Övningshäfte till kursen Regressionsanalys och tidsserieanalys

Linda Wänström

October 31, 2010

1 Enkel linjär regressionsanalys (baserad på uppgift 2.3 i Andersson, Jorner, Ågren (2009))

Antag att följande statistiska modell kan användas för att beskriva sambandet mellan veckoförsäljning och arbetstimmar per vecka: $Y = \beta_0 + \beta_1 X + E$. För en varuhuskedja har man samlat in data om försäljning och arbetstimmar per vecka och fått uppgifterna nedan.

1. Vilka variabler motsvarar, *troligtvis*, variablerna Y och X i modellen ovan? Motivera.

2. Rita ett spridningsdiagram mellan veckoförsäljning och arbetstimmar per vecka. Kommentera ett eventuellt samband.

Varuhus nr	Veckoförsäljning (1000-tals kr)	Arbetstimmar per vecka
1	180	170
2	210	190
3	165	170
4	300	200
5	120	160
6	240	220

3. Skatta β_0 och β_1 i modellen ovan. Tolka skattningarna.

4. Rita in den skattade linjen i spridningsdiagrammet.

5. Testa, på 5% signifikansnivå, om det finns ett samband mellan veckoförsäljning och arbetstimmar per vecka, d.v.s. testa $H_0 : \beta_1 = 0$. Tolka resultatet.

6. Diskutera skillnaden mellan korrelation och kausalitet. Kan du anta att X påverkar Y i det här fallet? Diskutera kritiskt.

7. Skatta veckoförsäljningen för ett varuhus med 180 arbetstimmar per vecka.

8. Beräkna ett 95%-igt prediktionsintervall runt din skattning i 7. ovan. Tolka intervallet.

9. Diskutera skillnaden mellan ett konfidensintervall och ett prediktionsintervall.

2 Korrelationskoefficienten (baserad på uppgift 2.3 i Andersson, Jorner, Ågren (2009)) forts. från uppgift 1 ovan

1. Beräkna stickprovskorrelationskoefficienten r mellan veckoförsäljning och arbetstimmar per vecka för materialet i uppgift 1 ovan.

2. Testa $H_0 : \rho = 0$ på 5% signifikansnivå. Tolka resultatet och jämför med uppgift 1.5 ovan.

3 Multipel regressionsanalys

Vi är intresserade av att undersöka relationen mellan konsumtion av lösgodis och pris för ett visst märke. Från 20 st områden samlar vi in uppgifter om konsumtion (mätt i mängden sålt lösgodis under en specifik vecka) samt pris. Vi har dessutom samlat in uppgifter om storleken på befolkningen i respektive område samt huruvida det finns någon affär i närheten som säljer lösgodis av ett annat märke. Vi antar följande modell: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 +$

Beroendevariabel: y

Antal lästa observationer	20
Antal använda observationer	20

Variansanalys

Källa	DF	Summa av kvadrater	Medel-kvadrat	F-värde
Modell		695.2577		
Fel				
Korrigerad total		1268.55		

E där Y motsvarar konsumtion (i kg), X_1 motsvarar hektopris (i kr.), X_2 motsvarar folkmängd (i 1000-tal) och X_3 är kodad 1 om det finns en affär inom en radie av 500 meter som säljer lösgodis och 0 annars. SAS-utskrifter från en regressionsanalys finns ovan.

1. Fyll i de uppgifter som saknas i SAS-utskriften (ANOVA-tablån).
2. Hur stor variation i Y kan förklaras med hjälp av X_1, X_2 och X_3? (tips: beräkna R^2)
3. Vilka statistiska antaganden bygger modellen ovan på?

4 Hypotestest forts. från uppgift 3 ovan.

1. Testa om modellen i uppgift 3 ovan som helhet är signifikant. Använd $\alpha = 0.05$. (tips: testa $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$)
2. Testa, på 5% signifikansnivå, om hektopriset behövs i modellen, d.v.s. testa om X_1 bidrar till att skatta Y givet att X_2 och X_3 finns med i modellen. (tips: genomför ett t-test). Använd SAS-utskriften på nästa sida till hjälp.
3. Beräkna ett 95%-igt konfidensintervall för den sanna parametern β_1. Tolka intervallet.

Parameterskattningar

Variabel	DF	Parameter-skattning	Standard-fel	t-värde	Pr > |t|
Skärning	1	60.1473	3.56479		
x1	1	-4.2189	0.71566		
x2	1	7.59239	1.70160		
x3	1	-10.580	3.54449		

5 Hypotestest

En mäklare vill planera försäljningen av villor. Hon är främst intresserad av sambandet mellan försäljning och annonsvolym. Hon samlar in uppgifter om försäljning, Y (milj. kr), folkmängd, X_1 (100 000 pers.) och annonsvolym, X_2 (10 000 kr) i 11 distrikt. Hon funderar på två möjliga modeller:

Modell 1: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$
Modell 2: $Y = \beta_0 + \beta_1 X_1 + E$

På nästa sida finns en SAS-utskrift från analys av modell 1.

1. Fyll i det som saknas i utskriften.
2. Testa, på 5% signifikansnivå, om modell 1 som helhet är signifikant.
3. Testa, på 5% signifikansnivå, om annonsvolym bidrar till att förklara variation i försäljning, utöver bidraget från befolkning.

Modell 1
Beroendevariabel: y

Antal lästa observationer	11
Antal använda observationer	11

Variansanalys

Källa	DF	Summa av kvadrater	Medel-kvadrat	F-värde	Sh. > F
Modell	?	?	?	28.04	0.0002
Fel	?	?	1.060039		
Korrigerad total	?	67.92727			

Rot MSE	?	R-kvadrat	?
Beroende medel	5.74545		
Koeff.var.	17.91927		

Parameterskattningar

Variabel	DF	Parameter-skattning	Standard-fel	t-värde	Pr > |t|
Skärning	1	?	0.78426	0.86	0.4137
x1	1	0.68849	?	2.23	0.0561
x2	1	0.34150	0.36737	?	0.3798

6 Korrelationer (baserad på exempel 3.2 i Andersson, Jorner, Ågren (2009))

På nästa sida finns uppgifter om oljeförbrukning (Y), medeltemperatur (X_1) och bostadsyta (X_2) för ett stickprov av villaägare under 10 månader. Där visas även en korrelationsmatris för variablerna.

1. Kommentera korrelationerna i korrelationsmatrisen.

2. Rita ett spridningsdiagram mellan oljeförbrukning och medeltemperatur. Kommentera diagrammet.

3. Beräkna, för hand, stickprovskorrelationen mellan oljeförbrukning och medeltemperatur och verifiera att den stämmer med korrelationsmatrisen pånästa sida.

4. Testa, på 5% signifikansnivå $H_0 : \rho_{Y1} = 0$. Jämför ditt resultat med motsvarande p-värde i korrelationsmatrisen.

Månad	Oljeförbrukning (liter)	Medeltemperatur (Celsius)	Bostadsyta (kvm)
Jul	70	17.8	170
Aug	100	16.6	210
Sep	185	12.2	150
Okt	300	7.1	190
Nov	310	2.8	110
Dec	650	0.1	250
Jan	525	-2.9	140
Feb	640	-3.1	155
Mar	550	-0.7	180
Apr	275	4.4	130

	Y	X_1	X_2
Y	1	-0.928 $p = 0.000$	0.178 $p = 0.623$
X_1		1	0.151 $p = 0.677$
X_2			1

7 Korrelationer (baserad på exempel 3.2 i Andersson, Jorner, Ågren (2009)) forts. från uppgift 6.

För materialet i uppgift 6 har man efter en regressionsanalys i SAS fått följande skattade ekvation: $\widehat{Y} = 219.37 - 27.23X_1 + 1.72X_2$.

1. Beräkna den multipla korrelationskoefficienten R. (tips: skapa en kolumn med $\widehat{Y}$ för varje observation i tabellen i uppgift 6 ovan)
2. Förklara vad R mäter, d.v.s. vad är det för korrelation som mäts?
3. Hur stor variation i Y kan förklaras med hjälp av X_1 och X_2?

8 Dummyvariabler

1. Ge ett exempel på en kategorisk variabel med två kategorier.
2. Definiera en eller flera dummyvariabler som kan användas om du vill ha med din kategoriska variabel (i 1 ovan) i en regressionsmodell.
3. Ge ett exempel på en kategorisk variabel med minst tre kategorier.
4. Definiera en eller flera dummyvariabler som kan användas om du vill ha med din kategoriska variabel (i 3 ovan) i en regressionsmodell.

9 Dummyvariabler

Du har antagit följande modell: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z + \beta_4 X_1 Z + E$ där Y=pris (1000-tals kr), X_1=yta (kvm), X_2=avgift (1000-tals kr) och Z=ort (Hammarby Sjöstad = 1, Haninge = 0). Du har samlat in uppgifter från Hemnet.se om lägenheter och genomfört en regressionsanalys (se SAS-utskrift på nästa sida).

1. Skriv modellen som två modeller, en för Hammarby Sjöstad och en för Haninge.
2. Är modellen (totala) som helhet signifikant? Testa på 5% signifikansnivå.

Dependent Variable: pris

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	99639481.8	24909870.4		
Error	44				
Corrected Total	48	107843713.4			

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-233.6921041	503.8277915		
yta	6.6037860	9.4855847		
avgift	185.0192623	126.0260655		
ort	-296.6078443	603.2678530		
yta*ort	30.2291286	7.1213234		

3. Hur många lägenheter har ingått i analysen?
4. Hur stor variation i pris förklaras av de oberoende variablerna?
5. Hur mycket förväntas priset öka per kvm i Hammarby Sjöstad om avgiften är konstant? I Haninge?
6. Skiljer sig sambandet mellan pris och yta åt i Hammarby Sjöstad och i Haninge? Genomför ett lämpligt hypotestest på 1% signifikansnivå.
7. Du tror att sambandet mellan pris och avgift ser olika ut i Hammarby Sjöstad och Haninge. Lägg till en term till modellen ovan så att det blir möjligt att testa detta.

10 Polynomregression (baserad på 4.1 i Andersson, Jorner, Ågren (2009))

Skissera, för $X = 0, 1, ..., 5$, utseendet av följande funktioner:

1. $\widehat{Y} = 5 + X - 0.2X^2$
2. $\widehat{Y} = 5 - 2X + 0.5X^2$

11 Polynomregression (baserad på 4.2 i Andersson, Jorner, Ågren (2009))

En butiksinnehavare i USA är intresserad av sambandet mellan antal reklamtillfällen per dag över det lokala radionätet (X) och den dagliga omsättningen, i dollar, av en viss vara (Y). Följande observationer föreligger:

X	4	5	6	7	8	9	10
Y	780	790	810	850	900	980	1100

Ett linjärt och ett kvadratiskt samband har anpassats till de givna observationerna med nedanstående resultat:

i) $\widehat{Y} = 529.6 + 51.1X$; $SSY = 82342.86$, $SSE = 9319.71$, $s_{\widehat{\beta}_X} = 8.16$
ii) $\widehat{Y} = 995.7 - 93.9X + 10.36X^2$; $SSE = 300.00$, $s_{\widehat{\beta}_X} = 13.33$, $s_{\widehat{\beta}_{X^2}} = 0.95$

1. Rita ett spridningsdiagram mellan X och Y. Kommentera diagrammet.
2. Testa, på 5% signifikansnivå, om andragradsmodellen som helhet är signifikant.
3. Testa, på 5% signifikansnivå, om den kvadratiska termen behövs i modellen (ii).
4. Välj en av modellerna utifrån ditt test i 3. ovan. Beräkna förklaringsgraden för denna modell. Kommentera kritiskt.

12 Logistisk regression

Antag att du vill undersöka sambandet mellan Y=betyg på sluttenta (godkänd = 1, underkänd = 0), och X_1=GPA (Grade Point Averge = genomsnittsbetyg vid kursstart), X_2=poäng på mitterminstenta och X_3=inlärningsmetod (1 = ny, 0 = gammal).

1. Skriv upp en lämplig modell.

2. Antag att du har genomfört en analys och fått följande skattningar: $\widehat{Y} = -13.02 + 2.83X_1 + 0.0951X_2 + 2.3786X_3$. Beräkna sannolikheten att en student som använder den nya inlärningsmetoden, har GPA=3 samt 20 poäng på mitterminstenta får godkänt på sluttentan.

3. Beräkna sannolikheten att en student som inte använder den nya inlärningsmetoden, har GPA=3 samt 20 poäng på mitterminstenta får godkänt på sluttentan.