LABORATION 2

732G52 - HT2024

Intro

Innehåll:

- Index
- Tidserieregression
- Prognoser och utvärdering av prognoser

Denna laboration är till för er övning på kursmaterialet. Det finns ingen obligatorisk inläming för den. Uppgifterna utgår från R och Rstudio. Om du behöver repetition i R:

- Titta på kurshemsidan för 732G33: här.
- Cheat sheets här

Material:

- Time Series for Data Science: se R-kod, dataset mm: länk
- Forecasting: Principles and Practice (FPP), tredje upplagan: länk
- Repo för kursen: länk
- Om man behöver repetition på vanlig linjär regression, se kursen Linjära modeller 1, titta tex här: länk

Se sektionen ”The `lm()` function” på sidan 4 för detaljer kring funktionen `lm()`.

Uppgifter

1. Index: Utgå från kompendiet om index som finns på kursrummet. Lös uppgifterna 1, 3 och 4 med hjälp av R. Förslag på lösning finns här: länk OBS försök att lösa uppgifterna själva först innan ni tittar på lösningen.

2. Stationäritet: gå igenom koden här: länk.

3. Regression med tidseriedata: När vi arbetar med regression på tidseriedata så är likt fallet när vi jobbar med "vanlig data" (tänk kursen Linjära modeller 1).

   (a) Det är vanligt att man skapar en tidsvariabel/tidsindex (kalla den *time* här) som vi använder för att modellera trender i tidserien. Detta eftersom vi inte kan räkna på "råa datum" (vilket värde har 4 mars 2003?). Vi kan låta trenden vara linjär, men vi kan använda kvadratisk eller kubisk trend om vi tror att det anpassar data bättre. Då skapar vi lämpliga transformationer på formen: $time$, $time^2$, $time^3$, om vi gör det vill vi ofta centrera eller standardisera $time$ först.

   (b) Dock så gäller fortfarande de vanliga antaganden på feltermen för att inferensen ska vara giltig. Vilka är dessa antaganden? Det är vanligt att det finns ett tidsberoende kvar i residualerna efter att man har anpassat en regressionsmodell på tidseriedata. Då ska vi inte göra vanlig inferens (test, konfidensintervall etc), då antagenden för inferensen inte är uppfyllda. Det finns metoder för att göra inferens här, mer om det senare i kursen.

   (c) Om vi gör tidserieregression och konstaterar att de vanliga antaganden för residualerna inte är uppfyllda, då kan en sådan modell ändå vara användbar i vissa situationer. Modellen kan producera bra prognoser. Även om vi inte kan använda signifikanstester eller konfidensintervall så kan vi använda de skattade regressionskoefficienter som beskrivande statistik för den givna tidserien.

   (d) Anta att vi har en regressionsmodell där vi har en variabel i designmatris, tex ett tidsvariabel $time$. Om vi i den designmatrisen har andra variabler som är en **funktion** av $time$, exempel kan vara $time^2$ och $time^3$, då är det inte meningsfullt att tolka de enskilda regressionskoefficienterna kopplade till $time$, $time^2$ och $time^3$. Om vi vill förstå dessa variablers gemensamma effekt på responsvariabeln, då får vi titta på den kurva som de tillsammans bildar och plotta den mot $time$. I detta fall så plottar vi kurvan $\beta_0 + \beta_1 \cdot time + \beta_2 \cdot time^2 + \beta_3 \cdot time^3$ mot $time$.

      - Det går inte att använda den vanliga tolkningen av β: Säg att $\beta_1 = 3.1$, om vi fösöker med "om $time$ ökar en enhet så kommer y att öka 3.1 enheter, givet att de andra variablerna hålls konstanta" $\rightarrow$ vi kan inte hålla $time^2$ och $time^3$ konstanta om

vi ökar *time* med en enhet → alltså är det meningslöst att tolka β_1 enskillt → vi kollar på $time$, $time^2$, $time^3$ gemensamma kurva

(e) Gå igenom koden som finns här: länk.

4. Ni ska nu analysera data över pappersproduktion. Data finns i filen "`pappersproduktion.csv`" och en beskrivning finns i "`sw_prod_paper_90-04.txt`". Filerna ligger här: länk

   (a) Läs in data som csv-fil i R.

   (b) Skapa en tidsvariabel och gör en tidseriesgraf. Visst finns säsongsvariation. Kommentera.

   (c) Skapa tolv indikatorvariabler, en för varje månad och ge dessa lämpliga namn. Använd 11 av dessa senare i regressionsmodellen.

   (d) Anpassa en regressionsmodell med `lm()`. Ta fram *Four in one* residualplottar, använd residual_diagnostics().

   (e) Tolka valfri dummy-variabel. Utför DW-testet och residualanalys. Ta fram SAC på residualerna. Tolka resultaten.

   (f) Gör en prognos för nästkommande fyra månader. Kommentera.

5. !!!!!!!!!!!!!!!!!!!!!!!!Fortsätt här!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

6. Nu ska ni jobba med ytterligare en tidsserie som innehåller antalet turister till Turkiet som du finner i filen "turisterTurkiet.csv". Beskrivningen av data finns i "tu_foreign_visitors_93-04.txt". Filerna ligger här: länk. Upprepa uppgift 4) för detta datamaterial.

7. Tidserieregression med `fpp3/fable`: gå igenom och återskapa koden koden i kap 2 i fpp3, länk.

8. Tidserieregression med `fpp3/fable`: gå igenom och återskapa koden koden i kap 5.XX i fpp3, länk.

The `lm()` function

The function `lm()` will be important during the course. Check out the documentation with `?lm()`.

Note that R is an object oriented language, and the `lm()` returns objects with class "`lm`", with has the form of a list, so you can easily fetch different part of the object when needed. These objects has several useful generic functions connected to it:

- `coef()`: Gives the regression coefficients
- `residuals()`: calculates the residuals of the model
- `fitted()`: Gives the fitted values of the model
- `summary()`: give detail summary and inference. It will return a object of class "summary.lm". `coef()` will work on this object.
- `anova()`: Calculates anova table for the model
- `predict()`: make predictions with the model for (new) data
- `plot()`: output diagnostics plots for the model

In general, for documentation for these methods run commands of the type `?summary.lm()` in the terminal. Another useful function is to use `str()` on the lm-object, to get detailed information about it.