Week 4: Course Organization

MATH-516 Applied Statistics

Tomas Masak

March 13th 2023

Section 1

Exponenial Family

Exponential Family

Definition. The distribution of Y is of exponential type if its density can be written as

$$f(y, \theta, \varphi) = \exp\left(\frac{y\theta - b(\theta)}{\varphi} + c(y, \varphi)\right)$$

where $\theta \in \mathbb{R}$ is the canonical parameter $\varphi \in (0, \infty)$ is the dispersion parameter, and b, c are real functions.

If $b \in C^2$, it can be shown using the moment generating function $m(t) = \mathbb{E}e^{tX}$ that

- $\mu := \mathbb{E}Y = b'(\theta)$
- $\mathrm{var}(Y) = \varphi b''(\theta)$
- $\mathrm{var}(Y) = \varphi V(\mu)$, where V is called variance function

Gaussian Distribution

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{for} \quad x, \mu \in \mathbb{R} \text{ and } \sigma^2 \in (0, \infty)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{x^2}{2\sigma^2}\right)$$

$$= \exp\left(\frac{x\mu - \mu^2/2}{\sigma^2} + \left[\frac{x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right]\right)$$

hence

- $b(\theta) = \mu^2/2$ and $c(x, \sigma^2) = \frac{x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$ with $\theta = \mu$ and $\varphi = \sigma^2$
- $\mathrm{var}(Y) = \varphi \cdot 1 \Rightarrow V(\mu) = 1$ (variance does not depend on expectation)

Bernoulli Distribution

$$
\begin{aligned}
f(x,p) &= p^x(1-p)^x \quad \text{for} \quad x \in \{0,1\} \text{ and } p \in (0,1) \\
&= \exp\left(x \log p + (1-x)\log(1-p)\right) \\
&= \exp\left(x \log \frac{p}{1-p} + \log(1-p)\right)
\end{aligned}
$$

hence

- $\theta = \log \frac{p}{1-p}$, $\varphi = 1$, $b(\theta) = -\log(1-p)$, and $c(x,\varphi) = 0$
- $\operatorname{var}(Y) = p(1-p)$ and $\mu = \mathbb{E}X = p \Rightarrow V(\mu) = \mu(1-\mu)$

MLE in Exponential Family

Firstly, assuming we know φ:

- $\ell(\theta) \propto \sum_n \exp\left(\frac{Y_n\theta - b(\theta)}{\varphi}\right)$
- $U_n(\theta) := \frac{\partial}{\partial\theta} \log f(Y_n, \theta, \varphi) = \frac{Y_n - b'(\theta)}{\varphi}$
- MLE: $\bar{U}_N(\theta) := \frac{1}{N}\sum_n U(Y_n, \theta) \stackrel{!}{=} 0 \Leftrightarrow \widehat{\theta} = (b')^{-1}(\bar{Y}_N)$ ... the MLE
- Fisher information: $\bar{I}_N(\theta) = \frac{1}{N}\sum_n -\frac{\partial}{\partial\theta} U(Y_n, \theta) = \frac{b''(\theta)}{\varphi} = I(\theta)$
  - since $\mathrm{var}(Y) = \varphi b''(\theta)$, $I(\theta) > 0$ and the MLE is unique

MLE theory $\Rightarrow \sqrt{N}(\widehat{\theta} - \theta) \to \mathcal{N}(0, I^{-1}(\theta))$

- if φ is unknown, nothing changes because the Fisher cross-covariance between θ and φ is 0
- we need some $\widehat{\varphi}$ for uncertainty quantification and MLE can be poorly behaved ... estimate it via method of moments from $\widetilde{E}_n = \frac{Y_n - b'(\widehat{\theta})}{\sqrt{b''(\widehat{\theta})}}$

Section 2

GLMs

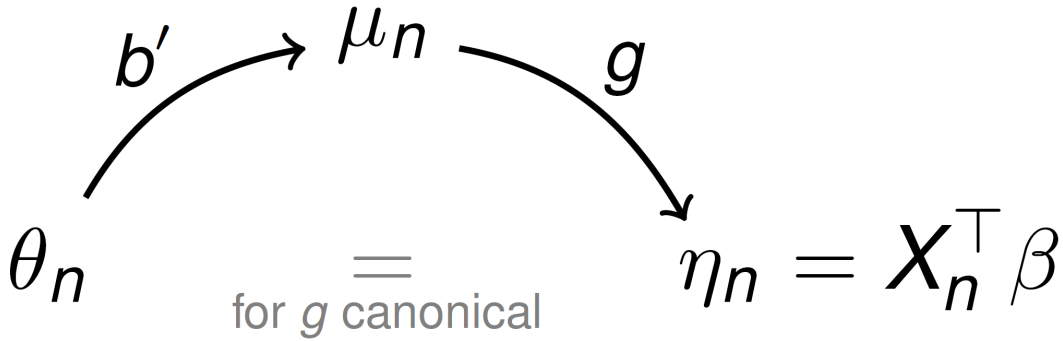
Definition of GLM

Definition. Data $(Y_1, X_1)^\top, \ldots, (Y_N, X_N)^\top$, where $Y_1, \ldots, Y_N$ are independent given $\mathbf{X} \in \mathbb{R}^{N \times p}$ with rows $X_1, \ldots, X_N$, satisfy a GLM if

- the density $f_{Y_n|X_n}(y, \theta_n, \varphi)$ is of exponential type with only θ_n depending on X_n,
- θ_n depends on X_n and β via the linear predictor $\eta_n = X_n^\top \beta$, and
- there exists $g \in C^2$ strictly monotone such that $g(\mu_n) = \eta_n$, where $\mu_n = \mathbb{E}[Y_n \mid X_n]$.

- g is called the link function, and it is canonical if $\eta_n = \theta_n$
  - Gaussian distribution has $g(x) = x$, i.e. the idenitty as the canonical link
  - Bernoulli distribution has the logistic link as the canonical link
  - Poisson distribution has the logarithmic link as the canonical link
  - canonical link is unique up to proportionality

Graph of GLM



- g is the link function
- b is from the canonical form of the exponential density

Note: Exponential densities are log-concave (i.e. log-likelihood is concave in θ) $\Rightarrow$ for a canonical link the log-likelihood of GLM (i.e. for β) is also concave.

Gaussian Distribution

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{for} \quad x, \mu \in \mathbb{R} \text{ and } \sigma^2 \in (0, \infty)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{x^2}{2\sigma^2}\right)$$

$$= \exp\left(\frac{x\mu - \mu^2/2}{\sigma^2} + \left[\frac{x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right]\right)$$

hence

- $b(\theta) = \mu^2/2$ and $c(x, \sigma^2) = \frac{x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$ with $\theta = \mu$ and $\varphi = \sigma^2$
- $\mathrm{var}(Y) = \varphi \cdot 1 \Rightarrow V(\mu) = 1$ (variance does not depend on expectation)
- canonical link must satisfy $g(\mu) = \theta = \mu$ and hence it is the identity

Bernoulli Distribution

$$
\begin{aligned}
f(x,p) &= p^x(1-p)^x \quad \text{for} \quad x \in \{0,1\} \text{ and } p \in (0,1) \\
&= \exp\left(x \log p + (1-x)\log(1-p)\right) \\
&= \exp\left(x \log \frac{p}{1-p} + \log(1-p)\right)
\end{aligned}
$$

hence

- $\theta = \log \frac{p}{1-p}$, $\varphi = 1$, $b(\theta) = -\log(1-p)$, and $c(x,\varphi) = 0$
- $\mathrm{var}(Y) = p(1-p)$ and $\mu = \mathbb{E}X = p \Rightarrow V(\mu) = \mu(1-\mu)$
- canonical link must satisfy $g(\mu) = \theta = \log \frac{\mu}{1-\mu}$

MLE in GLM

The marginal density of X can be dropped, so similarly to above:

- $\ell(\beta) \propto \sum_n \exp\left(\frac{Y_n\theta - b(\theta)}{\varphi}\right)$, where
  - $\theta_n = (b')^{-1}(\mu_n)$ and $\mu_n = g^{-1}(X_n^\top \beta)$
- $U_n(\beta) := \frac{1}{\varphi} w(\mu_i) g'(\mu_i)(Y_n - \mu_n) X_n$, where $w(\mu_n) = \frac{1}{V(\mu_n)[g'(\mu_n)]^2}$
  - shown using the chain and inverse function rules
- MLE: $\bar{U}_N(\theta) := \frac{1}{N}\sum_n U(Y_n, \theta) \stackrel{!}{=} 0$
  - does not have an analytic solution, solved using IRLS
- Fisher information: $I(\beta) = \frac{1}{\varphi}\mathbb{E} w(\mu_i) X_n X_n^\top$
  - can be shown >0 in case of a canonical link, then log-likelihood is concave and IRLS converges to the MLE

MLE in GLM

MLE theory $\Rightarrow$

1. $\sqrt{N}(\widehat{\beta} - \beta) \to \mathcal{N}_p(0, I^{-1}(\theta))$ [Wald]
2. $\sqrt{N}\bar{U}_N(\beta) \to \mathcal{N}_p(0, I(\theta))$ [Rao]
3. $2[\ell(\widehat{\beta}) - \ell(\beta)] \to \chi^2_p$ [likelihood ratio]
    - φ is hidden in all three ... estimate it consistently and use Cramer-Slutzsky

MLE theory with nuissance parameters $\Rightarrow$ all three can be used for model-submodel testing, but [likelihood ratio] is preferred:

Theorem. Let $H_0 : \beta_{p-m+1} = \ldots = \beta_p = 0$ hold in the GLM, $\widehat{\beta}$ denotes parameter estimates in the model, and $\widetilde{\beta}$ denotes parameter estimates in the submodel given by the linear constraints in H_0. Then

$$2[\ell(\widetilde{\beta}) - \ell(\widehat{\beta})] \to \chi^2_m$$

Notes

- all three statistics on the previous slide are asymptotically equivalent, but may differ in finite samples
  - [Wald] is considered the worst since it has been demonstrated in many simulation studies that the convergence is slowest
- [Wald] can be used to obtain CIs easily
  - but inverting [likelihood ratio] acceptance region provides better interval
  - [Wald]'s p-values (for $m = 1$) are provided by `summary()` in `R`
- [Rao] is somewhat simplest
  - the result above is just a CLT for i.i.d. variables $U_n(\beta)$
  - when used for model-submodel testing, nothing from the bigger model needs to be calculated

GLMs again

GLMs have 3 components:

- linear predictor: $\eta_n := X_n^\top \beta$
- link function : g
- response distribution from the exponential family
  - has some canonical parameter θ_n, which in not necessarily equal to the mean μ_n (every exp. family distribution has its bijective relationship between θ_n and μ_n)
  - θ_n and μ_n depend on η_n (that's why the subscripts)

The link function ties (or *links*) the mean μ_n to the linear predictor η_n (and back, since it is bijective).

- it introduces non-linearity
- if it is such that $\theta_n = \eta_n$, the link is called canonical
- wrong link can be a problem for prediction, not so so much for inference
- with a non-canonical link, the log-likelihood may not be concave
  - the IRLS may converge slowly and not to the true MLE

LM vs. GLM

Compared to linear models, GLMs are a bit harder:

- estimates cannot be expressed analytically (IRLS instead of LS)
- testing (and CIs) only asymptotically valid
  - theory based on maximum likelihood
- residual plots are not *that* useful
  - also no FWL theorem

But modelling is quite similar, except the link and distribution must also be chosen.

- but distribution is often clear and link is often taken as the canonical one

Diagnostic Tools

Using the following theorem (which is also the basis for IRLS), GLM diagnostics is adopted from linear models:

Theorem. The MLE $\widehat{\beta}$ in the GLM solves the system of equations

$$\mathbf{X}^\top \widehat{\mathbf{W}} \mathbf{X} \widehat{\beta} = \mathbf{X}^\top \widehat{\mathbf{W}} \widehat{Z},$$

where $\widehat{\mathbf{W}} = \mathrm{diag}(w(\widehat{\mu}_1), \ldots, w(\widehat{\mu}_N))$ and $\widehat{Z} \in \mathbb{R}^N$ has entires $\widehat{Z}_n = \widehat{\eta}_n + (Y_n - \widehat{\mu}_n) g'(\widehat{\mu}_n)$.

- it is a non-linear system of equations, because everything with a hat depends on $\widehat{\beta}$
  - solving it using a fixed point method (fixing in every iteration everything but $\widehat{\beta}$ on the LHS) is exactly what IRLS does and how the model is fitted

Diagnostic Tools

In this fixed point perspective, denoting

- $\widetilde{\mathbf{X}} := \widehat{\mathbf{W}}^{-1/2}\mathbf{X}$
- $\widetilde{Y} := \widehat{\mathbf{W}}^{-1/2}\widehat{Z}$

we have:

- $\widehat{\beta} = (\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}(\widetilde{\mathbf{X}}^\top\widetilde{Y})$
- $\widetilde{H} = \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top$
  - $\widetilde{h}_{nn}$ are leverages
- $\widehat{\widetilde{Y}} = \widetilde{\mathbf{X}}\widehat{\beta}$
- $\widetilde{E}_n := \widetilde{Y} - \widehat{\widetilde{Y}}$ are Pearson residuals
  - the default in R are deviance residuals
  - many other kinds of residuals available for GLMs, e.g. from the definition of $\widehat{Z}_n$ we read the working residuals $(Y_n - \widehat{\mu}_n)g'(\widehat{\mu}_n)$
- Cook's statistic: $C_n = \frac{\widetilde{E}_n^2 h_{nn}}{p(1-h_{nn})}$
- plot η_n against the working residuals to check the link function (plot should be linear)

Deviance

Definition.

1. The null model is the intercept-only model.
2. The saturated model is a model with the largest possible amount of parameters (i.e. $p = N$ if at least one regressor is continuous).
3. The statistic $D(Y, \widehat{\beta}) = 2\varphi[\widehat{\ell}(Y) - \ell(\widehat{\beta})]$, where $\widehat{\ell}(Y)$ denotes the maximized log-likelihood of the saturated model, is called the deviance.

- it is a goodness-of-fit measure
  - for linear model, it is equal to the residual sum of squares R^2
- lower the better
  - when adding a useless regressor we expect the deviance to reduce by 1, which is why likelihood-based model selection criteria (such as AIC) penalize for number of parameters
- model `summary()` in `R` provides:
  - null deviance: deviance of the intercept-only model ($N - 1$ df)
  - residual deviance: deviance of the provided model ($N - p$ df)

- historical remark: deviance is proportional to likelihood minus a constant: likelihood ratio test is sometimes called deviance test, AIC is sometimes calculated from the deviance, etc.

Residuals

- the Pearson statistic $\chi^2 = \sum_n \widetilde{E}_n^2/(N-p)$ is the moment estimator of φ ($\widetilde{E}_n$ are Pearson residuals)
  - notice the analogy with linear models - the Pearson statistic is used as a rough assessment of the model fit
  - if we know that $\varphi = 1$ (e.g. logistic or Poisson regression), this rough assessment is quite useful
- in a similar spirit, the deviance residuals D_n are defined such that the deviance satisfies $D(Y, \widehat{\beta}) = \sum_n D_n^2$
- in GLMs, residuals are used similarly as in linear models, but lack of fit should not be overinterpreted
  - e.g. when deviance residuals are approximately normal, this indicates a good model fit, but if they are not, it does not necessarily indicate problems (unless in some special cases with many observations, where we would expect the CLT to kick in)