Technical details on linear regression for proteomics

Lieven Clement

Statistical Genomics

1. Linear Regression

- Consider a vector of predictors $\mathbf{x} = (x_1, \ldots, x_{p-1})$ and
- a real-valued response Y
- then the linear regression model can be written as

$$Y = f(\mathbf{x}) + \epsilon = \beta_0 + \sum_{j=1}^{p-1} x_j \beta + \epsilon$$

with i.i.d. $\epsilon \sim N(0, \sigma^2)$

- n observations $(\mathbf{x}_1, y_1) \ldots (\mathbf{x}_n, y_n)$
- Regression in matrix notation

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$$

with $\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{np-1} \end{bmatrix}$, $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$

and $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$

1.1 Least Squares (LS)

- Minimize the residual sum of squares

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \end{aligned}$$

- or in matrix notation

$$\begin{aligned} RSS(\beta) &= (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) \\ &= \|\mathbf{Y} - \mathbf{X}\beta\|^2 \end{aligned}$$

with the L_2-norm of a p-dim. vector v $\|\mathbf{v}\| = \sqrt{v_1^2 + \ldots + v_p^2}$

$\rightarrow$ $\hat{\beta} = \text{argmin}_{\beta}\|\mathbf{Y} - \mathbf{X}\beta\|^2$

Minimize RSS

$$\begin{aligned}
\frac{\partial RSS}{\partial \boldsymbol{\beta}} &= \mathbf{0} \\
\frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \mathbf{0} \\
-2\mathbf{X}^T(\mathbf{Y}-\mathbf{X}\beta) &= \mathbf{0} \\
\mathbf{X}^T\mathbf{X}\beta &= \mathbf{X}^T\mathbf{Y} \\
\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}
\end{aligned}$$

```r
data<-readRDS("heartProtQ92736.rds")
fit <- lm(exprs~location+patient,data,x=TRUE)

head(fit$x,4)
```

```
##     (Intercept) locationLV locationRA locationRV patient4 pat
## LA3           1          0          0          0        0
## LA4           1          0          0          0        1
## LA8           1          0          0          0        0
## LV3           1          1          0          0        0
```

The model matrix can also be obtained without fitting the model:

```
X<-model.matrix(~location+patient,data)
head(X,4)
```

```
##     (Intercept) locationLV locationRA locationRV patient4 pat
## LA3           1          0          0          0        0
## LA4           1          0          0          0        1
## LA8           1          0          0          0        0
## LV3           1          1          0          0        0
```

```r
fit$coefficient
```

```
## (Intercept)  locationLV  locationRA  locationRV    patient4
## 27.50063357 -3.40997017  0.36748910  1.44473120  0.08573147 -
```

```r
sigma(fit)
```

```
## [1] 0.7812888
```

Variance Estimator?

$$
\begin{aligned}
\hat{\Sigma}_{\hat{\boldsymbol{\beta}}} &= \text{var}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\right] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{var}\left[\mathbf{Y}\right]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}\sigma^2)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{I}\quad\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2
\end{aligned}
$$

1.2 Contrasts

When we assess a contrast we assess a linear combination of model parameters:

$$H_0 : \mathbf{L^T}\beta = 0 \text{ vs } H_1 : \mathbf{L^T}\beta \neq 0$$

Estimator of Contrast?

$$\mathbf{L}^T\hat{\boldsymbol{\beta}}$$

Variance?

$$\boldsymbol{\Sigma}_{\mathbf{L}\hat{\boldsymbol{\beta}}} = \mathbf{L}^T\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}\mathbf{L}$$

1.3 Inference

- When the assumptions of the linear model hold

$$\hat{\boldsymbol{\beta}} \sim MVN\left[\boldsymbol{\beta}, \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma^2\right]$$

- Hence,

$$\mathbf{L}^T\hat{\boldsymbol{\beta}} \sim MVN\left[\mathbf{L}^T\boldsymbol{\beta}, \mathbf{L}^T\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma^2\right]\mathbf{L}\right]$$

- We estimate σ^2 by MSE

$$\hat{\sigma}^2 = \frac{\mathbf{e}^T\mathbf{e}}{n-p} \rightarrow \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\hat{\sigma}^2$$

- Statistic

$$\mathbf{F} = \hat{\boldsymbol{\beta}}^T\mathbf{L}\left(\mathbf{L}^T\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\mathbf{L}\right)^{-1}\mathbf{L}^T\hat{\boldsymbol{\beta}} \underset{H_0}{\sim} F_{r,n-p}$$

  follows an F distribution with r and n-p degrees of freedom under $H_0 : \mathbf{L}^T\hat{\boldsymbol{\beta}} = \mathbf{0}$

- Note, that r equals the number of contrasts or the rank of the contrast matrix

When we test one contrast at the time (e.g. the k^{th} contrast) the statistic reduces to

$$T = \frac{\mathbf{L}_k^T \hat{\beta}}{\sqrt{\left(\mathbf{L}_k^T \hat{\mathbf{\Sigma}}_{\hat{\beta}} \mathbf{L}_k\right)}} \underset{H_0}{\sim} t_{n-p}$$

follows a t distribution with n-p degrees of freedom under $H_0 : \mathbf{L}_k^T \hat{\beta} = 0$

```r
summary(fit)
```

```
##
## Call:
## lm(formula = exprs ~ location + patient, data = data, x = TRU
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8118 -0.3572 -0.1021  0.2641  1.0142
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.50063    0.55245  49.779 4.41e-09 ***
## locationLV  -3.40997    0.63792  -5.345  0.00175 **
## locationRA   0.36749    0.63792   0.576  0.58551
## locationRV   1.44473    0.63792   2.265  0.06413 .
## patient4     0.08573    0.55245   0.155  0.88177
## patient8    -0.31303    0.55245  -0.567  0.59152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 0.7813 on 6 degrees of freedom
```

```r
library(multcomp)
L<-matrix(0,nrow=length(fit$coefficient),ncol=2)
rownames(L)<-names(fit$coefficient)
L[2,1]<-1
L[3:4,2]<-c(-1,1)
L
```

```
##             [,1] [,2]
## (Intercept)    0    0
## locationLV     1    0
## locationRA     0   -1
## locationRV     0    1
## patient4       0    0
## patient8       0    0
```

```r
fit %>% glht(linfct=t(L)) %>% summary
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = exprs ~ location + patient, data = data, x
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
## 1 == 0  -3.4100     0.6379  -5.345  0.00335 **
## 2 == 0   1.0772     0.6379   1.689  0.25064
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
## (Adjusted p values reported -- single-step method)
```

2. Robust regression

- No normality assumption needed
- Robust fit minimises the maximal bias of the estimators
- CI and statistical tests are based on asymptotic theory
- If ϵ is normal, the M-estimators have a high efficiency!
- ordinary least squares (OLS): minimize loss function

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$$

- M-estimation: minimize loss function

$$\sum_{i=1}^{n} \rho \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right)$$

with

- ρ is symmetric, i.e. $\rho(z) = \rho(-z)$
- ρ has a minimum at $\rho(0) = 0$, is positive for all $z \neq 0$
- $\rho(z)$ increases as $|z|$ increases

The estimator $\hat{\mu}$ is also the solution to the equation

$$\sum_{i=1}^{n} \Psi(y_i - \mathbf{x}_i \boldsymbol{\beta}) = 0,$$

where Ψ is the derivative of ρ. For $\hat{\beta}$ possessing the robustness property, Ψ should be bounded.

Example: least squares

$\rho(z) = z^2$, and thus $\Psi(z) = 2z$ (unbounded!). $\hat{\boldsymbol{\beta}}$ is the solution of

$$\sum_{i=1}^{n} 2\mathbf{x}_i(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) = 0 \text{ or } \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$$

with $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_G]^T$

When a location and a scale parameter, say σ, have to be estimated simultaneously, we write

$$(\hat{\boldsymbol{\beta}}, \hat{\sigma}) = \mathrm{ArgMin}_{\boldsymbol{\beta},\sigma} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma}\right) \text{ and } \sum_{i=1}^{n} \Psi\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma}\right) = 0.$$

Define $u_i = \frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma}$. The last estimation equation is equivalent to

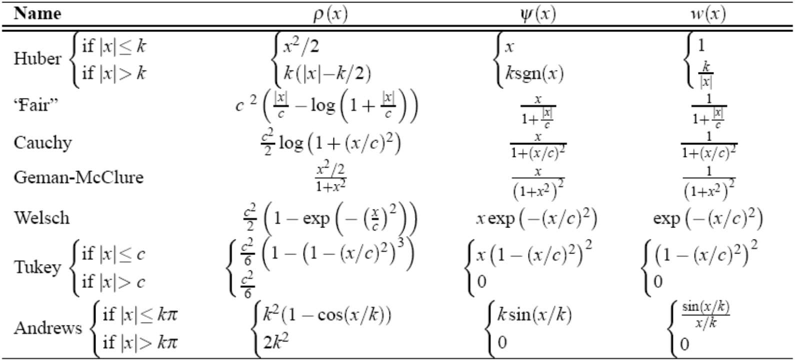
$$\sum_{i=1}^{n} w(u_i) u_i = 0,$$

with weight function $w(u) = \Psi(u)/u$. This is the typical form that appears when solving the *iteratively reweighted least squares problem*,

$$(\hat{\boldsymbol{\beta}}, \hat{\sigma}) = \mathrm{ArgMin}_{\mu,\sigma} \sum_{i=1}^{n} w(u_i^{(k-1)}) \left(u_i^{(k)}\right)^2,$$

where k represents the iteration number.

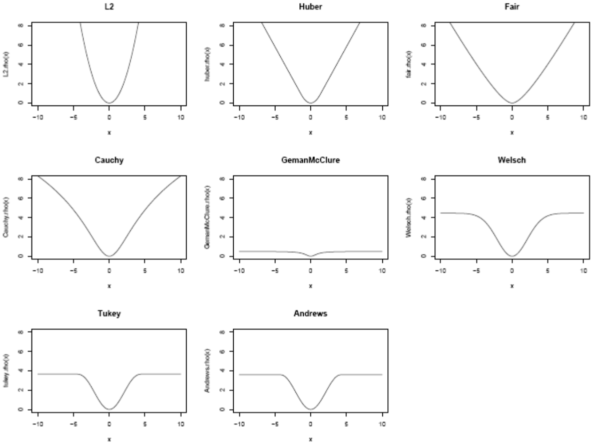
Some Examples of Robust Functions}



Name	$\rho(x)$	$\psi(x)$	$w(x)$
Huber $\begin{cases}\text{if } \lvert x\rvert \le k \\ \text{if } \lvert x\rvert > k\end{cases}$	$\begin{cases} x^2/2 \\ k(\lvert x\rvert - k/2)\end{cases}$	$\begin{cases} x \\ k\,\mathrm{sgn}(x)\end{cases}$	$\begin{cases} 1 \\ \frac{k}{\lvert x\rvert}\end{cases}$
‘Fair”	$c^{\,2}\left(\frac{\lvert x\rvert}{c} - \log\left(1+\frac{\lvert x\rvert}{c}\right)\right)$	$\frac{x}{1+\frac{\lvert x\rvert}{c}}$	$\frac{1}{1+\frac{\lvert x\rvert}{c}}$
Cauchy	$\frac{c^2}{2}\log\left(1+(x/c)^2\right)$	$\frac{x}{1+(x/c)^2}$	$\frac{1}{1+(x/c)^2}$
Geman-McClure	$\frac{x^2/2}{1+x^2}$	$\frac{x}{\left(1+x^2\right)^2}$	$\frac{1}{\left(1+x^2\right)^2}$
Welsch	$\frac{c^2}{2}\left(1-\exp\left(-\left(\frac{x}{c}\right)^2\right)\right)$	$x\exp\left(-(x/c)^2\right)$	$\exp\left(-(x/c)^2\right)$
Tukey $\begin{cases}\text{if } \lvert x\rvert \le c \\ \text{if } \lvert x\rvert > c\end{cases}$	$\begin{cases}\frac{c^2}{6}\left(1-\left(1-(x/c)^2\right)^3\right) \\ \frac{c^2}{6}\end{cases}$	$\begin{cases} x\left(1-(x/c)^2\right)^2 \\ 0\end{cases}$	$\begin{cases}\left(1-(x/c)^2\right)^2 \\ 0\end{cases}$
Andrews $\begin{cases}\text{if } \lvert x\rvert \le k\pi \\ \text{if } \lvert x\rvert > k\pi\end{cases}$	$\begin{cases} k^2(1-\cos(x/k)) \\ 2k^2\end{cases}$	$\begin{cases} k\sin(x/k) \\ 0\end{cases}$	$\begin{cases}\frac{\sin(x/k)}{x/k} \\ 0\end{cases}$

PhD thesis Bolstad 2004

The ρ functions



PhD thesis Bolstad 2004

Common Ψ-Functions

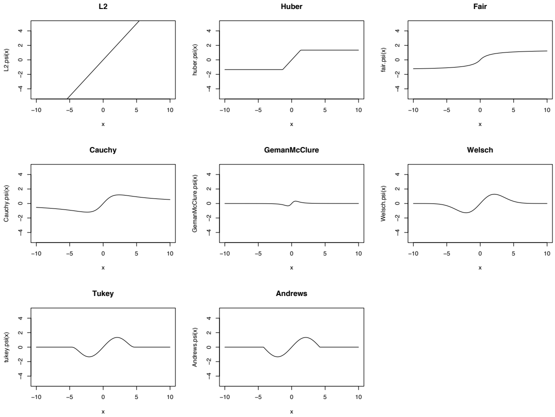


Figure 4.2: The ψ functions for some common M-estimators.

PhD thesis Bolstad 2004

Corresponding Weight Functions

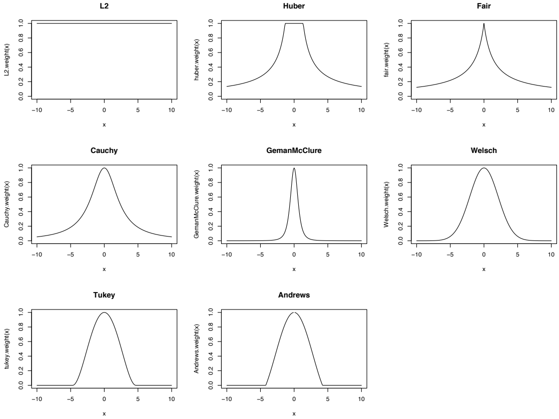
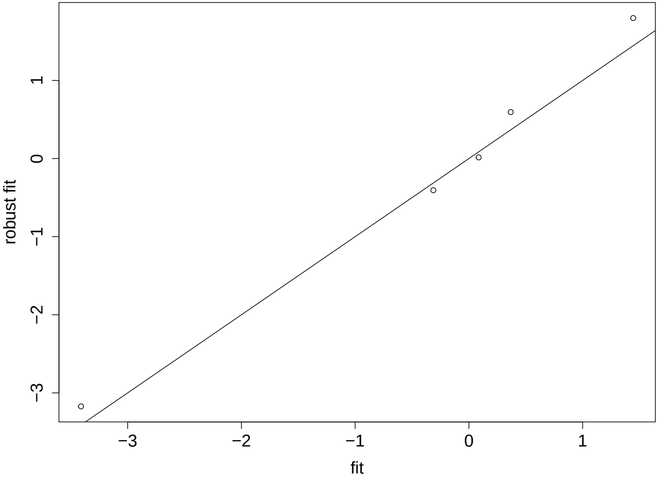


Figure 4.3: The weight functions for some common M-estimators.

PhD thesis Bolstad 2004

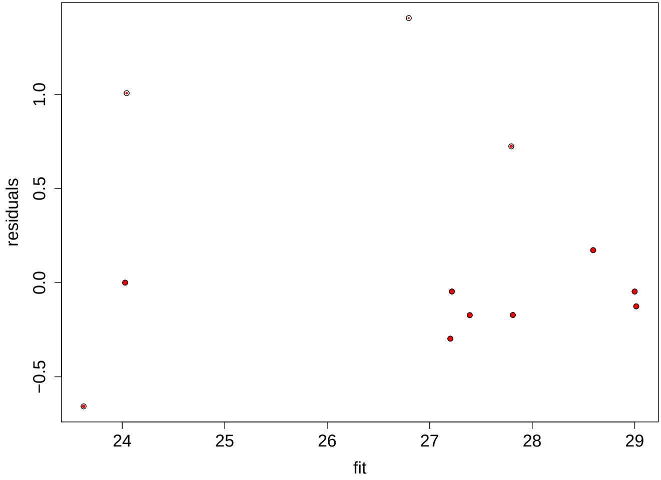
```
library("MASS")
rfit <- rlm(exprs~location+patient,data,maxit=500)
plot(fit$coefficient[-1],rfit$coefficient[-1],xlab="fit",ylab="robust fit",cex.axis=1.5,cex.lab=1.5)
abline(0,1)
```



```
rfit$w
```

```
##  [1] 1.0000000 1.0000000 0.2448895 1.0000000 0.3418904 0.5239307 0.4754051
##  [8] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
```

```
plot(rfit$fitted,rfit$res,cex=rfit$w,pch=19,col=2,cex.lab=1.5,cex.axis=1.5,ylab="residuals",xlab="fit")
points(rfit$fitted,rfit$res)
```



```r
summary(fit)
```

```
##
## Call:
## lm(formula = exprs ~ location + patient, data = data, x = TRU
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8118 -0.3572 -0.1021  0.2641  1.0142
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.50063    0.55245  49.779 4.41e-09 ***
## locationLV  -3.40997    0.63792  -5.345  0.00175 **
## locationRA   0.36749    0.63792   0.576  0.58551
## locationRV   1.44473    0.63792   2.265  0.06413 .
## patient4     0.08573    0.55245   0.155  0.88177
## patient8    -0.31303    0.55245  -0.567  0.59152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 0.7813 on 6 degrees of freedom
```

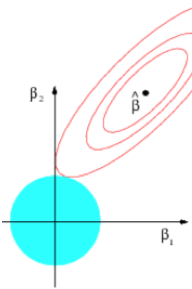
```r
summary(rfit)
```

```
##
## Call: rlm(formula = exprs ~ location + patient, data = data,
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65730 -0.17198 -0.04697  0.31060  1.40606
##
## Coefficients:
##             Value   Std. Error t value
## (Intercept) 27.2010  0.4518    60.2081
## locationLV  -3.1727  0.5217    -6.0817
## locationRA   0.5947  0.5217     1.1400
## locationRV   1.7986  0.5217     3.4478
## patient4     0.0150  0.4518     0.0333
## patient8    -0.4052  0.4518    -0.8970
##
## Residual standard error: 0.256 on 6 degrees of freedom
```

3. Penalized regression: ridge

1. Ridge penalty
2. Parameter estimation of ridge regression
3. Link between ridge regression and mixed models

3.1. Ridge Penalty



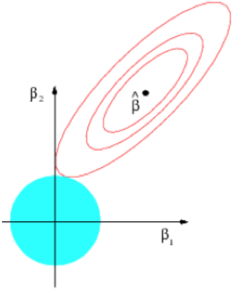
Hastie et al. 2008

- Add a ridge penalty

$$\hat{\beta} = \text{argmin}_{\beta} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2 \right\}$$

- λ: penalty parameter that controls the amount of penalisation

3.1. Ridge Penalty



Hastie et al. 2008

- Add a ridge penalty

$$\hat{\beta} = \text{argmin}_{\beta} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2 \right\}$$

- λ: penalty parameter that controls the amount of penalisation

- Equivalent to

$$\hat{\beta} = \text{argmin}_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|^2 \text{ subject to } \|\beta\|^2 \leq s$$

- Note, that s has a one-to-one correspondence with λ

3.2. Closed form solution

$$\hat{\beta}^{\text{ridge}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \right\}$$

Matrix form

- Let $\mathbf{D} = \left[\begin{array}{cc} 0 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & \mathbf{I}_{p\times p} \end{array} \right]$, which allows the criterion to be written in matrix form and to leave the intercept β_0 unpenalized.

$$\hat{\beta}^{\text{ridge}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \beta^T \mathbf{D} \beta \right\}$$

Minimization:

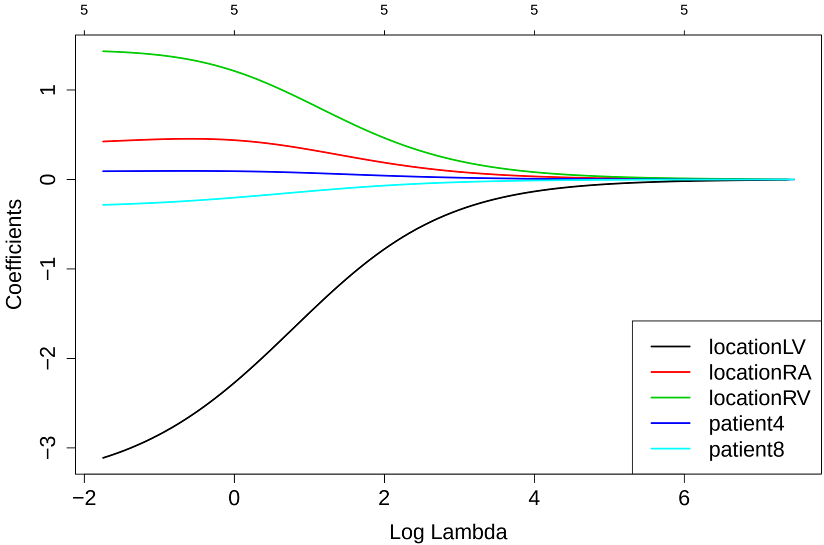
$$\frac{d \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \beta^T \mathbf{D} \beta \right\}}{d\boldsymbol{\beta}} = 0$$

$$\Leftrightarrow -\mathbf{X}^{\mathbf{T}}\mathbf{Y} + \mathbf{X}^{T}\mathbf{X}\beta + \lambda\mathbf{D}\beta = 0$$

$$\Leftrightarrow (\mathbf{X}^{T}\mathbf{X} + \lambda\mathbf{D})\beta = \mathbf{X}^{T}\mathbf{Y}$$

$$\Leftrightarrow \hat{\beta} = (\mathbf{X}^{T}\mathbf{X} + \lambda\mathbf{D})^{-1}\mathbf{X}^{T}\mathbf{Y}$$

```
library(glmnet)
ridgeFit<-glmnet(fit$x[,-1],data$exprs,family="gaussian", alpha=
plot(ridgeFit,xvar="lambda")
legend("bottomright",legend=colnames(fit$x)[-1],col=1:5,lty=1,ce
```


5
5
5
5
5
Coefficients
1
0
−1
−2
−3
−2
0
2
4
6
Log Lambda
locationLV
locationRA
locationRV
patient4
patient8

3.3 Tune ridge penalties

Tune the ridge penalties by exploiting the link between ridge regression and Mixed Models:

$$y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i$$

with

- $\beta_j \sim N\left(0, \frac{\sigma^2}{\lambda}\right)$
- $\epsilon_i \sim N\left(0, \sigma^2\right)$
- Variance components can be estimated using lme4 mixed model software and the predictions of the random effects $\boldsymbol{\beta}_j$ coincide with solution of ridge estimator.

Best linear unbiased predictor: BLUP

Optimize the joint log-likelihood $L(\mathbf{Y}, \beta)$ towards β

$$L(\mathbf{Y}, \beta) = \prod_{i=1}^{n} f(y_i|\beta) f(\beta)$$

Best linear unbiased predictor: BLUP

Optimize the joint log-likelihood $L(\mathbf{Y}, \beta)$ towards β

$$L(\mathbf{Y}, \beta) = \prod_{i=1}^{n} f(y_i|\beta) f(\beta)$$

$$-2l(\mathbf{Y}, \beta) \propto n \log(\sigma^2) + \frac{(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)}{\sigma^2} + p \log \frac{\sigma^2}{\lambda} + \frac{\lambda}{\sigma^2}\beta^T\beta$$

$$\begin{aligned} \hat{\beta} &= \text{argmin}_{\beta} \{ l(\mathbf{Y}, \beta) \} \\ &= \text{argmin}_{\beta} \left\{ ||\mathbf{Y} - \beta||_2^2 + \lambda \beta^T \boldsymbol{\beta} \right\} \end{aligned}$$

```r
library(lme4)
ridgeFit<-lmer(exprs~(1|location)+(1|patient),data)
```

```
## boundary (singular) fit: see ?isSingular
```

```r
summary(ridgeFit)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: exprs ~ (1 | location) + (1 | patient)
##    Data: data
## 
## REML criterion at convergence: 35.9
## 
## Scaled residuals: 
##      Min       1Q   Median       3Q      Max 
## -1.64229 -0.43326 -0.09407  0.42918  1.30037 
## 
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  location (Intercept) 4.2367   2.0583  
##  patient  (Intercept) 0.0000   0.0000  
##  Residual             0.5019   0.7084  
```

```r
ranef(ridgeFit)
```

```
## $location
##    (Intercept)
## LA   0.3842645
## LV  -2.8961752
## RA   0.7377943
## RV   1.7741165
##
## $patient
##   (Intercept)
## 3           0
## 4           0
## 8           0
##
## with conditional variances for "location" "patient"
```

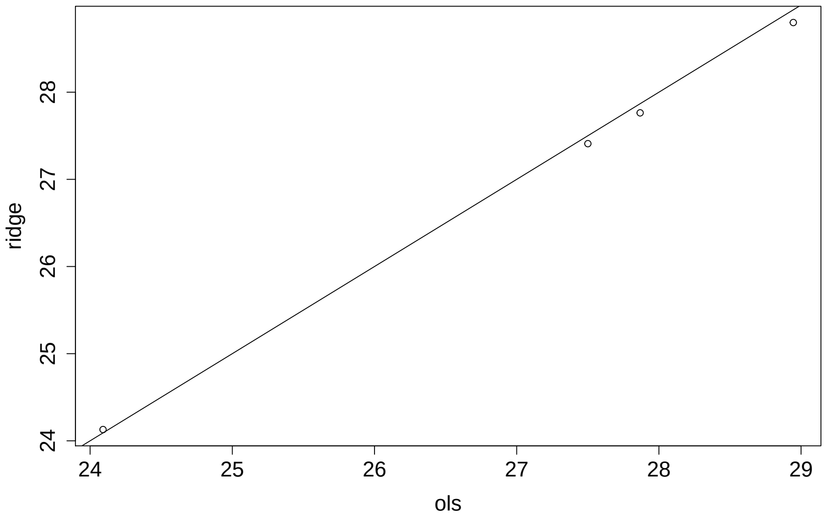
```
LG<-matrix(0,nrow=length(fit$coefficient),ncol=4)
rownames(LG)<-names(fit$coefficient)
LG[1,1]<-1
LG[c(1,2),2]<-1
LG[c(1,3),3]<-1
LG[c(1,4),4]<-1
sd(unlist(fit$coef%*%LG))
```

```
## [1] 2.09857
```

```
sd(unlist(fixef(ridgeFit)+ranef(ridgeFit)$location))
```

```
## [1] 2.018854
```

```
plot(unlist(fit$coef%*%LG),unlist(fixef(ridgeFit)+ranef(ridgeFit
abline(0,1)
```


ridge
ols