Statistik och dataanalys I, 15 hp

Inlämningsuppgift 2, 1.5 hp

Mattias Villani

7/24/23

Innehåll

⚠ Installation av nödvändiga paket

Den här inlämningsuppgiften förutsätter att följande paket finns installerade:

- `mosaic`
- `gplots`
- `remotes`
- `sda123`

De tre första paketen kan installeras som vanligt via kommandot `install.packages('packagename')`, där `'packagename'` är namnet på paketet, t.ex `'mosaic'`.
Det sista paketet, `sda123`, är SDA-kursernas egna R-paket och installeras med kommandot
`install_github("StatisticsSU/sda123")`
efter att du laddat in `remotes` paketet.

Introduktion

I denna andra inlämningsuppgift ska ni självständigt i grupper om tre analysera ett datamaterial i programmeringsspråket R, med fokus på sannolikhetslära och inferens. 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 med 1602 australiska hushålls elkonsumption[a], och finns i kursens R-paket `sda123` och heter `electricitycost`. När du installerat och laddat in `sda123`-paketet finns `electricitycost` tillgängligt som en dataframe, dvs en tabell där raderna är observationer (hushåll) och kolumnerna är variabler, t ex hushållets kostnad för el och information om hushållets storlek och utrustning. Se nedan för mer information.

Till skillnad från den tidigare inlämningsuppgiften **ska ni i denna inlämningsuppgift arbeta i ett separat Quarto-dokument där ni skriver alla svar**. Det här dokumentet som du läser nu innehåller alltså bara instruktioner och frågorna. Det Quarto-dokument som ni ska göra analysen och skriva svaren i finns här.

I många uppgifter vill jag att ni ska använda både formelsamlingen för att beräkna en sak (t ex ett hypotestest), men även färdiga funktioner i R (t ex `t.test` funktionen). När ni använder formelsamlingen får ni använda R för att beräkna de saker ni behöver i formlerna, t ex `sd`-funktionen för att beräkna standardavvikelsen, eller `qt`-funktionen för att beräkna ett kritiskt värde från t-fördelningen. På det sättet tränas ni både på att hantera och förstå formeln (tentan!) och hur man använder R i praktiken . Det kan också vara bra träning att leta upp alla kritiska värden i tabellerna, även om jag inte ber om det.

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 i Quarto-dokumentet istället för Namn 1, Namn 2 och Namn 3.

Alla gruppmedlemmar ska vara delaktiga och bidra till alla delar av rapporten och arbetet som leder upp till rapporten, dvs skriva kod, analysera data, tolka resultat, dra slutsatser och skriva rapporten.

[a] Bartels, R., Fiebig, D. and Plumb, M. (1996). Gas or electricity, which is cheaper? An econometric approach with application to Australian expenditure data, The Energy Journal 17(4): 33–58.

0. Läsa in data

Datamaterialet `electricitycost` läses in via kurspaketet `sda123`:

```
library(remotes)
#install_github("StatisticsSU/sda123")
library(sda123)
head(electricitycost)
```

```
  cost rooms people income onlysecondary waterheat cookel poolfilt airrev
1  545     7      4  29900             0         0      1        0      1
2  389     7      2  11700             0         0      1        0      0
3  390     8      2  16900             0         0      0        0      0
4  268     7      2   9750             1         0      1        0      1
5  543     6      2  24700             1         0      0        0      1
6  278     6      3   8450             1         0      0        0      0
  aircond microwave dish dryer
1       1         0    0     0
2       0         1    1     0
3       1         0    0     0
4       1         0    0     1
5       1         1    0     1
6       0         0    0     1
```

Varje rad i datamaterialet är ett av de 1602 australiska hushållen. Skriv `?electricitycost` i Console för att få en komplett beskrivning av alla variabler. För att inte behöva skriva det långa namnet `electricitycost` hela tiden kan vi definera en ny variabel `df` (förkortning av dataframe)

```
df = electricitycost
```

1. Poisson-modell för antal personer i hushållet

Uppgift 1.1

Variabeln `people` innehåller antal personer i hushållet. Definiera variabeln `extrapeople = df$people - 1`, som mäter antalet personer *utöver* ägaren (som vi antar är bara en person). I den här uppgiften ska vi modellera `extrapeople` som oberoende observationer från en Pois(λ)-fördelning. Skatta parametern λ från datamaterialet.

Uppgift 1.2

Undersök grafiskt om den skattade Poisson-modellen i Uppgift 1.1 anpassar data väl.

Uppgift 1.3

Använd den skattade Poisson-modellen för att beräkna sannolikheten för ett storhushåll, vilket vi definierar som ett hushåll med fler än 4 personer.

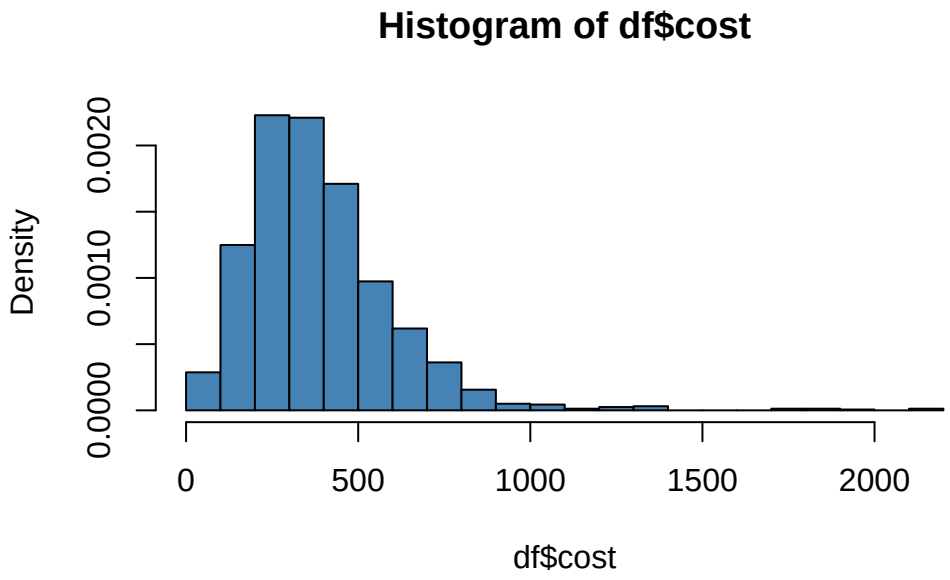
Uppgift 1.4

Poissonmodellen är en trevlig modell för räknedata, men är begränsad eftersom väntevärdet och variansen alltid måste vara lika in en Poissonfördelning. Verkar det vara ett problem för variabeln `extrapeople`?

2. (log)-normalmodell för elkostnad

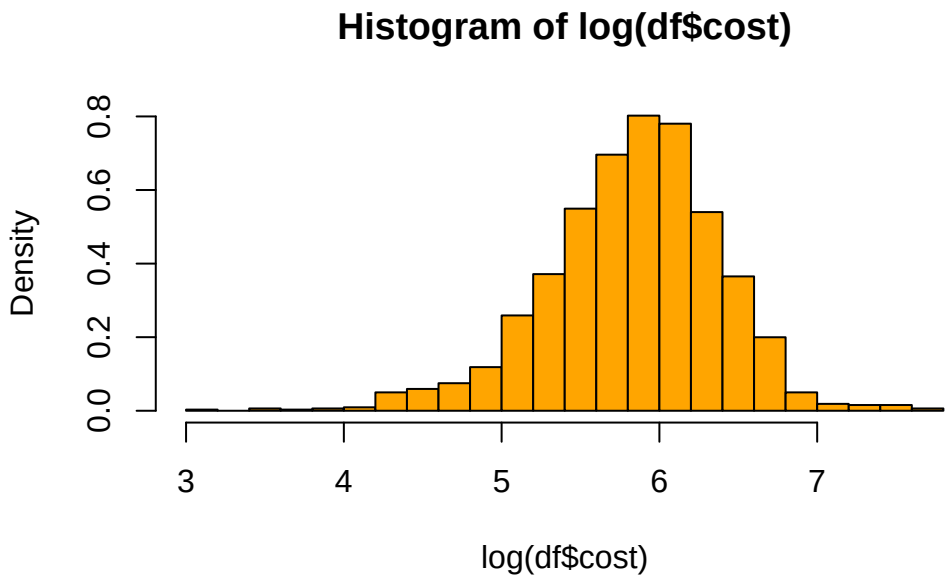
Hushållens totala elkostnad, `cost`, är rejält skev:

```
hist(df$cost, 30, freq = FALSE, col = "steelblue")
```



Logaritmen av `cost` har en fördelning som är mycket mer symmetrisk, även om viss skevhet verkar kvarstå:

```r
hist(log(df$cost), 30, freq = FALSE, col = "orange")
```



Uppgift 2.1

I den här uppgiften ska vi modellera variabeln `logcost = log(df$cost)` som oberoende observationer från en $N(\mu, \sigma)$ fördelning. Skatta μ och σ från data. [Obs! Ni behöver inte transformera tillbaka till originalskala, utan arbeta med hela Uppgift 2 på log-skala.]

Uppgift 2.2

Undersök grafiskt hur väl den skattade normalmodellen passar variabeln `logcost`.

Uppgift 2.3

Jämför median av variabeln `logcost` med medianen från den skattade sannolikhetsmodellen från Uppgift 2.1. [hint: lösningen blir väldigt enkel här, det är meningen. Jag vill att ni ska tänka på kopplingen mellan den teoretiska sannolikhetsmodellen och hur den relaterar till data.]

Uppgift 2.4

Gör ett 95%-igt konfidensintervall för μ i modellen för `logcost` från Uppgift 2.1, **både** genom att använda formelsamlingen och genom att använda en funktion i R. Tolka intervallet. [Som

jag skrev i instruktionerna ovan är det ok att använda R för att beräkna delar av konfidensintervallet även i fallet där jag ber om att ni ska använda formelsamling; t ex använda `sd()`-funktionen för att beräkna standardavvikelsen `s`. Men jag vill att ni använder formeln för konfidensintervall från formelsamlingen i den slutliga uträkningen. För att träna inför tentan. Och sen jämföra med det konfidensintervall ni får direkt från R].

Uppgift 2.5

Testa om den genomsnittliga `logcost` i modellen/populationen är mindre än 6. Ställ upp noll- och alternativhypotes och testa på 5% signifikansnivå. Gör beräkningarna både med hjälp av formelsamlingen och med R.

Uppgift 2.6

Beräkna p-värdet för testet i Uppgift 2.5 genom att använda R. Hade du förkastat nollhypotesen på 1% signifikansnivå?

Uppgift 2.7

Antag att vi tar den skevheten som vi ser i histogrammet över `logcost` på allvar. Beskriv (inga uträkningar) om vi ändå kan göra ett hypotestest utan att anta att `logcost` är normalfördelad.

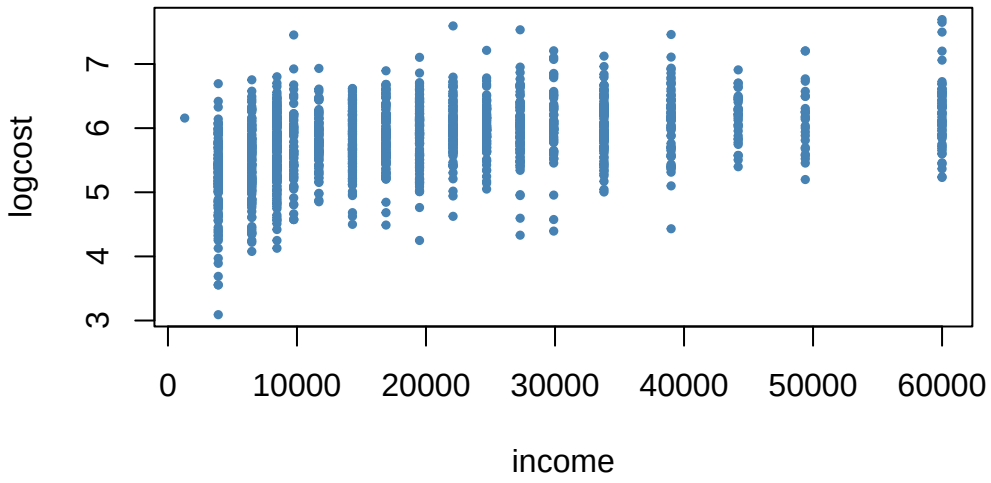
Uppgift 2.8

Testa på 5% signifikansnivå om det finns någon skillnad i genomsnittlig `logcost` i modell/populationen för hushåll med och utan air conditioner. Ställ upp noll- och mothypotest och beräkna teststatistikans värde med hjälp av formelsamlingen. Använd R för att beräkna frihetsgraderna i nollhypotesens t-fördelning (som är en komplicerad beräkning eftersom vi har olika antal observationer i de två grupperna).

3. Enkel och multipel linjär regression

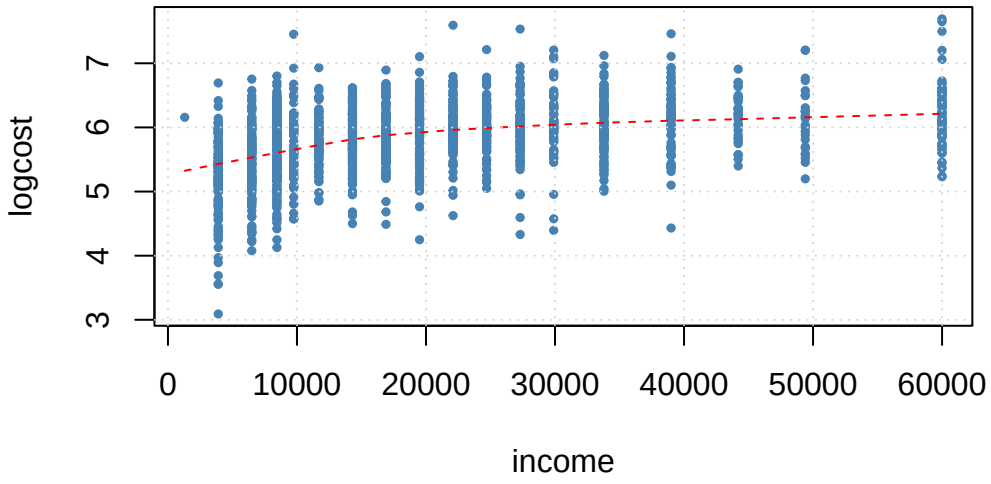
Ni ska nu analysera en regressionsmodell med `logcost` som responsvariabel. Vi börjar med hushållets inkomst `income` som förklarande variabel. Det är svårt att se om en linjär regression passar data eftersom `income` bara kan anta ett mindre antal värden (den verkar vara grupperad i inkomstklasser):

```
df$logcost = log(df$cost) # lägger in logcost i dataframen, blir mindre kod då.
plot(logcost ~ income, data = df, pch = 19, cex = 0.5, col = "steelblue")
```



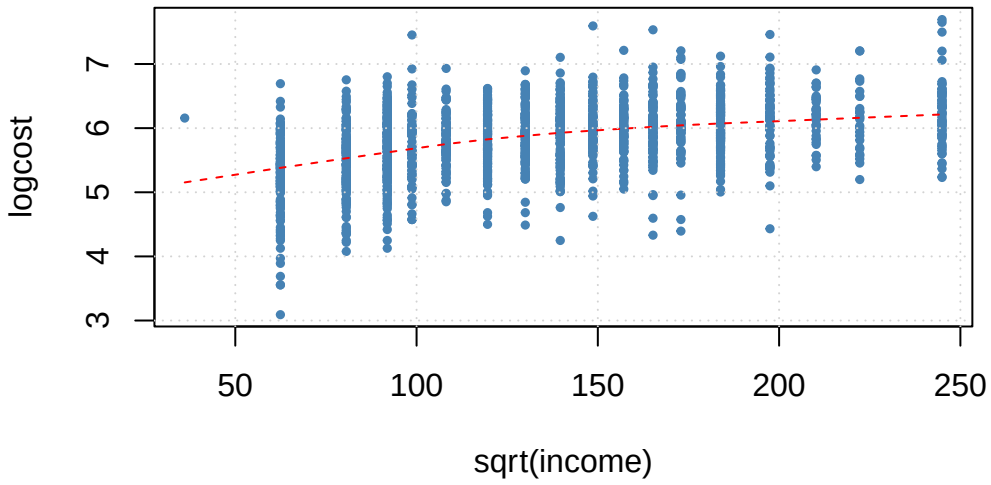
För att lättare se om det verkar linjärt anpassar jag en s k *lowess*-skattning (en slags icke-linjär regression) och plottar anpassningen, med funktionen `plotLowess` från `gplots` paketet:

```
plotLowess(logcost ~ income, data = df, pch = 19, cex = 0.5, col = "steelblue")
```



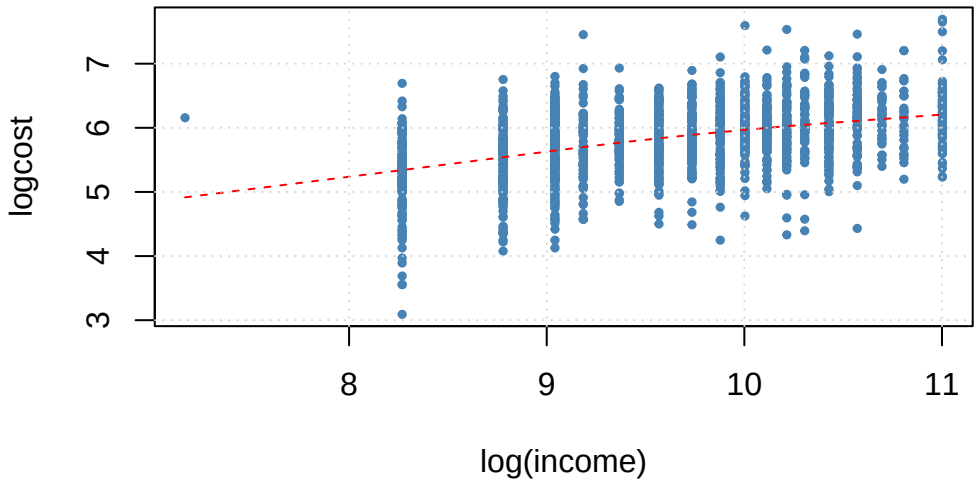
Det ser inte linjärt ut. Låt oss prova att gå ett steg nedåt på Tukey's transformationsstege (i Tukey's cirkel är vi i övre vänstra kvadraten, vilket indikerar att vi ska gå nedåt på stegen för X-variabeln) och göra en $\sqrt{}$-transformation (sqrt):

```
plotLowess(logcost ~ sqrt(income), data = df, pch = 19, cex = 0.5, col = "steelblue")
```



Hmm, inte riktigt linjärt ännu. Vi provar ett steg till ned på Tukey's stege, dvs att logaritmera x-variabeln `income` :

```
plotLowess(logcost ~ log(income), data = df, pch = 19, cex = 0.5, col = "steelblue")
```



Bingo! Rätt så linjärt! Vi kör på detta och lägger även in logaritmen av `income` i vår dataframe `df`.

```
df$logincome = log(df$income)
head(df)
```

```
cost rooms people income onlysecondary waterheat cookel poolfilt airrev
```

```
1  545     7      4  29900             0         0      1        0       1
2  389     7      2  11700             0         0      1        0       0
3  390     8      2  16900             0         0      0        0       0
4  268     7      2   9750             1         0      1        0       1
5  543     6      2  24700             1         0      0        0       1
6  278     6      3   8450             1         0      0        0       0
  aircond microwave dish dryer  logcost logincome
1       1         0    0     0 6.300786 10.305614
2       0         1    1     0 5.963579  9.367344
3       1         0    0     0 5.966147  9.735069
4       1         0    0     1 5.590987  9.185023
5       1         1    0     1 6.297109 10.114559
6       0         0    0     1 5.627621  9.041922
```

Uppgift 3.1

Använd R för att skatta modellen:

$$\texttt{logcost} = \beta_0 + \beta_1 \cdot \texttt{logincome} + \varepsilon, \qquad \varepsilon \overset{iid}{\sim} N(0, \sigma_\varepsilon)$$

och tolka skattningarna b_0 och b_1.

Uppgift 3.2

Använd formelsamlingen för att beräkna ett 99%-igt konfidensintervall för β_1. Kontrollera att ditt svar stämmer med R's direkta beräkning av detta konfidensintervall.

Uppgift 3.3

Använd formelsamlingen för att testa om `logincome` är en signifikant förklarande variabel på signifikansnivån 1%. Ställ upp noll- och alternativhypotes för testet och var noga med att dra en slutsats från testet.

Uppgift 3.4

Gör en prediktion med 95% prediktionsintervall för `logcost` vid `logincome=11` genom att använda R. [tips: argumentet `newdata` i `predict`-funktionen måste vara en dataframe, inte en vektor. Se min kod för lifespan data.]

Uppgift 3.5

Skatta nu en multipel linjär regression:

$$\texttt{logcost} = \beta_0 + \beta_1 \cdot \texttt{logincome} + \beta_2 \cdot \texttt{logrooms} + \beta_3 \cdot \texttt{logpeople} + \beta_4 \cdot \texttt{onlysecondary} + \beta_5 \cdot \texttt{poolfilt} + \beta_6 \cdot \texttt{aircond} + \varepsilon,$$

där `logrooms = log(rooms)` och `logpeople = log(people)` (lägg till dessa transformerade variabler i vår dataframe `df`). Tolka skattningarna av koefficienterna β_1 och β_6.

Uppgift 3.6

Vilka förklarande variabler är signifikanta på 5% nivån? På 1% signifikansnivå?

Uppgift 3.7

Gör en prediktion med 90% prediktionsintervall för `logcost` för ett hushåll med medianvärden på `logincome`, `logrooms`, `logpeople` och alla tre dummyvariabler satta till värdet noll.

Fördjupning/kuriosa

En av mina doktorander, Feng Li, har tillsammans med mig och professor Robert Kohn vid UNSW i Sydney analyserat det här datamaterialet i hans doktorsavhandling[1] vid statistiska institutionen vid SU. Feng utvecklade en flexibel regressionsmodell med fördelningar för felttermerna som bl a tillåts ha:

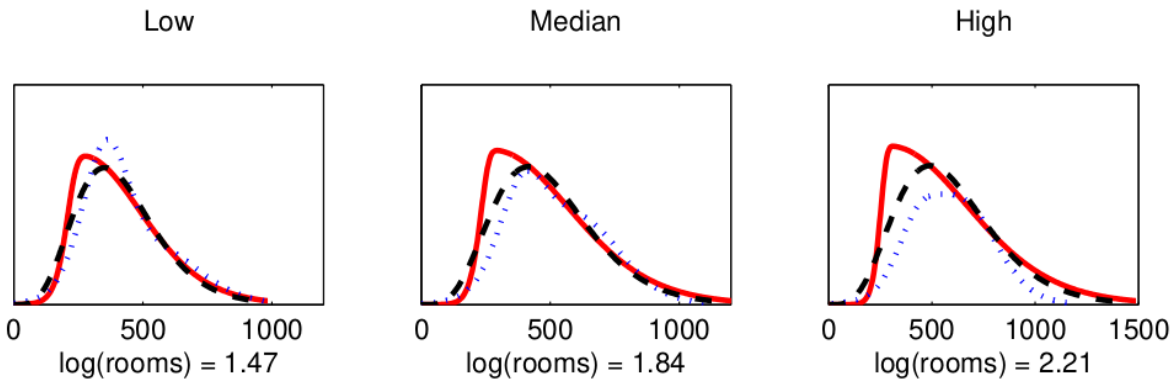
- heteroskedastisk varians (dvs olika varians för olika värden på t ex `logincome`)
- tunga svansar (likt t-fördelningen)
- skevhet

Hela artikeln är publicerad som ett kapitel in en bok[2], men finns även fritt tillgänglig som ett working paper.

Här är en bild från avhandlingen, där man ser den prediktiva sannolikhetsfördelningen för `cost` för olika värden på `logrooms`, för tre olika varianter av modellen (svarta och röda streckade linjer). Genom att utveckla en modell som kan modellera skevhet behövde vi inte transformera `cost` innan analysen, vilket blir trevligare att tolka.

[1] Li, F. (2013). *Bayesian Modeling of Conditional Densities*. Doktorsavhandling vid Statistiska institutionen, Stockholms universitet.

[2] Li, F., Villani, M. och Kohn, R. (2011). Modeling Conditional Densities Using Finite Smooth Mixtures, kapitel i boken *Mixtures: Estimation and Applications* (redaktörer:Mengersen, K., Robert, C. och Titterington, M.). Wiley.


Low
Median
High
0
500
1000
0
500
1000
0
500
1000
1500
log(rooms) = 1.47
log(rooms) = 1.84
log(rooms) = 2.21