Forelesning 7: Regresjon

Sok-2009 h24

Eirik Eriksen Heen & ChatGPT

25. Sep 2024

Regresjonsanalyse er en statistisk metode som brukes til å modellere og analysere sammenhengen mellom en avhengig variabel (ofte kalt responsvariabel) og en eller flere uavhengige variabler (forklaringsvariabler). Formålet med regresjonsanalyse er å forstå hvordan endringer i de uavhengige variablene påvirker den avhengige variabelen.

Hvorfor bruke regresjonsanalyse?

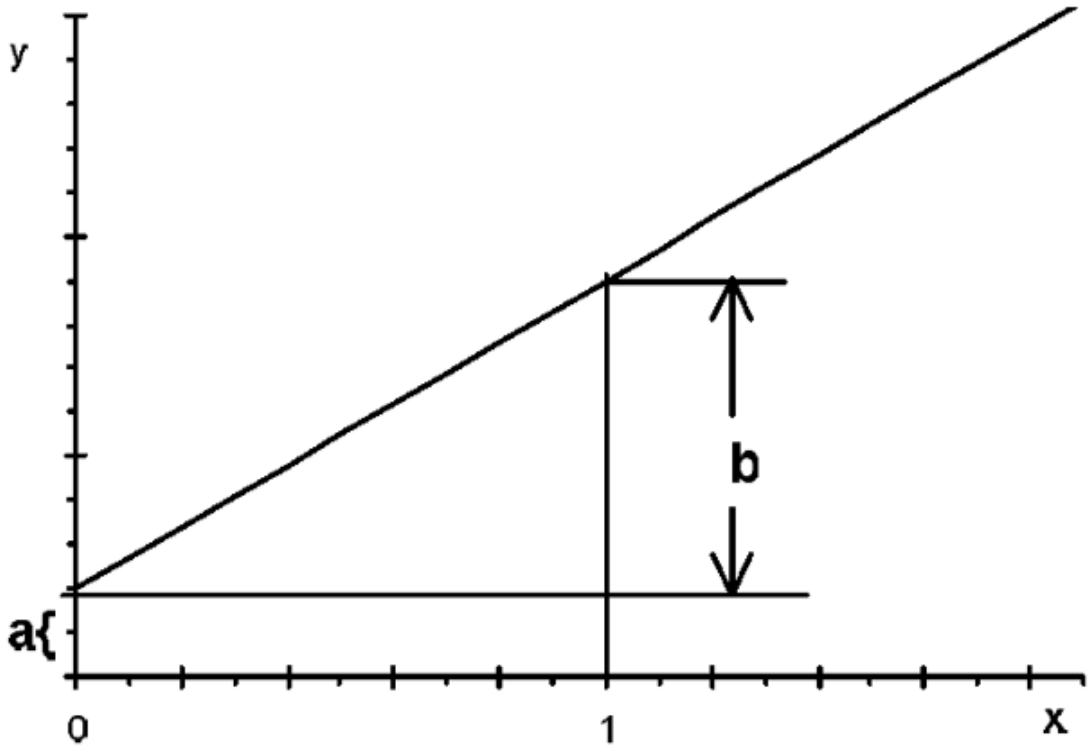
- **Forstå sammenhenger:** Regresjonsanalyse hjelper oss med å kvantifisere forholdet mellom variabler. For eksempel, hvordan endrer boligpriser seg med kvadratmeterpris eller antall rom?
- **Prediksjon:** Vi kan bruke modellen til å predikere verdier for den avhengige variabelen basert på gitte verdier av de uavhengige variablene. For eksempel, gitt en persons høyde, kan vi forutsi vekten.
- **Kausalitet:** Selv om regresjon alene ikke kan bevise kausalitet, kan den gi innsikt i om og hvordan en variabel potensielt kan påvirke en annen.

Rett linje

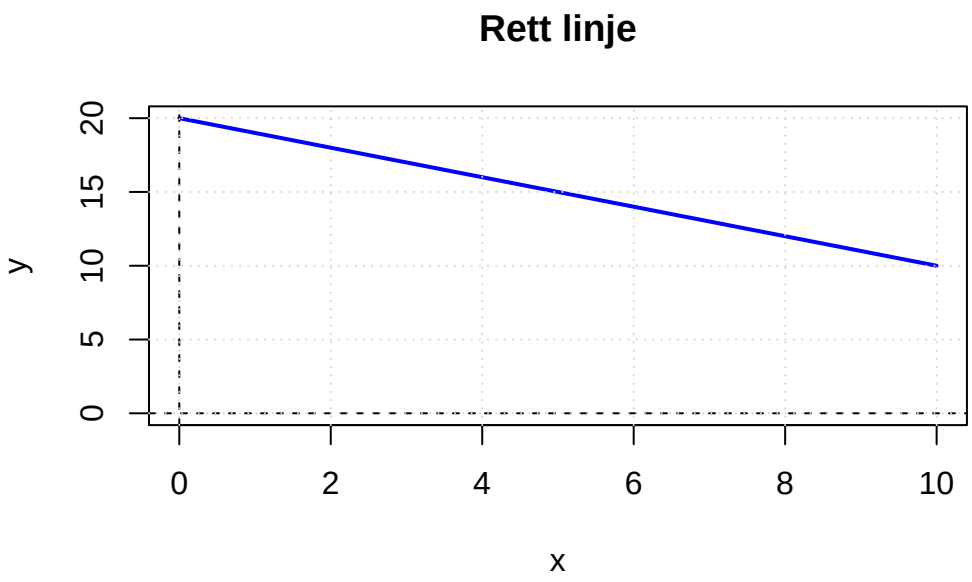
Rask oppfriskning av hva man tregner for en rett linje. En rett linje i mattematikken skrives

$$f(x) = y = b \times x + a$$

Her er b stigningstallet og a konstantleddet. For en visuell represntasjon:


y
b
a{
0
1
x

Grafisk represantasjon av Her er en linje med konstanledd på 20 og stigningstall på -1.



Lineær Regresjon

Den mest grunnleggende formen for regresjonsanalyse er lineær regresjon, hvor vi antar en lineær sammenheng mellom variablene. Den lineære regresjonsmodellen er gitt ved formelen:

$y = \beta_0 + \beta_1 x + \epsilon$

Her er:

- y: Den avhengige variabelen, variabelen vi ønsker å utforske
- x: Den uavhengige variabelen, variablenene som forklarer den avhengige variabelen
- β_0: Konstantleddet, som representerer skjæringspunktet med y-aksen
- β_1: Stigningstallet, som representerer effekten av x på y
- ϵ: Feilleddet, som representerer forskjellen mellom den observerte verdien og den verdien som er predikert av modellen.

Sammenheng mellom hus pris og størrelse

Her er en graf som viser sammenhengen mellom areal i kvadratfot og pris for et utvalg på 1000 hus. Den røde linjen representerer en enkel lineær regresjon for datasettet.

Virker det til å være en faktisk sammenheng eller er dette støy?

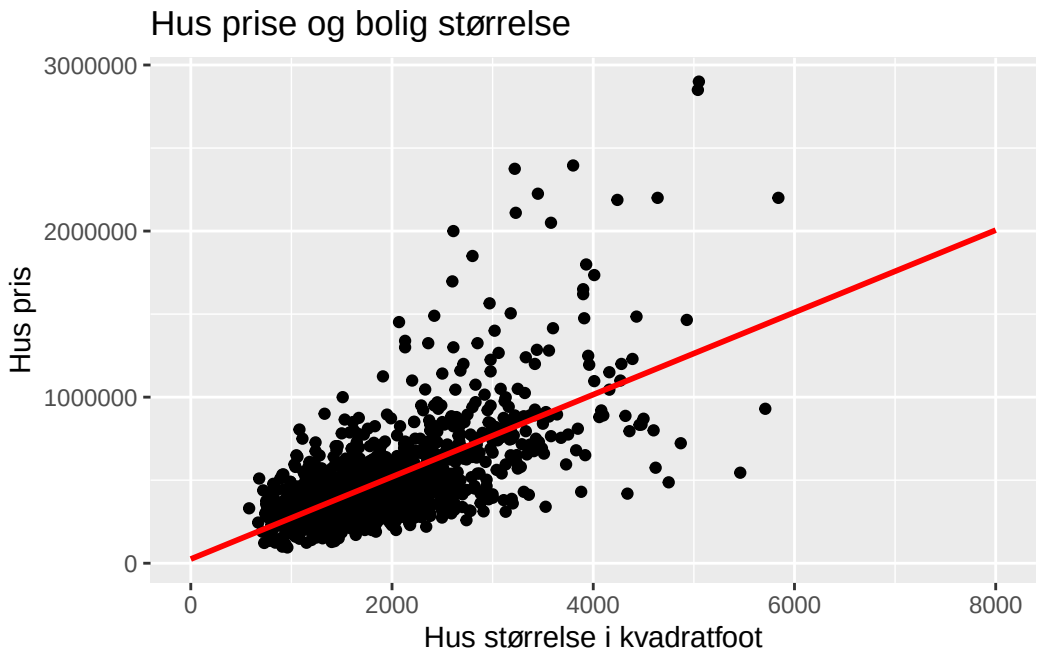
Hvis det er sammenheng. Påvirker prisen størrelsen eller størrelsen prisen?

Størrelse på huse er fast, men prisen kan varriere. Her påvirker størrelsen prisen på huset

$$\text{Hus pris} = \text{Pris per kvadratfoot} \times \text{kvadratfoot} + \text{konstant ledd}$$

```
set.seed(1337)

house_prices %>%
  slice_sample(n=1000) %>%
  ggplot(aes(x=sqft_living, y=price))+
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red", fullrange=TRUE) +
  labs(title = "Hus prise og bolig størrelse",
       x = "Hus størrelse i kvadratfoot",
       y = "Hus pris") +
  xlim(c(0,8000))
```



Finne regresjons linjen

For å finne regresjonslinjen, som representerer den lineære sammenhengen mellom to variabler, kan vi bruke funksjonen lm() i R. Denne funksjonen tilpasser en lineær modell til dataene dine ved å estimere koeffisientene for konstantleddet og stigningstallet.

I lm() koden legger vi inn avhengige variablen så ~ for å representere ærlik og legger vi inn de uavhengige variablene.

R-kode for å tilpasse en enkel lineær regresjon:

```
# En rett linje forklares av kun to variabler.
# Stigningstallet og konstantleddet er viktige elementer.
# Eksempel: lm(y ~ x, data = datasett)

# Eksempel:
model <- lm(price ~ sqft_living, data = house_prices)
model
```

```
Call:
lm(formula = price ~ sqft_living, data = house_prices)
```

```
Coefficients:
(Intercept)  sqft_living
-43580.7431     280.6236
```

Når vi bruker lm()-funksjonen, får vi en modell som gir en lineær prediksjon av price basert på sqft_living. For eksempel, hvis vi antar at den estimerte modellen er:

$$\text{Hus pris} = 281 \times \text{kvadratfoot} - 43580$$

Denne ligningen betyr at:

- **Stigningstallet (281)**: For hver ekstra kvadratfot øker husprisen med 281 dollar, gitt at alle andre faktorer holdes konstante. Dette er den gjennomsnittlige økningen i pris per kvadratfot.
- **Konstantleddet (-43,580)**: Dette representerer skjæringspunktet med y-aksen, altså den estimerte prisen når sqft_living er null. I dette tilfellet er verdien negativ, noe som ikke gir mening i en praktisk sammenheng. Det er viktig å merke seg at konstantleddet ofte ikke har en praktisk tolkning, spesielt når det er utenfor det observerte dataintervallet. Det er derfor viktig å være forsiktig med tolkningen av skjæringspunktet.

I praksis gir denne modellen oss en lineær prediksjon av boligpriser basert på størrelsen i kvadratfot, men vi må alltid vurdere konteksten og hvorvidt modellen gir meningsfulle resultater innenfor det observerte dataområdet.

Estimere Boligpris ved hjelp av Modellen

Når vi har tilpasset en lineær regresjonsmodell, kan vi bruke den til å estimere prisen på en bolig basert på antall kvadratfot. Dette kan enkelt gjøres ved å bruke funksjonen `makeFun()` i R.

For eksempel, la oss si at vi ønsker å estimere prisen på en bolig som er 2000 kvadratfot stor. Vi kan bruke `makeFun()`-funksjonen slik:

```r
# Gjør modelleom om til en funksjon
estimated_price <- makeFun(model)

# estimert prisen gitt at hus størrelsen er 2 000
estimated_price(2000)
```

```
          1
517666.3927
```

Med modellen vi har estimert, vil R bruke ligningen:

$$\text{Huspris} = 281 \times \text{kvadratfot} - 43,580$$

Til å beregne prisen:

$$\text{Huspris} = 281 \times 2000 - 43,580 = 517,666 \text{ dollars}$$

Dette betyr at den estimerte prisen for en bolig på 2000 kvadratfot er $517,666. Dette er en prediksjon basert på den lineære sammenhengen mellom kvadratfot og pris i datasettet vårt.

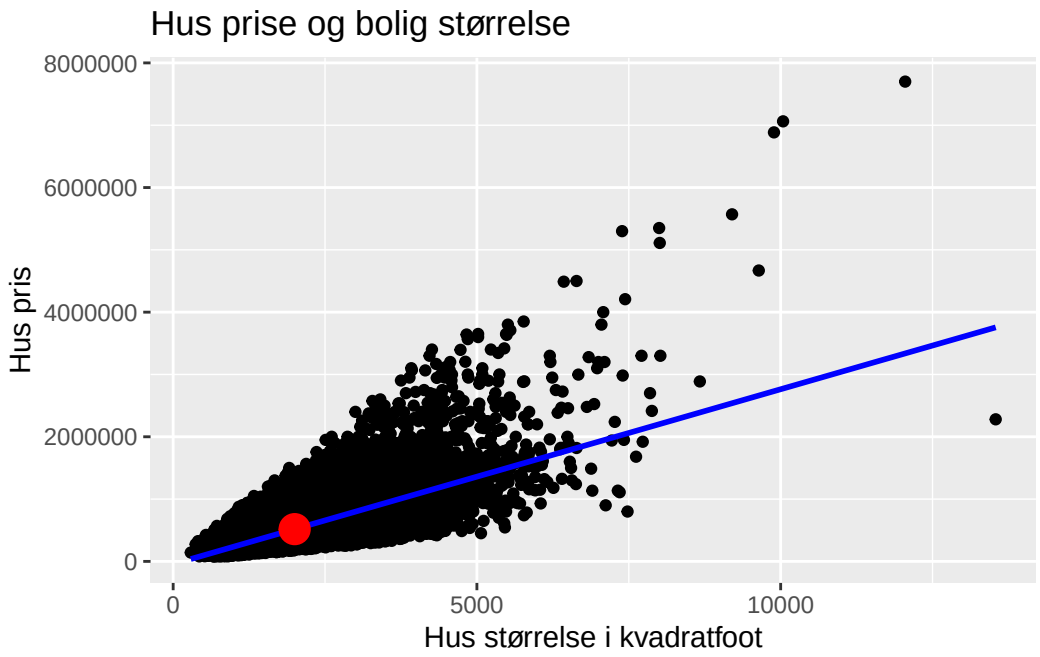
Alle prediksjonene er på regresjons linjen.

```
# lage en datafram med hus størrelse
Pris <- tibble(sqft_living = 2000)

# estimerer prisen til denne størrelsen
Pris <- Pris %>%
  mutate(price = estimated_price(sqft_living))




house_prices %>%
  ggplot(aes(x=sqft_living, y=price))+
  geom_point() +
  #koden for å tegne inn en regresjons linje
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Hus prise og bolig størrelse",
       x = "Hus størrelse i kvadratfoot",
       y = "Hus pris") +
  geom_point(data = Pris , color="red", size=5)
```

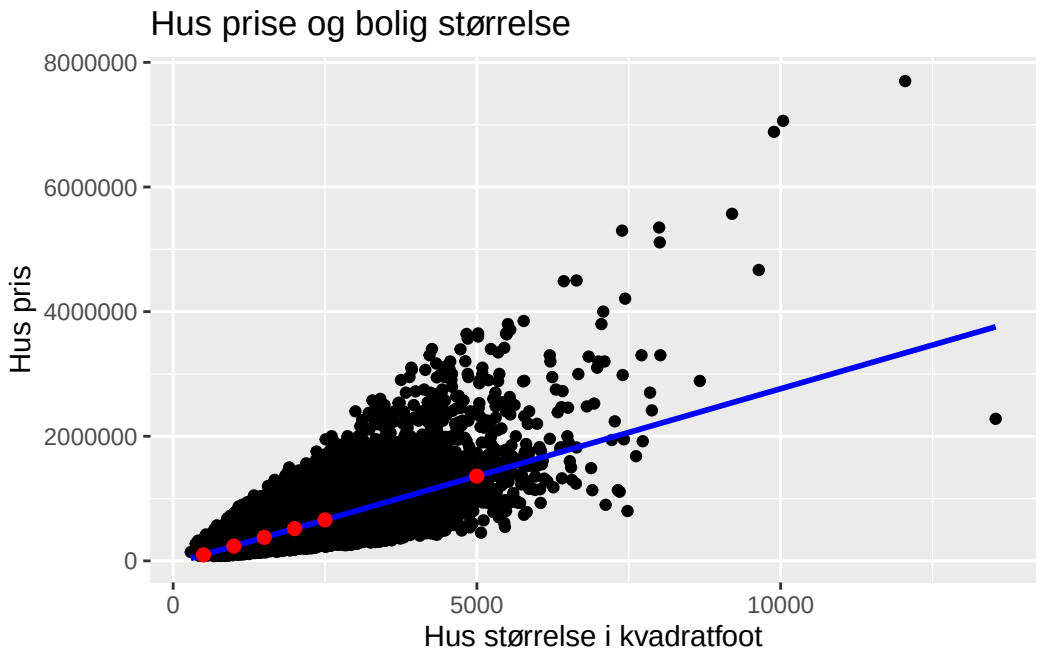

Hus prise og bolig størrelse
8000000
6000000
4000000
2000000
0
Hus pris
0
5000
10000
Hus størrelse i kvadratfoot

Dette gjelder for alle størrelser på huset.

```
# lage en datafram med hus størrelse
Pris <- tibble(sqft_living = c(500,1000,1500,2000,2500,5000))

# estimerer prisen til denne størrelsen
Pris <- Pris %>%
  mutate(price = estimated_price(sqft_living))


house_prices %>%
  ggplot(aes(x=sqft_living, y=price))+
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Hus prise og bolig størrelse",
       x = "Hus størrelse i kvadratfoot",
       y = "Hus pris") +
  geom_point(data = Pris , color="red", size=2)
```



Få ut mer informasjon av regersjonen

For å få en detaljert oversikt over regresjonsmodellen kan vi bruke funksjonen summary() i R. Denne funksjonen gir oss en omfattende rapport om modellen, inkludert estimater for koeffisientene, standardfeil, t-verdier, p-verdier, og andre viktige statistikker.

```
summary(model)
```

```
Call:
lm(formula = price ~ sqft_living, data = house_prices)

Residuals:
       Min          1Q      Median          3Q         Max
-1476062.4  -147486.0   -24042.8   106182.1  4362066.7

Coefficients:
                Estimate    Std. Error   t value   Pr(>|t|)
(Intercept) -43580.743094   4402.689690  -9.89866 < 2.22e-16 ***
sqft_living    280.623568      1.936399 144.92036 < 2.22e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 261452.9 on 21611 degrees of freedom
Multiple R-squared:  0.4928532, Adjusted R-squared:  0.4928298
F-statistic: 21001.91 on 1 and 21611 DF,  p-value: < 2.2204e-16
```

Denne kommandoen gir oss flere nyttige resultater, blant annet:

- **Koeffisienter (Estimate)**: Verdiene for konstantleddet (β_0) og stigningstallet (β_1), som forteller oss hvordan den avhengige variabelen (pris) endres i forhold til den uavhengige variabelen (kvadratfot).
- **Standardfeil (Std. Error)**: Et mål på usikkerheten i estimatene våre. Jo mindre standardfeil, desto mer presise er estimatene.
- **t-verdi (t value)**: Forholdet mellom en koeffisient og dens standardfeil. Det brukes til å teste hypotesen om at koeffisienten er lik null.
- **p-verdi (Pr(>|t|))**: P-verdien for hver koeffisient tester hvorvidt koeffisienten er statistisk signifikant forskjellig fra null. Dette er alltid en tosidig test. Hvis p-verdien er lav (typisk lavere enn 0.05), kan vi avvise nullhypotesen om at koeffisienten er null, og konkludere med at det er en signifikant sammenheng mellom den uavhengige og den avhengige variabelen.

Så her kan vi konkludere med at konstantleddet er mindre enn null og at stigningstallet er større enn null.

Evaluere Modelltilpasning

Selv om regresjonslinjen representerer en lineær sammenheng mellom variablene, er det viktig å merke seg at denne sammenhengen sjelden er perfekt. Det betyr at ikke alle datapunktene ligger på linjen; de vil typisk spre seg rundt linjen. For å forstå denne variasjonen bedre, kan vi bruke augment()-funksjonen fra broom-pakken i R.

Funksjonen augment() lar oss berike den opprinnelige datastrukturen med informasjon om modellen, som predikerte verdier, residualer, og andre målinger. Dette gir oss muligheten til å analysere hvert enkelt datapunkt i forhold til modellen.

Eksempel på bruk av augment():

```
augment(model)
```

```
# A tibble: 21,613 x 8
     price sqft_living  .fitted   .resid      .hat  .sigma    .cooksd .std.resid
     <dbl>       <int>    <dbl>    <dbl>     <dbl>   <dbl>      <dbl>      <dbl>
 1  221900        1180  287555.  -65655. 0.0000907 261459.    2.86e-6    -0.251
 2  538000        2570  677622. -139622. 0.0000594 261457.    8.48e-6    -0.534
 3  180000         770  172499.    7501. 0.000140  261459.    5.78e-8     0.0287
 4  604000        1960  506441.   97559. 0.0000471 261458.    3.28e-6     0.373
 5  510000        1680  427867.   82133. 0.0000550 261458.    2.72e-6     0.314
 6 1225000        5420 1477399. -252399. 0.000658  261453.    3.07e-4    -0.966
 7  257500        1715  437689. -180189. 0.0000536 261456.    1.27e-5    -0.689
 8  291850        1060  253880.   37970. 0.000103  261459.    1.09e-6     0.145
 9  229500        1780  455929. -226429. 0.0000512 261454.    1.92e-5    -0.866
10  323000        1890  486798. -163798. 0.0000482 261457.    9.47e-6    -0.627
# i 21,603 more rows
```

Denne funksjonen gir oss flere viktige kolonner:

- **.fitted**: Den predikerte verdien av den avhengige variabelen (pris) gitt modellen.
- **.resid**: Residualen, som er forskjellen mellom den faktiske og den predikerte verdien (dvs. $y_i - \hat{y}_i$). Residualene forteller oss hvor mye modellen bommer for hvert datapunkt. Hvis residualen er stor, betyr det at modellen ikke passer godt til dette spesifikke datapunktet.
- **.hat**: Hat-verdien som indikerer hvor mye innflytelse et gitt datapunkt har på tilpasningen av modellen. Høyere verdier indikerer at punktet har stor innflytelse.

- **.cooksd**: Cook's Distance, som måler påvirkningen til et datapunkt på hele modellen. Punkter med høye Cook's D-verdier bør undersøkes nærmere da de kan være uteliggere.

Første obersvasjon ser av vi at hadde boligen følgt dataen så burde den ha kostet 287 555, dette er en "bom" på -65 655.

Estimering av Regresjonslinjen

Når vi utfører en regresjonsanalyse, er målet å finne den linjen som best representerer sammenhengen mellom variablene i dataene våre. Denne linjen kalles ofte "best fit" linjen. Det er viktig å forstå at denne linjen ikke er den "sanne" linjen – fordi den sanne sammenhengen mellom variablene er ukjent og påvirkes av flere faktorer. I stedet er det vi estimerer, basert på dataene vi har tilgjengelig, en tilnærming til den sanne sammenhengen.

Minste Kvadraters Metode

For å finne den beste tilpasningen bruker vi en teknikk kalt minste kvadraters metode (Ordinary Least Squares, OLS). Denne metoden søker å minimere summen av de kvadratiske avvikene mellom de observerte verdiene og de verdiene som er predikert av modellen. Dette betyr at vi finner den linjen som gjør den totale "feilen" mellom de faktiske dataene og den predikerte linjen så liten som mulig.

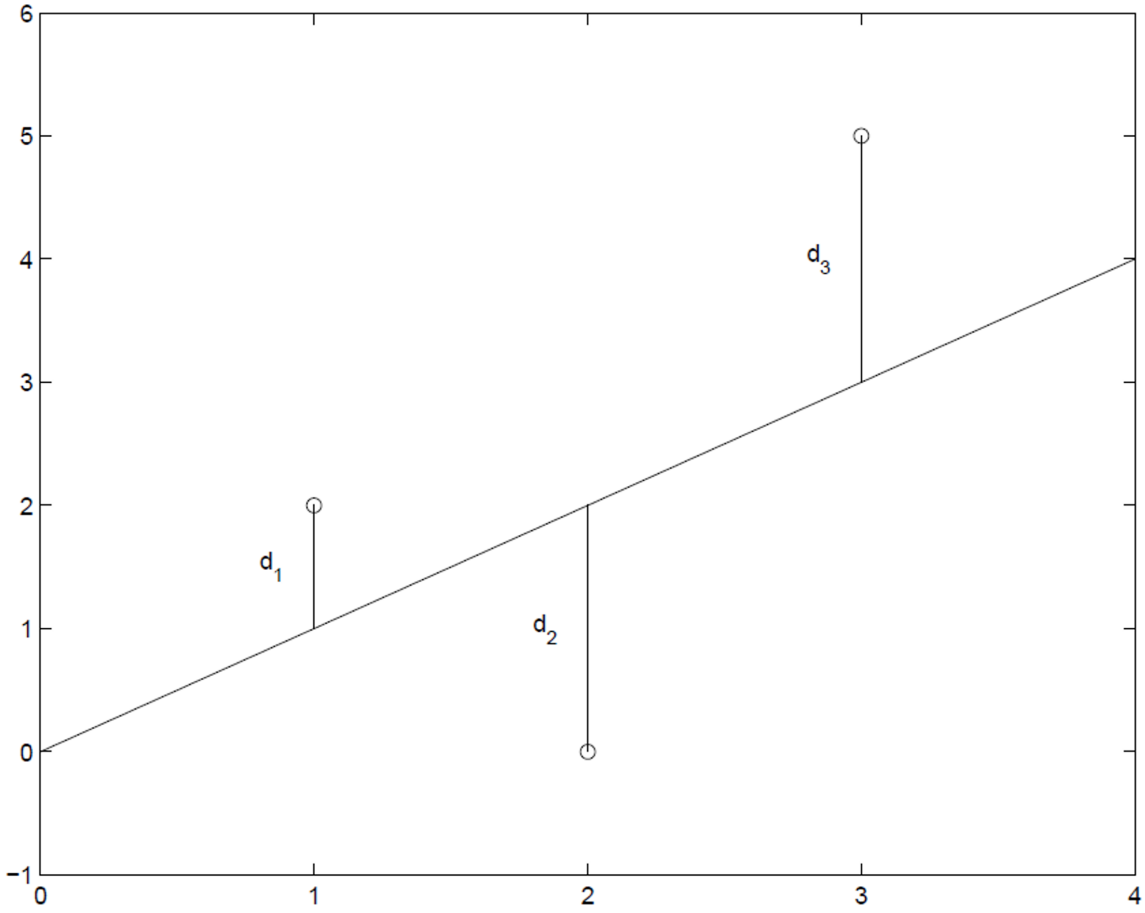
Matematisk kan dette uttrykkes slik:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

Her er:

- y_i den observerte verdien for den avhengige variabelen for observasjon i.
- $\widehat{y}_i = \beta_0 + \beta_1 x_i$ den predikerte verdien fra modellen for observasjon iii, gitt en uavhengig variabel x_i.
- β_0 og β_1 er koeffisientene vi estimerer for å finne den beste tilpasningen.
- Summen $\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$ er den totale kvadratiske avstanden mellom de faktiske og predikerte verdiene, ofte referert til som "residual sum of squares" (RSS).

Ved å minimere RSS, finner vi den linjen som best mulig tilpasser dataene våre i henhold til minste kvadraters metode. Dette er grunnen til at linjen kalles en "best fit"-linje.



Hva betyr dette i praksis?

I praksis betyr det at for hver punkt iii i datasettet, finner modellen en predikert verdi $\hat{y}_i$ Differansen mellom den faktiske verdien y_i og den predikerte verdien $\hat{y}_i$ kalles en residual. Ved å kvadrere disse residualene og summere dem, kan vi måle hvor godt modellen passer til dataene.

Modellen søker å gjøre denne summen så liten som mulig, noe som resulterer i en linje som best mulig representerer sammenhengen i datasettet. Når du ser på resultatene fra en regresjonsanalyse, er det viktig å huske på at denne linjen kun er en estimert tilnærming, og at den er basert på de observerte dataene. Naturlig variasjon og ukjente faktorer gjør at den sanne sammenhengen alltid vil ha en viss grad av usikkerhet.

Grunn til noe av avviken

Selv om størrelsen på en bolig er en viktig faktor i å bestemme prisen, er det viktig å erkjenne at vi ikke kan forutsi boligprisene nøyaktig basert kun på denne ene variabelen. Det er mange andre faktorer som også spiller en avgjørende rolle, for eksempel beliggenhet, boligtilstand, antall soverom, nærhet til fasiliteter, skolekvalitet og økonomiske trender. Disse faktorene bidrar til variabilitet i prisene som vår modell ikke fanger opp. Dette fører til at det oppstår feil eller avvik mellom de predikerte prisene og de faktiske prisene, fordi modellen er en forenkling av den komplekse virkeligheten.

Forklart varians R^2

R^2, også kjent som determinasjonskoeffisienten, er et mål på hvor godt regresjonsmodellen forklarer variansen i den avhengige variabelen y. R^2 beregnes ved å kvadrere korrelasjonskoeffisienten RRR (også kjent som Pearsons R), som kan variere fra -1 til 1. Når vi kvadrerer R, vil R^2 alltid variere mellom 0 og 1.

- R^2 på 0 indikerer at modellen ikke forklarer noen av variansen i den avhengige variabelen; modellen gir ingen bedre prediksjoner enn gjennomsnittet.
- R^2 på 1 indikerer at modellen forklarer all variansen i den avhengige variabelen; alle de observerte verdiene ligger nøyaktig på regresjonslinjen.

R^2 beskriver altså hvor stor andel av variansen i den avhengige variabelen som kan forklares av modellen.

```
summary(model)
```

```
Call:
lm(formula = price ~ sqft_living, data = house_prices)

Residuals:
       Min         1Q     Median         3Q        Max
-1476062.4  -147486.0   -24042.8   106182.1  4362066.7

Coefficients:
                Estimate    Std. Error   t value   Pr(>|t|)
(Intercept) -43580.743094   4402.689690  -9.89866 < 2.22e-16 ***
sqft_living    280.623568      1.936399 144.92036 < 2.22e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 261452.9 on 21611 degrees of freedom
Multiple R-squared:  0.4928532, Adjusted R-squared:  0.4928298
F-statistic: 21001.91 on 1 and 21611 DF,  p-value: < 2.2204e-16
```

I regresjons modellen vise den som *Multiple R-squared: 0.493*.

For eksempel, hvis R^2 for en lineær modell som predikerer pris basert på bolig størrelse er 0,493, betyr dette at modellen forklarer 49 % av variansen i pris. De resterende 51 % av variansen forklares av andre faktorer som modellen ikke tar hensyn til.

R^2 er derfor et viktig mål på modellens prediksjonsevne, hvor en høyere R^2 indikerer en bedre tilpasning til dataene.

Residual standard error

Residual Standard Error (RSE) er et mål på hvor mye de observerte verdiene avviker fra de predikerte verdiene i en regresjonsmodell. Det gir et gjennomsnittlig avvik eller "feil" mellom modellens prediksjoner og de faktiske dataene. En lav RSE indikerer at modellen passer godt til dataene, mens en høy RSE tyder på større avvik og mindre nøyaktige prediksjoner.

```r
summary(model)
```

```
Call:
lm(formula = price ~ sqft_living, data = house_prices)

Residuals:
       Min         1Q     Median         3Q        Max
-1476062.4  -147486.0   -24042.8   106182.1  4362066.7

Coefficients:
                Estimate    Std. Error   t value   Pr(>|t|)
(Intercept) -43580.743094   4402.689690  -9.89866 < 2.22e-16 ***
sqft_living    280.623568      1.936399 144.92036 < 2.22e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 261452.9 on 21611 degrees of freedom
Multiple R-squared:  0.4928532, Adjusted R-squared:  0.4928298
F-statistic: 21001.91 on 1 and 21611 DF,  p-value: < 2.2204e-16
```

Når vi forklarte boligpriser basert på boligens størrelse i vår modell, var Residual Standard Error 261,000. Dette betyr at den gjennomsnittlige avstanden mellom de faktiske boligprisene og de prisene som modellen predikerte, var 261,000 dollar.

Oppgave

I denne oppgaven bruk NANES datasettet.

1. Se på sammenhengen mellom høyde og vekt. Hva burde være avhengig variabel og hva burde være uavhengig variabel?
2. Plott dataen og legg ved en trendlinje i datasettet
3. Gjennomfør en regresjon med høyde og vekt
4. Nå sammenliken høyde og kjønn. Start med å regne ut gjennomsnittshøyd for kjønnene og legg til differanse mellom dem.
5. Gjennomfør nå er regresjon med Høyde som avhengig variabel og kjønn som uavhengig variabel? Hva skjer? sammenlikn svaret med svaret i oppgave 4.

Regresjon med nominal variabler

Regresjonsanalyse kan også gjennomføres med nominale variabler, som kategoriske variabler som representerer ulike grupper. Når en nominal variabel brukes i en regresjonsmodell, kodes den automatisk som en indikatorvariabel (dummy-variabel). Dette betyr at én kategori vil bli brukt som referansepunkt (skjæringspunktet eller intercept), mens de andre kategoriene viser forskjellen i gjennomsnitt mellom den aktuelle kategorien og referansekategorien.

Eksempel: La oss anta at vi vil undersøke forskjellen i boligpris mellom hus som ligger ved vannet og de som ikke gjør det.

```
gjen_pris <- house_prices %>%
  group_by(waterfront) %>%
  summarise(mean = mean(price))

gjen_pris
```

```
# A tibble: 2 x 2
  waterfront     mean
  <lgl>         <dbl>
1 FALSE       531564.
2 TRUE       1661876.
```

```r
diff(gjen_pris$mean)
```

```
[1] 1130312.425
```

I koden ovenfor grupperer vi først dataene basert på waterfront-variabelen (som angir om huset har vannutsikt eller ikke) og beregner gjennomsnittsprisen for hver gruppe. Deretter finner vi differansen i gjennomsnittspris mellom de to gruppene.

For å inkludere dette i en regresjonsmodell:

```r
model <- lm(price ~ waterfront, data = house_prices)
summary(model)
```

```
Call:
lm(formula = price ~ waterfront, data = house_prices)

Residuals:
       Min         1Q     Median         3Q        Max
-1376876.0  -211563.6   -81563.6   108436.4  7168436.4

Coefficients:
                 Estimate  Std. Error   t value   Pr(>|t|)
(Intercept)     531563.600    2416.194 220.00034 < 2.22e-16 ***
waterfrontTRUE 1130312.425   27822.465  40.62589 < 2.22e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 353871.4 on 21611 degrees of freedom
Multiple R-squared:  0.07095268,    Adjusted R-squared:  0.07090969
F-statistic: 1650.463 on 1 and 21611 DF,  p-value: < 2.2204e-16
```

I denne regresjonsmodellen:

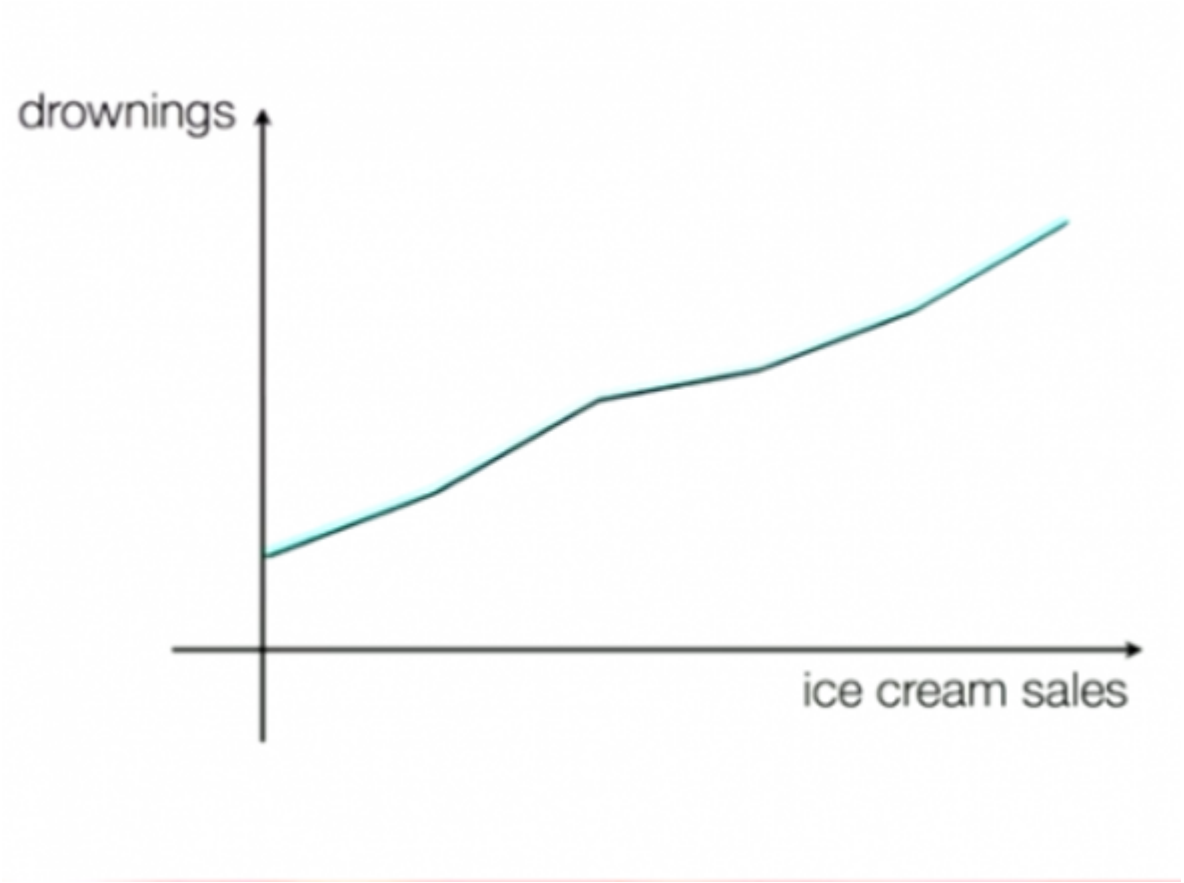
- **Intercept**: Dette representerer gjennomsnittsprisen for boliger uten vannutsikt (referansekategorien).
- **waterfrontTRUE**: Dette viser differansen i pris mellom boliger med vannutsikt og de uten. Verdien her vil tilsvare differansen vi tidligere beregnet manuelt. For eksempel, hvis `waterfrontTRUE` er 1,130,312, betyr dette at boliger med vannutsikt i gjennomsnitt koster 1,130,312 dollar mer enn boliger uten vannutsikt.

Denne metoden lar oss enkelt inkludere nominale variabler i regresjonsmodeller og tolke forskjeller mellom kategorier direkte fra modellresultatene.

Virkelig sammenheng?

Det er viktig å tenke seg litt om når vi gjør regresjoner. Vi kan altids finne at variabler er korrelet, men det betyr ikke at de er avhengig av hverander.

Fører mer solgt iskrem til at flere drukner?



Nei. Hvis det er pent vær så er det flere som drar på stranden og flere som kjøper iskrem.

Oppgave

I datasettet babies fra moderndive, skal vi se litt på sammenhenger.

1. Er det en sammenheng mellom vekt på baybien (wt) og om moren røkte (smoke)? Skal vi tror at det er noe sammenheng?

    1. wt Birth weight in ounces, marked 999 if unknown

    2. smoke Does mother smoke? 0=never, 1= smokes now, 2=until current pregnancy, 3=once did, not now, 9=unknown

2. Er det en sammenheng mellom kjønnet på baybien (sex) og om moren røkte (smoke)? Skal vi tror at det er noe sammenheng?

1. `sex Infant's sex, where 1 is male, 2 is female, and 9 is unknown`

3. Er det en sammenheng mellom vekt på baybien (wt) og om moren har vert gravid før (parity)? Skal vi tror at det er noe sammenheng?

    1. parity Total number of previous pregnancies including fetal deaths and stillbirths, marked 99 if unknown

Regresjon i en pipe

Du kan legge til en regresjon i en pipe med å bruke do() funksjonen.

```
model <- house_prices %>%
  filter(!is.na(waterfront)) %>%  # Eventuelt filtrere bort NA-verdier
  group_by(waterfront) %>%         # Gruppes for eventuell mer analyse, hvis nødvendig
  do(model = lm(price ~ sqft_living, data = .))

# For å se resultatene:
summary(model$model[[1]])
```

```
Call:
lm(formula = price ~ sqft_living, data = .)

Residuals:
       Min         1Q     Median         3Q        Max
-1284141.1  -141826.0   -23026.5   103790.6  4529857.7

Coefficients:
                Estimate    Std. Error   t value    Pr(>|t|)
(Intercept) -16223.750674   4147.370725  -3.91182 0.000091887 ***
sqft_living    264.428717      1.834197 144.16593  < 2.22e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 243445.1 on 21448 degrees of freedom
Multiple R-squared:  0.4921364, Adjusted R-squared:  0.4921128
F-statistic: 20783.82 on 1 and 21448 DF,  p-value: < 2.2204e-16
```