ST346 Week 6

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 6 Maximum likelihood estimation for GLMs

6.1 Review on maximum likelihood estimation

6.1.1 Likelihood

In the following we assume **suitable regularity conditions** that will be satisfied by the models considered in this module.

Suppose we observe independent random variables $Y_1, \dots, Y_n$, where the pdf of Y_i is

$$p_i(y \mid \boldsymbol{\beta}) \qquad \text{for } \boldsymbol{\beta} \in \mathbb{R}^p,$$

assumed to be a member of the exponential family of distributions.

- $Y_1, \dots, Y_n$ are independent but not identically distributed.
- The distribution of Y_i is parameterized by $\boldsymbol{\beta}$.

The likelihood $L(\boldsymbol{\beta})$ is the joint pdf considered as a function of the parameters:

$$\begin{aligned} L : \mathbb{R}^p &\to \mathbb{R} \\ \boldsymbol{\beta} &\mapsto \prod_{i=1}^{n} p_i(y_i \mid \boldsymbol{\beta}) \end{aligned}$$

We normally work in terms of the log likelihood, as the log likelihood is the sum of individual contributions from independent observations

$$l(\boldsymbol{\beta} \mid \boldsymbol{y}) \quad = \quad \log\Big(L(\boldsymbol{\beta} \mid \boldsymbol{y})\Big) \quad = \quad \sum_{i=1}^{n} \log\Big(p_i(y_i \mid \boldsymbol{\beta})\Big)$$

and this is usually more convenient than taking products.

Note The above also applies to discrete distributions where $p_i(y_i \mid \boldsymbol{\beta})$ is a pmf rather than a pdf.

The likelihood is a **relative** measure of consistency between the parameters $\boldsymbol{\beta}$ and the data $\boldsymbol{y}$.

- $l(\boldsymbol{\beta} \mid \boldsymbol{y})$ is defined up to an additive constant.
- Differences in log likelihood are always well defined.
- We may omit terms that are constant when deriving the log likelihood from the pdf/pmf. (NB: what is constant may depend on the context!)

Suppose $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ are two candidate values for the unknown parameter $\boldsymbol{\beta}$. If

$$l(\boldsymbol{\beta}^{(1)} \mid \boldsymbol{y}) \; - \; l(\boldsymbol{\beta}^{(2)} \mid \boldsymbol{y}) \quad > \quad 0,$$

then $\boldsymbol{\beta}^{(1)}$ has more **support** from the data $\boldsymbol{y}$.

(Under suitable regularity conditions) the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ satisfies

$$l(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{y}) - l(\boldsymbol{\beta} \mid \boldsymbol{y}) \quad > \quad 0 \qquad \forall \boldsymbol{\beta} \neq \widehat{\boldsymbol{\beta}},$$

so has the highest support from the data among all possible parameter values.

The **score function** $U : \mathbb{R}^p \to \mathbb{R}^p$ is the first derivative of the log likelihood:

$$U(\boldsymbol{\beta} \mid \boldsymbol{y}) \quad = \quad \frac{\partial l(\boldsymbol{\beta} \mid \boldsymbol{y})}{\partial \boldsymbol{\beta}}.$$

The maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ satisfies the **score equations**

$$U(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{y}) \quad = \quad \mathbf{0}.$$

Score equations generalize the normal equations for linear models.

The score function may be viewed as a random vector by replacing the observed data $y_1, \ldots, y_n$ with the corresponding random variables $Y_1, \ldots, Y_n$. This random vector has expectation zero:

$$\mathbb{E}\Big(U(\boldsymbol{\beta} \mid \boldsymbol{Y})\Big) \quad = \quad \mathbf{0}.$$

We say that the score function is an **unbiased estimating function.**

The variance of the score function is called the **Fisher (expected) information matrix**:

$$I(\boldsymbol{\beta}) \quad = \quad \mathbb{V}ar\Big(U(\boldsymbol{\beta} \mid \boldsymbol{Y})\Big) \quad = \quad \mathbb{E}\Big(U(\boldsymbol{\beta} \mid \boldsymbol{Y})\, U(\boldsymbol{\beta} \mid \boldsymbol{Y})^T\Big).$$

The Fisher information (or expected information) matrix $I(\boldsymbol{\beta})$ is positive semi-definite, that is

$$\boldsymbol{a}^T I(\boldsymbol{\beta}) \boldsymbol{a} \quad \geq \quad 0 \qquad \text{for any } \boldsymbol{a} \in \mathbb{R}^p.$$

It can be shown that

$$I(\boldsymbol{\beta}) \quad = \quad \mathbb{E}\Big(- \frac{\partial^2 l(\boldsymbol{\beta} \mid \boldsymbol{Y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\Big) \quad = \quad \mathbb{E}\Big(J(\boldsymbol{\beta} \mid \boldsymbol{Y})\Big),$$

where $J(\boldsymbol{\beta} \mid \boldsymbol{y})$ is the **observed information matrix** defined as the negative of the second derivative of the log likelihood, that is

$$J(\boldsymbol{\beta} \mid \boldsymbol{y}) \quad = \quad - \frac{\partial^2 l(\boldsymbol{\beta} \mid \boldsymbol{y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}.$$

Exercise 14 - Poisson maximum likelihood estimation

Suppose $y_1, \ldots, y_n$ is an iid sample from a Poisson distribution with mean μ.

a. Derive the likelihood function, score function and expectation of the score function.
b. Determine an expression for the observed information and for the Fisher information.

6.1.2 Properties of the maximum likelihood estimator

Suppose the number of parameters p is fixed as $n \to \infty$. Assuming suitable regularity conditions, the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ has the following properties:

1. **Consistency**

   As $n \to \infty$ we have $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

2. **Invariance under reparameterisation**

   Suppose $\boldsymbol{\gamma}$ is an alternative parameterization to $\boldsymbol{\beta}$.

   Then for some invertible function s we have

   $$\boldsymbol{\gamma} = s(\boldsymbol{\beta}) \qquad \text{and} \qquad \boldsymbol{\beta} = s^{-1}(\boldsymbol{\gamma}).$$

   The maximum likelihood estimates then satisfy

   $$\widehat{\boldsymbol{\gamma}} = s(\widehat{\boldsymbol{\beta}}) \qquad \text{and} \qquad \widehat{\boldsymbol{\beta}} = s^{-1}(\widehat{\boldsymbol{\gamma}}).$$

3. **Asymptotic unbiasedness**

   As $n \to \infty$,

   $$\sqrt{n}\Big(\mathbb{E}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta}\Big) \quad \to \quad \mathbf{0}.$$

4. **Asymptotic Efficiency (Generalization of Gauss-Markov theorem)**

   $\widehat{\boldsymbol{\beta}}$ is the unique asymptotically unbiased estimator with minimum variance.

5. **Asymptotic normality**

   For sufficiently large n we can use the approximation:

   $$\widehat{\boldsymbol{\beta}} \quad \sim \quad \mathcal{N}(\boldsymbol{\beta}, I(\boldsymbol{\beta})^{-1}).$$

For further details see Sections 4.4 - 4.9 in the recommended textbook by Dunn and Smyth.[1]

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

6.2 Maximum likelihood for GLMs

6.2.1 Recap

Note: to simplify notation we omit the explicit conditioning on $\boldsymbol{y}$ and $\boldsymbol{Y}$ but this is still assumed.

The maximum likelihood estimates $\widehat{\boldsymbol{\beta}}$ solve the score equations

$$U(\widehat{\boldsymbol{\beta}}) = \mathbf{0}.$$

We need an expression for the score function $U(\boldsymbol{\beta})$ for GLMs.

We also need an expression for the Fisher information matrix for GLMs, that is

$$I(\boldsymbol{\beta}) = \mathbb{E}\Big(- \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \Big)$$

so that we can calculate standard errors using the large sample approximation

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\Big(\boldsymbol{\beta}, I(\boldsymbol{\beta})^{-1}\Big).$$

6.2.2 Log likelihood for GLMs

Definition 6.1 (Generalized linear model)**.** A **generalized linear model** for outcomes $Y_1, \dots, Y_n$ and predictor variables $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ is defined by a combination of an exponential dispersion model and a link function

$$\begin{aligned} Y_i &\sim \text{EDM}(\mu_i, \phi/w_i) \\ g(\mu_i) &= \boldsymbol{x}_i^T \boldsymbol{\beta} \end{aligned}$$

where $\mathbb{E}(Y_i) = \mu_i = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta})$.

There is a common dispersion parameter ϕ which is modified by individual **prior weights** $(w_1, \dots, w_n)$.

Recall from Section 4.3 that Y_i as defined above has pdf/pmf given by

$$p(y_i|\theta_i, \phi) = a(y_i, \phi/w_i) \, \exp\Big(\frac{w_i \big[y_i\theta_i - b(\theta_i)\big]}{\phi}\Big).$$

Also recall that the mean parameter μ_i for observation i can be mapped onto the canonical parameter θ_i using the canonical link function. This allows us to write the log likelihood in canonical form:

$$\begin{aligned} l(\boldsymbol{\beta}, \phi) &= \sum_{i=1}^n l_i(\boldsymbol{\beta}, \phi) \\ &= \sum_{i=1}^n \Big[\frac{w_i \big[\theta_i y_i - b(\theta_i)\big]}{\phi} + \log\Big(a(y_i, \phi/w_i)\Big)\Big] \end{aligned}$$

where implicitly $\theta_i = \theta_i(\boldsymbol{\beta})$ is a function of $\boldsymbol{\beta}$. We can use this to derive the score function.

Lemma 6.1.

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{w_i\big[y_i - \mu_i\big]}{\phi} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}}.$$

Proof of Lemma 6.1

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} U_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial l_i(\theta_i)}{\partial \boldsymbol{\beta}}$$

$$= \sum_{i=1}^{n} \frac{w_i\big[y_i - \mu_i\big]}{\phi} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}}. \qquad \square$$

We consider two distinct cases:

- **Canonical link**. The score function and information matrix take a particularly simple form.
- **General link**. The score function is more complex but can be expressed in terms of μ_i and ϕ.

6.2.2.1 Overview on key results

If $g()$ is the canonical link function, then

$$\begin{aligned}
\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} &= \boldsymbol{x}_i, \\
U(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{w_i\big[y_i - \mu_i\big]}{\phi} \boldsymbol{x}_i, \\
I(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{w_i V(\mu_i)}{\phi} \boldsymbol{x}_i \boldsymbol{x}_i^T.
\end{aligned}$$

If $g()$ is a general link function, then

$$\begin{aligned}
\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} &= \frac{\boldsymbol{x}_i}{g'(\mu_i)V(\mu_i)}, \\
U(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{w_i\big[y_i - \mu_i\big]}{\phi} \frac{\boldsymbol{x}_i}{g'(\mu_i)V(\mu_i)} = \sum_{i=1}^{n} \left(\frac{W_i g'(\mu_i)}{\phi}\right)\big(y_i - \mu_i\big)\boldsymbol{x}_i, \\
I(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{w_i}{\phi[g'(\mu_i)]^2 V(\mu_i)} \boldsymbol{x}_i \boldsymbol{x}_i^T = \sum_{i=1}^{n} \left(\frac{W_i}{\phi}\right) \boldsymbol{x}_i \boldsymbol{x}_i^T,
\end{aligned}$$

where $W_i = \dfrac{w_i}{[g'(\mu_i)]^2 V(\mu_i)}$ is the ith working weight.

6.2.2.2 Canonical link

Proposition 6.1 (Score function for GLM with canonical link)**.** If $g()$ is the canonical link function, then

$$\begin{aligned} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} &= \boldsymbol{x}_i, \quad \text{and} \\ U(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{w_i\big[y_i - \mu_i\big]}{\phi} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{w_i\big[y_i - \mu_i\big]\boldsymbol{x}_i}{\phi}. \end{aligned}$$

Proof of Proposition 6.1

Recall that the canonical link maps the mean parameter onto the canonical parameter, that is $g(\mu) = \theta$. Hence

$$\begin{aligned} \theta_i = g(\mu_i) &= \boldsymbol{x}_i^T \boldsymbol{\beta}, \qquad \text{and so} \\ \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} &= \boldsymbol{x}_i. \end{aligned}$$

Therefore the score function is

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{w_i\big[y_i - \mu_i\big]}{\phi} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{w_i\big[y_i - \mu_i\big]}{\phi} \boldsymbol{x}_i. \qquad \square$$

Note:

- The maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ solves the score equation $U(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ which we can write as
$$\sum_{i=1}^{n} w_i\big[y_i - \mu_i(\widehat{\boldsymbol{\beta}})\big]\boldsymbol{x}_i = \mathbf{0}$$
independent of ϕ.
- Suppose the model has an intercept term, then the first column of the design matrix $\boldsymbol{X}$ is a column of ones ($x_{i1} = 1 \quad \forall i$). Writing $\widehat{\mu}_i = \mu_i(\widehat{\beta})$, the score equation for column 1 solves
$$\sum_{i=1}^{n} w_i\big[y_i - \widehat{\mu}_i\big] = 0.$$
This generalizes the result for linear models with an intercept term that the (weighted) sum of residuals is zero.

Proposition 6.2 (Fisher information for GLM with canonical link)**.**

$$I(\boldsymbol{\beta}) \quad = \quad \mathbb{E}\Big(-\sum_{i=1}^{n} \frac{\partial^2 l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \Big) \quad = \quad \sum_{i=1}^{n} \frac{w_i V(\mu_i) \boldsymbol{x}_i \boldsymbol{x}_i^T}{\phi}. \tag{6.1}$$

Proof of Proposition 6.2

The Fisher information $I(\boldsymbol{\beta}) = \mathbb{E}(J(\boldsymbol{\beta}))$, where $J(\boldsymbol{\beta})$ is the observed information. Using the fact that $\mu_i = b(\theta_i)$ we have

$$U(\boldsymbol{\beta}) \quad = \quad \sum_{i=1}^{n} \frac{w_i \Big[y_i - b'(\theta_i) \Big]}{\phi} \, \boldsymbol{x}_i.$$

Recall that $\frac{\partial \theta_i}{\partial \beta} = \boldsymbol{x}_i$. Then,

$$J(\boldsymbol{\beta}) \quad = \quad -\sum_{i=1}^{n} \frac{\partial^2 l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \quad = \quad -\sum_{i=1}^{n} \frac{\partial U_i}{\partial \boldsymbol{\beta}^T}$$

$$= \quad \sum_{i=1}^{n} \frac{w_i V(\mu_i) \boldsymbol{x}_i \boldsymbol{x}_i^T}{\phi}$$

This does not depend on $Y_1, \ldots, Y_n$, hence $I(\boldsymbol{\beta}) = J(\boldsymbol{\beta})$. $\square$

Exercise 15 - Score function for weighted normal linear model

Using the canonical form of the normal density, derive an expression for the score function and the Fisher information for a weighted normal linear model.

6.2.2.3 General link function

Define the **working weights** $W_1, \ldots, W_n$ as

$$W_i \quad = \quad \frac{w_i}{[g'(\mu_i)]^2 \, V(\mu_i)}.$$

The working weights should not be confused with the **prior weights** $w_1, \ldots, w_n$ defined by us when we fit the model.

Proposition 6.3 (Score function for GLM with a general link)**.** For a general link function $g()$ we have

$$\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \quad = \quad \frac{\boldsymbol{x}_i}{g'(\mu_i) V(\mu_i)} \qquad \text{and}$$

$$U(\boldsymbol{\beta}) \quad = \quad \sum_{i=1}^{n} \left(\frac{W_i g'(\mu_i)}{\phi} \right) \left(y_i - \mu_i \right) \boldsymbol{x}_i$$

Proof of Proposition 6.3 We have

$$\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \quad = \quad \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}.$$

As $\mu_i = b'(\theta_i)$ we have

$$\frac{\partial \mu_i}{\partial \theta_i} \quad = \quad b''(\theta_i) \quad = \quad V(\mu_i)$$

and so

$$\frac{\partial \theta_i}{\partial \mu_i} \quad = \quad \frac{1}{V(\mu_i)}.$$

Furthermore $g(\mu_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i$. Hence with the chain rule

$$g'(\mu_i) \, \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \quad = \quad \boldsymbol{x}_i$$

and so

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \quad = \quad \frac{\boldsymbol{x}_i}{g'(\mu_i)}.$$

Therefore it follows that

$$\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \quad = \quad \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \quad = \quad \frac{1}{V(\mu_i)} \times \frac{\boldsymbol{x}_i}{g'(\mu_i)}.$$

Next we derive the expression for the score function.

$$
\begin{aligned}
U(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{w_i\big[y_i - \mu_i\big]}{\phi} \, \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{n} \frac{w_i\big[y_i - \mu_i\big]}{\phi} \, \frac{1}{V(\mu_i)} \, \frac{\boldsymbol{x}_i}{g'(\mu_i)} \\
&= \sum_{i=1}^{n} \left(\frac{W_i g'(\mu_i)}{\phi}\right)\Big(y_i - \mu_i\Big)\boldsymbol{x}_i.
\end{aligned}
$$

where we used the working weights defined earlier as

$$
W_i = \frac{w_i}{[g'(\mu_i)]^2 \, V(\mu_i)}.
$$

Proposition 6.4 (Fisher information for a GLM with a general link)**.** For a general link function $g()$ the Fisher information matrix is

$$
I(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{W_i}{\phi}\right)\boldsymbol{x}_i \boldsymbol{x}_i^T.
$$

Proof of Proposition 6.4: Omitted.

6.3 Numerical maximum likelihood estimation for GLMs

6.3.1 The Newton-Raphson algorithm

Under suitable regularity conditions we can find the maximum likelihood estimate $\widehat{\gamma}$ of a parameter γ by solving $U(\gamma) = 0$. With an initial guess $\widetilde{\gamma}$, a first order Taylor expansion of the score function U gives gives

$$\begin{aligned} U(\widehat{\gamma}) &\approx U(\widetilde{\gamma}) + U'(\widetilde{\gamma})\Big[\widehat{\gamma} - \widetilde{\gamma}\Big] \\ &= U(\widetilde{\gamma}) - J(\widetilde{\gamma})\Big[\widehat{\gamma} - \widetilde{\gamma}\Big] \end{aligned}$$

as $U'(\widetilde{\gamma}) = -J(\widetilde{\gamma})$ where $J()$ is the observed information.

Now, using the fact that $U(\widehat{\gamma}) = 0$, we have

$$\widehat{\gamma} \approx \widetilde{\gamma} + \Big[J(\widetilde{\gamma})\Big]^{-1} U(\widetilde{\gamma}).$$

To determine $\widehat{\gamma}$ we can now compute an iterative sequence of approximations

$$\gamma^{(k+1)} \approx \gamma^{(k)} + \Big[J(\gamma^{(k)})\Big]^{-1} U(\gamma^{(k)}), \qquad k = 0, 1, 2, \ldots,$$

until convergence is reached. This is the so-called **Newton-Raphson algorithm.**

If we approximate the observed information by the Fisher information and so compute the sequence

$$\gamma^{(k+1)} \approx \gamma^{(k)} + \Big[I(\gamma^{(k)})\Big]^{-1} U(\gamma^{(k)}), \qquad k = 0, 1, 2, \ldots,$$

then this algorithm is referred to as **Fisher scoring.**

For GLMs, rather than inverting the Fisher information, we take a slightly different approach, namely the so-called **IWLS (iterated weighted least squares)** algorithm.

6.3.2 The IWLS algorithm for GLMs

Maximum likelihood (or quasi-likelihood) estimates for GLMs are obtained from the IWLS algorithm.

Take a local linear approximation to reduce a GLM to a linear model:

1. Start with an initial estimate $\tilde{\boldsymbol{\beta}}$.
2. Take a linear approximation for $\boldsymbol{\beta}$ "close" to $\tilde{\boldsymbol{\beta}}$:
   2.1 Approximate the likelihood using a weighted linear model.
   2.2 Obtain new estimate of $\boldsymbol{\beta}$ from this linear model.
3. Repeat step 2 until convergence.

Let $\tilde{\boldsymbol{\beta}}$ be our current estimate of $\boldsymbol{\beta}$. Using first oder Taylor approximation we can approximate the score function in the neighbourhood of $\tilde{\boldsymbol{\beta}}$:

$$\begin{aligned} U(\boldsymbol{\beta}) &\approx U(\tilde{\boldsymbol{\beta}}) + \frac{\partial U(\tilde{\boldsymbol{\beta}})}{\partial \beta^T}\left[\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right] \\ &= U(\tilde{\boldsymbol{\beta}}) - J(\tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \\ &\approx U(\tilde{\boldsymbol{\beta}}) - I(\tilde{\boldsymbol{\beta}})\left[\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right]. \end{aligned}$$

We previously derived that

$$\begin{aligned} U(\boldsymbol{\beta}) &= \frac{1}{\phi}\sum_{i=1}^{n} W_i g'(\mu_i)\Big(y_i - \mu_i\Big)\boldsymbol{x}_i \\ I(\boldsymbol{\beta}) &= \frac{1}{\phi}\sum_{i=1}^{n} W_i \boldsymbol{x}_i \boldsymbol{x}_i^T, \\ \text{where} \quad W_i &= \frac{w_i}{[g'(\mu_i)]^2\, V(\mu_i)}. \end{aligned}$$

Hence

$$\begin{aligned} U(\boldsymbol{\beta}) &\approx U(\tilde{\boldsymbol{\beta}}) - I(\tilde{\boldsymbol{\beta}})\left[\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right] \\ &= \sum_{i=1}^{n} \frac{\widetilde{W_i}}{\phi} g'(\widetilde{\mu_i})\Big(y_i - \widetilde{\mu_i}\Big)\boldsymbol{x}_i \; - \; \sum_{i=1}^{n} \frac{\widetilde{W_i}}{\phi}\boldsymbol{x}_i\boldsymbol{x}_i^T\left[\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right] \\ &= \sum_{i=1}^{n} \frac{\widetilde{W_i}}{\phi}\left[g'(\widetilde{\mu_i})\Big(y_i - \widetilde{\mu_i}\Big) + \boldsymbol{x}_i^T\widetilde{\boldsymbol{\beta}} - \boldsymbol{x}_i^T\boldsymbol{\beta}\right]\boldsymbol{x}_i \\ &= \sum_{i=1}^{n} \frac{\widetilde{W_i}}{\phi}\left[\widetilde{z}_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right]\boldsymbol{x}_i \end{aligned}$$

where $\widetilde{z}_i = g'(\widetilde{\mu_i})\Big(y_i - \widetilde{\mu_i}\Big) + \boldsymbol{x}_i^T\widetilde{\boldsymbol{\beta}}$ is the **working observation** for observation i.

We solve the approximate score equations

$$U(\boldsymbol{\beta}) \approx \sum_{i=1}^{n} \frac{\widetilde{W_i}}{\phi}\left[\widetilde{z}_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right]\boldsymbol{x}_i \quad = \quad \mathbf{0}.$$

This is equivalent to estimating $\boldsymbol{\beta}$ for a linear model on the working observations, that is

$$\widetilde{Z_i} \quad \sim \quad \mathcal{N}\left(\boldsymbol{x}_i^T\boldsymbol{\beta},\, \frac{\phi}{\widetilde{W_i}}\right).$$

The maximum likelihood estimate for this weighted linear model solves the approximate score equation and is given by

$$\widetilde{\boldsymbol{\beta}}^* \quad = \quad \left(\boldsymbol{X}^T\widetilde{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\widetilde{\boldsymbol{W}}\widetilde{z}_i$$

where $\widetilde{\boldsymbol{W}} = \text{diag}(\widetilde{W_1}, \ldots, \widetilde{W_n})$.

This new estimate $\widetilde{\boldsymbol{\beta}}^*$ becomes our new value of $\widetilde{\boldsymbol{\beta}}$ for the next iteration.

We continue until the new estimate is the same as the previous one.

Then $\widetilde{\boldsymbol{\beta}}$ solves the approximate score equations:

$$\sum_{i=1}^{n} \frac{\widetilde{W_i}}{\phi} \Big[\widetilde{z}_i - \boldsymbol{x}_i^T \widetilde{\boldsymbol{\beta}} \Big] \boldsymbol{x}_i \quad = \quad \mathbf{0}.$$

But then

$$\begin{aligned} \widetilde{z}_i - \boldsymbol{x}_i^T \widetilde{\boldsymbol{\beta}} \quad &= \quad g'(\widetilde{\mu_i})\Big(y_i - \widetilde{\mu_i}\Big) + \boldsymbol{x}_i^T \widetilde{\boldsymbol{\beta}} - \boldsymbol{x}_i^T \widetilde{\boldsymbol{\beta}} \\ &= \quad g'(\widetilde{\mu_i})\Big(y_i - \widetilde{\mu_i}\Big). \end{aligned}$$

and so $\widetilde{\boldsymbol{\beta}}$ also solves the exact score equations:

$$U(\widetilde{\boldsymbol{\beta}}) \quad = \quad \frac{1}{\phi} \sum_{i=1}^{n} \widetilde{W_i}\ g'(\widetilde{\mu_i})\ \Big(y_i - \widetilde{\mu_i}\Big)\ \boldsymbol{x}_i \quad = \quad \mathbf{0}.$$

Therefore $\widetilde{\boldsymbol{\beta}}$ is equal to the maximum likelihood estimate.

The information matrix from our approximate linear model is equal to

$$\frac{1}{\phi}\ \sum_{i=1}^{n} \widetilde{W_i} \boldsymbol{x}_i \boldsymbol{x}_i^T$$

and thus exactly equal to $I(\widetilde{\beta})$, the Fisher information matrix for our GLM.

Hence we can use the asymptotic result

$$\widehat{\beta} \quad \sim \quad \mathcal{N}\Big(\boldsymbol{\beta},\ \phi\big(\sum_{i=1}^{n} \widetilde{W_i} \boldsymbol{x}_i \boldsymbol{x}_i^T\big)^{-1}\Big).$$

Exercise 16 - IWLS algorithm

Show that for a normal linear model, the IWLS algorithm converges to the maximum likelihood estimate in one iteration, whatever starting value we use for $\boldsymbol{\beta}$.

6.3.3 Convergence in practice

In practice we stop the IWLS algorithm after a finite number of iterations.

- The convergence criterion is based on relative changes in the deviance.
- By maximizing the log likelihood over $\boldsymbol{\beta}$ for fixed ϕ we are **minimizing** the deviance.
- When relative changes to the deviance are sufficiently small, then we have converged.

The `glm` function in `R` derives starting values from the data.

We only need an initial estimate $\widetilde{\mu}_i^{(0)}$ for μ_i, for example

- normal, gamma, inverse Gaussian: $\widetilde{\mu}_i^{(0)} = y_i$
- Poisson: $\widetilde{\mu}_i^{(0)} = y_i + 0.1$
- Scaled binomial: $\widetilde{\mu}_i^{(0)} = \frac{(y_i m_i + 0.5)}{(m_i+1)}$

and then our initial working observations can be derived from

$$\widetilde{z}_i^{(0)} \quad = \quad g'(\widetilde{\mu}_i^{(0)})\Big(y_i - \widetilde{\mu}_i^{(0)}\Big) + g(\widetilde{\mu}_i^{(0)}).$$

This only works for certain link functions.

Exercise 17 - Computer Practical 3

Work through Computer Practical 3.

6.4 Estimating the dispersion parameter

6.4.1 Overview

- The maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ does not depend on ϕ.
- We estimate $\boldsymbol{\beta}$ first and then estimate ϕ in a second step.
- The IWLS algorithm gives us an estimator for ϕ based on the linear model.
- This estimator reduces to an intuitively clear form based on the sum of squares of the residuals.

6.4.2 Derivation

The last iteration of the IWLS algorithm is based on a linear approximation

$$\widehat{Z}_i \quad \sim \quad \mathcal{N}\Big(\boldsymbol{x}_i^T\boldsymbol{\beta},\ \frac{\phi}{\widehat{W}_i}\Big)$$

where

$$\widehat{z}_i \quad = \quad g'(\widehat{\mu}_i)\Big(y_i - \widehat{\mu}_i\Big) + g(\widehat{\mu}_i)$$

with

$$g(\widehat{\mu}_i) \quad = \quad \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}$$

and

$$\widehat{W}_i \quad = \quad \frac{w_i}{V(\widehat{\mu}_i)\Big[g'(\widehat{\mu}_i)\Big]^2}.$$

Under the normal weighted linear model approximation the estimator of ϕ is given as

$$\widehat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \widehat{W}_i \left(\widehat{z}_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} \right)^2.$$

Note the denominator is $n-p$ as we lose p degrees of freedom from estimating the p parameters.

Substituting expressions for $\widehat{W}_i$ and $\widehat{z}_i$:

$$\begin{aligned}
\widehat{\phi} &= \frac{1}{n-p} \sum_{i=1}^{n} \widehat{W}_i \left(\widehat{z}_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} \right)^2 \\
&= \frac{1}{n-p} \sum_{i=1}^{n} \frac{w_i}{V(\widehat{\mu}_i) \left[g'(\widehat{\mu}_i) \right]^2} \left(g'(\widehat{\mu}_i)\left(y_i - \widehat{\mu}_i \right) + \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} \right)^2 \\
&= \frac{1}{n-p} \sum_{i=1}^{n} \frac{w_i}{V(\widehat{\mu}_i) \left[g'(\widehat{\mu}_i) \right]^2} \left[g'(\widehat{\mu}_i) \right]^2 \left(y_i - \widehat{\mu}_i \right)^2 \\
&= \frac{1}{n-p} \sum_{i=1}^{n} \frac{w_i}{V(\widehat{\mu}_i)} \left(y_i - \widehat{\mu}_i \right)^2
\end{aligned}$$

The **Pearson residual** for observation i is

$$r_i^{(p)} = \sqrt{\frac{w_i}{V(\widehat{\mu}_i)}} \, (y_i - \widehat{\mu}_i).$$

We can thus write the estimator of ϕ as

$$\widehat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \left[r_i^{(p)} \right]^2.$$

In `R` we can get the Pearson residuals from a fitted GLM with the command

```
residuals(glm.out, type="pearson")
```

We will see later that there are many possible types of residual for a GLM.

The Pearson estimator for ϕ is the one used by the `summary()` function in `R`.

6.4.3 Other estimators for the dispersion parameter

Other possible estimators discussed by Dunn & Smyth in Section 6.8 of the recommended textbook.[2]

- **Modified profile likelihood.** Optimal estimator but requires stronger assumptions about the distribution of Y and usually requires numerical maximization.
- **Mean deviance.** Not suitable for Poisson or binomial models with small counts ($y < 3$ or, for binomial, $m - y < 3$.)

For normal linear models, all three estimators of ϕ are the same.

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

6.4.4 Example: cherry tree data

We illustrate the estimation of the dispersion parameter with the `trees` example from the `datasets` package. (You will recall the dataset from ST231!)

- We want to predict the volume (V) of wood from the height of the tree (H) and its diameter (G).
- Suppose the tree is a cylinder, then

$$\begin{aligned} V &= \pi H\Big(G/2\Big)^2 \\ \log(V) &= \log(\pi/4) + 2\log(G) + \log(H) \end{aligned}$$

  This suggests the following model for $\mu = \mathbb{E}(V)$:

$$\log(V) = \beta_0 + \beta_1 \log(G) + \beta_2 \log(H)$$

  As V is positive and real-valued we use the gamma EDM:

$$V \sim \Gamma(\mu, \phi).$$