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

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?

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?
b) Fit a regression model for `MPG.city` using the numerical variables `EngineSize`, `Weight`, `Passengers`, and `Price`.
c) Which variables are marked as statistically significant by the marginal t-test?
d) Which model is selected by AIC criteria?

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 and 2002.
a) Make boxplots of the `amount` and `log(amount)` broken by years.
b) Perform the complete analysis of variance of `log(amount)` by factor(`year`) for the years 2000, 2001 and 2002.

Q.N. 4) In the data set `mtcars` data set 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.

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.