AMS572: Project report

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 socioeconomic 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")
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.

```
data <- read.csv("AcademicSuccessData.csv")
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

```
sum(is.na(data))
```

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

There were no missing values in the dataset.

Key facts based on descriptive statistics

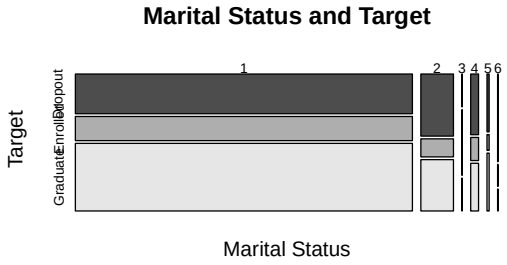
```
##                     Grade_Type         Mean Standard_Deviation
## 1 Previous Qualification Grade 1.326133e+02          13.188332
## 2              Admission Grade 1.269781e+02          14.482001
## 3  Curricular Units Sem1 Grade 1.064082e+01           4.843663
## 4  Curricular Units Sem2 Grade 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.

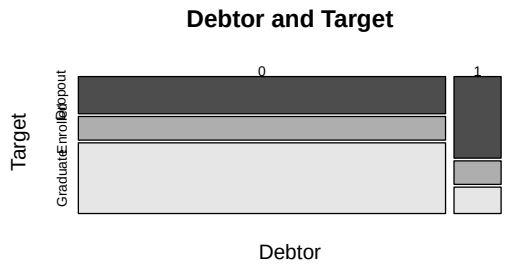
The average grades for the first and second semesters **Curricular_units_Sem1_grade** and **Curricular_units_Sem2_grade** were similar, but the standard deviation of semester-2 grades is noticebly higher. This suggests a varied academic performance across students.

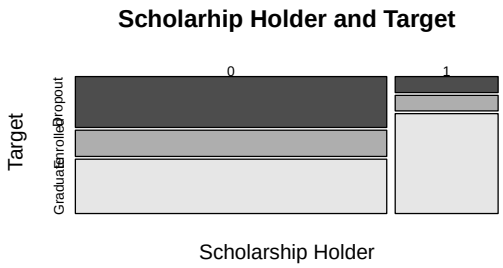
Some interesting plots



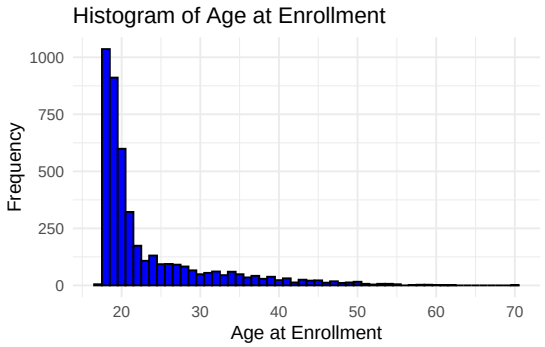
1 - Male, 0 - Female

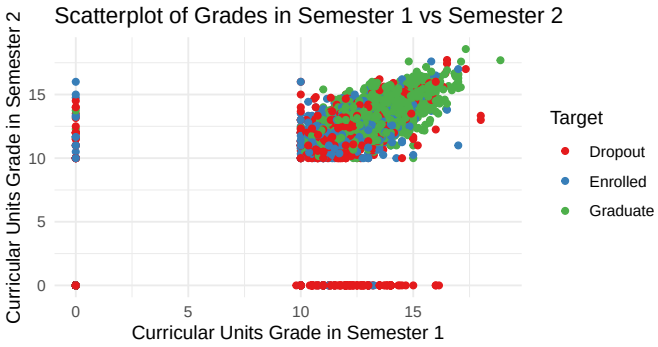
1 - Single, 2 - Married, 3 - Widower, 4 - Divorced, 5 - Facto Union, 6 - Legally Separated





0 - No, 1 - Yes





3. Hypotheses, Methodology and Testing

Hypothesis - 1

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

`Alternative hypothesis - Ha`: There is a significant relation 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`

```
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.

```
df <- 1
alpha <- 0.05

critical_value <- qchisq(1 - alpha, df)
```

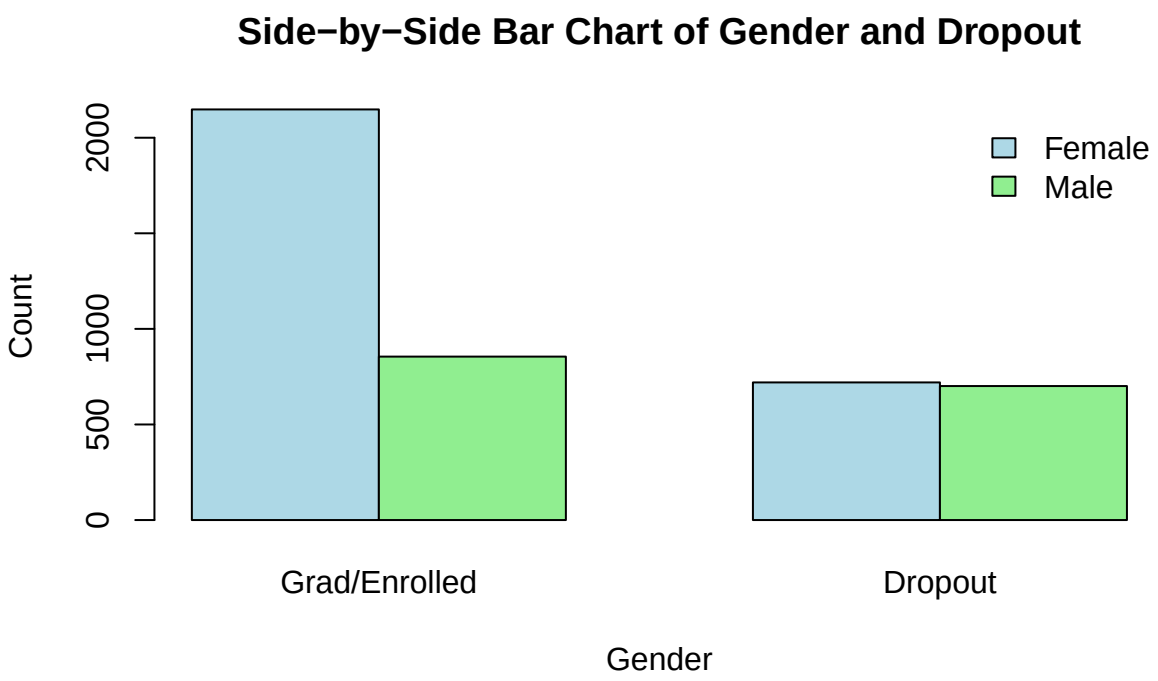
The critical region, C_α is

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

```
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.

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)

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.

```
MCAR_Chi_Test <- function(prop_missing){
  set.seed(123)

  data$Gender_MCAR <- data$Gender
  data$Dropout_MCAR <- data$dropout

  missing_indices_gender <- sample(1:nrow(data), size = round(prop_missing * nrow  (data)))
  missing_indices_dropout <- sample(1:nrow(data), size = round(prop_missing * nrow  (data)))

  data$Gender_MCAR[missing_indices_gender] <- NA
  data$Dropout_MCAR[missing_indices_dropout] <- NA

  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")

  x_values <- seq(0, critical_value + 10, by = 0.1)
  chi_sq_df <- data.frame(x = x_values, y = dchisq(x_values, df))
  ind_test_g_d <- chisq.test(contingency_table_MCAR)

   result <- sprintf("For %s%% of missing values, the chi-square value is %f", prop_missing * 100, ind_
  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.

```
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 not at random (MNAR)

TO DO

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` - $(\alpha) = 0.05$

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 (marital status, age, previous academic qualifications and grades) is less than 0.05, we reject null hypothesis.

```
logit_model <- glm(dropout ~ Course + Admission_grade + Curricular_units_Sem1_grade +   Unemployment_ra
, family = binomial(), data = data)

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.

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)

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.seed(123)

  data1<-data

  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",

  for (col in columns_to_miss) {
    num_missing <- round(prop_missing * nrow(data1))

    missing_indices <- sample(1:nrow(data1), size = num_missing)
```

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

  logit_model_MCAR <- glm(dropout ~ Course + Admission_grade + Curricular_units_Sem1_grade +    Unemplo
, 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.

```
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:

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

Data missing not at random (MNAR)

- **Scenario**: Data often needs to be stored which can be costly. In this scenario to lower operating costs the university decides to move inactive/old grading information to a different server, However, there was an error in the transfer of data which has damaged the original information. Those who have been in courses 33, 9119, 9130, 9991, 9853 have had their grades and personal identification info lost. These courses were the ones with the highest rates of dropping out (each >= 40%)

```
table(data$Course, data$dropout)
```

```
##
##          0   1
##   33     4   8
##   171  133  82
##   8014 144  71
##   9003 124  86
##   9070 175  51
##   9085 247  90
##   9119  78  92
##   9130  63  78
##   9147 246 134
##   9238 290  65
##   9254 156  96
##   9500 648 118
##   9556  53  33
##   9670 173  95
##   9773 230 101
##   9853 107  85
##   9991 132 136
```

```
course_data_alteration <- function(data,courses = c(33,9119, 9130, 9991, 9853)) {
  for (i in seq(1:nrow(data))) {
    if (data[i,c('Course')] %in% courses) {
      data[i, c("Curricular_units_Sem1_grade", "Curricular_units_Sem2_grade", "Gender", "Scholarship_ho
    }
  }
  return(data)
}
```

```
data_MNAR <- course_data_alteration(data)
```

The association between Gender and Dropout remain significant, though it appears the vast majority of those in the missing courses were female, under representing the strength of association between being a male and risk of dropping out.

```
contingency_table_MNAR <- table(data_MNAR[,c('Gender','dropout')])
names(dimnames(contingency_table_MNAR)) <- c('Gender', 'Dropout')
colnames(contingency_table_MNAR) <- c("Grad/Enrolled", "Dropout")
rownames(contingency_table_MNAR) <- c("Female","Male")
print(contingency_table_MNAR)
```

```
##         Dropout
## Gender   Grad/Enrolled Dropout
##   Female          1924     540
##   Male             695     482
```

```
chisq.test(contingency_table_MNAR)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table_MNAR
## X-squared = 142.01, df = 1, p-value < 2.2e-16
```

Checking the logistic regression model used previously to check for any changes in significant dropout predictors. As these courses contained all classes held at night, attendance has become singular and therefore has been dropped from the model. There appears to not be any change in significant predictors for dropout despite removal of half of the available courses.

```
logit_model_MNAR <- glm(dropout ~ Course + Admission_grade + Curricular_units_Sem1_grade +  Unemployment
, family = binomial(), data = data_MNAR)

summary(logit_model_MNAR)
```

```
##
## Call:
## glm(formula = dropout ~ Course + Admission_grade + Curricular_units_Sem1_grade +
##     Unemployment_rate + Inflation_rate + GDP + Gender, family = binomial(),
##     data = data_MNAR)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 0.655443   0.524217   1.250  0.21118
## Course8014                  2.825127   0.294948   9.578  < 2e-16 ***
## Course9003                  2.967638   0.293886  10.098  < 2e-16 ***
## Course9070                  2.574579   0.300752   8.560  < 2e-16 ***
## Course9085                  2.934160   0.286644  10.236  < 2e-16 ***
## Course9147                  2.682828   0.262953  10.203  < 2e-16 ***
## Course9238                  2.071423   0.282380   7.336 2.21e-13 ***
## Course9254                  2.981995   0.280757  10.621  < 2e-16 ***
```

```
## Course9500                    2.190568   0.269590   8.126 4.45e-16 ***
## Course9556                    3.079544   0.355663   8.659  < 2e-16 ***
## Course9670                    3.074357   0.285965  10.751  < 2e-16 ***
## Course9773                    3.001292   0.278792  10.765  < 2e-16 ***
## Admission_grade              -0.010129   0.003317  -3.054  0.00226 **
## Curricular_units_Sem1_grade -0.306883   0.015272 -20.094  < 2e-16 ***
## Unemployment_rate             0.016193   0.018092   0.895  0.37076
## Inflation_rate               -0.016278   0.032177  -0.506  0.61294
## GDP                          -0.025339   0.020785  -1.219  0.22279
## Gender                        0.606584   0.095298   6.365 1.95e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4322.6  on 3640  degrees of freedom
## Residual deviance: 3242.2  on 3623  degrees of freedom
##   (783 observations deleted due to missingness)
## AIC: 3278.2
##
## Number of Fisher Scoring iterations: 5
```

```
variables_to_impute <- c("Curricular_units_Sem1_grade", "Curricular_units_Sem2_grade", "Gender", "Schola

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

```
##
##  iter imp variable
##   1   1  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   1   2  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   1   3  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   1   4  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   1   5  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   2   1  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   2   2  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   2   3  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   2   4  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   2   5  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   3   1  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   3   2  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   3   3  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   3   4  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   3   5  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   4   1  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   4   2  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   4   3  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   4   4  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   4   5  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   5   1  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   5   2  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   5   3  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   5   4  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
##   5   5  Curricular_units_Sem1_grade  Curricular_units_Sem2_grade  Gender  Scholarship_holder  Debtor
```

```
## Warning: Number of logged events: 1
```

```
imputed_data <- complete(imputation_model)

data_imputed <- cbind(data_MNAR[, -which(names(data_MNAR) %in% variables_to_impute)], imputed_data)
```

The association between Gender and Dropout remain significant, the skew of female to male was not able to be recovered.

```
contingency_table_imputed <- table(data_imputed[,c('Gender','dropout')])
names(dimnames(contingency_table_imputed)) <- c('Gender', 'Dropout')
colnames(contingency_table_imputed) <- c("Grad/Enrolled", "Dropout")
rownames(contingency_table_imputed) <- c("Female","Male")
print(contingency_table_imputed)
```

```
##         Dropout
## Gender   Grad/Enrolled Dropout
##   Female          2192     808
##   Male             811     613
```

```
chisq.test(contingency_table_imputed)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table_imputed
## X-squared = 114.27, df = 1, p-value < 2.2e-16
```

MICE reinputed logistic regression, didn't have time to write anything.

```
logit_model_imputed <- glm(dropout ~  Course + Admission_grade + Curricular_units_Sem1_grade +  Unemploy
, family = binomial(), data = data_imputed)

summary(logit_model_imputed)
```

```
##
## Call:
## glm(formula = dropout ~ Course + Admission_grade + Curricular_units_Sem1_grade +
##     Unemployment_rate + Inflation_rate + GDP + Gender, family = binomial(),
##     data = data_imputed)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  3.914922   0.740480   5.287 1.24e-07 ***
## Course171                   -3.595858   0.652194  -5.513 3.52e-08 ***
## Course8014                  -1.844184   0.643592  -2.865 0.004164 **
## Course9003                  -1.725877   0.644829  -2.676 0.007440 **
## Course9070                  -2.169832   0.645979  -3.359 0.000782 ***
## Course9085                  -1.848215   0.637957  -2.897 0.003766 **
## Course9119                  -0.750311   0.643964  -1.165 0.243960
## Course9130                  -0.630773   0.648195  -0.973 0.330493
```

```
## Course9147                   -1.916432   0.634165  -3.022 0.002511 **
## Course9238                   -2.557067   0.640522  -3.992 6.55e-05 ***
## Course9254                   -1.677905   0.638760  -2.627 0.008619 **
## Course9500                   -2.560298   0.632166  -4.050 5.12e-05 ***
## Course9556                   -1.652900   0.670126  -2.467 0.013642 *
## Course9670                   -1.644407   0.638323  -2.576 0.009991 **
## Course9773                   -1.763225   0.635975  -2.772 0.005563 **
## Course9853                   -1.288664   0.643262  -2.003 0.045142 *
## Course9991                   -0.887903   0.636439  -1.395 0.162983
## Admission_grade              -0.008216   0.002696  -3.048 0.002307 **
## Curricular_units_Sem1_grade -0.208820   0.009938 -21.013  < 2e-16 ***
## Unemployment_rate             0.023333   0.014948   1.561 0.118551
## Inflation_rate                0.024894   0.026540   0.938 0.348254
## GDP                          -0.030666   0.017325  -1.770 0.076719 .
## Gender                        0.461550   0.079774   5.786 7.22e-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: 4537.8  on 4401  degrees of freedom
## AIC: 4583.8
##
## Number of Fisher Scoring iterations: 4
```