InterpretableML Spring 2022

Felix Hohne and Kevin Jiang

May 17, 2022

Problem: Linear Models are limiting for effective inference

- In Machine Learning, the key task is prediction
- In Inference, task is to understand the processes that generate our data
- The classic solution is Linear Regression and its extensions
- Compared to ML, these models are very under-powered
- But are interpretable, and we can do statistics with them
- Can we do better?

Exponential Families

A distribution is a member of the exponential family if it can be expressed as:

$$P(y;\theta,\phi) = a(y,\theta)\exp\left(\frac{y\theta - k(\theta)}{\phi}\right)$$

where $T(x), h(x), \eta(\theta)$ are known, and ϕ is the dispersion term. Examples include the Binomial, Poisson, Normal, and Gamma distributions.

Lemma

The normal distribution is a member of the exponential family.

$$\begin{aligned} P\left(y;\theta,\sigma^2\right) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{y^2}{2\sigma^2}\right)\exp\left(\frac{y\mu}{\sigma^2} - \frac{\left(\frac{\mu^2}{2}\right)}{\sigma^2}\right) \end{aligned}$$

$$\text{Let } \mathrm{a}(y,\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{y^2}{2\sigma^2}\right), \phi = \sigma^2, \theta = \mu, k(\theta) = \left(\frac{\mu^2}{2}\right)$$

The structure of a Generalized Linear Model (GLM)

The Linear Model assumption:

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_p x_p,\ Y|X \sim \mathcal{N}(X\beta, \sigma^2 I)$$

The limitations are:

- The prediction or the conditional mean is linear in the input features.
- The errors are normally distributed.

Generalized Linear Model: (i) Linear predictor.

$$g(\mathbb{E}(Y|X)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_p x_p$$

(ii) Random component.

$$Y|X \sim \text{ExpFamily}(\theta, \phi)$$

where the *link function* g connects the linear predictor with the random component.

Examples of GLMs

- (Ordinary) linear model: Identity link $g(x) = x$, $Y \sim N(\mu, \sigma^2)$
- Logistic regression: Logit link $g(x) = \log(\frac{x}{1-x})$, $Y \sim \text{Binomial}(\mu)$
- Poisson loglinear regression: Log link $g(x) = \log(x)$, $Y \sim \text{Poisson}(\mu)$
- Gamma regression: Log link $g(x) = \log(x)$, $Y \sim \text{Gamma}(\mu)$, shape parameter $\alpha > 0$ fixed
- And many others!

However, still limited in that our link function is *linear* in the parameters β.

The Generalized Additive Model

Generalized Additive Model:

$$g(\mathbb{E}(Y|X)) = b_0 + f_1(x_1) + f_2(x_2) + \dots f_p(x_p),\ Y|X \sim \text{ExpFamily}(\cdot)$$

where f_i are smooth functions.

Remaining problems:

- How do we choose f_i?
- How do we fit this model?
- How do we deal with interaction terms?

Constructing smooth functions from basis functions

UNIVARIATE SMOOTH FUNCTIONS 121

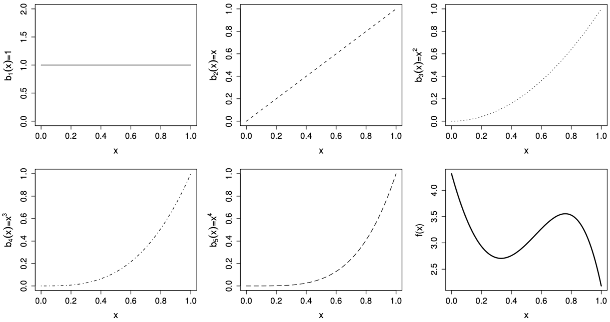


Figure 3.1 *Illustration of the idea of representing a function in terms of basis functions, using a polynomial basis. The first 5 panels (starting from top left), illustrate the 5 basis functions, $b_j(x)$, for a 4th order polynomial basis. The basis functions are each multiplied by a real valued parameter, β_j, and are then summed to give the final curve $f(x)$, an example of which is shown in the bottom right panel. By varying the β_j, we can vary the form of $f(x)$, to produce any polynomial function of order 4 or lower. See also figure 3.2*

LMs, GLMs, and GAMs

The Linear Model assumption:

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_p x_p,\, Y|X \sim \mathcal{N}(X\beta, \sigma^2 I)$$

Generalized Linear Model assumption:

$$g(\mathbb{E}(Y|X)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_p x_p,\, Y|X \sim \text{ExpFamily}(\theta, \phi)$$

Generalized Additive Model:

$$g(\mathbb{E}(Y|X)) = b_0 + f_1(x_1) + f_2(x_2) + \dots f_p(x_p),\, Y|X \sim \text{ExpFamily}(\cdot)$$

where f_i are smooth functions, and *link function* g connects the linear predictor with the random component. The *canonical link* $\theta = g(\mathbb{E}(Y|X)) = \beta X$ is used for easier interpretability and to simplify calculations.

Reducing GAMs to GLMs

To construct f_i:

- Select set of q functions as bases of function space.
- Given fitted $b_1, \dots b_q \in \mathbb{R}$:
- Define $f_i(x) = \sum_{i=1}^{q} b_i \cdot g_i(x)$

Given original data matrix X

- $\tilde{X}_i = [1, g_{1,1}(x_i), g_{1,2}(x_i)..., g_{1,q}(x_i), g_{2,1}(x_i), \dots g_{2,q}(x_i), \dots g_{p,q}(x_i)]$
- $\tilde{\beta} = [b_{1,1}, \dots b_{1,q}, b_{2,1}, \dots b_{2,q}, \dots b_{p,q}]^T$

We can re-write $f_1(x) + f_2(x) + \dots f_p(x) = \tilde{X}\tilde{\beta}$

Our GAM now reduces to a GLM: $g(\mathbb{E}(Y|X)) = \tilde{X}\tilde{\beta}$, $Y|X \sim \text{ExpFamily}(\cdot)$

Optimizing the Log-Likelihood to fit model parameters

We want to optimize the Log-Likelihood, so apply Newton for optimization.

$$x_{k+1} = x_k - (\nabla^2 f(x))^{-1} f'(x_k)$$

In practice, we replace $\nabla^2 f(x)$ with Fisher Information:

$$I(\theta) = -\mathbb{E}(\nabla^2 f(x))$$

Using properties of exponential family and much multivariate calculus:

$$\beta^{k+1} = \beta_k + (X^T W^k X)^{-1} X^T W^k (\widetilde{Y} - \widetilde{\mu})$$

where W^k is a per iteration weight matrix, $\widetilde{Y}, \widetilde{\mu}$ are pseudo-data.

To get the P-IRLS algorithm, add regularization:

iteratively minimize $\|\sqrt{W}(z - Xb)\|^2 + \sum_{i=1}^{i=p} \lambda_i S_i \beta_i \leftrightarrow \| \begin{bmatrix} \sqrt{W} & 0 \\ 0 & I \end{bmatrix} \left(\begin{bmatrix} z \\ 0 \end{bmatrix} \right) - \begin{bmatrix} X \\ \beta \end{bmatrix} \beta \|^2$

For classic linear regression, we use Least Squares:

$$\text{argmin}_{\beta}\|y - X\beta\|^2$$

Also recall Maximum Likelihood Estimation(MLE):

$$l(\theta; y) = \log(\prod_{i=1}^{n} f(y_i; \theta)) = \sum_{i=1}^{n} \log f(y_i; \theta)$$

Lemma

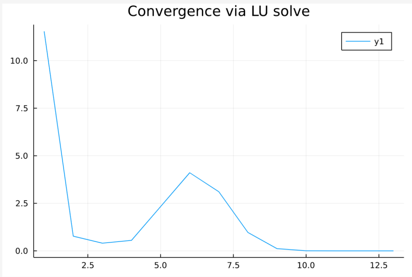
Least Squares is a special case of MLE under normal errors.

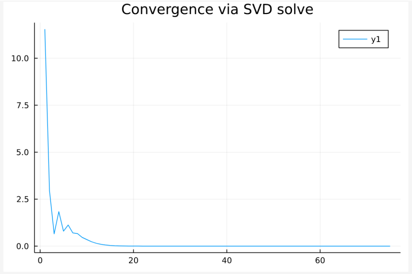
Proof.

$$\begin{aligned}
l(y; X, \beta) &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i - x_i\beta)^2}{2\sigma^2}\right) \\
&= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(\exp(-\frac{(y_i - x_i\beta)^2}{2\sigma^2}\right) \\
&= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_i - x_i\beta)^2}{2\sigma^2} \\
&= \sum_{i=1}^{n} -(y_i - x_i\beta)^2 \\
&= \|y - X\beta\|^2
\end{aligned}$$

Solving IRLS in practice

Solving IRLS inner loop possible via various Matrix Decompositions:





Inference for GLMs

- Large sample theory for MLEs to GLMs yields

$$\hat{\beta} \to \mathcal{N}(\beta, (X^T W X)^{-1})$$

Can be readily obtained from last iteration of IRLS algorithm.

- Delta method for fitted values $\hat{\eta}$ and $\hat{\mu}$

$$\hat{\eta} = X\hat{\beta} \to \mathcal{N}(X\beta, X(X^T W X)^{-1} X^T)$$
$$\mathrm{Var}(\hat{\mu}) \approx D\mathrm{Var}(\hat{\eta}) D^T = DX(X^T W X)^{-1} X^T D$$

- Standard test statistics from MLE theory can be used.

$$\textit{Likelihood ratio test:} \; -2\log\Lambda \to \chi_1^2$$
$$\textit{Wald test:} \; \hat{\beta} - \beta_0 / SE \to \mathcal{N}(0,1)$$
$$\textit{Score test:} \; \frac{[\partial L(\beta)/\partial\beta_0]^2}{-\mathbb{E}[\partial^2 L(\beta)/\partial\beta_0^2]} \to \chi_1^2$$

Inference for GAMs

- By Lindeberg-Feller CLT,

$$v := X^T W z \to \mathcal{N}(X^T W X \beta, X^T W X \phi)$$

where $z = X\beta + G(y - \mu)$ is 'working' response, $G = \mathrm{diag}(g'(\mu_i))$, and W is weight matrix as defined in IRLS.

- Frequentist approach: Recall $\hat{\beta} = (X^T W X + \sum_i \lambda_i S_i)^{-1} v$,

$$\hat{\beta} \to \mathcal{N}(\mathbb{E}[\hat{\beta}], V_f)$$

where V_f is known covariance matrix.

- Bayesian approach: for $\beta \propto \exp\{-\frac{1}{2}\beta^T(\sum_i \lambda_i S_i)\beta\}$ for some λ_i,

$$\beta | v \sim \mathcal{N}(\hat{\beta}, (X^T W X + \sum_i \lambda_i S_i)^{-1} \phi))$$

Fake News at Twitter (Osmundsen et. al., 2021)

Predict number of 'fake-news' articles a user on Twitter will click on based on their particular characteristics.

Features: Gender, ethnicity, income, political party, ideology.

Model via a Poisson loglinear model $Y \sim \mathrm{Poisson}(\lambda), \log \lambda = \beta_0 + \beta^T X$


How Ukraine's 'Ghost of Kyiv' legendary pilot was born

By Laurence Peter
BBC News

1 May

Russia-Ukraine war

Fake news in Twitter

OLS vs GLM Inference

Feature	OLS $\hat{\beta}$	OLS $\hat{\sigma}$	...	GLM $\hat{\beta}$	GLM $\hat{\sigma}$
(Intercept)	-2.614	2.538	·	-3.230	0.145
Female	-0.093	0.968	·	0.143	0.041
Caucasian	0.846	1.102	·	0.773	0.059
Income	0.016	0.021	·	*0.019*	0.0006
Cynicism	-0.731	1.266	·	*0.058*	0.0197
...	...	...	·	...	...
Ideology	1.272	0.387	·	0.8844	0.0194

Table 1: OLS and GLM coefficients and their std. errors

Fake news in Twitter

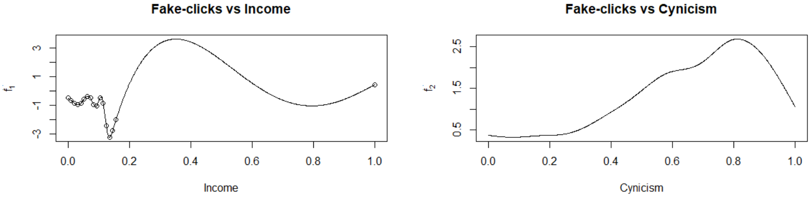


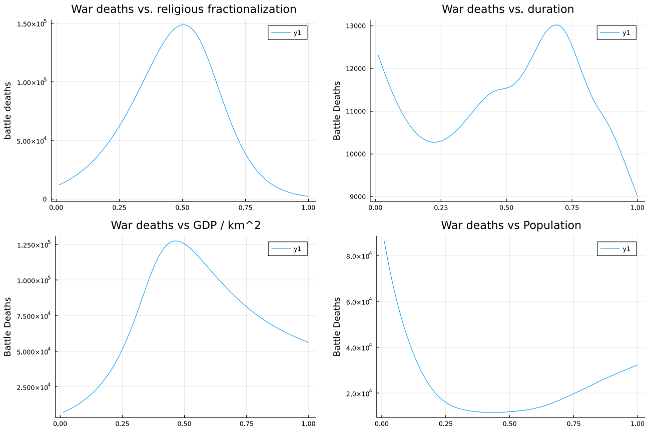
Figure 1: *Note:* fit from income not entirely trustworthy – tail is highly variable due to small sample size for larger income values. Cynicism plot more reliable though.

Predicting Civil War Deaths (Lacina, 2006)

- Contains 104 civil wars since 1945
  - 25th quantile deaths: 3246; 75th quantile deaths: 38825
- Paper uses Linear Regression with log transform
- We model with Gamma distribution due to large number of deaths, right-skew
- Model considered: deaths $\sim$ religious frac. $+$ (gdp/sq km) $+$ pop $+$ duration
- Do not include confidence intervals due to small sample size
  - In particular, GAM confidence intervals were extremely large
  - Resulted in numerical stability problems

Civil Wars death estimation results

Figure 2: GAM feature estimates



Ordinal Logistic Regression

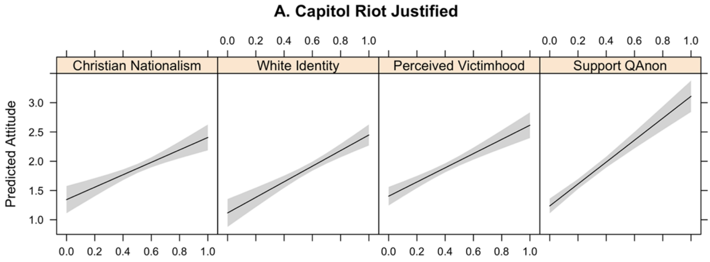
- (Latent-variable approach): A 'latent' or hidden variable Z determines the label Y:

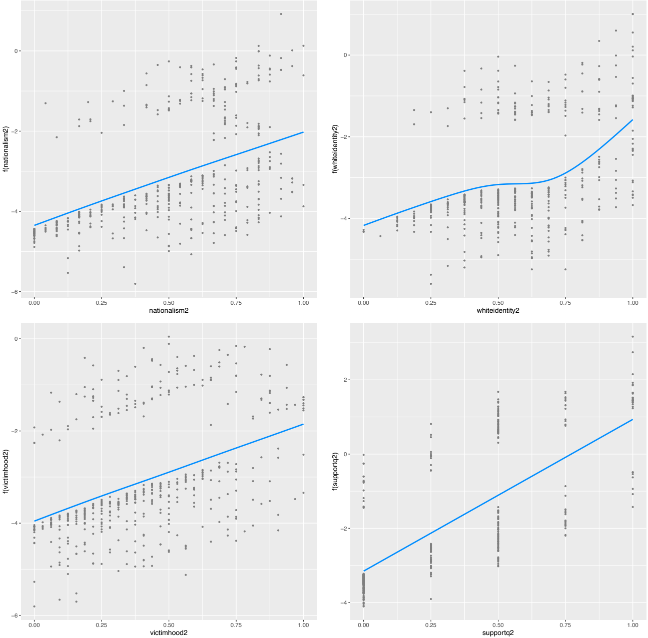
$$Z = \alpha + \beta^T X + \epsilon, \epsilon \perp X, \text{with } \epsilon \sim F(\epsilon)$$

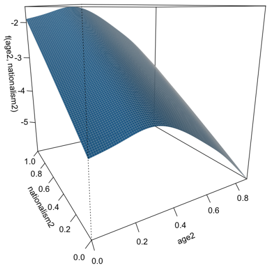
  $Y = j$ if and only if $\delta_{j-1} < Z < \delta_j$ for $j = 1, ..., c$.
- Different chosen $F(\epsilon)$ yields different models - a logistic distribution leads to the widely implemented *Ordinal Logistic Regression* model
- Linearity can be relaxed to smooth functions (e.g. $f_i(X)$) and include interaction between terms (e.g. $X_i X_j$)
- Fit with a more general framework of P-IRLS for smoothing models (implementation *very* involved, see Wood, Pya, & Safken, 2016)

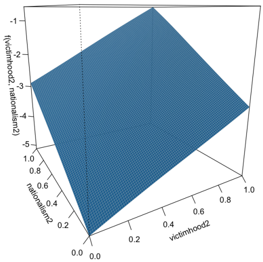
More Ordinal Regression: (Armaly et. al., 2021)

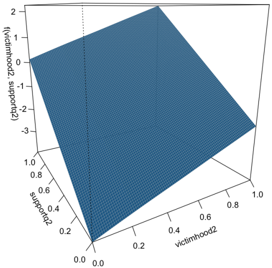
- Analyze public support for political science in US, including Jan 6. insurrection
- Conducted representative survey of 1100 Americans
- Dependent variable; ordinal response from 1 to 5
- Paper uses linear regression for analysis

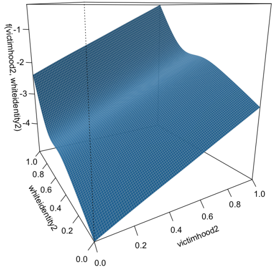



f(nationalism2)
0
-2
-4
-6
0.00
0.25
0.50
0.75
1.00
nationalism2
f(whiteidentity2)
0
-2
-4
0.00
0.25
0.50
0.75
1.00
whiteidentity2
f(victimhood2)
0
-2
-4
-6
0.00
0.25
0.50
0.75
1.00
victimhood2
f(supportq2)
2
0
-2
-4
0.00
0.25
0.50
0.75
1.00
supportq2


f(age2, nationalism2)
-2
-3
-4
-5
1.0
0.8
0.6
0.4
0.2
0.0
nationalism2
0.0
0.2
0.4
0.6
0.8
age2


f(victimhood2, nationalism2)
-1
-2
-3
-4
-5
1.0
0.8
0.6
0.4
0.2
0.0
nationalism2
0.0
0.2
0.4
0.6
0.8
1.0
victimhood2


f(victimhood2, supportq2)
2
1
0
-1
-2
-3
1.0
0.8
0.6
0.4
0.2
0.0
supportq2
0.0
0.2
0.4
0.6
0.8
1.0
victimhood2


f(victimhood2, whiteidentity2)
-1
-2
-3
-4
1.0
0.8
0.6
0.4
0.2
0.0
whiteidentity2
0.0
0.2
0.4
0.6
0.8
1.0
victimhood2

Conclusion: What we have learned

- In many ways, GAMs behave like ML models
  - Very sensitive to hyper-parameters
  - Need much more data than LMs
  - Confidence intervals are less and less based on statistics
- GAMs excellent at visualizing complex relationships
- Excellent for use as diagnostic
- Probably still default to GLMs for most practical applications
- Remaining open questions:
  - GAMs for time series to potentially model underlying seasonal patterns
  - GL*M*Ms and GA*M*Ms account for *mixed* or random/individual effect
  - Improving interpretability for social sciences + pharmaceutical studies