Statistik och dataanalys I, 15 hp

Inlämningsuppgift 1

Namn 1 Namn 2 Namn 3

1/25/23

Innehåll

> **Varning**
>
> Den här inlämningsuppgiften förutsätter att följande paket finns installerade:
>
> - `mosaic`
> - `dplyr`
> - `geosphere`
> - `leaflet`
>
> Paket kan installeras via kommandot `install.packages('packagename')`, där `'packagename'` är namnet på paketet, t.ex `'mosaic'`.

Introduktion

I den första inlämningsuppgiften ska ni självständigt i grupper om tre analysera ett dataset i programmeringsspråket R. Till skillnad från datorlaborationerna finns det minimalt med

kodexempel. Datorlaborationerna går igenom de flesta momenten som behandlas i inlämningsuppgiften, så se till att göra klart dessa innan.

Instruktioner

I denna inlämningsuppgift ska ni analysera ett datamaterial som innehåller en mängd olika variabler från en totalundersökning[a] i Boston 1970 som aggregerats till ca 500 censusdistrikt. Datasetet förekommer i många olika varianter. Här använder vi en modifierad version[b] av originaldata[c] som använts i en studie[d] där författarna predikterar medianhuspriset i ett censusdistrikt givet en uppsättning förklarande variabler.
Följande variabler finns i datasetet `boston_census_data.Rdata` (ladda ner) för 480 observationer. Notera att en observation motsvarar ett censusdistrikt:

- `town`: Stadsdel.
- `longitude`: Longitud koordinat.
- `latitude`: Latitud koordinat.
- `median_home_value`: Medianhuspriset (enhet 1K USD).
- `crime_rate`: Brott (per 1000 invånare).
- `zoned_25k_p`: Andel av stadsdelens bostadsmark ämnad för marklotter större än 25000 kvadratfot.
- `indust_p`: Andel tunnland ägd av företag utanför detaljhandel.
- `borders_charles`: Charles River dummy variabel (= 1 om området gränsar till floden, 0 annars).
- `NOx`: Koncentration av kväveoxider (andelar per 10 miljon).
- `n_rooms_avg`: Genomsnitt antal rum i ägda bostäder.
- `before_1940_p`: Andel ägda bostäder byggda före 1940.
- `employ_dist`: Viktat avstånd till fem arbetsförmedlingscentra i Boston.
- `radial_access`: Index som mäter tillgång till stadsmotorvägar.
- `tax_rate`: Fastighetsskatt per 10000 USD.
- `pupil_teacher_ratio`: Lärartäthet mätt som elev per lärare.
- `lower_stat_pct`: Procentandel underklass definerad som en av två: (i) andel vuxna utan gymnasieutbildning eller (ii) andel män som genomför okvalificerat arbete.

Bland de förklarande variablerna som använts i studien (ej med i datasetet) finns en icke-linjär interaktion av latitud och longitud koordinaterna för att modellera medianhuspriset spatiellt (dvs deras modell använder censusdistriktens geografiska platser för att fånga den spatiella variationen i huspriser, dvs geografisk variation). Det här sättet att modellera spatiellt beroende är överkurs[e], så ni kommer att få göra följande förenkling för att fånga det geografiska beroendet i medianhuspriset. Genom att använda latitud och longitud koordinaterna kan ni beräkna avståndet till en central plats i Boston. Ni

kan sedan inkludera detta avstånd som en förklarande variabel i en regressionsmodell, för att se om den förklarar variation i medianhuspriserna.
I sista uppgiften ska ni föreslå en prognosmodell för medianhuspriset där ni får välja vilka förklaringsvariabler ni vill ha med (ni vår välja bland en delmängd av de som listas ovanför, se Uppgift 5.4). Ni ska sedan använda er modell för att prognostisera medianhuspriset för tio censusdistrikt i datasetet `boston_districts_to_predict.Rdata` (ladda ner). Det här datasetet har endast de förklarande variablerna, dvs alla de variabler ni får använda förutom medianhuspriset. När vi rättar era inlämningsuppgifter kommer vi att jämföra prognoserna mot de faktiska värden (vi har tillgång till dessa). De tre bästa prognosmakarna kommer att publiceras på hemsidan.
Inlämningsuppgiften ska lämnas in i form av ett html dokument genererat av Quarto. **Kontrollera noga att du inte har några felmeddelande och att dokumentet kompileras utan problem**. Använd tydliga figurer och namnge axlarna med tydliga variabelnamn. Glöm inte att skriva era namn ovanför istället för Namn 1, Namn 2 och Namn 3.

[a] Kallas för census survey på engelska. En statistisk undersökning där hela populationen undersöks.
[b] Totalundersökningen trunkerade medianhusvärdet till 50K för de censusdistrikten som låg över. Vi har tagit bort dessa censusdistrikt. Vi har också tagit bort variabler som är irrelevanta.
[c] Harrison Jr, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. Journal of Environmental Economics and Management, 5(1), 81-102.
[d] Pace, R. K., & Gilley, O. W. (1997). Using the spatial configuration of the data to improve estimation. The Journal of Real Estate Finance and Economics, 14(3), 333-340.
[e] Fortsätt läsa Statistik och Dataanalys II så lär ni er hur man gör detta.

0. Ladda in data

Uppgift 0.1

Ladda in dataseten `Boston_census_data.Rdata` och `Boston_districts_to_predict.Rdata` (länkar för att ladda ner data finns i Instruktioner avsnittet ovan).

Uppgift 0.1 - Svar

```
# Write your code here
load(file = url("https://github.com/StatisticsSU/SDA1/blob/main/assignments/assignment1/
```

1. Kriminalitet i Boston

I detta avsnitt ska ni analysera kriminaliteten i Boston med hjälp av variabeln `crime_rate`.

Uppgift 1.1

Vad kan man generellt säga om kriminaliteten i censusdistrikten? Använd lämpliga figurer samt fördelningsmått som underlag.

Uppgift 1.1 - Svar

Skriv svaret här. Vid behov skrivs matematiska symboler inom dollartecken, till exempel $\overline{y} = \sum_{i=1}^{n} y_i$. Koden skrivs i R-rutan nedanför.

```
# Write your code here
```

Uppgift 1.2

Varierar brottsligheten i Boston beroende på den kategoriska variabeln `town`? Det finns 88 olika utfall av `town` (dvs 88 olika stadsdelar). Välj ut `Boston East Boston`, `Boston Downtown`,`Cambridge`, samt två valfria stadsdelar för att besvara frågan. Frågan besvaras med hjälp av lämpligt valda figurer och statistiska mått.

Tips

Skapa en ny data frame som filtrerar `Boston_census_data` (till exempel genom `filter()` funktionen) utefter de stadsdelarna ni är intresserade utav innan ni påbörjar analysen.

Uppgift 1.2 - Svar

Skriv svaret här.

```
# Write your code here
```

Uppgift 1.3

Vilka tre variabler i datasetet `Boston_census_data` korrelerar mest med brottslighet? Beskriv det parvisa sambandet mellan brottslighet och vardera av dessa tre variabler.

Tips

Kom ihåg att korrelation är ett beroendemått för *numeriska variabler*.

Uppgift 1.3 - Svar

Skriv svaret här.

```
# Write your code here
```

2. Fastighetsskatt i Boston

I detta avsnitt ska ni analysera fastighetsskatten i Boston med hjälp av variabeln `tax_rate`.

Uppgift 2.1

Vad kan man generellt säga om fastighetsskatten i censusdistrikten? Använd lämpliga figurer samt fördelningsmått som underlag.

Uppgift 2.1 - Svar

Skriv svaret här.

```
# Write your code here
```

Uppgift 2.2

Låt oss skapa en ny variabel `cat_tax` som anger om ett censusdistrikt betalar låg (`low`), medel (`medium`), eller hög (`high`) fastighetsskatt. Vi definerar skattekategorierna enligt

- `low`: `tax_rate` ≤ 250,
- `medium`: $250 <$ `tax_rate` ≤ 400,
- `high`: `tax_rate` > 400.

Följande kod skapar och lägger till variabeln `cat_tax` i `Boston_census_data`

```
Boston_census_data$cat_tax <- cut(Boston_census_data$tax_rate,
              breaks=c(0, 250, 400, 800),
              labels=c('Low', 'Medium', 'High'))
```

Finns det ett samband mellan vilken skattekategori ett censusdistrikt tillhör och dess angränsning till Charles River? Förklara med hjälp av lämplig tabell samt figur.

Uppgift 2.2 - Svar

Skriv svaret här.

```
# Write your code here
```

Uppgift 2.3

Hur många procent av alla censusdistrikt ligger i angränsning till Charles River och tillhör en hög skattekategori? Hur stor andel av censusdistrikten med hög skatt ligger inte i angränsning till Charles River?

Uppgift 2.3 - Svar

Skriv svaret här.

```
# Write your code here
```

Uppgift 2.4

Vilka två variabler i datasetet `Boston_census_data` korrelerar mest med `tax_rate`? Beskriv det parvisa sambandet mellan `tax_rate` och vardera av dessa två variabler. Är dessa korrelationssamband eller kausala samband?

Tips

Kom ihåg att korrelation är ett beroendemått för *numeriska variabler*.

Uppgift 2.4 - Svar

Skriv svaret här.

```
# Write your code here
```

3. Avstånd till Fenway park

I detta avsnitt ska ni skapa en ny variabel som mäter avståndet till Fenway park (stadion där basebollslaget Boston Red Sox spelar sina hemmamatcher). Genom variablerna `latitude`

och `longitude` kan vi beräkna det så kallade cirkelavståndet[1] till Fenway park för varje distrikt. Formeln för cirkelavståndet är ganska komplicerad, men den finns implementerad i funktionen `distHaversine()` i R-paketet `geosphere`. Följande kod beräknar avståndet till Fenway park för varje censusdistrikt och sparar den som en ny variabel `dist_fenway_park` i `Boston_census_data`.

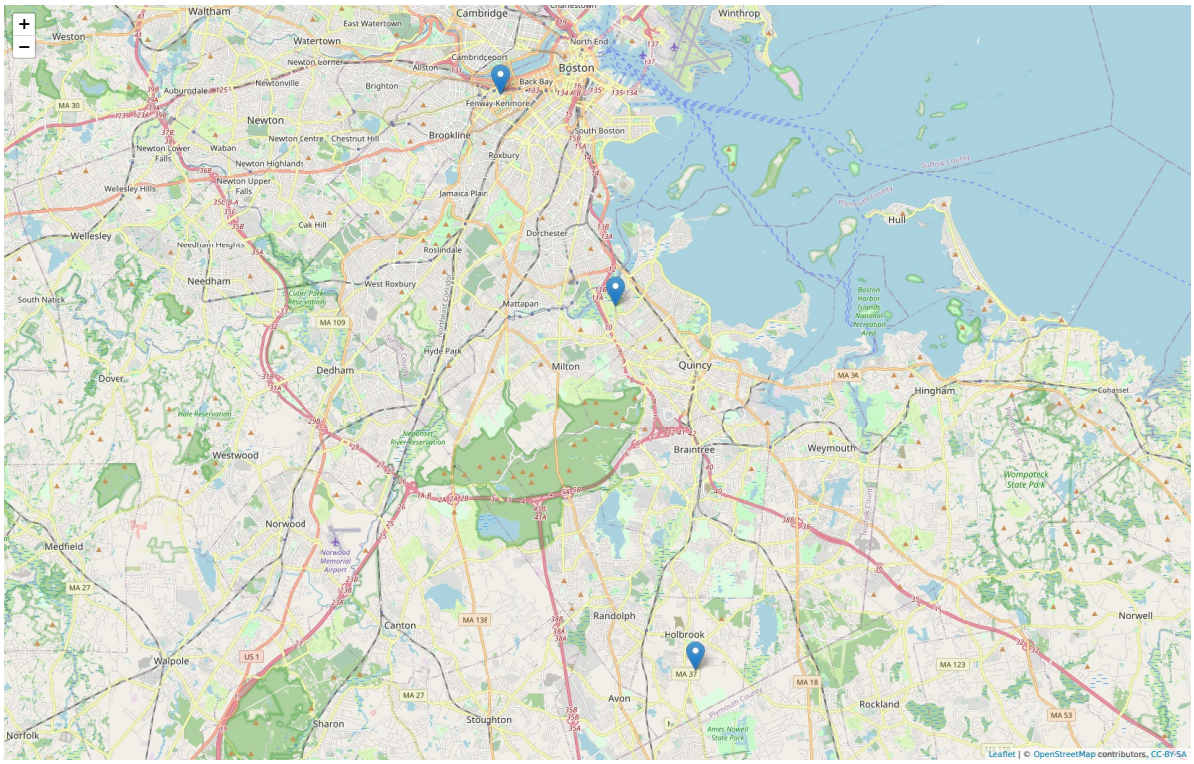
```
library(geosphere) # Install if not available
lat_long <- cbind(Boston_census_data$latitude, Boston_census_data$longitude)
fenway_park_lat_long <- c(42.346462, -71.097250) # latitude and longitude for Fenway_park
Boston_census_data$dist_fenway_park <- distHaversine(lat_long, fenway_park_lat_long)
```

Vi kan visualisera Fenway park samt censusdistrikten i en interaktiv karta med hjälp av R-paketet `leaflet`. Följande kod visualiserar Fenway park samt censusdistrikten för observationerna 30 och 45.

```
library(leaflet) # Install if not available
Boston_map <- leaflet() %>%
  addTiles() %>%
  addMarkers(lat = fenway_park_lat_long[1], lng = fenway_park_lat_long[2], popup="Fenway p
  addMarkers(lat = Boston_census_data$latitude[30], lng = Boston_census_data$longitude[30]
  addMarkers(lat = Boston_census_data$latitude[45], lng = Boston_census_data$longitude[45]

Boston_map # Show interactive map
```

[1] Ett avstånd mellan två punkter uttryckta i latitud och longitud koordinater som tar hänsyn till att jorden är rund. Se här för detaljer.



Uppgift 3.1

Gör ett histogram för variabeln `dist_fenway_park`. Vilket av censusdistrikten har längst respektive kortast avstånd till Fenway park? Markera ut dessa distrikt i en interaktiv karta tillsammans med Fenway park.

> **Tips**
>
> När ni vet vad det längsta respektive kortaste avståndet är så kan ni använda `filter()` funktionen för att filtrera `Boston_census_data` på ett lämpligt sätt.

Uppgift 3.1 - Svar

Skriv svaret här.

```
# Write your code here
```

Uppgift 3.2

Finns det ett samband mellan `dist_fenway_park` och `median_home_value`?

Uppgift 3.2 - Svar

Skriv svaret här.

```
# Write your code here
```

Uppgift 3.3

Finns det ett samband mellan `dist_fenway_park` och `crime_rate`?

Uppgift 3.3 - Svar

Skriv svaret här.

```
# Write your code here
```

4. Enkel linjär regression

I detta avsnitt ska ni anpassa och tolka några enkla linjära regressionsmodeller.

Uppgift 4.1

Anpassa en linjär regression med responsvariabel `NOx` och förklarande variabel `employ_dist`. Rita den anpassade regressionslinjen tillsammans med data i en lämplig figur. Beskriv resultaten och tolka modellen. Utför en modellvalidering via en residualanalys och kommentera modellens lämplighet. Om modellen inte anses lämplig, vilka antaganden har inte varit uppfyllda?

Uppgift 4.1 - Svar

Skriv svaret här.

```
# Write your code here
```

Uppgift 4.2

Använd modellen i Uppgift 4.1 för att prediktera genomsnittsutsläppet för observation 10 med `employ_dist`=10.5857 och beräkna dess residual.

Uppgift 4.2 - Svar

Skriv svaret här.

```
# Write your code here
```

Uppgift 4.3

Använd Tukeys cirkel för att transformera variablerna i Uppgift 4.1 (avgör själv vilken eller vilka av de två som behöver transformeras). Anpassa en ny linjär regression på de transformerade data. Utför en modellvalidering (efter transformation) via en residualanalys och kommentera modellens lämplighet jämfört med modellen i Uppgift 4.1. Plotta den anpassade regressionen i icke-transformerad skala tillsammans med data (också i icke-transformerad skala) i en lämplig figur.

Tips

Tänk på att ta hänsyn till eventuella transformationer!

Uppgift 4.3 - Svar

Skriv svaret här.

```
# Write your code here
```

Uppgift 4.4

Använd modellen i Uppgift 4.3 för att prediktera genomsnittsutsläppet för observation 10 med `employ_dist`=10.5857 och beräkna dess residual. Kommentera resultaten jämfört med Uppgift 4.2.

Tips

Tänk på att ta hänsyn till eventuella transformationer!

Uppgift 4.4 - Svar

Skriv svaret här.

```
# Write your code here
```

5. Multipel linjär regression

I detta avsnitt ska ni studera multipel linjära regression.

Uppgift 5.1

Anpassa en linjär regression med responsvariabel logaritmerad `median_home_value` samt förklarande variabler `lower_stat_pct` och dummy-variabeln `borders_charles`. Tolka koefficienten för `borders_charles`.

Uppgift 5.1 - Svar

Skriv svaret här.

```
# Write your code here
```

Uppgift 5.2

Anpassa en linjär regression med responsvariabel `NOx` samt förklarande variabler `lower_stat_pct` och dummy-variabeln `borders_charles`. Vad tror ni om den statistiska signifikansen för respektive förklarande variabel?

Uppgift 5.2 - Svar

Skriv svaret här.

```
# Write your code here
```

Uppgift 5.3

Använd modellen i Uppgift 5.1 för att prediktera `median_home_value` för observation 30 och beräkna dess residual.

Tips

Tänk på att ta hänsyn till log-transformationen i den anpassade modellen!

Uppgift 5.3 - Svar

Skriv svaret här.

```
# Write your code here
```

Uppgift 5.4

Ni ska nu utveckla en prognosmodell för medianhuspriset `median_home_value`. Ni får endast välja mellan följande förklarande variabler samt godtyckliga transformationer av dom (ni får även transformera responsen):

- `before_1940_p`
- `crime_rate`
- `radial_access`
- `NOx`
- `dist_fenway_park` (som skapades i Avsnitt 3).

Det finns $2^5 = 32$ olika sätt att inkludera de olika förklarande variabler och därmed 32 olika modeller man kan testa, plus i princip hur många som helst om vi också transformerar. Vi förväntar oss naturligtvis inte att ni går igenom varje möjlig modell, men vi förutsätter att ni testar er fram metodiskt.

För att utvärdera mellan olika modeller kan ni använda justerat R-kvadrat samt korsvalidering med 4 folds. **Sortera inte 'boston_census_data.Rdata slumpmässigt när ni korsvaliderar (data ligger redan i slumpmässig ordning)**. Dela upp datasetet i fyra delar när ni korsvaliderar (del 1: observationer 1-120, del 2: observationer 121-240, del 3: observationer 241-360, del 4: observationer 361-480).

Tips

Tänk på att ta hänsyn till eventuell transformation av responsvariabeln när ni utför korsvalideringen. Korsvalideringen använder prediktionen $\hat{y}$ som är prediktionen av y. Exempelvis, om ni har valt transformationen $\log(y)$ är modellens prediktion av responsen $\widehat{\log(y)}$. När ni korsvaliderar blir då $\hat{y} = \exp\left(\widehat{\log(y)}\right)$ prediktionen av y.
Om ni använder `reg_crossval()` funktionen från kurspaketet `sdakurs` tänk då på två saker:

- Använd argumentet `obs_order = 1:480` för att inte data ska sorteras slumpmässigt.
- Funktionen kan inte hantera en transformerad respons.

Vill man transformera responsen kan man följa korsvalideringsexemplet (med transformerad respons) i Lab 4.

Uppgift 5.4 - Svar

Skriv svaret här.

```
# Write your code here
```

Uppgift 5.5

Gör en residualanalys av den valda modellen i Uppgift 5.3.

Uppgift 5.5 - Svar

Skriv svaret här.

```
# Write your code here
```

Uppgift 5.6

Använd modellen i Uppgift 5.4 för att prediktera medianhuspriset för observationerna i datasetet `Boston_districts_to_predict.RData`. Skriv ut resultatet så att vi enkelt kan jämföra dina prognoser när vi rättar.

Tips

Tänk på att ta hänsyn till eventuella transformationer av de förklarande variablerna! Om ni har `dist_fenway_park` med i er prognosmodell behöver ni räkna ut dess värde för observationerna i datasetet `Boston_districts_to_predict.RData` (genom att använda latitud och longitud variablerna såsom i Avsnitt 3).

Uppgift 5.6 - Svar

Skriv svaret här.

```
# Write your code here
```