105
Arkiv

Försättsblad till skriftlig tentamen vid Linköpings Universitet

Datum för tentamen	2011-02-12
Sal (1) Om tentan går i flera salar ska du bifoga ett försättsblad till varje sal och ringa in vilken sal som avses	TER2
Tid	8-12
Kurskod	732G71
Provkod	TENT
Kursnamn/benämning **Provnamn/benämning**	Statistik B Tentamen
Institution	IDA
Antal uppgifter som ingår i tentamen	5
Jour/Kursansvarig Ange vem som besöker salen	Karl Wahlin
Telefon under skrivtiden	0709-719096
Besöker salen ca kl.	under fm
Kursadministratör/kontaktperson (namn + tfnr + mailaddress)	Carita Lilja, 1463, carita.lilja@liu.se
Tillåtna hjälpmedel	Valfri räknedosa
Övrigt	
Vilken typ av papper ska användas, rutigt eller linjerat	Rutigt
Antal exemplar i påsen	46

Tentamen

Linköpings Universitet, Institutionen för datavetenskap, Statistik

---

Kurskod och namn:	732G71 Statistik B
Datum och tid:	2011-02-12, 8-12
Jourhavande lärare:	Kalle Wahlin
Tillåtna hjälpmedel:	Valfri räknedosa.
Betygsgränser:	Tentamen omfattar totalt 15p. Godkänt från 9p, väl godkänt från 12p.
	Formelsamling och tabeller följer efter uppgifterna. Svarsformulär till uppgifterna 2-5 finns på slutet.
	Siffrorna i uppgifterna är påhittade.

Till uppgift 1 ska fullständig lösning inlämnas. Till uppgifterna 2-5 lämnas endast svar på svarsblankett.

Uppgift 1 (7.5p)

Sedlar tillverkas normalt av råbomull. I vissa länder, med Australien som föregångsland, har man dock gått över till att tillverka sedlarna av polymer, en plastförening. Fördelen är högre slitstyrka, nackdelen är högre tillverkningskostnad. Den betalningsmedelsemitterande myndigheten i ett visst land vill utreda om det finns någon förtjänst i att övergå till polymera sedlar, och utsätter därför bomulls- respektive polymerbaserade sedlar för artificiellt slitage, och mäter hur många månaders normalt bruk sedeln förväntas hålla för. I tabellen står B för bomull och P för polymer.

Typ	B	B	B	B	B	P	P	P	P	P
Hållbarhet (månader)	13	16	15	12	17	21	22	20	24	20

a) Åskådliggör data i ett spridningsdiagram. (0.5p)
b) Beräkna korrelationskoefficienten mellan sedeltyp och hållbarhet. (1p)
c) Beräkna b_0 och b_1 i en enkel linjär regressionsmodell. (1p)
d) Tolka b_0 och b_1 med ord. (0.5p)
e) Är β_1 signifikant skild från 0? Använd 5% signifikansnivå. (1.5p)
f) Ta fram modellens residualer, åskådliggör dem i en residualplott och tolka. Är modellen välanpassad? (1p)

g) Beräkna ett 95% konfidensintervall för skillnaden i hållbarhet mellan sedlar av bomull respektive polymer. (1p)

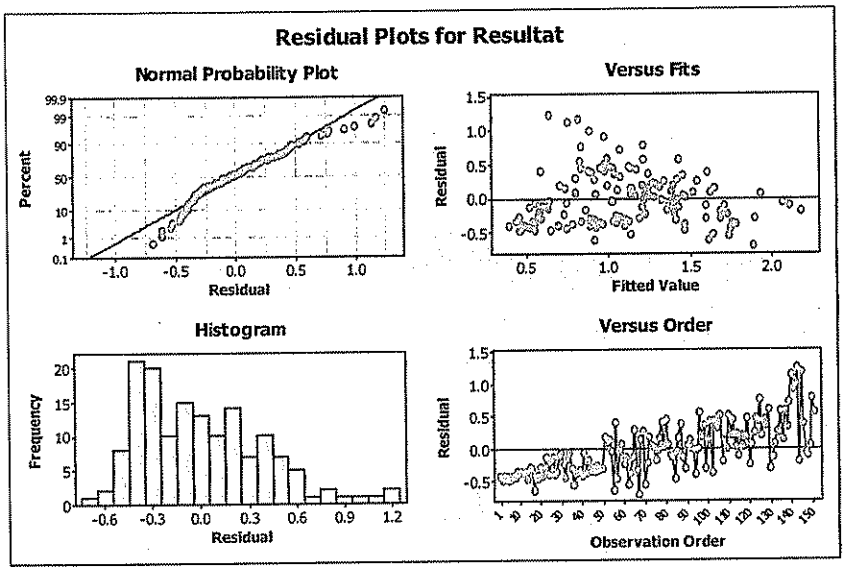
h) Beräkna ett lämpligt 95% intervall för den förväntade hållbarheten för sedlar av polymer. (1p)

Uppgift 2 (3p)

En forskare försöker utreda vilka faktorer som påverkar vilket resultat man fått på högskoleprovet. Hon drar ett slumpmässigt urval av personer som går eller nyss har slutat gymnasiet och samlar förutom provresultatet in information om ålder, kön (1 = kvinna 0 = man), senaste genomsnittsbetyg från gymnasieskolan samt huruvida man bor i tätort eller på landsbygden (1 = tätort 0 = landsbygd). Forskaren prövar flera olika regressionsmodeller, se utskrifter nedan.

Modell 1:

```
The regression equation is
Resultat = - 2.90 + 0.176 Ålder + 0.0628 Kön + 0.0693 Genomsnittsbetyg
           - 0.0705 Landsbygd/tätort


Predictor             Coef    SE Coef       T      P    VIF
Constant           -2.8991     0.4069   -7.12  0.000
Ålder              0.17595    0.02329    7.55  0.000  1.088
Kön                0.06280    0.06627    0.95  0.345  1.026
Genomsnittsbetyg  0.069336   0.009351    7.41  0.000  1.093
Landsbygd/tätort  -0.07046    0.07053   -1.00  0.319  1.005


S = 0.395384   R-Sq = 51.2%   R-Sq(adj) = 49.9%


Analysis of Variance

Source           DF       SS      MS      F
Regression        4  23.9859  5.9965  38.36
Residual Error  146  22.8240  0.1563
Total           150  46.8099


Source             DF   Seq SS
Ålder               1  15.0922
Kön                 1   0.0076
Genomsnittsbetyg    1   8.7301
Landsbygd/tätort    1   0.1560
```



Modell 2:

The regression equation is
Resultat = - 2.72 + 0.166 Ålder - 0.389 Kön + 0.0255 Kön*Ålder
+ 0.0692 Genomsnittsbetyg - 0.0740 Landsbygd/tätort

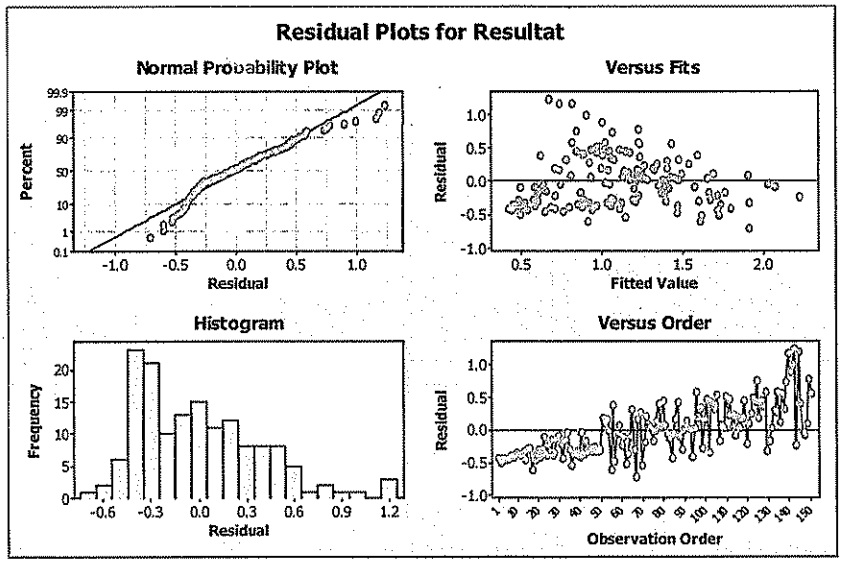
Predictor	Coef	SE Coef	T	P	VIF
Constant	-2.7238	0.5188	-5.25	0.000	
Ålder	0.16641	0.02915	5.71	0.000	1.695
Kön	-0.3886	0.8285	-0.47	0.640	159.665
Kön*Ålder	0.02550	0.04665	0.55	0.585	157.830
Genomsnittsbetyg	0.069189	0.009377	7.38	0.000	1.094
Landsbygd/tätort	-0.07395	0.07099	-1.04	0.299	1.013

S = 0.396337 R-Sq = 61.3% R-Sq(adj) = 59.8%

Analysis of Variance

Source	DF	SS	MS	F
Regression	5	24.0329	4.8066	30.60
Residual Error	145	22.7771	0.1571	
Total	150	46.8099		

Source	DF	Seq SS
Ålder	1	15.0922
Kön	1	0.0076
Kön*Ålder	1	0.0634
Genomsnittsbetyg	1	8.6992
Landsbygd/tätort	1	0.1705



Modell 3:

The regression equation is
Resultat = - 2.83 + 0.174 Ålder + 0.0684 Genomsnittsbetyg
- 0.0736 Landsbygd/tätort

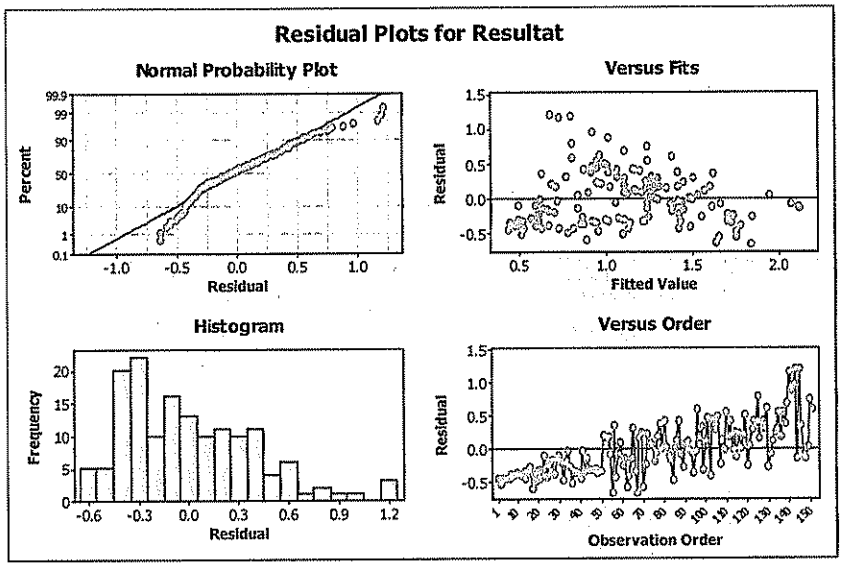
Predictor	Coef	SE Coef	T	P	VIF
Constant	-2.8273	0.3996	-7.07	0.000	
Ålder	0.17417	0.02321	7.51	0.000	1.080
Genomsnittsbetyg	0.068397	0.009295	7.36	0.000	1.081
Landsbygd/tätort	-0.07356	0.07043	-1.04	0.298	1.003

S = 0.395247 R-Sq = 50.9% R-Sq(adj) = 49.9%

Analysis of Variance

Source	DF	SS	MS	F
Regression	3	23.8455	7.9485	50.88
Residual Error	147	22.9644	0.1562	
Total	150	46.8099		

Source	DF	Seq SS
Ålder	1	15.0922
Genomsnittsbetyg	1	8.5829
Landsbygd/tätort	1	0.1704



Modell 4:

The regression equation is
Resultat = - 2.87 + 0.173 Ålder + 0.0688 Genomsnittsbetyg

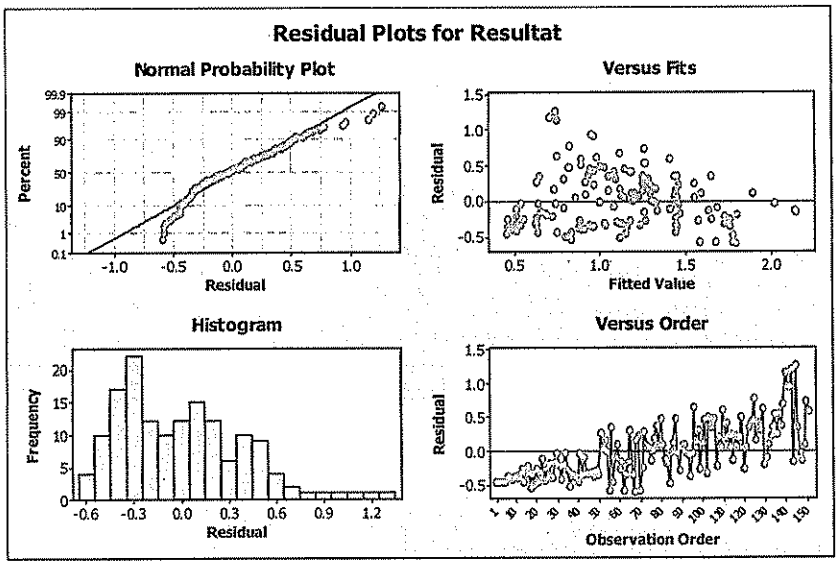
Predictor	Coef	SE Coef	T	P	VIF
Constant	-2.8657	0.3981	-7.20	0.000	
Ålder	0.17311	0.02319	7.46	0.000	1.078
Genomsnittsbetyg	0.068829	0.009289	7.41	0.000	1.078

S = 0.395369 R-Sq = 50.6% R-Sq(adj) = 49.9%

Analysis of Variance

Source	DF	SS	MS	F
Regression	2	23.675	11.838	75.73
Residual Error	148	23.135	0.156	
Total	150	46.810		

Source	DF	Seq SS
Ålder	1	15.092
Genomsnittsbetyg	1	



a) Vilket av följande påståenden stämmer bäst? (1p)

I. Modell 2 är den bästa eftersom den har högst förklaringsgrad.

II. Analyserna har baserats på data insamlade från 150 personer.

III. Enligt modell 1 gäller att en boende på landsbygden i genomsnitt har ett resultat som ligger 0.07 enheter lägre än boende i tätort.

IV. Residualplotten för modell 4 visar på tydliga tecken på multikollinearitet.

V. För modell 3 gäller att *F*-testet av om $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ ej är signifikant på 5% nivå.

VI. För modell 4 gäller att den sekventiella kvadratsumman för variabeln *Genomsnittsbetyg* givet att variabeln *Ålder* redan finns i modellen är 8.583.

b) Beräkna ett 95% konfidensintervall för variabeln *Ålder* i modell 4. (1p)

c) Testa på 5% nivå om variablerna *Genomsnittsbetyg* och *Landsbygd/tätort* tillsammans platsar i modell 3 eller ej på 5% signifikansnivå. (1p)

Uppgift 3 (1.5p)

Man vill mäta prisutvecklingen för inventarier och husgeråd mellan 2006 och 2008 inom en koncern som saluför dessa varutyper. Man delar in koncernens sortiment i tre varugrupper (möbler, mattor och armaturer) och väljer i varje grupp en representantvara.

Försäljningsvärde i Mkr

	Grupp 1: Möbler	Grupp 2: Mattor	Grupp 3: Armaturer
2006	2.9	0.8	1.6
2007	3.2	1.2	1.3
2008	4.5	1.7	0.4

Årsmedelpriser, kr

Representantvaror		2006	2007	2008
Grupp 1:	Tresitssoffa, fuskläderklädsel	7750	8925	9639
Grupp 2:	Wiltonmatta 140*200 cm	2950	3157	3284
Grupp 3:	Golvlampa i stål	950	1083	1104

Bilda ett kedjeprisindex med årslänkar enligt Laspeyre med basår 2006. (1.5p)

Uppgift 4 (2p)

En fondförvaltare har analyserat utvecklingen av ett visst belopp under fem års tid. Behållningen år t betecknas v_t och den modell man analyserat är

$$v_t = v_0 \cdot (1+a)^t \cdot \delta, t = 1, 2, 3, 4, 5$$

där a står för räntesatsekvivalenten. I analysen nedan har variabeln v logaritmerats med 10-logaritmen.

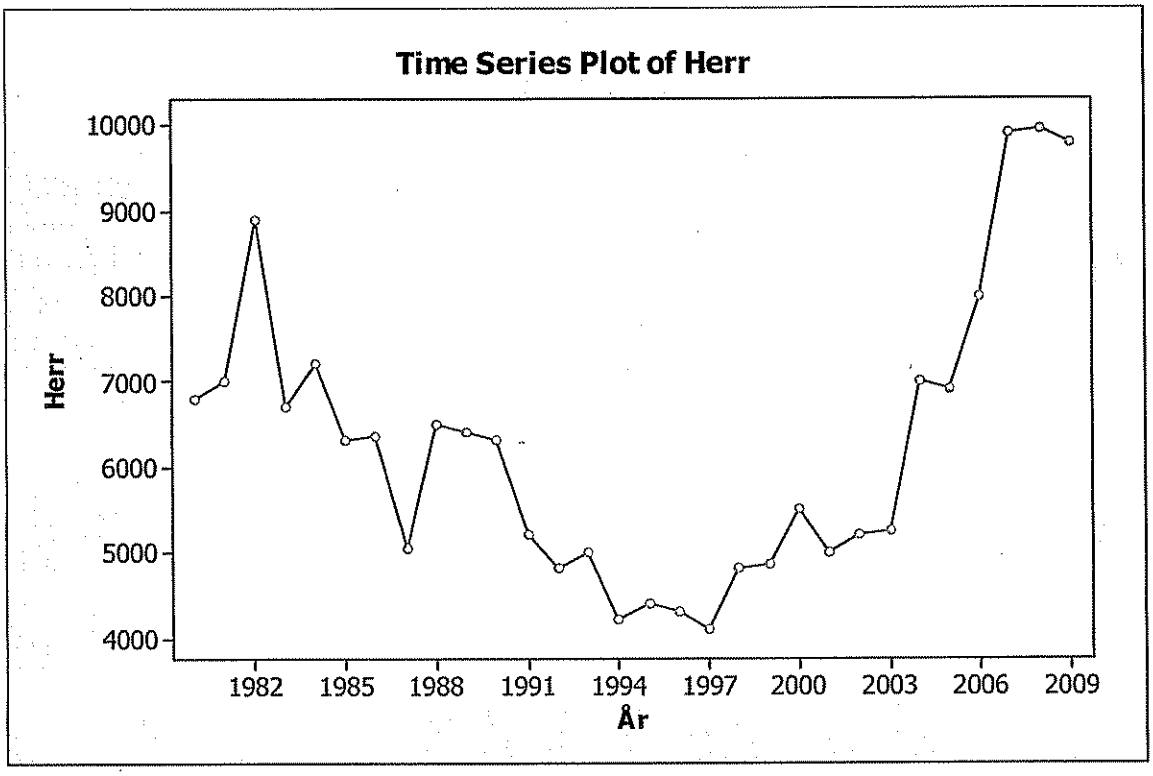
Regression Analysis: y versus x

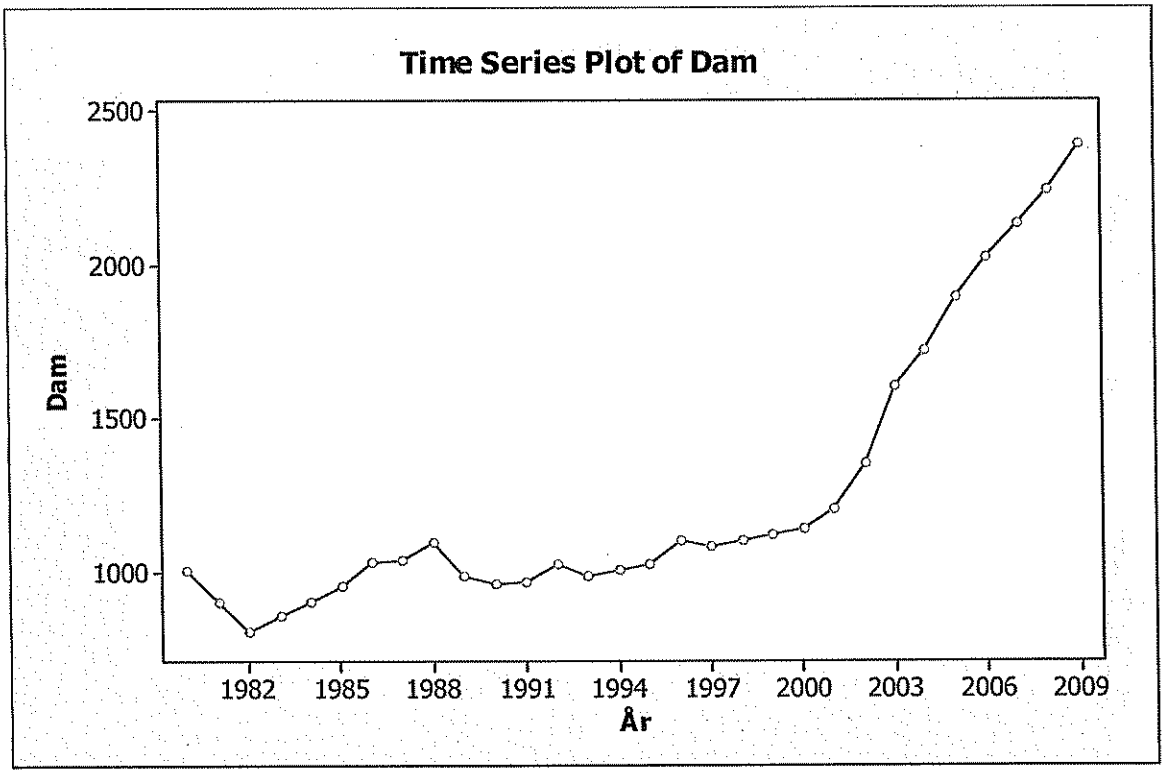
```
The regression equation is
log v = - 25.3 + 0.0133 t


Predictor       Coef  SE Coef
Constant     - 25.29    11.73
t          0.013320 0.005850


S = 0.0185004   R-Sq = 63.3%   R-Sq(adj) = 51.1%
```

a) Beräkna en skattning av räntesatsekvivalenten a. (1p)
b) Beräkna ett 95% konfidensintervall för. (1p)

Uppgift 5 (1p)

Följande figurer visar genomsnittligt årligt antal besökare vid hemmamatcher för LHC:s herrlag (överst) och damlag (underst).





a) Vilket av följande påståenden om de båda tidsserierna stämmer bäst? (0.5p)

I. Publiksiffrorna för herrmatcher visar på tydlig säsongsvariation, men detta mönster går inte att se för dammatcherna.

II. En linjär tidsseriegressionsmodell ger troligen bästa anpassningen för dammatcherna.

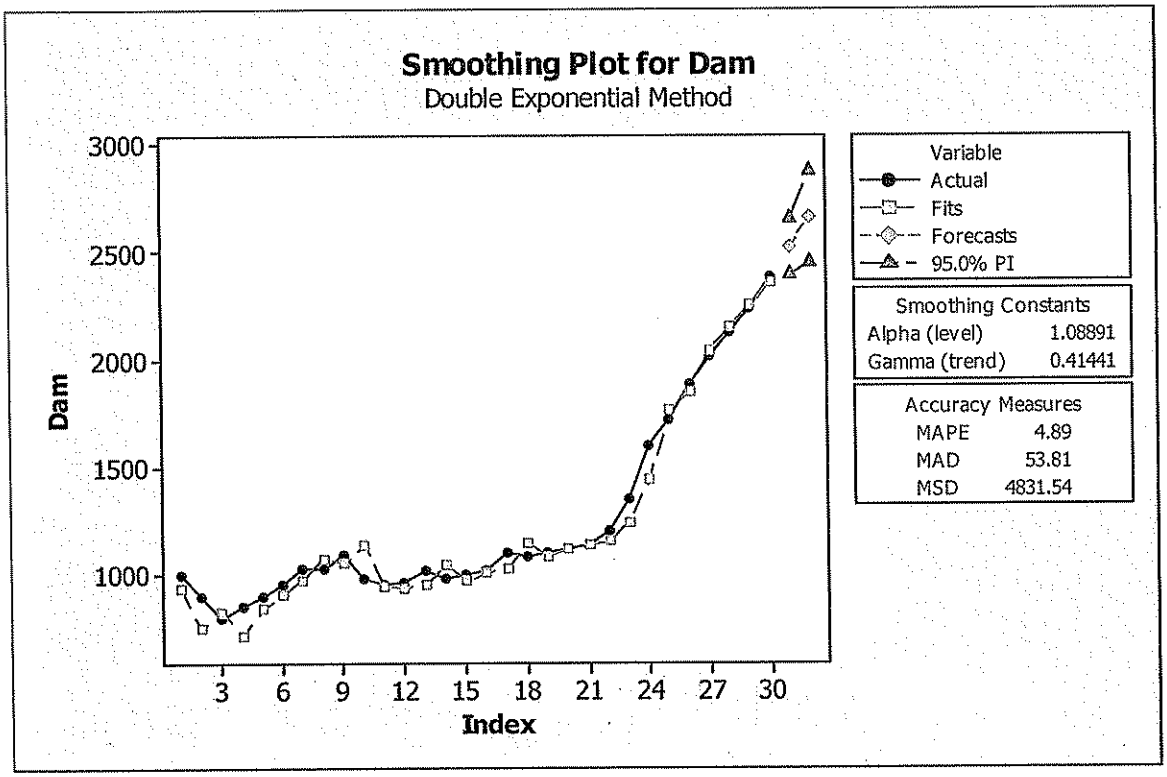
III. Publiksiffrorna för herrmatcher visar på cyklisk variation medan dammatchernas publiksiffror närmast visar på en exponentiell trend.

IV. Om tidsserieregression används ska dummyvariabler för kvartal användas som förklaringsvariabler.
V. Prognoser för de båda serierna görs lämpligtvis med Winters metod.
VI. Båda serierna är att bedöma som stationära.

En analys görs av dammatchernas publiksiffror.

Double Exponential Smoothing for Dam

```
Data     Dam
Length   30


Smoothing Constants
Alpha (level)   1.08891
Gamma (trend)   0.41441


Accuracy Measures
MAPE      4.89
MAD      53.81
MSD    4831.54


Forecasts
Period  Forecast    Lower    Upper
31       2527.33  2395.49  2659.17
32       2661.95  2448.16  2875.73
```



b) Vilket av följande påståenden stämmer bäst om analysen? (0.5p)

I. Måttet MAPE är närmast jämförbart med MSE i en tidsserieregression.
II. Prognoserna har beräknats genom att anpassa en rät linje till de 25% sista observationerna i tidsserien.
III. En logaritmering av värdena skulle ha resulterat i en negativ trend.
IV. Prognoserna gäller för kvartal 1 och 2 2010.
V. Den valda metoden är illa vald eftersom en av utjämningskonstanterna är större än 1.
VI. Det höga värdet på MSD indikerar att det föreligger multikollinearitet.