20/11/2021

MULTIPLE LINEAR REGRESSION

Migrova, Maria

NCIRL – STATISTICS FOR DATA ANALYTICS

INTRODUCTION

PREDICTIVE MODELING

Predictive modeling, also called predictive analysis, is a mathematical operation that aims to forecast future events or outcomes by examining patterns that are most likely to forecast future results. [1] The main motivation is: What is possible to happen in the future based on past behaviour? Predictive modeling may be used for a variety of purposes. For example, we can look up the future weather forecast, customer's next purchase, or setting ticket prices by airlines. Modelling is divided into two stages: 1. Determining the family of model. 2. Fitting the model (identifying the best parameters of the model that best fits the data). [2]

REGRESSION

Regression is a statistical technique for determining the strength and character of a relationship between a single dependent variable and a set of independent variables. It is frequently used in banking, investment and other fields. [3]

SIMPLE LINEAR REGRESSION

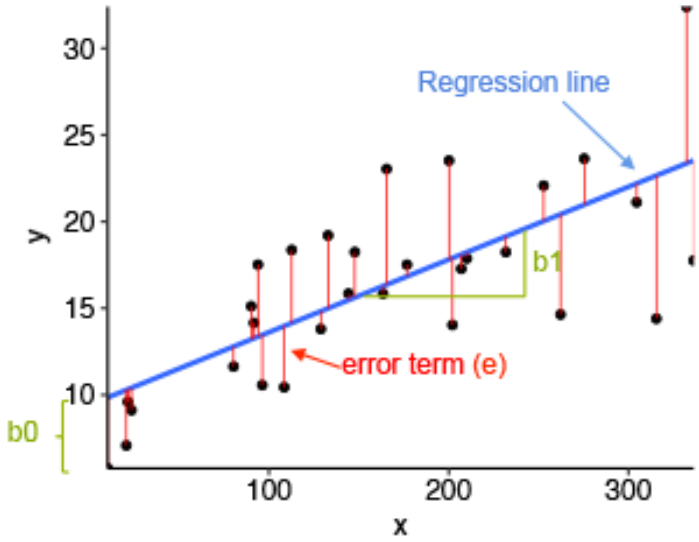
It is a statistical technique for modeling a response's dependence on a single explanatory variable. It is only useful if these teo objects have a straight-line relationship. The x-axis represents the explanatory variable, while the y-axis represents the response variable.

The following is a simple linear regression model:
Response variable = Intercept + slope * explanatory variable + random variability (random error)

$$y_i=\beta_0+\beta_1 x_i+\epsilon_i$$

Where the intercept is the expected mean value of y when x is equal to 0, and slope is the amount that the y value varies when the x value moves by 1 unit. We are looking for the best line equation to best describe the relationship between these two variables. The minimise the total of squared errors, the most widely used line of best fit is chosen (i.e. the squared distance of data points from the line). We're interested in the line with the smallest errors among the possible lines that can pass through the data cloud. [4] Using this equation, we may get the predicted values for y (labelled as y^)

$$\hat{y}=b_0+b_1 x$$



In this graph we can see an illustration of the linear regression model, with the blue line representing the line of best fit, b0 representing the intercept, and b1 representing the slope. The residuals are represented by the red lines.

Figure 1 Multiple Linear Regression

METHODOLOGY

DATA DESCRIPTION

For this project we are using a csv file called Credit_v2.csv. This dataset consists of 9 different variables of 687 observations. These variables are:

- ✓ age = Age in years (integer),
- ✓ ed = Level of education (integer),
- ✓ employ = Years with current employer (integer),
- ✓ address = Years at current address (integer),
- ✓ income = Household income in thousands (integer),
- ✓ debtinc = Debt to income ratio (x100) (double),
- ✓ othdebt = Other debt in thousands (double),
- ✓ default = Whether the customer has previously defaulted (double),
- ✓ creddebt = Amount of credit debt (integer),

```
      age              ed           employ         address          income          debtinc
 Min.   :20.00   Min.   :1.000   Min.   : 0.000   Min.   : 0.000   Min.   : 14.00   Min.   : 0.40
 1st Qu.:29.00   1st Qu.:1.000   1st Qu.: 3.000   1st Qu.: 3.000   1st Qu.: 24.00   1st Qu.: 5.00
 Median :34.00   Median :1.000   Median : 7.000   Median : 7.000   Median : 34.00   Median : 8.60
 Mean   :34.87   Mean   :1.731   Mean   : 8.362   Mean   : 8.285   Mean   : 45.46   Mean   :10.23
 3rd Qu.:40.00   3rd Qu.:2.000   3rd Qu.:12.000   3rd Qu.:12.000   3rd Qu.: 54.00   3rd Qu.:14.05
 Max.   :56.00   Max.   :5.000   Max.   :31.000   Max.   :34.000   Max.   :446.00   Max.   :41.30
    creddebt          othdebt          default
 Min.   : 0.0117   Min.   : 0.04558   Min.   :0.000
 1st Qu.: 0.3686   1st Qu.: 1.04306   1st Qu.:0.000
 Median : 0.8508   Median : 1.96171   Median :0.000
 Mean   : 1.5380   Mean   : 3.05196   Mean   :0.262
 3rd Qu.: 1.8877   3rd Qu.: 3.93045   3rd Qu.:1.000
 Max.   :20.5613   Max.   :27.03360   Max.   :1.000
Rows: 687
Columns: 9
$ age       <int> 52, 48, 36, 36, 43, 39, 41, 39, 47, 28, 29, 21, 25, 45, 43, 33, 26, 45, 30, 27, 25, 25, 26,~
$ ed        <int> 1, 1, 2, 2, 1, 1, 3, 1, 1, 1, 1, 2, 4, 2, 1, 2, 3, 1, 1, 3, 1, 1, 2, 2, 2, 2, 1, 1, 1, 2, 1~
$ employ    <int> 6, 22, 9, 13, 23, 6, 0, 22, 17, 3, 8, 1, 0, 9, 25, 12, 2, 3, 1, 2, 8, 8, 6, 10, 12, 1, 23, ~
$ address   <int> 9, 15, 6, 6, 19, 9, 21, 3, 21, 6, 6, 2, 2, 2, 26, 21, 8, 1, 15, 10, 7, 4, 1, 7, 4, 1, 8, 5, 2,~
$ income    <int> 29, 100, 49, 41, 72, 61, 26, 52, 43, 26, 27, 16, 32, 69, 64, 58, 37, 20, 22, 26, 27, 35, 45~
$ debtinc   <dbl> 16.3, 9.1, 8.6, 16.4, 7.6, 5.7, 1.7, 3.2, 5.6, 10.0, 9.8, 18.0, 17.6, 6.7, 16.7, 18.4, 14.2~
$ creddebt  <dbl> 1.715901, 3.703700, 0.817516, 2.918216, 1.181952, 0.563274, 0.099008, 1.154816, 0.587552, 0~
$ othdebt   <dbl> 3.011099, 5.396300, 3.396484, 3.805784, 4.290048, 2.913726, 0.342992, 0.509184, 1.820448, 2~
$ default   <int> 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0~
```

After summarising the data using summarise and glimpse functions, we can see the minimal, 1st quartile, median, mean, 3rd quartile and maximum values of all the variables. Most of the values seem to be ok, however there seems to be outliers for income and othdebt variables. We will examine that using graphical approach. We can also see that our dataset doesn't contain any NA values.

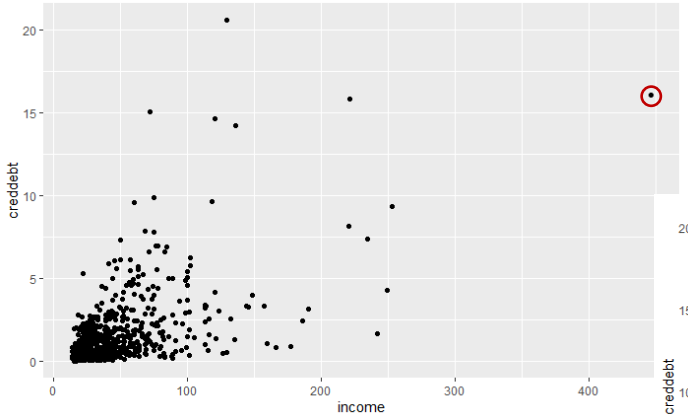


Figure 2 Income vs Creddebt

For income variable we can notice one outlier which is equal to 446. It is possible that this value is true, therefor we won't be removing it from the dataset.

For othdebt variable we can see two outliers. One is equal to 23 and the other one is equal to 23. These values are also very possible, therefor we will keep them in our dataset.

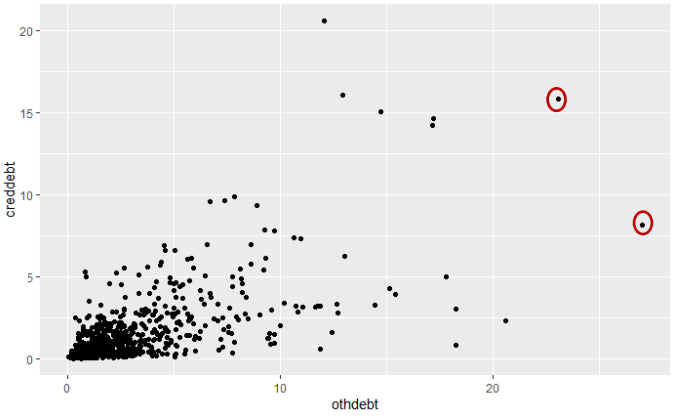
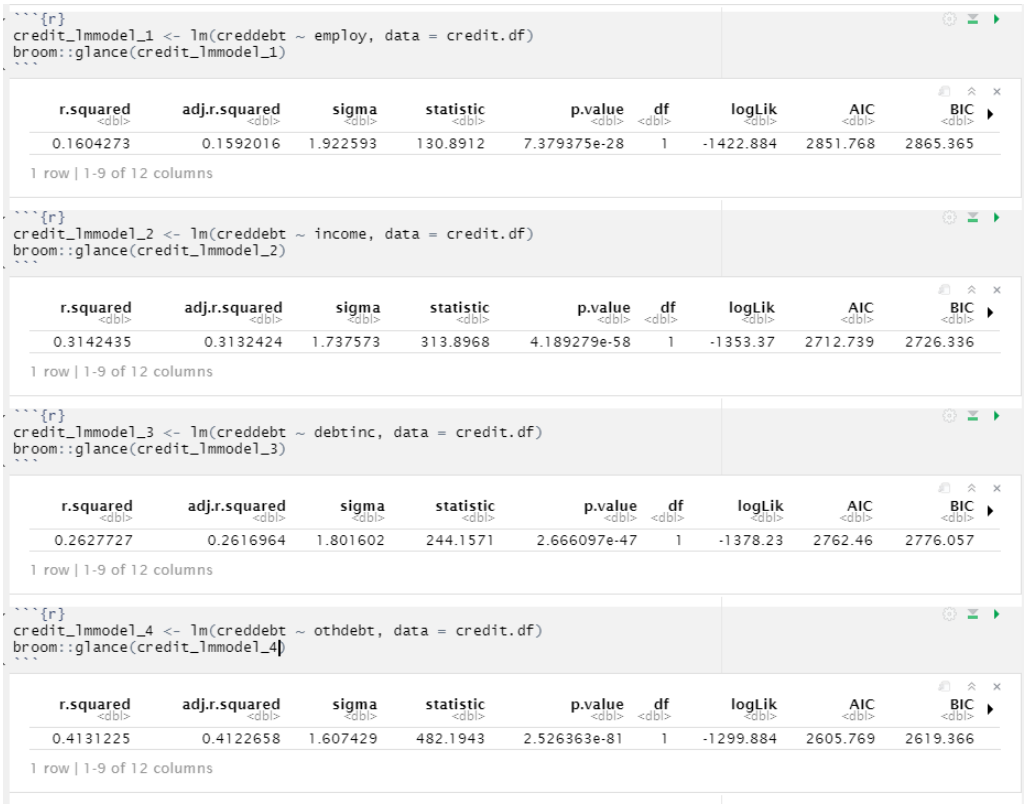


Figure 3 Othdebt vs Creddebt

BUILDING THE MODEL

In this project we will be building model to predict Creddebt by using other dependent variables.
We started our analysis by comparing a linear models for different variables. We're interested in the R-squared value. R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates **the percentage of the variance in the dependent variable that** the independent variables explain collectively. [5]

For our final we decided to choose this 4 variables: employ, income, debtinc and othdebt which R-squared value is the highest. Now it's time to find the best model by trying different combinations of these variables.



```r
credit_lmmodel_1 <- lm(creddebt ~ employ, data = credit.df)
broom::glance(credit_lmmodel_1)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC
0.1604273	0.1592016	1.922593	130.8912	7.379375e-28	1	-1422.884	2851.768	2865.365

1 row | 1-9 of 12 columns

```r
credit_lmmodel_2 <- lm(creddebt ~ income, data = credit.df)
broom::glance(credit_lmmodel_2)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC
0.3142435	0.3132424	1.737573	313.8968	4.189279e-58	1	-1353.37	2712.739	2726.336

1 row | 1-9 of 12 columns

```r
credit_lmmodel_3 <- lm(creddebt ~ debtinc, data = credit.df)
broom::glance(credit_lmmodel_3)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC
0.2627727	0.2616964	1.801602	244.1571	2.666097e-47	1	-1378.23	2762.46	2776.057

1 row | 1-9 of 12 columns

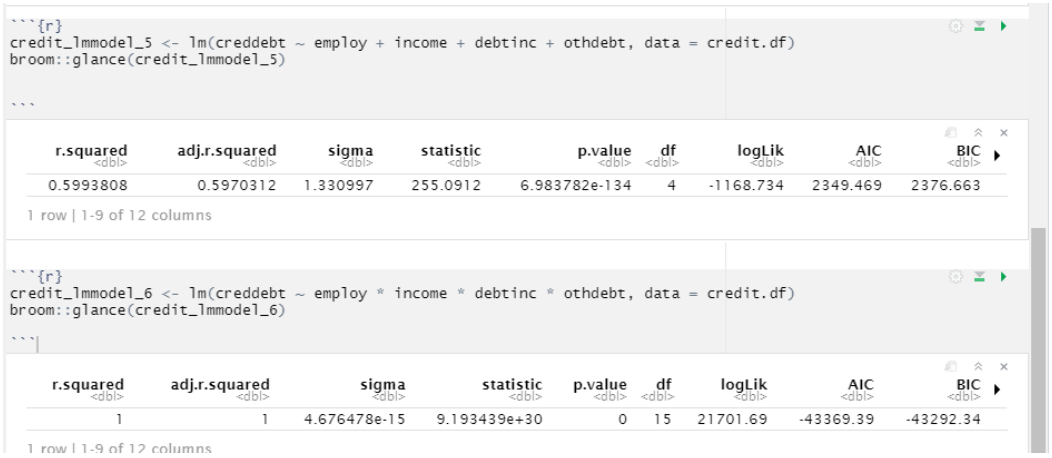
```r
credit_lmmodel_4 <- lm(creddebt ~ othdebt, data = credit.df)
broom::glance(credit_lmmodel_4)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC
0.4131225	0.4122658	1.607429	482.1943	2.526363e-81	1	-1299.884	2605.769	2619.366

1 row | 1-9 of 12 columns

Figure 4 Building the model

After trying different combinations of these variables we decided to include them all in our final model. We also tried using "+" and "*" in our model. R-squared of the model with "+" is equal to 0.559, but what is really surprising is that the r-squared of the model with "*" is equal to 1! Which means that the model perfectly fits the data. We will examine that later.



```r
credit_lmmodel_5 <- lm(creddebt ~ employ + income + debtinc + othdebt, data = credit.df)
broom::glance(credit_lmmodel_5)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC
0.5993808	0.5970312	1.330997	255.0912	6.983782e-134	4	-1168.734	2349.469	2376.663

1 row | 1-9 of 12 columns

```r
credit_lmmodel_6 <- lm(creddebt ~ employ * income * debtinc * othdebt, data = credit.df)
broom::glance(credit_lmmodel_6)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC
1	1	4.676478e-15	9.193439e+30	0	15	21701.69	-43369.39	-43292.34

1 row | 1-9 of 12 columns

Figure 5 Final models

PLOTTING THE RESULTS

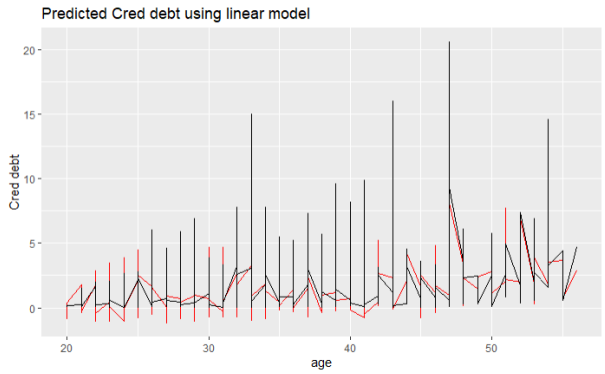


Figure 6 First model Predicted Creddebt

In this part of the project we decided to plot our results. Here we can see the Predicted Cred debt using linear model. The predictions are highlighted by the red colour.

Here we plotted the residuals of our model. We can see that this model fails mostly in the parts where the values reach over 5.

Figure 7 Firsst model Residuals

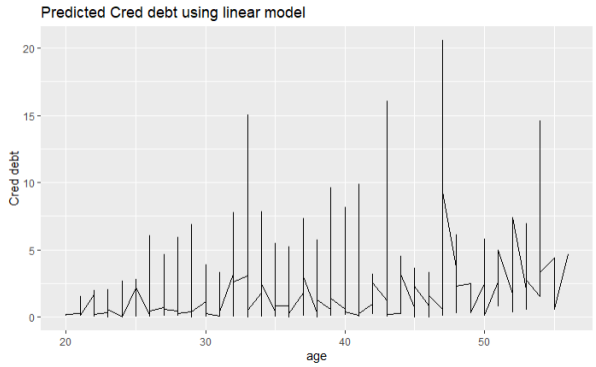


Figure 8 Second model Predicted Creddebt

These are the plotted results of our second model (the one with * instead +). We can notice that there is no red line, because the predictions are nearly perfect.

By looking at the residual graph we can see that the residuals reach very small nearly negligible values. The only part where the model fails (literally by a very tiny value) is when the values of Cred debt reach over 20.

Now we are sure that the second model perfectly fits the data.

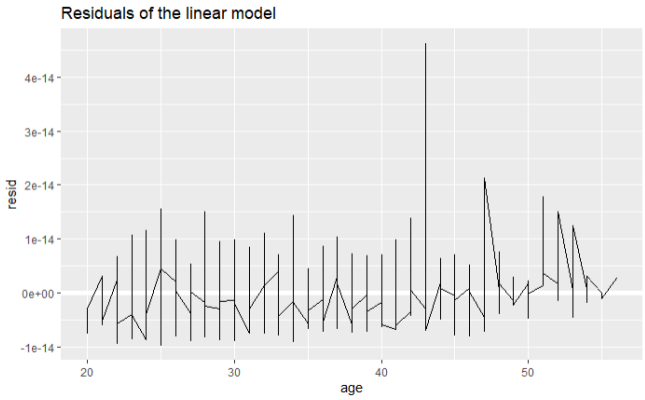


Figure 9 Second model Residuals

ASSUMPTIONS

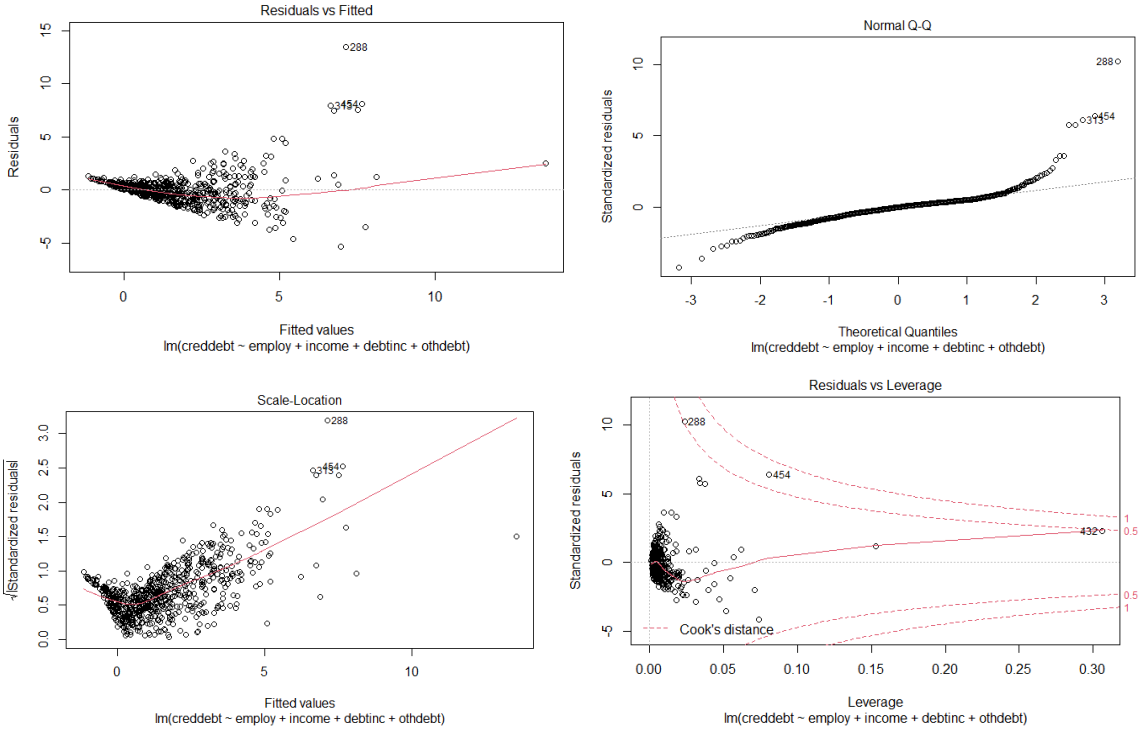


Figure 10 First model assumptions graphs

Here we can see the assumption graphs for the first model.

- ❖ At the first graph (Residuals vs Fited) we can check the linearity assumption. We can see that there is a systematic relationship between the residuals and the predicted values. Therefor the dependent variable is not linearly related to the independent variable. The assumption is suggesting a heteroscedasticity. We will check the other assumptions first.
- ✓ At the second graph (Normal Q-Q) we can check the normality assumption. Our points follow a straight line with some outliers around. This assumption is justified.
- ✓ At the third graph (Scale – Location) we can check the constant variance assumption. We can notice heteroskedasticity in our model as the points are not located around a horizontal line. Therefor we ran Breusch – Pagan test. The tiny p-value indicates heteroskedasticity as we th
ought.
Therefor we transformed the response variable using a log function. In this case we got even higher R2 value equal to 0.61. And this is how our Scale – Location graph looks like now.
- ✓ At the last graph (Residuals vs Leverage) we can spot outliers in our model. They are placed above the red dash line. There are 3 outliers: line 288, line 454 and line 432.

-

```
        studentized Breusch-Pagan test

data:  credit_lmmodel_5
BP = 121.59, df = 4, p-value < 2.2e-16
```

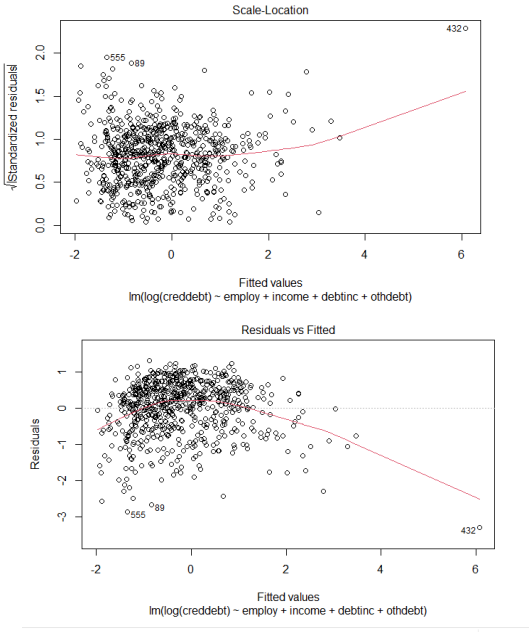


Figure 11 Model from log variable

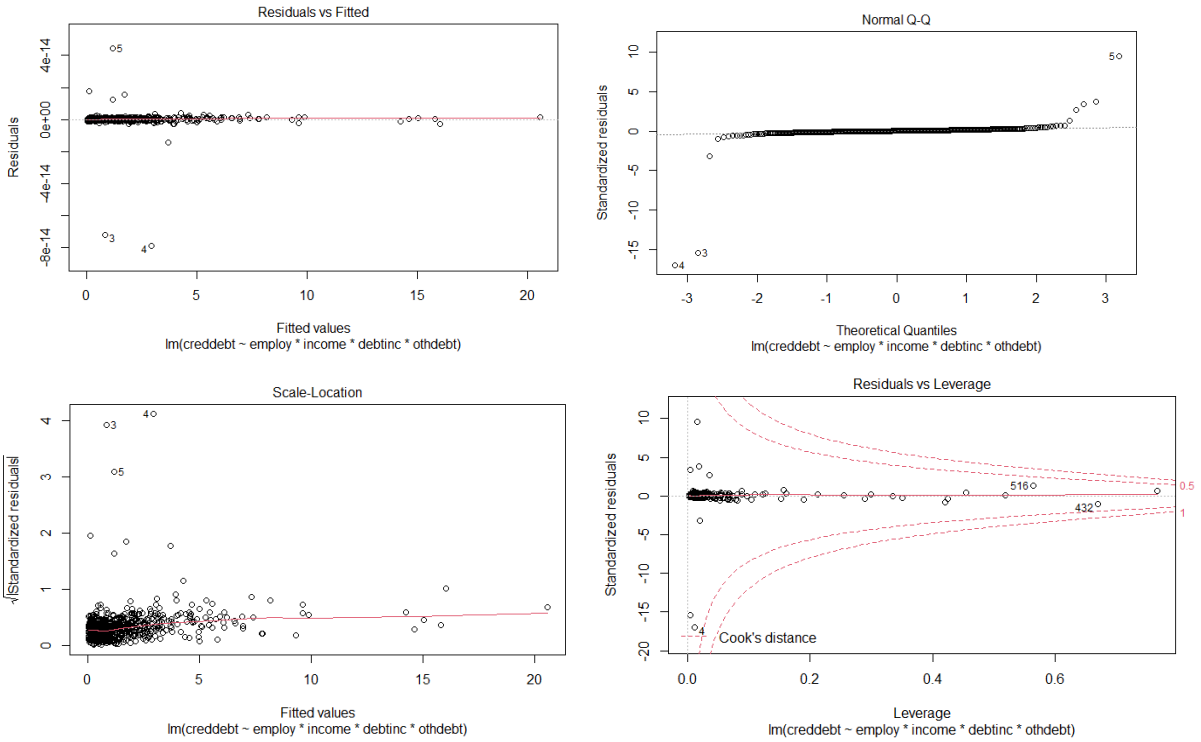


Figure 12 Second model assumptions graphs

Here we can see the assumption graphs for the second model.

- ✓ From the first graph (Residuals vs Fited) we can see that there is no systematic relationship between the residuals and the predicted values. That means that the dependent variable is linearly related to the independent variable. The assumption is justified.
- ✓ From the second graph (Normal Q-Q) we can check the normality assumption. However, from this graph it's very hard to say what is going on. Usually, the line should be 45 degrees, which is not in our case. If the 45 degrees don't matter, then this assumption would be justified as the points follow the line with only 3 outliers.
- ✓ From the third graph (Scale – Location) we can check the constant variance assumption. There is homoskedasticity in our model, therefor this assumption is justified.
- ✓ From the last graph (Residuals vs Leverage) we can notice that there are no outliers in our model.

DATA ACCURACY METRICS

We are evaluating our models using Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Mean Squared Error (MSE) metrics.

```r
#First model
MAPE(credit.lm1$pred,credit.lm1$creddebt)
MAE(credit.lm1$pred,credit.lm1$creddebt)
MSE(credit.lm1$pred,credit.lm1$creddebt)
```

```
[1] 3.923059
[1] 1.76938
[1] 5.392106
```

Figure 13 First model data accuracy metrics

The MAPE values is defined as actual to observed value minus the forecasted value. Value of 3.92 means that our model's predictions are, on average 3.92% off from the actual values.
The MAE value means that on average the forecasted distance from the true value is 1.77, which is a little bit high as our values range from 0 to 20.
The MSE refers to the mean of the squared difference between the predicted parameter and the observed .parameter. In our case the MSE is equal to 5.39.

```r
#Second model
MAPE(credit.lm2$pred,credit.lm2$creddebt)
MAE(credit.lm2$pred,credit.lm2$creddebt)
MSE(credit.lm2$pred,credit.lm2$creddebt)
```

```
[1] 1.474045e-14
[1] 3.672266e-15
[1] 2.402914e-29
```

Figure 14 Second model data accuracy metrics

For the second model the MAPE value is equal to 1.47 e-14, which is amazing. It means that our model's predictions are on average 0% off from the actual values.
The MAE value in this case is also very tiny. It means that on average the forecasted distance from the true value is 0.
The MSE which refers to the mean of squared difference between the predicted parameter and the observed parameter is equal to 0 as well.

From these values we can see that both of our models perform good. But the second model performs much better.

BIBLIOGRAPHY

[1] TechTarge, Predictive modeling, available online: https://searchenterpriseai.techtarget.com/definition/predictive-modeling

[2] Mason, K, 2021, CT5163 Applied Data Science in R lecture notes, Delivered at NUI Galway

[3] Investopedia, Regression Definition, available online: https://www.investopedia.com/terms/r/regression.asp

[4] Newell, J, 2021, ST2002 Statistics for Data Science 2 lecture notes, Delivered at NUI Galway

DECLERATION

I declare that the work I have submitted for this project is my own work and has been completed by me alone. I have acknowledged all material and sources that have been used in its preparation, whether they be books, publications, lecture notes, or any other kind of document. I have not copied or otherwise plagiarised any part of the work submitted for this project from other students and/or persons.

Date: 19/08/2021

Name: Maria Migrova

Signature: *Maria Migrova*