IST[A/G/T]1003: Statistisk læring og data science

Regresjon

Mette Langaas IMF/NTNU

19 October, 2020

Contents

Læringsmål

Etter du har lest dette kompendiet, sett videoene som er laget og deltatt på zoom-forelesningene skal du

- kunne forstå hva regresjon går ut på og
- kjenne igjen situasjoner der regresjon vil være en aktuell metode å bruke
- kjenne til modellen for multippel lineær regresjon, og kunne tolke utvalgte deler av en generell utskrift fra modelltilpasning
- forstå hvordan vi utfører multippel lineær regresjon i Python
- kunne svare godt på regresjonsoppgaven i den tellende prosjektoppgaven

Introduksjon

Videoressurser

Videoer som diskuterer innholdet av dette kompendiet:

For enkel lineær regresjon finnes 4 videoer i fellesdelen. Disse er

- Del 1
- Del 2
- Del 3
- Del 4

Det er i tillegg laget en video for utvidelsen fra enkel til multippel lineær regresjon

- Multippel lineær regresjon: introduksjon
- Multippel lineær regresjon: analyse av et datasett

Hvorfor regresjon?

Det er to hovedgrunner til at vi vil utføre en regresjonsanalyse:

1. Vi vil lage en modell for å hjelpe oss med å **forstå sammenhengen** mellom *en respons* og *en eller flere forklaringsvariabler*.
2. Vi vil lage en modell for å **predikere** en *respons* fra en eller flere *forklaringsvariabler* (mer eller mindre sort boks).

Lineær regresjon er en veiledet og parametrisk metode, og er byggesteinen for veiledede metoder innen statistisk læring og maskinlæring.

Eksempel: Leieindeks for leiligheter

I storbyer i Tyskland er det vanlig å leie en leilighet (ikke eie), og slik er det også i München. I Norge har vi en annen ordning, og der er det mer vanlig å eie enn å leie.

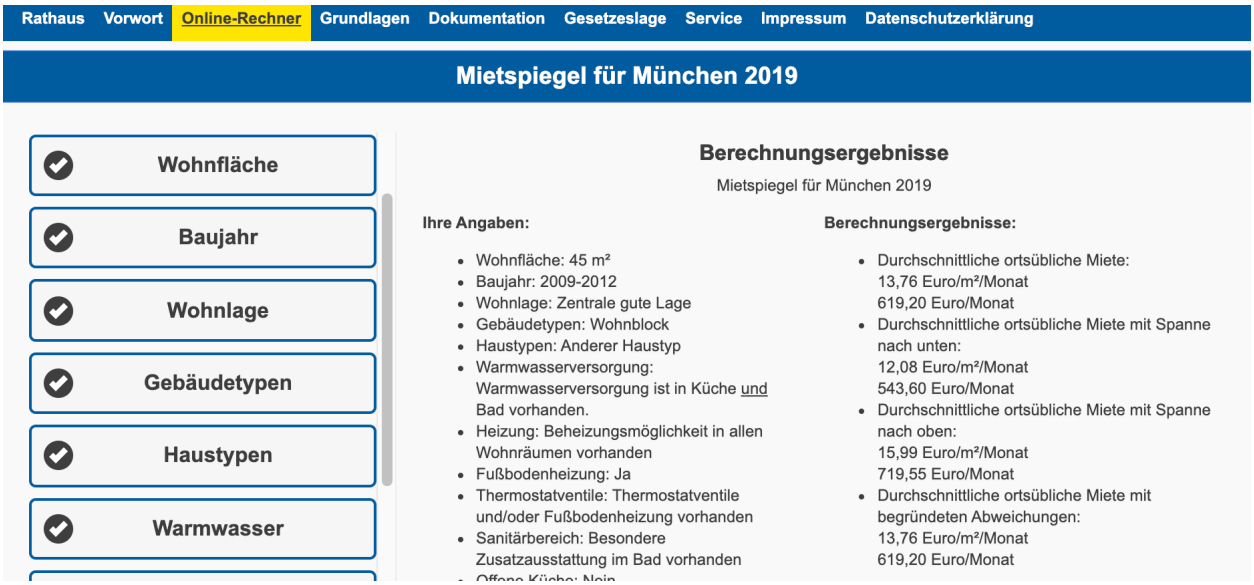
Myndighetene i München er svært opptatte av å veilede både leietaker og utleier slik at de kommer frem til en rettferdig markedspris for leiligheten som skal leies ut, og de har laget et "leie-speil" https://www.mietspiegel-muenchen.de der man kan fylle ut en laaaang rekke informasjon om leiligheten. Deretter får man laget en rapport med informasjon om hva gjennomsnittsleie er for en leilighet av denne typen.

Hvordan har myndighetene kommet frem til informasjonene de gir i leie-speilet?

De har over mange år samlet inn data fra utleiere, både detaljer om leiligheten som leies ut og hvilen leie som betales. Men, hvordan har de funnet frem til en utregning av gjennomsnittspris?

En mulighet er at de har brukt alle innsamlede data til å lage en prediksjon for hva leien bør være for en leilighet - basert på alle data om leiligheten. Siden det er leien vi vil predikere - og det er en kontinuerlig størrelse (her i Euro), kan vi tenke at vi kan bruke leien som respons i en regresjon - der vi har mange egenskaper til leiligheten som forklaringsvariabler (også kalt kovariater eller prediktorer).

Vi skal se på et datasett fra leieindeksen i München i 1999, presentert i Fahrmeir, Kneib, Lange, Marx (2013). Datasettet er et representativt utvalg av 3082 leiligheter fra de tilgjengelige dataene i 1999. Grunnen



til å bruke disse dataene er at datasettet innholder interessante muligheter til å utforske ulike aspekter av multippel lineær regresjon, og forstå hva regresjon kan brukes til i samfunnet.

VI skal se på følgende variabler som beskriver leiligheter i 1999:

- `rent`: leien (Euro)
- `area`: areal (m^2)
- `location`: beliggenhet (1=gjennomsnittlig, 2=god, 3=topp)
- `bath`: kvalitet av badet (0=standard, 1=premium)
- `kitchen`: kvalitet av kjøkkenet (0=standard, 1=premium)
- `cheating`: sentralvarme (0=ingen sentralvarme, 1=med sentralvarme)

Vi skal bruke dataene med `rent` som respons, og så en, flere eller alle de andre variablene som forklaringsvariabler.

Enkel lineær regresjon

Dette er for det meste repetisjon fra siste uke i fellesmodulen, husk at

- "enkel"=´{e}n forklaringsvariabel og at vi bruker ordene
- forklaringsvariabel, kovariat, prediktor for vår x. (Kjært barn har mange navn.)

Hvorfor lineær sammenheng?

Grunnen til å bruke en enkel lineær modell kan være at

- sammenhengen er en naturlov, som $v = v_0 + at$ eller $Pv = nRT$ (som blir lineær hvis vi tar logaritmen på begge sider)
- det er en lokal tilnærming
- vi har korrelert størrelser (høyde mot vekt eller høyde mot skonummer, areal av leilighet mot leiebeløp)
- det er trender over tid (nedgang i boligrenta?)

Enkel lineær modell

Vi har tidligere definert en *enkel lineær regresjonsmodell* - med to ulike skrivemåter. Vi vil her bruke følgende notasjon

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

- Y_i er en *kontinuerlig* responsvariabel
- β_0 er skjæringspunktet med den vertikale aksen (også kalt konstantledd). Når $x = 0$ blir vårt beste gjett på Y lik β_0.
- β_1 er stigningstallet og gir gjennomsnittlig økning i Y når x øker med en enhet
- x_i er kovariat (forklaringsvariabel, prediktor) og kan være kontinuerlig eller diskret (også kategorisk - men det krever litt omkoding)
- e_i feilleddet (kalles noen ganger støyledd) og representerer alle de kovariatene vi ikke har observert men som påvirker Y_i, eventuelt også ulike typer måleusikkerhet, målefeil eller generelt tilfeldig variasjon i Y_i

Vi antar at for observasjon i har vi observert paret (x_i, Y_i), og at observasjonsparene for observasjon $i = 1, \ldots, n$ er observert uavhengig av hverandre. For leieindeks-eksemplet blir det da at leien og (for eksempel) forklaringsvariabelen areal til de ulike leilighetene ikke er avhengig av hverandre.

Mulig misforståelse: her er målet vårt å se om Y_i kan forklares fra x_i så antar absolutt ikke at x_i og Y_i er uavhengige, det er observasjonsparene som er uavhengig av hverandre.

Egenskaper til feilleddene

Videre antok vi at feilleddene e_i var stokastiske variabler, med

- forventingsverdi $\mathrm{E}(e_i) = 0$ og
- standardavvik $\mathrm{SD}(e_i) = \sigma$.

Vi antok også at feilleddene var

- normalfordelte, så $e_i \sim N(0, \sigma)$.

Gitt verdien til forklaringsvariablen hadde vi da at den betingede fordelingen til responsen var normal, $Y_i \mid x_i \sim N(\beta_0 + \beta_1 x_i, \sigma)$.

(I den andre skrivemåten startet man heller med denne informasjonen om betinget fordeling.)

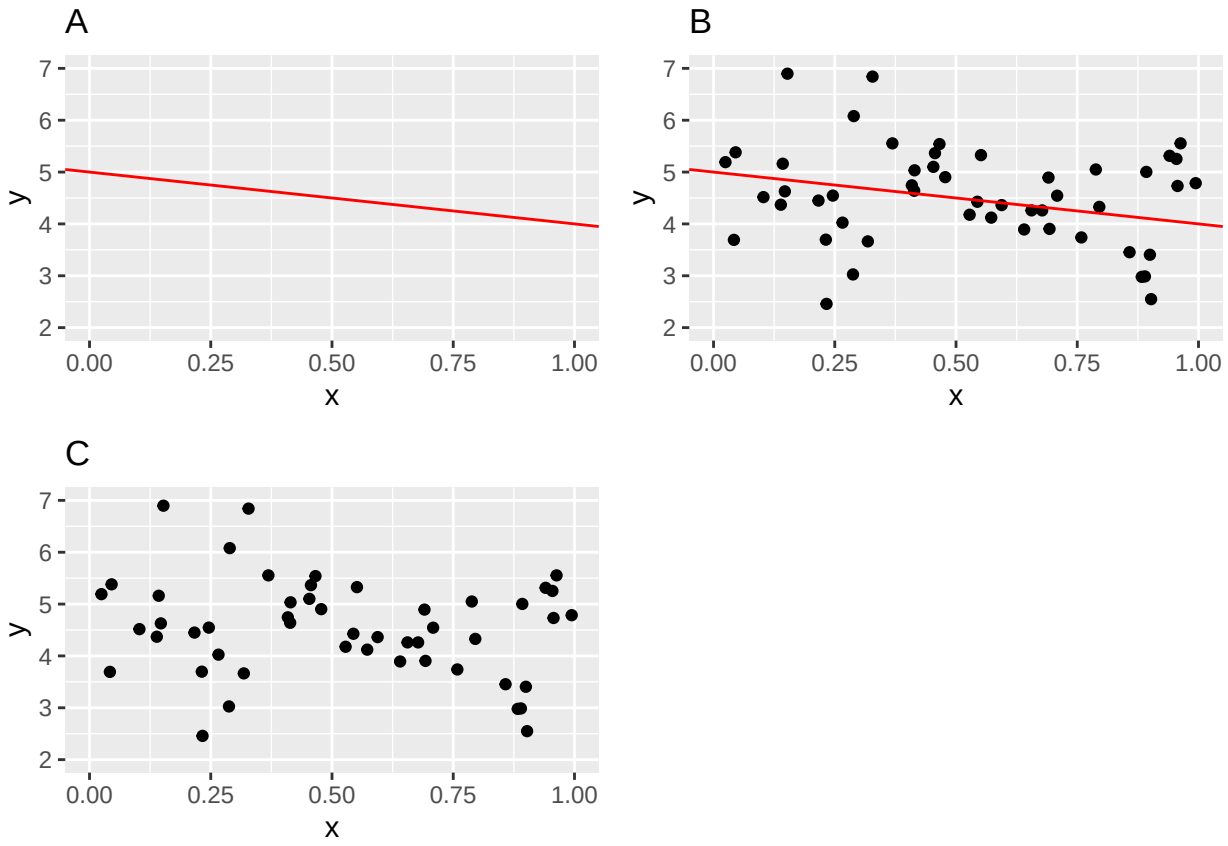
Hva er ukjent? Bare parametrene β_0, β_1, σ.

Simulert eksempel

A) Anta at vi kan bestemme hva den (røde) lineære sammenhengen er - her $\beta_0 + \beta_1 x_i = 5 - x_i$ (dvs. $\beta_0 = 5, \beta_1 = -1$).

B) Anta at vi kan trekke data fra denne lineære sammenhenge for $n = 50$ par av (x_i, y_i)
   - her trekker vi først x_i uniformt fra 0 til 1
   - så regner vi ut den forventede linjen $\mu_i = 5 - x_i$
   - så legger vi til normalfordelte feilledd med gitt varians, her $e_i \sim N(0, \sigma)$ med $\sigma = 1$

C) Men, så vet vi jo ikke den sanne linjen så den fjerner vi, og så er oppgaven å finne frem til den rette linjen som "best" passer til dataene - gitt at dataene er fra en slik modell som vi har satt opp - der vi ikke kjenner de tre ukjente parametrene β_0, β_1, σ.



Parameterestimering

Vi bruker en "hatt", $\hat{}$, over en parameter for å si at det er et estimat for parameteren, og en "hatt", $\hat{}$, over en stokastisk variabel for å si at vi har en prediksjon.

I enkel (og multippel) lineær regresjon estimerer vi regresjonsparameterne ved å minimere kvadratisk avvik mellom den tilpassede regresjonslinja og observasjonene.

Estimere regresjonsparameterne

Anta at vi har anslag $\hat{\beta}_0$ og $\hat{\beta}_1$ for regresjonsparameterne.

Notasjon:

- *predikert verdi*: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ er verdi svarende til x_i på regresjonslinja
- *residual*: $\hat{e}_i = y_i - \hat{y}_i$ er differensen mellom observert og predikert verdi av responsen.

Vi tenker på residualet som vårt beste gjett (prediksjon) for feilleddet (som vi ikke kan observere).

Da er vårt kvadratiske avvik SSE (sums of squares of error) definert som summen av kvadratet av residualene

$$\text{SSE} = \hat{e}_1^2 + \hat{e}_2^2 + ... + \hat{e}_n^2 = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Her ser vi at SSE er en funksjon av de to ukjente anslagene våre på regresjonsparametrene: $\hat{\beta}_0$ og $\hat{\beta}_1$. Vi har lært at en mulig løsning på å minimere en funksjon av to ukjente er å derivere og sette de deriverte lik 0.

Hvis man deriverer SSE med hensyn på hver av $\hat{\beta}_0$ og $\hat{\beta}_1$ og setter lik 0, får vi de såkalte normalligningene, som i dette tilfellet har en løsning som kan skrives ut. Dette blir

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

og

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

der $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ er gjennomsnitt for respons og kovariat i datasettet.

Simulert eksempel (forts.)

Først ser vi utskriften fra å "kjøre" en enkel lineær regresjon på de simulerte dataene i Python.

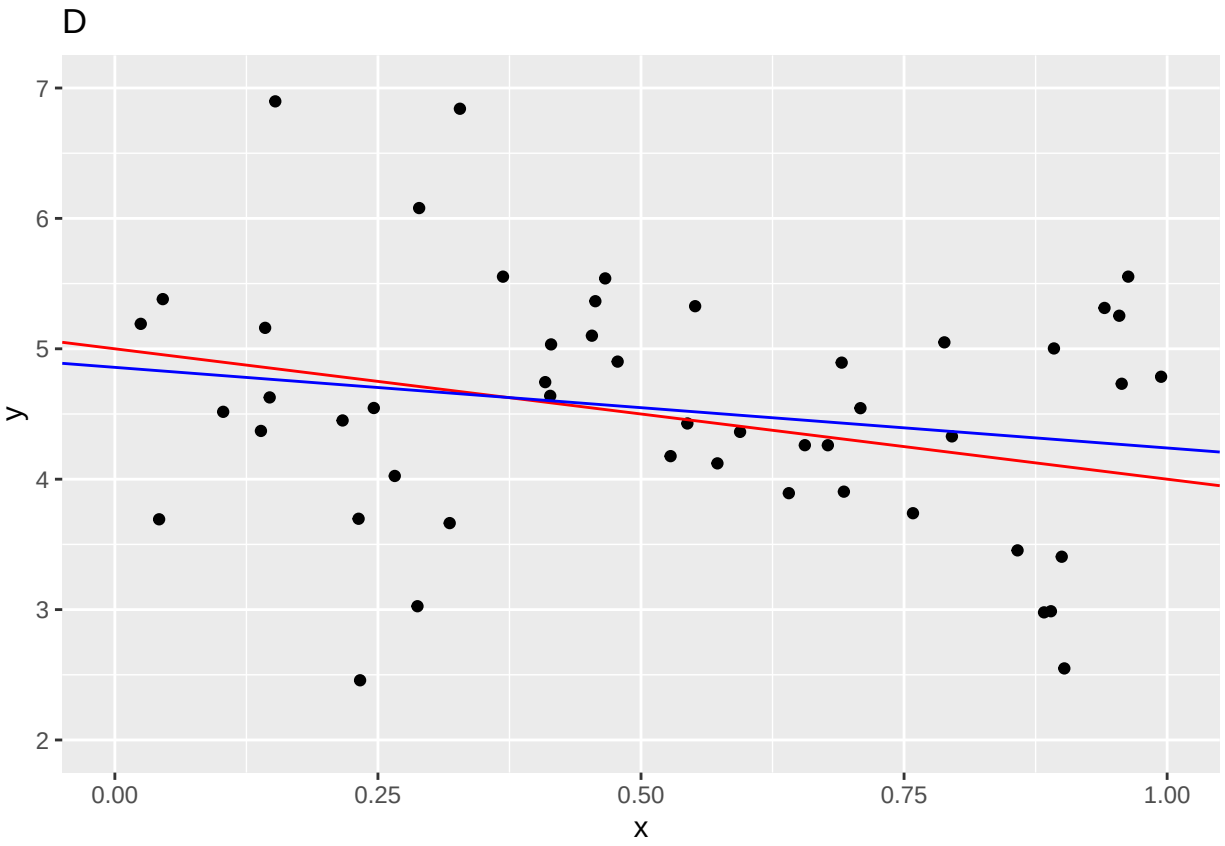
Utskriften har tre paneler, et øvre med informasjon om hva som er gjort og noen godhetsmål, deretter den midtre med fokus på estimater av skjæringspunkt og stigningstall, og til slutt den nedre delen som i hovedsak består av tester for om modellantagelser er oppfylt.

Nå skal vi bare se på den midtre delen og spesielt på kolonnen "coef", og observere at

- $\hat{\beta}_0$ (Intercept) er 4.86 og
- $\hat{\beta}_1$ (x) er -0.62.

Deretter er linjen $\hat{\beta}_0 + \hat{\beta}_1 x$ tegnet inn med blått sammen med observasjonene. I plottet er også den sanne linjen med (i rødt). Den sanne linjen vet vi bare akkurat i dette tilfellet fordi vi har simulert dataene og vet fasit!

```
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                      y   R-squared:                       0.037
## Model:                            OLS   Adj. R-squared:                  0.017
## Method:                 Least Squares   F-statistic:                     1.834
## Date:                Mon, 19 Oct 2020   Prob (F-statistic):              0.182
## Time:                        20:37:00   Log-Likelihood:                -66.852
## No. Observations:                  50   AIC:                             137.7
## Df Residuals:                      48   BIC:                             141.5
## Df Model:                           1
## Covariance Type:            nonrobust
## ==============================================================================
##                  coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## Intercept      4.8576      0.272     17.851      0.000       4.311       5.405
## x             -0.6183      0.456     -1.354      0.182      -1.536       0.300
## ==============================================================================
## Omnibus:                        0.270   Durbin-Watson:                   2.096
## Prob(Omnibus):                  0.874   Jarque-Bera (JB):                0.010
## Skew:                          -0.017   Prob(JB):                        0.995
## Kurtosis:                       3.059   Cond. No.                         4.43
## ==============================================================================
##
## Notes:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```



Usikkerhet i estimatorene for regresjonsparameterne

Vi kan regne ut variansen til $\hat{\beta}_0$ og $\hat{\beta}_1$, men kommer bare til å bruke variansen til $\hat{\beta}_1$ - og den blir:

$$\text{Var}(\hat{\beta}_1) = \text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

En ting å merke seg er at det ofte er slik at estimatet av skjæringspunkt og stigningstall ikke er uavhengige. Er det rart? Nei, endrer vi litt på stigningstallet vil skjæringspunktet til linja også endre seg - når vi skal på beste måte tilpasse en linje til data.

Og, kjenner vi σ egentlig? Nei, den kjenner vi ikke - og derfor må vi estimere den.

Estimere standardavviket σ til feilleddene

Feilleddene e_i er normalfordelte med forventingsverdi 0 og standardavvik σ. Vi skal bruke følgende estimator for σ:

$$s = \sqrt{\frac{1}{n-2}\text{SSE}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

Hvordan kan vi tenke på dette? Det har noe med hvor mye responsen avviker fra den estimerte regresjonslinjen, og standardavviket er jo definert som kvadratroten av gjennomsnittlig kvadratavvik fra forventningsverdien. Her er forventningsverdien regresjonslinja.

Hvorfor deler vi på $n-2$ og ikke på n hvis vi snakker om gjennomsnittlig kvadratavvik? Jo, det er litt som forklaringen på hvorfor vi delte på $n-1$ da vi regnet ut standardavviket i et utvalg (ikke regresjon). Da delte

vi på $n-1$ for at estimatoren skulle bli *forventningrett* og sa at vi hadde mistet en informasjonsenhet fordi vi hadde estimert forventningsverdi (en parameter). Nå har vi estimert to (skjæringspunkt og stigningstall).

Koble sammen for å få usikkerhet i estimatoren for stigningstallet

Da er det bare å putte inn estimatoren for σ inn i uttrykket for standardavviket til $\hat{\beta}_1$,

$$\widehat{\mathrm{SE}}(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Simulerte eksempel (forts.)

Hvor finner vi dette i utskriften fra Python? Jo, vi er fremdeles i den midtre delen, men nå har vi beveget oss over til "std err", og observere at $\widehat{\mathrm{SE}}(\hat{\beta}_1)$ (x) er \$0.456

Konfidensintervall

Vi har bare fokus på stigningstallet β_1. Det vi skal gjøre nå ligner på det vi har gjort før for en parameter for et normalfordelt utvalg, og så må få med de såkalte frihetsgradene $n-2$ fra estimatoren for standardavviket til feilleddet.

Nedre og øvre grense i et 95% konfidensintervall for stigningstallet β_1 er

$$\hat{\beta}_1 \pm t_{0.025,n-2} \cdot \widehat{\mathrm{SE}}(\hat{\beta}_1)$$

der $t_{0.025,n-2}$ er en kritisk verdi i t-fordelingen med $n-2$ frihetsgrader.

Før vi setter inn tall her er grensene til intervallet stokastisk, og det er 95% sjanse for at intervallet inneholder den sanne parameteren β_1. Så setter vi inn tall for grensene, og da vil i det lange løp 95% av gangene vi samler inn data og lager 95% konfidensintervall - den sanne β_1 dekkes av intervallet.

Når vi har mange observasjoner, det vil si at n er stor, kan vi bytte ut den kritisk verdien i t-fordelingen med tilsvarende kritisk verdi i normalfordeling.

Simulert eksempel (forts.)

Hvor finner vi dette i utskriften fra Python? Jo, vi er fremdeles i den midtre delen, men nå har vi beveget oss over til de to kolonnene som heter [0.025 og 0.975]. Vi husker at for å få et 95% konfidensintervall må vi ha en nedre grense - og den er kalt 0.025 - og en øvre grense - og denne er kalt 0.975. Leser vi av i raden som heter (x) finner vi at

- nedre grense $\hat{\beta}_1 - t_{\alpha/2,n-2} \cdot \mathrm{SE}(\hat{\beta}_1) = -1.5361$ og
- øvre grense $\hat{\beta}_1 + t_{\alpha/2,n-2} \cdot \mathrm{SE}(\hat{\beta}_1) = 0.2996$ og

Observer at tallet 0 ligger inne i intervallet, og det betyr at vi at 0 er et tall vi har stor tiltro til at det sanne stignignstallet kan være.

Konfidensintervall for regresjonslinja

Ved å bruke (estimert) standardavvik for det estimerte skjæringspunktet og stigningstallet - og også kovarians mellom dem - kan vi finne fordelingen til den estimerte regresjonslinja. Det kan vi bruke til å lage konfidensintevall for regresjonslinja.

Egentlig er det jo ett konfidensintervall for hver x-verdi! Og det tegner vi som en kurve for den nedre og øvre grensen i konfidensintervallet - for hver x-verdi.

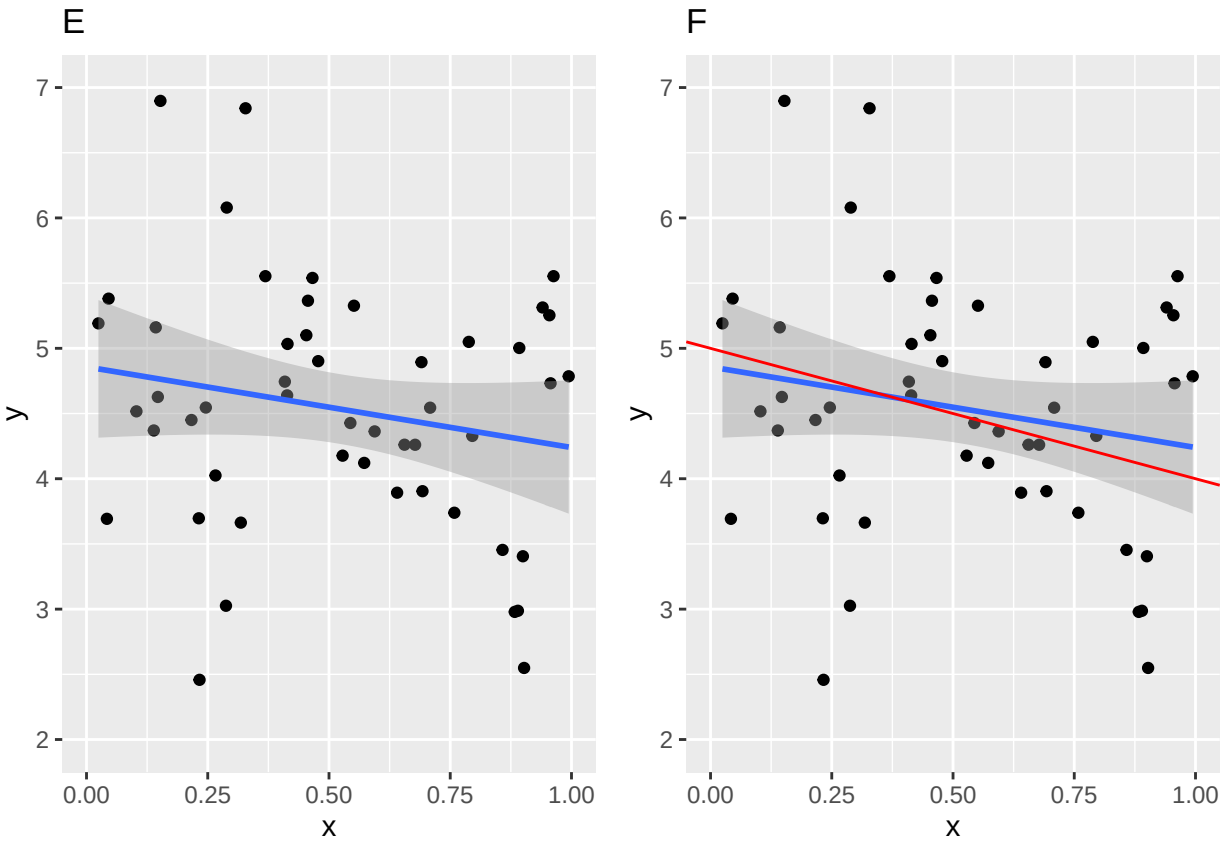
Legg merke til at regresjonslinja er mer usikker i endene enn på midten? Vi er mest sikre på linja i punktet $(\bar{x}, \bar{y})$. Er det rart?

Simulert eksempel (forts.)

Her har vi dataene våre, vår beste linje i blått, og så har vi et grått område mellom øvre og nedre grense i et 95% konfidensintervall for regresjonslinja. Det viser vi i figur E, og i figur F har vi i tillegg tegnet inn den "sanne" regresjonslinja.

Vil alltid den sanne røde linjen ligge inne i det grå området? Nei.

Tenk på følgende algoritme: vi har en sann linje, lager nye data, lager et slikt konfidensintervall for regresjonslinja. Hvis vi gjentar algoritmen 1000 ganger, vil vi i det lange løp i 950 tilfeller ha en rød linje som ligger innenfor det grå området vi lager.



Hypotesetest

Vi skal bare se på hypotesetest for stigningstallet

Hovedmålet med en enkel lineær regresjon er enten å forstå hvordan x påvirker y eller bruke x til å predikere y. Hvis den beste rette linjen gjennom data har stigningstall 0 vil vi ikke ha så stor tiltro til at det er noe vits med denne enkle lineære regresjonen mellom x og y.

Før vi har sett dataene setter vi opp følgende null- og alternativ hypotese.

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0$$

Vi vil altså teste om stigningstallet til den "sanne" linja (ja, den røde som vi egentlig ikke kjenner) er 0 eller ikke. Noen ganger sier vi at "vi vil teste om forklaringsvariabelen er signifikant". Det betyr det samme.

Merk: vi har ikke lyst til å teste hva skjæringspunktet med den vertikale aksen, β_0, er - det synes vi ofte ikke er så spennende.

To typer feil:

Husk at alt dette er basert på at vi er veldig redde for å gjøre noe galt, og det er to typer feil vi er redde for.

- "Forkaste H_0 når H_0 er sann"="falsk positivt funn" = "type-I-feil" ="justismord".

Dette er våre *fake news*, som vi vil unngå.

- "Unnlate å forkaste H_0 når H_1 er sann (og) H_0 er usann)"="falsk negativt funn" = "type-II-feil"= "la en skyldig gå fri".

Vi vil jo ikke la en skyldig gå fri, men i statistikk er vi mer redd for justismord enn å la en skyldig gå fri!

Vi velger da å forkaste H_0 ved et signifikansnivå α hvis p-verdien til testen (se under) er mindre enn det valgte signifikansnivået. Dette signifikansnivået har vi valgt før vi regnet ut p-verdien. Vi sier da at type-I-feilen er *kontrollert* på nivå α, og med det mener vi at sannsynligheten for å begå justismord (type-I-feilen) ikke overstiger α.

Testobservator

I en enkel lineær regresjon så er testobservatoren for å teste $H_0 : \beta_1 = 0$

$$T_0 = \frac{\hat{\beta}_1 - 0}{\widehat{\mathrm{SE}}(\hat{\beta}_1)}$$

og her er $\widehat{\mathrm{SE}}(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$ som vi regnet ut tidligere. Når vi setter inn numeriske verdier får vi t_0.

Neste steg er da å finne kritiske verdier for når vi synes T_0 er så stor i absoluttverdi at vi har nok bevis til å tro at nullhypotesen ikke er sann, eller å regne ut en p-verdi. Vi vil satse på en p-verdi fordi det er det som rapporteres i utskrift fra statistisk analyse med programvare (som Python).

p-verdi

Hvis vi har regnet ut en p-verdi så vil

- en liten p-verdi "gi bevis" for at H_0 er gal og dermed H_1 er sann
- hvis p-verdien er mindre en det vi har valgt som signifikansnivå α (vår valgte øvre grense for sannsynligheten for å gjøre en type-I-feil), da forkaster i nullhypotesen H_0.

I en enkel lineær regresjon bruker man en fordeling som heter t-fordelingen til å regne ut en p-verdi for vår tosidige test for stigningstallet. Anta at testobservatoren vår T_0 numerisk blir t_0. Siden t-fordelingen er symmetrisk om 0 regner vi da:

$$p\text{-verdi} = P(T_0 > \mid t_0 \mid) + P(T_0 < - \mid t_0 \mid) = 2 \cdot P(T_0 > \mid t_0 \mid).$$

Nei, vi trenger ikke regne det ut - det rapporteres i utskriften når vi tilpasser en enkel lineær regresjon UTEN at vi spør etter det.

Vi forkaster H_0 hvis p-verdien er *mindre* en det vi har valgt som signifikansnivå. Det er ekstremt populært å velge signifikansnivå $\alpha = 0.05$.

Jeg har laget en lang video (rundt 45 min) der jeg går i detalj for alt rundt inferens om stigningstallet i en enkel lineær regresjon, og liker du å ha kontroll på teorien så kan du se denne videoen. Du har ikke behov for det i dette emnet.

Simulert eksempel (forts.)

Vi er fremdeles i den midtre delen av summary-utskriften og ser på kolonnene "t" og "P>|t|", og vi er på raden som starter med x.

- "t" er den numeriske verdien til testobservatoren T_0 for a teste $H_0 : \beta_1 = 0$ mot $H_1 : \beta_1 \neq 0$, og den har verdi -1.3544
- "P>|t|" er p-verdien til testen, og den har verdi 0.182.

Her ser vi at p-verdien er større en det populære signifikansnivået 0.05, som betyr at vi ikke har bevis for at nullhypotesen er gal, og vi kan ikke forkaste H_0.

Nå vet vi jo at den sanne røde linjen har signingstall -1, men det klarer vi ikke å finne ut - fordi usikkerheten vi har i estimeringen av stigningstallet er for stort. Dette er på grunn av at variansen til feilleddet er stor og det bestemmer hvor mye observasjonene er spredt rundt den sanne linja - og her blir det mye i forhold til stigningstallet.

Siden det er et simulert datasett vet vi jo at nullhypotesen er falsk - og at stigningstallet er 1, men vi klarer ikke oppdage det med vår kombinasjon av bare $n = 50$ observasjoner og variansen til feilleddene som var 1.

(Vi skal etterpå se hva som hadde skjedd med p-verdien hvis antall observasjoner var mye større.)

Sjekk av modellantagelser

Vi har skrevet opp mange ligninger med greske bokstaver, men vi kan oppsummere det vi har antatt i fire punkter:

1. Det er en lineær sammenheng mellom forklaringsvariabelen og responsen.
2. Feilleddene har konstant varians - for alle verdier av forklaringsvariabelen.
3. Feilleddene er normalfordelte
4. Observasjonsparene (og da også feilleddene) er uavhengige.

For å sjekke disse antagelsene er det to plott som gjelder:

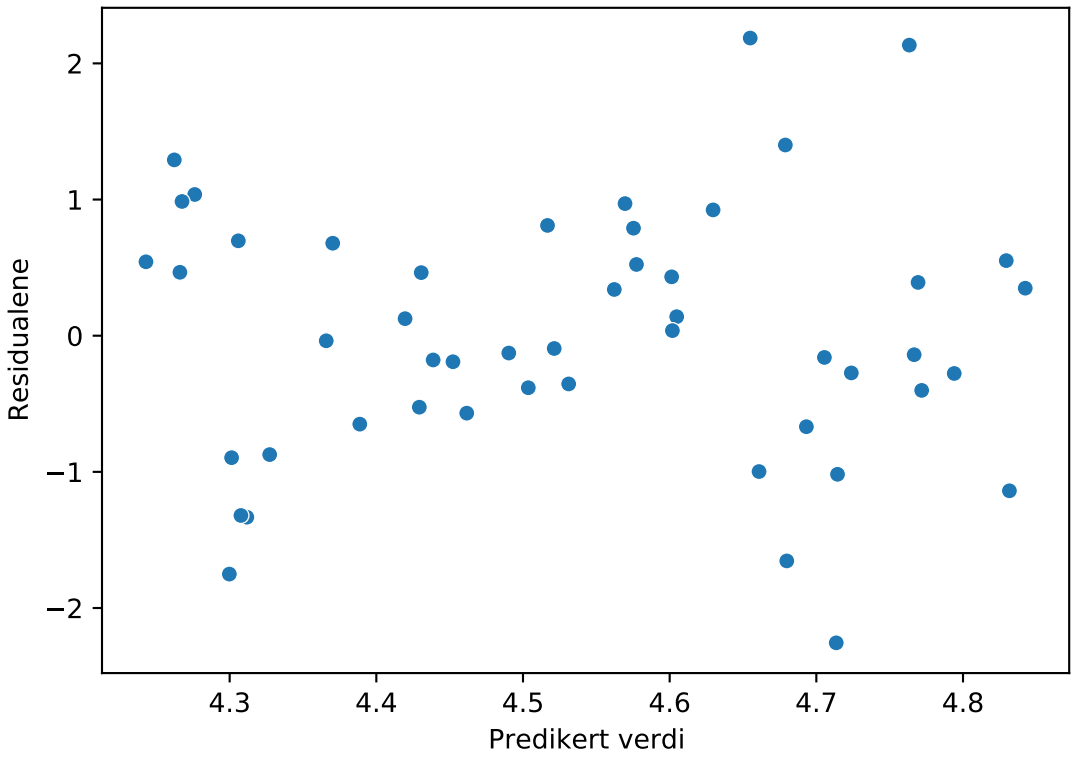
- Plott residualer $\hat{e}_i$ mot predikerte verdier $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Her sjekker vi punkt 1 og 2.
  - Punkt 1: Hvis det er en trend i residualene betyr det at regresjonsmodellen ikke har fått med seg alt av informasjon som ligger i forklaringsvariabelen x. Kanskje sammenhengen heller er kvadratisk eller av en annen type?
  - Punkt 2: Hvis bredden på området for residualer ikke er konstant, kan det være at variansen til feilleddene ikke er konstant. Da kan man bruke noe som heter variansstabiliserende tranformasjon, for eksempel ta logaritmen av responsen, men da må man passe på at den lineære sammenhengen ikke forsvinner - og kanskje forklaringsvariabelen også må transformeres.
- Lag et QQ-plott av residualene der vi sammenligner kvantiler for den empiriske fordelingen til residualene med kvantiler i normalfordelingen
  - Dette tester punkt 3. QQ-plottet er laget slik at hvis residualene kan sies å være normalfordelte vil de ligge på en rett linje. Avvik fra linja er oftest i halene av fordelingen til residualene og det vil vi se nede til høyre og oppe til venstre i plottet. Det er mulig å si akkurat hva som er problemet med fordelingene til residualene ved å studere plottet, men det løser ikke problemet. Det kreves erfaring med å se på slike plott for å virkelig se når vi må si at residualene virkelig ikke er normalfordelte.

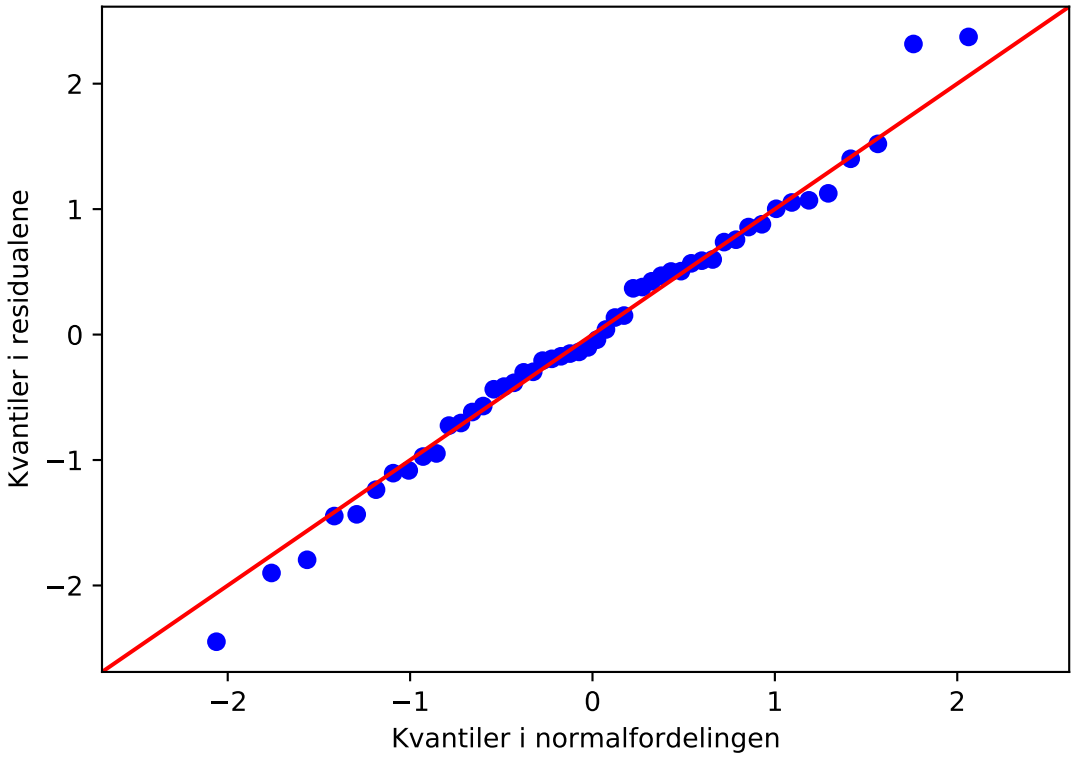
Punkt 4 er ofte vanskelig å sjekke, og det er i hovedsak når data er samlet over tid - at det finnes plotteteknikker. Da plotter man residualene mot observasjonsnummer, og ser om det er en trend. Hvis det er en trend så vil observasjonsrekkefølgen ha noe å si, og da må man se mer på hva akkurat observasjonsrekkefølgen er. Kanskje dette er måling av tid på en oppgave og man blir bedre til å gjøre oppgaven over tid?

I tillegg til at man kan sjekke modellantagelsene ved å plotte residualene så vil man også oppdage mulige utenforliggene observasjoner, eller *uteliggere (outliers)*, som kan indikere at noe har gått galt i innsamlingen - og det er alltid lurt å sjekke uteliggere. Uteliggere vil påvirke veldig hvilken rett linje som passer best til dataene når det er kvadratisk avvik fra linja vi modellerer. Det er ikke lov å fjerne en uteligger-observasjon fra et datasett uten at det er vist at det er noe galt med observasjonen.

Simulert eksempel (forts.)

Vi har jo simulert dataene med normalfordelte feilledd med konstant varians, og vi har en lineær sammenheng mellom forklaringsvariablen og responsen. Figurene under viser derfor hvordan slike plott skal se ut når modellantagelsene ER oppfylt.





Tester i utskrifter fra statistiske programmer

For å teste punkt 3 finnes det en rekke hypotesetetester for å sjekke om residualene er normalfordelte, og mange av disse skrives automatisk ut i Python når vi bruker statmodels.api og skriver summary. I vår utskrift er dette spesielt Omnibus og Jarque-Bera. Begge testene går ut på å teste

- H_0 : residualene er normalfordelte, mot
- H_1 : residualene er ikke normalfordelte,

men gjør det ved å legge vekt på ulike avvik fra normalfordelingen (da vi så på QQ-plottet så kunne vi for eksempel ha fokus på halene eller på midten av plottet).

I utskriften kan man lese ut verdien til testobservatoren som "Omibus" og "Jarque-Bera (JB), men det er heller"Prob(Omibus)" og "Prob(JB)" som er p-verdier fra denne testen - som vi ser på. Vi vil gjerne at p-verdien skal være større enn signifikansnivået som vi ønsker å bruke, fordi da vil vi ikke forkaste nullhypotesen at residualene er normalfordelte.

Vi er generelt litt tilbakeholdne med å legge mye vekt på tester for normalitet, fordi at hvis vi har få data vil en slik test aldri kunne finne at data ikke er normalfordelte og hvis vi har veldig mye data vil må avvik fra normalfordelingen (som er helt innafor for våre analyser) fremstå som veldig signifikante. I forskningen i dag forsøker man generelt å tone ned bruken av hypoteseteseter, og heller se på effekten av eventuelle avvik med plott.

Simulert eksempel (forts.)

Dette er bare for den interesserte leser.

I nedre del av summary-utskriften kan vi lese at p-verdien for Omnibustesten er 0.874 og for Jarque-Bera er 0.995, slik at vi ikke forkaster nullhypotesen at residualene er normalfordelte.

I utskriften er det også skrevet ut hva "Skew"=skjevheten til residualene er. Det er et mål på hvor usymmetrisk fordelingen til residualene er, og for normalfordelingen som er en symmetrisk fordeling så er dette målet 0. Vi vil derfor at dette tallet er nært 0. "Kurtosis" sier om hvor tunge halene i fordelingen til residulene er, og i en normalfordeling er dette tallet 3.

Til slutt står det "Durbin-Watson" og det er et mål på om det er korrelasjon i residualene mot observasjonsrekkefølgen. Er tallet rundt 2 er alt ok.

Hva gjør vi om noen modellantagelser ikke stemmer?

Hvis vi ser problemer i modellantagelsene kan vi gjøre endringer i modellen

- er det en transformasjon av forklaringsvariablen eller responsen som gjør sammenhengen lineær og dermed fjerner trend i residualene?
- kan en slik transformasjon også fjerne trend i variansen til residualene?

Det kan også være at vi må forlate den lineære regresjonsmodellen og heller bruke mer avanserte modeller som tar hensyn til ikke-lineæritet, korrelasjon mellom observasjoner og annet.

Hvor god er regresjonsmodellen?

Nå har vi sjekket at modellantagelsene er oppfylt, og vil nå se hvor stor andel av variasjonen i responsen som vi forklarer med regresjonsmodellen.

Hvis vi bare hadde sett på responsen - og vi visste ikke noe om forklaringsvariablen - så vet vi at en god estimator for variansen ville vært en skalert versjon (vi ville ha delt på $n-1$) av følgende kvadratsum:

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

SST står for "sums of squares total" - totalt sett - uten at forklaringsvariabler er med i bildet, og vi sier at det gir et anslag på mengden variasjon i dataene. Variasjonen er da i forhold til gjennomsnittet.

Vi vet fra før at avviket mellom hva vi har observert og hva regresjonsmodellen sier er gitt som summen av kvadratet av residualene, som vi kalte SSE:

$$\text{SSE} = \hat{e}_1^2 + \hat{e}_2^2 + ... + \hat{e}_n^2 = \sum_{i=1}^{n}\hat{e}_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2$$

Oppsummert:

- vi vet at total variabilitet i data kan presenteres som SST
- vi vet at vi i regesjonsmodellen ikke har forklart variabliliteten presentert som SSE

Derfor vet vi at regresjonsmodellen *har* forklart SST−SSE, så det er *forklart variabilitet.*

Vi har laget et tall som kalles *bestemmelseskoeffisienten* og noteres som R^2, som gir *andel variabilitet* forklart av regresjonsmodellen:

$$R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}.$$

Dette tallet er mellom 0 og 1, og uavhengig av skalaen på dataene. En god verdi av R^2 er nær 1, men det er stor forskjell fra situasjon til situasjon på hva det er mulig å oppnå for R^2. Noen ganger er man ikke fornøyd med mindre R^2 er godt over 0.9, mens andre ganger er en R^2 rundt 0.2 det beste man kan oppnå.

For en enkel lineær regresjonsmodell er R^2 lik kvadratet av den empiriske korrelasjonskoeffisienten mellom forklaringsvariabelen og responsen.

Simulert eksempel (forts.)

Nå er vi over til øvre panel, og der finner vi øverst til høyre "R-squared" som er 0.037 - et veldig lavt tall. Det ser i kanskje også av plottene av observasjonene og den sanne regresjonslinja, det er veldig lite av variabiliteten i dataen vi klarer å forklare.

Prediksjon

Selv om vi ofte har som mål med regresjonen å *forstå* sammenhengen som er mellom en forklaringsvariabel og en respons, er det flere situasjoner der vi vil bruke regresjonsmodellen til å predikere hvilken verdi vi kan få for responsen for en eller flere gitte verdier i for forklaringsvariabelen. På den måten kan den estimerte modellen brukes som en *prediksjonsmodell*. Da tenker vi litt på modellen som en sort boks.

Prediksjon for en observasjon x_0

Gitt at den nye observasjonen x_0 ligger innenfor området der vi har tilpasset modellen vår, bruker vi bare den estimerte regresjonslinjen til å predikere verdi for responsen for den nye observasjonen:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Prediksjonsintervall for en observasjon x_0

Vi har sett på usikkerheten i regresjonslinjen i et punkt x, og for en ny observasjon x_0 får vi i tillegg til usikkerheten i regresjonslinjen også usikkerheten i å observere en ny observasjon.

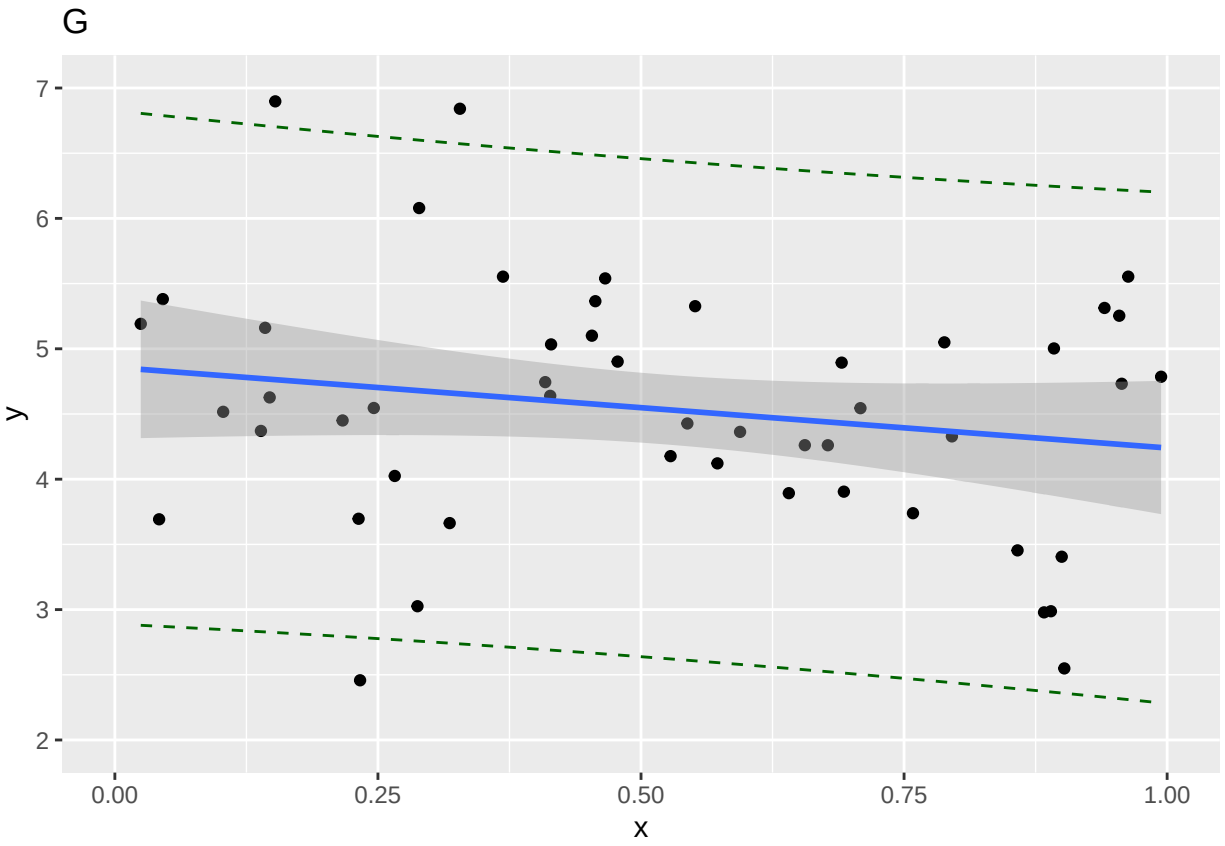
For å regne på formelen for et intervall for en ny observasjon (det heter et prediksjonsintervall) inngår

- estimert variabilitet i stigningstall og skjæringspunktet,
- korrelasjonen mellom disse to estimatene,
- hvor observasjonen ligger i forhold til observasjonene vi har brukt til å estimere regresjonslinjen
- antall observasjoner for å estimere regresjonslinja
- og nivået vi vil sette på intervallet (95% er vanlig).

I praksis vil man ikke regne en slik formel for hånd, men heller bruke statistisk programverktøy (som Python).

Simulert eksempel (forts.)

Her har vi tegnet inn et 95% prediksjonintervall sammen med beste regresjonlinje og et 95% konfidensintervall for regresjonslinja.



Kvadratavvik for prediksjon

Noen ganger vil man ha lyst til å sammenligne to metoder (eller sammenligne to regresjonsmodeller med samme respons men med to ulike forklaringsvariabler), og man har et sett med observasjoner man ikke har brukt til å lage regresjonsmodellene. Da kalles datasettet man har brukt til å lage regresjonsmodellene for *treningssettet* og så kalles datasettet med nye observasjoner for *testsettet*. Her tenker man da at man både kjenner til verdi for forklaringsvariabel og respons i testsettet.

Det er da vanlig å regne ut et såkalt *gjennomsnittlig kvadrert prediksjonsfeil*. Det vil gi et bedre bilde av fremtidig godhet av modellen enn om vi hadde brukt de samme observasjonene som vi laget modellen fra. Anta at vi har n_T observasjoner i testsettet, da regner vi ut

$$\frac{1}{n_T}\sum_{t=1}^{n_T}(y_{0l} - \hat{y}_{0l}(x_{0L}))^2$$

der y_{0l} er målt respons til observajon l i testsettet og $\hat{y}_{0l}$ er predikert respons for observasjon i testsettet som har forklaringsvariable x_{0l}.

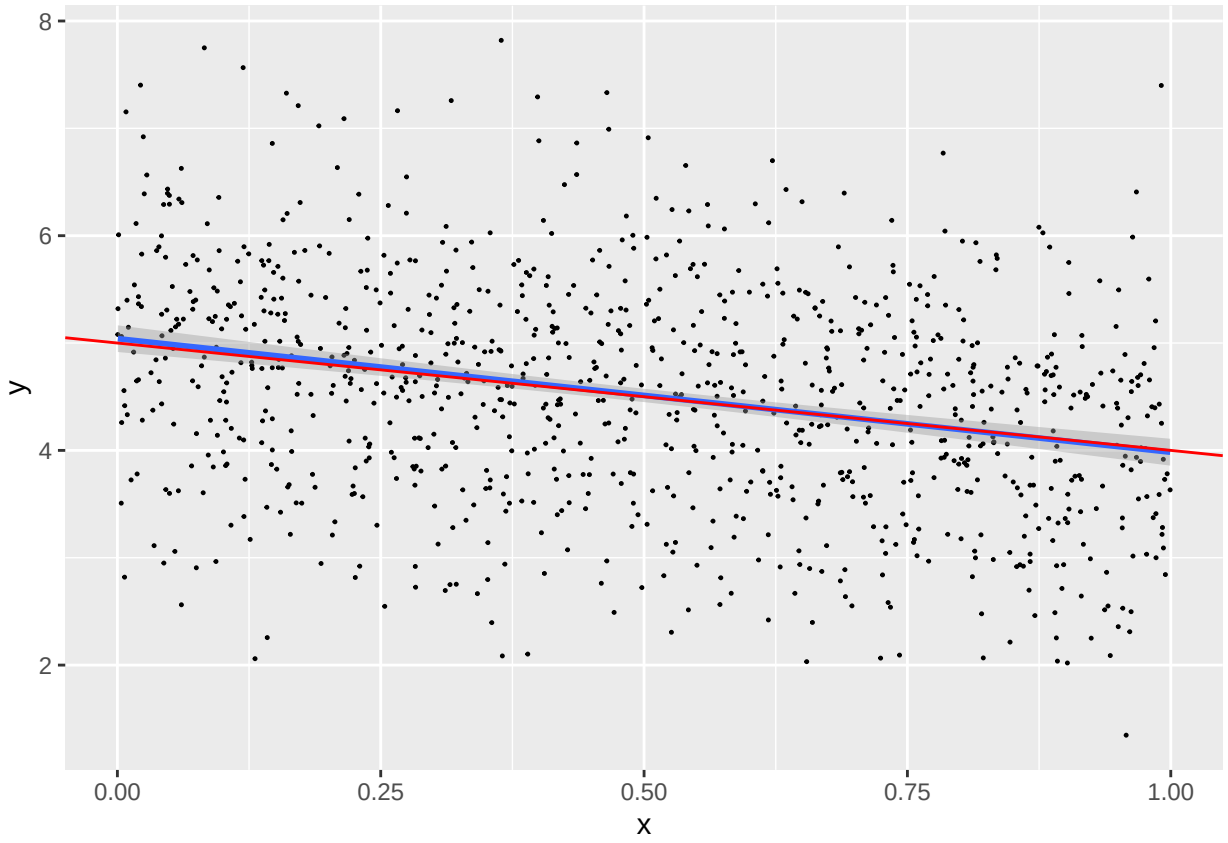
Effekt av antall observasjoner?

I vårt simulerte eksempel forklarte vi veldig lite av variasjonen i dataene, og vi var veldig usikre på om stigningstallet var forskjellig fra 0. Hva endrer seg hvis vi bruker samme modell, men nå genererer $n = 1000$ datapunkter?

Simulert eksempel med stor utvalgsstørrelse

Med mer data er vi veldig sikre på at stigningstallet ikke er 0, men andelen forklart variasjon er bare 0.08 - det ser vi også av figuren at det er lite vi har klart å forklare. I figuren vises den sanne linja i rød, den

estimert i blått og konfidensintervall for regresjonslinja som et grått område.



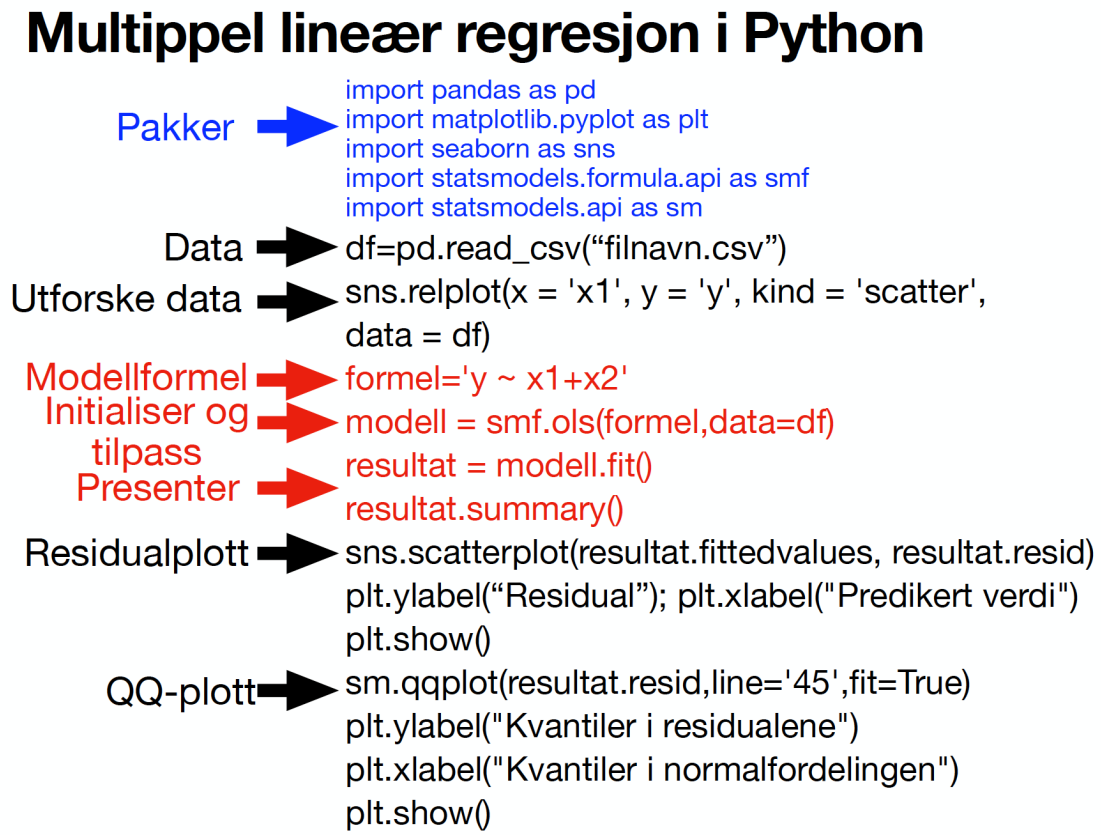

```
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                          y   R-squared:                       0.085
## Model:                                OLS   Adj. R-squared:                  0.084
## Method:                     Least Squares   F-statistic:                     92.13
## Date:                    Mon, 19 Oct 2020   Prob (F-statistic):           6.28e-21
## Time:                            20:37:03   Log-Likelihood:                -1419.8
## No. Observations:                    1000   AIC:                             2844.
## Df Residuals:                         998   BIC:                             2853.
## Df Model:                               1
## Covariance Type:                nonrobust
## ==============================================================================
##                  coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## Intercept      5.0409      0.063     79.604      0.000       4.917       5.165
## x             -1.0583      0.110     -9.599      0.000      -1.275      -0.842
## ==============================================================================
## Omnibus:                        0.291   Durbin-Watson:                   1.984
## Prob(Omnibus):                  0.865   Jarque-Bera (JB):                0.306
## Skew:                           0.041   Prob(JB):                        0.858
## Kurtosis:                       2.978   Cond. No.                         4.40
## ==============================================================================
##
## Notes:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Gjennomføre en regresjonsanalyse i fem steg i Python

Vi har lagt opp til å bruke statmodels.api og statmodels.formula.api for lineær regresjon i dette emnet, og analysene har vi lagt opp i fem steg.

- Steg 1: Bli kjent med dataaene ved å se på oppsummeringsmål og ulike typer plott
- Steg 2: Spesifiser en matematisk modell
- Steg 3: Initialiser og tilpass modellen
- Steg 4: Presenter resultater fra den tilpassede modellen
- Steg 5: Evaluer om modellen passer til dataene

I figuren vises hvordan steg 2-5 kan utføres.



Multippel lineær regresjon (MLR)

Nå skal vi *utvide* regresjonen vår til å ikke bare inkludere en forklaringsvariabel, men mange! Da er det *mye* som ikke endrer seg - og vi kan bare gjøre på akkurat samme måte som for enkel lineær regresjon. Men, så er det noen ting som er litt endret og noen få ting er helt nye.

Kort oppsummert:

- vi har en ny fortolkning av de estimerte regresjonskoeffientene
- dette påvirker også tolkningen av å teste om regresjonskoeffisienter er signfikante
- vi skal lære hvordan vi skal få med kategoriske forklaringsvariabler i regresjonen
- vi skal se at R^2 ikke kan brukes til å sammenligne to regresjonsmodeler, men at
- vi kan bruke en justert versjon av R^2, kalt justert R^2.

Vi skal bruke leieindeks-eksemplet for å illustrere gamle og nye begreper.

Multippel lineær regresjonsmodell

Det som er kjernen i alt er at vi nå skal lage en regresjonsmodell som ikke bare inneholder *en* forklaringsvariabel, men mange!

- Observasjonsenhetene våre består nå av p forklaringsvariabler $(x_{1i}, x_{2i}, x_{3i}, \ldots, x_{pi})$ og en respons y_i. Og vi antar fremdeles at vi har uavhengighet mellom observasjon i $(x_{1i}, x_{2i}, x_{3i}, \ldots, x_{pi}, y_i)$ og observasjon k for to ulike i og k.
- I tillegg til regresjonsparameterne β_0 og β_1 får vi nå $\beta_2, \beta_3, \ldots, \beta_p$.

Vår multiple lineære regresjonsmodell skriver vi for en gitt verdi av forklaringsvariablene $(x_{1i}, x_{2i}, x_{3i}, \ldots, x_{pi})$ som:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i$$

for n uavhengige observasjonsenheter $i = 1, \ldots, n$.

For feilleddene antar vi fremdeles det samme, nemlig at feilleddene er uavhengige og normalfordelte med forventning 0, og konstant standardavvik σ.

Gitt verdien til forklaringsvariablen har vi da at den betingede fordelingen til responsen er normal,

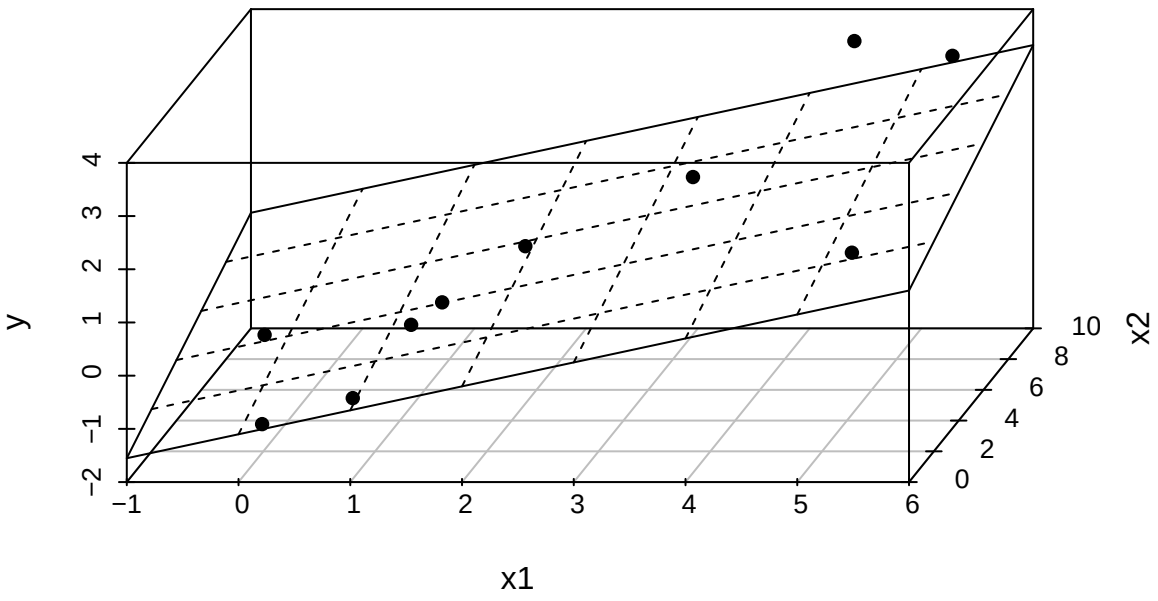
$$Y_i \mid (x_{1i}, x_{2i}, \ldots, x_{pi}) \sim N(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}, \sigma)$$

.

To forklaringsvariabler

Når vi har en enkel lineær regresjon vil vi estimere den beste rette linjen som beskriver dataene våre.

Når vi har to forklaringsvariabler vil vi finne den beste planet. Går vi til tre forklaringsvariabler blir det et hyperplan.



Estimere regresjonskoeffisienter

Vi vil finne estimatorer for våre ukjente regresjonsparametere. For enkel lineær regresjon er vi ute etter en rett linje og velger den linja der summen av kvadratet av residualene er så lite som mulig. Denne summen kaller vi SSE. Vi har sett at vi kan skrive opp formler for $\hat{\beta}_0$ og $\hat{\beta}_1$.

Hvor mye av dette endrer seg når vi går til multippel lineær regresjon?

Heldigvis ikke så mye. Nå har vi jo flere enn en forklaringsvariabel, og vi er ute etter det beste *hyperplanet* som passer til dataene våre. Anta at vi har en kandidat for det beste regresjonshyperplanet, da gjør vi det samme som for enkel lineær regresjon – vi predikerer en responsverdi for hver observasjon og regner ut residualer – og vi finner regresjonskoeffisientene som minimerer summen av de kvadrerte residualene.

For multippel lineær regresjon er det ganske lett å finne uttrykk for estimatorene $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, men vi trenger å bruke vektorer og matriser for at utrykkene skal bli enkle å lese. Vi skal ikke se på disse utrykkene her, bare vite at de finnes.

Tolke estimater av regresjonskoeffisienter

Når det gjelder tolkning av estimerte regresjonskoeffisienter er det en liten forskjell fra enkel lineær regresjon.

La oss si at vi har to forklaringsvariabler x_1 og x_2, med tilhørende estimert regresjonskoeffisienter $\hat{\beta}_1$ og $\hat{\beta}_2$.

Hvis vi vil forklare hva $\hat{\beta}_1$ betyr må vi nå si: hvis vi sammenligner to observasjoner som har samme verdi for x_2, men den ene observasjonen har verdi x_1 og den andre $x_1 + 1$. Da har den andre observasjonen i gjennomsnitt en responsverdi som er $\hat{\beta}_1$ større enn den første (mindre hvis $\hat{\beta}_1$ er negativ). Vi skal se på konkrete eksempler på hvordan dette skal forklares i leieindeks-eksemplet.

Leieindeks med MLR

For å repetere - datasettet består av følgende variabler som beskriver leiligheter i 1999:

- `rent`: leien (Euro)
- `area`: areal (m^2)
- `location`: beliggenhet (1=gjennomsnittlig, 2=god, 3=topp)
- `bath`: kvalitet av badet (0=standard, 1=premium)
- `kitchen`: kvalitet av kjøkkenet (0=standard, 1=premium)
- `cheating`: sentralvarme (0=ingen sentralvarme, 1=med sentralvarme)

Vi bruker `rent` som respons, og så har vi mange mulige forklaringsvariabler.

For dette datasettet er bare en av forklaringsvariablene kontinuerlig, og det er `area`. Alle de andre forklaringsvariablene er kategoriske.

Det er ikke nytt at forklaringsvariablene kan være kontinuerlige eller kategoriske, men vi har til nå bare sett på kontinuerlige forklaringsvariabler for enkel lineær regresjon.

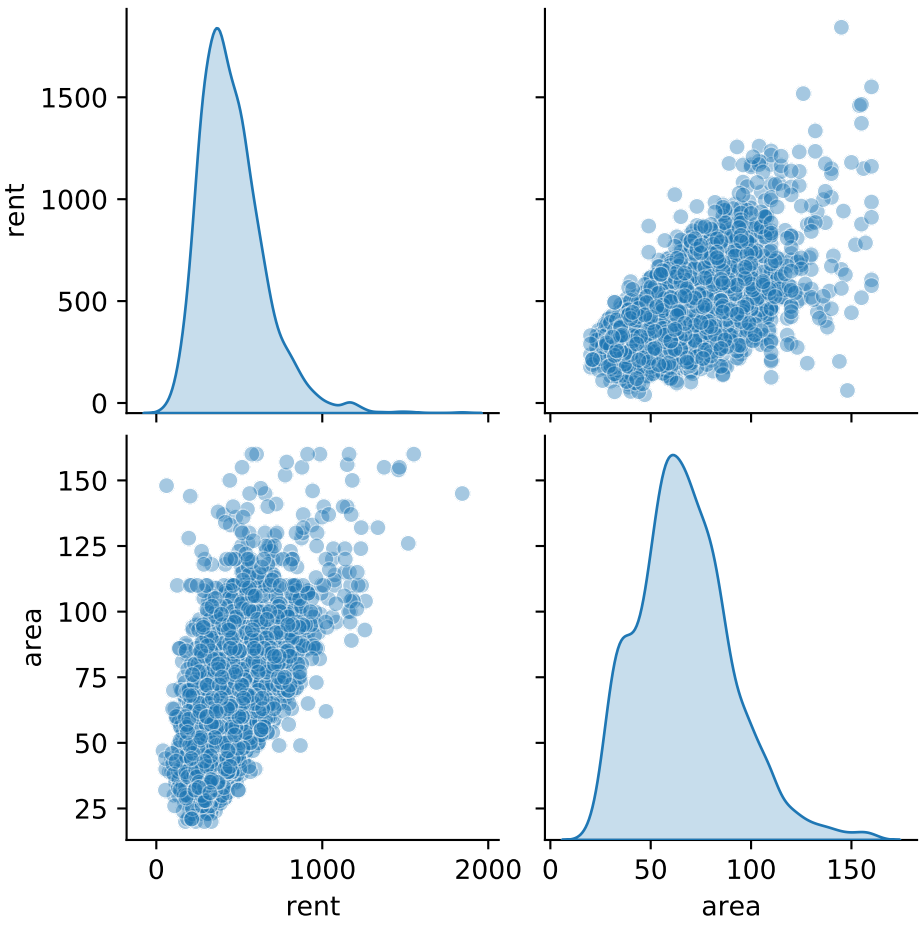
En kategorisk variabel kan ta verdi fra ulike kategorier, og vi har

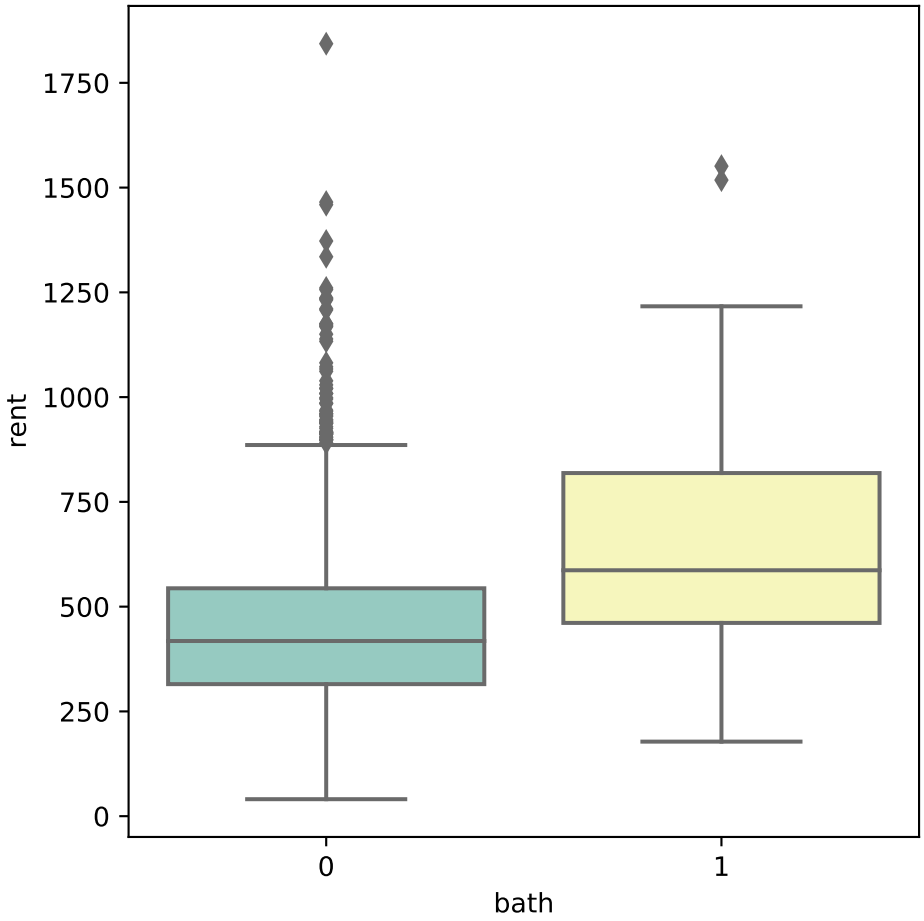
- `bath` og `kitchen` med to kategorier standard og premium
- `cheating` med to kategorier som er uten og med sentralvarme
- `location`: har tre katetorier som er gjennomsnittlig, god og topp

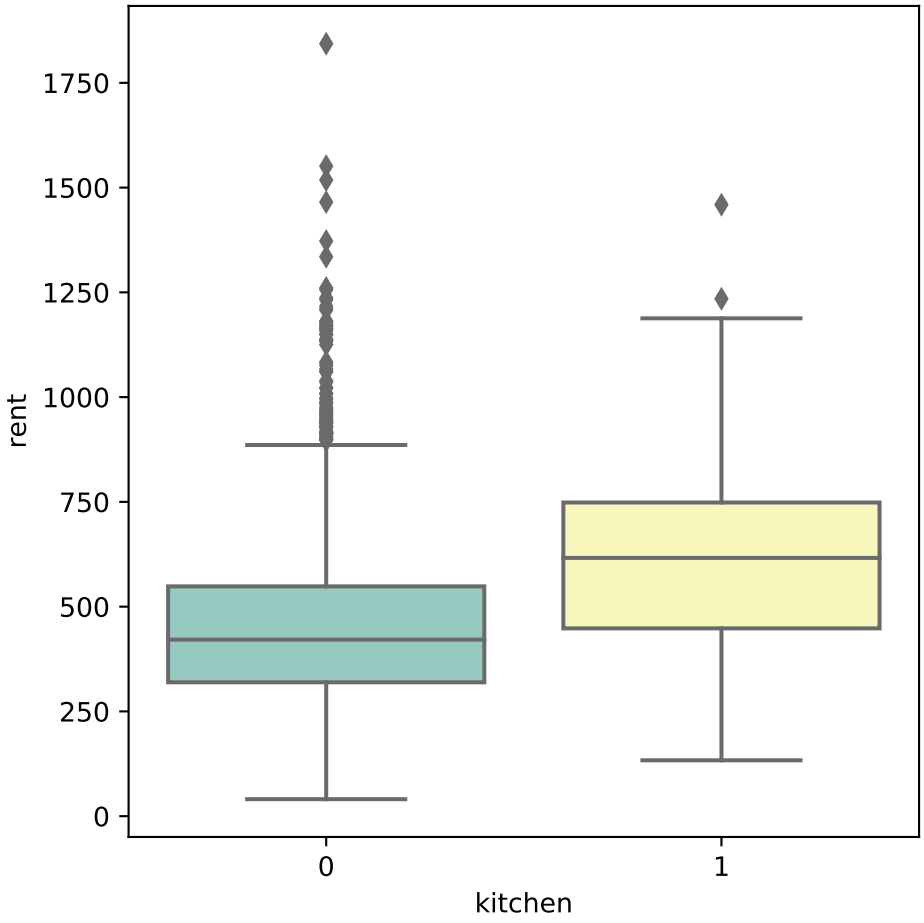
Under er kryssplott, boksplott og glattede histogrammer (tetthetsplott) av variablene og responsen.

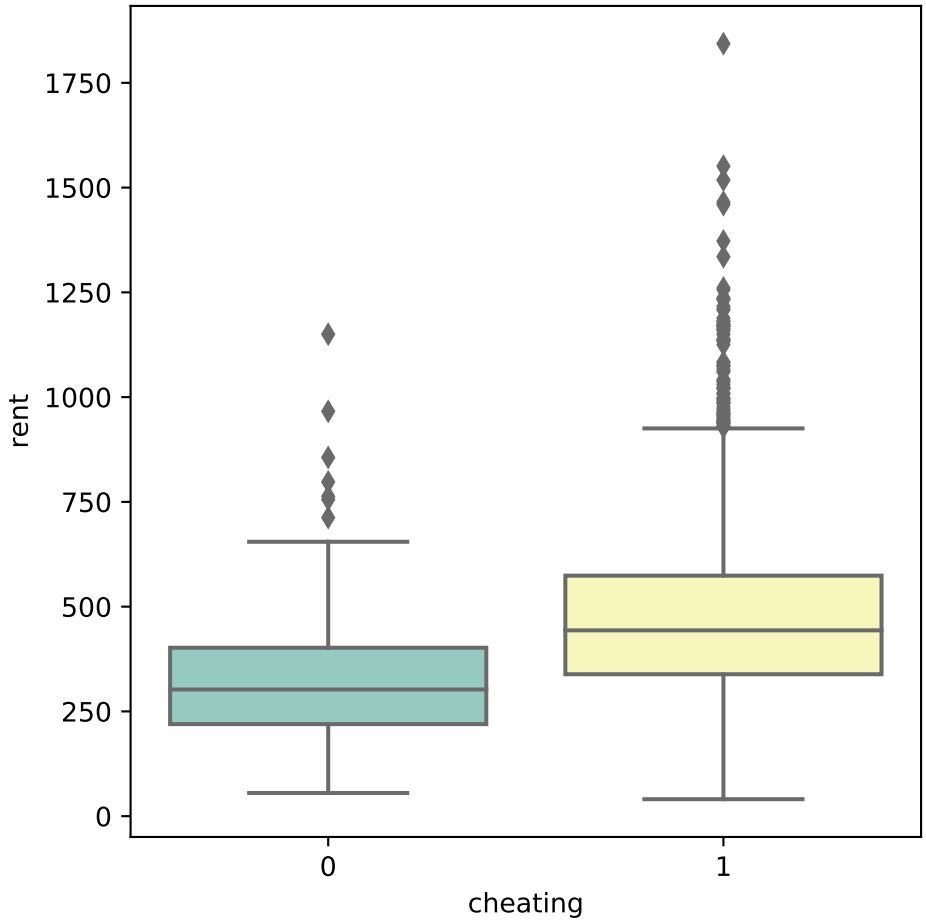
```
## Index(['rent', 'rentsqm', 'area', 'yearc', 'location', 'bath', 'kitchen',
##        'cheating', 'district'],
##       dtype='object')
```

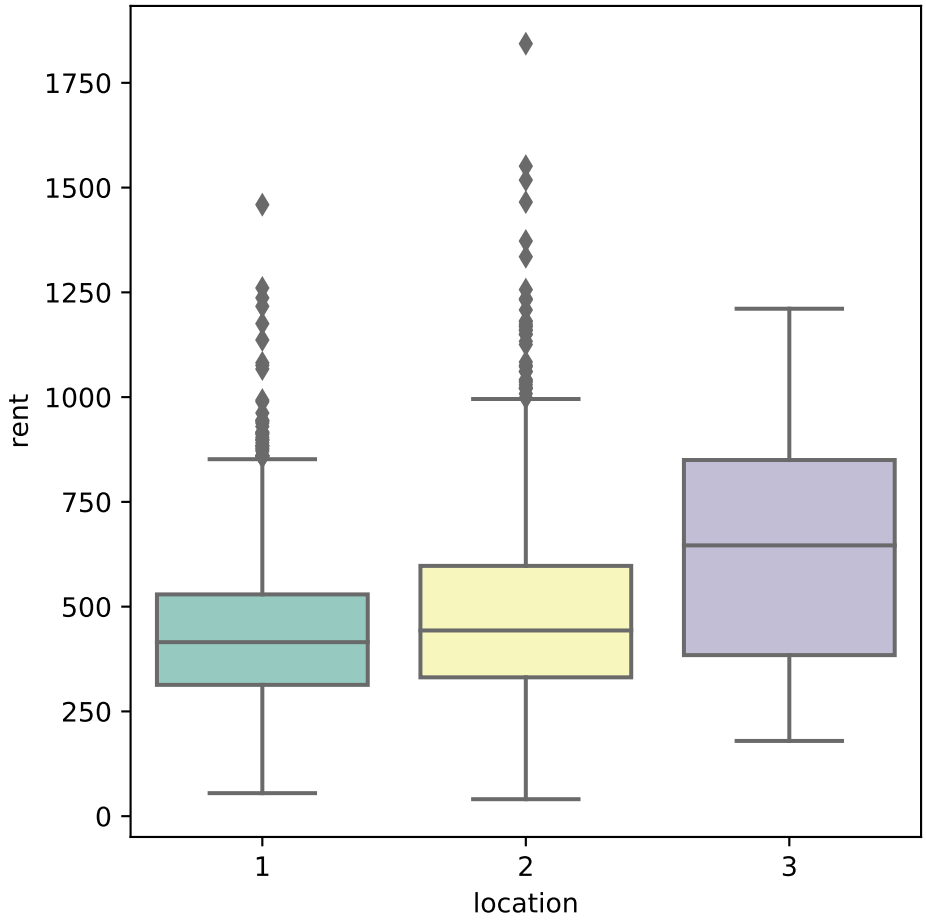
```
## <seaborn.axisgrid.PairGrid object at 0x1175eca90>
```


rent
area
0
500
1000
1500
25
50
75
100
125
150
0
1000
2000
rent
0
50
100
150
area


1750
1500
1250
1000
750
500
250
0
rent
0
1
bath


1750
1500
1250
1000
750
500
250
0
rent
0
1
kitchen


1750
1500
1250
1000
750
500
250
0
rent
0
1
cheating



Forklaringsvariablene

Kontinuerlige forklaringsvariabler

I enkel lineær regresjon har vi sett på en kontinuerlig forklaringsvariabel x mot en kontinuerlig respons y. Hvis ikke sammenhengen mellom x og y er lineær kan vi se på en transformasjon av x mot en transformasjon av y slik at vi får en lineær regresjonsmodell.

Kategoriske forklaringsvariabler med to nivå

For kategoriske variabler med to nivå, som kvalitet av baderom og kjøkken eller om man har sentralfyring eller ikke, er det vanlig å kode dette med såkalt *dummy-variabel* eller *one-hot* koding. Da velger man en *referansekategori* og koder den til det numeriske tallet 0. Den andre kategorien koder man som 1.

Dette betyr at hvis forklaringsvariabelen med to nivå er den eneste forklaringvariablen, kan man klare seg med enkel lineær regresjon (men er det flere nivå blir det multippel).

Se eksemplet under for hvordan man tolker regresjonskoeffisienten for en forklaringsvariabel med to kategorier.

Leieindeks med bare `bath` som forklaringsvariabel

I datasettet var opprinnelig `bath` kodet som 0 og 1, og da står det i utskriften (fra Python ved bruk av statsmodels.formula.api og statmodels.api) `bath[T.1]` for kategorien som er kodet originalt med 1 (premier bath). Vi tilpasser en enkel lineær regresjonsmodell med dummy-variabelkoding av `bath`.

```
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                   rent   R-squared:                       0.063
## Model:                            OLS   Adj. R-squared:                  0.063
## Method:                 Least Squares   F-statistic:                     207.9
## Date:                Mon, 19 Oct 2020   Prob (F-statistic):           1.17e-45
## Time:                        20:37:06   Log-Likelihood:                -20534.
## No. Observations:                3082   AIC:                          4.107e+04
## Df Residuals:                    3080   BIC:                          4.108e+04
## Df Model:                           1
## Covariance Type:            nonrobust
## ==============================================================================
##                  coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## Intercept    446.7929      3.523    126.838      0.000     439.886     453.700
## bath[T.1]    204.0295     14.150     14.419      0.000     176.285     231.774
## ==============================================================================
## Omnibus:                      638.976   Durbin-Watson:                   1.989
## Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1617.636
## Skew:                           1.122   Prob(JB):                         0.00
## Kurtosis:                       5.749   Cond. No.                         4.16
## ==============================================================================
##
## Notes:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Her leser vi av at estimert verdi - eller `coef` - for Intercept $\hat{\beta}_0$ =446.79. La oss kalle regresjonskoeffisienten for `bath[T.1]` for $\hat{\beta}_1$= 204.03. Da blir `bath[T.1]` vår x, som kan ta verdiene 0 og 1.

- Hvis vi skal predikere verdi for en leilighet uten bad setter vi x=`bath[T.1]=0` og da blir predikert verdi bare det som står som Intercept, fordi det er $\hat{\beta}_0 + 0 \cdot \hat{\beta}_1$=446.8.
- Hvis vi skal predikere verdi for en leilighet med bad setter vi `bath[T.1]=1` og får $\hat{\beta}_0 + 1 \cdot \hat{\beta}_1$=446.8+204 = 650.82.

Kategoriske forklaringsvariabler med tre (eller flere) nivå

For kategoriske variabler med tre nivå, som `location` (beliggenhet), brukes også dummy-variabelkoding, men nå trenger vi to dummy-variabler.

Vi starter med å velge et referansenivå. For `location` i leieindekseksemplet, velger vi "gjennomsnittlig" som referansekategori.

Vi må da "lage" to variabler.

- x_1="god beliggenhet": er 0 hvis beliggenheten er enten gjennomsnittlig eller topp, og så er den 1 hvis beliggenheten er god,
- x_2="topp beliggenhet": er 0 hvis beliggenheten er enten gjennomsnittlig eller god, og så er den 1 hvis beliggenheten er topp.

Dette skjer automatisk hvis vi bruker såkalt *modellformel* (i statmodels.formula.api) og den kategoriske variabelen er registrert som kategorisk.

Leieindeks med bare `location` som forklaringsvariabel

Vi vil at `location` er en kategorisk variabel. I datasettet var opprinnelig `loction` kodet som 1, 2 og 3, og da står det i utskriften `location[T.2]` (x_1 for oss) for kategorien som er kodet opprinnelig med 2

(god beliggenhet), og `location[T.3]` (x_2 for oss) for kategorien som opprinnelig var kodet som 3 (topp beliggenhet).

```
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                   rent   R-squared:                       0.035
## Model:                            OLS   Adj. R-squared:                  0.034
## Method:                 Least Squares   F-statistic:                     55.42
## Date:                Mon, 19 Oct 2020   Prob (F-statistic):           2.26e-24
## Time:                        20:37:06   Log-Likelihood:                -20580.
## No. Observations:                3082   AIC:                          4.117e+04
## Df Residuals:                    3079   BIC:                          4.118e+04
## Df Model:                           2
## Covariance Type:            nonrobust
## ==============================================================================
##                  coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## Intercept      435.8779      4.540     96.011      0.000     426.976     444.779
## location[T.2]   47.1074      7.153      6.585      0.000      33.082      61.133
## location[T.3]  200.1254     22.241      8.998      0.000     156.517     243.734
## ==============================================================================
## Omnibus:                      637.463   Durbin-Watson:                   1.951
## Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1649.790
## Skew:                           1.111   Prob(JB):                         0.00
## Kurtosis:                       5.812   Cond. No.                         7.03
## ==============================================================================
##
## Notes:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Her vi har nå estimert tre regresjonsparametere: $\hat{\beta}_0$= `Intercept`, $\hat{\beta}_1$ er koeffisienten for x_1= `location[T.2]` og $\hat{\beta}_2$ er koeffisienten for x_2= `location[T.3]`. Hvordan setter vi de sammen for å predikere leie for en leilighet på de tre beliggenhetene?

- Hvis vi skal predikere leie for en leilighet med gjennomsnittlig beliggenhet må både x_1=`location[T.2]` og x_2=`location[T.3]` settes til 0. Da blir predikert verdi bare verdien for Intercept, fordi formelen vi skal bruke blir $\hat{\beta}_0 + 0 \cdot \hat{\beta}_1 + 0 \cdot \hat{\beta}_2$=435.88.
- Hvis vi skal predikere leie for en leilighet med god beliggenhet er x_1=`location[T.2]`=1 og x_2=`location[T.3]`=0. Da blir predikert verdi $\hat{\beta}_0 + 1 \cdot \hat{\beta}_1 + 0 \cdot \hat{\beta}_2$=435.88 + 47.11 = 482.99.
- Hvis vi skal predikere leie for en leilighet med topp beliggenhet er x_1=`location[T.2]`=0 og x_2=`location[T.3]`=1. Da blir predikert verdi $\hat{\beta}_0 + 0 \cdot \hat{\beta}_1 + 1 \cdot \hat{\beta}_2$= 435.88 + 200.13 = 636.

Men, hva om vi heller bare hadde kodet `location` med en variabel som tok verdiene 1, 2, 3, hva hadde skjedd da? Jo, da hadde et vært en numerisk (kontinuerlig) kovariat og blitt håndtert som `area` ble, og det ville blitt estimert ett stigningstall. Det betyr at i denne modellen antar man at gjennomsnittlig differanse i leie mellom "gjennomsnittlig" `location` og "god" `location` er like stor som differansen i leie mellom "god" `location` og "topp" `location`. Hvis dette er rimelig kan man godt beholde `location` med en slik koding. Resultatene blir da som under.

```
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                   rent   R-squared:                       0.029
## Model:                            OLS   Adj. R-squared:                  0.029
## Method:                 Least Squares   F-statistic:                     92.14
## Date:                Mon, 19 Oct 2020   Prob (F-statistic):           1.61e-21
```

```
## Time:                        20:37:06   Log-Likelihood:                -20589.
## No. Observations:                3082   AIC:                         4.118e+04
## Df Residuals:                    3080   BIC:                         4.119e+04
## Df Model:                           1
## Covariance Type:            nonrobust
## ==============================================================================
##                  coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## Intercept    371.2006      9.826     37.776      0.000     351.934     390.468
## location      61.1388      6.369      9.599      0.000      48.651      73.627
## ==============================================================================
## Omnibus:                      631.779   Durbin-Watson:                   1.953
## Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1581.412
## Skew:                           1.115   Prob(JB):                         0.00
## Kurtosis:                       5.710   Cond. No.                         6.03
## ==============================================================================
##
## Notes:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Vi har estimert regresjonsparametere: $\hat{\beta}_0$= `Intercept`, $\hat{\beta}_1$ er koeffisienten for x= `location` kodet som en numerisk verdi.

- Hvis vi skal predikere leie for en leilighet med gjennomsnittlig beliggenhet settes x_1=`location` til 1. Da blir predikert verdi $\hat{\beta}_0 + 1 \cdot \hat{\beta}_1$=432.34.
- Hvis vi skal predikere leie for en leilighet med god beliggenhet er x=`location`=2 og da blir predikert verdi $\hat{\beta}_0 + 2 \cdot \hat{\beta}_1 = 371.2 + 261.14 = 493.48$.
- Hvis vi skal predikere leie for en leilighet med god beliggenhet er x= `location` =3 og da blir predikert verdi $\hat{\beta}_0 + 3 \cdot \hat{\beta}_1 =$ \$ 371.2 +3 61.14 = 554.62 \$.

Hvor forskjellig ble det å bruke `location` som kontinuerlig eller som kategorisk variabel? Det er ikke store forskjeller på predikert verdi for gjennomsnittlig eller god beliggenhet, men for topp beliggenhet gir modellen kategoriske beliggenhetsvariabel en mye høyere predikert leie. Det er ganske strengt å anta det (dvs. at gjennomsnittlig differanse i leie mellom "gjennomsnittlig" `location` og "god" `location` er like stor som differansen i leie mellom "god" `location` og "topp" `location`). Generelt er det lurt å undersøke om en dummy-variablekoding er mer hensiktsmessig enn en numerisk koding.

Prediksjoner

Når vi skal gi en prediksjon - eller beste gjett - på hva responsen for en ny observasjon vil være, gjør vi det akkurat på samme måte som for enkel lineær regresjon.

Hvis vi har en ny observasjon med forklaringsvariabler $(x_{10}, x_{20}, \ldots, x_{p0})$ blir beste anslag (prediksjon) for responsen $\hat{y}_0$:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{10} + \hat{\beta}_2 x_{20} + \cdots + \hat{\beta}_p x_{p0}$$

Vi skal se mer på dette i neste eksempel.

Leieindeks: kvalitet av bad, beliggenhet og areal

```
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                   rent   R-squared:                       0.370
## Model:                            OLS   Adj. R-squared:                  0.369
```

```
## Method:                 Least Squares   F-statistic:                     451.3
## Date:                Mon, 19 Oct 2020   Prob (F-statistic):          2.00e-306
## Time:                        20:37:06   Log-Likelihood:                -19923.
## No. Observations:                3082   AIC:                             3.986e+04
## Df Residuals:                    3077   BIC:                             3.989e+04
## Df Model:                           4
## Covariance Type:            nonrobust
## ==============================================================================
##                  coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## Intercept      137.4876      8.673     15.852      0.000     120.481     154.494
## bath[T.1]       99.3996     11.931      8.331      0.000      76.006     122.793
## location[T.2]   28.0179      5.802      4.829      0.000      16.641      39.394
## location[T.3]  128.6651     18.064      7.123      0.000      93.246     164.084
## area             4.4755      0.122     36.663      0.000       4.236       4.715
## ==============================================================================
## Omnibus:                      169.008   Durbin-Watson:                   2.014
## Prob(Omnibus):                  0.000   Jarque-Bera (JB):              380.129
## Skew:                           0.346   Prob(JB):                     2.86e-83
## Kurtosis:                       4.575   Cond. No.                         461.
## ==============================================================================
##
## Notes:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

I utskriften over har vi tilpasset en multippel lineær regresjon med `rent` som respons og `bath`, `location` og `area` som forklaringsvarabler.

Modellen er da

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{i4} + e_i$$

Estimatene av regrsjonsparameterne i utskriften er da:

- $\hat{\beta}_0$ ('Intercept') skjæringspunktet (også kalt *konstantledd*)
- $\hat{\beta}_1$ ('bath[T.1]') er gjennomsnittlig forskjell i leie mellom standard og premium bad
- $\hat{\beta}_2$ ('location[T.2]') er gjennomsnittlig forskjell i leie mellomgod og gjennomsnittlig beliggenhet
- $\hat{\beta}_3$ ('location[T.3]') er gjennomsnittlig forskjell i leie mellom topp og gjennomsnittlig beliggenhet
- $\hat{\beta}_4$ ('area') er gjennomsnittseffekten av areal på leien.

Husk at alle disse må tolkes gitt at vi sammenligner to leiligheter som ellers er like. Helt konkret - to eksempler:

a) Hvordan forklare hva "location[T.3] 128.6651" betyr?

Hvis vi sammenligner to leiligheter som har samme kvalitet på badet og er like store, så vil en leilighet på med topp beliggenhet i gjennomsnitt ha en leie som er 128.6651 Euro høyere enn en leilighet på gjennomsnittlig beliggenhet.

b) Hvordan forklare hva "area 4.4755" betyr?

Hvis vi sammenligner to leiligheter som har samme kvalitet på badet og samme beliggenhet, så vil en differanse på 1 m^2 for de to leilighetene i gjennomsnitt gi en forskjell i leien på 4.4755 Euro

c) Finn predikert leie for premium bath, god beliggenhet og 60m^2:

$$\hat{y} = 137.5 + 99.4 + 28.0 + 4.5 \cdot 60 = 534.9$$

- konstantleddet (skjæringspunktet med y-aksen) må alltid med - det er 137.5
- premium bad var kodet som 1 for bath[T.1], derfor 99.4

- god beliggenhet var kodet som 1 for location[T.2], derfor 28.0
- area er en kontinuerlig variabel og skal bare ganges med 60: derfor $4.5 \cdot 60$.

Intervaller og hypotesetest

Kort fortalt, estimerer vi standardavvik, regner ut konfidensintervaller og prediksjonsintervaller, og utfører hypotesetest på "samme" måte som for enkel lineær regresjon.

Estimere standardavvik til regresjonskoeffisientene

Vi skal ikke så i detalj her, men kort fortalt finnes det fine formler for å regne ut standardavviket til regresjonskoeffisientene.

Estimere standardavviket σ til feilleddene

Her er det kun en liten endring fra enkel lineær regresjon. Nå har vi ikke estimert to regresjonskoeffisienter ($\hat{\beta}_0$ og $\hat{\beta}_1$) men hele $p+1$ ($\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$), og da blir $n-2$ i nevneren for å estimere σ byttet ut med $n-p-1$.

Dette tallet, $n-p-1$, kalles ofte "frihetsgrader" og kan i utskrifter gis som "DF residual".

Estimatoren for σ er:

$$s = \sqrt{\frac{1}{n-p-1}\text{SSE}} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

Koble sammen for å få usikkerhet i estimatorene for regresjonskoeffisientene

På samme måte som for enkel lineær regresjon vil vi kunne lese av det estimerte standardavviket til regresjonskoeffisientene $\widehat{\text{SE}}(\hat{\beta}_j)$ i utskriften fra statistisk software. I statmodels.api i Python heter disse "sd err".

Hypotesetest for en regresjonskoeffisient

For tokning av hypotesetesten må vi legge til at vi gjør testen gitt at alle andre forklaringsvariabler er med i modellen vår.

P-verdi for testen regnes ut i en t-fordeling med $n-p-1$ frihetsgrader (mot $n-2$ i enkel lineær regresjon).

Leieindeks: kvalitet av bad, beliggenhet og areal (forts.)

I utskriften fra tilpassing av modellen kommer automatisk p-verdier fra testing av regresjonskoeffisentene i kolonnen "P>|t|".

a) For testing av koeffisienten for areal:

H_0 : I en modell der kvalitet av bad og beliggenhet er med, er det ingen effekt av areal: $\beta_4 = 0$

H_1 : I en modell der kvalitet av bad og beliggenhet er med, er det en effekt av areal: $\beta_4 \neq 0$

Denne nullhypotesen forkastes i eksemplet vårt på nivå 0.05 og vi konkluderer med at areal er viktig for å forklare leiepris.

b) For testing av koeffisienten for kvalitet av bad:

H_0 : I en modell der areal og beliggenhet er med, er det ikke forskjell i leie på å ha et standard og et premium bad: $\beta_2 = 0$

H_1 : I en modell der areal og beliggenhet er med, er det forskjell i leie på å ha et standard og et premium bad: $\beta_2 \neq 0$

Også denne nullhypotesen forkastes i eksemplet vårt på nivå 0.05 og vi konkluderer med at bad er viktig for å forklare leiepris.

c) For testing av koeffisientene for beliggenhet må det gjøres i to omganger, en gang der vi sammenligner gjennomsnittlig beliggenhet med god og en der vi sammenligner gjennomsnittlig beliggenhet med topp.

H_0 : I en modell der areal og kvalitet av bad er med, er det ikke forskjell i leie på gjennomsnittlig og god beliggenhet: $\beta_3 = 0$

H_1 : I en modell der areal og kvalitet av bad er med, er det forskjell i leie på gjennomsnittlig og god beliggenhet: $\beta_3 \neq 0$

Også denne nullhypotesen forkastes i eksemplet vårt på nivå 0.05 og vi konkluderer med at beliggenhet er viktig for å forklare leiepris.

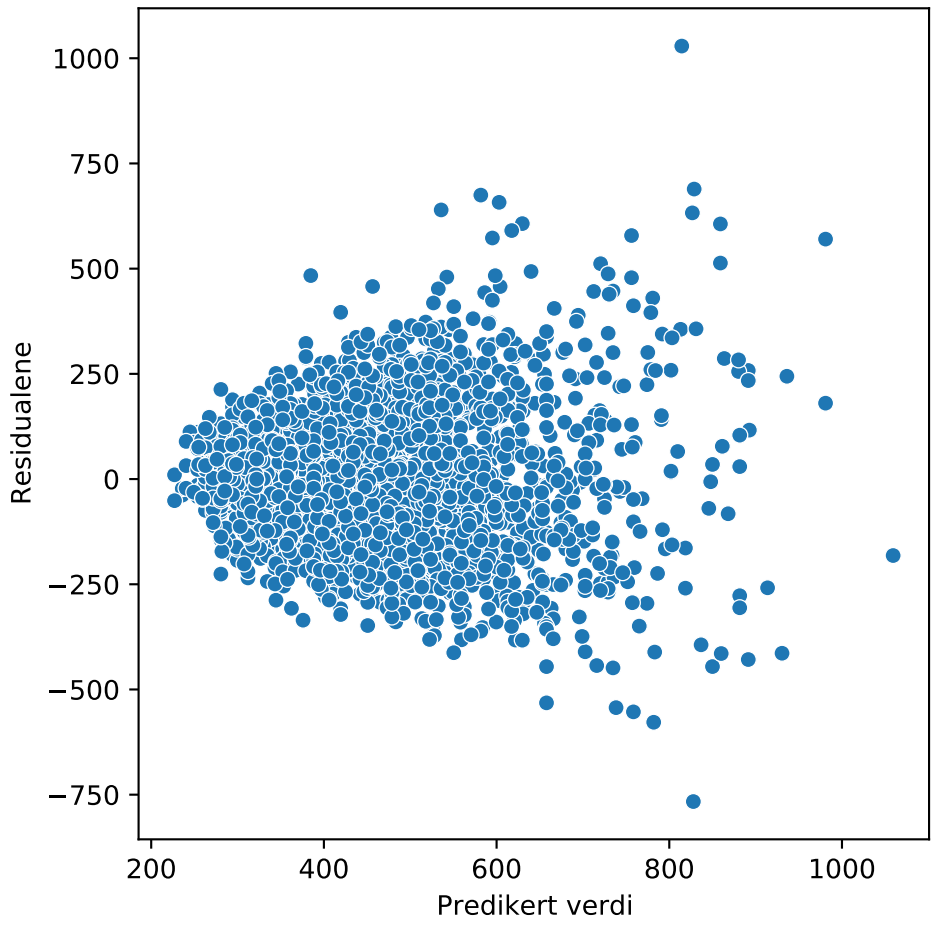
d) Hypotesentesten for å sammenligne gjennomsnittlig og topp beliggenhet gjøres tilsvarende som hypotesetesten for å sammenligne gjennomsnittlig mot god beliggenhet.

Vi skal ikke se på hvordan vi kan sammenligne god og topp beliggenhet. Det finnes også tester for å sjekke om minst en av beliggenhetene gir effekt på leien, men det skal vi ikke gå inn på her.

Sjekk av modellantagelser

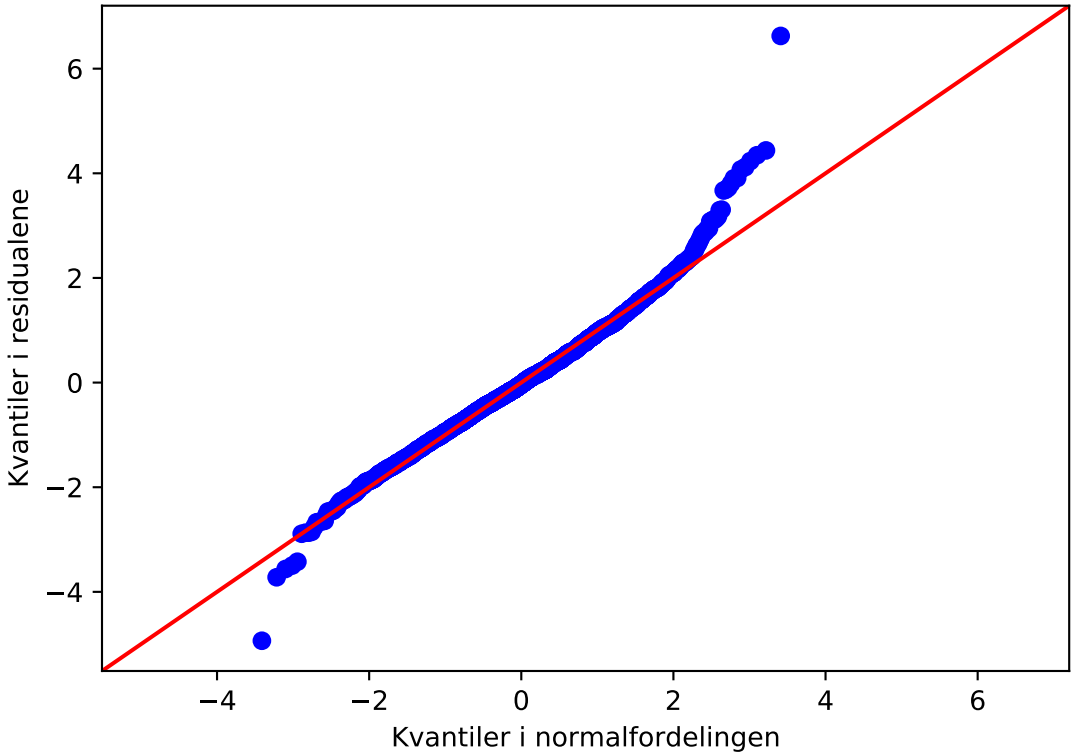
Vi sjekker modellantagelser på samme måte som for enkel lineær regresjon.

Først ser vi på plott av predikerte verdier mot residual. Når modellen passer viser plottet ingen trend og konstant bredde.



Vi har problemer med antagelsen om at variansen er den samme for alle kovariater. Dermed tror vi ikke at feilleddene e_i har konstant varians.

For å sjekke normalitet av residualene lager vi et QQ-plott.



Vi ser klart fra plottet at observasjonene ikke ligger på en rett linje og kan klart konkludere at residualene er ikke normalfordelte. Dermed tror vi ikke at feilleddene er normalfordelte.

Vi bør jobbe med å finne en bedre modell!

Hvor god er regresjonsmodellen

Gitt at vi har en modell der modellantagelsene stemmer, ønsker vi også å fortelle hvor god modellen - er, og da tenker vi på hvor stor andel av variasjonen i dataene vi har forklart med regresjonsmodellen.

Formelen for R^2 er helt lik som for enkel lineær regresjon:

$$R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}.$$

Leieindeks: kvalitet av bad, beliggenhet og areal (forts.)

Andelen forklart variansjon leser vi av i øvre del av utskriften, og finner at "R-squared" er lik 0.370, slik at vi har forklart 37% av variabiliteten i dataene.

Legg også merke til at det finnes noe som heter "Adj. R-squared" rett under "R-squared". Dette tallet skal vi nå ser litt mer på.

Modellvalg

Hva hvis vi har lyst til å sammenligne to regresjonsmodeller. Kanskje vi vil sammenligne modellen med "areal+kvalitet av bad+beliggenhet" med en modell der vi også har tatt med "kvalitet av kjøkken" og om "sentralfyring er tilstede". Kan vi bruke R^2 og velge den modellen som forklarer størst andel av variasjon i datasettet?

Leieindeks - mange forklaringsvariabler

```
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                   rent   R-squared:                       0.450
## Model:                            OLS   Adj. R-squared:                  0.449
## Method:                 Least Squares   F-statistic:                     420.0
## Date:                Mon, 19 Oct 2020   Prob (F-statistic):               0.00
## Time:                        20:37:07   Log-Likelihood:                -19712.
## No. Observations:                3082   AIC:                         3.944e+04
## Df Residuals:                    3075   BIC:                         3.948e+04
## Df Model:                           6
## Covariance Type:            nonrobust
## ==============================================================================
##                     coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## Intercept       -21.9733     11.655     -1.885      0.059     -44.826       0.879
## bath[T.1]        74.0538     11.209      6.607      0.000      52.076      96.031
## kitchen[T.1]    120.4349     13.019      9.251      0.000      94.908     145.962
## cheating[T.1]   161.4138      8.663     18.632      0.000     144.428     178.400
## location[T.2]    39.2602      5.447      7.208      0.000      28.580      49.941
## location[T.3]   126.0575     16.875      7.470      0.000      92.971     159.144
## area              4.5788      0.114     40.055      0.000       4.355       4.803
## ==============================================================================
## Omnibus:                      224.049   Durbin-Watson:                   2.004
## Prob(Omnibus):                  0.000   Jarque-Bera (JB):              580.302
## Skew:                           0.413   Prob(JB):                    9.75e-127
## Kurtosis:                       4.959   Cond. No.                        462.
## ==============================================================================
##
## Notes:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
## Justert-R2 med flere desimaler:
```

```
## 0.44935499762094155
```

Det kan vises at R^2 alltid vil øke eller være uforandret hvis en ny forklaringsvariabel legges til regresjonsmodellen. Dette kan vi forklare med at det alltid vil være små tilfeldige trender i dataene som fanges opp. For å vise dette har skal vi nå legge til en oppkonstruert kovariat som kunne ha vært IQ til utleier. Denne kovariaten har vi trukket fra en normalfordeling med forventningsverdi 100 og standardavvik 16, og så tilfeldig tilordnet tallene til observasjonene som en ny forklaringsvariabel. Under ser du utskrift fra å tilpasse den nye modellen der IQ er med som forklaringsvariabel.

Observer at den nye forklaringsvariablen iq ikke er signifikant (p-verdien er 0.469), men allikevel har R^2 har økt fra 0.4504 til 0.451. Det ser ut som vi kan forklare mer av variabiliteten i dataen når vi har med 'iq' som forklaringsvariabel - selv om vi vet at dette bare er tilfeldige tall.

Vi bruker ikke R^2 til å velge mellom modeller som har ulikt antall forklaringsvariabler, men bruker *justert* R^2.

```
##                             OLS Regression Results
```

```
## ==============================================================================
## Dep. Variable:                   rent   R-squared:                       0.451
## Model:                            OLS   Adj. R-squared:                  0.449
## Method:                 Least Squares   F-statistic:                     360.1
## Date:                Mon, 19 Oct 2020   Prob (F-statistic):               0.00
## Time:                        20:37:07   Log-Likelihood:                -19712.
## No. Observations:                3082   AIC:                         3.944e+04
## Df Residuals:                    3074   BIC:                         3.949e+04
## Df Model:                           7
## Covariance Type:            nonrobust
## ==============================================================================
##                     coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## Intercept       -10.3461     19.832     -0.522      0.602     -49.231      28.539
## bath[T.1]        74.1826     11.211      6.617      0.000      52.201      96.164
## kitchen[T.1]    120.1973     13.024      9.229      0.000      94.660     145.735
## cheating[T.1]   161.5706      8.667     18.643      0.000     144.578     178.563
## location[T.2]    39.1827      5.449      7.191      0.000      28.499      49.866
## location[T.3]   126.0694     16.876      7.470      0.000      92.980     159.159
## area              4.5777      0.114     40.040      0.000       4.354       4.802
## iq               -0.1167      0.161     -0.725      0.469      -0.432       0.199
## ==============================================================================
## Omnibus:                      224.468   Durbin-Watson:                   2.004
## Prob(Omnibus):                  0.000   Jarque-Bera (JB):              582.605
## Skew:                           0.413   Prob(JB):                    3.08e-127
## Kurtosis:                       4.963   Cond. No.                        945.
## ==============================================================================
##
## Notes:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

## Justert-R2 med flere desimaler:

## 0.44926994879740134
```

Justert R^2

Dette er en justering til R^2 der vi veier med antall regresjonsparametere vi har estimert i modellen. På denne måten straffer vi en modell med mange regresjonsparametere - fordi den trolig kan tilpasses til støyfulle data - slik som vi så at vi kunne med eksemplet med IQ.

$$\text{Justert-}R^2 = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

Regelen er at vi velger den modellen som har høyest justert-R^2.

Det finnes mange konkurrende mål til justert R^2, noen av dem er med på utskriften (AIC, BIC), og det er også populært å bruke et såkalt valideringssett til å bestemme den beste modellen. Dette er et eget fagfelt, og vi skal ikke gå mer inn i detalj her.

Leieindeks - mange forklaringsvariabler (forts)

Vi sammenligner justert R^2 for modellen med og uten IQ:

- Modellen uten IQ har justert R^2 0.4494
- Modellen med IQ har justert-R^2 0.4493

Vi velger modellen uten IQ.

Hva gjør vi hvis modellen ikke passer?

- vi undersøker om det hjelper å ta inn andre eller flere forklaringsvariabler?
- vi undersøker om det hjelper å transformere forklaringsvariablene eller responsen?
- vi kan velger mer avanserte metoder

Hva kan vi ikke stole på av resultatene fra regresjonen?

- vi kan ikke stole på konfidensintervaller, prediksjonsintervaller og p-verdier fra hypotesetest

Men kan vi stole på de estimerte regresjonskoeffisientene og bruke regresjonsmodellen til å lage en prediksjon?

- Ja, i mange tilfeller vil de være å stole på – ikke alltid – men hvis vi bruker dem til å lage en prediksjon kan vi ikke si noe om hvor usikre vi er på prediksjonen.

Annet vi ikke har diskutert

- Korrelerte forklaringsvariabler kan gjøre det vanskelig å skille effekten av hver forklaringsvariabel. Dette kalles multikollinearitet.
- Hvis effekten av at en leilighet har et fint bad og et fint kjøkken er mye større enn bare summere hver effekt, da kan man innføre noe som heter samspill.

Noen råd til slutt

- Husk at vi ikke kan stole på modellen vår utenfor områdene vi har observert forklaringsvariablene våre - det er ikke sikkert at det er den samme lineære sammenhengen der!
- Hvis vi finner en lineær sammenheng mellom forklaringsvariabene og responsen så trenger det ikke bety at det er en årsakssammenheng. “Korrelasjon medfører ikke kausalitet.”
- Selv om det ikke er en lineær sammenheng mellom en forklaringsvariabel og responsen kan det være en ikke-lineær sammenheng. Det er lurt å plotte dataene!
- Husk at en kategorisk variabel bør kodes ved å bruke dummy-variabel koding.
- Husk at R^2 vil alltid øke når vi legger til nye forklaringsvariabler, selv om de bare er tilfeldig tall, og derfor kan ikke R^2 brukes til å velge den beste modellen.

Referanser

- Notasjonen er til en viss grad forsøkt tilpasset Fellesmodulen og Løvås kapittel 7.
- München leieindeks-dataene er lastet ned fra: https://www.uni-goettingen.de/de/551625.html
- Fahrmeir, Kneib, Lange, Marx (2013) https://www.springer.com/gp/book/9783642343322 er en ebok fra Springer, som du kan laste ned hvis du er på NTNU-nettet. Den brukes som lærebok i et matematisk emne i regresjon (TMA4267 Lineære statistiske metoder)
- James et al. (2013): An introduction to statistical learning - with applications in R http://faculty.marshall.usc.edu/gareth-james/ISL/, som også kan lastes ned fra Springer (men pdf er under lenken). Denne brukes som lærebok i TMA4268 Statistisk læring.