Properties of Exp. families

— need this to come up with a fitting scheme.

① Log-likelihood $\ell(\theta, \varphi | y) = \sum_i \left(\frac{y_i \theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi) \right)$

② Score function $\ell_i'(\theta) = U[\theta_i] = \frac{\partial \ell}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\varphi)}$

Since $\ell(\theta) = \log f_y(\theta)$ ⇒ chain rule

$$\ell'(\theta) = \frac{f_y'(\theta)}{f_y(\theta)}$$

$$\Rightarrow E[\ell'(\theta)] = \int f_y(\theta) \frac{f_y'(\theta)}{f_y(\theta)} dy$$

"nice" $= \frac{d}{d\theta} \int f_y(\theta) dy = \frac{d}{d\theta} 1 = 0$

since f is a density

So $E(\text{Score}) = 0 \Rightarrow E\left(\frac{y_i - b'(\theta_i)}{a(\varphi)} \right) = 0$

$\Rightarrow E(y_i) = b'(\theta_i)$

(From 2nd derivative) $\Rightarrow V(y_i) = a(\varphi) b''(\theta_i)$

This $E \sim V$ relationship is assumed known and follows from choice of family as Poisson or Binomial

So — how to fit GLMs to data?

Two problems ⟶ $E(y_i) = \mu_i = g^{-1}(\theta_i) = g^{-1}(x_i'\beta)$ (A)

is often <u>non-linear</u> in β

Since $V(y_i)$ depends on $E(y_i)$ through $b''(\theta_i)$ we have non-constant error variance (B)

(A) → ILS like we did for nonlinear regression

(B) → <u>weighted</u> Least Squares (WLS)

(A)+(B) called

Iterative re-weighted Least Squares

or

IWLS

WLS

Let's say $V(y_i) = \sigma_i^2$ not constant σ^2

$\Rightarrow V(y) = \sigma^2 V$ varying components, y non-diagonal — y's are correlated as well.

$\Rightarrow$ Use LS $\Rightarrow$ $\hat{\beta} = (X'X)^{-1}X'y$

$E(\hat{\beta}) = \beta$, $V(\hat{\beta}) = \sigma^2 (X'X)^{-1}(X'VX)(X'X)^{-1}$

WLS

$$\min_{\beta} \sum_{i=1}^{n} w_i (y_i - x_i'\beta)^2 = \sum_{i=1}^{n} (w_i^{1/2}(y_i - x_i'\beta))^2 = \| W^{1/2} y - W^{1/2} X\beta \|^2$$

$$= \| \tilde{y} - \tilde{X}\beta \|^2$$

transformed variables

$$\begin{cases} \tilde{y} = W^{1/2} y \\ \tilde{X} = W^{1/2} X \end{cases}$$

$\Rightarrow \hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y}$, $\hat{y} = Hy = \tilde{X}'\hat{\beta}$

$= (X'WX)^{-1}X'Wy$ $= \underbrace{\left(W^{1/2} X (X'WX)^{-1} X' W^{1/2} \right)}_{H} y$

Now if $V(y) = \sigma^2 V$ and we use $W = V^{-1}$

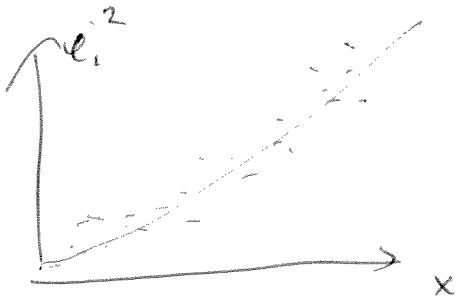
$\Rightarrow V(\hat{\beta}) = \sigma^2 (X'WX)^{-1}$ best possible

That is, use weights $w_i \sim \frac{1}{V(y_i)}$

What if don't know $V(y_i)$

Can learn from data

→ get first regression fit w/o weight



regression of e_i^2 on e.g. x

$\Rightarrow V(y_i) = x_i'\gamma$

- For GLMs

$$V(y_i) = a(\phi)\, b''(\theta_i)$$

depends on unknown mean parameters θ_i

So what we will do is — in each iterative step use previous values of $\hat{\theta}_i \Rightarrow \hat{V}(y_i) \Rightarrow w_i = \frac{1}{\hat{V}(y_i)}$.

Step 2: IWLS

(0) Remember $E(y_i) = \mu_i$, $g(\mu_i) = \eta_i = x_i'\beta$

$$g(\mu) \approx g(\mu_0) + (\mu - \mu_0) \left.\frac{\partial g(\mu)}{\partial \mu}\right|_{\mu_0}$$

linearize at value μ_0 (or $\eta_0 = x_i \hat{\beta}_0$)

$g(\mu_0) = \eta_0$; $\left.\frac{\partial g(\mu)}{\partial \mu}\right|_{\mu_0} = \left.\frac{\partial \eta}{\partial \mu}\right|_{\mu_0}$

(1) Initial values $\beta^0 \rightarrow \eta_i^0 = x_i'\beta^0 \rightarrow \mu_0 = g^{-1}(\eta_0)$

(2) "Working values" $Z_0 = \eta^0 + (y - \mu_0)\left.\frac{\partial \eta}{\partial \mu}\right|_0$

" linear approx of data near η^0 "

"Working weights" $W_0^{-1} = \left(\left.\frac{\partial \eta}{\partial \mu}\right|_{\mu_0}\right)^2 V_0$

$b''(\theta)\big|_{\mu_0} = \left.\frac{\partial \mu}{\partial \theta}\right|_0$

known variance for model family e.g. μ_i for Poisson, $\pi_i(1-\pi_i)$ for Binomial

(3) Regress Z_0 on X

$\rightarrow \hat{\beta}^{new} \rightarrow \hat{\eta}^{new} \rightarrow$ Go to (1).

Validation & Inference

LM & NLM $\Rightarrow$ RSS $\Rightarrow$ F-tests, R^2 etc

Here, everything is based on the Likelihood

Residual Deviance of a model M = $-2 * \log\text{likelihood}(y, \hat{\theta})$

plays the role of ~~the~~ RSS of a model in LM and NLM

- denote by $D(M)$

- $\frac{D(M)}{a(\phi)}$ is called the scaled deviance.

If normal error case

$$\frac{D(M)}{a(\phi)} = \frac{RSS(M)}{\sigma^2} \sim^d \chi^2_{n-p}$$

- Here, that is only true for very large data sets (asymptotically)

- Large n $\quad \boxed{\frac{D(M)}{a(\phi)} \sim^d \chi^2_{n-p}}$

Instead of an F-test for goodness-of-fit, we use a χ^2-test. If $\frac{D(M)}{a(\phi)} >> \chi^2_{n-p}(1-\alpha)$ we reject the fit.

$\rightarrow$ Try transformations, other links, other model families

We can also use the deviance to check if our model is appropriate in terms of scale

$$\hat{a}(\varphi) = \frac{D(M)}{n-p}$$

if $\hat{a}(\varphi) >>$ assumed $a(\varphi)$ (e.g. 1 for Poisson and Binomial)

$\Rightarrow$ we call the data <u>over-dispersed</u> $\rightarrow$ other model family needed.

<u>Model selection</u>

Simple model M_0, Complex model M – which to choose

e.g. x_1, x_2, x_3 — p_0 parameters; x_1, x_2, x_3, x_4, x_5 — p parameters

(A) if φ known $\rightarrow$ χ^2-test <u>Analysis of Deviance</u>

$$\frac{D(M_0) - D(M)}{\varphi} \sim \chi^2_{p-p_0} \quad \text{if } M_0 \text{ sufficient}$$

if $\frac{D(M_0) - D(M)}{\varphi} >> \chi^2_{p-p_0}(1-\alpha)$ reject M_0.

(B) φ not known → estimate it as above $\hat{\varphi} = \frac{D(M)}{n-p}$

→ use an approximate F-test

$$\frac{D(M_0) - D(M)}{\hat{\varphi}\,(p - p_0)} \sim F_{p-p_0,\, n-p} \quad \text{if } M_0 \text{ sufficient.}$$

- (This is pretty ad-hoc, but works ok in practice)

(C) Can always use CV, bootstrap, BIC etc

Diagnostics

- (can be tricky because the residuals look kind of weird.

- → Make sure that residuals on average ≈ 0 as a function of $\hat{y}$, x's.

Just like NLM → LTs based on last linear approximation

— now use Z-intervals (based on large n)

— check profile function and if not linear → bootstrap.

Common problems to be aware of

Over-dispersion

$\frac{D(M)}{n-p} >> \varphi$ assumed

↓

- Inflate all SE's by $\hat{\varphi}$
- Bootstrap
- Other models

Binomial data w. clear 0/1 separation

→ can get really weird $\hat{\beta}$'s → $\pm\infty$

→ can trust $\hat{\pi}_i$ but not $\hat{\beta}$ values.

Like in LM, be aware of high leverage / influential observations.