Stockholm Universitet
Statistiska Institutionen
Maria Anna Di Lucca

Regressions- och tidsserieanalys

Tentamen
Onsdagen den 30 november, 2022
Lösningar

Godkända hjälpmedel

- Miniräknare och språklexikon.
- Formelsamling och statistiska tabeller delas ut vid tentamen.

Info

- Tentamen består av fem uppgifter, uppdelade i deluppgifter.
- För full poäng på en uppgift krävs tydliga, utförliga och välmotiverade lösningar.
- Resultatet meddelas senast den **2022-12-21**.
- Lösningsförslag till tentamensuppgifterna läggs ut på Athena strax efter tentamen.
- Använd minst fem värdesiffror i dina beräkningar (1,2345 och 1234,5 är exempel på sådana tal). I dina beräkningar från R-utskrifter får du utgå från det som är givet. Du kan dock avrunda ditt slutliga svar.

Gräns för godkänt

- 50 poäng av totalt 100.

OBS. Fx och F är underkända betyg som kräver omexamination. Studenter som får betyget Fx kan alltså inte komplettera för högre betyg.

Uppgift 1. (35 poäng)

Från en undersökning om 44 kvinnor om sambandet mellan energiomsättning och kroppens vikt fick vi följande två diagram som visas i Figur 1 på vänster och höger sidan.

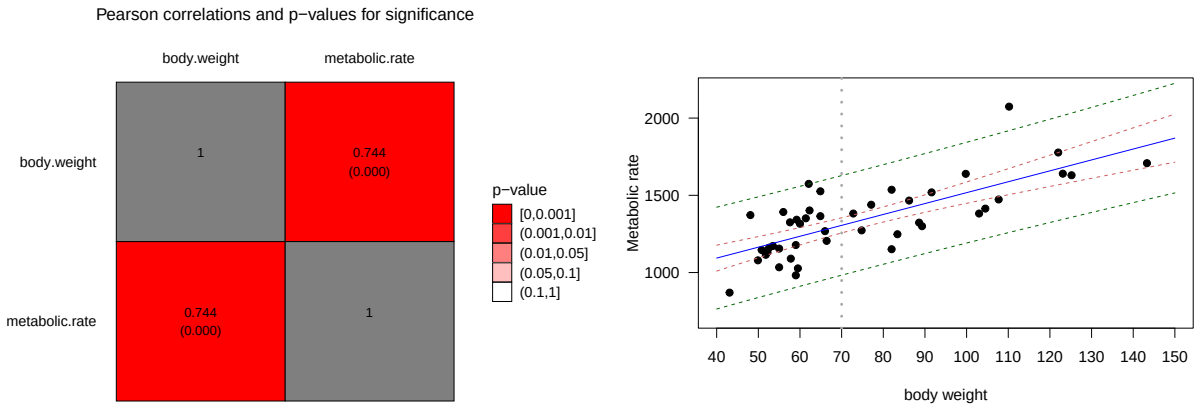


Figure 1: Samband mellan energiomsättning och vikt

Från Figur 1 åt höger ser vi att vi kan anta en enkel linjär regressionsmodell

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2)$$

där responsvariabeln y är energiomsättning (kilokaloriförbrukning per 24 timmar, kcal/24hr) och den förklarande variabeln x är kroppens vikt (i kilogram, kg). Nedan visas data

$$\sum x_i = 3294.7 \; \sum y_i = 58953 \; \sum x_i y_i = 4598556$$

$$\sum x_i^2 = 272795.5 \; \sum y_i^2 = 81335113$$

Vänligen lös och svara på följande deluppgifter (a-h):

a) Beräkna och tolka förklaringsgraden R^2 i ord. (5p)

Lösning

Det är en enkel regression och man kan använda formeln: $r^2 = R^2$. Från Figur 1 åt vänster har vi urvalskorrelationen: $0.744^2 = 0.553536$. Kvinnors kroppsvikt kan förklara 55% av den totala variationen i energiomsättning.

b) Den vänstra grafen i Figur 1 visar urvalskorrelationen. Förklara skillnaden mellan förklaringsgraden och korrelationen. (5p)

Lösning

Korrelationen och förklaringsgraden fungerar för linjära samnd. Korrelationen handlar om sambandet mellan vikt och energiomsättningen. I förklaringsgraden antar man att kroppsvikt påverkar energiomsättnigen. Man kan också lägga till att korrelationen är en indikator som varierar mellan -1 och 1. Förklaringsgraden, å andra sidan, är mellan 0 och 1.

c) Skatta koefficienterna β_0 och β_1 med minsta-kvadrat-metoden. (5p)

Lösning

Det finns flera vägar för att skatta lutningen, $\hat{\beta}_1 = b = r_{xy}\frac{s_y}{s_x}$

$\hat{\beta}_0 = a = \bar{y} - b\bar{x}$

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{3294.7}{44} = 74.87955$ och

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{58953}{44} = 1339.841$

$s_x^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{272795.5 - 44 * 74.87955^2}{43} = \frac{26089.83}{43} = 606.7402$

och då $s_x = \sqrt{606.740} = 24.6321$

$s_y^2 = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1} = \frac{81335113 - 44 * 1339.841^2}{43} = \frac{2347461}{43} = 54592.12$

och då $s_y = \sqrt{54592.12} = 233.6496$

så lutningen: $\hat{\beta}_1 = b = r_{xy}\frac{s_y}{s_x} = 0.744\frac{233.6496}{24.6321} = 7.057267$

eller man kunde beräkna med $\frac{s_{xy}}{s_x^2}$

För interceptet kan man stoppa in resultat i formeln:

$\hat{\beta}_0 = a = \bar{y} - b\bar{x} = 1339.841 - 7.057267 * 74.87955 = 811.396$

d) Tolka parameterskattningar från punkt c) i ord. (3p)

Lösning

Lutningen är positiv och för ett kilos vikteökning ökar energiomsättningen 7.05 (kg/kal). Interceptet är istället att man kan anta att man inte har resultat för kroppsvikt men man ser ändå att energiomsättning är 811.396 kilokalorier.

e) Testa om lutningen för regressionslinjen är skild från noll på $\alpha = 0.05$. (3p)

Lösning

$H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

Man förkastar nollhypotesen om $|t_{obs}| > t_{krit}$. I detta fall har vi

$t_{krit} = t_{0.025,42} = 2.018$ (den närmaste i t-tabell)

$t_{obs} = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}}$

För att beräkna $s_{\hat{\beta}_1}$ behöver man $s_e^2 = \frac{SSE}{n-k-1}$.

Vi har redan förklaringsgraden $R^2 = 0.553536 = 1 - \frac{SSE}{SST}$ och $SST = (n-1)s_y^2 = 43 * 54592.12 = 2347461$.

$0.553536 = 1 - \frac{SSE}{2347461}$ så

$SSE = (1 - 0.553536) * 2347461 = 1048057$ och då $s_e^2 = \frac{1048057}{42} = 24953.74$

$s_{\hat{\beta}_1}^2 = \frac{24953.74}{43*606.740} = 0.956455$ och då $s_{\hat{\beta}_1} = \sqrt{0.956455} = 0.9779852$

det följer

$t_{obs} = \frac{7.057267}{0.9779852} = 7.216129 > 2.018$

Man kan förkasta nollhypotesen dvs. kroppens vikt påverkar energiomsättningen.

f) Bestäm ett 95%-igt konfidensintervall för den genomsnittliga energiomsättningen för kvinnor med en kroppsvikt på 70 kg. (5p)

Lösning

Punktskattning för den genomsnittliga energiomsättningen för kvinnor med en kroppsvikt på 70 kg är $\hat{\mu}_{y|x=70} = 811.396 + 7.057267 * 70 = 1305.405$.

Från t-tabell: $t_{n-2,\alpha/2} = t_{44-2,0.05/2} = 2.018$.

För att beräkna konfidensintervall, konstruerar vi först standardavvikelse:

$\sqrt{s_e^2(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2})}$

och då blir den statistiska felmarginalen:

$2.018 * \sqrt{24953.74(\frac{1}{44} + \frac{23.81001}{26089.83})} =$

$2.018 * \sqrt{24953.74 * 0.02363989} = 2.018 * 24.28793 = 49.01304$

och då blir intervallet:

1305.405 ± 49.01304 eller $(1256.392; 1354.418)$.

g) Bestäm ett 95%-igt prognosintervall för energiomsättning för en slumpmässigt vald kvinna med vikten lika med 70 kg. (5p)

Lösning

Punktskattning för en kvinna med kroppsvikt lika med 70, energiomsättnings prognos är (den gråalinjen) $\hat{y}_{x=70} = 811.396 + 7.057267 * 70 = 1305.405$

Standardavvikelse ändrar i formeln med ettan:

$\sqrt{s_e^2(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2})}$

Vi använder samma t-värdet som tidigare:

$t_{n-2,\alpha/2} = t_{44-2,0.05/2} = 2.018$

och då blir den statistiska felmarginalen:

$2.018 * \sqrt{24953.74(1 + \frac{1}{44} + \frac{(70-74.87955)^2}{26089.83})} =$

$2.018 * \sqrt{24953.74(1 + \frac{1}{44} + \frac{23.81001}{26089.83})} =$

$2.018 * \sqrt{24953.74 * 1.02364} = 2.018 * 159.8238 = 322.5244$

och då blir intervallet:

1305.405 ± 322.5244 eller $(982.8806; 1627.929)$

h) Se resultat i punkt f) och i punkt g) till höger i Figur 1. Vilka är de röda och vilka är de grönfärgade linjerna i diagrammet? (4p)

Lösning

Punktskattningen är densamma men konfidensintervallet är annorlunda på grund av standardavvikelsen. Man kan beräkna bredd för intervallen: 98.026 och 645.0484 (ungefär) respektive. Det innebär att resultat i punkt g) refereras till de rödfärgade linjerna istället för de gröna med större bredd som är i resultatet i punkt h).

Uppgift 2. (30 poäng)

En linjär regressionsmodell med tre förklaringsvariabler, $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$, anpassades till ett datamaterial med $n = 67$ observationer. Den beroende variabeln är hyrespris per gräsbevuxen hektar (i dollar) och de förklarande variablerna är hyra per tunnland (X_1), mjölkkor per kvadratkilometer (X_2) och skillnad mellan betesmark och åkermark (X_3).

Vänligen lös och svara på följande deluppgifter (a-f):

a) Förklara varför Modell 2 är bättre än Modell 1. (5p)

Lösning

I Modell 2 har vi signifikanta skattningar av parametrar. Modell 1 är icke-signifikant skattning av parameter för förklarande variabeln x_3. Man kan fundera på standadavvikelse (Std.Error) och anse att de är mindre i modell 2 än i mondell 1.

Modell 1 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$

```
##
## Parameter estimates
## ------------------------------------------------
##              Estimate Std. Error t value   Pr(>|t|)
## (Intercept)  -4.07577   3.766912 -1.0820 2.8338e-01
## x1            0.89549   0.065132 13.7490 1.2419e-20
## x2            0.45189   0.091903  4.9170 6.6015e-06
## x3          -11.56709  11.235785 -1.0295 3.0719e-01
## ------------------------------------------------
```

Modell 2 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

```
##
## Parameter estimates
## ------------------------------------------------
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept) -6.61267   2.850396 -2.3199 2.3547e-02
## x1           0.93656   0.051508 18.1828 5.6825e-27
## x2           0.39234   0.071451  5.4910 7.3570e-07
## ------------------------------------------------
```

b) Flera tal i följande ANOVA-tabell som gäller den bästa modellen från punkt a) har ersatts med "*". Komplettera tabellen. (5p)

Lösning

Med hjälp av R kan man se resultat

```
##
## ANOVA table
## ------------------------------------------------
##       df SS      MS       F Pr(>F)
## Regr  *  *       14655.14 * 0
## Error 64 *       *
## Total 66 34371.7
##
## ------------------------------------------------
##
## -----Resultat med R---------------
##
## Analysis of variance - ANOVA
## ------------------------------------------------
##       df      SS        MS      F     Pr(>F)
```

```
## Regr   2 29310.3 14655.143 185.31 2.3892e-27
## Error 64  5061.4    79.085
## Total 66 34371.7
##
## Measures of model fit
## ---------------------------------------------
## Root MSE       R2   R2-adj
##  8.89295  0.85274  0.84814
```

Man kan beräkna från tabellen i den första kolumnen har vi antal frihetsgrader, $k = n - 1 - (n - k - 1) = 66 - 64 = 2$ redan här fattar vi att ANOVA är för modell 2. För att beräkna SSR i den andra kolumnen, kan man använda resultat MSR=14655.14. Då blir $SSR = MSR * k = 14655.14 * 2 = 29310.28$. För att beräkna SSE kan man använda SST-SSR och då blir det $34371.7 - 29310.28 = 5061.42$. Nu kan vi beräkna MSE som medelvärdet av residualtermer i kvadrat: $\frac{SSE}{n-k-1} = \frac{5061.42}{64} = 79.08469$. Det saknas F-värdet men nu har vi allt för att ha det: $\frac{MSR}{MSE} = \frac{14655.14}{79.08469} = 185.3094$

c) Beskriv F-testet i ANOVA-tabellen: förklara hypoteser, och vad det innebär att förkasta nollhypotesen. Genomför testet på signifikansnivån 5 procent. (5p)

Lösning

Vi refererar till den bästa modellen från punkt a).

$H_0 : \beta_1 = \beta_2$

$H_A : \beta_j \neq 0$ åtminstone en förklaringsvariabel kan förklara och påverka den beroende variabeln.

Man kan läsa direkt p-värdet och ser att det är noll och det är mindre än signifikansnivån, $0 < 0.05$. Vi kan förkasta nollhypotesen och modellen är i sin helhet signifikant. Man kan förklara på ett annat sätt med jämförelse mellan F från observationer och F från tabell. Man kan använda tidigare resultat: $F_{obs} = 185.3094$ och som tummregel vet vi att det är större än 1 och då kan vi förkasta nollhypotesen. Vi kan också lägga till att den delen av regressionen (MSR) är betydligt större än residualdelen (MSE). Då kan vi säga att parametrar skiljer sig från noll och förklaringsvariabler förklarar den beroende variabeln. Om vi vill titta på F-tabellen kan vi säga att värdet är mellan 1.50-1.52 baserade på förklaringsgrader mellan 60 och 70. Med R (inte på tentan!) kan vi beräkna värdet till 1.5073.

d) Förklara skillnaden mellan R^2 och VIF. (4p)

Lösning

Förklaringsgraden är en indikator mellan 0 och 1 och man beräknar för att se hur mycket en eller flera förklaringsvariabler samtidigt kan förklara den beroende variabeln. VIF är istället en indikator som kan variera från 0 till oändlighet. Den mäter sambandet mellan de två förklarande variablerna och om det är större än 5 eller 10 då kan man säga att det finns multikollinearitet. Det innebär att de två förklarande variablerna

kommer att säga samma information om den beroende variabeln och det betyder att det finns redundans i datasetet och vad man kan göra är att välja mellan dessa två förklaringsvariabler. Den förklarande variabeln som påverkade mest den beroende variabeln kommer att stanna kvar i modellen.

e) Beräkna den justerade förklaringsgraden, R^2_{adj} från ANOVA-tabellen och förklara varför man inte kan använda förklaringsgraden R^2 för att välja den bästa modellen mellan modeller med olika antal förklaringsvariabler. (3+3p)

Lösning

$R^2_{adj} = 1 - \frac{MSE}{MST} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{79.08469}{520.7833} = 1 - 0.1518572 = 0.8481428$

I justerat förklaringsgraden finns frihetsgrader, så tar man hand av antal parameter finns i modellen.

f) Beräkna ett 95%-igt konfidensintervall för den skattade parametern vid den förklarande variabeln som är icke-signifikant. (5p)

Lösning

Från formelsamlingen har vi $b_j \pm t_{n-k-1\frac{\alpha}{2}} s_{b_j}$

med n=67 k=3 och i t-tabell ska det vara medelvärdet: (2+1.997)/2=1.9985

men okej med avrundning.

$-11.56709 \pm 1.998341 * 11.23579$ och då

$(-34.02003; 10.88585)$.

(alternativ som passar på den här kursen-ingenting i teori om hat-matrix)

Från formelsamlingen får vi:

$b_j \pm z_{1-\frac{\alpha}{2}} s_{b_j}$

och i den här övningen har interceptet och x_3 i modell 1 som är icke-signifikanta. Men vi fokuserar på den förklarande variabeln som är icke-signifikant.

$-11.56709 \pm 1.96 * 11.23579$ och då

$(-33.58924; 10.45506)$

Uppgift 3. (10 poäng)

I nedanstående tabell visas kvartalsvisa observationer på produktionen (i miljoner kr) för ett företag från första kvartalet 2019 till fjärde kvartalet 2021.

År	Kvartal	y
2019	Kvartal 1	99
2019	Kvartal 2	63
2019	Kvartal 3	73
2019	Kvartal 4	167
2020	Kvartal 1	108
2020	Kvartal 2	76
2020	Kvartal 3	105
2020	Kvartal 4	174
2021	Kvartal 1	139
2021	Kvartal 2	99
2021	Kvartal 3	135
2021	Kvartal 4	197

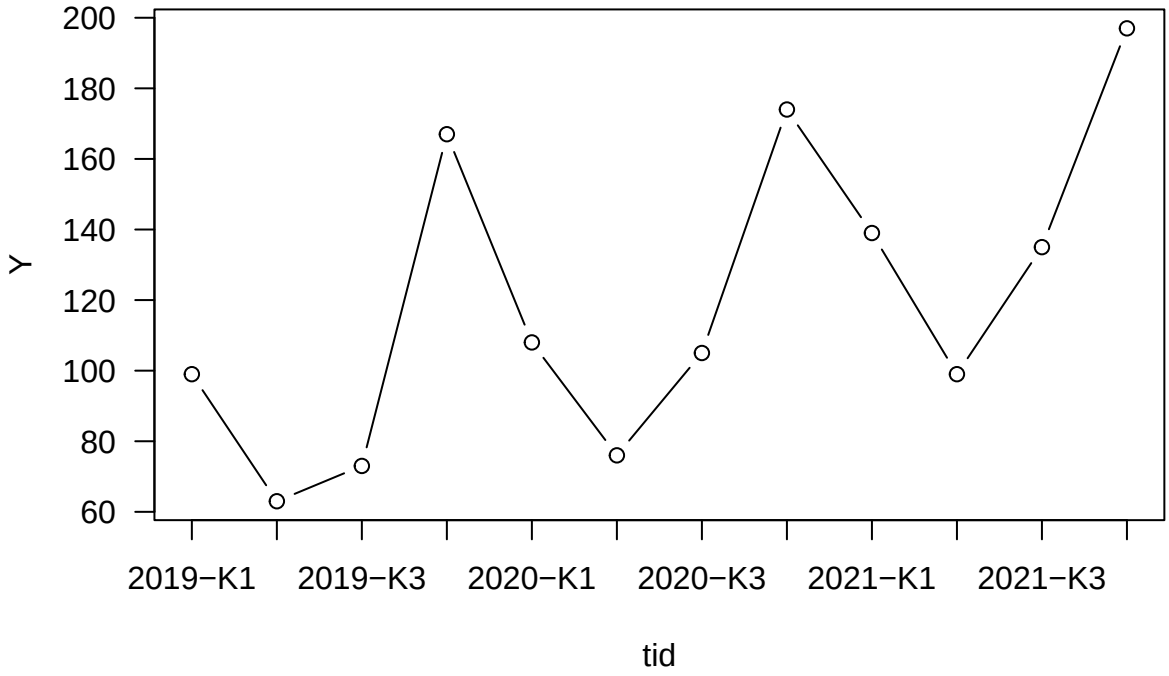
Svara på följande två frågor (a-b).

a) Skatta trenden för kvartalsdata med centrerade glidande medelvärden med säsongsvariation i en additiv modell. (5p)

Lösning

Man kan rita tidserien(inte efterfrågat på tentan!) för att vara säker på att den ser ut som en additiv modell.

Till exempel $T_3 = \frac{1}{8} * 99 + \frac{1}{4} * 63 + \frac{1}{4} * 73 + \frac{1}{4} * 167 + \frac{1}{8} * 108 = 101.625$



År	Kvartal	y	trend
2019	Kvartal 1	99	
2019	Kvartal 2	63	
2019	Kvartal 3	73	101.625
2019	Kvartal 4	167	104.375
2020	Kvartal 1	108	110
2020	Kvartal 2	76	114.875
2020	Kvartal 3	105	119.625
2020	Kvartal 4	174	126.375
2021	Kvartal 1	139	133
2021	Kvartal 2	99	139.625
2021	Kvartal 3	135	
2021	Kvartal 4	197	

b) Skatta säsongkomponenterna för modellen i a). (5p)

Lösning

Till exempel

$S_1 = \frac{-2+6}{2} = 2,$

$\bar{S} = \frac{2-39.75-21.625+55.125}{4} = -1.0625$

$S_1^+ = S_1 - (\bar{S}) = 2 - (-1.0625) = 3.0625$, $S_2^+ = -38.6875$, $S_3^+ = -20.5625$ och $S_4^+ = 56.1875$.

Det gäller

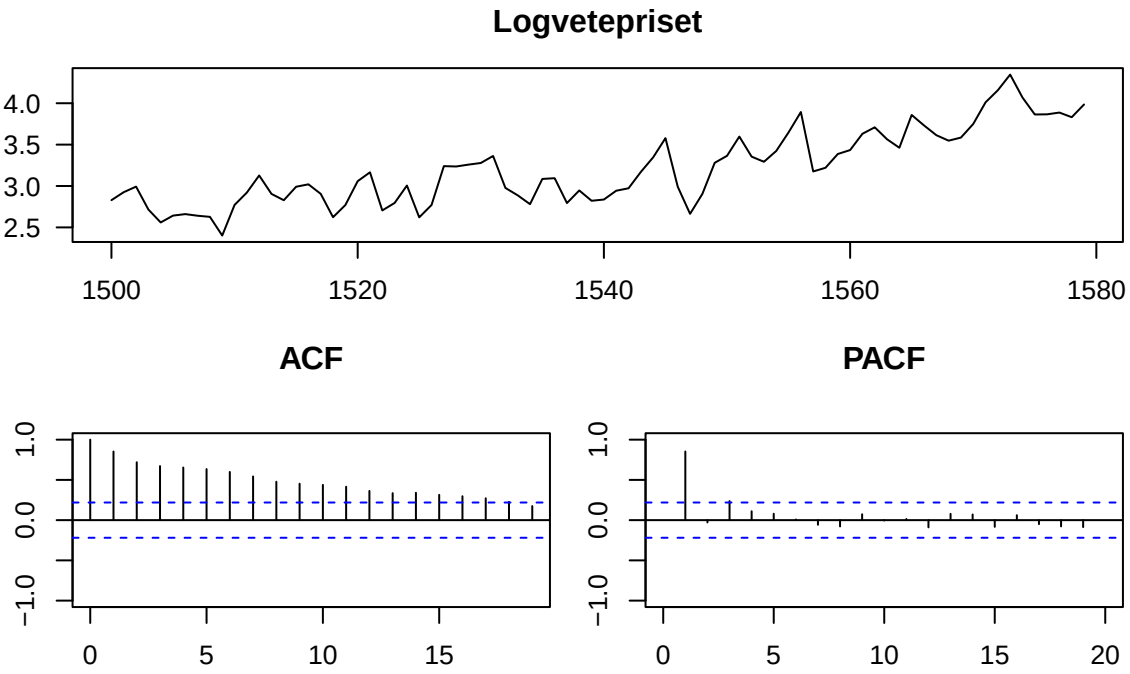
$\sum S_j^+ = 3.0625 - 38.6875 - 20.5625 + 56.1875 \approx 0.$

Det var inte efterfrågat men man kan beräkna $\hat{Y}_t = \hat{T}_t + \hat{S}_t$.

År	**Kvartal**	**y**	**trend**	**säsångsk**	**yhat**
2019	Kvartal 1	99		3.0625	102.0625
2019	Kvartal 2	63		-38.6875	24.3125
2019	Kvartal 3	73	101.625	-20.5625	52.4375
2019	Kvartal 4	167	104.375	56.1875	223.1875
2020	Kvartal 1	108	110	3.0625	111.0625
2020	Kvartal 2	76	114.875	-38.6875	37.3125
2020	Kvartal 3	105	119.625	-20.5625	84.4375
2020	Kvartal 4	174	126.375	56.1875	230.1875
2021	Kvartal 1	139	133	3.0625	142.0625
2021	Kvartal 2	99	139.625	-38.6875	60.3125
2021	Kvartal 3	135		-20.5625	114.4375
2021	Kvartal 4	197		56.1875	253.1875

Uppgift 4. (10 poäng)

I den här uppgiften analyseras förändringar i veteprisindex under 80 åren (1500-1579) på 50 olika platser i U.S.A. och sammanfattade som medelvärdet av dessa priser. I Figuren 4.1 nedan visas den ursprungliga tidserien redan logaritmerad, autokorrelationen och partiella autokorrelationen för datat.



Figur 4.1

Lös följande deluppgifter (a-b):

a) Med hjälp av R-utskrifter och residualdiagram i Figur 4.2 motivera om det är bättre en AR(1) modell eller en AR(3) modell. (4p)

Lösning

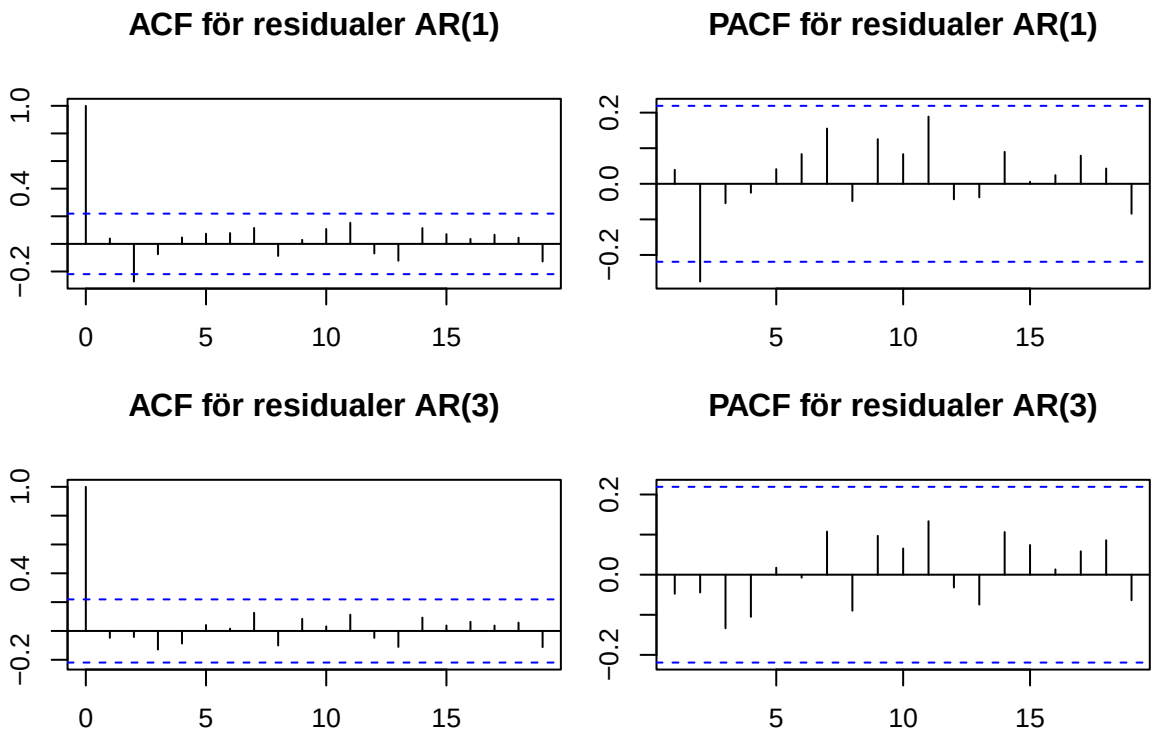
Man kan ge en titt på skattade modeller och ser att de har alla signifikanta skattningar. Dock, handlar det mer om residualanalys när det gäller vilken modell som bättre kan anpassas till datat. Det är bättre med AR(3) därför att det finns ingenting kvar i diagrammen med residualerna. I Figur 4.2 i residualerna för AR(1) har vi lag två som är signifikant i ACF och i PACF diagram. Man kan tänka på någon MA-komponent men man det är utanför efterfrågas på tentan. På den här mellan dessa två autoregressiva modeller ser vi att trelaggade tidsserien bättre representerar det logaritmerade vetepriset.

```
## --------------------------AR(1)-------------------------------

##
## Parameter estimates
## ------------------------------------------------------------
##      Estimate Std. Error z-ratio Pr(>|z|)   2.5 %  97.5 %
## ar1   0.88297   0.052748  16.739        0 0.77959 0.98636
## mean  3.24003   0.189259  17.120        0 2.86908 3.61098

## --------------------------AR(3) -----------------------------

##
## Parameter estimates
## ------------------------------------------------------------
##      Estimate Std. Error z-ratio Pr(>|z|)     2.5 %    97.5 %
## ar1   0.93862    0.10631  8.8294 0.000000  0.730259  1.146979
## ar2  -0.32399    0.14512 -2.2326 0.025575 -0.608412 -0.039559
## ar3   0.30501    0.10654  2.8627 0.004200  0.096181  0.513835
## mean  3.26527    0.24469 13.3446 0.000000  2.785680  3.744856
```


ACF för residualer AR(1)
PACF för residualer AR(1)
ACF för residualer AR(3)
PACF för residualer AR(3)
1.0
0.4
−0.2
0.2
0.0
−0.2
0
5
10
15

Figur 4.2

b) I nedanstående Tabell 4 visas data för år 1581-1586 och några prognosticerade vetepriser med AR(1) och respektive AR(3). Använd valfritt relevant mått som visar hur träffsäkra prognoserna är för AR(1) respektive AR(3) under tidsperioden. Vilken modell är att föredra utifrån det valda måttet? (6p=2+2+2)

Lösning

Man kan tänka på de olika måtten: MAD, MSE, MAPE. Man kan jämföra roten ur MSE och ser att det är mindre för AR(3) än för AR(1). Det innebär att man har en bättre prognos med AR(3) än med AR(1). Doch, har den skattade modellen med AR(3) mindre residualer (0.30) än i AR(1) (som är lika med 0.38).

```
## -------------Tabell 4--------------

## Time Series:
## Start = 81
## End = 86
## Frequency = 1
##      data  AR(1)  AR(3)
## 81 3.9914 3.8977 3.9471
## 82 4.0047 3.8208 3.8447
## 83 4.0027 3.7528 3.8077
## 84 4.0127 3.6928 3.7947
## 85 3.9479 3.6398 3.7631
## 86 4.3532 3.5930 3.7265

##            AR(1)      AR(3)
## MAD  0.07737048 0.05747822
## MAE  0.31928333 0.31928333
## MSE  0.14670243 0.09000257
## RMSE 0.38301753 0.30000428
```

Uppgift 5. (15 poäng)

Ett datamaterial innehåller 1070 köp där kunden antingen köpt Citrus Hill, (CH) eller Minute Maid Orange Juice (MM). Man satte upp en logistisk regressionsmodell för att modellera sannolikheten att en kund köper Minute Maid Orange Juice, $Y = 1$ som en funktion av endast en förklarande variabel: kundmärkeslojalitet för Citrus Hill (LoyalCH)

Vänligen lös och svara på följande deluppgifter (a-c):

a) Beräkna och tolka oddskvoten (OR). (4p)

Lösning

$OR(\hat{\beta}_0) = e^{2.755} = 15.72104$ och $OR(\hat{\beta}_1) = e^{-6.095} = 0.00225411$.

Givet att kundmärkeslojalitet är noll, då är sannolikheten att man köper MM 15 gånger större än sannolikheten att köpa CH. Givet att interceptet är noll, då är sannolikheten

att man köper MM (0.002) mindre än sannolikheten att man köper CH.

b) Beräkna 95%-igt konfidensintervall för oddskvoten (OR) för kundmärkeslojalitet för den enkla modellen. (5p)

Lösning

$KI_{OR(\hat{\beta}_0)} = e^{2.755 \pm 1.96 * 0.19959}$ eller $(10.63132; 23.24746)$ och

$KI_{OR(\hat{\beta}_1)} = e^{-6.095 \pm 1.96 * 0.36115}$ eller $(0.001110601; 0.004575012)$.

```
## ----------------------den beroende variabeln-----------------------
```

	CH	MM
Antal	653.0000	417.0000
Andel	0.6103	0.3897

```
## ------------------------------------------------------------

##
## Parameter estimates
## ------------------------------------------------
##             Estimate Std. Error z value   Pr(>|z|)   2.5 %  97.5 %
## (Intercept)   2.7554    0.19959  13.805 2.3686e-43  2.3754  3.1587
## LoyalCH      -6.0948    0.36115 -16.876 6.7504e-64 -6.8276 -5.4102
##
## Odds ratio estimates
## ------------------------------------------------
##               Estimate Std. Error z value   Pr(>|z|)      2.5 %    97.5 %
## (Intercept) 15.7272831     1.2209  13.805 2.3686e-43 10.6356073 23.256541
## LoyalCH      0.0022546     1.4350 -16.876 6.7504e-64  0.0011109  0.004576
```

c) Figur 5 visar datamaterialet med kundmärkeslojalitet på x-axeln och den binära responsvariabeln om kunderna köper Minute Maid i Orange Juice, Y = 1, på y-axeln. Figuren visar också den anpassade logistiska regressionsmodellen

$$P(Y = 1|x_1) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x_1}}$$

som en röd streckad linje. För vilken kundmärkeslojalitet är sannolikheten att köpa Minute Maid ska vara lika med 0.5, dvs $P(Y = 1|LoyalCH) = 0.5$? I Figur 5 är detta den grönfärgade vertikala linjen. (6p)

Lösning

$P(Y = 1|LoyalCH) = 0.5$ och

$P(Y = 1|x_1) = \frac{e^{2.755-6.095x_1}}{1+e^{2.755-6.095x_1}}$

$\frac{e^{2.755-6.095x_1}}{1+e^{2.755-6.095x_1}} = 0.5$

Man behöver logaritmera för att ta bort den exponentiella funktionen.

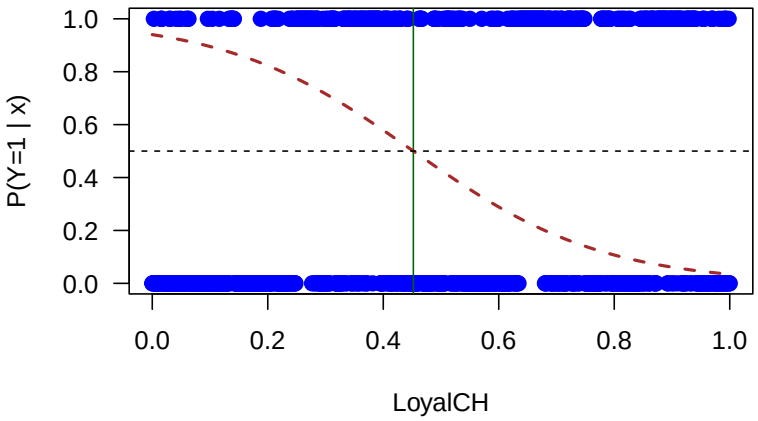
$\frac{log(e^{2.755-6.095x_1})}{log(1+e^{2.755-6.095x_1})} = log(0.5)$

$\frac{2.755-6.095x_1}{log(1+e^{2.755-6.095x_1})} = 0$

kvoten är noll om täljare är noll och då:

$2.755 - 6.095x_1 = 0$ och då $x_1 = \frac{2.755}{6.095} = 0.4520098$

Resultatet stämmer övens med den gröna linjen i Figur 5. Det betyder att kundmärkeslojalitet är ungefär 0.45 för att vara sannolikheten att köpa Minute Maid lika med 0.5.



Figur 5