# Multiple Linear Regression Analysis for Income Prediction

Akash Manjunatha
*School of Computing*
*National College of Ireland*
Dublin, Ireland
x21141797@student.ncirl.ie

***Abstract*— This project's goal is to forecast Income based on a variety of parameters. An effort has been made to develop a multiple regression model between Income and other variables. R Programming will be used to apply various regression models to the dataset.**

## I. OBJECTIVES

The project aims to improve comprehension of the dataset's variables by using descriptive statistics and related visuals. Using various Multiple regression models, we try to get the satisfactory Adjusted r squared and r squared value. Later the model will be diagnosed to verify the Gauss Markov and other relevant assumptions of multiple regression If not, the variables will be transformed and the analysis will be repeated.

**Multiple Linear Regression:** Multiple regression is a modeling technique that is used to determine the relationship between a single independent variable and several numbers of independent variables.[1]

Equation:

$$Y = \beta 0 + \beta 1 X1 + \beta 2 X2 + \cdots + \beta p Xp + \epsilon$$

Where $\beta 0$ is intercept, $X1$ – $Xp$ are the independent variables, $\beta 1$ - $\beta p$ are the regression coefficient, and the error associated with the regression is denoted by the letter $\epsilon$.

**R Square:** It is the proportion of variation in the dependent variable that can be explained by a variety of independent factors. It is easy to interpret the results a number close to 0 indicates that the set of independent factors and the dependent variable has a low correlation. A value close to 1 indicates a strong link and Squaring the value diminishes the negative value[2].

Formula:

$$R^2 = 1 - \frac{sum\ squared\ regression\ (SSR)}{total\ sum\ of\ squares\ (SST)}$$

**Adjusted R2:** It Quantifies the proportion of the variation in the independent variable that is influencing the dependent variable.

The limitation of R2 is that it increases when independent variables are added to the model, which is misleading because some added variables may be useless with minimal significance. Adjusted R2 overcomes this issue by imposing a penalty if we attempt to add independent variables that do not improve the model.

Formula:

$$R^{-2} = 1 - \frac{SSres/df_e}{SS_{tot}/df_t}$$

**Residual Standard Error:** It's the response's average divergence from the true regression line.

**Assumptions of Gauss Markov's:**

- Linearity – According to this assumption, the dependent variable is linearly related to the predictor variables. There should be no curves or grouping factors.
- Homoscedasticity – There should be no consistent relationship between the residuals and the fitted values, that is errors should have Constance variance. The distribution of values of the predictor variables can be shown in a graph of standard residuals vs fitted values.
- Independence of errors – Another key assumption is there should be no autocorrelation between errors. This specifies that noise should be independent of one another.
- Errors should be normally distributed.
- Absence of Multicollinearity – It states that independent variables with high multicollinearity should be avoided.
- No influential Data points – This assumption states to avoid altering data points(outliers).

## II. DESCRIPTION OF THE DATASET

Information about the data set is as shown in Fig.1

Target Variable: Income.

Number of records: 45085.

Number of columns: 13.

The structure of the data frame, dataset description, and its descriptive statistics are shown in the below Fig. 1, 2, and 3.

```
> str(Incomedata)
'data.frame':   4508 obs. of  13 variables:
 $ i..age  : int  45 67 68 75 38 49 52 61 62 68 ...
 $ yrsed   : int  6 6 6 6 7 7 7 7 7 7 ...
 $ edcat   : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ yrsempl : int  4 15 7 35 8 4 21 27 5 7 ...
 $ income  : int  17 12 9 16 37 21 44 15 32 31 ...
 $ creddebt: num  0.372 0.376 0.201 0.314 0.143 ...
 $ othdebt : num  1.294 0.392 0.789 0.758 0.412 ...
 $ default : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ jobsat  : Factor w/ 5 levels "1","2","3","4",..: 4 3 5 4 3 1 3 4 5 4 ...
 $ homeown : Factor w/ 2 levels "0","1": 2 2 1 1 1 2 2 2 2 2 ...
 $ address : int  22 28 21 11 11 14 11 35 29 18 ...
 $ cars    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ carvalue: num  9.1 5.9 5.8 5.8 22.1 10.8 19.8 4.9 14.6 13.6 ...
```

Fig. 1. Detail of dataset.

| Variable | Description |
|---|---|
| age | Age in years |
| yrsed | Years of education |
| edcat | Level of education<br>1=Did not complete high school, 2=High school degree, 3=Some college, 4=College degree, 5=Postgraduate degree |
| yrsempl | Years with current employer |
| creddebt | Credit card debt in thousands |
| othdebt | Other debt in thousands |
| default | Ever defaulted on a bank, loan 0=no, 1=yes |
| jobsat | Job satisfaction<br>1=Highly dissatisfied, 2=Somewhat dissatisfied, 3=Neutral, 4=Somewhat satisfied, 5=Highly satisfied |
| homeown | Home ownership, 0=rent, 1=own |
| address | Years at current address |
| cars | Number of cars owned/leased |
| carvalue | Value of primary vehicle |

Fig. 2. Dataset description.

```
> summary(Incomedata)
     i..age          yrsed       edcat       yrsempl
 Min.   :18.00   Min.   : 6.00   1: 859   Min.   : 0.000
 1st Qu.:32.00   1st Qu.:12.00   2:1418   1st Qu.: 2.000
 Median :46.00   Median :14.00   3: 917   Median : 7.000
 Mean   :46.93   Mean   :14.53   4: 991   Mean   : 9.719
 3rd Qu.:62.00   3rd Qu.:17.00   5: 323   3rd Qu.:15.000
 Max.   :79.00   Max.   :23.00            Max.   :52.000
     income          creddebt          othdebt        default
 Min.   :  9.00   Min.   :  0.0000   Min.   :  0.0000   0:3431
 1st Qu.:  24.00   1st Qu.:  0.3879   1st Qu.:  0.9828   1:1077
 Median :  38.00   Median :  0.9318   Median :  2.0816
 Mean   :  55.41   Mean   :  1.8979   Mean   :  3.6915
 3rd Qu.:  68.00   3rd Qu.:  2.0765   3rd Qu.:  4.4351
 Max.   :1073.00   Max.   :109.0726   Max.   :141.4591
 jobsat  homeown     address          cars          carvalue
 1:872   0:1675   Min.   : 0.00   Min.   :1.000   Min.   : 2.20
 2:936   1:2833   1st Qu.: 6.00   1st Qu.:2.000   1st Qu.:11.30
 3:987            Median :14.00   Median :2.000   Median :18.90
 4:907            Mean   :16.37   Mean   :2.367   Mean   :26.08
 5:806            3rd Qu.:25.00   3rd Qu.:3.000   3rd Qu.:34.00
                  Max.   :57.00   Max.   :8.000   Max.   :99.60
```

Fig. 3. Descriptive statistics of data.

## III. DATA VISUALIZATION

Before creating the model, visual analysis is required. Fig. 4 shows the code for scatterplots, histograms, and correlations between variables, and the visual depiction of the same can be seen in Fig. 5 [3].

```
install.packages("psych")
library(psych)
attach(Incomedata)
pairs.panels (Incomedata,
              method = "pearson", # correlation method
              hist.col = "#00AFBB",
              density = TRUE,  # show density plots
              ellipses = TRUE) # show correlation ellipses
```

Fig. 4. Code for plots.

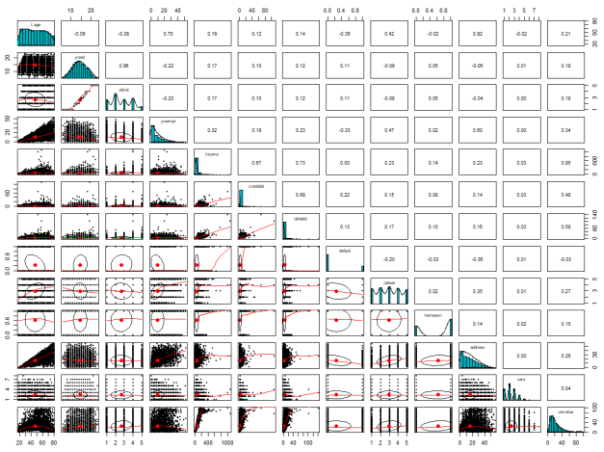

Fig. 5. Scatter-plot matrix

Correlation, scatter plot is plotted to check the relation between the variables, The value of the correlation varies from -1 to 1, where 1 is a positive correlation, -1 is a negative correlation and 0 represents no correlation.

If there is a linear relationship between the independent variables and the dependent variable, we need to use those variables in the modeling to forecast the response variable. These plots will also assist us in determining whether there is a significant association between the independent variables to foresee multicollinearity issues. In addition, the histogram plots indicate on which variable the transformation should be applied to meet the Gauss Markov Assumptions.

## IV. MODELS BUILDING PROCESS AND DESCRIPTION

Look at the correlation and scatter plot in Fig. 5. Creddebt, othdebt, and carvalue all have a stronger than 0.5 correlation with income. The "default" showed significance, according to spearman's correlation(used manly for categorical data variables), hence it is included in the model one evaluation.

### *A. Model 1*

Following are the Independent variables that are used to construct this model. creddebt, othdebt, carvalue and default, for predicting Income

```
model1<- lm (income ~ carvalue * creddebt * othdebt + carvalue
+                + creddebt + othdebt + default, data=Incomedata)
summary(model1)
```

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  6.5444287  0.6861272   9.538  < 2e-16 ***
carvalue                     1.6899538  0.0229057  73.779  < 2e-16 ***
creddebt                    -0.8465166  0.3666414  -2.309  0.02100 *
othdebt                     -0.6380884  0.2142252  -2.979  0.00291 **
default1                    -3.2862066  0.7651027  -4.295 1.78e-05 ***
carvalue:creddebt            0.0367833  0.0051251   7.177 8.29e-13 ***
carvalue:othdebt             0.0334603  0.0031919  10.483  < 2e-16 ***
creddebt:othdebt             0.0853904  0.0127254   6.710 2.18e-11 ***
carvalue:creddebt:othdebt -0.0005288  0.0001472  -3.593  0.00033 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.64 on 4499 degrees of freedom
Multiple R-squared:  0.8668,    Adjusted R-squared:  0.8666
F-statistic:  3661 on 8 and 4499 DF,  p-value: < 2.2e-16
```

Fig. 6. Summary of Model 1.

In this model, R squared value is found to be 0.866 and P-value is < 0.05, Overall the model is looking to be significant and each independent variable is contributing to the prediction of income and also carvalue*creddebt*othdebt

added in the code to check the interaction effect. Even though the model looks significant but it is failing due to the presence of the non-linearity and Heteroscedasticity this can be seen in the below visualization plots Fig. 7 and Fig. 8 respectively. Also, the NCV test has been performed to verify the Constance Variance assumption the P-value found to be significant which tells us the assumption has not been met.

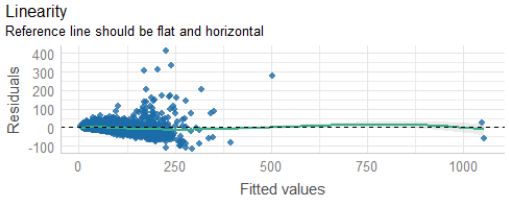


Fig. 7. Residuals vs Fitted values(Linearity).

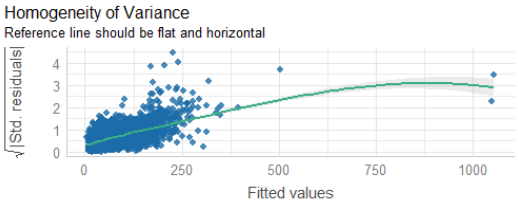


Fig. 8. Sd residuals vs Fitted values(Constance of variance).

```
> ncvTest(model102)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 16168, Df = 1, p = < 2.22e-16
```

Fig. 9. ncvTest run code.

## B. *Model 2*

This model is performed to correct the linearity assumption and rectify the heteroscedasticity this can be done by transforming the dependent and the independent variables by applying square root or log or square to the Y and X value respectively(dependent and predictor), in Fig. 5 histogram plot it can be seen that the income and carvalue are left-skewed log is used to transform dependent variable income and independent variable carvalue.

```
Model2 <- lm(log(income) ~ carvalue * creddebt * othdebt + log(carvalue)
+                  + creddebt + othdebt + default, data=Incomedata)
summary(model2)
```

```
Call:
lm(formula = log(income) ~ carvalue * creddebt * othdebt + log(carvalue) +
    creddebt + othdebt + default, data = Incomedata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55015 -0.13376 -0.01545  0.11411  1.11039

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.261e+00  2.331e-02  54.099  < 2e-16 ***
carvalue                  5.356e-03  5.291e-04  10.125  < 2e-16 ***
creddebt                  2.592e-02  3.636e-03   7.128 1.18e-12 ***
othdebt                   2.248e-02  2.227e-03  10.095  < 2e-16 ***
log(carvalue)             7.484e-01  1.207e-02  61.998  < 2e-16 ***
default1                 -4.968e-02  7.410e-03  -6.704 2.27e-11 ***
carvalue:creddebt        -1.444e-04  5.125e-05  -2.818  0.00486 **
carvalue:othdebt         -1.675e-04  3.405e-05  -4.921 8.94e-07 ***
creddebt:othdebt         -4.113e-04  1.257e-04  -3.272  0.00107 **
carvalue:creddebt:othdebt  4.560e-06  1.472e-06   3.097  0.00197 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1998 on 4498 degrees of freedom
Multiple R-squared:  0.9299,    Adjusted R-squared:  0.9298
F-statistic:  6634 on 9 and 4498 DF,  p-value: < 2.2e-16
```

Fig. 10. Summary of Model 2.

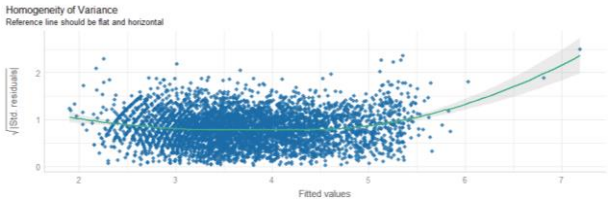


Fig. 11. Sd residuals vs Fitted values(Constance of variance).

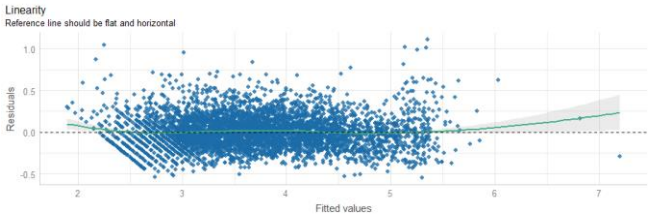


Fig. 12. Residuals vs Fitted values(Linearity)

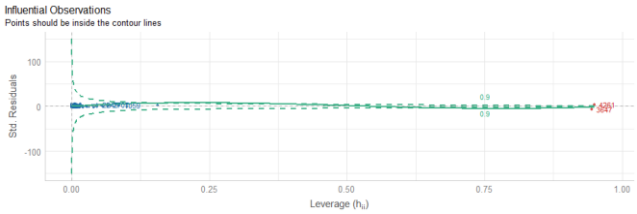


Fig. 13. Sd residuals vs Leverage(Outliers plot)

In the Model 2 summary, it can be observed that (Fig. 10) there is improvement in the adjusted r squared value 0.9298, a significant P value less than 0.05 and a fair scatter points (Fig. 11,12) in the linearity and Constance of variance plot further when we try to test the NCV test Model is found to be significant p = 0.011 which tells us that we have not met the constant variance assumption, To reverify this model, the variance inflation factor (VIF) is evaluated, and the value is found to be greater than 10 in a majority of the predictors, indicating the presence of high multicollinearity among them. Along with this Std. Residuals vs Leverage plotted Fig. 13 to check the Presence of Outliers.

## C. *Model 3 to 7*

Despite the fact that models 1 and 2 have greater R2 values, a high r-squared does not always imply a good regression model. We need to verify other parameters as well, so the outliers discovered in the plot Fig. 13 are eliminated from the data frame before we build the model. Multicollinearity observed in Model 2 directs us If an independent variable's VIF value is more than 5, the variable must be dropped.

```
Model 2 <- lm(log(income) ~ carvalue * creddebt * othdebt + log(carvalue)
+                  + creddebt + othdebt + default, data = Incomedata)
summary(model2)
model3<-update(model2,~.-carvalue:creddebt:othdebt)
summary (model3)
model4<-update(model3,~.-creddebt:othdebt)
summary (model4)
model5<-update(model4,~.-carvalue:creddebt)
summary (model5)
model6<-update(model5,~.-carvalue)
summary (model6)
model7<-update(model6,~.-othdebt:carvalue)
summary (model7)
```

Fig. 14. Summary of Model 3 to 7.

```
Call:
lm(formula = log(income) ~ carvalue + creddebt + othdebt + log(carvalue) +
    default + carvalue:creddebt + carvalue:othdebt + creddebt:othdebt,
    data = Incomedata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.57832 -0.13304 -0.01519  0.11579  1.10278

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.250e+00  2.304e-02  54.246  < 2e-16 ***
carvalue            4.866e-03  5.053e-04   9.630  < 2e-16 ***
creddebt            2.263e-02  3.481e-03   6.502 8.81e-11 ***
othdebt             1.985e-02  2.060e-03   9.633  < 2e-16 ***
log(carvalue)       7.579e-01  1.169e-02  64.820  < 2e-16 ***
default1           -4.873e-02  7.411e-03  -6.575 5.41e-11 ***
carvalue:creddebt -1.127e-04  5.027e-05  -2.242    0.025 *
carvalue:othdebt  -1.193e-04  3.031e-05  -3.937 8.38e-05 ***
creddebt:othdebt  -2.958e-05  2.470e-05  -1.197    0.231
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2 on 4499 degrees of freedom
Multiple R-squared:  0.9298,    Adjusted R-squared:  0.9297
F-statistic:  7448 on 8 and 4499 DF,  p-value: < 2.2e-16
```

Fig. 15. Summary of Model 3 to 7 (Continued).

When the variable is removed, it is evident that the majority of the predictors become statistically significant. The methods below show how this is done.

Step 1: Observed highest VIF score among the variable was 'carvalue:creddebt:othdebt' dropped from model 2,

Step 2: In the given model(Fig. 15), we can see that there are a lot of significant variables and a few that aren't. We start by eliminating the factors that are not significant.

Step 3: We follow the same step 2 procedure for further to arrive at model 7, in all the subsequent steps NCV test, independence of errors(Durbin-Watson test), VIF test and Cooks' distance (check for influential data points) has been performed to arrive at final model 7 also four columns that had outliers are removed from the data frame.

```
Call:
lm(formula = log(income) ~ creddebt + othdebt + log(carvalue) +
    default, data = Incomedata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.62855 -0.13661 -0.01697  0.11552  1.15236

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.0724884  0.0138619  77.369  < 2e-16 ***
creddebt       0.0179538  0.0013957  12.864  < 2e-16 ***
othdebt        0.0139316  0.0008695  16.022  < 2e-16 ***
log(carvalue)  0.8611666  0.0049740 173.134  < 2e-16 ***
default1      -0.0534599  0.0074077  -7.217 6.22e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.202 on 4497 degrees of freedom
Multiple R-squared:  0.9276,    Adjusted R-squared:  0.9275
F-statistic: 1.44e+04 on 4 and 4497 DF,  p-value: < 2.2e-16
```

Fig. 16. Summary of Model 7.

In comparison to the previous models, Model 7 has the lowest residual standard error of 0.2. R squared value of 0.9276, Adjusted R squared value of 0.9275, and a significant P-value of less than 0.05.

## V. DIAGNOSTICS CARRIED OUT TO ENSURE THAT THE GAUSS MARKOV ASSUMPTION HAS BEEN MET

Model 7 is determined to be superior to the preceding models, hence it is used to verify Gauss Markov assumptions using plots.

### *A. Linearity:*

There should be a linear relationship between the dependent and independent variables, this can be checked by scatter plot Residuals vs Fitted values in Fig. 17 it is evident that all the points are scattered across close to flat reference line. Proves the Linearity assumption.

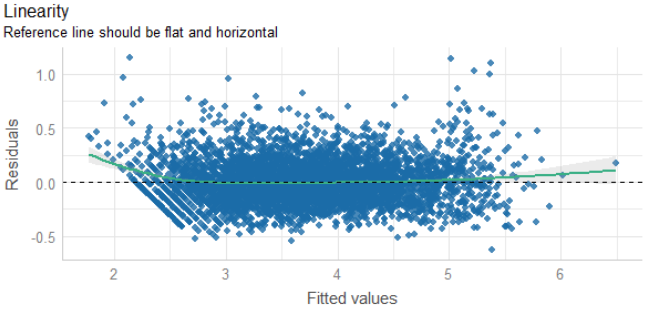


Fig. 17. Residuals vs Fitted values plot

### *B. Homoscedasticity:*

Upon plotting the Sq. root residual vs Fitted values scatter plot Fig18 There is no evidence of heteroscedasticity in the form of a Fan-out and Fan-in pattern, to verify the same ncv Test has been performed (Fig. 19) the P-value = 0.234 Found to be insignificant thus the Constance variance assumption has been met.

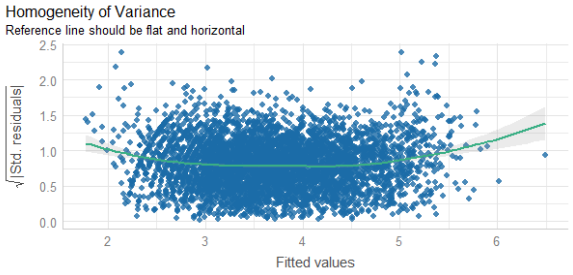


Fig. 18. Sqrt. Residuals vs Fitted values plot

```
> ncvTest(model7)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.413379, Df = 1, p = 0.2345
```

Fig. 19. ncv Test.

### *C. No autocorrelation between errors*

To check for non-autocorrelation between errors Durbin–Watson statistic test is performed, Its value should lie between 1 and 3 which determine whether or not the assumption of independent errors is valid.

```
> durbinWatsonTest(model7)
 lag Autocorrelation D-W Statistic p-value
   1      0.03384351       1.932099   0.024
 Alternative hypothesis: rho != 0
```

Fig. 20. Durbin–Watson Test.

For model 7 D-W (Fig. 20) statistic value is 1.93 which is close to 2 which is considered to be the better value.

### *D. Errors are normally distributed*

It can be observed in Fig. 21 that errors are regularly distributed and near to the normal curve. As a result, the normalcy assumption is met.

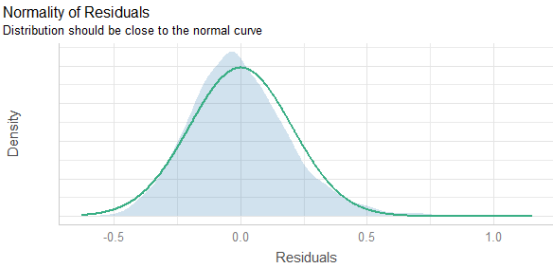


Fig. 21. Normality plot.

### *E. Testing for Multicollinearity*

To check the presence of multicollinearity among the variables, the variance inflation factor(VIF) test is performed if the predictor variables have more than 5 they are likely to be collinear, in Fig. 22 all the predictor variables for model 7 were found to be less than 5.

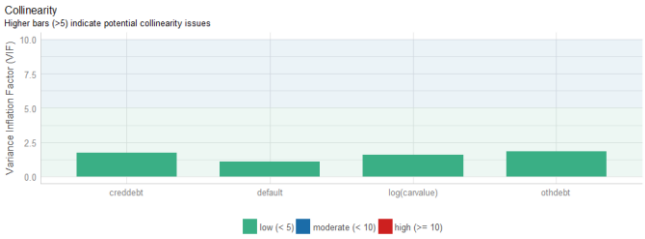


Fig. 22. VIF plot.

### *F. No influential data points*

In order to check if there are any influential data points the Std. Residuals vs Leverage plot is plotted (Fig. 23), the plot shows no outliers presence out of the contour lines which clears the assumption of no influence data point presence.

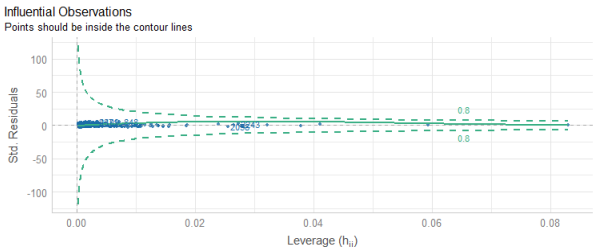


Fig. 23. plot.

## VI. SUMMARY OF THE FINAL MODEL

The main steps that are followed in this paper are

1. Analyze the model using the correlation matrix and scatter plot; the most important thing to remember here is to choose the right variables to begin building the model; variables that are highly correlated with the dependent variable are more likely to be considered for modeling, rather than variables that are correlated among themselves.
2. Gauss-assumption Markov's assumptions are utilized to verify each model, and diagnostics plots are employed to provide a clear image of the analysis.

After evaluating the correlation matrix and scatter plot, four independent variables creddebt, othdebt, carvalue, and default are used to forecast the income in model 1, Despite having a good r squared value of 0.866, this model was unable to create linearity and homoscedasticity.

The variables income and carvalues are transformed using log transformation to correct the linearity and homoscedasticity in model1. Although the linearity and homogeneity of variance plots showed a significant result, the model failed the ncv and variance inflation factor(VIF) tests due to the presence of multi-collinearity among the independent variables

Influential Columns 3647 and 4261, which were detected in the Std residual versus leverage plot, are removed from the data frame before Model 3 is built. carvalue:creddebt:othdebt variables with a higher VIF value of 77.92 are discarded due to the presence of multicollinearity. The creddebt:othdebt variable becomes insignificant in the following model, and the processes are repeated until model 7, plots and assumptions are validated in each subsequent phase.

Furthermore, model 7 which has independent variables creddept, otherdebt, log(carvalue), and default1 passes all of the gauss assumptions, indicating that our regression model is the Best Linear Unbiased Estimator (BLUE), model 7 was found to have a P-value of less than 0.05, indicating that it is significant. R squared and adjusted R-squared was found to be better than the earlier models, with values of 0.9276 and 0.9275, respectively. This means that the features 'creddept', 'default1', 'log(carvalue)', and 'othdebt' account for 92.7 percent of the variance in 'Income.'

## REFERENCES

[1] A. Z. Ul-Saufie, A. S. Yahya, en N. A. Ramli, “Improving multiple linear regression model using principal component analysis for predicting PM10 concentration in Seberang Prai, Pulau Pinang”, International Journal of Environmental Sciences, vol 2, no 2, bll 403–410, 2011.

[2] D. A. Lind, W. G. Marchal, S. A. Wathen, and R. D. Mason, *Statistical Techniques in Business & Economics*. Boston: McGraw-Hill Irwin, 2005.

[3] W. Chang, “R graphics cookbook, 2nd edition,” *5.13 Making a Scatter Plot Matrix*, 24-Mar-2022. [Online]. Available: https://r-graphics.org/recipe-scatter-splom. [Accessed: 25-Mar-2022].