```
###Assignment  1
###Instructor: Prof. Gunnar Heins
###Student:    C.H.(Chenhui Lu)
###UFID:       76982846

##4
library(AER)
library(data.table)

##5
data("NMES1988",package = "AER") #Imput the data
DataVisits <- as.data.table(NMES1988) # convert the data to data.table

##6
DataVisits$visits #selet each (sepcific) variable
mean(DataVisits$visits) #compute the mean of sepcific varibale: 5.774399
max(DataVisits$visits) #show the maximum of sepcific variable: 89
hist(DataVisits$visits) #plot the histogram of sepcific variable

##7
mean(DataVisits$age) # 7.402406
min(DataVisits$age) # 6.6
#Q: What does this tell you about the type of households the study focused on?
#A: This study focued on the elder households, with average age of 74.02 and minimum age
of 66

##8
summary(DataVisits) #summarize the details of Datavisits
#Q: What does the summary() function do?
#A: This function will summarize the details of each variables in 6 fields: 1.Minimum.
2.1st Quator. 3.Median. 4. Mean. 5.3rd Quator. 6.Maximum. different fields of dummy
variables.
#Q: How many people have insurance in the sample?
#A: 3421
#Q: How many medicaid?
#A: 402

##9
DataVisits_ins <- DataVisits[insurance == "yes"] # select the group of sample who has the
insurance
#Q: Do households with insurance go to the doctor more or less often on average compared
to the full sample?
mean(DataVisits_ins$visits) # 6.02
#A: 6.02 > 5.77 Thus, the insured sample group visits the doctor more frequently compared
to the full sample

##10
DataVisits_med <- DataVisits[medicaid == "yes"] # select the group of sample with
medicaid.
#Q: Do these households visit their doctor more or less often compared to the full sample?
mean(DataVisits_med$visits) # 6.71
#A: 6.71 > 5.77 Thus, the group of sample with medicaid visits the doctor more frequently.

##11
DataVisits$age = DataVisits$age * 10
mean(DataVisits$age)
#Q: What does this line do?
#A: Make the age variable to show the real age of each individuals (previous number * 10),
since the previous age variable show the really age / 10.

##12
DataVisits$income <- DataVisits$income * 10000 # Make the income variable to show the real
income of each individuals
#Q: How many households with an annual income above $30,000 have insurance?
```

```
DataVisits_high_inc <- DataVisits[income > 30000] #only keep the sample whose income above
30000
summary(DataVisits_high_inc$insurance) # show the details of sample on ownership of
insurance
#  no  yes
#  150 1034
#A: 1034
#Q: How many have medicaid?
summary(DataVisits_high_inc$medicaid) # show the details of sample on ownership of
medicaid
#  no  yes
#  1151   33
#A: 33


##13
reg <- lm(visits ~ insurance + medicaid, data = DataVisits) # regression visits on
insurance and medicaid
summary(reg) #summarize results of regression
#                Estimate Std. Error t value Pr(>|t|)
#  (Intercept)    4.1093     0.2549  16.122  < 2e-16 ***
#  insuranceyes   1.8718     0.2761   6.779 1.37e-11 ***
#  medicaidyes    2.3206     0.3995   5.808 6.75e-09 ***

#Q: Are the results of this regression qualitatively in line with your findings in (9) and
(10)?
#A: Yes they are. Form the results of regression, The possession of insurance and medicaid
is indeed significantly positively correlated with the frequency of individual visits to
the doctor.

##14
reg <- lm(visits ~ insurance + medicaid + health, data = DataVisits)# regression visits on
insurance, medicaid and health
summary(reg)#summarize results of regression
#                Estimate Std. Error t value Pr(>|t|)
#  (Intercept)       3.6931     0.2569  14.377  < 2e-16 ***
#  insuranceyes      2.1113     0.2714   7.779 9.06e-15 ***
#  medicaidyes       1.7932     0.3938   4.554 5.42e-06 ***
#  healthpoor        3.5043     0.3051  11.485  < 2e-16 ***
#  healthexcellent  -2.0850     0.3729  -5.591 2.40e-08 ***

#Q: Show that the results change only little if you also include the variable health in
the above regression.
#A: By comparison, after adding the variable of "health",  the coefficients form #13 and
#14 related to insuranceyes adn medicaidyes show similar pattern, positive with only
little difference.

##15
reg <- lm(visits ~ insurance + medicaid + health + age + income + school, data =
DataVisits)# regression visits on insurance, medicaid, health, age, income, and education
levels
summary(reg)#summarize results of regression
#                   Estimate Std. Error t value Pr(>|t|)
#  (Intercept)      2.762e+00  1.258e+00   2.196   0.0282 *
#  insuranceyes     1.763e+00  2.784e-01   6.332 2.66e-10 ***
#  medicaidyes      2.067e+00  3.967e-01   5.210 1.98e-07 ***
#  healthpoor       3.672e+00  3.065e-01  11.982  < 2e-16 ***
#  healthexcellent -2.196e+00  3.730e-01  -5.888 4.20e-09 ***
#  age             -5.827e-03  1.587e-02  -0.367   0.7134
#  income          -2.647e-06  3.528e-06  -0.750   0.4531
#  school           1.615e-01  2.930e-02   5.512 3.74e-08 ***

#Q: Very briefly describe how they affect the coefficients on insurance, medicaid, and
health.
#A: By comparison, the coefficents shows small decreases in variables of insurance,
```

medicaid, healthexcellent, while, small increases in healthpoor. Moreover, Insurance, Medicaid and health are still showing significant.

#Q: Why do you think R reports e.g. the coefficient on insurance as insuranceyes, but does not do so for age, income, and school?
#A: Since insurance is a dummy variable, which means this variable only shows binary value (0 for no or 1 or yes, for example). However, variables of age, income, and school are continuous variables, can be used and regressed directly.

#Q: And why does it report two coefficients for health?
#A: Since health is a dummy variable with three categories: Excellent, Average, and Poor. We splited health into two dummy variables: 1. poor or not (average), and 2. excellent or not (average). Then used these 2 dummy variables for regression. Therefore, we can find two names of dummy variables: healthpoor and healthexcellent.