UNDERVISNINGS
MINISTERIET
UNI•C – STYRELSEN
FOR IT OG LÆRING

Bilag 2: Den statistiske model

Den anvendte statistiske model er en såkaldt multilevel model:

$$y_{ij} = x_{ij}\beta + u_j + e_{ij}$$

hvor y_{ij} er karakteren for elev *i* på institution *j*, x_{ij} er elevens baggrundsvariable og β er de tilhørende parameterestimater, u_j er varianskomponenten svarende til variationen mellem skolerne, og e_{ij} er residualet svarende til variationen mellem eleverne på skolen.

Beregningerne er foretaget for såvel de bundne prøvefag som for prøvefag til udtræk (eksklusiv fransk, hvor relativt få elever aflægger prøve).

Ved anvendelse af modellen fås et estimat over, hvor stor en del af variationen mellem karaktererne der skyldes forskelle mellem eleverne, og hvor stor en del der skyldes forskelle mellem skolerne. Dette kaldes modellens forklaringsgrad.

Alle de socioøkonomiske baggrundsvariable beskrevet i Bilag 1: Baggrundsoplysninger medtages i modellen sammen med interaktionen mellem forældrenes uddannelse og elevens herkomst, interaktionen mellem elevens køn og herkomst samt andelen af indvandrere og efterkommere og andelen af elever med forældre med en mellemlang eller lang videregående uddannelse. Oplysningen om, hvorvidt eleven modtager specialundervisning, er først medtaget i modellerne vedrørende skoleåret 2012/2013 og frem.

Interaktionerne medtages for at undersøge, om effekten af forældrenes uddannelse og elevens køn afhænger af elevens herkomst.

Skoleniveauvariablen vedrørende andelen af indvandrere og efterkommere medtages for at undersøge, om der er en ekstra effekt af en høj andel af indvandrere og efterkommere på samme institution. Tilsvarende for andelen af elever med forældre, der har en mellemlang eller lang videregående uddannelse.

Øvrige interaktioner og skoleniveauvariable er ikke medtaget dels for ikke at øge modellernes kompleksitet, dels viser undersøgelser, at disse ikke øger modellernes forklaringskraft væsentligt.

Ikke alle elever har fyldestgørende oplysninger på de her benyttede baggrundsvariable. Der i disse tilfælde indsat en værdi, et såkaldt bedste gæt, på alle de steder, hvor oplysningerne er ukendte. For hver elev med ukendt baggrundsoplysning gættes 5 gange, således at modellen bliver estimeret på baggrund af en sandsynlig fordeling. Metoden kaldes 'multiple imputation'. Gættene baseres på fordelingen blandt elever med oplyst baggrundsvariabel, hvor der er taget hensyn til elevens køn, alder og herkomst. Den socioøkonomiske reference for hver elev er så et gennemsnit af de 5 beregnede værdier.

I den statistiske model for beregning af socioøkonomiske referencer for 3-års perioden medtages endvidere skoleåret som en forklarende variabel. Herved tages der højde for, at karakterniveauet kan være forskelligt fra år til år.

For hver elev beregnes en socioøkonomisk referenceværdi ud fra værdierne af elevens baggrundsvariable og de estimerede parametre for baggrundsvariablene. Herefter beregnes forskellen mellem opnået karakter og den socioøkonomiske reference, det såkaldte residual r_{ij}, for hver elev.

På skoleniveau er den estimerede forskel mellem karaktergennemsnittet og den socioøkonomiske reference i en ordinær regressionsmodel gennemsnittet af disse elevresidualer på den enkelte skole, $\bar{r}_j$. I de her anvendte to-niveau modeller, hvor vi både har variation mellem skoler ('Between'), og variation mellem elever indenfor skolen ('Within'), bliver residualerne på skoleniveau:

$$\hat{u}_j = c_j \bar{r}_j$$

hvor c_j er den såkaldte 'shrinkage' faktor, defineret ved

$$c_j = B \Big/ \left(B + W \Big/ n_j\right)$$

B er variationen mellem skoler ('Between'), *W* er residualvariationen mellem elever indenfor skolen ('Within') og n_j er antallet af elever på den *j*'te skole.

For forskellen mellem opnået karakter og den socioøkonomiske reference for hver skole er der endvidere udregnet et 95 % sikkerhedsinterval. Hvis dette sikkerhedsinterval ligger over 0, opnår skolens elever et statistisk signifikant højere karaktergennemsnit end elever på landsplan med lignende baggrundsforhold, mens det modsatte er tilfældet, hvis sikkerhedsintervallet ligger under 0. Indeholder 95 % sikkerhedsintervallet værdien 0, da opnår eleverne på den pågældende skole et karaktergennemsnit som elever på landsplan med lignende baggrundsforhold.

Sikkerhedsintervallet afhænger dels af den estimerede variation og dels af antallet af elever, der indgår i beregningerne for den enkelte skole. Således bliver sikkerhedsintervallerne mindre for den enkelte skole, når der betragtes en 3-årig periode, idet der her indgår 3 års elever i beregningerne.