Ridge regresjon og lasso – notat til STK2120

Ørulf Borgan februar 2016

I dette notatet vil vi se litt nærmere på noen alternativer til minste kvadraters metode ved lineær regresjon. Metodene er særlig aktuelle når vi har mange forklaringsvariabler i forhold til antall observasjoner. Notatet er et supplement til det som står om lineær regresjon i avsnittene 12.7 og 12.8 i boka til Devore & Berk (D&B). Når ikke annet er sagt bruker vi notasjonen i denne boka.

1. Balansen mellom skjevhet og varians

Vi vil anta at vi har observasjoner $(Y_i, \mathbf{x}_i)$; $i = 1, 2, \ldots, n$. Her er Y_i-ene uavhengige og normalfordelte med varians σ^2, mens $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})'$ er vektorer av forklaringsvariabler svarende til Y_i; $i = 1, 2, \ldots, n$. Vi vil anta $\mathbf{x}_i$-ene er gitte vektorer (dvs. ikke stokastiske) og at $E(Y_i) = f(\mathbf{x}_i)$ for en funksjon f. [Men merk at i mange anvendelser vil $\mathbf{x}_i$-ene være observerte verdier av stokastiske vektorer $\mathbf{X}_i$. I slike situasjoner ser vi på den betingete forventningen til Y_i gitt $\mathbf{X}_i = \mathbf{x}_i$.] På grunnlag av observasjonene $(Y_i, \mathbf{x}_i)$ bestemmer vi en modell $\hat{f}(\mathbf{x}_0)$ som kan predikere en ny observasjon Y_0 ut fra den tilhørende vektoren av forklaringsvariabler $\mathbf{x}_0$. En slik modell kan for eksempel være en lineær regresjonsmodell, slik at $\hat{f}(\mathbf{x}_0) = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$. For å se hvor god modellen er til å predikere Y_0, kan vi se på (den kvadratiske) prediksjonsfeilen

$$\mathrm{PE}(\mathbf{x}_0) = E\left(\left[Y_0 - \hat{f}(\mathbf{x}_0)\right]^2\right). \tag{1}$$

Vi kan skrive om prediksjonsfeilen som følger [siden Y_0 er uavhengig av $\hat{f}(\mathbf{x}_0)$ og $EY_0 = f(\mathbf{x}_0)$]:

$$\begin{aligned}
\mathrm{PE}(\mathbf{x}_0) &= E\left(\left[\{Y_0 - EY_0\} + \{f(\mathbf{x}_0) - E\hat{f}(\mathbf{x}_0)\} + \{E\hat{f}(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\}\right]^2\right) \\
&= E\left(\{Y_0 - EY_0\}^2\right) + E\left(\{E\hat{f}(\mathbf{x}_0) - f(\mathbf{x}_0)\}^2\right) + E\left(\{\hat{f}(\mathbf{x}_0) - E\hat{f}(\mathbf{x}_0)\}^2\right) \\
&= \sigma^2 + \mathrm{Bias}^2\{\hat{f}(\mathbf{x}_0)\} + \mathrm{V}(\hat{f}(\mathbf{x}_0),
\end{aligned} \tag{2}$$

der $\sigma^2 = V(Y_0)$ og $\mathrm{Bias}\{\hat{f}(\mathbf{x}_0)\} = E\hat{f}(\mathbf{x}_0) - f(\mathbf{x}_0)$ er skjevheten for estimatoren $\hat{f}(\mathbf{x}_0)$. Av (2) ser vi at prediksjonsfeilen avhenger både av skjevheten og variansen til estimatoren $\hat{f}(\mathbf{x}_0)$. For å få best mulig prediksjon må vi derfor finne en passende balanse mellom skjevhet og varians. Det er altså ikke nødvendigvis best å bruke en forventningsrett estimator for $f(x_0)$. Vi kan få bedre prediksjon ved å bruke en estimator som er litt skjev hvis den har (tilstrekkelig mye) mindre varians. Det gir en motivasjon for å se på andre estimatorer for den multiple lineære regresjonsmodellen enn minste kavadraters estimator.

2. Ridge regresjon

Vi har den lineære regresjonsmodellen

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \tag{3}$$

der x_{ij}-ene er gitte forklaringsvariabler og ϵ_i-ene er uavhengige og $N(0, \sigma^2)$-fordelte stokastiske variabler. Vi vil i dette notatet anta at alle forklaringsvariablene er sentrert, dvs. at

$$\sum_{i=1}^{n} x_{ij} = 0 \quad \text{for} \quad j = 1, 2, \ldots, k. \tag{4}$$

Merk at vi kan sentrere forklaringsvariablene ved å trekke gjennomsnittet fra av hver av dem. Merk videre at en slik sentrering ikke endrer $\beta_1, \ldots, \beta_k$ i (3). Det er bare konstantledded β_0 som blir et annet når vi sentrerer forklaringsvariablene.

Minste kvadraters estimatorer for $\beta_0, \beta_1, \ldots \beta_k$ er de verdiene av $b_0, b_1, \ldots b_k$ som minimerer kvadratsummen (jf. side 683 i D&B)

$$g(b_0, b_1, \ldots, b_k) = \sum_{i=1}^{n} \left[Y_i - (b_0 + b_1 x_{i1} + \cdots + b_k x_{ik})\right]^2 . \tag{5}$$

Hvis det er stor korrelasjon mellom forklaringsvariablene og/eller vi har mange forklaringsvariabler i forhold til antall observasjoner, vi minste kvadraters estimatorer ha stor varians. Da kan det være bedre å bruke estimatorer for β_j-ene som ikke er forventningsrette, men som har mindre varians. En slik mulighet er ridge regresjon. I stedet for å minimere kvadratsummen (5), minimerer vi for ridge regresjon den straffete kvadratsummen

$$h(b_0, b_1, \ldots, b_k) = \sum_{i=1}^{n} \left[Y_i - (b_0 + b_1 x_{i1} + \cdots + b_k x_{ik})\right]^2 + \lambda \sum_{j=1}^{k} b_j^2, \tag{6}$$

der vi "straffer" store verdier av b_j-ene. (Merk at vi ikke straffer store verdier av b_0.) I (6) er $\lambda \geq 0$ en parameter som angir hvor stor straff vi får for store verdier av b_j-ene. Når vi deriverer (6) med hensyn på b_0 og b_j for $j = 1, \ldots, k$, får vi

$$\frac{\partial}{\partial b_0} h(b_0, b_1, \ldots, b_k) = -2 \sum_{i=1}^{n} \left[Y_i - (b_0 + b_1 x_{i1} + \cdots + b_k x_{ik})\right]$$

$$\frac{\partial}{\partial b_j} h(b_0, b_1, \ldots, b_k) = -2 \sum_{i=1}^{n} \left[Y_i - (b_0 + b_1 x_{i1} + \cdots + b_k x_{ik})\right] x_{ij} + 2\lambda b_j$$

Vi setter de partiellderiverte lik null, og bruker (4). Da får vi ligningene:

$$n b_0 = \sum_{i=1}^{n} Y_i \tag{7}$$

og

$$b_1 \sum_{i=1}^{n} x_{i1} x_{ij} + \cdots + b_k \sum_{i=1}^{n} x_{ik} x_{ij} + \lambda b_j = \sum_{i=1}^{n} x_{ij} Y_i \tag{8}$$

for $j = 1, \ldots, k$. Ridge estimatoren er løsningen av disse $k + 1$ ligningene.

Ridge estimatoren for β_0 er dermed

$$\hat{\beta}_0 = \overline{Y} = \frac{1}{n} = \sum_{i=1}^{n} Y_i,$$

mens ridge estimatoren for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ er gitt som løsningen av ligningssystemet

$$\left(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}\right) \mathbf{b} = \mathbf{X}'\mathbf{Y}. \tag{9}$$

Her er $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, $\mathbf{b} = (b_1, \ldots, b_k)'$, $\mathbf{I}$ er identitetsmatrisen av dimensjon $k \times k$ og

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}. \tag{10}$$

Merk at matrisen (10) ikke er den samme $\mathbf{X}$-matrisen som den som er gitt på side 706 i D&B. Forskjellen er at $\mathbf{X}$-matrisen i D&B har en kolonne med 1-tall, mens matrisen (10) ikke har det. Tilsvarende skriver vi nå vektorene $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ og $\mathbf{b} = (b_1, \ldots, b_k)'$ uten konstantleddene β_0 og b_0. Vi løser (9) og får rigde estimatoren

$$\hat{\boldsymbol{\beta}}_\lambda = \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1} \mathbf{X}'\mathbf{Y}. \tag{11}$$

Merk at estimatoren avhenger av verdien av λ. Hvordan vi kan velge λ, tar vi opp i avsnitt 5.

Når vi bruker minste kvadraters metode, vil en endring av skalaen for den j-te forklaringsvariabelen bli kompensert ved en tilsvarende endring av estimatet $\hat{\beta}_j$ slik at produktet $\hat{\beta}_j x_{ij}$ blir det samme. (Hvis vi for eksempel endrer skalaen for alder fra år til måneder, vil minste kvadraters estimat når vi bruker måneder som tidsskala være tolv ganger så stort som det estimatet vi får når vi bruker år som tidsskala.) Men for ridge estimatoren er det ikke slik. Derfor er det for ridge regresjon vanlig å standardisere alle forklaringsvariablene slik at $\sum_{i=1}^n x_{ij}^2 = 1$ for $j = 1, \ldots, k$.

3. Egenskaper til ridge estimatoren

Vi vil studere egenskapene til rigde estimatoren $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\beta}_{\lambda 1}, \ldots, \hat{\beta}_{\lambda k})'$. Av (11) ser vi at $\hat{\beta}_{\lambda j}$ er en lineærkombinasjon av Y_i-ene, så $\hat{\beta}_{\lambda j}$ er normalfordelt. For å finne forventning og varians, ser vi først på hvordan ridge estimatoren henger sammen med minste kvadraters estimator.

Minste kvadraters estimator er gitt ved:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Y}. \tag{12}$$

Merk at (12) er minste kvadraters estimator for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$, dvs. uten konstantledded β_0. Ved en liten modifikasjon av resonementene på sidene 711 og 712 i D&B har vi at $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ og $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$.

Vi kan nå omforme ridge estimatoren (11):

$$\begin{aligned} \hat{\boldsymbol{\beta}}_\lambda &= \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1} \left(\mathbf{X}'\mathbf{X}\right) \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Y} \\ &= \left[(\mathbf{X}'\mathbf{X})(\mathbf{I} + \lambda(\mathbf{X}'\mathbf{X})^{-1})\right]^{-1} \left(\mathbf{X}'\mathbf{X}\right) \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Y} \\ &= \left(\mathbf{I} + \lambda(\mathbf{X}'\mathbf{X})^{-1}\right)^{-1} \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Y} \\ &= \left(\mathbf{I} + \lambda(\mathbf{X}'\mathbf{X})^{-1}\right)^{-1} \hat{\boldsymbol{\beta}}. \end{aligned} \tag{13}$$

Av (13) får vi dermed at

$$E(\hat{\boldsymbol{\beta}}_\lambda) = \left(\mathbf{I} + \lambda(\mathbf{X}'\mathbf{X})^{-1}\right)^{-1} E(\hat{\boldsymbol{\beta}}) = \left(\mathbf{I} + \lambda(\mathbf{X}'\mathbf{X})^{-1}\right)^{-1} \boldsymbol{\beta}. \tag{14}$$

og

$$\begin{aligned} \mathrm{Cov}(\hat{\boldsymbol{\beta}}_\lambda) &= \left(\mathbf{I} + \lambda(\mathbf{X}'\mathbf{X})^{-1}\right)^{-1} \mathrm{Cov}(\hat{\boldsymbol{\beta}}) \left(\mathbf{I} + \lambda(\mathbf{X}'\mathbf{X})^{-1}\right)^{-1} \\ &= \sigma^2 \left(\mathbf{I} + \lambda(\mathbf{X}'\mathbf{X})^{-1}\right)^{-1} \left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\mathbf{I} + \lambda(\mathbf{X}'\mathbf{X})^{-1}\right)^{-1}. \end{aligned} \tag{15}$$

4. Et eksempel

Formelene (14) og (15) er nokså komplekse, så det er vanskelig å se hva de gir oss. Som en liten illustrasjon ser vi derfor på situasjonen det forklaringsvariablene er ortonormale, dvs. at $\sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, \ldots, k$ og $\sum_{i=1}^{n} x_{ij}x_{i\ell} = 0$ for $j \neq \ell$. Da har vi at $\mathbf{X}'\mathbf{X} = \mathbf{I}$, slik at

$$E(\hat{\boldsymbol{\beta}}_\lambda) = (1+\lambda)^{-1}\boldsymbol{\beta} \tag{16}$$

og

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_\lambda) = \sigma^2(1+\lambda)^{-2}\,\mathbf{I}. \tag{17}$$

Merk at vi får $E(\hat{\boldsymbol{\beta}})$ og $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ ved å sette $\lambda = 0$ i (16) og (17).

Vi er interessert i å predikere en ny observasjon Y_0 ut fra den tilhørende vektoren av forklaringsvariabler $\mathbf{x}_0$. Det kan vi gjøre ved å bruke minste kvadraters prediktor:

$$\hat{f}(\mathbf{x}_0) = \mathbf{x}_0'\hat{\boldsymbol{\beta}} \tag{18}$$

eller ved å bruke ridge prediktoren:

$$\hat{f}_\lambda(\mathbf{x}_0) = \mathbf{x}_0'\hat{\boldsymbol{\beta}}_\lambda\,. \tag{19}$$

For minste kvadraters prediktor (18) har vi

$$E\left(\hat{f}(\mathbf{x}_0\right) = \mathbf{x}_0'E(\hat{\boldsymbol{\beta}}) = \mathbf{x}_0'\boldsymbol{\beta}$$
$$\mathrm{V}\left(\hat{f}(\mathbf{x}_0\right) = \mathbf{x}_0'\mathrm{Cov}(\hat{\boldsymbol{\beta}})\mathbf{x}_0 = \sigma^2\mathbf{x}_0'\mathbf{x}_0$$

Av (2) får vi da prediksjonsfeilen

$$\mathrm{PE}(\mathbf{x}_0) = \sigma^2 + \sigma^2\mathbf{x}_0'\mathbf{x}_0 \tag{20}$$

For ridge prediktoren (19) har vi

$$E\left(\hat{f}_\lambda(\mathbf{x}_0\right) = \mathbf{x}_0'E(\hat{\boldsymbol{\beta}}_\lambda) = (1+\lambda)^{-1}\mathbf{x}_0'\boldsymbol{\beta}$$
$$\mathrm{V}\left(\hat{f}_\lambda(\mathbf{x}_0\right) = \mathbf{x}_0'\mathrm{Cov}(\hat{\boldsymbol{\beta}}_\lambda)\mathbf{x}_0 = \sigma^2(1+\lambda)^{-2}\mathbf{x}_0'\mathbf{x}_0$$

og prediksjonsfeilen blir

$$\begin{aligned}\mathrm{PE}_\lambda(\mathbf{x}_0) &= \sigma^2 + \left[(1+\lambda)^{-1}\mathbf{x}_0'\boldsymbol{\beta} - \mathbf{x}_0'\boldsymbol{\beta}\right]^2 + \sigma^2(1+\lambda)^{-2}\mathbf{x}_0'\mathbf{x}_0 \\ &= \sigma^2 + \left(\frac{\lambda}{1+\lambda}\right)^2(\mathbf{x}_0'\boldsymbol{\beta})^2 + \sigma^2(1+\lambda)^{-2}\mathbf{x}_0'\mathbf{x}_0\end{aligned} \tag{21}$$

For å sammenligne prediksjonsfeilene, ser vi på differansen mellom (21) og (20):

$$\begin{aligned}\mathrm{PE}_\lambda(\mathbf{x}_0) - \mathrm{PE}_\lambda(\mathbf{x}_0) &= \sigma^2 + \left(\frac{\lambda}{1+\lambda}\right)^2(\mathbf{x}_0'\boldsymbol{\beta})^2 + \sigma^2(1+\lambda)^{-2}\mathbf{x}_0'\mathbf{x}_0 - \left\{\sigma^2 + \sigma^2\mathbf{x}_0'\mathbf{x}_0\right\} \\ &= \left(\frac{\lambda}{1+\lambda}\right)^2(\mathbf{x}_0'\boldsymbol{\beta})^2 - \frac{\lambda(2+\lambda)}{(1+\lambda)^2}\,\sigma^2\mathbf{x}_0'\mathbf{x}_0 \\ &= \sigma^2\mathbf{x}_0'\mathbf{x}_0\left\{\left(\frac{\lambda}{1+\lambda}\right)^2\left(\frac{\mathbf{x}_0'\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{x}_0'\mathbf{x}_0}}\right)^2 - \frac{\lambda(2+\lambda)}{(1+\lambda)^2}\right\}\end{aligned} \tag{22}$$

Av (22) ser vi at ridge estimatoren har mindre prediksjonsfeil enn minste kvadraters estimator når $0 < \lambda < 2\left\{\left(\frac{\mathbf{x}_0'\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{x}_0'\mathbf{x}_0}}\right)^2 - 1\right\}$.

5. Kryssvalidering

Vi ser igjen på den generelle situasjonen fra avsnitt 2. For å bestemme den beste verdien av λ kan vi bruke kryssvalidering. Vi ser her på "leave one out cross validation". Et alternativ er "k-fold cross validation".

Vi starter med å velge en sekvens av λ-verdier fra verdier nær null til store verdier. (Strengt tatt er vi interessert i alle verdier av $\lambda > 0$, men det er ikke praktisk mulig.) Så går vi fram på denne måten:

- For hver verdi av λ gjør vi følgende:
  - Bruk alle observasjonene unntatt den i-te til å bestemme ridge estimatoren $\hat{\boldsymbol{\beta}}_\lambda^{(-i)}$.
  - Bruk $\hat{\boldsymbol{\beta}}_\lambda^{(-i)}$ til å predikere Y_i, og la $\hat{Y}_i^{(\lambda)}$ være prediksjonen vi får.
- Beregn kryssvalideringsfeilen

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i^{(\lambda)}\right)^2 \tag{23}$$

Det beste valget av λ er det som gir minst verdi av $\mathrm{CV}(\lambda)$.

5. Et eksempel

Vi vil vise hvordan vi kan bruke `R` til å bestemme rigde estimatene og til å finne den beste verdien av λ ved kryssvalidering. Til illustrasjon vil vi bruke en studie med data for 442 pasienter med sukkersyke. Som en del av denne studien ønsket en å bruke ti forklaringsvariabler og 54 interaksjoner mellom dem til å predikere et mål for progresjon av sykdommen ett år fram i tid (`y`). De ti forklaringsvariablene er alder (`ald`), kjønn (`sex`), kroppsmasseindeks (`bmi`), gjennomsnittlig blodtrykk (`map`), og seks målinger av blodserum (`tc`, `ldl`, `hdl`, `tch`, `ltg`, `glu`). På datafilen er alle forklaringsvariablene sentrert og standardisert (slik at $\sum_{i=1}^n x_{ij} = 0$ og $\sum_{i=1}^n x_{ij}^2 = 1$ for alle j).

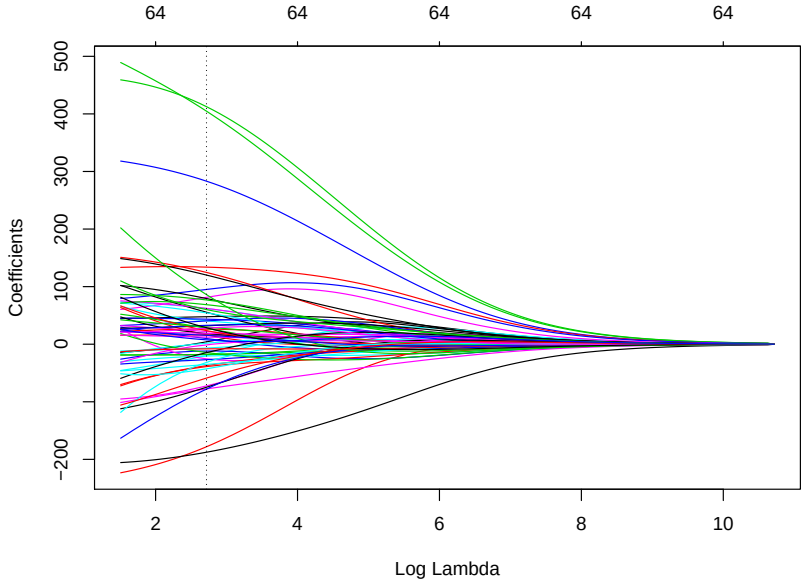
Vi leser dataene inn i en dataramme som vi kaller `diabetes` slik det er beskrevet på kurssiden. Vi vil bruke pakken `glmnet` til å utføre ridge regresjon. Når vi skal bruke denne pakken, må vi først definere vektoren av responser ($\mathbf{Y}$) og $\mathbf{X}$-matrisen (10):

```
> y=diabetes$y
> x=model.matrix(y~.,data=diabetes)[,-1]
```

Så bestemmer vi rigde estimatene (11) for en sekvens av λ-verdier (som kommandoen velger selv hvis vi ikke angir dem) og plotter dem som en funksjon av logaritmen til λ (opsjonen `alpha=0` gir oss ridge estimatene):

```
> ridge.mod=glmnet(x,y,alpha=0)
> plot(ridge.mod,xvar="lambda")
```

Vi får da plottet:

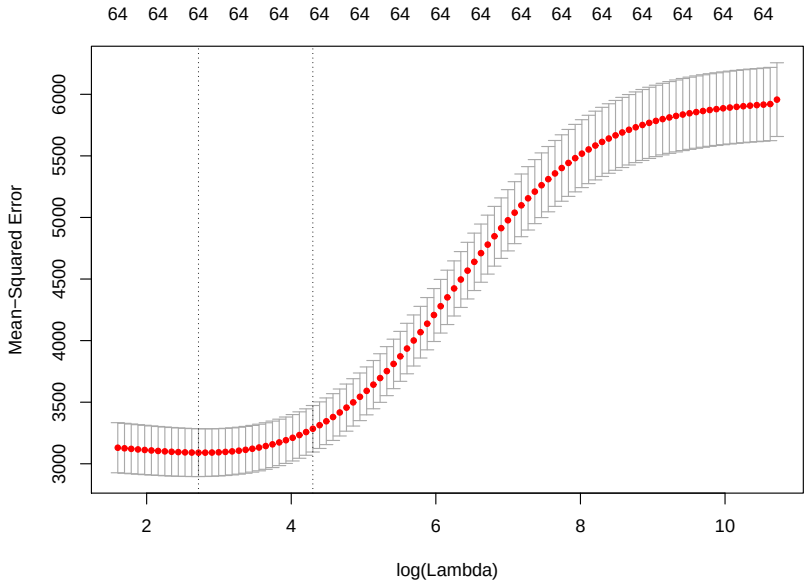


Vi ser at rigde estimatene avtar etter som vi øker verdien av λ, dvs. gir større straff til store verdier av estimatene.

For å bestemme beste verdi av λ gjør vi en kryssvalidering og plotter $CV(\lambda)$ som en funksjon av logaritmen til λ:

```
cv.ridge=cv.glmnet(x,y,alpha =0,nfolds=length(y),grouped=F)
plot(cv.cv.ridge)
```

Vi får da plottet:



Den λ-verdien som minimerer $CV(\lambda)$ er markert med den første av de to loddrette linjene på plottet. Ved å gi kommandoen `cv.ridge$lambda.min` finner vi verdien $\lambda = 15.14$. Denne verdien av λ har vi også markert på det første plottet ovenfor. Vi får de tilhørende ridge estimatene ved kommandoen `coef(cv.ridge,s="lambda.min")`.

6. Lasso

Et alternativ til ridge regresjon er lasso ("least absolute shrinkage and selection operator"). Også for lasso straffer vi store verdier av estimatene. Men for lasso bruker vi den straffete kvadratsummen

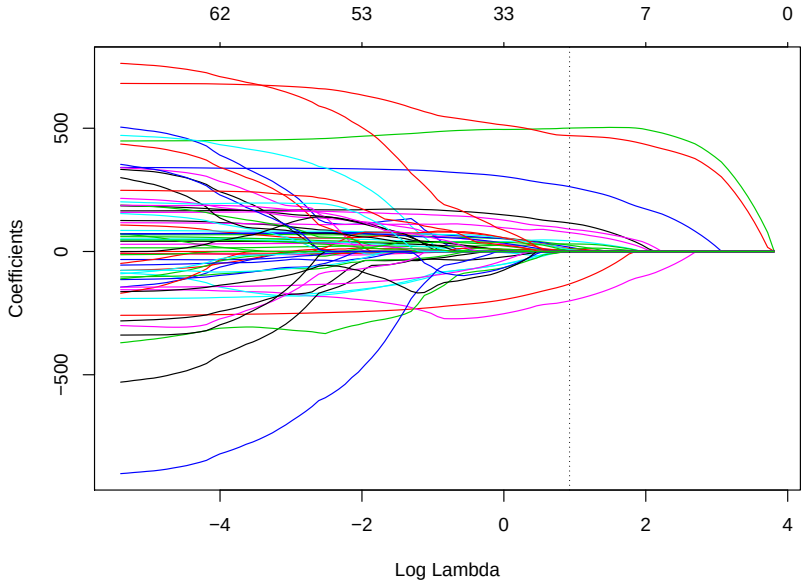
$$h(b_0, b_1, \ldots, b_k) = \sum_{i=1}^{n} [Y_i - (b_0 + b_1 x_{i1} + \cdots + b_k x_{ik})]^2 + \lambda \sum_{j=1}^{k} |b_j| \tag{24}$$

Teorien for lasso er noe vanskeligere enn for ridge regresjon, så vi vil ikke gå inn på denne. Men en viktig effekt av at lasso-straffen er definert ut fra absoluttverdiene til b_j-ene, er at flere av lasso estimatene vil bli eksakt lik null. Denne egenskapen til lasso, gjør estimatene lettere å fortolke enn estimatene for ridge regresjon.

Vi ser til slutt på eksemplet fra forrige avsnitt. For å få lasso estimatene for en sekvens av λ-verdier (som kommandoen velger selv hvis vi ikke angir dem) og plott dem som en funksjon av logaritmen til λ gir vi kommandoene (opsjonen `alpha=1` gir oss lasso):

```
> lasso.mod=glmnet(x,y,alpha=1)
> plot(lasso.mod,xvar="lambda")
```

Vi får da plottet:

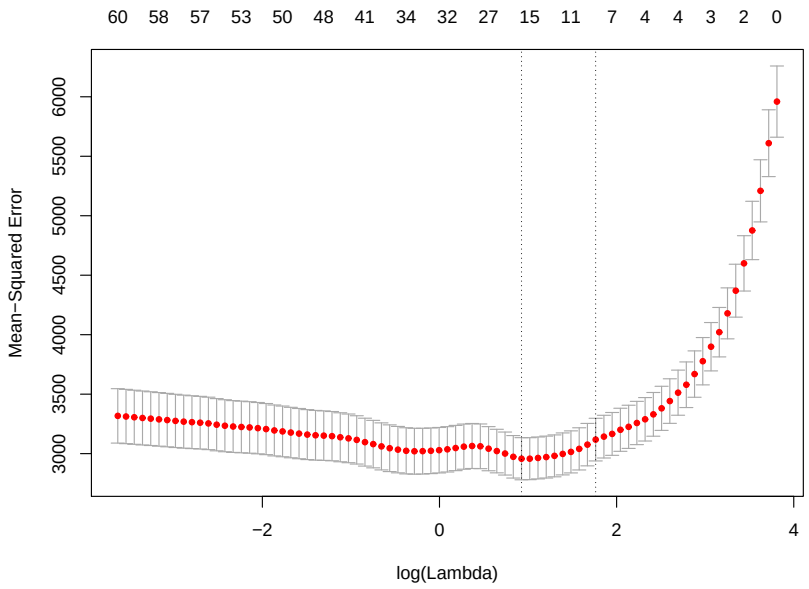


Vi ser at lasso estimatene avtar etter som vi øker verdien av λ og at de etterhvert blir lik null. (Tallene over plottet angir hvor mange estimater som ikke er lik null.)

For å bestemme beste verdi av λ gjør vi en kryssvalidering og plotter CV(λ) som en funksjom av logaritmen til λ:

```
cv.lasso=cv.glmnet(x,y,alpha=1,nfolds=length(y),grouped=F)
plot(cv.lasso)
```

Vi får da plottet:



Den λ-verdien som minimerer CV(λ) er markert med den første av de to loddrette linjene på plottet. Ved å gi kommandoen `cv.lasso$lambda.min` finner vi verdien $\lambda = 2.52$. Denne verdien av λ har vi også markert på det første plottet ovenfor. Vi får de tilhørende lasso estimatene ved kommandoen `coef(cv.lasso,s="lambda.min")`.