Week 2 Homework

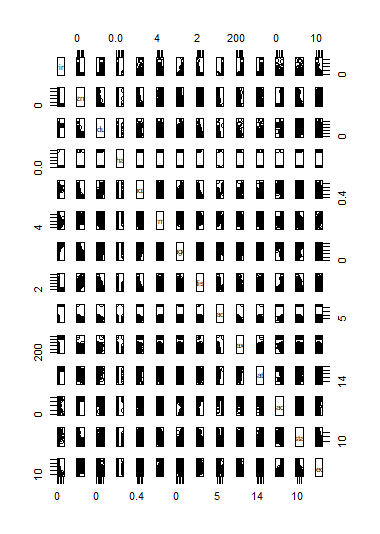
John Khwajazada

1. Describe your data set such as what is this data set about? How many observations? How many variables? etc.,What are those variables?

This data set contains the median house values for 506 neighborhoods around the Boston area. The variables are "crim" "zn" "indus" "chas" "nox" "rm" "age” "dis" "rad" "tax" "ptratio" "black" "lstat" "medv". There are 506 observations on 13 different variables.

1. Perform data exploration by plotting the graph(s), the distribution, and so on.  Then, interpret them.

> plot(Boston)



|  |
| --- |
| > summary(Boston)  crim zn indus  Min. : 0.00632 Min. : 0.00 Min. : 0.46  1st Qu.: 0.08204 1st Qu.: 0.00 1st Qu.: 5.19  Median : 0.25651 Median : 0.00 Median : 9.69  Mean : 3.61352 Mean : 11.36 Mean :11.14  3rd Qu.: 3.67708 3rd Qu.: 12.50 3rd Qu.:18.10  Max. :88.97620 Max. :100.00 Max. :27.74  chas nox rm  Min. :0.00000 Min. :0.3850 Min. :3.561  1st Qu.:0.00000 1st Qu.:0.4490 1st Qu.:5.886  Median :0.00000 Median :0.5380 Median :6.208  Mean :0.06917 Mean :0.5547 Mean :6.285  3rd Qu.:0.00000 3rd Qu.:0.6240 3rd Qu.:6.623  Max. :1.00000 Max. :0.8710 Max. :8.780  age dis rad  Min. : 2.90 Min. : 1.130 Min. : 1.000  1st Qu.: 45.02 1st Qu.: 2.100 1st Qu.: 4.000  Median : 77.50 Median : 3.207 Median : 5.000  Mean : 68.57 Mean : 3.795 Mean : 9.549  3rd Qu.: 94.08 3rd Qu.: 5.188 3rd Qu.:24.000  Max. :100.00 Max. :12.127 Max. :24.000  tax ptratio black  Min. :187.0 Min. :12.60 Min. : 0.32  1st Qu.:279.0 1st Qu.:17.40 1st Qu.:375.38  Median :330.0 Median :19.05 Median :391.44  Mean :408.2 Mean :18.46 Mean :356.67  3rd Qu.:666.0 3rd Qu.:20.20 3rd Qu.:396.23  Max. :711.0 Max. :22.00 Max. :396.90  lstat medv  Min. : 1.73 Min. : 5.00  1st Qu.: 6.95 1st Qu.:17.02  Median :11.36 Median :21.20  Mean :12.65 Mean :22.53  3rd Qu.:16.95 3rd Qu.:25.00  Max. :37.97 Max. :50.00 |
|  |
| |  | | --- | | > | |

3) Perform pairwise scatterplots and describe your findings, such as which pairs show positive correlation, negative correlation, or no correlation?

> cor(Boston)

crim zn indus chas

crim 1.00000000 -0.20046922 0.40658341 -0.055891582

zn -0.20046922 1.00000000 -0.53382819 -0.042696719

indus 0.40658341 -0.53382819 1.00000000 0.062938027

chas -0.05589158 -0.04269672 0.06293803 1.000000000

nox 0.42097171 -0.51660371 0.76365145 0.091202807

rm -0.21924670 0.31199059 -0.39167585 0.091251225

age 0.35273425 -0.56953734 0.64477851 0.086517774

dis -0.37967009 0.66440822 -0.70802699 -0.099175780

rad 0.62550515 -0.31194783 0.59512927 -0.007368241

tax 0.58276431 -0.31456332 0.72076018 -0.035586518

ptratio 0.28994558 -0.39167855 0.38324756 -0.121515174

black -0.38506394 0.17552032 -0.35697654 0.048788485

lstat 0.45562148 -0.41299457 0.60379972 -0.053929298

medv -0.38830461 0.36044534 -0.48372516 0.175260177

nox rm age dis

crim 0.42097171 -0.21924670 0.35273425 -0.37967009

zn -0.51660371 0.31199059 -0.56953734 0.66440822

indus 0.76365145 -0.39167585 0.64477851 -0.70802699

chas 0.09120281 0.09125123 0.08651777 -0.09917578

nox 1.00000000 -0.30218819 0.73147010 -0.76923011

rm -0.30218819 1.00000000 -0.24026493 0.20524621

age 0.73147010 -0.24026493 1.00000000 -0.74788054

dis -0.76923011 0.20524621 -0.74788054 1.00000000

rad 0.61144056 -0.20984667 0.45602245 -0.49458793

tax 0.66802320 -0.29204783 0.50645559 -0.53443158

ptratio 0.18893268 -0.35550149 0.26151501 -0.23247054

black -0.38005064 0.12806864 -0.27353398 0.29151167

lstat 0.59087892 -0.61380827 0.60233853 -0.49699583

medv -0.42732077 0.69535995 -0.37695457 0.24992873

rad tax ptratio black

crim 0.625505145 0.58276431 0.2899456 -0.38506394

zn -0.311947826 -0.31456332 -0.3916785 0.17552032

indus 0.595129275 0.72076018 0.3832476 -0.35697654

chas -0.007368241 -0.03558652 -0.1215152 0.04878848

nox 0.611440563 0.66802320 0.1889327 -0.38005064

rm -0.209846668 -0.29204783 -0.3555015 0.12806864

age 0.456022452 0.50645559 0.2615150 -0.27353398

dis -0.494587930 -0.53443158 -0.2324705 0.29151167

rad 1.000000000 0.91022819 0.4647412 -0.44441282

tax 0.910228189 1.00000000 0.4608530 -0.44180801

ptratio 0.464741179 0.46085304 1.0000000 -0.17738330

black -0.444412816 -0.44180801 -0.1773833 1.00000000

lstat 0.488676335 0.54399341 0.3740443 -0.36608690

medv -0.381626231 -0.46853593 -0.5077867 0.33346082

lstat medv

crim 0.4556215 -0.3883046

zn -0.4129946 0.3604453

indus 0.6037997 -0.4837252

chas -0.0539293 0.1752602

nox 0.5908789 -0.4273208

rm -0.6138083 0.6953599

age 0.6023385 -0.3769546

dis -0.4969958 0.2499287

rad 0.4886763 -0.3816262

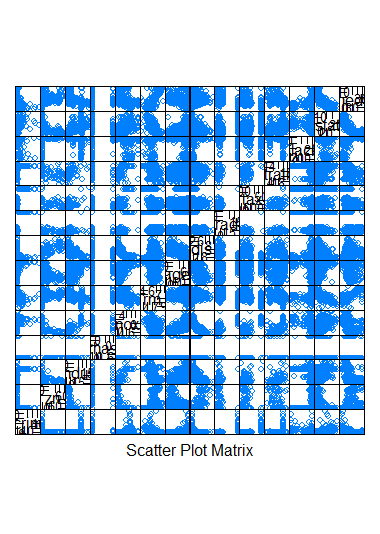
tax 0.5439934 -0.4685359

ptratio 0.3740443 -0.5077867

black -0.3660869 0.3334608

lstat 1.0000000 -0.7376627

medv -0.7376627 1.0000000



1. For this assignment, select one independent (a predictor or X variable) and one dependent variable (a response or Y variable). Explain the objective of selecting these variables. Do the scatter plot of the chosen variables.

|  |
| --- |
| > TestA <- subset(Boston, select = c(age,crim))  > summary(TestA)  age crim  Min. : 2.90 Min. : 0.00632  1st Qu.: 45.02 1st Qu.: 0.08204  Median : 77.50 Median : 0.25651  Mean : 68.57 Mean : 3.61352  3rd Qu.: 94.08 3rd Qu.: 3.67708  Max. :100.00 Max. :88.97620 |
| > confint(lm.fit)  2.5 % 97.5 %  (Intercept) 33.448457 35.6592247  lstat -1.026148 -0.8739505 |
| |  | | --- | |  | |
|  |
|  |
| |  | | --- | | > | |

 Now, perform linear regression on variables that you chose from step 4 using 5% significance level (or 95% confidence level).

> summary (lm(medv∼crim\*age ,data=Boston))

Call:

lm(formula = medv ~ crim \* age, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-13.957 -5.090 -2.411 2.316 31.622

Coefficients:

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

(Intercept) 30.086562 0.984255 30.568 < 2e-16 \*\*\*

crim -1.160123 0.512743 -2.263 0.0241 \*

age -0.092348 0.013859 -6.663 7.05e-11 \*\*\*

crim:age 0.008921 0.005372 1.661 0.0974 .

---

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

Residual standard error: 8.143 on 502 degrees of freedom

Multiple R-squared: 0.2208, Adjusted R-squared: 0.2162

F-statistic: 47.43 on 3 and 502 DF, p-value: < 2.2e-16

5) State both the null and the alternative hypothesis:

    H0: Age does not affect crime rates in Boston

    H1: Age does affect crime rates in Boston

6) Show the linear regression command in R and the corresponding results.

> lm.fit=lm(medv∼crim+age ,data=Boston )

> summary(lm.fit)

Call:

lm(formula = medv ~ crim + age, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-13.940 -4.991 -2.420 2.110 32.033

Coefficients:

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

(Intercept) 29.80067 0.97078 30.698 < 2e-16 \*\*\*

crim -0.31182 0.04510 -6.914 1.43e-11 \*\*\*

age -0.08955 0.01378 -6.499 1.95e-10 \*\*\*

---

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

Residual standard error: 8.157 on 503 degrees of freedom

Multiple R-squared: 0.2166, Adjusted R-squared: 0.2134

F-statistic: 69.52 on 2 and 503 DF, p-value: < 2.2e-16

7) Explain how do you read the result. Do you accept or reject the null hypothesis? Why? Discuss.

With this the p-vaule is less than 0.5 so we reject the null hypothesis and state that there is a correlation between age and crime rates in Boston.

8) Is there any evidence of a linear relationship between the predictor and response variable that you chose? Explain.

> cor(crim,age)

[1] 0.3527343

With a correlation of 0.3527343 there is a weak positive correlation between crime and age.

9) Verify the model assumptions (e.g. linearity, normality, variance) and specify which assumptions do not hold, if there is any

The assumption is that there is an additive relationship between the predictors and the responses. We also assumed that there is a normal distribution between the data points. We also do not know if the errors are normally distributed in this data set or not.

10) Provide insights/findings from the exercise. Address any other concerns you might have.

I was able to see that there is a correlation between age and crime. The concerns that I have had to do with the fact that we have no way of knowing if our assumptions are correct or not. This data set is good for a test example but would not be good to try and prove something without knowing more about how the data was gathered.