Contents

1 Generalized Linear Models **2**
1.1 Components of a GLM . . . 2
1.1.1 Random component . . . 2
1.1.2 Systematic component . . . 3
1.1.3 Link function . . . 3
1.1.4 Exponential Family . . . 3
1.1.5 For example: Normal random component . . . 4
1.2 GLMs for binary data . . . 4
1.2.1 Identity . . . 5
1.2.2 Logit . . . 7
1.2.3 Probit . . . 9
1.3 GLMs for count data . . . 12
1.3.1 Poisson distribution . . . 12
1.3.2 Poisson Regression/Loglinear Model for modelling count data . . . 13
1.4 Inference about Model Parameters in GLMs . . . 15
1.5 Model Fit and Comparison . . . 16
1.5.1 Model Utility Test . . . 16
1.5.2 Deviance . . . 17
1.5.3 Comparing Models . . . 18
1.5.4 Residuals for Comparing Observations to the fitted model . . . 19

1 Generalized Linear Models

Models (for describing the population):

- relates explanatory variables with response variable
- can handle complex questions
- describes pattern of association (interaction)
- parameters relate the nature and strength of the association
- test for association based on sample data while controlling for confounding variables
- model can be used for prediction

The Generalized Linear Model (GLM) is a model which can be specified to include a wide range of different models, e.g. ANOVA and multiple linear regression models are just special cases of this model.
We will be interested in the models that relate categorical response data to categorical and numerical explanatory variables.

1.1 Components of a GLM

- Random component
- Systematic component
- Link function to link random and systematic component

1.1.1 Random component

The random component identifies the response variable Y and gives the distribution of Y

- For a multiple linear regression model the distribution of Y is normal.
- Here we will study categorical outcomes and in particular binary outcomes (success/failure). In these cases it is not meaningful to assume normality. In this case the Y has a either a binomial or multinomial distribution.
- In general the different observations are assumed to be independent and are denoted by $Y_1, \ldots, Y_n$).

(Use uppercase Y for the response because it has a random effect)

1.1.2 Systematic component

The systematic component specifies the explanatory variables. Since this is a linear model they are entered in a linear combination into the model, the "linear predictor".

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Here we allow that some of the explanatory or predictor variables are interaction variables $x_3 = x_1 x_2$ or even $x_4 + x_1^2$ to allow polynomial relationships.
This is called the *linear predictor*.
In multiple linear regression this is the right hand side of the model equation (omitting the error).
(Use lowercase x for the explanatory variables to indicate that their outcome is not random)

1.1.3 Link function

The link function links the random component with the systematic component. More specific if μ is the mean of Y, then the link function relates the μ with the linear predictor.

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- In multiple linear regression $g(\mu) = \mu$ (identity)

  $$\mu = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

  because the mean error is assumed to be zero.

- In the case when g is the logarithm $g(\mu) = ln(\mu)$

  $$log(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

  the model is called a loglinear model used for count data (e.g.: number of parties per month).

- The logit link function is $g(\mu) = ln(\mu/(1-\mu))$. Since for binary variables $\mu = \pi$ the model relates the log odds with the predictor variables.

  $$log(\mu/(1-\mu)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

  The GLM is then called logistic regression model.

1.1.4 Exponential Family

The exponential family is a class of probability distributions, including the normal distribution, binomial distribution, and Poisson distribution.
It permits developing and proving results in very general settings. Tools do not have to be developed for each distribution separately, it can be done in general for the exponential family and then applied to each distribution of interest.
Usually to show, that a certain distribution is part of this family we have to rewrite the distribution function in a certain form (pattern). This form will then show a natural way, how the probabilities depend on the parameters of the distribution, these are then called the *natural parameters* of the distribution.

Probability distributions and their natural parameters:
normal distribution (σ known): μ, with μ = mean
Poisson distribution: $\ln(\mu)$, with μ = mean
binomial distribution: $\ln(\pi/(1-\pi))$, with π = success probability

The link function using the natural parameter is called the *canonical link*, and is the one most commonly applied.

1.1.5 For example: Normal random component

Models for numerical response variables, like ANOVA and Linear Regression are special cases of GLMs.
For these models the

- random component has a normal distribution
- systematic component $\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$
- link function = identity, i.e. $g(\mu) = \mu$.

GLMs generalize these models by allowing other than normal distributions for the error, and permitting to model a function of the mean.
Because of this transformation of the data to make it normal with constant standard deviation is not necessary any more. But now the problem is to find the proper distribution and link function.
The GLM is a "Super Model" which unifies many different models under one umbrella. Generally maximum likelihood estimators are used instead of least squares estimators.

1.2 GLMs for binary data

In this section binary response variables only allowing outcomes success or failure are studied. Such response variables have a binomial distribution. The probability for success is denoted by π, which implies the probability of failure to be $1 - \pi$.
Usually success is coded by a 1 and failure by 0. Single observations can be viewed as having a binomial distributions with $n = 1$, also called a Bernoulli distribution.

Modelling a binomial response variable Y with a GLM:

1. Random Component: Because Y has a binomial distribution the random component for binary data is binomial.

2. Systematic Component: The GLMs we will be looking at will model the success probability in dependency on explanatory variables. In this section, for introducing the basic principles, the easiest case with only one predictor, x, is considered.

   Therefore the systematic component for this section is the linear predictor in one variable

   $$\alpha + \beta x$$

   In order to express that the probability of success depends on explanatory variable x write $\pi(x)$.

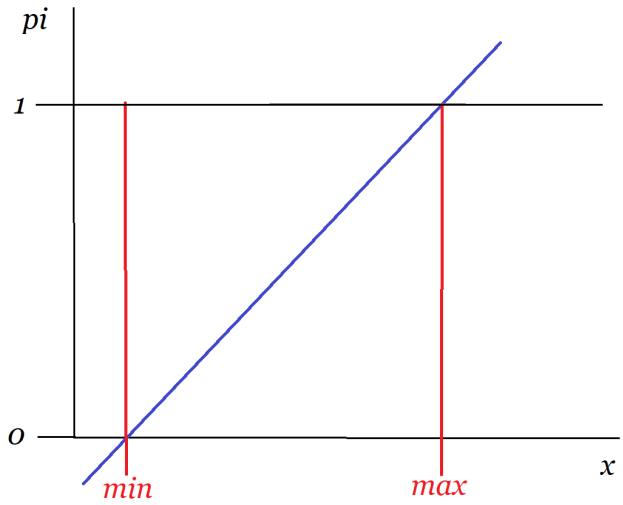
3. Link Function: The different choices for the link function, identity, logit, and probit, are presented in the following subsections.

1.2.1 Identity

When choosing the identity link function $g(\mu) = \mu$ and observing that for binomial distributions $\mu = \pi$, and that π depends on x, we get using the simple linear predictor the following model

$$g(\mu) = \mu = \pi(x) = \alpha + \beta x$$

which is called the *linear probability model*.
The big problem with this model is that probabilities are always numbers between 0 and 1, but that for sufficiently large values of x, we will get either values for $\alpha + \beta x$ greater than 1 ($\beta > 0$), or less than 0 ($\beta < 0$). Therefore this model only applies for a limited range of x.
In the diagram below only values of x between min and max result in reasonable values for π.



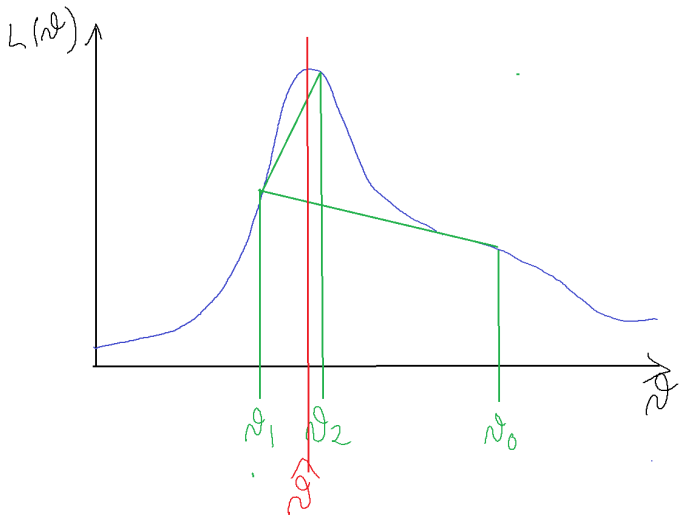
This becomes even more problematic in the presence of more than one predictor variable.

In general there do not exist closed form expressions (formulas) for calculating ML-estimates for the model parameters of generalized linear models (here: α and β) from data. The likelihood function is so complex that no closed form for its maximum can be determined.
Fortunately one area in mathematics, Numerical Analysis, deals with this type of problems and provides iterative methods for finding solutions for such situations. Such methods start with a value for the estimate which then in small steps is improved until not much improvement can be observed. The final result is then used as estimate.

To get a sense how this might work please refer to the illustration below:
The goal is to find $\hat{\vartheta}$, the value of the parameter ϑ, which makes the likelihood function, $L(\vartheta)$ the largest.

1. The algorithm starts with some guess, ϑ_0
2. Then by moving in the direction where the likelihood function is increasing, the algorithm finds a better value, ϑ_1
3. Moving from ϑ_1 again in the direction of increasing $L(\vartheta)$, another value can be found that is even closer to $\hat{\vartheta}$, ϑ_2

4. another iteration would find a value that is almost equal to $\hat{\vartheta}$, and would be used as an approximation to $\hat{\vartheta}$.



When reading the computer output for GLM analyses it will always state the number of iterations, which is the number of steps taken to improve the first guess.

Example 1
High school and Beyond (from C.J. Anderson, University of Idaho, introduced in a course on categorical data analysis)
The data has been collected from a large-scale longitudinal study conducted by The National Opinion Research Centre in 1980. The students in the sample are randomly selected high school seniors from the US.

- Question: Is the probability of attending an academic program related to achievement, x?
- Data from high school seniors (N=600).
- Response: Did students attend an academic high school program type or a non-academic program type (Y = program)? Coding: Program=1, if a student was enrolled in an academic preparatory program, otherwise Program = 0.
- Explanatory variables: Scores on 5 standardized achievement tests (Reading, Writing, Math, Science, and Civics), x being the total of the scores.

Let $\pi(x)$ be the probability to attend an academic program given a total score of x on the achievement tests. The the proposed model is

$$\pi(x) = \alpha + \beta x$$

The SPSS output for fitting the model to the data is

Iteration History

Iteration	Update Type	Number of Step-halvings	Log Likelihood[a]	Parameter			Number of Invalid Cases[b]
				(Intercept)	x	(Scale)	
0	Initial	0	-347.894	-.968450	.005700	1	4
1	Scoring	0	-347.556	-.848214	.005253	1	
2	Newton	0	-347.414	-.897262	.005432	1	2
3	Newton	1	-347.413	-.893459	.005419	1	2
4	Newton	5	-347.413[c]	-.893459	.005419	1	2

Redundant parameters are not displayed. Their values are always zero in all iterations.
Dependent Variable: program
Model: (Intercept), x

a. The full log likelihood function is displayed.

b. These cases are invalid because their mean estimates are out of range for the identity link function. Invalid cases are excluded from iterations in which they cause computational errors.

c. The maximum number of step-halvings was reached but the log likelihood value cannot be further improved.

After 4 iterations the estimated model is given as

$$\hat{\pi}(x) = -0.893 + .0054x$$

The interpretation of the model parameter is exactly like in Simple Linear Regression:
Intercept: The proportion of students in academic high school programs with a total score of $x = 0$ would be estimated to be -0.89. A negative proportion is not making sense, one of the reason that this is happening is that $x = 0$ is not within the observed range of x.
Slope: For every extra point in the total score of the achievement tests, x, the probability to be in an academic high school program increases by 0.0054.

When using the estimated model to find the probability for being in an academic program for the sample data, two cases have estimated probabilities outside the range from 0 to 1. Demonstrating the shortcomings of this model.
To avoid this from happening one should use a different model that makes it impossible to leave the range from 0 to 1.

1.2.2 Logit

The model based on the logit link function,

$$g(\mu) = logit(\mu) = \ln\left(\frac{\mu}{1-\mu}\right),$$

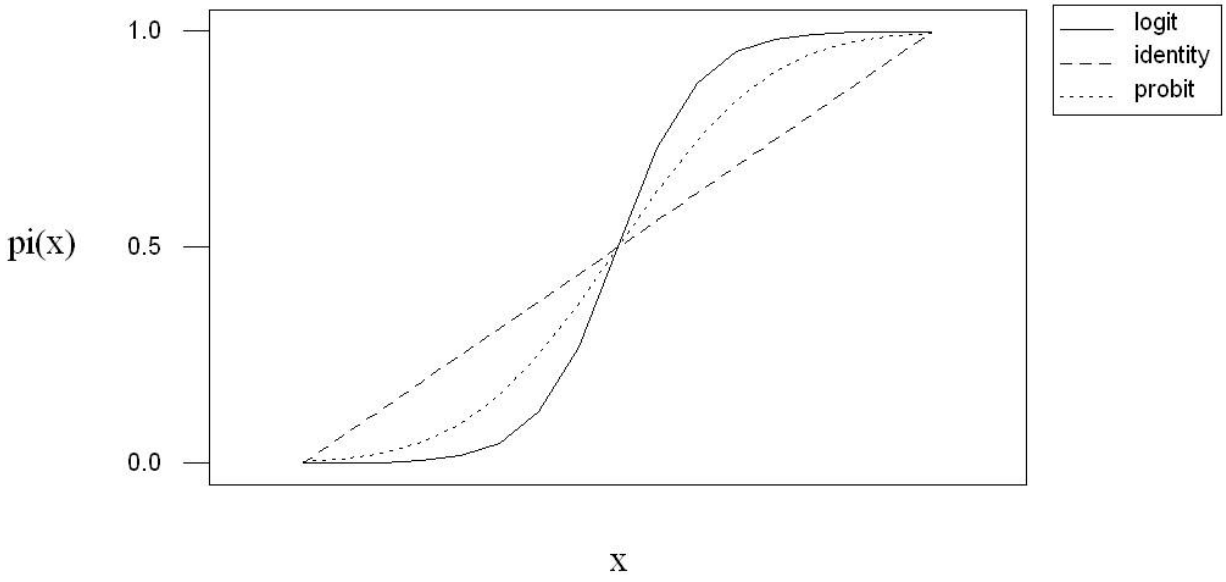
which combined with the linear predictor results in the following model for $\pi(x)$

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x,$$

which can be rewritten as

$$\pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$

The model is called the **logistic regression model**.
One can observe that the link function is the natural parameter of the binomial distribution.



Instead of a linear relationship between x and $\pi(x)$ the function is S-shaped.
The advantage of this model over the linear model, is that always

$$0 \leq \exp(\alpha + \beta x)/(1 + \exp(\alpha + \beta x)) \leq 1,$$

or $0 \leq \pi(x) \leq 1$ for all values of x.
Properties:

- When $\beta > 0$, then $\pi(x)$ increases as x increases.
- When $\beta < 0$, then $\pi(x)$ decreases as x increases.
- When $\beta = 0$ then the curve is a horizontal line.
- The larger $|\beta|$, the steeper the increase (decrease).

The *median effective level* is the value of the predictor for which the probability for success equals 0.5. For the simple logistic regression model the median effective level of x can be easily computed as

$$EL_{50} = -\frac{\alpha}{\beta}$$

Example 2
The maximum likelihood estimates for the logistic regression model using the HSB data.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-7.055	.6948	-8.417	-5.693	103.104	1	.000
x	.027	.0027	.022	.033	106.411	1	.000
(Scale)	1[a]						

Dependent Variable: program
Model: (Intercept), x

a. Fixed at the displayed value.

The estimated model equation is:

$$logit(\hat{\pi}(x)) = -7.055 + 0.027x$$

Since $\beta > 0$, this means the higher the total score the higher the probability a person attends an academic high school program. The intercept means, that the probability for a person to have attended an academic program having a total score of 0 equals $exp(-7.055)/(1 + exp(-7.055)) = 0.0009 \approx 0$.
The median effective level is estimated as $EL_{50} = 261.3$. On average a student with 261 point on the test has 50% chance to be in an academic program.

1.2.3 Probit

Another S-shaped curve used to link the probability for success with the linear predictor is the cumulative distribution function of a standard normal distribution, Φ (see graph above).
As a reminder: The cumulative distribution function, F, of a distribution of a random variable X is given by

$$F(x) = P(X \leq x), -\infty < x < \infty$$

Therefore this function is small for small values of x and increases as x increases. The normal cumulative distribution function shows the S-shape illustrated above.

The probit model with one predictor, x, and the linear predictor as systematic component can then be written as

$$\pi(x) = \Phi(\alpha + \beta x)$$

which is equivalent to

$$\Phi^{-1}\pi(x) = \alpha + \beta x$$

(Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution)
The last equation implies that the link function of the probit model is the Inverse Cumulative Distribution Function of the standard normal distribution, also called the probit function.
Φ^{-1} assigns to numbers,μ, between 0 and 1 the value, $g(\mu)$, such that
$P(Z \leq g(\mu)) = \mu$, which is the μ^{th} percentile of the standard normal distribution.

For example $probit(0.5) = 0$, since $\Phi(0) = P(Z \leq 0) = 0.5$.

From this discussion we derive that for the Probit Model the link function is

$$g(\mu) = probit(\mu) = \Phi^{-1}(\mu)$$

and the Probit Model can be stated as

$$\pi(x) = \Phi(\alpha + \beta x)$$

or

$$probit(\pi(x)) = \Phi^{-1}\pi(x) = \alpha + \beta x$$

Again the following properties hold:

Properties:

- When $\beta > 0$, then $\pi(x)$ increases as x increases.
- When $\beta < 0$, then $\pi(x)$ decreases as x increases.
- When $\beta = 0$ then the curve is a horizontal line.
- The larger $|\beta|$, the steeper the increase (decrease).

The median effective level of x is also for the probit model estimated by

$$EL_{50} = -\frac{\alpha}{\beta}$$

Example 3

Using the probit link function for the HSB data gives:

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-4.252	.3932	-5.023	-3.482	116.965	1	.000
x	.017	.0015	.014	.019	121.093	1	.000
(Scale)	1[a]						

Dependent Variable: program
Model: (Intercept), x

a. Fixed at the displayed value.

The estimated model equation is:

$$probit(\hat{\pi}(x)) = -4.252 + 0.017x$$

Since $\beta > 0$, this means the higher the total score the higher the probability a person attends an academic high school program. The intercept means, that the probability for a person to have attended an academic program having a total score of 0 equals $\pi(0) = F(-4.252) \approx 0$.
The median effective level of x estimate is

$$EL_{50} = -\frac{\alpha}{\beta} = -(-4.252)/0.017 = 250.12,$$

which is slightly different then the estimate from the Logistic Model.

At this point we have seen three different models for modelling the probability of a binomial distribution. All three models can be applied to a data set, as seen in the example. To figure out which model is the most appropriate for a given situation model fit statistics should be compared. We will later see different approaches for comparing such models.

1.3 GLMs for count data

Many discrete response data have counts as outcomes.
For example

- the number of deaths by horse kicking in the Prussian army (first application of the model we are going to see in this in this section)
- the number of family members,
- the number of courses taken in university,
- number of leucocytes per ml,
- number of students in course after first midterm,
- number of patients treated on a certain day,
- number of iterations to get "close" to the solution.
- number of Covid-19 cases

In this section we will introduce a model for response variables which represent counts.
The easiest GLMs for count data use the Poisson distribution as the random component.

1.3.1 Poisson distribution

The probability distribution of a Poisson random variable Y representing the number of successes occurring in a given time interval or a specified region of space is given by the formula:

$$P(Y = y) = \frac{\exp(-\mu)\mu^y}{y!}$$

for non negative whole numbers y.
Where μ = mean number of successes in a given time interval or region of space.
For a Poisson distribution

$$E(Y) = \mu, \text{ and st.dev.} = \sigma_Y = \sqrt{\mu}$$

Example 4
A computer salesman sells on average 3 laptops per day. Use Poisson's distribution to calculate the probability that

1. on a given day he will sell some laptops
2. on a given day he will sell 2 or more laptops but less than 5 laptops
3. assuming that there are 8 working hours per days, in a given hour he will sell one laptop?

Solution:

1. $P(Y > 0) = 1 - P(Y = 0) = 1 - \frac{\exp(-3)3^0}{0!} = 1 - \exp(-3) = 0.95$
2. $P(2 \leq Y \leq 4) = P(Y = 2) + P(Y = 3) + P(Y = 4) = e^{-3}\left(\frac{3^2}{2!} + \frac{3^3}{3!} + \frac{3^4}{4!}\right) = 0.616$
3. The average number of laptops sold per hour is $\mu = 3/8$, therefore
   $P(Y = 1) = \frac{\exp(-3/8)(3/8)^1}{1!} = \exp(-3/8) \cdot 3/8 = 0.26$

1.3.2 Poisson Regression/Loglinear Model for modelling count data

- Random component: Poisson distribution,
- Systematic component: linear predictor.
- Link function: $g(\mu) = \ln(\mu)$ the natural logarithm (ln).

When x is numerical the model is usually called a Poisson Regression Model, but if x is categorical it is called a Loglinear Model.
For a single explanatory variable the Poisson regression/loglinear model has the form

$$\ln(\mu) = \alpha + \beta x$$

which is equivalent to

$$\mu = \exp(\alpha + \beta x) = \exp(\alpha) exp(\beta)^x = e^\alpha (e^\beta)^x$$

For one unit increase of x the mean changes by a factor of $\exp(\beta)$.
When $x = 0$ then $\mu = e^\alpha$, so e^α is the mean count, when $x = 0$.

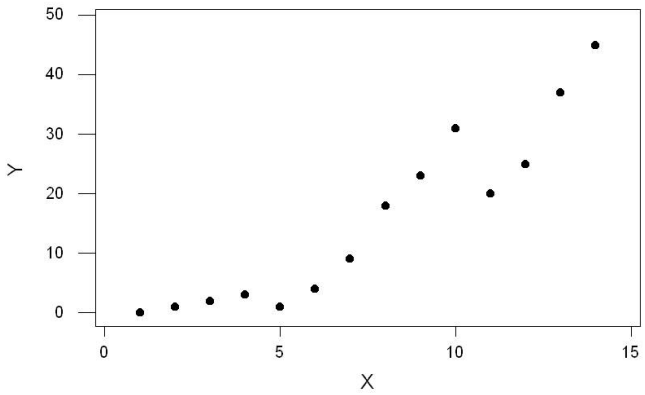
The link function is the natural choice for a Poisson distribution, since $\ln(\mu)$ is the natural parameter of a Poisson distribution.

Properties:

- If $\beta = 0$, then $\exp(\beta) = 1$ and the mean of Y is independent of x.
- If $\beta > 0$, then $\exp(\beta) > 1$ and the mean of Y increases when x increases.
- If $\beta \leq$, then $\exp(\beta) < 1$ and the mean of Y decreases when x increases.

Example 5
Classic Example: Whyte, et al 1987 (Dobson, 1990) reported the number of deaths due to AIDS in Australia per 3 month period from January 1983 - June 1986.
The data: x = time (measured in multiple of 3 month after January 1983)
Y = # of deaths in Australia due to AIDS

x_i	y_i	x_i	y_i
1	0	8	18
2	1	9	23
3	2	10	31
4	3	11	20
5	1	12	25
6	4	13	37
7	9	14	45



The graph shows a nonlinear relationship and a loglinear model seems to be a reasonable choice. Use SPSS for the analysis:

Parameter Estimates

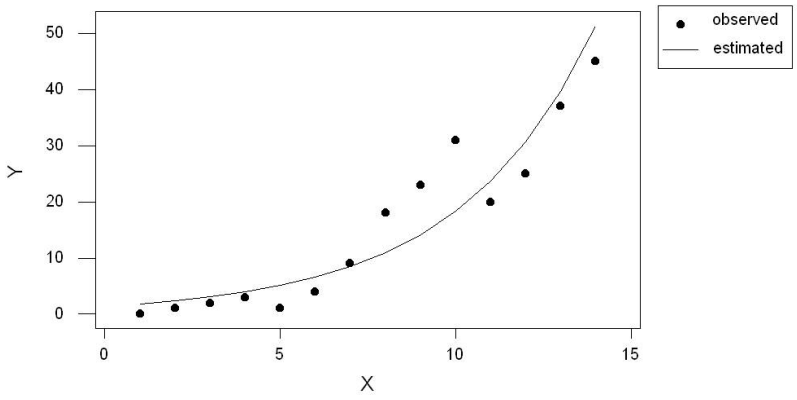
Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.340	.2512	-.153	.832	1.828	1	.176
X	.257	.0220	.213	.300	135.477	1	.000
(Scale)	1[a]						

Dependent Variable: Y
Model: (Intercept), X
a. Fixed at the displayed value.

The estimated equation is

$$\hat{\mu} = \exp(.340 + .257x)$$

In this period the number of deaths due to AIDS in a 3 month period year was on average $\exp(.257) = 1.29$ times higher than in the 3 month period before. According to the model the estimated number of death due to AIDS in the 3 month before the first observed numbers (time=x=0) is $e^{.340} = 1.404$, somewhere between 1 and 2.



The graph shows that in the beginning and the end of the time period the residuals are negative, but in the center period they are positive. This points to the fact that the model is not entirely appropriate. It can be improved by using $\ln(x)$ instead of x as an explanatory variable, but the interpretation of those results is less intuitive, though.

Also, for $x = 9$ and $x = 10$ the model gives not a very close fit.

1.4 Inference about Model Parameters in GLMs

For testing if the slope parameter, β, is different from zero, a Wald Z, Wald χ^2, or likelihood ratio test can be conducted. Remember if the slope is zero x drops out of the model, and the response becomes independent from x. So this test is a test if in the considered model the predictor has a notable effect on the response

1. $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$, choose α.
2. Assumptions: Random samples, samples have to be large for the ML- estimator to be approximately normal
3. Test statistic (three choices – usually only one is used):
   - Wald Z: $z_0 = \hat{\beta}/SE(\hat{\beta})$
   - Wald χ^2: $\chi_0^2 = (\hat{\beta}/SE(\hat{\beta}))^2$, $df = 1$
   - Likelihood ratio: $\chi_0^2 = -2\ln(l_0/l_1)$, $df = 1$
4. P-value:
   - Wald Z: $P = 2P(Z > |z_0|)$
   - Wald χ^2: $P = P(\chi^2 > \chi_0^2)$
   - Likelihood ratio: $P = P(\chi^2 > \chi_0^2)$
5. Conclusion

6. Context

Example 6
For the Aids in Australia example, test if time has an effect on the number of death due to Aids.

1. $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$, $\alpha = 0.05$.
2. Assumptions: Random samples, sample size is ok
3. Test statistic: Wald χ^2: $\chi_0^2 = 135.477$, $df = 1$ (from the output)
4. P-value: Wald χ^2: $P = P(\chi^2 > 135) < 0.001$.
5. Conclusion: Reject H_0.
6. Context: At significance level of 5% the data provide sufficient evidence that time has an effect on the number of death due to Aids in Australia.

1.5 Model Fit and Comparison

"Essentially, all models are wrong, but some are useful."
— Box, George E. P.; Norman R. Draper (1987).
Empirical Model-Building and Response Surfaces, p. 424, Wiley. ISBN 0471810339.

Before a model can be considered to give meaningful insight into the topic under investigation, it has to be assessed for model fit and its usefulness. Different tests and measures help with the assessment.

1.5.1 Model Utility Test

To test model fit or usefulness of a model with H_0:none of the predictors affects the response, a likelihood ratio test with the following test statistic should be conducted:

$$\chi^2 = -2\ln(l_0/l_1) = -2(\ln(l_0) - \ln(l_1)) = -2(L_0 - L_1)$$

where $L_0 = \ln(l_0)$ and $L_1 = \ln(l_1)$ are log-likelihoods:

- L_0 is the log-likelihood of the data based on the intercept-only model – model 0
- L_1 is the log-likelihood of the data based on the model including the proposed predictors – model 1

By comparing the two log-likelihood values the model fit of the two models are compared.
If the log-likelihood of model 0 is much smaller than the log-likelihood of model 1, the test statistic will be large, and lead to a rejection of H_0, indicating that the inclusion of the proposed predictors in the model leads to significantly better model fit.

The idea of comparing two models by comparing their log-likelihoods will later be generalized to allow the comparison of nested models.

Goodness of fit test based on the likelihood ratio statistic

1. Hyp.: H_0 : The proposed model fits as well as the intercept-only model H_a: H_0 is not correct. Choose α.

2. Assumptions: Random sample, large sample size

3. Test statistic: $\chi_0^2 = -2(L_0 - L_1)$, df=number of predictors

4. P-value=$P(\chi^2 > \chi_0^2)$

1.5.2 Deviance

The Deviance of a model is based on the difference between the log-likelihood of the model of interest, L_M, and the log-likelihood of the most complex model that perfectly fits the data (one parameter per "measurement", saturated model), L_S.

$$Deviance = -2(L_M - L_S)$$

The deviance is the likelihood-ratio statistic for testing if the saturated model gives a significant better fit than the model of interest. (This is equivalent to testing if all parameters in the saturated model, which are not in the model of interest are 0).

In this case the deviance is χ^2 distributed, the deviance therefore provides a test statistic for model fit. The df of the test statistic is the difference in the number of parameters in the saturated model and the model of interest.
If H_0 can be rejected this means that the model M provides a significant worse fit than the saturated model. Therefore we hope that this test is not significant. This would not mean that the model fits as well as the saturated model, but at least that there is no significant difference at this time.

Goodness of fit test based on the deviance

1. Hyp.: H_0 : The proposed model fits as well as the saturated model H_a: H_0 is not correct. Choose α.

2. Assumptions: Random sample, large sample size

3. Test statistic: $Deviance_0 = -2(L_M - L_S)$, df=difference in the number of parameters

4. P-value=$P(\chi^2 > Deviance_0)$

Example 7

Goodness of Fit[b]

	Value	df	Value/df
Deviance	29.654	12	2.471
Scaled Deviance	29.654	12	
Pearson Chi-Square	28.847	12	2.404
Scaled Pearson Chi-Square	28.847	12	
Log Likelihood[a]	-41.290		
Akaike's Information Criterion (AIC)	86.581		
Finite Sample Corrected AIC (AICC)	87.672		
Bayesian Information Criterion (BIC)	87.859		
Consistent AIC (CAIC)	89.859		

Dependent Variable: Y
Model: (Intercept), x

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

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

According to the SPSS output the Deviance for the loglinear model for the number of deaths due to Aids in Australia equals Deviance = 29.654, df=12(=n-2=n-(number of parameters in the model)). It is hard to judge this value, without knowing the distribution of the deviance. A better measure is Deviance/df=2.471, measures "close" to one indicate good model fit. Here the score is not close to one and can be interpreted as lack in model fit.
The P-value for a test of H_0, this model fits as well as the saturated model, equals
$P(\chi^2 > \chi_0^2) < 0.005$ (with $\chi_0^2 = 29.654, df = 12$). We would therefore reject H_0, and find that the saturated model fits significantly better than the proposed model.

Word of caution: When sample sizes are large this test is usually significant, rejecting the proposed model. In practise one most often uses the rule of thumbthat a model is found acceptable if Deviance/df<2.5

1.5.3 Comparing Models

Consider two models M_0 and M_1, such that M_0 is nested within M_1. This means M_1 has all the parameters as M_0 and more.
For example the linear regression model

$$M_0 : y = \alpha + \beta_1 x_1 + e$$

is nested within the linear regression model

$$M_1 : y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$$

We get M_0 from M_1 by setting $\beta_2 = 0$. M_0 is the special case of M_1 when $\beta_2 = 0$.
Then the likelihood ratio statistic for the null hypothesis that the fit of M_0 is as good as the fit of M_1 is given by

$$-2(L_0 - L_1) = Deviance_0 - Deviance_1$$

The test statistic would be large if M_0 fits poorly in comparison with M_1. It is approximately χ^2 distributed with df=difference in the number of parameters in the two models.

Example 8
For comparing the model including time with the model only including the intercept. For finding the test statistic we need the deviance for the model including the intercept only.
The SPSS output is:

Goodness of Fit[b]

	Value	df	Value/df
Deviance	207.272	13	15.944
Scaled Deviance	207.272	13	
Pearson Chi-Square	186.616	13	14.355
Scaled Pearson Chi-Square	186.616	13	
Log Likelihood[a]	-130.100		
Akaike's Information Criterion (AIC)	262.199		
Finite Sample Corrected AIC (AICC)	262.533		
Bayesian Information Criterion (BIC)	262.839		
Consistent AIC (CAIC)	263.839		

Dependent Variable: Y
Model: (Intercept)

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

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

1. H_0: The intercept only model fits as well as the model including time versus H_a: H_0 is not true
2. Random sample
3. $$Deviance_0 - Deviance_1 = 207.272 - 29.654 = 177.618, \quad df = 2 - 1$$
4. $P - value < 0.001$
5. Reject H_0
6. At significance level of 5% the data provide sufficient evidence that the model including the time fits better than the intercept only model.

1.5.4 Residuals for Comparing Observations to the fitted model

- Study residuals (differences between observed and estimated values to assess model fit).
- Like in the study of contingency tables it is better to use standardized residuals.

Pearson's residuals (to be calculated for each observation in the sample:

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\text{Var}}(y_i)}}$$

With grouped data the Pearson residuals are approximately normally distributed, but this is not the case with individual data. In both cases, however, observations with a Pearson residual exceeding two in absolute value may be worth a closer look.

Deviance residuals: alternative residuals are deviance residuals, these are measures how much each individual measurement adds to the deviance of a model, deviance residuals greater than 2 are indicative of an ill fitting model for this measurement.

Standardized Pearson's residuals are defined as

$$r_i = \frac{y_i - \hat{\mu}_i}{SE}$$

(we will see later how to find SE) These residuals have an approximate normal distribution and standardized residuals larger than 2 or 3 should be investigated.

Example 9

	x	Y	est	PearsonResidual	DevianceResidual	StdPearsonResidual
1	1.00	.00	1.82	-1.347	-1.905	-1.417
2	2.00	1.00	2.35	-.879	-.993	-.928
3	3.00	2.00	3.04	-.593	-.632	-.627
4	4.00	3.00	3.93	-.464	-.484	-.492
5	5.00	1.00	5.08	-1.806	-2.210	-1.916
6	6.00	4.00	6.57	-.995	-1.073	-1.054
7	7.00	9.00	8.49	.186	.184	.196
8	8.00	18.00	10.98	2.137	1.953	2.249
9	9.00	23.00	14.20	2.359	2.161	2.475
10	10.00	31.00	18.36	2.980	2.708	3.128
11	11.00	20.00	23.74	-.742	-.762	-.785
12	12.00	25.00	30.69	-.997	-1.029	-1.084
13	13.00	37.00	39.69	-.386	-.391	-.449
14	14.00	45.00	51.32	-.834	-.851	-1.134

The residuals for $x = 9$ and $x = 10$ are quite large, and indicate that at that time another factor influenced the increase in Aids in Australia.