Applications: Model

Build regression models with Statsmodels

Setup

In [1]:

```
%matplotlib inline

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.formula.api as smf

sns.set_theme(style="ticks", color_codes=True)
```

Data preparation

See notebook `10a-application-model-data-exploration.ipynb` for details about data preprocessing and data exploration.

In [2]:

```
ROOT = "https://raw.githubusercontent.com/kirenz/modern-statistics/main/data/"
DATA = "duke-forest.csv"
df = pd.read_csv(ROOT + DATA)

# Drop irrelevant features
df = df.drop(['url', 'address', 'type'], axis=1)

# Convert data types
df['heating'] = df['heating'].astype("category")
df['cooling'] = df['cooling'].astype("category")
df['parking'] = df['parking'].astype("category")

# drop column with too many missing values
df = df.drop(['hoa'], axis=1)
# drop remaining row with missing value
df = df.dropna()
```

Data splitting

In [3]:

```
train_dataset = df.sample(frac=0.8, random_state=0)
test_dataset = df.drop(train_dataset.index)
```

Modeling

In [4]:

```
# Fit Model
lm = smf.ols(formula='price ~ area', data=train_dataset).fit()
```

In [5]:

```
# Short summary
lm.summary().tables[1]
```

Out[5]:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.593e+04	6.21e+04	1.383	0.171	-3.78e+04	2.1e+05
area	167.7007	20.741	8.085	0.000	126.391	209.010

In [6]:

```
# Full summary
lm.summary()
```

Out[6]:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.462
Model:	OLS	**Adj. R-squared:**	0.455
Method:	Least Squares	**F-statistic:**	65.37
Date:	Sun, 19 Sep 2021	**Prob (F-statistic):**	7.56e-12
Time:	14:29:59	**Log-Likelihood:**	-1053.3
No. Observations:	78	**AIC:**	2111.
Df Residuals:	76	**BIC:**	2115.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.593e+04	6.21e+04	1.383	0.171	-3.78e+04	2.1e+05
area	167.7007	20.741	8.085	0.000	126.391	209.010

Omnibus:	26.589	**Durbin-Watson:**	2.159
Prob(Omnibus):		**Jarque-Bera (JB):**	

	0.000		107.927
Skew:	-0.862	**Prob(JB):**	3.66e-24
Kurtosis:	8.499	**Cond. No.**	9.16e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.16e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

To obtain single statistics:

In [7]:

```
# Adjusted R squared
lm.rsquared_adj
```

Out[7]:

```
0.4553434818683253
```

In [8]:

```
# R squared
lm.rsquared
```

Out[8]:

```
0.4624169431427626
```

In [9]:

```
# AIC
lm.aic
```

Out[9]:

```
2110.625966301898
```

In [10]:

```
train_dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 78 entries, 26 to 73
Data columns (total 9 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   price       78 non-null     int64
 1   bed         78 non-null     int64
 2   bath        78 non-null     float64
 3   area        78 non-null     int64
 4   year_built  78 non-null     int64
 5   heating     78 non-null     category
 6   cooling     78 non-null     category
 7   parking     78 non-null     category
 8   lot         78 non-null     float64
```

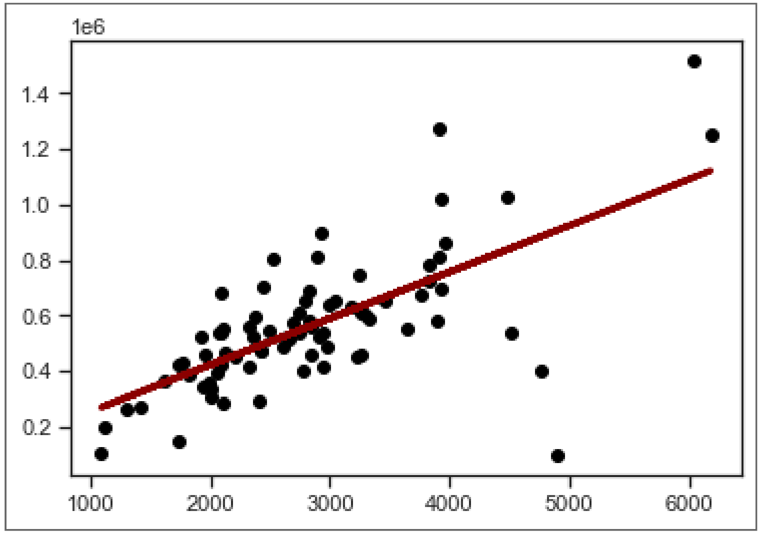
```
dtypes: category(3), float64(2), int64(4)
memory usage: 6.0 KB
```

In [11]:

```
# Add the regression predictions (as "pred") to our DataFrame
train_dataset['y_pred'] = lm.predict()
```

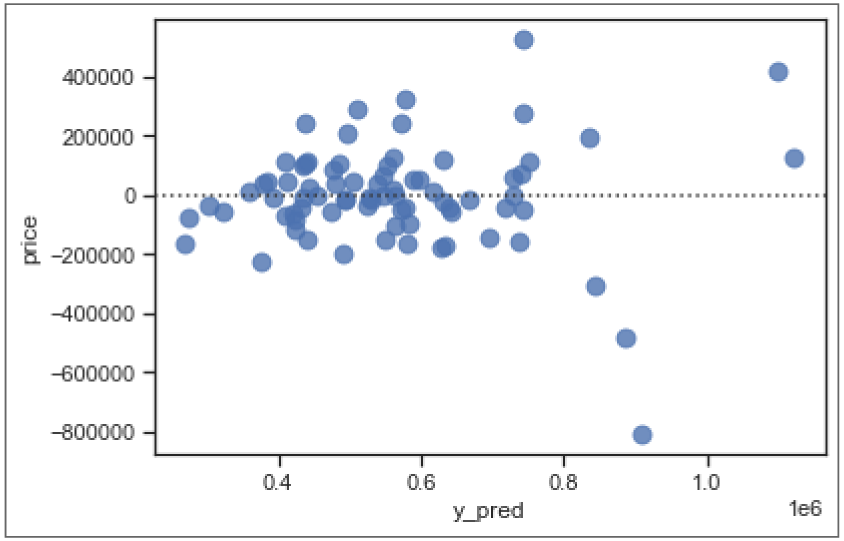
In [12]:

```
# Plot regression line
plt.scatter(train_dataset['area'], train_dataset['price'],  color='black')
plt.plot(train_dataset['area'], train_dataset['y_pred'], color='darkred', linewidth=3);
```



In [13]:

```
sns.residplot(x="y_pred", y="price", data=train_dataset, scatter_kws={"s": 80});
```



Multiple regression

In [14]:

```
lm_m = smf.ols(formula='price ~ area + bed + bath + year_built + cooling + lot', data=train_dataset).fit()
```

In [15]:

```
lm_m.summary()
```

Out[15]:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.626
Model:	OLS	**Adj. R-squared:**	0.595
Method:	Least Squares	**F-statistic:**	19.83
Date:	Sun, 19 Sep 2021	**Prob (F-statistic):**	1.86e-13
Time:	14:29:59	**Log-Likelihood:**	-1039.1
No. Observations:	78	**AIC:**	2092.
Df Residuals:	71	**BIC:**	2109.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.944e+06	2.26e+06	-1.302	0.197	-7.45e+06	1.56e+06
cooling[T.other]	-1.021e+05	3.67e+04	-2.778	0.007	-1.75e+05	-2.88e+04
area	111.8295	25.915	4.315	0.000	60.156	163.503
bed	5121.5208	3.1e+04	0.165	0.869	-5.68e+04	6.7e+04

bath	2.678e+04	2.94e+04	0.910	0.366	-3.19e+04	8.55e+04
year_built	1491.1176	1157.430	1.288	0.202	-816.732	3798.968
lot	3.491e+05	8.53e+04	4.094	0.000	1.79e+05	5.19e+05

Omnibus:	27.108	Durbin-Watson:	1.919
Prob(Omnibus):	0.000	Jarque-Bera (JB):	112.999
Skew:	-0.874	Prob(JB):	2.90e-25
Kurtosis:	8.632	Cond. No.	4.57e+05

Notes:

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

[2] The condition number is large, 4.57e+05. This might indicate that there are strong multicollinearity or other numerical problems.