Introduction and Objectives

This study focuses on identifying social and demographic predictors of pregnancy outcomes among women in the United States. The data used in this study come from the 2022–2023 Female Respondent Public Use File of the National Survey of Family Growth (NSFG), a nationally representative survey conducted by the National Center for Health Statistics (NCHS). The survey collects detailed information about family life, reproductive health, marriage, contraception, and related topics from women aged 15–49 in the U.S.

The primary objective of this analysis is to examine how a woman's background, education, marital status, and family structure relate to the total number of pregnancies a woman have. Using Poisson and negative binomial regression models, this study explores the associations between these predictors and pregnancy numbers. Additional statistical methods, including ordinal and multinomial logistic regression, chi-square tests, and data visualizations, are used to support a deeper understanding of patterns in the data.

Description of Dataset and Variables

Variable Name	Description	Reason for Inclusion
PREGNUM	Total number of pregnancies reported by the respondent.	Outcome variable; primary focus of the study to examine factors associated with pregnancy frequency.
FMARITAL	Current formal marital status (e.g., married, divorced, never married).	Marital status influences the likelihood of planned or unplanned pregnancies.
HISPRACE2	Respondent's race and Hispanic origin, categorized as Hispanic, White, Black, or Other.	To examine potential disparities in pregnancy outcomes by race and ethnicity.
HIEDUC	Highest level of education completed by the respondent (ordinal).	Education level may affect reproductive choices, family planning, and timing of pregnancies.
AGER	Age of the respondent at the time of the interview (continuous).	Older individuals have had more time to experience pregnancies, making age a natural covariate.
EDUCMOM	Education level of the respondent's mother or female caregiver.	Used as a proxy for childhood socioeconomic status, which may influence reproductive outcomes.
INTCTFAM	Indicates whether the respondent grew up in an intact family (lived with both parents).	Family structure during childhood might shape future relationship stability and reproductive decisions.

Result of Analysis

Poisson Regression

```
Call:
glm(formula = PREGNUM ~ FMARITAL + HISPRACE2 + HIEDUC + AGER +
    EDUCMOM + INTCTFAM, family = poisson(link = "log"), data = data)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.892588   0.066881 -13.346  < 2e-16 ***
FMARITALWidowed      -0.262698   0.127781  -2.056 0.039797 *
FMARITALDivorced     -0.158446   0.037558  -4.219 2.46e-05 ***
```

```
FMARITALSeparated     -0.049787   0.058604  -0.850 0.395579
FMARITALNever Married -0.783609   0.028664 -27.338  < 2e-16 ***
HISPRACE2White        -0.061340   0.031257  -1.962 0.049709 *
HISPRACE2Black         0.266027   0.036959   7.198 6.12e-13 ***
HISPRACE2Other        -0.150477   0.042451  -3.545 0.000393 ***
HIEDUC.L              -0.584296   0.065571  -8.911  < 2e-16 ***
HIEDUC.Q              -0.145761   0.056741  -2.569 0.010202 *
HIEDUC.C               0.320143   0.051383   6.231 4.65e-10 ***
... (etc.)
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 12582.0  on 5585  degrees of freedom
Residual deviance:  8731.9  on 5562  degrees of freedom
AIC: 17206

Number of Fisher Scoring iterations: 6
```

Negative Binomial Model

```
##
## Call:
## glm.nb(formula = PREGNUM ~ FMARITAL + HISPRACE2 + HIEDUC + AGER +
##     EDUCMOM + INTCTFAM, data = data, init.theta = 2.346770487,
##     link = log)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.171003   0.088576 -13.220  < 2e-16 ***
## FMARITALWidowed       -0.308018   0.184903  -1.666  0.09575 .
## FMARITALDivorced      -0.243232   0.054254  -4.483 7.35e-06 ***
## FMARITALSeparated     -0.083390   0.087683  -0.951  0.34158
## FMARITALNever Married -0.876554   0.037591 -23.318  < 2e-16 ***
## HISPRACE2White        -0.062713   0.042515  -1.475  0.14019
## HISPRACE2Black         0.323478   0.050895   6.356 2.07e-10 ***
## HISPRACE2Other        -0.165591   0.056771  -2.917  0.00354 **
## HIEDUC.L              -0.686856   0.089277  -7.694 1.43e-14 ***
## HIEDUC.Q              -0.218934   0.076444  -2.864  0.00418 **
## HIEDUC.C               0.415490   0.069475   5.980 2.23e-09 ***
## HIEDUC^4              -0.028068   0.074604  -0.376  0.70675
## HIEDUC^5               0.078777   0.073555   1.071  0.28417
## HIEDUC^6              -0.088840   0.070085  -1.268  0.20493
## HIEDUC^7              -0.071344   0.065955  -1.082  0.27938
## HIEDUC^8               0.019116   0.055448   0.345  0.73028
## HIEDUC^9              -0.016349   0.047437  -0.345  0.73036
## HIEDUC^10              0.042292   0.050256   0.842  0.40006
## AGER                   0.054487   0.001976  27.568  < 2e-16 ***
## EDUCMOM.L             -0.130601   0.079521  -1.642  0.10052
## EDUCMOM.Q              0.220303   0.070078   3.144  0.00167 **
## EDUCMOM.C              0.131053   0.047777   2.743  0.00609 **
## EDUCMOM^4              0.080844   0.033125   2.441  0.01466 *
## INTCTFAMNot Intact     0.259201   0.033146   7.820 5.28e-15 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.3468) family taken to be 1)
##
##     Null deviance: 8370.1  on 5585  degrees of freedom
## Residual deviance: 5709.3  on 5562  degrees of freedom
## AIC: 16496
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  2.347
##           Std. Err.:  0.131
##
##  2 x log-likelihood:  -16445.707
```

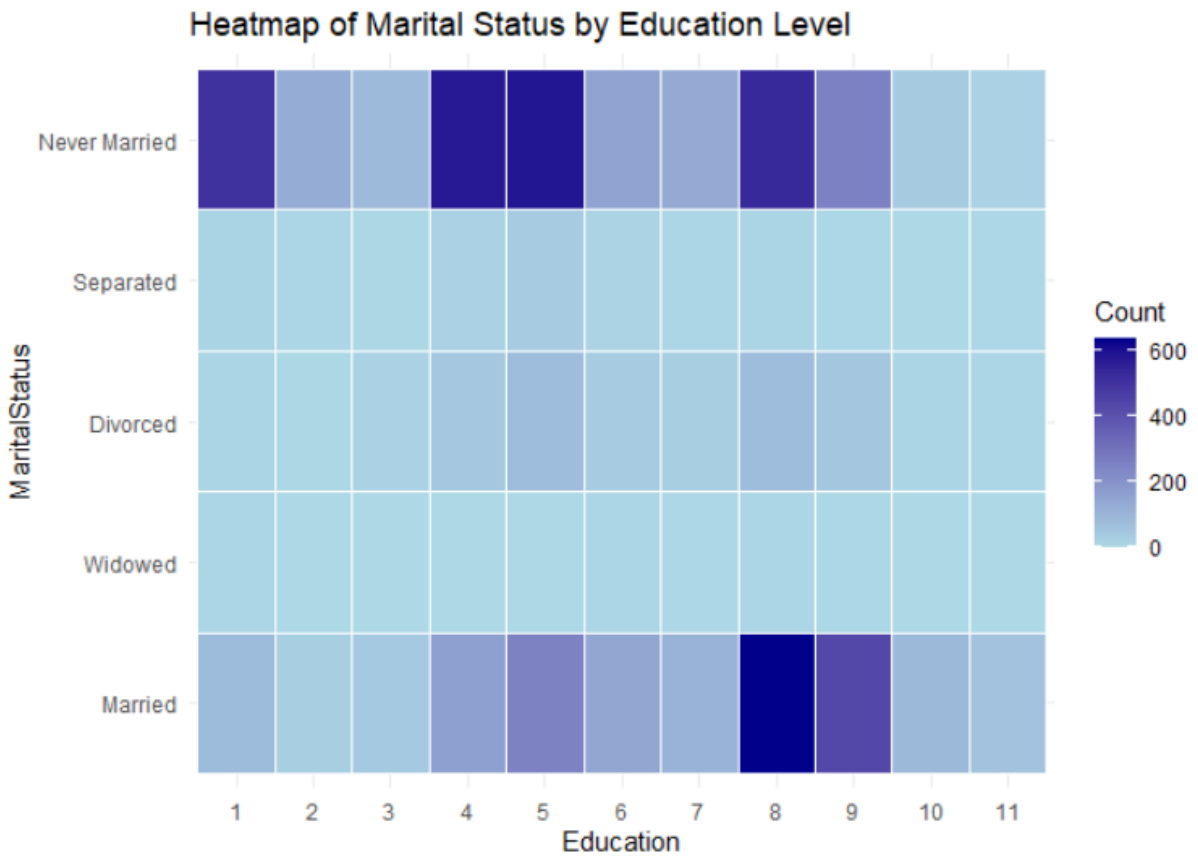
Contingency Table

```
##                HIEDUC
## FMARITAL          1   2   3   4   5   6   7   8   9  10  11
##   Married        85  30  44 164 254 147 110 637 437  90  62
##   Widowed         5   0   1   1   1   5   2   6   3   0   1
##   Divorced       12   4  16  45  80  34  28  80  51   9   6
##   Separated      15   6   4  18  33  13   7  11   3   0   2
##   Never Married 506 129  88 579 588 154 135 533 258  36  18
```

Pearson's Chi-squared test

```
##
##  Pearson's Chi-squared test
##
## data:  ct
## X-squared = 804.43, df = 40, p-value < 2.2e-16
```

Heatmap



Ordinal Logistic Regression (HIEDUC)

```
## Call:
## polr(formula = HIEDUC ~ AGER + HISPRACE2 + INTCTFAM + EDUCMOM,
##     data = data, Hess = TRUE)
##
## Coefficients:
##                       Value Std. Error t value
## AGER                0.08940    0.00282  31.700
## HISPRACE2White      0.41712    0.06688   6.237
## HISPRACE2Black     -0.08701    0.08194  -1.062
## HISPRACE2Other      0.82256    0.08788   9.360
## INTCTFAMNot Intact -0.61583    0.05362 -11.486
## EDUCMOM.L           0.59627    0.14481   4.118
## EDUCMOM.Q          -0.99603    0.12586  -7.914
## EDUCMOM.C          -0.49350    0.08275  -5.963
## EDUCMOM^4          -0.23277    0.05267  -4.420
##
## Intercepts:
##       Value    Std. Error t value
## 1|2     0.6613   0.1095     6.0382
## 2|3     0.9830   0.1088     9.0329
## 3|4     1.2296   0.1085    11.3299
## 4|5     2.2246   0.1101    20.2030
```

```
## 5|6     3.1571   0.1143    27.6332
## 6|7     3.4809   0.1157    30.0755
## 7|8     3.7430   0.1169    32.0222
## 8|9     5.1388   0.1242    41.3677
## 9|10    6.9035   0.1406    49.1129
## 10|11   7.8775   0.1634    48.2170
##
## Residual Deviance: 21710.97
## AIC: 21748.97
```

Multinomial Logistic Regression (FMARITAL)

```
## Call:
## polr(formula = HIEDUC ~ AGER + HISPRACE2 + INTCTFAM + EDUCMOM,
##     data = data, Hess = TRUE)
##
## Coefficients:
##                       Value Std. Error t value
## AGER                0.08940    0.00282  31.700
## HISPRACE2White      0.41712    0.06688   6.237
## HISPRACE2Black     -0.08701    0.08194  -1.062
## HISPRACE2Other      0.82256    0.08788   9.360
## INTCTFAMNot Intact -0.61583    0.05362 -11.486
## EDUCMOM.L           0.59627    0.14481   4.118
## EDUCMOM.Q          -0.99603    0.12586  -7.914
## EDUCMOM.C          -0.49350    0.08275  -5.963
## EDUCMOM^4          -0.23277    0.05267  -4.420
##
## Intercepts:
##       Value    Std. Error t value
## 1|2     0.6613   0.1095     6.0382
## 2|3     0.9830   0.1088     9.0329
## 3|4     1.2296   0.1085    11.3299
## 4|5     2.2246   0.1101    20.2030
## 5|6     3.1571   0.1143    27.6332
## 6|7     3.4809   0.1157    30.0755
## 7|8     3.7430   0.1169    32.0222
## 8|9     5.1388   0.1242    41.3677
## 9|10    6.9035   0.1406    49.1129
## 10|11   7.8775   0.1634    48.2170
##
## Residual Deviance: 21710.97
## AIC: 21748.97
```

Multiple Test

```
##
##  Kruskal-Wallis rank sum test
##
## data:  PREGNUM by HISPRACE2
## Kruskal-Wallis chi-squared = 38.536, df = 3, p-value = 2.177e-08

##
```

```
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  data$PREGNUM and data$HISPRACE2
##
##       Hispanic White   Black
## White 0.05785  -       -
## Black 0.01322  1.4e-05 -
## Other 0.00022  0.00563 4.2e-08
##
## P value adjustment method: fdr
```

Interpretation and Conclusions

Poisson Regression and Negative Binomial Model

The initial Poisson regression model was assessed for overdispersion by comparing the residual deviance to its degrees of freedom and examining the dispersion statistic. The presence of overdispersion was confirmed (dispersion statistic > 1). As a result, a Negative Binomial regression model was fitted to account for the extra-Poisson variation.

Model fit was compared using the Akaike Information Criterion (AIC). The Poisson model yielded an AIC of 17206.42, whereas the Negative Binomial model had a substantially lower AIC of 16495.71, indicating a better fit for the data when allowing for overdispersion. Based on these results, the Negative Binomial model was considered the more appropriate model for inference.

The negative binomial regression model revealed several important factors associated with the number of pregnancies among women. Compared to married women, those who were never married had 58% fewer pregnancies (IRR = 0.42), and those who were divorced had 22% fewer pregnancies (IRR = 0.78). Widowed women also had fewer pregnancies (IRR = 0.73).

In terms of race, Black women had 38% more pregnancies than Hispanic women (IRR = 1.38), while women classified as Other races had 15% fewer (IRR = 0.85). White women had slightly fewer pregnancies (IRR = 0.94).

Education showed a clear pattern: as education increased, the number of pregnancies generally decreased. For example, the main trend (HIEDUC.L) showed that each step up in education level was associated with a 50% decrease in the number of pregnancies (IRR = 0.50).

Age also had a strong positive effect: for each additional year of age, the expected number of pregnancies increased by about 5.6% (IRR = 1.06), as older women naturally have had more time to become pregnant.

Looking at background factors, women whose mothers had some college education had 25% more pregnancies (IRR = 1.25), while those whose mothers had less than high school had slightly fewer (IRR = 0.88). Additionally, women who did not grow up in an intact family had 30% more pregnancies than those who did (IRR = 1.30).

Ordinal Logistic Regression (HIEDUC)

An ordinal logistic regression revealed that age was positively associated with higher educational attainment (OR = 1.09, $p < .001$). Compared to the hispanics, White individuals (OR = 1.52, $p < .001$) and those of Other races (OR = 2.28, $p < .001$) had higher odds of achieving higher education levels. Coming from a non-intact family significantly reduced the odds of attaining higher education (OR = 0.54, $p < .001$). While maternal education showed a complex relationship, with both significant positive and negative coefficients.

Multinomial Logistic Regression (FMARITAL)

A multinomial logistic regression was used to examine how factors like age, education (both respondent's and mother's), race/ethnicity and family background affect marital status. The results showed that older age increases the chances of being widowed — for every additional year, the odds go up by 15% (OR = 1.15, p < .001). Black individuals were about 3.6 times more likely to be widowed (OR = 3.59, p = .009) compared to Hispanic individuals.

When it comes to being separated, White respondents were 3.6 times more likely (OR = 3.59, p = .050) and Black respondents were 1.9 times more likely (OR = 1.90, p = .005) than Hispanics. Additionally, people from non-intact families were over 4 times more likely to be separated (OR = 4.21, p < .001).

For those never married, each year of age slightly decreased the odds (OR = 0.88, p = .033). White individuals were much less likely to be never married (OR = 0.042, p < .001) compared to Hispanics. Higher maternal education also reduced the chances of never marrying — for example, a one-unit increase in a specific measure of maternal education cut the odds by about 67% (OR = 0.33, p = .011).

Pearson's Chi-squared Test

A Pearson's Chi-squared test of independence was conducted to examine the association between marital status (*FMARITAL*) and highest education level (*HIEDUC*). The test revealed a statistically significant association between the two variables, $\chi^2(40) = 804.43, p < 0.001$. This suggests that the distribution of education levels differs significantly across marital status categories.

Multiple Test

We also compared the total number of pregnancies among different race groups using a Kruskal-Wallis test, which showed a significant difference $\left(\chi^2 = 38.54, p < 0.001\right)$. To find out which of these groups differed, we conducted pairwise Wilcoxon tests with False Discovery Rate (FDR) adjustments to reduce false positives. Significant differences were found between Hispanic and Other $(p = 0.00022)$, White and Black $(p = 0.00001)$, Black and Other $(p < 0.00001)$, and Hispanic and Black $(p = 0.01322)$.

Limitation and Recommendation

One limitation of this study is that the data is cross-sectional. It captures information at a single point in time, so it cannot establish cause-and-effect relationships. Additionally, some variables like family structure and education levels are self-reported, which might introduce reporting bias. The sample also only includes women aged 15–49 in the United States, which limits the generalizability of the findings to other age groups or countries.

For future studies, it is recommended to use longitudinal data to better capture changes over time and establish causal relationships. Including more socio-economic and healthcare-related variables could help explain additional variability in pregnancy outcomes. Applying alternative modeling approaches such as machine learning algorithms may also improve predictive accuracy and uncover complex patterns not easily captured by traditional regression models.