Stockholm Universitet
Statistiska Institutionen
Maria Anna Di Lucca

Regressions- och tidsserieanalys

Tentamen

Måndagen den 9 januari, 2023

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 **2023-01-30**.
- 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)

Data från 16 offentliga elförsörjningsmyndigheter samlades in. En enkel linjär regressionsmodell används för att se om elproduktionen i miljoner kilowattimmar, Y, förklaras med energikostnaderna i US dollar, X, enligt:

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

Nedan visas data

$$\sum x_i = 786.9990 \; \sum y_i = 78.9550 \; \sum x_i y_i = 5556.6630$$
$$\sum x_i^2 = 57416.8702 \; \sum y_i^2 = 552.3916$$

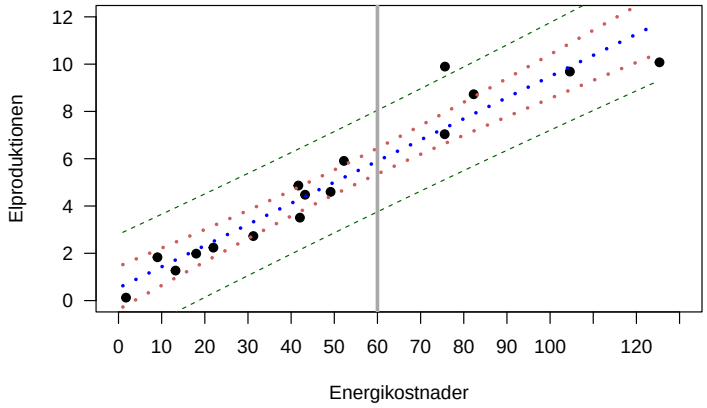


Figure 1: Samband mellan elproduktion och energikostnaderna

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

a) Beräkna urvalskorrelationen och tolka. (5p)

b) Testa om korrelationen är skild från noll på $\alpha = 0.05$. Utan beräkningar, kan man säga att lutningen är skild från noll? Varför? (5+3 p)

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

d) Skatta koefficienterna β_0 och β_1 med minsta-kvadrat-metoden. Tolka (endast) lutningen i ord. (5+3p)

e) Bestäm ett 95%-igt konfidensintervall för den genomsnittliga elproduktionen i kilowattimmar givet att energikostnaderna är lika med 60 (miljoner) dollar. (5p)

f) Med hjälp av formelsamling och resultat i punkt e) beskriv Figur 1. Fundera sen över om du ser något i diagrammet som skulle kunna påverka och för vilket sätt en eventuell residualanalys. (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 $n = 18$ observationer. Den beroende variabeln är antal mord per 1 000 000 invånare per år i US (murders) och de förklarande variablerna är invånare (inhabitants, X_1), procent familjer med inkomst mindre än 5000 dollar per månad (incomes, X_2) och procentandelen arbetslösa (unemployed, X_3).

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

a) Baserat på ANOVA-tabellerna, beskriv vilken modell som är den bästa. Motivera ditt svar. (5p)

b) Beräkna den justerade förklaringsgraden, R^2_{adj} från ANOVA-tabeller för att bekräfta att din valda modell i punkt a) stämmer överens med R^2_{adj}. (5p)

c) Beskriv F-testet för din vald modell: förklara hypoteser, och vad det innebär att förkasta nollhypotesen. Genomför testet på signifikansnivån 5 procent. (5p)

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

```
## --------Modell 1--------------------

##
## Analysis of variance - ANOVA
## ------------------------------------------------
##       df      SS       MS      F     Pr(>F)
## Regr   2 1439.45 719.7254 78.394 1.1435e-08
## Error 15  137.71   9.1809
## Total 17 1577.16

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

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

```
## --------Modell 2--------------------

##
## Analysis of variance - ANOVA
## ------------------------------------------------
##       df     SS       MS     F     Pr(>F)
## Regr   1 1338.1 1338.064 89.54 5.8911e-08
## Error 16  239.1   14.944
## Total 17 1577.2

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

Modell 3 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_3 x_3$

```
## --------Modell 3--------------------

##
```

```
## Analysis of variance - ANOVA
## ------------------------------------------------
##       df      SS       MS      F     Pr(>F)
## Regr   1 1282.02 1282.017 69.498 3.2358e-07
## Error 16  295.15   18.447
## Total 17 1577.16

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

d) I Tabell 2 finns delvis resultat av en regression mellan två forklarande variabler, beräkna VIF. Förklara om det är ett bra eller dåligt resultat och till vilken av tidigare modeller: Modell 1, Modell 2 eller Modell 3 resultatet berör. (5p)

Tabell 2 $\hat{x}_2 = \hat{\psi}_0 + \hat{\psi}_1 x_3$

```
## ----------------Tabell 2 -----------------------------------

##
## Measures of model fit
## ------------------------------------------------
## Root MSE       R2   R2-adj
##  1.95772  0.67588  0.65562
##
## Parameter estimates
## ------------------------------------------------
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)   4.7978    2.63602  1.8201 8.7506e-02
## unemployed    2.1691    0.37553  5.7762 2.8369e-05

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

e) Fyll i följande R-utskrift för Modell 1. (5p)

```
## Parameter estimates
## ------------------------------------------------

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.7095 4.4823     -7.9667 0
## income      *        0.3869     4.141   9e-04
## unemployed  3.3926   *          3.3231  0.0046
```

f) Lista antaganden om feltermer. Beskriv sedan vilka antaganden man testar i Figur 2. (5p=3+2)

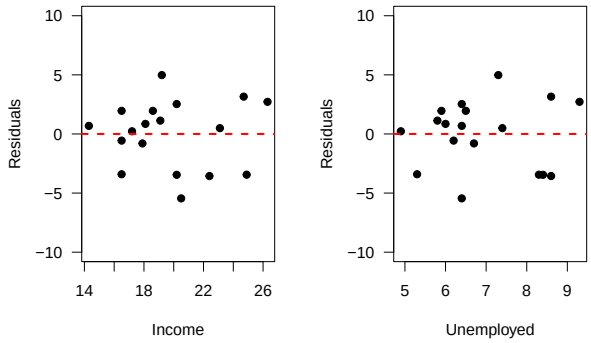


Figure 2: Regressorer och residualer

Uppgift 3. (15 poäng)

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

År	**Kvartal**	**Y**
2019	Kvartal 1	175
2019	Kvartal 2	75
2019	Kvartal 3	104
2019	Kvartal 4	56
2020	Kvartal 1	119
2020	Kvartal 2	9
2020	Kvartal 3	35
2020	Kvartal 4	26
2021	Kvartal 1	35
2021	Kvartal 2	5

Svara på följande frågor (a-c).

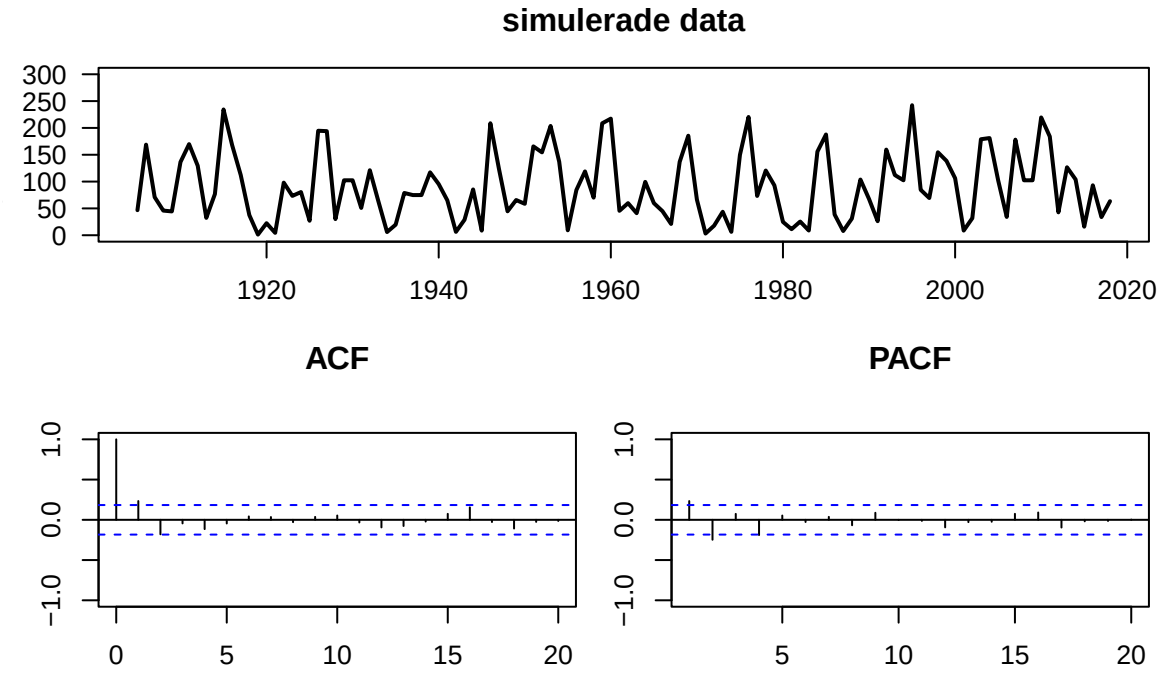
a) Skatta trenden för kvartalsdata med centrerade glidande medelvärden med säsongsvariation i en additiv eller multiplikativ modell. (5p)

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

c) Beräkna den skattade produktionen $\hat{y}_t = \hat{T}_t + \hat{S}_t$ eller $\hat{y}_t = \hat{T}_t\hat{S}_t$ med hjälp av resultat i punkt a) och b). (5p)

Uppgift 4. (10 poäng)

I den här uppgiften analyseras förändringar i simulerade antal lodjur fångade i Canada under de 114 åren 1905-2018. I Figuren 4.a nedan visas den ursprungliga tidserien, autokorrelationen och partiella autokorrelationen för datamaterialet.

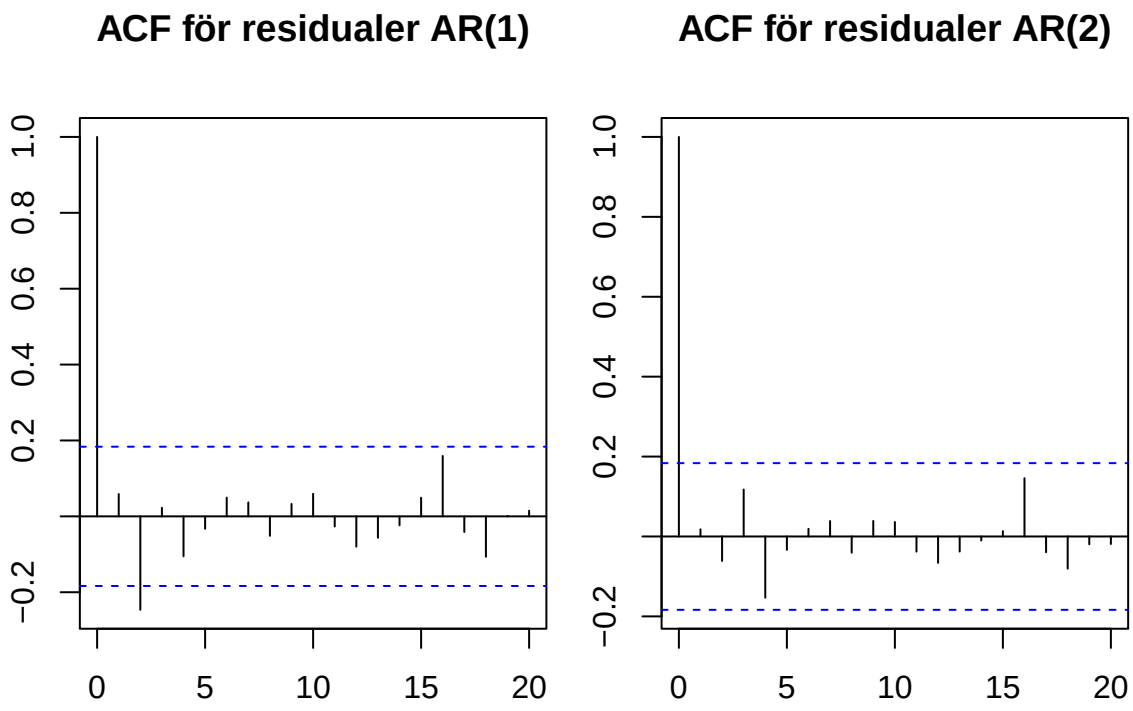


Figur 4.a

Lös följande deluppgifter (a-b):

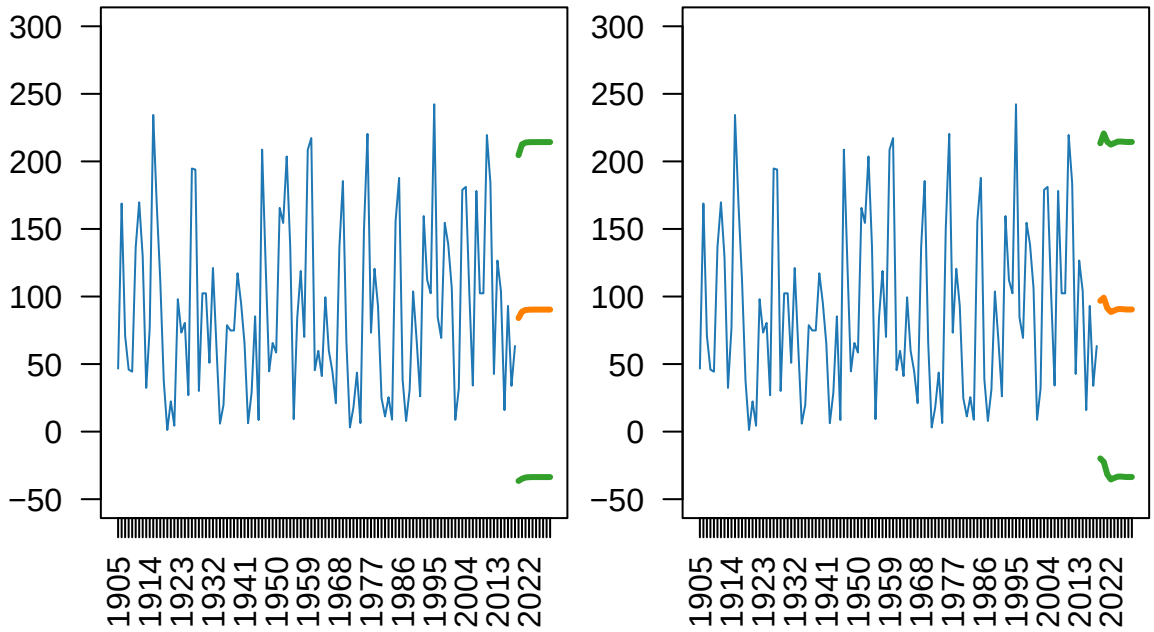
a) Med hjälp av R-utskrifter och residualdiagram i Figur 4.b motivera om det är bättre med en AR(1) modell eller med en AR(2) modell. (5p)

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

##
## Parameter estimates
## ------------------------------------------------------------
##      Estimate Std. Error z-ratio Pr(>|z|)     2.5 %    97.5 %
## ar1   0.23246   0.090867  2.5582 0.010521  0.054359   0.41056
## mean 90.32334   7.486709 12.0645 0.000000 75.649393 104.99729

## --------------------------AR(2) -------------------------------

##
## Parameter estimates
## ------------------------------------------------------------
##      Estimate Std. Error z-ratio  Pr(>|z|)    2.5 %     97.5 %
## ar1   0.29320   0.090751  3.2308 0.0012344  0.11533   0.471069
## ar2  -0.25241   0.091092 -2.7709 0.0055894 -0.43095  -0.073871
## mean 90.43414   5.822323 15.5323 0.0000000 79.02239 101.845893
```



Figur 4.b

b) I nedanstående Figur 4c och Tabell 4 visas prognosticerade fångade lodjur för de tio åren 2019-2028 med AR(1) och respektive AR(2). I tabellen har ett par tal ersatts med "*". Glöm inte att i R-utskrifterna finns medelvärdet, inte interceptet. I modell AR(1) är interceptet lika med 69.32678 och i AR(2) är lika med 86.74533. (5p)



Figur 4.c

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

## Time Series:
## Start = 2019
## End = 2028
## Frequency = 1
##         AR(1)   AR(2)
## 2019 84.0615 96.7869
## 2020 88.8677 99.1241
## 2021  89.985 91.3785
## 2022       * 88.5176
## 2023 90.3051 89.6338
## 2024 90.3191 90.6833
## 2025 90.3224 90.7092
## 2026 90.3231       *
## 2027 90.3233 90.3699
## 2028 90.3233 90.4108
```

Uppgift 5. (10 poäng)

Ett dataset i R innehåller information om förekomsten av diabetes hos n=392 patienter. En logistisk regression har använts för att modellera sannolikheten att en patient har diabetes, $Y = 1$ som en funktion av tre förklarande variabler: glukosnivån (glucose), X_1, Body Mass Index (BMI), X_2 och ålder (Age), X_3. BMI presenteras som en binär variabel. Det beskriver om en person är överviktig eller inte:

$x_2 = 1$, $BMI > 30$, patienten är överviktig

$x_2 = 0$, $BMI \leq 30$, patienten är inte överviktig

Resultatet från antal patienter med diabetes är i Tabell 5 a och resultater från en anpassad logistisk regression med diabetes som responsvariabel visas i Tabell 5b.

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

a) Beräkna sannolikheten för en patient med diabetes givet att patienten hade glukosen lika med 120, BMI är större än 30 och är 30 år gammal. (3p)

b) Tolka effekten av den förklarande variabeln BMI som en oddskvot (OR). (3p)

```
## ----------------------Tabell 5 a-----------------------
```

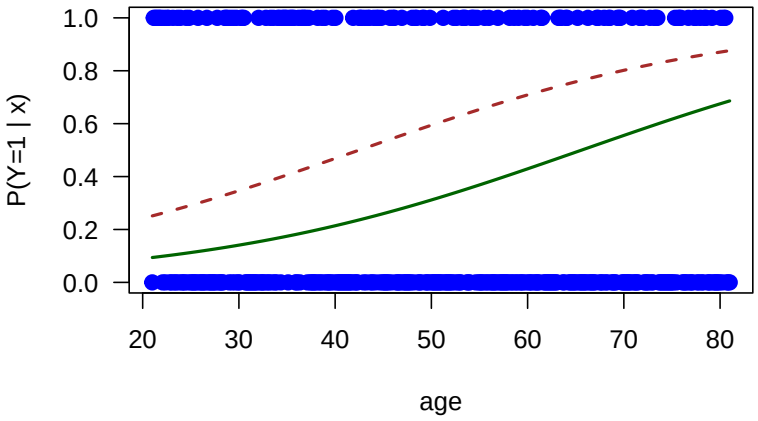
	neg	pos
Antal	262.0000	130.0000
Andel	0.6684	0.3316

```
## ----------------------Tabell 5 b-----------------------

##
## Parameter estimates
## ------------------------------------------------
##                 Estimate Std. Error z value   Pr(>|z|)
## (Intercept)   -7.805059  0.7832872 -9.9645 2.1799e-23
## glucose        0.036475  0.0049279  7.4019 1.3429e-13
## BMIöverviktig  1.174080  0.3161596  3.7136 2.0436e-04
## age            0.050789  0.0130973  3.8778 1.0539e-04
##
## Odds ratio estimates
## ------------------------------------------------
##                  Estimate Std. Error z value   Pr(>|z|)
## (Intercept)   0.00040767     2.1887 -9.9645 2.1799e-23
## glucose       1.03714871     1.0049  7.4019 1.3429e-13
## BMIöverviktig 3.23516654     1.3718  3.7136 2.0436e-04
## age           1.05210114     1.0132  3.8778 1.0539e-04
```

c) Figur 5 visar datamaterialet med ålder på x-axeln och den binära responsvariabeln om patienterna med diabetes, Y = 1, på y-axeln. Figuren visar också den anpassade logistiska regressionsmodellen

$$P(Y = 1|x_1, x_2, x_3) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\hat{\beta}_3 x_3}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\hat{\beta}_3 x_3}}$$

som en streckad linje och en heldragen linje. Förklara skillnaden mellan dessa två kurvor. (4p)



Figur 5