Onur Osmanoglu (66746)

Advanced Data Analysis in Python Final Project

This final project aims to establish a causal relationship between some house related aspects and house prices. Although the current work may seem to be a replication of the exercise, namely "Explaining House Prices" in Verbeek's (2008) book "A Guide to Modern Econometrics", the difference is that the whole data analysis in this project is conducted by using Python.

Verbeek (2008) defines the relationships concerning the elements of a house and their impact on house prices as hedonic price function in a way that price of a house is reflected by the "implicit price" of some characteristics related to that house. In line with this definition, the hypothesis of the project is the following: "*As the house related features increase, the price of a house increase as well.*"

Data: The data of Verbeek's (2008) exercise are actually based on the work of Anglin and Gencay (1996) that includes the sales prices of 546 houses between the time period July – September 1987 in Windsor, Canada. The summary and the descriptions of the variables used in the current final project is as follows:

```
Summary of Variables

          count          mean          std      min      25%      50%      75%      max
price     546.0  68121.593750  26702.669922  25000.0  49125.0  62000.0  82000.0   190000
lotsize   546.0   5150.265625   2168.160156   1650.0   3600.0   4600.0   6360.0    16200
bedrooms  546.0      2.965201      0.737387      1.0      2.0      3.0      3.0      6.0
bathrms   546.0      1.285714      0.502159      1.0      1.0      1.0      2.0      4.0
stories   546.0      1.807692      0.868203      1.0      1.0      2.0      2.0      4.0
driveway  546.0      0.858974      0.348369      0.0      1.0      1.0      1.0      1.0
recroom   546.0      0.177656      0.382573      0.0      0.0      0.0      0.0      1.0
fullbase  546.0      0.349817      0.477350      0.0      0.0      0.0      1.0      1.0
gashw     546.0      0.045788      0.209215      0.0      0.0      0.0      0.0      1.0
airco     546.0      0.316850      0.465676      0.0      0.0      0.0      1.0      1.0
garagepl  546.0      0.692308      0.861305      0.0      0.0      0.0      1.0      3.0
prefarea  546.0      0.234432      0.424033      0.0      0.0      0.0      0.0      1.0
```

price: house price

lotsize: lot size of the property in square feet

bedrooms: number of bedrooms

bathrms: number of full bathrooms

garagepl: number of garage places

stories: number of storys

driveway: (Dummy) the presence of a driveway near the house

recroom: (Dummy) the presence of recreational room

fullbase: (Dummy) the presence of full basement

airco: (Dummy) the presence of central air conditioning

prefarea: (Dummy) being located in a preferred area

gashw: (Dummy) using gas for hot water heating

First of all, in order to test the mentioned hypothesis, a linear regression is used as the base model with price as dependent variable, and the other variables explained above as independent variables. The result of the given model is the following:

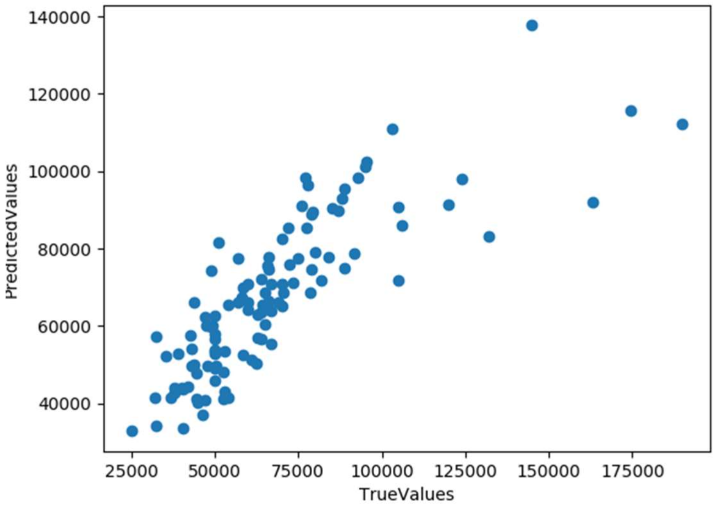
```
Linear Regression with Nonrobust Standard Errors
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.673
Model:                            OLS   Adj. R-squared:                  0.666
Method:                 Least Squares   F-statistic:                     99.97
Date:                Fri, 24 May 2019   Prob (F-statistic):          6.18e-122
Time:                        15:27:07   Log-Likelihood:                -6034.1
No. Observations:                 546   AIC:                         1.209e+04
Df Residuals:                     534   BIC:                         1.214e+04
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -4038.3504   3409.471     -1.184      0.237   -1.07e+04    2659.271
lotsize        3.5463      0.350     10.124      0.000       2.858       4.234
bedrooms    1832.0035   1047.000      1.750      0.081    -224.741    3888.748
bathrms     1.434e+04   1489.921      9.622      0.000    1.14e+04    1.73e+04
stories     6556.9457    925.290      7.086      0.000    4739.291    8374.600
driveway    6687.7789   2045.246      3.270      0.001    2670.065    1.07e+04
recroom     4511.2838   1899.958      2.374      0.018     778.976    8243.592
fullbase    5452.3855   1588.024      3.433      0.001    2332.845    8571.926
gashw       1.283e+04   3217.597      3.988      0.000    6510.706    1.92e+04
airco       1.263e+04   1555.021      8.124      0.000    9578.182    1.57e+04
garagepl    4244.8290    840.544      5.050      0.000    2593.650    5896.008
prefarea    9369.5132   1669.091      5.614      0.000    6090.724    1.26e+04
==============================================================================
Omnibus:                       93.454   Durbin-Watson:                   1.604
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              247.620
Skew:                           0.853   Prob(JB):                     1.70e-54
Kurtosis:                       5.824   Cond. No.                     3.07e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.07e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

The regression results show that all independent variables are significant predictors of house prices. Interestingly, the number of bedrooms seems to have a marginally significant impact on the house prices. According to the results, for instance, an increase in the amount of 100 square feet in the lot size of the house results in USD 355, on average, keeping everything else constant. In the same manner, one extra bedroom increases the house price by USD 1,832 on average, keeping everything else constant.

Additionally, to make sure that the model is not affected by heteroskedasticity, a linear regression with robust standard error is run, and its results table is the following:

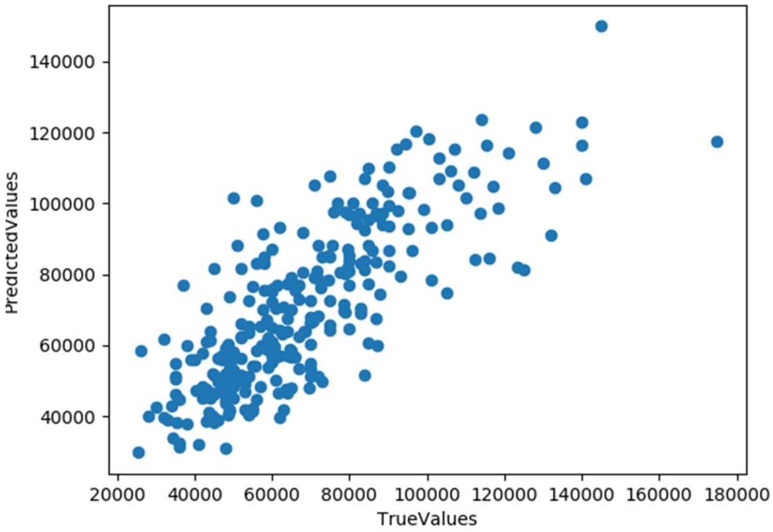
```
Linear Regression with Robust Standard Errors

"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.673
Model:                            OLS   Adj. R-squared:                  0.666
Method:                 Least Squares   F-statistic:                     87.32
Date:                Fri, 24 May 2019   Prob (F-statistic):          1.05e-111
Time:                        15:27:30   Log-Likelihood:                -6034.1
No. Observations:                 546   AIC:                         1.209e+04
Df Residuals:                     534   BIC:                         1.214e+04
Df Model:                          11
Covariance Type:                  HC1
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -4038.3504   3182.329     -1.269      0.205   -1.03e+04    2213.069
lotsize        3.5463      0.394      9.004      0.000       2.773       4.320
bedrooms    1832.0035   1038.158      1.765      0.078    -207.371    3871.378
bathrms     1.434e+04   1899.664      7.546      0.000    1.06e+04    1.81e+04
stories     6556.9457    869.607      7.540      0.000    4848.676    8265.215
driveway    6687.7789   1657.457      4.035      0.000    3431.843    9943.715
recroom     4511.2838   2144.416      2.104      0.036     298.757    8723.810
fullbase    5452.3855   1769.054      3.082      0.002    1977.227    8927.544
gashw       1.283e+04   4242.979      3.024      0.003    4496.428    2.12e+04
airco       1.263e+04   1666.225      7.582      0.000    9359.731    1.59e+04
garagepl    4244.8290    946.285      4.486      0.000    2385.930    6103.728
prefarea    9369.5132   1870.884      5.008      0.000    5694.319     1.3e+04
==============================================================================
Omnibus:                       93.454   Durbin-Watson:                   1.604
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              247.620
Skew:                           0.853   Prob(JB):                     1.70e-54
Kurtosis:                       5.824   Cond. No.                     3.07e+04
==============================================================================

Warnings:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 3.07e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

Additionally, in order to observe if the baseline model, which is linear regression in this project, can predict the rest of the housing prices by using the trained part of the data, “train_test_split” function of Python is used in two stages. Firstly, the amount of data used for training was 80% and testing was 20%, and the correlation coefficient yields to 0.822, meaning that the tested data can predict the house prices with a success rate of almost 82%, based on the training data.



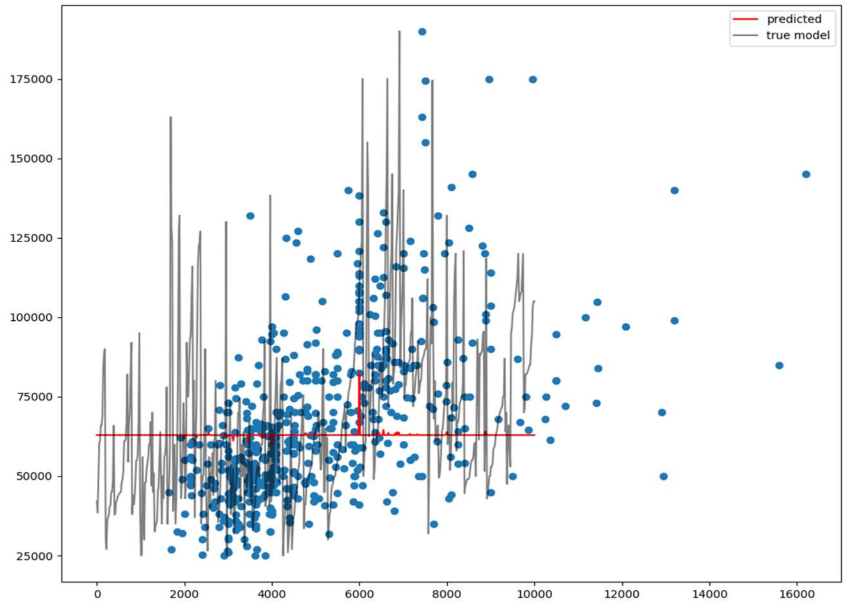
```
array([[1.        , 0.82230728],
       [0.82230728, 1.        ]])
```

The process is rerun by employing 50% of data for training, and the remaining 50% for testing the data, and the correlation coefficient is 0.794. In this situation, one may claim that the testing data can predict the housing prices based on the training data with approximately 79% of accuracy.



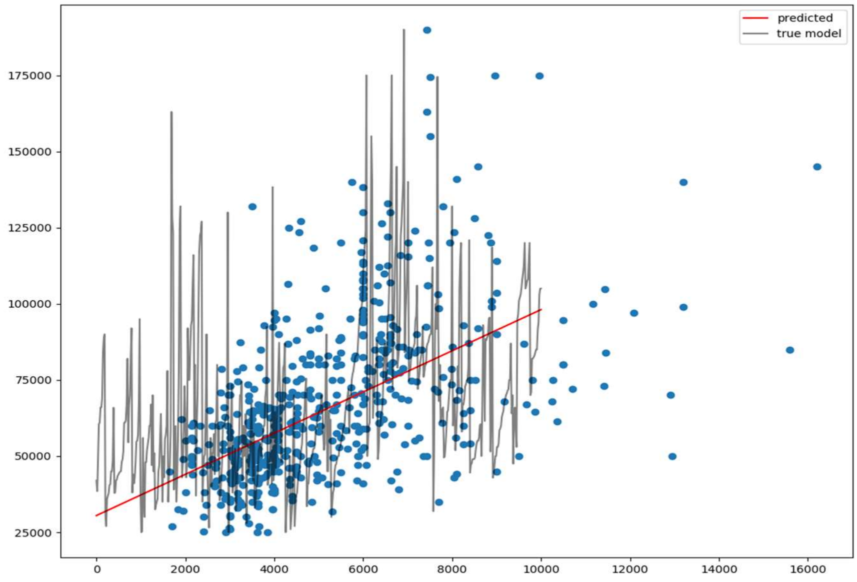
```
array([[1.        , 0.79378749],
       [0.79378749, 1.        ]])
```

Additionally, two Support Vector Regressions (SVR), with the difference of kernels and epsilons are conducted as another model to compare the results with the ones above. However, it is important to note that only the lot size is used as the independent variable in these regressions. The graphs, as well as the Mean Squared Errors (MSE) of the SVR's are the following:



The graph above shows the SVR conducted with C of 1000, as well as rbf as kernel and epsilon of 0.1 which are the default options (https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html). The mean squared error (MSE) for the given SVR is 735089635.13.

The graph below illustrates the SVR that is run with linear kernel, C of 1000, and epsilon of 1.0, and this version of SVR yields to the MSE of 716252642.74



References

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

Verbeek, M. (2008). *A guide to modern econometrics*. John Wiley & Sons.