College's predicted profit by sport

Standard Deviants - Ava Exelbirt, Yura Heo, Claire Li

2023-11-07

Introduction and data

In our research project, we aim to investigate and understand the factors influencing the profit generated by collegiate sport, specifically basketball, in the school year 2019 to 2020. We will be utilizing a dataset sourced from Tidy Tuesday, which provides a comprehensive collection of observations related to collegiate sports, including information about the schools, classifications, sport types, and various quantitative variables describing the sport players and the schools' financial investment in sports programs in the dataset called "sports.csv". Our primary motivation for this research is to gain insights into the determinants of collegiate basketball profit, which can be of significant interest to educational institutions and further local policy makers and sports enthusiasts. Understanding the factors that contribute to profit generation in collegiate basketball sports can inform decision-making, investment strategies in sports, and future planning for universities and colleges involved in sports programs. We chose to do profit as our response variable because expenditures and revenue are closely related, so expenditure would overpower the other predictor variables if we were to predict revenue or expenditure. Therefore, we will create a new variable, profit calculated from revenue minus expenditures and predict this.

Our primary research question is as follows: **In the school year 2019 to 2020, how can we predict the profit (revenue- expenditure) in USD of the collegiate sport basketball using participation rate, school sector name, gender ratio, total count of students, percent of expenditures towards women's sports, and school classification name**.

Our hypothesis is as follows: **Participation rate, sector name, and gender ratio will be the most influential predictors for the total profit generated by the collegiate sport basketball in USD.**

The data set was taken from TidyTuesday and was originally scraped from Equity in Athletics Data Analysis (EADA), a sector of the US Department of Education. The data is available on an online database found on the (EADA) website1.

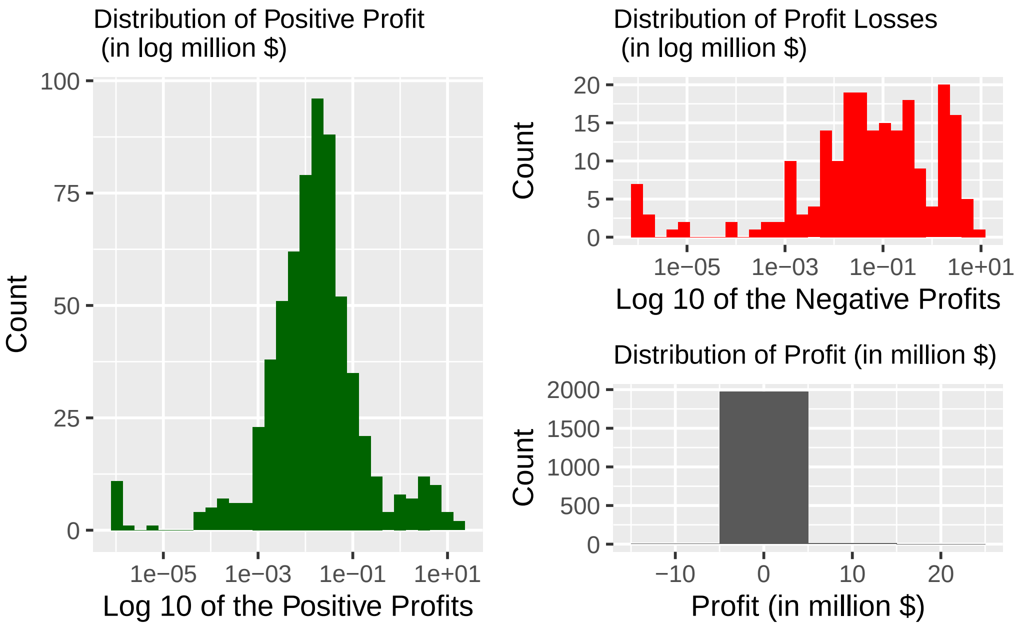
This data is submitted annually from colleges to the EADA. All co-educational postsecondary institutions that receive Title IV funding that have intercollegiate athletics programs are required by the Equity in Athletics Disclosure Act to submit this data. The original data files are also created immediately after the data collection for each school. These data are collected annually starting from 2003 to 2022, but for our specific data analysis we will only look at data taken from the school year of 2019 by cleaning the data to a new csv file which we will use for the rest of the project. The csv file from Tidy Tuesday contains thousands of observations from years 2015-2019, but we have adjusted the data file to only include the year 2019 and will filter the data to only include basketball as a sport, as this is our population we are analyzing.

The dataset we will use from the data scraped on Tidy Tuesday include many observations regarding collegiate sports. These observations include variables such as the name of the city which a school is in, the state, the school name, the classification of the school (like whether it is NCAA Division I, II, or III), the type/sector name of the school (like 4 year accredited university), sport, and many quantitative variables regarding characteristics about the specific school and their collegiate spending. Most of the quantitative variables are split between men and women, having two different observations for the same variable. For example, there is total male population, total female population, participation rate of women, participation rate of men, revenue for men, revenue for women, expenditures for men, and expenditures for women. There are also observations for the total amount for each of these above variables which includes both men and women; for example, total expenditures for both men and women together. The observations are therefore both quantitative and categorical and measure characteristics of different school's spending, revenue, populations, locations, and sports.

The key variables we will use are **participation rate** which is the total percentage of men and women students who participate in sports, **sector school name** which is the type of school for example, public, 4-year or above, **gender ratio** which is the male population count divided by female population count, **total count of students** which is the total amount of students enrolled in the college, women expenditure percent which is the percentage of expenditures that goes towards woman's sports (which we calculated by taking a school's expenditures of women and dividing it by the school's total expenditures), **school classification name** which is a school\'s sports classification for example NCAA Division I-FCS. Then, our response variable is **profit** which is calculated by the total revenue of the school for basketball minus the expenditures of the school for basketball.

For data cleaning, we need to filter for observations that have `Basketball` as the sports and filter out missing values. we also need to filter out observations that have missing values of the predictor variables we need in the model. We need to create new variables by mutation to turn variables involving revenue and expenditure into the unit of millions and total student count into the unit of hundreds, so that we can get larger coefficients for better modeling. We create the new response variable of profit and new predictor variable of woman's expenditure percent.

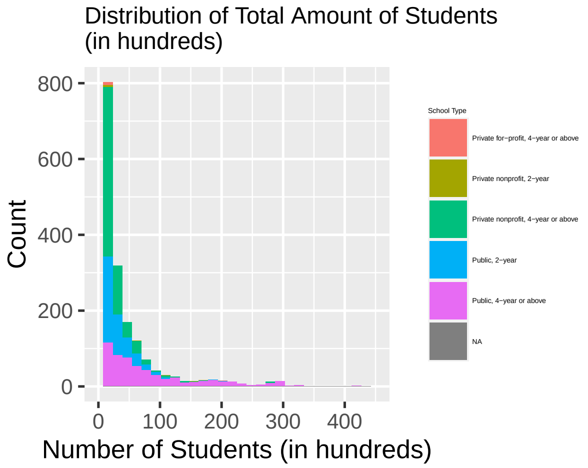
Distribution of the response variable: total profit generated by college basketball

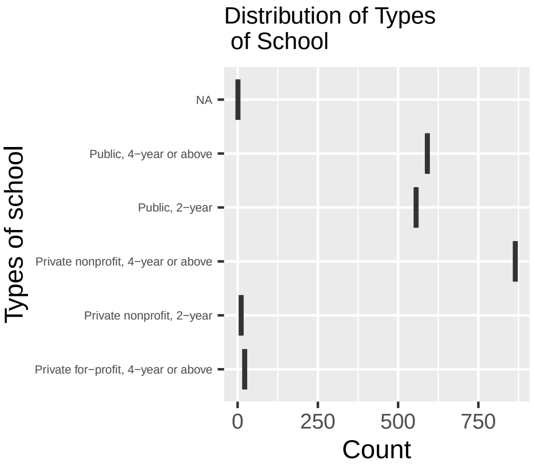


```
# A tibble: 1 x 8
  n_missing numeric.mean numeric.sd numeric.p0 numeric.p25 numeric.p50
      <int>        <dbl>      <dbl>      <dbl>       <dbl>       <dbl>
1        57       0.0274      0.919      -9.67           0           0
  numeric.p75 numeric.p100
        <dbl>        <dbl>
1     0.00417         16.9
```

The distribution of profit is approximately normally distributed with a slight right skew. However, many observations centered around 0$ in profit. The median of the data is at 0$, and the mean is about 0.027 million USD. Since the response is approximately normal, we can use the mean as the center of the data. The range is from -9.6738 million USD to 16.8938 million USD. We also split up the distribution into positive and negative profits and took the log of the x axis so better see a more detailed distribution. It seems that positive profits is normally distributed while negative profits are left skewed. There are 57 missing values of profit.

Distributions of total amount of students (potential quantitative predictor variable) and type of school (potential categorical predictor variable)





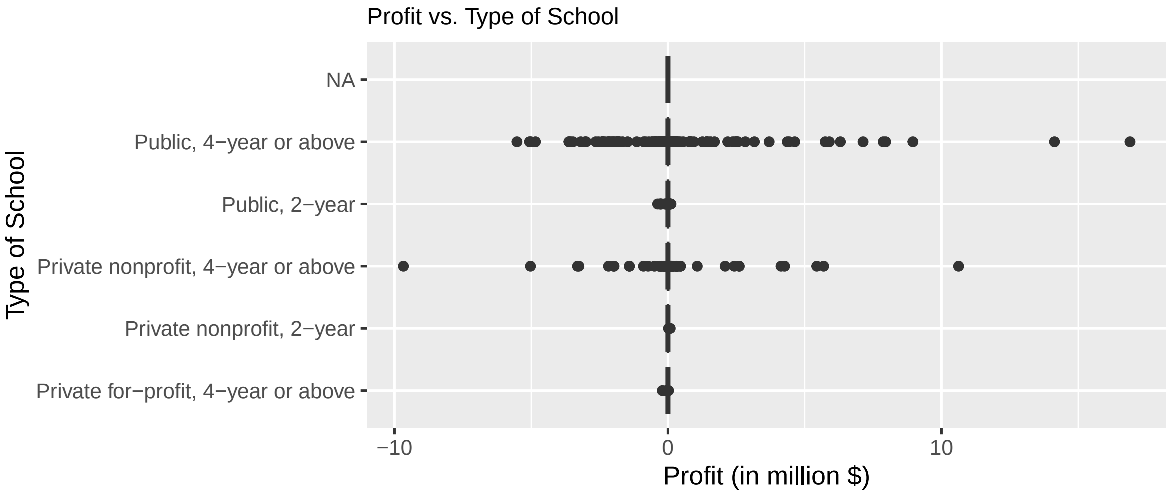
```
# A tibble: 1 x 8
  n_missing numeric.mean numeric.sd numeric.p0 numeric.p25 numeric.p50
      <int>        <dbl>      <dbl>      <dbl>       <dbl>       <dbl>
1         0         41.7       59.5        0.5        10.6        20.1
  numeric.p75 numeric.p100
        <dbl>        <dbl>
1        46.3          663.
```

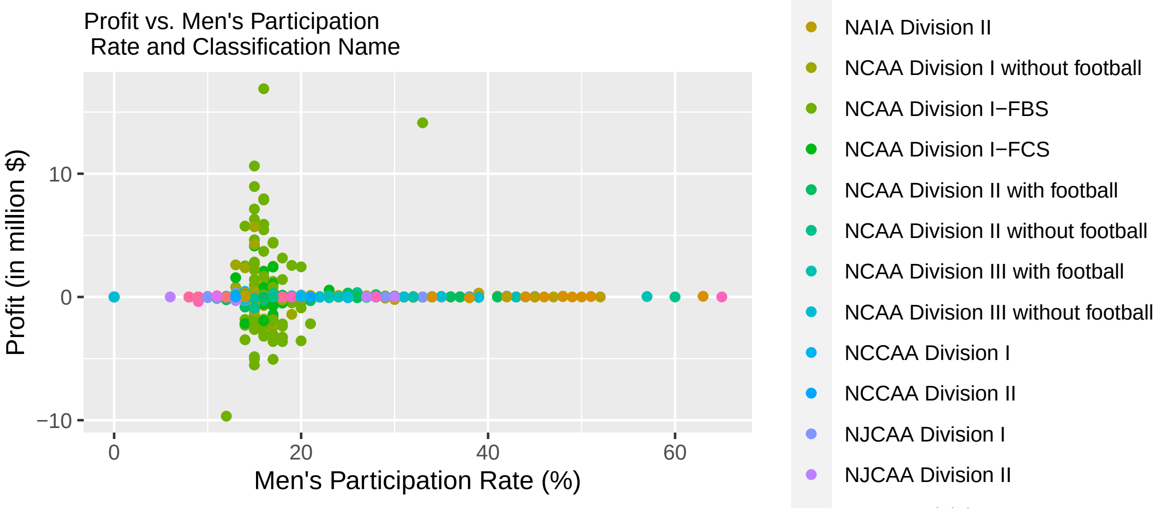
```
                                      n           p nmiss
Private for-profit, 4-year or above  22 0.010757946     1
Private nonprofit, 2-year            11 0.005378973    NA
Private nonprofit, 4-year or above  865 0.422982885    NA
Public, 2-year                      556 0.271882641    NA
Public, 4-year or above             591 0.288997555    NA
```

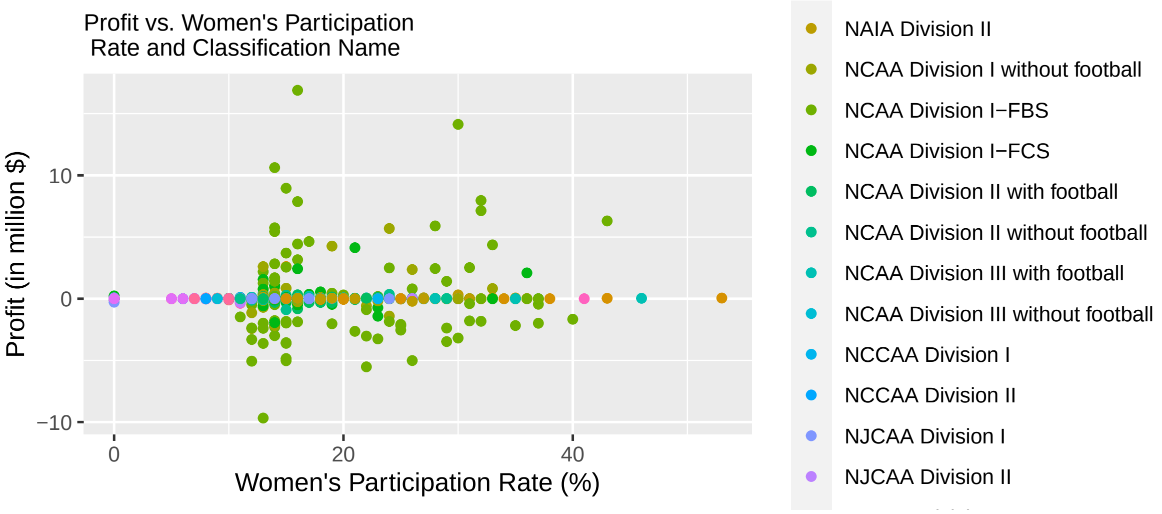
The distribution of total amount of students on college sports is right-skewed with most of the amount of student values in the lower range, while a number of observations have very high values that make them outliers. Given the apparent skewness, the center is the median of 2,005.5 students. Since the distribution is skewed, the IQR is used as a more reliable measure of spread which is Q3 - Q1 = 4,633 - 1,060.25 = 3,572.75 students. You can also see that most of the schools with very high number of students are public, 4 year or above colleges.

There are 5 types of school. Private for-profit, 2-year and private nonprofit, 2-year have low numbers of observations. Private nonprofit, 4-year or above, public, 2 year, and public, 4-year or above have comparably higher number of observations than the other two, with private nonprofit, 2-year having the highest number of observations. There is 1 missing value of type of school.

Relationship between Profit and a Categorical and Quantitative Predictor



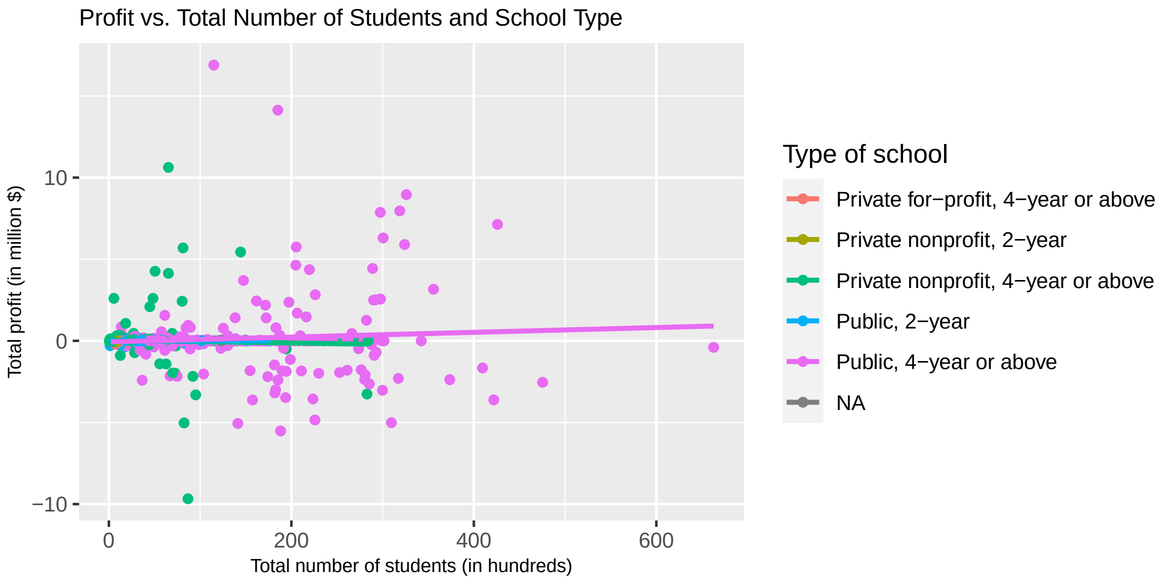




Looking at the relationships above, it seems that the medians of each type of school's profit is around $0. We can also see that Public, 4 year or above colleges have the highest recorded profit in USD, and that Private nonprofit, 4 year or above colleges have the lowest profit in USD.

Looking at the participation rate graphs, there seems to be slightly more participation of women than of men, as many of the data points are more spread out to the right of the graph for women while there is a conglomerate of data points centered around 20% for men's participation rate. There seems to be a very weak almost negative linear relationship between participation rate versus profit, as most of the data points are centered around $0 profit regardless of participation rate. However, looking at the mens participation rate specifically, it seems schools classified with sports as NCAA Division I-FBS have the highest profits and highest range of profits. Schools classified as NAIA Division II and NCAA Division II without football seem to have high mens participation rates. For womens participation rate, it shows the same thing as mens, however there tends to be greater participation rate of women than men in NCAA Division I-FBS schools.

A potential interaction effect between total number of student and type of school



The lines are not parallel indicating there is an interaction effect. The slope of total number of student differs based on the type of school.

Methodology

We plan to use a multiple linear regression to predict the profit (revenue - expenditure) in USD of the collegiate sport basketball using participation rate, school sector name, gender ratio, women expendtiture percent, total count of students, and school classification name. For the predictor gender ratio, we will mutate the data and divide the total male student count by total female student count. For the predictor women expenditure percent, we will mutate the data and divide a school's woman's expenditure by total expenditure.

For the response variable, we decided to predict the profit, subtracting the expenditure variable from the revenue variable, instead of predicting the revenue variable and using the expenditure variable as a predictor because we anticipate that the revenue variable and the expenditure variable will have strong correlations. Expenditure will be strongly correlated with the revenue of the sport basketball because schools often allocate significant financial resources to their sports programs, scholarships, and marketing. As these investments increase, the expectation is that they will have a direct impact on the overall revenue generated through different resources such as ticket sales.

For this reason, we decided that using expenditure to predict the revenue will not produce a meaningful result in choosing the best predictors as expenditure may already significantly affect the revenue, muting the effects of all the other predictors.

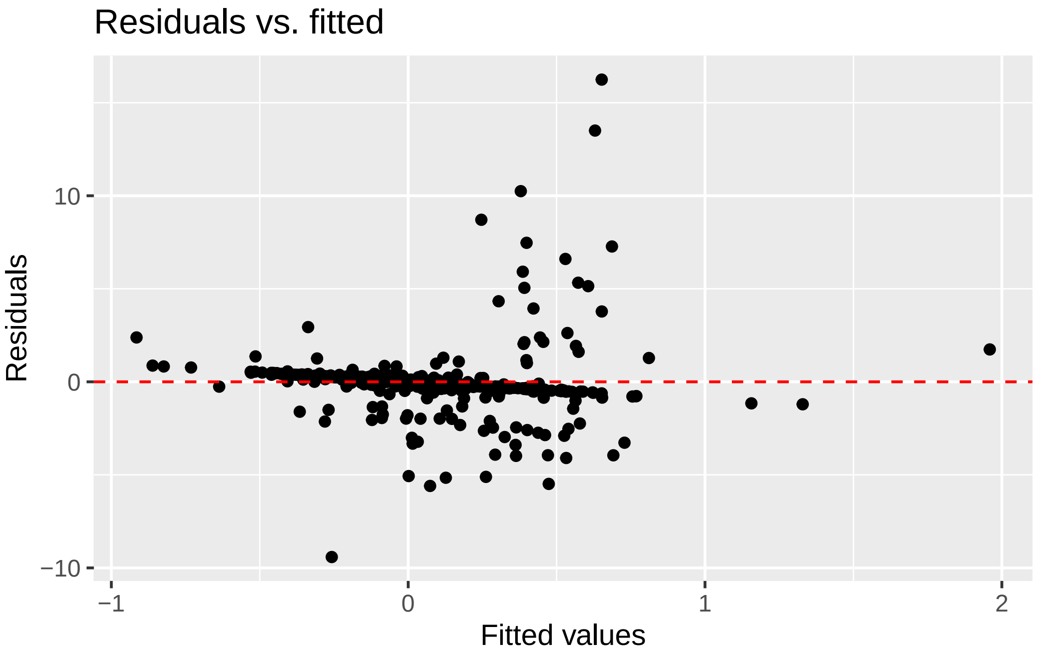
For the predictors, we anticipate that **participation rate** will be a key predictor of the total profit of the sport basketball because schools may allocate more money to this sport if there is more participation from the students. We also expect that the type of school **(sector name)** will be a strong predictor of profit because different types of schools have varying levels of resources, alumni support, and participation rates in sports depending on school size and program. **Gender ratio** is another predictor to consider because sports could be hugely dependent on the demographics of players.

We also anticipate interaction effects which we will do further analysis in our results. Specifically, we believe the participation rate and the school sector type will have a correlation. This is because we expect the participation rate to be very dependent on the levels of investment, competitive levels, institutional culture, and student demographics which vary based on the type of institution(sectorname).

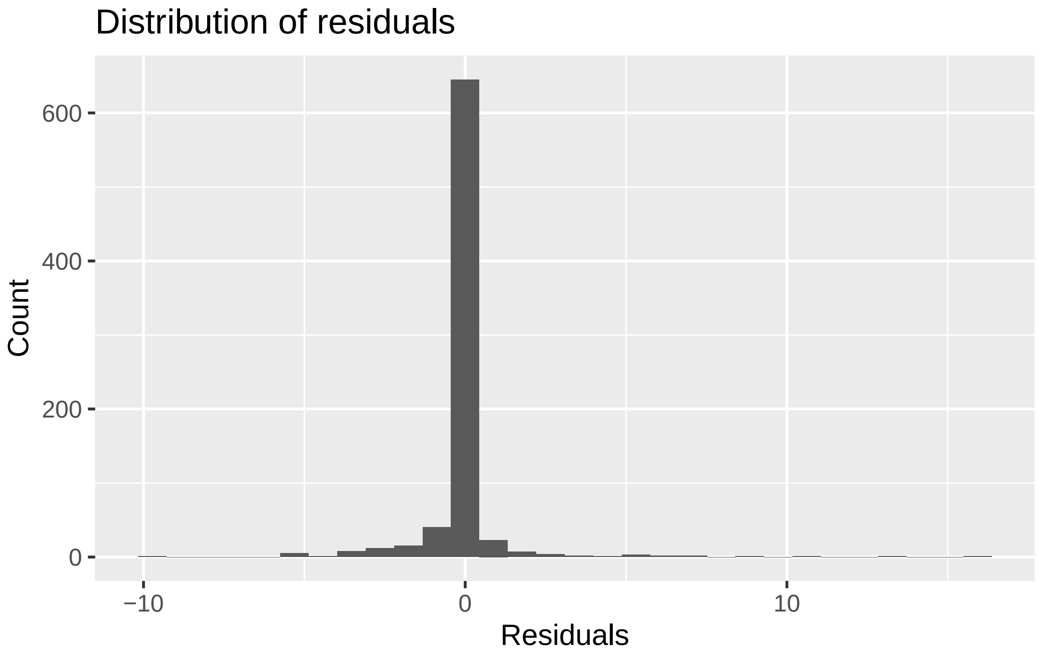
To prepare the variables for our analysis, we plan to use a recipe. First, we plan to drop all the NA values using `step_naomit()`. We decided to drop NA values to ensure that the values we do use are accurate. For example, the data set has observations with NA values in women's participation rate. However, when the sum participation rate for women is calculated, the data set represents this sum with 0. This 0 can skew the regression model when the true participation rate of women may not be 0. Since we are using total participation rate in our model, it is best to drop NA values. Also, considering that the data set has 775 observations after all NA's are dropped, we still have a lot of reliable data to use for modeling.

Then, we plan to use `step_zv()` to remove all predictors that contain only a single value. We plan to create dummy variables using `step_dummy()` for all nominal predictors which are Classification name and Sector name. Next, we will be using `step_center()` to mean center our quantitative predictors. Lastly, we plan to use `step_zv()` to remove all predictors that contain only a single value.

Checking Model Conditions:



Constant variance is met because there is an even amount of residuals above and below the horizontal line. Linearity may not be met. We can see that there is a conglomerate of points that shape a almost negative line around the y intercept. However, there is a random scatter of points around this. This should be kept in mind and show that a linear model may not be the best regression model to use.



We can see the normality condition is satisfied as the residuals are normally distributed and the sample has more than 30 observations. Independence is also satisfied because of how the data was taken. Collegiate information about one school is independent of another school.

Results

```
# A tibble: 2 x 6
  .metric .estimator  mean     n std_err .config
  <chr>   <chr>      <dbl> <int>   <dbl> <chr>
1 rmse    standard   1.25     20  0.180  Preprocessor1_Model1
2 rsq     standard   0.128    20  0.0283 Preprocessor1_Model1
```

```
# A tibble: 1 x 3
  mean_adj_rsq mean_aic mean_bic
         <dbl>    <dbl>    <dbl>
1       0.0169    1994.    2058.
```

```
# A tibble: 2 x 6
  .metric .estimator  mean     n std_err .config
  <chr>   <chr>      <dbl> <int>   <dbl> <chr>
1 rmse    standard   1.25     20  0.185  Preprocessor1_Model1
2 rsq     standard   0.232    20  0.0644 Preprocessor1_Model1
```

```
# A tibble: 1 x 3
  mean_adj_rsq mean_aic mean_bic
         <dbl>    <dbl>    <dbl>
1      0.00609    1995.    2038.
```

Our final model uses the predictor of total participation rate, total count of students, gender ratio, and profit by gender ratio.

term	estimate	std.error	statistic	p.value
(Intercept)	0.070	0.061	1.151	0.250
total_partc	0.013	0.011	1.180	0.239
ef_total_count	0.002	0.001	2.053	0.041
gender_r	-0.011	0.144	-0.076	0.940
profit_r	-0.002	0.001	-1.615	0.107

- Model fit statistics

```
# A tibble: 2 x 6
  .metric .estimator  mean     n std_err .config
  <chr>   <chr>      <dbl> <int>   <dbl> <chr>
1 rmse    standard   1.25     20  0.185  Preprocessor1_Model1
2 rsq     standard   0.213    20  0.0552 Preprocessor1_Model1
```

```
# A tibble: 1 x 3
  mean_adj_rsq mean_aic mean_bic
         <dbl>    <dbl>    <dbl>
1      0.00912    1990.    2016.
```

The final model has a AIC of 1989.851 and a BIC of 2015.731.

Our initial model including total participation rate, school sector name, gender ratio, total student count, classification name, and profit by gender ratio has a AIC of 2003.095 and a BIC of 2062.62. When we reduce the sector name variable, the AIC decreases to 1995.288 and the BIC decreases to 2037.56. We choose our final model that further reduce the classification name variable since it has the lowest AIC and BIC. We decide not to further reduce the terms in the model to have sufficient predictor variables for the response variable.

- Model diagnostics

```
   total_partc ef_total_count       gender_r       profit_r 
      1.024006       1.015160       1.015154       1.009525 
```

There is no variable with $VIF > 10$ that indicates concerning multicollinearity, so no apparent issue with multicollinearity is found.

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        1.48
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        1.39
```

The RMSE for the training data which is 1.454625 while the RMSE for testing data is 1.335227. Since we have a reasonably small difference in the RMSE, there is no apparent sign of model overfit. The model can generalize to new data.

Model Interpretations and Conclusions

Significant predictor is total count of students with a p-value smaller than 0.05. For each hundred increase in the total count of students, the school's basketball profit is expected to increase by 0.002 million dollars. Furthermore, when comparing the VIC and BIC of multiple models to predict profit, we concluded that the best model to fit the data was the model that includes total participation rate, total student count, gender ratio, and profit ratio.