IRAQI Academic Scientific Journals
العــراقية المـجلات الاكـاديمـيـة الـعلـمـيـة
Al-Kut University College Journal
K. U. C. J.
ISSN (E) : 2616 - 7808 I ISSN (P) : 2414 - 7419
Vol. 6 Issue 1 2021

Lasso Regression in Mediation Analysis with an Application

Zakariya Yahya Mohammed Al-Zubaidi **Assistant. Prof. Dr. Ahmad N. Flaih**

Department of Statistics, College of Administration and Economics, University of Al-Qadisiyah

Iraq – 0069 – Wasit Iraq – 0069 – diwaniyah

zakariyayahya90@yahoo.com ahmad.flaih@qu.edu.iq

Abstract

This paper came with the aim of knowing the causal mediation analysis and estimation methods in addition to the study data. Mediation is the basis for many fields, where at the present time the use of mediation analysis has become increasingly common in many areas of scientific research as well in other fields such as economic and social as well as medical sciences. Causal mediation analysis is one of the important methods in order to know the mechanism of influence between variables, which has had a large role among researchers in recent times. In order to arrive at a model that leads to accurate estimates, it is necessary to find or search for the method by which the important variables are chosen in order to be included in the model, especially when the study data suffers from the presence of the problem of linearity. Therefore, many methods of estimating the mediation variables were used, which are the least squares method and the Lasso method. In order to apply the estimation methods, a simple random sample of (50) women were drawn to study the factors affecting the number of births (response variable). The study data were analyzed using programming in R language.

Keywords: Mediation analysis of least squares, regression of lasso

انحدار لاسو في تحليل الوساطة مع التطبيق

زكريا يحيى محمد الزبيدي **أ. م. د. أحمد نعيم فليح**

قسم الإحصاء، كلية الإدارة والاقتصاد، جامعة القادسية

العراق - 0069 - واسط العراق - 0069 - الديوانية

zakariyayahya90@yahoo.com ahmad.flaih@qu.edu.iq

الملخص:

جاءت هذه الورقة بهدف معرفة طرق تحليل وتقدير الوساطة السببية بالإضافة إلى بيانات الدراسة. الوساطة هي الأساس للعديد من المجالات، حيث أصبح استخدام تحليل الوساطة في الوقت الحاضر شائعًا بشكل متزايد في العديد من مجالات البحث العلمي وكذلك في مجالات أخرى مثل العلوم الاقتصادية والاجتماعية وكذلك العلوم الطبية. يعتبر تحليل الوساطة السببية من الطرق المهمة لمعرفة آلية التأثير بين

المتغيرات والتي كان لها دور كبير بين الباحثين في الآونة الأخيرة. من أجل الوصول إلى نموذج يؤدي إلى تقديرات دقيقة ، من الضروري إيجاد أو البحث عن الطريقة التي يتم من خلالها اختيار المتغيرات المهمة ليتم تضمينها في النموذج ، خاصةً عندما تعاني بيانات الدراسة من وجود مشكلة الخطية. لذلك تم استخدام عدة طرق لتقدير متغيرات الوساطة وهي طريقة المربعات الصغرى وطريقة لاسو. لتطبيق طرق التقدير ، تم سحب عينة عشوائية بسيطة من (50) امرأة لدراسة العوامل المؤثرة على عدد المواليد (متغير الاستجابة). تم تحليل بيانات الدراسة باستخدام البرمجة بلغة R.

الكلمات المفتاحية : تحليل وساطة المربعات الصغرى، انحدار اللاسو

1. Introduction

In this paper we address the idea of causal mediation analysis, which is an important way to learn how to influence variables. We also describe the model of individual mediation and the model of multiple mediation, noting the equations of each model with the geometric shapes. In order to give a model with strong advantages and it includes the most important variables that increase the accuracy of the estimate and thus the power of prediction, which is the usual Lasso regression method in addition to the method of least squares.

2. Mediation analysis

Most of the research used in many fields focuses on the relationship between two variables, where the independent variable X (cause variable) and the other variable is deliberate variable Y (response) when we add a third variable (M) may be difficult then a new variable called median and be Between the independent variable (X) and the dependent variable (Y) called the mediation variable (M), which explains the effect of the cause variable X on the result variable Y, then the independent variable (X) is the cause of the mediation variable (M), which causes the result variable (Y) $(X \longrightarrow M \longrightarrow Y)$ (D.P.Mackinnon & Fairchild & Fritz, 2007) (Grotta & Bllocco, 2012).

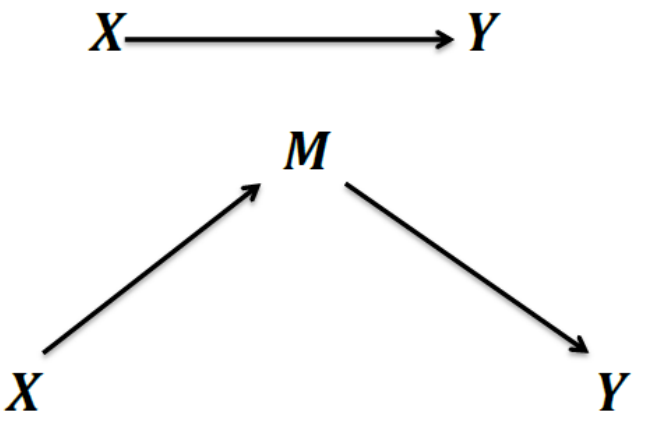

Figure (1) Shows the effects of the variables before and after adding the mediation variable (M).

3. Single - Mediator Model

The individual mediation model is one of the simplest models in the analysis of causal mediation. This model consists of three variables X is the independent variable (treatment variable), M mediation variable, Y dependent variable (result variable) so this model is used when the independent variable is changed to This one intermediate variable leads to an effect in the result. This model (Individual Mediator Model), also called individual mediation or simple mediation, has been applied in several areas, including behavioral, social, and educational sciences, which makes the researcher interested to learn about the side effects not studied (wen, 2013).

3.1 Regression equations used to evaluate mediation

Baron & Kenny (1986) is one of the most common methods used in the analysis of mediation. This was based on the modeling of linear structural equation (LSEM) in order to provide a mediation approach to an individual median model where the following regression equations are required:

$$\boldsymbol{Y} = \mathrm{r}_1 + \mathbf{cX} + \boldsymbol{e_1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$\boldsymbol{M} = \boldsymbol{r}_2 + \boldsymbol{aX} + \boldsymbol{e_3} \quad \ldots\ldots\ldots\ldots\ldots..\ldots\ldots (2)$$

$$\boldsymbol{Y} = \boldsymbol{r_3} + \boldsymbol{c'X} + \boldsymbol{bM} + \boldsymbol{e_2} \quad \ldots\ldots\ldots\ldots\ldots (3)$$

Whereas:

$\boldsymbol{Y}$**:** represents the result of interest.

$\boldsymbol{X}$**:** represents the independent variable (treatment variable).

$\boldsymbol{M}$**:** Represents the mediation variable.

$\boldsymbol{C}$**:** represents the relationship between the processing variable and the result in equation (1) where the total effect is called.

$\boldsymbol{C'}$**:** The parameter that binds as the process variable with the result variable under the influence of the mediation variable is represented in equation (3) where the partial effect is called.

$\boldsymbol{a}$**:** The parameter that transmits the effect of the processing variable on the mediation variable is in equation (2).

$\boldsymbol{b}$**:** The parameter that transmits the effect of the mediation variable on the result variable in equation (3).

$\boldsymbol{r_1, r_2, r_3}$**:** objections to each equation .

$\boldsymbol{e_1, e_2, e_3}$**:** errors .

$\boldsymbol{Equation\ (1):}$ represents the total effect in Figure (A).

$\boldsymbol{Equations\ (2), (3):}$ Determine the mediation model in Figure (B).

(D.P.Mackinnon & Pirlott , 2015) (Pachali & Mayer & landwehr , 2018) (Tofighi & Thoemmes , 2014) .

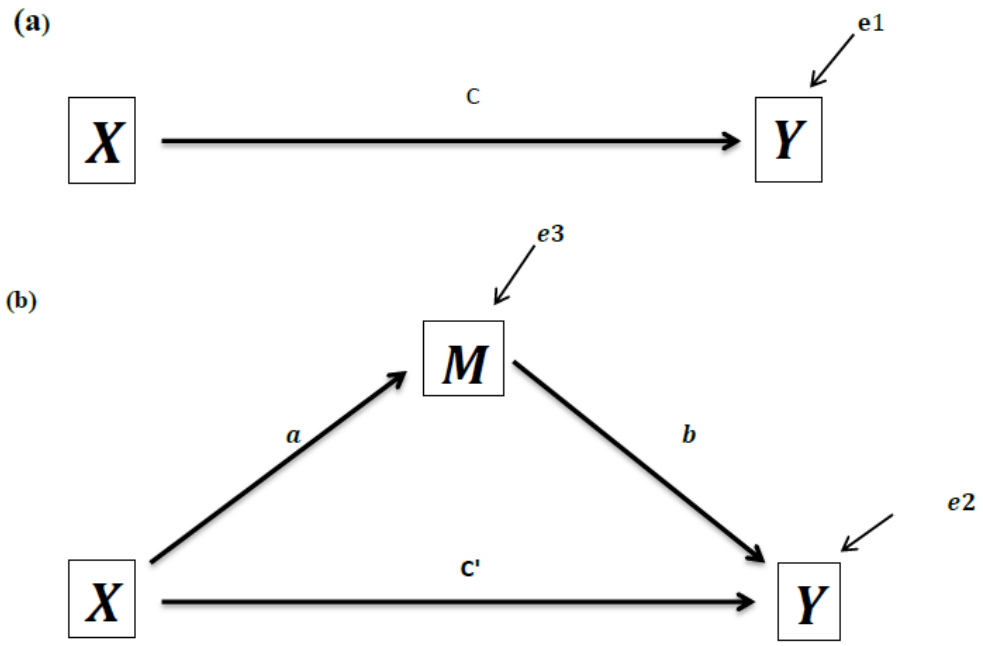

Figure (2) **(a)** - The direct effects of the X variable on the Y variable .

(b) - Indirect effects of the variable X on the variable Y by the mediation variable M .

The symbols (a, b, c) are considered of particular importance, where (a) denotes the effect of the treatment variable (X) on the mediation variable (M), (b) denotes the effect of the mediation variable (M) on the result variable (Y) , (c) denotes the effect of the treatment variable (X) on the outcome variable (Y) (David Mackinnon, 2012) (D.P.Mackinnon, 2007) (D.P.Mackinnon & Pirlott, 2015).

4. Multiple Mediator Model

After we describe the individual mediation model, we turn in this part to a broader and more complex mediation model than the previous model, which is a multi-mediation model with a description of its equations with its graphical forms.

4.1 Regression equations used to evaluate mediation

In order to provide a mediation approach to the multi-median model, the following regression equation is required:

$$\boldsymbol{Y = r_1 + cX + e_1} \quad \text{..................} \ \textbf{(4)}$$

$$\boldsymbol{Y = r_2 + c'X + b_1M_1 + \dots \dots + b_nM_n + e_2} \ \textbf{. (5)}$$

$$\boldsymbol{M_1 = r_3 + a_1X + e_3} \quad \text{..................} \ \textbf{(6)}$$

$$\boldsymbol{M_n = r_n + a_nX + e_n} \quad \text{..................} \ \textbf{(7)}$$

whereas

$\boldsymbol{Y}$: represents the result variable.

$\boldsymbol{X}$: represents a treatment variable.

$\boldsymbol{M_1}, \dots, \boldsymbol{M_n}$: Causal Mediation Variables.

$\boldsymbol{C}$: The parameter that binds the treatment variable with the outcome variable represents the absence of mediation variables.

$\boldsymbol{C'}$: The parameter that binds the treatment variable to the outcome variable is the effect of the mediation variables.

$\boldsymbol{a_1}$: The parameter that associates the transaction variable with the first broker variable.

$\boldsymbol{a_n}$: The parameter that links the transaction variable to other mediation variables.

$\boldsymbol{b_1}$: The parameter that binds the first mediation variable to the result variable

$\boldsymbol{b_n}$: The parameter that links the other mediation variables to the result variable.

$\boldsymbol{r_1}, \boldsymbol{r_2}, \boldsymbol{r_3}, \dots, \boldsymbol{r_n}$: are objections in each equation

$\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}, \dots, \boldsymbol{e_n}$: are errors.

$\boldsymbol{Equation}\ (\boldsymbol{4})$ is the estimation of the direct effect in the absence of mediation.

$\boldsymbol{Equations}\ (\boldsymbol{5})(\boldsymbol{6})(\boldsymbol{7})$ is the estimation of the indirect effect by the presence of the mediation model (David Mackinnon, 2012).

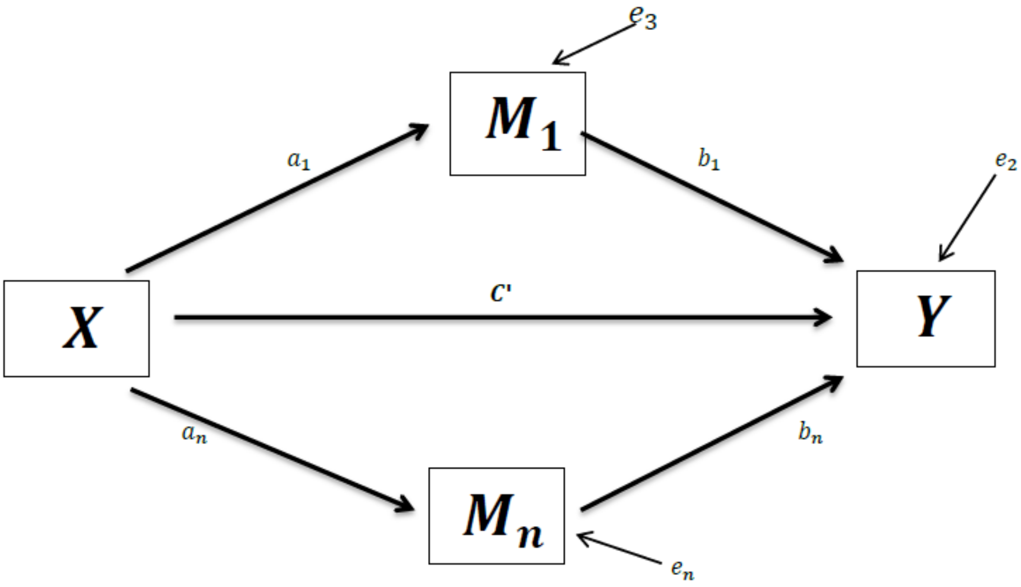

Figure (3) : Represents a diagram of the multi-mediation model using two variables, so the independent variable (X) is related to two arguments (M_1) and (M_n) which are related to the dependent variable (Y). Note that there are several other possible mediation models, for example X to $M_1$to $M_n$to y.

5. The Total Effect

The multiple mediation model in Figure (3) consists of two mediation variables whose total effect is the direct effect represented by parameter (C') and the indirect effect of the mediation variables $(M_1, \dots, M_n)$ (Hayes, 2009).

$$\boldsymbol{Total\ effect} = \boldsymbol{C'} + \boldsymbol{a_1 b_1} + \cdots + \boldsymbol{a_n b_n} \quad .. \quad (8)$$

6. Effect of mediation

We divide the total effect into direct effects, which are represented by the parameter (C') and others that are indirect and that are represented by the effects of the mediator $(a_1 b_1, \dots, a_n b_n)$ in the analysis of mediation, so it should be:

$$\boldsymbol{a_1 b_1} + \cdots + \boldsymbol{a_n b_n} = \boldsymbol{C} + \boldsymbol{C'} \qquad \dots \quad (9)$$

In the case of an unequal reason, this is due to the difference in the sample sizes between the equations Therefore, the parameters of the above models are estimated using different methods, such as (OLS) (Preacher & Hayes, 2008).

7. Testing the importance and confidence intervals of mediation effects

One of the most important methods used to test the significance of the mediation effect (ab) is to estimate the standard error and compare the resulting Z degrees with the critical value of the standard normal distribution (D.P.Mackinnon et al. ,2004).

The standard error and estimated causal mediation effect can also be used to establish confidence intervals for mediation effect. It is well known that confidence intervals use standard error in estimation, which is why we believe that the confidence intervals used may provide a number of effect values rather than a single value. Therefore, confidence intervals are tools with common uses in research because they ask the researcher to look for the value of the effect as well as its statistical significance (Harlow,nd) and to test the importance of the indirect effect (mediator), So we need to find the standard error of the sample by the median (ab) The most commonly used tests to estimate the standard errors of indirect effect are:

7.1 Strategy for causal action to establish mediation for the multi-median model

The multi-mediation form contains a set of steps:

1- The treatment variable (X) affects the outcome variable (Y) through (C) as in equation (4).

2- The treatment variable (X) affects the mediation of $(M_1, \dots, M_n)$ by (a_1) and (a_n) as in equations (6)(7).

3- Mediation variables affect the outcome variables (Y) after controlling the treatment variable (X) factors $(b_1, \dots, b_n)$ as in equation (5).

4- The treatment variable must have an unimportant effect (direct effect C') on the outcome variable in order to achieve full mediation as found in equation (5), but if there is an effect of the treatment variable on the outcome variable (C') here is a partial mediation (Wen , 2013).

7.2 Product of parameter coefficients testing

The standard error for the internal effect of a multiple mediation model is as follows (David Mackinnon , 2012).

$$S_{a1\ b1} = \sqrt{S_{a1}^2\ b_1^2 + S_{b1}^2\ a_1^2} \qquad (10)$$

Another variation of the standard error of the multiple mediation model can also be used:

$$S_{a1\ b1} + \cdots + S_{an\ bn} = \sqrt{S_{a1}^2 S_{b1}^2 + S_{b1}^2 a_1^2 + \cdots + S_{an}^2 b_n^2 + S_{bn}^2 a_n^2 + 2a_1 \ldots a_n\ \ S_{b1\ldots\ bn}} \qquad (11)$$

Rewrite the equation above as follows:

$$S_{a1\ b1} + \cdots + S_{an\ bn} = \sqrt{S_{a1\ b1}^2 + \cdots + S_{an\ bn}^2 + 2a_1, \ldots, a_n\ \ S_{b1,\ldots,bn}} \qquad (12)$$

$(2b_1, \ldots, b_n\ \ S_{a1,\ldots,an})$ Must be added to the above equations when there is a non-zero difference between $(a_1, \ldots, a_n)$

$(S_{b1,\ldots,bn})$: Represents the common difference

When the overall effect of the medium is present, the standard error is as follows:

8. Least Squares Estimation

The $\hat{\beta}$ in LSE is defined by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\boldsymbol{argmin}}\ \ \boldsymbol{RSS}(\boldsymbol{\beta})$$

Where RSS is the residuals sum of squares , i.e.,

$$RSS(\beta) = \sum_{i=1}^{n} (yi - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}))^2 \qquad (13)$$

So, the LSE $(\hat{B})$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\ \boldsymbol{X^T Y}$$

8.1 LSE Properties

The LSE properties are as follows:

1. $\hat{B}$ is unbiased, $E[\hat{\beta}] = \beta$.
2. The variance of $\hat{\beta}$ is $Var[\hat{\beta}] = \sigma^2 (X^T X)^{-1}$.
3. $\sum_{i=1}^{n} (\ y_i - \hat{y}_i) = 0$.
4. $\sum_{i=1}^{n} x_i e_i = 0,$ which means no correlation between x_i and e_i.
5. $\sum_{i=1}^{n} \hat{y}_i e_i = 0,$ which means no correlation between $\hat{y}_i$ and e_i.

(Montgomery et al. ,2013).

9.Lasso Regression

Lasso (least absolute shrinkage and selection operator) deals with several predictors, $k \gg n$, and with an unconditional model matrix X. However, it differs from RR, where lasso makes many parameters estimates equal to zero, which makes the interpretation of the statistical model more acceptable. Whereas, the Lasso was first introduced by scientist (Robert Tabsherani , 1996) and has been defined as:

Let:

$$\boldsymbol{L(\beta,\lambda) = RSS(\beta) + \lambda \textstyle\sum_{j=1}^{k} | \beta_j |,} \qquad (14)$$

$RSS\ (\beta)$ is defined

$$RSS(\beta) = \textstyle\sum_{i=1}^{n}(yi - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}))^2 \qquad (15)$$

Lasso estimator, referred to as $\hat{\beta}^L(\lambda)$, decreases (14) by β for a given λ. The setting parameter λ determines whether $\hat{\beta}^L(\lambda)$ is low or not (some parameter estimates are set to absolutely zero). When λ is large, $||\hat{\beta}^L(\lambda)||_1$ becomes smaller, resulting in a sparser solution. For $\lambda > 0$, lasso's unique estimates may not be , (Robert Tibshirani ,2012).

We also standardize X predictors if there are different units. The term intersection can be retrieved after standardization and the Lasso coefficient estimates can be obtained from

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{k} \bar{x}_j \hat{\beta}_j^L(\lambda),$$

Where that $\bar{y}$ and $\bar{x}_j$ for $j = 1, \ldots, k,$ are original means and $\hat{\beta}_j^L(\lambda)$ is the estimation of the lasso coefficient for $j = 1, \ldots, k$.

In the representation of the vector and vector norm, the lasso estimator $\hat{\beta}^L(\lambda)$ can be written as an appropriate solution to this following problem:

$$\min_{\beta \in R^k} L(\beta,\lambda) = \min_{\beta \in R^k} ||Y - X\beta||_2^2 + \lambda\, ||\beta||_1 \quad for\ som \quad \lambda \geq 0, \quad (16)$$

Where that $||\beta||_1 = \sum_{j=1}^{k} | \beta_j |$ and matrix X does not include a column after standardization.

In fact, the problem (15) is equivalent.

$$\min_{\beta \in R^k} \quad ||Y - X\beta||^2$$

$$subject\ to \quad ||\beta||_1 \leq r, \qquad (17)$$

10. Estimation Methods

10. 1 The Method Of Least Squares

This topic will include estimating the expected average of the response variable using the available information from the predictors variables $(X_1, X_2, \ldots, X_{19})$ through the use of the three regression models, namely:

First Model: Estimating the linear relationship of the regression model (Y/ X)

Here the parameters of the predicted variables (X) will be estimated and the expected mean of the response variable (Y) will be estimated through regression model (4).

Where it will be written as follows: $yi = r_1 + \sum_{i=1}^{19} C_i X_i + e_1 \qquad (18)$

where the vector (C) represents the total or direct effect of the predictors variables (X) on the response variable (Y) and the hypothesis test $H_0{:}\, \underline{C} = 0$, and the table below (1) shows a summary of statistics for the estimated model.

Table (1) shows a summary of the estimated model statistics

Variable	Estimate	S.E.	*t*	P-Value
X_1	1.875	2.363e-01	7.936	7.39e-09 ***
X_2	-1.562	1.864e-01	-8.379	2.37e-09 ***
X_3	2.072e-01	1.289e-01	1.607	0.11858
X_4	-1.785e-01	1.394e-01	-1.281	0.21008
X_5	-9.483e-02	1.673e-01	-0.567	0.57502
X_6	1.013e-01	1.302e-01	0.778	0.44273
X_7	-1.974e-01	1.275e-01	-1.548	0.13200
X_8	2.154e-01	1.414e-01	1.523	0.13814
X_9	1.841e-01	1.128e-01	1.632	0.11306
X_{10}	2.941e-01	1.017e-01	2.893	0.00705 **
X_{11}	-1.094e-01	1.020e-01	-1.073	0.29198
X_{12}	2.200e-01	1.058e-01	2.078	0.04634 *
X_{13}	1.253e-01	1.044e-01	1.200	0.23967
X_{14}	-3.432e-02	1.259e-01	-0.273	0.78699
X_{15}	8.100e-03	1.150e-01	0.070	0.94432
X_{16}	-8.956e-02	1.187e-01	-0.754	0.45659
X_{17}	6.112e-02	1.499e-01	0.408	0.68645
X_{18}	-5.024e-02	9.762e-02	-0.515	0.61055
X_{19}	-1.551e-01	1.233e-01	-1.258	0.21816

Table (1) shows the estimated values of the model parameters (4) in addition to the standard errors and the test (t) and the values of (P-Value) where we notice that the variables (Age of the woman) (Woman's age at marriage) (Number of deceased children) (Woman with thyroid disease) have a significant effect level of significance ($\alpha = 0.05$) and this is clear from the values of the column (P-Value) as well as we note from this The table shows that the rest of the variables do not have a significant effect in the study model, and this is evident from the values of (P-Value) in comparison with the level of significance ($\alpha =$ 0.05).

To test the significance of the estimated model (4) by using the least squares method, we will test the hypothesis

$$H_0: Ci = 0 \quad \text{and} \quad H_1: Ci \neq 0 \qquad i = 1, \dots, 19$$

Where the table below (2) shows the analysis of variance table for the estimated model.

Table (2) analysis of variance

S.O.V	D.F	Sum Sq	Mean Sq	F value	Pr(>F)	R^2	Adjusted R^2
Model	19	76.424	4.0223	11.999	2.916e-09***	0.8837	0.8101
Residuals	30	10.056	0.3352				
Total	49	86.480					

From the table (2) We note from the value of (Pr(>F)) the significance of the model estimated using the method of least squares, as its value was close to zero and it is less than the level of significance ($\alpha = 0.05$) and thus rejected H_0.

We also note that the value of (R^2) is equal to (%88), which is interpreted as the ratio of the total change occurring in the response variable (number of children born) as a result of its effect on the values of the predictors variables (X), where the value of (R^2) can be considered as an indicator of the suitability of the estimated model or an indicator of the strength of the linear relationship between the number of children born and predictors variables.

We also note the value of ($Adjusted\ R^2$), which is equal to (%81), which is another measure to judge the extent to which the estimated model matches the data, assuming a change in the number of predictors variables. We can use ($Adjusted\ R^2$) to compare models that have different numbers of predictors variables.

Second Model: Estimating the Regression Model $(Y/X, M)$

In this model (5) the parameters of the predictors variables (X) will be estimated $(C' = \text{partial effect})$ and the expected estimate of the response variable (Y) with another direct estimate called (the mediation effect) by the parameter (b) which links the mediation variable to the outcome variable in order to test the partial effect (C') , We will test the following null hypothesis : $H_0\colon \underline{C'} = 0\ and\ H_0\colon b = 0$, and the table (3) below shows a summary of statistics for the estimated model.

Table (3) shows a summary of the estimated model statistics

Variable	Estimate	S.E.	*t*	P-Value
X_1	1.898	2.343e-01	8.102	1 6.20e-09 ***
X_2	-1.667	2.015e-01	-8.275	4.01e-09 ***
X_3	2.279e-01	1.285e-01	1.773	0.0867
X_4	-1.859e-01	1.380e-01	-1.347	0.1883
X_5	-9.658e-02	1.654e-01	-0.584	0.5638
X_6	1.236e-01	1.299e-01	0.951	0.3493
X_7	-1.958e-01	1.261e-01	-1.553	0.1312
X_8	2.511e-01	1.425e-01	1.762	0.0886
X_9	2.299e-01	1.170e-01	1.965	0.0590
X_{10}	1.603e-01	1.440e-01	1.114	0.2746
X_{11}	-1.249e-01	1.016e-01	-1.230	0.2286
X_{12}	2.175e-01	1.047e-01	2.077	0.0467 *
X_{13}	1.410e-01	1.040e-01	1.356	0.1855
X_{14}	-7.211e-02	1.278e-01	-0.564	0.5770
X_{15}	3.036e-02	1.150e-01	0.264	0.7937
X_{16}	-1.037e-01	1.179e-01	-0.879	0.3865
X_{17}	4.531e-02	1.488e-01	0.305	0.7629
X_{18}	-1.043e-01	1.051e-01	-0.992	0.3293
X_{19}	-3.048e-01	1.678e-01	-1.816	0.0797
M	2.439e-01	1.880e-01	1.298	0.2046

Table (3) shows the estimated values of the model parameters (5) in addition to the standard errors and the test (t) and the values of (P-Value) where we notice that the variables (Age of the woman) (Woman's age at marriage) (Woman with thyroid disease) have a significant effect level of significance ($\alpha = 0.05$) and this is clear from the values of the column (P-Value) as well as we note from this The table shows that the rest of the variables do not have a significant effect in the study model, and this is evident from the values of (P-Value) in comparison with the level of significance ($\alpha = 0.05$).

To test the significance of the estimated model (5) by using the least squares method, we will test the hypothesis

$H_0{:}\,Ci = 0$ and $H_1{:}\,Ci \neq 0$ $\quad i = 1, \dots, 19$

$H_0{:}\,b = 0 \quad and \quad H_1{:}\,b \neq 0$

Where the table below (4) shows the analysis of variance table for the estimated model.

Table (4) analysis of variance

S.O.V	D.F	Sum Sq	Mean Sq	F value	Pr(>F)	R^2	Adjusted R^2
Model	20	76.976	3.8488	11.744	4.865e-09 ***	0.8901	0.8143
Residuals	29	9.504	0.3277				
Total	49	86.480					

From the table (4) We note from the value of (Pr(>F)) the significance of the model estimated using the method of least squares, as its value was close to zero and it is less than the level of significance ($\alpha = 0.05$) and thus rejected H_0.

We also note that the value of (R^2) is equal to (%89), which is interpreted as the ratio of the total change occurring in the response variable (number of children born) as a result of its effect on the values of the predictors variables (X) by the parameter (b) which links the mediation variable to the outcome variable in order to test the partial effect (C'), where the value of (R^2) can be considered as an indicator of the suitability of the estimated model or an indicator of the strength of the linear relationship between the number of children born and predictors variables.

We also note the value of ($Adjusted\ R^2$), which is equal to (%81), which is another measure to judge the extent to which the estimated model matches the data, assuming a change in the number of predictors variables. We can use ($Adjusted\ R^2$) to compare models that have different numbers of predictors variables.

<u>Third Model:</u> Estimating the Regression Model (M/X)

In this model (6), the effect of the parameters of the predictors variables (X) on the mediation variable (M) will be estimated in order to test the direct effect (a).

Null hypothesis : $H_0: a = 0$

and the table (5) below shows a summary of statistics for the estimated model.

Table (5) shows a summary of the estimated model statistics

Variable	Estimate	S.E.	*t*	P-Value
X_1	-6.357e-02	1.527e-01	-0.416	0.6801
X_2	2.917e-01	1.204e-01	2.422	0.0217 *
X_3	-5.705e-02	8.331e-02	-0.685	0.4987
X_4	2.022e-02	9.007e-02	0.224	0.8239
X_5	4.825e-03	1.081e-01	0.045	0.9647
X_6	-6.144e-02	8.413e-02	-0.730	0.4709
X_7	-4.346e-03	8.238e-02	-0.053	0.9583
X_8	-9.840e-02	9.137e-02	-1.077	0.2901
X_9	-1.262e-01	7.288e-02	-1.732	0.0936
X_{10}	3.689e-01	6.569e-02	5.616	4.10e-06 ***
X_{11}	4.277e-02	6.591e-02	0.649	0.5214
X_{12}	6.950e-03	6.840e-02	0.102	0.9197
X_{13}	-4.335e-02	6.749e-02	-0.642	0.5256
X_{14}	1.042e-01	8.133e-02	1.282	0.2098
X_{15}	-6.140e-02	7.432e-02	-0.826	0.4152
X_{16}	3.894e-02	7.673e-02	0.507	0.6155
X_{17}	4.362e-02	9.689e-02	0.450	0.6558
X_{18}	1.492e-01	6.308e-02	2.365	0.0247 *
X_{19}	4.128e-01	7.966e-02	5.182	1.39e-05 ***

Table (5) shows the estimated values of the model parameters (6) in addition to the standard errors and the test (t) and the values of (P-Value) where we notice that the variables (Woman's age at marriage) (Number of deceased children) (Match the blood) (Gestational diabetes) have a significant effect level of significance ($\alpha = 0.05$) and this is clear from the values of the column (P-Value) as well as we note from this The table shows that the rest of the variables do not have a significant effect in the study model, and this is evident from the values of (P-Value) in comparison with the level of significance ($\alpha = 0.05$).

To test the significance of the estimated model (6) by using the least squares method, we will test the hypothesis

$H_0: a = 0$ and $H_1: a \neq 0$ $\quad i = 1, \dots, 19$

Where the table below (6) shows the analysis of variance table for the estimated model.

Table (6) analysis of variance

S.O.V	D.F	Sum Sq	Mean Sq	F value	Pr(>F)	R^2	Adjusted R^2
Model	19	17.9813	0.94639	6.762	2.218e-06 ***	0.8107	0.6908
Residuals	30	4.1987	0.13996				
Total	49	22.1800					

From the table (6) We note from the value of (Pr(>F)) the significance of the model estimated using the method of least squares, as its value was close to zero and it is less than the level of significance ($\alpha = 0.05$) and thus rejected H_0.

We also note that the value of (R^2) is equal to (%81), which is interpreted the effect of the parameters of the predictors variables (X) on the mediation variable (M) which will be estimated in order to test the direct effect (a) , where the value of (R^2) can be considered as an indicator of the suitability of the estimated model or an indicator of the strength of the linear relationship between the number of children born and predictors variables.

We also note the value of ($Adjusted\ R^2$), which is equal to (%69), which is another measure to judge the extent to which the estimated model matches the data, assuming a change in the number of predictors variables. We can use ($Adjusted\ R^2$) to compare models that have different numbers of predictors variables.

10.2 Lasso regression method

In this study we use the lasso method and the similarity in its method of operation to the Ridge method as it reduces the squares of the mean residuals by the difference of the constraint placed on the penalty function as it is a method that deals with the existence of a large number of predictors variables compared to the sample size as well as deals with data that suffer from the problem of linear multiplicity. . What distinguishes the lasso method is to make some parameter estimates equal to zero, which adds explanatory ability to the model and creates a model with high predictive accuracy.

Below is an explanation of the process for estimating the proposed models using the lasso method.

First Model: Estimating the linear relationship of the regression model (Y/ X)

Here the parameters of the predicted variables (X) will be estimated and the expected mean of the response variable (Y) will be estimated through a regression model (18), As the vector (C) represents the total or direct effect of the predicted variables (X) on the response variable (Y) and the hypothesis test $H_0: \underline{C} = 0$, and the table (7) below shows a summary of statistics for the estimated model.

Table (7) shows a summary of the estimated model statistics

Variable	Estimate	S.E.	*t*	P-Value
X_1	1.317976	0.2778329	4.743771	2.097765e-06***
X_2	-9.263589e-01	0.1928095	-4.804530	1.551157e-06***
X_3	0.000000	0.1598062	0.000000	1.000000
X_4	0.000000	0.1754780	0.000000	1.000000
X_5	0.000000	0.2150542	0.000000	1.000000
X_6	0.000000	0.1666426	0.000000	1.000000
X_7	-1.427467e-01	0.1642538	-8.690620e-01	3.848132e-01
X_8	6.442416e-02	0.1794451	3.590188e-01	7.195810e-01
X_9	0.000000	0.1396024	0.000000	1.000000
X_{10}	3.388157e-01	0.1309480	2.587406	9.670144e-03**
X_{11}	0.000000	0.1294267	0.000000	1.000000
X_{12}	8.262912e-02	0.1331504	6.205696e-01	5.348828e-01
X_{13}	0.000000	0.1319095	0.000000	1.000000
X_{14}	0.000000	0.1624358	0.000000	1.000000e
X_{15}	0.000000	0.1485947	0.000000	1.000000
X_{16}	8.485125e-03	0.1517916	5.589984e-02	9.554216e-01
X_{17}	0.000000	0.1932266	0.000000	1.000000
X_{18}	0.000000	0.1256038	0.000000	1.000000
X_{19}	0.000000	0.1553468	0.000000	1.000000

Table (7) shows the estimated values of the model parameters (18) in addition to the standard errors and the test (t) and the values of (P-Value) where we notice that the variables (Age of the woman) (Woman's age at marriage) (Number of deceased children) have a significant effect level of significance ($\alpha = 0.05$) and this is clear from the values of the column (P-Value) as well as we note from this The table shows that the rest of the variables do not have a significant effect in the study model, and this is evident from the values of (P-Value) in comparison with the level of significance ($\alpha = 0.05$).

To test the significance of the estimated model (18) by using the lasso method, we will test the hypothesis

$H_0: Ci = 0$

$H_1: Ci \neq 0 \qquad i = 1, \dots, 19$

Where the table below (8) shows the analysis of variance table for the estimated model.

Table (8) analysis of variance

S.O.V	D.F	Sum Sq	Mean Sq	F value	Pr(>F)	R^2	Adjusted R^2
Model	19	69.47057	3.656346	6.4488	3.667236e-06***	0.8033	0.6787
Residuals	30	17.00943	0.566980				
Total	49	86.48					

From the table (8) We note from the value of (Pr(>F)) the significance of the model estimated using the method of lasso, as its value was close to zero and it is less than the level of significance ($\alpha = 0.05$) and thus rejected H_0.

We also note that the value of (R^2) is equal to (%80), which is interpreted as the ratio of the total change occurring in the response variable (number of children born) as a result of its effect on the values of the predictors variables (X), where the value of (R^2) can be considered as an indicator of the suitability of the estimated model or an indicator of the strength of the linear relationship between the number of children born and predictors variables.

We also note the value of ($Adjusted\ R^2$), which is equal to (%67), which is another measure to judge the extent to which the estimated model matches the data, assuming a change in the number of predictors variables. We can use ($Adjusted\ R^2$) to compare models that have different numbers of predictors variables.

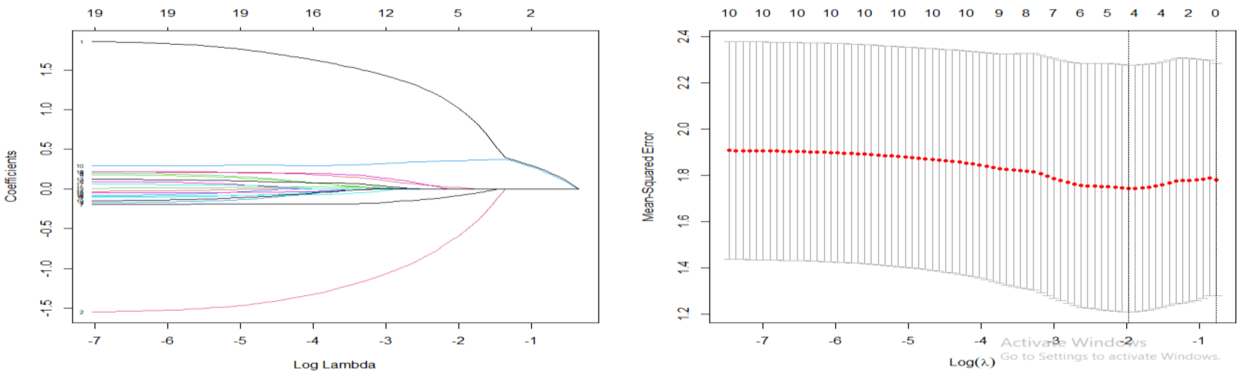

Figure (4) a schematic diagram of the first model using the lasso method

Second Model: Estimating the Regression Model $(Y/X, M)$

In this model (5), the parameters of the predicted variables (X) will be estimated and the expected estimate of the response variable (Y) with another direct estimate called (the mediation effect) by the parameter (b) which links the mediation variable to the outcome variable in order to test the partial effect (C').

Null hypothesis : $H_0{:}\,\underline{C}' = 0 \ and \ H_0{:}\,b = 0$, and the table (9) below shows a summary of statistics for the estimated model.

Table (9) shows a summary of the estimated model statistics

Variable	Estimate	S.E.	*t*	P-Value
X_1	1.406385	0.2666400	5.274472	1.331390e-07***
X_2	-1.039980	0.2000968	-5.197385	2.021112e-07***
X_3	0.000000	0.1503131	0.000000	1.000000
X_4	0.000000	0.1659017	0.000000	1.000000
X_5	-8.896786e-03	0.2050008	-4.339879e-02	9.653836e-01
X_6	0.000000	0.1590460	0.000000	1.000000
X_7	-1.637055e-01	0.1568829	-1.043489	2.967220e-01
X_8	9.500063e-02	0.1734834	5.476066e-01	5.839620e-01
X_9	1.056542e-02	0.1357568	7.782612e-02	9.379664e-01
X_{10}	3.255793e-01	0.1748693	1.861844	6.262507e-02
X_{11}	0.000000	0.1228906	0.000000	1.000000
X_{12}	1.254565e-01	0.1285116	9.762271e-01	3.289519e-01
X_{13}	2.656439e-02	0.1265667	2.098846e-01	8.337578e-01
X_{14}	0.000000	0.1582980	0.000000	1.000000
X_{15}	1.654417e-03	0.1431353	1.155841e-02	9.907779e-01
X_{16}	8.501076e-03	0.1443902	5.887572e-02	9.530511e-01
X_{17}	0.000000	0.1850213	0.000000	1.000000
X_{18}	0.000000	0.1285437	0.000000	1.000000
X_{19}	0.000000	0.1956980	0.000000	1.000000
M	0.000000	0.2266511	0.000000	1.000000

Table (9) shows the estimated values of the model parameters (5) in addition to the standard errors and the test (t) and the values of (P-Value) where we notice that the variables (Age of the woman) (Woman's age at marriage) have a significant effect level of significance ($\alpha = 0.05$) and this is clear from the values of the column (P-Value) as well as we note from this The table shows that the rest of the variables do not have a significant effect in the study model, and this is evident from the values of (P-Value) in comparison with the level of significance ($\alpha = 0.05$).

To test the significance of the estimated model (5) by using the lasso method, we will test the hypothesis.

$H_0: Ci = 0$ and $H_1: Ci \neq 0$ $i = 1, \dots, 19$

$H_0: b = 0$ *and* $H_1: b \neq 0$

Where the table below (10) shows the analysis of variance table for the estimated model.

Table (10) analysis of variance

S.O.V	D.F	Sum Sq	Mean Sq	F value	Pr(>F)	R^2	Adjusted R^2
Model	20	71.52987	3.576493	6.937617	1.887404e-06***	0.8271	0.7079
Residuals	29	14.95013	0.515521				
Total	49	86.48					

From the table (10) We note from the value of (Pr(>F)) the significance of the model estimated using the method of lasso, as its value was close to zero and it is less than the level of significance ($\alpha = 0.05$) and thus rejected H_0.

We also note that the value of (R^2) is equal to (%82), which is interpreted as the ratio of the total change occurring in the response variable (number of children born) as a result of its effect on the values of the predictors variables (X) by the parameter (b) which links the mediation variable to the outcome variable in order to test the partial effect (C'), where the value of (R^2) can be considered as an indicator of the suitability of the estimated model or an indicator of the strength of the linear relationship between the number of children born and predictors variables.

We also note the value of ($Adjusted\ R^2$), which is equal to (%70), which is another measure to judge the extent to which the estimated model matches the data, assuming a change in the number of predictors variables. We can use ($Adjusted\ R^2$) to compare models that have different numbers of predictors variables.

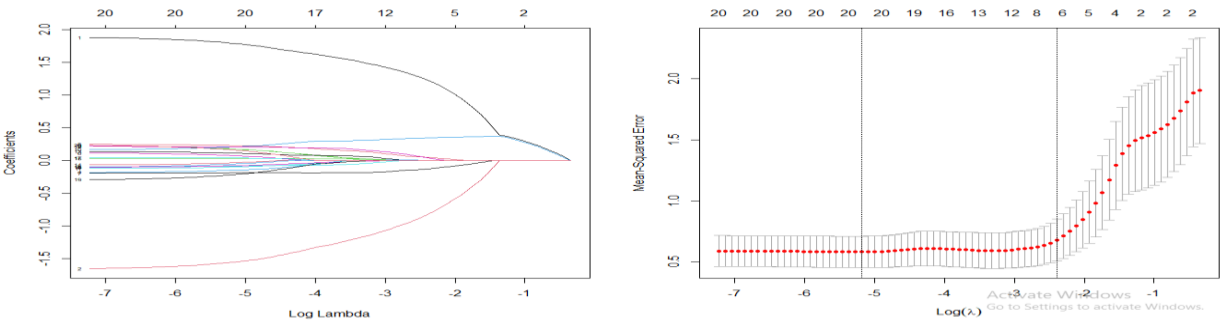

Figure (5) a schematic diagram of the second model using the lasso method

Third Model: Estimating the Regression Model (M/X)

In this model (6), the effect of the parameters of the predicted variables (X) on the mediation variable (M) will be estimated in order to test the direct effect (a), null hypothesis: $H_0: a = 0$ and the table below shows a summary of statistics for the estimated model.

Table (11) shows a summary of the estimated model statistics

Variable	Estimate	S.E.	*t*	P-Value
X_1	0.00000000	0.17174097	0.0000000	1.000000
X_2	0.17592840	0.13182218	1.3345888	1.820110e-01
X_3	0.00000000	0.09273353	0.0000000	1.000000
X_4	-0.01027589	0.10152239	-0.1012180	9.193774e-01
X_5	0.00000000	0.12231214	0.0000000	1.000000
X_6	-0.03211033	0.09480451	-0.3387004	7.348354e-01
X_7	0.00000000	0.09321728	0.0000000	1.000000
X_8	0.00000000	0.09914844	0.0000000	1.000000
X_9	-0.02889935	0.07720798	-0.3743053	7.081772e-01
X_{10}	0.28092102	0.06955552	4.0388025	5.372478e-05***
X_{11}	0.00000000	0.07349016	0.0000000	1.000000
X_{12}	0.00000000	0.07736865	0.0000000	1.000000
X_{13}	0.00000000	0.07527102	0.0000000	1.000000
X_{14}	0.01054264	0.08769974	0.1202128	9.043145e-01
X_{15}	-0.05507949	0.08408413	-0.6550521	5.124342e-01
X_{16}	0.00000000	0.08604772	0.0000000	1.000000
X_{17}	0.00000000	0.10886703	0.0000000	1.000000
X_{18}	0.06287865	0.06658401	0.9443505	3.449906e-01
X_{19}	0.34332972	0.08773636	3.9131976	9.108195e-05***

Table (11) shows the estimated values of the model parameters (6) in addition to the standard errors and the test (t) and the values of (P-Value) where we notice that the variables (Number of deceased children) (Gestational diabetes) have a significant effect level of significance ($\alpha = 0.05$) and this is clear from the values of the column (P-Value) as well as we note from this The table shows that the rest of the variables do not have a significant effect in the study model, and this is evident from the values of (P-Value) in comparison with the level of significance ($\alpha = 0.05$).

To test the significance of the estimated model (6) by using the lasso method, we will test the hypothesis

$H_0: a = 0$ and $H_1: a \neq 0$, where the table below (12) shows the analysis of variance table for the estimated model.

Table (12) analysis of variance

S.O.V	D.F	Sum Sq	Mean Sq	F value	Pr(>F)	R^2	Adjusted R^2
Model	19	16.73503	0.880791	4.852868	6.2717e-05***	0.7545	0.5990
Residuals	30	5.444971	0.181499				
Total	49	22.18					

From the table (12) We note from the value of (Pr(>F)) the significance of the model estimated using the method of lasso, as its value was close to zero and it is less than the level of significance ($\alpha = 0.05$) and thus rejected H_0.

We also note that the value of (R^2) is equal to (%75), which is interpreted the effect of the parameters of the predictors variables (X) on the mediation variable (M) which will be estimated in order to test the direct effect (a) , where the value of (R^2) can be considered as an indicator of the suitability of the estimated model or an indicator of the strength of the linear relationship between the number of children born and predictors variables.

We also note the value of ($Adjusted\ R^2$), which is equal to (%59), which is another measure to judge the extent to which the estimated model matches the data, assuming a change in the number of predictors variables. We can use ($Adjusted\ R^2$) to compare models that have different numbers of predictors variables.

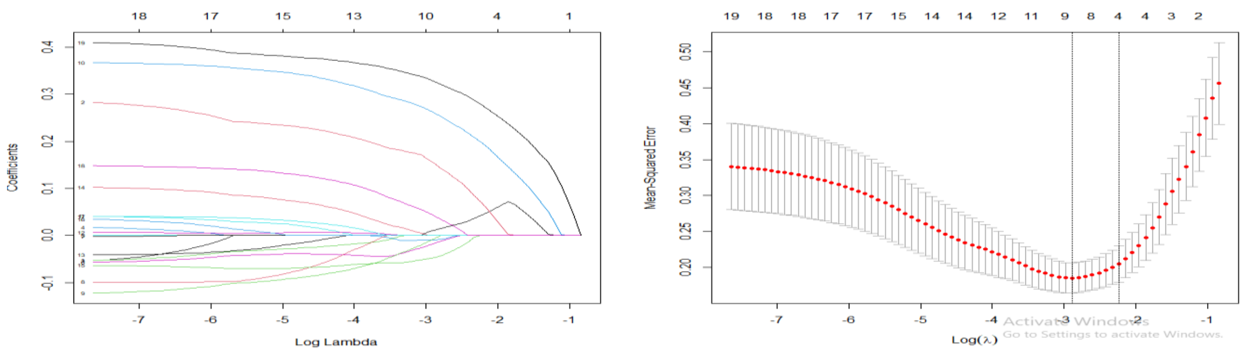

Figure (6) a schematic diagram of the third model using the lasso method

11. Conclusions

1- The effectiveness of the lasso regression method, which made the model more interpretable, in addition to choosing the easiest variables.

2- The emergence of the insignificance of some variables in the least squares method, and the lasso regression method despite its importance due to the presence of the problem of linear multiplicity and this is evident in the lack of significance of the mediation.

3- The emergence of some regression coefficients in the least squares method with an algebraic sign not identical to the reality of the studied phenomenon This is an indication of the inaccuracy of these methods in constructing a prediction model in the case of a multiplicity problem.

4- The preference for the Lasso regression method over the other methods, which is the least squares method.

5- In the lasso method, the less important independent variables in the model (Academic achievement of women, Academic achievement of the husband, Weight of woman, The length of the woman, Women smoking, The age of the husband, The profession of a husband, Previous use of contraceptives, The number of hours a woman sleeps a day, The duration of breastfeeding, Mother's food, Match the blood, Gestational diabetes, Psychological state) are reduced.

12. Recommendations

1- We recommend using the Lasso method when the matrix $(X'X)$ suffers from lack of an inverse value or when P> n. The Lasso method provides the best method for selecting the variables Variable Selection.

2- Not relying on the least squares method because they do not fully address the problem of linear multiplicity, especially in social studies, or when the number of independent variables is greater.

3- Study the Lasso Bayesian method, being one of the modern methods of estimating statistical models, which plays an effective role in determining the most important independent variables of the causal mediation model.

4- We recommend also studying the issue of causal mediation analysis in data in which more than one mediation variable is available.

Reference

[1] MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. Annu. Rev. Psychol., 58, 593–614.

[2] Grotta, A., & Bellocco, R. (2012). Causal mediation analysis on survival data: an application on the National March Cohort. PhD thesis, Univ. Milano-Bicocca, Milan.

[3] Wen, S. (2013). Estimation of multiple mediator model.

[4] MacKinnon, D. P., & Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis. Personality and Social Psychology Review, 19(1), 30–43.

[5] Otter, T., Pachali, M. J., Mayer, S., & Landwehr, J. (2018). Causal inference using mediation analysis or instrumental variables-full mediation in the absence of conditional independence. *Available at SSRN 3135313*.

[6] Tofighi, D., & Thoemmes, F. (2014). Single-level and multilevel mediation analysis. The Journal of Early Adolescence, 34(1), 93–119.

[7] MacKinnon, D. (2012). Introduction to statistical mediation analysis. Routledge.

[8] MacKinnon, D. P., Fritz, M. S., Williams, J., & Lockwood, C. M. (2007).

[9] Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. Communication Monographs, 76(4), 408–420.

[10] Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior Research Methods, 40(3), 879–891.

[11] MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. Multivariate Behavioral Research, 39(1), 99–128.

[12] Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51(6), 1173.

[13] Tibshirani, Robert, (1996) ," Regression Shrinkage and Selection via the Lasso" , J. R. Statist. Soc. B 58, No. 1, pp. 267-288.

[14] Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J. and Tibshirani, R.J. 2012. Strong rules for discarding predictors in LASSO-type problems. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74(2):245–266.

[15] Montgomery, D.C., Peck, E.A., and Vining, G.G. (2013). Introduction to Linear Regression Analysis. John Wiley & Sons, Inc., New Jersey, USA.