Generelle Statistiske Metoder

Erik Rybakken

December 11, 2017

Framingham Datasett

- Studie av 4434 pasienter, med fokus på hjerteproblemer.

Framingham Datasett

- Studie av 4434 pasienter, med fokus på hjerteproblemer.
- Hver pasient ble undersøkt tre ganger, med 6 års mellomrom.

Framingham Datasett

- Studie av 4434 pasienter, med fokus på hjerteproblemer.
- Hver pasient ble undersøkt tre ganger, med 6 års mellomrom.

To mulige problemstillinger:

1. Hvilke faktorer påvirker blodtrykket?

Framingham Datasett

- Studie av 4434 pasienter, med fokus på hjerteproblemer.
- Hver pasient ble undersøkt tre ganger, med 6 års mellomrom.

To mulige problemstillinger:

1. Hvilke faktorer påvirker blodtrykket?
2. Kan vi forutsi hvor lenge en pasient har igjen å leve basert på dataene som ble gjort i første sjekk?

Detaljer om analysen:

1. Kun pasientene fra den første sjekken ble brukt i analysen

Framingham Datasett

- Studie av 4434 pasienter, med fokus på hjerteproblemer.
- Hver pasient ble undersøkt tre ganger, med 6 års mellomrom.

To mulige problemstillinger:

1. Hvilke faktorer påvirker blodtrykket?
2. Kan vi forutsi hvor lenge en pasient har igjen å leve basert på dataene som ble gjort i første sjekk?

Detaljer om analysen:

1. Kun pasientene fra den første sjekken ble brukt i analysen
2. Paientene med minst én ukjent verdi ble fjernet fra analysen

Framingham Datasett

- Studie av 4434 pasienter, med fokus på hjerteproblemer.
- Hver pasient ble undersøkt tre ganger, med 6 års mellomrom.

To mulige problemstillinger:

1. Hvilke faktorer påvirker blodtrykket?
2. Kan vi forutsi hvor lenge en pasient har igjen å leve basert på dataene som ble gjort i første sjekk?

Detaljer om analysen:

1. Kun pasientene fra den første sjekken ble brukt i analysen
2. Paientene med minst én ukjent verdi ble fjernet fra analysen
3. Dette ble i alt 3885 pasienter

Lineær regresjon

Vi antar en lineær modell:

$$Y = X\beta + \epsilon \tag{1}$$

Lineær regresjon

Vi antar en lineær modell:

$$Y = X\beta + \epsilon \tag{1}$$

der

- Y er responsen
- X er prediktorene
- β er en vektor med koeffisienter
- ϵ er en normalfordelt variabel med forventningsverdi 0

Lineær regresjon

Vi antar en lineær modell:

$$Y = X\beta + \epsilon \tag{1}$$

der

- Y er responsen
- X er prediktorene
- β er en vektor med koeffisienter
- ϵ er en normalfordelt variabel med forventningsverdi 0

I vårt tilfelle er Y blodtrykket, mens X består av faktorene kjønn (mann/kvinne), alder, antall sigaretter røyket per dag, BMI, glukosenivå, utdanningsnivå og kolesterolnivå.

Minstre kvadraters metode

Vi har en $n \times p$-matrise $\mathbf{X}$ bestående av n observasjoner av p faktorer, og en $n \times 1$-matrise $\mathbf{Y}$ bestående av de korresponderene responsvariablene. Vi vil fra nå av anta at de observerte faktorene og responsene er normalisert til å ha gjennomsnitt 0 og varians 1.

Minstre kvadraters metode

Vi har en $n \times p$-matrise $\mathbf{X}$ bestående av n observasjoner av p faktorer, og en $n \times 1$-matrise $\mathbf{Y}$ bestående av de korresponderene responsvariablene. Vi vil fra nå av anta at de observerte faktorene og responsene er normalisert til å ha gjennomsnitt 0 og varians 1.

Minstre kvadraters metode estimerer koeffisientene β ved å minimere RSS (residual sum of squares):

$$(\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)^T \tag{2}$$

Minstre kvadraters metode

Vi har en $n \times p$-matrise $\mathbf{X}$ bestående av n observasjoner av p faktorer, og en $n \times 1$-matrise $\mathbf{Y}$ bestående av de korresponderene responsvariablene. Vi vil fra nå av anta at de observerte faktorene og responsene er normalisert til å ha gjennomsnitt 0 og varians 1.

Minstre kvadraters metode estimerer koeffisientene β ved å minimere RSS (residual sum of squares):

$$(\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)^T \tag{2}$$

Bruker alle faktorene, og kan føre til overfitting, og dermed dårligere prediksjon.

Tre forbedringer av minstre kvadraters metode

- Beste delmengde-utvalg velger ut en delmengde av faktorene og utfører minstre kvadrater på denne.

Tre forbedringer av minstre kvadraters metode

- Beste delmengde-utvalg velger ut en delmengde av faktorene og utfører minstre kvadrater på denne.
- Lasso-regresjon krymper absoluttverdien til koeffisientene β.

Tre forbedringer av minstre kvadraters metode

- Beste delmengde-utvalg velger ut en delmengde av faktorene og utfører minstre kvadrater på denne.
- Lasso-regresjon krymper absoluttverdien til koeffisientene β.
- Prinsipalkomponent-regresjon projiserer først faktorene til et lavere-dimensjonalt underrom og utfører deretter minstre kvadrater.

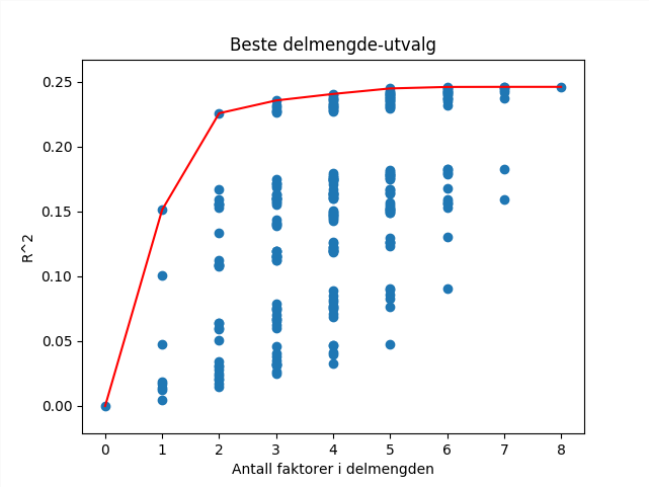
Beste delmengde-utvalg

Gitt en delmengde $S \subset \{1, \ldots, p\}$ kan vi danne matrisen $\mathbf{X}_S = (\mathbf{X}_{i_1} | \mathbf{X}_{i_2} | \ldots | \mathbf{X}_{i_k})_{i_* \in S}$.
Vi kan så utføre minst kvadrater på $\mathbf{X}_S$.

Beste delmengde-utvalg

Gitt en delmengde $S \subset \{1, \ldots, p\}$ kan vi danne matrisen $\mathbf{X}_S = (\mathbf{X}_{i_1} | \mathbf{X}_{i_2} | \ldots | \mathbf{X}_{i_k})_{i_* \in S}$. Vi kan så utføre minst kvadrater på $\mathbf{X}_S$. For en gitt $0 \leq k \leq p$ utfører vi minste kvadrater på den delmengden S med $|S| = k$ som gir lavest RSS.

Beste delmengde-utvalg



Lasso-regresjon

Lasso-regresjon finner koeffisientene β som minimerer uttrykket

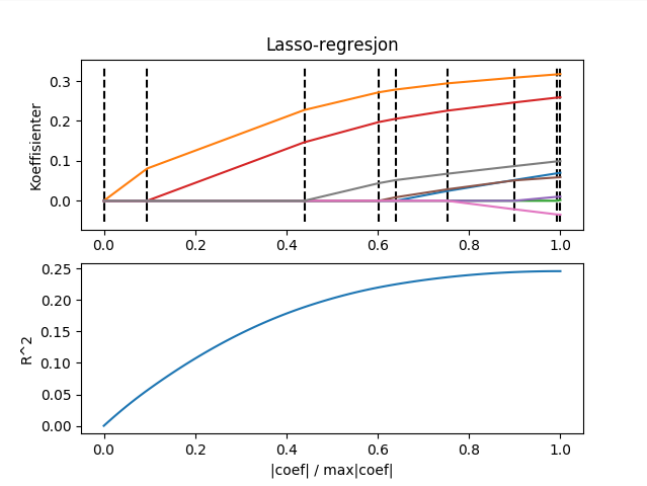
$$(\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)^T + \lambda \sum_{i=1}^{p} |\beta_i| \tag{3}$$

der λ er en parameter som bestemmer hvor mye store koeffisienter skal straffes. Dette er ekvivalent med å minimere uttrykket

$$(\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)^T \tag{4}$$

der vi krever at $\sum_{i=1}^{p} |\beta_i| \leq t$.

Lasso-regresjon



Prinsipalkomponent-regresjon

Matrisen $\mathbf{X}^T\mathbf{X}$ kan dekomponeres (egenverdi-dekomposisjon):

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T \tag{5}$$

der $\mathbf{V}$ er matrisen med egenvektorene til $\mathbf{X}^T\mathbf{X}$ som kolonnevektorer og $\mathbf{D}^2$ er diagonalmatrisa med de korresponderende egenverdiene $d_1^2 \geq d_2^2 \geq \cdots \geq d_p^2$ som diagonalelementer.

Prinsipalkomponent-regresjon

Matrisen $\mathbf{X}^T\mathbf{X}$ kan dekomponeres (egenverdi-dekomposisjon):

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T \tag{5}$$

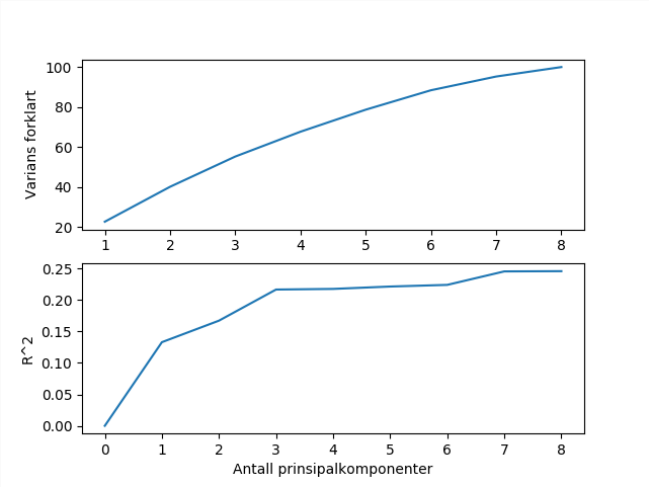
der $\mathbf{V}$ er matrisen med egenvektorene til $\mathbf{X}^T\mathbf{X}$ som kolonnevektorer og $\mathbf{D}^2$ er diagonalmatrisa med de korresponderende egenverdiene $d_1^2 \geq d_2^2 \geq \cdots \geq d_p^2$ som diagonalelementer.

Vi danner matrisa $\mathbf{Z} = \mathbf{X}\mathbf{V}$. Kolonnene i denne matrisa kalles *prinsipalkomponentene* til $\mathbf{X}$. Den n-te prinsipalkomponenten har maksimal varians gitt at den skal være ortogonal til de forrige $n-1$ prinsipalkomponentene.

Prinsipalkomponent-regresjon

Prinsipalkomponent-regresjon utføres ved at man velger de k første prinsipalkomponentene til X, dvs. de første k kolonnene til Z og gjør minste kvadrater på denne matrisa.

Prinsipalkomponent-regresjon



Parametervalg

Aller først delte jeg datasettet i to deler: Ett treningssett med 3108 (80%) pasienter og et valideringssett med 777 (20%) pasienter.

Parametervalg

Aller først delte jeg datasettet i to deler: Ett treningssett med 3108 (80%) pasienter og et valideringssett med 777 (20%) pasienter. For å bestemme parameterene til de tre metodene, brukte jeg kryss-validering.

Parametervalg

Aller først delte jeg datasettet i to deler: Ett treningssett med 3108 (80%) pasienter og et valideringssett med 777 (20%) pasienter. For å bestemme parameterene til de tre metodene, brukte jeg kryss-validering.

Parametervalg

Aller først delte jeg datasettet i to deler: Ett treningssett med 3108 (80%) pasienter og et valideringssett med 777 (20%) pasienter. For å bestemme parameterene til de tre metodene, brukte jeg kryss-validering.

- Treningssettet ble delt inn i 10 grupper.

Parametervalg

Aller først delte jeg datasettet i to deler: Ett treningssett med 3108 (80%) pasienter og et valideringssett med 777 (20%) pasienter. For å bestemme parameterene til de tre metodene, brukte jeg kryss-validering.

- Treningssettet ble delt inn i 10 grupper.
- Hver modell og valg av parameter ble trent på 9 grupper og testet på den siste.

Parametervalg

Aller først delte jeg datasettet i to deler: Ett treningssett med 3108 (80%) pasienter og et valideringssett med 777 (20%) pasienter. For å bestemme parameterene til de tre metodene, brukte jeg kryss-validering.

- Treningssettet ble delt inn i 10 grupper.
- Hver modell og valg av parameter ble trent på 9 grupper og testet på den siste.
- Dette ble gjentatt for hver gruppe.

Parametervalg

Aller først delte jeg datasettet i to deler: Ett treningssett med 3108 (80%) pasienter og et valideringssett med 777 (20%) pasienter. For å bestemme parameterene til de tre metodene, brukte jeg kryss-validering.

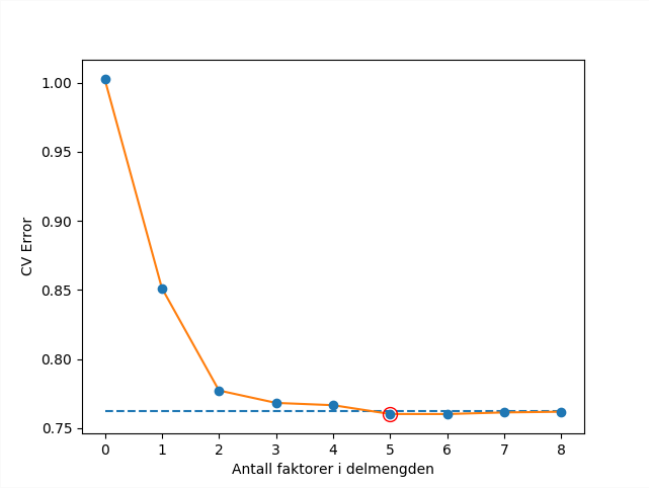
- Treningssettet ble delt inn i 10 grupper.
- Hver modell og valg av parameter ble trent på 9 grupper og testet på den siste.
- Dette ble gjentatt for hver gruppe.
- Gjennomsnittet av RSS ble beregnet for hver gruppe.

Parametervalg

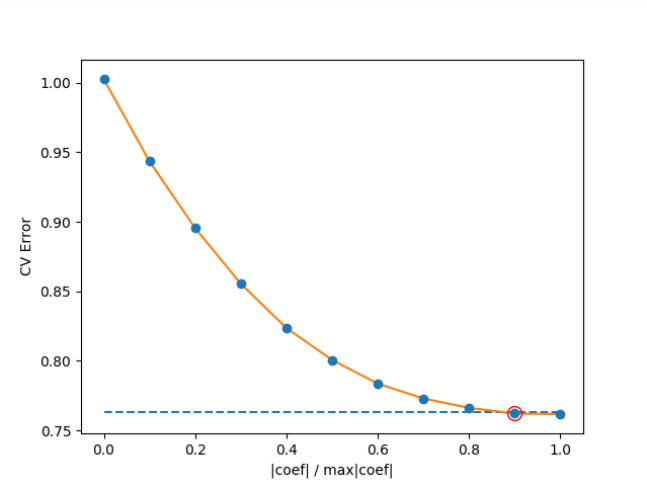
Aller først delte jeg datasettet i to deler: Ett treningssett med 3108 (80%) pasienter og et valideringssett med 777 (20%) pasienter. For å bestemme parameterene til de tre metodene, brukte jeg kryss-validering.

- Treningssettet ble delt inn i 10 grupper.
- Hver modell og valg av parameter ble trent på 9 grupper og testet på den siste.
- Dette ble gjentatt for hver gruppe.
- Gjennomsnittet av RSS ble beregnet for hver gruppe.
- Jeg valgte den parameteren med lavest gjennomsnittlig RSS.

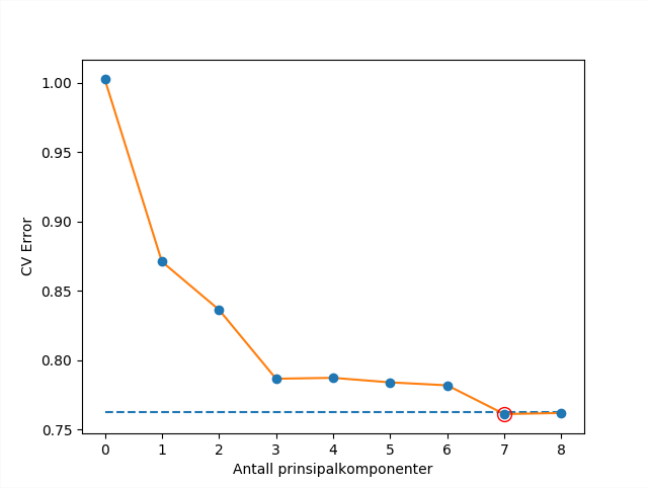
Beste delmengde-utvalg



Lasso-regresjon



Prinsipalkomponent-regresjon



Resultater

Faktor	Minste kvadrater	Beste delmengde	Lasso	PCR
Kjønn	0.070	0.070	0.052	0.071
Alder	0.317	0.323	0.309	0.317
Sigaretter per dag	0.002			0.002
BMI	0.259	0.264	0.259	0.246
Diabetiker	0.010		4.085	0.032
Glukose-nivå	0.059	0.066	0.051	0.037
Utdanningsnivå	-0.034		-0.021	-0.035
Kolesterol-nivå	0.099	0.097	0.087	0.100
Test Error	0.723	0.721	0.720	0.723
Std Error	0.039	0.039	0.039	0.039

Utvidelse av metodene

Jeg dannet nye faktorer ved å ta alle mulige produkter av de originale:

$$X_1 \cdot X_1, X_1 \cdot X_2, \ldots.$$

Utvidelse av metodene

Jeg dannet nye faktorer ved å ta alle mulige produkter av de originale:

$$X_1 \cdot X_1, X_1 \cdot X_2, \ldots.$$

Disse faktorene ble brukt til å danne en ny matrise $\mathbf{X}$ bestående av både de originale og de nye faktorene, tilsammen 44 faktorer.

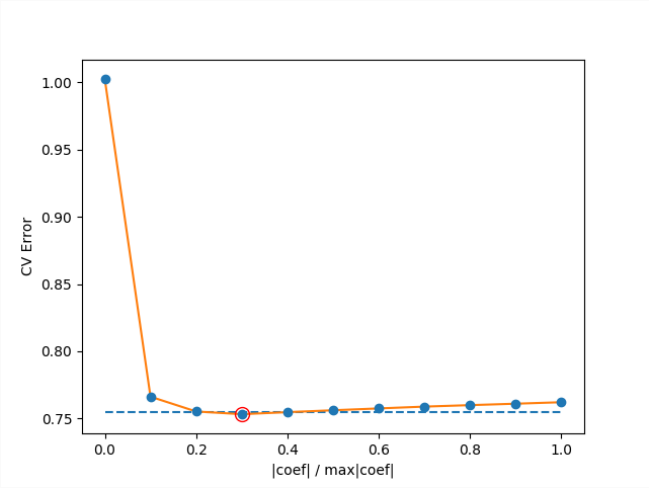
Utvidelse av metodene

Jeg dannet nye faktorer ved å ta alle mulige produkter av de originale:

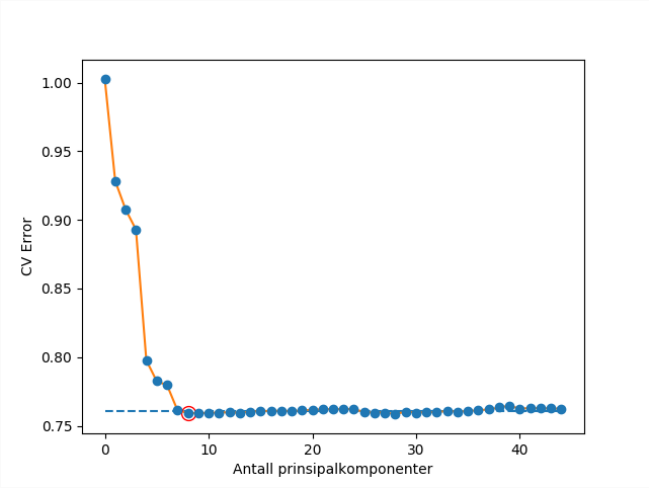
$$X_1 \cdot X_1, X_1 \cdot X_2, \ldots.$$

Disse faktorene ble brukt til å danne en ny matrise $\mathbf{X}$ bestående av både de originale og de nye faktorene, tilsammen 44 faktorer. Jeg utførte så Lasso-regresjon og PCR på denne nye matrisen.

Lasso-regresjon (del 2)



Prinsipalkomponent-regresjon (del 2)



Resultater

	Lasso	PCR
Test Error	0.703	0.720
Std Error	0.038	0.039

Vi ser at Lasso-regresjon ble forbedret fra 0.720 til 0.703.