

Øvelsestime 6

Dummy variable, interaktionsled og
Chow-test

Agenda

- 1) Introduktion til Dummy variable og PS6
- 2) Gruppearbejde: Diskussion
- 3) Opsamling
- 4) Gruppearbejde: Stataøvelser
- 5) Opsamling

Tobias Gabel Christiansen
tgc@econ.ku.dk
29-03-2021

KØBENHAVNS UNIVERSITET



1) Planen for semesteret

Dato	Opgave
Uge 7 - 15/02	PS1
Uge 8 - 22/02	PS2
Uge 9 - 01/03	PS3
Uge 10 - 08/03	OO1
Uge 11 - 15/03	PS4
Uge 12 - 22/03	PS5 + gennemgang af OO1
Uge 13 - 29/03	PS6
Uge 14 - 05/04	Påskeferie
Uge 15 - 12/04	OO2
Uge 15 - 19/04	PS7
Uge 16 - 26/04	PS8 + gennemgang af OO2
Uge 17 - 03/05	OO3
Uge 18 - 10/05	PS9
Uge 19 - 17/05	PS10 + gennemgang af OO3
	Spørgetime

 Her er vi :-)

1) Introduktion til Dummy variable

- I PS6 ser vi nærmere på hvordan vi kan bruge kvalitativ data i regressionsanalyser.
- I modsætning til hidtil er kvalitativ diskret (ikke-kontinuert) og tager form som enten:
Dummy variable (1/0): Fx kan køn kodes som en dummy variable hvor kvinde=1 og mand=0.
Kategoriske variable (1/2/3/4...): Fx kan kreditvurderinger kodes som en kategorisk variabel hvor AAA=1, AA+=2, AA=3 osv.
- I PS6 fokuserer vi på brugen af **dummy** variable.
- Ved at inkludere dummy variable kan vi bl.a. bruge modellen til at estimere forskelle på tværs af grupper uden at skulle restrikttere sample (husk fx at vi tidligere har brugt "reg y x if dmale==1").

Ekstra: Kategoriske variable kan opsplittes til flere dummy variable. Det er mindre restriktivt da man tillader at effekten af at gå fra fx AAA til AA+ er forskellig fra effekten ved at gå fra AA+ til AA osv. For en kategorisk variabel antages denne effekt at være ens. Ulempen ved at inkludere dummier for hver kategori er at antallet af variable i modellen ofte vokser hurtigt.

1) Introduktion til Dummy variable

Med dummy variable kan vi estimere **forskellige skæringer med y-aksen på tværs af grupper**. Fx kan vi finde forskelle på tværs af køn:

$$y_i = \beta_0 + \beta_1 x_i + \delta_1 Kvinde_i + u_i$$

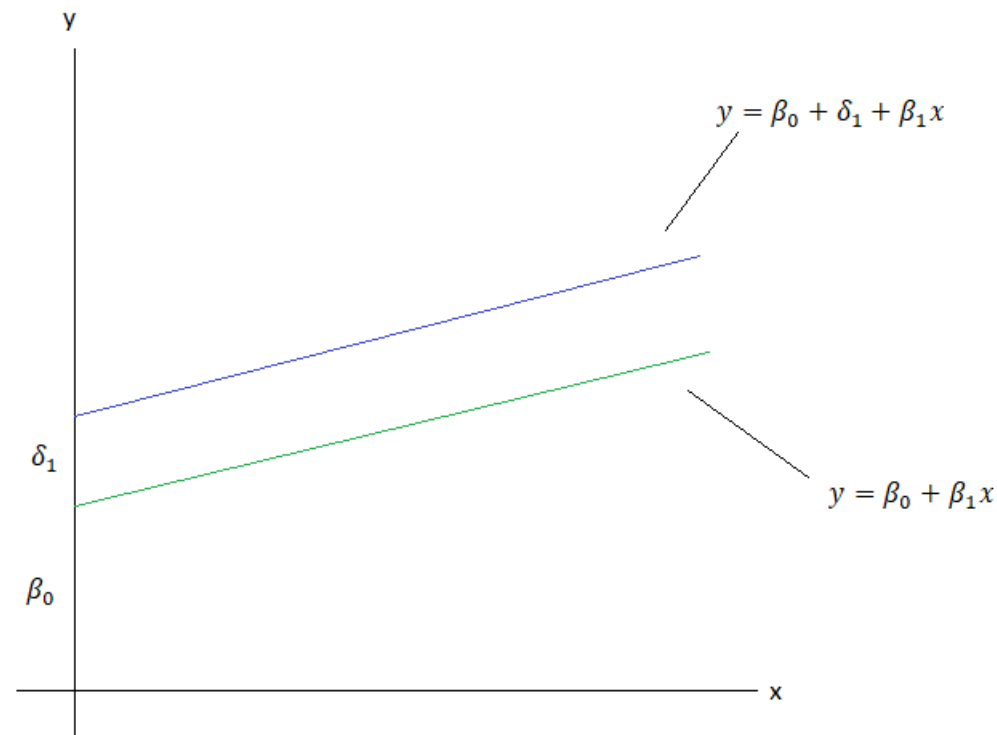
Hvor $Kvinde_i$ er en dummy der er lig 1 hvis individ i er en kvinde. I denne model er skæringen med y-aksen for mænd givet ved

$$E[y_i | x_i = 0, Kvinde_i = 0] = \beta_0$$

og for kvinder givet ved

$$E[y_i | x_i = 0, Kvinde_i = 1] = \beta_0 + \delta_1$$

Koefficienten på dummien (δ_1) fanger derfor forskelle i skæringen med y-aksen ift. referencegruppen (i dette tilfælde mænd). Da vi antager at hældningen er ens på tværs af køn angiver δ_1 faktisk niveausforskellen mellem mænd og kvinder for alle $x \in R$.

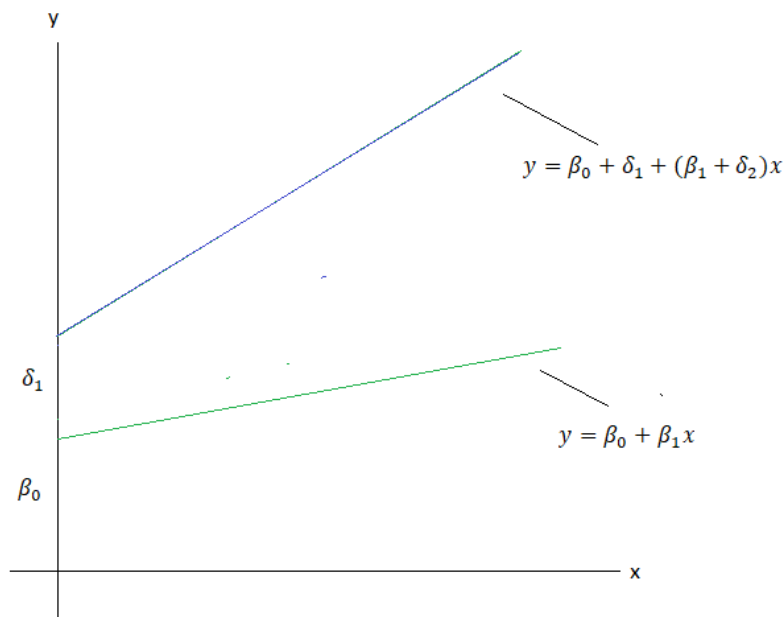


Introduktion til Dummy variable

På forrige slide var effekten af at øge x_i identisk på tværs af køn (β_1). Vi kan afslappe denne antagelse og estimere forskellige **hældninger på tværs af grupper** ved at bruge interaktionsled:

$$y_i = \beta_0 + \beta_1 x_i + \delta_1 Kvinde_i + \delta_2 (Kvinde_i \times x_i) + u_i \quad (A)$$

Nu er effekten af at øge x_i lig β_1 for mænd og lig $\beta_1 + \delta_2$ for kvinder. Bemærk at δ_1 nu kun angiver niveauforskellen for $x = 0$ (dvs. forskellen i skæringen med y-aksen)



Den fuldt interagerede model (dvs. model (A) ovenfor) svarer til at lave to regressioner af den simple model ($y_i = \beta_0 + \beta_1 x_i + u_i$) restrikeret til observationer hvor hhv. $i = mand$ og $i = kvinde$. Ulempen ved dette er at det er sværere at teste for forskelle på tværs af grupper (sværere at udregne de korrekte standardfejl, som automatisk findes i model (A) ovenfor).

1) Introduktion til Dummy variable

Direkte effekter og interaktionseffekter:

- Hvis vi ønsker at modellere interaktionseffekter på tværs af grupper bør vi altid inkludere den direkte effekt for grupperne.

Eksempel

Hvis vi fx ønsker at modellere hvordan effekten fra x på y afhænger af D bør vi modellere det som:

$$y_i = \beta_0 + \beta_1 x_i + \delta_1 D_i + \delta_2 (D_i \times x_i) + u_i \quad (1)$$

Fremfor

$$y_i = \beta_0 + \beta_1 x_i + \delta_2 (D_i \times x_i) + v_i \quad (2)$$

I (2) udelader vi den direkte effekt fra vores dummy ($\delta_1 D_i$). Hvis D_i har en effekt på y_i der ikke afhænger af x_i ($\delta_1 \neq 0$) vil estimatoren for interaktionsleddet ($\widehat{\delta_2}$) generelt være biased og inkonsistent i model (2). Skyldes at $v_i = u_i + \delta_1 D_i$ hvilket generelt vil korrelerer med interaktionsleddet.

Med andre ord er det en god ide at inkludere *forskellige skæringer* med y-aksen hvis vi fx ønsker at modellere *forskellige hældninger* på tværs af grupper.

1) Introduktion til Dummy variable

Dummytrap:

- Dummytrap opstår hvis vi ønsker at modellere g grupper som g dummy-variable samtidigt med at vi inkluderer et konstantled. Det vil skabe perfekt multikollinearitet mellem dummy-variablene og konstantleddet.

Eksempel

Hvis vi fx ønsker at modellere forskelle på tværs af køn og forsøger at estimere følgende model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 Mand_i + \beta_3 Kvinde_i + u_i$$

hvor $Mand_i = 1$ hvis i er mand og $Kvinde_i = 1$ hvis i er kvinde. Da alle individer kan kategoriseres som enten mand eller kvinde har vi at:

$$Mand_i + Kvinde_i = 1 = \text{konstantled}$$

- For at undgå dette kan vi enten **i)** udelade konstantleddet eller **ii)** udelade én af de to dummyvariable.
- Vi foretrækker som udgangspunkt **ii)** da vi derved lettere kan sammenligne og teste for forskelle på tværs af grupper.
- Generelt danner vi derfor $(g - 1)$ dummy-variable hvis vi har g grupper.
- Gruppen vi udelader fungerer som **referencegruppen**: Alle koefficienter for de inkluderede dummy-variable skal fortolkes relativt til denne gruppe.

1) Introduktion til PS6

- I PS6 skal vi estimere en hedonisk prismodel for lejlighedspriser.
- *Hedonisk prisfastsættelse*: Et gode (fx en bolig) kan værdiansættes på baggrund af dets nyttebærende karakteristika (fx størrelse, beliggenhed, mv.). Den samlede pris på boligen afspejles af de implicitte priser på de forskellige karakteristika. Vores mål er at estimere disse implicitte priser da det giver os mulighed for at prisfastsætte lejligheder.
- Data er et random sample bestående 988 lejlighedssalg i København i 2005.
- Data indeholder kvalitativ information om beliggenhed (KBH K, KBH V, KBH N, KBH Ø). Det er denne information vi vil modellere som dummy-variable i opgavesættet.
- Derudover indeholder datasættet information om størrelse, antal rum, antal toiletter, alder på lejlighed mv.

2) Gruppearbejde - Diskussion

To-do (20 minutter):

$$\log(\text{price}_i) = \beta_0 + \beta_1 m2_i + \beta_3 \text{rooms}_i + \beta_4 \text{toilets}_i + u_i$$

1. Discuss if it is reasonable to assume that assumptions MLR.1–MLR.5 are satisfied for model (1).
2. Discuss how the hedonic model can be extended with one or more dummy variables to investigate if the model's parameters differ across locations in Copenhagen? Note that the variable **location** takes on four categories. How would you construct the dummy variable(s)?
3. How can you test for level differences in apartment prices across Copenhagen? How can you test if the (implicit) price of an additional square meter is different across locations in Copenhagen?

Brug for hjælp? Skriv Jeres gruppenummer i Padlet (link i chatten)

3) Opsamling på diskussion

1. Discuss if it is reasonable to assume that assumptions MLR.1–MLR.5 are satisfied for model (1).

$$\log(\text{price}_i) = \beta_0 + \beta_1 \text{m2}_i + \beta_3 \text{rooms}_i + \beta_4 \text{toilets}_i + u_i$$

3) Opsamling på diskussion

1. Discuss if it is reasonable to assume that assumptions MLR.1–MLR.5 are satisfied for model (1).

$$\log(\text{price}_i) = \beta_0 + \beta_1 \text{m2}_i + \beta_3 \text{rooms}_i + \beta_4 \text{toilets}_i + u_i$$

- **MLR.1** (lineær i parametre): Modellen er lineær i parametre jf. ovenfor.
- **MLR.2** (tilfældig stikprøve): Opfyldt jf. databekrivelsen ("random sample of apartment sales").
- **MLR.3** (ingen multikollinearitet): Opfyldt da ingen af variablene umiddelbart kan skrives som en lineær kombination af de andre.
- **MLR.4** ($E[u|X] = 0$): Svært at argumentere for at vi ikke har udeladt relevante variable der også samvarierer med X . Fx kan lokation være en relevant variable som også samvarierer med størrelsen på boligen.
- **MLR.5** ($\text{var}(u|X) = \sigma^2 I_n$): Homoskedasticitet kan være brudt hvis prischok (u_i) for store lejligheder har større varians end for små lejligheder. Da er fejllidets varians større for store lejligheder end for små. (vi ser hvordan vi kan teste for og håndterer tilfælde hvor MLR.5 er brudt næste gang)

3) Opsamling på diskussion

2. Discuss how the hedonic model can be extended with one or more dummy variables to investigate if the model's parameters differ across locations in Copenhagen? Note that the variable **location** takes on four categories. How would you construct the dummy variable(s)?

3) Opsamling på diskussion

2. Discuss how the hedonic model can be extended with one or more dummy variables to investigate if the model's parameters differ across locations in Copenhagen? Note that the variable **location** takes on four categories. How would you construct the dummy variable(s)?

- Der er 4 beliggenhedskategorier:

København Ø
København V
København N
København K

- For at undgå dummy trap kan vi inkludere 3 beliggenhedsdummier:

$$\begin{aligned}KbhV &= 1(i \in \text{København V}) \\KbhN &= 1(i \in \text{København N}) \\KbhK &= 1(i \in \text{København K})\end{aligned}$$

Hvor $1(\cdot)$ er en indikatorfunktion.

- Ved at udelade København Ø fungerer denne beliggenhed som referencekategori.
 - i) Beliggenhedseffekten for lejligheder i KbhØ findes ved at sætte alle tre ovenstående dummier lig 0
 - ii) Koefficienterne på de tre ovenstående dummier tolkes som ændringer i prisen ved at gå fra KbhØ til fx KbhV.
- Valget af referencekategori er i dette tilfælde er arbitrært. Som huskeregel er det dog en god ide at vælge en referencekategori som vi har mange observationer for (vi bør kunne estimere effekten for referencekategorien med en vis sikkerhed for at gøre sammenligninger meningsfulde)

3) Opsamling på diskussion

- Ved at inkludere dummy variable for de tre bydele kan vi estimere niveau forskelle i lejlighedspriser der skyldes beliggenhed:

$$\log(\text{price}_i) = \beta_0 + \delta_1 \text{KbhK}_i + \delta_2 \text{KbhN}_i + \delta_3 \text{KbhV}_i + \beta_1 m2_i + \beta_2 \text{rooms}_i + \beta_3 \text{toilets}_i + v_i$$

- β_0 angiver $\log(\text{price}_i)$ for lejligheder på Østerbro (KbhØ) på 0 m^2 , med 0 rum og med 0 toiletter
- $\beta_0 + \delta_1$ angiver det samme blot for lejligheder i indre by (KbhK)
- Da vi antager at hældningerne er de ens for alle bydele fanger δ_1 niveauforskelle i den forventede pris for ens lejligheder i KbhØ og KbhK (dvs. ikke kun for lejligheder med $m2_i = 0$, $\text{rooms}_i = 0$, $\text{toilets}_i = 0$)
- Da vi arbejder med en log-level model svarer $100\delta_1\%$ approksimativt til den forventede stigning i prisen ved at gå fra en lejlighed på Østerbro til en tilsvarende lejlighed i Indre by

3) Opsamling på diskussion

- Ved yderligere at inkludere interaktionsled kan vi lade de implicitte priser på størrelse, antal rum og antal toiletter (hældningerne) variere på tværs af bydele:

$$\begin{aligned} \log(\text{price}_i) &= \beta_0 + \delta_1 KbhK_i + \delta_2 KbhN_i + \delta_3 KbhV_i + \delta_4 m2_i + \delta_5 (KbhK_i \times m2_i) + \delta_6 (KbhN_i \times m2_i) + \delta_7 (KbhV_i \times m2_i) \\ &+ \delta_8 \text{rooms}_i + \delta_9 (KbhK_i \times \text{rooms}_i) + \delta_{10} (KbhN_i \times \text{rooms}_i) + \delta_{11} (KbhV_i \times \text{rooms}_i) + \delta_{12} \text{toilets}_i \\ &+ \delta_{13} (KbhK_i \times \text{toilets}_i) + \delta_{14} (KbhN_i \times \text{toilets}_i) + \delta_{15} (KbhV_i \times \text{toilets}_i) + u_i \end{aligned}$$

- Den forventede stigning i prisen ved øge størrelsen på en lejlighed i KbhØ med $1m^2$ er (approx.) $100\delta_4\%$ når antal rum og antal toiletter holdes fast (husk KbhØ er referencekategorien).
- Den forventede stigning i prisen ved øge størrelsen på en lejlighed i KbhK med $1m^2$ er (approx.) $100(\delta_4 + \delta_5)\%$ når antal rum og antal toiletter holdes fast. δ_5 fanger derfor forskelle i (implicitte) m^2 -priser lejligheder i KbhØ og KbhK, når der kontrolleres for antal rum og toiletter.
- Den samlede forskel i prisen mellem lejligheder i KbhØ og KbhK er:

$$\delta_1 + \delta_5 m2_i + \delta_9 \text{rooms}_i + \delta_{13} \text{toilets}_i$$

3) Opsamling på diskussion

3. How can you test for level differences in apartment prices across Copenhagen? How can you test if the (implicit) price of an additional square meter is different across locations in Copenhagen?

3) Opsamling på diskussion

3. How can you test for level differences in apartment prices across Copenhagen? How can you test if the (implicit) price of an additional square meter is different across locations in Copenhagen?

- Hvis vi tager udgangspunkt i den fuldt interagerede model:

$$\begin{aligned} \log(\text{price}_i) &= \beta_0 + \delta_1 \mathbf{KbhK}_i + \delta_2 \mathbf{KbhN}_i + \delta_3 \mathbf{KbhV}_i + \delta_4 m2_i + \delta_5 (\mathbf{KbhK}_i \times m2_i) + \delta_6 (\mathbf{KbhN}_i \times m2_i) + \delta_7 (\mathbf{KbhV}_i \times m2_i) \\ &+ \delta_8 \text{rooms}_i + \delta_9 (\mathbf{KbhK}_i \times \text{rooms}_i) + \delta_{10} (\mathbf{KbhN}_i \times \text{rooms}_i) + \delta_{11} (\mathbf{KbhV}_i \times \text{rooms}_i) + \delta_{12} \text{toilets}_i \\ &+ \delta_{13} (\mathbf{KbhK}_i \times \text{toilets}_i) + \delta_{14} (\mathbf{KbhN}_i \times \text{toilets}_i) + \delta_{15} (\mathbf{KbhV}_i \times \text{toilets}_i) + u_i \end{aligned}$$

- Vi kan teste for niveauforskelle ved brug af følgende hypotese:

$$\begin{aligned} H_0: \delta_1 = \delta_2 = \delta_3 = 0 \\ H_a: \text{Et eller flere } \delta_i \neq 0, \quad i = 1, 2, 3 \end{aligned}$$

- Under MLR.1-5 og $n \rightarrow \infty$ kan vi bruge en F – test med tre restriktioner
- Bemærk: Det er ikke helt sigende at kalde dette for niveauforskelle da vi lader hældningerne variere på tværs af bydele. δ_1 , δ_2 og δ_3 angiver strengt taget kun niveauforskelle for lejligheder på 0 m^2 , med 0 rum og med 0 toiletter (dvs. forskelle i skæringen med y-aksen)

3) Opsamling på diskussion

3. How can you test for level differences in apartment prices across Copenhagen? How can you test if the (implicit) price of an additional square meter is different across locations in Copenhagen?

- Hvis vi tager udgangspunkt i den fuldt interagerede model:

$$\begin{aligned} \log(\text{price}_i) &= \beta_0 + \delta_1 KbhK_i + \delta_2 KbhN_i + \delta_3 KbhV_i + \delta_4 m2_i + \delta_5 (KbhK_i \times m2_i) + \delta_6 (KbhN_i \times m2_i) + \delta_7 (KbhV_i \times m2_i) \\ &+ \delta_8 \text{rooms}_i + \delta_9 (KbhK_i \times \text{rooms}_i) + \delta_{10} (KbhN_i \times \text{rooms}_i) + \delta_{11} (KbhV_i \times \text{rooms}_i) + \delta_{12} \text{toilets}_i \\ &+ \delta_{13} (KbhK_i \times \text{toilets}_i) + \delta_{14} (KbhN_i \times \text{toilets}_i) + \delta_{15} (KbhV_i \times \text{toilets}_i) + u_i \end{aligned}$$

- Vi kan teste for forskelle i (implicitte) kvadratmeterpriser ved brug af følgende hypotese:

$$\begin{aligned} H_0: \delta_5 &= \delta_6 = \delta_7 = 0 \\ H_a: &\text{Et eller flere } \delta_i \neq 0, \quad i = 5, 6, 7 \end{aligned}$$

- Under MLR.1-5 og $n \rightarrow \infty$ kan vi bruge en F – test med tre restriktioner

4) Stataøvelser

- **Arbejd på Stataøvelserne frem til 12:30**

- **Hints:**

- Generelt:
Stata kan godt drille når vi bruger bogstavet "Ø". For at undgå det kan i omdøbe "KBH Ø" til "KBH OE" med følgende kommando:

```
replace location = "KBH OE" if location!="KBH N" & location!="KBH V" & location!="KBH K"
```

- Opg. 3:
Chow-testen er givet ved:

$$Chow = \frac{\frac{SSR_p - (SSR_1 + \dots + SSR_m)}{(m-1)(k+1)}}{\frac{SSR_1 + \dots + SSR_m}{n - (m(k+1))}} \sim^a F((m-1)(k+1), n - m(k+1))$$

OBS: Brug for hjælp? Skriv gruppenummer i Padlet (link i chatten)

5) Stataøvelser - opsamling

1. Load CPHAPTS.dta into STATA and provide a descriptive analysis of sales prices and apartment sizes across different locations in Copenhagen.
- Data indeholder 988 observationer. Størstedelen af observationerne kommer fra KbhØ
 - Godt argument for at bruge KbhØ som referencegruppe.
 - Vi ser at kvadratmeterpriserne varierer på tværs af beliggenheder (højest i KbhK og lavest i KbhN). Indikerer at geografisk placering kan have indflydelse på prisen.
 - Størrelserne på lejlighederne varierer dog også en del på tværs af beliggenheder (mindst i KbhN og størst i KbhK). Måske skyldes prisforskellene på tværs af geografisk placering blot sådanne forskelle. God ide at inkludere størrelse I vores model.

tabulate location				
location	Freq.	Percent	Cum.	
KBH K	223	22.57	22.57	
KBH N	229	23.18	45.75	
KBH V	138	13.97	59.72	
KBH Ø	398	40.28	100.00	
Total	988	100.00		

gen m2price=price/m2 tabstat m2price price m2 rooms, by(location) stat(mean)				
location	m2price	price	m2	rooms
KBH K	30137.7	2625185	90.44843	2.753363
KBH N	25475.39	1672342	66.44541	2.292576
KBH V	26976.55	2252893	85.92029	2.811594
KBH Ø	27497.29	2404902	89.65578	2.922111
Total	27551.88	2263596	83.9332	2.722672

5) Stataøvelser - opsamling

2. Assume model (1) satisfies MLR.1–MLR.5. Estimate model (1) by OLS and comment on the parameter estimates. How much of the variation in $\log(\text{prices})$ can the regression model explain? Is the sign of $\hat{\beta}_3$ surprising?

- Under MLR.1-5 er OLS estimatoren unbiased og konsistent. Hvis vi tror på antagelserne kan vi altså troværdigt estimere alt-andet-lige effekter.
- $\hat{\beta}_1$: En ekstra kvadratmeter øger den forventede pris med (approx.) 0,8 pct. (når størrelse, antal rum, antal toiletter og alt andet holdes fast).
- $\hat{\beta}_2$: Et ekstra rum øger den forventede pris med (approx.) 1,5 pct. (alt andet lige). Lille effekt og ikke statistisk signifikant. Skal ses i lyset af at størrelsen på lejligheden holdes konstant – måske mindre fordelagtigt at have flere små rum end færre store?
- $\hat{\beta}_3$: Et ekstra toilet reducere den forventede pris med (approx.) -12 pct. (alt andet lige). Igen, når størrelsen på lejligheden holdes konstant er et ekstra toilet måske ikke at foretrække. Desuden er der kun én lille del af sample der har mere end ét toilet (måske er disse lejligheder systematisk forskellige fra resten på uobserverede faktorer? Hvis dette er tilfældet og det samvarierer med de inkluderede variable vil det bryde med vores initiale MLR.4 antagelse)
- R^2 : Modellen forklarer 56,2% af variationen i $\log(\text{price}_i)$

```
gen logprice=log(price)
regress logprice m2 rooms toilets
outreg2 using results.tex, replace tex(frag) ctitle(Model 1)
```

VARIABLES	(1)
	Model 1
m2	0.00864*** (0.000458)
rooms	0.0152 (0.0140)
toilets	-0.122*** (0.0427)
Constant	13.90*** (0.0423)
Observations	988
R-squared	0.562
Standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

5) Stataøvelser - opsamling

3. We want to test if the parameters in model (1) are the same for Copenhagen K, N, V and Ø using a Chow test. What is the null and alternative hypothesis? How many restrictions are there? Use the test results to argue which model is the preferred one.

- Model (1) er givet ved

$$\log(\text{price}_i) = \beta_0 + \beta_1 m2_i + \beta_3 \text{rooms}_i + \beta_4 \text{toilets}_i + u_i$$

- Denne model restrikerer alle beliggenheder til at have identiske implicitte priser og skæring med y-aksen. Den urestrikerede model er givet ved:

$$\begin{aligned} \log(\text{price}_i) = & \beta_0 + \beta_1 m2_i + \beta_2 \text{rooms}_i + \beta_3 \text{toilets}_i + \delta_1 KbhK_i + \delta_2 KbhN_i + \delta_3 KbhV_i + \delta_4 (KbhK_i \times m2_i) + \delta_5 (KbhN_i \times m2_i) + \delta_6 (KbhV_i \times m2_i) \\ & + \delta_7 (KbhK_i \times \text{rooms}_i) + \delta_8 (KbhN_i \times \text{rooms}_i) + \delta_9 (KbhV_i \times \text{rooms}_i) + \delta_{10} (KbhK_i \times \text{toilets}_i) + \delta_{11} (KbhN_i \times \text{toilets}_i) + \delta_{12} (KbhV_i \times \text{toilets}_i) + v_i \end{aligned}$$

- Hypotesen for ingen forskelle i modellens parameter på tværs af beliggenhed er derfor:

$$\begin{aligned} H_0: & \delta_1 = \delta_2 = \dots = \delta_{12} = 0 \\ H_A: & \text{mindst en } \delta_j \neq 0, \quad j = 1, 2, \dots, 12 \end{aligned}$$

- Vi kan teste dette enten ved at bruge en **F-test** eller en **Chow-test**.
- F-testen sammenligner sum of squared residuals for model (1) og den fuldt urestrikerede model ovenfor.
- Chow-testen deler denne sammenligning op i bidder ved at estimere model (1) betinget på hver beliggenhed for derefter at summere disse.
- Begge test tjekker altså om forklaringsgraden (målt ved SSR) ændrer sig significant ved at gå fra restrikerede til urestrikerede model.

5) Stataøvelser - opsamling

- Chow-testen er formelt givet ved (valid under MLR.1-5 + $n \rightarrow \infty$):

$$Chow = \frac{\frac{SSR_p - (SSR_1 + \dots + SSR_m)}{(m-1)(k+1)}}{\frac{SSR_1 + \dots + SSR_m}{n - (m(k+1))}}$$

- Hvor SSR_p er sum of squared residuals for den restrikerede model (dvs. på hele sample), og SSR_i , $i \in KbhK, KbhV, Kbh\emptyset, KbhM$ er sum of squared residuals for model (1) estimerede for hver beliggenhed.

$$Chow = \frac{\frac{78,20 - (15,73 + 11,67 + 30,40 + 12,39)}{(4-1)(3+1)}}{\frac{15,73 + 11,67 + 30,40 + 12,39}{988 - (4(3+1))}} = 9,24$$

- Under H_0 følger $Chow$ en F-fordeling med $(m-1)(k+1)$, $n - (m(k+1))$ eller (12, 972) frihedsgrader. Den kritiske værdi for 5% signifikansniveau er 1,76.
- Vi forkaster derfor H_0 og finder empirisk belæg for at alle skæringer med y-aksen og implicitte priser ikke er identiske på tværs af beliggenheder (vi kan dog ikke med testen sige hvilke parametre der er forskellige)

```
regress logprice m2 rooms toilets if location=="KBH K"
gen SSR_1=e(rss)
```

```
regress logprice m2 rooms toilets if location=="KBH V"
gen SSR_2=e(rss)
```

```
regress logprice m2 rooms toilets if location=="KBH OE"
gen SSR_3=e(rss)
```

```
regress logprice m2 rooms toilets if location=="KBH N"
gen SSR_4=e(rss)
```

```
regress logprice m2 rooms toilets
gen SSR_p=e(rss)
```

```
su SSR*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
SSR_1	988	15.72656	0	15.72656	15.72656
SSR_2	988	11.67154	0	11.67154	11.67154
SSR_3	988	30.40185	0	30.40185	30.40185
SSR_4	988	12.38888	0	12.38888	12.38888
SSR_p	988	78.19934	0	78.19934	78.19934

```
gen chow=(SSR_p-(SSR_1+SSR_2+SSR_3+SSR_4))/(SSR_1+SSR_2+SSR_3+SSR_4)*((988-4*4)/(3*4))
gen c=invFtail(12,988-4*4,0.05)
su chow c
```

Variable	Obs	Mean	Std. Dev.	Min	Max
chow	988	9.244364	0	9.244364	9.244364
c	988	1.762126	0	1.762126	1.762126

5) Stataøvelser - opsamling

- Vi kan også finde frem til dette ved en F-test (valid under MLR.1-5 + $n \rightarrow \infty$):

$$F = \frac{\frac{(SSR_R - SSR_{UR})}{q}}{\frac{SSR_{UR}}{n - k - 1}}$$

- Hvor SSR_R er sum of squared residuals for den restrikerede model (model (1)) og SSR_{UR} er sum of squared residuals for den fuldt interagerede model.

$$F = \frac{\frac{78,20 - 70,19}{12}}{\frac{70,19}{988 - 15 - 1}} = 9,24$$

- Under H_0 følger F en F-fordeling med $q, n - k - 1$ eller (12, 972) frihedsgrader. Den kritiske værdi for 5% signifikansniveau er 1,76.
- Præcis samme statistik og konklusion som ved Chow-test

```

gen KbhK=1 if location=="KBH K"
replace KbhK=0 if KbhK==.

gen KbhN=1 if location=="KBH N"
replace KbhN=0 if KbhN==.

gen KbhV=1 if location=="KBH V"
replace KbhV=0 if KbhV==.

foreach x in KbhK KbhN KbhV {
    gen `x'm2=`x'*m2
    gen `x'rooms=`x'*rooms
    gen `x'toilets=`x'*toilets
    gen `x'sqm2=`x'*sqm2
}

regress logprice m2 rooms toilets KbhK KbhN KbhV KbhKm2 KbhNm2 KbhVm2 KbhKrooms ///
    KbhNrooms KbhVrooms KbhKtoilets KbhNtoilets KbhVtoilets

test KbhK KbhN KbhV KbhKm2 KbhNm2 KbhVm2 KbhKrooms KbhNrooms KbhVrooms KbhKtoilets ///
    KbhNtoilets KbhVtoilets

( 1) KbhK = 0
( 2) KbhN = 0
( 3) KbhV = 0
( 4) KbhKm2 = 0
( 5) KbhNm2 = 0
( 6) KbhVm2 = 0
( 7) KbhKrooms = 0
( 8) KbhNrooms = 0
( 9) KbhVrooms = 0
(10) KbhKtoilets = 0
(11) KbhNtoilets = 0
(12) KbhVtoilets = 0

F( 12, 972) = 9.24
Prob > F = 0.0000

```

5) Stataøvelser - opsamling

4. Write up an econometric model based on (1), which allows for differences in price levels across different locations in Copenhagen relative to Copenhagen Ø. Assume that apartment size and the number of rooms and toilets have the same effect on prices across locations. Estimate the model by OLS and interpret the regression results. Test the new regression model against the preferred model from question 3. State which model is now the preferred one.

- Modellen vi bliver bedt om at estimere er givet ved:

$$\log(\text{price}_i) = \beta_0 + \delta_1 \text{KbhK}_i + \delta_2 \text{KbhN}_i + \delta_3 \text{KbhV}_i + \beta_1 m2_i + \beta_2 \text{rooms}_i + \beta_3 \text{toilets}_i + \epsilon_i$$

- Lejligheder i KbhK har en forventet pris der (approx.) er $100\delta_1\% = 8,6\%$ højere end end KbhØ (når størrelse, antal rum, antal toiletter og alt andet holdes fast). For KbhN er den forventede pris ift. KbhØ (approx.) 15,1% lavere og i KbhV er den forventede pris (approx.) 4,5% lavere (alt andet lige). Alle estimater er signifikante på et 10% signifikansniveau.

- Modellen svarer til den fuldt interagerede model med følgende restriktioner:

$$H_0: \delta_4 = \dots = \delta_{12} = 0$$

$$H_A: \text{mindst en } \delta_j \neq 0, \quad j = 4, \dots, 12$$

- Vi tester den nye model (ovenfor) med den fuld interagerede (forrige slide) med en F-test: dette med en F-test med 9 restriktioner

```
regress logprice m2 rooms toilets KbhK KbhN KbhV KbhKm2 KbhNm2 KbhVm2 KbhKrooms ///
      KbhNrooms KbhVrooms KbhKtoilets KbhNtoilets KbhVtoilets
test KbhKm2 KbhNm2 KbhVm2 KbhKrooms KbhNrooms KbhVrooms KbhKtoilets KbhNtoilets KbhVtoilets
F( 9, 972) = 2.69
Prob > F = 0.0043
```

- Vi forkaster H_0 og konkluderer at de (implicitte) priser er forskellige på tværs af beliggenheder. Vi foretrækker derfor stadig den fuldt interagerede model.

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3
m2	0.00864***	0.00820***	0.00787***
rooms	0.0152	-0.00576	0.0230*
toilets	-0.122***	-0.0683	-0.109***
KbhK		0.217**	0.0858***
KbhN		-0.496**	-0.151***
KbhV		-0.0761	-0.0450*
KbhKm2		0.000176	
KbhNm2		-3.14e-05	
KbhVm2		0.000373	
KbhKrooms		0.00298	
KbhNrooms		0.106***	
KbhVrooms		-0.00139	
KbhKtoilets		-0.150	
KbhNtoilets		0.0954	
KbhVtoilets		0.00294	
Constant	13.90***	13.97***	13.96***
Observations	988	988	988
R-squared	0.562	0.607	0.597

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

5) Stataøvelser - opsamling

5. Extend the econometric model from the previous question to allow apartment size to have a different effect on prices across locations. Estimate the model by OLS and interpret the regression results. Test the extended regression model against the preferred model from question 4. State which model is now the preferred one.

- Ift. opg. 4 tillader vi nu også (implicitte) kvadratmeterpriser at variere på tværs af beliggenheder:

$$\log(\text{price}_i) = \beta_0 + \delta_1 \text{KbhK}_i + \delta_2 \text{KbhN}_i + \delta_3 \text{KbhV}_i + \beta_1 m2_i + \delta_4 (\text{KbhK}_i \times m2_i) + \delta_5 (\text{KbhN}_i \times m2_i) + \delta_6 (\text{KbhV}_i \times m2_i) + \beta_2 \text{rooms}_i + \beta_3 \text{toilets}_i + \eta_i$$

- $100\hat{\beta}_1\% = 0,77\%$ angiver (approx.) den forventede stigning i prisen når man øger lejlighedsstørrelsen med én kvadratmeter i KbhØ (når antal rum, toiletter og alt andet holdes fast).
- For en lejlighed i KbhN er dette $100(\hat{\beta}_1 + \hat{\delta}_5)\% = (0,77 + 0,29)\% = 1,06\%$. Dette er signifikant forskelligt fra KbhØ. Den forventede pris for lejligheder i KbhN er derfor overordnet set billigere end i KbhØ (lavere skæring med y-aksen), men den forventede prisstigning for én ekstra kvadratmeter er større end i KbhØ.
- De implicitte priser på ekstra kvadratmeter er ikke signifikant forskellige for KbhØ og KbhV/KbhK
- Vi tester den nye model mod den fuldt interagerede model ved

$$H_0: \delta_7 = \dots = \delta_{12} = 0$$

$$H_A: \text{mindst en } \delta_j \neq 0, \quad j = 7, \dots, 12$$
- Igen bruger vi en F-test, nu med 6 restriktioner:

`test KbhKrooms KbhNrooms KbhVrooms KbhKtoilets KbhNtoilets KbhVtoilets`

`F( 6, 972) = 2.11`
`Prob > F = 0.0494`

- Vi forkaster (marginalt) H_0 på et 5% signifikansniveau og finder empirisk belæg for at de implicitte priser for antal rum og toiletter er forskellige på tværs af beliggenheder. Vi foretrækker derfor stadig den fuldt interagerede model.

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4
m2	0.00864***	0.00820***	0.00787***	0.00765***
rooms	0.0152	-0.00576	0.0230*	0.0200
toilets	-0.122***	-0.0683	-0.109***	-0.0924**
KbhK		0.217**	0.0858***	0.0991*
KbhN		-0.496**	-0.151***	-0.352***
KbhV		-0.0761	-0.0450*	-0.0626
KbhKm2		0.000176		-0.000146
KbhNm2		-3.14e-05		0.00294***
KbhVm2		0.000373		0.000202
KbhKrooms		0.00298		
KbhNrooms		0.106***		
KbhVrooms		-0.00139		
KbhKtoilets		-0.150		
KbhNtoilets		0.0954		
KbhVtoilets		0.00294		
Constant	13.90***	13.97***	13.96***	13.97***
Observations	988	988	988	988
R-squared	0.562	0.607	0.597	0.602

Standard errors in parentheses

*** p<0.01. ** p<0.05. * p<0.1

5) Stataøvelser - opsamling

6. Model (1) assume that apartment prices depend linearly on apartment size, but this may be a restrictive assumption. You are therefore asked to include a quadratic term of $m2$ in your preferred model from question 5. Moreover, the model should allow the effect of the quadratic term to be different across location, while making the simplifying assumption that *rooms* and *toilets* have the same effect on sales prices across location. Test the new model with quadratic terms against a restricted model where all slope parameters are the same across locations in Copenhagen. Which model do you prefer in this case? What is the expected effect of increasing $m2$ by one unit on prices?

- Vi udvider nu modellen fra forrige spørgsmål med et kvadreret led af kvadratmeter ($m2_i$)², samt interaktioner af ($m2_i$)² med beliggenhederne:

$$\log(price_i) = \beta_0 + \delta_1 KbhK_i + \delta_2 KbhN_i + \delta_3 KbhV_i + \beta_1 m2_i + \delta_4 (KbhK_i \times m2_i) + \delta_5 (KbhN_i \times m2_i) + \delta_6 (KbhV_i \times m2_i) + \beta_2 rooms_i + \beta_3 toilets_i + \beta_4 \widetilde{m2_i^2} + \gamma_1 (KbhK_i \times \widetilde{m2_i^2}) + \gamma_2 (KbhN_i \times \widetilde{m2_i^2}) + \gamma_3 (KbhV_i \times \widetilde{m2_i^2}) + \psi_i$$

- For at gøre modellen lettere at læse har vi skaleret variablen med 1000:

$$\widetilde{m2_i^2} = \frac{m2_i^2}{1000}$$

- Det gør at de estimerede koefficienter nu er 1000 gange større (undgår at vi får output hvor de estimerede koefficienter fx er -3.67e-05 hvilket er svært at læse). Vi skal dog huske at dividere koefficienterne med 1000 hvis vi vil fortolke på ændringer på én kvadratmeter (se nedenfor).
- Modellen tillader for en ikke-lineær sammenhæng mellem implicite kvadratmeterpriser og forventet pris.
- Fx er den forventede ændring i prisen på en lejlighed i KbhK når denne øges én m^2 (for givent antal rum, toiletter og alt andet):

$$\left. \frac{\partial \log(price_i)}{\partial m2_i} \right|_{KbhK=1} = (\beta_1 + \delta_4) + 2(\beta_4 + \gamma_1) \frac{m2_i}{1000}$$

- Effekten af én ekstra kvadratmeter afhænger nu af størrelsen på lejligheden. Vi observerer at effekten initialt er positiv ($\widehat{\beta_1} + \widehat{\delta_4} > 0$) men aftagende ($\widehat{\beta_4} + \widehat{\gamma_1} < 0$).

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5
m2	0.00864***	0.00820***	0.00787***	0.00765***	0.0169***
rooms	0.0152	-0.00576	0.0230*	0.0200	0.00162
toilets	-0.122***	-0.0683	-0.109***	-0.0924**	-0.00172
KbhK		0.217**	0.0858***	0.0991*	-0.0198
KbhN		-0.496**	-0.151***	-0.352***	-0.339**
KbhV		-0.0761	-0.0450*	-0.0626	-0.202
KbhKm2		0.000176		-0.000146	0.00301
KbhNm2		-3.14e-05		0.00294***	0.00580
KbhVm2		0.000373		0.000202	0.00371
KbhKrooms		0.00298			
KbhNrooms		0.106***			
KbhVrooms		-0.00139			
KbhKtoilets		-0.150			
KbhNtoilets		0.0954			
KbhVtoilets		0.00294			
sqm2					-0.0367***
KbhKsqm2					-0.0175**
KbhNsqm2					-0.0337
KbhVsqm2					-0.0172*
Constant	13.90***	13.97***	13.96***	13.97***	13.44***
Observations	988	988	988	988	988
R-squared	0.562	0.607	0.597	0.602	0.658

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

5) Stataøvelser - opsamling

6. Model (1) assume that apartment prices depend linearly on apartment size, but this may be a restrictive assumption. You are therefore asked to include a quadratic term of $m2$ in your preferred model from question 5. Moreover, the model should allow the effect of the quadratic term to be different across location, while making the simplifying assumption that *rooms* and *toilets* have the same effect on sales prices across location. Test the new model with quadratic terms against a restricted model where all slope parameters are the same across locations in Copenhagen. Which model do you prefer in this case? What is the expected effect of increasing $m2$ by one unit on prices?

- Når effekten afhænger af den forklarende variabel (i dette tilfælde $m2_i$) bør man rapportér effekten for gennemsnitsstørrelsen eller medianstørrelsen eller andre relevante værdier.
- Desuden kan man, hvis den "lineære" og "ikke-lineære" del har forskelligt fortegn, også overveje at finde punktet hvor effekten går fra positiv til negative (eller omvendt). Dette punkt kan godt vise sig at ligge udenfor de værdier vi observerer og i så fald skal I passe på med tolke for meget på det.
- Indsætter vi de estimerede værdier for effekten af en ekstra $m2_i$ i KbhK (og husker at tage højde for at vi skalerede $m2_i^2$ med 1000) finder vi at:

$$\left. \frac{\partial \log(\text{price}_i)}{\partial m2_i} \right|_{KbhK=1} = (0,017 + 0,003) + 2 \left(-\frac{0,037}{1000} - \frac{0,018}{1000} \right) m2_i = 0,02 - 0,00011m2_i$$

- For en lejlighed på 90 kvadratmeter (gennemsnittet i KbhK) er den marginale effekt på den forventede pris ved én ekstra kvadratmeter altså 0,0101 (eller approx. 1,01%). For en lejlighed på 150 kvadratmeter er dette 0,0035 (eller approx. 0,35%). Som nævnt på forrige slide er der altså en aftagende marginaleffekt.
- Ønsker vi at finde punktet hvor effekten går fra positiv til negativ kan vi finde denne ved at løse:

$$0,02 - 0,00011m2_i = 0 \rightarrow m2_i = \frac{0,02}{0,00011} = 182$$

- Effekten på den forventede pris ved én ekstra kvadratmeter for lejligheder i KbhK der er større eller lig 182 m2 er altså negativ. Der er dog kun 6 lejligheder i KbhK i vores sample som er større end dette. Estimatet skal derfor ses med betydelig usikkerhed og ikke overfortolkes.

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5
m2	0.00864***	0.00820***	0.00787***	0.00765***	0.0169***
rooms	0.0152	-0.00576	0.0230*	0.0200	0.00162
toilets	-0.122***	-0.0683	-0.109***	-0.0924**	-0.00172
KbhK		0.217**	0.0858***	0.0991*	-0.0198
KbhN		-0.496**	-0.151***	-0.352***	-0.339**
KbhV		-0.0761	-0.0450*	-0.0626	-0.202
KbhKm2		0.000176		-0.000146	0.00301
KbhNm2		-3.14e-05		0.00294***	0.00580
KbhVm2		0.000373		0.000202	0.00371
KbhKrooms		0.00298			
KbhNrooms		0.106***			
KbhVrooms		-0.00139			
KbhKtoilets		-0.150			
KbhNtoilets		0.0954			
KbhVtoilets		0.00294			
sqm2					-0.0367***
KbhKsqm2					-0.0175**
KbhNsqm2					-0.0337
KbhVsqm2					-0.0172*
Constant	13.90***	13.97***	13.96***	13.97***	13.44***
Observations	988	988	988	988	988
R-squared	0.562	0.607	0.597	0.602	0.658

Standard errors in parentheses
 *** p<0.01, ** p<0.05, * p<0.1

5) Stataøvelser - opsamling

6. Model (1) assume that apartment prices depend linearly on apartment size, but this may be a restrictive assumption. You are therefore asked to include a quadratic term of $m2$ in your preferred model from question 5. Moreover, the model should allow the effect of the quadratic term to be different across location, while making the simplifying assumption that *rooms* and *toilets* have the same effect on sales prices across location. Test the new model with quadratic terms against a restricted model where all slope parameters are the same across locations in Copenhagen. Which model do you prefer in this case? What is the expected effect of increasing $m2$ by one unit on prices?

- Hypotesen om at de (implicitte) kvadratmeterpriser er identisk på tværs af beliggenheder er nu givet ved:

$$H_0: \delta_4 = \delta_5 = \delta_6 = \gamma_1 = \gamma_2 = \gamma_3 = 0$$

$$H_A: H_0 \text{ ikke sand}$$

- Vi benytter en F-test med 6 restriktioner og finder:

```
test KbhKm2 KbhNm2 KbhVm2 KbhKsqm2 KbhNsqm2 KbhVsqm2
F( 6, 974) = 1.89
Prob > F = 0.0793
```

- Da $p = 0,0793 > 0,05$ kan vi ikke forkaste H_0 på et 5% signifikansniveau. Vi kan derfor ikke afvise at modellen bør være givet ved:

$$\log(\text{price}_i) = \beta_0 + \delta_1 KbhK_i + \delta_2 KbhN_i + \delta_3 KbhV_i + \beta_1 m2_i + \beta_4 m2_i^2 + \beta_2 rooms_i + \beta_3 toilets_i + \kappa_i$$

- Effekten fra yderligere rum og yderligere toiletter i denne model er dog insignifikante, så vi kunne yderligere teste om koefficienterne på begge disse (samlet) er nul.

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
m2	0.00864***	0.00820***	0.00787***	0.00765***	0.0169***	0.0181***
rooms	0.0152	-0.00576	0.0230*	0.0200	0.00162	0.00542
toilets	-0.122***	-0.0683	-0.109***	-0.0924**	-0.00172	-0.000347
KbhK		0.217**	0.0858***	0.0991*	-0.0198	0.0835***
KbhN		-0.496**	-0.151***	-0.352***	-0.339**	-0.119***
KbhV		-0.0761	-0.0450*	-0.0626	-0.202	-0.0347
KbhKm2		0.000176		-0.000146	0.00301	
KbhNm2		-3.14e-05		0.00294***	0.00580	
KbhVm2		0.000373		0.000202	0.00371	
KbhKrooms		0.00298				
KbhNrooms		0.106***				
KbhVrooms		-0.00139				
KbhKtoilets		-0.150				
KbhNtoilets		0.0954				
KbhVtoilets		0.00294				
sqm2					-0.0367***	-0.0436***
KbhKsqm2					-0.0175**	
KbhNsqm2					-0.0337	
KbhVsqm2					-0.0172*	
Constant	13.90***	13.97***	13.96***	13.97***	13.44***	13.38***
Observations	988	988	988	988	988	988
R-squared	0.562	0.607	0.597	0.602	0.658	0.654

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Tak for i dag

- **Påskeferie:** Ingen undervisning næste gang (mandag d. 5 marts)
- **Anden aflevering** er blevet gjort tilgængelig på Absalon. Skal afleveres d. 19 april
- Husk igen at aflevere rapport (.pdf) og do-file (.txt)
- Undervisningen mandag d. 12 april er sat af til at arbejde på den.