مجلة تنمية الرافدين

College of Adminstration and Economics Tanmyat AL-Rafidain 106(34) 2012
P.P [7-21]

A Suggested Method of Detecting Multicollinearity in Multiple Regression Models

Abdalla M. EL-HABIL (PhD)
Department of Applied Statistics
University of Gaza (Palastine)
abdalla20022002@yahoo.com

ABSTRACT

In literature, several methods suggested for the detection of multicollinearity in multiple regression models, and one of the multicollinearity problems solutions is to omit the explanatory variables in the model, which cause the multicollinearity. In this paper, we concentrated on the extra sum of squares method as a suggested method that can be used for detecting multicollinearity. The method of extra sum of squares is applied to real data on the annually surveys about smoking were conducted by the American Federal Trade Commission (FTC). In this data, we detected multicollinearity, then we solved this problem by using the ridge regression and we got the new estimates of the new model without omitting any of the explanatory variables.

Keywords: Multicollinearity , Ridge regression, Ordinary least square.

طريقة مقترحة لاكتشاف الازدواج الخطي في نماذج الانحدار المتعدد

الدكتور عبدالله الهبيل
قسم الإحصاء التطبيقي
جامعة غزة فلسطين

المستخلص

في الدراسات الأدبية، تم اقتراح العديد من الأساليب للكشف عن مشكلة الازدواج الخطي في نماذج الانحدار الخطي المتعدد، وكان من ضمن الحلول لهذه المشكلة هو حذف بعض المتغيرات المفسرة التي تتسبب في إيجاد مشكلة الازدواج الخطي من النموذج. في هذه الورقة البحثية، ركزنا على طريقة مجموع مربعات الانحدار كطريقة مقترحة يمكن استعمالها في اكتشاف الازدواج الخطي. ومن أجل ذلك قمنا بتحليل بيانات حقيقية صادرة عن المؤسسة الفيدرالية الأمريكية للتجارة والتي تعنى بعمل دراسات ميدانية سنوية عن التدخين، حيث تم اكتشاف الازدواج الخطي في النموذج المقدر، ثم بعد ذلك تم علاج تلك المشكلة باستخدام انحدار ردج دون حذف أي من المتغيرات المفسرة.

الكلمات المفتاحية: الارتباط الخطي المتعدد، طريقة انحدار الحرف، طريقة المربعات الصغرى الاعتيادية.

تأريخ استلام البحث ٢٠١٠/١١/٢٩ تأريخ قبول النشر ٢٠١١/٣/٣١

Introduction

Data with multicollinearity frequently arise and cause problems in many applications of linear regression such as in econometrics, oceanography, geophysics and other fields that rely on no experimental data. Multicollinearity is a natural flaw in the data set due to the uncontrollable operations of the data generating mechanism. In multiple linear regressions two or more of independent variables used in the model, the multicollinearity word has been used to represent a near exact relationship between two or more variables. Thomas P. Rayan (2009). In estimating the parameters in the regression model, it is often stated that multicollinearity can cause the signs of the parameter estimator to be wrong. The presence of multicollinearity will also mislead with the significance test telling us that some important variables are not needed in the model; multicollinearity causes a reduction of statistical power in the ability of statistical tests. Neter (1989) said that in the process of fitting regression model, when one independent variable is nearly combination of other independent variables, the combination would affect parameter estimates. Multicollinearity is the extreme problem for regression models, because it violates the assumptions of the model that is the explanatory variables should be independent. Belsley (1980) stated that, in case of existing of multicolliearity, it becomes difficult to infer the separate influence of such explanatory variables on the response variable. Weismann & Helge& Shalabh (2007) said that various diagnostic tools such as condition number, singular value decomposition method, Belsley condition indices, variance decomposition method, variance inflation factors, and Belsley's perturbation analysis etc., have been suggested in the literature for the detection of multicollinearity and identification of variables causing the linear relationships. Therefore, detecting multicollinearity is very important in regression analysis. The paper is organized as follows. Section 2 recalls the technical background of multicollinearity. Section 3 the extra sum of squares method Section 4 data analysis. Section 5 concludes.

1. Multicollinearity

Multicollinearity is defined as the existence of nearly linear dependency among the independent variables. The presence of serious multicollinearity would reduce the accuracy of the parameters estimate in a linear regression model and affect the independency of the independent variables of the regression model. Multicollinearity can cause serious problem in estimation and prediction, increasing the variance of least squares of the regression coefficients and tending to produce least squares estimates that are too large in absolute value. Theoretically, we have two types of multicollinearity; these types are partial multicollinearity and

perfect multicollinearity or full multicollinearity. In addition, multicollinearity has two cases: scalar case and matrix case.

Scalar case

Any population model looks like the below model:

$$Y = B_0 + B_1X_1 + B_2X_2 + ... + B_nX_n + e \qquad (2.1)$$

Where

y: The dependent variable

B_0: is the intercept

$B_1, B_2, ..., B_n$: are the slopes of coefficients for their respective explanatory variables.

$X_1, X_2, ..., X_n$: are the explanatory variables.

e : is the random error term.

Note: with absent of multicollinearity then Y is a linear function of the explanatory variables and a random error term.

If we suppose that, we have just only two explanatory variables X_1, X_2 and X_2 is a multiple of X_1 then $X_2 = dX_1$ for simplicity.

The regression would need to find the coefficient estimates b_1, b_2 that produce the best $\hat{Y}$.

Then $\hat{Y} = b_0 + b_1X_1 + b_2X_2$, we can substitute X_2 by dX_1 as:

$$\hat{Y} = b_0 + b_1X_1 + b_2(dX_1)$$

$$\hat{Y} = b_0 + X_1(b_1 + db_2) \qquad (2.2)$$

The above result is true also for infinite number of coefficient pairs; also these pairs produce the same value of $\hat{Y}$.

Any small change in b_1 from one possible value to another (δb_1) is matched by corresponding change in b_2.

By compensation, we get the below result:

$$\delta b_2 = -\frac{\delta b_1}{d}$$

by repeating all the coefficient pairs and keeping the linearity we can minimize sum of square errors $[Y - \hat{Y}]^2$.

Mathematically, the standard errors for coefficients S_j equals:

$$S_j = \sqrt{\frac{\sum_1^n (Y - \hat{Y})^2}{\sum (x_i - \bar{x}_i)^2 (1 - R_j^2)(n - k)}}$$ where

n: is the numbers of predictors.

$\bar{x}_i$: is the mean of x_i

R_j^2: is the square of the multiple correlation coefficients that result when the predictor variable X_j is regressed against the other entire predictor variable.

k: is the number of explanatory variables.

At full multicollinearity $R^2=1$, then $(1- R_j^2) = 0$

Therefore S_j is undefined, then there is no linear regression.

Matrix case

In a matrix case the explanatory variable X can be take the following form:

$$X = \begin{bmatrix} 1 & X_{12} & X_{13} & ..X_{1k} \\ 1 & X_{21} & X_{22} & ..X_{2k} \\ . & & & \\ & . & . & . \\ . & & & \\ 1 & X_{n2} & X_{n3} & X_{nk} \end{bmatrix}$$

Where:

n: number of observations

k: number of explanatory variables.

The first column is intercept or constant.

Matrix linear model is:

$$Y = B_0 + \sum_{i=1}^{j} X_i^T B_i + e \qquad (2.3)$$

$$\hat{Y} = b_0 + \sum_{i=1}^{j} X_i^T b_i$$

Where Y is a vector $n \times 1$.

The coefficient b can be calculated as:

$X^T X\, b = X^T Y$

$\therefore\ b = (X^T X)^{-1} X^T Y$

At multicollinearity the determinant of $(X^T X)$ is equal zero, therefore the inverse will not existing.

Detection of multicollinearity

Multicollinearity can be detected by examining one of two qualities: Variance Inflation Factor "VIF" and Tolerance.

We can detect the multicollinearity by examining a quality called Variance Inflation Factor (VIF).

$VIF_j = \frac{1}{1-R_j^2} = \text{diag}(\mathrm{X}^{-1}X)^{-1}$, $j = 1, ..., p$ where:

R_j^2: is the square of the multiple correlation coefficients that result when the predictor variable X_j is regressed against the other entire predictor variables.

p : is the number of predictor variables.

X: in matrix case

At multicollinearity R^2 would be closed to 1 then VIF_j would be large,

When VIF_j greater than 10 , the data have collinearity problems.

And we can detect the multicollinearity by examining a quality called Tolerance which equals $\frac{1}{VIF} = (1-R_j)^2$, and the tolerance in multicollinearity should be small.

Also, we can detect multicollinearity by using extra sum of squares method.

2. Extra Sum of squares method for detecting multicollinearity

We can define the extra sums of squares as the marginal increase in regression sum of squares when one or more independent variables are added to a regression model. In general, we use extra sums of squares to determine whether specific variables are making substantial contributions to our model. Extra sums of squares provides a means of formally testing whether one set of predictors is necessary given that another set is already in the model. In regression analysis, we can use hypothesis test to check the significance of the fitted model. Analysis of variance gives the information on regression sum of squares (SSR), residual sum of squares (SSE), total sum of squares (SST) and the F value for the hypothesis test. Regression sum of squares account for the variation in y that is explained by the variation of x_i. In regression analysis, the regression sum of squares will always increase while the residual sum of squares will decrease when a new independent variable is added to the model, because the total sum of squares unchanged**. Decomposition of SSR into extra sum of squares** In many applications such as stepwise regression, additional sums of squares are needed to measure the variation of y on some independent variables when a certain set of independent variables are already in the model. Here, we use $SSR(x_i \mid x_j, x_k)$ to represent the additional sum of squares which account for the variation in y when x_i is added in the model that already contains independent variables x_i and x_k**.** The $SSR(x_i \mid x_j, x_k)$ can be calculated as:

$$SSR(x_i \mid x_j, x_k) = SSR(x_i, x_j, x_k) - SSR(x_j, x_k) \qquad (3.1)$$

If we have a linear regression model is constructed based on p independent variables $x_1.x_2,\ldots,x_p$ and if multicollinearity among the

regressors $x_1, x_2, \ldots, x_p$ does not exist. However, at least one of the equality statements below does not hold if the regressors $x_1, x_2, \ldots, x_p$ are correlated.

$$SSR(xi|\, xj) = SSR(xi),\ i, j \in \{1, 2,\ldots, p\},\ i \neq j \qquad (3.2)$$

$$SSR(xi|\, xj, xk) = SSR(xi),\ i, j, k \in \{1, 2,\ldots, p\},\ i \neq j \neq k \qquad (3.3)$$

$$SSR(xi \,|\, xj, xk, xm) = SSR(xi),\ i, j, k, m \in \{1, 2,\ldots, p\},\ i \neq j \neq k \neq m \qquad (3.4)$$

$$SSR(xi|\, xj, xk, xm,\ldots) = SSR(xi),\ i, j, k, m,\ldots \in \{1, 2,\ldots, p\},\ i \neq j \neq \ldots \qquad (3.5)$$

Chin (2006).

Here, the independent variables are (p – 1), equation (3.5) represents the additional sum of squares accounted for the variation in y when x_i is added in a model that already contains p-1 independent variables.

Uses of extra sum of squares

One of the major uses of extra sum of squares is for conducting tests concerning regression coefficients without fitting both of the full and reduced models separately. For example if we have MLR model with two independent variables and we want to test whether or not $\beta_2 = 0$, here actually we don't need to fit the reduced model since the partial F test statistic can be calculated immediately from the relation below:

$$F^* = \frac{SSR(X_2 / X_1)}{1} \div \frac{SSE(X_1, X_2)}{n-3}$$

Also, we can use the extra sum of squares to measure the coefficient of partial determination between y and any independent variable in the MLR model, for example if we have a model with two independent variables, we have the following: $SSE(X_2)$ measures the variation in Y when X2 is included in the model, $SSE(X_1,X_2)$ measures the variation in Y when X_1 and X_2 are included in the model. The relative marginal reduction in the variation in Y associated with X_1 when X_2 is already in the model is:

$$\frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)}$$

This measure is the coefficient of partial determination between Y and

X_1 given that X_2 is in the model. We can denote this measure by $r^2_{y\,1.2}$

$$r^2_{y\,1.2} = \frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)} = 1 - \frac{SSE(X_1, X_2)}{SSE(X_2)}$$

We can express the coefficient of partial determination by extra sum of squares by:

$$r^2_{y\,1.2} = \frac{SSR(X_1 / X_2)}{SSE(X_2)}$$

3. Data analysis

In this section, we will conduct an application using real data and try to detect multicollinearity by using sum of squares method. The data is from the American Federal Trade Commission (FTC), which annually ranks varieties domestic cigarettes according to their tar, nicotine, and carbon monoxide contents. The U.S. surgeon general considers each of these three substances hazardous to a smoker's health. Past studies have shown that increases in the tar and nicotine contents of a cigarette are accompanied by an increase in the carbon monoxide emitted from the cigarette smoke. Table (4.1) lists tar, nicotine, and carbon monoxide contents (in milligrams) and weight (in grams) for a sample of 25 (filter) brands tested in recent years.

Table (4.1) FTC Cigarette Data

BRAND	Carbon Monoxide (y)	TAR (x_1)	Nicotine (x_2)	Weight (x_3)
Alpine	13.6	14.1	.86	0.9853
Benson &Hedges	16.6	16	1.06	1.0938
Bull Durham	23.5	29.8	2.03	1.1650
Camel Lights	10.2	8	0.67	0.9280
Carlton	5.4	4.1	0.40	0.9462
Chesterfield	15	15	1.04	0.8885
Golden Lights	9	8.8	0.76	1.0267
Kent	12.3	12.4	0.95	0.9225
Kool	16.3	16.6	1.12	0.9372
L&M	15.4	14.9	1.02	0.8858
Lark Lights	13.0	13.7	1.01	0.9643
Marlboro	14.4	15.1	0.90	0.9316
Merit	10	7.8	0.57	0.9705
Multifilter	10.2	11.4	0.78	1.1240
Newport Lights	9.5	9	0.74	0.8517
Now	1.5	1	0.13	0.7851
Old Gold	18.5	17	1.26	0.9186
Pall Mall Lights	12.6	12.8	1.08	1.0395
Raleigh	17.5	15.8	0.96	0.9573

BRAND	Carbon Monoxide (y)	TAR (x_1)	Nicotine (x_2)	Weight (x_3)
Salem Ultra	4.9	4.5	0.42	0.9106
Tareyton	15.9	14.5	1.01	1.0070
True	8.5	7.3	0.61	0.9806
Viceroy Rich Lights	10.6	8.6	0.69	0.9693
Virginia Slims	13.9	15.2	1.02	0.9496
Winston Lights	14.9	12	0.82	1.1184

Here, we need to model carbon monoxide content, y, as a function of tar content, x_1, nicotine, x_2, and weight, x_3, using the linear model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

We used SPSS (Statistical Package for Social Sciences) program to analyze the data as the following:

Table (4.2) Correlations

		carbon monoxide Y, milligram	Weight x3 grams	Nicotine x2 milligrams	Tar x1 milligrams
carbon monoxide y, milligrams	Pearson Correlation	1	.464*	.926*	.957*
	Sig. (2-tailed)		.019	.000	.000
Weight x3 grams	Pearson Correlation	.464*	1	.500*	.491*
	Sig. (2-tailed)	.019		.011	.013
Nicotine x2 milligrams	Pearson Correlation	.926*	.500*	1	.977*
	Sig. (2-tailed)	.000	.011		.000
Tar x1 milligrams	Pearson Correlation	.957*	.491*	.977*	1
	Sig. (2-tailed)	.000	.013	.000	

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

a. Listwise N=25

From table (4.2), we can conclude that all the correlation coefficients are significant at (0.05) significance level.

Table (4.3) model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.958(a)	.919	.907	1.44573	.919	78.98	3	21	.000

a Predictors: (Constant), Tar x_1 milligrams, Weight x_3 grams, Nicotine x_2 milligrams

From table (4.3), we can see that the coefficient of determination is 0.919; therefore, about 91.9% of the variation in the Carbon Monoxide in cigarettes is explained by Tar, Nicotine, and Weight. The regression equation appears to be very useful for making predictions since the value of R^2 is close to 1, but may be this indicator indicates that multicollinearity problem is exist.

Table (4.4) ANOVA table[(b)]

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	495.258	3	165.086	78.984	.000[(a)]
	Residual	43.893	21	2.090		
	Total	539.150	24			

Predictors: (Constant), Tar x1 milligrams, Weight x3 grams, Nicotine x2 milligrams b Dependent variable: carbon monoxide y, milligrams

Since p-value < 0.01, so at the $\alpha = 0.05$ level of significance, there exists enough evidence to conclude that at least one of the predictors is useful for predicting the Carbon Monoxide; therefore the model is useful.

Table (4.5) Coefficients (t) test

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.200	3.461		.924	.366
	Weight x_3 grams	-.128	3.885	-.002	-.033	.974
	Nicotine x_2 milligrams	-2.632	3.901	-.197	-.675	.507
	Tar x_1 milligrams	.963	.242	1.151	3.974	.001

a Dependent Variable: carbon monoxide y, milligrams

From table (4.5), the model will be as:

Carbon monoxide(y) = 3.2 - 0.128weight(x_1) – 2.632nicotine(x_2) +0.963tar(x_3)

and we can conclude that the slope of the Tar variable is not zero since p-value < 0.001and, hence, that Tar is useful (with Nicotine and Weight) as a predictor of Carbon Monoxide.

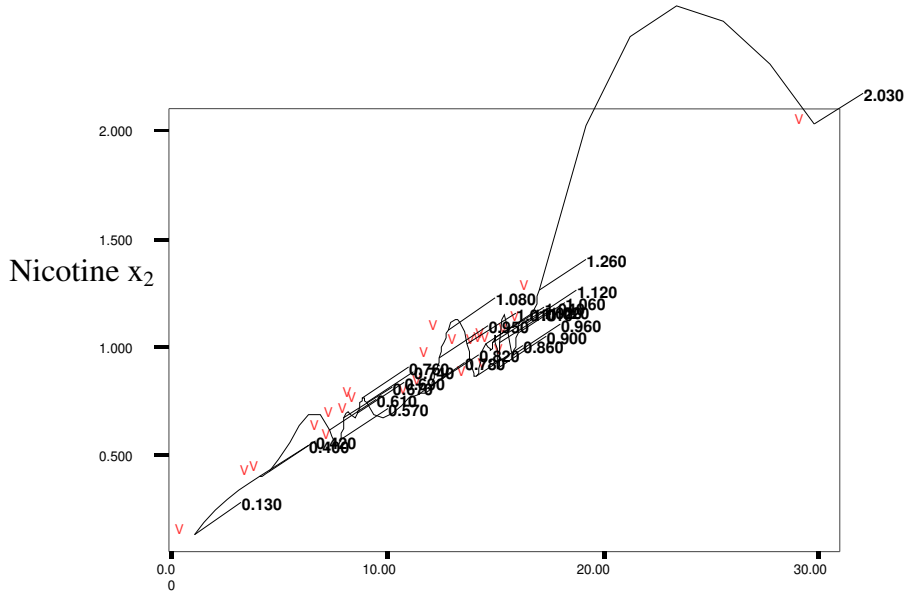
But the both slopes of Weight and the Nicotine are equal to zero.

This means that there something is not normal in our analysis, so we should test for the multicollinearity.

Table (4.6) multicollinearity diagnosing

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3.200	3.461		.924	.366		
	Weight x3 grams	-.128	3.885	-.002	-.033	.974	.750	1.334
	Nicotine x2 milligrams	-2.632	3.901	-.197	-.675	.507	.046	21.900
	Tar x1 milligrams	.963	.242	1.151	3.974	.001	.046	21.631

From table (4.6), since the two predictors, Nicotine and Tar have a variance inflation factor (VIF) greater than ten, there are apparent multicollinearity problems in the model.



Fig(4.1)
Relation between tar and nicotine

Also, from figure (4.1), we can conclude that there is aproximately quadratic relation between tar and nicotine, which indicates to multicollinearity problem.

Table (4.7) Coefficient Correlations [(a)]

		Weight x_3 grams	**Nicotine x_2 milligrams**	**Tar x_1 milligrams**
Weight x_3 grams	Pearson Correlation	1	.500(*)	.491(*)
	Sig. (2-tailed)		.011	.013
	N	25	25	25
Nicotine x_2 milligrams	Pearson Correlation	.500(*)	1	.977(**)
	Sig. (2-tailed)	.011		.000
	N	25	25	25
Tar x_1 milligrams	Pearson Correlation	.491(*)	.977(**)	1
	Sig. (2-tailed)	.013	.000	
	N	25	25	25

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

From table (4.7), we can see that tar content x_1 and nicotine content x_2 appear to be highly correlated (r = 0.977), whereas weight x_3 appears to be moderately correlated with both tar content (r = 0.491) and nicotine content

(r = 0.500). In fact, all three-sample correlations are significantly different from zero based on the small *p*-values as shown in table (4.7).

Now, we will rerun the analysis of the model systematically to get at extra sum of squares $SSR(X_2/X_1)$ and $SSR(X_1/X_2, X3)$ by depending on SSE.

We run the analysis with one independent variable.

Table (4.8) ANOVA y/x_1

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	494.281	1	494.281	253.370	.000[(a)]
	Residual	44.869	23	1.951		
	Total	539.150	24			

a Predictors: (Constant), Tar x_1 milligrams

b Dependent Variable: carbon monoxide y, milligrams

From table (4.8) we get $SSE(x_1) = 44.869$, then we run the analysis with two independent variables, the result shown in table (4.9)

Table (4.9) ANOVA $y/x_1, x_2$

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	495.255	2	247.628	124.110	.000[(a)]
	Residual	43.895	22	1.995		
	Total	539.150	24			

a Predictors: (Constant), Nicotine x_2 milligrams, Tar x_1 milligrams

b Dependent Variable: carbon monoxide y, milligrams

From table (4.9), we get $SSE(x_1,x_2) = 43.895$ and

$SSR(x_2/x_1) = SSE(x_1) - SSE(x_1,x_2) = 44.869 – 43.895 = 0.974$

Now we can find $SSR(x_1/x_2)$ from table (4.10)

Table (4.10) ANOVA y/x_2

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	462.256	1	462.256	138.266	.000[(a)]
	Residual	76.894	23	3.343		
	Total	539.150	24			

a Predictors: (Constant), Nicotine x_2 milligrams

b Dependent Variable: carbon monoxide y, milligrams

$$SSR(x_1/x_2) = SSE(x2) - SSE(x_1,x_2)$$
$$= 76.894 - 43.895 = 32.999$$

Table (4.11) sum of squares

SSR	SSE
$SSR(x_1) = 494.281$	$SSE(x_1) = 44.869$
$SSR(x_2) = 462.256$	$SSE(x_2) = 76.894$
$SSR(x_1, x_2) = 495.255$	$SSE(x_1, x_2) = 43.895$
$SSR(x_1 \mid x_2) = 32.999 \neq SSR(x_1)$	
$SSR(x_2 \mid x_1) = 0.974 \neq SSR(x_2)$	

Therefore, by the extra sum of squares, we can conclude from the table (4.11) that there is a severe multicollinearity problem in the model, and we cannot omit any independent variable from the model because logically there is very strong relation between the three independent variables in the model.

So the question, which appears here, is how can we solve this problem? One of the remedial methods of multicollinearity is a ridge regression, which first introduced by Hoerl and Kennard (1970), it is one of the most popular methods that have been suggested for the multicollinearity problem. This regression enables us to inference on values of predictor variables that follow the same pattern of multicollinearity and this aspect is very important.

Solving multicollinearity by ridge regression

By using R program, we got the following:

C	β_0	β_1	β_2	β_3
0.070	3.106198	-0.1283334	-1.676919	0.9017659

modified HKB estimator is 0.07112338

The program reached to this result after 70 iterations, and the model was:

Carbon monoxide(y) = 3.106198 - 0.1283334weight(x_1) – 1.676919nicotine(x_2) +0.9017659tar(x_3)

To make sure of these results, we plot these results as follows:

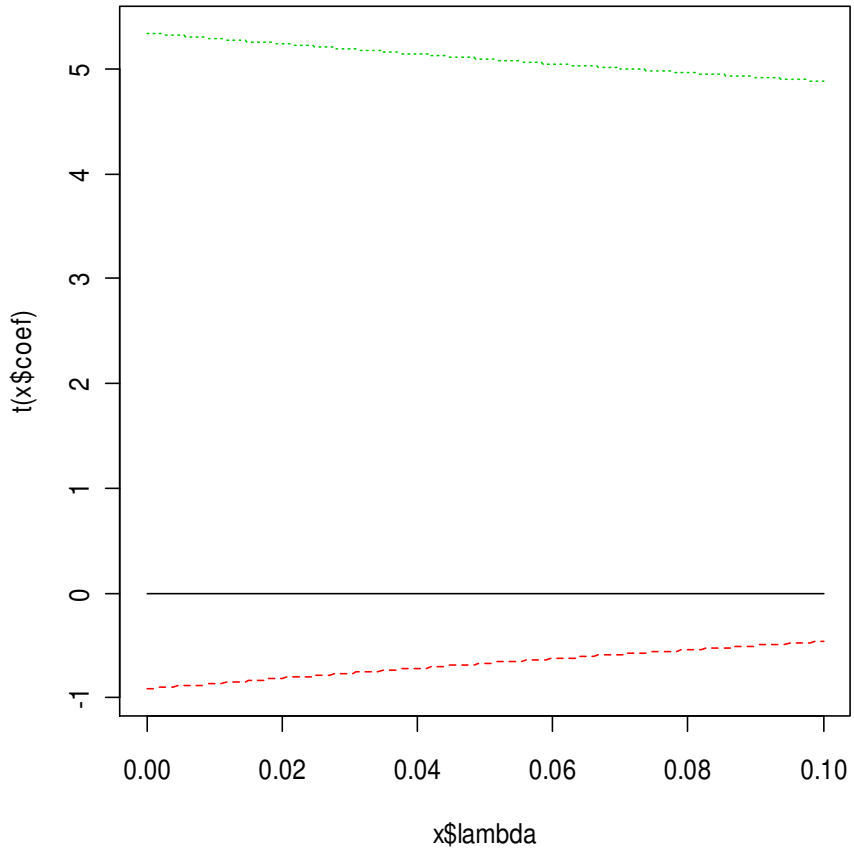


Fig (4.2) coefficients plot

We see from figure (4.2), that the modified value was approximately 0.07.

Therefore, by using ridge regression, we got the best model for the data without omitting any of the explanatory variables.

4. Conclusion

In this paper, we concentrated on the extra sum of squares method as a suggested method that can be used for detecting multicollinearity. The method of extra sum of squares is applied to real data on the annually surveys about smoking were conducted by the American Federal Trade Commission (FTC). In this data, we detected multicollinearity, then we solved this problem by using the ridge regression and we got the new estimates of the new model without omitting any of the explanatory variables.

References

1. Belsley, D.A., Kuh, E. and Welsch, R.E. 1980, "Regression Diagnostics: Identifying influential Data and Sources of Collinearity", John Wiley and Sons., New York.
2. Hoerl, A. E., and R. W. Kennard, 1970, "Ridge regression: biased estimation for Non-orthogonal problems", Technometrics, 12, 55-67.
3. Low Chin 2006, "A Study on How Sum of Squares Can be Used to Detect Multicollinearity", School of Mathematical Sciences Penang, Malaysia. June 13- 15, 2006.
4. M. Wissmann, H. Toutenburg and Shalabh 2007. "Role of Categorical Variables in Multicollinearity in the Linear Regression Model", Technical Report Number 008, 2007, Department of Statistics University of Munich.
5. Neter, Wasserman and Kutner 1989. "Applied Linear Regression Models", 2nd edition. Irwin. Homewood IL.
6. Thomas P. Rayan 2009, "Modern Regression Methods", John Wiley and Sons. New York.