The Income-Childbirth Paradox: A Statistical Analysis of NHANES Data

Aung Nyein

Background/Motivation for the Study

Birth rate has been decreasing in many developed countries (Nargund G. 2009). Despite having some of the highest GDP per capita in the world, countries like Japan and Korea have some of the lowest birth rates. Couples who choose not to have children or have only very few children often state that the reason is because they cannot afford the cost of living for the extra individuals. The motivation for this study is to find out the effects that income have on the number of children middle age females have. This would give insights to help policy makers come up with the right tools to address this issue.

Research Question and Hypothesis

For females between ages 30 and 60 years in the United States, what is the relationship between their household income and the number of pregnancies they have had? Based on the reasoning that people cannot afford the cost of living for new individuals, it is expected that income and the number of pregnancies will have a positive correlation. I am trying to show the effect, if any, of income on the number of children people have.

Data Description and Exploratory Data Analysis

The targeted variables are middle value of total annual gross income in each category (HHIncomeMid), and the number of times participants have been pregnant (nPregnanceis). The data set is filtered to include only data for females because the data for pregnancies is only available for females. The range for age starts at 30 years although the data set includes ages starting 20 years old because it is assumed that it is more common for females to have more children after 20 than they do after 30. And the range stops at 60 years old because income might be compromised after retirement.

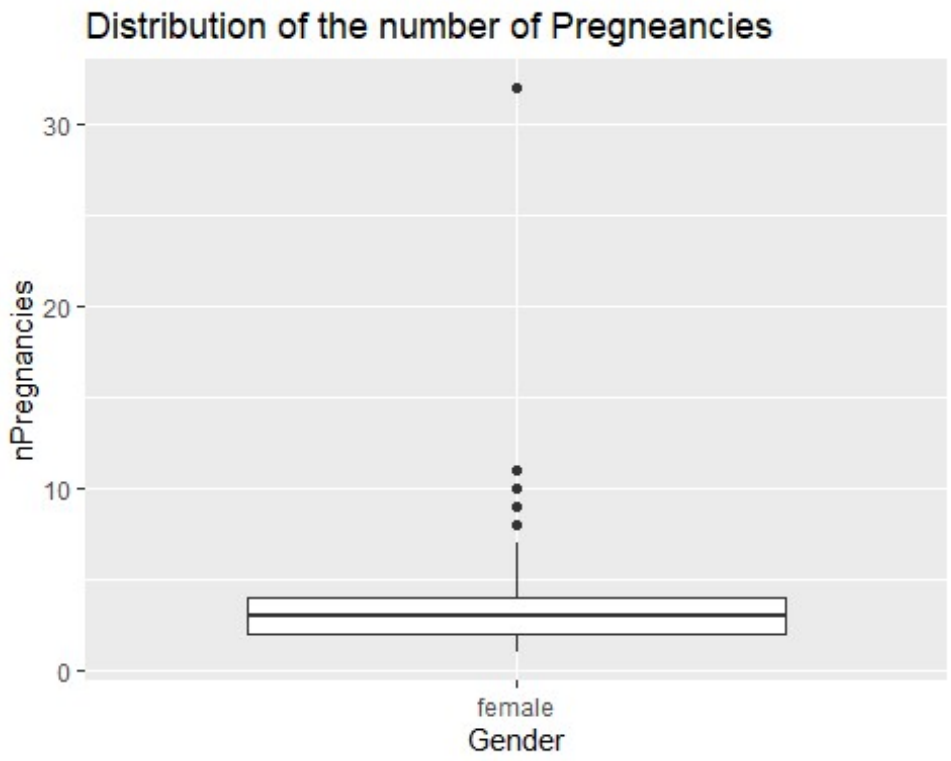
```
library(NHANES)

library(dplyr)
Females<- NHANES %>% filter(Gender=="female", Age%in%(30:60) )
View (Females)

#Basic summary statistics here, sample size etc...
#Summary for household median income
summary(Females$HHIncomeMid)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   2500   30000   60000   62136  100000  100000     171
```

```
#Summary for the number of pregnancies they have had
summary(Females$nPregnancies)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.000   2.000   3.000   2.955   4.000  32.000     514
```

The data for income ranges from $2500 to $100,000 and the data for pregnancies ranges from 1 to 32. The data for zero pregnancies might have been recorded as NA but the report does not specify this so the study will include only data starting from 1 pregnancy. In addition, there seems to be outliers in the data, such as 32 pregnancies. The following box plot is created to see the distribution of the number of pregnancies.

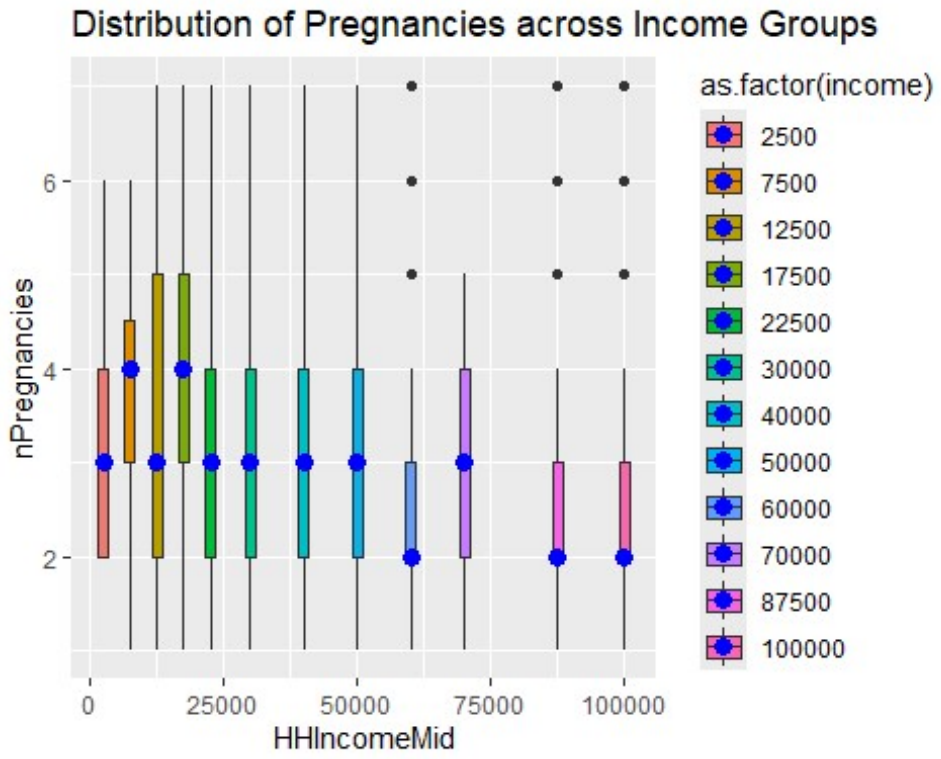
```
# Distribution of the number of Pregneancies
library(ggplot2)
ggplot(Females, aes( x= Gender, y=nPregnancies)) + geom_boxplot() +
  ggtitle("Distribution of the number of Pregneancies ") +
  ylab("nPregnancies")
```



(Figure1)The typical value for the number of pregnancies ranges between 1 and 7. In order to avoid the outliers the data is filtered to have only pregnancies less than 8. After filtering the the data set has 1501 observations. Using “ggplot” for visualization removes 96 rows of data that have “NA” in either income or pregnancies. Therefore the final sample size for the study will be 1405.

```
Females<- Females %>% filter(nPregnancies<8, !is.na(HHIncomeMid),
!is.na(nPregnancies))
```

```
income<-as.factor(Females$HHIncomeMid)
ggplot( Females, aes( x= HHIncomeMid, y=nPregnancies ,
fill=as.factor(income))) + geom_boxplot() + stat_summary(fun = median, geom =
"point", size = 3, color = "blue")+
  ggtitle("Distribution of Pregnancies across Income Groups ") +
  ylab("nPregnancies")
```



(Figure 2)The above figure shows the distribution of pregnancies across income groups. The blue point in the middle of the plots is the median value, which decreases from 3 to 2 in the two highest income groups. Some low income groups also have higher number of pregnancies. This indicates that females with higher income have lower number of pregnancies.

Analysis

Since household income (HHIncomeMid) and the number of pregnancies are recorded as categorical variables, it is not appropriate to do a regression analysis. A chi-square analysis which analyzes whether there is a relationship between two categorical variables is most appropriate. The null hypothesis for this test is that household income and the number of pregnancies are independent.

```
#Chi square analysis
res2=chisq.test(table(Females$HHIncomeMid, Females$nPregnancies))
res2
```

```
    Pearson's Chi-squared test

data:  table(Females$HHIncomeMid, Females$nPregnancies)
X-squared = 239.16, df = 66, p-value < 2.2e-16
```

```
round(res2$stdres,5)
```

```
               1        2        3        4        5        6        7
  2500   -2.06843  0.21304 -0.32016  1.12073  0.14046  2.27320 -0.62428
  7500   -1.90606 -1.60335  0.15903  1.96976  2.38696  0.73440 -0.72687
  12500  -1.71868 -1.17476 -0.47905  0.73612 -0.31216  5.72174  0.78103
  17500  -2.04793 -2.91555  0.45151  1.62195  4.30596  0.35225  0.98193
  22500  -0.57093  0.37530 -1.99752  1.61590 -0.66735  0.18772  3.76151
  30000  -0.41191 -4.32541  3.87762 -0.97209  3.00828 -0.83826  1.52817
  40000   1.33401 -1.55646  0.75890  1.18228 -0.38709 -1.91951 -0.07182
  50000  -1.82192 -1.93071 -0.77350  2.41821  0.94450  4.56369 -0.59733
  60000  -0.34800  2.30279  0.98495 -2.19277 -2.37082  0.46205 -0.57633
  70000   0.98974 -0.76072 -0.37049  1.50548  0.60093 -1.89722 -1.21257
  87500   2.08886  2.31315 -1.84506 -1.25002 -0.81940 -1.13066 -1.37731
  100000  1.79402  4.25832 -0.35768 -2.94435 -2.78590 -3.26396 -0.73235
```

We can also analyze the household income as a numerical variable and number of pregnancies as a factor. One-way-ANOVA and Tukey test can be used to test whether the income groups are different in regards to the response variable: the number of pregnancies. The null hypothesis for the one-way-ANOVA is that all means are equal, whereas the alternative hypothesis is that at least two of the income group means are different. And Tukey test will compare each group to all other groups.

```
anovaIncome =aov(nPregnancies~as.factor(HHIncomeMid ), data=Females)
summary(anovaIncome)
```

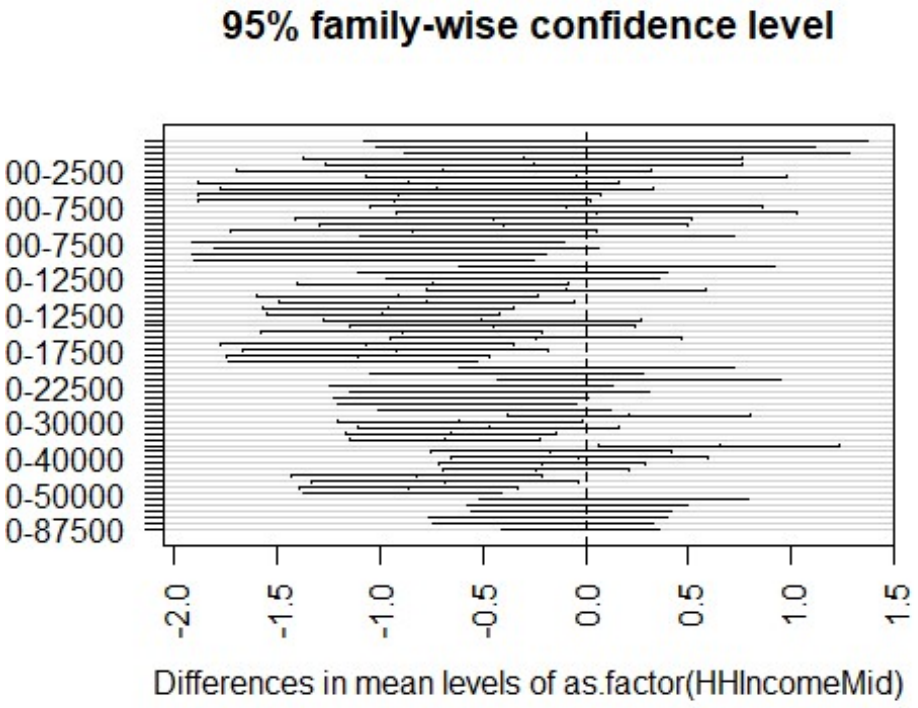
```
                        Df Sum Sq Mean Sq F value Pr(>F)
as.factor(HHIncomeMid)   11  223.5  20.316      11 <2e-16 ***
Residuals              1391 2568.7   1.847
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(anovaIncome, conf.level=.95)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = nPregnancies ~ as.factor(HHIncomeMid), data = Females)

$`as.factor(HHIncomeMid)`
                    diff         lwr         upr     p adj
7500-2500     0.14586255 -1.07840155  1.37012665 0.9999998
12500-2500    0.05132850 -1.01417440  1.11683140 1.0000000
```

```
17500-2500    0.20013803 -0.88363424  1.28391029 0.9999826
22500-2500   -0.30434783 -1.37544395  0.76674830 0.9988152
30000-2500   -0.25144928 -1.26404590  0.76114735 0.9996683
40000-2500   -0.69259511 -1.70008934  0.31489913 0.5129570
50000-2500   -0.04225924 -1.06470168  0.98018320 1.0000000
60000-2500   -0.86335404 -1.88751783  0.16080975 0.1987616
70000-2500   -0.72393924 -1.77220878  0.32433031 0.5054148
87500-2500   -0.90741942 -1.88665022  0.07181137 0.1000739
100000-2500  -0.93353573 -1.88736319  0.02029174 0.0616724
12500-7500   -0.09453405 -1.05017448  0.86110638 1.0000000
17500-7500    0.05427547 -0.92169303  1.03024398 1.0000000
22500-7500   -0.45021038 -1.41208310  0.51166234 0.9315850
30000-7500   -0.39731183 -1.29358419  0.49896053 0.9529282
40000-7500   -0.83845766 -1.72896136  0.05204604 0.0875147
50000-7500   -0.18812180 -1.09550306  0.71925946 0.9999439
60000-7500   -1.00921659 -1.91853704 -0.09989614 0.0152388
70000-7500   -0.86980179 -1.80618912  0.06658555 0.0981099
87500-7500   -1.05328198 -1.91167883 -0.19488513 0.0035838
100000-7500  -1.07939828 -1.90869861 -0.25009795 0.0013049
17500-12500   0.14880952 -0.61864625  0.91626530 0.9999715
22500-12500  -0.35567633 -1.10512477  0.39377211 0.9249969
30000-12500  -0.30277778 -0.96593536  0.36037981 0.9421865
40000-12500  -0.74392361 -1.39926374 -0.08858348 0.0113171
50000-12500  -0.09358775 -0.77168405  0.58450856 0.9999992
60000-12500  -0.91468254 -1.59537154 -0.23399354 0.0007155
70000-12500  -0.77526774 -1.49171389 -0.05882159 0.0209072
87500-12500  -0.95874793 -1.56974629 -0.34774957 0.0000206
100000-12500 -0.98486423 -1.55426118 -0.41546728 0.0000012
22500-17500  -0.50448585 -1.27968832  0.27071661 0.6008316
30000-17500  -0.45158730 -1.14371726  0.24054265 0.5968200
40000-17500  -0.89273313 -1.57737653 -0.20808974 0.0012625
50000-17500  -0.24239727 -0.94885356  0.46405902 0.9936650
60000-17500  -1.06349206 -1.77243734 -0.35454678 0.0000647
70000-17500  -0.92407726 -1.66742172 -0.18073281 0.0028977
87500-17500  -1.10755745 -1.74988516 -0.46522974 0.0000013
100000-17500 -1.13367375 -1.73656576 -0.53078174 0.0000001
30000-22500   0.05289855 -0.61920894  0.72500604 1.0000000
40000-22500  -0.38824728 -1.05264263  0.27614806 0.7511512
50000-22500   0.26208858 -0.42476299  0.94894016 0.9847890
60000-22500  -0.55900621 -1.24841756  0.13040514 0.2506187
70000-22500  -0.41959141 -1.14432969  0.30514687 0.7622827
87500-22500  -0.60307160 -1.22377240  0.01762920 0.0660626
100000-22500 -0.62918790 -1.20898387 -0.04939192 0.0201953
40000-30000  -0.44114583 -1.00641231  0.12412064 0.3068914
50000-30000   0.20919003 -0.38230815  0.80068822 0.9918234
60000-30000  -0.61190476 -1.20637345 -0.01743607 0.0370289
70000-30000  -0.47248996 -1.10758884  0.16260892 0.3826260
87500-30000  -0.65597015 -1.16917040 -0.14276989 0.0018044
100000-30000 -0.68208645 -1.14497779 -0.21919512 0.0000992
50000-40000   0.65033586  0.06761569  1.23305604 0.0141829
```

```
60000-40000  -0.17075893 -0.75649413  0.41497627 0.9985000
70000-40000  -0.03134413 -0.65827578  0.59558752 1.0000000
87500-40000  -0.21482432 -0.71788217  0.28823354 0.9640159
100000-40000 -0.24094062 -0.69256113  0.21067990 0.8461569
60000-50000  -0.82109479 -1.43218369 -0.21000590 0.0007165
70000-50000  -0.68167999 -1.33236210 -0.03099788 0.0305045
87500-50000  -0.86516018 -1.39752389 -0.33279647 0.0000078
100000-50000 -0.89127648 -1.37532710 -0.40722586 0.0000001
70000-60000   0.13941480 -0.51396880  0.79279840 0.9999252
87500-60000  -0.04406539 -0.57972763  0.49159685 1.0000000
100000-60000 -0.07018169 -0.55785773  0.41749435 0.9999987
87500-70000  -0.18348019 -0.76390383  0.39694345 0.9968901
100000-70000 -0.20959649 -0.74605291  0.32685993 0.9816631
100000-87500 -0.02611630 -0.41057618  0.35834358 1.0000000

plot(TukeyHSD(anovaIncome, conf.level=.95), las = 2)
```



Conclusions

The p-value for the chi-square analysis is less than 0.05. We can reject the null hypothesis and conclude that household income and the number of pregnancies that females of ages between 30 and 60 years are associated. Moreover, standardized residuals show that there are more than expected females in the high income and 2 pregnancies , and low income and high pregnancies of 6 and 7.

The p-value for one-way ANOVA is less than 0.05 so we can reject the null hypothesis and conclude that at least two of the income groups are different. The results from Tukey test

agree with the trend shown in figure 2. The figure and the test show a trend of decreasing pregnancies as income increase.

While it is generally hypothesized that inability to afford living expenses is the reason why people are having less children, our study suggests otherwise. Females who have higher pregnancies tend to have less household income than those who have lower pregnancies. Therefore, policy makers aiming to increase birth rates should look for other factors influencing the number of children people have.

Citations:

Nargund G. Declining birth rate in Developed Countries: A radical policy re-think is required. Facts Views Vis Obgyn. 2009;1(3):191-3. PMID: 25489464; PMCID: PMC4255510.