Miniprojekt: Programmering i R

February 10, 2017

Som en del av kursen i R programmering ska en analys av data göras med Rmarkdown genomföras. Miniprojektet är uppdelat i två delar. Den första delen handlar om att läsa in och bearbeta data från externa datakällor och beskriva dessa data.

I den andra delen av miniprojektet ska mer utförlig analys genomföras samt bearbeta och analysera denna data vidare.

För båda delarna gäller att:

- R-markdown ska användas. En mall kan ni hitta **här**.
- Rapporterna ska lämnas in som både **PDF** och **.Rmd**-fil. Det är ok att skapa en HTML som ni sedan sparar/skriver ut som PDF.
- Samtliga material ska laddas in i R från webben som **externa datakällor**. Vill ni använda ett eget material får ni lägga upp det öppet på github, dropbox, google docs eller dylikt och läsa in det därifrån i R. Syftet är att rapporten ska vara helt reproducerbar och kunna återskapas på godtycklig dator.
- **Inget output från R console ska visas i dokumentet.** Antingen skapar ni tabeller (med `kable()`) eller grafer.
- **Rmd**-filen ska kunna köras och reproducera era resultat. D.v.s. den ska innehålla all er kod som behövs för analysen.
- **Namn**, **liu-id** och **gruppnummer** ska framgå i början av rapporten.

1 Del I: Deskriptiv analys

Den första delen av miniprojektet är att samla in datamaterial och beskriva materialet kortfattat i en första del av rapporten.

Till den miniprojektet behöver ni totalt **tre** datamaterial, två som innehåller kommunala data och ett material som innehåller en tidsserie.

Tänk på att välja material ni själva tycker är intressant!

Kommunala data Två datamaterial (data.frames) ska innehålla data på kommunnivå (d.s.v. för alla 290 kommuner). Ett exempel skulle kunna vara antal invånare i varje kommun. Dataseten ska ha minst **2 variabler** utöver kommunnamn. Ni väljer själv vilka variabler som ska ingå och vilken område data ska komma ifrån. Tanken är att i nästa del av miniprojektetgöra en enklare sambandsanalys mellan dessa variabler.

Tidsserie Hitta ett dataset som innehåller en **tidserie**, det innebär att det finns en variabel som har observerats över tiden. Kravet är att data ska innehålla data på **månadsnivå** och innehålla data från minst 5 år (60 månder). Här ska ni alltså hitta en variabel som observerats under minst 60 tidpunkter, men fler går bra. Data ska alltså innehålla två kolumner, en med variabeln som vi är intresserade av och en med tidpunkterna.

Obs! Tidsperioden ska vara fix, d.v.s ex. jan 2005 - jan 2012.

1.1 Inlämning av del I

Den första inlämningsuppgiften handlar om att läsa in i R och beskriva de material ni valt med R-markdown. Ni ska beskriva era material i text samt sammanfatta de variabler ni valt med de beskrivande statistiska mått som ni själva finner lämpliga. Ta fram beskrivande statistik för **alla** variabler i data. Beroende på hur data ser ut så kan det vara medelvärden, frekvenstabeller mm. Ni kan göra relevanta transformationer av era variabler om ni vill, tex göra en numeriska variabel till en binär och räkna med andelar eller dela in kommunerna i stora, medelstora och små när det gäller befolkning.

Följande saker ska ni göra med data med basgrafiken i R:

1. Ni ska minst ha ett histogram eller barplot per variabel i kommun-materialen
2. En tidsseriegraf/linjediagram för tidseriematerialet
3. En "riktig" tabell, inte bara R output. (**Tips!** `kable()` i paketet `knitr`)

Lämna in rapporten både som en fullt reproducerbar **Rmd**-fil och som **PDF** i LISAM.

- I denna del ska samtliga grafer vara skapade med basgrafiken i R.
- Tabeller ska vara "riktiga" tabeller (med ex. `kable()` i paketet `knitr`), inte utskrifter av R-kod.

2 Del II: Analys

I den första delen av minprojektet har ni valt ut och beskrivit två datamaterial. Nu ska vi fortsätta detta arbete med analyser av materialen. Ni som grupp kommer att ha en del frihet i hur ni utför datanalysen som beskrivs nedan. Det ni ska göra är att bearbeta data, några enkla analyser, lite olika grafer i `ggplot2` och en linjär regression.

2.1 Inlämning del II

Den fulla rapporten ska lämnas in som en fullt reproducerbar **Rmd**-fil och som ett **PDF**-dokument i LISAM. Nedan framgår exakt vilka analyser som ska genomföras.

- I denna del ska samtliga grafer vara skapade med `ggplot2`
- Tabeller ska vara "riktiga" tabeller (med ex. `kable()`), inte utskrifter i R-kod.

2.1.1 Dataanalys av kommundata

Slå samman de två dataseten med kommundata så det blir ett dataset som innehåller variablerna från båda dataseten. Om ni gör rätt här så ska ni få ett dataset med en variabel över kommun och minst 4 andra variabler. Detta kan göras på olika sätt, ett är att använda funktionen `merge()`. Här finns en video för hur ni kan använda `merge()`.

Följande saker ska ni göra med data:

1. Producera minst en scatterplot i `ggplot2` mellan två variabler. Beskriv i text vad ni drar för slutsats.
2. Producera minst ett histogram i `ggplot2`. Beskriv i text vad ni drar för slutsats.
3. Producera minst en barplot, om ni bara har kontinuerliga funktioner kan ni använda `cut()`. Beskriv i text vad ni drar för slutsats.
4. Gör minst ett hypotestest, där ni ställer upp en nollhypotes och sen testar om ni kan förkasta den. Beroende på hur er data ser ut så kan det vara ett t-test eller ett χ^2-test. Har ni inte några kategoriska variabler kan ni använda funktionen `cut()`. Beskriv i text vad ni drar för slutsats.
5. Beräkna korrelationer mellan två variabler och beskriv hur ni tolkar resultatet.
6. Beräkna två olika linjära regressionsmodeller med funktionen `lm()`. Ni väljer en responsvariabel (y-variabel) ni tror kan bero/ha ett samband med de andra variablerna. Välj sedan ut två olika variabler som ni kan ha som förklarande variabler (x-variabler). Skatta sedan två regressionsmodeller

med de två olika förklarande variablerna i var sin regressionsmodell, dvs om era förklarande variabler heter X_1 och X_2, så blir modellerna: $y = \beta_0 + \beta_1 * X_1$ och $y = \beta_0 + \beta_1 * X_2$. Det är okej att använda frekvensdata i regressionmodellerna . Följande saker ska vara med från modellerna:

   (a) Scatterplot mellan föklarande variabel och responsvariabel, tillsammans med skattad regressionlinje i `ggplot2`.

   (b) Histogram över residualerna (felen) **Tips!** `resid()`

   (c) Hurvida lutningen på regressionslinjen är signifikant skild från noll, dvs om p-värdet för β_1 i de olika modellern är mindre är 0.05.

7. I beskrivningen ska ni beskriva er regressionsmodell med LaTeX, ex: $y = \beta_0 + \beta_1 * X_1$. **Här** finns en LaTeX editor för att skapa matematiska uttryck. Beskriv i text vad ni drar för slutsats från de två modellerna.

2.1.2 Dataanalys av tidseriedata

Låt `X` vara er variabel i tidsseriematerialet. Utför nu följande:

1. Gör en linjeplot mellan `X` och er tidsvariabel. Skalan på x-axeln ska vara en lämplig tidsskala.

2. Beräkna medelvärden per månad och spara dessa i `month_means`. **Tips!** `aggregate()`

3. Använd funktionen `summary()` för att fram beskrivande statistik för varje år (det ska vara minst fem år i data)

4. Subtrahera månadsmedelvärden från `X`, så ni tar bort säsongsvariationen i data. Månadsmedelvärdet för januari ska subtraheras från alla januarivärden i data, och likadant för de andra månaderna. Spara den nya tidserie som `new_X`.

5. Gör en linjeplot mellan `new_X` och tid i `ggplot2`. Lägg också till `X` i samma graf som jämförelse.

6. Använd er funktion `my_moving_average()` från tidigare labb och beräkna `moving_average_X`. Lägg till variabel i samma graf som ovan. Totalt ska grafen ha tre linjer i olika färger. Det ska framgå i en legend eller i texten vilken färg som är vilken linje.

7. Verkar det finnas någon trend i data? Dvs ökar/minskar data med tiden, eller är data konstant över tid. Dra er slutsats och skriv ned den i dokumentet.