# CSC570 Machine Learning

**Homework 4**

**Linear Regression**

In this homework, you need to implement a linear regression model and answer the followed questions.

1. Load Boston.cvs data set. It records 14 variables for 506 neighborhoods around Boston:

crim: per capita crime rate by town.

zn: proportion of residential land zoned for lots over 25,000 sq.ft.

indus: proportion of non-retail business acres per town.

chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox: nitrogen oxides concentration (parts per 10 million).

rm: average number of rooms per dwelling.

age: proportion of owner-occupied units built prior to 1940.

dis: weighted mean of distances to five Boston employment centers.

rad: index of accessibility to radial highways. tax: full-value property-tax rate per \$10,000. ptratio: pupil-teacher ratio by town.

black: *1000(Bk - 0.63)^2* where *Bk* is the proportion of blacks by town.

lstat: lower status of the population (percent).

medv: median value of owner-occupied homes in \$1000s.

1. Fit a multiple regression model to predict the response variable medv using all other variables. For which variables the corresponding regression coefficients are likely to be significant? (40 points)

> cor(house[c("medv","crim", "zn", "indus", "chas", "nox", "rm", "age", "dis", "rad", "tax", "ptratio","black", "lstat")])

medv crim zn indus chas nox rm age dis

medv 1.0000000 -0.38830461 0.36044534 -0.48372516 0.175260177 -0.42732077 0.69535995 -0.37695457 0.24992873

crim -0.3883046 1.00000000 -0.20046922 0.40658341 -0.055891582 0.42097171 -0.21924670 0.35273425 -0.37967009

zn 0.3604453 -0.20046922 1.00000000 -0.53382819 -0.042696719 -0.51660371 0.31199059 -0.56953734 0.66440822

indus -0.4837252 0.40658341 -0.53382819 1.00000000 0.062938027 0.76365145 -0.39167585 0.64477851 -0.70802699

chas 0.1752602 -0.05589158 -0.04269672 0.06293803 1.000000000 0.09120281 0.09125123 0.08651777 -0.09917578

nox -0.4273208 0.42097171 -0.51660371 0.76365145 0.091202807 1.00000000 -0.30218819 0.73147010 -0.76923011

rm 0.6953599 -0.21924670 0.31199059 -0.39167585 0.091251225 -0.30218819 1.00000000 -0.24026493 0.20524621

age -0.3769546 0.35273425 -0.56953734 0.64477851 0.086517774 0.73147010 -0.24026493 1.00000000 -0.74788054

dis 0.2499287 -0.37967009 0.66440822 -0.70802699 -0.099175780 -0.76923011 0.20524621 -0.74788054 1.00000000

rad -0.3816262 0.62550515 -0.31194783 0.59512927 -0.007368241 0.61144056 -0.20984667 0.45602245 -0.49458793

tax -0.4685359 0.58276431 -0.31456332 0.72076018 -0.035586518 0.66802320 -0.29204783 0.50645559 -0.53443158

ptratio -0.5077867 0.28994558 -0.39167855 0.38324756 -0.121515174 0.18893268 -0.35550149 0.26151501 -0.23247054

black 0.3334608 -0.38506394 0.17552032 -0.35697654 0.048788485 -0.38005064 0.12806864 -0.27353398 0.29151167

lstat -0.7376627 0.45562148 -0.41299457 0.60379972 -0.053929298 0.59087892 -0.61380827 0.60233853 -0.49699583

rad tax ptratio black lstat

medv -0.381626231 -0.46853593 -0.5077867 0.33346082 -0.7376627

crim 0.625505145 0.58276431 0.2899456 -0.38506394 0.4556215

zn -0.311947826 -0.31456332 -0.3916785 0.17552032 -0.4129946

indus 0.595129275 0.72076018 0.3832476 -0.35697654 0.6037997

chas -0.007368241 -0.03558652 -0.1215152 0.04878848 -0.0539293

nox 0.611440563 0.66802320 0.1889327 -0.38005064 0.5908789

rm -0.209846668 -0.29204783 -0.3555015 0.12806864 -0.6138083

age 0.456022452 0.50645559 0.2615150 -0.27353398 0.6023385

dis -0.494587930 -0.53443158 -0.2324705 0.29151167 -0.4969958

rad 1.000000000 0.91022819 0.4647412 -0.44441282 0.4886763

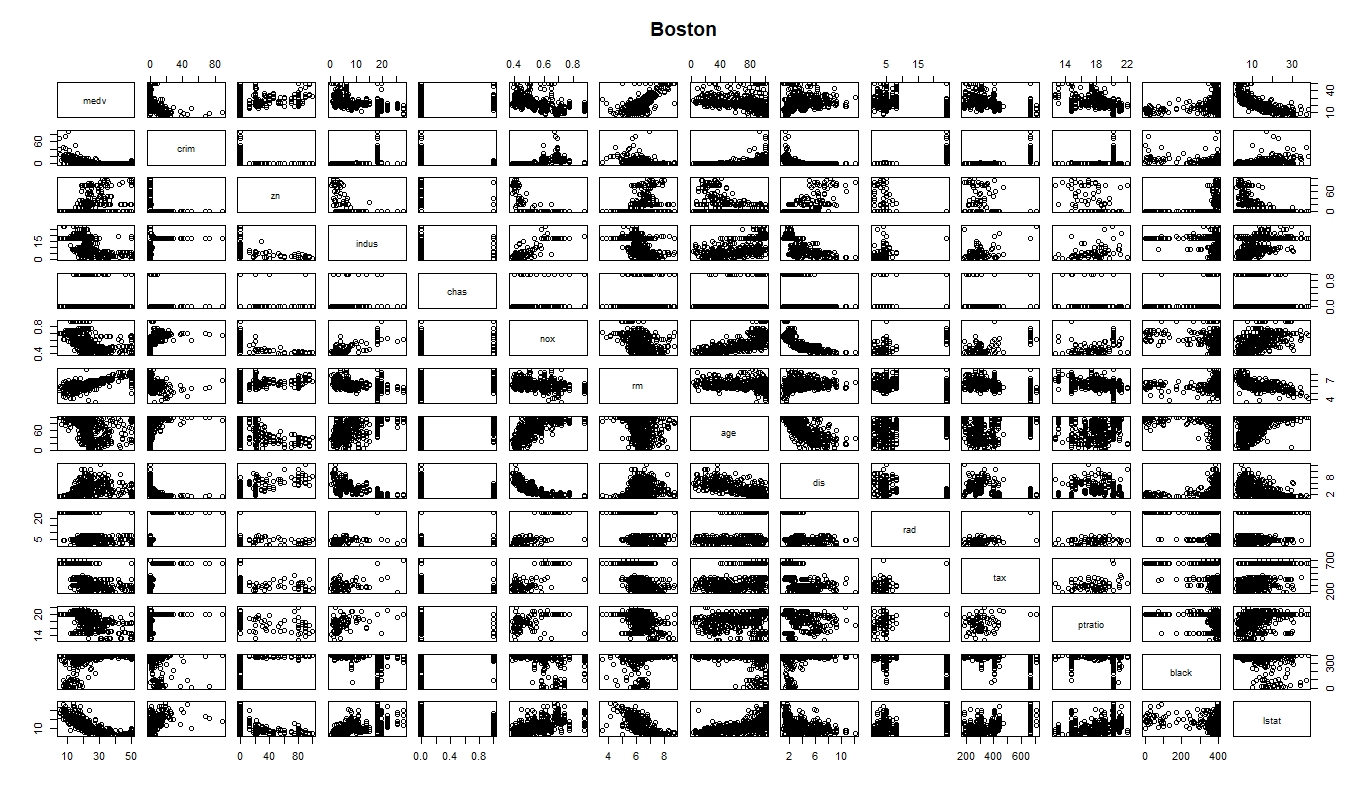
tax 0.910228189 1.00000000 0.4608530 -0.44180801 0.5439934

ptratio 0.464741179 0.46085304 1.0000000 -0.17738330 0.3740443

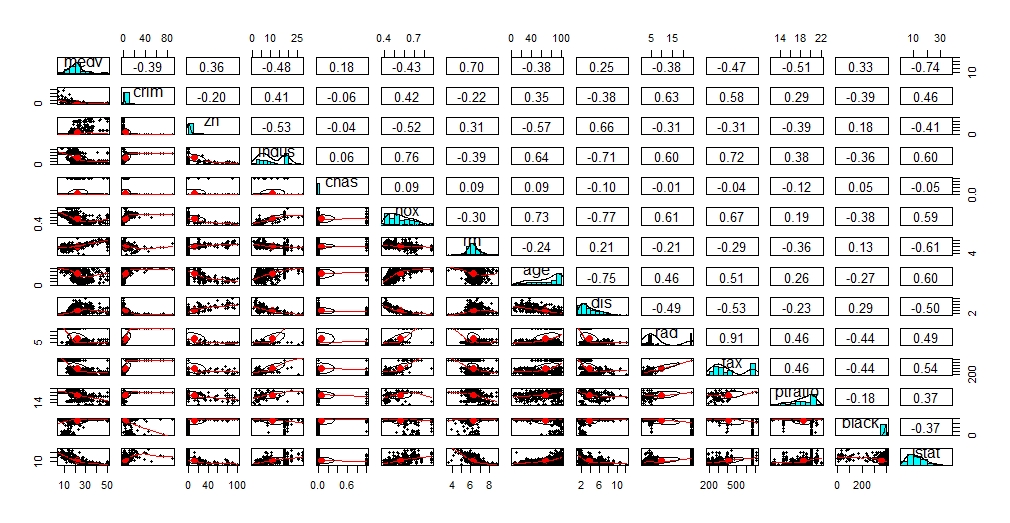
black -0.444412816 -0.44180801 -0.1773833 1.00000000 -0.3660869

lstat 0.488676335 0.54399341 0.3740443 -0.36608690 1.0000000

> pairs(~ medv + x + crim + zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio + black + lstat, data = house, main = "Boston")



> pairs.panels(house[c("medv","crim", "zn", "indus", "chas", "nox", "rm", "age", "dis", "rad", "tax", "ptratio","black", "lstat")])



> multi = lm(formula = medv ~ .,data = house)

> summary(multi)

Call:

lm(formula = medv ~ ., data = house)

Residuals:

Min 1Q Median 3Q Max

-15.8948 -2.7585 -0.4663 1.7963 26.0911

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 36.461352 5.100994 7.148 3.21e-12 \*\*\*

X -0.002526 0.002080 -1.215 0.225046

crim -0.108762 0.032855 -3.310 0.001000 \*\*

zn 0.048031 0.013785 3.484 0.000538 \*\*\*

indus 0.019932 0.061468 0.324 0.745871

chas 2.705245 0.861298 3.141 0.001786 \*\*

nox -17.541602 3.822390 -4.589 5.66e-06 \*\*\*

rm 3.839225 0.418422 9.175 < 2e-16 \*\*\*

age -0.001938 0.013380 -0.145 0.884866

dis -1.493304 0.199892 -7.471 3.68e-13 \*\*\*

rad 0.324925 0.068111 4.771 2.43e-06 \*\*\*

tax -0.011598 0.003807 -3.046 0.002443 \*\*

ptratio -0.947985 0.130822 -7.246 1.67e-12 \*\*\*

black 0.009357 0.002685 3.485 0.000536 \*\*\*

lstat -0.526184 0.050704 -10.377 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.743 on 491 degrees of freedom

Multiple R-squared: 0.7414, Adjusted R-squared: 0.734

F-statistic: 100.6 on 14 and 491 DF, p-value: < 2.2e-16

The variables Indus and age are not significant.

1. On the basis of your response to the previous question, fit a smaller linear regression model that only uses predictors for which there is evidence of association with medv. (10 points)

> cor(house[c("medv", "crim", "chas", "rm", "dis", "ptratio","black", "lstat")])

medv crim chas rm dis ptratio black lstat

medv 1.0000000 -0.38830461 0.17526018 0.69535995 0.24992873 -0.5077867 0.33346082 -0.7376627

crim -0.3883046 1.00000000 -0.05589158 -0.21924670 -0.37967009 0.2899456 -0.38506394 0.4556215

chas 0.1752602 -0.05589158 1.00000000 0.09125123 -0.09917578 -0.1215152 0.04878848 -0.0539293

rm 0.6953599 -0.21924670 0.09125123 1.00000000 0.20524621 -0.3555015 0.12806864 -0.6138083

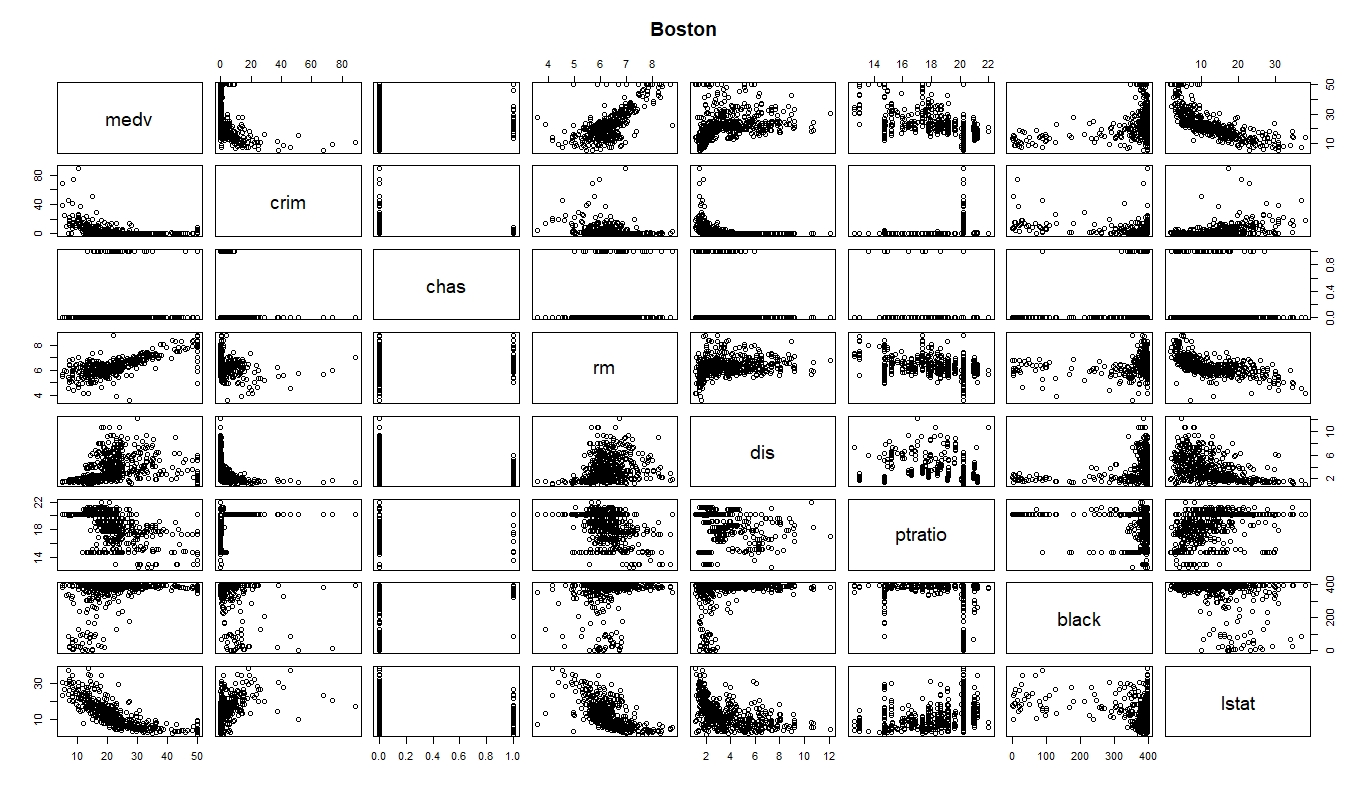
dis 0.2499287 -0.37967009 -0.09917578 0.20524621 1.00000000 -0.2324705 0.29151167 -0.4969958

ptratio -0.5077867 0.28994558 -0.12151517 -0.35550149 -0.23247054 1.0000000 -0.17738330 0.3740443

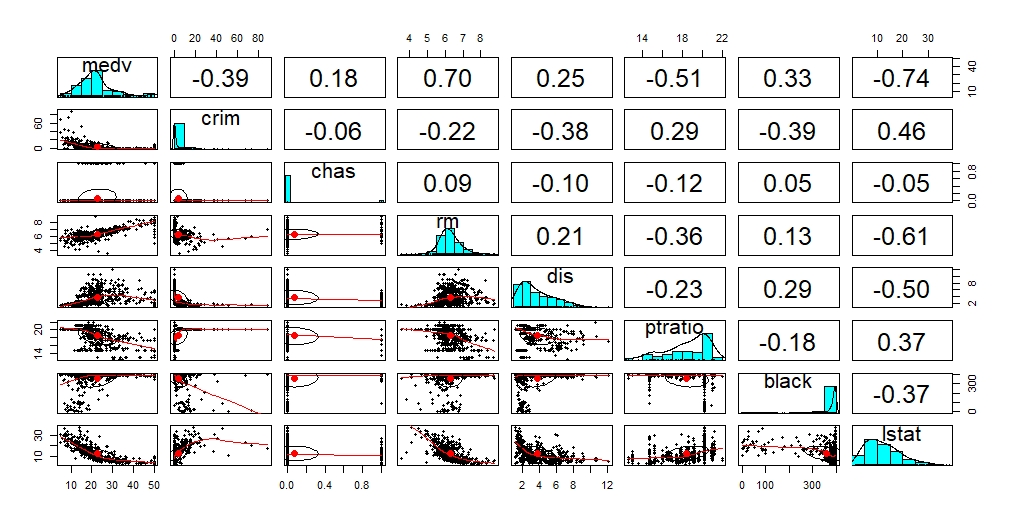
black 0.3334608 -0.38506394 0.04878848 0.12806864 0.29151167 -0.1773833 1.00000000 -0.3660869

lstat -0.7376627 0.45562148 -0.05392930 -0.61380827 -0.49699583 0.3740443 -0.36608690 1.0000000

> pairs(medv ~ crim + chas + rm + dis +ptratio + black + lstat, data = house, main = "Boston")



> pairs.panels(house[c("medv", "crim", "chas", "rm", "dis", "ptratio","black", "lstat")])



> predictors = lm(medv ~ crim + chas + rm + dis +ptratio + black + lstat, data=house)

> summary(predictors)

Call:

lm(formula = medv ~ crim + chas + rm + dis + ptratio + black +

lstat, data = house)

Residuals:

Min 1Q Median 3Q Max

-18.0908 -2.8068 -0.7533 1.5149 27.9661

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 16.698091 4.246147 3.933 9.60e-05 \*\*\*

crim -0.060019 0.030853 -1.945 0.052299 .

chas 2.646537 0.893817 2.961 0.003213 \*\*

rm 4.453120 0.415810 10.710 < 2e-16 \*\*\*

dis -0.596983 0.127166 -4.695 3.46e-06 \*\*\*

ptratio -0.874969 0.114598 -7.635 1.16e-13 \*\*\*

black 0.010302 0.002738 3.763 0.000188 \*\*\*

lstat -0.583148 0.048361 -12.058 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.99 on 498 degrees of freedom

Multiple R-squared: 0.7097, Adjusted R-squared: 0.7056

F-statistic: 173.9 on 7 and 498 DF, p-value: < 2.2e-16

1. How well the models in 2 and 3 fit the data? (10 points)

Model 2: Multiple R-squared: 0.7414, Adjusted R-squared: 0.734

*R*2 is a measure of the model’s quality. Bigger is better. Mathematically, it is the fraction of the variance of *y* that is explained by the regression model. The remaining variance is not explained by the model, so it must be due to other factors (i.e., unknown variables or sampling variability). In this case, the model explains 0.741409 (74.14%) of the variance of *y* and remaining is unexplanined.

Model 3 :Multiple R-squared: 0.7097, Adjusted R-squared: 0.7056

*R*2 is a measure of the model’s quality. Bigger is better. Mathematically, it is the fraction of the variance of *y* that is explained by the regression model. The remaining variance is not explained by the model, so it must be due to other factors (i.e., unknown variables or sampling variability). In this case, the model explains 0.7097 (740.97%) of the variance of *y* and remaining is unexplanined.

1. Try to reduce the set of predictors by making correlation and scatterplot matrices and fit even a smaller model using the predictors of your choice. (30 points)

rm, lstat,ptratio are the variables associated with medv

> cor(house[c("medv", "rm","ptratio","lstat")])

medv rm ptratio lstat

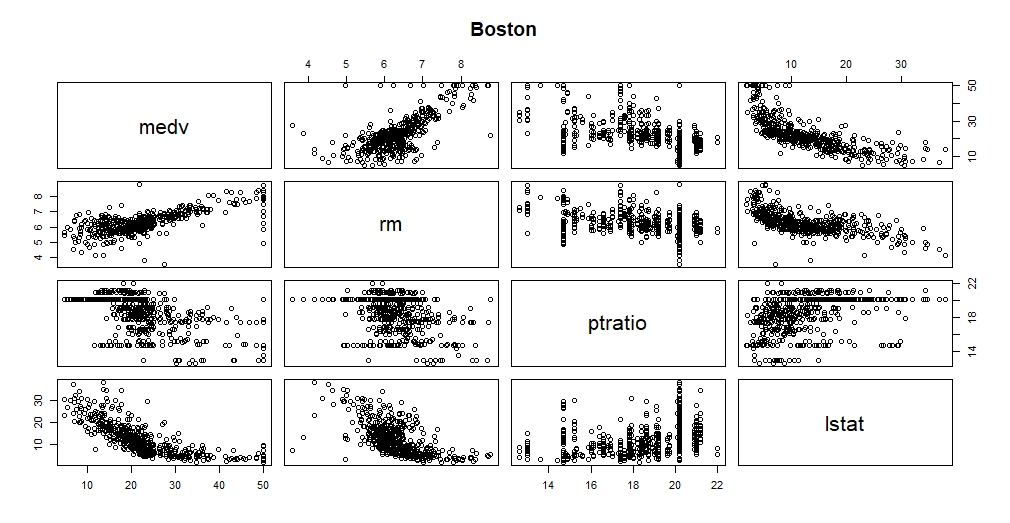
medv 1.0000000 0.6953599 -0.5077867 -0.7376627

rm 0.6953599 1.0000000 -0.3555015 -0.6138083

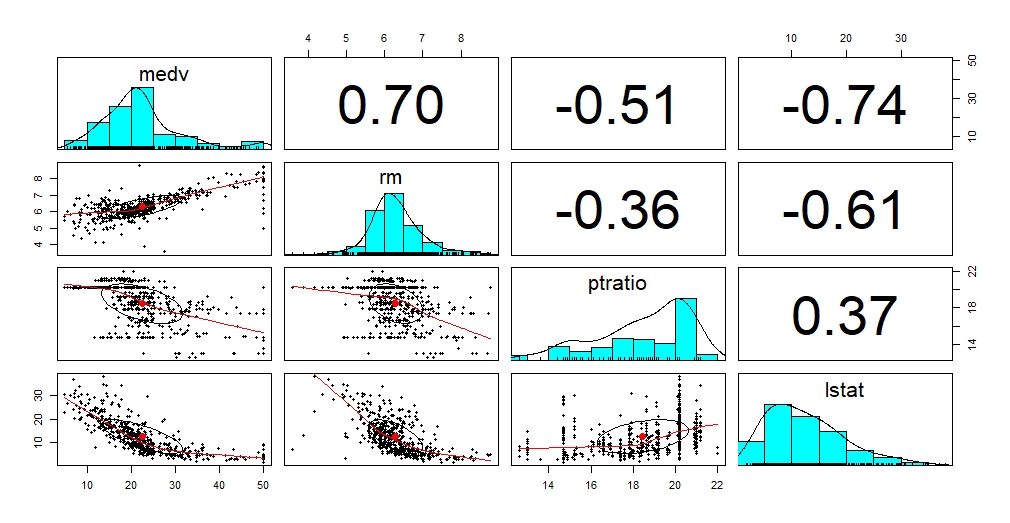
ptratio -0.5077867 -0.3555015 1.0000000 0.3740443

lstat -0.7376627 -0.6138083 0.3740443 1.0000000

> pairs(medv ~ rm + ptratio + lstat, data = house, main = "Boston")



> pairs.panels(house[c("medv", "rm","ptratio","lstat")])



> var1<-lm(medv~rm+lstat+ptratio, data = house)

> summary(var1)

Call:

lm(formula = medv ~ rm + lstat + ptratio, data = house)

Residuals:

Min 1Q Median 3Q Max

-14.4871 -3.1047 -0.7976 1.8129 29.6559

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 18.56711 3.91320 4.745 2.73e-06 \*\*\*

rm 4.51542 0.42587 10.603 < 2e-16 \*\*\*

lstat -0.57181 0.04223 -13.540 < 2e-16 \*\*\*

ptratio -0.93072 0.11765 -7.911 1.64e-14 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.229 on 502 degrees of freedom

Multiple R-squared: 0.6786, Adjusted R-squared: 0.6767

F-statistic: 353.3 on 3 and 502 DF, p-value: < 2.2e-16

1. Investigate possible interactions between the variables in the last model. Try to find a model with the maximum of R2 value (10 points)

> ind1=lm(medv~lstat,data=house)

> summary(ind1)

Call:

lm(formula = medv ~ lstat, data = house)

Residuals:

Min 1Q Median 3Q Max

-15.168 -3.990 -1.318 2.034 24.500

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 34.55384 0.56263 61.41 <2e-16 \*\*\*

lstat -0.95005 0.03873 -24.53 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.216 on 504 degrees of freedom

Multiple R-squared: 0.5441, Adjusted R-squared: 0.5432

F-statistic: 601.6 on 1 and 504 DF, p-value: < 2.2e-16

> ind3=lm(medv~rm,data=house)

> summary(ind3)

Call:

lm(formula = medv ~ rm, data = house)

Residuals:

Min 1Q Median 3Q Max

-23.346 -2.547 0.090 2.986 39.433

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -34.671 2.650 -13.08 <2e-16 \*\*\*

rm 9.102 0.419 21.72 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.616 on 504 degrees of freedom

Multiple R-squared: 0.4835, Adjusted R-squared: 0.4825

F-statistic: 471.8 on 1 and 504 DF, p-value: < 2.2e-16

> ind4=lm(medv~ptratio,data=house)

> summary(ind4)

Call:

lm(formula = medv ~ ptratio, data = house)

Residuals:

Min 1Q Median 3Q Max

-18.8342 -4.8262 -0.6426 3.1571 31.2303

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 62.345 3.029 20.58 <2e-16 \*\*\*

ptratio -2.157 0.163 -13.23 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.931 on 504 degrees of freedom

Multiple R-squared: 0.2578, Adjusted R-squared: 0.2564

F-statistic: 175.1 on 1 and 504 DF, p-value: < 2.2e-16

maximum of R2 value is lstat with 0.5441