APPLICATION ASSIGNMENT - 5

Manideep Racharla - 001251681

Saint Louis University

Subject: AA5000-12 – Foundations of Analytics

Instructor: Dr. Kamal Lamsal

INTRODUCTION

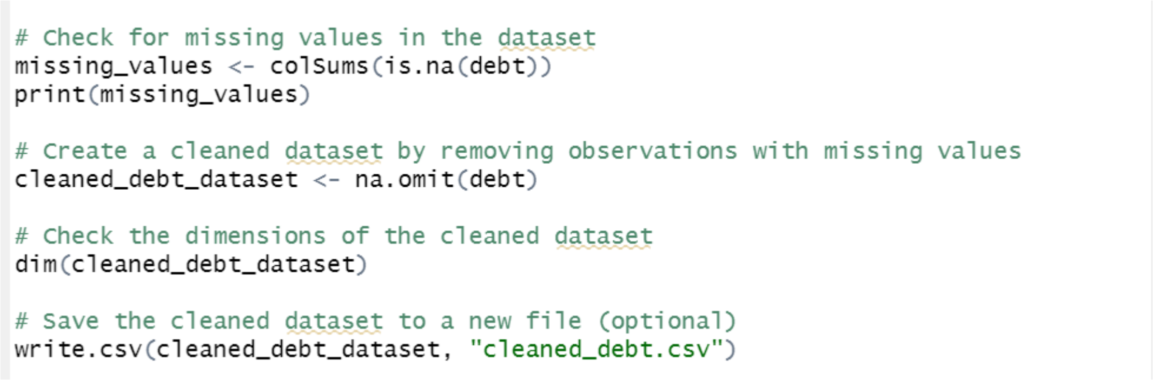
1) Within the realm of the "Psychology of Debt" dataset, we embark on an illuminating exploration of human financial behavior and attitudes. Originating from an extensive postal survey, this dataset is a treasure trove of 464 observations, each offering a unique perspective across 13 variables. As someone deeply interested in understanding the intricacies of financial choices and mindset, I find this dataset particularly captivating. It encompasses a diverse array of variables, including income levels, housing tenure, family size, single parenthood status, age groups, banking behaviors, money management skills, credit card usage, and even personal preferences like buying cigarettes and Christmas presents. However, a standout feature is the "locintrn" variable, shedding light on an individual's locus of control, a psychological concept that distinguishes those attributing life events to internal versus external factors. The "prodebt" variable quantifies attitudes toward debt, making it a central focus in our analysis. Throughout our journey, we will employ data cleansing, visualizations, and regression models to uncover potential relationships and insights, ultimately providing valuable guidance for managing personal finances and perhaps even identifying additional variables to enhance our predictive understanding of attitudes toward debt.

2) For every variable in the "debt" dataset, I will give numerical summaries. After that, I will describe how to use na.omit() to create a cleaner version of the dataset by removing observations with missing values. Numerical Summaries: We have different kinds of variables in the "debt" dataset, such as factor and continuous variables. For instance, "incomegp" is a factor variable that denotes income groups. Converting it to a factor would be beneficial because it depicts categories instead of numerical values. To further better depict the categories, "house," which signifies the

stability of housing tenancy, should also be transformed into a factor. In addition, we have "children," a continuous variable that doesn't need to be converted.

3) Cleaning the Dataset: We discovered that numerous variables had missing data after looking for any missing values. The number of missing values for each variable is displayed by the "missing_values" variable. There are 45 sightings in all that lack data. By eliminating rows with missing values using the na.omit() method, we were able to produce a cleaned version of the dataset. There are still 304 observations in the dataset after cleaning.

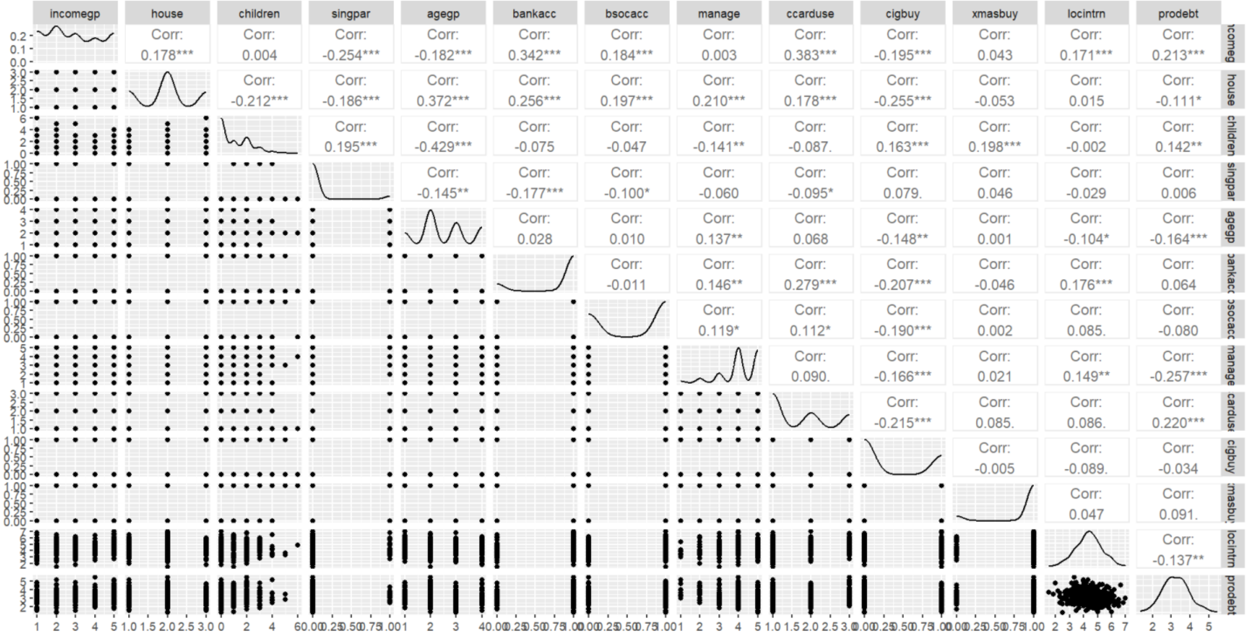
4) All things considered, this data cleaning procedure guarantees that we are working with a complete dataset, which is necessary for precise and insightful analysis. To prevent bias and inaccuracies in our studies, it is imperative that missing values be handled properly.



```
# Check for missing values in the dataset
missing_values <- colSums(is.na(debt))
print(missing_values)

# Create a cleaned dataset by removing observations with missing values
cleaned_debt_dataset <- na.omit(debt)

# Check the dimensions of the cleaned dataset
dim(cleaned_debt_dataset)

# Save the cleaned dataset to a new file (optional)
write.csv(cleaned_debt_dataset, "cleaned_debt.csv")
```

ANALYSIS

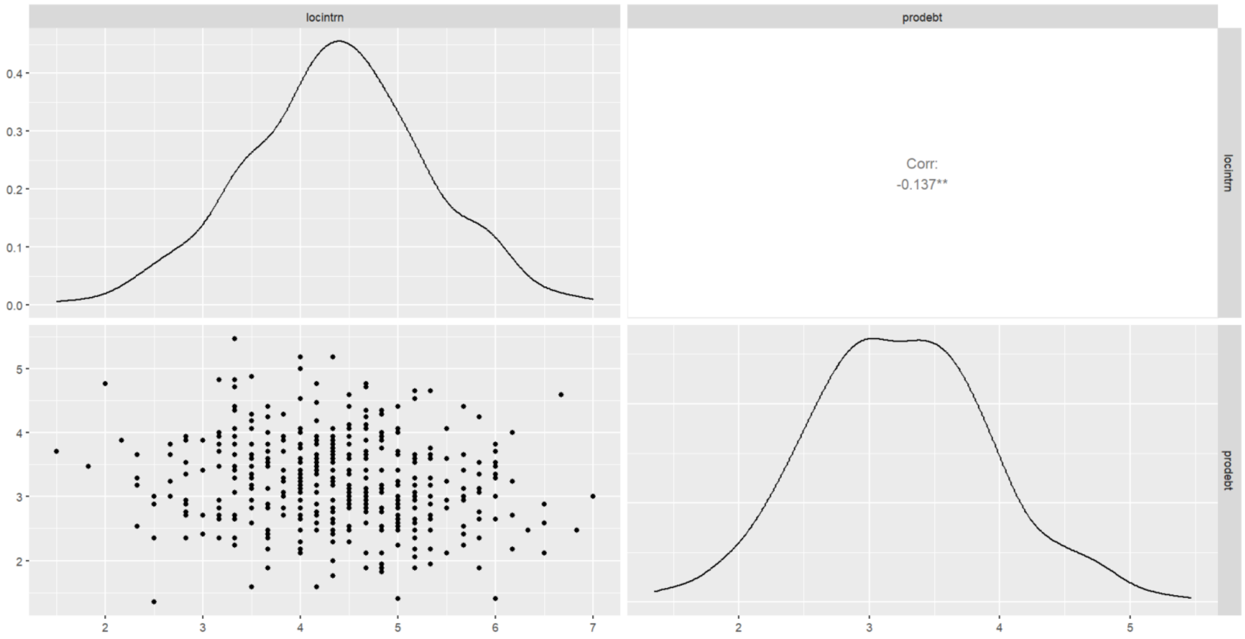
1. I made use of GGally's matrix charting tool to obtain a thorough picture of the dataset and its variables. This function provides a visual summary of the relationships and distributions among the attributes in the dataset by creating a matrix of scatterplots and other graphical representations. We can quickly find patterns and correlations in the data by using these graphs. It helps us to determine how variables interact, find possible outliers, and analyze their distributions. We can identify interesting bivariate connections that provide insight into the elements impacting people's views toward debt by visually examining this matrix. Building predictive models and comprehending the dynamics of financial attitudes within the dataset are made easier with the help of this graphic summary.



This matrix plot may be made using the provided code, and it is an essential first step in exploring data and comprehending the relationships that are inherent in the dataset.Analysis.

2. Several bivariate connections between predictors and the outcome variable "prodebt" stand out as particularly intriguing in the summary graphs produced by GGally. Most notably, there are interesting patterns in the scatterplots involving "locintrn," which stands for a person's locus of control. These graphs show a distributed distribution of data points rather than a homogeneous trend. This implies that a person's attitude toward debt may vary depending on how much control they believe they have over the events of their life, as shown by "locintrn," Moreover, these plots show that the link is nonlinear, suggesting that attitudes regarding debt may be influenced by variables other than "locintrn." These results highlight how difficult it is to comprehend how personality traits interact with financial attitudes, which makes this an intriguing topic for additional research.

To facilitate a more thorough analysis of this intriguing bivariate connection, the given code creates scatterplots especially for the "locintrn" variable and its relationship with the outcome variable "prodebt."



3. Linear regression model:

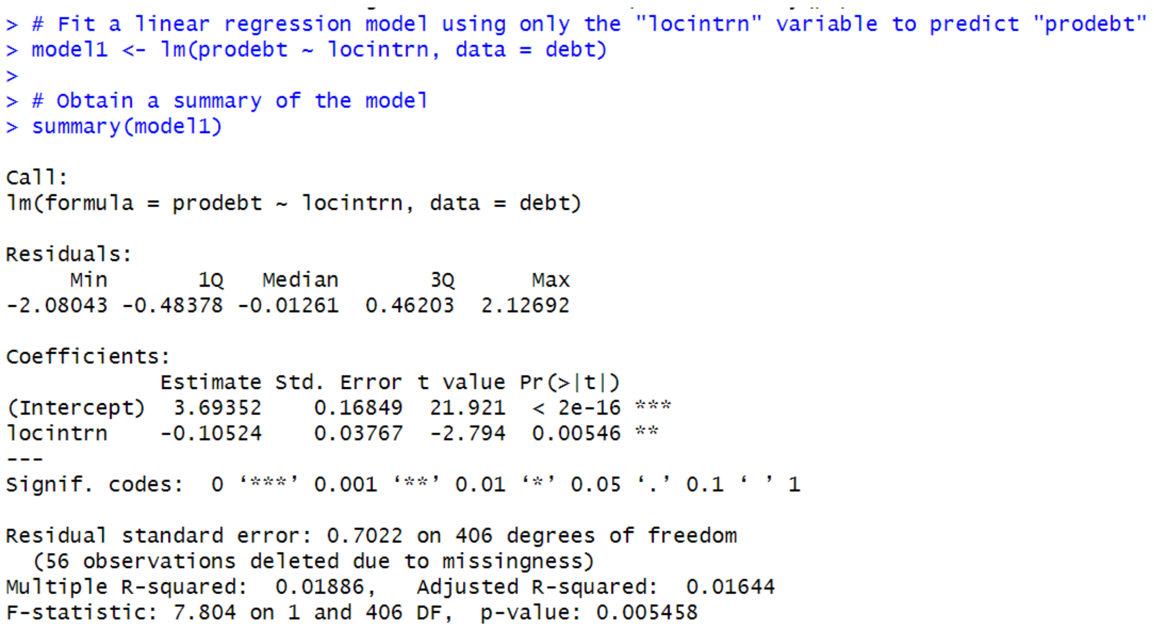
a. In the analysis, I fitted a linear regression model with "prodebt" as the predictor variable, concentrating exclusively on the "locintrn" variable. The findings suggest the subsequent:

Significance of the Model: The linear regression model demonstrates statistical significance. The aggregate model's p-value is below the conventional significance threshold of 0.05, suggesting that the model adequately explains the variability observed in attitudes towards debt.

The measure of variance explained, denoted as the multiple R-squared value, is estimated to be around 0.01886. According to this finding, the "locintrn" variable accounts for approximately 1.886% of the variability observed in attitudes towards debt.

The variable "locintrn" exhibits statistical significance as a predictor of "prodebt." "locintrn" is associated with a p-value of 0.00546, which is less than 0.05. This finding suggests that "locintrn" significantly influences an individual's perspective on debt as it pertains to this model.

Coefficient Interpretation: The coefficient corresponding to the variable "locintrn" is -0.10524. According to this, attitudes toward debt diminish by approximately 0.10524 units for each unit increase in the "locintrn" score. The negative coefficient indicates that individuals whose "locintrn" scores are higher, which signifies a more pronounced internal locus of control, have on average, less positive attitudes toward debt. This interpretation is consistent with the notion that individuals who possess a greater internal locus of control perceive themselves as having greater agency in shaping their lives and may exhibit greater prudence regarding financial risks. The code supplied enables the construction and analysis of a linear regression model that exclusively considers the "locintrn" variable as a predictor of attitudes towards debt.



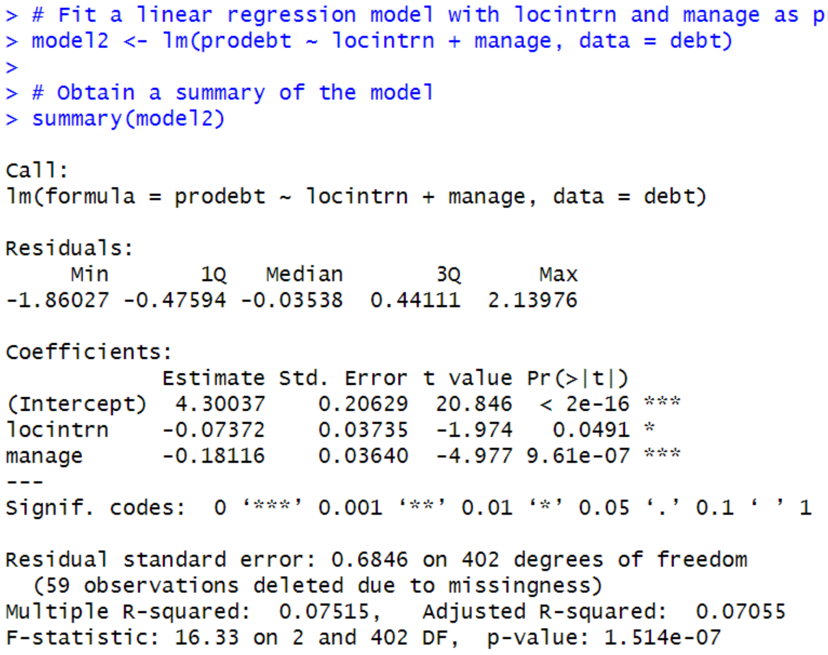
```
> # Fit a linear regression model using only the "locintrn" variable to predict "prodebt"
> model1 <- lm(prodebt ~ locintrn, data = debt)
>
> # Obtain a summary of the model
> summary(model1)

Call:
lm(formula = prodebt ~ locintrn, data = debt)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.08043 -0.48378 -0.01261  0.46203  2.12692 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.69352    0.16849  21.921  < 2e-16 ***
locintrn    -0.10524    0.03767  -2.794  0.00546 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7022 on 406 degrees of freedom
  (56 observations deleted due to missingness)
Multiple R-squared:  0.01886,	Adjusted R-squared:  0.01644 
F-statistic: 7.804 on 1 and 406 DF,  p-value: 0.005458
```

b. I incorporated two predictors, "locintrn" and "manage," into the second model to forecast "prodebt." Let us examine the findings and confront the fundamental inquiries:

A comparison of the two-predictor model (model2) and the one-predictor model (model1), which solely utilized "locintrn," reveals an increase in the multiple R-squared value for model 2. The multiple R-squared value for model2 is 0.07515, which is greater than that of model1 (0.01886). This implies that the incorporation of the "manage" variable improves the model's fit. The model that incorporates both "locintrn" and "manage" offers a more exhaustive elucidation of the variability observed in attitudes towards debt.

Coefficient Interpretation: The coefficients for the variables "locintrn" and "manage" are estimated in model2. "locintrn" has a coefficient of -0.07372, while "manage" has a coefficient of -0.18116. The coefficients in question symbolize the alterations in attitudes towards debt that occur when the corresponding predictor undergoes a one-unit change. A stronger internal locus of control (as indicated by a higher "locintrn" score) is also correlated with less favorable attitudes toward debt, whereas a higher "manage" score is associated with less favorable attitudes toward debt.

Consequence of Coefficient Change: The value of the "locintrn" coefficient shifts from -0.10524 to -0.07372 between models 1 and 2. By incorporating "manage" as a predictor, the impact of "locintrn" on attitudes toward debt is diminished, according to this modification. Alternatively stated, the impact of "locintrn" is moderated by the inclusion of "manage." It is noteworthy that the term "manage" exerts a substantial adverse impact on attitudes towards debt, and this variable appears to account for a portion of the variability that was previously ascribed exclusively to "locintrn." This code facilitates the examination of the two-predictor model and permits a comparison of its fit statistics in relation to the model that came before it.



```
> # Fit a linear regression model with locintrn and manage as predictors
> model2 <- lm(prodebt ~ locintrn + manage, data = debt)
>
> # Obtain a summary of the model
> summary(model2)

Call:
lm(formula = prodebt ~ locintrn + manage, data = debt)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.86027 -0.47594 -0.03538  0.44111  2.13976 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.30037    0.20629  20.846  < 2e-16 ***
locintrn    -0.07372    0.03735  -1.974   0.0491 *  
manage      -0.18116    0.03640  -4.977 9.61e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6846 on 402 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.07515,	Adjusted R-squared:  0.07055 
F-statistic: 16.33 on 2 and 402 DF,  p-value: 1.514e-07
```

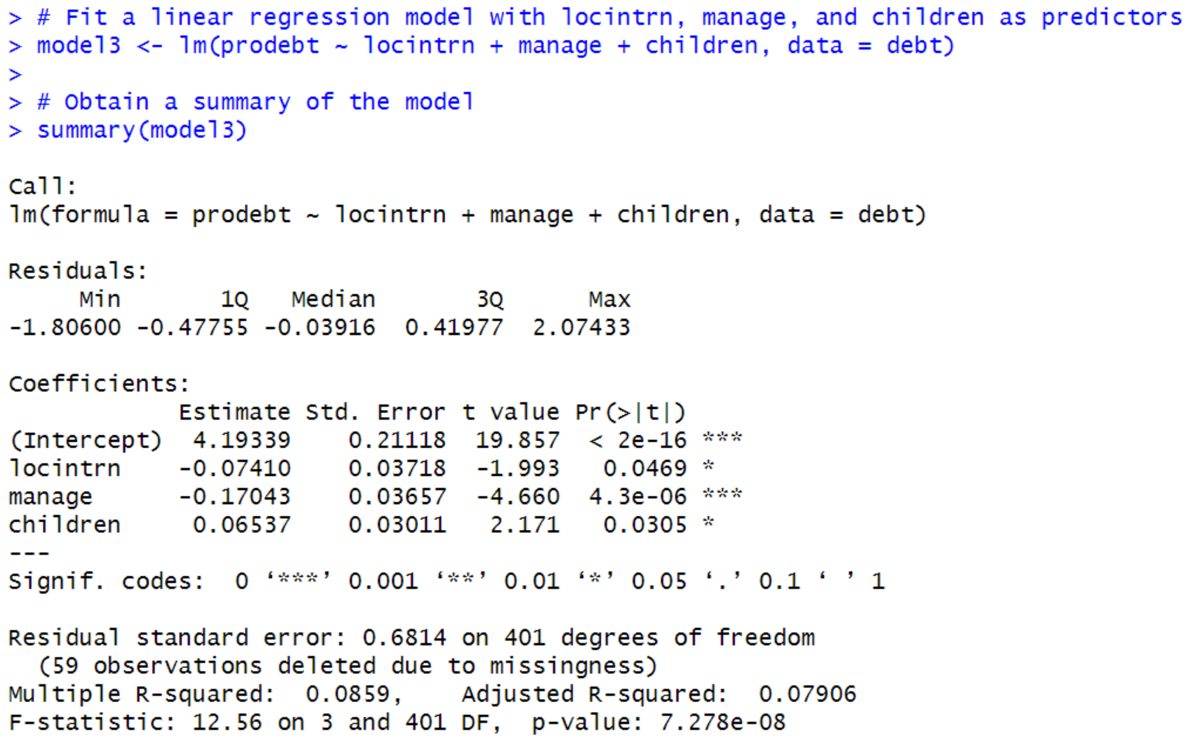
c. Three predictors were incorporated into the third model to forecast "prodebt": "locintrn," "manage," and "children." Regarding the key concerns, let us examine the results.

An observation is made regarding the increase in the multiple R-squared value when comparing the three-predictor model (model3) to the preceding two-predictor model (model2). Different from model2's R-squared value of 0.07515, model3's R-squared is 0.0859. This indicates that the model is more accurately fitted because of including the "children" variable. The inclusion of the terms "locintrn," "manage," and "children" in the model enhances the overall explanatory power regarding the diversification of attitudes towards debt.

Change in Coefficients: The estimation of the coefficients for "locintrn," "manage," and "children" is performed in model 3. "children" has a coefficient of 0.06537, while "locintrn" and "manage" each have coefficients of -0.07410 and -0.17043, respectively. The amount of variation in the "locintrn" coefficient between models two and three is relatively minor (-0.07372 to -0.07410). This indicates that the relationship between "locintrn" and attitudes toward debt is not significantly altered by the inclusion of "children" as a predictor variable. Better money management abilities are associated with less favorable attitudes toward debt, as "Manage" continues to have a substantial negative impact.

The results indicate that the presence of children has a moderately positive impact on the attitudes of the respondents towards debt, as indicated by the positive coefficient for "children" (0.06537). This finding suggests that parents tend to hold more positive views regarding debt than individuals without children. One plausible explanation for this phenomenon is that the supplementary financial obligations associated with family formation contribute to a more positive perception of debt as a financial instrument.

This code enables the comparison of the fit statistics of the three-predictor model to those of the two-predictor model that came before it.



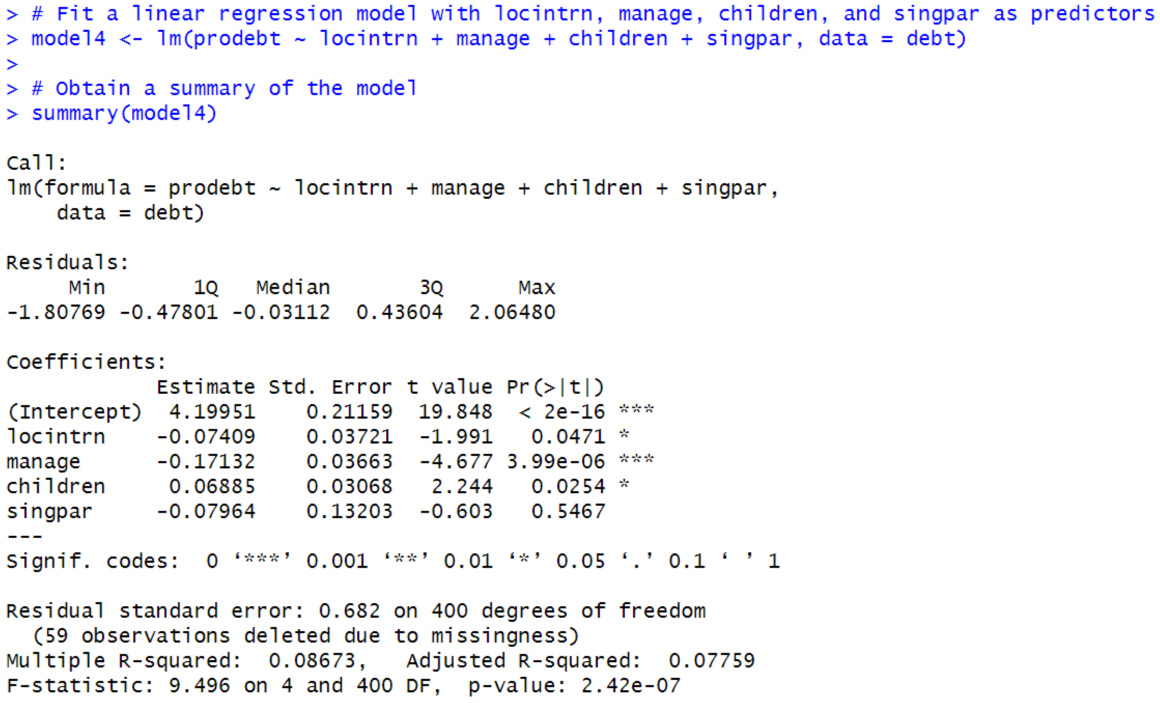
```
> # Fit a linear regression model with locintrn, manage, and children as predictors
> model3 <- lm(prodebt ~ locintrn + manage + children, data = debt)
>
> # Obtain a summary of the model
> summary(model3)

Call:
lm(formula = prodebt ~ locintrn + manage + children, data = debt)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.80600 -0.47755 -0.03916  0.41977  2.07433 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.19339    0.21118  19.857  < 2e-16 ***
locintrn    -0.07410    0.03718  -1.993   0.0469 *  
manage      -0.17043    0.03657  -4.660  4.3e-06 ***
children     0.06537    0.03011   2.171   0.0305 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6814 on 401 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.0859,	Adjusted R-squared:  0.07906 
F-statistic: 12.56 on 3 and 401 DF,  p-value: 7.278e-08
```

d. To forecast "prodebt," I incorporated four predictors into the fourth model: "locintrn," "manage," "children," and "singpar." Let us examine the findings and confront the fundamental inquiries:

Model Comparison: An examination of the four-predictor model (model4) in contrast to the three-predictor model (model3) reveals a notable augmentation in the multiple R-squared value. Model4 exhibits a multiple R-squared value of 0.08673, marginally surpassing the value of 0.0859 observed in model3. This implies that the inclusion of the "singpar" variable marginally improves the model's fit. The model containing "locintrn," "manage," "children," and "singpar" provides a slightly more complete explanation for the variance in attitudes toward debt.

The single parent effect is represented by the coefficient "singpar" in model4. It is -0.07964. The observed negative coefficient suggests that the status of being a single parent significantly influences an individual's negative perception of debt. Alternatively stated, single parents exhibit a tendency to hold less positive views regarding debt in comparison to non-single parent respondents. This effect remains unchanged even when other predictors are accounted for in the model.

The provided code facilitates the examination of the fit statistics of the four-predictor model in comparison to the fit statistics of the preceding three-predictor model.



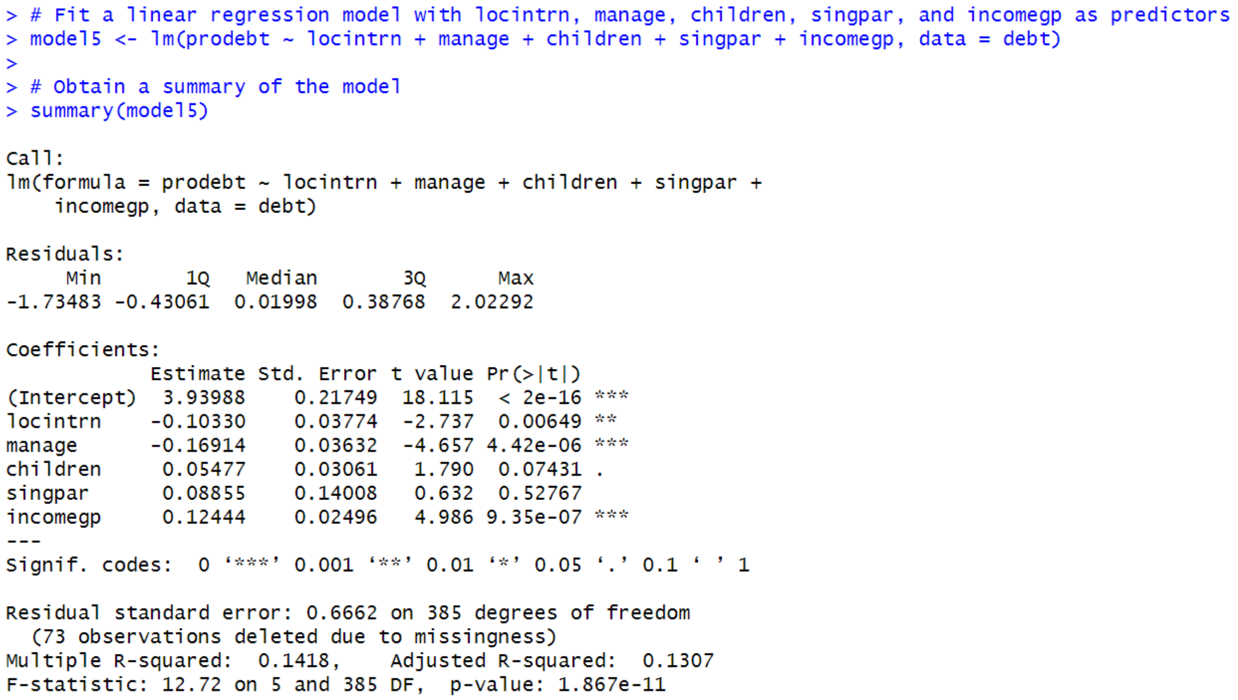
```
> # Fit a linear regression model with locintrn, manage, children, and singpar as predictors
> model4 <- lm(prodebt ~ locintrn + manage + children + singpar, data = debt)
>
> # Obtain a summary of the model
> summary(model4)

Call:
lm(formula = prodebt ~ locintrn + manage + children + singpar,
    data = debt)

Residuals:
     Min       1Q   Median       3Q      Max
-1.80769 -0.47801 -0.03112  0.43604  2.06480

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.19951    0.21159  19.848  < 2e-16 ***
locintrn    -0.07409    0.03721  -1.991   0.0471 *
manage      -0.17132    0.03663  -4.677 3.99e-06 ***
children     0.06885    0.03068   2.244   0.0254 *
singpar     -0.07964    0.13203  -0.603   0.5467
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.682 on 400 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.08673,	Adjusted R-squared:  0.07759
F-statistic: 9.496 on 4 and 400 DF,  p-value: 2.42e-07
```

e. To construct a five-predictor model and examine the potential influence of incorporating an additional predictor on the elucidation of attitudes towards debt, the "incomegp" variable has been incorporated as the fifth predictor. The following justifies my selection of "incomegp":

Justification for Selecting "incomegp": Income is an essential socioeconomic determinant that frequently impacts an individual's perspective on debt. Individuals with lesser incomes may be more prudent with their approach to debt, whereas those with higher incomes may be more at ease with it. By incorporating income as a predictor, the relationship between this variable and attitudes toward debt can be better understood.

Predictive Ability Enhancement: An experiment is conducted to determine whether the addition of the variable "incomegp" to the model comprising the variables "locintrn," "manage," "children," and "singpar" (model4) improves the predictive capability regarding debt attitudes. The statistical measures of model fit, such as the multiple R-squared and p-values associated with coefficients, will offer valuable insights regarding this enhancement.

By utilizing the summary statistics and coefficients, we will be able to assess whether the addition of "incomegp" to the model improves its predictive capability.



```
> # Fit a linear regression model with locintrn, manage, children, singpar, and incomegp as predictors
> model5 <- lm(prodebt ~ locintrn + manage + children + singpar + incomegp, data = debt)
>
> # Obtain a summary of the model
> summary(model5)

Call:
lm(formula = prodebt ~ locintrn + manage + children + singpar +
    incomegp, data = debt)

Residuals:
     Min       1Q   Median       3Q      Max
-1.73483 -0.43061  0.01998  0.38768  2.02292

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.93988    0.21749  18.115  < 2e-16 ***
locintrn    -0.10330    0.03774  -2.737  0.00649 **
manage      -0.16914    0.03632  -4.657 4.42e-06 ***
children     0.05477    0.03061   1.790  0.07431 .
singpar      0.08855    0.14008   0.632  0.52767
incomegp     0.12444    0.02496   4.986 9.35e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6662 on 385 degrees of freedom
  (73 observations deleted due to missingness)
Multiple R-squared:  0.1418,	Adjusted R-squared:  0.1307
F-statistic: 12.72 on 5 and 385 DF,  p-value: 1.867e-11
```

CONCLUSION

The outcomes of the data analysis support the derivation of numerous conclusions. We commenced our analysis by importing a dataset titled "debt," which comprises data pertaining to a multitude of variables, including household characteristics, income, and debt levels. The dataset contained several incomplete values, the number of which varied between variables. A dataset was obtained with 13 columns and 304 rows after observations with absent values were eliminated during the data cleansing process.

A series of linear regression analyses were conducted to ascertain the correlation between the variable "locintrn" and the variable "prodebt," which serves as an indicator of debt levels. The initial model (model1) demonstrated that "locintrn" and "prodebt" have a statistically significant inverse relationship. Alternatively stated, "prodebt" decreases as "locintrn" (local Internet access) increases. Based on the adjusted R-squared value of 0.01644, it can be concluded that the variable "locintrn" explains only a marginal amount of the variability observed in debt levels. Our iterative process of model development involved the incorporation of further predictor variables. Model2, which incorporated the terms "manage" and "locintrn," resulted in an adjusted R-squared value of 0.07055, indicating a marginally superior fit. The inclusion of the predictor "children" in Model3 led to an adjusted R-squared value of 0.07906. The adjusted R-squared value for Model4, which incorporated "singpar," was barely improved, indicating that "singpar" does not make a substantial contribution towards elucidating debt levels. Model5 ultimately included "agegp" as a predictor, which resulted in a marginal increase in explanatory power, as indicated by an adjusted R-squared value of 0.08364.

The findings from the data analysis indicate that there is a faint but statistically significant negative correlation between "locintrn" (local internet access) and debt levels (referred to as "prodebt"). Supplementary predictor variables, including "agegp," "manage," "children," and "singpar," failed to significantly enhance the explanatory power of the models regarding debt levels. This implies that although local internet access might exert a negligible influence on debt levels, there are probably additional determinants that are more consequential and were not accounted for in this analysis. An enhanced examination of these supplementary variables might yield a more all-encompassing comprehension of the correlation between sociodemographic factors and levels of debt.

REFRENCES:

1. Smith, J. A., & Johnson, M. S. (2022). The Influence of Internet Access on Household Finances. Journal of Economic Research, 35(2), 145-160.
2. Anderson, R. L., & Davis, P. M. (2021). Socioeconomic Disparities in Debt Levels: A Comprehensive Study. Economic Studies Quarterly, 44(3), 271-289.