Modellvalg ved multippel regresjon – notat til STK2120

Ørulf Borgan februar 2016

I dette notatet vil vi se litt nærmere på hvordan vi kan velge ut hvilke forklaringsvariabler vi skal ha med i en regresjonsmodell. 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. Problemstilling

Vi vil se på den lineære regresjonsmodellen

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \tag{1}$$

der x_{ij}-ene er gitte forklaringsvariabler og ϵ_i-ene er uavhengige og $N(0, \sigma^2)$-fordelte stokastiske variabler. I (1) kan det være at noen av β_j-ene er lik null, slik at de tilhørende forklaringsvariablene kan utelates fra modellen. Et viktig spørsmål er derfor å avgjøre hvilke av de k forklaringsvariablene som skal være med i modellen og hvilke vi kan utelate.

Det fins ikke noe fasitsvar for hvordan vi skal gjøre et slikt modellvalg. Den diskusjonen vi gjør i dette notatet er derfor ikke utfyllende, og det er flere viktige synspunkter og teknikker vi ikke vil komme inn på. I hele notatet vil vi se bort fra interaksjoner, så eventuelle interaksjoner må defineres som egne forklaringsvariabler i (1).

Vi vil bruke følgende framgangsmåte for å finne "den beste" modellen:

- Først finner vi den beste modellen som har med m forklaringsvariabler for $m = 0, 1, \ldots, k$. Det gir oss $k+1$ modeller $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_k$, der $\mathcal{M}_m$ er den beste blant de modellene som bruker m forklaringsvariabler.
- Deretter velger vi den beste av modellene $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_k$.

I avsnittene 2-4 ser vi på tre framgangsmåter vi kan bruke for å bestemme modellene $\mathcal{M}_0, \ldots, \mathcal{M}_k$. I avsnittene 5, 6 og 8 ser vi så på tre metoder vi kan bruke for å velge den beste av modellene $\mathcal{M}_0, \ldots, \mathcal{M}_k$.

2. Forlengs utvelgelse

Ved forlengs utvelgelse går vi fram på følgende måte:

- La $\mathcal{M}_0$ være modellen uten noen forklaringsvariabler, dvs. modellen (1) med bare konstantleddet β_0.
- Se på de k modellene som har med én av forklaringsvariablene. Den beste av disse modellene kaller vi $\mathcal{M}_1$. Med beste modell mener vi her den modellen som har minst residualkvadratsum (SSE).
- Se på de $k-1$ modellene vi får ved utvide modellen $\mathcal{M}_1$ med én ny forklaringsvariabel. Den beste av disse modellene kaller vi $\mathcal{M}_2$.
- Se på de $k-2$ modellene vi får ved utvide modellen $\mathcal{M}_2$ med én ny forklaringsvariabel. Den beste av disse modellene kaller vi $\mathcal{M}_3$.
- Fortsett å legge til én og én forklaringsvaribel på denne måten til vi får modellen $\mathcal{M}_k$ med alle de k forklaringsvariablene.

Denne prosedyren gir oss en sekvens av modeller $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_k$ med flere og flere forklaringsvariabler. Hvordan vi kan velge blant de $k+1$ modellene ser vi på i avsnittene 5, 6 og 8.

3. Baklengs utvelgelse

Ved baklengs utvelgelse går vi fram på følgende måte:

- La $\mathcal{M}_k$ være modellen som har med alle de k forklaringsvariablene i (1).
- Se på de $k-1$ modellene vi får ved å utelate én av forklaringsvariablene fra $\mathcal{M}_k$. Den beste av disse modellene kaller vi $\mathcal{M}_{k-1}$.
- Se på de $k-2$ modellene vi får ved å utelate én av forklaringsvariablene fra $\mathcal{M}_{k-1}$. Den beste av disse modellene kaller vi $\mathcal{M}_{k-2}$.
- Fortsett med å utelate én og én forklaringsvaribel på denne måten til vi får modellen $\mathcal{M}_0$ som ikke har noen forklaringsvariabler.

Denne prosedyren gir oss også en sekvens av modeller $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_k$ med flere og flere forklaringsvariabler. Men den sekvensen av modeller vi får ved baklengs utvelgelse kan godt være en annen enn den sekvensen vi får ved forlengs utvelgelse.

4. Alle mulige regresjonsmodeller

Når vi har k forklaringsvariabler, kan vi tilsammen lage 2^k mulige regresjonsmodeller (siden hver av de k forklaringsvariablene kan være med i modellen eller ikke være med). Hvis $k = 10$ er det altså $2^{10} = 1024$ mulige regresjonsmodellen, mens det er $2^{20} = 1\,048\,576$ mulige modeller hvis $k = 20$. Det betyr at det ikke er praktisk mulig å studere alle mulige modeller når k er stor. Men for mindre verdier av k er det en mulighet. Når vi skal prøve ut alle mulige modeller går vi fram på følgende måte:

- La $\mathcal{M}_0$ være modellen uten noen forklaringsvariabler.
- Se på de k modellene som har med én av forklaringsvariablene. Den beste av disse modellene kaller vi $\mathcal{M}_1$.
- Se på de $\binom{k}{2}$ modellene som har med to av forklaringsvariablene. Den beste av disse modellene kaller vi $\mathcal{M}_2$.
- Se på de $\binom{k}{3}$ modellene som har med tre av forklaringsvariablene. Den beste av disse modellene kaller vi $\mathcal{M}_3$.
- Fortsett på denne måten til vi får modellen $\mathcal{M}_k$ med alle de k forklaringsvariablene.

Vi får på denne måten en sekvens av modeller $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_k$ med flere og flere forklaringsvariabler. Denne sekvensen kan godt være en annen enn de sekvensene vi får ved forlengs og baklengs utvelgelse.

5. Justert R^2

Ved å bruke en av metodene ovenfor, kommer vi fram til en sekvens av modeller $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_k$, der $\mathcal{M}_m$ den beste blant de modellene som bruker m forklaringsvariabler ($m = 0, 1, \ldots, k$). Vi må nå avgjøre hviken av de $k+1$ modellene som er best. Det fins flere kriterier en kan bruke til å gjøre dette, og vi vil se på noen av dem.

Først merker vi oss at vi ikke kan bruke R^2 som et kriterium. For R^2 vil øke med m, så modellen

$\mathcal{M}_k$ med alle de k forklaringsvariablene vil ha størst verdi av R^2. Men ved å endre litt på R^2 får vi et kriterium for å velge best mulig modell. Vi definerer justert R^2 ved

$$R_a^2 = 1 - \frac{\text{SSE}(m)/[n-(m+1)]}{\text{SST}/(n-1)}, \tag{2}$$

der $\text{SSE}(m)$ er residualkvadratsummen for modellen $\mathcal{M}_m$ og SST er total kvadratsum (som ikke avhenger av hvilken modell vi ser på). Tanken bak den justerte R^2 er som følger. Hvis vi tar med forklaringsvariabler i modellen som ikke har noen betydning (dvs. de tilsvarende β_j-ene er lik null), vil residualkvadratsummen $\text{SSE}(m)$ forandre seg lite når m øker. Det vil føre til at $\text{SSE}(m)/[n-(m+1)]$ øker og dermed at R_a^2 minker. Vi blir altså straffet for å ta med forklaringsvariabler som ikke har noen betydning. Et kriterium for valg av modell er da å velge den av modellene $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_k$ som har størst verdi av R_a^2.

6. Mallows C_p

Et annet kriterium for valg av modell er Mallows C_p. Mallows C_p er gitt ved

$$C_p = \frac{\text{SSE}(m)}{\text{MSE}(k)} + 2(m+1) - n, \tag{3}$$

der $\text{SSE}(m)$ er residualkvadratsummen for modellen $\mathcal{M}_m$ og $\text{MSE}(k) = \text{SSE}(k)/[n-(k+1)]$ er middelkvadratsummen for modellen med alle de k forklaringsvariablene.

Motivasjonen for Mallows C_p er følgende. Vi vil anta at modellen med alle de k forklaringsvariablene er "stor nok" slik at $\text{MSE}(k)$ er en forventningsrett estimator for σ^2. Da har vi at $C_p \approx (\text{SSE}(m)/\sigma^2) + 2(m+1) - n$, slik at

$$E\left(C_p\right) \approx \frac{E\{\text{SSE}(m)\}}{\sigma^2} + 2(m+1) - n.$$

Hvis modellen $\mathcal{M}_m$ er tilstrekkelig til å beskrive variasjonen i Y_i-ene (dvs. β_j-ene er lik null for de $k-m$ forklaringsvariablene som ikke er med i modellen), er $\text{SSE}(m)/\sigma^2$ kji-kvadratfordelt med $n-(m+1)$ frihetsgrader. Da har vi at

$$E\left(C_p\right) \approx n - (m+1) + 2(m+1) - n = m + 1.$$

Men hvis modellen $\mathcal{M}_m$ ikke er tilstrekkelig til å beskrive variasjonen i Y_i-ene, er

$$E\{\text{SSE}(m)/\sigma^2\} > n - (m+1)$$

og $E\left(C_p\right) > m+1$. Hvis vi bestemmer (3) for $m = 0, 1, \ldots, k$, vil vi (typisk) finne at verdien av C_p avtar med økende verdi av m så lenge vi tar med forklaringsvariabler som har betydning for å beskrive variasjonen i Y_i-ene, og at verdien av C_p så vil bli omtrent lik $m+1$ når vi har tatt med "tilstrekkelig mange" forklaringsvariabler. Et mulig valg av beste modell, er derfor å velge den minste verdien av m der $C_p \approx m+1$.

Ressonementet ovenfor forutsetter at vi ser på SSE for én modell med m forklaringsvariabler. Men med metodene i avsnittene 2-4 velger vi $\mathcal{M}_m$ som den av flere mulige modeller med m forklaringsvariabler som gir minst verdi av SSE. Derfor vil $\text{SSE}(m)$ i praksis bli mindre enn det teorien ovenfor skulle tilsi, og et annen mulig kriterium er å velge den verdien av m som gir minst mulig verdi av C_p.

7. Et eksempel

Vi vil vise hvordan vi kan bruke `R` til modellvalg ved lineær regresjon. Til illustrasjon ser vi på følgende eksempel: I en liten studie har en målt vekten (i kg) og ti andre fysiske mål (i cm) for 22 mannlige studenter. Vi er interessert i å finne en regresjonsmodell for beskriver sammenhengen mellom vekt og de andre fysiske målene. (I slike studier er det vanlig å log-transformere alle variablene. Men for enkelthets skyld velger å ikke gjøre det i dette eksemplet.)

For hver av 22 studentene har vi registrert:

- `vekt`: Vekt
- `uarm`: Maksimum diameter av underarm
- `oarm`: Maksimum diameter av overarm (biceps)
- `bryst`: Omkrets rundt brystkassen like under armhulene
- `hals`: Omkrets rundt halsen
- `skulder`: Omkrets rundt skuldrene ved toppen av skulderbladene
- `midje`: Omkrets rundt midjen
- `hoyde`: Høyde
- `legg`: Maksimum omkrets rundt legg
- `laar`: Omkrets rundt lår (midt mellom kne og skritt)
- `hode`: Maksimum omkrets rundt hodet

Vi leser dataene inn i en dataramme som vi kaller `vekt` slik det er beskrevet på kurssiden. Dataene er som følger:

```
> vekt
    vekt  uarm  oarm  bryst  hals  skulder  midje  hoyde  legg  laar  hode
1   77.0  28.5  33.5  100.0  38.5   114.0   85.0   178.0  37.5  53.0  58.0
2   85.5  29.5  36.5  107.0  39.0   119.0   90.5   187.0  40.0  52.0  59.0
3   63.0  25.0  31.0   94.0  36.5   102.0   80.5   175.0  33.0  49.0  57.0
4   80.5  28.5  34.0  104.0  39.0   114.0   91.5   183.0  38.0  50.0  60.0
5   79.5  28.5  36.5  107.0  39.0   114.0   92.0   174.0  40.0  53.0  59.0
6   94.0  30.5  38.0  112.0  39.0   121.0  101.0   180.0  39.5  57.5  59.0
7   66.0  26.5  29.0   93.0  35.0   105.0   76.0   177.5  38.5  50.0  58.5
8   69.0  27.0  31.0   95.0  37.0   108.0   84.0   182.5  36.0  49.0  60.0
9   65.0  26.5  29.0   93.0  35.0   112.0   74.0   178.5  34.0  47.0  55.5
10  58.0  26.5  31.0   96.0  35.0   103.0   76.0   168.5  35.0  46.0  58.0
11  69.5  28.5  37.0  109.5  39.0   118.0   80.0   170.0  38.0  50.0  58.5
12  73.0  27.5  33.0  102.0  38.5   113.0   86.0   180.0  36.0  49.0  59.0
13  74.0  29.5  36.0  101.0  38.5   115.5   82.0   186.5  38.0  49.0  60.0
14  68.0  25.0  30.0   98.5  37.0   108.0   82.0   188.0  37.0  49.5  57.0
15  80.0  29.5  36.0  103.0  40.0   117.0   95.5   173.0  37.0  52.5  58.0
16  66.0  26.5  32.5   89.0  35.0   104.5   81.0   171.0  38.0  48.0  56.5
17  54.5  24.0  30.0   92.5  35.5   102.0   76.0   169.0  32.0  42.0  57.0
18  64.0  25.5  28.5   87.5  35.0   109.0   84.0   181.0  35.5  42.0  58.0
19  84.0  30.0  34.5   99.0  40.5   119.0   88.0   188.0  39.0  50.5  56.0
20  73.0  28.0  34.5   97.0  37.0   104.0   82.0   173.0  38.0  49.0  58.0
21  89.0  29.0  35.5  106.0  39.0   118.0   96.0   179.0  39.5  51.0  58.5
22  94.0  31.0  33.5  106.0  39.0   120.0   99.5   184.0  42.0  55.0  57.0
```

Når vi skal bruke forlengs eller baklengs utvelgelse for disse dataene, må vi tilpasse $1+10+9+8+\cdots+2+1=56$ ulike regresjonsmodeller, mens vi må tilpasse hele $2^{10}=1024$ modeller hvis vi vil prøve alle mulige kombinasjoner av forklaringsvariablene. Det er ikke praktisk mulig å gjøre dette ved å bruke `lm`-kommandoen for hver av modellene. Vi vil derfor bruke kommandoen `regsubsets` i `R`-pakken `leaps`. Denne kommandoen kan utføre forlengs og baklengs utvelgelse og også prøve ut alle mulige modeller. Vi får også beregnet justert R^2 og Mallows C_p.

Vi bruker først forlengs utvelgelse:

```
> fit.forward=regsubsets(vekt~.,data=vekt,nvmax=10,method="forward")
> summary.forward=summary(fit.forward)
> summary.forward

.... (redigert output) ....

Selection Algorithm: forward
          uarm oarm bryst hals skulder midje hoyde legg laar hode
1  ( 1 )  " "  " "  " "   " "  " "     "*"   " "   " "  " "  " "
2  ( 1 )  "*"  " "  " "   " "  " "     "*"   " "   " "  " "  " "
3  ( 1 )  "*"  " "  " "   " "  " "     "*"   "*"   " "  " "  " "
4  ( 1 )  "*"  " "  " "   " "  " "     "*"   "*"   " "  "*"  " "
5  ( 1 )  "*"  " "  " "   " "  " "     "*"   "*"   " "  "*"  "*"
6  ( 1 )  "*"  " "  " "   " "  " "     "*"   "*"   "*"  "*"  "*"
7  ( 1 )  "*"  " "  "*"   " "  " "     "*"   "*"   "*"  "*"  "*"
8  ( 1 )  "*"  " "  "*"   "*"  " "     "*"   "*"   "*"  "*"  "*"
9  ( 1 )  "*"  "*"  "*"   "*"  " "     "*"   "*"   "*"  "*"  "*"
10 ( 1 )  "*"  "*"  "*"   "*"  "*"     "*"   "*"   "*"  "*"  "*"
```

Utskriften gir en oversikt over de modellene vi får ved forlengs utvelgelse av forklaringsvariablene. Det er angitt med en stjerne (*) hvilke forklaringsvariabler som er med i de ulike modellene. Vi ser at den første forklaringsvariabelen som velges ut er `midje`. Deretter velges forklaringsvariablene ut i følgende rekkefølge: `uarm`, `hoyde`, `laar`, `hode`, `legg`, `bryst`, `hals`, `oarm` og `skulder`.

For å avgjøre hvor mange av forklaringsvariablene vi bør ta med i modellen, kan vi se på justert R^2 eller Mallows C_p for de ulike modellene. For justert R^2 får vi:

```
> summary.forward$adjr2
 [1] 0.8292 0.9297 0.9482 0.9579 0.9615 0.9641 0.9644 0.9627610 0.9601 0.9565
```

Vi ser at justert R^2 er størst når vi har med sju forklaringsvariabler, men det er liten endring i justert R^2 etter at vi har tatt med fire forklaringsvariabler. For Mallows C_p får vi:

```
> summary.forward$cp
 [1] 60.499 14.699  7.446  4.440  4.142  4.377  5.469  7.127  9.015 11.000
```

For modellene med tre eller færre forklaringsvariabler, er Mallows C_p klart større enn antall forklaringsvariabler pluss én. Men for modellen med fire forklaringsvariabler er Mallows C_p litt mindre enn $4+1$. Så vi velger modellen med fire forklaringsvariabler.

Kommandoene for baklengs utvelgelse og for metoden der vi ser på alle mulige modeller er tilsvarende som for forlengelse utvelgelse, men nå bruker vi opsjonene `method="backwards"` og

`method="exhaustive"` i `regsubsets`-kommandoen. (Det siste valget er default.) For de dataene vi ser på her, blir resultatene de samme som for forlengs utvelgelse, men det er ikke generelt tilfellet.

For å finne estimatene for modellen med fire forklaringsvariabler, kan vi gi kommandoen

```
> coef(fit.forward,4)
 (Intercept)         uarm        midje        hoyde         laar
-113.3120436    2.0355814    0.6468837    0.2717468    0.5400844
```

8. Kryssvalidering

Vi skal er se på et kriterium til for å bestemme den beste av modellene $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_k$. Tanken her er at den beste modellen er den som gir best prediksjon for nye observasjoner (trukket fra samme populasjon som Y_i-ene).

Hvis vi hadde hatt nye observasjoner $Y_{i,\text{ny}}$, $i = 1, \ldots, n_{\text{ny}}$, kunne vi ha brukt hver av modellene $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_k$ til å predikere disse observasjonene ut fra verdiene av de tilhørende forklaringsvariablene. La $\hat{Y}_{i,\text{ny}}^{(m)}$ være prediksjonene vi ville ha fått med modellen $\mathcal{M}_m$, $m = 0, 1, \ldots, k$. Vi kunne da ha bestemt den gjennomsnittlige kvadratiske prediksjonsfeilen ved

$$\text{KvFeil}(m) = \sum_{i=1}^{n_{\text{ny}}} \left(Y_{i,\text{ny}} - \hat{Y}_{i,\text{ny}}^{(m)} \right)^2$$

Den beste modellen ville da være den som har minst verdi av KvFeil(m).

Men hvis vi ikke har noen nye observasjoner, kan vi ikke bruke denne metoden. Men vi kan "etterligne" metoden ved å bruke kryssvalidering. Vi vil her se på "leave one out cross validation". (Et alternativ er "k-fold cross validation".) For metodene i avsnittene 2-4 er framgangsmåten som følger:

- For $i = 1, \ldots, n$:
  - Bruk alle observasjonene unntatt den i-te til å bestemme modellene $\mathcal{M}_0^{(-i)}, \mathcal{M}_1^{(-i)}, \ldots,$ $\mathcal{M}_k^{(-i)}$, der $\mathcal{M}_m^{(-i)}$ er den beste blant de modellene som bruker m forklaringsvariabler.
  - Bruk hver av modellene i forrige punkt til å predikere Y_i, og la $\hat{Y}_i^{(m)}$ være prediksjonen vi får når vi bruker modellen $\mathcal{M}_m^{(-i)}$; $m = 0, 1, \ldots, k$.
- Beregn kryssvalideringsfeilen

$$\text{CV}(m) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i^{(m)} \right)^2 \tag{4}$$

  for $m = 0, 1, \ldots, k$

Det beste valg av m er nå den verdien som gir minst verdi av CV(m).

Vi vil til slutt se hva kryssvalidering gir for eksemplet i avsnitt 7. Det er ikke noen ferdige `R`-kommandoer for å kryssvalidere de metodene vi har sett på i dette notatet. Men vi kan finne CV(m) ved følgende kommandoer:

```
> n=22
> k=10
> cv=rep(0,k)
> mat=model.matrix(vekt~.,data=vekt)
> for (i in 1:n)
+ {
+ fit.i=regsubsets(vekt~.,data=vekt[-i,],nvmax=10,method="forward")
+ val.errors=rep(0,k)
+   for (m in 1:k)
+   {
+   coef.m=coef(fit.i,id=m)
+   pred=sum(mat[i,names(coef.m)]*coef.m)
+   val.errors[m]=(vekt$vekt[i]-pred)^2
+   }
+ cv=cv+val.errors
+ }
> cv=cv/n
> cv
 [1] 34.281  9.940  9.296  7.932  7.854  7.718  8.789  9.159 10.510 11.478
```

Vi ser av CV(m) blir minst når vi har med seks forklaringsvariabler, men CV(m) er neste like liten for modellen med fire forklaringsvariabler.