**Financial and Other Barriers to FAFSA Completion**

**Prepared by Matthew Wilkinson**

**December 19, 2023**

**EXECUTIVE SUMMARY**

The primary goal of this project is to identify potential financial and other barriers to Free Application for Federal Student Aid (FAFSA) completion. This exploratory analysis indicates communities with a higher rate of education, higher poverty, lower unemployment, and demographics associate with FAFSA completion rates among high schoolers. In brief, the statistical model presented provides a glimpse into the complexities of the financial and other barriers that affect FAFSA completion. Although not all high school graduates decide to attend college, it is important to recognize that barriers to FAFSA completion exist and to develop methods to minimize their impact.

**BACKGROUND**

The Free Application for Federal Student Aid (FAFSA) allows students to apply for need-based aid from the federal government to reduce the college cost of attendance. Nonetheless, there are barriers students encounter when completing the form such as a lack of support from parents, financial costs, incorrect assumptions of ineligibility, and other factors which may ultimately deter students from completing the FAFSA or applying to college. Therefore, this report examines the relationships between measures of socioeconomic health and community demographics to identify potential barriers to FAFSA completion. It is hypothesized that higher levels of education, greater poverty, and higher unemployment will relate to greater FAFSA completion rates. Exploratory analyses will examine associations between race and population counts with FAFSA completion rates.

**METHODOLOGY**

Data is aggregated at the geographic level of county among data sets provided by the US Department of Education, US Census Bureau, and National Center for Education Statistics. That is, each of the data sets were joined on the unique value of state and county for this report.

First, the US Department of Education provided data on **FAFSA completion counts** by school which serves as the primary measure of interest. FAFSA completions are as of May 1, 2022. Next, the U.S. Census Bureau’s American Community Survey collects information on educational attainment, school enrollment, demographics, and other topics and makes the data available across multiple tables. Data is aggregated across five years to provide reliable population estimates for regions such as county with populations below 65,000 people. The tables used in this report provide measures of socioeconomic health and general demographics: **poverty rate**, **unemployment rate**, **household median income**, **educational attainment** which refers to the percentage of adults with a bachelor's degree or higher, **population counts**, and **race** which was measured as the percentage of the population that is white non-Hispanic. Each of these variables are deemed relevant as measures of socioeconomic health and key community characteristics, and thus fit the project goals.

Furthermore, the National Center for Education Statistics was referenced to estimate **grade 12 enrollment counts** to calculate the percentage of FAFSA completions among eligible high schoolers. Additionally, StatsAmerica, a tool provided by Indiana University, was used to develop a city-to-county crosswalk by school as listed in the US Department of Education’s FAFSA data set. City, town, and other places do not match always perfectly to county. Therefore, when a city overlapped with multiple counties, the county with the greatest percentage of overlap with the city boundaries was entered as the county. This process led to most cities being linked with a county. Any remaining cities without a county were determined via data provided from the National Center of Education Statistics and ChatGPT with that data being entered manually into the crosswalk spreadsheet.

After data cleaning, a total of 3,026 US counties are included in the data set. Counties with fewer than 10 completed FAFSA forms were removed from the data set to minimize the potential unobservable confound of counties with low interest in higher education impacting the interpretation of results. Additionally, three counties had FAFSA completion percentages greater than 100.0%. As this is expectedly due to grade 12 enrollment estimation error, these counties were capped at 100.0% for FAFSA completion. Lastly, log-transformation of population counts and income was done to reduce to impact of outliers in the regression analysis.

All data sets including the final data set as well as the python code to prepare the data are available on a GitHub repository: <https://github.com/mwlknsn/fafsa_2022.git>

See Appendix C for source code and notes regarding data creation and cleaning.

**RESULTS**

On average, 46.8% (*SD* = 13.6%) of grade 12 high school students by county completed the FAFSA.

Bivariate correlations were then done to determine which predictors to enter in the regression model. Poverty rate (r = -0.04), educational attainment (r = 0.12), percentage of white non-Hispanic (r = 0.04), log-transformed household median income (r = 0.03), log-transformed adult population (r = -0.08), and unemployment rate (r = -0.09) were all significantly related to FAFSA completion rates, *p*s < 0.05. Additionally, household income was strongly correlated with poverty rate (r = -0.85) and educational attainment (r = 0.71). To reduce the potential for multicollinearity, household income was not included in the final regression model. See Appendix A for the correlation matrix.

The final regression model included first-order terms educational attainment, percentage of white non-Hispanic, poverty rate, population counts, and unemployment rate and the second-order term for percentage of white non-Hispanic. The model explained approximately 5.0% of the variance in FAFSA completion rates. This low amount of explained variance indicates further work is necessary to identify predictors of FAFSA completion. Nonetheless, as educational attainment and poverty increase so does FAFSA completion, *p*s < 0.001. Also, as unemployment and total population increase FAFSA completion rates decrease, *p*s < 0.05. Furthermore, as the percentage of white non-Hispanic increases the FAFSA completion rate decreases until a completion rate of about 70% at which point as the percentage of white non-Hispanic increases the FAFSA completion rate decreases, *p*s < 0.001. See Appendix B for the regression table.

**CONCLUSION**

In conclusion, the hypothesis was supported by the data and exploratory analyses yield areas for further research. FAFSA completion rates are generally higher in communities with greater financial need, however, differences in completion according to race and ethnicity are worrisome and may indicate inequity in access or other barriers. There are a few important observations to note following these analyses. FAFSA completion rates increase when educational attainment and poverty rates increase for counties. The positive relationship between the percentage of adults with a bachelor's degree or higher and FAFSA completion may indicate these communities are able to better support students on completing the FAFSA or be an area with greater interest in higher education. The positive relationship between poverty rates and FAFSA completion also indicate students in need are completing the FAFSA form, however, the weak association may be viewed as troublesome as not enough students may be seeking out this resource. Second, the negative association between unemployment and FAFSA completion is interesting and may indicate family hardship may deter high school students from completing the FAFSA. Lastly, the curvilinear relationship between the percentage of white non-Hispanic in a community and FAFSA completion is worrisome as this may indicate inequity based on race and ethnicity and FAFSA completion. More research is necessary to further probe this relationship and identify the barriers to FAFSA completion in these communities.

There are some limitations that should be noted about this data set and analysis. First, 12th grade enrollment data is estimated using the most recently available NCES data: 2021-22 public school and 2019-20 private school data. Other estimation techniques for 12th grade enrollments are recommended to validate FAFSA completion percentages such as comparing to estimated 12th grade counts provided by the US Census Bureau or using 12th grade enrollment counts projected using multiple years and other grade levels. Second, as noted throughout the report, not all high school graduates intend to pursue college. A follow-up analysis should seek to estimate non-college bound students and remove them from estimated 12th grade enrollment counts for calculating FAFSA completion percentages. Finally, the model explains a low amount of variance. Further exploration of variables that might explain FAFSA completion is recommended to improve model performance and create action plans based on this information.

**APPENDIX A: CORRELATION MATRIX**

A screenshot of a computer screen

Description automatically generated

**APPENDIX B: REGRESSION ANALYSIS**

A screenshot of a computer

Description automatically generated

**APPEDNIX C: PYTHON CODE**

#need to install packages to read files, work with the data, and conduct regression analysis

!pip install xlrd==2.0.1

!pip install openpyxl

!pip install statsmodels==0.14.1

import pandas as pd

import numpy as np

import xlrd as xlrd

import matplotlib.pyplot as plt

import seaborn as sns

import statsmodels.api as sm

#packages ready

#need to load the primary file (fafsa data) and then clean

df = pd.read\_excel('HSARCHIVE04302023.xls', skiprows = 1, header = 2)

df.drop(df.columns[[4, 5, 6, 8, 9, 10, 11]], axis = 1, inplace = True)

df.columns = ['school\_code', 'school\_name', 'school\_city', 'school\_state', 'completed\_fafsa']

df['school\_city\_state'] = df['school\_city'] + ', ' + df['school\_state']

df = df[~df['school\_state'].str.contains('PR|MP|AS|FM|GU|MH|PW|VI|FC')]

df['completed\_fafsa'] = df['completed\_fafsa'].replace('<5', 1)

df['completed\_fafsa'] = df['completed\_fafsa'].astype(int)

df\_crosswalk = pd.read\_excel('school\_city\_to\_county\_crosswalk.xls', skiprows = 0, header = 0)

df\_crosswalk.columns = ['fips', 'city\_state', 'school\_code']

df\_fafsa = pd.merge(df, df\_crosswalk, on = 'school\_code', how = 'left')

df\_fafsa = df\_fafsa.groupby('fips')['completed\_fafsa'].sum().reset\_index()

df\_fafsa

#primary file is loaded and cleaned, i.e., variables renamed, US territories removed from the data set, county assigned to school, and fafsa completion data summed by county; for schools with "<5" completions this value was changed to 1 in the data set

#need to get estimated grade 12 enrollments by county to calculate fafsa completion percentages as the outcome variable

df\_elsi = pd.read\_excel('ELSI\_Schools.xlsx', skiprows = 0, header = 1)

df\_elsi.columns = ['school\_name', 'school\_state\_name', 'school\_code', 'fips', 'grade\_12\_enrollment']

df\_elsi = df\_elsi.groupby('fips')['grade\_12\_enrollment'].sum().reset\_index()

df\_fafsa\_and\_nces = pd.merge(df\_fafsa, df\_elsi, on = 'fips', how = 'left')

df\_fafsa\_and\_nces['completed\_fafsa\_perc'] = df\_fafsa\_and\_nces['completed\_fafsa'] / df\_fafsa\_and\_nces['grade\_12\_enrollment'] \* 100

df\_fafsa\_and\_nces['completed\_fafsa\_perc'] = df\_fafsa\_and\_nces['completed\_fafsa\_perc'].round(4)

df\_fafsa\_and\_nces = df\_fafsa\_and\_nces[df\_fafsa\_and\_nces['completed\_fafsa'] >= 10]

df\_fafsa\_and\_nces.loc[df\_fafsa\_and\_nces['fips'].isin([17151, 19119, 31089]), 'completed\_fafsa\_perc'] = 100

df\_fafsa\_and\_nces

#school enrollment data is joined to fafsa data set and fafsa completion percentages calculated. three counties had percentages greater than 100.0%: 17151 16/10 = 160%; 19119 145/150 = 102%; 31089 94/92 = 102%; percentages capped at 100%

#further, a decision is made to drop counties with fewer than 10 completed fafsas to minimize potential estimation error

#now that fafsa data is compiled, socioeconomic health and demographic data needs to be joined to the data set

#data from four files from the US Census Bureau are joined in

df\_education = pd.read\_excel('Education.xlsx', skiprows = 1, header = 2)

df\_education = df\_education.drop(columns = df\_education.columns[3:54])

df\_education.columns = ['fips', 'state', 'county', 'edu\_attainment']

df\_pov = pd.read\_excel('PovertyEstimates.xlsx', skiprows = 1, header = 3)

df\_pov = df\_pov.drop(columns = df\_pov.columns[1:7])

df\_pov = df\_pov.drop(columns = df\_pov.columns[2:4])

df\_pov = df\_pov.drop(columns = df\_pov.columns[3:17])

df\_pov = df\_pov.drop(columns = df\_pov.columns[4:14])

df\_pov.columns = ['fips', 'total\_poverty', 'poverty\_rate', 'hh\_med\_inc']

df\_pov['hh\_med\_inc\_log'] = np.log(df\_pov['hh\_med\_inc']+1).round(4)

df\_unemp = pd.read\_excel('Unemployment.xlsx', skiprows = 1, header = 3)

df\_unemp = df\_unemp.drop(columns = df\_unemp.columns[1:93])

df\_unemp = df\_unemp.drop(columns = df\_unemp.columns[2:8])

df\_unemp.columns = ['fips', 'unemp\_rate']

df\_race = pd.read\_csv('B03002 Race.csv', skiprows = 0, header = 1)

df\_race = df\_race.drop(columns = df\_race.columns[1])

df\_race = df\_race.drop(columns = df\_race.columns[2:5])

df\_race = df\_race.drop(columns = df\_race.columns[3:1000])

df\_race.columns = ['fips', 'total\_pop', 'white\_nonhisp']

df\_race['white\_nonhisp\_perc'] = (df\_race['white\_nonhisp'] / df\_race['total\_pop'] \* 100).round(4)

df\_race['total\_pop\_log'] = np.log(df\_race['total\_pop']+1).round(4)

df\_census = pd.merge(pd.merge(pd.merge(df\_education, df\_pov, on = 'fips', how = 'left'), df\_unemp, on = 'fips', how = 'left'), df\_race, on = 'fips', how = 'left')

df\_fafsa\_census = pd.merge(df\_fafsa\_and\_nces, df\_census, on = 'fips', how = 'left')

df\_fafsa\_census['grade\_12\_enrollment\_log'] = np.log(df\_fafsa\_census['grade\_12\_enrollment']+1).round(4)

df\_fafsa\_census['white\_nonhisp\_perc\_sq'] = df\_fafsa\_census['white\_nonhisp\_perc'] \* df\_fafsa\_census['white\_nonhisp\_perc']

df\_fafsa\_census

#the final data set is ready after joining US Census data

#need to perform initial exploratory analyses, i.e., bivariate correlations

df\_correl = pd.DataFrame(df\_fafsa\_census)

correl\_matrix = df\_correl[['completed\_fafsa\_perc', 'edu\_attainment', 'total\_pop\_log', 'white\_nonhisp\_perc', 'poverty\_rate', 'hh\_med\_inc\_log', 'unemp\_rate']].corr()

plt.figure(figsize = (8, 6))

sns.heatmap(correl\_matrix, annot = True, cmap = 'coolwarm', fmt = '.2f')

plt.title('Bivariate Correlations')

plt.show()

#potential mulitcollinearity when examining associations between income and both poverty and educational attainment, thus, income not to be included in regression

#need to develop a model to examine associations between socioeconomic health indicators, demographics, and fafsa completion rate

predictors = ['edu\_attainment', 'total\_pop\_log', 'white\_nonhisp\_perc', 'poverty\_rate', 'unemp\_rate', 'white\_nonhisp\_perc\_sq']

target = 'completed\_fafsa\_perc'

X = df\_fafsa\_census[predictors]

y = df\_fafsa\_census[target]

X = sm.add\_constant(X)

model = sm.OLS(y, X).fit()

print(model.summary())

#ultimately, a model with first-order terms and a second-order race and ethnicity term was obtained; model performs poorly in terms of explained variance indicating more work to be done in terms of identifying factors that relate to fafsa completion rates

#want to report some basic statisics on the primary measure of interest: FAFSA completion rate

fafsa\_mean = df\_fafsa\_census['completed\_fafsa\_perc'].mean()

fafsa\_mean

fafsa\_stdev = df\_fafsa\_census['completed\_fafsa\_perc'].std()

fafsa\_stdev

#calculations of the mean and standard deviation for fafsa completion percentage