19MIS1018_LAB-4_ML_Multiple Linear Regression

August 18, 2022

Name : B DEVI PRASAD

Reg No: 19MIS1018

Slot : L13+L14

Faculty: Dr. G. Bharadwaja Kumar

1 Multiple Linear Regression

[1]:
```
import pandas as pd
import numpy as np
import statsmodels.api as sm
df_adv = pd.read_csv('Advertising.csv', index_col=0)
```

[2]:
```
X = df_adv[['TV', 'Radio']]
y = df_adv['Sales']
df_adv.head()
```

[2]:

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

[3]:
```
X = df_adv[['TV', 'Radio']]
y = df_adv['Sales']
## fit a OLS model with intercept on TV and Radio
X = sm.add_constant(X)
est = sm.OLS(y, X).fit()
est.summary()
```

[3]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.897
Model:                            OLS   Adj. R-squared:                  0.896
```

```
Method:                 Least Squares   F-statistic:                     859.6
Date:                Thu, 18 Aug 2022   Prob (F-statistic):           4.83e-98
Time:                        18:51:59   Log-Likelihood:                -386.20
No. Observations:                 200   AIC:                             778.4
Df Residuals:                     197   BIC:                             788.3
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.9211      0.294      9.919      0.000       2.340       3.502
TV             0.0458      0.001     32.909      0.000       0.043       0.048
Radio          0.1880      0.008     23.382      0.000       0.172       0.204
==============================================================================
Omnibus:                       60.022   Durbin-Watson:                   2.081
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              148.679
Skew:                          -1.323   Prob(JB):                     5.19e-33
Kurtosis:                       6.292   Cond. No.                         425.
==============================================================================

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

2 Handling Categorical Variables(Sales and Newspaper as independent variables)

[5]:
```
import statsmodels.formula.api as smf
# formula: response ~ predictor + predictor
est = smf.ols(formula='Sales ~ TV + Radio', data=df_adv).fit()
```

[6]:
```
import pandas as pd
df = pd.read_csv('Advertising.csv', index_col=0)# copy data and separate␣
 ↪predictors and response
X = df.copy()
y = X.pop('Newspaper')
df.head()
```

[6]:

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

[7]:
```
y.groupby(X.Sales).mean()
```

[7]:
```
Sales
1.6      8.7
3.2      5.7
4.8      1.0
5.3     17.5
5.5     41.4
        …
24.7     3.2
25.4    33.5
25.5    66.2
26.2    71.8
27.0    41.8
Name: Newspaper, Length: 121, dtype: float64
```

[8]:
```
import statsmodels.formula.api as smf
# encode df.famhist as a numeric via pd.Factor
#df['Sales'] = pd.Categorical((df.Sales).labels
est = smf.ols(formula="Newspaper ~ Sales", data=df_adv).fit()
```

[13]:
```
# fit OLS on categorical variables children and occupation
est = smf.ols(formula='Sales ~ Newspaper', data=df).fit()
#short_summary(est)
```

[14]:
```
est = sm.OLS(y, X).fit()
est.summary()
```

[14]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:              Newspaper   R-squared (uncentered):
0.683
Model:                            OLS   Adj. R-squared (uncentered):
0.678
Method:                 Least Squares   F-statistic:
141.4
Date:                Thu, 18 Aug 2022   Prob (F-statistic):
6.89e-49
Time:                        18:52:44   Log-Likelihood:
-893.75
No. Observations:                 200   AIC:
1794.
Df Residuals:                     197   BIC:
1803.
```

```
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
TV            -0.0388      0.042     -0.912      0.363      -0.123       0.045
Radio          0.3594      0.192      1.875      0.062      -0.019       0.737
Sales          1.8528      0.736      2.517      0.013       0.401       3.304
==============================================================================
Omnibus:                       10.613   Durbin-Watson:                   1.952
Prob(Omnibus):                  0.005   Jarque-Bera (JB):               12.125
Skew:                           0.432   Prob(JB):                      0.00233
Kurtosis:                       3.841   Cond. No.                         86.6
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

3 TV and Newspaper as independent variables

[27]:
```
import statsmodels.formula.api as smf
# formula: response ~ predictor + predictor
est = smf.ols(formula='TV ~ Sales + Radio', data=df_adv).fit()
```

[28]:
```
import pandas as pd
df = pd.read_csv('Advertising.csv', index_col=0)# copy data and separate␣
 ↪predictors and response
X = df.copy()
y = X.pop('TV')
df.head()
```

[28]:

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

[29]:
```
y.groupby(X.Newspaper).mean()
```

[29]:
```
Newspaper
0.3      265.6
```

```
0.9       69.0
1.0        8.6
1.7      184.9
1.8      293.6
         …
79.2     125.7
84.8     234.5
89.4      16.9
100.9    296.4
114.0     67.8
Name: TV, Length: 172, dtype: float64
```

[30]:
```
import statsmodels.formula.api as smf
# encode df.famhist as a numeric via pd.Factor
#df['Sales'] = pd.Categorical((df.Sales).labels
est = smf.ols(formula="TV ~ Newspaper", data=df_adv).fit()
```

[31]:
```
est = smf.ols(formula='Newspaper ~ TV', data=df).fit()
```

[32]:
```
est = sm.OLS(y, X).fit()
est.summary()
```

[32]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                     TV   R-squared (uncentered):
0.957
Model:                            OLS   Adj. R-squared (uncentered):
0.956
Method:                 Least Squares   F-statistic:
1458.
Date:                Thu, 18 Aug 2022   Prob (F-statistic):
3.49e-134
Time:                        19:05:51   Log-Likelihood:
-996.73
No. Observations:                 200   AIC:
1999.
Df Residuals:                     197   BIC:
2009.
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Radio         -3.3814      0.216    -15.651      0.000      -3.807      -2.955
```

```
Newspaper     -0.1086      0.119     -0.912      0.363      -0.343       0.126
Sales         16.5974      0.410     40.507      0.000      15.789      17.405
==============================================================================
Omnibus:                        3.564   Durbin-Watson:                   2.137
Prob(Omnibus):                  0.168   Jarque-Bera (JB):                3.153
Skew:                           0.283   Prob(JB):                        0.207
Kurtosis:                       3.240   Cond. No.                         8.14
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

4 Comparing results:

By Seeing the different combinations of the independent variable to detect the Sales as per the advertisment we got a model that uses the Newspaper and TV only as the independent variable and Sales newspaper . Multiple linear regression is a regression model that estimates the relationship between a quantitative dependent variable. 1. Newspaper Vs Sales. Adjusted R-squared: 0.897 Predicted R-squared:0.683

2. TV vs Newspaper: Adjusted R-squared: 0.897 Predicted R-squared:0.957