9. Introduction to Generalized Linear Models I

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

$$g\left(E(y)\right) = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k$$

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 $\sigma^2 = \mu^0 \cdot \sigma^2$. 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 $f(y \mid \theta, \phi) = \exp\left\{\dfrac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$, 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:

$$E(y) = \mu = b'(\theta)$$
$$Var(y) = \sigma^2 = b''(\theta).a(\phi)$$

For the normal distribution, for example, $f(y) = \exp\left\{\dfrac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} + c(y, \sigma^2)\right\}$, that shows $b(\theta) = \frac{1}{2}\theta^2$ since ($\theta = \mu$) and $a(\phi) = \phi = \sigma^2$.

Note that as the link function is monotonic, it can be inverted to obtain $\mu = g^{-1}(\boldsymbol{X\beta})$. In practice, any link function could be used but the most useful and mathematically convenient ones are *canonical* link functions that relate μ to the canonical parameter θ.

Canonical Link Function

An alternative parameterisation for the general form for a density or probability mass function for an exponential family distribution is $f(y\,|\,\theta) = h(y)\cdot c(\theta)\cdot e^{t(y).W(\theta)}$. 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):

$$f(y\,|\,p) = p^{y}\cdot(1-p)^{(1-y)} = p^{y}\cdot(1-p)^{1}\cdot(1-p)^{-y}$$

$$= \frac{p^{y}\cdot(1-p)}{(1-p)^{y}} = (1-p)\cdot I_{y}\cdot\left(\frac{p}{1-p}\right)^{y} = (1-p)\cdot I_{y}\cdot e^{\ln\left(\frac{p}{1-p}\right)^{y}}$$

$$= (1-p)\cdot I_{y}\cdot e^{y\cdot\ln\left(\frac{p}{1-p}\right)} \qquad \text{where } I_{y} = \begin{cases} 1 & \text{for } y = 0,1 \\ 0 & \text{otherwise} \end{cases}$$

In this format you can see that the canonical link is, $\ln\left(\frac{p}{1-p}\right)$, where the location parameter is p and the scale parameter is 1.

Example 9.2

The probability distribution function for Poisson distribution is given by:

$$f(\mathrm{y}\,|\,\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \qquad y = 0, 1, 2, \cdots$$

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

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

$$f(y\,|\,\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} = \frac{e^{-\lambda}\cdot e^{\ln(\lambda^{y})}}{y!} = e^{-\lambda}\cdot\frac{1}{y!}\cdot e^{y(\ln(\lambda))}$$

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.

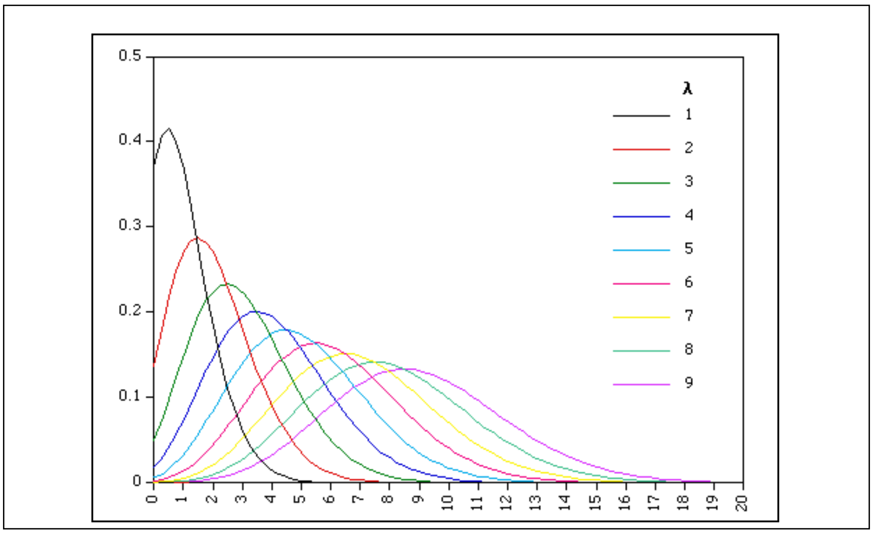
In addition to being skewed, the sample distribution should have a fairly small mean if Poisson regression is the method of choice. The mean should certainly be below 10, preferably below 5, and ideally in the neighbourhood of 1.

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:

$$\Pr(X = x \mid \lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \qquad x = 0, 1, 2, \ldots$$

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 $\lambda = \exp\{\boldsymbol{X\beta}\}$ and thus for observation *i*:

$$\Pr(Y = y_i \mid \boldsymbol{X};\boldsymbol{\beta}) = \frac{e^{-\exp\{\boldsymbol{X\beta}\}}\exp\{\boldsymbol{X\beta}\}^{y_i}}{y_i!}$$

$$E(y) = \lambda = \exp\{\beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k\}$$

$$\ln\big(E(y)\big) = \lambda = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

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 $100(e^{b_j} - 1)$, we obtain the percentage change in the expected number of events with each one-unit increase in the predictor variable. For example, if $e^{b_j} = 1.20$, then a one-unit increase in x_j yields a 20% increase in the estimated mean and if $e^{b_j} = 0.80$, then a one-unit increase in x_j 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

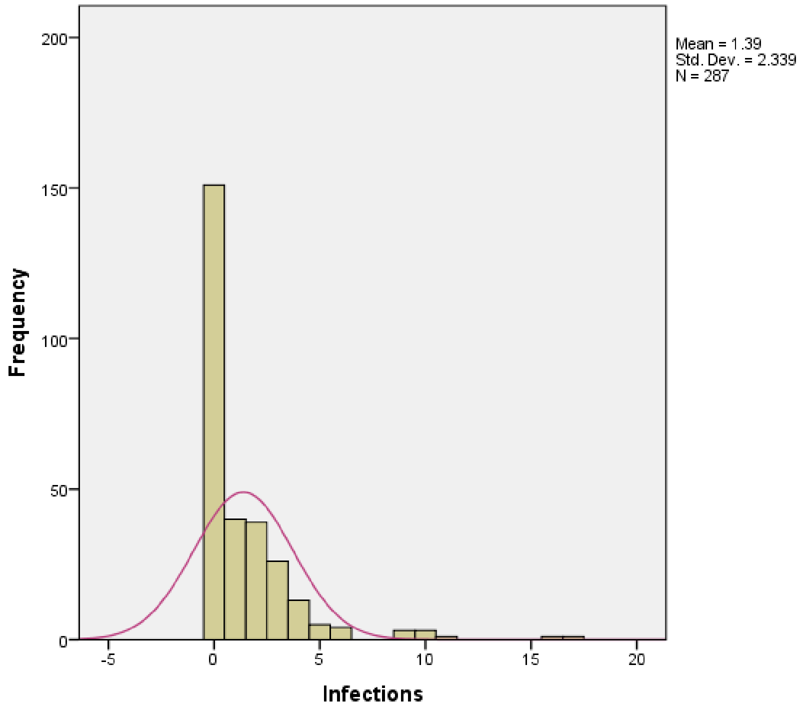
Swimmer swimmer's perception of whether he or she is a frequent ocean swimmer or occasional ocean swimmer

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

Age age in years

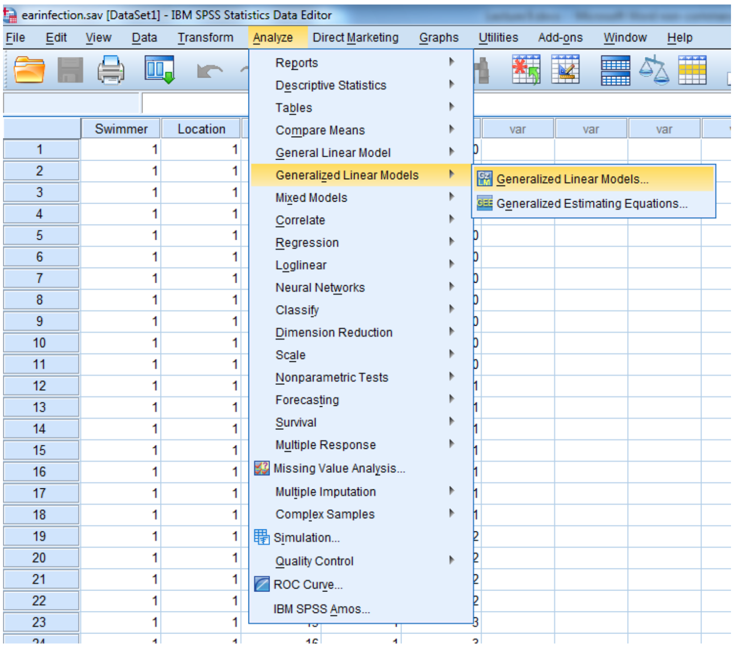
Sex sex of swimmer.

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

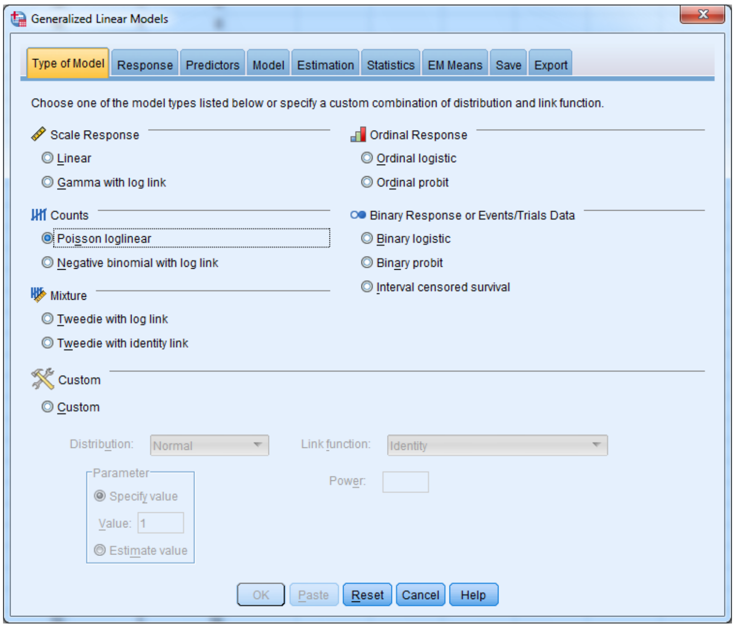


It is obvious that the response distribution is far away from normal and no appropriate transformation can resolve this problem.

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



	Swimmer	Location
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1
10	1	1
11	1	1
12	1	1
13	1	1
14	1	1
15	1	1
16	1	1
17	1	1
18	1	1
19	1	1
20	1	1
21	1	1
22	1	1
23	1	1



Choose one of the model types listed below or specify a custom combination of distribution and link function.

Scale Response
○ Linear
○ Gamma with log link

Ordinal Response
○ Ordinal logistic
○ Ordinal probit

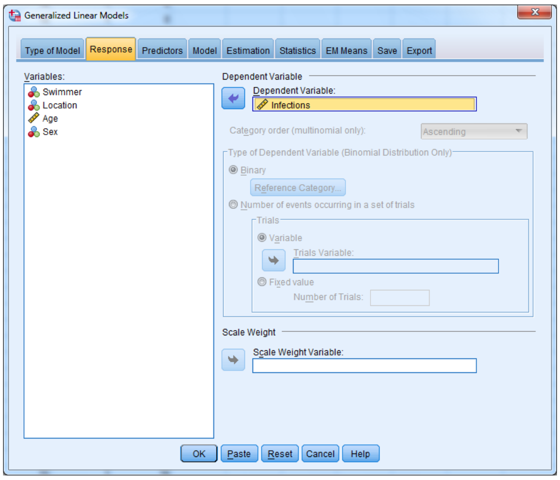
Counts
● Poisson loglinear
○ Negative binomial with log link

Binary Response or Events/Trials Data
○ Binary logistic
○ Binary probit
○ Interval censored survival

Mixture
○ Tweedie with log link
○ Tweedie with identity link

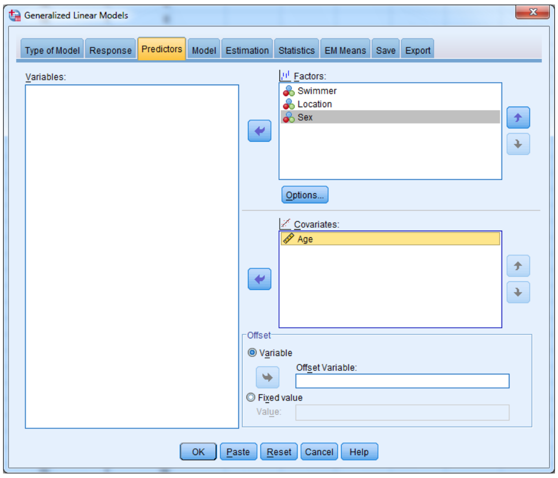
Custom
○ Custom

Distribution: Normal
Link function: Identity
Parameter
● Specify value
Value: 1
○ Estimate value
Power:



Variables:
Swimmer
Location
Age
Sex

Dependent Variable
Dependent Variable: Infections
Category order (multinomial only): Ascending
Type of Dependent Variable (Binomial Distribution Only)
○ Binary
Reference Category...
○ Number of events occurring in a set of trials
Trials
● Variable
Trials Variable:
○ Fixed value
Number of Trials:
Scale Weight
Scale Weight Variable:

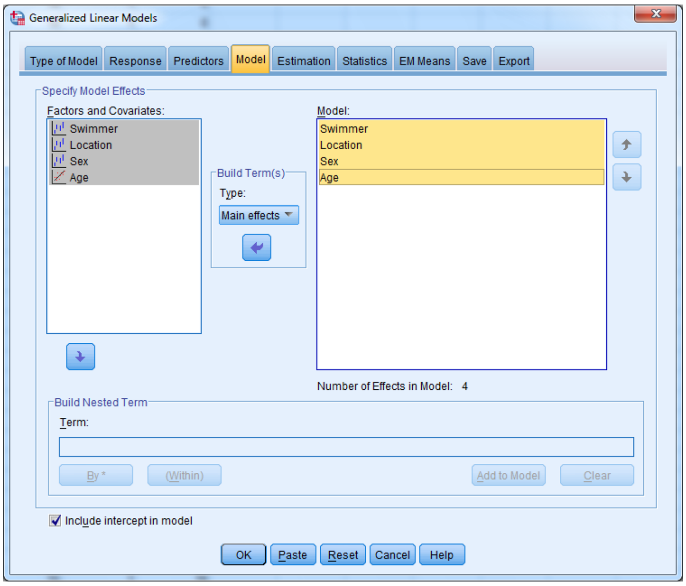


Variables:

Factors:
Swimmer
Location
Sex

Covariates:
Age

Offset
● Variable
Offset Variable:
○ Fixed value
Value:



Specify Model Effects
Factors and Covariates:
Swimmer
Location
Sex
Age

Build Term(s)
Type: Main effects

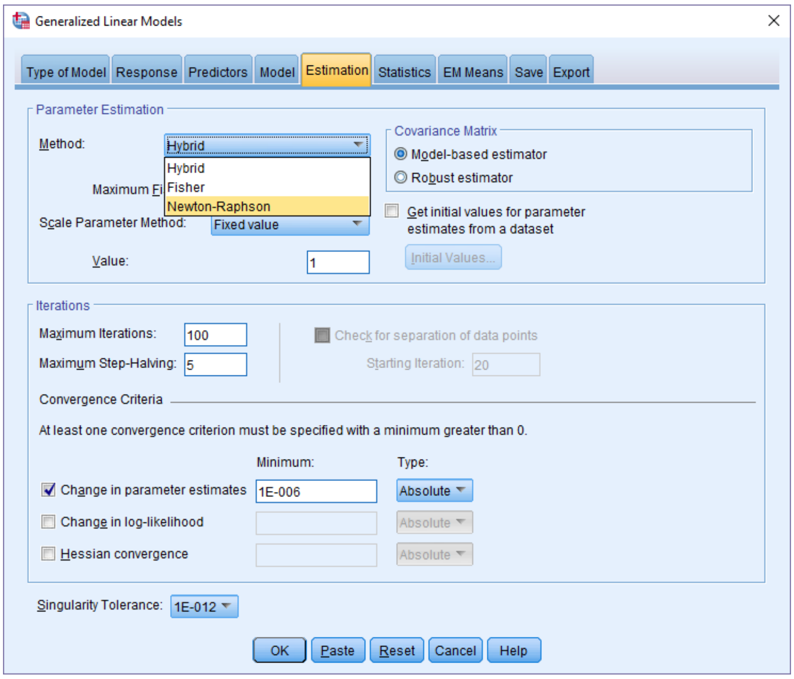
Model:
Swimmer
Location
Sex
Age

Number of Effects in Model: 4

Build Nested Term
Term:

☑ Include intercept in model

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



Parameter Estimation

Method: Hybrid

Maximum Fi

Scale Parameter Method: Fixed value

Value: 1

Covariance Matrix

● Model-based estimator

○ Robust estimator

□ Get initial values for parameter estimates from a dataset

Iterations

Maximum Iterations: 100

Maximum Step-Halving: 5

□ Check for separation of data points

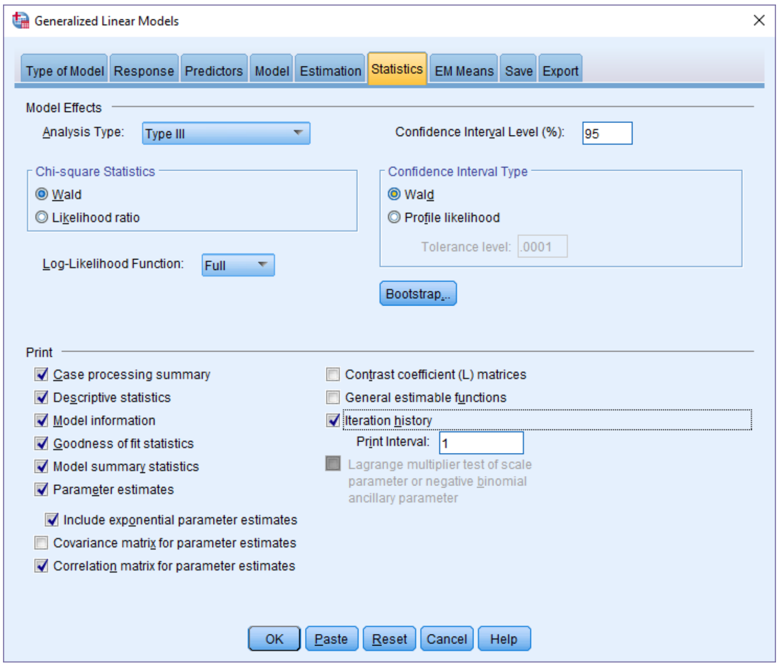
Starting Iteration: 20

Convergence Criteria

At least one convergence criterion must be specified with a minimum greater than 0.

	Minimum:	Type:
☑ Change in parameter estimates	1E-006	Absolute
□ Change in log-likelihood		Absolute
□ Hessian convergence		Absolute

Singularity Tolerance: 1E-012



Model Effects

Analysis Type: Type III

Confidence Interval Level (%): 95

Chi-square Statistics

● Wald

○ Likelihood ratio

Log-Likelihood Function: Full

Confidence Interval Type

● Wald

○ Profile likelihood

Tolerance level: .0001

Print

☑ Case processing summary

☑ Descriptive statistics

☑ Model information

☑ Goodness of fit statistics

☑ Model summary statistics

☑ Parameter estimates

☑ Include exponential parameter estimates

□ Covariance matrix for parameter estimates

☑ Correlation matrix for parameter estimates

□ Contrast coefficient (L) matrices

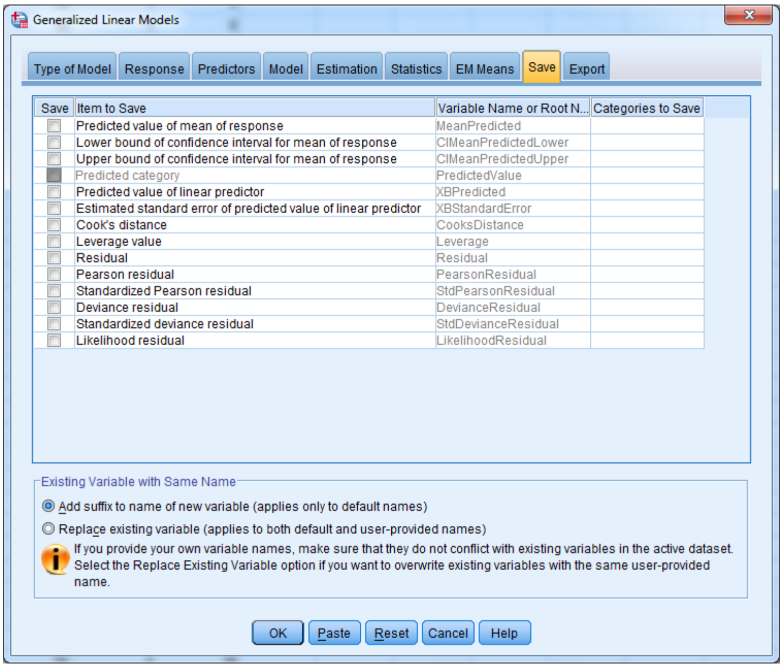
□ General estimable functions

☑ Iteration history

Print Interval: 1

□ Lagrange multiplier test of scale parameter or negative binomial ancillary parameter

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



Save	Item to Save	Variable Name or Root N...	Categories to Save
□	Predicted value of mean of response	MeanPredicted	
□	Lower bound of confidence interval for mean of response	CIMeanPredictedLower	
□	Upper bound of confidence interval for mean of response	CIMeanPredictedUpper	
□	Predicted category	PredictedValue	
□	Predicted value of linear predictor	XBPredicted	
□	Estimated standard error of predicted value of linear predictor	XBStandardError	
□	Cook's distance	CooksDistance	
□	Leverage value	Leverage	
□	Residual	Residual	
□	Pearson residual	PearsonResidual	
□	Standardized Pearson residual	StdPearsonResidual	
□	Deviance residual	DevianceResidual	
□	Standardized deviance residual	StdDevianceResidual	
□	Likelihood residual	LikelihoodResidual	

Existing Variable with Same Name

● Add suffix to name of new variable (applies only to default names)

○ Replace existing variable (applies to both default and user-provided names)

If you provide your own variable names, make sure that they do not conflict with existing variables in the active dataset. Select the Replace Existing Variable option if you want to overwrite existing variables with the same user-provided name.

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

				Parameter					
Iteration	Update Type	Number of Step-halvings	Log Likelihood[b]	(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	Newton[a]	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.

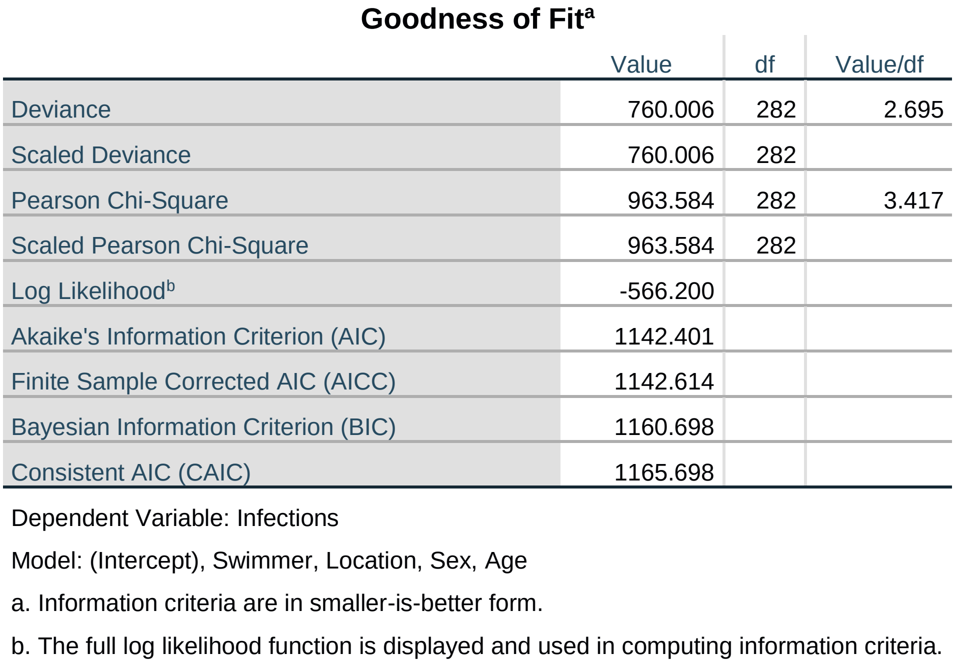
b. The full log likelihood function is displayed.

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 Fit[a]

	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 Likelihood[b]	-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

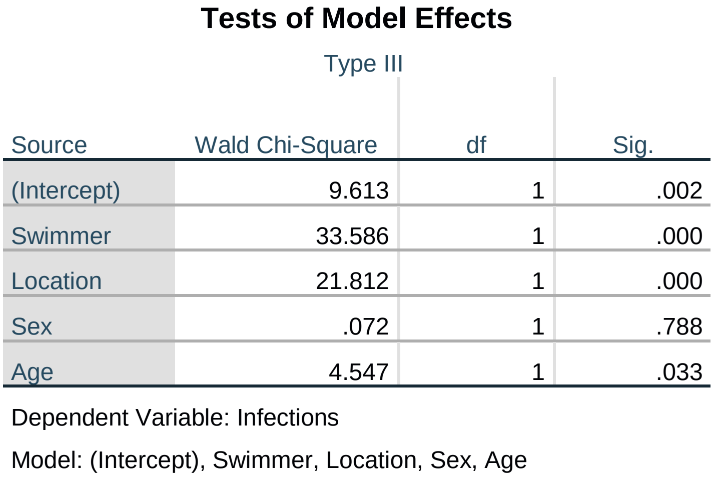
a. Information criteria are in smaller-is-better form.

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.



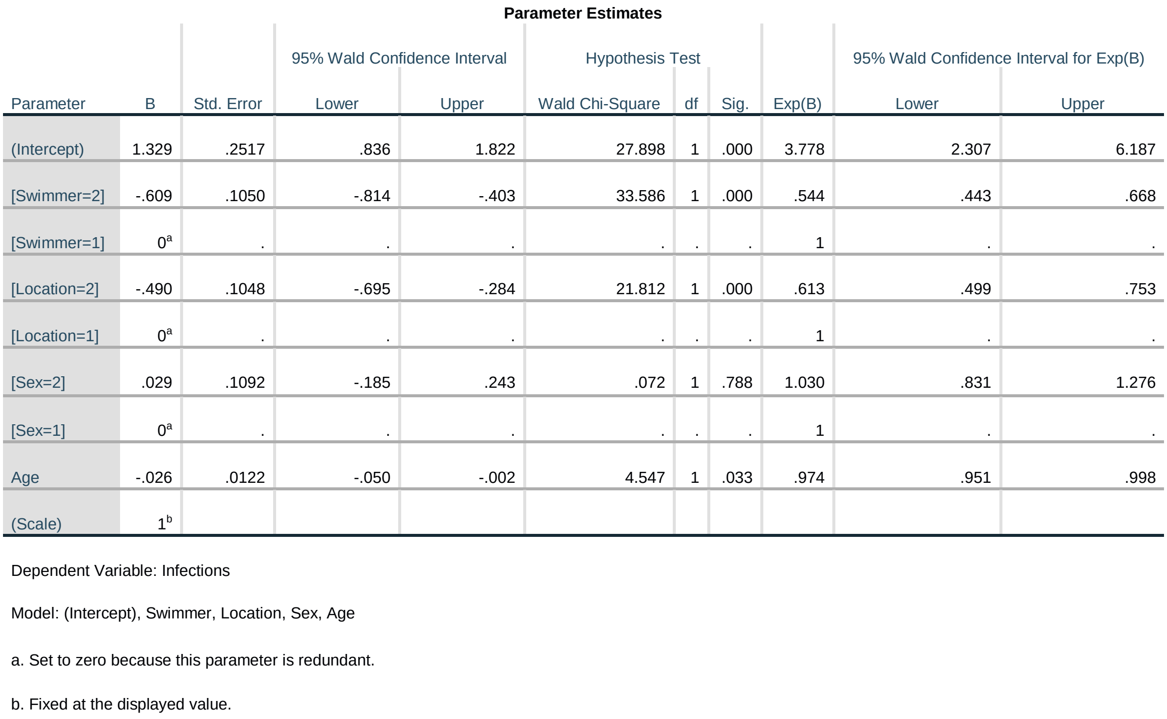
Tests of Model Effects

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

			95% Wald Confidence Interval		Hypothesis Test				95% Wald Confidence Interval for Exp(B)	
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.	Exp(B)	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]	0[a]	.	.	.	.	.	.	1	.	.
[Location=2]	-.490	.1048	-.695	-.284	21.812	1	.000	.613	.499	.753
[Location=1]	0[a]	.	.	.	.	.	.	1	.	.
[Sex=2]	.029	.1092	-.185	.243	.072	1	.788	1.030	.831	1.276
[Sex=1]	0[a]	.	.	.	.	.	.	1	.	.
Age	-.026	.0122	-.050	-.002	4.547	1	.033	.974	.951	.998
(Scale)	1[b]									

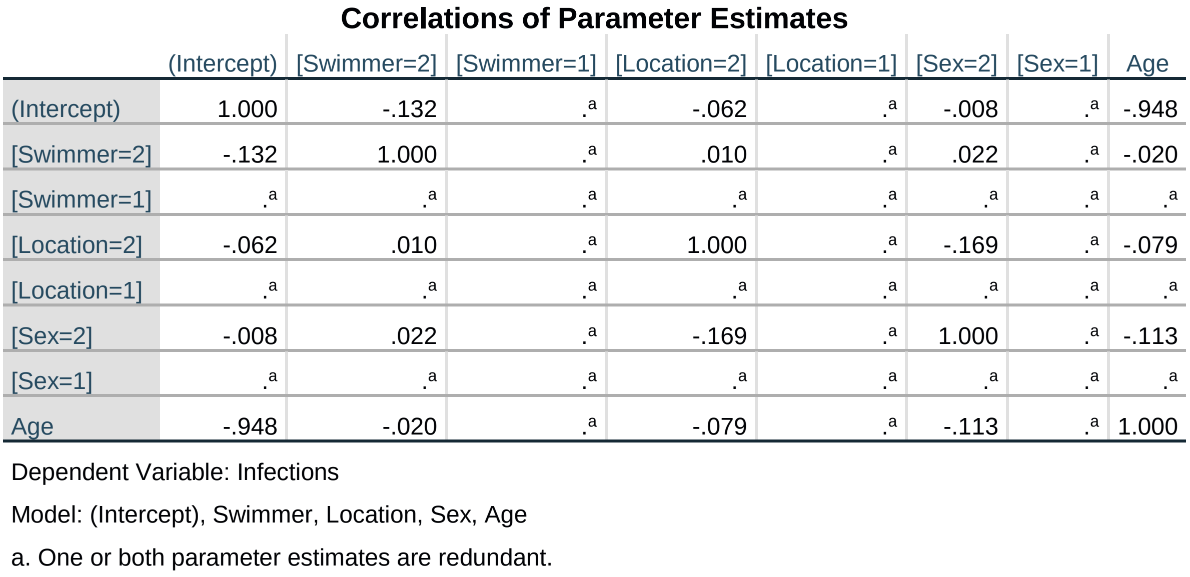
Dependent Variable: Infections

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.