1. Introduction

The Women, Infant and Children (WIC) Nutrition Program is a federal assistance program provided by the Food and Nutrition Service of the United States Department of Agriculture (USDA) and aims to provide services to families 185% below the federal poverty level.

This policy report will examine the effect of the WIC program participation during pregnancy on child mathematics achievement in 1997. Additionally, this report will determine the extent to which parental influence contributes significantly toward child mathematics achievement.

2. Methods

Data and Sample

The primary dataset leveraged in this analysis is the Panel Study of Income Dynamics (PSID). The PSID dataset consists of a total 3536 observations across 112 variables. This report however, will focus on a selection of variables to analyse the effect of the WIC program on math levels.

Operationalization of relevant variables

The primary dependent variable in consideration is the Woodcock-Johnson Revised Mathematics Achievement Test Raw Score is labelled 'math97' (minimum = 0, maximum = 98).

The analyses in this report will focus on five independent variables. WIC program participation during pregnancy is labelled 'WICpreg' (1 = yes, or 0 = no). A child's age in 1997 is labelled 'AGE97' (minimum = 3 years, maximum = 13 years). Total family income in 1997 (in 2002 constant dollars) is labelled 'faminc97' (minimum = $-72296.26[1], maximum = $784610.59). Low birth weight status of the child, labelled 'bthwht,' is coded as a binary, where 0 = non-low birth weight child, 1 = low birth weight child. Finally, the composite total score of emotional and cognitive stimulation at home is labelled 'HOME97' (minimum = 7, maximum = 27).

Child age, family income, parenting practices and math scores are continuous variables. Categorical variables comprise of WIC program participation and low birth weight status and are coded as binaries.

Rationale of analyses

Ancillary enrollment i.e. participation in programs in addition to WIC (such as the Aid to Families with Dependent Children 'AFDC' Program) are included in our analyses. Missing values in the PSID dataset have not been imputed in this analysis, only omitted from regressions. All multivariate analyses were conducted using RStudio v1.1.456.

3. Results and Analysis

Descriptive Statistics

[1] The minimum value for FAMINC97 is negative, indicating a potential inputting error. This is omitted from the regressions in this report.

The PSID dataset we are considering comprises of 3564 observations. 242 observations were removed from the initial dataset, as these rows do not indicate either a 'yes' or 'no' participation status in the WIC program. After removal, the PSID dataset contains of 3322 observations. 1440 observations indicate participation in only the WIC program, 1882 indicate non-WIC status. Additionally, rows with missing fields are removed, leaving 2036 observations in the cleaned PSID dataset. Table 1 contains relevant descriptive statistics for the cleaned PSID dataset.

Table 1: Descriptive statistics summary

Measure	Mean	SD	Min	Max
WIC participation	0.434	N/A	0.00	1.00
Low birth weight status	0.364	N/A	0.00	1.00
Child age in 1997	7.38	2.92	3.00	13.00
Family income in 1997	50267.00	50577.00	-72,296.00	784611.00

Standard deviations for WIC participation and low birth status are not included in the table above as they are binary measures. Frequencies for the each of the above independent variables are detailed in Figure 1[2]. Correlations between independent variables (Table 2) indicate relatively weak correlations between our independent variables. Plots of each measure against math scores are also detailed in the Appendix (Figure 3).

Table 2: Correlations of independent measures[3]

Measure	1	2	3	4	5
1. WIC participation	-				
2. Low birth status	0.104	-			
3. Child's age in 1997	N/A	N/A	-		
4. Family income in 1997	-0.393	-0.0994	N/A	-	
5. Parenting practices	-0.304	0.0580	N/A	0.302	-

[2] All figures containing plots are in the Appendix.
[3] The correlations for age are N/A due to initial missingness in the data.

Table 3: Linear regressions

Independent Variable	Model 1	Model 2	Final Model
WIC participation	-3.11*** (0.402)	-1.83*** (0.457)	-2.67*** (0.451)
Child's age in 1997	7.01*** (0.0639)	**0.396***** (0.00407)	**0.409***** (0.00400)
Family income in 1997	3.24e-05*** (3.71e-06)	**0.809***** (0.221)	**1.93***** (0.229)
Birth weight status	-2.15*** (0.383)	-3.192*** (0.414)	-3.41*** (0.410)
Parenting practices	-	0.806*** (0.0705)	-
Model R^2	0.865	0.846	0.850
Adjusted R^2	0.865	0.845	0.849
Model N	2041	2041	2020

Standard errors are included in parentheses
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Transformed variables highlighted bold

Model 1 Observations:

From our initial regression (Model 1), we can note the following: WIC participation decreases math scores by -3.11 points than not participating in the program, holding all other variables in the model constant. As a child ages by one year, math scores increase by 7.01 points, holding all other variables constant. For every dollar increase in family income, we see a marginal increase in math scores by 3.24e-05 points, holding all other variables constant. Low birth weight children tend to have math scores 2.15 points lower than non-low birth weight children, holding all other variables constant.

Significance levels ($p < .05$) indicate that all independent variables this regression model (Table 3) have a significant impact. In general, participation in the WIC program's effect on child math achievement is 0.865, indicating that our pre-diagnosed model explains 86.5% of the total variance in a child's math scores in 1997.

Diagnosis of Generalised Linear Model (GLM) assumptions

In order to assess the robustness of Model 1, evaluation of general linear model (GLM) assumptions is required in order to validate the model's estimation capabilities. Evaluation is considered across three verticals: linearity, homoscedasticity, and normality.

A close examination of the residuals-fitted plot and Q-Q plots (Figure 2), we can note a relatively flat fitted line, indicating that the linearity assumption is slightly violated. From our residuals-fitted plot (Appendix), we can see that variance increases - larger predicted values are associated with larger errors or residuals. The unequal variance here therefore suggests a slightly heteroscedastic model. When considering normality assumptions, we can see significant deviations at the tails of our normal Q-Q plot (Appendix), indicating violation of GLM normality.

The primary cost of these violations pertains to biased regression coefficients and standard errors and in particular, inflated standard errors. While reliability is a predictor-related issue, the outcomes generated from our model do not reflect true estimated relationships. As such, it is necessary to correct any violations, having identified and diagnosed the violations.

Proposed model corrections

Having identified violations of linearity, homoscedasticity and normality, correctional measures can be implemented to modify our initial regression model.

1. ***Transformation of select variables***

   The relationship between AGE97 and mathraw97 is a nonlinear. While the linearity of Model 1 is not violated significantly, we can correct the non-linearity of the age variable by centering and squaring AGE97. Similarly, family income is curvilinear and right skewed, we log the family income variable. Together, both corrections address violations of linearity, homoscedasticity and normality. Coding for each transformation is detailed in the Appendix.

2. ***Considering an additional variable***

   Another relevant variable we can consider from the PSID dataset is the nature of parenting practices, HOME97. This might reveal further insight into how home environment and upbringing impact a child's math scores, in addition to the other relevant variables included in Model 1. Post modification of these two variables, we generate Model 2. Regression results for this amended model are detailed in Table 3 above.

   a. ***Added variable plot***

      An AV plot of parenting practices (HOME97) as a dependent variable helps us assess the effect of this variable in our linear model. The lack of slope in the plot (Figure 4) indicates that HOME97 is not an omitted variable and therefore does not bias the predictive power of our model. Therefore, we omit this variable from our third, final model (Table 3).

   b. ***Assessing multicollinearity***

      Additionally, we assess the multicollinearity (Table 4) of independent variables in Model 2, we can note that there since variance inflation factor (VIF) values are between 1 and 5, the variables are weakly correlated and is therefore not a cause for concern.

Table 4: VIF results for Model 2

Independent Variable	VIF
WIC participation	1.35
Child's age in 1997	1.11
Family income in 1997	1.43
Birth weight status	1.10
Parenting practices	1.40

3. Outlier removal

When examining histogram plots for family income and age, we see highly skewed normality distributions. To remedy these normality violations, these variables are transformed for our proposed linear model. We use Cook's Distance as a measure of influence to gauge how many observations cross the threshold of $4/n$ to determine outliers of our corrected Model 2.

Upon closer examination of the distribution of extreme values of Cook's Distance, we see a notable increase in Cook's Distance values between the 75th and 85th quartile. As there is a marked increase from the continuum of Cook's Distance values up till the 75th percentile, we remove cases above the 85th quartile, i.e. CD = 0.006988796 (Figure 5).

Final Model Observations:

After the three model corrections detailed above, we have our final model. Regression results are detailed in Table 3. We can now consider key differences between Models 1, 2 and the final model.

First, it is worth noting that model Ns are slightly lower, indicating removal of 1) infinite values post log transformation of the family income variable, and 2) removal of outliers using the Cook's Distance method detailed above.

WIC participation decreases math scores by -2.67 points than not participating in the program, holding all other variables in the model constant. As a child's ages by one year, math scores increase by 2.82 points[4], holding all other variables constant. For every dollar increase in logged family income, we see a marginal increase in math scores by 1.93 points, holding all other variables constant - a more pragmatic way to consider this relationship is a 1% increase in family income increases math scores by 0.00193 points, holding all other variables constant. Low birth weight children tend to have math scores 3.41 points lower than non-low birth weight children, holding all other variables constant.

Standard errors are fairly consistent across the independent variables across the models, however we see significantly larger standard errors post-transformation of variables 'faminc97' and 'AGE97.' Significance levels ($p < .05$) indicate that all independent variables this regression model (Table 3) have a significant impact. In general, participation in the WIC program's effect on child math achievement is

[4] Since AGE97 is transformed into $(AGE97 + 1)^2$, we interpret the coefficients in the new variable's derivative form.

0.850, indicating that our final model explains 85.0% of the total variance in a child's math scores in 1997. While the goodness of fit in our final model is slightly lower than the initial model, the R-squared value is marginally improved after outliers are removed from Model 2, indicating an improvement in predictive accuracy between model 2 and the final model.

When analysing the final model plots (Figure 6), we see that our previously diagnosed violations are corrected. Our fitted residuals is slightly curved downward, potentially due to variable transformations. The normal Q-Q plot displays less variation at the tails, as previously seen in Model 1's Q-Q plot. Similarly, we can note the outlier removal when examining the leverage plots for Model 1 and the final.

4. Concluding Discussion

As detailed in our analyses above, we conclude that the addition of the parenting practices variable has little impact on child math scores. Low birth weight status, followed by WIC participation are the more significant factors influencing math achievement. It is also worth considering some of the study's limitations, and how they might influence our interpretation of results:

1) ***Omitted data***: Observation with missing fields (namely age and reading scores) are removed from our regression models - imputing reading scores in particular results in further predictive inaccuracies. For further analyses, we might consider imputing missing fields, however this could reduce some variability in the data, and could artificially inflate the 'useable' sample size for regression analysis.
2) ***Adjunctive effects:*** While our study focuses on the impact of WIC participation, we include observations where some participants were enrolled in adjunct programs such as the Aid to Families with Dependent Children (AFDC) program. Determining the 'pure' effect of WIC participation therefore varies, given the additional program overlap.
3) ***Duration of WIC participation:*** It is also unclear how long participants stayed enrolled in the program. Early departure (i.e. immediately after delivery) might yield significantly different outcomes than those who continued to stay in the program post-partum.

In all, these results have meaningful policy implications, in terms of redesigning the WIC program. In particular, our analyses underscore the importance of establishing robust initiatives geared toward low birth weight prevention. Example efforts may include subsidised nutrition programs from expectant mothers during pregnancy, routine prenatal advising, promotion of breastfeeding post-delivery etc.

Appendix

Data cleaning and preparation

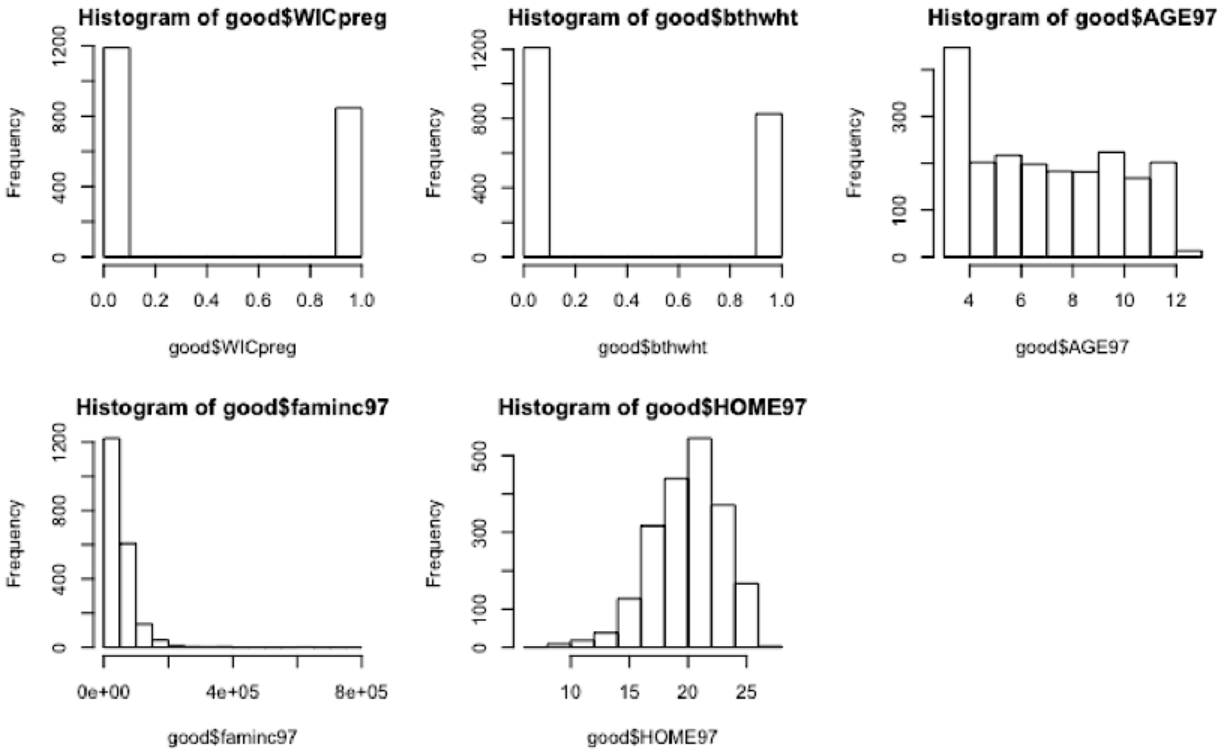
```
library(plyr)
library(dplyr)

#Loading the dataset:
good = read.csv("~/Desktop/good.csv")


#Selecting relevant variables
good = good %>% select("mathraw97", "WICpreg", "AGE97", "faminc97","bthwht",
"HOME97") %>% na.omit(good[,c("mathraw97","WICpreg","AGE97","faminc97","bthwh
t", "HOME97")])
```

Descriptive statistics

```
#Descriptive statistics:
##means, frequencies, standard deviations, and correlations of the variables
summary(good)

##    mathraw97        WICpreg           AGE97           faminc97
##  Min.   : 0.00   Min.   :0.0000   Min.   : 3.000   Min.   :     0
##  1st Qu.:15.00   1st Qu.:0.0000   1st Qu.: 5.000   1st Qu.: 20686
##  Median :36.00   Median :0.0000   Median : 7.000   Median : 40705
##  Mean   :35.83   Mean   :0.4167   Mean   : 7.395   Mean   : 52060
##  3rd Qu.:54.00   3rd Qu.:1.0000   3rd Qu.:10.000   3rd Qu.: 67476
##  Max.   :98.00   Max.   :1.0000   Max.   :13.000   Max.   :784611
##      bthwht          HOME97
##  Min.   :0.000   Min.   : 7.90
##  1st Qu.:0.000   1st Qu.:18.00
##  Median :0.000   Median :20.50
##  Mean   :0.406   Mean   :20.19
##  3rd Qu.:1.000   3rd Qu.:22.40
##  Max.   :1.000   Max.   :27.00

#Standard deviations of relevant variables
##Child age in 1997
sd(good$AGE97)

## [1] 2.922137

##Family income in 1997
sd(good$faminc97)

## [1] 53175.04

##Math scores in 1997
sd(good$mathraw97)

## [1] 22.28801
```

```
##Parenting practices
sd(good$HOME97)

## [1] 3.068104

#Correlation matrix of relevant variables
corr = cor(good)
corr

##            mathraw97     WICpreg       AGE97    faminc97      bthwht
## mathraw97  1.0000000 -0.18295165  0.91987832  0.15834114  0.12361334
## WICpreg   -0.1829516  1.00000000 -0.08530692 -0.38393247  0.12644926
## AGE97      0.9198783 -0.08530692  1.00000000  0.05414526  0.20405292
## faminc97   0.1583411 -0.38393247  0.05414526  1.00000000 -0.10282942
## bthwht     0.1236133  0.12644926  0.20405292 -0.10282942  1.00000000
## HOME97     0.3122676 -0.40249151  0.19893080  0.39448347 -0.09637508
##               HOME97
## mathraw97  0.31226757
## WICpreg   -0.40249151
## AGE97      0.19893080
## faminc97   0.39448347
## bthwht    -0.09637508
## HOME97     1.00000000

#Frequencies of independent variables
par(mfrow=c(2,3))
##WIC Participation
hist(good$WICpreg)
##Low birth weight status
hist(good$bthwht)
##Child age in 1997
hist(good$AGE97)
##Family income in 1997
hist(good$faminc97)
##Parenting practices
hist(good$HOME97)
```

Figure 1: Frequency plots of relevant variables

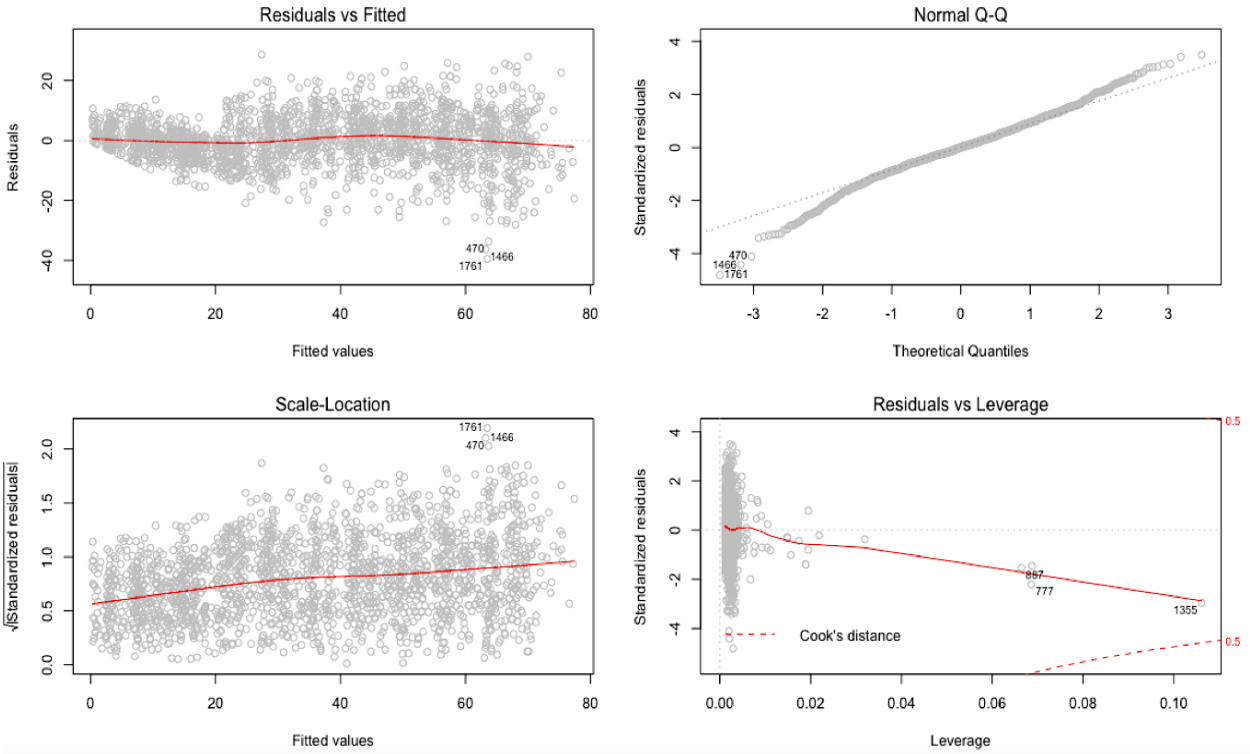


Model 1

```
#Model 1 regression
lm = lm(mathraw97 ~ WICpreg + AGE97 + faminc97 + bthwht, data=good)
summary(lm)

##
## Call:
## lm(formula = mathraw97 ~ WICpreg + AGE97 + faminc97 + bthwht,
##     data = good)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.478  -4.573   0.087   4.994  28.613
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.555e+01  5.863e-01 -26.517  < 2e-16 ***
## WICpreg     -3.113e+00  4.023e-01  -7.737 1.59e-14 ***
## AGE97        7.013e+00  6.394e-02 109.683  < 2e-16 ***
## faminc97     3.238e-05  3.707e-06   8.733  < 2e-16 ***
## bthwht      -2.149e+00  3.826e-01  -5.616 2.22e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.204 on 2037 degrees of freedom
```

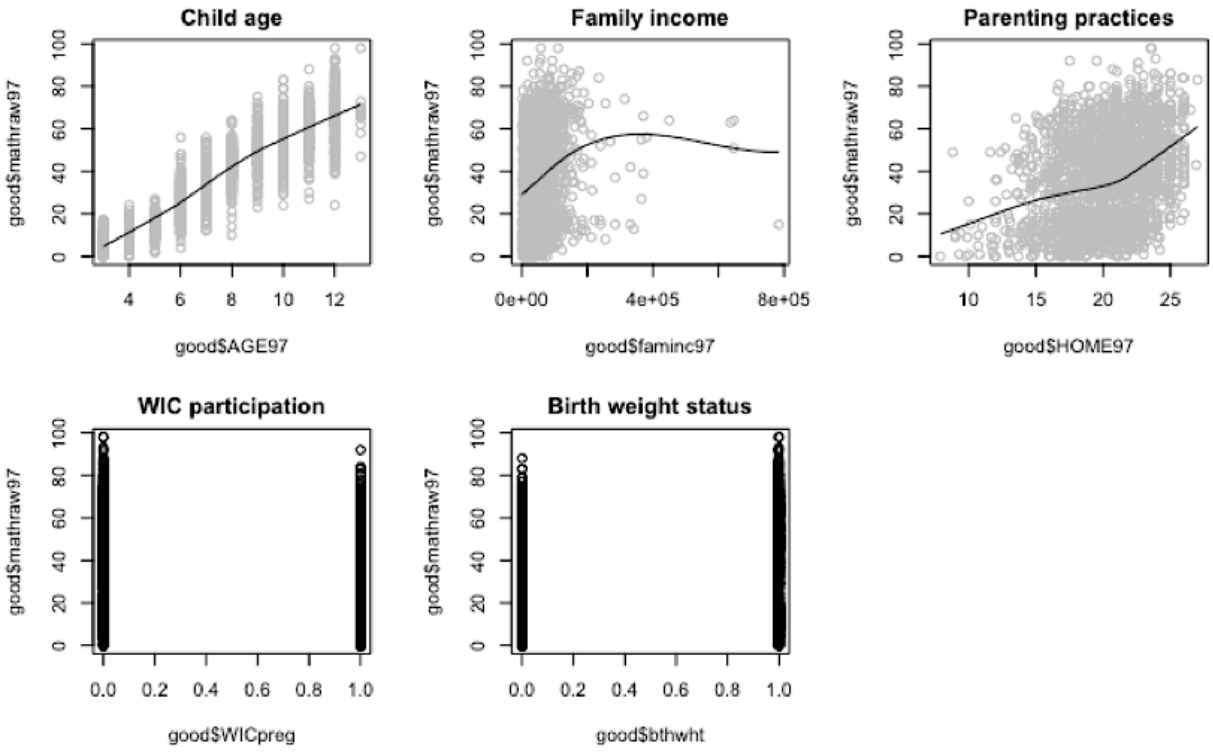
```
## Multiple R-squared:  0.8648, Adjusted R-squared:  0.8645
## F-statistic:  3257 on 4 and 2037 DF,  p-value: < 2.2e-16

#Relevant plots for Model 1:
par(mfrow=c(2,2))
plot(lm, col='grey')
```

Figure 2: Model 1 plots




```
#Correlation plots for Model 1:
par(mfrow=c(2,3))
scatter.smooth(good$AGE97, good$mathraw97, main="Child age", col='grey')
scatter.smooth(good$faminc97, good$mathraw97, main="Family income", col='grey
')
scatter.smooth(good$HOME97,good$mathraw97, main="Parenting practices", col='g
rey')
plot(good$WICpreg, good$mathraw97, main="WIC participation")
plot(good$bthwht,good$mathraw97, main="Birth weight status")
```

Figure 3: Model 1 correlation plots of independent variables



Model 1 Corrections

```
#Transformation of faminc97
good$loginc = log(good$faminc97)
good = good[is.finite(rowSums(good)),]

#Transformation of AGE97
good$AGE97R = (good$AGE97+1)^2
```

Model 2

```
lm2 = lm(mathraw97 ~ WICpreg + AGE97R + loginc + bthwht + HOME97, data=good)
summary(lm2)

#
## Call:
## lm(formula = mathraw97 ~ WICpreg + AGE97R + loginc + bthwht +
##     HOME97, data = good)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.374  -5.313  -0.349   5.404  28.533
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18.17309    2.40319  -7.562 5.97e-14 ***
## WICpreg      -1.83349    0.45684  -4.013 6.20e-05 ***
```

```
## AGE97R        0.39646    0.00407  97.420  < 2e-16 ***
## loginc        0.80901    0.22077   3.665 0.000254 ***
## bthwht       -3.19244    0.41364  -7.718 1.84e-14 ***
## HOME97        0.80611    0.07505  10.742  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.75 on 2030 degrees of freedom
## Multiple R-squared:  0.8465, Adjusted R-squared:  0.8461
## F-statistic:  2239 on 5 and 2030 DF,  p-value: < 2.2e-16
```

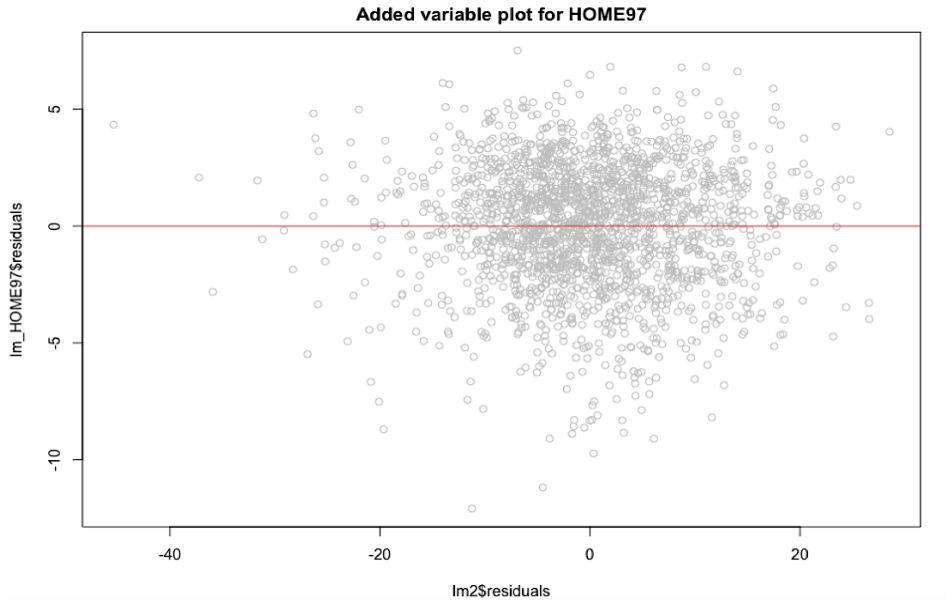
Model 2: Multicollinearity Tests

```
#VIF for Model 2
vif(lm2)

##  WICpreg   AGE97R   loginc   bthwht   HOME97
## 1.348239 1.112939 1.427937 1.097375 1.396471
```

Model 2: Added Variable Plot

```
#Residuals from Model 2
lm2$residuals

#Residuals where HOME97 is the dependent
lm_HOME97 = lm(HOME97 ~ WICpreg + AGE97R + faminc97 + bthwht, data=good)


#Plotting the AVP
plot(lm2$residuals, lm_HOME97$residuals, main='Added variable plot for HOME97
', col='grey')
abline(h=0, col="red")
```

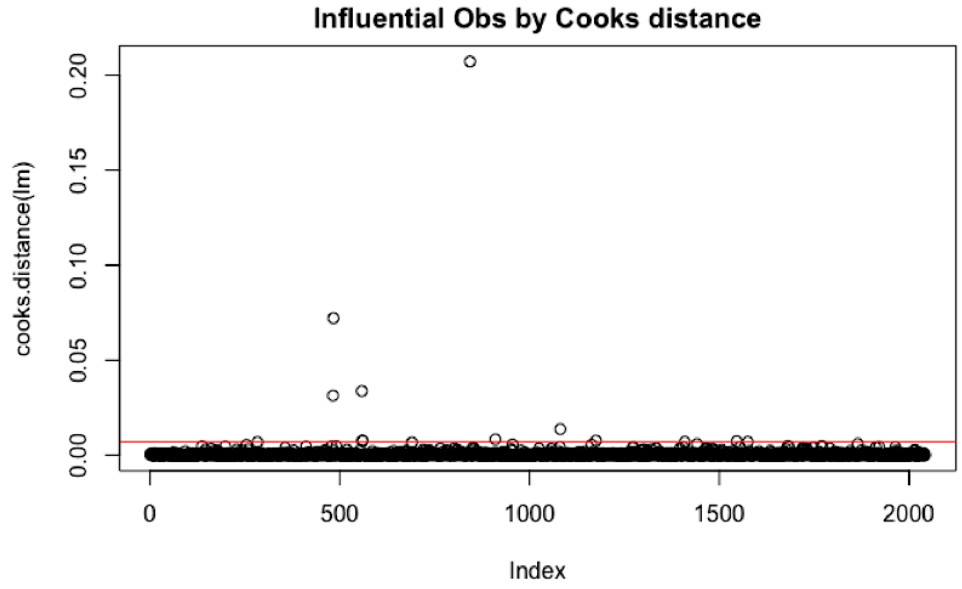
Figure 4: Added variable plot of HOME97



Outlier removal

```
#outlier information
outliers = good[,c("mathraw97","WICpreg","AGE97R","loginc","bthwht")]
outliers$cd = cooks.distance(lm2)

#Identification of all cases with Cook's D above 4/1964.
large_cd = subset(outliers, cd > (4/1964))

#Setting the cutoff to 85%
large_cd2 = subset(outliers, cd > 0.006988796)
#Use the View() function to examine observations in this object
View(large_cd2)

#Plotting Cook's D with newly set cutoff
cd = cooks.distance(lm)
plot(cooks.distance(lm), main="Influential Obs by Cooks distance")
abline(h = 0.006988796, col="red")
```

Figure 5: Identification of outliers using Cook's Distance



Model 3

```
lm3<-lm(mathraw97 ~ WICpreg + AGE97R + loginc + bthwht, data=subset(outliers,
cd < 0.006988796))
summary(lm3)

##
## Call:
## lm(formula = mathraw97 ~ WICpreg + AGE97R + loginc + bthwht,
##     data = subset(outliers, cd < 0.006988796))
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.499  -5.483  -0.445   5.439  31.345
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14.059908   2.514705  -5.591 2.56e-08 ***
## WICpreg      -2.674033   0.451090  -5.928 3.60e-09 ***
## AGE97R        0.409269   0.004004 102.212  < 2e-16 ***
## loginc        1.929271   0.229266   8.415  < 2e-16 ***
## bthwht       -3.406883   0.409806  -8.313  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.674 on 2016 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8494
## F-statistic:  2850 on 4 and 2016 DF,  p-value: < 2.2e-16
```

```
#Model plots for final model:
par(mfrow=c(2,2))
plot(lm3, col='grey')
```

Figure 6: Final model plots