# Generalized linear models

Sofia Olhede

EPFL

November 23, 2020

1. Generalized linear models

2. Measures of Fit

3. Residuals

4. Examples

# GLM

Theorem (Asymptotic Normality of MLE in GLM)

*In the same context and notation as before, assume that:*

Reminder

$\beta$: true parameter value

(C1) $\boldsymbol{\beta} \in B$ *for* $B$ *an open convex subset of* $\mathbb{R}^p$.

(C2) *The* $p \times p$ *matrix* $\boldsymbol{X}_n^\top \boldsymbol{X}_n$ *is of full rank for all* $n$.

(C3) *The information diverges, i.e.* $\lambda_{min}\Big(\underbrace{\mathcal{I}_n(\boldsymbol{\beta})}_{\text{Fisher's information}}\Big) \to \overset{0}{\cancel{\infty}}$ *as* $n \to \infty$ *for* $\lambda_{\min}(\cdot)$ *the smallest eigenvalue.*

(C4) *Given any parameter* $\boldsymbol{\beta} \in \mathbb{R}^p$ *it holds that*

$$\sup_{\boldsymbol{\alpha} \in N_\delta(\boldsymbol{\beta})} \left\| \mathcal{I}_n^{-1/2}(\boldsymbol{\beta}) \mathcal{I}_n^{1/2}(\boldsymbol{\alpha}) - \boldsymbol{I}_{p\times p} \right\| \to 0$$

$$\forall\, \delta > 0, \text{ where } N_\delta(\boldsymbol{\beta}) = \left\{ \boldsymbol{\alpha} \in \mathbb{R}^p : (\boldsymbol{\alpha} - \boldsymbol{\beta})^\top \mathcal{I}_n(\boldsymbol{\beta})(\boldsymbol{\alpha} - \boldsymbol{\beta}) \le \delta \right\}.$$

*Then, as* $n \to \infty$*, provided it exists, the MLE* $\hat{\boldsymbol{\beta}}_n$ *of* $\boldsymbol{\beta}_0$ *is unique & satisfies*

$$\mathcal{I}_n^{1/2}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} \underline{N(0, \boldsymbol{I}_{p\times p})}. \longrightarrow \text{Standard Gaussian}$$

- Recall that under canonical link $\underline{\mathcal{I}_n(\boldsymbol{\beta}) = \boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta}) \boldsymbol{X}_n}$.

# GLM

Comments on conditions:

- (C1) implies that $\boldsymbol{X}_n\boldsymbol{\beta}$ ranges over an open set, and so $\gamma$ is infinitely differentiable, and our exponential family possesses all moments.
- (C2) is as with our linear model, and essentially means that our covariates should not be perfectly correlated.
- (C3) is similar to the "balanced design" assumption we had for the asymptotics of a non-Gaussian linear model.
- (C1-C3) are up to us: they depend on the design matrix $\boldsymbol{X}_n$, which in principle is for us to choose.
- (C4) asks that the (root) information matrix converge uniformly on compact ellipsoids centred at the true parameter.

Comments on conclusion:

- Can also be read as saying that for $n$ sufficiently large,

$$\hat{\boldsymbol{\beta}}_n \overset{d}{\approx} N(\boldsymbol{\beta}, \mathcal{I}_n^{-1}(\boldsymbol{\beta}))$$

- Conclusion also immediately implies that

$$(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^\top \mathcal{I}_n(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \chi^2_p.$$

- Allows adapting testing/CI developed before using LR and Wald statistics.

## How do we get there?

- So for a particular example we recall although a very useful framework, there are scenarios where the linear model does not work.
- The range of $Y_i$ is restricted. We can only take some values which makes the Gaussian assumption unrealistic.
- If incorrectly phrased as Gaussian the higher order moments would need to depend on the mean.
- Sometimes we could transform $Y_i$ instead of using the GLM framework. Say use a Box–Cox transformation:

ie "take the observations to some power, and choose the power so it looks like gaussian. If the power is zero, we take the logs"

$$U_i(\alpha) = \frac{Y_i^{\alpha} - 1}{\alpha}, \quad \alpha > 0,$$

and $U_i(\alpha) = \log(Y_i)$ if $\alpha = 0$. If $Y_i = \sigma W_i$ and $W_i > 0$ we see for example how the log transformation helps. But the GLM is a more complete framework.

# How do we get there?

- We can illustrate the problem by starting from the standard linear model. Say we take for Gaussian $Y_i$

$$Y_i = \mathrm{x}_i^T \boldsymbol{\beta} + \epsilon_i.$$

- If $Y_i$ is Bernoulli obviously this does not work, as when $\mathrm{x}_i^T \boldsymbol{\beta}$ changes outside the range $(0, 1)$ we have a problem.
- We need to de-couple $Y_i$ from $\mathrm{x}_i^T \boldsymbol{\beta}$. Define $\pi_i = \mathbb{E}\, Y_i$ associated with covariate $x_i$.
- One such example is toxicology experiments, where each animal is subjected to a known level $x$ of a stimulant or dose.
- Normally with the probit model one sets

$$\pi_x = \overset{\text{Gaussian CDF}}{\Phi}(\alpha + \beta x).$$

- The argument can here range over the entire real line but we know $\Phi(y)$ ranges between nought and unity.

# How do we get there?

- This model can also be used if $Y_i = Z_i/m_i$ is a proportion, namely the number $Z_i$ surviving out of the original $m_i$ is recorded, together with the dose $x_i$ administered.
- We can also model the proportion using a logit model

$$\mathbb{E}\, Y_i = \pi_i, \quad \log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \mathrm{x}_i^T \boldsymbol{\beta}.$$

- Because of the logit model we no longer need to restrict $\mathrm{x}_i^T \boldsymbol{\beta}$.
- We see how $\pi_i$ ranges from zero to unity makes $g(\pi_i) = \pi_i/(1-\pi_i)$ range over the entire real line.

# Measures of Fit

In Gaussian linear regression we used sums of squares to measure fit and compare nested models. What about in GLM?

Idea: compare best possible to observed maximised loglikelihood

- Let $\ell_n(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{Y} + \sum_{i=1}^n \gamma(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})$ be the maximised loglikelihood.
- Define the saturated model to be that which has

$$\#\text{parameters} = \#\text{observations}$$

  i.e. where we replace $\boldsymbol{X}_n\boldsymbol{\beta}$ by some unconstrained $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)^\top \in \mathbb{R}^n$. (and thus we also replace $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ by $\eta_i$).
- Let $\hat{\boldsymbol{\eta}}$ be the maximiser of $\ell_n(\boldsymbol{\eta}) = \boldsymbol{\eta}^\top \boldsymbol{Y} + \sum_{i=1}^n \gamma(\eta_i)$ w.r.t. $\boldsymbol{\eta}$.
- Define the saturated loglikelihood as $\ell_n(\hat{\boldsymbol{\eta}})$.

# Measures of Fit II

Definition ((Scaled) Deviance)

$$D = 2\Big(\overset{\text{saturated Loglikelihood}}{\ell_n(\hat{\boldsymbol{\eta}})} - \underset{\text{maximized loglikelihood}}{\ell_n(\hat{\boldsymbol{\beta}})}\Big) = 2\left((\hat{\boldsymbol{\eta}} - \boldsymbol{X}_n\hat{\boldsymbol{\beta}})^\top \boldsymbol{Y} + \sum_{i=1}^{n}(\gamma(\hat{\eta}_i) - \gamma(\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}))\right)$$

Comments:

- Always $D \geq 0$ (why?)
- Small $D$ implies a good model fit ($\boldsymbol{X}_n\hat{\boldsymbol{\beta}} \approx \hat{\boldsymbol{\eta}}$).
- Large $D$ implies poor fit.
- In Gaussian case: deviance $\equiv$ residual sum of squares (exercise).
- Can now use deviance differences to mimic sum-of-square ratios and construct a GLM variant of the F-test.

# Measures of Fit III

- Consider the problem of comparing two nested models:
  - Model $A$: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ vary freely — MLE $\hat{\boldsymbol{\beta}}^A$
  - Model $B$: for $q < p$, $(\beta_1, \ldots, \beta_q) \in \mathbb{R}^q$ vary freely, but $\beta_{q+1}, \ldots, \beta_p$ are fixed — hence only $q$ free parameters, with MLE $\hat{\boldsymbol{\beta}}^B$
- Model $B$ is *nested within* model $A$: $B$ can be obtained by restrictions on $A$
  - More generally, could have Model $B$ with $\boldsymbol{\beta}$ constrained to vary in a subspace $\mathcal{V}$ of dimension $q < p$, which we can write as $\boldsymbol{\beta} = \boldsymbol{Q}_{p\times q} \underbrace{\zeta}_{q\times 1}$ for $\mathcal{M}(\boldsymbol{Q}) = \mathcal{V}$.
- Likelihood ratio test statistic for comparing the models is

$$2\big(\ell_n(\hat{\boldsymbol{\beta}}^A) - \ell_n(\hat{\boldsymbol{\beta}}^B)\big) = D_B - D_A,$$

and when model $B$ is correct $D_B - D_A \xrightarrow{d} \chi^2_{p-q}$.

# Residuals

Main idea: use deviance instead of sums-of-squares and use final iterate of IWLS to get hat matrix

- Leverage $h_{jj}$ defined as $j$th diagonal element of

$$\boldsymbol{H} = \boldsymbol{V}^{1/2}(\hat{\boldsymbol{\beta}})\boldsymbol{X}_n(\boldsymbol{X}_n^\top \boldsymbol{V}(\hat{\boldsymbol{\beta}})\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top \boldsymbol{V}^{1/2}(\hat{\boldsymbol{\beta}}),$$

- Cook statistic now becomes the change in deviance

$$2p^{-1}\left\{\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_{-j})\right\},$$

  where $\hat{\boldsymbol{\beta}}_{-j}$ is MLE when $j$th case $(\boldsymbol{x}_j^\top, Y_j)$ is dropped.

- Cook statistic can be approximated by

$$C_j = \frac{h_{jj}}{p(1-h_{jj})} r_{Pj}^2,$$

  where $r_{Pj}$ is standardised Pearson residual (to be defined in next slide).

# Residuals II

- Deviance residual:

$$d_j = \operatorname{sgn}(\hat{\eta}_j - \boldsymbol{x}_i^\top \hat{\beta}_j)\Big[2\{\underbrace{\eta_j\, Y_j + \gamma(\eta_j)}_{\ell_j(\hat{\boldsymbol{\eta}})} - \underbrace{[(\boldsymbol{x}_j^\top \boldsymbol{\beta})\, Y_j + \gamma(\boldsymbol{x}_j^\top \boldsymbol{\beta})]}_{\ell_j(\hat{\boldsymbol{\beta}})}\}\Big]^{1/2},$$

  for which we note that

$$\sum_{j=1}^{n} d_j^2 = D$$

  gives the deviance (in analogy with RSS in Gaussian linear regression).

- Pearson residual:

$$p_j = \frac{Y_i - \gamma'(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})}{\sqrt{\gamma''(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})}} = \frac{Y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \quad \text{so } \boldsymbol{r}_P = \boldsymbol{V}^{-1/2}(\hat{\boldsymbol{\beta}})(\boldsymbol{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}})).$$

- Standardised versions:

$$r_{Dj} = \frac{d_j}{(1 - h_{jj})^{1/2}} \qquad \& \qquad r_{Pj} = \frac{p_j}{(1 - h_{jj})^{1/2}}.$$

# Special examples

We will now consider two fundamental specific GLM families:

1. Logistic Regression for Binary Data (Bernoulli GLM with natural link)
2. Loglinear Regression for Count Data (Poisson GLM with natural link)

These will give us concrete situations to keep in mind, demonstrating concepts already presented in generality.

We will conclude with remarks on the notion of a scale parameter.

# Special examples II

$$\text{Very often have response} \to Y = \begin{cases} 1, & \text{“success”} \\ 0, & \text{“failure”} \end{cases}$$

So $Y$ has a very simple Bernoulli structure:

$$\mathbb{P}[Y = 1] = \pi = 1 - \mathbb{P}[Y = 0], \quad \mathbb{E}\{Y\} = \pi$$

- Regression: need to connect response $Y$ with an explanatory $\boldsymbol{x}$.
- Use GLM. Can postulate that $g(\pi) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ for some link $g$. Why?
- Intuition: Depending on circumstances, can imagine $Y$ arising as

$$Y = \mathbf{1}\{Z > 0\} \implies \mathbb{P}[Y = 1] = \pi = 1 - F_Z(0)$$

i.e. describing the level of a “hidden” variable $Z$:

- Now suppose $Z = \boldsymbol{x}^\top \boldsymbol{\alpha} + \sigma\varepsilon$. Then

$$\pi_i = 1 - F_Z(0) = 1 - F_\varepsilon\left(-\boldsymbol{x}^\top(\sigma^{-1}\boldsymbol{\alpha})\right) = 1 - F_\varepsilon(-\boldsymbol{x}^\top \boldsymbol{\beta})$$

$((\boldsymbol{\alpha}, \sigma)$ unidentifiable, but $\boldsymbol{\beta} = \sigma^{-1}\boldsymbol{\alpha}$ is ok)

# Special examples III

So $g(x) = -F^{-1}(1-x)$ can serve as a link function

$$1-\pi = F(-\boldsymbol{x}^\top\boldsymbol{\beta}) \implies -F^{-1}(1-\pi) = \boldsymbol{x}^\top\boldsymbol{\beta}$$

$$\boxed{\text{Choice of Link} \iff \text{Choice of Error Distribution } F_\varepsilon}$$

| Distribution $F_\varepsilon(u)$ | | Link function $g(\pi)$ | |
|---|---|---|---|
| Logistic | $e^u/(1+e^u)$ | Logit | $\log\{\pi/(1-\pi)\}$ |
| Normal | $\Phi(u)$ | Probit | $\Phi^{-1}(\pi)$ |
| Log Weibull | $1-\exp(-\exp(u)\}$ | Log-log | $-\log\{-\log(\pi)\}$ |
| Gumbel | $\exp\{-\exp(-u)\}$ | Complementary log-log | $\log\{-\log(1-\pi)\}$ |

- Logit and probit symmetric, hard to distinguish in practice
- Log-log and complementary log-log are asymmetric
- Logit (canonical link) is usual choice, with nice interpretation