Data Analysis Project   28.25/31 Points

13/03/2022

| Attempt 1   ⌄ | ◯   **REVIEW FEEDBACK** 13/03/2022 |

Attempt 1 Score: 🗩 View feedback
   28.25/31

1 Attempt Allowed

⌄ **Details**

For this project, you are required to obtain and analyze a dataset and produce a report. The report should have the structure described below, with each section containing meaningful responses to the questions included in each section.

Preparatory steps:

1. Install the R package *faraway*. This package, produced by Julian Faraway, consists of several datasets and R functions related to analyzing data using a set of approaches commonly known as the *generalized linear modeling* approach.

2. From this package, import the dataset *debt* into the R environment.

3. Refer to the documentation associated with *debt* (use **?debt** for looking up the R documentation on the dataset) to get basic descriptions of the variables in the dataset. Be sure to look at both the descriptions of the individual variables, and the description of the personality measure *locintrn*, which represents the locus of control (you will see its description in the "Details" section of the R documentation page of *debt*.

4. Make sure that you understand what each column represents.

Next, create a report that has a structure described below and contains your analysis of the dataset, where *prodebt* is the variable that is going to be treated at the outcome of interest. Within each section of the report, you will need to provide appropriate graphical and/or numerical output, and address the questions contained in each section. In the text that follows, the section titles in italic font will be the section titles that you would use in report. And within each section in your report, you would provide explanations to address the points indicated. Note: your explanations should be in a narrative form, distributed into paragraphs; please do not use a question-answer format in your report.

Introduction

1. Provide an introduction to the dataset:

   a. its name

   b. how the data were collected

   c. how you obtained the dataset

   d. details of each column present in the dataset, summarized in a table with four columns: *name of the variable*; *description of the variable*; *measurement type* (factor or continuous); *role – predictor* or *outcome.*

2. Provide numerical summaries of all of the variables in the dataset. Identify clearly any variables that ought to be

| > | | > |

(https://canvas.slu.edu/courses/19708/modules/items/491544)   **(https://canvas.slu.edu/courses/197**

4. State how many observations have been excluded from the original dataset in creating a "cleaned" version of the dataset with missing values excluded.

Analyses

1. Using GGally's matrix plotting function, create a summary of all of the variables.

2. Comment on what you find in the summary plots. Specifically, are there any bivariate relationships between the predictors and the outcome that look "interesting"? What makes them interesting?

3. Create a linear regression model to address the following questions:

   a. Ignoring all predictors and using only *locintrn*, fit a linear regression model to predict the outcome *prodebt*.

      i. Is the model significant?

      ii. What is the amount of variance explained?

      iii. Is *locintrn* a statistically significant predictor of *prodebt*?

      iv. Provide an interpretation of the value of the coefficient associated with *locintrn.*

   b. Next, create a second model, where you have *locintrn*, and *manage* as predictors. Then, answer the following questions:

      i. Is the new two-predictor model a better fit when compared to the fit statistics associated with model 1. Explain by drawing upon relevant evidence.

      ii. Interpret the coefficients of *locintrn* and *manage*.

      iii. Explain the implication of the change in the coefficient of *locintrn* from model 1 to model 2.

   c. Next, build a three predictor model, adding *children* as a third predictor and answer the following questions:

      i. Is the new three-predictor model a better fit when compared to the fit statistics associated with model 2? Explain by drawing upon relevant evidence.

      ii. Explain what the change in the coefficient of *locintrn* and *manage* from model 2 to model 3 would imply, taking into account any changes in the model fit.

      iii. Explain, on the basis of the coefficient of *children*, the effect that having children appears to have on a respondent's attitude towards debt.

   d. Next, build a four-predictor model, adding *singpar* as a fourth predictor and answer the following questions:

      i. Is the new four-predictor model a better fit when compared to the fit statistics associated with model 3. Explain by drawing upon relevant evidence.

      ii. Explain, on the basis of the coefficient of *singpar*, the effect that being a single parent appears to have on a respondent's attitude towards debt, taking into account any changes in the model fit.

   e. Next, build a five-predictor model, adding another predictor that you find interesting and expect to add to the explanation of which set of predictors (including the ones included in the previous four models, and the most recent one you have added) are good indicators of an individual's attitude towards debt. Then,

      i. Explain your rationale for choosing the particular fifth variable you have chosen.

1. Based on the results of your analyses, what conclusions would you provide to an individual on how they could change their attitude towards debt?

2. Are there other variables that you could add to improve your ability to predict an individual's attitude towards debt? If so, identify one additional one (additional, in comparison to the variables you have already included in the five models you have built) and explain why you think this might be a useful predictor.

Deliverables:

1. A .R (not .Rdata) file that contains all of the R code you have used for completing the requirements of the project.

2. A PDF document, whose structure meets the requirements specified above. In this document, you are to cover all of the points mentioned above by including narrative explanations alongside numerical and graphical information.

∨ **View Rubric**

Select Grader

Srikanth Mudigonda ⌄

>
(https://canvas.slu.edu/courses/19708/modules/items/491544)

>
(https://canvas.slu.edu/courses/197

Rubric for data analysis project

Criteria	Ratings		Points
Deliverables [view longer description](#)	**2 pts** **Full Marks** ▲	**0 pts** **No Marks**	2 / 2 pts
Document format [view longer description](#)	**1 pts** **Full Marks** ▲	**0 pts** **No Marks**	1 / 1 pts
Clarity [view longer description](#)	**3.5 pts** **Full Marks** ▲	**0 pts** **No Marks**	3.5 / 3.5 pts
Name of the dataset	**0.25 pts** **Full Marks** ▲	**0 pts** **No Marks**	0.25 / 0.25 pts
How the data were collected	**0.25 pts** **Full Marks** ▲	**0 pts** **No Marks**	0.25 / 0.25 pts
How you obtained the dataset	**0.5 pts** **Full Marks** ▲	**0 pts** **No Marks**	0.5 / 0.5 pts
Details of each column present in the dataset, summarized in a table with four columns: name of the variable; description of the variable; measurement type (factor or continuous); role – predictor or outcome.	**2 pts** **Full Marks** ▲	**0 pts** **No Marks**	2 / 2 pts

> ❯   ❯

[(https://canvas.slu.edu/courses/19708/modules/items/491544)](https://canvas.slu.edu/courses/19708/modules/items/491544)   **[(https://canvas.slu.edu/courses/197](https://canvas.slu.edu/courses/197)**

Rubric for data analysis project		
any variables that ought to be converted into a factor form. **Criteria**	**Ratings**	**Points**
Create a "cleaned" version of the dataset by excluding observations that have one or more variables with a value of NA. Use the na.omit() function for this.	**1 pts** **Full Marks** ▲	**0 pts** **No Marks**	1 / 1 pts
State how many observations have been excluded from the original dataset in creating a "cleaned" version of the dataset with missing values excluded.	**1 pts** **Full Marks** ▲	**0 pts** **No Marks**	1 / 1 pts
Using GGally's matrix plotting function, create a summary of all of the variables.	**1 pts** **Full Marks** ▲	**0 pts** **No Marks**	1 / 1 pts
Comment on what you find in the summary plots. Specifically, are there any bivariate relationships between the predictors and the outcome that look "interesting"? What makes them interesting?	**2 pts** **Full Marks** ▲	**0 pts** **No Marks**	2 / 2 pts
Ignoring all predictors and	**0.5 pts**	**0 pts**	0.5 / 0.5 pts

>   >

(https://canvas.slu.edu/courses/19708/modules/items/491544)   **(https://canvas.slu.edu/courses/197**

Rubric for data analysis project		
Criteria	**Ratings**	**Points**
predict the outcome prodebt. i. Is the model significant?		
What is the amount of variance explained?	**0.5 pts** **Full Marks** ▲ · **0 pts** **No Marks**	0.5 / 0.5 pts
Is locintrn a statistically significant preditor of prodebt?	**0.5 pts** **Full Marks** ▲ · **0 pts** **No Marks**	0.5 / 0.5 pts
Provide an interpretation of the value of the coefficient associated with locintrn.	**0.5 pts** **Full Marks** ▲ · **0 pts** **No Marks**	0.5 / 0.5 pts
Next, create a second model, where you have locintrn, and manage as predictors. Then, answer the following questions: i. Is the new two-predictor model a better fit when compared to the fit statistics associated with model 1. Explain by drawing upon relevant evidence.	**5 pts** **Full Marks** · **0 pts** **No Marks** **Comments** Improvement in model fit is determined by checking via anova() for the change in F-statistic being significant. This was not done.	3.5 / 5 pts
Interpret the coefficients of locintrn and	**0.5 pts** **Full Marks** · **0 pts** **No Marks**	0.5 / 0.5 pts

> [(https://canvas.slu.edu/courses/19708/modules/items/491544)](https://canvas.slu.edu/courses/19708/modules/items/491544)   > [(https://canvas.slu.edu/courses/197](https://canvas.slu.edu/courses/197)

Rubric for data analysis project

Criteria	Ratings		Points
Explain what the change in the coefficient of locintrn from model 1 to model 2 would imply.	**0.5 pts** **Full Marks** ▲	**0 pts** **No Marks**	0.5 / 0.5 pts
Next, build a three predictor model, adding children as a third predictor and answer the following questions: i. Is the new three-predictor model a better fit when compared to the fit statistics associated with model 2. Explain by drawing upon relevant evidence.	**0.5 pts** **Full Marks**	**0 pts** **No Marks**	0.25 / 0.5 pts
Explain what the change in the coefficient of locintrn and manage from model 2 to model 3 would imply, taking into account any changes in the model fit.	**0.5 pts** **Full Marks** ▲	**0 pts** **No Marks**	0.5 / 0.5 pts
Explain, on the basis of the coefficient of children, the effect that having children appears to have on a respondent's attitude towards	**0.5 pts** **Full Marks** ▲	**0 pts** **No Marks**	0.5 / 0.5 pts

> **[(https://canvas.slu.edu/courses/19708/modules/items/491544)](https://canvas.slu.edu/courses/19708/modules/items/491544)**

> **[(https://canvas.slu.edu/courses/197](https://canvas.slu.edu/courses/197)**

Rubric for data analysis project			
Criteria	**Ratings**		**Points**
adding singpar as a fourth predictor and answer the following questions: i. Is the new four-predictor model a better fit when compared to the fit statistics associated with model 3. Explain by drawing upon relevant evidence.	**Full Marks**	**No Marks**	
	Comments Improvement in model fit is determined by checking via anova() for the change in F-statistic being significant. This was not done.		
Explain, on the basis of the coefficient of singpar, the effect that being a single parent appears to have on a respondent's attitude towards debt, taking into account any changes in the model fit.	**0.75 pts Full Marks** ▲	**0 pts No Marks**	0.75 / 0.75 pts
Next, build a five-predictor model, adding another predictor that you find interesting and expect to add to the explanation of which set of predictors (including the ones included in the previous four models, and the most recent one you have added)	**0.75 pts Full Marks** ▲	**0 pts No Marks**	0.75 / 0.75 pts

> **(https://canvas.slu.edu/courses/19708/modules/items/491544)**

> **(https://canvas.slu.edu/courses/197**

Rubric for data analysis project

Criteria	Ratings		Points
attitude towards debt.			
Explain whether the fifth variable you have added allowed you to improve your ability to predict an individual's attitude towards debt.	**0.75 pts** Full Marks	**0 pts** No Marks	0.25 / 0.75 pts
Based on the results of your analyses, what conclusions would you provide to an individual on how they could change their attitude towards debt?	**1 pts** Full Marks ▲	**0 pts** No Marks	1 / 1 pts
Are there other variables that you could add to improve your ability to predict an individual's attitude towards debt? If so, identify one additional one (additional, in comparison to the variables you have already included in the five models you have built) and explain why you think this would be a useful predictor.	**1 pts** Full Marks ▲	**0 pts** No Marks	1 / 1 pts
			Total points: 28.25

File name	Size	
🔗 Srijesh_D...k8_5000.R	2.21 KB	✅
📄 srijesh_y...oject.pdf	405 KB	✅
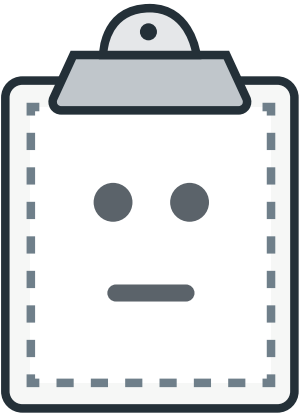
Preview Unavailable

Srijesh_Data_Analysis_Project_week8_5000.R

⬇ Download

**(https://canvas.slu.edu/files/1215308/download?
download_frd=1&verifier=Zqyg3g3awUVwFli5TrWFXrBDwJ2N4DenXlc0FaqX)**

You are unable to submit to this assignment as your enrolment in this course has been concluded.