Arbetsgivarverket

ARBETSGIVARORGANISATIONEN FÖR STATLIGA ARBETSGIVARE

PM
2021-05-17

Metodbeskrivning, bilaga till Arbetsgivarverkets rapportering av löneskillnader mellan män och kvinnor i staten

Regressionsanalys

En regressionsanalys syftar till att undersöka om det förekommer några former av statistiska korrelationer eller samvariationer mellan två eller flera variabler. Utöver korrelationen kan regressionen användas för att studera en eller en grupp variablers påverkan på en annan variabel. Regressionsanalysen tillhör en av de vanligaste metoderna som tillämpas inom vetenskapliga studier av skillnader i utfall mellan olika individer eller sociala grupper.

Vår regressionsmodell

En regressionsmodell består av minst två variabler. Begreppet variabel är ett samlingsnamn för olika objekt, som i sin tur består av olika värden, som t.ex. lön eller kön. Vår basmodell har följande utseende:

$$log(Lön) = \beta_0 + \beta_1 kön + \varepsilon$$

I basmodellen ställer vi upp heltidslön som en funktion av kön. Med modellen vill vi studera om kön har en direkt påverkan på lön. Den variabel som är föremål för undersökningen, i detta fall lön, kallas normalt för beroende variabel. Variabeln som ska hjälpa oss att förklara lönen, i detta fall kön, kallas förklarande variabel. Det kan endast finnas en beroende variabel, medan det inte finns någon begränsning för antalet förklarande variabler.

I basmodellen visas lönen i form av logaritmen, log(lön). Genom att lönen transformeras till logaritmen kan vi med fördel redovisa löneskillnader mellan kvinnor och män i form av procentsatser istället för kronor.

Dessutom ger transformationen log(lön) en betydligt tydligare normalfördelad profil än lön i kronor, och detta gör log(lön) mer anpassad till regressionsanalysen.

Föremålen för regressionsanalysen är β-värden för förklarande variabler, eller koefficienter. Koefficienten β_1 från basmodellen signalerar könets påverkan på lönen. Om β_1 är negativ så betyder det att kvinnor tjänar mindre än män. Till koefficienten medföljer ett testvärde som signalerar om koefficienten är statistiskt signifikant. Icke-signifikant koefficient saknar statistisk relevans.

B_0 kallas konstant och kan likställas med medelvärde för den beroende variabeln. Regressionens ekvation avslutas med en felterm, ε, som symboliserar övriga faktorer eller effekter som inte fångas upp av förklarande variabler i modellen. Normalt studerar man varken konstanten eller feltermen vid regressionsanalyser. Dessutom vet vi, med hjälp av en lång rad tidigare vetenskapliga studier, att löner i högsta grad påverkas av andra faktorer än bara kön. I regressionen har vi utökat vår modell med ytterligare ett antal variabler, och dessa variabler kallas för kontrollvariabler, se ekvationen nedan.

$$log(lön) = \beta_0 + \beta_1\, kön + \beta_2\, arbetstidsomfattning + \beta_3\, ålder + \beta_4\, anställningsår + \beta_5\, chef + \beta_6\, arbetsområde + \beta_7\, grupperingsnivå + \beta_8\, cofog + \beta_9\, utbildningsinriktning + \beta_{10}\, utbildningsnivå + \beta_{11}\, erfarenhet + \varepsilon$$

I regel finns det inga begränsningar för hur många kontrollvariabler man kan ha i en regression. Dock kan man säkerställa att kontrollvariabler tillför relevant information till regressionen genom att, dels kontrollera att koefficienter till kontrollvariablerna är statistiskt signifikanta, och dels studera värdena på korrelationskoefficienten, R^2

R^2 får man efter varje regressionsberäkning och värdet visar hur mycket av variationen i den beroende variabeln som kan förklaras av de förklarande variablerna. Om R^2 är 0, då finns det ingen korrelation mellan beroende och förklarade variabler. Om R^2 är 1, då har vi en perfekt modell. R^2 tenderar att öka i takt med antalet förklarande variabler. Om antalet förklarande variabler ökar medan R^2 inte ökar eller minskar så kan det vara en indikation på att regressionen innehåller för många kontrollvariabler.

Tillsammans med kontrollvariablerna i regressionen mäter man det så kallade ceteris paribus, dvs. allt annat lika, på kön. Det innebär att vi får veta könets påverkan på lönen givet att personerna har samma arbetsuppgifter, arbetstidsomfattning och utbildning och så vidare.