ELECTRONIC & TELECOMMUNICATION ENGINEERING
UNIVERSITY OF MORATUWA

EN3150 Pattern Recognition

Assignment 01

Learning From Data and Related Challenges and Linear Models for Regression

A.A.W.L.R. Amarasinghe

210031H

1 Data Pre-Processing

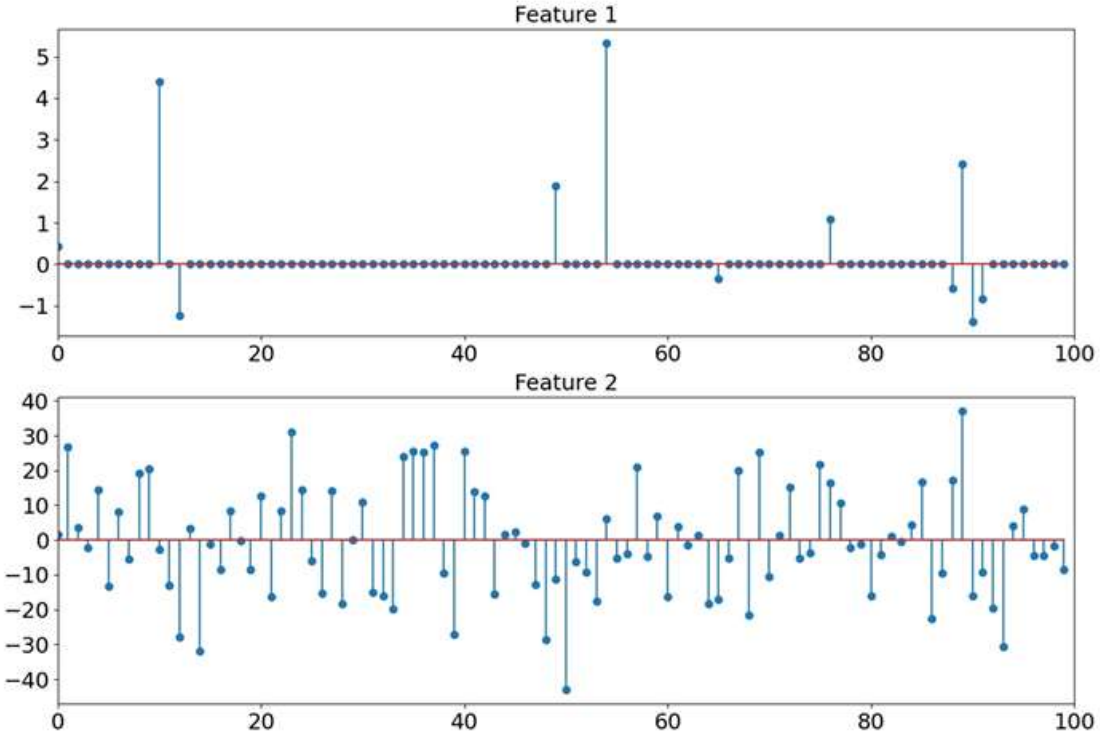


Figure 1: Feature values of a dataset.

Standard Scaling (Z-score normalization) is the most suitable method for both features, as it maintains the structure and characteristics of each while applying a uniform transformation.

Application to Feature 1: Although most of the values in Feature 1 are near zero with some outliers, standard scaling is effective because it centers the data around zero and scales it based on the standard deviation. This approach reduces the influence of outliers by adjusting them relative to the entire distribution.

Application to Feature 2: In the case of Feature 2, which has a wider range and greater variation, standard scaling normalizes the data so it has a mean of zero and a standard deviation of one, making it well-suited for this feature.

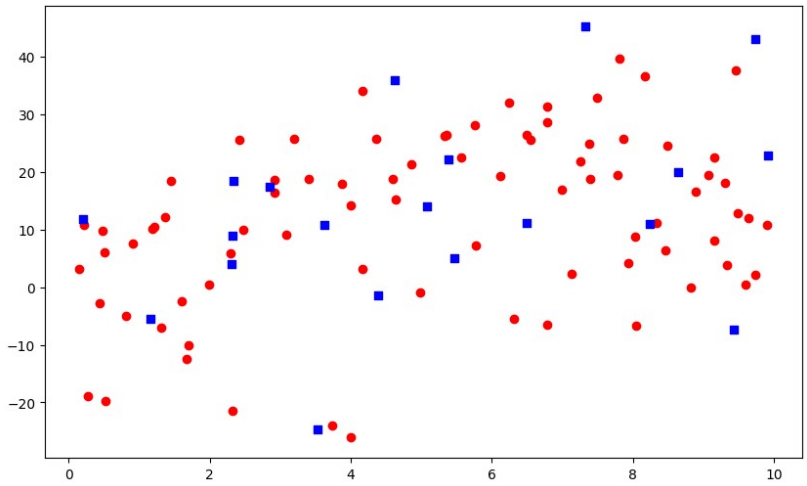
Justification:

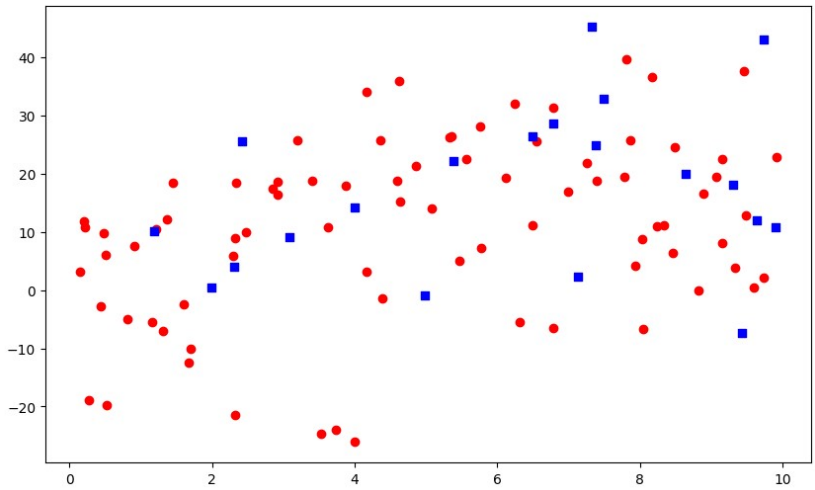
- **Consistency**: Using the same scaling method across both features ensures uniformity, which is important in machine learning models that expect features to be on a similar scale.
- **Outlier Impact**: Standard scaling helps minimize the effect of outliers by normalizing the distribution.
- **Feature Comparison**: Since Feature 2 has more variation and a broader range, standard scaling brings it to a comparable scale with Feature 1, facilitating better integration and comparison in subsequent analysis.

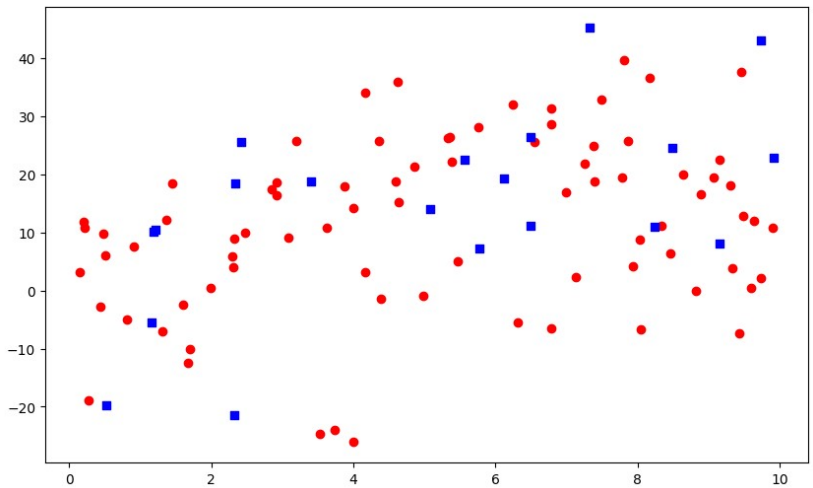
Therefore, Standard Scaling is the best choice for both features, as it standardizes their distributions, retains their relative variances, and effectively manages differences in their value ranges.

2 Learning from Data

2.Observations from Running Listing 2 Multiple Times



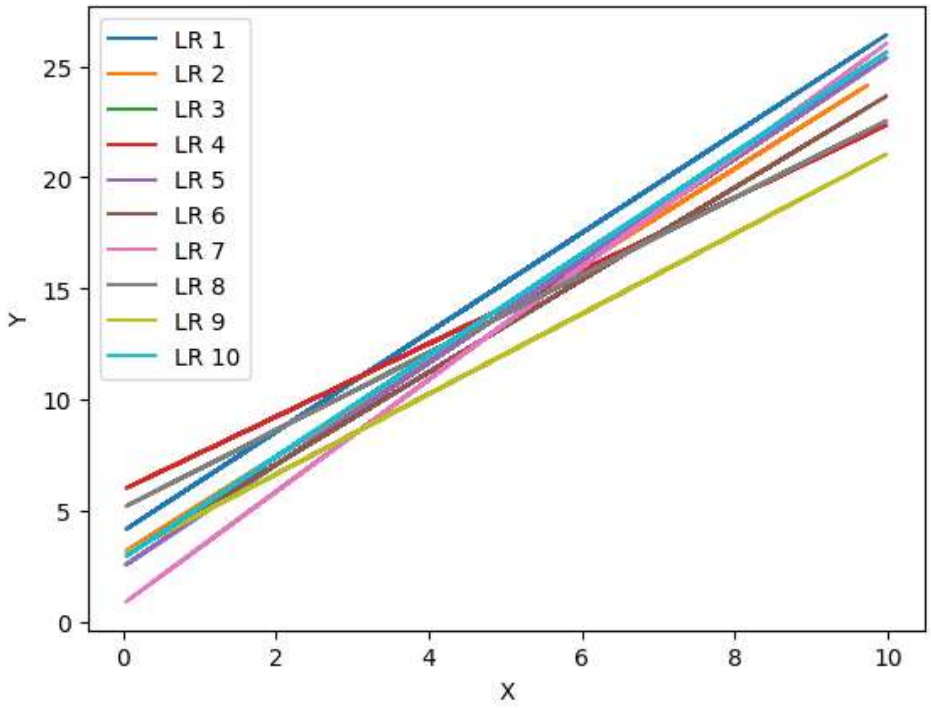




2.Reason for Differences in training and testing data in Each Run:

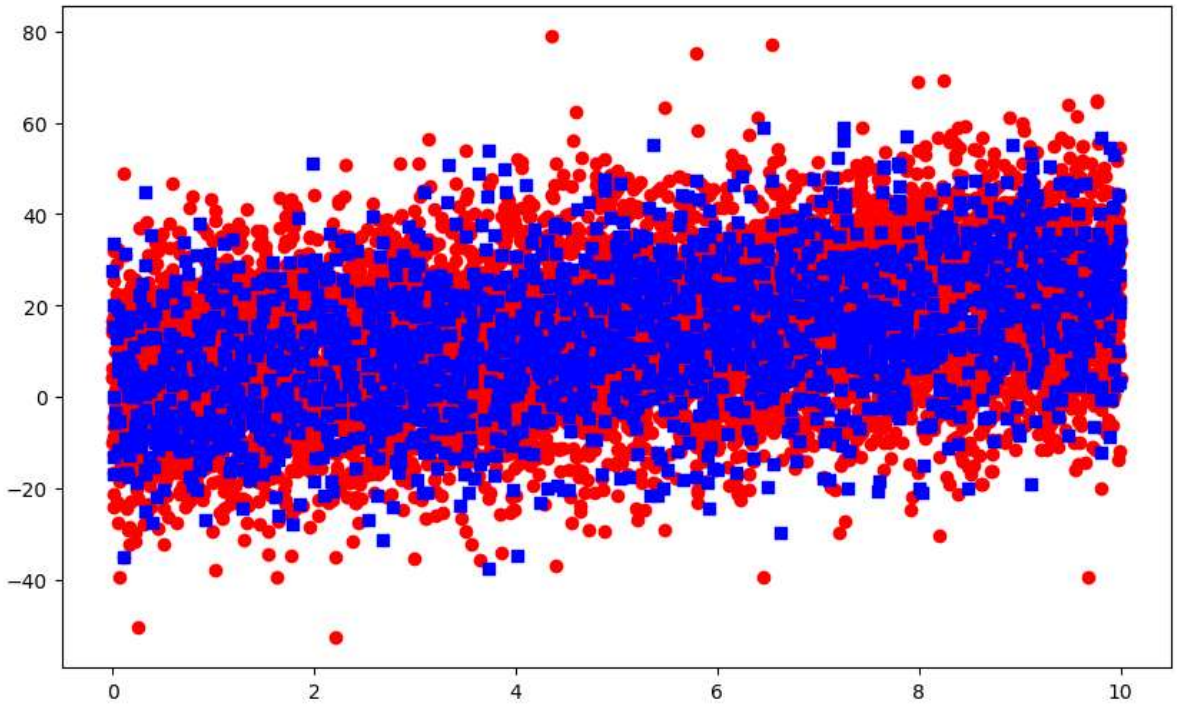
- **Random Seed (r):** The random_state=r ensures that the split is reproducible for the same value of r. However, since r is a random integer generated by np.random.randint(104) with a range of possible values (0 to 103), the split changes in each run.
- **Random Splitting:** The random seed affects how the dataset is divided. Since the seed varies, the exact samples that end up in the training and testing sets will also vary.

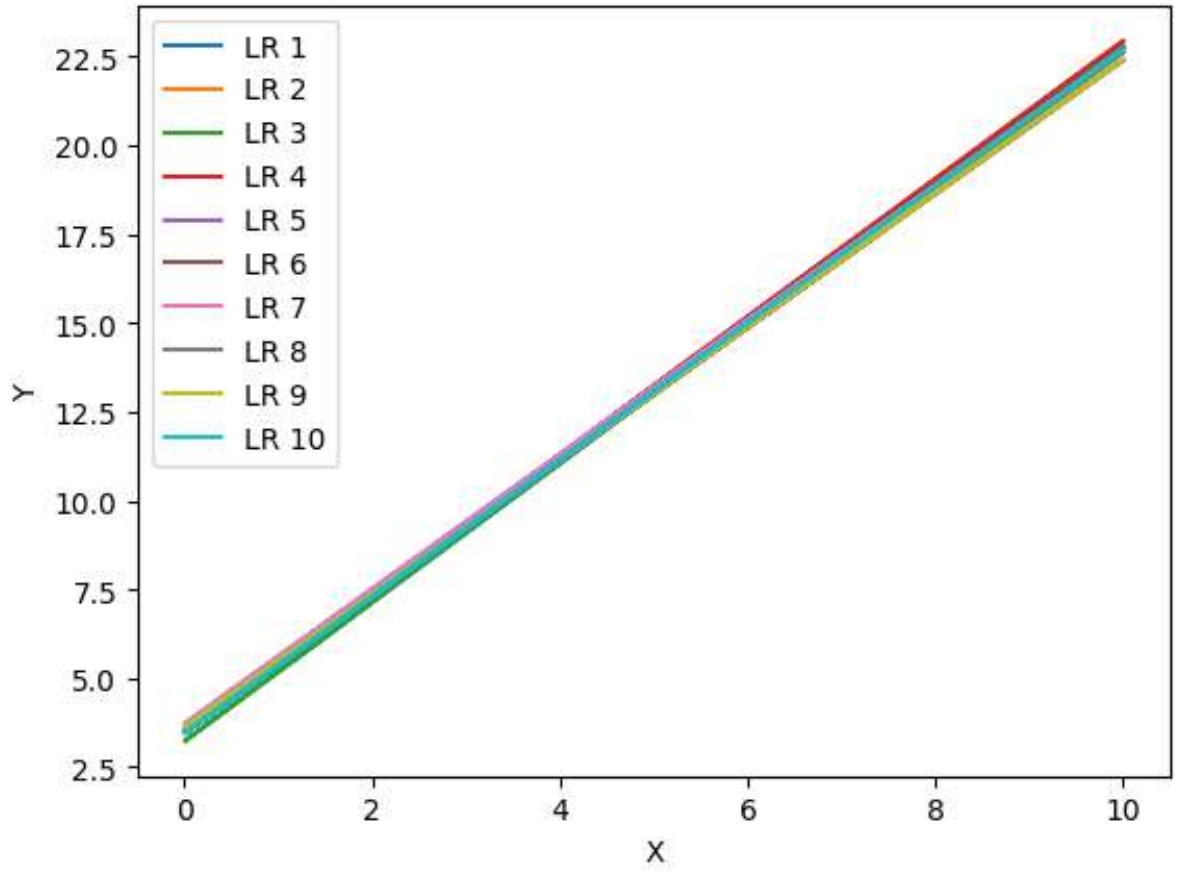
3.Why is the Linear Regression Model Different in Each Instance?



- **Random Data Splitting:**
  - Each iteration of the loop uses a different training dataset due to the random split
  - Since the model is fitted on different subsets of the data, the resulting linear regression model (the line's slope and intercept) varies with each run.
- **Data Sensitivity:**
  - Training Data Influence**:** Linear regression finds the line of best fit based on the training data. Different training datasets lead to different "best fit" lines.
  - Variability in Predictions: Because the training data is different, the predictions (Y_pred_train) and thus the regression line for each instance differ.
- **Effect of Noise:**
  - The data generation includes a noise component (epsilon), which causes variability in the Y values.
  - The noise impacts how well the model fits the data in each instance, contributing to differences in the regression line.

4.Observations with 10000 samples





Observations

Variability in Regression Lines:

- **With 100 Samples:**
  - The regression lines varied significantly across different instances. This was due to the small sample size, which made each random split quite different from the others
- **With 10,000 Samples:**
  - The regression lines become much more consistent and overlap significantly across different instances. With a larger dataset, the impact of random splitting diminishes because each subset (training or testing) is more representative of the overall data distribution.

Model Stability:

- **With 100 Samples:**
  - The model was sensitive to small changes in the training data, leading to different regression lines and indicating instability due to a lack of sufficient data.
- **With 10,000 Samples:**
  - The model becomes much more stable. Larger datasets provide more information to the model, making it less susceptible to the randomness in data splitting. The model's intercept and coefficient will have less variance, leading to more stable and reliable predictions.

Reason for Different Behaviour compared to 100 data samples:

Larger Sample Size Reduces Variance:

As the sample size increases, the sample mean and other statistics (like regression coefficients) tend to converge to their true population values. This reduces the variance between different model instances. With 10,000 samples, each training set (even when split randomly) contains enough data to accurately reflect the underlying relationship between X and Y. This consistency results in similar models across different instances.

3 Linear Regression on Real World Data

2. How many independent variables and dependent variables are there in the data set?

Number of independent variables: 33
Number of dependent variables: 2

3. Is it Possible to Apply Linear Regression on this Dataset?

Since some independent features (age) are categorical, we cannot directly apply linear regression to the dataset. To make the dataset suitable for linear regression, we need to convert the categorical features into numerical format. This can be achieved using one-hot encoding.

4.Issues with the Provided Code

It is an incorrect approach. The line X = X.dropna() only removes rows in X that contain NaN values, but it does not remove rows where y contains NaN values. Similarly, the line y = y.dropna() only removes rows in y that contain NaN values, but does not account for NaN values in X. Therefore, we must first concatenate X and y, then remove any rows containing NaN values. Finally, we can split the dataset into the new X and y.

5. Dependent feature and independent features

Dependent feature - aveOralM

Independent features - Age , T_atm , Humidity , Distance , T_RC_Max1

7. Estimated coefficients

	Estimated Coefficients
Age_18-20	-0.070356
Age_21-25	-0.026691
Age_21-30	0.016431
Age_26-30	-0.076694
Age_31-40	-0.138249
Age_41-50	0.002548
Age_51-60	-0.085890
Age_>60	0.378902
T_atm	-0.059292
Humidity	0.001348
Distance	-0.082847
T_LC1	0.751796

8. Which independent variable contributes highly for the dependent feature?

T_LC1

9. Estimate the coefficient corresponds to independent variables.

T_OR1	0.50338995
T_OR_Max1	0.02169014
T_FHC_Max1	-0.06022472
T_FH_Max1	0.35936308

10.Calculations

RSS	18.80409309153579
RSE	0.3073970235838663
MSE	0.09217692691929309
R^2	0.6495141698846425

Standard Errors for Each Feature

T_OR1	1.831723
T_OR_Max1	1.827018
T_FHC_Max1	0.089940
T_FH_Max1	0.094565

t-statistics for Each Feature

T_OR1	-0.371945
T_OR_Max1	0.716412
T_FHC_Max1	-0.728860
T_FH_Max1	2.996086

p-values for Each Feature:

T_OR1	0.710330
T_OR_Max1	0.474577
T_FHC_Max1	0.466944
T_FH_Max1	0.003082

11 Will you be able to discard any features based on p-value?

P-values indicate the significance of each feature in relation to the target variable. Typically, a feature with a p-value below a certain threshold (commonly 0.05) is considered statistically significant. It's likely to have a meaningful relationship with the dependent variable.

T_OR1 (p-value: 0.710330)
T_OR_Max1 (p-value: 0.474577)
T_FHC_Max1 (p-value: 0.466944)

These features have high p-values, implying they are not statistically significant. We can discard them.

4 Performance Evaluation of Linear Regression

Consider the linear regression models Model A: $y = w_0 + w_1x_1 + w_2x_2$ and Model B:$y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4$. Sum of squared errors (SSE) and total sum of squares (TSS) of these models are given in Table 1.

Table 1: SSE and TSS of linear regression models.

	Model A	Model B
SSE= $\sum_{i=1}^{N}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2$	9	2
TSS= $\sum_{i=1}^{N}(y_i - \bar{y}_i)^2$	90	10
Number of data samples (N)	10000	10000

2.Compute residual standard error (RSE) for models A and B. Based on RSE for which model performs better?

② $RSE = \sqrt{\frac{SSE}{N-p}}$

model A

$SSE = 9 \qquad N=10\,000 \qquad p=3$

$RSE_A = \sqrt{\frac{9}{10000-3}} = 0.030$

model B

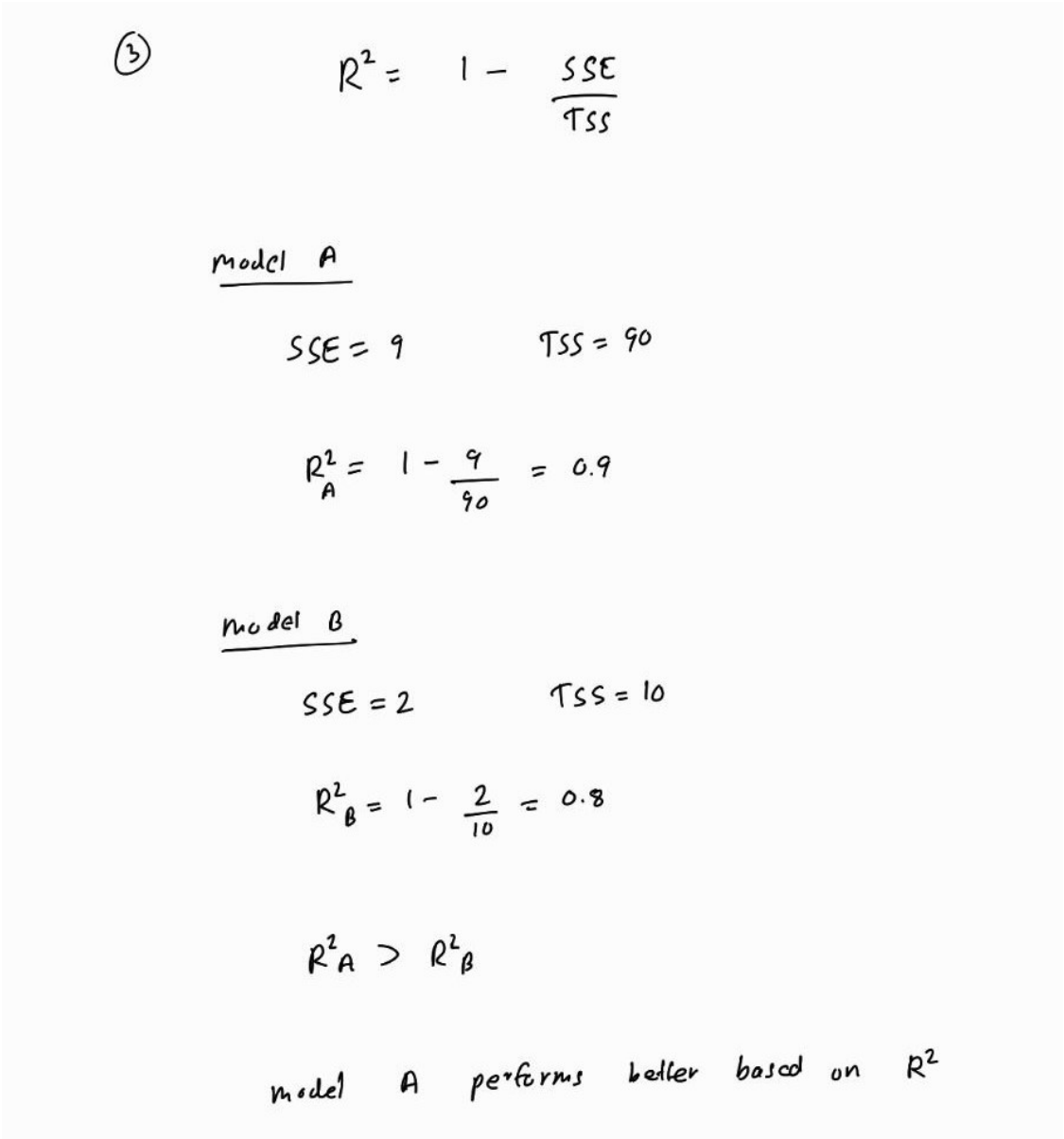
$SSE = 2 \qquad N=10\,000 \qquad p=5$

$RSE_B = \sqrt{\frac{2}{10000-5}} = 0.014$

$RSE_B < RSE_A$

model B performs better based on RSE.

3.Compute R-squared (R2) for models A and B. Based on R2 for which model performs better?



4. Between RSE and R-squared, which performance metric is more fair for comparing two models and why?

RSE is generally more reliable for comparing models when there is a different number of predictors, as it accounts for the number of predictors and penalizes models with more predictors unless they significantly reduce the error.

In contrast, R^2 doesn't penalize additional predictors unless the improvement in SSE is not substantial. Therefore, RSE is often more fair when comparing models with different numbers of predictors because it balances the goodness of fit against model complexity.

RSE is more fair for comparing the two models, especially because Model B has more predictors than Model A. The higher R^2 for Model A doesn't necessarily mean it's a better model since it might be overfitting compared to Model B.

5 Linear Regression Impact on Outliers

2. What happens when a → 0 ?

$a \to 0$?

$$L_1(w) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{r_i^2}{a^2 + r_i^2} \right)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left(\frac{r_i^2}{0 + r_i^2} \right)$$

$$= \frac{1}{N} \sum_{i=1}^{N} (1) \qquad r_i \neq 0$$

$$= \frac{1}{N} \times N = 1$$

$$L_2(w) = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \exp\left(-\frac{2|r_i|}{a} \right) \right)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left(1 - \exp(-\infty) \right)$$

$$= \frac{1}{N} \sum_{i=1}^{N} (1 - 0)$$

$$= \frac{1}{N} \sum_{i=1}^{N} (1) \qquad r_i \neq 0$$

$$= \frac{1}{N} \times N = 1$$

As a approaches 0, for non-zero residuals, both L1(w) and L2(w) become constant functions with values close to 1.

This implies that the loss functions lose sensitivity to the actual residual values and loss functions become less sensitive to the magnitude of residuals and focus instead on whether residuals are zero or non-zero.

3. Suppose we need to minimize the influence of data points with $|r_i| \geq 40$. What value(s) of "a" and what function(s) would you choose, and why?

To minimize the influence of outliers:

- **For L1(w):** Observing the graph, L1(w) shows less sensitivity to large residuals as a increases. For a=100, L1(w) flattens out significantly for large r, reducing the penalty for outliers.
- **For L2(w):** L2(w) also decreases the influence of large residuals as a increases. For a=100, L2(w) values are relatively small, effectively down-weighting the contribution of outliers to the overall loss.

To minimize the influence of data points with $|r_i| \geq 40$, we can choose a larger value of a, such as a=100. Both L1(w) and L2(w) functions can be effective, but L2(w) tends to penalize large residuals less than L1(w) making L2(w) with a=100 a potentially better choice for minimizing the impact of outliers.