Report for Ovary data analysis

Xiaoqing Ye

Institute of Statistics and Big Data

April 8, 2019

Contents

```
> head(expr)
# A tibble: 6 x 17,816
  bcr_patient_bar…   ELMO2  CREB3L1  RPS11  PNMA1   MMP2 C10orf90   ZHX3  ERCC5  GPR98 RXFP3  APBB2
  <chr>              <dbl>    <dbl>  <dbl>  <dbl>  <dbl>    <dbl>  <dbl>  <dbl>  <dbl> <dbl>  <dbl>
1 TCGA-01-0628-11… -0.248  -0.142    1.37   1.09  -1.24     -1.96 -0.738 0.270  -1.09  0.583 -1.19
2 TCGA-01-0639-11…  0.0951 -0.0465   0.983  1.39  -0.808    -1.61 -0.652 0.425  -1.78  0.524 -1.61
3 TCGA-01-0642-11… -0.145   0.036    0.866  1.12  -0.316    -2.00 -0.517 0.380  -1.38  0.382 -1.08
4 TCGA-04-1331-01… -0.620  -0.00325  0.567  1.29  -0.328    -2.64 -0.821 0.631  -0.766 0.222  0.840
5 TCGA-04-1332-01…  0.185   1.01     0.968  1.16   0.416    -2.75  0.099 0.0612 -1.4   0.249  0.572
6 TCGA-04-1335-01… -0.672   1.22    -0.234 -0.842 -1.11     -1.66 -0.186 0.384  -1.73  1.08   0.158
# … with 17,804 more variables: PRO0478 <dbl>, KLHL13 <dbl>, PRSSL1 <dbl>, PDCL3 <dbl>,
#   DECR1 <dbl>, SALL1 <dbl>, CADM4 <dbl>, RPS18 <dbl>, HNRPD <dbl>, CFHR5 <dbl>, SLC10A7 <dbl>,
#   OR2K2 <dbl>, LMAN1 <dbl>, SUHW1 <dbl>, CHD8 <dbl>, SUMO1 <dbl>, GP1BA <dbl>, DDB1 <dbl>,
#   MYO9B <dbl>, MMP7 <dbl>, CRNKL1 <dbl>, C9orf45 <dbl>, XAB2 <dbl>, RTN1 <dbl>, KLHL14 <dbl>,
#   TBX10 <dbl>, CENPQ <dbl>, UTY <dbl>, ZBTB12 <dbl>, DTNBP1 <dbl>, KBTBD8 <dbl>, ZEB1 <dbl>,
#   ZG16 <dbl>, MIER1 <dbl>, ADAM5P <dbl>, CHD9 <dbl>, STK16 <dbl>, KIAA1486 <dbl>, TOB2 <dbl>,
#   BANK1 <dbl>, OR2V2 <dbl>, GRM2 <dbl>, PROSC <dbl>, SPIN2B <dbl>, PIR <dbl>, IPO9 <dbl>,
#   EVC <dbl>, CXCL13 <dbl>, KIAA1199 <dbl>, SORL1 <dbl>, NAT10 <dbl>, CHD1 <dbl>, SYN3 <dbl>,
#   SLC22A2 <dbl>, SERPINF1 <dbl>, WDR34 <dbl>, OR7A17 <dbl>, C9orf11 <dbl>, RNF216L <dbl>,
#   LHB <dbl>, STK25 <dbl>, TAOK3 <dbl>, LOC152573 <dbl>, C3orf39 <dbl>, C14orf108 <dbl>,
#   CDC25B <dbl>, BMP3 <dbl>, TMEM180 <dbl>, MAP1LC3C <dbl>, CRYGC <dbl>, POU3F1 <dbl>,
#   C20orf32 <dbl>, CCDC95 <dbl>, HIGD1B <dbl>, USP6NL <dbl>, ABCD4 <dbl>, DIMT1L <dbl>, TEK <dbl>,
#   SLC25A46 <dbl>, LARP7 <dbl>, CD160 <dbl>, MT1JP <dbl>, PHF20 <dbl>, CPNE4 <dbl>, GTPBP1 <dbl>,
#   RAB33B <dbl>, ALDOC <dbl>, ZNF212 <dbl>, NUDT1 <dbl>, RFPL2 <dbl>, ZNF83 <dbl>, GDPD5 <dbl>,
#   PDCD4 <dbl>, CEP350 <dbl>, OR10A2 <dbl>, CST7 <dbl>, CIAO1 <dbl>, SELL <dbl>, OR8J3 <dbl>,
#   LTBP4 <dbl>, …
```

Figure: 1

```
> head(expr)
# A tibble: 6 x 17,816
  bcr_patient_bar…   ELMO2  CREB3L1  RPS11  PNMA1   MMP2 C10orf90   ZHX3  ERCC5  GPR98 RXFP3  APBB2
  <chr>              <dbl>    <dbl>  <dbl>  <dbl>  <dbl>    <dbl>  <dbl>  <dbl>  <dbl> <dbl>  <dbl>
1 OV1              -0.248   -0.142    1.37   1.09  -1.24     -1.96 -0.738 0.270  -1.09  0.583 -1.19
2 OV2               0.0951  -0.0465   0.983  1.39  -0.808    -1.61 -0.652 0.425  -1.78  0.524 -1.61
3 OV3              -0.145    0.036    0.866  1.12  -0.316    -2.00 -0.517 0.380  -1.38  0.382 -1.08
4 OV4              -0.620   -0.00325  0.567  1.29  -0.328    -2.64 -0.821 0.631  -0.766 0.222  0.840
5 OV5               0.185    1.01     0.968  1.16   0.416    -2.75  0.099 0.0612 -1.4   0.249  0.572
6 OV6              -0.672    1.22    -0.234 -0.842 -1.11     -1.66 -0.186 0.384  -1.73  1.08   0.158
# … with 17,804 more variables: PRO0478 <dbl>, KLHL13 <dbl>, PRSSL1 <dbl>, PDCL3 <dbl>,
#   DECR1 <dbl>, SALL1 <dbl>, CADM4 <dbl>, RPS18 <dbl>, HNRPD <dbl>, CFHR5 <dbl>, SLC10A7 <dbl>,
#   OR2K2 <dbl>, LMAN1 <dbl>, SUHW1 <dbl>, CHD8 <dbl>, SUMO1 <dbl>, GP1BA <dbl>, DDB1 <dbl>,
#   MYO9B <dbl>, MMP7 <dbl>, CRNKL1 <dbl>, C9orf45 <dbl>, XAB2 <dbl>, RTN1 <dbl>, KLHL14 <dbl>,
#   TBX10 <dbl>, CENPQ <dbl>, UTY <dbl>, ZBTB12 <dbl>, DTNBP1 <dbl>, KBTBD8 <dbl>, ZEB1 <dbl>,
#   ZG16 <dbl>, MIER1 <dbl>, ADAM5P <dbl>, CHD9 <dbl>, STK16 <dbl>, KIAA1486 <dbl>, TOB2 <dbl>,
#   BANK1 <dbl>, OR2V2 <dbl>, GRM2 <dbl>, PROSC <dbl>, SPIN2B <dbl>, PIR <dbl>, IPO9 <dbl>,
#   EVC <dbl>, CXCL13 <dbl>, KIAA1199 <dbl>, SORL1 <dbl>, NAT10 <dbl>, CHD1 <dbl>, SYN3 <dbl>,
#   SLC22A2 <dbl>, SERPINF1 <dbl>, WDR34 <dbl>, OR7A17 <dbl>, C9orf11 <dbl>, RNF216L <dbl>,
#   LHB <dbl>, STK25 <dbl>, TAOK3 <dbl>, LOC152573 <dbl>, C3orf39 <dbl>, C14orf108 <dbl>,
#   CDC25B <dbl>, BMP3 <dbl>, TMEM180 <dbl>, MAP1LC3C <dbl>, CRYGC <dbl>, POU3F1 <dbl>,
#   C20orf32 <dbl>, CCDC95 <dbl>, HIGD1B <dbl>, USP6NL <dbl>, ABCD4 <dbl>, DIMT1L <dbl>, TEK <dbl>,
#   SLC25A46 <dbl>, LARP7 <dbl>, CD160 <dbl>, MT1JP <dbl>, PHF20 <dbl>, CPNE4 <dbl>, GTPBP1 <dbl>,
#   RAB33B <dbl>, ALDOC <dbl>, ZNF212 <dbl>, NUDT1 <dbl>, RFPL2 <dbl>, ZNF83 <dbl>, GDPD5 <dbl>,
#   PDCD4 <dbl>, CEP350 <dbl>, OR10A2 <dbl>, CST7 <dbl>, CIAO1 <dbl>, SELL <dbl>, OR8J3 <dbl>,
#   LTBP4 <dbl>, …
```

Figure: 2

```
> expr[1:5, (dim(expr)[2] - 2): dim(expr)[2]]
# A tibble: 5 x 3
     AQP7    CTSC dataset
    <dbl>   <dbl> <chr>
1 -0.0005 -0.0925 OV.mRNA
2 -0.190   0.196  OV.mRNA
3  0.068  -0.106  OV.mRNA
4 -0.22   -0.990  OV.mRNA
5 -0.310   0.681  OV.mRNA
```

Figure: 2

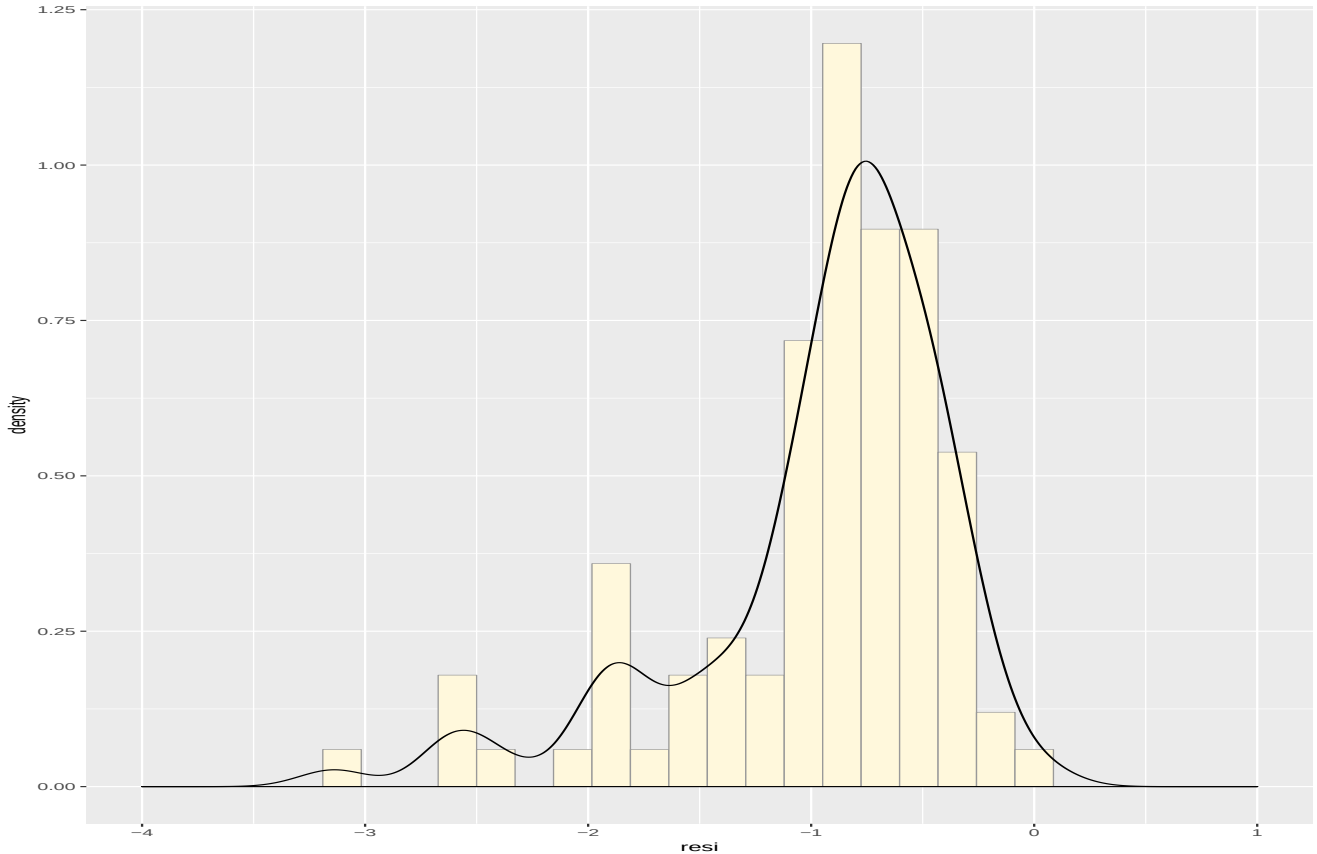

Figure: 4

Assumed that $p \gg n$, we establish the subject-specified linear model

$$y_i = \mu_i + \mathrm{x}_i^T \beta + \epsilon_i, i = 1, \ldots, n$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p-$dimension vector of the parameters. And ϵ_i is error term independent with covariates x_i with

$$\mathbb{E}(\epsilon_i) = 0, \mathbb{V}(\epsilon_i) = \sigma^2$$

.

Applied MCP[Zhang, 2010] to the diffrences of intercepts $\mu_i - \mu_j, 1 \leqslant i <\leqslant j$and the parameters $\beta_k, k = 1, \ldots, p$, we can obtain

$$Q_n(\mu, \beta; \lambda, \omega) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu_i - \mathrm{x}_i^T\beta)^2 + \sum_{1\leqslant i<j\leqslant n} p_\gamma(|\mu_i - \mu_j|, \lambda) + \sum_{k=1}^{p} p_\gamma(|\beta_k|, \omega)$$

where

$$p_\gamma(t, \lambda) = \lambda \int_0^t (1 - \frac{x}{\gamma\lambda})_+ dx, \gamma > 1$$

And the sign $(x)_+$ means that

$$(x)_+ = \begin{cases} x, x > 0 \\ 0, x \leqslant 0 \end{cases}$$

Define

$$\eta_{ij} = \mu_i - \mu_j$$

Denote

$$\eta = \{\eta_{ij}, i < j\}^T$$

Then

$$\min S_n(\mu, \beta, \eta; \lambda, \omega) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu_i - \mathrm{x}_i^T\beta)^2 + \sum_{1 \leqslant i < j \leqslant n} p_\gamma(|\eta_{ij}|, \lambda) + \sum_{k=1}^{p} p_\gamma(|\beta_k|, \omega)$$

$$s.t.\ \mu_i - \mu_j = \eta_{ij}$$

Trough Augmented Lagrange multiplier method, we can obtain

$$L(\mu, \beta, \eta, \nu) = S_n(\mu, \eta, \beta) + \sum_{1 \leqslant i < j \leqslant n} \nu_{ij}(\mu_i - \mu_j - \eta_{ij}) + \frac{\vartheta}{2} \sum_{1 \leqslant i < j \leqslant n} (\mu_i - \mu_j - \eta_{ij})^2$$

where ϑ is the penalty parameter and $\nu = \{\nu_{ij}\, i < j\}^T$ are Lagrange multipliers.

Compute $\hat{\eta}_{ij}$

$$L(\mu, \eta, \beta, \nu) = \sum_{1 \leqslant i < j \leqslant n} \left\{ \frac{\vartheta}{2}(\mu_i - \mu_j + \vartheta^{-1}\nu_{ij} - \eta_{ij})^2 + p_\gamma(|\eta_{ij}|, \lambda) \right\} + C_1$$

Therefore

$$\frac{\vartheta}{2}(\mu_i - \mu_j + \vartheta^{-1}\nu_{ij} - \eta_{ij})^2 + p_\gamma(|\eta_{ij}|, \lambda)$$

Let

$$\delta_{ij} = \mu_i - \mu_j + \vartheta^{-1}\nu_{ij}$$

Denote

$$\delta = \{\delta_{ij}, i < j\}^T$$

We can obtain given $\delta^{(m+1)}$ at the $m-$step, we can update $\eta^{(m+1)}$ whose elements are the minimizer of (2.7) as

$$\eta^{(m+1)} = \begin{cases} \frac{ST(\delta^{(m+1)}, \lambda/\vartheta)}{1-1/\gamma\vartheta}, |\delta^{(m+1)}| \leqslant \gamma\lambda \\ \delta^{(m+1)}, |\delta^{(m+1)}| > \gamma\lambda \end{cases}$$

where $ST(t, \lambda) = sign(t)(|t| - \lambda)_+$ is the soft-therosholding function. And given $\nu^{(m)}$ at the $m-$step and$\mu^{(m+1)}$ at the $(m+1)-$step, the update $\delta^{(m+1)}$ is

$$\delta^{(m+1)} = \Delta\mu^{(m+1)} - \frac{1}{\vartheta}\nu^{(m)}$$

where$\Delta = \{e_i - e_j, i < j\}^T$, e_i is a n-dimension vector with the $i-$th component 1 and others 0.

Compute $\hat{\mu}$

$$L(\mu,\eta,\beta,\nu) = \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu_i-\mathrm{x}_i^T\beta)^2 + \sum_{1\leqslant i<j\leqslant n}\nu_{ij}(\mu_i-\mu_j-\eta_{ij}) + \frac{\vartheta}{2}\sum_{1\leqslant i<j\leqslant n}(\mu_i-\mu_j-\eta_{ij})^2 + C_2$$

Then

$$\begin{aligned} L(\mu,\eta,\beta,\nu) &= \frac{1}{2}\|y-\mu-X\beta\|^2 + \frac{\vartheta}{2}\sum_{1\leqslant i<j\leqslant n}(\mu_i-\mu_j-\eta_{ij}+\vartheta^{-1}\nu_{ij})^2 + C_3 \\ &= \frac{1}{2}\|y-\mu-X\beta\|^2 + \frac{\vartheta}{2}\sum_{1\leqslant i<j\leqslant n}\left\{(e_i-e_j)\mu-\eta_{ij}+\vartheta^{-1}\nu_{ij}\right\}^2 + C_3 \\ &= \frac{1}{2}\|y-\mu-X\beta\|^2 + \frac{\vartheta}{2}\|\Delta\mu-\eta+\vartheta^{-1}\nu\|^2 + C_3 \end{aligned}$$

Let

$$\frac{\partial L(\mu, \eta, \beta, \nu)}{\partial \mu} = -(y - \mu - X\beta) + \vartheta \Delta^T (\Delta\mu - \eta + \vartheta^{-1}\nu) = 0$$

Given $\eta^{(m)}, \nu^{(m)}, \beta^{(m)}$ at the $m-$step, we obtain

$$\mu^{(m+1)} = (E_{n\times n} + \vartheta \Delta^T \Delta)^{-1} (\vartheta \Delta^T \eta^{(m)} - \Delta^T \nu^{(m)} + y - X\beta^{(m)})$$

where $E_{n\times n}$ is n-dimension unit matrix.

Compute $\hat{\beta}$

$$L(\mu, \eta, \beta, \nu) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu_i - \mathrm{x}_i^T\beta)^2 + \sum_{k=1}^{p} p_\gamma(|\beta_k|, \omega) + C_4$$
$$= \frac{1}{2}\|y - \mu - X\beta\|^2 + \sum_{k=1}^{p} p_\gamma(|\beta_k|, \omega) + C_4$$

Let $\tilde{y} = y - \mu$, then (10) is equivalent to

$$L(\mu, \eta, \beta, \nu) = \frac{1}{2}\|\tilde{y} - X\beta\|^2 + \sum_{l=1}^{p} p_\gamma(|\beta_l|, \omega) + C_4$$

The solution of (2.11) is given by Patrick B.(2011), that is, given $\mu^{(m+1)}$ at the $(m+1)-$step, the update is

$$\beta^{(m+1)} = \begin{cases} \frac{ST(n^{-1}X^T\tilde{y}^{(m+1)},\omega)}{1-1/\gamma}, |n^{-1}X^T\tilde{y}^{(m+1)}| \leqslant \gamma\omega \\ n^{-1}X^T\tilde{y}^{(m+1)}, |n^{-1}X^T\tilde{y}^{(m+1)}| > \gamma\omega \end{cases}$$

where $\tilde{y}^{(m+1)} = y - \mu^{(m+1)}$. And this can be estimated in the ncvreg R-package by Breheny P. and Huang J. (2011).

Stopping criterion

Define

$$D_{\mu}^{(m+1)} = \left\{ (\Delta\mu^{(m+1)} - \eta^{(m+1)})^T (\Delta\mu^{(m+1)} - \eta^{(m+1)}) \right\}^{\frac{1}{2}}$$

$$D_{\beta}^{(m+1)} = \left\{ (\beta^{(m+1)} - \beta^{(m)})^T (\beta^{(m+1)} - \beta^{(m)}) \right\}^{\frac{1}{2}}$$

Let $\|s\|_w$ be the weighted average norm by weighting the values of norms applied to every elements of s. In this paper, we terminate the iteraction when $\|D^{(m+1)}\|_w$ is less than a pre-defined tolerance, where $D^{(m+1)} = (D_{\mu}^{(m+1)}, D_{\beta}^{(m+1)})$,
and

$$\|D^{(m+1)}\|_w = \omega_0 \|D_{\mu}^{(m+1)}\| + \omega_1 \|D_{\beta}^{(m+1)}\|, \omega_0 + \omega_1 = 1$$

.

Algorithm

By ADMM, we can obtain the iteration function of ν is

$$\nu^{(m+1)} = \nu^{(m)} + \vartheta(\Delta\mu_{(m+1)} - \eta^{(m+1)})$$

Therefore, the algorithm is as follows.

Selection of tunig parameters

$$BIC(\lambda,\omega) = \log\left[\frac{\sum_{i=1}^{n}\left(y_i - \hat{\mu}_i(\lambda,\omega) - \mathrm{x}_i^T\hat{\beta}(\lambda,\omega)\right)^2}{n}\right] + C_n\frac{\log n}{n}\left(\hat{K}(\lambda,\omega) + p\right)$$

where C_n is a positive numbers depending on sample size n. With $C_n = 1$, the revised BIC is indeed the traditional BIC. Wang, LI, and Leng(2009) suggests

$$C_n = c\log(\log(n+p))$$

where c is a positive number. The pair (λ,ω) minimized $BIC(\lambda,\omega)$ is applied to the model fit.

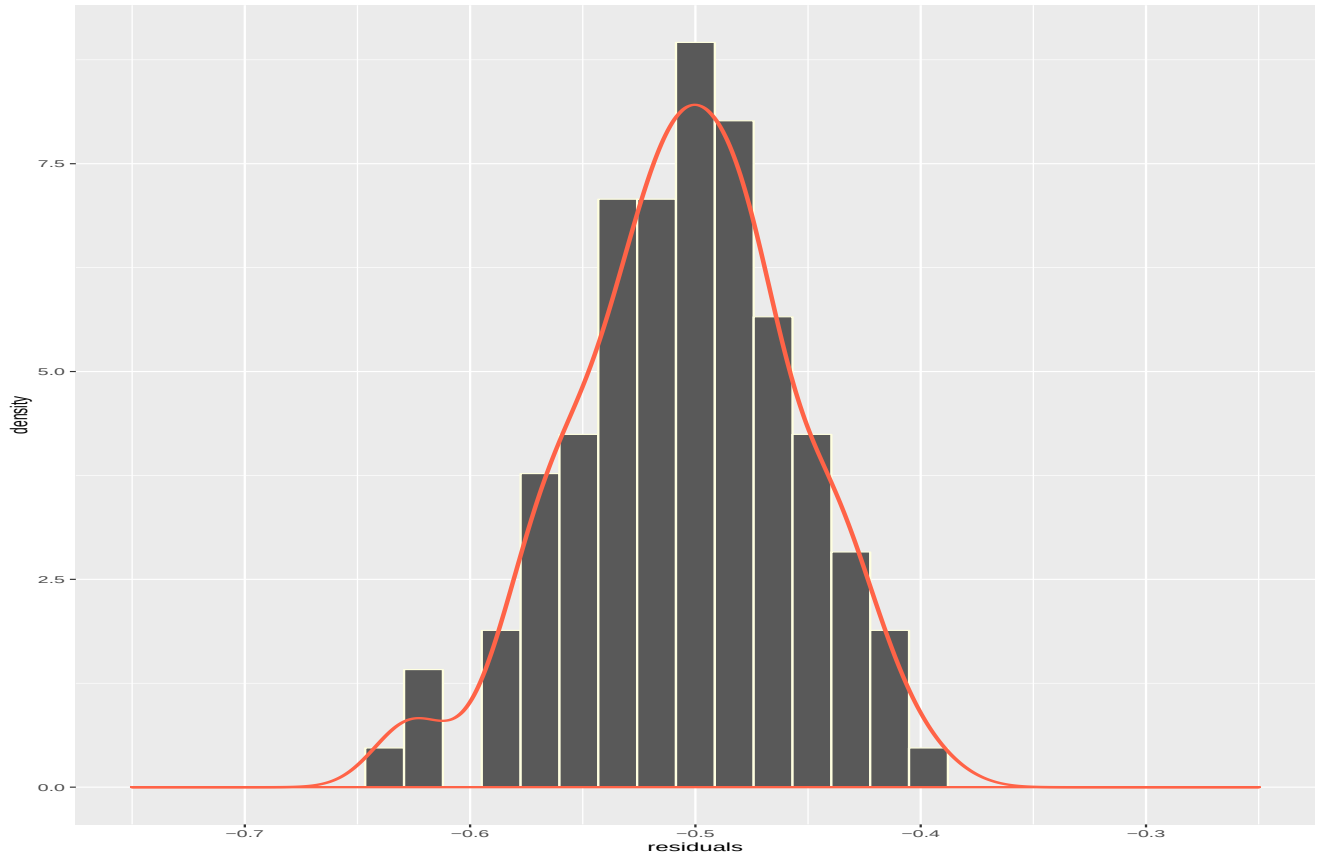

Figure: group 1

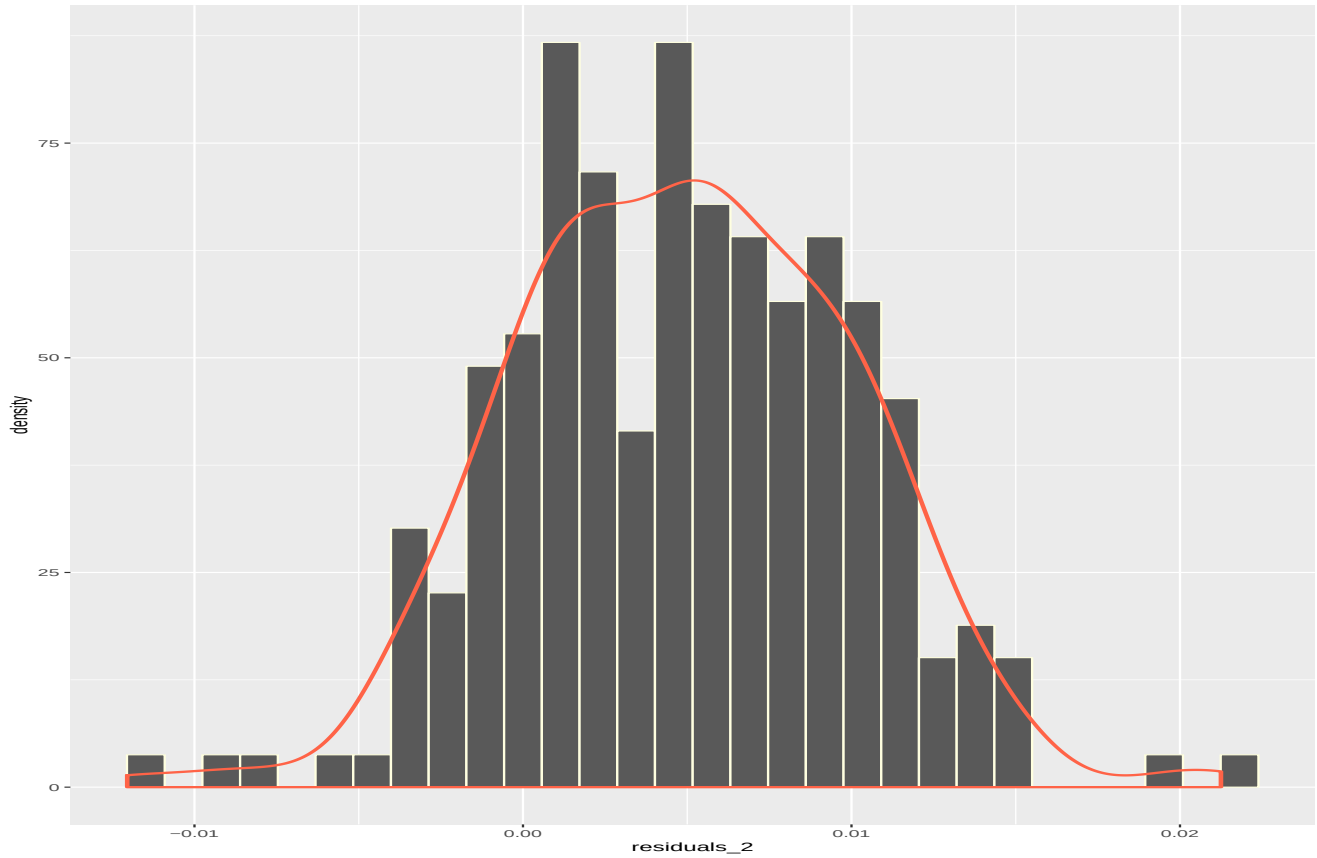

Figure: group 2

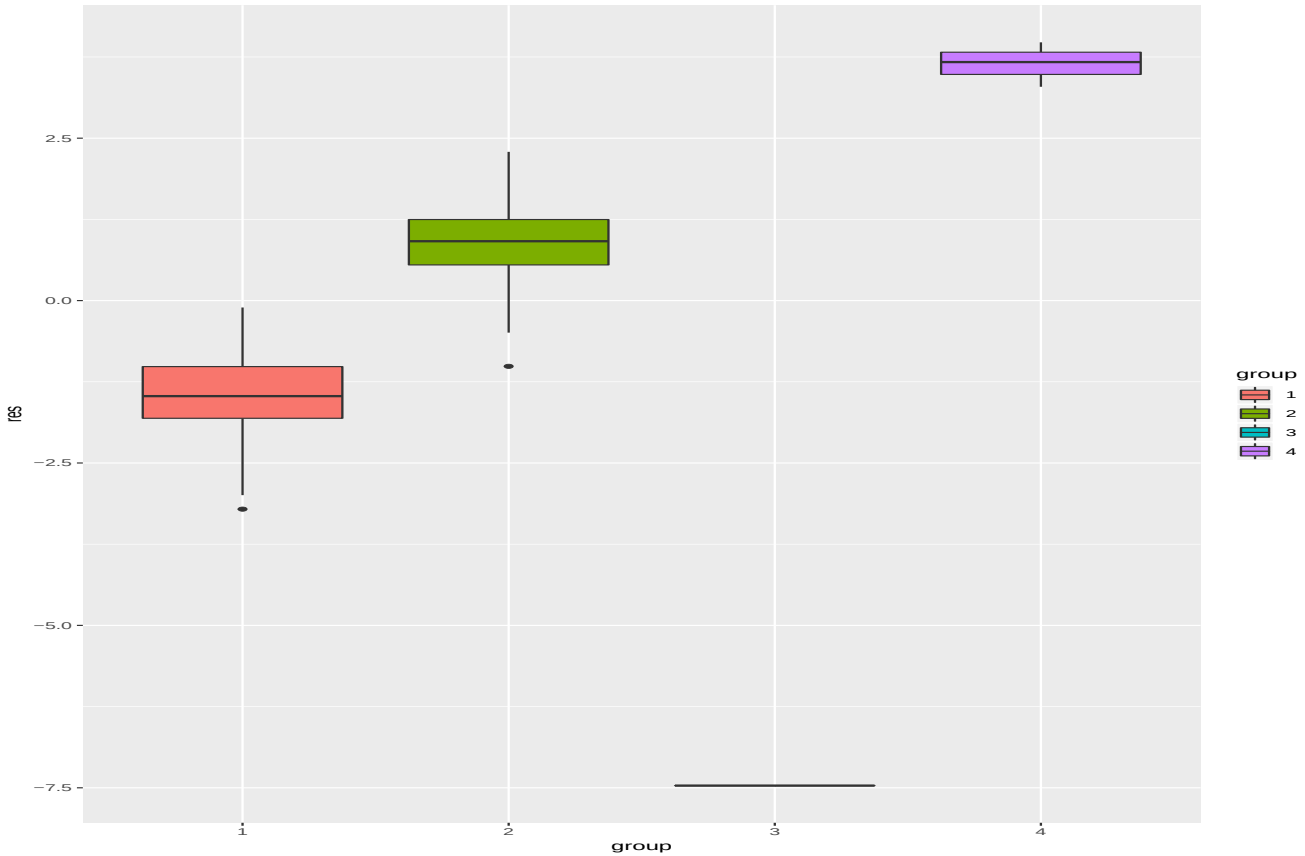

Figure: Boxplot of the divided groups

Thanks

Thank you!