第三章 指数族分布的统计推断

Parameter Estimation

- In the GLM model we are interesting in estimating β, the parameters in the linear predictor term.
- Consider maximum likelihood.
- The likelihood function is

$$
\begin{aligned}
L(\beta) &= \prod_{i=1}^{n} f(y_i|\theta_i, \phi) \\
&= \prod_{i=1}^{n} exp(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)) \qquad (1) \\
&= exp(\frac{\sum_{i=1}^{n}(y_i\theta_i - b(\theta_i))}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi)).
\end{aligned}
$$

Parameter Estimation

- Hence the log-likelihood is

$$
\begin{aligned}
l(\beta) &= \frac{1}{a(\phi)} \sum_{i=1}^{n} (y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(y_i, \phi) \\
&= \sum_{i=1}^{n} l_i(\beta),
\end{aligned}
\tag{2}
$$

say.

The score functions

- Taking derivatives of the log likelihood with respect to β_j, $j = 1, ..., p$ we obtain the score functions

$$U_j(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j}, \quad j = 1, ..., p.$$

- Remember that

$$\mu_i = b^{'}(\theta_i). \qquad \eta_i = g(\mu_i) = \mathbf{x_i}^T \boldsymbol{\beta}.$$

- Using the chain rule we have

$$\begin{aligned} U_j(\boldsymbol{\beta}) &= \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta})}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}. \end{aligned} \tag{3}$$

The derivatives

- The derivatives are:

$$\frac{\partial l_i(\boldsymbol{\beta})}{\partial \theta_i} = \frac{1}{a(\phi)}(y_i - b^{'}(\theta_i)) = \frac{y_i - \mu_i}{a(\phi)};$$

$$\begin{aligned} \frac{\partial \theta_i}{\partial \mu_i} = (\frac{\partial \mu_i}{\partial \theta_i})^{-1} &= (\frac{\partial b^{'}(\theta_i)}{\partial \theta_i})^{-1} \\ &= (b^{''}(\theta_i))^{-1} \\ &= (V(\mu_i))^{-1}; \end{aligned} \tag{4}$$

$$\begin{aligned} \frac{\partial \mu_i}{\partial \eta_i} = (\frac{\partial \eta_i}{\partial \mu_i})^{-1} &= (\frac{\partial g(\mu_i)}{\partial \mu_i})^{-1} \\ &\equiv (g^{'}(\mu_i))^{-1}; \end{aligned} \tag{5}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mathbf{x_i}^T \boldsymbol{\beta}}{\partial \beta_j} = x_{ij},$$

where x_{ij} is the jth element of $\mathbf{x_i}$, or equivalently the (i,j) element of matrix $\mathbf{X}$.

The score equations

- To solve for β we need to solve the score equations,that is,

$$U_j(\beta) = 0, \quad j = 1, ..., p.$$

- In our case we solve

$$\begin{aligned} U_j(\beta) &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)} (V(\mu_i))^{-1} (g^{'}(\mu_i))^{-1} x_{ij} \\ &= \sum_{i=1}^{n} \frac{1}{V(\mu_i) a(\phi) (g^{'}(\mu_i))^2} x_{ij} (y_i - \mu_i) g^{'}(\mu_i) \\ &= 0, \end{aligned} \tag{6}$$

for $j = 1, ..., p$.

Using adjusted dependent variables

- Let

$$\omega_i = \frac{1}{V(\mu_i)a(\phi)(g'(\mu_i))^2}.$$

- Then the score equations become

$$U_j(\beta) = \sum_{i=1}^{n} \omega_i x_{ij}(y_i - \mu_i)g'(\mu_i) = 0,$$

for $j = 1, ..., p$.

- Also define

$$z_i = \eta_i + (y_i - \mu_i)g'(\mu_i).$$

- Thus

$$U_j(\beta) = \sum_{i=1}^{n} \omega_i x_{ij}(z_i - \eta_i) = 0,$$

for $j = 1, ..., p$.

Matrix notation

- Let $\mathbf{W} = diag(\omega_1, ..., \omega_n)$, $\mathbf{X}$ denote the design matrix, $\mathbf{z} = (z_1, ..., z_n)^T$ and $\boldsymbol{\eta} = (z_1, ..., z_n)^T$. Also let $\mathbf{U}(\beta) = (U_1(\beta), ..., U_n(\beta))^T$. Then

$$\mathbf{U}(\beta) = \mathbf{X}^T\mathbf{W}(\mathbf{z} - \boldsymbol{\eta}) = 0.$$

Since $\boldsymbol{\eta} = \mathbf{X}\beta$,we have that

$$\mathbf{X}^T\mathbf{W}\mathbf{z} = \mathbf{X}^T\mathbf{W}\mathbf{X}\beta.$$

This is the weighted least squares (WLS) problem. When $\mathbf{X}$ and $\mathbf{W}$ are known, the estimate of β which solves the normal equations are

$$\hat{\beta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}.$$

Problem:$\mathbf{W}$ and $\mathbf{z}$ depend on β!

Iteratively weighted least squares(IWLS)

- Solution: we iterate!
- Start with a guess for η:

$$\eta^{(0)} = g(\mathbf{y}).$$

  (we may need to adjust this slightly in practice).
- Iteratively calculate the following for $j = 1, 2, \ldots$
  1. $\mu = h(\eta^{(j-1)})$ (where we let $h(\cdot)$ denote the inverse link function).
  2. $\mathbf{W} = diag([V(\mu)a(\phi)(g^{'}(\mu))^2]^{-1})$.
  3. $\mathbf{z} = \eta + (\mathbf{y} - \mu)g^{'}(\mu)$.
  4. $\beta = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}$.
  5. $\eta^{(j)} = \mathbf{X}\beta$.
- We stop iterating when $\beta^{(j)} - \beta^{(j-1)}$ is "small" (equivalently we can look at changes in η).

Remarks on parameter estimation

- We could also estimate β using a Fisher scoring or Newton-Raphson scheme.

  -**Fisher scoring** is

$$\beta^{(j)} = \beta^{(j-1)} + [\mathbf{I}(\beta^{(j-1)})^{-1}]\mathbf{U}(\beta^{(j-1)}),$$

  where **Fisher's information matrix** is

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

  This can be shown to be equivalent to IWLS.

  -The **Newton-Raphson algorithm** is

$$\beta^{(j)} = \beta^{(j-1)} + [\mathbf{i}(\beta^{(j-1)})^{-1}]\mathbf{U}(\beta^{(j-1)}),$$

  where the **observed information matrix** is

$$\mathbf{i}(\beta) = -(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}).$$

  This is equivalent to IWLS, Fisher scoring for canonical links only.

The second derivative

- We have already shown that

$$\frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_j} = \omega_i g'(\mu_i) x_{ij}(y_i - \mu_i).$$

- Now

$$\begin{aligned}
\frac{\partial l_i^2(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} &= \frac{\partial}{\partial \beta_k}[\omega_i g'(\mu_i) x_{ij}(y_i - \mu_i)] \\
&= [\frac{\partial}{\partial \beta_k}\omega_i g'(\mu_i)] x_{ij}(y_i - \mu_i) + \omega_i g'(\mu_i) x_{ij}[\frac{\partial}{\partial \beta_k}(y_i - \mu_i)].
\end{aligned} \tag{7}$$

- Start by ignoring the first term!
- In the second term, we know that

$$\frac{\partial \mu_i}{\partial \beta_k} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_k} = (g'(\mu_i))^{-1} x_{ik},$$

and so

$$\frac{\partial l_i^2(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} = [\frac{\partial}{\partial \beta_k}\omega_i g'(\mu_i)] x_{ij}(y_i - \mu_i) - \omega_i g'(\mu_i) x_{ij}(g'(\mu_i))^{-1} x_{ik}.$$

The Fisher information matrix

- Simplifying we have
$$\frac{\partial l_i^2(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} = [\frac{\partial}{\partial \beta_k} \omega_i g^{'}(\mu_i)] x_{ij}(y_i - \mu_i) - \omega_i x_{ij} x_{ik}.$$
- Since $E(Y_i) = \mu_i$, the Fisher information for observation i is
$$-E(\frac{\partial l_i^2(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}) = 0 + \omega_i x_{ij} x_{ik}.$$
- The Fisher information for the whole sample is
$$-E(\frac{\partial l^2(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}) = \sum_{i=1}^n \omega_i x_{ij} x_{ik}.$$
- In the matrix notation already defined,
$$\mathbf{I}(\beta) = \mathbf{X}^T \mathbf{W} \mathbf{X}.$$

Estimating the Fisher information matrix

- $\mathbf{X}$ is known in practice, but, $\mathbf{W}$ is a diagonal matrix with entries

$$\omega_i = \frac{1}{V(\mu_i)a(\phi)(g'(\mu_i))^2}.$$

- We can estimate μ_i using

$$\mu_i = h(\eta_i) = h(\mathbf{x}_i^T \hat{\beta}),$$

- where we let $h(\cdot)$ denote the inverse of the link function, $g(\cdot)$, and $\hat{\beta}$ denote the resulting estimator of β obtained from the IWLS.
- In certain cases we do not need to estimate ϕ.

-Why?

按照上述迭代模拟过程（二项分布）

```
#生成x1,x2
m=2000
x1=rnorm(m)
x2=rnorm(m)
eta=rep(1,m)+x1+x2
mu=exp(eta)/(1+exp(eta))
#生成y
y=NULL
for(i in 1:m){
  y[i]=rbinom(1,1,mu[i])
}
# summary(y)
# table(y)

X=model.matrix(~x1+x2)#设计矩阵
```

```
############迭代
beta=c(0,0,0)
eta=X%*%beta

n=0
beta0=c()
repeat{
  beta_old=beta
  mu=exp(eta)/(1+exp(eta))
  w=diag(as.vector(mu*(1-mu)))
  z=eta+(y-mu)/(mu*(1-mu))
  beta=solve((t(X)%*%w%*%X))%*%t(X)%*%w%*%z
  #print(beta)#输出每一次迭代得到的beta
  eta=X%*%beta
  D=max(abs(beta-beta_old))
  n=n+1
  if(D<1e-8)
    break
}
n#迭代次数
beta

data=as.data.frame(cbind(y,x1,x2))
y.glm=glm(y~x1+x2,family=binomial(link="logit"),data=data)
summary(y.glm)
```

Statistical inference on β.

- We obtain the estimate of β, $\hat{\beta}$, from IWLS.
- Asymptotically, for large sample sizes we will have

$$\hat{\beta} \xrightarrow{d} N_p(\beta, \mathbf{I}(\beta)^{-1}).$$

- We also estimate $\mathbf{I}(\beta)$ from the data (see previous slide).
- As with the linear model we can now write down a table of the estimated coefficients.

Goodness of fit for GLMs

- In linear models we often assess the goodness of fit by looking at sums of squares,e.g.,RSS
- How do we assess the fit for GLMs?
- We measure the goodness of fit of a GLM using the **deviance** and **Pearson χ^2 statistic**
- Both statistics look at how the data, **y**, estimated from the GLM.
- The deviance compares log likelihoods, whereas the Pearson statistic is a sum of squares (with an appropriate scaling for the mean-variance relationship).

The deviance

- The **deviance** compares the fit of the **full** model (when we n parameters,on for each observation) to the fit of the **reduced** model (when we fit p parameters). It compares log likelihoods.
- Remember the log likelihood is

$$l(\beta) = \frac{\sum_{i=1}^{n}(y_i\theta_i - b(\theta_i))}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi)$$

- We will evaluate this log likelihood (in terms of θ_i or $\mu_i = b^{'}_{\theta_i}$) for the full and reduced model.

The deviance(cont.)

- Let $\widetilde{\theta}_i$ denote the estimate of the canonical parameter, θ_i, when we fit the full model, that is, when we estimate μ_i using $\widetilde{\mu} = y_i$.
- Let $\hat{\theta}$ be the estimate of θ_i in the reduced model, and μ_i denote the associated estimate of μ_i
- Then the **deviance** is defined to be

$$
\begin{aligned}
D(y, \hat{\mu}) &= 2a(\phi)[l(\tilde{\theta}) - l(\hat{\theta})] \\
&= 2a(\phi)\{\frac{\sum_{i=1}^{n}[y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i))]}{a(\phi)}\} \\
&+ a(\phi)\sum_{i=1}^{n}[c(y_i, \phi) - c(y_i, \phi)] \\
&= 2\sum_{i=1}^{n}[y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i))]
\end{aligned}
\tag{8}
$$

Example:Normal

- For normally distributed data we have

$$a(\phi) = \sigma^2$$

$$b(\theta_i) = \frac{\theta_i^2}{2}$$

$$\theta_i = \mu_i$$

- In the full model we have $\widetilde{\theta_i} = \widetilde{\mu_i} = y_i$,and in the reduced model $\widehat{\theta_i} = \widehat{\mu_i}$.
- The deviance is

$$\begin{aligned} D(y, \hat{\mu}) &= 2\sum_{i=1}^{n}[y_i(y_i - \widehat{\mu_i}) - \frac{y_i^2}{2} + \frac{\widehat{\mu_i}^2}{2}] \\ &= \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2. \end{aligned} \tag{9}$$

Example:Binomial

- For data following a binomial distribution we have.

Example:Possion

- For Possion distributed data we have

$$a(\phi) = 1$$

$$b(\theta_i) = e^{\theta_i}$$

$$\theta_i = log\,\mu_i$$

- In the full model we have $\hat{\theta}_i = log\,\tilde{\mu}_i = log y_i$,and in the reduced model $\hat{\theta}_i = log\,\hat{\mu}_i$.
- The deviance is

$$\begin{aligned} D(y,\hat{\mu}) &= 2\sum_{i=1}^{n}[y_i(log y_i - log\,\widehat{\mu_i}) - (y_i - \hat{\mu}_i)] \\ &= 2\sum_{i=1}^{n}[y_i log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]. \end{aligned} \tag{10}$$

Using deviance

- Note that the deviance dose **not depend** on ϕ.
- Suppose we fit two GLMs,with one model being a subset of the other.We can compare the fit of these two models by calculating **the difference of the deviances**.This is equivalent to a likelihood ratio test(see next page).
- Similar to the **analysis of variance** table we produce for linear models we can create an **analysis of deviance** table for GLMs
  1. Except for the normal cases,the **distribution** of differences in the deviance are **asymptotic**.
  2. Thus,we need to interpret the analysis of deviance table carefully.

Likelihood ratios tests(LRTs)

- Consider a partition of the p dim coefficient vector, β

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

where β_1 is a q dim vector,andβ_2 is a p-q dim vector.

- Suppose we want to test $H_0 : \beta_2 = \beta_2^0$.
- Then the likelihood ratio statistic,

$$\lambda(\hat{\beta}_2) = 2[l(\hat{\beta}_1, \hat{\beta}_2) - l(\hat{\beta}_1, \beta_2^0)]$$

is asymptotically χ^2_{p-q}.

Pearsonχ^2statistic

- As before,suppose that we have estimated $\hat{\mu}_i$ from the data.
- Then the **Pearson** χ^2 **statistic** is defined by

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

- Some examples:
  1. Normal:$\chi^2 = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$,the RSS.
  2. Binomial: χ^2
  3. Poisson:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

Residual plots for GLMs

Similar to linear models we want to:

- Plot residuals versus fitted:
  - Check for appropriateness of the fit.
  - Do we need to transform the response?
  - Check for constancy of the variance of errors.
  - Look for outliers.
- Residual versus time or collection order.
  - Check for systematic problems in the residuals(e.g.,serial correlation).
- $Q-Q$ plot of residuals.
  - Check distributional assumptions of the errors.

Residual plots for GLMs(cont.)

- Plot residuals versus each predictor/covariate in the model.
  - Check adequacy of the fit for each predictor.
  - Curvature may indicate the need to transform predictors.
- Plot residual versus potential predictors NOT in the model.
  - Have any variables been omitted from the model?

But,what do we use for residuals in a GLM?

- Residuals for GLMs are harder to interpret.
- Checking the adequacy of variance and link functions are important for GLMs.

The deviance and Pearson residuals

- We showed that the deviances are given by

$$
\begin{aligned}
D(y,\hat{\mu}) &= 2\sum_{i=1}^{n}[y_i(\tilde{\theta}_i-\hat{\theta}_i)-(b(\tilde{\theta}_i)-b(\hat{\theta}_i))] \\
&= \sum_{i=1}^{n} D_i .
\end{aligned} \tag{11}
$$

- The **deviance residuals** are defined by

$$
(r_D)_i = sign(y_i-\hat{\mu}_i)\sqrt{D_i}
$$

- The **Pearson residuals** are defined by

$$
(r_P)_i = \frac{y_i-\hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}
$$

- Thus $\chi^2 = \sum^{n} (r_P)^2$

The working and Anscomebe residuals

- We obtain the **working residuals** from the **IWLS algorithm**:

$$\begin{aligned}(r_W)_i &= z_i - \hat{\eta}_i \\ &= [\hat{\eta}_i + (y_i - \hat{\mu}_i)g^{'}(\hat{\mu}_i)] - \hat{\eta}_i \\ &= y_i - \hat{\mu}_i)g^{'}(\hat{\mu}_i\end{aligned} \tag{12}$$

- For the **Anscombe residuals**, see McCullagh and Nelder, page 38.
  – The general idea:**transform** the Pearson residual to be **less skewed**.