6 Count response

6.1 Representations of count response data

In the count data, the possible values the response variable Y_i are integers $0, 1, 2, \ldots$. In practical situations there is sometimes an upper limit for Y_i but this can be ignored if $\mathrm{E}(Y_i)$ is far below the the upper limit. It is often assumed that the counted events arise from a Poisson process whose intensity depends on the covariates.

When the number of events for each individual are observed, the data can be presented in the form of data matrix

x_1	x_2	y
250	A	7
250	A	1
350	B	1
300	A	0
250	B	8
300	B	3
⋮	⋮	⋮

or weighted data matrix

x_1	x_2	y	frequency
250	A	0	4
250	A	1	2
250	A	3	1
250	B	0	4
250	B	1	5
250	B	2	3
250	B	3	2
300	A	0	6
⋮	⋮	⋮	⋮

When the number of events are observed only for each covariate class, the data can be presented in the form of frequency table (crosstabulation)

	$x_2 = A$	$x_2 = B$
$x_1 = 250$	23	12
$x_1 = 300$	21	19
$x_1 = 350$	7	13

6.2 Link functions for count data

Count data can be modeled by a GLM with the logarithmic link function

$$\log(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}. \tag{77}$$

These models are also called Poisson regression or log-linear models.

6.3 Likelihood

The log-likelihood Poisson distributed response has the form

$$l(\mathbf{y};\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log(\mu_i) - \mu_i - \log(y_i!) = \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \beta_j x_{ij} y_i - \exp\left(\sum_{j=1}^{p} \beta_j x_{ij} \right) - \log(y_i!) \right). \tag{78}$$

Example: Full likelihood for missing data Count response y_i, continuous covariate x_{i1} and binary covariate x_{i2} are recorded for the sample $i = 1, 2, \ldots, 1000$. y_i and x_{i1} are observed for all i but x_{i2} is missing for $i = 1, 2, \ldots, 250$. We can assume that the values x_{i2} are missing at random (MAR), i.e. the missingness does not depend on the unobserved value itself but it may depend on the observed response or on the other covariate. The full likelihood for the data can be written

$$\begin{aligned} &L(\mathbf{y};\beta_1,\beta_2) = \\ &\prod_{i=1}^{250} \left(p(X_{i2}=1) p(x_{i1}|X_{i2}=1) \frac{\exp(\beta_1 x_{i1} + \beta_2)^{y_i} \exp(-\exp(\beta_1 x_{i1} + \beta_2))}{y_i!} + \right. \\ &\left. p(X_{i2}=0) p(x_{i1}|X_{i2}=0) \frac{\exp(\beta_1 x_{i1})^{y_i} \exp(-\exp(\beta_1 x_{i1}))}{y_i!} \right) \\ &\times \prod_{i=251}^{1000} p(x_{i2}) p(x_{i1}|x_{i2}) \frac{\exp(\beta_1 x_{i1} + \beta_2 x_{i2})^{y_i} \exp(-\exp(\beta_1 x_{i1} + \beta_2 x_{i2}))}{y_i!}. \end{aligned} \tag{79}$$

In order to calculate this likelihood we need to specify the marginal distribution $p(x_2)$ and the conditional distribution $p(x_1|x_2)$. This may require specification of some additional parameters that are considered as nuisance parameters and are estimated together with the parameters of interest β_1 and β_2.

6.4 Offset

The intensity of a process is defined in the form events/time and the observed number of events depends on the system has been followed up. In the GLM, the time is taken into account adding an offset term

$$\log(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij} + \log(t_i), \tag{80}$$

where t_i is the offset, i.e. the follow-up time or the time under exposure. The regression coefficient of the offset is fixed to 1. In R, this can done with the model term `offset`.

6.5 Overdispersion

For the Poisson distribution the variance is equal to expected value. In real-world data, it is common that the observed variance is higher, i.e. there is overdispersion.

6.6 Example: Follow-up for cardiovascular diseases

Cohort studies are important in epidemiological research. A population cohort represents the population of specified age range living in a certain geographical area. The FINRISK cohort 1982 consists of men and women who were 25-64 years old at baseline year 1982 (the beginning of the follow-up). The cohort has been followed up for deaths and fatal and non-fatal events of cardiovascular diseases until the end of year 2006. The example data set contains the number of recorded events grouped by year, age group, sex and area (Eastern Finland / Western Finland). The number of events should be considered in proportion to the person years of follow-up.

7 Nominal and ordinal response

7.1 Representations of nominal response data

Data matrix

x	y
250	A
250	C
350	C
300	A
250	B
300	B
⋮	⋮

Weighted data matrix

x	y	frequency
250	A	23
250	B	12
250	C	4
300	A	21
300	B	19
300	C	7
350	A	7
350	B	13
350	C	3

Frequency table (crosstabulation)

	$Y =' A'$	$Y =' B'$	$Y =' C'$
$x = 250$	23	12	4
$x = 300$	21	19	7
$x = 350$	7	13	3

In nominal response data, the response is one of the categories. In ordinal response data, the categories have a natural order. Binomial response is a special case of both nominal and ordinal response.

When there are more than two categories, nominal and ordinal response have a multivariate nature. For the ith covariate class m_i responses are recorded and the number of responses in the categories $1, 2, \ldots, Q$ is denoted

by a vector $\left(K_{i1}K_{i2}\ldots K_{iQ}\right)$. The response is then defined as vector

$$\mathbf{Y}_i = \left(\frac{K_{i1}}{m_i} \quad \frac{K_{i2}}{m_i} \ldots \frac{K_{iQ}}{m_i}\right). \tag{81}$$

7.2 Multinomial distribution

Multinomial distribution is a generalization of binomial distribution in the case when there are more than two categories

$$f_i(k_{i1}, k_{i2}, \ldots, k_{iQ}; m_i, \pi_{i1}, \pi_{i2}, \ldots, \pi_{iQ}) = \frac{m_i!}{k_{i1}!k_{i2}!\cdots k_{iQ}!}\pi_{i1}^{k_{i1}}\pi_{i2}^{k_{i2}}\cdots\pi_{iQ}^{k_{iQ}}, \tag{82}$$

where π_{iq}'s are category probabilities for which $\sum_{q=1}^{Q}\pi_{iq} = 1$.

Multinomial distribution does not belong to the exponential family defined in equation (18) but it can derived from the Poisson distribution. Let $K_1, K_2, \ldots, K_Q$ be independent Poisson distributed random variables with means $\lambda_1, \lambda_2, \ldots, \lambda_Q$. The sum $m = K_1 + K_2 + \ldots + K_Q$ is a random variable that follows Poisson distribution with the mean $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_Q$. Then the conditional distribution

$$\begin{aligned} f(k_1, k_2, \ldots, k_Q; m) =& \frac{\prod_{q=1}^{Q}\lambda_q^{k_q}e^{-\lambda_q}}{k_q} \Bigg/ \frac{\lambda^m e^{-\lambda}}{m} = \\ & \left(\frac{\lambda_1}{\lambda}\right)^{k_1}\left(\frac{\lambda_2}{\lambda}\right)^{k_2}\cdots\left(\frac{\lambda_Q}{\lambda}\right)^{k_Q}\frac{m!}{k_1!k_2!\cdots k_Q!} \end{aligned} \tag{83}$$

has the form of the multinomial distribution.

7.3 Regression models for nominal and ordinal response

There are alternative ways to parameterize regression models for nominal and ordinal response. For nominal response data, one of the categories is typically chosen as the reference category and the model has form

$$\log\left(\frac{\pi_{iq}}{\pi_{i1}}\right) = \beta_{0q} + \sum_{j=1}^{p}\beta_{jq}x_{ij}, \tag{84}$$

where the first category serves as reference. For ordinal response data, the model can be written using cumulative probabilities $\gamma_{iq} = \pi_{i1} + \pi_{i2} + \ldots + \pi_{iq}$

$$\log\left(\frac{\gamma_{iq}}{1-\gamma_{iq}}\right) = \beta_{0q} + \sum_{j=1}^{p}\beta_{jq}x_{ij}. \tag{85}$$

7.4 Proportional odds model

In proportional odds model it is assumed that the effect of the covariates is the same between all response categories on the logarithmic scale. Model (84)is then written as

$$\log\left(\frac{\pi_{iq}}{\pi_{i1}}\right) = \beta_{0q} + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{86}$$

and model (85) as

$$\log\left(\frac{\gamma_{iq}}{1-\gamma_{iq}}\right) = \beta_{0q} + \sum_{j=1}^{p} \beta_j x_{ij}. \tag{87}$$

7.5 Latent variable interpretation for ordinal regression

In some situations, the actual variable of interest is a continuous response that is difficult or impossible to measure. Instead, an ordinal response variable is measured. It can be assumed that there are unobserved cutpoints that divide the continuous response into categories.

7.6 Nominal and ordinal response data in R

Because of the multivariate nature of the response, function `glm` cannot be directly applied to nominal or ordinal response data in the general case. When the data can be presented in the form a frequency table, log-linear models can be fitted using `glm(...,family=poisson(link=log))`. Function `multinom` from the package `nnet` can be used to fit multinomial log-linear models via neural networks. Function `lopr` from the package `MASS` can be used to fit proportional odds models with logit, probit or cloglog links. Function `vglm` from the package `VGAM` fits a large variety of vector GLMs including multinomial logit models and proportional odds models.

8 Positive response

8.1 Characteristics of positive response data

The response variable is continuous and obtains only positive values. Non-negative response may also obtain value 0. Typically, the distribution of the response is skewed. We may identify three situations:

1. All observed values are positive and "far" from zero.
2. All observed values are positive and some values are relatively close to zero.
3. All observed values are non-negative and a number of them are exactly zero.

The first situation is the easiest to handle whereas the third situation often requires two models, one for the probability of zero response and one for the positive response.

Models for positive response data often assume that the coefficient of variation (CV), the ratio of the the standard deviation to the expectation,

$$CV = \frac{\sqrt{\mathrm{Var}(Y)}}{\mathrm{E}(Y)}. \tag{88}$$

is constant. This is equivalent to assuming that the variance is proportional to square of the expectation

$$\mathrm{Var}(Y_i) \propto \mu_i^2. \tag{89}$$

8.2 Gamma distribution

The gamma distribution is a distribution with a constant coefficient of variation. The gamma distribution has the density function

$$f(y; \lambda, \nu) = \frac{1}{\lambda^\nu \Gamma(\nu)} y^{\nu-1} e^{-y/\lambda}, \quad y > 0, \tag{90}$$

where $\nu > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter. Alternative parameterizations exist. Exponential distribution is a special case of the gamma distribution with the shape parameter $\nu = 1$. The

gamma distribution belongs to the exponential family (18) with $\theta_i = 1/(\nu\lambda_i)$, $b(\theta) = \log(-\theta)$ and $\phi = 1/\nu$ and has the mean $\mathrm{E}(Y_i) = \nu\lambda_i$ and variance $\mathrm{Var}(Y_i) = \nu\lambda_i^2$. Usually the dispersion parameter ϕ is not known and need to be estimated.

8.3 Link functions for gamma distributed response

The canonical link for the gamma distribution is the inverse link (reciprocal)

$$g(\mu_i) = \frac{1}{\mu_i}. \tag{91}$$

Because $g(\mu_i)$ is always positive, the regression parameters $\boldsymbol{\beta}$ need to be restricted so that the linear predictor is positive. Identity link can be also for the gamma distribution. Restrictions for the regression parameters $\boldsymbol{\beta}$ are needed also the identity link.

Log-link is often used for gamma distributed response. The use of log-link implies multiplicative effect of the covariates. Restrictions for the regression parameters are not needed.

8.4 Lognormal distribution

An alternative to the GLM with gamma distributed response is to take the logarithm of the response and assume that $\log(Y)$ follows normal distribution. Then Y follows log-normal distribution

$$f(y;\mu,\sigma) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right), \quad y > 0, \tag{92}$$

where μ is the mean of $\log(Y)$ and σ^2 is the variance of $\log(Y)$.

8.5 Inverse Gaussian distribution

The inverse Gaussian distribution is another standard distribution for positive responses in GLM. The pdf has form

$$f(y;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left(\frac{-\lambda(y-\mu)^2}{2\mu^2 y}\right), \quad y > 0, \tag{93}$$

where $\mu > 0$ is location parameter and $\lambda > 0$ is the shape parameter. The inverse Gaussian distribution belongs to the exponential family (18) and has mean μ and variance μ^3/λ. When λ tends to infinity the pdf of the inverse Gaussian distribution approaches the pdf of the normal distribution.

Link functions used with the inverse Gaussian distribution include identity, log, inverse and inverse squared link,

$$g(\mu_i) = \frac{1}{\mu_i^2}, \tag{94}$$

which is the canonical link.

8.6 Compound Poisson model

In the compound Poisson model, the response is a sum of K independent identically distributed random variables and K follows Poisson distribution. The compound Poisson model can be used, for instance, to model the total amount of claims for an insurance company.

8.7 Weibull distribution

The Weibull distribution has the pdf

$$f(y; \nu, \lambda) = \frac{\nu}{\lambda} \left(\frac{y}{\lambda}\right)^{\nu-1} \exp\left(-\left(\frac{y}{\lambda}\right)^{\nu}\right), \quad y \geq 0, \tag{95}$$

where $\nu > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter. The Weibull distribution belongs to the exponential family only if the shape parameter is known. The special case $\nu = 1$ is the exponential distribution.

8.8 Pareto distribution

The Pareto distribution has the pdf

$$f(y; \nu, c) = \frac{\nu c^{\nu}}{y^{\nu+1}}, \quad y \geq c, \tag{96}$$

where $\nu > 0$ is the shape parameter and $c > 0$ is the scale parameter.

9 Time-to-event response

9.1 Representations of time-to-event data

Data matrix – one time variable: the time under follow-up

x	t	d
5.4	10	0
6.0	7.7	1
4.3	10	0
4.4	10	0
5.1	4.5	1
5.9	10	0
⋮	⋮	

Data matrix – two time variables: the start and the end of the follow-up

x	t_1	t_2	d
5.4	41.3	51.3	0
6.0	60.3	68.0	1
4.3	54.4	64.4	0
4.4	62.9	72.9	0
5.1	45.0	49.5	1
5.9	58.7	68.7	0
⋮	⋮		

Analysis of time-to-event data is also known by names survival analysis, lifetime data analysis, failure time analysis, reliability analysis and duration analysis.

9.2 Censoring and truncation

In time-to-event data, the exact event times are not always available for all observations:

Right censoring : It is only known that $T_i > c_i$. c_i is observed.

Left censoring : It is only known that $T_i < c_i$. c_i is observed.

Interval censoring : It is only known that $c_{li} < T_i < c_{ui}$. c_{li} and c_{ui} are observed.

Some observations may be completely missing:

Left truncation : If $T < c$ the observation is not present in the data set. c is known.

Right truncation : If $T > c$ the observation is not present in the data set. c is known.

9.3 Prospective and retrospective studies

In a prospective study, a cohort of subjects is followed up for the future events. In a retrospective study, data on the past events are collected. The study illustrated in Figure 1 is a prospective study with some retrospective characteristics.

9.4 Survival function and hazard function

Consider event time T with the cdf $F(t)$ and the pdf $f(t)$. Survival (or survivor) function gives the probability the time of the event is later than a specified time

$$S(t) = 1 - F(t) = P(T > t). \tag{97}$$

Hazard function is defined as the event rate at time t conditional on survival until time t or later

$$\lambda(t) = \lim_{h \to 0^+} \frac{P(t < T < t + h \mid T \geq t)}{h} = \frac{f(t)}{S(t)}. \tag{98}$$

Integration over time leads to cumulative hazard function

$$\Lambda(t) = \int_0^t \lambda(u)du = -\log(S(t)). \tag{99}$$

For example, exponential function with the pdf $f(t) = \lambda e^{-\lambda t}$ and the cdf $F(t) = 1 - e^{-\lambda t}$ has the hazard function $\lambda(t) = \lambda$, i.e. the hazard does not depend on the time.

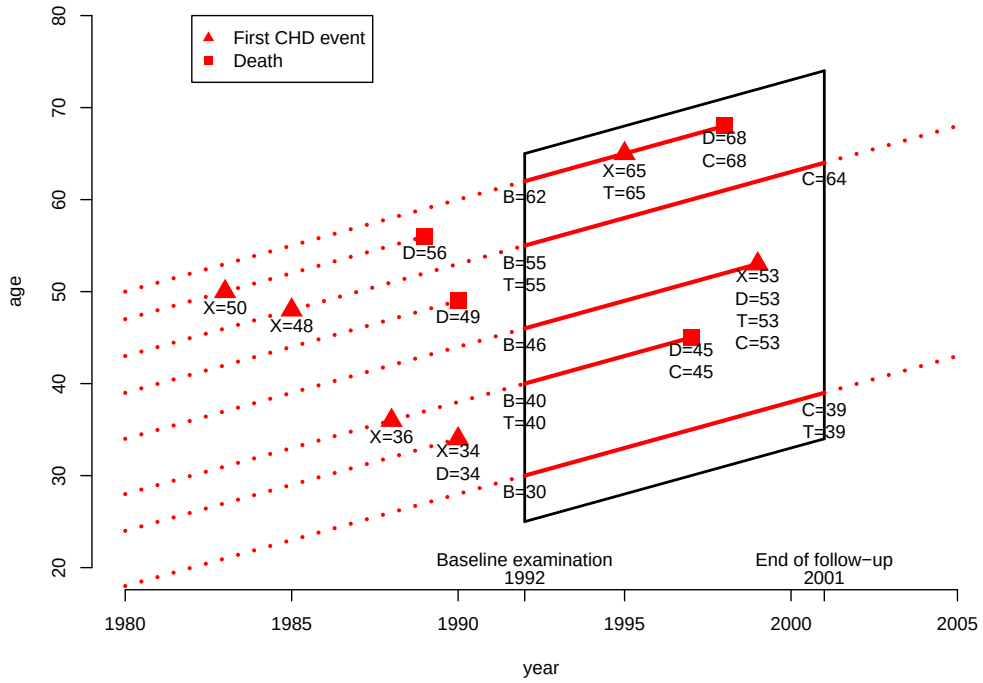

Figure 1: Illustration of a study design leading to left and right censored data with left truncation. The Lexis diagram of a cohort study is displayed. The follow-up period is from the year 1992 to the year 2001 and the age of the subjects is 25–65 years at the baseline examination. The following variables are presented: B = age at baseline examination, X = time of first coronary heart disease (CHD) event, C = censoring time, T = observed time and D = time of death. In the diagram, the data of eight subjects are presented. Two subjects have an event observed during the follow-up ($X = 65$ and $X = 53$). One of the events is fatal ($X = 53$ and $D = 53$) and the other is non-fatal ($X = 65$ and $D = 68$). One subject is right censored ($C = 39$). Two subjects have a left censored event ($X = 48$ and $X = 36$). At the baseline examination, the existence of a left censored event is recorded but the exact time of an event remains unknown. One of the subjects with left censored event dies during the follow-up period ($D = 45$); the other survives up to the end of follow-up ($C = 64$). Three subjects are completely unobserved ($D = 56$, $D = 49$ and $D = 34$). One of them had fatal CHD event ($X = 34$ and $D = 34$), one had a non-fatal event ($X = 50$) and died later ($D = 56$) and one died ($D = 49$) without a preceding CHD event.

9.5 Proportional hazards model

In proportional hazards model, the covariates have a multiplicative effect on the hazard function

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}), \tag{100}$$

where $\lambda_0(t)$ defines the baseline hazard. The baseline hazard function may be defined parametrically, for example, the hazard function of the Weibull distribution is often used. However, the most popular choice is the Cox model where $\lambda_0(t)$ is specified as a function that changes only at observed events times $\{t_i : d_i = 1\}$. The Cox model can be estimated via partial likelihood. The likelihood contribution of an event at time t_i equals

$$\frac{R_i(t_i)\lambda_i(t_i)}{\sum_{j=1}^{N} R_j(t_i)\lambda_j(t_i)}, \tag{101}$$

where $R_i(t)$ is the at-risk indicator. Censored events contribute to the partial likelihood only through their presence in the at-risk set.

In R, Cox models can be fitted with `cph` from the package `Design` or `coxph` from the package `survival` and Weibull models can be fitted with `weibreg` from the package `eha`. The response is defined as a survival object which can be created with the function `Surv`.