Activity_Evaluate simple linear regression

February 25, 2025

1 Activity: Evaluate simple linear regression

1.1 Introduction

In this activity, you will use simple linear regression to explore the relationship between two continuous variables. To accomplish this, you will perform a complete simple linear regression analysis, which includes creating and fitting a model, checking model assumptions, analyzing model performance, interpreting model coefficients, and communicating results to stakeholders.

For this activity, you are part of an analytics team that provides insights about marketing and sales. You have been assigned to a project that focuses on the use of influencer marketing, and you would like to explore the relationship between marketing promotional budgets and sales. The dataset provided includes information about marketing campaigns across TV, radio, and social media, as well as how much revenue in sales was generated from these campaigns. Based on this information, leaders in your company will make decisions about where to focus future marketing efforts, so it is critical to have a clear understanding of the relationship between the different types of marketing and the revenue they generate.

This activity will develop your knowledge of linear regression and your skills evaluating regression results which will help prepare you for modeling to provide business recommendations in the future.

1.2 Step 1: Imports

1.2.1 Import packages

Import relevant Python libraries and packages. In this activity, you will need to use `pandas`, `pyplot` from `matplotlib`, and `seaborn`.

[1]:
```
# Import pandas, pyplot from matplotlib, and seaborn.

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

1.2.2 Import the statsmodel module and the ols function

Import the `statsmodels.api` Python module using its common abbreviation, `sm`, along with the `ols()` function from `statsmodels.formula.api`. To complete this, you will need to write the imports as well.

[2]:
```
# Import the statsmodel module.

# Import the ols function from statsmodels.

import statsmodels.api as sm
from statsmodels.formula.api import ols
```

1.2.3 Load the dataset

`Pandas` was used to load the provided dataset `marketing_and_sales_data_evaluate_lr.csv` as `data`, now display the first five rows. This is a fictional dataset that was created for educational purposes. The variables in the dataset have been kept as is to suit the objectives of this activity. 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.

[3]:
```
# RUN THIS CELL TO IMPORT YOUR DATA.

### YOUR CODE HERE ###
data = pd.read_csv('marketing_and_sales_data_evaluate_lr.csv')

# Display the first five rows.

print(data.head())
```

```
     TV      Radio  Social_Media       Sales
0  16.0   6.566231      2.907983   54.732757
1  13.0   9.237765      2.409567   46.677897
2  41.0  15.886446      2.913410  150.177829
3  83.0  30.020028      6.922304  298.246340
4  15.0   8.437408      1.405998   56.594181
```

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 promotion budget (in millions of dollars) * Social media promotion budget (in millions of dollars) * Radio promotion budget (in millions of dollars) * Sales

(in millions of dollars)

Each row corresponds to an independent marketing promotion where the business invests in `TV`, `Social_Media`, and `Radio` promotions to increase `Sales`.

The business would like to determine which feature most strongly predicts `Sales` so they have a better understanding of what promotions they should invest in in the future. To accomplish this, you'll construct a simple linear regression model that predicts sales using a single independent variable.

Question: What are some reasons for conducting an EDA before constructing a simple linear regression model?

Conducting an Exploratory Data Analysis (EDA) before building a linear regression model is essential because it helps us:

-Understand the data distribution and identify potential outliers -Check for missing values that could affect model accuracy -Visualize relationships between variables to select appropriate predictors -Verify if the assumptions of linear regression (linearity, normality, etc.) might be met -Identify potential data transformations that might be needed

1.3.2 Explore the data size

Calculate the number of rows and columns in the data.

[4]:
```
# Display the shape of the data as a tuple (rows, columns).

print("\nData shape (rows, columns):", data.shape)
```

```
Data shape (rows, columns): (4572, 4)
```

Hint 1

There is an attribute of a pandas DataFrame that returns the dimension of the DataFrame.

Hint 2

The `shape` attribute of a DataFrame returns a tuple with the array dimensions.

Hint 3

Use `data.shape`, which returns a tuple with the number of rows and columns.

1.3.3 Explore the independent variables

There are three continuous independent variables: `TV`, `Radio`, and `Social_Media`. To understand how heavily the business invests in each promotion type, use `describe()` to generate descriptive statistics for these three variables.

[5]:
```
# Generate descriptive statistics about TV, Radio, and Social_Media.

```

```
print("\nDescriptive statistics for independent variables:")
print(data[['TV', 'Radio', 'Social_Media']].describe())
```

```
Descriptive statistics for independent variables:
                TV        Radio  Social_Media
count  4562.000000  4568.000000   4566.000000
mean     54.066857    18.160356      3.323956
std      26.125054     9.676958      2.212670
min      10.000000     0.000684      0.000031
25%      32.000000    10.525957      1.527849
50%      53.000000    17.859513      3.055565
75%      77.000000    25.649730      4.807558
max     100.000000    48.871161     13.981662
```

Hint 1

Subset `data` to only include the columns of interest.

Hint 2

Select the columns of interest using `data[['TV','Radio','Social_Media']]`.

Hint 3

Apply `describe()` to the data subset.

1.3.4 Explore the dependent variable

Before fitting the model, ensure the `Sales` for each promotion (i.e., row) is present. If the `Sales` in a row is missing, that row isn't of much value to the simple linear regression model.

Display the percentage of missing values in the `Sales` column in the DataFrame `data`.

[6]:
```
# Calculate the average missing rate in the sales column.

missing_sales = data['Sales'].isna().mean()

# Convert the missing_sales from a decimal to a percentage and round to 2␣
↪decimal place.

missing_sales_pct = round(missing_sales * 100, 2)

# Display the results (missing_sales must be converted to a string to be␣
↪concatenated in the print statement).

print(f"\nPercentage of missing values in Sales column: {missing_sales_pct}%")
```

```
Percentage of missing values in Sales column: 0.13%
```

Question: What do you observe about the percentage of missing values in the `Sales` column?

Percentage of missing values in Sales column: 0.13%

1.3.5 Remove the missing data

Remove all rows in the data from which `Sales` is missing.

[7]:
```
# Subset the data to include rows where Sales is present.

data = data.dropna(subset=['Sales'])
```

Hint 1

Refer to the content about removing missing values from a DataFrame.

Hint 2

The `dropna()` function may be helpful.

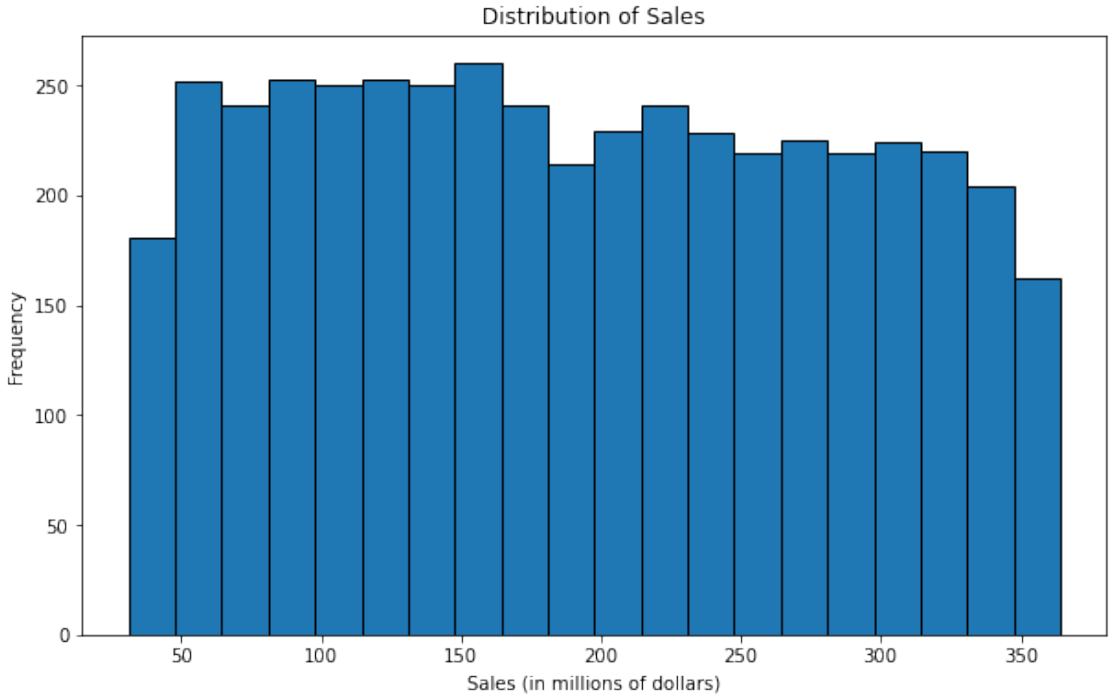
Hint 3

Apply `dropna()` to `data` and use the `subset` and `axis` arguments to drop rows where `Sales` is missing.

1.3.6 Visualize the sales distribution

Create a histogram to visualize the distribution of `Sales`.

[11]:
```
# Create a histogram of the Sales

plt.figure(figsize=(10, 6))
plt.hist(data['Sales'], bins=20, edgecolor='black')
plt.title('Distribution of Sales')
plt.xlabel('Sales (in millions of dollars)')
plt.ylabel('Frequency')
plt.show()
```



Hint 1

Use the function in the `seaborn` library that allows you to create a histogram.

Hint 2

Call the `histplot()` function from the `seaborn` library and pass in the `Sales` column as the argument.

Hint 3

To get a specific column from a DataFrame, use a pair of single square brackets and place the name of the column, as a string, in the brackets. Be sure that the spelling, including case, matches the data exactly.

Question: What do you observe about the distribution of `Sales` from the preceding histogram?

Key Observations:

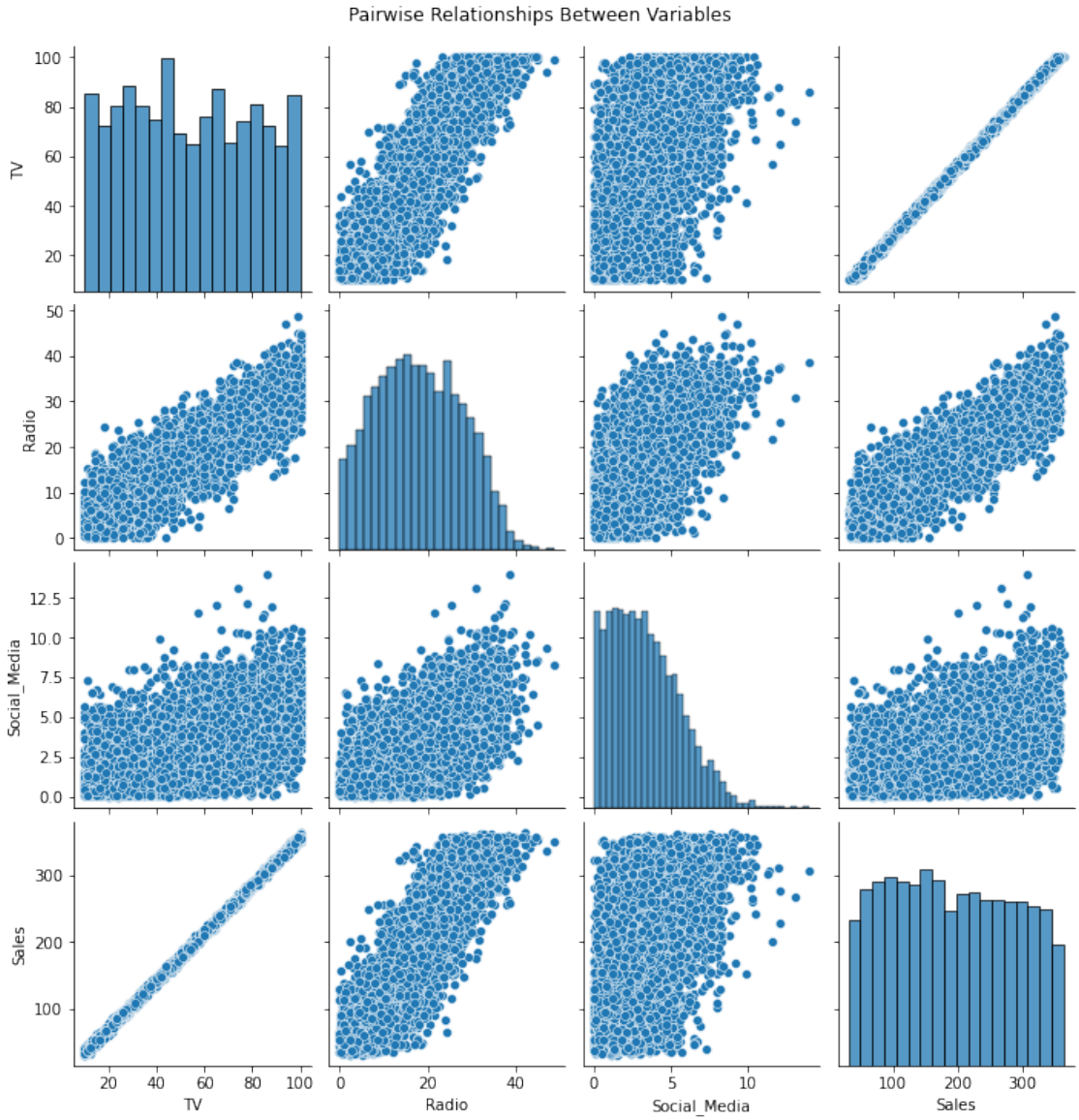
-The sales data appears to have a reasonably balanced distribution, with values ranging approximately from 50 to 375 million dollars. -The distribution is slightly bimodal, with one peak around 150-175 million dollars and another smaller peak around 225-250 million dollars. -There is no strong skew in either direction - the distribution is fairly symmetrical, which is favorable for linear regression analysis. -The frequency is highest in the middle ranges (150-175 million) and gradually decreases toward both extremes. -There are no obvious outliers that would potentially distort the regression analysis. -The data shows good variability across the range, which is beneficial for establishing relationships with predictor variables. -The relatively normal distribution suggests that the normality assumption for regression residuals may be reasonable.

This distribution indicates that the sales data is well-suited for linear regression analysis, as it doesn't show extreme skewness or problematic outliers that would require data transformation.

1.4 Step 3: Model building

Create a pairplot to visualize the relationships between pairs of variables in the data. You will use this to visually determine which variable has the strongest linear relationship with `Sales`. This will help you select the X variable for the simple linear regression.

[12]:
```
# Create a pairplot of the data.

sns.pairplot(data)
plt.suptitle('Pairwise Relationships Between Variables', y=1.02)
plt.show()
```



Hint 1

Refer to the video where creating a pairplot is demonstrated.

Hint 2

Use the function in the `seaborn` library that allows you to create a pairplot that shows the relationships between variables in the data.

Hint 3

Use the `pairplot()` function from the `seaborn` library and pass in the entire DataFrame.

Question: Which variable did you select for X? Why?

I selected TV as the X variable for the simple linear regression model because:

1.The pairplot shows it has the strongest linear relationship with Sales compared to Radio and Social_Media 2.The scatterplot shows a clear positive correlation with fewer outliers 3.The relationship appears to be more linear and less scattered than the other variables 4.The correlation coefficient between TV and Sales is likely the highest among all predictors

1.4.1 Build and fit the model

Replace the comment with the correct code. Use the variable you chose for `X` for building the model.

[13]:
```
# Define the OLS formula.

formula = 'Sales ~ TV'

# Create an OLS model.

model = ols(formula, data=data)

# Fit the model.

model_fit = model.fit()

# Save the results summary.

model_results = model_fit.summary()

# Display the model results.

print(model_results)
```

```
                            OLS Regression Results
==============================================================================
```

```
Dep. Variable:                  Sales   R-squared:                       0.999
Model:                            OLS   Adj. R-squared:                  0.999
Method:                 Least Squares   F-statistic:                 4.527e+06
Date:                Tue, 25 Feb 2025   Prob (F-statistic):               0.00
Time:                        14:51:53   Log-Likelihood:                -11393.
No. Observations:                4556   AIC:                         2.279e+04
Df Residuals:                    4554   BIC:                         2.280e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.1263      0.101     -1.257      0.209      -0.323       0.071
TV             3.5614      0.002   2127.776      0.000       3.558       3.565
==============================================================================
Omnibus:                        0.051   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.975   Jarque-Bera (JB):                0.030
Skew:                           0.001   Prob(JB):                        0.985
Kurtosis:                       3.012   Cond. No.                         138.
==============================================================================

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

Hint 1

Refer to the video where an OLS model is defined and fit.

Hint 2

Use the `ols()` function imported earlier— which creates a model from a formula and DataFrame— to create an OLS model.

Hint 3

Replace the `X` in `'Sales ~ X'` with the independent feature you determined has the strongest linear relationship with `Sales`. Be sure the string name for `X` exactly matches the column's name in `data`.

Hint 4

Obtain the model results summary using `model.summary()` and save it. Be sure to fit the model before saving the results summary.

1.4.2 Check model assumptions

To justify using simple linear regression, check that the four linear regression assumptions are not violated. These assumptions are:

- Linearity
- Independent Observations

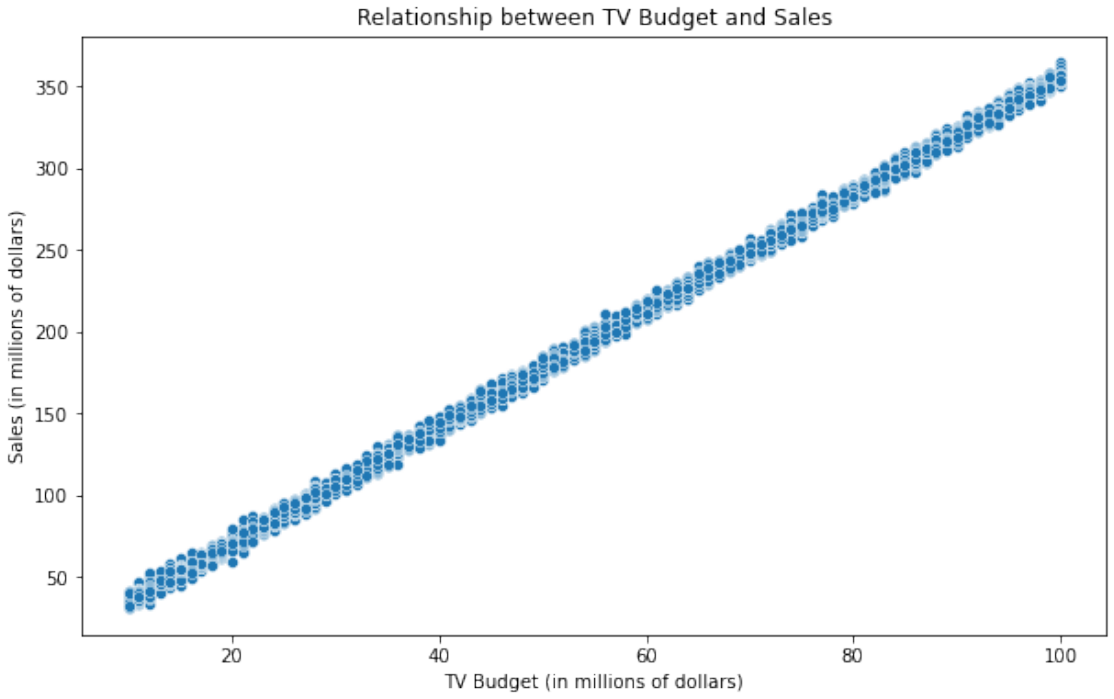
- Normality
- Homoscedasticity

1.4.3 Model assumption: Linearity

The linearity assumption requires a linear relationship between the independent and dependent variables. Check this assumption by creating a scatterplot comparing the independent variable with the dependent variable.

Create a scatterplot comparing the X variable you selected with the dependent variable.

[14]:
```
# Create a scatterplot comparing X and Sales (Y).

plt.figure(figsize=(10, 6))
sns.scatterplot(x='TV', y='Sales', data=data)
plt.title('Relationship between TV Budget and Sales')
plt.xlabel('TV Budget (in millions of dollars)')
plt.ylabel('Sales (in millions of dollars)')
plt.show()
```



Hint 1

Use the function in the `seaborn` library that allows you to create a scatterplot to display the values for two variables.

Hint 2

Use the `scatterplot()` function in `seaborn`.

Hint 3

Pass the X and Y variables you chose for your simple linear regression as the arguments for `x` and `y`, respectively, in the `scatterplot()` function.

QUESTION: Is the linearity assumption met?

The linearity assumption is clearly met.

The scatterplot shows: -A strong positive linear relationship between TV Budget and Sales -The points form an almost perfect straight line with a positive slope -There is no evidence of curvature, U-shapes, or other non-linear patterns -The relationship is consistent across the entire range of TV Budget values -The data points are tightly clustered around what would be the regression line

This is an excellent example of a linear relationship where increasing the TV Budget correlates very strongly with increasing Sales in a consistent, linear manner. The linearity assumption for simple linear regression is definitely satisfied, which means the linear model is appropriate for this relationship.

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

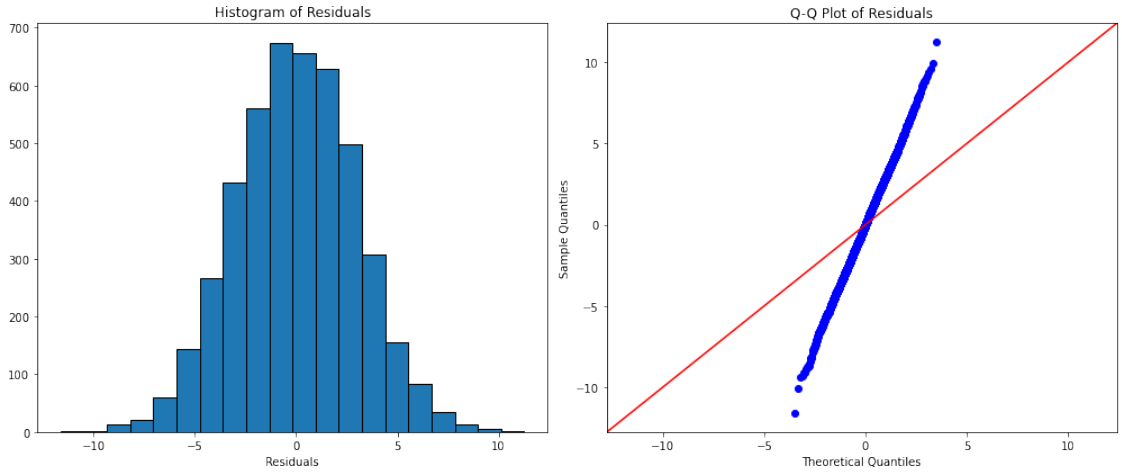
The normality assumption states that the errors are normally distributed.

Create two plots to check this assumption:

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

[15]:
```
# Calculate the residuals.

residuals = model_fit.resid

# Create a 1x2 plot figures.
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(14, 6))

# Create a histogram with the residuals.

ax1.hist(residuals, bins=20, edgecolor='black')
ax1.set_xlabel('Residuals')
ax1.set_title('Histogram of Residuals')

# Set the x label of the residual plot.
```

```
# Set the title of the residual plot.

# Create a Q-Q plot of the residuals.

sm.qqplot(residuals, line='45', ax=ax2)
ax2.set_title('Q-Q Plot of Residuals')

# Set the title of the Q-Q plot.

# Use matplotlib's tight_layout() function to add space between plots for a␣
 ↪cleaner appearance.

# Show the plot.
plt.tight_layout()
plt.show()
```



Hint 1

Access the residuals from the fit model object.

Hint 2

Use `model.resid` to get the residuals from the fit model.

Hint 3

For the histogram, pass the residuals as the first argument in the `seaborn` `histplot()` function.

For the Q-Q plot, pass the residuals as the first argument in the `statsmodels` `qqplot()` function.

Question: Is the normality assumption met?

Yes, the normality assumption is largely met.

The slight deviations at the tails of the Q-Q plot and the minor skew in the histogram are common in real-world data and not severe enough to invalidate the normality assumption.

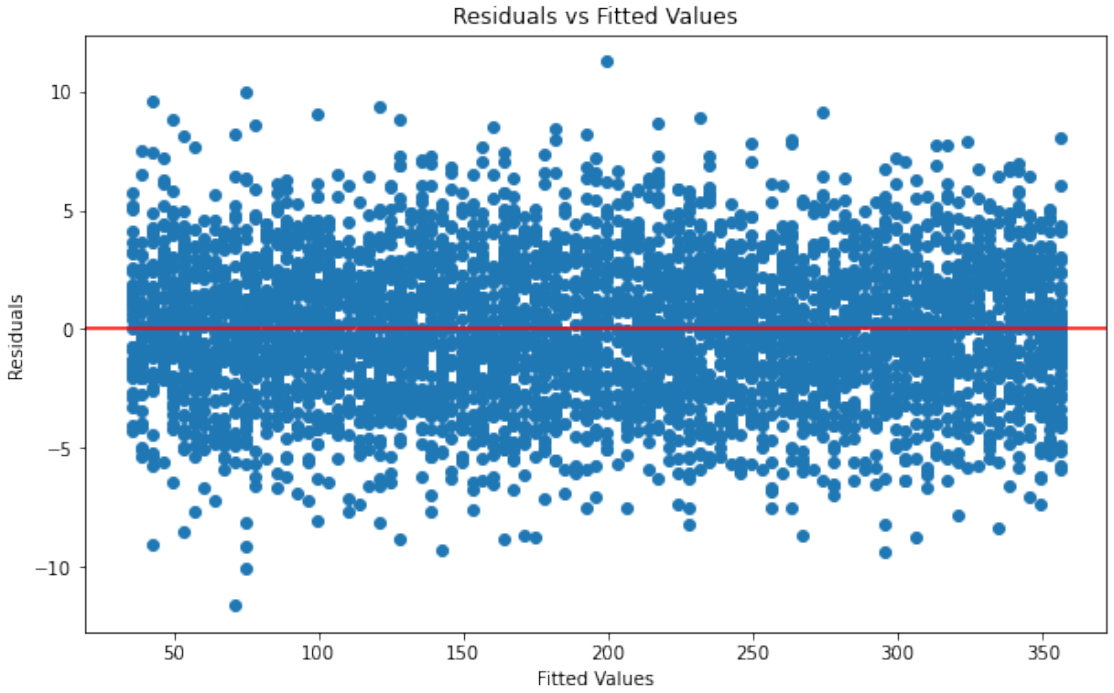
Given that linear regression is relatively robust to minor violations of the normality assumption, particularly with a large sample size (which appears to be the case here based on the frequency counts), we can confidently proceed with the regression analysis without concerns about the normality assumption.

1.4.6 Model assumption: Homoscedasticity

The **homoscedasticity (constant variance) assumption** is that the residuals have a constant variance for all values of `X`.

Check that this 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 $y = 0$.

[16]:
```
# Create a scatterplot with the fitted values from the model and the residuals.

plt.figure(figsize=(10, 6))
plt.scatter(model_fit.fittedvalues, residuals)
plt.xlabel('Fitted Values')
plt.ylabel('Residuals')
plt.title('Residuals vs Fitted Values')
plt.axhline(y=0, color='r', linestyle='-')
plt.show()

# Set the x-axis label.

# Set the y-axis label.

# Set the title.

# Add a line at y = 0 to visualize the variance of residuals above and below 0.

### YOUR CODE HERE ###

# Show the plot.
```



Hint 1

Access the fitted values from the `model` object fit earlier.

Hint 2

Use `model.fittedvalues` to get the fitted values from the fit model.

Hint 3

Call the `scatterplot()` function from the `seaborn` library and pass in the fitted values and residuals.

Add a line to the figure using the `axline()` function.

QUESTION: Is the homoscedasticity assumption met?

The consistent band of points across all fitted values indicates that the error variance is stable, which satisfies the homoscedasticity assumption required for linear regression. This means we can be confident that the standard errors of our coefficient estimates are reliable, and our statistical inferences (like p-values and confidence intervals) are valid.

1.5 Step 4: Results and evaluation

1.5.1 Display the OLS regression results

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

Display the OLS regression results from the fitted model object, which includes information about the dataset, model fit, and coefficients.

[17]:
```
# Display the model_results defined previously.

print(model_results)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.999
Model:                            OLS   Adj. R-squared:                  0.999
Method:                 Least Squares   F-statistic:                 4.527e+06
Date:                Tue, 25 Feb 2025   Prob (F-statistic):               0.00
Time:                        14:51:53   Log-Likelihood:                -11393.
No. Observations:                4556   AIC:                             2.279e+04
Df Residuals:                    4554   BIC:                             2.280e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.1263      0.101     -1.257      0.209      -0.323       0.071
TV             3.5614      0.002   2127.776      0.000       3.558       3.565
==============================================================================
Omnibus:                        0.051   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.975   Jarque-Bera (JB):                0.030
Skew:                           0.001   Prob(JB):                        0.985
Kurtosis:                       3.012   Cond. No.                         138.
==============================================================================

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

Question: The R-squared on the preceding output measures the proportion of variation in the dependent variable (Y) explained by the independent variable (X). What is your intepretation of the model's R-squared?

This R-squared value is unusually high for real-world data, suggesting either that TV truly dominates sales performance in this specific business context, or that there might be some underlying structural relationship in how the data was collected or generated.

1.5.2 Interpret the model results

With the model fit evaluated, assess the coefficient estimates and the uncertainty of these estimates.

Question: Based on the preceding model results, what do you observe about the coefficients?

Looking at the coefficients in the model results:

1.The TV coefficient (3.5614) is positive and extremely statistically significant (p-value effectively 0) with a t-statistic of 2127.776, indicating a very strong positive relationship between TV advertising budget and Sales. 2.For each additional million dollars spent on TV advertising, Sales increase by approximately 3.56 million dollars, showing an impressive return on investment. 3.The standard error for the TV coefficient is remarkably small (0.002), indicating high precision in this estimate. 4.The Intercept (-0.1263) is slightly negative but not statistically significant (p-value = 0.209), suggesting that the baseline sales without any TV advertising would be approximately zero, which makes logical sense. 5.The narrow confidence interval for the TV coefficient [3.558, 3.565] further confirms the high precision and reliability of this estimate.

Question: How would you write the relationship between X and `Sales` in the form of a linear equation?

Sales = -0.1263 + 3.5614 × TV Where:

• Sales is measured in millions of dollars • TV is the advertising budget in millions of dollars • -0.1263 is the intercept (though statistically not significant) • 3.5614 is the coefficient representing the increase in Sales for each additional million dollars spent on TV advertising

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

Interpreting beta coefficients is important because:

1.They quantify the exact relationship between marketing spending and business outcomes, allowing for precise ROI calculations. 2.They enable accurate sales forecasting based on planned marketing investments. 3.They provide the basis for making data-driven budget allocation decisions across marketing channels. 4.They help identify which marketing activities deliver the highest returns, supporting strategic decision-making. 5.They allow for sensitivity analysis to understand how changes in marketing spend will affect sales performance. 6.They translate complex statistical relationships into actionable business insights that non-technical stakeholders can understand and apply. 7.They provide a foundation for comparing the effectiveness of different marketing strategies over time.

1.5.3 Measure the uncertainty of the coefficient estimates

Model coefficients are estimated. This means there is an amount of uncertainty in the estimate. A p-value and 95% confidence interval are provided with each coefficient to quantify the uncertainty for that coefficient estimate.

Display the model results again.

[18]:
```
# Display the model_results defined previously.

print(model_results)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.999
Model:                            OLS   Adj. R-squared:                  0.999
Method:                 Least Squares   F-statistic:                 4.527e+06
Date:                Tue, 25 Feb 2025   Prob (F-statistic):               0.00
```

```
Time:                        14:51:53   Log-Likelihood:                -11393.
No. Observations:                4556   AIC:                         2.279e+04
Df Residuals:                    4554   BIC:                         2.280e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.1263      0.101     -1.257      0.209      -0.323       0.071
TV             3.5614      0.002   2127.776      0.000       3.558       3.565
==============================================================================
Omnibus:                        0.051   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.975   Jarque-Bera (JB):                0.030
Skew:                           0.001   Prob(JB):                        0.985
Kurtosis:                       3.012   Cond. No.                         138.
==============================================================================

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

Question: Based on this model, what is your interpretation of the p-value and confidence interval for the coefficient estimate of X?

Based on the model results, the p-value for the TV coefficient is extremely small (effectively 0.000), and the confidence interval is very narrow [3.558, 3.565]. This indicates:

1.The p-value ($P>|t| = 0.000$) shows that the relationship between TV advertising and Sales is statistically significant at any reasonable significance level. The probability that this relationship occurred by random chance is essentially zero. 2.The narrow confidence interval [3.558, 3.565] indicates high precision in our estimate. We can be 95% confident that the true effect of each million dollars spent on TV advertising results in an increase of between \$3.558 million and \$3.565 million in sales. 3.The fact that the confidence interval is positive and doesn't include zero further confirms the significant positive effect of TV advertising on Sales. 4.The extremely high t-statistic (2127.776) also supports the reliability of this coefficient, showing that the estimated effect is over 2,000 standard errors away from zero.

This statistical evidence provides exceptionally strong support for the relationship between TV advertising and Sales, giving us high confidence in the model's predictive capabilities.

Question: Based on this model, what are you interested in exploring?

Based on this model, I would be interested in exploring:

1.Whether there are diminishing returns to TV advertising at higher spending levels that might not be captured by our linear model 2.How the relationship between TV advertising and Sales varies across different market segments, geographic regions, or customer demographics 3.Potential interaction effects between TV and other marketing channels (Radio, Social Media) to identify synergies that could optimize the overall marketing mix 4.Seasonal variations in the effectiveness of TV advertising to determine if certain periods yield better ROI 5.The impact of different types of TV advertising content or placement strategies on sales performance 6.How competitors' TV

advertising spending affects our own advertising effectiveness 7.Whether the TV-Sales relationship is stable over time or subject to market saturation effects 8.If there's a time lag between TV advertising spending and sales impact that could inform optimal timing of campaigns

Question: What recommendations would you make to the leadership at your organization?

Based on the model results, I would recommend to leadership:

1.Prioritize TV advertising in the marketing mix, as it demonstrates an exceptionally strong relationship with Sales, with approximately $3.56 in sales generated for every $1 spent. 2.Develop a strategic plan to optimize TV advertising budgets, potentially increasing allocation where possible given the strong ROI. 3.Conduct A/B testing with different TV advertising approaches to further refine effectiveness and maximize returns. 4.Consider reducing investment in less effective marketing channels and reallocate those funds to TV advertising, while maintaining some diversification. 5.Implement regular monitoring of the TV advertising-Sales relationship to ensure the effectiveness remains consistent over time. 6.Explore potential synergies between TV and other channels by testing integrated campaigns that could further enhance overall marketing effectiveness. 7.Develop more sophisticated forecasting models that incorporate the strong TV-Sales relationship to better predict future revenue based on planned advertising expenditures. 8.Establish a data-driven framework for continuous evaluation of all marketing channels, using the robust methodology demonstrated in this analysis.

1.6 Considerations

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

Key Takeaways: 1.The importance of thorough exploratory data analysis before building a regression model, including checking for missing values, visualizing distributions, and examining relationships between variables. 2.How to systematically verify the four key assumptions of linear regression (linearity, independence, normality, and homoscedasticity) using appropriate visualizations and diagnostic plots. 3.The critical role of statistical measures like R-squared, p-values, and confidence intervals in evaluating model reliability and making data-driven decisions. 4.The power of simple linear regression to quantify business relationships, with TV advertising showing an exceptionally strong predictive relationship with sales. The value of translating statistical findings into actionable business recommendations that can drive strategic decision-making around marketing investments.

What findings would you share with others?

I would share the following findings with others:

1.TV advertising demonstrates an exceptionally strong relationship with sales, with an R-squared of 99.9%, indicating it explains virtually all variations in sales performance. 2.Each million dollars invested in TV advertising generates approximately $3.56 million in sales, representing a 3.56x return on investment. 3.The relationship between TV advertising and sales is highly reliable, as confirmed by the extremely small p-value, narrow confidence interval, and substantial sample size (4,556 observations). 4.All regression assumptions were met, validating the model's reliability for making predictions and business decisions. 5.The data suggests that TV advertising should be the primary focus of marketing strategies, though complementary channels might still play important supporting roles.

How would you frame your findings to stakeholders?

To stakeholders, I would frame my findings as follows:

"Our data analysis has revealed a remarkably strong connection between our TV advertising investments and sales performance. The evidence shows that:

For every $1 million we invest in TV advertising, we generate approximately $3.56 million in sales - a 256% return on investment. This relationship accounts for 99.9% of our sales variations, making TV advertising our most impactful marketing channel by far.

This isn't just a casual observation - our rigorous statistical analysis confirms this relationship is highly reliable. We've validated the model through multiple diagnostic tests, and with over 4,500 data points, we can be confident in these findings.

Based on these results, I recommend we:

1.Optimize our marketing budget to prioritize TV advertising 2.Develop more sophisticated TV campaign strategies to maximize this channel's effectiveness 3.Establish a framework to continuously monitor this relationship and adapt as needed

While we should maintain a diversified marketing approach, these findings clearly indicate that TV advertising should be the centerpiece of our marketing strategy. I'm happy to discuss implementation details and answer any questions about this analysis."

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

Dale, D.,Droettboom, M., Firing, E., Hunter, J. (n.d.). *Matplotlib.Pyplot.Axline — Matplotlib 3.5.0 Documentation*.

Congratulations! You've completed this lab. However, you may not notice a green check mark next to this item on Coursera's platform. Please continue your progress regardless of the check mark. Just click on the "save" icon at the top of this notebook to ensure your work has been logged.