Advanced Regression Methods for Independent Data

STAT/BIOST 570, 2020

Generalized Linear Models

Mauricio Sadinle

Department of Biostatistics

UNIVERSITY of WASHINGTON

Our Data Structure

- Y_i: response variable for unit i
- $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ik})$: row vector of covariates for unit i, $x_{i0} = 1$
- We observe n independent pairs $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^n$
- We organize the data as

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

Generalized Linear Models

Motivation:

- In a normal linear model we assume $Y \mid \boldsymbol{x} \sim N(\boldsymbol{x}\boldsymbol{\beta}, \sigma^2)$
- Without restrictions on $\boldsymbol{\beta}$ or on $\boldsymbol{x}$, we have that $\boldsymbol{x}\boldsymbol{\beta} \in \mathbb{R}$, which is what we want if $Y \in \mathbb{R}$ is unrestricted
- But what if $Y \in \{0, 1\}$, $Y \in \{0, 1, 2, \dots\}$, or $Y \in \mathbb{R}^+$?
- *Generalized linear models* (GLMs) use known parametric distributions for Y that are more appropriate for each of these cases

GLMs were introduced by Nelder & Wedderburn (1972) as a way of unifying several existing types of parametric regression

Generalized Linear Models

The main characteristics of a *generalized linear model* (GLM) are:

- Response follows a distribution H in the *exponential family*:

$$Y_i \mid \boldsymbol{x}_i \sim H[\,\mu_i, \alpha_i],$$

where μ_i is the mean for observation i, and α_i represents other parameters of the distribution H

- Dependence on covariates only through the mean:

$$\mu_i := \mu(\boldsymbol{x}_i), \quad \alpha_i = \alpha/\phi_i \text{ for known } \phi_i$$

- Linearity of the regression function on a transformed scale:

$$g(\mu_i) = g[E(Y_i \mid \boldsymbol{x}_i)] = g[\mu(\boldsymbol{x}_i)] = \boldsymbol{x}_i\boldsymbol{\beta},$$

where *linearity* refers to $g[\mu(\boldsymbol{x}_i)] = g[\mu(\boldsymbol{x}_i; \boldsymbol{\beta})]$ being linear as a function of $\boldsymbol{\beta}$, for some invertible function $g(\cdot)$

Exponential Family

- $Y_i \mid \boldsymbol{x}_i$ is assumed to follow a member of the *exponential family*, so that the density is of the form

$$f(y_i \mid \theta_i, \alpha_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\alpha_i} + c(y_i, \alpha_i)\right),$$

  for functions $b(\cdot)$ and $c(\cdot, \cdot)$, where θ_i and α_i are scalars.

- This is often called the *exponential dispersion family*
  - θ_i: *natural parameter*
  - $\alpha_i > 0$: *dispersion parameter*; $\alpha_i = \alpha$ or $\alpha_i = \alpha/\phi_i$ for known ϕ_i
  - If $\alpha_i = 1$ and $c(y_i, \alpha_i) = c(y_i)$ we obtain the *natural exponential family*

Exponential Family

It is easy to show that:

- $$\mathsf{E}[Y_i \mid \theta_i, \alpha_i] = b'(\theta_i) := \mu_i,$$

  which means that

  $$\theta_i = (b')^{-1}(\mu_i)$$

- We can parameterize the exponential family in terms of μ_i:

$$f(y_i \mid \mu_i, \alpha_i) = \exp\left(\frac{y_i[(b')^{-1}(\mu_i)] - b[(b')^{-1}(\mu_i)]}{\alpha_i} + c(y_i, \alpha_i)\right)$$

Exponential Family

Also,

- $$\text{var}(Y_i \mid \theta_i, \alpha_i) = \alpha_i b''(\theta_i) := \alpha_i V(\mu_i),$$

  where $V(\cdot) = b''[(b')^{-1}(\cdot)]$

- Note that V itself is not the variance, but it captures the dependence of the variance on the mean

Exponential Family

Examples of members of the exponential dispersion family:

	$\mathsf{N}(\mu, \sigma^2)$	$\mathsf{Poisson}(\mu)$	$\mathsf{Bern}(\mu)$	$\mathsf{Ga}(1/\alpha, 1/[\mu\alpha])$
θ	μ	$\log(\mu)$	$\log(\frac{\mu}{1-\mu})$	$-1/\mu$
α	σ^2	1	1	α
$b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\log[1+\exp(\theta)]$	$-\log(-\theta)$
$c(y, \alpha)$	$\frac{y^2}{2\alpha} + \log(2\pi\alpha)/2$	$-\log(y!)$	0	$\frac{\log(y/\alpha)}{\alpha} - \log[y\Gamma(1/\alpha)]$
Mean $\mathsf{E}[Y \mid \theta]$	θ	$\exp(\theta)$	$\frac{\exp(\theta)}{1+\exp(\theta)}$	$-1/\theta$
Var. $\alpha V(\mu)$	$\alpha \times 1$	$1 \times \mu$	$1 \times \mu(1-\mu)$	$\alpha \times \mu^2$

Link Functions

- A *link function* $g(\cdot)$ provides the connection between the mean function $\mu_i = \mathrm{E}[Y_i \mid \theta_i, \alpha_i]$ and the *linear predictor* $\boldsymbol{x}_i\boldsymbol{\beta}$, via

$$g(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta},$$

  where

  - $\boldsymbol{x}_i$ is a $1 \times (k+1)$ vector of explanatory variables (including a 1 for the intercept)
  - $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_k]^{\mathrm{T}}$ is a $(k+1) \times 1$ vector of regression parameters.

- g^{-1} maps the linear predictor to the mean

$$\mu_i = g^{-1}(\boldsymbol{x}_i\boldsymbol{\beta})$$

- $g(\cdot)$ needs to be monotonic and differentiable

Link Functions

Note that now we have:

- The natural parameter of the exponential family can be written in terms of the mean as
$$\theta_i = (b')^{-1}(\mu_i)$$

- On the other hand, the mean depends on covariates as
$$\mu_i = g^{-1}(\boldsymbol{x}_i\boldsymbol{\beta})$$

- This leads to a dependence of the natural parameter on covariates as
$$\theta_i = (b')^{-1}[g^{-1}(\boldsymbol{x}_i\boldsymbol{\beta})]$$

Link Functions

$$\theta_i = (b')^{-1}[g^{-1}(\boldsymbol{x}_i\boldsymbol{\beta})]$$

- If we take $g(\cdot) = (b')^{-1}(\cdot)$ then $\theta_i = g(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta}$
- The *canonical link* is a function $g(\cdot)$ such that

$$g(\mu_i) = \theta_i = \boldsymbol{x}_i\boldsymbol{\beta},$$

  that is, $g(\cdot) = (b')^{-1}(\cdot)$
- Canonical links provide great simplifications in terms of computation

Distribution	$\mathsf{N}(\mu, \sigma^2)$	$\mathsf{Poisson}(\mu)$	$\mathsf{Bern}(\mu)$	$\mathsf{Ga}(1/\alpha, 1/[\mu\alpha])$
Canonical link $g(\mu)$	μ	$\log(\mu)$	$\log(\frac{\mu}{1-\mu})$	$-1/\mu$

Links and Parameter Interpretation

Can we rescue the interpretation we use in linear models?

- We assume $g(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta}$, which is a linear model in the transformed mean space. Interpretations in this transformed space such as "*a 1 unit difference in x_j is associated with a β_j difference in $g(\mu_i)$*" are technically correct but not as clear in general

- In a linear model, we have that for a continuous x_j

$$\frac{\partial \mathrm{E}(Y \mid \boldsymbol{x})}{\partial x_j} = \beta_j,$$

  which characterizes the rate of change of $\mathrm{E}(Y \mid \boldsymbol{x})$ as a function of x_j for fixed values of the other covariates.

- However, in a GLM this is not the case any more

$$\frac{\partial \mathrm{E}(Y \mid \boldsymbol{x})}{\partial x_j} = \beta_j \times (g^{-1})'(\boldsymbol{x}\boldsymbol{\beta}),$$

  which will generally depend on the values of the other covariates.

- *Moral*: interpretation of the regression parameters in a GLM is specific to the link function.

Links and Parameter Interpretation

Identity link: $g(\mu) = \mu$

- This is what we have been using thus far with linear models.
- You may find people who use the identity link even if Y is in a restricted range, why?

Links and Parameter Interpretation

- The identity link leads to an appealing parameter interpretation
- Say $Y \in \{0, 1\}$. If we take

$$\mathrm{E}(Y \mid x, z) = P(Y = 1 \mid x, z) = \beta_0 + \beta_1 x + \beta_2 z$$

  we obtain

$$P(Y = 1 \mid x + 1, z) - P(Y = 1 \mid x, z) = \beta_1$$

- This is, the *risk difference* can be summarized by a single number
- This might be OK if linearity of $P(Y = 1 \mid x, z)$ is justifiable in a restricted range of the covariates.
- However, in general, your regression function will be misspecified, in particular with continuous covariates, but you are relying on correct specification for parameter interpretation!

Links and Parameter Interpretation

Log link: $g(\mu) = \log(\mu)$

- This is the canonical link for the Poisson GLM, and a common choice when $Y_i \geq 0$.

- If we take

$$\log[\mathrm{E}(Y \mid x, z)] = \beta_0 + \beta_1 x + \beta_2 z,$$

this is equivalent to

$$\mathrm{E}(Y \mid x, z) = e^{\beta_0} e^{\beta_1 x} e^{\beta_2 z}$$

- The parameter $\exp(\beta_1)$ has a relatively straightforward interpretation:

$$\mathrm{E}(Y \mid x + 1, z) = e^{\beta_1} \mathrm{E}(Y \mid x, z),$$

i.e., $\exp(\beta_1)$ is the multiplicative change in the average response associated with a one unit increase in x, with z held constant.

Links and Parameter Interpretation

- The *relative risk* is
$$\frac{\mathrm{E}(Y \mid x+1, z)}{\mathrm{E}(Y \mid x, z)} = e^{\beta_1}$$

- Note that the relative risk interpretation leads people to use the log link when Y is binary

- Taking
$$\log P(Y = 1 \mid x, z) = \beta_0 + \beta_1 x + \beta_2 z$$
leads to
$$\frac{P(Y = 1 \mid x+1, z)}{P(Y = 1 \mid x, z)} = e^{\beta_1}$$

- When $Y \in \{0, 1\}$, while this approach is appealing due to its simplicity, your regression function in general will be misspecified, but you are relying on correct specification for parameter interpretation!

Links and Parameter Interpretation

Logit link: $g(\pi) = \log[\pi/(1-\pi)] := \text{logit}(\pi), \quad \pi \in (0,1)$

- This is the canonical link for the Bernoulli GLM.
- For $Y \in \{0,1\}$, if we take

$$\text{logit}[P(Y=1 \mid x,z)] = \beta_0 + \beta_1 x + \beta_2 z,$$

this is equivalent to

$$P(Y=1 \mid x,z) = \frac{\exp(\beta_0 + \beta_1 x + \beta_2 z)}{1+\exp(\beta_0 + \beta_1 x + \beta_2 z)} := \text{expit}(\beta_0 + \beta_1 x + \beta_2 z)$$

- The parameter $\exp(\beta_1)$ has a relatively straightforward interpretation:

$$\frac{P(Y=1 \mid x+1,z)/P(Y=0 \mid x+1,z)}{P(Y=1 \mid x,z)/P(Y=0 \mid x,z)} = e^{\beta_1},$$

i.e., so the odds of $Y=1$ when $X=x+1$ is $\exp(\beta_1)$ times the odds of $Y=1$ when $X=x$.

Links and Parameter Interpretation

Other links for binary regression: $g : (0, 1) \to \mathbb{R}$

- Basically any inverse CDF of a continuous variable could be used
- For example, a common choice is the *probit link*, which is the inverse CDF of the standard normal, so that

$$P(Y = 1 \mid \boldsymbol{x}) := \pi(\boldsymbol{x}) = \Phi(\boldsymbol{x}\boldsymbol{\beta}).$$

  However, a direct interpretation of the β_j's is difficult

- This motivates a *latent variable representation*

Links and Parameter Interpretation

- Let $Z = \boldsymbol{x}\beta + \epsilon$, where ϵ follows a distribution with CDF F
- We do not observe Z, but we observe $Y = 1$ when Z exceeds a *tolerance level* t, and $Y = 0$ otherwise
- This construction implies a model for Y:

$$\begin{aligned} P(Y = 1 \mid \boldsymbol{x}) &= P(Z > t \mid \boldsymbol{x}) \\ &= P(\boldsymbol{x}\beta + \epsilon > t \mid \boldsymbol{x}) \\ &= P(\epsilon > t - \boldsymbol{x}\beta \mid \boldsymbol{x}) \\ &= 1 - F(t - \boldsymbol{x}\beta) \end{aligned}$$

- Since $\boldsymbol{x}\beta$ includes an intercept, for identifiability we take $t = 0$

Links and Parameter Interpretation

- If the distribution of ϵ is symmetric about zero,

$$1 - F(-\boldsymbol{x}\boldsymbol{\beta}) = F(\boldsymbol{x}\boldsymbol{\beta}),$$

  and $F^{-1}[\pi(\boldsymbol{x})] = \boldsymbol{x}\boldsymbol{\beta}$ provides a link function

  - $\epsilon \sim \mathrm{N}(0, 1)$ leads to the *probit link*: $g(\pi) = \Phi^{-1}(\pi)$
  - $\epsilon \sim \mathrm{Logistic}(0, 1)$ leads to the *logit link*: $g(\pi) = \log[\pi/(1-\pi)]$

- Otherwise, if the distribution of ϵ is not symmetric about zero $-F^{-1}[1 - \pi(\boldsymbol{x})] = \boldsymbol{x}\boldsymbol{\beta}$ still provides a link function

  - $\epsilon \sim \mathrm{Gumbel}(0, 1)$ leads to the *complementary-log-log link*: $g(\pi) = \log[-\log(1-\pi)]$

- Interpretation of β_j's could be obtained by thinking about the latent linear model $Z = \boldsymbol{x}\boldsymbol{\beta} + \epsilon$, but this latent variable representation may not make scientific sense in many contexts

Likelihood Inference for GLMs: Estimation

- For an independent sample, the likelihood function is

$$L(\boldsymbol{\beta}, \alpha) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta}, \alpha) = \prod_{i=1}^{n} f(y_i \mid \theta_i, \alpha_i)$$

- The log-likelihood is given by

$$\log L(\boldsymbol{\beta}, \alpha) := l(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} \left[\frac{y_i \theta_i - b(\theta_i)}{\alpha_i} + c(y_i, \alpha_i) \right]$$

- Here the vector of canonical parameters is a function of $\boldsymbol{\beta}$

$$\begin{aligned} \boldsymbol{\theta} &= \boldsymbol{\theta}(\boldsymbol{\beta}) \\ &= [\theta_1, ..., \theta_n]^T \\ &= [\theta_1(\mu_1(\boldsymbol{\beta})), ..., \theta_n(\mu_n(\boldsymbol{\beta}))]^T \end{aligned}$$

- For maximum likelihood estimation we take the derivatives of the log-likelihood function and set them to zero

Likelihood Inference for GLMs: Estimation

- The derivatives of the log-likelihood are known as the *score vector*
- To simplify notation, we will denote $l := l(\boldsymbol{\beta}, \alpha)$ and $l_i := l_i(\boldsymbol{\beta}, \alpha)$
- Using the chain rule, the score function (for $\boldsymbol{\beta}$) is

$$\begin{aligned}
\boldsymbol{S}(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}} &= \begin{pmatrix} \frac{\partial l}{\partial \beta_0} \\ \vdots \\ \frac{\partial l}{\partial \beta_k} \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^n \frac{\partial l_i}{\partial \theta_i} \frac{d\theta_i}{d\mu_i} \frac{\partial \mu_i}{\partial \beta_0} \\ \vdots \\ \sum_{i=1}^n \frac{\partial l_i}{\partial \theta_i} \frac{d\theta_i}{d\mu_i} \frac{\partial \mu_i}{\partial \beta_k} \end{pmatrix} \\
&= \sum_{i=1}^n \frac{\partial l_i}{\partial \theta_i} \frac{d\theta_i}{d\mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}},
\end{aligned}$$

where $\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = (\frac{\partial \mu_i}{\partial \beta_0}, \ldots, \frac{\partial \mu_i}{\partial \beta_k})^T$

Likelihood Inference for GLMs: Estimation

The score

$$\boldsymbol{S}(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i} \frac{d\theta_i}{d\mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}},$$

can be rewritten noticing that

$$\begin{aligned} \frac{\partial l_i}{\partial \theta_i} &= \frac{y_i - b'(\theta_i)}{\alpha_i}, \\ \frac{d\theta_i}{d\mu_i} &= \frac{d(b')^{-1}(\mu_i)}{d\mu_i} = \frac{1}{b''[(b')^{-1}(\mu_i)]} = \frac{1}{b''(\theta_i)} = \frac{1}{V(\mu_i)}, \end{aligned}$$

and we had seen before that

$$\begin{aligned} \mathrm{E}(Y_i \mid \boldsymbol{x}_i) &= b'(\theta_i) \\ \mathrm{var}(Y_i \mid \boldsymbol{x}_i) &= \alpha_i V(\mu_i) \end{aligned}$$

Likelihood Inference for GLMs: Estimation

Hence,

$$
\begin{aligned}
\boldsymbol{S}(\beta) &= \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i} \frac{d\theta_i}{d\mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{n} \frac{[y_i - \mathrm{E}(Y_i \mid \boldsymbol{x}_i)]}{\mathrm{var}(Y_i \mid \boldsymbol{x}_i)} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right) \\
&= \begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_0} & \cdots & \frac{\partial \mu_n}{\partial \beta_0} \\ & \vdots & \\ \frac{\partial \mu_1}{\partial \beta_k} & \cdots & \frac{\partial \mu_n}{\partial \beta_k} \end{pmatrix} \begin{pmatrix} \frac{y_1 - E(Y_1|\boldsymbol{X}_1)}{\mathrm{var}(Y_1|\boldsymbol{X}_1)} \\ \vdots \\ \frac{y_n - E(Y_n|\boldsymbol{X}_n)}{\mathrm{var}(Y_n|\boldsymbol{X}_n)} \end{pmatrix} \\
&= \boldsymbol{D}^{\mathsf{T}} \boldsymbol{V}^{-1} [\boldsymbol{y} - \boldsymbol{\mu}(\beta)] / \alpha
\end{aligned}
$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and $\boldsymbol{\mu}(\beta) = [\mu_1(\beta), \ldots, \mu_n(\beta)]^T$

Likelihood Inference for GLMs: Estimation

$$\boldsymbol{S}(\boldsymbol{\beta}) \quad = \quad \boldsymbol{D}^T \boldsymbol{V}^{-1} \left[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right] / \alpha$$

- In these slides $\boldsymbol{V}$ is the $n \times n$ diagonal matrix with i-th diagonal element $V(\mu_i)/\phi_i$ [1] and $\text{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \alpha \boldsymbol{V}$.

- $\boldsymbol{D}$ is the $n \times (k+1)$ matrix with elements $\partial \mu_i / \partial \beta_j$, $i = 1, ..., n$, $j = 0, ..., k$

- Denoting the linear predictor $\eta_i = \boldsymbol{x}_i \boldsymbol{\beta}$, then $\mu_i = g^{-1}(\eta_i)$ and we can further write

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \frac{d\mu_i}{d\eta_i} \boldsymbol{x}_i^T = \frac{1}{g'(g^{-1}(\eta_i))} \boldsymbol{x}_i^\top,$$

and in matrix form

$$\boldsymbol{D} = \text{diag}\{\frac{d\mu_i}{d\eta_i}\}\boldsymbol{X} = [\text{diag}\{g'(g^{-1}(\eta_i))\}]^{-1}\boldsymbol{X}$$

[1] Wakefield's book takes $\phi_i = 1$, Agresti's book absorbs α into $\boldsymbol{V}$

Likelihood Inference for GLMs: Estimation

- The MLE $\hat{\boldsymbol{\beta}}_n$ is obtained by solving the score equations

$$\boldsymbol{S}(\boldsymbol{\beta}) \;=\; \boldsymbol{D}^T \boldsymbol{V}^{-1} \left[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right] / \alpha \;=\; \boldsymbol{0}$$

  which we will do using numerical methods

- From traditional likelihood theory we obtain that, under our model, the MLE $\hat{\boldsymbol{\beta}}_n$ has asymptotic distribution

$$\mathcal{I}_n(\boldsymbol{\beta})^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \;\to_d\; \mathrm{N}_{k+1}(\boldsymbol{0}, \boldsymbol{I}_{k+1}),$$

  where the expected Fisher information is given by the covariance matrix of the score

$$\begin{aligned}
\mathcal{I}_n(\boldsymbol{\beta}) &= \mathrm{E}[\boldsymbol{S}(\boldsymbol{\beta})\boldsymbol{S}(\boldsymbol{\beta})^T] \\
&= \boldsymbol{D}^T \boldsymbol{V}^{-1} \mathrm{E}[(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))^T] \boldsymbol{V}^{-1} \boldsymbol{D} / \alpha^2 \\
&= \boldsymbol{D}^T \boldsymbol{V}^{-1} \boldsymbol{D} / \alpha \\
&= \boldsymbol{X}^T \mathrm{diag} \left\{ \frac{\phi_i (d\mu_i / d\eta_i)^2}{\alpha V(\mu_i)} \right\} \boldsymbol{X} \\
&= \boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{\beta}) \boldsymbol{X}
\end{aligned}$$

Likelihood Inference for GLMs: Estimation

- In practice we use

$$\mathcal{I}_n(\hat{\beta}_n) = \hat{\boldsymbol{D}}^T \hat{\boldsymbol{V}}^{-1} \hat{\boldsymbol{D}}/\alpha = \boldsymbol{X}^T \boldsymbol{W}(\hat{\beta}_n)\boldsymbol{X},$$

  where $\hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{D}}$ are evaluated at $\hat{\beta}_n$

- The asymptotic variance of the estimator is

$$\widehat{\text{var}}(\hat{\beta}_n) = \mathcal{I}_n(\hat{\beta}_n)^{-1}$$

- In some cases α is unknown (e.g., normal and gamma), so it needs to be estimated

Likelihood Inference for GLMs: Estimation

- A common approach is to use a method of moments estimator

- *Remark*: let $\boldsymbol{Z}$ be a $n \times 1$ random vector with $\mathrm{E}[\boldsymbol{Z}] = \boldsymbol{\mu}$, $\mathrm{var}(\boldsymbol{Z}) = \Sigma$ and $\boldsymbol{A}$ be a symmetric $n \times n$ matrix. Then $\mathrm{E}[\boldsymbol{Z}^\top \boldsymbol{A} \boldsymbol{Z}] = \mathrm{tr}(\boldsymbol{A}\Sigma) + \boldsymbol{\mu}^\top \boldsymbol{A} \boldsymbol{\mu}$.

- Then $\mathrm{E}[(\boldsymbol{Y} - \boldsymbol{\mu})^\top \boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})/\alpha] = n$, and an unbiased estimator of α would be (with $\boldsymbol{\mu}$ known)

$$\hat{\alpha} = (\boldsymbol{Y} - \boldsymbol{\mu})^\top \boldsymbol{V}^{-1}(\boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})/n$$

- A degrees of freedom corrected estimator is then:

$$\hat{\alpha} = \frac{1}{n-k-1} \sum_{i=1}^{n} \frac{\phi_i (Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

where $\hat{\mu}_i = \hat{\mu}_i(\hat{\boldsymbol{\beta}})$.

Likelihood Inference for GLMs: Computation

- Computation is tipically easy for GLMs, since the log-likelihood surface is well-behaved for commonly used link functions
- A variant of *Newton-Raphson* (a generic method for root-finding) known as *Fisher scoring* is commonly used to find the MLEs
- The score $\boldsymbol{S}(\boldsymbol{\beta})$ represent a $(k+1) \times 1$ vector of functions that are themselves functions of a $(k+1) \times 1$ vector $\boldsymbol{\beta}$
- We wish to find $\boldsymbol{\beta}$ such that

$$\boldsymbol{S}(\boldsymbol{\beta}) = \boldsymbol{D}^{\mathsf{T}} \boldsymbol{V}^{-1} \left[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right] / \alpha = \mathbf{0}$$

Likelihood Inference for GLMs: Computation

- A first-order Taylor series expansion of $\boldsymbol{S}(\boldsymbol{\beta})$ about $\boldsymbol{\beta}^{(0)}$ gives:

$$\boldsymbol{S}(\boldsymbol{\beta}) \approx \boldsymbol{S}(\boldsymbol{\beta}^{(0)}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})^{\mathsf{T}} \boldsymbol{S}'(\boldsymbol{\beta}^{(0)})$$

where

$$\boldsymbol{S}'(\boldsymbol{\beta}) = \left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \beta_0}, \ldots, \frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \beta_k}\right]$$

- $\boldsymbol{S}'(\boldsymbol{\beta})$ is the *Hessian* of the log-likelihood $l(\boldsymbol{\beta})$, that is, its (a, b) entry is

$$[\boldsymbol{S}'(\boldsymbol{\beta})]_{a,b} = \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_a \partial \beta_b}$$

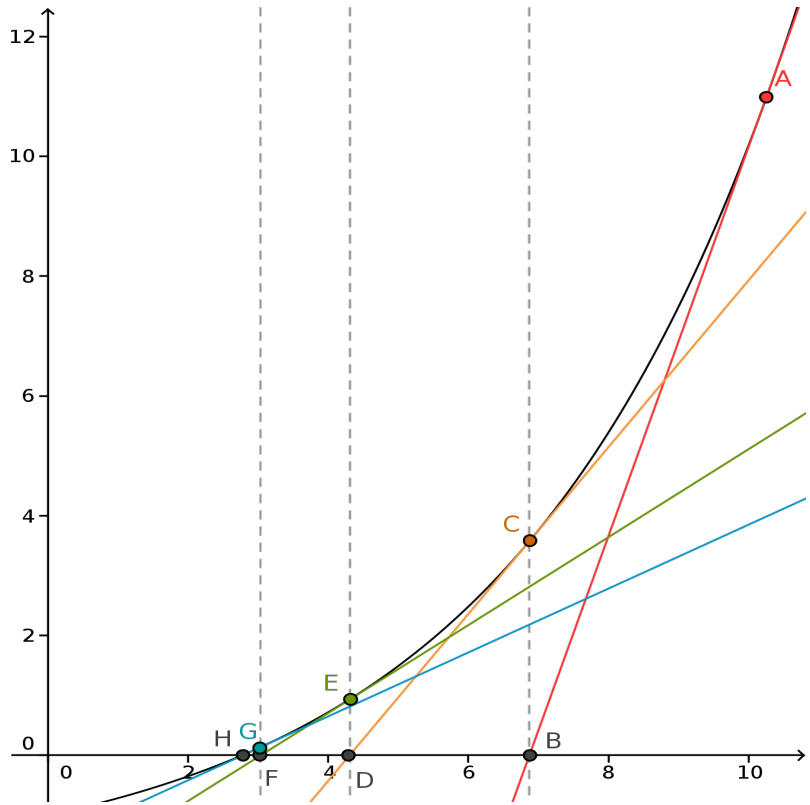
- Setting the left-hand side of the approximation to zero yields

$$\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)} - \boldsymbol{S}'(\boldsymbol{\beta}^{(0)})^{-1} \boldsymbol{S}(\boldsymbol{\beta}^{(0)})$$

Likelihood Inference for GLMs: Computation

- The Newton-Raphson method iterates for $t = 0, 1, 2, ...$:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \boldsymbol{S}'(\boldsymbol{\beta}^{(t)})^{-1}\boldsymbol{S}(\boldsymbol{\beta}^{(t)})$$



The univariate case for root-finding. Taken from `https://commons.wikimedia.org/wiki/File:Newton%E2%80%93Raphson.png`

Likelihood Inference for GLMs: Computation

- We derived Newton-Raphson above from iterating an approximate solution to the exact problem we need to solve: $\boldsymbol{S}(\beta) = \mathbf{0}$

- However, it can also be derived from iterating an exact solution to an approximation of the maximum likelihood problem

- A second-order Taylor series expansion of $l(\beta)$ about $\beta^{(0)}$ leads to

$$l(\beta) \approx l(\beta^{(0)}) + \boldsymbol{S}(\beta^{(0)})^T(\beta - \beta^{(0)}) + (\beta - \beta^{(0)})^T \boldsymbol{S}'(\beta^{(0)})(\beta - \beta^{(0)})/2$$

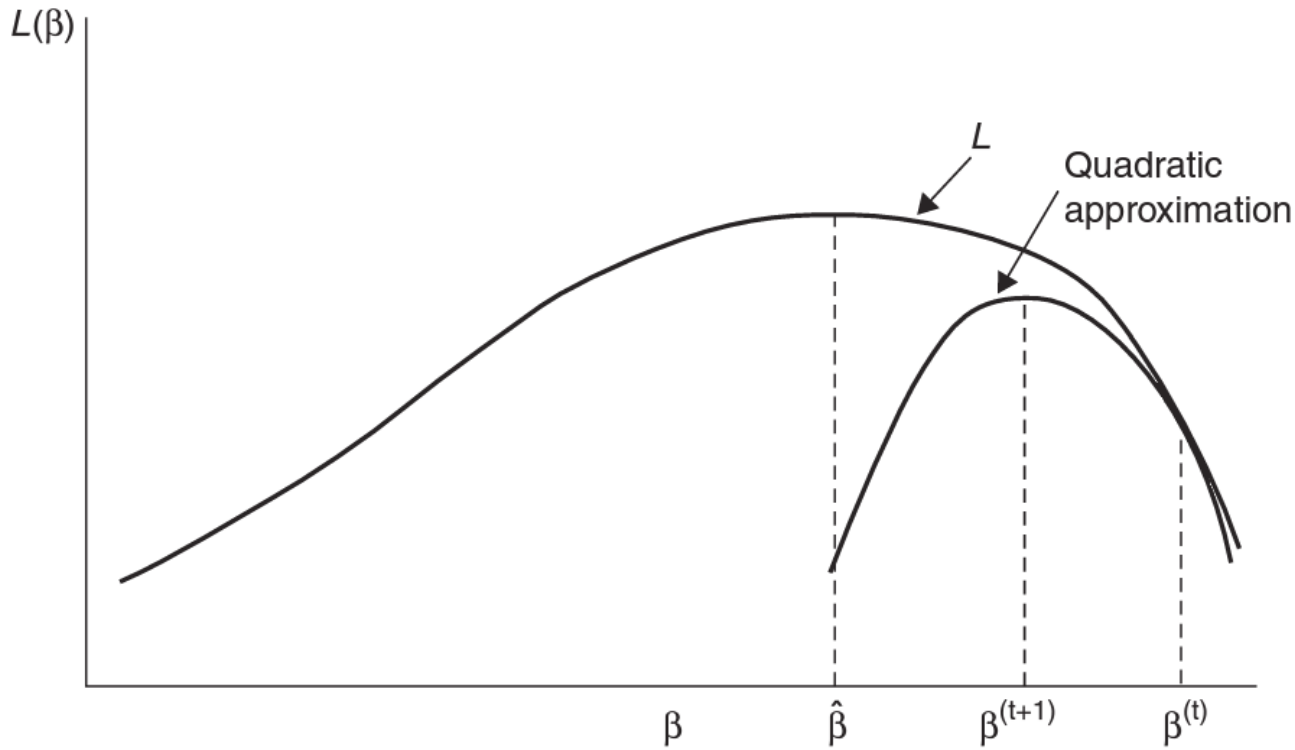
- After taking derivatives of this quadratic approximation, and setting them to zero, we find that the exact minimizer of this approximation of $l(\beta)$ is again

$$\beta = \beta^{(0)} - \boldsymbol{S}'(\beta^{(0)})^{-1}\boldsymbol{S}(\beta^{(0)})$$

Likelihood Inference for GLMs: Computation

- The Newton-Raphson method iterates for $t = 0, 1, 2, ...$:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \boldsymbol{S}'(\boldsymbol{\beta}^{(t)})^{-1}\boldsymbol{S}(\boldsymbol{\beta}^{(t)})$$



A cycle of Newton-Raphson for maximization. Modified from Agresti (2015)

Reminder: Fisher Information

- *Reminder*: the *Fisher information* is defined as the $p \times p$ covariance matrix of the score:

$$\mathcal{I}_n(\beta) = \mathrm{E}\left[\boldsymbol{S}(\beta)\boldsymbol{S}(\beta)^T\right]$$

- If $l(\beta)$ is twice differentiable with respect to β, and under the regularity conditions for $\mathrm{E}[\boldsymbol{S}(\beta)] = \boldsymbol{0}$, we have

$$\mathcal{I}_n(\beta) = -\mathrm{E}\left[\frac{\partial^2}{\partial\beta\partial\beta^T} l(\beta)\right] = -\mathrm{E}\left[\frac{\partial \boldsymbol{S}(\beta)}{\partial\beta}\right] = -\mathrm{E}\left[\boldsymbol{S}'(\beta)\right]$$

- We sometimes refer to $-\mathrm{E}\left[\boldsymbol{S}'(\beta)\right]$ as the *expected* Fisher information, and to $-\boldsymbol{S}'(\beta)$ as the *observed* Fisher information

Likelihood Inference for GLMs: Computation

- *Fisher scoring* is Newton-Raphson applied to the score equation, but with the observed information, $-\boldsymbol{S}'(\beta)$, replaced by the expected information $-\mathrm{E}[\boldsymbol{S}'(\beta)] = \mathcal{I}_n(\beta)$:

$$\beta^{(t+1)} = \beta^{(t)} + \mathcal{I}_n(\beta^{(t)})^{-1}\boldsymbol{S}(\beta^{(t)})$$

- Each update is calculated based on the score and Fisher information evaluated at the previous value

- This general procedure for maximizing the likelihood takes a particular form for GLMs, as we will see now

- Recall that for a GLM, $\mathcal{I}_n(\beta) = \boldsymbol{X}^T\boldsymbol{W}(\beta)\boldsymbol{X}$, with $\boldsymbol{W}(\beta) = \mathrm{diag}\left\{\frac{\phi_i(d\mu_i/d\eta_i)^2}{\alpha V(\mu_i)}\right\}$

- Also, $\boldsymbol{S}(\beta) = \boldsymbol{X}^T\mathrm{diag}\{\frac{d\mu_i}{d\eta_i}\}\boldsymbol{V}^{-1}\left[\boldsymbol{y} - \boldsymbol{\mu}(\beta)\right]/\alpha$

Likelihood Inference for GLMs: Computation

Using $\boldsymbol{W}^{(t)} = \boldsymbol{W}(\boldsymbol{\beta}^{(t)})$, we write

$$\begin{aligned}
\boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} + (\boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}^T \text{diag}\left\{ \left.\frac{d\mu_i}{d\eta_i}\right|_{\boldsymbol{\beta}^{(t)}} \right\} \boldsymbol{V}^{(t)-1}\left[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta}^{(t)})\right]/\alpha \\
&= (\boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(t)}\left[\boldsymbol{X}\boldsymbol{\beta}^{(t)} + (\boldsymbol{W}^{(t)})^{-1}\boldsymbol{u}^{(t)}\right] \\
&= (\boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)}
\end{aligned}$$

where $\boldsymbol{u}^{(t)}$ and $\boldsymbol{z}^{(t)}$ are $n \times 1$ vectors with i-th elements

$$u_i^{(t)} = \frac{\phi_i(y_i - \mu_i^{(t)})}{V_i^{(t)}} \left.\frac{d\mu_i}{d\eta_i}\right|_{\boldsymbol{\beta}^{(t)}},$$

and

$$z_i^{(t)} = \boldsymbol{x}_i\boldsymbol{\beta}^{(t)} + (y_i - \mu_i^{(t)}) \left.\frac{d\eta_i}{d\mu_i}\right|_{\boldsymbol{\beta}^{(t)}}.$$

Likelihood Inference for GLMs: Computation

- The Fisher scoring updates

$$\beta^{(t+1)} = (\boldsymbol{X}^T \boldsymbol{W}^{(t)} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}^{(t)} \boldsymbol{z}^{(t)}$$

  have the form of the solution to a weighted least squares problem:

$$\beta^{(t+1)} = \arg\min_{\beta} \; (\boldsymbol{z}^{(t)} - \boldsymbol{X}\beta)^T \boldsymbol{W}^{(t)} (\boldsymbol{z}^{(t)} - \boldsymbol{X}\beta)$$

  with "working", or "adjusted" response $\boldsymbol{z}^{(t)}$.

- This method is therefore also known as *iteratively reweighted least squares* (IRLS).

- For canonical links, the observed and expected information coincide, so that the Fisher scoring and Newton-Raphson methods are identical.

Likelihood Inference for GLMs: Computation

- Iteratively reweighted least squares can also be justified by weighted least squares of a linearized the response

- Taking a first-order Taylor series approximation of the link function about the mean $\mu_i^{(0)} = g^{-1}(\boldsymbol{x}_i\boldsymbol{\beta}^{(0)})$ leads to

$$g(\mu_i) \approx g(\mu_i^{(0)}) + (\mu_i - \mu_i^{(0)})g'(\mu_i^{(0)}) = \boldsymbol{x}_i\boldsymbol{\beta}^{(0)} + (\mu_i - g[\boldsymbol{x}_i\boldsymbol{\beta}^{(0)}]) \left.\frac{d\eta_i}{d\mu_i}\right|_{\boldsymbol{\beta}^{(0)}}$$

  where $g(\mu_i) = \eta_i = \boldsymbol{x}_i\boldsymbol{\beta}$

- Replacing y_i for μ_i leads to (as before)

$$z_i^{(t)} = \boldsymbol{x}_i\boldsymbol{\beta}^{(t)} + (y_i - \mu_i^{(t)}) \left.\frac{d\eta_i}{d\mu_i}\right|_{\boldsymbol{\beta}^{(t)}}$$

- From this, $\text{var}(\boldsymbol{z}^{(t)}) = \boldsymbol{W}(\boldsymbol{\beta}^{(t)})^{-1} = \text{diag}\left\{\frac{\phi_i(d\mu_i/d\eta_i)^2}{\alpha V(\mu_i)}\right\}^{-1}|_{\boldsymbol{\beta}^{(t)}}$

- IRLS solves

$$\boldsymbol{\beta}^{(t+1)} = \arg\min_{\boldsymbol{\beta}} \; (\boldsymbol{z}^{(t)} - \boldsymbol{X}\boldsymbol{\beta})^T \text{var}(\boldsymbol{z}^{(t)})^{-1}(\boldsymbol{z}^{(t)} - \boldsymbol{X}\boldsymbol{\beta})$$

Likelihood Inference for GLMs: Computation

Canonical links facilitate computations

- If we take the canonical link, $g(\mu) = (b')^{-1}(\mu)$, then

$$\eta_i = \boldsymbol{x}_i\boldsymbol{\beta} = g(\mu_i) = g(b'[\theta_i]) = \theta_i$$

- Then

$$\frac{d\mu_i}{d\eta_i} = \frac{d\mu_i}{d\theta_i} = \frac{db'(\theta_i)}{d\theta_i} = b''(\theta_i) = V(\mu_i) = \left(\frac{d\theta_i}{d\mu_i}\right)^{-1}$$

- This implies

$$\frac{\partial l_i}{\partial \boldsymbol{\beta}} = \frac{\partial l_i}{\partial \theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \frac{1}{\alpha_i}\boldsymbol{x}_i^T\left[Y_i - \mu_i(\boldsymbol{\beta})\right]$$

- So that the score simplifies to

$$\boldsymbol{S}(\boldsymbol{\beta}) \quad = \quad \boldsymbol{X}^T \text{diag}\{\phi_i\}\left[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right]/\alpha$$

and the likelihood equations are $\boldsymbol{X}^T\text{diag}\{\phi_i\}\boldsymbol{Y} = \boldsymbol{X}^T\text{diag}\{\phi_i\}\boldsymbol{\mu}(\boldsymbol{\beta})$

Likelihood Inference for GLMs: Computation

Some comments on the shape of the log-likelihood:

- For a GLM, the log-likelihood is given by $l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta})$, with

$$l_i(\boldsymbol{\beta}) = \frac{y_i\theta_i - b(\theta_i)}{\alpha_i} + c(y_i, \alpha_i)$$

- We can ignore $c(y_i, \alpha_i)$ as it doesn't depend on $\boldsymbol{\beta}$. Replacing θ_i

$$\begin{aligned} l_i(\boldsymbol{\beta}) &= \frac{y_i\ (b')^{-1}(\mu_i) - b((b')^{-1}(\mu_i))}{\alpha_i} \\ &= \frac{y_i\ (b')^{-1}(g^{-1}(\eta_i)) - b((b')^{-1}(g^{-1}(\eta_i)))}{\alpha_i} \end{aligned}$$

- Concavity cannot be guaranteed in general for $l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta})$ with arbitrary link functions (see, e.g., Fahrmeir and Kaufmann, 1985); e.g., log-likelihood of gamma GLM with identity link may not be concave

- Wedderburn (1976) showed existence and uniqueness of MLE for the normal, Poisson, gamma, and binomial with common link functions

- We'll see some important particular results that are somewhat easy to derive: canonical links and links for binary responses

Likelihood Inference for GLMs: Computation

If we work with the canonical link, i.e. $g(\cdot) = (b')^{-1}(\cdot)$,

$$l_i(\boldsymbol{\beta}) = \frac{y_i\eta_i - b(\eta_i)}{\alpha_i} = \frac{y_i\boldsymbol{x}_i\boldsymbol{\beta} - b(\boldsymbol{x}_i\boldsymbol{\beta})}{\alpha_i}$$

- Facts about convex functions
  - If $f''(x) \geq 0$ for all x then f is convex
  - If f is convex, then $h(x) = f(Ax + b)$ is convex ($Ax + b$ is an *affine function*)
  - If f is convex, $-f$ is concave
  - If f_1 and f_2 are convex (concave), then $w_1 f_1(\cdot) + w_2 f_2(\cdot)$, $w_1, w_2 \geq 0$, is convex (concave)
- Remember that in the exponential dispersion family, $\text{var}(Y_i \mid \theta_i, \alpha_i) = \alpha_i b''(\theta_i)$, so $b''(\cdot) \geq 0$, meaning that the *log-partition* function $b(\cdot)$ is convex
- Therefore, $l_i(\boldsymbol{\beta}) = (y_i\boldsymbol{x}_i\boldsymbol{\beta} - b(\boldsymbol{x}_i\boldsymbol{\beta}))/\alpha_i$ is concave, and so is $l(\boldsymbol{\beta}) = \sum_{i=1}^n l_i(\boldsymbol{\beta})$

Likelihood Inference for GLMs: Computation

Now we'll consider link functions for binary regression corresponding to the inverse CDF of a *log-concave* distribution

- *Log-concave* distribution: log of PDF is concave (implies log of CDF also concave)
  - Examples: normal, logistic, Gumbel, and many others
- Facts about log-concave functions
  - Product of log-concave functions is log-concave
  - If $f(x, y)$ is log-concave then $\int f(x, y)dy$ is log-concave

Likelihood Inference for GLMs: Computation

- Let $g = F^{-1}$, where F is the CDF of a log-concave distribution
- Under Bernoulli likelihood

$$\begin{aligned} l_i(\beta) &= \log\{[g^{-1}(\boldsymbol{x}_i\beta)]^{y_i}[1 - g^{-1}(\boldsymbol{x}_i\beta)]^{1-y_i}\} \\ &= y_i \log F(\boldsymbol{x}_i\beta) + (1 - y_i)\log[1 - F(\boldsymbol{x}_i\beta)] \end{aligned}$$

- $\log F(\boldsymbol{x}_i\beta)$: composition of concave and affine functions, therefore concave
- $1 - F(z) = \int I(z \leq x \leq \infty) f(x) dx$: log-concave since both $I(z \leq x \leq \infty)$ and $f(x)$ are log-concave
- $\log[1 - F(\boldsymbol{x}_i\beta)]$: composition of concave and affine functions, therefore concave
- Therefore, $l_i(\beta)$ is concave, and so is $l(\beta) = \sum_{i=1}^{n} l_i(\beta)$

Reminder: Likelihood-Based Hypothesis Testing

There are three generic recipes for tests under likelihood-based inference:

- Score tests [2]
- Wald tests [3]
- Likelihood ratio tests [4]

Since it is difficult, in general, to obtain exact finite sample inferences, these tests rely on asymptotic results

[2] a.k.a. *Rao's Score test* after C. R. Rao, and as *Lagrange Multiplier test* in econometrics)

[3] Due to Abraham Wald

[4] Due to Samuel S. Wilks

Reminder: Score Tests

- Notice that

$$\boldsymbol{S}(\beta) = \frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta},$$

  and so under i.i.d. data $\boldsymbol{S}(\beta)/n$ converges to a normal distribution

- We also had seen that under the model $\mathsf{E}[\boldsymbol{S}(\beta)] = \boldsymbol{0}$ and

$$\mathsf{var}[\boldsymbol{S}(\beta)] = \mathsf{E}[\boldsymbol{S}(\beta)\boldsymbol{S}(\beta)^T] = \mathcal{I}_n(\beta)$$

- We then have

$$\mathcal{I}_n(\beta)^{-1/2}\boldsymbol{S}_n(\beta) \ \rightarrow_d \ \mathsf{N}_{k+1}(\boldsymbol{0}, \boldsymbol{I}_{k+1}),$$

  and

$$\boldsymbol{S}_n(\beta)^T\mathcal{I}_n(\beta)^{-1}\boldsymbol{S}_n(\beta) \ \rightarrow_d \ \chi^2_{k+1}$$

Reminder: Score Tests

- The simple hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ can then be tested using

$$\boldsymbol{S}_n(\boldsymbol{\beta}_0)^T \mathcal{I}_n(\boldsymbol{\beta}_0)^{-1} \boldsymbol{S}_n(\boldsymbol{\beta}_0) \ \rightarrow_d \ \chi^2_{k+1}$$

- Intuition: since $\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, this quadratic form is small when the MLE $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$ are close

Reminder: Wald Tests

- The Wald statistic is based upon the asymptotic distribution

$$\mathcal{I}_n(\boldsymbol{\beta})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \ \rightarrow_d \ \mathrm{N}_{k+1}(\mathbf{0}, \boldsymbol{I}_{k+1})$$

  which implies

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathcal{I}_n(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \ \rightarrow_d \ \chi^2_{k+1}$$

- Therefore, under the simple null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, the Wald statistic is given by

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \mathcal{I}_n(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \ \rightarrow_d \ \chi^2_{k+1}$$

Reminder: Likelihood Ratio Tests

- Consider the *likelihood ratio* for testing $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ given by

$$R_n(\boldsymbol{\beta}_0) = \frac{L_n(\boldsymbol{\beta}_0)}{L_n(\hat{\boldsymbol{\beta}})},$$

  where $L_n(\cdot)$ is the likelihood function and $\hat{\boldsymbol{\beta}}$ is the MLE so that $R(\boldsymbol{\beta}_0) \leq 1$.

- A second order Taylor expansion of $\log L_n(\boldsymbol{\beta}_0) = l_n(\boldsymbol{\beta}_0)$ about $\hat{\boldsymbol{\beta}}$ gives

$$l_n(\boldsymbol{\beta}_0) = l_n(\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})^T \boldsymbol{S}(\hat{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})^T \boldsymbol{S}'(\boldsymbol{\beta}^*)(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}),$$

  where $\boldsymbol{\beta}^*$ is between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$

- In this expression, $\boldsymbol{S}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, and under H_0

$$\boldsymbol{S}'(\boldsymbol{\beta}^*) = \left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}^*} \to_p -\mathcal{I}_n(\boldsymbol{\beta}_0)$$

Reminder: Likelihood Ratio Tests

- Hence,

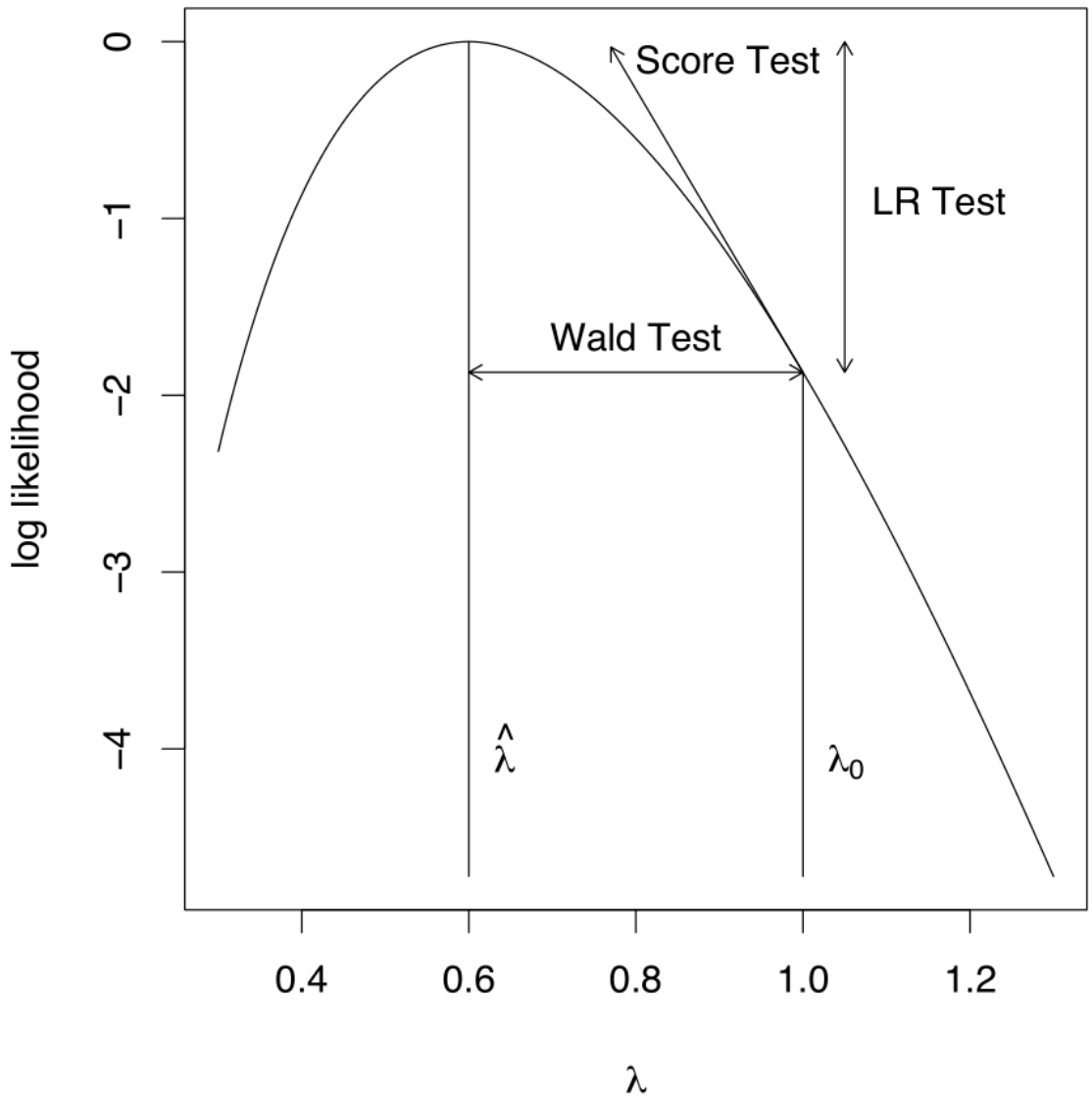
$$-2\{l_n(\boldsymbol{\beta}_0) - l_n(\hat{\boldsymbol{\beta}})\} \approx (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathsf{T}} \mathcal{I}_n(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

  and so

$$-2\{l_n(\boldsymbol{\beta}_0) - l_n(\hat{\boldsymbol{\beta}})\} \ \rightarrow_d \ \chi^2_{k+1}$$

- This can therefore be used to test $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$

Reminder: Likelihood-Based Hypothesis Testing



Geometric interpretation of score, Wald and likelihood ratio (LR) statistics. Taken from Wakefield (2013).

Likelihood Inference for GLMs: Hypothesis Testing

- As in the linear model, we are interested in testing more general hypotheses

$$H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b},$$

where $\boldsymbol{A}$ is $q \times (k+1)$ matrix of rank q

- For example,

$$H_0 : \beta_1 = \cdots = \beta_q = 0, \text{ for } 1 \leq q \leq k,$$

or

$$H_0 : \beta_1 = \cdots = \beta_{q+1}, \text{ for } 1 \leq q < k$$

Likelihood Inference for GLMs: Wald Tests

- From traditional likelihood theory we obtain the approximate distribution of the MLE as

$$\hat{\boldsymbol{\beta}}_n \approx \mathsf{N}_{k+1}[\ \boldsymbol{\beta},\ \mathcal{I}_n(\boldsymbol{\beta})^{-1}\ ]$$

- From this,

$$\boldsymbol{A}\hat{\boldsymbol{\beta}}_n \approx \mathsf{N}_q[\ \boldsymbol{A}\boldsymbol{\beta},\ \boldsymbol{A}\mathcal{I}_n(\boldsymbol{\beta})^{-1}\boldsymbol{A}^T\ ]$$

- This implies

$$(\boldsymbol{A}\hat{\boldsymbol{\beta}}_n - \boldsymbol{A}\boldsymbol{\beta})^T[\boldsymbol{A}\mathcal{I}_n(\boldsymbol{\beta})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_n - \boldsymbol{A}\boldsymbol{\beta}) \approx \chi^2_q$$

Likelihood Inference for GLMs: Wald Tests

- The Wald test for $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$ is then based on

$$Q = (\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})^T [\boldsymbol{A}\mathcal{I}_n(\hat{\boldsymbol{\beta}}_0)^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b}) \approx \chi^2_q,$$

  where we replaced $\boldsymbol{\beta}$ for its MLE under H_0, denoted $\hat{\boldsymbol{\beta}}_0$

- In the normal linear model, we used Q/q which follows an $F_{q,n-k-1}$ distribution

- Since $qF_{q,n-k-1} \overset{n\to\infty}{\to} \chi^2_q$, then the approximate Wald test is asymptotically equivalent to the F test in the normal linear model

Likelihood Inference for GLMs: Score Tests

- Denote the MLE of $\boldsymbol{\beta}$ under $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$ as $\hat{\boldsymbol{\beta}}_0$
- $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ can then be tested using

$$\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_0)^T \mathcal{I}_n(\hat{\boldsymbol{\beta}}_0)^{-1} \boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_0) \rightarrow_d \chi^2_q$$

- Intuition: since $\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$, this quadratic form is small when the MLE $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_0$ are close

Likelihood Inference for GLMs: Score Tests

- The score test can be motivated using Lagrange multipliers, which is why it receives the name *Lagrange multiplier test* in econometrics
- Consider the null hypothesis $H_0 : h_j(\boldsymbol{\beta}) = 0, \quad j = 1, \ldots, q$
- The MLE under the null is obtained from solving the constrained optimization problem

$$\arg\max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) \quad \text{s.t.} \quad h_j(\boldsymbol{\beta}) = 0, \quad j = 1, \ldots, q$$

- The corresponding Lagrangian is

$$l(\boldsymbol{\beta}) + \sum_{j=1}^{q} \lambda_j h_j(\boldsymbol{\beta})$$

Likelihood Inference for GLMs: Score Tests

- Differentiating the Lagrangian with respect to $\boldsymbol{\beta}$ we obtain the first-order conditions

$$\boldsymbol{S}(\boldsymbol{\beta}) + H(\boldsymbol{\beta})\boldsymbol{\lambda} = \mathbf{0}$$

where H is a $(k+1) \times q$ matrix containing $\partial h_j / \partial \beta_h$ in its (h, j) entry, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)^T$

- The Lagrangian multiplier test appears in econometrics as

$$\hat{\boldsymbol{\lambda}}^T H(\hat{\boldsymbol{\beta}}_0)^T \mathcal{I}_n(\hat{\boldsymbol{\beta}}_0)^{-1} H(\hat{\boldsymbol{\beta}}_0)\hat{\boldsymbol{\lambda}} = \boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_0)^T \mathcal{I}_n(\hat{\boldsymbol{\beta}}_0)^{-1} \boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_0) \ \rightarrow_d \ \chi^2_q$$

- Aitcheson and Silvey (*Annals of Mathematical Statistics*, 1958) proposed the test based on the Lagrange multipliers

- See Breusch and Pagan (*The Review of Economic Studies*, 1980) for more details

Likelihood Inference for GLMs: Likelihood Ratio Tests

- Denote the MLE of $\boldsymbol{\beta}$ under $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$ as $\hat{\boldsymbol{\beta}}_0$
- The *likelihood ratio* statistic is given by

$$R_n(\hat{\boldsymbol{\beta}}_0) = \frac{L_n(\hat{\boldsymbol{\beta}}_0)}{L_n(\hat{\boldsymbol{\beta}})} \leq 1,$$

- Under H_0 we obtain the likelihood ratio statistic,

$$-2 \log R(\widehat{\boldsymbol{\theta}}_0) \rightarrow_d \chi^2_q,$$

  where where q is the number of constraints imposed by H_0

- For details see van der Vaart (1998, ch. 16)

Comparison of Test Statistics

- The score, Wald, and likelihood ratio test statistics are asymptotically equivalent but are not equally well-behaved in finite samples
- An advantage of the Wald statistic is that confidence intervals can be derived directly from the statistic and so there is a direct link between estimation and testing. Interpretation is also more straightforward
- A major drawback of the Wald statistic is that it is not invariant under reparameterization
- The score test statistic is invariant under reparameterization, provided the expected, rather than the observed, information is used
- The likelihood ratio statistic is invariant under reparameterization

Comparison of Test Statistics

- The score statistic requires the value of the score at the null but the MLE is not required
- If $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_0$ are close then the three statistics will tend to agree
- Confidence intervals derived from likelihood ratio tests always preserve the support of the parameter, unlike score- and Wald-based intervals

Likelihood Inference for GLMs: Deviance

- The *deviance* is a commonly used measure to assess the fit of a GLM and to compare models
- Main idea: compare your model's fit versus a model with perfect fit
- A model with a perfect fit to the data (and therefore, with the largest log-likelihood) is obtained by creating an indicator covariate per observation, that is, taking $\boldsymbol{X} = \boldsymbol{I}_n$
- Such a model is called the *saturated model*, and it involves a vector $\boldsymbol{\beta}$ of length n
- The *deviance* measures if your model has a fit to the data similar to that of the saturated model

Likelihood Inference for GLMs: Deviance

- Remember that, regardless of the model, we can write the natural parameters as a function of $\boldsymbol{\beta}$: $\theta_i[\mu_i(\boldsymbol{\beta})]$

- We write $\tilde{\boldsymbol{\theta}} = [\tilde{\theta}_1, ..., \tilde{\theta}_n]$ to represent the MLEs under the saturated model

- Similarly, let $\hat{\boldsymbol{\theta}} = [\hat{\theta}_1, ..., \hat{\theta}_n]$ denote the MLEs under a reduced model containing $k + 1$ parameters

- The *deviance* can be obtained as a log-likelihood ratio statistic for H_0 : Reduced model versus H_1 : Saturated model:

$$-2\left[l(\hat{\boldsymbol{\theta}}) - l(\tilde{\boldsymbol{\theta}})\right] = \frac{2}{\alpha}\sum_{i=1}^{n}\phi_i\left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\right] = \frac{D}{\alpha},$$

where D is known as the *deviance* (associated with the saturated model) and D/α is the *scaled deviance*

Likelihood Inference for GLMs: Deviance

- The greater the deviance, the poorer the fit of the model
- The main use of the deviance is for inferential comparisons of models with fixed number of parameters
- The deviance leads to an attractive additivity property of the likelihood ratio test statistic for nested models

Likelihood Inference for GLMs: Deviance

- Consider two models M_0 and M_1 with q_0 and q_1 parameters, where M_0 can be seen as a special case of M_1
- Consider testing $H_0 : M_0$ versus $H_1 : M_1$
- Let $\hat{\boldsymbol{\beta}}_0$, $\hat{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}$ represent the MLEs of $\boldsymbol{\beta}$ under the null, alternative and saturated models, respectively
- Under H_0:

$$\begin{aligned} -2\left[l(\hat{\boldsymbol{\beta}}_0) - l(\hat{\boldsymbol{\beta}}_1)\right] &= 2\left\{l(\tilde{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}}_0) - [l(\tilde{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}}_1)]\right\} \\ &= \frac{1}{\alpha}(D_0 - D_1) \;\rightarrow_d\; \chi^2_{q_1 - q_0}, \end{aligned}$$

where D_j is the deviance representing the fit under hypothesis j, relative to the saturated model, $j = 0, 1$.

Likelihood Inference for GLMs: Pearson X^2

- An alternative statistic for testing $H_0 : M_0$ versus $H_1 : M_1$ is the Pearson statistic

$$X^2 = \sum_{i=1}^{n} \phi_i \frac{(\hat{\mu}_{1,i} - \hat{\mu}_{0,i})^2}{\hat{\alpha} V(\hat{\mu}_{0,i})},$$

  where $\hat{\mu}_{j,i}$ is the fitted mean under model j

- Similarly as with the likelihood ratio test, $X^2 \to_d \chi^2_{q_1 - q_0}$ under H_0

- The Pearson statistic does not share the additivity property of the deviance

Assessment of Assumptions for GLMs

We need to specify the distribution of the response, the link function, in addition to the linear predictor form

- There is often a default distribution and link function based on the type of response, so these choices are often based on convenience, but are they OK for our data?
- To specify the mean model, in an initial data exploration, it is common to plot the response, transformed to the linear predictor scale, against covariates
- For example, with log link, one may plot $\log y$ versus covariates x, although we know $\mathrm{E}(\log Y \mid \boldsymbol{x}) \neq \log \mathrm{E}(Y \mid \boldsymbol{x})$

Assessment of Assumptions for GLMs

It is common to start with a default specification of the GLM, and then find evidence against its assumptions, using residuals

- With GLMs, the definition of a residual is more ambiguous, for discrete outcomes in particular
- Various attempts exist to provide a general definition of residuals that possess zero mean, constant variance and a symmetric distribution

Assessment of Assumptions for GLMs

- The obvious definition of a residual is

$$e_i = y_i - \hat{\mu}_i$$

but clearly in a GLM such residuals will generally have unequal variances, so that some form of standardization is required

- Pearson residuals, are defined as:

$$e_i^\star = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\text{var}}(Y_i)}} = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}_i},$$

where $\hat{\text{var}}(Y_i) = \hat{\alpha} V(\hat{\mu}_i)/\phi_i$, and $\hat{\mu}_i$ are the fitted values from the model

- Squaring and summing these residuals reproduces Pearson's X^2 statistic seen before:

$$X^2 = \sum_{i=1}^{n} e_i^{\star 2}$$

Assessment of Assumptions for GLMs

- Deviance residuals are given by

$$e_i^{\dagger} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{D_i}$$

where $D_i = 2\phi_i \left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right]$ where $\tilde{\theta}_i$ is obtained under the saturated model, so that $D = \sum_{i=1}^{n} e_i^{\dagger 2}$.

- For discrete data with small means, residuals are difficult to interpret since the response can only take on a small number of discrete values

- One strategy to aid in interpretation is to simulate data with the same covariate matrix $\boldsymbol{X}$, and under the parameter estimates from the fitted model

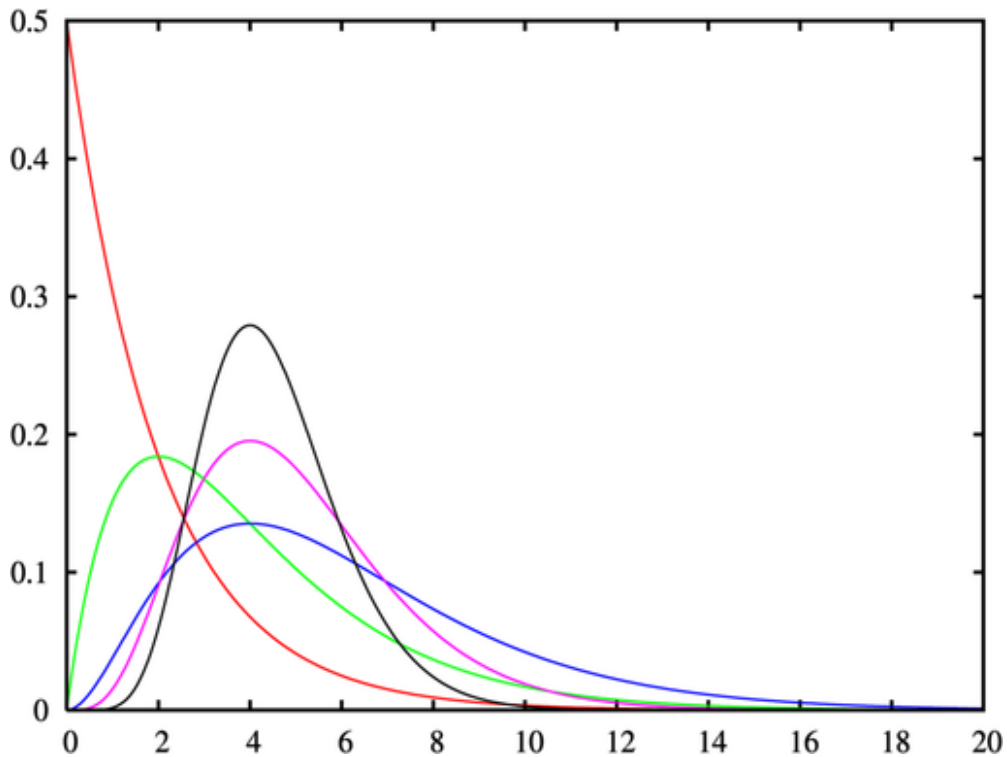
- One may then examine residual plots to see their form when the model is correct

Assessment of Assumptions for GLMs

- As with linear model residuals, Pearson or deviance residuals can be plotted against covariates to suggest possible model forms
- They may also be plotted against fitted values, or some function of the fitted values, to access mean-variance relationships
- If the spread is not constant then this suggests that the assumed mean-variance relationship is not correct

Example: The Gamma GLM

The gamma distribution is suitable for modeling outcomes that are continuous, positive, and potentially right-skewed.



Some possible shapes of the gamma density. Modified from `https://en.wikipedia.org/wiki/Gamma_distribution`

Example: The Gamma GLM

The gamma GLM arises from assuming:

- Response follows a gamma distribution:

$$Y_i \mid \boldsymbol{x}_i \sim \mathit{Gamma}[\, \mu_i, \alpha_i],$$

where μ_i is the mean and α_i is the dispersion parameter for observation i

- Dependence on covariates only through the mean:

$$\mu_i = \mu(\boldsymbol{x}_i), \quad \alpha_i = \alpha$$

- Linearity of the regression function on a transformed scale:

$$g(\mu_i) = g[\mu(\boldsymbol{x}_i)] = \boldsymbol{x}_i\boldsymbol{\beta},$$

where the most commonly used link function is $g(\cdot) = \log(\cdot)$, that is,

$$\log \mu(\boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$$

Example: The Gamma GLM

- A common parameterization of the gamma density is

$$f(y \mid u, v) = \frac{v^u}{\Gamma(u)} y^{u-1} \exp(-vy)$$

such that $\mathrm{E}(Y) = u/v$ and $\mathrm{var}(Y) = u/v^2$

- Taking $\mu = u/v$ and $\alpha = 1/u$ leads to $u = 1/\alpha$ and $v = 1/\alpha\mu$

$$f(y \mid \mu, \alpha) = \frac{(1/\alpha\mu)^{1/\alpha}}{\Gamma(1/\alpha)} y^{1/\alpha-1} \exp(-y/\alpha\mu)$$

such that $\mathrm{E}(Y) = \mu$ and

$$\mathrm{var}(Y) = \alpha\mu^2 = \alpha V(\mu)$$

for $V(\mu) = \mu^2$

Example: The Gamma GLM

We can write the density in exponential dispersion family form:

$$
\begin{aligned}
f(y \mid \mu, \alpha) &= \frac{(1/\alpha\mu)^{1/\alpha}}{\Gamma(1/\alpha)} y^{1/\alpha - 1} \exp(-y/\alpha\mu) \\
&= \exp\left\{ \frac{y(-1/\mu) - [-\log(-[-1/\mu])]}{\alpha} - \frac{\log(y/\alpha)}{\alpha} - \log[y\Gamma(1/\alpha)] \right\} \\
&= \exp\left\{ \frac{y\theta - b(\theta)}{\alpha} + c(y, \alpha) \right\},
\end{aligned}
$$

where

- The natural parameter is $\theta = -1/\mu$
- The dispersion parameter is α
- The log-partition function is $b(\theta) = -\log(-\theta)$

Example: The Gamma GLM

- Given n independent pairs $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^n$, and the log link, we can directly write down the likelihood function

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} f(y_i \mid \mu(\boldsymbol{x}_i), \alpha) \\ &= \prod_{i=1}^{n} \frac{(1/\alpha \exp[\boldsymbol{x}_i \boldsymbol{\beta}])^{1/\alpha}}{\Gamma(1/\alpha)} y_i^{1/\alpha - 1} \exp(-y_i/\alpha \exp[\boldsymbol{x}_i \boldsymbol{\beta}]) \end{aligned}$$

  since $\mu(\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i \boldsymbol{\beta})$

- We could take this as the starting point for deriving all the pieces we need to obtain MLEs and inferences (score, the Fisher info., IRLS)

- However, we can directly use our general formulae for GLMs!

Example: The Gamma GLM

Before that, can we ensure a unique maximizer of the log-likelihood? The MLE of $\boldsymbol{\beta}$ is obtained from

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} (1/\alpha) \sum_{i=1}^{n} [\boldsymbol{x}_i\boldsymbol{\beta} - y_i \exp(-\boldsymbol{x}_i\boldsymbol{\beta})] + K(\boldsymbol{y}, \alpha)$$

$$= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{\beta} - y_i \exp(-\boldsymbol{x}_i\boldsymbol{\beta})$$

where the latter is a concave function of $\boldsymbol{\beta}$ since

- $\boldsymbol{x}_i\boldsymbol{\beta}$ is an affine transformation of $\boldsymbol{\beta}$
- $-y_i \exp(\eta_i)$ is a concave function of η_i, since $\exp(\cdot)$ is convex and $y_i > 0$
- $-y_i \exp(-\boldsymbol{x}_i\boldsymbol{\beta})$ is a concave function of $\boldsymbol{\beta}$ from the concavity of the composition of concave and affine functions
- Sum of concave functions is concave

We conclude that the gamma GLM with log link leads to a unique maximum

Example: The Gamma GLM

A central quantity of interest is the score

$$\boldsymbol{S}(\boldsymbol{\beta}) \quad = \quad \boldsymbol{D}^T \boldsymbol{V}^{-1} \left[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right] / \alpha$$

where

- $$\boldsymbol{V} = \text{diag}\{V(\mu_i)/\phi_i\} = \text{diag}\{\exp(2\boldsymbol{x}_i\boldsymbol{\beta})\}$$

  since $\phi_i = 1$ and $V(\mu_i) = \mu_i^2 = \mu(\boldsymbol{x}_i)^2 = [\exp(\boldsymbol{x}_i\boldsymbol{\beta})]^2$

- $$\boldsymbol{D} = \text{diag}\{\exp(\boldsymbol{x}_i\boldsymbol{\beta})\}\boldsymbol{X}$$

  since $[\boldsymbol{D}]_{i,j} = \partial\mu_i/\partial\beta_j$, where $\mu_i = \exp(\eta_i)$ with $\eta_i = \boldsymbol{x}_i\boldsymbol{\beta}$, and therefore

  $$\frac{\partial\mu_i}{\partial\boldsymbol{\beta}} = \frac{d\mu_i}{d\eta_i}\frac{\partial\eta_i}{\partial\boldsymbol{\beta}} = \frac{d\mu_i}{d\eta_i}\boldsymbol{x}_i^T = \exp(\boldsymbol{x}_i\boldsymbol{\beta})\boldsymbol{x}_i^T$$

Example: The Gamma GLM

Therefore we obtain the score for the gamma GLM with log link as

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{\beta}) &= \boldsymbol{D}^T \boldsymbol{V}^{-1} \left[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right] / \alpha \\ &= \boldsymbol{X}^T \text{diag}\{\exp(-\boldsymbol{x}_i \boldsymbol{\beta})\} \left[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right] / \alpha \\ &= \sum_{i=1}^{n} \frac{y_i - \mu_i(\boldsymbol{\beta})}{\alpha \mu_i(\boldsymbol{\beta})} \boldsymbol{x}_i^T \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\mu}(\boldsymbol{\beta}) &= [\mu_1(\boldsymbol{\beta}), \ldots, \mu_n(\boldsymbol{\beta})]^T \\ &= [\exp(\boldsymbol{x}_1 \boldsymbol{\beta}), \ldots, \exp(\boldsymbol{x}_n \boldsymbol{\beta})]^T \end{aligned}$$

Example: The Gamma GLM

The expected Fisher information is given by

$$\begin{aligned}
\mathcal{I}_n(\boldsymbol{\beta}) &= \boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{\beta}) \boldsymbol{X} \\
&= \boldsymbol{X}^T \text{diag} \left\{ \frac{\phi_i (d\mu_i/d\eta_i)^2}{\alpha V(\mu_i)} \right\} \boldsymbol{X} \\
&= \boldsymbol{X}^T \boldsymbol{X}/\alpha
\end{aligned}$$

since we had seen that

$$\begin{aligned}
\phi_i &= 1, \\
d\mu_i/d\eta_i &= \exp(\eta_i) = \exp(\boldsymbol{x}_i \boldsymbol{\beta}), \\
V(\mu_i) &= \exp(2\boldsymbol{x}_i \boldsymbol{\beta})
\end{aligned}$$

Example: The Gamma GLM

We denote

- $\hat{\boldsymbol{\theta}}$: estimated natural parameters under a given model
- $\tilde{\boldsymbol{\theta}}$: estimated natural parameters under the saturated model

Since $\theta_i = -1/\mu_i$ and $b(\theta_i) = -\log(-\theta_i) = \log(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta}$, the scaled deviance is

$$\begin{aligned}\frac{D}{\alpha} &= -2\left[l(\hat{\boldsymbol{\theta}}) - l(\tilde{\boldsymbol{\theta}})\right] \\ &= \frac{2}{\alpha}\sum_{i=1}^{n}\phi_i\left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\right] \\ &= \frac{2}{\alpha}\sum_{i=1}^{n}\left[y_i/\hat{\mu}_i - 1 + \log(\hat{\mu}_i/y_i)\right]\end{aligned}$$

with $\hat{\mu}_i = \exp(\boldsymbol{x}_i\hat{\boldsymbol{\beta}})$ and $\tilde{\mu}_i = y_i$

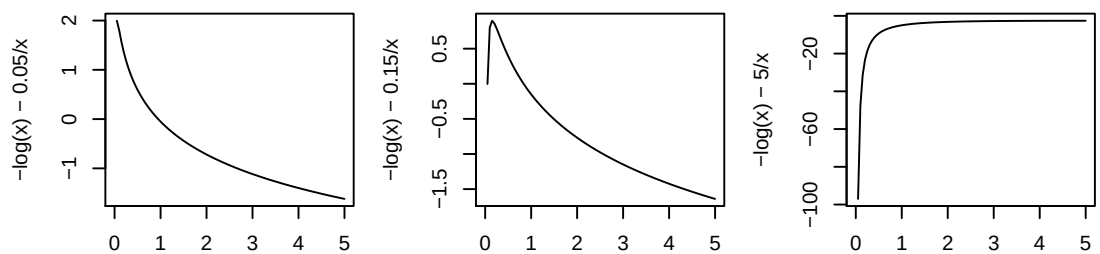
Example: The Gamma GLM

How about a gamma GLM with identity link, i.e. $\mu(\boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$?

- If the Y_i's are away from zero, using the identity link might be reasonable as it will usually hold that $\hat{\mu}(\boldsymbol{x}_i) = \boldsymbol{x}_i\hat{\boldsymbol{\beta}} > 0$ *in the observed range of the covariates*

- It is easy to see that with the identity link the MLE is obtained from

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} -\log(\boldsymbol{x}_i\boldsymbol{\beta}) - y_i/\boldsymbol{x}_i\boldsymbol{\beta}$$



- There might be numerical issues if the Y_i's are close to zero, but in that case it doesn't make sense to use this approach

Example: The Bernoulli GLM

- What is a Bernoulli random variable?
- If we take

$$Y = I(\text{event A occurs}),$$

then

$$P(A) = P(Y = 1) := p, \quad P(A^c) = P(Y = 0) := 1 - p,$$

which can be summarized as

$$P(Y = y \mid p) := p^y (1 - p)^{1-y}, \quad y \in \{0, 1\}$$

- Therefore, a binary variable *has to* follow a Bernoulli distribution

Example: The Bernoulli GLM

- Given n independent pairs $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^n$, where $Y_i \in \{0, 1\}$ we usually have

$$Y_i = I(\text{event or characteristic A occurs for unit } i),$$

then

$$P(Y_i = 1) := p_i, \quad P(Y_i = 0) := 1 - p_i,$$

which can be summarized as

$$P(Y_i = y_i \mid p_i) := p_i^{y_i}(1 - p_i)^{1-y_i}, \quad y_i \in \{0, 1\}$$

- Therefore, when $Y_i \in \{0, 1\}$, Bernoulli is the *correct* distribution
- Until now, the only assumption is that we have n *independent* pairs
- Since $E(Y_i) = P(Y_i = 1)$, denote $\mu_i = p_i$ from now on

Example: The Bernoulli GLM

How is a Bernoulli GLM parametric?

- The means (probabilities) are determined by fully observed covariates:

$$\mu_i = \mu(\boldsymbol{x}_i)$$

- Probabilities (regression function) are a linear function of $\boldsymbol{\beta}$ on a transformed scale:

$$g(\mu_i) = g[\mu(\boldsymbol{x}_i)] = \boldsymbol{x}_i\boldsymbol{\beta},$$

Example: The Bernoulli GLM

We can write the density in exponential dispersion family form:

$$\begin{aligned} f(y_i \mid \mu_i, \alpha_i) &= \mu_i^{y_i}(1-\mu_i)^{1-y_i} \\ &= \exp\left\{ y_i \log \frac{\mu_i}{1-\mu_i} + \log(1-\mu_i) \right\}, \end{aligned}$$

where

- The natural parameter is $\theta_i = \log[\mu_i/(1-\mu_i)]$
- The dispersion parameter is $\alpha = 1$
- The log-partition function is $b(\theta_i) = \log[1+\exp(\theta_i)]$

and the GLM specification comes from $\mu_i = g^{-1}(\boldsymbol{x}_i\boldsymbol{\beta})$

Example: The Bernoulli GLM

- The score, Fisher information, etc, can be easily obtained from the general expressions, which will depend on the link function (e.g., logit, probit, etc)

- The only part that deserves especial attention is the deviance:

$$D = -2\left[l(\hat{\boldsymbol{\theta}}) - l(\tilde{\boldsymbol{\theta}})\right] = -2\sum_{i=1}^{n}\left[l_i(\hat{\theta}_i) - l_i(\tilde{\theta}_i)\right]$$

$$= 2\sum_{i=1}^{n}\left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\right]$$

where $\tilde{\theta}_i$ is not well defined as $\tilde{\theta}_i = \log[\tilde{\mu}_i/(1-\tilde{\mu}_i)]$ with $\tilde{\mu}_i = y_i$

- Instead, we note that $l_i(\tilde{\theta}_i) = y_i \log \tilde{\mu}_i + (1-y_i)\log(1-\tilde{\mu}_i) = 0$ and obtain

$$D = -2l(\hat{\boldsymbol{\theta}}) \;=\; -2\sum_{i=1}^{n}\left[y_i\hat{\theta}_i - b(\hat{\theta}_i)\right]$$

$$= -2\sum_{i=1}^{n}\left[y_i \log \hat{\mu}_i + (1-y_i)\log(1-\hat{\mu}_i)\right]$$

Example: The Binomial GLM

- *Reminder*: if $Z_1, \ldots, Z_N$ are independent and distributed as *Bernoulli*$[p]$ (common probability p) then

$$Y = \sum_{i=1}^{N} Z_i \sim \mathit{Binomial}[N, p]$$

- The binomial distribution is commonly used for *grouped binary data*
  - Unit i is a group of N_i individuals
  - Y_i measures the number individuals that have an event

Example: The Binomial GLM

How do grouped binary data arise?

- Predefined groups and covariates measured at the group level
  - e.g., i: cat litter; N_i: kittens in the litter, Y_i: number of kittens that survive first month, $\boldsymbol{x}_i$: characteristics of the mother
- Binary responses with repeated values of covariates
  - Say $Z_i \sim Bernoulli[p(\boldsymbol{x}_i)]$, and the covariate space takes M fixed different values $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(M)}$
  - This defines M groups, each of size $N_m = \sum_{i=1}^{n} I(\boldsymbol{x}_i = \boldsymbol{x}_{(m)})$, and within each we have

$$Y_m = \sum_{i:\boldsymbol{x}_i = \boldsymbol{x}_{(m)}} Z_i \sim Binomial[N_m, p(\boldsymbol{x}_{(m)})]$$

  - This can happen especially if the covariates are categorical

Example: The Binomial GLM

- Assuming
$$Y_i \mid N_i, p_i \sim Binomial[N_i, p_i],$$
leads to $\mathrm{E}(Y_i) = N_i p_i$

- We are mainly interested in modeling the probabilities p_i as a function of covariates

- It is common to derive the binomial GLM taking the outcome variable as
$$Y_i' = Y_i / N_i, \quad \text{so that} \quad \mathrm{E}(Y_i') = p_i$$

- We denote $\mu_i = p_i = \mathrm{E}(Y_i')$ to go back to the usual notation

Example: The Binomial GLM

We can write the density for $Y_i' = Y_i/N_i$ in exponential dispersion family form as:

$$f(y_i' \mid \mu_i, \alpha_i) = \binom{N_i}{N_i y_i'} \mu_i^{N_i y_i'} (1-\mu_i)^{N_i - N_i y_i'}$$

$$= \exp\left\{ \frac{y_i' \log \frac{\mu_i}{1-\mu_i} + \log(1-\mu_i)}{1/N_i} + \log \binom{N_i}{N_i y_i'} \right\},$$

where

- The natural parameter is $\theta_i = \log[\mu_i/(1-\mu_i)]$
- The dispersion parameter is $\alpha_i = 1/N_i$ (i.e., $\alpha = 1$, $\phi_i = N_i$)
- The log-partition function is $b(\theta_i) = \log[1 + \exp(\theta_i)]$

The GLM is fully specified by a link function $g(\cdot)$ and a linear predictor $\boldsymbol{x}_i\boldsymbol{\beta}$ so that

$$g(\mu_i) = g[\mu(\boldsymbol{x}_i)] = \boldsymbol{x}_i\boldsymbol{\beta}$$

Example: The Binomial GLM

Grouped vs ungrouped binary data?

- Say we originally have binary responses, what should we use?
  - Bernoulli GLM with the original binary responses
  - Binomial GLM with responses summarized according to the different values of the covariates
- MLEs and Fisher information will be identical under both cases
- Differences arise in the deviance, and reflect philosophical interpretations of asymptotics

Example: The Binomial GLM

How should we think about asymptotics?

- With grouped binary data obtained from predefined groups, *more observations means more groups*
  - "Asymptotically" means, e.g., the number of cat litters $\to \infty$
- With individual binary responses, *more observations means more individuals*, however, what happens as we observe more individuals?

a) If expect to increasingly observe new values of covariates (e.g. with continuous covariates) $\implies$ stick to Bernoulli GLM

b) If the combination of covariate values takes a finite number and stays fixed as we obtain new individuals (e.g. with categorical covariates) $\implies$ move to binomial GLM

Example: The Binomial GLM

- If the covariate space takes M fixed different values $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(M)}$, we can organize the data in a $2 \times M$ contingency table

	$\boldsymbol{x}_{(1)}$	$\boldsymbol{x}_{(2)}$	$\ldots$	$\boldsymbol{x}_{(M)}$
"Successes"	y_1	y_2	$\ldots$	y_M
"Failures"	$N_1 - y_1$	$N_2 - y_2$	$\ldots$	$N_M - y_M$

- In this case, the *saturated model* involves M parameters, one per unique value of the covariates, and it returns the observed cell values as its fitted values

- The deviance for a given model can be written as

$$
\begin{aligned}
D &= -2\left[l(\hat{\boldsymbol{\theta}}) - l(\tilde{\boldsymbol{\theta}})\right] \\
&= 2\sum_{m=1}^{M} y_m \log \frac{y_m}{N_m \hat{\mu}_m} + 2\sum_{m=1}^{M} (N_m - y_m) \log \frac{N_m - y_m}{N_m(1-\hat{\mu}_m)}
\end{aligned}
$$

where $\hat{\mu}_m = g^{-1}(\boldsymbol{x}_{(m)}\hat{\boldsymbol{\beta}})$ obtained from the model

Example: The Binomial GLM

- Observed table:

	$\boldsymbol{x}_{(1)}$	$\boldsymbol{x}_{(2)}$	$\cdots$	$\boldsymbol{x}_{(M)}$
"Successes"	y_1	y_2	$\cdots$	y_M
"Failures"	$N_1 - y_1$	$N_2 - y_2$	$\cdots$	$N_M - y_M$

- Fitted table:

	$\boldsymbol{x}_{(1)}$	$\boldsymbol{x}_{(2)}$	$\cdots$	$\boldsymbol{x}_{(M)}$
"Successes"	$N_1\hat{\mu}_1$	$N_2\hat{\mu}_2$	$\cdots$	$N_M\hat{\mu}_M$
"Failures"	$N_1(1-\hat{\mu}_1)$	$N_2(1-\hat{\mu}_2)$	$\cdots$	$N_M(1-\hat{\mu}_M)$

- The deviance for the fitted model can be written as

$$D = 2\sum_{m=1}^{M} y_m \log\frac{y_m}{N_m\hat{\mu}_m} + 2\sum_{m=1}^{M}(N_m - y_m)\log\frac{N_m - y_m}{N_m(1-\hat{\mu}_m)}$$

$$= 2\sum_{\text{cells}} \text{observed} \times \log\left(\frac{\text{observed}}{\text{fitted}}\right)$$

Example: The Binomial GLM

- In this case, since the saturated model has a finite number of parameters M, the deviance D can be used to test the overall fit of the model with $k+1$ parameters
  - H_0: the data follows the model with $k+1$ parameters
  - H_1: the data follows the saturated model
- The test is based on

$$D \to \chi^2_{M-k-1}$$

where the asymptotics assume that the observed counts in the cells grow to infinity, referred to as *small-dispersion asymptotics* [5]

[5] See Agresti (2015, ch. 4)

Example: The Binomial GLM

- The test can also be conducted based on a Pearson X^2 statistic

$$\begin{aligned} X^2 &= \sum_{m=1}^{M} \frac{(y_m - N_m\hat{\mu}_m)^2}{N_m\hat{\mu}_m(1-\hat{\mu}_m)} \\ &= \sum_{m=1}^{M} \frac{(y_m - N_m\hat{\mu}_m)^2}{N_m\hat{\mu}_m} + \sum_{m=1}^{M} \frac{[(N_m - y_m) - N_m(1-\hat{\mu}_m)]^2}{N_m(1-\hat{\mu}_m)} \\ &= \sum_{\text{cells}} \frac{(\text{observed} - \text{fitted})^2}{\text{fitted}} \end{aligned}$$

- We also have that

$$X^2 \to \chi^2_{M-k-1}$$

  which again relies on *small-dispersion asymptotics*

- X^2 is known to reach its asymptotic distribution faster than D

Example: The Multinomial GLM

The multinomial GLM constitutes an extension of the Bernoulli and binomial GLM

- Let each unit i have a categorical response taking $c \geq 2$ categories
- Let

$$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ic})$$

where

$$Y_{ij} = I(\text{unit } i\text{'s response belongs to category } j)$$

so that $\sum_{j=1}^{c} Y_{ij} = 1$

- Let $p_{ij} = P(Y_{ij} = 1)$ and $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{ic})$
- We seek to model $\boldsymbol{p}_i$ as a function of covariates, that is, $\boldsymbol{p}_i = \boldsymbol{p}(\boldsymbol{x}_i)$

Example: The Multinomial GLM

- The most common multinomial GLM uses $c-1$ logits and a baseline category

$$\log \frac{p_{i1}}{p_{ic}}, \quad \log \frac{p_{i2}}{p_{ic}}, \quad \ldots, \quad \log \frac{p_{i,c-1}}{p_{ic}}$$

where

$$\log \frac{p_{ij}}{p_{ic}} = \text{logit}[P(Y_{ij} = 1 \mid Y_{ij} = 1 \text{ or } Y_{ic} = 1)]$$

- The *multinomial logit model* is specified taking

$$\log \frac{p_{ij}}{p_{ic}} = \boldsymbol{x}_i \boldsymbol{\beta}_j, \quad j = 1, \ldots, c-1$$

where $\boldsymbol{\beta}_j = (\beta_{j0}, \ldots, \beta_{jk})^T$, so that now there are $(k+1)(c-1)$ parameters

Example: The Multinomial GLM

- Comparisons for other pairs of categories can be obtained easily since

$$\log \frac{p_{ij}}{p_{ij'}} = \log \frac{p_{ij}}{p_{ic}} - \log \frac{p_{ij'}}{p_{ic}} = \boldsymbol{x}_i(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'})$$

- The probabilities are obtained as

$$P(Y_{ij} = 1 \mid \boldsymbol{x}_i) = p_{ij} = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta}_j)}{1 + \sum_{h=1}^{c-1} \exp(\boldsymbol{x}_i\boldsymbol{\beta}_h)}$$

  where for the baseline category we have $\boldsymbol{\beta}_c = \mathbf{0}$

- For more details see Agresti (2015, ch. 6)

Example: The Poisson GLM

The Poisson GLM is commonly used for count outcomes. These counts commonly arise in two ways:

- Each unit has an associated count
  - e.g., i: US voter; Y_i: number of alcoholic drinks during election night; $\boldsymbol{x}_i$: political affiliation and demographic covariates
- Units are cross-classified in a contingency table according to categorical features, where each combination of categories leads to a count
  - e.g., a contingency table containing the population (or a sample) of a county cross-classified by sex, educational status, religious and political affiliations

Example: The Poisson GLM

The most common Poisson GLM is a *log-linear model*

- $Y_i \mid \boldsymbol{x}_i \sim \mathit{Poisson}(\mu_i)$
- $\mu_i = \mu(\boldsymbol{x}_i)$
- $\log \mu(\boldsymbol{x}_i) = \boldsymbol{x}_i \boldsymbol{\beta}$

This is a GLM with canonical link, so we obtain the simplified version of the score equations, Fisher information, etc

Example: The Poisson GLM

- The deviance for a Poisson GLM can be written as

$$D = -2\left[l(\hat{\boldsymbol{\theta}}) - l(\tilde{\boldsymbol{\theta}})\right]$$
$$= 2\sum_{i=1}^{n}\left[y_i \log\frac{y_i}{\hat{\mu}_i} - y_i + \hat{\mu}_i\right]$$

  where $\hat{\mu}_i = g^{-1}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}})$ obtained from the model

- When we use a log-linear model with an intercept, we obtain from the score equations $\sum_{i=1}^{n}\hat{\mu}_i = \sum_{i=1}^{n} y_i$, and the deviance simplifies to

$$D = 2\sum_{i=1}^{n} y_i \log\frac{y_i}{\hat{\mu}_i} = 2\sum_{\text{units}} \text{observed} \times \log\left(\frac{\text{observed}}{\text{fitted}}\right)$$

Example: The Poisson GLM

- The Pearson X^2 statistic can also be defined here as

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} = \sum_{\text{units}} \frac{(\text{observed} - \text{fitted})^2}{\text{fitted}}$$

- If n is fixed, and the means $\mu_i \to \infty$ then both D and X^2 converge to χ^2_{n-k-1} where $k+1$ is the length of $\boldsymbol{\beta}$ in the GLM

- This is again *small-dispersion asymptotics* and it's a reasonable set-up when modeling contingency tables

Example: The Poisson GLM

The log-linear model is easily adapted to model rates

- Let Y_i be the number of events that occur for unit i over a period of time t_i (or some other measure of exposure)
- We expect the larger t_i the larger Y_i, so we want to account for this by modeling the rates Y_i/t_i
- Conditioning on the t_i's, we can write a log-linear model for the rates as

$$\log \mathrm{E}(Y_i/t_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$$

from which

$$\log \mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta} + \log t_i$$

which indicates that modeling rates can be accomplished by including $\log t_i$ as an *offset* [6] in the linear predictor

[6] *Offset*: a covariate for which you do not estimate a coefficient

Example: The Poisson GLM

The Poisson GLM can also be used as an approximation for the binomial GLM

- The $binomial(N, p)$ converges to the $Poisson(\mu)$ if $N \to \infty$ and $p \to 0$ such that $Np = \mu$ is fixed
- This supports using the Poisson GLM when the groups sizes N_i are large and the probabilities of the events p_i are very small
  - e.g., i: county; N_i: number of eligible voters in county; Y_i: votes for the 2020 Libertarian Party (or Green Party) presidential nominee; $\boldsymbol{x}_i$: county-level information; $\log N_i$: offset

Example: The Poisson GLM for Contingency Tables

Another context where a Poisson log-linear model appears is in the *log-linear analysis of contingency tables*

- Let U, V, W be three categorical variables taking J, K, L categories
- Let $\{U_i, V_i, W_i\}_{i=1}^n$ be a random sample from their joint distribution
- Let

$$n_{jkl} = \sum_{i=1}^{n} I(U_i = j, V_i = k, W_i = l)$$

		W		
U	V	1	...	L
1	1	n_{111}	...	n_{11L}
	⋮			
	K	n_{1K1}	...	n_{1KL}
⋮	⋮			
J	1	n_{J11}	...	n_{J1L}
	⋮			
	K	n_{JK1}	...	n_{JKL}

Example: The Poisson GLM for Contingency Tables

For example

		Political Affiliation	
Sex	Educational Level	Republican	Democrat
Male	$<$ High School	n_{111}	n_{112}
	High School	n_{121}	n_{122}
	$\geq$ College	n_{131}	n_{132}
Female	$<$ High School	n_{211}	n_{212}
	High School	n_{221}	n_{222}
	$\geq$ College	n_{231}	n_{232}

Example: The Poisson GLM for Contingency Tables

Interest lies in modeling the joint distribution of U, V, W

- Let

$$\mathrm{E}(n_{jkl}) = \mu_{jkl} = n\pi_{jkl},$$

where

$$\pi_{jkl} = P(U_i = j, V_i = k, W_i = l)$$

- Assuming the counts are independent, they are modeled as

$$n_{jkl} \sim Poisson(\mu_{jkl})$$

- Models for the joint distribution of U, V, W impose different conditional independence and association structures on the π_{jkl}

- These models can be written in log-linear form

Example: The Poisson GLM for Contingency Tables

- Under the *mutual independence model*, $U \perp\!\!\!\perp V \perp\!\!\!\perp W$, and

$$\pi_{jkl} = \pi_{j++}\pi_{+k+}\pi_{++l}$$

where $\pi_{j++} = \sum_{k,l} \pi_{jkl}$, likewise for π_{+k+} and π_{++l}

- This implies that we can write

$$\begin{aligned} \log \mu_{jkl} &= \log n + \log \pi_{j++} + \log \pi_{+k+} + \log \pi_{++l} \\ &:= \lambda + \lambda_j^U + \lambda_k^V + \lambda_l^W \end{aligned}$$

for some parameters $\lambda, \lambda_j^U, \lambda_k^V, \lambda_l^W$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, $l = 1, \ldots, L$ subject to constraints, e.g.,

$$\lambda_J^U = \lambda_K^V = \lambda_L^W = 0$$

Example: The Poisson GLM for Contingency Tables

Note that we can write

$$\log \mu_{jkl} := \lambda + \lambda_j^U + \lambda_k^V + \lambda_l^W,$$

with $\lambda_J^U = \lambda_K^V = \lambda_L^W = 0$ as $\log(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\lambda}$:

$$\begin{pmatrix} \log \mu_{111} \\ \log \mu_{211} \\ \vdots \\ \log \mu_{J11} \\ \log \mu_{121} \\ \log \mu_{221} \\ \vdots \\ \log \mu_{J21} \\ \vdots \\ \log \mu_{1KL} \\ \log \mu_{2KL} \\ \vdots \\ \log \mu_{JKL} \end{pmatrix} = \begin{pmatrix} \lambda + \lambda_1^U + \lambda_1^V + \lambda_1^W \\ \lambda + \lambda_2^U + \lambda_1^V + \lambda_1^W \\ \vdots \\ \lambda + \lambda_1^V + \lambda_1^W \\ \lambda + \lambda_1^U + \lambda_2^V + \lambda_1^W \\ \lambda + \lambda_2^U + \lambda_2^V + \lambda_1^W \\ \vdots \\ \lambda + \lambda_2^V + \lambda_1^W \\ \vdots \\ \lambda + \lambda_1^U \\ \lambda + \lambda_2^U \\ \vdots \\ \lambda \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1\,0\,\ldots\,0 & 1\,0\,\ldots\,0 & 1\,0\,\ldots\,0 \\ 1 & 0\,1\,\ldots\,0 & 1\,0\,\ldots\,0 & 1\,0\,\ldots\,0 \\ & & \vdots & \\ 1 & 0\,0\,\ldots\,0 & 1\,0\,\ldots\,0 & 1\,0\,\ldots\,0 \\ 1 & 1\,0\,\ldots\,0 & 0\,1\,\ldots\,0 & 1\,0\,\ldots\,0 \\ 1 & 0\,1\,\ldots\,0 & 0\,1\,\ldots\,0 & 1\,0\,\ldots\,0 \\ & & \vdots & \\ 1 & 0\,0\,\ldots\,0 & 0\,1\,\ldots\,0 & 1\,0\,\ldots\,0 \\ & & \vdots & \\ 1 & 1\,0\,\ldots\,0 & 0\,0\,\ldots\,0 & 0\,0\,\ldots\,0 \\ 1 & 0\,1\,\ldots\,0 & 0\,0\,\ldots\,0 & 0\,0\,\ldots\,0 \\ & & \vdots & \\ 1 & 0\,0\,\ldots\,0 & 0\,0\,\ldots\,0 & 0\,0\,\ldots\,0 \end{pmatrix}}_{1+(J-1)+(K-1)+(L-1)\text{columns}} \begin{pmatrix} \lambda \\ \lambda_1^U \\ \lambda_2^U \\ \vdots \\ \lambda_{J-1}^U \\ \lambda_1^V \\ \lambda_2^V \\ \vdots \\ \lambda_{K-1}^V \\ \lambda_1^W \\ \lambda_2^W \\ \vdots \\ \lambda_{L-1}^W \end{pmatrix}$$

Example: The Poisson GLM for Contingency Tables

- Under the *joint independence model*, $(U, V) \perp\!\!\!\perp W$, and

$$\pi_{jkl} = \pi_{jk+}\pi_{++l}$$

- This implies that we can write

$$\begin{aligned}\log \mu_{jkl} &= \log n + \log \pi_{jk+} + \log \pi_{++l} \\ &:= \lambda + \lambda_j^U + \lambda_k^V + \lambda_l^W + \lambda_{jk}^{UV}\end{aligned}$$

for some parameters $\lambda, \lambda_j^U, \lambda_k^V, \lambda_l^W, \lambda_{jk}^{UV}$ subject to constraints, e.g.,

$$\lambda_J^U = \lambda_K^V = \lambda_L^W = \lambda_{Jk}^{UV} = \lambda_{jK}^{UV} = 0$$

Example: The Poisson GLM for Contingency Tables

- More generally, the log-linear model

$$\log \mu_{jkl} = \lambda + \lambda_j^U + \lambda_k^V + \lambda_l^W + \lambda_{jk}^{UV} + \lambda_{jl}^{UW} + \lambda_{kl}^{VW} + \lambda_{jkl}^{UVW}$$

  represents a saturated model for the contingency table, subject to some parameter constraints

- Submodels represent different types of marginal or conditional independence among U, V, W or other types of structure

- *Note*: a log-linear model for a contingency table is a GLM where the responses are the counts in the cells, and the covariates are built from the dimensions of the table

- For more on this see Bishop, Fienberg, and Holland (1974), Agresti (2002)

Example: The Negative Binomial GLM

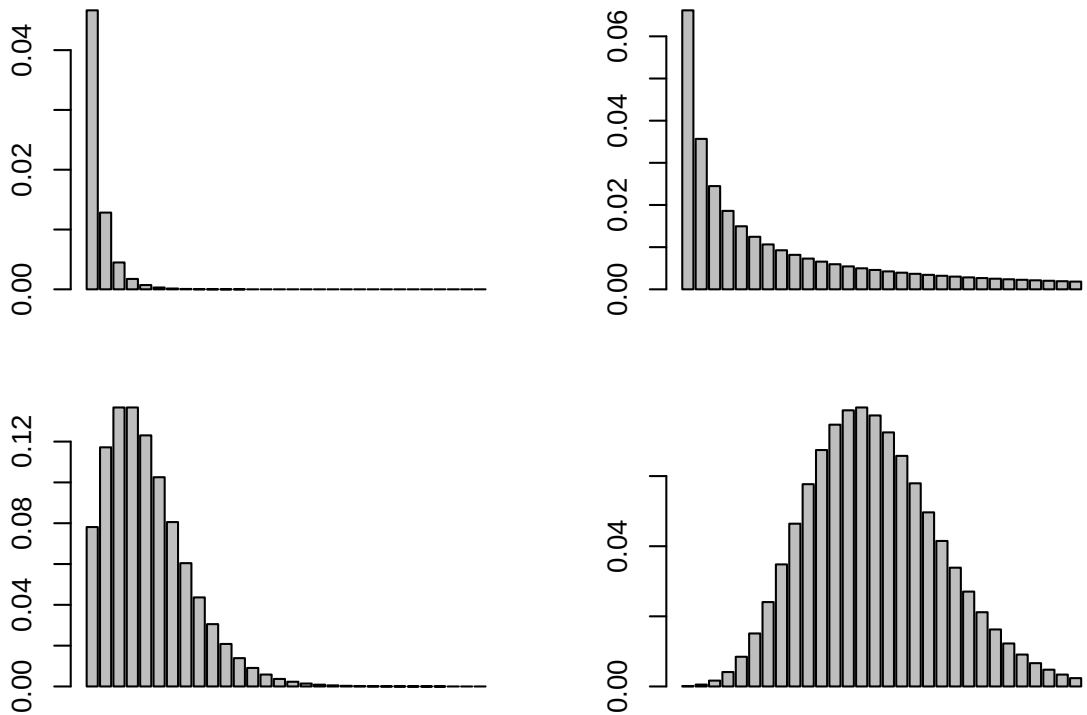
- If $Y \sim \textit{Poisson}(\mu)$, then $\mathrm{E}(Y) = \mathrm{var}(Y) = \mu$
- This property transfers to the Poisson GLM as

$$\mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \mathrm{var}(Y_i \mid \boldsymbol{x}_i),$$

  but it is common to find data where this doesn't hold
- *Overdispersion*: when the variance of the responses exceeds the variance predicted by a model
- Overdispersion is commonly used to indicate larger variability than encoded by the Poisson
- The negative binomial distribution offers an option for handling overdispersion where $\mathrm{var}(Y) > \mathrm{E}(Y)$

Example: The Negative Binomial GLM

In general, the negative binomial is suitable for modeling counts that are potentially heavy-tailed, potentially zero-inflated.



Some possible shapes of the negative binomial probability mass function

Example: The Negative Binomial GLM

- The most common parameterization of the *NegativeBinomial*(r, p) PMF is

$$f(y \mid r, p) = \frac{\Gamma(y + r)}{y!\,\Gamma(r)}(1 - p)^r p^y \quad \text{for } y = 0, 1, 2, \ldots$$

where $\mathrm{E}(Y) = rp/(1 - p)$ and $\mathrm{var}(Y) = rp/(1 - p)^2$

- For regression, we reparameterize in terms of the mean

$$\mu = rp/(1 - p)$$

obtaining

$$f(y \mid r, \mu) = \frac{\Gamma(r + y)}{y!\,\Gamma(r)} \left(\frac{r}{r + \mu}\right)^r \left(\frac{\mu}{r + \mu}\right)^y \quad \text{for } y = 0, 1, 2, \ldots$$

so that

$$\mathrm{var}(Y) = \mu + \mu^2/r$$

Example: The Negative Binomial GLM

For a fixed r, we can write the PMF in exponential dispersion family form:

$$
\begin{aligned}
f(y \mid r, \mu) &= \frac{\Gamma(r+y)}{y!\,\Gamma(r)} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^y \\
&= \exp\left\{ y \log\left(\frac{\mu}{r+\mu}\right) - r \log\left(\frac{r+\mu}{r}\right) + \log \frac{\Gamma(r+y)}{y!\,\Gamma(r)} \right\}
\end{aligned}
$$

where

- The natural parameter is $\theta = \log\left[\mu/(r+\mu)\right]$
- The log-partition function is $b(\theta) = -r \log\left(1 - \exp\theta\right)$
- The dispersion parameter is $\alpha = 1$ (*but this assumes r is fixed!*)

If r is unknown, the negative binomial does not belong to the exponential dispersion family

Example: The Negative Binomial GLM

- In practice, the parameter r controls the overdispersion, and it is unknown
- Technically, with unknown r, negative binomial (NB) regression *is not* a GLM
- The NB regression model still has the same structure as a GLM:
  - $Y_i \mid \boldsymbol{x}_i \sim NB(\mu_i, r)$
  - $\mu_i = \mu(\boldsymbol{x}_i)$
  - $g[\mu(\boldsymbol{x}_i)] = \boldsymbol{x}_i\boldsymbol{\beta}$, where usually $g(\cdot) = \log(\cdot)$
- NB regression behaves like a GLM in many ways

Example: The Negative Binomial GLM

- The log-likelihood $l(\boldsymbol{\beta}, r)$ can be written based on n independent pairs $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^n$
- To obtain the MLEs of $\boldsymbol{\beta}$ and r, note
  - The likelihood equations obtained from $\partial l(\boldsymbol{\beta}, r)/\partial \boldsymbol{\beta} = \mathbf{0}$ have the shape of the score equations obtained for GLMs
  - Given a value of $r^{(t)}$, IRLS (Fisher scoring) can be used to obtain $\hat{\boldsymbol{\beta}}^{(t)}$
  - Given $\hat{\boldsymbol{\beta}}^{(t)}$, univariate optimization methods can be used to obtain $\hat{r}^{(t+1)}$
  - These two steps are iterated until convergence

Example: The Negative Binomial GLM

To obtain inferences on $\boldsymbol{\beta}$, we have the following

- It's easy to check that $\mathrm{E}[\partial^2 l(\boldsymbol{\beta}, r)/\partial\boldsymbol{\beta}\partial r] = \mathbf{0}$
- Therefore the Fisher information for $(\boldsymbol{\beta}, r)$ is block-diagonal, and so is its inverse
- The MLEs $\hat{\boldsymbol{\beta}}$ and $\hat{r}$ are asymptotically independent
- The asymptotic variance of $\hat{\boldsymbol{\beta}}$ is the same regardless of whether r is known or replaced by a consistent estimator
- For inference on $\boldsymbol{\beta}$, we can treat NB regression as a GLM conditional on $\hat{r}$
- The asymptotic variance of the estimator $\hat{\boldsymbol{\beta}}$ is

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T \boldsymbol{W}(\hat{\boldsymbol{\beta}}, \hat{r})\boldsymbol{X})^{-1},$$

with

$$\boldsymbol{W}(\hat{\boldsymbol{\beta}}, \hat{r}) = \mathrm{diag}\left\{(d\mu_i/d\eta_i)^2|_{\hat{\boldsymbol{\beta}}} \;/\; [\mu_i(\hat{\boldsymbol{\beta}}) + \mu_i(\hat{\boldsymbol{\beta}})^2/\hat{r}]\right\}$$

Example: The Negative Binomial GLM

An appealing way of motivating the NB "GLM" is as follows:

- In the Poisson GLM we assume that the means μ_i can be written as a deterministic function of covariates $\boldsymbol{x}_i$ as $\mu_i = \mu(\boldsymbol{x}_i)$
- However, with unmeasured covariates that influence the μ_i's, there is a distribution of μ_i's per unique value of measured covariates $\boldsymbol{x}_i$
- We can then think of a model for the distribution of the *latent* means μ_i as a function of observed covariates $\boldsymbol{x}_i$
- *Remember*: If $Y \mid \mu \sim \textit{Poisson}(\mu)$ and $\mu \sim \textit{Gamma}(\bar{\mu}, \alpha)$ then Y is marginally NB with $\mathrm{E}(Y) = \bar{\mu}$ and $\mathrm{var}(Y) = \bar{\mu} + \alpha\bar{\mu}^2$

Example: The Negative Binomial GLM

The NB "GLM" can then be derived as follows:

- Assume a gamma GLM for the latent means

$$\mu_i \mid \boldsymbol{x}_i \sim \mathit{Gamma}[\ \bar{\mu}(\boldsymbol{x}_i), \alpha]$$

- Assume that the observed counts are

$$Y_i \mid \mu_i \sim \mathit{Poisson}[\ \mu_i]$$

- Since the individual μ_i's are not observed, we integrate them out, obtaining

$$Y_i \mid \boldsymbol{x}_i \sim \mathit{NB}[\ \bar{\mu}(\boldsymbol{x}_i), 1/\alpha]$$

where $1/\alpha = r$ above

Some Comments on Nonlinear Models

Motivation:

- So far we have focused on models with

$$g[\mathrm{E}(Y_i \mid \boldsymbol{x}_i)] = \boldsymbol{x}_i\boldsymbol{\beta}$$

- What about general models with

$$\mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i;\ \boldsymbol{\beta}) := \mu_i(\boldsymbol{\beta}),$$

  where $\mu_i(\cdot)$ is not a linear function of $\boldsymbol{\beta}$?

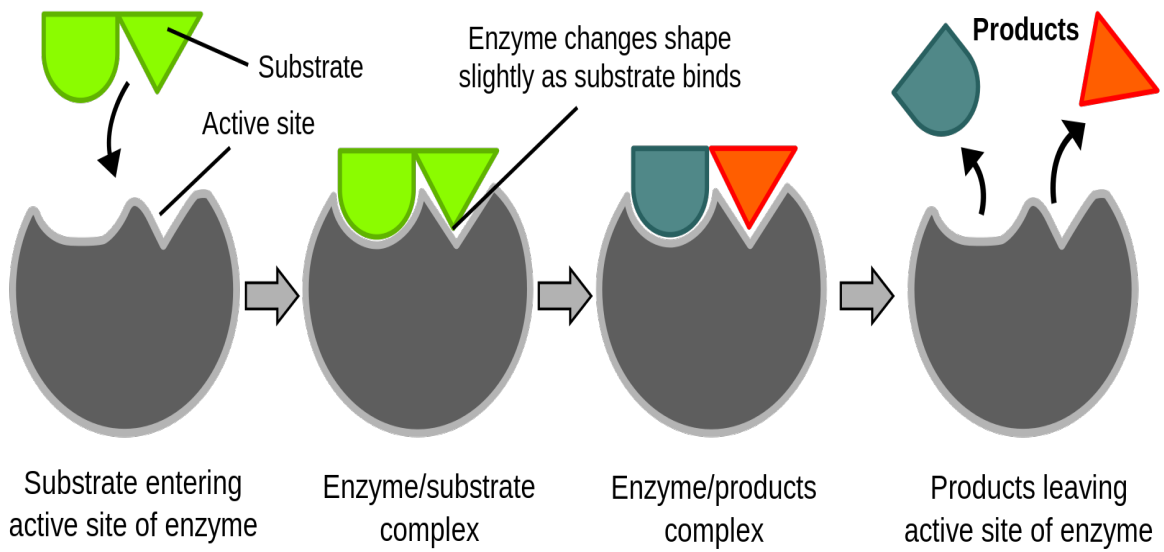
- Such models sometimes appear based on strong scientific motivation

- See Wakefield (2013, ch. 1.3.4, 6.10) for examples in pharmacokinetics and biochemistry

Some Comments on Nonlinear Models

Some background on the *Michaelis-Menten model*:

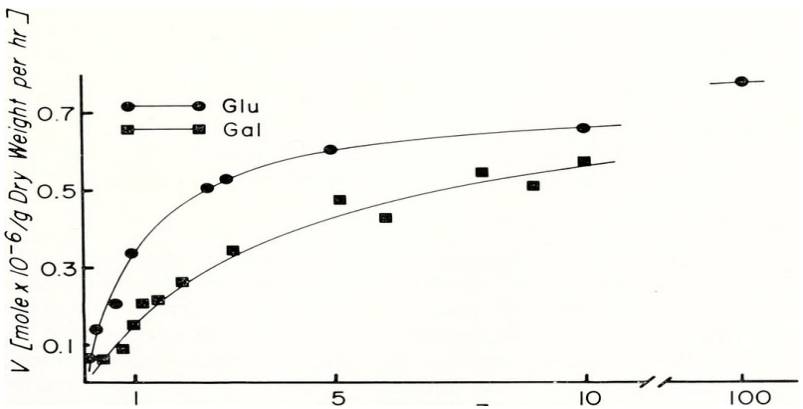
- A *chemical reaction* is a process that leads to the chemical transformation of one set of chemical substances to another
- In biochemistry, biochemical reactions are mainly controlled by *enzymes*

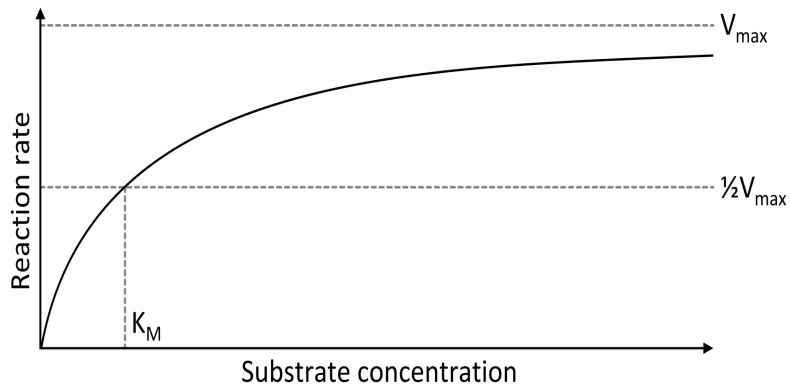


Taken from `https://en.wikipedia.org/wiki/Chemical_reaction#/media/File:Induced_fit_diagram.svg`

Some Comments on Nonlinear Models

- It has been shown experimentally that if the amount of the enzyme is kept constant and the substrate concentration is then gradually increased, the reaction velocity will increase until it reaches a maximum (the enzyme becomes saturated with substrate).





Taken from https://www.flickr.com/photos/internetarchivebookimages/19756411313/ and https://commons.wikimedia.org/wiki/File:Michaelis_Menten_curve_2.svg

Some Comments on Nonlinear Models

- The Michaelis-Menten model is commonly used for enzyme kinetics (the study of the chemical reactions that are catalysed by enzymes)

- Describes expected rate of reaction Y (counts/min) as a function of substrate concentration Z (parts per million)

- The Michaelis-Menten model corresponds to the nonlinear regression function

$$\mu(z) = \frac{\alpha_0 z}{\alpha_1 + z}.$$

- Parameter interpretation is obtained by recognizing that as $z \to \infty$, $\mu(z) \to \alpha_0$ and at α_1, $\mu(\alpha_1) = \alpha_0/2$

- Note that the Michaelis-Menten model only describes the *systematic component* of the statistical model, but not the *stochastic component*

Some Comments on Nonlinear Models

The definition of nonlinear models simply say that $\mu_i(\cdot)$ is not a linear function of $\boldsymbol{\beta}$, so the implementations can take different flavors

- Wakefield (2013, ch. 6.10) presents an example with

$$\mathsf{E}(Y \mid \boldsymbol{x}) = \mu(\boldsymbol{x}; \boldsymbol{\beta}),$$

with $\mathsf{var}[Y \mid \boldsymbol{x}] = \sigma^2 \mu(\boldsymbol{x})^r$, with $r = 0, 1$ or 2, and

$$Y_i \mid \boldsymbol{x}_i \sim \mathsf{N}[\mu_i(\boldsymbol{\beta}), \sigma^2 \mu_i(\boldsymbol{\beta})^r]$$

- This leads to a fully parametric model for the data, from which we can obtain a likelihood function

Some Comments on Nonlinear Models

- The corresponding log-likelihood function is

$$l(\boldsymbol{\beta}, \sigma) = -n \log \sigma - \frac{r}{2} \sum_{i=1}^{n} \log \mu_i(\boldsymbol{\beta}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{[Y_i - \mu_i(\boldsymbol{\beta})]^2}{\mu_i^r(\boldsymbol{\beta})}.$$

- Differentiation with respect to $\boldsymbol{\beta}$ and σ yields the score equations

$$\begin{aligned}
\boldsymbol{S}_1(\boldsymbol{\beta}, \sigma) &= \frac{\partial l}{\partial \boldsymbol{\beta}} \\
&= \frac{r}{2\sigma^2} \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \frac{1}{\mu_i(\boldsymbol{\beta})} \left\{ \frac{[Y_i - \mu_i(\boldsymbol{\beta})]^2}{\mu_i^r(\boldsymbol{\beta})} - \sigma^2 \right\} + \frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{[Y_i - \mu_i(\boldsymbol{\beta})]}{\mu_i(\boldsymbol{\beta})^r} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \\
S_2(\boldsymbol{\beta}, \sigma) &= \frac{\partial l}{\partial \sigma} \\
&= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} \frac{[Y_i - \mu_i(\boldsymbol{\beta})]^2}{\mu_i^r(\boldsymbol{\beta})}
\end{aligned}$$

- In general, the MLEs for $\boldsymbol{\beta}$ are not available in closed form but they need to be obtained via numerical methods, like the ones we saw for GLMs

Some Comments on Nonlinear Models

- The MLE for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{[Y_i - \mu_i(\hat{\boldsymbol{\beta}})]^2}{\mu_i^r(\hat{\boldsymbol{\beta}})}$$

- By analogy with the linear model case, it is more usual to use the degrees-of-freedom-adjusted estimator

$$\widetilde{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\frac{[Y_i - \mu_i(\widehat{\boldsymbol{\beta}})]^2}{\mu_i^r(\widehat{\boldsymbol{\beta}})}$$

Some Comments on Nonlinear Models

- Under the usual regularity conditions

$$\mathcal{I}(\boldsymbol{\theta})^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \to \mathrm{N}_{k+2}(\mathbf{0}, \boldsymbol{I}_{k+2}).$$

  where $\boldsymbol{\theta} = [\boldsymbol{\beta}, \sigma]$ and $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher's expected information.

- As usual, different tests of hypothesis may be conducted using the Wald, score or likelihood ratio statistics.

- *Moral*: likelihood-based inference is not limited to GLMs!

Some Comments on Nonlinear Models

Final remark:

- Much of what we did for GLMs was simply to adapt general results for likelihood-based inference to the structure of GLMs:
  - Response Y in exponential family
  - Dependence of Y on covariates $\boldsymbol{x}$ only through the mean
  - Linearity of means on a transformed scale as a function of $\boldsymbol{\beta}$
- However, some of the formulae that we obtained applies even if we change the third component, e.g., in

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i} \frac{d\theta_i}{d\mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \\ &= \boldsymbol{D}^{\mathsf{T}} \boldsymbol{V}^{-1} \left[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right] / \alpha \end{aligned}$$

only $\boldsymbol{D}$, where $[\boldsymbol{D}]_{i,j} = \partial \mu_i / \partial \beta_j$ depends on the mean modeling, which can be used as the basis for general nonlinear modeling with responses in the exponential family