ESP Analysis Draft

Haojia Li, Zihan Yang, Tianran Zhang

June 23, 2020

Menu

[Data Preprocessing 1](#_Toc43778000)

[Empire State Poll (ESP) 1](#_Toc43778001)

[PIAAC 7](#_Toc43778002)

[Exploration among ESP Variables 9](#_Toc43778003)

[Summary 9](#_Toc43778004)

[Numeracy and Covariates 9](#_Toc43778005)

[Numaracy and Health Status 10](#_Toc43778006)

[Numeracy and Care Provider Visits 12](#_Toc43778007)

[Health Status and Covariates 12](#_Toc43778008)

[Regression Models 14](#_Toc43778009)

[Modeling Framework 14](#_Toc43778010)

[Results 14](#_Toc43778011)

## Data Preprocessing

### Empire State Poll (ESP)

#### Data Description

**Study Population**

New York State residents aged 18 and over.

**Data Source**

Empire State Poll (ESP) 2019, a random-digit-dial telephone survey conducted by the Survey Research from February to April in 2019 that covers 800 completed interviews. 400 (50%) of the participants were from downstate, and the rest were from upstate.

**Study Variables**

1. Outcomes

* JAq4 - self-reported health status
* 1 = Excellent
* 2 = Very Good
* 3 = Good
* 4 = Fair
* 5 = Poor
* RVq2 - visits to care providers in the past 12 months
* 1 = Never
* 2 = 1-2 times
* 3 = Once every few months
* 4 = Once a month
* 5 = Twice a month
* 6 = Once a week
* 7 = 2-3 times a week

1. Predictors

* SNS-3: 3-item version of the Subjective Numeracy Scale (SNS), which was validated having good internal reliability and correlated highly with the full SNS.
* JAq1 - rate competence related to fractions (1 = not good at all to 6 = extremely good)
* JAq2 - rate competence related to percentages (1 = not good at all to 6 = extremely good)
* JAq3 - rate usefulness related to numerical information (1 = never to 6 = very often)

1. Covariates

* Demographical variables
* gender 1=Male, 2=Female, 3=Other, -1=Refused
* age 18-20, 21-44, 45-64, 65+, -1=Refused
* educ 1-3=“High school or less”, 4-5=“Some college or associate’s degree”, 6-7=“Bachelor’s degree or higher”, -1=Refused
* hhinc household income in 2018 before taxes ($): 1-5=“<50k”, 6-7=“50k-100k”, 8=“100k-150k”, 9=“150k+”, NA=“No Data”
* race White, Black, Natam, Asian, Other, More than 1 race, -1=Refused
* hisp ethnicity: 0=Non-Hispanic, 1=Hispanic, -1=Refused

**Table 1: Demographical characteristics of ESP**

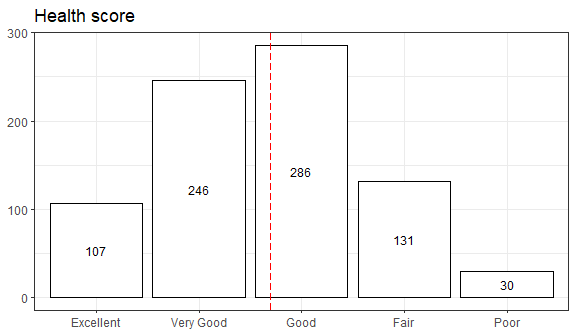
|  |  |
| --- | --- |
| Demographics | N(%) |
| Gender |  |
| Male | 403(50.38%) |
| Female | 393(49.12%) |
| Other | 2(0.25%) |
| Refused | 2(0.25%) |
| Age |  |
| 18-20 | 42(5.25%) |
| 21-44 | 298(37.25%) |
| 45-64 | 286(35.75%) |
| 65+ | 168(21%) |
| Refused | 6(0.75%) |
| Education |  |
| High school or less | 184(23%) |
| Some college or associate’s degree | 224(28%) |
| Bachelor’s degree or higher | 390(48.75%) |
| Refused | 2(0.25%) |
| Income |  |
| <$50k | 253(31.62%) |
| $50k-$100k | 262(32.75%) |
| $100k-$150k | 112(14%) |
| $150k+ | 143(17.88%) |
| No Data | 30(3.75%) |
| Race |  |
| White | 543(67.88%) |
| Black | 103(12.88%) |
| Asian | 44(5.5%) |
| Natam | 10(1.25%) |
| More than 1 race | 38(4.75%) |
| Other | 51(6.38%) |
| Refused | 11(1.38%) |
| Ethnicity |  |
| Hispanic | 126(15.75%) |
| Non-hispanic | 671(83.88%) |
| Refused | 3(0.38%) |

**Distribution of Self-Reported Health Status Scores**

The distribution of self-reported health status is right-skewed. The median is 3 (“Good”), and more than 75% of the interviewees had health status equal to or better than “Good”.

**Distribution of health status**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0% | 25% | 50% | 75% | 100% |
| 1 | 2 | 3 | 3 | 5 |

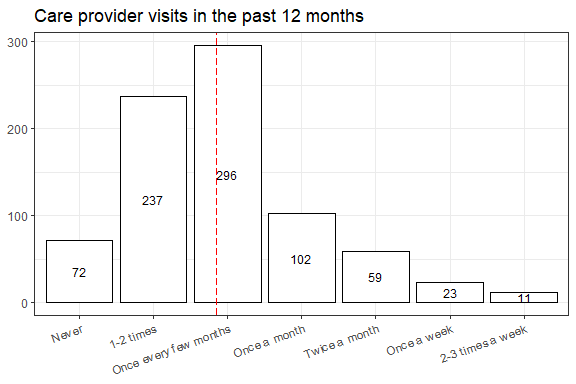


**Distribution of Healthcare Provider Visits**

The distribution of healthcare provider visits is right-skewed. And the median is 3 (“Once every few months”), and more than 75% of the interviewees had the visit frequency equal to or less than “Once every few months”.

**Distribution of care provider visits**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0% | 25% | 50% | 75% | 100% |
| 1 | 2 | 3 | 3 | 7 |



**Distribution of Self-Reported Health Status by Healthcare Provider Visits**

According to the heatmap, the top 4 cells of care provider visits and health status are: (Once every few months, Good), (Once every few months, Very Good), (1-2 times, Good), and (1-2 times, Very Good). This result is consistent with the previous results of the distribution of health status or care provider visits alone.

Horizontally reading the plot, we infer that people with excellent health status tend to have a lower frequency than the other groups and that people under poor health condition will visit care providers more than others. However, especially for the poor group, we are not 100% confident to draw such a conclusion since the number of interviewees involved in this subgroup was limited. On the other hand, based on the distribution of health status in each subgroup of visits, there is no obvious difference among the first 6 groups, but only people who visit care providers the most frequently (2-3 times a week) are more likely to have not a good health condition.

A close up of a piece of paper

Description automatically generated

#### Data Preprocessing for Statistical Analysis

**Refusal Rate**

As the table show below, the refusal rate of race and age are higher than the other variables.

**Summary of refusal rate**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Race | Age | Ethnicity | Gender | Education | JAq1 | JAq2 | JAq3 |
| Number | 11.00 | 6.00 | 3.000 | 2.00 | 2.00 | 2.00 | 3.000 | 3.000 |
| % | 1.38 | 0.75 | 0.375 | 0.25 | 0.25 | 0.25 | 0.375 | 0.375 |

**Calculation of Numeracy Score**

The responding scores to the three SNS-3 questions (JAq1, JAq2, and JAq3) has been added together serving as a continuous predictor named numeracy score for statistical analysis.

Though almost all answers of these three questions have “correct”, there are a few “incorrect” answers with negative score value -1 (Refused) and -2 (Don’t know). Specifically, there are 2 subjects had more than half (2 of the 3) wrong answers, they are dropped from the dataset. And there are 11 subjects have only 1 (1 of the 3) wrong answer, and such negative values will be imputed using 4.74, the average score of all correct answers of the all three questions.

**Demographical Comparison Table: New York State vs. ESP**

This comparison aims to compare the sample distribution of ESP with the population of New York State (NYS), which was collected from tables of [United States Census Bureau](https://www.census.gov). Because the data of each demographical characteristic was collected from different tables, the total count number (in thousand) of each characteristic were pretty close but may not exactly equal to each other.

The original ESP data has 800 observations, there are 763 observations for further analysis after dropping missing values collapsing race into 3 categories (White, Black, and Other).

To calculate the weights in the regression analysis, Fisher’s test is firstly performed to evaluate the difference between ESP sample and NYS population. Based on the table below, those demographical characteristics with significant p-values are worth adding weights (the proportion of each subgroup in NYS) into the regression models.

Briefly, between ESP and NYS:

* There is no significant difference in gender, age, and ethnicity between.
* There is a significant difference in education, income, and race.
* Subjects with higher education/income levels tend to have a higher will to be interviewed.
* The proportion of white people was closed to that in NYS. The reason why there is a significant difference is that ESP has a lower percentage of black people involved in.

**Comparison table between NYS and ESP**

|  |  |  |  |
| --- | --- | --- | --- |
| Demographics | NYS | ESP | pval |
| Gender |  |  | 0.157 |
| Male | 121301(48.41%) | 389(50.98%) |  |
| Female | 129262(51.59%) | 374(49.02%) |  |
| Age |  |  | 0.136 |
| 18-20 | 11851(4.73%) | 41(5.37%) |  |
| 21-44 | 103469(41.29%) | 288(37.75%) |  |
| 45-64 | 82455(32.91%) | 274(35.91%) |  |
| 65+ | 52788(21.07%) | 160(20.97%) |  |
| Education |  |  | <0.001 |
| High school or less | 97506(38.92%) | 173(22.67%) |  |
| Some college or associate’s degree | 69577(27.77%) | 216(28.31%) |  |
| Bachelor’s degree or higher | 83478(33.32%) | 374(49.02%) |  |
| Income |  |  | <0.001 |
| <$50k | 51303(33.03%) | 252(33.03%) |  |
| $50k-$100k | 64932(41.8%) | 259(33.94%) |  |
| $100k-$150k | 19158(12.33%) | 109(14.29%) |  |
| $150k+ | 19930(12.83%) | 143(18.74%) |  |
| Race |  |  | <0.001 |
| White | 13559(69.7%) | 522(68.41%) |  |
| Black | 3424(17.6%) | 98(12.84%) |  |
| Other | 2471(12.7%) | 143(18.74%) |  |
| Ethnicity |  |  | 0.494 |
| Hispanic | 41217(16.47%) | 118(15.47%) |  |
| Non-hispanic | 209046(83.53%) | 645(84.53%) |  |
|  |  |  |  |

### PIAAC

The PIAAC U.S. 2017 public-use file contains individual unit data, including both responses to the background questionnaire and the cognitive assessment from the third U.S. PIAAC data collection.

Originally, there were 3660 observations in the PIAAC dataset. After dropping the missing values, we finally got 2609 complete observations. The variables of interest are shown below.

* Outcome: health\_score I\_Q08
* Predictor: numeracy PVNUM, literacy PVLIT, problem\_solving PVPSL
* Covariates: education B\_Q01A, income J\_Q09USX, race RACETHN\_5CAT

Variables including health\_score, education, income were recoded in the same way as ESP. We also collapsed the levels of the initiate race into both hisp and race. The comparison table between ESP and PIAAC is as below. All the common covariates in both datasets have significantly different distributions except race.

* ESP interviewees had a higher education/income level.
* The percentage of Hispanics in ESP was higher than PIAAC.

**Comparison table between ESP and PIAAC**

|  |  |  |  |
| --- | --- | --- | --- |
| Demographics | ESP | PIAAC | pval |
| Education |  |  | 0.001 |
| High school or less | 173(22.67%) | 1106(42.39%) |  |
| Some college or associate’s degree | 216(28.31%) | 508(19.47%) |  |
| Bachelor’s degree or higher | 374(49.02%) | 995(38.14%) |  |
| Income |  |  | <0.001 |
| <$50k | 252(33.03%) | 984(37.72%) |  |
| $50k-$100k | 259(33.94%) | 889(34.07%) |  |
| $100k-$150k | 109(14.29%) | 411(15.75%) |  |
| $150k+ | 143(18.74%) | 325(12.46%) |  |
| Race |  |  | 0.9 |
| White | 522(68.41%) | 1800(68.99%) |  |
| Black | 98(12.84%) | 319(12.23%) |  |
| Other | 143(18.74%) | 490(18.78%) |  |
| Ethnicity |  |  | <0.001 |
| Hispanic | 118(15.47%) | 281(10.77%) |  |
| Non-hispanic | 645(84.53%) | 2328(89.23%) |  |

## Exploration among ESP Variables

### Summary

**Numeracy Score**

* The distributions of numeracy are significantly different in the subgroups of all covariates. (education, gender race, age categories, income categories).
* The distributions of numeracy are significantly different in the subgroups of original health status (5 levels), as well as its dichotomized variables health status (good/ not good) and health status (not poor/ poor). Numeracy score and health status have positive significant correlation no matter using “not-poor/poor” (corr = 0.136, p-value < 0.001) or “good/not good” (corr = 0.175, p-value < 0.001) as the health status outcome variable.
* The distributions of numeracy scores are not significantly different in the subgroups of healthcare visiting (often/not often).

**Health Status**

* Health status is always positively associated with education and income no matter in which way we dichotomize the outcome (good or not poor as cut point). That is, people with a higher education and income levels tend to have a better self-reported health condition. Moreover, age and ethnicity are associated with health status (good/not good) but not associated with health status (not poor/poor).

**Healthcare Visits**

* All the covariates have significant effects on the proportion of healthcare visiting often(education, gender race, age categories, income categories).

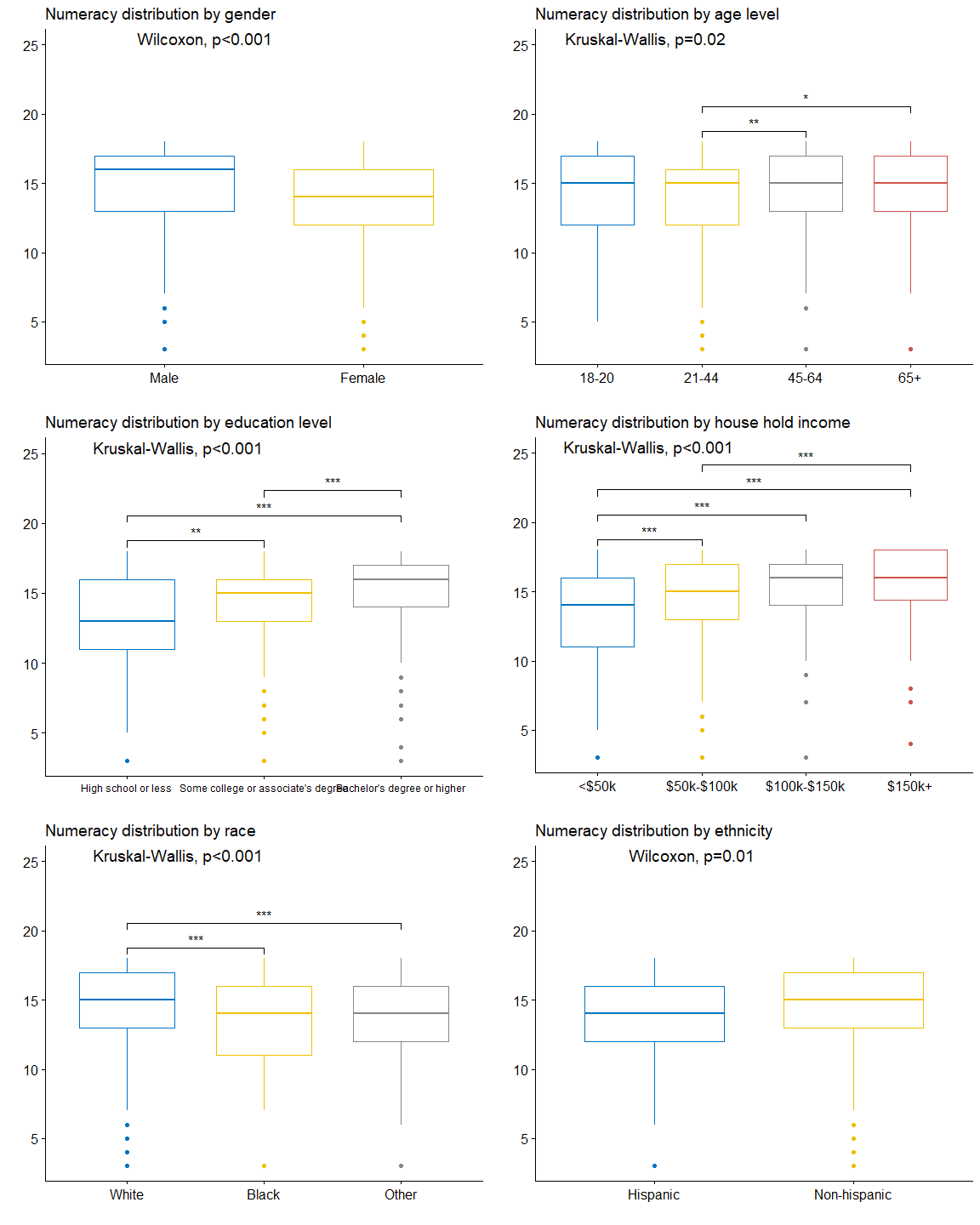
### Numeracy and Covariates

To explore the relationship between the predictor numeracy and the covariates, boxplots are drawn, showing the distribution of numeracy in each demographical subgroup. Since the distribution of numeracy avoids the normality assumption, appropriate non-parametric has been performed to test if there is a significant difference across the groups.

Based on the number of categories, we consider two situations:

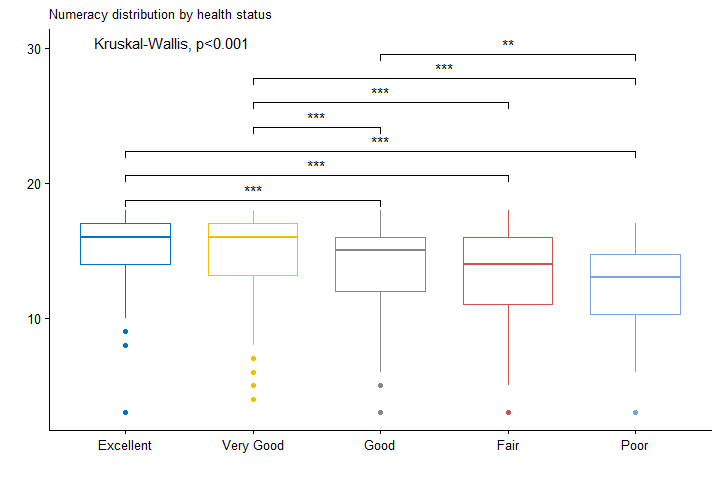
* Covariates with two categories, including gender and ethnicity: We use the Wilcoxon test.
* Covariates with more than two categories, including age, education, income, race: We use Kruskal-Wallis test. Once the global p-value is less than 0.05, which indicates the existence of a significant difference, we will do a pairwise comparison using the Wilcoxon test, and mark the pairs with significant differences (Holm-Bonferroni adjusted p-value < 0.05) in the plot.

According to the plots, there are significant differences in the distribution of numeracy stratified by each covariate.



### Numaracy and Health Status

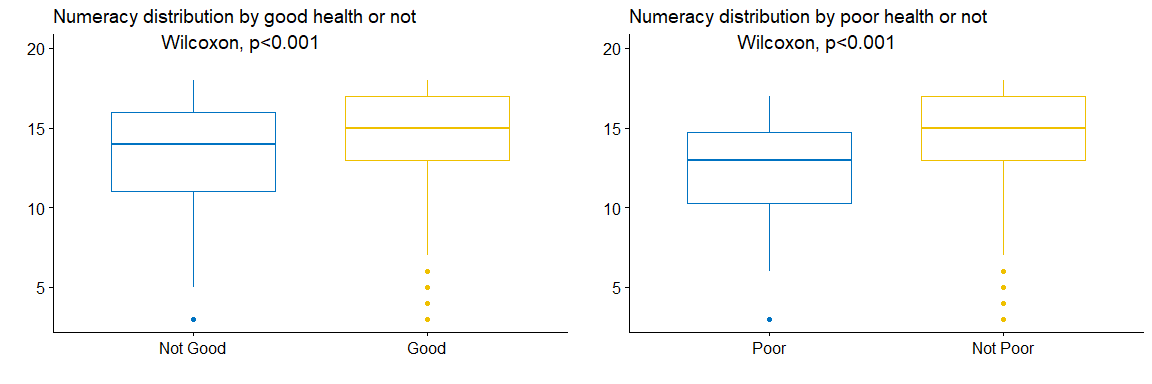
The global Kruskal-Wallis p-value less than 0.001 shows that there is significant difference in numeracy across the subgroups in health status. According to the result of pairwise comparison, 7 out of 10 pairs have significant differences. We find that two adjacent subgroups are less likely to, but the pair of “Very Good” and “Good” do have a significant difference in numeracy.



Due to power concerns, we dichotomize health\_status in two ways:

* health\_status\_good
* Good = Excellent, Very Good, Good
* Not Good = Fair, Poor
* health\_status\_poor
* Not Poor = Excellent, Very Good, Good, Fair
* Poor = Poor

No matter splitting health status into Not-Good/Good or Poor/Not-Poor, there is strong evidence to reject the null hypothesis that the distributions of numeracy are the same in the subgroups. And the associations are both significant that people in better health status tend to have a higher numeracy level.



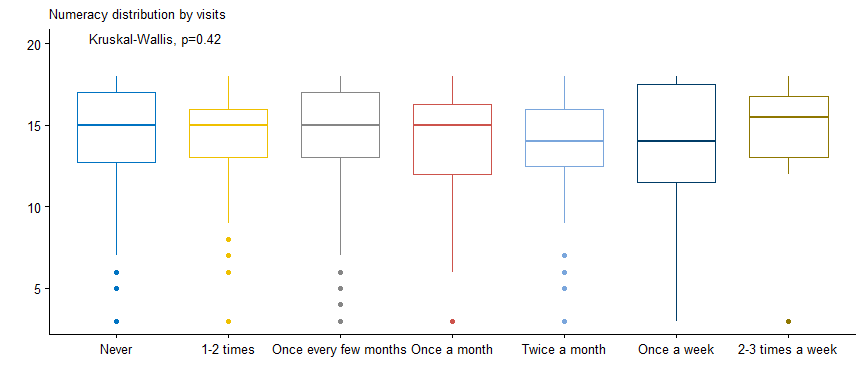
**Correlation between health status and numeracy**

|  |  |  |
| --- | --- | --- |
| Health Status | Point-Biserial Corr | P-value |
| Poor/Not Poor | 0.136 | <0.001 |
| Not Good/Good | 0.175 | <0.001 |

### Numeracy and Healthcare Provider Visits

Due to power concerns, we dichotomize visits into once every few months or more frequently as “Often” and others as “Not Often”.

There is no evidence to reject the null hypothesis that the distribution of numeracy scores are the same among the visits subgroups or between visits often and not-often group.



**Correlation between care provider visits and numeracy**

|  |  |  |
| --- | --- | --- |
| Care Provider Visits | Point-Biserial Corr | P-value |
| Often/Not Often | 0.004 | 0.912 |

### Health Status and Covariates

Since all the covariates are categorical variables, we will use the Chi-squared test to check if there are significant differences in health\_status across different subgroups, including gender, age, education, income, race and hisp separately. What’s more, polychoric rho will be calculated to evaluate the correlation between the outcome and the covariates.

#### Good/Not Good

The polychoric rho is negatively associated with Chi-squared p-value. education and income are the two variables sharing rather high correlation with health\_status\_good than the others with p-values less than 0.001 at the two-sided significance level of 0.05. age and ethnicity have a lower but still significant correlation. At the same time, gender and race don’t show significant relationships with the binary outcome.

**Chi-squared test and polychoric rho between health status good and covarites**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Gender | Age | Education | Income | Race | Ethnicity |
| Polycor | 0.041 | -0.146 | 0.383 | 0.319 | -0.133 | 0.208 |
| Chisq pval | 0.525 | 0.025 | <0.001 | <0.001 | 0.057 | 0.004 |
| Signif |  | \* | \*\*\* | \*\*\* |  | \*\* |

#### Not Poor/Poor

Education and income are still significantly correlated with health\_status\_poor, and the correlations are even higher. But the rest of the covariates do not show any significant relationship with the outcome.

**Chi-squared test and polychoric rho between health status poor and covarites**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Gender | Age | Education | Income | Race | Ethnicity |
| Polycor | -0.142 | -0.202 | 0.444 | 0.415 | -0.019 | -0.002 |
| Chisq pval | 0.194 | 0.101 | <0.001 | <0.001 | 0.079 | 0.991 |
| Signif |  |  | \*\*\* | \*\*\* |  |  |

#### Healthcare Provider Visits and Covariates

All the covariates have significant correlations with visits\_often.

**Chi-squared test and polychoric rho between often visits and covarites**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Gender | Age | Education | Income | Race | Ethnicity |
| Polycor | 0.169 | 0.195 | 0.187 | 0.177 | -0.208 | 0.237 |
| Chisq pval | 0.003 | 0.001 | 0.002 | 0.004 | 0.001 | <0.001 |
| Signif | \*\* | \*\* | \*\* | \*\* | \*\* | \*\*\* |

## 

## Regression Models

### Modeling Framework

Since all the outcomes, including health status and visits to care providers in ESP as well as health status in PIAAC, are dichotomized variables, we will use logistic regression to evaluate the effect of the predictors on the outcomes, after controlling demographical covariates of interest. Steps developing our model are:

* Decide the variables in the initial model by doing univariate regression. Variables with significant (p-value < 0.05) odds ratio will be selected.
* Conduct stepwise model selection with both directions and arrive at the final model.
* Evaluate the performance of the final model, including AUC, power, and prediction error after randomly divide the dataset into training and test data by a proportion of 7:3. (PIAAC analysis ends)
* Based on the comparison table between ESP and NYS, education, income, and race are three variables that have different distributions in ESP data and the NYS dataset. Add weights to adjust the proportion of these three variables and refit the multi-regression model.
* Compare the performance of the unweighted and weighted final model, including AUC, power, and prediction error. (ESP analysis ends)

### Results

#### Numeracy

The original health status has 5 levels after dropping all unknowns: Excellent/Very good/Good/Fair/Poor. Taking clinical meaning into consideration, we dichotomize the outcome in two ways: Good(611 individuals)/Not Good(152); Poor(737)/Not Poor(26). Since the situation is quite unbalanced in the second one, and the regression result shows that numeracy is not a significant predictor, we choose the first dichotomizing method health\_status\_good(Good/Not Good).

**1. ESP**

Univariate regression results show that 6 variables are significant covariates of people’s health status: numeracy, age, education, income, race, and ethnicity. For multivariate regressions, a stepwise selection method was conducted to choose the final model. Numeracy has a significant effect on the health\_status\_good with a power of 81% after controlling education, age, and income. This model has an AUC of 72% and a prediction error of 0.219.

Weights are applied to education and income since their distributions are significantly different from that of NYS data, and they are both significant covariates of health status. The weighted regression results are quite similar to unweighted results: Numeracy has a significant effect on the health\_status\_good with a power of 57.4%, after controlling education, age, income, and ethnicity. This model has an AUC of 72% and a prediction error of 0.219.

The regression summary table for health\_status\_good in the ESP dataset is shown in the table below.

**ESP’s multivariates regression on health status: Good vs Not Good**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Weight | Performance | | | Predictor | Covariates | | | | | |
|  |  | AUC | Power | Error | numeracy | gender | age | educ | income | race | hisp |
| 1 |  | 0.724 | 0.81 | 0.219 | \* |  | \* | \* | \* |  |  |
| 2 | educ  income | 0.72 | 0.574 | 0.219 | \* |  | \* | \* | \* |  | \* |

**2. PIAAC**

A stepwise selection method is conducted to choose the multivariate regression model. Numeracy has a significant effect on the health\_status\_good with a power of 75% after controlling education and income. The final model has an AUC of 67.6% and a prediction error rate of 14.8%.

Additional stepwise regression is conducted with predictors, literacy and problem\_solving. Regression result shows that numeracy and problem solving have significant effects on the health\_status\_good with a power of 77% after controlling education and income. This model has an AUC of 69% and a prediction error of 14.8%.

The regression summary table for health\_status\_good in the PIAAC dataset is shown in the table below.

**PIAAC’s multivariates regression on health status: Good vs. Not Good**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Performance | | | Predictor | Covariates | | | | Additional Pred | |
|  | AUC | Power | Error | numeracy | educ | income | race | hisp | prob.solv | literacy |
| 1 | 0.676 | 0.75 | 0.1475 | \* | \* | \* |  |  | / | / |
| 2 | 0.69 | 0.77 | 0.1475 | \* | \* | \* |  |  | \* |  |

#### Healthcare Provider Visits

The outcome visits is dichotomized as visits\_often: “Never” or “1-2 times” as “Not Often” and others as “Often”. Univariate regression results show that 6 variables are significant covariates of visits\_often: gender, age, education, income, race, and ethnicity. Numeracy is not a significant predictor for visits\_often. For multivariate regressions, a stepwise selection method is conducted to choose the final model. Numeracy does not have a significant effect on visits\_often, although gender, age, income, and ethnicity are significant covariates. The final model has an AUC of 64.3% and a prediction error of 34.97%.

Additional multinomial regression is conducted after adjusting the distribution of education and income with weights. Numeracy seems not significantly associated with visits\_often, while gender, age, education, and income are significant covariates. The final model has an AUC of 64.6% and a prediction error of 37.25%.

The regression summary table for visits is shown in the table below.

**ESP’s multivariates regression on visits to care providers: Often vs. Not Often**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Weight | Performance | |  | Predictor | Covariates | | | | |
|  |  | AUC | Power | Error | numeracy | gender | age | educ | income | race | hisp |
| 1 |  | 0.643 | / | 0.3497 |  | \* | \* |  | \* |  | \* |
| 2 | educ income | 0.636 | / | 0.3725 |  | \* | \* | \* | \* |  |  |