NewGlobe Case Study - Analyst M&E/Data Analytics Teams

Kamran Ahmed

2023-10-18

# The Data

You have received four files, all in .dta and .xlsx formats, so you can use whichever format you prefer. These files are the following:

● “Lesson completion”: file provided at the teacher level, meaning that there is a unique row for each teacher. The file contains the grade that each teacher teaches, and the average lesson completion rate over the term of interest.

● “Pupil attendance”: file provided at the pupil level (that means that there is a unique row for each pupil). This file includes the unique school ID, unique pupil ID, the pupil’s grade, the attendance records, and the present records.

○ The attendance records means the total number of times that a pupil’s teacher took attendance.

○ The present records means the total number of times that a pupil was present, out of the attendance

● “Pupil scores”: file provided at the pupil\*subject level (that means that there are more than one row per pupil). This file includes the unique school ID, unique pupil ID, the pupil’s grade, the subject for this assessment, and the score obtained in this assessment.

● “School information”: file provided at the school-level. It includes the region and province where each school is located, the unique school ID, and the “treatment status” (yes/no) for a given tutoring program.

# Load libraries

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(readxl)

# Load Data

Let’s load the data and have a quick look of it to see how the data look like

Lesson\_completion <- read\_excel("Lesson completion.xlsx")  
head(Lesson\_completion)

## # A tibble: 6 × 4  
## school\_id teacher\_id grade lesson\_completion\_rate  
## <dbl> <dbl> <chr> <dbl>  
## 1 416 505 Grade 1 0.568  
## 2 416 202 Grade 2 0.681  
## 3 416 124 Grade 3 0.250  
## 4 416 516 Grade 4 0.359  
## 5 416 145 Grade 5 0.397  
## 6 792 545 Grade 1 0.809

Pupil\_attendance <- read\_excel("Pupil attendance.xlsx")  
head(Pupil\_attendance)

## # A tibble: 6 × 5  
## school\_id pupil\_id grade attendance\_records present\_records  
## <dbl> <dbl> <chr> <dbl> <dbl>  
## 1 35175 1 Grade 1 91 69  
## 2 40580 7 Grade 2 92 86  
## 3 9342 8 Grade 5 43 39  
## 4 858450 10 Grade 5 86 62  
## 5 792 13 Grade 3 104 81  
## 6 324884 14 Grade 4 90 67

Pupil\_scores <- read\_excel("Pupil scores.xlsx")  
head(Pupil\_scores)

## # A tibble: 6 × 5  
## school\_id pupil\_id grade subject score  
## <dbl> <dbl> <chr> <chr> <dbl>  
## 1 35175 1 Grade 1 Fluency 65   
## 2 35175 1 Grade 1 Kiswahili 0.943  
## 3 35175 1 Grade 1 Math 1   
## 4 40580 7 Grade 2 Math 0.933  
## 5 40580 7 Grade 2 Kiswahili 0.943  
## 6 40580 7 Grade 2 Fluency 117

School\_information <- read\_excel("School\_information.xlsx")  
head(School\_information)

## # A tibble: 6 × 4  
## region province school\_id tutoring\_program  
## <chr> <chr> <dbl> <chr>   
## 1 Mombasa Coast 136992 No   
## 2 Kilifi Coast 687400 Yes   
## 3 Mombasa Coast 609982 Yes   
## 4 Eastern Eastern 223941 No   
## 5 Isiolo Eastern 34092 No   
## 6 Isiolo Eastern 46684 No

# Step 1: Data cleaning

Please create a file at the student-level which has information about their test scores, school information, their attendance, and their teacher’s lesson completion rate. Note that this is the main data set that we expect you to share with us.

Hint: note that the four data sets you will use are all presented at different “levels” of the data (e.g., “School information” is at the level of the school, but “Pupil scores” is at the level of the student). Therefore, we suggest that you start by reshaping the “Pupil scores” file so that each student only has one row in the data, with different columns for their scores in math, fluency, and Kiswahili. Use this as your “base file”, and start merging all the other files to this. Be careful with how you merge things: since there are many students to a school or even a teacher, some of these merges will need to be “many-to-one” (but not all).

# Create a base file by reshaping the “Pupil scores” file so that each student only has one row in the data, with different columns for their scores in math, fluency, and Kiswahili  
Pupil\_scores <- Pupil\_scores %>%  
 pivot\_wider(names\_from = subject, values\_from = score)  
head(Pupil\_scores)

## # A tibble: 6 × 6  
## school\_id pupil\_id grade Fluency Kiswahili Math  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl>  
## 1 35175 1 Grade 1 65 0.943 1   
## 2 40580 7 Grade 2 117 0.943 0.933  
## 3 9342 8 Grade 5 144 0.850 0.700  
## 4 858450 10 Grade 5 211 1 0.720  
## 5 792 13 Grade 3 221 0.857 0.967  
## 6 324884 14 Grade 4 267 0.921 0.900

# Merge Pupil\_attendance file to the base file i.e, Pupil\_scores file  
pupil\_df <- merge(Pupil\_scores, Pupil\_attendance, by = c("pupil\_id", "school\_id", "grade"))  
head(pupil\_df)

## pupil\_id school\_id grade Fluency Kiswahili Math attendance\_records  
## 1 1 35175 Grade 1 65 0.9428571 1.0000000 91  
## 2 10 858450 Grade 5 211 1.0000000 0.7200000 86  
## 3 100 32940 Grade 2 170 0.7142857 0.7333333 61  
## 4 10000 49404 Grade 1 7 0.6285715 0.7000000 93  
## 5 10002 223941 Grade 1 0 0.6571429 0.8333333 92  
## 6 10005 822894 Grade 2 137 0.7739512 0.6362270 92  
## present\_records  
## 1 69  
## 2 62  
## 3 49  
## 4 46  
## 5 44  
## 6 80

# Merge teacher's data  
pupil\_teacher\_df <- merge(pupil\_df, Lesson\_completion, by = c("school\_id", "grade"))  
head(pupil\_teacher\_df)

## school\_id grade pupil\_id Fluency Kiswahili Math attendance\_records  
## 1 108210 Grade 1 6430 41 0.4000000 0.9 89  
## 2 108210 Grade 1 10987 32 0.5428572 1.0 83  
## 3 108210 Grade 1 22350 NA 0.1428571 1.0 85  
## 4 108210 Grade 1 5572 41 0.7428572 1.0 89  
## 5 108210 Grade 1 21191 12 0.4000000 0.9 89  
## 6 108210 Grade 1 10184 33 0.8857143 1.0 57  
## present\_records teacher\_id lesson\_completion\_rate  
## 1 72 323 0.3953488  
## 2 74 323 0.3953488  
## 3 26 323 0.3953488  
## 4 87 323 0.3953488  
## 5 59 323 0.3953488  
## 6 55 323 0.3953488

#merge school information  
pupil\_teacher\_school\_df <- merge(pupil\_teacher\_df, School\_information, by = "school\_id")  
head(pupil\_teacher\_school\_df)

## school\_id grade pupil\_id Fluency Kiswahili Math attendance\_records  
## 1 416 Grade 1 23222 43 0.6571429 0.9666666 85  
## 2 416 Grade 1 8377 11 0.1428571 0.8666667 85  
## 3 416 Grade 1 11313 26 0.1428571 0.7666667 85  
## 4 416 Grade 1 5052 38 0.5428572 1.0000000 85  
## 5 416 Grade 1 6151 21 0.1428571 0.7666667 85  
## 6 416 Grade 1 2097 10 0.2000000 0.6666667 85  
## present\_records teacher\_id lesson\_completion\_rate region province  
## 1 76 505 0.5684008 Kirinyaga Central  
## 2 69 505 0.5684008 Kirinyaga Central  
## 3 64 505 0.5684008 Kirinyaga Central  
## 4 77 505 0.5684008 Kirinyaga Central  
## 5 59 505 0.5684008 Kirinyaga Central  
## 6 62 505 0.5684008 Kirinyaga Central  
## tutoring\_program  
## 1 No  
## 2 No  
## 3 No  
## 4 No  
## 5 No  
## 6 No

This is the the main data set that we will work with. Let’s export this as a csv file and name it “main\_data”

write.csv(pupil\_teacher\_school\_df, file = "main\_data.csv", row.names = FALSE)

# Step 2: Calculating KPIs

One of our main KPIs within the Schools Vertical is “Percent Pupils Present”. The “layman’s definition” of this KPI is “The percentage of pupils who were present, out of all pupils - across all days in the term to date”. In other words, the percentage of pupils who were present (for each pupil in the “Pupil attendance” file, this is displayed in the “present\_records” variable), out of pupils who had attendance records (the “attendance\_records” variable in the same file).

● The first task is to translate this KPI into the data. We will calculate this KPI in two different ways. First, calculate this KPI for all pupils at once. What is the network-level average Percent Pupils Present (use two decimal points)?

# Network-Level Average Percent Pupils Present (All Pupils)  
network\_level\_average\_kpi <- round(sum(pupil\_teacher\_school\_df$present\_records)/sum(pupil\_teacher\_school\_df$attendance\_records), 2)  
network\_level\_average\_kpi

## [1] 0.76

● Now, please calculate this percentage for each school, and create an average at the school-level. What is the average Percent Pupils Present now (use two decimal points)?

# School-Level Average Percent Pupils Present  
school\_level\_average\_kpi <- pupil\_teacher\_school\_df %>%  
 select(school\_id, present\_records, attendance\_records) %>%  
 group\_by(school\_id) %>%  
 summarise(total\_present = sum(present\_records), total\_records = sum(attendance\_records)) %>%  
 mutate(school\_kpi = round(total\_present/total\_records, 2)) %>%  
 summarise(round(mean(school\_kpi), 2))  
school\_level\_average\_kpi

## # A tibble: 1 × 1  
## `round(mean(school\_kpi), 2)`  
## <dbl>  
## 1 0.76

● How does the interpretation of the KPI change between the two approaches? Does it matter in this case? When would it matter, (i.e., when would one be more appropriate than the other?) 2-4 sentences max.

# The way we interpret the KPI shifts with these two approaches due to their scope. When we calculate the network-level average, we're looking at the big picture, assessing how well the entire network (Bridge Kenya programme) is performing by considering all pupils across all schools. On the other hand, the school-level average narrows our focus to individual school performance, helping us pinpoint differences between schools. In this case it doesn't matter much because we get the same values for the KPIs through both approaches when rounded to two decimal places. However, it would matter when there is more heterogeneity across schools in terms of attendance rate and number of pupils. If there is more variation in attendance rate across schools and number of pupils, merely calculating percentage for each school and creating a simple average at the school-level would give the same weightage to each scool regardless of the number of pupils in that school, hence the value will deviate from the network-level average. The choice between these approaches hinges on the specific analysis or decision-making context. The network-level approach would one be more appropriate when we want to gauge the overall network performance, whereas the school-level approach is valuable for recognizing variations and addressing specific issues within each school. Ultimately, the choice depends on the specific objectives of the analysis or decision-making process.

# Step 3: Descriptives

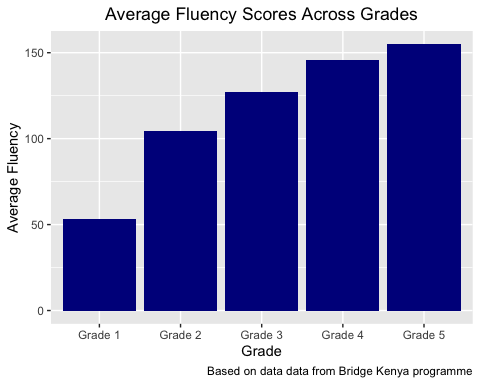
Let’s dig into the reading fluency scores in your current data set. These came from the “Pupil scores” data, but you will need the data set you created in Step 1 above to answer these questions. Please answer the following questions as succinctly as possible.

Please create a figure or a table, whichever you prefer, which shows average fluency scores for each of the five grades.

pupil\_teacher\_school\_df %>%  
 select(grade, Fluency) %>%  
 group\_by(grade) %>%  
 summarise(mean(Fluency, na.rm = TRUE))

## # A tibble: 5 × 2  
## grade `mean(Fluency, na.rm = TRUE)`  
## <chr> <dbl>  
## 1 Grade 1 53.1  
## 2 Grade 2 104.   
## 3 Grade 3 127.   
## 4 Grade 4 145.   
## 5 Grade 5 155.

pupil\_teacher\_school\_df %>%  
 select(grade, Fluency) %>%  
 group\_by(grade) %>%  
 summarise(avg\_fluency = mean(Fluency, na.rm = TRUE)) %>%  
 ggplot(aes(x = grade, y=avg\_fluency))+  
 geom\_col(fill= "dark blue")+  
 labs(title = "Average Fluency Scores Across Grades",  
 x = "Grade",  
 y = "Average Fluency",  
 caption = "Based on data data from Bridge Kenya programme")+  
 theme(plot.title = element\_text(hjust = 0.5))



● Which regions (using the “region” variable) have the lowest and highest average fluency score across all grades?

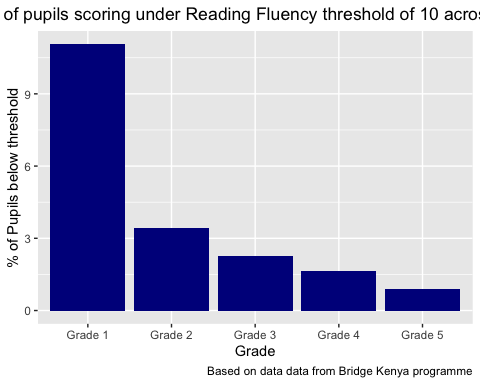
pupil\_teacher\_school\_df %>%  
 select(region, Fluency) %>%  
 group\_by(region) %>%  
 summarise(avg\_fluency = mean(Fluency, na.rm = TRUE)) %>%  
 filter(avg\_fluency == max(avg\_fluency, na.rm=TRUE) | avg\_fluency == min(avg\_fluency, na.rm=TRUE))

## # A tibble: 2 × 2  
## region avg\_fluency  
## <chr> <dbl>  
## 1 Kirinyaga 60.3  
## 2 Machakos 158.

Kirinyaga has the lowest average fluency score across all grades. Machakos has the highest average fluency score across all grades.

● Please create a binary variable that is 1 if a given child reads at 10 or lower, and 0 otherwise. Please create a bar chart with grades on the x-axis, and the share of pupils scoring under this threshold for each grade.

pupil\_teacher\_school\_df$not\_fluent <- ifelse(pupil\_teacher\_school\_df$Fluency <=10, 1, 0)  
  
pupil\_teacher\_school\_df %>%  
 select(grade, not\_fluent) %>%  
 group\_by(grade) %>%  
 summarise(proportion\_not\_fluent = mean(not\_fluent, na.rm = TRUE)) %>%  
 ggplot(aes(x = grade, y=proportion\_not\_fluent\*100))+  
 geom\_col(fill= "dark blue")+  
 labs(title = "Share of pupils scoring under Reading Fluency threshold of 10 across Grades",  
 x = "Grade",  
 y = "% of Pupils below threshold",  
 caption = "Based on data data from Bridge Kenya programme")+  
 theme(plot.title = element\_text(hjust = 0.5))



● What school has the highest share of pupils scoring under this threshold in grade 3?

pupil\_teacher\_school\_df %>%  
 filter(grade == "Grade 3") %>%  
 select(school\_id, not\_fluent) %>%  
 group\_by(school\_id) %>%  
 summarise(proportion\_not\_fluent = mean(not\_fluent, na.rm = TRUE)) %>%  
 filter(proportion\_not\_fluent == max(proportion\_not\_fluent, na.rm=TRUE))

## # A tibble: 1 × 2  
## school\_id proportion\_not\_fluent  
## <dbl> <dbl>  
## 1 223941 0.342

school\_id 223941 is the one that has the highest share of pupils scoring under this threshold in grade 3.

# Step 4: Impact evaluation

During this term, we rolled out an intensive after-school tutoring program in 55 schools. The selection to be a part of the 55 schools was randomly assigned - in other words, these schools were part of a randomized controlled trial (RCT). The “School Information” data set has a binary variable for whether each school was part of the program or not.

● Our Chief Academic Officer would like to know whether this program had any effects on test scores in math, Kiswahili, fluency, and/or student attendance. Please conduct any calculations you see fit to answer his questions.

Let’s do the following calculations to transform the columns first and prepare the data for regression analysis.

# Create a column for attendance performance at the student level  
pupil\_teacher\_school\_df$student\_attendance\_score <- 100\*(pupil\_teacher\_school\_df$present\_records/pupil\_teacher\_school\_df$attendance\_records)  
  
# The column for Math, Kiswahili and lesson completion rate currently have values between 0 and 1. Let's multiply them by 100 to show them in percentage form   
  
pupil\_teacher\_school\_df$Math <- pupil\_teacher\_school\_df$Math\*100  
pupil\_teacher\_school\_df$Kiswahili <- pupil\_teacher\_school\_df$Kiswahili\*100  
pupil\_teacher\_school\_df$lesson\_completion\_rate <- pupil\_teacher\_school\_df$lesson\_completion\_rate\*100

summary(lm(Math ~ tutoring\_program, data = pupil\_teacher\_school\_df))

##   
## Call:  
## lm(formula = Math ~ tutoring\_program, data = pupil\_teacher\_school\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -74.010 -14.375 2.735 19.402 29.402   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 70.5982 0.2852 247.56 <2e-16 \*\*\*  
## tutoring\_programYes 3.7767 0.4026 9.38 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 22.13 on 12085 degrees of freedom  
## (614 observations deleted due to missingness)  
## Multiple R-squared: 0.007227, Adjusted R-squared: 0.007145   
## F-statistic: 87.98 on 1 and 12085 DF, p-value: < 2.2e-16

summary(lm(Kiswahili ~ tutoring\_program, data = pupil\_teacher\_school\_df))

##   
## Call:  
## lm(formula = Kiswahili ~ tutoring\_program, data = pupil\_teacher\_school\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -80.660 -12.948 6.338 19.340 32.052   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 67.9482 0.2923 232.4 <2e-16 \*\*\*  
## tutoring\_programYes 12.7121 0.4127 30.8 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 22.7 on 12099 degrees of freedom  
## (600 observations deleted due to missingness)  
## Multiple R-squared: 0.07272, Adjusted R-squared: 0.07264   
## F-statistic: 948.8 on 1 and 12099 DF, p-value: < 2.2e-16

summary(lm(Fluency ~ tutoring\_program, data = pupil\_teacher\_school\_df))

##   
## Call:  
## lm(formula = Fluency ~ tutoring\_program, data = pupil\_teacher\_school\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -130.01 -56.02 -12.02 49.98 252.99   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 97.2551 0.9118 106.66 <2e-16 \*\*\*  
## tutoring\_programYes 32.7602 1.2810 25.57 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 70.09 on 11974 degrees of freedom  
## (725 observations deleted due to missingness)  
## Multiple R-squared: 0.05179, Adjusted R-squared: 0.05171   
## F-statistic: 654 on 1 and 11974 DF, p-value: < 2.2e-16

summary(lm(student\_attendance\_score ~ tutoring\_program, data = pupil\_teacher\_school\_df))

##   
## Call:  
## lm(formula = student\_attendance\_score ~ tutoring\_program, data = pupil\_teacher\_school\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -77.166 -8.931 3.993 13.030 25.300   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 74.7002 0.2232 334.613 < 2e-16 \*\*\*  
## tutoring\_programYes 2.4660 0.3159 7.806 6.35e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.8 on 12699 degrees of freedom  
## Multiple R-squared: 0.004776, Adjusted R-squared: 0.004697   
## F-statistic: 60.94 on 1 and 12699 DF, p-value: 6.348e-15

# By running simple linear regression with Math, Kiswahili, Fluency, and Attendance score as the dependent variable and tutoring program as the independent variable we saw that tutoring program has a statistically significant positive effect on students' test scores in Math, Kiswahili, Fluency, and attendance on average. The results are highly statistically significant even at a significance level as low as 0.001.

● After conducting the impact evaluation, we have heard anecdotally that teachers in schools that received tutoring felt more motivated and were completing their lessons at a faster pace. Hence, we could worry that the effects that we see are not (solely) due to the tutoring program, but also due to the higher lesson completion rate. Does this hypothesis hold up in the data?

# If teachers in schools that received tutoring indeed felt more motivated and were completing their lessons at a faster pace then tutoring program also has an effect on teachers completion rate. In that case teacher's lesson completion rate would be a confounding variable that also effects the outcome variable and is being effected by the treatment variable (tutoring program). If this is true, omitting teachers lesson completion rate from the regressions would make the results biased as the effects of teachers' higher lesson completion rate would also be wrongly attributed to the tutoring program.  
#Let's first run a diagnostic regression regression of teachers lesson completion rate on tutoring program to see if there is any effect of tutoring program on lesson completion rate. Then, we will modify the above regressions by controlling for the effects of lesson completion rate by including this variable in our regression analysis as a covariate.

summary(lm(lesson\_completion\_rate ~ tutoring\_program, data = pupil\_teacher\_school\_df))

##   
## Call:  
## lm(formula = lesson\_completion\_rate ~ tutoring\_program, data = pupil\_teacher\_school\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -60.854 -16.288 4.733 17.321 39.588   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 60.4115 0.2810 214.996 <2e-16 \*\*\*  
## tutoring\_programYes 0.4425 0.3976 1.113 0.266   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 22.41 on 12699 degrees of freedom  
## Multiple R-squared: 9.753e-05, Adjusted R-squared: 1.879e-05   
## F-statistic: 1.239 on 1 and 12699 DF, p-value: 0.2657

# In the above regression results, we see that tutoring program does not have a statistically significant effect on lesson completion rate. This suggests that it is not a significant confounder. However, it might be confounding in combination with the tutoring program variable so we should still use it as a control variable. Below, we control for lesson\_completion\_rate to see if the results of regression vary from the above ones.

summary(lm(Math ~ tutoring\_program+lesson\_completion\_rate, data = pupil\_teacher\_school\_df))

##   
## Call:  
## lm(formula = Math ~ tutoring\_program + lesson\_completion\_rate,   
## data = pupil\_teacher\_school\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -76.050 -14.325 2.933 19.017 36.814   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 63.140229 0.609769 103.548 <2e-16 \*\*\*  
## tutoring\_programYes 3.733093 0.399538 9.344 <2e-16 \*\*\*  
## lesson\_completion\_rate 0.122980 0.008907 13.807 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21.96 on 12084 degrees of freedom  
## (614 observations deleted due to missingness)  
## Multiple R-squared: 0.02265, Adjusted R-squared: 0.02249   
## F-statistic: 140 on 2 and 12084 DF, p-value: < 2.2e-16

summary(lm(Kiswahili ~ tutoring\_program+lesson\_completion\_rate, data = pupil\_teacher\_school\_df))

##   
## Call:  
## lm(formula = Kiswahili ~ tutoring\_program + lesson\_completion\_rate,   
## data = pupil\_teacher\_school\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -81.88 -12.76 6.12 18.44 35.05   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 64.953350 0.629297 103.216 < 2e-16 \*\*\*  
## tutoring\_programYes 12.693886 0.412233 30.793 < 2e-16 \*\*\*  
## lesson\_completion\_rate 0.049367 0.009189 5.372 7.92e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 22.67 on 12098 degrees of freedom  
## (600 observations deleted due to missingness)  
## Multiple R-squared: 0.07492, Adjusted R-squared: 0.07477   
## F-statistic: 489.9 on 2 and 12098 DF, p-value: < 2.2e-16

summary(lm(Fluency ~ tutoring\_program+lesson\_completion\_rate, data = pupil\_teacher\_school\_df))

##   
## Call:  
## lm(formula = Fluency ~ tutoring\_program + lesson\_completion\_rate,   
## data = pupil\_teacher\_school\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -148.52 -55.43 -10.80 49.43 240.96   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 115.75903 1.95379 59.25 <2e-16 \*\*\*  
## tutoring\_programYes 32.76266 1.27499 25.70 <2e-16 \*\*\*  
## lesson\_completion\_rate -0.30357 0.02839 -10.70 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 69.76 on 11973 degrees of freedom  
## (725 observations deleted due to missingness)  
## Multiple R-squared: 0.06076, Adjusted R-squared: 0.06061   
## F-statistic: 387.3 on 2 and 11973 DF, p-value: < 2.2e-16

summary(lm(student\_attendance\_score ~ tutoring\_program+lesson\_completion\_rate, data = pupil\_teacher\_school\_df))

##   
## Call:  
## lm(formula = student\_attendance\_score ~ tutoring\_program + lesson\_completion\_rate,   
## data = pupil\_teacher\_school\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -77.481 -8.920 4.001 12.993 26.241   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 73.759424 0.480803 153.409 < 2e-16 \*\*\*  
## tutoring\_programYes 2.459136 0.315868 7.785 7.49e-15 \*\*\*  
## lesson\_completion\_rate 0.015573 0.007049 2.209 0.0272 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.8 on 12698 degrees of freedom  
## Multiple R-squared: 0.005158, Adjusted R-squared: 0.005001   
## F-statistic: 32.92 on 2 and 12698 DF, p-value: 5.501e-15

# By including lesson completion rate variable in our regression analysis as a covariate, we saw that the coefficients for tutoring program in all of the four regressions remained almost the same as before and were not effected much. Although we do see a statistically significant effect of lesson completion rate on students performance and attendance, this variable is independently assocaited with the outcome variables and does not mediate the effects of the tutoring program.  
# The hypothesis that the effects that we saw are not (solely) due to the tutoring program, but also due to the higher lesson completion rate does not hold up in the data.