Kvalitative variable (dummier)

Økonometri A

Bertel Schjerning

Program

Motivation (W7.1)

To kategorier (W7.2)

Mange kategorier (W7.3)

Interaktionseffekter (W7.4)

- Chow testet

Dummier som outcome variable (W7.5)

- Den lineære sandsynlighedsmodel

Opsummering

Motivation

Motivation

Indtil nu: Kontinuerte og kardinale variable

- Eksempler: Forbrug, priser, timeløn, uddannelse, erfaring, etc.
- Afstande mellem værdier er meningsfulde

Nu: Kategoriske og ordinale variable

- Køn (mand/kvinde)
- Sektor (offentlig/privat)
- Helbred (dårligt, middel, godt)
- Beskæftiget (i job/ikke i job)
- Bilmærke (Ford, Volvo, etc.), Motortype (El/Gas/Benzin/Hybrid)
- Udstyrsniveau (basis, mellemklasse, luksus)

Spørgsmål: Hvordan inkluderer vi dem i en regressionsmodel?

To kategorier

Dummy variable

Vi ønsker at undersøge effekten af kvalitativ variable med to kategorier, fx biologisk køn (mand/kvinde) eller sektor (offentlig/privat).

Vi kan opsummere informationen i en **dummy variabel**, som kun antager værdier 0 og 1.

For eksempel:

$$kvinde_i = \begin{cases} 1 & \text{hvis person i er en kvinde} \\ 0 & \text{ellers} \end{cases}$$

$$offentlig_i = \begin{cases} 1 & \text{hvis person i er ansat i det offentlige} \\ 0 & \text{ellers} \end{cases}$$

Referencekategorien er den kategori, hvor variablen er lig 0.

Dummy variable kaldes også **indikator** eller **binære** variable.

Dummy variable i den simple regressionsmodel

Eksempel: Gender wage gap

$$\log(\textit{timeløn}) = \beta_0 + \beta_1 \textit{kvinde} + u, \quad E[u|\textit{kvinde}] = 0$$

$$E(\log(\textit{timeløn})|\textit{kvinde}) = \beta_0 + \beta_1 \textit{kvinde}$$

Fortolkning:

- For kvinder ($\textit{kvinde} = 1$): $E(\log(\textit{timeløn})|\textit{kvinde} = 1) = \beta_0 + \beta_1$
- For mænd ($\textit{kvinde} = 0$): $E(\log(\textit{timeløn})|\textit{kvinde} = 0) = \beta_0$

Procentvis forskel:

$$\beta_1 = E(\log(\textit{timeløn})|\textit{kvinde} = 1) - E(\log(\textit{timeløn})|\textit{kvinde} = 0)$$

Dvs. β_1 er den forventede procentvise lønforskel mellem kvinder og mænd. Ofte estimeres $\hat{\beta}_1 < 0$, da kvinder tjener mindre i gennemsnit .

Dummy variable i den simple regressionsmodel

Parameterestimatet til en dummy i SLR har en særlig fortolkning:

$$\log(\textit{timeløn}) = \beta_0 + \beta_1 \textit{kvinde} + u.$$

OLS-estimat for $\hat{\beta}_1$:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\frac{1}{n}\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\frac{1}{n}\left[\sum_{x_i=1} y_i(1 - \bar{x}) + \sum_{x_i=0} y_i(0 - \bar{x})\right]}{\bar{x}(1 - \bar{x})} \\
&= \frac{\bar{y}_{x=1}(1 - \bar{x}) - \bar{y}_{x=0}\bar{x}}{\bar{x}(1 - \bar{x})} = \bar{y}_{x=1} - \bar{y}_{x=0}
\end{aligned}$$

Fortolkning: OLS-estimatet for dummyvariablen $\hat{\beta}_1$ er den (procentvise) forskel i gennemsnitslønnen mellem kvinder og mænd.

Dummy variable i den multiple regressionsmodel

Model:

$$\log(\textit{timeløn}) = \beta_0 + X\beta + \alpha \textit{kvinde} + u,$$

hvor X indeholder uddannelse, erfaring, osv.

Antagelse: Under MLR.4:

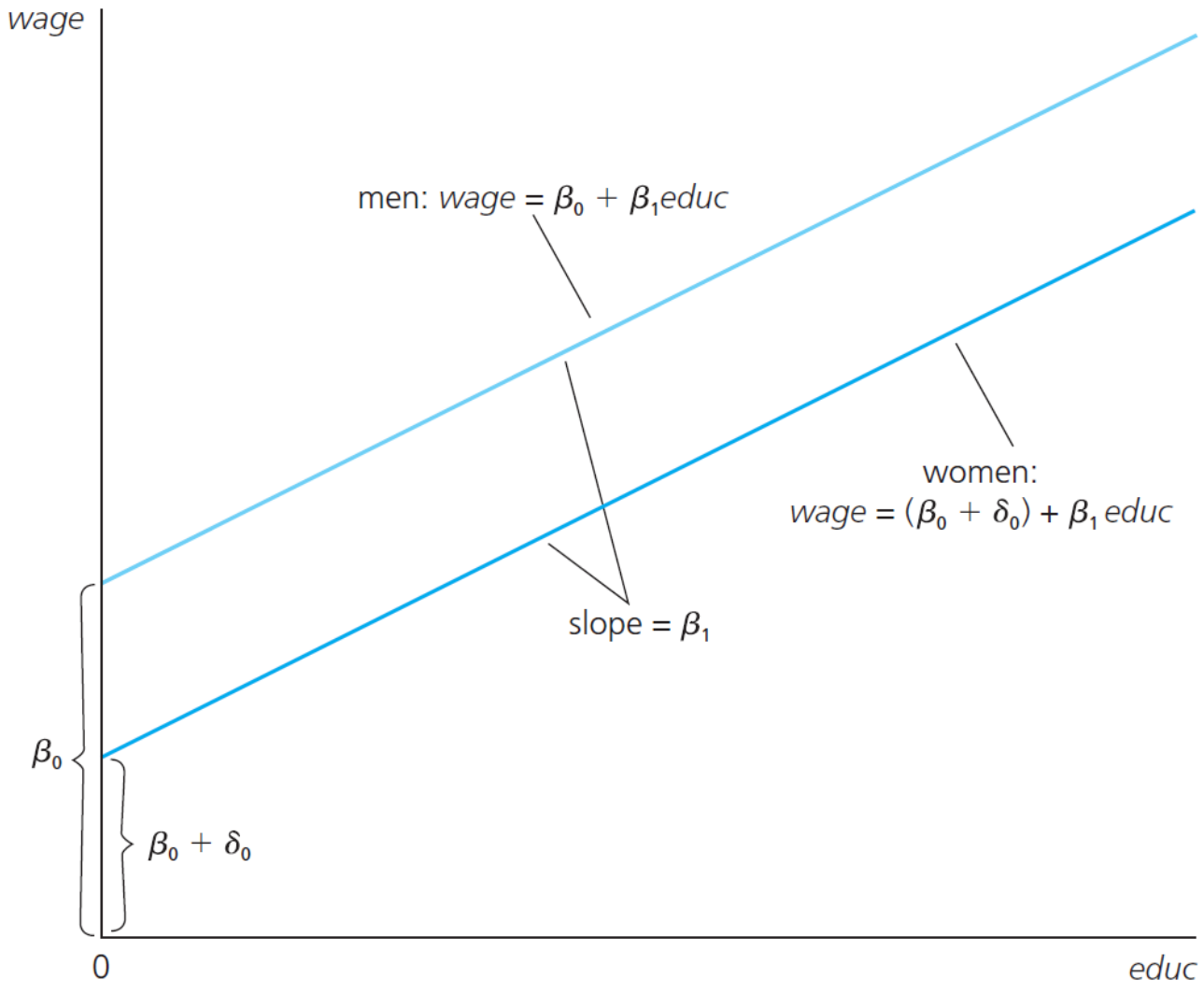
$$E(u|X, \textit{kvinde}) = 0 \implies$$
$$\alpha = E(\log(\textit{timeløn})|X, \textit{kvinde} = 1) - E(\log(\textit{timeløn})|X, \textit{kvinde} = 0)$$

α er forskellen i løn mellem mænd og kvinder med samme X.

Bemærk: Antagelsen medfører, at lønforskellen er uafhængig af X, så modellen tillader kun en konstant niveauforskel mellem mænd og kvinder.

Dummy variable i den multiple regressionsmodel

FIGURE 7.1 Graph of $wage = \beta_0 + \delta_0\, female + \beta_1\, educ$ for $\delta_0 < 0$.



Valg af referencegruppe

Hvad hvis vi i stedet havde inkluderet en dummy for "mand"?

Omskrivning af modellen:

$$mand_i + kvinde_i = 1 \quad \Longleftrightarrow \quad kvinde_i = 1 - mand_i$$

Resultat:

$$\begin{aligned} \log(timeløn) &= \beta_0 + \beta_1 uddannelse + ... + \alpha kvinde + u \\ &= \beta_0 + \beta_1 uddannelse + ... + \alpha(1 - mand_i) + u \\ &= (\beta_0 + \alpha) + \beta_1 uddannelse + ... - \alpha mand_i + u \end{aligned}$$

Begge dummy variable kan ikke inkluderes samtidig—dette kaldes "**dummy fælden**".

Let's Code!

- **Jupyter Notebook:** `03_dummyvar_examples.ipynb`
- **Python Module:** `mymlr.py`
- **Part 1:** Lønforskelle mellem mænd og kvinder

Dummy variable i den multiple regressionsmodel

Dummy variable ændrer ikke OLS-egenskaber:

- **Kausal effekt**: Parameteren fortolkes som effekten af at skifte fra én gruppe til en anden **(KUN hvis MLR.1-4 er opfyldt)**.
- **Illustration**: Effekten af en dummy vises som et skift i konstantleddet, når de øvrige variable holdes faste.
- **Afhængig variabels form**:
  - **I niveau**: Absolut forskel mellem grupperne.
  - **I log**: Procentvis forskel mellem grupperne.

Regression med to dummy variable

OLS estimatet for en dummy er forskellen i gennemsnit mellem to grupper.

Nu tilføjer vi en dummy for at arbejde i den offentlige sektor:

$$\log(\textit{timeløn}) = \beta_0 + \beta_1 x + \beta_2 z + u$$

Hvor:

- x: dummy for køn ($x = 1$ for kvinde, $x = 0$ for mand)
- z: dummy for sektor ($z = 1$ for offentlig, $z = 0$ for privat)

Vha. Frisch-Waughs teorem kan vi vise, at:

$$\hat{\beta}_1 = \frac{w_{z=1}(\bar{y}_{x=1,z=1} - \bar{y}_{x=0,z=1}) + w_{z=0}(\bar{y}_{x=1,z=0} - \bar{y}_{x=0,z=0})}{w_{z=1} + w_{z=0}} \quad (1)$$

Hvor w er vægte, som afhænger af størrelsen og variansen i grupperne.

Regression model med to dummy variable

OLS estimatet er et vægtet gennemsnit af effekten af kvinde ($x = 1$) i den offentlige ($z = 1$) og private sektor ($z = 0$):

$$\hat{\beta}_1 = \frac{w_{z=1}(\bar{y}_{x=1,z=1} - \bar{y}_{x=0,z=1}) + w_{z=0}(\bar{y}_{x=1,z=0} - \bar{y}_{x=0,z=0})}{w_{z=1} + w_{z=0}}$$

Vægtene er givet ved

$$w_{z=1} = \bar{z} \cdot \bar{x}_{z=1} \cdot (1 - \bar{x}_{z=1})$$
$$w_{z=0} = (1 - \bar{z}) \cdot \bar{x}_{z=0} \cdot (1 - \bar{x}_{z=0})$$

Dvs. vægtene er en funktion af to faktorer:

- Den relative størrelse af de to grupper ($\bar{z}$), og
- Den relative varians ($\bar{x}_z(1 - \bar{x}_z)$)

Hvorfor vægter OLS mindre ved større varians?

OLS vægter effekten af en variabel i forskellige grupper baseret på variansen af den uafhængige variabel.

Hvorfor?

- OLS minimerer residualvariansen, så grupper med mere præcis information får større indflydelse på estimaterne.
- Grupper med **højere varians** i x har større usikkerhed og får **mindre vægt**
- Grupper med **lav varians** giver mere præcis information og får **større vægt**.

Mange kategorier

Kvalitative variable med mere end to kategorier

Kvalitative variable kan også antage flere end to niveauer. Fx uddannelse, branche eller region.

Hvis den kvalitative variabel har m kategorier:

- Konstrueres $m - 1$ dummy variable, som antager værdien 0/1.
- Kategorien uden dummy variabel kaldes **referencekategorien**.
- Medtages m dummy variable ender man i **dummyfælden**.
- Parameteren til dummy variablen fortolkes som (niveau) forskellen til referencekategorien.
- Valget af referencekategorien:
  - har ingen betydning for modelegenskaberne,
  - men ændrer fortolkningen af parameterne

Kvalitative variable: Eksempel

Dansk Branchekode DB07, v3:2014-

Navn: Dansk Branchekode DB07, v3:2014-

Beskrivelse:
Dansk Branchekode 2007 (DB07), er en 6-cifret branchenomenklatur, der først og fremmest er udarbejdet til statistisk brug. De indledende afsnit til DB07 beskriver regler og retningslinier for tildeling af branchekode til virksomhederne og deres produktionsenheder (arbejdssteder). Dermed sikres, at brancheplaceringen foretages på en ensartet måde i Danmarks Statistiks Erhvervsstatistiske Register (ESR) og dermed også i Det Centrale Virksomhedsregister (CVR).

DB07 er en dansk underopdeling af EU's fælles branchenomenklatur, den 4-cifrede NACE rev. 2. NACE rev. 2 er en revideret udgave af NACE rev. 1.1 fra 2002. NACE rev. 2 bygger på FN´s branchenomenklatur ISIC rev. 4. NACE rev. 2 gælder fra 1. januar 2008 i alle EU-medlemslande ifølge forordning nr. 1893/2006.

Gyldig fra: 1. januar 2014

Kontor: Erhvervsindberetning og –registre

Kontaktperson: Birgit Nielsen, bgn@dst.dk, tlf. 39 17 38 69

Koder og kategorier

LUK HIERAKIET | DOWNLOAD

- A: Landbrug, jagt, skovbrug og fiskeri
  - 01: Plante- og husdyravl, jagt og serviceydelser i forbindelse hermed
    - 01.1: Dyrkning af etårige afgrøder
      - 01.11: Dyrkning af korn (undtagen ris), bælgfrugter og olieholdige frø
        - 011100: Dyrkning af korn (undtagen ris), bælgfrugter og olieholdige frø
      - 01.12: Dyrkning af ris
        - 011200: Dyrkning af ris
      - 01.13: Dyrkning af grøntsager og meloner, rødder og rodknolde
        - 011300: Dyrkning af grøntsager og meloner, rødder og rodknolde
      - 01.14: Dyrkning af sukkerrør
        - 011400: Dyrkning af sukkerrør
      - 01.15: Dyrkning af tobak
        - 011500: Dyrkning af tobak

Kardinale variable som kategorier

Vi kan omforme kvalitative variable til dummier for at inkludere dem i vores estimationer.

Kardinale variable kan også omdannes til kategorier (eksempelvis uddannelse).

- **Fordele**: Mere fleksibel funktionel form.
- **Ulemper**: Flere variable $\rightarrow$
  - Øget varians af OLS estimaterne, da færre frihedsgrader gør estimaterne mindre præcise.
  - Beregning af $X'X$ og inversion af store matricer kræver mere CPU-tid og hukommelse.
  - For matricer større end 10000×10000 stiger beregningstiden og hukommelseskravene markant.

Mange moderne studier sammenligner estimater fra kardinale variable med dummykodede versioner.

Kardinale variable som kategorier: Eksempel

I stedet for at måle uddannelse i år, kunne vi også betragte det som en kategoriseret variabel:

- Minimum uddannelse (edu < 10).
- 10. klasse (edu=10).
- Ungdomsuddannelser (edu $\in [11; 13]$).
- Videregående uddannelser (edu> 13).

[Mere interessant kunne man også opdele i fagområde]

Variablen har 4 kategorier, så vi skal have 3 dummy variable.

Regressionsmodel:

$$\begin{aligned}\log(\text{timeløn}) = \beta_0 &+ \delta_1 \mathit{klasse10} + \delta_2 \mathit{ung.udd} + \delta_3 \mathit{videreg.udd} \\ &+ \beta_2 \mathit{erfaring} + \beta_3 \mathit{erfaring}^2 + \beta_4 \mathit{kvinde} + u.\end{aligned}$$

Kardinale variable som kategorier: Eksempel

Hvordan kan vi teste hypotesen om at afkastet af uddannelse er lineært i antallet af års uddannelse?

To muligheder:

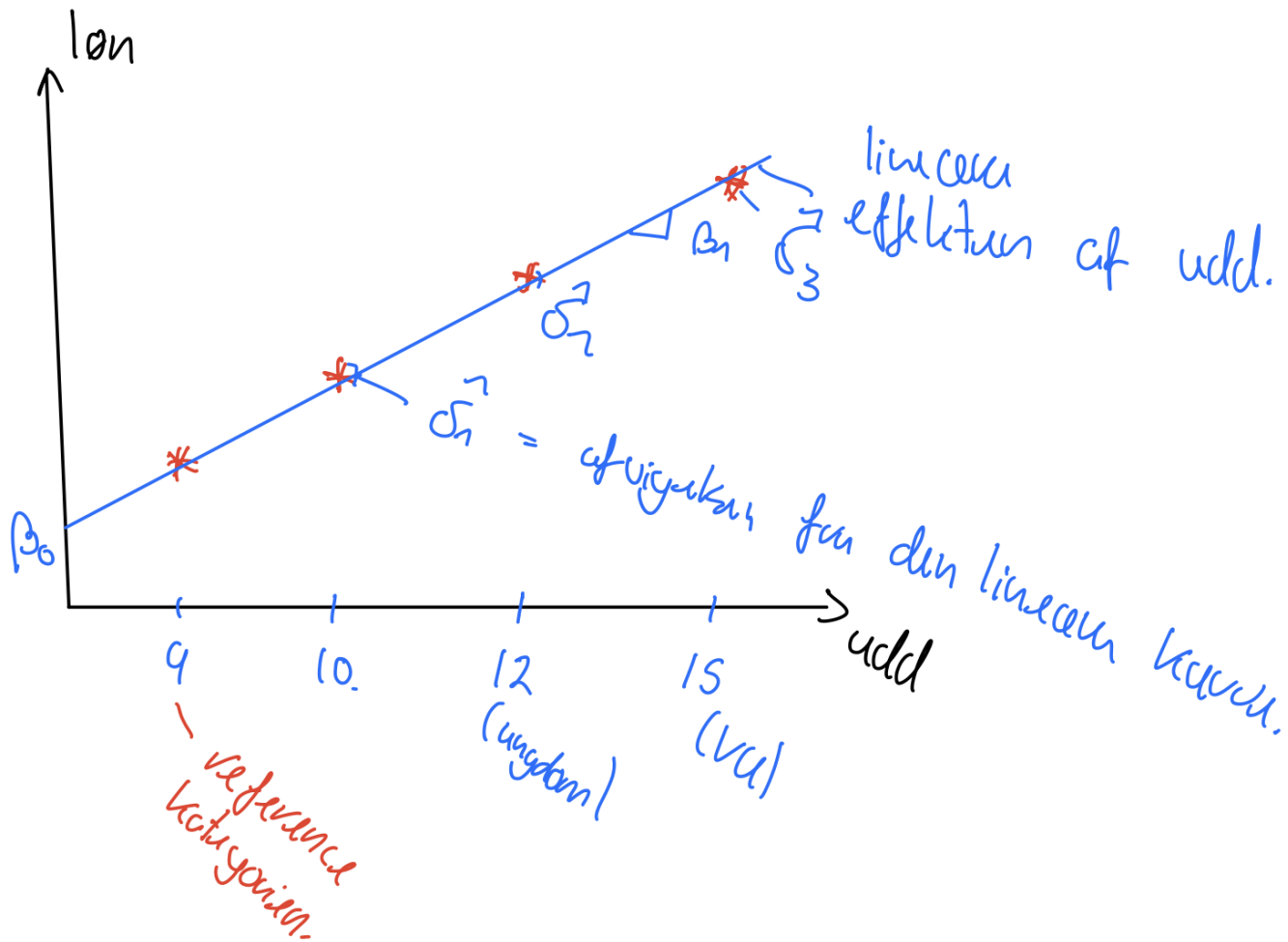
1. Test om sammenhængen mellem dummy estimaterne er "linear" (se Wooldridge s. 230).
2. Inkl. den lineære effekt af uddannelse og test om dummy estimaterne er 0.

Regressionsmodel:

$$\begin{aligned}\log(\text{timeløn}) = \beta_0 &+ \delta_1 \mathit{klasse10} + \delta_2 \mathit{ung.udd} + \delta_3 \mathit{videreg.udd} \\ &+ \beta_1 \mathit{uddannelse} + \beta_2 \mathit{erfaring} + \beta_3 \mathit{erfaring}^2 + \beta_4 \mathit{kvinde} + u.\end{aligned}$$

OBS: Dummy fælden

Kardinale variable som kategorier: test for linearitet



Let's Code!

- **Jupyter Notebook:** `03_dummyvar_examples.ipynb`
- **Python Module:** `mymlr.py`
- **Part 2:** Lønforskelle på tværs af uddannelseskategorier
- Specifickations test (F-test: Dummies eller lineær effekt)

Kardinale variable som kategorier: Eksempel

Boserup et al. (2018)[1] undersøger sammenhængen mellem forældre og børns formue

Dvs. en model a la:

$$\text{Børns formue} = \beta_0 + \beta_1 \text{Forældres formue} + u$$

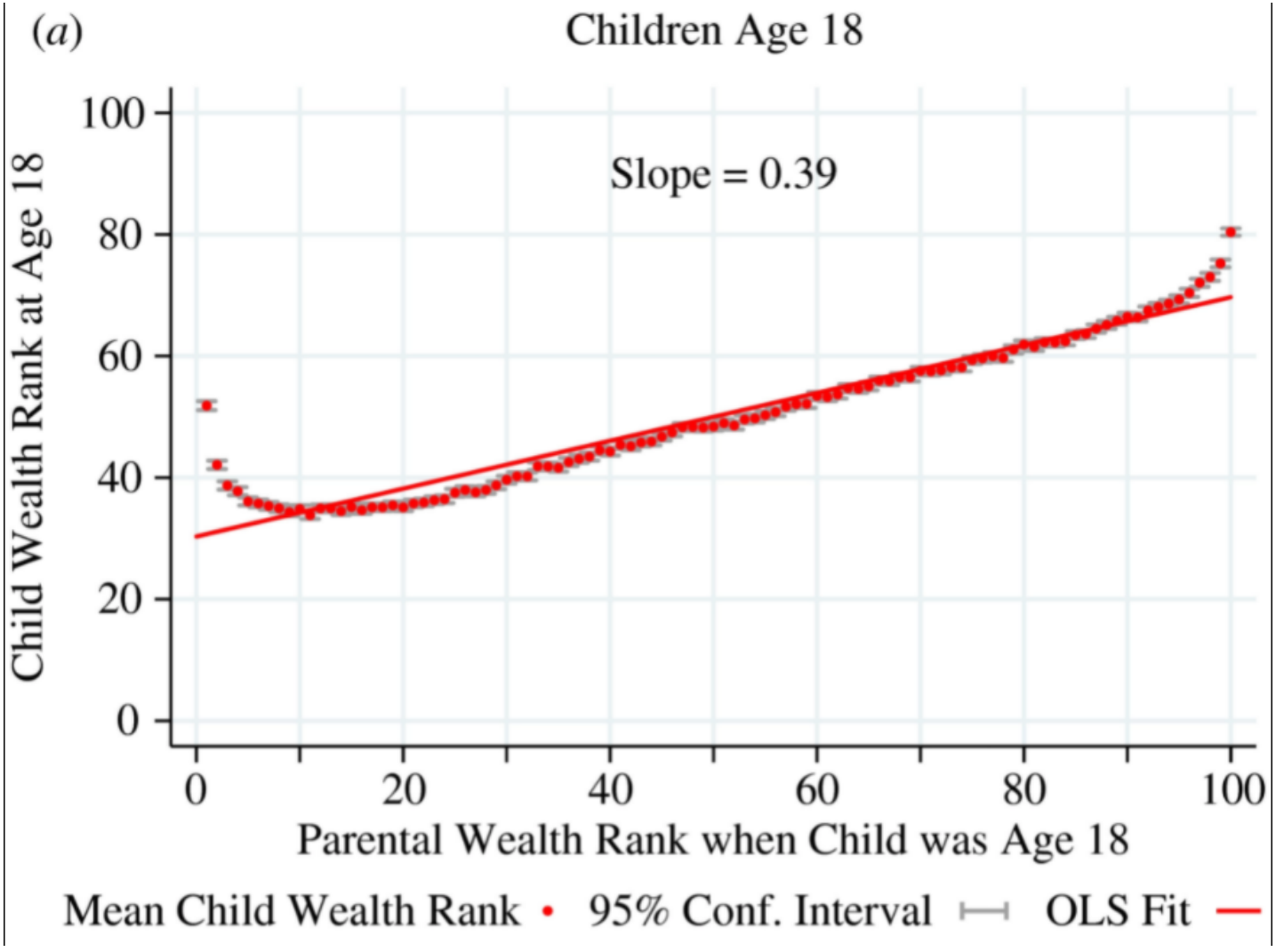
Spørgsmålet er om vi kan antage at modellen er lineær? Løsning:

- Grupper forældrene i "formue ranks", dvs. 100 lige store grupper efter formue og regresser

$$\text{Børns formue} = \beta_0 + \sum_j \beta_j \text{Forældres formue rank}_j + u$$

[1] Kilde: Boserup, S.H., Kopczuk, W. and Kreiner, C.T. (2018), Born with a Silver Spoon? Danish Evidence on Wealth Inequality in Childhood. Econ J, 128

Kardinale variable som kategorier: Eksempel



Interaktionseffekter

Interaktionsled

Når vi inkluderer en dummy variabel "for sig selv" tillader vi for **niveauskift** mellem to grupper.

Nogle gange vil vi også tillade for forskellige i den marginal effekt af en anden variable: **hældningsskift**.

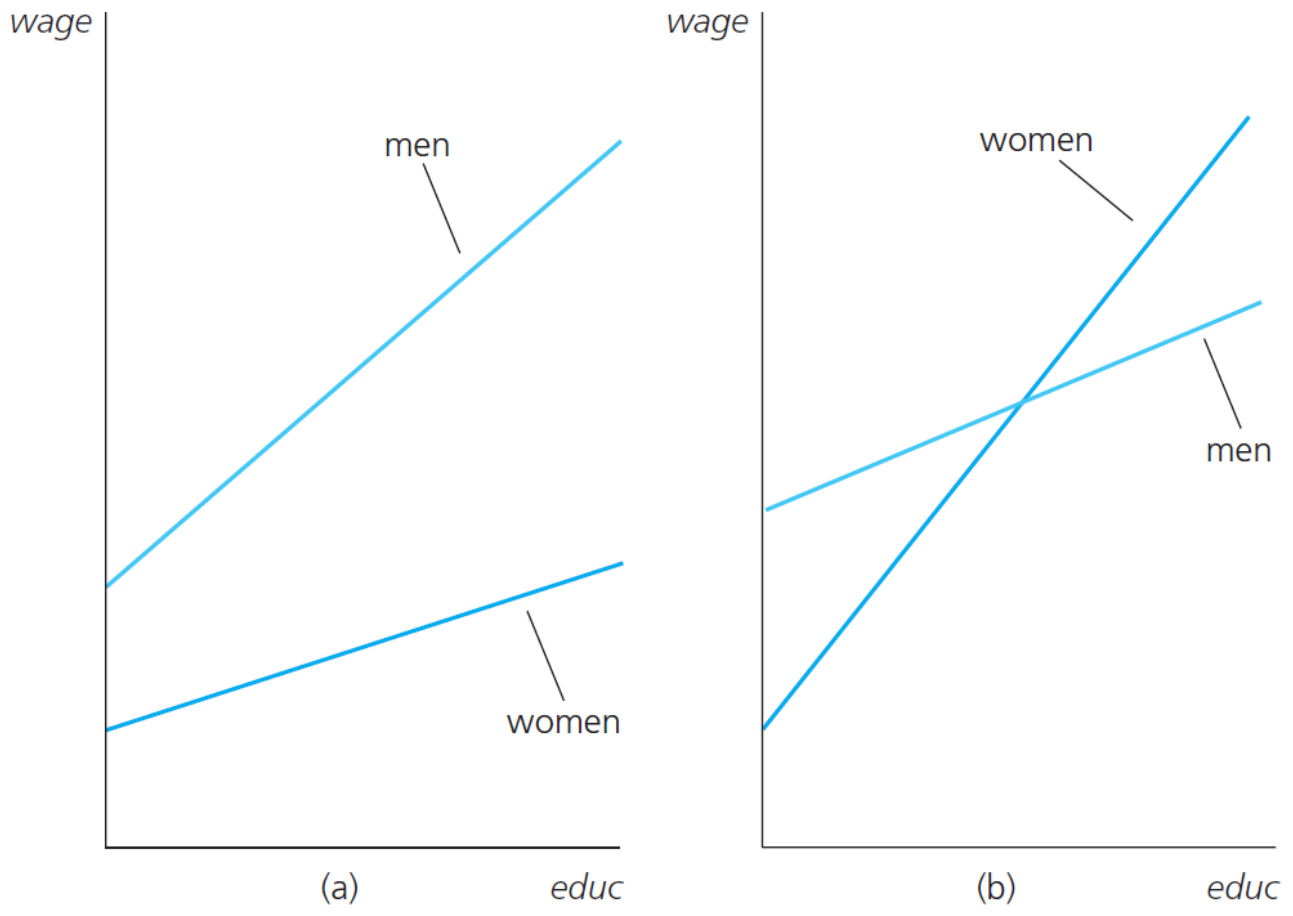
Dette kan vi gøre ved at lave interaktionsled mellem variablene:

$$\begin{aligned}\log(\textit{timeløn}) = \quad & \beta_0 + (\beta_1 + \beta_5 \textit{kvinde}) \cdot \textit{uddannelse} + \beta_2 \textit{erfaring} \\ & + \beta_3 \textit{erfaring}^2 + \beta_4 \textit{kvinde} + u\end{aligned}$$

Bemærk at vi "altid" har niveauet med, når vi inkluderer interaktionsled.

Interaktionsled: Afkastet af uddannelse afhænger af køn

FIGURE 7.2 Graphs of equation (7.16): (a) $\delta_0 < 0$, $\delta_1 < 0$; (b) $\delta_0 < 0$, $\delta_1 > 0$.



Interaktionsled

Endelig kan vi interagere en dummy variable med en anden dummy

Derved tillader vi niveauskiftet mellem to gruppe varierer opdelt på to andre grupper:

Interaktionsled mellem single og køn:

$$\begin{aligned}\log(\textit{timeløn}) = \quad & \beta_0 + \beta_1 \textit{uddannelse} + \beta_2 \textit{erfaring} + \beta_3 \textit{erfaring}^2 \\ + \quad & \beta_4 \textit{kvinde} + \beta_5 \textit{single} + \beta_6 \textit{single} \cdot \textit{kvinde} + u.\end{aligned}$$

Implicit antal grupper?

4: Single mænd, single kvinder, gifte mænd og gifte kvinder.

Hvad er referencegruppen her?

Interaktionsled

Hvordan skal parameterne fortolkes?

$$E[\log(\textit{timeløn})|\textit{erfaring}, \textit{uddannelse}, \textit{kvinde} \quad ,\textit{single} \quad] -$$
$$E[\log(\textit{timeløn})|\textit{erfaring}, \textit{uddannelse}, \textit{kvinde} = 0 \quad ,\textit{single} = 0 \quad] =$$

$$\textit{kvinde} = 1 \text{ og } \textit{single} = 0 \Rightarrow \beta_4$$
$$\textit{kvinde} = 0 \text{ og } \textit{single} = 1 \Rightarrow \beta_5$$
$$\textit{kvinde} = 1 \text{ og } \textit{single} = 1 \Rightarrow \beta_4 + \beta_5 + \beta_6$$

Let's Code!

- **Jupyter Notebook:** `03_dummyvar_examples.ipynb`
- **Python Module:** `mymlr.py`
- **Part 3:** Interaktionseffekter

Interaktionsled

Quiz

Betragt den følgende regressionsmodel:

$$\log(\textit{timeløn}) = \beta_0 + \beta_1 \textit{uddannelse} + \beta_2 \textit{erfaring} + \beta_3 \textit{erfaring}^2 + \beta_4 \textit{mand} + \beta_5 \textit{børn} + \beta_6 \textit{børn} \cdot \textit{mand} + u.$$

Hvad er den forventede forskel mellem mænd med børn og kvinder med børn?

1. β_4
2. β_5
3. $\beta_4 + \beta_5$
4. $\beta_5 + \beta_6$
5. $\beta_4 + \beta_6$
6. $\beta_4 + \beta_5 + \beta_6$

Interaktionseffekter eller separate estimationer

Hvad hvis vi interagerer *alle* variable med en dummy?

Simple model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Interageret med en dummy d:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \delta_0 d + \delta_1 x_1 \cdot d + \delta_2 x_2 \cdot d + u$$

Hvad er den betingede middelværdi af y for hver gruppe $d \in 0, 1$

$$\begin{aligned} E(y|d=0) &= \beta_0 & &+\beta_1 x_1 & &+\beta_2 x_2 \\ E(y|d=1) &= \beta_0 + \delta_0 & &+(\beta_1 + \delta_1) x_1 & &+(\beta_2 + \delta_2) x_2 \end{aligned}$$

Dvs. den betingede middelværdi i den interagerede model svarer til den betingede middelværdi fra to separate estimationer for de to gruppe.

Lønregression: Stata eksempel

```
reg lwage c.experience c.experience2 c.educ
estimates store simpel


reg lwage c.experience c.experience2 c.educ if kvinde == 0
estimates store mand


reg lwage c.experience c.experience2 c.educ if kvinde == 1
estimates store kvinde


reg lwage (c.experience c.experience2 c.educ)##i.kvinde
estimates store samlet


estimates table simpel mand kvinde samlet, stats(N r2)
```

Lønregression: Stata eksempel

```
estimates table alle mand kvinde interakt, stats(N r2)
 ----------------------------------------------------------------------
            Variable |    simpel         mand        kvinde      samlet
---------------------+------------------------------------------------
          experience |  .02536316    .03331748      .024342    .03331748
         experience2 | -.00040779   -.00068712   -.00050032   -.00068712
                educ |  .02743474    .03077454    .02341758    .03077454
              kvinde |
                   1 |                                        -.02779329
 kvinde#c.experience |
                   1 |                                        -.00897548
kvinde#c.experience2 |
                   1 |                                         .00018681
       kvinde#c.educ |
                   1 |                                        -.00735696
                     |
               _cons |  4.3154561    4.3300165    4.3022232    4.3300165
---------------------+------------------------------------------------
                   N |       1078          561          517         1078
                  r2 |  .20329622    .21341236    .17351194    .28787611
 ----------------------------------------------------------------------
```

Interaktionseffekter eller separate estimationer

Separate modeller er typisk mere overskuelige. Både i forhold til at

- Præsentere regressionsmodellen.
- Fortolke parameterestimaterne.

Men de gør det ikke umiddelbart muligt at teste om der er signifikante forskelle i parameterestimaterne for grupperne.

Det kan vi gøre på to måder

- Kører den fulde model og bruge en standard F-test.
- Kører de to separate estimationer og bruge **Chow testet**.

Chow testet med to grupper

Regressionsmodeller

$$y = \beta_{0,g} + \beta_{1,g} x_1 + ... + \beta_{k,g} x_k \quad \text{for} \quad g = 1, 2$$

$$\Leftrightarrow$$

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \delta_0 d_{g=2} + \delta_1 x_1 d_{g=2} + ... + \delta_k x_k d_{g=2}$$

Modellerne har tilsammen 2(k + 1) parametre.

Hypotese for ingen forskel på grupperne:

$$H_0 : \beta_{0,1} = \beta_{0,2}, \beta_{1,1} = \beta_{1,2}, ..., \beta_{k,1} = \beta_{k,2}$$
$$\Leftrightarrow \delta_0 = 0, \delta_1 = 0, ..., \delta_k = 0$$
$$H_1 : H_0 \text{ ikke opfyldt.}$$

Chow testet med to grupper

Procedure for Chow testet

1. Estimer den simple model for $g = 1, 2$ sammen og noter SSR_{simpel}.
2. Estimer modellen for gruppe 1 og noter SSR_1.
3. Estimer modellen for gruppe 2 og noter SSR_2.
4. Udregn teststørrelsen som

$$F = \frac{(SSR_{simpel} - (SSR_1 + SSR_2))/(k+1)}{(SSR_1 + SSR_2)/(n - 2(k+1))}.$$

5. Under H_0 er teststørrelsen F-fordelt med $(k+1)$ og $(n - 2(k+1))$ frihedsgrader.

OBS. Chow testet antager, at variansen af fejlleddet er den samme for de to grupper.

Chow testet med to grupper: Eksempel

Tal for SSR fra lønregressionerne ovenfor:

1. $SSR_{simpel} = 88.634$
2. $SSR_1 = 45.069$
3. $SSR_2 = 34.155$
4. Teststørrelsen

$$F = \frac{(88.634 - (45.069 + 34.155))/(3 + 1)}{(45.069 + 34.155)/(1078 - 2(3 + 1))} = 31.773$$

5. Under H_0 er teststørrelsen F-fordelt med 4 og 1070 frihedsgrader. Kritisk værdi ved 5% significansniveau = 2.380.

Forholdning af Chow teststørrelsen

Husk at $SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Med to grupper kan vi skrive SSR for den fuldt interagerede model som

$$SSR = \sum_{g=1}(y_i - \hat{y}_i)^2 + \sum_{g=2}(y_i - \hat{y}_i)^2$$

Da $\hat{y}_i$ fra den fuldt interagerede model er de samme som for de separate estimationer gælder det således at $SSR_{fuld} = SSR_1 + SSR_2$.

Dvs. Chow teststørrelsen er lig

$$F = \frac{(SSR_{simpel} - SSR_{fuld})/(k+1)}{(SSR_{fuld})/(n-2(k+1))}$$

Som netop er F-testet ved sammenligning af den fulde og samlede (restrikterede) model.

Chow testet med flere to grupper

Chow testet kan uden de store anstrengelser udvides til mere en to grupper:

Procedure

1. Estimer modellen for $g = 1, 2, .., G$ samlet og noter SSR_{samlet}.
2. Estimer modellen for hver gruppe g og noter SSR_g.
3. Udregn teststørrelsen som

$$F = \frac{(SSR_{samlet} - \sum_g SSR_g)/((G-1)(k+1))}{\sum_g SSR_g)/(n - G(k+1))}.$$

4. Under H_0 er teststørrelsen F-fordelt med $(G-1)(k+1)$ og $(n - G(k+1))$ frihedsgrader.

Dummier som outcome variable

Dummier som outcome variable

Indtil nu har vi set på modeller, hvor den afhængige variabel er kontinuert (f.eks. løn, forbrug, test score, BNP).

Hvad hvis den afhængige variabel er binære (antager to værdier)?

Eksempler:

- Arbejde eller ikke arbejde.
- Selvstændig eller lønmodtager.
- Består eksamen i økonometri A eller ej.
- Købe en bil eller ikke købe en bil.
- Investere i aktier eller ej.
- Blive skilt eller ej.
- Virksomheder eksporterer eller ej.
- Får et barn eller ej.
- Tag et SU lån, så man kan tage med DatØk på skiferie.

Den lineære sandsynlighedsmodel

Antag at y kun tager værdierne 0 eller 1.

Regressionsmodellen er nu (som den plejer):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_k x_k + u$$
$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_k x_k$$

Da y er binær gælder der, at

$$E(y|x) = \Pr(y = 0|x) \cdot 0 + \Pr(y = 1|x) \cdot 1 = \Pr(y = 1|x).$$

Dvs. at vi har lavet en model for sandsynligheden

$$\Pr(y = 1|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_k x_k.$$

Den lineære sandsynlighedsmodel

Modellen kaldes den **lineære sandsynlighedsmodel**.

$$\Pr(y = 1|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_k x_k.$$

Vi kan også udregne sandsynligheden for at $y = 0$

$$\Pr(y = 0|x) = 1 - \Pr(y = 1|x) = 1 - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - .. - \beta_k x_k.$$

Den lineære sandsynlighedsmodel

Fortolkningen af parametrene:

$$\Delta \Pr(y = 1|x) = \beta_j \Delta x_j.$$

Fortolkning β:

- y er diskret, så man kan ikke tale om en marginal ændring i y som følge af en ændring i x på en enhed.
- β angiver ændringen i sandsynligheden for $y = 1$ som følge af en ændring i x på en enhed.

Bortset fra fortolkningen er egenskaberne ved OLS de samme.

Den lineære sandsynlighedsmodel

Ulemper ved den lineære sandsynlighedsmodel:

- De prædikterede værdier af y ligger ikke nødvendigvis i intervallet $[0, 1]$.
- Dvs. de prædikterede sandsynligheder kan være mindre end 0 eller større end 1.
- I praksis er det typisk kun et problem for "atypiske" observationer med x'er langt fra gennemsnittet i stikprøven.

Gauss-Markow antagelserne:

- MLR.1-MLR.4 er muligvis opfyldt. Det må man - som altid - afgøre i hvert enkelt tilfælde).
- MLR.5 er per definition **ikke opfyldt**.

Variansen i den lineære sandsynlighedsmodel

Da y er en binær variabel, vil u også være binær for givet x:

$$y = 1 \Rightarrow u = 1 - \beta_0 - \beta_1 x_1 - \cdots - \beta_k x_k = 1 - p(x)$$
$$y = 0 \Rightarrow u = -\beta_0 - \beta_1 x_1 - \cdots - \beta_k x_k = -p(x)$$

Variansen af u for givet x:

$$\begin{aligned}
\text{Var}(u|x) &= E(u^2|x) \\
&= E(u^2|x, y = 1)p(x) + E(u^2|x, y = 0)(1 - p(x)) \\
&= (1 - p(x))^2 p(x) + (-p(x))^2 (1 - p(x)) \\
&= (1 - p(x))p(x)[(1 - p(x)) + p(x)] \\
&= (1 - p(x))p(x)
\end{aligned}$$

Da $\text{Var}(u|x)$ afhænger af $x \Rightarrow$ heteroskedasticitet.

Simple lineære sandsynlighedsmodel

Quiz: Antag vi har følgende model hvor d og y er dummy variable

$$y = \beta_0 + \beta_1 d + u$$

Vi har 100 observationer som fordeler sig således

	$d = 0$	$d = 1$
$y = 0$	20	40
$y = 1$	20	20

A: **Hvad er estimatet af** β_0?	B: **Hvilket fortegn vil** $\hat{\beta}_1$ **være?**
a1. $\hat{\beta}_0 = \frac{1}{4}$	b1. $\hat{\beta}_1 > 0$
a2. $\hat{\beta}_0 = \frac{1}{2}$	b2. $\hat{\beta}_1 = 0$
a3. $\hat{\beta}_0 = \frac{2}{3}$	b3. $\hat{\beta}_1 < 0$

Udregn OLS estimatoren for β_1

Den lineære sandsynlighedsmodel: Eksempel

Vi ønsker at modellere sandsynligheden for at blive arresteret

$$arr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + u$$

hvor

*arr*86	Arresteret i 1986 (binær variabel)
pcnv	Andel af tidligere arrestationer, som medførte dom
avgsen	Gennemsnitlig længde af tidligere straffe
tottime	Total tid i fængsel siden 18 års alderen
*ptime*86	Måneder i fængsel i 1986
*qemp*86	Antal kvartaler i beskæftigelse i 1986

Population: Mænd i Californien født 1960-61 og som mindst en gang før 1986 har været arresteret.

Kriminalitet: Stata eksempel

```
reg arr86 pcnv avgsen tottime ptime86 qemp86

      Source |       SS           df       MS      Number of obs   =     2,725
-------------+----------------------------------   F(5, 2719)      =     27.03
       Model |  25.8452455         5  5.16904909   Prob > F        =    0.0000
    Residual |  519.971268     2,719  .191236215   R-squared       =    0.0474
-------------+----------------------------------   Adj R-squared   =    0.0456
       Total |  545.816514     2,724   .20037317   Root MSE        =    .43731

------------------------------------------------------------------------------
       arr86 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |  -.1624448   .0212368    -7.65   0.000    -.2040866    -.120803
      avgsen |   .0061127    .006452     0.95   0.344    -.0065385     .018764
     tottime |  -.0022616   .0049781    -0.45   0.650    -.0120229    .0074997
     ptime86 |  -.0219664   .0046349    -4.74   0.000    -.0310547   -.0128781
      qemp86 |  -.0428294   .0054046    -7.92   0.000    -.0534268   -.0322319
       _cons |   .4406154   .0172329    25.57   0.000     .4068246    .4744063
------------------------------------------------------------------------------
```

Outcome variable med flere end to kategorier

Hvad gør vi, hvis outcome variablen har mere en to kategorier? Fx

- Arbejdsløs, Studerende, Beskæftiget.

En løsning er at danne to dummier, fx

- $D_1 = 1$ hvis studerende og 0 ellers.
- $D_2 = 1$ hvis beskæftiget og 0 ellers.

Og kører en separat LPM for hver dummy.

I lærer om mere advancerede sandsynlighedsmodeller i Økonometri B.

Opsummering

Opsummering

- Dummy variable gør det muligt at arbejde med kvalitativ information i OLS modeller.
- Dummier ændrer ikke grundlæggende på OLS egenskaberne.
- Dummier for sig selv tillader niveauforskelle mellem grupper (level shifts).
- Interaktionsled tillader forskelle i hældning mellem grupper (slope shifts)
- Variable med flere end 2 kategorier kan omdannes til flere (0/1)-dummy variable.
- En dummy, som outcome variabel, medfører den lineære sandsynlighedsmodel.