An Analytical Study of Graduate Earnings

~~Shaobo Qin, Tao Xiang & Aashi Gupta~~

1. Introduction

College is a major commitment of time and money, and the returns can be high since a college degree would help open doors to higher-paying jobs. The price and earning of colleges have always been a subject of universal interest. Soon-to-be college students usually place a high priority on college price and earnings, meanwhile, recent college graduates also need these data to help them make a prediction of their future income. Data shows that in 2019, full-time workers with high school graduates (no college) had weekly earnings of $749, meanwhile workers with college degrees had weekly earnings of $874. Though college would generally increase graduate earnings, things may differ among different colleges, especially between public and private colleges.

In this study, we focused on a survey of 706 colleges, which includes both public and private colleges. This survey studies the graduate earnings, SAT and ACT score, college price, price with aid, need fraction, and merit aided. Graduate earnings are the median earnings of college graduates 5 years after graduation. SAT and ACT scores represent the difficulty level of admission. College price is the average tuition, and the price with aid is the tuition after tuition reduction. Need fraction is the portion of students that applied for tuition reduction and merit aided means the proportion of students that have been granted the tuition reduction.

In our analysis, we first focused on the price and earning relationship. Is college worth the expense? Which of public and private schools are better investments in terms of graduate earnings? We do so by means of Multiple Linear Regression and study the effects that the missing values in our dataset have on the results of the model.

2. Data Visualization

In this part, we focused on the price and earnings of 706 graduate schools, which include 438 private colleges and 268 public colleges. The first plot (fig.1) is an earning and price-scatter plot. The green dots represent public schools, and the blue dots represent private schools. The X-axis is the tuition of colleges, and the y-axis is the median graduate earning of each college. This plot shows that public schools and private schools clearly form two main populations. Meanwhile, with a price of 66600 and an earning of 53400, Amherst College is an obvious outlier from the public schools that are mixed in the private school population.

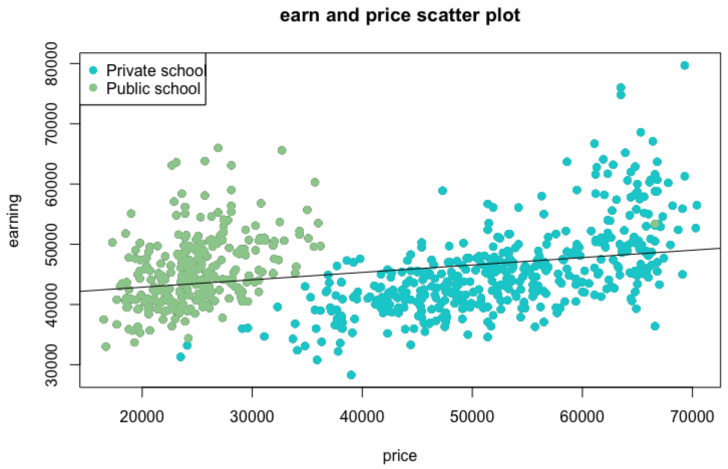


Fig1. Earning and price-scatter plot. Best fit line: y = 40420 + 0.1227x

Next, we used the Shapiro-Wilk test to examine whether the price and earning are normally distributed. In this test, the null-hypothesis is that the population is normally distributed. The P-value is 2.2e-16 for the price and 8.262e-16 for the earning, which indicates that neither of them is normally distributed. This result can be confirmed by the shape of the bar plot (fig.2).

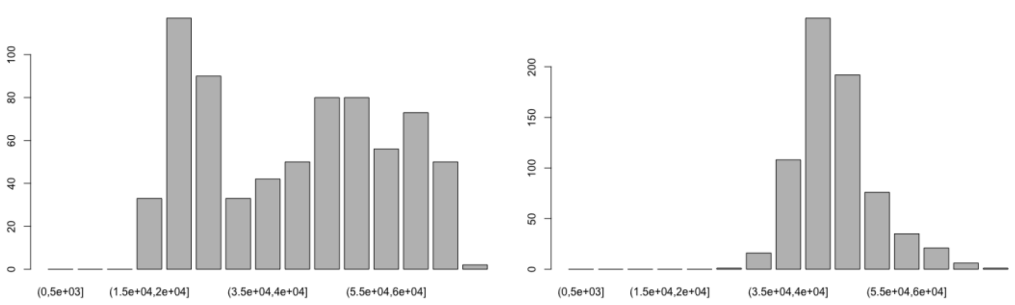


Fig.2 The frequency of price and earnings. Bat plot on the left side is the frequency distribution of the price. The bar plot on the right side is the frequency distribution of the earnings. The X-axis is the amount of money and the y-axis is the frequency of colleges.

Next, to better understand the difference between public and private schools, we introduce a new variable: ratio. The ratio is defined by earnings divided by price. We plot a boxplot to compare the ratio (Fig.3). In the box plot, the ratio of private schools is evidently lower than in public schools. The lowest ratio in private schools is Bennington College (0.55), while the highest ratio is Martin Luther College (1.38). The average is 0.88. Among public schools, the lowest ratio is Amherst College (0.80) and the highest ratio is North Carolina A&T State University (2.91). The average is 1.87. Z-test is performed to check whether the difference is statistically significant (table.1).

2.2 The First Hypothesis

The null hypothesis is that the ratio between earnings and the price is the same between private and public schools. Through the result of a t-test, the null hypothesis is rejected, and public colleges have significantly higher earning-to-price ratios compared with private schools.

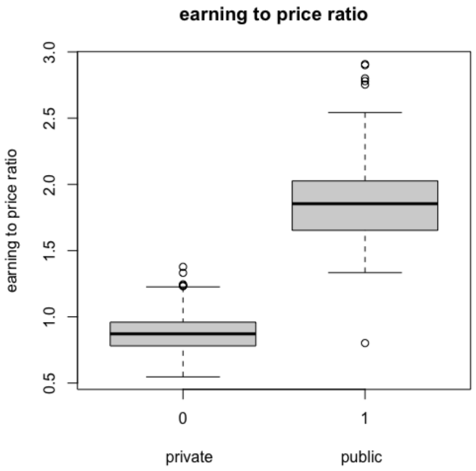


Fig.3 Boxplot of the earning to price ratio between private and public colleges

Table 1. z-test of the earning to price ratio between private and public colleges

	private	***public***
Mean	0.87858447	1.86705224
Variance	0.13	0.31
Observations	438	268
Hypothesized Mean Difference	0	
z	-25.926984	
P(Z<=z) one-tail	0	
z Critical one-tail	1.64485363	
P(Z<=z) two-tail	0	
z Critical two-tail	1.95996398	

3. Multiple Linear Regression

In this part, we are looking for multiple variables that are affecting each other and fit into multiple linear relationships. The hypothesis is to form the possible multiple linear regression model between the earning and the other predictor variables. There are seven predictor variables: public/private school, SAT score, ACT score, Tuition Price, Tuition price with aid, needing ratio of aid, and the merit aided ratio. It is expected to test which variables contribute to the future earning of college students. By conducting the least square technique, the multiple linear regression is formed, and then the goodness of the fit of the model is tested. Next, variables are evaluated using the stepwise regression method. Last, the missing data effect is discussed.

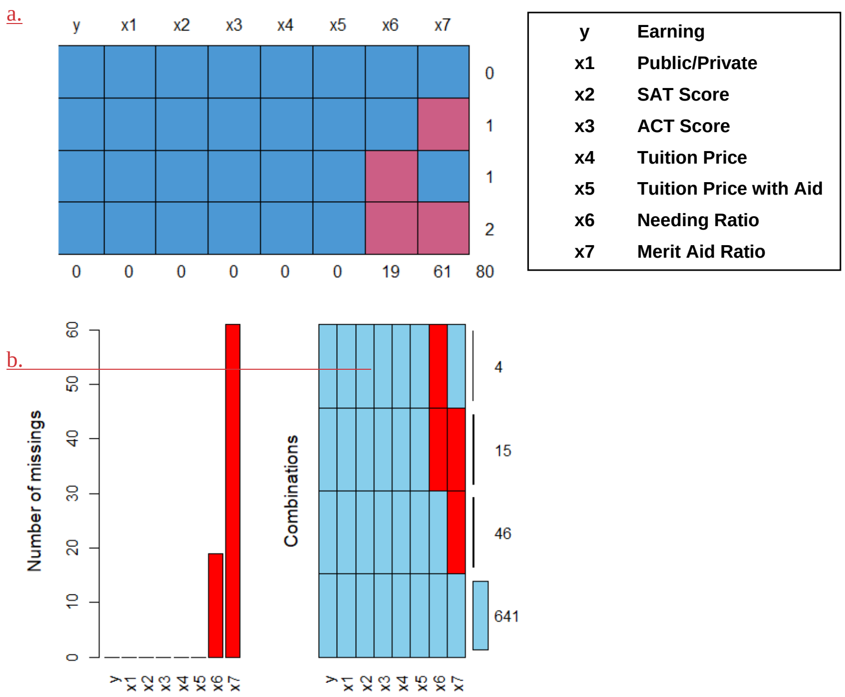


Fig. 4 Missing values distribution in the original dataset

3.1 The Second Hypothesis

The original data contain a certain number of missing values. Analyzing the data set in the program R commend, it was noticed that the original data contain 80 missing values, which is about 1.4% of the entire 7060 data values, the missing items belong to variables “Need Fraction” and “Merit Aided”. This can be shown in the plots of the data distribution in Fig 4. In the figure “y” represents earning, x1~x7 represents for Public/Private school, SAT score, ACT score, tuition price, tuition price with aid, needing ratio, and merit aid ratio. The red boxes represent missing data, and the number of missing data is 19 for x6(needing ratio) and 61 for x7(merit aided ratio).

For the initial analysis, the missing data rows are removed, therefore, 641 complete data groups are left. The effect of missing values is discussed in the later chapters. Apply the least square method in R to fit a multiple linear regression model to discover the effects of each variable ($x_1 \sim x_7$) to the future earing (y).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \epsilon$$

Set the hypothesis for multiple linear regression as:

$$H_o: \beta_0 \sim \beta_7 = 0 \text{ vs.} H_1: \text{at least one coefficient: } \beta_0 \sim \beta_7 \neq 0$$

The results of the analysis using R are shown as follow:

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = DATA_noNA)

Residuals:
    Min          1Q          Median       3Q          Max
-12184.7       -3242.1      -496.3       2602.4      24930.0

Coefficients:
              Estimate     Std.Error     t value   Pr(>|t|)
 (Intercept)  2.24E+04     3.06E+03      7.314     7.88E-13      ***
 x1           7.76E+03     1.05E+03      7.383     4.91E-13      ***
 x2           -8.02E-02    4.22E+00      -0.019    0.984831
 x3           5.03E+02     1.69E+02      2.986     0.002936      **
 x4           2.83E-01     4.05E-02      6.994     6.81E-12      ***
 x5           -2.27E-02    5.22E-02      -0.435    0.663612
 x6           -6.24E+03    1.66E+03      -3.758    0.000187      ***
 x7           -1.30E+03    2.26E+03      -0.576    0.564995
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4902 on 633 degrees of freedom
Multiple R-squared:  0.3888,      Adjusted R-squared:  0.382
F-statistic: 57.52 on 7 and 633 DF, p-value: < 2.2e-16
```

Formatted: Space After: 12 pt, Line spacing: single

	~~Estimate~~	~~Std.Error~~	~~t value~~	~~Pr(>|t|)~~	
~~(Intercept)~~	~~2.24E+04~~	~~3.06E+03~~	~~7.314~~	~~7.88E-13~~	~~***~~
~~x1~~	~~7.76E+03~~	~~1.05E+03~~	~~7.383~~	~~4.91E-13~~	~~***~~
~~x2~~	~~-8.02E-02~~	~~4.22E+00~~	~~-0.019~~	~~0.984831~~	

```
~~x3          5.03E+02     1.69E+02     2.986      0.002936      **~~
~~x4          2.83E-01     4.05E-02     6.994      6.81E-12      ***~~
~~x5         -2.27E-02     5.22E-02    -0.435      0.663612~~
~~x6         -6.24E+03     1.66E+03    -3.758      0.000187      ***~~
~~x7         -1.30E+03     2.26E+03    -0.576      0.564995~~
~~---~~
~~Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1~~

~~Residual standard error: 4902 on 633 degrees of freedom~~
~~Multiple R-squared:  0.3888,        Adjusted R-squared:  0.382~~
~~F-statistic: 57.52 on 7 and 633 DF, p-value: < 2.2e-16~~
```

According to the analysis, the equation of the least square of the multiple linear regression is given as:

$$y = 22400 + 7760x_1 - 0.0802x_2 + 503x_3 + 0.283x_4 - 0.0227x_5 - 6240x_6 - 1300x_7 + \epsilon.$$

Notice that the variable of x_1 which represents whether the data is from a public school or not is a dummy variable, which equals 1 for~~ data from a~~ public schools~~ data~~ and 0 for~~ a~~ private schools.

In the summary, the P-value is less than 2.2e-16 which is less $\alpha = 0.05$ level. In ~~an~~other words, ~~-~~the F statistics is 142.2, which is significantly large than the F distribution with $k = 7$ and $n - (k + 1) = 641 - (7 + 1) = 633$ degree of freedom, $f_{7,633,\alpha}$. To be convenient, the calculation results are summarized in the form of an analysis of variance table (ANOVA) in table 2. Therefore, reject the null hypothesis, and concluded that there is a statistical significance that at least one coefficient among $\beta_0 \sim \beta_7$ does not equal 0. On the other hand, the R-square value is 0.3888 and the adjusted R-square value is 0.382, which indicated that this multiple linear regression model is not ideally fitted.

Table 2 ANOVA Table for Multiple Regression Model

Source of Variance (Source)	**Sum of Squares (SS)**	**Degree of Freedom (d.f.)**	**Mean Square (MS)**	**F**
Regression	SSR = 1.5209e10	k = 7	MSR = SSR/k =2.17e9	MSR/MSE=142.2
Error	SSE = 9.6737e9	n-(k+1) = 633	MSE = SSE/(n-(k+1)) =1.53e7	
Total	SST = SSR+SSE = 2.49e10	n-1 = 640		

3.2 Distribution Diagnostics

Check the distribution of the residual ϵ using the Shapiro-Wilk test in R, which set the hypothesis:

H_o: ϵ follows normal distribution vs. H_1: ϵ does not follow a ~~a~~ normal distribution

The result is shown as follow:

```
           Shapiro-Wilk normality test
data:  y.res
W = 0.96031, p-value = 3.851e-12
```

Formatted: Space After: 12 pt

In the result, the p-value is 3.85e-12 which is significantly smaller than the significant level $\alpha = 0.05$,

In the result, the p-value is 3.85e-12 which is significantly smaller than the significant level $\alpha = 0.05$, that the null hypothesis cannot be rejected. Therefore, the Shapiro-Wilk test indicated that the residual ϵ does not follow the normal distribution. Fig 5a shows the plot of fitted values versus the residuals, the horizontal axis is the fitted values, and the vertical axis is the residuals. The figure indicated that data points are relatively concentrated. The red line represents the average value of the residuals which is not always around 0, but higher than zero on the left side, which shows that the distribution of the residual is not normal. Fig 5c is the QQ-plot of residuals, a normal distributed residual will form a straight line around the diagonal line. However, the current residual distribution does not cross the diagonal and there are large deviations on the two sides, the same conclusion can be drawn from Fig 5b, the distribution of the residual. Therefore, the current regression model using the seven variables is not a good representation of the relationship with future earnings.

In the result, the p-value is 3.85e-12 which is significantly smaller than the significant level $\alpha = 0.05$, that the null hypothesis cannot be rejected. Therefore, the Shapiro-Wilk test indicated that the residual ϵ does not follow the normal distribution. Fig 5a shows the plot of fitted values versus the residuals, the horizontal axis is the fitted values, and the vertical axis is the residuals. The figure indicated that data points are relatively concentrated. The red line represents the average value of the residuals which is not always around 0, but higher than zero on the left side, which shows that the distribution of the residual is not normal. Fig 5c is the QQ-plot of residuals, a normal distributed residual will form a straight line around the diagonal line. However, the current residual distribution does not cross the diagonal and there are large deviations on the two sides, the same conclusion can be drawn from Fig 5b, the distribution of the residual. Therefore, the current regression model using the seven variables is not a good representation of the relationship with future earnings.

In the result, the p-value is 3.85e-12 which is significantly smaller than the significant level $\alpha = 0.05$, that the null hypothesis cannot be rejected. Therefore, the Shapiro-Wilk test indicated that the residual ϵ does not follow the normal distribution. Fig 5a shows the plot of fitted values versus the residuals, the horizontal axis is the fitted values, and the vertical axis is the residuals. The figure indicated that data points are relatively concentrated. The red line represents the average value of the residuals which is not always around 0, but higher than zero on the left side, which shows that the distribution of the residual is not normal. Fig 5c is the QQ-plot of residuals, a normal distributed residual will form a straight line around the diagonal line. However, the current residual distribution does not cross the diagonal and there are large deviations on the two sides, the same conclusion can be drawn from Fig 5b, the distribution of the residual. Therefore, the current regression model using the seven variables is not a good representation of the relationship with future earnings.

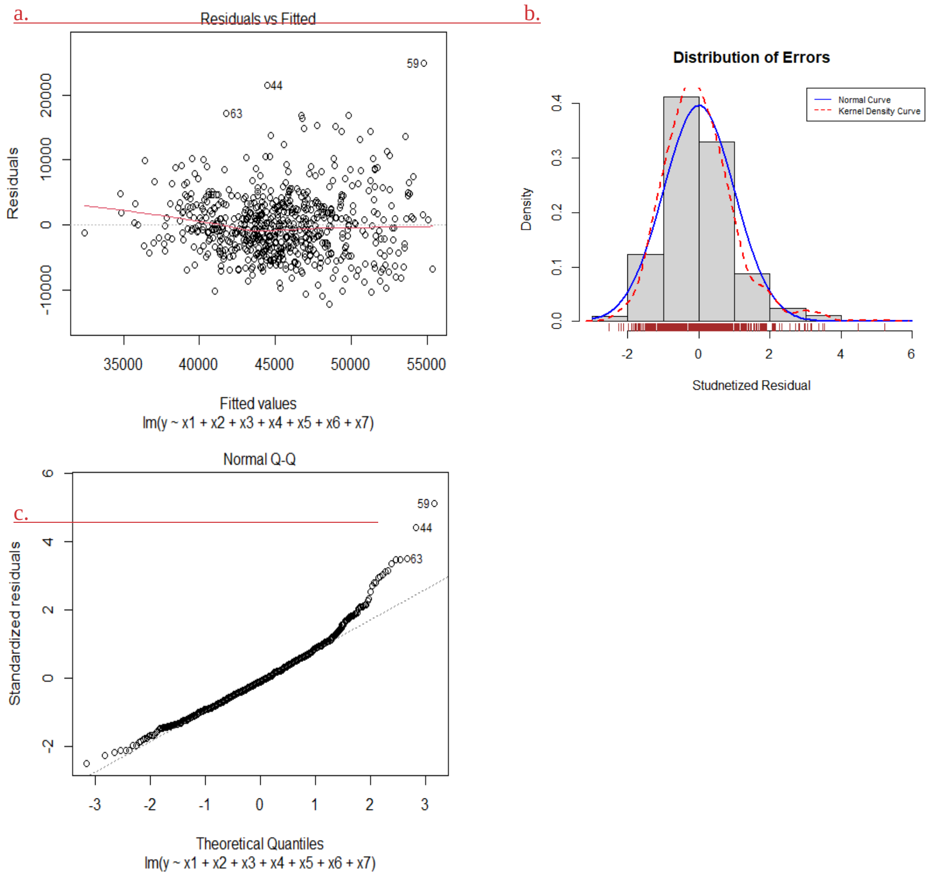


Fig 5. Residual distribution plots

3.3 Variable Selection

Similar to Fig. 5a which shows the relations between fitted values and the residuals, Fig. 6 shows the components residual plots to check the linearities of relationships between the earning(y) and the seven variables (x_i). From each plot, nonlinearities existing in each of the variables, therefore, the linear regression model needs to be improved. There are several ways that can improve the regression model, for example: removing extreme values, variable transformations, variable selections, and so on. For the current study, the variable selection methods are tested using R, specifically, backwards stepwise regression is tested.

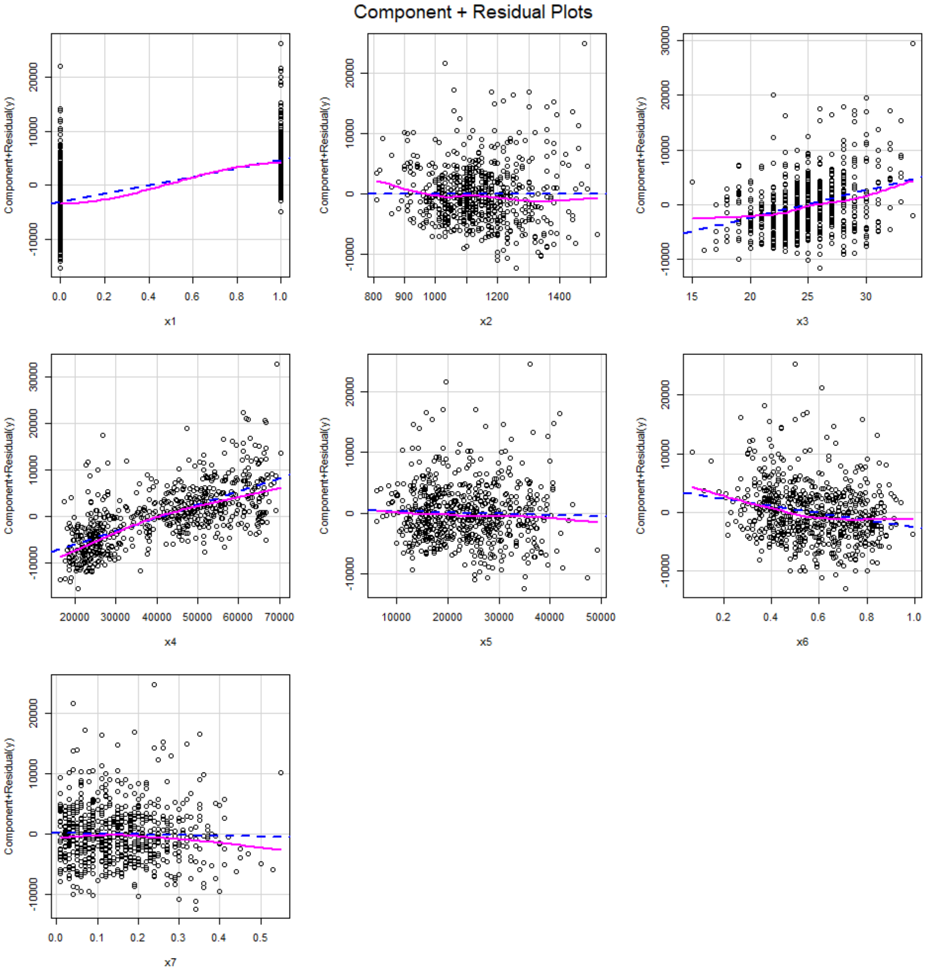


Fig. 6 Component-Residual Plot

3.4 Backward Stepwise Regression

Stepwise regression is a method that screens the candidate predictor variables from a multiple linear regression model. By removing variables one at a time, the backward stepwise regression can evaluate the multiple regression models with different predictor variables by applying the Partial F-test or the Akaike Information Criterion (AIC). For the current study, using the build-in R command, the Akaike

Information Criterion (AIC) is used to compare models with a reducing number of the variables. In each step, the model with the lowest AIC value is selected.

```
Start:  AIC=10901.54
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7
            Df      Sum of Sq    RSS        AIC
 x2         1       8692         1.52E+10   10900
 x5         1       4549300      1.52E+10   10900
 x7         1       7964289      1.52E+10   10900
 <none>                          1.52E+10   10902
 x3         1       2.14E+08     1.54E+10   10908
 x6         1       3.39E+08     1.55E+10   10914
 x4         1       1.18E+09     1.64E+10   10947
 x1         1       1.31E+09     1.65E+10   10952

Step:  AIC=10899.54
y ~ x1 + x3 + x4 + x5 + x6 + x7
           Df       Sum of Sq   RSS        AIC
 x5        1        4544792     1.52E+10   10898
 x7        1        8044855     1.52E+10   10898
 <none>                         1.52E+10   10900
 x6        1        3.43E+08    1.56E+10   10912
 x3        1        7.57E+08    1.60E+10   10929
 x4        1        1.2E+09     1.64E+10   10946
 x1        1        1.34E+09    1.65E+10   10952

Step:  AIC=10897.73
y ~ x1 + x3 + x4 + x6 + x7
           Df      Sum of Sq    RSS        AIC
 x7        1       14124811     1.52E+10   10896
 <none>                         1.52E+10   10898
 x6        1       3.39E+08     1.56E+10   10910
 x3        1       7.84E+08     1.60E+10   10928
 x1        1       1.34E+09     1.66E+10   10950
 x4        1       1.73E+09     1.69E+10   10965

Step:  AIC=10896.32
y ~ x1 + x3 + x4 + x6
           Df       Sum of Sq    RSS         AIC
 <none>                          1.52E+10    10896
 x6        1        3.28E+08     1.56E+10    10908
 x3        1        7.7E+08      1.60E+10    10926
 x1        1        1.52E+09     1.67E+10    10955
 x4        1        1.76E+09     1.70E+10    10965

Call:
lm(formula = y ~ x1 + x3 + x4 + x6, data = DATA_noNA)

Coefficients:
 (Intercept)   x1         x3         x4        x6
 21916.85      7966.072   494.8219   0.2746    -5997.22
```

	~~Df~~	~~Sum of Sq~~	~~RSS~~	~~AIC~~
~~x2~~	~~1~~	~~8692~~	~~1.52E+10~~	~~10900~~
~~x5~~	~~1~~	~~4549300~~	~~1.52E+10~~	~~10900~~
~~x7~~	~~1~~	~~7964289~~	~~1.52E+10~~	~~10900~~
~~<none>~~			~~1.52E+10~~	~~10902~~
~~x3~~	~~1~~	~~2.14E+08~~	~~1.54E+10~~	~~10908~~
~~x6~~	~~1~~	~~3.39E+08~~	~~1.55E+10~~	~~10914~~
~~x4~~	~~1~~	~~1.18E+09~~	~~1.64E+10~~	~~10947~~
~~x1~~	~~1~~	~~1.31E+09~~	~~1.65E+10~~	~~10952~~

~~Step: AIC=10899.54~~
~~y ~ x1 + x3 + x4 + x5 + x6 + x7~~

	~~Df~~	~~Sum of Sq~~	~~RSS~~	~~AIC~~
~~x5~~	~~1~~	~~4544792~~	~~1.52E+10~~	~~10898~~
~~x7~~	~~1~~	~~8044855~~	~~1.52E+10~~	~~10898~~
~~<none>~~			~~1.52E+10~~	~~10900~~
~~x6~~	~~1~~	~~3.43E+08~~	~~1.56E+10~~	~~10912~~
~~x3~~	~~1~~	~~7.57E+08~~	~~1.60E+10~~	~~10929~~
~~x4~~	~~1~~	~~1.2E+09~~	~~1.64E+10~~	~~10946~~
~~x1~~	~~1~~	~~1.34E+09~~	~~1.65E+10~~	~~10952~~

~~Step: AIC=10897.73~~
~~y ~ x1 + x3 + x4 + x6 + x7~~

	~~Df~~	~~Sum of Sq~~	~~RSS~~	~~AIC~~
~~x7~~	~~1~~	~~14124811~~	~~1.52E+10~~	~~10896~~
~~<none>~~			~~1.52E+10~~	~~10898~~
~~x6~~	~~1~~	~~3.39E+08~~	~~1.56E+10~~	~~10910~~
~~x3~~	~~1~~	~~7.84E+08~~	~~1.60E+10~~	~~10928~~
~~x1~~	~~1~~	~~1.34E+09~~	~~1.66E+10~~	~~10950~~
~~x4~~	~~1~~	~~1.73E+09~~	~~1.69E+10~~	~~10965~~

~~Step: AIC=10896.32~~
~~y ~ x1 + x3 + x4 + x6~~

	~~Df~~	~~Sum of Sq~~	~~RSS~~	~~AIC~~
~~<none>~~			~~1.52E+10~~	~~10896~~
~~x6~~	~~1~~	~~3.28E+08~~	~~1.56E+10~~	~~10908~~
~~x3~~	~~1~~	~~7.7E+08~~	~~1.60E+10~~	~~10926~~
~~x1~~	~~1~~	~~1.52E+09~~	~~1.67E+10~~	~~10955~~
~~x4~~	~~1~~	~~1.76E+09~~	~~1.70E+10~~	~~10965~~

~~Call:~~
~~lm(formula = y ~ x1 + x3 + x4 + x6, data = DATA_noNA)~~

~~Coefficients:~~

~~(Intercept)~~	~~x1~~	~~x3~~	~~x4~~	~~x6~~

~~21916.85 7966.072 494.8219 0.2746 -5997.22~~

After backward stepwise regression, four of the predictor variables are selected, i.e. Public/private school, ACT score, tuition price, and needing ratio. Summarize the calculation results in the ANOVA table 3. The new least-square relation is given as:

$$y = 21916.85 + 7966.1x_1 + 494.8x_3 + 0.2746x_4 - 5997.2x_6 + \epsilon.$$

Table 3 ANOVA table for the new regression equation

Source of Variance (Source)	Sum of Squares (SS)	Degree of Freedom (d.f.)	Mean Square (MS)	F
Regression	SSR = 1.5228e10	k = 4	MSR = SSR/k =3.81e9	MSR/MSE=249.98
Error	SSE = 9.686e9	n-(k+1) = 636	MSE = SSE/(n-(k+1)) =1.52e7	
Total	SST = SSR+SSE = 2.49e10	n-1 = 640		

The new least-square relationship has a larger F-statistics, yet the R square value and adjusted R square value are 0.388 and 0.384 respectively, which do not change much. To compare the two least square models, a Partial-F test has been conducted, and the result is shown as follow:

New variables: $y = 21916.85 + 7966.1x_1 + 494.8x_3 + 0.2746x_4 - 5997.2x_6 + \epsilon$

Old model: $y = 22400 + 7760x_1 - 0.0802x_2 + 503x_3 + 0.283x_4 - 0.0227x_5 - 6240x_6 - 1300x_7 + \epsilon$.

```
Model 1: y ~ x1 + x3 + x4 + x6
Model 2: y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7
  Res.Df       RSS          Df     Sum of Sq       F       Pr(>F)
1   636        1.5228e+10
2   633        1.5209e+10   3      18678295        0.2591  0.8548
```

Formatted: Line spacing: Multiple 1.15 li

~~Model 1: y ~ x1 + x3 + x4 + x6~~

~~Model 2: y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7~~

~~Res.Df RSS Df Sum of Sq F Pr(>F)~~

~~1 636 1.5228e+10~~

~~2 633 1.5209e+10 3 18678295 0.2591 0.8548~~

Formatted: Space After: 12 pt

The F statistics is less than 1, smaller than the F distribution value, $f_{3,633,\alpha} = 2.60$ at $\alpha = 0.05$ level, which indicated that the removed variables (x2, x5, and x7) are not contributing much to earning(y), the existence of the other four variables (x1, x3, x4 and x6) are already accounted for them. This result makes sense because some of the variables have repeated representations. For example, the SAT score (x2) and the ACT score(x3) should play a similar role in the model, which means for a certain school, the criteria for both SAT and ACT scores will be similar, in another word, the two variables are correlated, therefore, only one is needed. Similar to the other variables: tuition price is correlated with tuition price with aid, as well as the needing ratio and the aided ratio. This can be proved by the correlations between the predictor variables shown in fig. 7. All the variables in the data are paired up to find the linear relationship. From the plot, the SAT and ACT scores have a strong linear correlation, so does price and price with aid. Therefore, for the current study, the multiple linear regression can be simplified from seven variables to four variables, yet the goodness of the fit of the model cannot be further improved.

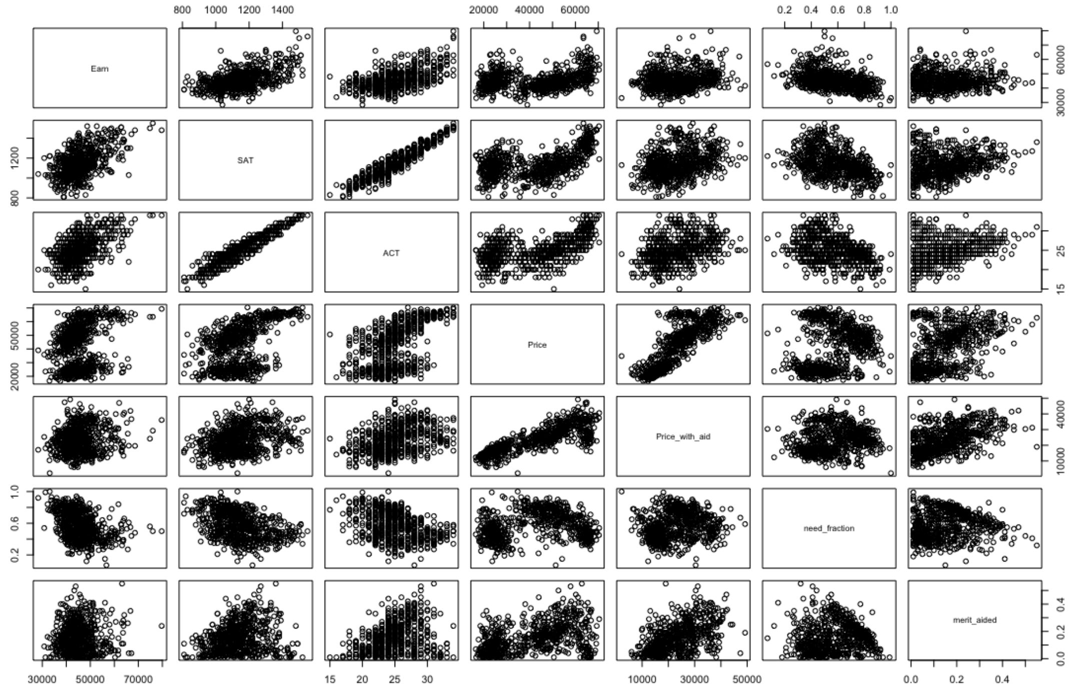
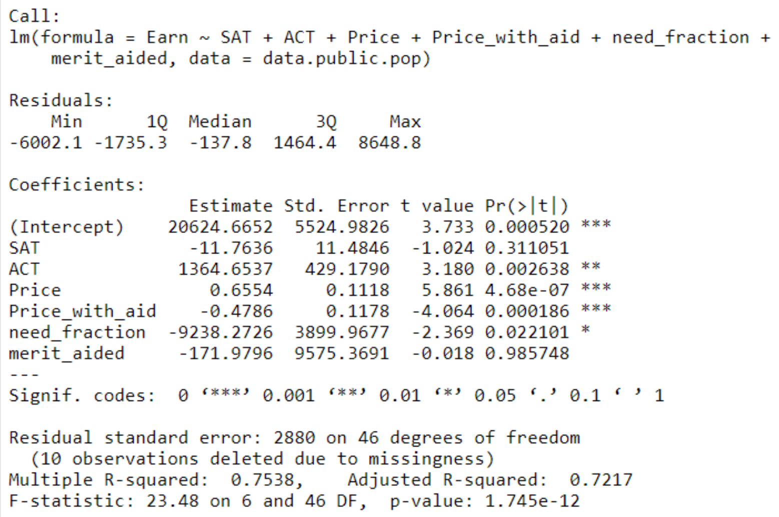


Fig.7 Scatter plot of all numerical variables pairing up forming correlation.

3.5 Multiple Linear Regression in the Pop Region (NY, CA & PA)

In this section, the attention is focused on public and private colleges in NY, CA and PA, the three states that have a great number of colleges, and assess which features are critical in determining graduate earnings. The hypothesis is that graduate earnings depend on prices. It is natural that schools help graduates land well-paying jobs are in demand, and hence pricey to attend. The result of the multiple linear regression models for the Public schools, and the Private schools in Pop Regions (NY, CA & PA) are shown as follow. The results indicate that, to predict Graduate Earnings for Public Schools in NY, CA and PA regions, Price, Price_with_aid and ACT are valuable features that produce significant results. For Private Schools in these pop regions too, the Price variable along with need_fraction plays a pivotal role in generating significant results. Thus, it can be concluded that Graduate Earnings for the states of NY, CA and PA greatly depend on the cost of education.

Public Schools:



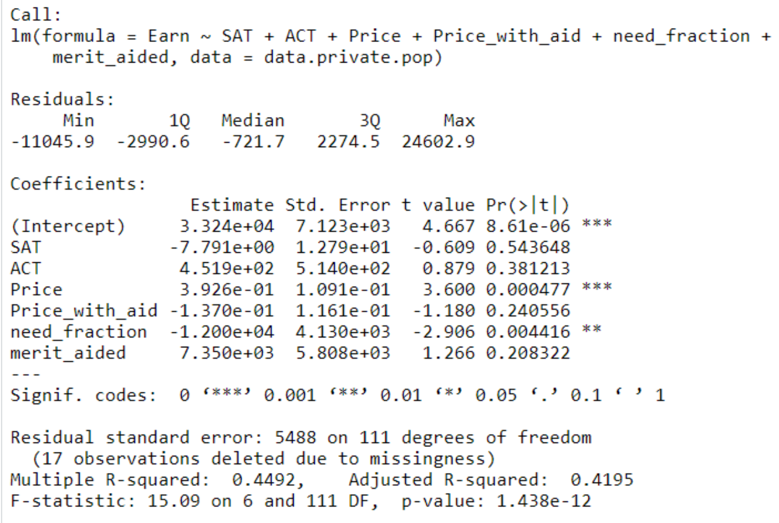
```
Call:
lm(formula = Earn ~ SAT + ACT + Price + Price_with_aid + need_fraction +
    merit_aided, data = data.public.pop)

Residuals:
    Min      1Q  Median      3Q     Max 
-6002.1 -1735.3  -137.8  1464.4  8648.8 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    20624.6652  5524.9826   3.733 0.000520 ***
SAT              -11.7636    11.4846  -1.024 0.311051    
ACT             1364.6537   429.1790   3.180 0.002638 ** 
Price              0.6554     0.1118   5.861 4.68e-07 ***
Price_with_aid    -0.4786     0.1178  -4.064 0.000186 ***
need_fraction  -9238.2726  3899.9677  -2.369 0.022101 *  
merit_aided     -171.9796  9575.3691  -0.018 0.985748    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2880 on 46 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.7538,	Adjusted R-squared:  0.7217 
F-statistic: 23.48 on 6 and 46 DF,  p-value: 1.745e-12
```

Private Schools:



```
Call:
lm(formula = Earn ~ SAT + ACT + Price + Price_with_aid + need_fraction +
    merit_aided, data = data.private.pop)

Residuals:
     Min       1Q   Median       3Q      Max 
-11045.9  -2990.6   -721.7   2274.5  24602.9 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     3.324e+04  7.123e+03   4.667 8.61e-06 ***
SAT            -7.791e+00  1.279e+01  -0.609 0.543648    
ACT             4.519e+02  5.140e+02   0.879 0.381213    
Price           3.926e-01  1.091e-01   3.600 0.000477 ***
Price_with_aid -1.370e-01  1.161e-01  -1.180 0.240556    
need_fraction  -1.200e+04  4.130e+03  -2.906 0.004416 ** 
merit_aided     7.350e+03  5.808e+03   1.266 0.208322    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5488 on 111 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared:  0.4492,	Adjusted R-squared:  0.4195 
F-statistic: 15.09 on 6 and 111 DF,  p-value: 1.438e-12
```

4. Effect of Missing values

In this section, the effect of the missing values is discussed. In reality, missing data is inevitable in almost every research. Take the current study for example, as mentioned in the previous section, the original data has about 1.4% missing values from the variables needed ratio and merited aided ratio. In this section, two missing values scenarios will be discussed: the missing data that are completely random and non-ignorable missing values.

4.1 Missing Completely at Random (MCAR)

By definition, missing completely at random (MCAR) means that the missing data is irrelevant to either the observable variables and the unobservable parameters of interest. The missed values and observations tend to have similar distributions. (Bhaskaran and Smeeth 2014). In addition to the existing missing values in the original dataset, 20% more of the data is removed randomly from the original dataset using R. A comparison of the randomly-removed-datasets and the original one is shown in Fig. 8, the red blocks represent the locations of the missing data. The missing data are randomly distributed amount the variables. The variables x6 and x7 have a higher number of missing values in the new dataset, this is due to the original missing data.

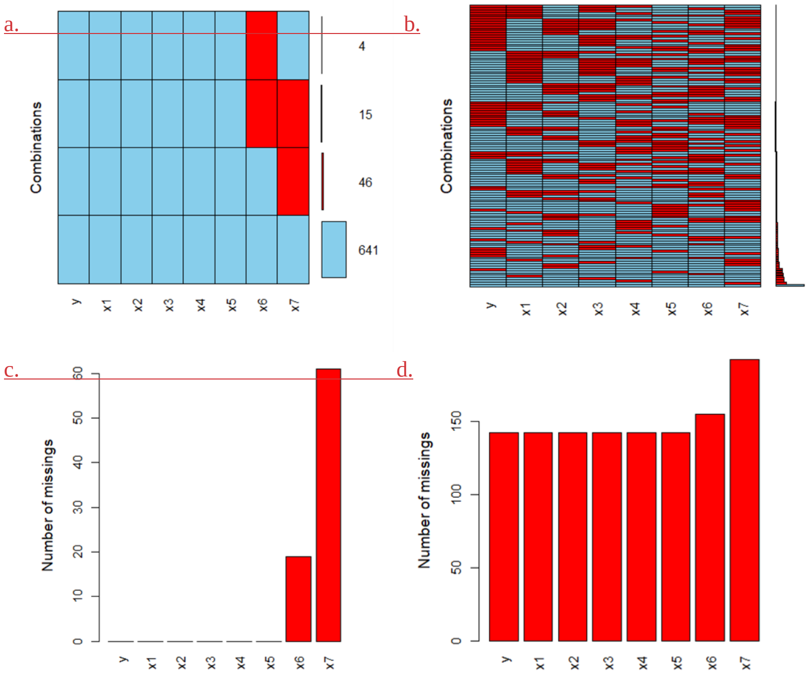


Fig. 8 Comparison of the missing data between the original data and the 20% randomly missing data

Rerun the least square analysis to the 20% missing value model. The calculation is summarized in table 4, and the result is listed as follow:

```
Call:
lm(formula = y ~ ., data = DATA_20_NA)

Residuals:
    Min      1Q  Median      3Q     Max 
-13027.7  -3398.9    -14.4   2859.1  12192.3 

Coefficients:
               Estimate    Std. Error   t value     Pr(>|t|)
 (Intercept)   1.13E+04     7.32E+03      1.547      0.124842
 x1            9.91E+03     2.36E+03      4.205      5.61E-05    ***
 x2            1.04E+01     1.03E+01      1.013      0.313551
 x3            1.85E+02     3.99E+02      0.463      0.644672
 x4            3.70E-01     9.48E-02      3.907      0.000168    ***
 x5            6.84E-02     1.35E-01      0.506      0.614172
 x6           -4.24E+03     3.72E+03     -1.139      0.257534
 x7           -4.15E+03     5.73E+03     -0.725      0.470073

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4765 on 102 degrees of freedom
  (596 observations deleted due to missingness)
Multiple R-squared:  0.5009,	Adjusted R-squared:  0.4666 
F-statistic: 14.62 on 7 and 102 DF,  p-value: 4.594e-13
```

~~Call:~~
~~lm(formula = y ~ ., data = DATA_20_NA)~~

~~Residuals:~~
~~Min 1Q Median 3Q Max~~
~~-13027.7 -3398.9 -14.4 2859.1 12192.3~~

~~Coefficients:~~

	~~Estimate~~	~~Std. Error~~	~~t value~~	~~Pr(>|t|)~~	
~~(Intercept)~~	~~1.13E+04~~	~~7.32E+03~~	~~1.547~~	~~0.124842~~	
~~x1~~	~~9.91E+03~~	~~2.36E+03~~	~~4.205~~	~~5.61E-05~~	~~***~~
~~x2~~	~~1.04E+01~~	~~1.03E+01~~	~~1.013~~	~~0.313551~~	
~~x3~~	~~1.85E+02~~	~~3.99E+02~~	~~0.463~~	~~0.644672~~	
~~x4~~	~~3.70E-01~~	~~9.48E-02~~	~~3.907~~	~~0.000168~~	~~***~~
~~x5~~	~~6.84E-02~~	~~1.35E-01~~	~~0.506~~	~~0.614172~~	

Formatted: Line spacing: Multiple 1.15 li

Formatted: Font: (Default) +Headings (Calibri), 10.5 pt, Font color: Text 2

~~x6 -4.24E+03 3.72E+03 -1.139 0.257534~~
~~x7 -4.15E+03 5.73E+03 -0.725 0.470073~~

~~---~~
~~Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1~~

~~Residual standard error: 4765 on 102 degrees of freedom~~
~~(596 observations deleted due to missingness)~~
~~Multiple R-squared: 0.5009, Adjusted R-squared: 0.4666~~
~~F-statistic: 14.62 on 7 and 102 DF, p-value: 4.594e-13~~

Formatted: Left

Table 4. ANOVA table for the regression equation

Source of Variance (Source)	Sum of Squares (SS)	Degree of Freedom (d.f.)	Mean Square (MS)	F
Regression	SSR = 2.316e9	k = 7	MSR = SSR/k =3.309e8	MSR/MSE=14.52
Error	SSE = 2.324e9	n-(k+1) = 102	MSE = SSE/(n-(k+1)) =2.279e7	
Total	SST = SSR+SSE = 4.64e9	n-1 = 108		

The results indicated that an additional 20% completely random data will have a significant influence on the multiple regression model. The new relationship has only two noticeable correlated prediction variables: x1 and x4. The R-square value and adjusted R-square value are 0.5009 and 0.4666 respectively. However, the degree of freedom reduced from 631 to 102 which reduced the statistical power. Therefore, the missing data up to 20% should be treated with caution, because the conclusion that draw from the fragmentary dataset could be misleading.

4.2 Nonignorable ~~missing~~ Missing ~~data~~Data

Formatted: Space After: 12 pt

In certain situation, the missing data mechanism is nonignorable, for example, Franks et. al. (2020) introduced two modeling approaches: pattern-mixture models: the distributions of missing data and observation data are different; and the selection models: select observed data under a missing-data mechanism with the distribution of the data pre-observation. One of the techniques to deal with the nonignorable missing data, according to Rubin (1974) and Rubin (1976) is the missingness mechanism, which is to generate a standard data complete model first and then select observed data from the complete data. Another technique is to find the separate distribution for the observed data and the missing data, and then prove the explicit assumption can be avoided. (Little 1993). The sensitivity analysis is also an important technique, which can increase the insight into the stability of the result. (Buuren 2018) Therefore, for the current study to deal with nonignorable missing data, the missingness mechanism, and the sensitivity analysis could be used first, and the separate distribution technique could be applied to the missing data as a comparison.

5. Conclusion

This study focused on a survey of 706 colleges all over the Unite States, which includes both public and private colleges. This survey studies the several variables, including graduate earnings, SAT and ACT scores, college price, price with aid, need fraction and merit aided. Based on the survey data, two statistical studies have been conducted: firstly, test the hypothesis that the ratio between earning and the price is the same between private and public schools, and secondly, generate multiple linear regression models between the earnings and the other predictor variables. And the mainly conclusions are listed as follow:

Formatted: Font:

1. The public colleges have significantly higher earning-to-price ratios compared with private schools.

2. The multiple linear regression model has been generated, and through the backward stepwise regression method, four of the critical predictor variables are selected: Public/private school, ACT score, tuition price, and needing ratio.

Formatted: Font color: Auto

3. Within the Pop region (NY, CA & PA), the multiple linear regression models show that the Price variable along with need_fraction are important to the graduate earning, which emphasizes the importance of the cost of education.

Formatted: Font: (Default) Times New Roman

4. The effect of the missing data completely at random (MCAR) up to 20% could cause misleading of the statistical conclusion.

Formatted: Font: (Default) Times New Roman

5. By doing literature reviews about the nonignorable missing data, it has been learned the techniques such as the missingness mechanism and the sensitivity analysis can be applied to handle this type of missing data.

6. Reference

Rubin, D. B. (1974), “Characterizing the estimation of parameters in incomplete data problems,” Journal of the American Statistical Association, 69, 467–474.

Rubin, D. B. (1976), “Inference and Missing Data,” Biometrika, 63, 581.

Little, R. J. (1993), “Pattern-mixture models for multivariate incomplete data,” Journal of the American Statistical Association, 88, 125–134.

Alexander M. Franks, Edoardo M. Airoldi, Donald B. Rubin Proceedings of the National Academy of Sciences Aug 2020, 117 (32) 19045-19053; DOI: 10.1073/pnas.1815563117

Formatted: Font: (Default) Times New Roman

Formatted: Font: (Default) Times New Roman

Van Buuren, S. (2018). Flexible Imputation of Missing Data, Second Edition. New York: Chapman and Hall/CRC, https://doi.org/10.1201/9780429492259

Formatted: Font: (Default) Times New Roman

7. Source code :

Part 1 R code:

```
> setwd("~/Desktop/data analysis")
> earnings = read.delim("graduate-earnings.txt", header = TRUE)
> attach(earnings)

> plot(earnings$Price, earnings$Earn, xlab='price', ylab='earning', main='earn and price scatter plot',
  pch=19)
> points(earnings$Price[earnings$Public=="0"], earnings$Earn[earnings$Public=="0"], pch=19,
  col="darkturquoise")
> points(earnings$Price[earnings$Public=="1"], earnings$Earn[earnings$Public=="1"], pch=19,
  col="darkseagreen3")
> legend("topleft", pch=19, col=c("darkturquoise", "darkseagreen3"), c("Private school", "Public
  school"))

> barplot(table(cut(earnings$Price, breaks = seq(0,75000, by = 5000))))
> barplot(table(cut(earnings$Earn, breaks = seq(0,75000, by = 5000))))
> summary(lm(earnings$Earn~earnings$Price))
Call:
lm(formula = earnings$Earn ~ earnings$Price)
Residuals:
    Min       1Q   Median       3Q      Max
-16905.1  -4183.1   -921.5   3217.7  30777.6
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.042e+04  6.951e+02  58.150  < 2e-16 ***
earnings$Price 1.227e-01  1.544e-02   7.948 7.55e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 6446 on 704 degrees of freedom
Multiple R-squared:  0.08234, Adjusted R-squared:  0.08103
F-statistic: 63.17 on 1 and 704 DF,  p-value: 7.552e-15

> splitearnings=split(earnings,earnings$public)
> publicschools=splitearnings[1]
> privateschools=splitearnings[2]
> boxplot(earnings$ratio~earnings$Public, xlab=c("private public"), ylab='earning to price ratio',
   main='earning to price ratio')
```

Part 2 R code

```
Call:
lm(formula = SAT ~ Price + merit_aided)

Residuals:
    Min      1Q  Median      3Q     Max
-330.43  -80.15   -3.39   77.16  305.51
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.858e+02  1.296e+01  76.063   <2e-16 ***
Price       3.443e-03  3.219e-04  10.699   <2e-16 ***
merit_aided 3.387e+01  4.647e+01   0.729    0.466
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 113.4 on 642 degrees of freedom
  (61 observations deleted due to missingness)
Multiple R-squared:  0.1883,   Adjusted R-squared:  0.1857
F-statistic: 74.44 on 2 and 642 DF,  p-value: < 2.2e-16

Call:
lm(formula = Earn ~ SAT + ACT)
Residuals:
    Min       1Q   Median       3Q      Max
-15689.4  -3542.0   -278.7   3132.2  24360.7
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 15508.123   1801.657   8.608  < 2e-16 ***
SAT           15.154      4.374   3.464 0.000564 ***
ACT          511.855    173.103   2.957 0.003211 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 5572 on 703 degrees of freedom
Multiple R-squared:  0.3152,   Adjusted R-squared:  0.3132
F-statistic: 161.8 on 2 and 703 DF,  p-value: < 2.2e-16

Call:
lm(formula = SAT ~ merit_aided + Price_with_aid)
Residuals:
    Min      1Q  Median      3Q     Max
-302.81  -81.69  -11.83   70.82  364.69
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.003e+03  1.402e+01  71.510  < 2e-16 ***
merit_aided    3.195e+01  5.078e+01   0.629     0.53
Price_with_aid 5.357e-03  6.581e-04   8.139 2.07e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 117.1 on 642 degrees of freedom
  (61 observations deleted due to missingness)
Multiple R-squared:  0.133,     Adjusted R-squared:  0.1303
F-statistic: 49.24 on 2 and 642 DF,  p-value: < 2.2e-16

Call:
lm(formula = Price_with_aid ~ merit_aided + Price)
Residuals:
     Min       1Q   Median       3Q      Max
-19036.1  -2285.0    199.8   2317.9  16857.7
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.753e+03  4.348e+02   8.631   <2e-16 ***
merit_aided 1.381e+04  1.559e+03   8.859   <2e-16 ***
Price       4.250e-01  1.080e-02  39.357   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3803 on 642 degrees of freedom
  (61 observations deleted due to missingness)
Multiple R-squared:  0.7859,   Adjusted R-squared:  0.7852
F-statistic:  1178 on 2 and 642 DF,  p-value: < 2.2e-16

> step(fit_before_no_NA,direction="backward")
Start:  AIC=10901.54
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7

       Df  Sum of Sq        RSS   AIC
```

```
- x2    1       8692 1.5209e+10 10900
- x5    1    4549300 1.5213e+10 10900
- x7    1    7964289 1.5217e+10 10900
<none>               1.5209e+10 10902
- x3    1  214221235 1.5423e+10 10908
- x6    1  339325275 1.5548e+10 10914
- x4    1 1175270813 1.6384e+10 10947
- x1    1 1309537107 1.6518e+10 10952

Step:  AIC=10899.54
y ~ x1 + x3 + x4 + x5 + x6 + x7

       Df  Sum of Sq        RSS   AIC
- x5    1    4544792 1.5213e+10 10898
- x7    1    8044855 1.5217e+10 10898
<none>               1.5209e+10 10900
- x6    1  343336237 1.5552e+10 10912
- x3    1  756775161 1.5966e+10 10929
- x4    1 1203814650 1.6413e+10 10946
- x1    1 1336748902 1.6546e+10 10952

Step:  AIC=10897.73
y ~ x1 + x3 + x4 + x6 + x7

       Df  Sum of Sq        RSS   AIC
- x7    1   14124811 1.5228e+10 10896
<none>               1.5213e+10 10898
- x6    1  338898169 1.5552e+10 10910
- x3    1  783856540 1.5997e+10 10928
- x1    1 1336485748 1.6550e+10 10950
- x4    1 1731663593 1.6945e+10 10965

Step:  AIC=10896.32
y ~ x1 + x3 + x4 + x6

       Df  Sum of Sq        RSS   AIC
<none>               1.5228e+10 10896
- x6    1  327506009 1.5555e+10 10908
- x3    1  770046199 1.5998e+10 10926
- x1    1 1517503539 1.6745e+10 10955
- x4    1 1763550039 1.6991e+10 10965

Call:
lm(formula = y ~ x1 + x3 + x4 + x6, data = DATA_noNA)

Coefficients:
(Intercept)           x1           x3           x4           x6  
 21916.8537    7966.0722     494.8219       0.2746   -5997.2193  
```

```
> shapiro.test(y.res)

	Shapiro-Wilk normality test

data:  y.res
W = 0.96031, p-value = 3.851e-12
```

Missing data R code

```
> nac <- NAControl(NArate = 0.2)

> DATA_20_NA <- setNA(DATA3,nac)
```