# Modeling Household Poverty in Connecticut: A Data-Driven Analysis of Socioeconomic and Housing Determinants

Victor Cazabal

2024-12-01

## Introduction

In the wake of national outrage over rising inflation and mounting concerns about the cost of living over the past couple of years, understanding the factors that drive household poverty has never been more critical. As policymakers and community organizations seek effective strategies to alleviate economic hardship, it is essential to identify and quantify the social and economic determinants that most strongly influence whether a household lives below the poverty line. This report contributes to that understanding by examining household-level data from Connecticut sourced from the 2023 American Community Survey (ACS) Public Use Microdata Sample (PUMS), accessed via the tidycensus package in R.

In this analysis, we consider a variety of explanatory variables that reflect both structural and individual-level characteristics. These include the number of persons in a household (household size), single-parent status, educational attainment of the householder, race of the householder, housing cost burden, and health insurance coverage. By including measures of housing cost burden and healthcare access, we capture critical dimensions of financial strain that have proven especially salient in an era of increasing expenses and wage stagnation.

Through this approach, the report aims to inform both policy and practice by highlighting which characteristics most strongly predict household poverty, thereby guiding interventions that target the root causes of economic instability in the region.

## Exploratory Data Analysis

As mentioned previously, the dataset used in this study was extracted from the 2023 American Community Survey (ACS) Public Use Microdata Sample (PUMS) and includes individual and household-level data for residents across the entire state of Connecticut. The data, accessed via the tidycensus package in R, provides a comprehensive snapshot of demographic, socioeconomic, and housing-related characteristics, which are key to understanding poverty dynamics in the state.

That being said, the dataset used in this analysis contains 5000 rows and 8 columns, with each row representing a household head from Connecticut. The dataset includes a combination of variables capturing demographic, socioeconomic, and housing-related characteristics, all relevant to understanding poverty risk in the region. Below is an overview of the key variables included:

- **POV (Below Poverty or Not)**: A binary variable indicating whether a household is below the federal poverty line (1 = below poverty, 0 = not). This is the primary dependent variable for the analysis.
- **TEN (Tenure)**: Indicates whether the household is renter-occupied or owner-occupied. This variable provides insight into housing dynamics and socioeconomic status, as renters and owners often face different financial pressures.

- **HICOV (Health Insurance Coverage)**: A binary variable indicating whether any member of the household has health insurance (1 = insured, 2 = uninsured). Lack of health insurance can be a significant driver of financial instability, making this variable critical for understanding household vulnerability.
- **NP (Number of People in Household)**: A continuous variable representing household size. Larger households often face greater financial demands, and this variable helps capture the potential strain on household resources.
- **SP (Single Parent in Household or Not)**: A binary variable indicating whether the household is headed by a single parent (1 = single parent, 0 = not). Single-parent households are often more financially vulnerable, as they typically rely on a single income source.
- **HCB (Housing Cost Burden)**: A continuous variable capturing the percentage of household income spent on housing costs. This variable was calculated using two ACS variables:
  - GRPIP: Gross rent as a percentage of income, applicable to renter households.
  - OCPIP: Selected monthly owner costs as a percentage of income, applicable to owner households. The HCB variable combines these measures to create a unified indicator of housing cost burden, allowing analysis across both renters and owners.
- **SCHL (Educational Attainment of Householder)**: A categorical variable indicating the highest level of education completed by the household head. The categories include "High School or Less," "Some College or Associates," and "Bachelor's Degree or Higher." Education level often correlates with income and employment opportunities, making it an important factor in poverty analysis.
- **RACE (Race of Householder)**: A categorical variable indicating the race of the household head. The categories include "White," "Black or African American," "Asian," "Other Race," and "Two or More Races." This variable provides insight into racial disparities in economic vulnerability.

Table 1 below displays the summary statistics for the continuous variables in our analysis. In order to check for any outliers, we have displayed each of these variables graphically in a histogram which can be seen in the figure below Table 1.

Table 1: Summary of the Continuous Variables

| | mean | sd | median | iqr | min | max |
|---|---|---|---|---|---|---|
| NP | 2.40 | 1.36 | 2.00 | 2.00 | 1.00 | 12.00 |
| HCB | 29.32 | 25.43 | 21.00 | 22.00 | 0.00 | 101.00 |

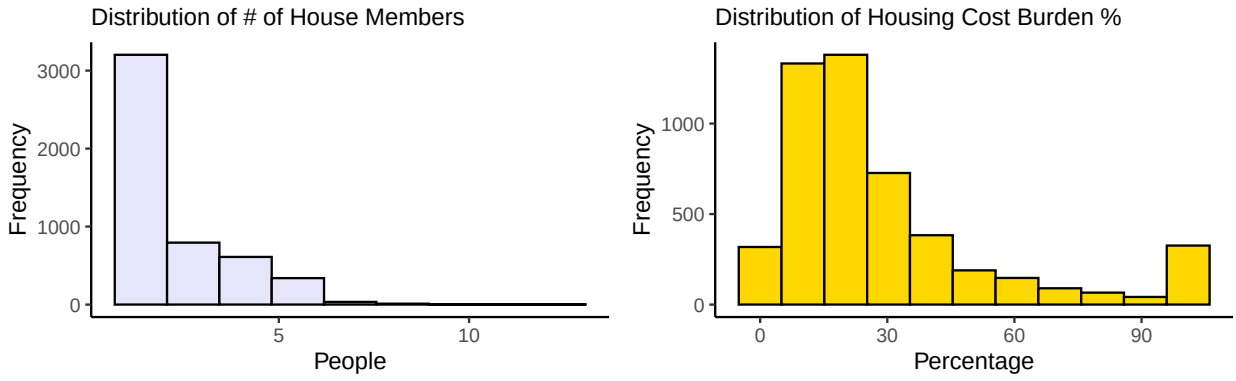


Figure 1: Histograms of Continuous Variables

The distribution of number of house members is unimodal, right skewed, with a couple of potential outliers. Similarly, the distribution of housing cost burden percentage is right skewed, but bimodal. This distribution doesn't have any significant outliers.

Figure 2 illustrates the distributions of our categorical variables. The barplot for poverty status highlights that while the majority of people are not living under the poverty line, a notable portion are. The tenure barplot reveals that most individuals in our sample own a home, though a significant number rent. Educational attainment data shows nearly half of head householders have a college degree or higher. Regarding race, the majority of individuals identify as white. Similarly, most households are not single-parent households, and the majority have health insurance coverage.

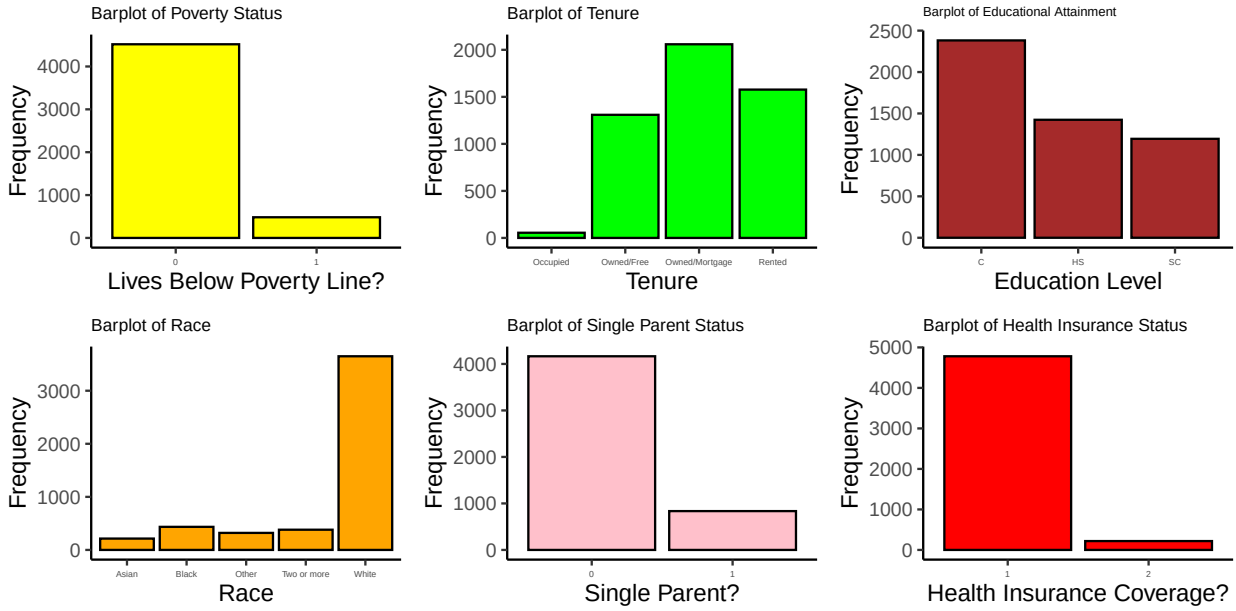


Figure 2: Barplots of Categorical Variables

Next, we explore the relationship between each continuous variable and our response variable. We show this relationship using boxplots.

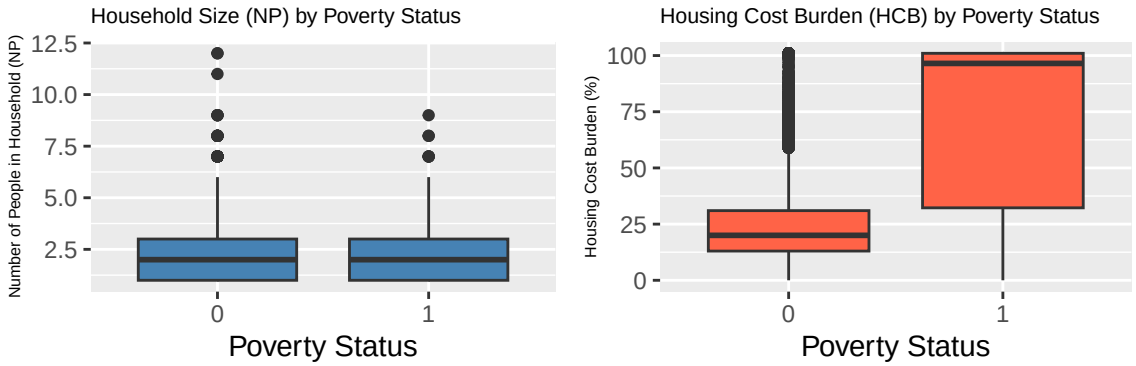


Figure 3: Boxplots of Continuous Variables

Surprisingly, we don't see a clear difference between number of people in households among poverty status. These boxplots are nearly identical. On the other hand, housing cost burden is very high for those living under poverty compared to those who aren't. This is to be expected, however it's interesting to see that the

3rd quartile of the first boxplot is the 1st quartile of the second. This goes to show how much housing cost burden might affect poverty status.

For our last part of our EDA, we explore the relationship of our categorical variables against the response variable.

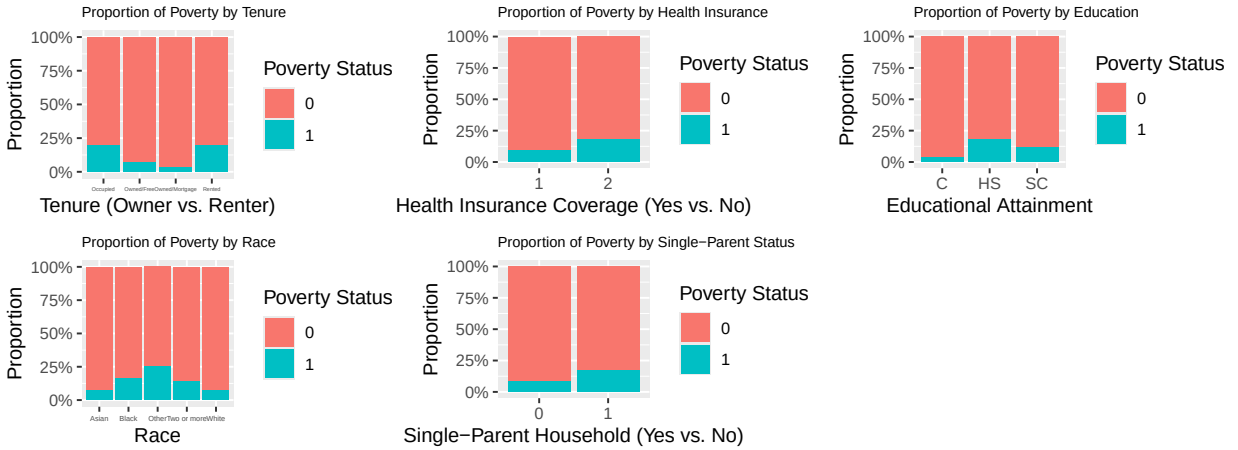


Figure 4: Barplots of Categorical Variables

The barplots reveal notable trends in poverty rates across various categorical variables. Poverty rates are significantly higher among renters and those occupying a living space without paying rent. Interestingly, individuals who own their homes outright experience higher poverty rates than those with a mortgage. As anticipated, poverty rates are elevated for those lacking health insurance. Educational attainment shows a clear pattern, with the lowest poverty rates among those with a college degree or higher and the highest among those with a high school degree or less. Poverty rates are also disproportionately higher among individuals identifying as Black, of another race, or of two or more races. Finally, households led by single parents exhibit significantly higher poverty rates compared to other household types.

## Modeling the data

Through a greedy method, we found that the variable "NP" was the only variable that didn't significantly affect the model fit. Below you can find when examining the impact of household size (NP) on the model, the likelihood ratio test returned a p-value of 0.4, indicating no significant improvement in fit after including this variable. In other words, NP does not appear to provide meaningful additional explanatory power for poverty status beyond the other predictors already in the model.

```
## Analysis of Deviance Table
##
## Model 1: POV ~ SP + SCHL + HCB + HICOV + TEN + RACE + NP
## Model 2: POV ~ SP + SCHL + HCB + HICOV + TEN + RACE
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4986     1959.6
## 2      4987     1960.3 -1  -0.6952   0.4044
```

Additionally, incorporating interaction terms revealed that the relationship between tenure (TEN) and housing cost burden (HCB), as well as the relationship between HCB and educational attainment (SCHL), significantly improved the model's fit. The likelihood ratio tests for these interactions yielded extremely low p-values, indicating that these combined effects provide meaningful additional insights into how socioeconomic factors influence poverty status.

```
## Analysis of Deviance Table
##
## Model 1: POV ~ SP + SCHL + HCB + HICOV + TEN + RACE
## Model 2: POV ~ SP + SCHL + HCB + HICOV + TEN + RACE + HCB * TEN + HCB *
##     SCHL
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4987     1960.3
## 2      4983     1887.9  4   72.394 7.085e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus, the logistic regression model that we will use to predict poverty status is as follows:

$\text{logit}(P(POV = 1)) = \beta_0 + \beta_1(SP) + \beta_2(SCHL) + \beta_3(HCB) + \beta_4(HICOV) + \beta_5(TEN) + \beta_6(RACE) + \beta_7(HCB \times TEN) + \beta_8(HCB \times SCHL)$

## Diagnostics

Model diagnostics were conducted to evaluate the performance and validity of the selected logistic regression model. This process ensures that the model appropriately fits the data and meets the assumptions underlying logistic regression. Below you will find a binned residual plot used to evaluate the fit of the logistic regression model.

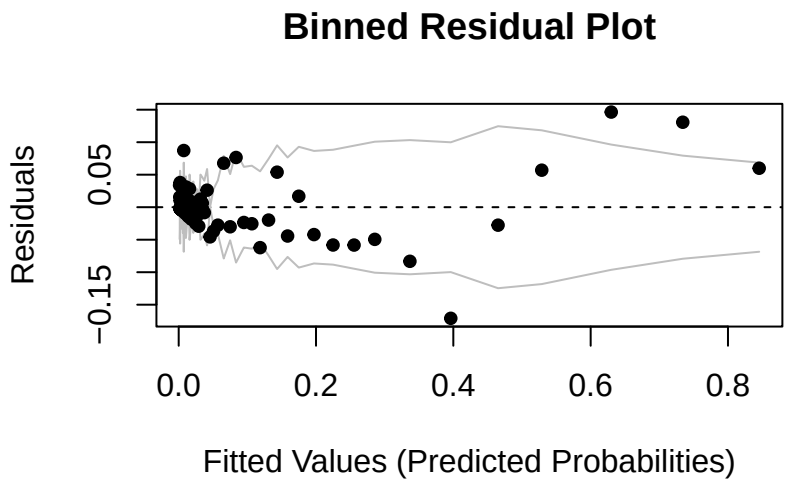


Figure 5: Binned Residual Plot

The binned residual plot shows most residuals falling within the expected bounds, with no evident systematic patterns. While a slight fanning shape is observed, this is consistent with the natural behavior of residuals in logistic regression and does not indicate a lack of model fit. Next, we plot the residuals vs categorical predictors.

```
## # A tibble: 4 x 2
##   TEN            mean_resid
##   <fct>               <dbl>
## 1 Owned/Mortgage  -7.69e-11
```

```
## 2 Occupied        -7.52e-11
## 3 Owned/Free      -7.71e-12
## 4 Rented          -1.42e-11
```

```
## # A tibble: 5 x 2
##   RACE        mean_resid
##   <fct>            <dbl>
## 1 White        -3.73e-11
## 2 Asian        -3.66e-11
## 3 Black        -3.82e-11
## 4 Other        -5.07e-11
## 5 Two or more  -4.77e-11
```

```
## # A tibble: 2 x 2
##   HICOV mean_resid
##   <fct>      <dbl>
## 1 1      -3.94e-11
## 2 2      -3.11e-11
```

```
## # A tibble: 3 x 2
##   SCHL  mean_resid
##   <fct>      <dbl>
## 1 C      -4.84e-11
## 2 HS     -3.85e-11
## 3 SC     -2.10e-11
```

```
## # A tibble: 2 x 2
##   SP    mean_resid
##   <fct>      <dbl>
## 1 0      -3.80e-11
## 2 1      -4.42e-11
```

Above, we have calculated average residual for each level of predictor. All means are close to 0, so we are confident residuals are not systematically biased for certain levels of categorical predictors. Finally, to better understand the interactions included in the model, I will explore how the relationship between the continuous variable (housing cost burden) and the predicted probability of poverty changes across the levels of the categorical variables (tenure and educational attainment). This will provide insights into the nuanced effects these predictors have when combined.

The interaction plots provide valuable insights into the relationships captured by the model and support the validity of the included interaction terms. In the plot of housing cost burden (HCB) by tenure (TEN), the predicted probability of poverty increases more rapidly for renters compared to homeowners at lower HCB values, as indicated by the higher y-intercept and steeper initial slope of the renter line. The homeowner line, in contrast, exhibits a more gradual and curved increase, with both lines converging to similar slopes around HCB values of 75. Similarly, in the plot of HCB by educational attainment (SCHL), the predicted probabilities for individuals with a high school education (HS) start slightly higher than for those with a college education (C), but both lines follow a curved pattern. The HS line increases at a faster rate initially, before aligning with the slope of the C line at around HCB values of 65. These patterns highlight the differential effects of HCB across categories of TEN and SCHL, validating the inclusion of these interaction terms in the model.

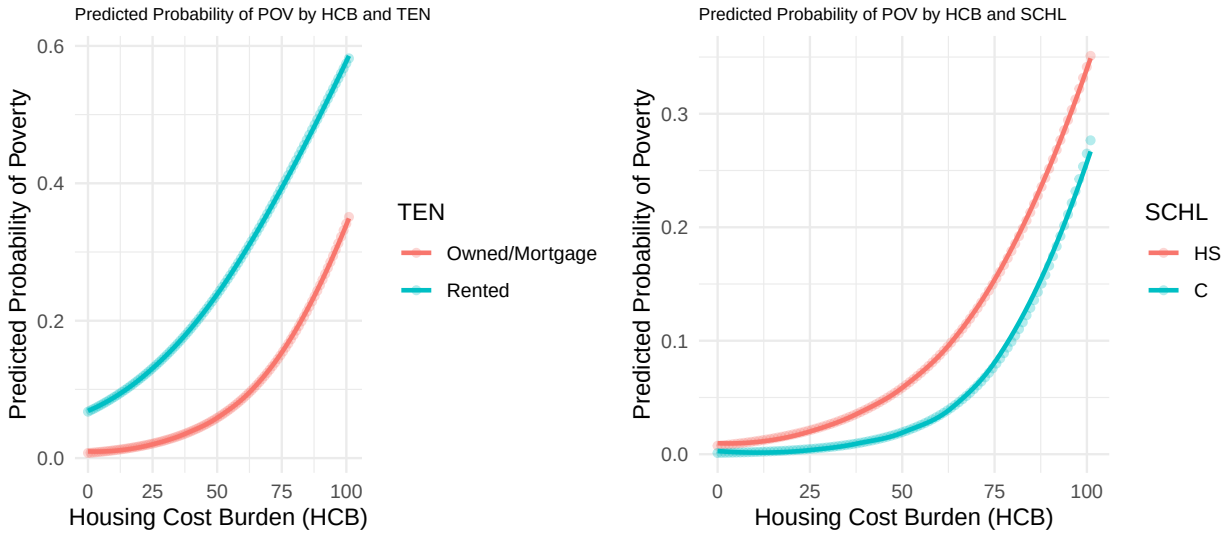


Figure 6: Binned Residual Plot

## Results

In this results section, we present the key findings from the fitted logistic regression model. After presenting the model summary, we report odds ratios and 95% confidence intervals for each predictor, providing an intuitive measure of how each factor influences the likelihood of poverty. Next, we evaluate the model's overall fit using a likelihood ratio test to confirm that the chosen predictors significantly improve the explanation of the data. Finally, we compare the predictive accuracy of the full model against the null model by examining classification error rates, offering insight into how well the model distinguishes between households below and above the poverty line.

Below you can find the model summary, stating estimates, standard errors and p-values for each predictor.

```
##
## Call:
## glm(formula = POV ~ SP + SCHL + HCB + HICOV + TEN + RACE + HCB *
##     TEN + HCB * SCHL, family = binomial(link = "logit"), data = ct_sample)
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -6.830470   0.391825 -17.432  < 2e-16 ***
## SP1                0.268274   0.139461   1.924 0.054398 .
## SCHLHS             1.935677   0.310867   6.227 4.76e-10 ***
## SCHLSC             0.398805   0.375109   1.063 0.287705
## HCB                0.058110   0.005165  11.252  < 2e-16 ***
## HICOV2            -0.214389   0.224635  -0.954 0.339887
## TENOccupied        4.401595   0.491580   8.954  < 2e-16 ***
## TENOwned/Free      0.030083   0.474996   0.063 0.949502
## TENRented          2.270215   0.365586   6.210 5.31e-10 ***
## RACEAsian         -0.047310   0.343427  -0.138 0.890430
## RACEBlack          0.247345   0.183820   1.346 0.178437
## RACEOther          0.774680   0.189613   4.086 4.40e-05 ***
## RACETwo or more    0.425087   0.203800   2.086 0.036996 *
## HCB:TENOccupied          NA         NA      NA       NA
## HCB:TENOwned/Free  0.021289   0.006694   3.180 0.001471 **
```

```
## HCB:TENRented     -0.013123   0.004962  -2.644 0.008182 **
## SCHLHS:HCB        -0.015731   0.004574  -3.439 0.000583 ***
## SCHLSC:HCB         0.004233   0.005320   0.796 0.426179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3171.0  on 4999  degrees of freedom
## Residual deviance: 1887.9  on 4983  degrees of freedom
## AIC: 1921.9
##
## Number of Fisher Scoring iterations: 7
```

For those coefficients with very small p-values the interpretations of their estimates are:

- **(Intercept)**: The negative intercept (-6.830) indicates that, when all predictors are at their baseline categories and set to zero, the log-odds of being below the poverty line are very low. In other words, in the reference scenario (e.g., college-educated, homeowner with mortgage, White race, etc.), the likelihood of poverty is quite small.
- **SCHLHS**: A coefficient of 1.9357 for high school or less education compared to college education suggests that having lower educational attainment substantially increases the log-odds of poverty. This translates into a much higher chance of being below the poverty line for households headed by individuals with a high school degree or less.
- **HCB (Housing Cost Burden)**: The coefficient 0.0581 means that for each one-unit increase in housing cost burden (e.g., a 1
- **TENOccupied (Occupying Without Payment)**: A large coefficient of 4.4016 indicates that households occupying a dwelling without paying rent are at a dramatically higher risk of poverty compared to the reference group (owners with a mortgage). This aligns with a very precarious housing situation often associated with severe economic hardship.
- **TENRented (Renting)**: The coefficient 2.2702 shows that renters also face higher log-odds of poverty compared to owners with a mortgage, though not as extreme as those who occupy without payment. Still, renting is strongly linked to increased economic vulnerability.
- **RACEOther**: With a coefficient of 0.7747, household heads identifying as "Other" race face higher log-odds of poverty than White household heads. Although less dramatic than some other predictors, this indicates a racial disparity in economic well-being.
- **HCB:TENOwned/Free (Interaction)**: The small positive coefficient (0.0213) suggests that for households owning their home free and clear, the effect of HCB on poverty differs slightly compared to the reference category. Essentially, as HCB increases, the impact on poverty odds changes marginally depending on this tenure status.
- **HCB:TENRented (Interaction)**: The negative coefficient (-0.0131) indicates that for renters, the relationship between HCB and poverty is slightly tempered compared to the reference category. While HCB still increases poverty risk, it does so at a somewhat reduced rate for renters.
- **SCHLHS:HCB (Interaction)**: The negative interaction (-0.0157) means that for households with a high school degree or less, the increase in poverty odds with rising HCB is somewhat muted compared to those with a college education. Although they start off at a higher risk, the incremental effect of housing cost burden is slightly less steep for this group.

In summary, these coefficients and their interactions confirm that a household's likelihood of poverty is shaped by a combination of education level, housing cost burden, housing tenure, and race. The interactions indicate that the effect of increasing housing cost burden varies depending on a household's education and tenure status, reinforcing the complexity of factors influencing economic vulnerability.

```
##                              Predictor          OR     CI_lower     CI_upper
## (Intercept)               (Intercept)  0.00108035 4.796906e-04 2.237288e-03
## SP1                               SP1  1.30770550 9.930567e-01 1.716074e+00
## SCHLHS                         SCHLHS  6.92873497 3.831809e+00 1.300863e+01
## SCHLSC                         SCHLSC  1.49004332 7.135885e-01 3.123336e+00
## HCB                               HCB  1.05983150 1.049385e+00 1.070903e+00
## HICOV2                         HICOV2  0.80703421 5.145020e-01 1.242804e+00
## TENOccupied               TENOccupied 81.58092549 3.083755e+01 2.147514e+02
## TENOwned/Free           TENOwned/Free  1.03053983 4.028655e-01 2.625197e+00
## TENRented                   TENRented  9.68148330 4.876304e+00 2.056990e+01
## RACEAsian                   RACEAsian  0.95379126 4.713935e-01 1.821286e+00
## RACEBlack                   RACEBlack  1.28062065 8.890272e-01 1.828738e+00
## RACEOther                   RACEOther  2.16989793 1.492307e+00 3.139876e+00
## RACETwo or more     RACETwo or more  1.52972384 1.019229e+00 2.268006e+00
## HCB:TENOccupied     HCB:TENOccupied          NA           NA           NA
## HCB:TENOwned/Free HCB:TENOwned/Free  1.02151727 1.008346e+00 1.035257e+00
## HCB:TENRented         HCB:TENRented  0.98696320 9.771877e-01 9.964342e-01
## SCHLHS:HCB               SCHLHS:HCB  0.98439231 9.754835e-01 9.931601e-01
## SCHLSC:HCB               SCHLSC:HCB  1.00424198 9.938315e-01 1.014817e+00
```

Some odds ratios worth mentioning include the odds ratio of 1.31 for the predictor "SP1". Holding all other predictors constant, this means that households with a single parent (SP = 1) are 31% more likely to be below the poverty line compared to households without a single parent (SP = 0).

Another is the odds ratio of 6.93 for the predictor SCHLHS. Holding all other predictors constant, households where the head of the household has a high school education or less are 6.93 times more likely to be below the poverty line compared to households where the head of the household has a college education.

Additionally, an odds ratio 1.06 for the predictor HCB means holding all other predictors constant, for every 1-unit increase in HCB (e.g., a 1% increase in the percentage of income spent on housing), the odds of being below the poverty line increase by 6%.

A notable odds ratio is 81.58 for the predictor TENOccupied, which means, holding all other predictors constant, households occupying a living space without paying rent are 81.58 times more likely to be below the poverty line compared to households that own their home with a mortgage. Similarly, holding all other predictors constant, households that rent are 9.68 times more likely to be below the poverty line compared to households that own their home with a mortgage.

Finally, holding all other predictors constant, households where the head identifies as "Other" race are 2.17 times more likely to be below the poverty line compared to households where the head identifies as White. Similarly, there's a 28% higher likelihood of poverty for Black-headed households compared to White-headed households.

Next, we evaluate the model's overall model fit using the following likelihood ratio test.

```
## Analysis of Deviance Table
##
## Model 1: POV ~ 1
## Model 2: POV ~ SP + SCHL + HCB + HICOV + TEN + RACE + HCB * TEN + HCB *
##     SCHL
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
```

```
## 1      4999     3171.0
## 2      4983     1887.9 16   1283.1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood ratio test comparing the null model (intercept-only) to the full model with all chosen predictors shows a highly significant improvement in fit. The difference in deviance between the two models is 1283.1 over 16 degrees of freedom, and the p-value is less than 2.2e-16, indicating that the included predictors collectively explain the variation in poverty status far better than a model with no predictors. In other words, the full model provides a substantially improved understanding of the factors influencing whether a household is below the poverty line.

Finally, we compare the predictive accuracy of the full model against the null model by calculating error rates.

```
## Null Model Error Rate: 0.0964
```

```
## Full Model Error Rate: 0.0632
```

The comparison of error rates between the null model and the full model provides evidence that incorporating the selected predictors enhances the model's predictive accuracy. The null model, which includes no predictors, misclassifies approximately 9.64% of households. By contrast, the full model with the chosen predictors and interaction terms reduces the error rate to about 6.32%. This improvement in classification performance indicates that the included factors add meaningful explanatory power, allowing the model to more accurately distinguish between households that are below the poverty line and those that are not.

## Conclusion

This analysis provides a comprehensive look at the socioeconomic and demographic factors that influence household poverty risk in Connecticut. By fitting a logistic regression model and incorporating key predictors—educational attainment, housing cost burden, race, tenure status, health insurance coverage, and single-parent household status—we identified clear patterns associated with a household's likelihood of living below the poverty line. The inclusion of interaction terms further highlighted that the impact of housing cost burden depends not only on the household's educational background but also on its tenure status.

The results clearly show that lower educational attainment, unstable or costly housing arrangements, and lack of health insurance coverage are closely tied to higher poverty risk. Additionally, households headed by individuals identifying as "Other" race, as well as those identifying as Black, are more vulnerable to poverty compared to their White counterparts. These findings underscore longstanding disparities and the need for targeted interventions.

For the state of Connecticut, understanding these relationships is vital. With cost-of-living and housing expenses rising, this research can inform policymakers, community organizations, and social service agencies in crafting strategies that address the structural vulnerabilities—such as high housing costs and inadequate educational support—that drive households into poverty. By focusing resources on improving educational access, ensuring affordable housing, and expanding health coverage, Connecticut can take meaningful steps toward reducing poverty and fostering long-term economic stability for its residents.