The Exponential Family and Generalized Linear Models

Rebecca C. Steorts (adaptation from notes of Yue Jiang)

March 3, 2025

Feedback on Homework 4

Issues with 1c (interpretation of λ).

Assuming poisson likelihood for the response, λ is the expected number of fish caught during one-week. In our data, the sample estimate of λ would be the sample mean, which is 21.5.

The parameter λ represents the mean rate for the number of fish caught per week, which is 21.5.

Feedback on Homework 4

In Question 3c, some students used deviance residuals when calculating the dispersion parameter.

In Question 5d, some students interpreted the interaction for one of the two parties only.

The Exponential Family

In the next few lectures, we will tie together exponential families and generalized linear models.

1. Each generalized linear model can be expressed as an exponential family.
2. Exponential families have nice properties that allow us to calculate certain quantities, such as the MLE, quite easily.
3. We will not explore all the properties in this class as the mathematics goes beyond the scope, however, it's important to know what's going behind the hood of the R engine each time you call the glm function.
4. We could derive (very tediously) general expressions for GLM's using the exponential family form that involve the regression parameter estimates (and other quantities that we calculate), however, we won't do this in this class as the mathematics is tedious and ugly. If you'd like to see this or know more about it, please let me know!

The Exponential Family

The exponential family of probability distributions are those that can be expressed in a specific form.

Suppose X is a random variable with a distribution that depends on parameter(s) θ.

A random variable X belongs to the exponential family if its density (or mass) function can be written as:

$$f(x|\theta) = h(x) \exp\Big(\eta(\theta)^T T(x) - \psi(\theta)\Big).$$

The Exponential Family

$$f(x|\theta) = h(x)\exp\Big(\eta(\theta)^T T(x) - \psi(\theta)\Big).$$

Note that the term $\eta(\theta)^T T(x)$ represents $\sum_{i=1}^{k} \eta_i(\theta) T_i(x)$.

Here, $\eta_i(\theta)$ and $\psi(\theta)$ are real-valued functions of the parameters (θ). Also, each of $T_i(x)$ and $h(x)$ are real-valued functions of the data.

In the case of a single parameter (θ), we have a member of a one-parameter exponential family distribution as follows:

$$f(x|\theta) = h(x)\exp\Big(\eta(\theta) T(x) - \psi(\theta)\Big).$$

One-Parameter Exponential Family

For simplicity, we will first consider the one-parameter exponential family of distributions as follows:

$$f(x|\theta) = h(x)\exp\Big(\eta(\theta)T(x) - \psi(\theta)\Big).$$

The Binomial Distribution

Suppose $X \sim \text{Bin}(n, p)$, where n is known and $0 < p < 1$.

Questions:

- What is the probability mass function $f(x|p)$ corresponding to X?
- Is X a member of the one-parameter exponential family?
- If yes, identify the components $\eta(p)$, $\psi(p)$, $T(x)$, and $h(x)$. If not, explain why not.

The Binomial Distribution

The probability mass function can be written as

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} \, 1\{x \in \{0, 1, \ldots, n\}\} \tag{1}$$

$$= \binom{n}{x} \left(\frac{p}{1-p}\right)^x (1-p)^n \, 1\{x \in \{0, 1, \ldots, n\}\} \tag{2}$$

$$= \binom{n}{x} \exp\left(x \log \frac{p}{1-p} + n \log(1-p)\right) 1\{x \in \{0, 1, \ldots, n\}\}. \tag{3}$$

Is X a member of the one-parameter exponential family? If yes, identify the components $\eta(p), \psi(p), T(x), h(x)$. If not, explain.

The Binomial Distribution

Recall that

$$f(x|p) = \binom{n}{x} \exp\left(x \log \frac{p}{1-p} + n \log(1-p)\right) 1\{x \in \{0, 1, \ldots, n\}\}.$$

Yes, X is a member of the exponential family with the following parameters:

$$\eta(p) = \log \frac{p}{1-p},$$

$$T(x) = x,$$

$$\psi(p) = -n \log(1-p),$$

$$h(x) = \binom{n}{x} 1\{x \in \{0, 1, \ldots, n\}\}.$$

The Binomial Distribution

Suppose $X \sim \text{Bin}(n, p)$, where n is known and $0 < p < 1$.
The probability mass function is given by

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} \, 1\{x \in \{0, 1, \ldots, n\}\}.$$

The sufficient statistic is $T(x) = x$. A sufficient statistic is one that provides all the information about θ that the entire sample could have provided.

What does this mean in terms of the binomial distribution?

More on Sufficiency

Sufficient statistics must be functions of the data (and cannot depend on the parameter value).

Sufficiency suggests that because $T(x)$ contains all the information about θ, then it suffices to utilize $T(x)$ for inference. Thus, we don't need the individual data points themselves (just the sufficient statistic).

More on Sufficiency

Intuitively, for $X \sim \text{Bin}(n, p)$, knowing the number of successes x is all that is needed for inference on p, rather than knowing which specific observations were successes or failures.

The Normal Distribution

Suppose $X \sim N(\mu, \sigma^2)$ with parameter $\theta = (\mu, \sigma)$.

Questions:

- What is the dimension of θ in this case?
- What is the probability density function $f(x|\theta)$? Again, pay attention to the support of X.
- Is X a member of the exponential family? If yes, identify the components $\eta_i(\theta)$, $\psi(\theta)$, $T_i(x)$, and $h(x)$. If not, explain.

The Normal Distribution (Density)

The density is given by:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} 1\{x \in \mathbb{R}\}.$$

The Normal Distribution (Exponential Family Form)

Rewriting the density:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma \right) \right\} 1\{x \in \mathbb{R}\}.$$

Thus, we identify:

$$\eta_1(\theta) = -\frac{1}{2\sigma^2}, \qquad T_1(x) = x^2,$$

$$\eta_2(\theta) = \frac{\mu}{\sigma^2}, \qquad T_2(x) = x,$$

$$\psi(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma,$$

$$h(x) = \frac{1}{\sqrt{2\pi}} 1\{x \in \mathbb{R}\}.$$

i.i.d. Sampling from the Exponential Family

Suppose we have n i.i.d. samples $x_1, x_2, \ldots, x_n$ from an exponential family distribution.

The joint density is

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} h(x_i) \exp\Big(\eta(\theta)^T T(x_i) - \psi(\theta)\Big).$$

This can be rewritten as:

$$f(x_1, \ldots, x_n|\theta) = \Big(\prod_{i=1}^{n} h(x_i)\Big) \exp\Big(\eta(\theta)^T \sum_{i=1}^{n} T(x_i) - n\psi(\theta)\Big).$$

Sufficient Statistics for the Normal Distribution

For an i.i.d. sample from $N(\mu, \sigma^2)$, the statistics

$$\sum_{i=1}^{n} x_i^2 \quad \text{and} \quad \sum_{i=1}^{n} x_i$$

are sufficient for the parameters (μ, σ^2).

Question: What does this mean in plain English?

Canonical Form

The exponential family can be written in canonical form as:

$$f(x|\theta) = h(x)\exp\Big(\eta(\theta)^T T(x) - \psi(\theta)\Big).$$

If $\eta(\cdot)$ is invertible, we define the canonical parameter $\eta = \eta(\theta)$ so that $\theta = \eta^{-1}(\eta)$. (This is not optimal notation in terms of writing η as both a function and a variable. We will remedy this soon).

Canonical Form (Restated)

We can re-write the exponential family in its canonical form using η (the canonical parameters):

$$f(x|\theta) = h(x)\exp\Big(\eta(\theta)^T T(x) - \psi(\theta)\Big).$$

$$f(x|\eta) = h(x)\exp\Big(\eta^T T(x) - \psi\big(\eta^{-1}(\eta)\big)\Big).$$

This rewriting shows that in the canonical form the parameters directly multiply with the sufficient statistics, and the function $\psi(\theta)$ is composed with $\eta^{-1}(\cdot)$ as it acts on the (untransformed) parameters θ.

The Binomial Distribution in Canonical Form

Recall that

$$f(x|p) = \binom{n}{x} \exp\left\{ x \log \frac{p}{1-p} + n \log(1-p) \right\} 1\{x \in \{0, 1, \ldots, n\}\}.$$

Taking

$$\eta = \log \frac{p}{1-p} \quad \Longrightarrow \quad 1 - p = \frac{1}{1+e^{\eta}}.$$

We can re-express the binomial distribution in the canonical form as follows:

$$f(x|\eta) = \binom{n}{x} \exp\left\{ \eta x - n \log\left(1 + e^{\eta}\right) \right\} 1\{x \in \{0, 1, \ldots, n\}\}.$$

Canonical Components for the Binomial

Recall that

$$f(x|\eta) = \binom{n}{x} \exp\Big\{\eta x - n\log\big(1+e^{\eta}\big)\Big\}\, 1\{x \in \{0,1,\ldots,n\}\}.$$

In canonical form we have:

$$\eta = \log\frac{p}{1-p}, \quad T(x) = x, \quad A(\eta) = n\log\big(1+e^{\eta}\big), \quad h(x) = \binom{n}{x}.$$

The Log-Partition Function

The function composition $\psi\big(\eta^{-1}(\eta)\big)$ is called the log-partition function. Define $A(\eta) = \psi\big(\eta^{-1}(\eta)\big)$, where we will work with $A(\eta)$ moving forward.

This function is useful as we can easily calculate the mean and variance of distributions in the exponential family by differentiating the log-partition (instead of proceeding with messy integration).

Specifically, the following hold:

$$\frac{\partial A}{\partial \eta_i} = E\big[T_i(x)\big] \quad \text{and} \quad \frac{\partial^2 A}{\partial \eta_i \partial \eta_j} = \text{Cov}\big(T_i(x), T_j(x)\big).$$

The first and second derivatives of $A(\eta)$ are the mean and variances, respectively, of the sufficient statistic.

Log-Partition Function for the Binomial

Recall in canonical form for the binomial distribution, the log-partition function is $A(\eta) = n\log(1 + e^{\eta})$, where $\eta = \log(\frac{p}{1-p})$. Then,

$$E(x) = \frac{\partial A}{\partial \eta} = \frac{ne^{\eta}}{1 + e^{\eta}} = np,$$

and

$$\text{Var}(x) = \frac{\partial^2 A}{\partial \eta^2} = \frac{ne^{\eta}}{(1 + e^{\eta})^2} = np(1 - p).$$

This corresponds to the mean and variance of the binomial distribution.

Maximum Likelihood Estimation

For n i.i.d. samples from an exponential family in canonical form, the joint density is:

$$f(x_1, \ldots, x_n|\eta) = \prod_{i=1}^{n} h(x_i) \exp\Big(\eta^T T(x_i) - A(\eta)\Big) \tag{4}$$

$$= \Big(\prod_{i=1}^{n} h(x_i)\Big) \exp\Big(\sum_{i=1}^{n} \eta^T T(x_i) - nA(\eta)\Big) \tag{5}$$

Question: What important fact do you notice about this formulation?

Maximum Likelihood Estimation

The product of exponential family distributions is also in the exponential family with sufficient statistic $\sum_{i=1}^{n} T(x_i)$.

For example, the sufficient statistic for the joint distribution of n iid Binomial random variables is $\sum_{i=1}^{n} x_i$. Thus, all that is required for inference on the parameter p from the n observations is the sum of all the observations.

Maximum Likelihood Estimation (Derivation)

$$f(x|\eta) = \Big(\prod_{i=1}^{n} h(x_i)\Big) \exp\Big(\sum_{i=1}^{n} \eta^T T(x_i) - nA(\eta)\Big) \tag{6}$$

$$\log L = = \log\Big(\prod_{i=1}^{n} h(x_i)\Big) + \sum_{i=1}^{n} \eta^T T(x_i) - nA(\eta) \tag{7}$$

$$\nabla_\eta \log L = \sum_{i=1}^{n} T(x_i) - n\nabla_\eta A(\eta) := 0 \tag{8}$$

This implies that

$$E[T(x)) = \nabla_\eta A(\eta) = \frac{1}{n}\sum_{i=1}^{n} T(x_i),$$

which is known as a method of moments estimator.

Generalized Linear Models

A generalized linear model has three components:

- An outcome Y that follows a distribution from the exponential family
- A linear predictor $X\beta$
- A link function g that connects the conditional expectation of Y to the linear predictor

$$E(Y \mid X) = g^{-1}(X\beta).$$

Dispersion Parameters

We often incorporate a dispersion parameter into exponential families

$$f(x \mid \theta) = h(x) \exp\Big(\eta(\theta)^T T(x) - \psi(\theta)\Big).$$

The dispersion parameter often gets at the notion of variance, and we essentially have a two-parameter exponential family, where θ corresponds to some notion of the mean and ϕ for the variance such that

$$f(x \mid \theta, \phi) = h(x, \phi) \exp\Big(\frac{\eta(\theta)^T T(x) - \psi(\theta)}{c(\phi)}\Big).$$

This is called an exponential dispersion family, where our prior formulation is a special case of it.[1]

[1] For more details about this, see Dunn and Smith (2018), Chapter 5).

Generalized Linear Models

For an exponential family that is in the canonical form has many convenient properties, as we have already demonstrated.

Generalized Linear Models

Recall that a link function g that connects the conditional expectation of Y to the linear predictor via

$$E(Y \mid X) = g^{-1}(X\beta).$$

- The canonical link is often used since it directly relates the canonical parameter to the linear predictor.
- The canonical link is also often computationally convenient as the MLE's are easily found under this formulation.
- There are reason for using non-canonical links that are often problem specific and could be computationally driven.

Binary Regression

Assume the response Y follows a Bernoulli distribution or rather Binomial distribution with $n = 1$. Let's formulate binary regression as a GLM by using the fact that we know that Y is a member of the exponential family.

Binary Regression

Recall that we previously showed that

$$f(x|\eta) = \exp\Big\{\eta y - \log\big(1+e^{\eta}\big)\Big\}\, 1\{y \in \{0,1,\ldots,n\}\}, \text{where}$$

$$\eta = \log\frac{p}{1-p}, \quad T(x) = x, \quad A(\eta) = \log\big(1+e^{\eta}\big),$$
$$h(y) = 1\{y \in \{0,1,\ldots,n\}\}.$$

Recall that
$p = E(Y \mid X)$ and $p = \frac{e^{\eta}}{(1+e^{\eta})}$.

Binary Regression

It follows that

$$E(Y \mid X) = g^{-1}(X\beta) = \frac{\exp X\beta}{(1 + \exp X\beta)} \implies$$

$$\log \frac{E(Y \mid X)}{1 - E(Y \mid X)} = X\beta,$$

which corresponds to logistic regression. The form of the link function is the canonical link of the Bernoulli distribution.

Logistic Regression

- The outcome Y follows a Bernoulli distribution and is a special case of the Binomial distribution with $n = 1$, which we showed is in the exponential family.
- We assume the function form of the predictors is a linear combination.
- We will use the canonical link function, known as the logit link, which has the following form

$$g\Big(E(Y \mid X)\Big) = X\beta \tag{9}$$

$$\log \frac{E(Y \mid X)}{1 - E(Y \mid X)} = X\beta \tag{10}$$

$$\text{logit}\Big(E(Y \mid X)\Big) = X\beta \tag{11}$$

Why a non-canonical link?

Why would we perhaps choose a non-canonical link over a canonical link?

Choosing a non-canonical link can sometimes improve the fit of the model to the data, especially when the underlying theoretical relationship is better captured by that transformation. This should always be done with care and understanding of the data and application at hand.

Why a non-canonical link?

While the canonical link for a Poisson distribution is the log link, sometimes a logit-link might be used if the data shows over-dispersion (variance greater than the mean).

Why a non-canonical link?

- The probit link function is the inverse of the cumulative distribution function (CDF) of the standard normal distribution, meaning it converts a probability value into a "standard normal deviate" (z-score).
- When analyzing binary outcomes (like yes/no) where you believe the underlying latent variable (the "true" continuous variable that determines the binary outcome) follows a normal distribution, using the probit link can be more theoretically appropriate than the canonical logit link.

Probit regression

- The outcome Y follows a Bernoulli distribution and is a special case of the Binomial distribution with $n = 1$, which we showed is in the exponential family.
- We assume the function form of the predictors is a linear combination.
- We will use a non-canonical link function, Φ^{-1} (the inverse of the normal CDF) to link the conditional mean of Y with $X\beta$.

$$g\Big(E(Y \mid X)\Big) = X\beta \tag{12}$$

$$\Phi^{-1}\frac{E(Y \mid X)}{1 - E(Y \mid X)} = X\beta \tag{13}$$

$$\tag{14}$$

This is called probit regression and Φ is the probit link.

Probit regression

The use of the probit regression model dates back to Bliss (1934).

Bliss was interested in finding an effective pesticide to control insects that fed on grape leaves (Greenberg, 1980).

He applied the probit transformation to transform the shape of the dose-response curve to a linear relationship.

His ideas were later generalized in a book by Finney (1985) where the applications of probit analysis in toxicological experiments were explored.

According to some sources, probit analysis remains the preferred method in understanding dose-response relationships.

Summary

- We have introduced the exponential family of distributions.
- We have provided the connection with exponential families to GLMs through the link function.
- We have introduced some convenient properties under GLMs that belong the the exponential family with the canonical link.
- Next, we will seek to understand how we derive the parameter estimates (β) for generalized linear models.

Exercise 1

1. Consider a uniform distribution on $(0, \theta)$, i.e.,

$$f(x) = \frac{1}{\theta} \quad \text{for } x \in (0, \theta).$$

- Is this a member of the exponential family?
- If so, identify the components in canonical form and use the log-partition function to calculate the MLE for θ from an i.i.d. sample.
- If not, explain why not.

Exercise 2

2. Consider a normal distribution $N(\mu, \mu)$ for $\mu > 0$ (i.e., the variance equals the mean).

- Is this a member of the exponential family?
- If so, identify the components in canonical form and use the log-partition function to calculate the MLE for μ from an i.i.d. sample.
- If not, explain why not.

Exercise 3

3. Consider a distribution with

$$f(x|\lambda) = \lambda\, x^{-1-\lambda} \quad \text{for } \lambda > 0,\ x > 1.$$

- Is this a member of the exponential family?
- If so, identify the components in canonical form and use the log-partition function to calculate the MLE for λ from an i.i.d. sample.
- If not, explain why not.

Solution: Exercise 1

Consider a uniform distribution on $(0, \theta)$, that is,

$$f_X(x) = \frac{1}{\theta} I(0 < x < \theta).$$

No, this is not a member of the exponential family. There is no way to disentangle x from θ in the term $I(0 < x < \theta)$; there is no possible function $h(x)$ that can be free of θ as required by the exponential family form.

Solution: Exercise 2

The density is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\mu}} \exp\left(-\frac{(x-\mu)^2}{2\mu}\right) I(x \in \mathbb{R})$$

$$= I(x \in \mathbb{R}) \frac{\exp(x)}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\mu} - \frac{\mu}{2} - \frac{1}{2}\log\mu\right).$$

Yes, this is a member of the one-parameter exponential family. The components in canonical form are:

$$h(x) = I(x \in \mathbb{R}) \frac{\exp(x)}{\sqrt{2\pi}},$$

$$\eta(\mu) = -\frac{1}{2\mu},$$

$$T(x) = x^2,$$

$$\psi(\mu) = \frac{1}{2}(\mu + \log\mu).$$

Solution: Exercise 2 (Continued)

In canonical form, if we let $\eta = -\frac{1}{2\mu}$, then $\mu = -\frac{1}{2\eta}$ and the log-partition function (or cumulant function) is given by:

$$A(\eta) = \frac{1}{2}\left(-\frac{1}{2\eta} + \log\left(-\frac{1}{2\eta}\right)\right).$$

Note that

$$\nabla_\eta A(\eta) = \frac{1}{4\eta^2} - \frac{1}{2\eta}.$$

Recall $\eta = -\frac{1}{2\mu}$ and plug this in the expression above to find

$$\mu^2 + \mu = \frac{1}{n}\sum_i x_i^2.$$

Solution: Exercise 2 (Continued)

We have a quadratic form in terms of μ as follows:

$$\mu^2 + \mu = \frac{1}{n}\sum_i x_i^2$$

The general solution for a quadratic equation is:

$$\mu = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

This is a quadratic equation:

$$a\mu^2 + b\mu + c = 0$$

where:

- $a = 1$
- $b = 1$
- $c = -\frac{1}{n}\sum_{i=1}^{n} x_i^2$

$$\mu = \frac{-1 + \sqrt{1^2 - 4\,(1)\left(-\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)}}{2}$$

Solution: Exercise 3 (Continued)

We can rewrite the density as:

$$f_X(x) = \lambda\, x^{-\lambda-1} I(x > 1)$$
$$= \exp\Big(-(\lambda+1)\log(x) + \log(\lambda)\Big) I(x > 1).$$

Yes, this is a member of the one-parameter exponential family with components:

$$h(x) = I(x > 1),$$
$$\eta(\lambda) = -(\lambda + 1),$$
$$T(x) = \log(x),$$
$$\psi(\lambda) = -\log(\lambda).$$

Solution: Exercise 3 (Continued)

In canonical form, let $\eta = -(\lambda + 1)$. Then, solving for λ,

$$\lambda = -\eta - 1.$$

The log-partition function is:

$$A(\eta) = -\log(-\eta - 1).$$

To find the MLE for λ, note that the derivative of $A(\eta)$ with respect to η is:

$$\nabla_\eta A(\eta) = \frac{1}{-\eta - 1} = \frac{1}{\lambda}.$$

Thus, equating the sample average of the sufficient statistic to the derivative of the log-partition function, we have:

$$\frac{1}{n}\sum_{i=1}^{n} T(x_i) = \frac{1}{n}\sum_{i=1}^{n} \log(x_i) = \frac{1}{\lambda}.$$

Therefore, the MLE for λ is:

$$\widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} \log(x_i)}.$$