# MODULE 5 UNIT 3

## Activity submission

Learning outcomes:

LO4: Analyse the output from a multiple regression model.

LO5: Evaluate the use of multiple regression analysis.

### Name: Matlotlo Magasa

#### 1. Instructions and guidelines (Read carefully)

##### Instructions

1. Insert your name and surname in the space provided above, as well as in the **file name.** Save the file as: **First name Surname M5U3 Activity Submission** – **e.g. Lilly Smith M5U3 Activity Submission.** **NB:** *Please ensure that you use the name that appears in your student profile on the Online Campus.*

2. Write all your answers in this document. There is an instruction that says, “Start writing here” under each question. Please type your answer there.

3. Submit your assignment in **Microsoft Word only**. No other file types will be accepted.

4. You will be required to include visualisations that you have made in Tableau in this activity. To do this, export them as an image file and paste them into this document. Tableau has a page that details how to export your Tableau view as an image file.

5. Do **not delete the plagiarism declaration** or the **assignment instructions and guidelines**. They must remain in your assignment when you submit.

**PLEASE NOTE:** **Plagiarism cases will be investigated in line with the Terms and Conditions for Students.**

**IMPORTANT NOTICE:** Please ensure that you have checked your course calendar for the due date for this assignment.

##### Guidelines

1. There are 8 pages and 4 questions in this assignment.

2. This activity submission relies heavily on work done in the IDE notebook in the previous unit. If you have not done so already, return to the IDE and complete this activity before attempting to answer the questions.

3. You will be required to download a data file titled LOANS.sav in order to interact with the data in Tableau. Return to the Module 5 downloads folder and save a copy of the data file before attempting to answer the following questions.

4. Make sure that you have carefully read and fully understood the questions before answering them. Answer the questions fully, but concisely, and as directly as possible. Follow all specific instructions for individual questions (e.g. “list”, “in point form”).

5. Answer all questions in your own words. Do not copy any text from the notes, readings, or other sources. **The assignment must be your own work only.**

|  |
| --- |
| **Plagiarism declaration:** |
| **1. I know that plagiarism is wrong. Plagiarism is to use another’s work and pretend that it is one’s own.**  **2. This assignment is my own work.**  **3. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.**  **4. I acknowledge that copying someone else’s assignment (or part of it) is wrong and declare that my assignments are my own work.** |

#### 2. Mark allocation

Each question receives a mark allocation. However, you will only receive a final percentage mark and will not be given individual marks for each question. The mark allocation is there to show you the weighting and length of each question.

Question 1 10

Question 2 10

Question 3 10

Question 4 10

**TOTAL 40**

#### 3. Questions

The management team at Speedy Loans is interested in trying to predict the risk score of a potential client using information they have obtained from the client’s loan application. Given that multiple variables will be used to predict the client’s risk score, you decided that a multiple regression analysis would be most appropriate. After reviewing the results, the management team has asked you to answer specific questions related to the regression output.

##### Question 1

In the IDE activity, you performed a multiple regression analysis to determine the variables that could help predict a potential client’s credit risk score. The results of the analysis indicated that the variables shown in Table 1 were significant within the final model.

Table 1: Significant independent variables in the final regression model.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Age | Age of the applicant (in years) |
| Yearend | The number of years with current employer |
| Other debt | The average monthly other debts for applicants (in ZAR) |
| Education Some Tertiary | A dummy variable representing applicants with “Some Tertiary” education |
| Education Full Tertiary | A dummy variable representing applicants with “Full Tertiary” education |
| Location Durban | The geographical location of the applicant, categorised as “Durban” |
| Location Cape Town | The geographical location of the applicant, categorised as “Cape Town” |

The Speedy Loans management team has now asked you to provide some informative visuals of the relationship between the risk score and some of these variables, specifically the potential client’s age, the number of years they have been with their current employer, and any other debts they have.

The management team has asked you to create scatter plots with trend lines to visualise these relationships. To do this, drag “Risk Score” into the rows field, and convert the variable from a “Measure” to a “Dimension”. Following this, drag each independent variable of interest (“Age”, “Year\_Emp”, and “Other\_debt”) into the columns field. Make sure that you change each of these variables from a “Measure” to a “Dimension”. This will ensure that each applicant’s data points are shown and will allow for the effective use of a scatter plot in visualising the data. If desired, you may change the colour of each graph by dragging the variable onto the “Color” button in the “Marks” pane. Rename the worksheet “Relationship with risk score”.

To add the trend lines for each visualisation, you can right-click on each graph, select “Trend Lines” and click on “Show Trend Lines”. If you hover over each trend line, Tableau shows some informative details about each relationship, including the regression equation, the R2 value, and the p-value.

After creating your visualisations, the team has asked you to answer the following questions:

* Is the relationship between risk score and the independent variables positive or negative? Refer to the visualisation to support your answer.
* When hovering over the trend line for each graph, interpret the results of the p-value and R2 value to provide insight into how much each independent variable helps to explain the client’s risk score.

(Max. 300 words)

Start writing here:

The relationship between the Risk Score and all the independent variables is a negative one (inversely proportional). This means that the increase of a client’s age, their debt (other debt) and the number of years employed at their current employment contributes to a decrease in the clients risk score and thus the client becomes likely to be one who pays back their monthly credit contributions within the record time of payment. However, it is important to note that the degree of strength these independent variables have on the client’s risk score differs. This is easily distinguished by the slope of the trend line. Visually, one may see the major influence on the reduction of a client’s risk score is their age, followed by the number of years they are employed and lastly the amount of debt they have.

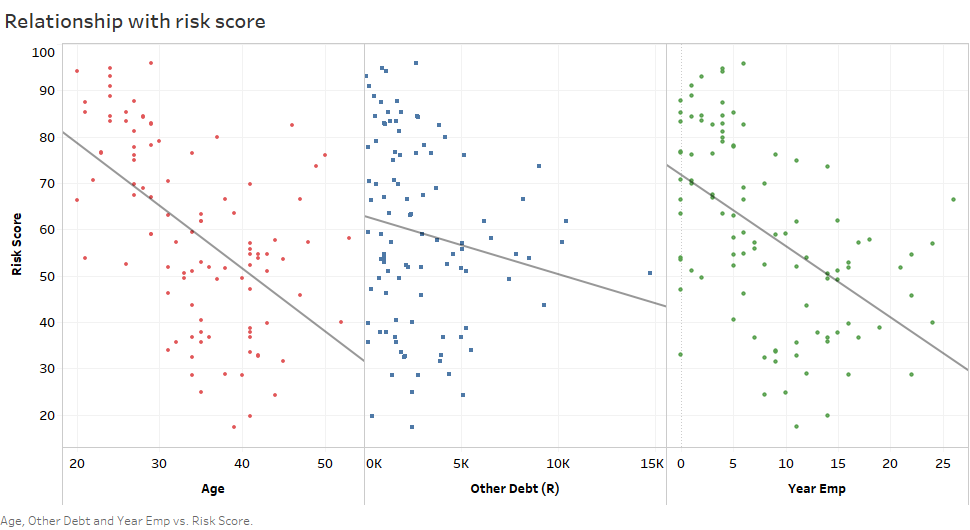
Table 2: The R2 and the p-values of the independent variables

|  |  |  |  |
| --- | --- | --- | --- |
| Indep. Variable | Age | Other Debt | No. of Year Employed |
| R2 Value | 0.296 | 0.028 | 0.267 |
| p-Value | <0.0001 | 0.093 | <0.0001 |

The R2 coefficient measures the overall significance of the regression model. The R2 of all the independent variables respectively, are relatively low with that of ‘Other debt’ being the lowest. For the simple linear regression of the three variables, the variation in the age of the client can only explain 29.6% of the variation in risk score and the variation in number of years the client has worked for their employer explains 26.7% of the risk score variation. Variation in ‘other debt’ explains a very low variation in the risk score (2.8%).

The p-values for all these independent variables are below the significance level of 0.1 (10%), meaning that there is statistical significance between these variables and the risk score and therefore theses independent variables each have influence on the predictive ability of the risk score.

Paste your Tableau visualisation here:



##### Question 2

The Speedy Loans management team has asked you to provide some insight into whether multicollinearity may exist between certain independent variables. You decided to test this using a correlation matrix and a heatmap in the analysis, but the team is unsure of how to interpret the results of the visualisation. Therefore, they have asked you to address the following:

* Use the heatmap to identify at least three pairs of independent variables that appear to be highly correlated with each other.
* Discuss how multicollinearity might impact the results of the regression model.
* Discuss how multicollinearity was addressed in the analysis. In other words, how were variables removed to account for these associations?

(Max. 200 words)

Start writing here:

Multicollinearity is the extent to which the independent variables are correlated. The three independent variables that are highly correlated are, ‘other debt’ and ‘debt-to-income ratio’; ‘other debt’ and ‘credit debt’; ‘income’ and ‘year employed’. It’s undesirable to have multicollinearity between independent variables because these variables may influence each other within the model, causing disproportionate weighting of observations and thus skewing the result of the output variable (i.e. risk score). This means an unreliable regression model to predict risk score which may also be difficult to interpret.

To address multicollinearity, highly correlated variables should be removed from the model. To do this we use the Stepwise Regression technique or use regularization. The backward stepwise regression method was used in this model. This method includes all independent variable to the regression model and will be removed one at a time based on the output of the p-values of each independent variable upon each step of this technique. The independent variable which is least significant (p-value > significance level) is first removed, regression is applied again, and the next independent variable is removed. The method is repeated until we have all independent variables with a p-values below the significance level of 0.1.

##### Question 3

The management team has asked you to interpret the results of the final regression model, specifically referencing how the potential client’s age, years employed at their current employer, and other debt can help predict their risk score. The management team would like you to address the following in your interpretation:

* The interpretation of the beta coefficients for each independent variable in relation to the risk score.
* Assuming that the beta coefficients reported have already been standardised, the team would like you to comment on the relative importance of each of the variables. In other words, you should rank the three independent variables mentioned in order of importance.

(Max. 150 words)

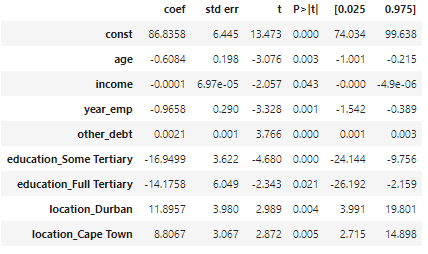
Start writing here:

The beta coefficient for the client’s age and year employed are both inversely proportional to the client’s risk score. The negative sign indicates that the older the client becomes, their risk score decreases by a factor of 0.608 and for every year the client stays at their current place of employment their risk score deceases by a factor of 0.966, assuming all other independent variables are held constant respectively. Increasing these means a lower risk client to company.

The client’s other debt has a positive relationship to their risk score. Increasing the amount of the client’s other debt will increase their credit risk score with a factor of 0.0021, keeping other variables constant.

The beta coefficients of the mentioned independent variables are relatively low compared to the other independent variables, therefore have the least impact on the risk score. In order of most to least importance: ‘years employed’, age, then ‘other debt’.

Table 3: Beta Coefficients of independent variables of final regression model



##### Question 4

At the beginning of your analysis in the previous component, you were asked to create dummy variables for the “Location” and “Education” variables given that they were categorical in nature and had more than two categories per variable. For the “Location” variable, you used “Johannesburg” as the reference variable, or category. For the “Education” variable, you used “Primary” education as the reference variable, or category.

The Speedy Loans management team is now interested in the interpretation of these dummy variables in the final regression model. They have asked you to address the following:

* “Some Tertiary” and “Full Tertiary” education were significant in the final regression model. Interpret the result output in reference to the reference variable, or “Primary” education.
* “Durban” and “Cape Town” were significant in the final regression model. Interpret the result output in reference to the reference variable, or “Johannesburg”.

(Max. 200 words)

Start writing here:

The education dummy variables used both display an inverse/negative relationship with the risk score. The independent variable of some tertiary education has more influence on the other education levels in this model because the its beta coefficient is the highest of the three. A client with some tertiary education will have 16.95 risk score points less than a client with primary education holding all else equal.

Clients with a full tertiary education has a risk score of 14. 18 less than the client with primary education (holding all else equal). Ceteris paribus, this means that the client with some tertiary education is a lower risk than the client with primary school and tertiary education.

The location dummy variables used both display a positive relationship with the risk score. Being from Durban has a greater influence on risk score than other locations. A client from Durban will have 11.90 more risk rating than a person from Johannesburg and a client from Cape Town will have 8.81 more risk rating points than a person from Johannesburg, holding all else equal respectively. Based *solely* on location, this means that client from Johannesburg, has the least risk score and is thus a lower risk.

**4. Rubric**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **No submission** | **Poor** | **Good** | **Excellent** |
| **Question 1**  *The image supplied contains the correct graphs and adheres to the instructions. The graphs contain the correct variable(s). There is a full interpretation of the results.* | No submission or attempt. (0) | The image supplied contains the incorrect graphs, or uses the incorrect variable(s). The interpretation of the results is poor. (3) | The image supplied contains the correct graphs and uses the correct variable(s), but not all instructions have been adhered to. The interpretation of the results is acceptable. (6) | The image supplied contains the correct graphs and uses the correct variable(s). All instructions have been adhered to. The interpretation of the results is exceptional. (10) |
| **Question 2**  *The student correctly identified at least three pairs of highly-correlated variables and describes how multicollinearity can impact regression results, providing insight into how this was handled in the analysis.* | No submission or attempt. (0) | The student identified fewer than three pairs of highly-correlated variables.  OR  The student correctly identified three pairs of variables, but failed to discuss how multicollinearity impacts the regression results or provide insight into how this was addressed. (3) | The student correctly identified at least three pairs of highly-correlated variables, attempted to discuss how multicollinearity impacts the regression results and provided some insight into how this was addressed. (6) | The student correctly identified at least three pairs of highly-correlated variables, clearly discussed how multicollinearity impacts the regression results and provided the necessary insight into how this was addressed. (10) |
| **Question 3**  *The student interpreted the regression output for the specified variables only, and interpreted the beta coefficients and their relative importance correctly.* | No submission or attempt. (0) | The student interpreted all the variables.  OR  Student interpreted the specified variables only, but has failed to interpret the beta coefficients or the relative importance. (3) | The student interpreted the specified variables only, but has not fully interpreted the beta coefficients or the relative importance. (6) | The student interpreted the specified variables only, and clearly and correctly interpreted the beta coefficients and the relative importance. (10) |
| **Question 4**  *The student interpreted the dummy variables correctly and specified the result in relation to the correct reference variable.* | No submission or attempt. (0) | The student attempted to answer the question, but has either provided an unsatisfactory interpretation of one of the dummy variables, or failed to address them entirely. (3) | The student provided a satisfactory interpretation of both dummy variables, but failed to include information in relation to the reference variable. (6) | The student provided an excellent interpretation of both dummy variables, and specified the results in relation to the correct reference variables. (10) |

**Total:** 40 marks