

## Tentamen *Tillämpad statistik A5 (15hp) 2014-08-26*

### UPPLYSNINGAR

**A. Tillåtna hjälpmedel:**

Miniräknare

Formelsamlingar: *A4/A8 Tabell- och formelsamling samt  
A5 Kompletterande Tabell- och formelsamling.*  
Inga anteckningar är tillåtna i formelsamlingarna.

- B. Skrivtid: 8<sup>00</sup>-13<sup>00</sup> Skrivningen omfattar 6 uppgifter om sammanlagt 100 poäng.
- C. För varje uppgift anges den maximala poäng som kan erhållas. Om en uppgift är uppdelad på deluppgifter anges den maximala poängen för varje deluppgift. Ibland kan inte deluppgifterna bedömas oberoende av varandra, vilket kan innebära att poäng inte utdelas på en senare uppgift om inte tidigare deluppgift lösts på ett i princip riktigt sätt. Dock gäller att utdelad poäng för varje deluppgift aldrig kan vara negativ.
- D. Om Du känner Dig osäker på någonting (skrivningens genomförande, någon formulering i en uppgift, om något hjälpmedel är otillåtet), fråga då jourhavande skrivningsvakt eller den skrivningsansvariga läraren (besök, alternativt telefon).
- E. Efter skrivningens slut får Du behålla sidorna med frågeställningarna. Preliminära lösningar anslås på Pingpong.

### UPPMANINGAR

- A. Följ noga de anvisningar som finns på skrivningsförsättsbladet.
- B. Redovisa Dina lösningar i en form som gör det lätt att följa Din tankegång! (Det dunkelt uttryckta förutsätts av rättaren vara dunkelt tänkt). Motivera alla väsentliga steg i lösningen. Ange alla antaganden Du gör och alla förutsättningar Du utnyttjar.
- C. Vid konfidensintervall måste Du dessutom ange vad intervallet avser att täcka samt teckna intervallet i symbolform innan de numeriska uppgifterna insätts. Verbal slutsats av det framräknade intervallet krävs för full poäng.
- D. Vid hypotestest måste Du utöver vad som sägs i punkt B ovan ange  $H_0$ ,  $H_1$ , signifikansnivå, testfunktion (inklusive antal frihetsgrader), förkastelseområde, resultat och verbal slutsats.
- E. Vid variansanalys måste Du utöver vad som sägs ovan ange modell.

### Uppgift 1 (26)

- A. För nedanstående fem observationspar beräkna  $a$  och  $b$  i regressionslinjen  $\hat{Y} = a + bX$  med minsta-kvadratmetoden.
- B. Beräkna residualerna och residualkvadratsumman (SSE).
- C. Beräkna den totala variationen i  $Y$  (SST).
- D. Beräkna förklaringsgraden.
- E. Beräkna standardfelet för skattningen  $b$  och ange förutsättningarna för detta.
- F. Om du i nedanstående material utesluter observation ett eller observation fem eller båda dessa kommer du att få samma regressionslinje som du får om du använder alla fem observationerna. Varför?

### Data Display

$x$	$y$	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$	$x \cdot y$	$x^2$
1	6	-2	-2	4	4	6	1
2	9	-1	1	-1	1	18	4
3	4	0	-4	0	0	12	9
4	11	1	3	3	1	44	16
5	10	2	2	4	4	50	25
<b>15</b>	<b>40</b>	<b>0</b>	<b>0</b>	<b>10</b>	<b>10</b>	<b>130</b>	<b>55</b>

### Uppgift 2 (12)

Nedan presenteras tre skattade modeller för att förklara variationen i en given uppsättning av observationer på en och samma  $Y$ -variabel tillsammans med tre antaganden.

$$\text{Skattad modell 1: } Y_i = a + bX_{1i} + e_{1i}.$$

$$\text{Skattad modell 2: } Y_i = a_2 + b_{12}X_{1i} + b_{22}X_{2i} + e_{2i}.$$

$$\text{Skattad modell 3: } Y_i = a_3 + b_{13}X_{1i} + b_{33}X_{3i} + e_{3i}.$$

Antagande 1:  $X_2$  är starkt korrelerad med  $X_1$  och okorrelerad med  $e_1$ .

Antagande 2:  $X_3$  är okorrelerad med  $X_1$  och har en korrelation = 0,71 med  $e_1$ .

Antagande 3: Förklaringsgraden är 50% vid skattning av modell 1.

- A. Givet ovanstående hur stor blir förklaringsgraden vid skattning av modell 2. Hur blir skattningen av variansen för  $b_{12}$  jämfört med skattningen av variansen för  $b$ ?
- B. Givet ovanstående hur blir förklaringsgraden vid skattning av Modell 3 jämfört med skattningen av Modell 1? Hur blir skattningen av variansen för  $b_{13}$  jämfört med skattningen av variansen för  $b$ ?

### Uppgift 3 (14)

Denna uppgift baseras på kvartalsdata för andel av antal sysselsatta i Sverige som utgörs av kvinnor 1994:1 t o m 2011:4. Fram till och med första kvartalet 2005 omfattar antal sysselsatta åldersgruppen 16-64 år, därefter åldersgruppen 15-74 år.

- A. Finns någon anledning att tro att effekten av definitionsförändringen är olika för män och för kvinnor? Specificera den använda regressionsmodellen och genomför ett formellt test av frågeställningen på 5%-nivån med hjälp av nedanstående regressionsutskrift.
- B. Är förutsättningarna för testet i A-uppgiften uppfyllda. Kommentera kortfattat.

Variabeldefinitioner:

*Andel sysselsatta* är andel av antal sysselsatta i Sverige som utgörs av kvinnor 1994:1 t o m 2011:4.

*tid*: 1, 2, 3, ..., 72

*Dummy(15-74)* = 0 t o m första kvartalet 2005, = 1 från och med andra kvartalet 2005

Variablerna *Kv2*, *Kv3* och *Kv4* är säsongdummies.

### Regression Analysis: Andel sysselsatta versus tid; Dummy(15-74); Kv2; Kv3; Kv4

The regression equation is

$\text{Andel sysselsatta} = 0,483 - 0,000052 \text{ tid} - 0,00542 \text{ Dummy}(15-74) - 0,000902 \text{ Kv2} - 0,00196 \text{ Kv3} - 0,00142 \text{ Kv4}$

Predictor	Coef	SE Coef	T	P
Constant	0,483222	0,000787	614,38	0,000
tid	-0,00005167	0,00002371	-2,18	0,033
Dummy(15-74)	-0,005417	0,001018	-5,32	0,000
Kv2	-0,0009015	0,0007587	-1,19	0,239
Kv3	-0,0019606	0,0007584	-2,59	0,012
Kv4	-0,0014167	0,0007587	-1,87	0,066

S = 0,00227325    R-Sq = 74,0%    R-Sq(adj) = 72,1%

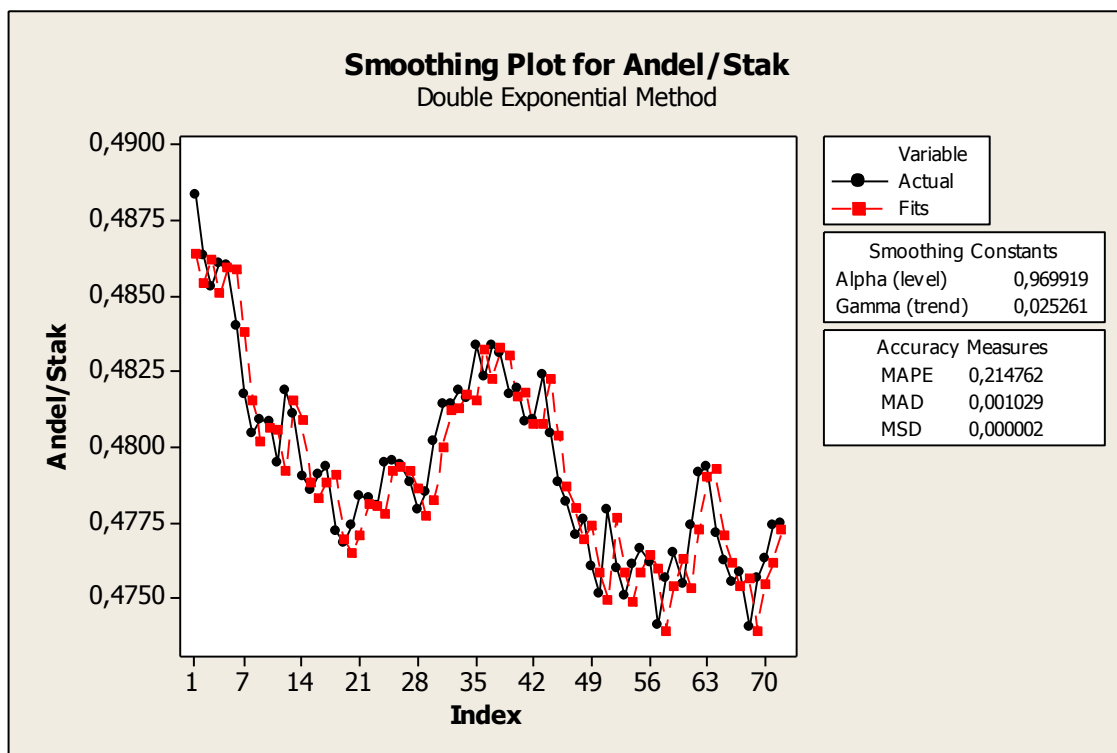
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	0,00097243	0,00019449	37,64	0,000
Residual Error	66	0,00034107	0,00000517		
Total	71	0,00131350			

Durbin-Watson statistic = 0,326024

### Uppgift 4 (13)

Denna uppgift baseras på säsongrensade kvartalsdata för andel av antal sysselsatta i Sverige som utgörs av kvinnor 1994:1 t o m 2011:4, åldersgrupp 16-64 år. Den studerade variabeln kallas  $Andel/Stak = Y_t$ .



**Data Display Andel/Stak = Säsongrensad andel sysselsatta som är kvinnor av totala antalet sysselsatta. 16-64 år.**

Row	År	Kvartal	$Y_t$	$S_t$	$T_t$
1	1994	1	0,488327	0,488269	-0,0028687
2	1994	2	0,486309	0,486281	-0,0000763
3	1994	3	0,485266	0,485295	-0,0001838
4	1994	4	0,486106	0,486076	-0,0001594
5	1995	1	0,486016	0,486013	-0,0001570
-----					
67	2010	3	0,475823	0,475810	-0,0001383
68	2010	4	0,474043	0,474092	-0,0001782
69	2011	1	0,475659	0,475606	-0,0001354
70	2011	2	0,476278	0,476254	-0,0001157
71	2011	3	0,477373	0,477336	-0,0000854
72	2011	4	0,477453	0,477447	-0,0000805

- Demonstrera hur  $S_{72}$  och  $T_{72}$  beräknats och gör prognoser för den säsongrensade tidsserien ( $Y_t$ ) en, två och tre tidpunkter framåt.
- Som du märker är utjämningskonstanten  $\alpha$  nästan lika med ett. Vilken är tolkningen av att  $\alpha = 1$ ?

### Uppgift 5 (15)

Bärgningsfirman i en region noterar för ett antal incidenter under vinterhalvåret om de förolyckade bilarna haft dubbdäck, friktionsdäck eller sommardäck. För varje incident har också väglaget noterats, enligt A: isgata, B: snömodd eller C: gott skick.

	A	B	C
Dubbdäck	38	30	34
Friktionsdäck	31	26	28
Sommardäck	31	10	8

Finns det något samband mellan typ av däck på det förolyckade fordonet och väglag? Besvara frågan med hjälp av ett hypotestest med 5% signifikansnivå.

### Uppgift 6 (20)

- A. Antag att vi önskar skatta ett medelvärde i en population med hjälp av ett OSU från en ändlig population (dragning utan återläggning). Populationsstorleken är  $N = 2000$  och den kända standardavvikelsen är  $\sigma = 30$  enheter. Hur stor ska stickprovsstorleken,  $n$ , vara om medelfelet,  $\sqrt{V(\bar{X})}$ , får vara högst 5 enheter?
- B. Energimyndigheten önskar uppskatta medelvärdet per hushåll av den energiåtgång i watt (W) som TV-apparater i ett distrikt drar på en vecka. För att undersöka detta måste ett urval av hushåll väljas vilka sedan erhåller en dosa som mäter energiåtgången för enbart TV-apparater. Hushållen i distriktet kan naturligt indelas i tre stycken strata som är tätorterna A, B och C. Energimyndigheten genomför ett stratifierat urval med lika allokering ( $n_j = 200$ ) med OSU från respektive stratum. Resultatet från undersökningen är följande:

Stratum	Populationsstorlek	Medelvärde (W)	Standardavvikelse
	$N_j$	$\bar{x}$	$s$
A	1510	1100	140
B	987	950	222
C	2543	1200	123

- a) Skatta medelvärdet per hushåll av energiåtgången i watt (W) som TV-apparater i det aktuella distriktet drar på en vecka.
- b) Beräkna medelfelet för skattningen i uppgift a).
- c) Beräkna ett 95% konfidensintervall för den aktuella genomsnittliga energiåtgången per hushåll i distriktet. Tolka intervallet!

Var noga med att ange förutsättningar.