Applied Bayesian Data Analysis — Chapter 15

Kim Albertsson

CERN and Luleå University of Technology

kim.albertsson@ltu.se

December 3, 2019

Chapter 15

Overview of the Generalised Linear Model

“Generalised Linear Model” (GLM)

Estimate parameters as first-order functions of input variables (as opposed to zeroth-order). For one input variable:

$$\begin{aligned} y &\sim \mathcal{N}(\mu, \sigma) \\ \mu &= \beta_0 \end{aligned} \quad \text{cmp.} \quad \begin{aligned} y &\sim \mathcal{N}(\mu, \sigma) \\ \mu &= \beta_0 + \beta_1 x_1 \end{aligned}$$

The linear model is the simplest model. The GLM in full:

$$\begin{aligned} \operatorname{lin}_\beta(\widehat{x}) = \beta_0 + \sum_i^K \beta_i x_i + \sum_i^K \sum_{j=i+1}^K \beta_i x_i x_j \\ \mu = f(\operatorname{lin}_\beta(\widehat{x}), \theta_A) \\ y \sim \operatorname{pdf}(\mu, \theta_B) \end{aligned}$$

Note: Other models are possible e.g. non-linear function approximators (Restricted Boltzmann Machines/Deep Belief Networks). Analysis becomes more involvled.

Types of variables

Input/Output

Predictor
: x.

Predicted
: y.

Scale types

Metric
: Distances make sense. Ordinal. Often continuous. E.g. tone frequency.

Ordinal
: Ordinal. E.g. "first", "second".

Nominal
: Non-ordinal. Discrete. E.g. labels: "plane", "cat".

Count
: Discrete metric.

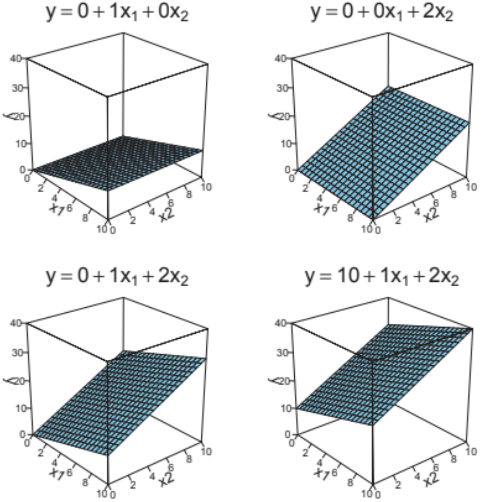
Linear combination of metric predictors (I)

Linear function: Additivity $\left(f(x+a) = f(x) + f(a)\right)$ and Homogeneity $\left(f(ax) = a \cdot f(x)\right)$

No interactions:

$$\mathrm{lin}_\beta(x) = \beta_0 + \sum_k^K \beta_k x_k$$

Linear combination of metric predictors (I)



Linear combination of metric predictors (II)

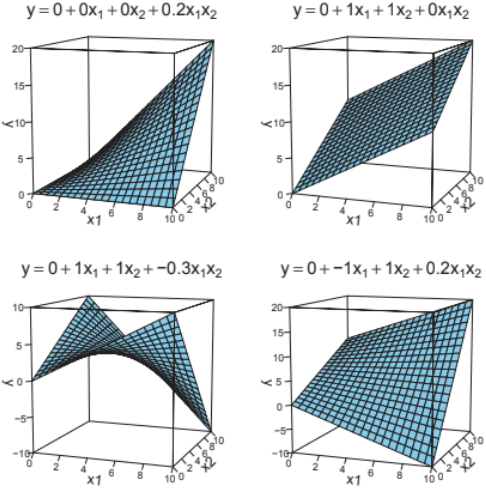
Bilinear function: $\sim$ polynomial function with no self-interactions:

$$\begin{aligned} f(x+a, y) &= f(x, y) + f(a, y) \\ f(x, y+a) &= f(x, y) + f(x, a) \\ f(ax, y) &= f(x, ay) = a \cdot f(x, y) \end{aligned}$$

With interactions (conditionally linear):

$$\operatorname{lin}_\beta(x) = \beta_0 + \sum_k^K \beta_k x_k + \sum_j^K \sum_{k=j+1}^K \beta_{jk} x_j x_k$$

Linear combination of metric predictors (II)



Aside: Notation for linear combinations (I)

Common approach: write down the power series expansion

$$f(x_1) = \beta_0 + \beta_1 x_1 = \beta_0 x_1^0 + \beta_1 x_1^1$$

Works for two variables:

$$\begin{aligned} f(x_1, x_2) &= \beta_{00} x_1^0 x_2^0 + \beta_{10} x_1^1 x_2^0 + \beta_{01} x_1^0 x_2^1 + \beta_{11} x_1^1 x_2^1 + \dots \\ &= \sum_{i=0}^{P} \sum_{j=0}^{P} \beta_{ij} x_1^i x_2^j \end{aligned}$$

Problem: Scaling to arbitrary dimensions, notation unwieldy:

$$f(\widehat{x}) = \sum_{i=0}^{P} \sum_{j=0}^{P} \sum_{k=0}^{P} \dots \left(\beta_{ijk\dots} x_1^i x_2^j x_3^k \dots \right)$$

Aside: Notation for linear combinations (II)

(One) solution: Introduce bias in feature vector:
$\widehat{x} =< x_1.x_2, \dots > \rightarrow \widehat{x} =< 1, x_1, x_2, \dots >$:

$$
\begin{aligned}
f(x_1) &= \beta_0 + \beta_1 x_1 = \beta_0 x_0 + \beta_1 x_1 : x_0 = 1 \\
f(x_1, x_2) &= \beta_{00} x_0 x_0 + \beta_{01} x_0 x_1 + \beta_{02} x_0 x_2 + \beta_{12} x_1 x_2 \\
&= \beta_{00} + \beta_{01} x_1 + \beta_{02} x_2 + \beta_{12} x_1 x_2 \\
f(\widehat{x}) &= \sum_{i=0}^{K} \sum_{j=i+1}^{K} \beta_{ij} x_i x_j : x_0 = 1
\end{aligned}
$$

Note: This is a more compact representation of
$\mathrm{lin}_\beta(x) = \beta_0 + \sum_i^K \beta_i x_i + \sum_i^K \sum_{j=i+1}^K \beta_i x_i x_j$
Note: If $j = i + 1$ is changed to $j = i$, self-interactions are taken into consideration (but this violates conditional linearity).

Nominal predictors

One-hot coding: Each feature is vector (instead of scalar) $\widehat{x} =< 0, 0, 1 >$
For metric predictors each sample has a number of features:

feature 0	feature 1	feature 2	...
x_0	x_1	x_2	...

For nominal predictors each feature is a vector:

	feature 0	feature 1	feature 2	...
category 0	x_{00}	x_{01}	x_{02}	...
category 1	x_{10}	x_{11}	x_{12}	...
category 2	x_{20}	x_{21}	x_{22}	...

Linear combinations of Nominal predictors

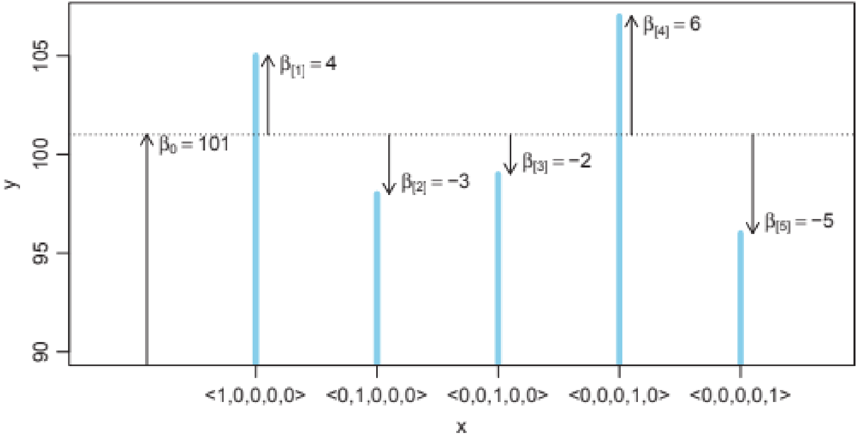
Linear combination:

$$lin(\widehat{x}) = \beta_0 + \sum_j \widehat{\beta}_j \cdot \widehat{x}_j$$

where $\cdot$ denotes the dot product.

Intuition: Think of β_0 as population "average", and β_j as "average" of group j.

Linear combinations of Nominal predictors



Summary Linear Combination

Form of linear function in GLM for different scale types (of predictor x):

Scale Type of Predictor x					
		Metric		Nominal	
Single Group	Two Groups	Single Predictor	Multiple Predictors	Single Factor	Multiple Factors
β_0	$\beta_{x=1}$ $\beta_{x=2}$	β_0 $+\beta_1 x$	β_0 $+\sum_k \beta_k x_k$ $+\sum_{j,k} \beta_{j\times k} x_j x_k$ $+\left[\text{higher-order interactions}\right]$	β_0 $+\vec{\beta} \cdot \vec{x}$	β_0 $+\sum_k \vec{\beta}_k \cdot \vec{x}_k$ $+\sum_{j,k} \vec{\beta}_{j\times k} \cdot \vec{x}_{j\times k}$ $+\left[\text{higher-order interactions}\right]$

Linking from Combined Predictors to (noisy) predicted data

$$y = f(\mathrm{lin}(x)) \tag{15.11}$$

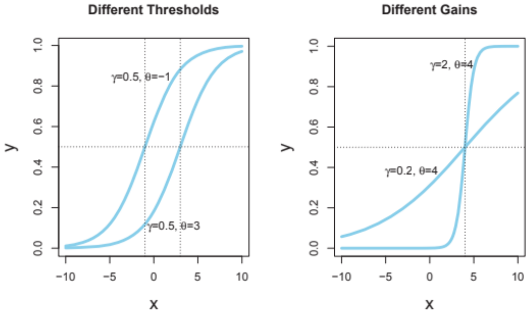
Here f is called the (inverse) link function and converts the combined predictors to an apporopriate output scale.

$$\text{Input Space} \xrightarrow{\mathrm{lin}} \text{Combined predictor Space} \xrightarrow{f} \text{Output Space}$$

Scale Type	**Task**
Metric	Regression
Nominal	Classification

Logistic function

Link function common for classification: the *logistic* function. Unrestricted domain, range $(0, 1)$. $\text{logistic}(\text{x}) = 1/(1 + \exp(\text{-x}))$

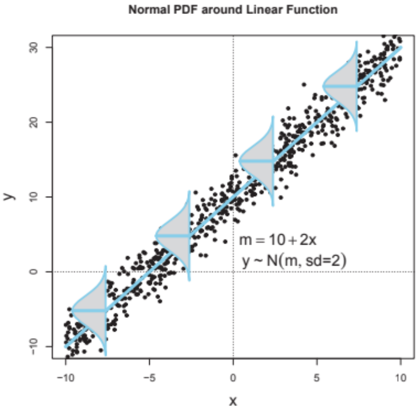


Note: Another reasonable choice: Φ, the cumulative normal distribution (convenicent for e.g. ordinal data).

Noisy output

$$y \sim \mathrm{pdf}(\mu, [\mathrm{parameters}])$$
$$\mu = f(\mathrm{lin}(x))$$

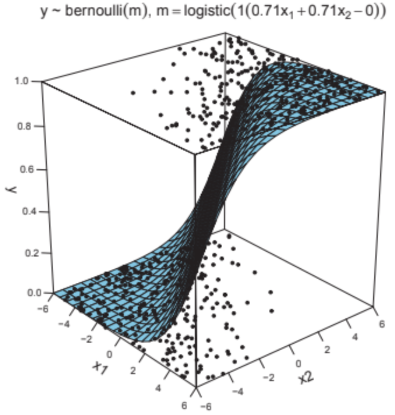
If f is indentity: Linear regression.



Noisy output (II)

$$y \sim \text{pdf}(\mu, [\text{parameters}])$$
$$\mu = f(\text{lin}(x))$$

If f is logistic: Logistic regression.
(Special case of binary classification).



Typical distributions and link functions

Scale Type of Predicted y	Typical Noise Distribution $y \sim \text{pdf}(\mu, [\text{parameters}])$	Typical Inverse-Link Function $\mu = f(\text{lin}(x), [\text{parameters}])$
Metric	$y \sim \text{normal}(\mu, \sigma)$	$\mu = \text{lin}(x)$
Dichotomous	$y \sim \text{bernoulli}(\mu)$	$\mu = \text{logistic}(\text{lin}(x))$
Nominal	$y \sim \text{categorical}(\ldots, \mu_k, \ldots)$	$\mu_k = \dfrac{\exp(\text{lin}_k(x))}{\sum_c \exp(\text{lin}_c(x))}$
Ordinal	$y \sim \text{categorical}(\ldots, \mu_k, \ldots)$	$\mu_k = \begin{array}{l} \Phi((\theta_k - \text{lin}(x))/\sigma) \\ -\Phi((\theta_{k-1} - \text{lin}(x))/\sigma) \end{array}$
Count	$y \sim \text{poisson}(\mu)$	$\mu = \exp(\text{lin}(x))$

Aside: Link function and loss function

Note: $p(y = \mathrm{A}|\widehat{x}) = \eta = f^{-1}(v) = 1 - p(y = \mathrm{B}|\widehat{x})$. I.e. binary classification given some input variable x.

Loss name	$\phi(v)$	$C(\eta)$	$f^{-1}(v)$	$f(v)$
Exponential	e^{-v}			
Logistic	$\frac{\log(1+e^{-v})}{\log(2)}$	-	$\frac{e^v}{1+e^v}$	$\log(\frac{\eta}{1-\eta})$
Square		-		
Savage		-		
Tangent		-		

en.wikipedia.org/wiki/Loss_functions_for_classification