Multiple_Linear_Regression

September 11, 2024

1 Activity: Perform multiple linear regression

1.1 Introduction

As you have learned, multiple linear regression helps you estimate the linear relationship between one continuous dependent variable and two or more independent variables. For data science professionals, this is a useful skill because it allows you to compare more than one variable to the variable you're measuring against. This provides the opportunity for much more thorough and flexible analysis.

For this activity, you will be analyzing a small business' historical marketing promotion data. Each row corresponds to an independent marketing promotion where their business uses TV, social media, radio, and influencer promotions to increase sales. They previously had you work on finding a single variable that predicts sales, and now they are hoping to expand this analysis to include other variables that can help them target their marketing efforts.

To address the business' request, you will conduct a multiple linear regression analysis to estimate sales from a combination of independent variables. This will include:

- Exploring and cleaning data
- Using plots and descriptive statistics to select the independent variables
- Creating a fitting multiple linear regression model
- Checking model assumptions
- Interpreting model outputs and communicating the results to non-technical stakeholders

1.2 Step 1: Imports

1.2.1 Import packages

Import relevant Python libraries and modules.

[18]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

1.2.2 Load dataset

`Pandas` was used to load the dataset `marketing_sales_data.csv` as `data`, now display the first five rows. The variables in the dataset have been adjusted to suit the objectives of this lab. As shown in this cell, the dataset has been automatically loaded in for you. You do not need to download the .csv file, or provide more code, in order to access the dataset and proceed with this lab. Please continue with this activity by completing the following instructions.

[27]:
```
data = pd.read_csv('marketing_sales_data.csv')

data.head()
```

[27]:

	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 Familiarize yourself with the data's features

Start with an exploratory data analysis to familiarize yourself with the data and prepare it for modeling.

The features in the data are:

- TV promotional budget (in "Low," "Medium," and "High" categories)
- Social media promotional budget (in millions of dollars)
- Radio promotional budget (in millions of dollars)
- Sales (in millions of dollars)
- Influencer size (in "Mega," "Macro," "Micro," and "Nano" categories)

Question: What are some purposes of EDA before constructing a multiple linear regression model?

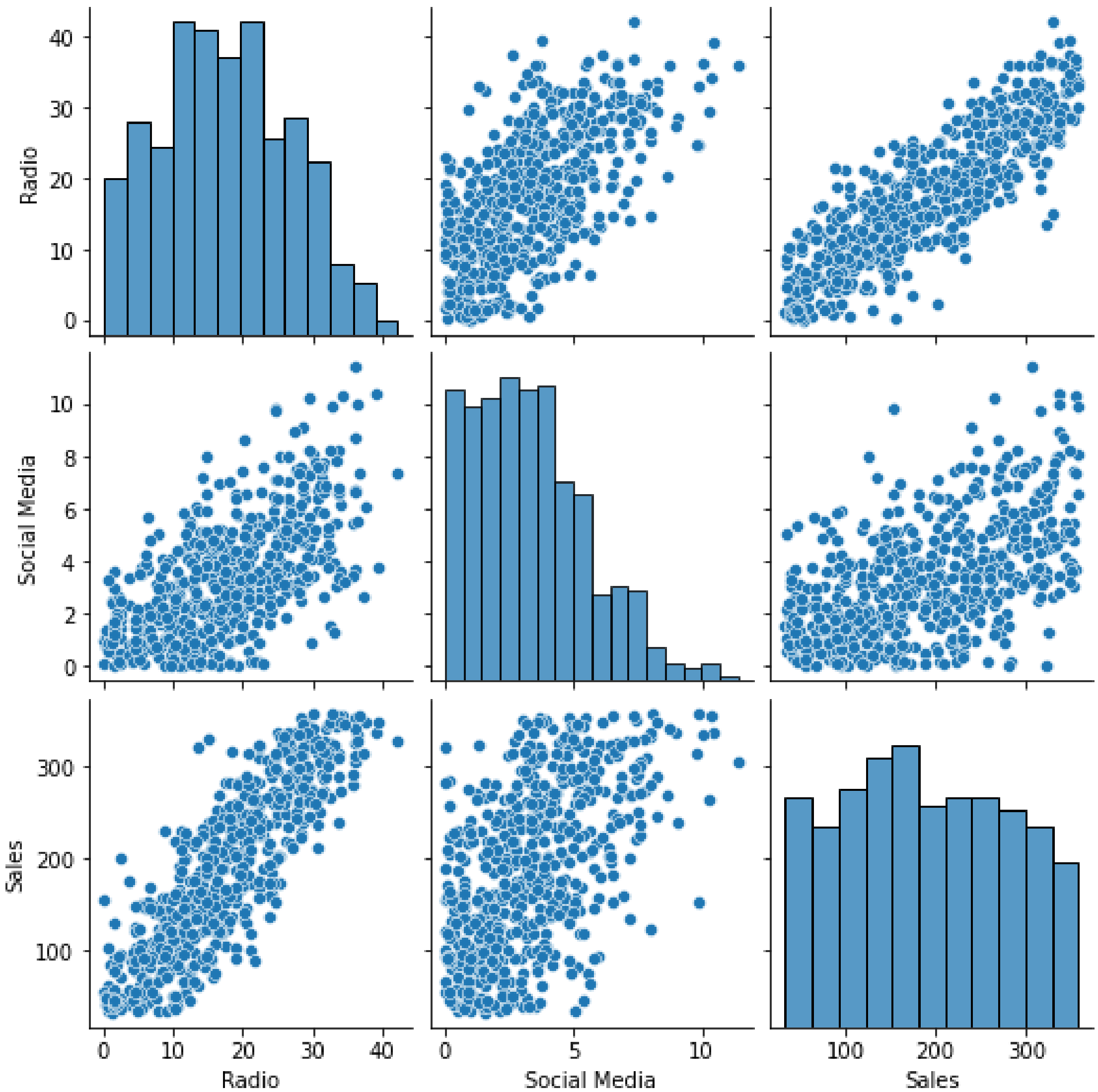
Potential reasons include:

- Understanding which variables are present in the data
- Reviewing the distribution of features, such as minimum, mean, and maximum values
- Plotting the relationship between the independent and dependent variables to visualize which features have a linear relationship
- Identifying issues with the data, such as incorrect values (e.g., typos) or missing values

1.3.2 Create a pairplot of the data

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

[28]:
```
sns.pairplot(data);
```



Question: Which variables have a linear relationship with `Sales`? Why are some variables in the data excluded from the preceding plot?

`Radio` and `Social Media` both appear to have linear relationships with `Sales`. Given this, `Radio` and `Social Media` may be useful as independent variables in a multiple linear regression model estimating `Sales`.

`TV` and `Influencer` are excluded from the pairplot because they are not numeric.

1.3.3 Calculate the mean sales for each categorical variable

There are two categorical variables: `TV` and `Influencer`. To characterize the relationship between the categorical variables and `Sales`, find the mean `Sales` for each category in `TV` and the mean `Sales` for each category in `Influencer`.

[29]:
```python
print(data.groupby('TV')['Sales'].mean())

print('')

print(data.groupby('Influencer')['Sales'].mean())
```

```
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
```

Question: What do you notice about the categorical variables? Could they be useful predictors of `Sales`?

The average `Sales` for `High` `TV` promotions is considerably higher than for `Medium` and `Low` `TV` promotions. `TV` may be a strong predictor of `Sales`.

The categories for `Influencer` have different average `Sales`, but the variation is not substantial. `Influencer` may be a weak predictor of `Sales`.

These results can be investigated further when fitting the multiple linear regression model.

1.3.4 Remove missing data

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

[14]:
```python
data.dropna(axis=0)
```

[14]:

	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]
```

1.3.5 Clean column names

The `ols()` function doesn't run when variable names contain a space. Check that the column names in `data` do not contain spaces and fix them, if needed.

[30]:
```python
data = data.rename(columns={'Social Media': 'Social_Media'})
```

1.4 Step 3: Model building

1.4.1 Fit a multiple linear regression model that predicts sales

Using the independent variables of your choice, fit a multiple linear regression model that predicts `Sales` using two or more independent variables from `data`.

[31]:
```python
ols_formula = 'Sales ~ C(TV) + Radio'

OLS = ols(formula = ols_formula, data = data)

model = OLS.fit()

model_results = model.summary()

model_results
```

[31]:
```
<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:                Tue, 13 Aug 2024   Prob (F-statistic):          1.63e-288
Time:                        16:41:54   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.
"""
```

Question: Which independent variables did you choose for the model, and why?

- `TV` was selected, as the preceding analysis showed a strong relationship between the `TV` promotional budget and the average `Sales`.
- `Radio` was selected because the pairplot showed a strong linear relationship between `Radio` and `Sales`.
- `Social Media` was not selected because it did not increase model performance and it was later determined to be correlated with another independent variable: `Radio`.
- `Influencer` was not selected because it did not show a strong relationship to `Sales` in the preceding analysis.

1.4.2 Check model assumptions

For multiple linear regression, there is an additional assumption added to the four simple linear regression assumptions: **multicollinearity**.

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

1.4.3 Model assumption: Linearity

Create scatterplots comparing the continuous independent variable(s) you selected previously with `Sales` to check the linearity assumption. Use the pairplot you created earlier to verify the linearity assumption or create new scatterplots comparing the variables of interest.

[32]:
```
fig, axes = plt.subplots(1, 2, figsize = (8,4))

sns.scatterplot(x = data['Radio'], y = data['Sales'],ax=axes[0])

axes[0].set_title("Radio and Sales")

sns.scatterplot(x = data['Social_Media'], y = data['Sales'],ax=axes[1])

axes[1].set_title("Social Media and Sales")

axes[1].set_xlabel("Social Media")

plt.tight_layout()
```

Question: Is the linearity assumption met?

The linearity assumption holds for `Radio`, as there is a clear linear relationship in the scatterplot between `Radio` and `Sales`. `Social Media` was not included in the preceding multiple linear regression model, but it does appear to have a linear relationship with `Sales`.

1.4.4 Model assumption: Independence

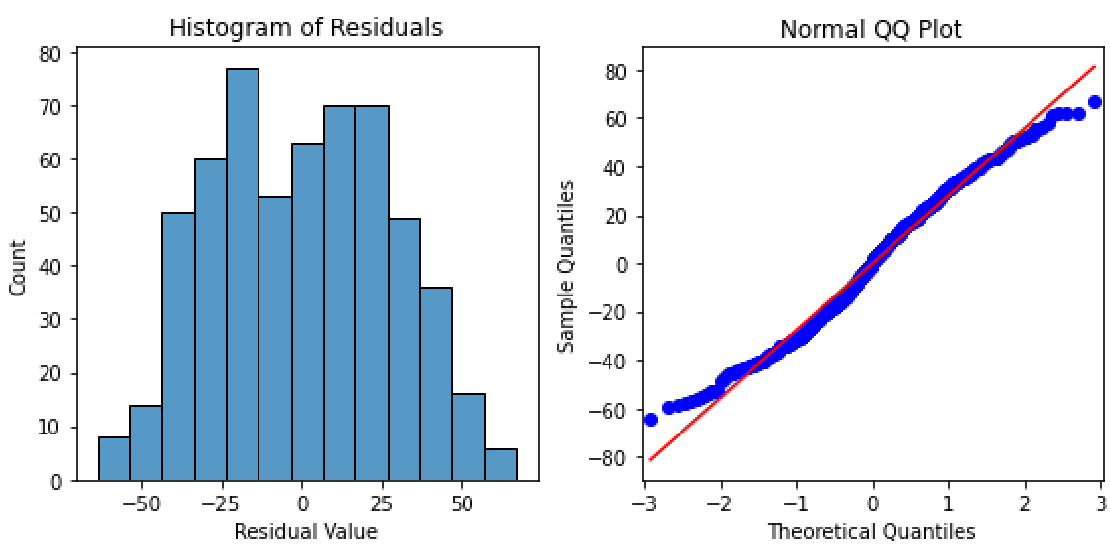
The **independent observation assumption** states that each observation in the dataset is independent. As each marketing promotion (i.e., row) is independent from one another, the independence assumption is not violated.

1.4.5 Model assumption: Normality

Create the following plots to check the **normality assumption**:

- **Plot 1**: Histogram of the residuals
- **Plot 2**: Q-Q plot of the residuals

[33]:
```
residuals = model.resid

fig, axes = plt.subplots(1, 2, figsize = (8,4))

sns.histplot(residuals, ax=axes[0])

axes[0].set_xlabel("Residual Value")

axes[0].set_title("Histogram of Residuals")

sm.qqplot(residuals, line='s',ax = axes[1])

axes[1].set_title("Normal QQ Plot")

plt.tight_layout()

plt.show()
```



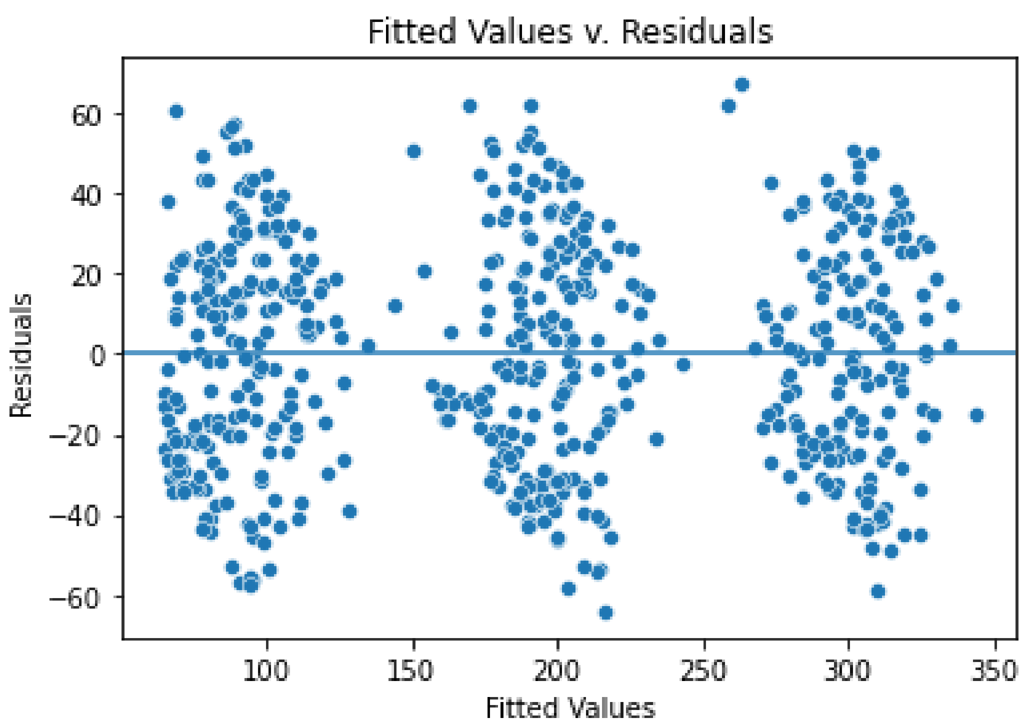
Question: Is the normality assumption met?

The histogram of the residuals are approximately normally distributed, which supports that the normality assumption is met for this model. The residuals in the Q-Q plot form a straight line, further supporting that this assumption is met.

1.4.6 Model assumption: Constant variance

Check that the **constant variance assumption** is not violated by creating a scatterplot with the fitted values and residuals. Add a line at $y = 0$ to visualize the variance of residuals above and below $y = 0$.

[34]:
```
fig = sns.scatterplot(x = model.fittedvalues, y = model.resid)

fig.set_xlabel("Fitted Values")

fig.set_ylabel("Residuals")

fig.set_title("Fitted Values v. Residuals")

fig.axhline(0)

plt.show()
```



Question: Is the constant variance assumption met?

The fitted values are in three groups because the categorical variable is dominating in this model, meaning that TV is the biggest factor that decides the sales.

However, the variance where there are fitted values is similarly distributed, validating that the assumption is met

1.4.7 Model assumption: No multicollinearity

The **no multicollinearity assumption** states that no two independent variables (X_i and X_j) can be highly correlated with each other.

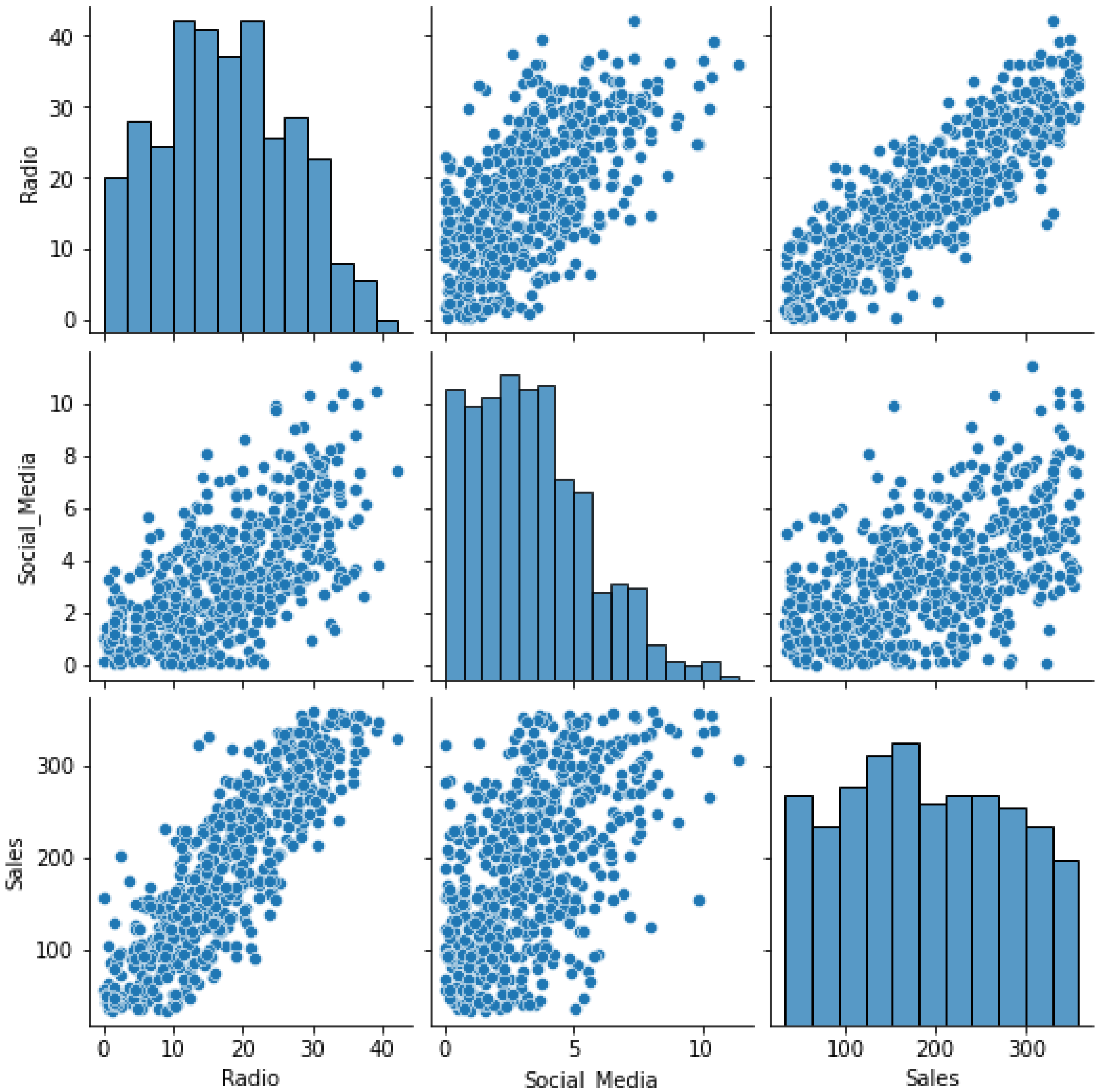
Two common ways to check for multicollinearity are to:

- Create scatterplots to show the relationship between pairs of independent variables
- Use the variance inflation factor to detect multicollinearity

Use one of these two methods to check your model's no multicollinearity assumption.

[35]:
```
sns.pairplot(data)
```

[35]:
```
<seaborn.axisgrid.PairGrid at 0x7c530b9d4e90>
```



[36]:
```python
from statsmodels.stats.outliers_influence import variance_inflation_factor

X = data[['Radio','Social_Media']]

vif = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]

df_vif = pd.DataFrame(vif, index=X.columns, columns = ['VIF'])

df_vif
```

[36]:

	VIF
Radio	5.170922
Social_Media	5.170922

Question 8: Is the no multicollinearity assumption met?

The preceding model only has one continous independent variable, meaning there are no multicollinearity issues. If a model used both `Radio` and `Social_Media` as predictors, there would be a moderate linear relationship between `Radio` and `Social_Media` that violates the multicollinearity assumption. Furthermore, the variance inflation factor when both `Radio` and `Social_Media` are included in the model is 5.17 for each variable, indicating high multicollinearity.

1.5 Step 4: Results and evaluation

1.5.1 Display the OLS regression results

If the model assumptions are met, you can interpret the model results accurately.

First, display the OLS regression results.

[37]:
```
model_results
```

[37]:
```
<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:                Tue, 13 Aug 2024   Prob (F-statistic):          1.63e-288
Time:                        16:41:54   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.
"""
```

Question: What is your interpretation of the model's R-squared?

Using `TV` and `Radio` as the independent variables results in a multiple linear regression model with $R^2 = 0.904$. In other words, the model explains 90.4% of the variation in `Sales`. This makes the model an excellent predictor of `Sales`.

1.5.2 Interpret model coefficients

With the model fit evaluated, you can look at the coefficient estimates and the uncertainty of these estimates.

Again, display the OLS regression results.

[38]:
```python
# Display the model results summary.

### YOUR CODE HERE ###

model_results
```

[38]:
```
<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:                Tue, 13 Aug 2024   Prob (F-statistic):          1.63e-288
Time:                        16:41:54   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.
"""
```

Question: What are the model coefficients?

When `TV` and `Radio` are used to predict `Sales`, the model coefficients are:

- $\beta_0 = 218.5261$
- $\beta_{TVLow} = -154.2971$
- $\beta_{TVMedium} = -75.3120$
- $\beta_{Radio} = 2.9669$

Question: How would you write the relationship between `Sales` and the independent variables as a linear equation?

$\text{Sales} = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3$

$\text{Sales} = \beta_0 + \beta_{TVLow} * X_{TVLow} + \beta_{TVMedium} * X_{TVMedium} + \beta_{Radio} * X_{Radio}$

$\text{Sales} = 218.5261 - 154.2971 * X_{TVLow} - 75.3120 * X_{TVMedium} + 2.9669 * X_{Radio}$

Question: What is your intepretation of the coefficient estimates? Are the coefficients statistically significant?

The default `TV` category for the model is `High` since there are coefficients for the other two `TV` categories, `Medium` and `Low`. Because the coefficients for the `Medium` and `Low TV` categories are negative, that means the average of sales is lower for `Medium` or `Low TV` categories compared to the `High TV` category when `Radio` is at the same level.

For example, the model predicts that a `Low TV` promotion is 154.2971 lower on average compared to a `high TV` promotion given the same `Radio` promotion.

The coefficient for `Radio` is positive, confirming the positive linear relationship shown earlier during the exploratory data analysis.

The p-value for all coefficients is 0.000, meaning all coefficients are statistically significant at $p = 0.05$. The 95% confidence intervals for each coefficient should be reported when presenting results to stakeholders.

For example, there is a 95% chance that the interval $[-163.979, -144.616]$ contains the true parameter of the slope of β_{TVLow}, which is the estimated difference in promotion sales when a `Low TV` promotion is chosen instead of a `High TV` promotion.[Write your response here. Double-click (or enter) to edit.]

Question: Why is it important to interpret the beta coefficients?

Beta coefficients allow you to estimate the magnitude and direction (positive or negative) of the effect of each independent variable on the dependent variable. The coefficient estimates can be converted to explainable insights, such as the connection between an increase in TV promotional budgets and sales mentioned previously.

Question: What are you interested in exploring based on your model?

- Providing the business with the estimated sales given different TV promotions and radio budgets
- Additional plots to help convey the results, such as using the `seaborn regplot()` to plot the data with a best fit regression line

Question: Do you think your model could be improved? Why or why not? How?

Yes, by getting a more granular view of the `TV` promotions, such as by considering more categories or the actual `TV` promotional budgets, and getting additional variables, such as the location of the marketing campaign or the time of year, could increase model accuracy.

1.6 Conclusion

What are some key takeaways that you learned from this lab?

- Multiple linear regression is a powerful tool to estimate a dependent continous variable from several independent variables.
- Exploratory data analysis is useful for selecting both numeric and categorical features for multiple linear regression.
- Fitting multiple linear regression models may require trial and error to select variables that fit an accurate model while maintaining model assumptions.

What findings would you share with others?

According to the model, high TV promotional budgets result in significantly more sales than medium and low TV promotional budgets. For example, the model predicts that a `Low TV` promotion is 154.2971 lower on average than a `high TV` promotion given the same `Radio` promotion.

The coefficient for radio is positive, confirming the positive linear relationship shown earlier during the exploratory data analysis.

The p-value for all coefficients is 0.000, meaning all coefficients are statistically significant at $p = 0.05$. The 95% confidence intervals for each coefficient should be reported when presenting results to stakeholders.

For example, there is a 95% chance the interval $[-163.979, -144.616]$ contains the true parameter of the slope of β_{TVLow}, which is the estimated difference in promotion sales when a low TV promotional budget is chosen instead of a high TV promotion budget.

How would you frame your findings to stakeholders?

High TV promotional budgets have a substantial positive influence on sales. The model estimates that switching from a high to medium TV promotional budget reduces sales by \$75.3120 million (95% CI $[-82.431, -68.193]$), and switching from a high to low TV promotional budget reduces sales by \$154.297 million (95% CI $[-163.979, -144.616]$). The model also estimates that an increase of \$1 million in the radio promotional budget will yield a \$2.9669 million increase in sales (95% CI $[2.551, 3.383]$).

Thus, it is recommended that the business allot a high promotional budget to TV when possible and invest in radio promotions to increase sales.

References Saragih, H.S. (2020). *Dummy Marketing and Sales Data*.