Dacanay, Jordan

Lansangan, Romand

Ramilo, Zion

**Analyzing Reproductive Health and Behavior Patterns Using NSFG 2022–2023**

DSC 1105 Summative Assessment 2

May 20, 2025

3rd Year Applied Mathematics (Data Science)

# Introduction

This report analyzes reproductive health and behavior patterns among female respondents using data from the National Survey of Family Growth (NSFG) 2022-2023. The goal of the said report is analyzing relationship and nature of (selected) variables inside the named dataset. Using different statistical techniques like Posson regression, contingency tables and categorical response modeling, the modelers was able to understand and articulate the complexity of the relationship between different behavioral patterns among female respondents.

# Dataset Description

The data for this analysis is drawn from the female respondent subset of the National Survey of Family Growth (NSFG). The NSFG is a national United States based representational survey from National Center for Health Statistics (2024) done with 5586 female respondents to collect 1912 behaviors with regards to their lives, marriages, divorces, pregnancies, contraception, and general health. The dataset along with the male counterpart can be found on the website of [U.S Centers for Disease Control and Prevention](https://www.cdc.gov/nchs/nsfg/nsfg-2022-2023-puf.htm).

The specific variables used will differ for each part of the report.

# Statistical Analyses

Variables to be used:

|  |  |  |  |
| --- | --- | --- | --- |
| Original Variable | Description | Type | New Name |
| OPPLIFENUM | Number of opposite-sex partners in lifetime for all types of sex (top-coded). | Count Variable | sex\_partners |
| GRFSTSX | Grade Respondent Was in at First Sexual Intercourse (bottom-coded). | Ordered categorical | first\_sex |
| MANRASDU | Man Respondent thinks of as raised her during teens | Nominal  categorical | man\_guardian |
| WOMRASDU | Woman Respondent thinks of as raised her during teens. | Nominal  categorical | woman\_guardian |
| LMARSTAT | Respondent's legal marital status relative to opposite-sex spouses. | Demographics | marital\_status |

Table 1.a: Variables to be used for statistical analyses

A screenshot of a computer code

AI-generated content may be incorrect.

Table 1.b: General Information on the Cleaned Data

Table 1.b shows the general information of the dataset after cleaning where it can be observed that we have the variables of the number of sex partners, the stage where they had any sexual experience, what mother or father figure they have in their household, and lastly the marital status of the respondent.

Table 1.b also shows that there are a total of five thousand five hundred eighty-six observations within the data where it can be seen that there are no null entries which implies that our dataset is complete and does not suffer from missing data issues, thereby ensuring the reliability of any statistical analysis.

Within this data, it is determined that the number of sex partners shall be the dependent variable being evaluated within our statistical analysis, while other variables that are identified to have a data type of object shall become our independent variables.

A white paper with black text

AI-generated content may be incorrect.

Figure 1: Unique Values per Variable

Figure 1 shows the unique values per variable where it can be observed that within our independent variable, it is composed mostly of discrete values. First-sex experience has thirteen unique values, Male and Female guardian has six unique variables, and lastly, Marital Status has seven unique values.

A graph with blue squares

AI-generated content may be incorrect.

Figure 2: Histogram Plot of Sex Partners

The plot shows that the number of Sex Partners takes on a distribution that is skewed to the right, which implies most of the values present within the data set take a smaller value. In addition, the data has values like 999 or 998 which was removed out of the data set under the grounds that this was suspicious enough of an outlier that would constitute its removal.

Figure 2 shows that the data does not pass the assumption of normality that is needed for powerful statistical analysis. In this case, the transformation of the data set shall be applied. The statisticians shall be using the Yeo-Johnson transformation to address the presence of zero values within the data and ensure that our data is primed for statistical analyses.

A graph with blue and purple squares

AI-generated content may be incorrect.

Figure 3: Bar plot of the frequency of the First Sex Experience

Figure 3 shows the distribution of the frequency of the observations for each category of the first sex experience, where within the figure it is shown that more observations were recorded that it was inapplicable for them, with grade ten being the second most, and eleventh grade being the third most. Furthermore, it has been identified that there are less respondents of the category of sophomore college and above.

A graph with blue squares

AI-generated content may be incorrect.A graph with blue squares

AI-generated content may be incorrect.

Figure 4: Frequency of Male Guardian Categories (Left) and Frequency of Female Guardian Categories (Right)

Figure 4 shows the distribution of the respondents about the presence of male (left) and female (right) guardians. Where Figure 4 shows that on both male and female guardians most of the respondents have both of their presents present within their lives whilst, for the male guardians the least refused to say, and the least for the female guardian shows that there is a low proportion of respondents either don’t know or refused to say.

A graph with different colored squares

AI-generated content may be incorrect.

Figure 5: Box plot of First Sex Experience by the number of Sex Partners

Figure 5 shows the distribution of the number of sex partners for every single category of first sex experience, whereas derived from Figure 1 most of the distribution lies within lower values, which can be seen by their interquartile ranges of each category. Furthermore, the distribution for each different category shows that there is much variation where some data is more spread than others.

Figure 5 also shows that there are a lot of outliers for each first-sex experience category which can be a big source that can cause type II or type I errors such that there would be a large probability that it will produce a false positive or false negative. Moreover, another that would be a cause for type I and II errors is the few samples that are considered for each group, this is an issue as ANOVA requires a certain threshold for each between-group value to enable the model to perform accurately. To mitigate this the statisticians reduced the data into the following: “College”, “Senior Highschool”, “Highschool”, “Middle School”, “Inapplicable”, “Refused”, and “Don’t know”. In addition to this the observations that were identified to take on a value of either “Refused” or “Don’t know” were removed due to the few observations which would not be ideal for one-way ANOVA testing.

A graph with different colored squares

AI-generated content may be incorrect.A graph with colorful squares and text

AI-generated content may be incorrect.

Figure 6: Box plot of Male Guardian Categories (Left) and Box plot of Female Guardian Categories (Right) by the number of Sex Partners

Figure 6 shows the distribution of the respondents within each variable for male and female guardians where it can be seen that there is still a presence of outliers within each value, where four of the categorical values for both the male and female guardians have outliers. In addition, similar to the distribution for the first-sex experience, it shows a distribution that leans more toward lower values.

Within Figure 6 it has been identified that there are categorical values that have fewer observations which would not be ideal for one-way ANOVA. In order to mitigate this problem, these observations were removed from the main dataset.

A close-up of numbers

AI-generated content may be incorrect.

Table 2: One-Way ANOVA Test for First-Sex Experience

A one-way ANOVA was conducted to examine the effect of first-sex experience on the dependent variable. There was a statistically significant effect of group, F(5, 5251) = 361.96, p < .001.

A screenshot of a computer screen

AI-generated content may be incorrect.

Table 3: Multiply Comparisons of Means for First-Sex Experience with p-value adjustments Tukey's Honestly Significant Difference (HSD) controlling Family-wise error rate (FWER)

Post-hoc comparisons were used to explore the pair-wise mean difference within groups of the variable through Tukey’s HSD test indicated that several categories of first sex experience differed significantly in their mean scores which are the following:

* College scores were significantly higher than high school (mean difference = 0.30, *p* = .007).
* College scores were significantly lower than Inapplicable (mean difference = –0.58, *p* < .001).
* Highschool differed significantly from Inapplicable, Not-school, and Senior Highschool (all *p* < .001).
* Inapplicable had significantly higher scores than Middle School & Below and Senior Highschool.
* Not-school was significantly different from multiple groups, including College and Middle School & Below.

A number and numbers on a white background

AI-generated content may be incorrect.

Table 4: One-Way ANOVA Test for Male Guardian

A one-way ANOVA was conducted to examine the effect of the presence of a male guardian on the dependent variable. There was a statistically significant effect of group, F(4, 5265) = 10.011252, p = 4.537223e-08.

A screenshot of a computer

AI-generated content may be incorrect.

Table 5: Multiply Comparisons of Means for the presence of Male Guardian with p-value adjustments Tukey's Honestly Significant Difference (HSD) controlling Family-wise error rate (FWER)

Post-hoc comparisons were used to explore the pair-wise mean difference within groups of the variable through Tukey’s HSD test indicating that there were only select categories of presence of Male Guardian that differed significantly in their mean scores which are the following:

* Respondents with both parents had significantly higher scores than the rest of the groups except for those groups that were identified to have another father figure.

A close-up of numbers

AI-generated content may be incorrect.

Table 6: One-Way ANOVA Test for Female Guardian

A one-way ANOVA was conducted to examine the effect of the presence of a female guardian on the dependent variable. There was a statistically significant effect of group, F(3, 5269) = 11.76112, p = 1.122049e-07.

A black and white paper with numbers and lines

AI-generated content may be incorrect.

Table 7: Multiply Comparisons of Means for the presence of a Female Guardian with p-value adjustments Tukey's Honestly Significant Difference (HSD) controlling Family-wise error rate (FWER)

Post-hoc comparisons were used to explore the pair-wise mean difference within groups of the variable through Tukey’s HSD test indicating that there were only select categories of the presence of a female guardian that differed significantly in their mean scores which are the following:

* Respondents with both parents had significantly higher scores than the rest of the groups except for those groups that were identified to have no mother figure.

A number and a number

AI-generated content may be incorrect.

Table 8: One-Way ANOVA Test for Marital Status

A one-way ANOVA was conducted to examine the effect of the marital status of the respondents on the dependent variable. There was a statistically significant effect of group, F(4, 5262) = 98.232524, p = 9.466236e-81.

A screenshot of a computer screen

AI-generated content may be incorrect.

Table 9: Multiply Comparisons of Means for Marital Status with p-value adjustments Tukey's Honestly Significant Difference (HSD) controlling Family-wise error rate (FWER)

Post-hoc comparisons were used to explore the pair-wise mean difference within groups of the variable through Tukey’s HSD test indicating that several categories of marital status differed significantly in their mean scores which are the following:

* Never-married groups have a higher mean score of sexual partners than those who were identified to be separated and widowed.
* Groups that are married have a lower mean score of sexual partners than those who were identified to have never married and separated.
* Groups that are divorced or annulled have a lower mean score of sexual partners than those who were identified to have never married and married.

To reduce the risk of false positives due to multiple pairwise comparisons, we used Tukey’s Honestly Significant Difference (HSD) test which controls the family-wise error rate at 0.05, in order to ensure the robustness of the statistical test, ensuring that the overall probability of creating Type I errors are controlled, leading to more trustworthy conclusions and interpretations.

# Poisson Regression

| **Variable Name** | **Description** |
| --- | --- |
| OPPLIFENUM | Number of opposite-sex partners in lifetime for all types of sex (computed in FC J-14d) (top-coded) |
| **Predictors:** |  |
| MARSTAT | AD-7b R's marital or cohabiting status |
| ONOWN | AG-0a (before age 18) R ever live away from parents/guardians? |
| LVSIT14F | AG-3 female parent (figure) living with at age 14 - fam not intact thru 18 |
| LVSIT14M | AG-4 male parent (figure) living with at age 14 - fam not intact thru 18 |
| MENARCHE | BA-1 Age at first menstrual period (bottom-coded) |
| GRFSTSX | CE-8 Grade R Was in at First Sexual Intercourse (bottom-coded) |
| WHOFSTPR | CG-3 Who Was Rs First Sexual Partner |
| RELIGION | Current religious affiliation |
| ECTIMESX | EA-12 Number of times R used emergency contraception |
| HIEDUC | Highest completed year of school or highest degree received (bottom-coded) |
| RELDLIFE | IC-7 How important is religion in R's daily life |
| DRINK12 | JC-4 Last 12 mos: how often drank alcoholic beverages |
| POT12 | JC-6 Last 12 mos: how often smoked marijuana |
| CONDSEXL | JD-11 Was condom used at last sex of any kind with a male partner |
| MALSHT12 | JF-6 Last 12 mos: R had sex with male intravenous drug user |
| ABORTION | Number of completed pregnancies ending in induced abortion |
| CURRPRTT | Number of Current Male Sexual Partners-including curr H/P |
| PARTS1YR | Number of opposite-sex sexual partners in last 12 months (top-coded) |
| AGE\_R | R's age at interview (FC A-2b) |
| RSCRRACE | R's race as reported in screener |
| TOTINCR | Total income of R's family |

Table 10: A table containing the predictors and response variables (of female respondents) to be used in Poisson Regression Modeling.

## 2. Exploratory Data Analysis

### Missing Data and Cardinality

A table of numbers and letters

AI-generated content may be incorrect.

Table 11: A table representing variable characteristics for sexual behavior analysis of female respondents

Based on Table 11, the data set of *Female Respondents* contains many null values. Null values are placeholders for the respondents who are not applicable to the specific questions. There are significant number of instances where a respondent “Don’t Know” (9,99,999) or “Refused” (8,88,88) to answer the question.

Each variable required different handling since some of the seemingly unusable values may contain key information.

### Distributions

#### 2.2.1 Continuous Variable

A group of graphs showing distribution of menarche

AI-generated content may be incorrect.

Figure 6: Distribution of Continuous Variables (with cardinality of 21 and above)

Without the invalid values (“Don’t Knows” and “Refused”), Figure 1 shows the distribution of continuous variables. The dependent variable “OPPLIFENUM,” is skewed to the right which is to be expected of a count variable. The same is the case for “Menarche” and “VRY1STAG.” “AGE\_R” appeared to be relatively normally distributed (compared to others).

A diagram of a distribution of continuous variable

AI-generated content may be incorrect.

Figure 7: Box-Plot Distribution of Continuous Variables (with cardinality of 21 and above)

Figure 7 showed the Box-Plot distribution of the continuous variables and the presence of some significant outliers. A normal regression will not be suitable for this dataset for even the “Opplifenum” have numerous significant outliers. Since that is normal behavior for count variables, these outliers could not be removed. *Poisson Regression* was used instead of normal regression. Although the presence of outliers may be an early warning of *overdispersion*. This will be problematic with *Poisson Regression* because the primary assumption of it is that the *mean is equal to variance*. Overdispersion occurs when this assumption is violated. In such cases, *Negative Binomial Regression* will be used.

#### 2.2.1 Categorical Variable

A group of blue and black bars

AI-generated content may be incorrect.

Figure 8: Distribution of Discrete/Categorical Variable. Note that the histogram is horizontally oriented.

Figure 3 indicates a highly imbalanced distribution of categorical variables. Some features have significant “nan” or respondents whose inapplicable for a specific question, sometimes even greater than those who are applicable. This is an indication that one-hot-encoding is necessary to counteract the imbalance.

### Variable Transformation Details

| **Variable** | **Wrangling Steps** |
| --- | --- |
| **ABORTION** | * NaN values filled with 0. * Renamed to Abortion\_num. |
| **CONDSEXL** | * NaN values replaced with "inapplicable". * Value 5 replaced with 0.<p> * One-Hot Encoded. |
| **DRINK12** | * One-Hot Encoded. * Rows where DRINK12 is 9 are dropped. |
| **ECTIMESX** | * NaN values replaced with "Never" * Non-"Never" values are clustered using a KMedoids clustering model * Original values are replaced with cluster labels. |
| **GRFSTSX** | * NaN values filled with "nosex\_u18 * Rows where GRFSTSX is 98 or 99 are dropped * Values 14, 15, and 16 are mapped to 13. |
| **LVSIT14F** | * Rows where LVSIT14F is 8 or 9 are dropped. * Values mapped using dictionary: {1: "Bio/adoptive mother", 2: "Other mother figure", 3: "No mother figure", None: "both parents"}. * One-Hot Encoded. * The column LVSIT14F\_both parents is dropped. |
| **LVSIT14M** | * Rows where LVSIT14M is 8 or 9 are dropped. * Values mapped using dictionary: {1: "Bio/adoptive mother", 2: "Bio/adoptive father", 3: "Step father", 4: "Other father figure", None: "both parents"}. * One-Hot Encoded. |
| **MALSHT12** | * Each unique value transformed into separate binary indicator column.<p>• Original column dropped. * Rows where MALSHT12 is 8 or 9 are dropped. |
| **MARSTAT** | * Rows where MARSTAT is 8 or 9 are dropped. |
| **MENARCHE** | * Value 96 replaced with 37 * Rows where MENARCHE is 98 or 99 are dropped. |
| **ONOWN** | * Rows where ONOWN is 8 or 9 are dropped. |
| **OPPLIFENUM** | * NaN values filled with 0 * Rows where OPPLIFENUM is 998 or 999 are dropped (Invalid "Refused and Don't Know" responses). |
| **PARTS1YR** | * Value 5 replaced with 0 * Transformed into two binary columns: PARTS1YR\_no and PARTS1YR\_yes. * Original column dropped. |
| **POT12** | * One-Hot Encoded. * Rows where POT12 is 9 are dropped. |
| **RELDLIFE** | * Values replaced: 8 → 4, 9 → 5, NaN → 6. |
| **WHOFSTPR** | * Each unique value transformed into separate binary indicator column.<p>• Original column dropped. * Rows where WHOFSTPR is 98 or 99 are dropped. |
| **vry1stag** | * Rows where vry1stag is 97 are dropped. * Binned into categories: '≤14 years', '15-17 years', '18-19 years', '20+ years'.<p> * NaN values become "Never Had sex". * Binned version is One-Hot Encoded. |

Table 12: A table representing the processing done to each variable. Code in appendix.

### Correlation

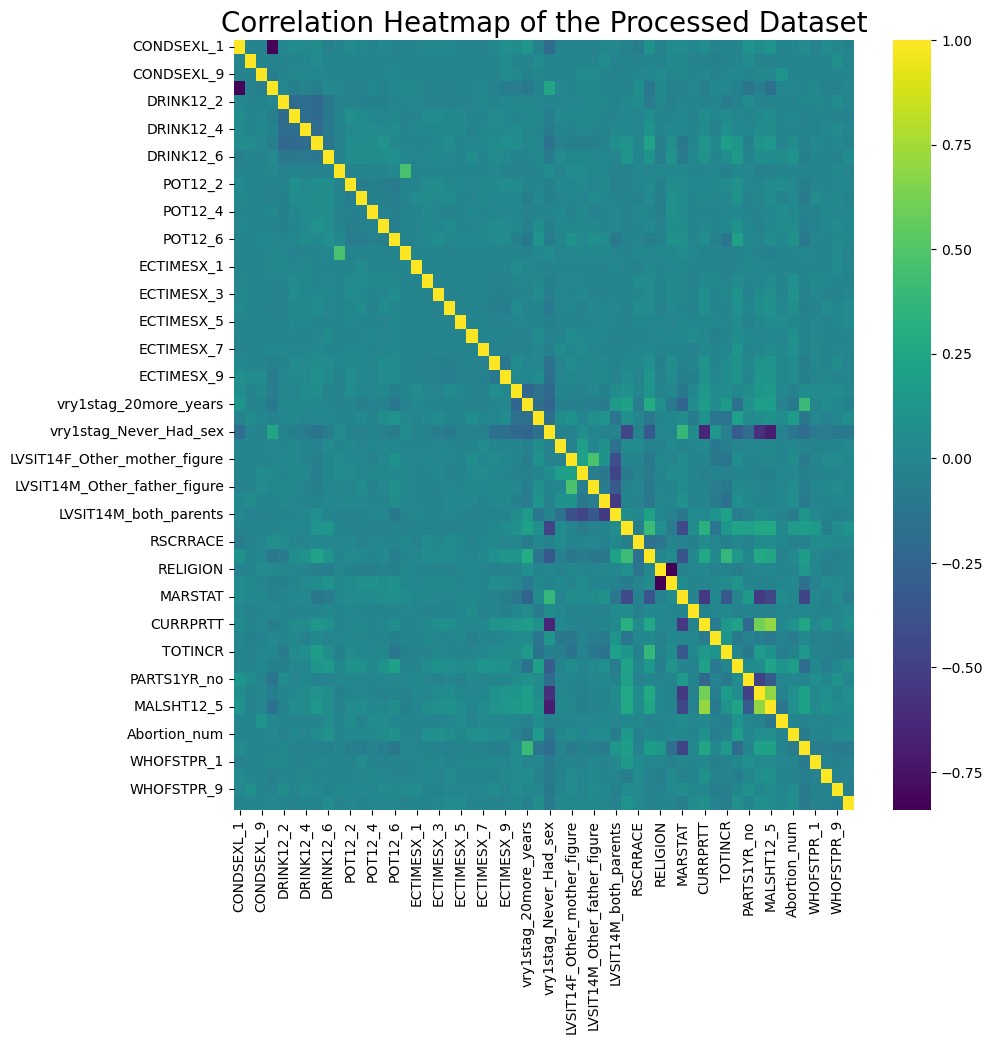


Figure 9: Correlational heatmap of the processed dataset. Note that the highest Pearson correlation is “CURRPRTT” and “MALSHT12\_5” with 0.70. This correlation is not significant enough to flag multicollinearity.

## 3. Split

The original data were loaded again and was split into two sets, training set (80%) and test set (20%). The reloading of original state was done to prevent data leakage.

## 4. Cross Validation

A cross validation was conducted for each combination of parameters:

For Poisson Regression, the parameters “alpha” and “L1\_wt” of regularization was tuned with the following set of values:

* alphas = [0.01, 0.1, 0.5, 1.0, 10.0]
* l1\_wts = [0.1, 0.5, 0.9, 1.0]

For Negative Binomial Regression, “alpha\_binom” was added to tune the best alpha for the Negative Binomial Regression object.

* alphas\_binom\_to\_tune = [0.1, 0.5, 1.0, 2.0]

Note that calculation of deviance for a regularized model (“elastic\_net” was used) is not yet implemented in python, or any statistical summary for that matter. Two probable reason for this are: (a) the fact that there is no consensus to the right computation some statistical summary of a regularized model (including deviance) and (b) the regularization fitting is primarily used for model selection that relies more heavily on cross-validated predictive performance rather than statistical metric performance.

To get around calculation of overdispersion (“overdis\_params”), the modeler took the non-zero coefficient parameters of the regularized model and train a unregularized model to calculate for possible overdispersion.

|  |  |
| --- | --- |
| **Poisson Regression** | **Negative Binomial Regression** |
| A table of numbers and letters  AI-generated content may be incorrect. | A table of numbers and letters  AI-generated content may be incorrect. |
| Table 13.1: Result of hyperparameter tuning of Poisson Regression (top 10 Mean Square Error or MSE) | Table 13.2: Result of hyperparameter tuning of Negative Binomial Regression (top 10 MSE) |

Both Table 13.1 and 13.2 showed of the hyperparameter tuning and cross-validation for Poisson and Negative Binomial Regression, respectively. The Poisson Regression did have the higher mean square error. Although, as was suspected earlier (see Figure 7), overdispersion occur in all of them. For that reason, Negative Binomial Regression was imperative. Negative Binomial Regression did have some model that have over dispersed.

## 5. Model Selection

* Poisson Regression – {“alpha”: 0.01, “L1\_wt”: 0.1} or the highest MSE (overdispersion =4.29).
* Negative Binomial Regression – {“alpha”: 0.01, “L1\_wt”:0.5, “alpha\_binom”: 0.5} or the highest MSE that does not over dispersed (overdispersion =0.97).

## 6. Model Evaluation

A table with numbers and letters

AI-generated content may be incorrect.

Table 14: MSE Result of each model. Note that the baseline is basically using the mean of the response variable as prediction for all values of predictors.

Table 14 clearly indicates that both Poisson Regression and Negative Binomial model performed significantly better than the baseline. Predictive accuracy speaking, it is ideal to pick the model with lowest MSE. In this case, Poisson Regression. But as was said earlier, all Poisson Regression model have resulted to overdispersion, including the best parameter. Thus, it becomes imperative to choose the **Negative Binomial model even though it performed worse since it didn’t over dispersed.**

## 7. Communication

A screenshot of a computer

AI-generated content may be incorrect.

Table 15: Summary of unregularized Negative Binomial Regression trained using parameters from regularized Negative Binomial Regression

A table of numbers and letters

AI-generated content may be incorrect.

Table 16: Coefficients of unregularized Negative Binomial Regression trained using parameters from regularized Negative Binomial Regression

A graph of a graph

AI-generated content may be incorrect.

Figure 10: Variable Importance for selected parameters of Negative Binomial Regression. Note that Percent Change (%) was calculated by

****

Figure 10 shows how different factors affect the the final model, measured as percent change in expected count. Most variables in the model are statistically significant (marked with \*\*\*).

**Strong Positive Effects:**

- “MALSHT12\_1” has the largest impact, increasing expected counts by ~200%. These are the people who had sex with male drug user in the past 12 months.

- “DRINK12\_8” increases counts by ~150%. These are the people who have refused to disclosed how often they drank alcohol in the past 12 months.

- ECTIMESX\_6 both increase counts by ~100%.  These are the cluster of people who have used emergency contraceptives about 15-20 times.

**Strong Negative Effects:**

- “vry1stag\_Never\_Had\_sex” decreases counts by ~75%. These are the people who never had sex.

- “WHOFSTPR” variables (7 and 8) decrease counts by ~50%. WHOFSTPR\_7 and WHOFSTPR\_8 are the people who have had their first sexual encounter with their husband or current partner.

- “vry1stag\_20more\_years” decreases counts by ~50%. These are the people who has had their sexual encounter when their 20 and older.

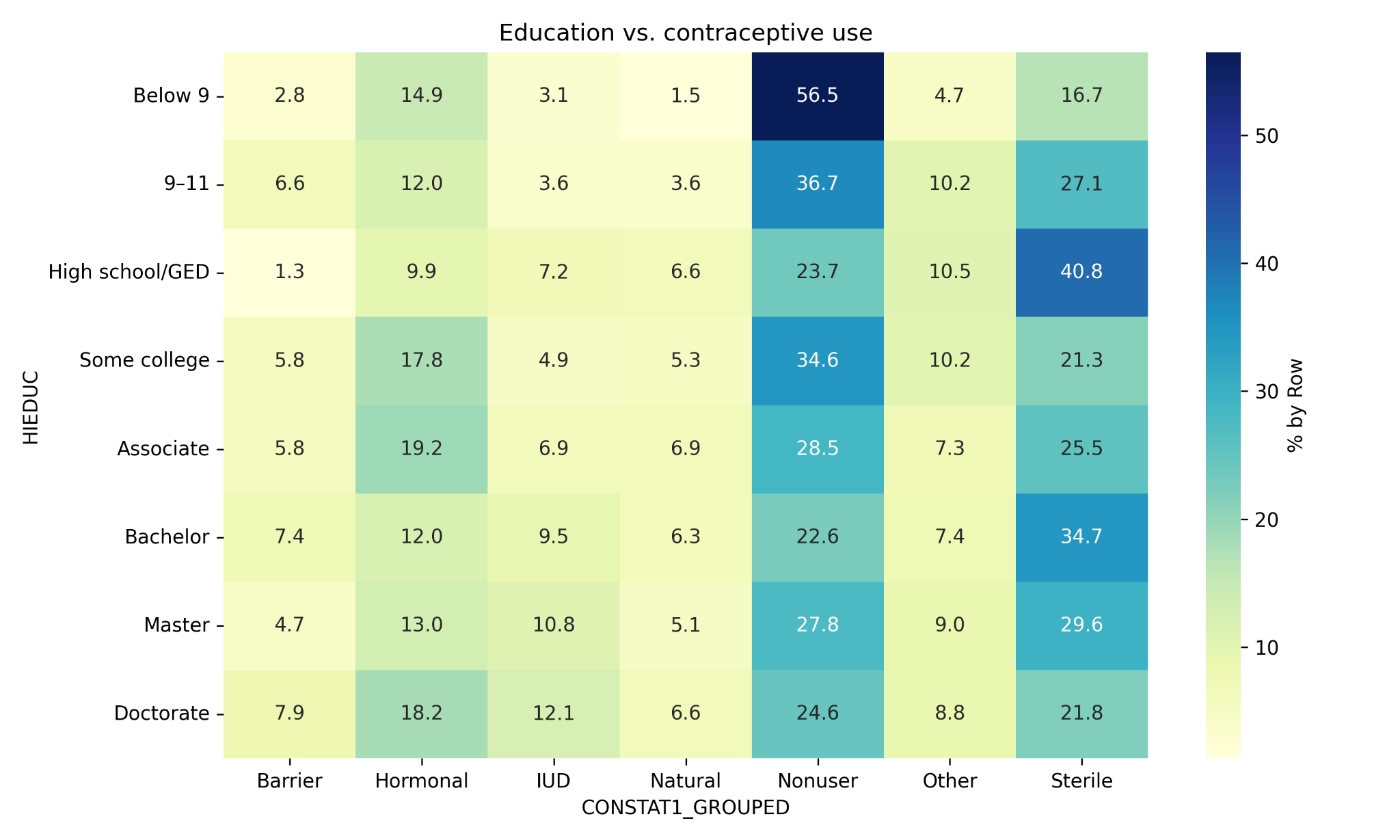
# Contingency Tables

|  |  |
| --- | --- |
| Variable Name | Description |
| AGE\_R | Age of respondent |
| HIEDUC | Highest educational attainment |
| POVERTY | Family income as percentage of federal poverty threshold |
| RELIGION | Current religious affiliation |
| INTCTFAM | Intact status of childhood family |
| vry1stag | Age at first sexual intercourse |
| FMARITAL | Marital Status |
| CONSTAT1 | Current contraceptive status |
| GENHEALT | General health status (scaled from Poor to Excellent) |
| OPPLIFENUM | Number of opposite-sex partners in lifetime |

Table 17: Variables Used in Contingency Tables Section

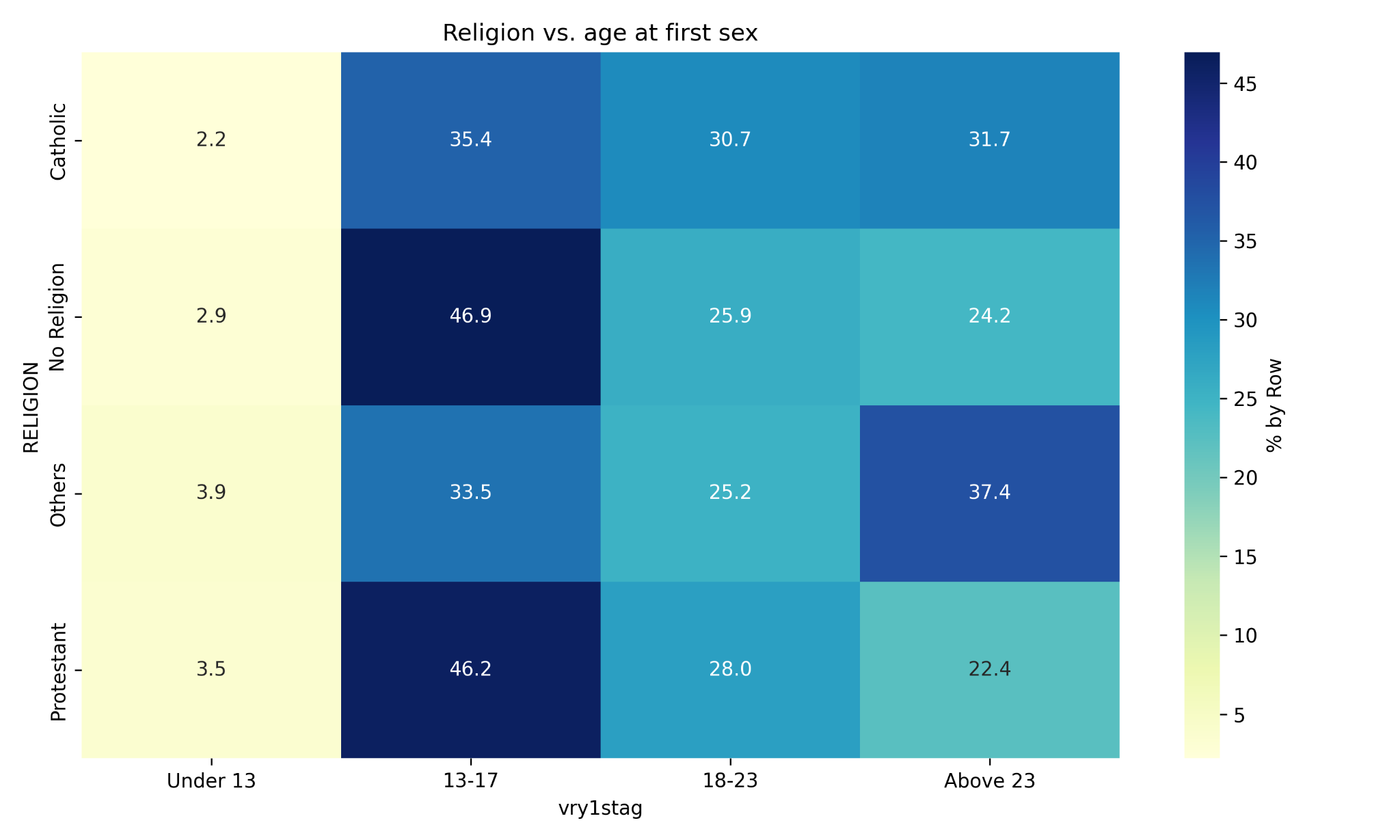
1. **Education vs. Contraceptive Use**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| HIEDUC | Below 9 | 9–11 | High school/GED | Some college | Associate | Bachelor | Master | Doctorate |
| CONSTAT1\_GROUPED |  |  |  |  |  |  |  |  |
| Barrier | 17 | 11 | 2 | 46 | 55 | 26 | 13 | 99 |
| Hormonal | 92 | 20 | 15 | 141 | 181 | 42 | 36 | 228 |
| IUD | 19 | 6 | 11 | 39 | 65 | 33 | 30 | 152 |
| Natural | 9 | 6 | 10 | 42 | 65 | 22 | 14 | 83 |
| Nonuser | 349 | 61 | 36 | 274 | 269 | 79 | 77 | 309 |
| Other | 29 | 17 | 16 | 81 | 69 | 26 | 25 | 110 |
| Sterile | 103 | 45 | 62 | 169 | 241 | 121 | 82 | 274 |

**  
Figure 11.** Education vs. Contraceptive Use (Row-wise %)

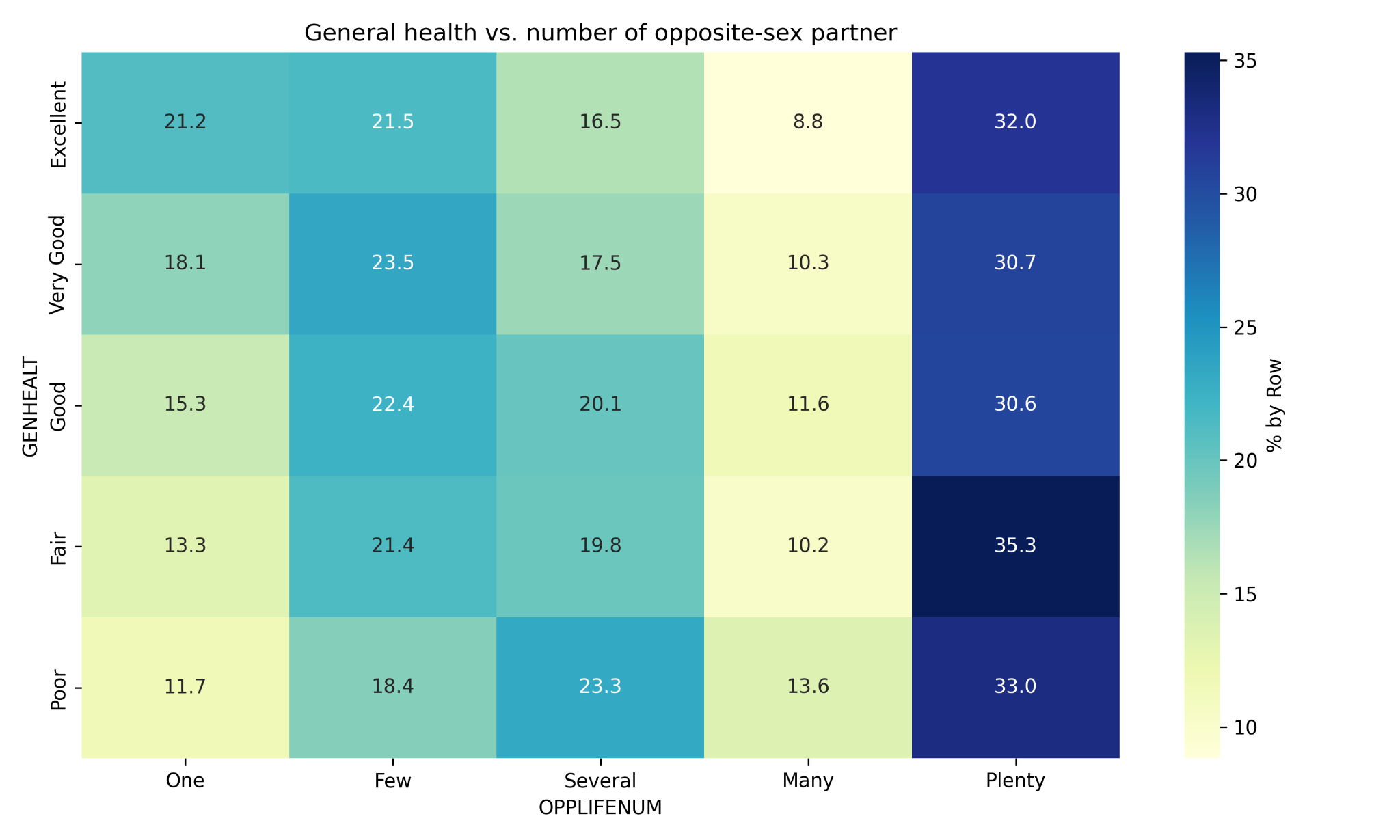
1. **Religion vs. Age at First Sexual Intercourse**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| vry1stag | Under 13 | 13-17 | 18-23 | Above 23 |
| RELIGION |  |  |  |  |
| Catholic | 18 | 289 | 251 | 259 |
| No Religion | 46 | 736 | 406 | 380 |
| Others | 16 | 138 | 104 | 154 |
| Protestant | 61 | 811 | 492 | 393 |

**  
Figure 12.** Religion vs. Age at First Sexual Intercourse (Row-wise %)

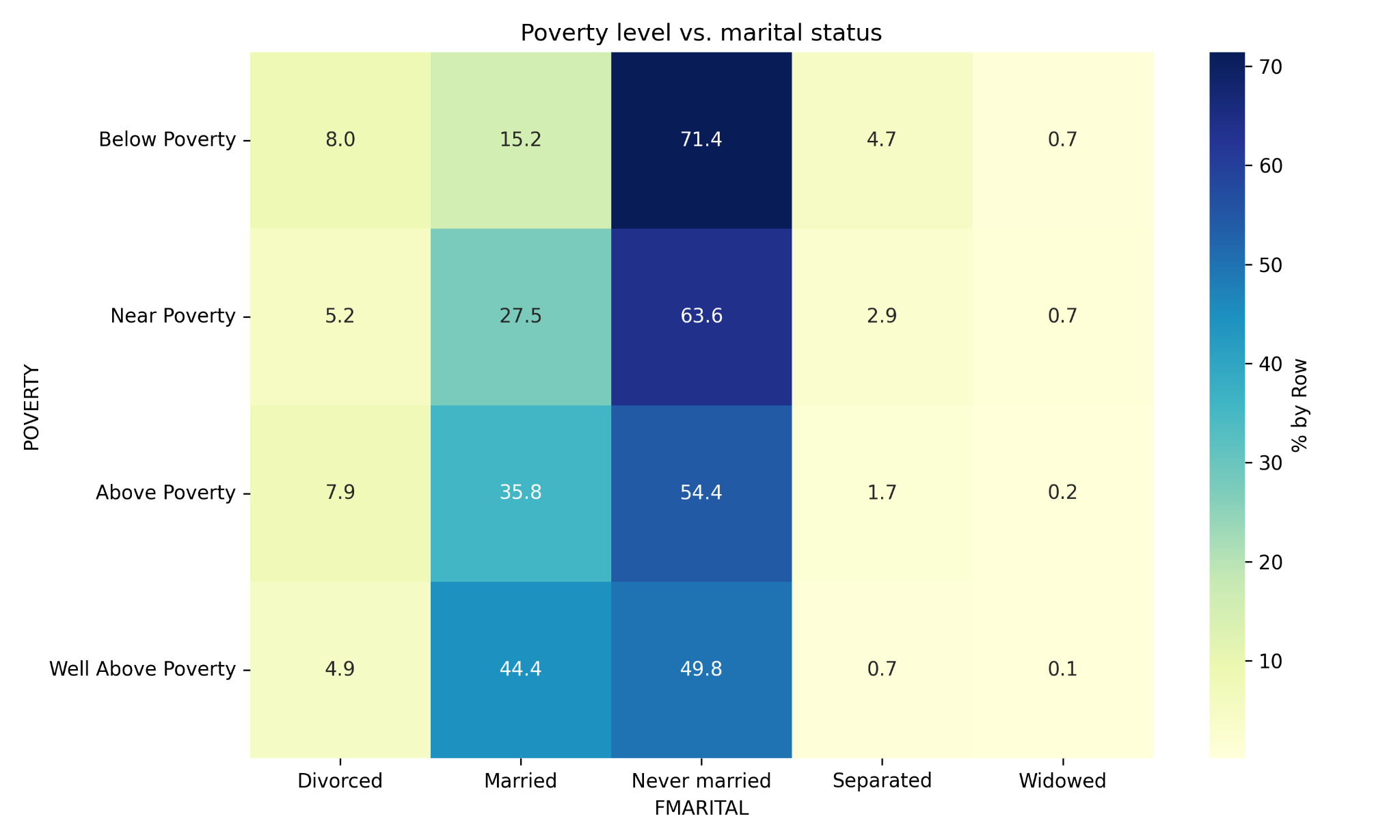
1. **General Health vs. Number of Opposite-Sex Partners**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| OPPLIFENUM | One | Few | Several | Many | Plenty |
| GENHEALT |  |  |  |  |  |
| Excellent | 197 | 200 | 153 | 82 | 297 |
| Very Good | 285 | 370 | 275 | 162 | 483 |
| Good | 227 | 334 | 299 | 173 | 455 |
| Fair | 61 | 98 | 91 | 47 | 162 |
| Poor | 12 | 19 | 24 | 14 | 34 |

**  
Figure 13.** General Health vs. Number of Opposite-Sex Partners (Row-wise %)

1. **Poverty Level vs. Marital Status**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| FMARITAL | Divorced | Married | Never married | Separated | Widowed |
| POVERTY |  |  |  |  |  |
| Below Poverty | 83 | 157 | 737 | 48 | 7 |
| Near Poverty | 50 | 263 | 609 | 28 | 7 |
| Above Poverty | 97 | 440 | 669 | 21 | 3 |
| Well Above Poverty | 66 | 593 | 665 | 9 | 2 |

**  
Figure 14.** Poverty Level vs. Marital Status (Row-wise %)

**Chi-Square Test**

|  |  |  |  |
| --- | --- | --- | --- |
| **Comparison** | **Chi-square** | **p-value** | **df** |
| Education vs. Contraceptive Use | 371.582669 | 2.266653e-54 | 42 |
| Religion vs. Age at First Sexual Intercourse | 79.875308 | 1.710974e-13 | 9 |
| General health vs.  Number of opposite-sex partner | 35.777128 | 3.106796e-03 | 16 |
| Poverty level vs. marital status | 286.406587 | 3.338153e-54 | 12 |

1. **Education vs. Contraceptive Use**

There is a high and statistically significant association between education level and contraceptive method used. The very low p-value indicates that contraceptive preferences vary significantly across educational backgrounds. This makes sense—education often affects knowledge, access, and attitudes toward reproductive health.

1. **Religion vs. Age at First Sex**

There is a strong and significant association between religious affiliation and the age at first sexual intercourse. Religious beliefs may influence norms around abstinence, sexual behavior, and the timing of first sex. The relationship is unlikely to be due to chance.

1. **General Health vs. Number of Opposite-Sex Partners**

There is a statistically significant association, though weaker than others. General health status may be related to the number of opposite-sex partners—possibly reflecting behavioral or socioeconomic factors affecting both.

1. **Poverty Level vs. Marital Status**

There is an extremely strong and significant association between poverty level and marital status. Lower-income individuals may have different patterns in marriage or separation compared to those above the poverty line—possibly due to economic barriers, social norms, or stress-related factors.

# Categorical Response Modeling

The variables used are the same as the previous section, “Contingency Tables.”

## Ordinal Logistic Regression Analysis

Using ordinal logistic regression, we will examine the relationship between age, education, income level, religion, childhood family status, age at first sexual intercourse, marital status, contraceptive use, general health, and the likelihood of having multiple opposite-sex partners. We will use Python libraries, such as statsmodels, to apply the regression model.

1. **Model Fitting**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Log-Likelihood | -6696.79871 |
| AIC | 13439.597419 |
| Converged | True |

1. **Log-Likelihood**

The model has a log-likelihood of -6696.8, which indicates an acceptable fit for ordinal logistic regression on social data.

1. **AIC (Akaike Information Criterion)**

The AIC value of 13,440 suggests a balance between model fit and simplicity. This can be used for comparison if multiple models are tested.

1. **Convergence**

The model converged successfully, indicating the estimated parameters are statistically reliable.

1. **Odds Ratios**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Coefficient** | **Odds Ratio** | **CI Lower (95%)** | **CI Upper (95%)** | **p-value** |
| AGE\_R | -0.16511 | 0.8478 | 0.783705 | 0.917138 | 3.85E-05 |
| HIEDUC | -0.028148 | 0.972245 | 0.945658 | 0.999578 | 4.66E-02 |
| POVERTY | 0.086903 | 1.090791 | 1.034576 | 1.150062 | 1.29E-03 |
| vry1stag | 0.211503 | 1.235534 | 1.149587 | 1.327907 | 8.95E-09 |
| GENHEALT | 0.099107 | 1.104184 | 1.0452 | 1.166497 | 4.03E-04 |
| RELIGION\_No Religion | 0.507521 | 1.661168 | 1.417341 | 1.946941 | 3.69E-10 |
| RELIGION\_Others | 0.154429 | 1.166992 | 0.93359 | 1.458746 | 1.75E-01 |
| RELIGION\_Protestant | 0.349327 | 1.418113 | 1.215232 | 1.654865 | 9.22E-06 |
| INTCTFAM\_Had two parents | -0.250334 | 0.778541 | 0.692835 | 0.874849 | 2.59E-05 |
| FMARITAL\_Married | -1.019873 | 0.360641 | 0.290479 | 0.44775 | 2.48E-20 |
| FMARITAL\_Never married | -0.2419 | 0.785134 | 0.632387 | 0.974777 | 2.84E-02 |
| FMARITAL\_Separated | -0.457586 | 0.63281 | 0.4351 | 0.920359 | 1.67E-02 |
| FMARITAL\_Widowed | -0.631696 | 0.531689 | 0.232401 | 1.216404 | 1.35E-01 |
| CONSTAT1\_GROUPED\_Hormonal | 0.347272 | 1.415201 | 1.10535 | 1.81191 | 5.88E-03 |
| CONSTAT1\_GROUPED\_IUD | 0.710979 | 2.035983 | 1.53973 | 2.692177 | 6.10E-07 |
| CONSTAT1\_GROUPED\_Natural | 0.576414 | 1.779645 | 1.320849 | 2.397804 | 1.51E-04 |
| CONSTAT1\_GROUPED\_Nonuser | 1.217933 | 3.380193 | 2.660314 | 4.294871 | 2.11E-23 |
| CONSTAT1\_GROUPED\_Other | 0.512494 | 1.66945 | 1.265039 | 2.203144 | 2.93E-04 |
| CONSTAT1\_GROUPED\_Sterile | 0.792502 | 2.208915 | 1.73735 | 2.808477 | 9.92E-11 |

1. **AGE\_R** (Age)
   1. For each increase in age group, the odds of being in a higher category of opposite-sex partners decrease by 15.2%.
   2. Older respondents are less likely to report more partners.
2. **HIEDUC** (Education Level)
   1. Not statistically significant at 0.05 (p = 0.047 borderline).
   2. Slight negative effect: more education may be associated with fewer partners, but this effect is weak.
3. **POVERTY** (Poverty Level)
   1. For each increase in poverty level, the odds of reporting more partners increase by 9.1%.
   2. Suggests possible link between economic status and sexual behavior.
4. **vry1stag** (Age at First Sex)
   1. Each increase in age group of first sex increases odds of more partners by 23.6%.
   2. This seems counterintuitive — may reflect categorization. Possibly, “older first sex” may cluster with specific behaviors later.
5. **GENHEALT** (General Health)

Better general health is associated with 10.4% higher odds of reporting more partners.

1. **RELIGION**
   1. **No Religion**: 66% higher odds of having more partners than Catholics (baseline).
   2. **Protestant**: 41.8% higher odds.
2. **INTCTFAM\_Had two parents** (Had two parents)

Individuals from two-parent households are 22.2% less likely to report higher partner counts.

1. **FMARITAL**\*\* (Marital Status)
   1. **Married**: 64% lower odds of more partners.
   2. **Never married**: 33% lower odds
   3. **Separated/Widowed**: Even lower odds, suggesting marital status strongly predicts behavior.
2. **CONSTAT1\_GROUPED**
   1. **Hormonal**: Users are 68% more likely to be in higher categories
   2. **IUD**: Double the odds of reporting more partners
   3. **Natural**: Strong positive association
   4. **Nonuser**: Very strong predictor of high partner count
   5. **Other**: Similar to hormonal methods
   6. **Sterile**: Strong positive odds

Overall, the ordinal logistic regression model significantly predicts the number of opposite-sex partners based on variables such as age, poverty level, marital status, religion, and contraceptive method. For example, individuals using IUDs had 2.0 times higher odds of reporting more partners compared to the baseline group. Nonusers had 3.4 times higher odds. Additionally, being older, married, or from a two-parent household decreased the likelihood of reporting higher partner counts. The model converged successfully, and fit statistics (AIC = 13,440) indicate acceptable performance for social science data.

## Multinomial Logistic Regression Analysis

Using multinomial logistic regression, we will examine the relationship between age, education, income level, religion, childhood family status, age at first sexual intercourse, marital status, number of sex partners had, general health, and the likelihood of using a specific contraceptive method. We will use Python libraries again, such as statsmodels, to apply the regression model.

1. **Model Fitting**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Log-Likelihood | -6696.79871 |
| AIC | 13439.597419 |
| Converged | True |

The multinomial logistic regression model converged successfully, indicating that the optimization algorithm reached a stable solution and that the estimated coefficients are statistically valid. The model produced a log-likelihood of -6858.41, which reflects the degree of model fit—the higher (i.e., less negative) the value, the better the model fits the observed data.

The Akaike Information Criterion (AIC) value of 13,896.82 suggests an acceptable balance between model fit and complexity. Although AIC is not interpretable in absolute terms, it serves as a key comparative metric; lower AIC values are generally preferred when comparing alternative models.

1. **Odds Ratios**

A positive coefficient indicates increased odds of choosing that method over the base category, while a negative coefficient means decreased odds. Below are variables with their respective metrics. Only statistically significant predictors (p < 0.05) are highlighted.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Outcome | Coefficient | Odds Ratio | p-value |
| AGE\_R | Hormonal | -0.2363 | 0.7895437753 | 0.026 |
| POVERTY | Hormonal | 0.2443 | 1.276727291 | 0.001 |
| OPPLIFENUM | Hormonal | 0.1913 | 1.210822644 | 0.001 |
| AGE\_R | IUD | 0.2818 | 1.325513591 | 0.023 |
| POVERTY | IUD | 0.1723 | 1.18803419 | 0.039 |
| GENHEALT | IUD | 0.2168 | 1.242095658 | 0.013 |
| OPPLIFENUM | IUD | 0.2702 | 1.31022647 | 0 |
| vry1stag | Natural | -0.4037 | 0.6678444446 | 0.001 |
| HIEDUC | Nonuser | -0.1314 | 0.8768669574 | 0 |
| vry1stag | Nonuser | 0.7372 | 2.090075103 | 0 |
| OPPLIFENUM | Nonuser | 0.402 | 1.494811333 | 0 |
| HIEDUC | Other | -0.0873 | 0.9164021338 | 0.042 |
| OPPLIFENUM | Other | 0.2239 | 1.250945919 | 0 |
| AGE\_R | Sterile | 1.0026 | 2.725358557 | 0 |
| HIEDUC | Sterile | -0.1731 | 0.8410535053 | 0 |
| GENHEALT | Sterile | 0.2816 | 1.325248515 | 0 |
| OPPLIFENUM | Sterile | 0.2939 | 1.341649732 | 0 |

1. **Hormonal Method (vs. Base)**
   1. **AGE\_R**: Coef = -0.2363, OR ≈ 0.79  
      Older respondents are **21% less likely** to choose hormonal methods. *(p = 0.026)*
   2. **POVERTY**: Coef = 0.2443, OR ≈ 1.28  
      Higher poverty level is associated with a **28% increase** in odds of choosing hormonal methods. *(p = 0.001)*
   3. **OPPLIFENUM**: Coef = 0.1913, OR ≈ 1.21  
      Those with more opposite-sex partners have **21% higher odds** of choosing hormonal methods. *(p < 0.001)*
2. **IUD**
   1. **AGE\_R**: Coef = 0.2818, OR ≈ 1.33  
      Older respondents are **33% more likely** to choose IUD. *(p = 0.023)*
   2. **POVERTY**: Coef = 0.1723, OR ≈ 1.19  
      More poverty → higher odds of IUD use. *(p = 0.039)*
   3. **GENHEALT**: Coef = 0.2168, OR ≈ 1.24  
      Better health associated with IUD use. *(p = 0.013)*
   4. **OPPLIFENUM**: Coef = 0.2702, OR ≈ 1.31  
      Strong predictor: more partners → **31% higher odds** of using IUD. *(p < 0.001)*
3. **Natural Methods**
   1. **vry1stag**: Coef = -0.4037, OR ≈ 0.67  
       → Later age at first sex **reduces odds** of using natural methods by **33%**. *(p = 0.001)*
4. **Nonuser**
   1. **HIEDUC**: Coef = -0.1314, OR ≈ 0.88  
      More educated respondents are **12% less likely** to be nonusers. *(p < 0.001)*
   2. **vry1stag**: Coef = 0.7372, OR ≈ 2.09  
      Later sexual debut strongly increases likelihood of being a nonuser. *(p < 0.001)*
   3. **OPPLIFENUM**: Coef = 0.4020, OR ≈ 1.50  
      More sexual partners = **50% more likely** to be nonusers. *(p < 0.001)*
5. **Other Methods**
   1. **HIEDUC**: Coef = -0.0873, OR ≈ 0.92  
      Higher education **reduces odds** of choosing other methods by **8%**. *(p = 0.042)*
   2. **OPPLIFENUM**: Coef = 0.2239, OR ≈ 1.25  
      More partners = **25% more likely** to choose "Other". *(p < 0.001)*
6. **Sterile**
   1. **AGE\_R**: Coef = 1.0026, OR ≈ 2.72  
      Older respondents are nearly **3x more likely** to use sterilization. *(p < 0.001)*
   2. **HIEDUC**: Coef = -0.1731, OR ≈ 0.84  
      More educated = **16% less likely** to choose sterilization. *(p < 0.001)*
   3. **GENHEALT**: Coef = 0.2816, OR ≈ 1.33  
      Better health = **33% higher odds** of sterilization. *(p < 0.001)*
   4. **OPPLIFENUM**: Coef = 0.2939, OR ≈ 1.34  
      Strong positive predictor. *(p < 0.001)*

Overall, the results of the multinomial logistic regression suggest that several demographic and behavioral factors significantly influence contraceptive method choice. For example, individuals with a higher number of opposite-sex partners were consistently more likely to choose IUDs, hormonal methods, and sterilization. Conversely, higher educational attainment was associated with a decreased likelihood of being a nonuser or choosing sterilization.

Age, poverty level, general health, and age at first sexual intercourse also emerged as significant predictors in multiple contraceptive categories. Notably, increasing age was strongly associated with higher odds of selecting long-term or permanent methods such as IUDs and sterilization, while later age at first sex was linked to reduced reliance on natural methods and increased likelihood of nonuse.

# Conclusion

The analysis of female respondents from the NSFG dataset successfully identified several key factors influencing reproductive health behaviors.

Predictors such as age, education level, poverty status, age at first sexual intercourse, religious affiliation, family structure during childhood, and current contraceptive use are some of the more significant predictors in predicting the number of opposite-sex partners of female respondents. It is also worth noting that behaviors such as prior sex with a male drug user and high emergency contraceptive use were associated with a higher number of sexual partners, while never having had sex or having first sexual experiences later in life or with a husband/current partner were linked to fewer partners.

With regards to contraceptive choice, demographic and behavioral factors like age, poverty level, general health, age at first sexual intercourse, and the number of opposite-sex partners significantly influenced the type of contraceptive method used. For instance, individuals with more partners were more likely to use IUDs, hormonal methods, or sterilization. Higher educational attainment was also generally associated with a decreased likelihood of being a nonuser or choosing sterilization.

# Recommendations

It is important to **develop tailored educational programs and resources** focusing on groups identified with higher-risk behaviors or specific contraceptive needs (e.g., younger individuals, those with earlier sexual debut, those with a higher number of partners).

Given the influence of age at first sex, **further explore how to best support informed decision-making around sexual debut.**

**Enhance** comprehensive **sexual health education**, emphasizing various contraceptive methods, their effectiveness, and addressing socioeconomic barriers that may limit access to preferred options, particularly considering the influence of poverty.

Explore the **underlying reasons for the observed associations**, particularly with the strong links between specific behaviors (e.g., emergency contraceptive use frequency, substance use in partners) and the number of sexual partners.

**Investigate the dynamics influencing contraceptive choices in more detail**, like looking into the male perspective (dataset) and cross-analyzing with the female perspective. Even just digging deeper to the objective of this report like how life transitions and changing partner numbers affect method selection over time.

# Code Appendix

The annotated codes or notebook file, divided by sections, can be found here:

<https://github.com/RomandRapido/DSC1105_EDA/blob/main/SA2_Ramilo-Lansangan-Dacanay/Delivarables_CODES-SEC%201-SA2%20GROUP%201-DACANAY%2C%20J%3B%20LANSANGAN%2C%20R%3B%20RAMILO%2C%20Z.ipynb>