LISBOA SCHOOL OF ECONOMICS & MANAGEMENT

MASTER IN ACTUARIAL SCIENCE

Risk Models

01/02/2019

1st part of the exam

Time allowed: 2 hours

Instructions:

1. This paper contains 7 questions and comprises 4 pages including the title page.
2. Enter all requested details on the cover sheet.
3. You have 10 minutes of reading time. You must not start writing your answers until instructed to do so.
4. Number the pages of the paper where you are going to write your answers.
5. Attempt all 7 questions.
6. Begin your answer to each of the 7 questions on a new page.
7. Marks are shown in brackets. Total marks: **140**.
8. Show calculations where appropriate.
9. An approved calculator may be used.
10. The distributed formulary and the Formulae and Tables for Actuarial Examinations (the 2002 edition) may be used. Note that the parametrization used for the different distributions is that of the distributed formulary.

1. An experiment is conducted to compare the starting salaries of male and female college graduates who find jobs. Pairs are formed by choosing a male and a female with the same major and similar grade point averages. Suppose a random sample of 5 pairs is formed in this manner and the starting annual salary of each person is recorded. Assume that the salaries are normally distributed for both males and females. The results are

Female	28.8	41.6	39.8	38.5	42.6
Male	29.3	41.5	40.4	38.5	43.5

a. **[5]** Compute a 90% confidence interval for the expect value of a male salary.

b. **[10]** Test if the expected value of starting salary for males and females are equal against the alternative that it is higher for males. Use $\alpha = 0.05$ and comment.

2. The following random sample of 20000 policyholders has been observed where each observation represents the number of claims per year.

Nº Claims	0	1	2	3	4	5
Nº Policies	17220	1920	793	50	15	2

a. **[10]** Obtain a non-parametric estimator for the probability that a policy originates 3 claims. Is this estimator unbiased? Justify.

b. **[10]** Compute a 95% confidence interval for the probability that a policy originates 3 claims.

c. **[15]** Now, assume that the number of claims per year follows a Poisson distribution with mean θ. Obtain a maximum likelihood estimate for θ and also obtain a ML estimate for the probability that a policy originates 3 claims.

3. From a continuous population with density function f, you are given the following sorted random sample (1; 1; 2; 2; 2; 4; 4; 4; 6; 6).

a. **[15]** Estimate $f(5)$ using a Pareto kernel function with parameters $\alpha = 3$ and $\theta = 2y$, i.e. $k_y(x) = 24y^3(x+2y)^{-4}$, $x > 0$.

b. **[5]** Explain why, once $\alpha = 3$ was chosen, the value $2y$ was obtained for the parameter θ

4. You are given a sorted random sample with $n=120$ observations. You know that the first $n_1=44$ of these observations are **less than or equal** to 10 (and you have no access to the observed values) and you also know that for the remaining $n_2=76$ observations you have $\sum_{i=n_1+1}^{n}(1/x_i)^3=0.0296$. Assume that losses follow an inverse Weibull distribution with $\tau=3$, i.e. with **distribution function** $F(x\mid\theta)=e^{-(\theta/x)^3}$, $x>0$, $\theta>0$. Let $p=P(X\le 10)$.

   a. **[10]** Show that the maximum likelihood estimate for θ is 10.1075 and show that the asymptotic variance of the corresponding estimator is 0.1494 using the 2nd derivative of the log-likelihood function.

   b. **[10]** Obtain a maximum likelihood estimate for p and, using the delta method, calculate the standard error of $\hat{p}$. Obtain a 95% asymptotic confidence interval for p.

5. **[10]** Let $f(x\mid\theta)$, $x>0$ be the density function (and $F(x\mid\theta)$ the distribution function) of the claim amounts for a given risk. Assume that for each policy an ordinary deductible (the payments are made in excess of the deductible) can be in force and that each policy can have a different capital amount insured.. You are given the following information about a sample of five payments

Deductible value	0	10	25	0	0
Policy limit	1000	500	100	1000	1000
Payment	850	15	100	1000	500

   Write the log likelihood function that should be maximized to obtain a maximum likelihood estimate of θ as a function of F and f. No additional computations are required.

6. Let $X\mid\theta \sim Po(2\theta)$ and assume that, using a Bayesian approach, we define our prior for θ as being an exponential distribution with mean 1/4. From population X we observe a sample with 20 observations where $t=\sum_{i=1}^{20}x_i=49$.

   a. **[15]** Show that the posterior for θ is a gamma distribution with parameters 50 and $1/44$ and obtain a Bayes estimate for θ against a 0-1 loss function.

   b. **[10]** Obtain a 95% HPD interval for θ using an exact distribution. If you are unable to do so, use Bayesian Central limit Theorem.

7. **[15]** An actuary wants to use simulation to compare two estimators for the expected value of a Pareto distribution when it is known that α should be close to the value 2 (although greater than 2). The first estimator, T, is the sample mean while the second estimator, V, is the sample trimmed mean, i.e., the sample average after removing the largest and the smallest observations.

His idea is to assume that X is Pareto distributed with parameters $\alpha = 2.1$ and $\theta = 110$ and to use a sample with size $n = 10$. The performance of the estimators will be compared using the mean square errors.

Explain how to use simulation to perform this simulation.

Solutions

Exercise 1

a)

Let X be the salaries for males.

Pivotal quantity: $T = \dfrac{\bar{X} - \mu}{S/\sqrt{5}} \sim t_{(4)}$

The 90% CI is given by $\bar{x} \pm t_{0.05}\left(s/\sqrt{5}\right)$, i.e. $38.64 \pm 2.132\left(5.5261/\sqrt{5}\right)$ → (33.371, 43.909)

b)

Paired sample test. Let X be the salaries for males and Y be the salaries for females. Define $D = X - Y$

$H_0 : \mu_D = 0 \qquad H_1 : \mu_D > 0$

Test Statistic: $\dfrac{\bar{D}}{S_D/\sqrt{5}} \sim t_{(4)}$

The observed values of D are: 0.5, -0.1, 0.6, 0, 0.9 and then $\bar{d} = 0.38$, $s_D = 0.4207$

$T_{obs} = \dfrac{0.38}{0.4207/\sqrt{5}} = 2.0197$ p-value=0.05677 or $t_{0.05} = 2.132$

At 5% we do not reject the null, i.e. statistical evidence is not enough to conclude that starting salaries for males are higher than that for females. However, the p-value close to the rejection border underlines that this conclusion should be taken carefully.

Exercise 2

a)

Let $p = P(X = 3)$. The non-parametric estimator of p is $\tilde{p} = \dfrac{N_3}{n} = \dfrac{N_3}{20000}$ where n is the sample size and $N_3 = \#\{X_i = 3, i = 1, 2, \cdots, n\}$.

As $N_3 \sim b(n, p)$ we have $E(N_3) = n\,p$ and then $E(\tilde{p}) = p$. The estimator is unbiased.

b)

As $\operatorname{var}(N_3) = n\,p\,(1-p)$ we get $\operatorname{var}(\tilde{p}) = \dfrac{p\,(1-p)}{n} = \dfrac{p\,(1-p)}{20000}$. Applying the CLT we obtain the CI $\tilde{p} \pm z_{0.025}\sqrt{\dfrac{\tilde{p}(1-\tilde{p})}{20000}}$ where $\tilde{p} = \dfrac{50}{20000} = 0.0025$ and $z_{0.025} = 1.96$. The interval is then (0.0018, 0.0032).

c)

$$\ell(\theta) = \ln L(\theta) = \sum_{i=1}^{20000} \ln\left(\frac{e^{-\theta}\theta^{x_i}}{x_i\,!}\right) = \sum_{i=1}^{20000}\left(-\theta + x_i \ln\theta - \ln(x_i\,!)\right)$$

$$\ell'(\theta) = \sum_{i=1}^{20000}\left(-1 + \frac{x_i}{\theta}\right) = -20000 + \frac{\sum_{i=1}^{20000} x_i}{\theta} = -20000 + \frac{1920 + 2\times 793 + 3\times 50 + 4\times 15 + 5\times 2}{\theta}$$

$$= -20000 + \frac{1920 + 15886 + 150 + 60 + 10}{\theta} = -20000 + \frac{3726}{\theta}$$

$$\ell'(\theta)=0 \Leftrightarrow 20000=\frac{3726}{\theta} \Leftrightarrow \theta=\frac{3726}{20000} \Leftrightarrow \theta=0.1863$$

As $\ell''(\theta)=-\frac{3726}{\theta^2}<0$, the ML estimate for θ is $\hat{\theta}=0.1863$.

Then the ML estimate for $p=P(X=3)=\frac{e^{-\theta}\,\theta^3}{3!}$ is $\hat{p}=\frac{e^{-0.1863}\,0.1863^3}{6}=0.0008945$

Exercise 3

a)

Sample (1; 1; 2; 2; 2; 4; 4; 4; 6; 6) →

j	1	2	3	4
y_j	1	2	4	6
$p(y_j)$	0.2	0.3	0.3	0.2

$$\begin{aligned}\hat{f}(5)&=\sum_{j=1}^{4} p(y_j)\,k_{y_j}(5)=0.2\times k_1(5)+0.3\times k_2(5)+0.3\times k_4(5)+0.2\times k_6(5)\\&=0.2\times\left(24\times1^3\times7^{-4}\right)+0.3\times\left(24\times2^3\times9^{-4}\right)+0.3\times\left(24\times4^3\times13^{-4}\right)+0.2\times\left(24\times6^3\times17^{-4}\right)\\&=0.2\times0.009996+0.3\times0.029264+0.3\times0.053780+0.2\times0.062068\\&=0.039326\end{aligned}$$

b)

As it is well known the expected value of a Pareto random variable (when $\alpha>1$ as it is the case) is given by $\theta/(\alpha-1)$. As $\alpha=3$, the expected value is $\theta/2$. In order to guarantee that the expected value at each point y_j is equal to y_j (so that the kernel preserves the expected value) we equate y_j to $\theta/2$ and then we use $\theta=2y_j$

Exercise 4

a)

$F(x\,|\,\theta)=e^{-(\theta/x)^3}$, $x>0$, $\theta>0$.

$f(x\,|\,\theta)=3\theta^3x^{-4}\,e^{-(\theta/x)^3}$

$$\begin{aligned}\ell(\theta)&=\sum_{i=1}^{n_1}\ln F(10\,|\,\theta)+\sum_{i=n_1+1}^{n}\ln f(x_i\,|\,\theta)\\&=-n_1\,\theta^3\,10^{-3}+\sum_{i=n_1+1}^{n}\left(\ln 3+3\ln\theta-4\ln x_i-\theta^3\,x_i^{-3}\right)\end{aligned}$$

$$\begin{aligned}\ell'(\theta)&=-3n_1\,10^{-3}\theta^2+\sum_{i=n_1+1}^{n}\left(3\theta^{-1}-3\theta^2\,x_i^{-3}\right)=-3n_1\,10^{-3}\theta^2+3n_2\,\theta^{-1}-3\theta^2\sum_{i=n_1+1}^{n}x_i^{-3}\\&=-3\theta^2\left(n_1\,10^{-3}+\sum_{i=n_1+1}^{n}x_i^{-3}\right)+3n_2\,\theta^{-1}\end{aligned}$$

$$\ell'(\theta)=0\Leftrightarrow-3\theta^2\left(n_1\,10^{-3}+\sum_{i=n_1+1}^{n}x_i^{-3}\right)+3n_2\,\theta^{-1}=0\Leftrightarrow\theta^2\left(n_1\,10^{-3}+\sum_{i=n_1+1}^{n}x_i^{-3}\right)=n_2\,\theta^{-1}$$

$$\Leftrightarrow\theta^3=\frac{n_2}{\left(n_1\,10^{-3}+\sum_{i=n_1+1}^{n}x_i^{-3}\right)}\Leftrightarrow\theta=\left(\frac{n_2}{\left(n_1\,10^{-3}+\sum_{i=n_1+1}^{n}x_i^{-3}\right)}\right)^{1/3}=10.1075$$

As $\ell''(\theta) = -6\,\theta\left(n_1\,10^{-3} + \sum_{i=n_1+1}^{n} x_i^{-3}\right) - 3n_2\,\theta^{-2} < 0$, the ML estimate for θ is $\hat{\theta} = 10.1075$.

$\hat{\text{var}}(\hat{\theta}) \approx -\dfrac{1}{\ell''(\hat{\theta})} = \dfrac{1}{6.6952} = 0.1494$ as

$\ell''(\hat{\theta}) = -6\,\hat{\theta}\left(n_1\,10^{-3} + \sum_{i=n_1+1}^{n} x_i^{-3}\right) - 3n_2\,\hat{\theta}^{-2} = -6.6952$

b)

$p = P(X \le 10 \mid \theta) = e^{-(\theta/10)^3}$ The ML estimate is $\hat{p} = e^{-(10.1075/10)^3} = 0.3561$

Delta method

$g(\theta) = e^{-(\theta/10)^3}$; $g'(\theta) = -3\times 10^{-3} \times \theta^2 \times e^{-(\theta/10)^3} = -0.003\,\theta^2\, e^{-(\theta/10)^3}$; $g'(\hat{\theta}) = -0.1091328$

$\hat{\text{var}}(\hat{p}) \approx \left(0.003\ \hat{\theta}^2\, e^{-(\hat{\theta}/10)^3}\right)^2\ \hat{\text{var}}(\hat{\theta}) = 0.0119 \times 0.1494 = 0.00178$

The asymptotic 95% CI is given by $0.3561 \pm 1.96\sqrt{0.00178}$ →(0.2734, 0.4387)

Exercise 5

The individual contributions to the log likelihood vary according to the existence of a deductible (truncation) and of a policy limit (censoring value for the claim amounts). In any case, when a deductible is in force, the observed claim amounts are given by the sum of the deductible and the payment.

$$\ell(\theta) = \ln f(850 \mid \theta) + \ln\left(\frac{f(25 \mid \theta)}{1 - F(10 \mid \theta)}\right) + \ln\left(\frac{1 - F(125 \mid \theta)}{1 - F(25 \mid \theta)}\right) + \ln\left(1 - F(1000 \mid \theta)\right) + \ln f(500 \mid \theta)$$

Exercise 6

a) $L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = \prod_{i=1}^{n} \dfrac{e^{-2\theta}\,(2\theta)^{x_i}}{x_i\,!} = \dfrac{e^{-2n\theta}\,2^t\,\theta^t}{\prod_{i=1}^{n} x_i\,!} \propto \theta^t\, e^{-2n\theta}$

$\pi(\theta) = 4\,e^{-\theta/(1/4)} \propto e^{-4\theta}$

And then

$\pi_{\underline{x}}(\theta) \propto e^{-4\theta}\theta^t\, e^{-2n\theta} = \theta^t\, e^{-(4+2n)\theta}$ which is the core of a gamma distribution with parameters $t+1$ and $1/(4+2n)$, i.e. 50 and 1/44.

Bayes estimate against 0-1 loss function → Mode of the posterior

The mode of the gamma distribution is $\theta^B = (1/44)\times(50-1) = 1.1136$

b) The approximate 95% HPD interval (q_1, q_2) for θ is such that $P(\theta < q_1 \mid \underline{x}) = 0.025$ and $P(\theta > q_2 \mid \underline{x}) = 0.025$.

We know that if $\theta \sim \gamma(\alpha,\beta)$ then $2\theta/\beta \sim \chi^2_{(2\alpha)}$, i.e. $88\theta \sim \chi^2_{(100)}$ and then

$P(\theta < q_1 \mid \underline{x}) = P(88\theta < 88q_1 \mid \underline{x}) = 0.025 \Rightarrow q_1 = 74.22/88 = 0.8434$

$P(\theta > q_2 \mid \underline{x}) = P(88\theta > 88q_2 \mid \underline{x}) = 0.025 \Rightarrow q_2 = 129.6/88 = 1.4727$

Exercise 7

- Define NR, the number of replicas and define, 2 arrays $\mathbf{t}$ and $\mathbf{v}$ both with size NR.
- For each replica, $j = 1, 2, \cdots, NR$
  - Generate 10 Pareto distributed variables $\left(x_1, x_2, \cdots, x_n\right)$ To generate each x_i: generate u_i as a Uniform(0,1) random variable and use the inverse method, i.e. compute $x_i = \theta\left(\left(1-u_i\right)^{-1/\alpha} - 1\right)$ where $\alpha = 2.1$ and $\theta = 110$.
  - Let $t_i = \overline{x}$ and $v_i = \dfrac{\sum_{i=1}^{n} x_i - x_{(1)} - x_{(n)}}{n-2}$ and keep these values as element i of the arrays $\mathbf{t}$ and $\mathbf{v}$ respectively ($n = 10$)
- Now, as $E(X) = \dfrac{\theta}{\alpha - 1} = \dfrac{110}{1.1} = 100$, compute $mse(T) \approx \dfrac{\sum_{j=1}^{NR} (t_j - 100)^2}{NR}$ and $mse(V) \approx \dfrac{\sum_{i=1}^{NR} (v_i - 100)^2}{NR}$ and compare.

MASTER IN ACTUARIAL SCIENCE

Risk Models

01/02/2019

2nd part of the exam

Time allowed: 1 hour

Instructions:

1. This paper contains 2 questions and comprises 2 pages including the title page.
2. Enter all requested details on the cover sheet.
3. You have 10 minutes of reading time. You must not start writing your answers until instructed to do so.
4. During the reading time you can download and install all the R packages that you need to answer the questions. Once the reading time is over no more access to the internet is allowed.
5. You are requested to summarize your answers on the cover sheet. You can add, using another paper sheets, any comments you think necessary to understand your answers. At the end of the exam you should submit your R files to Aquila using your usual username and password.
6. Attempt all questions.
7. Marks are shown in brackets. Total marks: **60**.
8. The distributed formulary and the Formulae and Tables for Actuarial Examinations (the 2002 edition) may be used. Note that the parametrization used for the different distributions is that of the distributed formulary.

1. In file 04cars.csv (source JSE Web site) specifications are given for a set of vehicles produced in year 2004. The variables recorded are:
   - Name of the vehicle
   - Retail Price (US$) - suggested retail price by the manufacturer, including adequate profit for the automaker and the dealer
   - Dealer Cost (US$) - Price paid by the dealership to the manufacturer
   - Engine size (liters)
   - Cylinders – Number of cylinders
   - Horsepower
   - City Miles (per gallon of fuel)
   - Highway Miles (per gallon of fuel)
   - Weight (pounds)
   - Wheel base (inches)
   - Length (inches)
   - Width (inches)

   a. **[10]** Assuming that the observations of the Retail Price are a random sample picked from a normal population test $H_0 : \mu = 30000$ against $H_1 : \mu > 30000$ where μ is the expected value of the Retail Price in the population (use $\alpha = 0.05$)

   b. **[15]** Now, run a PCA to **answer** the following questions

      i. How many principal components should you use according to Kaiser's criterion?

      ii. What are the loadings for each PC retained?

      iii. Plot the PC scores using the first 2 PC. Can you identify any outliers?

2. Let X be the ratio between a claim payment and insured capital. Let us assume that $X \sim beta(\alpha, \beta)$, i.e.

$$f_X(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha+1}(1-x)^{\beta+1}, \ 0 < x < 1, \ \alpha, \beta > 0$$

   File beta.csv presents a random sample with $n = 120$ observations. As it is well known,

$$E(X) = \frac{\alpha}{\alpha+\beta} \text{ and } E(X^2) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}.$$

   a) **[15]** Obtain the maximum likelihood estimates for α and β and determine a 95% confidence interval for α.

   b) **[10]** Using delta method, determine a 95% interval for μ.

   c) **[10]** What are the moment estimates for α and β?

Part 2 solutions

Exercise 1

a)

$H_0 : \mu = 30000 \qquad H_1 : \mu > 30000$

Test Statistic: $T = \dfrac{\bar{X} - 30000}{S / \sqrt{387}} \sim t_{(386)}$

$T_{obs} = 3.2226 \qquad p - value = 0.0007$ Reject the null.

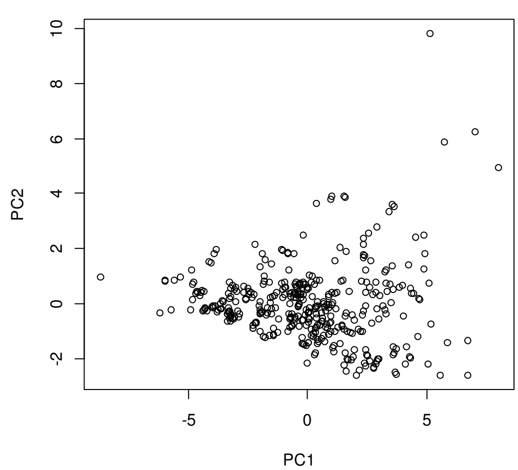
b)

According to Kaiser's criterion (eigenvalues greater than one for scaled data) we should use 2 Principal Components (PC)

The standardized loadings are

	PC1	PC2
Retail.Price	0.7030143	0.643056949
Dealer.Cost	0.6991979	0.645305050
Engine.size	0.9251267	-0.021064956
Cylinders	0.8907643	0.107103725
Horse.Power	0.8492193	0.401080935
City.Miles	-0.8275744	-0.004619953
Highway.Miles	-0.8171975	-0.015049395
Weight	0.8964700	-0.229854032
Wheel.base	0.7095703	-0.573973751
Length	0.6844621	-0.560569692
Width	0.7891195	-0.429462586

Plot the PC scores



Yes, we can identify 4 outliers in the upper right corner

Exercise 2

a)

$\hat{\alpha} = 2.1052$ and $\hat{\beta} = 5.4194$. Furthermore $\hat{\text{var}}(\hat{\alpha}, \hat{\beta}) = \begin{bmatrix} 0.06472 & 0.15213 \\ 0.15213 & 0.49544 \end{bmatrix}$

Then the 95% asymptotic CI for α is (1.6066, 2.6038)

b)

$$g(\alpha, \beta) = E(X) = \frac{\alpha}{\alpha + \beta}$$

$$\frac{\partial g}{\partial \alpha} = \frac{1 \times (\alpha + \beta) - 1 \times \alpha}{(\alpha + \beta)^2} = \frac{\beta}{(\alpha + \beta)^2} \quad \text{and} \quad \frac{\partial g}{\partial \beta} = \frac{-\alpha}{(\alpha + \beta)^2}.$$

Using the results presented in a),

$$g(\hat{\alpha}, \hat{\beta}) = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} = 0.2798$$

$$\hat{\text{var}}\left(g\left(\hat{\alpha}, \hat{\beta}\right)\right) \approx \begin{bmatrix} \dfrac{\hat{\beta}}{\left(\hat{\alpha} + \hat{\beta}\right)^2} & \dfrac{-\hat{\alpha}}{\left(\hat{\alpha} + \hat{\beta}\right)^2} \end{bmatrix} \begin{bmatrix} 0.06472 & 0.15213 \\ 0.15213 & 0.49544 \end{bmatrix} \begin{bmatrix} \dfrac{\hat{\beta}}{\left(\hat{\alpha} + \hat{\beta}\right)^2} \\ \dfrac{-\hat{\alpha}}{\left(\hat{\alpha} + \hat{\beta}\right)^2} \end{bmatrix} = 0.000195$$

And then the 95% asymptotic CI is (0.2524 0.3071)

c)

$\tilde{\alpha} = 2.1384$ and $\tilde{\beta} = 5.5157$

R programs and outputs

Exercise 1

```
> ########### Exam B - part 2
>
> #### Exercise 1
> rm(list=ls(all=TRUE))
> direct="C:/Users/joaoas/Desktop/trial/"
> file="04cars.csv"
>
> dta=read.csv(paste(direct,file,sep=""),header=T,sep=";") # read data set
> # Check reading
> head(dta)
                        Name Retail.Price Dealer.Cost Engine.size Cylinders
1          Chevrolet Aveo 4dr        11690       10965         1.6         4
2 Chevrolet Aveo LS 4dr hatch        12585       11802         1.6         4
3      Chevrolet Cavalier 2dr        14610       13697         2.2         4
4      Chevrolet Cavalier 4dr        14810       13884         2.2         4
5   Chevrolet Cavalier LS 2dr        16385       15357         2.2         4
6           Dodge Neon SE 4dr        13670       12849         2.0         4
  Horse.Power City.Miles Highway.Miles Weight Wheel.base Length Width
1         103         28            34   2370         98    167    66
2         103         28            34   2348         98    153    66
3         140         26            37   2617        104    183    69
4         140         26            37   2676        104    183    68
5         140         26            37   2617        104    183    69
6         132         29            36   2581        105    174    67
> attach(dta)
>
>
> ## a)
> t.test(Retail.Price,mu=30000,alternative="greater")

        One Sample t-test

data:  Retail.Price
t = 3.2226, df = 386, p-value = 0.0006891
alternative hypothesis: true mean is greater than 30000
95 percent confidence interval:
 31577.99      Inf
sample estimates:
mean of x
 33231.18

>
> ## b)
> x=cbind(Retail.Price,Dealer.Cost,Engine.size,Cylinders,Horse.Power,
+         City.Miles,Highway.Miles,Weight,Wheel.base,Length,Width)
> out1=prcomp(x,center=T,scale=T)
> out1
Standard deviations (1, .., p=11):
 [1] 2.66545276 1.37256139 0.92180708 0.59750773 0.52481958 0.44490866
 [7] 0.37485892 0.29434472 0.25765865 0.19229499 0.02811325

Rotation (n x k) = (11 x 11):
                     PC1          PC2         PC3         PC4         PC5
Retail.Price   0.2637504  0.468508698  0.25497414  0.27988372  0.04974519
Dealer.Cost    0.2623186  0.470146585  0.25725039  0.28771958  0.03681996
Engine.size    0.3470805 -0.015347186  0.04719422 -0.52537085  0.05199547
Cylinders      0.3341888  0.078032011  0.08144926 -0.63979956 -0.12572880
Horse.Power    0.3186023  0.292213476  0.07638087 -0.05834309 -0.11991697
City.Miles    -0.3104817 -0.003365936  0.53506296 -0.18627557  0.32572837
Highway.Miles -0.3065886 -0.010964460  0.59899766 -0.12574440  0.03974688
Weight         0.3363294 -0.167463572 -0.11221865  0.11954611  0.39721698
Wheel.base     0.2662100 -0.418177107  0.26445345  0.22070397 -0.22487841
Length         0.2567902 -0.408411381  0.34470795  0.16849408 -0.45631524
Width          0.2960546 -0.312891350  0.08759176  0.09065658  0.66258913
                       PC6         PC7          PC8          PC9         PC10
Retail.Price  -0.032492780  0.22241966 -0.051733338 -0.090697742  0.024222804
Dealer.Cost   -0.048557764  0.21944375 -0.066924964 -0.086684420  0.028296680
Engine.size   -0.005366507  0.05172424  0.358461569 -0.683332423 -0.010997289
Cylinders     -0.092893743  0.23997429 -0.423251772  0.452827436 -0.004140785
Horse.Power    0.208117703 -0.81352906  0.180257941  0.227458995  0.010574748
City.Miles    -0.248202294 -0.09254946  0.166248559  0.119762499  0.603612624
Highway.Miles  0.081731932 -0.05687409 -0.004478466 -0.010063971 -0.720867745
```

```
Weight        -0.541582756  0.03774491  0.417049636  0.335965459 -0.304350252
Wheel.base    -0.449967732 -0.29947371 -0.458409091 -0.284378470  0.041367733
Length         0.314058641  0.27751645  0.406364770  0.228686539  0.136883007
Width          0.530956852 -0.01888721 -0.275417442 -0.006361882  0.035872189
                       PC11
Retail.Price  -0.7092944127
Dealer.Cost    0.7046473100
Engine.size    0.0095506165
Cylinders     -0.0038579751
Horse.Power    0.0010013109
City.Miles    -0.0004696922
Highway.Miles  0.0026644793
Weight         0.0020139958
Wheel.base    -0.0129305351
Length         0.0025480564
Width          0.0090242027
> summary(out1)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.6655 1.3726 0.92181 0.59751 0.52482 0.44491 0.37486
Proportion of Variance 0.6459 0.1713 0.07725 0.03246 0.02504 0.01799 0.01277
Cumulative Proportion  0.6459 0.8171 0.89439 0.92685 0.95189 0.96988 0.98266
                           PC8     PC9    PC10    PC11
Standard deviation     0.29434 0.25766 0.19229 0.02811
Proportion of Variance 0.00788 0.00604 0.00336 0.00007
Cumulative Proportion  0.99053 0.99657 0.99993 1.00000
>
> # How many components should be retained?
> # Kaiser criterion -> 2 (81.71% of the total variance explained)
>
> ## loadings
> #1st alternative
> W=out1$rotation[,1:2]
> sqrt.lamb=out1$sdev[1:2]
> cbind(W[,1]*sqrt.lamb[1],W[,2]*sqrt.lamb[2])
                    [,1]         [,2]
Retail.Price   0.7030143  0.643056949
Dealer.Cost    0.6991979  0.645305050
Engine.size    0.9251267 -0.021064956
Cylinders      0.8907643  0.107103725
Horse.Power    0.8492193  0.401080935
City.Miles    -0.8275744 -0.004619953
Highway.Miles -0.8171975 -0.015049395
Weight         0.8964700 -0.229854032
Wheel.base     0.7095703 -0.573973751
Length         0.6844621 -0.560569692
Width          0.7891195 -0.429462586
> # 2nd alternative
> cor(x,out1$x[,1:2])
                     PC1          PC2
Retail.Price   0.7030143  0.643056949
Dealer.Cost    0.6991979  0.645305050
Engine.size    0.9251267 -0.021064956
Cylinders      0.8907643  0.107103725
Horse.Power    0.8492193  0.401080935
City.Miles    -0.8275744 -0.004619953
Highway.Miles -0.8171975 -0.015049395
Weight         0.8964700 -0.229854032
Wheel.base     0.7095703 -0.573973751
Length         0.6844621 -0.560569692
Width          0.7891195 -0.429462586
>
> ## Plotting scores
> plot(out1$x[,1:2])
>
```

Exercise 2

```
> #### Exercise 2
> rm(list=ls(all=TRUE))
> # data generated using a beta distribution with alpha=2.3 and beta=5.1
> direct="C:/Users/joaoas/Desktop/trial/"
```

```
> file="beta.csv"
>
> dta=read.csv(paste(direct,file,sep=""),header=T) # read data set
> # Check reading
> head(dta)
           x
1 0.16618834
2 0.07277618
3 0.20775820
4 0.47662196
5 0.24840555
6 0.63443500
> attach(dta)
>
> # a)
> minusloglikbeta=function(param,x){
+   alpha=param[1]; beta=param[2]
+   return(-sum(dbeta(x,alpha,beta,log=T)))
+   }
>
> out=nlm(minusloglikbeta,c(.5,.5),hessian=T,x=x)
Warning messages:
1: In dbeta(x, alpha, beta, log = T) : NaNs produced
2: In nlm(minusloglikbeta, c(0.5, 0.5), hessian = T, x = x) :
  NA/Inf replaced by maximum positive value
3: In dbeta(x, alpha, beta, log = T) : NaNs produced
4: In nlm(minusloglikbeta, c(0.5, 0.5), hessian = T, x = x) :
  NA/Inf replaced by maximum positive value
> out
$`minimum`
[1] -62.36104
$estimate
[1] 2.105201 5.419423
$gradient
[1] -3.375178e-09 -9.833281e-08
$hessian
          [,1]       [,2]
[1,]  55.53646 -17.053206
[2,] -17.05321   7.254821
$code
[1] 1
$iterations
[1] 14

> vcov=solve(out$hessian); vcov
           [,1]      [,2]
[1,] 0.06472013 0.1521314
[2,] 0.15213135 0.4954398
>
cbind(out$estimate[1]+qnorm(0.025)*sqrt(vcov[1,1]),out$estimate[1]+qnorm(0.975)*sqrt(vco
v[1,1]))
         [,1]     [,2]
[1,] 1.606583 2.603819
>
> # b)
> dg=matrix(c(out$estimate[2]/(out$estimate[1]+out$estimate[2])^2,-
out$estimate[1]/(out$estimate[1]+out$estimate[2])^2),nrow=2,ncol=1)
> var.mu=t(dg)%*%vcov%*%dg; var.mu
             [,1]
[1,] 0.0001950327
> g.hat=out$estimate[1]/(out$estimate[1]+out$estimate[2]); g.hat
[1] 0.2797749
> cbind(g.hat+qnorm(0.025)*sqrt(var.mu),g.hat+qnorm(0.975)*sqrt(var.mu))
          [,1]      [,2]
[1,] 0.2524032 0.3071466
>
> # c)
> met.mom=function(param,m){
+   alpha=param[1]; beta=param[2]
+   eq1=m[1]-alpha/(alpha+beta)
+   eq2=m[2]-alpha*(alpha+1)/((alpha+beta)*(alpha+beta+1))
+   return(c(eq1,eq2))
+   }
>
> m=c(mean(x),mean(x^2))
>
> require(nleqslv);
```

```
Loading required package: nleqslv
> nleqslv(c(.5,.5),met.mom,m=m)
$`x`
[1] 2.138413 5.515694
$fvec
[1] -5.604143e-09 -2.924847e-09
$termcd
[1] 1
$message
[1] "Function criterion near zero"
$scalex
[1] 1 1
$nfcnt
[1] 40
$njcnt
[1] 1
$iter
[1] 23
>
> # or
>
> nlmfunc=function(param,m){
+   return(crossprod(met.mom(param,m),met.mom(param,m)))
+   }
>
> nlm(nlmfunc,c(.5,.5),m=m)
$`minimum`
[1] 9.900847e-12
$estimate
[1] 2.138018 5.514719
$gradient
[1]  1.909861e-08 -6.983457e-09
$code
[1] 1
$iterations
[1] 28
```

R programs and outputs – Part 2

Exercise 1

```
> rm(list=ls(all=TRUE))
> Female=c(28.8,41.6,39.8,38.5,42.6)
> Male=c(29.3,41.5,40.4,38.5,43.5)
> n=length(Male)
> #a)
> t.test(Male,alternatve="two.sided",conf=0.9)

        One Sample t-test

data:  Male
t = 15.635, df = 4, p-value = 9.772e-05
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 33.37145 43.90855
sample estimates:
mean of x
    38.64

> #b)
> D=Male-Female
> t.test(D,alternative="greater")

        One Sample t-test

data:  D
t = 2.0197, df = 4, p-value = 0.05677
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 -0.02110458         Inf
sample estimates:
mean of x
     0.38
```

Exercise 2

```
> rm(list=ls(all=TRUE))
>
> # b)
> x=0:5
> n.x=c(17220,1920,793,50,15,2); n=sum(n.x)
> p.est=n.x[4]/n; p.est
[1] 0.0025
> cbind(p.est+qnorm(0.025)*sqrt(p.est*(1-p.est)/n),p.est+qnorm(0.975)*sqrt(p.est*(1-
p.est)/n))
            [,1]        [,2]
[1,] 0.001807915 0.003192085
>
> # c)
> theta.hat=crossprod(x,n.x)/sum(n.x); theta.hat
       [,1]
[1,] 0.1863
> p.hat=dpois(3,theta.hat); p.hat
[1] 0.0008944957
>
> minusloglik.Poisson=function(theta,x){
+   return(-sum(dpois(x,theta,log=T)))
+   }
> xx=c(rep(0,17220),rep(1,1920),rep(2,793),rep(3,50),rep(4,15),rep(5,2))
> out=nlm(minusloglik.Poisson,1,hessian=T,x=xx); out
There were 14 warnings (use warnings() to see them)
$`minimum`
[1] 10683.66
$estimate
[1] 0.1862995
$gradient
[1] 0.004343747
$hessian
         [,1]
[1,] 107239.1
$code
[1] 1
```

```
$iterations
[1] 10

>
> minusloglik.Poisson1=function(theta,x,n){
+   p=dpois(x,theta,log=T)
+   return(-crossprod(n,p))
+   }
> xxx=0:5;
> out=nlm(minusloglik.Poisson1,1,hessian=T,x=xxx,n=n.x); out
There were 14 warnings (use warnings() to see them)
$`minimum`
[1] 10683.66
$estimate
[1] 0.1862995
$gradient
[1] 0.004340109
$hessian
         [,1]
[1,] 107239.1
$code
[1] 1
$iterations
[1] 10
```

Exercise 3

```
> rm(list=ls(all=TRUE))
> # a)
> y=c(1,1,2,2,2,4,4,4,6,6)
> sum((1/length(y))*(24*y^3*(5+2*y)^(-4)))# f5.hat
[1] 0.03932585
> f.ker=function(x) return(sum((1/length(y))*(24*y^3*(x+2*y)^(-4))))
> f.ker(5)
[1] 0.03932585
```

Exercise 4

```
> rm(list=ls(all=TRUE))
> # a)
> mloglik1=function(theta){
+   loglik=-0.044*theta^3+3*76*log(theta)-0.0296*theta^3
+   return(-loglik)
+   }
> nlm(mloglik1,1,hessian=T)
$`minimum`
[1] -451.4281
$estimate
[1] 10.10753
$gradient
[1] 1.687161e-08
$hessian
         [,1]
[1,] 6.695231
$code
[1] 1
$iterations
[1] 5

>
> theta.hat=(228/0.2208)^(1/3); theta.hat # valor tirado da eq.
[1] 10.10754
> (76/(0.044+0.0296))^(1/3) # verificação
[1] 10.10754
>
> hess=-0.4416*theta.hat-228/(theta.hat^2); hess
[1] -6.695231
> var.theta.hat=-1/hess; var.theta.hat
[1] 0.14936
>
> # b)
> p.hat=exp(-(theta.hat/10)^3); p.hat
[1] 0.3560769
>
> der.g.theta=-0.003*(theta.hat^2)*p.hat; der.g.theta
[1] -0.1091329
> var.p.hat=(der.g.theta^2)*var.theta.hat; var.p.hat
```

```
[1] 0.001778875
> cbind(p.hat+qnorm(0.025)*sqrt(var.p.hat),p.hat+qnorm(0.975)*sqrt(var.p.hat))
         [,1]      [,2]
[1,] 0.273412 0.4387417
```

Exercise 6

```
> require(nleqslv);
>
> HPD.gamma=function(bounds,param,prob){
+   alpha=param[1]; beta=param[2]
+   a=bounds[1]; b=bounds[2];
+   eq1=pgamma(b,shape=alpha,scale=beta)-pgamma(a,shape=alpha,scale=beta)-prob
+   eq2=dgamma(b,shape=alpha,scale=beta)-dgamma(a,shape=alpha,scale=beta)
+   return(c(eq1,eq2))
+   }
>
> par.g=c(50,1/44)
> p=0.95;
> a=qgamma((1-p)/2,shape=par.g[1],scale=par.g[2]); a
[1] 0.843431
> b=qgamma((1+p)/2,shape=par.g[1],scale=par.g[2]); b
[1] 1.472286
>
> out=nleqslv(c(a,b),HPD.gamma,param=par.g,prob=p); out
$`x`
[1] 0.8297868 1.4558367
$fvec
[1] -1.528504e-10 -1.411124e-09
$termcd
[1] 1
$message
[1] "Function criterion near zero"
$scalex
[1] 1 1
$nfcnt
[1] 8
$njcnt
[1] 1
$iter
[1] 6
>
> a=out$x[1]; b=out$x[2]
> pgamma(b,shape=par.g[1],scale=par.g[2])-pgamma(a,shape=par.g[1],scale=par.g[2])
[1] 0.95
> dgamma(a,shape=par.g[1],scale=par.g[2])
[1] 0.3643579
> dgamma(b,shape=par.g[1],scale=par.g[2])
[1] 0.3643579
```