UNIVERSITY OF MORATUWA, SRI LANKA

Faculty of Engineering

Department of Electronic and Telecommunication Engineering

Semester 5 (Intake 2020)

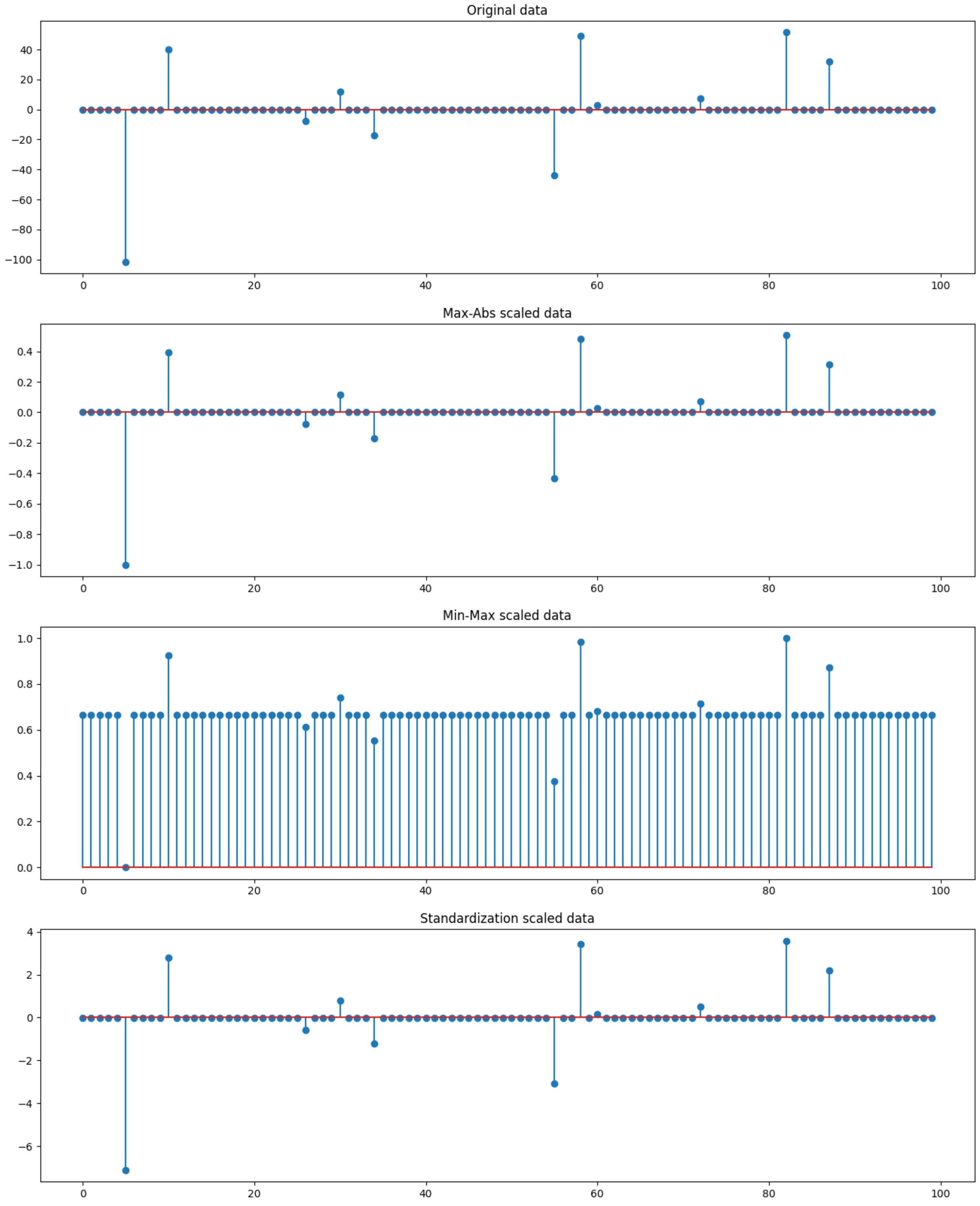
EN3150 – Pattern recognition

Assignment 01
Learning from data and related challenges and linear models for regression

AMARASEKARA A. T. P.
200023C

1. Data Preprocessing

Visualization of data before and after each normalization method:



No. of non-zero elements in the data before the normalization : 11
No. of non-zero elements in the data after the MaxAbs Scaler normalization : 11
No. of non-zero elements in the data after the Min-Max scaler normalization : 99
No. of non-zero elements in the data after the Standard normalization : 100

MaxAbs Scaler:
This method scales features by dividing each value by the maximum absolute value in that feature column.
Impact on Data:

- The range of the scaled data will be between -1 and 1, preserving the relative relationships between data points.

Min-Max Scaler:
This method scales features to a specified range, typically [0, 1], by subtracting the minimum value and then dividing by the range (max - min).
Impact on Data:

- It maps the data to a fixed range, which can be useful when you need to compare or visualize data on a consistent scale.
- Preserves the relative relationships between data points.
- Sensitive to outliers since it depends on both the minimum and maximum values in the column.

Standard Normalization:
This method scales features to have a mean of 0 and a standard deviation of 1 by subtracting the mean and dividing by the standard deviation.
Impact on Data:

- Centers the data around zero and scales it to have a standard deviation of 1, making it suitable for algorithms that assume normal, Gaussian distributions or require zero-centered data.
- The relative relationships between data points may change, especially if the original data has a skewed distribution.
- It is sensitive to outliers since it relies on the mean and standard deviation.

	MaxAbs Scaler	Min-Max Scaler	Standard Normalization
Distribution	Preserves the distribution shape, as it scales each value relative to the maximum absolute value in the signal.	Changes the original distribution shape	Alters the original distribution shape
Structure	The relative relationships between data points remain same	Preserves the relative relationships between data points	May change the relative relationships between data points, especially if the data was not normally distributed
Scale	Scaled to a range of [-1, 1]	Scales the data to a fixed range, [0, 1]	Data is centered around zero with a standard deviation of 1

As the MaxAbs scaler scales the data obtained to a range of [-1,1] which is a convenient range to work with, preserves the distribution shape and the relationship between data points and conserve the sign of the data points, MaxAbs scaler is recommended for this collection of data.

2. Linear regression on real world data

Data was loaded, split into training and testing data and the training data set was used to train a linear regression model which resulted in the following intercept and coefficients.

Intercept: 2.979067338122629
Coefficients:

TV	radio	newspaper
0.04472952	0.18919505	0.00276111

The trained model was evaluated using the testing data and the following statistics were obtained.

Training Data:	Testing Data:
RSS : 432.8207	RSS : 126.9639
RSE : 1.6551	RSE : 1.8279
MSE : 2.7051	MSE : 3.1741
R^2 : 0.8957	R^2 : 0.8994
Standard Errors :	Standard Errors :
w_0 : 0.3513	w_0 : 0.7106
TV : 0.0016	TV : 0.0033
radio : 0.0096	radio : 0.0200
newspaper : 0.0070	newspaper : 0.0114
t-Statistics :	t-Statistics :
w_0 : 8.4808	w_0 : 4.1924
TV : 28.7260	TV : 13.7492
radio : 19.6427	radio : 9.4377
newspaper : 0.3943	newspaper : 0.2425
p-Values :	p-Values :
w_0 : 1.5099e-14	w_0 : 0.0002
TV : 0.0	TV : 2.2204e-16
radio : 0.0	radio : 1.6629e-11
newspaper : 0.6939	newspaper : 0.8097

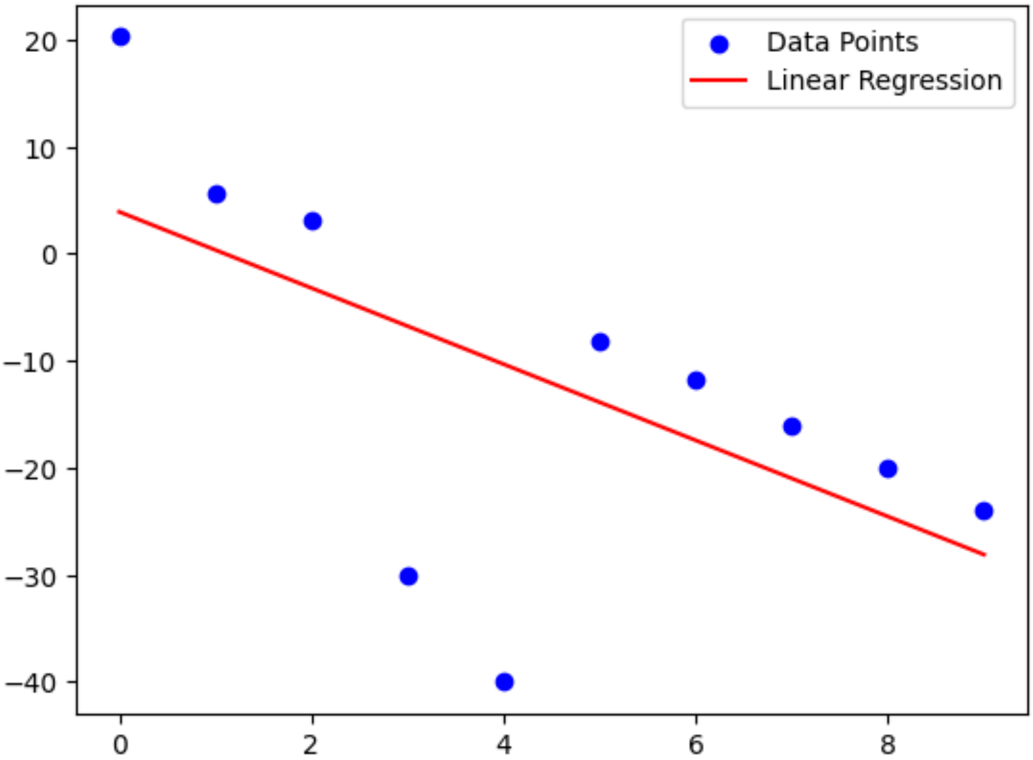
There is a relationship between advertising budgets and sales, as indicated by the high R-squared (R^2) value for both the training and testing data sets. However, when considering the p-values, for TV and radio advertising budgets, p-values are significantly low indicating a significant impact on the sales, whereas the p-value for newspaper advertising budget is considerably higher indicating the presence of no statistical significance on the sales.

Both TV and radio advertising budgets have extremely low p-values, indicating that they are highly statistically significant in predicting sales. The t-statistics for TV and radio are also considerably high, indicating a strong positive relationship between these variables and sales. As the units of the two features are the same, the regression coefficients can be used to evaluate which variable contributes the most. Therefore, as the coefficient for the **radio advertising budget** is highest, it has the most contribution to the sales.

Estimated sales by allocating 25,000 dollars each for both television and radio advertising: [8.85141627]
Estimated sales by allocating 50,000 dollars for television advertising : [5.16304126]

Estimated sales by allocating 50,000 dollars for radio advertising: [12.53979129]
By the estimated sales, the sales are highest when 50,000 dollars budget is allocated for the radio advertising in comparison to allocating 25,000 dollars each for both television and radio advertising. Therefore, the **claim is false**.

3. Linear regression impact on outliers



Plot of Data points and Linear regression model

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

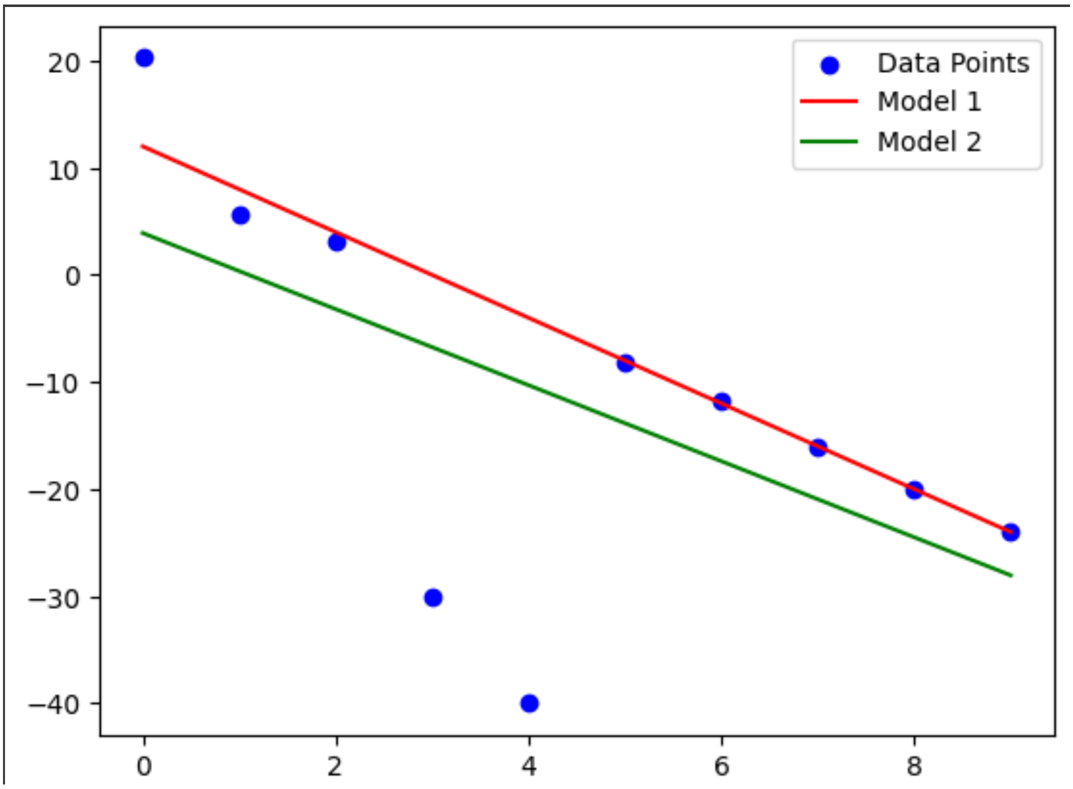
By loss function,
Loss for Model 1: [0.43541626]
Loss for Model 2: [0.97284705]

By the loss function values, the loss of model 1 is lower compared to model 2, meaning model 1 performs better in reducing the error of the predicted value. Therefore, **model 1 is most suitable for the dataset**.

Robust estimator reducing the impact of the outliers
Normally, we utilize Ordinary Least Square method to find the model parameter which minimize the Mean Squared Error (MSE). Here, as the difference between actual and predicted values are squared, the MSE will be higher of there's an outlier. Hence, we will not be able to identify the most ideal model in the presence of outliers.
In the robust estimator, we try to find the model parameters which minimizes the following loss function,

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

Where in the presence of an outlier the effect of that outlier on the loss function will be downweighed by the term in the denominator., by which we can obtain the most suitable model parameters. Therefor this robust estimator reduces the impact of the outliers.



Plot of Data points and Models

Regularization parameter

β serves as a regularization parameter in the context of the given loss function.

Smaller β values results in a loss function more sensitive to outliers, meaning outliers have a greater influence on the model parameters (θ). Smaller β values result in a less robust estimator.

When β is large, the regularization term β^2 dominates the loss function. This reduces the sensitivity of the loss function to outliers, effectively reducing their impact. Larger β values result in a more robust estimator that is less influenced by outliers.