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

Last updated: 1 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.

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

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

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

[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.

We can **drop additive terms** that do not involve the parameters; the value of $\vec{\theta}$ that maximizes the likelihood remains unchanged.

$$\ell\left(\vec{\theta} \mid \mathbf{X}, \vec{y}\right) = \sum_{i=1}^{n} \log f_Y\left(y_i \mid \vec{x}_i, \vec{\theta}\right) = \sum_{i=1}^{n} \left(h_1\left(y_i, \vec{x}_i\right) + h_2\left(y_i, \vec{x}_i, \vec{\theta}\right)\right) \doteq \sum_{i=1}^{n} h_2\left(y_i, \vec{x}_i, \vec{\theta}\right)$$

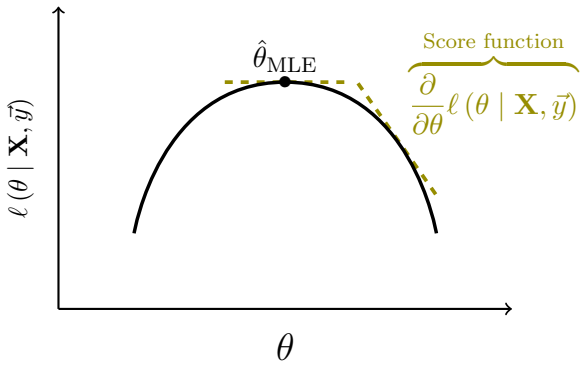
Climbing the mountain

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.

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

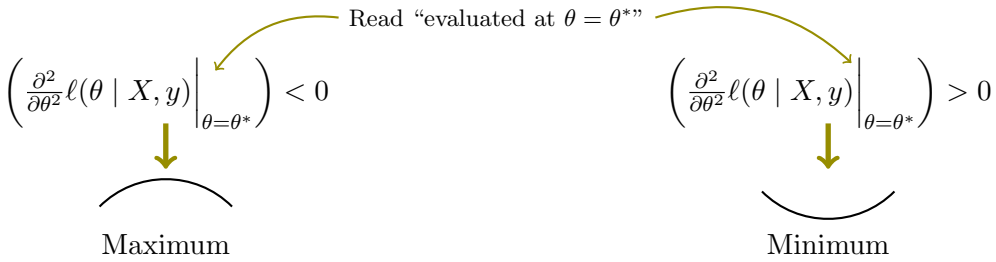


If $\vec{\theta}$ has many elements, the first derivative is called the **gradient** (denoted ∇) and captures the slope along each coordinate. The gradient is also called the **score function**.

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

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.



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

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 $\hat{\theta}_{\text{MLE}}$

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. This suggests we are quite sure of the MLE estimate.

The negative Hessian is called the **Fisher information**. When the Hessian is very negative, this indicates the score changes quickly from positive to negative near the MLE, so we are quite certain of our estimate. Correspondingly, a large positive Fisher information implies more information and thus more certainty about our estimate.

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

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

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.

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

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

For more details on inference on MLE theory, we recommend Casella and Berger (2002).

[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.

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.[4] 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.[5][6]

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

3. **Calculate the linear predictor** for each observation for each draw. Depending on you 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.

[4] 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).

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

[6] 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).

Model	Response type	Unknown parameters	Linear 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.

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:

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

Ex. For the Bernoulli,

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

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,

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

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

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

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).

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. In this case,

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

Because the normalizing constant does not involve the unknown parameters θ, we can ignore it. Maximizing the likelihood thus agrees with finding the coefficients that maximize the posterior distribution under a flat prior (a prior $P(\theta)$ that takes the same value at all values θ). A "flat prior" is a bit of a misnomer, though, as it is not generally flat under transformations of the parameters. Further, we usually have some prior knowledge that can improve efficiency and estimation by avoiding parameter values that are highly implausible. For those interested in Bayesian inference, we recommend 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.

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.

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.

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