ST346 Chapter 5

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.

These materials are solely for your own use and you must not distribute these in any format. **Do not upload these materials to the internet or any filesharing sites nor provide them to any third party or forum.**

Chapter 5 Generalized linear models

5.1 The variance function

5.1.1 Introduction

Recall that the mean and variance of an EDM can be derived from the cumulant function $b(\theta)$.

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

Furthermore, there is a one-to-one correspondence between the canonical parameter θ and the mean parameter $\mu = \mathbb{E}(Y)$ via the canonical link.

Hence we can rewrite the variance in terms of μ and ϕ as

$$\mathbb{V}ar(Y \mid \mu, \phi) = \phi V(\mu)$$

where $V(\mu)$ is the **variance function.**

5.1.2 Properties of the variance function

Proposition 5.1. An EDM is uniquely defined by the relation

$$\mathbb{V}ar(Y \mid \mu, \phi) = \phi V(\mu).$$

To prove the proposition we need the following lemmas.

Lemma 5.1.

$$\theta(\mu) = \int \frac{1}{V(\mu)} d\mu.$$

Proof. (Lemma 5.1) Recall that $\mu = b'(\theta)$ and so

NB: Integrals are indefinite, so we can add an arbitrary constant C to θ. □

Lemma 5.2.

$$b\Big(\theta(\mu)\Big) \quad = \quad \int \frac{\mu}{V(\mu)} d\mu.$$

Proof. (Lemma 5.2)

$$\begin{aligned} b'(\theta) &= \mu \\ \Longrightarrow \qquad &= \\ &= \\ &= \end{aligned}$$

NB: Integrals are indefinite, so we can add an arbitrary constant C to $b(\theta)$. □

Next, the proof of Proposition 5.1.

Proof. (Proposition 5.1) Lemmas 5.1 and 5.2 show that we can construct the canonical parameter θ and the cumulant function $b(\theta)$ from the variance function. As we have seen earlier, the cumulant generating function is then defined as

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

If the cumulant generating function exists then it uniquely determines the distribution.

NB: Any constant terms added to $b(\theta)$ cancel out. □

Proposition 5.1 means that when modelling real-world data, we need to check the mean-variance relationship. This will determine which EDM to use.

5.1.3 Example - normal distribution

For the normal distribution $V(\mu) = 1$. Hence

$$\theta(\mu) \quad = \quad \int \frac{1}{V(\mu)} d\mu \quad =$$

and

$$b(\theta(\mu)) \quad = \quad \int \frac{\mu}{V(\mu)} d\mu \quad =$$

Substituting $\mu = \theta$ gives

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

Exercise 12 - variance function

Derive the cumulant function for the scaled binomial and the Poisson distributions from their variance functions. Recall

- the scaled binomial variance function is $V(\mu) = \mu(1-\mu)$;
- The Poisson variance function is $V(\mu) = \mu$.

5.2 Quasi-likelihood

5.2.1 Introduction

It is possible that an EDM does not exist for a given combination of dispersion parameter ϕ and variance function $V(\mu)$.

For example, the Poisson and binomial distributions assume $\phi = 1$. But if we allow ϕ to be free, then there is no EDM for $\phi \neq 1$.

Surprisingly, this does not stop us. We can define a **quasi-likelihood**:

- Quasi-Poisson: ϕ free, $V(\mu) = \mu$.
- Quasi-binomial: ϕ free, $V(\mu) = \mu(1-\mu)$.

Later we will see that the maximum likelihood estimator and its large sample properties depend only on:

- the dispersion parameter ϕ,
- the variance function $V(\mu)$, and
- the link function $g(\mu)$.

So we can still conduct inference for quasi-likelihoods even when there is no corresponding EDM.

5.2.2 Example - Noisy miner data

A study[1] examined the effect of the density of eucalyptus trees on the invasion of woodland habitats by noisy miners, a small but aggressive Australian bird. Data from the study is available as the dataframe `nminer` in the `GLMsData` package and discussed in Example 5.9 in the textbook by Dunn and Smyth.[2]

The noisy miner data is an example of over-dispersion, where the Poisson variance function $V(\mu) = \mu$ fits the data, but the dispersion parameter $\phi > 1$.

[1] M. Maron (2007): Threshold effect of eucalypt density on an aggressive avian competitor. Biological Conservation, 136, 100–107.

[2] Dunn, P. K. and Smyth, G.K (2018): Generalized linear models with examples in R Vol. 53. New York: Springer.

5.3 The gamma distribution

5.3.1 Properties

The gamma distribution with rate $r > 0$ and shape $s > 0$

$$Y \quad \sim \quad \Gamma(r, s)$$

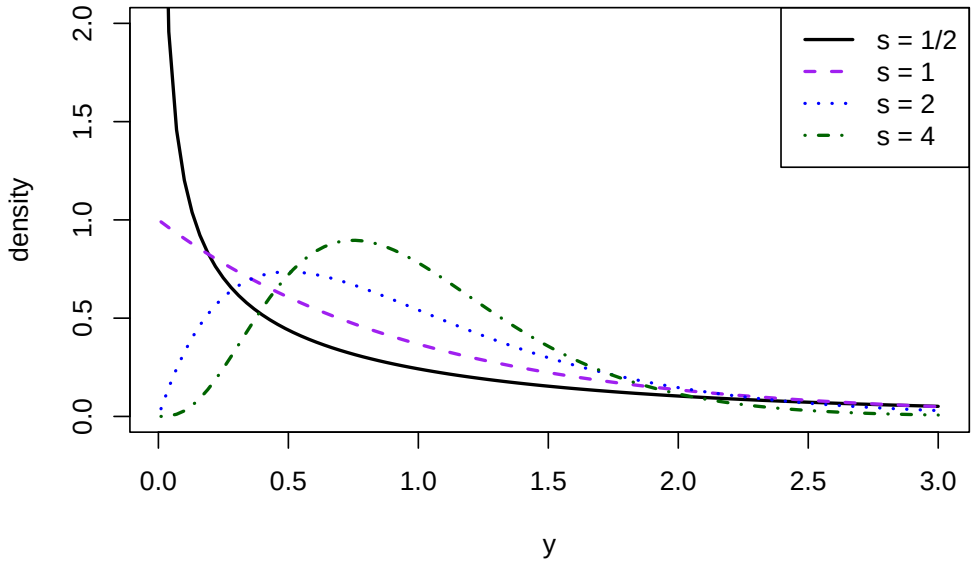
has probability density function

$$p(y \mid r, s) \quad = \quad \frac{\exp(-ry) r^s y^{s-1}}{\Gamma(s)}$$

for $y > 0$.

The graph below shows the gamma density for different values of the shape parameter but with a fixed mean.

- For $s = 1$, the gamma distribution is the exponential distribution.
- For $s > 1$, the density is unimodel and becomes increasingly symmetric as $s \to \infty$.
- For $s < 1$, the density tends to ∞ as $y \to 0$.



The gamma distribution is an EDM with

$$\mu \; = \; \frac{s}{r} \qquad \text{and} \qquad \phi \; = \; \frac{1}{s}.$$

The density can be rewritten as

$$p(y \mid \mu, \phi) \quad = \quad \left(\frac{y}{\phi\mu}\right)^{\frac{1}{\phi}} \frac{1}{y} \exp\left(-\frac{y}{\phi\mu}\right) \frac{1}{\Gamma(\frac{1}{\phi})}$$

[You are not expected to memorize this density.]

The cumulant function of the gamma distribution is

$$b(\theta) \quad = \quad -\log(-\theta).$$

This yields mean μ and variance function $V(\mu)$ as

$$\begin{aligned} \mu &= \\ V(\mu) &= \end{aligned}$$

The canonical link function is

$$g(\mu) \quad = \quad -\frac{1}{\mu}.$$

The default link for the gamma distribution in `R` is slightly different as it omits the minus sign:

$$g(\mu) \quad = \quad \frac{1}{\mu}.$$

5.3.2 Implementation in R

Use the argument `family=Gamma` in the command `glm()`!

Note the capital letter **G**.

In practice we do not use the canonical link. The log link is more numerically stable and makes it easier to interpret parameters.

For a log link function use `family=Gamma(link="log")`.

5.3.3 Example: the lime dataset

The `lime` data set in the package `GLMsData` concerns measurements of leaf biomass from small-leaved lime trees (*Tilia cordata*).

See Sections 11.2 and 11.3 in the textbook by Dunn and Smyth.[3]

[3]Dunn, P. K. and Smyth, G.K (2018): Generalized linear models with examples in R Vol. 53. New York: Springer.

5.4 The inverse Gaussian distribution

5.4.1 Properties

The inverse Gaussian distribution

$$Y \quad \sim \quad IG(\mu, \phi)$$

has density function

$$p(y \mid \mu, \phi) = \sqrt{\frac{1}{2\pi\phi y^3}} \; \exp\Big(- \frac{(y-\mu)^2}{2\phi\mu^2 y} \Big)$$

for $y > 0, \mu > 0$ and $\phi > 0$.

The inverse Gaussian is an EDM with

$$\theta \; = \; -\frac{1}{2\mu^2} \qquad \text{and} \qquad b(\theta) \; = \; -\sqrt{-2\theta}.$$

Its variance function is

$$V(\mu) \quad = \quad \mu^3.$$

The inverse Gaussian distribution is useful for continuous distributions on the positive real line where the variance increases very rapidly with the mean.

5.4.2 Relationship to Brownian Motion

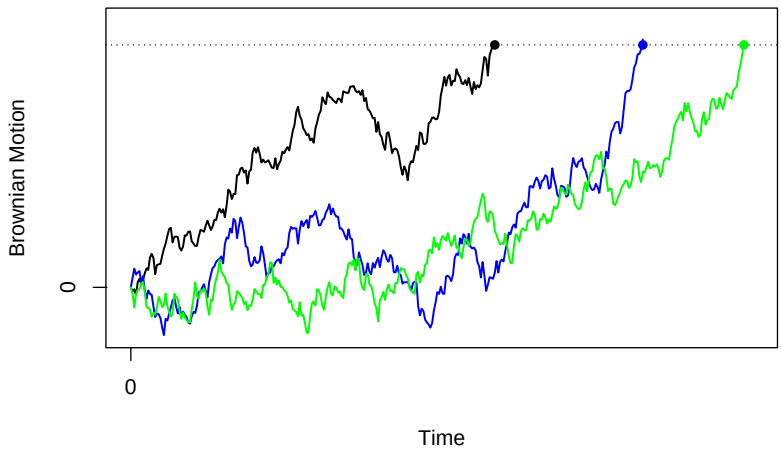
Consider a stochastic process with

$$\begin{aligned} X_0 \quad &= \quad 0, \\ X_t \quad &= \quad \frac{t}{\mu} + \sigma W_t. \end{aligned}$$

where W_t is a Brownian motion (or Wiener process). This is a Brownian motion with drift $\frac{1}{\mu}$. A random walk is superimposed on top of a systematic tendency for the process $\{X_t\}$ to increase with time t. The hitting time for the process to hit fixed level $X_t = 1$ has an inverse Gaussian distribution:

$$T \; = \; \inf\{t > 0 \mid X_t = 1\} \; \sim \; IG(\mu, \sigma^2).$$

The plot below shows three realisations of the Brownian motion with drift, using the same parameters for μ and ϕ. The dots show the hitting times for the three realisations.



The canonical link is

$$g(\mu) \quad = \quad -\frac{1}{2\mu^2}.$$

The default link for the inverse Gaussian in `R` is slightly different as it omits the minus sign

$$g(\mu) \quad = \quad -\frac{1}{2\mu^2}.$$

In practice it is rarely used. The log link is numerically more stable and makes it easier to interpret parameters.

5.4.3 Example: the perm dataset

The `perm` dataset in the `R` package `GLMsData` concerns measurements of permeability of building materials.

- Permeability is a measure of the ease of passage of liquids or gases through the material.
- The Brownian motion with drift is a useful physical model for permeability and hence the inverse Gaussian is a useful statistical model for the results of the permeability tests.

See Sections 11.4 and 11.7.1 in the recommended textbook by Dunn and Smyth.[4]

5.5 The Poisson limit of the binomial distribution

5.5.1 Binomial distribution for rare outcomes

Suppose that Y has an **unscaled** binomial distribution

$$Y \quad \sim \quad \text{Binomial}(m, \mu).$$

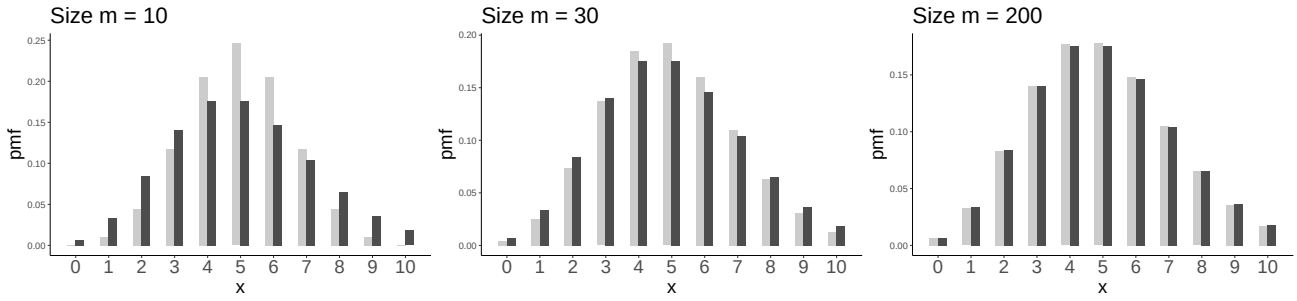
Then Y has expectation $\mathbb{E}(Y) \quad = \quad m\mu$.

Suppose we increase the size parameter m but keep $\mathbb{E}(Y)$ constant. What happens to the distribution of Y?

[4]Dunn, P. K. and Smyth, G.K (2018): Generalized linear models with examples in R Vol. 53. New York: Springer.

Example: Consider $\mathbb{E}(Y) = 5$.

- The dark bars show the probability mass function (pmf) for the binomial distribution as the size parameter m increases.
- The light bars show the pmf of the Poisson distribution.



More formally, suppose

$$Y \quad \sim \quad \text{Binomial}(m, \mu)$$

with $\mu = \lambda/m$ so that

$$\mathbb{E}(Y) \quad = \quad m\mu \quad = \quad \lambda.$$

Then the limiting distribution of Y as $m \to \infty$ with λ fixed is

$$Y \quad \sim \quad \text{Poisson}(\lambda).$$

Proof Consider the moment generating function (mgf) of the (unscaled) binomial distribution

$$\begin{aligned} M(t) \quad &= \quad \mathbb{E}\Big(\exp(tY) \Big) \\ &= \quad \sum_{r=0}^{m} \binom{m}{r} \Big(\exp(t)\mu \Big)^r (1-\mu)^{m-r} \\ &= \quad \sum_{r=0}^{m} \binom{m}{r} \Big(\exp(t)\mu \Big)^r (1-\mu)^{m-r} \\ &= \quad \Big(\exp(t)\mu + 1 - \mu \Big)^m \\ &= \quad \Big(1 + \mu \Big[\exp(t) - 1 \Big] \Big)^m. \end{aligned}$$

Substituting $\mu = \lambda/m$ we obtain

$$\begin{aligned} M(t) \quad &= \quad \Big(1 + \frac{\Big[\exp(t) - 1 \Big] \lambda}{m} \Big)^m \\ &= \quad \Big(1 + \frac{a}{m} \Big)^m \qquad \text{for } a = \lambda[\exp(t) - 1]. \end{aligned}$$

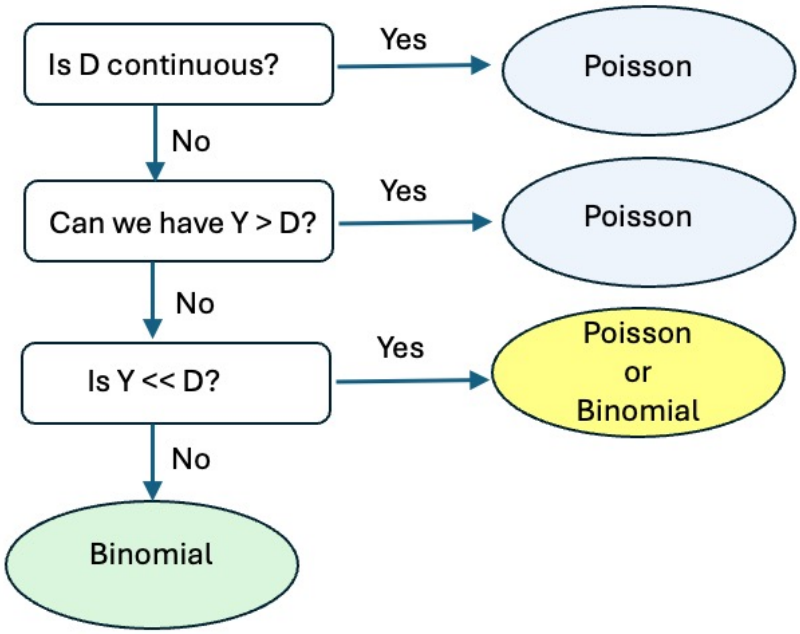
Taking the limit $m \to \infty$ with t, λ fixed we obtain

$$M(t) \quad \longrightarrow \quad \exp(a) = \exp\Big(\lambda[\exp(t) - 1]\Big)$$

which is the mgf of the Poisson distribution with mean λ.

5.6 Choosing between the Poisson and Binomial GLM

We have seen that we can use a Poisson GLM to model rates and a Binomial GLM to model proportions. For rare events we have a choice between the two. Prof Martyn Plummer designed the following flowchart to help you decide. In the flow chart Y is the count response variable and D the denominator which turns the count into a rate or a proportion.



Exercise 13 - GLM scenarios

Consider the following scenarios. Determine which model might be appropriate to use in each case. Give the exponential dispersion model, the link function, the outcome variable, and the predictor variables.

a. A cohort of 18 year-old school leavers were surveyed regarding their future plans, with specific interest in higher education. Several variables were measured, including age, gender, smoking status, as well as whether or not they had a place at university or other higher education institute. We wish to investigate the variables which are associated with higher education attendance.

b. Sixty rats of the same age were divided into 2 groups of 30. At the start of the experiment, all animals were weighed. Then one group was fed a control diet, whilst the other was fed a diet supplemented with vitamin D. After 6 weeks, all rats were weighed again. We wish to understand how vitamin D supplementation affects weight gain.

c. A car manufacturer conducted a study to investigate the reliability of their cars. They measured the number of times that each of 1000 vehicles had broken down in the 10 years since it was made. They also measured the number of miles that each car had been driven as well as the model of each vehicle and the number of times that it had been serviced.