Part II: Generalized Linear Models

Chapter II.2

Theory of Generalized Linear Models

The Link Function

The Link Function

The link function g relates the linear predictor $\mathbf{X}\boldsymbol{\beta}$ to the expected value of the response vector $\boldsymbol{\mu} = \mathsf{E}(\boldsymbol{Y})$:

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = g[\mathsf{E}(\boldsymbol{Y})] = \mathbf{X}\boldsymbol{\beta} \ .$$

- In LMs, the expected value equals the linear predictor, i.e. the link is the identity function. In this case, the linear predictor and the expected value can take any real value.

Theoretically, the **link function** g can be *any* monotonic and differentiable function. Practically, the options are limited, since the link is chosen so that the inverse $g^{-1}(\boldsymbol{\eta}) = \boldsymbol{\mu}$ leads to *admissible values* for $\boldsymbol{\mu}$ (and simple functions of $\boldsymbol{\vartheta}$).

II.2.9 Remark

> The link function g specifies the nature of the distribution considered for the error term ε.

II.2.10 Example (Binomial distributed response variable)

If $Y \sim \mathcal{B}(m,\pi)/m$, then $\mu = \pi$ (s. Example II.2.5) and we have $\mu \in (0,1)$. Thus a link function should map the interval $(0,1)$ on the whole real line. The links that are more often used for binomial data are

1. the *logit* $\eta = g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$,
2. the *probit* $\eta = g(\pi) = \Phi^{-1}(\pi)$, where Φ is the Normal cumulative density function (cdf), and
3. the *complementary log-log* $\eta = g(\pi) = \log\left(-\log(1-\pi)\right)$.

II.2.11 Example (Poisson distributed response variable)

For $Y \sim \mathcal{P}(\mu)$, it must hold $\mu > 0$. A suitable link function is the log-link

$$\eta = g(\mu) = \log(\mu) \ .$$

Data Transformation vs. GLM

Traditionally, the response variable Y is often transformed in order to have approximately a normal distribution with constant variance. Then, on the transformed scale $Y^* = g(Y)$, ordinary least squares methods for linear models are applicable.

II.2.12 Example (Poisson distributed count data)

Consider $Y_i \sim \mathcal{P}(\mu_i)$, $i = 1, \ldots, n$. The distribution of Y_i is skewed to the right with $\mathsf{Var}(Y_i) = \mathsf{E}(Y_i) = \mu_i$ but In case $Y_i^* = g(Y_i) = \sqrt{Y_i}$ is nearly normal distributed with $\mathsf{Var}(Y_i^*) \approx 1/4$.

In a GLM, a useful link function g enables the fit of a linear model for this link without necessarily stabilizing the variance or producing normality.

II.2.13 Remark

The model parameters of a GLM with link function g describe $g\left[\mathsf{E}(Y)\right]$. These parameters describe also effects on the expected response $E(Y)$, after applying the inverse function g^{-1}. On the other hand, the parameters of a LM applied on the transformed data, by a transformation function g, describe $\mathsf{E}\left[g(Y)\right]$ and these effects cannot be transformed to effects on $E(Y)$.

The Canonical Link

A convenient link with nice properties is the canonical link.

II.2.14 Definition

The *canonical link* is a link function g such that $g(\mu_i) = \vartheta_i$, $i = 1, \ldots, n$. It connects the μ_i directly to the natural parameter ϑ_i and thus

$$\boldsymbol{\vartheta} = g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta},$$

with $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_n)'$.

II.2.15 Example

The canonical link $g(\mu) = \theta$ for a normal, Poisson or Binomial response (s. Example II.2.5):

	θ	$\mathsf{E}(Y) = b'(\theta)$	$g(\mu)$
$Y \sim \mathcal{N}(\mu,\sigma^2)$	μ	$\mu = \theta$	μ
$Y \sim \mathcal{P}(\mu)$	$\log(\mu)$	$\mu = e^\theta$	$\log(\mu)$
$Y \sim \mathcal{B}(m,\pi)/m$	$\log(\frac{\pi}{1-\pi})$	$\mu = \pi = \frac{e^\theta}{1+e^\theta}$	$\log(\frac{\mu}{1-\mu})$