PS 6: Linear Regression

Pooja Sadarangani

2022-11-19

Collaborator: Pranali Oza

#1 Housing Values in Boston

1. Describe the data and variables that are part of the Boston dataset. Are there any missings? Any unreasonable values? Clean data as necessary.

```
## [1] 506  14
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

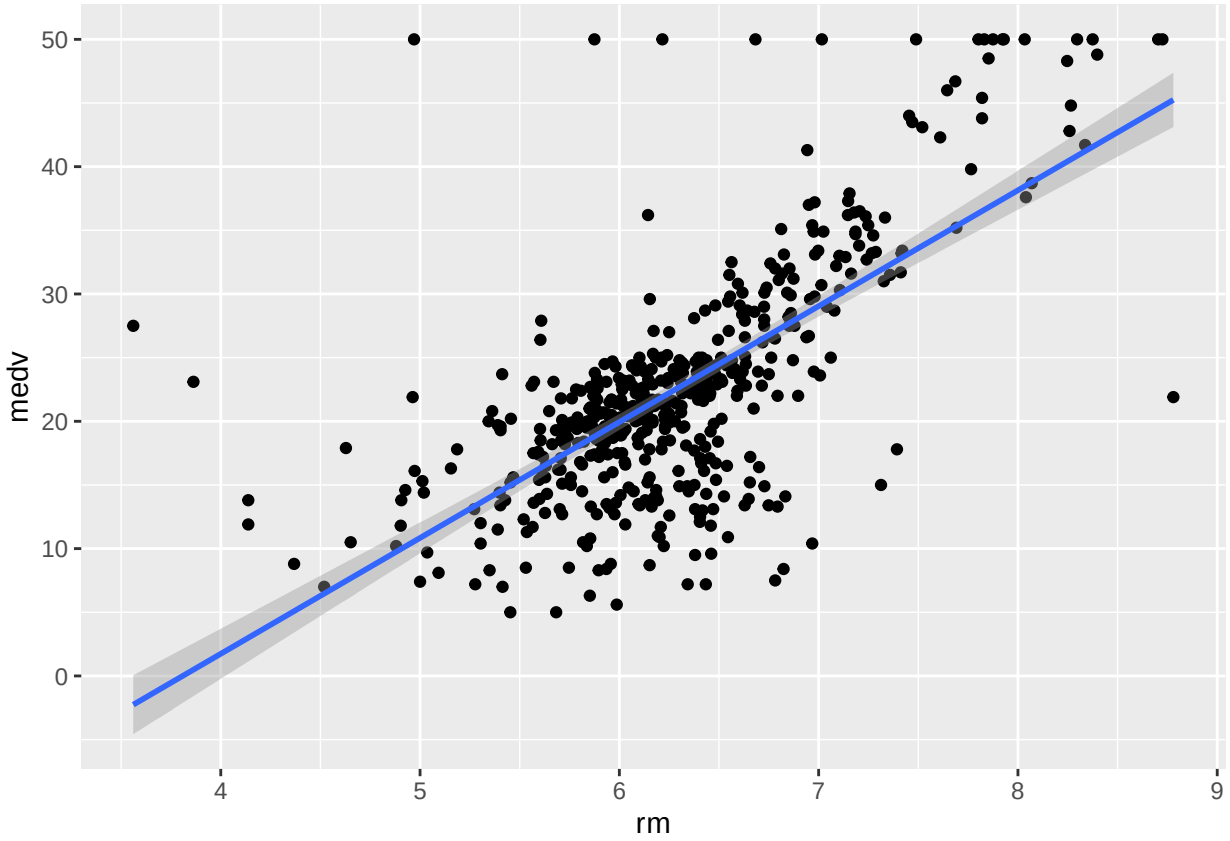
```
## [1] "There are no missing values in the dataset"
```

2. Use the following predictors: rm, lstat, and add an additional predictors of your choice, something that you consider might be interesting to analyze. For each predictor do the following:

a. Make a scatterplot that displays how medv is related to that predictor and add regression line to that plot. Comment the result: do you see any relationship? Anything else interesting you see?

Scatterplot for predictor = rm:

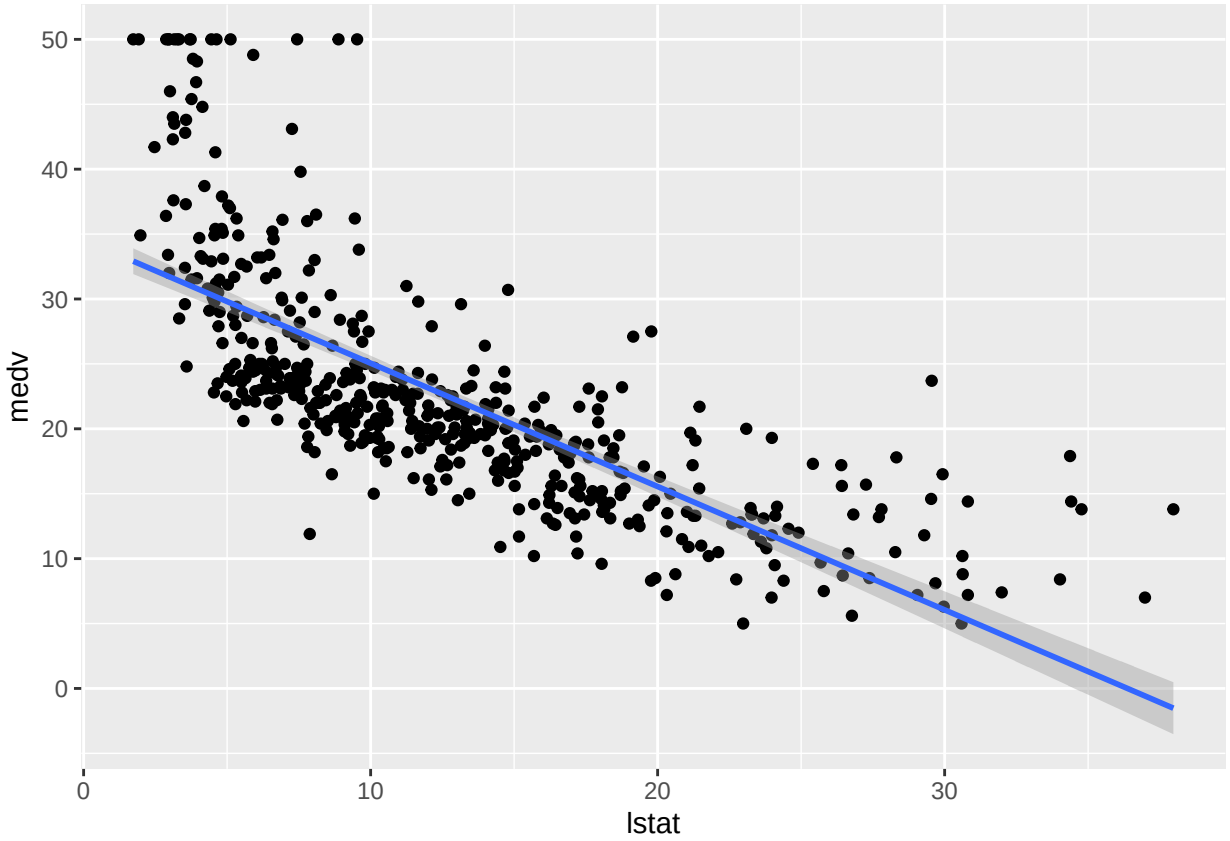
```
## `geom_smooth()` using formula 'y ~ x'
```



Ans. The variable medv is positively and strongly correlated to the variable rm. That means that the value of medv increases with the increase in value of rm.

Scatterplot for predictor = lstat:

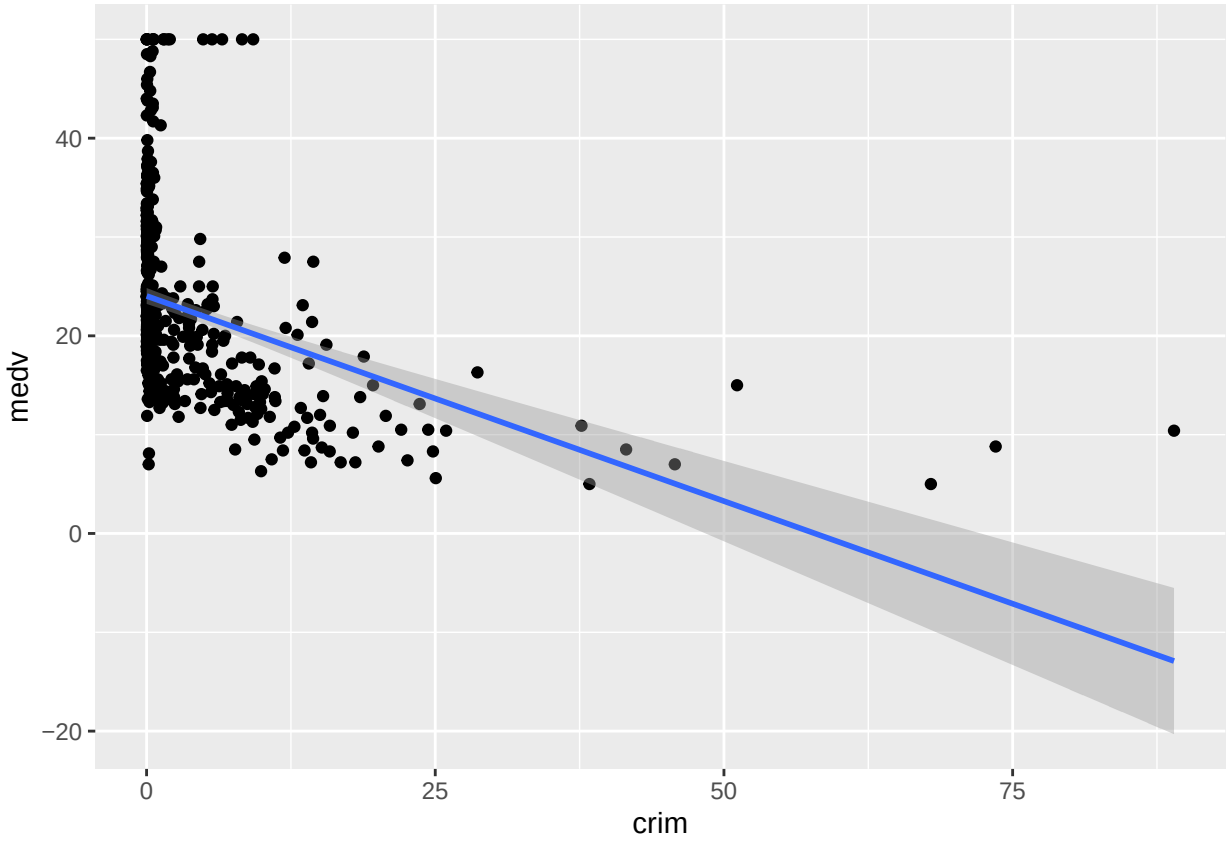
```
## `geom_smooth()` using formula 'y ~ x'
```



Ans. The variable medv is negatively and strongly correlated to the variable lstat. That means that the value medv increases with the decrease in value of lstat.

Scatterplot for predictor = crim:

```
## `geom_smooth()` using formula 'y ~ x'
```



Ans. The variable medv is negatively correlated to the variable crim. That means that the value medv increases with the decrease in value of crim.

b. Fit a simple linear regression model to predict the response. Show the regression output.

Model for predictor = rm:

```
##
## Call:
## lm(formula = medv ~ rm, data = boston)
##
## 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
```

Model for predictor = 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
```

Model for predictor = crim:

```
##
## Call:
## lm(formula = medv ~ crim, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.957  -5.449  -2.007   2.512  29.800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.03311    0.40914   58.74   <2e-16 ***
## crim        -0.41519    0.04389   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.484 on 504 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491
## F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

c. Interpret the slope (the effect of your explanatory variable). Is it statistically significant?

For predictor = rm

Ans: The slope is 9.102. This indicates that the two variables are positively correlated which means medv increases as rm increases. It is statistically significant as the p-value is very small (less than zero) i.e 2e-16.

For predictor = lstat

Ans: The slope is -0.95. This indicates that the two variables are negatively correlated which means medv increases as lstat decreases. It is statistically significant as the p-value is very small (less than zero) i.e 2e-16.

For predictor = crim

Ans: The slope is -0.4152. This indicates that the two variables are negatively correlated which means medv increases as crim decreases. It is statistically significant as the p-value is very small (less than zero) i.e 2e-16.

d. Explain why do you think you see (or don't see) the relationship on the figure/- model. Try to think about the possible social processes that make certain neighborhoods more or less expensive. For instance, why do you see that neighborhoods with more lower status people have lower house prices.

For predictor = rm (average number of rooms per dwelling)

As the number of rooms increases, the house price should also increase as it means that the house is bigger and can accomodate more people. With this logic, the relationship determined between medv and rm above seems correct.

For predictor = lstat (lower status of the population (percent))

The neighbourhoods with more lower status people have lower house prices because people will buy the house only if they can afford it. This suggests that houses prices in those regions must be affordable. With this logic, the relationship determined between medv and lstat above seems correct.

For predictor = crim (per capita crime rate by neighborhood)

People will tend to live in areas which have a lower crime rate. Thus, the houses in these areas will be pricey due to the high demand. With this logic, the relationship determined between medv and crim above seems correct.

3. Comment the results: are plots where you clearly can see a relationship related to models where the effect is statistically significant?

Ans. Yes, the plots are related to the models. We are able to deduct the same relationship between the predictor and the response outcome from both.

4. If you do this correctly, you find t-value for rm to be 21.72. How is the t-value computed? What is it testing–what is H0 here? What would be the critical t-value here? Enough of simple regression. Now let's move to multiple regression.

```
## [1] 1.964682
```

```
## function (p, df, ncp, lower.tail = TRUE, log.p = FALSE)
## {
##     if (missing(ncp))
##         .Call(C_qt, p, df, lower.tail, log.p)
##     else .Call(C_qnt, p, df, ncp, lower.tail, log.p)
## }
## <bytecode: 0x1220ff008>
## <environment: namespace:stats>
```

The t-value for the coefficient rm is computed by dividing the estimate of the parameter rm by the std. error of the parameter rm. The t-value is testing the H0 (null hypothesis), which in this case is that the outcome variable and the predictor are not correlated. Using the qt function, we get the critical t value as 1.964682.

5. Fit a multiple regression model to predict the response using all the predictors. Describe your results. For which predictors can we reject the null hypothesis H0 :Bj =0

```
##
## Call:
## lm(formula = medv ~ rm + lstat + crim + indus + zn + chas + nox +
##     age + dis + rad + tax + ptratio + black, 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 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## indus        2.056e-02  6.150e-02   0.334 0.738288
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## chas         2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## 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 ***
## ---
## 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
```

Ans. According to the results, all parameters except 'age' and 'indus' are statistically significant. Thus,we can reject the null hypothesis for the following predictors: crim, zn, chas, nox, rm, dis, rad, tax, ptratio, black, and lstat.

6. Interpret the results for rm, lstat and indus. Are the results statistically significant? Here just statistical interpretation is enough.

Ans. According to the above results, the parameters rm and lstat are statistically significant while the parameter indus is not statistically significant. We can tell this by looking at the p-values.

7. How do your results from 2 compare to your results from 5? Compare the results for those predictors you used for simple regression above. Explain why do the values differ. Do they still tell the same basic story?

The results tell the same basic story; medv is positively correlated to rm and negatively correlated to lstat and crim.

The results from 2 differ with the results from 5 for the parameter 'crim' as the p value is greater than what it was before.

2 Interpret Regression Results

1. Do neighborhoods with more evictions see more or less 311 calls? By how much?

Ans. The number of 311 calls is positively correlated with the number of evictions.Thus, neighborhoods with more evictions see more number of 311 calls. The number of 311 calls increases by 0.048 with every unit increase in the number of evictions.

2. Is the figure statistically significant (at 5% level)?

Ans. No, the figure is not statistically significant at 5% level but instead, it is statistically significant at 1% level as $p<.01$

3. How is poverty rate associated with 311 calls? How much more (or less) calls there are in neighborhoods with 10 pct point more poverty?

Ans. The number of 311 calls are negatively correlated to poverty rate. Thus,higher the poverty rate, lesser the 311 calls. From the coefficient that we get from the table i.e -0.14, we can tell that every pct point increase in poverty rate causes decrease in number of 311 calls by 0.14. Thus, when poverty increases by 10 pct point, the number of 311 calls decreases by 1.4.

4. What can you tell about association of race (white) and calls?

Ans. The number of 311 calls in negatively correlated to the % of whites. Thus, more the number of whites in the neighborhood, lesser the 311 calls.

5. Is older median age associated with more or less 311 calls? At which level is this statistically significant?

Ans. Older median age is associated with more calls as they are positively correlated which is clear from the positive coefficient i.e 0.0067 . This is statistically significant at level 1% as $p<.01$.

6. The value for housing density is -0.13. What does thin number mean?

Ans. The value -0.13 is the regression coefficient (coefficient B0 in the regression line function y= B0 + B1x). It denotes that the number of calls to 311 is