

9 Generalized Linear Models

The Generalized Linear Model (GLM) is a model which has been built to include a wide range of different models you already know, e.g. ANOVA and multiple linear regression models are just special cases of this model. But it also includes models permitting the response to be categorical, and/or the error not being normal.

9.1 Components of a GLM

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

9.1.1 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.
- We will study GLMs for binary outcomes (success/failure) and outcomes resulting from counting. In such cases it is not meaningful to assume a normal random component, but for binary outcomes Y has a binomial distribution and for counts the Poisson distribution is often a good choice for the random component.
- In general the different observations are assumed to be independent and are denoted by $Y_1, \dots, Y_n$.

9.1.2 Systematic component

The systematic component specifies the explanatory variables. Since this is a linear model the systematic component is linear in the coefficients $\beta_0, \dots, \beta_k$.

The linear predictor is:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

The regressors can be interaction terms or transformations of the predictor variables as we have seen in MLRMs. In multiple linear regression this is the right hand side of the model equation (omitting the error).

9.1.3 Link function

The link function links the random component with the systematic component. More specific: let μ be the mean of Y , then the link function g relates the mean μ with the linear predictor.

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

- For MLRMs $g(\mu) = \mu$ (identity)

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k.$$

- For count data the logarithm $g(\mu) = \ln(\mu)$ is used

$$\ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

The model is called a loglinear model.

- For binary data the logit link function $g(\mu) = \ln(\mu/(1 - \mu))$ is often used.

Since for binary variables $\mu = \pi$, the success probability, the model relates the log odds with the predictor variables.

$$\log(\mu/(1 - \mu)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

The resulting GLM is then called logistic regression model.

The GLM is a "Super Model" which unifies many different models under one umbrella. Maximum Likelihood estimators are used instead of least squares estimators.

9.2 Maximum Likelihood

A Maximum-Likelihood Estimator (MLE) for a parameter is chosen, such that the chance of the data occurring is maximal if the true value of the parameter is equal to the value of the Maximum Likelihood Estimator.

The likelihood function L

$$L : \Theta \times \mathbb{R}^n \rightarrow [0, 1]$$

assigns to parameter value $\theta \in \Theta$ and sample data $\vec{y} \in \mathbb{R}^n$ the likelihood to observe the data $\vec{y}$, if θ is the true parameter value describing the population.

Let $f(\vec{y}|\theta)$ be the density function for random variable $\vec{Y}$ (representing the random sample) with parameter θ , then the Likelihood function is:

$$L(\theta|\vec{y}) = f(\vec{y}|\theta)$$

The MLE for θ based on data $\vec{y}$ is the value $\tilde{\theta}$ which maximizes the likelihood function for the given $\vec{y}$.

In most cases it will be easier to maximize the function of $l := \ln(L)$. This is valid because the $\ln$ function is monotonic increasing.

9.3 GLM for binary data

We will assume in this section a binary response variable (success/failure), Y . Then Y has a binomial distribution with probability for success of π , and the probability of failure is $1 - \pi$.

Usually success is coded by a 1 and failure by 0. Single observations can be viewed as binomial distributions with $n = 1$, also called a Bernoulli distribution.

Because Y has a binomial distribution the random component for binary data is binomial.

The GLMs for binary data model the success probability, π in dependency on explanatory variables. The systematic component is the linear predictor in one variable.

$$\beta_0 + \sum_{i=1}^k \beta_i x_i$$

In order to express that the probability of success depends on explanatory variables $x_1, \dots, x_k$ write $\pi(x_1, \dots, x_k)$.

In the following sections link functions logit, and probit are presented for binary response.

9.3.1 Logit

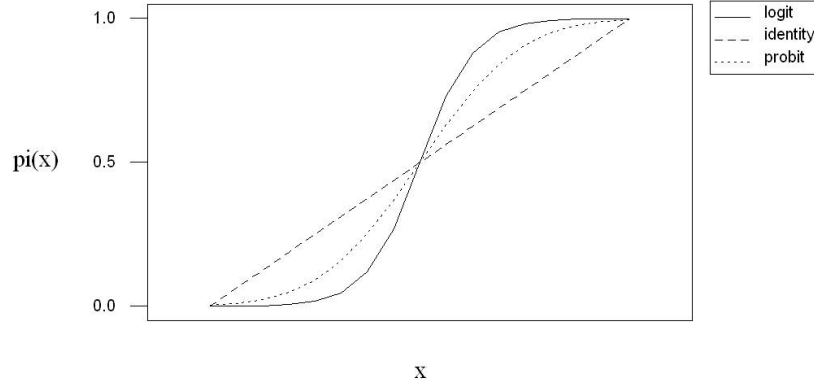
The model based on the logit link function combined with the linear predictor results in the following model for the logit of $\pi(x_1, \dots, x_k)$:

$$\ln(\pi(x_1, \dots, x_k)/(1 - \pi(x_1, \dots, x_k))) = \beta_0 + \sum_{i=1}^k \beta_i x_i$$

this is equivalent to:

$$\pi(x) = \frac{\exp(\beta_0 + \sum_{i=1}^k \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^k \beta_i x_i)}$$

This model is called the **logistic regression model**.



The relationship between a predictor x_i and $\pi(x_i)$ is S-shaped. This choice of link function enforces:

$$0 \leq \exp(\beta_0 + \sum_{i=1}^k \beta_i x_i)/(1 + \exp(\beta_0 + \sum_{i=1}^k \beta_i x_i)) \leq 1,$$

or $0 \leq \pi(x_1, \dots, x_k) \leq 1$ for all values of $x_1, \dots, x_k$.

Properties:

- When $\beta_i > 0$, then $\pi(x_1, \dots, x_k)$ increases as x_i increases.
- When $\beta_i < 0$, then $\pi(x_1, \dots, x_k)$ decreases as x_i increases.
- When $\beta_i = 0$ then the curve is a horizontal line.
- The larger $|\beta_i|$, the steeper the increase (decrease).

The MLEs can not be given in closed form and have to be found using an iterative method, like we have seen in nonlinear regression.

Example 9.1.

To illustrate use data from Koch & Edwards (1988) from a double-blind clinical trial investigating a new treatment for rheumatoid arthritis. (included in the "vcd" package – data(Arthritis))

The data include information on 84 patients; if they were treated (yes/no), their sex(Female, Male) and age, and if they have improved(None, Some, Marked).

In order to make the outcome binary we will distinguish between patients who got better (Improved = Some or Improved = Marked), who didn't.

The main question is to find out if the treatment helped when correcting for age and sex.

Let π be the probability for getting better, then the model is

$$\text{logit}(\pi) = \beta_0 + \beta_1 \text{treat}_y + \beta_2 \text{age} + \beta_3 \text{sex}_m$$

The R code and partial output:

```
library(vcd)
Arthritis$better <- 1-(Arthritis$Improved=="None")
attach(Arthritis)
model.glm <- glm(better~Sex+Age+Treatment,family=binomial("logit"))
anova(model.glm)
summary(model.glm)

Analysis of Deviance Table
Model: binomial, link: logit
Response: better

Terms added sequentially (first to last)
```

		Df	Deviance	Resid. Df	Resid. Dev
NULL				83	116.449
Sex	1	4.6921		82	111.757
Age	1	7.4973		81	104.259
Treatment	1	12.1965		80	92.063

```
> summary(model.glm)

Call: glm(formula = better ~ Sex + Age + Treatment, family = binomial("logit"))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.01546    1.16777  -2.582  0.00982
SexMale       -1.48783    0.59477  -2.502  0.01237
Age           0.04875    0.02066   2.359  0.01832
TreatmentTreated 1.75980    0.53650   3.280  0.00104
---
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 116.449  on 83  degrees of freedom
Residual deviance: 92.063  on 80  degrees of freedom
AIC: 100.06

Number of Fisher Scoring iterations: 4
```

The estimated model equation has been found after 4 iterations:

$$\text{logit}(\pi) = -3.02 + 1.76\text{treat}_y + 0.05\text{age} - 1.49\text{sex}_m$$

Since $\beta_1 > 0$ we conclude that the treatment had on positive effect on the chance of getting better, and since $\beta_3 > 0$ we find that older patients had a better chance of getting better in the study and since $\beta_3 < 0$ we find that women had a greater chnce of improving in this study.

The intercept would apply to patients who did not get treated, were of age 0 and female. The chance of getting better for such patients would be estimated to be $\exp(-3.02)/(1 + \exp(-3.02)) = 0.047$. Not of interest here.

9.3.2 Probit

Another S-shaped curve used to link the probability for success with the linear predictor is the cumulative distribution function of a normal distribution, Φ .

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{x^2}{2}} dx$$

The normal cumulative distribution function shows the S-shape shown in the diagram above.

The link function is the Inverse Cumulative Distribution Function of the standard normal distribution, Φ^{-1} , called the probit function, which assigns to numbers between 0 and 1 the percentile of the standard normal distribution.

The same properties hold as for the logit function.

The probit model is given by

$$probit(\pi(x)) = \beta_0 + \sum_{i=1}^k \beta_i x_i$$

Therefore

$$\pi(x_1, \dots, x_k) = \Phi(\beta_0 + \sum_{i=1}^k \beta_i x_i)$$

Example 9.2.

Using the probit link function for the Arthritis data, the model is

$$probit(\pi) = \beta_0 + \beta_1 treat_y + \beta_2 age + \beta_3 sex_m$$

R code and output

```
model.glm.p <- glm(better~Sex+Age+Treatment,family=binomial("probit"))
anova(model.glm.p)
summary(model.glm.p)
```

Analysis of Deviance Table
Model: binomial, link: probit
Response: better

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev
NULL			83	116.449
Sex	1	4.6921	82	111.757
Age	1	7.1730	81	104.584
Treatment	1	12.6550	80	91.929

```
> summary(model.glm.p)
```

Call: glm(formula = better ~ Sex + Age + Treatment, family = binomial("probit"))

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.79023	0.67852	-2.638	0.008329
SexMale	-0.90071	0.34390	-2.619	0.008816
Age	0.02874	0.01207	2.381	0.017265
TreatmentTreated	1.07894	0.31130	3.466	0.000528

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 116.449 on 83 degrees of freedom
Residual deviance: 91.929 on 80 degrees of freedom
AIC: 99.929

Number of Fisher Scoring iterations: 4

Again after 4 iteration the estimation process halted. The estimated model equation is:

$$\text{logit}(\pi) = -1.79 + 1.08\text{treat}_y + 0.03\text{age} - 0.90\text{sex}_m$$

The interpretation of the signs of the coefficient is consistent with the logistic regression model.

To answer if the probit or logit model is the better choice for this data, model fit measures have to be compared.

9.4 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)
- 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.

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

9.4.1 Poisson distribution

The density 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 random variable Y with Poisson distribution with parameter μ :

$$E(Y) = \mu, \text{ and } Var(Y) = \mu$$

Example 9.3.

A computer salesman sells on average 3 laptops per day. Use Poisson's distribution to calculate the probability that on a given day he will sell

1. some laptops on a day,
2. 2 or more laptops but less than 5 laptops.
3. Assuming that there are 8 working hours per days, what is the probability that 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
 $P(Y = 1) = \frac{\exp(-3/8)(3/8)^1}{1!} = \exp(-3/8) \cdot 3/8 = 0.26$

9.4.2 Poisson Regression/Loglinear Model

- Random component is the Poisson distribution.
- The systematic component is as usual the linear predictor.

When the predictor is numerical the model is usually called a Poisson Regression Model, but if it is categorical it is called a Loglinear Model, or in general the Poisson loglinear model.

- The link function is the natural logarithm (ln).

The Poisson regression/loglinear model has the form

$$\ln(\mu) = \beta_0 + \sum_{i=1}^k \beta_i x_i$$

which is equivalent to

$$\mu = \exp(\beta_0 + \sum_{i=1}^k \beta_i x_i) = \exp(\beta_0) \exp(\beta_1)^{x_1} \exp(\beta_2)^{x_2} \cdots \exp(\beta_k)^{x_k}$$

For one unit increase in x_i the mean changes by a factor of $\exp(\beta_i)$.

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

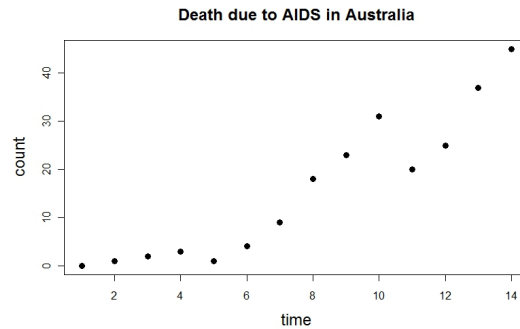
Example 9.4.

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. To fit the Poisson Regression model use

```
time <- 1:14
count <- c(0,1,2,3,1,4,9,18,23,31,20,25,37,45)

AIDS <- data.frame(time,count)

glm.Aids <- glm(count~time, family=poisson(), data=AIDS)
anova(glm.Aids)
summary(glm.Aids)

Part of the output

> anova(glm.Aids)
Analysis of Deviance Table
Model: poisson, link: log
Response: count

Terms added sequentially (first to last)

      Df Deviance Resid. Df Resid. Dev
NULL                                13    207.272
time  1   177.62          12     29.654
> summary(glm.Aids)
Call:
glm(formula = count ~ time, family = poisson())

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.33963   0.25119   1.352   0.176
time         0.25652   0.02204  11.639 <2e-16
---
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 207.272  on 13  degrees of freedom
Residual deviance:  29.654  on 12  degrees of freedom
AIC: 86.581

Number of Fisher Scoring iterations: 5
```

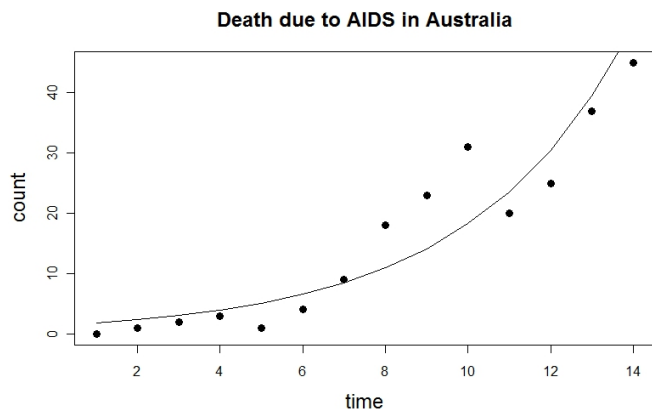
The coefficient table shows the parameter estimates, their standard errors and test statistic for testing if the coefficient is zero.

The estimated equation is

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

In this period the number of deaths due to AIDS in a year was in average $\exp(.257) = 1.29$ times higher than in the year before.

The deviance is a measure of fit we will look at shortly.



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 $\ln(\text{time})$ instead of time as an explanatory variable, but the interpretation of those results is less intuitive.

Also, for times 9 and 10 the model does not give a good fit.

9.5 Inference about Model Parameters

For testing if the model parameters (β_i $0 \leq i \leq k$) are different from zero, a Wald Z or likelihood ratio test can be done.

This

1. $H_0 : \beta_i = 0$ versus $H_a : \beta_i \neq 0$, choose α .
2. Assumptions: Random samples, samples have to be large for the ML- estimator to be approximately normal
3. Test statistic:
 - Wald Z : $z_0 = \hat{\beta}/SE(\hat{\beta})$
 - Likelihood ratio: $\chi_0^2 = -2\ln(l_0/l_1)$, $df = 1$, where l_0 is the log likelihood value for the model not including the parameter, and l_1 is the log likelihood value for the model including the function.
4. P-value:
 - Wald Z : $P = 2P(Z > |z_0|)$
 - Likelihood ratio: $P = P(\chi^2 > \chi_0^2)$
5. Conclusion
6. Context

Example 9.5.

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

1. $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, $\alpha = 0.05$.
2. Assumptions: Random samples, sample size is ok

3. Test statistic: Wald Z : $Z_0 = 11.639$ (from the output)
4. P-value: $P = P(Z_0 > 11.639) < 0.001$.
5. Conclusion: Reject H_0 .
6. Context: At significance level of 5% the data provide sufficient evidence that the time has an effect on the number of death due to Aids in Australia.

9.6 Model Fit and Comparison

As mentioned above the likelihood ratio test statistic is:

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

where $L_0 = \ln(l_0)$ and $L_1 = \ln(l_1)$ are the log-likelihoods for the model not including x_i ($\beta_i = 0$) and the model including x_i ($\beta_i \neq 0$), respectively.

It is based on the maximum log-likelihood values for two different models, and allows us in that way to compare the two models.

This idea is generalized to obtain measures of model fit and comparing models in general.

9.6.1 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 which 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).

The deviance is χ^2 distributed and therefore provides a test statistic of 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 proposed 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 9.6.

For the Aids in Australia example:

In the R-output the deviance is stated as the residual deviance = 29.654, df=12. The P-value for the test would be less than 0.001, and therefore indicate that the saturated model fits the data significantly better than the model only including time, at $\alpha = 0.05$

This residual deviance is indicating lack of model fit for the model only including time to predict death due to Aids in Australia.

To judge model fit without formally conducting the test the good measure is Deviance/df (here=2.471). If close to one it indicates good model fit. Here the score is not close enough and can be interpreted as lack in model fit.

Example 9.7.

For the Arthritis data:

The residual deviance for the probit model is lower than for the logistic model, indicating better model fit for the probit model.

The residual deviance is 91.9, df=80, resulting in a p-value=0.17, at significance level of 5% the data do not indicate that the saturated model has a significant better model fit, supporting the chosen model.

deviance/df=1.15 is close to 1 and therefore support the finding of proper model fit.