Key topics in Soc 401/504: Advanced Social Statistics
GLMs, maximum likelihood, and quantities of interest

Last updated: 15 February 2018 Ian Lundberg (ilundberg at princeton dot edu)

Defining GLMs[1]

A **generalized linear model** (GLM) extends the linear model to various response types: binary, count, ordinal, duration, etc. The **data generating process** in a GLM involves three elements:

$$\overbrace{\vec{X}_i\vec{\beta} = \eta_i}^{\text{Linear predictor}} \qquad \overbrace{\eta_i = g(\mu_i)}^{\text{Link function } g} \qquad \overbrace{Y_i \sim f_Y(\mu_i, \gamma)}^{\text{Stochastic component}}$$

GLMs are defined for data generated from distributions f_Y in the **exponential family** (see Supplement A). We use $\vec{\theta}$ to denote the full set of parameters to be estimated, which include coefficients $\vec{\beta}$ and, if relevant, a parameter γ related to the variance. Table 1 provides examples of GLMs.

Maximum likelihood

Unlike OLS, there is no general analytic formula for the optimal parameter estimates $\hat{\beta}$. Instead, we choose the parameters under which the data we observe would be most likely: we search for the parameters that maximize the **likelihood**:[2]

$$\overbrace{L\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right)}^{\text{Likelihood}} = \overbrace{f_Y\left(\vec{y} \mid \mathbf{X}, \vec{\theta}\right)}^{\substack{\text{Probability density of} \\ \text{data given parameters}}}$$

We often assume **conditional independence**, thereby allowing us to factor the likelihood.

$$L\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) = f_Y\left(\vec{y} \mid \mathbf{X}, \vec{\theta}\right) = f_Y\left(y_1 \mid \vec{x}_1, \vec{\theta}\right) \times \cdots \times f_Y\left(y_n \mid \vec{x}_n, \vec{\theta}\right) = \prod_{i=1}^n f_Y\left(y_i \mid \vec{x}_i, \vec{\theta}\right)$$

Often, the likelihood factors. We can **drop terms** that do not involve the parameters $\vec{\theta}$ to produce a function proportional to the likelihood; the value of $\vec{\theta}$ that maximizes the likelihood remains unchanged.

$$L\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) = \prod_{i=1}^n f_Y\left(y_i \mid \vec{x}_i, \vec{\theta}\right) = \prod_{i=1}^n h_1\left(y_i, \vec{x}_i\right) h_2\left(y_i, \vec{x}_i, \vec{\theta}\right) \propto \prod_{i=1}^n h_2\left(y_i, \vec{x}_i, \vec{\theta}\right)$$

The log is a monotone function, so the argument $\vec{\theta}$ that maximizes L also maximizes the **log likelihood**.

$$\ell\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) = \log L\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) = \log\left(\prod_{i=1}^n f_Y\left(y_i \mid \vec{x}_i, \vec{\theta}\right)\right) = \sum_{i=1}^n \log f_Y\left(y_i \mid \vec{x}_i, \vec{\theta}\right) = \sum_{i=1}^n \log h_2\left(y_i, \vec{x}_i, \vec{\theta}\right) + c$$

where the constant $c = \sum_{i=1}^n h_1\left(y_i, \vec{x}_i\right)$ can be ignored.

[1] For alternative presentations of GLMs, we recommend Agresti (2015) and Powers and Xie (2008). A classic reference is McCullagh and Nelder (1989). Handout color scheme inspired by Efron and Hastie (2016).

[2] Maximum likelihood estimation has a close connection to Bayesian inference. See Supplement B.

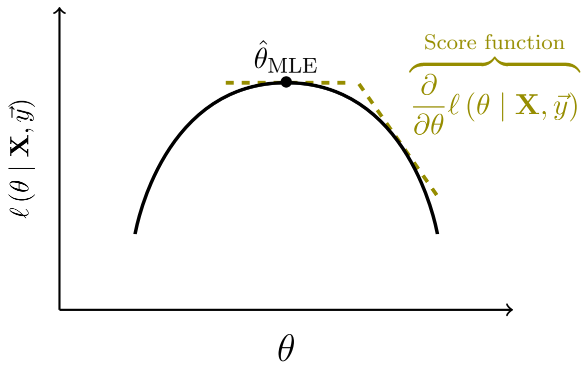
Climbing the mountain

The log likelihood is a mountain. Each point on the mountain is represented by coordinates that correspond to the parameter vector $\vec{\theta}$. We want to find the maximum likelihood estimate: the peak.

We will assume temporarily that the mountain has only one dimension: $\vec{\theta} = \theta$ has just one element.

The **first derivative** $\frac{\partial}{\partial\theta}\ell\left(\theta \mid \mathbf{X}, \vec{y}\right)$ captures the slope of the mountain. At the peak, the mountain is flat. We start by finding a candidate point θ^* at which the first derivative is 0.

$$\theta^* = \theta \text{ such that } \frac{\partial}{\partial\theta}\ell\left(\theta \mid \mathbf{X}, \vec{y}\right) = 0$$

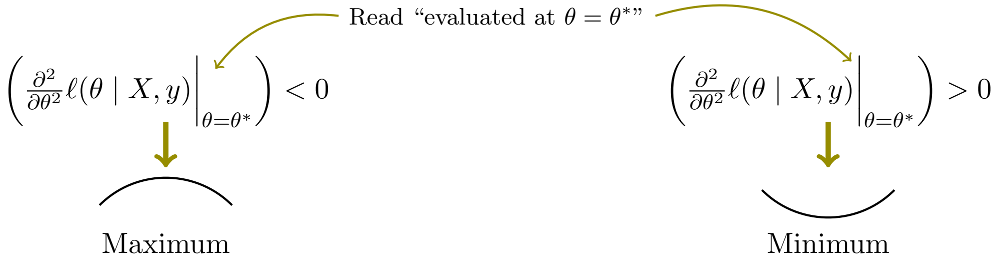

$\hat{\theta}_{\text{MLE}}$
Score function
$\overbrace{\frac{\partial}{\partial\theta}\ell\left(\theta \mid \mathbf{X}, \vec{y}\right)}$
$\ell\left(\theta \mid \mathbf{X}, \vec{y}\right)$
θ

If $\vec{\theta}$ has many elements, the first derivative is called the **gradient** (denoted ∇) and captures the slope along each coordinate. The gradient is also called the **score function**.

$$\nabla\ell\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) = \begin{bmatrix} \frac{\partial}{\partial\theta_1}\ell\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) \\ \vdots \\ \frac{\partial}{\partial\theta_p}\ell\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) \end{bmatrix}$$

In the multivariate case, a place where every element of the score vector is 0 is a candidate peak.

A flat place could be a peak or a valley. To see whether we have found a maximum, we take the **second derivative** and evaluate it at our candidate θ^*. It tells us the direction of the curvature.


Read “evaluated at $\theta = \theta^*$”

$$\left(\frac{\partial^2}{\partial\theta^2}\ell(\theta \mid X, y)\bigg|_{\theta=\theta^*}\right) < 0$$

$$\left(\frac{\partial^2}{\partial\theta^2}\ell(\theta \mid X, y)\bigg|_{\theta=\theta^*}\right) > 0$$

If the first derivative is 0 and the second derivative is negative, then θ^* is our **maximum likelihood estimate** $\hat{\theta}_{\text{MLE}}$.[3]

[3] Technical note: In GLMs, the objective function is convex so there is no risk of a local maximum; there is only one maximum. In more complex models you may worry whether your maximum is a global maximum.

When $\vec{\theta}$ has many dimensions, we get a matrix of second derivatives called the **Hessian**.

$$H = \nabla\nabla^T \ell\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) = \begin{bmatrix} \frac{\partial^2}{\partial\theta_1^2}\ell\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) & \frac{\partial}{\partial\theta_1}\frac{\partial}{\partial\theta_2}\ell\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) & \cdots & \frac{\partial}{\partial\theta_1}\frac{\partial}{\partial\theta_p}\ell\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) \\ \frac{\partial}{\partial\theta_1}\frac{\partial}{\partial\theta_2}\ell\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) & \frac{\partial^2}{\partial\theta_2^2}\ell\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) & \cdots & \frac{\partial}{\partial\theta_2}\frac{\partial}{\partial\theta_p}\ell\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial\theta_1}\frac{\partial}{\partial\theta_p}\ell\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) & \frac{\partial}{\partial\theta_2}\frac{\partial}{\partial\theta_p}\ell\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) & \cdots & \frac{\partial^2}{\partial\theta_p^2}\ell\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) \end{bmatrix}$$

Variance of $\hat{\theta}_{\textbf{MLE}}$

The variance of the MLE estimator relates to the amount of curvature at the peak: the Hessian.

Heuristically, a very negative Hessian at the MLE (the peak) suggests that the score changes quickly from positive to negative near the MLE; we have a lot of information about the MLE. For this reason, the negative Hessian is called the **Fisher information**.

The **variance of the MLE estimate** is the inverse of the Fisher information:

$$\mathrm{V}\left(\hat{\theta}_{\text{MLE}}\right) = \left(\mathcal{I}_n(\theta)\right)^{-1}\Big|_{\theta=\theta_{\text{MLE}}}$$

In practice, we don't know the true θ, so we estimate the variance by evaluating the inverse Fisher information at our MLE estimate. This is sometimes called the observed Fisher information.

$$\hat{\mathrm{V}}\left(\hat{\theta}_{\text{MLE}}\right) = \left(\mathcal{I}_n(\theta)\right)^{-1}\Big|_{\theta=\hat{\theta}_{\text{MLE}}}$$

Remember: Sharper peak → more negative Hessian → more positive Fisher information → lower variance estimate

Properties of the MLE[4]

1. **Invariance**: If $\hat{\vec{\theta}}_{\text{MLE}}$ is the MLE for $\vec{\theta}$, then for any function $g(\vec{\theta})$ the MLE of $g(\vec{\theta})$ is $g(\hat{\vec{\theta}}_{\text{MLE}})$.
2. **Consistency:** The MLE is sometimes biased $\left(\mathbf{E}\left[\hat{\vec{\theta}}_{\text{MLE}}\right] \neq \vec{\theta}\right)$ but is always consistent: $\hat{\vec{\theta}}_{\text{MLE}} \xrightarrow{P} \vec{\theta}$.
3. **Asymptotic normality**: As $n \to \infty$, the sampling distribution of $\hat{\theta}_{\text{MLE}}$ converges to Normal.
$$\sqrt{n}\left(\hat{\vec{\theta}} - \vec{\theta}\right) \xrightarrow{D} \text{Normal}\left(0, [\mathcal{I}_1(\theta)]^{-1}\right)$$
4. **Efficiency:** The Cramér-Rao lower bound states that any unbiased estimator for $\vec{\theta}$ has variance at least as great as $(\mathcal{I}_n(\theta))^{-1}$, which is $\mathrm{V}\left(\hat{\vec{\theta}}_{\text{MLE}}\right)$. This means:
   - If the MLE is unbiased, it is the most efficient (lowest variance) among unbiased estimators. However, the MLE is often biased.
   - The MLE is asymptotically efficient: even biased MLEs asymptotically achieve the CRLB.

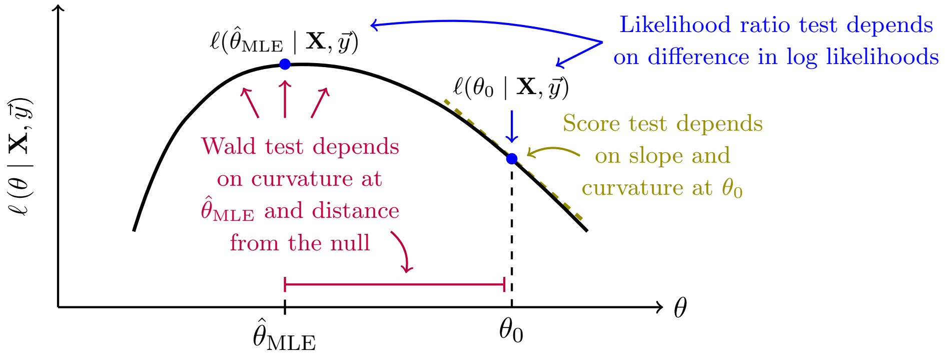
[4] For more details on inference on MLE theory, we recommend Casella and Berger (2002). Invariance is Thm. 7.2.10, consistency is Thm. 10.1.6, normality is Thm. 10.1.12, the CRLB is Thm. 7.3.9, and asymptotic efficiency is defined in Def. 10.1.11 and related to MLE in Thm. 10.1.12.

Hypothesis testing[5]

Suppose we want to test the null hypothesis that a subset of the coefficients are zero.

$$\vec\beta = \begin{bmatrix}\vec\beta_A\\ \vec\beta_B\end{bmatrix}, \qquad H_0 : \vec\beta = \vec\beta_0 \equiv \begin{bmatrix}\vec 0\\ \vec\beta_B\end{bmatrix}$$

There are three main methods for conducting this test. We will denote the number of elements in $\vec\theta_A$ by k, often using the $k = 1$ case to build intuition.


$\ell(\theta \mid \mathbf{X}, \vec y)$
$\ell(\hat\theta_{\text{MLE}} \mid \mathbf{X}, \vec y)$
Likelihood ratio test depends
on difference in log likelihoods
$\ell(\theta_0 \mid \mathbf{X}, \vec y)$
Score test depends
on slope and
curvature at θ_0
Wald test depends
on curvature at
$\hat\theta_{\text{MLE}}$ and distance
from the null
$\hat\theta_{\text{MLE}}$
θ_0
θ

The **Wald test** (Wald, 1943) relies on the asymptotic normality of $\hat\beta_{\text{MLE}}$:

For $k = 1$

$$\frac{\hat\beta_{A,\text{MLE}}}{\widehat{\text{SE}}(\hat\beta_{A,\text{MLE}})} \xrightarrow{D} N(0,1)$$

Standardized deviation from the null
Squared would be $\sim \chi^2_1$

For any k

$$\left[\hat{\vec\beta}_{A,\text{MLE}}\right]^T \left[\hat{\text{V}}\left(\hat{\vec\beta}_{A,\text{MLE}}\right)\right]^{-1} \hat{\vec\beta}_{A,\text{MLE}} \xrightarrow{D} \chi^2_k$$

Intuition in special case with independent coefficient estimates:
Sum of squared standardized deviations from the null

The **likelihood ratio test** (Neyman and Pearson, 1933; Wilks, 1938) compares nested models: a restricted model estimating $\vec\beta_B$ and assuming $\vec\beta_A = 0$ with likelihood L^*_R, and an unrestricted model with likelihood L^*.

$$-2\log\left(\frac{L^*_R}{L^*}\right) \xrightarrow{D} \chi^2_k$$

Intuition: We have more evidence in favor of the unrestricted model if the likelihood (L^*_R) under the restricted model is much smaller than the likelihood under the unrestricted model L^*. We need more evidence if k is larger.

The **score test** (Rao, 1948) is based on the score function and the Fisher information evaluated at $\vec\beta_0$.

For $k = 1$

$$\left.\frac{\left(\frac{\partial}{\partial\beta_A}\ell(\vec\beta|y)\right)^2}{\frac{\partial^2}{\partial{\beta_A}^2}\ell(\vec\beta|y)}\right|_{\beta_A=0} \xrightarrow{D} \chi^2_1$$

$\frac{\text{Squared slope}}{\text{Rate of change of slope}}$

Evaluated at
the null hypothesis

Intuition: A steep likelihood with
minimal curvature is far from the maximum.

For any k

$$\left.\left[s\left(\tilde{\vec\beta}\right)\right]^T \left[\mathcal{I}\left(\tilde{\vec\beta}\right)\right]^{-1} \left[s\left(\tilde{\vec\beta}\right)\right]\right|_{\tilde{\vec\beta}=\vec\beta_0} \xrightarrow{D} \chi^2_k$$

Matrix of rate of change
of slopes in all directions

Score vector of slopes
in all directions

Evaluated at
the null hypothesis

[5] These tests can be generalized to the null hypothesis $H_0 : h(\vec\theta\,) = \vec c$, where h is a function that maps the vector $\vec\theta \in \mathbb{R}^p$ to a vector of constraints $c \in \mathbb{R}^k$. For an overview in this more general framework, see Rao (2005). For a textbook treatment including a discussion of inverting the tests to produce confidence intervals, see Agresti (2015) Section 4.3.

Reporting results

The parameters $\vec{\theta}$ are rarely of interest; regression coefficients are always difficult to interpret. Instead, you should always report **quantities of interest** that clearly summarize your finding.[6] These might include predicted probabilities, first differences, the average treatment effect, etc.

The entire process can be summarized in a few simple steps:

1. **Fit a model**. Assume a model, write the log likelihood, and estimate $\hat{\theta}_{\text{MLE}}$ and the Hessian.
2. **Simulate estimation uncertainty**. We are uncertain about the true MLE $\vec{\theta}$. We want to incorporate this uncertainty in our estimate. We approximate the sampling distribution of $\hat{\vec{\theta}}_{\text{MLE}}$ with thousands of draws from a multivariate normal distribution.[7][8]

$$\tilde{\theta} \sim N\left(\hat{\vec{\theta}}_{\text{MLE}}, \quad \underbrace{\hat{\text{V}}\left(\hat{\vec{\theta}}_{\text{MLE}}\right)}_{\text{Variance-covariance matrix}}\right)$$

3. **Calculate the linear predictor** for each observation for each draw. Depending on your quantity of interest, you may want to change the $\vec{X}_i$ values of some observations so that the observations you compare are informative for your theory.

$$\left\{\tilde{\eta}_i = \vec{X}_i\tilde{\vec{\beta}}\right\}_{i=1}^{n}$$

4. Transform by the **inverse link function**. $\left\{\tilde{\mu}_i = g^{-1}\left(\tilde{\eta}_i\right)\right\}_{i=1}^{n}$
   - The link function is $g(\mu_i) = \vec{X}_i\vec{\beta}$. The inverse link function does the reverse: $\mu_i = g^{-1}\left(\vec{X}_i\vec{\beta}\right)$.
5. **Simulate fundamental uncertainty**. Even if we knew the true $\vec{\theta}$, a component would remain that is fundamentally stochastic. Draw from the distribution of Y to simulate this.

$$\left\{\tilde{Y}_i \sim f_Y\left(\tilde{\mu}_i, \tilde{\gamma}\right)\right\}_{i=1}^{n}$$

6. **Calculate your quantity of interest**. This is the thing you want to report to your readers.

$$\tilde{\tau} = \overbrace{h\left(\tilde{Y}_1, \ldots, \tilde{Y}_n\right)}^{\substack{\text{Any quantity you} \\ \text{want to report}}}$$

7. **Repeat** steps 2-6 thousands of times.
8. **Summarize** the resulting distribution of the $\tilde{\tau}$ samples in a clear, informative graph.

[6] This strategy was originally advocated by King et al. (2000). Some functions to fit common GLMs are available in the `Zelig` package in `R` (Imai et al., 2008; Choirat et al., 2017).

[7] It is important that all parameters are simulated together $\left(\tilde{\vec{\theta}} \text{ includes both } \tilde{\vec{\beta}} \text{ and } \tilde{\gamma}\right)$.

[8] The log likelihood is asymptotically normal. The proof invokes the Central Limit Theorem and the fact that the log likelihood is a sum of independent quantities. Draws from the likelihood are the likelihoodist analog of draws from the posterior distribution in Bayesian inference. For more on the likelihood theory of inference, see King (1998).

		Unknown	Linear			
Model	Response type	parameters	predictor	=	Link function g	Stochastic component
OLS	Normal	$\vec{\theta} = \left\{\vec{\beta}, \gamma\right\}$	$\vec{X}_i\vec{\beta} = \eta_i$	=	$g(\mu_i) = \mu_i$	$Y_i \sim \text{Normal}(\mu_i, \sigma^2 = \gamma)$
Logit	Binary	$\vec{\theta} = \vec{\beta}$	$\vec{X}_i\vec{\beta} = \eta_i$	=	$g(\mu_i) = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$	$f_Y(\mu_i) = \text{Bernoulli}(\pi_i = \mu_i)$
Probit	Binary	$\vec{\theta} = \vec{\beta}$	$\vec{X}_i\vec{\beta} = \eta_i$	=	$g^*(\mu_i) = \text{probit}(\mu_i) = \Phi^{-1}(\mu_i)$	$Y_i \sim \text{Bernoulli}(\pi_i = \mu_i)$
Complementary log-log	Binary	$\vec{\theta} = \vec{\beta}$	$\vec{X}_i\vec{\beta} = \eta_i$	=	$g^*(\mu_i) = \log(-\log(1-\mu_i))$	$Y_i \sim \text{Bernoulli}(\pi_i = \mu_i)$
Poisson	Count	$\vec{\theta} = \vec{\beta}$	$\vec{X}_i\vec{\beta} = \eta_i$	=	$g(\mu_i) = \log(\mu_i)$	$Y_i \sim \text{Poisson}(\lambda_i = \mu_i)$
Neg. Binomial	Count	$\vec{\theta} = \left\{\vec{\beta}, \gamma\right\}$	$\vec{X}_i\vec{\beta} = \eta_i$	=	$g^*(\mu_i) = \log(\mu_i)$	$Y_i \sim \text{NegBin}\left(\gamma, \pi_i = \frac{\mu_i}{\gamma+\mu_i}\right)$
Exponential	Duration	$\vec{\theta} = \vec{\beta}$	$\vec{X}_i\vec{\beta} = \eta_i$	=	$g^*(\mu_i) = \log\left(\frac{1}{\mu_i}\right) = \log(\lambda_i)$ $g(\mu_i) = \frac{1}{\mu_i} = \lambda_i$	$Y_i \sim \text{Exponential}\left(\lambda_i = \frac{1}{\mu_i}\right)$
Gamma	Duration	$\vec{\theta} = \left\{\vec{\beta}, \gamma\right\}$	$\vec{X}_i\vec{\beta} = \eta_i$	=	$g^*(\mu_i) = \log\left(\frac{1}{\mu_i}\right) = \log(\lambda_i)$ $g(\mu_i) = \frac{1}{\mu_i} = \lambda_i$	$Y_i \sim \text{Gamma}\left(\alpha = \gamma, \lambda_i = \frac{\gamma}{\mu_i}\right)$
Multinomial logit	Categorical	$\boldsymbol{\theta} = \boldsymbol{\beta} = \left[\vec{\beta}_2 \cdots \vec{\beta}_k\right]$	$\vec{X}_i\boldsymbol{\beta} = \vec{\eta}_i$	=	$g(\vec{\mu}_i) = \log\left(\frac{\vec{\mu}_i}{1-\vec{\mu}_i^T\vec{1}}\right)$	$Y_i \sim \text{Multinomial}\left(\vec{\pi}_i = \begin{bmatrix} 1-\vec{\mu}_i^T\vec{1} \\ \vec{\mu}_i \end{bmatrix}\right)$

Table 1: Common generalized linear models (GLMs). List is not exhaustive. Alternative parameterizations and link functions exist; these were chosen to maximize consistency of notation within the table. Link functions denoted g^* are not the canonical link for the given response type (see Supplement A). There are various reasons to choose a non-canonical link in some cases; one such reason is that the inverse of the canonical link $g^{-1}(\eta_i) = \mu_i$ does not always map all possible values of η_i (the real line) into the support of μ_i.

Supplement A. Exponential family

The distributions used in generalized linear models all come from the **exponential family**. The probability density function of each distribution can be written in the following form:

$$f(y \mid \vec{\theta}) = \overbrace{h(y)}^{\substack{\text{Normalizing}\\\text{constant}}} \exp\left(\overbrace{\vec{\theta}^T}^{\substack{\text{Natural}\\\text{parameter}}} \overbrace{\vec{\phi}(y)}^{\substack{\text{Sufficient}\\\text{statistics}}} - \overbrace{A\left(\vec{\theta}\right)}^{\substack{\text{Cumulant}\\\text{function}}} \right)$$

Ex. For the Bernoulli,

$$\begin{aligned} P(y \mid p) &= \pi^y (1-\pi)^{1-y} \\ &= \exp\left(\log\left(\pi^y (1-\pi)^{1-y}\right) \right) \\ &= \exp\Big(y\log(\pi) + (1-y)\log(1-\pi) \Big) \\ &= \exp\left(\underbrace{y}_{=\phi(y)} \underbrace{\log\left(\frac{\pi}{1-\pi}\right)}_{=\theta} + \underbrace{\log(1-\pi)}_{=\log(1-\text{logit}^{-1}[\theta])=-A(\theta)} \right) \end{aligned}$$

The **canonical link function** is the one that transforms the mean of the distribution to the natural parameter θ. In this case, the canonical link function is $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \text{logit}(\pi)$. This is why the logit is popular!

The exponential family is nice because:

1. Sufficient statistics $\vec{\phi}(y)$ are finite in dimension even in an infinite sample. To calculate the MLE, you only need the sufficient statistics and not actually all of the data. Information is compressed.
2. Conjugate priors exist for Bayesian inference.
3. The exponential family has maximum entropy (most diffuse) among distributions subject to some moment constraints.
4. The mean has a known formula: $\mathbf{E}(Y) = \nabla A(\theta)$. For the Bernoulli example,

$$\begin{aligned} \mathbf{E}(Y) &= \nabla A(\theta) \\ &= \frac{\partial}{\partial\theta}\Big[-\log(1-\text{logit}^{-1}[\theta]) \Big] \\ &= \frac{\partial}{\partial\theta}\left[-\log\left(\frac{1}{1+e^\theta}\right) \right] \\ &= -\left(\left(1+e^\theta\right) \left(-e^\theta \left[1+e^\theta\right]^{-2}\right) \right) \\ &= \left(\frac{e^\theta}{1+e^\theta}\right) \\ &= \pi \end{aligned}$$

5. The variance has a known formula: $\mathrm{V}(Y) = \nabla\nabla^T A(\theta)$. For the Bernoulli example,

$$\begin{aligned}
\mathrm{V}(Y) &= \nabla\nabla^T A(\theta) \\
&= \frac{\partial^2}{\partial^2\theta}\left[-\log(1-\text{logit}^{-1}[\theta])\right] \\
&= \frac{\partial}{\partial\theta}\left(\frac{e^\theta}{1+e^\theta}\right) \\
&= \frac{e^\theta}{(1+e^\theta)^2} \\
&= \left(\frac{e^\theta}{1+e^\theta}\right)\left(\frac{1}{1+e^\theta}\right) \\
&= \pi(1-\pi)
\end{aligned}$$

For more complex distributions, having a known function for the mean and variance is nice! For more on the exponential family, we recommend Murphy (2012) Ch. 9. A Bayesian technique called **variational inference** approximates distributions with the closest possible match among the exponential family (see Murphy 2012 Ch. 21). There are also many fun connections between the distributions within the exponential family; for this we recommend Blitzstein and Hwang (2014).

Supplement B. Connection to Bayesian inference

We will not cover Bayesian inference in this class. In this class, the unknown parameters (β) are treated as fixed constants. In Bayesian inference, the unknown parameters are treated as random variables.

$$P(\vec{\theta} \mid \vec{y}) = \frac{\overbrace{P(\vec{y} \mid \vec{\theta})}^{\text{Likelihood}}\,\overbrace{P(\vec{\theta})}^{\text{Prior}}}{\underbrace{P(\vec{y})}_{\text{Normalizing constant}}}$$

Because the normalizing constant does not involve the unknown parameters θ, we can ignore it. The two key components in Bayesian inference are the **prior distribution** on the unknown parameters and the **likelihood**. What you learn about likelihood in this course will prepare you to jump into Bayesian inference in the future.

As an example of Bayesian inference, consider Bernoulli draws for which we assume a Beta(α,β) prior on the probability of success π. This corresponds to our beliefs about the distribution of π before seeing the data.

$$\pi \sim \text{Beta}(\alpha,\beta)$$
$$Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi_i)$$

Bayesian inference produces a **posterior distribution** that summarizes our updated beliefs given the data. In the Bernoulli case, the Beta distribution is a **conjugate prior** because the posterior will also follow a Bernoulli distribution.

$$\begin{aligned}\underbrace{f(\pi \mid \vec{y})}_{\text{Posterior}} &\propto \underbrace{f(\pi)}_{\text{Prior}}\,\underbrace{\mathrm{P}(\vec{y} \mid \pi)}_{\text{Likelihood}}\\ &\propto \pi^{\alpha}(1-\pi)^{\beta}\prod_{i=1}^{n}\pi^{y_i}(1-\pi)^{1-y_i}\\ &= \pi^{\alpha+\sum_{i=1}^{n} y_i}(1-\pi)^{\beta+n-\sum_{i=1}^{n} y_i}\\ &\propto \text{Beta}\left(\alpha+\sum_{i=1}^{n} y_i, \beta+n-\sum_{i=1}^{n} y_i\right)\end{aligned}$$

We often summarize the posterior distribution by the **posterior mean**.

$$\hat{\pi}_{\text{Posterior Mean}} = \mathbf{E}\left[\pi \mid \vec{y}\right] = \mathbf{E}\left[\text{Beta}\left(\alpha+\sum_{i=1}^{n} y_i, \beta+n-\sum_{i=1}^{n} y_i\right)\right] = \frac{\alpha+\sum_{i=1}^{n} y_i}{\alpha+\beta+n}$$

Rearranging terms, we can express the posterior mean as a weighted average of the prior mean and the MLE estimate. As the $n \to \infty$, the data overwhelm the prior and the posterior mean will converge to the true value. The posterior mean is therefore biased but consistent.

$$\hat{\pi}_{\text{Posterior Mean}} = \underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{Prior mean}}\underbrace{\left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)}_{\text{Weight on prior}} + \underbrace{\frac{\sum_{i=1}^{n} y_i}{n}}_{\hat{\pi}_{\text{MLE}}}\underbrace{\left(\frac{n}{\alpha+\beta+n}\right)}_{\text{Weight on likelihood}}$$

We can often **re-interpret frequentist results** as Bayesian results with a particular prior. In the case of the Bernoulli, the posterior mean converges to the frequentist $\hat{\pi}_{\text{MLE}}$ in the limit as α and β go to 0.

$$\lim_{\{\alpha,\beta\}\to 0^{+}} \hat{\pi}_{\text{Posterior Mean}} = \lim_{\{\alpha,\beta\}\to 0^{+}} \frac{\alpha+\sum_{i=1}^{n} y_i}{\alpha+\beta+n} = \frac{\sum_{i=1}^{n} y_i}{n} = \hat{\pi}_{\text{MLE}}$$

What would that prior look like? To build intuition for the above, we can interpret a Beta(α, β) prior in terms of **pseudocounts**. The posterior mean is the same as what a frequentist would conclude if α successes and β failures were added to the observed data. As the pseudocounts go to 0, the prior becomes more and more diffuse (see Fig. 1). The limit as the pseudocounts go to 0 corresponds to high prior probabilities on parameters near 0 and 1 and negligible prior probability in the middle of the region; this is the prior effectively assumed by the MLE procedure.

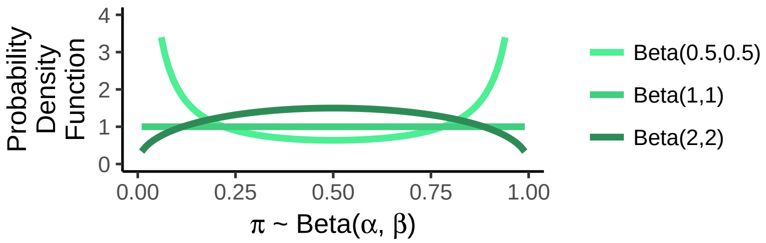


Figure 1: Beta distribution probability density function. Lighter green indicates a more diffuse prior that can be interpreted as having fewer pseudocounts before seeing the data.

Because we often have some knowledge of π before seeing the data, it is often helpful to assume a weakly informative prior, which will bias results toward the prior but improve efficiency. Researchers often mistakenly state that they have chosen a "non-informative" or "flat" prior. In our example, for instance, you might place a Beta(1,1) or uniform prior on π. This would be a reasonable choice, but it would not be flat if you reparameterized to focus on $\gamma = \pi^2$. Under this reparameterization, a uniform prior on π would correspond to much greater prior probabilities on γ near 0 than γ near 1. Because priors can be sensitive to parameterization, one should be cautious in claiming that any prior is flat.

To avoid this problem, a common diffuse prior distribution is the **Jeffreys prior**

$$f(\theta) \propto \sqrt{\underbrace{\left|\mathcal{I}\left(\vec{\theta}\right)\right|}_{\substack{\text{Determinant of}\\ \text{Fisher information}}}}$$

The Jeffreys prior is invariant to reparameterizations of θ because the parameterization of θ is taken into account in the Fisher information. In the Bernoulli example, the Jeffreys prior is a Beta $\left(\frac{1}{2}, \frac{1}{2}\right)$. The Jeffreys prior is still informative, however: the posterior mean in this case is biased toward $\frac{1}{2}$.

$$\hat{\pi}_{\text{Posterior Mean}} = \mathbf{E}\left[\text{Beta}\left(\frac{1}{2} + \sum_{i=1}^{n} y_i, \frac{1}{2} + n - \sum_{i=1}^{n} y_i\right)\right] = \underbrace{\left(\frac{1}{2}\right)}_{\substack{\text{Prior}\\ \text{mean}}}\left(\frac{1}{1+n}\right) + \underbrace{\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)}_{\hat{\pi}_{\text{MLE}}}\left(\frac{n}{1+n}\right)$$

Rather than arguing about the choice of a flat prior, it may be wiser to choose a "weakly informative" prior which gains efficiency by placing lower probability on highly improbable regions of the parameter space (Gelman et al., 2008). We chose a conjugate prior in this example so that the math would be nice, but one can use Markov Chain Monte Carlo (MCMC) methods to sample from the posterior distribution even when it does not have an analytical solution (see Brooks et al. 2011 and Robert and Casella 2010 for introductions). MCMC represents one of the foremost computational breakthroughs in statistics in recent decades, and new methods for sampling from the posterior are continually being developed (i.e. Carpenter et al. 2017). Bayesian methods are likely to become even more accessible to applied researchers in the future as computation improves. For those interested in Bayesian inference, we recommend Hoff (2009) and Gelman et al. (2014).

References

Agresti, A. 2015. *Foundations of Linear and Generalized Linear Models.* John Wiley & Sons.

Blitzstein, J. K. and J. Hwang 2014. *Introduction to Probability.* CRC Press.

Brooks, S., A. Gelman, G. Jones, and X.-L. Meng, eds. 2011. *Handbook of markov chain monte carlo.* CRC press.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell 2017. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76.

Casella, G. and R. L. Berger 2002. *Statistical Inference*, 2 edition. Duxbury Pacific Grove, CA.

Choirat, C., J. Honaker, K. Imai, G. King, and O. Lau 2017. *Zelig: Everyone's Statistical Software.* Version 5.1.5.

Efron, B. and T. Hastie 2016. *Computer Age Statistical Inference.* Cambridge University Press.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin 2014. *Bayesian Data Analysis*, volume 2. CRC press Boca Raton, FL.

Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su 2008. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.

Hoff, P. D. 2009. *A First Course in Bayesian Statistical Methods.* Springer Science & Business Media.

Imai, K., G. King, and O. Lau 2008. Toward a common framework for statistical analysis and development. *Journal of Computational Graphics and Statistics*, 17(4):892–913.

King, G. 1998. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* University of Michigan Press.

King, G., M. Tomz, and J. Wittenberg 2000. Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science*, Pp. 347–361.

McCullagh, P. and J. Nelder 1989. Generalized linear models. *Chapman and Hall: New York.*

Murphy, K. P. 2012. *Machine Learning: A Probabilistic Perspective.* MIT press.

Neyman, J. and E. S. Pearson 1933. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A*, 231(694-706):289–337.

Powers, D. and Y. Xie 2008. *Statistical Methods for Categorical Data Analysis.* Emerald Group Publishing.

Rao, C. 2005. Score test: historical review and recent developments. In *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*, Pp. 3–20. Springer.

Rao, C. R. 1948. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 44, Pp. 50–57. Cambridge University Press.

Robert, C. P. and G. Casella 2010. *Introducing Monte Carlo Methods with R*, volume 18. Springer.

Wald, A. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482.

Wilks, S. S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.