ST346 Chapter 4

Contents

Preface

These slides are a slight adaptation from the original slides developed by Prof Martyn Plummer for the module.

If you find any typos, please inform the module leader.

Chapter 4 Exponential dispersion models

4.1 Motivation

Before we define the class of exponential dispersion models, we take a moment to understand why they are needed.

Consider a simple problem of estimating a common mean.

- Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with mean μ.
- We propose a family of probability density functions $p(y \mid \mu)$ parameterized by μ.

We may estimate the mean μ using two alternative approaches.

1. Sample mean:
$$\overline{\mu} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} y_i.$$
2. Maximum likelihood estimate:
$$\widehat{\mu} \quad = \quad \arg\max_{\mu} \sum_{i=1}^{n} \log\Big(p(y_i \mid \mu)\Big).$$

Question: Is the maximum likelihood estimate $\widehat{\mu}$ identical to the sample mean $\overline{\mu}$?

Let's perform an `R` demo to test this for the following distributions:

- Normal,
- Gamma,
- t.

The maximum likelihood estimate $\widehat{\mu}$ is identical to the sample mean $\overline{\mu}$ for some probability models (e.g. normal, gamma). But this is not true for all probability models (e.g. t-distribution).

- Exponential dispersion models (EDMs) are probability models for which $\widehat{\mu} = \overline{\mu}$ for i.i.d. observations with common mean μ.
- This property uniquely characterizes EDMs.
- For non-EDMs, the sample mean may still be a consistent and efficient estimator of μ.
- Hence $\overline{\mu}$ may be "close to" $\widehat{\mu}$, but not identical.

4.2 Definition of an EDM

An EDM is a distribution from the exponential family of distributions. The probability density function (or probability mass function) of an EDM can be put in the canonical form:

$$p(y \mid \theta, \phi) \quad = \quad a(y, \phi) \ \exp\left(\frac{\theta y - b(\theta)}{\phi}\right)$$

where

- $\theta \in \Theta$ is the **canonical parameter** and $\Theta = \{\theta \in \mathbb{R} : |b(\theta)| < \infty\}$,
- $\phi \in \mathbb{R}^+$ is the **dispersion parameter**. The dispersion parameter may be free, in which case it is an additional parameter to be estimated, or it may be fixed to a known value (usually $\Phi = 1$).
- $a(y, \phi)$ is the **normalizing function.** It ensures that

  $$\int_{y \in \mathcal{S}} p(y \mid \theta, \phi) \quad = \quad 1$$

  where $\mathcal{S}$ is the support (the permitted values of y). The normalizing function does not depend on θ and plays no role in inference on θ.

The support $\mathcal{S}$ of an EDM is determined by its normalizing function $a(y, \phi)$.

Different EDMs have different support.

Distribution	Support $\mathcal{S}$
Normal	$\mathbb{R}$
Poisson	$\mathbb{N}_0$
Scaled Binomial	$\{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\}$

So far we have considered regression models for three distributions:

- Normal (Gaussian),
- Binomial,
- Poisson.

These are all examples of Exponential Dispersion Models (EDM). Other examples include:

- Gamma,
- Inverse Gaussian,
- Negative Binomial.

Distribution	Support $\mathcal{S}$
Negative Binomial	$\mathbb{N}_0$
Gamma	$\mathbb{R}^+$
Inverse Gaussian	$\mathbb{R}^+$

4.3 Weighted EDMs

Suppose we have independent $Y_1, \ldots, Y_n$ from an EDM with the same canonical parameter θ but different dispersion parameters $\phi_1, \ldots, \phi_n$.

We can extend our definition of EDMs to include this case if we assume

$$\phi_i = \frac{\phi}{w_i}$$

for known weights $w_1, \ldots, w_n$ and common dispersion parameter ϕ.

The density function is then

$$p(y_i \mid \theta, \phi) = a(y_i, \phi/w_i) \exp\left(\frac{w_i\big[\theta y_i - b(\theta)\big]}{\phi}\right).$$

For fixed ϕ, the log likelihood of θ is

$$\log\Big(L(\theta \mid \phi, \boldsymbol{y})\Big) = \frac{1}{\phi} \sum_{i=1}^{n} w_i\big[\theta y_i - b(\theta)\big] + \ldots$$

where terms depending on the normalizing function $a(y, \phi/w_i)$ have been omitted.

An observation with weight $w_i \in \mathbb{N}$ makes the same contribution to the log likelihood as w_i identical observations with weight 1.

4.4 Examples

4.4.1 Normal distribution

Recall the canonical form of the pdf/pmf of an EDM:

$$p(y \mid \theta, \phi) \quad = \quad a(y, \phi) \ \exp\left(\frac{\theta y - b(\theta)}{\phi}\right).$$

The density of a $\mathcal{N}(\mu, \sigma^2)$ can be written in canonical form as

$$\begin{aligned} p(y \mid \mu, \sigma^2) \quad &= \quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\ &= \quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-y^2 + 2\mu y - \mu^2}{2\sigma^2}\right) \\ &= \quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-y^2}{2\sigma^2}\right) \exp\left(\frac{\mu y - \frac{1}{2}\mu^2}{\sigma^2}\right). \end{aligned}$$

Comparing this to the general canonical form

$$p(y \mid \theta, \phi) \quad = \quad a(y, \phi) \ \exp\left(\frac{\theta y - b(\theta)}{\phi}\right)$$

we deduce $\theta = \mu$ and $\phi = \sigma^2$.

Now

$$p(y \mid \mu, \sigma^2) \quad = \quad p(y \mid \theta, \phi) \quad = \quad \frac{1}{\sqrt{2\pi\phi}} \exp\left(\frac{-y^2}{2\phi}\right) \exp\left(\frac{\theta y - \frac{1}{2}\theta^2}{\phi}\right)$$

is a pmf in canonical form with

$$\begin{aligned} \phi \quad &= \quad \sigma^2, \\ b(\theta) \quad &= \quad \frac{\theta^2}{2}, \text{ and} \\ a(y, \phi) \quad &= \quad \frac{1}{\sqrt{2\pi\phi}} \exp\left(\frac{-y^2}{2\phi}\right). \end{aligned}$$

4.4.2 Scaled binomial distribution

We have

$$
\begin{aligned}
p(y \mid \mu, m) &= \binom{m}{my} \mu^{my} (1-\mu)^{m(1-y)} \\
&= \\
&= \\
&=
\end{aligned}
$$

Set

$$\theta = \log(\frac{\mu}{1-\mu}), \quad w = m, \quad \phi = 1 \quad \text{and} \quad a(y, \phi) = \binom{m}{my}.$$

To determine $b(\theta)$ in terms of θ we need to express μ as a function of θ.

We have

$$\theta \quad = \quad \log\left(\frac{\mu}{1-\mu}\right) \qquad \text{if and only if} \qquad \mu \quad = \quad \frac{\exp(\theta)}{1+\exp(\theta)}.$$

Therefore

$$
\begin{aligned}
b(\theta) &= -\log(1-\mu) \\
&= -\log\left(1 - \frac{\exp(\theta)}{1+\exp(\theta)}\right) \\
&= -\log\left(\frac{1}{1+\exp(\theta)}\right) \\
&= \log\Big(1+\exp(\theta)\Big).
\end{aligned}
$$

Exercise 9: canonical form

Derive the canonical form of the probability mass function for the Poisson distribution.

4.5 Cumulants for EDMs

Recall the canonical form of an EDM

$$p(y \mid \theta, \phi) \quad = \quad a(y, \phi) \ \exp\left(\frac{\theta y - b(\theta)}{\phi}\right).$$

The function $b(\theta)$ is called the **cumulant function.**

The **cumulant generating function** of an EDM is

$$K(t) \quad = \quad \frac{b(\theta + t\phi) - b(\theta)}{\phi}.$$

Note that the recommended textbook by Dunn and Smyth uses $\kappa(\theta)$ (kappa) for the cumulant function, whereas we will use $b(\theta)$.

We can obtain the mean and variance of the EDM from the derivatives of the cumulant generating function:

$$\begin{aligned}
K(t) &= \frac{b(\theta + t\phi) - b(\theta)}{\phi}, \\
K'(t) &= \frac{\phi\ b'(\theta + t\phi)}{\phi} \quad = \quad b'(\theta + t\phi), \\
K''(t) &= \phi b''(\theta + t\phi).
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{E}(Y \mid \phi, \theta) &= K'(0) &= b'(\theta) \\
\mathbb{V}ar(Y \mid \phi, \theta) &= K''(0) &= \phi b''(\theta)
\end{aligned}$$

Therefore, the mean is independent of ϕ and the variance is proportional to ϕ (hence the name "dispersion parameter".)

Next we prove that for EDMs the cumulant generating function is given by

$$K(t) \quad = \quad \frac{b(\theta + t\phi) - b(\theta)}{\phi}.$$

Proof Recall the canonical form of an EDM

$$p(y \mid \theta, \phi) \quad = \quad a(y, \phi)\ \exp\left(\frac{\theta y - b(\theta)}{\phi}\right).$$

Then,

$$\begin{aligned}
M(t) &= \mathbb{E}\Big(\exp(tY)\Big) \\
\\
&= \\
\\
&= \\
\\
&=
\end{aligned}$$

Let $\theta^* = \theta + t\phi$, then

$$\begin{aligned}
M(t) &= \int_{y\in\mathcal{S}} a(y,\phi) \exp\left(\frac{\theta^* y - b(\theta)}{\phi}\right) dy \\
&= \int_{y\in\mathcal{S}} a(y,\phi) \exp\left(\frac{\theta^* y - b(\theta^*) + b(\theta^*) - b(\theta)}{\phi}\right) dy \\
&= \int_{y\in\mathcal{S}} a(y,\phi) \exp\left(\frac{\theta^* y - b(\theta^*)}{\phi}\right) \exp\left(\frac{b(\theta^*) - b(\theta)}{\phi}\right) dy \\
&= \exp\left(\frac{b(\theta^*) - b(\theta)}{\phi}\right) \int_{y\in\mathcal{S}} p(y \mid \theta^*, \phi) dy \\
&= \exp\left(\frac{b(\theta^*) - b(\theta)}{\phi}\right).
\end{aligned}$$

Hence the cumulant generating function is given by

$$\begin{aligned} K(t) &= \log\Big(M(t)\Big) \\ &= \frac{b(\theta^*) - b(\theta)}{\phi} \\ &= \frac{b(\theta + t\phi) - b(\theta)}{\phi}. \end{aligned}$$

4.6 The canonical link

4.6.1 Definition

Recall that the mean and variance of an EDM can be derived from the cumulant function:

$$\begin{aligned} \mathbb{E}(Y \mid \theta) &= b'(\theta) \\ \mathbb{V}ar(Y \mid \phi, \theta) &= \phi b''(\theta) \end{aligned}$$

Let $\mu = \mathbb{E}(Y \mid \theta) = b'(\theta)$, then

$$\frac{d\mu}{d\theta} = b''(\theta) = \frac{\mathbb{V}ar(Y \mid \phi, \theta)}{\phi} > 0.$$

So μ is a strictly increasing function of θ (and vice versa) and thus, there is a one-to-one correspondence between the canonical parameter θ and the mean μ.

For every EDM there is a function g that maps μ onto the canonical parameter θ

$$g(\mu) = \theta.$$

This is the **canonical link function**.

We can derive the canonical link from the cumulant function. The canonical mean function $h(\theta)$ is the inverse of the canonical link function:

$$h(\theta) = \mu = b'(\theta).$$

So we invert $h()$ to get

$$\theta = g(\mu).$$

4.7 Examples of canonical link functions

4.7.1 The normal distribution

The cumulant function is

$$b(\theta) = \frac{\theta^2}{2}.$$

Hence

$$\mu = b'(\theta) = \theta.$$

Therefore, the canonical link for the normal distribution is the **identity link**.

$$\theta = \mu.$$

4.7.2 The Poisson distribution

As shown in Exercise 9, the cumulant function for the Poisson distribution is

$$b(\theta) \quad = \quad \exp(\theta).$$

Hence

$$\mu \quad = \quad b'(\theta) \quad = \quad \exp(\theta).$$

Solving for θ gives

$$\theta \quad = \quad \log(\mu).$$

Therefore, the canonical link for the Poisson distribution is the **log link**.

4.7.3 The scaled Binomial distribution

The cumulant function is

$$b(\theta) \quad = \quad \log\Big(1 + \exp(\theta)\Big).$$

Hence

$$\mu \quad = \quad b'(\theta) \quad = \quad \frac{\exp(\theta)}{1 + \exp(\theta)}.$$

Solving for θ gives

$$\theta \quad = \quad \log\left(\frac{\mu}{1 - \mu}\right).$$

Therefore, the canonical link for the (scaled) Binomial distribution is the **logit link**.

Exercise 10: canonical link

Find the canonical link given the cumulant function for the following distributions:

- The gamma distribution

$$b(\theta) \quad = \quad -\log(-\theta) \quad \text{for } \theta < 0.$$

- The negative binomial distribution

$$b(\theta) \quad = \quad -k \log\Big(1 - \exp(\theta)\Big) \quad \text{for } \theta < 0.$$

- The inverse Gaussian distribution

$$b(\theta) \quad = \quad -\sqrt{-2\theta} \quad \text{for } \theta < 0.$$

4.8 Deviance

In ST231 we derived our parameter estimates by minimizing the residual sum of squares function, or **deviance.**

$$D(\boldsymbol{y}, \boldsymbol{\mu}) = \sum_{i=1}^{n} (y_i - \mu_i)^2.$$

We saw that, because the normal density is given by

$$p(y \mid \mu, \sigma^2) \propto \exp\Big(- \frac{(y-\mu)^2}{2\sigma^2} \Big),$$

the likelihood

$$p(\boldsymbol{y} \mid \boldsymbol{\mu}, \sigma^2) \propto \exp\Big(- \frac{\sum_{i=1}^{n} (y_i - \mu_i)^2}{2\sigma^2} \Big)$$

is maximized when the deviance is minimized.

In week 1 of ST346 we extended this to weighted models:

$$D(\boldsymbol{y}, \boldsymbol{\mu}) = \sum_{i=1}^{n} w_i (y_i - \mu_i)^2.$$

The concept of deviance arises naturally from EDMs, but different EDMs will have different formulae for the deviance.

The one-to-one correspondence between θ and μ implies that we can re-write the density function of an EDM in terms of μ, ϕ instead of θ, ϕ.

Let

$$t(y, \mu) = \theta y - b(\theta),$$

then the canonical form of the EDM is given by

$$\begin{aligned}
p(y \mid \theta, \phi) &= a(y, \phi) \; \exp\Big(\frac{\theta y - b(\theta)}{\phi} \Big) \\
&= a(y, \phi) \; \exp\Big(\frac{t(y, \mu)}{\phi} \Big) \\
&= a(y, \phi) \; \exp\Big(\frac{t(y, y)}{\phi} \Big) \; \exp\Big(\frac{t(y, \mu) - t(y, y)}{\phi} \Big) \\
&= a^*(y, \phi) \; \exp\Big(- \frac{2\big[t(y, y) - t(y, \mu) \big]}{2\phi} \Big)
\end{aligned}$$

where $a^*(y, \phi) = a(y, \phi) \exp\Big(t(y, y)/\phi \Big)$.

Setting

$$d(y, \mu) \quad = \quad 2\Big(t(y, y) - t(y, \mu)\Big).$$

gives the **dispersion model** form of the EDM:

$$p(y \mid \mu, \phi) \quad = \quad a^*(y, \phi) \exp\Big(-\frac{d(y, \mu)}{2\phi}\Big)$$

where $d(y, \mu)$ is the **unit deviance.**

Proposition 4.1 (Unit deviance)**.** *The unit deviance $d(y, \mu)$ is non-negative and exactly zero if and only if $\mu = y$. It thus is a measure of the discrepancy between the expected value μ and the observed value y.*

Proof

If $t(y, \mu)$ has a unique maximum at $\mu = y$, then the unit deviance

$$d(y, \mu) \quad = \quad 2\Big(t(y, y) - t(y, \mu)\Big)$$

is zero at $\mu = y$ and positive otherwise.

Consider t as a function of θ. Then

$$\frac{dt}{d\theta} \;=\; y - b'(\theta) \;=\; y - \mu$$

and so

$$\frac{dt}{d\theta} \;=\; 0 \qquad \Longleftrightarrow \qquad \mu \;=\; y.$$

Moreover

$$\frac{d^2t}{d\theta^2} \;=\; -b''(\theta) \;=\; -\mathbb{V}ar(Y)/\phi \;<\; 0.$$

Hence t is a strictly concave function of θ with a unique maximum at $y = \mu$.

It follows that the unit deviance is non-negative and is exactly zero if and only if $\mu = y$.

4.8.1 Example: normal unit deviance

Let's derive the unit deviance for the normal distribution from the canonical form.

Recall that $b(\theta) = \theta^2/2$ and the canonical linke is $\mu = \theta$. Then

$$\begin{aligned} t(y,\mu) &= y\theta - b(\theta) \\ &= y\theta - \theta^2/2 \\ &= y\mu - \mu^2/2. \end{aligned}$$

This is maximised at

$$t(y,y) \quad = \quad y^2 - y^2/2 \quad = \quad y^2/2.$$

Hence the unit deviance for the normal distribution is

$$\begin{aligned} d(y,\mu) &= 2\Big(t(y,y) - t(y,\mu)\Big) \\ &= 2\Big(y^2/2 - y\mu + \mu^2/2\Big) \\ &= y^2 - 2y\mu + \mu^2 \\ &= (y-\mu)^2. \end{aligned}$$

4.8.2 Unit deviance on the boundary

If the parameter space for μ is bounded we need to take extra care.

If y lies on the boundary of the possible values of μ, it is possible that $t(y,y)$ is not defined.

So we modify the definition of the unit deviance.

Bounded below:

$$d(y,\mu) \quad = \quad \lim_{\epsilon \to 0} 2\Big(t(y, y+\epsilon) - t(y,\mu)\Big) \quad \text{for } \epsilon > 0.$$

Bounded above

$$d(y,\mu) \quad = \quad \lim_{\epsilon \to 0} 2\Big(t(y, y-\epsilon) - t(y,\mu)\Big) \quad \text{for } \epsilon > 0.$$

4.8.3 Example: Poisson unit deviance

Recall that the cumulant function is $b(\theta) = \exp(\theta)$ and the canonical link is $\log(\mu) = \theta$.

Hence

$$t(y, \mu) = y\theta - b(\theta) = y \log(\mu) - \mu.$$

Note that $t(y, y)$ is not defined for $y = 0$.

First case: if $y > 0$, then

$$\begin{aligned} d(y, \mu) &= 2\Big(t(y, y) - t(y, \mu)\Big) \\ &= 2\Big(y \log(y) - y - y \log(\mu) + \mu\Big) \\ &= 2\Big(y \log(y/\mu) - (y - \mu)\Big) \end{aligned}$$

So the unit deviance depends partly on the ratio y/μ and partly on the difference $y - \mu$.

Second case: if $y = 0$, then

$$\begin{aligned} d(0, \mu) &= \lim_{\epsilon \to 0} 2\Big(t(0, \epsilon) - t(0, \mu)\Big) \\ &= \lim_{\epsilon \to 0} 2\Big(0 \times \log(\epsilon) - 0 - 0 \times \log(\mu) + \mu\Big) \\ &= 2\mu. \end{aligned}$$

4.8.4 Total deviance

Suppose we have independent $Y_i \sim \text{EDM}(\mu_i, \phi/w_i)$ for $i = 1, \ldots, n$, where $w_1, \ldots, w_n$ are fixed weights.

The **total deviance** is

$$D(\boldsymbol{y}, \boldsymbol{\mu}) = \sum_{i=1}^{n} w_i \, d(y_i, \mu_i).$$

The **scaled deviance** is

$$D^*(\boldsymbol{y}, \boldsymbol{\mu}) = \frac{D(\boldsymbol{y}, \boldsymbol{\mu})}{\phi}.$$

- The total deviance $D(\boldsymbol{y}, \boldsymbol{\mu})$ and the scaled deviance $D^*(\boldsymbol{y}, \boldsymbol{\mu})$ measure the discrepancy between the observed values $y_1, \ldots, y_n$ and the corresponding mean values predicted by the model $\mu_1, \ldots, \mu_n$.
- The smaller the deviance the better the fit. Hence the total deviance measures **relative** goodness-of-fit of the model.
- Later we will see that we can compare the deviance of nested models and generalize Analysis of Variance (ANOVA) to Analysis of Deviance.

Exercise 11 - Binomial unit deviance

Derive the unit deviance for the scaled binomial distribution. There are three cases to consider

1. $y = 0$,
2. $y = r/m$ for $r = 2, \ldots, m - 1$,
3. $y = 1$.

You will find the expressions for the cumulant function and the canonical link in previous lecture slides.