**Interaction Effects & Incomplete Case Analysis**

*A study collaborated with the WIC Nutrition Program*

Linya Liu

Professor: Dr. Ezekiel Dixon-Roman

University of Pennsylvania, Philadelphia, PA

**Introduction**

This literature review examines the effect of the Women, Infant and Children (WIC) Nutrition Program participation during pregnancy on the raw scores of child mathematics achievement in 1997 using the dataset from *Child Development Supplement to the Panel Study of Income Dynamics* named “good.csv”. The effect of WIC program participation during pregnancy is moderated by (1) family income, (2) race, and (3) the current age of the child.

**Methods**

In paper 2, I evaluated the assumptions of linearity, homoscedasticity, and normality of residuals and diagnosis outliers. Homoscedasticity and normality are met in the model, but linearity is violated with three independent variables *faminc*, *race*,and *Age97*. Specifically, this will consist of creating a centered age variable, a centered binary race variable, a squared age term, and log transformed income variable. After data transformation, I first construct a multiple regression model (lm1) without interaction effects as the following.

*lm1:*

*mathraw*97 = Β0 + Β1 *Age*97 *+* Β2 *Age*972 + Β3 *race* + Β4*WICpreg* +Β5 *log(faminc97)* + ε

The second model (lm2) contains an interaction term between centered value of *logfaminc* and *WICpreg* (cincWIC).

*lm2:*

*mathraw*97 = Β0 + Β1 *Age*97 *+* Β2 *Age*972 + Β3 *race* + Β4*WICpreg* +Β5 *log(faminc97)*

+ Β6 *WICpreg\* log(faminc97)* + ε

The third model (lm3) contains an interaction term between *race* and *WICpreg* (raceWIC)

*lm3:*

*mathraw*97 = Β0 + Β1 *Age*97 *+* Β2 *Age*972 + Β3 *race* + Β4*WICpreg* +Β5 *log(faminc97)*

+ Β6 *WICpreg\* race* + ε

The fourth model (lm4) contains one interaction term between *Age97* and *WICpreg* (age97cWIC) and one between *squared Age97* and *WICpreg* (age97c2WIC).

*lm4:*

*mathraw*97 = Β0 + Β1 *Age*97 *+* Β2 *Age*972 + Β3 *race* + Β4*WICpreg* +Β5 *log(faminc97)*

+ Β6 *WICpreg\* Age*97 + Β7 *WICpreg\* Age*972 + ε

**Results**

Table 1 demonstrates the descriptive statistics (i.e. N, means/medians/proportions, standard deviations, frequencies, and observed ranges) of all variables. 36.33 is the mean child math test scores; 43% of the sample are WIC program participants; 7.47 is the mean age in years for children; 74 is the number of white more than black; 52060.10 is the mean family income in the sample, etc.

**Table 1.**

**CATEGORIES, VAIRABLES TYPES, N, SCALE RANGES, MEANS, STANDARD DEVIATIONS FOR ALL INVESTIGATED VARIABLES.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Category** | **Variable Names** | **Variable** | **Variable Types** | **N** | **Scale Range** | **Mean** | **Standard Deviation** |
| Outcome Variable of Interest | Child math test score | mathraw97 | Continuous | 2211 | 0-98 | 36.33 | 22.27 |
| Primary Independent Variable | WIC participation status | WICpreg | Binary | 2042 | 0=No, 1=Yes | 0.43 | 0.50 |
| Control Variables | Age | Age97 | Ordinal | 2042 | 3-13  (Unit: year) | 7.47 | 2.93 |
| Race | race | Binary | 1848 | -0.5 = Black, 0.5 = White | 0.02 | 0.50 |
| Family income | faminc97 | Continuous | 2042 | 0-784610.59 (Unit: dollar) | 52060.10 | 53175.04 |

On the top of diagonal, Table 2 shows the value of the correlation and the significance level of all variables. The correlation between *WICpreg* and *chrace*, *mathraw97*, *age97*, and *faminc97* are all significant. From the histogram distribution of each variable shown on the diagonal, we can tell that *age97* and *faminc97* are right skewed and nonlinear with *mathraw97*. Therefore, data transformation is applied for those three variables before the regression model as mentioned at the beginning of the Methods section.

**Table 2.**

**PERFORMANCE MATRIXS FOR ALL INVESTIGATED VARIABLES.**

A screenshot of a cell phone

Description automatically generated

Table 3 include both the variables before and after centering and standardizing the variables, which interaction effects avoid problems of multicollinearity. The first model tells us the expected math score (37.94) for the "average" child (i.e., a child having the average age of children in the sample (~ 7 years old), a child who did not receive WIC, and a child who comes from a family with the average income in the sample). The age term tells us that math score increases on average 7.02 points for each year increase in age. The effect of the age-squared variable is the negative (-0.15) and tells us that this increase in math score decelerates by 0.30 (i.e., -0.15 \* 2) points for each additional year (i.e. reading scores increase at a decreasing rate with age).

**Table 3.**

**ESTIMATES AND STANDARD ERRORS MULTIPLE REGRESSION MODELS BEFORE AND AFTER CENTERING AND STANDARDIZING VARIABLES**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficients:** | **lm1** | **lm2** | **lm3** | **lm4** |
| (Intercept) | 37.94 | 37.73 | 37.72 | 38.18 |
| (0.34) | (0.35) | (0.35) | (0.39) |
| AGE97c | 7.02 | 7.02 | 7.02 | 7.22 |
| (0.07) | (0.07) | (0.07) | (0.09) |
| AGE97c2 | -0.15 | -0.15 | -0.15 | -0.18 |
| (0.03) | (0.03) | (0.03) | (0.03) |
| race | 3.67 | 3.56 | 4.63 | 3.66 |
| (0.45) | (0.45) | (0.56) | (0.45) |
| cinc | 0.95 | 1.46 | 0.94 | 0.97 |
| (0.18) | (0.26) | (0.18) | (0.18) |
| WICpreg | -1.94 | -2.05 | -2.17 | -2.52 |
| (0.47) | (0.47) | (0.47) | -0.65 |
| cincWIC | - | -0.97 | - | - |
| - | (0.35) | - | - |
| raceWIC | - | - | -2.53 | - |
| - | - | (0.90) | - |
| AGE97cWIC | - | - | - | -0.47 |
| - | - | - | (0.13) |
| AGE97c2WIC | - | - | - | 0.06 |
| - | - | - | (0.05) |

In model 2, the p-value for the *cinc\*WICpreg* interaction term is significant (*p*<0.05), meaning that there is evidence for *cinc* being moderated by *WICpreg* (i.e., the impact of *WICpreg* on children’s math scores depends on the level of family income). In such interaction model, the parameter estimate for WIC (-2.05) can be interpreted as the effect of WIC participation on math scores when *cinc*=0 (i.e., when the interaction term *cinc\*WICpreg* = 0). Since *cinc* is the centered value of *logfaminc*, the parameter estimate for *WICpreg* is giving us the estimated effect of WIC participation at the average score of *logfaminc*. On average, children whose mothers participated in the WIC program have math scores that are 2.05 points lower than children whose mothers did not participate in the WIC program. However, this disparity is not uniform across income levels. For every 1-unit increase in the logged family income score, a child whose mother did not participate in the WIC program should expect to earn an additional 1.46 points, while a child whose mother did participate in the WIC program should only expect to earn 0.49 points. The predicted gg-plot loess smoother of interaction between WIC and Centered Log Family Income is shown in Figure 1.

In model 3, the p-value for the *race\*WICpreg* interaction term is significant (*p*<0.05), meaning that there is evidence for race being moderated by *WICpreg* (i.e., the impact of *WICpreg* on children’s math scores depends on race). If the child is black, children whose mothers participated in the WIC program have math scores that are 2.17 points lower than children whose mothers did not participate in the WIC program. If the child is white, children whose mothers participated in the WIC program have math scores that are 2.53 points lower than children whose mothers did not participate in the WIC program. The predicted gg-plot loess smoother of interaction between WIC and Centered Race is shown in Figure 2.

In model 4, the p-value for the *AGE97c\*WIC* interaction term are significant (*p*<0.05), meaning that there is evidence for race being moderated by *WICpreg* (i.e., the impact of *WICpreg* on children’s reading scores depends on the level of age). On average, children whose mothers participated in the WIC program have math scores that are 2.52 points lower than children whose mothers did not participate in the WIC program. However, this disparity is not uniform across age levels. For every 1-year increase in age, a child whose mother did not participate in the WIC program should expect to earn an additional 7.22 points, while a child whose mother did participate in the WIC program should only expect to earn 6.75 points. According to the coefficient of the square term, each additional year of age reduces the slope by .18 points.

**Figure 1. Figure 2.**

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

**Conclusion**

The effect of WIC program participation during pregnancy is moderated significantly by all the following three terms: (1) family income, (2) race, and (3) the current age of the child. Detailed interpretation is seen above.

**Appendix:**

## Step 1: Start the setup

install.packages("Ecdat")

install.packages("lmSupport")

install.packages("pastecs")

library("PerformanceAnalytics")

library(ggplot2)

library(lmSupport)

library(car)

library(pastecs)

setwd("/Users/Leah/Downloads")

good <- read.csv("good.csv",,header=TRUE, sep=",")

## Create a dataset with no missing observations

goodR <- na.omit(good[,c("WICpreg","CHRACE","mathraw97","AGE97","faminc97")])

* #WICpreg – Women, Infant and Children (WIC) Nutrition Program participant during pregnancy: 0 = No, 1 = Yes.
* #Race – Centered Binary Coding of Race: -0.5 = Black, 0.5 = White.
* #mathraw97 – Woodcock-Johnson Revised Mathematics Achievement Test Raw Score. Minimum = 0, Maximum = 98.
* #age97 – The child’s age in 1997. Minimum = 3, Maximum = 13.
* #faminc97 – Total family income in 1997 (in 2002 constant dollars). Minimum = $-72296.26, Maximum = $784610.59.

## Step2: Display descriptive data, correlations and frequencies

stat.desc(goodR, basic=TRUE, desc=TRUE, norm=FALSE, p=0.95)

chart.Correlation(goodR, histogram=TRUE, pch=19)

## Step2: Perform data transformations for variables that violate OLS assumptions

# Create a centered race varaible, a centered age variable, a squared age term, and log transformed income variable

goodR$race<- ifelse(goodR $CHRACE== 9, NA, ifelse(goodR$CHRACE== 1, .5, ifelse(goodR$CHRACE== 2, -.5, NA)))

goodR$AGE97c <- goodR$AGE97 -mean(goodR$AGE97)

goodR$logfaminc <- ifelse(goodR$faminc97 <= 1, 0, ifelse( goodR$faminc97 > 1, log(goodR$faminc97),NA))

goodR$AGE97c2 <- goodR$AGE97c\*\*2

# Center continuous variables to avoid multicollinearity

goodR$cinc <- goodR$logfaminc - mean(goodR$logfaminc)

goodR$cincWIC <- goodR$cinc \* goodR$WICpreg

goodR$raceWIC <- goodR$race \* goodR$WICpreg

goodR$AGE97cWIC <- goodR$AGE97c \* goodR$WICpreg

goodR$AGE97c2WIC <- goodR$AGE97c2 \* goodR$WICpreg

# New Standardized Variables

goodR$zmath <- (goodR$mathraw97-mean(goodR$mathraw97))/sd(goodR$mathraw97)

goodR$zinc <- (goodR$logfaminc-mean(goodR$logfaminc))/sd(goodR$logfaminc)

goodR$zAGE97 <- (goodR$AGE97c - mean(goodR$AGE97c))/sd(goodR$AGE97c)

goodR$zAGE972 <- (goodR$AGE97c2 - mean(goodR$AGE97c2))/sd(goodR$AGE97c2)

## Step 3: Construct Main Effects Model

lm1<-lm(mathraw97 ~ AGE97c + AGE97c2 + race+ cinc + WICpreg, data=goodR)

summary(lm1)

lm2<-lm(mathraw97 ~ AGE97c + AGE97c2 + race + cinc + WICpreg+ cincWIC , data= goodR)

summary(lm2)

lm3<-lm(mathraw97 ~ AGE97c + AGE97c2 + race + cinc + WICpreg+ raceWIC , data= goodR)

summary(lm3)

lm4<-lm(mathraw97 ~ AGE97c + AGE97c2 + race + cinc + WICpreg+ AGE97cWIC+ AGE97c2WIC , data= goodR)

summary(lm4)

## Step 4: Plot Interaction Effect & Set Up the Dataset with Predicted Y Values for Model 2

# First we create a new data frame with all of the WIC, chome variables used in lm2 (above),

# as well as the math scores, and untransformed age variable (AGE97)

goodplot2 <- na.omit(goodR[,c("AGE97","race","cinc","cincWIC","WICpreg","mathraw97")])

# We want to control for several other covariates. However, we have to choose at what value/level to hold these covariates constant at.

# In this code, AGE97c & AGE97c2= 0 (or mean which is ~ 7 year old child). Keep this in mind when interpretting the plot

goodplot2$AGE97c<-0

goodplot2$AGE97c2<-0

# Now we can use the predict function to produce fitted values for our goodplot data set using coefficients from the lm2 model (above)

goodplot2$fit<-predict(lm2,goodplot2)

# Create a duplicate of WICpreg that is a factor

goodplot2$WicR<-factor(goodplot2$WICpreg, levels=c(0,1), labels=c("non participant","participant"))

# Plot using loess smoother, which is short for "local regression smoother".

# A smoother that can be used to fit a smoothed curve through points in a scatter plot.

# Note: We are plotting untransformed age values (AGE97) against predicted Y values (fit).

# the scale x/y continuous funcitons allow you to set the minimum/maximum values and major units the x and y axes

# labs is used to set the x/y axis labels

# ggtitle is used to set the title of the plot and the final line centers the title

ggplot(goodplot2,aes(x = cinc, y =fit)) + geom\_smooth(method="loess", aes(colour= WicR)) + scale\_x\_continuous(limits=c(-12, 4), breaks=c(-12,-10,-8,-6,-4,-2,0,2,4)) + scale\_y\_continuous(limits=c(0,100), breaks=c(0,10,20,30,40,50,60,70,80,90,100)) + labs(x="Centered log family income", y="Math Scores in 1997") + ggtitle("Interaction between WIC and Centered Log Family Income") + theme(plot.title = element\_text(hjust = 0.5))

# Plot the Interaction Effect & Set Up the Dataset with Predicted Y Values for Model 3

goodplot3 <- na.omit(goodR[,c("AGE97","race","cinc","raceWIC","WICpreg","mathraw97")])

goodplot3$AGE97c<-0

goodplot3$AGE97c2<-0

goodplot3$fit<-predict(lm3,goodplot3)

goodplot3$WicR<-factor(goodplot3$WICpreg, levels=c(0,1), labels=c("non participant","participant"))

ggplot(goodplot3,aes(x = cinc, y =fit)) + geom\_smooth(method="loess", aes(colour= WicR)) + scale\_x\_continuous(limits=c(-0.5, 0.5), breaks=c(-0.5,0,0.5)) + scale\_y\_continuous(limits=c(0,100), breaks=c(0,10,20,30,40,50,60,70,80,90,100)) + labs(x="Centered Race", y="Math Scores in 1997") + ggtitle("Interaction between WIC and Centered Race") + theme(plot.title = element\_text(hjust = 0.5))