Problem A Report

Introduction

This paper will discuss the methods I used to integrate two distinct datasets. Each dataset comprises a column for the subject ID and the independent or dependent variable value. My goal is to sort and merge these datasets based on subject ID and methodically handle missing values.

Some of my research questions were: how many observations, the fraction of missing data in the independent variable and dependent variable, and the imputation of missing data. The background of this work is to implement efficient practices and statistical tools to find missing data.

Methods

Using R, I sorted the two sets of data by ID, then merged the data into a completed file. There were 542 observations total and 10 observations had missing independent and dependent values. I eliminated the 10 observations as they provide no information, decreasing the number of observations to 532. To impute my data, I used the `MICE` library to implement linear regression with the bootstrap method. Then, I used the `knitr` library to construct my ANOVA table. Finally, I plotted the data using the code below.

Results

The mean for the independent and dependent variables are 4.971 and 50.563, respectively. The standard deviations for the independent and dependent variables are 0.974 and 3.991. The minimum, first quartile, median, third quartile, and maximum for the independent variable are 2.193, 4.309, 4.932, 5.591, and 9.230; for the dependent variable they are 38.722, 48.041, 50.425, 52.870, and 62.953. The correlation coefficient of 0.649 indicates a moderately strong positive linear relationship. The R^2 is 0.4326, so about 43.26% of the variation in the dependent variable is explained. The 95% confidence intervals for the slope and intercept are [2.381, 2.900] and [36.230, 38.831], respectively, while the 99% confidence intervals are [2.300, 2.977] and [35.819, 39.242]. At significance levels of 0.05 and 0.01, the null hypothesis would be rejected.

Conclusion and Discussion

The association between the independent and dependent variables is high and statistically significant ($p = 0.000$). Also, 43.26% of the variance is explained ($R^2 = 0.4326$), which indicates that a moderate proportion of variability in the dependent variable is accounted for by the independent variable.

Appendix B (Tables and Plot)

ANOVA Table

	df	Sum of Squares	Mean Square	F	Sig
Regression (IV)	1	3629.660	3629.65994	405.905	0
Residual	530	4739.334	8.94214	–	–
Total	531	8368.994			

Table 1: ANOVA Table for Problem A

Model Summary

Model	R	R^2	Adjusted R^2	Std. Error	F	df1	df2	Sig F Change
	0.658	0.4337	0.4326	2.99	405.9	1	530	0

Table 2: Model Summary for Problem A

Coefficients

	B	Std. Error	t value	$Pr(>\|t\|)$	Sig
Constant	37.5304	0.6619	56.70	$< 2e^{-16}$	0
IV	2.6383	0.1310	20.15	$< 2e^{-16}$	0

Table 3: Coefficients for Problem A

Confidence Interval

	95% CI		99% CI	
Model	Lower Bound	Upper Bound	Lower Bound	Upper Bound
Constant	36.230106	38.830723	35.819261	39.241568
IV	2.381067	2.895568	2.299786	2.976848

Table 4: Confidence Intervals for Problem A

Residual Statistics

Minimum	First Quartile	Median	Third Quartile	Maximum
-7.971	-2.023	-0.073	2.122	8.075

Table 5: Residual Statistics for Problem A

Graph

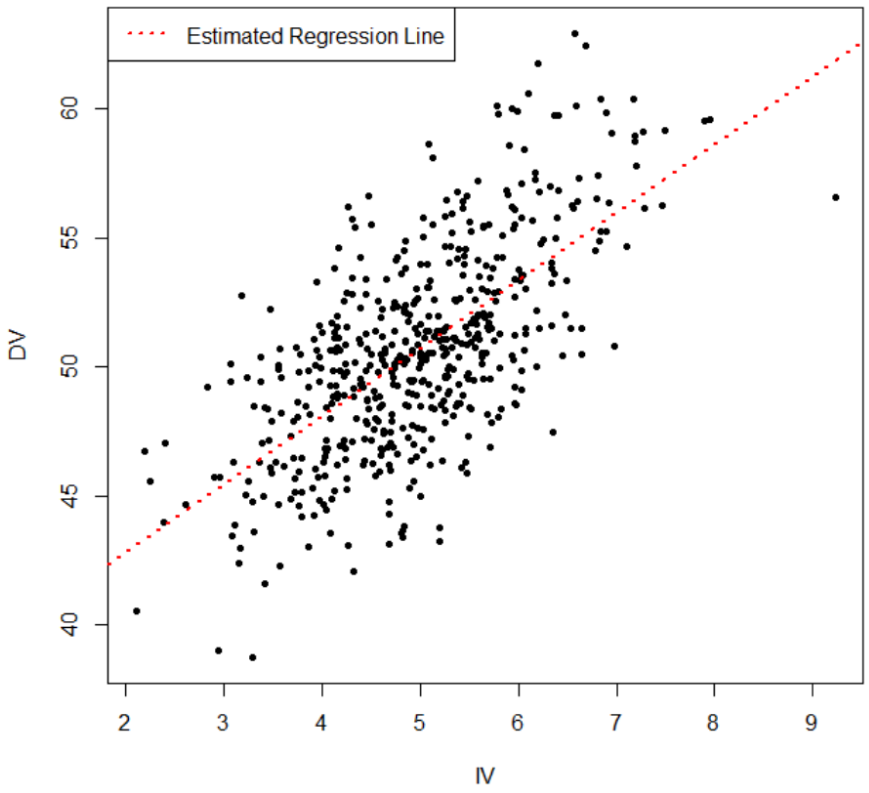

Figure 1: Graph for Problem A

Problem B Report

Introduction

This paper will discuss the methods I used to recover the function that generated the dependent variable based on the value of the independent variable. I analyzed the data in IV and DV by constructing LOF ANOVA tables for each transformation I performed and compared each result to find the best linear model. Some of my research questions were: how many observations, finding the best transformation, and which variable(s) should be transformed. The background of this work is to figure out the trends in the data to analyze the relationships between the independent and dependent variables.

Methods

Using R, I plotted the IV-DV scatter plot to determine an appropriate transformation. For each transformation, I created groups and used the `cut()` function to bin points within a certain interval into one group. I tried to create about 40 groups for each transformation. By using the `remotes` and `alr3` libraries, I constructed pure error ANOVA tables to determine if the linear regression model is satisfactory.

For my first transformation I wanted to transform DV, so I tried IV-DV$^{-\frac{2}{3}}$, and the results were satisfactory. Next, I transformed IV to try to get better results: IV^{-1}-DV; these results were also satisfactory. Then, I decided to transform both IV and DV, and my best result was IV^2-DV$^{-\frac{2}{3}}$.

Results

There are 539 observations. The correlation coefficient for the untransformed scatter plot is -0.634, indicating a moderately strong negative linear relationship. The fitted function with an IV-DV$^{-\frac{2}{3}}$ transformation was:

$$Y = 1.07303\, IV + 0.71955$$

with an R^2 value of 0.4639. The 95% confidence intervals for the slope and intercept are [0.975, 1.171] and [0.586, 0.853], respectively. The lack of fit p-value of 0.5754 shows that the transformation is adequate. The correlation coefficient is 0.682.

Conclusion and Discussion

Overall, my best model is DV =0.39504IV + 1.43968, $IV^2 - DV^{-\frac{2}{3}}$. With a correlation coefficient of 0.681, it indicates a moderately strong positive linear relation. Also, there is a highly significant association between the variables (p = 0.000), and 46.22% of the dependent variables are explained by the independent variables ($R^2 = 0.4622$) which indicates the regression model explains a moderate proportion of variability in the dependent variable. For these reasons, $IV^2 - DV^{-\frac{2}{3}}$ transformation describes the relationship between the two variables in an adequate way.

Appendix B (Tables and Plot)

Table 6: ANOVA for IV-DV$^{-\frac{2}{3}}$ Transformation

	df	Sum of Squares	Mean Square	F
Regression (IV)	1	13.41715	13.41715	466.5148
Residual	537	15.4433	0.02876	–
Lack of fit	47	1.286	0.0274	0.9471
Pure Error	490	14.157	0.0289	–

Table 7: Model Summary for IV-DV$^{-\frac{2}{3}}$ Transformation

R	R^2	Adjusted R^2	Std. Error	F-stat	df1	df2
0.682	0.4649	0.4639	0.1696	466.5	1	537

Table 8: Coefficients for IV-DV$^{-\frac{2}{3}}$ Transformation

	B	Std. Error	t value	$Pr(>\|t\|)$	95% CI Lower	95% CI Upper
Constant	0.71955	0.06775	10.62	$< 2e^{-16}$	0.586	0.853
IV	1.07303	0.04968	21.60	$< 2e^{-16}$	0.975	1.171

Table 9: Residual Statistics for IV-DV$^{-\frac{2}{3}}$ Transformation

Minimum	First Quartile	Median	Third Quartile	Maximum
-0.65818	-0.10470	0.01359	0.11577	0.43896

Table 10: ANOVA for IV^{-1}-DV Transformation

	df	Sum of Squares	Mean Square	F
Regression (IV)	1	0.75008	0.75008	367.9586
Residual	537	1.09467	0.00203	–
Lack of fit	35	0.06949	0.00199	0.9719
Pure Error	502	1.02539	0.00204	–

Table 11: Model Summary for IV^{-1}-DV Transformation

R	R^2	Adjusted R^2	Std. Error	F-stat	df1	df2
0.638	0.4066	0.4055	0.04515	367.96	1	537

Table 12: Coefficients for IV^{-1}-DV Transformation

	B	Std. Error	t value	$Pr(>\|t\|)$	95% CI Lower	95% CI Upper
Constant	-0.01926	0.01776	-1.084	0.279	-0.054	0.016
IV	0.45368	0.02365	19.182	$< 2e^{-16}$	0.407	0.500

Table 13: Residual Statistics for IV^{-1}-DV Transformation

Minimum	First Quartile	Median	Third Quartile	Maximum
-0.10290	-0.02686	-0.00597	0.02000	0.29039

Table 14: ANOVA for IV^2-$DV^{-\frac{2}{3}}$ Transformation

	df	Sum of Squares	Mean Square	F
Regression (IV)	1	13.3880	13.3880	455.0720
Residual	537	15.4734	0.0288	–
Lack of fit	65	1.2754	0.0196	0.6523
Pure Error	472	14.1980	0.0301	–

Table 15: Model Summary for IV^2-$DV^{-\frac{2}{3}}$ Transformation

R	R^2	Adjusted R^2	Std. Error	F-stat	df1	df2
0.681	0.4632	0.4622	0.1699	463.4	1	537

Table 16: Coefficients for IV^2-$DV^{-\frac{2}{3}}$ Transformation

	B	Std. Error	t value	$Pr(>\|t\|)$	95% CI Lower	95% CI Upper
Constant	1.43968	0.03491	41.24	$< 2e^{-16}$	1.371	1.508
IV	0.39504	0.01835	21.53	$< 2e^{-16}$	0.359	0.431

Table 17: Residual Statistics for IV^2-$DV^{-\frac{2}{3}}$ Transformation

Minimum	First Quartile	Median	Third Quartile	Maximum
-0.66273	-0.10390	0.01307	0.11543	0.44436

Graphs for Problem B

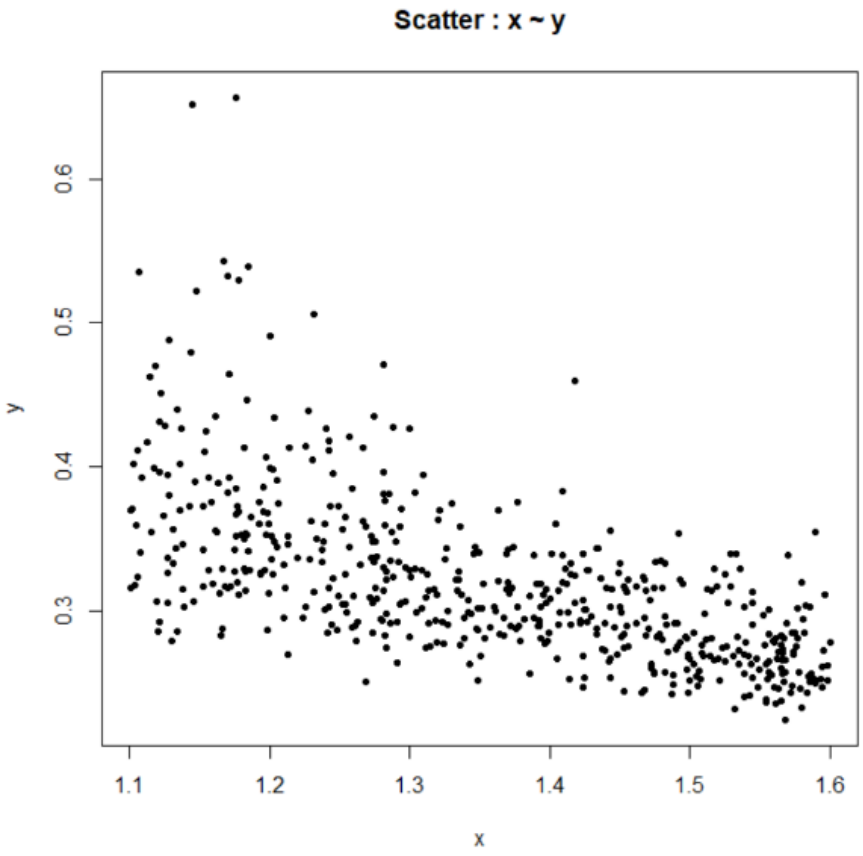

Figure 2: Untransformed Scatter Plot

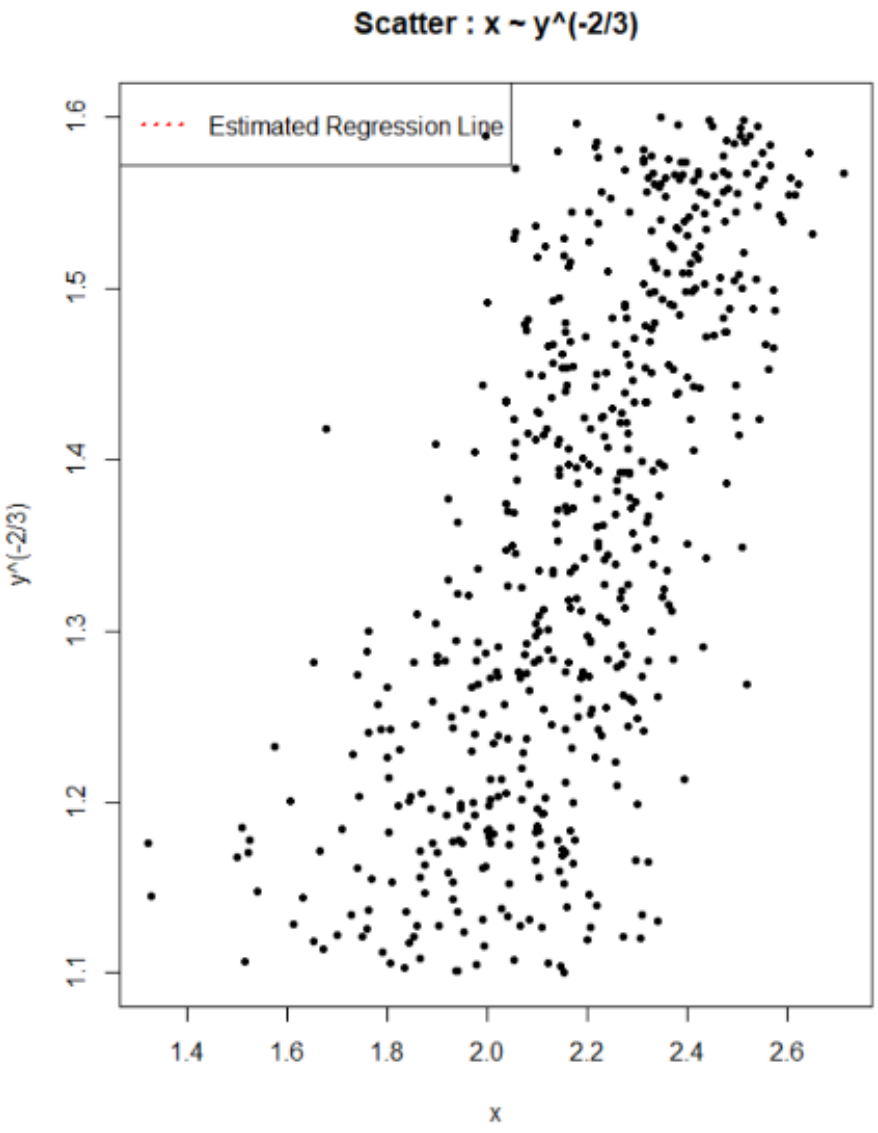

Figure 3: IV vs. DV$^{-\frac{2}{3}}$ Scatter Plot

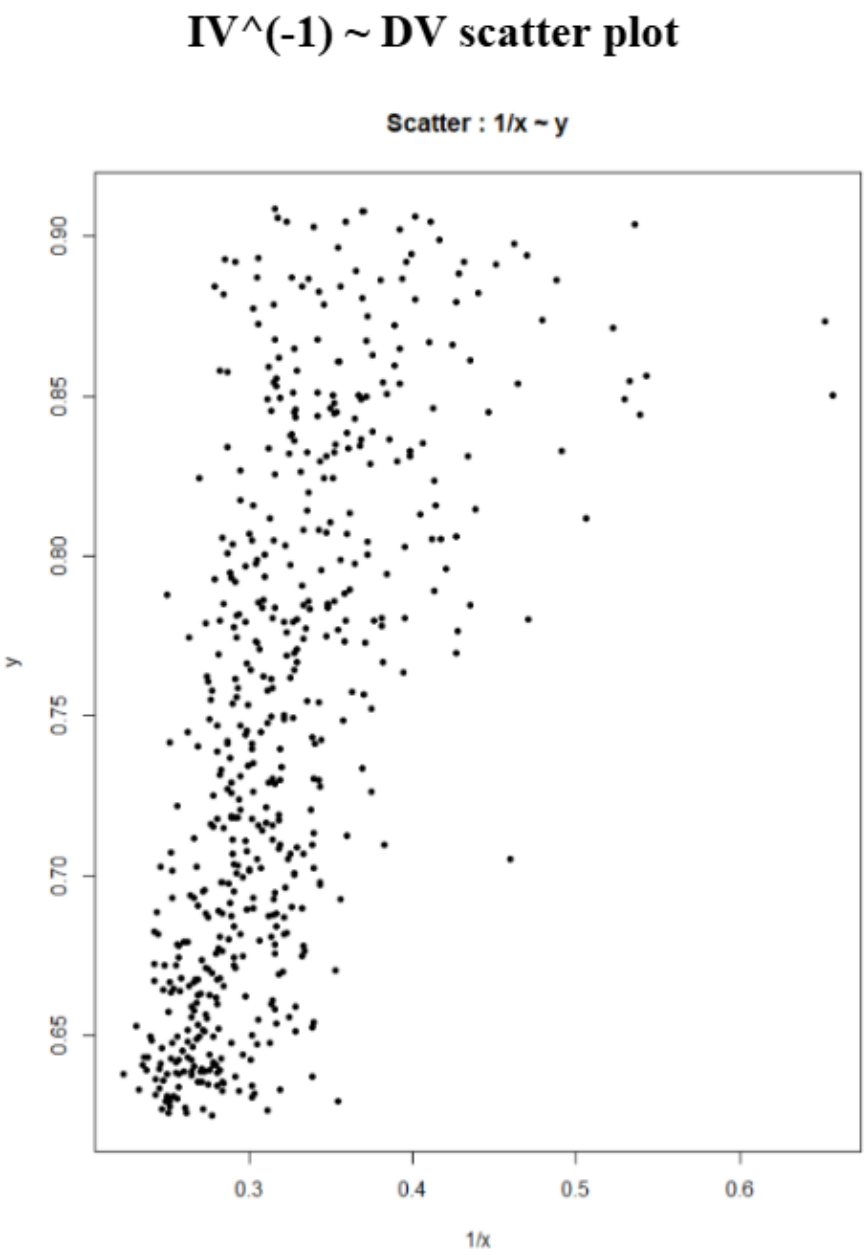

Figure 4: IV^{-1} vs. DV Scatter Plot

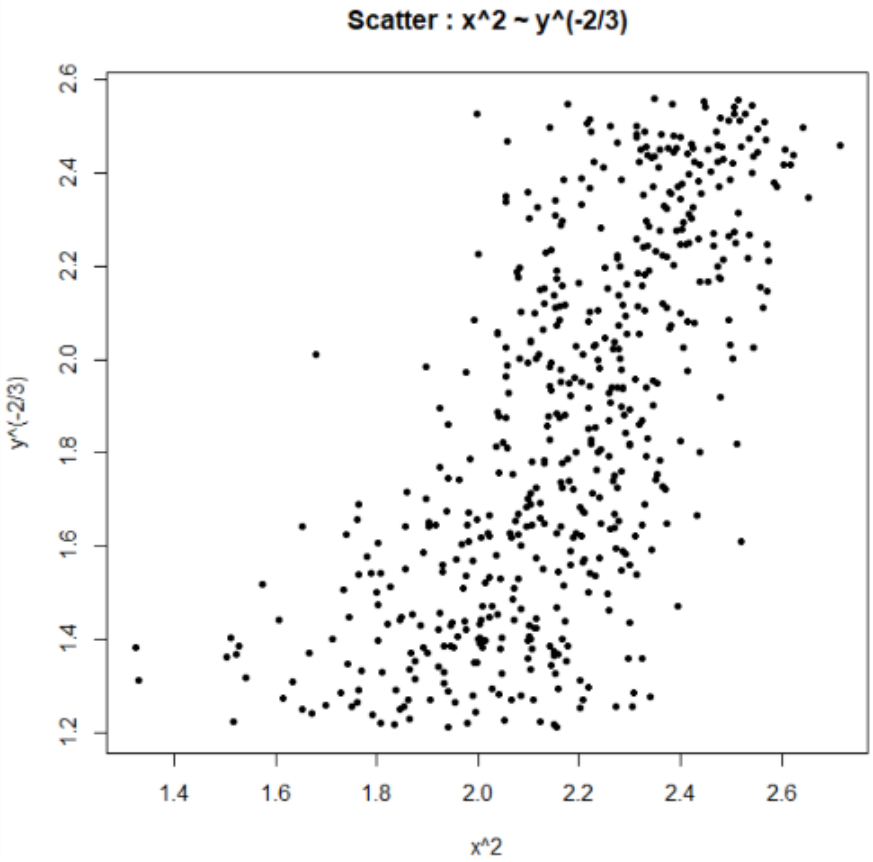

Figure 5: IV^2 vs. $\mathrm{DV}^{-\frac{2}{3}}$ Scatter Plot