Part IV:

Theory of Generalized Linear Models

Lung cancer surgery

Q: Is there an association between time spent in the operating room and post-surgical outcomes?

- Could choose from a number of possible response variables, including:
  - ⋆ hospital stay of > 7 days
  - ⋆ number of major complications during the hospital stay
- The scientific goal is to characterize the joint distribution between both of these responses and a p-vector of covariates, X
  - ⋆ age, co-morbidities, surgery type, resection type, etc
- The first response is *binary* and the second is a *count* variable
  - ⋆ $Y \in \{0, 1\}$
  - ⋆ $Y \in \{0, 1, 2, \ldots\}$

Q: Can we analyze such response variables with the linear regression model?

- ⋆ specify a mean model

$$\mathrm{E}[Y_i|X_i] \;=\; X_i^T\boldsymbol{\beta}$$

- ⋆ estimate $\boldsymbol{\beta}$ via least squares and perform inference via the CLT

- Given continuous response data, least squares estimation works remarkably well for the linear regression model
  - ⋆ assuming the mean model is correctly specified, $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ is unbiased
  - ⋆ OLS is generally robust to the underlying distribution of the error terms
    - ∗ Homework #2
  - ⋆ OLS is 'optimal' if the error terms are homoskedastic
    - ∗ MLE if $\epsilon \sim$ Normal(0, σ^2) and BLUE otherwise

- For a binary response variable, we could specify a linear regression model:

$$
\begin{aligned}
\mathsf{E}[Y_i|X_i] &= X_i^T\boldsymbol{\beta} \\
Y_i|X_i &\sim \text{Bernoulli}(\mu_i)
\end{aligned}
$$

where, for notational convenience, $\mu_i = X_i^T\boldsymbol{\beta}$

- As long as this model is correctly specified, $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ will still be unbiased

- For the Bernoulli distribution, there is an implicit mean-variance relationship:

$$
\mathsf{V}[Y_i|X_i] = \mu_i(1-\mu_i)
$$

  - ⋆ as long as $\mu_i \neq \mu \; \forall \; i$, study units will be heteroskedastic
  - ⋆ non-constant variance

- Ignoring heteroskedasticity results in invalid inference
  - ⋆ naïve standard errors (that assume homoskedasticity) are incorrect

- We've seen three possible remedies:

  (1) transform the response variable

  (2) use OLS and base inference on a valid standard error

  (3) use WLS

- Recall, $\widehat{\boldsymbol{\beta}}_{\text{WLS}}$ is the solution to

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta}; \boldsymbol{W}) \\
0 &= \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} w_i (y_i - X_i^T \boldsymbol{\beta})^2 \\
0 &= \sum_{i=1}^{n} X_i w_i (y_i - X_i^T \boldsymbol{\beta})
\end{aligned}
$$

- For a binary response, we know the form of $\mathsf{V}[Y_i]$
  - ⋆ estimate $\boldsymbol{\beta}$ by setting $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$, a diagonal matrix with elements:

$$w_i = \frac{1}{\mu_i(1-\mu_i)}$$

- From the Gauss-Markov Theorem, the resulting estimator is BLUE

$$\widehat{\boldsymbol{\beta}}_{\text{GLS}} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

- Note, the least squares equations become

$$0 = \sum_{i=1}^{n} \frac{X_i}{\mu_i(1-\mu_i)}(y_i - \mu_i)$$

  - ⋆ in practice, we use the IWLS algorithm to estimate $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ while simultaneously accommodating the mean-variance relationship

- We can also show that $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$, obtained via the IWLS algorithm, is the MLE
  - ⋆ firstly, note that the likelihood and log-likelihood are:

$$\mathcal{L}(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{n} \mu_i^{y_i}(1-\mu_i)^{1-y_i}$$

$$\ell(\boldsymbol{\beta}|\boldsymbol{y}) = \sum_{i=1}^{n} y_i \log(\mu_i) + (1-y_i)\log(1-\mu_i)$$

  - ⋆ to get the MLE, we take derivatives, set them equal to zero and solve
  - ⋆ following the algebra trail we find that

$$\frac{\partial}{\partial \boldsymbol{\beta}}\ell(\boldsymbol{\beta}|\boldsymbol{y}) = \sum_{i=1}^{n} \frac{X_i}{\mu_i(1-\mu_i)}(Y_i - \mu_i)$$

- The score equations are equivalent to the least squares equations
  - ⋆ $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ is therefore ML

- So, least squares estimation can accommodate implicit heteroskedasticity for binary data by using the IWLS algorithm
  - ⋆ assuming the model is correctly specified, WLS is in fact optimal!

- However, when modeling binary or count response data, the linear regression model doesn't respect the fact that the outcome is bounded
  - ⋆ the functional that is being modeled is bounded:
    - ∗ binary: $\text{E}[Y_i|X_i] \in (0,1)$
    - ∗ count: $\text{E}[Y_i|X_i] \in (0,\infty)$
  - ⋆ but our current specification of the mean model doesn't impose any restrictions

$$\text{E}[Y_i|X_i] = X_i^T\boldsymbol{\beta}$$

Q: Is this a problem?

Summary

- Our goal is to develop statistical models to characterize the relationship between some response variable, Y, and a vector of covariates, X

- Statistical models consist of two components:
  - ⋆ a *systematic* component
  - ⋆ a *random* component

- When moving beyond linear regression analysis of continuous response data, we need to be aware of two key challenges:

  (1) sensible specification of the systematic component

  (2) proper accounting of any implicit mean-variance relationships arising from the random component

Generalized Linear Models

Definition

- A *generalized linear model* (GLM) specifies a parametric statistical model for the conditional distribution of a response Y_i given a p-vector of covariates X_i
- Consists of three elements:

  (1) probability distribution, $Y \sim f_Y(y)$

  (2) linear predictor, $X_i^T \boldsymbol{\beta}$

  (3) link function, $g(\cdot)$

  ⋆ element (1) is the random component

  ⋆ elements (2) and (3) jointly specify the systematic component

Random component

- In practice, we see a wide range of response variables with a wide range of associated (possible) distributions

Response type	Range	Possible distribution
Continuous	$(-\infty,\ \infty)$	Normal($\mu,\ \sigma^2$)
Binary	$\{0,\ 1\}$	Bernoulli(π)
Polytomous	$\{1,\ \ldots,\ K\}$	Multinomial(π_k)
Count	$\{0,\ 1,\ \ldots,\ n\}$	Binomial($n,\ \pi$)
Count	$\{0,\ 1,\ \ldots\}$	Poisson(μ)
Continuous	$(0,\ \infty)$	Gamma($\alpha,\ \beta$)
Continuous	$(0, 1)$	Beta($\alpha,\ \beta$)

- Desirable to have a single framework that accommodates all of these

Systematic component

- For a given choice of probability distribution, a GLM specifies a model for the *conditional mean*:

$$\mu_i \;=\; \mathsf{E}[Y_i|X_i]$$

Q: How do we specify reasonable models for μ_i while ensuring that we respect the appropriate range/scale of μ_i?

- Achieved by constructing a linear predictor $X_i^T\boldsymbol{\beta}$ and relating it to μ_i via a link function $g(\cdot)$:

$$g(\mu_i) \;=\; X_i^T\boldsymbol{\beta}$$

  ⋆ often use the notation $\eta_i = X_i^T\boldsymbol{\beta}$

The random component

- GLMs form a class of statistical models for response variables whose distribution belongs to the *exponential dispersion family*
  - ⋆ family of distributions with a pdf/pmf of the form:

$$f_Y(y;\theta,\phi) \;=\; \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} \;+\; c(y,\phi)\right\}$$

  - ⋆ θ is the *canonical parameter*
  - ⋆ ϕ is the *dispersion parameter*
  - ⋆ $b(\theta)$ is the *cumulant function*
- Many common distributions are members of this family

- $Y \sim \text{Bernoulli}(\pi)$

$$
\begin{aligned}
f_Y(y;\pi) &= \mu^y(1-\mu)^{1-y} \\
f_Y(y;\theta,\phi) &= \exp\left\{y\theta \;-\; \log\left(1+\exp\{\theta\}\right)\right\} \\
\theta &= \log\left(\frac{\pi}{1-\pi}\right) \\
a(\phi) &= 1 \\
b(\theta) &= \log\left(1+\exp\{\theta\}\right) \\
c(y,\phi) &= 0
\end{aligned}
$$

- Many other common distributions are also members of this family

- The canonical parameter has key relationships with both E$[Y]$ and V$[Y]$
  - ⋆ typically varies across study units
  - ⋆ index θ by i: θ_i

- The dispersion parameter has a key relationship with V$[Y]$
  - ⋆ may but typically does not vary across study units
  - ⋆ typically no unit-specific index: ϕ
  - ⋆ in some settings we may have $a(\cdot)$ vary with i: $a_i(\phi)$
    - ∗ e.g. $a_i(\phi) = \phi/w_i$, where w_i is a prior weight

- When the dispersion parameter is known, we say that the distribution is a member of the *exponential family*

Properties

- Consider the likelihood function for a single observation

$$\mathcal{L}(\theta_i, \phi; y_i) \;=\; \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} \;+\; c(y_i, \phi) \right\}$$

- The log-likelihood is

$$\ell(\theta_i, \phi; y_i) \;=\; \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} \;+\; c(y_i, \phi)$$

- The first partial derivative with respect to θ_i is the score function for θ_i and is given by

$$\frac{\partial}{\partial \theta_i} \ell(\theta_i, \phi; y_i) \;=\; U(\theta_i) \;=\; \frac{y_i - b'(\theta_i)}{a_i(\phi)}$$

- Using standard results from likelihood theory, we know that under appropriate regularity conditions:

$$\mathsf{E}\left[U(\theta_i)\right] \;=\; 0$$

$$\mathsf{V}\left[U(\theta_i)\right] \;=\; \mathsf{E}\left[U(\theta_i)^2\right] \;=\; -\mathsf{E}\left[\frac{\partial U(\theta_i)}{\partial \theta_i}\right]$$

  ⋆ this latter expression is the $(i,i)^{th}$ component of the Fisher information matrix

- Since the score has mean zero, we find that

$$\mathsf{E}\left[\frac{Y_i - b'(\theta_i)}{a_i(\phi)}\right] \;=\; 0$$

and, consequently, that

$$\mathsf{E}[Y_i] \;=\; b'(\theta_i)$$

- The second partial derivative of $\ell(\theta_i, \phi; y_i)$ is

$$\frac{\partial^2}{\partial \theta_i^2} \ell(\theta_i, \phi; y_i) \;=\; -\,\frac{b''(\theta_i)}{a_i(\phi)}$$

  ⋆ the observed information for the canonical parameter from the i^{th} observation

- This is also the expected information and using the above properties it follows that

$$\mathsf{V}\left[U(\theta_i)\right] \;=\; \mathsf{V}\left[\frac{Y_i - b'(\theta_i)}{a_i(\phi)}\right] \;=\; \frac{b''(\theta_i)}{a_i(\phi)},$$

  so that

$$\mathsf{V}[Y_i] \;=\; b''(\theta_i)a_i(\phi)$$

- The variance of Y_i is therefore a function of both θ_i and ϕ
- Note that the canonical parameter is a function of μ_i

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

so that we can write

$$\mathsf{V}[Y_i] \ = \ b''(\theta(\mu_i))a_i(\phi)$$

- The function $V(\mu_i) = b''(\theta(\mu_i))$ is called the *variance function*
    - ⋆ specific form indicates the nature of the (if any) mean-variance relationship
- For example, for $Y \sim$ Bernoulli(μ)

$$a(\phi) \ = \ 1$$

$$
\begin{aligned}
b(\theta) &= \log\left(1+\exp\{\theta\}\right)
\end{aligned}
$$

$$
\begin{aligned}
\mathsf{E}[Y] &= b'(\theta) \\
&= \frac{\exp\{\theta\}}{1+\exp\{\theta\}} \;=\; \mu
\end{aligned}
$$

$$
\begin{aligned}
\mathsf{V}[Y] &= b''(\theta)a(\phi) \\
&= \frac{\exp\{\theta\}}{(1+\exp\{\theta\})^2} \;=\; \mu(1-\mu)
\end{aligned}
$$

$$
V(\mu) = \mu(1-\mu)
$$

The systematic component

- For the exponential dispersion family, the pdf/pmf has the following form:

$$f_Y(y_i; \theta_i, \phi) \; = \; \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} \; + \; c(y_i, \phi) \right\}$$

  - ⋆ this distribution is the random component of the statistical model

- We need a means of specifying how this distribution depends on a vector of covariates X_i
  - ⋆ the systematic component

- In GLMs we model the conditional mean, $\mu_i = \mathsf{E}[Y_i | X_i]$
  - ⋆ provides a connection between X_i and distribution of Y_i via the canonical parameter θ_i and the cumulant function $b(\theta_i)$

- Specifically, the relationship between μ_i and X_i is given by

$$g(\mu_i) \;=\; X_i^T \boldsymbol{\beta}$$

  ⋆ we 'link' the linear predictor to the distribution of of Y_i via a transformation of μ_i

- Traditionally, this specification is broken down into two parts:

  (1) the linear predictor, $\eta_i = X_i^T \boldsymbol{\beta}$

  (2) the link function, $g(\mu_i) = \eta_i$

- You'll often find the linear predictor called the 'systematic component'
  ⋆ e.g., McCullagh and Nelder (1989) *Generalized Linear Models*

- In practice, one cannot consider one without the other
  ⋆ the relationship between μ_i and X_i is *jointly* determined by $\boldsymbol{\beta}$ and $g(\cdot)$

The linear predictor, $\eta_i = X_i^T \boldsymbol{\beta}$

- Constructing the linear predictor for a GLM follows the same process one uses for linear regression

- Given a set of covariates X_i, there are two decisions
  - ⋆ which covariates to include in the model?
  - ⋆ how to include them in the model?

- For the most part, the decision of which covariates to include should be driven by scientific considerations
  - ⋆ is the goal estimation or prediction?
  - ⋆ is there a primary exposure of interest?
  - ⋆ which covariates are predictors of the response variable?
  - ⋆ are any of the covariates effect modifiers? confounders?

- In some settings, practical or data-oriented considerations may drive these decisions
  - ⋆ small sample sizes
  - ⋆ missing data
  - ⋆ measurement error/missclassification
- How one includes them in the model will also depend on a mixture of scientific and practical considerations
- Suppose we are interested in the relationship between birth weight and risk of death within the first year of life
  - ⋆ infant mortality
- Note: birth weight is a continuous covariate
  - ⋆ there are a number of options for including a continuous covariate into the linear predictor

- Let X_w denote the continuous birth weight measure
- A simple model would be to include X_w via a linear term

$$\eta \;=\; \beta_0 \;+\; \beta_1 X_w$$

  - ⋆ a 'constant' relationship between birth weight and infant mortality
- May be concerned that this is too restrictive a model
  - ⋆ include additional polynomial terms

$$\eta \;=\; \beta_0 \;+\; \beta_1 X_w \;+\; \beta_2 X_w^2 \;+\; \beta_3 X_w^3$$

  - ⋆ more flexible than the linear model
  - ⋆ but the interpretation of β_2 and β_3 is difficult

- Scientifically, one might only be interested in the 'low birth weight' threshold
  - let $X_{lbw} = 0/1$ if birth weight is >2.5kg/≤2.5kg

$$\eta \;=\; \beta_0 \;+\; \beta_1 X_{lbw}$$

  - impact of birth weight on risk of infant mortality manifests solely through whether or not the baby has a low birth weight

- The underlying relationship may be more complex than a simple linear or threshold effect, although we don't like the (lack of) interpretability of the polynomial model
  - categorize the continuous covariates into $K + 1$ groups
  - include in the linear predictor via K dummy variables

$$\eta \;=\; \beta_0 \;+\; \beta_1 X_{cat,1} \;+\; \ldots \;+\; \beta_K X_{cat,K}$$

The link function, $g(\cdot)$

- Given the form of linear predictor $X_i^T\boldsymbol{\beta}$ we need to specify how it is related to the conditional mean μ_i

- As we've noted, the range of values that μ_i can take on may be restricted
  - ⋆ binary data: $\mu_i \in (0, 1)$
  - ⋆ count data: $\mu_i \in (0, \infty)$

- One approach would be to estimate $\boldsymbol{\beta}$ subject to the constraint that all (modeled) values of μ_i respect the appropriate range

Q: What might the drawbacks of such an approach be?

- An alternative is to permit the estimation of $\boldsymbol{\beta}$ to be 'free' but impose a functional form of the relationship between μ_i and $X_i^T\boldsymbol{\beta}$
  - ⋆ via the link function $g(\cdot)$

$$g(\mu_i) \;=\; X_i^T\boldsymbol{\beta}$$

- We interpret the link function as specifying a transformation of the conditional mean, μ_i
  - ⋆ we are <u>not</u> specifying a transformation of the response Y_i

- The inverse of the link function provides the specification of the model on the scale of μ_i

$$\mu_i \;=\; g^{-1}\left(X_i^T\boldsymbol{\beta}\right)$$

  - ⋆ link functions are therefore usually monotone and have a well-defined inverse

- In linear regression we specify

$$\mu_i = X_i^T \boldsymbol{\beta}$$

  ⋆ $g(\cdot)$ is the identity link

- In logistic regression we specify

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = X_i^T \boldsymbol{\beta}$$

  ⋆ $g(\cdot)$ is the logit or logistic link

- In Poisson regression we specify

$$\log(\mu_i) = X_i^T \boldsymbol{\beta}$$

  ⋆ $g(\cdot)$ is the log link

- For linear regression also we have that

$$\mu_i \;=\; X_i^T \boldsymbol{\beta}$$

  ⋆ $g^{-1}(\eta_i) = \eta_i$ is the identity function

- For logistic regression

$$\mu_i \;=\; \frac{\exp\left\{X_i^T \boldsymbol{\beta}\right\}}{1 + \exp\left\{X_i^T \boldsymbol{\beta}\right\}}$$

  ⋆ $g^{-1}(\eta_i) = \text{expit}(\eta_i)$ is the expit function

- For Poisson regression

$$\mu_i \;=\; \exp\left\{X_i^T \boldsymbol{\beta}\right\}$$

  ⋆ $g^{-1}(\eta_i) = \exp(\eta_i)$ is the exponential function

The canonical link

- Recall that the mean and the canonical parameter are linked via the derivative of the cumulant function
  - ⋆ $\mathsf{E}[Y_i] = \mu_i = b'(\theta_i)$
- An important link function is the *canonical* link:

$$g(\mu_i) \;=\; \theta(\mu_i)$$

  - ⋆ the function that results by viewing the canonical parameter θ_i as a function of μ_i
  - ⋆ inverse of $b'(\cdot)$
- We'll see later that this choice results in some mathematical convenience

Choosing $g(\cdot)$

- In practice, there are often many possible link functions

- For binary response data, one might choose a link function from among the following:

$$
\begin{aligned}
\text{identity:} \qquad & g(\mu_i) = \mu_i \\
\text{log:} \qquad & g(\mu_i) = \log(\mu_i) \\
\text{logit:} \qquad & g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) \\
\text{probit:} \qquad & g(\mu_i) = \text{probit}(\mu_i) \\
\text{complementary log-log:} \qquad & g(\mu_i) = \log\left\{-\log(1-\mu_i)\right\}
\end{aligned}
$$

⋆ note the logit link is the canonical link function

- We typically choose a specific link function via consideration of two issues:

  (1) respect of the range of values that μ_i can take

  (2) impact on the interpretability of $\boldsymbol{\beta}$

- There can be a trade-off between mathematical convenience and interpretability of the model

- We'll spend more time on this later on in the course

Frequentist estimation and inference

- Given an i.i.d sample of size n, the log-likelihood is

$$\ell(\boldsymbol{\beta}, \phi; \boldsymbol{y}) \;=\; \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} \;+\; c(y_i, \phi)$$

  where θ_i is a function of $\boldsymbol{\beta}$ and is determined by

  ⋆ the form of $b'(\theta_i) = \mu_i$

  ⋆ the choice of the link function via $g(\mu_i) = \eta_i = X_i^T\boldsymbol{\beta}$

- The primary goal is to perform estimation and inference with respect to $\boldsymbol{\beta}$

- Since we've fully specified the likelihood, we can proceed with likelihood-based estimation/inference

Estimation

- There are $(p+2)$ unknown parameters: $(\boldsymbol{\beta}, \phi)$
- To obtain the MLE we need to solve the score equations:

$$\left(\frac{\partial \ell(\boldsymbol{\beta}, \phi; \boldsymbol{y})}{\partial \beta_0}, \ \cdots, \ \frac{\partial \ell(\boldsymbol{\beta}, \phi; \boldsymbol{y})}{\partial \beta_p}, \ \frac{\partial \ell(\boldsymbol{\beta}, \phi; \boldsymbol{y})}{\partial \phi}\right)^T = \mathbf{0}$$

  ⋆ system of $(p+2)$ equations

- The contribution to the score for ϕ by the i^{th} unit is

$$\frac{\partial \ell(\boldsymbol{\beta}, \phi; y_i)}{\partial \phi} = -\frac{a_i'(\phi)}{a_i(\phi)^2}\left(y_i\theta_i - b(\theta_i)\right) + c'(y_i, \phi)$$

- We can use the chain rule to obtain a convenient expression for the i^{th} contribution to the score function for β_j:

$$\frac{\partial \ell(\boldsymbol{\beta}, \phi; y_i)}{\partial \beta_j} = \frac{\partial \ell(\boldsymbol{\beta}, \phi; y_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

- Note the following results:

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\beta}, \phi; y_i)}{\partial \theta_i} &= \frac{y_i - \mu_i}{a_i(\phi)} \\
\frac{\partial \mu_i}{\partial \theta_i} &= b''(\theta_i) \\
&= \frac{\mathsf{V}[Y_i]}{a_i(\phi)} \\
\frac{\partial \eta_i}{\partial \beta_j} &= X_{j,i}
\end{aligned}$$

- The score function for β_j can therefore be written as

$$\frac{\partial \ell(\boldsymbol{\beta}, \phi; \boldsymbol{y})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \eta_i} \frac{X_{j,i}}{V(\mu_i) a_i(\phi)} (y_i - \mu_i)$$

  - ⋆ depends on the distribution of Y_i solely through $\mathsf{E}[Y_i] = \mu_i$ and $\mathsf{V}[Y_i] = V(\mu_i) a_i(\phi)$

- Suppose $a_i(\phi) = \phi / w_i$. The score equations become

$$\frac{\partial \ell(\boldsymbol{\beta}, \phi; \boldsymbol{y})}{\partial \phi} = \sum_{i=1}^{n} - \frac{w_i \left(y_i \theta_i - b(\theta_i) \right)}{\phi^2} + c'(y_i, \phi) = 0$$

$$\frac{\partial \ell(\boldsymbol{\beta}, \phi; \boldsymbol{y})}{\partial \beta_j} = \sum_{i=1}^{n} w_i \frac{\partial \mu_i}{\partial \eta_i} \frac{X_{j,i}}{V(\mu_i)} (y_i - \mu_i) = 0$$

- Notice that the $(p+1)$ score equations for $\boldsymbol{\beta}$ do not depend on ϕ
- Consequently, obtaining the MLE of $\boldsymbol{\beta}$ doesn't require knowledge of ϕ
  - ⋆ ϕ isn't required to be known or estimated (if unknown)
  - ⋆ for example, in linear regression we don't need σ^2 (or $\hat{\sigma}^2$) to obtain

$$\widehat{\boldsymbol{\beta}}_{\text{MLE}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

  - ⋆ inference does require an estimate of ϕ (see below)

Asymptotic sampling distribution

- From standard likelihood theory, subject to appropriate regularity conditions,

$$\sqrt{n}((\widehat{\boldsymbol{\beta}}_{\text{MLE}}, \widehat{\phi}_{\text{MLE}}) - (\boldsymbol{\beta}, \phi)) \longrightarrow \text{MVN}\left(\mathbf{0},\ \mathcal{I}(\boldsymbol{\beta}, \phi)^{-1}\right)$$

- To get the asymptotic variance, we first need to derive expressions for the second partial derivatives:

$$\frac{\partial^2 \ell(\boldsymbol{\beta}, \phi; y_i)}{\partial \beta_j \partial \beta_k} = \frac{\partial}{\partial \beta_k}\left\{\frac{\partial \mu_i}{\partial \eta_i} \frac{X_{j,i}}{V(\mu_i) a_i(\phi)}(y_i - \mu_i)\right\}$$

$$= (y_i - \mu_i)\frac{\partial}{\partial \beta_k}\left\{\frac{\partial \mu_i}{\partial \eta_i} \frac{X_{j,i}}{V(\mu_i) a_i(\phi)}\right\} - \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{X_{j,i} X_{k,i}}{V(\mu_i) a_i(\phi)}$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta}, \phi; y_i)}{\partial \beta_j \partial \phi} = \frac{\partial}{\partial \phi} \left\{ \frac{\partial \mu_i}{\partial \eta_i} \frac{X_{j,i}}{V(\mu_i) a_i(\phi)} (y_i - \mu_i) \right\}$$

$$= - \frac{a_i'(\phi)}{a_i(\phi)^2} \frac{\partial \mu_i}{\partial \eta_i} \frac{X_{j,i}}{V(\mu_i)} (y_i - \mu_i)$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta}, \phi; y_i)}{\partial \phi \partial \phi} = \frac{\partial}{\partial \phi} \left\{ - \frac{a_i'(\phi)}{a_i(\phi)^2} (y_i \theta_i - b(\theta_i)) + c'(y_i, \phi) \right\}$$

$$= - \left\{ \frac{a_i(\phi)^2 a_i''(\phi) - 2 a_i(\phi) a_i'(\phi)^2}{a_i(\phi)^4} \right\} (y_i \theta_i - b(\theta_i)) + c''(y_i, \phi)$$

$$= - K(\phi) (y_i \theta_i - b(\theta_i)) + c''(y_i, \phi)$$

- Upon taking expectations with respect to Y, we find that

$$- \mathsf{E}\left[\frac{\partial^2 \ell(\boldsymbol{\beta}, \phi; \boldsymbol{y})}{\partial \beta_j \partial \beta_k}\right] = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{X_{j,i} X_{k,i}}{V(\mu_i) a_i(\phi)}$$

- The second expression has mean zero, so that

$$- \mathsf{E}\left[\frac{\partial^2 \ell(\boldsymbol{\beta}, \phi; \boldsymbol{y})}{\partial \beta_j \partial \phi}\right] = 0$$

- Taking the expectation of the negative of the third expression gives:

$$- \mathsf{E}\left[\frac{\partial^2 \ell(\boldsymbol{\beta}, \phi; \boldsymbol{y})}{\partial \phi \partial \phi}\right] = \sum_{i=1}^{n} K(\phi) \left(b'(\theta_i)\theta_i - b(\theta_i)\right) - \mathsf{E}[c''(Y_i, \phi)]$$

- The expected information matrix can therefore be written in block-diagonal form:

$$\mathcal{I}(\boldsymbol{\beta}, \phi) \;=\; \begin{bmatrix} \mathcal{I}_{\beta\beta} & \mathbf{0} \\ \mathbf{0} & \mathcal{I}_{\phi\phi} \end{bmatrix}$$

  where the components of $\mathcal{I}_{\beta\beta}$ are given by the first expression on the previous slide and the $\mathcal{I}_{\phi\phi}$ is given by the last expression on the previous slide

- The inverse of the information matrix is gives the asymptotic variance

$$\mathsf{V}[\widehat{\boldsymbol{\beta}}_{\mathsf{MLE}}, \widehat{\phi}_{\mathsf{MLE}}] \;=\; \mathcal{I}(\boldsymbol{\beta}, \phi)^{-1} = \begin{bmatrix} \mathcal{I}_{\beta\beta}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathcal{I}_{\phi\phi}^{-1} \end{bmatrix}$$

- The block-diagonal structure $\text{V}[\widehat{\boldsymbol{\beta}}_{\text{MLE}}, \widehat{\phi}_{\text{MLE}}]$ indicates that for GLMs valid characterization of the uncertainty in our estimate of $\boldsymbol{\beta}$ does not require the propagation of uncertainty in our estimation of ϕ

- For example, for linear regression of Normally distributed response data we plug in an estimate of σ^2 into

$$\text{V}[\widehat{\boldsymbol{\beta}}_{\text{MLE}}] \;=\; \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

  ⋆ we typically don't plug in $\hat{\sigma}^2_{\text{MLE}}$ but, rather, an unbiased estimate:

$$\hat{\sigma}^2 \;=\; \frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - X_i^T\widehat{\boldsymbol{\beta}}_{\text{MLE}})^2$$

  ⋆ further, we don't worry about the fact that what we plug in is *an estimate* of σ^2

- For GLMs, therefore, estimation of the variance of $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$ proceeds by plugging in the values of $(\widehat{\boldsymbol{\beta}}_{\text{MLE}}, \widehat{\phi})$ into the upper $(p+1)\times(p+1)$ sub-matrix:

$$\widehat{\mathsf{V}}[\widehat{\boldsymbol{\beta}}_{\text{MLE}}] \;=\; \widehat{\mathcal{I}}_{\beta\beta}^{-1}$$

where $\widehat{\phi}$ is *any* consistent estimator of ϕ

Matrix notation

- If we set

$$W_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{1}{V(\mu_i)a_i(\phi)}$$

then the $(j,k)^{th}$ element of $\mathcal{I}_{\beta\beta}$ can be expressed as

$$\sum_{i=1}^{n} W_i X_{j,i} X_{k,i}$$

- We can therefore write:

$$\mathcal{I}_{\beta\beta} = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}$$

where $\boldsymbol{W}$ is an $n \times n$ diagonal matrix with entries W_i, $i = 1, \ldots, n$, and $\boldsymbol{X}$ is the design matrix from the specification of the linear predictor

Special case: canonical link function

- For the canonical link function, $\eta_i = g(\mu_i) = \theta_i(\mu_i)$, so that

$$\frac{\partial\theta_i}{\partial\eta_i} = 1 \qquad \Rightarrow \qquad \frac{\partial\mu_i}{\partial\eta_i} = \frac{\partial\mu_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\eta_i} = \frac{\mathsf{V}[Y_i]}{a_i(\phi)} = V(\mu_i)$$

- The score contribution for β_j by the i^{th} unit simplifies to

$$\frac{\partial\ell(\boldsymbol{\beta},\phi;y_i)}{\partial\beta_j} = \frac{\partial\mu_i}{\partial\eta_i}\frac{X_{j,i}}{V(\mu_i)a_i(\phi)}(y_i-\mu_i) = \frac{X_{j,i}}{a_i(\phi)}(y_i-\mu_i)$$

and the components of the sub-matrix for $\boldsymbol{\beta}$ of the expected information matrix, $\mathcal{I}_{\beta\beta}$, are the summation of

$$-\mathsf{E}\left[\frac{\partial^2\ell(\boldsymbol{\beta},\phi;y_i)}{\partial\beta_j\partial\beta_k}\right] = \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\frac{X_{j,i}X_{k,i}}{V(\mu_i)a_i(\phi)} = \frac{V(\mu_i)X_{j,i}X_{k,i}}{a_i(\phi)}$$

Hypothesis testing

- For the linear predictor $X_i^T\boldsymbol{\beta}$, suppose we partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and we are interested in testing:

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{1,0} \quad \text{vs} \quad H_a : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_{1,0}$$

  - ⋆ length of $\boldsymbol{\beta}_1$ is $q \leq (p+1)$
  - ⋆ $\boldsymbol{\beta}_2$ is left arbitrary

- In most settings, $\boldsymbol{\beta}_{1,0} = \mathbf{0}$ which represents some form of 'no effect'
  - ⋆ at least given the structure of the model

- Following our review of asymptotic theory, there are three common hypothesis testing frameworks

- Wald test:

  ⋆ let $\widehat{\boldsymbol{\beta}}_{\text{MLE}} = (\widehat{\boldsymbol{\beta}}_{1,\text{MLE}}, \widehat{\boldsymbol{\beta}}_{2,\text{MLE}})$

  ⋆ under H_0

$$(\widehat{\boldsymbol{\beta}}_{1,\text{MLE}} - \boldsymbol{\beta}_{1,0})^T \widehat{\text{V}}[\widehat{\boldsymbol{\beta}}_{1,\text{MLE}}]^{-1} (\widehat{\boldsymbol{\beta}}_{1,\text{MLE}} - \boldsymbol{\beta}_{1,0}) \longrightarrow_d \chi^2_q$$

  where $\widehat{\text{V}}[\widehat{\boldsymbol{\beta}}_{1,\text{MLE}}]$ is the inverse of the $q \times q$ sub-matrix of $\mathcal{I}_{\beta\beta}$ corresponding to $\boldsymbol{\beta}_1$, evaluated at $\widehat{\boldsymbol{\beta}}_{1,\text{MLE}}$

- Score test:

  ⋆ let $\widehat{\boldsymbol{\beta}}_{0,\text{MLE}} = (\boldsymbol{\beta}_{1,0}, \widehat{\boldsymbol{\beta}}_{2,\text{MLE}})$ denote the MLE under H_0

  ⋆ under H_0

$$\boldsymbol{U}(\widehat{\boldsymbol{\beta}}_{0,\text{MLE}}; \boldsymbol{y}) \mathcal{I}(\widehat{\boldsymbol{\beta}}_{0,\text{MLE}})^{-1} \boldsymbol{U}(\widehat{\boldsymbol{\beta}}_{0,\text{MLE}}; \boldsymbol{y}) \longrightarrow_d \chi^2_q$$

- Likelihood ratio test:

  ⋆ obtain the 'best fitting model' without restrictions: $\widehat{\boldsymbol{\theta}}_{\text{MLE}}$

  ⋆ obtain the 'best fitting model' under H_0: $\widehat{\boldsymbol{\theta}}_{0,\text{MLE}}$

  ⋆ under H_0

$$2(\ell(\widehat{\boldsymbol{\beta}}_{\text{MLE}};\boldsymbol{y}) - \ell(\widehat{\boldsymbol{\beta}}_{0,\text{MLE}};\boldsymbol{y})) \longrightarrow_d \chi^2_q$$

Iteratively re-weighted least squares

- We saw that the score equation for β_j is

$$\frac{\partial \ell(\boldsymbol{\beta}, \phi; \boldsymbol{y})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \eta_i} \frac{X_{j,i}}{V(\mu_i) a_i(\phi)} (y_i - \mu_i) = 0$$

  - ⋆ estimation of $\boldsymbol{\beta}$ requires solving $(p+1)$ of these equations simultaneously
  - ⋆ tricky because $\boldsymbol{\beta}$ appears in several places

- A general approach to finding roots is the Newton-Raphson algorithm
  - ⋆ iterative procedure based on the gradient

- For a GLM, the gradient is the derivative of the score function with respect to $\boldsymbol{\beta}$
  - ⋆ these form the components of the observed information matrix $\boldsymbol{I}_{\beta\beta}$

- *Fisher scoring* is an adaptation of the Newton-Raphson algorithm that uses the expected information, $\mathcal{I}_{\beta\beta}$, rather than $\boldsymbol{I}_{\beta\beta}$, for the update

- Suppose the current estimate of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}^{(r)}$
  - ⋆ compute the following:

$$
\begin{aligned}
\eta_i^{(r)} &= X_i^T \widehat{\boldsymbol{\beta}}^{(r)} \\
\mu_i^{(r)} &= g^{-1}\left(\eta_i^{(r)}\right) \\
W_i^{(r)} &= \left(\left. \frac{\partial \mu_i}{\partial \eta_i} \right|_{\eta_i^{(r)}} \right)^2 \frac{1}{V\left(\mu_i^{(r)}\right)} \\
z_i^{(r)} &= \eta_i^{(r)} + \left(y_i - \mu_i^{(r)}\right) \left. \frac{\partial \eta_i}{\partial \mu_i} \right|_{\mu_i^{(r)}}
\end{aligned}
$$

  - ⋆ W_i is called the 'working weight'
  - ⋆ z_i is called the 'adjusted response variable'

- The updated value of $\widehat{\boldsymbol{\beta}}$ is obtained as the WLS estimate to the regression of Z on X:

$$\widehat{\boldsymbol{\beta}}^{(r+1)} = (\boldsymbol{X}^T\boldsymbol{W}^{(r)}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{W}^{(r)}\boldsymbol{Z}^{(r)})$$

  - ⋆ $\boldsymbol{X}$ is the $n \times (p+1)$ design matrix from the initial specification of the model
  - ⋆ $\boldsymbol{W}^{(r)}$ is a diagonal $n \times n$ matrix with entries $\{W_1^{(r)}, \ldots, W_n^{(r)}\}$
  - ⋆ $\boldsymbol{Z}^{(r)}$ is the n-vector $(z_1^{(r)}, \ldots, z_n^{(r)})$

- Iterate until the value of $\widehat{\boldsymbol{\beta}}$ converges
  - ⋆ i.e. the difference between $\widehat{\boldsymbol{\beta}}^{(r+1)}$ and $\widehat{\boldsymbol{\beta}}^{(r)}$ is 'small'

Fitting GLMs in R with `glm()`

- A generic call to `glm()` is given by

```
fit0 <- glm(formula, family, data, ...)
```

  - ⋆ many other arguments that control various aspects of the model/fit
  - ⋆ `?glm` for more information

- '`data`' specifies the data frame containing the response and covariate data

- '`formula`' specifies the structure of linear predictor, $\eta_i = X_i^T \boldsymbol{\beta}$
  - ⋆ input is an object of class 'formula'
  - ⋆ typical input might be of the form:

```
Y ~ X1 + X2 + X3
```

  - ⋆ `?formula` for more information

- '`family`' jointly specifies the probability distribution $f_Y(\cdot)$, link function $g(\cdot)$ and variance function $V(\cdot)$
  - ⋆ most common distributions have already been implemented
  - ⋆ input is an object of class 'family'
    - ∗ object is a list of elements describing the details of the GLM

- The call for a standard logistic regression for binary data might be of the form:

```
glm(Y ~ X1 + X2, family=binomial(), data=myData)
```

or, more simply,

```
glm(Y ~ X1 + X2, family=binomial, data=myData)
```

- A more detailed look at family objects:

```
> ##
> ?family
> poisson()

Family: poisson
Link function: log
> ##
> myFamily <- binomial()
> myFamily

Family: binomial
Link function: logit
> names(myFamily)
 [1] "family"     "link"        "linkfun"    "linkinv"    "variance"
     "dev.resids" "aic"
 [8] "mu.eta"     "initialize" "validmu"    "valideta"   "simulate"
> myFamily$link
[1] "logit"
```

```
> myFamily$variance
function (mu)
mu * (1 - mu)
>
> ## Changing the link function
> ##  * for a true 'log-linear' model we'd need to make appropriate
> ##    changes to the other components of the family object
> ##
> myFamily$link <- "log"
>
> ## Standard logistic regression
> ##
> fit0 <- glm(Y ~ X, family=binomial)
>
> ## log-linear model for binary data
> ##
> fit1 <- glm(Y ~ X, family=binomial(link = "log"))
>
> ## which is (currently) not the same as
> ##
> fit1 <- glm(Y ~ X, family=myFamily)
```

- Once you've fit a GLM you can examine the components of the `glm` object:

```
> ##
> names(fit0)
 [1] "coefficients"      "residuals"         "fitted.values"     "effects"
 [5] "R"                 "rank"              "qr"                "family"
 [9] "linear.predictors" "deviance"          "aic"               "null.deviance"
[13] "iter"              "weights"           "prior.weights"     "df.residual"
[17] "df.null"           "y"                 "converged"         "boundary"
[21] "model"             "call"              "formula"           "terms"
[25] "data"              "offset"            "control"           "method"
[29] "contrasts"         "xlevels"
>
> ##
> names(summary(fit0))
 [1] "call"           "terms"          "family"         "deviance"       "aic"
 [6] "contrasts"      "df.residual"    "null.deviance"  "df.null"        "iter"
[11] "deviance.resid" "coefficients"   "aliased"        "dispersion"     "df"
[16] "cov.unscaled"   "cov.scaled"
```

The deviance

- Recall, the contribution to the log-likelihood by the i^{th} study unit is

$$\ell(\theta_i, \phi; y_i) \;=\; \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} \;+\; c(y_i, \phi)$$

- Implicitly, θ_i is a function of μ_i so we could write the log-likelihood contribution as a function of μ_i:

$$\ell(\theta_i, \phi; y_i) \;\Rightarrow\; \ell(\mu_i, \phi; y_i)$$

- Given $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$, we can compute each $\hat{\mu}_i$ and evaluate

$$\ell(\widehat{\boldsymbol{\mu}}, \phi; \boldsymbol{y}) \;=\; \sum_{i=1}^{n} \ell(\hat{\mu}_i, \phi; y_i),$$

  ⋆ the maximum log-likelihood

- $\ell(\widehat{\boldsymbol{\mu}}, \phi; \boldsymbol{y})$ is the maximum achievable log-likelihood given the structure of the model
  - ⋆ μ_i is modeled via $g(\mu_i) = \eta_i = X_i^T \boldsymbol{\beta}$
  - ⋆ any other value of $\boldsymbol{\beta}$ would correspond to a lower value of the log-likelihood

- The overall maximum achievable log-likelihood, however, is one based on a *saturated model*
  - ⋆ same number of parameters as observations
  - ⋆ each observation is its own mean: $\mu_i = y_i$

$$\ell(\boldsymbol{y}, \phi; \boldsymbol{y}) = \sum_{i=1}^{n} \ell(y_i, \phi; y_i),$$

  - ⋆ this represents the 'best possible fit'

- The difference

$$D^*(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}) \;=\; 2\left[\ell(\boldsymbol{y}, \phi; \boldsymbol{y}) \;-\; \ell(\widehat{\boldsymbol{\mu}}, \phi; \boldsymbol{y})\right]$$

  is called the *scaled deviance*

- Let
  - ⋆ $\tilde{\theta}_i$ be the value of θ_i based on setting $\mu_i = y_i$
  - ⋆ $\hat{\theta}_i$ be the value of θ_i based on setting $\mu_i = \hat{\mu}_i$

- If we take $a_i(\phi) = \phi/w_i$, then

$$D^*(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}) \;=\; \sum_{i=1}^{n} \frac{2w_i}{\phi}\left[y_i(\tilde{\theta}_i - \hat{\theta}_i) \;-\; b(\tilde{\theta}_i) \;+\; b(\hat{\theta}_i)\right] \;=\; \frac{D(\boldsymbol{y}, \widehat{\boldsymbol{\mu}})}{\phi}$$

- $D(\boldsymbol{y}, \widehat{\boldsymbol{\mu}})$ is the *deviance* for the current model

- $D(\boldsymbol{y}, \widehat{\boldsymbol{\mu}})$ is used as a measure of goodness of fit of the model to the data
  - ⋆ measures the 'discrepancy' between the fitted model and the data

- For the Normal distribution, the deviance is the sum of squared residuals:

$$D(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}) \;=\; \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$$

  - ⋆ has an exact χ^2 distribution
  - ⋆ compare two nested models by taking the difference in their deviances
    - ∗ distribution of the difference is still a χ^2
    - ∗ the likelihood ratio test

- Beyond the Normal distribution the deviance is not χ^2

- But we still can rely on a χ^2 approximation to the asymptotic sampling distribution of the *difference* in the deviance between two models

Residuals

- In the context of regression modeling, residuals are used primarily to
  - ⋆ examine the adequacy of model fit
    - ∗ functional form for terms in the linear predictor
    - ∗ link function
    - ∗ variance function
  - ⋆ investigate potential data issues
    - ∗ e.g. outliers
- Interpreted as representing variation in the outcome that is not explained by the model
  - ⋆ variation once the systematic component has been accounted for
  - ⋆ residuals are therefore *model-specific*

- An ideal residual would look like an i.i.d sample when the correct mean model is fit

- For linear regression, we often consider the *raw* or *response residual*

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

  - ⋆ if the ϵ_i are homoskedastic then $\{r_1, \ldots, r_n\}$ will be i.i.d

- For GLMs the underlying probability distribution is often skewed and exhibits a mean-variance relationship

- *Pearson residuals* account for the heteroskedasticity via standardization

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

  - ⋆ Pearson χ^2 statistic for goodness-of-fit is equal to $\sum_i (r_i^p)^2$

- The *deviance residual* is defined as

$$r_i^d \;=\; \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$$

where d_i is the contribution to $D(\boldsymbol{y}, \widehat{\boldsymbol{\mu}})$ from the i^{th} study unit

  - ⋆ why is this a reasonable quantity to consider?

- Pierce and Schafer (JASA, 1986) examined various residuals for GLMs
  - ⋆ conclude that deviance residuals are 'a very good choice'
  - ⋆ very nearly normally distributed after one allows for the discreteness
  - ⋆ continuity correction which replaces

$$y_i \;\Rightarrow\; y_i \pm \frac{1}{2}$$

  in the definition of the residual

  ∗ $+/-$ chosen to move the value closer to $\hat{\mu}_i$

- All three types of residuals are returned by `glm()` in R:

```
> ## generic (logistic regression) model
> fit0 <- glm(Y ~ X, family=binomial)
>
> args(residuals.glm)
function (object, type = c("deviance", "pearson", "working",
    "response", "partial"), ...)
NULL
>
> ## deviance residuals are the default
> residual(fit0)
...
>
> ## extracting the pearson residuals
> residual(fit0, type="pearson")
...
```

The Bayesian solution

- A GLM is specified by:

$$
\begin{aligned}
Y_i|X_i &\sim f_Y(y;\mu_i,\phi) \\
\mathsf{E}[Y_i|X_i] &= g^{-1}(X_i^T\boldsymbol{\beta}) = \mu_i \\
\mathsf{V}[Y_i|X_i] &= V(\mu_i)a_i(\phi)
\end{aligned}
$$

  - ⋆ $f_Y(\cdot)$ is a member of the exponential dispersion family
  - ⋆ $\boldsymbol{\beta}$ is a vector of regression coefficients
  - ⋆ ϕ is the dispersion parameter

- $(\boldsymbol{\beta}, \phi)$ are the unknown parameters
  - ⋆ note there might not necessarily be a dispersion parameter
  - ⋆ e.g. for binary or Poisson data

- Required to specify a prior distribution for $(\boldsymbol{\beta}, \phi)$ which is often factored into

$$\pi(\boldsymbol{\beta}, \phi) \;=\; \pi(\boldsymbol{\beta}|\phi)\pi(\phi)$$

- For $\boldsymbol{\beta}|\phi$, strategies include
  - ⋆ a flat, non-informative prior
    - ∗ recover the classical analysis
    - ∗ posterior mode corresponding to a uniform prior density is the MLE
  - ⋆ an informative prior
    - ∗ e.g., $\boldsymbol{\beta} \sim$ MVN($\boldsymbol{\beta}_0$, Σ_β)
    - ∗ convenient choice given the computational methods described below

- Unfortunately, specifying a prior for ϕ is less prescriptive
  - ⋆ consider specific models in Parts V-VII of the notes

- Given an independent sample Y_1, ..., Y_n, the likelihood is the product of n terms:

$$\mathcal{L}(\boldsymbol{\beta}, \phi | \boldsymbol{y}) \; = \; \prod_{i=1}^{n} f_Y(y_i | \mu_i, \phi)$$

- Apply Bayes' Theorem to get the posterior:

$$\pi(\boldsymbol{\beta}, \phi | \boldsymbol{y}) \; \propto \; \mathcal{L}(\boldsymbol{\beta}, \phi | \boldsymbol{y}) \pi(\boldsymbol{\beta}, \phi)$$

Computation

- For most GLMs, the posterior won't be of a convenient form
  - ⋆ analytically intractable
- Use Monte Carlo methods to summarize the posterior distribution
- We've seen that the Gibbs sampler and the Metropolis-Hastings algorithm are powerful tools for generating samples from the posterior distribution
  - ⋆ need to specify a proposal distribution
  - ⋆ need to specify starting values for the Markov chain(s)
- Towards this, let $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\beta}}, \tilde{\phi})$ denote the posterior mode

- Consider a Taylor series expansion of the log-posterior at $\widetilde{\boldsymbol{\theta}}$:

$$
\begin{aligned}
\log \pi(\boldsymbol{\theta}|\boldsymbol{y}) \;=\; & \log \pi(\widetilde{\boldsymbol{\theta}}|\boldsymbol{y}) \\
& + \; (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}) \, \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}|\boldsymbol{y}) \Big|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}} \\
& + \; \frac{1}{2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^T \left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}|\boldsymbol{y}) \right]_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}) \\
& + \; \ldots
\end{aligned}
$$

- Ignore the $\log \pi(\widetilde{\boldsymbol{\theta}}|\boldsymbol{y})$ term because, as a function of $\boldsymbol{\theta}$, it is constant

- The linear term in the expansion disappears because the first derivative of the log-posterior at the mode is equal to 0

- The middle component of the quadratic term is approximately the negative observed information matrix, evaluated at the mode

- We therefore get

$$\log \pi(\boldsymbol{\theta}|\boldsymbol{y}) \;\approx\; -\frac{1}{2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^T \boldsymbol{I}(\widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})$$

  which is the log of the kernel for a Normal distribution

- So, towards specifying a proposal distribution for the Metropolis-Hastings algorithm, we can consider the following Normal approximation to the posterior

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \;\approx\; \text{Normal}\left(\widetilde{\boldsymbol{\theta}},\; \boldsymbol{I}(\widetilde{\boldsymbol{\theta}})^{-1}\right)$$

Q: How can we make use of this for sampling from the posterior $\pi(\boldsymbol{\beta}, \phi|\boldsymbol{y})$?

  ⋆ there are many approaches that one could take

  ⋆ we'll describe three

- First, we need to find the mode, $(\widetilde{\boldsymbol{\beta}}, \tilde{\phi})$
  - ⋆ the value that maximizes $\pi(\boldsymbol{\beta}, \phi | \boldsymbol{y})$
  - ⋆ given a non-informative prior:

$$(\widetilde{\boldsymbol{\beta}}, \tilde{\phi}) \ \equiv \ (\widehat{\boldsymbol{\beta}}_{\mathsf{MLE}}, \hat{\phi}_{\mathsf{MLE}})$$

    - ∗ obtain the mode via the IRLS algorithm
  - ⋆ otherwise, use any other standard optimization technique
    - ∗ e.g. Newton-Raphson
    - ∗ could use $(\widehat{\boldsymbol{\beta}}_{\mathsf{MLE}}, \hat{\phi}_{\mathsf{MLE}})$ as a starting point

- Next, recall the block-diagonal structure of the information matrix for a GLM:

$$\mathcal{I}(\boldsymbol{\beta}, \phi) \ = \ \begin{bmatrix} \mathcal{I}_{\beta\beta} & \mathbf{0} \\ \mathbf{0} & \mathcal{I}_{\phi\phi} \end{bmatrix}$$

- Exploit this and consider the approximation:

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}) \;\approx\; \text{Normal}\left(\widetilde{\boldsymbol{\beta}},\; V_\beta(\widetilde{\boldsymbol{\beta}}, \tilde{\phi})\right)$$

  to the marginal posterior of $\boldsymbol{\beta}$

  ⋆ $V_\beta(\widetilde{\boldsymbol{\beta}}, \tilde{\phi}) = \boldsymbol{I}^{-1}_{\beta\beta}$ evaluated at the mode

  ⋆ denote the approximation by $\widetilde{\pi}(\boldsymbol{\beta}; \boldsymbol{y})$

- Also consider the approximation:

$$\pi(\phi|\boldsymbol{y}) \;\approx\; \text{Normal}\left(\tilde{\phi},\; \widetilde{V}_\phi(\widetilde{\boldsymbol{\beta}}, \tilde{\phi})\right)$$

  to the marginal posterior of ϕ

  ⋆ $V_\phi(\widetilde{\boldsymbol{\beta}}, \tilde{\phi}) = \boldsymbol{I}^{-1}_{\phi\phi}$ evaluated at the mode

  ⋆ denote the approximation by $\widetilde{\pi}(\phi|\boldsymbol{y})$

Approach #1

- If we believe that $\widetilde{\pi}(\boldsymbol{\beta}|\boldsymbol{y})$ is a good approximation, we could simply report summary statistics directly from the multivariate Normal distribution

$$\boldsymbol{\beta}|\boldsymbol{y} \sim \text{Normal}\left(\widetilde{\boldsymbol{\beta}},\ V_\beta(\widetilde{\boldsymbol{\beta}}, \tilde{\phi})\right)$$

  - ⋆ report the posterior mean (equivalently, the posterior median)
  - ⋆ posterior credible intervals using the components of $V_\beta(\widetilde{\boldsymbol{\beta}}, \tilde{\phi})$

- The approach conditions on $\tilde{\phi}$
  - ⋆ uncertainty in the true value of ϕ is ignored
  - ⋆ this is what we do in classical estimation/inference for linear regression anyway

- Similarly, we could summarize features of the posterior distribution of ϕ using the $\tilde{\pi}(\phi|\boldsymbol{y})$ Normal approximation

Approach #2

- We may not be willing to believe that the approximation is good enough to summarize features of $\pi(\boldsymbol{\beta}; \boldsymbol{y})$
  - ⋆ approximation may not be good in small samples
  - ⋆ approximation may not be good in the tails of the distribution
    - ∗ away from the posterior mode

- We could use $\widetilde{\pi}(\boldsymbol{\beta}|\boldsymbol{y})$ as a proposal distribution in a Metropolis-Hastings algorithm to sample from the exact posterior $\pi(\boldsymbol{\beta}; \boldsymbol{y})$

- Let $\boldsymbol{\beta}^{(r)}$ be the current state in the sequence

(1) generate a proposal $\boldsymbol{\beta}^*$ from $\widetilde{\pi}(\boldsymbol{\beta}|\boldsymbol{y})$
   - ∗ straightforward since this is a multivariate Normal distribution

(2) evaluate the acceptance ratio

$$a_r = \min\left(1, \frac{\pi(\boldsymbol{\beta}^*|\boldsymbol{y}, \tilde{\phi})}{\pi(\boldsymbol{\beta}^{(r)}|\boldsymbol{y}, \tilde{\phi})} \frac{\widetilde{\pi}(\boldsymbol{\beta}^{(r)}|\boldsymbol{\beta}^*)}{\widetilde{\pi}(\boldsymbol{\beta}^*|\boldsymbol{\beta}^{(r)})}\right)$$

$$= \min\left(1, \frac{\pi(\boldsymbol{\beta}^*|\boldsymbol{y}, \tilde{\phi})}{\pi(\boldsymbol{\beta}^{(r)}|\boldsymbol{y}, \tilde{\phi})} \frac{\widetilde{\pi}(\boldsymbol{\beta}^{(r)})}{\widetilde{\pi}(\boldsymbol{\beta}^*)}\right)$$

(3) generate a random $U \sim$ Uniform(0, 1)

- *reject* the proposal if $a_r < U$:

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)}$$

- *accept* the proposal if $a_r \geq U$:

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^*$$

Approach #3

- While approach #2 facilitates sampling from the exact posterior distribution of $\boldsymbol{\beta}$, $\pi(\boldsymbol{\beta}|\boldsymbol{y})$, uncertainty in the value of ϕ is still ignored
  - ⋆ condition on $\phi = \tilde{\phi}$

- To sample from the full exact posterior $\pi(\boldsymbol{\beta}, \phi; \boldsymbol{y})$ we could implement a Gibbs sampling scheme and iterate between the full conditionals
  - ⋆ for each, implement a Metropolis-Hastings step using the approximations we've developed
  - ⋆ for the r^{th} sample:

    (1) sample $\boldsymbol{\beta}^{(r)}$ from $\pi(\boldsymbol{\beta}|\ \phi^{(r-1)}; \boldsymbol{y})$ with $\widetilde{\pi}(\boldsymbol{\beta}|\boldsymbol{y})$ as a proposal

    (2) sample $\phi^{(r)}$ from $\pi(\phi|\ \boldsymbol{\beta}^{(r)}; \boldsymbol{y})$ with $\widetilde{\pi}(\phi|\boldsymbol{y})$ as a proposal

- Use the approximations to generate starting values for the Markov chain(s)