AMS572: Project Report

Determinations of Academic Success

Brandon Bernstein, Pushpa Ugma Vemula, Saiteja Kalam, Sanath Kumar Bajjuri

1. Introduction

The landscape of higher education has witnessed profound changes, spurred by globalization and a diversification of student demographics. Amidst this evolving educational milieu, understanding the determinants of academic success has become paramount for educators, policymakers, and researchers alike. This project, rooted in the heart of this context, aims to delve into the complex dynamics of academic achievement, focusing on students from various countries. The pivotal objective is to unravel the intricate interplay of factors that sculpt educational outcomes, thereby providing insightful contributions to the field of educational research.

The pursuit of academic success is not merely an individual endeavor but is deeply intertwined with broader societal and demographic factors. Among these, gender has emerged as a significant variable. Extensive research has demonstrated that gender differences in educational achievement are pervasive, yet the underlying causes and implications of these disparities remain a topic of intense debate. By examining the relationship between gender and academic success.

Furthermore, the project transcends mere correlation by adopting a robust analytical approach through the use of Generalized Linear Models (GLMs). GLMs offer a flexible framework for analyzing data with varied distributions, making them particularly suited for educational data, which often encompasses binary outcomes, count data, and continuous measurements. By incorporating a range of demographic and socio-economic variables into the GLM, this study endeavors to construct a comprehensive model of academic outcomes.

Setup

Required Packages

```
library("ggplot2")
library("vcd")
library("dplyr")
library("knitr")
library("caret")
library("leaps")
library("corrplot")
library("tidyverse")
library("mice")
```

```r
library("ranger")
library("patchwork")
library("gridExtra")
```

2. Exploratory Data Analysis

Data

The dataset originates from a higher education institution and is compiled from various separate databases. Each instance (each row) represents a student, it encompasses data about students enrolled in diverse undergraduate programs, including fields like agronomy, design, education, nursing, journalism, management, social service, and technology. This dataset captures details available at the point of student admission, covering their academic history, demographic background, and socio-economic factors. It also includes records of their academic achievements at the conclusion of their first and second semesters. The primary use of this data is in developing classification models aimed at predicting student attrition and academic success. The classification task is divided into three categories, with a notable imbalance favoring one of the categories.

```r
# Read data from a CSV file
data <- read.csv("AcademicSuccessData.csv")

# Convert the 'Course' column in the 'data' dataframe to a factor.
data$Course <- as.factor(data$Course)
```

The dataset comprises of 4424 instances (rows) and 36 features (columns). Columns listed below are important columns of data:

`Student_ID` - Integer - Uniquely identify each student

`Marital_status` - Categorical - Describes marital status of student

`Course` - Categorical - Describes course in which student is enrolled

`Attendance` - Categorical - Describes whether student attendance is in daytime or evening

`Previous_qualification` - Categorical - Describes highest education level attained by student

`Previous_qualification_grade` - Categorical - Describes grade achieved by student in his previous qualification

`Nationality` - Categorical - Describes the nationality of the student

`Mother_qualification` - Categorical - Describes highest education level attained by mother of the student

`Father_qualification` - Categorical - Describes highest education level attained by father of the student

`Mother_occupation` - Categorical - Describes occupation of mother of the student

`Father_occupation` - Categorical - Describes occupation of father of the student

`Admission_grade` - Decimal - Describes the grade achieved by student in previous qualification

Displaced - Categorical - Describes if student is displaced

Educational_special_needs - Categorical - Describes if student have special education needs in reading, writing, speaking or understanding

Debtor - Categorical - Describes if student is on education loan to complete pursue the degree

Tuition_fees_up_to_date - Categorical - Describes if student is paying tuition fee on time

Gender - Categorical - Describes the gender of the student

Scholarship_holder - Categorical - Describes if student is receiving any cholarship from the university

Age_at_enrollment - Numeric - Describes age of the student at the time of enrollment

International - Categorical - Describes if the student is an international student at university

Curricular_units_Sem1_grade - Decimal - Describes the grade average of the student in 1^{st} semester

Curricular_units_Sem2_grade - Decimal - Describes the grade average of the student in 2^{nd} semester

Unemployment_rate - Decimal - Unemployment rate in the country of student nationality

GDP - Decimal - GDP of the country of student nationality

Target - Categorical - Describes if the student is a dropout or graduated or still enrolled

```r
# Count the total number of missing values in the 'data' dataframe.
sum(is.na(data))
```

```
## [1] 0
```

There were no missing values in the dataset.

Key facts based on descriptive statistics

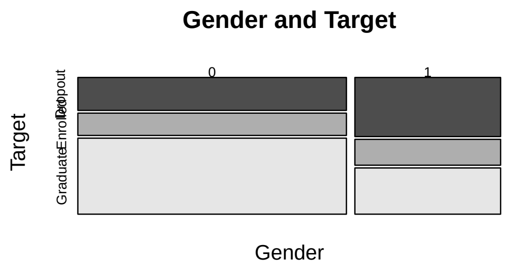
```
##                    Grade_Type         Mean Standard_Deviation
## 1 Previous Qualification Grade 1.326133e+02          13.188332
## 2              Admission Grade 1.269781e+02          14.482001
## 3  Curricular Units Sem1 Grade 1.064082e+01           4.843663
## 4  Curricular Units Sem2 Grade 1.023021e+01           5.210808
## 5            Unemployment Rate 1.156614e+01           2.663850
## 6                          GDP 1.968807e-03           2.269935
```

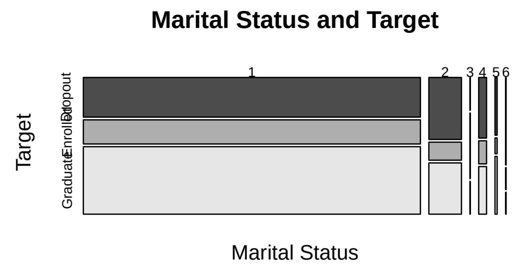
The average **Previous_qualification_grade** was around `132.61` with a standard deviation of approximately `13.2`, indicating a moderate range variability of academic backgrounds among students.

The average **Admission_grade** was around `126.97` with a standard deviation of approximately `14.48`, indicating a high range of variability academic backgrounds among students.

The average grades for the first and second semesters **Curricular_units_Sem1_grade** and **Curricular_units_Sem2_grade** were similar, but the standard deviation of semester-2 grades is noticebly higher. This suggests a varied academic performance across students.

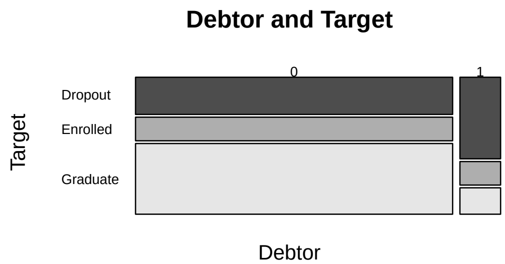
Some interesting plots

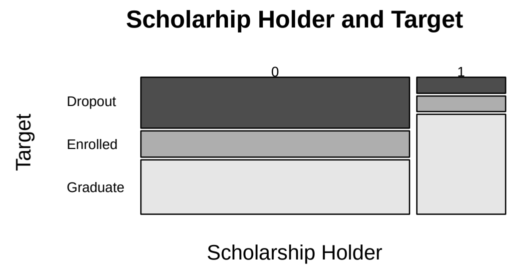




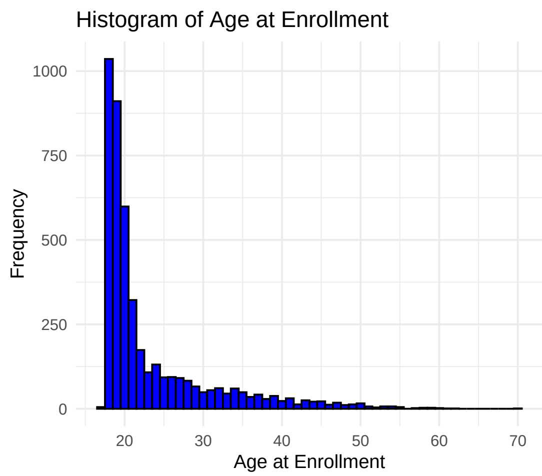
1 - Male, 0 - Female

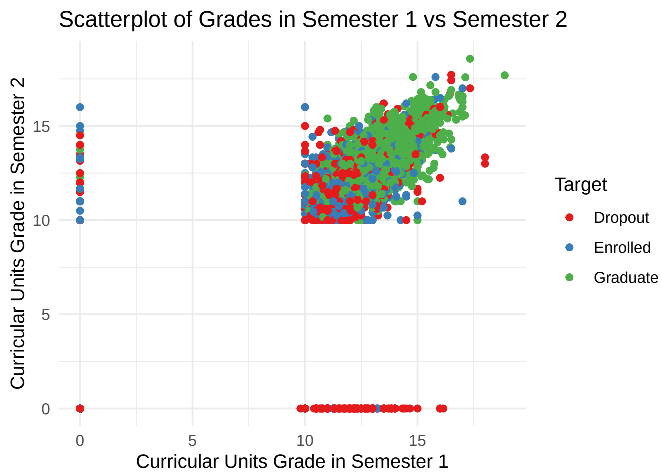
1 - Single, 2 - Married, 3 - Widower, 4 - Divorced, 5 - Facto Union, 6 - Legally Separated





0 - No, 1 - Yes





3. Hypotheses, Methodology and Testing

Hypothesis - 1

`Null hypothesis - H0`: There is no significant relation (independent) between gender of a student and their academic success.

`Alternative hypothesis - Ha`: There is a significant relation (dependent) between gender of a student and their academic success.

To investigate the relationship between the **`Gender`** and **`Target`** columns, which are both categorical, a Chi-square test would be appropriate. Hence, we will employ χ^2 as our test statistic. The chosen significance level, α, is `0.05`

```
# Create a new binary column 'dropout' in the 'data' dataframe, where
# 'Dropout' in 'Target' is marked as 1, and all others as 0.
data$dropout <- ifelse(data$Target == "Dropout", 1, 0)
```

Created a new column **`dropout`** with integer encoding of the **`Target`** such that **`dropout=`** `1` when student's **`Target`** variable is 'dropout', **`dropout=`** `0` otherwise.

There are 1421 dropouts and 3003 students who are graduated or still enrolled.

```
##         Dropout
## Gender   Grad/Enrolled Dropout
##   Female          2148     720
##   Male             855     701
```

`Assumptions:`

- The data in the cells should be frequencies, or counts of cases rather than percentages or some other transformation of the data.
- The levels categories of the variables are mutually exclusive. That is, a particular subject fits into one and only one level of each of the variables.
- Each subject may contribute data to one and only one cell in the χ^2.
- The study groups must be independent.
- There are 2 variables, and both are measured as categories, usually at the nominal level.
- Large sample sample size with small percentage of expected cell counts less than 5

Since all the assumptions for a ^{2} are satisfied, we proceed with the test.

The degrees of freedom for a χ^2 test is, $df = (r - 1) \times (c - 1)$

where r is the number of categories in one variable, and c is the number of categories in another. In **`Gender`** there are two categories (`0` - Male, `1` - Female), but in **`Target`** we will consider only two categories (`1` - dropout, `0` - not_dropout). Hence the value of $df = (2 - 1) \times (2 - 1)$ which is, 1.

```r
df <- 1
alpha <- 0.05
# Calculate the critical value for a chi-squared test with a
# significance level of 0.05 and 1 degree of freedom.
critical_value <- qchisq(1 - alpha, df)
```

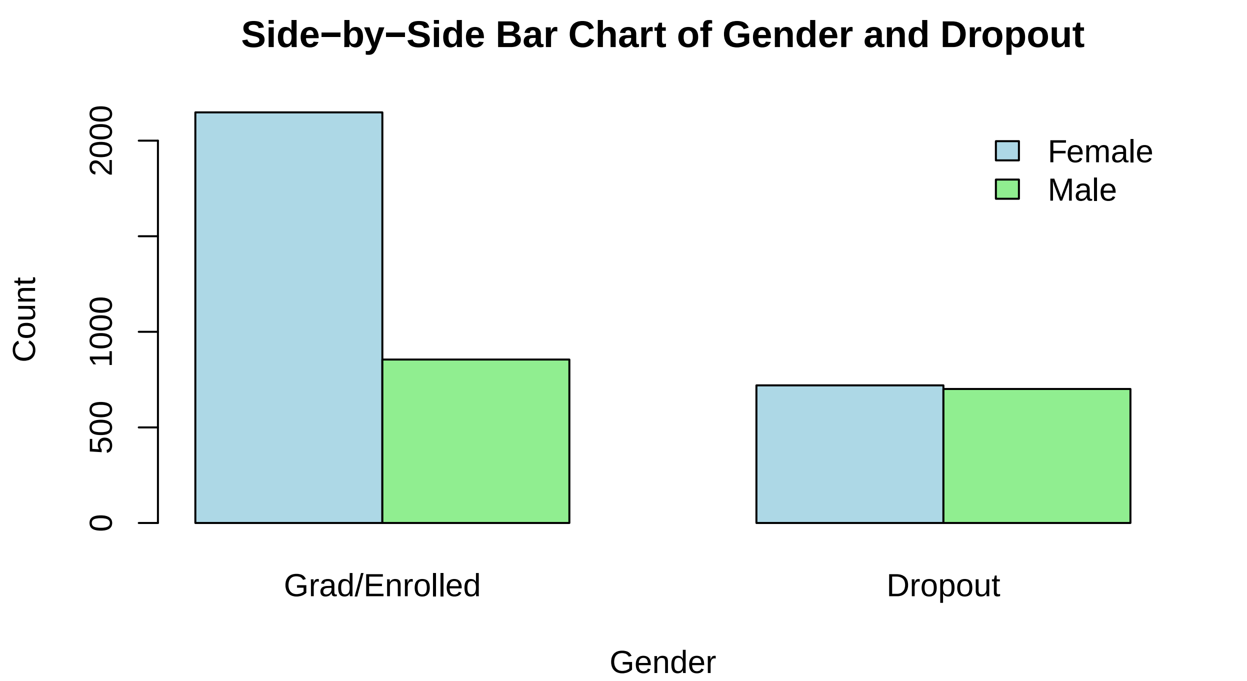
The critical region, C_α is

At the significance level, $\alpha = 0.5$, we reject the `H0` in favor of `Ha` if $\chi^2 > 3.8415$

```r
# Perform a chi-squared test on the data in 'contingency_table'.
ind_test_g_d <- chisq.test(contingency_table)
print(ind_test_g_d)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table
## X-squared = 183.16, df = 1, p-value < 2.2e-16
```

Since, $\chi^2 (=183.16) > 3.8415$ we reject the null hypothesis `H0` in favor of `Ha` and conclude that there exists a significant dependence of **`Target`** column on **`Gender`** column.



Despite there being an overwhelmingly higher amount of Females enrolled/graduated compared to Males, the number of dropouts are the same. Males have a higher association with dropping out than females.

Hypothesis - 2

`Null hypothesis - H0` : The likelihood of a student dropping out is not impacted by economic climate when courses, gender, and grades are equal.

`Alternative hypothesis - Ha` : The likelihood of a student dropping out is impacted by economic climate when courses, gender, and grades are equal.

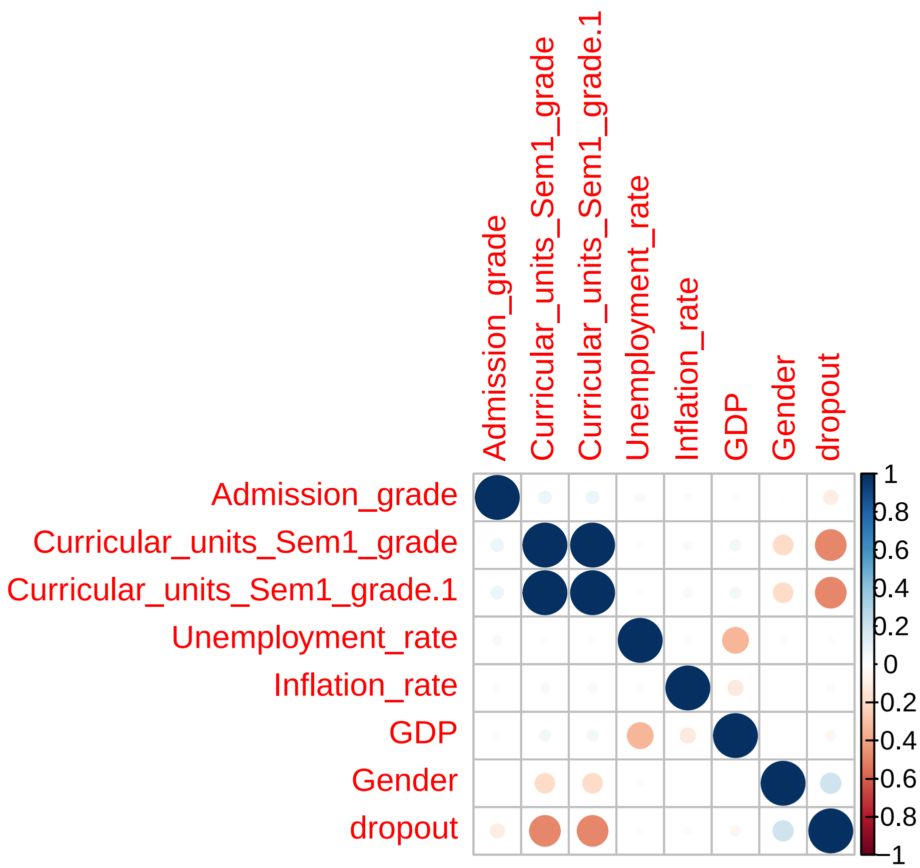
`Significance Level` - $(\alpha) = 0.01$

To identify the influence of independent variable on dependent variable we use `Generalized Linear Model (GLM)`. A Generalized Linear Model (GLM) is a flexible generalization of ordinary linear regression that allows for dependent variables that have error distribution models other than a normal distribution. GLM generalizes linear regression by allowing the linear model to be related to the dependent variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

`Critical region`, for the test, is defined based on the p-values of the coefficients in the logistic regression model. If the p-value for any of the coefficients (GDP, Unemployment rate and Inflation rate) is less than 0.01, we reject null hypothesis.

`Assumptions` - Data is independent with no outlier's. There is no multi-colinearity and there exists a linear relationship between the variables used and the logit function.

Checking multi-colinearity: Due to semester grades being co-linear, only one will be used in the model. Besides that no co linearity is detected.



```r
# Fit a logistic regression model (logit_model) predicting 'dropout'
# from various predictors.
# The model uses a binomial family, appropriate for binary response
# data, on the 'data' dataset.
logit_model <- glm(dropout ~ Course + Admission_grade +
                     Curricular_units_Sem1_grade+ Unemployment_rate +
                     Inflation_rate + GDP + Gender,
                   family = binomial(), data = data)

# Summarize the fitted logistic regression model.
model_summary <- summary(logit_model)
print(model_summary)
```

```
##
## Call:
## glm(formula = dropout ~ Course + Admission_grade + Curricular_units_Sem1_grade +
##     Unemployment_rate + Inflation_rate + GDP + Gender, family = binomial(),
##     data = data)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  4.315143   0.808690   5.336 9.50e-08 ***
## Course171                   -3.992147   0.729216  -5.475 4.39e-08 ***
## Course8014                  -1.313382   0.715365  -1.836  0.06636 .
## Course9003                  -1.233787   0.715483  -1.724  0.08463 .
## Course9070                  -1.596564   0.716290  -2.229  0.02582 *
## Course9085                  -1.231680   0.710478  -1.734  0.08299 .
## Course9119                  -1.085860   0.720259  -1.508  0.13166
## Course9130                  -0.235117   0.721497  -0.326  0.74452
## Course9147                  -1.457458   0.704846  -2.068  0.03866 *
## Course9238                  -2.067548   0.713219  -2.899  0.00374 **
## Course9254                  -1.164978   0.709236  -1.643  0.10047
## Course9500                  -1.976086   0.704172  -2.806  0.00501 **
## Course9556                  -1.099791   0.741903  -1.482  0.13824
## Course9670                  -1.076576   0.708320  -1.520  0.12854
## Course9773                  -1.171605   0.706998  -1.657  0.09749 .
## Course9853                  -0.343165   0.713055  -0.481  0.63033
## Course9991                  -0.855154   0.709674  -1.205  0.22820
## Admission_grade             -0.009053   0.002865  -3.160  0.00158 **
## Curricular_units_Sem1_grade -0.291481   0.012878 -22.633  < 2e-16 ***
## Unemployment_rate            0.029320   0.015828   1.852  0.06396 .
## Inflation_rate               0.005808   0.028146   0.206  0.83651
## GDP                         -0.017262   0.018370  -0.940  0.34740
## Gender                       0.620201   0.086590   7.163 7.92e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5554.5  on 4423  degrees of freedom
## Residual deviance: 4127.3  on 4401  degrees of freedom
## AIC: 4173.3
##
## Number of Fisher Scoring iterations: 5
```

`Conclusions:` None of the economic KPI's were significant. Unemployment rate, Inflation rate, and GDP all had p values greater than 0.05. The variables that had significant relationships with dropout were Gender, admission grades and the students semester 1 grade (P values < 0.01). Specific courses also showed significant differences in the likelihood of dropping out, notably course 171 had the lowest odds of dropping out.

4. Effects of missing values

Now, we will investigate the effect of missing values on data analysis for the following scenarios:

- Data missing completely at random (MCAR)
- Data missing not at random / non-ignorable missing values (MNAR)

Data missing completely at random (MCAR) for hypothesis I

Data can be considered Missing Completely at Random (MCAR) when the likelihood of data being missing is the same for all the observations. In other words, the missingness of data is entirely unrelated to the observed data or any of the unobserved data.

Here are some criteria to consider data as MCAR:

`No Systematic Differences:` There are no systematic differences between the missing values and the observed values. This means that the missing data points are a random subset of the data.

`No Relationship with Other Variables:` The probability that a value is missing is not related to the value of the variable itself or to the value of any other variables. For instance, if you're looking at test scores and gender, the missingness of test scores should not be related to gender or the scores themselves.

`Random Dropouts:` In longitudinal studies, if participants drop out of the study for reasons unrelated to the study or their characteristics, the missing data due to dropout can be considered MCAR.

`Missingness Due to Random Events:` If the missingness is due to a random event (like a survey respondent accidentally skipping a question) and not due to any inherent characteristic of the respondent or the survey design, then it can be considered MCAR.

We don't have any missing values in our dataset, let's simulate a dataset with data missing at random.

```r
MCAR_Chi_Test <- function(prop_missing){
  # Set a seed for reproducibility of random processes
  set.seed(123)

  # Create copies of the Gender and dropout columns
  data$Gender_MCAR <- data$Gender
  data$Dropout_MCAR <- data$dropout

  # Calculate the number of values to be set as
  # missing based on the specified proportion
  size <- round(prop_missing * nrow(data))

  # Randomly select indices to be set as
  # missing for both Gender and Dropout columns
  missing_indices_gender <- sample(1:nrow(data), size = size)
  missing_indices_dropout <- sample(1:nrow(data), size = size)

  # Set the selected values to NA (missing) in the Gender_MCAR
  # and Dropout_MCAR columns
  data$Gender_MCAR[missing_indices_gender] <- NA
  data$Dropout_MCAR[missing_indices_dropout] <- NA

  # Create a contingency table with the modified columns
  contingency_table_MCAR <- table(data[,c('Gender_MCAR','Dropout_MCAR')])
  names(dimnames(contingency_table_MCAR)) <- c('Gender_MCAR', 'Dropout_MCAR')
  colnames(contingency_table_MCAR) <- c("Grad/Enrolled", "Dropout")
  rownames(contingency_table_MCAR) <- c("Female","Male")

  # Prepare for chi-squared distribution plot
  x_values <- seq(0, critical_value + 10, by = 0.1)
  chi_sq_df <- data.frame(x = x_values, y = dchisq(x_values, df))

  # Perform chi-squared test of independence on the contingency table
  ind_test_g_d <- chisq.test(contingency_table_MCAR)

  # Format and print the result
  result <- sprintf("For %s%% of missing values, the chi-square value is %f"
                    , prop_missing * 100, ind_test_g_d$statistic)
  print(result)
}
```

The above function, modifies the dataset by adding new columns **`Gender_MCAR`** and **`Target_MCAR`** for variable percentages of missing values (`e.g 10%,20%,30%,40%,50%`) , these columns consists of the same data as the columns **`Gender`** and **`Target`** but also null values for the students. Also, performs the hypothesis testing on newly created columns and prints χ^2 of each test.

```r
# Call MCAR_Chi_Test function with increasing missing data proportions
# (10% to 50%) to assess impact on chi-squared statistic.
for (i in 1:5) {
  MCAR_Chi_Test(0.1 * i)
}
```

```
## [1] "For 10% of missing values, the chi-square value is 156.110383"
## [1] "For 20% of missing values, the chi-square value is 117.398484"
## [1] "For 30% of missing values, the chi-square value is 89.389119"
## [1] "For 40% of missing values, the chi-square value is 66.865717"
## [1] "For 50% of missing values, the chi-square value is 45.909804"
```

The chi-square values decrease as the percentage of missing values increases. This suggests that as you introduce more missing data, the association between **`Gender_MCAR`** and **`Dropout_MCAR`** becomes weaker or less significant. But, association between them still exists as all the χ^2 value greater than critical value ($= 3.841459$).

Data missing completely at random (MCAR) for hypothesis II

We don't have any missing values in our dataset, let's simulate a dataset with data missing at random. #

```r
MCAR_Chi_TestII <- function(prop_missing){
  # Set a seed for reproducibility
  set.seed(123)

  # Make a copy of the original data
  data1 <- data

  # Define a list of columns where missing data will be introduced
  columns_to_miss <- c("Marital_status", "Course", "Attendance",
                       "Previous_qualification", "Previous_qualification_grade",
                       "Admission_grade", "Educational_special_needs", "Debtor",
                       "Gender", "Scholarship_holder", "Age_at_enrollment",
                       "Curricular_units_Sem1_grade", "Curricular_units_Sem2_grade",
                       "Unemployment_rate", "Inflation_rate", "GDP")

  # Loop through each column and introduce missing values
  for (col in columns_to_miss) {
    # Calculate number of missing values
    num_missing <- round(prop_missing * nrow(data1))

    # Randomly select indices
    missing_indices <- sample(1:nrow(data1), size = num_missing)

    # Assign NA to the selected indices
```

```r
    data1[[col]][missing_indices] <- NA
  }

  # Fit logistic regression model on data with introduced missing values
  logit_model_MCAR <- glm(dropout ~ Course + Admission_grade +
                             Curricular_units_Sem1_grade + Unemployment_rate +
                             Inflation_rate + GDP + Gender,
                           family = binomial(), data = data1)

  model_summary_MCAR <- summary(logit_model_MCAR)
  print(model_summary_MCAR)
}
```

The above function, modifies the dataset columns by inserting null values for all the columns by varied percentages of missing values `(10%,20%,30%)` . Also, builds **`GLM`** on newly modified dataset.

```r
# Execute MCAR_Chi_TestII function with 10%, 20%, and 30%
# missing data to evaluate logistic model's robustness.
for (i in 1:3) {
  MCAR_Chi_TestII(0.1 * i)
}
```

```
##
## Call:
## glm(formula = dropout ~ Course + Admission_grade + Curricular_units_Sem1_grade +
##     Unemployment_rate + Inflation_rate + GDP + Gender, family = binomial(),
##     data = data1)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  3.136963   1.230607   2.549  0.01080 *
## Course171                   -3.241515   1.115949  -2.905  0.00368 **
## Course8014                  -0.084048   1.097337  -0.077  0.93895
## Course9003                   0.156520   1.100987   0.142  0.88695
## Course9070                  -0.229479   1.102464  -0.208  0.83511
## Course9085                  -0.039135   1.094709  -0.036  0.97148
## Course9119                   0.005315   1.102753   0.005  0.99615
## Course9130                   1.096158   1.102081   0.995  0.31992
## Course9147                  -0.417569   1.084777  -0.385  0.70029
## Course9238                  -0.767864   1.093210  -0.702  0.48243
## Course9254                   0.104468   1.087268   0.096  0.92345
## Course9500                  -0.487808   1.081338  -0.451  0.65191
## Course9556                   0.111642   1.139518   0.098  0.92195
## Course9670                   0.117142   1.089032   0.108  0.91434
## Course9773                   0.102598   1.085466   0.095  0.92470
## Course9853                   1.049037   1.092144   0.961  0.33679
## Course9991                   0.154251   1.092058   0.141  0.88767
```

```
## Admission_grade            -0.008518   0.004346  -1.960  0.04998 *
## Curricular_units_Sem1_grade -0.323485   0.020568 -15.728  < 2e-16 ***
## Unemployment_rate            0.044154   0.023397   1.887  0.05913 .
## Inflation_rate              -0.010505   0.042556  -0.247  0.80502
## GDP                         -0.030493   0.027551  -1.107  0.26838
## Gender                       0.617039   0.127936   4.823 1.41e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2606.6  on 2099  degrees of freedom
## Residual deviance: 1896.5  on 2077  degrees of freedom
##   (2324 observations deleted due to missingness)
## AIC: 1942.5
##
## Number of Fisher Scoring iterations: 5
##
##
## Call:
## glm(formula = dropout ~ Course + Admission_grade + Curricular_units_Sem1_grade +
##     Unemployment_rate + Inflation_rate + GDP + Gender, family = binomial(),
##     data = data1)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  16.595324 474.412733   0.035  0.97210
## Course171                   -16.761506 474.412054  -0.035  0.97182
## Course8014                  -14.533790 474.411972  -0.031  0.97556
## Course9003                  -14.048645 474.411977  -0.030  0.97638
## Course9070                  -14.857616 474.412067  -0.031  0.97502
## Course9085                  -14.177906 474.411953  -0.030  0.97616
## Course9119                  -13.922134 474.411983  -0.029  0.97659
## Course9130                  -13.183682 474.411964  -0.028  0.97783
## Course9147                  -14.742115 474.411902  -0.031  0.97521
## Course9238                  -15.208521 474.411991  -0.032  0.97443
## Course9254                  -13.964449 474.411879  -0.029  0.97652
## Course9500                  -14.742740 474.411883  -0.031  0.97521
## Course9556                  -14.195370 474.412175  -0.030  0.97613
## Course9670                  -13.753446 474.411891  -0.029  0.97687
## Course9773                  -13.686074 474.411883  -0.029  0.97699
## Course9853                  -12.383552 474.411993  -0.026  0.97918
## Course9991                  -13.698336 474.411922  -0.029  0.97696
## Admission_grade              -0.005997   0.006571  -0.913  0.36145
## Curricular_units_Sem1_grade  -0.316691   0.029531 -10.724  < 2e-16 ***
## Unemployment_rate             0.056648   0.037432   1.513  0.13019
## Inflation_rate                0.033147   0.065992   0.502  0.61546
## GDP                           0.065455   0.042327   1.546  0.12201
```

```
## Gender                         0.599694   0.192631   3.113  0.00185 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1141.91  on 921  degrees of freedom
## Residual deviance:  807.27  on 899  degrees of freedom
##   (3502 observations deleted due to missingness)
## AIC: 853.27
##
## Number of Fisher Scoring iterations: 13
##
##
## Call:
## glm(formula = dropout ~ Course + Admission_grade + Curricular_units_Sem1_grade +
##     Unemployment_rate + Inflation_rate + GDP + Gender, family = binomial(),
##     data = data1)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  3.062562   1.974167   1.551   0.1208
## Course171                   -3.908280   1.676233  -2.332   0.0197 *
## Course8014                  -0.210881   1.586197  -0.133   0.8942
## Course9003                  -1.002897   1.621788  -0.618   0.5363
## Course9070                  -0.361954   1.601565  -0.226   0.8212
## Course9085                  -0.983655   1.588656  -0.619   0.5358
## Course9119                  -1.530031   1.671864  -0.915   0.3601
## Course9130                  -0.634593   1.575236  -0.403   0.6871
## Course9147                  -0.813378   1.560670  -0.521   0.6022
## Course9238                  -1.178090   1.548714  -0.761   0.4468
## Course9254                  -1.069074   1.558496  -0.686   0.4927
## Course9500                  -1.931080   1.556514  -1.241   0.2147
## Course9556                  -0.163225   1.732488  -0.094   0.9249
## Course9670                  -1.363889   1.548610  -0.881   0.3785
## Course9773                  -0.693059   1.540126  -0.450   0.6527
## Course9853                   0.240800   1.626514   0.148   0.8823
## Course9991                  -0.589419   1.537284  -0.383   0.7014
## Admission_grade             -0.000711   0.010670  -0.067   0.9469
## Curricular_units_Sem1_grade -0.254143   0.038477  -6.605 3.97e-11 ***
## Unemployment_rate           -0.048597   0.055708  -0.872   0.3830
## Inflation_rate               0.118498   0.100633   1.178   0.2390
## GDP                         -0.121301   0.064429  -1.883   0.0597 .
## Gender                       0.738128   0.312883   2.359   0.0183 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 452.27  on 367  degrees of freedom
## Residual deviance: 335.72  on 345  degrees of freedom
##   (4056 observations deleted due to missingness)
## AIC: 381.72
##
## Number of Fisher Scoring iterations: 5
```

Above are the observations made from outputs of GLM for following MCAR percentages `( 10%, 20%, 30% )`. As the level of MCAR increases, the model's ability to identify significant predictors and compute reliable estimates changes, particularly evident in the 30% MCAR model.

Overall takeaway:

- The more data that was randomly removed correlated with less predictors being significant, until only the curricular semester 1 grade remained significant at alpha of 0.01. Economic factors did not become significant so the results of testing hypothesis 2 under MCAR did not change.

Model with 10% MCAR:

- **Significant Predictors**: Curricular semester 1 grade, gender and course 171 are the only variables remaining significant at alpha of 0.01.

Model with 20% MCAR:

- **Significant Predictors**: Curricular semester 1 grade and gender are the only variables remaining significant at alpha of 0.01.
- **Increased Missing Data**: The change in significant predictors suggests that the increased missing data might be affecting the reliability and consistency of the model.

Model with 30% MCAR:

- **Significant Predictors**: Curricular semester 1 grade is the only variable remaining significant at alpha of 0.01.

Data missing not at random (MNAR)

- **Scenario**: This scenario used will be a replication of a recession or pandemic where students have an increased dropout rate in their second semester with a negative change in GDP. The second semester has been chosen due to a lack of data in later semesters but also allows for a previous data point for imputation. When a student drops out their 2nd-semester grade will become NA. The values chosen are 20% of dropouts will have a null when the GDP is positive in the 2nd-semester grades and when the GDP is negative this rate will be 50%. This creates 462 nulls.

```r
course_data_alteration <- function(data) {
  set.seed(123)

  for (i in seq(1:nrow(data))) {

    # Condition 1: If a student dropped out and GDP is negative
    if (data[i, 'dropout'] == 1 && data[i, 'GDP'] < 0 && rbinom(1, 1, 0.50)) {
      # With a 50% probability, set 'Curricular_units_Sem2_grade' to NA
      data[i, 'Curricular_units_Sem2_grade'] = NA
    }

    # Condition 2: If a student dropped out and GDP is non-negative
    else if (data[i, 'dropout'] == 1 && data[i, 'GDP']>=0 &&rbinom(1, 1, 0.2)) {
      # With a 20% probability, set 'Curricular_units_Sem2_grade' to NA
      data[i, 'Curricular_units_Sem2_grade'] = NA
    }
  }

  return(data)
}
```

```r
# Apply missing data simulation to 'data' and
# calculate the total number of missing values in the altered dataset
data_MNAR <- course_data_alteration(data)
sum(is.na(data_MNAR))
```

```
## [1] 462
```

Hypothesis 1 Re-test MNAR

Does not change due to the overall counts of gender and dropout not being impacted.

```
##          Dropout
## Gender   Grad/Enrolled Dropout
##   Female          2148     720
##   Male             855     701
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table_MNAR
## X-squared = 183.16, df = 1, p-value < 2.2e-16
```

Hypothesis 2 Re-test MNAR

`Null hypothesis - H0` : The likelihood of a student dropping out is not impacted by economic climate when courses, gender, and grades are equal.

`Alternative hypothesis - Ha`: The likelihood of a student dropping out is impacted by economic climate when courses, gender, and grades are equal.

The initial model used the first semester grade as the predictor instead of second semester due to co-linearity. This model will use the second semester to fit into the scenario. While this may change some p values slightly, the variables that are/aren't significant should not change.

All other variables that were significant previously when testing hypothesis 2 remain significant. That is gender, specific courses, semester grade and admission grade remain significant. However, GDP has now become significant, so we can reject the null hypothesis in favor of the alternative. That is there is a significant increase in the likelihood of a student dropping out due to poor economic climate.

```r
# Fit a logistic regression model on the 'data_MNAR'
# dataset (with Missing Not At Random data).

logit_model_MNAR <- glm(dropout ~ Course + Admission_grade +
                          Curricular_units_Sem2_grade + Unemployment_rate +
                          Inflation_rate + GDP + Gender,
                        family = binomial(), data = data_MNAR)

summary(logit_model_MNAR)
```

```
##
## Call:
## glm(formula = dropout ~ Course + Admission_grade + Curricular_units_Sem2_grade +
##     Unemployment_rate + Inflation_rate + GDP + Gender, family = binomial(),
##     data = data_MNAR)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  4.330121   0.896484   4.830 1.36e-06 ***
## Course171                   -4.208647   0.775923  -5.424 5.83e-08 ***
## Course8014                  -1.547805   0.776713  -1.993 0.046287 *
## Course9003                  -1.357676   0.774028  -1.754 0.079424 .
## Course9070                  -1.789127   0.774885  -2.309 0.020949 *
## Course9085                  -1.488620   0.765478  -1.945 0.051812 .
## Course9119                  -0.800860   0.769318  -1.041 0.297876
## Course9130                   0.026386   0.772647   0.034 0.972757
## Course9147                  -1.447064   0.754289  -1.918 0.055054 .
## Course9238                  -1.974675   0.766113  -2.578 0.009951 **
## Course9254                  -1.178060   0.760385  -1.549 0.121311
## Course9500                  -1.880984   0.754637  -2.493 0.012682 *
## Course9556                  -1.721950   0.848266  -2.030 0.042360 *
## Course9670                  -0.921490   0.757735  -1.216 0.223942
## Course9773                  -1.178497   0.758184  -1.554 0.120097
## Course9853                  -0.397277   0.764636  -0.520 0.603368
## Course9991                  -0.402673   0.757233  -0.532 0.594886
```

```
## Admission_grade            -0.009806   0.003579  -2.740 0.006148 **
## Curricular_units_Sem2_grade -0.335215   0.013827 -24.243  < 2e-16 ***
## Unemployment_rate            0.026826   0.020275   1.323 0.185788
## Inflation_rate               0.028079   0.037225   0.754 0.450663
## GDP                          0.089265   0.023795   3.751 0.000176 ***
## Gender                       0.557807   0.106793   5.223 1.76e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4385.4  on 3961  degrees of freedom
## Residual deviance: 2881.3  on 3939  degrees of freedom
##   (462 observations deleted due to missingness)
## AIC: 2927.3
##
## Number of Fisher Scoring iterations: 5
```

Re-imputing with MICE

We will attempt to recover the data using MICE to re-impute the missing nulls.

```r
# Perform multiple imputation on 'data_MNAR' using
# random forests method with the 'mice' package.

imputation_model <- mice(data_MNAR, method = "rf")
```

```
##
##  iter imp variable
##   1   1  Curricular_units_Sem2_grade
##   1   2  Curricular_units_Sem2_grade
##   1   3  Curricular_units_Sem2_grade
##   1   4  Curricular_units_Sem2_grade
##   1   5  Curricular_units_Sem2_grade
##   2   1  Curricular_units_Sem2_grade
##   2   2  Curricular_units_Sem2_grade
##   2   3  Curricular_units_Sem2_grade
##   2   4  Curricular_units_Sem2_grade
##   2   5  Curricular_units_Sem2_grade
##   3   1  Curricular_units_Sem2_grade
##   3   2  Curricular_units_Sem2_grade
##   3   3  Curricular_units_Sem2_grade
##   3   4  Curricular_units_Sem2_grade
##   3   5  Curricular_units_Sem2_grade
##   4   1  Curricular_units_Sem2_grade
##   4   2  Curricular_units_Sem2_grade
##   4   3  Curricular_units_Sem2_grade
```

```
##   4   4  Curricular_units_Sem2_grade
##   4   5  Curricular_units_Sem2_grade
##   5   1  Curricular_units_Sem2_grade
##   5   2  Curricular_units_Sem2_grade
##   5   3  Curricular_units_Sem2_grade
##   5   4  Curricular_units_Sem2_grade
##   5   5  Curricular_units_Sem2_grade
```

```
## Warning: Number of logged events: 26
```

```
# Impute missing values using the specified imputation
# model and store the completed dataset in 'imputed_data'.

imputed_data <- complete(imputation_model)
```

Hypothesis 1 with MICE

Hypothesis does not change.

```
##         Dropout
## Gender   Grad/Enrolled Dropout
##   Female          2148     720
##   Male             855     701
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table_imputed
## X-squared = 183.16, df = 1, p-value < 2.2e-16
```

Hypothesis 2 with MICE

All other variables that were significant previously when testing hypothesis 2 remain significant. That is gender, specific courses, semester grade and admission grade remain significant. MICE was able to recover the relationship between the semester 2 grades GDP and dropout. We can no longer reject the null hypothesis that economic factors impact the likelihood of dropping out. .

```
# Fit a logistic regression model on the imputed dataset 'imputed_data'.

logit_model_imputed <- glm(dropout ~ Course + Admission_grade +
                             Curricular_units_Sem2_grade + Unemployment_rate +
                             Inflation_rate + GDP + Gender,
                           family = binomial(), data = imputed_data)

# Prints summary of model with imputed data

summary(logit_model_imputed)
```

```
##
## Call:
## glm(formula = dropout ~ Course + Admission_grade + Curricular_units_Sem2_grade +
##     Unemployment_rate + Inflation_rate + GDP + Gender, family = binomial(),
##     data = imputed_data)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  4.703446   0.825066   5.701 1.19e-08 ***
## Course171                   -4.241384   0.727141  -5.833 5.45e-09 ***
## Course8014                  -1.446394   0.722783  -2.001  0.04538 *
## Course9003                  -1.227150   0.720290  -1.704  0.08844 .
## Course9070                  -1.688775   0.721872  -2.339  0.01931 *
## Course9085                  -1.271886   0.712899  -1.784  0.07441 .
## Course9119                  -0.717986   0.720714  -0.996  0.31915
## Course9130                  -0.034570   0.724077  -0.048  0.96192
## Course9147                  -1.391621   0.706420  -1.970  0.04884 *
## Course9238                  -2.021293   0.717034  -2.819  0.00482 **
## Course9254                  -1.082022   0.711141  -1.522  0.12813
## Course9500                  -1.822154   0.705837  -2.582  0.00984 **
## Course9556                  -1.271667   0.764569  -1.663  0.09626 .
## Course9670                  -1.021448   0.711091  -1.436  0.15087
## Course9773                  -1.125366   0.709188  -1.587  0.11255
## Course9853                  -0.359748   0.714901  -0.503  0.61481
## Course9991                  -0.487122   0.710276  -0.686  0.49283
## Admission_grade             -0.009745   0.003141  -3.103  0.00192 **
## Curricular_units_Sem2_grade -0.339673   0.013147 -25.837  < 2e-16 ***
## Unemployment_rate            0.030015   0.017106   1.755  0.07932 .
## Inflation_rate               0.006564   0.030700   0.214  0.83070
## GDP                          0.005529   0.019959   0.277  0.78177
## Gender                       0.554708   0.094028   5.899 3.65e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5554.5  on 4423  degrees of freedom
## Residual deviance: 3622.3  on 4401  degrees of freedom
## AIC: 3668.3
##
## Number of Fisher Scoring iterations: 5
```

5. Conclusion

Hypothesis I

χ^2 is used to test independence between two categorical columns (i.e., gender and dropout) and test resulted that there is significant relation (dependence) between gender of a student and their academic success. As percentage of missing values increases their dependence decreases.

Hypothesis II

GLM is used to test whether student dropping out is impacted by economic climate when courses, gender, and grades are equal and it resulted in unemployment rate, Inflation rate, and GDP all had p values greater than 0.05. The variables that had significant relationships with dropout were Gender, admission grades and the students semester 1 grade (P values < 0.01). Specific courses also showed significant differences in the likelihood of dropping out, notably course 171 had the lowest odds of dropping out.

6. References

https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

https://education.rstudio.com/learn/beginner/

https://www.w3schools.com/r/

https://www.geeksforgeeks.org/exploratory-graphs-for-eda-in-r/

https://www.rdocumentation.org/packages/mice/versions/3.16.0/topics/mice

https://www.ncbi.nlm.nih.gov/books/NBK493614/

https://www.youtube.com/watch?v=rDZWtgOH124

https://bookdown.org/yihui/rmarkdown-cookbook/raw-latex.html

https://stackoverflow.com/questions/29274501/r-markdown-changing-font-size-and-font-type-in-html-output

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/