Stat 511 Final Exam

May 7, 2008
Prof. Vardeman

Edited/Corrected Version

I have neither given nor received unauthorized assistance on this exam.

_____
Name

_____
Name Printed

1. Below is a small table of fake (x, y) data. Use them to compute smoothed/nonparametric regression fitted values for $x = 2.5$ making use of the two indicated smoothers.

y	1	2	3	1	0	1
x	0	1	2	3	4	5

a) Use a kernel smoother with kernel $K(u) = \exp(-u^2)$ and bandwidth $b = 2$. (You must plug in completely, but there is no need to do arithmetic.)

b) Use the loess smoother with neighborhood size $k = 4$ and the tri-cube weight function. (Be careful to convince me that your weighted regression is correct.)

2. Attached to this exam is an R printout of an analysis of part of the famous "German Credit" data set (due originally to Dr. Hans Hofmann of the University of Hamburg) consisting of information on $n = 234$ applications for new car loans. The response variable is

$$y = \begin{cases} 0 & \text{if the loan was denied} \\ 1 & \text{if the loan was granted} \end{cases}$$

and explanatory variables are

$$chk = \begin{cases} 0 & \text{if applicant's checking account balance is negative} \\ 1 & \text{if applicant's checking account balance is between 0 and 200 DM} \\ 2 & \text{if applicant's checking account balance is above 200 DM} \\ 3 & \text{if the applicant has no checking account} \end{cases}$$

$dur =$ duration of credit in months

$percent =$ installment payment as a % of disposable income

$$res = \begin{cases} 1 & \text{if the applicant has lived less than 1 year at current residence} \\ 2 & \text{if the applicant has lived between 1 and 2 years at current residence} \\ 3 & \text{if the applicant has lived between 2 and 3 years at current residence} \\ 4 & \text{if the applicant has lived more than 3 years at current residence} \end{cases}$$

$age =$ applicant age

$$install = \begin{cases} 0 & \text{if the applicant has no other installment plan credit} \\ 1 & \text{if the applicant has other installment plan credit} \end{cases}$$ and

$$foreign = \begin{cases} 0 & \text{if the applicant is a foreign worker} \\ 1 & \text{if the applicant is not a foreign worker} \end{cases}$$

The R analysis is based on a "probit" model for the responses y_i that treats them as independent Bernoulli(p_i) variables with $p_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$ for appropriate vectors of predictors $\mathbf{x}_i$ and $\boldsymbol{\beta}$ a vector of parameters. (Φ is the standard normal cdf.)

a) Which of the 7 factors above least clearly impacts the loan granting probability? Explain.

Which are apparently the least favorable and most favorable levels of the factors in terms of getting a loan approved (for quantitative variables just list "small" and "large"). Fill in the table below.

Factor	Least Favorable Level	Most Favorable Level
Checking Account		
Duration of the Loan		
Percent of Disposable Income		
Residence		
Age		
Other Installment Credit		
Foreign Worker Status		

b) Find an approximate 95% confidence interval for the probability that a person of checking account status 1, seeking a 36 month loan, with proposed payments 10% of disposable income, that has lived more than 3 years at a current residence, is 50 years old, has no other installment credit, and is a foreign worker is granted a new auto loan.

c) The quantity $\mathbf{x}'\boldsymbol{\beta}$ can be thought of as "the z-value corresponding to the probability a loan is granted under conditions $\mathbf{x}$." All other factors being equal, give a 95% confidence interval for the difference in such z-values for a 25 year-old applicant with a negative checking balance and a 60 year-old applicant who has no checking account. (You must plug in but need not simplify.)

3. The Cauchy distribution is one that is commonly used to make counter-examples in statistics. It is a bell-shaped distribution centered at the origin that has no moments (the integrals that would define mean and variance diverge). It has the peculiar property that the sample mean of n iid standard Cauchy variables has the same distribution as any one of the summands.

Suppose that in standard fashion, one assumes that $X_1, X_2, \ldots, X_n$ are iid with distribution F and wants to approximate the F probability that $\bar{X}_n < 0$ via a nonparametric bootstrap argument. So a given bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ will be processed to produce a sample mean $\bar{X}_n^*$ and the faction of B such values that are less than 0 will be used to estimate $P_F\left(\bar{X}_n < 0\right)$. Consider what will happen if in fact F is (unbeknownst to the hapless statistician) the standard Cauchy distribution. (If the statistician knew this he or she would know that $P_F\left(\bar{X}_n < 0\right) = P_F\left(X_1 < 0\right) = .5$.)

For $\bar{X}_n$ and S_n respectively the (real) sample mean and standard deviation, why is it plausible to expect the bootstrap estimate to be approximately $\Phi\left(\dfrac{-\bar{X}_n}{S_n/\sqrt{n}}\right)$?

Attached to this exam is another R printout showing simulated distributions for $\dfrac{-\bar{X}_n}{S_n/\sqrt{n}}$ for n of various sizes. Together with the above, what do these suggest about the likely effectiveness of the bootstrap in estimating $P_F\left(\bar{X}_n < 0\right) = .5$ for large n if in fact observations are Cauchy distributed?

4. At the end of this exam there is another R printout, this one concerning analysis of data on the age of onset of the mental illness schizophrenia. Estimate the bias in the difference of sample medians for males and females as an estimator of the difference in population medians. Then find a standard error for the difference in sample medians.

5. Consider a full rank linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{X}_j$ is the $n \times j$ sub-matrix of the $n \times k$ matrix $\mathbf{X}$ consisting of the first j columns of $\mathbf{X}$. With $\mathbf{x}_j$ the jth column of $\mathbf{X}$, let $\mathbf{x}_j^* = \left(\mathbf{I} - \mathbf{P}_{\mathbf{X}_{j-1}}\right)\mathbf{x}_j$. For simplicity in what follows, you may restrict your attention to the $k = 3$ case.

Define $\mathbf{X}^* = \left(\mathbf{x}_1 \,|\, \mathbf{x}_2^* \,|\, \mathbf{x}_3^*\right)$. Argue carefully that $\mathbf{X}$ and $\mathbf{X}^*$ have the same column space. Unless $\mathbf{X} = \mathbf{X}^*$, what computational and other reasons might there be for preferring a model specified in terms of $\mathbf{X}^*$ to one specified in terms of $\mathbf{X}$?

6. An approximate "power law" relationship between x and y of the form

$$y \approx ax^b$$

could potentially be handled statistically by taking logs and then using ordinary linear model theory with the relationship

$$\ln y_i = \ln a + b \ln x_i + \varepsilon_i \qquad (*)$$

(i.e. by doing simple linear regression of $y_i' = \ln y_i$ on $x_i' = \ln x_i$). The usual $\varepsilon_i \sim \text{iid N}\left(0, \sigma^2\right)$ assumption in (*) leads to a model for the y_is that says they are independent with both means and standard deviations proportional to x_i^b. But of course, the distributions of y_i at various x_i are not normal, nor are they symmetric about their means.

a) Motivated by these considerations, one might consider the model

$$y_i = Cx_i^b + \delta_i \qquad (**)$$

where the δ_i are independent $\text{N}\left(0, \left(x_i^b \sigma\right)^2\right)$ variables, and $b, C,$ and σ^2 are model parameters. Is this an instance of the "nonlinear (regression) model" (either exactly as presented in class or in an "Aitken" form)? Explain.

b) For fixed b, what C and σ^2 maximize the loglikelihood for model (**),

$$\ell\left(b, C, \sigma^2\right) = -\frac{n}{2}\ln\left(2\pi\right) - n\ln\left(\sigma\right) - b\sum_{i=1}^{n}\ln\left(x_i\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\frac{y_i}{x_i^b} - C\right)^2$$

7. A classical hierarchical data set originally appearing in Snedecor and Cochran concerns measured glycogen content of rat livers. Rats were each given one of 3 different treatments, 2 rats per treatment. Rats were sacrificed and livers cut into 3 pieces each. Then 2 analyses were made for glycogen content on each piece (for $3\times 2\times 3\times 2 = 36$ measured contents). With

$$y_{ijkl} = \text{measurment } l \text{ from piece } k \text{ from rat } j \text{ given treatment } i$$

the final `R` printout attached to this exam is for an analysis of the data based on a model

$$y_{ijkl} = \mu + \tau_i + \rho_{ij} + \theta_{ijk} + \varepsilon_{ijkl}$$

where $\mu, \tau_1, \tau_2,$ and τ_3 are fixed effects, the $\rho_{ij} \sim$ iid $\mathrm{N}\left(0, \sigma_\rho^2\right)$ are random rat effects, independent of random piece effects $\theta_{ijk} \sim$ iid $\mathrm{N}\left(0, \sigma_\theta^2\right)$, independent of random analysis errors $\varepsilon_{ijkl} \sim$ iid $\mathrm{N}\left(0, \sigma^2\right)$. Use it to answer the following.

a) Are clear differences indicated between the treatment effects $\tau_1, \tau_2,$ and τ_3? Explain.

b) Is it clear whether the largest part of the response variability for a given treatment comes from rat-to-rat differences, piece-to-piece differences, or from analysis-to-analysis differences? Explain.

c) What is a point estimate of $\mathrm{Var}\left(\bar{y}_{1\cdots} - \bar{y}_{2\cdots}\right)$ (the variance of the difference in treatment 1 and 2 sample means)? (Give a number.)

Printout for Problem 2

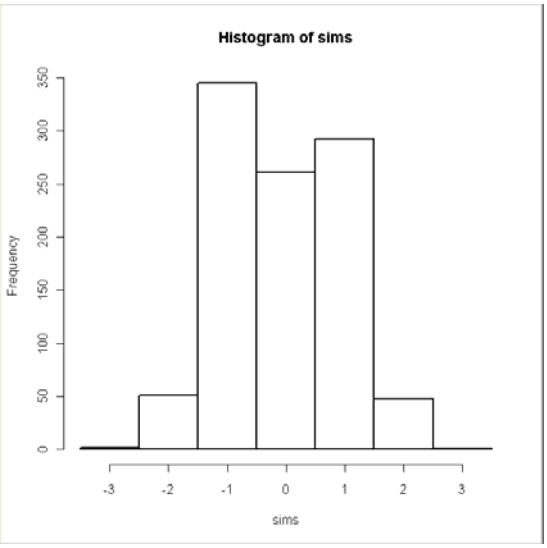
```
> creditdata[1:10,]
   CHK DUR PERCENT RES AGE INSTALL FOREIGN Y
1    0  24       3   4  53       0       0 0
2    1  30       4   2  28       0       0 0
3    1  12       3   1  25       0       0 0
4    0  24       4   4  60       0       0 0
5    0  15       2   4  28       0       0 1
6    3   9       4   4  48       0       0 1
7    0  10       1   3  48       0       1 1
8    1  18       2   2  30       0       0 1
9    3  11       4   4  35       0       0 1
10   3  11       1   4  39       0       0 1

> Ystar<-cbind(Y,1-Y)

> creditscore<-glm(Ystar~CHK+DUR+PERCENT+RES+AGE+INSTALL+FOREIGN,binomial(link=probit))

> summary(creditscore)

Call:
glm(formula = Ystar ~ CHK + DUR + PERCENT + RES + AGE + INSTALL +
    FOREIGN, family = binomial(link = probit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8357  -0.8524   0.3556   0.8329   2.1534

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.196396   0.514942  -0.381   0.7029
CHK1         0.358001   0.239017   1.498   0.1342
CHK2         0.945421   0.438231   2.157   0.0310 *
CHK3         1.355881   0.251993   5.381 7.42e-08 ***
DUR         -0.014610   0.008430  -1.733   0.0831 .
PERCENT     -0.165267   0.083718  -1.974   0.0484 *
RES2        -0.253368   0.334666  -0.757   0.4490
RES3         0.359118   0.374558   0.959   0.3377
RES4        -0.119002   0.337890  -0.352   0.7247
AGE          0.024283   0.009577   2.535   0.0112 *
INSTALL1    -1.065777   0.263834  -4.040 5.35e-05 ***
FOREIGN1     1.175505   0.499522   2.353   0.0186 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 310.86  on 233  degrees of freedom
Residual deviance: 229.51  on 222  degrees of freedom
AIC: 253.51

Number of Fisher Scoring iterations: 6
```

```
> C<-vcov(creditscore)
> C
             (Intercept)          CHK1          CHK2          CHK3           DUR
(Intercept)  0.265164861 -0.0317084378 -4.028574e-02 -1.750415e-02 -1.040997e-03
CHK1        -0.031708438  0.0571290056  2.676758e-02  2.661143e-02 -3.022826e-04
CHK2        -0.040285744  0.0267675793  1.920460e-01  2.706532e-02  2.024910e-04
CHK3        -0.017504148  0.0266114288  2.706532e-02  6.350062e-02 -2.695720e-04
DUR         -0.001040997 -0.0003022826  2.024910e-04 -2.695720e-04  7.106650e-05
PERCENT     -0.017976287  0.0008686049 -2.570812e-04 -4.982245e-04 -8.624759e-05
```

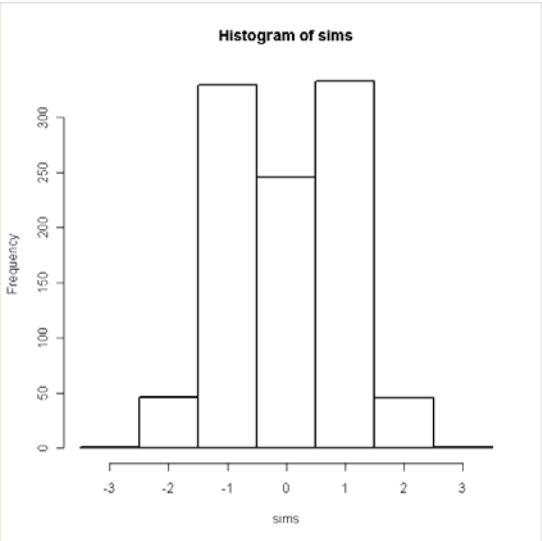
```
RES2        -0.081544351  0.0031955597  1.083984e-02 -1.898002e-03 -2.643520e-05
RES3        -0.081169344  0.0071020930  2.284705e-02  6.457004e-03  1.996011e-04
RES4        -0.056110335  0.0056257124  1.739269e-02 -1.700728e-03 -1.345407e-04
AGE         -0.002903189  0.0001392074  1.025394e-05 -7.252210e-06  2.475103e-06
INSTALL1     0.002166719 -0.0050637968 -2.445417e-02 -8.761880e-03 -1.089239e-04
FOREIGN1    -0.026666204  0.0191270672  1.582144e-02  1.344582e-02  4.449407e-04
                  PERCENT          RES2          RES3          RES4           AGE
(Intercept) -1.797629e-02 -8.154435e-02 -0.0811693435 -0.0561103349 -2.903189e-03
CHK1         8.686049e-04  3.195560e-03  0.0071020930  0.0056257124  1.392074e-04
CHK2        -2.570812e-04  1.083984e-02  0.0228470463  0.0173926886  1.025394e-05
CHK3        -4.982245e-04 -1.898002e-03  0.0064570038 -0.0017007278 -7.252210e-06
DUR         -8.624759e-05 -2.643520e-05  0.0001996011 -0.0001345407  2.475103e-06
PERCENT      7.008731e-03 -2.188098e-04 -0.0014550684 -0.0015680297 -2.002773e-05
RES2        -2.188098e-04  1.120016e-01  0.0837794877  0.0845269907 -3.139996e-05
RES3        -1.455068e-03  8.377949e-02  0.1402934549  0.0863189032 -2.171223e-04
RES4        -1.568030e-03  8.452699e-02  0.0863189032  0.1141696980 -7.273258e-04
AGE         -2.002773e-05 -3.139996e-05 -0.0002171223 -0.0007273258  9.172652e-05
INSTALL1     4.079270e-03 -5.448572e-03 -0.0069044185 -0.0018871858 -4.209554e-04
FOREIGN1    -7.515452e-04 -6.780356e-03  0.0063741934 -0.0011883942  1.183635e-04
                 INSTALL1      FOREIGN1
(Intercept)  0.0021667186 -0.0266662043
CHK1        -0.0050637968  0.0191270672
CHK2        -0.0244541666  0.0158214358
CHK3        -0.0087618800  0.0134458157
DUR         -0.0001089239  0.0004449407
PERCENT      0.0040792703 -0.0007515452
RES2        -0.0054485717 -0.0067803560
RES3        -0.0069044185  0.0063741934
RES4        -0.0018871858 -0.0011883942
AGE         -0.0004209554  0.0001183635
INSTALL1     0.0696082845 -0.0144602283
FOREIGN1    -0.0144602283  0.2495222762

> t(c(1,1,0,0,36,10,0,0,1,50,0,0))%*%coef(creditscore)
          [,1]
[1,] -0.921903

> t(c(1,1,0,0,36,10,0,0,1,50,0,0))%*%C%*%c(1,1,0,0,36,10,0,0,1,50,0,0)
          [,1]
[1,] 0.3921053
```

Printout for Problem 3

```
> value<-function(x) {(n^.5)*(-mean(x)/sd(x))}

> i<-1
> n=100
> while (i<1001) {
+ x<-c(rcauchy(n))
+ sims[i]<-value(x)
+ i<-i+1
+ }
> hist(sims,breaks=c(-3.5,-2.5,-1.5,-.5,.5,1.5,2.5,3.5))
```

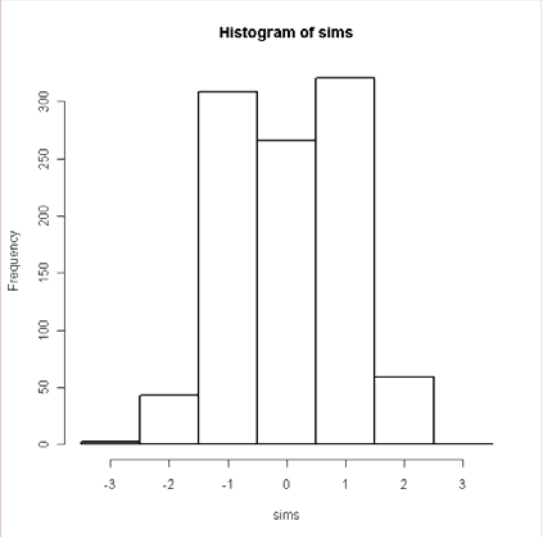



```
> i<-1
> n<-1000
> while (i<1001) {
+ x<-c(rcauchy(n))
+ sims[i]<-value(x)
+ i<-i+1
+ }
> hist(sims,breaks=c(-3.5,-2.5,-1.5,-.5,.5,1.5,2.5,3.5))
```

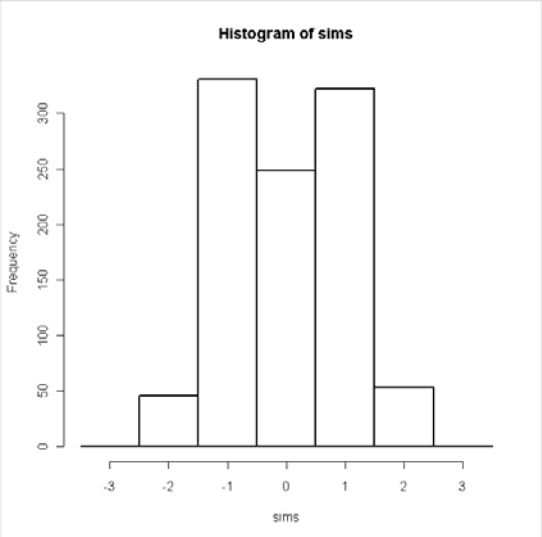



```
> i<-1
> n<-10000
> while (i<1001) {
```

```
+ x<-c(rcauchy(n))
+ sims[i]<-value(x)
+ i<-i+1
+ }
> hist(sims,breaks=c(-3.5,-2.5,-1.5,-.5,.5,1.5,2.5,3.5))
```



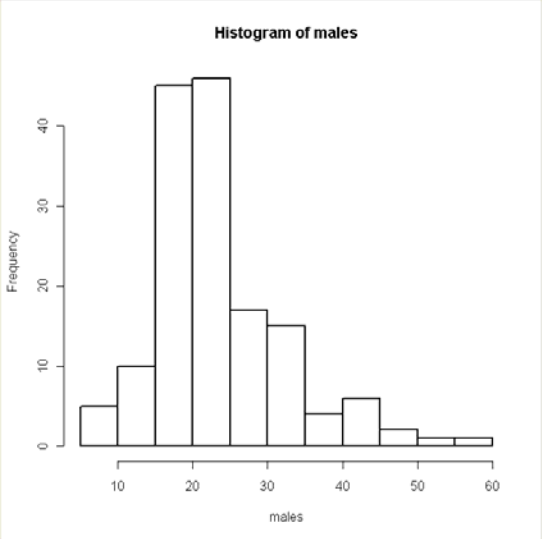

```
> i<-1
> n<-100000
> while (i<1001) {
+ x<-c(rcauchy(n))
+ sims[i]<-value(x)
+ i<-i+1
+ }
> hist(sims,breaks=c(-3.5,-2.5,-1.5,-.5,.5,1.5,2.5,3.5))
```



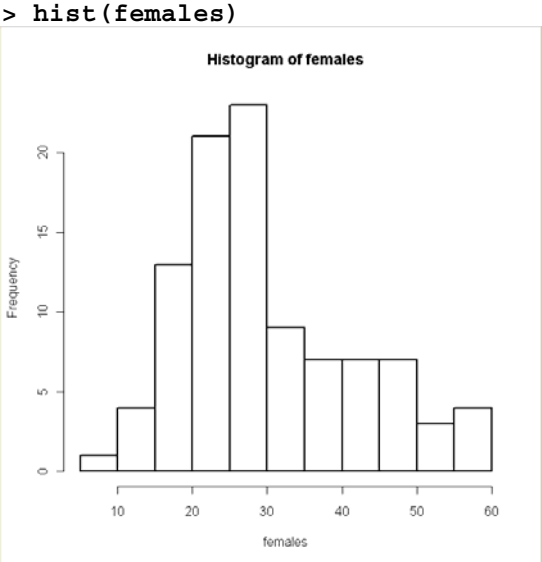
Printout for Problem 4

```
> males
  [1] 21 18 23 21 27 24 20 12 15 19 21 22 19 24  9 19 18 17 23 17 23
 [22] 19 37 26 22 24 19 22 19 16 16 18 16 33 22 23 10 14 15 20 11 25
 [43]  9 22 25 20 19 22 23 24 29 24 22 26 20 25 17 25 28 22 22 23 35
```

```
 [64] 16 29 33 15 29 20 29 24 39 10 20 23 15 18 20 21 30 21 18 19 15
 [85] 19 18 25 17 15 42 27 18 43 20 17 21  5 27 25 18 24 33 32 29 34
[106] 20 21 31 22 15 27 26 23 47 17 21 16 21 19 31 34 23 23 20 21 18
[127] 26 30 17 21 19 22 52 19 24 19 19 33 32 29 58 39 42 32 32 46 38
[148] 44 35 45 41 31

> summary(males)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   19.00   22.00   23.91   27.25   58.00
> hist(males)
```




```
> females
 [1] 20 30 21 23 30 25 13 19 16 25 20 25 27 43  6 21 15 26 23 21 23 23
[23] 34 14 17 18 21 16 35 32 48 53 51 48 29 25 44 23 36 58 28 51 40 43
[45] 21 48 17 23 28 44 28 21 31 22 56 60 15 21 30 26 28 23 21 20 43 39
[67] 40 26 50 17 17 23 44 30 35 20 41 18 39 27 28 30 34 33 30 29 46 36
[89] 58 28 30 28 37 31 29 32 48 49 30

> summary(females)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00   21.00   28.00   30.47   38.00   60.00
> hist(females)
```




```
> B<-10000

> males.boot<-bootstrap(males,B,median)
> summary(males.boot$thetastar)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   22.00   22.00   21.95   22.00   25.00
```

```
> sd(males.boot$thetastar)
[1] 0.6102509

> females.boot<-bootstrap(females,B,median)
> summary(females.boot$thetastar)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  23.00   28.00   28.00   28.37   29.00   34.00
> sd(females.boot$thetastar)
[1] 1.123143
```

Printout for Problem 7

```
> ratsdata
     y Treat Rat Piece
1  131     1   1     1
2  130     1   1     1
3  131     1   1     2
4  125     1   1     2
5  136     1   1     3
6  142     1   1     3
7  150     1   2     4
8  148     1   2     4
9  140     1   2     5
10 143     1   2     5
11 160     1   2     6
12 150     1   2     6
13 157     2   3     7
14 145     2   3     7
15 154     2   3     8
16 142     2   3     8
17 147     2   3     9
18 153     2   3     9
19 151     2   4    10
20 155     2   4    10
21 147     2   4    11
22 147     2   4    11
23 162     2   4    12
24 152     2   4    12
25 134     3   5    13
26 125     3   5    13
27 138     3   5    14
28 138     3   5    14
29 135     3   5    15
30 136     3   5    15
31 138     3   6    16
32 140     3   6    16
33 139     3   6    17
34 138     3   6    17
35 134     3   6    18
36 127     3   6    18


> summary(ratsdata)
       y         Treat       Rat          Piece
 Min.   :125.0   1:12   Min.   :1.0   Min.   : 1.0
 1st Qu.:135.8   2:12   1st Qu.:2.0   1st Qu.: 5.0
 Median :141.0   3:12   Median :3.5   Median : 9.5
 Mean   :142.2          Mean   :3.5   Mean   : 9.5
 3rd Qu.:150.0          3rd Qu.:5.0   3rd Qu.:14.0
 Max.   :162.0          Max.   :6.0   Max.   :18.0

> rats<-lmer(y~Treat+(1|Rat)+(1|Piece))

> summary(rats)

Linear mixed-effects model fit by REML
Formula: y ~ Treat + (1 | Rat) + (1 | Piece)
   AIC   BIC logLik MLdeviance REMLdeviance
```

```
 229.6 237.5 -109.8      234.3        219.6
Random effects:
 Groups   Name        Variance Std.Dev.
 Piece    (Intercept) 14.162   3.7632
 Rat      (Intercept) 36.084   6.0070
 Residual             21.168   4.6008
number of obs: 36, groups: Piece, 18; Rat, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  140.500      4.708  29.842
Treat2        10.500      6.658   1.577
Treat3        -5.333      6.658  -0.801

Correlation of Fixed Effects:
       (Intr) Treat2
Treat2 -0.707
Treat3 -0.707  0.500

> vcov(rats)

3 x 3 Matrix of class "dpoMatrix"
            (Intercept)    Treat2    Treat3
(Intercept)    22.16642 -22.16642 -22.16642
Treat2        -22.16642  44.33283  22.16642
Treat3        -22.16642  22.16642  44.33283


> fitted(rats)

 [1] 131.9920 131.9920 130.5613 130.5613 136.8565 136.8565 148.1495 148.1495
 [9] 143.8573 143.8573 151.5833 151.5833 150.5358 150.5358 148.8190 148.8190
[17] 149.9635 149.9635 152.6088 152.6088 149.1750 149.1750 154.8979 154.8979
[25] 131.6336 131.6336 136.4981 136.4981 135.0673 135.0673 137.6506 137.6506
[33] 137.3644 137.3644 132.7861 132.7861

> ranef(rats)

An object of class “ranef.lmer”
[[1]]
   (Intercept)
1  -1.99642195
2  -3.42715046
3   2.86805498
4   1.13798485
5  -3.15420069
6   4.57173327
7   0.62108201
8  -1.09579220
9   0.04879061
10  0.52350079
11 -2.91024763
12  2.81266641
13 -2.85480836
14  2.00966857
15  0.57894006
16  1.80560746
17  1.51946175
18 -3.05886948

[[2]]
  (Intercept)
1  -6.5115290
2   6.5115290
3  -1.0852548
4   1.0852548
5  -0.6782843
6   0.6782843
```

```
> sim<-mcmcsamp(rats ,50000)

> HPDinterval(sim)
                     lower      upper
(Intercept)    128.141693 152.622306
Treat2          -6.765762  27.448308
Treat3         -22.137104  12.268336
log(sigma^2)     2.704367   4.276008
log(Piec.(In)) -56.427038   4.897371
log(Rat.(In))  -55.426521   7.836563
attr(,"Probability")
[1] 0.95
```