Using the GSS to Visualize People’s Opinions on Traditional Gender Roles in the United States

Asma Syed

2023-12-16

# About the Dataset

## Study Population

The data set that I used was extracted from the General Social Survey (GSS) Data Explorer database. I selected my variables of interest after considering their possible influence on people’s opinions on traditional gender roles. After siphoning the data and loading it on R studio, the data set consisted of 72,390 observations and 20 variables. Since I wanted to exclusively focus my analysis on the population surveyed in the year 2022, I filtered the original data to exclude all other years from 1972 to 2021. The result was a data set that had a total of 3,544 respondents from the 2022 survey.

The table below displays the counts of the respondents who were administered either Ballot A, B, or C:

Counts of the Ballots Adminstered in 2022

| BALLOT | count |
| --- | --- |
| 1 | 1173 |
| 2 | 1203 |
| 3 | 1168 |

Considering that the response variables, which were the survey questions regarding traditional gender roles, were asked to respondents administered Ballot A, I refined the data set even further to include just Ballot A recipients. This produced a data set of 1,173 observations. Lastly, I decided on removing the variable RELIG (religion) from my analysis since its data was unavailable on the GSS database. I also decided to exclude respondents who answered with a “Don’t know/unsure” on the two survey questions (the response variables) and the variable MARTYPE. My final omissions led to a study population of 803 observations and 14 variables. The explanatory variables that were included in the analysis was as follows:

1. AGE
2. BORN
3. DEGREE
4. INCOME
5. HRS1
6. MARITAL
7. MARTYPE
8. POLVIEWS
9. RACE
10. REGION
11. SEX
12. WRKSTAT

Data management was my next steps.

# Data Management

I first started by re-coding the negative codes (-100, -99, -98 and -97) which were assigned values for missing or unanswered responses. Next, I re-categorized the explanatory variables, their new categorizations are shown in the count tables.

In addition, I re-coded the response variables, HUBBYWK1 and FAMSUFFR, both of which were coded on a Likert scale of 1-5: (1) Strongly Agree, (2) Agree, (3) **Neutral responses removed** (4) Disagree, (5) Strongly Disagree. I condensed all the agree and disagree responses into two separate groups: *1* for *agree* and *2* for *disagree*. The table below displays the counts under each category, agree and disagree, as well as the missing values for the two response variables.

Counts For The Variable HUBBYWK1

| HUBBYWK1 | count |
| --- | --- |
| 1 | 142 |
| 2 | 627 |
| NA | 34 |

Counts For The Variable FAMSUFFR

| FAMSUFFR | count |
| --- | --- |
| 1 | 147 |
| 2 | 622 |
| NA | 34 |

## Descriptive Statistics

Below are tables representing the counts

Counts For The Variable AGE

| AGE | count |
| --- | --- |
| 20 to 29 Years Old | 123 |
| 30 to 39 Years Old | 144 |
| 40 to 49 Years Old | 131 |
| 50 to 59 Years Old | 113 |
| 60 to 69 Years Old | 130 |
| 70 to 79 Years Old | 75 |
| 80 to 89 Years Old | 24 |
| NA | 63 |

Counts For The Variable BORN

| BORN | count |
| --- | --- |
| Born in the US | 707 |
| Not Born in the US | 92 |
| NA | 4 |

Counts For The Variable DEGREE

| DEGREE | count |
| --- | --- |
| Associate Degree | 64 |
| Bachelor’s | 193 |
| Graduate | 129 |
| High School | 344 |
| Less than High School | 73 |

Counts For The Variable INCOME

| INCOME | count |
| --- | --- |
| $25,000 or More | 566 |
| From $1,000 to $4,999 | 11 |
| From $10,000 to $24,999 | 104 |
| From $5,000 to $9,999 | 16 |
| Under $1,000 | 6 |
| NA | 100 |

Counts For The Variable HRS1

| HRS1 | count |
| --- | --- |
| Worked 0-15 Hours | 29 |
| Worked 16-30 Hours | 57 |
| Worked 31-40 Hours | 236 |
| Worked 41-50 Hours | 114 |
| Worked 51-60 Hours | 45 |
| Worked 61-75 Hours | 13 |
| Worked 76-89 Hours | 9 |
| NA | 300 |

Counts For The Variable MARITAL

| MARITAL | count |
| --- | --- |
| Divorced | 133 |
| Married | 319 |
| Never Married | 272 |
| Seperated | 21 |
| Widowed | 55 |
| NA | 3 |

Counts For The Variable MARTYPE

| MARTYPE | count |
| --- | --- |
| Marriage between a man and woman | 164 |
| Same Sex Marriage | 1 |
| NA | 638 |

Counts For The Variable POLVIEWS

| POLVIEWS | count |
| --- | --- |
| Conservative | 102 |
| Extremely Conservative | 36 |
| Extremely Liberal | 68 |
| Liberal | 134 |
| Moderate | 265 |
| Slightly Conservative | 80 |
| Slightly Liberal | 95 |
| NA | 23 |

Counts For The Variable RACE

| RACE | count |
| --- | --- |
| Black | 145 |
| Other | 71 |
| White | 576 |
| NA | 11 |

Counts For The Variable REGION

| REGION | count |
| --- | --- |
| Middle Atlantic | 84 |
| Midwest | 183 |
| Mountain | 70 |
| New England | 40 |
| Pacific | 120 |
| South | 126 |
| South Atlantic | 180 |

Counts For The Variable SEX

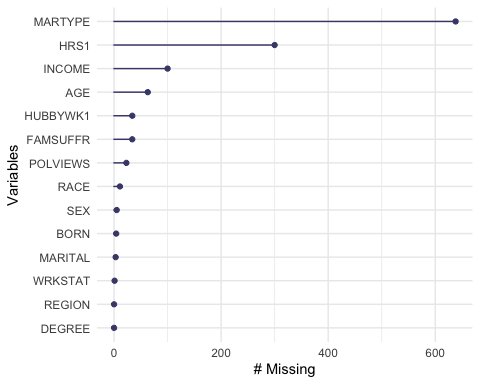
| SEX | count |
| --- | --- |
| Female | 419 |
| Male | 379 |
| NA | 5 |

Counts For The Variable WRKSTAT

| WRKSTAT | count |
| --- | --- |
| In School | 22 |
| Keeping House | 44 |
| Other | 35 |
| Retired | 147 |
| Unemployed, Laid Off, Looking for Work | 32 |
| With Job But Not at Work Due to Certain Reasons | 18 |
| Working Full-Time | 425 |
| Working Part-Time | 79 |
| NA | 1 |

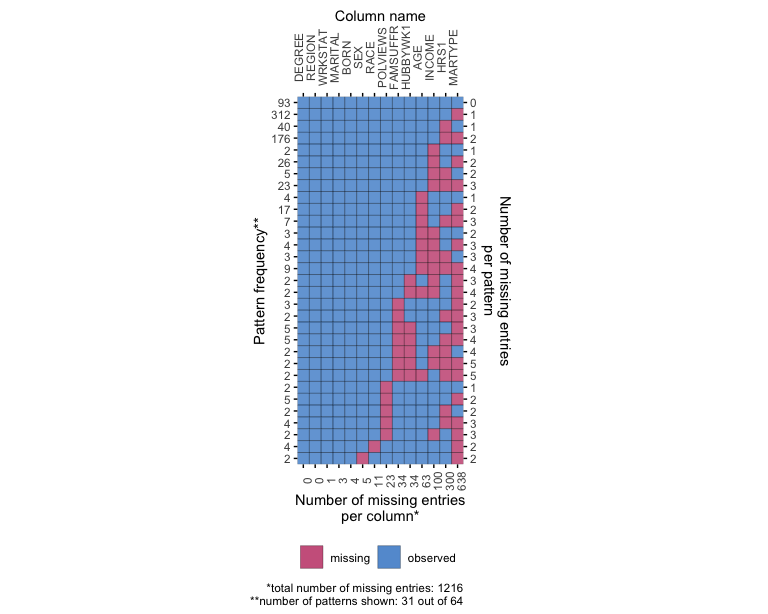
## Multiple Imputation For Missing Values

For the missing values in my data set, I decided to perform multiple imputation using the MICE function in R. Before performing the imputation, I first converted all the variables into factors, and then built some visuals to assess how many missing values there were and whether there was any pattern.

The Missing Variables Plot below shows the number of missing values under each variable:  As apparent from the plot, the variable with the highest number of missing values is MARTYPE, with over 600 missing values in comparison to 165 complete cases. Despite the large missing to complete cases ratio, I decided to proceed with imputation on this variable.

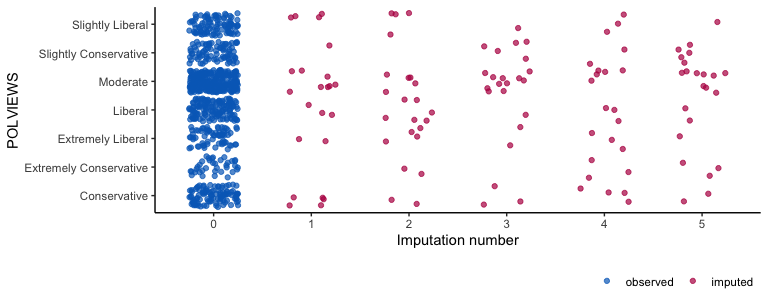
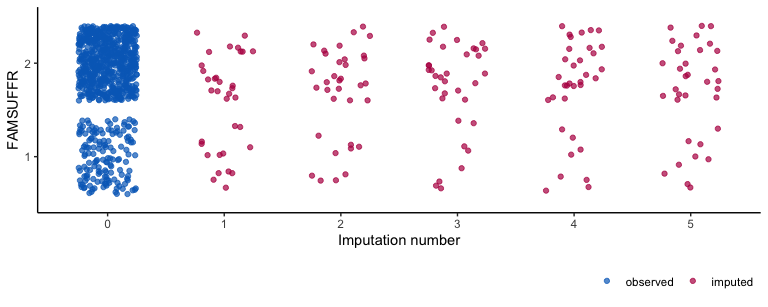
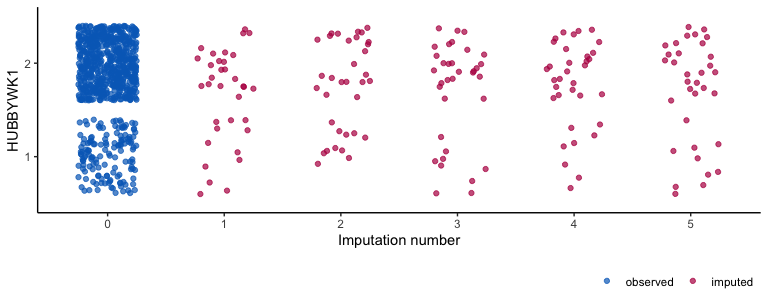
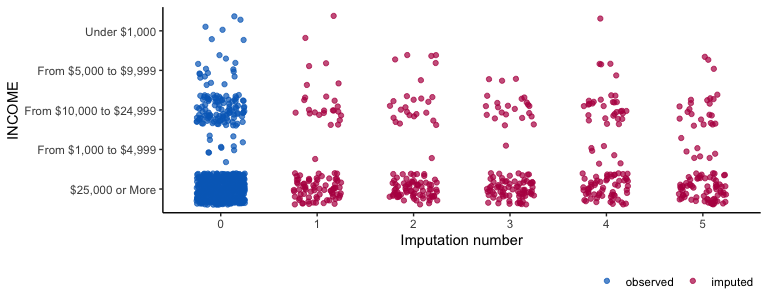
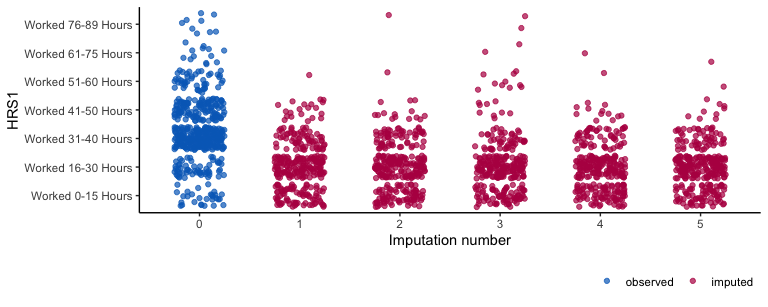
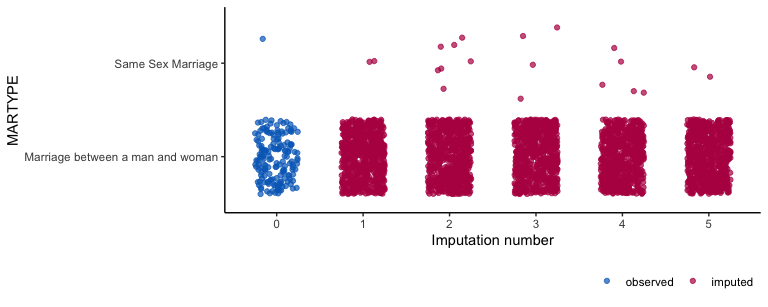
Next, I created a pattern plot which visualizes the number of missing entries and their pattern frequency throughout the data. The plot displays 25 patterns that were found within the data set. Essentially, the plot can be read both horizontally and vertically. First, each row is analyzed to determine which variables have missing values (pink squares). The numbers on the right of the plot directly correlate with the number of pink squares along that row. The values on the left represent the number of observations that display that particular missing pattern frequency of that row.

For example, in the plot, the first row is completely blue; there are no missing values. The 93 on the left represents that 93 observations have 0 missing values. The second row is read as follows: 317 observations have missing values under the variable MARTYPE. The fourth row: 178 observations have missing values under both HRS1 and MARTYPE. The last row is read as follows: 2 observations have missing values under POLVIEWS and MARTYPE. The numbers along the bottom of the plot represent how many missing values each variable has.



Then I performed multiple imputation on the data; I used the Predictive Mean Matching Method for the imputations. I created 5 data sets with 25 iterations which were then pooled into one data set.

After obtaining the imputed model, I created split plots for the variables that had missing values. The split plots help determine how well the imputed data fits with the complete cases.



# Statistical Analysis

## Question 1

My hypothesis was as follows: There is a relationship between People’s views on traditional gender roles and their demographics. The variables that I hypothesize as having the strongest relationship with the response are: POLVIEWS, BORN, RACE, AGE, INCOME, WRKSTAT and REGION.

To test my hypothesis, I built a logistic regression model with the imputed data. My first model tests the significance between the response, HUBBYWK1 and the explanatory variables.

Estimates from Log Reg Model of Response HUBBYWK1

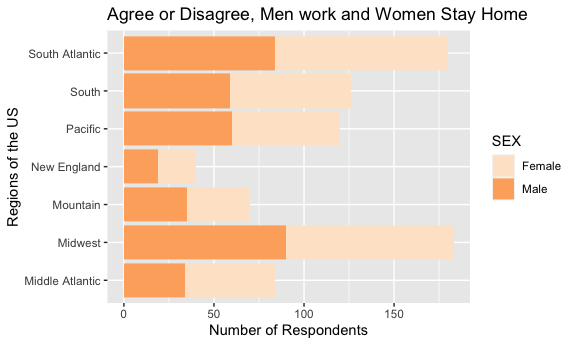
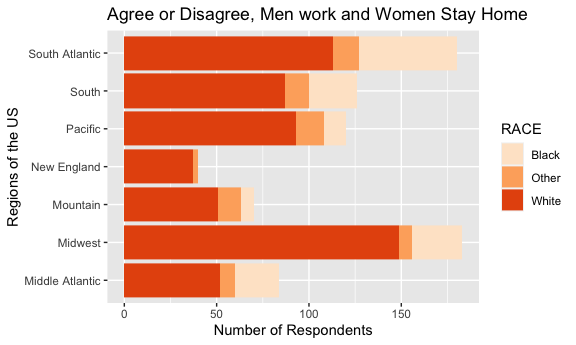
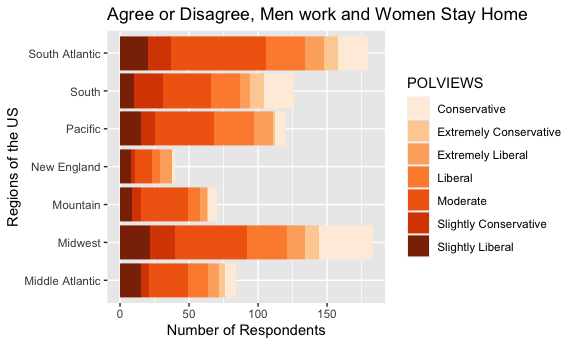
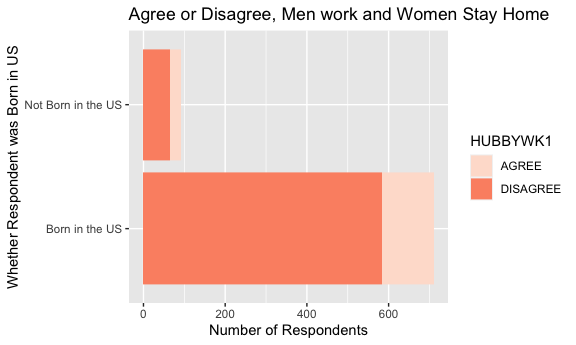
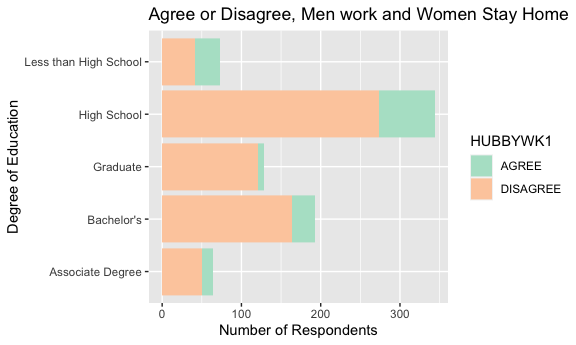
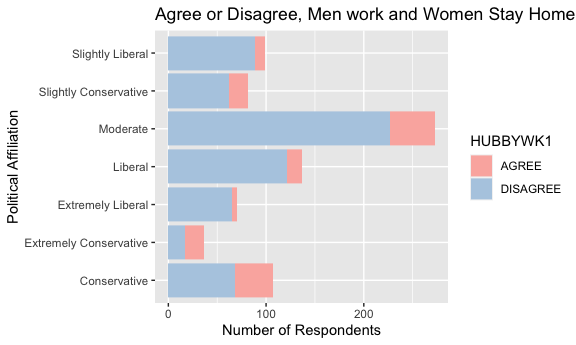
| term | estimate | std.error | statistic | df | p.value |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | 1.379 | 1.299 | 1.061 | 47.843 | 0.294 |
| WRKSTATKeeping House | -1.444 | 0.900 | -1.604 | 487.392 | 0.109 |
| WRKSTATOther | -0.008 | 1.098 | -0.007 | 58.129 | 0.994 |
| WRKSTATRetired | -1.341 | 0.888 | -1.510 | 661.650 | 0.131 |
| WRKSTATUnemployed, Laid Off, Looking for Work | 0.208 | 1.081 | 0.192 | 98.271 | 0.848 |
| WRKSTATWith Job But Not at Work Due to Certain Reasons | -0.096 | 1.213 | -0.079 | 467.213 | 0.937 |
| WRKSTATWorking Full-Time | -0.579 | 1.006 | -0.576 | 41.451 | 0.568 |
| WRKSTATWorking Part-Time | -0.602 | 0.873 | -0.689 | 440.305 | 0.491 |
| HRS1Worked 16-30 Hours | 0.350 | 0.536 | 0.652 | 12.806 | 0.526 |
| HRS1Worked 31-40 Hours | -0.199 | 0.818 | -0.243 | 6.950 | 0.815 |
| HRS1Worked 41-50 Hours | 0.526 | 0.862 | 0.610 | 9.434 | 0.556 |
| HRS1Worked 51-60 Hours | 0.284 | 0.974 | 0.291 | 12.874 | 0.776 |
| HRS1Worked 61-75 Hours | 0.013 | 1.248 | 0.010 | 20.229 | 0.992 |
| HRS1Worked 76-89 Hours | 0.183 | 1.598 | 0.114 | 24.305 | 0.910 |
| MARITALMarried | -0.305 | 0.354 | -0.863 | 527.358 | 0.389 |
| MARITALNever Married | -0.634 | 0.398 | -1.596 | 313.326 | 0.112 |
| MARITALSeperated | -0.578 | 0.650 | -0.889 | 473.623 | 0.374 |
| MARITALWidowed | -0.662 | 0.514 | -1.288 | 163.605 | 0.200 |
| AGE30 to 39 Years Old | 0.296 | 0.425 | 0.696 | 65.316 | 0.489 |
| AGE40 to 49 Years Old | 0.241 | 0.453 | 0.533 | 55.884 | 0.596 |
| AGE50 to 59 Years Old | -0.165 | 0.535 | -0.307 | 26.764 | 0.761 |
| AGE60 to 69 Years Old | 1.032 | 0.573 | 1.802 | 42.312 | 0.079 |
| AGE70 to 79 Years Old | -0.036 | 0.617 | -0.059 | 134.388 | 0.953 |
| AGE80 to 89 Years Old | -0.782 | 0.745 | -1.049 | 231.341 | 0.295 |
| DEGREEBachelor’s | 0.211 | 0.428 | 0.494 | 577.874 | 0.621 |
| DEGREEGraduate | 1.402 | 0.553 | 2.533 | 512.549 | 0.012 |
| DEGREEHigh School | 0.295 | 0.396 | 0.746 | 435.523 | 0.456 |
| DEGREELess than High School | -0.525 | 0.462 | -1.136 | 743.481 | 0.256 |
| SEXMale | -0.583 | 0.240 | -2.430 | 231.922 | 0.016 |
| RACEOther | 0.786 | 0.462 | 1.701 | 418.747 | 0.090 |
| RACEWhite | 0.696 | 0.317 | 2.197 | 96.430 | 0.030 |
| BORNNot Born in the US | -0.930 | 0.348 | -2.677 | 593.389 | 0.008 |
| INCOMEFrom $1,000 to $4,999 | -1.715 | 0.889 | -1.929 | 26.619 | 0.064 |
| INCOMEFrom $10,000 to $24,999 | -0.010 | 0.388 | -0.026 | 32.105 | 0.979 |
| INCOMEFrom $5,000 to $9,999 | -0.811 | 0.915 | -0.886 | 22.868 | 0.385 |
| INCOMEUnder $1,000 | 0.454 | 1.518 | 0.299 | 22.567 | 0.768 |
| REGIONMidwest | -0.194 | 0.474 | -0.408 | 336.016 | 0.683 |
| REGIONMountain | -0.268 | 0.568 | -0.472 | 258.229 | 0.637 |
| REGIONNew England | 0.436 | 0.908 | 0.481 | 437.553 | 0.631 |
| REGIONPacific | -0.869 | 0.513 | -1.694 | 218.781 | 0.092 |
| REGIONSouth | -1.002 | 0.482 | -2.079 | 213.789 | 0.039 |
| REGIONSouth Atlantic | -0.677 | 0.479 | -1.415 | 150.346 | 0.159 |
| POLVIEWSExtremely Conservative | -0.402 | 0.491 | -0.819 | 105.105 | 0.414 |
| POLVIEWSExtremely Liberal | 2.563 | 0.705 | 3.636 | 78.390 | 0.000 |
| POLVIEWSLiberal | 1.904 | 0.421 | 4.526 | 225.500 | 0.000 |
| POLVIEWSModerate | 1.408 | 0.347 | 4.058 | 92.810 | 0.000 |
| POLVIEWSSlightly Conservative | 0.618 | 0.394 | 1.568 | 288.969 | 0.118 |
| POLVIEWSSlightly Liberal | 2.124 | 0.500 | 4.251 | 199.907 | 0.000 |
| MARTYPESame Sex Marriage | 8.867 | 440.707 | 0.020 | 751.769 | 0.984 |

When using 0.05 as the significance level, the following variables are shown to have a significant relationship with the response, HUBBYWK1:

* DEGREE (Graduate) -> p.value (0.012)
* SEX (Male) -> p.value (0.016)
* RACE (White) -> p.value (0.030)
* BORN (Not born in US) -> p.value(0.008)
* REGION (South) -> p.value(0.039)
* POLVIEWS (Liberal) -> p.value(< 0.001)
* POLVIEWS (Extremely Liberal) -> p.value(< 0.001)
* POLVIEWS (Slightly Liberal) -> p.value(< 0.001)
* POLVIEWS (Moderate) -> p.value(< 0.001)

After computing the Hosmer-Lemeshow Goodness of Fit test on the first model, the test results show that the logistic regression model is a good fit for the data, since the p.value is larger than 0.05 and it is pretty close to 1.

## Combination of Chi Square Statistics for Multiply Imputed Data  
## Using 5 Imputed Data Sets  
## F(8, 62.6)=0.423 p=0.90286

Below are stacked bar plots which represent the significant explanatory variables: 

Note the larger margin of discrepancy among the people who identify themselves under the umbrella of the Liberal political party and their opinions on the response variable. There is a significantly larger proportion of Liberals who disagree with the survey question than those who agree. Also note a similar pattern under the GRADUATE level of the variable DEGREE. People who have a graduates degree more often disagree with traditional gender roles than those who agree. Similarly, for the variable BORN, even though a smaller portion of the population consists of people not born in the US, a large amount of them disagree with gender roles compared to those who do. Finally, in order to assess what demographics may contribute to the significance of the variable REGION, the explanatory variable has been compared with certain demographics like RACE, POLVIEWS, and SEX. It seems that other than the fact that there are a larger proportion of white people in the South compared to any other race, none of the other variables, POLVIEWS or SEX, seem to contribute to the reason why that particular level has the most significance out of all the US regions. This alludes to another, more important demographic or factor here which could potentially be lending a part into why the South has a significance with the response.

## Question 2

Finally, I created a second logestic regression model for the second response, FAMSUFFR and it yielded the following results:

Estimates from Log Reg Model of Response FAMSUFFR

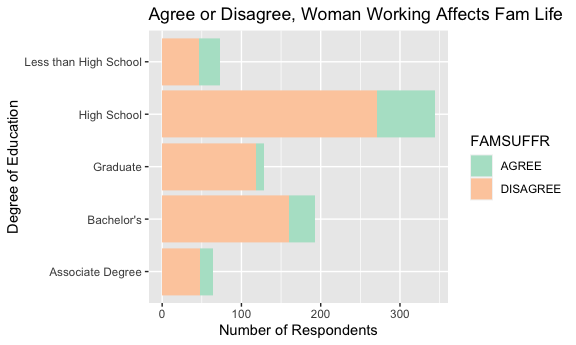
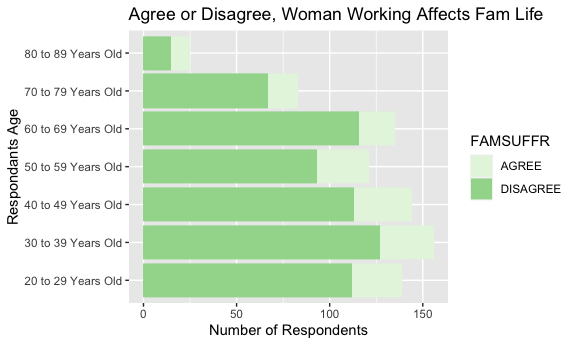
| term | estimate | std.error | statistic | df | p.value |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | 0.297 | 0.988 | 0.300 | 305.326 | 0.764 |
| WRKSTATKeeping House | -0.230 | 0.713 | -0.323 | 729.283 | 0.747 |
| WRKSTATOther | 0.564 | 0.823 | 0.685 | 239.520 | 0.494 |
| WRKSTATRetired | -0.275 | 0.724 | -0.379 | 553.057 | 0.704 |
| WRKSTATUnemployed, Laid Off, Looking for Work | 1.577 | 0.952 | 1.656 | 229.765 | 0.099 |
| WRKSTATWith Job But Not at Work Due to Certain Reasons | 0.692 | 1.018 | 0.680 | 740.399 | 0.497 |
| WRKSTATWorking Full-Time | 0.394 | 0.790 | 0.499 | 47.109 | 0.620 |
| WRKSTATWorking Part-Time | -0.298 | 0.673 | -0.443 | 718.375 | 0.658 |
| HRS1Worked 16-30 Hours | 0.531 | 0.384 | 1.382 | 60.364 | 0.172 |
| HRS1Worked 31-40 Hours | 0.019 | 0.816 | 0.023 | 6.332 | 0.982 |
| HRS1Worked 41-50 Hours | 0.488 | 0.833 | 0.586 | 8.396 | 0.573 |
| HRS1Worked 51-60 Hours | -0.252 | 0.909 | -0.277 | 10.074 | 0.787 |
| HRS1Worked 61-75 Hours | 0.577 | 1.205 | 0.479 | 22.401 | 0.637 |
| HRS1Worked 76-89 Hours | 0.048 | 1.384 | 0.034 | 41.005 | 0.973 |
| MARITALMarried | -0.119 | 0.312 | -0.380 | 626.481 | 0.704 |
| MARITALNever Married | -0.039 | 0.363 | -0.108 | 302.602 | 0.914 |
| MARITALSeperated | -0.601 | 0.623 | -0.965 | 361.203 | 0.335 |
| MARITALWidowed | -0.412 | 0.513 | -0.803 | 64.130 | 0.425 |
| AGE30 to 39 Years Old | 0.047 | 0.365 | 0.129 | 446.021 | 0.897 |
| AGE40 to 49 Years Old | -0.062 | 0.372 | -0.168 | 631.456 | 0.867 |
| AGE50 to 59 Years Old | -0.312 | 0.427 | -0.730 | 184.892 | 0.466 |
| AGE60 to 69 Years Old | 1.042 | 0.471 | 2.210 | 505.085 | 0.028 |
| AGE70 to 79 Years Old | 0.363 | 0.578 | 0.628 | 203.581 | 0.531 |
| AGE80 to 89 Years Old | -0.537 | 0.719 | -0.747 | 153.576 | 0.456 |
| DEGREEBachelor’s | 0.419 | 0.405 | 1.035 | 406.786 | 0.301 |
| DEGREEGraduate | 1.227 | 0.479 | 2.560 | 707.853 | 0.011 |
| DEGREEHigh School | 0.314 | 0.374 | 0.841 | 281.870 | 0.401 |
| DEGREELess than High School | -0.128 | 0.471 | -0.272 | 168.301 | 0.786 |
| SEXMale | -0.532 | 0.224 | -2.375 | 237.822 | 0.018 |
| RACEOther | 0.902 | 0.473 | 1.904 | 196.676 | 0.058 |
| RACEWhite | 0.180 | 0.278 | 0.646 | 502.508 | 0.519 |
| BORNNot Born in the US | -1.096 | 0.337 | -3.253 | 144.980 | 0.001 |
| INCOMEFrom $1,000 to $4,999 | -0.637 | 0.822 | -0.775 | 49.731 | 0.442 |
| INCOMEFrom $10,000 to $24,999 | -0.176 | 0.380 | -0.463 | 24.967 | 0.647 |
| INCOMEFrom $5,000 to $9,999 | -0.957 | 0.783 | -1.222 | 41.016 | 0.229 |
| INCOMEUnder $1,000 | -0.400 | 1.240 | -0.322 | 53.642 | 0.748 |
| REGIONMidwest | 0.266 | 0.382 | 0.695 | 576.490 | 0.487 |
| REGIONMountain | 0.239 | 0.458 | 0.521 | 738.981 | 0.602 |
| REGIONNew England | 0.077 | 0.578 | 0.134 | 736.036 | 0.894 |
| REGIONPacific | 0.258 | 0.419 | 0.616 | 697.502 | 0.538 |
| REGIONSouth | 0.139 | 0.402 | 0.347 | 460.832 | 0.729 |
| REGIONSouth Atlantic | 0.393 | 0.378 | 1.040 | 685.473 | 0.299 |
| POLVIEWSExtremely Conservative | -0.325 | 0.485 | -0.670 | 163.741 | 0.504 |
| POLVIEWSExtremely Liberal | 1.943 | 0.637 | 3.048 | 315.907 | 0.002 |
| POLVIEWSLiberal | 1.040 | 0.386 | 2.697 | 670.893 | 0.007 |
| POLVIEWSModerate | 0.264 | 0.314 | 0.840 | 234.041 | 0.402 |
| POLVIEWSSlightly Conservative | 0.137 | 0.389 | 0.353 | 574.996 | 0.724 |
| POLVIEWSSlightly Liberal | 0.888 | 0.427 | 2.081 | 284.626 | 0.038 |
| MARTYPESame Sex Marriage | 0.210 | 226.935 | 0.001 | 751.323 | 0.999 |

The following variables are shown to have the most significance with FAMSUFFR:

* AGE (60 to 69 Years) -> p.value (0.028)
* DEGREE (Graduate) -> p.value (0.011)
* SEX (Male) -> p.value (0.018)
* BORN (Not born in US) -> p.value (0.001)
* POLVIEWS (Liberal) -> p.value (0.007)
* POLVIEWS (Extremely Liberal) -> p.value (0.002)
* POLVIEWS (Slightly Liberal) -> p.value (0.038)

After computing the Hosmer Lemeshow Goodness of fit test, I can conclude that the insignificant p-value shows that the logistic regression model is a good fit for our data.

## Combination of Chi Square Statistics for Multiply Imputed Data  
## Using 5 Imputed Data Sets  
## F(8, 100.24)=0.729 p=0.66602

Below is a bar graph of the variables AGE and DEGREE with the response, FAMSUFFR: 

The variable, DEGREE, here shows a similar result as was shown for the bar plot of the variable DEGREE and the response FAMSUFFR. We can deduce here then that people who have a graduates degree tend to oppose traditional gender roles at a larger discrepancy in comparison to people who don’t have a graduates degree. Also, since the variable POLVIEWS, particularly the Liberal levels, were highly significant, we can determine that people who fall under the Liberal political affilation are strongly opposed to traditional gender views in comparison to people who identify as conservataives, as there seems to be a mix of people who agree and disagree.

# Conclusion

In conclusion, my hypothesis was partially correct. While I was unable to use the variable RELIG in my analysis, the assumption that some of the other variables that I predicted would have a significant relationship with the response variables HUBBYWK1, FAMSUFFR turned out be true. This is namely the case for the variables POLVIEWS, AGE, BORN, RACE, and REGION. In other words, we can conclude the following: ”Certain demographics, like a person’s political views, their race, sex, age, the highest degree they earned, and whether they were born in the US, are all factors that contribute to their opinion on traditional gender roles in the US.”