En regresjonslinje kan beskrives med formelen $y=m + k\times x$. Vi kjenner denne formelen fra matematikken og vi kan lese den som at $y$ er lik skjæringspunktet ($m$) pluss $k$ (stigningstall) enheter per hver enhets endring i $x$. I statistikken bruker man ofte andre tegn for å beskrive skjæringspunkt og stigningstall. Den samme ekvasjonen kan se ut slik i statistikkboken

$$y=\beta\_0 + \beta\_1 \times x$$

En regresjonsmodell fungerer best der hvor vi faktisk har data. En enkel regresjonsmodell begrenses også til rette linjer. Dette gjør at resultatene fra en slik analyse bør behandles med skepsis når vi kan gjøre antagelser om en ikke rett sammenheng og når modellen brukes for å predikere utenfor variasjonsvidden til dataene som bruktes til å lage modellen.

I eksemplet over så vi en hvordan vi kan lage en matematisk modell for sammenhengen mellom to kontinuerlige variabler. Modellen gir oss muligheter til å predikere en avhengig variabel ved hjelp av en uavhengige variabel. Prediksjonen antyder at vi ser på sammenhengen som at den ene variablene (uavhengig) påvirker den andre (avhengig). Sammenhengen beskrives også med enheter som vi finner igjen i dataene, for eksempel så gav 1 cm forandring i høyde en estimert forskjell i vekt på `r round(coef(mod)[2], 2)` kg. Regresjonsanalysen gir oss altså mye informasjon som kan brukes til flere formål. Men modellen kan også forenkles.

Sammenhengen mellom to kontinuerlige variabler kan beskrives som en korrelasjonskoeffisient. Denne korrelasjonskoeffisienten beskriver sammenhengen mellom to variabler som et tall mellom -1 og +1 hvor estimat som nærmer seg -1 eller +1 indikerer en sterk korrelasjon og estimat som nærmer seg 0 indikerer en svak eller ingen korrelasjon (eller sammenheng).

Den kanskje vanligste måten å beskrive en korrelasjon på er ved hjelp av Pearson’s korrelasjonskoeffisient. Denne gis vanligvis symbolet \*R\* og noen eksempler finnes å se i @fig-korr.

Korrelasjonskoeffisienten påvirkes ikke av rekkefølgen som vi gir variablene til dataprogrammet noe som er viktig for å tolke en regresjonsanalyse på en korrekt måte. Korrelasjonsanalysen og regresjonsanalysen påvirkes begge av ekstreme verdier og ikke linjere forhold mellom variabler.

Avvik fra antagelser om dataene kan gi feilaktige tolkninger av en regresjonskoeffisient

Ikke-lineær

En linje

Vi kan enkelt omformulere regresjonsmodellen gjennom å bruke en nominal eller ordinal variabel som uavhengig variabel. La oss si at vi ønsker å estimere sammenheng mellom alderskategoriene \*barn\*/\*voksen\* og vekt. I tabellen under (@tbl-simpsons) har vi identifisert voksne og barn, og vi uttrykker dette som to forskjellige \*indikatorvariabler\*, også kallet \*dummyvariabler\*. En \*dummyvariabel\* kan ta to verdier, 0 og 1, og her kan vi forstå den ene som, voksen ja = 1/nei = 0, og den andre som, barn ja = 1/nei = 0. Vi trenger ikke begge variablene, men kodingen avgjør hvordan i tolker resultatene fra en regresjonsmodell.