

# **Machine learning - DTU**

Rapport

**Anna Louise Hansen**

Fra Udviklings- og Forenklingsstyrrelsen

# Chapter 1

## Part I

### 1.1 Beskrivelse af data

Alle boligejere i Danmark betaler en skat, ejendomsværdiskat, som er baseret på værdien af deres ejendom. Dette vil sige hele ejendommen inkl. grunden som boligen ligger på. For at kunne gøre dette laver den danske stat offentlige ejendomsvurderinger som disse skatter bliver baseret på. Det er derfor vigtigt at disse vurderinger er retvisende og ikke mindst forklarbare, således at en borgers kan forstå hvilke parametre der ligger til grund for ejendomsvurderingen. Til dette projekt har jeg valgt at arbejde med anonymiseret data fra mit arbejde i udviklings- og forenklingsstyrrelsen, hvor jeg til dagligt arbejder med netop dette. Datasættet består af ejendomssalg fra en 6 årig periode. Ud over selve huspriserne består data også af en lang række attributter som beskriver karakteristika ved selve boligen. Det kan f.eks. være tagmateriale, boligens opførelsesår, information om størrelsen af huset og grunden eller bbr koder som dækker over boligens anvendelse. Der ud over består data også af en lang række attributter som fortæller noget om hvor boligens beliggenhed. Det kan f.eks. være boligens koordinater eller information om afstanden til kyst og skov eller afstand til motorvej og jernbane. Data kommer fra en række forskellige registre og offentlige styrrelser som eks. BBR og Styrrelsen for Dataforsyning og Effektivisering.

Til dette projekt vil jeg overordnet set prøve at se hvor godt man kan forudsige ejendomsværdier ud fra salgspriserne fra en 6-årig periode.

Jeg vil med Principal Component Analysis få et overblik over de data der er til rådighed og få et visuelt overblik over attributterne. Herefter vil jeg med en unsupervised learning forsøge at gruppere det data jeg har til at generer yderlige attributer som kan indgå i modellen. Jeg vil her specifikt prøve at se om det er muligt at gruppere salgene i forskellige boligtyper. Jeg vil i samme omgang også forsøge at frasorterer outliers i data med anomaly detection. Herefter vil jeg med regressions model forsøge at kaste lys over projektets overordnede problem ved at forsøge at forudsige huspriserne ud fra salgspriser. I tilfælde af at modellen ikke ikke kan komme med en god prædiktion af en given ejendom vil det være

muligt at denne ejendom bliver manuelt værdiansat af en sagsbehandler. Jeg vil derfor til slut med en classifikation forsøge at estimerer om en ejendom skal ud til manuel sagsbehandling baseret på dens estimerede ejendomsværdi.

## 1.2 Detaljeret beskrivelse af data

Det salgsdata som jeg har valgt at arbejde med dækker i udgangspunktet ‘r nrow(train)‘ observationer med ‘r ncol(train)‘ attributter. Inden jeg går i gang med at kigge på data har jeg valgt at lave en oprydning i data. Det har jeg gjort fordi mange af attributterne bliver i mit daglige arbejde brugt i forbindelse med imødekomme diverse forretningskrav. Desuden dækker observationerne mange forskellige typer af ejendomssalg. Det er en blanding af parcelhussalg, rækkehus-salg, sommerhussalg, salg af ejerlejligheder mm. og ud fra et forretningsmæssigt perspektiv giver det ikke mening at træne en model på alle salg og ejendomstypen vil påvirke salgsprisen. F.eks. vil der på sommerhuse være restriktioner på hvor meget om året man må bo i sommerhuset og der kan være i sommerhusområder være andre regler for hvad man må bruge sin grund til end der er i et parcelhusområde. Jeg har derfor ligeledes valgt at reducerer antallet af observationer således at de kun dækker almindelige parcelhus. Dette er gjort ved kun at beholde alle de ejendomssalg, hvor ejendommen i BBR er registreret med enheds- og bygningsanvendelsen 120.

Herefter er der ‘r nrow(train)‘ observationer tilbage og ‘r ncol(train)‘ attributter.

Der er for en del af attributterne en stor andel af manglende værdier. Det er især i forhold til variable fra BBR, som beskriver forskellige karakteristika ved selve boligen. Her har jeg har valgt at fjerne alle de attributter som har mere end 95% manglende værdier.

|      | attributes                                 | descrete | continous | att    |
|------|--|----------|-----------|--------|
| :--- | :-----                                     | :-----   | :-----    | :----- |
| 72   | aux.adresse.etage                          | descrete |           | nom    |
| 99   | enhed.supplerendevarme                     | descrete |           | nom    |
| 98   | enhed.opvarmningsmiddel                    | descrete |           | nom    |
| 101  | enhed.varmeinstallation                    | descrete |           | nom    |
| 28   | aux.ice_info.adresse.afstand_mellem_soe    | descrete |           | rat    |
| 86   | bygning.afloeksforhold                     | descrete |           | nom    |
| 27   | aux.ice_info.adresse.afstand_lille_soe     | descrete |           | rat    |
| 89   | bygning.supplerendevarme                   | descrete |           | nom    |
| 95   | enhed.energiforsyning                      | descrete |           | nom    |
| 88   | bygning.opvarmningsmiddel                  | descrete |           | nom    |
| 105  | fremskreven_pris                           |          | continous | int    |
| 106  | fremskreven_pris_M2                        |          | continous | int    |
| 104  | aux.vurbenyttelseskode                     |          | descrete  | nom    |
| 34   | aux.ice_info.adresse.udsigtslinjer_hav     |          | descrete  | rat    |
| 36   | aux.ice_info.adresse.udsigtslaengde_samlet |          | descrete  | rat    |

|     |   |          |     |
|-----|---|----------|-----|
| 37  | aux.ice_info.adresse.udsigtslinjer_soe                      | descrete | rat |
| 1   | aux.adresse.etr89koordinatnord                              | descrete | int |
| 2   | aux.adresse.etr89koordinatoest                              | descrete | int |
| 69  | EV_NN_M2  | descrete | int |
| 73  | aux.adresse.regionskode                                     | descrete | nom |
| 94  | enhed.badeforhold   | descrete | nom |
| 100 | enhed.toiletforhold   | descrete | nom |
| 92  | bygning.varmeinstallation                                   | descrete | nom |
| 3   | aux.ice_info.adresse.afstand_kyst                           | descrete | rat |
| 4   | aux.ice_info.adresse.afstand_lokalvej                       | descrete | rat |
| 5   | aux.ice_info.adresse.afstand_trafikvej_fordeling            | descrete | rat |
| 6   | aux.ice_info.adresse.afstand_trafikvej_gennemfart           | descrete | rat |
| 7   | aux.ice_info.adresse.afstand_motorvej_motortrafikvej        | descrete | rat |
| 8   | aux.ice_info.adresse.afstand_jernbane_hovedbane             | descrete | rat |
| 9   | aux.ice_info.adresse.afstand_jernbane_any                   | descrete | rat |
| 10  | aux.ice_info.adresse.afstand_jernbane_lokalbane             | descrete | rat |
| 11  | aux.ice_info.adresse.afstand_jernbane_metro                 | descrete | rat |
| 12  | aux.ice_info.adresse.afstand_jernbane_regionalfbane         | descrete | rat |
| 13  | aux.ice_info.adresse.afstand_jernbane_region_privat         | descrete | rat |
| 14  | aux.ice_info.adresse.afstand_jernbane_s_bane                | descrete | rat |
| 15  | aux.ice_info.adresse.afstand_stor_skov                      | descrete | rat |
| 16  | aux.ice_info.adresse.afstand_lille_skov                     | descrete | rat |
| 17  | aux.ice_info.adresse.afstand_lokalnet_hoejspaending         | descrete | rat |
| 18  | aux.ice_info.adresse.afstand_regionalfnet_hoejspaending     | descrete | rat |
| 19  | aux.ice_info.adresse.afstand_transmissionsnet_hoejspaending | descrete | rat |
| 20  | aux.ice_info.adresse.afstand_lille_vandloeb                 | descrete | rat |
| 21  | aux.ice_info.adresse.afstand_mellem_vandloeb                | descrete | rat |
| 22  | aux.ice_info.adresse.afstand_stort_vandloeb                 | descrete | rat |
| 23  | aux.ice_info.adresse.afstand_ukendt_vandloeb                | descrete | rat |
| 24  | aux.ice_info.adresse.afstand_station_metro                  | descrete | rat |
| 25  | aux.ice_info.adresse.afstand_station_s_tog                  | descrete | rat |
| 26  | aux.ice_info.adresse.afstand_station_tog                    | descrete | rat |
| 29  | aux.ice_info.adresse.afstand_stor_soe                       | descrete | rat |
| 30  | aux.ice_info.adresse.afstand_lille_vindmoelle               | descrete | rat |
| 31  | aux.ice_info.adresse.afstand_mellem_vindmoelle              | descrete | rat |
| 32  | aux.ice_info.adresse.afstand_stor_vindmoelle                | descrete | rat |
| 33  | aux.ice_info.adresse.areal_samlet_skov                      | descrete | rat |
| 35  | aux.ice_info.adresse.udsigtslaengde_hav                     | descrete | rat |
| 38  | aux.ice_info.adresse.udsigtslaengde_soe                     | descrete | rat |
| 39  | aux.ice_info.bygning.etager.arealaflovligbeboelseikaelder   | descrete | rat |
| 40  | aux.ice_info.bygning.etager.arealafudnyttetdelaftagetage    | descrete | rat |
| 41  | aux.ice_info.bygning.etager.etagensadgangsareal             | descrete | rat |
| 42  | aux.ice_info.bygning.etager.kaelderareal                    | descrete | rat |
| 43  | aux.ice_info.bygning.etager.samletarealaftagetage           | descrete | rat |
| 44  | aux.ice_info.byzoneareal_opsummeret_delj                    | descrete | rat |
| 45  | aux.ice_info.landzoneareal_opsummeret_delj                  | descrete | rat |

|     |   |           |     |
|-----|---|-----------|-----|
| 46  | aux.ice_info.sommerhuszoneareal_opsummeret_delj               | descrete  | rat |
| 47  | aux.ice_info.jordstykker.registreretareal_fratrukket_vejareal | descrete  | rat |
| 48  | aux.ice_info.jordstykker.vejareal                             | descrete  | rat |
| 49  | bolig_alder   | descrete  | rat |
| 50  | bolig_areal   | descrete  | rat |
| 51  | bygning.antaletager   | descrete  | int |
| 52  | bygning.arealindbyggetcarport                                 | descrete  | rat |
| 53  | bygning.arealindbyggetgarage                                  | descrete  | rat |
| 54  | bygning.bebyggetareal   | descrete  | rat |
| 55  | bygning.bygningenssamledeboligareal                           | descrete  | rat |
| 56  | bygning.bygningenssamledeerhvervsareal                        | descrete  | rat |
| 57  | bygning.omtilbygningsaar                                      | descrete  | int |
| 58  | bygning.opfoerelsesaar  | descrete  | int |
| 59  | bygning.samletbygningsareal                                   | descrete  | rat |
| 60  | enhed.antalbadevaarelser                                      | descrete  | rat |
| 61  | enhed.antalvaarelser  | descrete  | rat |
| 62  | enhed.antalvandskylledetoiletter                              | descrete  | rat |
| 63  | enhed.arealtilbeboelse  | descrete  | rat |
| 64  | enhed.arealtilerhverv   | descrete  | rat |
| 65  | enhed.enhedenssamledeareal                                    | descrete  | rat |
| 66  | etage.arealafudnyttedelaftagetaage                            | descrete  | rat |
| 67  | etage.etagensadgangsareal                                     | descrete  | rat |
| 68  | etage.kaelderareal  | descrete  | rat |
| 70  | ice_info.min_koteletratiobuff250                              | continous | int |
| 71  | ombyg_alder   | descrete  | int |
| 74  | aux.ice_info.bb_anvendelse_0                                  | descrete  | rat |
| 75  | aux.ice_info.bb_anvendelse_910                                | descrete  | rat |
| 76  | aux.ice_info.bb_anvendelse_920                                | descrete  | rat |
| 77  | aux.ice_info.bb_anvendelse_930                                | descrete  | rat |
| 78  | aux.ice_info.bb_anvendelse_940                                | descrete  | rat |
| 79  | aux.ice_info.bb_anvendelse_950                                | descrete  | rat |
| 80  | aux.ice_info.bb_anvendelse_960                                | descrete  | rat |
| 81  | aux.ice_info.bb_anvendelse_970                                | descrete  | rat |
| 82  | ice_info.bb_per_delgrund                                      | descrete  | rat |
| 83  | ice_info.en_per_delgrund                                      | descrete  | rat |
| 84  | ice_info.et_per_delgrund                                      | descrete  | rat |
| 85  | ice_info.js_per_delgrund                                      | descrete  | rat |
| 87  | bygning.bygningensanvendelse                                  | descrete  | nom |
| 90  | bygning.tagdaekningsmateriale                                 | descrete  | nom |
| 91  | bygning.vandforsyning   | descrete  | nom |
| 93  | bygning.ydervaeggensmateriale                                 | descrete  | nom |
| 96  | enhed.enhedensanvendelse                                      | descrete  | nom |
| 97  | enhed.koekkenforhold  | descrete  | nom |
| 102 | etage.bygningensetagebetegnelse                               | descrete  | nom |
| 103 | region_nr   | descrete  | nom |

Som udgangspunkt vil jeg træne en model som skal være i stand til at kunne prædiktere værdien af et standard parcelhus. Data som der skal trænes på skal derfor også være salg af standard parcelhuse. Data er derfor blevet ensrettet på følgende måde:

Slutteligt er der blevet taget et valg om kun at udvælge de rå attributter som menes at være de vigtigste i forhold til at forudsige værdien af et standad parcelhus.

| attributes  | descrete_continous | attribut |
|---|--------------------|----------|
| -----   | -----              | -----    |
| fremskreven_pris_M2   | continous          | interval |
| bygning.ydervaeggensmateriale                                 | descrete           | nomial   |
| bygning.tagdaekningsmateriale                                 | descrete           | nomial   |
| ombyg_alder   | descrete           | interval |
| enhed.antalvaarelser  | descrete           | ratio    |
| enhed.antalbadevaarelser                                      | descrete           | ratio    |
| enhed.antalvandskylledetoiletter                              | descrete           | ratio    |
| bolig_areal   | descrete           | ratio    |
| bolig_alder   | descrete           | ratio    |
| aux.ice_info.jordstykker.registreretareal_fratrukket_vejareal | descrete           | ratio    |
| aux.ice_info.adresse.afstand_motorvej_motortrafikvej          | descrete           | ratio    |
| aux.ice_info.adresse.afstand_kyst                             | descrete           | ratio    |
| EV_NN_M2  | descrete           | interval |

De to attributter der dækker over tagtypematriale og ydervæggsmateriale er diskrete variable som fordel kan normaliseres med en one-out-of-k transformering.

Afstand til kyst og afstand til motorvej er to variable som jeg har valgt at binariserer. Det har jeg da de to features forud for denne rapport er blevet imputeret. For afstand til kyst er afstanden op til 1500 meter målt. Alt herover er imputeret til 1501 meter. Ligeledes er gjort for afstand til motorvej. For at håndtere disse variable har jeg som sagt valgt at binariserer dem. For afstand til kyst har jeg ud fra et forretningsmæssigt synspunkt valgt at en ejendom ligger så tæt på kysten at det har en effekt i dens værdi hvis den ligger inden for 300 meter af kysten. Det samme er gjort for afstand til vej. Her er grænsen dog blot 100 meter.

### 1.3 Data visualisering heriblandt Principal Component Analysis (PCA)

Principal Component Analysis (PCA) er en metode som kan bruges til at reducere dimensionerne i det data der arbejdes med. Man kan have mange dimensioner i sit datasæt, men hvis de alle sammen er med til at forklare sammen tendens er det 'sande' antal af dimensioner lavere end antallet af attributter. Målet med at lave PCA er at reducere dimensionerne i data uden at reducere

variationen, således at man ender op med data som med færre dimensioner men som kan forklare det samme som datasættet med flere dimensioner. PCA fungerer kun ud fra antagelsen om at der er en linear forklaring i data med færre dimensioner. De bedste projektion af data ned på et subspace er dem hvor observationer er spredt ud (Høj varians), mens samtidig hvor residualerne reduceres. Vektoren bliver valgt ud fra at den skal være en eigenvektor til vores datamatrix som har den højeste eigenværdi. Singular Value Decomposition (SVD) er en metode som for en hvilken som helst  $N * M$  matrix udregner eigenvektoren med den højeste eigenværdi.

Ejendomsdata er blevet klargjort. Data er blevet transformeret. Nogle variable er blevet transformeret med one-out-of-K transformation, mens enkelte er blevet binariseret. Til PCA er det første trin at standardisere data, således at attributernes værdier er på samme skala. Selve standardiseringen består i at trække gennemsnittet fra hver attribut, hvorefter der også er blevet divideret med standardafvigelsen. For data betyder det at hver attribut reskaleres således at de får et gennemsnit på 0 og en standardafvigelse på 1. Årsagen til at en reduktion af dimensionerne er ønskværdig er at det for nogle typer af algoritmer kan være med til at forøge deres nøjagtighed. Dette er eksempelvis tilfældet med xgboost algoritmen.

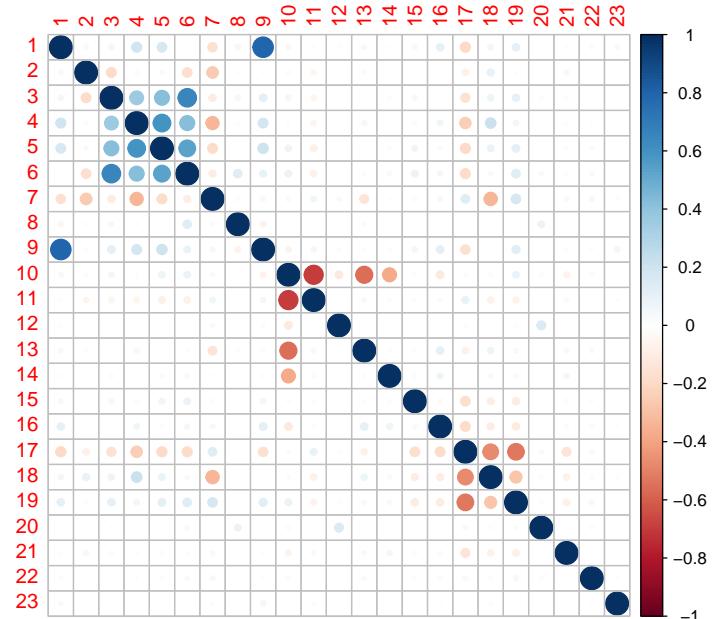
Efter alle datatransformationerne består data af 185018 observationer (N) med 23 features (M). Dette data skal senere danne grundlaget for regressionsanalysen, men inden da bliver der med en korrelationsanalyse og en PCA taget stilling til hvorvidt det er muligt at reducere dimensionerne i data. Resultatet af korrelationsanalysen er vist i et korrelationsplot. Resultatet af Korrelationsanalysen viser at der er en stor positiv korrelation mellem den fremskrevne kvadratmeter pris og den vægtede gennemsnitspris for de nærmeste naboor. Der er desuden også en større positiv sammenhæng mellem antallet af værelser og boligarealet. Disse to positive sammenhænge giver logik rigtig god mening. Salgspriser er i høj grad styret af det område som ejendommen ligger i. Ligger ejendommen i et dyrt område, vil naboorne blive solgt til høje handelspriser og det samme vil højst sandsynligt også gælde for den specifikke ejendom. Samtidig vil der typisk også være flere væresler jo større boligarealet en ejendommene har.

```
[1] "fremskreven_pris_M2"
[2] "ombyg_alder"
[3] "enhed.antalvaarelser"
[4] "enhed.antalbadevaarelser"
[5] "enhed.antalvandskyllede toiletter"
[6] "bolig_areal"
[7] "bolig_alder"
[8] "aux.ice_info.jordstykker.registreretareal_fratrukket_vejareal"
[9] "EV_NN_M2"
[10] "ydervaegsmateriale_mursten"
[11] "ydervaegsmateriale_gasbeton"
[12] "ydervaegsmateriale_bindingsvaerk"
```

```

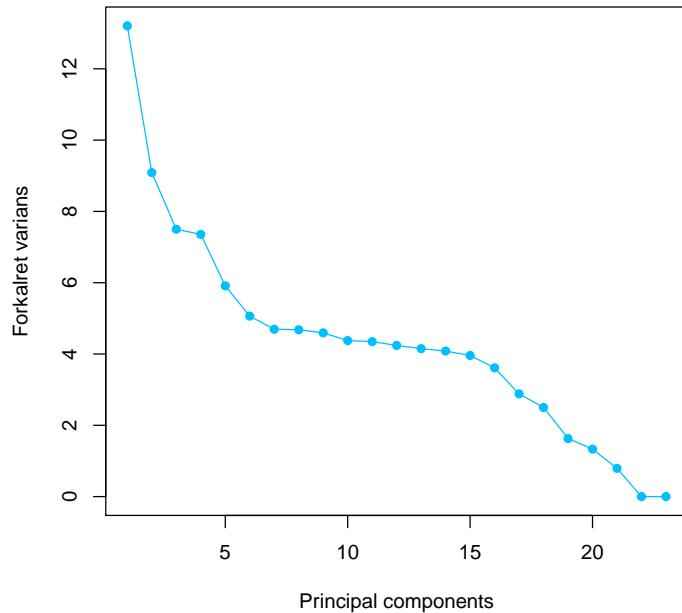
[13] "ydervaegsmateriale_traebeklaedning"
[14] "ydervaegsmateriale_andet"
[15] "tagtype_builtin"
[16] "tagtype_tagpap"
[17] "tagtype_fibercement"
[18] "tagtype_cementsten"
[19] "tagtype_tegl"
[20] "tagtype_straatag"
[21] "tagtype_andet"
[22] "taet_paa_kyst"
[23] "taet_paa_motorvej"

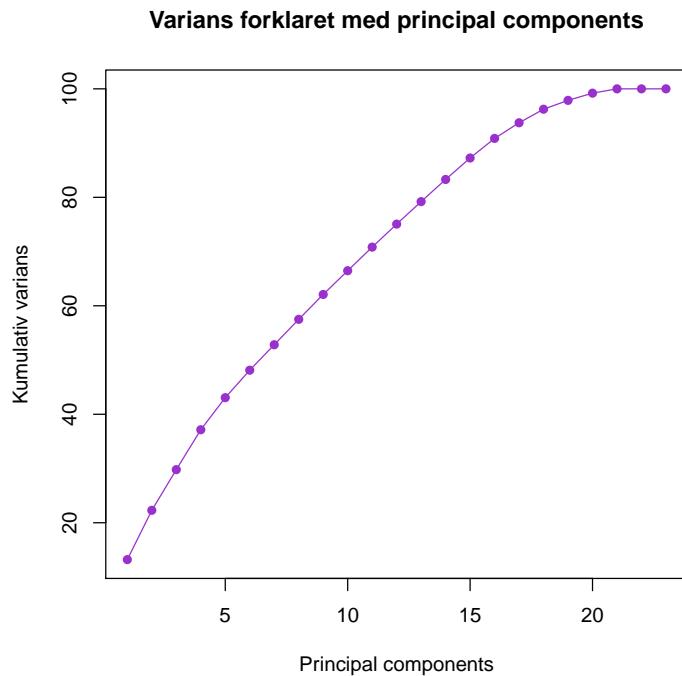
```



Som en del af PCA udregnes herefter Singular Value Decomposition (SVD).

### Varians forklaret med principal components





De føste 16 principal components kan forklare 90% af variationen i data. For at komme over 95% skal man have de 18 første komponenter. Ud af de i alt 23 mulige komponenter er det med dette data ikke muligt at reducerer mange komponenter væk uden også at miste variation i data. Ved at have antallet af komponenter som er mindre end antallet af attributter i ens datasæt bliver information tabt, og hvorvidt man med fordel kan bruge PCA skal bestemmes ud fra den pågældende problemstilling. I den videre opgave har jeg valgt at gå videre med mit originale datasæt som det så ud før PCA.

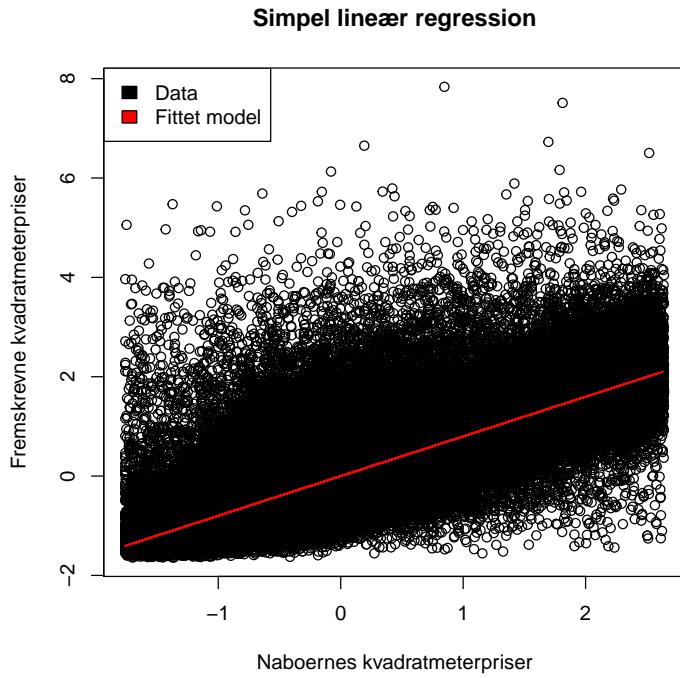
# Chapter 2

## Part II

### 2.1 Regression - part A

I part II er formålet at bruge det rensede data fra part I til at forudsige fremskrevne handelspriser ud fra forskellige variablene. Håbet med denne regressionsanalyse er at man ud fra relativt få variable og en relativt simpel model vil kunne forudsige ejendomspriserne. Forud for regressionsanalysen er data blevet transformeret. For faktorvariablene tagtype og vægmateriale har jeg valgt at transformere med en one-of-k transformering. Herefter er alle attributer blevet standardiseret, således at de har en gennemsnit på 0 og en standardafvigelse på 1.

Den første lineaære model der fittes er en univariate linear regressionsmodel. Her er naboernes områdepris den eneste variabel som bruges til at forudsige de fremskrevne handelspriser. Denne simple lineaære regression er vist i figuren nedenfor.



Den anden lineære model der fittes er en multivariate lineær regressionsmodel. Variablene som bliver brugt i modellen er områdepriser i form af naboernes kvadratmeterpriser. Det er boligens opførelsesalder og ombygningsår, og det er boligens og grundens areal. I en multivariate lineær regressionsmodel kan man ikke på samme måde plotte den fitte model på det todimensionelle plot. Her kan man i stedet estimerer vide hvor godt modellen fitter til data ved at man ved at minimere summen af de kvadrerede afvigelser (RSS). Der findes flere forskellige typer af algoritmer hvis formål er at finde de parametre/vægte som laver det bedste fit til data ved at minimerer 'cost'.

|   | fremskrevnen_pris_M2 | predicted  | residuals  |
|---|----------------------|------------|------------|
| 1 | 0.9571538            | 0.5640135  | 0.3931403  |
| 2 | -0.6903684           | 1.5952953  | -2.2856636 |
| 4 | -1.3992891           | -0.9897822 | -0.4095069 |
| 5 | -0.4448076           | -0.3181251 | -0.1266825 |
| 6 | -1.0156162           | -0.3257869 | -0.6898292 |
| 8 | -0.8392054           | -0.6190015 | -0.2202038 |

Call:  
`lm(formula = fremskrevnen_pris_M2 ~ EV_NN_M2 + bolig_areal + aux.ice_info.jordstykker.registreringer.bolig_alder + bolig_alder, data = train)`

Residuals:

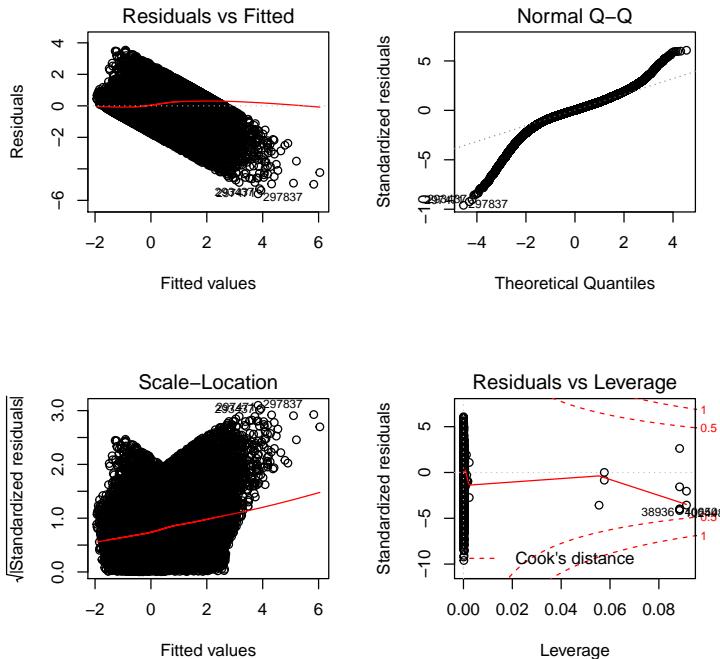
|  | Min     | 1Q      | Median | 3Q     | Max    |
|--|---------|---------|--------|--------|--------|
|  | -5.5851 | -0.2914 | 0.0178 | 0.3322 | 3.5360 |

Coefficients:

|   | Estimate         |
|---|------------------|
| (Intercept)   | 3.834e-15        |
| EV_NN_M2  | 8.010e-01        |
| bolig_areal   | -8.561e-02       |
| aux.ice_info.jordstykker.registreretareal_fratrukket_vejareal | 8.883e-03        |
| bolig_alder   | -1.371e-01       |
|   | Std. Error       |
| (Intercept)   | 1.352e-03        |
| EV_NN_M2  | 1.363e-03        |
| bolig_areal   | 1.378e-03        |
| aux.ice_info.jordstykker.registreretareal_fratrukket_vejareal | 1.369e-03        |
| bolig_alder   | 1.360e-03        |
|   | t value Pr(> t ) |
| (Intercept)   | 0.000 1          |
| EV_NN_M2  | 587.889 < 2e-16  |
| bolig_areal   | -62.144 < 2e-16  |
| aux.ice_info.jordstykker.registreretareal_fratrukket_vejareal | 6.487 8.76e-11   |
| bolig_alder   | -100.860 < 2e-16 |
| <br>  |                  |
| (Intercept)   | ***              |
| EV_NN_M2  | ***              |
| bolig_areal   | ***              |
| aux.ice_info.jordstykker.registreretareal_fratrukket_vejareal | ***              |
| bolig_alder   | ***              |
| <br>  |                  |
| ---   |                  |
| Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |                  |

Residual standard error: 0.5815 on 185013 degrees of freedom  
Multiple R-squared: 0.6618, Adjusted R-squared: 0.6618  
F-statistic: 9.052e+04 on 4 and 185013 DF, p-value: < 2.2e-16

[1] 0.3381834



Evalueringen af den lineære model kan foregå ved at man træner på et datasæt og prædikterer på et andet. I følgende regression har jeg nabokvadratmeterpriserne og boligarealet med som prediktorvariable. Ud fra den estimerede model kan man plotte sin afhængige variabel mod de prædikterede værdier. Forskellen mellem den faktiske afhængige variabel (i dette tilfælde de fremskrevne salgspiser) og den prædikterede variabel (modellens estimerede  $y$ ) er residualet. Selve den lineære regressionsmodel kan anvances ved at transformere inputvariablene. Det kan eksempelvis være at opløfte variablene i anden potens og bruge dem som prediktor sammen med den originale version af variablen. Målet for denne simple lineære regressionsmodel er reducere 'cost-funktionen' så meget som muligt. For en simpel lineær regressionsmodel er cost-funktionen en squared error, hvilket vil sige at målet her er at reducere 'mean squared error' så meget som muligt - dog uden at overfitte. Andre cost-funktioner kan bruges til andre problemstillinger. En logistisk funktion kan således bruges til at hvis man arbejder med et logistisk regressionsproblem.

Ridge regression - ridge regression er regressionsmetode som bruges for at undgå at man overfitter. Overfitting sker når den fittede model fitter træningsdata rigtig godt, men at den model der trænes ikke generalisere godt på ukendt data. Overfitting kan ske både hvis der er for mange prædiktor variable eller for få observationer. Et godt fit er kendtegnet ved et lav RSS, men også størrelsen på koefficenterne - disse to til sammen udgør den samlede 'cost'. I ridge regression bruges l2-normen som et mål for størrelsen på koefficenterne

og målet med ridge regression er minimere den samlede 'cost'. Dette kan gøres ved at indføre en regulariseringsparameter (lambda), og den kan bruges til at styre kompleksiteten af sin model.

## **2.2 Regression - part B**

## **2.3 Classification**

## **Chapter 3**

### **Part III**