# Income Prediction Using Multiple Linear Regression

Abhinav Thapa
Student ID: 20259409
Statistics for Data Analytics, MSc in Data Analytics
National College of Ireland, Dublin
Email: x20259409@student.ncirl.ie

***Abstract*— This report will determine the Income of a particular individual using a Multiple Linear Regression Model employing suitable feature selection and Regression diagnostics on a raw dataset that provides the financial profile for over four thousand individuals.**

***Keywords—Multiple Linear Regression, Dataset, Feature Selection, Regression Diagnostics, Dummy variables, Pairplots, Correlation, Predictors, Response, Factor. Tools used- R,SPSS.***

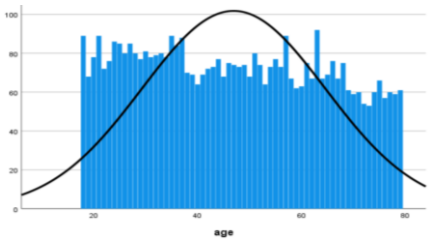
## I. INTRODUCTION

The dataset in question, IncomeData.csv, contains 9 continuous variables and 4 categorical variables. It provides a detailed information about individual demographics, education, property ownership, job-satisfaction and financial data for four-thousand five-hundred and eight people. Continuous information is captured in variables with names: age, yrsed, yrsempl, creddebt, othdebt, address, cars, carvalue and income, while categorical data is captured in variables: edcat, default, jobsat and homeown. Our goal for this project is to utilize the given predictors and devise a generalized linear model using regression technique to predict the income (in thousands of Euros) of an individual based on the given variables with high accuracy.

## II. DESCRIIPTION OF VARIABLES

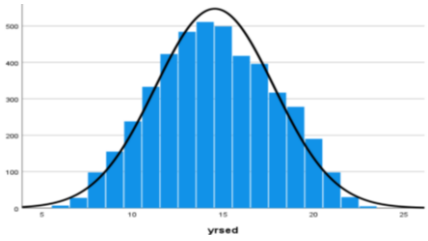
### A. *age*

The variable represents the continuous data for age of each person in years and is represented as an integer in R. It will be considered as one of the predictors of income (response) for model building.



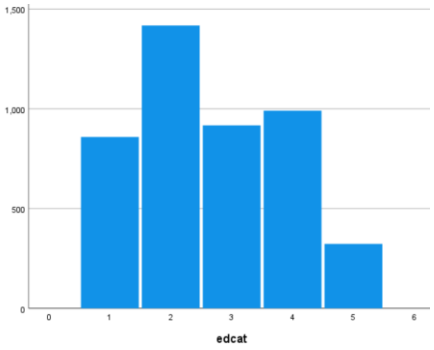

### B. *yrsed*

It provides the years of education that a person has recevied in years and is represented as an integer in R. It will be one of the predictors of income as well.



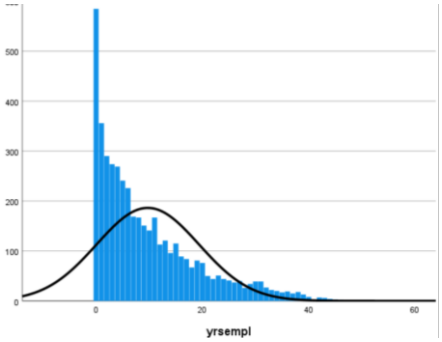

### C. *edcat*

It provides information about the highest level of education one has received. It will be among the predictors of income and is represented as a factor with 5 levels : 1 - Did not complete high school, 2 – High school degree, 3 – Some college, 4 – College degree, 5 – Postgraduate degree.



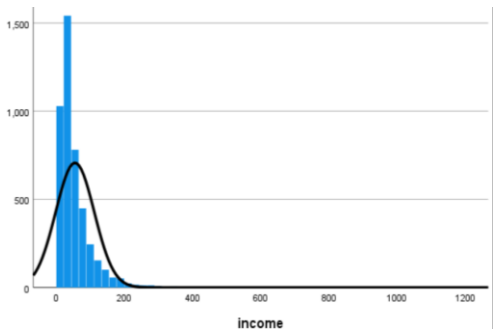

### D. *yrsempl*

It provides the number of years a person has spent working for his/her current employer in years and is represented as an integer in R.



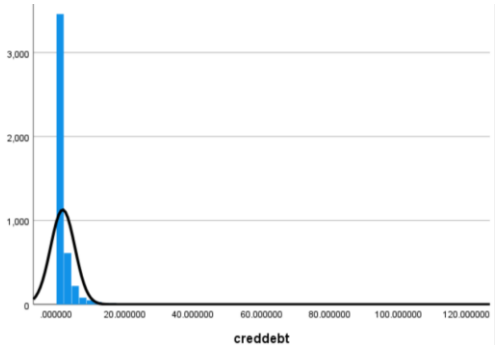

### E. *income*

It is the **response** variable for our analysis and is measured in thousands of euros. It's represented as an integer in R.



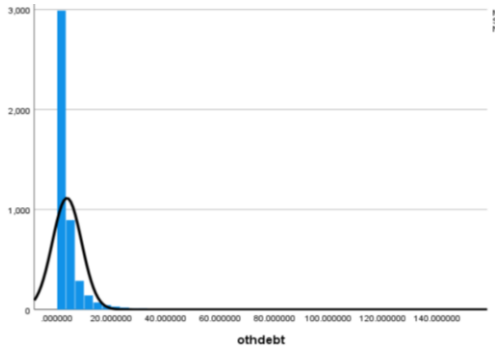
F. *creddebt*

It's one of the predictor variables and gives the amount of credit card debt that a person owes in thousands of euros. It is represented as a continuous variable in R and it's type is numeric.



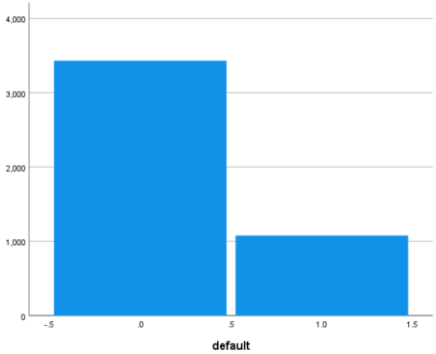

G. *othdebt*

It represents the total debt except credit card in thousands of euros and is identified as a numeric type data in R.



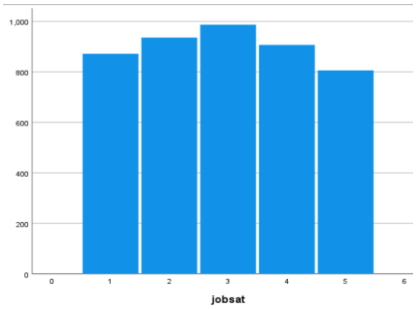

H. *default*

This variable shows whether the person has ever defaulted on any bank loan before. It is represented as a factor with 2 levels in R: '0' – No and '1' – Yes.



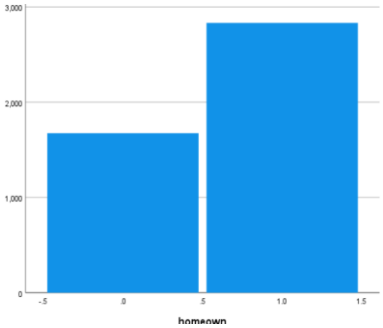

I. *jobsat*

It gives the job-satisfaction level of an individual mentioned under 5 categories. Given as a factor in R with 5 levels: 1 – Highly dissatisfied, 2 – Somewhat dissatisfied, 3 – Neutral, 4 – Somewhat satisfied, 5 – Highly satisfied.



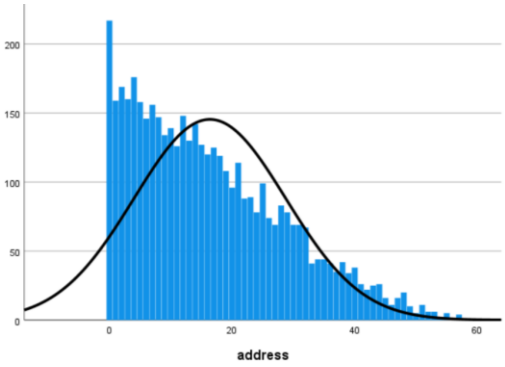

J. *homeown*

It tells us whether the person owns the house he/she lives in or rents it. It's given as a factor with 2 levels in R: '0' – Rents, '1' – Owns.



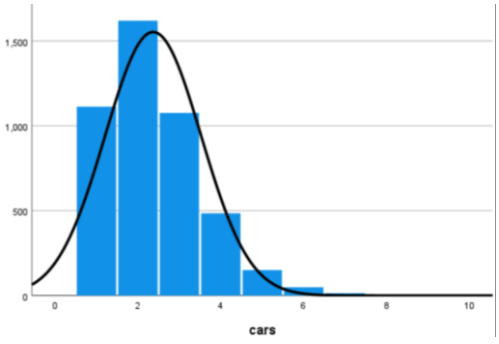

K. *address*

It tells us the number of years the person has lived in his current house.



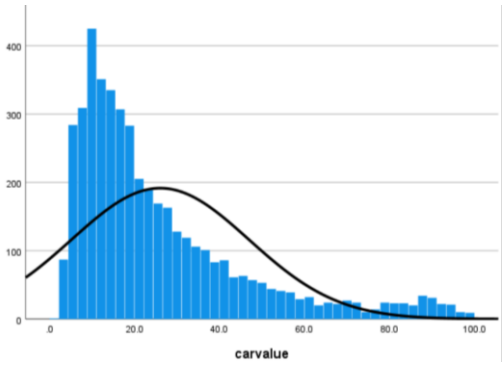

L. *cars*

It shows the number of cars a person has leased/owns in numbers. It is an integer variable that will help as predictor variable for income.



*M. Carvalue*

It gives the value of the primary car in thousands of euros for each record. It's represented as a continuous variable in R.




## III. EXPLORATORY DATA ANALYSIS

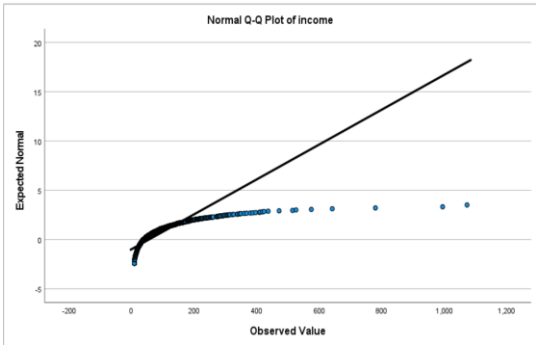
Data understanding is the first step of model building in regression analysis. Once we get a fair idea about the relevant features and their practical influence on income, our concerned response variable, we will build on an approach to tackle the issues that the dataset may possess. Summary of data is captured using structure - str() and summary - summary() functions in R in-built packages (refer Figure 1).

```
> str(findata)
'data.frame':   4508 obs. of  13 variables:
 $ 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 ...
> summary(findata)
      age            yrsed       edcat       yrsempl          income          creddebt
 Min.   :18.00   Min.   : 6.00   1: 859   Min.   : 0.000   Min.   :   9.00   Min.   : 0.0000
 1st Qu.:32.00   1st Qu.:12.00   2:1418   1st Qu.: 2.000   1st Qu.:  24.00   1st Qu.: 0.3879
 Median :46.00   Median :14.00   3: 917   Median : 7.000   Median :  38.00   Median : 0.9318
 Mean   :46.93   Mean   :14.53   4: 991   Mean   : 9.719   Mean   :  55.41   Mean   : 1.8979
 3rd Qu.:62.00   3rd Qu.:17.00   5: 323   3rd Qu.:15.000   3rd Qu.:  68.00   3rd Qu.: 2.0765
 Max.   :79.00   Max.   :23.00            Max.   :52.000   Max.   :1073.00   Max.   :109.0726
    othdebt         default  jobsat  homeown     address          cars         carvalue
 Min.   :  0.0000   0:3431   1:872   0:1675   Min.   : 0.00   Min.   :1.000   Min.   :  2.20
 1st Qu.:  0.9828   1:1077   2:936   1:2833   1st Qu.: 6.00   1st Qu.:2.000   1st Qu.: 11.30
 Median :  2.0816            3:987            Median :14.00   Median :2.000   Median : 18.90
 Mean   :  3.6915            4:907            Mean   :16.37   Mean   :2.367   Mean   : 26.08
 3rd Qu.:  4.4351            5:806            3rd Qu.:25.00   3rd Qu.:3.000   3rd Qu.: 34.00
 Max.   :141.4591                             Max.   :57.00   Max.   :8.000   Max.   : 99.60
```

*Figure 1: Descriptive Statistics*

From the initial descriptive statistics in **R** and **SPSS**, we can make the following observations:

i. *income* is not normal in nature and will need to be transformed for model building.




ii. *yrsed* and *cars* are fairly normally distributed about the mean.

iii. *yrsempl, income, creddebt, othdebt, address,* and *carvalue* are positively skewed about their corresponding mean due to presence of extreme outliers and will need some transformations while *age* is generally limited between a range.

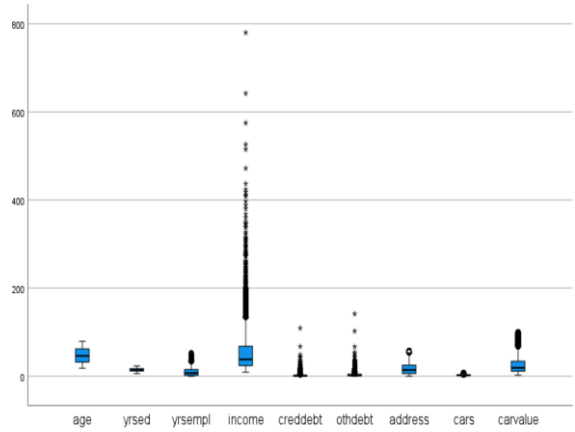



*Figure 2: Boxplots for each numerical variable*

iv. *edcat* has lowest count of category '5' records, *default* has lower '1' *category* records, *jobsat* has almost equally distributed levels, and *homeown* has fairly lower '0' category records. Here note that *edcat* and *jobsat* are factors with more than 2 levels and they will need to be transformed to dummy variables for model building.

We can also determine the correlation among the numerical variables by using the corelation function and its corresponding plots using pairs() function in R (Figure 3 and 4).

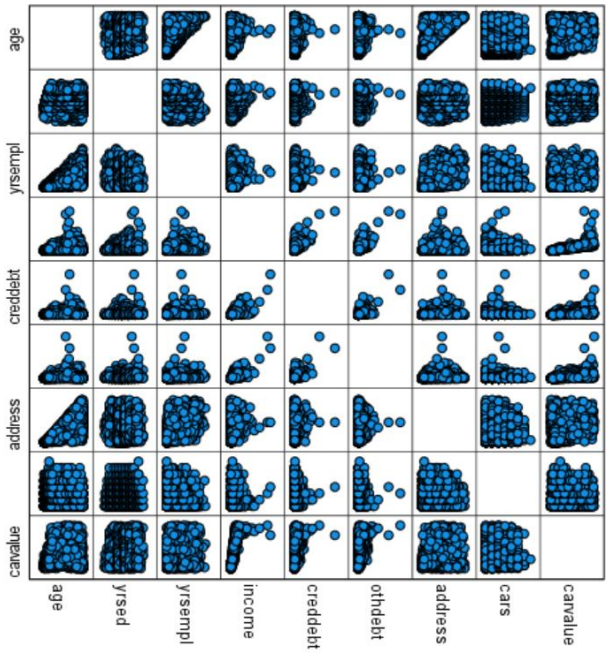



*Figure 3: Pearson correlation among numerical variables in the dataset in SPSS.*

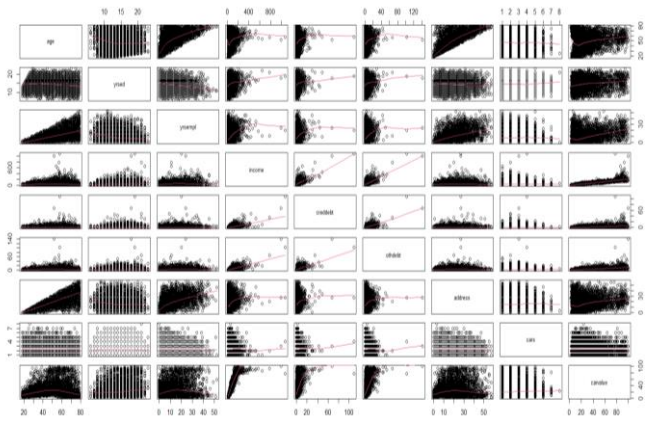


*Figure 4: Correlation in R*

Careful evaluation of the pairplots provides us with the following observations:

i. *age, yrsed, yrsempl* and *address* are not linearly related to *income.*

ii. *creddebt, othdebt, cars and carvalue* may be approximately linearly related with *income:* dependant variable, post transformations.

iii. **Highly correlated independent variables***: age* is highly correlated with *yrsempl* and *address; yrsempl* is highly correlated with *age, address* and *carvalue; creddebt* with *othdebt* and *carvalue; othdebt* is with *creddebt* and *carvalue; address* with *age, yrsempl* and *carvalue;* and *carvalue* with *yrsempl, creddebt, othdebt* and *address* (any correlation above 0.25 is taken as high).

Now that we have a fair understanding of the dataset in question, we can conclude that we are dealing with a multi-collinear features and skewed data that will need preparation prior to model building. We observed in R that using 'log' transformation on *income* corrects its problem of non-normality. Hence, we will be using ***log(income)*** as our response variable for model building and generate the corresponding dummy variables for *edcat* and *jobsat* (factors with more than 2 levels). Let's proceed to the data transformation, model building and diagnostic step.

## IV. Model Building and Diagnostics

Model building step is an iterative process in which the final suitable model is achieved by capturing the effects of each independent (or its transformations) on the response variable and eliminating insignificant features using statistical tests while satisfying each of the assumptions (linearity between predictors and response, homoscedasticity, no autocorrelation among errors, no omitted variable bias, normally distributed errors, normality of errors, no multicollinearity among predictors and no influential data points which may be addressed in no particular order) of Multiple Linear Regression technique.

### *A. Iteration 1 (Base model)*

We will start by regressing **all the independent variables against *income*** (response) and see how does it fair in terms of the metrics and initial diagnostics. This will be our base model. Figure 5 shows the results obtained on the 1st iteration of our model building process in R(studio). **RSE is 0.3086 while the R2 score is 83.26 %**.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.9497825  0.1159379  25.443  < 2e-16 ***
age         -0.0031916  0.0005415  -5.894 4.03e-09 ***
yrsed        0.0061950  0.0054773   1.131 0.258102
yrsempl     -0.0026907  0.0007485  -3.595 0.000328 ***
creddebt     0.0102879  0.0018452   5.575 2.62e-08 ***
othdebt      0.0091025  0.0012760   7.134 1.13e-12 ***
default1    -0.0795847  0.0124999  -6.367 2.12e-10 ***
homeown1     0.0172391  0.0099700   1.729 0.083860 .
address      0.0017812  0.0006838   2.605 0.009219 **
cars         0.0087030  0.0039813   2.186 0.028867 *
carvalue     0.0304174  0.0002977 102.175  < 2e-16 ***
edcat_1     -0.0075116  0.0616781  -0.122 0.903074
edcat_2     -0.0188436  0.0450349  -0.418 0.675657
edcat_3      0.0133733  0.0341961   0.391 0.695759
edcat_4     -0.0018321  0.0245324  -0.075 0.940472
jobsat_1    -0.1116249  0.0171286  -6.517 7.97e-11 ***
jobsat_2    -0.0520983  0.0158403  -3.289 0.001013 **
jobsat_3    -0.0259141  0.0151870  -1.706 0.088016 .
jobsat_4    -0.0112234  0.0151246  -0.742 0.458088
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3086 on 4489 degrees of freedom
Multiple R-squared:  0.8333,    Adjusted R-squared:  0.8326
F-statistic:  1246 on 18 and 4489 DF,  p-value: < 2.2e-16
```

*Figure 5: Base model Metrics*

We also observed that *edcat* is highly insignificant while *yrsed* and *homown* are only slightly insignificant for predicting *income* and hence *edcat* will be removed first from the model.

Figure 6 gives the diagnostic plots for our base model that clearly indicates (from Residuals vs Fitted; plot at- top right) that the **assumption of linearity between dependent and independent variables is violated**. The model is also **heteroscedastic** in nature (non- constant variance), fact that is supported by a significant ncvTest. Variables are also highly **auto-correlated** (significant Durbin Watson Test), and *yrsed* and *edcat* are **multi-collinear**.

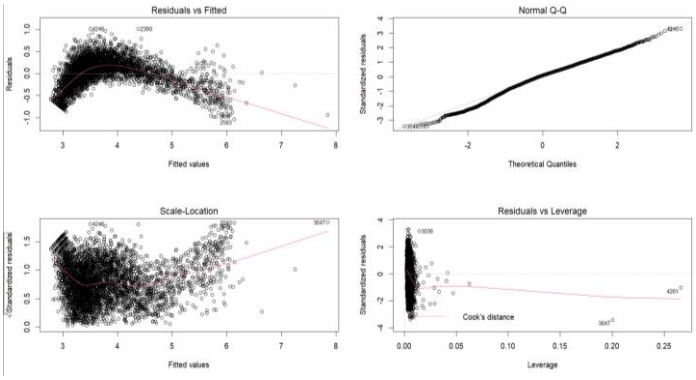



*Figure 6: Base model Diagnostic plots.*

In conclusion, we need to start by addressing the non-linearity issue between predictors and response first and build a model by removing *edcat* simultaneously.

**Steps for satisfying the assumption of Linearity between predictors and response for Regression**:

We will be handling this issue by applying transformations (log/root/squared) on the concerned predictors prior to next iteration of model building.

As observed earlier in EDA from figures 3 and 4, we found that *age, yrsed, yrsempl* and *address* were not linearly related to *income* which could be contributing to the violation of this assumption. We will address this by applying log and squared transformations on the mentioned predictors and replace them in the model. We can expect that it may also fix the subsequent

problems of possible heteroscedasticity and Influential outlier as observed in diagnostic plot and the rest of the Regression assumptions. Nevertheless, we will be analysing each of them statistically to achieve a **BLUE** (Best Linear Unbiased Estimator) model.

### *B. Iteration 2 (Model with transformed Independent variables)*

We replaced each of the non-linearly related predictors with their corresponding transformations as shown in the below figure 7.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.042e+00  1.949e-02  53.497  < 2e-16 ***
creddebt          1.331e-02  1.194e-03  11.147  < 2e-16 ***
othdebt           1.209e-02  8.126e-04  14.874  < 2e-16 ***
default1         -3.905e-02  8.166e-03  -4.783 1.79e-06 ***
homeown1          3.403e-03  6.441e-03   0.528   0.5973
cars              1.571e-03  2.587e-03   0.607   0.5437
jobsat_1         -9.123e-03  1.117e-02  -0.817   0.4143
jobsat_2         -1.907e-02  1.029e-02  -1.854   0.0638 .
jobsat_3         -1.055e-02  9.864e-03  -1.070   0.2849
jobsat_4         -1.213e-02  9.827e-03  -1.235   0.2170
I(age^2)         -2.847e-05  3.598e-06  -7.913 3.13e-15 ***
I(yrsed^2)        2.261e-04  3.311e-05   6.829 9.66e-12 ***
I(yrsempl)        3.312e-03  4.890e-04   6.774 1.41e-11 ***
I(address^0.5)    1.877e-02  3.336e-03   5.626 1.95e-08 ***
I(log(carvalue))  8.498e-01  5.152e-03 164.955  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2005 on 4493 degrees of freedom
Multiple R-squared:  0.9295,    Adjusted R-squared:  0.9293
F-statistic:  4234 on 14 and 4493 DF,  p-value: < 2.2e-16
```

*Figure 7: Model with transformed Predictors*

Here we can see that the **RSE** has reduced to **0.2005** and **R2 score** increased to **92.93%** which means that the model performance and fit has improved. However, the results of the test of non-constant variance, variance inflation factor and test for auto-correlation of errors indicate that there might be a little heteroscedasticity and minutely significant multi-collinearity between *age* and *address*, and *creditdebt* and *othdebt* (refer figure 8 and 9) but no auto-correlation among residuals was found.

Also, from Normal Q-Q plot in the diagnostics plot we can observe a little departure of residuals from normal curve which might be due to a missing variable or extreme outliers (Influential outliers).

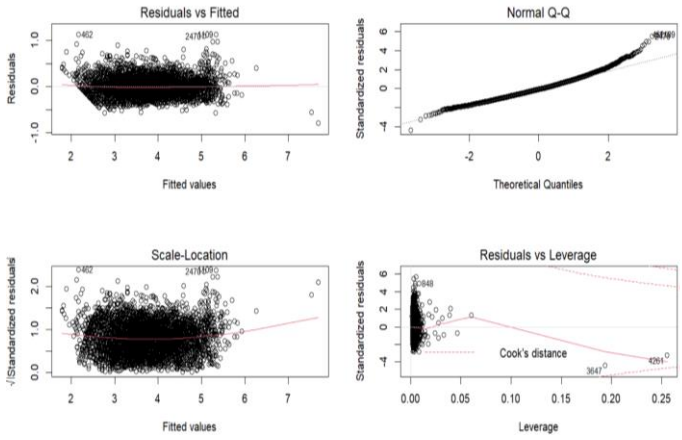



*Figure 8: Diagnostic plot for Iteration 2 model.*

```
> vif(mymodel_2)
      creddebt        othdebt        default        homeown           cars
      2.005809       2.141427       1.359504       1.086137       1.005759
      jobsat_1       jobsat_2       jobsat_3       jobsat_4      I(age^2)
      2.183712       1.951791       1.865471       1.740357       4.247443
  I(yrsed^2)     I(yrsempl)  I(address^0.5) I(log(carvalue))
      1.145235       2.496364       3.609292       1.729473
> durbinWatsonTest(mymodel_2)
 lag Autocorrelation D-W Statistic p-value
   1      0.02390656      1.952027   0.102
 Alternative hypothesis: rho != 0
> ncvTest(mymodel_2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 5.753037, Df = 1, p = 0.01646
```

*Figure 9: Results for vif(), durbinWatsonTest() and ncvTest()*

From the above results, we conclude that we need to introduce another variable to improve the homoscedasticity of the model.

**Steps for satisfying the assumption of Homoscedasticity for Regression**:

From a practical point of view, it is curious to think whether there is a combined effect of the *age* of a person and how long a person has lived at his current *address*, on our response variable *income*. Investigating it on the model may be worthwhile in the next iteration of model building.

### *C. Iteration 3 (Evaluating combined effect of age and address)*

We introduced an interaction term between *age* and *address* into our current model and following results were obtained.

```
Call:
lm(formula = log(income) ~ creddebt + othdebt + default + homeown +
    cars + jobsat_1 + jobsat_2 + jobsat_3 + jobsat_4 + I(age^2) +
    I(yrsed^2) + I(yrsempl) + I(address^0.5) + I(log(carvalue)) +
    I(age^2 + address), data = findata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.78767 -0.13472 -0.01599  0.11759  1.13239

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.017e+00  2.087e-02  48.710  < 2e-16 ***
creddebt            1.333e-02  1.193e-03  11.176  < 2e-16 ***
othdebt             1.210e-02  8.117e-04  14.907  < 2e-16 ***
default1           -3.622e-02  8.198e-03  -4.417 1.02e-05 ***
homeown1            5.596e-03  6.465e-03   0.865 0.386818
cars                1.629e-03  2.584e-03   0.630 0.528583
jobsat_1           -4.356e-03  1.125e-02  -0.387 0.698516
jobsat_2           -1.768e-02  1.028e-02  -1.720 0.085535 .
jobsat_3           -1.060e-02  9.852e-03  -1.076 0.282182
jobsat_4           -1.277e-02  9.817e-03  -1.301 0.193458
I(age^2)            3.150e-03  9.311e-04   3.383 0.000723 ***
I(yrsed^2)          2.188e-04  3.314e-05   6.602 4.52e-11 ***
I(yrsempl)          3.399e-03  4.890e-04   6.951 4.15e-12 ***
I(address^0.5)      3.992e-02  7.035e-03   5.674 1.48e-08 ***
I(log(carvalue))    8.473e-01  5.201e-03 162.910  < 2e-16 ***
I(age^2 + address) -3.176e-03  9.305e-04  -3.413 0.000647 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2003 on 4492 degrees of freedom
Multiple R-squared:  0.9297,    Adjusted R-squared:  0.9295
F-statistic:  3962 on 15 and 4492 DF,  p-value: < 2.2e-16
```

*Figure 10: Metrics for Model 3*

The model's overall performance improved a little in an attempt to achieve better homoscedasticity. Figure 10 illustrates the influence of the combined effect variable in the model as highly significant and this will be incorporated in the model. This issue will be treated further later on.

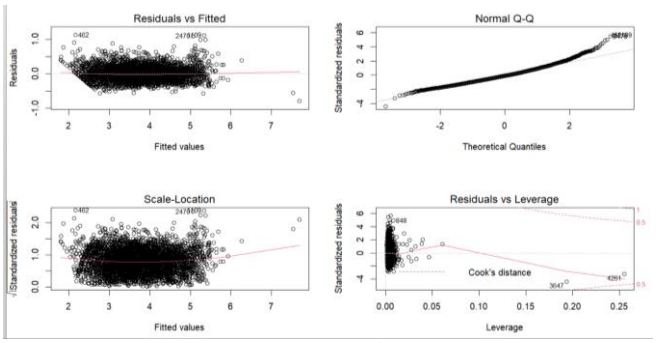



*Figure 11: Diagnostics for Model 3*

### D. *Iteration 4 (Elimination of insignificant predictors)*

We have achieved a decent model with a high Accuracy (**RSE is very low**) and high fit (**high Adj. R2 score**), however, model building approach also advises to follow the **rule of parsimony**. This means that model simplicity also needs to be considered while model building. Due to this, we will be removing the non-significant predictors (*jobsat, cars* and *homeown*) that have little to no practical effect on *income*, one by one while analysing the corresponding scores.

```
Call:
lm(formula = log(income) ~ creddebt + othdebt + default + I(age^2) +
    I(yrsed^2) + I(yrsempl) + I(address^0.5) + I(log(carvalue)) +
    I(age^2 + address), data = findata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.79456 -0.13480 -0.01654  0.11808  1.14317

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.013e+00  1.671e-02  60.633  < 2e-16 ***
creddebt            1.338e-02  1.192e-03  11.231  < 2e-16 ***
othdebt             1.210e-02  8.113e-04  14.912  < 2e-16 ***
default1           -3.605e-02  8.190e-03  -4.402 1.10e-05 ***
I(age^2)            2.965e-03  9.129e-04   3.247  0.00117 **
I(yrsed^2)          2.188e-04  3.307e-05   6.616 4.11e-11 ***
I(yrsempl)          3.508e-03  4.746e-04   7.391 1.73e-13 ***
I(address^0.5)      3.903e-02  6.888e-03   5.667 1.54e-08 ***
I(log(carvalue))    8.479e-01  5.162e-03 164.265  < 2e-16 ***
I(age^2 + address) -2.992e-03  9.124e-04  -3.279  0.00105 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2003 on 4498 degrees of freedom
Multiple R-squared:  0.9296,    Adjusted R-squared:  0.9295
F-statistic:  6603 on 9 and 4498 DF,  p-value: < 2.2e-16
```

*Figure 12: Model 4 metrics*

Even though there was no valuable difference in the performance metrics of the model, simplistic model is preferred over a complex one because only the presence significant predictors will provide a more generalized and accurate model.

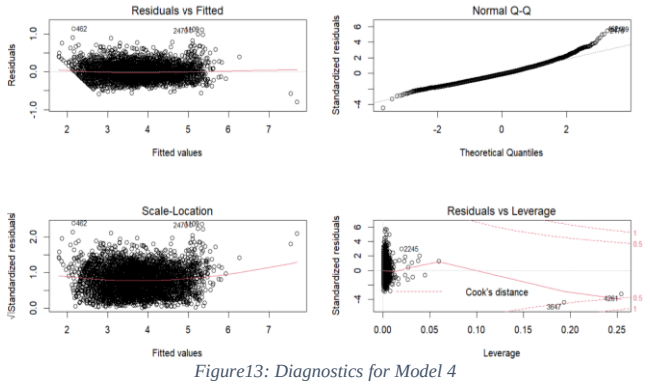



*Figure13: Diagnostics for Model 4*

```
> vif(mymodel_4)
       creddebt          othdebt          default          I(age^2)
   2.002590e+00     2.139834e+00     1.371120e+00     2.741728e+05
     I(yrsed^2)       I(yrsempl)   I(address^0.5) I(log(carvalue))
   1.145679e+00     2.357909e+00     1.542569e+01     1.740536e+00
I(age^2 + address)
   2.770693e+05
> durbinWatsonTest(mymodel_4)
 lag Autocorrelation D-W Statistic p-value
   1       0.0253461      1.949125   0.084
 Alternative hypothesis: rho != 0
> ncvTest(mymodel_4)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 6.9563, Df = 1, p = 0.0083525
```

*Figure 14: Tests for Multi-collinearity, Auto-correlation and Homoscedasticity*

Finally, we will be removing redundant effects of any predictors, if any, in the model. Figures 13 and 14 show that the model is now stable and reliable now that the assumptions of Regression have been addressed with the exception of a little departure from linearity of errors in Q-Q plot which is acceptable to a certain extent.

### E. *Iteration 5 (treating all remaining assumptions for model optimisation)*

In this step, we removed the final redundancy (even though significant per model) that may lead to an ineffective model i.e. *age ^2* term and the *address^0.5* term. So, the final model is illustrated below with plots in Figure 15.

```
Call:
lm(formula = log(income) ~ creddebt + othdebt + default + I(yrsed^2) +
    I(yrsempl) + I(log(carvalue)) + I(age^2 + address), data = findata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.79259 -0.13361 -0.01645  0.11651  1.15019

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.041e+00  1.553e-02  67.016  < 2e-16 ***
creddebt            1.354e-02  1.197e-03  11.308  < 2e-16 ***
othdebt             1.212e-02  8.152e-04  14.869  < 2e-16 ***
default1           -4.769e-02  8.048e-03  -5.926 3.35e-09 ***
I(yrsed^2)          2.313e-04  3.316e-05   6.974 3.53e-12 ***
I(yrsempl)          3.385e-03  4.766e-04   7.102 1.42e-12 ***
I(log(carvalue))    8.600e-01  4.871e-03 176.532  < 2e-16 ***
I(age^2 + address) -1.422e-05  2.546e-06  -5.587 2.45e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2012 on 4500 degrees of freedom
Multiple R-squared:  0.9289,    Adjusted R-squared:  0.9288
F-statistic:  8401 on 7 and 4500 DF,  p-value: < 2.2e-16
```

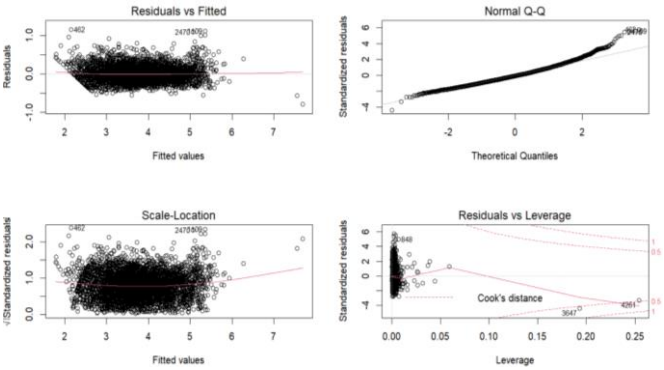



*Figure 15: Final Model metrics and diagnostic plots.*

The final model is also free of multi-collinearity (as all below 5 variance inflation factors), any auto-correlated error terms (as durbin Watson test is insignificant), influential outliers (cooks distance less than 1) and heteroscedasticity (as ncvTest is insignificant) as mentioned below.

```
> vif(mymodel_5)
       creddebt          othdebt          default          I(yrsed^2)
       2.001416         2.139576         1.310966           1.140868
     I(yrsempl)   I(log(carvalue)) I(age^2 + address)
       2.354224         1.535091         2.136304
> durbinWatsonTest(mymodel_5)
 lag Autocorrelation D-W Statistic p-value
   1       0.0254892      1.948852   0.068
 Alternative hypothesis: rho != 0
> ncvTest(mymodel_5)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.527278, Df = 1, p = 0.060367
> influencePlot(mymodel_5)
      StudRes         Hat      CookD
462   5.743778 0.002749812 0.01129088
1109  5.539593 0.003828706 0.01464630
3647 -4.392990 0.192899277 0.57421029
4261 -3.320058 0.254062849 0.46824597
```

## V. Summary of Final Model

Our final model from Iteration 5, exhibits high accuracy of 0.2 residual errors and utilizes almost all of the reliable predictors (Adj. R2 score of 92.8%) for Income prediction from the given dataset satisfying all of the assumptions of the regression technique to become a best linear unbiased estimator model. Further analysis and model optimization can be done using cross-validation techniques which will explored in detail at a later stage.

## References


[1] An Introduction to Statistical Learning (second edition), https://www.statlearning.com/resources-second-edition