

Generelle Statistiske Metoder

Erik Rybakken

December 12, 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 prediktorer 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 prediktorer påvirker blodtrykket?
2. Kan vi forutsi hvor lenge en pasient har igjen å leve basert på dataene som ble gjort i første sjekk?

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 prediktorer 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 prediktorer 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. Pasientene 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 prediktorer 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. Pasientene med minst én ukjent verdi ble fjernet fra analysen
3. Dette ble i alt 3885 pasienter

Lineær regresjon

Vi antar en lineær modell:

$$Y = X\beta + \epsilon \quad (1)$$

Lineær regresjon

Vi antar en lineær modell:

$$Y = X\beta + \epsilon \quad (1)$$

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:

$$Y = X\beta + \epsilon \quad (1)$$

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 prediktorene 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 prediktorer, og en $n \times 1$ -matrise \mathbf{Y} bestående av de korresponderende responsvariablene. Vi vil fra nå av anta at de observerte prediktorene 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 prediktorer, og en $n \times 1$ -matrise \mathbf{Y} bestående av de korresponderende responsvariablene. Vi vil fra nå av anta at de observerte prediktorene 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):

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

Tre modifikasjoner av minste kvadraters metode

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

Tre modifiseringer av minste kvadraters metode

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

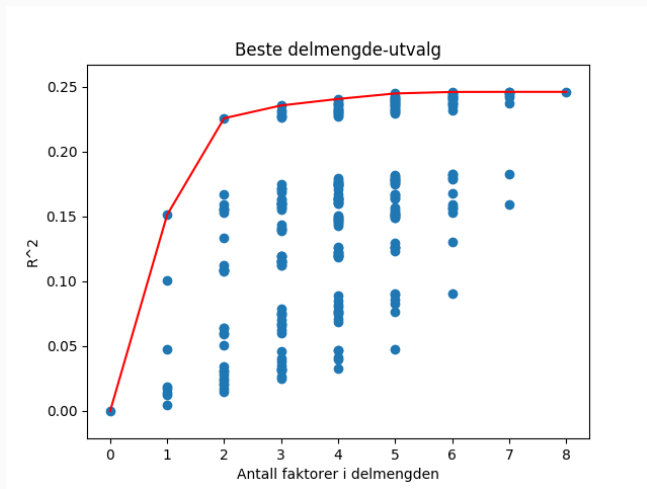
Tre modifiseringer av minste kvadraters metode

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

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

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

Beste delmengde-utvalg



Lasso-regresjon finner koeffisientene β som minimerer uttrykket

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

der λ er en parameter som bestemmer hvor mye koeffisientene skal krympes.

Lasso-regresjon finner koeffisientene β som minimerer uttrykket

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

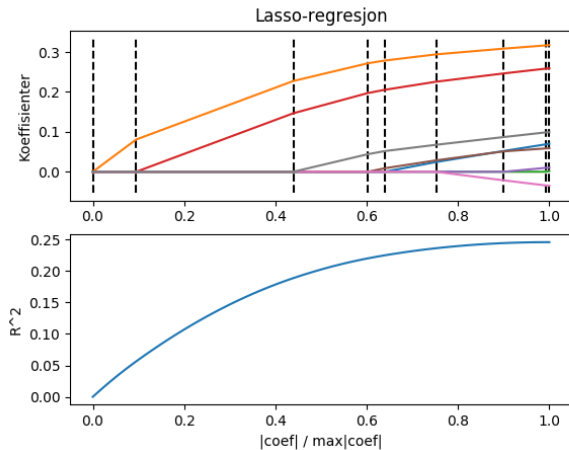
der λ er en parameter som bestemmer hvor mye koeffisientene skal krympes.

Dette er ekvivalent med å minimere uttrykket

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

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

Lasso-regresjon



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

$$\mathbf{X}^T \mathbf{X} = \mathbf{V} \mathbf{D}^2 \mathbf{V}^T \quad (5)$$

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 $d_1^2 \geq d_2^2 \geq \dots \geq d_p^2$ som diagonalelementer.

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

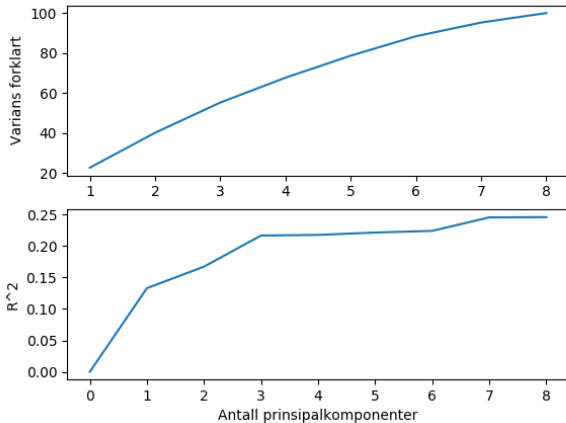
$$\mathbf{X}^T \mathbf{X} = \mathbf{V} \mathbf{D}^2 \mathbf{V}^T \quad (5)$$

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 $d_1^2 \geq d_2^2 \geq \dots \geq d_p^2$ som diagonalelementer.

Vi danner matrisa $\mathbf{Z} = \mathbf{X} \mathbf{V}$. Kolonnene i denne matrisa kalles *prinsipalkomponentene* til \mathbf{X} .

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



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

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.

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.

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.

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.

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.

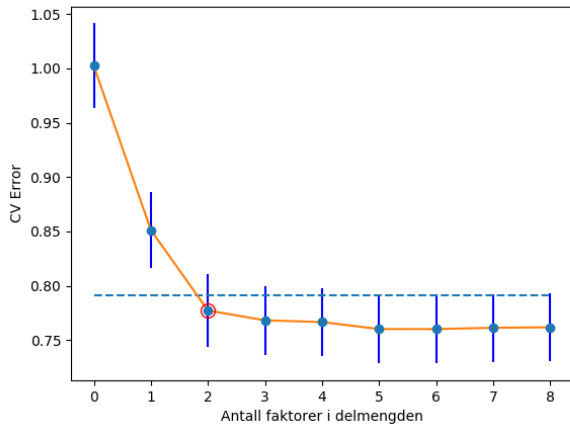
Aller først delte jeg datasettet i to deler: Ett treningssett med 3108 (80%) pasienter og et valideringssett med 777 (20%) pasienter. For å bestemme parameterene til de tre metodene, brukte jeg kryss-validering.

- 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.
- Vi estimerte prediksjonserroreren og standard-erroreren til denne estimatoren.

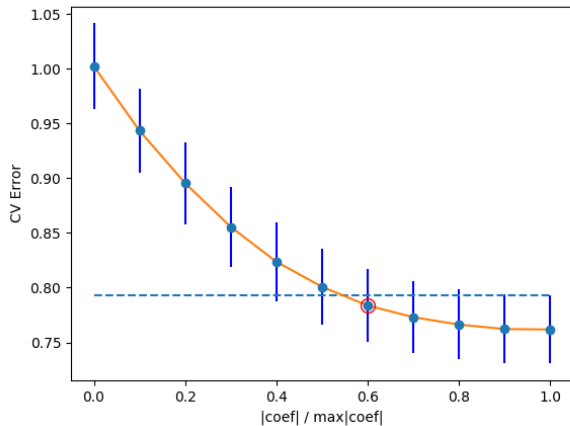
Aller først delte jeg datasettet i to deler: Ett treningssett med 3108 (80%) pasienter og et valideringssett med 777 (20%) pasienter. For å bestemme parameterene til de tre metodene, brukte jeg kryss-validering.

- 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.
- Vi estimerte prediksjonserroreren og standard-erroreren til denne estimatoren.
- Jeg bestemte parameteren ved å bruke *én standard-error-regelen*.

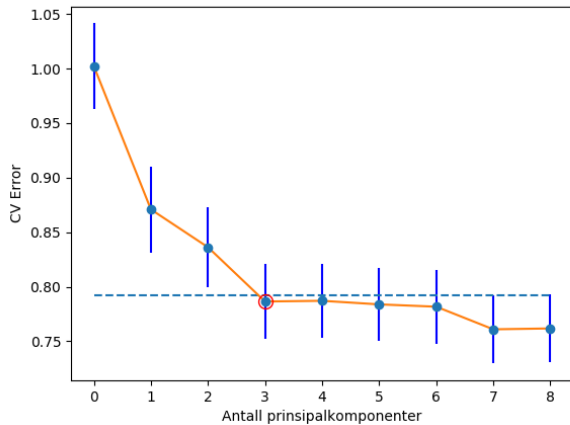
Beste delmengde-utvalg



Lasso-regresjon



Prinsipalkomponent-regresjon



Resultater

Prediktor	Minste kvadrater	Beste delmengde	Lasso	PCR
Kjønn	0.070			0.008
Alder	0.317	0.356	0.271	0.214
Sigaretter per dag	0.002			-0.080
BMI	0.259	0.275	0.196	0.171
Diabetiker	0.010			0.040
Glukose-nivå	0.059			0.047
Utdanningsnivå	-0.034			-0.148
Kolesterol-nivå	0.099		0.043	0.176
Test Error	0.723	0.720	0.733	0.776
Std Error	0.039	0.039	0.040	0.041

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

$$X_1 \cdot X_1, X_1 \cdot X_2, \dots$$

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

$$X_1 \cdot X_1, X_1 \cdot X_2, \dots$$

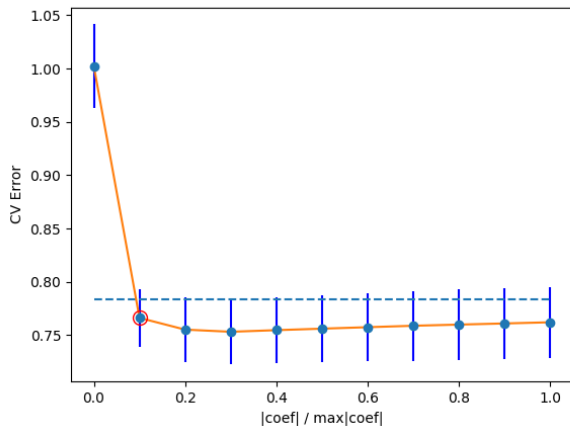
Disse prediktorene ble brukt til å danne en ny matrise **X** bestående av både de originale og de nye prediktorene, tilsammen 44 prediktorer.

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

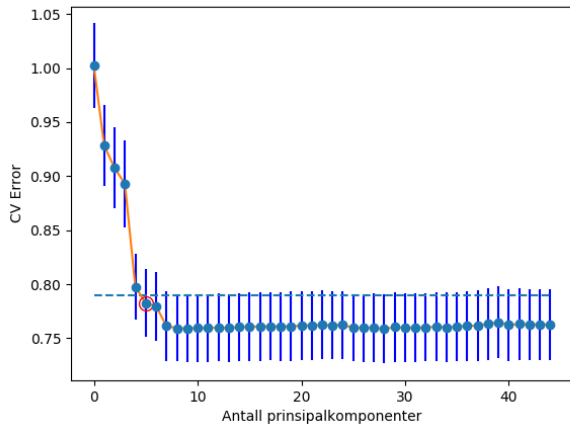
$$X_1 \cdot X_1, X_1 \cdot X_2, \dots$$

Disse prediktorene ble brukt til å danne en ny matrise **X** bestående av både de originale og de nye prediktorene, tilsammen 44 prediktorer. Jeg utførte så Lasso-regresjon og PCR på denne nye matrisen.

Lasso-regresjon (del 2)



Prinsipalkomponent-regresjon (del 2)



	Lasso	PCR
Test Error	0.716	0.773
Std Error	0.039	0.039

Vi ser at test-erroren fra lasso-regresjonen ble forbedret fra 0.733 til 0.716.