HD Dataanalyse

Note 10: Regression med FLERE forklarende variabel

Copenhagen Business School

Emne i dette notesæt

Når vi betragter ***en række forskellige variable*** i et datamateriale, er vi ofte interesseret i at vurdere, om der er en ***sammenhæng mellem dem*** eller ej.

Mere specifikt er vi interesseret i, hvorvidt én bestemt KVANTITATIV variabel kan forklares ud fra de øvrige variable.

I den simple lineære regressionsmodel (note 9) så vi på, hvordan én kvantitativ variabel Y kunne forklares ud fra én kvantitativ variabel X.

I denne note skal vi se på, hvordan én kvantitativ variabel Y kan forklares ud fra en hel række både kategoriske og kvantitative variable $X_1, \ldots, X_m$.

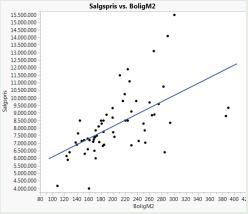
Idéen er, at hver af variablene $X_1, \ldots, X_m$ indeholder information, der er relevant for at forstå værdien af Y. Spørgsmålet vi skal se på er, hvordan vi på én og samme tid bedst kombinerer informationen fra alle variablene $X_1, \ldots, X_m$ til én samlet model, der kan forklare værdien af Y.

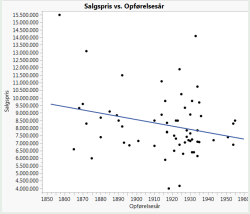
Emne i dette notesæt

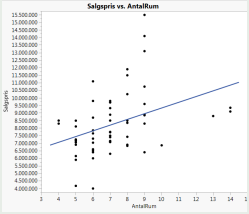
Eksempel: Boligpriser

Vi ser igen på datamaterialet med salgsinformation om boliger solgt i perioden 2014-2015. Salgsprisen på en bolig må formodes at afhænge af en række forskellige karakteristika så som beliggenhed, størrelse, alder etc.

Tegner vi figurer af boligens salgspris mod henholdsvis boligareal, opførelsesår og antal rum i boligen, så ser vi ikke overraskende, at alle tre variable (boligareal, opførelsesår, antal rum) hver især lader til at kunne forklare en del af boligens salgspris.







(vi har her igen for overskuelighedens skyld begrænset os til at se på villaer solgt på Frederiksberg).

Spørgsmålet er nu, hvordan vi kombinerer informationen fra alle tre variable til én samlet model, der kan forklare boligens salgspris?

Indholdsfortegnelse

Når vi...

- betragter en ***kvantitativ*** variabel Y
- betragter en række forskellige kvantitative eller kategoriske variable $X_1, \ldots X_m$
- formoder at variablen Y ***afhænger af*** variablene $X_1, \ldots, X_m$

vil vi gerne undersøge, hvordan sammenhængen mellem variablene $X_1, \ldots, X_m$ og Y helt præcist er.

Variablen Y kaldes for **responsvariablen**, og variablene $X_1, \ldots, X_m$ kaldes for de **forklarende variable**.

Hvis vi kun bruger én forklarende variable (f.eks. X_1) til at forklare værdien af responsvariablen Y med, har vi en simpel lineær regressionsmodel

$$Y = \beta_0 + \beta_1 \cdot X_1 + \varepsilon$$

Hvis vi bruger en hel række forklarende variable til at forklare værdien af Y med, har vi i stedet følgende model

$$Y = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_m \cdot X_m + \varepsilon$$

Vores model

$$Y = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_m \cdot X_m + \varepsilon$$

til forklaring af værdierne af variablen Y ud fra kendskab til værdierne af variablene $X_1, \ldots, X_m$ kalder vi for en **multipel lineær regressionsmodel**.

Logikken bag betegnelsen "multipel lineær regression" er følgende:

- Ordet ***regression*** henviser til, at modellen forsøger at forklare Y ud fra $X_1, \ldots, X_m$, dvs. forsøger at "føre Y tilbage til $X_1, \ldots, X_m$"
  ("regressere" betyder at "føre tilbage")
- Ordet ***lineær*** henviser til, at modellen forklarer Y ud fra en lineær sammenhæng med $X_1, \ldots, X_m$
- Ordet ***multipel*** henviser til, at modellen forklarer Y ud fra en hel række forskellige variable $X_1, \ldots, X_m$

I den multiple lineære regressionsmodel angiver...

- β_0 den forventede værdi af variablen Y i det tilfælde, hvor alle de forklarende variable $X_1, \ldots, X_m$ har værdien 0
- β_i hvor meget den forventede værdi af Y ændres, hver gang den i'te forklarende variabel X_i stiger med 1

På samme måde som i den simple lineære regressionsmodel bruger vi også her et residual ε til at beskrive den del af variationen i Y, som ikke kan forklares af variablene $X_1, \ldots, X_m$.

I den multiple lineære regressionsmodel

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

beskriver ε tilfældig/uforklarlig variation. Formelt set antager vi, at residualet ε er beskrevet ved en normalfordeling $N(0, \sigma)$.

Residualet har en forventet værdi på $E(\varepsilon) = 0$, således at $\beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_m \cdot X_m$ beskriver vores forventning til værdien af Y givet vores kendskab til $X_1, \ldots, X_m$, hvilket vi også skriver som

$$E(Y | X_1, \ldots, X_m) = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_m \cdot X_m$$

Standardafvigelsen σ beskriver hvor meget/lidt vi regner med, at værdierne af Y vil afvige fra den forventede værdi $\beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_m \cdot X_m$.

Når vi skal gætte på hvilken multipel lineær regressionsmodel, der på baggrund af datamaterialet bedst beskriver sammenhængen mellem $X_1, \ldots, X_m$ og Y, så foregår det på følgende måde:

- For hver af datamaterialets n observationer $i = 1, \ldots, n$ (dvs. for hver række i vores JMP-fil) har vi en værdi $X_{1,i}, \ldots, X_{m,i}$ og en værdi Y_i
- For hver af datamaterialets n observationer $i = 1, \ldots, n$ kan vi derfor beregne den kvadrerede afvigelse mellem variablen Y og modellen

$$\Big(Y_i - \big(\beta_0 + \beta_1 \cdot X_{1,i} + \ldots + \beta_m \cdot X_{m,i}\big)\Big)^2$$

- Som estimerede værdier for $\beta_0, \beta_1, \ldots, \beta_m$ vælger vi de værdier, som giver den mindste samlede afvigelse mellem datamateriale og modellen, dvs. de værdier som minimerer summen af alle de kvadrerede afvigelser

$$\sum_{i=1}^{n} \Big(Y_i - \big(\beta_0 + \beta_1 \cdot X_{1,i} + \ldots + \beta_m \cdot X_{m,i}\big)\Big)^2$$

- De estimerede værdier $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ angiver dermed den forventede værdi $\hat{\beta}_0 + \hat{\beta}_1 \cdot X_1 + \ldots + \hat{\beta}_m \cdot X_m$ der bedst beskriver sammenhængen mellem $X_1, \ldots, X_m$ og Y

Eksempel: Boligpriser

Når vi estimerer den rette linje, der bedst beskriver sammenhængen mellem boligpris (Y) og boligareal (X_1) og opførelsesår (X_2)

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \varepsilon$$

hvor ε er normalfordelt $N(0, \sigma)$, så finder vi at

$$\hat{\beta}_0 = 28.480.276 \qquad \hat{\beta}_1 = 19.322 \qquad \hat{\beta}_2 = -12.643 \qquad \hat{\sigma} = 1.665.129$$

Fortolkningen af de tre parametre er, at...

- for hver 1 m^2 større boligarealet (X_1) er, så stiger den forventede salgspris med 19.322 kr.
- for hver 1 år nyere boligen er (X_2), så falder den forventede salgspris med 12.643 kr.
- salgsprisen på en bolig vil med ca. 95% sandsynlighed variere med $\pm 2 \cdot \hat{\sigma} = 3.330.258$ kr.

I en multipel regressionsanalyse har $\hat{\beta}_0$ ofte ikke nogen naturlig fortolkning, idet den svarer til tilfældet, hvor alle de forklarende variable (her X_1, X_2) antager værdien 0 (dvs. i denne model en bolig med 0 m^2 boligareal opført i år 0).

Effect Summary

Source	LogWorth	PValue
BoligM2	5,660	0,00000
Opførelsesår	0,778	0,16678

Remove Add Edit FDR

Summary of Fit

RSquare	0,361332
RSquare Adj	0,339683
Root Mean Square Error	1665129
Mean of Response	8234466
Observations (or Sum Wgts)	62

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	9,2551e+13	4,628e+13	16,6899
Error	59	1,6359e+14	2,773e+12	Prob > F
C. Total	61	2,5614e+14		<,0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	28480276	17454804	1,63	0,1081
BoligM2	19322,031	3681,28	5,25	<,0001*
Opførelsesår	-12643,22	9031,589	-1,40	0,1668

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
BoligM2	1	1	7,6384e+13	27,5491	<,0001*
Opførelsesår	1	1	5,4335e+12	1,9597	0,1668