<u>Applied Econometrics project</u>:

A Structural Model of Birth Weight:

Factors Influencing Birth Weight

Yu JIANG Health Economics

Xiaolan WU Applied Economics

2022 - 2023

We declare on our honors that this econometrics project on Excel has been written personally, without any unauthorized external help, and that it has never been presented before or published anywhere, even partially.

Date: 10/11/2022

Signature:

Yu JIANG Yu JIANG

Xiaolan WU Xiaolan Wu

Professors:

Rosalind Bell-Aldeghi

Mira Rahal

Université Paris Cité

Table of Contents

The factors which influence the birth weight of a newborn

I. INTRODUCTION

When a newborn arrives, one of the first things you are told is the baby's weight. Knowing the baby's weight not only helps to inform others but is also a clear indicator of the baby's health, but not all babies weigh the same, in well to being a measure of health condition, birth weight is also an indication of economic and social welfare, so, we were interested in looking at what factors affect the baby's weight.

Using representative data collected in the early 1990s, this paper investigates the birth weight of infants and several factors that influence it and discusses their relevance. We found that maternal behavior during pregnancy was strongly associated with infant birth weight, and other factors significantly associated with birth weight included the mother's living environment, such as family income level, and the gender of the infant.

In this paper, we have selected several variables related to the birth weight of infants, including family income, the gender of the infant, the mother's smoking level and the mother's education level, and we investigate the correlation between each explanatory variable and the explanatory variable through some simple linear regression equations.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature on the determinants of birth weight, followed by the justification of the econometric model, presentation of the data, description of the data, and econometric analysis which includes the discussion of the results in sections 3 to 5 respectively. This is followed by the conclusion of the paper.

II. RELATED LITERATURE

The literature on birth weight is immense and we have found a selection of economic studies on this topic. In general, since the structural production model of birth weight was first proposed, many studies have been carried out along the same lines.

The first is the variable that underlies our study, cigarette, and the variable that is of most interest to us. The number of mothers' daily smoking during pregnancy was used as one of the explanatory variables to explain not only his negative effect on baby's birth weight, cause the harmful effects of smoking on the mother's body can affect the health of the fetus (Shengfang et.al, 2020), but also progresses to analyze the effect on other health indicators of the infant from a medical professional perspective (Magdalena Chełchowska et.al, 2020).

The baby's birth weight also depends on the level of the family income, there is some variation in birth weight due to differences in family income, and the cost of raising a child is one of the main factors determining the fertility rate (Esther Toledano et al,. 2013). For example, having a family allowance is equivalent to a certain increase in family income, which reduces the cost of raising a child, so that the quality of life of the mother during pregnancy is improved and is a certain possibility to increase the baby's birth weight, with low-birth-weight babies usually coming from low-income families (Roni Frish et al,. 2013).

Gender is another explanatory variable that influences baby's birth weight, some biological differences between males and females result in most male infants weighing more than female infants, and its correlates different national situation which affects the whole social environment and economic development, presenting the social nature of gender differences between men and women, and finally affecting the biological nature of gender (Germano Mwabu, 2008).

Finally, maternal education is strongly associated with poor perinatal results, including preterm birth, and poorer health at birth among infants of mothers with low levels of education (Arnaud

Chevalier and Vincent O'Sullivan, 2007). Further, higher educated pregnant women are more likely than lower educated pregnant women to change their minds about the physical, psychosocial, and behavioral factors that influence pregnancy, for example, poor nutrition, smoking and manual labor (B. K. L. Walawalkar Rural Medical College et al,. 2019).

III. Justification of the econometric model

Simple regression linear equation:

1) Variable: cigarette

We identified the first simple linear regression model based on the explanatory variable X = the number of cigarettes smoked per day by the mother, the explanatory variable Y = birth weight, and the error term U:

$$\mathbf{Y_i=\beta_0+\beta_1X_{1i}+U_i \qquad i = 1, 2, \ldots\ldots, n \qquad (1)}$$

The hypothesis of this relation:

We know that smoking is harmful to health. Based on reading the relevant literature and general knowledge of life, we can propose the hypothesis that the number of cigarettes smoked per day by the mother has a negative effect on the baby's birth weight, and we will test the hypothesis and the significance in the following by analyzing specific data from models of Cigarette Smoking Behavior.

Multiple Linear Regression:

2) Variable: family income

As family income is also a factor that influences birth weight, we added the variable of family income to examine the relationship between the variable of family income and the birth weight and go on to discuss whether and what correlation still existed between the number of cigarettes smoked per day by the mother and the birth weight.

$$\mathbf{Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+U_i \qquad i = 1, 2, \ldots\ldots, n \qquad (2)}$$

The hypothesis of this relation:

Based on our reading of the literature, we suggest that family income is positively associated with a baby's birth weight. With the addition of a variable, at the same time, we hypothesized the correlation between the number of cigarettes smoked per day by the mother and the baby's birth weight would change. However, the coefficient for the smoking correlation will change with the addition of the second variable, and we suspect that the correlation will be relatively lower but does not change the positive or negative correlation itself.

3) Variable: male

It is well known that there is a certain difference in body size between males and females, so it is reasonable to suppose that there is a difference in the weight basis between males and females at birth.

$$\mathbf{Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}+U_i \qquad i = 1, 2, \ldots, n \qquad (3)}$$

The hypothesis of this relation:

In this model, in addition to the number of cigarettes and family income already studied, we add a new variable - the sex of the child. We know from extensive data that the birth weight of male infants is generally greater than that of female infants. Then, we predict that the effect of the other two variables on the study objective of birth weight will decrease with the addition of the child's sex variable but does not change the positive or negative correlation itself.

4) Variable: mother's education

The last variable we selected to correlate with the birth weight of the infant was the mother's level of education.

$$\mathbf{Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}+\beta_4X_{4i}+U_i \qquad i = 1, 2, \ldots, n \qquad (4)}$$

The hypothesis of this relation:

Based on the literature we read, we suggest that the lower the mother's level of education, the higher the possibility that the infant will be born in poor health, so we suspect that there is a positive correlation between the two, and we will continue to examine whether the correlation between the previous variables on birth weight changes with the inclusion of this variable. We predict that the effect of the other variables will decrease but does not change the positive or negative correlation itself.

IV. DATA

1) Presentation of data

The data we use are derived from Wooldridge Source: J. Mullahy (1997), “Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior,” Review of Economics and Statistics 79, 596-593. Professor Mullahy kindly provided the data. He obtained them from the 1988 National Health Interview Survey. Data loads lazily. A data. Frame with 1388 observations on 14 variables. It contains 13 explanatory variables and 1 explanatory variable (bwght), and some log transformations of the variables.

Variable	Description of the variable
bwght	birth weight, ounces
cigs	cigs smoked per day while preg
cigtax	cig. tax in home state, 1988
cigprice	cig. price in home state, 1988
faminc	1988 family income, $1000s
male	=1 if male child
motheduc	mother's yrs of educ
fatheduc	father's yrs of educ
parity	birth order of child
white	=1 if white
lbwght	log of bwght
bwghtlbs	birth weight, pounds
packs	packs smoked per day while preg
lfaminc	log(faminc)

Table 1

In this project, we have selected four variables out of these fourteen variables, which are: 10^{th} cigarettes; 1^{st} family income; 8^{th} male; 6^{th} mother’s education. Because we think about these four variables are of interest to us, particularly the number of cigarettes smoked by the mother in a day during pregnancy. We will create four equations to explain these four variables with the explanatory variable baby's birth weight.

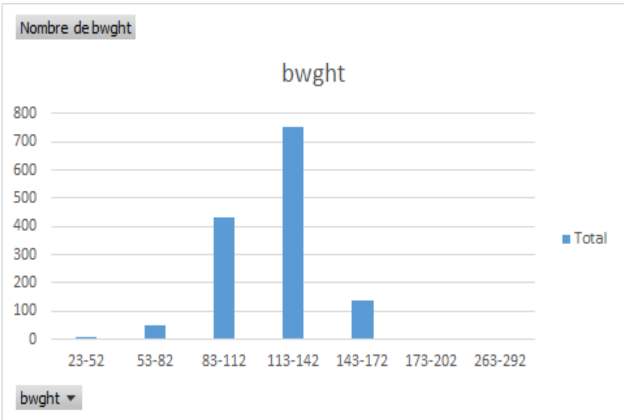
2) Description of data

Dictionary of variables	Definition	Nature	Unit
bwght	birth weight	discrete quantitative[23, 271]	ounces
cig	cigarettes smoked per day while pregnacy	discrete quantitative[0, 50]	1 piece
faminc	1988 family income	continuous quantitative[0.5, 65]	$1000
male	gender	nominal qualitative	1 if male child
motheduc	mother's years of education	discrete quantitave[2, 18]	1 year

Table 2

Y: bwght	
Moyenne	118.6995677
Erreur-type	0.546329028
Médiane	120
Mode	120
Écart-type	20.35396434
Variance de l'échantillon	414.2838643
Kurstosis (Coefficient d'aplatissement)	3.163342525
Coefficient d'asymétrie	-0.146023516
Plage	248
Minimum	23
Maximum	271
Somme	164755
Nombre d'échantillons	1388
Niveau de confiance(95.0%)	1.07172044

Table 3

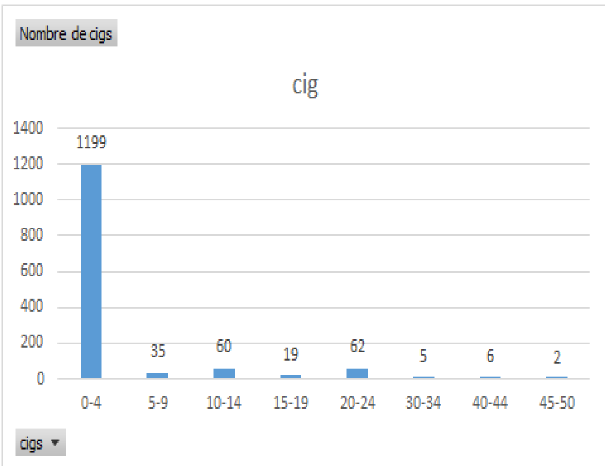


Graph 1

According to descriptive statistical analysis, the mean weight of the 1388 newborns observed was 118.69 ounces, with a maximum value of 271 ounces and a minimum value of 23 ounces. Half of the infants had a birth weight of fewer than 120 ounces, while the other half were greater than 120 ounces.

X1: cig	
Moyenne	2.087175793
Erreur-type	0.160315342
Médiane	0
Mode	0
Écart-type	5.972687881
Variance de l'échantillon	35.67300053
Kurstosis (Coefficient d'aplatissement)	14.99224562
Coefficient d'asymétrie	3.564300836
Plage	50
Minimum	0
Maximum	50
Somme	2897
Nombre d'échantillons	1388
Niveau de confiance(95.0%)	0.31448673

Table 3

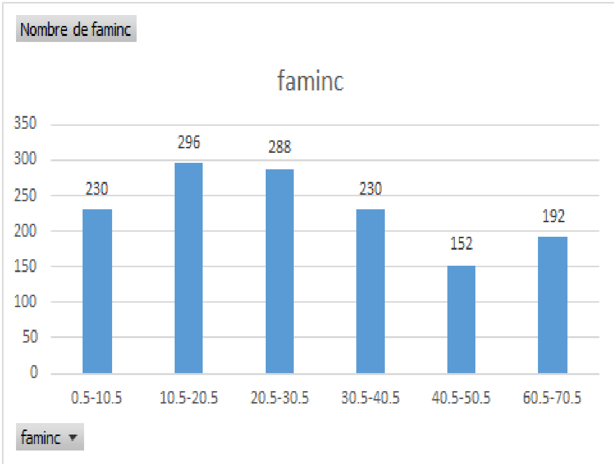


Graph 2

The average number of cigarettes smoked per day during pregnancy was 2.09 for 1388 mothers, with a maximum of 50 cigarettes per day and a minimum of no smoking, with a large variation. The median was 0 indicating that half of the mothers did not smoke during pregnancy, while the other half had a daily smoking habit.

X2: faminc	
Moyenne	29.02665706
Erreur-type	0.502988754
Médiane	27.5
Mode	65
Écart-type	18.73928463
Variance de l'échantillon	351.1607885
Kurstosis (Coefficient d'aplatissement)	-0.524170256
Coefficient d'asymétrie	0.618288847
Plage	64.5
Minimum	0.5
Maximum	65
Somme	40289
Nombre d'échantillons	1388
Niveau de confiance(95.0%)	0.986700873

Table 4

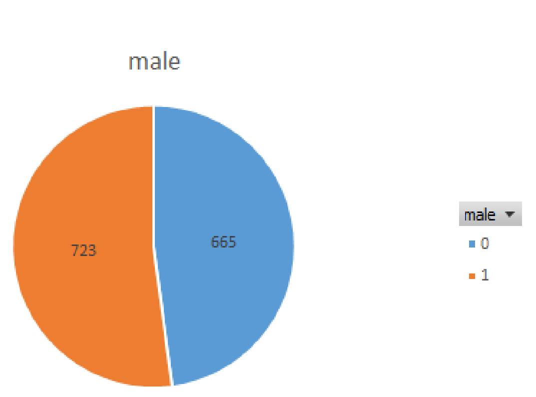


Graph 3

The data show that in 1988 the average annual income of 1388 households was $29,026.66, with a maximum annual income of $65,000 and a minimum annual income of $500. Half of the households had annual incomes below $27,500, while the other half had incomes greater than $27,500, with large differences in income between households.

X3: male	
Moyenne	0.520893372
Erreur-type	0.013413814
Médiane	1
Mode	1
Écart-type	0.499743332
Variance de l'échantillon	0.249743397
Kurstosis (Coefficient d'aplatissement)	-1.995866072
Coefficient d'asymétrie	-0.083737069
Plage	1
Minimum	0
Maximum	1
Somme	723
Nombre d'échantillons	1388
Niveau de confiance(95.0%)	0.026313554

Table 5

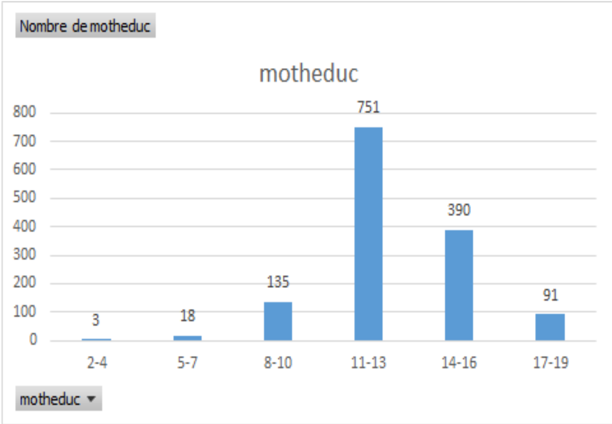


Graph 4

This is a binary variable, which is a quality variable. When the infant is male, the condition is met and is denoted by 1. When the infant is female, the target condition is not met and is denoted by 0. The mean is 0.52, with a similar ratio of males to females, and 1388 infants are roughly equally divided about males and females.

X4: motheduc	
Moyenne	12.93299712
Erreur-type	0.063834742
Médiane	12
Mode	12
Écart-type	2.378218986
Variance de l'échantillon	5.655925546
Kurstosis (Coefficient d'aplatissement)	0.648383261
Coefficient d'asymétrie	-0.031761988
Plage	16
Minimum	2
Maximum	18
Somme	17951
Nombre d'échantillons	1388
Niveau de confiance(95.0%)	0.12522307

Table 6



Graph 5

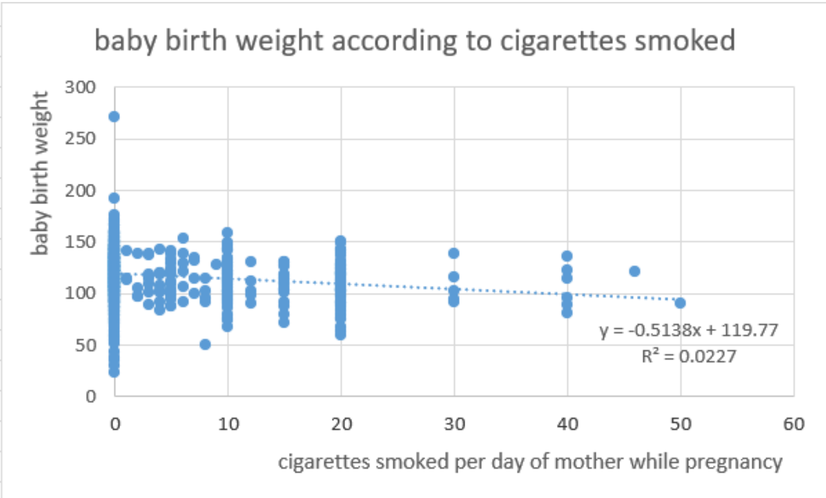
The average number of years of education for the 1388 mothers was 12.93 years, with the highest level of education being 18 years (PhD level) and the lowest level of education being 2 years (not completed elementary school). The median was 12, indicating that half of the mothers had less than 12 years of education and the other half had more than 12 years (high school graduation).

V. Econometric analysis

1) The first equation: simple linear regression **(1)**

$$Y_i = \beta_0 + \beta_1 X_{1i} + U_i \qquad \textbf{avec } i = 1, 2, \ldots, n$$

There are 2 ways to analyze the equation of the fitting line by the OLS (ordinary least squares) between the 2 variables: by the scatterplot or by using the analysis utility in Excel, because we have just one independent variable in this model.



Graph 6

According to the scatterplot (Graph 6), birth weight is negatively correlated with the number of cigarettes smoked per day during the mother's pregnancy: the higher the number of cigarettes smoked, the lower the birth weight. If the mother did not smoke during pregnancy, the infant's birth weight was 119.77 ounces, and the infant's weight decreased by 0.5138 ounces for each additional cigarette smoked, all other things being equal. The coefficient of determination is 0.0227, which means that the model can explain 2.27% of the variations in the situation.

In other words, we can analyze the fit line and the graph of residuals by the analysis utility. In this case, we get more detailed information between the variables of the model.

RAPPORT DÉTAILLÉ						
Statistiques de la régression						
Coefficient de détermination multiple	0.150761803					
Coefficient de détermination R^2	0.022729121					
Coefficient de détermination R^2	0.022024019					
Erreur-type	20.1285784					
Observations	1388					
ANALYSE DE VARIANCE						
	Degré de liberté	*Somme des carrés*	*Moyenne des carrés*	*F*	*Valeur critique de F*	
Régression	1	13060.41937	13060.41937	32.23524054	1.66154E-08	
Résidus	1386	561551.3004	405.1596684			
Total	1387	574611.7197				
	Coefficients	*Erreur-type*	*Statistique t*	*P-value*	*Limite inférieure α = 5%*	*Limite supérieure α = 95%*
Constante	119.7719004	0.572340664	209.2668018	0	118.6491529	120.8946479
cig	-0.513772093	0.09049093	-5.677608699	1.66E-08	-0.691286074	-0.336258112

Table 6

We have the following equation of the regression line:

$$\widehat{bwght_i} = \mathbf{119.771 - 0.514cig_i}$$

$$\mathbf{N = 1388}$$

Interpretation of the estimate:

if the number of cigarette consumption increases by one, the birth weight of the baby decreases by 0.514 ounces, all other things being equal. The weight of a baby whose mother does not smoke during pregnancy is 119.771 ounces.

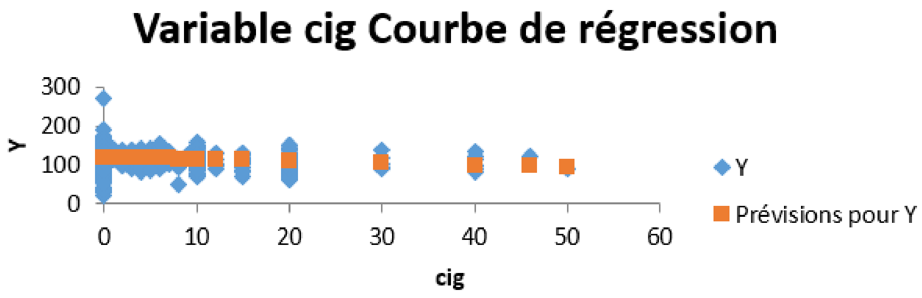
Student's t-test:

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

In table 6 we can find t-statistics for the variable of cigarette which is -5.678, its absolute value is higher than 1,96 (according to the student test for $N > 120$), it is not in the confidence interval [-0.691, -0.336], so we reject H_0 at the 5% risk and this variable is significant.

Since the p-value is 1.66E-08 much less than 0.05, indicating that this variable is significant in this simple linear regression model. Then maternal smoking status during pregnancy has a negative effect on the weight of the newborn.



Graph 7

According to graph 7 of the residuals of the variable cig in relation to the birth weight, it can be observed that the true values of the weight are not perfectly equal to the predicted values of the weight, and therefore the differences between the two values are the residuals.

$$u_i = Y_i - \widehat{Y_i}$$

In the analysis of the variance table (table 6), we have R^2 =0.0227. So, the goodness of fit of the estimated model is not good enough, it explains only 2.27% of the situations. The OLS method offers an imperfect fit, and the knowledge of the values of x (cigarette) does not allow us to predict well the values of y (baby's birth weight). We then try to use a "log-level" model to see if the quality of fit of the model improves.

$$\mathbf{\log (Y_i) = \beta_0+\beta_1X_{1i}+U_i}$$

We were able to get a new analysis table:

RAPPORT DÉTAILLÉ				
Statistiques de la régression				
Coefficient de détermination multiple	0.140674235			
Coefficient de détermination R^2	0.019789241			
Coefficient de détermination R^2	0.019082018			
Erreur-type	0.188834348			
Observations	1388			
ANALYSE DE VARIANCE				
	Degré de liberté	*Somme des carrés*	*Moyenne des carrés*	*F*
Régression	1	0.997780189	0.997780189	27.9816224
Résidus	1386	49.42255745	0.035658411	
Total	1387	50.42033763		
	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*
Constante	4.769403816	0.00536936	888.2630622	0
cig	-0.004490652	0.000848932	-5.289765819	1.4218E-07

Table 7

We have the following equation of the regression line:

$$\widehat{logbwght}_i = 4.77 - 0.0045cig_i$$

$$N = 1388$$

Interpretation of the estimate:

If the number of cigarette consumption increases by one, the birth weight of the baby decreases by 0.45 %, all other things being equal.

Since the p-value is 1.42E-07 much less than 0.05, indicating that the effect of variable cig on the log of the baby's birth weight is very significant in this simple linear regression model.

As the 2 models have the same observations of y and the same number of explanatory variables and, therefore, we can consider the same variance to be explained and directly compare the 2 determination coefficients. In this log-level model, the $R^2 = 0.0198$ is almost equal to that of the level-level model (0.0227), even smaller, which means the number of cigarettes smoked explains 1.98% of the total variation of the log of baby's birth weight observed on the 1388 individuals of the sample. So, the goodness of fit of the model to the empirical data is not very good either. Therefore, we choose the model with the larger coefficient of determination, the level-level model which measures the better quality of adjustment of the model in this case.

In both the original and the changed models, we can conclude that the negative effect of the number of cigarettes smoked by the mother during pregnancy, as the independent variable X_1, on the dependent variable Y - baby birth weight is significant. However, the shortcomings of this model are also evident in that R^2 is small, so the variation explained is minimal. This becomes the reason for us to continue studying other factors next.

2) The second equation: multiple linear regression with 2 variables **(2)**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i \quad \textbf{avec } i = 1, 2, \ldots\ldots, n$$

As we have added a second variable, different from the first simple linear regression equation, we cannot get the linear equation and R^2 directly from the scatterplot, so we directly use the analysis utility.

RAPPORT DÉTAILLÉ

Statistiques de la régression				
Coefficient de détermination multiple	0,172640775		AIC	403,3866043
Coefficient de détermination R^2	0,029804837		SIC	407,9773254
Coefficient de détermination R^2	0,028403833			
Erreur-type	20,06281671			
Observations	1388			

ANALYSE DE VARIANCE

	Degré de liberté	Somme des carrés	Moyenne des carrés	F	Valeur critique de F
Régression	2	17126,20883	8563,104417	21,27391544	7,942E-10
Résidus	1385	557485,5109	402,5166144		
Total	1387	574611,7197			

	Coefficients	Erreur-type	Statistique t	P-value	lower 95%	upper 95%
Constante	116,9741305	1,04898413	111,5118209	0	114,9163611	119,0318999
Cig	-0,46340754	0,091576823	-5,060314648	4,74744E-07	-0,643051806	-0,283763274
Faminc	0,092764738	0,029187869	3,178194982	0,001514708	0,03550753	0,150021947

Table 8

We have the following equation of the multiple linear regression:

$$\widehat{\mathbf{bwght}} = \mathbf{116.974 - 0.463Cig + 0.093faminc}$$

N=1388

Interpretation of the estimate:

The mother doesn't smoke during her pregnancy and her baby's birth weight increases by 0.093 ounces when her family income increases by \$1,000; if family income is 0, the infant's birth weight decreases by 0.463 ounces with one unit increase in the number of cigarettes smoked; if a mother does not smoke and her family income is zero then the birth weight of her child is 116.974ounces, all other things being equal. Compared to the first equation containing only one variable of cigarette smoking, the absolute value of the cigarette smoking coefficient decreases a bit, which indicates that the effect of maternal cigarette smoking on infant birth weight is less significant than before, influenced by the variable of family income.

Fisher's test:

H_0: $\beta_1 = 0$, $\beta_2=0$, non-significant model

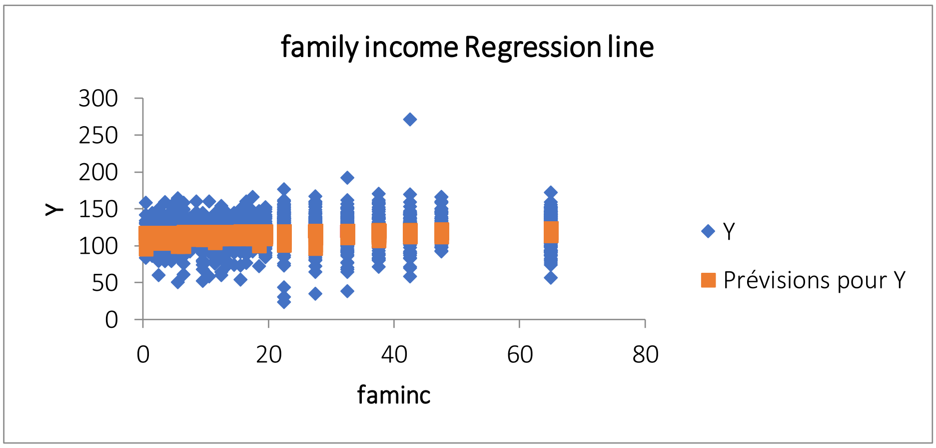
H_1: $\beta_1 \neq 0$, $\beta_2 \neq 0$, significant model

In table 8 we can find F for the variable of family income which is 21.274 >2.99 (according to the Fisher for N > 120, k=2) so we reject H_0 at the 5% risk and this model is significant.

The P-value of the variable cigarette is 4.74744E-07 and the P-value of the variable family income is 0.001514708. The p-value of every variable is much less than 0.05, indicating that 2 variables are significant in this linear regression model.

We then compare the bivariate and univariate level-level models. Even though the 2 models have the same observations of y, they do not have the same number of explanatory variables, so, we cannot directly compare the 2 determination coefficients. But we can calculate other predictive quality indicators like the adjusted R-squared. The adjusted R-Squared of this equation is 0.0284, The greater adjusted R-Squared compared to the previous equation indicates that with the inclusion of the family income variable, this bivariate model is somewhat more significantly representative than the univariate.

Then maternal smoking status during pregnancy has a negative effect on the weight of the newborn, and it is the opposite for family income. But as influenced by the addition of the second variable, the absolute value of the correlation coefficient between smoking and birth weight changed from 0,51 to 0,46, indicating a decrease in the correlation between the mother's smoking during pregnancy and the baby's birth weight, influenced by the family income.



Graph 8

From table 8 and graph 8 we can see that there is some difference between the estimated and real values of birth weight, and the goodness of fit of the estimated model is not good enough, it explains only 2.29% of the situations.

We then try to use a "log-level" model to see if the quality of the model improves. We will use a logistic function model to investigate the percentages that affect birth weight:

$$\mathbf{Log\ (Y_i)=\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i \quad avec\ i = 1, 2, \ldots.. , n}$$

We were able to get a new analysis table:

RAPPORT DÉTAILLÉ						
Statistiques de la régression						
Coefficient de détermination multiple	0,162666362					
Coefficient de détermination R^2	0,026460345		AIC	0,0355179		
Coefficient de détermination R^2	0,025054512		SIC	0,03592211		
Erreur-type	0,188258594					
Observations	1388					
ANALYSE DE VARIANCE						
	Degré de liberté	Somme des carrés	Moyenne des carrés	F	Valeur critique de F	
Régression	2	1,334139543	0,667069772	18,821821	8,60795E-09	
Résidus	1385	49,08619809	0,035441298			
Total	1387	50,42033763				
	Coefficients	Erreur-type	Statistique t	Probabilité	Limite inférieure α = 5%	Limite supérieure α = 5%
Constante	4,743956539	0,009843098	481,95765	0	4,724647547	4,763265532
Cigs	-0,004032558	0,000859307	-4,692801271	2,9616E-06	-0,005718243	-0,002346874
faminc	0,000843747	0,000273883	3,080682519	0,00210595	0,000306476	0,001381018

Table 9

$$\mathbf{\log(\widehat{bwght}) = 4.744 - 0.004Cig + 0.00084faminc}$$

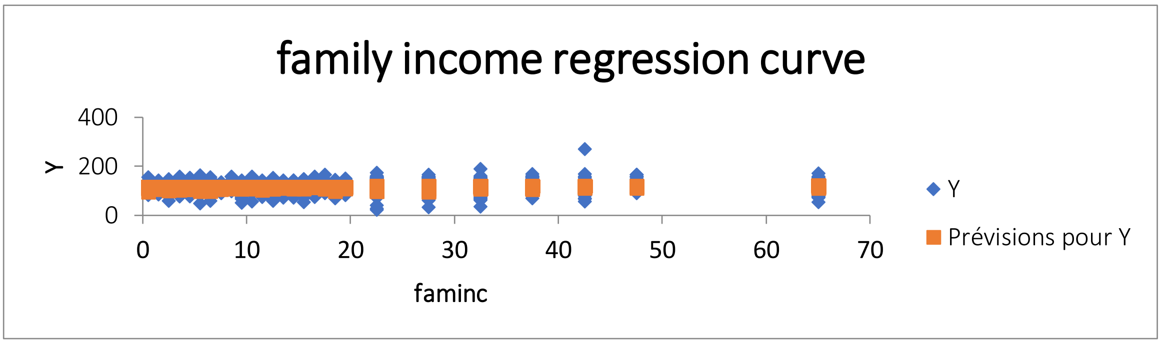
N=1388

Interpretation of the estimate:

The mother does not smoke during her pregnancy and her baby's birth weight increases by 0.084% when her family income increases by $1,000; if family income is zero, the infant's birth weight decreases by 0.04% with one unit increase in the number of cigarettes smoked, all other things being equal.

The effect of cigarettes on the log of birth weight is significant in this simple linear regression model because the p-value is 2.9616E-06, and family income on the log of birth weight is significant also because the p-value is 0.00210595.

At the same time, we can see that the difference between the R^2 =0,0265 of the log-level model and the R^2 =0,0298 of the level-level model is not too much, or even smaller, which means that the fit of the level-level model is relatively higher, and the error term is smaller and closer to the true value.



Graph 9

To investigate whether there are equations that make the estimates more like the real values, so that these explanatory variables can better explain the variation in Y, we tried to use a 'log–log' model:

$$\mathbf{Log\ (Y_i)=\beta_0+\beta_1X_{1i}+\beta_2\log\left(X_{2i}\right)+U_i \qquad avec\ i = 1, 2, \ldots.. , n}$$

We were able to get a new analysis table:

RAPPORT DÉTAILLÉ						
Statistiques de la régression						
Coefficient de détermination multiple	0,160496985					
Coefficient de détermination R^2	0,025759282					
Coefficient de détermination R^2	0,024352436					
Erreur-type	0,188326366					
Observations	1388					
ANALYSE DE VARIANCE						
	Degré de liberté	Somme des carrés	Moyenne des carrés	F	Valeur critique de F	
Régression	2	1,298791701	0,649395851	18,3099541	1,41709E-08	
Résidus	1385	49,12154593	0,03546682			
Total	1387	50,42033763				
	Coefficients	Erreur-type	Statistique t	P-value	lower 95%	upper 95%
Constante	4,71859367	0,018244497	258,631057	0	4,682803835	4,7543835
Cig	-0,004081578	0,000858214	-4,755900294	2,1819E-06	-0,005765117	-0,002398
lfaminc	0,016265688	0,005583314	2,913267922	0,00363405	0,005313023	0,02721835

Table 10

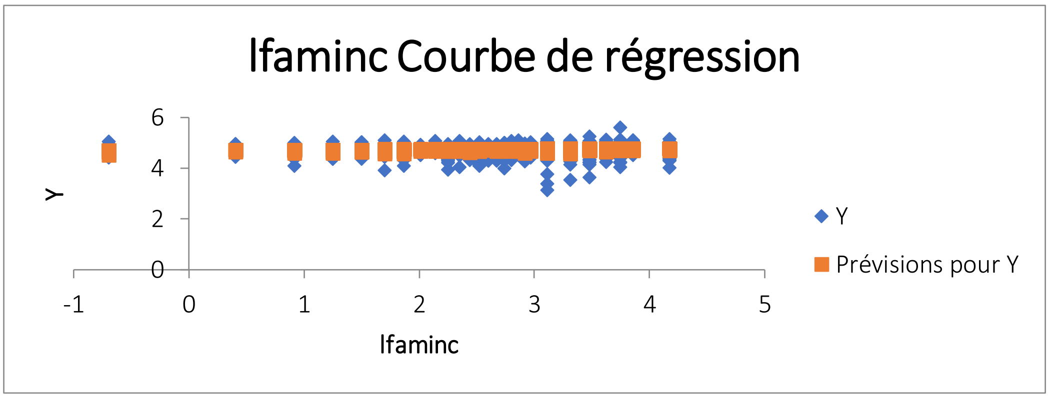
$$\mathbf{\log(\widehat{bwght}) = 4.7186 - 0.004Cig + 0.0163logfaminc}$$

N=1388

Interpretation of the estimate:

The mother doesn't smoke during her pregnancy and her baby's birth weight increases by 0.0163% when her family income increases by 1% ($10); if family income is zero, infant birth weight decreases by 0.04% with one unit increase in number of cigarettes smoked also, all other things being equal.

The effect of cigarettes on the log of birth weight is very significant in this simple linear regression model because the p-value is 2.9616E-06, and family income on the log of birth weight is significant also because the p-value is 0.00210595.



Graph 10

From table 10 and graph 10, we can see that $R^2 = 0.02576$, which is still in a decreasing situation compared to the previous log level model with $R^2 = 0{,}02646$, therefore, by transforming the three models, we believe that the level-level model is relatively accurate and the goodness fit of the regression equation is better.

In both the original and the changed models, we can conclude that the negative effect of the number of cigarettes smoked by the mother during pregnancy, as the independent variable X1, on the dependent variable Y— baby birth weight is significant, and the positive effect of the family income on baby's birth weight. The shortcomings of this model are also evident in that R2 is small, so the variation explained is minimal, although the coefficient of determination has increased compared to the previous two models. In the next step, we continue to add new independent variables and investigate whether the new model can further improve the quality.

3) The third equation with 3 variables: **(3)**

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}+U_i$$

We have the following equation of the regression line:

$$\widehat{bwght}_i = 115.23 - 0.461cig_i + 0.0969faminc_i + 3.114male_i$$

$$N = 1388$$

RAPPORT DÉTAILLÉ

Statistiques de la régression			
Coefficient de détermination multiple	0.188776182		
Coefficient de détermination R^2	0.035636447	AIC	401.540116
Coefficient de détermination R^2	0.033546063	SIC	407.644587
Erreur-type	20.00965446		
Observations	1388		

ANALYSE DE VARIANCE

	Degré de liberté	Somme des carrés	Moyenne des carrés	F	Valeur critique de F
Régression	3	20477.11998	6825.70666	17.0478039	7.10494E-11
Résidus	1384	554134.5998	400.3862715		
Total	1387	574611.7197			

	Coefficients	Erreur-type	Statistique t	P-value	Limite inférieure α = 5%	Limite supérieure α = 5%
Constante	115.2277082	1.207880175	95.39663828	0	112.8582344	117.597182
cig	-0.4610457	0.091337813	-5.047698062	5.0662E-07	-0.640221217	-0.281870183
faminc	0.096879835	0.02914526	3.324034017	0.00091045	0.039706175	0.154053494
male	3.113967891	1.076396465	2.892956258	0.00387607	1.002422981	5.225512801

Table 11

Interpretation of the estimate:

The weight of a female baby whose mother does not smoke during pregnancy and has zero family income is 115.23 ounces, if the number of cigarette consumption increases by one, the family income is zero and the baby is female, the birth weight of the baby decreases by 0.46 ounces; if the family income increases by $1,000, the mother does not smoke and baby is female, the birth weight of baby increases by 0.097 ounces; if the baby is male, the mother does not smoke and the family income is zero, the birth weight of baby increases by 3.11 ounces, all other things being equal.

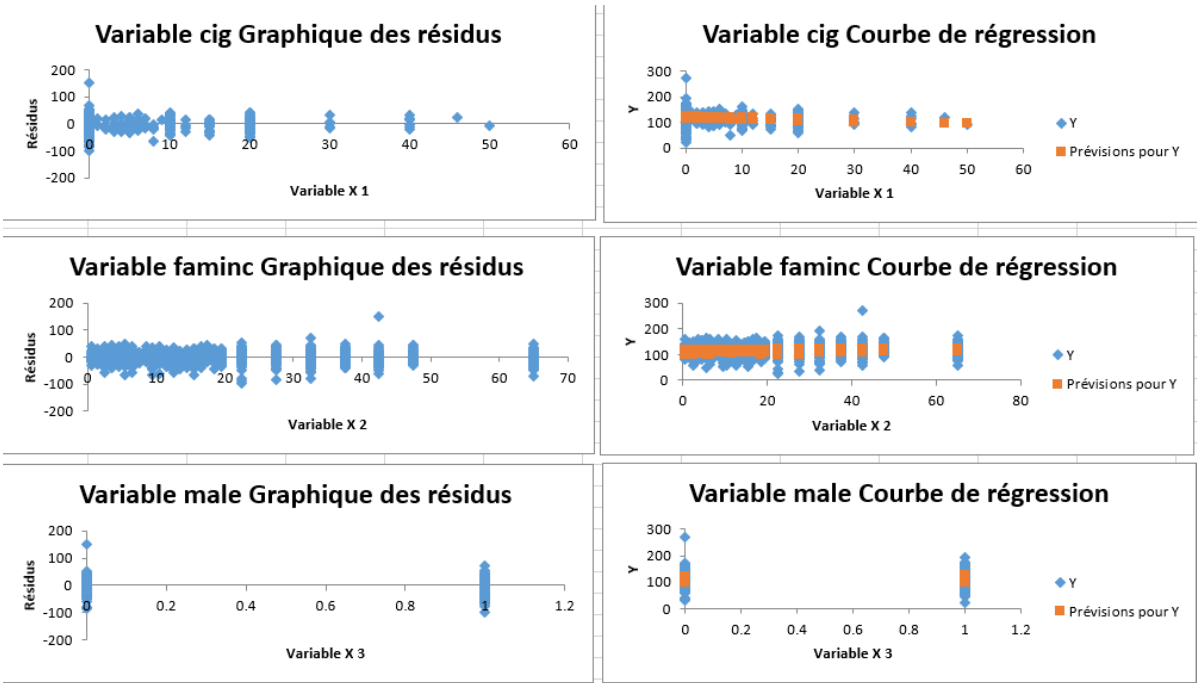
Fisher's test:

H_0: $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$ non-significant model

H_1: $\beta_1 \neq 0$, $\beta_2 \neq 0$, $\beta_3 \neq 0$ significant model

In table 11 we can find F for the variable of family income which is 17.048 > 2.60 (according to the Fisher for N > 120, k=3) so we reject H_0 at the 5% risk and this model is significant.

Since the p-value of every variable is much less than 0.05, indicating that 3 variables cig, faminc and male are significant in this linear regression model. Then maternal smoking status during pregnancy has a negative effect on the weight of the newborn, and it's the opposite for family income and male infants.



Graph 11

According to the regression curve and the graph of the residuals of the variables---cig, faminc and male in relation to the baby's birth weight, it can be observed that the true values of the weight are not perfectly equal to the predicted values of the weight, and therefore the differences between two values are the residuals.

In the internal comparisons of this model, gender has the most direct and important effect on weight, followed by the number of mothers smoking and finally by family income. This result can be interpreted as the infant's weight is the most correlated with its own nature, followed by the close physiological correlation between mother and infant, where the mother's health status tends to be reflected in the child, and finally by external factors similar to family status, where a higher family income the more likely the mother is to have a healthy and stable life status, then there is some positive effect on the birth weight of the newborn.

Compared to the previous two univariate and bivariate models, the trivariate model has increased quality and the ability to explain the changes explained. Comparing the coefficients of different variables across models, the coefficient on smoking decreased continuously from -0.51 to -0.4634 to -0.4610, and the coefficient on family income increased from 0.0928 to

0.0969, indicating that the addition of new variables reduced the effect of smoking on body weight and increased the effect of income and gender on body weight, but the overall correlation remained the same.

In the analysis of the variance table, we have R^2 = 0.0356. So, the goodness of fit of the estimated model is not good enough, it explains only 3.56% of the situations. The OLS method offers an imperfect fit, and the knowledge of the values of cigarettes, family income, and gender of the baby does not allow for predict well the values of the baby's birth weight. We then try to use a "log-level" model to see if the quality of the model improves.

$$\log(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i$$

RAPPORT DÉTAILLÉ				
Statistiques de la régression				
Coefficient de détermination multiple	0.176333088			
Coefficient de détermination R^2	0.031093358			
Coefficient de détermination R^2	0.028993127			
Erreur-type	0.187877943			
Observations	1388			
ANALYSE DE VARIANCE				
	Degré de liberté	*Somme des carrés*	*Moyenne des carrés*	*F*
Régression	3	1.567737614	0.522579205	14.8047313
Résidus	1384	48.85260002	0.035298121	
Total	1387	50.42033763		
	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*
Constante	4.729375031	0.011341227	417.0073345	0
cig	-0.004012838	0.000857604	-4.679127179	3.163E-06
faminc	0.000878105	0.000273655	3.208799012	0.00136346
male	0.025999638	0.010106679	2.572520411	0.01019956

Table 12

We have the following equation of the regression line:

$$\widehat{logbwght_i} = 4.73 - 0.004cig_i + 0.0009faminc_i + 0.0259male_i$$

N = 1388

Interpretation of the estimate:

If the number of cigarette consumption increases by one, the family income is zero and baby is female, the birth weight of the baby decreases by 0.4 %; if the family income increases by \$1,000, the mother does not smoke and baby is female, the birth weight of baby increases by 0.09 %; if the baby is male, the mother does not smoke and the family income is zero, the birth weight of baby increases by 2.59 %, all other things being equal.

Since p-value of every variable much less than 0.05, indicating that the effect of 3 variables cig, faminc and male on the log of the baby's birth weight is very significant in this linear regression model. The different coefficients represent the closeness of the association between the different factors and the dependent variable: the effect of the infant's sex on its birth weight is the most direct, with male infants usually being 3.11ounces heavier than female infants; followed by the negative effect of the mother's smoking habit during pregnancy on the health of the newborn; in contrast, the effect of household income on the infant's weight is present, but the relationship is not as strong.

As the 2 models have the same observations of y and the same number of explanatory variables and, therefore, we can consider the same variance to be explained and directly compare the 2 determination coefficients. In this log-level model, the $R^2 = 0.0311$ is almost equal to that of the level-level model (0.0356), even smaller, which means the model explains 3.11% of the total variation of the log of baby's birth weight observed on the 1388 individuals of the sample. So, the goodness of fit of the model to the empirical data is not very good either. Therefore, we choose the model with the larger coefficient of determination, the level-level model which measures the better quality of adjustment of the model in this case.

We then compare the bivariate and trivariate level-level models. Even the 2 models have the same observations of y, but they do not have the same number of explanatory variables, so, we cannot directly compare the 2 determination coefficients. But calculate other predictive quality indicators just like the adjusted R-squared, Akaike's Information Criterion (AIC) and Schwartz's Information Criterion (SIC) could be a solution.

The adjusted R^2 of bivariate models is 0.0284, and the adjusted R^2 of trivariate models is 0.0335.
The AIC of bivariate models is 403.39, and the AIC of trivariate models is 401.54.
The SIC of bivariate models is 407.98, and the SIC of trivariate models is 407.64.

Increasing the number of explanatory variables improved the explanatory power of the linear fit, and the more the model minimizes the variance of the residual, the better the fit will be. So, the preferred model will be the model with the highest adjusted R-squared, lowest AIC and SIC. The consideration of these three additional criteria confirms our choice of model with 3 explanatory variables.

In both the original and the changed models, we can conclude that the negative effect of the number of cigarettes smoked by the mother during pregnancy, as the independent variable X_1, on the dependent variable Y— baby birth weight is significant, and the positive effect of the family income and the sex of baby on baby's birth weight. The shortcomings of this model are also evident in that R^2 is small, so the variation explained is minimal, although the coefficient of determination has increased compared to the previous two models. In the next step, we continue to add new independent variables and investigate whether the new model can further improve the quality.

4) The fourth equation with 4 variables: **(4)**

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}+\beta_4X_{4i}+U_i$$

RAPPORT DÉTAILLÉ

Statistiques de la régression			
Coefficient de détermination multiple	0,188776189	AIC	402,11912
Coefficient de détermination R^2	0,03563645	SIC	409,7751778
Coefficient de détermination R^2	0,032847256		
Erreur-type	20,01688727		
Observations	1388		

ANALYSE DE VARIANCE

	Degré de liberté	Somme des carrés	Moyenne des carrés	F	Valeur critique de F
Régression	4	20477,12157	5119,280393	12,77661566	3,24688E-10
Résidus	1383	554134,5982	400,675776		
Total	1387	574611,7197			

	Coefficients	Erreur-type	Statistique t	Probabilité	Limite inférieure α = 5%	Limite supérieure α= 5%
Constante	115,2335505	3,168783834	36,36522922	9,4124E-204	109,0174082	121,4496928
cigs	-0,46107391	0,092459191	-4,986782866	6,91679E-07	-0,642449328	-0,279698493
faminc	0,096908074	0,032412054	2,989877581	0,002840202	0,033325971	0,160490177
male	3,114025115	1,077167752	2,890937934	0,003900939	1,000965851	5,22708438
motheduc	-0,00051287	0,257154407	-0,001994404	0,998408985	-0,504967724	0,503941985

Table 12

The equation of the regression linear equation is as follows:

$$\widehat{\mathbf{bwght}} = \mathbf{115.2336 - 0.4611Cig + 0.0969faminc + 3.1140male}$$
$$\mathbf{- 0.0005motheduc}$$

N=1388

Interpretation of the estimate:

If a mother is a non-smoker, her family income is 0 and the baby is female, the baby's weight decreases by 0.0005 ounces when her year of education increases by one unit (one year). If the mother is a non-smoker, the family income is zero, the sex of the infant is male, and the mother's education level is zero, the infant's birth weight increases by 3.114 ounces from 115.2336 ounces, and if the mother is a non-smoker and the sex of the infant is female, when the family income increases by $1,000, the infant's birth weight increases by 0,0969 ounces, and when the family income is zero, the mother's education level is also zero and the infant is female, the infant's birth weight decreases by 0,4611 ounces when her number of cigarettes per day increases by one, all other things being equal.

The negative coefficient on the variable related to mother's education in this model was surprising to us because in our earlier reviews of the literature, our assumption was that mother's education was positively related to the birth weight of the infant, whereas our data showed a negative relationship between mother's education and birth weight in the absence of testing errors.

Fisher's test:

H_0: $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$, $\beta_4 = 0$ non-significant model

H_1: $\beta_1 \neq 0$, $\beta_2 \neq 0$, $\beta_3 \neq 0$, $\beta_4 \neq 0$ significant model

In the table 12 we can find F for the variable of family income which is $12.7766 > 2.37$ (according to the Fisher for $N > 120$, $k=4$) so we reject H_0 at the 5% risk and this model which have four variables is significant.

However, since the p-value of the variable cig, faminc and male less than 0.05, indicating that variables cigarette, family income and sex of are significant in this linear regression model. But the p-value of variable motheduc is more than 0.05, so the variable the mother's education is non-significant of this model. And when compared with the previous equation, we can see that the variable of the mother's education does not affect the correlation coefficients of the other three variables with infant birth weight.

Therefore, although we hypothesized in our speculation that the higher the education of the mother, the higher the birth weight of the baby would also increase, as more educated pregnant women are more likely to change their perceptions of the physical, psychosocial and behavioral factors that influence pregnancy than less educated pregnant women, we did not reach this conclusion through our data analysis and we were unable to conclude that there was a definite correlation between the two.

We suppose that the reason for the difference may be a limitation of the data, which comes from 1988, when perhaps most women were at essentially the same level of education and all at a low level, so it is possible that the social environment was such that the level of education of the mothers of that period did not affect the birth weight of their babies.

Since the Student t test concludes that the coefficients are not significant coefficients while the Fisher test concludes that the model is significance of the model, it is the index of multicollinearity, so we make the test of Klein as the detection of the multicollinearity.

The general model:

$bwght=\beta_0+\beta_1 cigs+\beta_2 faminc+\beta_3 male+\beta_4 motheduc+U_i$

The auxiliary models:

a) $Cigs=\beta_0+\beta_1 faminc+\beta_2 male+\beta_3 motheduc$

b) $Faminc=\beta_0+\beta_1 cigs+\beta_2 male+\beta_3 motheduc$

c) $Male=\beta_0+\beta_1 cigs+\beta_2 faminc+\beta_3 motheduc$

d) $Motheduc=\beta_0+\beta_1 cigs+\beta_2 faminc+\beta_3 male$

	R^2	TOL	VIF
general	0,0356364496	0.96436355	1.03695333
a)	0.0527234083	0.94727659	1.05565788
b)	0.2169355637	0.7830644	1.27703412
c)	0.0030900653	0.99690993	1.00309964
d)	0.2276307072	0.77236929	1.29471745

The R^2 of the auxiliary models for cigarette, family income and mother's education are greater than the R^2 of the general model, so there is a multicollinearity problem among the three variables. An R2 of the auxiliary model with male as the explained variable which close to 0 indicates that the independent variables do not explain most of the variation in the dependent variable, which confirms the suspicion that the male variable is not multicollinear with the other variables.

But after calculating the results of the tolerance and variance inflation factor (VIF) for each auxiliary regression, we see that the TOL are close to 1 and VIF are small. This means that a large part of the variance of an explanatory variable that is not explained by the other explanatory variables which does not provide evidence on multicollinearity.

Heteroscedasticity of disturbances

The initial model is estimated by the OLS:

$bwght=\beta_0+\beta_1 cigs+\beta_2 faminc+\beta_3 male+\beta_4 motheduc+U_i$

The auxiliary model is estimated:

$$\hat{u}_i^2 = \lambda_0 + \lambda_1 \text{cigs} + \lambda_2 \text{faminc} + \lambda_3 \text{male} + \lambda_4 \text{motheduc} + v_i$$

Breusch – Pagan test :

H_0 : $V(U_i) = \sigma^2, \forall n$ (Homoscedasticity hypothesis)

H_1 : $V(U_i)=\sigma_i^2 = \lambda_0+\lambda_1 cigs+\lambda_2 faminc+\lambda_3 male+\lambda_4 motheduc$ (Heteroscedasticity hypothesis)

R^2	0,00027650
BP_C	0,38378347
Chi2(4ddl)	9,487729037

BP_C < Chi2, the null hypothesis of homoscedasticity is not rejected.

The random error terms in the overall regression function satisfy the homoscedasticity, that is, they all have the same variance, which conforms to one of the important assumptions of the classical linear regression model and ensures that the regression parameter estimators have good statistical properties.

VI. Conclusion

The purpose of this paper is to study the factors influencing the birth weight of newborns and to select four out of a large amount of data as the main study subjects: the number of mothers smoking during pregnancy, household income, infant gender, and mother's education level. The first simple linear regression equation was established with the number of cigarettes smoked as the base variable, and then some new multiple linear regressions was established by adding variables one after another, and a comparative analysis was made based on the differences within and between the models to draw conclusions and explain the causes.

From the analysis of the data, we conclude that the effects of the first three selected explanatory variables on the study population were significant same as our hypothesis, and that a reduction in smoking by the mother during pregnancy, an increase in household income and male infants were responsible for increasing the birth weight of the infant. The underlying linear relationship was constant regardless of model changes. However, as the variables increase, the parameters of each variable change, either increasing or decreasing, representing a change in the degree of influence of each variable on the study population, and the contribution may increase or decrease. Unexpectedly the fourth variable, the mother's education, unlike our hypothesis, does not reflect the significance of the variable and has the opposite sign of correlation to what we expected.

But in the process of analyzing the results, we found a common problem, the values of R^2 are small, regardless of the number of variables or the type of model (level-level, log-level, log-log). Among all the models studied, the largest one does not exceed 5%, indicating that the variables do not explain more than 5% of the variation. This is not only a conclusion for us but also a limitation and a question for our study. Another shortcoming is that the data empirical tests differ from the hypothesis conclusions, excluding the reasons for analytical errors, and the solution may lie in the selection of data and model building. These will be the questions to be addressed with attention as we continue our study.

VII. References

Daniel Biftu Bekalo & Zelalem Mehari Nigussie & Belay Birlie Yimer & Zelalem Alamrew Anteneh, 2021. "Modeling Determinants of Low Birth Weight for Under Five-Children in Ethiopia," Annals of Data Science, Springer, vol. 8(3), pages 475-496, September.

Del Bono, Emilia & Ermisch, John F & Francesconi, Marco, 2008. "Intrafamily Resource Allocations: A Dynamic Model of Birth Weight," CEPR Discussion Papers 6970, C.E.P.R. Discussion Papers.

Jeffrey M. Wooldridge Introductory Econometrics A Modern Approach 2012.

Ghouse, Ghulam & Zaid, Muhammad, 2016. "Determinants of Low-Birth-Weight a Cross Sectional Study: In Case of Pakistan," MPRA Paper 70660.

Germano Mwabu, 2008. "The Production of Child Health in Kenya: A Structural Model of Birth Weight," Working Papers 963, Economic Growth Center, Yale University.

Rasheed A. Adeyemi & Temesgen Zewotir & Shaun Ramroop, 2016. "Semiparametric Multinomial Ordinal Model to Analyze Spatial Patterns of Child Birth Weight in Nigeria," IJERPH, MDPI, vol. 13(11), pages 1-22, November.

Reuven Shnaps, 2001. "Estimating the Effect of Smoking on Birth Weight in a Dynamic Model when Fertility is a Choice," Computing in Economics and Finance 2001 242.

Shengfang Tang & Zongwu Cai & Ying Fang & Ming Lin, 2020. "A New Quantile Treatment Effect Model for Studying Smoking Effect on Birth Weight During Mother's Pregnancy," WORKING PAPERS SERIES IN THEORETICAL AND APPLIED ECONOMICS 202003, University of Kansas, Department of Economics, revised Feb 2020.

Esther Toledano & Roni Frish & Noam Zussman & and Daniel Gottlieb, Discussion Paper

Series 2009.13, December 2009. “The Effect of Child Allowances on Fertility”. Israel Economic Review Vol. 10, No. 2 (2013), 107-145.

Chevalier, Arnaud, O'Sullivan, Vincent "Mother's education and birth weight” , 2007-06-12 UCD Geary Institute Discussion Paper Series; WP/24/2007. University College Dublin. Geary Institute
http://www.ucd.ie/geary/static/publications/workingpapers/GearyWp200724.pdf

Choudhary V V1, Moruskar A D2. Associate Professor, Assistant Professor, Department of Pediatrics, B. K. L. Walawalkar Rural Medical College, Chiplun, Ratnagiri, Maharashtra, INDIA. “Correlation between maternal education and birth weight of the baby- A hospital-based study” 14/08/2019.

Bansal, *et al*.: Prevalence of low birth weight and its association with risk factors Asian Journal of Medical Sciences, Jan-Feb 2019 | Vol 10 | Issue 1.

Deshmukh JS, Motghare DD, Zodpey SP, Wadhva SK. Low birth weight and associated maternal factors in an urban area. Indian Pediatr 1998;35:33-6.