Marketing and Sales Data Multiple Linear Regression

August 7, 2023

1 Marketing and Sales Data Multiple Linear Regression

1.1 Introduction

1.2 Step 1: Imports

[1]:
```
# Import libraries and modules.
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from statsmodels.formula.api import ols
import statsmodels.api as sm
```

1.2.1 Load dataset

[2]:
```
#Import the data.
data = pd.read_csv('marketing_sales_data_2.csv')

# Display the first five rows.
data.head()
```

[2]:

	TV	Radio	Social Media	Influencer	Sales
0	Low	3.518070	2.293790	Micro	55.261284
1	Low	7.756876	2.572287	Mega	67.574904
2	High	20.348988	1.227180	Micro	272.250108
3	Medium	20.108487	2.728374	Mega	195.102176
4	High	31.653200	7.776978	Nano	273.960377

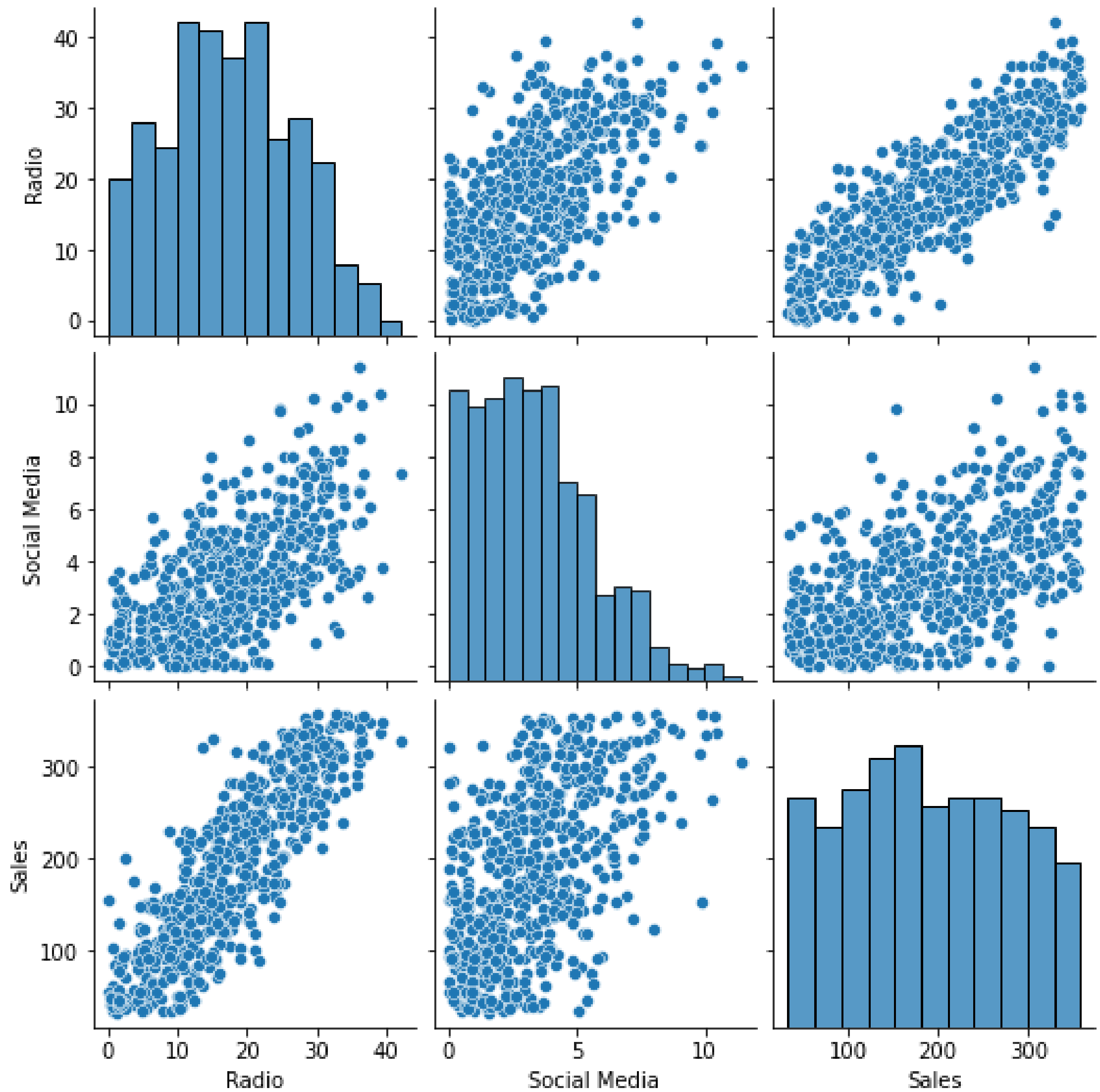
1.3 Step 2: Data exploration

1.3.1 Create a pairplot of the data

Create a pairplot to visualize the relationship between the continous variables in `data`.

[3]:
```python
# Create a pairplot of the data.
sns.pairplot(data)
```

[3]: <seaborn.axisgrid.PairGrid at 0x7f3342700cd0>



1.3.2 Calculate the mean sales for each categorical variable

[4]:
```python
# Calculate the mean sales for each TV category.
TV_sales=data.groupby("TV")["Sales"].mean()
print(TV_sales)

# Calculate the mean sales for each Influencer category.
```

```
Inf_sales=data.groupby("Influencer")["Sales"].mean()
print(Inf_sales)
```

```
TV
High      300.853195
Low        90.984101
Medium    195.358032
Name: Sales, dtype: float64
Influencer
Macro    181.670070
Mega     194.487941
Micro    188.321846
Nano     191.874432
Name: Sales, dtype: float64
```

1.3.3 Remove missing data

This dataset contains rows with missing values. To correct this, drop all rows that contain missing data.

[5]:
```
# Drop rows that contain missing data and update the DataFrame.
data.dropna(axis=0)
```

[5]:
```
         TV      Radio  Social Media Influencer       Sales
0       Low   3.518070      2.293790      Micro   55.261284
1       Low   7.756876      2.572287       Mega   67.574904
2      High  20.348988      1.227180      Micro  272.250108
3    Medium  20.108487      2.728374       Mega  195.102176
4      High  31.653200      7.776978       Nano  273.960377
..      ...        ...           ...        ...         ...
567  Medium  14.656633      3.817980      Micro  191.521266
568    High  28.110171      7.358169       Mega  297.626731
569  Medium  11.401084      5.818697       Nano  145.416851
570  Medium  21.119991      5.703028      Macro  209.326830
571     Low  13.221237      3.660566      Micro  135.773151

[572 rows x 5 columns]
```

1.3.4 Clean column names

The `ols()` function doesn't run when variable names contain a space.

[6]:
```
# Rename all columns in data that contain a space.
data.columns=["TV", "Radio", "Social_Media", "Influencer", "Sales"]
data.columns
```

[6]: Index(['TV', 'Radio', 'Social_Media', 'Influencer', 'Sales'], dtype='object')

1.4 Step 3: Model building

1.4.1 Fit a multiple linear regression model that predicts sales

[7]:
```python
# Define the OLS formula.
ols_formula="Sales ~ Radio + C(TV)"

# Create an OLS model.
OLS=ols(data=data, formula=ols_formula)

# Fit the model.
model=OLS.fit()

# Save the results summary.
result=model.summary()

# Display the model results.
result
```

[7]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.904
Model:                            OLS   Adj. R-squared:                  0.904
Method:                 Least Squares   F-statistic:                     1783.
Date:                Mon, 07 Aug 2023   Prob (F-statistic):         1.63e-288
Time:                        12:56:18   Log-Likelihood:                -2714.0
No. Observations:                 572   AIC:                             5436.
Df Residuals:                     568   BIC:                             5453.
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
===
                      coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
---
Intercept         218.5261      6.261     34.902      0.000     206.228
230.824
C(TV)[T.Low]     -154.2971      4.929    -31.303      0.000    -163.979
-144.616
C(TV)[T.Medium]   -75.3120      3.624    -20.780      0.000     -82.431
-68.193
```

```
Radio                  2.9669      0.212     14.015      0.000       2.551
3.383
==============================================================================
Omnibus:                       61.244   Durbin-Watson:                   1.870
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               18.077
Skew:                           0.046   Prob(JB):                     0.000119
Kurtosis:                       2.134   Cond. No.                         142.
==============================================================================

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

1.4.2 Check model assumptions

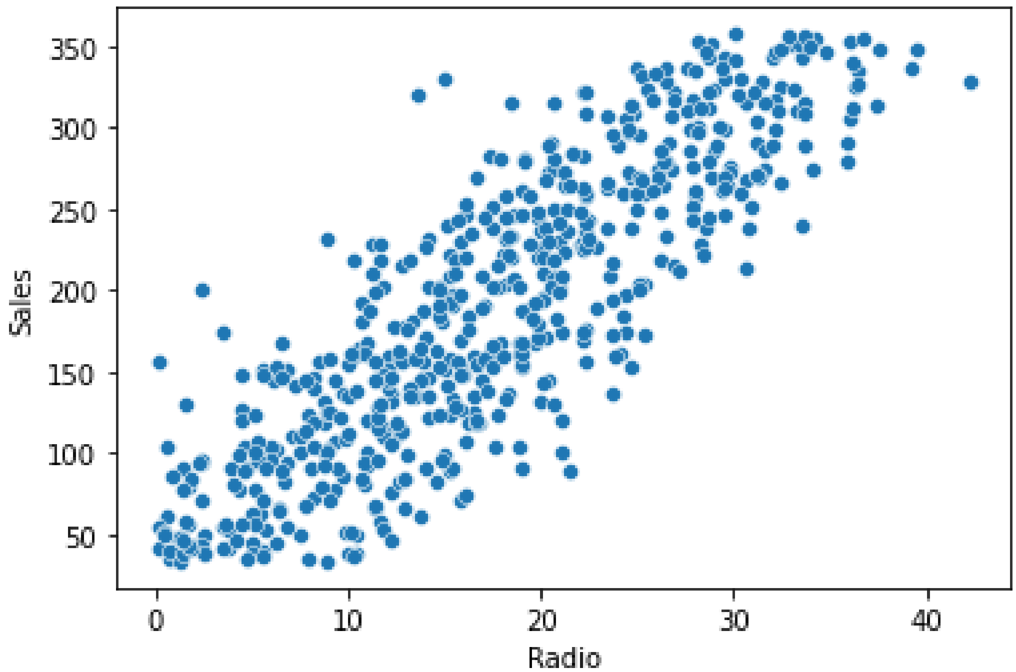
For multiple linear regression, there are five assumptions to be met: Linearity, Independent Observations, Normality, Homoscedasticity and Multicollinearity.

- Linearity requires a linear relationship between the independent and dependent variables.
- Independent Observations states that each observation in the dataset is independent.
- Normality states that the errors are normally distributed.
- Homoscedasticity is that the residuals have a constant variance for all values of independent variables.
- The no multicollinearity assumption states that no two independent variables (and) can be highly correlated with each other.

Check that all five multiple linear regression assumptions are upheld for the model.

1.4.3 Model assumption: Linearity

[8]:
```
# Create a scatterplot for continuous independent variable and the dependent␣
↪variable.
sns.scatterplot(data=data, x="Radio", y="Sales")
plt.show()
```



1.4.4 Model assumption: Independence

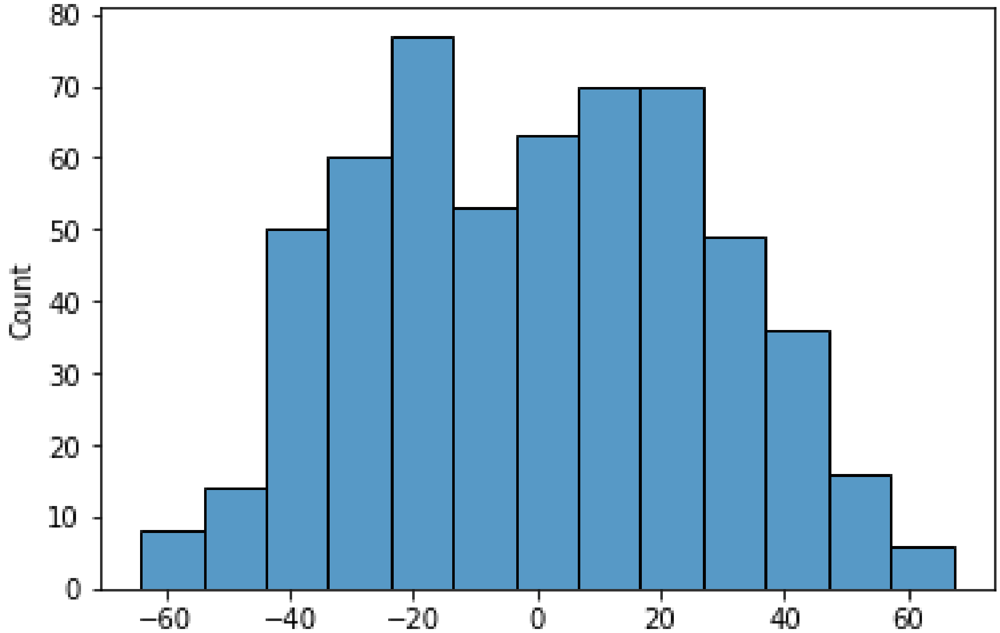
As each marketing promotion (i.e., row) is independent from one another, the independence assumption is not violated.

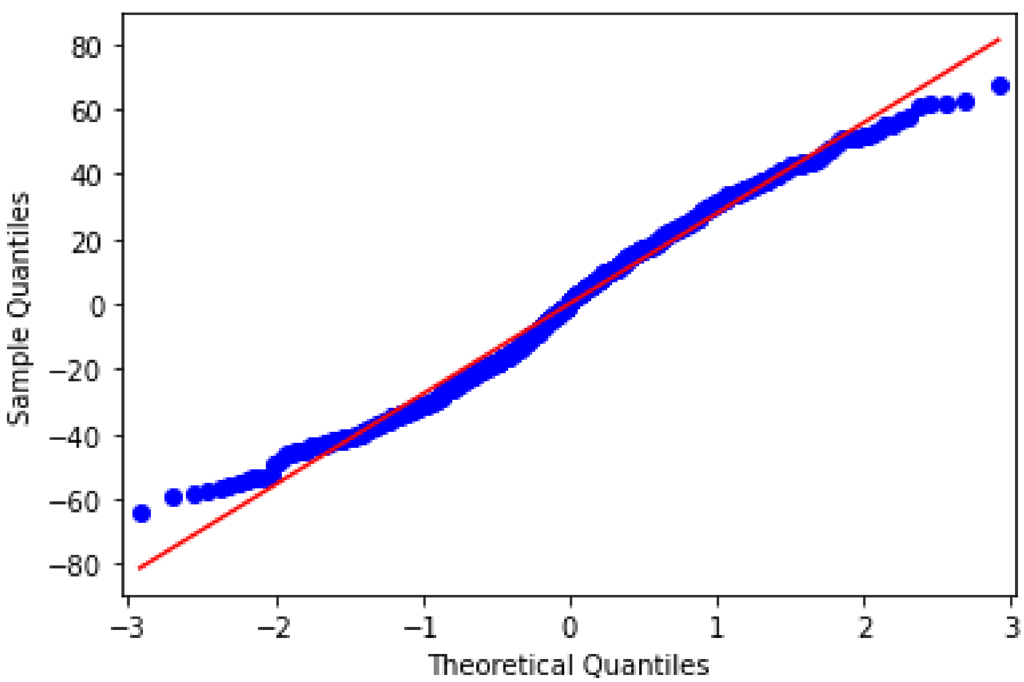
1.4.5 Model assumption: Normality

[9]:
```
# Calculate the residuals.
residuals=model.resid

# Create a histogram with the residuals.
sns.histplot(residuals)
plt.show()

# Create a Q-Q plot of the residuals.
sm.qqplot(data=residuals, line="s")
plt.show()
```

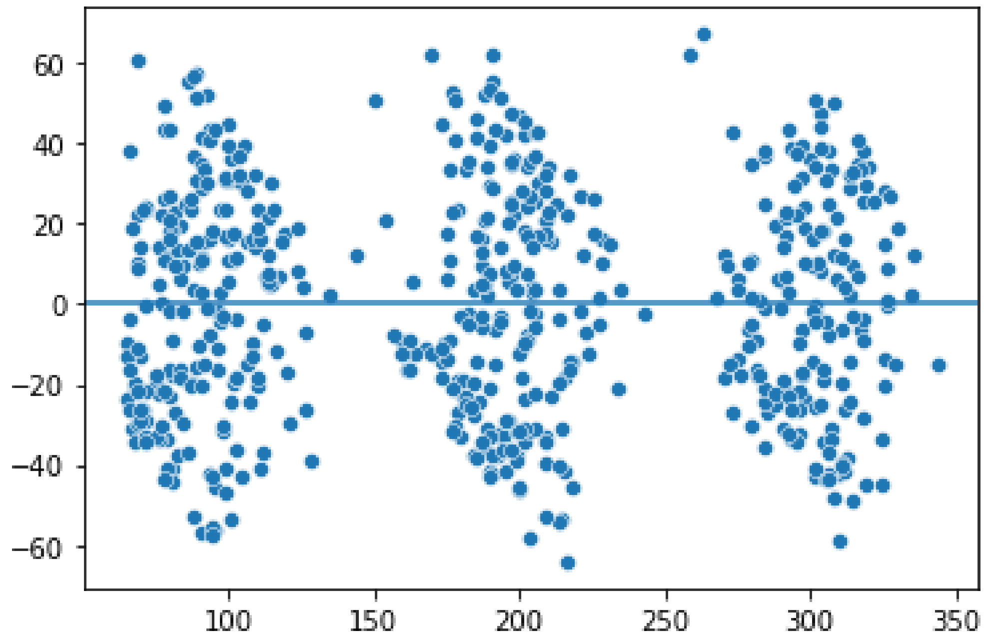

Count
80
70
60
50
40
30
20
10
0
−60
−40
−20
0
20
40
60


Sample Quantiles
80
60
40
20
0
−20
−40
−60
−80
−3
−2
−1
0
1
2
3
Theoretical Quantiles

1.4.6 Model assumption: Homoscedasticity-Constant variance

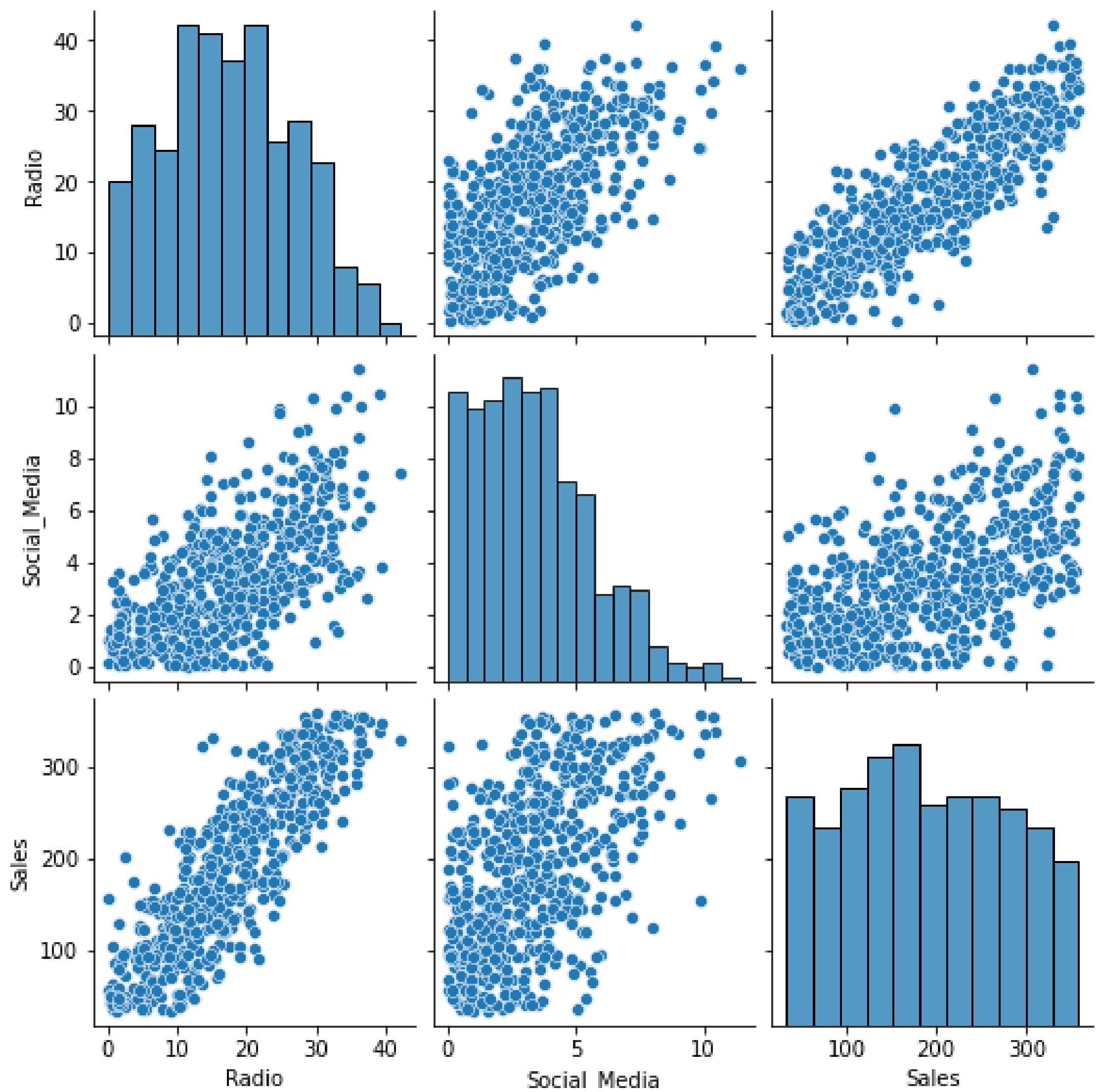
[10]:
```
# Create a scatterplot with the fitted values from the model and the residuals.
fig=sns.scatterplot(x=model.fittedvalues, y=residuals)

# Add a line at y = 0 to visualize the variance of residuals above and below 0.
fig.axhline(0)
plt.show()
```



1.4.7 Model assumption: No multicollinearity

[11]:
```
# Create a pairplot of the data.
sns.pairplot(data)
plt.show()
```



1.5 Step 4: Results and evaluation

1.5.1 Display the OLS regression results

[13]:
```
# Display the model results summary.
result
```

[13]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
```

```
Dep. Variable:                  Sales   R-squared:                       0.904
Model:                            OLS   Adj. R-squared:                  0.904
Method:                 Least Squares   F-statistic:                     1783.
Date:                Mon, 07 Aug 2023   Prob (F-statistic):          1.63e-288
Time:                        12:56:18   Log-Likelihood:                -2714.0
No. Observations:                 572   AIC:                             5436.
Df Residuals:                     568   BIC:                             5453.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
===
                      coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
---
Intercept          218.5261      6.261     34.902      0.000     206.228
230.824
C(TV)[T.Low]      -154.2971      4.929    -31.303      0.000    -163.979
-144.616
C(TV)[T.Medium]    -75.3120      3.624    -20.780      0.000     -82.431
-68.193
Radio                2.9669      0.212     14.015      0.000       2.551
3.383
==============================================================================
Omnibus:                       61.244   Durbin-Watson:                   1.870
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               18.077
Skew:                           0.046   Prob(JB):                     0.000119
Kurtosis:                       2.134   Cond. No.                         142.
==============================================================================

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

1.6 Conclusion

- Especially for the TV promotion spendings, it can be inferred that the higher the TV promotion spending, the higher the sale revenue. However, as for the categories of influencer, the sale in revenue is closer to each other and it cannot be established whether there is an upward (micro-mega) or downward (nano-micro) linearity among groups.
- Default TV spending is "High," and reducing it to "Low" results in an average decrease of 154 million in sales revenue. The medium spending category leads to an average decrease of 75.31 million compared to the "High" category.
- For every 1 million dollars spent on radio promotion, sales increase by approximately 2.96 million dollars.

- The model's R-squared value is close to 1 (0.90), indicating that sales can mostly be explained by TV and radio promotion spending. The high adjusted R-squared suggests no overfitting issues with the model.
- All independent variables have a statistically significant impact on sales, as indicated by their p-values of 0.000 and with the confidence level of 95%.
- Beta coefficients indicate the direction and magnitude of the effect of each independent variable on sales.
- It is recommended to allocate a high promotional budget to TV and invest in radio promotions to increase sales.