# Housing Price Structure

September 14, 2023

[99]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import math
import matplotlib.pyplot as plt
import statsmodels.api as sm
from sklearn.model_selection import train_test_split, GridSearchCV, KFold,␣
↪cross_val_score
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.metrics import classification_report
```

[100]:
```
# Creating a function to print output in green and bold
# ANSI escape code for green color and bold font
GREEN_BOLD = '\033[1;32m'

# ANSI escape code to reset colors and font style
RESET = '\033[0m'

def print_green_bold(*args):
    text = ' '.join(str(arg) for arg in args)
    print(GREEN_BOLD + text + RESET)
```

## 1 Housing Price Structure

**The file MidCity.xls contains data on 128 recent sales of houses in a town. For each sale, the file shows the neighborhood in which the house is located, the number of offers made on the house, the squarefootage, whether the house is made out of brick, the number of bathrooms, the number of bedrooms, and the sellingprice. Neighborhoods 1 and 2 are more traditional whereas 3 is a more modern, newer and more prestigious part of town. Use regression models to estimate the pricing structure ofhouses in this town and answer the following questions:**

**1. Is there a premium for brick houses everything else being equal?**

[170]:
```
cdata=pd.read_csv('MidCity.csv')
cdata.head()
```

[170]:

| | Home | Nbhd | Offers | SqFt | Brick | Bedrooms | Bathrooms | Price |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 2 | 1790 | No | 2 | 2 | 114300 |
| 1 | 2 | 2 | 3 | 2030 | No | 4 | 2 | 114200 |
| 2 | 3 | 2 | 1 | 1740 | No | 3 | 2 | 114800 |
| 3 | 4 | 2 | 3 | 1980 | No | 3 | 2 | 94700 |
| 4 | 5 | 2 | 3 | 2130 | No | 3 | 3 | 119800 |

[171]:

```
# One-hot encoding on Nbhd and Brick variables
cdata_1 = pd.get_dummies(cdata, columns=['Nbhd', 'Brick'], drop_first=True)
cdata_1.head()
```

[171]:

| | Home | Offers | SqFt | Bedrooms | Bathrooms | Price | Nbhd_2 | Nbhd_3 | Brick_Yes |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 1790 | 2 | 2 | 114300 | 1 | 0 | 0 |
| 1 | 2 | 3 | 2030 | 4 | 2 | 114200 | 1 | 0 | 0 |
| 2 | 3 | 1 | 1740 | 3 | 2 | 114800 | 1 | 0 | 0 |
| 3 | 4 | 3 | 1980 | 3 | 2 | 94700 | 1 | 0 | 0 |
| 4 | 5 | 3 | 2130 | 3 | 3 | 119800 | 1 | 0 | 0 |

[172]:

```
# Independent variables
X_cols_cdata = ['Offers', 'SqFt', 'Nbhd_2', 'Nbhd_3', 'Brick_Yes', 'Bedrooms',␣
↪'Bathrooms']

# Dependent variable
y_cdata=cdata_1['Price']

# Adding a constant term
X_cdata= sm.add_constant(cdata_1[X_cols_cdata])

# Fit the multiple linear regression model
mlr_model_cdata = sm.OLS(y_cdata, X_cdata)
mlr_results_cdata = mlr_model_cdata.fit()

# Print the regression summary
print_green_bold(mlr_results_cdata.summary())
```

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Price | R-squared: | 0.869 |
| Model: | OLS | Adj. R-squared: | 0.861 |
| Method: | Least Squares | F-statistic: | 113.3 |
| Date: | Sun, 30 Jul 2023 | Prob (F-statistic): | 8.25e-50 |
| Time: | 15:24:49 | Log-Likelihood: | -1356.7 |
| No. Observations: | 128 | AIC: | 2729. |
| Df Residuals: | 120 | BIC: | 2752. |
| Df Model: | 7 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 2159.4982 | 8877.810 | 0.243 | 0.808 | -1.54e+04 | 1.97e+04 |
| Offers | -8267.4883 | 1084.777 | -7.621 | 0.000 | -1.04e+04 | -6119.706 |
| SqFt | 52.9937 | 5.734 | 9.242 | 0.000 | 41.640 | 64.347 |
| Nbhd_2 | -1560.5791 | 2396.765 | -0.651 | 0.516 | -6306.008 | 3184.850 |
| Nbhd_3 | 2.068e+04 | 3148.954 | 6.568 | 0.000 | 1.44e+04 | 2.69e+04 |
| Brick_Yes | 1.73e+04 | 1981.616 | 8.729 | 0.000 | 1.34e+04 | 2.12e+04 |
| Bedrooms | 4246.7939 | 1597.911 | 2.658 | 0.009 | 1083.042 | 7410.546 |
| Bathrooms | 7883.2785 | 2117.035 | 3.724 | 0.000 | 3691.696 | 1.21e+04 |

| | | | |
|---|---|---|---|
| Omnibus: | 3.026 | Durbin-Watson: | 1.921 |
| Prob(Omnibus): | 0.220 | Jarque-Bera (JB): | 2.483 |
| Skew: | 0.268 | Prob(JB): | 0.289 |
| Kurtosis: | 3.421 | Cond. No. | 2.03e+04 |

Notes:

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

[2] The condition number is large, 2.03e+04. This might indicate that there are strong multicollinearity or other numerical problems.

We note that the Brick_Yes dummy variable has a positive coefficent and a very low p-value (<0.001). This means that the Brick_Yes variable has a positive impact on Prices and this impact is statistically significant. This implies that, on average, brick houses have a higher price compared to non-brick houses when taking into account the other factors included in the model. Thus, we can conclude that there is a premium for brick houses when everything else is equal.

**2. Is there a premium for houses in neighborhood 3?**

Similar to the Brick_Yes dummy variable, the Nbhd_3 dummy has a positive coefficient with a very low p-value. This means it has a positive and statistically significant impact on Price and thus, there is a premium for houses in neighbourhood 3.

**3. Is there an extra premium for brick houses in neighborhood 3?**

To determine if there is an extra premium for brick houses in neighborhood 3, we need to look at the interaction effect between the Brick_Yes variable and the Nbhd_3 variable. In order to do this, we run a regression model with an additional interaction term added.

[173]:
```python
# Adding an interaction term between Brick_Yes and Nbhd_3
cdata_1['Interaction_Brick_Nbhd3'] = cdata_1['Brick_Yes'] * cdata_1['Nbhd_3']

# Adding a constant term
X_cdata_new = sm.add_constant(cdata_1[['Offers', 'SqFt', 'Nbhd_2', 'Nbhd_3',
 ↪'Brick_Yes', 'Bedrooms', 'Bathrooms', 'Interaction_Brick_Nbhd3']])

# Running the OLS regression
mlr_model_cdata_new = sm.OLS(y_cdata, X_cdata_new)
mlr_results_cdata_new = mlr_model_cdata_new.fit()
print_green_bold(mlr_results_cdata_new.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.875
Model:                            OLS   Adj. R-squared:                  0.866
Method:                 Least Squares   F-statistic:                     104.0
Date:                Sun, 30 Jul 2023   Prob (F-statistic):           5.05e-50
Time:                        15:25:04   Log-Likelihood:                -1353.5
No. Observations:                 128   AIC:                             2725.
Df Residuals:                     119   BIC:                             2751.
Df Model:                           8
Covariance Type:            nonrobust
================================================================================
===========
                             coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
-----------
const                   3009.9934   8706.264      0.346      0.730   -1.42e+04
2.02e+04
Offers                 -8401.0879   1064.370     -7.893      0.000   -1.05e+04
-6293.529
SqFt                      54.0648      5.636      9.593      0.000      42.905
65.225
Nbhd_2                  -673.0283   2376.477     -0.283      0.778   -5378.691
4032.634
Nbhd_3                  1.724e+04   3391.347      5.084      0.000    1.05e+04
2.4e+04
Brick_Yes               1.383e+04   2405.556      5.748      0.000    9063.223
1.86e+04
Bedrooms                4718.1634   1577.613      2.991      0.003    1594.333
7841.994
Bathrooms               6463.3650   2154.264      3.000      0.003    2197.708
1.07e+04
Interaction_Brick_Nbhd3 1.018e+04   4165.274      2.444      0.016    1933.918
1.84e+04
==============================================================================
```

The coefficient for the interaction term is positive with a statistically significant p-value of 0.016. This suggests that, on average, houses that are both brick houses and located in neighborhood 3 have a higher price when taking into account the other factors included in the model. This positive coefficient indicates that there is an additional premium for brick houses in neighborhood 3 beyond the individual effects of being in neighborhood 3 or having a brick house.

**4. For the purposes of prediction could you combine the neighborhoods 1 and 2 into a single "older" neighborhood?**

[174]:
```
# Adding the older neighbourhood flag
cdata_2=cdata.copy()
cdata_2['Older_Neighbourhood_flag']=cdata_2['Nbhd'].apply(lambda x:1 if x in␣
 ↪[1,2] else 0)
cdata_2 =cdata_2.drop(['Nbhd'], axis=1)
cdata_2 = pd.get_dummies(cdata_2, columns=['Brick'], drop_first=True)

# Independent variables
X_cdata_new_2 = ['Offers', 'SqFt', 'Older_Neighbourhood_flag', 'Brick_Yes',␣
 ↪'Bedrooms', 'Bathrooms']
X_cdata_new_2 = sm.add_constant(cdata_2[X_cdata_new_2])

# Fitting the multiple linear regression model with the new flag
mlr_model_cdata_new_2 = sm.OLS(y_cdata, X_cdata_new_2)
mlr_results_cdata_new_2 = mlr_model_cdata_new_2.fit()

# Regression summary
print_green_bold(mlr_results_cdata_new_2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.868
Model:                            OLS   Adj. R-squared:                  0.862
Method:                 Least Squares   F-statistic:                     132.8
Date:                Sun, 30 Jul 2023   Prob (F-statistic):           8.44e-51
Time:                        15:25:09   Log-Likelihood:                -1356.9
No. Observations:                 128   AIC:                             2728.
Df Residuals:                     121   BIC:                             2748.
Df Model:                           6
Covariance Type:            nonrobust
================================================================================
============
                               coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
------------
const                     2.501e+04   9658.403      2.589      0.011    5883.686
4.41e+04
Offers                   -8019.0028   1013.011     -7.916      0.000      -1e+04
-6013.480
SqFt                        52.1492      5.572      9.359      0.000      41.117
63.181
Older_Neighbourhood_flag -2.194e+04   2482.393     -8.837      0.000   -2.69e+04
-1.7e+04
Brick_Yes                 1.706e+04   1942.805      8.780      0.000    1.32e+04
2.09e+04
Bedrooms                  4070.0049   1570.921      2.591      0.011     959.952
7180.058
Bathrooms                 7810.6983   2109.060      3.703      0.000    3635.257
1.2e+04
==============================================================================
Omnibus:                        2.787   Durbin-Watson:                   1.902
Prob(Omnibus):                  0.248   Jarque-Bera (JB):                2.238
Skew:                           0.270   Prob(JB):                        0.327
Kurtosis:                       3.357   Cond. No.                     2.22e+04
```

The coefficient for the Older_Neighbourhood_flag is negative, with a very low p-value which indicates that the coefficient is statistically significant. Since the coefficient is negative, it suggests that on average, houses in older neighborhoods have a lower price compared to houses in newer neighborhoods when taking into account the other factors included in the model. This is in line with what we observed previously in part 2, where houses in Neighborhood 3 (a new neighborhood) were more expensive.

```
[ ]:
```