University of Moratuwa, Sri Lanka

Faculty of Engineering

Department of Electronic and Telecommunication Engineering

Semester 5

EN3150 – Pattern recognition

Assignment 01

Learning from data and related challenges and linear models for egression

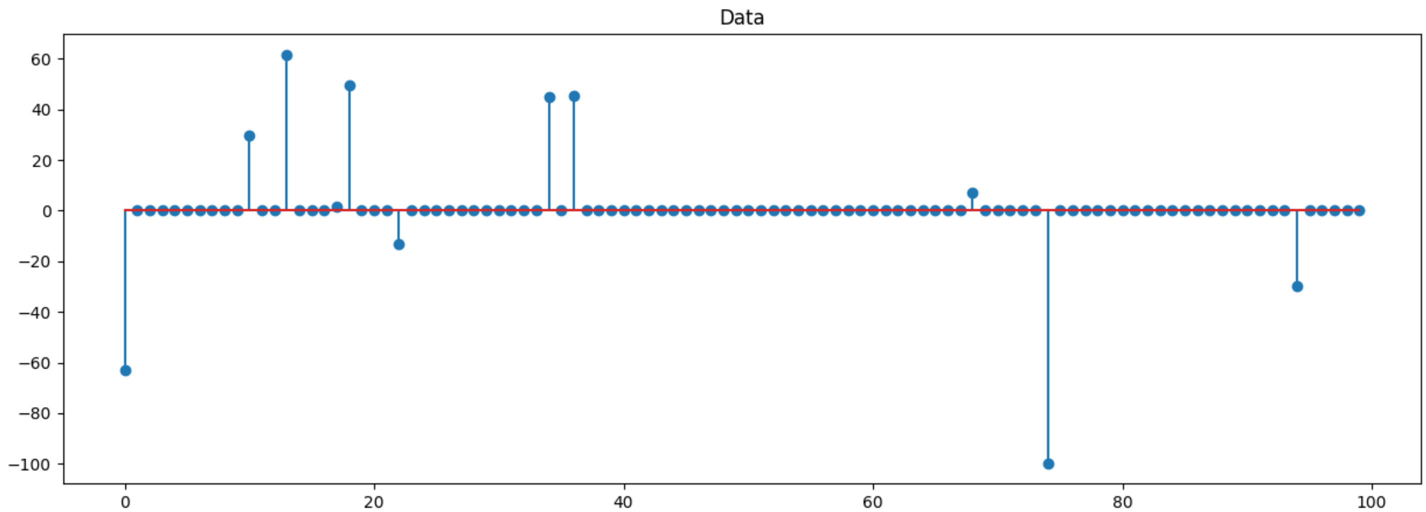
G.K.M.I.D.Rajarathna

200500L

Data Preprocessing

Question 01

Index number = 200500L



Question 03

```
min of data -99.74731349246915 max of data 61.57082897086394
mean of data 0.34175454785447973 std of data 16.19075843901647
```

Data normalization

- <u>MaxAbsScaler</u>

  We can apply this scaling method by using the sklearn.preprocessing library

- <u>Min-Max Scaler</u>

Min-Max scaling equation

$$x_{\text{scaled}} = \frac{x - \min(x)}{\max(x) - \min(x)}.$$

Where:

x_{scaled} is the scaled value of x.
x is the original value.
$\min(x)$ is the minimum value in the data.
$\max(x)$ is the maximum value in the data.

NumPy library contain min( ), and max( ) functions.

- <u>Standard Scaler</u>

Standardization scaling equation

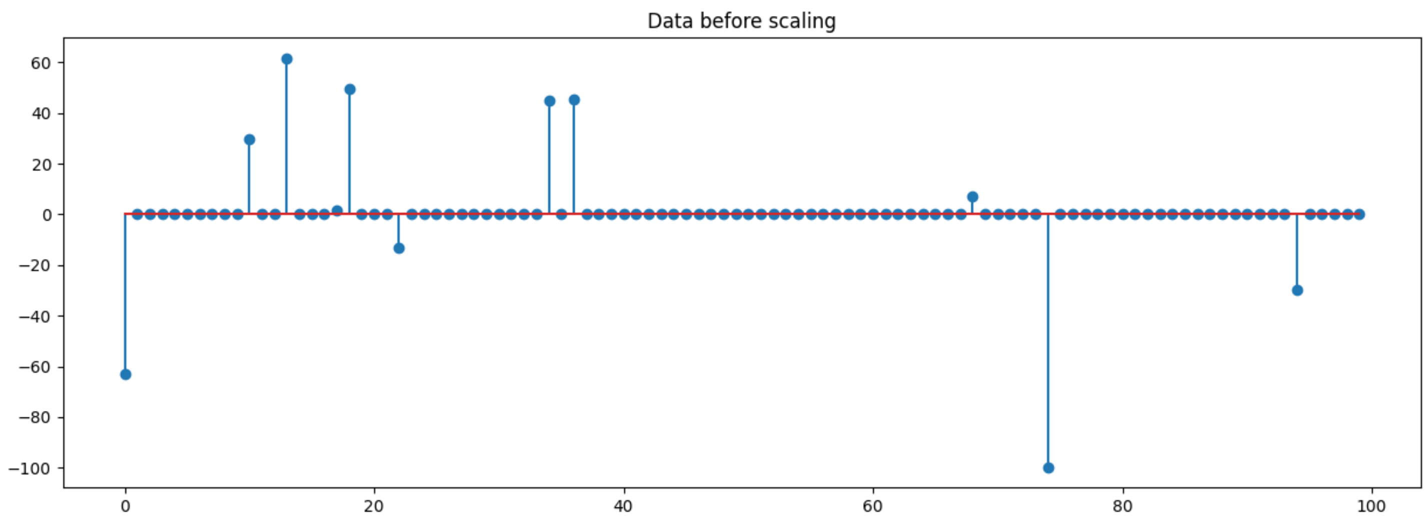
$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}.$$

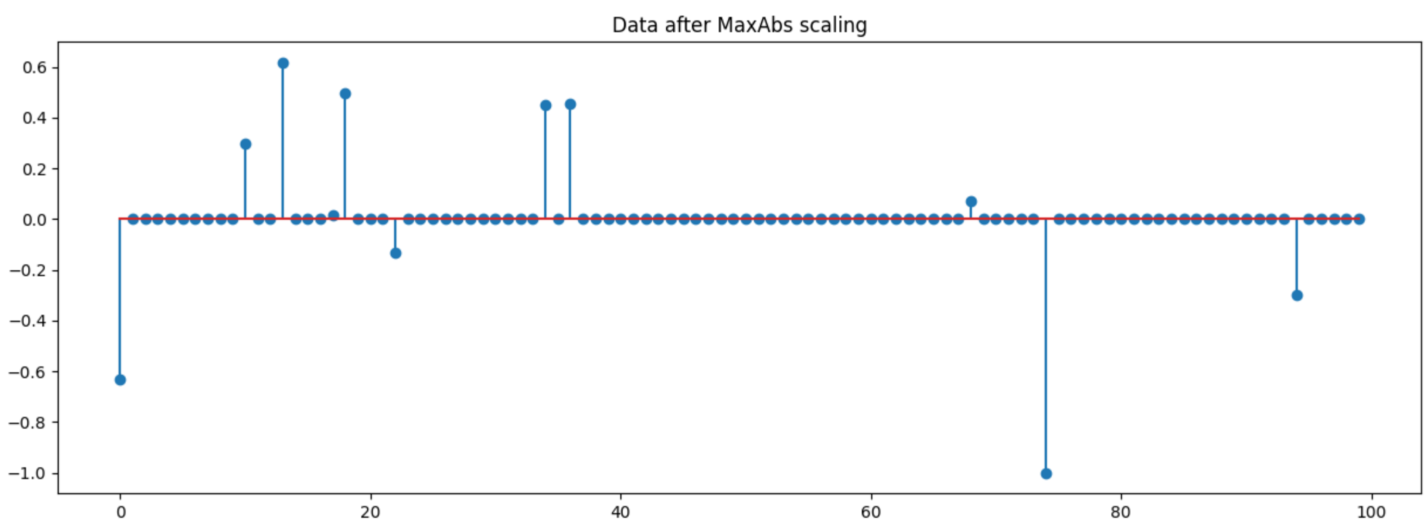
Where:

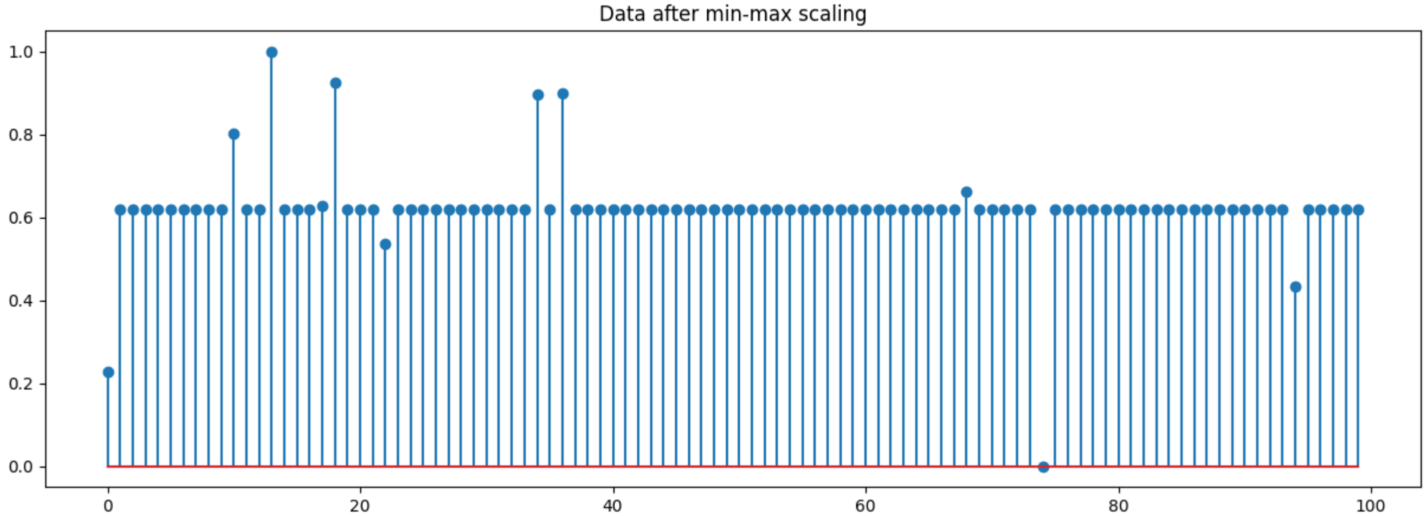
x_{scaled} is the scaled value of x.
x is the original value.
μ is the mean (average) of the data.
σ is the standard deviation of the data.

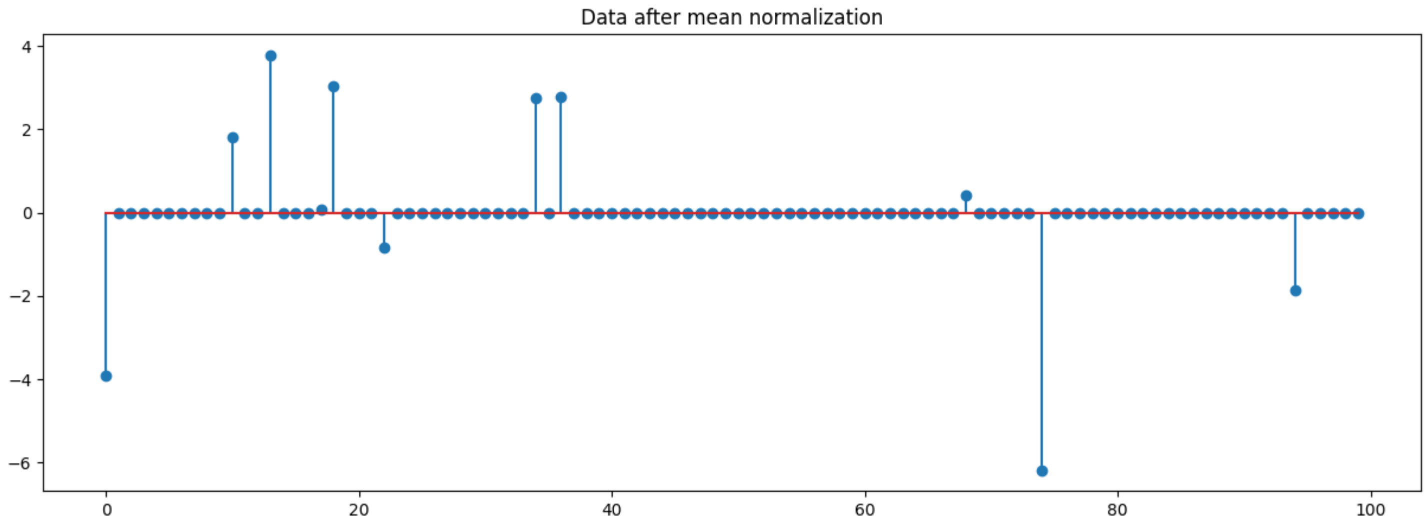
Used mean( ) and std( ) function used that available in NumPy library.

Question 04









Question 05

```
Number of non-zero elements in the signal:  11
Number of non-zero elements in the scaled_data:  11
Number of non-zero elements in the scaled_data_min_max:  99
Number of non-zero elements in the scaled_data_mean_norm:  100
```

Question 06

<u>MaxAbs Scaler</u>

MaxAbs Scaler is a data preprocessing technique that scales data within the [-1, 1] range, preserving the sign of each data point and not altering the data distribution. While it can be useful for some applications, especially when the sign of the data is important, it does not have the property of reducing the influence of outliers. Outliers will still have a significant impact on the scaled data, and their presence may affect subsequent data analysis or modeling.

<u>Min-Max Scaler</u>

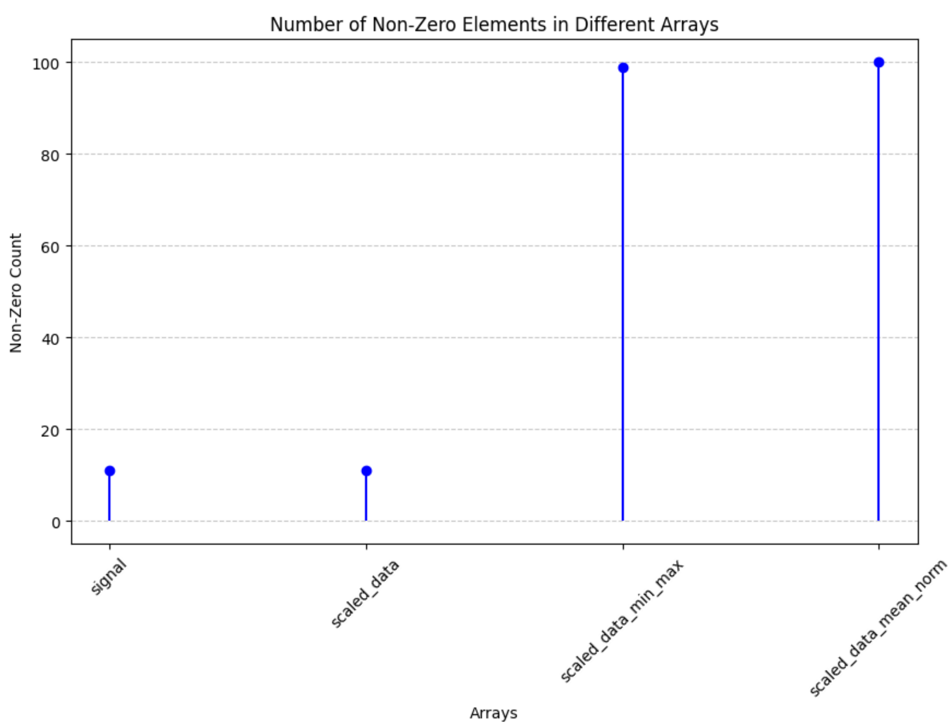
MinMaxScaler is a useful technique for scaling data to a specific range, such as [0, 1], while preserving data distribution. However, it is sensitive to outliers and can lead to skewed scaling when outliers are present.

<u>Standard Scaler</u>

standard scaling is a useful technique for centering data around zero and standardizing the spread of data, which is often beneficial for various machine learning algorithms. However, it should be used with caution when dealing with datasets that contain outliers, as the presence of outliers can distort the computed mean and standard deviation, affecting the quality of the scaling. In such cases, robust scaling techniques that are less sensitive to outliers, such as the Robust Scaler, might be more appropriate.

Code outputs

```
           signal
count  100.000000
mean     0.341755
std     16.272324
min    -99.747313
25%      0.000000
50%      0.000000
75%      0.000000
max     61.570829

       scaled_data_min_max
count           100.000000
mean              0.620445
std               0.100871
min               0.000000
25%               0.618327
50%               0.618327
75%               0.618327
max               1.000000

       scaled_data_mean_norm
count           1.000000e+02
mean            6.591949e-18
std             1.005038e+00
min            -6.181864e+00
...
25%       0.000000
50%       0.000000
75%       0.000000
max       0.617268
```



Question 07

In this context, where there are no apparent outliers in the dataset, our primary goal is to transform the data into a more convenient range for computational processing. To achieve this, the MaxAbs Scaler is selected as the preferred scaling method. The MaxAbs Scaler scales the data to fit within the range of [-1, 1], which can simplify subsequent calculations and model training.

Furthermore, this dataset exhibits a significant number of zero values, indicating a degree of sparsity. Preserving these zero values is advantageous for two key reasons. Firstly, it ensures accurate calculations, as zero values may carry specific meaning or contribute to the overall data characteristics. Secondly, retaining zero values reduces computational expenses, particularly at the hardware level, which can be crucial for efficiency.

Linear Regression on Real World Data

Question 01

```
   sample index     TV  radio  newspaper  sales
0             1  230.1   37.8       69.2   22.1
1             2   44.5   39.3       45.1   10.4
2             3   17.2   45.9       69.3    9.3
3             4  151.5   41.3       58.5   18.5
4             5  180.8   10.8       58.4   12.9
```

Question 02

This part of the code is used to split the data set

```
# Split the data into training (80%) and testing (20%) sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# The 'test_size' parameter specifies the proportion of data to be allocated to the test set.
# 'random_state' is used to ensure reproducibility; we can choose any integer value.
```

```
X_train dimensions:  (160, 3)
y_train dimensions:  (160,)
X_test dimensions:  (40, 3)
y_test dimensions:  (40,)
```

Question 03

Linear Model that generated by the LinearRegression() function,

$$y = w_0 + w_1x_1 + w_2x_2 + w_3x_3$$

$$x_1 = TV$$

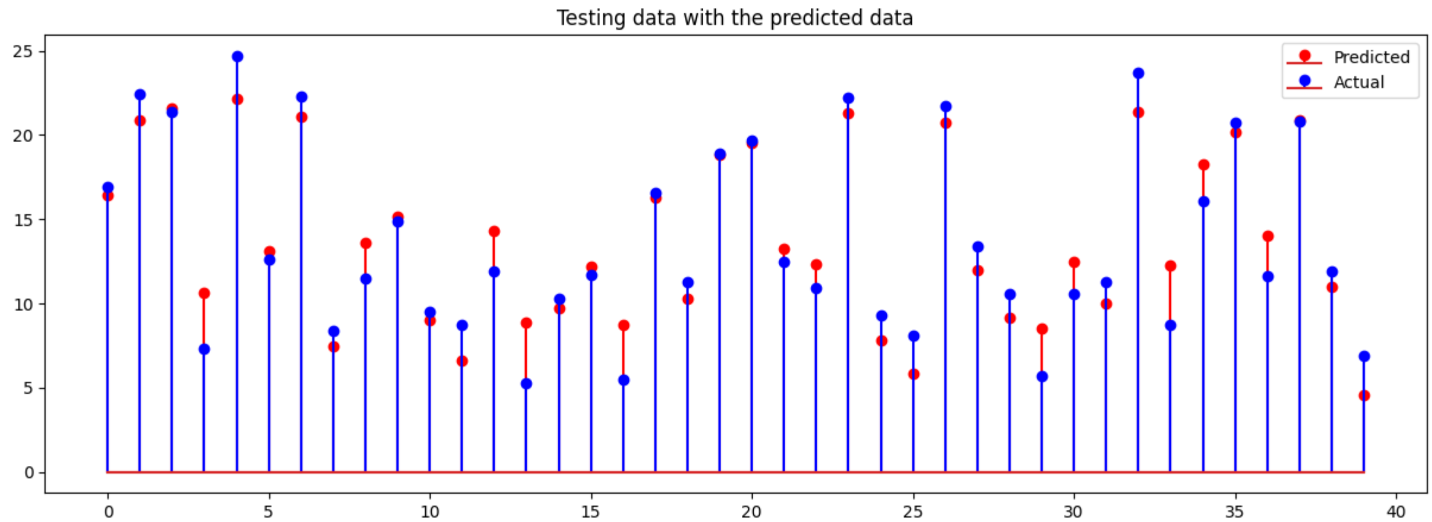
$$x_2 = radio$$

$$x_3 = newspaper$$

Model parameters,

```
Coefficients:
TV: 0.044729517468716326
Radio: 0.18919505423437652
Newspaper: 0.0027611143413671935
Intercept: 2.979067338122629
```

Question 04



Now apply generated model to testing data,

```
d: 3

RSS: 126.96389415904413

RSE: 1.8779709363435915

MSE: 3.34115510944853

r^2: 0.899438024100912

Standard error of coefficients:
TV: 0.0007271180958166518
Radio: 0.005424195529624446
Newspaper: 0.003955416709697839

t-statistic:
TV: 61.51616597917156
Radio: 34.87983668750153
Newspaper: 0.6980590274085482

p-values:
TV: 0.0
Radio: 0.0
Newspaper: 0.48939013301949563
```

Note that RSE is given by

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{N-d}}. \tag{3}$$

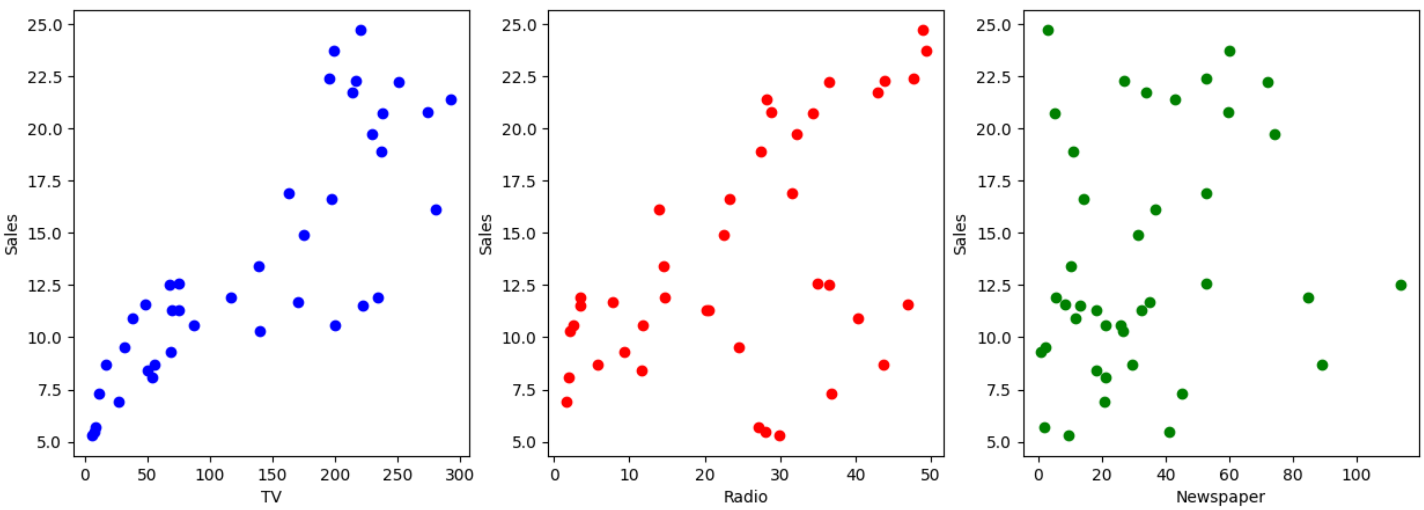
Here, N is the total number of data samples and d is the number of model parameters. As an example for a model $y = w_0 + w_1 x + \epsilon$, there are two model parameters namely w_0 and w_1.

```
d = X_test.shape[1] # number of features
print("d:", d)
print()
# residual sum of squares (RSS)
RSS = np.sum(np.square(y_test - testing))
print("RSS:", RSS)
print()

# residual standard error (RSE)
RSE = np.sqrt(RSS / (len(y_test) - 4))
print("RSE:", RSE)
print()

# mean squared error (MSE)
MSE = RSS / (len(y_test) - 2)
print("MSE:", MSE)
print()
```

Question 05

When analyzing the p-values associated with the model parameters $w1$, $w2$, and $w3$, representing the advertising budgets for TV, radio, and newspaper, important insights can be gleaned. Notably, $w1$ and $w2$ exhibit remarkably low p-values, signifying their high statistical significance concerning both the training and testing datasets. This statistical significance implies a meaningful and robust relationship between the advertising budgets for TV and radio and the resulting sales.

Conversely, the p-value for $w3$, corresponding to the advertising budget for newspapers, fails to reach the conventional levels of statistical significance (typically set at 0.05 or 0.01). Consequently, there is insufficient evidence to reject the Null Hypothesis ($H0$) in this context. This suggests that the advertising budget for newspapers does not demonstrate a substantial and statistically significant impact on sales. In other words, the relationship between newspaper advertising expenditures and sales may be weaker or non-existent compared to TV and radio advertising, as indicated by the p-value analysis.



Question 06

according to the provided linear model, radio advertising has the highest impact on sales, followed by TV advertising, while newspaper advertising has the least influence. These conclusions are based on the respective coefficients, which indicate the strength and direction of the relationship between each advertising channel's budget and the resulting sales.

Question 07

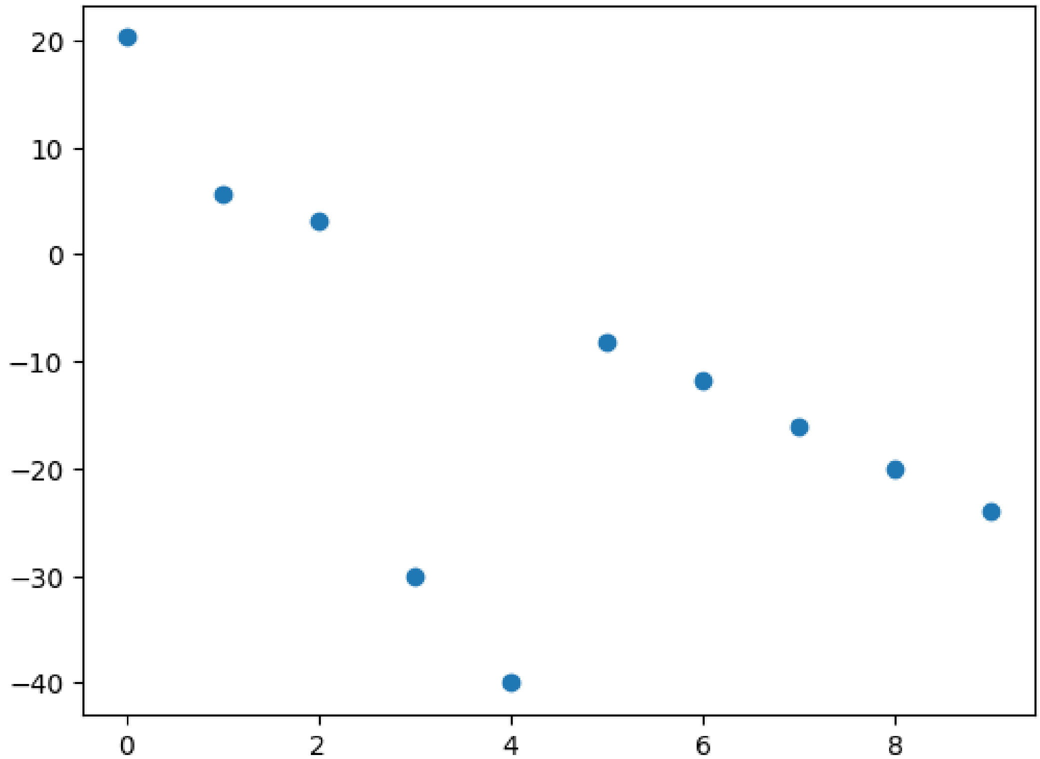
scenario_1 = $25,000 allocated to TV, $25,000 to radio, $0 to newspaper
scenario_2 = $50,000 allocated to TV, $0 to radio, $0 to newspaper

```
Scenario 1 Predicted Sales ($): [8827.1816307]
Scenario 2 Predicted Sales ($): [5215.54321156]
$25,000 allocated to TV, $25,000 to radio, $0 to newspaper is better.
```

based on the output scenario_1 is better.

Linear regression impact on outliers

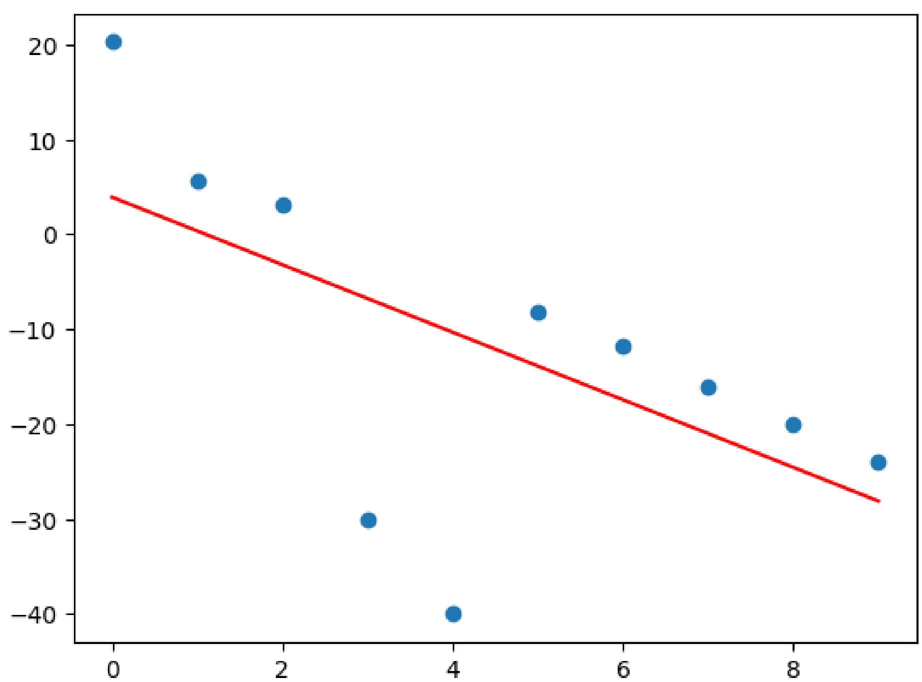
Question 01



Question 02

```
Intercept: 3.916727272727277
Coefficient: [-3.55727273]
```

So generated model is $\boxed{y = -3.55 + 3.91}$



Question 03

Model 1: y = −4x+12
Model 2: y = −3.55x+3.91

$$L(\theta,\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{(y_i-\hat{y}_i)^2}{(y_i-\hat{y}_i)^2+\beta^2}\right).$$

```
def new_loss_reduce_outliers(theta, Beta, x, y, N):
    y_hat = theta[0] + theta[1] * x  # Calculate predicted y values for all data points
    residuals = y - y_hat
    squared_residuals = residuals**2

    # Calculate the loss for all data points using vectorized operations
    loss = np.sum(squared_residuals / (squared_residuals + Beta)) / N

    return loss
```

Question 04

```
Loss for w1: 0.435416262490386
Loss for w2: 0.9728470518681676
```

Loss for model 1 = 0.43542
Loss for model 1 = 0.97285

Question 05

Since the loss for the data set is smaller for the model 1, model 1 is better than model 2 for the data set.

Question 06

A robust estimator reduces the impact of outliers by downweighting or giving less influence to data points that deviate significantly from the majority of the data. This is achieved through various techniques and loss functions that are less sensitive to extreme values compared to traditional least squares estimators. One common robust estimator is the Huber loss, which is less affected by outliers than the mean squared error (MSE) used in ordinary least squares (OLS) regression. Here's how a robust estimator reduces the impact of outliers:

1. **Outlier-Resistant Loss Function:** Robust estimators use loss functions that have a bounded influence on outliers. The loss function is designed to penalize large errors less severely than the squared error loss used in OLS. For example, the Huber loss combines a quadratic loss for small errors and a linear loss for large errors. This means that outliers have less impact on the overall loss.

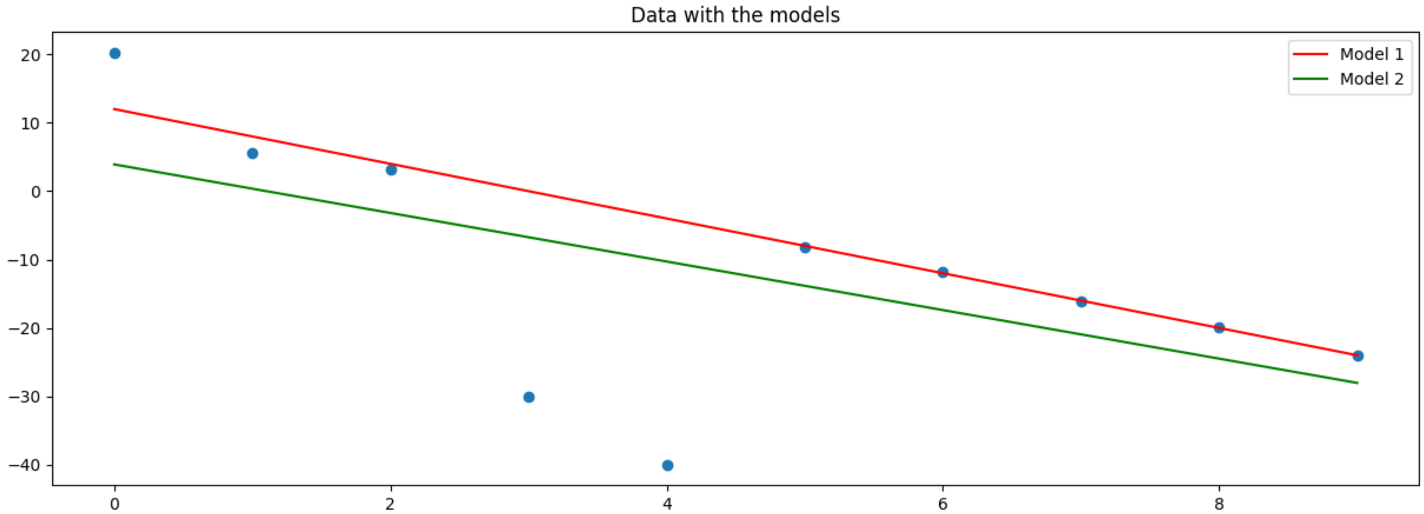
2. **Weighted Least Squares:** In some robust regression techniques, such as weighted least squares (WLS), data points are assigned different weights. Outliers are assigned lower weights, which means they contribute less to the estimation of model parameters. This downweighting effectively reduces the influence of outliers on the parameter estimates.

3. **Trimming or Winsorizing:** Another approach is to identify and exclude extreme outliers from the analysis. For example, you might trim the top and bottom 5% of data points or apply Winsorization, which caps the extreme values. By doing this, you reduce the impact of outliers on the parameter estimates.

4. **M-Estimation:** Robust estimators often use M-estimation, which is a general method for estimating parameters that minimize a robust objective function. The objective function is chosen to be robust against outliers, making the parameter estimates less sensitive to extreme values.

5. **Robust Residuals:** Some robust estimators use robust residuals, which are less affected by outliers. For example, the weighted residual sum of squares (WRRSS) gives less weight to the residuals associated with outliers, leading to more robust parameter estimates.

Question 07



Question 08

1. **β Helps Downweight Outliers:** The primary purpose of the hyperparameter β is to control the extent to which outliers in the data influence the loss function. In robust regression, outliers can significantly affect the parameter estimates, and β is used to adjust this impact.

2. **Effect of High β:** When β is set to a high value, the loss function becomes similar to the Mean Squared Error (MSE) loss function, which is commonly used in ordinary least squares (OLS) regression. In other words, as β increases, the contribution of the squared difference between the observed (actual) values and the predicted values (residuals) increases. This means that the loss function becomes more sensitive to outliers, similar to the way MSE is.

3. **Effect of Small β:** Conversely, when β is set to a small value, the loss function behaves differently. In this case, the loss function becomes less influenced by the β component, and it mainly relies on the squared difference between observed and predicted values. This can be problematic because it doesn't effectively downweight outliers. The loss function behaves similarly to a standard squared error loss, which is sensitive to outliers.

4. **Choosing an Appropriate β Value:** To effectively downweight outliers and balance the influence of the squared differences and the β component, it is crucial to choose an appropriate value for β based on the nature of the data and the distribution of the residuals. The example provided illustrates that if the squared differences fall within a certain range (e.g., -10 to 10), a specific β value may be suitable. However, if the squared differences are larger (e.g., -100 to 100), a different β value may be more appropriate.