A real estate appraiser is interested in predicting residential home prices in a mid-western city as a function of various features. For that purpose a regression model is to be constructed from a sample of 522 houses. Download the `homes.xls` data set from blackboard.

Consider the predictors
x_1: lot size (square feet), x_2: area (square feet), x_3: number of bedrooms,
x_4: number of bathrooms, x_5: year of construction, x_6: garage size (number of cars).

1. What are the predictors with the highest correlation?

2. What is the area (not lot size) of the most expensive house?

   Fit the full model.

3. If there are outliers find the largest one (in absolute value).

4. Find a 99% confidence interval for β_2

5. Find a 95% confidence interval for the mean price of a house with garage for two cars, area of 2650 square feet, built in 1990, 24500 square feet size, three bedrooms, three bathrooms

6. Find the predicted price when all predictors are equal to their median values.

   Fit the model with the best subset of predictors (in terms of adj-R^2).

7. Find the best and worst predictors

   Fit a model with only x_3, the number of bedrooms as the predictor

8. Interpret the slope value b_1.

   Fit a full model for houses having between two to four bedrooms

9. Interpret adequacy values (MSE, R^2).

10. Find a 95% prediction interval for the price of a house with a garage for two cars, area of 3150 square feet, built in 1996, 26250 square feet size, two bedrooms, three bathrooms.

```
library(MASS)     # stepAIC()
library(leaps)    # regsubsets()
library(PASWR2)   # checking.plots()

d0=read.csv("homes.csv",header=T)

# 1) correlation
#-------------------------------------------------------------

d1=d0[,c("price","lotsize","area","beds","baths","year","garage")]
cor(d1)
#             price    lotsize      area      beds     baths       year    garage
# price   1.0000000  0.2241685 0.8194701 0.4133239 0.6836854  0.5555164 0.5777863
# lotsize 0.2241685  1.0000000 0.1575247 0.1265384 0.1470066 -0.1004519 0.1522193
# area    0.8194701  0.1575247 1.0000000 0.5578378 0.7552729  0.4411967 0.5337665
# beds    0.4133239  0.1265384 0.5578378 1.0000000 0.5834469  0.2686924 0.3168137
# baths   0.6836854  0.1470066 0.7552729 0.5834469 1.0000000  0.5128410 0.4898981
# year    0.5555164 -0.1004519 0.4411967 0.2686924 0.5128410  1.0000000 0.4617604
# garage  0.5777863  0.1522193 0.5337665 0.3168137 0.4898981  0.4617604 1.0000000

# area and bath, most highly correlated

# 2) most expensive house
#-------------------------------------------------------------
which.max(d1$price)     #[1] 73
d1[73,]
#     price lotsize area beds baths year garage
# 73 920000   32793 3857    4     5 1997      3

# most expensive house has 3857 squared-feet (area)

# 3) full model
#-------------------------------------------------------------
m2=lm(price~.,d1)
checking.plots(m2)   #shows some large std residuals
res = rstandard(m2)
# largest residual
b=which.max(res)
d1[b,]
    price lotsize area beds baths year garage
73 920000   32793 3857    4     5 1997      3

# 4) CI on beta2
#-------------------------------------------------------------
confint(m2,level=0.99)
#                     0.5 %        99.5 %
#(Intercept) -4.649785e+06 -2.485632e+06
#lotsize      8.517208e-01  2.258260e+00
#area         1.077357e+02  1.437415e+02
```

```
# 5) CI
#-------------------------------------------------------------
newval=data.frame(garage=2,area=2650,year=1990,lotsize=24500,beds=3,baths=3)
predict(m2,newval,interval="conf")
#        fit      lwr      upr
# 1 374920.5 362128.4 387712.6

# 6) Prediction at the median
#-------------------------------------------------------------
apply(d1,2,median)
#  price lotsize    area    beds   baths    year  garage
# 229900   22200    2061       3       3    1966       2
newval=data.frame(garage=2,area=2061,year=1966,lotsize=22200,beds=3,baths=3)
predict(m2,newval)
# 254573.4

# 7) best subset predictors with adj-R2
#-------------------------------------------------------------
library(leaps)
m3 = regsubsets(price~.,data=d1)
summary(m3)         # table of selected predictors
# Selection Algorithm: exhaustive
#         lotsize area beds baths year garage
#1  ( 1 ) " "     "*"  " "  " "   " "  " "
#2  ( 1 ) " "     "*"  " "  " "   "*"  " "
#3  ( 1 ) "*"     "*"  " "  " "   "*"  " "
#4  ( 1 ) "*"     "*"  " "  " "   "*"  "*"
#5  ( 1 ) "*"     "*"  "*"  " "   "*"  "*"
#6  ( 1 ) "*"     "*"  "*"  "*"   "*"  "*"

# best predictor area, worst predictor baths

summary(m3)$adjr2
# [1] 0.6708995 0.7171632 0.7364263 0.7434685 0.7476413 0.7484930
# full model gives best adjr2

# 8) SLR model with bedrooms
#-------------------------------------------------------------
m1=lm(price~beds,d1)
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)    82809      19634   4.218 2.91e-05 ***
# beds           56200       5430  10.351  < 2e-16 ***

# Residual standard error: 125700 on 520 degrees of freedom
# Multiple R-squared:  0.1708,    Adjusted R-squared:  0.1692
# F-statistic: 107.1 on 1 and 520 DF,  p-value: < 2.2e-16

# mean price increases 56200 dollars with each additional bedroom
```

```
# 9) full model - 2 to 4 beds
#-------------------------------------------------------
table(d1$beds)
#  0   1   2   3   4   5   6   7
#  1   9  64 202 179  52  12   3

d2=d1[d1$beds<5,]
table(d2$beds)
#  0   1   2   3   4
#  1   9  64 202 179
d3=d2[d2$beds>1,]
table(d3$beds)
#  2   3   4
# 64 202 179

m3=lm(price~.,d3)
summary(m3)

# Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
# (Intercept) -3.241e+06  4.214e+05  -7.690 9.80e-14 ***
# lotsize      1.581e+00  2.839e-01   5.570 4.46e-08 ***
# area         1.327e+02  7.537e+00  17.613  < 2e-16 ***
# beds        -1.274e+04  5.078e+03  -2.509 0.012474 *
# baths        8.812e+03  5.049e+03   1.745 0.081622 .
# year         1.604e+03  2.171e+02   7.387 7.68e-13 ***
# garage       2.207e+04  5.949e+03   3.710 0.000234 ***

# Residual standard error: 65130 on 438 degrees of freedom
# Multiple R-squared:  0.7584,    Adjusted R-squared:  0.7551
# F-statistic: 229.1 on 6 and 438 DF,  p-value: < 2.2e-16

# Estimated regression variance is MSE = 65130^2 squared-dollars
# Average distance from all points to the fitted plane is S = 65130 dollars
# Model explains 75.5% of price variability

# 10) predict
#-------------------------------------------------------

newval=data.frame(garage=2,area=3150,year=1996,lotsize=26250,beds=2,baths=3)
predict(m3,newval,interval="pred")
#        fit      lwr      upr
# 1 465134.5 334969.9 595299.1
```