Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

This document illustrates fitting generalized linear models (GLM) with various error types and link functions.

Test files and data formats

From the main SIMFIT menu choose [Statistics], [Generalized linear models], then [Comprehensive GLM options], and after selecting an error and link type view the test file provided which will be one of these.
`glm.tf1 (≡ g02gaf.tf1)`: normal error and reciprocal link
`glm.tf2 (≡ g02gbf.tf1)`: binomial error and logistic link (logistic regression)
`glm.tf3 (≡ g02gcf.tf1)`: Poisson error and log link
`glm.tf4 (≡ g02gdf.tf1)`: gamma error and reciprocal link

Here the data format for k variables, observations y and weightings s is

$$x_1, x_2, \ldots, x_k, y, s$$

except for the binomial error which has

$$x_1, x_2, \ldots, x_k, y, N, s$$

for y successes in N independent Bernoulli trials.

It is absolutely essential to have a final column of s values in the data as the number of columns is used to indicate the number of covariates. In most cases these values would be $s = 1$, but note that the weights w used are actually $w = 1/s^2$ if advanced users wish to employ weighting, e.g., using s as the reciprocal of the square root of the number of replicates for replicate weighting, except that when $s \leq 0$ the corresponding data points are suppressed. Also, observe the alternative measures of goodness of fit, such as residuals, leverages and deviances. The residuals r_i, sums of squares SSQ and deviances d_i and overall deviance depend on the error types as indicated in the examples.

GLM example 1: G02GAF, normal errors and reciprocal link

The test file `glm.tf1` contains the following data.

x	y	s
1.0	25.0	1
2.0	10.0	1
3.0	6.0	1
4.0	4.0	1
5.0	3.0	1

The next table has the results from fitting a reciprocal link with mean but no offsets to `glm.tf1`,

No. parameters = 2, Rank = 2, No. points = 5, Deg. freedom = 3

Parameter	Value	Lower95%cl	Upper95%cl	Std. error	p
Constant	-0.0238725	-0.0327174	-0.0150276	0.00277926	0.0033
B(1)	0.0638107	0.0554160	0.0722054	0.00263782	0.0002

$WSSQ = 0.387173$, $S = 0.129058$, $A = 1$

while the table of deviance residuals and leverages was as follows.

Number	Y-value	Theory	Dev-resid	Leverage
1	25.0	25.0387	-0.038665	0.995407
2	10.0	9.63865	0.361348	0.457746
3	6.0	5.96802	0.031977	0.268103
4	4.0	4.32207	-0.322074	0.166606
5	3.0	3.38775	-0.387751	0.112138

Note that the scale factor ($S = \sigma^2$) can be input or estimated using the residual sum of squares SSQ defined as follows

$$\text{For normal errors: } d_i = y_i - \hat{\mu}_i$$
$$\text{Deviance residuals: } r_i = d_i$$
$$SSQ = \sum_{i=1}^{n} r_i .$$

GLM example 2: G02GBF, binomial errors with logistic link

The next table shows the results from fitting a logistic link and mean but no offsets to test file `glm.tf2` which contains the following data for covariate x, number of successes y in N Bernoulli trials, with no weighting (i.e. all $s = 1$).

x	y	N	s
1.0	19	516	1
0.0	29	560	1
-1.0	24	293	1

No. parameters = 2, Rank = 2, No. points = 3, Deg. freedom = 1

Parameter	Value	Lower95%cl	Upper95%cl	Std. error	p
Constant	-2.86822	-4.41463	-1.32180	0.121705	0.0270
B(1)	-0.42637	-2.45654	1.60380	0.159778	0.2283 ***

Deviance = 0.0735389

Number	Y-value	Theory	Dev-resid	Leverage
1	19.0	18.4508	0.129596	0.768720
2	29.0	30.0984	-0.207027	0.422046
3	24.0	23.4508	0.117828	0.809234

The estimates are defined as follows

$$\text{For binomial errors: } d_i = 2\left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (t_i - y_i)\log\left(\frac{t_i - y_i}{t_i - \hat{\mu}_i}\right)\right\}$$
$$\text{Deviance residuals: } r_i = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$$
$$\text{Deviance } = \sum_{i=1}^{n} d_i .$$

Note that, unlike the situation with normal errors as in Example 1, the deviance residuals in the column headed as Dev-resid in the previous residuals table are not the same as the usual residuals from a regression.

GLM example 3: G02GCF, Poisson errors with a log link

This example illustrates using the choice for Poisson error and a log link to analyze a contingency table. and the test file for this option is `glm.tf3` which has columns for 8 variables x_i, then a column y for the Poisson variable, and a final column of weights $s = 1$. However, to understand the format for these data it must be pointed out that this is a representation of a 3 by 5 contingency table contained in test file `loglin.tf1`. Because there are 3 rows and 5 columns in the contingency table there will be 8 categorical variables with a 1 representing true and a 0 representing false. To clarify the situation consider the following table displaying the contingency table along with equivalent data file

Test file loglin.tf1 Test file glm.tf3

	c_1	c_2	c_3	c_4	c_5	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	y	s
r_1	141	67	114	79	39	1	0	0	1	0	0	0	0	141	1
r_2	131	66	143	72	35	1	0	0	0	1	0	0	0	67	1
r_3	36	14	38	28	16	1	0	0	0	0	1	0	0	114	1
						1	0	0	0	0	0	1	0	79	1
						1	0	0	0	0	0	0	1	39	1
						0	1	0	1	0	0	0	0	131	1
						0	1	0	0	1	0	0	0	66	1
						0	1	0	0	0	1	0	0	143	1
						0	1	0	0	0	0	1	0	72	1
						0	1	0	0	0	0	0	1	35	1
						0	0	1	1	0	0	0	0	36	1
						0	0	1	0	1	0	0	0	14	1
						0	0	1	0	0	1	0	0	38	1
						0	0	1	0	0	0	1	0	28	1
						0	0	1	0	0	0	0	1	16	1

Hence, because cell 1,1 indicates 141 number of times that category 1,1 occurred then row 1 of the data file will have a 0 everywhere except for $x_1 = 1$ and $x_4 = 1$ indicating row 1 and column 1 of the contingency table. In other words variables x_1, x_2, x_3 represent rows 1, 2, 3 in the contingency table, while variables x_4, x_5, x_6, x_7, x_8 represent columns 1, 2, 3, 4, 5 in the contingency table.

To summarize. If a contingency table T has r rows and c columns then the equivalent data file D will have rc rows and $r + c + 2$ columns. The value in contingency table cell T_{ij} will be the value in data cell D_{kl} with $k = (i - 1)c + j$ and $l = r + c + 1$. However, all the data cells D_{kl} will be zero for $l \leq r + c$ except for $l = i$ and $l = j + r$ which will be one.

The next tables show the results from fitting a log link and mean but no offsets to `glm.tf3`.

No. parameters = 9, Rank = 7, No. points = 15, Deg. freedom = 8

Parameter	Value	Lower95%cl	Upper95%cl	Std. error	p
Constant	2.59766	2.53813	2.65719	0.0258152	0.0000
B(1)	1.26195	1.16091	1.36299	0.0438171	0.0000
B(2)	1.27773	1.17714	1.37833	0.0436224	0.0000
B(3)	0.05798	-0.09595	0.21190	0.0667511	0.4104 ***
B(4)	1.03069	0.90365	1.15773	0.0550913	0.0000
B(5)	0.29102	0.12229	0.45976	0.0731714	0.0041
B(6)	0.98757	0.85859	1.11654	0.0559316	0.0000
B(7)	0.48798	0.33224	0.64371	0.0675352	0.0001
B(8)	-0.19960	-0.40795	0.00875	0.0903524	0.0582 *

Deviance = 9.03788, A = 1

Number	Y-value	Theory	Dev-resid	Leverage
1	141	132.993	0.68750	0.603533
2	67	63.4740	0.43857	0.513759
3	114	127.380	-1.20721	0.596285
4	79	77.2915	0.19363	0.531602
5	39	38.8616	0.02218	0.481976
6	131	135.109	-0.35531	0.608326
7	66	64.4838	0.18808	0.519638
8	143	129.406	1.17492	0.601167
9	72	78.5211	-0.74647	0.537265
10	35	39.4799	-0.72715	0.488239
11	36	39.8979	-0.62759	0.392649
12	14	19.0422	-1.21309	0.255123
13	38	38.2139	-0.03464	0.381546
14	28	23.1874	0.96754	0.282457
15	16	11.6585	1.20279	0.206435

The definitions are

$$\text{For Poisson errors: } d_i = 2\left\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right\}$$

$$\text{Deviance residuals: } r_i = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$$

$$\text{Deviance } = \sum_{i=1}^{n} d_i,$$

but note that an error message is output to warn you that the solution is overdetermined, i.e., the parameters and standard errors are not unique.

Thus, in order to obtain unique parameter estimates, it is necessary to impose constraints so that the resulting constrained system is of full rank. Let the singular value decomposition (SVD) P^* be represented, as in G02GKF, by

$$P^* = \begin{pmatrix} D^{-1}P_1^T \\ P_0^T \end{pmatrix},$$

and suppose that there are m parameters and the rank is r, so that there need to be $n_c = m - r$ constraints, for example, in a m by n_c matrix C where

$$C^T\beta = 0.$$

Then the constrained estimates $\hat{\beta}_c$ are given in terms of the SVD parameters $\hat{\beta}_{svd}$ by

$$\begin{aligned}\hat{\beta}_c &= A\hat{\beta}_{svd} \\ &= (I - P_0(C^T P_0)^{-1}C^T)\hat{\beta}_{svd},\end{aligned}$$

while the variance-covariance matrix V is given by

$$V = AP_1D^{-2}P_1^T A^T,$$

provided that $(C^T P_0^{-1})$ exists.

This approach is commonly used in log-linear analysis of contingency tables, but it can be tedious to first fit the overdetermined Poisson GLM model then apply a matrix of constraints as just described. For this reason SimFIT provides an automatic procedure to calculate the dummy indicator matrix from the contingency table then fit a log-linear model and apply the further constraints that the sum of row effects and sum of column effects are zero.

This simplified GLM log-linear analysis of contingency tables is available from the SIMFIT main menu [Statistics] option using either the [Standard statistical tests] sub-menu or the [Generalized linear models] options.

For instance, the next table illustrates how this is done with `loglin.tf1` using the GLM log-linear contingency table analysis procedure to read in a contingency table, fit a Poisson model, then apply the correction to apply the equations of constraint

$$\sum_{i=1}^{ncol} \text{Column parameter}_i = 0$$

$$\sum_{i=1}^{nrow} \text{Row parameter}_i = 0$$

to obtain well-defined parameter estimates.

No. rows = 3, No. columns = 5
Deviance (D) = 9.03788E+00, Deg. freedom = 8
$P(\chi^2 \geq D)$ = 0.3391

Parameter	Value	Lower95%cl	Upper95%cl	Std. error	p	
Constant	3.98308	0.0395833	3.89180	4.07435	0.0000	
Row 1	0.39606	0.0458291	0.29038	0.50175	0.0000	
Row 2	0.41185	0.0456995	0.30646	0.51723	0.0000	
Row 3	-0.80791	0.0621905	-0.95132	-0.66450	0.0000	
Col 1	0.51116	0.0561557	0.38166	0.64065	0.0000	
Col 2	-0.22851	0.0727114	-0.39618	-0.06084	0.0137	*
Col 3	0.46804	0.0569148	0.33679	0.59933	0.0000	
Col 4	-0.03156	0.0675080	-0.18723	0.12412	0.6527	***
Col 5	-0.71913	0.0887225	-0.92373	-0.51454	0.0000	

Data	Model	Delta	Dev-resid	Leverage
141	132.9931	8.0069	0.6875	0.6035
67	63.4740	3.5260	0.4386	0.5138
114	127.3798	-13.3798	-1.2072	0.5963
79	77.2915	1.7085	0.1936	0.5316
39	38.8616	0.1384	0.0222	0.4820
131	135.1089	-4.1089	-0.3553	0.6083
66	64.4838	1.5162	0.1881	0.5196
143	129.4063	13.5937	1.1749	0.6012
72	78.5211	-6.5211	-0.7465	0.5373
35	39.4799	-4.4799	-0.7271	0.4882
36	39.8979	-3.8979	-0.6276	0.3926
14	19.0422	-5.0422	-1.2131	0.2551
38	38.2139	-0.2139	-0.0346	0.3815
28	23.1874	4.8126	0.9675	0.2825
16	11.6585	4.3415	1.2028	0.2064

GLM example 4: G02GDF, gamma errors with a reciprocal link

The next tables show the results from fitting a reciprocal link and mean but no offsets to `glm.tf4`.

No. parameters = 2, Rank = 2, No. points = 10, Deg. freedom = 8

Parameter	Value	Lower95%cl	Upper95%cl	Std. error	p
Constant	1.44085	-0.08812	2.96981	0.663037	0.0615 *
B(1)	-1.28653	-2.82436	0.25131	0.666882	0.0898 *

Adjusted Deviance = 35.0344, S = 1.07418, A = 1

Number	Y-value	Theory	Dev-resid	Leverage
1	1.00	6.48000	-1.39085	0.2
2	0.30	6.48000	-1.92278	0.2
3	10.5	6.48000	0.52365	0.2
4	9.70	6.48000	0.43179	0.2
5	10.9	6.48000	0.56784	0.2
6	0.62	0.69404	-0.11071	0.2
7	0.12	0.69404	-1.32870	0.2
8	0.09	0.69404	-1.48152	0.2
9	0.50	0.69404	-0.31063	0.2
10	2.14	0.69404	1.36648	0.2

Note that with gamma errors, the scale factor (ν^{-1}) can be input or estimated using the degrees of freedom, k, and

$$\hat{\nu}^{-1} = \sum_{i=1}^{n} \frac{[(y_i - \hat{\mu}_i]/\hat{\mu}_i]^2}{N-k}$$

$$\text{For gamma errors: } d_i = 2\left\{\log(\hat{\mu}_i) + \left(\frac{y_i}{\hat{\mu}_i}\right)\right\}$$

$$\text{Deviance residuals: } r_i = \frac{3(y_i^{\frac{1}{3}} - \hat{\mu}_i{}^{\frac{1}{3}})}{\hat{\mu}_i{}^{\frac{1}{3}}}$$

$$\text{Deviance: } = \sum_{i=1}^{n} d_i$$