Forelesning 9: Multippel regresjon

Sok-2009 h24

Eirik Eriksen Heen & ChatGPT

Invalid Date

Multippel regresjon er en utvidelse av enkel lineær regresjon, hvor vi kan inkludere flere uavhengige variabler for å forklare variasjonen i den avhengige variabelen. Dette gir oss en mer nyansert forståelse av forholdene mellom variablene, spesielt i komplekse datasett som **NHANES**, hvor flere faktorer kan påvirke utfallet vi undersøker.

I alle eksemplene under hvor vi skal bruke NHANES settet skal vi kun bruke personer som er 20 år eller høyere.

```
# Lager et nytt datasett
NHANES20 <- NHANES %>%
    # Bruker kun personer som er 20 år eller eldre
    filter(Age >= 20)
```

Formål

I enkel lineær regresjon estimerer vi forholdet mellom en uavhengig og en avhengig variabel. Men ofte er det flere faktorer som påvirker det vi ønsker å undersøke. For eksempel kan både høyde og vekt påvirke blodtrykk, og vi ønsker å modellere dette samtidig.

Formel for multippel regresjon

Den generelle modellen for multippel regresjon kan skrives som:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

- y er den avhengige variabelen (utfallet vi ønsker å forklare, f.eks. blodtrykk).
- $x_1, x_2, \ldots, x_n$ er de uavhengige variablene (f.eks. høyde, vekt, alder).

- $\beta_1, \beta_2, \ldots, \beta_n$ er koeffisientene som viser hvor mye hver uavhengig variabel påvirker yyy, når de andre variablene holdes konstante.
- ϵ er feilleddet som representerer tilfeldig variasjon eller ukjente faktorer.

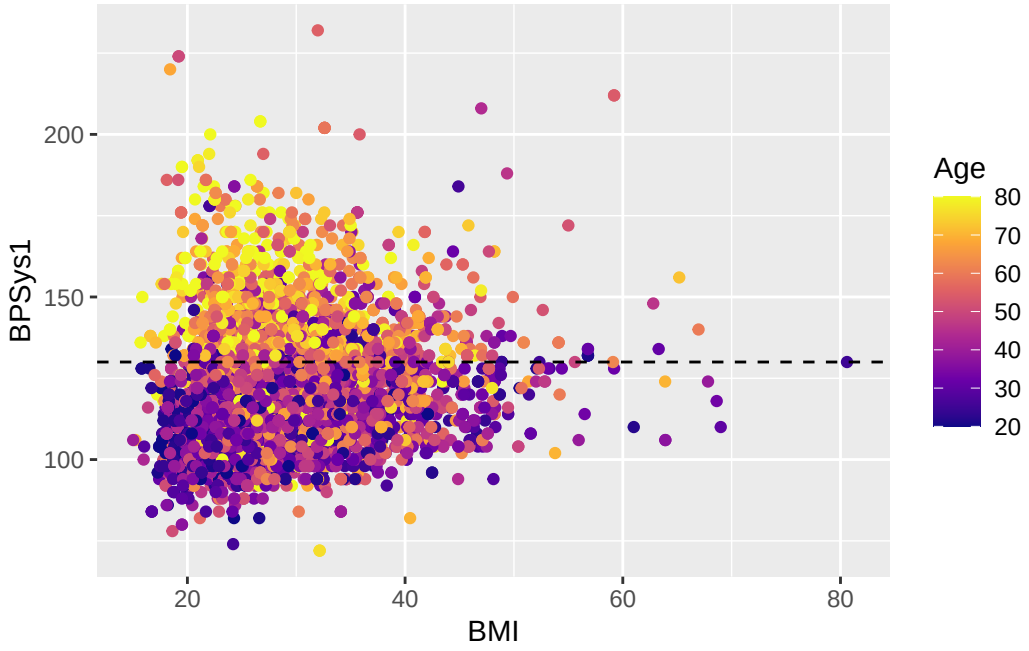
Eksempel fra NHANES

La oss se på et eksempel fra NHANES-datasettet, der vi er interessert i å modellere systolisk blodtrykk som en funksjon av BMI og alder. I dette tilfellet kan vi sette opp modellen som:

$$\text{Blodtrykk} = \alpha + \beta_1 \cdot \text{BMI} + \beta_2 \cdot \text{Alder} + \epsilon$$

Her ønsker vi å undersøke hvordan både BMI og alder påvirker blodtrykket, mens vi kontrollerer for den andre variabelen. Det gir oss en forståelse av hvor mye blodtrykket øker eller synker med endringer i BMI og alder, uavhengig av hverandre. Når vi skal se på blodtrykk ser vi hovedsakelig på personer som er 20 år eller eldre. Et blodtrykk på 130 eller høyere er ansett til å være høytbodtrykk. Vi starter med å plotte dataen.

```
ggplot( NHANES20, aes(x=BMI,y=BPSys1 , color = Age)) +
  geom_point() +
  # Lager en "plasma" farger for å få litt mer kontrast
  scale_color_viridis_c(option = "plasma") +
  # Legger inn en linje for hva som er høyt blodtrykk
  geom_hline(yintercept = 130, color = "black", linetype = "dashed")
```



Det ser ut som at de som er eldre har høyere boldtrykk. Det ser også ut til å være en effekt av BMI også, men ikke like mye effekt som alder.

For å estimere dette bruker vi `lm` funksjonen i R. Vi legger kun et pluss mellom de uavhengige variablene.

```
# gjennomfører regresjon
reg1 <- lm(BPSys1 ~ BMI + Age, data = NHANES20)

summary(reg1)
```

```
Call:
lm(formula = BPSys1 ~ BMI + Age, data = NHANES20)

Residuals:
   Min     1Q Median     3Q    Max
-63.65  -9.72  -0.89   8.31 105.62

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  93.3718     0.9822   95.06   <2e-16 ***
BMI           0.2767     0.0289    9.59   <2e-16 ***
Age           0.4392     0.0113   39.00   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.5 on 6663 degrees of freedom
  (569 observations deleted due to missingness)
Multiple R-squared:  0.198, Adjusted R-squared:  0.198
F-statistic:  825 on 2 and 6663 DF,  p-value: <2e-16
```

Tolkning av koeffisienter

- **Interceptet** (α) representerer det forventede blodtrykket for en person med BMI og alder lik null (som ikke er realistisk, men en nødvendig komponent i modellen).
- **Koeffisientene** (β_1 og β_2) viser hvor mye blodtrykket forventes å endre seg med en enhets endring i BMI og alder, henholdsvis, mens den andre variabelen holdes konstant. β_1 for BMI er positiv, betyr det at blodtrykket øker med økt BMI, forutsatt at alderen er konstant.

- Så β_1 eller gitt som BMI i tabellen, på gjennomsnitt gitt at alderen er konstant så fører økning på BMI på 1 til 0.28 økning i blodtrykk. Siden p-verdien er lavere enn 5% så kan vi forkaste nullhypotesen at det IKKE er effekt av BMI på blodtrykk. Vi tar i bruk alternativ hypotesen at BMI påvirker blodtrykk.
- Så β_2 eller gitt som Age i tabellen, på gjennomsnitt gitt at BMI er konstant så fører økning på alder på 1 år til 0.44 økning i blodtrykk. Siden p-verdien er lavere enn 5% så kan vi forkaste nullhypotesen at det IKKE er effekt av alder på blodtrykk. Vi tar i bruk alternativ hypotesen at alder påvirker blodtrykk.

Multippel regresjon gir altså en kraftigere modell når vi har flere faktorer som potensielt påvirker utfallet vårt, noe som er spesielt nyttig i komplekse helsedata som NHANES.

Vi ser at *Residual standard error* er 15.5, eller på gjennomsnitt bommer modellen med 15.5 i boldtrykk. R^2 er 0.198 som ikke er spesielt høy, vi klarer å forklare rundt 20% av variabiliteten til boldtrykk, med BMI og alder.

Oppgave

Hva andre faktorer kan påvirke blodtrykk?

```
glimpse(NHANES20)
```

```
Rows: 7,235
Columns: 76
$ ID              <int> 51624, 51624, 51624, 51630, 51647, 51647, 51647, 5165~
$ SurveyYr        <fct> 2009_10, 2009_10, 2009_10, 2009_10, 2009_10, 2009_10,~
$ Gender          <fct> male, male, male, female, female, female, female, mal~
$ Age             <int> 34, 34, 34, 49, 45, 45, 45, 66, 58, 54, 58, 50, 33, 6~
$ AgeDecade       <fct>  30-39,  30-39,  30-39,  40-49,  40-49,  40-49,  40-4~
$ AgeMonths       <int> 409, 409, 409, 596, 541, 541, 541, 795, 707, 654, 700~
$ Race1           <fct> White, White, White, White, White, White, White, Whit~
$ Race3           <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ Education       <fct> High School, High School, High School, Some College, ~
$ MaritalStatus   <fct> Married, Married, Married, LivePartner, Married, Marr~
$ HHIncome        <fct> 25000-34999, 25000-34999, 25000-34999, 35000-44999, 7~
$ HHIncomeMid     <int> 30000, 30000, 30000, 40000, 87500, 87500, 87500, 3000~
$ Poverty         <dbl> 1.36, 1.36, 1.36, 1.91, 5.00, 5.00, 5.00, 2.20, 5.00,~
$ HomeRooms       <int> 6, 6, 6, 5, 6, 6, 6, 5, 10, 6, 10, 4, 11, 5, 10, 10, ~
$ HomeOwn         <fct> Own, Own, Own, Rent, Own, Own, Own, Own, Rent, Rent, ~
$ Work            <fct> NotWorking, NotWorking, NotWorking, NotWorking, Worki~
$ Weight          <dbl> 87.4, 87.4, 87.4, 86.7, 75.7, 75.7, 75.7, 68.0, 78.4,~
$ Length          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

```
$ HeadCirc          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ Height            <dbl> 164.7, 164.7, 164.7, 168.4, 166.7, 166.7, 166.7, 169.~
$ BMI               <dbl> 32.22, 32.22, 32.22, 30.57, 27.24, 27.24, 27.24, 23.6~
$ BMICatUnder20yrs <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ BMI_WHO           <fct> 30.0_plus, 30.0_plus, 30.0_plus, 30.0_plus, 25.0_to_2~
$ Pulse             <int> 70, 70, 70, 86, 62, 62, 62, 60, 62, 76, 94, 74, 96, 8~
$ BPSysAve          <int> 113, 113, 113, 112, 118, 118, 118, 111, 104, 134, 127~
$ BPDiaAve          <int> 85, 85, 85, 75, 64, 64, 64, 63, 74, 85, 83, 68, 74, 1~
$ BPSys1            <int> 114, 114, 114, 118, 106, 106, 106, 124, 108, 136, NA,~
$ BPDia1            <int> 88, 88, 88, 82, 62, 62, 62, 64, 76, 86, NA, 66, 80, 9~
$ BPSys2            <int> 114, 114, 114, 108, 118, 118, 118, 108, 104, 132, 134~
$ BPDia2            <int> 88, 88, 88, 74, 68, 68, 68, 62, 72, 88, 82, 74, 74, 9~
$ BPSys3            <int> 112, 112, 112, 116, 118, 118, 118, 114, 104, 136, 120~
$ BPDia3            <int> 82, 82, 82, 76, 60, 60, 60, 64, 76, 82, 84, 62, NA, 1~
$ Testosterone      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ DirectChol        <dbl> 1.29, 1.29, 1.29, 1.16, 2.12, 2.12, 2.12, 0.67, 0.96,~
$ TotChol           <dbl> 3.49, 3.49, 3.49, 6.70, 5.82, 5.82, 5.82, 4.99, 4.24,~
$ UrineVol1         <int> 352, 352, 352, 77, 106, 106, 106, 113, 163, 215, 29, ~
$ UrineFlow1        <dbl> NA, NA, NA, 0.094, 1.116, 1.116, 1.116, 0.489, NA, 0.~
$ UrineVol2         <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ UrineFlow2        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ Diabetes          <fct> No, No, No, No, No, No, No, No, No, No, No, No, No, N~
$ DiabetesAge       <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ HealthGen         <fct> Good, Good, Good, Good, Vgood, Vgood, Vgood, Vgood, V~
$ DaysPhysHlthBad   <int> 0, 0, 0, 0, 0, 0, 0, 10, 0, 4, NA, 0, 3, 7, 3, 3, 0, ~
$ DaysMentHlthBad   <int> 15, 15, 15, 10, 3, 3, 3, 0, 0, 0, NA, 0, 7, 0, 0, 0, ~
$ LittleInterest    <fct> Most, Most, Most, Several, None, None, None, None, No~
$ Depressed         <fct> Several, Several, Several, Several, None, None, None,~
$ nPregnancies      <int> NA, NA, NA, 2, 1, 1, 1, NA, NA, NA, NA, NA, NA, NA, 4~
$ nBabies           <int> NA, NA, NA, 2, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ Age1stBaby        <int> NA, NA, NA, 27, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ SleepHrsNight     <int> 4, 4, 4, 8, 8, 8, 8, 7, 5, 4, 5, 7, 6, 6, 7, 7, 8, 6,~
$ SleepTrouble      <fct> Yes, Yes, Yes, Yes, No, No, No, No, No, Yes, No, No, ~
$ PhysActive        <fct> No, No, No, No, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Ye~
$ PhysActiveDays    <int> NA, NA, NA, NA, 5, 5, 5, 7, 5, 1, 2, 7, NA, NA, 7, 7,~
$ TVHrsDay          <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ CompHrsDay        <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ TVHrsDayChild     <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ CompHrsDayChild   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ Alcohol12PlusYr   <fct> Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, NA,~
$ AlcoholDay        <int> NA, NA, NA, 2, 3, 3, 3, 1, 2, 6, NA, NA, 3, 6, 1, 1, ~
$ AlcoholYear       <int> 0, 0, 0, 20, 52, 52, 52, 100, 104, 364, NA, 0, 104, 3~
$ SmokeNow          <fct> No, No, No, Yes, NA, NA, NA, No, NA, NA, Yes, NA, No,~
```

```
$ Smoke100         <fct> Yes, Yes, Yes, Yes, No, No, No, Yes, No, No, Yes, No,~
$ Smoke100n        <fct> Smoker, Smoker, Smoker, Smoker, Non-Smoker, Non-Smoke~
$ SmokeAge         <int> 18, 18, 18, 38, NA, NA, NA, 13, NA, NA, 17, NA, NA, 1~
$ Marijuana        <fct> Yes, Yes, Yes, Yes, Yes, Yes, Yes, NA, Yes, Yes, NA, ~
$ AgeFirstMarij    <int> 17, 17, 17, 18, 13, 13, 13, NA, 19, 15, NA, NA, NA, N~
$ RegularMarij     <fct> No, No, No, No, No, No, No, NA, Yes, Yes, NA, No, No,~
$ AgeRegMarij      <int> NA, NA, NA, NA, NA, NA, NA, NA, 20, 15, NA, NA, NA, N~
$ HardDrugs        <fct> Yes, Yes, Yes, Yes, No, No, No, No, Yes, Yes, NA, No,~
$ SexEver          <fct> Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, NA,~
$ SexAge           <int> 16, 16, 16, 12, 13, 13, 13, 17, 22, 12, NA, NA, 27, 2~
$ SexNumPartnLife  <int> 8, 8, 8, 10, 20, 20, 20, 15, 7, 100, NA, 9, 1, 1, 2, ~
$ SexNumPartYear   <int> 1, 1, 1, 1, 0, 0, 0, NA, 1, 1, NA, 1, 1, NA, 1, 1, 1,~
$ SameSex          <fct> No, No, No, Yes, Yes, Yes, Yes, No, No, No, NA, No, N~
$ SexOrientation   <fct> Heterosexual, Heterosexual, Heterosexual, Heterosexua~
$ PregnantNow      <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

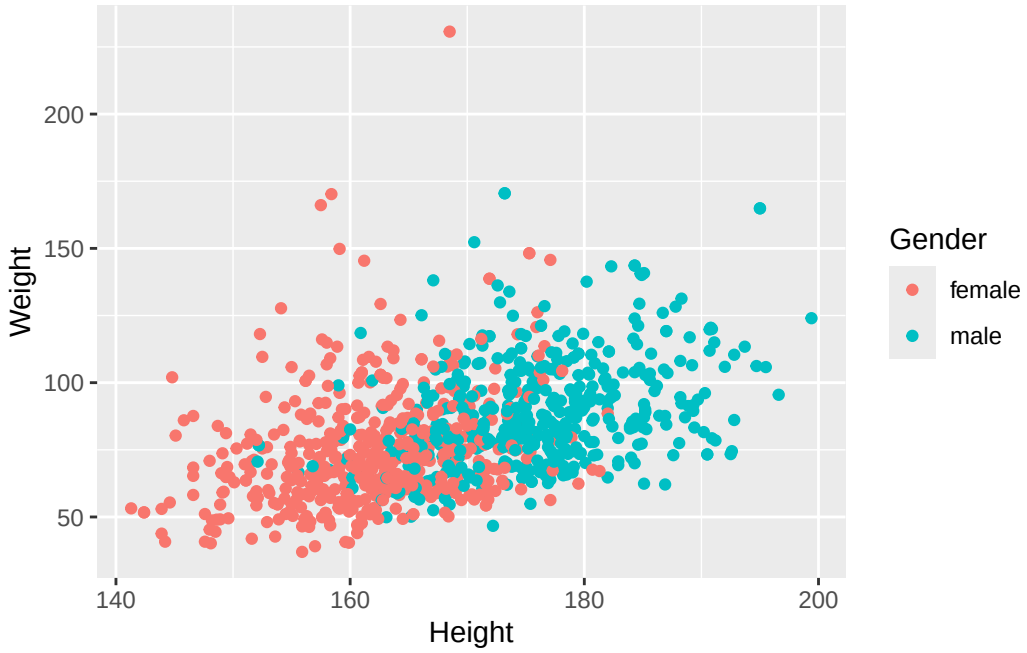
```r
# gjennomfører regresjon
reg2 <- lm(BPSys1 ~ BMI + Age, data = NHANES20)

summary(reg2)
```

```
Call:
lm(formula = BPSys1 ~ BMI + Age, data = NHANES20)

Residuals:
   Min     1Q Median     3Q    Max
-63.65  -9.72  -0.89   8.31 105.62

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  93.3718     0.9822   95.06   <2e-16 ***
BMI           0.2767     0.0289    9.59   <2e-16 ***
Age           0.4392     0.0113   39.00   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.5 on 6663 degrees of freedom
  (569 observations deleted due to missingness)
Multiple R-squared:  0.198, Adjusted R-squared:  0.198
F-statistic:  825 on 2 and 6663 DF,  p-value: <2e-16
```

Spørsmål: Er det forskjell i vekt mellom menn og kvinner?

Når vi skal se på forskjellen i vekt mellom menn og kvinner. Vi starter med å visualser dataen.

```
set.seed(1337)
NHANES20 %>%
  slice_sample(n=1000) %>%
  ggplot(aes(x=Height, y=Weight, color=Gender))+
  geom_point()
```



Vi ser at menn er høyere og tyngere, men det er vanskelig å si om det er noe forskjell mellom kjønnene. Kan videre see på gjennomsnitt vekten.

```
NHANES20 %>%
  group_by(Gender) %>%
  summarise(gjenn_vekt = mean(Weight, na.rm = TRUE))
```

```
# A tibble: 2 x 2
  Gender gjenn_vekt
  <fct>       <dbl>
1 female       75.5
2 male         89.2
```

Først ønsker vi å undersøke om det er en signifikant forskjell i vekt mellom menn og kvinner. Dette kan vi gjøre ved å kjøre en enkel lineær regresjon med vekt som avhengig variabel og kjønn som forklarende variabel.

$\text{Vekt} = \alpha + \beta_1 \cdot \text{Kjønn} + \epsilon$

```r
# Enkel lineær regresjon: vekt som funksjon av kjønn
model1 <- lm(Weight ~ Gender, data = NHANES20)
summary(model1)
```

```
Call:
lm(formula = Weight ~ Gender, data = NHANES20)

Residuals:
   Min     1Q Median     3Q    Max
 -43.0  -14.0   -3.6   10.9  155.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   75.496      0.332   227.3   <2e-16 ***
Gendermale    13.705      0.474    28.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.1 on 7176 degrees of freedom
  (57 observations deleted due to missingness)
Multiple R-squared:  0.104, Adjusted R-squared:  0.104
F-statistic:  835 on 1 and 7176 DF,  p-value: <2e-16
```

- Koeffisienten for ´Gendermale´ vil vise om at menn har signifikant høyeree vekt enn kvinner. Altså menn veier, 13.7 kg mer enn kvinner.
- Siden p-verdien for kjønn er lav (< 0.05), betyr det at det er en signifikant forskjell i vekt mellom menn og kvinner.

Her er veier kvinner på gjennomsnitt 75.5 kg mens menn veier 75.5+13.7=89.2 kg.

Fjerne skjæringspuntet

Hvis vi legger til "+0" i regresjon fjerner vi intercept. Dette gjør at kan lette lese de kategoriske variablene.

$\text{Vekt} = \beta_1 \cdot \text{Kjønn} + \epsilon$

For å estimere dette bruker vi `lm` funksjonen i R.

```
# Enkel lineær regresjon: vekt som funksjon av kjønn uten skjæringspunkt
model1 <- lm(Weight ~ Gender+0, data = NHANES20)
summary(model1)
```

```
Call:
lm(formula = Weight ~ Gender + 0, data = NHANES20)

Residuals:
   Min     1Q Median     3Q    Max
 -43.0  -14.0   -3.6   10.9  155.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
Genderfemale   75.496      0.332     227   <2e-16 ***
Gendermale     89.201      0.339     263   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.1 on 7176 degrees of freedom
  (57 observations deleted due to missingness)
Multiple R-squared:  0.944, Adjusted R-squared:  0.944
F-statistic: 6.05e+04 on 2 and 7176 DF,  p-value: <2e-16
```

Her ser vi at vi får samme verdier som tidligeremen nå regnet R dem ut for oss.

Samsvar mellom vekt og høyde

Neste trinn er å undersøke om det er et samsvar mellom vekt og høyde. Vi forventer at høyere personer veier mer, uavhengig av kjønn.

$\text{Vekt} = \alpha + \beta_1 \cdot \text{Høyde} + \epsilon$

```
# Enkel lineær regresjon: vekt som funksjon av høyde
model2 <- lm(Weight ~ Height, data = NHANES20)
summary(model2)
```

```
Call:
```

```
lm(formula = Weight ~ Height, data = NHANES20)

Residuals:
   Min     1Q Median     3Q    Max
-41.56 -13.10  -2.82   9.62 148.74

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -80.6294     3.7350   -21.6   <2e-16 ***
Height        0.9649     0.0221    43.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.9 on 7170 degrees of freedom
  (63 observations deleted due to missingness)
Multiple R-squared:  0.21,  Adjusted R-squared:  0.21
F-statistic: 1.91e+03 on 1 and 7170 DF,  p-value: <2e-16
```

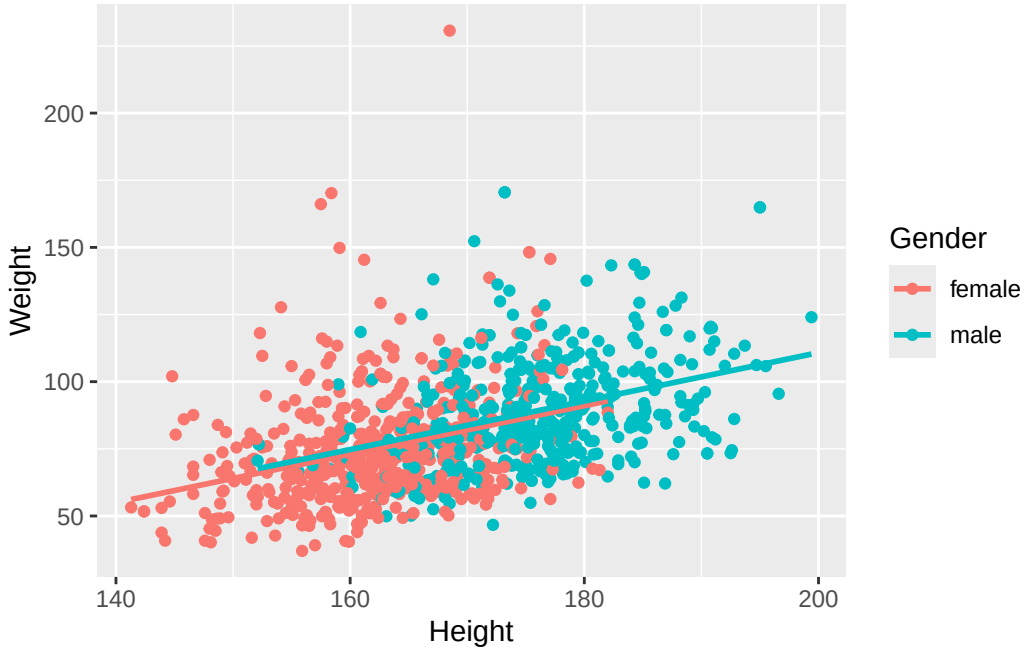
Tolkning:

- Koeffisienten for høyde viser hvor mye vi forventer at vekten øker for hver ekstra centimeter i høyde. Eller vi forventer 0.965 ekstra kg per ekstra cm.
- Siden p-verdien for høyde er lav (< 0.05), betyr det at høyde er en god prediktor for vekt.

Kontroll for høyde: Er det fortsatt forskjell i vekt mellom kjønnene?

Til slutt undersøker vi om kjønn fortsatt er signifikant for å forklare vekt når vi kontrollerer for høyde. Først plotter vi dataen og sammenlikner trendlinjene.

```r
set.seed(1337)
NHANES20 %>%
  slice_sample(n=1000) %>%
# Scatter plott av vekt og kroppshøyde og kjønn
ggplot( aes(x=Height, y=Weight, color=Gender))+
  geom_point() +
  # denne koden lager trendlinjer for hver gruppe
  geom_parallel_slopes(se = FALSE) # trenger pakken "moderndive"
```



Det er vanskelig å se om der noe er noe høyde forskjell mellom linjene. Vi kjører en multippel regresjon hvor både høyde og kjønn er forklarende variabler.

$$\text{Vekt} = \alpha + \beta_1 \cdot \text{Høyde} + \beta_2 \cdot \text{Kjønn} + \epsilon$$

```r
# Enkel lineær regresjon: vekt som funksjon av høyde
model3 <- lm(Weight ~ Height+Gender, data = NHANES20)
summary(model3)
```

```
Call:
lm(formula = Weight ~ Height + Gender, data = NHANES20)

Residuals:
   Min     1Q Median     3Q    Max
-41.52 -13.01  -2.93   9.63 149.14

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -76.3000     4.8959  -15.58   <2e-16 ***
Height        0.9369     0.0302   31.07   <2e-16 ***
Gendermale    0.8316     0.6081    1.37     0.17
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.9 on 7169 degrees of freedom
  (63 observations deleted due to missingness)
Multiple R-squared:  0.21,  Adjusted R-squared:  0.21
F-statistic:  955 on 2 and 7169 DF,  p-value: <2e-16
```

Tolkning:

- Her ser vi på koeffisientene for både høyde og kjønn.
- Koeffisienten for kjønn blir ikke-signifikant når vi kontrollerer for høyde, betyr det at forskjellen i vekt mellom menn og kvinner hovedsakelig skyldes høyde, ikke kjønn i seg selv. En mann og kvinne av samme høyde er forventet til å veie det samme.
- Dette kan tolkes som at høyere personer, som ofte er menn, veier mer, men når vi holder høyden konstant, er ikke kjønn lenger en viktig faktor for å forklare vekt.

Vi kan se at ved å legge til kjønn som ikke er statistisk signifikant, andres ikke *Residual standard error* eller R^2 seg for modellen uten kjønn.

Konklusjon:

Kun høyde og ikke kjønn påvirker vekt. Kjønn påvirker på gjennomsnittet kun 856 gram, mens høyden teller for alt annet forskjelli i vekt.

Et viktig poeng her er at vi *antar* at menn og kvinner får ekstra vekt i samme rate for en ekstra cm med høyde. Det er ingen forskjell mellom kjønn. Dette skal vi se næremere på under.

Prediksjon

Nå som vi har bygd en regresjonsmodell basert på høyde og kjønn, kan vi også bruke modellen til å predikere vekten til en person, gitt deres høyde og kjønn. Prediksjon i en regresjonsmodell fungerer på samme måte som når vi finner forventningen i et sannsynlighetseksperiment. For eksempel, når vi kaster en terning, er forventningen for en enkelt kast 3.5, fordi dette er gjennomsnittsverdien av alle mulige utfall.

I vår modell kan vi finne den forventede gjennomsnittsvekten for en person basert på høyde og kjønn ved å sette inn verdiene i modellen. Dette gir oss en estimert vekt, som er vår beste gjetning basert på de observerte dataene.

Vi kan bruke funksjonen **`makeFun`** fra `mosaic`-pakken for å lage en enkel funksjon som gir oss prediksjoner.

Eksempel:

La oss si vi ønsker å predikere vekten for en mann som er 174 cm høy. Ved å bruke modellen vår kan vi estimere den forventede vekten for en slik person.

R-kode for prediksjon:

```
# Lager en funksjon til å gi prediksjoner basert på modellen
library(mosaic)
Funksjon <- makeFun(model3)  # model3 inneholder høyde og kjønn

# Prediksjon for en mann som er 174 cm høy
Funksjon(174, "male")
```

```
    1
87.54
```

Tolkning av resultatet:

Hvis vi trekker mange menn som er 174 cm høye, forventer vi at deres gjennomsnittsvekt skal være rundt **87.54 kg** (avhengig av koeffisientene i modellen). Dette betyr at vår modell predikerer at menn med denne høyden, i gjennomsnitt, vil veie omtrent 87.54 kg.

Prediksjonen er basert på gjennomsnittsverdier i datasettet, så enkeltindivider kan naturligvis ha avvik fra denne prediksjonen, men over mange individer med samme høyde og kjønn, vil gjennomsnittsvekten nærme seg dette estimatet.

Oppgave

Prediker prisen til en bolig i `saratoga_houses` som uavhengig variabel bruk price (pris). Tolk koeffisientene når du bruker de uavhengige variablene:

```
glimpse(saratoga_houses )
```

```
Rows: 1,057
Columns: 8
$ price       <dbl> 142212, 134865, 118007, 138297, 129470, 206512, 50709, 108~
$ living_area <dbl> 1982, 1676, 1694, 1800, 2088, 1456, 960, 1464, 1216, 1632,~
$ bathrooms   <dbl> 1.0, 1.5, 2.0, 1.0, 1.0, 2.0, 1.5, 1.0, 1.0, 1.5, 2.5, 2.5~
$ bedrooms    <dbl> 3, 3, 3, 2, 3, 3, 2, 2, 2, 3, 3, 3, 3, 2, 3, 3, 4, 3, 4, 4~
$ fireplaces  <dbl> 0, 1, 1, 2, 1, 0, 0, 0, 0, 0, 2, 1, 0, 0, 1, 1, 0, 1, 1, 1~
$ lot_size    <dbl> 2.00, 0.38, 0.96, 0.48, 1.84, 0.98, NA, 0.11, 0.61, 0.23, ~
$ age         <dbl> 133, 14, 15, 49, 29, 10, 12, 87, 101, 14, 9, 0, 16, 17, 0,~
$ fireplace   <lgl> FALSE, TRUE, TRUE, TRUE, TRUE, FALSE, FALSE, FALSE, FALSE,~
```

1. living_area (størrelse) & fireplace (ildsted)

```r
summary(
  lm(price ~ living_area + fireplace, data = saratoga_houses)
  )
```

```
Call:
lm(formula = price ~ living_area + fireplace, data = saratoga_houses)

Residuals:
    Min      1Q  Median      3Q     Max
-262826  -24856   -4102   16992  400822

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    7884.18    4496.77    1.75  0.07984 .
living_area      83.63       2.64   31.66  < 2e-16 ***
fireplaceTRUE 13194.87    3565.63    3.70  0.00023 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50000 on 1054 degrees of freedom
Multiple R-squared:  0.581, Adjusted R-squared:  0.58
F-statistic:  731 on 2 and 1054 DF,  p-value: <2e-16
```

2. fireplace (ildsted) & bathrooms (bad)
3. living_area (størrelse), fireplace (ildsted) & bathrooms (bad)
4. living_area (størrelse), fireplace (ildsted), bathrooms (bad), bedrooms (soverom), lot_size (tomt størelse), age (alder)

Interaksjonseffekter i regresjonsmodeller

I regresjonsanalyse brukes **interaksjonseffekter** for å undersøke hvordan effekten av én uavhengig variabel på den avhengige variabelen avhenger av nivået på en annen uavhengig variabel. Interaksjonseffekter lar oss modellere komplekse sammenhenger der to variabler påvirker hverandre i måten de påvirker utfallet.

Hva er en interaksjonseffekt?

Når vi inkluderer interaksjonseffekter i en regresjonsmodell, tester vi om to variabler i kombinasjon har en annen effekt på den avhengige variabelen enn de ville hatt hver for seg. Interaksjonseffekter kan hjelpe oss å forstå om forholdet mellom den avhengige og en uavhengig variabel varierer avhengig av nivået på en annen uavhengig variabel.

Eksempel på interaksjon: Høyde og kjønn

I regresjonen du kjørte, ser vi på forholdet mellom **høyde** og **kjønn** i forklaringen av **vekt**. Vi er interessert i å undersøke om effekten av høyde på vekt er forskjellig for menn og kvinner, altså om det finnes en interaksjon mellom høyde og kjønn.

Modellen er som følger:

$$\text{Weight} = \alpha + \beta_1 \cdot \text{Height} + \beta_2 \cdot \text{Gender} + \beta_3 \cdot (\text{Height} \times \text{Gender}) + \epsilon$$

```r
# Enkel lineær regresjon: vekt som funksjon av høyde
model3 <- lm(Weight ~ Height+Gender+Height*Gender, data = NHANES20)
summary(model3)
```

```
Call:
lm(formula = Weight ~ Height + Gender + Height * Gender, data = NHANES20)

Residuals:
   Min     1Q Median     3Q    Max
-42.79 -13.05  -3.03   9.57 150.00

Coefficients:
                  Estimate Std. Error t value          Pr(>|t|)
(Intercept)       -54.8076     6.9291   -7.91 0.000000000000003 ***
Height              0.8042     0.0427   18.83           < 2e-16 ***
Gendermale        -43.6937    10.1888   -4.29 0.000018233107397 ***
Height:Gendermale   0.2637     0.0602    4.38 0.000012158668337 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.8 on 7168 degrees of freedom
  (63 observations deleted due to missingness)
Multiple R-squared:  0.212, Adjusted R-squared:  0.212
F-statistic:  645 on 3 and 7168 DF,  p-value: <2e-16
```

- α: Interceptet, som er den forventede vekten for kvinner (referansekategorien) ved 0 cm høyde (ikke realistisk, men nødvendig for beregningen).
- β_1: Koeffisienten for høyde som viser effekten av høyde på vekt for kvinner.
- β_2: Koeffisienten for kjønn, som viser den gjennomsnittlige forskjellen i vekt mellom menn og kvinner når høyden er null.
- β_3: Interaksjonseffekten mellom høyde og kjønn, som viser hvordan effekten av høyde på vekt endres mellom menn og kvinner.

Her ser vi at *Residual standard error* er lavere og R^2 er høyere.

Tolkning av regresjonsresultatet:

Fra resultatene av regresjonen din ser vi følgende:

- **Intercept (-54.81)**: Dette er den forventede vekten for en kvinne (referansekategorien for kjønn) ved 0 cm høyde. Selv om dette tallet ikke gir praktisk mening, brukes det som et grunnlag for beregningene i modellen.
- **Høyde (0.80)**: Koeffisienten for høyde er 0.80. Dette betyr at for hver ekstra centimeter i høyde, forventes vekten for **kvinner** å øke med 0.80 kg. Dette er hovedvirkningen av høyde på vekt for kvinner.
- **Kjønn (Gendermale, -43.69)**: Denne koeffisienten viser at menn, ved null høyde (som er en teoretisk verdi), forventes å veie 43.69 kg mindre enn kvinner. Denne tolkningen er også teoretisk, ettersom ingen har en høyde på null, men det gir informasjon om hvordan modellen justerer for kjønn ved lave høyder.
- **Interaksjonseffekt mellom høyde og kjønn (0.26)**: Denne koeffisienten viser at effekten av høyde på vekt er **0.26 kg større for menn enn for kvinner** for hver ekstra centimeter i høyde. Med andre ord, mens vekten for kvinner øker med 0.80 kg per ekstra cm høyde, øker vekten for menn med $0.80 + 0.26 = 1.06$ kg per ekstra cm høyde.

Samlet Tolkning:

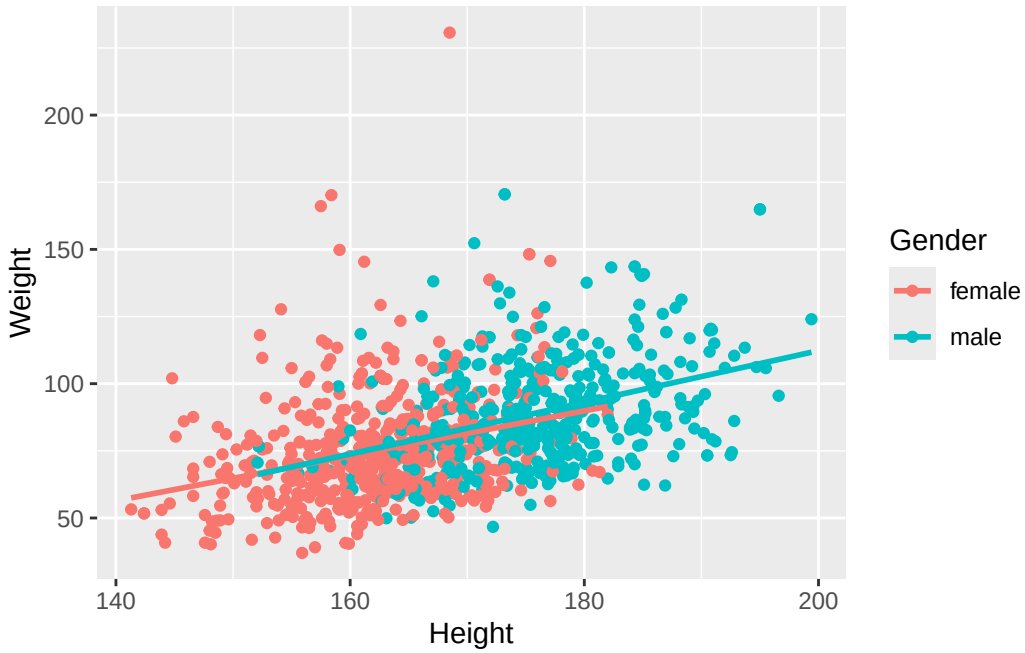
- For **kvinner**, øker vekten med 0.80 kg for hver ekstra centimeter høyde.
- For **menn**, øker vekten med 1.06 kg for hver ekstra centimeter høyde (0.80 kg for høydeeffekten + 0.26 kg for interaksjonseffekten).
- **Forskjellen mellom menn og kvinner** i vekt øker med høyden. Selv om menn ved lav høyde (når høyden nærmer seg null) kan veie mindre enn kvinner (som indikert av koeffisienten for kjønn), vil menn etter hvert som høyden øker veie betydelig mer enn kvinner med tilsvarende høyde.

Dette viser at kjønn modifiserer effekten av høyde på vekt. Modellen tar hensyn til hvordan forholdet mellom høyde og vekt er forskjellig for menn og kvinner. Dette er en klassisk måte å bruke interaksjonseffekter på for å modellere komplekse forhold mellom variabler.

Visualisering av interaksjonseffekter

For å få en bedre forståelse av hvordan interaksjonen mellom høyde og kjønn påvirker vekt, kan det være nyttig å visualisere interaksjonseffektene. Dette kan gjøres ved å lage et plott som viser hvordan vekten endres med høyden for både menn og kvinner.

```
set.seed(1337)
NHANES20 %>%
  slice_sample(n=1000) %>%
# Scatter plott av vekt og kroppshøyde og kjønn
ggplot( aes(x=Height, y=Weight, color=Gender))+
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



Dette vil gi deg et plott som viser hvordan vekten øker med høyde for både menn og kvinner. Hvis de to linjene har forskjellige helninger (som vi forventer basert på interaksjonseffekten), viser dette at vekten øker raskere med høyden for menn enn for kvinner.

Konklusjon:

Interaksjonseffekter er et kraftig verktøy i regresjonsanalyse for å modellere komplekse forhold mellom variabler. I denne analysen ser vi at effekten av høyde på vekt er forskjellig for menn og kvinner, og denne forskjellen fanges opp ved å inkludere en interaksjonseffekt mellom høyde og kjønn. Dette gir oss en mer presis modell som kan forklare variasjoner i vekt bedre enn en modell uten interaksjonseffekt.

R^2 og justert R^2 i regresjonsanalyse

Når vi bygger en regresjonsmodell, ønsker vi å vite hvor godt modellen passer dataene, og hvor mye av variasjonen i den avhengige variabelen (for eksempel vekt) som kan forklares av de uavhengige variablene (som høyde og kjønn). To vanlige måter å måle dette på er $\mathbf{R^2}$ og **justert** $\mathbf{R^2}$.

Hva er R^2 (R-squared)?

R^2 forteller oss hvor mye av variasjonen i den avhengige variabelen som kan forklares av de uavhengige variablene i modellen. Verdien av R^2 ligger mellom 0 og 1:

- En R^2-verdi nær 0 betyr at modellen forklarer svært lite av variasjonen.
- En R^2-verdi nær 1 betyr at modellen forklarer nesten all variasjon i den avhengige variabelen.

Eksempel fra din modell:

I din modell, der du ser på vekt som avhengig variabel og høyde og kjønn som forklarende variabler, har vi en R^2-verdi på 0.212. Dette betyr at omtrent **21.2%** av variasjonen i vekt kan forklares av høyde og kjønn i denne modellen. Resten av variasjonen skyldes faktorer som ikke er inkludert i modellen.

Hva er justert R^2 (adjusted R^2)?

Justert R^2 er en litt mer avansert versjon av R^2, som tar hensyn til hvor mange variabler du har lagt til i modellen. Når vi legger til flere variabler, vil R^2 vanligvis øke, selv om de nye variablene ikke egentlig bidrar til å forklare variasjonen. Justert R^2 korrigerer for dette, slik at du ikke får en falsk følelse av at modellen er bedre bare fordi du har lagt til flere variabler.

Justert R^2 kan være litt lavere enn R^2 hvis de nye variablene ikke bidrar mye til modellen.

Når skal du bruke justert R^2?

- Hvis du legger til en variabel, og **R^2 går opp**, men **justert R^2 går ned**, betyr det at den nye variabelen ikke bidrar til modellen på en god måte. I slike tilfeller bør du vurdere å fjerne denne variabelen, siden den ikke gir nyttig informasjon.

Eksempel

La oss si du legger til en ny variabel som for eksempel antall søsken til hver person i modellen din. Hvis R^2 går litt opp, men justert R^2 går ned, betyr det at antall søsken ikke er en nyttig variabel for å forklare variasjonen i vekt. Det er da en indikator på at variabelen kanskje bør fjernes fra modellen.

Hvordan tolke resultatene fra din modell

I din modell har både R^2 og justert R^2 en verdi på 0.212, noe som betyr at høyde og kjønn er viktige faktorer for å forklare variasjonen i vekt, og at antall variabler i modellen er passende i forhold til antall observasjoner. Hadde du lagt til en irrelevant variabel, kunne justert R^2 ha falt mens R^2 fortsatt gikk opp.

Oppsummering:

- **R^2** forteller deg hvor mye av variasjonen i den avhengige variabelen (vekt) som kan forklares av modellen. En høy R^2-verdi betyr at modellen forklarer mye av variasjonen.
- **Justert R^2** tar hensyn til antall variabler i modellen og hjelper deg med å avgjøre om det er lurt å legge til flere variabler. Hvis justert R^2 går ned når du legger til en ny variabel, bør du vurdere å fjerne den.

Ved å bruke både R^2 og justert R^2 kan du sørge for at modellen din forklarer variasjonen i dataene på en effektiv måte, uten å overkomplikere modellen med unødvendige variabler.

Korrelasjon betyr ikke nødvendigvis kausalitet

Når vi ser på sammenhenger i data, som i regresjonsmodeller, kan vi ofte finne at to variabler ser ut til å være relatert til hverandre. Dette kalles **korrelasjon** – det betyr at når én variabel endres, har den en tendens til å følges av en endring i den andre variabelen. Men det er viktig å huske på at **korrelasjon ikke nødvendigvis betyr at den ene variabelen forårsaker endringen i den andre**. Dette er et vanlig misforståelse i dataanalyse.

Hva betyr det?

Bare fordi to variabler ser ut til å bevege seg sammen, betyr det ikke at den ene forårsaker den andre. Det kan være flere andre grunner til at de to variablene er relatert, for eksempel:

- **Tredjevariabler**: Det kan være en tredje, skjult variabel som påvirker begge de to variablene du ser på, og skaper en tilsynelatende sammenheng. For eksempel kan det se ut som at iskremsalg og drukninger er korrelert fordi begge øker om sommeren, men det er temperaturen som er den skjulte faktoren her.
- **Tilfeldighet**: Noen ganger kan korrelasjoner oppstå ved en ren tilfeldighet. Hvis du har et stort datasett med mange variabler, er det nesten garantert at noen vil se ut til å være korrelert, selv om det ikke er noen reell sammenheng.

Eksempel i konteksten av regresjon

I regresjonsmodellen din kan det for eksempel se ut som at kjønn og høyde har en sterk sammenheng med vekt, noe som kan være intuitivt forståelig. Men det betyr ikke nødvendigvis at kjønn eller høyde direkte forårsaker vektendringer – det kan være mange andre biologiske, genetiske eller miljømessige faktorer som også spiller inn. Modellene våre kan vise oss sammenhenger, men de kan ikke alltid bekrefte årsak-virkning-forhold uten videre bevis.

Hvordan skille mellom korrelasjon og kausalitet?

For å kunne si at en variabel forårsaker en annen, må vi kunne vise at:

1. **Det er en sammenheng** mellom variablene (korrelasjon).
2. **Årsaken skjer før effekten**.
3. **Sammenhengen er ikke spuriøs** (den skyldes ikke en skjult tredjevariabel eller tilfeldighet).

Det kreves vanligvis eksperimentell design eller dypere analyser for å kunne trekke konklusjoner om kausalitet.

Konklusjon

Selv om regresjonsmodeller og korrelasjoner kan hjelpe oss å finne mønstre og sammenhenger i data, er det viktig å huske at de ikke nødvendigvis viser oss årsak-virkning. Vi må alltid være forsiktige med å trekke konklusjoner om kausalitet uten videre analyser eller eksperimenter.

Boligpriser med interaksjoneffekt

Er det mulig at boareale og antall soverom påvikrker hverandre?

```
summary(
  lm(price ~ living_area + bedrooms + fireplace, data = saratoga_houses)
  )
```

```
Call:
lm(formula = price ~ living_area + bedrooms + fireplace, data = saratoga_houses)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-268380  -23848   -3880   17838  397774

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   21129.28    6777.09    3.12  0.00187 **
living_area      89.06       3.36   26.52  < 2e-16 ***
bedrooms      -7216.51    2768.46   -2.61  0.00927 **
fireplaceTRUE 12949.61    3557.12    3.64  0.00029 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49900 on 1053 degrees of freedom
Multiple R-squared:  0.584, Adjusted R-squared:  0.583
F-statistic:  492 on 3 and 1053 DF,  p-value: <2e-16
```

```
summary(
  lm(price ~ living_area * bedrooms + fireplace, data = saratoga_houses)
  )
```

```
Call:
lm(formula = price ~ living_area * bedrooms + fireplace, data = saratoga_houses)

Residuals:
    Min      1Q  Median      3Q     Max
-272561  -24036   -3948   17992  389881

Coefficients:
                     Estimate Std. Error t value   Pr(>|t|)
(Intercept)          61330.91   19641.23    3.12    0.00184 **
living_area             63.26      12.30    5.14 0.00000032 ***
bedrooms             -18809.52    5992.66   -3.14    0.00174 **
fireplaceTRUE         13793.49    3571.83    3.86    0.00012 ***
living_area:bedrooms      7.05       3.24    2.18    0.02946 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49800 on 1052 degrees of freedom
Multiple R-squared:  0.586, Adjusted R-squared:  0.584
F-statistic:  372 on 4 and 1052 DF,  p-value: <2e-16
```

Koeffisientene:

1. **Intercept (61330.91)**:
    - Dette er den forventede prisen når alle forklaringsvariablene er lik null. Siden både boareal og antall soverom ikke kan være null i virkeligheten, har denne koeffisienten liten praktisk betydning. Den gir oss bare et startpunkt for prediksjonen i modellen.
2. **Living_area (63.26)**:
    - For hver ekstra kvadratfot i boareal, forventes boligprisen å øke med **63.26 dollar**, gitt at antall soverom og peis holdes konstant. Denne koeffisienten er sterkt signifikant (p-verdi = 0.00000032), noe som betyr at boareal er en viktig prediktor for boligpris.
3. **Bedrooms (-18809.52)**:
    - Koeffisienten for antall soverom er **-18809.52**, noe som betyr at økning i antall soverom fører til en reduksjon i boligprisen med 18809.52 dollar, alt annet likt. Dette virker kontraintuitivt, men det kan forklares ved at større antall soverom ofte er assosiert med mindre boareal per soverom, noe som senker prisen. Denne effekten er også sterkt signifikant (p-verdi = 0.00174).
4. **FireplaceTRUE (13793.49)**:
    - Hvis huset har peis (TRUE), øker boligprisen med **13793.49 dollar** sammenlignet med et hus uten peis, alt annet likt. Denne effekten er signifikant (p-verdi = 0.00012), så peis er en viktig faktor i prissettingen av husene.
5. **Living_area:bedrooms (7.05)**:
    - Denne interaksjonseffekten på **7.05** betyr at effekten av boareal på boligpris endres med **7.05 dollar** for hver ekstra soverom. Dette innebærer at flere soverom reduserer den negative effekten av boareal, noe som gir en mer nyansert forståelse av hvordan boareal og soverom sammen påvirker prisen. Denne effekten er signifikant (p-verdi = 0.02946).

Oppgaver

Se på datasettet `house_prices`, hva tror du er viktig i å påvirke pris? Gjennomfør en multippel regresjon (uten interaksjon effekter) med variblene du tror er viktig å ha med og tolk modellen.

Gjennomfør en ny regresjon men denne gangen legg med minst en interaksjon effekt.