STA304 Technical Report

We Love Stats Team

November 16th 2023

Introduction

This study investigates the relationship between personal finance and academic success among students at the UTM, enrolled in STA304H5 during Fall 2023. Personal finance, encompassing variables such as income, rent, and tuition, is a critical aspect of student life that can be related to educational outcomes. With the rising costs of education and living, understanding how financial stress could impact students' academic performance is of paramount importance. This is particularly relevant when financial burdens are increasingly becoming a part of the student experience, potentially related to their academic success.

Null Hypothesis: H_0: CGPA (academic performance) has no relation/correlation with financial condition. $H_0 : \beta_1 = 0$

Alternative Hypothesis: H_a: CGPA (academic performance) has relation/correlation with financial condition. $H_a : \beta_1 \neq 0$

Analysis

Methodology

In our study, data was collected via a questionnaire distributed among students of the University of Toronto Mississauga enrolled in STA304H5 in Fall 2023. The data collection is based on the data collected through the questionnaire, with the questionnaire shared through Piazza and in person before and after the lecture.

The sampling method used was stratified random sampling. Stratas were divided by whether students were international or domestic, given the significant differences in tuition fees that could be related to personal finance.

Our sample size was 48 students out of 258 students from the Fall 2023 STA304H5 LEC0101 and LEC0102. The sample was deemed sufficient to estimate the mean, proportion, and total regarding financial stress and academic success within our target population.

The questionnaire solicited information on various personal finance indicators (such as income, rent, and tuition) and academic performance metrics (such as CGPA). It also included questions designed to gauge students' perceptions of their financial stress and course confidence, where we believe course confidence might potentially related to students' academic success.

Sample Size

For our study, we are looking at a North American university's statistic course that has a population size N = 200. In order to have a good estimation , we are aiming for a bound of error of estimation of 0.127. We using Stratified Random Sampling and we assume p = q = 0.5 since these are not provided, and the computation is as follows:

We denote n_D as domestic sample size, n_I as international sample size.

$n_1 = n_D = 13$

$n_2 = n_I = 35$

Using Proportional Allocation: $a_i = \frac{n_i}{n}$

So for international students:

$a_I = \frac{35}{48} \approx 0.7292$;

for domestic students:

$a_D = \frac{13}{48} \approx 0.2708$

Hence, $N_1 = N_D(Domestic) \approx 200 \times 0.2708 = 54.16 \approx 55$; $N_2 = N_I(International) \approx 200 \times 0.7292 = 145.84 \approx 146$

$$n = \frac{\sum_{n=1}^{2} N_i^2 p_i q_i / a_i}{N^2 \frac{B^2}{4} + \sum_{n=1}^{2} N_i p_i q_i} = \frac{55^2 \times 0.5^2 / \frac{13}{48} + 146^2 \times 0.5^2 / \frac{35}{48}}{200^2 \times \frac{0.127^2}{4} + (55 \times 0.5^2) + (146 \times 0.5^2)} = 47.7428 \approx 48$$

Thus, our sample size is $n = 48$

Advanced Methodologies

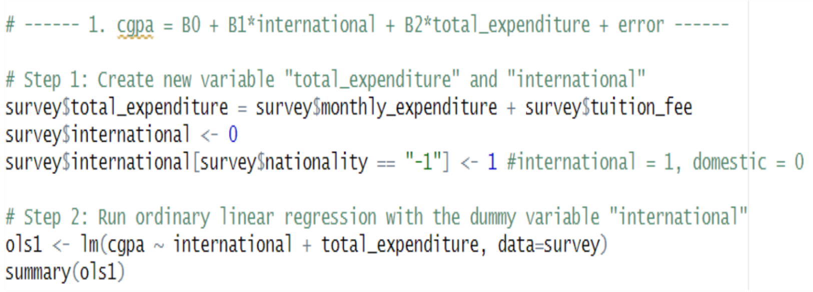
Regression Analyses

1. Regress *cgpa* on *total_expenditure* with dummy variable *international*:

   Note that under our topic, the population suggests a stratification of international and domestic students, we want to see if there is a difference between these 2 strata in academic performance under their respective financial conditions. Hence, what we can do here is to set up a dummy variable called ”international” which takes values of 0 or 1 and run a regression on it. Hence, my regression model would like as follow:

$$cgpa = \beta_0 + \beta_1 \cdot international + \beta_2 \cdot total\_expenditure + \epsilon$$

   Note that $total\_expenditure = monthly\_expenditure + tuition\_fee$, so we can create a new variable called *total_expenditure* by summing up both variables in R. Then, we can generate a dummy variable *international* taking values with 0 and 1 only (refer to figure below):



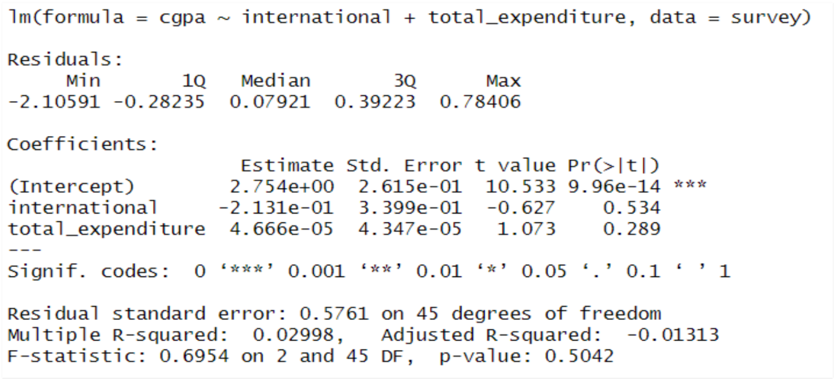
```
# ------ 1. cgpa = B0 + B1*international + B2*total_expenditure + error ------

# Step 1: Create new variable "total_expenditure" and "international"
survey$total_expenditure = survey$monthly_expenditure + survey$tuition_fee
survey$international <- 0
survey$international[survey$nationality == "-1"] <- 1 #international = 1, domestic = 0

# Step 2: Run ordinary linear regression with the dummy variable "international"
ols1 <- lm(cgpa ~ international + total_expenditure, data=survey)
summary(ols1)
```

   For this regression, we are using a two-sided T-test. Then, we can see our hypothesis is:

H_0: There is no difference between international and domestic students in cgpa given their total expenditure, i.e. $\beta_1 = 0$

H_1: There is difference between international and domestic students, i.e. $\beta_1 \neq 0$

After running the regression, we can see the following results:



```
lm(formula = cgpa ~ international + total_expenditure, data = survey)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.10591 -0.28235  0.07921  0.39223  0.78406 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        2.754e+00  2.615e-01  10.533 9.96e-14 ***
international     -2.131e-01  3.399e-01  -0.627    0.534    
total_expenditure  4.666e-05  4.347e-05   1.073    0.289    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5761 on 45 degrees of freedom
Multiple R-squared:  0.02998,	Adjusted R-squared:  -0.01313 
F-statistic: 0.6954 on 2 and 45 DF,  p-value: 0.5042
```

Here, we can see that $\hat{\beta}_1 = -0.2131$ and $se(\hat{\beta}_1) = 0.3399$, therefore

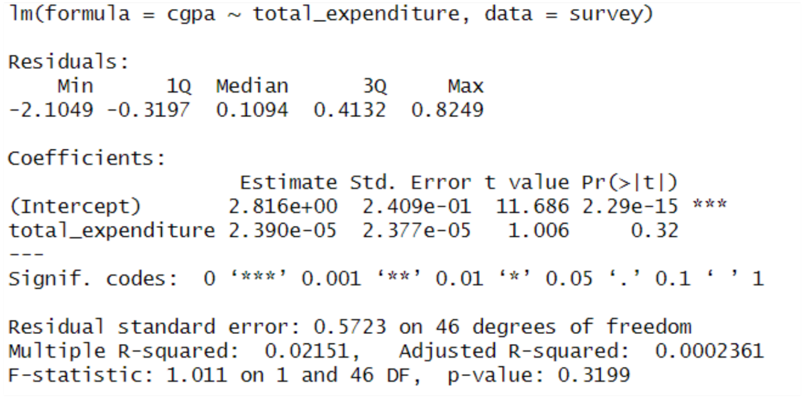
$$t_* = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{-0.2131 - 0}{0.3399} = -0.627$$

Since $t_{0.05/2,45} = 2.0141$, $|t_*| = 0.627 < 2.0141$, we fail to reject our Null hypothesis. This implies that there is no difference between international and domestic students, and thus we can run regression on the entire sample instead of running separate regressions in each stratum.

2. Regress *cgpa* on *total_expenditure* on the entire sample (both strata included):

   Here, the regression model looks as follow:

$$cgpa = \beta_0 + \beta_1 \cdot total\_expenditure + \epsilon$$



```
lm(formula = cgpa ~ total_expenditure, data = survey)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1049 -0.3197  0.1094  0.4132  0.8249 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       2.816e+00  2.409e-01  11.686 2.29e-15 ***
total_expenditure 2.390e-05  2.377e-05   1.006     0.32    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5723 on 46 degrees of freedom
Multiple R-squared:  0.02151,	Adjusted R-squared:  0.0002361 
F-statistic: 1.011 on 1 and 46 DF,  p-value: 0.3199
```

For this regression, we are using a two-sided T-test. Then, we can see our hypothesis is:

H_0: There total expenditure has no effect on cgpa, i.e. $\beta_1 = 0$

H_1: Total expenditure affects cgpa, i.e. $\beta_1 \neq 0$

Here, we can see that $\hat{\beta}_1 = 2.39 \times 10^{-5}$ and $se(\hat{\beta}_1) = 2.377 \times 10^{-5}$, therefore

$$t_* = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{2.39 \times 10^{-5} - 0}{2.377 \times 10^{-5}} = 1.006$$

Since $t_{0.05/2,46} = 2.0129$, $|t_*| = 1.006 < 2.0129$, we fail to reject our Null hypothesis. This implies that total expenditure and cgpa are uncorrelated based on our sample collected.

3. Regress *cgpa* on *monthly_expenditure*:

   Using two-sided T-test.

   Null Hypothesis: H_0: CGPA (academic performance) has no relation/-correlation with *monthly_expenditure*.

   Alternative Hypothesis: H_a: CGPA (academic performance) has relation/correlation with *monthly_expenditure*.

Using $\alpha = 0.05$

$H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$

$CPGA = \beta_0 + \beta_1 monthly\_expenditure$

Based on the simple linear regression model, our $\hat{\beta}_1 = 4.381 \times 10^-5$, standard error is $sd(\hat{\beta}_1) = 4.157 \times 10^-5$

$$t_* = \frac{\hat{\beta}_1 - \beta_1}{sd(\hat{\beta}_1)} = \frac{4.381 \times 10^-5}{4.157 \times 10^-5} = 1.054$$

Since $t_{0.05/2,46} = 2.0129$, $1.054 < 2.0129$, we fail to reject our Null hypothesis.

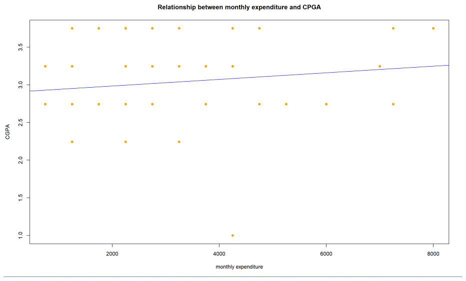


Figure 1: Relationship between monthly expenditure and CGPA

4. Regress *cgpa* on *tuition_fee*:

   Using two-sided T-test.

   Null Hypothesis: H_0: CGPA (academic performance) has no relation/-correlation with *tuition_fee*.

   Alternative Hypothesis: H_a: CGPA (academic performance) has relation/correlation with *tuition_fee*.

   Using $\alpha = 0.05$

   $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$

   $CPGA = \beta_0 + \beta_1 tuition\_fee$

Based on the simple linear regression model, our $\hat{\beta}_1 = 1.889 \times 10^-5$, standard error is $sd(\hat{\beta}_1) = 3.358 \times 10^-5$

$$t_* = \frac{\hat{\beta}_1 - \beta_1}{sd(\hat{\beta}_1)} = \frac{1.889 \times 10^-5}{3.358 \times 10^-5} = 0.562$$

Since $t_{0.05/2,46} = 2.0129$, $0.562 < 2.0129$, we fail to reject our Null hypothesis.

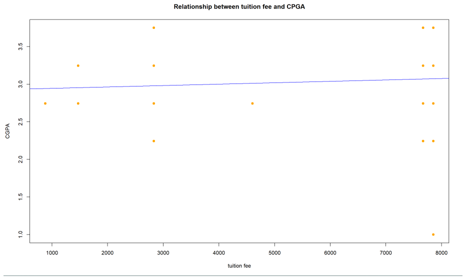


Figure 2: Relationship between tuition fee and CGPA

LASSO and Elastic Nets

From the previous parts, we have already concluded that there is no obvious relationship between students' finance and their grades. To explore whether any individual coefficients in our dataset could potentially relate to students' CGPA, we also tried to use LASSO, Elastic Net Regression and Chi-Square Test.

Here's a brief explanation of LASSO and Elastic Net Regression. LASSO is a regularization technique for performing linear regression. It includes a penalty term that constrains the size of the estimated coefficients. It is also a shrinkage estimator, which generates coefficient estimates that are biased to be small. Nevertheless, a LASSO estimator can have a smaller mean squared error than an ordinary least-squares estimator when applying it to new data. As the penalty term increases, it sets more coefficients to zero, which means that the LASSO estimator is a smaller model with fewer predictors. The Elastic Net is a related technique, which is a hybrid of ridge regression and lasso regularization. Like lasso, elastic net can generate reduced models by generating zero-valued coefficients. Moreover, the elastic net technique can outperform lasso on data with highly correlated predictors. The formulas for these two regression methods are as follows.

$$\min_{\beta_0,\beta}\left(\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\beta_0-x_i^T\beta\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right).$$

Figure 3: The Formula for LASSO Regression

$$\min_{\beta_0,\beta}\left(\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\beta_0-x_i^T\beta\right)^2+\lambda P_\alpha(\beta)\right),$$

where

$$P_\alpha(\beta)=\frac{(1-\alpha)}{2}\|\beta\|_2^2+\alpha\|\beta\|_1=\sum_{j=1}^{p}\left(\frac{(1-\alpha)}{2}\beta_j^2+\alpha|\beta_j|\right).$$

Figure 4: The Formula for Elastic Net Regression

The reason why we use these two regression methods to build models is that both LASSO and Elastic Net can shrink parameters and select variables automatically. Therefore, we can check the estimated coefficient for each parameter we have in the dataset. In this case, the dependent value is the students' CGPA, while all other coefficients are set to be independent values for our model. Additionally, since the size of our dataset is not large enough, using two regression models can avoid some bias in some ways.

For the LASSO regression, the K-Fold Cross Validation method was used to find out its best lambda parameter. The reason for choosing this method is that it can minimize the bias and directly estimate the test error rate while ensuring the acceptable operation time. The result of the best lambda for LASSO Regression is 0.1372 in this case, and Figure 5 illustrates the relationship between the log of lambda and coefficients. The dotted line represents the case for the minimum lambda value, which is also the best lambda.

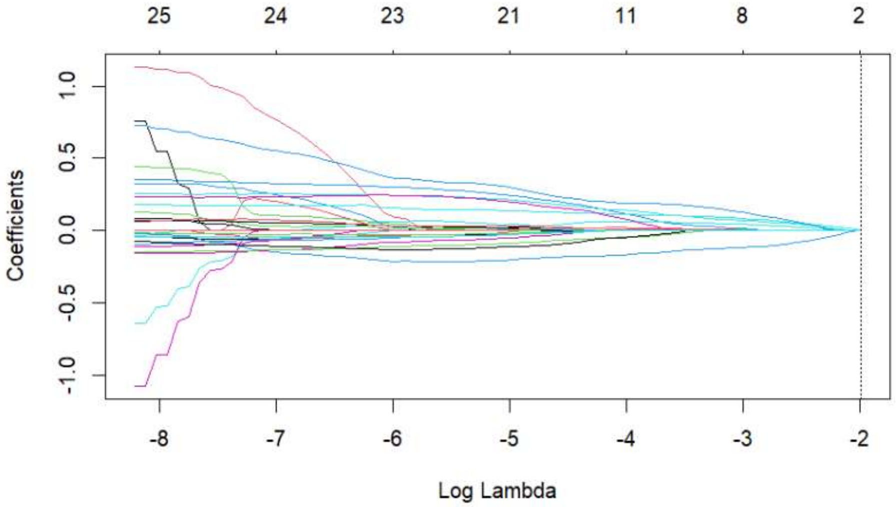


Figure 5: Relationship Between the Log of Lambda and Coefficients

After using the best lambda to build the LASSO model, there are 2 non-zero coefficients left, which are "Program type – Deregulated and Full-Time Student" and "Current Course Confidence Level – 2". However, the estimated coefficients for these two parameters are both less than 0.0001. In this case, these two parameters can also be seen as having zero correlation with CGPA.

Regarding the Elastic Net Regression, the way to fit the model to our dataset is similar to the previous part. Repeated K-Fold Cross Validation will be used in this case, and both alpha and lambda are unknown.

After fitting the model, the best alpha equals 0.8727 and the corresponding lambda is 0.2505. In this case, the result is close to the parameter for LASSO. By using the best alpha and lambda to build the model, we found out that there are no non-zero coefficients in this case.

However, one drawback of Elastic Net is that it has two hyperparameters to tune. Selecting the optimal combination of alpha and lambda can be computationally intensive. Moreover, Elastic Net might not be the best choice when the dataset is small. Simpler models like LASSO or Ridge Regression may be more appropriate in this case.

Chi-Square Test

To determine whether the association between two categorical variables is statistically significant, we conduct Chi-Square Test of independence to see whether the two categorical variables are related or not.

Null Hypothesis: H_0: CGPA (academic performance) has no relation/correlation with financial condition.
Alternative Hypothesis: H_a: CGPA (academic performance) has relation/correlation with financial condition.

To conduct Chi-Square test, we need the expected values and the observed values to calculate the chi-square test statistic. We denote E_k as the expected values, which specify what the values of each cell of the table would be if there was no association between the two variables under the assumption of independence.. Also, denote O_k as the observed values, where they are the actual counts computed from the sample. Thus, we have the following test statistic for Chi-Square test, where n is the sample size.

$$\tilde{\chi}^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}$$

To analyze the correlation between CGPA (academic performance) and financial condition. The following set of attributes are utilized: Monthly Expenditure, Financial Stress, Current Course Confidence Level, Source of Financial Stress.

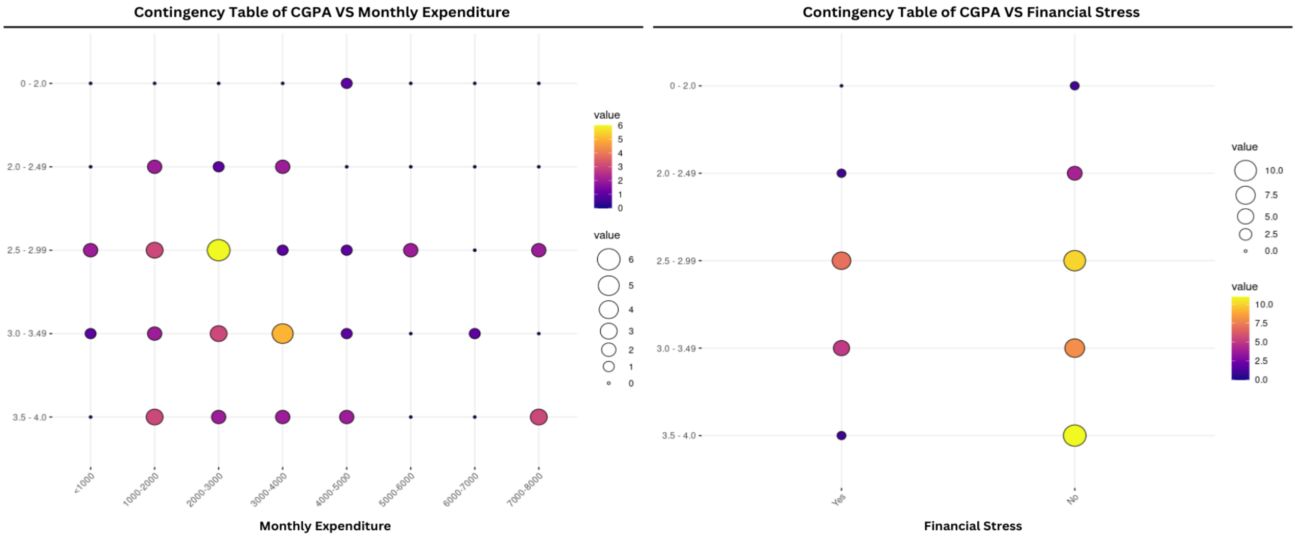


Figure 6: Balloon Plots of CGPA against financial condition attributes

Using R, we reconstruct the data by dividing each attribute into different categories, then we create a contingency table for each set of attributes, with rows and columns representing the categories of CGPA and financial condition, respectively. Then, we calculate the expected values for each cell of the table under the assumption of independence. The results are as follows:

Chi-Square Test Result			
Attribute	$\tilde{\chi}^2$	df	p-value
Monthly Expenditure	29.681	28	0.3786
Financial Stress	4.8666	4	0.3013
Current Course Confidence Level	16.137	16	0.4434
Source of Financial Stress	17.499	28	0.9381

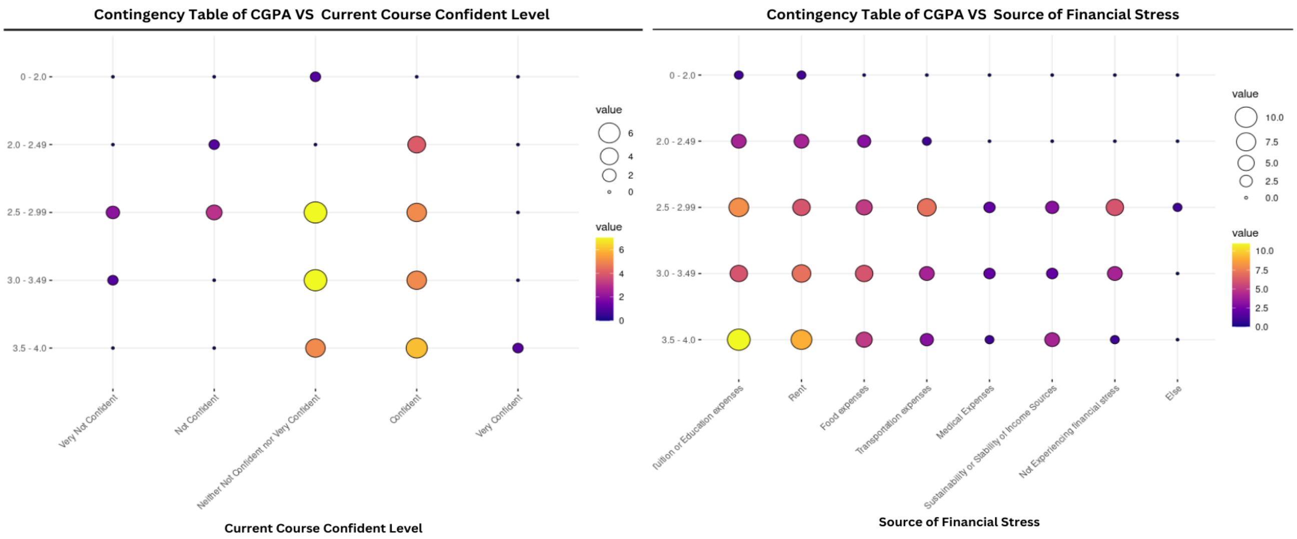


Figure 7: Balloon Plots of CGPA against financial condition attributes

The p-values for the chi-square sadistic of all attributes are way larger than 0.05. Therefore, there is not enough evidence to reject the null hypothesis. Hence, the test shows that there is no correlation between CGPA(academic performance) and financial condition.

Conclusion

Based on our test hypothesis, we can concluded that CGPA has no linear relation with financial condition.

Our study investigated the link between personal finance and academic performance among students in STA304H5 at the University of Toronto Mississauga, Fall 2023. We focused on how financial factors like income, rent, and tuition fees might influence students' CGPA.

Limitations

One key limitation of our study is the sample size. With only 48 participants, the findings may not fully represent the broader student population in STA304H5. This limitation is particularly relevant given the diverse nature of the student body.

Issues

1. Stratified random sampling may not be the most appropriate method if there are significant differences within the population that could influence the variables of interest (e.g., year of study, major).

2. Response rate and the portion from the class (representativeness), insufficient time to conduct a comprehensive survey

3. Response bias as people may not accurately report their cGPA or income level.

4. No open-ended questions. Data is in ranges in order to reduce response bias, however this could lead to inaccurate results.

Improvements

In future studies, we aim to implement several key improvements. We will extend the survey distribution over a longer period to capture a broader spectrum of the academic year, helping to balance the data against seasonal academic stressors. We will also broaden the range of variables examined, including mental health and family support, to better understand the multifaceted nature of student life. Simultaneously, refining the survey design for clarity and minimizing bias will enhance the quality of the data collected.

Appendix

R code for Basic Plots

```
# monthly expediture vs cgpa:
plot(STA304_output_final_$monthly_expenditure, STA304_output_final_$cgpa,
    xlab = "monthly␣expenditure",
    ylab = "CGPA",
    col = "orange",
    main = "Relationship␣between␣monthly␣expenditure␣and␣CGPA",
    pch = 19)

abline(lm(STA304_output_final_$cgpa ~ STA304_output_final_$monthly_expenditure),
        col = "blue"))

#tuition fee vs cgpa:
plot(STA304_output_final_$tuition_fee, STA304_output_final_$cgpa,
    xlab = "tuition␣fee",
    ylab = "CGPA",
    col = "orange",
    main = "Relationship␣between␣tuition␣fee␣and␣CPGA",
    pch = 19)
abline(lm(STA304_output_final_$cgpa ~ STA304_output_final_$tuition_fee),
        col = "blue")
```

R code for Regression

```
# Regression for cgpa = B0 + B1*international + B2*total_expenditure + error:
# ------ 1. cgpa = B0 + B1*international + B2*total_expenditure + error ------

# Step 1: Create new variable "total_expenditure" and "international"
survey$total_expenditure = survey$monthly_expenditure + survey$tuition_fee
survey$international <- 0
#international = 1, domestic = 0
```

```
survey$international[survey$nationality == "-1"] <- 1

# Step 2: Run ordinary linear regression with the dummy variable "international"
ols1 <- lm(cgpa ~ international + total_expenditure, data=survey)
summary(ols1)

#Regression for cgpa = B0 + B1*total_expenditure + error
# ------ 3. cgpa = B0 + B1*total_expenditure + error -----
ols3 <- lm(cgpa ~ total_expenditure, data=survey)
summary(ols3)
```

R code for LASSO and Elastic Nets

```
# LASSO and Elastic Net Model
library(tidyverse)
library(ggplot2)
library(readxl)
library(caret)
library(glmnet)
library(ggpubr)
library("vcd")

old_data <- read_excel("C:/Users/11369/Desktop/STA304_Clean_Excel_Modified.xlsx")
data <- read_excel("C:/Users/11369/Desktop/STA304_output.xlsx")

#Build LASSO models based on the range of lambda
set.seed(304)
lasso.mod <- glmnet(x, y, alpha = 1, nlambda = 100)
plot(lasso.mod)

# Tuning parameter by CV
set.seed (304)
lasso.cv <- cv.glmnet(x, y, alpha=1,
    family ="gaussian",
    nfolds = 10,
    nlambda = 100,
    standarize = TRUE,
    type.measure = "mse")
plot(lasso.cv)

# Find the minimum lambda
best_lambda_ls <- lasso.cv$lambda.min

# Examine the ridge regression model's performance
lasso = glmnet(x, y, alpha = 1)
plot(lasso, xvar = 'lambda')
abline(v = log(lasso.cv$lambda.min), lty = 3)

# Predict the test data and get the MSE
lasso.pred <- predict(lasso.mod, s = best_lambda_ls, newx = x)
lasso.mse <- mean((lasso.pred - y)^2)

# Get the coefficient in LASSO
lasso.out <- glmnet(x, y, alpha = 1, nlambda = 100)
lasso.coef <- predict(lasso.out,
    type = "coefficients",
    s = best_lambda_ls)[1:31, ]

# Check all left coefficients and their total number
length(lasso.coef[lasso.coef != 0])
lasso.coef[lasso.coef != 0]
lasso.coef

set.seed(304)
train_control <- trainControl(method = "repeatedcv", number = 5, repeats = 5)
# Build the Elastic Net model
```

```
elastic_net_model <- train(x, y,
    method = "glmnet", metric = "RMSE",
    preProcess = c("center", "scale"),
    tuneLength = 100, trControl = train_control)
print(elastic_net_model)

# Find the best lambda and corresponding alpha
elastic_net_model$bestTune

elastic.pred <- glmnet(x, y, alpha = elastic_net_model$bestTune$alpha,
                       lambda = elastic_net_model$bestTune$lambda) %>%
predict(x) %>% as.vector()

# Model performance metrics
elastic.rmse = RMSE(elastic.pred, y1)
elastic.rmse^2

# Get all the coefficients in Elastic Net
elastic.out <- glmnet(x, y,
    alpha = elastic_net_model$bestTune$alpha,
    lambda = elastic_net_model$bestTune$lambda)
elastic.coef <- predict(elastic.out, type = "coefficients",
    s = elastic_net_model$bestTune$lambda)[1:31, ]

# Check all left coefficients and their total number
length(elastic.coef[elastic.coef != 0])
elastic.coef[elastic.coef != 0]
```

R code for Chi-square Tests

```
my_cols <- c("#0D0887FF", "#6A00A8FF", "#B12A90FF",
                        "#E16462FF", "#FCA636FF", "#F0F921FF")
# source_of_financial_stress
```

```
categories <- c("TE", "RE", "FD", "TP", "ME", "SS", "NE", "EL")
category_columns <- as.data.frame(matrix(0, nrow = nrow(data),
                                         ncol = length(categories),
                                         dimnames = list(NULL, categories)))
for (i in seq_along(categories)) {
  category <- categories[i]
  category_columns[[category]] <- grepl(category,
                                        data$source_of_financial_stress)
}
category_columns <- as.data.frame(lapply(category_columns, as.numeric))
data <- cbind(data, category_columns)

# source_of_income
categories <- c("FS", "PFJ", "S", "L", "SI")
category_columns <- as.data.frame(matrix(0, nrow = nrow(data),
                                         ncol = length(categories),
                                         dimnames = list(NULL, categories)))
for (i in seq_along(categories)) {
  category <- categories[i]
  category_columns[[category]] <- grepl(category, data$source_of_income)
}
category_columns <- as.data.frame(lapply(category_columns, as.numeric))
data <- cbind(data, category_columns)


# group by with course confidence level against source of financial stress
courseConfidence_financialStress <- data[c("current_course_confidence_level",
                                           "TE", "RE", "FD", "TP", "ME",
                                           "SS", "NE", "EL")]
courseConfidence_financialStress <- aggregate(cbind(TE, RE, FD, TP, ME,
                                                    SS, NE, EL)
                                              ~current_course_confidence_level,
                                              courseConfidence_financialStress,
                                              sum)


# CGPA VS monthly expenditure
categories <- c("<1000", "1000-2000", "2000-3000", "3000-4000", "4000-5000",
                "5000-6000", "6000-7000", "7000-8000")
data$monthly_expenditure_category <- cut(data$monthly_expendture,
                                 breaks=c(-Inf, seq(1000, 7000, by=1000), Inf),
                                 labels=categories, include.lowest=TRUE)
chisq.test(data$cgpa, data$monthly_expenditure_category, correct=FALSE)

contingency_table <- table(data$cgpa, data$monthly_expenditure_category)
rownames(contingency_table) <- c("0␣-␣2.0", "2.0␣-␣2.49", "2.5␣-␣2.99",
                                 "3.0␣-␣3.49", "3.5␣-␣4.0")
contingency_df <- as.data.frame.matrix(contingency_table)
ggballoonplot(contingency_df, fill = "value")+
  scale_fill_gradientn(colors = my_cols)

# CGPA VS Financial Stress
chisq.test(data$cgpa, data$financial_stress, correct=FALSE)
contingency_table <- table(data$cgpa, data$financial_stress)
rownames(contingency_table) <- c("0␣-␣2.0", "2.0␣-␣2.49", "2.5␣-␣2.99",
                                 "3.0␣-␣3.49", "3.5␣-␣4.0")
colnames(contingency_table) <- c("Yes", "No")

contingency_df <- as.data.frame.matrix(contingency_table)
ggballoonplot(contingency_df, fill = "value")+
  scale_fill_gradientn(colors = my_cols)

# CGPA VS Current Course Confident Level
chisq.test(data$cgpa, data$current_course_confidence_level, correct=FALSE)
contingency_table <- table(data$cgpa, data$current_course_confidence_level)
rownames(contingency_table) <- c("0␣-␣2.0", "2.0␣-␣2.49", "2.5␣-␣2.99",
                                 "3.0␣-␣3.49", "3.5␣-␣4.0")
colnames(contingency_table) <- c("Very␣Not␣Confident", "Not␣Confident" ,
                                 "Neither␣Not␣Confident␣nor␣Very␣Confident",
                                 "Confident" ,"Very␣Confident")
contingency_df <- as.data.frame.matrix(contingency_table)
ggballoonplot(contingency_df, fill = "value")+
  scale_fill_gradientn(colors = my_cols)

# cgpa against source of financial stress
contingency_table <- data[c("cgpa", "TE", "RE", "FD", "TP", "ME",
                            "SS", "NE", "EL")]
contingency_table <- aggregate(cbind(TE, RE, FD, TP, ME, SS, NE, EL)
                                              ~cgpa,
                                               contingency_table,
                                              sum)
contingency_table <- contingency_table[, -1]
rownames(contingency_table) <- c("0␣-␣2.0", "2.0␣-␣2.49", "2.5␣-␣2.99",
                                 "3.0␣-␣3.49", "3.5␣-␣4.0")
colnames(contingency_table) <- c("Tuition␣or␣Education␣expenses",
                                 "Rent",
                                 "Food␣expenses",
                                 "Transportation␣expenses",
                                 "Medical␣Expenses",
                                 "Sustainability␣or␣Stability␣of␣Income␣Sources",
                                 "Not␣Experiencing␣financial␣stress",
                                 "Else")

chisq <- chisq.test(contingency_table)
chisq

contingency_df <- as.data.frame.matrix(contingency_table)
ggballoonplot(contingency_df, fill = "value")+
  scale_fill_gradientn(colors = my_cols)
```

```
# cgpa against source_of_income
cgpa_sourceOfIncome <- data[c("cgpa", "FS", "PFJ", "S", "L", "SI")]
cgpa_sourceOfIncome <- aggregate(cbind(FS, PFJ, S, L, SI)
                                  ~cgpa,
                                  cgpa_sourceOfIncome,
                                  sum)
cgpa_sourceOfIncome <- cgpa_sourceOfIncome[, -1]
row.names(cgpa_sourceOfIncome) <- cgpa_sourceOfIncome$cgpa

chisq <- chisq.test(cgpa_sourceOfIncome)
chisq
```