CollegeDistance dataset: Effect Of Background Factors On Individuals' Education

Robin Tran

May 3, 2022

```
library(AER)
library(dplyr) # functions like summarize
library(ggplot2) # for making plots
library(readr)
library(gridExtra)
library(GGally)
library(gmodels)
```

```
data("CollegeDistance")
```

Introduction

This paper constructs statistical models to identify two main problems of interest, using the CollegeDistance dataset which first appeared in Rouse C.E's journal (1995) on the effects of community colleges on educational attainment. In their journal, Rouse discussed the effect of college access on the educational achievement of an individual, by analyzing data collected about members of the high school senior class in different areas. In this paper, the focus will be on how an individual's background can affect their education level, considering factors such as distance from home to the nearest 4-year college, ethnicity, and family income. The main hypotheses are:

(1) Distance from home to a 4-year college of a high school student does affect the number of education years they receive.

(2) Ethnicity of a high school student does affect the number of education years they receive.

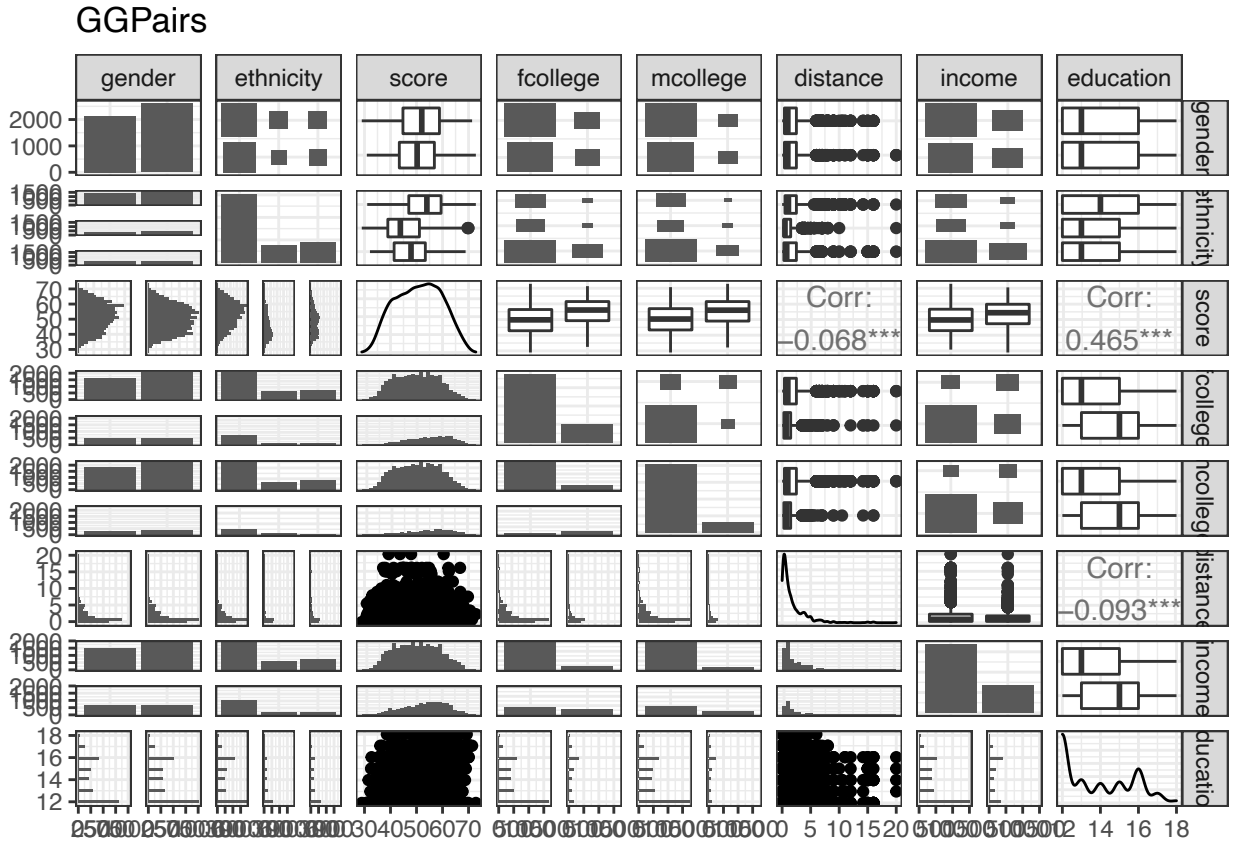
(3) Family income of a high school student does affect the number of education years they receive.

This paper will use all subsets regression methods to come up with the most effective linear regression model to test the hypotheses. Transformation will be applied to satisfy multiple linear regression conditions as much as possible. Lastly, results from the tests will be stated and concluded in text.

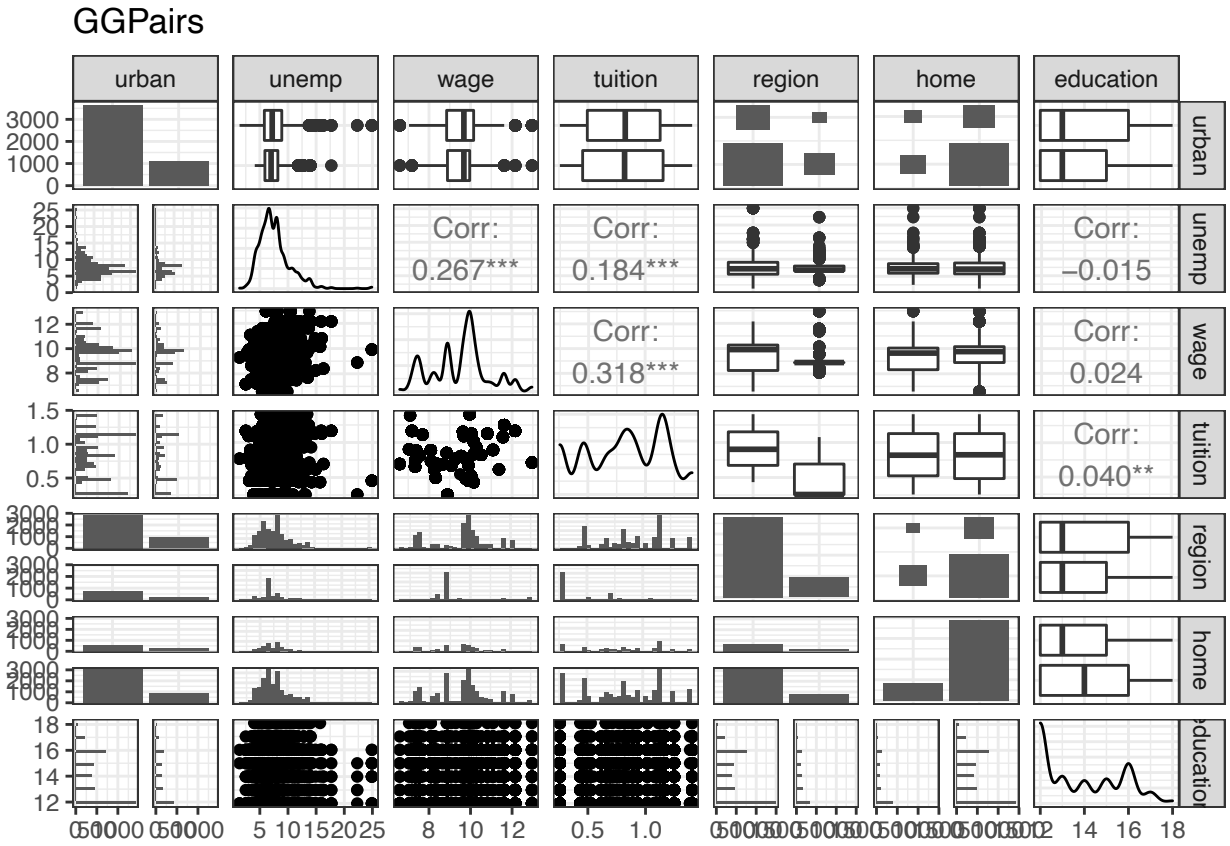
Variables selection

The CollegeDistance dataset contains 14 variables, including information about high school senior individuals, as well as information about the areas they were living in. In this paper, we will discuss gender, ethnicity, score (base year composite test score), fcollege (whether the father is a college graduate), mcollege (whether the mother is a college graduate), distance (distance from 4-year college), income (whether the family income is above $25000 per year), home (whether family income above USD 25,000 per year), urban (whether the school is in urban area), unemp (county unemployment rate in 1980), wage (state hourly wage in manufacturing in 1980), tuition (average state 4-year college tuition), region (West or other), and years of education.

```
library(GGally)
ggpairs(CollegeDistance %>% select(gender, ethnicity, score, fcollege, mcollege, distance, income, educa
  ggtitle("GGPairs ") +
  theme_bw()
```




```
ggpairs(CollegeDistance %>% select(urban, unemp, wage, tuition, region, home, education)) +
  ggtitle("GGPairs") +
  theme_bw()
```



Summary Statistics for Numerical Variables

```
options(knitr.table.format = "latex", knitr.kable.NA = "")
```

```
library(kableExtra)

kable2 <- function(data, ...) {
  knitr::kable(data, ..., booktabs = TRUE, escape = FALSE, digits = 3) %>%
    kable_styling(position = "center", latex_options = "HOLD_position") }
kable2(CollegeDistance_sumstats)
```

variable	Mean	SD	Min	Max
score	50.889	8.702	28.950	72.810
distance	1.803	2.297	0.000	20.000
unemp	7.597	2.764	1.400	24.900
wage	9.501	1.343	6.590	12.960
tuition	0.815	0.340	0.258	1.404
education	13.808	1.789	12.000	18.000

Model assumptions

Firstly, the model with all variables will be fitted, along with a summary of the model. We will check the assumptions about multiple linear regression of this model. Independent observations might not be satisfied in this case when data was collected in nested structure – schools, counties, and states. Students from the same high school might have similar characteristics to each other, students from the same area (county, state)

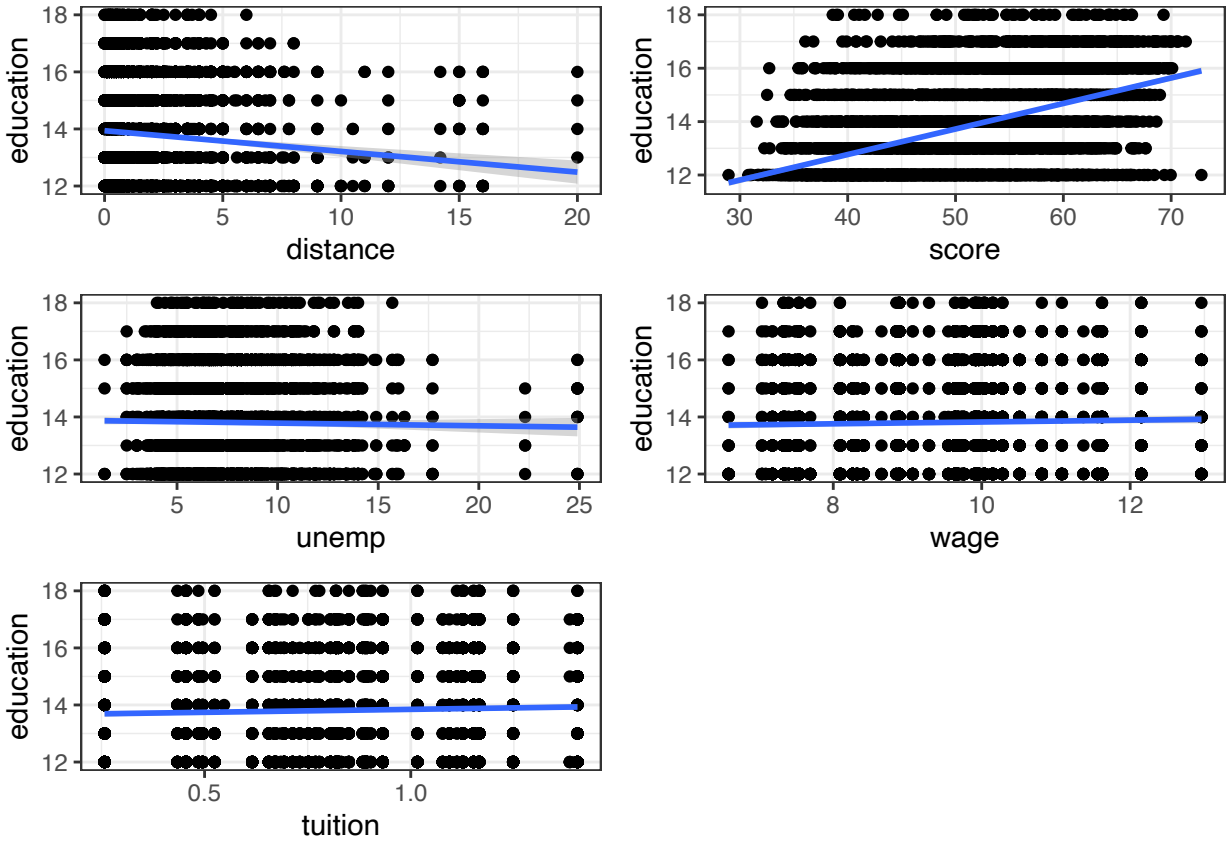
have the same values of average wage, etc. There the observations might not be independent. Also, the data was collected as a survey – a convenient method. For further analysis, we will assume that independence is satisfied, but it is noted that there is a problem of nested effect in this dataset.

```
allvariables_fit <- lm(data = CollegeDistance, education ~ gender + ethnicity + score + fcollege + mcoll
summary(allvariables_fit)
```

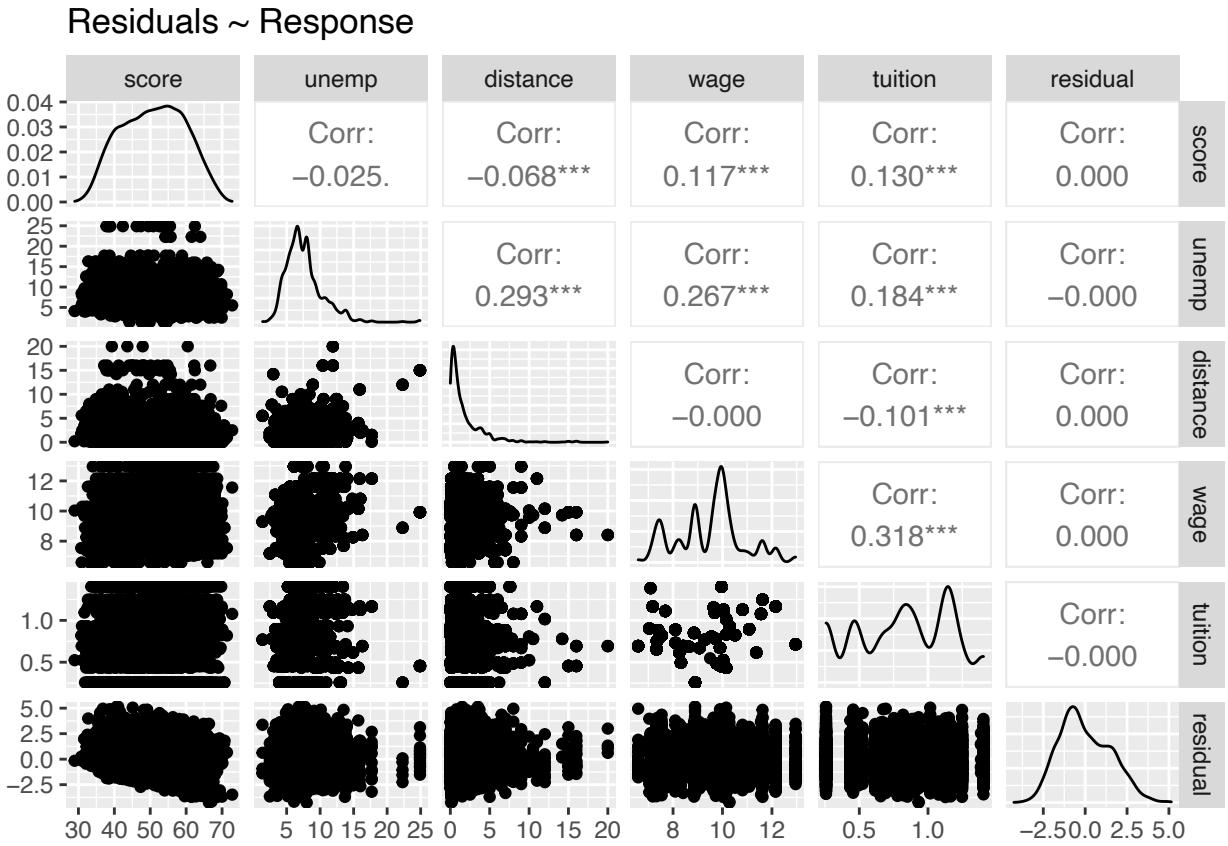
```
##
## Call:
## lm(formula = education ~ gender + ethnicity + score + fcollege +
##     mcollege + home + urban + unemp + wage + distance + tuition +
##     income + region, data = CollegeDistance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2586 -1.1251 -0.2188  1.1335  5.1428
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.044064   0.227614  39.734  < 2e-16 ***
## genderfemale       0.129014   0.044891   2.874 0.004072 **
## ethnicityafam      0.308261   0.067611   4.559 5.26e-06 ***
## ethnicityhispanic  0.317455   0.063675   4.986 6.40e-07 ***
## score              0.089834   0.002847  31.558  < 2e-16 ***
## fcollegeyes        0.553740   0.064308   8.611  < 2e-16 ***
## mcollegeyes        0.383678   0.072338   5.304 1.18e-07 ***
## homeyes            0.145662   0.059265   2.458 0.014015 *
## urbanyes           0.040734   0.056521   0.721 0.471138
## unemp              0.030046   0.009023   3.330 0.000876 ***
## wage              -0.037013   0.018299  -2.023 0.043162 *
## distance          -0.036449   0.010822  -3.368 0.000763 ***
## tuition           -0.209261   0.089600  -2.335 0.019559 *
## incomehigh         0.381584   0.053728   7.102 1.41e-12 ***
## regionwest        -0.188441   0.069813  -2.699 0.006975 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.528 on 4724 degrees of freedom
## Multiple R-squared:  0.273,  Adjusted R-squared:  0.2708
## F-statistic: 126.7 on 14 and 4724 DF,  p-value: < 2.2e-16
```

Plots showing the relationship between numerical explanatory variables and the response variables education are included for a clearer view.

```
grid.arrange(p1, p2, p3, p4, p5, ncol = 2)
```



There are linear relationships between the numerical explanatory variable (distance, score, unemp, wage, tuition) and the response variable (education). Another limitation of this model is that education is not technically a continuous numerical variable, or can be categorized as a categorical variable. Therefore, the graphs are not ideal, and the linear relationship is not clear for education ~ wage, education ~ unemp, and education ~ tuition plots.



Residuals are quite normally distributed, but can still be improved. We proceed by checking the equal standard deviation of response for all values for our numerical explanatory variables. Points scatter around line 0.0 without a particular pattern, but it still looks problematic. One approach is to apply a transformation to the variables; in this case, we can use log transformation.

```
CollegeDistance <- CollegeDistance %>%
  mutate(
    log_education = log(education),
    log_score = log(score),
    log_distance = log(distance+1),
    log_wage = log(wage),
    log_tuition = log(tuition),
    log_unemp = log(unemp)
  )
```

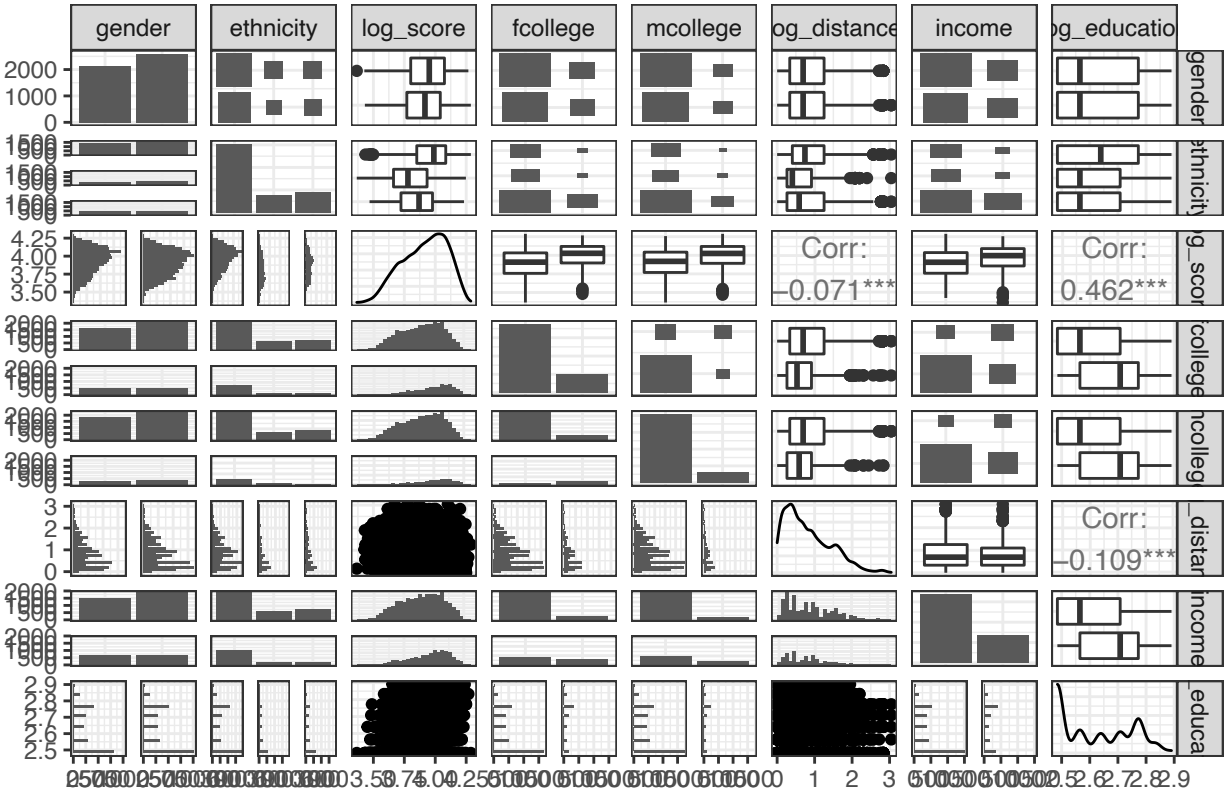
We fit the model again with the transformed variables and view its summary.

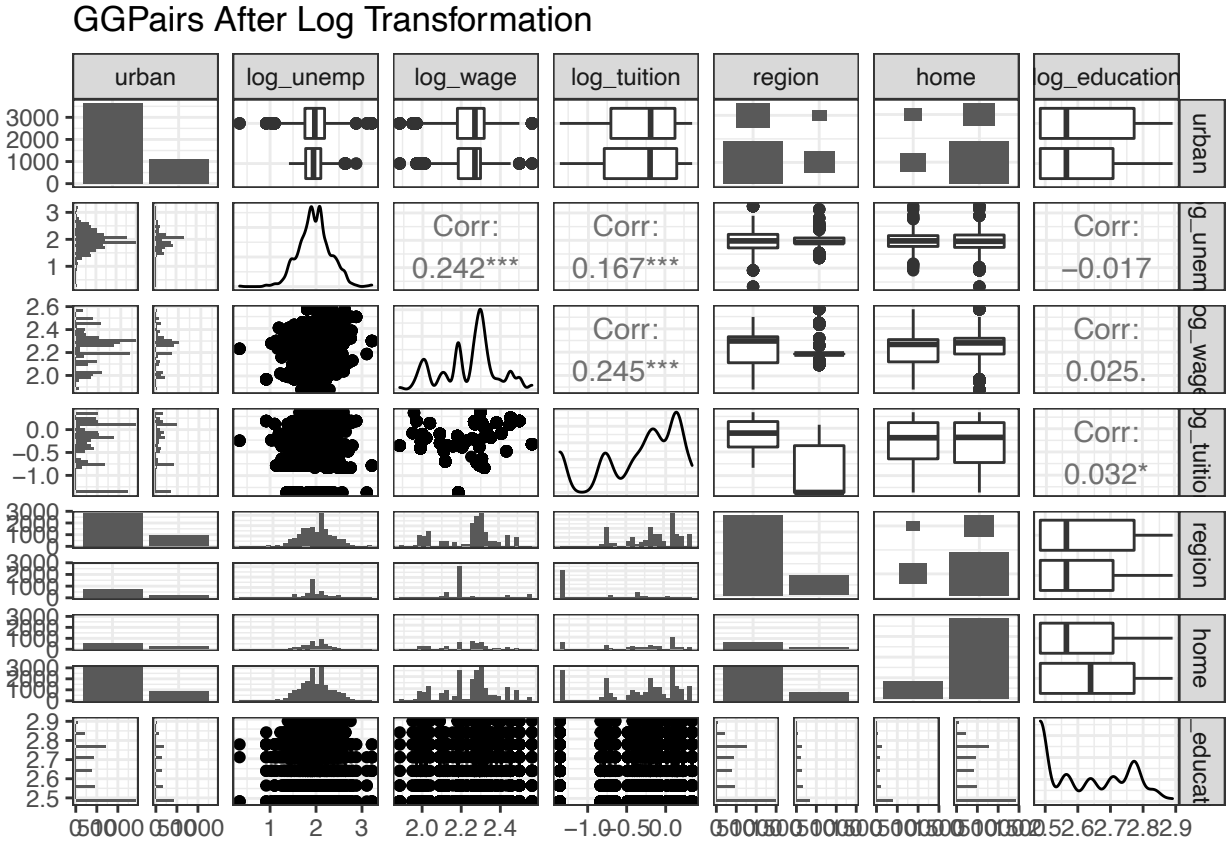
```
allvariables_fit <- lm(data = CollegeDistance, log_education ~ gender + ethnicity + log_score + fcolleg
summary(allvariables_fit)
```

```
##
## Call:
## lm(formula = log_education ~ gender + ethnicity + log_score +
##     fcollege + mcollege + home + urban + log_unemp + log_wage +
##     log_distance + log_tuition + income + region, data = CollegeDistance)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30011 -0.08305 -0.01018  0.08371  0.34430
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.378684   0.047303  29.146  < 2e-16 ***
## genderfemale       0.008672   0.003192   2.717 0.006610 **
## ethnicityafam      0.022771   0.004839   4.705 2.61e-06 ***
## ethnicityhispanic  0.021752   0.004529   4.803 1.61e-06 ***
## log_score          0.313428   0.010028  31.256  < 2e-16 ***
## fcollegeyes        0.039944   0.004583   8.715  < 2e-16 ***
## mcollegeyes        0.028231   0.005144   5.488 4.28e-08 ***
## homeyes            0.010122   0.004217   2.400 0.016415 *
## urbanyes           0.001745   0.004121   0.423 0.671967
## log_unemp          0.016612   0.004980   3.336 0.000857 ***
## log_wage          -0.024415   0.011879  -2.055 0.039896 *
## log_distance      -0.010143   0.002906  -3.490 0.000487 ***
## log_tuition       -0.010742   0.004509  -2.383 0.017229 *
## incomehigh         0.027150   0.003824   7.100 1.43e-12 ***
## regionwest        -0.015393   0.005462  -2.818 0.004848 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1087 on 4724 degrees of freedom
## Multiple R-squared:  0.2724, Adjusted R-squared:  0.2702
## F-statistic: 126.3 on 14 and 4724 DF,  p-value: < 2.2e-16
```

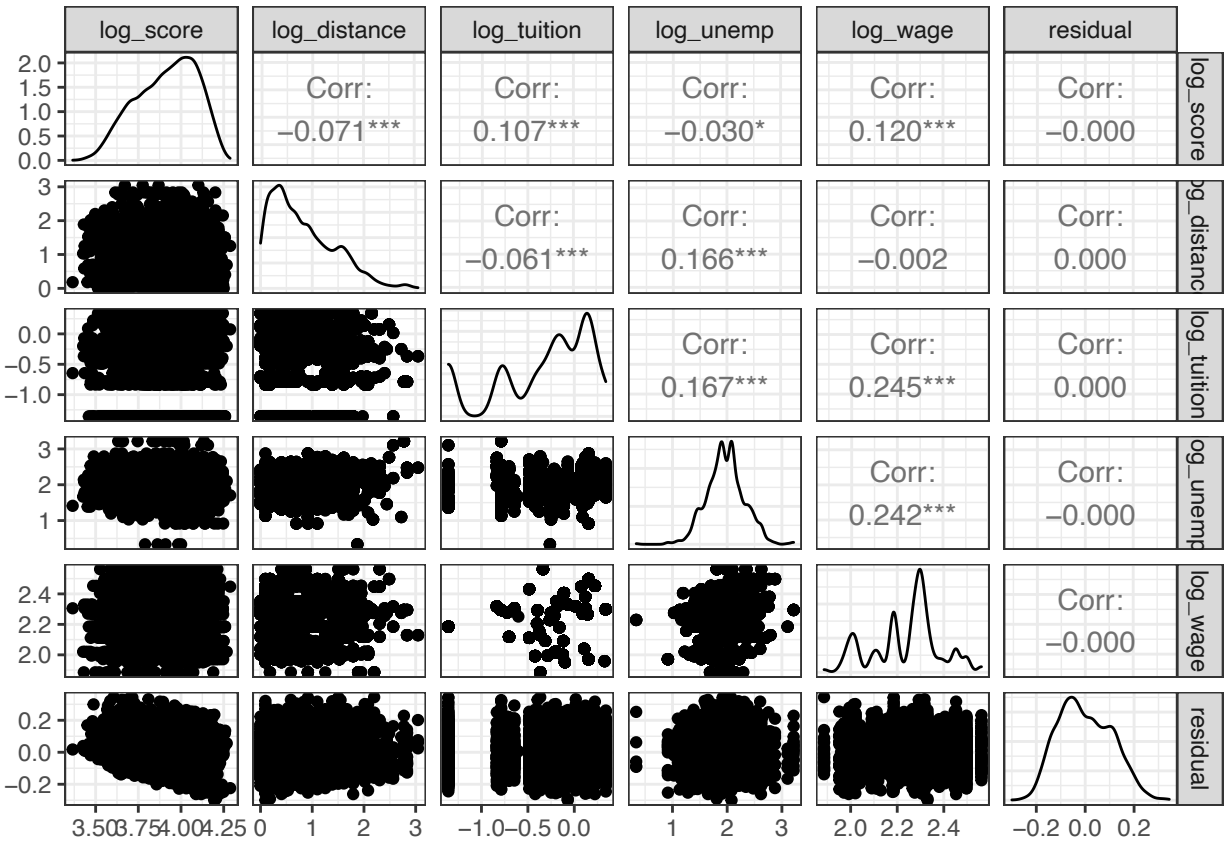
GGPairs After Log Transformation





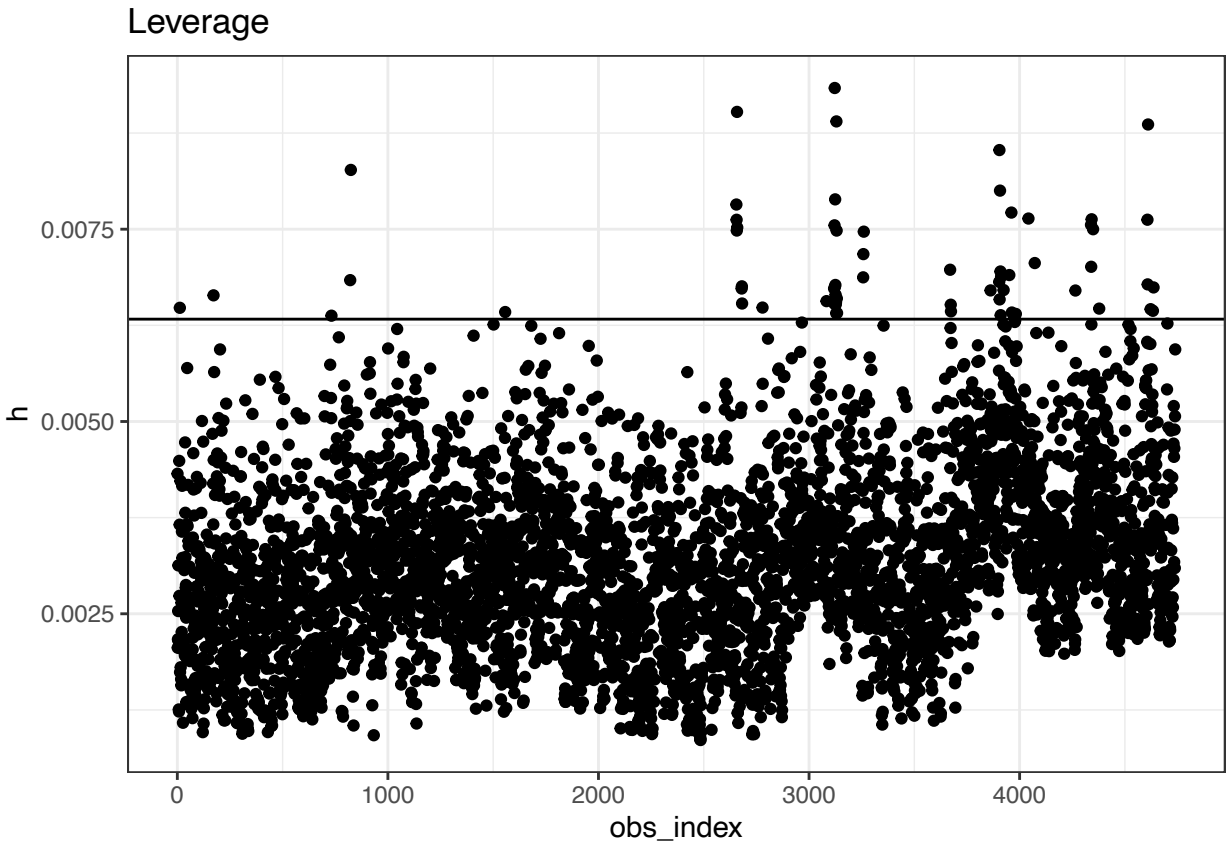
The linear relationships between the numerical explanatory variable (log_distance, log_score, log_unemp, log_wage, log_tution) and the response variable (log_education) are improved, but there is still problem of categorical response variable.

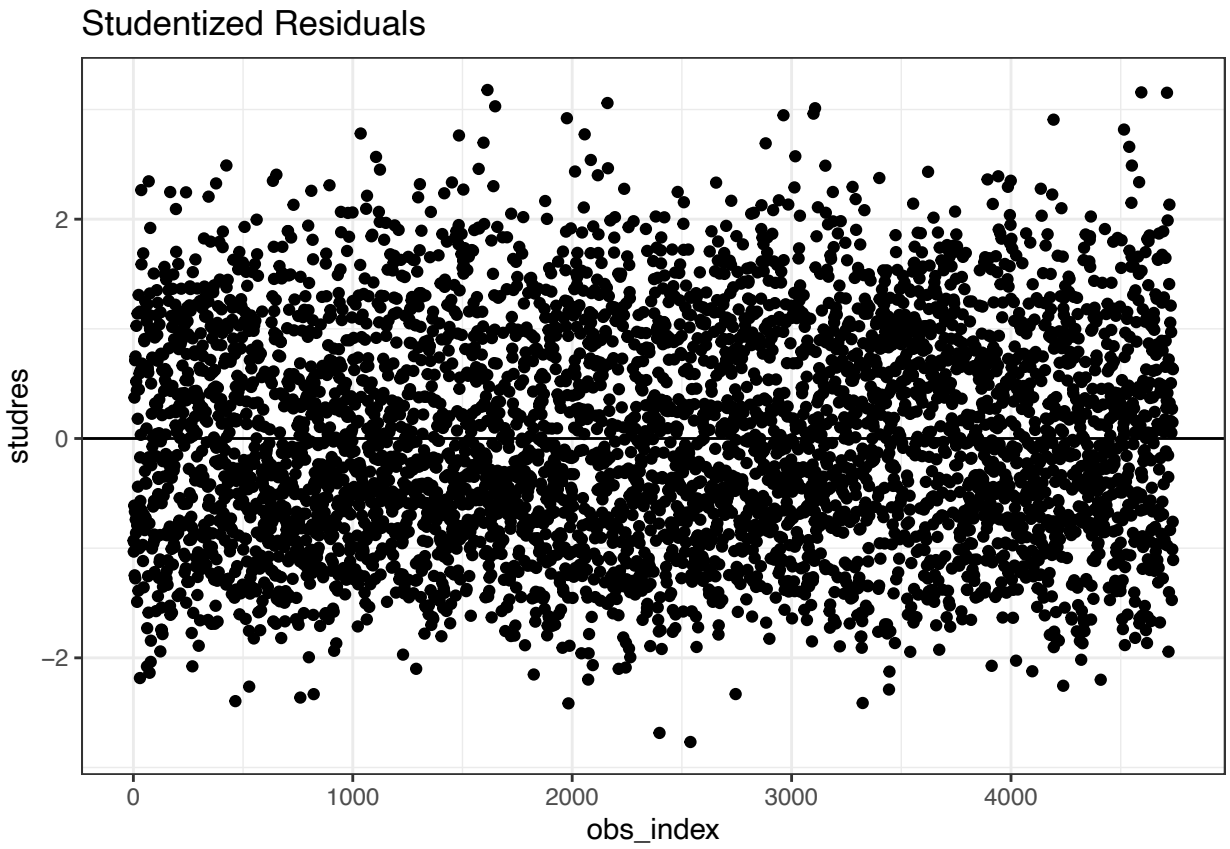
```
CollegeDistance <- CollegeDistance %>%
  mutate(
    residual = residuals(allvariables_fit)
  )
```

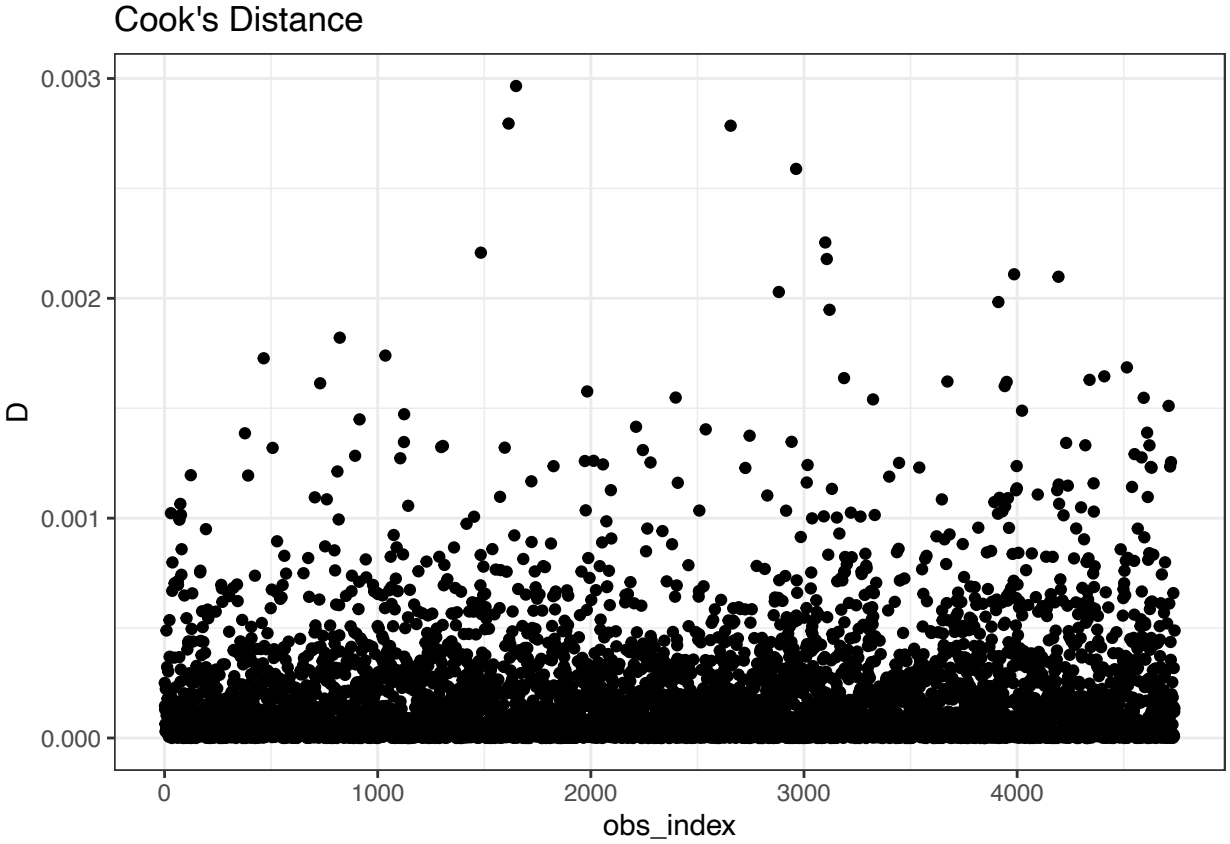


After the log transformation on score, distance, tuition, unemp, wage, and education, the residual plot looks more bell-curved, and the equal standard deviation of response assumption has improved (it can be seen in the explanatory ~ residual plots). The last condition that requires considering is influential points and leverage, which can be tested using diagnostic plots, including leverage plot, studentized residuals plot, and Cook's distance plot.

```
CollegeDistance <- CollegeDistance %>%
  mutate(
    obs_index = row_number(),
    h = hatvalues(allvariables_fit),
    studres = rstudent(allvariables_fit),
    D = cooks.distance(allvariables_fit)
  )
```


Leverage
h
0.0075
0.0050
0.0025
0
1000
2000
3000
4000
obs_index

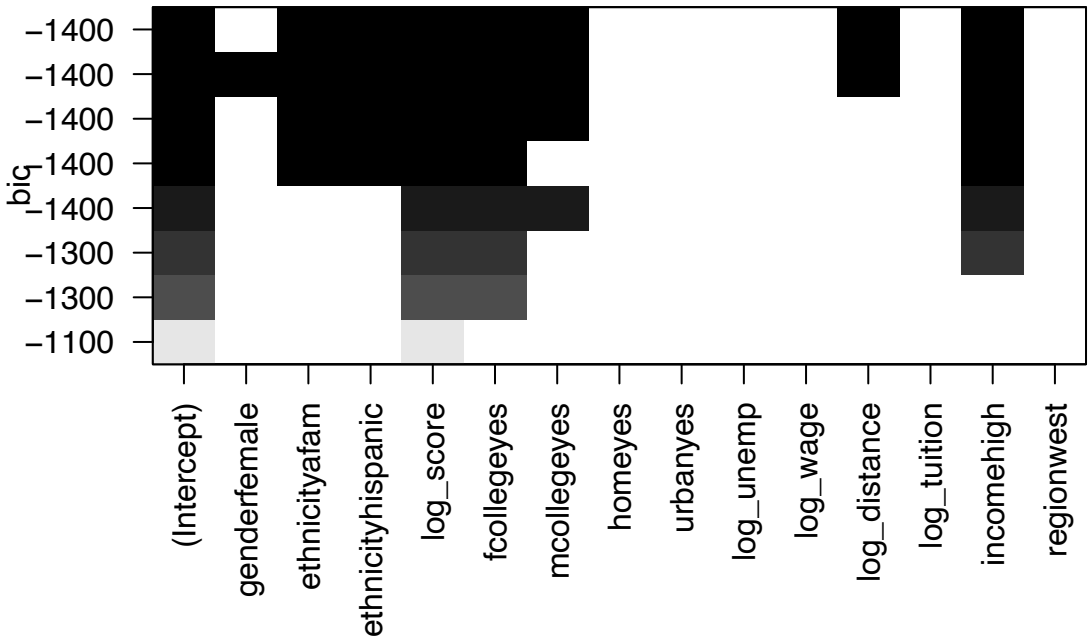

Studentized Residuals
studres
2
0
−2
0
1000
2000
3000
4000
obs_index



As observed in the diagnostic plots, no point specifically stands out at an influential point. We can proceed with the model selection.

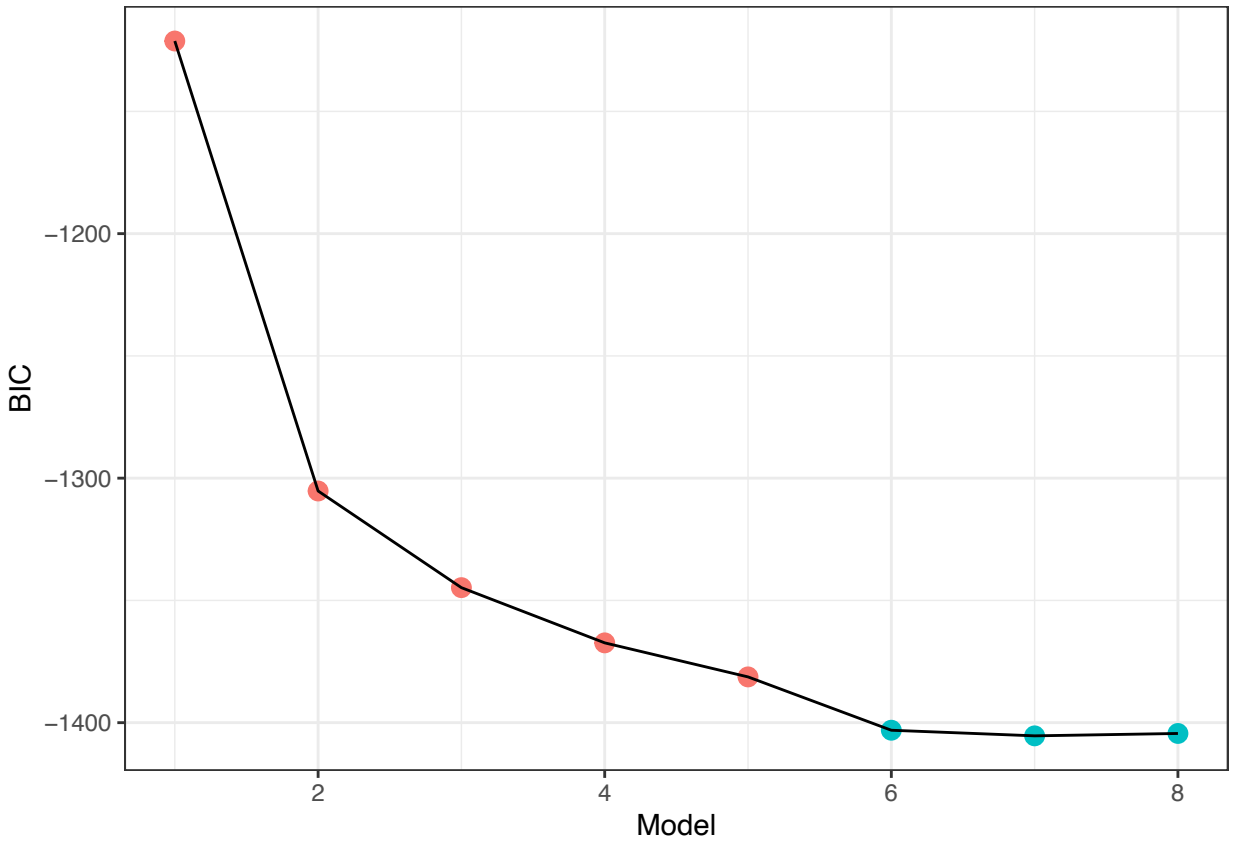
Model selection

For model selection, this paper will use all subsets regression and determine the model that contains both our variables of interest and the smallest BIC (Schwarz's Bayesian Information Criterion).



```
summary(CollegeDistance_model1)
```

```
## Subset selection object
## Call: regsubsets.formula(data = CollegeDistance, log_education ~ gender +
##     ethnicity + log_score + fcollege + mcollege + home + urban +
##     log_unemp + log_wage + log_distance + log_tuition + income +
##     region)
## 14 Variables  (and intercept)
##                   Forced in Forced out
## genderfemale          FALSE      FALSE
## ethnicityafam         FALSE      FALSE
## ethnicityhispanic     FALSE      FALSE
## log_score             FALSE      FALSE
## fcollegeyes           FALSE      FALSE
## mcollegeyes           FALSE      FALSE
## homeyes               FALSE      FALSE
## urbanyes              FALSE      FALSE
## log_unemp             FALSE      FALSE
## log_wage              FALSE      FALSE
## log_distance          FALSE      FALSE
## log_tuition           FALSE      FALSE
## incomehigh            FALSE      FALSE
## regionwest            FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          genderfemale ethnicityafam ethnicityhispanic log_score fcollegeyes
## 1  ( 1 ) " "          " "           " "                "*"       " "
## 2  ( 1 ) " "          " "           " "                "*"       "*"
## 3  ( 1 ) " "          " "           " "                "*"       "*"
## 4  ( 1 ) " "          " "           " "                "*"       "*"
## 5  ( 1 ) " "          "*"           "*"                "*"       "*"
## 6  ( 1 ) " "          "*"           "*"                "*"       "*"
## 7  ( 1 ) " "          "*"           "*"                "*"       "*"
## 8  ( 1 ) "*"          "*"           "*"                "*"       "*"
##          mcollegeyes homeyes urbanyes log_unemp log_wage log_distance
## 1  ( 1 ) " "         " "     " "      " "       " "      " "
## 2  ( 1 ) " "         " "     " "      " "       " "      " "
## 3  ( 1 ) " "         " "     " "      " "       " "      " "
## 4  ( 1 ) "*"         " "     " "      " "       " "      " "
## 5  ( 1 ) " "         " "     " "      " "       " "      " "
## 6  ( 1 ) "*"         " "     " "      " "       " "      " "
## 7  ( 1 ) "*"         " "     " "      " "       " "      "*"
## 8  ( 1 ) "*"         " "     " "      " "       " "      "*"
##          log_tuition incomehigh regionwest
## 1  ( 1 ) " "         " "        " "
## 2  ( 1 ) " "         " "        " "
## 3  ( 1 ) " "         "*"        " "
## 4  ( 1 ) " "         "*"        " "
## 5  ( 1 ) " "         "*"        " "
## 6  ( 1 ) " "         "*"        " "
## 7  ( 1 ) " "         "*"        " "
## 8  ( 1 ) " "         "*"        " "
```




```
summary(CollegeDistance_model1)$bic
```

```
## [1] -1121.217 -1305.247 -1344.778 -1367.357 -1381.285 -1403.094 -1405.389
## [8] -1404.415
```

According to the BIC plot, Model 6, Model 7, and Model 8 have roughly similar performances.

Model 6: ethnicity, log_score, fcollege, mcollege, income (BIC = -1403.094)

Model 7: ethnicity, log_score, fcollege, mcollege, log_distance, income (BIC = -1405.389)

Model 8: gender, ethnicity, log_score, fcollege, mcollege, log_distance, income (BIC = -1404.415)

Model 7 has the lowest BIC and also includes all of the variables we are interested in. Therefore, we will select Model 7 for further analysis. Using all subsets regression, we can see that home, urban, unemp, wage, tuition, and region variables are not significant for the model. On the other hand, the variables of our interest - ethnicity, score, fcollege, mcollege, and income are significant and consistent among 3 considered models with the lowest BIC. Distance appeared in 2 out of 3 considered models.

After considering the assumptions and model selection, we come up with a model which presumably satisfies the conditions for multiple linear regression, as well as includes the variables of interest.

```
college_fit <- lm(data = CollegeDistance, log_education ~ ethnicity + log_score + fcollege + mcollege +
summary(college_fit)
```

```
##
## Call:
## lm(formula = log_education ~ ethnicity + log_score + fcollege +
```

```
##     mcollege + income + log_distance, data = CollegeDistance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29334 -0.08366 -0.01074  0.08526  0.35264
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.376957   0.039607  34.765  < 2e-16 ***
## ethnicityafam      0.024572   0.004725   5.201 2.07e-07 ***
## ethnicityhispanic  0.024077   0.004295   5.605 2.20e-08 ***
## log_score          0.311417   0.009978  31.212  < 2e-16 ***
## fcollegeyes        0.038819   0.004583   8.471  < 2e-16 ***
## mcollegeyes        0.027997   0.005153   5.433 5.82e-08 ***
## incomehigh         0.026479   0.003800   6.968 3.66e-12 ***
## log_distance      -0.008774   0.002676  -3.279  0.00105 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.109 on 4731 degrees of freedom
## Multiple R-squared:  0.2672, Adjusted R-squared:  0.2661
## F-statistic: 246.4 on 7 and 4731 DF,  p-value: < 2.2e-16
```

The equation for the model selected:

$\hat{\mu} = 1.376957 + 0.024572 \times$ ethnicityafam $+ 0.024077 \times$ ethnicityhispanic $+ 0.311417 \times$ log_score $+ 0.038819 \times$ fcollegeyes $+ 0.027997 \times$ mcollegeyes $+ 0.026479 \times$ incomehigh - $0.008774 \times$ log_distance

ethnicityafam = 1 if ethnicity = afam, 0 otherwise.

ethnicityhispanic = 1 if ethnicity = hispanic, 0 otherwise.

log_score = log base year composite test score

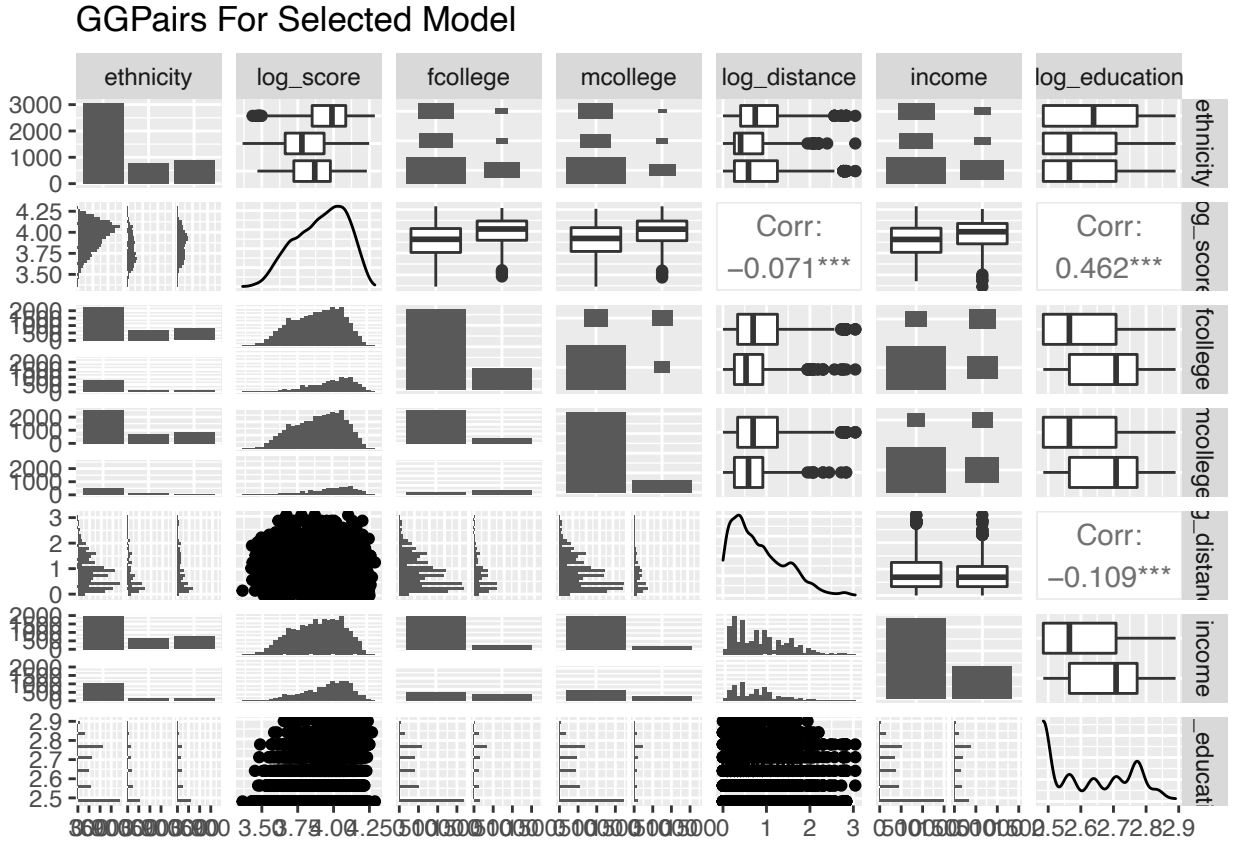
fcollegeyes = 1 if fcollege = yes, 0 otherwise.

mcollegeyes = 1 if mcollege = yes, 0 otherwise.

incomelow = 1 if income = high, 0 otherwise.

log_distance = log distance from home to a 4-year college

A ggpairs plot for selected model is provided below.



Hypotheses

The first relationship we are interested in from this model is the effect of distance from a high school student's home to a college on the number of education years they receive. This is also one focus of Rouse C.E when they first used the dataset and discussed the effect of college accessibility on education opportunities. In this paper, we will use F-test to see if log_distance has significant effect on education.

$H_0 : \beta_0 distance = 0$

$H_A : \beta_0 distance \neq 0$

We will fit a model without the log_distance variable, and use ANOVA method on the full model and the reduced model.

```
model_without_distance <- lm(log_education ~ ethnicity + log_score + fcollege + mcollege + income, data
anova(model_without_distance, college_fit)
```

```
## Analysis of Variance Table
##
## Model 1: log_education ~ ethnicity + log_score + fcollege + mcollege +
##     income
## Model 2: log_education ~ ethnicity + log_score + fcollege + mcollege +
##     income + log_distance
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1   4732 56.315
## 2   4731 56.187  1   0.12771 10.753 0.001048 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for this test is 0.001048, which says there is strong evidence against the null hypothesis that distance has no effect on education. The F-statistic is large and positive, which supports the argument that distance is a contributor of education.

Another hypothesis we are interested in is whether ethnicity affects the number of years of education a high school student achieve. We will also use F-test in this case.

$H_0 : \mu_{afam} = \mu_{hispanic} = \mu_{other}$

There is no difference between the number of years of education among the ethnic groups.

H_A : At least one is different. There is difference between the years of education among the ethnic groups.

```
model_without_ethnicity <- lm(log_education ~ log_score + fcollege + mcollege + log_distance + income, c
anova(model_without_ethnicity, college_fit)
```

```
## Analysis of Variance Table
##
## Model 1: log_education ~ log_score + fcollege + mcollege + log_distance +
##     income
## Model 2: log_education ~ ethnicity + log_score + fcollege + mcollege +
##     income + log_distance
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4733 56.726
## 2   4731 56.187  2   0.53922 22.701 1.542e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for this test is 1.542e-10, which is proof of very strong evidence against the null hypothesis that ethnicity has no effect on education. The F-statistic is large and positive, which supports the argument that ethnicity affects the number of years of education a high school student in the given population receive.

Whether an individual student's family income is high or low can affect the number of years of education a student can receive. F-test will be applied to see the relationship.

```
model_without_income <- lm(log_education ~ ethnicity + log_score + fcollege + mcollege + log_distance, c
anova(model_without_income, college_fit)
```

```
## Analysis of Variance Table
##
## Model 1: log_education ~ ethnicity + log_score + fcollege + mcollege +
##     log_distance
## Model 2: log_education ~ ethnicity + log_score + fcollege + mcollege +
##     income + log_distance
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4732 56.764
## 2   4731 56.187  1   0.57658 48.549 3.665e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for this test is 3.665e-12, which is proof of very strong evidence against the null hypothesis that ethnicity has no effect on education. The F-statistic is large and positive, which supports the argument that family income affects the number of years of education a high school student in the given population receive.

Confidence Intervals For fcollege and mcollege

Having considered other factors of an student's background, such as ethnicity, income, it can be beneficial to see how mother's education level and father's education level can affect a student's education.

```
confint(college_fit)
```

```
##                         2.5 %       97.5 %
## (Intercept)        1.29930785  1.454605271
## ethnicityafam      0.01530933  0.033835098
## ethnicityhispanic  0.01565572  0.032497765
## log_score          0.29185654  0.330977817
## fcollegeyes        0.02983532  0.047802990
## mcollegeyes        0.01789440  0.038099084
## incomehigh         0.01902872  0.033929258
## log_distance      -0.01401921 -0.003528458
```

```
exp(0.02983532)
```

```
## [1] 1.030285
```

```
exp(0.047802990)
```

```
## [1] 1.048964
```

```
exp(0.01789440)
```

```
## [1] 1.018055
```

```
exp(0.038099084)
```

```
## [1] 1.038834
```

We are 95% confident that for high school students whose father was a college graduate and whose father was not, students whose father was a college graduate receive between 1.03028 and 1.048964 more years of education, on average, holding other variables constant.

Similarly, we are 95% confident that for high school students whose mother was a college graduate and whose mother was not, students whose mother was a college graduate receive between 1.018055 and 1.038834 more years of education, on average, holding other variables constant.

Results/Conclusions

The analysis mainly focuses on how different aspects of an individual high school student's background can affect their education level (indicated by years), and the significance of these aspects on the response variable. From all subsets regression application, we were able to determine the most efficient model that presumably satisfies the multiple linear regression conditions except for independence, which is a limitation of this model that will be discussed later.

From the model we selected, we can conclude that the students' distance from a 4-year college will affect people's years of education in populations similar to our study. Distance can serve as a natural experiment in a variety of applications (Rouse C.E, 1980). As the distance increased, the number of years of education for a student decreased. This can also apply in today's world when many people are disconnected from higher

education because of the lack of physical educational institutions (Victoria Rosenboom & Kristin Blagg, 2018) in the local area.

The results also demonstrated that other background factors like ethnicity, family income, whether that individual's mother went to college, and whether that individual's father went to college are also significant contributors to educational achievement. If policymakers aim to improve people's education rate/levels, it is helpful to provide more educational accessibility to some college-desert areas and provide supportive schemes for students with low family incomes. The focus should also be on racial equality in the educational system.

Limitations of model and future work

As discussed earlier in the paper, this model as well as this dataset have several limitations that can affect the credibility of the results. First of all, we employed a categorical variable as the main response variable.

When the response variable is categorical, a standard linear regression model can't be used, but we can use logistic regression models instead. Furthermore, the data was collected in a nested structure, which leads to dependence among the observations. In this case, one method is to use mixed-effects models, which allow both fixed and random effects, and are used when there is a non-independence in the data. The data was collected through a survey, a convenient method, which cannot ensure the credibility of the data. For future work, data collectors can choose populations randomly from random areas to achieve impartial results.

Rouse, C.E. (1995). Democratization or Diversion? The Effect of Community Colleges on Educational Attainment. Journal of Business \& Economic Statistics, 12, 217–224.

Victoria, Rosenboom & Kristin, Blagg (2018). Three million Americans are disconnected from higher education. Urban Institute. https://www.urban.org/urban-wire/three-million-americans-are- disconnected-higher-education.

Stock, J.H. and Watson, M.W. (2007). Introduction to Econometrics, 2nd ed. Boston: Addison Wesley.

Zoe Chao, Catherine Paredes Amaya, Robin Tran. (2022). CollegeDistance Project.