Important Factors Contribute to U.S High School Drop out Rate

Professor: Wei-Yin Loh

TA: Xiaomao Li

Name: Qiuying Li

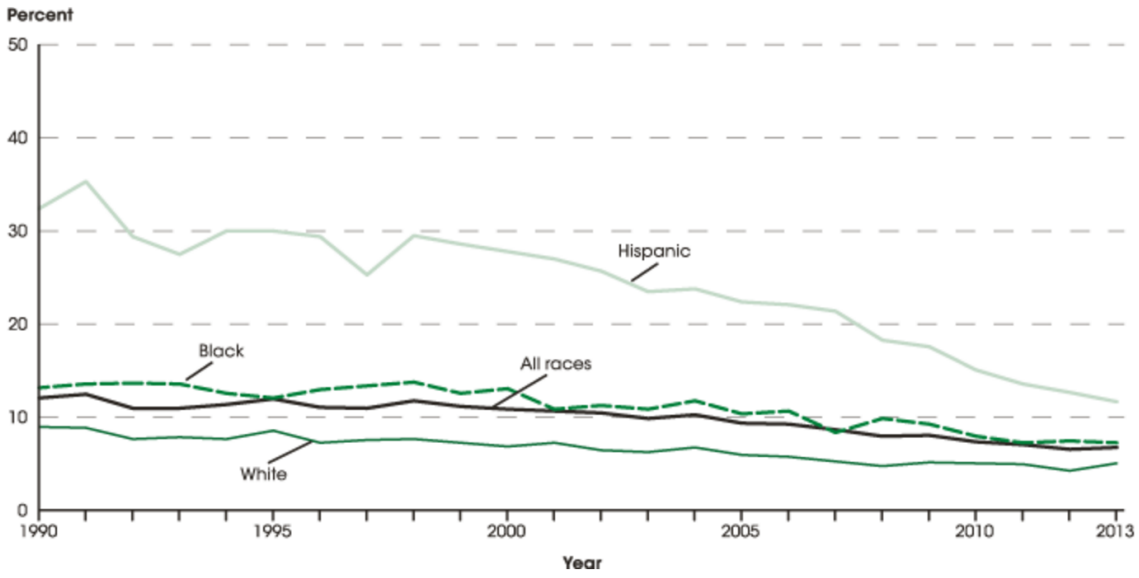
November 8, 2015

-

Introduction

Even though school completion rates have continually grown during much of past 100 years, dropping out of school persists as a problem that interferes with educational system efficiency and the most straightforward and satisfying route to individual educational goals for young people.

The status dropout rate represents the percentage of 16- through 24-year-olds who are not enrolled in school and have not earned a high school credential (either a diploma or an equivalency credential such as a General Educational Development.



Status dropout rates of 16- through 24-year-olds, by race/ethnicity: 1990 through 2013

From the grace above, we can see that the ethnicity plays an important part in the high school drop out rates. Hispanic has higher percent than the average high school drop out rates, while white has lower high school drop out rates compared to the average. Besides the ethic, it would be interesting to explore what other factors play significant roles in high school drop pout rates?

Data Description

The data set contains graduation data joined with the maximum overlapping Census data. In addition, the data set includes information about every Census tract that overlaps each school district. This is a huge data set, which contains more than 500 variables, and near 10,000 values for each variable. All the variables can be categorized into four parts, population variables, household variables, operation variables, and calculated percentages.

Research Questions

We already understand ethnicity is one of the reasons students don't make it through high school. However, what are the other factors affecting graduation rates that we don't know about?

Income, family dynamics, and pregnancy?

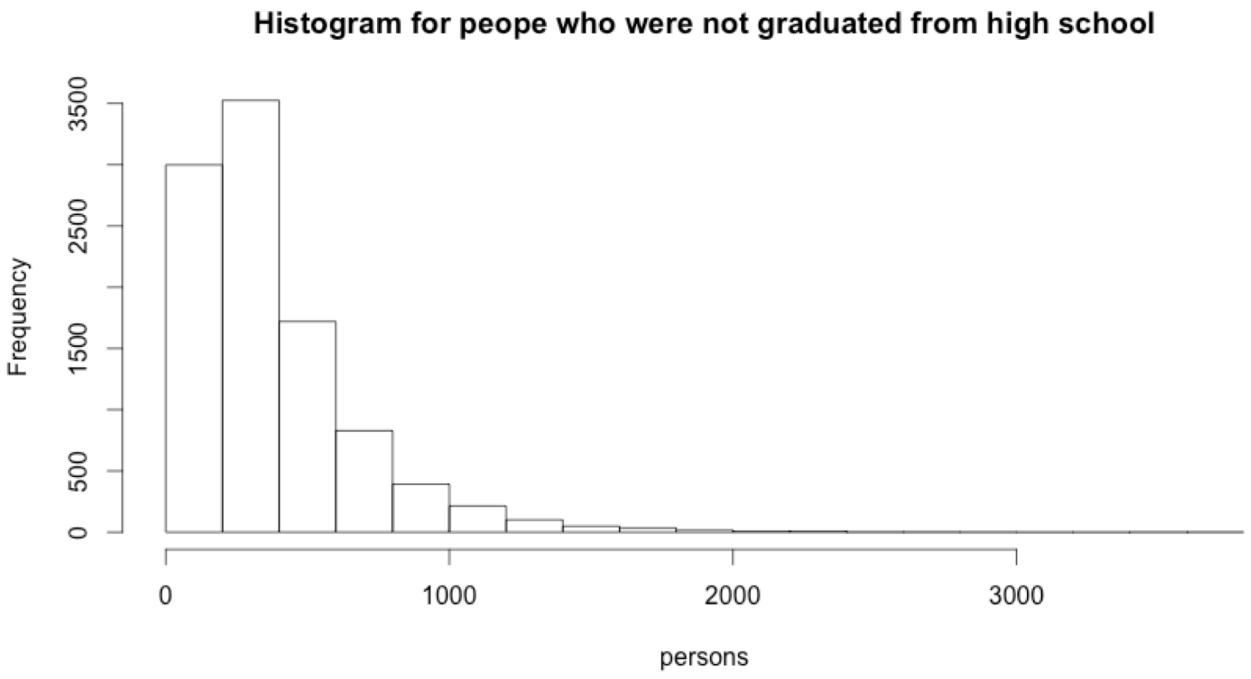
Does bullying play a part? Crime?

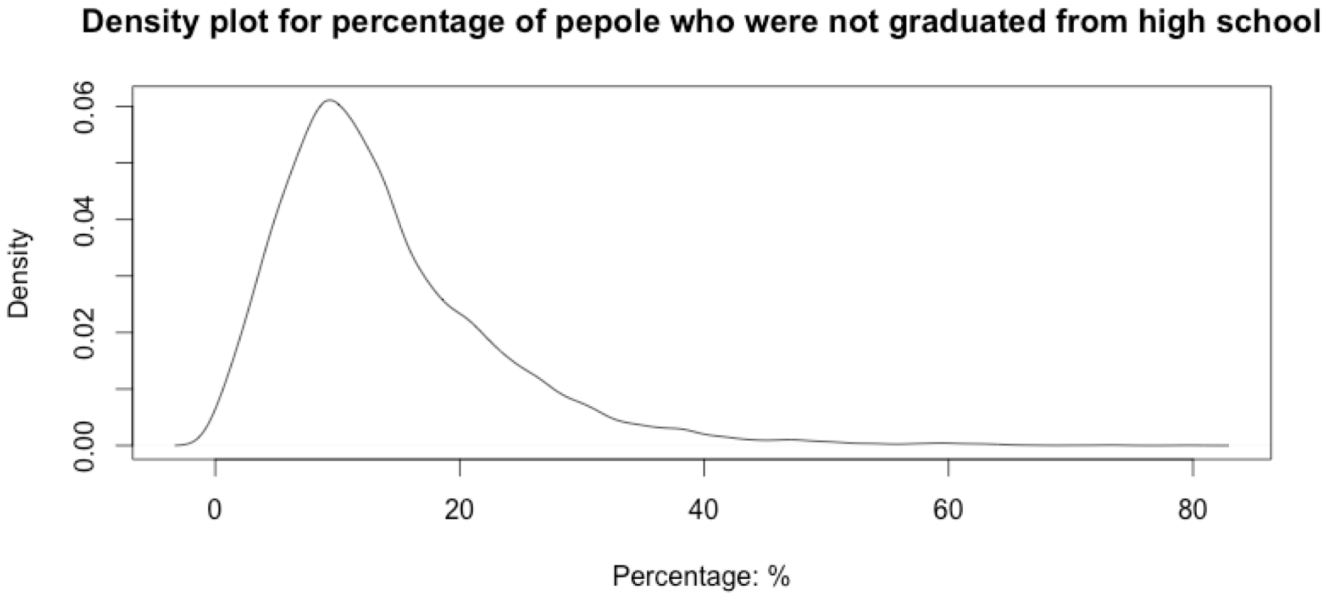
What about local gas prices & transportation, or the layout of a city?

Hopefully, we would be able to predict the high school graduate rate based on the model we have.

Response Variable Selection

Considering that the research is about High school dropout rates, so I choose Not_HS_Grad_ACS_08_12(Number of people 25 years old and over who are not high school graduates) as my response variable.





The mean of people who were not graduated from high school is 376,6, and the percentage for people who were not graduated from high school is 13.89%. However, from the histogram and the density plot, we can see the Not_HS_Grad_ACS_08_12 is heavily skewed.

Predictor Variables Selection

1. Deleted Variables

Even though there are more than 500 variables in the data set, some of them are not useful.

a. I deleted All the Margin of Error variables;

b. I deleted all the calculated percentage variables.

c. I deleted Repeated Variables: for some variables, both 2010 Census and ASC provide values for them, so two different variables actually measure the same thing, which means one of variable is repeated.

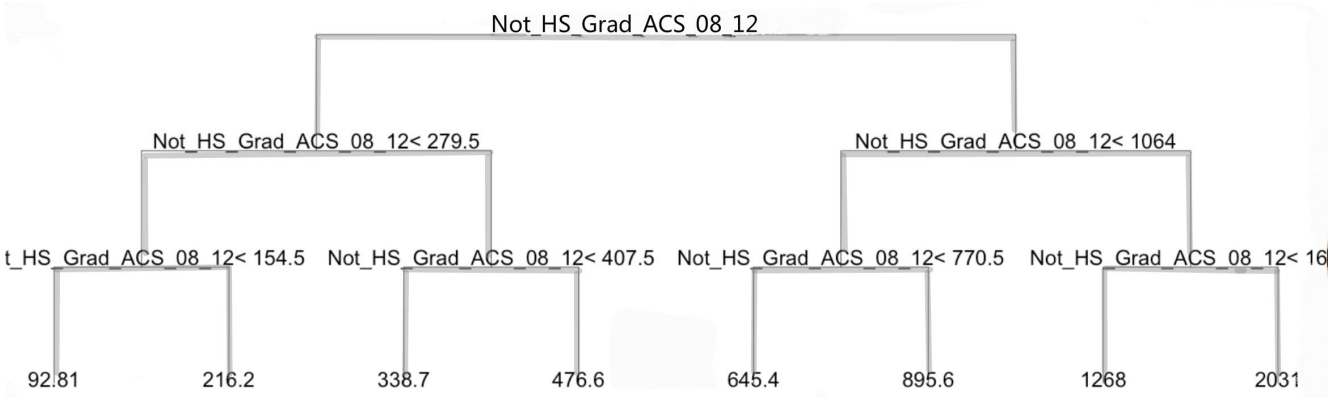
d. I deleted all unrelated population variables. The research question is about the high school dropout rate, so I deleted the variables for 5-year old children, and people who are older than 24 years old.

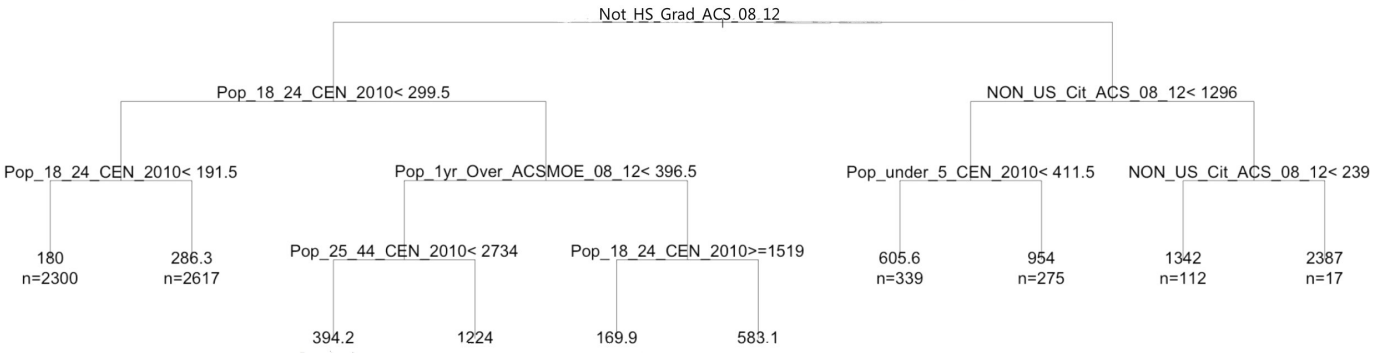
2. Rank Variable importance

After deleting all the unimportant variables, there were still more than 100 variables left. So it is essential to rank the importance of the predictor variables by using Rpart.

Rpart output:

Regression Trees






```
Regression tree:
rpart(formula = drop ~ ., data = data1, method = "anova", control = rpart.control(minsplit = 20,
    cp = 0.01, maxcompete = 4, maxsurrogate = 5, usesurrogate = 2,
    xval = 10, surrogatestyle = 0, maxdepth = 30))

Variables actually used in tree construction:
[1] Not_HS_Grad_ACS_08_12

Root node error: 908137212/9904 = 91694

n=9904 (3 observations deleted due to missingness)

        CP nsplit rel error   xerror      xstd
1 0.609592      0  1.000000 1.000079 0.0333451
2 0.144758      1  0.390408 0.391753 0.0183725
3 0.110917      2  0.245650 0.248105 0.0095669
4 0.032768      3  0.134733 0.135595 0.0094352
5 0.025020      4  0.101965 0.103933 0.0049821
6 0.019130      5  0.076945 0.078986 0.0048673
7 0.016276      6  0.057815 0.057795 0.0048602
8 0.010000      7  0.041539 0.043095 0.0048505
```

```
Variable importance
     Not_HS_Grad_ACS_08_12 Aggregate_HH_INC_ACS_08_12  Prs_Blw_Pov_Lev_ACS_08_12
                        35                         33                         8
         Hispanic_CEN_2010      ENG_VW_SPAN_ACS_08_12     Crowd_Occp_U_ACS_08_12
                         7                          6                         4
      NON_US_Cit_ACS_08_12      Female_No_HB_CEN_2010          ENG_VW_ACS_08_12
                         2                          1                         1
  Not_MrdCple_HHD_CEN_2010     Mobile_Homes_ACS_08_12
                         1                          1
```

Summary of the Rpart output:

The analysis of the Rpart shows the regression trees of the high school drop out rate. In addition, the R-part rank the 10 of the important variables among all of the other variables. The specific information of the these 10 variables showed in the table below:

Important Variables		
Not_HS_Grad_ACS_08_12	Drop	Number of people 25 years old and over who are not high school graduates
Aggregate_HH_INC_ACS_08_12	X1	Sum of all incomes in the household
Prs_Blw_Pov_Lev_ACS_08_12	X2	Number of people classified as below the poverty level in the ACS
Hispanic_CEN_2010	X3	Persons of Hispanic Origin in the 2010 Census
ENG_VW_SPAN_ACS_08_12	X4	Spanish - Households where no one 14 and over speaks English only or speaks English very well in the ACS
Crowd_Occp_U_ACS_08_12	X5	Occupied Units with more than 1.01 persons per room in the ACS
NON_US_Cit_ACS_08_12	X6	Persons who are not US citizens in the ACS
Female_No_HB_C	X7	Households with a female householder, no husband present in the 2010 Census
ENG_VW_ACS_08_12	X8	. Households where no one 14 and over speaks English only or speaks English very well in the ACS
Not_MrdCple_HHD_CEN_2010	X9	Households with no Married Couple present in the 2010 Census
Mobile_Homes_A	X10	Mobile Homes in the ACS

As we have the 10 important variables, it is essential to fit these 10 variables into a multiple linear model, to see which variable is a significant part to the high school dropout rate.

3.Multiple Linear Regression

a. fit the multiple linear model

```
lm(formula = drop ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10, data = data00)

Residuals:
     Min       1Q   Median       3Q      Max
-1087.98   -88.44   -18.08    69.57  2059.33

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 68.4010052  5.5518766  12.320  < 2e-16 ***
x1           0.0012421  0.0006579   1.888   0.0591 .
x2           0.1602848  0.0061733  25.964  < 2e-16 ***
x3           0.0561973  0.0042180  13.323  < 2e-16 ***
x4           0.7752407  0.0782954   9.901  < 2e-16 ***
x5           0.3704688  0.0496029   7.469 8.78e-14 ***
x6           0.1814936  0.0118066  15.372  < 2e-16 ***
x7           0.7247226  0.0297508  24.360  < 2e-16 ***
x8          -0.3261924  0.0660766  -4.937 8.08e-07 ***
x9          -0.0567218  0.0086076  -6.590 4.63e-11 ***
x10          0.3389354  0.0082243  41.212  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 171.7 on 9893 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.679,     Adjusted R-squared:  0.6787
F-statistic:  2093 on 10 and 9893 DF,  p-value: < 2.2e-16
```

b. Step-wise : model selection

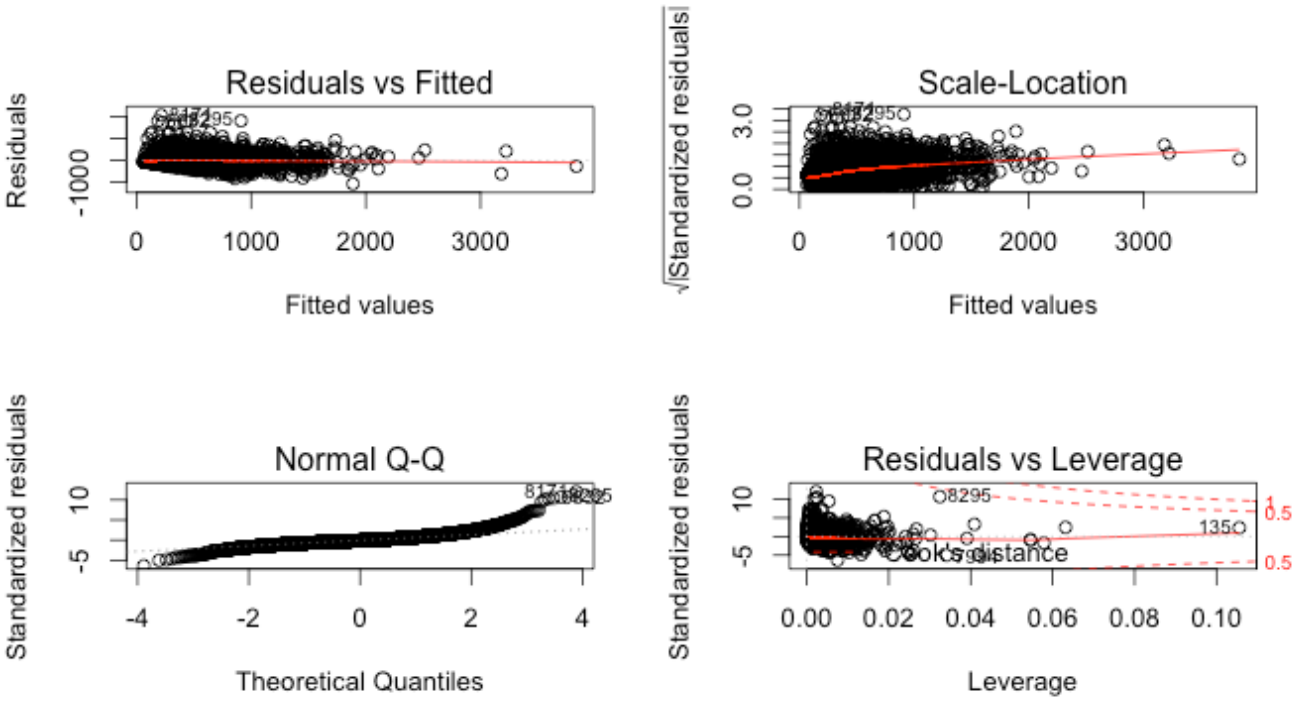
```
Start:  AIC=101933.3
drop ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10

       Df Sum of Sq       RSS    AIC
<none>              291515661 101933
- x1    1    105037 291620697 101935
- x8    1    718102 292233763 101956
- x9    1   1279591 292795252 101975
- x5    1   1643704 293159365 101987
- x4    1   2888910 294404571 102029
- x3    1   5230514 296746174 102107
- x6    1   6963135 298478795 102165
- x7    1  17485627 309001287 102508
- x2    1  19865013 311380673 102584
- x10   1  50046439 341562100 103500
> fit12$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
drop ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10

Final Model:
drop ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10


  Step Df Deviance Resid. Df Resid. Dev      AIC
1                       9893  291515661 101933.3
```

c. Summary of the multiple linear regression

From the results above, we can see p-values for all the 10 variables are significant if we choose a=0.01 as significant level, which means the slope for the predictor variables are not zero. In addition, the multiple R-square is 0.67, which means 67% of variation in response variable: high school dropout rate, can be explained by these 10 variables.

More importantly, we used stepwise method in both forward and outward direction to select model, and the final model which has the smallest AIC value, is same to the initial model. So we can conclude that the all the 10 variables play important part to explain the variation of the high school drop out rates.

d. Model checking:



- Residuals in the residual plot are **not** scatter around the 0-line, which indicates that independence assumption can not be held.
- QQ plot shows a heavy tails on the two side, which means data set is not perfectly normal

e. Model Interpretation

The model we have is below:

High_school_dropout = **68.4** + **0.012*(**Aggregate_HH_INC_ACS_08_12) + **0.16*** **(**Prs_Blw_Pov_Lev_ACS_08_12)+ **0.056*(**Hispanic_CEN_2010**) + 0.775+** ***(** ENG_VW_SPAN_ACS_08_12) + **0.37* (**Crowd_Occp_U_ACS_08_12) + **0.18** * (NON_US_Cit_ACS_08_12) + **0.724*** (Female_No_HB_CEN_2010) - **0.237** * (ENG_VW_ACS_08_12) - **0.057** * (Not_MrdCple_HHD_CEN_2010)+ **0.339** * (Mobile_Homes_A)

Variable Interpretation

1. **Aggregate_HH_INC_ACS_08_12**: The regression coefficient associated with high school dropout rate is 0.012suggesting that each one unit increase in is associated with a 0.056 unit increase in high school drop out rate.

2. Prs_Blw_Pov_Lev_ACS_08_12 : The regression coefficient associated with high school dropout rate is 0.16, suggesting that each one unit increase in **poverty** is associated with a 0.16 unit increase in high school drop out rate.

3. **Hispanic_CEN_2010**: The regression coefficient associated with high school dropout rate is **0.056**, suggesting that each one person increase in **Hispanic Origin** is associated with a **0.056** unit increase in high school drop out rate.

4. **ENG_VW_SPAN_ACS_08_12:**The regression coefficient associated with high school dropout rate is **0.775**, suggesting that each one person increase in **Spanish household** is associated with a 0.775 unit increase in high school drop out rate.

5. Crowd_Occp_U_ACS_08_12: The regression coefficient associated with high school dropout rate is **0.37**, suggesting that each one unit increase in **Occupied Units with more than 1.01 persons per room** is associated with a **0.37** unit increase in high school drop out rate.

6. NON_US_Cit_ACS_08_12: The regression coefficient associated with high school dropout rate is **0.18**, suggesting that each one unit increase **Persons who are not US citizens** in is associated with a **0.18** unit increase in high school drop out rate.

7. Female_No_HB_CEN_2010: The regression coefficient associated with high school dropout rate is **0.724**, suggesting that each one unit increase **Households with a female householder who has no husband** in is associated with a **0.724** unit increase in high school drop out rate.

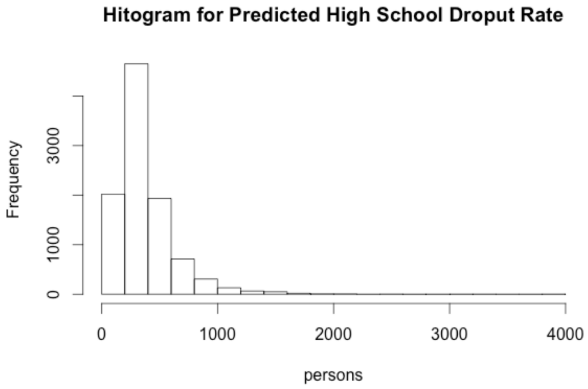
8. ENG_VW_ACS_08_12: The regression coefficient associated with high school dropout rate is **-0.237**, suggesting that each one unit increase in **Households speak English very well** in is associated with a **0.237** unit decrease in high school drop out rate.

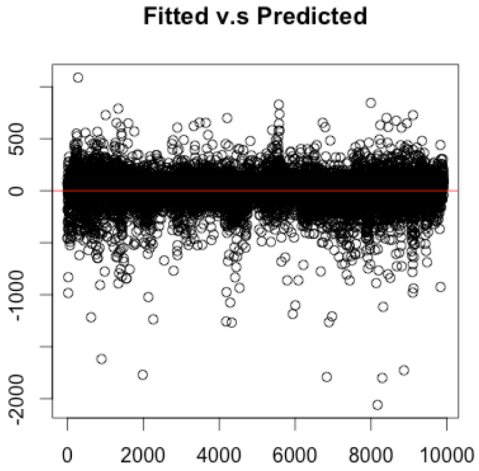
9. Not_MrdCple_HHD_CEN_2010: The regression coefficient associated with high school dropout rate is **-0.057**, suggesting that each one unit increase in **Households with no Married Couple** in is associated with a **0.057** unit decrease in high school drop out rate.

10. Mobile_Homes_A: The regression coefficient associated with high school dropout rate is **0.339**, suggesting that each one unit increase in Mobile Homes in the ACS in is associated with a **0.339** unit increase in high school drop out rate

4. Prediction

```
summary(prediction)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
63.99  215.60  308.80  376.60  457.40 3835.00       3
```





The mean value of prediction is very close to the current data, both are around the 376 persons. Both of the High school drop out rate are near 20%. In addition, we can notice that most of the points in fitted value v.s predicted value are around the 0 line, which means the prediction is reliable.

Conclusion :

From all the analysis above, we can see that the average high school drop out rate is near 20%. Except from the ethnicity, for example, there would be a high drop out rates if the kids were not born in U.S, or the kids from Spanish family. There are other important factors contribute to high school dropout, such as economic factors, family education background and stability of the family.

In term of the economic factors, if a family has low sum income, or the family lives below the poverty line, the high school dropout rate is higher.

Besides the economic issue, the family education background is also very important. If family members can speak English well, then the high school drop out rates is lower.

More importantly, stability is very essential to a family. If family without a husband, or the family tend to move a lot, then there will be higher high school dropout.