Yizhen Jia, Yuhua Qiao, and Andrew Chen

AMS572 Group 10 Project Report

Project Background:

For this analysis, we are interested in studying graduate earnings at several U.S. universities. A college education is an expensive investment, and we are interested in studying if the cost of education is justified. We have obtained our data from the Data and Story Library. The dataset contained data from 706 US universities and colleges. In the data we obtained, schools were split into two categories, public and private schools. It has 438 public schools and 268 private schools. We also have statistics on the school's enrollment average SAT and ACT scores. In addition, we have data on the average income of the student after graduating college and the tuition and tuition assistance for their studies in school. There are missing values in the merit_aided column of our data. According to the topic of this group project· We will use the knowledge learned in the course to supplement and experiment on the data, and we will present the results of our data analysis and the experimental procedure below.

Missing Data:

According to the project requirements we need to introduce the application of Missing Completely At Random (MCAR) and Missing Not At Random(MNAR) topics on our data separately. In our raw data, there were 706 values (8.64%) in merit_aided and 19 out of 706 values (2.69%) in need_fraction are missing. First, we analyze the data by determining which kind of missing data.

Checking Missing Completely At Random (MCAR):

At first glance, it seems that there were more missing values for private schools than public schools. 16 missing need_fraction for public schools and 3 missing need_fraction for private schools. We did a two-sample Z proportion test to determine if the public and private have the same true proportion of missing need_fraction values.

Setup:

let p1=true proportion of missing values for need_fraction based for public schools

let p2=true proportion of missing values need_fraction based for private schools

H0: p1 = p2

HA: p1 != p2

```
        2-sample test for equality of proportions with continuity correction

data:  c(x1, x2) out of c(n1, n2)
X-squared = 3.1653, df = 1, p-value = 0.07522
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0007104866 0.0499608144
sample estimates:
    prop 1     prop 2
0.03652968 0.01119403
```

Result:

There is not enough evidence to show that there is a statistically significant difference between the true proportion of missing values for need_fraction based on public schools and private schools. Since the p-value = 0.07522 > alpha=0.05. For this reason, we claim that the missing value for need_fraction is a type of MCAR type of missing.

Similarly, we check to see if the missingness of merit_aid is at random using the same methods. There are 44 missing merit_aided for public schools and 17 missing merit_aided for private schools.

Setup:

let p1=true proportion of missing values for merit_aided based for public schools

let p2=true proportion of missing values for merit_aided based for private schools

H0: p1 = p2

HA: p1 != p2

```
        2-sample test for equality of proportions with continuity correction

data:  c(x1, x2) out of c(n1, n2)
X-squared = 2.4373, df = 1, p-value = 0.1185
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.006530957  0.080578528
sample estimates:
    prop 1     prop 2
0.10045662 0.06343284
```

Result:

There is not enough evidence to show that there is a statistically significant difference between the true proportion of missing values for merit_aided based on public schools and private schools. Since the p-value = 0.1185 > alpha=0.05. For this reason, we claim that the missing value for merit_aided is also a type of MCAR type missing.

Checking Missing Not At Random(MNAR):

After applying the two-sample Z proportion test twice, we claim that our dataset has only MCAR type of missing values. To simulate the effect of MNAR, we have to decide a column upon which we should enforce missing values. For this, we selected private schools to enforce 20% of missing values for earning if they are from private schools. We chose income because we assumed that some schools did not complete the corresponding graduate student income statistics, or some schools did not want to release this part of data to the public. Thus, the MNAR is simulated.

Hypothesis 1:

For our first hypothesis, we are interested in studying if schools are giving enough aid to students to attend college. Colleges are expensive and fundings are granted from the government to help this. However, we would like to know if the needs are satisfied to help students to go to college.

In the data, we have the need_fraction with missing values of type MCAR which measures how much money a student needs to attend school with values ranging from 0 and 1. The money a student would need can be calculated as need_index*price. For example, if a student's need_fraction is 0.25 and the price is $45,000. They would need $11,250 of money in aid to help them attend school.

To measure if the need is satisfied, we defined a new data field called aid_percent which is calculated by (price-price_with_aid)/price. This field shows the percent of aid given by the school also ranging from 0 and 1.

We used a paired sample t-test to test if the true mean aid_percent is greater than or equal to the true mean of need_fraction. We used a paired sample t-test because we are interested in seeing if each school has given enough aid to its students to attend school since aid is given on a school basis.

Dealing with MCAR:

Before conducting the paired sample t-test, we already know that the need_fraction contains MCAR type of missing values. We need to have a way of dealing with missing values. We know that missingness completely at random is that the data miss some values that have no relationship between the observed variables data and missing variable data. If the type of the missingness is missing completely at random, there is no systematic fault of the data as long as the percent of missing data is not too large to use the imputation method appropriately. In addition, what we can do to deal with this kind of missingness is to use the package called

MICE. MICE imputes the missing data for both univariate and multivariate data. Since our data is multivariate, we decided to use a predictive mean matching method. We chose this method because it imputes plausible data and the data are not too extreme. Predictive Mean Matching (PMM) is a semi-parametric imputation approach. It is similar to the regression method except that for each missing value, it fills in a value randomly from among the observed donor values from an observation whose regression-predicted values are closest to the regression-predicted value for the missing value from the simulated regression model."

Therefore, we have 3 ways to deal with the MCAR type of missing values as proposed. We can ignore the missing values completely by dropping NAs since it's only 19 out of 706 values (2.69%) in need_fraction, impute by replacing the missing values with the mean of the need_fractioin column, and using the Predictive Mean Matching algorithm provided by the MICE package to impute the missing values.

Paired T-Tests:

Setup:

Let d be the true mean difference of aid_percent and need fraction

d ranges from -1 and 1.

-1 meaning no aid given despite having 100% need

0 meaning given is given sufficiently to satisfy student's need

1 meaning 100% aid given despite having no need

H0: $d \geq 0$

HA: $d < 0$

Method 1: Ignore the missing values by dropping NA's

```
        One Sample t-test

data:  d
t = -21.176, df = 686, p-value < 2.2e-16
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
       -Inf -0.1357403
sample estimates:
 mean of x
-0.1471885
```

Result:

Reject H0, the d be true mean difference is less than 0 with p-value <2.2e-16 < alpha=0.05. Thus, we conclude that the aid is not given sufficiently to students.

Method 2: Impute by replacing the missing values with the mean of the need_fraction

```
        One Sample t-test

data:  d
t = -21.831, df = 705, p-value < 2.2e-16
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
       -Inf -0.1372232
sample estimates:
 mean of x
-0.1484204
```

Result:

Reject H0, the d be true mean difference is less than 0 with p-value <2.2e-16 < alpha=0.05. Thus, we conclude that the aid is not given sufficiently to students.

Method 3: Impute by replacing the missing values with the Predictive Mean Matching algorithm

```
        One Sample t-test

data:  d
t = -22.01, df = 705, p-value < 2.2e-16
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
       -Inf -0.1402592
sample estimates:
 mean of x
-0.1516037
```

Result:

Reject H0, the d be true mean difference is less than 0 with p-value <2.2e-16 < alpha=0.05. Thus, we conclude that the aid is not given sufficiently to students.

Summary:

In all three ways of dealing with MCAR, we see that the conclusions are the same. Aid is not given sufficiently to students. Despite dealing with missingness with different methods, we came to the same conclusion. This conclusion aligns with the opinion that U.S. colleges are expensive even though aids are provided by schools.

Hypothesis 2:

For our second hypothesis, we are interested in modeling the average earnings by the college. We are interested in this because we would like to know if certain characteristics of a college can help make better earnings as college is an expensive investment and one would like to have better returns on their investment. We started by examining the correlation between the variables in our model. We do this by calculating the correlation coefficient.

	Earn	Public	State	SAT	ACT	Price	Price_with_aid	need_fraction	merit_aided	Aid_amount	Aid_percent
Earn	1.00000000	0.08598003	-0.02348949	0.50910382	0.51118842	0.39267717	0.31425053	-0.48182567	0.047645824	0.316113390	0.07097941
Public	0.08598003	1.00000000	0.06277247	-0.02994785	-0.07402089	-0.65778688	-0.48416685	-0.36787645	-0.226936940	-0.566284635	-0.39011345
State	-0.02348949	0.06277247	1.00000000	0.01785238	-0.01770141	-0.08609315	-0.08489542	-0.06714313	0.080749145	-0.055389638	-0.02043064
SAT	0.50910382	-0.02994785	0.01785238	1.00000000	0.93235089	0.54448933	0.40701977	-0.55627313	0.134394857	0.463313495	0.11992799
ACT	0.51118842	-0.07402089	-0.01770141	0.93235089	1.00000000	0.56475212	0.42758354	-0.55695723	0.202401837	0.475842858	0.13362443
Price	0.39267717	-0.65778688	-0.08609315	0.54448933	0.56475212	1.00000000	0.76972593	-0.12689193	0.228823194	0.831600016	0.39258666
Price_with_aid	0.31425053	-0.48416685	-0.08489542	0.40701977	0.42758354	0.76972593	1.00000000	-0.15850914	0.393305772	0.285566846	-0.25784963
need_fraction	-0.48182567	-0.36787645	-0.06714313	-0.55627313	-0.55695723	-0.12689193	-0.15850914	1.00000000	-0.091763453	-0.052596053	0.16217344
merit_aided	0.04764582	-0.22693694	0.08074915	0.13439486	0.20240184	0.22882319	0.39330577	-0.09176345	1.000000000	0.001351117	-0.16960498
Aid_amount	0.31611339	-0.56628463	-0.05538964	0.46331349	0.47584286	0.83160002	0.28556685	-0.05259605	0.001351117	1.000000000	0.81369499
Aid_percent	0.07097941	-0.39011345	-0.02043064	0.11992799	0.13362443	0.39258666	-0.25784963	0.16217344	-0.169604976	0.813694986	1.00000000

Table 1: correlation coefficients between variables

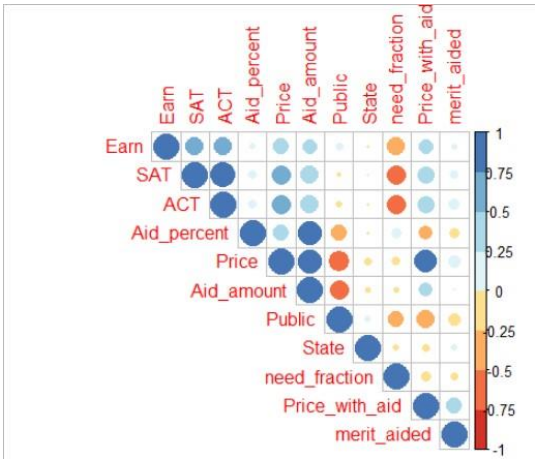

Figure 1: Heat map showing the correlation coefficient between variables

Looking at the correlation coefficient matrix, we see quite a bit of high correlations between several variables such as SAT and ACT, Public and Price, etc. Thus, we think that multiple regression models may not fit as well for our modeling purpose. We chose to go with a LASSO regression model instead. The reason for choosing LASSO is because LASSO regression has the features of variable selection and complexity adjustment while fitting a generalized linear model. Therefore, LASSO regression can be modeled regardless of whether the target-dependent variable is continuous, binary, or multivariate discrete. We could do a model selection using criteria like Akaike information criterion, Bayesian information criterion, etc. However, these model selection methods are more computationally expensive since we need to consider an exponential amount of cases. In the computational sense, we would like to avoid these kinds of computations, thus using a LASSO regression would be the best choice.

Dealing with MNAR:

Before modeling the earning we need to consider how to deal with the missing values in the Earning column which we simulated as an MNAR type of missingness. The way we chose to

deal with this is by utilizing the MICE package once again with the predictive mean matching method. Unlike MCAR, MNAR is that the missingness depends on missed data. The type of missingness affects the data and test decision heavily if there is a high percentage of this kind of missingness. If the type of the missingness is missingness not at random, it is hard to identify the missingness at first hand using statistical methods because if the data are missed not at random, there is no data presented that can help us find. Through the use of the predictive mean matching method, MNAR can be dealt with. It takes in information from other columns to impute the data. Since the data is missing based on the public school column, it can use this information to help fill in with means of public school earnings. In addition, it can go even further, utilizing earnings of public school and location, or means of public school, location, and SAT, etc. This kind of imputation is more specific and can better simulate what the missing data can be.

LASSO Regression:

After using the predictive mean matching method, we were able to fill in the earnings column and run the LASSO regression. We tried different values of lambda ranging from 0.001 to 100 to see what the results are and plotted the coefficients vs log lambda as follows.

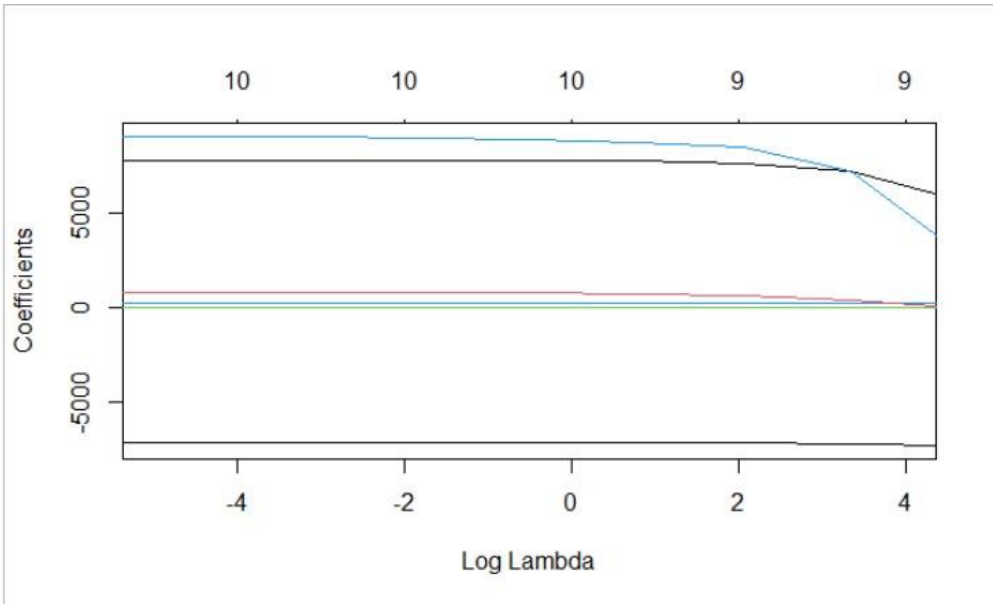

Figure 2: Showing the coefficients for respective Log Lambda values

To select the best lambda for the LASSO regression we utilized cross-validation with 10 folds and the final best lambda value from cross-validation was 45.16499.

```
Call:  cv.glmnet(x = x, y = y, nfolds = 10, alpha = 1)

Measure: Mean-Squared Error

    Lambda Index  Measure      SE Nonzero
min   45.2    48 28215426 1874375       9
1se  318.6    27 29999823 2311744       7
```

The final LASSO Regression model that we obtained at the end was:

Earn = 20405.38 + 6849.57*Public - 21.51*State + 7.35e*SAT + 241.01*ACT + .20*Price + 0.07*Price_with_aid - 0.01*need_fraction + 147.27*merit_aided + 6132.19*Aid_percent

The variable that got eliminated was Aid_amount which makes sense since Aid_amount is highly correlated with both price and Aid_percent. Looking at the coefficients of our LASSO regression model, it seems like the field Public schools have the highest coefficient which could mean that private school students make less than public school students. We can verify this by doing a 2 sample t-test.

Setup:

Let X be the earnings of public school students

Let Y be the earnings of private school students

First, we need to check for equal variance. We do this by conducting the F test.

H0: Var(X) = Var(Y)

HA: Var(X) != Var(Y)

```
        F test to compare two variances

data:  x and y
F = 0.66284, num df = 267, denom df = 437, p-value = 0.0002528
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5357631 0.8247447
sample estimates:
ratio of variances
         0.6628427
```

By the F-test, reject the null hypothesis with p-value=0.0002528 < alpha=0.05. The variance of the earnings of public school students is not equal to the variance of the earnings of private school students.

Two Sample T-test with unequal variances:

H0: true mean of X >= true mean of Y

HA: true mean of X < true mean of Y

```
        Welch Two Sample t-test

data:  x and y
t = 0.21762, df = 649.37, p-value = 0.5861
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 927.4528
sample estimates:
mean of x mean of y
 45637.69  45529.45
```

Result:

By the two-sample t-tests with unequal variances, fail to reject the null hypothesis with p-value=0.5861 > alpha=0.05. We cannot conclude that the true mean earnings of public school students are less than the true mean earnings of private school students.

Conclusion:

From the two hypotheses, we can conclude that colleges don't satisfy the need of students when giving out student aids from our first hypothesis, and earnings after graduating college don't depend upon public or private schools. We see that colleges are expensive for students who need financial support in our first hypothesis and post-college earnings depend on quite a few factors in the LASSO regression. Something that would be interesting to investigate further would be the expected time to pay off college loans with the earnings as people do say that college is a good investment. If we were to continue this study, we would also obtain further information about loans that were taken out by college students as well as model the time it would take to pay off student loans.

Works Cited:

"6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples)." Towards Data Science,https://towardsdatascience.com/6-different-ways-to-Compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779. Accessed 11 Nov. 2021.

Graduate Earnings. The Data And Story Library, https://dasl.datadescription.com/datafile/graduate-earnings/?_sfm_cases=500+10000&sf_paged=3. Accessed 11 Nov. 2021.

Package 'mice.' The Comprehensive R Archive Network,https://cran.r-project.org/web/packages/mice/mice.pdf. Accessed 11 Nov. 2021.