

Økonometri I: Ugeseddel 6

Levi van Boekel

Uge 12, 2025

Opgave 1

1.1

MLR1 er opfyldt, idet model (1) er lineær i parametrene. MLR2 – der tilsiger uafhængighed mellem observationerne – er jf. opgavebeskrivelsen opfyldt, men det virker også rimeligt at antage, at der kan være en indbyrdes effekter (f.eks. at hvis en bolig når en bestemt pris, stiger alle boliger i nærheden i pris), taget i betragtning, at hele datasættet fra København. MLR3 (ingen perfekt multikollinearitet) er opfyldt, da ingen af regressorene har den samme værdi for alle observationerne, og heller ej kan dannes som lineære kombinationer af hinanden. MLR4 er straks mere kritisk; antagelsen om $E(u | x) = 0$ indebærer nemlig, at fejldet (som her f.eks. indeholder lokation, etage, antallet af lejligheder i bygningen, kvalitet af byggeriet, osv) er ukorreleret med de forklarende variable. Det vil altså sige, kvaliteten af et byggeri – bredt forstået – *ikke* må være korreleret med m^2 , antallet af værelser eller antallet af toiletter, hvilket virker usandsynligt. Bias afhænger så af, hvor stærkt samvariationen er mellem u og x , samt hvor stor en forklaringskraft den udeladte variabel har. MLR5 angiver, at variansen af fejldet skal være konstant (homoskedasticitet), hvilket også kan diskuteres. Det må da forekomme rimeligt at antage, at variansen for større lejligheder er større end for mindre.

1.2

Ved at konstruere tre dummy variable for hver bydel i København (og benytte Østerbro som referencegruppen) kan man udvide modellen til:

$$\begin{aligned}\log(\text{price}_i) = & \beta_0 + \delta_1 N + \delta_2 K + \delta_3 V + \beta_1 m^2 + \delta_4 N m^2 + \delta_5 K m^2 + \delta_6 V m^2 \\ & + \beta_3 \text{rooms} + \beta_4 \text{toilets} + u\end{aligned}$$

Hvor N, K, V er dummy variable for Nørrebro, København K og Vesterbro. Bemærk, at dummy variabelernes koefficienter for niveaueffekterne nu kan fortolkes som de absolutte forskelle i prisen,

relativt til referencegruppen, Østerbro. Interaktionsleddene, dvs. $\delta_4, \delta_5, \delta_6$ angiver i stedet den ... den marginale effekt af en ekstra kvadratmeter på boligprisen, relativt til den effekt, som en ekstra kvadratmeter har på Østerbro.

1.3

Ved at definere nulhypotese til at, $\mathcal{H}_0 : \delta_1 = \delta_2 = \delta_3 = 0$, dvs. at der er ingen forskelle relativt til Østerbro, kan vi teste $\mathcal{H}_0$ ved et F-test og afgøre, om den parametrene tilsammen er statistisk signifikante. Det samme kan gøres ved interaktionsleddene for at teste den marginale effekt af m^2 er ens for alle bydele, dvs. $\mathcal{H}_0 : \delta_4 = \delta_5 = \delta_6 = 0$.

Opgave 2

2.1

Vi ser, at datasættet indeholder 988 observationer. Der er flest fra Østerbro (ca. 40 pct.) og færrest fra Vesterbro (ca. 14 pct.). Observationerne fra København K og Nørrebro udgør ca. lige store dele af datasættet.

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	

Beskrivende statistisk er vist i nedenstående tabel, hvor den første observation udfra et område viser gennemsnit og den anden viser standardafvigelsen. Vi ser, at kvadratmeterprisen er højest i København K, men ligeså er standardafvigelsen (7342). Den laveste kvadratmeterpris findes på Nørrebro, hvor lejlighederne og standardafvigelsen på pris (og størrelse) er mindst. De største lejligheder findes i gns. I København K og Østerbro, hvor standardafvigelserne dog også er betydelige (ca. 38 m²). Hvis dataene antages at være normalfordelte, vil det altså, at omkring 68% af ejendommene ligger inden for en standardafvigelse fra gennemsnittet, dvs. lejlighederne 68 af ejendommene har en m² størrelse på ca [52, 128].

location	m2price	price	m2	rooms
-----+-----				
KBH K	30137.7	2625185	90.44843	2.753363

		7342.949	1064166	38.83919	1.113866
-----+-----					
KBH N		25475.39	1672342	66.44541	2.292576
		5733.5	611105.5	21.49914	.8870347
-----+-----					
KBH V		26976.55	2252893	85.92029	2.811594
		6837.385	947438.6	38.10915	1.258842
-----+-----					
KBH Ø		27497.29	2404902	89.65578	2.922111
		6785.383	995128	38.88949	1.138501
-----+-----					
Total		27551.88	2263596	83.9332	2.722672
		6876.651	991269.6	36.74623	1.123576

2.2

Regressionsresultaterne er vist nedenfor:

Tabel 1: Regressionsresultater for model (1)

(1)	
	log_price
m2	0.00864*** (0.000458)
rooms	0.0152 (0.0140)
toilets	-0.122** (0.0427)
_cons	13.90*** (0.0423)
<i>N</i>	988
<i>R</i> ²	0.562

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Vi ser, at modellen kan forklare ca. 56.2 pct. af variationen i $\log(\text{prices})$, samtidig med, at såvel m^2 og $rooms$ har en positiv effekt, mens $toilets$ har en negativ effekt. Da modellen er en log-level model svarer det til, at $\log(\text{prices})$ falder med -0.122 pct, når $toilets$ stiger med en enhed, når værelser og størrelse holder konstant.

2.3

Nullhypotesen er, parametrene er de samme for alle modeller, dvs.

$$\mathcal{H}_0 : \beta_{1,0} = \beta_{1,1} = \beta_{3,0} = \beta_{4,0}, \dots, \beta_{3,4} = \beta_{4,4}$$

Hvor alternativhypotesen er, at de ikke er ens:

$$\mathcal{H}_A : \beta_{1,0} \neq \beta_{1,1} \neq \beta_{3,0} \neq \beta_{4,0}, \dots, \beta_{3,4} \neq \beta_{4,4}$$

Vi starter med at notere $SSR_{all} = 78.1993433$ fra regressionen i 2.2. Herefter regresserer vi modellen for de forskellige bydele og får:

$$SSR = 29.5386436$$

$$SSR_N = 12.3888799$$

$$SSR_V = 11.6715445$$

$$SSR_K = 15.7265596$$

Dermed bliver test-statistikken:

$$F = \frac{78.1993433 - (30.401849 + 12.3888799 + 11.6715445 + 15.7265596)/(4-1)(3+1)}{(30.401849 + 12.3888799 + 11.6715445 + 15.7265596)/(988 - 4(3+1))} \approx 9.28$$

Ovenstående er F fordelt med $(4-1)(3+1)$ og $988 - 4(3+1)$ frihedsgrader. Dermed bliver det kritiske område:

$$\text{display invF}(12, 988-12, 1-0.05) = 1.762$$

Da test-statistikken er strengt over det kritiske område kan nullhypotesen afvises på et 5 pct. konfidensinterval. Dermed kan vi afvise, at der *ikke* er nogen forskel mellem parametrene på tværs af bydelene (dvs. de varierer).

2.4

Den nye model bliver blot:

$$\log(\text{price}_i) = \beta_0 + \delta_1 N + \delta_2 K + \delta_3 V + \beta_1 m^2 + \beta_3 \text{rooms} + \beta_4 \text{toilets} + u \quad (1)$$

Som er estimeret nedenfor:

Tabel 2: Regressionsresultater for model (2)

	(1)
	log_price
KbhK	0.0858*** (0.0228)
KbhV	-0.0450 (0.0268)
KbhN	-0.151*** (0.0232)
m2	0.00787*** (0.000448)
rooms	0.0230 (0.0135)
toilets	-0.109** (0.0411)
_cons	13.96*** (0.0436)
<i>N</i>	988
<i>R</i> ²	0.597

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Hvor *KbhV* og *KbhN* er mindre end 0 (omend V er insignifikant), svarende til, lejligheder her er hhv. 4.5 og 15.1 pct. billigere, for fastholdt størrelse, rum og toiletter. For *KbhK* er effekten omvendt; her er lejlighederne ca. 8.5 pct. dyrere. Den generelle model kan vi opfatte som en model, der indebærer en dummy variabel for alle led, dvs.

$$\begin{aligned}\log(\text{price}_i) = & \beta_0 + \delta_1 N + \delta_2 K + \delta_3 V + \beta_1 m2 + \delta_4 Nm2 + \delta_5 Km2 + \delta_6 Vm2 \\ & + \beta_3 \text{rooms} + \delta_7 \text{rooms}N + \delta_8 \text{rooms}K + \delta_9 \text{rooms}V \\ & + \beta_4 \text{toilets} + \delta_{10} \text{toilets}N + \delta_{11} \text{toilets}K + \delta_{12} \text{toilets}V + u\end{aligned}$$

Testen mellem de to modeller er nu, at alle led udover $\delta_1, \delta_2, \delta_3$ er lig 0. Altså:

$$\mathcal{H}_0 : \delta_4, \dots, \delta_{12} = 0$$

Mens alterantiv hypotesen er, at mindst en af parametrene *ikke* er lig nul.

$$\mathcal{H}_A : \delta_j \neq 0, j = 4, \dots, 12$$

Ovenstående er et almindeligt F-test med 9 restriktioner, der kan testes i STATA ved:

```
test KbhN_m2 KbhK_m2 KbhV_m2 KbhN_rooms KbhK_rooms KbhV_rooms KbhN_toilets KbhK_toilets KbhV_toilets

( 1)  KbhN_m2 = 0
( 2)  KbhK_m2 = 0
( 3)  KbhV_m2 = 0
( 4)  KbhN_rooms = 0
( 5)  KbhK_rooms = 0
( 6)  KbhV_rooms = 0
( 7)  KbhN_toilets = 0
( 8)  KbhK_toilets = 0
( 9)  KbhV_toilets = 0
```

```
F( 9, 972) = 2.69
Prob > F = 0.0043
```

Her fås, at F-test statistikken er 2.69, samt p -værdien er 0.0043. Det betyder, at sandsynligheden for at få et mere ekstremt udfald er 0.43 pct, hvilket er under vores signifikantniveau, på 5 pct. Dermed kan vi forkaste nullhypotesen, og fastholde, at parametrene for størrelse, antal rum og toiletter varierer på tværs af bydelene. Den generelle model foretrækkes dermed.

2.5

Vi udvider nu modellen til:

$$\log(\text{price}_i) = \beta_0 + \delta_1 N + \delta_2 K + \delta_3 V + \beta_1 m2 + \delta_4 Nm2 + \delta_5 Km2 + \delta_6 Vm2 + \beta_3 \text{rooms} + \beta_4 \text{toilets} + u$$

Source	SS	df	MS	Number of obs	=	988
-----+-----				F(9, 978)	=	164.27
Model	107.48795	9	11.9431055	Prob > F	=	0.0000
Residual	71.1043977	978	.072703883	R-squared	=	0.6019
-----+-----				Adj R-squared	=	0.5982
Total	178.592347	987	.180944628	Root MSE	=	.26964

log_price	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
-----+-----						
KbhK	.0990791	.057174	1.73	0.083	-.0131189	.211277
KbhV	-.0626264	.0665686	-0.94	0.347	-.19326	.0680073
KbhN	-.3519945	.0679719	-5.18	0.000	-.4853821	-.2186069
m2	.0076497	.0005283	14.48	0.000	.0066129	.0086864
KbhN_m2	.002936	.0009113	3.22	0.001	.0011478	.0047243
KbhK_m2	-.0001458	.0005855	-0.25	0.803	-.0012949	.0010032
KbhV_m2	.0002017	.0007041	0.29	0.775	-.0011801	.0015834
rooms	.0199746	.0135996	1.47	0.142	-.0067132	.0466624
toilets	-.0923734	.0415462	-2.22	0.026	-.1739032	-.0108435
_cons	13.96658	.0469168	297.69	0.000	13.87452	14.05865

Vi ser, at kun $KbhNm2$ er statistisk signifikant (p værdierne giver $0.001 < 0.05$, fordi STATA altid tester $\mathcal{H}_0 : KbhNm2 = 0$, hvilket kan afvises på 5 pct. signifikansniveau), og her angiver, at effekten af en ekstra kvadratmeter, for fastholdt antal rum og toilet, er 0.294 pct. højere på Nørrebro, end på Østerbro (selvom prisniveauet er lavere jf. fortegnet på $KbhN$ variabelen). Variablen $m2$ angiver effekten af en ekstra kvadratmeter på Østerbro.

Testen mellem de to modeller er nu, at alle led udover $\delta_1, \dots, \delta_6$ er lig 0. Altså:

$$\mathcal{H}_0 : \delta_7, \dots, \delta_{12} = 0$$

Mens alterantiv hypotesen er, at mindst en af parametrene *ikke* er lig nul.

$$\mathcal{H}_A : \delta_j \neq 0, j = 7, \dots, 12$$

Ovenstående er et almindeligt F-test med 6 restriktioner, der kan testes i STATA ved:

```
. test KbhN_rooms KbhK_rooms KbhV_rooms KbhN_toilets KbhK_toilets KbhV_toilets
```

```
( 1) KbhN_rooms = 0
```

```
( 2)  KbhK_rooms = 0
( 3)  KbhV_rooms = 0
( 4)  KbhN_toilets = 0
( 5)  KbhK_toilets = 0
( 6)  KbhV_toilets = 0
```

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

Vi ser, at p -værdien er 0.0494 svarende til, at sandsynligheden for at få et mere ekstremt udfald er 4.94 pct. Vi afviser derfor lige akkurat $\mathcal{H}_0$, da vi kræver, at sandsynligheden for at få et mere ekstremt udfald skal være større end signifikansniveauet på 5 pct. for at bibeholde nulhypotesen. Modellen med interaktionsled på kvadratmeterpris er altså foretrukket.

2.6

Modellen er nu:

$$\begin{aligned}\log(\text{price}_i) = & \beta_0 + \delta_1 N + \delta_2 K + \delta_3 V + \beta_1 m2 + \delta_4 N m2 + \delta_5 K m2 + \delta_6 V m2 \\ & + \beta_2 m2 + \delta_7 N m2^2 + \delta_8 K m2^2 + \delta_9 V m2^2 \\ & + \beta_3 \text{rooms} + \beta_4 \text{toilets} + u\end{aligned}$$

Som giver følgende estimat:

Source	SS	df	MS	Number of obs	=	988
-----+-----				F(13, 974)	=	143.95
Model	117.459257	13	9.03532748	Prob > F	=	0.0000
Residual	61.1330902	974	.06276498	R-squared	=	0.6577
-----+-----				Adj R-squared	=	0.6531
Total	178.592347	987	.180944628	Root MSE	=	.25053

log_price	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
-----+-----						
KbhK	-.0197653	.1119949	-0.18	0.860	-.2395443	.2000138
KbhV	-.2024935	.1272729	-1.59	0.112	-.4522542	.0472673
KbhN	-.3388977	.1701453	-1.99	0.047	-.6727913	-.0050041
m2	.0169196	.0013044	12.97	0.000	.0143599	.0194793
KbhN_m2	.0057962	.0042002	1.38	0.168	-.0024464	.0140388
KbhK_m2	.0030124	.0020817	1.45	0.148	-.0010727	.0070974

KbhV_m2		.0037141	.0023802	1.56	0.119	-.0009569	.0083851
m2_2		-.0000367	4.79e-06	-7.66	0.000	-.0000461	-.0000273
KbhN_m2_2		-.0000337	.0000242	-1.39	0.165	-.0000812	.0000139
KbhK_m2_2		-.0000175	8.58e-06	-2.05	0.041	-.0000344	-7.09e-07
KbhV_m2_2		-.0000172	9.76e-06	-1.76	0.079	-.0000363	1.99e-06
rooms		.0016181	.012869	0.13	0.900	-.023636	.0268723
toilets		-.0017218	.039438	-0.04	0.965	-.0791151	.0756715
_cons		13.44125	.0779203	172.50	0.000	13.28833	13.59416

Effekten af at øge m^2 med en kvadratmeter på Østerbro er:

$$\frac{\delta \log(\text{prices})}{\delta m2} = 0.0169196 + 2(-0.0000367) \cdot m2$$

Dvs. effekten varierer, alt afhængigt af, hvor stor den nuværende lejlighed er.

Hypotesen om, at alle hældningsparametre er de samme kan skrives som:

$$\delta_4 = \dots = \delta_9 = 0$$

Mens alternativ hypotesen er, at det ikke er tilfældet, dvs.

$$\delta_4 \neq \dots \neq \delta_9 \neq 0$$

Ovenstående kan testes ved et F-test med 6 restriktioner. Det er gjort i STATA nedenfor:

```
test KbhN_m2 KbhK_m2 KbhV_m2 KbhN_m2_2 KbhK_m2_2 KbhV_m2_2

( 1)  KbhN_m2 = 0
( 2)  KbhK_m2 = 0
( 3)  KbhV_m2 = 0
( 4)  KbhN_m2_2 = 0
( 5)  KbhK_m2_2 = 0
( 6)  KbhV_m2_2 = 0
```

```
F( 6, 974) = 1.89
Prob > F = 0.0793
```

Vi ser, at p-værdien nu er 0.0793, svarende til, at sandsynligheden for at få et mere ekstremt udfald er ca. 7.93 pct. Dermed kan nullhypotesen *ikke* afvises, $0.0793 > 0.05$. Vi foretrækker derfor

modellen, hvor $\mathcal{H}_A$ er opfyldt, dvs.

$$\log(\text{price}_i) = \beta_0 + \delta_1 N + \delta_2 K + \delta_3 V + \beta_1 m + \beta_2 m^2 + \beta_3 \text{rooms} + \beta_4 \text{toilets} + u$$