Statistics formulas for use in the exam of "Applied mathematics and biostatistics"

1 Preliminaries

This document may be used during the written exam of the course "Applied mathematics and biostatistics". It can be printed out one or double sided on A4 paper. Arbitrary handwritten notes may be added on both sides. The exam covers the entire subject taught during the course, not only the content of this document!

2 Random variables

Random variables

In statistics observed measurement variables are modelled by *random variables* (RV), often denoted by X:

discrete: sample space is finite or countably infinite, e.g., set of integers, subsets of positive integers, ...: $\{x_1, x_2, \ldots\}$

continous: sample space is uncountably infinite, e.g., real numbers, intervals, ...

Discrete random variables

- The sample space of X is at least countably infinite
$$M_X = \{x_1, x_2, \ldots\} \ .$$
- *Probability mass function*
$$p_X(x_i) = P(X = x_i)$$
- The sum of probabilities of all possible realisations x_i has to be equal to 1, i.e.:
$$\sum_{x_i \in M_X} p_X(x_i) = 1$$
Note: The probabilities of possible realisations x_i do not have to be necessarily equal!

Continuous random variables

- The sample space of X is uncountably infinite, e.g., interval of real numbers.
- Because there exists uncountably infinite many elements x in the sample space of X, we have
$$P(X = x) = 0 \ .$$
Positive probabilities, i.e., probabilites ≥ 0, can only be assigned to intervals!
- *Density function*:
non-negative function $f_X : \boldsymbol{R} \to \boldsymbol{R}_+$
- Properties:
$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\, dt$$
$$P(a < X \leq b) = \int_a^b f_X(x)\, dx, \qquad \int_{-\infty}^{\infty} f_X(x)\, dx = 1$$
- The function $F_X(x) = P(X \leq x)$ is called *Cumulative distribution function* (*CDF*); it can also be defined for discrete random variables in an analogue way.

Expectation and variance

$$\mathrm{E}(X) = \mu_X \quad = \quad \begin{cases} \sum_{x_i \in M_X} x_i\, p_X(x_i) & \text{if } X \text{ discrete} \\ \int_{-\infty}^{\infty} x\, f_X(x)\, dx & \text{if } X \text{ continuos} \end{cases}$$

$$\mathrm{Var}(X) = \mathrm{E}((X - \mu_X)^2) \quad = \quad \begin{cases} \sum_{x_i \in M_X} (x_i - \mu_X)^2\, p_X(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} (x - \mu_X)^2\, f_X(x)\, dx & \text{continuous} \end{cases}$$

Univariate normal distributions

A continuous random variable X is called normally distributed,
$X \sim N(\mu, \sigma^2)$, if it has the following density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) .$$

$$\begin{aligned} \mathrm{E}(X) &= \mu \\ \mathrm{Var}(X) &= \sigma^2 \end{aligned}$$

A univariate normal distribution with $\mu = 0$ and $\sigma^2 = 1$ is called standard normal distribution.

Lognormal distribution

A log-normal distribution is a continuous probability distribution of a non-negative random variable whose logarithm is normally distributed. If Y is a random variable with a normal distribution, then $X = \exp(Y)$ has a log-normal distribution; likewise, if X is log-normally distributed, $X \sim LN(\mu, \sigma^2)$, then $Y = \log(X)$ is $N(\mu, \sigma^2)$-distributed.

The density function of X is defined by

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-(\log(x)-\mu)^2/2\sigma^2\right), \qquad x > 0.$$

$$\begin{aligned} \mathrm{E}(X) &= \exp(\mu + \sigma^2/2), \\ \mathrm{Var}(X) &= \exp(2\mu + \sigma^2) \cdot (\exp(\sigma^2) - 1). \end{aligned}$$

Confidence interval for the mean

Let's assume the sample is *arbitrarily* distributed with *known variance* σ^2. Then

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

is *approximately* standard normally distributed and the limits of the *approximate* confidence interval are as follows

$$\begin{aligned} \theta_l &= \bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \\ \theta_u &= \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \end{aligned}$$

where z_γ is the γ-quantile of $N(0,1)$. If the variance is unknown we may again substitute σ^2 by $\hat{\sigma}^2$. However, this approximation is only valid for large n.

Hypothesis testing

Let's sort out the *formal steps* of statistical hypothesis testing:

1. Fix the significance level (or size) α (which is equal to fixing the confidence level).
2. Formulate the null hypothesis H_0 and the alternative hypothesis H_1.
3. Identify a test statistic that will assess the evidence against the null hypothesis. With the test statistic one can estimate how likely a certain property of the sample is if the null hypothesis were true. Under the null hypothesis the test statistic has a certain pre-defined probability distribution which can be used to construct a confidence interval for the test statistic. Hence, the decision problem is reduced to a single value.
4. If the test statistic is outside the confidence intervall (= acceptance region) the null hypothesis is rejected. (The null hypothesis is too unlikely given the sample.)
5. The used test statistic specifies which property of the sample is examined (location, spread, ...).

Type I error: The null hypothesis is true, although it is rejected.
Type II error: The alternative hypothesis is true, but we fail to reject the null hypothesis.

	Truth	
Test decision	H_0	H_1
do not reject H_0	1 - size	type II error
reject H_0	type I error	power

Optimal test: maximizing the power given the size; minimizing both types of errors is not possible.

The balance between both types of errors depends usually on the practical application.

The most important tests

t-test: testing means, one-sample t-test, independent two-sample t-test, t-test for paired samples; for small sample(s) we have to assume that the data generating process follows a normal distribution

rank tests: non-parametric test testing medians, e.g., Wilcoxon signed-rank test, Wilcoxon rank-sum test; assuming arbitrary, but identically shaped distribution(s).

more than 2 groups: ANOVA (assuming normal distribution), Kruskal-Wallis test (based on ranks).

variances: F-test (two groups), Levene's test (more groups).

contingency tables: χ^2-test of independence.

Margin of error for yes/no-questions

Let's ask n persons a yes/no-question. Then, the number of "yes" is binomially distributed with parameters n and p, where p is the true (but usually unknown) percentage of persons answering "yes".

A single person can answer "yes" (1) or "no" (0); this corresponds to a random variable X with

$$\begin{aligned} \mathrm{E}(X) &= p * 1 + (1-p) * 0 = p \\ \mathrm{Var}(X) &= \mathrm{E}(X^2) - (\mathrm{E}X)^2 = \\ &= p * 1^2 + (1-p) * 0^2 - p^2 = \\ &= p - p^2 = p(1-p) \end{aligned}$$

The precentage for n persons is calculated as the sample mean

$$\hat{p} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad .$$

According to the central limit theorem it is asymptotically normal distributed with expectation p and variance $p(1-p)/n$:

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \quad \sim \quad N(0,1)$$

As 1.96 is the 0.975-quantile of the standard normal distribution, z is within the range [-1.96,+1.96] with probability 95%. This can be used to calculate 1. the margin of error given the sample size, or 2. the sample size given the desired margin of error:

$$\begin{aligned} \Delta p &\approx \frac{1}{\sqrt{n}} \\ n &\approx \frac{1}{(\Delta p)^2} \end{aligned}$$

3 Linear models

Data: $(y_i, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, with a metric variable y and several metric or (binary coded) categorial regressors $x_1, \ldots, x_k$.

Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i \,, \quad i = 1, \ldots, n \,.$$

The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed (i.i.d.) with

$$\mathrm{E}(\varepsilon_i) = 0 \,, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2.$$

In matrix form we have

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \boldsymbol{I}$. The design matrix $\boldsymbol{X}$ has full (column) rank, i.e., $rg(\boldsymbol{X}) = k + 1 = p$.

The estimated linear function

$$\hat{y}_i = \hat{f}(x_1, \ldots, x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

can be interpreted as an estimate $\widehat{\mathrm{E}}(y|x_1, \ldots, x_k)$ of the conditional expectation of y given the covariates $x_1, \ldots, x_k$ and hence used to predict y.

Estimation of parameters and residuals

In order to estimate the unknown parameter β_i we use the ordinary least squares (OLS) method:

$$\text{SQR} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\Big(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik})\Big)^2 \quad \rightarrow \quad \min$$

The residuals are computed by $\hat{\varepsilon}_i = y_i - \hat{y}_i$.

1. The average of the residuals is equal to zero:

$$\sum_{i=1}^{n}\hat{\varepsilon}_i = 0 \text{ bzw. } \bar{\hat{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i = 0.$$

2. The average of the estimated values $\hat{y}_i$ is equal to the average of the observed values y_i:

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \bar{y}$$

.

3. The center of the data is an element of the fitted hyperplane:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \cdots + \hat{\beta}_k\bar{x}_k.$$

Linear influence of covariates

At a first glimpse, it seems to be very restrictive regarding only linear models of covariates. Nevertheless, using linear models, we can model nonlinear relationships too. E.g.:

$$y_i = \beta_0 + \beta_1 \log(z_i) + \varepsilon_i,$$

Here, the influence of the explanatory variable z_i is a logarithmic one. With $x_i = \log(z_i)$ we get a linear model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. In general, we can transform nonlinear relationships into linear models as long as they are *linear in the parameters.*

Estimation in R

In R parameters of linear models are estimated using the function `lm`:

```
> meinmodell <- lm(y ~ x1 + x2 + x3, data=meinedaten)
```

The first argument of the function, i.e., the part with the symbol $\sim$, is the so-called model formula. On the left hand side we put the name of the measurement variable (column in the data set), on the right hand side we put all predictors:

1. Variables seperated by the symbol "`+`" will not be added, but included in our model as single predictors.
2. Using terms like "`x1 * x2`" will lead to models where the main effects as well as the interactions are included in the model (cf. later).
3. A "`.`" in the model formula refers to all variables (main effects) in the data set not used on the left hand side. E.g.: `y ~ .`
4. A "`-`" in front of a variable means to exclude this variable from the model. E.g.: `y ~ . - x1`

An example: The call

```
> meinmodell <- lm(y ~ x1 + x2 * x3, data=meinedaten)
```

corresponds to the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_2 x_3 + \epsilon \ .$$

Basic mathematical functions can be directly used within formulas.
The call

```
> meinmodell <- lm(log(y) ~ x1 + x2 * sin(x3), data=meinedaten)
```

corresponds to the model

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \sin(x_3) + \beta_4 x_2 \sin(x_3) + \epsilon \ .$$

Variance decomposition

Question: How well does the hyperplane fit the data? The variance is a measure of the variability of the response variable Y:

$$\hat{s}_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

In regression analysis we usually use the following sum of squares

$$\text{SQT} = (n-1)\hat{s}_Y^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

(SQT = "**S**um of s**Q**uares **T**otal").

We have

$$\text{SQT} = \text{SQE} + \text{SQR}$$

with

- **S**um of s**Q**uares **T**otal
$$\text{SQT} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
- **S**um of s**Q**uares **E**xplained
$$\text{SQE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$
- **S**um of s**Q**uares **R**esidual
$$\text{SQR} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Explained variability

R^2:

$$R^2 = \frac{\text{SQE}}{\text{SQT}} = 1 - \frac{\text{SQR}}{\text{SQT}} \quad \in [0,1]$$

We have $R^2 = \text{Cor}(y,\hat{y})^2 = r_{y\hat{y}}^2$.

$R^2 \approx 0$**:** The variance of the residuals is as large as the variance of Y, there is no linear (!) influence of X on Y.

$R^2 \approx 1$**:** The variance of the residuals is (almost) equal to zero, the observed data are (almost) part of the fitted hyperplane.

Hypothesis testing

- The parameter estimators are *random variables* and depend on the sample.
- If the errors are normally distributed the parameter vector $\hat{\boldsymbol{\beta}}$ is multivariate normally distributed (with true but unknown location parameter $\boldsymbol{\beta}$ which is equal to the expected value of $\hat{\boldsymbol{\beta}}$, i.e., $\text{E}(\hat{\beta}_j) = \beta_j$, and covariance matrix that can be estimated using the data).
- Without normality assumption of the errors the parameter vector $\hat{\boldsymbol{\beta}}$ is approximately multivariate normally distributed only in case of large samples.

The normal distribution of the parameters can be used to provide statistical inference for linear models:

- t-tests for location
- z-tests for location in case of large samples
- F-tests to compare variances

t-test

Using t-tests we may test if a single coefficient $\hat{\beta}_j$ differs significantly from a prespecified value. Usually we test $\beta_j = 0$, because, if we cannot reject the null hypothesis, we then may exclude the corresponding variable from the model.

Let se_j denote the square root of the j-th diagonal element of $\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. We then have

$$t = \frac{\hat{\beta}_j}{se_j} \sim t_{n-p}.$$

The null hypothesis $\beta_j = 0$ can be rejected if

$$|t| > t_{n-p}(1 - \alpha/2).$$

We know that the t-distribution converges to the standard normal distribution as $(n - p) \to \infty$. Hence, the t-test coincide with the z-test for sufficiently large n (and fixed p).

Confidence and prediction intervals

If we want to construct confidence intervals for the response variable y we have to consider two sources of uncertainty:

- The true but unknown vector of regression coefficients $\boldsymbol{\beta}$ is approximated by $\hat{\boldsymbol{\beta}}$.
- The observed data scatter around the regression line (or hyperplane) with variance σ^2.

The *confidence interval for the expected value* $\mu_0 = E(y_0)$ of a new observed value y_0 at $\boldsymbol{x}_0$ with confidence level $1 - \alpha$ is given by

$$\hat{y}_0 \pm t_{n-p}(1 - \alpha/2)\hat{\sigma}\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}.$$

The *prediction interval for a new observed value* y_0 at $\boldsymbol{x}_0$ with confidence level $1 - \alpha$ is given by

$$\hat{y}_0 \pm t_{n-p}(1 - \alpha/2)\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}.$$

Categorical covariates: dummy coding

To model the influence of a c-categorical covariate $x \in \{1, \ldots, c\}$ using dummy coding we define $c-1$ dummy variables

$$x_{i1} = \begin{cases} 1 & x_i = 2 \\ 0 & \text{else} \end{cases} \quad \ldots \quad x_{i,c-1} = \begin{cases} 1 & x_i = c \\ 0 & \text{else} \end{cases}$$

with $i = 1, \ldots, n$ and include them as explanatory variables in our regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{i,c-1} x_{i,c-1} + \ldots + \varepsilon_i$$

To keep the coefficients identifiable the dummy variable of one category—here the first one—is not included in the model. This category is called the reference category. The values of the regression estimates are then interpreted in comparison to the omitted category.

Comparing different models

Two models are called *nested* if all variables of the smaller model are also included in the larger one. In this case, the (mean squared) error of the smaller model using the training data set is always equal to or larger than the one of the larger model.

Nested models can be compared using F-tests (cf. ANOVA): Is the error sum of squares (i.e., the variance) of the larger model significantly smaller given the additional number of estimated parameters?

The two most important approaches are:
ANOVA type 1: nested models with variables in the same order as given in the model formula.
ANOVA type 2: compares the complete model with all models where one variable is excluded at a time.

Adjusted R^2

The usage of R^2 is suitable to only a limited extent comparing different models because it increases if new covariates are included in the model.

To cope with this problem the so called *adjusted* R^2 is used. Here, according to the number of parameters the R^2 is adjusted in a way that it does not necessarily increase if additional covariates are included in the model:

$$\bar{R}^2 = 1 - \frac{n-1}{n-p}(1 - R^2).$$

Very popular; it is calculated per default by all statistical software packages, but it penalises new covariates too little.

Information criterion AIC

The most used model selection criterion in the framework of maximum likelihood estimation is Akaike's information criterion (AIC). Generally, the AIC is defined by

$$\mathrm{AIC} = -2 \cdot l(\hat{\beta}, \hat{\sigma}^2) + 2\,|p|$$

where $l(\hat{\beta}, \hat{\sigma}^2)$ is the maximal value of the log-likelihood function, and p is the number of model parameters. The derivation of the AIC is based on a Taylor expansion of the expected error and ignores constant terms of the likelihood function. Hence, the AIC can only be used to compare *nested models of the same model family*; its value is only relative. Models with a smaller AIC value are preferred.

Bayesian information criterion BIC

The Bayesian information criterion BIC is defined by

$$\mathrm{BIC} = -2 \cdot l(\hat{\beta}, \hat{\sigma}^2) + \log(n)\,|p|,$$

Again, models with a smaller BIC value are preferred. The derivation of AIC and BIC is differently motivated. From a practical point of view the main difference is that the BIC penalises complex models more than the AIC (because $\log(n) > 2$ for $n \leq 8$).

4 Generalized linear models

Binomial regression models

Goal: Modelling and estimating the influence of covariates on the (conditional) probability

$$\pi_i = \mathrm{P}(y_i = 1 \,|\, x_{i1}, \ldots, x_{ik}) = \mathrm{E}(y_i \,|\, x_{i1}, \ldots, x_{ik})$$

in case $y_i = 1$ given the covariate values $x_{i1}, \ldots, x_{ik}$. The response variables are assumed to be (conditionally) independent.

Common solution approach for all usually used binomial regression models:
linking of the probability π_i and the linear predictor η_i by

$$\pi_i = h(\eta_i) = h(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}) \quad .$$

- *Response function*: the function h is strictly increasing and maps onto the interval $[0, 1]$, i.e., $h(\eta) \in [0, 1]$, $\forall \eta \in \mathbb{R}$. Especially, many cumulative distribution functions can be used as response functions.

- *Link function*: This is the inverse $g = h^{-1}$ of the response function, i.e. $\eta_i = g(\pi_i)$.
  Logit model

$$\pi = \frac{\exp(\eta)}{1 + \exp(\eta)} \qquad \Longleftrightarrow \qquad \log \frac{\pi}{1 - \pi} = \eta.$$

  Probit model

$$\pi = \Phi(\eta) \qquad \Longleftrightarrow \qquad \Phi^{-1}(\pi) = \eta.$$

  complementary log-log model

$$\pi = 1 - \exp(-\exp(\eta)) \qquad \Longleftrightarrow \qquad \log(-\log(1 - \pi)) = \eta.$$

Interpretation of the logit model

Using the linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

the *odds*

$$\frac{\pi_i}{1 - \pi_i} = \frac{\mathrm{P}(y_i = 1 \,|\, \boldsymbol{x}_i)}{\mathrm{P}(y_i = 0 \,|\, \boldsymbol{x}_i)}$$

are equal to the multiplicative model

$$\frac{\mathrm{P}(y_i = 1 \,|\, \boldsymbol{x}_i)}{\mathrm{P}(y_i = 0 \,|\, \boldsymbol{x}_i)} = \exp(\beta_0) \cdot \exp(x_{i1}\beta_1) \cdot \ldots \cdot \exp(x_{ik}\beta_k).$$

If we increase x_{i1} by 1, i.e., $x_{i1} + 1$, the odds ratio is

$$\frac{\mathrm{P}(y_i = 1 \,|\, x_{i1} + 1, \dots)}{\mathrm{P}(y_i = 0 \,|\, x_{i1} + 1, \dots)} \Big/ \frac{\mathrm{P}(y_i = 1 \,|\, x_{i1}, \dots)}{\mathrm{P}(y_i = 0 \,|\, x_{i1}, \dots)} = \exp(\beta_1).$$

$$\begin{aligned}
\beta_1 > 0 \quad &: \quad \text{the odds } \mathrm{P}(y_i = 1)/\mathrm{P}(y_i = 0) \text{ increase,} \\
\beta_1 < 0 \quad &: \quad \text{the odds } \mathrm{P}(y_i = 1)/\mathrm{P}(y_i = 0) \text{ decrease,} \\
\beta_1 = 0 \quad &: \quad \text{the odds } \mathrm{P}(y_i = 1)/\mathrm{P}(y_i = 0) \text{ stay the same.}
\end{aligned}$$

Maximum likelihood estimation

Independently of the used link function the likelihood of the model cannot be solved for $\boldsymbol{\beta}$ in closed form. Hence, we have to maximize the (log-)likelihood function numerically → Fisher Scoring. We can proof: as $n \to \infty$ the maximum likelihood estimator (MLE) exits, it is a consistent estimator and asymptotically normally distributed. (It is sufficient that the sample size $n \to \infty$.)

Significance of parameters

Testing

$$H_0 : \beta_j = 0 \quad \text{against} \quad H_1 : \beta_j \neq 0,$$

with β_j an element of $\boldsymbol{\beta}$, we again check if we can exclude the relevant variable from the model. As $\hat{\beta}$ is only asymptotically normally distributed, we always compare the test statistic t_j with quantiles of the standard normal distribution → "z-statistic", "z-test".

Linearly separable classes

A nasty feature of binomial regression (whether logit or probit or ...) is that the easiest case of linearly separable classes leads to 'infinite' coefficients, i.e., the MLE $\hat{\beta} = \pm\infty$.

Hence, in all (reasonable implemented) statistical software packages the numerical maximization of the likelihood function will be stopped after a maximal number of Fisher Scoring iterations and an additional warning will be given. In this case, the estimated parameters will simply be 'very large'.

In such a situation, e.g., Fisher's discriminant analysis will yield a suitable model and, especially, the separating hyper planes.

Poisson regression for count data

The response variables y_i take values in $\{0, 1, 2, \dots\}$ and are (conditional) independent given the covariates $x_{i1}, \dots, x_{ik}$.

Log-linear Poisson model: $y_i \,|\, \boldsymbol{x}_i \sim Po(\lambda_i)$ with

$$\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \quad \text{bzw.} \quad \log \lambda_i = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Model with overdispersion:

$$\mathrm{E}(y_i \,|\, \boldsymbol{x}_i) = \lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}), \quad \mathrm{Var}(y_i \,|\, \boldsymbol{x}_i) = \phi\lambda_i$$

with overdispersion parameter ϕ. The usual Poisson distribution has only one parameter λ; λ is equal to the expectation as well as to the variance of the distribution:

$$P\{X = k\} = \frac{\lambda^k e^{-\lambda}}{k!}, \quad E(X) = \mathrm{Var}(X) = \lambda \; .$$