SPIDERFINANCIAL
www.spiderfinancial.com

Phone: 1-888-427-9486
1-312-324-0367
Fax: 1-312-238-9092
info@spiderfinancial.com

# Tutorial: GLM with NumXL

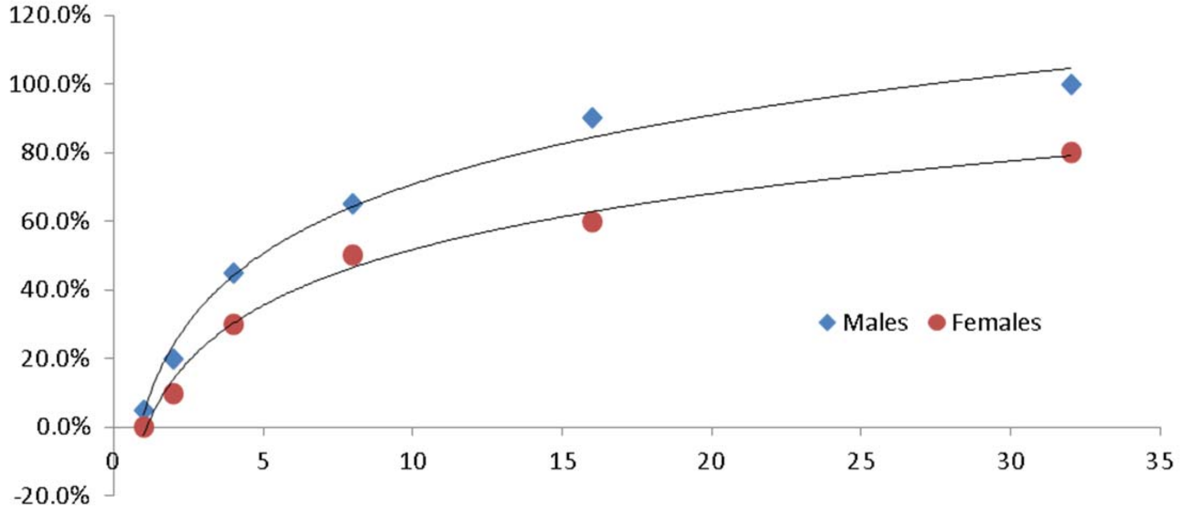
In this tutorial, we will use a sample data gathered during a clinical trial of a new chemical/pesticide on tobacco Budworms. The subjects (i.e. budworms) are grouped into batches of 20, and exposed to different doses of the chemical. The results are summarized below:

| Batch | Dose | Gender | Death |
|---|---|---|---|
| 1 | 1 | 0 | **1** |
| 2 | 2 | 0 | **4** |
| 3 | 4 | 0 | **9** |
| 4 | 8 | 0 | **13** |
| 5 | 16 | 0 | **18** |
| 6 | 32 | 0 | **20** |
| 7 | 1 | 1 | **0** |
| 8 | 2 | 1 | **2** |
| 9 | 4 | 1 | **6** |
| 10 | 8 | 1 | **10** |
| 11 | 16 | 1 | **12** |
| 12 | 32 | 1 | **16** |

## Data preparation

Our objective here is to model (and forecast) the effectiveness of the new chemical using different dosages, and explain, to some extent, any variation based on the gender of the budworm. Furthermore, we want to express the results in term of the worm mortality rates (i.e. probability).

| Batch | Dose | Gender | Death | Rate |
|---|---|---|---|---|
| 1 | 1 | 0 | **1** | 5.0% |
| 2 | 2 | 0 | **4** | 20.0% |
| 3 | 4 | 0 | **9** | 45.0% |
| 4 | 8 | 0 | **13** | 65.0% |
| 5 | 16 | 0 | **18** | 90.0% |
| 6 | 32 | 0 | **20** | 100.0% |
| 7 | 1 | 1 | **0** | 0.0% |
| 8 | 2 | 1 | **2** | 10.0% |
| 9 | 4 | 1 | **6** | 30.0% |
| 10 | 8 | 1 | **10** | 50.0% |
| 11 | 16 | 1 | **12** | 60.0% |
| 12 | 32 | 1 | **16** | 80.0% |



We plot the data into two separate curves: males and females. It is apparent that mortality rate is affected by those two factors: gender and dosage.

We will make two assumptions: (1) the results for each trial (i.e. batch) are drawn from a Binomial distributed population; we would like to estimate p - the probability of success (i.e. worm's death). The probability (p) is allowed to vary across different trials (batches). (2) The probability of success is affected by two factors: gender of the subject and administered dosage of the drug.

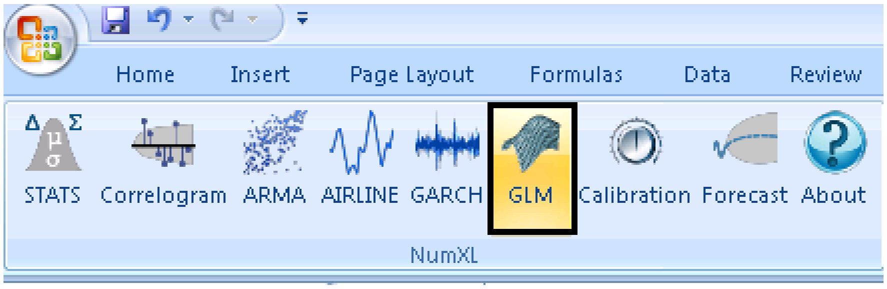
Based on these two assumptions, we would model this relationship:

$$P = f(X,Y) = E[p \mid X,Y]$$

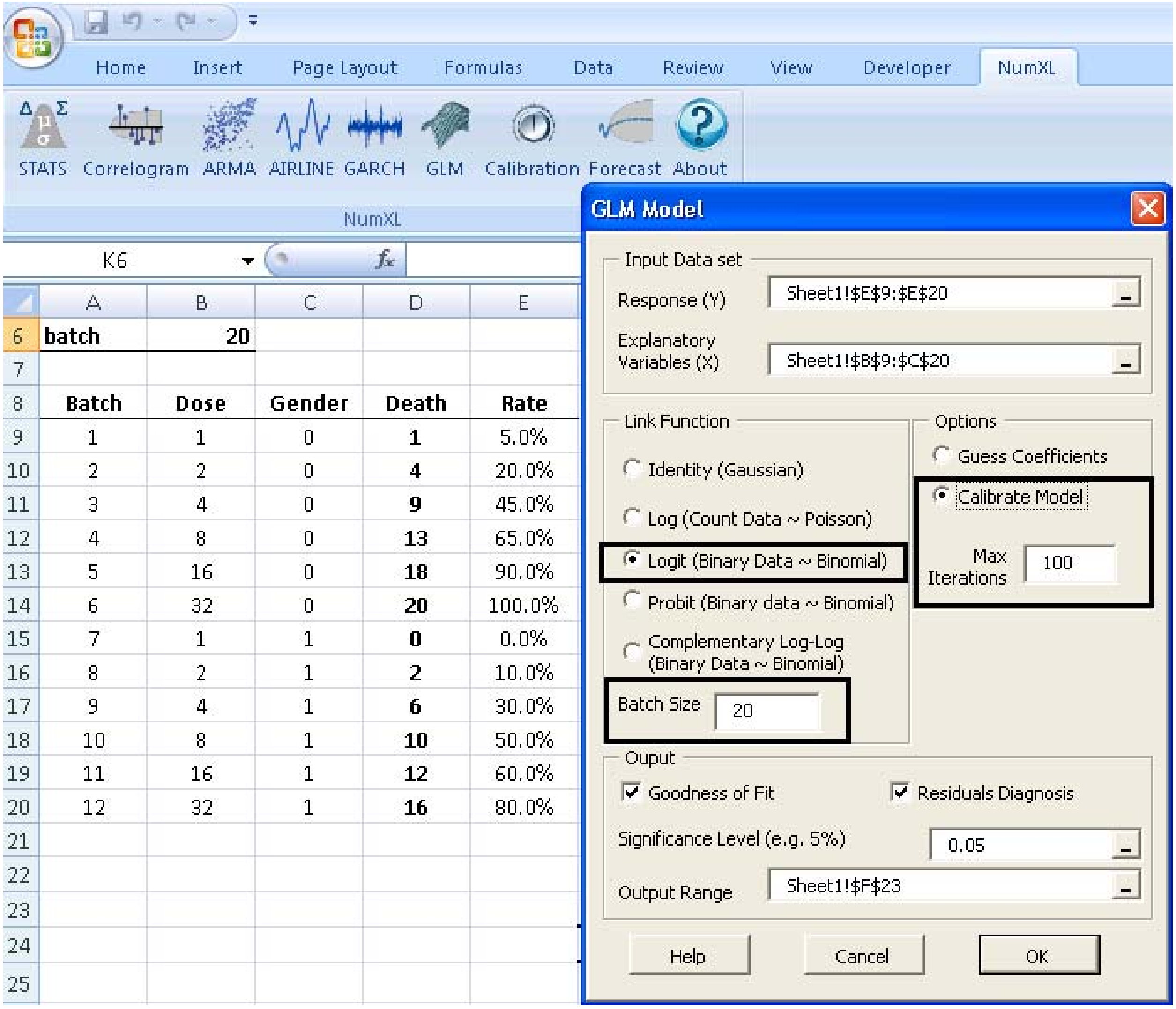
## Modeling

We are ready now to propose a statistical model: generalized linear model with residuals following the Binomial distribution.

Using NumXL toolbar, click on the "GLM" icon.



The GLM wizard will pop up. Initially, all the controls are disabled until we specify a valid range for response and explanatory variables. The number of rows of the two cells ranges must match.



|  | A | B | C | D | E |
|---|---|---|---|---|---|
| 6 | batch | 20 |  |  |  |
| 7 |  |  |  |  |  |
| 8 | **Batch** | **Dose** | **Gender** | **Death** | **Rate** |
| 9 | 1 | 1 | 0 | **1** | 5.0% |
| 10 | 2 | 2 | 0 | **4** | 20.0% |
| 11 | 3 | 4 | 0 | **9** | 45.0% |
| 12 | 4 | 8 | 0 | **13** | 65.0% |
| 13 | 5 | 16 | 0 | **18** | 90.0% |
| 14 | 6 | 32 | 0 | **20** | 100.0% |
| 15 | 7 | 1 | 1 | **0** | 0.0% |
| 16 | 8 | 2 | 1 | **2** | 10.0% |
| 17 | 9 | 4 | 1 | **6** | 30.0% |
| 18 | 10 | 8 | 1 | **10** | 50.0% |
| 19 | 11 | 16 | 1 | **12** | 60.0% |
| 20 | 12 | 32 | 1 | **16** | 80.0% |

GLM Model

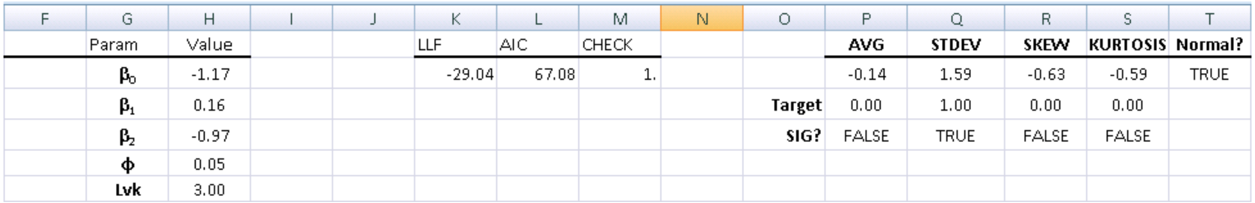
Input Data set
- Response (Y): Sheet1!$E$9:$E$20
- Explanatory Variables (X): Sheet1!$B$9:$C$20

Link Function
- ○ Identity (Gaussian)
- ○ Log (Count Data ~ Poisson)
- ● Logit (Binary Data ~ Binomial)
- ○ Probit (Binary data ~ Binomial)
- ○ Complementary Log-Log (Binary Data ~ Binomial)
- Batch Size: 20

Options
- ○ Guess Coefficients
- ● Calibrate Model
- Max Iterations: 100

Ouput
- ☑ Goodness of Fit
- ☑ Residuals Diagnosis
- Significance Level (e.g. 5%): 0.05
- Output Range: Sheet1!$F$23

For now, we choose “Logit” as our link (transform) function, specify the trial or batch size(20), and instruct the Wizard to calibrate (i.e. compute optimal values for the coefficients). Leave the Goodness-of-fit and residual diagnosis options checked.



| F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | Param | Value |  |  | LLF | AIC | CHECK |  |  | **AVG** | **STDEV** | **SKEW** | **KURTOSIS** | **Normal?** |
|  | $\beta_0$ | -1.17 |  |  | -29.04 | 67.08 | 1. |  |  | -0.14 | 1.59 | -0.63 | -0.59 | TRUE |
|  | $\beta_1$ | 0.16 |  |  |  |  |  |  | **Target** | 0.00 | 1.00 | 0.00 | 0.00 |  |
|  | $\beta_2$ | -0.97 |  |  |  |  |  |  | **SIG?** | FALSE | TRUE | FALSE | FALSE |  |
|  | $\phi$ | 0.05 |  |  |  |  |  |  |  |  |  |  |  |  |
|  | **Lvk** | 3.00 |  |  |  |  |  |  |  |  |  |  |  |  |

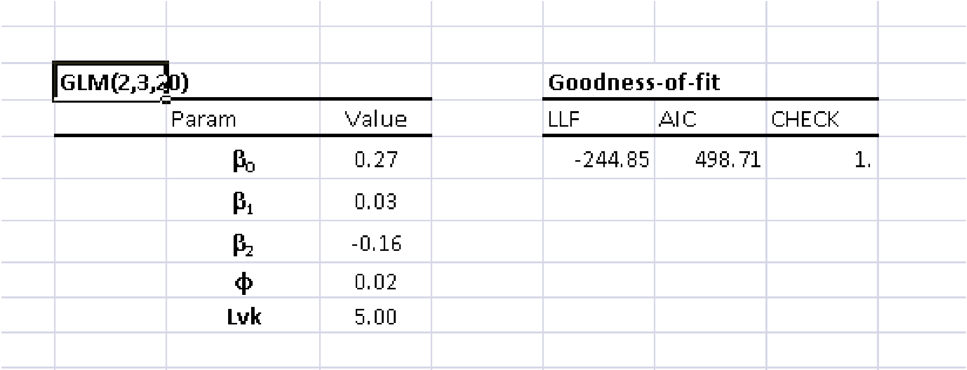
## Calibration

In this case, the GLM Wizard has calibrated the model's coefficients, so we can skip this step.

But, in the event we wish to experiment with different link functions: LOGIT, PROBIT or LOG-LOG, then we need to re-calibrate the model. To do so, we can either:

(1) Create a new model with the wizard, or,
(2) Change the **"Lvk"** parameter in an existing model table, and run the calibration using NumXL toolbar.

## Step 1: Select the cell that acts as a header for the model table



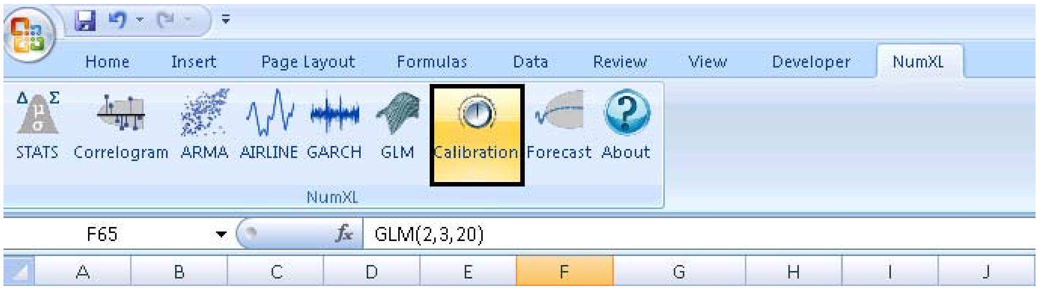
GLM(2,3,20)

| Param | Value |
|---|---|
| $\beta_0$ | 0.27 |
| $\beta_1$ | 0.03 |
| $\beta_2$ | -0.16 |
| $\phi$ | 0.02 |
| Lvk | 5.00 |

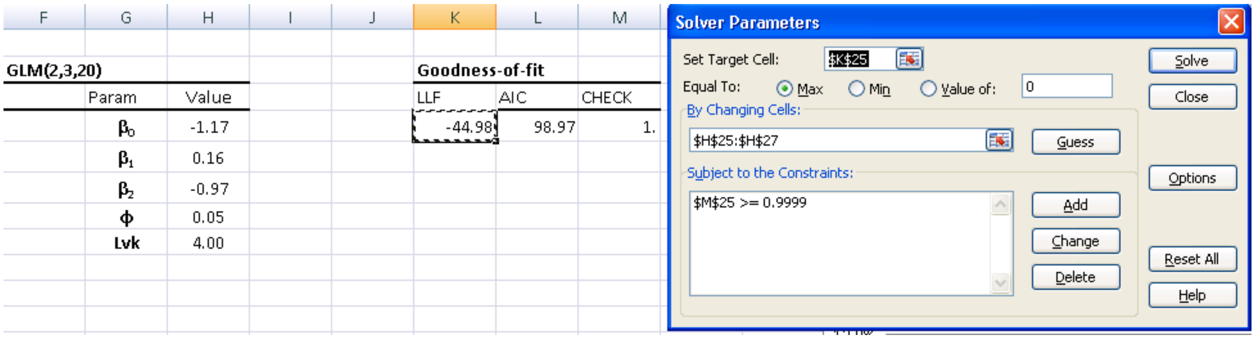
Goodness-of-fit

| LLF | AIC | CHECK |
|---|---|---|
| -244.85 | 498.71 | 1. |

## Step 2: Click on the calibration icon/menu (Excel 2003)



## Step 3: Click on "Solve" button in the Solver window



| GLM(2,3,20) | |
|---|---|
| Param | Value |
| $\beta_0$ | -1.17 |
| $\beta_1$ | 0.16 |
| $\beta_2$ | -0.97 |
| $\phi$ | 0.05 |
| Lvk | 4.00 |

| Goodness-of-fit | | |
|---|---|---|
| LLF | AIC | CHECK |
| -44.98 | 98.97 | 1. |

Set Target Cell: $K$25

Equal To: ● Max ○ Min ○ Value of: 0

By Changing Cells: $H$25:$H$27

Subject to the Constraints: $M$25 >= 0.9999

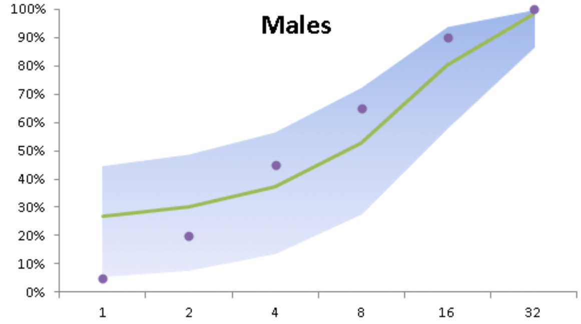
## Forecast

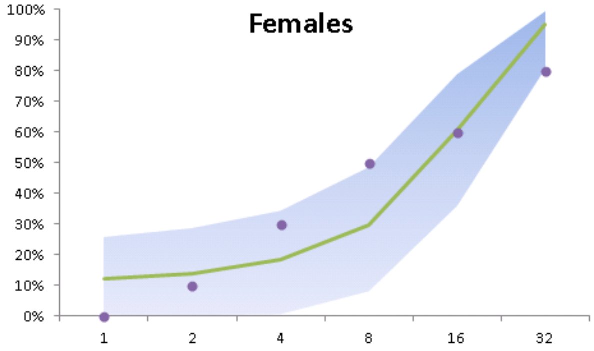
Once the model is calibrated, and we are happy with the residuals, we can use it to construct our forecast mean (and confidence interval around it).

Using NumXL function (GLM_FORE), we can compute the mean. Using GLM_FORECI, we can compute the upper and lower limit of the confidence interval.

| batch | 20 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Forecast | | | | |
| Batch | Dose | Gender | Death | Rate | | Mean | STD | UL | LL | Range |
| 1 | 1 | 0 | **1** | 5.0% | | 27% | 10% | 45% | 6% | 39% |
| 2 | 2 | 0 | **4** | 20.0% | | 30% | 10% | 49% | 8% | 41% |
| 3 | 4 | 0 | **9** | 45.0% | | 37% | 11% | 57% | 14% | 43% |
| 4 | 8 | 0 | **13** | 65.0% | | 53% | 11% | 72% | 28% | 44% |
| 5 | 16 | 0 | **18** | 90.0% | | 80% | 9% | 94% | 59% | 35% |
| 6 | 32 | 0 | **20** | 100.0% | | 98% | 3% | 100% | 87% | 13% |
| 7 | 1 | 1 | **0** | 0.0% | | 12% | 7% | 26% | 0% | 25% |
| 8 | 2 | 1 | **2** | 10.0% | | 14% | 8% | 29% | 0% | 29% |
| 9 | 4 | 1 | **6** | 30.0% | | 18% | 9% | 34% | 0% | 34% |
| 10 | 8 | 1 | **10** | 50.0% | | 30% | 10% | 49% | 8% | 41% |
| 11 | 16 | 1 | **12** | 60.0% | | 60% | 11% | 79% | 36% | 43% |
| 12 | 32 | 1 | **16** | 80.0% | | 95% | 5% | 100% | 81% | 19% |

Plotting the data again (actual) versus the model values.





The dots represent the sample data, while the center line is the forecast mean. The shaded regions in the graphs are the 95% confidence intervals.

Notes:

1. The forecast error decrease as we increase the dosage (C.I. gets tighter). This is evident in male and female batches.
2. The logarithmic relation detected when we plot the raw data can be merely a data anomaly; the GLM shows more like a quadratic-type of relationship.
3. The mean is not exactly the center of the confidence interval due to the discrete-nature of the underlying binomial distribution, and the small batch/trial size.