Stockholm Universitet
Statistiska Institutionen
Maria Anna Di Lucca

Regressions- och tidsserieanalys

Tentamen

Onsdagen den 30 november, 2022

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.

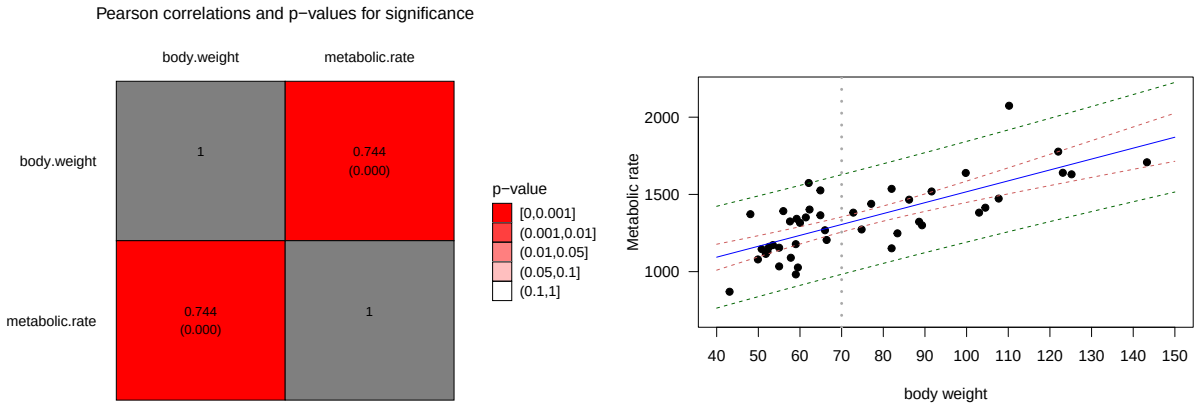


Figure 1: Samband mellan energiomsättning och vikt

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)

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

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

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

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

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

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

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)

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)

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)

```
##
## ANOVA table
## ------------------------------------------------
##       df SS      MS       F Pr(>F)
## Regr  *  *       14655.14 * 0
## Error 64 *       *
## Total 66 34371.7
##
## ------------------------------------------------
```

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

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

e) Beräkna justerat 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)

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

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

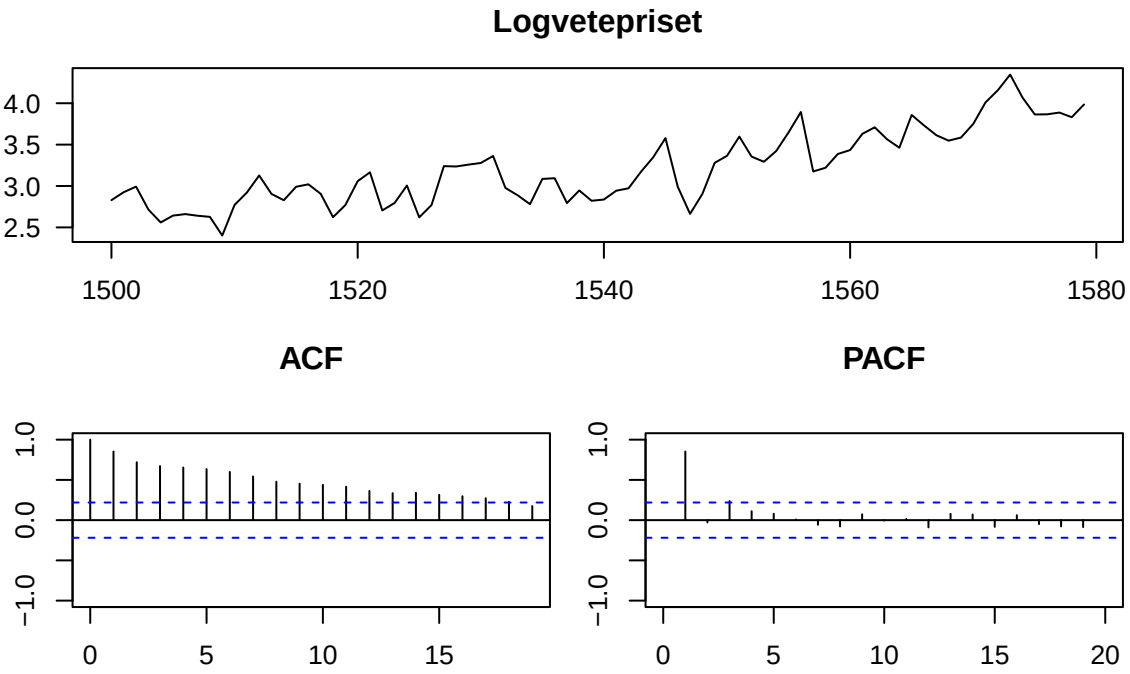
Svara på följande två frågor (a-b).

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

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

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):

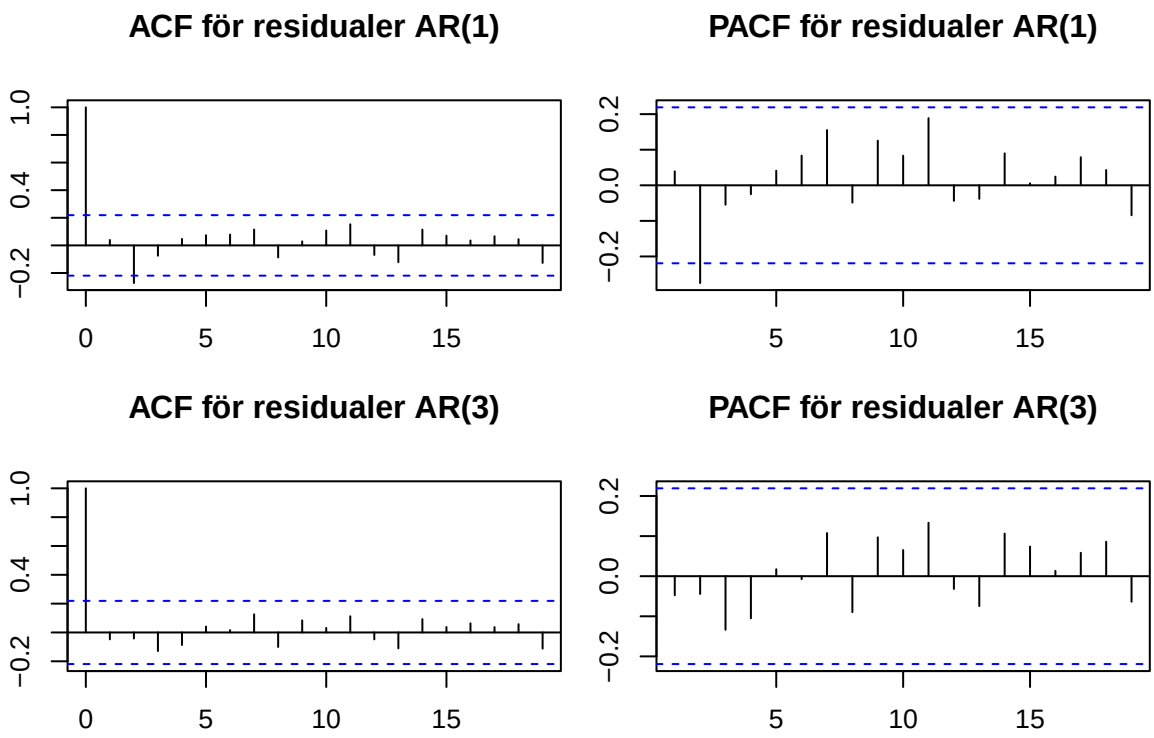
a) Med hjälp av R-utskrifter och residualdiagram i Figur 4.2 motivera om det är bättre en AR(1) modell eller en AR(3) modell. (4p)

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



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)

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

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)

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

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

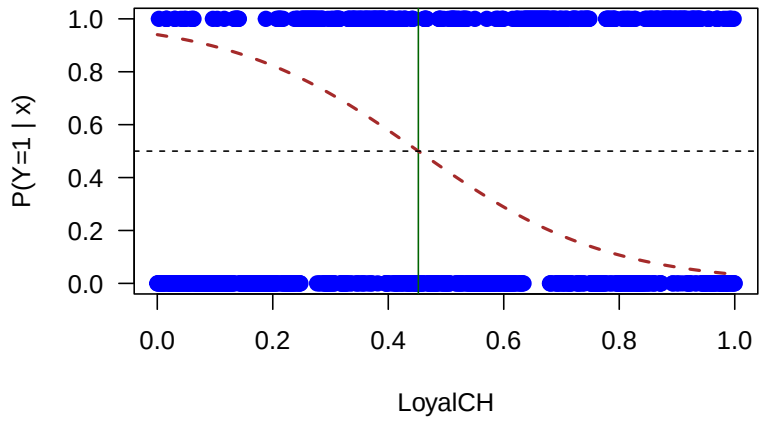
	CH	MM
Antal	653.0000	417.0000
Andel	0.6103	0.3897

```
## ------------------------------------------------------
##
## Parameter estimates
## ---------------------------------------------
##             Estimate Std. Error z value   Pr(>|z|)
## (Intercept)   2.7554    0.19959  13.805 2.3686e-43
## LoyalCH      -6.0948    0.36115 -16.876 6.7504e-64
```

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 lika med 0.5, dvs $P(Y = 1|LoyalCH) = 0.5$? I Figur 5 är detta den grönfärgade vertikala linjen. (6p)



Figur 5