**Assumptions Diagnostics & Violation Correction of**

**Children’s Mathematics Achievement Model**

*A study collaborated with the WIC Nutrition Program & the AFDC Program*

Linya Liu

Professor: Dr. Ezekiel Dixon-Roman

University of Pennsylvania, Philadelphia, PA

**Introduction**

In this paper, I still examine the effect of the Women, Infant and Children (WIC) Nutrition Program participation during pregnancy, but now move to focus 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”.

**Methods**

I first construct a general linear model as the following:

*Model 1:*

*mathraw*97 = Β0 + Β1 *Age*97 + Β2 *faminc*97 + Β3*bthwht* + Β4*WICpreg* + ε

Using Table 1 to 3, I evaluate 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 two independent variables and .

Before data transformation, I would like to check whether is an omitted relevant variable, I used the method of added variable plots and construct model 2 as below.

*Model 2:*

*mathraw*97 = Β0 + Β1 *Age*97 + Β2 *faminc*97 + Β3*bthwht* + Β4*WICpreg* + ε

*HOME97* = Β0 + Β1 *Age*97 + Β2 *faminc*97 + Β3*bthwht* + Β4*WICpreg* + ε

After ensuring that is an omitted relevant variable, I specify a new model with transformation of nonlinear variables and . Specifically, I want to do is log-transform our faminc97 variable and center AGE97.

*Model 3:*

*mathraw*97 = Β0 + Β1 *Age*97 + Β2 *log(faminc97)* + Β3*bthwht*

+ Β4*WICpreg* +Β5*HOME97 +* Β6 + ε

**Results**

Table 1 demonstrates the descriptive statistics (i.e. N, means/medians/proportions, standard deviations, frequencies, observed ranges, and correlations) of all variables.

**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** |
| Independent Variable | Child math test score | mathraw97 | Continuous | 2211 | 0-98 | 36.33 | 22.27 |
| Dependent Variables | WIC participation status | WICpreg | Categorical | 3322 | 0=No, 1=Yes | 0.43 | 0.50 |
| Control Variables | Low birth weight status | bthwht | Categorical | 3563 | 0 = non-low birth weight, 1 = low birth weight child | 0.39 | 0.49 |
| Age | AGE97 | Ordinal | 2223 | 3-13 (Unit: year) | 7.47 | 2.93 |
| Family income | faminc97 | Continuous | 3563 | 0-784610.59  (Unit: dollar) | 49841.25 | 49751.07 |
| Parenting practices | HOME97 | Continuous | 3563 | 7-27 | 18.92 | 3.62 |

On the top of diagonal, Table 2 shows the value of the correlation and the significance level of all variables. From the histogram distribution of each variable shown on the diagonal, we can tell that and are right skewed and nonlinear with . Table 3 even shows an obviously upside-down U-shaped loess curve between and .

**Table 2.**

**PERFORMANCE MATRIXS FOR ALL INVESTIGATED VARIABLES.**

A screenshot of a cell phone

Description automatically generated

**Table 3.**

**SCATTERPLOTS OF DV AGAINST NON-LINEAR IVS WITH LOESS CURVE AND REGRESSION LINE**

A close up of a map

Description automatically generated A close up of a map

Description automatically generated

In Table 4’s Residuals vs. Fitted plot, since the residuals is spread randomly and form an approximate horizontal band around the 0 line, homoscedasticity is met in the model. And table 3’s Normal Q-Q plot shows that most residuals are normally distributed despite observations numbered as 487, 1533, and 1840 look a little off.

**Table 4.**

**RESIDUALS VS FITTED PLOT AND NORMAL Q-Q PLOT OF MODEL 1**

A close up of a map

Description automatically generated A close up of a map

Description automatically generated

In Table 5, the Residuals vs Leverage plot shows no outlier influential case as all cases are well inside of the Cook’s distance lines. And the residuals appear randomly spread in the Scale-Location plot.

**Table 5.**

**RESIDUALS VS LEVERAGE PLOT AND SCALE-LOCATION PLOT OF MODEL 1**

A close up of a map

Description automatically generated A close up of a map

Description automatically generated

In Table 5, I used added variable plots to visualize the evaluation of potential omitted relevant variable bias. Since the regression/lowess line is not horizontal, I can suspect that HOME97 is an omitted variable.

**Table 5.**

**PLOTTED RESIDUALS OF MODEL 1 AND MODEL 2 ADDING HOME97 AS IV**

**WITH LOESS CURVE AND REGRESSION LINE**

A close up of a map

Description automatically generated

As Table 6 shows, all independent variables are significant as all the p-values are less 0.01. Sampled coefficients interpretation involve with and . tells us that a one percent increase in family income is associated with a 0.009-point increase in math score, holding all other independent variables constant. indicates that on average a child’s math scores will increase by 6.92 points for every additional year of age above the mean, holding all other independent variables constant.

**Table 6.**

**MULTIPLE REGRESSION MODELS BEFORE ASSUMPTION DIAGNOSTICS AND CORRECTIONS AND AFTER RESPECIFICATION AND CORRECTIONS**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model 1** | | | | **Model 3** | | |
|  | **Estimate** | **Std. Error** | **p- value** | **Estimate** | **Std. Error** | **p-value** |
| **(Intercept)** | -15.55 | 0.59 | \*\*\* | 12.53 | 1.74 | \*\*\* |
| **WICpreg** | -3.11 | 0.40 | \*\*\* | - | - | - |
| **bthwht** | -2.15 | 0.38 | \*\*\* | -1.80 | 0.39 | \*\*\* |
| **AGE97** | 7.01 | 0.06 | \*\*\* | - | - | - |
| **faminc97** | 0.00 | 0.00 | \*\*\* | - | - | - |
| **logfaminc** | - | - | - | 0.90 | 0.17 | \*\*\* |
| **AGE97c** | - | - | - | 6.92 | 0.06 | \*\*\* |
| **AGE97c2** | - | - | - | -0.08 | 0.02 | \*\* |
| **HOME97** | - | - | - | 0.76 | 0.07 | \*\*\* |
| **Adjusted R-squared** | 0.87 | | | 0.86 | | |
| Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 | | | | | | |

Conclusion

implications of results, particularly in relation to the research questions, and/or the methodological limitations of the study

one to two paragraph concluding statement your thoughts on the changes in the results of the before and after analyses and what they suggests

**Appendix:**

# initial setup

install.packages("pastecs")

install.packages("lmSupport")

install.packages("PerformanceAnalytics")

library(pastecs)

library(lmSupport)

library("PerformanceAnalytics")

library(ggplot2)

setwd("/Users/Leah/Downloads")

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

# descriptive data

mydata <- na.omit(good[,c("mathraw97","AGE97","faminc97","bthwht","WICpreg","HOME97")])

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

# display correlations and frequencies

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

summary(lm)

# check linearity

ggplot(mydata, aes(x = faminc97, y = mathraw97))

+ geom\_point(size = 0.6) +

+ xlab("Respondent's 1997 Family Income")

+ ylab("Math Achievement")

+ theme\_bw()

+ geom\_smooth(method = "loess")

ggplot(mydata, aes(x = HOME97, y = mathraw97))

+ geom\_point(size = 0.6)

+ xlab("Respondent's Age")

+ ylab("Reading Achievement")

+ theme\_bw()

+ geom\_smooth(method = "loess")

# check homoscedasticity

mydata.res<- resid(lm)

fitted.res<-fitted(lm)

plot(fitted.res, mydata.res)

abline(0, 0, col= "red")

lines(lowess(mydata.res ~ fitted.res), col="green")

# check normality by Normal Q-Q plot

plot(lm)

# check added variable

lm<-lm(mathraw97 ~ WICpreg + bthwht + AGE97 + faminc97, data = mydata)

summary(lm)

resid\_math <- as.data.frame(lm$residuals)

lm2<-lm(HOME97 ~ WICpreg + bthwht + AGE97 + faminc97, data = mydata)

resid\_home <- as.data.frame(lm2$residuals)

sq <- data.frame(resid\_math,resid\_home)

plot(sq)

abline(lm(lm$residuals~ lm2$residuals), col="red")

lines(lowess(lm$residuals ~ lm2$residuals), col="blue")

# respecifying

min(mydata$faminc97) #check min is 0

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

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

mydata$AGE97c2 <- (mydata$AGE97c)^2

lm3 <- lm(mathraw97 ~ logfaminc + AGE97c + AGE97c2 +bthwht+HOME97,data = mydata)

fitted <- lm3$fitted.values

residuals <- lm3$residuals

pred1<-data.frame(fitted,residuals)

plot(mydata $logfaminc , pred1$residuals)

abline(lm(pred1$residuals~ mydata $logfaminc ), col="red")

lines(lowess(pred1$residuals~ mydata$logfaminc), col="blue")

plot(mydata$HOME97, pred1$residuals)

abline(lm(pred1$residuals~mydata$HOME97), col="red")

lines(lowess(mydata$HOME97, pred1$residuals), col="blue")

summary(lm3)