Part IV:

Theory of Generalized Linear Models
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

\[ E[Y_i|X_i] = X_i^T \beta \]

* estimate \( \beta \) 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, \( \hat{\beta}_{OLS} \) 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 \( \epsilon \sim \text{Normal}(0, \sigma^2) \) and BLUE otherwise
For a binary response variable, we could specify a linear regression model:

\[ E[Y_i|X_i] = X_i^T \beta \]
\[ Y_i|X_i \sim \text{Bernoulli}(\mu_i) \]

where, for notational convenience, \( \mu_i = X_i^T \beta \)

As long as this model is correctly specified, \( \hat{\beta}_{\text{OLS}} \) will still be unbiased.

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

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

- 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, \( \hat{\beta}_{\text{WLS}} \) is the solution to

\[
0 = \frac{\partial}{\partial \beta} \text{RSS}(\beta; W)
\]

\[
0 = \frac{\partial}{\partial \beta} \sum_{i=1}^{n} w_i (y_i - X_i^T \beta)^2
\]

\[
0 = \sum_{i=1}^{n} X_i w_i (y_i - X_i^T \beta)
\]
• For a binary response, we know the form of $V[Y_i]$
  
  - estimate $\beta$ by setting $W = \Sigma^{-1}$, a diagonal matrix with elements:
    
    $$w_i = \frac{1}{\mu_i(1 - \mu_i)}$$

• From the Gauss-Markov Theorem, the resulting estimator is BLUE
  
  $$\hat{\beta}_{\text{GLS}} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$

• 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 $\hat{\beta}_{\text{GLS}}$ while simultaneously accommodating the mean-variance relationship
• We can also show that $\hat{\beta}_{\text{GLS}}$, obtained via the IWLS algorithm, is the MLE

  ★ firstly, note that the likelihood and log-likelihood are:

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

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

  ★ to get the MLE, we take derivatives, set them equal to zero and solve

  ★ following the algebra trail we find that

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

• The score equations are equivalent to the least squares equations

  ★ $\hat{\beta}_{\text{GLS}}$ 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: \( \mathbb{E}[Y_i | X_i] \in (0, 1) \)
- count: \( \mathbb{E}[Y_i | X_i] \in (0, \infty) \)

but our current specification of the mean model doesn’t impose any restrictions

\[
\mathbb{E}[Y_i | X_i] = X_i^T \beta
\]

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, $X_i^T \beta$
  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

<table>
<thead>
<tr>
<th>Response type</th>
<th>Range</th>
<th>Possible distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>$(-\infty, \infty)$</td>
<td>Normal($\mu, \sigma^2$)</td>
</tr>
<tr>
<td>Binary</td>
<td>{0, 1}</td>
<td>Bernoulli($\pi$)</td>
</tr>
<tr>
<td>Polytomous</td>
<td>{1, \ldots, K}</td>
<td>Multinomial($\pi_k$)</td>
</tr>
<tr>
<td>Count</td>
<td>{0, 1, \ldots, n}</td>
<td>Binomial($n, \pi$)</td>
</tr>
<tr>
<td>Count</td>
<td>{0, 1, \ldots}</td>
<td>Poisson($\mu$)</td>
</tr>
<tr>
<td>Continuous</td>
<td>(0, $\infty$)</td>
<td>Gamma($\alpha, \beta$)</td>
</tr>
<tr>
<td>Continuous</td>
<td>(0, 1)</td>
<td>Beta($\alpha, \beta$)</td>
</tr>
</tbody>
</table>

- 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:

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

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

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

$$g(\mu_i) = X_i^T \beta$$

* often use the notation $\eta_i = X_i^T \beta$
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:

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

- \(\theta\) is the *canonical parameter*
- \(\phi\) is the *dispersion parameter*
- \(b(\theta)\) is the *cumulant function*

- Many common distributions are members of this family
\( Y \sim \text{Bernoulli}(\pi) \)

\[
f_Y(y; \pi) = \mu^y (1 - \mu)^{1-y}
\]

\[
f_Y(y; \theta, \phi) = \exp \{ y\theta - \log (1 + \exp \{ \theta \}) \}
\]

\[
\theta = \log \left( \frac{\pi}{1 - \pi} \right)
\]

\[a(\phi) = 1\]

\[b(\theta) = \log (1 + \exp \{ \theta \})\]

\[c(y, \phi) = 0\]
• 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 $\theta$ by $i$: $\theta_i$

• The dispersion parameter has a key relationship with $V[Y]$
  * may but typically does not vary across study units
  * typically no unit-specific index: $\phi$
  * 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 \( \theta_i \) is the score function for \( \theta_i \) 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:

\[
E[U(\theta_i)] = 0
\]

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

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

Since the score has mean zero, we find that

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

and, consequently, that

\[
E[Y_i] = b'(\theta_i)
\]
The second partial derivative of \( \ell(\theta_i, \phi; y_i) \) is

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

\( \star \) 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

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

so that

\[
V[Y_i] = b''(\theta_i)a_i(\phi)
\]
• The variance of $Y_i$ is therefore a function of both $\theta_i$ and $\phi$

• Note that the canonical parameter is a function of $\mu_i$

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

so that we can write

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

• 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 $Y \sim \text{Bernoulli}(\mu)$

$$a(\phi) = 1$$
\[ b(\theta) = \log (1 + \exp\{\theta\}) \]

\[ \mathbb{E}[Y] = b'(\theta) \]
\[ = \frac{\exp\{\theta\}}{1 + \exp\{\theta\}} = \mu \]

\[ \mathbb{V}[Y] = b''(\theta)a(\phi) \]
\[ = \frac{\exp\{\theta\}}{(1 + \exp\{\theta\})^2} = \mu(1 - \mu) \]

\[ V(\mu) = \mu(1 - \mu) \]
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, \( \mu_i = \mathbb{E}[Y_i | X_i] \)
  ★ provides a connection between \( X_i \) and distribution of \( Y_i \) via the canonical parameter \( \theta_i \) and the cumulant function \( b(\theta_i) \)
Specifically, the relationship between $\mu_i$ and $X_i$ is given by

$$g(\mu_i) = X_i^T \beta$$

we ‘link’ the linear predictor to the distribution of $Y_i$ via a transformation of $\mu_i$

Traditionally, this specification is broken down into two parts:

1. the linear predictor, $\eta_i = X_i^T \beta$
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 $\mu_i$ and $X_i$ is *jointly* determined by $\beta$ and $g(\cdot)$
The linear predictor, $\eta_i = X_i^T \beta$

- 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 $\beta_2$ and $\beta_3$ is difficult
Scientifically, one might only be interested in the ‘low birth weight’ threshold

- let \( X_{lbw} = 0/1 \) if birth weight is \( >2.5\text{kg}/\leq 2.5\text{kg} \)

\[
\eta = \beta_0 + \beta_1 X_{lbw}
\]

- 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 $X_i^T \beta$ we need to specify how it is related to the conditional mean $\mu_i$

- As we’ve noted, the range of values that $\mu_i$ 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 $\beta$ subject to the constraint that all (modeled) values of $\mu_i$ respect the appropriate range

**Q:** What might the drawbacks of such an approach be?
An alternative is to permit the estimation of $\beta$ to be ‘free’ but impose a functional form of the relationship between $\mu_i$ and $X_i^T \beta$

- via the link function $g(\cdot)$

$$g(\mu_i) = X_i^T \beta$$

We interpret the link function as specifying a transformation of the conditional mean, $\mu_i$

- we are not specifying a transformation of the response $Y_i$

The inverse of the link function provides the specification of the model on the scale of $\mu_i$

$$\mu_i = g^{-1}(X_i^T \beta)$$

- link functions are therefore usually monotone and have a well-defined inverse
In linear regression we specify

\[ \mu_i = X_i^T \beta \]

* \( g(\cdot) \) is the identity link

In logistic regression we specify

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

* \( g(\cdot) \) is the logit or logistic link

In Poisson regression we specify

\[ \log(\mu_i) = X_i^T \beta \]

* \( g(\cdot) \) is the log link
• For linear regression also we have that

\[ \mu_i = X_i^T \beta \]

\[ g^{-1}(\eta_i) = \eta_i \] is the identity function

• For logistic regression

\[ \mu_i = \frac{\exp \{X_i^T \beta\}}{1 + \exp \{X_i^T \beta\}} \]

\[ g^{-1}(\eta_i) = \expit(\eta_i) \] is the expit function

• For Poisson regression

\[ \mu_i = \exp \{X_i^T \beta\} \]

\[ g^{-1}(\eta_i) = \exp(\eta_i) \] is the exponential function
The canonical link

• Recall that the mean and the canonical parameter are linked via the derivative of the cumulant function
  \[ E[Y_i] = \mu_i = b'(\theta_i) \]

• An important link function is the canonical link:
  \[ g(\mu_i) = \theta(\mu_i) \]
  * the function that results by viewing the canonical parameter \( \theta_i \) as a function of \( \mu_i \)
  * 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:

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

☆ 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 $\mu_i$ can take
  
  2. impact on the interpretability of $\beta$

- 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

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

where $\theta_i$ is a function of $\beta$ and is determined by

- the form of $b'(\theta_i) = \mu_i$
- the choice of the link function via $g(\mu_i) = \eta_i = X_i^T \beta$

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

- Since we’ve fully specified the likelihood, we can proceed with likelihood-based estimation/inference
• There are \((p+2)\) unknown parameters: \((\beta, \phi)\)

• To obtain the MLE we need to solve the score equations:

\[
\left( \frac{\partial \ell(\beta, \phi; y)}{\partial \beta_0}, \ldots, \frac{\partial \ell(\beta, \phi; y)}{\partial \beta_p}, \frac{\partial \ell(\beta, \phi; y)}{\partial \phi} \right)^T = 0
\]

\(\star\) system of \((p+2)\) equations

• The contribution to the score for \(\phi\) by the \(i^{th}\) unit is

\[
\frac{\partial \ell(\beta, \phi; y_i)}{\partial \phi} = - \frac{a'_i(\phi)}{a_i(\phi)^2} (y_i \theta_i - b(\theta_i)) + c'(y_i, \phi)
\]
• We can use the chain rule to obtain a convenient expression for the $i^{th}$ contribution to the score function for $\beta_j$:

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

• Note the following results:

$$\frac{\partial \ell(\beta, \phi; y_i)}{\partial \theta_i} = \frac{y_i - \mu_i}{a_i(\phi)}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)$$

$$= \frac{\text{Var}[Y_i]}{a_i(\phi)}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = X_{j,i}$$
• The score function for $\beta_j$ can therefore be written as

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


* depends on the distribution of $Y_i$ solely through $E[Y_i] = \mu_i$ and $V[Y_i] = V(\mu_i)a_i(\phi)$

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

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

$$\frac{\partial \ell(\beta, \phi; y)}{\partial \beta_j} = \sum_{i=1}^{n} w_i \frac{\partial \mu_i}{\partial \eta_i} \frac{X_{j,i}}{V(\mu_i)}(y_i - \mu_i) = 0$$
• Notice that the \((p+1)\) score equations for \(\beta\) do not depend on \(\phi\)

• Consequently, obtaining the MLE of \(\beta\) doesn’t require knowledge of \(\phi\)
  \begin{itemize}
    \item \(\phi\) isn’t required to be known or estimated (if unknown)
    \item for example, in linear regression we don’t need \(\sigma^2\) (or \(\hat{\sigma}^2\)) to obtain
    \[
    \hat{\beta}_{MLE} = (X^T X)^{-1} X^T Y
    \]
    \item inference does require an estimate of \(\phi\) (see below)
  \end{itemize}
Asymptotic sampling distribution

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

\[ \sqrt{n}( (\hat{\beta}_{\text{MLE}}, \hat{\phi}_{\text{MLE}}) - (\beta, \phi) ) \rightarrow \text{MVN}(0, I(\beta, \phi)^{-1}) \]

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

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

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

\[
\frac{\partial^2 \ell (\beta, \phi; y_i)}{\partial \phi \partial \phi} = \frac{\partial}{\partial \phi} \left\{ - \frac{a_i'(\phi)}{a_i(\phi)^2} (y_i \theta_i - b(\theta_i)) + c'(y_i, \phi) \right\} \\
= - \left\{ \frac{a_i(\phi)^2 a_i''(\phi) - 2a_i(\phi)a_i'(\phi)^2}{a_i(\phi)^4} \right\} (y_i \theta_i - b(\theta_i)) + c''(y_i, \phi) \\
= - K(\phi) (y_i \theta_i - b(\theta_i)) + c''(y_i, \phi)
\]
Upon taking expectations with respect to $Y$, we find that

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

The second expression has mean zero, so that

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

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

$$- \mathbb{E} \left[ \frac{\partial^2 \ell(\beta, \phi; y)}{\partial \phi \partial \phi} \right] = \sum_{i=1}^{n} K(\phi) \left( b'(\theta_i)\theta_i - b(\theta_i) \right) - \mathbb{E}[c''(Y_i, \phi)]$$
The expected information matrix can therefore be written in block-diagonal form:

\[
\mathcal{I}(\beta, \phi) = \begin{bmatrix}
\mathcal{I}_{\beta\beta} & 0 \\
0 & \mathcal{I}_{\phi\phi}
\end{bmatrix}
\]

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

\[
\text{V}[\hat{\beta}_{\text{MLE}}, \hat{\phi}_{\text{MLE}}] = \mathcal{I}(\beta, \phi)^{-1} = \begin{bmatrix}
\mathcal{I}_{\beta\beta}^{-1} & 0 \\
0 & \mathcal{I}_{\phi\phi}^{-1}
\end{bmatrix}
\]
The block-diagonal structure $V[\hat{\beta}_{\text{MLE}}, \hat{\phi}_{\text{MLE}}]$ indicates that for GLMs valid characterization of the uncertainty in our estimate of $\beta$ does not require the propagation of uncertainty in our estimation of $\phi$.

For example, for linear regression of Normally distributed response data we plug in an estimate of $\sigma^2$ into

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

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

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

further, we don’t worry about the fact that what we plug in is an estimate of $\sigma^2$
For GLMs, therefore, estimation of the variance of $\hat{\beta}_{\text{MLE}}$ proceeds by plugging in the values of $(\hat{\beta}_{\text{MLE}}, \hat{\phi})$ into the upper $(p+1) \times (p+1)$ sub-matrix:

$$\hat{V}[\hat{\beta}_{\text{MLE}}] = \hat{I}_{\hat{\beta}\hat{\beta}}^{-1}$$

where $\hat{\phi}$ is any consistent estimator of $\phi$
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:

\[ \mathcal{I}_{\beta \beta} = X^T W X \]

where \(W\) is an \(n \times n\) diagonal matrix with entries \(W_i, i = 1, \ldots, n\), and \(X\) 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

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

- The score contribution for \( \beta_j \) by the \( i^{th} \) unit simplifies to

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

and the components of the sub-matrix for \( \beta \) of the expected information matrix, \( I_{\beta\beta} \), are the summation of

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

- For the linear predictor $X_i^T \beta$, suppose we partition $\beta = (\beta_1, \beta_2)$ and we are interested in testing:

  $$H_0 : \beta_1 = \beta_{1,0} \quad \text{vs} \quad H_a : \beta_1 \neq \beta_{1,0}$$

  * length of $\beta_1$ is $q \leq (p + 1)$
  * $\beta_2$ is left arbitrary

- In most settings, $\beta_{1,0} = 0$ 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 $\hat{\beta}_{\text{MLE}} = (\hat{\beta}_{1,\text{MLE}}, \hat{\beta}_{2,\text{MLE}})$
  
  * under $H_0$

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

  where $\hat{V}[\hat{\beta}_{1,\text{MLE}}]$ is the inverse of the $q \times q$ sub-matrix of $\mathcal{I}_{\beta\beta}$ corresponding to $\beta_1$, evaluated at $\hat{\beta}_{1,\text{MLE}}$

• **Score test:**

  * let $\hat{\beta}_{0,\text{MLE}} = (\beta_{1,0}, \hat{\beta}_{2,\text{MLE}})$ denote the MLE under $H_0$
  
  * under $H_0$

  $$
  U(\hat{\beta}_{0,\text{MLE}}; y) \mathcal{I}(\hat{\beta}_{0,\text{MLE}})^{-1} U(\hat{\beta}_{0,\text{MLE}}; y) \longrightarrow_d \chi^2_q
  $$
• Likelihood ratio test:

- obtain the ‘best fitting model’ without restrictions: $\hat{\theta}_{MLE}$
- obtain the ‘best fitting model’ under $H_0$: $\hat{\theta}_{0,MLE}$
- under $H_0$

$$2(\ell(\hat{\beta}_{MLE}; y) - \ell(\hat{\beta}_{0,MLE}; y)) \xrightarrow{d} \chi^2_q$$
Iteratively re-weighted least squares

- We saw that the score equation for $\beta_j$ is

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

- estimation of $\beta$ requires solving $(p + 1)$ of these equations simultaneously
- tricky because $\beta$ 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 $\beta$
  - these form the components of the observed information matrix $I_{\beta\beta}$
Fisher scoring is an adaptation of the Newton-Raphson algorithm that uses the expected information, $\mathcal{I}_{\beta \beta}$, rather than $I_{\beta \beta}$, for the update.

Suppose the current estimate of $\beta$ is $\hat{\beta}^{(r)}$.

- compute the following:

  \[
  \eta_i^{(r)} = X_i^T \hat{\beta}^{(r)}
  \]
  \[
  \mu_i^{(r)} = g^{-1} \left( \eta_i^{(r)} \right)
  \]
  \[
  W_i^{(r)} = \left( \frac{\partial \mu_i}{\partial \eta_i} \bigg|_{\eta_i^{(r)}} \right)^2 \frac{1}{V \left( \mu_i^{(r)} \right)}
  \]
  \[
  z_i^{(r)} = \eta_i^{(r)} + (y_i - \mu_i^{(r)}) \frac{\partial \eta_i}{\partial \mu_i} \bigg|_{\mu_i^{(r)}}
  \]

- $W_i$ is called the ‘working weight’
- $z_i$ is called the ‘adjusted response variable’
The updated value of $\hat{\beta}$ is obtained as the WLS estimate to the regression of $Z$ on $X$:

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

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

Iterate until the value of $\hat{\beta}$ converges

- i.e. the difference between $\hat{\beta}^{(r+1)}$ and $\hat{\beta}^{(r)}$ is ‘small’
Fitting GLMs in R with \texttt{glm()}

- A generic call to \texttt{glm()} is given by

\[
\text{fit0} \leftarrow \texttt{glm(formula, family, data, ...)}
\]

★ many other arguments that control various aspects of the model/fit
★ \texttt{?glm} for more information

- \texttt{‘data’} specifies the data frame containing the response and covariate data

- \texttt{‘formula’} specifies the structure of linear predictor, \( \eta_i = X_i^T \beta \)

★ input is an object of class ‘formula’
★ typical input might be of the form:

\[
Y \sim X1 + X2 + X3
\]

★ \texttt{?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:

\[
\text{glm}(Y \sim X_1 + X_2, \text{family}=\text{binomial}(), \text{data}=\text{myData})
\]

or, more simply,

\[
\text{glm}(Y \sim X_1 + X_2, \text{family}=\text{binomial}, \text{data}=\text{myData})
\]
• A more detailed look at family objects:

```r
> ##
> ?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:

```r
> ##
> 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, $\theta_i$ is a function of $\mu_i$ so we could write the log-likelihood contribution as a function of $\mu_i$:

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

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

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

\begin{itemize}
  \item the maximum log-likelihood
\end{itemize}
• $\ell(\hat{\mu}, \phi; y)$ is the maximum achievable log-likelihood given the structure of the model
  
  * $\mu_i$ is modeled via $g(\mu_i) = \eta_i = X_i^T \beta$
  * any other value of $\beta$ 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: $\mu_i = y_i$

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

* this represents the ‘best possible fit’
The difference

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

is called the scaled deviance.

Let

\* \( \tilde{\theta}_i \) be the value of \( \theta_i \) based on setting \( \mu_i = y_i \)
\* \( \hat{\theta}_i \) be the value of \( \theta_i \) based on setting \( \mu_i = \hat{\mu}_i \)

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

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

\( D(y, \hat{\mu}) \) is the deviance for the current model.
• $D(y, \hat{\mu})$ 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:

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

  * has an exact $\chi^2$ distribution
  * compare two nested models by taking the difference in their deviances
    * distribution of the difference is still a $\chi^2$
    * the likelihood ratio test

• Beyond the Normal distribution the deviance is not $\chi^2$

• But we still can rely on a $\chi^2$ 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 \( \epsilon_i \) 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

\[ r_{pi} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \]

★ Pearson \( \chi^2 \) statistic for goodness-of-fit is equal to \( \sum_i (r_{pi})^2 \)
The *deviance residual* is defined as

\[ r^d_i = \text{sign}(y_i - \hat{\mu}_i) \sqrt{d_i} \]

where \( d_i \) is the contribution to \( D(y, \hat{\mu}) \) 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:

```r
### generic (logistic regression) model
fit0 <- glm(Y ~ X, family=binomial)

### deviance residuals are the default
residual(fit0)

### extracting the pearson residuals
residual(fit0, type="pearson")
```
The Bayesian solution

- A GLM is specified by:

\[
Y_i \mid X_i \sim f_Y(y; \mu_i, \phi)
\]

\[
E[Y_i \mid X_i] = g^{-1}(X_i^T \beta) = \mu_i
\]

\[
V[Y_i \mid X_i] = V(\mu_i)a_i(\phi)
\]

- \(f_Y(\cdot)\) is a member of the exponential dispersion family
- \(\beta\) is a vector of regression coefficients
- \(\phi\) is the dispersion parameter

- \((\beta, \phi)\) 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 \((\beta, \phi)\) which is often factored into

\[
\pi(\beta, \phi) = \pi(\beta|\phi)\pi(\phi)
\]

• For \(\beta|\phi\), 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., \(\beta \sim \text{MVN}(\beta_0, \Sigma_\beta)\)
  * convenient choice given the computational methods described below

• Unfortunately, specifying a prior for \(\phi\) is less prescriptive
  * consider specific models in Parts V-VII of the notes
Given an independent sample $Y_1, \ldots, Y_n$, the likelihood is the product of $n$ terms:

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

Apply Bayes' Theorem to get the posterior:

$$\pi(\beta, \phi | y) \propto L(\beta, \phi | y) \pi(\beta, \phi)$$
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 \( \tilde{\theta} = (\tilde{\beta}, \tilde{\phi}) \) denote the posterior mode
Consider a Taylor series expansion of the log-posterior at \( \tilde{\theta} \):

\[
\log \pi(\theta|y) = \log \pi(\tilde{\theta}|y) \\
+ (\theta - \tilde{\theta}) \frac{\partial}{\partial \theta} \log \pi(\theta|y) \bigg|_{\theta=\tilde{\theta}} \\
+ \frac{1}{2} (\theta - \tilde{\theta})^T \left[ \frac{\partial^2}{\partial \theta \partial \theta} \log \pi(\theta|y) \right] \bigg|_{\theta=\tilde{\theta}} (\theta - \tilde{\theta}) \\
+ \ldots
\]

- Ignore the \( \log \pi(\tilde{\theta}|y) \) term because, as a function of \( \theta \), 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

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

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

$$\pi(\theta|y) \approx \text{Normal}(\tilde{\theta}, I(\tilde{\theta})^{-1})$$

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

* there are many approaches that one could take
* we’ll describe three
• First, we need to find the mode, \((\tilde{\beta}, \tilde{\phi})\)
  ★ the value that maximizes \(\pi(\beta, \phi | y)\)
  ★ given a non-informative prior:

\[
(\tilde{\beta}, \tilde{\phi}) \equiv (\hat{\beta}_{\text{MLE}}, \hat{\phi}_{\text{MLE}})
\]

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

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

\[
\mathcal{I}(\beta, \phi) = \begin{bmatrix}
\mathcal{I}_{\beta\beta} & 0 \\
0 & \mathcal{I}_{\phi\phi}
\end{bmatrix}
\]
• Exploit this and consider the approximation:

\[ \pi(\beta | y) \approx \text{Normal} \left( \tilde{\beta}, V_\beta(\tilde{\beta}, \tilde{\phi}) \right) \]

to the marginal posterior of \( \beta \)

\[ \ast \quad V_\beta(\tilde{\beta}, \tilde{\phi}) = I_{\beta\beta}^{-1} \text{ evaluated at the mode} \]

\[ \ast \quad \text{denote the approximation by } \tilde{\pi}(\beta; y) \]

• Also consider the approximation:

\[ \pi(\phi | y) \approx \text{Normal} \left( \tilde{\phi}, V_\phi(\tilde{\beta}, \tilde{\phi}) \right) \]

to the marginal posterior of \( \phi \)

\[ \ast \quad V_\phi(\tilde{\beta}, \tilde{\phi}) = I_{\phi\phi}^{-1} \text{ evaluated at the mode} \]

\[ \ast \quad \text{denote the approximation by } \tilde{\pi}(\phi | y) \]
Approach #1

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

\[
\beta|y \sim \text{Normal}(\tilde{\beta}, V_\beta(\tilde{\beta}, \tilde{\phi}))
\]

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

- The approach conditions on \( \tilde{\phi} \)
  - uncertainty in the true value of \( \phi \) 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 \( \phi \) using the \( \tilde{\pi}(\phi|y) \) Normal approximation
Approach #2

• We may not be willing to believe that the approximation is good enough to summarize features of $\pi(\beta; y)$
  * 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 $\tilde{\pi}(\beta|y)$ as a proposal distribution in a Metropolis-Hastings algorithm to sample from the exact posterior $\pi(\beta; y)$

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

  (1) generate a proposal $\beta^*$ from $\tilde{\pi}(\beta|y)$
    * straightforward since this is a multivariate Normal distribution
(2) evaluate the acceptance ratio

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

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

(3) generate a random \( U \sim \text{Uniform}(0, 1) \)

* reject the proposal if \( a_r < U \):

\[ \beta^{(r+1)} = \beta^{(r)} \]

* accept the proposal if \( a_r \geq U \):

\[ \beta^{(r+1)} = \beta^* \]
Approach #3

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

- To sample from the full exact posterior $\pi(\beta, \phi; y)$ 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 $\beta^{(r)}$ from $\pi(\beta|\phi^{(r-1)}; y)$ with $\tilde{\pi}(\beta|y)$ as a proposal
    2. sample $\phi^{(r)}$ from $\pi(\phi|\beta^{(r)}; y)$ with $\tilde{\pi}(\phi|y)$ as a proposal

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