Non-Gaussian responses (week 2)

30 Jan 2023

Table of contents

Non-Gaussian responses 1
Some answers 1
Why not linear? 2
Logistic regression (ESL § 4.4) 2
Log-likelihood 3
Newton step 3
Newton vs IRLS 3
Families 4
Regularized versions 4
proximal gradient descent/Newton 4
`makeX` 4

Non-Gaussian responses

- Why worry about it?
- Isn't least-squares good enough?
- **poll** (polleverywhere)

Some answers

- heteroscedasticity (Gauss-Markov only applies to homog. variance)
- still unbiased but no longer minimum variance

- maybe we shouldn't (e.g. **linear probability model** in econometrics)
  - adjust for heteroscedasticity with **robust/sandwich estimators** etc. (White):

$$\hat{\mathbf{V}} = (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{G} \mathbf{X}) (\mathbf{X}^\top \mathbf{X})^{-1}$$

    where $\mathbf{G} = \mathrm{Diag}(\hat{\varepsilon}_i^2)$ (contrast with $s^2 (\mathbf{X}^\top \mathbf{X})^{-1}$)
- if we have
- if we have **nonlinear** models, MLEs are no longer unbiased

Why not linear?

- actual nonlinear patterns (but can handle these by transformation/basis expansion)
- unrealistic predictions (e.g. probabilities outside of $[0, 1]$
- varying effects (e.g. effect of a 1-unit change in x on probability must differ depending on baseline probability)
- Why not transform? **poll** (polleverywhere)

Logistic regression (ESL § 4.4)

- Worst-case scenario (farthest from Gaussian)
- ESL starts with a *multinomial* model:

$$\log \left(\frac{\Pr(G = i | X = x)}{\Pr(G = K | X = x)} \right) = \beta_{i0} + \beta_i^\top x, \quad i \in 1 \ldots K-1$$

(and so $\Pr(G = K | X = x) = 1 / \left(1 + \sum_{i=1}^{K-1} \exp(\beta_{i0} + \beta_i^\top x) \right)$)

- independent of baseline/reparameterization
- log-likelihood $\sum \log p_i(x_i; \theta)$ where θ is the complete set of parameters

Log-likelihood

- for two categories, log-likelihood simplifies to

$$\begin{aligned} & \sum \left(y_i \beta^\top x_i - \log\left(1 + e^{\beta^\top x_i}\right)\right) \\ = & \sum \left(y_i \eta_i - \log\left(1 + \exp(\eta_i)\right)\right) \end{aligned}$$

- **weight matrix** $\mathbf{W} = \text{Diag}(p(1-p))$
  - more generally, $\text{Diag}(1/\text{Var}(\mu))$
- score equation:
  - $\sum_i = 1^N x_i(y_i - p(x_i; \beta))$
  - Newton update is $\beta^* - \mathbf{H}^{-1}\mathbf{g}$
  - gradient: $\mathbf{X}^\top(\mathbf{y} - \mathbf{p})$
  - generally $\mathbf{X}^\top(\mathbf{y} - \mu) = \mathbf{X}^\top(\mathbf{y} - g^{-1}(\eta))$
- Hessian: $-\mathbf{X}^\top\mathbf{W}\mathbf{X}$
- solution is $(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{z}$
- where $\mathbf{z} = \mathbf{X}\beta_0 + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})$ is the adjusted response

Newton step

- **iteratively reweighted least squares**
- solve

$$\mathbf{X}^\top\mathbf{W}\mathbf{X}beta^* = \mathbf{X}^\top\mathbf{W}$$

$$\beta$$

- 4C03 notes
- 4C03 notes 2

Newton vs IRLS

- Newton vs *Fisher scoring* (expected value of the Hessian); equivalent for the *canonical link*
  (e.g. logistic for binary data, log for Poisson data
- link mostly important for interpretation
- can be disregarded (?) if we are going to handle nonlinearity by basis expansion
- convergence? (Mount 2012)

Mount, John. 2012. "How Robust Is Logistic Regression?" *Win Vector LLC*. https://win-vector.com/2012/08/23/how-robust-is-logistic-regression/.

Families

- Gaussian, Poisson, binomial (binary)
- May need to compute *scale*/*dispersion* parameter
  - for exponential families, calculate as $\sqrt{D/(n-p)}$ where D is the *deviance* (-2 log likelihood, equal to SSQ for Gaussian)
  - not exactly the MLE but good enough
- **over-dispersion**: quasi-likelihood
- more complex familes (negative binomial etc.) have an additional, non-collapsible parameters, need to estimate by MLE (or **profiling**)

Regularized versions

- lasso, ridge, or elasticnet
- score equations: $\mathbf{x}^{|}top(\mathbf{y}-\mathbf{p}) = \lambda \cdot \text{sign}(\beta_j)$ for **active** variables (non-zero coeffs)
- ridge should still be solvable by data expansion

proximal gradient descent/Newton

- simpler strategies (cyclic coordinate descent) may not work as well
- **proximal** gradient descent or **proximal** IRLS:

glmnet family docs

`makeX`