Oaxaca-Blinder dekomposition på CLASS datasættet fra SASHELP.

Oaxaca-Blinder dekompositionen er en teknik, der vil forsøge at reducere forskellen mellem to grupper ved at inddrage forklarende variable.

Eksempelvis er der ofte forskel på mænd og kvinders løn og en del af denne forskel skyldes, at mænd har længere anciennitet, flere kurser etc. Ved at udføre klassisk regression på de to grupper kan man enkelt dekomponere forskellen mellem de grupper, i noget der skyldes de forklarende variable, og noget der skyldes resten, som så kaldes den uforklarede forskel. Den er så ofte den uforklarede forskel, som rapporteres gerne i procent.

Når man vurderer, om der er ligeløn mellem mænd og kvinder er det ofte Oaxaca-Blinder dekompositionen der bruges. Generelt tror jeg, at der er en uforklaret forskel på mænd og kvinders løn på ca. 10 %.

Nedenstående illustreres Oaxaca-Blinder dekompositionen på et lille overskueligt datasæt.

Et meget ”berømt” datasæt er CLASS som er hentet fra hjælpen i SAS. Det omhandler 19 unge mennesker, der har fået målt vægt, højde og alder.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Sex** | **Age** | **Height** | **Weight** |
| Alfred | M | 14 | 69,0 | 112,5 |
| Alice | F | 13 | 56,5 | 84,0 |
| Barbara | F | 13 | 65,3 | 98,0 |
| Carol | F | 14 | 62,8 | 102,5 |
| Henry | M | 14 | 63,5 | 102,5 |
| James | M | 12 | 57,3 | 83,0 |
| Jane | F | 12 | 59,8 | 84,5 |
| Janet | F | 15 | 62,5 | 112,5 |
| Jeffrey | M | 13 | 62,5 | 84,0 |
| John | M | 12 | 59,0 | 99,5 |
| Joyce | F | 11 | 51,3 | 50,5 |
| Judy | F | 14 | 64,3 | 90,0 |
| Louise | F | 12 | 56,3 | 77,0 |
| Mary | F | 15 | 66,5 | 112,0 |
| Philip | M | 16 | 72,0 | 150,0 |
| Robert | M | 12 | 64,8 | 128,0 |
| Ronald | M | 15 | 67,0 | 133,0 |
| Thomas | M | 11 | 57,5 | 85,0 |
| William | M | 15 | 66,5 | 112,0 |

Kilde: SAShelp

Gennemsnittene for de to køn er

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sex** | **Antal** | **Weight** | **Height** | **Age** |
| F | 9 | 90,11 | 60,59 | 13,22 |
| M | 10 | 108,95 | 63,91 | 13,40 |
| Forskel/gab |  | 18,84 | 3,32 | 0,18 |

Så drengene (M) er i gennemsnit højere, tungere og ældre end pigerne (F).

Gabet er forskellen mellem drenge og pigers vægt, som her bliver 18,84.

I første omgang fokuseres på vægt og højde. Man vil gerne forklare vægten som funktion af højden.

Så vi vurderer om vægten afhænger lineært af højden og det gøres for hvert køn for sig. Se nedenstående tegninger. Første tegning er for piger.





Resultaterne af de to traditionelle regressionsanalyser bliver:

|  |  |  |
| --- | --- | --- |
| **Sex** | **Intercept**  **(afskæring)** | **Slope**  **(hældning)** |
| F | -117,37 | 3,424 |
| M | -141,10 | 3,913 |

Bemærk at vi kan skrive gennemsnitsvægten som funktion af parametrene og gennemsnitshøjden

F: 90,11= -117,37 + 3,424\*60,59

M: 108,95= -141,10 + 3,913\*63,91

”Gabet” mellem drenge og piger bliver: 108,95-90,11= 18,8. Dette gab vil vi gerne reducere som følge af at drengene er højere end pigerne, vi har jo konstateret at højden er korreleret med vægten for begge køn. Så med lidt matematisk manipulation kan vi skrive gabet som:

(I notatet Oaxaca er det beskrevet mere matematisk).

Gab= 18,8= (-141,10 - -117,37) +60,59\*(3,91-3,42) + 3,91\*(63,91-60,59) =

5,8 + 13,0 = uforklaret + forklaret

Så af gabet på 18,8 kan de 13,0 forklares ved at højderne for de to grupper er forskellige, de resterende 5,8 er dermed uforklaret. Kigger man lidt mere matematisk på dette, så består den uforklarede del af forskellen mellem parameterestimater for de to grupper. Mens den forklarede del består af hældningskoefficienten (slope) for drenge ganget med forskellen på gruppernes højde.

Hvis man inddrager alder (age) i modellen dvs.

Model Weight = height age

Så bliver

Gabet =18,8 = 4,4 + 14,4.

Så inddragelse af alder øgede den forklarede del fra 13,0 til 14,4 af det samlede gab på 18,8.