Multiple Linear Regression

Using Categorical Variables

Content

1. MLR using Categorical Independent Variables
2. Statistical Model Using Dummy Variables
3. MLR with Categorical Dummy Variables in Python
4. Changing the base category

Case Study – Predicting Restaurant Sales

Background

- A city-based association of restaurants and cafes records all sorts of transactions and descriptive data for the purpose of industry-level analysis. The association wishes to find out if this data can be used to determine sales of restaurants.

Objective

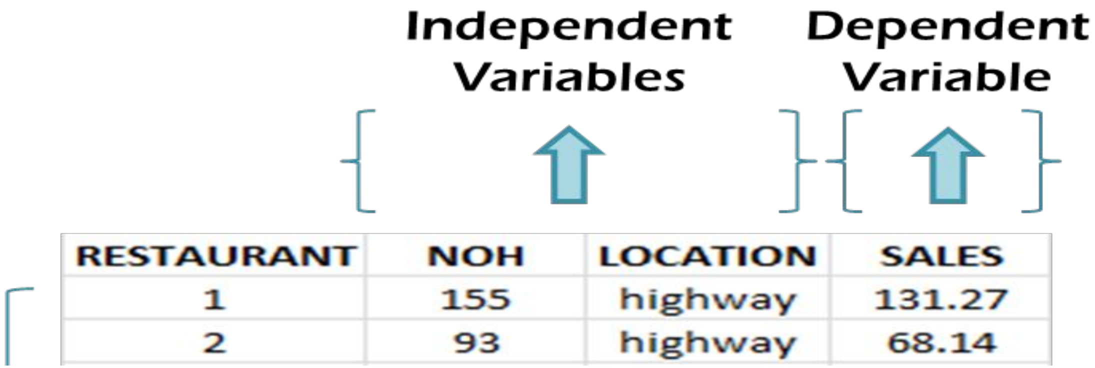
- To predict sales of restaurants

Available Information

- Sample size is 16
- Independent Variables: **Location of the Restaurant** – Categorical Variable with 3 Categories – Mall, Street and Highway and **Number of Households in the Area**
- Dependent Variable: **Sales of the Restaurant**

Data Snapshot

RESTAURANT SALES DATA



Independent Variables

Dependent Variable

RESTAURANT	NOH	LOCATION	SALES
1	155	highway	131.27
2	93	highway	68.14

Columns	Description	Type	Measurement	Possible values
RESTAURANT	Restaurant Number	numeric	-	-
NOH	Number of Households in the Vicinity of the Restaurant	numeric	-	positive values
LOCATION	Whether the Restaurant is Situated in a Mall, on a Street or on a Highway	Categorical	highway, mall, street	3
SALES	Annual Sales of the Restaurant	numeric	-	positive values

MLR using Categorical Independent Variables

If there is a categorical independent variable with K categories we must define only K – 1 dummy variables

In the case study,

Dependent Variable	Sales of a Restaurant
Independent Variables	1. Location of the Restaurant 2. Number of Households in the Area

- Location is a categorical variable with 3 categories – Mall, Street and Highway

  Therefore, there will be 3-1=2 dummy variables
- The category for which dummy variable is not defined is called ‘Base Category’

Data Snapshot – With Dummy Variables

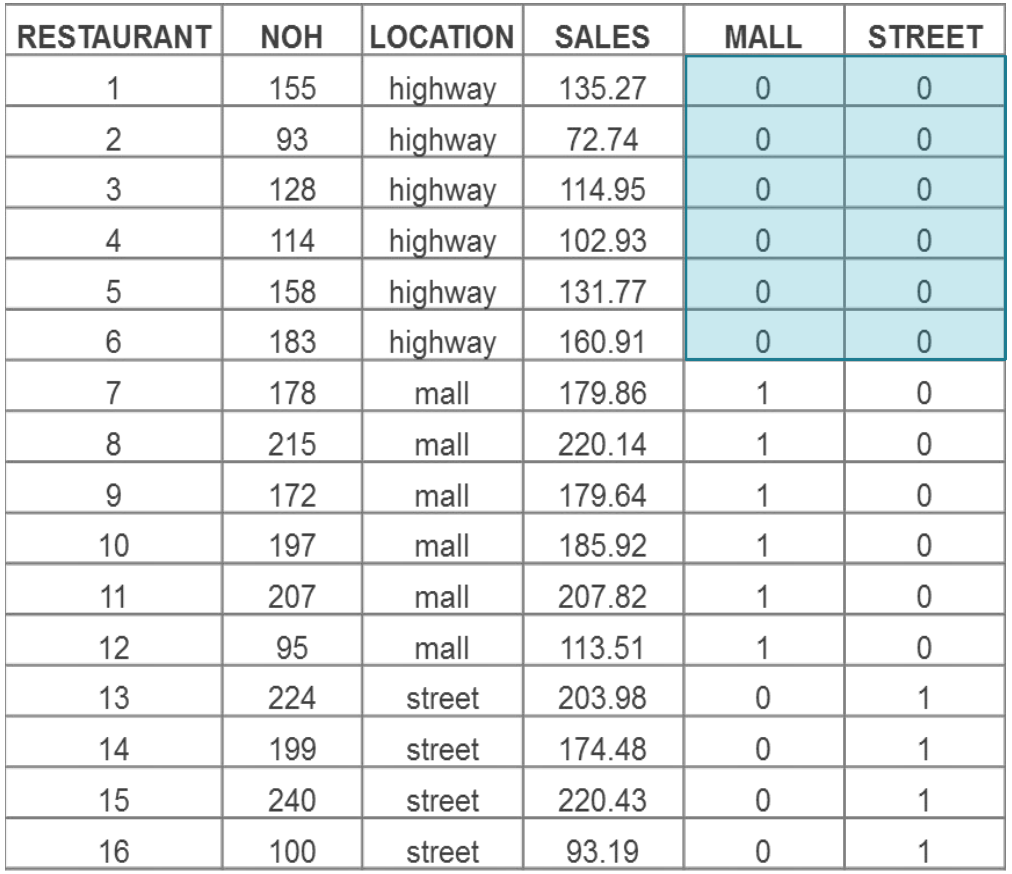
Y : Sales of a Restaurant

X1 : No. of Households

X2 : 1 if location is 'Mall' and 0 otherwise

X3 : 1 if location is 'Street' and 0 otherwise

- The first six rows have 0 under MALL and STREET columns
- This clearly implies that the restaurant is located neither in mall nor on street
- It is located on HIGHWAY



RESTAURANT	NOH	LOCATION	SALES	MALL	STREET
1	155	highway	135.27	0	0
2	93	highway	72.74	0	0
3	128	highway	114.95	0	0
4	114	highway	102.93	0	0
5	158	highway	131.77	0	0
6	183	highway	160.91	0	0
7	178	mall	179.86	1	0
8	215	mall	220.14	1	0
9	172	mall	179.64	1	0
10	197	mall	185.92	1	0
11	207	mall	207.82	1	0
12	95	mall	113.51	1	0
13	224	street	203.98	0	1
14	199	street	174.48	0	1
15	240	street	220.43	0	1
16	100	street	93.19	0	1

Why Not K Dummy Variables?

Can there be as many dummy variables as categories?

- If k dummy variables are created for k categories, there will be perfect multicollinearity – The Dummy Variable Trap
- In order to avoid falling into this trap, model with k categories and k dummy variables must have no intercept
- In such a model, coefficients will directly represent mean value of that variable

However, it is desirable to stick to the rule of k categories = k – 1 Dummy Variables

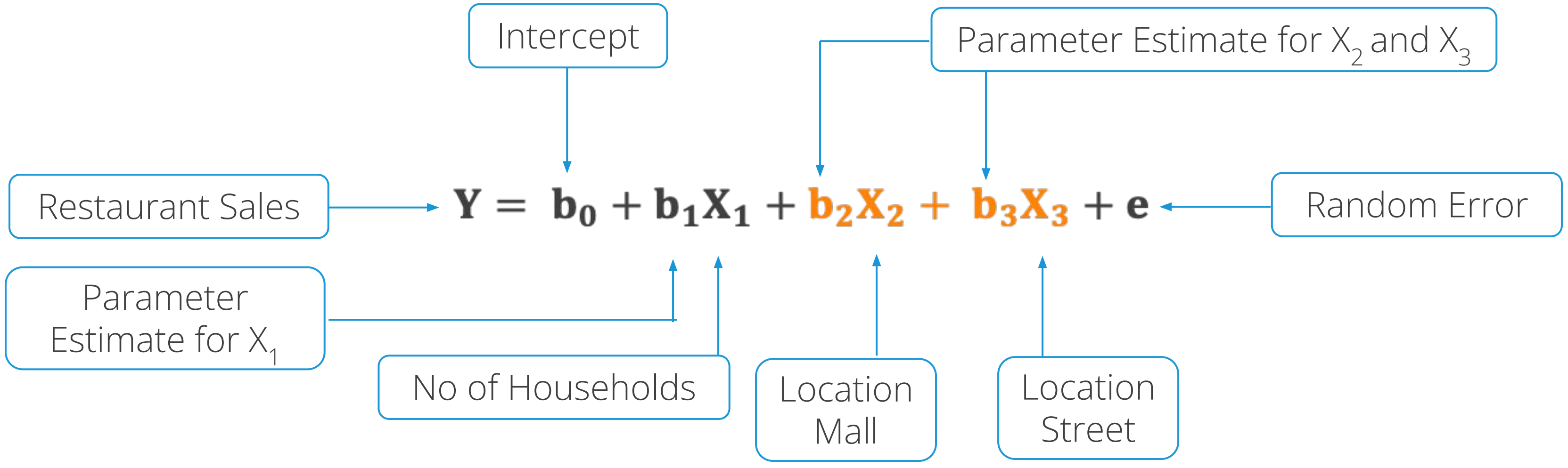
Statistical Model Using Dummy Variables

Basic Multiple Linear Regression Model

$$\boxed{Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p + e}$$

where,

Y	:	Dependent Variable
$X_1, X_2, \ldots, X_p$	:	Independent Variables
$b_0, b_1, \ldots, b_p$	:	Parameters of Model
e	:	Random Error Component



Intercept

Parameter Estimate for X_2 and X_3

Restaurant Sales

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e$$

Random Error

Parameter Estimate for X_1

No of Households

Location Mall

Location Street

Interpretation of Results

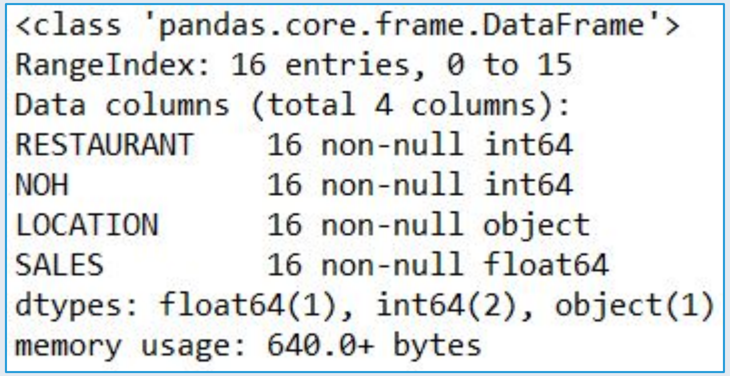
Regression Coefficients of categorical dummy variables are interpreted **relative to the base category**

- The positive beta coefficient of mall (b_2) implies that if the restaurant is located in a mall, the sale amount will be higher than sale amount of restaurant on highway by $b_{2.}$ units.
- If the coefficient is negative b_2 then it implies that restaurant located in mall will have lower sales than restaurant on highway by b_2 units.
- The same applies to street v/s highway
- Remember, dummy variable inferences are useful only if the variable is significant

MLR with Categorical Dummy Variables in Python

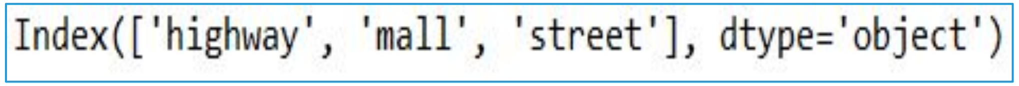
```
#Importing the Data
```

```
import pandas as pd
restaurantsales = pd.read_csv("RESTAURANT SALES DATA.csv")
restaurantsales.info()
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16 entries, 0 to 15
Data columns (total 4 columns):
RESTAURANT    16 non-null int64
NOH           16 non-null int64
LOCATION      16 non-null object
SALES         16 non-null float64
dtypes: float64(1), int64(2), object(1)
memory usage: 640.0+ bytes
```

```
restaurantsales['LOCATION']=restaurantsales['LOCATION'].astype('category')
restaurantsales['LOCATION'].cat.categories
```



```
Index(['highway', 'mall', 'street'], dtype='object')
```

- ❑ *info() shows class and levels of variables in the data.*
- ❑ *.astype is used to convert object type into "category"*
- ❑ *cat.categories() is used to check categorical variable's levels and their order.*

MLR with Categorical Dummy Variables in Python

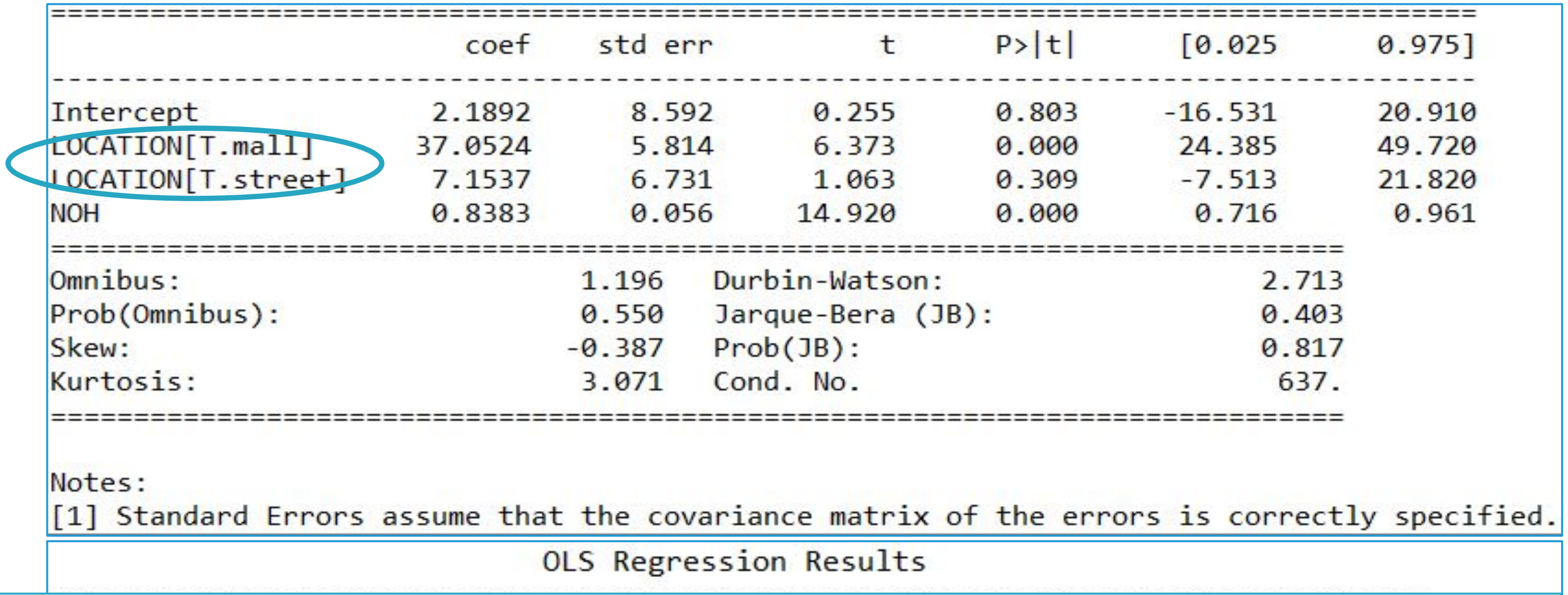
```
#Fitting Multiple Linear Regression Model
import statsmodels.formula.api as smf
salesmodel=smf.ols('SALES~NOH+LOCATION',data=restaurantsales).fit()

salesmodel.summary()
```

summary() generates a detailed description of the model.

MLR with Categorical Dummy Variables in Python

#Output



```
==================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept             2.1892      8.592      0.255      0.803     -16.531      20.910
LOCATION[T.mall]     37.0524      5.814      6.373      0.000      24.385      49.720
LOCATION[T.street]    7.1537      6.731      1.063      0.309      -7.513      21.820
NOH                   0.8383      0.056     14.920      0.000       0.716       0.961
==============================================================================
Omnibus:                        1.196   Durbin-Watson:                   2.713
Prob(Omnibus):                  0.550   Jarque-Bera (JB):                0.403
Skew:                          -0.387   Prob(JB):                        0.817
Kurtosis:                       3.071   Cond. No.                         637.
==============================================================================

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

Interpretation:

- *Python orders factor levels alphabetically and takes the first level as the base category by default*
- *NOH and Mall are significant variables*
- *Beta coefficient for mall is 37.0524. This implies that if a restaurant is located in a mall, its sales will be more than the restaurant located on highway by 37.0524.*
- *Street too has a positive coefficient, implying that sales of restaurant located on street will be 7.1537 times higher than highway.*

Changing the Base Category in Python

```
#Importing the required library
```

```
from patsy.contrasts import Treatment
```

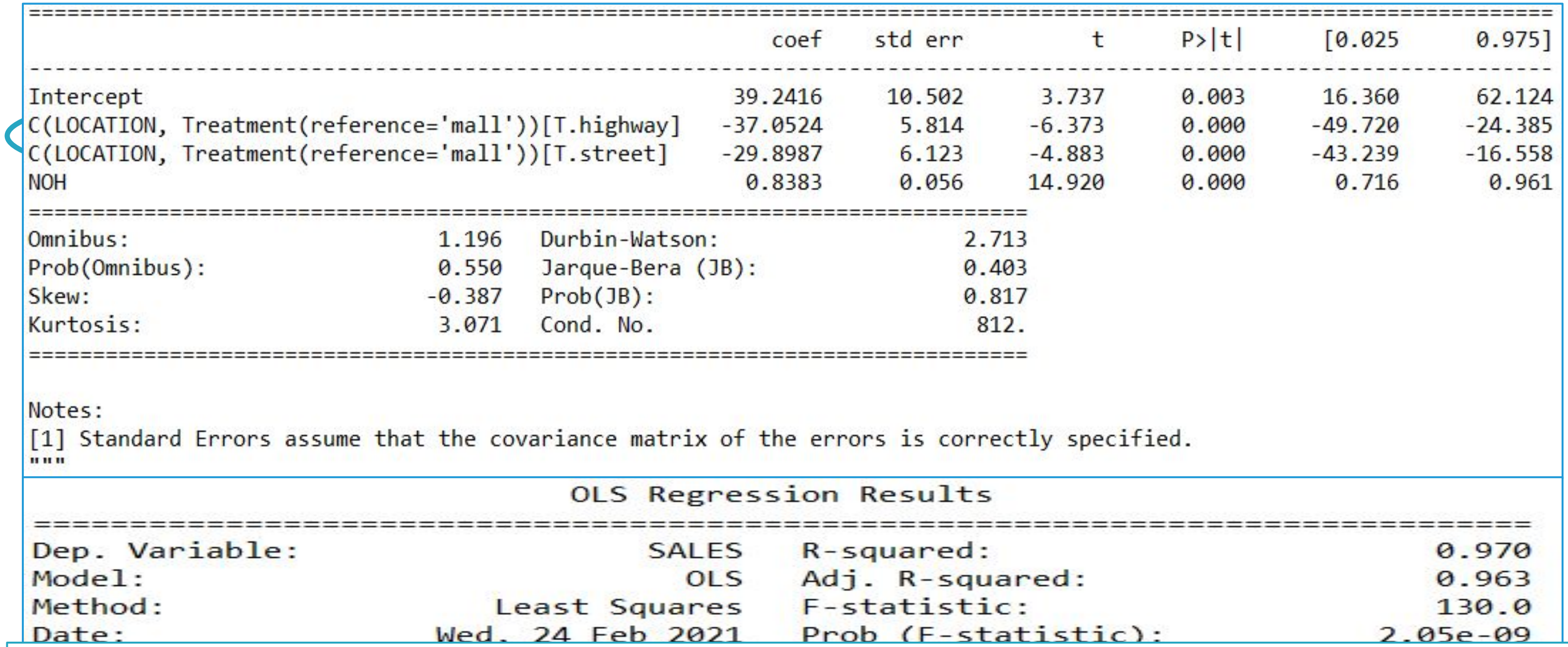
```
#Fitting Model on Data with Reordered Levels
```

```
salesmodel=smf.ols("SALES ~ C(LOCATION, Treatment(reference='mall')) +
NOH",data=restaurantsales).fit()
salesmodel.summary()
```

- ❑ *Treatment() reorders levels of a factor variables.*
- ❑ *reference= is used to specify changed reference (base) level.*

Changing the Base Category in Python

```
#Output
```



```
==========================================================================================================
                                                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
Intercept                                         39.2416     10.502      3.737      0.003      16.360      62.124
C(LOCATION, Treatment(reference='mall'))[T.highway]  -37.0524      5.814     -6.373      0.000     -49.720     -24.385
C(LOCATION, Treatment(reference='mall'))[T.street]   -29.8987      6.123     -4.883      0.000     -43.239     -16.558
NOH                                                0.8383      0.056     14.920      0.000       0.716       0.961
==============================================================================
Omnibus:                        1.196   Durbin-Watson:                   2.713
Prob(Omnibus):                  0.550   Jarque-Bera (JB):                0.403
Skew:                          -0.387   Prob(JB):                        0.817
Kurtosis:                       3.071   Cond. No.                         812.
==============================================================================

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

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  SALES   R-squared:                       0.970
Model:                            OLS   Adj. R-squared:                  0.963
Method:                 Least Squares   F-statistic:                     130.0
Date:                Wed, 24 Feb 2021   Prob (F-statistic):           2.05e-09
```

Interpretation:

- *Mall has now become the base category*
- *Coefficient of highway is just negative of coefficient of mall observed in previous model (Degree of association between mall and highway is the same, changed sign indicates that relativity has reversed)*
- *Note that street is also significant*

Quick Recap

In this session, we learnt how to **handle categorical variables in multiple linear regression by introducing Dummy Variables**

Number of Dummy Variables	• The number of dummy variables must be one less than the number of levels in the categorical variable
Interpretation	• The coefficient attached to the dummy variables must always be interpreted in relation to the base or reference group.
Dummy Variables in Python	• Python automatically assigns dummies to categorical variables in **`ols()`** • Use **`Treatment()`** to change the base category for modeling