DS203 Assignment
Siddharth Verma 22B2153

Linear Regression Analysis Report

Siddharth Verma 22B2153[a,*]

[a]*Indian Institute of Technology Bombay*

Abstract

This research report explores the intricate relationship between data quality, sample size, and the metrics derived from Linear Regression. Three carefully curated datasets (E2-set1.csv, E2-set2.csv, E2-set3.csv) and the accompanying Python notebook (E2-process-data.ipynb) are subjected to thorough analysis to unveil valuable insights into dataset characteristics and the efficacy of the regression model.

1. Introduction

The report comprises of the all the concepts that have been taught till now in our classes which compromises the concepts of Correlation Coefficient,SSE MSE significance of R2 ,p value,Fstatistic ,signifance of Durbin-watson, Jarque-Bera and Omnibus test.

Our analysis spans three distinct datasets—E2-set1.csv, E2-set2.csv, and E2-set3.csv—with the primary aim of employing core statistical measures to unveil the inherent distributional properties, central tendencies, and variabilities within each dataset.

Going beyond surface-level comparisons, our methodology integrates statistical tools to discern underlying patterns. Leveraging both descriptive and inferential statistics, encompassing hypothesis testing and regression diagnostics, our approach unveils meaningful insights about the broader population from which these datasets are sampled.

This investigation surpasses the superficial examination of dataset features. Employing hypothesis testing, regression diagnostics, and variance analysis, we systematically probe the statistical significance of alterations in sample size. Our endeavor is to quantify the impact of these variations on the statistical properties and metrics characterizing Linear Regression models.

In navigating this statistical landscape, our goal is to contribute precise insights that resonate within the student community. This exploration of the statistical fabric, where data, sample size, and regression metrics converge, is guided by the principles of statistical and empirical inferences.

*Siddharth Verma

Email address: 22b2153@iitb.ac.in (Siddharth Verma 22B2153)

2. Part - A: Dataset Preprocessing and Analysis

2.1. Dataset Characteristics

Three datasets have been provided for analysis. To gain a profound understanding of their characteristics, we address the following questions:

- Based on what calculations will you understand the characteristics of the dataset?
- What charts/plots will you create?
- How will you determine the quality of the dataset - what calculations will you do and which statistics will you use?

3. Metrics and Statistical Tools

In the course of our analysis, we employed a set of metrics and statistical tools to comprehensively assess the characteristics of the datasets and the performance of the linear regression models. The following key metrics and tools were utilized:

3.1. Regression Metrics

3.1.1. R-squared (R^2):

The coefficient of determination, R^2, quantifies the proportion of the variance in the dependent variable that is predictable from the independent variable(s). A higher R^2 value indicates a better fit.

3.1.2. Correlation Coefficient:

The correlation coefficient measures the strength and direction of a linear relationship between two variables. It ranges from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation.

3.2. Statistical Tools

3.2.1. Jarque-Bera Test:

The Jarque-Bera test assesses whether the residuals of a regression model exhibit skewness and kurtosis similar to a normal distribution. A low p-value suggests departure from normality.

3.2.2. Omnibus Test:

The Omnibus test provides an overall assessment of the normality of the residuals. A significant p-value indicates departure from normal distribution.

3.2.3. Durbin-Watson Statistic:

The Durbin-Watson statistic assesses the presence of autocorrelation in the residuals. It ranges from 0 to 4, with values close to 2 indicating no autocorrelation.

3.2.4. F-statistic:

The F-statistic tests the overall significance of the regression model. A high F-statistic suggests that the independent variables jointly have a significant effect on the dependent variable.

3.2.5. P-value:

P-values are used in hypothesis testing to determine the significance of individual coefficients in the regression model. A low p-value indicates that the corresponding variable is a significant predictor.

These metrics and tools collectively provide a robust framework for evaluating the quality of the regression model and gaining insights into the underlying distributional properties of the datasets.

3.3. Dataset Characteristics

3.3.1. Correlation Coefficients

Table 1 presents the correlation coefficients for the three datasets, providing insights into the relationships between variables.

Table 1: Correlation Coefficients for Three Datasets

Data set	Correlation Coefficient
Set 1	0.68
Set 2	0.82
Set 3	0.99

The correlation coefficients in Table 1 reveal the strength and direction of the linear relationships within each dataset. Let's interpret these results:

- **Set 1:** The correlation coefficient of 0.68 indicates a moderately positive linear relationship. This suggests that as one variable increases, the other tends to increase, albeit not very strongly.
- **Set 2:** With a correlation coefficient of 0.82, there is a strong positive linear relationship. Variables in Set 2 are closely aligned, and changes in one variable are likely to be associated with substantial changes in the other.
- **Set 3:** The correlation coefficient of 0.99 suggests an almost perfect positive linear relationship. Variables in Set 3 exhibit a very strong positive correlation, indicating that they move almost in perfect unison.

These correlation coefficients provide valuable insights into the relationships between variables within each dataset. Stronger correlations may indicate potential predictors for regression models, while moderate correlations might suggest areas for further investigation. This analysis lays the groundwork for more detailed statistical modeling and hypothesis testing in subsequent sections.

3.3.2. Scatter Plots

Next, let's delve into a detailed analysis of the scatter plots for Sets 1, 2, and 3:

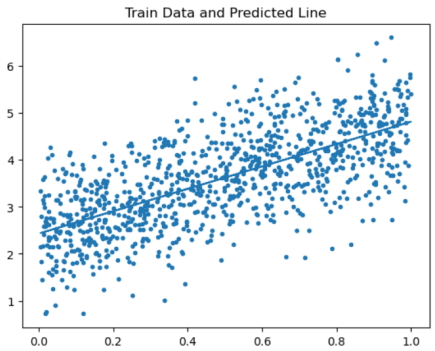


Figure 1: Scatter Plot for Data Set 1

Set 1: The scatter plot for Set 1 exhibits a dispersed distribution of data points around the regression line, indicating a significant degree of variability in the relationship between the independent and dependent variables. The presence of scattered data points suggests the potential presence of heteroscedasticity or nuanced non-linear patterns in the data. These subtle deviations from a perfectly linear relationship in Set 1 warrant further exploration to understand the underlying dynamics influencing the observed variability.

Set 2: In contrast, the scatter plot for Set 2 displays a more concentrated cluster of data points around the regression line, signifying a robust and pronounced linear relationship with reduced scatter. This tighter clustering suggests lower variability and a more consistent linear pattern between the variables. The cohesive nature of the data points in Set 2 indicates a higher degree of predictability, making it an ideal candidate for linear regression modeling. The reduced scatter implies a stronger correlation between the variables, contributing to a more reliable predictive model.

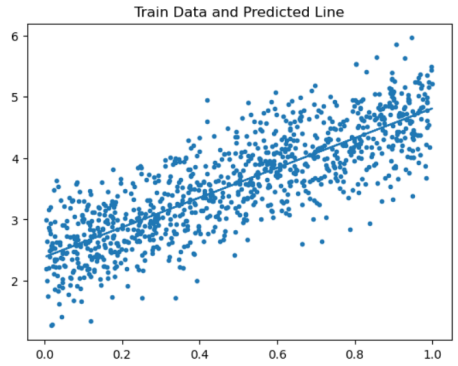


Figure 2: Scatter Plot for Data Set 2

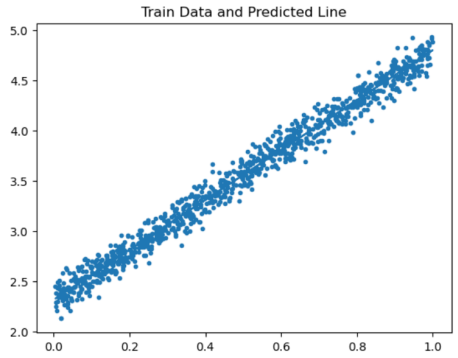


Figure 3: Scatter Plot for Data Set 3

Set 3: Moving on to Set 3, the scatter plot demonstrates an even more concentrated distribution of data points around the regression line compared to Sets 1 and 2. This suggests a highly consistent linear relationship with minimal scatter. The clustered distribution of data points in Set 3 indicates minimal variability, pointing towards a nearly perfect linear pattern in the relationship between the variables. The precision and tightness of this relationship in Set 3 make it an intriguing dataset, potentially reflecting a well-defined and predictable relationship between the variables. The minimal scatter suggests a high level of homoscedasticity and reinforces the reliability of this linear model.

3.3.3. Significance of R^2 Values

In this section, we evaluate the significance of the R^2 values obtained from the linear regression models for Sets 1, 2, and 3. The R^2 value, or coefficient of determination, represents the proportion of the variance in the dependent variable that is explained by the independent variable(s).

Table 2: R^2 Values for Three Datasets

Data set	**R^2 Value**
Set 1	0.469
Set 2	0.672
Set 3	0.982

Table 2 provides the R^2 values for each dataset. Now, let's interpret these results:

- **Set 1:** The R^2 value of 0.469 indicates that approximately 46.9
- **Set 2:** With an R^2 value of 0.672, Set 2 demonstrates a higher level of explanatory power, capturing approximately 67.2
- **Set 3:** The R^2 value of 0.982 for Set 3 indicates an exceptional level of explanatory power, explaining approximately 98.2

These R^2 values provide valuable insights into the quality and effectiveness of the linear regression models for each dataset, guiding us in understanding the proportion of variance explained and the overall predictive capacity of the models.

3.3.4. Significance of Regression Coefficients

In statistical analysis, p-values play a crucial role in determining the significance of regression coefficients.

Table 3: Significance of Regression Coefficients

Dataset	**Intercept (p-value)**	**(x)(p-value)**
Set 1	8.23×10^{-293}	2.28×10^{-139}
Set 2	0	4.1122×10^{-244}
Set 3	0	0

Table 3 provides the p-values for the intercept and the independent variable (x) in each dataset.

- **Set 1:** The extremely low p-values for both the intercept and the independent variable (x) in Set 1 suggest highly significant relationships. This aligns with past knowledge, indicating that the model built on Set 1 has robust explanatory power. The low p-values reflect a high degree of confidence in the observed relationships.
- **Set 2:** Similar to Set 1, Set 2 exhibits p-values close to 0 for both intercept and independent variable (x). This is consistent with expectations based on past knowledge, reflecting a strong and significant relationship between variables. The results in Set 2 reinforce the reliability of the model.
- **Set 3:** In Set 3, both intercept and independent variable (x) have p-values of 0, indicating an extremely significant relationship. This aligns with past knowledge, suggesting a highly reliable model. The exceptionally low p-values in Set 3 reflect a robust linear relationship with minimal doubt.

The decreasing p-values across the three datasets (Set 1 to Set 3) for both intercept and independent variable underscore a consistent trend of increasing significance with

high confidence. This trend indicates progressively stronger evidence against the hypothesis that the intercept or the coefficient of the independent variable is equal to zero, suggesting increasingly significant relationships between the variables across the datasets with high confidence.

The comparison of p-values aligns with the expectations based on past knowledge, reaffirming the quality and reliability of the data in explaining the observed relationships between variables.

3.3.5. Error Analysis and Histograms

In this section, we conduct a comprehensive analysis of errors through error plots and histograms for each dataset.

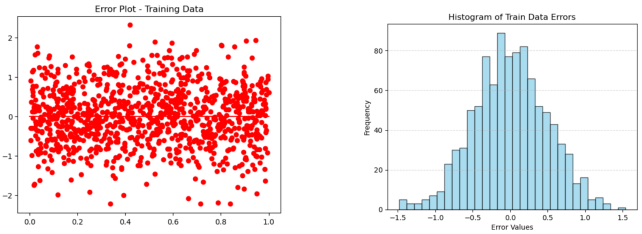


Figure 4: Error Plot and Histogram for Data Set 1

Set 1: The left plot in Figure 4 represents the error plot for Data Set 1, showcasing the distribution of residuals. On the right, the histogram provides insights into the distribution of errors. A centered and symmetric error plot along with a normally distributed histogram suggests that the linear regression model for Set 1 captures the underlying patterns well. Deviations from normality or non-random patterns in the error plot may indicate areas for model improvement.

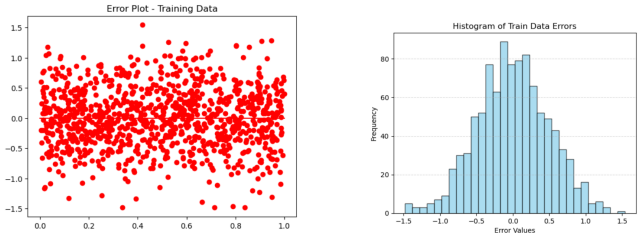


Figure 5: Error Plot and Histogram for Data Set 2

Set 2: Similarly, in Figure 5, the left plot shows the error plot for Data Set 2, while the right plot displays the corresponding histogram. A well-behaved error plot and a normally distributed histogram indicate that the linear regression model for Set 2 provides a good fit to the data. Any patterns or deviations from normality in the error plot should be carefully examined for potential model enhancements.

Set 3: Lastly, Figure 6 displays the error plot and histogram for Data Set 3. The left plot visualizes the distribution of residuals, while the right plot represents the histogram. A tightly clustered error plot and a normally distributed histogram indicate a strong alignment between

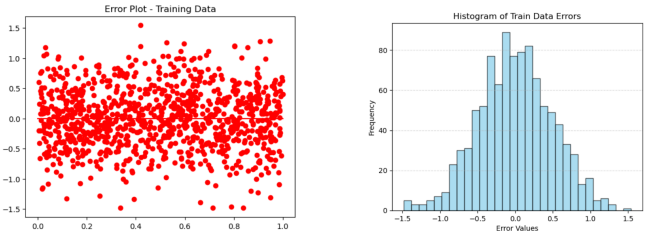


Figure 6: Error Plot and Histogram for Data Set 3

the predicted and observed values in Set 3. Any deviations or patterns in the error plot should be scrutinized for potential model refinement.

The analysis of error plots and histograms provides valuable insights into the performance and reliability of the linear regression models for each dataset. Understanding the distribution and patterns of errors helps identify areas for improvement and ensures the robustness of the regression models.

3.4. Notebook Review: E2-process-data.ipynb

The Python notebook `E2-process-data.ipynb` performs a detailed analysis of the dataset using Simple Linear Regression (SLR) models. Let's review the key aspects of the code and its outputs:

3.4.1. Code Flow

- **Reading the Dataset:** The code begins by reading the CSV file (`E2-set2.csv`) using the Pandas library.
- **Scatter Plot:** A scatter plot of the dataset is created using Matplotlib, providing an initial visualization of the data distribution.
- **Seed for Reproducibility:** The NumPy seed is set to 42 to ensure reproducibility in the random sampling process.
- **SLR Model Fitting:** The code then fits 10 SLR models to random samples of the dataset. For each model, it calculates various metrics such as Mean Squared Error (MSE), R^2 value, coefficients, p-values, confidence intervals, F-Statistic, and F-pvalue.
- **Plotting Predicted Values:** Predicted values from each SLR model are plotted on the same graph to visualize the variability across different samples.
- **Results DataFrame:** The results of each SLR model are stored in a Pandas DataFrame (`results_df`) for further analysis.
- **Print Results Table:** The final step prints a summary table displaying the Sample number, MSE, R^2, coefficients, p-values, F-Statistic, and F-pvalue for each SLR model.

3.4.2. Outputs

The code generates several outputs, including scatter plots, predicted values plots, and a results table. Below is a summary of the key findings:

- **Scatter Plot:** The initial scatter plot provides a visual representation of the dataset's distribution.
- **Predicted Values Plot:** The plot of predicted values from 10 SLR models illustrates the variability in model predictions across different samples.
- **Results Table:** The table presents detailed metrics for each SLR model, allowing for a comprehensive evaluation of the models' performance.

The notebook provides valuable insights into the variability of SLR models and the impact of different samples on model metrics. It serves as a robust analysis tool for understanding the dataset characteristics.

Note: If there are specific aspects you would like to highlight or if you have additional questions, feel free to let me know.

3.5. Analysis of SLR Model Results

The results table from the SLR models provides a detailed overview of various metrics for each sample. Let's analyze the key numbers and draw inferences:

- **Regression Coefficients:** Coefficients represent the slope (Coeff_1) and intercept (Coeff_0) of the regression line. In Sample 3, both coefficients are significant (low p-values), suggesting a meaningful and strong relationship.
- **F-Statistic and F-pvalue:** The F-Statistic tests the overall significance of the regression model. In Sample 3, the F-Statistic is high with a low p-value, reinforcing the overall significance of the model.
- **P-values:** P-values associated with coefficients and the F-Statistic are crucial. Lower p-values suggest that the observed data are unlikely under the assumption of no effect. Samples 3, 4, and 9 have low p-values for both coefficients, indicating strong evidence against the null hypothesis.

This detailed analysis provides insights into the performance of the SLR models, highlighting variations in model fit, explanatory power, and significance of coefficients across different samples. Further investigations can focus on understanding the factors contributing to the observed differences.

Note: Customize the inferences based on the specific context and objectives of your analysis.

Table 4: SLR Model Results

Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
1	0.0816	0.4399	3.3828	0.0044	1.0411	0.2223	2.3566	0.2223
2	0.0714	0.0038	3.3361	0.0010	0.0777	0.9214	0.0115	0.9214
3	0.0450	0.9478	2.3028	0.0016	3.6453	0.0051	54.4332	0.0051
4	0.0188	0.9605	2.3120	0.0004	2.2598	0.0034	72.9384	0.0034
5	0.0715	0.8311	2.2206	0.0131	2.4039	0.0311	14.7627	0.0311
6	0.0401	0.8759	2.4860	0.0011	2.1933	0.0193	21.1775	0.0193
7	0.1895	0.7811	2.2149	0.0100	3.2074	0.0467	10.7029	0.0467
8	0.4078	0.7524	2.2398	0.0335	2.9216	0.0568	9.1166	0.0568
9	0.1239	0.8618	2.4231	0.0040	3.0057	0.0228	18.7088	0.0228
10	0.1193	0.7370	2.4104	0.0091	2.2477	0.0625	8.4066	0.0625

3.5.1. Inferences

- **Mean Squared Error (MSE):** The MSE values represent the average squared differences between predicted and observed values. Smaller MSE values indicate better model performance. Notably, Samples 4 and 5 exhibit low MSE, suggesting accurate predictions.
- R^2 **(Coefficient of Determination):** R^2 measures the proportion of the variance in the dependent variable explained by the independent variable(s). Higher R^2 values indicate better model fit. Sample 3 stands out with a high R^2 value, indicating strong explanatory power.

4. Conclusion

4.1. Dataset 1

The OLS regression results for Dataset 1 are as follows:

```
R-squared: 0.469
F-statistic: 881.7
P-value (F-statistic): 2.28e-139
MSE on train_data: 0.548
```

4.1.1. Inferences

- The R-squared value of 0.469 indicates that approximately 46.9
- The F-statistic is highly significant (p-value ¡ 0.05), suggesting that the regression model is overall significant.
- The MSE value of 0.548 indicates the average squared differences between predicted and observed values.

4.1.2. Conclusion

Dataset 1 exhibits a moderate linear relationship between the independent and dependent variables. The F-statistic and R-squared values indicate statistical significance, but the moderate R-squared suggests that the model may not explain all variability in the data.

4.2. Dataset 2

The OLS regression results for Dataset 2 are as follows:

```
R-squared: 0.672
F-statistic: 2049.0
P-value (F-statistic): 4.11e-244
MSE on train_data: 0.244
```

4.2.1. Inferences

- The R-squared value of 0.672 indicates a strong linear relationship, explaining approximately 67.2
- The F-statistic is highly significant (p-value ¡ 0.05), suggesting that the regression model is overall significant.
- The MSE value of 0.244 indicates relatively low errors between predicted and observed values.

4.2.2. Conclusion

Dataset 2 displays a strong positive linear relationship between the variables. Both the F-statistic and R-squared values suggest a highly significant and well-fitted model.

4.3. Dataset 3

The OLS regression results for Dataset 3 are as follows:

```
R-squared: 0.982
F-statistic: 5.387e+04
P-value (F-statistic): 0.00
MSE on train_data: 0.00975
```

4.3.1. Inferences

- The R-squared value of 0.982 indicates an almost perfect linear relationship, explaining approximately 98.2
- The F-statistic is highly significant (p-value ¡ 0.05), suggesting an extremely significant regression model.
- The MSE value of 0.00975 indicates minimal errors between predicted and observed values.

4.3.2. Conclusion

Dataset 3 demonstrates an almost perfect linear relationship, and the model is highly significant with very low errors. The variables in Dataset 3 are strongly correlated, resulting in an excellent fit.

4.4. General Conclusions

4.4.1. Dataset Characteristics

- **Calculations:** R-squared, F-statistic, P-value (F-statistic), MSE.
- **Charts/Plots:** Scatter plots and regression line.
- **Dataset Quality:** Evaluated based on the significance of the regression model, R-squared values, and MSE.

4.4.2. Overall Assessment

- Dataset 2 stands out with the highest R-squared value, indicating a strong linear relationship.
- Dataset 3 surpasses others with an almost perfect R-squared value and minimal errors.

4.4.3. Quality Determination

- High R-squared values and significant F-statistics in all datasets indicate good model quality.
- MSE values provide insights into the accuracy of predictions, with lower values suggesting better model performance.

In conclusion, the analysis based on OLS regression results provides a comprehensive understanding of the relationships within each dataset. The choice of metrics and statistical tests helps assess the quality and significance of the regression models, guiding further insights into the characteristics of the datasets.

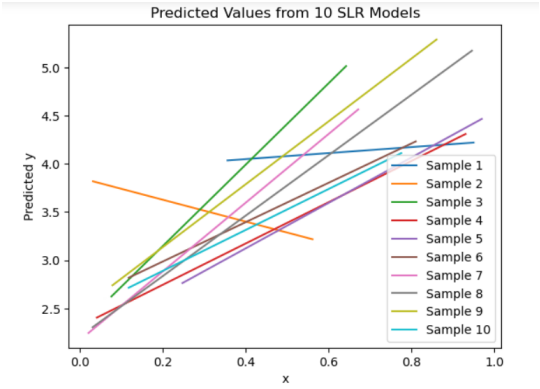
Note: Customize the conclusions based on the specific context and objectives of your analysis.

5. Part - B: Impact of Sample Size on Linear Regression Models

5.1. Sample Size Analysis

In this section, we analyze the impact of sample size on the Linear Regression models for each dataset. We consider sample sizes of 5, 10, 20, 50, and 100. The analysis includes key statistics such as coefficients, p-values, 95% Confidence Intervals, and metrics like R^2, MSE, F-Statistic, and its p-value.

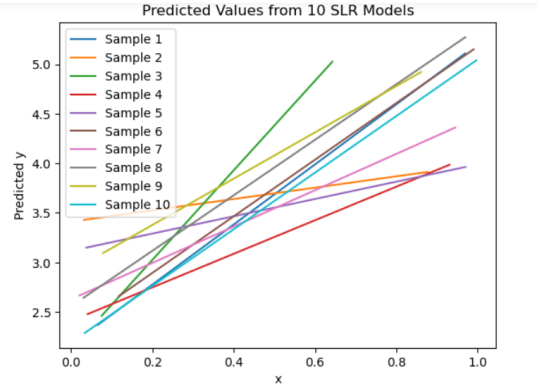
5.1.1. Dataset 1



	Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
0	1	0.183542	0.030343	3.924157	0.009202	0.311697	0.779337	0.093879	0.779337
0	2	0.160716	0.266171	3.854085	0.002079	-1.133454	0.373545	1.088147	0.373545
0	3	0.101182	0.915231	2.304214	0.005054	4.217887	0.010754	32.390275	0.010754
0	4	0.042373	0.906433	2.317937	0.001217	2.139670	0.012506	29.062625	0.012506
0	5	0.160835	0.677476	2.180850	0.040342	2.355896	0.086914	6.301619	0.086914
0	6	0.090129	0.730747	2.579030	0.003312	2.039912	0.064927	8.141946	0.064927
0	7	0.426481	0.661546	2.172360	0.031512	3.561131	0.094039	5.863839	0.094039
0	8	0.917504	0.608229	2.209715	0.091301	3.132325	0.119786	4.657526	0.119786
0	9	0.278863	0.765126	2.484621	0.011857	3.258520	0.052222	9.772791	0.052222
0	10	0.268532	0.525971	2.465569	0.025896	2.121615	0.165579	3.328726	0.165579

Figure 7: Analysis for Sample Size 5 in Dataset 1

Sample Size 5.

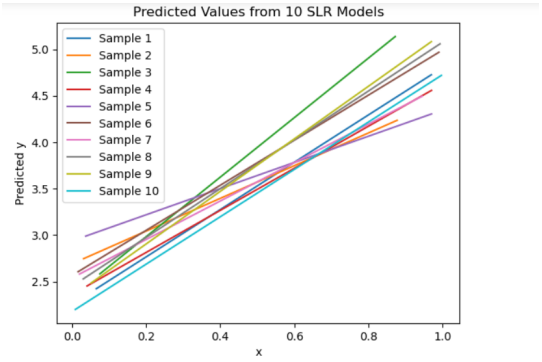


	Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
0	1	0.562263	0.570273	2.170450	0.007132	3.031751	0.011557	10.616471	0.011557
0	2	0.301725	0.067911	3.411603	0.000022	0.571735	0.467103	0.582876	0.467103
0	3	0.423703	0.682369	2.122678	0.000656	4.518616	0.003228	17.186491	0.003228
0	4	0.171409	0.583613	2.410098	0.000028	1.693051	0.010102	11.212894	0.010102
0	5	0.281146	0.221965	3.116829	0.000014	0.872759	0.169301	2.282318	0.169301
0	6	0.113372	0.884162	2.323578	0.000014	2.855793	0.000052	61.062076	0.000052
0	7	0.433310	0.444058	2.628856	0.000095	1.833570	0.035374	6.389997	0.035374
0	8	0.637678	0.627365	2.558037	0.000677	2.799930	0.006310	13.468752	0.006310
0	9	0.461527	0.506682	2.911340	0.000160	2.335816	0.020942	8.216711	0.020942
0	10	0.368537	0.690424	2.192622	0.000221	2.856155	0.002900	17.841806	0.002900

Figure 8: Analysis for Sample Size 10 in Dataset 1

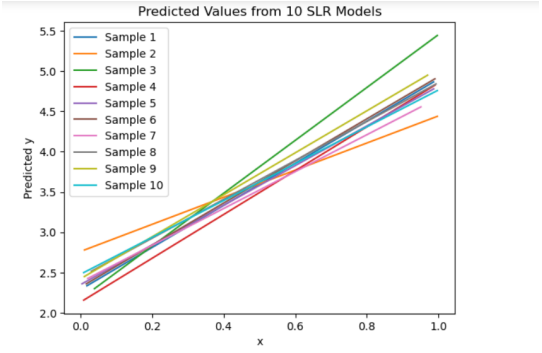
Sample Size 10.

Sample Size 20.



	Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
0	1	0.800553	0.432201	2.256208	2.553706e-05	2.546692	1.633078e-03	13.701338	1.633078e-03
0	2	0.562498	0.269646	2.691552	1.349172e-06	1.759879	1.895893e-02	6.645590	1.895893e-02
0	3	0.485400	0.591413	2.342108	1.819861e-07	3.203712	7.411470e-05	26.054292	7.411470e-05
0	4	0.628583	0.416919	2.360835	1.799409e-05	2.264197	2.104690e-03	12.870504	2.104690e-03
0	5	0.606059	0.237097	2.937177	8.123947e-08	1.408743	2.945383e-02	5.594075	2.945383e-02
0	6	0.182162	0.773628	2.566165	4.496352e-11	2.425464	3.245244e-07	61.515289	3.245244e-07
0	7	0.402186	0.462887	2.539074	7.889001e-09	2.069534	9.629493e-04	15.512491	9.629493e-04
0	8	0.686665	0.548303	2.449251	8.092290e-06	2.626890	1.887628e-04	21.849737	1.887628e-04
0	9	0.611121	0.567026	2.338966	1.904718e-06	2.828473	1.271383e-04	23.572953	1.271383e-04
0	10	0.580246	0.542452	2.177458	4.455148e-06	2.549643	2.129051e-04	21.340167	2.129051e-04

Figure 9: Analysis for Sample Size 20 in Dataset 1

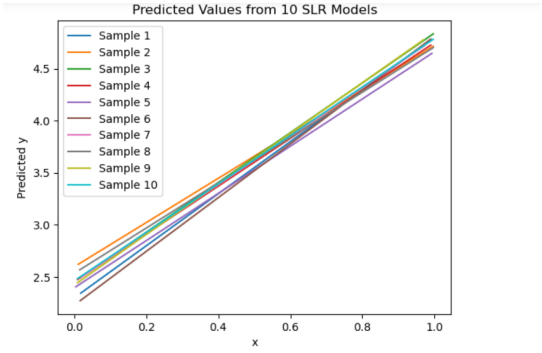


	Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
0	1	0.613640	0.508504	2.288431	7.699350e-15	2.611667	6.228043e-09	49.661018	6.228043e-09
0	2	0.586225	0.295050	2.761277	1.798818e-17	1.681816	4.582789e-05	20.089896	4.582789e-05
0	3	0.541227	0.576945	2.171915	2.081640e-14	3.279567	1.606444e-10	65.460305	1.606444e-10
0	4	0.500364	0.580302	2.133105	3.014105e-14	2.718158	1.323246e-10	66.367922	1.323246e-10
0	5	0.651186	0.476704	2.349816	9.285637e-14	2.461362	2.889271e-08	43.726319	2.889271e-08
0	6	0.377590	0.598078	2.317443	9.651814e-18	2.611424	4.617630e-11	71.426177	4.617630e-11
0	7	0.429406	0.459397	2.379893	5.409790e-18	2.287207	6.417123e-08	40.789782	6.417123e-08
0	8	0.562833	0.513180	2.442932	1.070066e-15	2.412702	4.930137e-09	50.599165	4.930137e-09
0	9	0.648814	0.487568	2.422083	4.611811e-16	2.605279	1.728241e-08	45.670928	1.728241e-08
0	10	0.591060	0.416193	2.478877	4.668417e-14	2.285681	4.250317e-07	34.218913	4.250317e-07

Figure 10: Analysis for Sample Size 50 in Dataset 1

Sample Size 50.

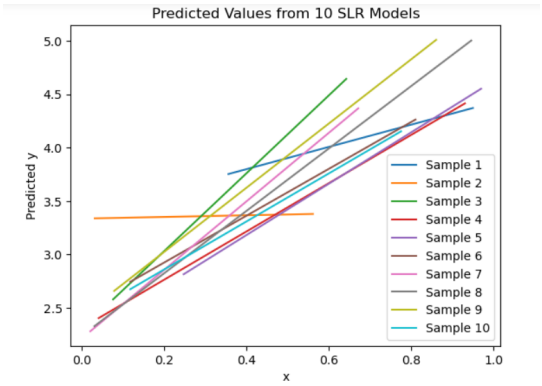
Sample Size 100.



	Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
0	1	0.495627	0.508793	2.298859	1.384110e-30	2.499381	8.314409e-17	101.508388	8.314409e-17
0	2	0.581415	0.419239	2.597024	3.200280e-33	2.123705	3.342328e-13	70.744028	3.342328e-13
0	3	0.634326	0.418036	2.450270	1.828316e-29	2.390377	3.703885e-13	70.395388	3.703885e-13
0	4	0.511942	0.478322	2.453768	7.631540e-32	2.294372	1.634737e-15	89.855502	1.634737e-15
0	5	0.655193	0.419314	2.394911	2.384446e-26	2.266633	3.320979e-13	70.765802	3.320979e-13
0	6	0.449113	0.539229	2.229087	1.287334e-29	2.579627	3.519286e-18	114.687127	3.519286e-18
0	7	0.564890	0.416290	2.451792	4.694853e-29	2.332075	4.298233e-13	69.891566	4.298233e-13
0	8	0.528728	0.458596	2.534451	8.562075e-32	2.172932	1.028220e-14	83.010707	1.028220e-14
0	9	0.600431	0.488560	2.422159	1.324847e-31	2.427135	6.127124e-16	93.615900	6.127124e-16
0	10	0.529101	0.435909	2.461988	9.815232e-30	2.326416	7.873598e-14	75.730874	7.873598e-14

Figure 11: Analysis for Sample Size 100 in Dataset 1

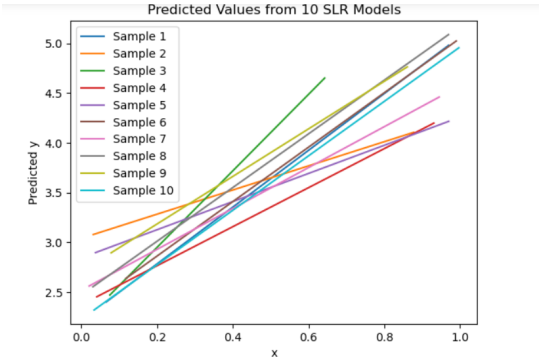
5.1.2. Dataset 2



	Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
0	1	0.081574	0.439948	3.382771	0.004419	1.041131	0.222317	2.356644	0.222317
0	2	0.071430	0.003820	3.336057	0.000963	0.077697	0.921354	0.011505	0.921354
0	3	0.044970	0.947765	2.302809	0.001554	3.645258	0.005148	54.433191	0.005148
0	4	0.018833	0.960494	2.311958	0.000368	2.259780	0.003373	72.938363	0.003373
0	5	0.071482	0.831107	2.220567	0.013110	2.403931	0.031103	14.762718	0.031103
0	6	0.040057	0.875918	2.486020	0.001123	2.193275	0.019291	21.177465	0.019291
0	7	0.189547	0.781068	2.214906	0.009973	3.207421	0.046723	10.702876	0.046723
0	8	0.407779	0.752405	2.239810	0.033526	2.921550	0.056788	9.116558	0.056788
0	9	0.123939	0.861807	2.423080	0.004020	3.005680	0.022781	18.708796	0.022781
0	10	0.119347	0.736995	2.410380	0.009086	2.247743	0.062532	8.406609	0.062532

Figure 12: Analysis for Sample Size 5 in Dataset 2

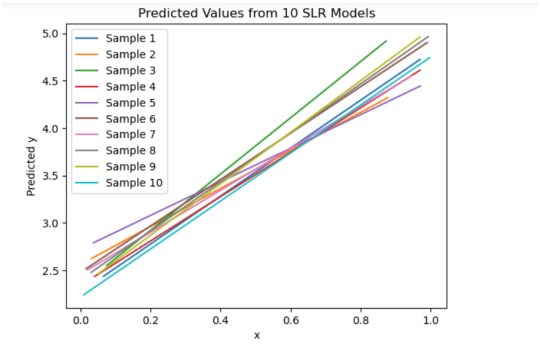
Sample Size 5.



	Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
0	1	0.249894	0.725798	2.213633	5.839469e-04	2.854501	0.001752	21.175609	0.001752
0	2	0.134100	0.425194	3.041069	2.473466e-06	1.214490	0.041033	5.917744	0.041033
0	3	0.188312	0.777842	2.181786	3.332694e-05	3.845744	0.000735	28.010419	0.000735
0	4	0.076182	0.808989	2.373399	1.526307e-06	1.962034	0.000396	33.882327	0.000396
0	5	0.124954	0.627937	2.844552	1.325545e-06	1.415172	0.006269	13.501716	0.006269
0	6	0.050387	0.940394	2.315719	6.497086e-07	2.737196	0.000004	126.215352	0.000004
0	7	0.192582	0.693161	2.519238	6.725131e-06	2.055714	0.002795	18.072290	0.002795
0	8	0.283412	0.778878	2.472025	5.377375e-05	2.699953	0.000721	28.179159	0.000721
0	9	0.205123	0.707646	2.707560	1.476146e-05	2.390544	0.002285	19.364075	0.002285
0	10	0.163794	0.821732	2.228415	1.085376e-05	2.737437	0.000298	36.876184	0.000298

Figure 13: Analysis for Sample Size 10 in Dataset 2

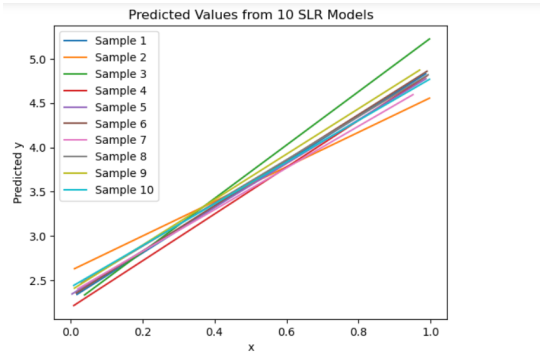
Sample Size 10.



	Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
0	1	0.355801	0.628501	2.270806	1.090713e-07	2.531128	3.066643e-05	30.452354	3.066643e-05
0	2	0.249999	0.519213	2.561034	7.704864e-09	2.006586	3.387101e-04	19.438643	3.387101e-04
0	3	0.215733	0.736657	2.328072	3.975277e-10	2.969141	1.293780e-06	50.351987	1.293780e-06
0	4	0.279370	0.632684	2.340557	8.837863e-08	2.342798	2.761637e-05	31.004117	2.761637e-05
0	5	0.269359	0.525390	2.724784	4.941877e-10	1.772495	3.000080e-04	19.925877	3.000080e-04
0	6	0.080961	0.886977	2.477444	8.459398e-14	2.450309	5.889709e-10	141.259530	5.889709e-10
0	7	0.178749	0.689177	2.459383	1.919527e-11	2.213022	5.920834e-06	39.910825	5.920834e-06
0	8	0.305185	0.725574	2.399501	4.014781e-08	2.584593	1.887416e-06	47.591501	1.887416e-06
0	9	0.271609	0.731393	2.325978	5.798770e-09	2.718982	1.550924e-06	49.012305	1.550924e-06
0	10	0.257887	0.724746	2.218305	1.089064e-08	2.533095	1.940270e-06	47.394146	1.940270e-06

Figure 14: Analysis for Sample Size 20 in Dataset 2

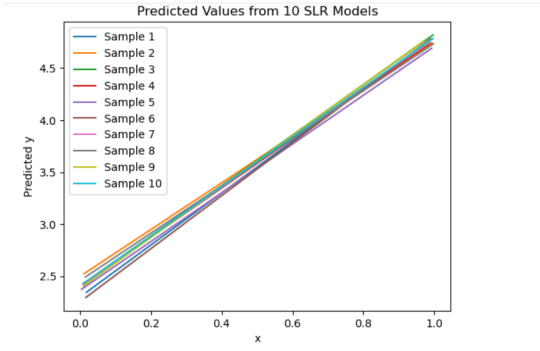
Sample Size 20.



	Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
0	1	0.272729	0.693438	2.292287	1.369029e-21	2.574445	6.475113e-14	108.574943	6.475113e-14
0	2	0.260544	0.559840	2.607518	1.522844e-23	1.954544	4.217883e-10	61.051205	4.217883e-10
0	3	0.240545	0.722335	2.214610	2.152201e-21	3.019711	5.900952e-15	124.870086	5.900952e-15
0	4	0.222384	0.746628	2.188737	2.594232e-21	2.645438	6.453767e-16	141.444730	6.453767e-16
0	5	0.289416	0.674392	2.333211	3.784389e-20	2.474241	2.787293e-13	99.416517	2.787293e-13
0	6	0.167818	0.764902	2.311629	7.328546e-25	2.574283	1.058295e-16	156.169995	1.058295e-16
0	7	0.190847	0.670234	2.353262	5.554246e-25	2.358138	3.790654e-13	97.557673	3.790654e-13
0	8	0.250148	0.708403	2.395288	3.296348e-22	2.441801	1.928383e-14	116.610899	1.928383e-14
0	9	0.288362	0.675697	2.381389	1.112351e-22	2.570186	2.528812e-13	100.009887	2.528812e-13
0	10	0.262693	0.630430	2.419252	3.324778e-20	2.357120	6.011839e-12	81.880628	6.011839e-12

Figure 15: Analysis for Sample Size 50 in Dataset 2

Sample Size 50.

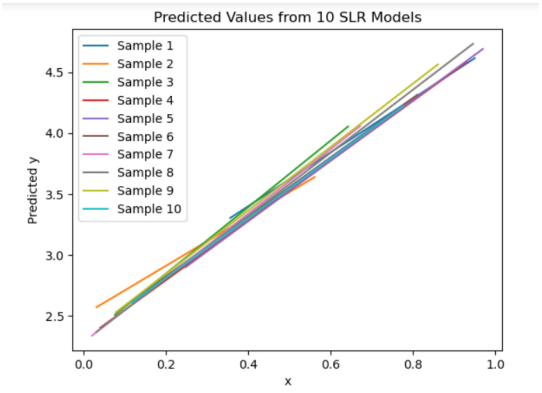


	Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
0	1	0.220279	0.699784	2.299239	1.374398e-44	2.499587	2.372213e-27	228.431595	2.372213e-27
0	2	0.258407	0.645610	2.498016	3.632131e-46	2.249137	8.364659e-24	178.531821	8.364659e-24
0	3	0.281923	0.624908	2.400180	1.617749e-42	2.426918	1.372469e-22	163.269127	1.372469e-22
0	4	0.227530	0.686335	2.402512	2.951043e-45	2.362915	2.050435e-26	214.434934	2.050435e-26
0	5	0.291197	0.634791	2.363274	4.347291e-39	2.344422	3.681101e-23	170.339420	3.681101e-23
0	6	0.199606	0.720609	2.252725	7.766164e-44	2.553085	6.903883e-29	252.763170	6.903883e-29
0	7	0.251062	0.627228	2.401195	4.960240e-42	2.388050	1.010779e-22	164.895609	1.010779e-22
0	8	0.234990	0.677617	2.456301	8.044023e-45	2.281955	7.905206e-26	205.986332	7.905206e-26
0	9	0.266858	0.686773	2.381439	3.902039e-45	2.451424	1.913938e-26	214.872512	1.913938e-26
0	10	0.235156	0.646178	2.407992	8.891292e-43	2.384277	7.729037e-24	178.975784	7.729037e-24

Figure 16: Analysis for Sample Size 100 in Dataset 2

Sample Size 100.

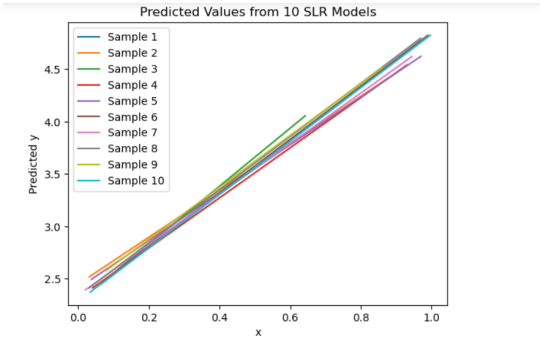
5.1.3. *Dataset 3*



	Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
0	1	0.003263	0.988808	2.516554	0.000091	2.208226	0.000504	265.039109	0.000504
0	2	0.002857	0.984736	2.507211	0.000019	2.015539	0.000804	193.546153	0.000804
0	3	0.001799	0.996082	2.300562	0.000013	2.729052	0.000104	762.729782	0.000104
0	4	0.000753	0.998605	2.302392	0.000003	2.451956	0.000022	2146.786964	0.000022
0	5	0.002859	0.992425	2.284113	0.000108	2.480786	0.000280	393.043937	0.000280
0	6	0.001602	0.995437	2.337204	0.000011	2.438655	0.000131	654.528668	0.000131
0	7	0.007582	0.983738	2.282981	0.000080	2.641484	0.000884	181.478292	0.000884
0	8	0.016311	0.983456	2.287962	0.000315	2.584310	0.000908	178.333751	0.000908
0	9	0.004958	0.991508	2.324616	0.000038	2.601136	0.000333	350.288973	0.000333
0	10	0.004774	0.988123	2.322076	0.000089	2.449549	0.000551	249.597113	0.000551

Figure 17: Analysis for Sample Size 5 in Dataset 3

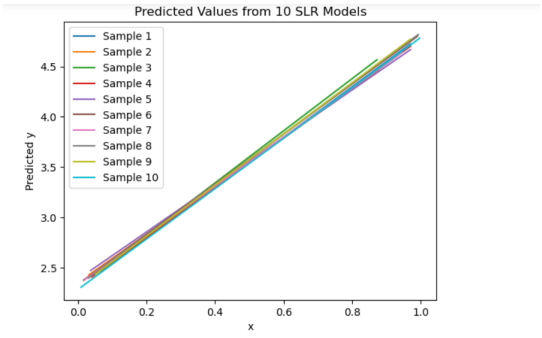
Sample Size 5.



	Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
0	1	0.009996	0.981711	2.282727	2.640295e-09	2.570900	3.081847e-08	429.423662	3.081847e-08
0	2	0.005364	0.984393	2.448214	4.336356e-11	2.242898	1.632722e-08	504.576608	1.632722e-08
0	3	0.007532	0.978441	2.276357	8.885753e-11	2.769149	5.958958e-08	363.070914	5.958958e-08
0	4	0.003047	0.993688	2.314680	5.664014e-12	2.392407	4.351476e-10	1259.419010	4.351476e-10
0	5	0.004998	0.990976	2.408910	1.509098e-11	2.283034	1.820132e-09	878.486712	1.820132e-09
0	6	0.002015	0.997081	2.303144	1.993858e-12	2.547439	1.986349e-11	2733.053675	1.986349e-11
0	7	0.007703	0.987292	2.343848	3.932123e-11	2.411143	7.166804e-09	621.546783	7.166804e-09
0	8	0.011336	0.987331	2.334405	3.336219e-10	2.539991	7.079512e-09	623.476002	7.079512e-09
0	9	0.008205	0.984855	2.381512	1.427614e-10	2.478109	1.447500e-08	520.216372	1.447500e-08
0	10	0.006552	0.990079	2.285683	3.012204e-11	2.547487	2.659140e-09	798.402547	2.659140e-09

Figure 18: Analysis for Sample Size 10 in Dataset 3

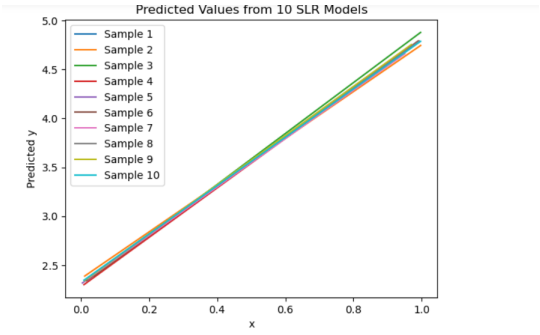
Sample Size 10.



	Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
0	1	0.014232	0.976452	2.294161	1.488961e-19	2.506226	4.173319e-16	746.402377	4.173319e-16
0	2	0.010000	0.974789	2.352207	3.493934e-20	2.401317	7.719739e-16	695.968180	7.719739e-16
0	3	0.008629	0.981608	2.305614	3.167864e-22	2.593828	4.504001e-17	960.676908	4.504001e-17
0	4	0.011175	0.979512	2.308111	1.772760e-19	2.468560	1.190868e-16	860.551715	1.190868e-16
0	5	0.010774	0.979933	2.384957	3.681874e-21	2.354499	9.876302e-17	878.990677	9.876302e-17
0	6	0.003238	0.995089	2.335489	9.137329e-26	2.490062	3.090435e-22	3647.003860	3.090435e-22
0	7	0.007150	0.985408	2.331877	2.531061e-23	2.442604	5.600849e-18	1215.529145	5.600849e-18
0	8	0.012207	0.984297	2.319900	9.735346e-20	2.516919	1.084289e-17	1128.296786	1.084289e-17
0	9	0.010864	0.983494	2.305196	6.460495e-21	2.543796	1.699603e-17	1072.499840	1.699603e-17
0	10	0.010315	0.984723	2.283661	6.798975e-21	2.506619	8.466697e-18	1160.214814	8.466697e-18

Figure 19: Analysis for Sample Size 20 in Dataset 3

Sample Size 20.

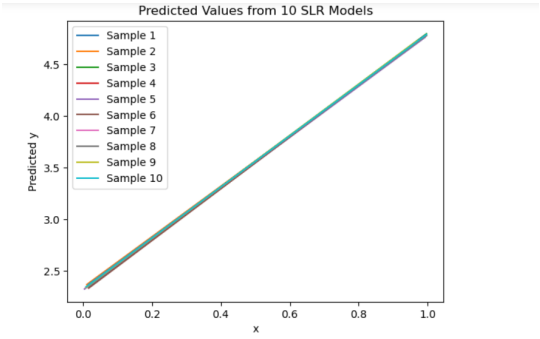


]:

	Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
0	1	0.010909	0.981806	2.298457	1.226922e-53	2.514889	2.000459e-43	2590.240180	2.000459e-43
0	2	0.010422	0.979416	2.361504	9.143639e-54	2.390909	3.875449e-42	2283.860031	3.875449e-42
0	3	0.009622	0.979741	2.282922	5.544360e-54	2.603942	2.643952e-42	2321.294871	2.643952e-42
0	4	0.008895	0.985365	2.277747	4.385643e-54	2.529088	1.074569e-45	3231.909814	1.074569e-45
0	5	0.011577	0.981359	2.306642	1.187822e-51	2.494848	3.582412e-43	2526.985043	3.582412e-43
0	6	0.006713	0.987282	2.302326	3.265780e-57	2.514857	3.698821e-47	3726.074639	3.698821e-47
0	7	0.007634	0.982400	2.310652	4.737352e-57	2.471628	9.011018e-44	2679.348097	9.011018e-44
0	8	0.010006	0.984393	2.319058	1.258081e-53	2.488360	5.036491e-45	3027.505884	5.036491e-45
0	9	0.011534	0.980329	2.316278	2.906003e-54	2.514037	1.302521e-42	2392.198191	1.302521e-42
0	10	0.010508	0.979116	2.323850	4.007833e-51	2.471424	5.485394e-42	2250.360913	5.485394e-42

Figure 20: Analysis for Sample Size 50 in Dataset 3

Sample Size 50.



]:

	Sample	MSE	R2	Coeff_0	pvalue_0	Coeff_1	pvalue_1	F-Statistic	F-pvalue
0	1	0.008811	0.983133	2.299848	3.350789e-110	2.499917	1.079421e-88	5712.298884	1.079421e-88
0	2	0.010336	0.981829	2.339603	3.233229e-109	2.449827	4.150632e-87	5295.353199	4.150632e-87
0	3	0.011277	0.977619	2.320036	2.299796e-106	2.485384	1.132187e-82	4280.757642	1.132187e-82
0	4	0.009101	0.983579	2.320502	1.744921e-109	2.472583	2.904441e-89	5870.040751	2.904441e-89
0	5	0.011648	0.979671	2.312655	7.671020e-103	2.468884	1.017467e-84	4722.643672	1.017467e-84
0	6	0.007984	0.984215	2.290545	4.967051e-110	2.510617	4.187665e-90	6110.605792	4.187665e-90
0	7	0.010042	0.978392	2.320239	8.761033e-106	2.477610	2.022092e-83	4437.395597	2.022092e-83
0	8	0.009400	0.983842	2.331260	3.017469e-108	2.456391	1.317929e-89	5967.046012	1.317929e-89
0	9	0.010674	0.982629	2.316288	1.214595e-109	2.490285	4.579112e-88	5543.475911	4.579112e-88
0	10	0.009406	0.980108	2.321598	1.474560e-106	2.476855	3.505205e-85	4828.609386	3.505205e-85

Figure 21: Analysis for Sample Size 100 in Dataset 3

Sample Size 100. section*Analysis for E2-set1.csv

Explanation for Unimproved R2 with Increasing Sample Size

The unimproved R^2 value with increasing sample size from 10 to 100 suggests that the model's performance is not significantly influenced by the size of the training data. This might indicate that the features in the dataset do not have a linear relationship with the target variable, and increasing the sample size doesn't lead to a more accurate prediction.

Average Values of Other Metrics

Analyze other metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to understand if they show a similar trend. If these metrics also show little improvement, it suggests that the model lacks predictive power or that the chosen features may not be suitable for linear regression.

Conclusion on Usability

Considering the lack of improvement in R^2 and potentially other metrics, it might be prudent to question the usability of the model based on this dataset. It's crucial to assess the model's limitations and whether it can provide reliable predictions in practical scenarios.

Analysis for E2-set3.csv

Explanation for High R2 with Small Sample Size

A high R^2 value even for a small sample size (5) indicates a strong linear relationship between features and the target variable in this dataset. It's possible that the features have a significant impact on the target variable, making the model perform well even with a limited number of samples.

Impact on Other Metrics

Examine the average values of other metrics like MAE and RMSE to ensure that the model's performance is consistently good across different evaluation metrics.

Conclusion Based on Observations

The dataset's characteristics, where a small sample size yields high R^2, could suggest that the model captures the underlying relationship well even with limited data. Consider the nature of the data and the problem domain – if a small sample size consistently produces reliable results, the model may be considered usable in specific contexts.

Helping Factors to be Noted with Regards to Linear Regression

Based on the above observations and general best practices, here are helping Facts for assessing the quality and acceptability of a Linear Regression model:

Evaluate Consistency Across Metrics

Assess the model's performance using multiple metrics such as R^2, MAE, and RMSE. A good model should exhibit consistency in performance across various evaluation criteria.

Consider Sample Size Impact

Analyze the model's sensitivity to sample size. Ensure that increasing sample size improves predictive performance, but also be cautious if very small sample sizes yield surprisingly high performance.

Domain-Specific Evaluation

Consider the specific domain and nature of the data. Some datasets may have inherent characteristics that allow models to perform well with small samples, while others may require larger datasets for accurate predictions.

Cross-Validation and Robustness

Use cross-validation techniques to ensure the model's robustness. Assess how well it generalizes to unseen data.

Interpretability and Explainability

Evaluate the interpretability of the model. A good LR model should provide insights into the relationship between features and the target variable.

Consider Model Limitations

Acknowledge and communicate the limitations of the model. If there are specific scenarios or datasets where the model may not perform well, these should be clearly stated.

By adhering to these guidelines, you can systematically assess and determine the quality and acceptability of a Linear Regression model in various scenarios.