GLMs and Causal Inference

Sofia Olhede

EPFL

December 1, 2020

1. GLM examples
2. Causal Inference

GLM Interlude

- Some more GLM examples...
- Before looking at details for non-parametrics, let us re-visit the details of the GLM specification.
- Recall for a Bernoulli random variable has pmf

$$f_Y(y) = \theta^y\{1-\theta\}^{1-y}$$
$$= \exp\{y \log \frac{\theta}{1-\theta} + \log(1-\theta)\}. \quad (1)$$

Clearly here we set $\phi = \log \frac{\theta}{1-\theta}$. We can solve for $\exp(\phi) = \frac{\theta}{1-\theta}$, with

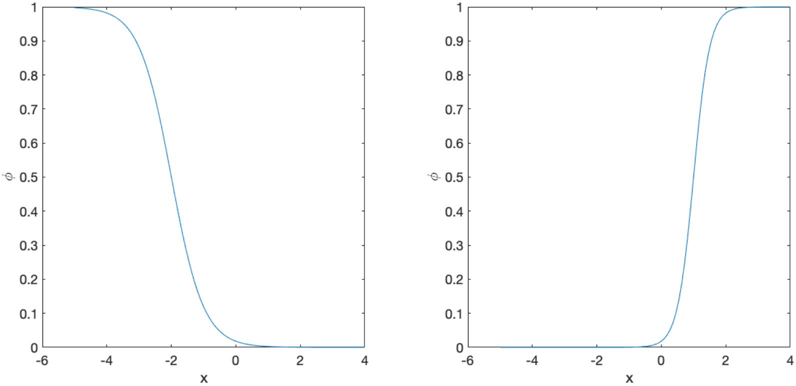
$$\theta = \frac{1}{1+\exp(\phi)} \Rightarrow 1-\theta = \frac{\exp(\phi)}{1+\exp(\phi)}.$$

For example we could look at a single covariate x_i and set

$$\phi_i = \beta_0 + \beta_1 x_i.$$

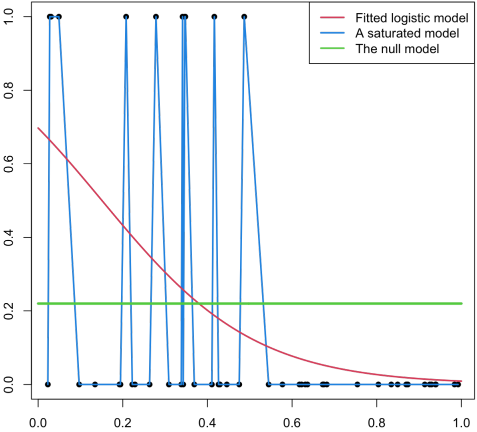
We see directly that as ϕ ranges across any value, θ is constrained to lie between zero and unity.

GLM Interlude

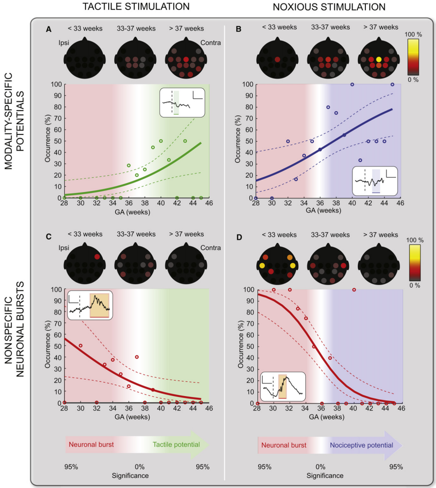


GLM Interlude

- We can examine the improvement using the deviance. Recall that
$$D = 2\{\ell_n(\widehat{\phi}) - \ell_n(\widehat{\beta})\}.$$
So we can fit both models and compare



Bernoulli and Binomial



Poisson observations

- Looking at the Poisson pmf we have

$$f_Y(y) = \frac{e^{-\mu}\mu^y}{y!} = \exp\{-\mu + y\log(\mu) - \ln y!\}. \tag{2}$$

  Here we clearly set $\phi = \log(\mu)$. Solving for μ just gives us $\mu = \exp(\phi)$. We only need the mean to remain positive so this will fix our problem.
- Again we use the deviance to assess the fit; and would compare to the model μ_i is different for each value of i.

What about the sparse GLM?

- Hastie and Park (2007) estimate the parameters of the GLM using

$$\hat{\beta}_L(\lambda) = \arg\min_{\beta}\{-\log L(\beta) + \lambda\|\beta\|_1.\}$$

- This mimics using the Lasso for the Gaussian linear model.
- We can study the geometry of this space in β. Unfortunately unlike the LASSO it is not a convex optimisation problem. This means we are not seeing the possibility of a polynomial-time algorithm solving our problem. We could also end up with multiple optima.
- Hastie and Park also extended the elastic net to this setting

$$\hat{\beta}_{EN}(\lambda_1, \lambda_2) = \arg\min_{\beta}\left\{-\log L(\beta) + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2\right\}. \quad (3)$$

- This has two penalties. Problems that arise when X has linearly dependent columns; the coefficient estimates are highly unstable.

What about the sparse GLM?

- When λ_2 is a constant, and λ_1 varies in an open set, such that the current active set remains the same, a unique, continuous and differentiable function.
- The additional penalization of the elastic net, is not either yielding a convex problem.
- Just optimizing the GLM likelihood can be problematic on its own.
- This brings us back to the penalized GLM. Augugliaro et al (2013) looked at the differential geometry of this problem.

GLM Nonparametric relationships with x_i

So far: how to estimate $g : \mathbb{R} \to \mathbb{R}$ (assumed smooth) in

$$Y_i = g(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \text{ given data } \quad \{(Y_i, x_i)\}_{i=1}^n.$$

Can extend to GLM setting as:

$$Y_i | x_i \overset{indep}{\sim} \exp\{g(x_i)y - \gamma(g(x_i)) + S(y)\}$$

- Parametrise candidate g via spline

$$s(x) = \sum_{j=1}^{n} \gamma_j B_j(x).$$

- Define matrices $\boldsymbol{B}$ and $\boldsymbol{\Omega}$ as before,

$$B_{ij} = B_j(x_i), \quad \Omega_{ij} = \int B_i''(x) B_j''(x)\, dx$$

- And consider penalised likelihood, similarly as with penalised GLM

$$\ell_n(\gamma) + \lambda \gamma^\top \boldsymbol{\Omega} \gamma = \gamma^\top \boldsymbol{B}^\top \boldsymbol{Y} - \sum_{i=1}^{n} \gamma(\boldsymbol{b}_i^\top \gamma) + \lambda \gamma^\top \boldsymbol{\Omega} \gamma.$$

GLM Nonparametric relationships with x_i

How can we generalise to multivariate covariates?

- "Immediate" Generalisation: $g : \mathbb{R}^p \to \mathbb{R}$ (smooth)

$$Y_j = g(x_{j1}, \dots, x_{jp}) + \varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Estimation by (e.g.) multivariate kernel method.
- Two basic drawbacks of this approach ...

↪ Shape of kernel? (definition of *local*)

↪ *Curse of dimensionality*

GLM Nonparametric relationships with x_i

What is "local" in $\mathbb{R}^p$, though?

$\rightarrow$ Need some definition of "local" in the space of covariates

$\hookrightarrow$ Use some metric on $\mathbb{R}^p \ni (x_1, \ldots, x_p)^\top$!

But which one?

- Choice of metric $\iff$ choice of geometry

  $\hookrightarrow$ e.g., curvature reflects intertwining of dimensions

- Geometry $\implies$ reflects structure in the covariates
  - $\underline{\text{potentially different units of measurement}}$ (variable stretching of space)
  - $\underline{g \text{ may be of higher variation in some dimensions}}$ (need finer neighbourhoods there)
  - $\underline{\text{statistical dependencies present in the covariates}}$ ("local" should reflect these)

GLM Nonparametric relationships with x_i

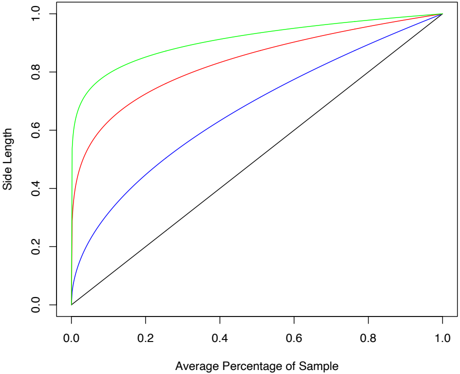


Figure: Curse of Dimensionality $(\mathrm{Unif}[0,1]^p)$: $p = 1$, $p = 2$, $p = 5$, $p = 10$

GLM Nonparametric relationships with x_i

Curse of Dimensionality

> *"neighbourhoods with a fixed number of points become less local as the dimensions increase"*
>
> *Bellman (1961)*

- Notion of local in terms of % of data: fails in high dimensions
  $\hookrightarrow$ There is too much space!
- Hence to allow for reasonably small bandwidths
  $\hookrightarrow$ Density of sampling must increase.
- Need to have ever larger samples as dimension grows.

GLM Nonparametric relationships with x_i

Attempt to find a link/compromise between:

- our mastery of 1D case (at least we can do that well ...),
- and higher dimensional covariates (and associated difficulties).

Perhaps something that can be fitted/interpreted variable-by-variable?

- Compromise: Additive Model

$$Y_i = \alpha_i + \sum_{k=1}^{p} f_k(x_{ik}) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

with f_k's univariate smooth functions, $\sum_i f_k(x_{ik}) = 0$.

- Can extend to Generalised Additive Model:

$$Y_i | x_i^\top \overset{indep}{\sim} \exp\left\{ \alpha_i y + y \sum_{k=1}^{p} f_k(x_{ik}) - \gamma\left(\alpha_i + \sum_{k=1}^{p} f_k(x_{ik}) \right) + S(y) \right\}$$

GLM Nonparametric relationships with x_i

- How to fit additive model? Consider Gaussian case only for simplicity.
  - ↪ Know how to fit each f_k separately quite well
  - ↪ Take advantage of this ...
- Consider ith response:

$$\mathbb{E}\left[Y_i - \alpha - \sum_{m \neq k} f_m(x_{im})\right] = f_k(x_{ik})$$

- Suggests the *Backfitting Algorithm*:

(1) `Initialise:` $\alpha = \bar{Y}$, $f_k = f_k^0$, $k = 1, \ldots, p$.

(2) `Cycle:` Get f_k by 1D smoothing of partial residual scatterplot

$$\left\{\left(Y_i - \alpha - \sum_{m \neq k} f_m(x_{im}), x_{ik}\right)\right\}_{i=1}^{n} = \{e_{ik}, x_{ik}\}_{i=1}^{n}.$$

(3) `Stop:` when individual functions don't change

- Any smoother can be used, usually splines.

GLM Nonparametric relationships with x_i

A different approach is inspired by tomography. Model Gaussian response as:

$$Y_i = \underbrace{\sum_{k=1}^{K} h_k(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)}_{=g(\boldsymbol{x}_i^\top)} + \varepsilon_i, \quad ||\boldsymbol{\beta}_k|| = 1,\ \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

- Also additively decomposes g into smooth functions $h_k : \mathbb{R} \to \mathbb{R}$.
- But each function now depends on a global linear feature $\boldsymbol{x}_i^\top \boldsymbol{\beta}_k$
  ↪ a linear combination of the covariates
  ↪ $||\boldsymbol{\beta}_k|| = 1$ for identifiability.
- Projections directions to be chosen for best fit (nonlinear problem)
- Each h_k is a ridge function of $\boldsymbol{x}_i^\top$: varies only in the direction defined by $\boldsymbol{\beta}_k$

Pros and Cons:

(+) By classical Fourier series, can show that any $C^1([0,1]^p) \to \mathbb{R}$ function is uniformly approximated arbitrarily well as $K \to \infty$. Useful for prediction.

(−) Interpretability? What do terms mean within problem?

GLM Nonparametric relationships with x_i

How is the model fitted to data?

Assume only one term, $K = 1$ and consider penalized likelihood:

$$\min_{h_1 \in C^2[0,1], ||\beta||=1} \left\{ \sum_{i=1}^{n} \{Y_i - h_1(\boldsymbol{x}_i^\top \boldsymbol{\beta})\}^2 + \int_0^1 \{h_1''(t)\}^2 dt \right\}.$$

Two steps:

- *Smooth*: Given a direction $\boldsymbol{\beta}$, fitting $h_1(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ is done via 1D smoothing.
- *Pursue*: Given h_1, have a non-linear regression problem w.r.t. $\boldsymbol{\beta}$.

Hence, iterate between the two steps

↪ Complication is that h_1 not explicitly known, so need numerical derivatives.

↪ Computationally intensive (impractical in the '80's but doable today).

↪ Can separate second step by looking for non-Gaussian projection directions.

Further terms added in forward stepwise manner.

GLM Nonparametric relationships with x_i

If $\boldsymbol{\beta}_k$ needs to be estimated non-linearly anyway...

$$g(\boldsymbol{x}_i^\top) \approx \sum_{k=1}^{K} h_k(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)$$

... do we really need to estimate the h_k or can we fix them?

Theorem (Nonlinear Sigmoidal Approximation)

Let $\Psi : \mathbb{R} \to [0,1]$ be a strictly increasing distribution function and $g : [0,1]^p \to \mathbb{R}$ be an arbitrary continuous function. Then, for any $\epsilon > 0$, there exists $K < \infty$ and vectors $\boldsymbol{\alpha}, \boldsymbol{t} \in \mathbb{R}^K$ and $\{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K\} \subset \mathbb{R}^p$ such that

$$\sup_{x \in [0,1]^d} \left| g(\boldsymbol{x}) - \sum_{k=1}^{K} \alpha_k \Psi(t_k + \boldsymbol{x}^\top \boldsymbol{\beta}_k) \right| < \epsilon.$$

- Can take h_k to be translations of the same known function Ψ!
- The tradeoff is that K may need to be quite large (interpretability?)
- Called a (single layer) neural network by analogy to synaptic function.
- A parametric model with many parameters – fit by nonlinear least squares (gradient descent)

GLM Nonparametric relationships with x_i

What about including transformations of the original covariates?

1. Can of course include J transformations $w_j : \mathbb{R}^p \to \mathbb{R}$

$$(u_1, ..., u_p) \mapsto w_j(u_1, ..., u_n), \qquad j = 1, ..., J,$$

   of the original variables as additional covariates by suitably enlarging the design matrix $\boldsymbol{X}$.

2. We simply adjoin to $\boldsymbol{X}$ another J columns of dimension $n \times 1$ each:

$$\begin{pmatrix} w_j(\boldsymbol{x}_1^\top) \\ \vdots \\ w_j(\boldsymbol{x}_n^\top) \end{pmatrix} \qquad j = 1, ..., J.$$

3. Which functions w_j should we pick though?

Since we've gone nonlinear anyway,

why not attempt to learn which transformations to include from the data?

GLM Nonparametric relationships with x_i

How?

- Instead of including our original covariates (p columns of $\boldsymbol{X}$)...
- ... use q derived covariates (q can be larger than p)

$$\begin{pmatrix} w_1(\boldsymbol{x}_1^\top) \\ \vdots \\ w_1(\boldsymbol{x}_n^\top) \end{pmatrix}, \begin{pmatrix} w_2(\boldsymbol{x}_1^\top) \\ \vdots \\ w_2(\boldsymbol{x}_n^\top) \end{pmatrix}, \quad \cdots \quad, \begin{pmatrix} w_q(\boldsymbol{x}_1^\top) \\ \vdots \\ w_q(\boldsymbol{x}_n^\top) \end{pmatrix}$$

- ... where the q transformations $\{w_j\}_{j=1}^{q}$ are to be estimated from the data.

Recycling our nonlinear approximation theorem, write

$$w_j(\boldsymbol{x}^\top) \approx \sum_{m=1}^{M_j} \delta_{m,j} \Psi(s_{m,j} + \boldsymbol{x}^\top \gamma_{m,j})$$

using the same Ψ, and needing to estimate $(\boldsymbol{\delta}_j, \boldsymbol{s}_j, \gamma_{1,j}, ..., \gamma_{M_j,j})$, for $j = 1, ..., q$.

GLM Nonparametric relationships with x_i

Assuming that we've constructed our new variables, we have a new design matrix

$$\begin{pmatrix} w_1(\boldsymbol{x}_1^\top) & \dots & w_q(\boldsymbol{x}_1^\top) \\ \vdots & & \vdots \\ w_1(\boldsymbol{x}_n^\top) & \dots & w_q(\boldsymbol{x}_n^\top) \end{pmatrix}.$$

Summarising, we have defined a hierarchical nonlinear regression model:

$$Y_i = \sum_{k=1}^{K} \alpha_k \Psi\Big(t_k + (w_i(\boldsymbol{x}_1^\top), ..., w_i(\boldsymbol{x}_n^\top))\boldsymbol{\beta}_k\Big) + \varepsilon_i =$$

$$= \sum_{k=1}^{K} \alpha_k \Psi \left(t_k + \left(\sum_{m=1}^{M_1} \delta_{m,1} \Psi(s_{m,1} + \boldsymbol{x}^\top \gamma_{m,1}), ..., \sum_{l=1}^{M_q} \delta_{l,q} \Psi(s_{l,q} + \boldsymbol{x}^\top \gamma_{l,q}) \right) \boldsymbol{\beta}_k \right) + \varepsilon_i$$

... known these days as a two-layer neural network.

- Can add more layers ("deep neural network").
- Highly non-linear and non-convex – cascade of simple nonlinearities applied to linear transformations.
- More easily perceived visually through a graphical representation

Causal Inference

- If we say "X causes Y"; mathematically this means *changing* the value of x *changes* the distribution of Y.
- When X causes Y then X and Y will be associated (one type of association is correlation), but the converse is generally <u>not</u> true.
- We shall discuss this in terms of <u>counterfactual</u> random variables.
- Let us start by a simple binary setup. Let $X = 1$ denote the event that a unit was "treated" and $X = 0$ denote the event that a unit was not "treated".
- We use the term "treated" in a very broad sense. Instead we might have used "exposed" and "not–exposed".
- Let Y be some <u>outcome variable</u>. To distinguish between association and causation we need to enhance our vocabulary.

Causal Inference II

- Two new symbols C_0 and C_1 are introduced to denote potential outcomes.
- C_0 is the outcome if the unit was not treated, and similarly, C_1 is the outcome if the unit was treated. These are both random variables. Thus
$$Y = C_X. \tag{4}$$
This is the consistency relationship.
- Note that many things are unobserved in this model. When $X = 1$ then we do not observe C_0 for those cases; also when $X = 0$ we do not observe C_1. We call those outcomes counterfactual.
- Thus (C_0, C_1) are hidden or latent variables.

Causal Inference III

- Define the average causal effect to be

$$\theta = \mathbb{E}\{C_1\} - \mathbb{E}\{C_0\}. \tag{5}$$

θ is the difference in effect if everyone was treated versus if everyone was not. If C_0 and C_1 were binary then we can define the causal odds ratio

$$\frac{\frac{\Pr\{C_1=1\}}{\Pr\{C_1=0\}}}{\frac{\Pr\{C_0=1\}}{\Pr\{C_0=0\}}}.$$

- We also define the causal relative risk:

$$\frac{\Pr\{C_1 = 1\}}{\Pr\{C_0 = 1\}}.$$

- Define the association of Y with X to be

$$\alpha = \mathbb{E}\{Y \mid X = 1\} - \mathbb{E}\{Y \mid X = 0\}. \tag{6}$$

Causal Inference III

- Theorem (Association is not causation): In general $\theta \neq \alpha$.
- Example: Suppose that we have observed the following units for a treatment:

Table: Causation vs association.

X	Y	C_0	C_1
0	0	0	0^*
0	0	0	0^*
0	0	0	0^*
0	0	0	0^*
1	1	1^*	1
1	1	1^*	1
1	1	1^*	1
1	1	1^*	1

Asterisks are indicating unobserved values.

Causal Inference IV

- For every experimental unit $C_0 = C_1$ and so the "treatment" has no effect.

$$\theta = \mathbb{E}\{C_1\} - \mathbb{E}\{C_0\} \tag{7}$$
$$= \frac{1}{8}\sum_{i=1}^{8} C_{1i} - \frac{1}{8}\sum_{i=1}^{8} C_{0i} = \frac{1}{8}\sum\{C_{1i} - C_{0i}\} = 0 \tag{8}$$

Thus the average causal effect is zero.

- We can also estimate the association:

$$\alpha = \mathbb{E}\{Y \,|\, X = 1\} - \mathbb{E}\{Y \,|\, X = 0\} = \frac{1+1+1+1}{4} - \frac{0+0+0+0}{4}$$
$$= 1.$$

Thus in this example $\theta \neq \alpha$.