 | -2838.889 | -3209.724 | -2744.139 | -169025.328 |
| The last evaluation of the gradient vector and Hessian matrix are displayed.  Redundant parameters are not displayed. | | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Goodness of Fita** | | | |
|  | Value | df | Value/df |
| Deviance | 760.006 | 282 | 2.695 |
| Scaled Deviance | 760.006 | 282 |  |
| Pearson Chi-Square | 963.584 | 282 | 3.417 |
| Scaled Pearson Chi-Square | 963.584 | 282 |  |
| Log Likelihoodb | -566.200 |  |  |
| Akaike's Information Criterion (AIC) | 1142.401 |  |  |
| Finite Sample Corrected AIC (AICC) | 1142.614 |  |  |
| Bayesian Information Criterion (BIC) | 1160.698 |  |  |
| Consistent AIC (CAIC) | 1165.698 |  |  |
| Dependent Variable: Infections  Model: (Intercept), Swimmer, Location, Sex, Age | | | |
| a. Information criteria are in smaller-is-better form. | | | |
| b. The full log likelihood function is displayed and used in computing information criteria. | | | |

Goodness of Fit table provides statistics for testing the goodness of fit of the model. The measures are the **Deviance** and the **Pearson Chi-Square** statistic. The values of these statistics divided by the squared scale parameter (that is, the dispersion parameter) are called scaled deviance and scaled Pearson chi-squared. Because the scale parameter by definition is 1 for Poisson regression, the statistics (original and scaled) are equal.

The Value/df values are computed by dividing the goodness-of-fit statistics by the degrees of freedom. These values for the scaled deviance or the scaled Pearson chi-square are useful for assessing the goodness of model fit. Values close to 1 indicate good model fit. The Value/df column in the table has 2.695 for scaled deviance and 3.417 for scaled Pearson chi-square. They are not close to 1. This might indicate over-dispersed data, which occurs frequently in Poisson regression, and occasionally in logistic regression. Over-dispersion does not affect the parameter estimates, but it causes the estimates of the standard error of the parameter estimates to be under-estimated which will increase type I error.

Other fit statistics include AIC, AICC and BIC. Each is a measure of goodness of model fit that balances model fit against model simplicity. These criteria are useful in selecting among models, with smaller values representing better model fit.

|  |  |  |  |
| --- | --- | --- | --- |
| **Tests of Model Effects** | | | |
| Source | Type III | | |
| Wald Chi-Square | df | Sig. |
| (Intercept) | 9.613 | 1 | .002 |
| Swimmer | 33.586 | 1 | .000 |
| Location | 21.812 | 1 | .000 |
| Sex | .072 | 1 | .788 |
| Age | 4.547 | 1 | .033 |
| Dependent Variable: Infections  Model: (Intercept), Swimmer, Location, Sex, Age | | | |

The LR Statistics for Type III Analysis gives the tests of significance for each of the parameters. All independent variables, except Sex, are significant. However, because the data exhibits over-dispersion, these results might not be reliable.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | | | | | | |
| Parameter | B | Std. Error | 95% Wald Confidence Interval | | Hypothesis Test | | | Exp(B) | 95% Wald Confidence Interval for Exp(B) | |
| Lower | Upper | Wald Chi-Square | df | Sig. | Lower | Upper |
| (Intercept) | 1.329 | .2517 | .836 | 1.822 | 27.898 | 1 | .000 | 3.778 | 2.307 | 6.187 |
| [Swimmer=2] | -.609 | .1050 | -.814 | -.403 | 33.586 | 1 | .000 | .544 | .443 | .668 |
| [Swimmer=1] | 0a | . | . | . | . | . | . | 1 | . | . |
| [Location=2] | -.490 | .1048 | -.695 | -.284 | 21.812 | 1 | .000 | .613 | .499 | .753 |
| [Location=1] | 0a | . | . | . | . | . | . | 1 | . | . |
| [Sex=2] | .029 | .1092 | -.185 | .243 | .072 | 1 | .788 | 1.030 | .831 | 1.276 |
| [Sex=1] | 0a | . | . | . | . | . | . | 1 | . | . |
| Age | -.026 | .0122 | -.050 | -.002 | 4.547 | 1 | .033 | .974 | .951 | .998 |
| (Scale) | 1b |  |  |  |  |  |  |  |  |  |
| Dependent Variable: Infections  Model: (Intercept), Swimmer, Location, Sex, Age | | | | | | | | | | |
| a. Set to zero because this parameter is redundant. | | | | | | | | | | |
| b. Fixed at the displayed value. | | | | | | | | | | |

Parameter estimates and the *p*-values for testing whether the estimates are significantly different from zero. Again, be aware of under-estimated standard errors!

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Correlations of Parameter Estimates** | | | | | | | | |
|  | (Intercept) | [Swimmer=2] | [Swimmer=1] | [Location=2] | [Location=1] | [Sex=2] | [Sex=1] | Age |
| (Intercept) | 1.000 | -.132 | .a | -.062 | .a | -.008 | .a | -.948 |
| [Swimmer=2] | -.132 | 1.000 | .a | .010 | .a | .022 | .a | -.020 |
| [Swimmer=1] | .a | .a | .a | .a | .a | .a | .a | .a |
| [Location=2] | -.062 | .010 | .a | 1.000 | .a | -.169 | .a | -.079 |
| [Location=1] | .a | .a | .a | .a | .a | .a | .a | .a |
| [Sex=2] | -.008 | .022 | .a | -.169 | .a | 1.000 | .a | -.113 |
| [Sex=1] | .a | .a | .a | .a | .a | .a | .a | .a |
| Age | -.948 | -.020 | .a | -.079 | .a | -.113 | .a | 1.000 |
| Dependent Variable: Infections  Model: (Intercept), Swimmer, Location, Sex, Age | | | | | | | | |
| a. One or both parameter estimates are redundant. | | | | | | | | |

If any of the correlation coefficients are above 0.98 or 0.99, the parameters are excessively correlated. This can be an indication that the model is over-parametrised.