I regresjonsmodellene som ble presentert i tidligere avsnitt antok vi for eksempel at feilleddet var symmetrisk fordelt kring 0. Dette er en gyldig antagelse når vi prøver å forklare en nærmest kontinuerlig avhengig variabel som alder eller treningstid. Når den avhengige variabelen kan ta enten verdien 0 (f.eks. ikke medlem i et idrettslag) eller 1 (medlem i et idrettslag) kan vi ikke si att antakelsen stemmer om et normalfordelt feilledd stemmer. En binær variabel (0/1) kan for eksempel beskrive hvis en person har en sykdom eller er død, hvis et fotball-lag vinner en kamp eller bruk av narkotiske stoffer. Når vi ønsker å forstå disse variablene ved hjelp av en regresjonsmodell må vi forandre modellen for å unngå å bli lurt av dataene.

Alder kan tenkes være en variabel som påvirker (direkte eller indirekte) forekomst av hjertesykdom[^hosmer]. Når vi skal prøve oss på å beskrive sammenhengen mellom disse to variablene kan vi starte med å gjøre en figur hvor alder settes på x-aksel og sykdomsstatus (forekomst av hjertesykdom = 1, ikke hjertesykdom = 0) settes på y-aksel. For å estimere sammenhengen legger vi også inn en regresjonslinje (@fig-chd1). Modellen er ikke en god nok beskrivelse av dataene da den mener at færre en null individer under 25 har hjertesykdom. Vi kan gjøre bedre.

[^hosmer]: Eksemplet kommer fra @hosmer\_2013.

Som et neste forsøk å beskrive dataene kan vi beregne andelen med hjertesykdom i hver av noen aldersgrupper. Aldersgruppene som går fra 19-29, 29-39 osv. brukes for å regne ut andelen med hjertesykdom per gruppe. Andelen med hjertesykdom øker med økende alder, men forholdet er ikke en rett linje. Når prosentandelen nærmer seg 0 eller 1 flater kurven ut (@fig-chd2). Vi kan gjøre det enda bedre.

Den vanligste modellen for å beskrive sammenhenger mellom uavhengige variabler og en binær avhengig variabel er den logistiske regresjonsmodellen (@hosmer\_2013). Denne modellen er en av flere modeller i kategorien \*Generaliserte lineære modeller\*. Disse modellene kombinerer regresjonsmodellens bruk av uavhengige variabler for å beskrive en avhengig variabel med en «linkfunksjon» som transformere resultatet av regresjonsmodellen til en skala hvor den avhengige variabelen kan beskrives som (en lineær) kombinasjon av uavhengige variabler. Linkfunksjonen for en logistisk regresjonsmodell er logit-linken. Denne funksjonen gjør at vi kan lage en modell som forteller oss om sannsynligheten for, i vårt eksempel, hjertesykdom.

Ratioen i odds for utfallet mellom $x = 0$ og $x = 1$ er altså et ratio av ratioer.