Contents

§2 Exponential family **47**
§2.1 Generalized Linear Models (GLM) 47
§2.2 Exponential Family 49
§2.3 Weighted Least Squares (WLS) 54
§2.4 Iterative reweighted LS (IRLS) or M-estimation 57
§2.5 Quantile regression 60
§2.5.1 Asymmetric loss function 60
§2.5.2 Parameter estimation 61
§2.5.3 Bayesian approach 65
§2.6 ML estimate for GLM 67
§2.7 Quasi-Likelihood 73
§2.8 Generalized linear mixed model (GLMM) 79
§2.8.1 Random effects model 79
§2.8.2 Generalized random effect model 81
§2.9 Appendix 88

§2 Exponential family

§2.1 Generalized Linear Models (GLM)

Generalized linear model (GLM) is a useful generalization and extension of linear regression models to accommodate both non-normal data distributions and transformations to linearity in a straightforward way.

Suppose that the *outcomes* $Y_i,\ i = 1, \ldots, n$ are realisations of a random variable which is observed *independently*. Together with values of $p < n$ *explanatory variables* $x_{i1}, ..., x_{ip}$, a GLM contains the following components and assumptions:

1. **Random component:** The distribution $f(\cdot)$ of Y_i condition on x_{ij} is a member of the *exponential family* (EF) with mean μ_i and scale ϕ.

2. **Systematic component:** the explanatory variables affect the mean of Y_i through a *linear function of predictors*:

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}.$$

   or in matrix form,

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$$

   where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is a $n \times p$ *design* matrix and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is a p-dimensional vector of regression coefficients.

3. **Link function:** the linear function of predictors η_i is related to the mean μ_i of the distribution with the probability density function (pdf) function $f(\cdot)$ through a smooth invertible function called the *link* function $g(\cdot)$ such that $g(\mu_i) = \eta_i$ or in matrix form,

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} \ \Rightarrow g^{-1}(\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\mu} = E(\boldsymbol{Y})$$

where $\boldsymbol{X}\boldsymbol{\beta}$ is a n-dimensional vector of the linear functions of predictors η_i. For example, the canoncial link function is *log* for Poisson and *logit* for binomial.

§2.2 Exponential Family

Nelder and Wedderburn (1972, JRSSA 135, 221-32) stated that many common statistical models including regression for continuous data, logistic model for binary data, and log-linear models for count data were members of a general model, the exponential family (EF). They showed that the parameters could be estimated via a simple algorithm that produced the maximum likelihood estimates. The probability density function in the exponential family can be expressed as

$$f(y) = \exp\left[\boldsymbol{T}(y) \cdot \boldsymbol{\eta}(\boldsymbol{\theta}) - B(\boldsymbol{\theta}) + C(y)\right], \tag{1}$$

where
$\boldsymbol{\eta}(\boldsymbol{\theta})$ is the natural parameter,
$B(\boldsymbol{\theta})$ is the normalization factor and
$\boldsymbol{T}(y)$ is the sufficient statistic.

In general, exponential families cannot have a support that varies according to a parameter; rather, the support must remain the same across all distributions in the family. If $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, then the exponential family is said to be in *canonical* form. By defining a transformed parameter $\boldsymbol{\theta}' = \boldsymbol{\eta}(\boldsymbol{\theta})$ or simply $\boldsymbol{\theta}$, it is always possible to convert an exponential family to canonical form. Furthermore, if $\boldsymbol{\theta} = \theta$, (1) can be written as

$$f(y) = \exp\left\{\frac{T(y)\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$$

where θ and $\phi > 0$ are *location* and *scale* parameters respectively and $a(\phi)$, and $c(y, \phi)$ are known functions. In all models considered the function $a(\phi)$ has the form

$$a(\phi) = \phi/w,$$

where w is a known *prior weight*, usually 1 and ϕ also denoted by σ^2 is the *dispersion* parameter which is constant over observations.

The log likelihood function is

$$\ell(\theta; y) = [T(y)\theta - b(\theta)]/a(\phi) + c(y, \phi). \tag{2}$$

where

$$\frac{\partial \ell}{\partial \theta} = [T(y) - b'(\theta)]/a(\phi), \quad \frac{\partial^2 \ell}{\partial \theta^2} = -b''(\theta)/a(\phi) \tag{3}$$

and $b'(\theta)$ and $b''(\theta)$ are the first and second order derivatives of $b(\theta)$. Moreover the mean and variance satisfy

$$E\left(\frac{\partial \ell}{\partial \theta}\right) = 0 \text{ and } E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) + E\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] = 0.$$

Then from (3),

$$\begin{aligned} E\{[T(Y) - b'(\theta)]/a(\phi)\} = 0 &\Rightarrow E[T(Y)] = \mu = b'(\theta), \\ -\frac{b''(\theta)}{a(\phi)} + \frac{\text{var}[T(Y)]}{a^2(\phi)} = 0 &\Rightarrow \text{var}[T(Y)] = \sigma^2 = b''(\theta)a(\phi). \end{aligned}$$

Hence the normal equation $\frac{\partial \ell}{\partial \theta} = 0$ in (3) matches the sample mean with the fitted mean. For one parameter distributions, $a(\phi) = 1$ and if $T(Y) = Y$, $E(Y) = b'(\theta)$ and $\text{Var}(Y) = b''(\theta)$. Thus the mean and variance for members of the exponential family except for the normal (2-parameter) are functionally related.

When $a(\phi) = \phi/w$, the variance has the simpler form

$$\text{var}(Y) = \sigma^2 = b''(\theta)\phi/w.$$

Exponential families include normal, Bernoulli, binomial (fixed n), Poisson, geometric, negative binomial (fixed number of failures), exponential, gamma, multinomial (with fixed number of trials), inverse Gaussian, chi-square, beta, Weibull (fixed shape parameter), Dirichlet, Wishart, Inverse Wishart distributions as well as the family of Pareto distributions

(fixed minimum bound). The Cauchy, Laplace, t, uniform families and mixtures distributions are not exponential families.

Exponential families have many desirable properties including *sufficient statistics* that can summarize arbitrary amounts of iid data and *conjugate priors*, an important property in Bayesian statistics.

Example: Rayleigh distribution:

The pdf is $f(y)=\frac{y}{\sigma^2}\exp\left(-\frac{y^2}{2\sigma^2}\right)=\exp\left(-\frac{y^2}{2\sigma^2}+\ln y-\ln\sigma^2\right),\ \ y\geq 0$

Hence $T(y)=y^2$, $\eta(\theta)=\theta=\frac{-1}{2\sigma^2}\Rightarrow\sigma^2=\frac{1}{-2\theta}$,
$b(\theta)=\ln\sigma^2=\ln(\frac{1}{-2\theta})=-\ln(-2\theta)$, and $c(t)=\ln y$.

For a sample of $y_i,\ i=1,\ldots,n$, the sufficient statistic is $\sum_{i=1}^{n}y_i^2$. Then

$$\begin{aligned} E[T(y)] &= E(y^2)=\frac{\partial b(\theta)}{\partial\theta}=-\frac{-2}{-2\theta}=-\frac{1}{\theta}=2\sigma^2 \\ Var[T(y)] &= Var(y^2)=\frac{\partial^2 b(\theta)}{\partial\theta^2}=\frac{\partial}{\partial\theta}\left(-\frac{1}{\theta}\right)=\frac{1}{\theta^2}=4\sigma^4 \end{aligned}$$

Example: Weibull distribution:

$$\begin{aligned} f(y)&=\frac{k}{\lambda}\left(\frac{y}{\lambda}\right)^{k-1}e^{-(\frac{y}{\lambda})^k}=\exp[\ln k+(k-1)\ln y-k\ln\lambda-\frac{y^k}{\lambda^k}] \\ &= \exp\left[y^k\left(-\frac{1}{\lambda^k}\right)-k\ln\lambda+\ln k+(k-1)\ln y\right] \end{aligned}$$

Hence $a(\phi)=\phi=1$, $T(y)=y^k$, $\theta=\frac{-1}{\lambda^k}\Rightarrow\lambda^k=(-\frac{1}{\theta})$,
$b(\theta)=k\ln\lambda=\ln\lambda^k=\ln(-\frac{1}{\theta})$, $c(y,\phi)=\ln k+(k-1)\ln y$,

$$E[T(Y)] = E(Y^k)=b'(\theta)=\frac{1}{-\frac{1}{\theta}}\cdot(-1)(-1)\theta^{-2}=-\frac{1}{\theta}=\lambda^k \text{ and}$$

$$\text{Var}[T(Y)] = \text{Var}(Y^k)=a(\phi)b''(\theta)=\frac{1}{\theta^2}=\lambda^{2k}.$$

This agrees with the n-th moment of Y being $E(Y^n) = \lambda^n \Gamma(1+\frac{n}{k})$ such that $E(Y) = \lambda\Gamma(1+\frac{1}{k})$ and $E(Y^k) = \lambda^k\Gamma(1+\frac{k}{k}) = \lambda^k\Gamma(2) = \lambda^k$.

Table: Summary of distributions in the exponential family

Dist.	Range	Dis. ϕ	Cum. $b(\theta)$	Other $c(y;\phi)$	Link θ $b'^{-1}, g(\mu)$	Mean μ $b'(\theta)$	Var $V(\mu)$ $b''(\theta)a(\phi)$
Normal $\mathcal{N}(\mu,\sigma^2)$	$(-\infty,\infty)$ conts.	σ^2	$\theta^2/2$	$-\frac{1}{2}[\ln(2\pi\phi) + \frac{y^2}{\phi}]$	μ identity	θ identity	σ^2
Poisson $\mathcal{P}(\mu)$	$0(1)\infty$ count	1	$\exp(\theta)$	$-\ln y!$	$\ln(\mu)$ log	$\exp(\theta)$ exp	μ
Binomial $\mathcal{B}(n,\pi)$	$0(1)n$ binary	1	$n\ln(1+e^\theta)$	$\ln\binom{n}{y}$	$\ln(\frac{\pi}{1-\pi})$ logit	$\frac{ne^\theta}{1+e^\theta} = n\pi$	$n\pi(1-\pi)$
Gamma $\mathcal{G}(\alpha,\beta)$	$(0,\infty)$ + conts.	$\frac{1}{\alpha}$	$-\ln(-\alpha\theta)$	$(\alpha-1)\ln y - \ln\Gamma(\alpha)$	$\frac{-1}{\mu} = \frac{-\beta}{\alpha}$ inverse	$\frac{-1}{\theta} = \frac{\alpha}{\beta}$ inverse	$\frac{\mu^2}{\alpha} = \frac{\alpha}{\beta^2}$
Inv. Gau. $\mathcal{I}G(\mu,\sigma^2)$	$(0,\infty)$ + conts.	σ^2	$-(-2\theta)^{\frac{1}{2}}$	$-\frac{1}{2}[\ln(2\pi\phi y^3) + \frac{1}{\phi y}]$	$-\frac{1}{2}\mu^{-2}$ inv. sq.	$(-2\theta)^{-\frac{1}{2}}$ inv. sq. r.	μ^3
Neg. Bin. $\mathcal{N}B(r,\pi)$	$0(1)\infty$	1	$-r\ln(1-e^\theta)$	$\ln\Gamma(r+y) - \ln\Gamma(r) - \ln y!$	$\ln(1-\pi)$	$\frac{re^\theta}{1-e^\theta} = \frac{r(1-\pi)}{\pi}$	$\frac{re^\theta}{(1-e^\theta)^2} = \frac{r(1-\pi)}{\pi^2}$

If the canoncial link is used, there is a simple set of minimal sufficient statistics on $(\boldsymbol{Y}, \boldsymbol{X})$ for $\boldsymbol{\beta}$. The likelihood function is

$$L(\boldsymbol{\beta},\phi) = \prod_{i=1}^{n} \exp\left\{\frac{1}{a(\phi)}\left[y_i\left(\sum_{j=1}^{p} x_{ij}\beta_j\right) - b(\boldsymbol{x}_i'\boldsymbol{\beta})\right] + c(y_i,\phi)\right\}$$

$$= \exp\left\{\frac{1}{a(\phi)}\left[\sum_{j=1}^{p}\beta_j\left(\sum_{i=1}^{n} y_i x_{ij}\right) - \sum_{i=1}^{n} b(\boldsymbol{x}_i'\boldsymbol{\beta})\right] + \sum_{i=1}^{n} c(y_i,\phi)\right\}$$

so $\left\{\sum_{i=1}^{n} y_i x_{ij}\right\}$, $j = 1,\ldots,p$ are sufficient for $\boldsymbol{\beta}$. Hence if x_{ij} are 0-1 factor codes then the marginal totals (the sum of the y_i's with attribute j) are sufficient for the $\boldsymbol{\beta}$.

§2.3 Weighted Least Squares (WLS)

The least square (LS) estimates for the model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \overset{iid}{\sim} \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \tag{4}$$

based on the log-likelihood function:

$$\ln L(\boldsymbol{y}; \boldsymbol{\beta}, \sigma^2) = \frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2}{2\sigma^2} \tag{5}$$

are

$$\widehat{\boldsymbol{\beta}} \;=\; (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

which is also a best linear unbiased estimator (BLUE). However if

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ is a known positive definite matrix with diagonals

$$\boldsymbol{\Sigma} = \text{diag}(c_1^2, \ldots, c_n^2),\ c_i^2 = \sigma_i^2/\sigma^2 \text{ and } \text{Cov}(\epsilon_i, \epsilon_j) = 0,\ i \neq j,$$

we consider transformation to a standard linear model in (4). Let $\boldsymbol{Y}^* = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{Y}$ and $\boldsymbol{X}^* = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}$ where $\boldsymbol{\Sigma}^{\frac{1}{2}}$ is a matrix such that $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}} = \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-\frac{1}{2}} = (\boldsymbol{\Sigma}^{\frac{1}{2}})^{-1}$. Then

$$\boldsymbol{Y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*, \qquad \boldsymbol{\epsilon}^* = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$$

since

$$\text{Var}(\boldsymbol{\epsilon}^*) = \boldsymbol{\Sigma}^{-\frac{1}{2}}\text{Var}(\boldsymbol{\epsilon})\boldsymbol{\Sigma}'^{-\frac{1}{2}} = \sigma^2\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}'^{-\frac{1}{2}} = \sigma^2\boldsymbol{I}.$$

The LS estimates for $\boldsymbol{\beta}$ are thus

$$(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^{*\prime}\boldsymbol{Y}^*, \quad \Rightarrow \quad (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

The WLS estimator is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1})$$

by the Gauss-Markov Theorem. The weight is $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1} = \text{diag}(w_1, .., w_n)$ where $w_i = 1/c_i^2$. When $p = 2$, we have

$$\widehat{\beta}_1 = \frac{(\sum_i w_i)(\sum_i w_i X_i Y_i) - (\sum_i w_i X_i)(\sum_i w_i Y_i)}{(\sum_i w_i)(\sum_i w_i X_i^2) - (\sum_i w_i X_i)^2}$$

It is unbiased for $\boldsymbol{\beta}$ and is a *best linear unbiased estimator* (BLUE). If standard LS is used, the estimate $\widehat{\boldsymbol{\beta}}_o = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ is still unbiased but it will not have minimum variance since

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}_o) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta} \text{ and} \\ \text{Var}(\widehat{\boldsymbol{\beta}}_o) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'[\text{Var}(\boldsymbol{Y})]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

as compared to $\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$. The residuals are

$$\boldsymbol{\epsilon}^* = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}) = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1})\boldsymbol{Y}.$$

Example: In economic data, the variability of the data can be modelled in terms of time, e.g.

$$y_i = \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, kX_i), \ i = 1, \ldots, n.$$

Let $\sigma_i^2 = kX_i$. Consider $Y_i^* = \frac{\sigma}{\sigma_i}Y_i = \sqrt{w_i}Y_i$, $X_i^* = \frac{\sigma}{\sigma_i}X_i = \sqrt{w_i}X_i$, $\epsilon_i^* = \frac{\sigma}{\sigma_i}\epsilon_i = \sqrt{w_i}\epsilon_i$ where

$$w_i = \frac{\sigma^2}{\sigma_i^2} = \frac{\sigma^2}{kX_i},$$

we have $Y_i^* = \beta_1 X_i^* + \epsilon_i^*$ where $\epsilon_i^* \sim \mathcal{N}(0, \sigma^2)$. It can be shown that

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{i=1}^n X_i^* Y_i^*}{\sum_{i=1}^n X_i^{*2}} = \frac{\sum_{i=1}^n w_i X_i Y_i}{\sum_{i=1}^n w_i X_i^2} = \frac{\sum_{i=1}^n \frac{X_i Y_i}{kX_i}}{\sum_{i=1}^n \frac{X_i^2}{kX_i}} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i} = \frac{\overline{Y}}{\overline{X}}, \text{ and} \\ \text{Var}(\widehat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^n w_i X_i^2} = \frac{\sigma^2}{\frac{\sigma^2}{k}\sum_{i=1}^n \frac{X_i^2}{X_i}} = \frac{k}{\sum_{i=1}^n X_i} = \frac{k}{n\overline{X}}. \end{aligned}$$

If $\text{Var}(Y_i) = kX_i^2$, $w_i = \dfrac{\sigma^2}{kX_i^2}$. Hence

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n w_i X_i Y_i}{\sum_{i=1}^n w_i X_i^2} = \frac{\frac{\sigma^2}{k}\sum_{i=1}^n \frac{X_i Y_i}{X_i^2}}{\frac{\sigma^2}{k}\sum_{i=1}^n \frac{X_i^2}{X_i^2}} = \frac{\sum_{i=1}^n \frac{Y_i}{X_i}}{\sum_{i=1}^n 1} = \frac{1}{n}\sum_{i=1}^n \frac{Y_i}{X_i} \quad \text{and} \\
\text{Var}(\widehat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^n w_i X_i^2} = \frac{\sigma^2}{\frac{\sigma^2}{k}\sum_{i=1}^n \frac{X_i^2}{X_i^2}} = \frac{k}{n}.
\end{aligned}$$

Remarks: For restricted LS with restriction $\boldsymbol{C\beta} = \boldsymbol{a}$ on the parameters, one can minimize the Lagrangean function

$$(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta}) + \lambda(\boldsymbol{a} - \boldsymbol{C\beta}). \tag{6}$$

The solution is

$$\tilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} + (\boldsymbol{X'X})^{-1}\boldsymbol{C}'[\boldsymbol{C}(\boldsymbol{X'X})^{-1}\boldsymbol{C}']^{-1}(\boldsymbol{a} - \boldsymbol{C}\widehat{\boldsymbol{\beta}}) \tag{7}$$

where $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$ is the unrestricted estimator. The proof is in the Appendix.

Check that

$$\boldsymbol{C}\tilde{\boldsymbol{\beta}} = \boldsymbol{C}\widehat{\boldsymbol{\beta}} + \boldsymbol{C}(\boldsymbol{X'X})^{-1}\boldsymbol{C}'[\boldsymbol{C}(\boldsymbol{X'X})^{-1}\boldsymbol{C}']^{-1}(\boldsymbol{a} - \boldsymbol{C}\widehat{\boldsymbol{\beta}}) = \boldsymbol{C}\widehat{\boldsymbol{\beta}} + \boldsymbol{a} - \boldsymbol{C}\widehat{\boldsymbol{\beta}} = \boldsymbol{a}$$

§2.4 Iterative reweighted LS (IRLS) or M-estimation

LS estimates can behave badly when the error distribution is not normal, particularly when the errors are heavy-tailed. One remedy is to remove influential observations. Another approach, termed *robust regression*, is to employ a fitting criterion that is not as vulnerable as least squares to unusual data. The most common general method of robust regression is M-estimation, introduced by Huber (1964). It refers to the LS with an objective function

$$\sum_{i=1}^{n} \rho(\epsilon_i) = \sum_{i=1}^{n} \rho(Y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) \tag{8}$$

where the objective function $\rho(\cdot)$ satisfies:
1. $\rho(\epsilon_i) \geq 0$,
2. $\rho(0) = 0$,
3. $\rho(\epsilon_i) = \rho(-\epsilon_i)$ and
4. $\rho(\epsilon_i) \geq \rho(\epsilon_j)$ for $|\epsilon_i| > |\epsilon_j|$,

For ordinary LS, $\rho(\epsilon_i) = \epsilon_i^2$. Differentiate (8) w.r.t. $\boldsymbol{\beta}$, the p estimating equations for $\boldsymbol{\beta}$ are

$$\sum_{i=1}^{n} \rho'(\epsilon_i)\boldsymbol{x}_i' = \mathbf{0} \Rightarrow \sum_{i=1}^{n} \frac{\rho'(\epsilon_i)}{\epsilon_i}\epsilon_i\boldsymbol{x}_i' = \mathbf{0} \Rightarrow \sum_{i=1}^{n} w_i(Y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i' = \mathbf{0} \tag{9}$$

writing the weight function $w_i = w(\epsilon_i) = \rho'(\epsilon_i)/\epsilon_i$. The weights w_i, however, depend on residuals ϵ_i, which in terms depend on the estimated coefficients $\boldsymbol{\beta}$ and then weights w_i again. Hence iterative procedures are required. The IRLS procedures are

1. Set $\boldsymbol{\beta}^{(0)}$ to be the LS estimates.
2. Calculate $\epsilon_i^{(k)}$ and associated weights $w_i^{(k)}$ at iteration k.

3. Solve for the new WLS estimates

$$\boldsymbol{\beta}^{(k+1)} = [\boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{Y}$$

where $\boldsymbol{W}^{(k)} = \text{diag}\{w_i^{(k)}\}$ is the current weight matrix.

Steps 2 and 3 are then repeated until $\widehat{\boldsymbol{\beta}}$ converge. Then

$$\text{Var}(\boldsymbol{\beta}) = \frac{E(\rho'^2)}{[E(\rho'')]^2}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

using $\frac{1}{n}\sum_{i=1}^{n}[\rho'(\epsilon_i)]^2$ to estimate $E(\rho'^2)$ and

$\left[\frac{1}{n}\sum_{i=1}^{n}\rho''(\epsilon_i)\right]^2$ to estimate $[E(\rho'')]^2$. The proof is given in Huber (1964).

If $\rho(\epsilon) = \epsilon^2$ for LS estimates, $\rho'(\epsilon) = 2\epsilon$ and $\rho''(\epsilon) = 2$. Then

$$\text{Var}(\boldsymbol{\beta}) = \frac{E(\rho'^2)}{[E(\rho'')]^2}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{4\sigma^2}{4}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

For the choice of objective functions, the following figure and table gives the objective function $\rho(\epsilon)$, $\psi(\epsilon) = \rho'(\epsilon)$ and weight functions $w(\epsilon) = \rho'(\epsilon)/\epsilon$ for 3 M-estimators: the LS estimator; the Huber (H) estimator; and the Tukey bisquare (TB) estimator.

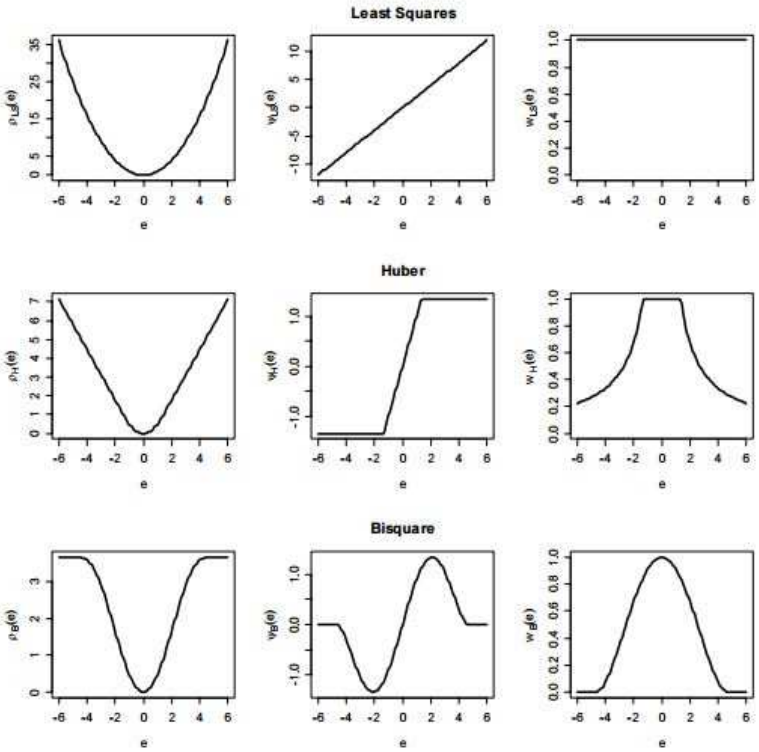


Table: Objective and weight functions for least-square, Huber and bisquare estimator

Method	Objective function ρ	Weight function $\rho'(\epsilon)/\epsilon$
Least-sq.	$\rho_{LS}(\epsilon) = \epsilon^2$	$w_{LS}(\epsilon) = 1$
Huber	$\rho_H(\epsilon) = \begin{cases} \frac{1}{2}\epsilon^2 & \text{for } \lvert\epsilon\rvert \le k \\ k\lvert\epsilon\rvert - \frac{1}{2}k^2 & \text{for } \lvert\epsilon\rvert > k \end{cases}$	$w_H(\epsilon) = \begin{cases} 1 & \text{for } \lvert\epsilon\rvert \le k \\ k/\lvert\epsilon\rvert & \text{for } \lvert\epsilon\rvert > k \end{cases}$
Bisquare	$\rho_{BS}(\epsilon) = \begin{cases} \frac{k^2}{6}\left[1 - \left(1 - \left(\frac{\epsilon}{k}\right)^2\right)^3\right] & \text{for } \lvert\epsilon\rvert \le k \\ k^2/6 & \text{for } \lvert\epsilon\rvert > k \end{cases}$	$w_{BS}(\epsilon) = \begin{cases} \left[1 - \left(\frac{\epsilon}{k}\right)^2\right]^2 & \text{for } \lvert\epsilon\rvert \le k \\ 0 & \text{for } \lvert\epsilon\rvert > k \end{cases}$

Both the LS and H objective functions increase without bound as the residual ϵ departs from 0, but the LS increases faster. In contrast, the TB objective function levels off for $|\epsilon| > k$, called a *tuning constant*. The weights are equal for LS estimator; decline when $|\epsilon| > k$ for H estimator and decline as soon as ϵ departs from 0, and are 0 for $|\epsilon| > k$ for TB estimator.

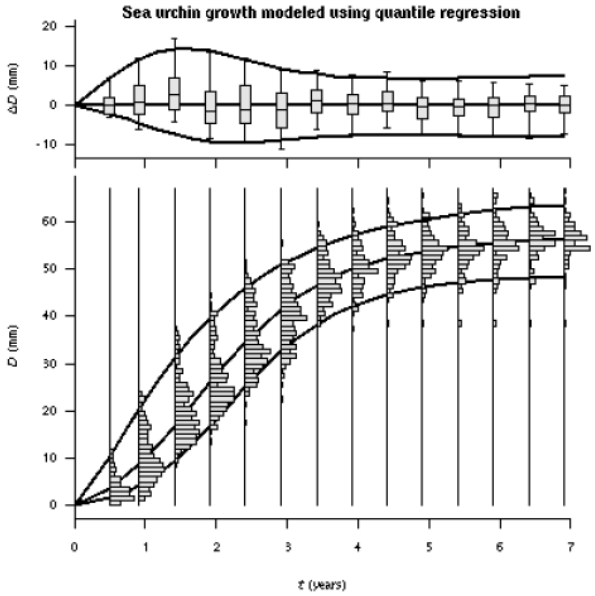
Smaller k produce more resistance to outliers, but at the expense of lower efficiency for normal errors. Hence k is generally picked to give reasonably high (95%) efficiency in the normal case: $k = 1.345\sigma$ and 4.685σ (σ is the sd of ϵ) respectively for the H and TB estimator, and still offer protection against outliers. A robust

estimate of σ (preferable to sd) is to take $\sigma = MAR/0.6745$, where MAR is the median absolute residual.

§2.5 Quantile regression

§2.5.1 Asymmetric loss function

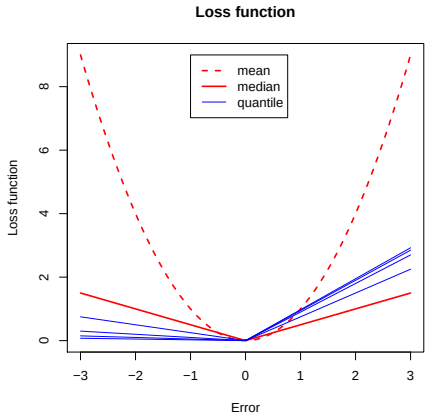
Mean regression models cannot differentiate covariate effects that drive high to low responses.

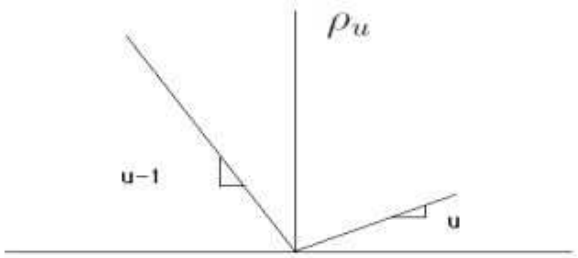


Koenker and Bassett (1978) proposed to estimate parameters $\boldsymbol{\beta}_u$ *across quantile level* $u \in (0,1)$ by minimizing the *asymmetric loss* function:

$$\sum_{t=1}^{n} \rho_u(\epsilon_t) \quad \text{where} \quad \rho_u(\epsilon_t) = \epsilon_t[u - I(\epsilon_t < 0)] \tag{10}$$

and $\epsilon_t = y_t - \boldsymbol{x}_t\boldsymbol{\beta}_u$ are residuals. The estimate $\widehat{\boldsymbol{\beta}}_u$ is u-th conditional quantile in *nonparametric quantile regression*. The slope is $u-1$ when $\epsilon < 0$ and is u when $\epsilon < 0$. When $u = 0.5$, the objective function becomes symmetric and it corresponds to median regression. The objective function is graphed below:





Nonparametric quantile regression is also *robust* to distributional assumptions, e.g. homogeneous variance across risk factors, independence and symmetric and light tails. Moreover parameter estimates $\boldsymbol{\beta}_u$ have an asymptotic normal distribution

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_u - \boldsymbol{\beta}_u) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_u),$$

§2.5.2 Parameter estimation

The minimization of (10) can be performed using the `R` package `quantreg`

```
library(quantreg)
rq(y~x,tau=taus,method="br")
```

contrinbuted by Koenker where `taus` is a vector of quantile levels τ and `"br"` is the default method of estimation called the Simplex method.

Example: In loss reserving, mean estimator is not statistically robust and therefore sensitive to outlier claims. Insurance provision requires a higher risk margin, say 75%. The following table reports the observed and predicted claims in the run-off triangle using 0.75 quantile level for the Israel loss reserves data. Each row shows claims made across `development year` for all policies made in a certain `policy year` as shown in the first column.

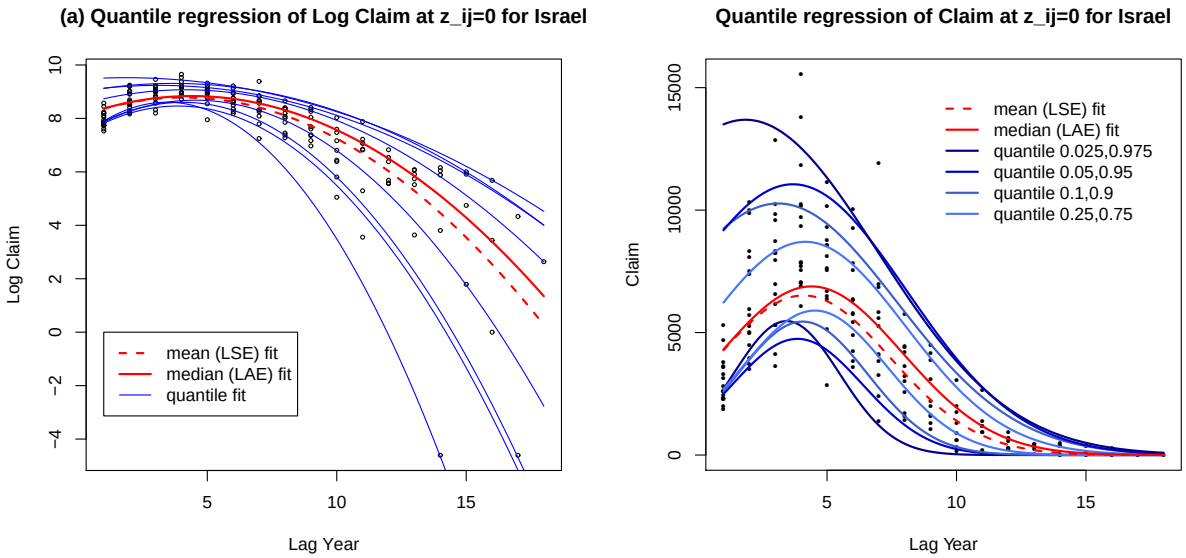
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1978	3323	8332	9572	10172	7631	3855	3252	4433	2188	333	199	692	311	0.01	405	293	76	14
1979	3785	10342	8330	7849	2839	3577	1404	1721	1065	156	35	259	250	420	6	1	0.01	*14*
1980	4677	9989	8746	10228	8572	5787	3855	1445	1612	626	1172	589	438	473	370	31	*35*	14
1981	5288	8089	12839	11829	7560	6383	4118	3016	1575	1985	2645	266	38	45	115	*81*	35	14
1982	2294	9869	10242	13808	8775	5419	2424	1597	4149	1296	917	295	428	359	*165*	77	34	14
1983	3600	7514	8247	9327	8584	4245	4096	3216	2014	593	1188	691	368	*339*	169	79	34	14
1984	3642	7394	9838	9733	6377	4884	11920	4188	4492	1760	944	921	*636*	340	170	79	35	14
1985	2463	5033	6980	7722	6702	7834	5579	3622	1300	3069	1370	*1087*	621	331	165	77	34	14
1986	2267	5959	6175	7051	8102	6339	6978	4396	3107	903	*1780*	1087	621	331	165	77	34	14
1987	2009	3700	5298	6885	6477	7570	5855	5751	3871	*2691*	1757	1073	613	327	163	76	33	14
1988	1860	5282	3640	7538	5157	5766	6862	2572	*3834*	2678	1749	1068	609	325	162	76	33	13
1989	2331	3517	5310	6066	10149	9265	5262	*5207*	3890	2717	1774	1083	618	330	165	77	34	14
1990	2314	4487	4112	7000	11163	10057	*6515*	5205	3888	2716	1773	1083	618	330	165	77	34	14
1991	2607	3952	8228	7895	9317	*7683*	6565	5245	3918	2736	1787	1091	623	332	166	77	34	14
1992	2595	5403	6579	15546	*8405*	7681	6563	5243	3917	2736	1786	1091	623	332	166	77	34	14
1993	3155	4974	7961	*8708*	8511	7778	6646	5310	3967	2770	1809	1105	631	337	168	78	34	14
1994	2626	5704	*8232*	8605	8411	7687	6568	5247	3920	2738	1788	1092	623	333	166	77	34	14
1995	2827	*7398*	8271	8646	8451	7723	6599	5273	3939	2751	1796	1097	626	334	167	78	34	14

LossReserveData.txt:

```
3323 1 1 3323
8332 1 2 3323
9572 1 3 3323
...
```

```
> library(quantreg)
> taus <- c(.025,.05,.1,.25,.5,.75,.9,.95,.975)
> data <- read.table("LossReserveData.txt",header=FALSE)  #Israel data
> Claim=data[,1]
> n=length(Claim); np=4; nd=max(data[,2]); nt=length(taus)
> PolicyYear=data[,2]
> LagYear=data[,3]
> Level=log(data[,4])  #development 1st yr claim for each policy yr
> StLevel=(Level-mean(Level))/sd(Level)
> LnClaim=log(Claim)
> LagYearSq=LagYear^2
> qreg=rq(LnClaim~LagYear+LagYearSq+StLevel,tau=taus) #quantile reg
> beta=coef(qreg)
> beta
              tau= 0.025  tau= 0.050  tau= 0.100  tau= 0.250  tau= 0.500
(Intercept)  7.2016085  7.28969705  7.29686599  7.38761540  8.05382754
LagYear      0.8176366  0.60901409  0.64388055  0.57174407  0.35619636
LagYearSq   -0.1188986 -0.07891841 -0.07933669 -0.06310696 -0.04049392
StLevel      0.2295675  0.12337026  0.12451423  0.04194221  0.01621570
```

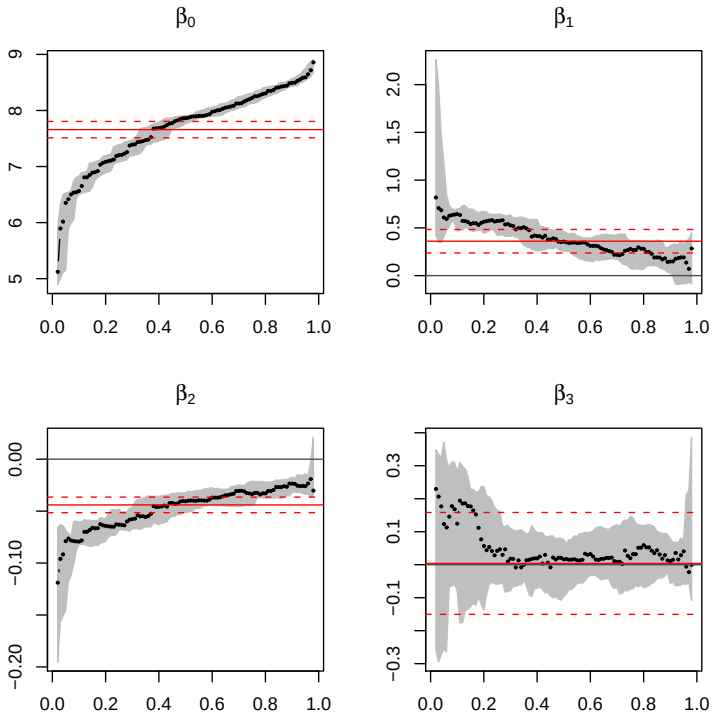
```
             tau= 0.750  tau= 0.900  tau= 0.950  tau= 0.975
(Intercept)  8.49019163  9.00849287  8.95903434  9.45856712
LagYear      0.27962915  0.14693654  0.19086638  0.07107803
LagYearSq   -0.03360740 -0.02361571 -0.02589786 -0.01918251
StLevel      0.01974775  0.01085010  0.04039713 -0.02240238

> for(k in 1:nt){
> yhat[,k]=exp(beta[1,k]+beta[2,k]*LagYear+beta[3,k]*LagYearSq+
               beta[4,k]*StLevel)} #quantile fcn
>
> summary(qreg,se="nid")

Call: rq(formula = LnClaim ~ LagYear + LagYearSq + StLevel, tau = taus)

tau: [1] 0.025
Coefficients:
             Value    Std. Error t value  Pr(>|t|)
(Intercept)  7.20161  0.32411   22.21960  0.00000
LagYear      0.81764  0.18411    4.44101  0.00002
LagYearSq   -0.11890  0.02046   -5.81094  0.00000
StLevel      0.22957  0.13375    1.71640  0.08794

Call: rq(formula = LnClaim ~ LagYear + LagYearSq + StLevel, tau = taus)
...
tau: [1] 0.975
Coefficients:
             Value    Std. Error t value  Pr(>|t|)
(Intercept)  9.45857  0.14282   66.22875  0.00000
LagYear      0.07108  0.05362    1.32561  0.18678
LagYearSq   -0.01918  0.00329   -5.83043  0.00000
StLevel     -0.02240  0.07102   -0.31545  0.75281
```



The left plot gives the fitted quantile functions for the log-transformed data and the right plot gives the fitted quantile functions for the original data. Note that the quantile functions *cross over* at extreme level quantile when the data are *rare and are heavily weighted* by the asymmetric loss function.

```
> LagYear.d=LagYear - mean(LagYear)
> LagYearSq.d=LagYearSq - mean(LagYearSq)
> StLevel.d=StLevel - mean(StLevel)
> fit1=summary(rq(LnClaim~LagYear.d+LagYearSq.d+StLevel.d,tau = 2:98/100))
> win.graph()
> plot(fit1,main=expression(beta[0],beta[1],beta[2],beta[3])) #coeff plot
```

The plot shows how the parameter estimates and their 95% confidence intervals changes across quantiles.



§2.5.3 Bayesian approach

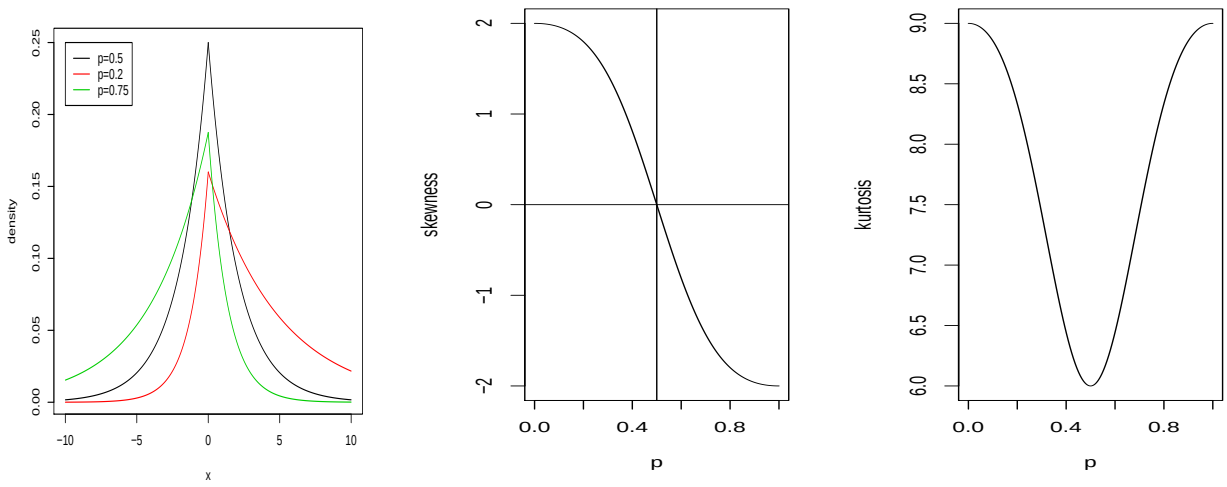
Koenker and Machado (1999) proposes *minimising the loss function is equivalent to maximising the likelihood function of the Asymmetric Laplace* (AL; double exponential when $u = 0.5$) distribution with pdf:

$$f_{AL}(y_t|\psi_t, \tau^2, u) = \frac{u(1-u)}{\tau} \exp\left\{ -\frac{y_t - \boldsymbol{x}_t\boldsymbol{\beta}_u}{\tau} \left[u - I(y_t \leq \boldsymbol{x}_t\boldsymbol{\beta}_u)\right] \right\}$$

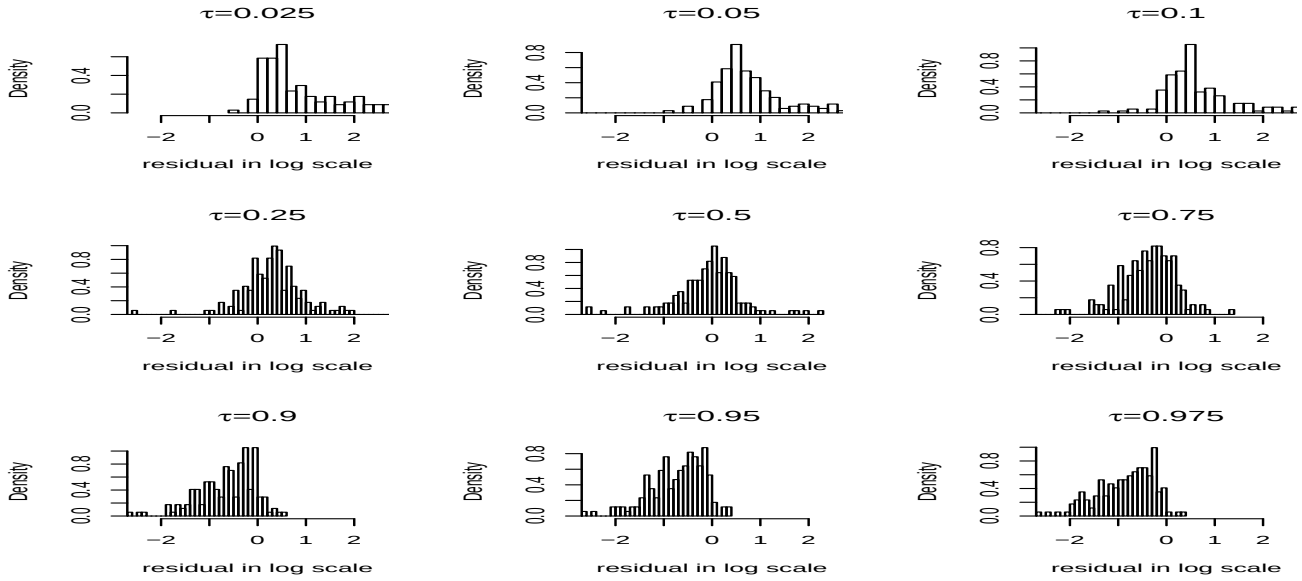
where $\boldsymbol{x}_t\boldsymbol{\beta}_u \in \Re$, $\tau > 0$ and $u = \Pr(\ln Y_t < \psi_t) \in (0,1)$ are location (mode), scale and shape parameters since the pdf contains the loss function (10). AL distribution can be expressed as scale mixture of uniform (SMU) to facilitate implementation in Bayesian MCMC approach:

$$\begin{aligned} f_{AL}(y_t) &= \int_0^\infty f_U\left(y_t \middle| \ \boldsymbol{x}_t\boldsymbol{\beta}_u - \frac{\tau\lambda_t}{1-u},\ \boldsymbol{x}_t\boldsymbol{\beta}_u + \frac{\tau\lambda_t}{u}\right) f_G(\lambda_t|\,2,1)\, d\lambda_t \\ &= \int_0^\infty \frac{u(1-u)\exp(-\lambda_t)}{\tau(1-2u)\Gamma(2)} I\left(y_t \in \left(\boldsymbol{x}_t\boldsymbol{\beta}_u - \frac{\tau\lambda_t}{1-u}, \boldsymbol{x}_t\boldsymbol{\beta}_u + \frac{\tau\lambda_t}{u}\right)\right) d\lambda_t \end{aligned}$$

where $f_U(\cdot|l_t, u_t)$ denotes $U(l_t, u_t)$ and $f_G(\cdot|c, d)$ denotes $Gamma(c, d)$ with scale parameter d and mean $\frac{c}{d}$. The density, skewness (drop with u) and kurtosis (min at $u = 0.5$) of AL distribution are shown below:



Note that the residual plot below shows asymmetric distributions which agree with the shape of AL distribution for extreme quantile levels.



§2.6 ML estimate for GLM

An important practical feature of GLM models is that they all can be fitted to data using the same algorithm, the *iteratively re-weighted least squares* (IRLS) algorithm. McCullagh and Nelder (1989) prove that this algorithm is equivalent to Fisher scoring and leads to ML estimates. IRLS is applied to the modified dependent variable

$$z_i = \eta_i + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i} \tag{11}$$

by mapping y_i to z_i which equals to the *transformed mean* η_i plus a *transformed error* $(y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$. The mapping corresponds to a linearized form :

$$g(y) = g(\mu) + (y - \mu)g'(\mu)$$

using the link function $g(\mu) = \eta$. The weight is

$$W_i = Var(z_i)^{-1} = \left[V_i \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2\right]^{-1} = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{1}{V_i} \tag{12}$$

from (11) where $\text{Var}(Y_i) = V_i$.

In summary,

	Obs	Mean	Error	Var
Observed space	y_i	μ_i	$y_i - \mu_i$	V_i
Transformed space	z_i	$\eta_i = \boldsymbol{x}_i\boldsymbol{\beta}$	$(y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$	$\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2 V_i$

To show the Fisher scoring corresponds to IRLS algorithm, we consider the first order derivative of the log-likelihood function (2), that is the

score vector $\boldsymbol{u} = \dfrac{\partial \ell}{\partial \boldsymbol{\beta}}$ with:

$$\begin{aligned} u_r = \frac{\partial \ell}{\partial \beta_r} &= \sum_i \frac{\partial \ell}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_r} \\ &= \sum_i \frac{(y_i-\mu_i)}{a(\phi)}\frac{1}{b''(\theta_i)}\frac{\partial \mu_i}{\partial \eta_i}x_{ir} \qquad (13) \\ &= \sum_i \frac{(y_i-\mu_i)}{V_i}\frac{\partial \mu_i}{\partial \eta_i}x_{ir} \\ &= \sum_i W_i\,(y_i-\mu_i)\frac{\partial \eta_i}{\partial \mu_i}x_{ir} \qquad (14) \end{aligned}$$

which is *the weighted sum of transformed residuals* in (9) by the chain rule, since

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) \Rightarrow \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)}, \qquad \frac{\partial \eta_i}{\partial \beta_r} = \frac{\partial}{\partial \beta_r}\sum_j \beta_r x_{ij} = x_{ir},$$

$$V_i = a(\phi)b''(\theta_i) \quad \text{and} \quad W_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{1}{V_i} \Rightarrow \frac{1}{V_i}\frac{\partial \mu_i}{\partial \eta_i} = W_i\frac{\partial \eta_i}{\partial \mu_i}.$$

The information matrix is $\boldsymbol{A} = -E\left(\dfrac{\partial^2 \ell}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right)$ with

$$\begin{aligned} A_{rs} &= -E\left(\frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s}\right) = -E\left(\frac{\partial u_r}{\partial \beta_s}\right) \\ &= -E\left\{\sum_i \left[W_i\,(y_i-\mu_i)\,x_{ir}\frac{\partial^2 \eta_i}{\partial \mu_i \partial \beta_s} + W_i\frac{\partial \eta_i}{\partial \mu_i}x_{ir}\frac{\partial}{\partial \beta_s}(y_i-\mu_i)\right]\right\} \\ &= \sum_i W_i\frac{\partial \eta_i}{\partial \mu_i}x_{ir}\frac{\partial \mu_i}{\partial \beta_s} = \sum_i W_i\,x_{ir}\,x_{is} \qquad (15) \end{aligned}$$

since $\dfrac{\partial \eta_i}{\partial \mu_i}\dfrac{\partial \mu_i}{\partial \beta_s} = \dfrac{\partial \eta_i}{\partial \beta_s} = x_{is}$ and $E(y_i-\mu_i) = 0$. Thus $\boldsymbol{A} = \boldsymbol{X'WX}$ is the *weighted sums-of-squares-and-products matrix* with weights $\boldsymbol{W}$.

From the Fisher scoring (FS) method,

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - E\left(\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right)^{-1} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}}.$$

That is

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \boldsymbol{A}^{-1}\boldsymbol{u} \Rightarrow \boldsymbol{A}\boldsymbol{\beta}^{(k+1)} = \boldsymbol{A}\boldsymbol{\beta}^{(k)} + \boldsymbol{u} \tag{16}$$

where

$$\begin{aligned}(\boldsymbol{A}\boldsymbol{\beta}^{(k)})_r &= \sum_s A_{rs}\beta_s^{(k)} = \sum_s \sum_i W_i\, x_{ir}\, x_{is}\beta_s^{(k)} \\ &= \sum_i W_i\, x_{ir} \left(\sum_s x_{is}\beta_s^{(k)}\right) = \sum_i W_i x_{ir}\eta_i^{(k)}\end{aligned}$$

from (15) and

$$\begin{aligned}(\boldsymbol{A}\boldsymbol{\beta}^{(k+1)})_r &= (\boldsymbol{A}\boldsymbol{\beta}^{(k)})_r + u_r = \sum_i W_i x_{ir}\left[\eta_i^{(k)} + (y_i - \mu_i^{(k)})\frac{\partial \eta_i}{\partial \mu_i}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(k)}}\right] \\ &= \sum_i W_i x_{ir} z_i^{(k)} = (\boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{z}^{(k)})_r\end{aligned}$$

from (14) and (16) implying that $\boldsymbol{A}\boldsymbol{\beta}^{(k+1)} = \boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{z}^{(k)}$ and that $\boldsymbol{\beta}^{(k+1)} = \boldsymbol{A}^{-1}\boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{z}^{(k)}$ where $\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{X}$. Hence the ML estimates by FS method is the same as the weighted least-squares (WLS) estimates

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = (\boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{z}^{(k)} \tag{17}$$

with the weights $W_i^{(k)}$ in the diagonals of $\boldsymbol{W}^{(k)}$ given by (12) and the working dependent variable z_i given by (11). The procedure is repeated until successive estimates change by less than a specified small amount.

Remark:

1. The MLE's are asymptotically unbiased, efficient and normal. The large sample distribution for $\widehat{\boldsymbol{\beta}}$ is

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\phi)$$

and the information matirx is

$$\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}/\phi = \left(-E\left[\frac{\partial \ell}{\partial \beta_r \partial \beta_s}\right]\right).$$

2. With normal errors and identity link $\eta_i = \mu_i$, the derivative is $\partial\eta_i/\partial\mu_i = 1$ and the working dependent variable is

$$z_i = \eta_i + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i} = y_i$$

itself. Since in addition,

$$b(\theta_i) = \frac{\theta_i^2}{2},\ b'(\theta_i) = \theta_i,\ b''(\theta_i) = 1,\ V_i = a(\phi) = \phi = \sigma^2 \text{ and}$$

the weights $W_i = \dfrac{1}{V_i}\left(\dfrac{\partial \mu_i}{\partial \eta_i}\right)^2 = \sigma^{-2}$, $\boldsymbol{W} = \sigma^{-2}\boldsymbol{I}$ is a constant. Hence no iteration is required. The exact distribution of $\widehat{\boldsymbol{\beta}}$ is multivariate normal with the mean $\boldsymbol{\beta}$ and variance covariance matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$.

For binomial and count data with logit and log links respectively, iterations are necessary.

3. Equation (13) gives a set of estimation equations (EE) to match the first order moment. If the second order moment is also matched, one may have:

$$EE_1 :\ \sum_i (y_i - \mu_i)\frac{E(\frac{\partial \epsilon_i}{\partial \theta})}{E(\epsilon_i^2)}$$

$$EE_2 :\ \sum_i (\epsilon_i^2 - \sigma_i^2)\frac{E(\frac{\partial v_i}{\partial \theta})}{E(v_i^2)} \text{ where } v_i = (\epsilon_i^2 - \sigma_i^2).$$

Example: (AIDS data) For the Poisson count with $\eta = \boldsymbol{x}'\boldsymbol{\beta}$,

$\mu = \exp(\eta)$, $\dfrac{\partial \mu}{\partial \eta} = \exp(\eta) = \mu$,

$z = \eta + (y-\mu)\dfrac{\partial \eta}{\partial \mu} = \eta + \dfrac{(y-\mu)}{\mu}$ and $W = \left(\dfrac{\partial \mu_i}{\partial \eta_i}\right)^2 V_i^{-1} = \dfrac{\mu^2}{\mu} = \mu.$

```
> y=c(0,1,2,3,1,5,10,17,23,31,20,25,37,45)
> x=c(1:14)
> n=length(x)
> one=c(rep(1,n))
> X=cbind(one,x)
> beta=matrix(0.3,2,1)  #starting values
>
> for (i in 1:10){
+ eta=X%*%beta
+ mu=exp(eta)
+ va=mu
+ W=matrix(0,n,n)
+ for (j in 1:n)  {W[j,j]=va[j]}
+ z=eta+(y-mu)/va
+ XWX=t(X)%*%W%*%X
+ XWXI=solve(XWX)
+ XWZ=t(X)%*%W%*%z
+ beta=XWXI%*%XWZ
+ VA=diag(XWXI)
+ se=sqrt(VA)
+ names(se) = NULL
+ result=c(i,beta[1],se[1],beta[2],se[2])
+ print(result)
+ }
[1] 1.00000000 0.31265005 0.22369592 0.26745248 0.01911643
[1] 2.00000000 0.37099248 0.24626829 0.25450048 0.02145492
[1] 3.00000000 0.37569641 0.24879148 0.25365113 0.02185838
[1] 4.00000000 0.37571105 0.24884170 0.25364851 0.02187525
[1] 5.00000000 0.37571105 0.24884184 0.25364851 0.02187530
[1] 6.00000000 0.37571105 0.24884184 0.25364851 0.02187530
```

```
[1] 7.00000000 0.37571105 0.24884184 0.25364851 0.02187530
[1] 8.00000000 0.37571105 0.24884184 0.25364851 0.02187530
[1] 9.00000000 0.37571105 0.24884184 0.25364851 0.02187530
[1] 10.00000000  0.37571105  0.24884184  0.25364851  0.02187530
```

§2.7 Quasi-Likelihood

Sometimes there is insufficient information from the data Y_i to specify a distribution for the data. We may want to model just the mean $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and variance $\text{Cov}(\boldsymbol{Y}) = \phi \boldsymbol{V}(\boldsymbol{\mu})$ where ϕ is a constant and $\boldsymbol{V}(\boldsymbol{\mu}) = \text{diag}(V_1(\mu_1), \ldots, V_n(\mu_n))$ but not a specific data distribution.

Consider the variable

$$U = \frac{Y - \mu}{\phi V(\mu)},$$

acting like a score function $u(\mu) = \dfrac{\partial \ell}{\partial \mu}$ with $E(U) = 0$,

$\text{Var}(U) = \dfrac{\text{Var}(Y)}{[\phi V(\mu)]^2} = \dfrac{1}{\phi V(\mu)}$ as $\text{Var}(Y) = \phi V(\mu)$ and

$$\begin{aligned} \frac{\partial U}{\partial \mu} &= \frac{-1}{\phi V(\mu)} + \frac{Y-\mu}{\phi} \frac{-1}{V(\mu)^2} \frac{\partial V(\mu)}{\partial \mu} \\ \Rightarrow E\left(\frac{\partial U}{\partial \mu}\right) &= \frac{-1}{\phi V(\mu)} = -\frac{\phi V(\mu)}{\phi^2 V^2(\mu)} = -E\left[\left(\frac{Y-\mu}{\phi V(\mu)}\right)^2\right] = -E(U^2) = -\text{Cov}(U) \end{aligned} \tag{18}$$

since $E(Y - \mu) = 0$. Recall that for the score function $u(\mu)$,

$$E\left(\frac{\partial u}{\partial \mu}\right) = E\left(\frac{\partial^2 \ell(\mu)}{\partial \mu^2}\right) = -E\left(\frac{\partial \ell(\mu)}{\partial \mu} \frac{\partial \ell(\mu)}{\partial \mu}\right) = -\text{Cov}(u)$$

which is the result for ML estimation using a log-likelihood function. Since most first order asymptotic theory connected with the likelihood functions depends on these properties, the integral

$$Q(\mu, y) = \int_y^\mu \frac{Y - t}{\phi V(t)} dt,$$

if exists, should behave like a log-likelihood function for μ. The function $Q(\mu, y)$ is called the *quasi-log likelihood* function. Note that $U = \dfrac{Y - \mu}{\phi V(\mu)} = \dfrac{\partial}{\partial \mu} Q(\mu, y)$.

Example: If $E(Y) = \mu = \text{Var}(Y)$,

$$Q(\mu; y) = \int_y^\mu \frac{y-t}{t} dt = \int_y^\mu \left(\frac{y}{t} - 1\right) dt = [y \ln t - t]_y^\mu = y \ln \mu - \mu - y \ln y + y$$

as compared to the Poisson distribution with

$$\ell(\mu; y) = \ln f(y) = y \ln \mu - \mu - \ln y!.$$

Hence $Q(\mu; y)$ and $\ell(\mu; y)$ differ by a constant and give the same derivatives.

Example: If $E(Y) = \mu = \frac{\alpha}{\beta}$, $\text{Var}(Y) = \frac{\mu^2}{\alpha}$ and α is known,

$$Q(\mu, y) = \alpha \int_y^\mu \frac{y-t}{t^2} dt = \alpha \int_y^\mu \left(\frac{y}{t^2} - \frac{1}{t}\right) dt = \alpha \left[-\frac{y}{t} - \ln t\right]_y^\mu = -\alpha \frac{y}{\mu} + \alpha - \alpha \ln \mu + \alpha \ln y$$

as compared to the Gamma(α, β) distribution with

$$\ell(\mu; y) = \ln f(y) = -\alpha \left(\frac{y}{\mu} + \ln \mu\right) + (\alpha - 1) \ln y + \alpha \ln \alpha - \ln \Gamma(\alpha).$$

Hence $Q(\mu; y)$ and $\ell(\mu; y)$ differ by a constant and give the same derivatives.

For n independent observations $Y_1, \ldots, Y_n$, the quasi-log-likelihood is

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

To estimate $\boldsymbol{\beta}$, one solves

$$U_Q(\beta_j) = \frac{\partial Q(\boldsymbol{\mu}, \boldsymbol{y})}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial Q}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} = \sum_{i=1}^n \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0$$

as compared to the score vector $\sum_{i=1}^n \frac{(y_i - \mu_i)}{V_i} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = 0$ in (14) for ML estimation. These equations can be written in a vector form as

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = U_Q(\boldsymbol{\beta}) = \phi^{-1} \cdot {}^{\text{p}}\boldsymbol{D'}^{\text{n}} \boldsymbol{V}^{-1\,\text{n}} (\boldsymbol{Y} - \boldsymbol{\mu})^1 \qquad (19)$$

where $\boldsymbol{D}$ is an $n \times p$ matrix of $D_{ij} = \dfrac{\partial \mu_i}{\partial \beta_j} \overset{or}{=} \dfrac{\partial \mu_i}{\partial \eta_i}\dfrac{\partial \eta_i}{\partial \beta_j} = \dfrac{\partial \mu_i}{\partial \eta_i} x_{ij}$ and $\boldsymbol{V}$ is a diagonal matrix of V_i. Note that from (19),

$$\begin{aligned} E[U_Q(\boldsymbol{\beta})] &= \mathbf{0} \text{ and} \\ \text{Cov}[U_Q(\boldsymbol{\beta})] &= \phi^{-2}\boldsymbol{D}'\boldsymbol{V}^{-1}(\phi\boldsymbol{V})\boldsymbol{V}^{-1}\boldsymbol{D} = \phi^{-1}\boldsymbol{D}'\boldsymbol{V}^{-1}\boldsymbol{D}. \end{aligned}$$

Solving $U_Q(\boldsymbol{\beta}) = \mathbf{0}$ for $\boldsymbol{\beta}$ using the Fisher scoring algorithm, we have

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{(k+1)} &= \widehat{\boldsymbol{\beta}}^{(k)} - E\left[\frac{\partial^2 Q(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'}\right]^{-1} \frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(k)}} \\ &= \widehat{\boldsymbol{\beta}}^{(k)} + (\boldsymbol{D}'\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})\bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(k)}} \end{aligned}$$

since (19) and from (18)

$$E\left[\frac{\partial^2 Q(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'}\right] = E\left[\frac{\partial U_Q(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right] = -\text{Cov}[U_Q(\boldsymbol{\beta})] = -\phi^{-1}\boldsymbol{D}'\boldsymbol{V}^{-1}\boldsymbol{D}.$$

The iterative procedures follow as the FS algorithm. Also it is natural to relax the independence assumption and allow $\boldsymbol{V}$ to be a *positive definite* covariance matrix instead of a *diagonal* matrix. This leads to the method of *Generalized Estimating Equations* (GEE) for possibly dependent data proposed by Zeger and Liang in 1986.

The following gives Quasi-likelihood associated with some simple variance functions.

Distribution	Variance fcn. $V(\mu)$	Canonical par. θ	Restricted range	Quasi-likelihood $Q(\mu; y)$
Normal	1	μ	-	$-\frac{(y-\mu)^2}{2}$
Poisson	μ	$\ln \mu$	$\mu > 0, y \geq 0$	$y \ln \mu - \mu$
Gamma	μ^2	$-\frac{1}{\mu}$	$\mu > 0, y > 0$	$-y/\mu - \ln \mu$
Inv. Gauss.	μ^3	$\frac{-1}{2\mu^2}$	$\mu > 0, y > 0$	$\frac{-y}{2\mu^2} + \frac{1}{\mu}$
Bin./m	$\mu(1-\mu)$	$\ln \frac{\mu}{1-\mu}$	$0 < \mu < 1,\ 0 \leq y \leq 1$	$y \ln \frac{\mu}{1-\mu} + \ln(1-\mu)$
Neg. bin.	$\mu + \frac{\mu^2}{r}$	$\ln \frac{\mu}{r+\mu}$	$\mu > 0, y \geq 0$	$y \ln \frac{\mu}{r+\mu} + r \ln \frac{r}{r+\mu}$

Example: (Mice data) We fit a logistic regression model using a simplified GEE method with a working correlation matrix $\boldsymbol{C}(\rho)$ of AR1 (lag-1 autoregressive) type:

$$E(Y_i)=\mu_i=\frac{\exp(\eta_i)}{1+\exp(\eta_i)},\ \eta_i=\boldsymbol{x}_i'\boldsymbol{\beta},\ \mathrm{SE}(Y_i)=s_i=\sqrt{\mu_i(1-\mu_i)},\ \phi=1,$$

$$\hat{\rho}=\frac{1}{n-1}\sum_{i=1}^{n-1}\frac{(y_i-\mu_i)(y_{i+1}-\mu_{i+1})}{se(y_i)\,se(y_{i+1})},$$

$$D_{ij}=\frac{\partial\mu_i}{\partial\beta_j}=\frac{\partial}{\partial\beta_j}\frac{\exp(\eta_i)}{1+\exp(\eta_i)}=\mu_i(1-\mu_i)x_{ij}$$

$\boldsymbol{V}(\hat{\rho})=\boldsymbol{S}\,\boldsymbol{C}(\hat{\rho})\,\boldsymbol{S}$ (the working corr. matrix $\boldsymbol{C}$ is AR1 instead of $\boldsymbol{I}$)

$$=\begin{pmatrix} s_1 & 0 & 0 & \dots & 0\\ 0 & s_2 & 0 & \dots & 0\\ 0 & 0 & s_3 & \dots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \dots & s_n \end{pmatrix}\begin{pmatrix} 1 & \hat{\rho} & \hat{\rho}^2 & \dots & \hat{\rho}^{n-1}\\ \hat{\rho} & 1 & \hat{\rho} & \dots & \hat{\rho}^{n-2}\\ \hat{\rho}^2 & \hat{\rho} & 1 & \dots & \hat{\rho}^{n-3}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \hat{\rho}^{n-1} & \hat{\rho}^{n-2} & \hat{\rho}^{n-3} & \dots & 1 \end{pmatrix}\begin{pmatrix} s_1 & 0 & 0 & \dots & 0\\ 0 & s_2 & 0 & \dots & 0\\ 0 & 0 & s_3 & \dots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \dots & s_n \end{pmatrix}$$

```
> x=c(0:25)/10
> y=c(rep(0,5),1,rep(0,4),1,0,1,0,0,rep(1,11))
> n=length(x)
> p=2
> iter=10
> one=c(rep(1,n))
> X=cbind(one,x)
> results=matrix(0,iter,6)
> beta=matrix(0.1,2,1)  #starting values
>
> for (k in 1:iter){
+ eta=X%*%beta
+ mu=exp(eta)/(1+exp(eta))
+ se=sqrt(mu*(1-mu))
+ S=matrix(0,n,n)
+ D=matrix(0,n,p)
```

```
+
+ C=diag(one)
+ rhov=c(rep(0,n-1))
+
+ for (i in 1:n) {S[i,i]=se[i]}
+
+ for (i in 1:n)
+ { for (j in 1:p) { D[i,j]=se[i]^2*X[i,j] } }  #D matrix
+
+ for (i in 1:n-1)
+ { rhov[i]=(y[i]-mu[i])*(y[i+1]-mu[i+1])/(se[i]*se[i+1]) }  #rho hat
+ rho=sum(rhov)/(n-1)  #rho hat
+
+ for (i in 1:n)
+ { for (j in 1:n) { C[i,j]=rho^(abs(i-j)) } } #working corr. C (AR1)
+
+ V=S%*%C%*%S  #working cov matrix
+ VI=solve(V)
+ z=y-mu
+ DVID=t(D)%*%VI%*%D
+ DVIDI=solve(DVID)
+ DVIZ=t(D)%*%VI%*%z
+ beta=beta+DVIDI%*%DVIZ
+ VA=diag(DVIDI)
+ se=sqrt(VA)
+ names(se) = NULL
+ results[k,]=c(k,beta[1],se[1],beta[2],se[2],rho)
+ }
> results
      [,1]      [,2]      [,3]     [,4]      [,5]        [,6]
 [1,]    1 -2.275958 1.0896443 1.948302 0.7407660  0.39060747
 [2,]    2 -3.403097 0.9914335 2.944301 0.7328844 -0.04969473
 [3,]    3 -3.977545 1.2287111 3.461214 0.9590696 -0.11222065
 [4,]    4 -4.100869 1.4015116 3.573096 1.1205824 -0.12621644
 [5,]    5 -4.105567 1.4431492 3.577451 1.1594105 -0.12807356
 [6,]    6 -4.105575 1.4447882 3.577462 1.1609613 -0.12812993
```

```
 [7,]    7 -4.105575 1.4447918 3.577462 1.1609656 -0.12812974
 [8,]    8 -4.105575 1.4447918 3.577462 1.1609656 -0.12812974
 [9,]    9 -4.105575 1.4447918 3.577462 1.1609656 -0.12812974
[10,]   10 -4.105575 1.4447918 3.577462 1.1609656 -0.12812974
```

as compared to $\hat{a} = -4.111$ and $\hat{b} = 3.581$ with independence assumption. The changes are small due to the weak correlation $\hat{\rho} = -0.128$.

§2.8 Generalized linear mixed model (GLMM)

§2.8.1 Random effects model

Consider a 1-way ANOVA model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

$i = 1, \ldots, m;\ j = 1, \ldots, n_i$ where $\alpha_1, \ldots, \alpha_m$ are independent observations from a population of factor effects. We assume $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$. Testing $H_0 : \alpha_i = 0$ is the same as testing $H_0 :\ \sigma_\alpha^2 = 0$, that is there is no variability in the factor population.
Note that

$$E(Y_{ij}) = \mu \ \text{ and } \ \text{Var}(Y_{ij}) = \sigma_\alpha^2 + \sigma^2$$

and σ^2 and σ_α^2 are called the components of variance. Further the rv Y_{ij} are dependent and

$$\text{Cov}(Y_{ij}, Y_{ij'}) = \sigma_\alpha^2.$$

Assume that $n_i = n$ and $\sum_{i=1}^{m} n_i = N$. The normality assumption for α_i needs to be justified in practice. Under the assumption,

$$\frac{SSR}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{m} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\cdot})^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{\epsilon_{ij} - \bar{\epsilon}_{i\cdot}}{\sigma} \right)^2 \sim \chi^2_{N-m},$$

$$SSA = \sum_{i=1}^{m} n(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{m} n[\alpha_i + \bar{\epsilon}_{i\cdot} - (\bar{\alpha} + \bar{\epsilon}_{\cdot\cdot})]^2 = \sum_{i=1}^{m} n(\alpha_i + \bar{\epsilon}_{i\cdot})^2.$$

Hence

$$\frac{SSA}{\sigma^2+n\sigma_\alpha^2}=\sum_{i=1}^{m}\frac{n(\alpha_i+\bar{\epsilon}_{i\cdot})^2}{n(\sigma_\alpha^2+\sigma^2/n)}=\sum_{i=1}^{m}\frac{(\alpha_i+\bar{\epsilon}_{i\cdot})^2}{\sigma_\alpha^2+\sigma^2/n}\sim\chi^2_{m-1}$$

since $\bar{\alpha}=\bar{\epsilon}_{\cdot\cdot}=0$ and $\alpha_i+\bar{\epsilon}_{i\cdot}\sim\mathcal{N}(0,\sigma_\alpha^2+\sigma^2/n)$. Then

$$\begin{aligned} f_0 &= \frac{SSA/(m-1)}{SSR/(N-m)}=\frac{(\sigma^2+n\sigma_\alpha^2)(\frac{SSA}{\sigma^2+n\sigma_\alpha^2})/(m-1)}{\sigma^2(\frac{SSR}{\sigma^2})/(N-m)} \\ &= \left(\frac{\sigma^2+n\sigma_\alpha^2}{\sigma^2}\right)\frac{\chi^2_{m-1}/(m-1)}{\chi^2_{N-m}/(N-m)}\sim\left(1+\frac{n\sigma_\alpha^2}{\sigma^2}\right)F_{m-1,N-m}. \end{aligned}$$

Hence the classical F-test is appropriate for testing $H_0:\ \sigma_\alpha^2=0$.

The above ideas can be extended to higher order ANOVA models and linear models. It is possible to have a mixed model where some factors are random and others fixed.

§2.8.2 Generalized random effect model

For mixed effect models, consider

$$^{n}\boldsymbol{Y}^{1} = {}^{n}\boldsymbol{X}^{p}\boldsymbol{\beta}^{1} + {}^{n}\boldsymbol{Z}^{q}\boldsymbol{u}^{1} + {}^{n}\boldsymbol{\epsilon}^{1}$$

where $\boldsymbol{X}$ is a $n \times p$ design matrix and $\boldsymbol{Z}$ is a $n \times q$ design matrix, $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ independently of $\boldsymbol{u} \sim \mathcal{N}_q(\boldsymbol{0}, \sigma^2\boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is positive definite.

The common method for estimating $\boldsymbol{\beta}$ and predicting $\boldsymbol{u}$ is the *Best Linear Unbiased Predictor* (BLUP) procedure which produces estimates that maximize the joint likelihood of $\boldsymbol{y}$ and $\boldsymbol{u}$ whereas the ML procedure produces estimates that maximize the marginal likelihood of $\boldsymbol{y}$ after integrating out the unobserved $\boldsymbol{u}$. The joint likelihood of $\boldsymbol{\epsilon}$ and $\boldsymbol{u}$ is

$$\begin{aligned} & L(\boldsymbol{\theta}) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(\frac{-1}{2\sigma^2}\boldsymbol{\epsilon}'\boldsymbol{\epsilon}\right) (2\pi\sigma^2)^{-\frac{q}{2}} |\boldsymbol{\Sigma}|^{\frac{-1}{2}} \exp\left(\frac{-1}{2\sigma^2}\boldsymbol{u}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u}\right) \\ &= (2\pi\sigma^2)^{-\frac{n+q}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{\frac{-1}{2\sigma^2}[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) + \boldsymbol{u}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u}]\right\} \end{aligned}$$

To maximize $L(\boldsymbol{\theta})$, we need to minimize the sum of squares (SS)

$$SS = \boldsymbol{Y}'\boldsymbol{Y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{u} - 2\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{Y}'\boldsymbol{Z}\boldsymbol{u} + 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u}$$

The normal equations are

$$\begin{aligned} \frac{\partial SS}{\partial \boldsymbol{\beta}} &= 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{u} = \boldsymbol{0}, \\ \frac{\partial SS}{\partial \boldsymbol{u}} &= 2\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{u} - 2\boldsymbol{Z}'\boldsymbol{Y} + 2\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{\beta} + 2\boldsymbol{\Sigma}^{-1}\boldsymbol{u} = \boldsymbol{0}. \end{aligned}$$

Note that $\dfrac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{a}'\boldsymbol{\beta} = \boldsymbol{a}$, $\dfrac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{\beta}'\boldsymbol{A}\boldsymbol{\beta} = 2\boldsymbol{A}\boldsymbol{\beta}$, and $\boldsymbol{A}' = \boldsymbol{A}$. The BLUP equations are

$$\boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{X}'\boldsymbol{Z}\tilde{\boldsymbol{u}} = \boldsymbol{X}'\boldsymbol{Y} \tag{20}$$

$$\boldsymbol{Z}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} + (\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{\Sigma}^{-1})\tilde{\boldsymbol{u}} = \boldsymbol{Z}'\boldsymbol{Y}, \tag{21}$$

that is

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}' \\ \boldsymbol{Z}' \end{pmatrix} \boldsymbol{Y}$$

Alternatively we can write

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\delta}, \quad \boldsymbol{\delta} = \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2(\boldsymbol{I} + {}^{\mathrm{n}}\boldsymbol{Z}^{\mathrm{q}}\boldsymbol{\Sigma}^{\mathrm{q}}\boldsymbol{Z}'^{\mathrm{n}})).$$

The best estimator for $\boldsymbol{\beta}$ is then the WLS estimator $\widehat{\boldsymbol{\beta}}$ where

$$\boldsymbol{X}'(\boldsymbol{I}+\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}')^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'(\boldsymbol{I}+\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}')^{-1}\boldsymbol{Y}.$$

We will show that $\tilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}$. Multiply (21) by $\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}$ and subtract from (20),

$$\begin{aligned}
& \boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{X}'\boldsymbol{Z}\tilde{\boldsymbol{u}} - \boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}[\boldsymbol{Z}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} + (\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})\tilde{\boldsymbol{u}}] = \\
& \boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{Z}'\boldsymbol{Y}) \\
\Rightarrow\ & \boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{X}'\boldsymbol{Z}\tilde{\boldsymbol{u}} - \boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} - \\
& \boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})\tilde{\boldsymbol{u}} = \boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{Z}'\boldsymbol{Y}) \\
\Rightarrow\ & \boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{X}'\boldsymbol{Z}\tilde{\boldsymbol{u}} - \boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}'\boldsymbol{Z}\tilde{\boldsymbol{u}} = \\
& \boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{Z}'\boldsymbol{Y}) \\
\Rightarrow\ & \boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{Y} \\
\Rightarrow\ & \boldsymbol{X}'[\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{Z}']\boldsymbol{X}\tilde{\boldsymbol{\beta}} = \boldsymbol{X}'[\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{Z}']\boldsymbol{Y} \\
\Rightarrow\ & \boldsymbol{X}'(\boldsymbol{I}+\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}')^{-1}\boldsymbol{X}\tilde{\boldsymbol{\beta}} = \boldsymbol{X}'(\boldsymbol{I}+\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}')^{-1}\boldsymbol{Y}
\end{aligned}$$

Thus $\tilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}$, a WLS estimator since

$$\begin{aligned}
& (\boldsymbol{I}+\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}')[\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{Z}'] \\
= \ & \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}' - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{Z}' - \boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{Z}' \\
= \ & \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}' - \boldsymbol{Z}[(\boldsymbol{\Sigma}\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{I})(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}]\boldsymbol{Z}' \\
= \ & \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}' - \boldsymbol{Z}[\boldsymbol{\Sigma}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}]\boldsymbol{Z}' \\
= \ & \boldsymbol{I}
\end{aligned}$$

implies $\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{Z}' = (\boldsymbol{I}+\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}')^{-1}$.

Here $\tilde{\boldsymbol{\beta}}$ is BLUE (and hence BLUP). We can show that the linear combination of any other estimators

$$\text{Var}(\boldsymbol{b}'\widehat{\boldsymbol{\beta}}^* + \boldsymbol{c}'(\widehat{\boldsymbol{u}}^* - \boldsymbol{u})) \geq \text{Var}(\boldsymbol{b}'\widehat{\boldsymbol{\beta}} + \boldsymbol{c}'(\widehat{\boldsymbol{u}} - \boldsymbol{u}))$$

have larger variance in order to verify that the estimates are 'best' linear. The estimators are unbiased too, i.e. $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $E(\widehat{\boldsymbol{u}}) = \boldsymbol{0}$.

Note that the BLUP estimates for $\boldsymbol{u}$ are not the simple LS estimates from the model $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$, that is

$$\boldsymbol{Z}'\boldsymbol{Z}\tilde{\boldsymbol{u}} = \boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}).$$

From (21), we have

$$(\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{\Sigma}^{-1})\tilde{\boldsymbol{u}} = \boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \Rightarrow \tilde{\boldsymbol{u}} = (\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}).$$

This takes the variance of the $\boldsymbol{u}$ variables into account.

Summary of types of estimate:

Type	OLS	$\rightarrow$	WLS	$\rightarrow$	$IRLS$	$\rightarrow$
Cov	$\sigma^2\boldsymbol{I}$		$\sigma^2\boldsymbol{\Sigma}$ (diagonal)		$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{W}(\boldsymbol{\theta})^{-1}$	
RSS	$RSS = \sum_i \epsilon_i^2$		$RSS = \sum_i \epsilon_i^2$		$RSS = \sum_i \rho(\epsilon_i)$	
Est.	$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$		$(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$		$(\boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{Y}$	
Dist.	normal		normal		normal	

Type	WLS (MLE)	$\rightarrow$	GEE (QL)
	$z = \eta + (y - \mu)\frac{\partial \eta}{\partial \mu}$		$\boldsymbol{V}$ not diagonal
	$W = V^{-1}(\frac{\partial \eta}{\partial \mu})^{-2}$		$D = \frac{\partial \mu}{\partial \beta}$
Est. $\widehat{\boldsymbol{\beta}}^{(k+1)}$	$(\boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{z}^{(k)}$		$\widehat{\boldsymbol{\beta}}^{(k)} + (\boldsymbol{D}'\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$
Dist.	non-normal		non-normal

Example: (random effects) $n = 60$, $q = 10$, $p = 1$.

Group i	n_i	$\bar{y}_{i\cdot}$	**Observation** y_{ij}						
1	6	5.990	5.058	8.611	3.739	5.201	6.944	6.387	
2	6	4.843	3.901	6.735	3.030	6.539	5.508	3.345	
3	5	5.672	7.623	6.876	3.451	5.468	4.942		
4	5	3.946	2.405	5.007	3.153	2.856	6.309		
5	5	5.591	7.839	5.476	2.892	6.143	5.605		
6	7	4.287	3.957	1.025	5.463	3.399	6.285	5.280	4.602
7	6	6.564	9.057	6.325	6.394	3.464	7.541	6.603	
8	7	6.357	5.909	7.404	7.937	4.275	6.391	4.023	8.560
9	7	14.799	17.779	13.841	14.033	14.526	15.599	15.848	11.967
10	6	15.439	17.839	15.476	12.894	16.143	15.604	14.678	

The model is

$$Y_{ij} = \beta_0 + u_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2); \quad u_i \sim \mathcal{N}(0, \sigma^2)$$

u_i are independent, i.e. $\boldsymbol{\Sigma} = \boldsymbol{I}_{10}$.

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= -\frac{n+q}{2}\ln(2\pi\sigma^2) - \frac{1}{2}|\boldsymbol{\Sigma}| - \\
&\quad \frac{1}{2\sigma^2}[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) + \boldsymbol{u}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u}] \\
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{n+q}{2\sigma^2} + \frac{1}{2\sigma^4}[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) + \boldsymbol{u}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u}] \\
\widehat{\sigma}^2 &= \frac{1}{n+q}[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) + \boldsymbol{u}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u}]
\end{aligned}$$

```
> dat=read.csv("data/random.csv")
> dat
        y u1 u2 u3 u4 u5 u6 u7 u8 u9 u10
1   5.058  1  0  0  0  0  0  0  0  0   0
2   8.611  1  0  0  0  0  0  0  0  0   0
3   3.739  1  0  0  0  0  0  0  0  0   0
4   5.201  1  0  0  0  0  0  0  0  0   0
5   6.944  1  0  0  0  0  0  0  0  0   0
```

```
6   6.387  1  0  0  0  0  0  0  0  0   0
7   3.901  0  1  0  0  0  0  0  0  0   0
8   6.735  0  1  0  0  0  0  0  0  0   0
9   3.030  0  1  0  0  0  0  0  0  0   0
10  6.539  0  1  0  0  0  0  0  0  0   0
11  5.508  0  1  0  0  0  0  0  0  0   0
12  3.345  0  1  0  0  0  0  0  0  0   0
...
55 17.839  0  0  0  0  0  0  0  0  0   1
56 15.476  0  0  0  0  0  0  0  0  0   1
57 12.894  0  0  0  0  0  0  0  0  0   1
58 16.143  0  0  0  0  0  0  0  0  0   1
59 15.604  0  0  0  0  0  0  0  0  0   1
60 14.678  0  0  0  0  0  0  0  0  0   1
> y=dat[,1]
> n=length(y)
> q=10
> one=c(rep(1,q))
> X=c(rep(1,n))    #n by 1 vector
> Z=as.matrix(dat[,-1])    #n by q matrix
> S=diag(one)    #q by q identity matrix
> SI=solve(S)
> I=diag(X)
> IZ=I+Z%*%S%*%t(Z)
```

```
> IZ  #n by n block diagonal matrix  var(y)=s2+s2  cov(yij,yik)=s2 within cluster and
      #0 across clusters
   1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 ...
1  2 1 1 1 1 1 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
2  1 2 1 1 1 1 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
3  1 1 2 1 1 1 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
4  1 1 1 2 1 1 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
5  1 1 1 1 2 1 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
6  1 1 1 1 1 2 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
7  0 0 0 0 0 0 2 1 1  1  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
8  0 0 0 0 0 0 1 2 1  1  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
9  0 0 0 0 0 0 1 1 2  1  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
10 0 0 0 0 0 0 1 1 1  2  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
11 0 0 0 0 0 0 1 1 1  1  2  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
12 0 0 0 0 0 0 1 1 1  1  1  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

```
13 0 0 0 0 0 0 0 0 0  0  0  0  2  1  1  1  1  0  0  0  0  0  0  0  0  0  0  0
14 0 0 0 0 0 0 0 0 0  0  0  0  1  2  1  1  1  0  0  0  0  0  0  0  0  0  0  0
15 0 0 0 0 0 0 0 0 0  0  0  0  1  1  2  1  1  0  0  0  0  0  0  0  0  0  0  0
16 0 0 0 0 0 0 0 0 0  0  0  0  1  1  1  2  1  0  0  0  0  0  0  0  0  0  0  0
17 0 0 0 0 0 0 0 0 0  0  0  0  1  1  1  1  2  0  0  0  0  0  0  0  0  0  0  0
18 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  2  1  1  1  1  0  0  0  0  0  0
19 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  1  2  1  1  1  0  0  0  0  0  0
20 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  1  1  2  1  1  0  0  0  0  0  0
21 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  1  1  1  2  1  0  0  0  0  0  0
22 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  1  1  1  1  2  0  0  0  0  0  0
23 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  2  1  1  1  1  0
24 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  1  1  1  0
25 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  2  1  1  0
26 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  1  2  1  0
27 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  1  1  2  0
...

> IZ=solve(IZ)
> XIZY=t(X)%*%IZ%*%y
> XIZXI=solve(t(X)%*%IZ%*%X)
> beta=XIZXI%*%XIZY
> beta
         [,1]
[1,] 7.374953
> ZR=t(Z)%*%(y-X*beta) #sum of ui and eij for each cluster
> ZZSI=t(Z)%*%Z+solve(S)
> ZZSI #no of components
    u1 u2 u3 u4 u5 u6 u7 u8 u9 u10
u1   7  0  0  0  0  0  0  0  0   0
u2   0  7  0  0  0  0  0  0  0   0
u3   0  0  6  0  0  0  0  0  0   0
u4   0  0  0  6  0  0  0  0  0   0
u5   0  0  0  0  6  0  0  0  0   0
u6   0  0  0  0  0  8  0  0  0   0
u7   0  0  0  0  0  0  7  0  0   0
u8   0  0  0  0  0  0  0  8  0   0
u9   0  0  0  0  0  0  0  0  8   0
u10  0  0  0  0  0  0  0  0  0   7

> U=solve(ZZSI)%*%ZR #divide the sum of ui and eij by the no. of components
```

```
> U
           [,1]
u1  -1.1871023
u2  -2.1702452
u3  -1.4191272
u4  -2.8574606
u5  -1.4866272
u6  -2.7017086
u7  -0.6951023
u8  -0.8907086
u9   6.4960414
u10  6.9120406

> sum(U) #sum to zero
[1] -3.863576e-14
> Res=y-X%*%beta-Z%*%U
> sigma2=(t(Res)%*%Res+t(U)%*%SI%*%U)/(n+q) #sum over n eij^2 and q ui^2
> sigma2
[1,] 4.068185
```

Note:

1. The last two groups have very large random effect estimates.
2. The matrix $\boldsymbol{I}+\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}'$ shows that observations within a group *sharing the same random effect* are *correlated*.

§2.9 Appendix

To prove the result (7), differentiate (6) with respect to $\boldsymbol{\beta}$ and λ.

$$\begin{aligned} -2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{C}'\lambda &= \mathbf{0} \\ \boldsymbol{a} - \boldsymbol{C}\boldsymbol{\beta} &= \mathbf{0} \end{aligned} \tag{22}$$

Multiply (22) by $\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ to obtain

$$\begin{aligned} -2\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\lambda &= \mathbf{0} \\ -2\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\lambda &= \mathbf{0} \end{aligned}$$

Solving for λ,

$$\begin{aligned} -\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\lambda &= 2\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} - 2\boldsymbol{C}\boldsymbol{\beta} \\ \lambda &= -2[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}']^{-1}[\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{a}] \end{aligned} \tag{23}$$

since $\boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{a}$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \hat{\boldsymbol{\beta}}$. Substitute (23) into (22),

$$\begin{aligned} &-2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}'2[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}']^{-1}[\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{a}] = \mathbf{0} \\ &\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{C}'[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}']^{-1}[\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{a}] \\ &\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}']^{-1}[\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{a}] \\ &\quad = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}']^{-1}[\boldsymbol{a} - \boldsymbol{C}\hat{\boldsymbol{\beta}}] \end{aligned}$$

Reference

Huber, P.J. (1964) Robust estimation of a location parameter, *Annals of Mathematical Statistics*, **35**, 73-101.

Koenker, R.W. and Basset, G.J. (1978). Regression Quantiles, *Econometrica*, **46**, 33-50.