LINKÖPINGS UNIVERSITET
Institutionen för datavetenskap
Avdelningen för statistik
Josef Wilzén & Måns Magnusson

2017-06-07
Programmering i R, 7.5 hp
732G33

Tentamen i Programmering i R, 7.5 hp

Skrivtid: 8.00-12.00
Hjälpmedel: Inget tryckt material, dock finns "R reference card v.2" och några andra referenskort tillgängliga elektroniskt.
Betygsgränser: Tentamen omfattar totalt 20 poäng. 12 poäng ger Godkänt, 16 poäng ger Väl godkänt.

Tänk på följande:
Skriv dina lösningar i **fullständig och läsbar kod**.
Lösningen skrivs i en körbar R-fil med namnet **tentaXX.R** där XX är ditt tenta-ID
Tex: tenta01.R om ditt tenta-ID är 01. Lämna *inte* in något Word-dokument!
Se filen **DocStudent.pdf** för hur tentan ska lämnas in.
Kommentera direkt i R-filen när något behöver förklaras eller diskuteras.
Eventuella grafer som skapas under tentans gång behöver **INTE** skickas in för rättning, det räcker med att **skicka in den kod som producerar figurerna**.

OBS: Glöm inte att spara din fil ofta! Om R krashar kan kod förloras.

1. Datastrukturer (4p)

   (a) Gör följande beräkning $(1 - cos\,(\pi \cdot 2.3))^{(1-\sqrt{7})}$ och avrunda till 4 decimaler. **1p**

   (b) Skapa en lista `my_list` med 3 element. Elementen ska innehålla de inbyggda dataseten `trees`, `AirPassengers` och `iris` (i given ordning). Elementnamnen ska vara samma som namnen på dataseten. Om du gjort rätt ska det se ut enligt nedan. **1p**

```
str(my_list)

List of 3
 $ trees        :'data.frame': 31 obs. of  3 variables:
  ..$ Girth : num [1:31] 8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...
  ..$ Height: num [1:31] 70 65 63 72 81 83 66 75 80 75 ...
  ..$ Volume: num [1:31] 10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...
 $ AirPassengers: Time-Series [1:144] from 1949 to 1961: 112 118 132 129 121 135 148 14
 $ iris         :'data.frame': 150 obs. of  5 variables:
  ..$ Sepal.Length: num [1:150] 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
  ..$ Sepal.Width : num [1:150] 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
  ..$ Petal.Length: num [1:150] 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
  ..$ Petal.Width : num [1:150] 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
  ..$ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ..
```

(c) Skapa en array `my_array` med dimensionerna 5,3,2. `my_array` ska vara fylld med slumptal från en normalfördelad variabel med medelvärde 14 och standardavvikelse 2. **1p**

```
dim(my_array)

[1] 5 3 2
```

(d) Utgå från det inbyggda datasetet `iris`. Skapa en ny data.frame kallad `my_frame`, som innehåller variablerna `Sepal.Length`, `Petal.Length` och `Species`. Men bara de rader som `Species` är lika med "versicolor" ska vara med. **1p**

2. Kontrollstrukturer (4p)

(a) Skapa en tom matris X av storlek 3×4 och fyll den sedan matrisen med elementen $x_{i,j} = (i - 2.5)^j$ där i är radnumret och j är kolumnnumret. Du ska använda en nästlad for-loop. **2p**

(b) Skapa en while-loop som genomför följande beräkning: Summerar arean av cirklar med radien $r = 1, 2, 3, 4, ...$ och skriver ut summan till konsolen. Loopen ska avbrytas om summan blir större än 1000 och ingen summa ska då skrivas ut. Arean av en cirkel ges av: $area = \pi \cdot r^2$. Om du gjort rätt ska resultatet se ut enligt nedan: **2p**

```
[1] 3.14159
[1] 15.708
[1] 43.9823
[1] 94.2478
[1] 172.788
[1] 285.885
[1] 439.823
[1] 640.885
[1] 895.354
```

3. Strängar och datum (4p)

(a) Läs in paketen `lubridate` och `stringr` i R. Läs is textfilen "wiki_robot.txt" till R och spara som en vektor som du kallar `robot`. **0.5p**

(b) Använd funktioner ur `stringr` för att välja ut de element ur `robot` som innehåller minst ett ord som är längre än är lika med 14 tecken. Ord kan bestå av tecknen A-Z, a-z och 0-9. **1.5p**

(c) Läs in datamaterialet "kaffe.csv". Data innehåller information om antal sålda koppar kaffe på ett kaffe för alla dagar mellan 1990-01-01 till 2009-12-31. Svara på frågorna nedan. **2p**

i. Vilken månad såldes det mest kaffe i genomsnitt?

ii. Den dagen som det såldes minst kaffe, vilken veckodag var det då?

iii. Hur många hela veckor är det mellan dagarna då det såldes mest respektive minst kaffe? (inkludera gränserna)

iv. Vad är standardavvikelsen för antalet koppar kaffe i perioden 1994-01-01 till 1995-12-31? (inkludera gränserna)

4. Funktioner: (**4p**)

(a) Ska du du skapa en funktion som kan applicera kvaderingsreglerna på en textsträng. Funktionen ska heta `quad_rule(x)`, och ha argumentet `x`, som är en textvektor. Kvaderingsreglerna är $(a+b)^2 = a^2 + 2ab + b^2$ och $(a-b)^2 = a^2 - 2ab + b^2$. Funktionen ska ta textsträngar på formen "`(a+b)^2`" och returnera uttrycket som ges efter att kvaderingsreglerna har använts. "`a`" och "`b`" ska kunna vara tecken (a-z) av längd 1-3 tecken eller numeriska tal bestående av 1-3 tecken. Om `a` och/eller `b` är numeriska så ska de utrycken där de ingår beräknas. Se exempeln nedan för hur funktionen ska fungera.

```
x1<-c("(a+b)^2","(d-t)^2","(2-b)^2","(e+921)^2")
x2<-c("(11+3)^2","(312-z)^2","(q+49)^2","(w+10)^2","(2+6)^2")
quad_func(x = x1)

[1] "(a+b)^2 = a^2+2*a*b+b^2"        "(d-t)^2 = d^2-2*d*t+t^2"
[3] "(2-b)^2 = 4-2*2*b+b^2"          "(e+921)^2 = e^2+2*e*921+848241"

test2<-quad_func(x = x2)
test2

[1] "(11+3)^2 = 121+66+9"           "(312-z)^2 = 97344-2*312*z+z^2"
[3] "(q+49)^2 = q^2+2*q*49+2401"    "(w+10)^2 = w^2+2*w*10+100"
[5] "(2+6)^2 = 4+24+36"
```

5. Linjär algebra och grafik (4p)

(a) Den linjära regressionsmodellen defineras som: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + + \beta_p x_p + \epsilon$ eller på matrisform $y = X\beta + \epsilon$. Läs in datamaterialet "fmri_data.csv" till R, kalla det `fmri`. Du ska beräkna två skattningar[1] på vektorn β som vi benämner med $\hat{\beta}_1$ och $\hat{\beta}_2$. De ska beräknas enligt följande formel:

$$\hat{\beta}_1 = \left(X^T X + \lambda_1 \mathbb{I}_P\right)^{-1} X^T y$$
$$\hat{\beta}_2 = \left(X^T X + \lambda_2 \mathbb{I}_P\right)^{-1} X^T y$$

Där X består av variablerna `X1` till `X10` i `fmri`. y är variabeln `y` i `fmri`. Låt P vara antalet kolumner i X. $\mathbb{I}_P$ är en diagnalmatris[2] av storleken P. Låt $\lambda_1 = 5$ och $\lambda_2 = 100$.

[1] Denna typ av skattning kallas Ridge regression.

[2] En kvadratisk matris som har nollar överallt utom på huvuddiagonalen, där den har ettor.

Beräkna $\hat{\beta}_1$ och $\hat{\beta}_2$ och kalla dem `beta_hat1` och `beta_hat2` respektive. Beräkna sedan anspassade värden ($\hat{y}$) enligt :

$$\hat{y}_1 = X\hat{\beta}_1$$
$$\hat{y}_2 = X\hat{\beta}_2$$

Kalla dem `y_hat1` och `y_hat2` respektive. **2p**

(b) Utgå från `fmri`. Lägg till variabeln `time<-1:72` till `fmri`. och gör nu följande plot i ggplot2. **2p**

i. Gör scatterplot mellan `y` och `time`. Texten på x-axeln ska vara "tid" och på y-axeln "BOLD fmri"

ii. Lägg till `y_hat1` som en röd linje.

iii. Lägg till `y_hat2` som en blå linje.

Kom ihåg: Lösningen skrivs i en körbar R-fil med namnet **tentaXX.R** där XX är ditt tenta-ID tex: tenta01.R om du har tenta-ID är 01. Lämna *inte* in något Word-dokument!

Lycka till!