Begüm Arslan

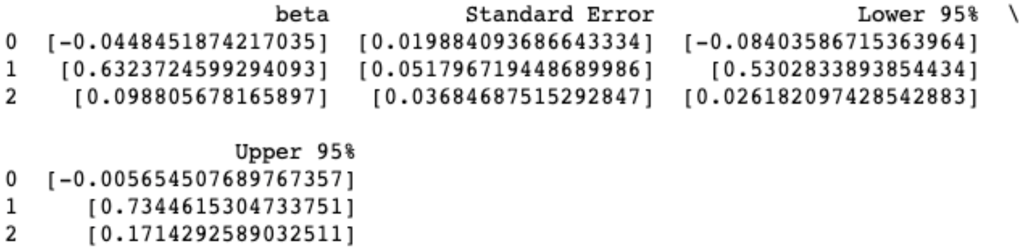
Homework 2 Report

The objective of this homework is to implement a linear regression function from scratch and observe the relstionship between the housing prices and the age and area of the houses. Hürriyet Emlak’s website was used in order to scrape and get the data. First 30 pages of the current for sale houses in İstanbul was taken and scraped for the areas, building ages and their prices.

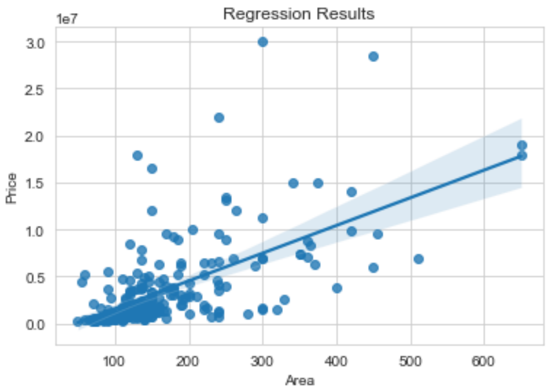
In this project, the null hypothesis is that there is no linear relationship between the area, age of the houses and prices. On the other hand, the alternative hypothesis is that there is a linear relationship between the area, age of the houses and their prices. Here, the dependent variable is the price of the houses and the independent variables are age and area of the houses.

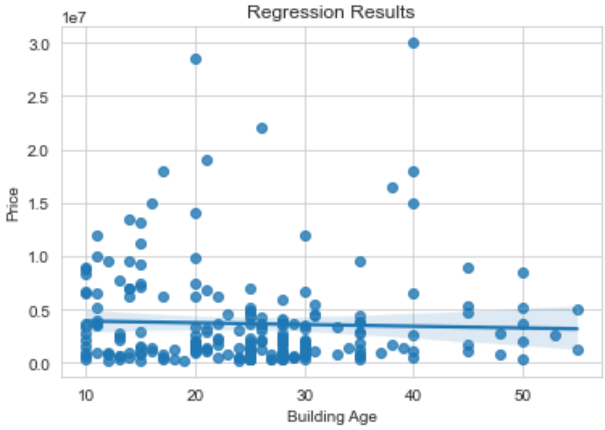
Before performing the linear regression, the data was adjusted in order to perform a successful fit for the linear regression. Outliers had been removed from the data with the help of the z-score and data was scaled by MinMaxScaler.

After performing the linear regression, results are shown below:



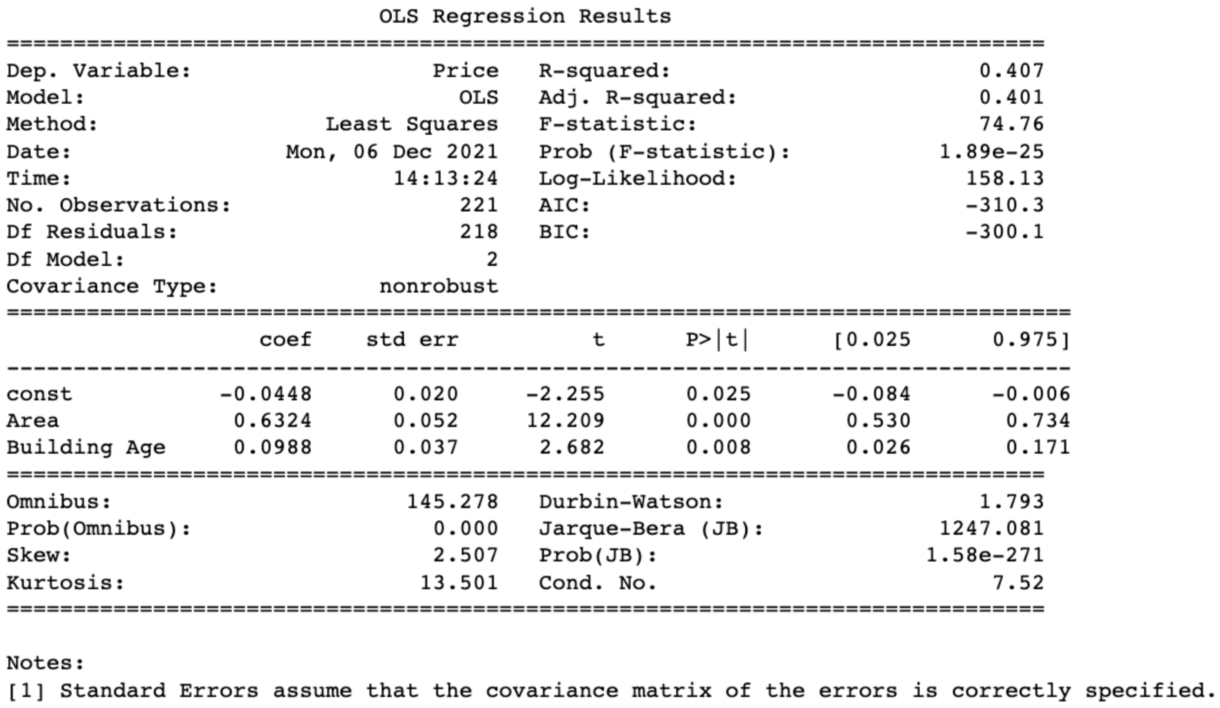
```
                   beta         Standard Error             Lower 95%  \
0  [-0.0448451874217035]  [0.019884093686643334]  [-0.08403586715363964]
1   [0.6323724599294093]  [0.051796719448689986]    [0.5302833893854434]
2    [0.098805678165897]   [0.03684687515292847]  [0.026182097428542883]

                Upper 95%
0  [-0.005654507689767357]
1     [0.7344615304733751]
2     [0.1714292589032511]
```





The final formula is: Price = -0.0448 + 0.632*Area + 0.0989*Age.

In order to check the results of the linear regression from scratch function, OLS statsmodels was used. Summary table of the OLS Statsmodels is shown below:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.407
Model:                            OLS   Adj. R-squared:                  0.401
Method:                 Least Squares   F-statistic:                     74.76
Date:                Mon, 06 Dec 2021   Prob (F-statistic):           1.89e-25
Time:                        14:13:24   Log-Likelihood:                 158.13
No. Observations:                 221   AIC:                            -310.3
Df Residuals:                     218   BIC:                            -300.1
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.0448      0.020     -2.255      0.025      -0.084      -0.006
Area             0.6324      0.052     12.209      0.000       0.530       0.734
Building Age     0.0988      0.037      2.682      0.008       0.026       0.171
==============================================================================
Omnibus:                      145.278   Durbin-Watson:                   1.793
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1247.081
Skew:                           2.507   Prob(JB):                    1.58e-271
Kurtosis:                      13.501   Cond. No.                         7.52
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

As it can be seen from the OLS Regression Results, the coefficients and credible intervals are same with the linear regression scratch model which shows that the model performs successfully. Also, when looked at the F-statistics and its p-value, F-statistics is way larger than one and its p value is less than 0.05 which shows that there is a good linear relationship between the predictor variable and features. In order to infer if a given feature is relevant to the target variable or not, t-statistics can be observed. In our case, t values of Area and building age is high and their p-values are significant which shows us that both area and building age are relevant features for predicting prices of the houses. Thus, this again shows that the null hypothesis is being rejected and alternative hypothesis is being accepted.