Build a Housing Simple Linear Regression Model Using SAS Studio

Didem Bulut Aykurt

MIS500-1 – Foundations of Data Analytics

Colorado State University-Global Campus

Dr. Steve Chung

August 21, 2022

Introduction

In this case, I will work on The Boston Housing Dataset to apply a simple linear regression model by a dependent variable “MEDV” with countries variable “CRIM,” using SAS Studio, which describes concerns housing values in the suburbs of Boston. The data sources originate from the StatLib library at Carnegie Mellon University on July 7, 1993, as follows housing data set available at the UCI Machine Learning repository. Five hundred six intakes represent aggregated data and 13 attributes(features) with the dependent variable(price).

- CRIM - per capita crime rate by town
- ZN – the proportion of residential land zoned for over 25,000 sq. ft.
- INDUS – the proportion of non-retail business acres per town.
- CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX - nitric oxides concentration (parts per 10 million)
- RM - average number of rooms per dwelling
- AGE – the proportion of owner-occupied units built before 1940
- DIS - weighted distances to five Boston employment centers
- RAD - index of accessibility to radial highways
- TAX - full-value property-tax rate per $10,000
- PTRATIO - pupil-teacher ratio by town

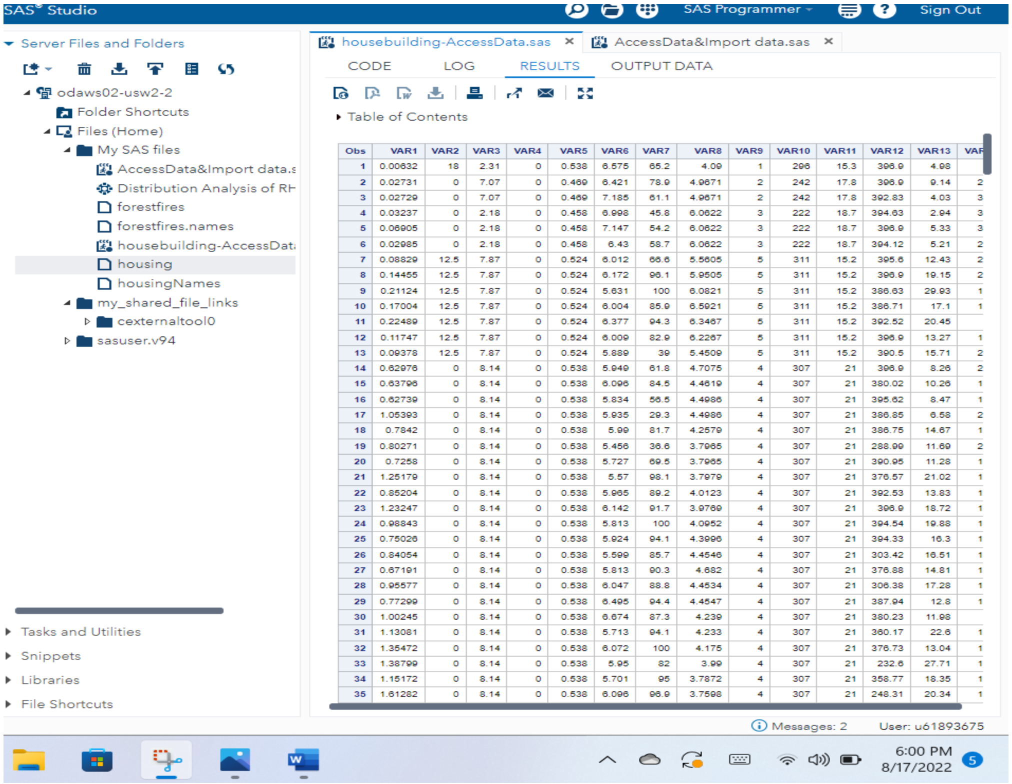
- B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT - % lower status of the population
- MEDV - Median value of owner-occupied homes in $1000's

Uploading and importing the housing. Data CSV(Comma-delimited) file to SAS Studio

The first step to uploading CSV data from my computer into SAS Studio is right click on the existing folder “My SAS files” under the main list “Files(Home)” and upload a file from my laptop. Now is the “housing. data” dataset online on SAS Studio; change the file name right, click then rename “housing” Screenshots 1 and 2 have details.

Screenshot 1

To upload the housing data set in SAS Studio.

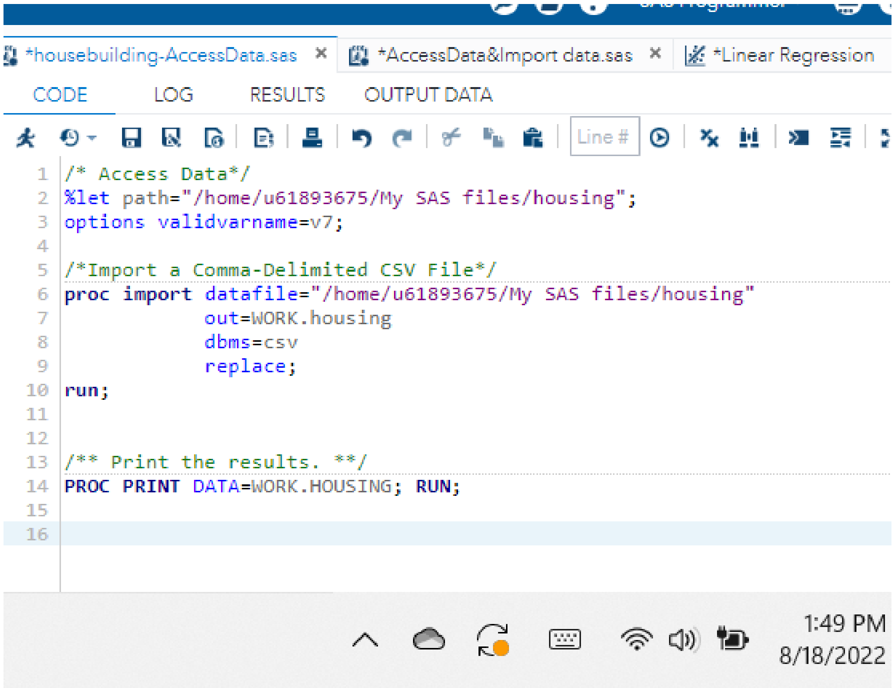


Obs	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7	VAR8	VAR9	VAR10	VAR11	VAR12	VAR13	VAF
1	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	
2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	2
3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	3
4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	3
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	3
6	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	2
7	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	2
8	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	2
9	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	1
10	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	1
11	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	
12	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	1
13	0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	2
14	0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	2
15	0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	1
16	0.62739	0	8.14	0	0.538	5.834	56.5	4.4986	4	307	21	395.62	8.47	1
17	1.05393	0	8.14	0	0.538	5.935	29.3	4.4986	4	307	21	386.85	6.58	2
18	0.7842	0	8.14	0	0.538	5.99	81.7	4.2579	4	307	21	386.75	14.67	1
19	0.80271	0	8.14	0	0.538	5.456	36.6	3.7965	4	307	21	288.99	11.69	2
20	0.7258	0	8.14	0	0.538	5.727	69.5	3.7965	4	307	21	390.95	11.28	1
21	1.25179	0	8.14	0	0.538	5.57	98.1	3.7979	4	307	21	376.57	21.02	1
22	0.85204	0	8.14	0	0.538	5.965	89.2	4.0123	4	307	21	392.53	13.83	1
23	1.23247	0	8.14	0	0.538	6.142	91.7	3.9769	4	307	21	396.9	18.72	1
24	0.98843	0	8.14	0	0.538	5.813	100	4.0952	4	307	21	394.54	19.88	1
25	0.75026	0	8.14	0	0.538	5.924	94.1	4.3996	4	307	21	394.33	16.3	1
26	0.84054	0	8.14	0	0.538	5.599	85.7	4.4546	4	307	21	303.42	16.51	1
27	0.67191	0	8.14	0	0.538	5.813	90.3	4.682	4	307	21	376.88	14.81	1
28	0.95577	0	8.14	0	0.538	6.047	88.8	4.4534	4	307	21	306.38	17.28	1
29	0.77299	0	8.14	0	0.538	6.495	94.4	4.4547	4	307	21	387.94	12.8	1
30	1.00245	0	8.14	0	0.538	6.674	87.3	4.239	4	307	21	380.23	11.98	
31	1.13081	0	8.14	0	0.538	5.713	94.1	4.233	4	307	21	360.17	22.6	1
32	1.35472	0	8.14	0	0.538	6.072	100	4.175	4	307	21	376.73	13.04	1
33	1.38799	0	8.14	0	0.538	5.95	82	3.99	4	307	21	232.6	27.71	1
34	1.15172	0	8.14	0	0.538	5.701	95	3.7872	4	307	21	358.77	18.35	1
35	1.61282	0	8.14	0	0.538	6.096	96.9	3.7598	4	307	21	248.31	20.34	1

Messages: 2 User: u61893675

Screenshot 2

Access and import the housing data set in SAS Studio.



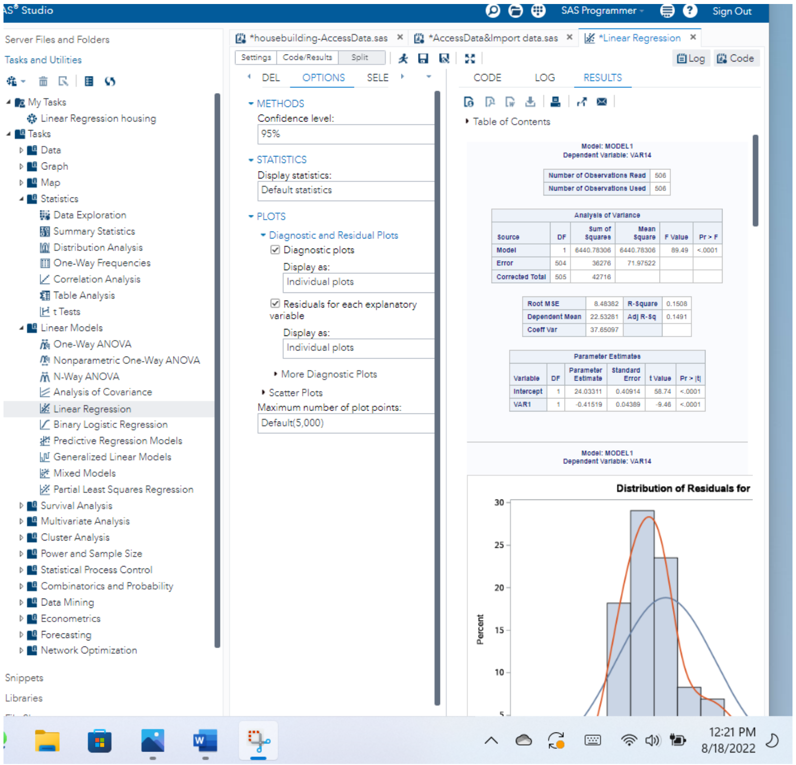
```
/* Access Data*/
%let path="/home/u61893675/My SAS files/housing";
options validvarname=v7;

/*Import a Comma-Delimited CSV File*/
proc import datafile="/home/u61893675/My SAS files/housing"
            out=WORK.housing
            dbms=csv
            replace;
run;


/** Print the results. **/
PROC PRINT DATA=WORK.HOUSING; RUN;
```

To apply a simple linear regression model to the "housing" dataset.

Now housing dataset is ready to analyze, and I will do the first step to click "Tasks and Utilities" on the main menu and then double click "Linear Regression" under the "Linear Model." It will pop up a new Linear regression page. The middle of the page has a "DATA" organizer to choose the dependent variable VAR14(MEDV) and the continuous variable VAR1(CRIM). The next step is to edit the "Model" effect to a variable of VAR1. The final part is "OPTION" to create diagnostic plots and residuals for each variable with individual stories. Screenshot 3 shows the detail.

Screenshot 3

The simple linear regression model of housing dataset in SAS Studio.



METHODS
Confidence level: 95%

STATISTICS
Display statistics: Default statistics

PLOTS
Diagnostic and Residual Plots
☑ Diagnostic plots
Display as: Individual plots
☑ Residuals for each explanatory variable
Display as: Individual plots
More Diagnostic Plots
Scatter Plots
Maximum number of plot points: Default(5,000)

Model: MODEL1
Dependent Variable: VAR14

Number of Observations Read	506
Number of Observations Used	506

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6440.78306	6440.78306	89.49	<.0001
Error	504	36276	71.97522		
Corrected Total	505	42716			

Root MSE	8.48382	R-Square	0.1508
Dependent Mean	22.53281	Adj R-Sq	0.1491
Coeff Var	37.65097		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	24.03311	0.40914	58.74	<.0001
VAR1	1	-0.41519	0.04389	-9.46	<.0001

Model: MODEL1
Dependent Variable: VAR14

The result for linear regression model of housing data set in SAS Studio.

The first result table has statistical detail as we can say one Way ANOVA table. That table has great points like R-square, P-value, F-value, Sum of squares, and RMSE.

Look at each result to see how our data fit a linear and normal distribution. **SS**’s best result is zero, the case result of 6440. A higher number to tell not to do the housing dataset to the linear regression model.

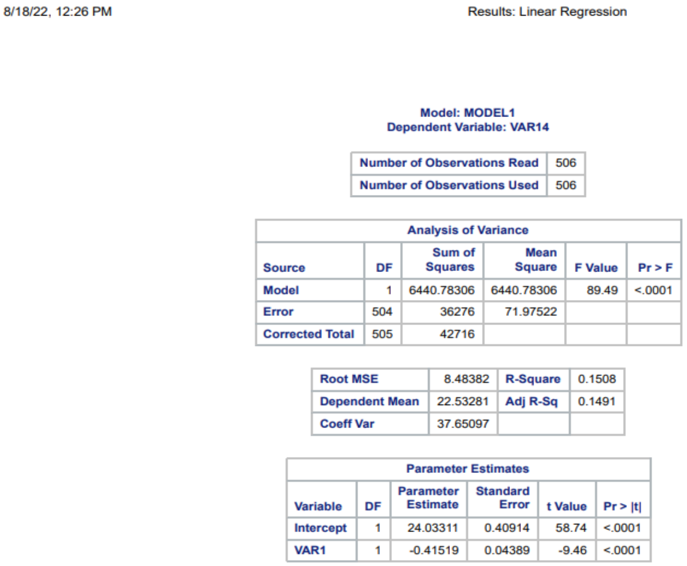
F value, one-way ANOVA measures how a group of variables is jointly significant. The result from the critical of the F value table of 2.47.

R-square also tells us how strong the relation between two variables is. The result of 0.15 is 15%, pretty low to apply these two variables. That other way to say VAR1 is not a prediction for this model result.

The P-value shows how strongly this model supported the housing data set. The result of 0.001 is low. The simple linear regression model strongly supports the housing dataset as the linear regression model is statistically significant—all the table detail in screenshot 4.

The following table has the parameter estimate of VAR1 -0.42, which tells us VAR14 and VAR 1 have a negative correlation as VAR14 increases every unit, and VAR1 decreases by -0.42.

Screenshot 4



8/18/22, 12:26 PM Results: Linear Regression

Model: MODEL1
Dependent Variable: VAR14

Number of Observations Read	506
Number of Observations Used	506

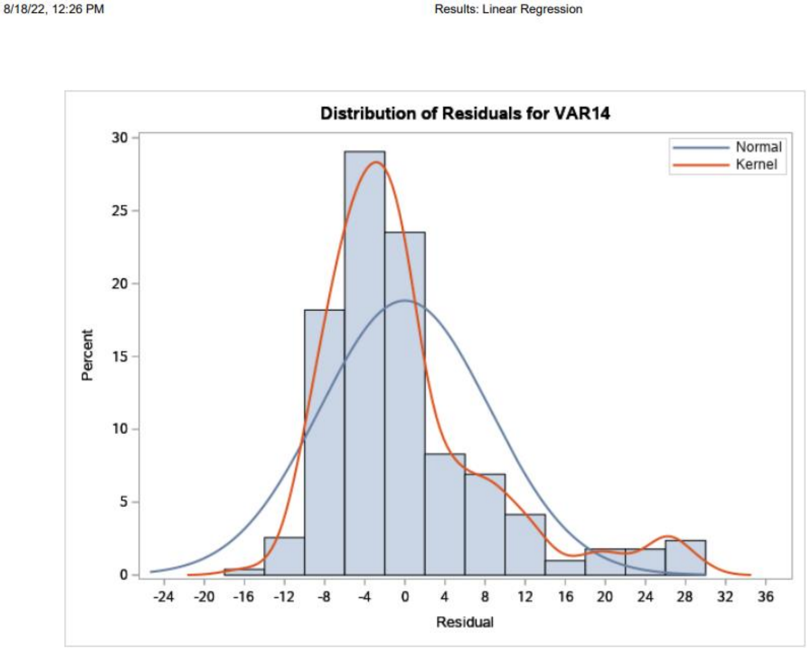
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6440.78306	6440.78306	89.49	<.0001
Error	504	36276	71.97522		
Corrected Total	505	42716			

Root MSE	8.48382	R-Square	0.1508
Dependent Mean	22.53281	Adj R-Sq	0.1491
Coeff Var	37.65097		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	24.03311	0.40914	58.74	<.0001
VAR1	1	-0.41519	0.04389	-9.46	<.0001

Screenshot 5

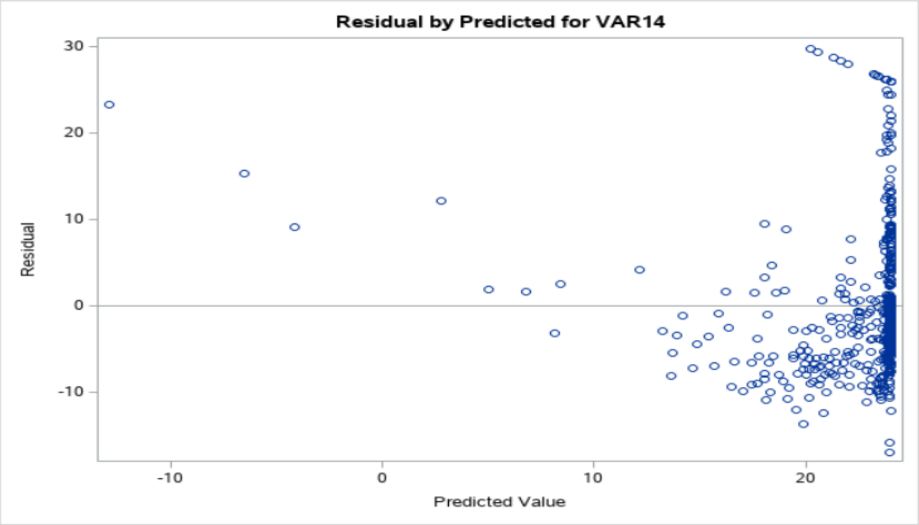
The outcome of distribution of residual for VAR14



8/18/22, 12:26 PM Results: Linear Regression

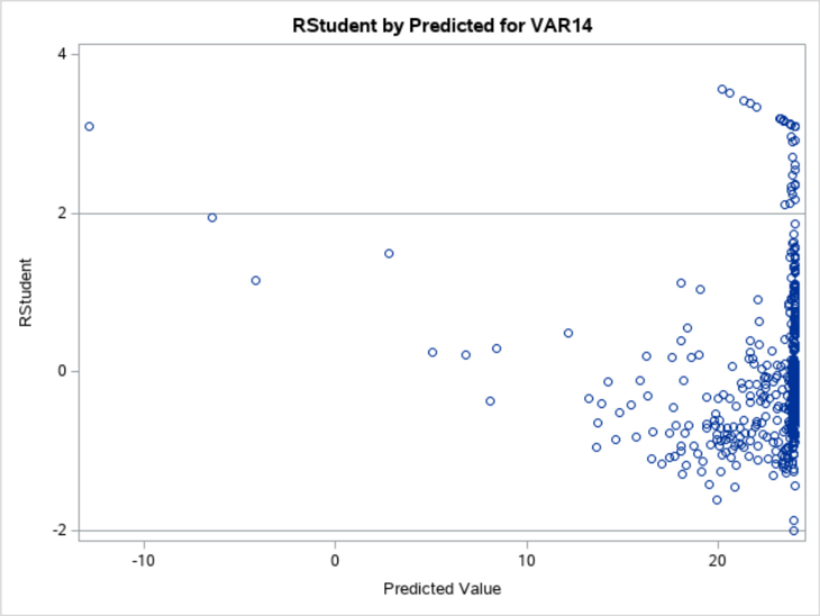
Screenshot 6

8/18/22, 12:26 PM Results: Linear Regression



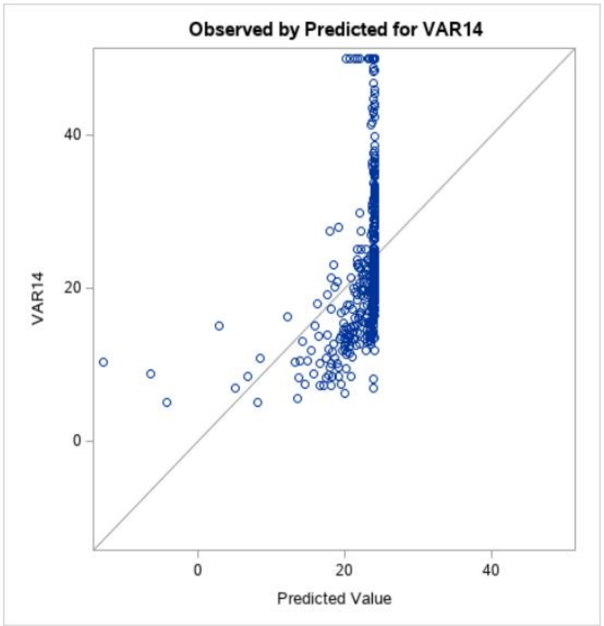
Screenshot 7

8/18/22, 12:26 PM Results: Linear Regression



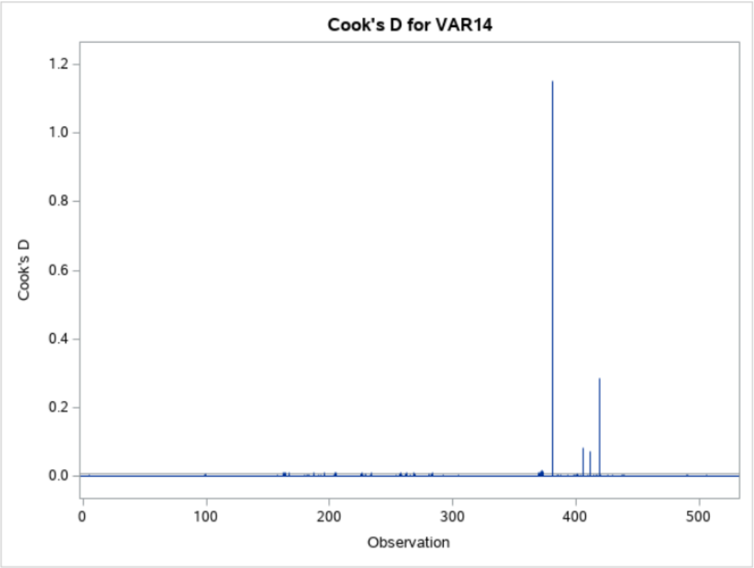
Screenshot 8

8/18/22, 12:26 PM Results: Linear Regression



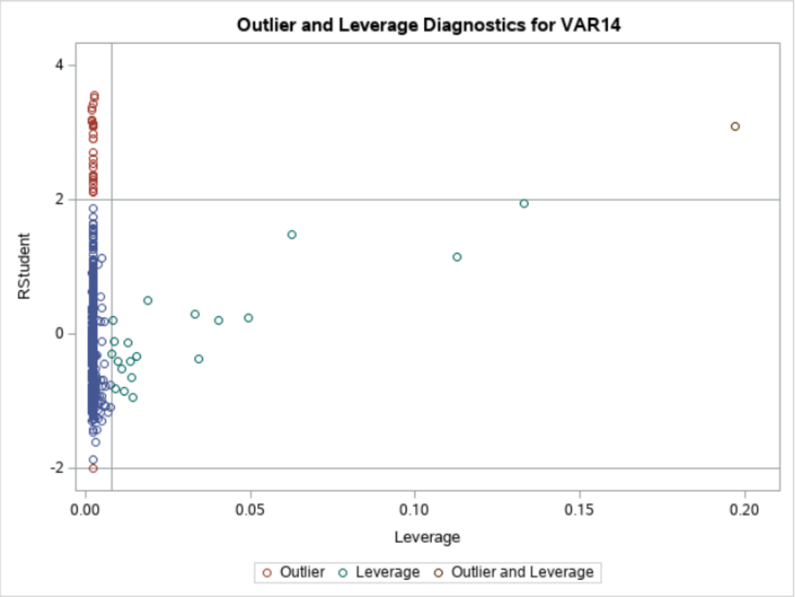
Screenshot 9

8/18/22, 12:26 PM Results: Linear Regression



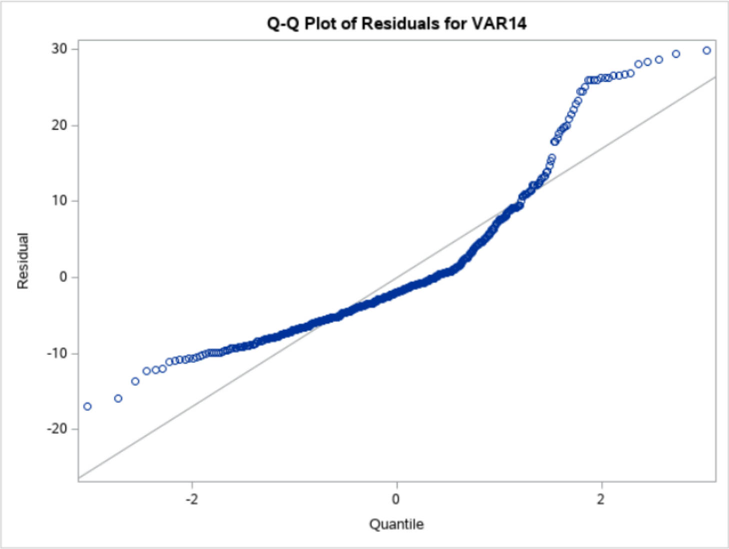
Screenshot 10

8/18/22, 3:05 PM Results: Linear Regression



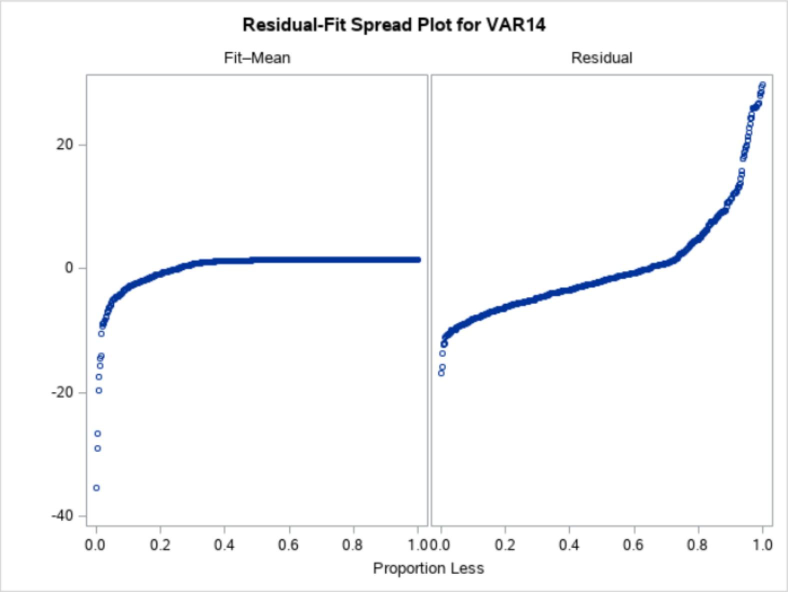
Screenshot 11

8/18/22, 3:05 PM Results: Linear Regression



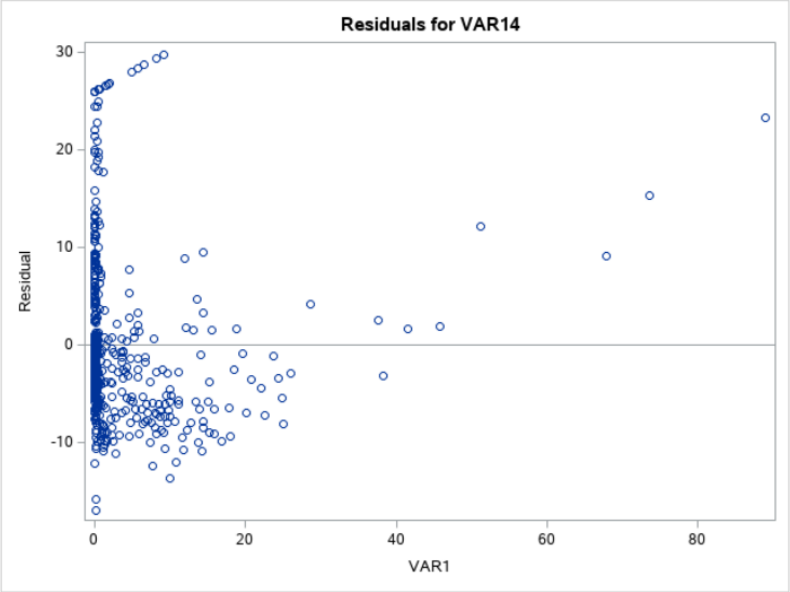
Screenshot 12

8/18/22, 3:05 PM Results: Linear Regression

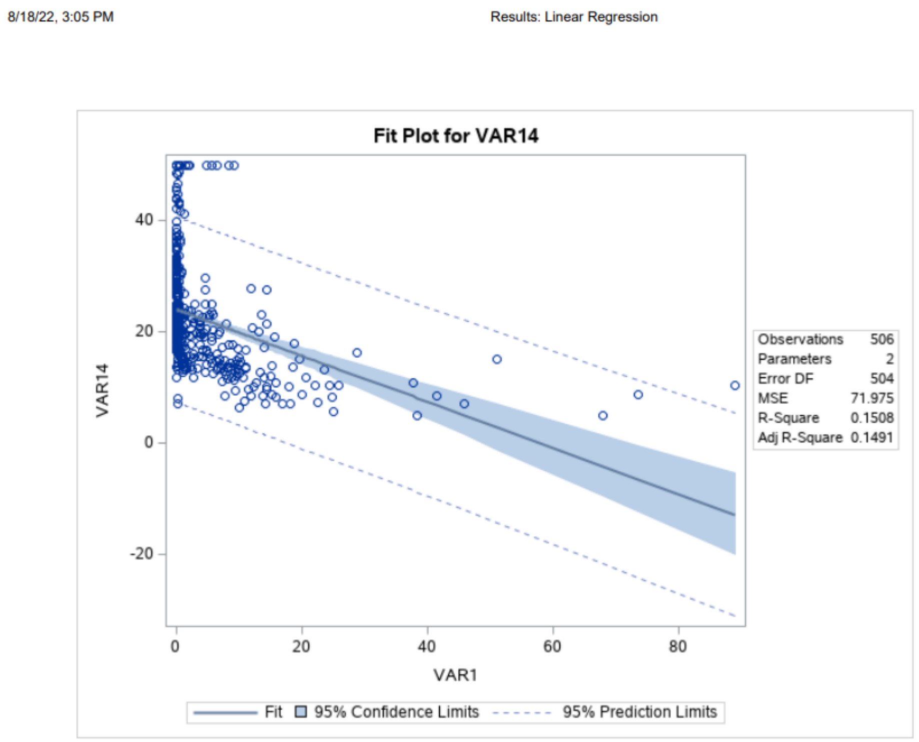


Screenshot 13

8/18/22, 3:05 PM Results: Linear Regression



Screenshot 14



8/18/22, 3:05 PM Results: Linear Regression

Fit Plot for VAR14

Observations	506
Parameters	2
Error DF	504
MSE	71.975
R-Square	0.1508
Adj R-Square	0.1491

As a result of the definition, **screenshot 5** tells us the residual of the distribution plot for VAR14 has a heavy-tailed residual, which means there are many positive and negative residuals.

Screenshot 11 shows a quantile plot of residuals for VAR14. The relationship between the residual and quantile is not linear.

Screenshot 8 tells how VAR14 and VAR1 don't have a linear relationship that does not have a linear connection.

Screenshots 6, 7, and 13 tell the linear regression model is not a good choice because all the variables are not distributed homogeneously around the residual line or are not a random scatter against the predicted value as the residual should equal zero but not in this case. That means our model's not picking up all the signals of the dataset. It also implies that variability increases as the expected value increases.

Screenshot 14, the final chart, is the most interesting. The blue line represents the confidence interval for the average and is mainly created using y-intercept and that slope with a blue area. Those confidence intervals allow us to be confident in what we can say for the individual to predict value-dependent value. That also shows outlines out the outside lines.

Conclusion

Thus, I used the sample linear regression model for VAR1 and VAR14. An essential point for the linear model that is dependent and independent should be linear, independency, normal distribution, and equal error. In this case, two variables are not a linear and normal distribution. Other, not an independency and equal variance. That result might be outliers affecting.

Reference

Housing Dataset

Index of /ml/machine-learning-databases/housing (uci.edu)

Normal Probability Plot of Residuals

4.6 - Normal Probability Plot of Residuals | STAT 462 (psu.edu)