Multiple Regression Analysis Report on Income Prediction and Understanding of Income Factors

Viplav Vijay Gadewar
Statistics for Data Analytics
Master of Science in Data Analytics
National College of Ireland
Dublin, Ireland
x21164274@student.ncirl.ie

***Abstract*—A variety of things influence the amount of money you earn. In order to better understand the multiple factors that influence income, this study used a Multiple Regression Model. This model has a large number of independent variables as well as one dependent variable, which is income. A dataset was subjected to a multiple regression approach using Python and R.**

***Index Terms*—Multiple Linear Regression, Assumptions, Ordinary Least Squares, Descriptive Statistics**

I. Introduction

The primary determinant of a person's level of life and financial condition is their income. This study aims to provide useful insights that may be utilized as the foundation for many better decisions made by the nation's or industries' administrators, taking into consideration the importance and influence on defining growth in their respective sectors. Age, years of education, years of employment, credit debt, other debts, bank default status, automobile, car value, job happiness, and years at the current residence are just a few of the numerous characteristics that define a person's income. This study used modules to cope with a vast dataset with only a few properties that are relevant to us.

To test alternative combinations and carefully pick the primary critical traits, the Ordinary Least Squares estimator is utilized. The attributes are not only extracted, but also normalized, and the results are compared.
Only when all of the following assumptions are met can a Multiple Linear Regression Test be performed:

- Linearity is the connection between X (Independent Variables) and Y (Dependent Variables).
- Multicollinearity should be avoided at all costs.
- Homoscedasticity means that the residual variance is the same for all X values.
- Residuals are dispersed regularly.
- Errors do not have an auto-correlation.

Multiple Regression : It is used to determine the relationship between two or more independent variables and one dependent variable.
Equation:

$$y = mx1 + mx2 + .. + mxn + b$$

Where,
y = Dependent variable
m = Slope
x1 = 1 st Independent variable
x2 = 2 nd Independent variable
x3 = 3 rd Independent variable
b = Constant

R-Squared Number : The distance between the data and the best-fit line. In multiple regression, it's also known as the proportion of determination or the coefficient of multiple determination.
Equation:

$$R^2 = 1 - \frac{SumofSquaredRegression}{TotalSumofSquares}$$

Adjusted R-Squared Number : is a modified R-squared that takes into account the number of predictors. The adjusted R-squared rises when the additional term improves the model more than would be expected by chance. A prediction is discarded if it weakens the model less than predicted.
Equation:

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(N-1)}{N-p-1}\right]$$

Where,
R2 = Sample R-squared
p = Number of Predictor
N = Total Size of Sample

II. DESCRIPTION OF DATA SET

The following are the characteristics of variables/columns from the "IncomeData.csv" dataset that are used to forecast a person's income:

1) **age:** Number of years an individual has lived.
2) **yrsed:** Years of Education of an individual.

3) **edcat:** Level of education of an individual
   - 1 = Did not finished high school
   - 2 = Completed high school degree
   - 3 = Passed some college
   - 4 = Pursued college degree
   - 5 = Completed post graduation
4) **yrsempl:** Years spend with the current employer of an individual.
5) **creddept:** Credit card loan in thousand Euros of an individual.
6) **othdept:** Other loan in thousand Euros of an individual.
7) **default:** If the bank has defaulted by an individual
   - 0 = no
   - 1 = yes
8) **jobstat:** Job Satisfaction of an individual
   - 1 = extremely unsatisfied
   - 2 = slightly dissatisfied
   - 3 = neutral
   - 4 = satisfied
   - 5 = extremely satisfied
9) **homeown:** House ownership of an individual
   - 0 = rent
   - 1 = own
10) **address:** Years spend at current address by an individual.
11) **cars:** Number of cars owned by an individual.
12) **carvalue:** Basic value to car owned by an individual in thousand Euros.

```
> summary(incomedata)
    ï..age          yrsed           edcat          yrsempl           income            creddebt
 Min.   :18.00   Min.   : 6.00   Min.   :1.000   Min.   : 0.000   Min.   :   9.00   Min.   :  0.0000
 1st Qu.:32.00   1st Qu.:12.00   1st Qu.:2.000   1st Qu.: 2.000   1st Qu.:  24.00   1st Qu.:  0.3879
 Median :46.00   Median :14.00   Median :2.000   Median : 7.000   Median :  38.00   Median :  0.9318
 Mean   :46.93   Mean   :14.53   Mean   :2.667   Mean   : 9.719   Mean   :  55.41   Mean   :  1.8979
 3rd Qu.:62.00   3rd Qu.:17.00   3rd Qu.:4.000   3rd Qu.:15.000   3rd Qu.:  68.00   3rd Qu.:  2.0765
 Max.   :79.00   Max.   :23.00   Max.   :5.000   Max.   :52.000   Max.   :1073.00   Max.   :109.0726
    othdebt            default           jobsat          homeown           address           cars          carvalue
 Min.   :  0.0000   Min.   :0.0000   Min.   :1.000   Min.   :0.0000   Min.   : 0.00   Min.   :1.000   Min.   : 2.20
 1st Qu.:  0.9828   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:0.0000   1st Qu.: 6.00   1st Qu.:2.000   1st Qu.:11.30
 Median :  2.0816   Median :0.0000   Median :3.000   Median :1.0000   Median :14.00   Median :2.000   Median :18.90
 Mean   :  3.6915   Mean   :0.2389   Mean   :2.964   Mean   :0.6284   Mean   :16.37   Mean   :2.367   Mean   :26.08
 3rd Qu.:  4.4351   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.:25.00   3rd Qu.:3.000   3rd Qu.:34.00
 Max.   :141.4591   Max.   :1.0000   Max.   :5.000   Max.   :1.0000   Max.   :57.00   Max.   :8.000   Max.   :99.60
```

Fig. 1. Data-set Summary

The Minimum, Maximum, Median, Mean, 1st Quratiles, 2nd Quratiles, and 3rd Quratiles for all the variables/columns in the data frame are shown in the diagram above.

III. DATA VISUALISATION

We must undertake data visual analysis, such as correlation between variables, scatter plots, and histograms, before commencing the model construction phases. Analyze the variables. Correlation explains the significance of a link between two variables.

Correlation values vary from -1 to 1, with -1 being the lowest and 1 being the highest.

- -1 indicates a negative connection (indirect correlation)
- 0 indicates that there is no connection.
- 1 indicates a positive connection (direct correlation)

Figure 2 depicts a portion of the Scatter plot as well as the relationship between variables.

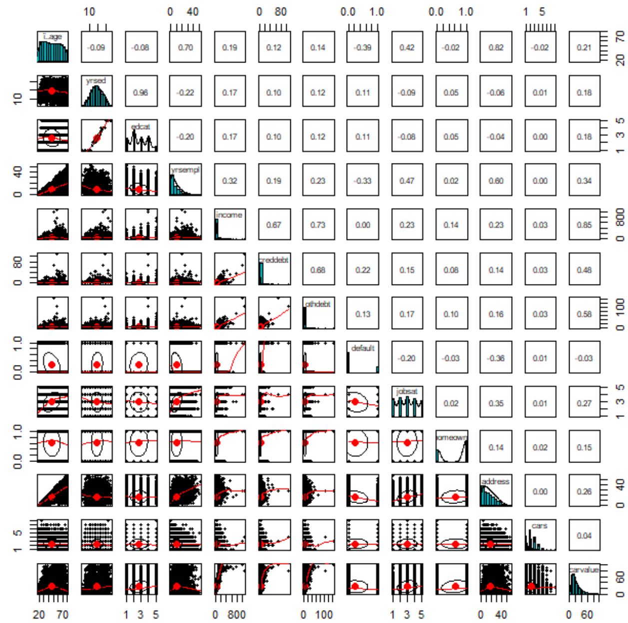

Fig. 2. Co-relation matrix histogram

	age	yrsed	yrsempl	income	creddebt	othdebt
age	1.000000	-0.094420	0.700842	0.193835	0.116871	0.139503
yrsed	-0.094420	1.000000	-0.217510	0.170060	0.104107	0.120735
yrsempl	0.700842	-0.217510	1.000000	0.322178	0.192129	0.225459
income	0.193835	0.170060	0.322178	1.000000	0.667599	0.732583
creddebt	0.116871	0.104107	0.192129	0.667599	1.000000	0.681329
othdebt	0.139503	0.120735	0.225459	0.732583	0.681329	1.000000
default	-0.392285	0.114260	-0.325454	0.004280	0.223253	0.130549
homeown	-0.019608	0.045031	0.018356	0.137967	0.080530	0.098668
address	0.821497	-0.057216	0.598004	0.234224	0.137001	0.164843
cars	-0.021859	0.008946	-0.004001	0.034918	0.032391	0.031832
carvalue	0.208119	0.177940	0.340973	0.848027	0.482614	0.584588

	default	homeown	address	cars	carvalue
age	-0.392285	-0.019608	0.821497	-0.021859	0.208119
yrsed	0.114260	0.045031	-0.057216	0.008946	0.177940
yrsempl	-0.325454	0.018356	0.598004	-0.004001	0.340973
income	0.004280	0.137967	0.234224	0.034918	0.848027
creddebt	0.223253	0.080530	0.137001	0.032391	0.482614
othdebt	0.130549	0.098668	0.164843	0.031832	0.584588
default	1.000000	-0.026729	-0.356768	0.014652	-0.033412
homeown	-0.026729	1.000000	0.137685	0.017819	0.152546
address	-0.356768	0.137685	1.000000	0.000301	0.257557
cars	0.014652	0.017819	0.000301	1.000000	0.043543
carvalue	-0.033412	0.152546	0.257557	0.043543	1.000000

Fig. 3. Descriptive view of Co-relation in variables

IV. MODEL BUILDING STEPS

The entire data set will be visually summarized using a box plot, which allows us to quickly assess data set dispersion, mean values, and skewness. The box plot below demonstrates that the variable income has multiple outliers. As a result, we will eliminate outliers from variable income in order to accurately see and predict the data.

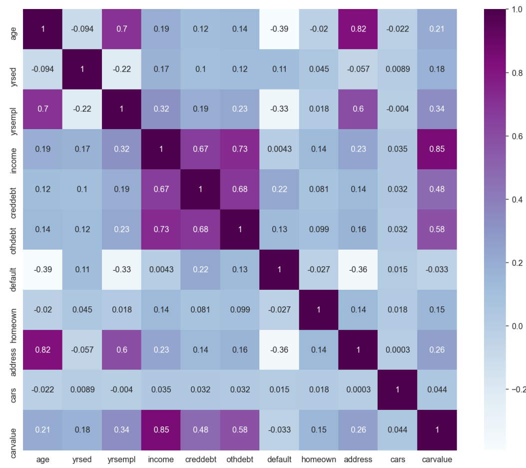

Fig. 4. Co-relation Heat-map

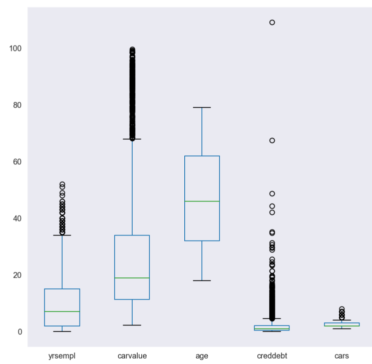

Fig. 5. Box plot before removing outliers

The outliers were eliminated using the Inter-Quartile Range (IQR) approach, which measures the difference between the third and first quartiles of the data. Because we had a number of outliers for the variable "income," we used the values of the 3rd and 1st Quartiles from Fig 1 (68.00 and 24.00, respectively) to construct the values displayed as dots in Fig 5.

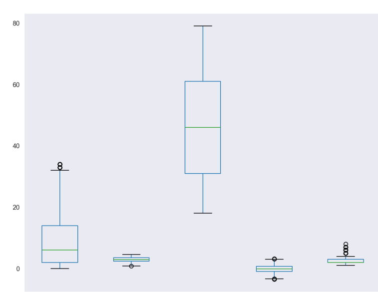

Fig. 6. Outliers have been removed from the box plot

We used Python's built-in function "hist" to present the individual frequency distributions of each variable after removing the outliers. On the other side, to better understand the relationship between the variables, we will generate a heatmap of the dataset. Below is a representation of histograms.

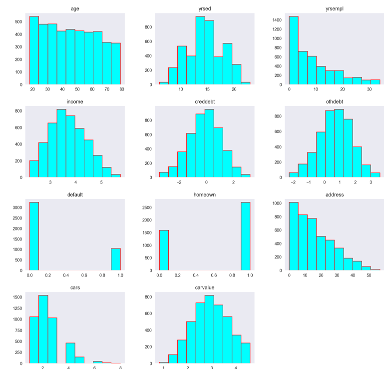

Fig. 7. Histogram after removing outliers

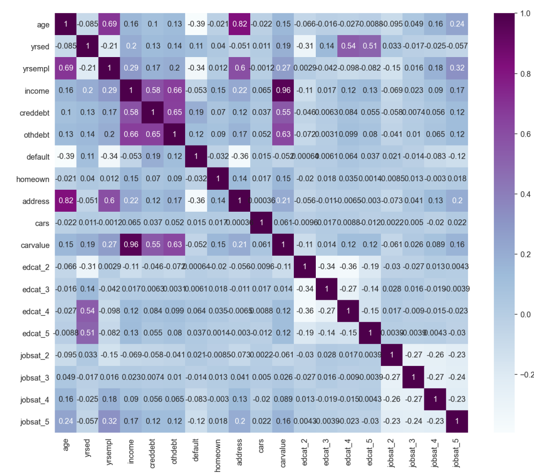

Fig. 8. heatmap after removing outliers

From the heat-map and histogram above, we can derive the following:

1) The majority of predictors in the first model had a p-value higher than 0.05, with `'edcat_2'` having the highest p-value of 0.955 . We remove the `'edcat_2'` variable and re-run OLS to obtain regression findings to enhance our model even further.

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.1126	0.045	24.756	0.000	1.025	1.201
age	-0.0015	0.000	-3.768	0.000	-0.002	-0.001
yrsed	0.0059	0.004	1.522	0.128	-0.002	0.013
yrsempl	0.0040	0.001	6.703	0.000	0.003	0.005
creddebt	0.0296	0.004	7.785	0.000	0.022	0.037
othdebt	0.0515	0.004	11.587	0.000	0.043	0.060
default	-0.0415	0.009	-4.753	0.000	-0.059	-0.024
homeown	0.0138	0.007	1.980	0.048	0.000	0.028
address	0.0007	0.000	1.377	0.169	-0.000	0.002
cars	0.0027	0.003	0.976	0.329	-0.003	0.008
carvalue	0.8342	0.006	134.938	0.000	0.822	0.846
edcat_2	0.0009	0.016	0.056	0.955	-0.030	0.031
edcat_3	0.0068	0.024	0.283	0.777	-0.040	0.054
edcat_4	0.0050	0.032	0.153	0.878	-0.059	0.069
edcat_5	0.0245	0.043	0.566	0.571	-0.060	0.109
jobsat_2	-0.0113	0.010	-1.092	0.275	-0.032	0.009
jobsat_3	-0.0078	0.011	-0.738	0.461	-0.028	0.013
jobsat_4	-0.0014	0.011	-0.130	0.896	-0.023	0.020
jobsat_5	0.0063	0.012	0.522	0.602	-0.017	0.030

Fig. 9. Model 1

2) Because the `'jobsat_4'` variable has the greatest p-value of 0.897 in our second model, we exclude it from the following model.

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.1112	0.037	30.109	0.000	1.039	1.184
age	-0.0015	0.000	-3.774	0.000	-0.002	-0.001
yrsed	0.0060	0.002	2.554	0.011	0.001	0.011
yrsempl	0.0040	0.001	6.705	0.000	0.003	0.005
creddebt	0.0296	0.004	7.787	0.000	0.022	0.037
othdebt	0.0515	0.004	11.588	0.000	0.043	0.060
default	-0.0415	0.009	-4.754	0.000	-0.059	-0.024
homeown	0.0138	0.007	1.981	0.048	0.000	0.028
address	0.0007	0.000	1.379	0.168	-0.000	0.002
cars	0.0027	0.003	0.976	0.329	-0.003	0.008
carvalue	0.8342	0.006	134.966	0.000	0.822	0.846
edcat_3	0.0056	0.012	0.468	0.640	-0.018	0.029
edcat_4	0.0034	0.016	0.208	0.835	-0.029	0.035
edcat_5	0.0225	0.024	0.938	0.348	-0.025	0.070
jobsat_2	-0.0113	0.010	-1.094	0.274	-0.032	0.009
jobsat_3	-0.0078	0.011	-0.739	0.460	-0.028	0.013
jobsat_4	-0.0014	0.011	-0.130	0.897	-0.023	0.020
jobsat_5	0.0063	0.012	0.523	0.601	-0.017	0.030

Fig. 10. Model 2

3) The third model is constructed similarly, but the `'edcat_4'` variable is removed owing to its 0.837 significance value.

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.1108	0.037	30.191	0.000	1.039	1.183
age	-0.0015	0.000	-3.806	0.000	-0.002	-0.001
yrsed	0.0061	0.002	2.559	0.011	0.001	0.011
yrsempl	0.0040	0.001	6.749	0.000	0.003	0.005
creddebt	0.0296	0.004	7.787	0.000	0.022	0.037
othdebt	0.0515	0.004	11.591	0.000	0.043	0.060
default	-0.0415	0.009	-4.754	0.000	-0.059	-0.024
homeown	0.0138	0.007	1.981	0.048	0.000	0.028
address	0.0007	0.000	1.381	0.167	-0.000	0.002
cars	0.0027	0.003	0.978	0.328	-0.003	0.008
carvalue	0.8341	0.006	135.533	0.000	0.822	0.846
edcat_3	0.0056	0.012	0.466	0.641	-0.018	0.029
edcat_4	0.0034	0.016	0.206	0.837	-0.029	0.035
edcat_5	0.0225	0.024	0.936	0.349	-0.025	0.070
jobsat_2	-0.0106	0.009	-1.213	0.225	-0.028	0.007
jobsat_3	-0.0070	0.009	-0.809	0.419	-0.024	0.010
jobsat_5	0.0072	0.010	0.728	0.466	-0.012	0.027

Fig. 11. Model 3

4) Furthermore, the `'edcat_3'` variable has been eliminated from our fourth because its significant value in the fourth was 0.648

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.1058	0.027	40.232	0.000	1.052	1.160
age	-0.0015	0.000	-3.801	0.000	-0.002	-0.001
yrsed	0.0065	0.001	5.233	0.000	0.004	0.009
yrsempl	0.0040	0.001	6.749	0.000	0.003	0.005
creddebt	0.0296	0.004	7.786	0.000	0.022	0.037
othdebt	0.0515	0.004	11.595	0.000	0.043	0.060
default	-0.0414	0.009	-4.750	0.000	-0.058	-0.024
homeown	0.0138	0.007	1.981	0.048	0.000	0.028
address	0.0007	0.000	1.380	0.168	-0.000	0.002
cars	0.0027	0.003	0.980	0.327	-0.003	0.008
carvalue	0.8342	0.006	135.699	0.000	0.822	0.846
edcat_3	0.0038	0.008	0.456	0.648	-0.013	0.020
edcat_5	0.0186	0.015	1.246	0.213	-0.011	0.048
jobsat_2	-0.0105	0.009	-1.209	0.227	-0.028	0.007
jobsat_3	-0.0070	0.009	-0.805	0.421	-0.024	0.010
jobsat_5	0.0072	0.010	0.729	0.466	-0.012	0.027

Fig. 12. Model 4

5) Finally, predictors like 'age', 'yrsed', 'yrsemp', 'creddebt', 'otherdebt', 'default', 'homeown', 'address', 'cars' and 'carvalue' are included in our 5th Model.

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.1045	0.027	40.415	0.000	1.051	1.158
age	-0.0015	0.000	-3.791	0.000	-0.002	-0.001
yrsed	0.0066	0.001	5.493	0.000	0.004	0.009
yrsempl	0.0040	0.001	6.743	0.000	0.003	0.005
creddebt	0.0296	0.004	7.785	0.000	0.022	0.037
othdebt	0.0515	0.004	11.591	0.000	0.043	0.060
default	-0.0414	0.009	-4.753	0.000	-0.059	-0.024
homeown	0.0139	0.007	1.987	0.047	0.000	0.028
address	0.0007	0.000	1.374	0.170	-0.000	0.002
cars	0.0027	0.003	0.986	0.324	-0.003	0.008
carvalue	0.8342	0.006	135.730	0.000	0.822	0.846
edcat_5	0.0169	0.014	1.170	0.242	-0.011	0.045
jobsat_2	-0.0104	0.009	-1.198	0.231	-0.028	0.007
jobsat_3	-0.0068	0.009	-0.790	0.430	-0.024	0.010
jobsat_5	0.0072	0.010	0.733	0.463	-0.012	0.027

Omnibus:	243.513	Durbin-Watson:	2.072
Prob(Omnibus):	0.000	Jarque-Bera (JB):	336.919
Skew:	0.580	Prob(JB):	6.90e-74
Kurtosis:	3.933	Cond. No.	484.

Fig. 13. Model 5

V. CHECKING MULTIPLE REGRESSION ASSUMPTIONS

1) To validate the lack of multicollinearity, utilize the Variance Inflation Factor (VIF) test. The VIF determines this if the independent variables are multicollinear. To demonstrate that there is no multicollinearity among independent variables, the VIF value between independent variables should be less than 5.

	feature	VIF
0	const	71.919433
1	age	4.210563
2	yrsed	1.485028
3	yrsempl	2.308473
4	creddebt	1.991175
5	othdebt	2.184618
6	default	1.346349
7	homeown	1.091582
8	address	3.490824
9	cars	1.005961
10	carvalue	1.994947
11	edcat_5	1.350623
12	jobsat_2	1.215861
13	jobsat_3	1.233084
14	jobsat_5	1.331171

Fig. 14. Variance Inflation Factor

2) The variance must be constant, which is the second most important condition in a regression model. Heteroscedascity refers to a lack of homoscedasticity, which suggests that standard errors are smaller than they should be. We can achieve a terrific result and meet the homoscedasticity condition if the data sample is randomly distributed about the line. This need is met by the logarithmic change of variables performed before, as seen in Fig. 15.

Fig. 15. Homoscedasticity satisfied by Residual Plots

3) The Durbin-Watson test can be used to see if the assumption of independent mistakes is correct. The value of the D-W statistics should be between 2 and 3, with no value less than 1 and no value more than 3. The following is the Durbin-Watson formula:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

Omnibus:	243.513	Durbin-Watson:	2.072
Prob(Omnibus):	0.000	Jarque-Bera (JB):	336.919
Skew:	0.580	Prob(JB):	6.90e-74
Kurtosis:	3.933	Cond. No.	484.

Fig. 16. Durbin Watson Results

4) The presence of errors with a normal distribution can be confirmed by the presence of residuals with a normal distribution (Fig. 17) They appear to be evenly scattered.

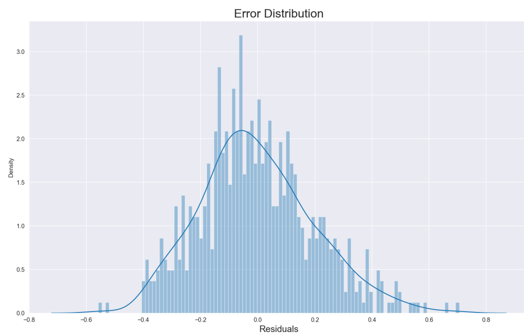

Fig. 17. Error distribution graph

VI. FINAL MODEL SUMMARY

Figure 18 shows the outcomes of a code snippet that can be regarded a decent prediction model. Our model agrees with 92.66% of the data points, according to the R Square score of 0.9266. The difference between actual and projected values is calculated using the Mean Absolute Error, and the closer the value comes to 0, the better our model is at forecasting. Our model's Mean Absolute Error is 0.1575, which is close to zero and acceptable. Our model's mean square error is 0.0395, which is acceptable.

```
from sklearn.metrics import mean_squared_error,r2_score,mean_absolute_error

mae = mean_absolute_error(Y_test,Y_pred)
mse = mean_squared_error(Y_test,Y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(Y_test,Y_pred)

print("Mean Absolute Error is :", mae)
print("Mean Squared Error is :", mse)
print("Root Mean Squared Error is :", rmse)
print("R2 score is :", r2)

Mean Absolute Error is : 0.1575480878642651
Mean Squared Error is : 0.03958977866219549
Root Mean Squared Error is : 0.19897180368634018
R2 score is : 0.9266947549538166
```

Fig. 18. Regression Results

The Predicted Values versus Actual Values for the dependent variable 'income' are shown in Figs. 19 and 20 below, along with a Q-Q plot for the final.

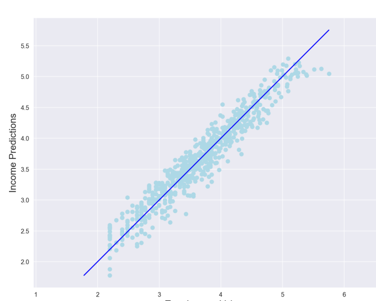

Fig. 19. Actual Values vs Predicated values

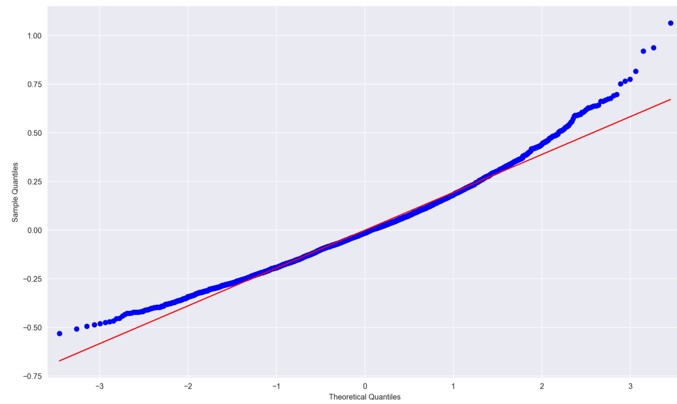

Fig. 20. Q-Q Plot

VII. CONCLUSION

Finally, a person's age, years of education, years of work, credit and other debt value, bank defaulter outcome, current automobile worth, and current property ownership are all factors that influence their income. Other components in the data set had no influence on our predictor variables or were ineffective factors.

VIII. REFERENCES

1) "A Statistical Approach to Adult Census Income Level Prediction" Navoneel Chakrabarty; Sanket Biswas 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) Year: 2018 — Conference Paper — Publisher: IEEE

2) "Prediction Model of Financial Income of Listed Companies Based on Grey Model" Li Tao 2020 IEEE International Conference on Industrial Application of Artificial Intelligence (IAAI) Year: 2020 — Conference Paper — Publisher: IEEE