Multiple Linear Regression

Using Categorical Variables

Content

1. Categorical Variables
2. Dummy Variables
3. MLR using Categorical Independent Variables
4. Changing the base category

Categorical Variables (Nominal Scale)

- Categorical variables with a nominal scale define sub-groups in the data.
  Example: Region, department etc.

- Even if these variables are represented using numerical values such as 1,2,3,4 etc, the numbers do not have any mathematical meaning.

- A Regression model can include these variables in the model

What Are Dummy Variables?

- Regression analysis requires numerical variables.
- So when there are categorical variables in a regression model , we create dummy variables for them.
- A dummy variable is a binary variable which takes the values 1 or zero.
- Dummy variables are useful because they enable us to use a single regression equation to represent multiple groups. This means that we do not need to write out separate equation models for each subgroup.

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

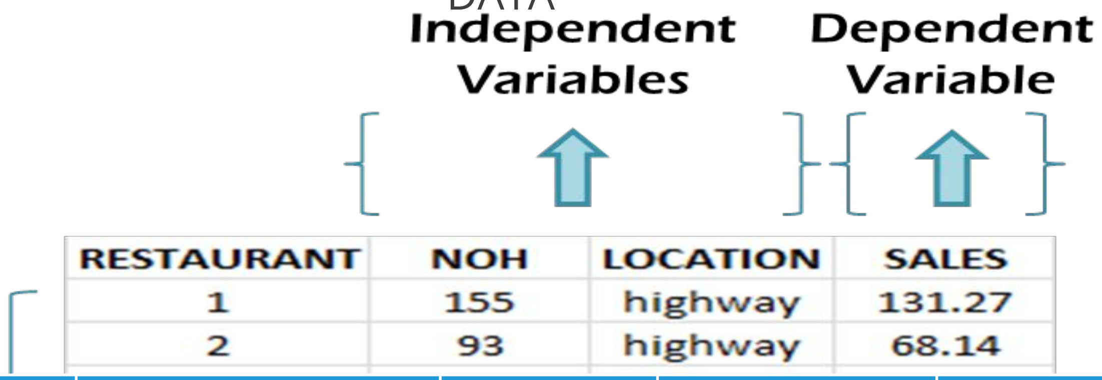
- To predict sales of restaurants

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	mall, street, highway	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

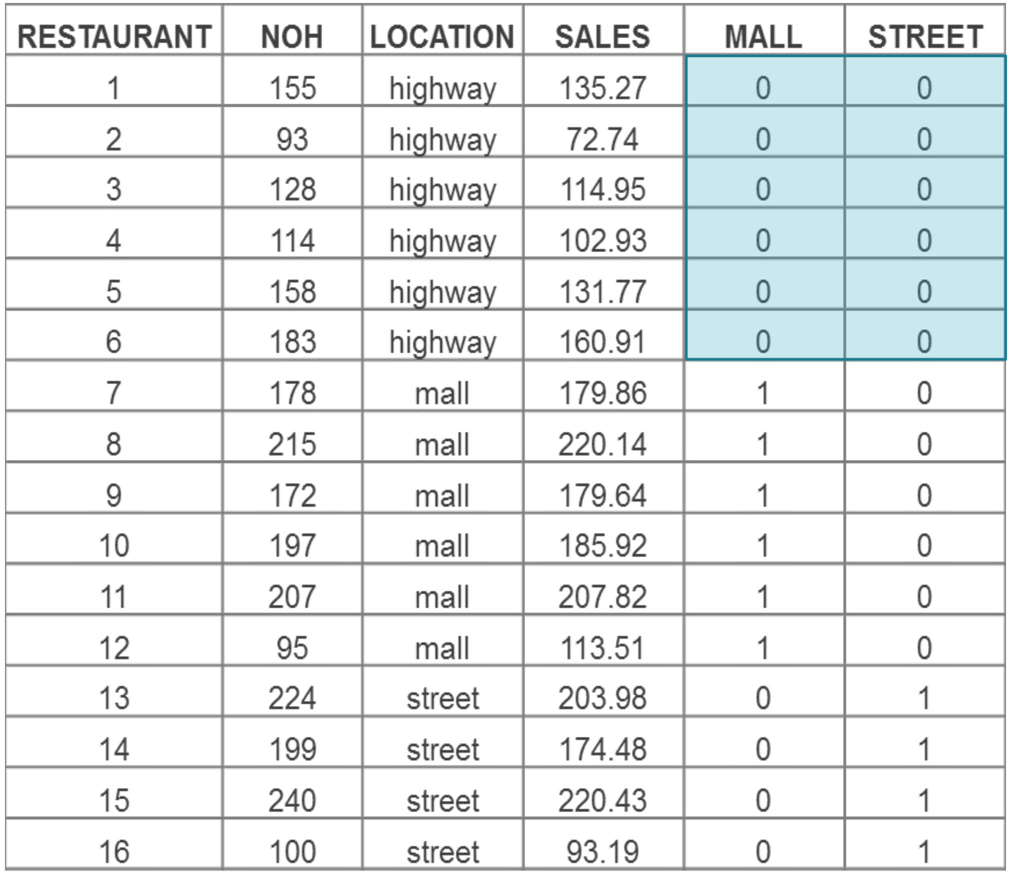
Y : Sales of a Restaurant

X1 : No. of Households

X2 : 1 if location is ‘Mall’
and 0 otherwise

X3 : 1 if location is ‘Street’ and
0 otherwise

- The location is “highway” if MALL=0 and STREET=0



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

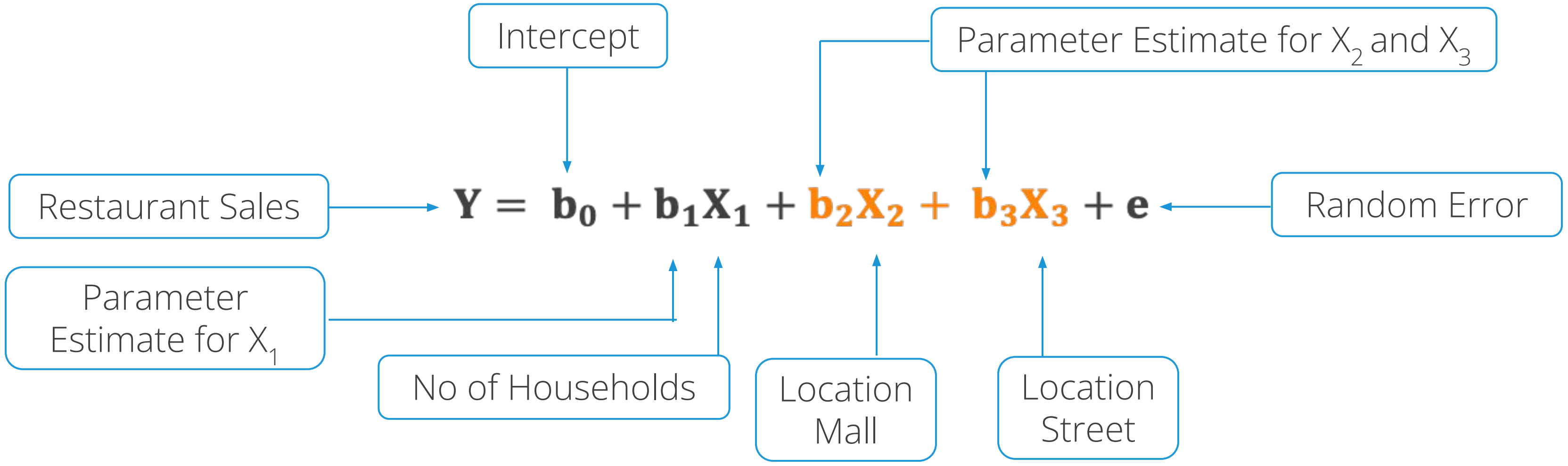
Statistical Model Using Dummy Variables

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

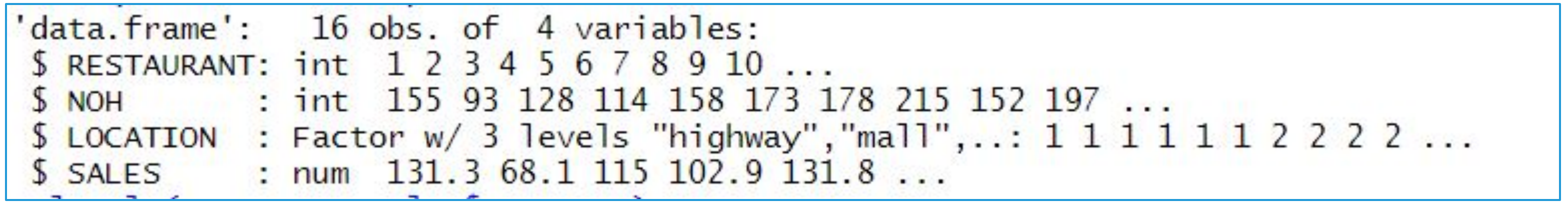
MLR with Categorical Dummy Variables in R

```
#Importing the Data
```

```
restaurantsales<-read.csv("RESTAURANT SALES DATA.csv",header=TRUE)

str(restaurantsales)
```

- *str() shows class and levels of variables in the data.*



```
'data.frame':   16 obs. of  4 variables:
 $ RESTAURANT: int  1 2 3 4 5 6 7 8 9 10 ...
 $ NOH       : int  155 93 128 114 158 173 178 215 152 197 ...
 $ LOCATION  : Factor w/ 3 levels "highway","mall",..: 1 1 1 1 1 1 2 2 2 2 ...
 $ SALES     : num  131.3 68.1 115 102.9 131.8 ...
```

```
levels(restaurantsales$LOCATION)

[1] "highway" "mall"    "street"
```

- *levels() to check categorical variable's levels and their order*

```
#Fitting Multiple Linear Regression Model
```

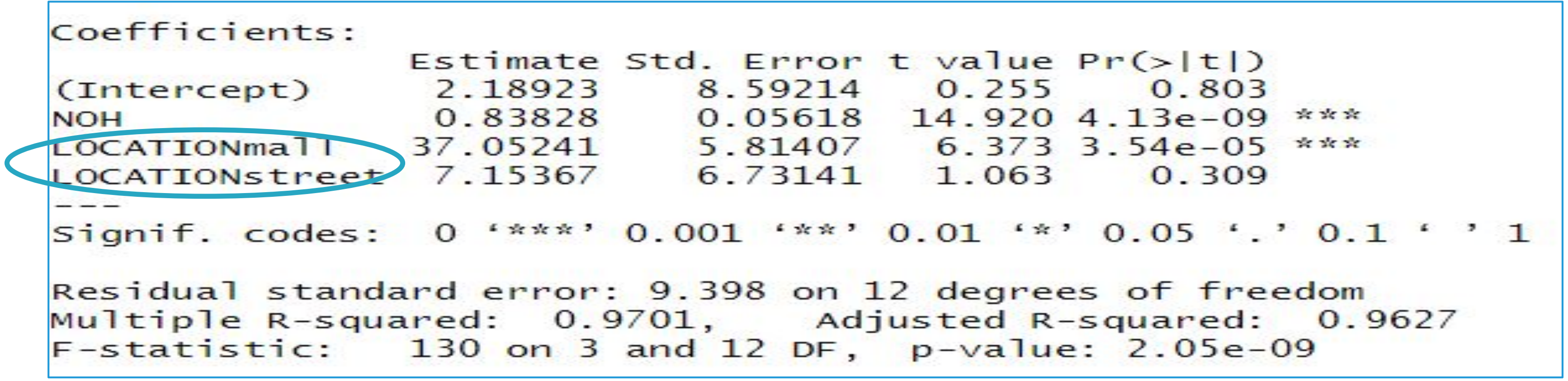
```
salesmodel<-lm(SALES~NOH+LOCATION,data=restaurantsales)

summary(salesmodel)
```

summary() generates a detailed description of the model.

MLR with Categorical Dummy Variables in R

#Output



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2.18923    8.59214   0.255    0.803    
NOH             0.83828    0.05618  14.920 4.13e-09 ***
LOCATIONmall   37.05241    5.81407   6.373 3.54e-05 ***
LOCATIONstreet  7.15367    6.73141   1.063    0.309    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.398 on 12 degrees of freedom
Multiple R-squared:  0.9701,    Adjusted R-squared:  0.9627 
F-statistic:   130 on 3 and 12 DF,  p-value: 2.05e-09
```

Interpretation:

- *R 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 37.05241. This implies that if a restaurant is located in a mall, its sales will be more than the restaurant located on highway by 37.05241.*
- *Street too has a positive coefficient, implying that sales of restaurant located on street will be 7.15367 times higher than highway.*

Changing the Base Category in R

```
#Changing the Base Category to “mall”
restaurantsales$LOCATION<-relevel(restaurantsales$LOCATION,ref="mall")
```

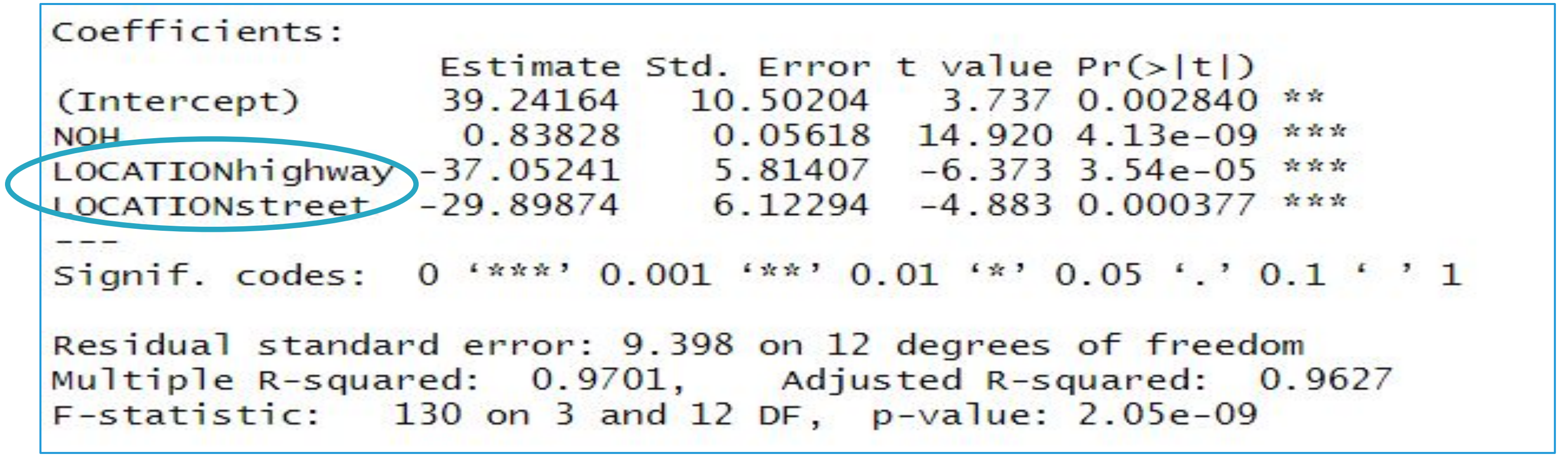
relevel() reorders levels of a factor variables.
ref= is used to specify changed reference (base) level.

```
#Fitting Model on Data with Reordered Levels
salesmodel2<-lm(SALES~NOH+LOCATION, data=restaurantsales)

summary(salesmodel2)
```

Changing the Base Category in R

#Output



```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      39.24164   10.50204   3.737 0.002840 ** 
NOH               0.83828    0.05618  14.920 4.13e-09 ***
LOCATIONhighway -37.05241    5.81407  -6.373 3.54e-05 ***
LOCATIONstreet  -29.89874    6.12294  -4.883 0.000377 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.398 on 12 degrees of freedom
Multiple R-squared:  0.9701,	Adjusted R-squared:  0.9627 
F-statistic:   130 on 3 and 12 DF,  p-value: 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—that is, the group that receives the value of zero. The base chosen will depend on the purpose of research at hand

Dummy Variables v/s Data Observations

- If a model has several qualitative variables with several classes, introduction of dummy variables can consume a large number of degrees of freedom. Therefore, one should always weigh the number of dummy variables to be introduced against the total number of observations available for analysis

Dummy Variables in R

- R automatically assigns dummies to categorical variables in **`lm()`**
- Use **`relevel()`** to change the base category for modeling