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

Multiple Linear Regression Analysis with Advertising Data

Objective: To perform multiple linear regression analysis using two models: one with limited predictors (TV and Radio) and one with all predictors (TV, Radio, and Newspaper). The experiment involves fitting models, extracting key metrics (coefficients, residuals, confidence intervals), and visualizing the results.

Task:

1. **Data Exploration and Preparation:**
    - Load the advertising dataset from the given CSV file.
    - Display the first few rows of the dataset.
2. **Limited Model (TV and Radio):**
    - Define the predictors as TV and Radio, and set the response variable as Sales.
    - Fit a multiple linear regression model using only TV and Radio as predictors.
    - Display the intercept and coefficients for the model.
    - Predict Sales using the model and calculate the residuals.
    - Use the statsmodels library to perform regression and display the model summary.
    - Extract and display the confidence intervals and standard errors of the model.
3. **Full Model (TV, Radio, and Newspaper):**
    - Define the predictors as TV, Radio, and Newspaper, and set the response variable as Sales.
    - Fit a multiple linear regression model using all three predictors.
    - Display the intercept and coefficients for the model.
    - Predict Sales using the model and calculate the residuals.
    - Use the statsmodels library to perform regression and display the model summary.
    - Extract and display the confidence intervals and standard errors of the model.
4. **Visualization:**
    - Create a 3D plot to visualize the regression plane and the residuals for the limited model.
    - Generate residual plots for both the limited and full models.

Discuss the results of the models, including the significance of the coefficients, residuals, and confidence intervals.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm

# Load Advertising Data from CSV file
data = pd.read_csv('/Advertising_data.csv')

# Display the first few rows of the dataset
print("First few rows of the data:")
print(data.head())

# Define predictors (X) and response (y) using only TV and Radio
X_limited = data[['TV', 'Radio']]
y = data['Sales']

# Fit multiple linear regression model using sklearn (limited predictors)
model_limited = LinearRegression()
model_limited.fit(X_limited, y)

# Print coefficients and intercept for the limited model
print("\nLimited Model Intercept:", model_limited.intercept_)
print("Limited Model Coefficients:", model_limited.coef_)

# Predict sales and calculate residuals for the limited model
data['Predicted_Sales_Limited'] = model_limited.predict(X_limited)
data['Residuals_Limited'] = data['Sales'] - data['Predicted_Sales_Limited']
```

```
# Model summary using statsmodels (limited predictors)
X_limited_with_intercept = sm.add_constant(X_limited)  # Add intercept manually for statsmodels
ols_model_limited = sm.OLS(y, X_limited_with_intercept).fit()

# Display the model summary for the limited model
print("\nLimited Model Summary:\n", ols_model_limited.summary())
```

```
# Regression analysis using all predictors (TV, Radio, and Newspaper)
X_full = data[['TV', 'Radio', 'Newspaper']]

# Fit the full model using sklearn
model_full = LinearRegression()
model_full.fit(X_full, y)

# Print coefficients and intercept for the full model
print("\nFull Model Intercept:", model_full.intercept_)
print("Full Model Coefficients:", model_full.coef_)

# Predict sales and calculate residuals for the full model
data['Predicted_Sales_Full'] = model_full.predict(X_full)
data['Residuals_Full'] = data['Sales'] - data['Predicted_Sales_Full']

# Model summary using statsmodels (full model)
X_full_with_intercept = sm.add_constant(X_full)
ols_model_full = sm.OLS(y, X_full_with_intercept).fit()

# Display the model summary for the full model
print("\nFull Model Summary:\n", ols_model_full.summary())

# Extract confidence intervals and standard errors for both models
print("\nLimited Model Confidence Intervals:\n", ols_model_limited.conf_int())
print("Limited Model Standard Errors:\n", ols_model_limited.bse)
print("\nFull Model Confidence Intervals:\n", ols_model_full.conf_int())
print("Full Model Standard Errors:\n", ols_model_full.bse)
```

```
# 3D Visualization of the regression plane with residuals for limited model
fig = plt.figure(figsize=(10, 6))
ax = fig.add_subplot(111, projection='3d')

# Plot the observed data points
ax.scatter(data['TV'], data['Radio'], data['Sales'], color='red', label='Observed Data')

# Generate grid for plotting the regression plane for the limited model
tv_range = np.linspace(data['TV'].min(), data['TV'].max(), 10)
radio_range = np.linspace(data['Radio'].min(), data['Radio'].max(), 10)
TV_grid, Radio_grid = np.meshgrid(tv_range, radio_range)

# Compute the predicted Sales for the grid points (limited model)
Sales_pred_limited = model_limited.intercept_ + model_limited.coef_[0] * TV_grid + model_limited.coef_[1] * Radio_grid

# Plot the regression plane for the limited model
ax.plot_surface(TV_grid, Radio_grid, Sales_pred_limited, color='lightblue', alpha=0.5)
```

```
# Plot residuals as vertical lines for the limited model
for i in range(len(data)):
    ax.plot([data['TV'][i], data['TV'][i]],
            [data['Radio'][i], data['Radio'][i]],
            [data['Sales'][i], data['Predicted_Sales_Limited'][i]],
            color='purple')

# Labels and legend
ax.set_xlabel('TV Advertising')
ax.set_ylabel('Radio Advertising')
ax.set_zlabel('Sales')
ax.legend()
plt.title('3D Regression Plane with Residuals (Limited Model)')
plt.show()
```

```
# Residual plot for the limited model
plt.figure(figsize=(10, 5))
plt.scatter(range(len(data)), data['Residuals_Limited'], color='orange', label='Residuals (Limited Model)')
plt.axhline(y=0, color='gray', linestyle='--')
plt.title('Residuals for Limited Model')
plt.xlabel('Data Points')
plt.ylabel('Residuals')
plt.legend()
plt.show()

# Residual plot for the full model
plt.figure(figsize=(10, 5))
plt.scatter(range(len(data)), data['Residuals_Full'], color='blue', label='Residuals (Full Model)')
plt.axhline(y=0, color='gray', linestyle='--')
plt.title('Residuals for Full Model')
plt.xlabel('Data Points')
plt.ylabel('Residuals')
plt.legend()
plt.show()
```

OUTPUT:

```
First few rows of the data:
      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

Limited Model Intercept: 4.630879464097768
Limited Model Coefficients: [0.05444896 0.10717457]
```

Limited Model Summary:

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.903
Model:	OLS	Adj. R-squared:	0.902
Method:	Least Squares	F-statistic:	912.7
Date:	Sat, 08 Feb 2025	Prob (F-statistic):	2.39e-100
Time:	04:00:35	Log-Likelihood:	-383.34
No. Observations:	200	AIC:	772.7
Df Residuals:	197	BIC:	782.6
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.6309	0.290	15.952	0.000	4.058	5.203
TV	0.0544	0.001	39.726	0.000	0.052	0.057
Radio	0.1072	0.008	13.522	0.000	0.092	0.123

Omnibus:	16.227	Durbin-Watson:	2.252
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27.973
Skew:	-0.434	Prob(JB):	8.43e-07
Kurtosis:	4.613	Cond. No.	425.

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

Full Model Intercept: 4.625124078808653
Full Model Coefficients: [0.05444578 0.10700123 0.00033566]

Full Model Summary:

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.903
Model:	OLS	Adj. R-squared:	0.901
Method:	Least Squares	F-statistic:	605.4
Date:	Sat, 08 Feb 2025	Prob (F-statistic):	8.13e-99
Time:	04:00:35	Log-Likelihood:	-383.34
No. Observations:	200	AIC:	774.7
Df Residuals:	196	BIC:	787.9
Df Model:	3		
Covariance Type:	nonrobust		

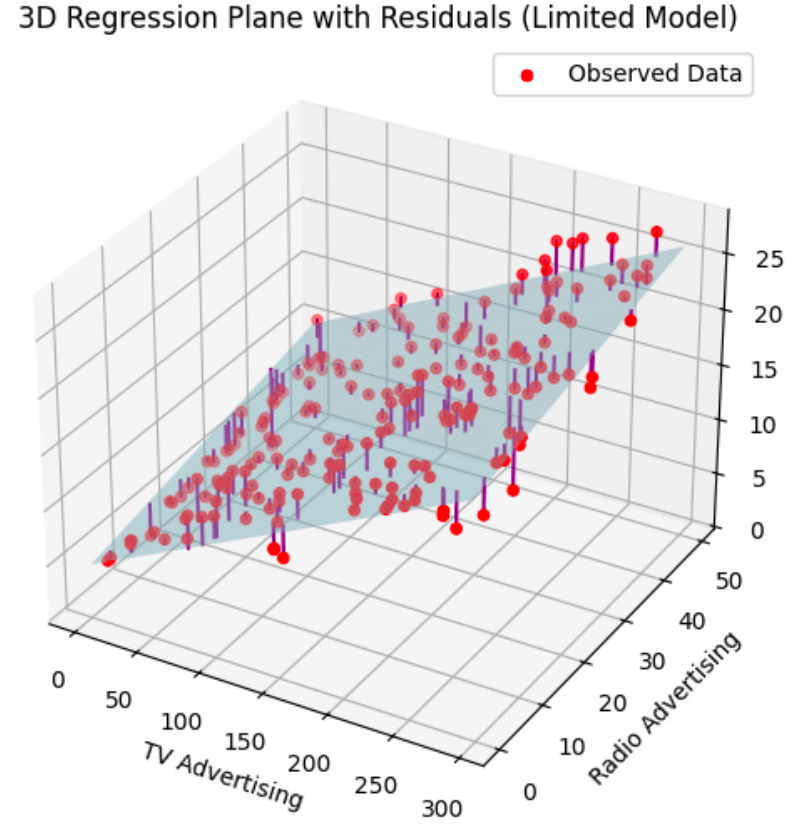
	coef	std err	t	P>|t|	[0.025	0.975]
const	4.6251	0.308	15.041	0.000	4.019	5.232
TV	0.0544	0.001	39.592	0.000	0.052	0.057
Radio	0.1070	0.008	12.604	0.000	0.090	0.124
Newspaper	0.0003	0.006	0.058	0.954	-0.011	0.012

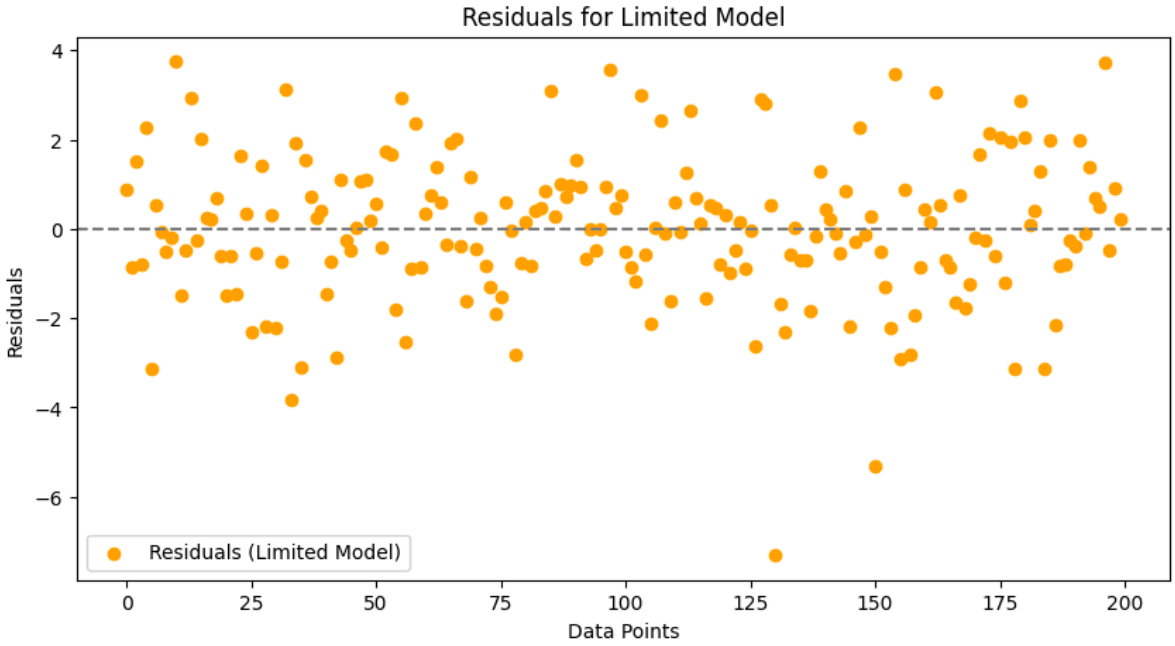
Omnibus:	16.081	Durbin-Watson:	2.251
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27.655
Skew:	-0.431	Prob(JB):	9.88e-07
Kurtosis:	4.605	Cond. No.	454.

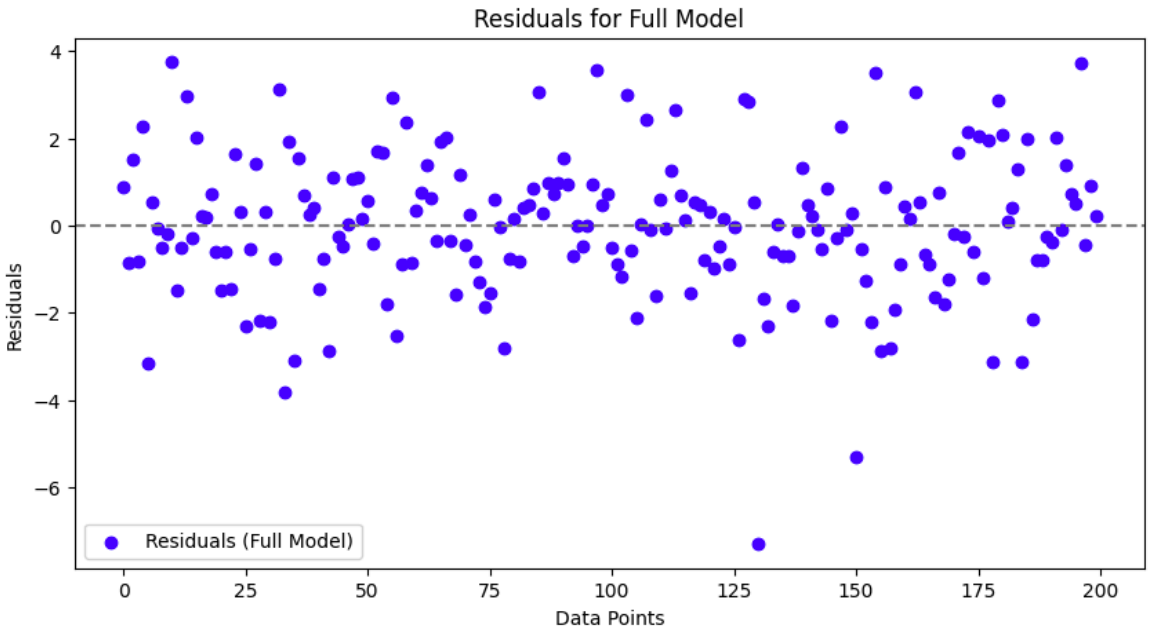
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```
Limited Model Confidence Intervals:
                0         1
const   4.058369  5.203390
TV      0.051746  0.057152
Radio   0.091544  0.122805
Limited Model Standard Errors:
 const     0.290308
TV        0.001371
Radio     0.007926
dtype: float64

Full Model Confidence Intervals:
                    0         1
const       4.018688  5.231560
TV          0.051734  0.057158
Radio       0.090259  0.123744
Newspaper -0.011079  0.011751
Full Model Standard Errors:
 const         0.307501
TV            0.001375
Radio         0.008490
Newspaper     0.005788
dtype: float64
```







RESULT ANALYSIS:

1. Limited Model Analysis (TV and Radio)

- **Intercept (4.6309)**: When there is no investment in TV and Radio advertising, the baseline expected sales are approximately **4.63 units**.
- **Coefficient for TV (0.0544)**: For every additional unit spent on TV advertising, sales increase by approximately **0.0544 units**, holding all other factors constant.

- **Coefficient for Radio (0.1072)**: For every additional unit spent on Radio advertising, sales increase by approximately **0.1072 units**, holding other factors constant.
- **Model Statistics:**
    - **R-squared (0.903)**: The model explains **90.3%** of the variance in sales, indicating a very strong fit.
    - **Adjusted R-squared (0.902)**: Even after adjusting for the number of predictors, the model maintains high explanatory power.
    - **F-statistic (912.7)**: A large F-statistic with a near-zero p-value implies that the model is statistically significant.
    - **Standard Errors** for TV and Radio coefficients are low, implying precise estimates.
- **Confidence Intervals:**
    - TV (0.0517, 0.0571)
    - Radio (0.0915, 0.1228)
      These narrow intervals show that both predictors are highly reliable.

---

2. Full Model Analysis (TV, Radio, Newspaper)

- **Intercept (4.6251)**: Similar to the limited model, the baseline expected sales are around **4.63 units** when advertising expenditures are zero.
- **Coefficient for TV (0.0544)**: This value is consistent with the limited model, indicating a positive relationship between TV advertising and sales.
- **Coefficient for Radio (0.1070)**: Similar to the limited model, Radio remains an important predictor.
- **Coefficient for Newspaper (0.0003)**: The effect of newspaper advertising on sales is negligible and statistically insignificant ($p = 0.954$).
- **Model Statistics:**
    - **R-squared (0.903)**: The model explains **90.3%** of the variance in sales, the same as the limited model.
    - **Adjusted R-squared (0.901)**: Slightly lower than the limited model, indicating that adding Newspaper does not improve model performance.
    - **F-statistic (605.4)**: Still statistically significant, but not as high as in the limited model.
- **Confidence Intervals:**
    - Newspaper (-0.011, 0.012): The interval includes zero, reinforcing the insignificance of this predictor.

Comparison of Models

Metric	Limited Model	Full Model
Intercept	4.63	4.63
TV Coef	0.0544	0.0544
Radio Coef	0.1072	0.107
Newspaper Coef	—	0.0003
R-squared	0.903	0.903
Adj R-squared	0.902	0.901
AIC	772.7	774.7
BIC	782.6	787.9

Final Conclusion:

- **Model Selection:** The **Limited Model** (with TV and Radio only) is the better choice.
- **Reasons:**
    1. **Insignificance of Newspaper:** The p-value for Newspaper is 0.954, indicating it does not contribute meaningfully to the prediction of sales.
    2. **Simplicity:** The limited model is simpler and easier to understand while still explaining the data effectively.
    3. **AIC/BIC Comparison:** The lower AIC (772.7) and BIC (782.6) values for the limited model indicate better model efficiency.
    4. **Adjusted R-squared:** The limited model slightly outperforms the full model here, reaffirming that the additional predictor does not add value.

The **Limited Model** is the preferred model due to its simplicity, statistical significance, and efficient fit without overfitting the data.