LTH: Fastighetsekonomi 23-24 sep 2008

Enkel och multipel linjär regressionsanalys

HYPOTESPRÖVNING

Hypotesprövning (statistisk inferensteori)

Statistisk hypotesprövning innebär att man med hjälp av slumpmässiga urval bedömer trovärdigheten i *hypoteser* – antaganden – angående populationen.

Inferens - hypotesprövning
Testa hypotesen att $\beta_1 = 0$ mot någon av mothypoteserna $\beta_1 \neq 0$, $\beta_1 < 0$ eller $\beta_1 > 0$.

Tre alternativ för statistisk hypotesprövning:

1. *Klassisk hypotesprövning*
2. *Konfidensintervallmetoden*
3. *p-värdesmetoden.*

Excel ”rapporterar” alla tre alternativen. Men värdena för alternativ 2 och 3 som presenteras i Excel är det för tvåsidiga hypotestester. För ensidiga tester rekommenderas att använda alternativ 1.

Låt oss välja *signifikansnivå 5%* (vanligt).

Tvåsidig hypotesprövning

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$.

Alt 1 – klassisk hypotesprövning
Nollhypotesen förkastas om absolutbeloppet av den observerade *t*-kvoten (värdet på testfunktionen) är större än det kritiska värdet hämtat från *t*-tabell:

H_0 förkastas om

$$|t| = \left| \frac{b_1}{s_{b1}} \right| > t_{0,025}(n-(k+1))$$

Tumregel **(mer eller mindre fin/grov) för att förkasta noll-hypotes:**
t-kvoten ska vara > 2 för "tillräckligt" många observationer (jmf frihetsgrader)

Språkbruk (gäller alla tester):
Om H_0 *inte* kan förkastas säger man "vi kan inte förkasta H_0 på 5% signifikansnivå" (men vi säger aldrig att "H_0 accepteras på 5% signifikansnivå".)

Man kan också säga att variabeln "x_1 är inte statistisk signifikant".

Om H_0 däremot *kan* förkastas säger man att variabeln "x_1 är statistisk signifikant".

Alt 2 – konfidensintervallmetoden
Studera om det 95-procentiga konfidensintervallet innehåller talet 0. H_0 kan förkastas om 0 inte ingår i intervallet.

Alt 3 – p-värdesmetoden
H_0 förkastas om *p*-värdet < 0,05.

Tillämpning av hypotesprövning på enkel linjär regressionsanalys.

Exempel 1: samband huspriser och ålder.

Problem:
Kalle P hävdar att det inte finns något *kasualt samband* mellan ålder (här definierad som försäljningsår minus värdeår) och huspriser för hus sålda i Norra och Södra Ängby.

Lena Q hävdar å andra sidan att det finns ett sådant samband, dock osäker på om ålder har en negativ eller positiv inverkan på husspriser.

Urval: 104 försäljningar ett visst år.

Enkel linjär regressionsmodell:

Populationen:
$pris = \beta_0 + \beta_1 * \text{ålder} + u$

Vill skatta β_0 och β_1 med urval (104 sålda hus):

Skattad linje baserad på urvalet (OBS! Annan notation för skattad parameter):

$pris = b_0 + b_1 * \text{ålder}$

Tvåsidig hypotesprövning:

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$.

Regressionsstatistik	
Multipel-R	0,080
R-kvadrat	0,006
Justerad R-kvadrat	-0,003
Standardfel	1207123,733
Observationer	104

ANOVA

	fg	*KvS*	*MKv*	*F*	*p-värde för F*
Regression	1	9,69633E+11	9,69633E+11	0,665431961	0,416549631
Residual	102	1,48629E+14	1,45715E+12		
Totalt	103	1,49599E+14			

	Koefficienter	*Standardfel*	***t-kvot***	*p-värde*	*Nedre 95%*	*Övre 95%*
Konstant	3 525 628	1010673	3,49	0,00	1520964	5530293,361
Ålder	**-13 376**	**16397**	**-0,82**	0,42	-45900	19148,2793

$$\boldsymbol{pris = b_0 + b_1 * ålder = 3\ 525\ 628 - 13\ 376 * Ålder}$$

$$t = \frac{b_1}{s_{b1}} = \frac{-13376}{16397} = -0,82$$

Vi kan direkt konstatera att absolutbeloppet av det observerade värdet på t-kvoten (0,82) är mindre än det kritiska värdet 2 (1,98), vilket innebär att vi *inte* kan förkasta nollhypotesen på 5% signifikansnivå.

Resultatet är *icke*-signifikant.

Med andra ord, vi konstaterar att ålder inte har någon inverkan på försäljningspris.

Exempel 2: samband huspriser och värdeyta

Problem: Undersöka om det finns ett *kasualt samband* mellan värdeyta och huspriser för hus sålda i Norra och Södra Ängby.

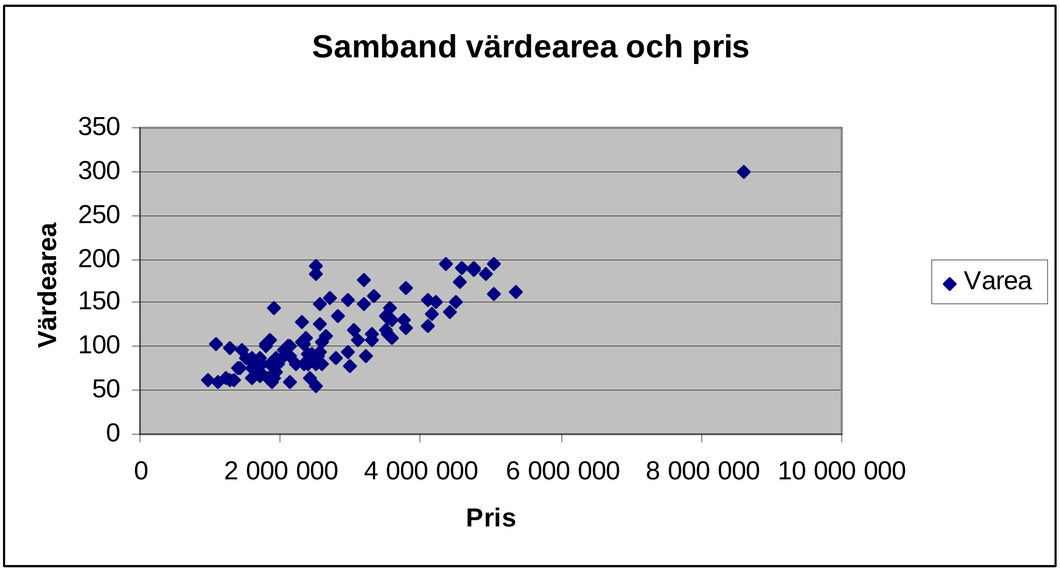
Urval: 104 försäljningar ett visst år (2000).

Enkel linjär regressionsmodell:

Populationen:
$pris = \beta_0 + \beta_1 * Värdeyta + \varepsilon$

Vill skatta β_0 och β_1 med urval (104 sålda hus):

Skattad linje baserad på urvalet:
$pris = b_0 + b_1 * Värdeyta$



1. Formulera hypoteser

Enkelsidig hypotesprövning:

H_0: $\beta_1 = 0$
H_1: $\beta_1 > 0$.

$$t = \frac{b_1}{s_{b1}}$$

Regressionsstatistik	
Multipel-R	0,827
R-kvadrat	0,683
Justerad R-kvadrat	0,680
Standardfel	681504,512
Observationer	104,000

ANOVA

	fg	*KvS*	*MKv*	*F*	*p-värde för F*
Regression	1	1,02E+14	1,02E+14	2,20E+02	3,22E-27
Residual	102	4,74E+13	4,64E+11		
Totalt	103	1,50E+14			

§

	Koefficienter	*Standardfel*	*t-kvot*	*p-värde*	*Nedre 95%*	*Övre 95%*
Konstant	112665	187195	0,60	0,55	-258636	483966
Värdearea	**23 416**	**1578**	**14,84**	0,00	20285	26546

pris = b_0 + b_1*värdearea =112 665 + 23 416*Värdearea

$$t = \frac{b_1}{s_{b1}} = \frac{23416}{1578} = 14,84$$

Resultatet är *signifikant* (t-kvoten är större än 2). Nollhypotesen förkastas därmed på 5% signifikansnivå till förmån för mothypotesen. Med andra ord, vi konstaterar att *värdearea* har inverkan på försäljningspris.

OBS! Vi har ovan inte studerat

- hur du slår upp kritiska värden i en t-tabell mht antal frihetsgrader.
- Vi har heller inte i detalj studerat enkelsidiga- och tvåsidiga hypotesprövningar.
- Etc.