Matthew Abruzzese Ott (111520701)
Professor Stephen Finch

AMS 315 Project 1 - Part A

Introduction

The objective of Part A is to merge, sort, and impute data that was generated by a simulation program. This data must then be used to determine the fitted linear function model that best describes the dependent variable value based upon the independent variable value. Various statistical packages (Excel, SPSS) will be used during this data processing.

Methodology

The two original sets of data were given to us in separate spreadsheets, one containing the ID number and the corresponding independent variable (IV) value, and the other containing the ID number and the corresponding dependent variable (DV) value. Both spreadsheets contained 688 ID numbers, along with their associated IV and DV values. Upon being imported to Excel, the two data sets were then sorted by increasing ID number from 1-688 and merged together into one spreadsheet with three columns (ID, IV, DV). In some instances, data entries were missing either their IV or DV values, or they were missing both of their IV and DV values. In other cases, data entries were complete. To identify and record these different data types, multiple codes were used in Excel (see Appendix A). There were 490 data entries with IV and DV values, 145 data entries with an IV but no DV value, 46 data entries with a DV but no IV value, and 7 data entries with neither an IV nor DV value. The data was then imported into SPSS to be analyzed. To account for missing data, the multiple imputation method in SPSS was used. Using the Linear Regression tab in SPSS, an ANOVA table was generated along with a scatter plot and fitted linear regression function for the data.

Results

The fitted linear function model $Y=B_0+B_1(X)$ was expressed as DV=12.24+4.05(IV) with respect to the data. The R-Squared value was 0.560, which explains 56% of the variation of the dependent variable. The 95% confidence interval for the slope (B_1) was [3.775 , 4.315] and the 95% confidence interval for the intercept (B_0) was [10.525 , 13.951]. The ANOVA table (see Appendix A) shows that the independent and dependent variables have a highly significant association (linear relationship), ($R=0.748$, $sig.(p)=0.000 < 0.05$).

Conclusion

For Part A, it was observed that there was a strong association (notable relationship) between the independent and dependent variables ($R=0.748$, $sig.(p)=0.000 < 0.05$). The R-Squared value explained 56% of the dependent variable variation. The linear regression fitted function model DV=12.24+4.05(IV) properly describes the relationship between the independent and dependent variables. The validity of this model can be verified through the scatter plot, residual vs. predicted value plot, and data tables (ANOVA, etc.) in Appendix A.

Appendix A

Code (Excel):

[=MIN(B2:C689)] To find the minimum value out of all IV and DV values.
[=COUNTIFS(B2:B689, ">-1", C2:C689, ">-1")] For the count of ID with IV and DV.
[=COUNTIFS(B2:B689, ">-1", C2:C689, "NA")] For the count of ID with IV but no DV.
[=COUNTIFS(B2:B689, "NA", C2:C689, ">-1")] For the count of ID with DV but no IV.
[=COUNTIFS(B2:B689, "NA", C2:C689, "NA")] For the count of ID with no IV nor DV.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	Change Statistics			
						F Change	df1	df2	Sig. F Change
1	.748[a]	.560	.559	7.1208435668635	.560	864.238	1	679	.000

a. Predictors: (Constant), IV

b. Dependent Variable: DV

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	43822.399	1	43822.399	864.238	.000[b]
	Residual	34429.654	679	50.706		
	Total	78252.054	680			

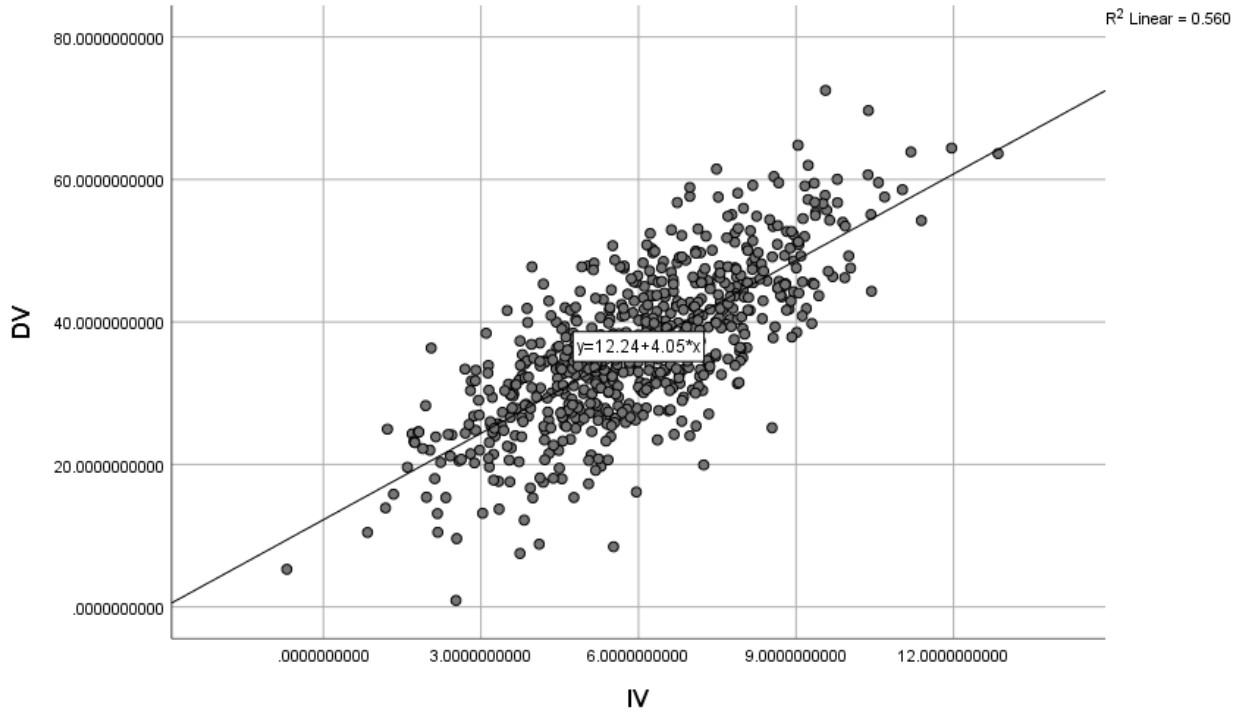
a. Dependent Variable: DV

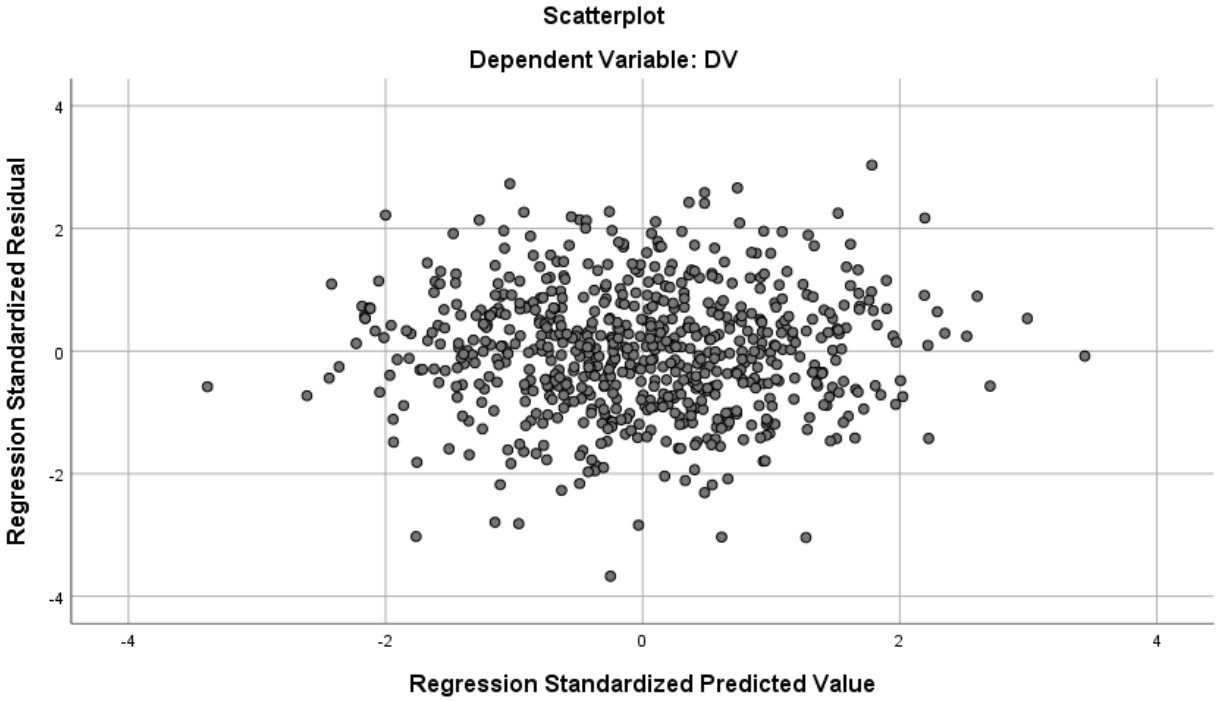
b. Predictors: (Constant), IV

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	12.238	.872		14.027	.000	10.525	13.951
	IV	4.045	.138	.748	29.398	.000	3.775	4.315

a. Dependent Variable: DV





End of Report

Matthew Abruzzese Ott (111520701)
Professor Stephen Finch

AMS 315 Project 1 – Part B

Introduction

Given a data set that was previously merged (contains ID, IV, DV columns) and sorted (by increasing ID), the objective of Part B is to identify the fitted linear function model that best describes the dependent variable value as a result of the independent variable value. In order to optimize the fitted linear function, the independent and dependent variables will undergo multiple transformations. The statistical packages SPSS and Minitab will be used during this data processing.

Methodology

Upon importing the data set into SPSS, eleven DV values were missing, so multiple imputation was used to account for these missing values. Next, the data was transformed in multiple ways using the transform tab in SPSS. This yielded three new variables, sqrt(DV), ln(DV), and $(IV)^2$. The data set, including the three transformed variables was then imported into Minitab. Upon creating a scatter plot of IV vs. DV and generating its ANOVA table, it was observed that using one of the transformed variables might yield a stronger correlation between IV and DV. The scatter plots and statistical tables (ANOVA, etc.) were generated (using Stat/Graph tabs in Minitab) to analyze the linear regression and observe the fitted functions between IV and sqrt(DV), IV and ln(DV), $(IV)^2$ and DV. $(IV)^2$ and DV displayed the strongest correlation, so using SPSS, the data points on its scatter plot were binned and using Minitab, an approximate lack of fit (LOF) test was conducted on the model to see if it adequately represented the data and its correlation. (See Appendix B for all plots, tables, etc.)

Results

The fitted linear function model $Y=B_0+B_1(X)$ was expressed as $DV=7.129+0.4155((IV)^2)$ with respect to the data. An R-Squared value of 0.6173, explained 61.73% of the dependent variable variation in the model. The 95% confidence intervals for the intercept (B_0) and slope (B_1) were [6.732 , 7.525] and [0.3835 , 0.4475] respectively. This model's data tables convey the significant association between the independent and dependent variables (R=0.786). The model's approximate LOF test F-Value=1.05, which is very close to 1. Therefore, the null hypothesis that the model is adequate can be accepted.

Conclusion

It was observed that the correlation between $(IV)^2$ and DV was the most significant when compared to IV and DV, and the other transformations. The strong linear correlation between $(IV)^2$ and DV can be confirmed quantitatively through their statistical tables, as well as visually through the scatter plot of $(IV)^2$ vs. DV and their residual plot. The linear regression between $(IV)^2$ and DV is adequately described by the model $DV=7.129+0.4155((IV)^2)$.

Appendix B

IV and DV

Model Summary

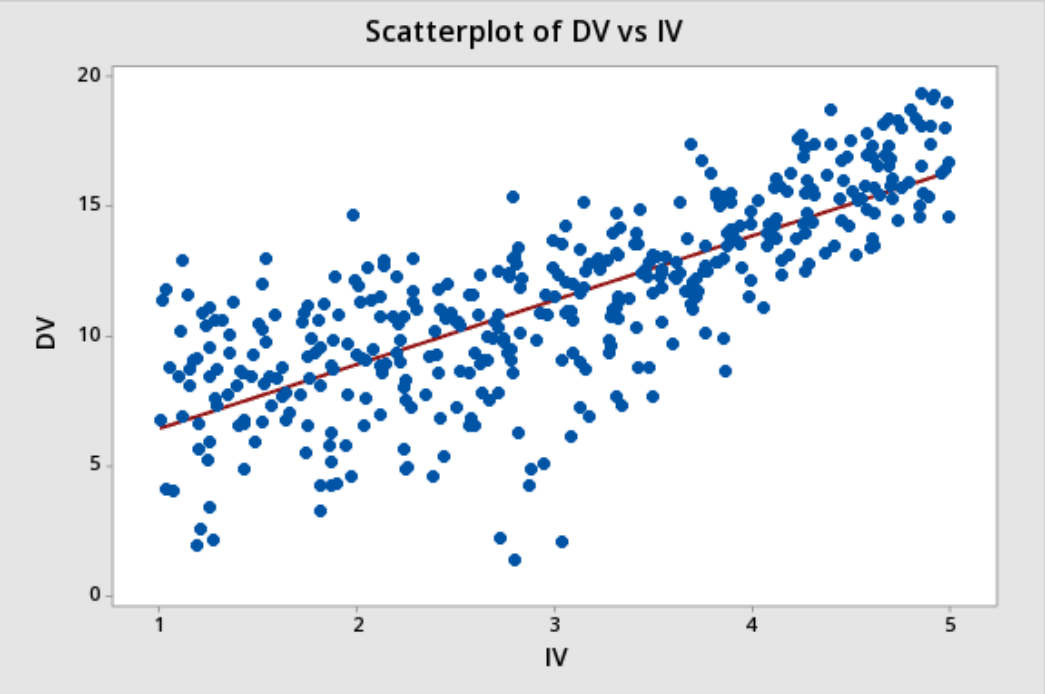
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
2.34880	58.50%	58.39%	2255.79	58.10%	1854.15	1866.11

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	3149	58.50%	3149	3148.98	570.79	0.000
IV	1	3149	58.50%	3149	3148.98	570.79	0.000
Error	405	2234	41.50%	2234	5.52		
Total	406	5383	100.00%				

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	3.997	0.332	(3.345, 4.649)	12.05	0.000	
IV	2.466	0.103	(2.264, 2.669)	23.89	0.000	1.00



IV and sqrt(DV)

Model Summary

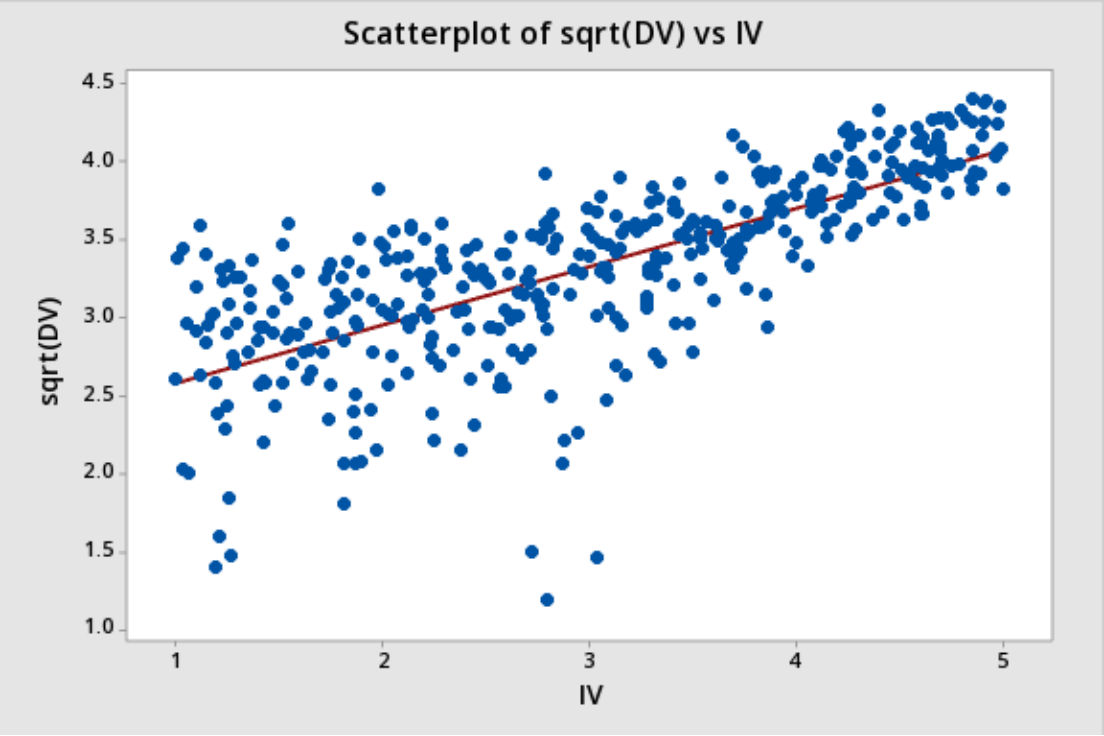
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.395756	53.07%	52.95%	64.0402	52.62%	404.53	416.50

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	71.73	53.07%	71.73	71.7264	457.96	0.000
IV	1	71.73	53.07%	71.73	71.7264	457.96	0.000
Error	405	63.43	46.93%	63.43	0.1566		
Total	406	135.16	100.00%				

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	2.2096	0.0559	(2.0998, 2.3194)	39.55	0.000	
IV	0.3722	0.0174	(0.3381, 0.4064)	21.40	0.000	1.00



IV and ln(DV)

Model Summary

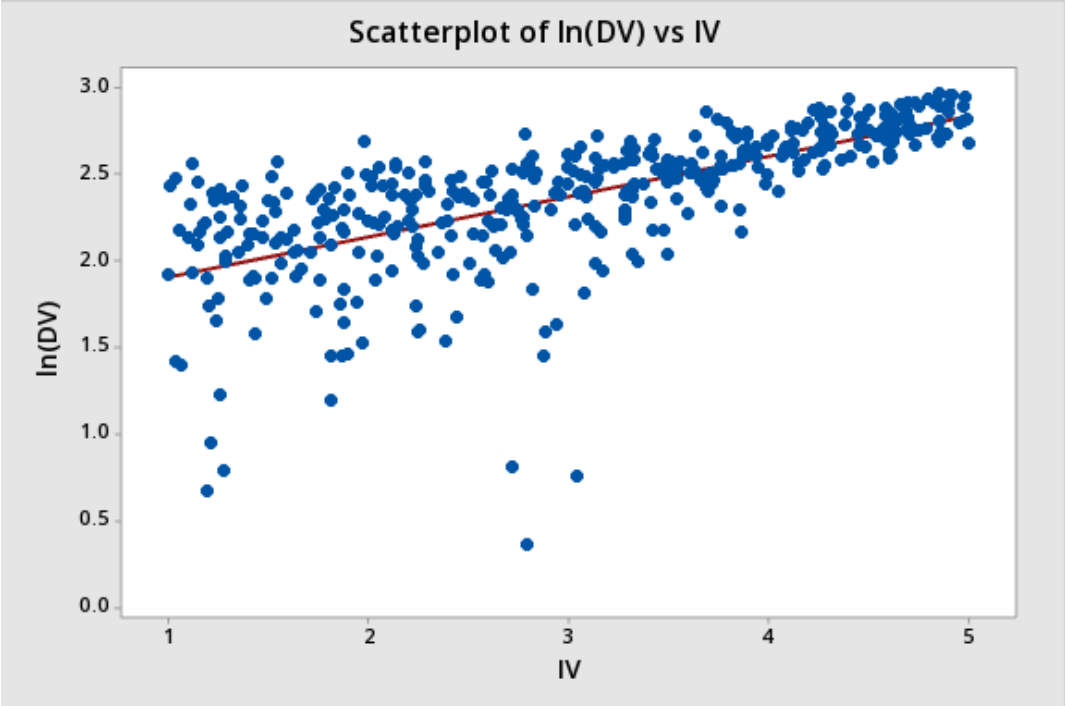
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.292250	44.49%	44.35%	34.9262	43.95%	157.73	169.70

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	27.72	44.49%	27.72	27.7229	324.59	0.000
IV	1	27.72	44.49%	27.72	27.7229	324.59	0.000
Error	405	34.59	55.51%	34.59	0.0854		
Total	406	62.31	100.00%				

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	1.6745	0.0413	(1.5934, 1.7556)	40.59	0.000	
IV	0.2314	0.0128	(0.2062, 0.2567)	18.02	0.000	1.00



$(IV)^2$ and DV
(proper model)

Model Summary

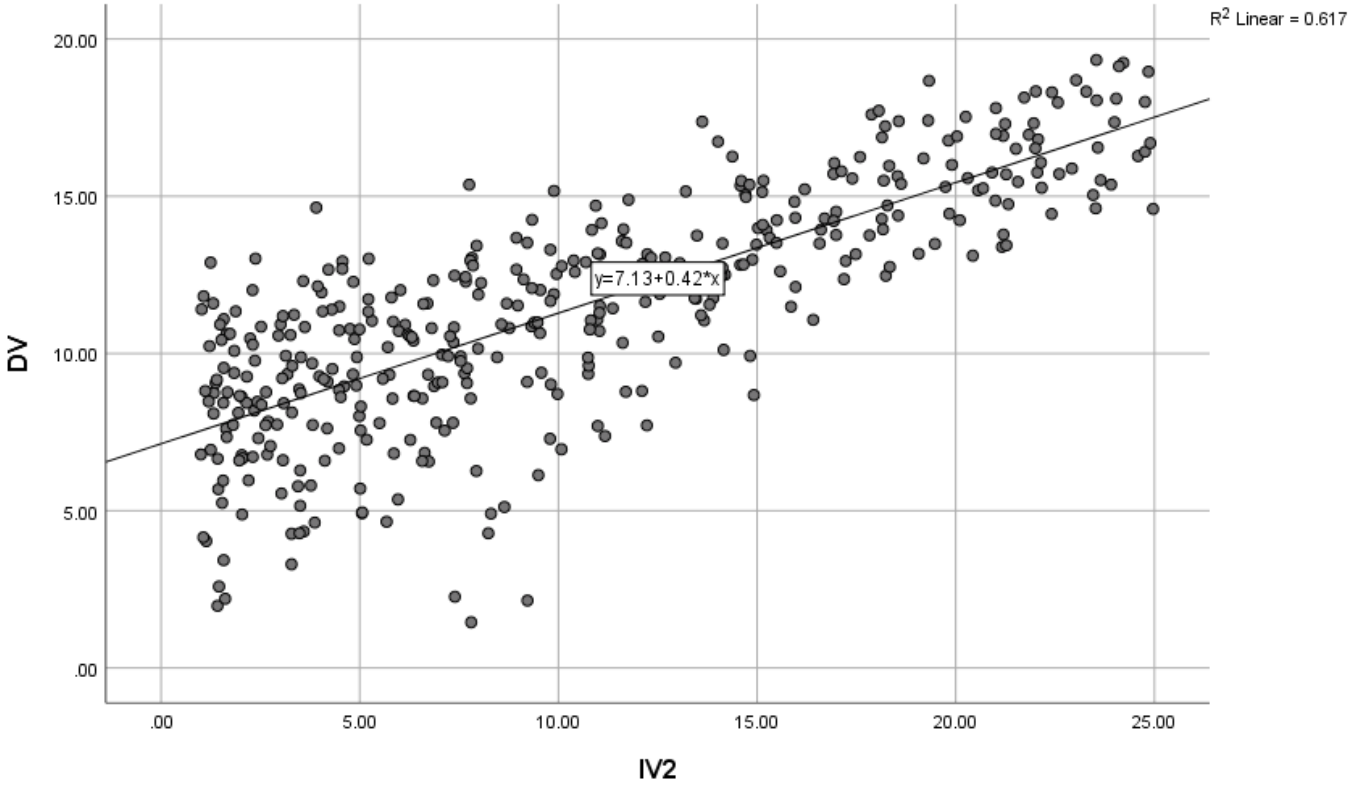
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
2.25558	61.73%	61.63%	2078.98	61.38%	1821.18	1833.15

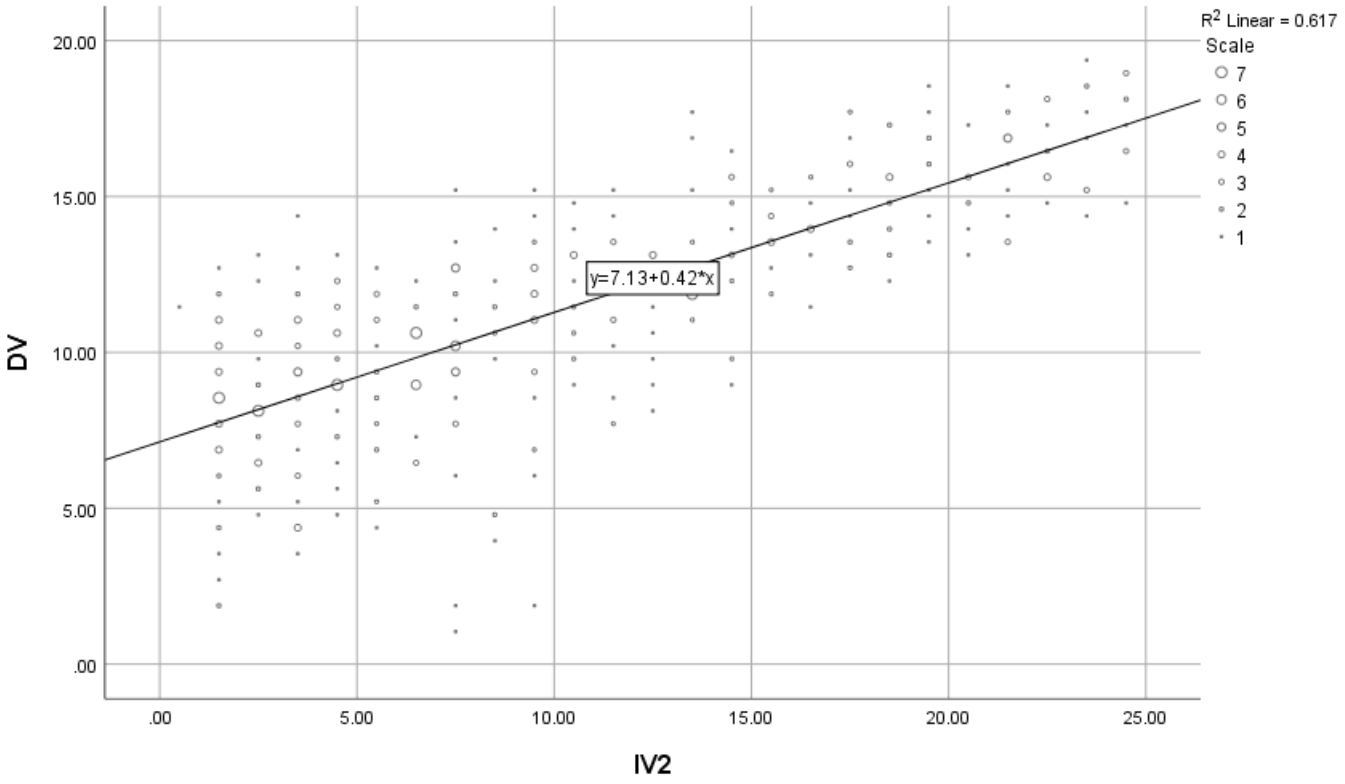
Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	3323.0	61.73%	3323.0	3322.95	653.14	0.000
IV^2	1	3323.0	61.73%	3323.0	3322.95	653.14	0.000
Error	405	2060.5	38.27%	2060.5	5.09		
Lack-of-Fit	370	1889.9	35.11%	1889.9	5.11	1.05	0.453
Pure Error	35	170.6	3.17%	170.6	4.87		
Total	406	5383.4	100.00%				

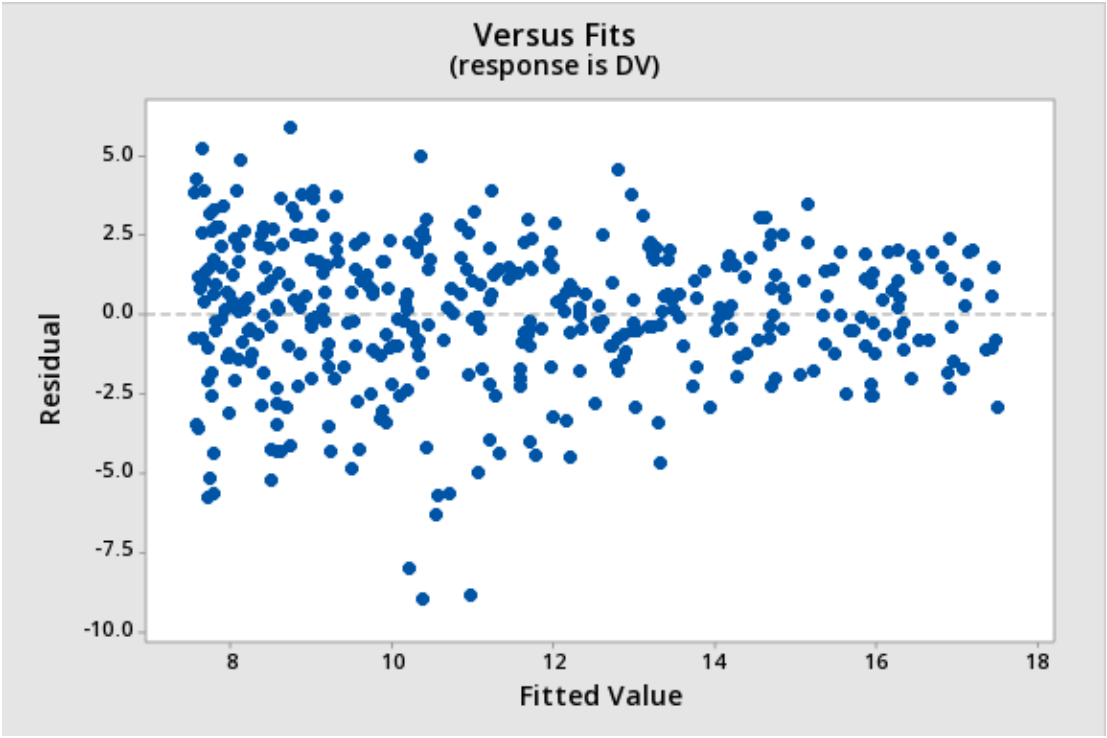
Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	7.129	0.202	(6.732, 7.525)	35.37	0.000	
IV^2	0.4155	0.0163	(0.3835, 0.4475)	25.56	0.000	1.00





(binned data above)



End of Report