Linear regression

Benjamin Lacroix

18/02/2020

1 Introduction

1.1 Objectives

- Understand basic concepts of supervised learning techniques in R
- Use simple and multiple linear regression models in R
- Interpret simple and multiple linear regression model in R

1.2 Prerequiste

Libraries needed:

- stats : the stats package is loaded in R by default.
- corrplot : to visualise correlation matrix.
- MASS :

```
library(corrplot)
```

```
library(MASS) # Where the Boston dataset is
data(Boston)
```

The Boston, records medv (median house value) for 506 neighborhoods around Boston. It has 13 predictors including 12 numeric variables and one qualitative one 'chas'. 'chas' is actually encoded as a number (0 and 1).

```
# Have a look at the description of the dataset
?Boston
```

2 Simple linear regression

2.1 First look

First analyse the correlation between variable pairs. you can either

- visualise the pairwise correlation

```
pairs(Boston)
```

- calculate the correlation between variables

```
corrplot(cor(Boston))
```

Let's see if we can fit a linear model to predict the median house value (mdev) from the lower status of the population (lstat). We use the lm() command which fits linear models to data. It requires two arguments:

- 'formula' uses the format 'Y ~ X' which indicates the response variable Y and the predictor variable X.
- 'data' is the data frame used.

```
lm.medv.lstat <- lm(formula = medv ~ lstat, data = Boston)
```

We can analyse the results:

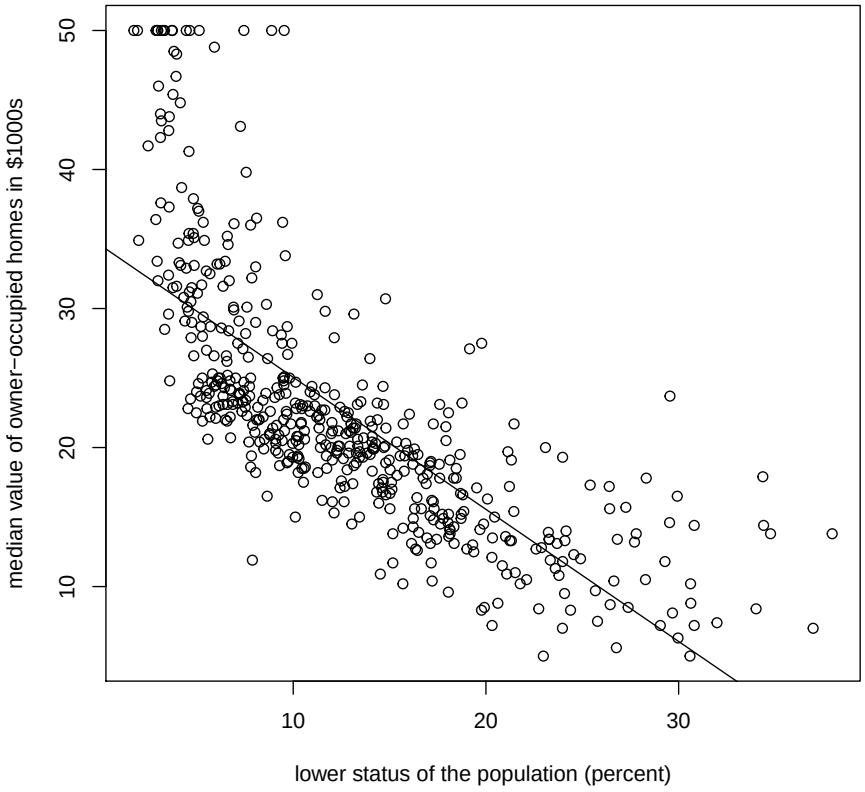
```
summary(lm.medv.lstat)

##
## Call:
## lm(formula = medv ~ lstat, data = Boston)
##
## 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
```

Here, we can see for instance the lower status of the population explains 54.44% of the house value (multiple $R^2 = 0.5441$). We can also analyse the coeficients of the model. The model obtained is of the form medv = -0.95005*lstat + 34.55384

And visualise:

```
plot(x= Boston$lstat, y = Boston$medv, xlab = "lower status of the population (percent)", yl
abline(lm.medv.lstat)
```



2.2 Your turn

- Use every other variables to predict the medv.
- Identify the best and the worst variables to predict medv. Check out particularly the multiple R^2 and the residual error.
- What can you learn from the coefficient obtained for each variables.

3 Multiple linear regression

3.1 Formulas in R

Formulas in R are used in supervised learning to represent the independant and dependant variables (predictors and response). For multiple regression models, formulas take the form of '$Y \sim X_1 + X_2 ... + X_n$'. You can also use different operators such as the dot '$Y \sim .$' which indicates every variables in your data. The minus symbol - removes the following variable from the set of seleced variables. For instance, '$Y \sim . - X_1$' means every variable BUT X_1.

3.2 First look

Here, we will try to predict the medium house value using every the other variables. Note here that in the formula, you can select every variable in your data set by usin '.' for the predictors

```
lm.medv.all <- lm(medv ~ . , Boston)
summary(lm.medv.all)

##
## Call:
## lm(formula = medv ~ ., data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## chas         2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## black        9.312e-03  2.686e-03   3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406,Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

We can see that 74.06% of house value can be explained by the other variables in the data set. The t-test for each variables indicates for each variable if it makes a significant contribution to explain medv. Here.

Task :

- Identify the variables that do not have a significant contribution to the prediction of the house value.
- rewrite the formula and fit a linear model using only the variables that make a significant contribution. How do the results compare with using all the variables?

4 Your turn

Load the 'Carseat' dataset that can be found in the ISLR library:

```
library(ISLR)
data("Carseats")
```

In this data set, we aim at predicting the number of child car seat sold in each store based on various indicator. Have a look at the dataset and repeat the process of the previous section.