| **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 1: A table containing the predictors and response variables (of female respondents) to be used in Poisson Regression Modeling.

## **Missing Data and Cardinality**

A table of numbers and letters

AI-generated content may be incorrect.

Table 2: A table representing variable characteristics for sexual behavior analysis of female respondents

Based from Table 2, 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 1: 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 2: Box-Plot Distribution of Continuous Variables (with cardinality of 21 and above)

Figure 2 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.

A group of blue and black bars

AI-generated content may be incorrect.

Figure 3: 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 3: A table representing the processing done to each variable. Code in appendix.

# Correlation

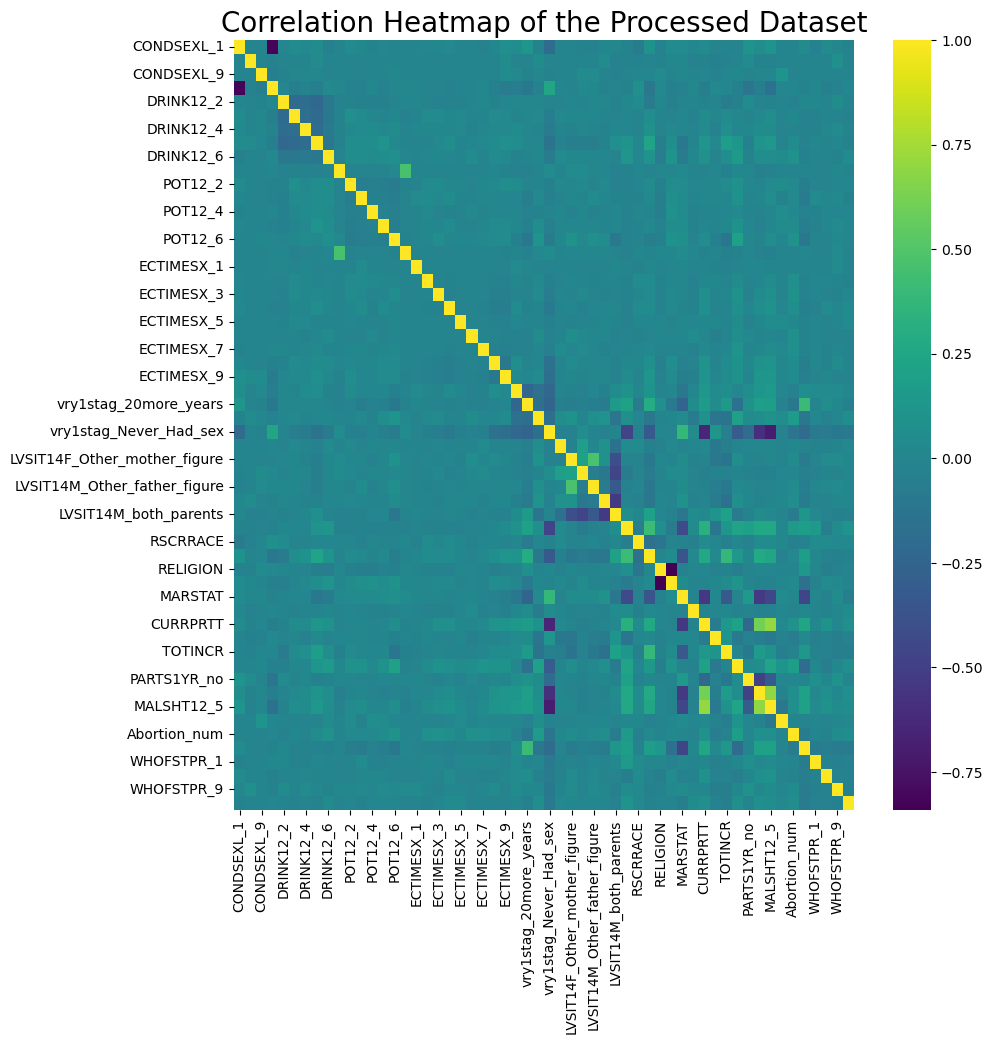


Figure 3: Correlational heatmap of the processed dataset. Note that the highest Pearson correlation is “CURRPRTT” and “MALSHT12\_5” with 0.70. This correlation is not significant enough to flag multicollinearity.

# 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.

# 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 4.1: Result of hyperparameter tuning of Poisson Regression (top 10 Mean Square Error or MSE) | Table 4.2: Result of hyperparameter tuning of Negative Binomial Regression (top 10 MSE) |

Both Table 4.1 and 4.2 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 2), 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.

# Model Selection

* Poisso 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).

# Model Evaluation

A table with numbers and letters

AI-generated content may be incorrect.

Table 5: 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 5 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.**

A screenshot of a computer

AI-generated content may be incorrect.

Table 6: 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 6: 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 4: Variable Importance for selected parameters of Negative Binomial Regression. Note that Percent Change (%) was calculated by



Figure 4 shows how different factors affect the of 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%

- “DRINK12\_8” increases counts by ~150%

- “MALSHT12\_5” and ECTIMESX\_6 both increase counts by ~100%

**Strong Negative Effects:**

- “vry1stag\_Never\_Had\_sex” decreases counts by ~75%

- “WHOFSTPR” variables (7 and 8) decrease counts by ~50%

- “vry1stag\_20more\_years” decreases counts by ~50%