Public school education in Uttar Pradesh

Sannat Mengi

Motivation

Our country is still struggling with public school education system. Public school systems are the most fundamental sources of education for many low income communities and people who live in remote areas. Not to mention, economic aims for a country are dependent on education systems and in a country where around 70% of population is rural, public schools are the only resort for millions of people to gain edcuation. However, as we all know public education systems are not the best producers of knowledge or even uplifters of low-income class. Governments always boast about their education funding and newer learning aids available to public schools yet this sector is not growing since last decade or two. In this assignment, I will try to analyse what is it that affects education the most ? Is it teachers, infrastructure, hygiene, enrollment, community representation etc. For the analysis, I have chosen to study Uttar Pradesh public schools and look for factors that really affect education. Also, this analysis includes region wise (urban/rural) effects and representation of girls in schools.

Model

Education is not as linear as we imagine it to be. Quanitfying education is a tough task and cannot be done by some simple metric. However, for our analysis we will choose results as a metric which tells us about the impact of a particular school. There are a lot of other parameters which are used to describe education (Learning outcomes, attendance etc.) but I will use results as a target variable here as it is easily quanitifiable. The features we will be using with results would be-

- **Student teacher ratio (STR)**- measures effect of teachers on students
- **Hygiene (hyg)** - A ratio of total bathrooms in school to total rooms in school. It highlights funding as well.
- **Percentage of girls (prcntgrl)**- Measures respresentation of girls in school.
- **Urban (urban)**- a dummy variable with value 1 for urban schools and value 0 for rural schools.

A linear model would be used for the modelling education in UP. Our aim would be to find parameters from OLS method (Ordinary least squares). Model can be written as-

$$result = \beta_0 + \beta_1 STRnorm + \beta_{11} STR^2 norm + \beta_2 hyg + \beta_3 prcntgrl + \delta_0 urban + u$$

u signifies error term for our regression model and it captures all the variables apart from feature variables listed in model.For e.g. **u** can signify the affect of society, economic status of parents, school teacher's credibility etc.

Linear regression model is computed using R.

Data

Data for districts is compiled from DISE (District Information System of Education) under National University of Education and Planning (NUEPA). Preliminary data merging, manipulation and cleaning is done on MS Excel.

Importing Data

```
data <- read.csv('/Users/sannatmengi/Desktop/EconData/EdData/FinalUP.csv',header=TRUE)
head(data)
```

```
##        SCHCD TotalStudents Total_boys Percntage_girls Total_Pass
## 1 9150202502            29         12       0.5862069         11
## 2 9151704701            40         16       0.6000000          0
## 3 9150206905            76         32       0.5789474         16
## 4 9151104501            74         35       0.5270270         20
## 5 9151717115           152         82       0.4605263          0
## 6 9151006902            78         38       0.5128205         90
##   Result_Percentage Tot_Teachers  Hyg_infra Urban North South East West Central
## 1         0.3793103             4 0.08695652     0     0     0    0    1       0
## 2         0.0000000             5 0.08000000     0     0     0    0    1       0
## 3         0.2105263             3 0.12500000     1     0     0    0    1       0
## 4         0.2702703             6 0.09090909     0     0     0    0    1       0
## 5         0.0000000             4 0.09523809     0     0     0    0    1       0
## 6         1.1538462             6 0.11764706     1     0     0    0    1       0
```

Generating feature columns of interest from data

```
#Generate a new column STR- Student teacher ratio
data$STR = data$TotalStudents/data$Tot_Teachers
#View a few rows of data
head(data)
```

```
##        SCHCD TotalStudents Total_boys Percntage_girls Total_Pass
## 1 9150202502            29         12       0.5862069         11
## 2 9151704701            40         16       0.6000000          0
## 3 9150206905            76         32       0.5789474         16
## 4 9151104501            74         35       0.5270270         20
## 5 9151717115           152         82       0.4605263          0
## 6 9151006902            78         38       0.5128205         90
##   Result_Percentage Tot_Teachers  Hyg_infra Urban North South East West Central
## 1         0.3793103             4 0.08695652     0     0     0    0    1       0
## 2         0.0000000             5 0.08000000     0     0     0    0    1       0
## 3         0.2105263             3 0.12500000     1     0     0    0    1       0
## 4         0.2702703             6 0.09090909     0     0     0    0    1       0
## 5         0.0000000             4 0.09523809     0     0     0    0    1       0
## 6         1.1538462             6 0.11764706     1     0     0    0    1       0
##        STR
## 1  7.25000
## 2  8.00000
## 3 25.33333
## 4 12.33333
## 5 38.00000
## 6 13.00000
```

```
#Dimensions of dataframe
dim(data)
```

```
## [1] 306  15
```

After observing, there is faulty data too in our file. Such as many rows have Result_percentage=0. However, this might be faulty data or can be truth as well because it is highly unlikely for a school to have 0% as it's result. So for uniformity of our analysis, we will drop the rows with Result_Percentage=0.

```r
# remove the rows with 0% result
data_clean <- data[!(data$Result_Percentage==0),]
#remove rows with result percentage >1
data_clean <- data_clean[!(data_clean$Result_Percentage>1),]
#remove rows with hyg_infra > 1
data_clean <- data_clean[!(data_clean$Hyg_infra >1),]
# view a few rows of cleaned data
head(data_clean)
```

```
##         SCHCD TotalStudents Total_boys Percntage_girls Total_Pass
## 1  9150202502            29         12       0.5862069         11
## 3  9150206905            76         32       0.5789474         16
## 4  9151104501            74         35       0.5270270         20
## 11 9151006006            77         47       0.3896104         36
## 13 9151509701           114         57       0.5000000         26
## 16 9151501801            79         30       0.6202532         18
##    Result_Percentage Tot_Teachers  Hyg_infra Urban North South East West
## 1          0.3793103            4 0.08695652     0     0     0    0    1
## 3          0.2105263            3 0.12500000     1     0     0    0    1
## 4          0.2702703            6 0.09090909     0     0     0    0    1
## 11         0.4675325           12 0.10000000     0     0     0    0    1
## 13         0.2280702            2 0.11764706     0     0     0    0    1
## 16         0.2278481            2 0.12500000     0     0     0    0    1
##    Central       STR
## 1        0  7.250000
## 3        0 25.333333
## 4        0 12.333333
## 11       0  6.416667
## 13       0 57.000000
## 16       0 39.500000
```

```r
# view new dimension of data frame
dim(data_clean)
```

```
## [1] 219  15
```

No. of rows get reduced from 306 to 219 after removing spurious rows.

Exploratory Data Analysis

```
##      SCHCD           TotalStudents      Total_boys      Percntage_girls
##  Min.   :9.070e+09   Min.   :  13.0   Min.   :   0.00   Min.   :0.0000
##  1st Qu.:9.360e+09   1st Qu.:  63.5   1st Qu.:  30.50   1st Qu.:0.4319
##  Median :9.360e+09   Median : 101.0   Median :  51.00   Median :0.4878
##  Mean   :9.438e+09   Mean   : 163.0   Mean   :  87.39   Mean   :0.4907
##  3rd Qu.:9.671e+09   3rd Qu.: 154.5   3rd Qu.:  81.50   3rd Qu.:0.5489
##  Max.   :9.671e+09   Max.   :2884.0   Max.   :2004.00   Max.   :1.0000
##    Total_Pass     Result_Percentage  Tot_Teachers      Hyg_infra
##  Min.   :  2.00   Min.   :0.03343   Min.   : 1.000   Min.   :0.00000
##  1st Qu.: 19.50   1st Qu.:0.23217   1st Qu.: 3.500   1st Qu.:0.09524
##  Median : 36.00   Median :0.34483   Median : 5.000   Median :0.12500
##  Mean   : 52.25   Mean   :0.37803   Mean   : 6.484   Mean   :0.13633
##  3rd Qu.: 58.50   3rd Qu.:0.48422   3rd Qu.: 8.000   3rd Qu.:0.16905
##  Max.   :599.00   Max.   :0.96154   Max.   :32.000   Max.   :0.43478
```

```
##      Urban            North       South            East            West
##  Min.   :0.0000   Min.   :0   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.0000   Median :0   Median :0.0000   Median :0.000   Median :0.0000
##  Mean   :0.1918   Mean   :0   Mean   :0.4658   Mean   :0.379   Mean   :0.1553
##  3rd Qu.:0.0000   3rd Qu.:0   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :0   Max.   :1.0000   Max.   :1.000   Max.   :1.0000
##     Central       STR
##  Min.   :0   Min.   :  1.533
##  1st Qu.:0   1st Qu.: 11.625
##  Median :0   Median : 18.500
##  Mean   :0   Mean   : 35.282
##  3rd Qu.:0   3rd Qu.: 33.438
##  Max.   :0   Max.   :547.000
```

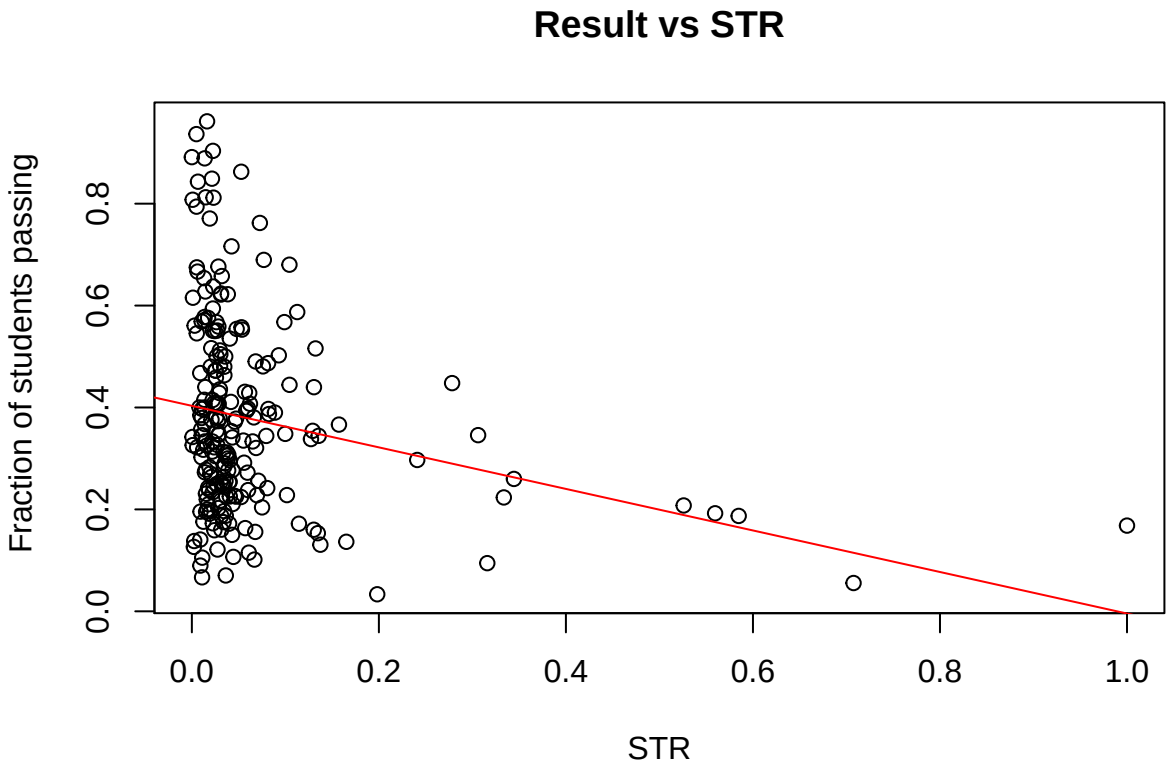
Variable STR needs to be normalized to be in scale with other variables. Also, variable Result_precentage cannot be grater than 1.Result percentage greater than 100% must be due to error in data entry, hence it needs to be removed as well.

```
## function for normalizing data columns
normalize <- function(x) {
return ((x - min(x)) / (max(x) - min(x)))
}
#normalizing STR
data_clean$STR_norm <- normalize(data_clean$STR)
data_clean$STR_norm_sqr <- (data_clean$STR_norm)^2
head(data_clean)
```

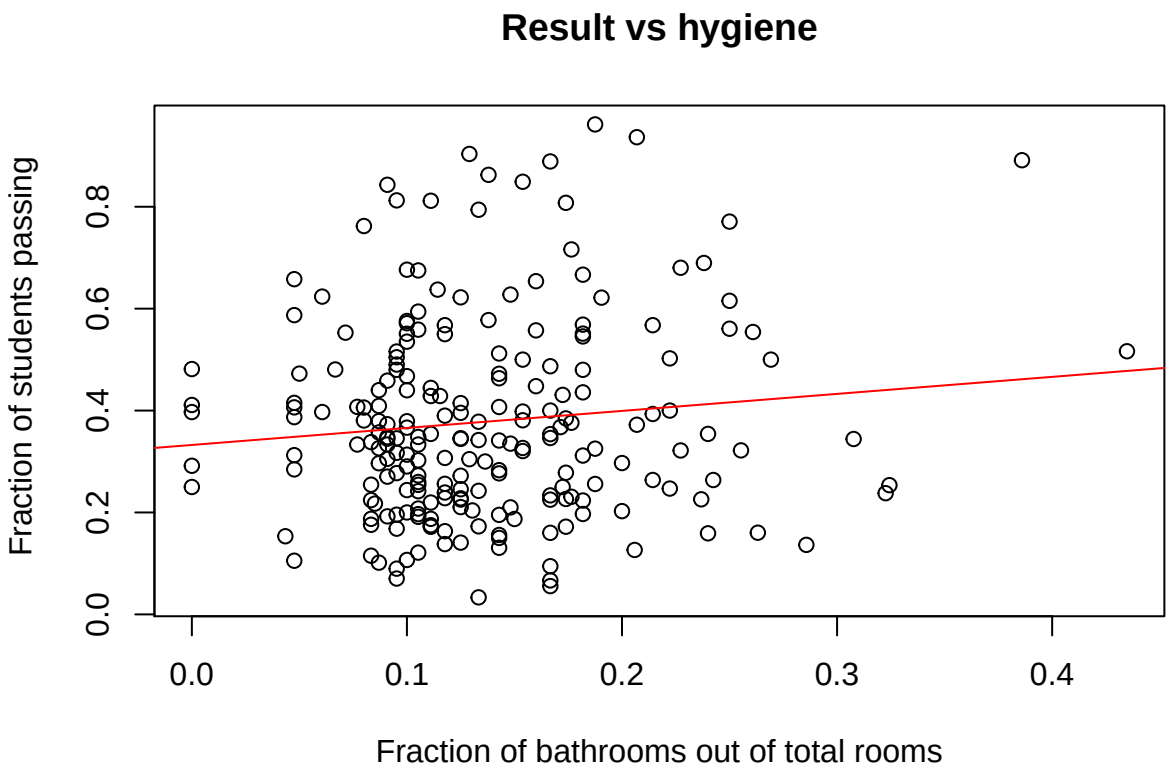
```
##         SCHCD TotalStudents Total_boys Percntage_girls Total_Pass
## 1  9150202502            29         12       0.5862069         11
## 3  9150206905            76         32       0.5789474         16
## 4  9151104501            74         35       0.5270270         20
## 11 9151006006            77         47       0.3896104         36
## 13 9151509701           114         57       0.5000000         26
## 16 9151501801            79         30       0.6202532         18
##    Result_Percentage Tot_Teachers  Hyg_infra Urban North South East West
## 1          0.3793103             4 0.08695652     0     0     0    0    1
## 3          0.2105263             3 0.12500000     1     0     0    0    1
## 4          0.2702703             6 0.09090909     0     0     0    0    1
## 11         0.4675325            12 0.10000000     0     0     0    0    1
## 13         0.2280702             2 0.11764706     0     0     0    0    1
## 16         0.2278481             2 0.12500000     0     0     0    0    1
##    Central       STR    STR_norm STR_norm_sqr
## 1        0  7.250000 0.010480323 1.098372e-04
## 3        0 25.333333 0.043632364 1.903783e-03
## 4        0 12.333333 0.019799560 3.920226e-04
## 11       0  6.416667 0.008952579 8.014867e-05
## 13       0 57.000000 0.101686629 1.034017e-02
## 16       0 39.500000 0.069604009 4.844718e-03
```

Finally, the dimension of data frame after more cleaning gets reduced to 230

```
#visualising result vs str
plot(data_clean$STR_norm,data_clean$Result_Percentage,xlab="STR",ylab=" Fraction of students passing", m
abline(lm(data_clean$Result_Percentage~data_clean$STR_norm), col='red')
```

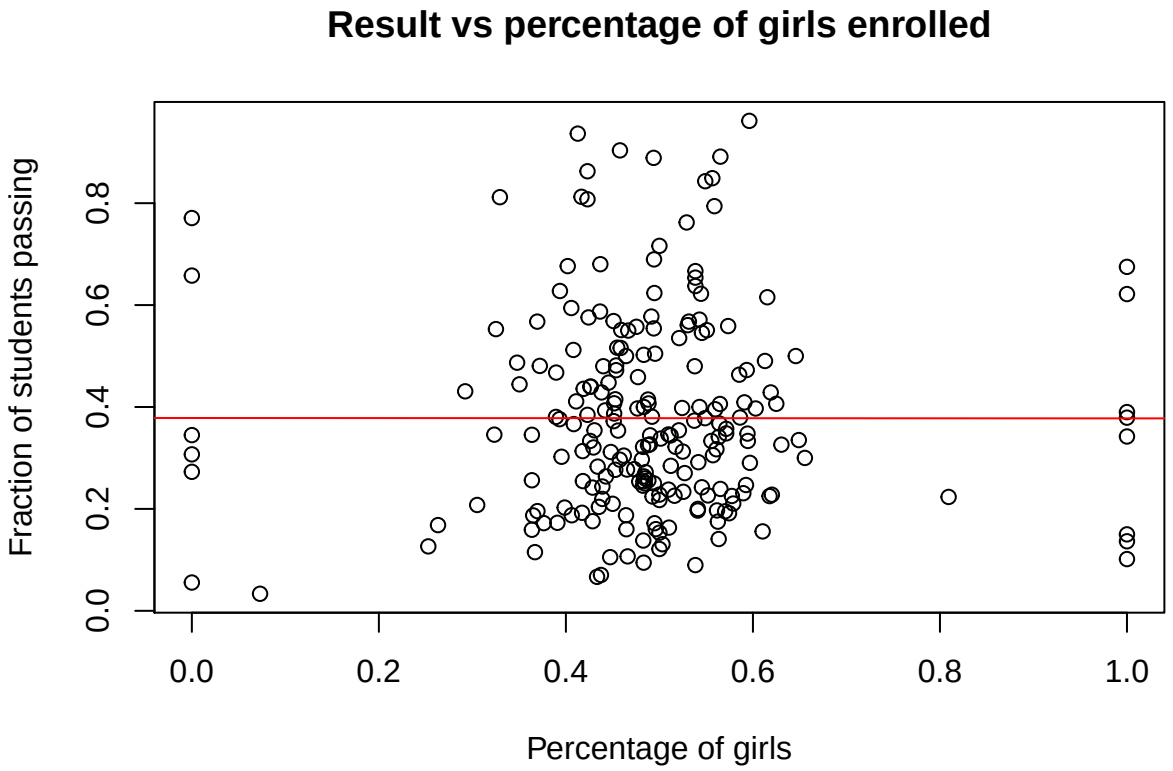



```
#visualising result with hygiene
plot(data_clean$Hyg_infra,data_clean$Result_Percentage,xlab="Fraction of bathrooms out of total rooms",
abline(lm(data_clean$Result_Percentage~data_clean$Hyg_infra), col='red')
```




```
plot(data_clean$Percntage_girls, data_clean$Result_Percentage, xlab="Percentage of girls", ylab="Fracti
```

```
abline(lm(data_clean$Result_Percentage~data_clean$Percntage_girls),col='red')
```



From preliminary data analysis, following insights can be inferred-

- Results increase with increase in hygiene facilities
- Results decrease with increase in Student teacher ratio
- No particular comment can be inferred from Result vs girl enrollment curve. The line is having a very less slope or maybe insignificant.

Results:Regressing the model

Multiple variable linear regression of model-

```
# renaming variables for ease of reference
y <- data_clean$Result_Percentage
x1 <- data_clean$STR_norm
x2 <- data_clean$STR_norm_sqr
x3<- data_clean$Hyg_infra
x4 <- data_clean$Percntage_girls
x5 <- data_clean$Urban

reg <- lm(y ~ x1+x2+x3+x4+x5)

#Regression results
print(reg)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5)
##
```

```
## Coefficients:
## (Intercept)           x1           x2           x3           x4           x5
##     0.40291     -0.77113      0.53484      0.24032     -0.06387      0.06617
```

NA values for North and south indicate that there were very few dummy variables/data for northern and southern region and reression library didn't compute the parameters attached to it.

```
summary(reg)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37430 -0.14260 -0.02210  0.09378  0.56404
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.40291    0.05269   7.647 6.98e-13 ***
## x1          -0.77113    0.28358  -2.719  0.00708 **
## x2           0.53484    0.37779   1.416  0.15832
## x3           0.24032    0.20171   1.191  0.23482
## x4          -0.06387    0.08544  -0.748  0.45553
## x5           0.06617    0.03286   2.014  0.04526 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1866 on 213 degrees of freedom
## Multiple R-squared:  0.09509,    Adjusted R-squared:  0.07385
## F-statistic: 4.477 on 5 and 213 DF,  p-value: 0.0006712
```

Inferences

Following can be observed form summary of regression-

- Only STR(Student teacher ratio), intercept and urban/rural dummy variable are statistically significant variable. i.e t-test statistic > critical value at 5% significance level.
- Other variables are not that statistically significant for result prediction.
- Coefficient of STR indicates that an increase of 0.1 units in normalized STR leads to reduction of 0.077 units in fraction of students passing. This coefficient is quite huge and highlights the importance of a healthy student teacher ratio.
- STR^2, hyg, prcntgirl are not statistically significant variables and can be removed from model as well.
- Coefficient of **urban** shows that there is a difference of 0.066 units in fraction of students passing from ubran and rural schools. While this is quite intuitive, but the difference is not that large. This also indicates that the society influence on education is not very different in Uttar Pradesh.
- We can use the model to predict any fraction of students passing from a school

$$result = 0.402 - 0.77 STRnorm + 0.066 urban + error$$

Improvements

- OLS estimates are unbiased and consistent only when independent and identically distributed random samples are taken. The data chosen is truly random or biased is not clear and depends upon the source as well. Sample data extracted can be randomised further. Features maybe highly correlated as well.
- Functional form of linear model can be imporved using interaction terms, log values, polynomial expressions.
- The fit of data is very poor. Adjusted R-squared for model is just 7.3%. Some more investigation into model complexity needs to be done.
- Other feature variables such as avaialbility of extra teaching aids, extra-curricular activities, economic status (funding) of school can also be explored to study effects of these variables as well.