UNIVERSITY OF PENNSYLVANIA

SCHOOL OF SOCIAL POLICY AND PRACTICE

MASTER OF SCIENCE IN SOCIAL POLICY AND DATA ANALYTICS

POLICY RESEARCH REPORT 1

MULTIPLE REGRESSION

Autor: Rose Barragan

Professor: Dr. Momin Malik

Course: MSSP 8970

Applied Linear Modeling

Fall 2023

October 7th, 2023

INDEX

1. INTRODUCTION

In recent years, an increasing number of research studies have highlighted the critical role that early childhood nutrition and family income play in shaping the developmental trajectories of children from birth to age 5 (Birch and Gussow, 1970; Duncan and Brooks-Gunn, 1997).

The following report will comprehensively analyze the impact of two key programs: 1) the Women, Infants, and Children (WIC) Nutrition Program and 2) the Aid to Families with Dependent Children (AFDC) Program] on children's reading achievement.

The report will analyze in detail 1) participation in the WIC program during pregnancy (WICpreg), 2) participation in the AFDC program during pregnancy (AFDCpreg), 3) child age in 1997 (AGE97), 4) family income (faminc97), 5) low birth weight (bthwht) and 6) parenting practices (HOME97), thus facilitating an assessment of the effects on the improvement of children's reading achievement, which in turn, will have important implications for public policy formulation.

Therefore, before analyzing the variables that affect the Women, Infants, and Children (WIC) Nutrition Program and Aid to Families with Dependent Children (AFDC) program participation during pregnancy on child reading achievement, it is necessary to understand what variables constitute our Dataset.

1.1. INDEPENDENT VARIABLES

Independent variables are understood as those that influence or predict the dependent variable. Among the key independent variables, there are:

- **WICpreg** is a binary variable, where 0 indicates "no" participation in the program, and 1 indicates active participation in the Nutrition Program for Women, Infants, and Children during pregnancy).

- **AFDCpreg** is another binary variable, where 0 indicates "no" participation in the Aid to Families with Dependent Children during Pregnancy program, and 1 indicates active participation in the aid program.

- **Bthwht** is another binary variable, where 0 indicates that the child was born with normal weight, while 1 indicates that the child was born with low weight.

- **AGE97** is a continuous variable indicating the infant's age in 1997, ranging from 3 to 13 years old.

- **Faminc97** is a continuous variable indicating total family income in 1997. The characteristic feature of this variable is that we can find both positive and negative values.

- **HOME97** is a continuous variable since it indicates a score that measures the emotional and cognitive stimulation provided at home, ranging from 7 (minimum score achieved) to 27 (maximum score achieved).

1.2. DEPENDENT VARIABLES

Finally, we identify a dependent variable. We understand dependent variables as those that seek to explain the independent variables. In its analysis, the dependent variable (readss97) is a continuous variable very relevant to the analysis since it represents the score obtained by a child in the reading test in 1997, ranging from a minimum value of 47.5 to a maximum of 165.5.

Therefore, the Dataset that collects the data for the Women, Infant, and Children (WICpreg) Nutrition Program and Aid to Families with Dependent Children (AFDCpreg) program participation during pregnancy on child reading achievement collects six independent variables (WICpreg, AFDCpreg, age97, faminc97, bthwht, and HOME97) that have effects on one dependent variable (readss97).

2. METHODOLOGY

The observations in the study represent a total of 3563, i.e., the study collected data on 3563 children between the ages of 3 and 13 years in 1997, whose families have a total income between 72296.16 and 784610.59 dollars.

In this research, the data will be analyzed through linear regression models for each of the assumptions of the independent variables regarding the dependent variable to avoid assumptions in the recommendations on the importance of each policy for child achievement. We will also assess and compare the results of a multiple regression model through nested models (Control, WIC model, and AFDC Model) and draw some conclusions upon the comparative results.

3. RESULTS

3.1. DESCRIPTIVE STATISTICS

3.1.1 Family income (Famic97 - independent variable)

The family income ranges from a low of $-72,296 to $784,611, where the average is $49,841, and the median is $39,118 with a standard deviation of $71,491.87. This standard deviation indicates that income levels are spread out from the mean amount of $49841, thus making a wide variety of households' incomes available to the Dataset.

3.1.2. Age of children (age97 - independent variable)

The age of children values ranges from 3 to 13 years old, where the mean is 7.467 years old, the median is 7.000 years old, and the standard deviation is 1.97, making the values skew to the right.

3.1.3. Composite total emotional and cognitive stimulation score at home (home97 - independent variable).

The composite of total emotional and cognitive stimulation scores ranges from a score of 7.00 to 27.00, where the mean (18.92) and the median (19.20) are pretty close to one another.

3.1.4. Read achievement score (Readss97 – dependent variable)

For this variable, it is worth clarifying that all the children were submitted to a reading test known as the "Woodcock-Johnson (WJ) Revised Reading Achievement Test Age Standardized Score" to know the child's intellectual development, which means that average performance is represented by a score of 100. A score over 131 means a high performance, and a score under 69 is a significantly below performance.

The mean for reading achievement scores among children is 102.5 (average score within the WJ Test), and the median is 101.5. However, the Dataset collects scores ranging from 47.5 (considered below the average) to 165.5 (considered a high-performance score).

3.1.5. Low birth weight status (Bthwht - independent variable)

For the Bthwht variable, we obtain a mean of 0.3882, meaning that 38.82% of the children have low weight at birth.

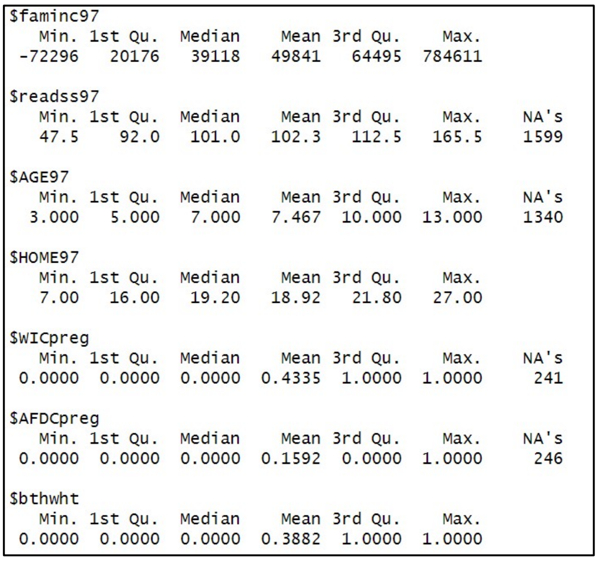
3.1.6. Women, Infants, and Children (WICpreg - independent variable)

We obtained a mean of 0.4335 for the WIC variable, meaning that there was a 43.3 5% active participation in the Nutrition Program for Women, Infants, and Children during pre gnancy.

3.1.7. AFDC program participation during pregnancy (AFDCpreg - independent variabl e)

For the AFDC program participation during pregnancy, we obtained a mean of 0.1592, meaning there was only a 15.92% active participation in the Aid to Families with Dependent Children program during pregnancy. This is such a lower participation rate and even lower if we compare it with the WIC program.

Table #1 Descriptive analytics



```
$faminc97
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -72296   20176   39118   49841   64495  784611 

$readss97
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   47.5    92.0   101.0   102.3   112.5   165.5    1599 

$AGE97
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  3.000   5.000   7.000   7.467  10.000  13.000    1340 

$HOME97
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   7.00   16.00   19.20   18.92   21.80   27.00 

$WICpreg
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 0.0000  0.0000  0.0000  0.4335  1.0000  1.0000     241 

$AFDCpreg
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 0.0000  0.0000  0.0000  0.1592  0.0000  1.0000     246 

$bthwht
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.3882  1.0000  1.0000 
```

3.2.CORRELATION REGRESSION MATRIX

The most relevant insights from the correlation matrix are the following [Table #2]:

- **WICpreg*AFDCpreg**: there is a positive correlation (0.384) between the WICpreg and the AFDC variable, meaning it is likely that the individuals who have participated in the WIC program (during pregnancy) have also participated in the AFDC.

- **WICPreg*faminc97** presents a negative correlation (-0.2524), meaning households with higher revenue income have lower participation in the WIC program during pregnancy.

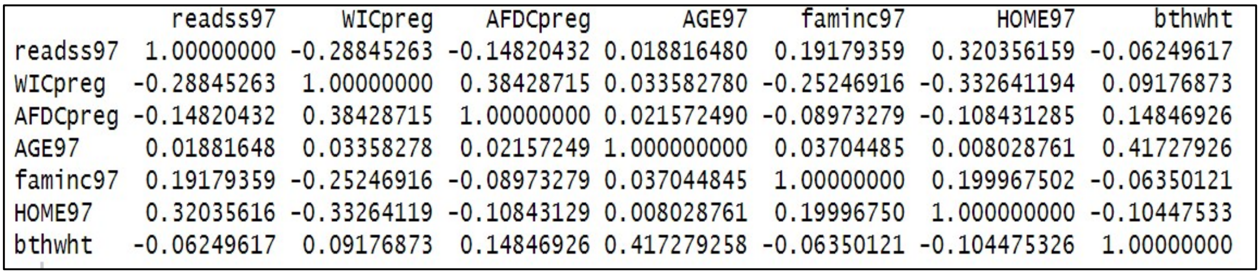
- **WICpreg*readss97** presents a negative correlation (-0.288), meaning that individuals who participated in the Women, Infant, and Children program during pregnancy are less likely to achieve more outstanding reading scores [Table #3].

- **WICpreg*HOME97** presents a negative correlation (-0.3203), meaning that households that implement positive parenting practices regarding emotional and cognitive stimulation are less likely to occur in the Women, Infant, and Children Nutrition program.

- **HOME97*readss97** presents a positive correlation (0.32), meaning that households that implement positive parenting practices regarding emotional and cognitive stimulation are more likely to have children scoring higher results on reading achievement tests [Table #4].

- **AGE97*bthwht** presents a positive correlation coefficient (0.4172), meaning that the variable weight birth increases in children that grow older.

Table #2 – Correlation regression matrix



```
            readss97     WICpreg    AFDCpreg       AGE97    faminc97      HOME97      bthwht
readss97  1.00000000 -0.28845263 -0.14820432 0.018816480  0.19179359  0.320356159 -0.06249617
WICpreg  -0.28845263  1.00000000  0.38428715 0.033582780 -0.25246916 -0.332641194  0.09176873
AFDCpreg -0.14820432  0.38428715  1.00000000 0.021572490 -0.08973279 -0.108431285  0.14846926
AGE97     0.01881648  0.03358278  0.02157249 1.000000000  0.03704485  0.008028761  0.41727926
faminc97  0.19179359 -0.25246916 -0.08973279 0.037044845  1.00000000  0.199967502 -0.06350121
HOME97    0.32035616 -0.33264119 -0.10843129 0.008028761  0.19996750  1.000000000 -0.10447533
bthwht   -0.06249617  0.09176873  0.14846926 0.417279258 -0.06350121 -0.104475326  1.00000000
```

Table #3 - Women, Infant and Children program during pregnancy and Reading Scores (negative correlation)

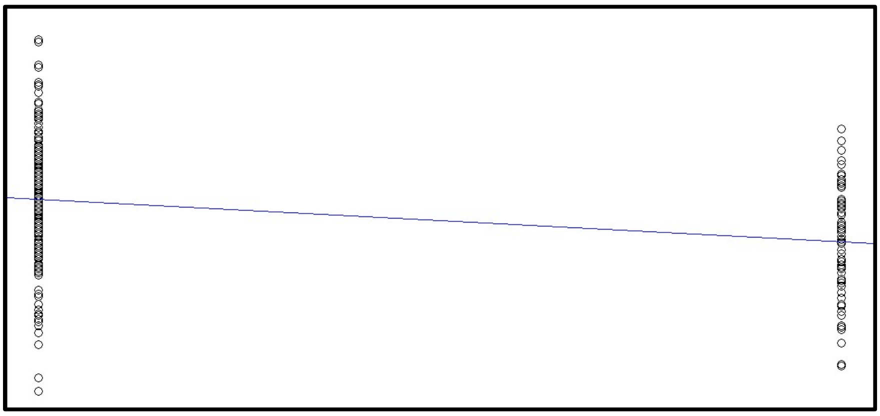
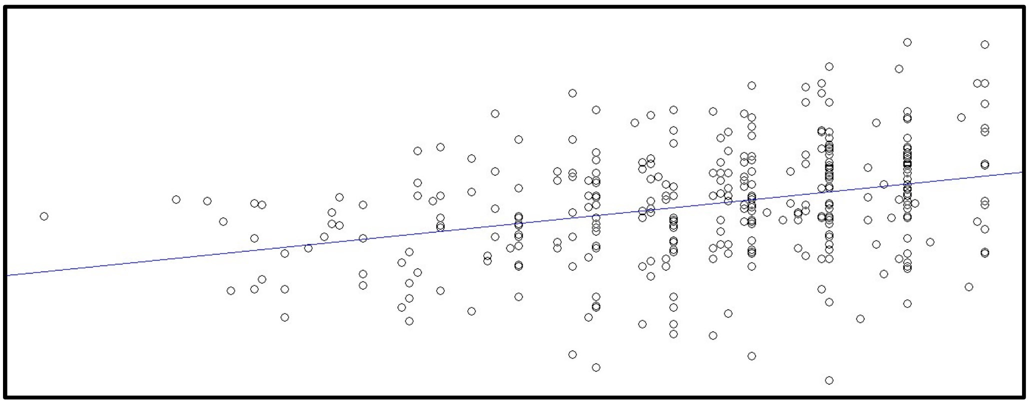


Table #4 - Emotional and cognitive stimulation at home and Reading Scores (positive correlation)



3.3. MULTIPLE REGRESSION MODEL AND NESTED MODELS

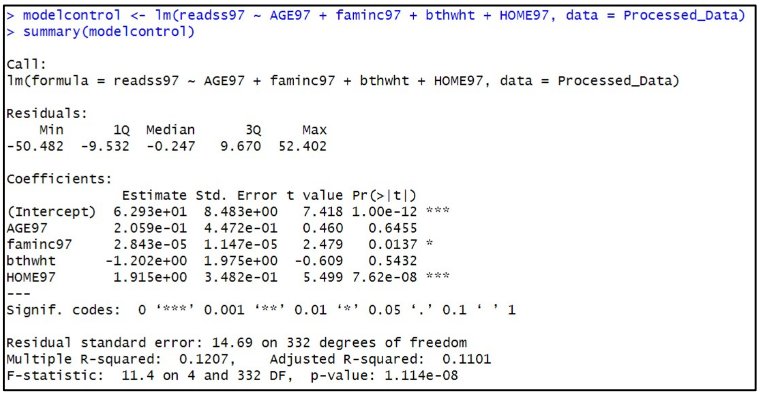
3.3.1. Control Model

The "control" model is a linear regression that collects the AGE97, faminc97, bthwht, and HOME97 variables to see the effect on readss97.

- The coefficient of AGE97 is 0.2059 with a standard error of 0.447 and a P-value of 0.6455, indicating that AGE97 is not a relevant variable and does not have a considerable weight on readss97.
- The coefficient of bthwht is -1.202, with a standard error of 1.975. However, unlike the other two variables, bthwht does not significantly affect obtaining a reading score, the p-value being 0.5432).
- The coefficient of faminc97 is approximately 0.00002843 with a standard error of 0.00001147 and a P-value equal to 0.0137, indicating that the variable faminc97 is relevant (positive effect) for achieving a good reading test score.
- The coefficient of HOME97 is 1.915, with a standard error of 0.348 and a P-value of 7.62e-08), indicating that the variable HOME97 positively affects readss97.

In summary, the model presents a multiple R-squared of 0.1207, i.e., the model explains 12.07% of the variance of readss97. Thus, the model suggests that family income (faminc97) and the emotional and cognitive stimulation provided at home (HOME97) score are relevant variables with a positive effect on the reading score (readss97). While on the other hand, the variables of age (AGE97) and low birth weight (bthwht) are not relevant to the model [Table #5].

Table #5 Linear regression for the Control Model.



```
> modelcontrol <- lm(readss97 ~ AGE97 + faminc97 + bthwht + HOME97, data = Processed_Data)
> summary(modelcontrol)

Call:
lm(formula = readss97 ~ AGE97 + faminc97 + bthwht + HOME97, data = Processed_Data)

Residuals:
    Min      1Q  Median      3Q     Max 
-50.482  -9.532  -0.247   9.670  52.402 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.293e+01  8.483e+00   7.418 1.00e-12 ***
AGE97        2.059e-01  4.472e-01   0.460   0.6455    
faminc97     2.843e-05  1.147e-05   2.479   0.0137 *  
bthwht      -1.202e+00  1.975e+00  -0.609   0.5432    
HOME97       1.915e+00  3.482e-01   5.499 7.62e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.69 on 332 degrees of freedom
Multiple R-squared:  0.1207,	Adjusted R-squared:  0.1101 
F-statistic:  11.4 on 4 and 332 DF,  p-value: 1.114e-08
```

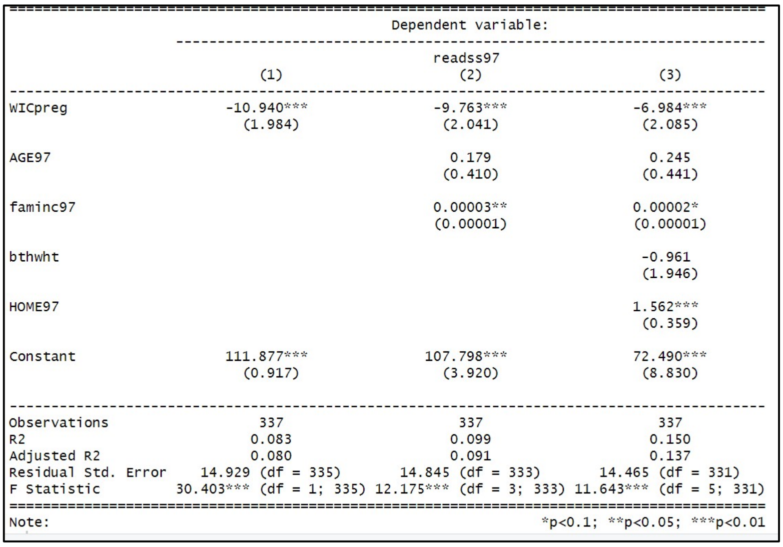
3.3.2 Nested Model WICpreg Model

We will make a triple comparison to see which variables have greater weight and relevance when assessing the reading achievement score.

- The first model (WICPreg1) contains only the variable WICpreg, offering a highly significant coefficient (given that the p-value is less than 0.01). Moreover, the model explains an R-squared equivalent to 0.083, i.e., 8.3% of the variance in readss97 is explained in the WICpreg.
- The second model (WICPreg2) integrates two new independent variables (AGE97 & faminc97) in addition to the previous one; the p-value of the second one is equal to 0.05, i.e., that the faminc97 variable is highly relevant to the model. With an R-squared of 0.099, i.e., the model explains approximately 9.9% of the variance of readss97.
- Finally, the third model (WICPreg3) contains the six independent variables studied so far. The p-value of HOME97 should be highlighted since it is less than 0.01, which makes HOME97 a very relevant variable for estimating the reading score in children. The third net model presents an R-squared of 0.150, i.e., the model explains 15% of the variance of readss97.

Thus, comparing the three models, we see that the WICpreg3 model (the model that integrates the six independent variables) is the best, with the highest R2 of 15%, an adjusted R1 (13.7%), the lowest residual error, and the lowest p-value of the three. Thus, faminc97 and HOME97 are the most relevant variables for the model, while bthwght, Age97, and WICpreg are not significant [Table #6].

Table #6 – Comparison among three multiple regression models (WICpreg) using Stargazer



```
===========================================================================================
                                              Dependent variable:
                    -----------------------------------------------------------------------
                                                   readss97
                              (1)                     (2)                     (3)
-------------------------------------------------------------------------------------------
WICpreg                   -10.940***              -9.763***               -6.984***
                           (1.984)                 (2.041)                 (2.085)

AGE97                                               0.179                   0.245
                                                   (0.410)                 (0.441)

faminc97                                          0.00003**               0.00002*
                                                  (0.00001)               (0.00001)

bthwht                                                                     -0.961
                                                                           (1.946)

HOME97                                                                    1.562***
                                                                           (0.359)

Constant                  111.877***              107.798***              72.490***
                           (0.917)                 (3.920)                 (8.830)

-------------------------------------------------------------------------------------------
Observations                 337                     337                     337
R2                          0.083                   0.099                   0.150
Adjusted R2                 0.080                   0.091                   0.137
Residual Std. Error   14.929 (df = 335)       14.845 (df = 333)       14.465 (df = 331)
F Statistic        30.403*** (df = 1; 335) 12.175*** (df = 3; 333) 11.643*** (df = 5; 331)
===========================================================================================
Note:                                                           *p<0.1; **p<0.05; ***p<0.01
```

3.3.3. Nested Model AFDCpreg Model

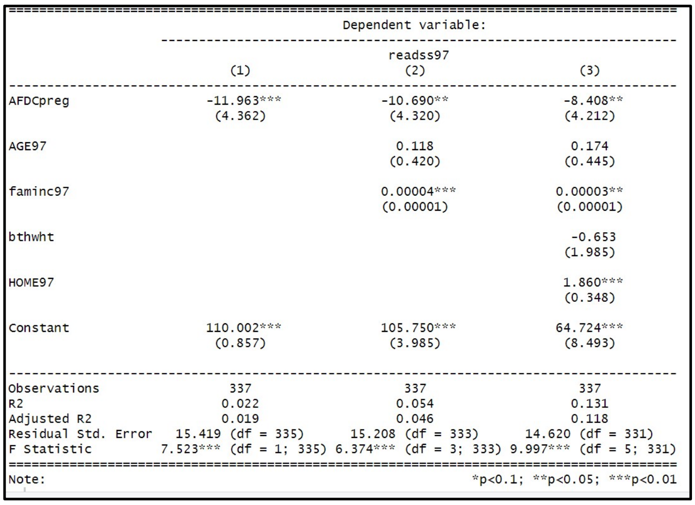
We will make a triple comparison to see which variables have greater weight and relevance when assessing the reading achievement score.

- The first model (AFDCpreg1) contains only the variable AFDCpreg, offering a highly significant coefficient (given that the p-value is less than 0.01). Moreover, the model explains an R-squared equivalent to 0.022, i.e., only 2,2% of the variance in readss97 is explained in the AFDCpreg model.
- The second model (AFDCpreg2) integrates two new independent variables (AGE97 & faminc97) in addition to the previous one. The p-value of the second one is equal to 0.01, i.e., the faminc97 variable is highly relevant to the model. With an R-squared of 0.054, i.e., the model explains approximately 5.4% of the variance of readss97.
- Finally, the third model (AFDCpreg3) contains the six independent variables seen so far. It is interesting to see that, as what happened in the other two models (control model and

WICpreg model), both HOME97 and famic97 are the most relevant variables for the model. The R-squared (R2) value is 0.131, indicating that approximately 13.1% of the variance in readss97 is explained by this model.

To sum up, the third model (AFDCpreg3) has the highest R-squared (13.1%) from all three models, indicating it explains the most variance of reads by the model. However, The second model (AFDCpreg2) has more significant values for HOME97 and faminc97 but has a lower R-squared (5.4%). While the first model has the lowest R-squared overall (2,2%), making it irrelevant [Table #7].

Table #7 Comparison among three multiple regression models (AFDCpreg) using Stargazer



```
========================================================================================
                                         Dependent variable:
                    --------------------------------------------------------------------
                                              readss97
                             (1)                    (2)                     (3)
----------------------------------------------------------------------------------------
AFDCpreg                 -11.963***             -10.690**               -8.408**
                          (4.362)                (4.320)                 (4.212)

AGE97                                             0.118                   0.174
                                                 (0.420)                 (0.445)

faminc97                                        0.00004***              0.00003**
                                                (0.00001)               (0.00001)

bthwht                                                                   -0.653
                                                                         (1.985)

HOME97                                                                   1.860***
                                                                         (0.348)

Constant                 110.002***             105.750***              64.724***
                          (0.857)                (3.985)                 (8.493)

----------------------------------------------------------------------------------------
Observations                337                    337                     337
R2                         0.022                  0.054                   0.131
Adjusted R2                0.019                  0.046                   0.118
Residual Std. Error   15.419 (df = 335)      15.208 (df = 333)       14.620 (df = 331)
F Statistic         7.523*** (df = 1; 335) 6.374*** (df = 3; 333) 9.997*** (df = 5; 331)
========================================================================================
Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

4. RECOMMENDATIONS & CONCLUSION

This report has analyzed 1) child age in 1997 (AGE97), 2) family income (faminc97), 3) low birth weight (bthwht), and 4) parenting practices (HOME97), thus facilitating an assessment of the effects on the improvement of children's reading achievement (readss97) on regards two public policies: WIC program during pregnancy (WICpreg), and the AFDC program during pregnancy (AFDCpreg).

The results have shown that from all three models (Control Model, nested model WICpreg3, and nested modelAFDCpreg3), the best performing model is WICpreg3 with an R-squared of 15%; in second place the AFDCpreg3 with an R-squared of 13.1%; and finally in the last position the Control Model with an R-squared of 12.07%. Although the control model has the lowest R-squares among the three, there is little difference against the other two [2.93% and 1.03%, respectively].

On the other hand, the worst performing models are nested model WICpreg1 (R-squared of 0.083), nested modelAFDCpreg1 (R-squared equivalent to 0.022), and nested modelAFDCpreg2 (R-squared equivalent to 0.054), meaning that WICpreg and AFDCpreg are not significant factors that have a relevant impact of the reading score.

The variables with a higher significance for the model are the emotional and cognitive stimulation at home (HOME97) and the family income (faminc97). In other words, households with greater income revenue and more investment in the child's emotional and cognitive support tend to perform better in the reading test, and they tend to have lower participation in the Women, Infant, and Children (WIC) Nutrition Program as well as Aid to Families with Dependent Children (AFDC) program.

The conclusion is that programs such as the Women, Infant, and Children (WIC) Nutrition Program and the Aid to Families with Dependent Children (AFDC) program have a very slight positive impact on the reading achievement score but do not provide excellent results.

Thus, policymakers need to look at all the contributor factors to understand whether the investment in fostering two programs (WIC & AFDC) will have the return of the investment policymakers expect and the positive impact society needs. On the other hand, policymakers need to start implementing programs that allow households to increase their income levels and provide a system that keeps fostering emotional and cognitive stimulation because those two factors are the ones worth developing and improving to achieve more outstanding results in reading scores.

5. R CODING

```
#Setting the working directory

setwd()

getwd()

Complete_Dataset <- read.csv("good.csv")


#Exploring the Dataset

head(Complete_Dataset)

names(Complete_Dataset)


#Display information about the Dataset’s structure,

str(Complete_Dataset)


#check null variales in the Dataset

Null_Values <- sum(is.na(Complete_Dataset))

print(Null_Values)


#The Dataset has a total of 4778 null values, so I need to remove them by cleaning the datsset

Processed_Data<- na.omit(Complete_Dataset)


#3.1. Descriptive Analytics

vars_to_summarize <- c(“faminc97”, “readss97”, “AGE97”, “HOME97”, “WICpreg”, “AFDCpreg”, “bthwht”)
```

```
summary_stats <- sapply(Processed_Data[vars_to_summarize], summary)

print(summary_stats)


sd(Processed_Data$faminc97)

sd(Processed_Data$AGE97)


#3.2 correlation matrix

library(dplyr)


cor_vars <- Processed_Data %>%

  select(readss97, WICpreg, AFDCpreg, AGE97, faminc97, HOME97, bthwht)

correlation_matrix <- cor(cor_vars, use = "pairwise.complete.obs")

print(correlation_matrix)


# Linear regression model

Regression_Model <- lm(readss97 ~ WICpreg, data=Processed_Data)

summary(Regression_Model)

plot(Processed_Data$WICpreg, Processed_Data$readss97,

    main= “Women, Infant and Children Nutrition Program participation during pregnancy
and Reading Scores”,

    xlab= “AWomen, Infant and Children Nutrition Program participation during pregnancy”,
ylab= “Reading Scores”)

abline(Regression_Model, col="blue")
```

```
Regression_Model <- lm(readss97 ~ HOME97, data=Processed_Data)

summary(Regression_Model)

plot(Processed_Data$HOME97, Processed_Data$readss97,

    main= “Emotional and cognitive stimulation at home and Reading Scores”,

    xlab= “emotional and cognitive stimulation”, ylab= “Reading Scores”)

abline(Regression_Model, col="blue")


#3.3. Multiple regression models & nested models

install.packages('stargazer')

library(stargazer)

#model control

modelcontrol <- lm(readss97 ~ AGE97 + faminc97 + bthwht + HOME97, data = Processed_Data)

summary(modelcontrol)


#modelWIC

modelWIC1 <- lm(readss97 ~ WICpreg, data = Processed_Data)

summary(modelWIC1)

modelWIC2 <- lm(readss97 ~ WICpreg + AGE97 + faminc97, data = Processed_Data)

summary(modelWIC2)

modelWIC3 <- lm(readss97 ~ WICpreg + AGE97 + faminc97 + bthwht + HOME97, data = Processed_Data)

summary(modelWIC3)
```

```
stargazer(modelWIC1,modelWIC2,modelWIC3, header = F, type="text")


#AFDCpreg

modelAFDC1 <- lm(readss97 ~ AFDCpreg, data = Processed_Data)

summary(modelAFDC1)

modelAFDC2 <- lm(readss97 ~ AFDCpreg + AGE97 + faminc97, data = Processed_Data)

summary(modelAFDC2)

modelAFDCpreg3 <- lm(readss97 ~ AFDCpreg + AGE97 + faminc97 + bthwht + HOME97,
data = Processed_Data)

summary(modelAFDCpreg3)


stargazer(modelAFDC1,modelAFDC2,modelAFDCpreg3, header = F, type="text")
```