ECO204 (Section 6) - R lab

Simple Linear Regression

Tanvir

2023-11-20

First we clear the environment

```r
rm(list = ls())
```

1 About the Data and Summary Measures

We will do some simple linear regression analysis using a data set in the file `Boston.xlsx`. This data set has information about housing values and other information about Boston census tracts. Let's first load the data set. Be careful with the `setwd()` , adjust this with your own directory path.

```r
setwd("/home/tanvir/Documents/ownCloud/Git_Repos/EWU_repos/3_Fall_2023/eco_204/ewu-eco204.githu

# load the library for reading the excel file
library(readxl)

# load the data set
boston <- read_excel("Boston.xlsx")
```

We can view the data set just by clicking in the environment or by the following command

```r
boston
```

```
## # A tibble: 506 x 13
##       crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio lstat
##      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl>
##  1 0.00632  18    2.31     0 0.538  6.58  65.2  4.09     1   296    15.3  4.98
##  2 0.0273    0    7.07     0 0.469  6.42  78.9  4.97     2   242    17.8  9.14
##  3 0.0273    0    7.07     0 0.469  7.18  61.1  4.97     2   242    17.8  4.03
##  4 0.0324    0    2.18     0 0.458  7.00  45.8  6.06     3   222    18.7  2.94
##  5 0.0690    0    2.18     0 0.458  7.15  54.2  6.06     3   222    18.7  5.33
##  6 0.0298    0    2.18     0 0.458  6.43  58.7  6.06     3   222    18.7  5.21
##  7 0.0883   12.5  7.87     0 0.524  6.01  66.6  5.56     5   311    15.2 12.4
##  8 0.145    12.5  7.87     0 0.524  6.17  96.1  5.95     5   311    15.2 19.2
##  9 0.211    12.5  7.87     0 0.524  5.63 100    6.08     5   311    15.2 29.9
## 10 0.170    12.5  7.87     0 0.524  6.00  85.9  6.59     5   311    15.2 17.1
## # i 496 more rows
## # i 1 more variable: medv <dbl>
```

So we have 13 variables and the sample size is 506. Here are some details about the variables.

- `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 centres.
- `rad` index of accessibility to radial highways.
- `tax` full-value property-tax rate per $10,000.
- `ptratio` pupil-teacher ratio by town.
- `lstat` lower status of the population (percent).
- `medv` median value of owner-occupied homes in $1000s.

Let's see some summary statistics, this is simple to see with the `summary()` function in R

```r
summary(boston)
```

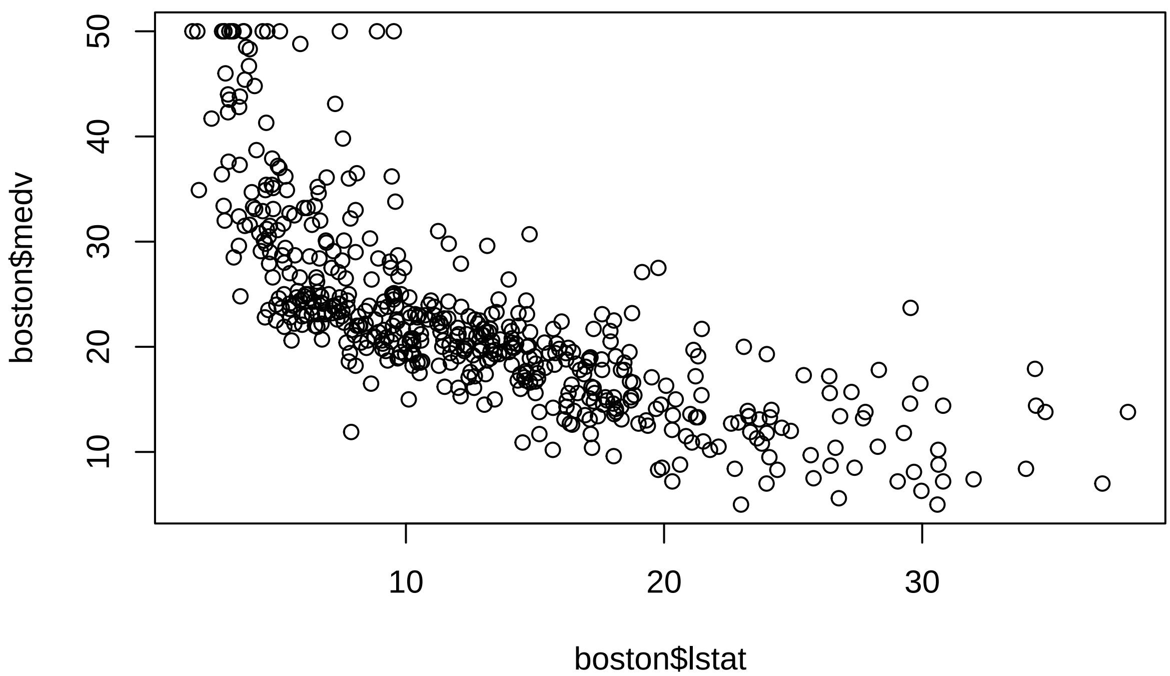
```
##       crim                zn             indus            chas
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##       nox               rm             age              dis
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       rad              tax           ptratio          lstat
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   : 1.73
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.: 6.95
##  Median : 5.000   Median :330.0   Median :19.05   Median :11.36
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :12.65
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:16.95
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :37.97
##       medv
##  Min.   : 5.00
```

```
##   1st Qu.:17.02
##   Median :21.20
##   Mean   :22.53
##   3rd Qu.:25.00
##   Max.   :50.00
```

2 Running Simple Linear Regression (SLR)

Now we will start with the simple linear regression modeling. Our goal is to predict `medv` using `lstat`. This means we want to predict the median value of the homes (in 1000$) using the percentage of the people who are in the lower status. You can expect that there should be a negative association between `lstat` and `medv`. Let's see this with the scatter plot,

```r
plot(boston$lstat, boston$medv)
```



so our guess is correct. We can also check the sample correlation

```r
cor(boston$lstat, boston$medv)
```

```
## [1] -0.7376627
```

the sample correlation shows high negative correlation in the data. Now let's fit a regression line. Fit we will fit the line using `lm()` function. Always remember the syntax is `lm(dependent variable ~ indepndent variable)`. Then we will save the output of this function as an object in R. It's important that here variables don't have space in their labels.

```r
model_fit <- lm(medv ~ lstat, data = boston)
```

3 Getting the results

Now everything that we need from the regression results are hidden in the object `model_fit`. We can see the summary of the regression results using the `summary()` function.

```r
summary(model_fit)
```

```
##
## 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
```

3.1 Estimated Coefficients, Equation of the Estimated Regression Line and R^2

The output of the `summary()` function is very long. So let's break it down one by one, first note that the the estimated regression coefficients are

$$\hat{\beta}_0 = 34.55 \tag{1}$$
$$\hat{\beta}_1 = -0.95 \tag{2}$$

Equation of fitted regression line is $\hat{y}_i = 34.55 - 0.95x_i$. Or if we want it with the variable names, we can write

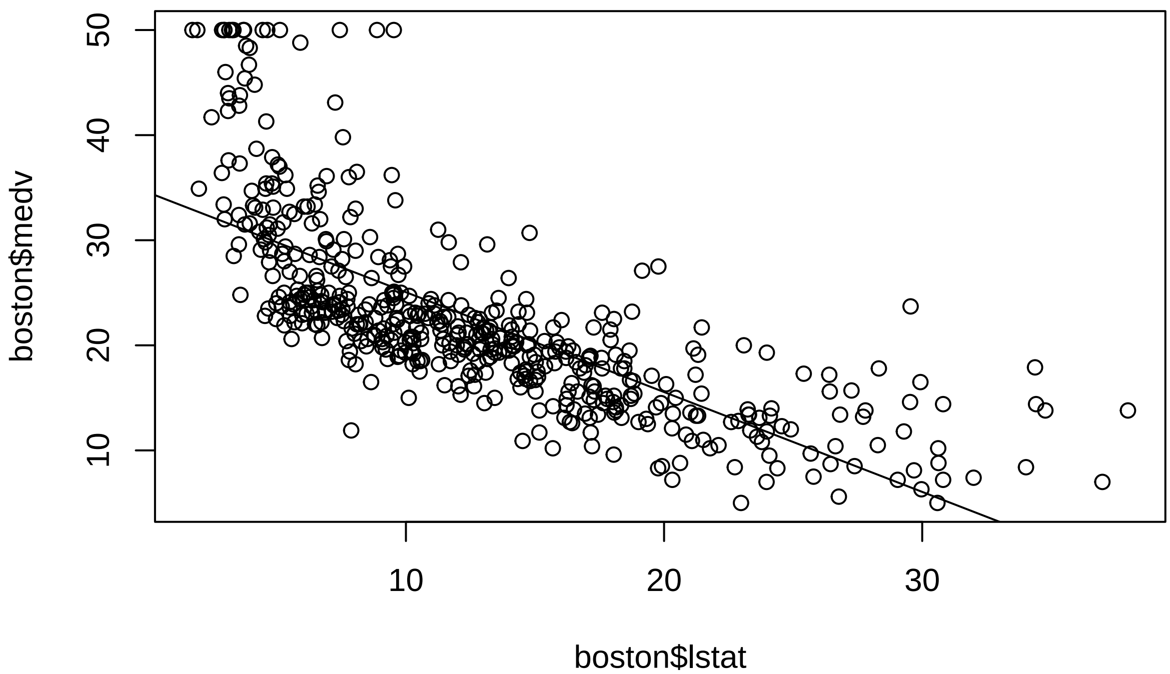
$$\widehat{medv} = 34.55 - 0.95 \times lstat \tag{3}$$

What's the interpretation of $\hat{\beta}_1 = -0.95$?

The interpretation of -0.95 is that if `lstat` or the lower status of the population increases by 1 percent then the `medv` or median value of the home is predicted to decrease by 950 dollars. OR *a* *decrease of 950 dollars in the median value of the home is associated with a 1 percent increase in the lower status of the population.*

The R^2 is 0.5441. The line is not a perfect fit, but it's not bad either. We can see the scatter plot with the fitted line using the `plot()` function. We will use the `abline()` function to add the fitted line in the scatter plot.

```
plot(boston$lstat, boston$medv)
abline(model_fit)
```



Also we can write that the 54.4% of the variation in the median value of the home is explained by the variation in the lower status of the population.

3.2 Model Assumptions and Significance Testing

For the simple linear regression model, our model assumption is

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where $f(X) = \beta_0 + \beta_1 X$ is our linear CEF, and

- Y: dependent variable median value of the home or `medv`.
- X: - independent variable lower status of the population or `lstat`
- ϵ: error term
- β_0: Unknown Population Intercept Coefficient
- β_1: Unknown Population Slope Coefficient

Using linear model assumption, we can easily show that

$$\mathbb{E}(\epsilon|X = x) = 0$$

This means conditional mean of error is 0. This is not a direct assumption, rather this is a result that we get if we assume linear CEF.

There is also another very important assumption that we will use, that is

$$\mathbb{V}(\epsilon|X = x) = \sigma^2$$

This means that the variance of the error term is constant for values of X. This is called the **homokcedasticity** assumption. The standard errors of the estimated coefficients, that we use here are based on this assumption. If this assumption is violated, then the standard errors of the estimated coefficients will be different, which we don't cover here, that is called **heteroskedasticity**. Under homoskedasticitty we can also show that

$$\mathbb{V}(\epsilon|X = x) = \mathbb{V}(\epsilon)$$

So this means $\mathbb{V}(\epsilon) = \sigma^2$.

Notice if $\beta_1 = 0$, this means in population there is no relationship between the two variables X and Y. In other words there is no relationship between the median value of the home and the lower status of the population. If we test this claim, this is called **significance testing**. So mathematically we want to do the test

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

3.2.1 t-test

Now we will do t test to check the claims in the hypotheses. This testing procedure is same as t-test that we saw before. We will use the t-test statistic to test this claim. The t-test statistic is

$$t_{calc} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

where $SE(\hat{\beta}_1)$ is the standard error of the slope coefficient $\hat{\beta}_1$. The $SE(\hat{\beta}_1)$ is given in the regression output. So we can calculate the t-test statistic as

```r
tcalc <- (-0.95 - 0)/0.03873
tcalc
```

```
## [1] -24.52879
```

Notice the value of the t-statistic is also given in the output. So now we will compare this with two critical values $t_{\alpha/2}$ and $t_{1-\alpha/2}$ coming from t distribution with $n-2$ degrees of freedom. Here n is the sample size. We will use the `qt()` function to get the critical values. leet's

```r
alpha <- 0.05
n <- nrow(boston)
n
```

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

```r
qt(1 - alpha/2, n - 2)
```

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

```r
qt(alpha/2, n - 2)
```

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

In this case notice our $t_{calc} < -1.964682$, so we can reject the Null.

Recall we can do the same test using the p-value.

```r
abs_tcalc <-abs(tcalc)
pvalue <- 2 * (1-pt(abs_tcalc, n - 2))
pvalue
```

```
## [1] 0
```

The p value looks very very. In fact it is so small that R actually shows 0, so we can reject the Null at $\alpha = 0.01$ or $\alpha = 0.05$ or $\alpha = .10$

Interestingly for this test p value is also calculated by R. It shows that p value is smaller that $2e-16$, this is a scientific printing, this means it's smaller than $2 \times 10^{-16} = 0.0000000000000002$. So we can reject the Null. If you want to stop this scientific printing, you can use the following command

```r
options(scipen = 999)
```

Now run the summary function again

```r
summary(model_fit)
```

```
##
## 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 <0.0000000000000002 ***
## lstat       -0.95005    0.03873  -24.53 <0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```

3.2.2 F - test

The same testing can be done using another test statistic called F-statistic, which can be calculated using

$$F_{calc} = \frac{\text{MSR}}{\text{MSE}}$$

where MSR is the mean square regression and MSE is the mean square error. The MSR is given by

$$\text{MSR} = \frac{\text{SSR}}{\text{Df for SSR}} = \frac{\text{SSR}}{1} = \text{SSR}$$

where SSR is the sum of squares regression. The MSE is given by

$$\text{MSE} = \frac{\text{SSE}}{\text{Df for SSE}} = \frac{\text{SSE}}{n-2}$$

where SSE is the sum of squares error. More details about this sum of squared errors are given in the slides. We can get all SS and MS using `anova()` function in R

```r
anova(model_fit)
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df Sum Sq Mean Sq F value                Pr(>F)
## lstat       1  23244 23243.9  601.62 < 0.00000000000000022 ***
## Residuals 504  19472    38.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here the testing procedure is same, if the p-value of the F test statistic is less than α, then we can reject the Null. In the simple linear linear regression in case of F test we are doing the same test as t-test and the procedure is also same. If the $p < \alpha$, then we can reject the Null.

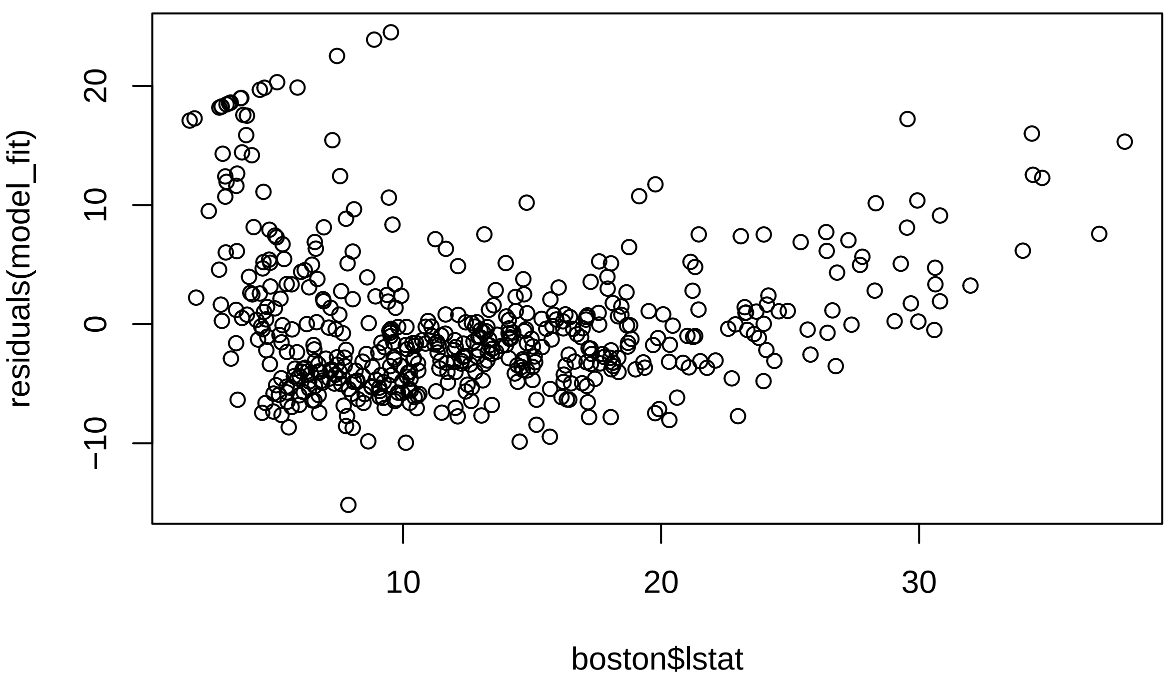
4 Cheking Model Assumptions

There are some assumptions when we fit a linear regression model. See the model assumptions section in the slides. Now it is possible to check whether our assumptions are correct or not using the regression results, sometimes this is known as **model diagnostics / diagnostic test**. We will now check Assumption 2 (linear model assumption) and Assumption 3 (homoskedasticity) using the residual plots. There are some ways to check other model assumptions, but we will not cover them here. .

4.1 Checking Linear Model Assumption

Checking linear model assumption is important. Recall one of the important implications of the linear model assumption is that the conditional mean of the error term is 0. So we can check this assumption by plotting the residuals against the fitted values. If the linear model assumption is correct, then we should not see any pattern in the plot. Let's plot the residuals against the fitted values. First note that we can get the residuals by using the function `residuals(model_fit)`. Now we can plot this with the independent variable.

```
plot(boston$lstat,  residuals(model_fit))
```



Interstingly the picture shows a pattern. This means the conditional mean of the error term is not always 0. So we can conlcude that the linear model assumption is probably not correct. What is the solution? One solution is fit a nonlinear line and the other solution is taking more variables as inputs, this is called multiple linear regression. We will see this in the next chapter.

4.2 Checking Homoskedasticity

Using the same plot we can also check the homoskedasticity assumption. Recall the homoskedasticity assumption is that the variance of the error term is constant. The plot suggests the variancesa are probably same for different values of the independent variable. So we may conclude that the homoskedasticity assumption is probably correct.

5 Prediction Interval for the Response and Confidence Interval for the Mean Response