STAT 40001/MA 59800 Statistical Computing/ Computational Statistics Fall 2013
Homework 7-Solution

Name:
Due : December 5, 2013 PUID:

Instruction: Please submit your R code along with a brief write-up of the solutions (do not submit raw R codes with Errors!). Some of the questions below can be answered with very little or no programming. However, write code that outputs the final answer and does not require any additional paper calculations.

Q.N. 1) Data below gives the amount of chemical yield($\boldsymbol{y}$) on using another chemical($\boldsymbol{x}$)

$\boldsymbol{x}$	23.1	32.8	31.8	32.0	30.4	24.0	39.5	24.2	52.5	37.9	30.5	25.1	12.4	35.1	31.5	21.1
$\boldsymbol{y}$	10.5	16.7	18.2	17.0	16.3	10.5	23.1	12.4	24.9	22.8	14.1	12.9	8.8	17.4	14.9	10.5

a) Fit a simple linear regression of $\boldsymbol{y}$ as a function of $\boldsymbol{x}$. List the assumptions that you make.
b) Calculate a **90%** confidence interval for the slope of your model.
c) In the context of the property of the chemical when $\boldsymbol{x = 0}$ then $\boldsymbol{y = 0}$**,** fit a simple linear regression model.
d) Which model (model in(a) or model in (c)) is appropriate for the representation of the given data?

Solution:
a) Using the R code below we have the simple regression model of the subject data is

$$\boldsymbol{y = 0.51802 + 0.50157x}$$

```
> x=c(23.1,32.8,31.8,32.0,30.4,24.0,39.5,24.2,52.5,37.9,30.5,25.1,12.4,35.1,31.5,21.1)
> y=c(10.5,16.7,18.2,17.0,16.3,10.5,23.1,12.4,24.9,22.8,14.1,12.9,8.8,17.4,14.9,10.5)
> model1=lm(y~x)
> summary(model1)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0558 -1.4643 -0.2629  0.8336  3.2723

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.51802    1.56746    0.33    0.746
x            0.50157    0.04977   10.08  8.5e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.747 on 14 degrees of freedom
Multiple R-squared: 0.8788,     Adjusted R-squared: 0.8702
F-statistic: 101.5 on 1 and 14 DF,  p-value: 8.496e-08
```

We assume that the data fits a simple linear regression model $\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \epsilon_i}$*, where* $\boldsymbol{\epsilon_i}$ *are independent and normally distributed with mean 0 and constant variance.*

b)

```
> confint(model1,level=0.9)
                    5 %      95 %
(Intercept) -2.2427613 3.2788045
x            0.4139049 0.5892431
```

90% *confidence interval for* $\boldsymbol{\beta_1}$ *is (0.4139049,0.5892431).*

c)Using the R code below we have the simple regression model through origin of the subject data $\boldsymbol{y = 0.5174x}$ *and a* **90%** *CI for* $\boldsymbol{\beta_1}$ *is given by (0.4937915, 0.5409532).*

```
> model2=lm(y~-1+x)
> summary(model2)

Call:
lm(formula = y ~ -1 + x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2620 -1.4107 -0.1951  0.8658  3.1916

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x  0.51737    0.01345   38.46   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.694 on 15 degrees of freedom
Multiple R-squared:  0.99,      Adjusted R-squared: 0.9893
F-statistic:  1479 on 1 and 15 DF,  p-value: < 2.2e-16
```

d) In examining the summary of each model and $\boldsymbol{R^2}$ *and* $\boldsymbol{R^2_{Adj.}}$*, we decide that the no-intercept model is more appropriate.*

We may perform the residual analysis and draw the same conclusion.

Q.N. 2) The data set `Cars93` provided in the library `MASS` contains data on cars sold in the United States in the year 1993.

a) How many variables are included in the data set?

Solution: Based on the R code below it appears that there are 27 variables included in the data.

```
> library(MASS)
> data(Cars93)
> attach(Cars93)
> dim(Cars93)
[1] 93 27
```

b) Fit a regression model for `MPG.city` using the numerical variables `EngineSize`, `Weight`, `Passengers`, and `Price`.

Solution: We use R code below to estimate the model parameters

```
> library(MASS)
> data(Cars93)
> attach(Cars93)
> model=lm(MPG.city~EngineSize+Weight+Passengers+Price)
> model
```

```
Call:
lm(formula = MPG.city ~ EngineSize + Weight + Passengers + Price)

Coefficients:
(Intercept)   EngineSize       Weight   Passengers        Price
  46.389413     0.196119    -0.008207     0.269622    -0.035804
```

Hence, the desired regression model is

$$\boldsymbol{MPG.city = 46.3894 + 0.1961 \times EngineSize - 0.0082 \times Weight + 0.2696 \times Passengers - 0.0358 \times Price}$$

c) Which variables are marked as statistically significant by the marginal t-test?
Solution: We use R code below to test the significance of the model parameters

```
> library(MASS)
> data(Cars93)
> attach(Cars93)
> model=lm(MPG.city~EngineSize+Weight+Passengers+Price)
> summary(model)
Call:
lm(formula = MPG.city ~ EngineSize + Weight + Passengers + Price)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1207 -1.9098  0.0522  1.1294 13.9580

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.389413   2.097516  22.116  < 2e-16 ***
EngineSize   0.196119   0.588880   0.333    0.740
Weight      -0.008207   0.001343  -6.111 2.63e-08 ***
Passengers   0.269622   0.424951   0.634    0.527
Price       -0.035804   0.049179  -0.728    0.469
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.06 on 88 degrees of freedom
Multiple R-squared: 0.7165,     Adjusted R-squared: 0.7036
F-statistic: 55.59 on 4 and 88 DF,  p-value: < 2.2e-16
```

It appears that the Weight is the significance variable to determine the City MPG.

d) Which model is selected by AIC criteria?
Solution: Using R code below we can perform the model selection using AIC criteria

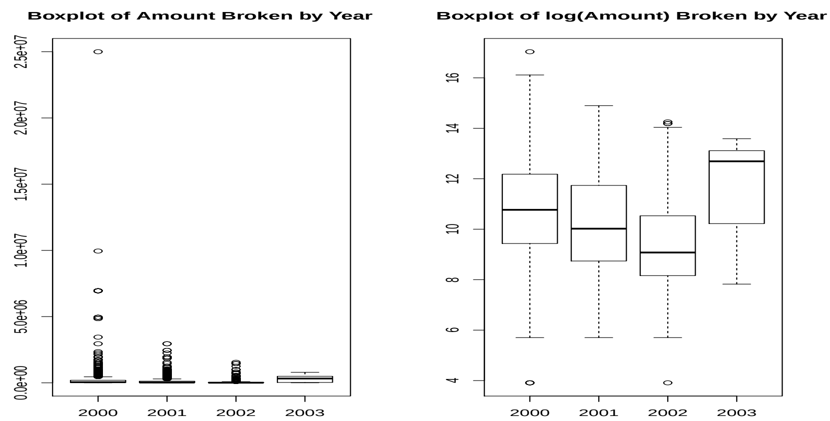
```
> best=stepAIC(model)
Start:  AIC=212.87
MPG.city ~ EngineSize + Weight + Passengers + Price

             Df Sum of Sq     RSS    AIC
- EngineSize  1      1.04  824.89 210.99
- Passengers  1      3.77  827.62 211.29
- Price       1      4.96  828.82 211.43
<none>                     823.85 212.87
- Weight      1    349.67 1173.52 243.77
```

```
Step:  AIC=210.99
MPG.city ~ Weight + Passengers + Price
             Df Sum of Sq     RSS    AIC
- Passengers  1      3.20  828.10 209.35
- Price       1      4.84  829.74 209.53
<none>                     824.89 210.99
- Weight      1    627.12 1452.01 261.57

Step:  AIC=209.35
MPG.city ~ Weight + Price
         Df Sum of Sq     RSS    AIC
- Price   1     11.96  840.05 208.68
<none>                 828.10 209.35
- Weight  1   1050.34 1878.44 283.52

Step:  AIC=208.68
MPG.city ~ Weight
         Df Sum of Sq     RSS    AIC
<none>                 840.05 208.68
- Weight  1    2065.5 2905.57 322.09
```

It appears that the model containing only the the regressor variable Weight is the best model

Q.N. 3) The data set National Practitioner Data Bank (`npdb`) included in the `UsingR` package contains malpractice award information. The variable `amount` contains the amount of settlement and the variable `year` contains the year of the award. We wish to investigate whether the dollar amount awarded was steady during the years 2000, 2001, 2002 and 2003.
a) Make boxplots of the `amount` and `log(amount)` broken by years.
Solution: We can use R code below to display the information using box plot.

```
> library(UsingR)
> data(npdb)
> attach(npdb)
> par(mfrow=c(1,2))
> boxplot(amount~factor(year),main="Boxplot of Amount Broken by Year")
> boxplot(log(amount)~factor(year),main="Boxplot of log(Amount) Broken by Year")
```



Note that the log transformation helped to better visualize the information contained in the data.

b) Perform the complete analysis of variance of `log(amount)` by factor(`year`) for the years 2000, 2001 and 2002.
Solution: We can extract only two variables "year" and "amount" and then perform the analysis using R code below

```
> library(UsingR)
> data(npdb)
> attach(npdb)
> data<- subset(npdb, select=c("amount","year"))
> summary(aov(log(amount)~factor(year)))
               Df Sum Sq Mean Sq F value Pr(>F)
factor(year)    3    827  275.74   79.02 <2e-16 ***
Residuals    6793  23705    3.49
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

It appears that there is a significance difference in the amount of settlement from year to year.
In order to perform the pairwise comparison using Tukey's method we use R code below

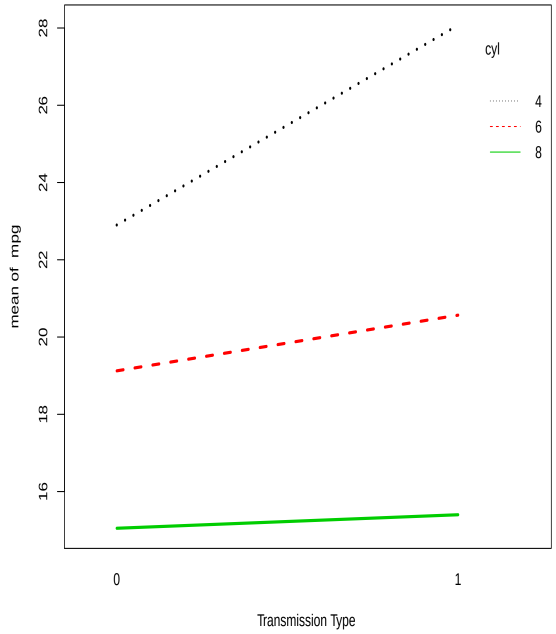
```
> TukeyHSD(aov(log(amount)~factor(year)))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log(amount) ~ factor(year))

$`factor(year)`
                diff         lwr        upr     p adj
2001-2000 -0.4872275 -0.62058561 -0.3538695 0.0000000
2002-2000 -1.2850610 -1.53063132 -1.0394906 0.0000000
2003-2000  0.7794277 -1.36849742  2.9273528 0.7874041
2002-2001 -0.7978334 -1.05859390 -0.5370730 0.0000000
2003-2001  1.2666552 -0.88305953  3.4163700 0.4290051
2003-2002  2.0644886 -0.09509324  4.2240705 0.0670148
```

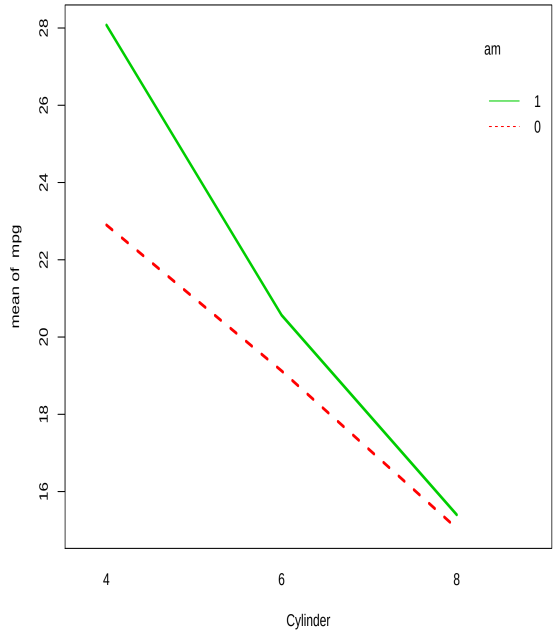
Note that at $\boldsymbol{\alpha = 0.05}$ *there is a significant difference in the settlement amount in 2000 and 2001, 2000 and 2002 and also in 2001 and 2002.*

Q.N. 4) In the data set `mtcars` in the `UsingR` package the variable `mpg`, `cyl` and `am` indicates the miles per gallon, the number of cylinder and the type of transmission respectively. Perform a two way ANOVA modeling `mpg` by the `cyl` and `am`, each treated as categorical variable.
Solution: First we will access and draw interaction plots using R code below

```
> library(UsingR)
> data(mtcars)
> attach(mtcars)
> dim(mtcars)
[1] 32 11
> data<- subset(mtcars, select=c("mpg","cyl", "am"))
> head(data,5)
                   mpg cyl am
Mazda RX4         21.0   6  1
Mazda RX4 Wag     21.0   6  1
Datsun 710        22.8   4  1
Hornet 4 Drive    21.4   6  0
Hornet Sportabout 18.7   8  0
> interaction.plot(am,cyl,mpg,col=c(1,2,3),lwd=3,xlab="Transmission Type",main="Interaction Plot")
> interaction.plot(cyl,am, mpg, col=c(2,3), lwd=3, xlab="Cylinder", main="Interaction Plot")
```

To perform the analysis of the variables we use the R code below






```
>  model1=lm(mpg~factor(cyl)+factor(am))
> model1

Call:
lm(formula = mpg ~ factor(cyl) + factor(am))

Coefficients:
 (Intercept)  factor(cyl)6  factor(cyl)8   factor(am)1
      24.802        -6.156       -10.068         2.560

> model2=lm(mpg~factor(cyl)*factor(am))

> summary(aov(model2))
                       Df Sum Sq Mean Sq F value   Pr(>F)
factor(cyl)             2  824.8   412.4  44.852 3.73e-09 ***
factor(am)              1   36.8    36.8   3.999   0.0561 .
factor(cyl):factor(am)  2   25.4    12.7   1.383   0.2686
Residuals              26  239.1     9.2
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> anova(model1,model2)
Analysis of Variance Table

Model 1: mpg ~ factor(cyl) + factor(am)
Model 2: mpg ~ factor(cyl) * factor(am)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     28 264.50
2     26 239.06  2    25.436 1.3832 0.2686
```

It appears that the interaction is not significant. We can construct pairwise confidence intervals for the treatment factors using Tukey method

```
> TukeyHSD(aov(mpg~factor(cyl)*factor(am)))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mpg ~ factor(cyl) * factor(am))

$`factor(cyl)`
          diff        lwr       upr     p adj
6-4  -6.920779 -10.563826 -3.277732 0.0002015
8-4 -11.563636 -14.599509 -8.527764 0.0000000
8-6  -4.642857  -8.130809 -1.154905 0.0075037

$`factor(am)`
        diff        lwr      upr     p adj
1-0 1.860708 -0.3827415 4.104157 0.1001455

$`factor(cyl):factor(am)`
              diff         lwr       upr     p adj
6:0-4:0  -3.775000 -10.8905739  3.340574 0.5871784
8:0-4:0  -7.850000 -13.8637575 -1.836242 0.0054390
4:1-4:0   5.175000  -1.1322821 11.482282 0.1546661
6:1-4:0  -2.333333  -9.9402018  5.273535 0.9315095
8:1-4:0  -7.500000 -16.0047375  1.004737 0.1072775
8:0-6:0  -4.075000  -9.4538683  1.303868 0.2192160
4:1-6:0   8.950000   3.2448487 14.655151 0.0006955
6:1-6:0   1.441667  -5.6739072  8.557241 0.9883098
8:1-6:0  -3.725000 -11.7933024  4.343302 0.7158963
4:1-8:0  13.025000   8.7726313 17.277369 0.0000000
6:1-8:0   5.516667  -0.4970909 11.530424 0.0859484
8:1-8:0   0.350000  -6.7655739  7.465574 0.9999875
6:1-4:1  -7.508333 -13.8156155 -1.201051 0.0129262
8:1-4:1 -12.675000 -20.0403187 -5.309681 0.0002083
8:1-6:1  -5.166667 -13.6714041  3.338071 0.4436999
```

It can be observed that all three cylinder types are significanly differenet each other whereas the transmission type is not different.

Q.N. 5) According to the web site http://www.keepkidshealthy.com, risk factors associated with premature births include smoking and maternal malnutrition. A birth is consider premature if the gestation period is less than 37 full weeks. Also note that the body mass index(BMI) can be used as a measure of malnutrition. Do you find this to be true with the data in `babies` provided in the `UsingR` package?
Tasks to perform:
a) Extract the variables of interest: gestation, smoking status, mother's height and weight, and birth weight of the babies.
b) Clean the data set as there are some missing values coded as 9, 99, or 999.
c) Calculate the BMI of mothers.
d) Create indicator variable(1 for premature and 0 for not premature) babies.
e) Fit a logistic regression model with `smoke` and `BMI` as a predictor variable and `premature` as a response variable.
Solution: We use R code below to extract the variables of interest

```
> library(UsingR)
> data(babies)
```

```
> data=subset(babies,select=c("gestation","smoke","wt1","ht","wt"))
> dim(data)
[1] 1236    5
> head(data,5)
  gestation smoke wt1 ht  wt
1       284     0 100 62 120
2       282     0 135 64 113
3       279     1 115 64 128
4       999     3 190 69 123
5       282     1 125 67 108
```

We clean the data set using the following R code:

```
> library(UsingR)
> data(babies)
> data=subset(babies,select=c("gestation","smoke","wt1","ht","wt"))
> Clean=subset(data, gestation !=999&smoke!=9 & wt1!=999 & ht!=99 & wt!=999)
> dim(Clean)
[1] 1175    5
```

We calculate the BMI of mothers

```
> library(UsingR)
> data(babies)
> data=subset(babies,select=c("gestation","smoke","wt1","ht","wt"))
> Clean=subset(data, gestation !=999&smoke!=9 & wt1!=999 & ht!=99 & wt!=999)
> attach(Clean)
> BMI=wt1/(ht)^2*703
> BMI[1:10]
 [1] 18.28824 23.17017 19.73755 19.57563 17.00806 32.55307 23.29467 22.86030
 [9] 21.94858 18.24394
```

We create an indicator variable premature using the R code below.

```
> library(UsingR)
> data(babies)
> data=subset(babies,select=c("gestation","smoke","wt1","ht","wt"))
> Clean=subset(data, gestation !=999&smoke!=9 & wt1!=999 & ht!=99 & wt!=999)
> dim(Clean)
[1] 1175    5
> preemie=as.numeric(Clean$gestation<7*37)
> table(preemie)
preemie
   0    1
1079   96
```

We can now model the variable preemie by the levels of smoke and the variable BMI.

```
> model=glm(preemie~factor(Clean$smoke)+BMI, family=binomial)
> summary(model)

Call:
glm(formula = preemie ~ factor(Clean$smoke) + BMI, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6306  -0.4262  -0.4040  -0.3810   2.3891
```

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.42458    0.71159  -4.813 1.49e-06 ***
factor(Clean$smoke)1  0.19353    0.23569   0.821    0.412
factor(Clean$smoke)2  0.31370    0.38896   0.806    0.420
factor(Clean$smoke)3  0.10114    0.40499   0.250    0.803
BMI                   0.04023    0.03050   1.319    0.187
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 664.83  on 1174  degrees of freedom
Residual deviance: 662.34  on 1170  degrees of freedom
AIC: 672.34

Number of Fisher Scoring iterations: 5
```

Note that none of the variables are flagged as significant. This indicates that the model with no effect is , perhaps, preferred. In order to check which model is preferred by AIC we use R code below

```
> library(MASS)
> stepAIC(model)
Start:  AIC=672.34
preemie ~ factor(Clean$smoke) + BMI

                      Df Deviance    AIC
- factor(Clean$smoke)  3   663.35 667.35
- BMI                  1   663.98 671.98
<none>                     662.34 672.34

Step:  AIC=667.35
preemie ~ BMI

       Df Deviance    AIC
- BMI   1   664.83 666.83
<none>      663.35 667.35

Step:  AIC=666.83
preemie ~ 1
Call:  glm(formula = preemie ~ 1, family = binomial)

Coefficients:
(Intercept)
     -2.419

Degrees of Freedom: 1174 Total (i.e. Null);  1174 Residual
Null Deviance:      664.8
Residual Deviance: 664.8        AIC: 666.8
```

Since, the model of constant mean is chosen, this data don't indicates that neither smoking status nor BMI are the risk factors for premature babies.