VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

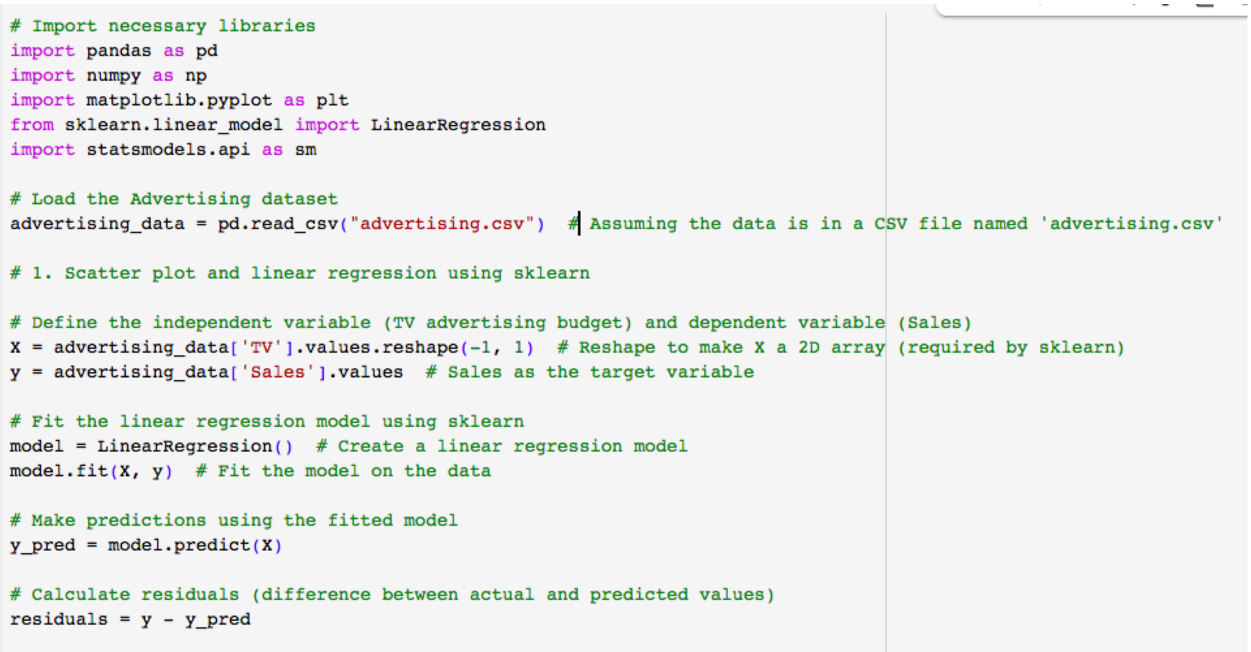
Winter Semester 2024 –2025
PMDTS 504L – Regression Analysis and Predictive Modelling
Experiment IV

Linear Regression Analysis on Advertising Dataset

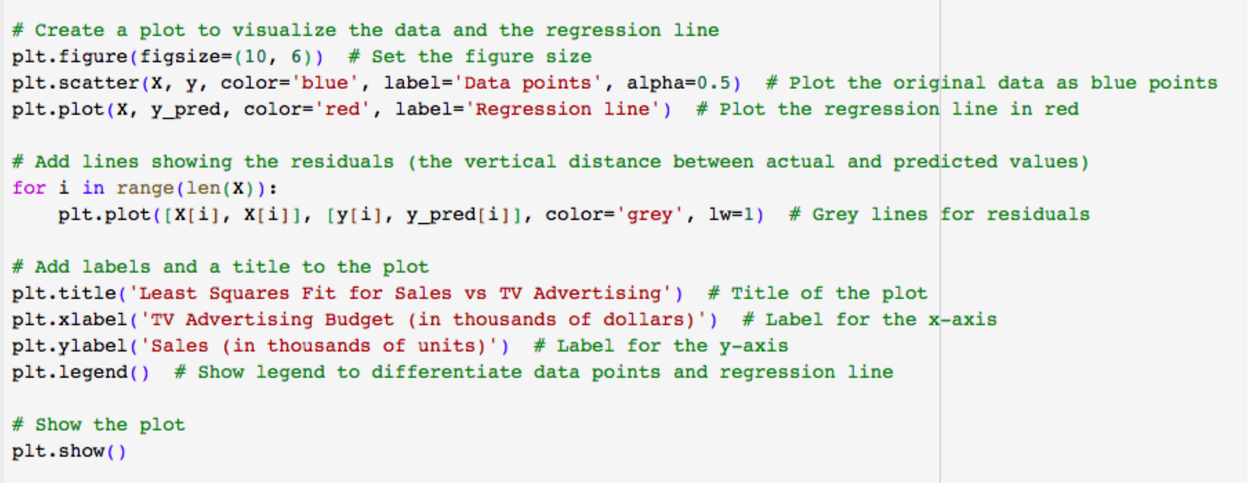
You are provided with a dataset advertising.csv, which contains information about the advertising budget spent on TV (in thousands of dollars) and the corresponding sales (in thousands of units) for a company. Your task is to analyze the relationship between TV advertising and sales using linear regression.

Tasks:

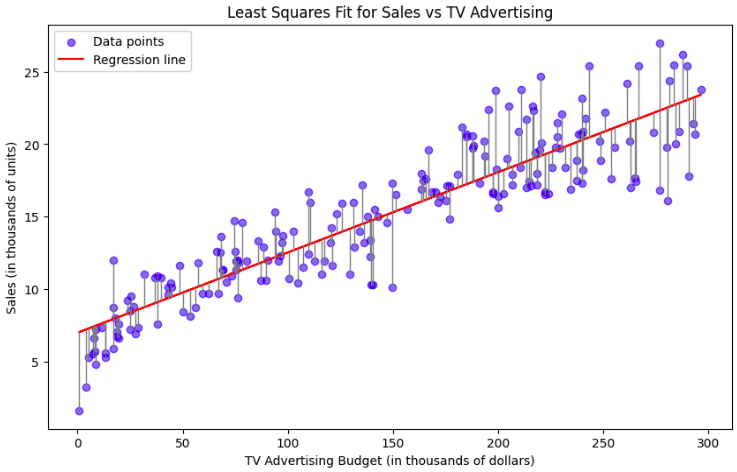
1. Data Exploration:
   - Load the dataset advertising.csv into a DataFrame.
   - Display the first few rows of the dataset to understand its structure.
2. Perform Linear Regression using sklearn:
   - Define TV as the independent variable (X) and Sales as the dependent variable (y).
   - Fit a linear regression model using sklearn and plot the following:
     - A scatter plot of TV vs. Sales.
     - The regression line on top of the scatter plot.
     - Grey lines showing the residuals (the difference between actual and predicted sales).
3. Perform Linear Regression using statsmodels:
   - Add an intercept term to the independent variable TV (using sm.add_constant).
   - Fit an Ordinary Least Squares (OLS) linear regression model using statsmodels.
   - Print the summary of the fitted model. This summary should include:
     - The coefficients (intercept and slope).
     - R-squared value.
     - P-values and standard errors.
   - Calculate and display the 95% confidence intervals for the coefficients (intercept and slope).
   - Extract and print the standard errors for the intercept and the coefficient of TV advertising.

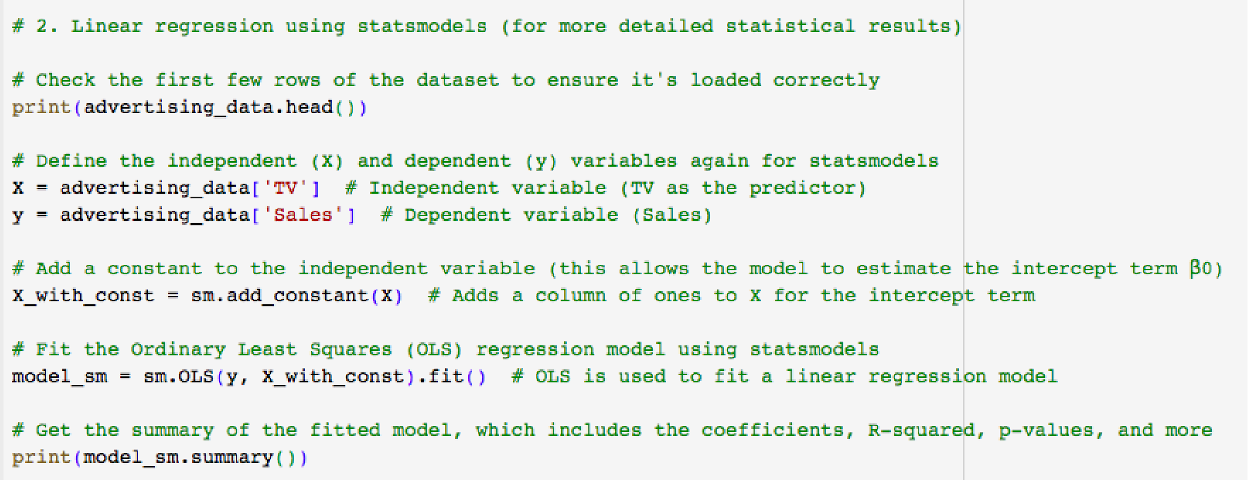


```python
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm

# Load the Advertising dataset
advertising_data = pd.read_csv("advertising.csv")  # Assuming the data is in a CSV file named 'advertising.csv'

# 1. Scatter plot and linear regression using sklearn

# Define the independent variable (TV advertising budget) and dependent variable (Sales)
X = advertising_data['TV'].values.reshape(-1, 1)  # Reshape to make X a 2D array (required by sklearn)
y = advertising_data['Sales'].values  # Sales as the target variable

# Fit the linear regression model using sklearn
model = LinearRegression()  # Create a linear regression model
model.fit(X, y)  # Fit the model on the data

# Make predictions using the fitted model
y_pred = model.predict(X)

# Calculate residuals (difference between actual and predicted values)
residuals = y - y_pred
```

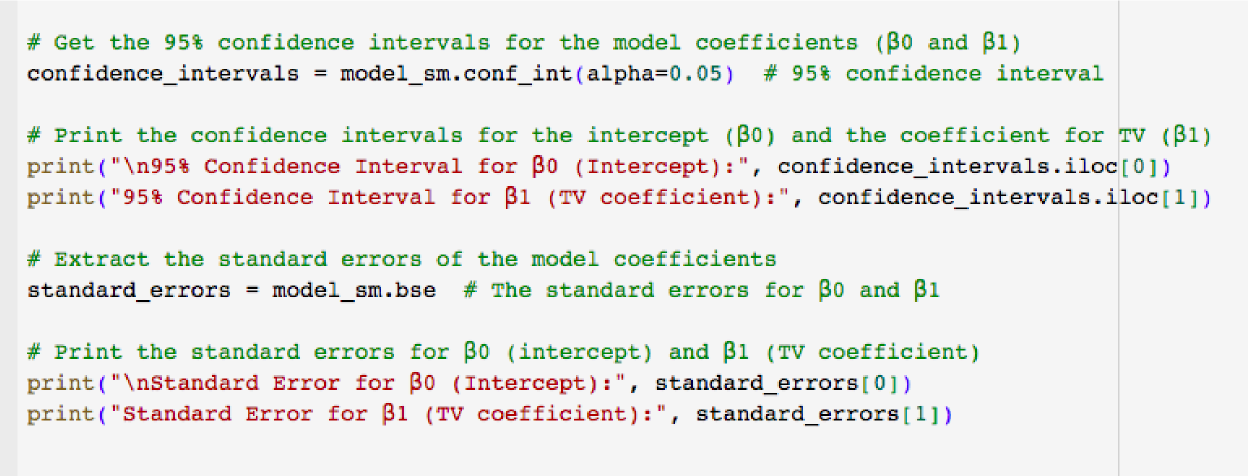


```python
# Create a plot to visualize the data and the regression line
plt.figure(figsize=(10, 6))  # Set the figure size
plt.scatter(X, y, color='blue', label='Data points', alpha=0.5)  # Plot the original data as blue points
plt.plot(X, y_pred, color='red', label='Regression line')  # Plot the regression line in red

# Add lines showing the residuals (the vertical distance between actual and predicted values)
for i in range(len(X)):
    plt.plot([X[i], X[i]], [y[i], y_pred[i]], color='grey', lw=1)  # Grey lines for residuals

# Add labels and a title to the plot
plt.title('Least Squares Fit for Sales vs TV Advertising')  # Title of the plot
plt.xlabel('TV Advertising Budget (in thousands of dollars)')  # Label for the x-axis
plt.ylabel('Sales (in thousands of units)')  # Label for the y-axis
plt.legend()  # Show legend to differentiate data points and regression line

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

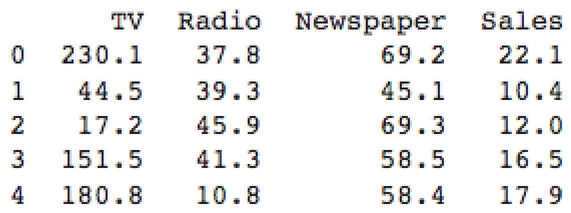




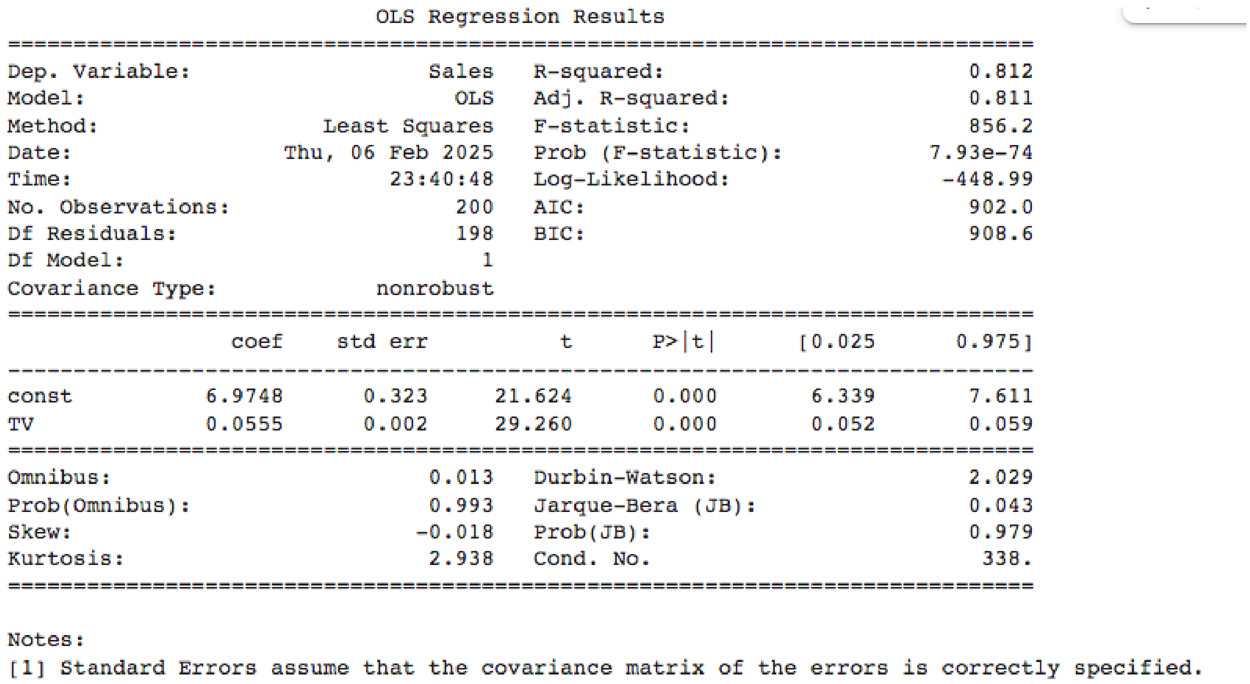
```python
# 2. Linear regression using statsmodels (for more detailed statistical results)

# Check the first few rows of the dataset to ensure it's loaded correctly
print(advertising_data.head())

# Define the independent (X) and dependent (y) variables again for statsmodels
X = advertising_data['TV']  # Independent variable (TV as the predictor)
y = advertising_data['Sales']  # Dependent variable (Sales)

# Add a constant to the independent variable (this allows the model to estimate the intercept term β0)
X_with_const = sm.add_constant(X)  # Adds a column of ones to X for the intercept term

# Fit the Ordinary Least Squares (OLS) regression model using statsmodels
model_sm = sm.OLS(y, X_with_const).fit()  # OLS is used to fit a linear regression model

# Get the summary of the fitted model, which includes the coefficients, R-squared, p-values, and more
print(model_sm.summary())
```



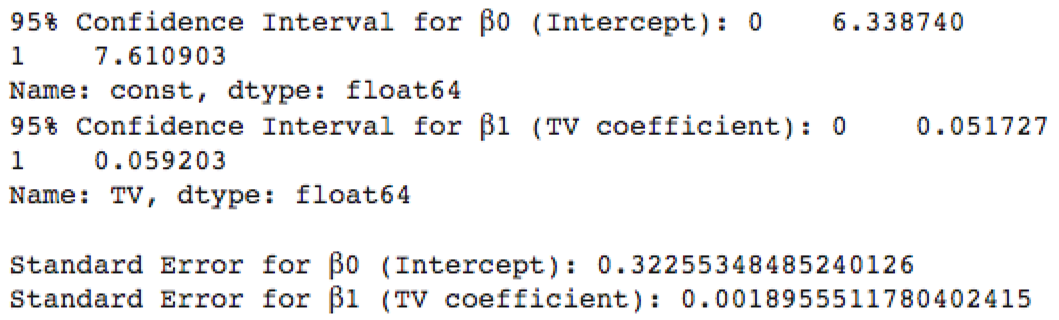
```python
# Get the 95% confidence intervals for the model coefficients (β0 and β1)
confidence_intervals = model_sm.conf_int(alpha=0.05)  # 95% confidence interval

# Print the confidence intervals for the intercept (β0) and the coefficient for TV (β1)
print("\n95% Confidence Interval for β0 (Intercept):", confidence_intervals.iloc[0])
print("95% Confidence Interval for β1 (TV coefficient):", confidence_intervals.iloc[1])

# Extract the standard errors of the model coefficients
standard_errors = model_sm.bse  # The standard errors for β0 and β1

# Print the standard errors for β0 (intercept) and β1 (TV coefficient)
print("\nStandard Error for β0 (Intercept):", standard_errors[0])
print("Standard Error for β1 (TV coefficient):", standard_errors[1])
```



```
      TV  Radio  Newspaper  Sales
0  230.1   37.8       69.2   22.1
1   44.5   39.3       45.1   10.4
2   17.2   45.9       69.3   12.0
3  151.5   41.3       58.5   16.5
4  180.8   10.8       58.4   17.9
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.812
Model:                            OLS   Adj. R-squared:                  0.811
Method:                 Least Squares   F-statistic:                     856.2
Date:                Thu, 06 Feb 2025   Prob (F-statistic):           7.93e-74
Time:                        23:40:48   Log-Likelihood:                -448.99
No. Observations:                 200   AIC:                             902.0
Df Residuals:                     198   BIC:                             908.6
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          6.9748      0.323     21.624      0.000       6.339       7.611
TV             0.0555      0.002     29.260      0.000       0.052       0.059
==============================================================================
Omnibus:                        0.013   Durbin-Watson:                   2.029
Prob(Omnibus):                  0.993   Jarque-Bera (JB):                0.043
Skew:                          -0.018   Prob(JB):                        0.979
Kurtosis:                       2.938   Cond. No.                         338.
==============================================================================

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



```
95% Confidence Interval for β0 (Intercept): 0    6.338740
1    7.610903
Name: const, dtype: float64
95% Confidence Interval for β1 (TV coefficient): 0    0.051727
1    0.059203
Name: TV, dtype: float64

Standard Error for β0 (Intercept): 0.32255348485240126
Standard Error for β1 (TV coefficient): 0.0018955511780402415
```

Regression Summary:

- **Dependent Variable:** Sales (in thousands of units)
- **Independent Variable:** TV Advertising Budget (in thousands of dollars)

Key Results:

1. **R-squared = 0.812:**
    - The R-squared value represents the proportion of the variance in the dependent variable (Sales) that is explained by the independent variable (TV advertising). An R-squared of 0.812 indicates that about 81.2% of the variance in sales can be explained by the TV advertising budget. This is a strong fit, suggesting that TV advertising has a significant impact on sales.
2. **Adj. R-squared = 0.811:**

   - The adjusted R-squared takes into account the number of predictors and adjusts for the potential overfitting. It's very close to the R-squared value, indicating that the model is well-fitting and that including just one predictor (TV) is appropriate for the data.

3. **F-statistic = 856.2 (p-value = 7.93e-74):**
   - The F-statistic tests the overall significance of the regression model. A very large F-statistic with an extremely low p-value (close to zero) indicates that the model is statistically significant, and the independent variable (TV advertising budget) is a significant predictor of sales.
4. **Coefficients:**
   - **Intercept (β_0) = 6.9748:** This is the estimated sales when the TV advertising budget is zero. It suggests that if the company spends nothing on TV advertising, the expected sales will be approximately 6.97 thousand units.
   - **TV (β_1) = 0.0555:** This is the slope of the regression line, which shows the effect of an additional thousand dollars spent on TV advertising. For every additional $1,000 spent on TV advertising, the sales are expected to increase by 0.0555 thousand units, or 55.5 units.
5. **Standard Errors:**
   - The standard errors of the coefficients give us a measure of the uncertainty of the estimated coefficients.
   - **Standard Error for β_0 (Intercept) = 0.323**: This means there is some uncertainty in the estimate of the intercept. A lower value indicates a more precise estimate.
   - **Standard Error for β_1 (TV coefficient) = 0.002:** This indicates a very precise estimate of the slope (coefficient for TV), suggesting that the effect of TV advertising on sales is estimated with high accuracy.
6. **Confidence Intervals for the Coefficients:**
   - **95% Confidence Interval for β_0 (Intercept) = [6.339, 7.611]:** We are 95% confident that the true intercept lies between 6.339 and 7.611. This range includes the point estimate of 6.9748, which suggests the intercept is well-estimated.
   - **95% Confidence Interval for β_1 (TV coefficient) = [0.0517, 0.0592]:** We are 95% confident that the true effect of TV advertising on sales lies between 0.0517 and 0.0592. The interval does not include zero, reinforcing the idea that TV advertising has a statistically significant effect on sales.
7. **P-values for Coefficients:**
   - **P-value for β_0 (Intercept) = 0.000:** This value is extremely low, indicating that the intercept is highly statistically significant.
   - **P-value for β_1 (TV coefficient) = 0.000:** This p-value is also very low, indicating that the relationship between TV advertising and sales is statistically significant.

Interpretation:

- The model suggests a positive relationship between TV advertising and sales. For every additional $1,000 spent on TV advertising, sales are expected to increase by approximately 55.5 units (0.0555 thousand units).
- The R-squared value of 0.812 shows a strong fit of the model, meaning that TV advertising is a significant predictor of sales.
- The standard errors of the coefficients are small, indicating that the estimates are precise.
- The p-values for both the intercept and the coefficient of TV advertising are very small, which confirms the statistical significance of both.
- The confidence intervals provide a range for the estimated coefficients, and since the intervals do not contain zero, they further support the significance of the relationship between TV advertising and sales.