econotest7

December 1, 2020

1 Econometrics Final Case Project

Jacci Ferrantino

```
In [1]: import statsmodels.api as sm
        import statsmodels.stats as sms
        from statsmodels.stats import outliers_influence as oi
        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
```

2 Background

This project is of an applied nature and uses data that are available in the data file Capstone-HousePrices. The source of these data is Anglin and Gencay, "Semiparametric Estimation of a Hedonic Price Function"(Journal of Applied Econometrics 11, 1996, pages 633-648). We consider the modeling and prediction of house prices. Data are available for 546 observations of the following variables:

1. sell: Sale price of the house
2. lot: Lot size of the property in square feet
3. bdms: Number of bedrooms
4. fb: Number of full bathrooms
5. sty: Number of stories excluding basement
6. drv: Dummy that is 1 if the house has a driveway and 0 otherwise
7. rec: Dummy that is 1 if the house has a recreational room and 0 otherwise
8. ffin: Dummy that is 1 if the house has a full finished basement and 0 otherwise
9. ghw: Dummy that is 1 if the house uses gas for hot water heating and 0 otherwise
10. ca: Dummy that is 1 if there is central air conditioning and 0 otherwise
11. gar: Number of covered garage places
12. reg: Dummy that is 1 if the house is located in a preferred neighborhood of the city and 0 otherwise
13. obs: Observation number, needed in part (h)

```
In [2]: # Import data that will be used to construct confidence interval of population mean
        df = pd.read_excel("test7.xls")
```

```
In [3]: print(df.head(2))
```

```
   obs   sell   lot  bdms  fb  sty  drv  rec  ffin  ghw  ca  gar  reg
0    1  42000  5850     3   1    2    1    0     1    0   0    1    0
1    2  38500  4000     2   1    1    1    0     0    0   0    0    0
```

3 Questions

4 (a)

Consider a linear model where the sale price of a house is the dependent variable and the explanatory variables are the other variables given above. Perform a test for linearity. What do you conclude based on the test result?

```
In [4]: df.dropna()
        model = sm.OLS.from_formula("sell ~ lot+bdms+fb+sty+drv+rec+ffin+ghw+ca+gar+reg", data=
        res = model.fit()
        print(res.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   sell   R-squared:                       0.673
Model:                            OLS   Adj. R-squared:                  0.666
Method:                 Least Squares   F-statistic:                     99.97
Date:                Tue, 01 Dec 2020   Prob (F-statistic):          6.18e-122
Time:                        00:58:52   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]
------------------------------------------------------------------------------
Intercept  -4038.3504   3409.471     -1.184      0.237   -1.07e+04    2659.271
lot            3.5463      0.350     10.124      0.000       2.858       4.234
bdms        1832.0035   1047.000      1.750      0.081    -224.741    3888.748
fb          1.434e+04   1489.921      9.622      0.000    1.14e+04    1.73e+04
sty         6556.9457    925.290      7.086      0.000    4739.291    8374.600
drv         6687.7789   2045.246      3.270      0.001    2670.065    1.07e+04
rec         4511.2838   1899.958      2.374      0.018     778.976    8243.592
ffin        5452.3855   1588.024      3.433      0.001    2332.845    8571.926
ghw         1.283e+04   3217.597      3.988      0.000    6510.706    1.92e+04
ca          1.263e+04   1555.021      8.124      0.000    9578.182    1.57e+04
gar         4244.8290    840.544      5.050      0.000    2593.650    5896.008
reg         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.
```

```
In [5]: reset=oi.reset_ramsey(res, degree=2)
        print(reset)
```

```
<F test: F=array([[26.98603521]]), p=2.922110452788588e-07, df_denom=533, df_num=1>
```

4.0.1 Conclusion:

The Durbin-Watson stat is used to test for autocorrelation. It is about 1.497. Usually, if the Durbin-Watson stat is between 1.5 and 2.5, it is a good sign that there is no autocorrelation. The Jarque-Bera test statistic is not significantly different than zero. This may indicate that our data does not have a normal distribution. The RESET test has an F-Test of approx 27 and a p-value of approx 0, indicating that the model may be misspecified. The R2 (percentage explained) of the model is about 67.3%.

5 (b)

Now consider a linear model where the log of the sale price of the house is the dependent variable and the explanatory variables are as before. Perform again the test for linearity. What do you conclude now?

```
In [6]: df["log_sell"]=np.log(df["sell"])
        model2 = sm.OLS.from_formula("log_sell ~ lot+bdms+fb+sty+drv+rec+ffin+ghw+ca+gar+reg",
        res2 = model2.fit()
        print(res2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.677
Model:                            OLS   Adj. R-squared:                  0.670
Method:                 Least Squares   F-statistic:                     101.6
Date:                Tue, 01 Dec 2020   Prob (F-statistic):          3.67e-123
Time:                        00:58:55   Log-Likelihood:                 73.873
No. Observations:                 546   AIC:                            -123.7
Df Residuals:                     534   BIC:                            -72.11
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
```

```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     10.0256      0.047    212.210      0.000       9.933      10.118
lot         5.057e-05   4.85e-06     10.418      0.000     4.1e-05    6.01e-05
bdms           0.0340      0.015      2.345      0.019       0.006       0.063
fb             0.1678      0.021      8.126      0.000       0.127       0.208
sty            0.0923      0.013      7.197      0.000       0.067       0.117
drv            0.1307      0.028      4.610      0.000       0.075       0.186
rec            0.0735      0.026      2.792      0.005       0.022       0.125
ffin           0.0994      0.022      4.517      0.000       0.056       0.143
ghw            0.1784      0.045      4.000      0.000       0.091       0.266
ca             0.1780      0.022      8.262      0.000       0.136       0.220
gar            0.0508      0.012      4.358      0.000       0.028       0.074
reg            0.1271      0.023      5.496      0.000       0.082       0.173
==============================================================================
Omnibus:                        7.621   Durbin-Watson:                   1.510
Prob(Omnibus):                  0.022   Jarque-Bera (JB):                8.443
Skew:                          -0.199   Prob(JB):                       0.0147
Kurtosis:                       3.461   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.
```

```
In [7]: reset2=oi.reset_ramsey(res2, degree=2)
        print(reset2)
```

```
<F test: F=array([[0.27031433]]), p=0.6033368567016177, df_denom=533, df_num=1>
```

5.0.1 Conclusion:

Durbin-Watson stat is between 1.5 and 2.5, it is a good sign that there is no autocorrelation. The Jarque-Bera test statistic is not significantly different than zero. This may indicate that our data does not have a normal distribution. The RESET test has an F-Test of approx .27 and a p-value of approx 0.6033, indicating that the model may be correctly specified. The R2 (percentage explained) of the model is about 67.7%.

6 (c)

Continue with the linear model from question (b). Estimate a model that includes both the lot size variable and its logarithm, as well as all other explanatory variables without transformation. What is your conclusion, should we include lot size itself or its logarithm?

```
In [8]: df["log_lot"]=np.log(df["lot"])
        model3 = sm.OLS.from_formula("log_sell ~ log_lot+lot+bdms+fb+sty+drv+rec+ffin+ghw+ca+g
```

```
res3 = model3.fit()
print(res3.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.687
Model:                            OLS   Adj. R-squared:                  0.680
Method:                 Least Squares   F-statistic:                     97.51
Date:                Tue, 01 Dec 2020   Prob (F-statistic):          6.43e-126
Time:                        00:58:56   Log-Likelihood:                 82.843
No. Observations:                 546   AIC:                            -139.7
Df Residuals:                     533   BIC:                            -83.75
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7.1505      0.683     10.469      0.000       5.809       8.492
log_lot        0.3827      0.091      4.219      0.000       0.205       0.561
lot         -1.49e-05   1.62e-05     -0.918      0.359   -4.68e-05     1.7e-05
bdms           0.0349      0.014      2.442      0.015       0.007       0.063
fb             0.1659      0.020      8.161      0.000       0.126       0.206
sty            0.0912      0.013      7.224      0.000       0.066       0.116
drv            0.1068      0.028      3.752      0.000       0.051       0.163
rec            0.0547      0.026      2.078      0.038       0.003       0.106
ffin           0.1052      0.022      4.848      0.000       0.063       0.148
ghw            0.1791      0.044      4.079      0.000       0.093       0.265
ca             0.1643      0.021      7.657      0.000       0.122       0.206
gar            0.0483      0.011      4.203      0.000       0.026       0.071
reg            0.1344      0.023      5.884      0.000       0.090       0.179
==============================================================================
Omnibus:                        7.927   Durbin-Watson:                   1.525
Prob(Omnibus):                  0.019   Jarque-Bera (JB):                9.364
Skew:                          -0.180   Prob(JB):                      0.00926
Kurtosis:                       3.531   Cond. No.                     4.27e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.27e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
In [9]: reset3= oi.reset_ramsey(res3, degree=2)
        print(reset3)
```

```
<F test: F=array([[0.06769]]), p=0.7948309729185185, df_denom=532, df_num=1>
```

6.0.1 Conclusion:

Durbin-Watson stat is between 1.5 and 2.5, it is a good sign that there is no autocorrelation. The Jarque-Bera test statistic is still not significantly different than zero. This may indicate that our data does not have a normal distribution. The RESET test has an F-Test of approx .067 and a p-value of approx 0.79483, indicating that the model may be correctly specified. This is larger than the previous models. The R2 (percentage explained) of the model is about 68.7%, also the highest of the models so far, indicating this model is the most relevant thus far and we should include its logarithm instead of lot itself.

7 (d)

Consider now a model where the log of the sale price of the house is the dependent variable and the explanatory variables are the log transformation of lot size, with all other explanatory variables as before. We now consider interaction effects of the log lot size with the other variables. Construct these interaction variables. How many are individually significant?

```
In [10]: model4 = sm.OLS.from_formula("log_sell ~ log_lot+bdms+fb+sty+drv+rec+ffin+ghw+ca+gar+
         res4 = model4.fit()
         print(res4.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.695
Model:                            OLS   Adj. R-squared:                  0.683
Method:                 Least Squares   F-statistic:                     56.89
Date:                Tue, 01 Dec 2020   Prob (F-statistic):          2.26e-120
Time:                        00:58:58   Log-Likelihood:                 89.971
No. Observations:                 546   AIC:                            -135.9
Df Residuals:                     524   BIC:                            -41.28
Df Model:                          21
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept        8.9665      1.071      8.375      0.000       6.863      11.070
log_lot          0.1527      0.128      1.190      0.235      -0.099       0.405
bdms             0.0191      0.327      0.058      0.953      -0.623       0.661
fb              -0.3682      0.429     -0.858      0.391      -1.211       0.475
sty              0.4889      0.310      1.579      0.115      -0.120       1.097
drv             -1.4634      0.717     -2.040      0.042      -2.872      -0.054
rec              1.6740      0.656      2.552      0.011       0.385       2.963
ffin            -0.0318      0.446     -0.071      0.943      -0.907       0.843
ghw             -0.5059      0.903     -0.560      0.575      -2.279       1.268
ca              -0.3403      0.496     -0.686      0.493      -1.315       0.634
gar              0.4019      0.259      1.554      0.121      -0.106       0.910
reg              0.1185      0.480      0.247      0.805      -0.824       1.061
log_lot:bdms     0.0021      0.039      0.054      0.957      -0.074       0.078
```

```
log_lot:fb        0.0620      0.050      1.237      0.217      -0.036       0.161
log_lot:sty      -0.0464      0.036     -1.290      0.198      -0.117       0.024
log_lot:drv       0.1915      0.087      2.193      0.029       0.020       0.363
log_lot:rec      -0.1885      0.076     -2.468      0.014      -0.338      -0.038
log_lot:ffin      0.0159      0.053      0.301      0.763      -0.088       0.120
log_lot:ghw       0.0811      0.107      0.759      0.448      -0.129       0.291
log_lot:ca        0.0595      0.058      1.026      0.305      -0.054       0.174
log_lot:gar      -0.0414      0.030     -1.372      0.171      -0.101       0.018
log_lot:reg       0.0015      0.056      0.027      0.978      -0.108       0.112
==============================================================================
Omnibus:                        7.141   Durbin-Watson:                   1.524
Prob(Omnibus):                  0.028   Jarque-Bera (JB):                8.203
Skew:                          -0.173   Prob(JB):                       0.0165
Kurtosis:                       3.491   Cond. No.                     4.77e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.77e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
In [11]: reset4= oi.reset_ramsey(res4, degree=2)
         print(reset4)
```

```
<F test: F=array([[0.01157078]]), p=0.9143799774375326, df_denom=523, df_num=1>
```

7.0.1 Conclusion:

With the addition of the new interaction terms, the model seems to be performing better with a R2 of 69.5% and a RESET F-Test value of 0.0115. However, when looking at the individual t-stats and corresponding p-values of the coefficients, only the driveway and rec-room dummy variables (drv and rec) and their interaction terms (multiplied by the log of lot size) drv* log_lot and rec* log lot, are statistically significant.

8 (e)

Perform an F-test for the joint significance of the interaction effects from question (d).

```
In [29]: #restricted model
         model5 = sm.OLS.from_formula("log_sell ~ log_lot+bdms+fb+sty+drv+rec+ffin+ghw+ca+gar+r
         res5 = model5.fit()
         print(res5.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.692
```

```
Model:                            OLS   Adj. R-squared:                  0.684
Method:                 Least Squares   F-statistic:                     91.79
Date:                Tue, 01 Dec 2020   Prob (F-statistic):          1.32e-126
Time:                        01:11:27   Log-Likelihood:                 86.878
No. Observations:                 546   AIC:                            -145.8
Df Residuals:                     532   BIC:                            -85.52
Df Model:                          13
Covariance Type:            nonrobust
================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept       8.7419      0.629     13.906      0.000       7.507       9.977
log_lot         0.1791      0.077      2.323      0.021       0.028       0.330
bdms            0.0388      0.014      2.714      0.007       0.011       0.067
fb              0.1615      0.020      7.971      0.000       0.122       0.201
sty             0.0908      0.013      7.242      0.000       0.066       0.115
drv            -1.1900      0.665     -1.790      0.074      -2.496       0.116
rec             1.5025      0.626      2.402      0.017       0.274       2.731
ffin            0.1028      0.022      4.763      0.000       0.060       0.145
ghw             0.1845      0.044      4.223      0.000       0.099       0.270
ca              0.1653      0.021      7.792      0.000       0.124       0.207
gar             0.0469      0.011      4.107      0.000       0.024       0.069
reg             0.1326      0.023      5.880      0.000       0.088       0.177
log_lot:drv     0.1594      0.081      1.962      0.050      -0.000       0.319
log_lot:rec    -0.1683      0.073     -2.314      0.021      -0.311      -0.025
==============================================================================
Omnibus:                        7.976   Durbin-Watson:                   1.526
Prob(Omnibus):                  0.019   Jarque-Bera (JB):                9.237
Skew:                          -0.189   Prob(JB):                      0.00987
Kurtosis:                       3.513   Cond. No.                     1.23e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.23e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [ ]:
```python
#sum of squared residuals
ssr_unrestricted=res4.ssr
ssr_restricted=res5.ssr

f_testa=(ssr_restricted - ssr_unrestricted)/10
f_testb=ssr_unrestricted/524
f_test=f_testa/f_testb

print("F-Test for Joint Significance:",f_test,'with p>.05 indicates joint significance
```

9 (f)

Now perform model specification on the interaction variables using the general-to-specific approach. (Only eliminate the interaction effects.)

```
In [13]: ModelA=sm.OLS.from_formula("log_sell ~ log_lot+bdms+fb+sty+drv+rec+ffin+ghw+ca+gar+reg
         res4 = model4.fit()
         print(res4.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.695
Model:                            OLS   Adj. R-squared:                  0.683
Method:                 Least Squares   F-statistic:                     56.89
Date:                Tue, 01 Dec 2020   Prob (F-statistic):         2.26e-120
Time:                        00:59:00   Log-Likelihood:                 89.971
No. Observations:                 546   AIC:                            -135.9
Df Residuals:                     524   BIC:                            -41.28
Df Model:                          21
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept        8.9665      1.071      8.375      0.000       6.863      11.070
log_lot          0.1527      0.128      1.190      0.235      -0.099       0.405
bdms             0.0191      0.327      0.058      0.953      -0.623       0.661
fb              -0.3682      0.429     -0.858      0.391      -1.211       0.475
sty              0.4889      0.310      1.579      0.115      -0.120       1.097
drv             -1.4634      0.717     -2.040      0.042      -2.872      -0.054
rec              1.6740      0.656      2.552      0.011       0.385       2.963
ffin            -0.0318      0.446     -0.071      0.943      -0.907       0.843
ghw             -0.5059      0.903     -0.560      0.575      -2.279       1.268
ca              -0.3403      0.496     -0.686      0.493      -1.315       0.634
gar              0.4019      0.259      1.554      0.121      -0.106       0.910
reg              0.1185      0.480      0.247      0.805      -0.824       1.061
log_lot:bdms     0.0021      0.039      0.054      0.957      -0.074       0.078
log_lot:fb       0.0620      0.050      1.237      0.217      -0.036       0.161
log_lot:sty     -0.0464      0.036     -1.290      0.198      -0.117       0.024
log_lot:drv      0.1915      0.087      2.193      0.029       0.020       0.363
log_lot:rec     -0.1885      0.076     -2.468      0.014      -0.338      -0.038
log_lot:ffin     0.0159      0.053      0.301      0.763      -0.088       0.120
log_lot:ghw      0.0811      0.107      0.759      0.448      -0.129       0.291
log_lot:ca       0.0595      0.058      1.026      0.305      -0.054       0.174
log_lot:gar     -0.0414      0.030     -1.372      0.171      -0.101       0.018
log_lot:reg      0.0015      0.056      0.027      0.978      -0.108       0.112
==============================================================================
Omnibus:                        7.141   Durbin-Watson:                   1.524
Prob(Omnibus):                  0.028   Jarque-Bera (JB):                8.203
```

```
Skew:                          -0.173   Prob(JB):                       0.0165
Kurtosis:                       3.491   Cond. No.                     4.77e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.77e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [14]:
```python
#log_lot*reg eliminated
modelb = sm.OLS.from_formula("log_sell ~ log_lot+bdms+fb+sty+drv+rec+ffin+ghw+ca+gar+
resb = modelb.fit()
print(resb.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.695
Model:                            OLS   Adj. R-squared:                  0.683
Method:                 Least Squares   F-statistic:                     59.85
Date:                Tue, 01 Dec 2020   Prob (F-statistic):          2.88e-121
Time:                        00:59:02   Log-Likelihood:                 89.970
No. Observations:                 546   AIC:                            -137.9
Df Residuals:                     525   BIC:                            -47.59
Df Model:                          20
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept        8.9679      1.068      8.394      0.000       6.869      11.067
log_lot          0.1525      0.128      1.191      0.234      -0.099       0.404
bdms             0.0195      0.326      0.060      0.952      -0.621       0.660
fb              -0.3677      0.428     -0.859      0.391      -1.209       0.474
sty              0.4872      0.303      1.607      0.109      -0.108       1.083
drv             -1.4679      0.697     -2.106      0.036      -2.837      -0.098
rec              1.6747      0.655      2.558      0.011       0.388       2.961
ffin            -0.0349      0.430     -0.081      0.935      -0.880       0.810
ghw             -0.5043      0.900     -0.560      0.575      -2.272       1.264
ca              -0.3395      0.495     -0.686      0.493      -1.312       0.633
gar              0.4023      0.258      1.560      0.119      -0.104       0.909
reg              0.1315      0.023      5.705      0.000       0.086       0.177
log_lot:bdms     0.0020      0.039      0.052      0.958      -0.074       0.078
log_lot:fb       0.0620      0.050      1.238      0.216      -0.036       0.160
log_lot:sty     -0.0462      0.035     -1.312      0.190      -0.115       0.023
log_lot:drv      0.1921      0.085      2.259      0.024       0.025       0.359
log_lot:rec     -0.1885      0.076     -2.473      0.014      -0.338      -0.039
log_lot:ffin     0.0163      0.051      0.319      0.750      -0.084       0.116
log_lot:ghw      0.0809      0.107      0.759      0.448      -0.128       0.290
```

```
log_lot:ca       0.0595      0.058      1.027      0.305      -0.054       0.173
log_lot:gar     -0.0414      0.030     -1.377      0.169      -0.100       0.018
==============================================================================
Omnibus:                        7.117   Durbin-Watson:                   1.524
Prob(Omnibus):                  0.028   Jarque-Bera (JB):                8.170
Skew:                          -0.172   Prob(JB):                       0.0168
Kurtosis:                       3.490   Cond. No.                     4.73e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.73e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [15]:
```python
#log_lot*bdrms removed
modelc = sm.OLS.from_formula("log_sell ~ log_lot+bdms+fb+sty+drv+rec+ffin+ghw+ca+gar+
resc = modelc.fit()
print(resc.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.695
Model:                            OLS   Adj. R-squared:                  0.684
Method:                 Least Squares   F-statistic:                     63.12
Date:                Tue, 01 Dec 2020   Prob (F-statistic):          3.57e-122
Time:                        00:59:02   Log-Likelihood:                 89.969
No. Observations:                 546   AIC:                            -139.9
Df Residuals:                     526   BIC:                            -53.89
Df Model:                          19
Covariance Type:            nonrobust
================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept      8.9349      0.861     10.381      0.000       7.244      10.626
log_lot        0.1565      0.103      1.515      0.130      -0.046       0.359
bdms           0.0366      0.015      2.506      0.013       0.008       0.065
fb            -0.3775      0.386     -0.979      0.328      -1.135       0.380
sty            0.4820      0.286      1.685      0.093      -0.080       1.044
drv           -1.4625      0.689     -2.123      0.034      -2.816      -0.109
rec            1.6759      0.654      2.564      0.011       0.392       2.960
ffin          -0.0374      0.427     -0.088      0.930      -0.877       0.802
ghw           -0.5016      0.898     -0.559      0.577      -2.265       1.262
ca            -0.3387      0.494     -0.685      0.493      -1.309       0.632
gar            0.4010      0.257      1.563      0.119      -0.103       0.905
reg            0.1314      0.023      5.710      0.000       0.086       0.177
log_lot:fb     0.0631      0.045      1.405      0.161      -0.025       0.151
log_lot:sty   -0.0456      0.033     -1.372      0.171      -0.111       0.020
```

```
log_lot:drv      0.1914      0.084      2.277      0.023      0.026      0.357
log_lot:rec     -0.1887      0.076     -2.479      0.014     -0.338     -0.039
log_lot:ffin     0.0166      0.051      0.328      0.743     -0.083      0.116
log_lot:ghw      0.0806      0.106      0.758      0.449     -0.128      0.289
log_lot:ca       0.0594      0.058      1.027      0.305     -0.054      0.173
log_lot:gar     -0.0413      0.030     -1.380      0.168     -0.100      0.017
==============================================================================
Omnibus:                        7.132   Durbin-Watson:                   1.524
Prob(Omnibus):                  0.028   Jarque-Bera (JB):                8.190
Skew:                          -0.173   Prob(JB):                       0.0167
Kurtosis:                       3.491   Cond. No.                     2.83e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.83e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
In [16]: #log_lot*ffin removed
         modelc = sm.OLS.from_formula("log_sell ~ log_lot+bdms+fb+sty+drv+rec+ffin+ghw+ca+gar+
         resc = modelc.fit()
         print(resc.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.695
Model:                            OLS   Adj. R-squared:                  0.685
Method:                 Least Squares   F-statistic:                     66.73
Date:                Tue, 01 Dec 2020   Prob (F-statistic):          4.54e-123
Time:                        00:59:02   Log-Likelihood:                 89.913
No. Observations:                 546   AIC:                            -141.8
Df Residuals:                     527   BIC:                            -60.08
Df Model:                          18
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept       8.8765      0.841     10.550      0.000       7.224      10.529
log_lot         0.1636      0.101      1.621      0.106      -0.035       0.362
bdms            0.0365      0.015      2.507      0.012       0.008       0.065
fb             -0.3819      0.385     -0.992      0.322      -1.138       0.375
sty             0.4885      0.285      1.713      0.087      -0.072       1.049
drv            -1.4502      0.687     -2.110      0.035      -2.801      -0.100
rec             1.6214      0.632      2.567      0.011       0.380       2.862
ffin            0.1023      0.022      4.691      0.000       0.059       0.145
ghw            -0.5160      0.896     -0.576      0.565      -2.276       1.244
ca             -0.3545      0.491     -0.722      0.471      -1.320       0.611
```

```
gar             0.4015      0.256      1.566      0.118      -0.102       0.905
reg             0.1323      0.023      5.786      0.000       0.087       0.177
log_lot:fb      0.0636      0.045      1.417      0.157      -0.025       0.152
log_lot:sty    -0.0464      0.033     -1.403      0.161      -0.111       0.019
log_lot:drv     0.1899      0.084      2.264      0.024       0.025       0.355
log_lot:rec    -0.1822      0.073     -2.481      0.013      -0.326      -0.038
log_lot:ghw     0.0825      0.106      0.778      0.437      -0.126       0.291
log_lot:ca      0.0612      0.057      1.066      0.287      -0.052       0.174
log_lot:gar    -0.0413      0.030     -1.383      0.167      -0.100       0.017
==============================================================================
Omnibus:                        7.017   Durbin-Watson:                   1.525
Prob(Omnibus):                  0.030   Jarque-Bera (JB):                8.046
Skew:                          -0.170   Prob(JB):                       0.0179
Kurtosis:                       3.487   Cond. No.                     2.78e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.78e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [17]:
```python
#log_lot*ghw removed
modeld = sm.OLS.from_formula("log_sell ~ log_lot+bdms+fb+sty+drv+rec+ffin+ghw+ca+gar+
resd= modeld.fit()
print(resd.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.695
Model:                            OLS   Adj. R-squared:                  0.685
Method:                 Least Squares   F-statistic:                     70.67
Date:                Tue, 01 Dec 2020   Prob (F-statistic):          7.16e-124
Time:                        00:59:03   Log-Likelihood:                 89.600
No. Observations:                 546   AIC:                            -143.2
Df Residuals:                     528   BIC:                            -65.75
Df Model:                          17
Covariance Type:            nonrobust
===============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept      8.8086      0.836     10.530      0.000       7.165      10.452
log_lot        0.1716      0.100      1.710      0.088      -0.025       0.369
bdms           0.0366      0.015      2.513      0.012       0.008       0.065
fb            -0.3764      0.385     -0.978      0.329      -1.132       0.380
sty            0.4909      0.285      1.722      0.086      -0.069       1.051
drv           -1.4326      0.687     -2.086      0.037      -2.782      -0.083
rec            1.6306      0.631      2.583      0.010       0.390       2.871
```

```
ffin            0.1036      0.022      4.766      0.000       0.061       0.146
ghw             0.1799      0.044      4.098      0.000       0.094       0.266
ca             -0.3397      0.491     -0.692      0.489      -1.304       0.624
gar             0.3973      0.256      1.551      0.121      -0.106       0.900
reg             0.1311      0.023      5.750      0.000       0.086       0.176
log_lot:fb      0.0630      0.045      1.405      0.161      -0.025       0.151
log_lot:sty    -0.0467      0.033     -1.412      0.159      -0.112       0.018
log_lot:drv     0.1878      0.084      2.241      0.025       0.023       0.352
log_lot:rec    -0.1832      0.073     -2.496      0.013      -0.327      -0.039
log_lot:ca      0.0593      0.057      1.034      0.302      -0.053       0.172
log_lot:gar    -0.0408      0.030     -1.368      0.172      -0.099       0.018
==============================================================================
Omnibus:                        7.161   Durbin-Watson:                   1.532
Prob(Omnibus):                  0.028   Jarque-Bera (JB):                7.984
Skew:                          -0.186   Prob(JB):                       0.0185
Kurtosis:                       3.462   Cond. No.                     2.73e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.73e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [18]:
```python
#log_lot*ca removed
modele = sm.OLS.from_formula("log_sell ~ log_lot+bdms+fb+sty+drv+rec+ffin+ghw+ca+gar+r
rese= modele.fit()
print(rese.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.694
Model:                            OLS   Adj. R-squared:                  0.685
Method:                 Least Squares   F-statistic:                     75.01
Date:                Tue, 01 Dec 2020   Prob (F-statistic):          1.38e-124
Time:                        00:59:03   Log-Likelihood:                 89.048
No. Observations:                 546   AIC:                            -144.1
Df Residuals:                     529   BIC:                            -70.95
Df Model:                          16
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept       8.7822      0.836     10.503      0.000       7.140      10.425
log_lot         0.1748      0.100      1.743      0.082      -0.022       0.372
bdms            0.0352      0.015      2.428      0.016       0.007       0.064
fb             -0.3803      0.385     -0.988      0.324      -1.136       0.376
sty             0.4720      0.285      1.659      0.098      -0.087       1.031
```

```
drv             -1.4286      0.687     -2.080      0.038      -2.778      -0.079
rec              1.5567      0.627      2.482      0.013       0.324       2.789
ffin             0.1034      0.022      4.756      0.000       0.061       0.146
ghw              0.1772      0.044      4.043      0.000       0.091       0.263
ca               0.1672      0.021      7.880      0.000       0.125       0.209
gar              0.3385      0.250      1.355      0.176      -0.152       0.829
reg              0.1330      0.023      5.848      0.000       0.088       0.178
log_lot:fb       0.0636      0.045      1.418      0.157      -0.025       0.152
log_lot:sty     -0.0443      0.033     -1.344      0.180      -0.109       0.020
log_lot:drv      0.1873      0.084      2.235      0.026       0.023       0.352
log_lot:rec     -0.1746      0.073     -2.395      0.017      -0.318      -0.031
log_lot:gar     -0.0339      0.029     -1.164      0.245      -0.091       0.023
==============================================================================
Omnibus:                        6.946   Durbin-Watson:                   1.529
Prob(Omnibus):                  0.031   Jarque-Bera (JB):                7.810
Skew:                          -0.177   Prob(JB):                       0.0201
Kurtosis:                       3.467   Cond. No.                     2.71e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.71e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [19]:
```python
#log_lot*gar removed
modelf = sm.OLS.from_formula("log_sell ~ log_lot+bdms+fb+sty+drv+rec+ffin+ghw+ca+gar+
resf= modelf.fit()
print(resf.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.693
Model:                            OLS   Adj. R-squared:                  0.685
Method:                 Least Squares   F-statistic:                     79.87
Date:                Tue, 01 Dec 2020   Prob (F-statistic):          2.96e-125
Time:                        00:59:04   Log-Likelihood:                 88.350
No. Observations:                 546   AIC:                            -144.7
Df Residuals:                     530   BIC:                            -75.86
Df Model:                          15
Covariance Type:            nonrobust
================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept       8.7739      0.836     10.490      0.000       7.131      10.417
log_lot         0.1758      0.100      1.753      0.080      -0.021       0.373
bdms            0.0353      0.015      2.432      0.015       0.007       0.064
fb             -0.3402      0.383     -0.887      0.375      -1.093       0.413
```

```
sty              0.4682      0.285      1.645      0.101      -0.091       1.027
drv             -1.2369      0.667     -1.854      0.064      -2.547       0.073
rec              1.5141      0.626      2.417      0.016       0.283       2.745
ffin             0.1028      0.022      4.727      0.000       0.060       0.145
ghw              0.1800      0.044      4.112      0.000       0.094       0.266
ca               0.1670      0.021      7.869      0.000       0.125       0.209
gar              0.0480      0.011      4.200      0.000       0.026       0.070
reg              0.1299      0.023      5.750      0.000       0.086       0.174
log_lot:fb       0.0590      0.045      1.321      0.187      -0.029       0.147
log_lot:sty     -0.0439      0.033     -1.331      0.184      -0.109       0.021
log_lot:drv      0.1645      0.082      2.018      0.044       0.004       0.325
log_lot:rec     -0.1694      0.073     -2.327      0.020      -0.312      -0.026
==============================================================================
Omnibus:                        7.433   Durbin-Watson:                   1.523
Prob(Omnibus):                  0.024   Jarque-Bera (JB):                8.418
Skew:                          -0.186   Prob(JB):                       0.0149
Kurtosis:                       3.482   Cond. No.                     2.60e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.6e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
In [20]: #log_lot*fb removed
         modelg = sm.OLS.from_formula("log_sell ~ log_lot+bdms+fb+sty+drv+rec+ffin+ghw+ca+gar+
         resg= modelg.fit()
         print(resg.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.692
Model:                            OLS   Adj. R-squared:                  0.684
Method:                 Least Squares   F-statistic:                     85.33
Date:                Tue, 01 Dec 2020   Prob (F-statistic):          7.43e-126
Time:                        00:59:04   Log-Likelihood:                 87.452
No. Observations:                 546   AIC:                            -144.9
Df Residuals:                     531   BIC:                            -80.37
Df Model:                          14
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept       8.2985      0.755     10.984      0.000       6.814       9.783
log_lot         0.2309      0.091      2.529      0.012       0.052       0.410
bdms            0.0362      0.015      2.497      0.013       0.008       0.065
fb              0.1655      0.021      8.030      0.000       0.125       0.206
```

```
sty             0.3842      0.278      1.384      0.167      -0.161       0.930
drv            -1.2546      0.667     -1.880      0.061      -2.566       0.056
rec             1.4725      0.626      2.352      0.019       0.243       2.702
ffin            0.1004      0.022      4.631      0.000       0.058       0.143
ghw             0.1809      0.044      4.130      0.000       0.095       0.267
ca              0.1662      0.021      7.831      0.000       0.125       0.208
gar             0.0475      0.011      4.155      0.000       0.025       0.070
reg             0.1313      0.023      5.812      0.000       0.087       0.176
log_lot:sty    -0.0340      0.032     -1.058      0.291      -0.097       0.029
log_lot:drv     0.1669      0.082      2.047      0.041       0.007       0.327
log_lot:rec    -0.1647      0.073     -2.263      0.024      -0.308      -0.022
==============================================================================
Omnibus:                        8.037   Durbin-Watson:                   1.525
Prob(Omnibus):                  0.018   Jarque-Bera (JB):                9.196
Skew:                          -0.195   Prob(JB):                       0.0101
Kurtosis:                       3.501   Cond. No.                     2.18e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.18e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
In [21]: #log_lot*sty removed
         modelh = sm.OLS.from_formula("log_sell ~ log_lot+bdms+fb+sty+drv+rec+ffin+ghw+ca+gar+
         resh= modelh.fit()
         print(resh.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.692
Model:                            OLS   Adj. R-squared:                  0.684
Method:                 Least Squares   F-statistic:                     91.79
Date:                Tue, 01 Dec 2020   Prob (F-statistic):          1.32e-126
Time:                        00:59:04   Log-Likelihood:                 86.878
No. Observations:                 546   AIC:                            -145.8
Df Residuals:                     532   BIC:                            -85.52
Df Model:                          13
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept       8.7419      0.629     13.906      0.000       7.507       9.977
log_lot         0.1791      0.077      2.323      0.021       0.028       0.330
bdms            0.0388      0.014      2.714      0.007       0.011       0.067
fb              0.1615      0.020      7.971      0.000       0.122       0.201
sty             0.0908      0.013      7.242      0.000       0.066       0.115
```

```
drv            -1.1900      0.665     -1.790      0.074      -2.496       0.116
rec             1.5025      0.626      2.402      0.017       0.274       2.731
ffin            0.1028      0.022      4.763      0.000       0.060       0.145
ghw             0.1845      0.044      4.223      0.000       0.099       0.270
ca              0.1653      0.021      7.792      0.000       0.124       0.207
gar             0.0469      0.011      4.107      0.000       0.024       0.069
reg             0.1326      0.023      5.880      0.000       0.088       0.177
log_lot:drv     0.1594      0.081      1.962      0.050      -0.000       0.319
log_lot:rec    -0.1683      0.073     -2.314      0.021      -0.311      -0.025
==============================================================================
Omnibus:                        7.976   Durbin-Watson:                   1.526
Prob(Omnibus):                  0.019   Jarque-Bera (JB):                9.237
Skew:                          -0.189   Prob(JB):                      0.00987
Kurtosis:                       3.513   Cond. No.                     1.23e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.23e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
In [22]: #log_lot*rec removed
         modeli = sm.OLS.from_formula("log_sell ~ log_lot+bdms+fb+sty+drv+rec+ffin+ghw+ca+gar+
         resi= modeli.fit()
         print(resi.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.689
Model:                            OLS   Adj. R-squared:                  0.682
Method:                 Least Squares   F-statistic:                     98.19
Date:                Tue, 01 Dec 2020   Prob (F-statistic):          1.83e-126
Time:                        00:59:06   Log-Likelihood:                 84.143
No. Observations:                 546   AIC:                            -142.3
Df Residuals:                     533   BIC:                            -86.35
Df Model:                          12
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept       8.8348      0.630     14.026      0.000       7.597      10.072
log_lot         0.1696      0.077      2.194      0.029       0.018       0.321
bdms            0.0346      0.014      2.429      0.015       0.007       0.063
fb              0.1642      0.020      8.091      0.000       0.124       0.204
sty             0.0918      0.013      7.290      0.000       0.067       0.116
drv            -1.1161      0.667     -1.674      0.095      -2.426       0.193
rec             0.0561      0.026      2.157      0.031       0.005       0.107
```

```
ffin            0.1029      0.022      4.748      0.000       0.060       0.145
ghw             0.1790      0.044      4.087      0.000       0.093       0.265
ca              0.1658      0.021      7.787      0.000       0.124       0.208
gar             0.0468      0.011      4.083      0.000       0.024       0.069
reg             0.1307      0.023      5.777      0.000       0.086       0.175
log_lot:drv     0.1500      0.081      1.841      0.066      -0.010       0.310
==============================================================================
Omnibus:                        7.294   Durbin-Watson:                   1.527
Prob(Omnibus):                  0.026   Jarque-Bera (JB):                8.228
Skew:                          -0.184   Prob(JB):                       0.0163
Kurtosis:                       3.476   Cond. No.                     1.22e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.22e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

9.0.1 Conclusion:

After ruling out interaction terms using the general to specific approach the final model should only include the interaction term: log of lot size* driveway dummy variable (see model i above).

10 (g)

One may argue that some of the explanatory variables are endogenous and that there may be omitted variables. For example, the 'condition' of the house in terms of how it is maintained is not a variable (and difficult to measure) but will affect the house price. It will also affect, or be reflected in, some of the other variables, such as whether the house has an air conditioning (which is mostly in newer houses). If the condition of the house is missing, will the effect of air conditioning on the (log of the) sale price be over- or underestimated? (For this question no computer calculations are required.)

10.0.1 Conclusion:

Since "condition of the home" is not included in the model, along with other factors which are correlated with the error term, epsilon, we conclude that house price (and log of house price) is endogenous, as well as several of the explanatory variables. The effect of air conditioning in this case will be overestimated as home condition will be included in the marginal effect of having air conditioning, overinflating our home price estimates.

11 (h)

Finally we analyze the predictive ability of the model. Consider again the model where the log of the sale price of the house is the dependent variable and the explanatory variables are the log transformation of lot size, with all other explanatory variables in their original form (and no interaction effects). Estimate the parameters of the model using the first 400 observations. Make predictions on the log of the price and calculate the MAE for the other 146 observations. How good is the predictive power of the model (relative to the variability in the log of the price)?

11.1 Let "train" = first 400 observations of dataset (df) & let "test"= final 146 observations of dataset (df):

In [23]:
```
train=df[0:400]
test=df[400:]
model_train = sm.OLS.from_formula("log_sell ~ log_lot+lot+bdms+fb+sty+drv+rec+ffin+ghw
res_train = model_train.fit()
print(res_train.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.671
Model:                            OLS   Adj. R-squared:                  0.660
Method:                 Least Squares   F-statistic:                     65.64
Date:                Tue, 01 Dec 2020   Prob (F-statistic):           1.64e-85
Time:                        00:59:09   Log-Likelihood:                 37.367
No. Observations:                 400   AIC:                            -48.73
Df Residuals:                     387   BIC:                             3.156
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7.9123      0.879      9.006      0.000       6.185       9.640
log_lot        0.2818      0.116      2.422      0.016       0.053       0.511
lot         5.934e-06   2.06e-05      0.289      0.773   -3.45e-05    4.63e-05
bdms           0.0376      0.017      2.153      0.032       0.003       0.072
fb             0.1520      0.025      6.140      0.000       0.103       0.201
sty            0.0885      0.018      4.853      0.000       0.053       0.124
drv            0.0878      0.032      2.759      0.006       0.025       0.150
rec            0.0560      0.034      1.634      0.103      -0.011       0.123
ffin           0.1144      0.027      4.274      0.000       0.062       0.167
ghw            0.1987      0.053      3.744      0.000       0.094       0.303
ca             0.1780      0.027      6.520      0.000       0.124       0.232
gar            0.0530      0.015      3.577      0.000       0.024       0.082
reg            0.1503      0.042      3.553      0.000       0.067       0.233
==============================================================================
Omnibus:                        0.901   Durbin-Watson:                   1.476
Prob(Omnibus):                  0.637   Jarque-Bera (JB):                0.716
Skew:                          -0.090   Prob(JB):                        0.699
Kurtosis:                       3.103   Cond. No.                     4.23e+05
==============================================================================

Warnings:
```

```
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.23e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
In [24]: model_test = sm.OLS.from_formula("log_sell ~ log_lot+lot+bdms+fb+sty+drv+rec+ffin+ghw+
         res_test = model_test.fit()
         print(res_test.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_sell   R-squared:                       0.731
Model:                            OLS   Adj. R-squared:                  0.707
Method:                 Least Squares   F-statistic:                     30.13
Date:                Tue, 01 Dec 2020   Prob (F-statistic):           3.41e-32
Time:                        00:59:10   Log-Likelihood:                 70.558
No. Observations:                 146   AIC:                            -115.1
Df Residuals:                     133   BIC:                            -76.33
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.7505      0.972      4.890      0.000       2.829       6.672
log_lot        0.6643      0.127      5.217      0.000       0.412       0.916
lot        -7.859e-05    2.3e-05     -3.416      0.001      -0.000   -3.31e-05
bdms           0.0232      0.024      0.951      0.343      -0.025       0.072
fb             0.1955      0.034      5.691      0.000       0.128       0.263
sty            0.0841      0.015      5.608      0.000       0.054       0.114
drv            0.5375      0.113      4.736      0.000       0.313       0.762
rec            0.0355      0.035      1.011      0.314      -0.034       0.105
ffin           0.0787      0.033      2.354      0.020       0.013       0.145
ghw            0.1028      0.072      1.431      0.155      -0.039       0.245
ca             0.1145      0.030      3.757      0.000       0.054       0.175
gar            0.0379      0.017      2.289      0.024       0.005       0.071
reg            0.1083      0.031      3.511      0.001       0.047       0.169
==============================================================================
Omnibus:                        1.734   Durbin-Watson:                   1.712
Prob(Omnibus):                  0.420   Jarque-Bera (JB):                1.381
Skew:                          -0.048   Prob(JB):                        0.501
Kurtosis:                       3.467   Cond. No.                     4.70e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.7e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
In [25]: mse_train=res_train.mse_model
         print("Mean-Squared Error for train (n=400) data:", mse_train)
```

```
Mean-Squared Error for train (n=400) data: 3.2955043626927334
```

```
In [26]: print(res_train.summary2())
```

```
                 Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.660
Dependent Variable: log_sell         AIC:                -48.7330
Date:               2020-12-01 00:59 BIC:                3.1560
No. Observations:   400              Log-Likelihood:     37.367
Df Model:           12               F-statistic:        65.64
Df Residuals:       387              Prob (F-statistic): 1.64e-85
R-squared:          0.671            Scale:              0.050204
--------------------------------------------------------------------
             Coef.    Std.Err.     t      P>|t|     [0.025   0.975]
--------------------------------------------------------------------
Intercept    7.9123     0.8786   9.0057   0.0000    6.1849   9.6397
log_lot      0.2818     0.1164   2.4216   0.0159    0.0530   0.5107
lot          0.0000     0.0000   0.2887   0.7729   -0.0000   0.0000
bdms         0.0376     0.0175   2.1528   0.0320    0.0033   0.0720
fb           0.1520     0.0248   6.1399   0.0000    0.1033   0.2007
sty          0.0885     0.0182   4.8528   0.0000    0.0527   0.1244
drv          0.0878     0.0318   2.7595   0.0061    0.0253   0.1504
rec          0.0560     0.0343   1.6338   0.1031   -0.0114   0.1235
ffin         0.1144     0.0268   4.2736   0.0000    0.0618   0.1671
ghw          0.1987     0.0531   3.7444   0.0002    0.0944   0.3031
ca           0.1780     0.0273   6.5200   0.0000    0.1243   0.2317
gar          0.0530     0.0148   3.5768   0.0004    0.0239   0.0821
reg          0.1503     0.0423   3.5529   0.0004    0.0671   0.2335
------------------------------------------------------------------
Omnibus:               0.901        Durbin-Watson:           1.476
Prob(Omnibus):         0.637        Jarque-Bera (JB):        0.716
Skew:                  -0.090       Prob(JB):                0.699
Kurtosis:              3.103        Condition No.:           422770
==================================================================
* The condition number is large (4e+05). This might indicate
strong multicollinearity or other numerical problems.
```

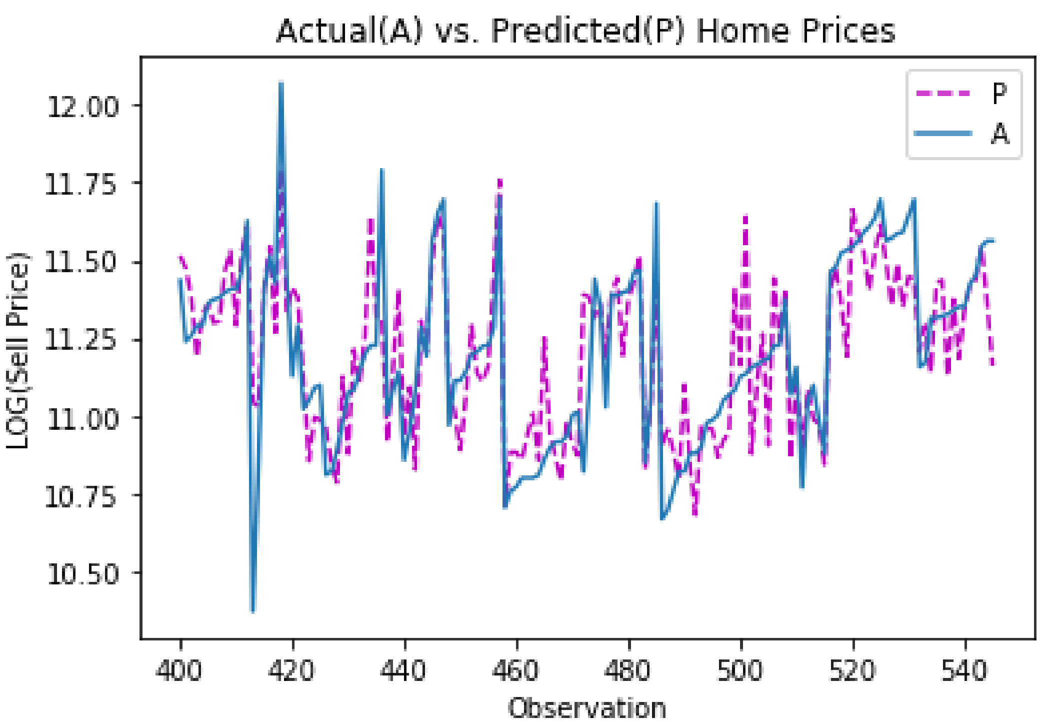
```
In [27]: pred=res_train.predict(test)
         print("Predicted Log of Home Price for Test Data:", pred)
```

```
Predicted Log of Home Price for Test Data: 400    11.513638
401    11.474370
402    11.381008
```

403	11.186940
404	11.325803
405	11.358523
406	11.298190
407	11.304389
408	11.466191
409	11.528824
410	11.290817
411	11.523928
412	11.605500
413	11.042359
414	11.029946
415	11.413355
416	11.542744
417	11.267786
418	11.792698
419	11.331574
420	11.413486
421	11.379393
422	11.124759
423	10.854466
424	10.994233
425	10.986283
426	10.968728
427	10.840954
428	10.785195
429	11.125551
	...
516	11.404139
517	11.474269
518	11.362657
519	11.190575
520	11.662842
521	11.583297
522	11.527259
523	11.402234
524	11.527259
525	11.613323
526	11.494522
527	11.354422
528	11.489631
529	11.348101
530	11.446655
531	11.439996
532	11.203667
533	11.290028
534	11.138776
535	11.425899

```
536    11.436641
537    11.126084
538    11.373627
539    11.183051
540    11.338372
541    11.422668
542    11.436641
543    11.545669
544    11.368590
545    11.160546
Length: 146, dtype: float64
```

```
In [28]: plt.plot(pred, "--m")
         plt.plot(test["log_sell"])
         plt.title("Actual(A) vs. Predicted(P) Home Prices")
         plt.ylabel("LOG(Sell Price)")
         plt.xlabel("Observation")
         plt.legend("PA")
         plt.show()
```



11.1.1 Conclusion:

As you can see from the performance metrics above (including mse, mae, r-squared, and adjusted r-squared) and visually by observing the predicted values on the "test" data (last 146 observations) versus the actual values, our model does have some predictive ability, albeit not perfect. There are still many improvements that could be made to improve the model's predictive accuracy.