COMMUNICATIONS

DE LA FACULTÉ DES SCIENCES
DE L'UNIVERSITÉ D'ANKARA

Série A_1 : Mathématiques

TOME 30 — **ANNÉE 1981**

ALA MEMOIRE D'ATATÜRK AU CENTENAIRE DE SA NAISSANCE

The Effects Of Multicollinearity –A Geometric View–

by

Fikri AKDENİZ and Fikri ÖZTÜRK

3

Faculté des Sciences de l'Université d'Ankara
Ankara, Turquie

Communications de la Faculté des Sciences de l'Université d'Ankara

---

La Revue "Communications de la Faculté des Sciences de l'Université d'Ankara" est un organe de publication englobant toutes les disciplines scientifiques représentées à la Faculté.

La Revue, jusqu'a 1975 à l'exception des tomes I, II, III, était composée de trois séries:

Série A: Mathématiques, Physique et Astronomie.

Série B: Chimie.

Série C: Sciences naturelles.

A partir de 1975 la Revue comprend sept séries:

Série A_1: Mathématiques
Série A_2: Physique
Série A_3: Astronomie
Série B : Chimie
Série C_1: Géologie
Série C_2: Botanique
Série C_3: Zoologie

En principe, la Revue est réservée aux mémoires originaux des membres de la Faculté. Elle accepte cependant, dans la mesure de la place disponible, les communications des auteurs étrangers. Les langues allemande, anglaise et française sont admises indifféremment. Les articles devront être accompagnés d'un bref sommaire en langue turque.

Adres: Fen Fakültesi Tebliğler Dergisi Fen Fakültesi, Ankara, Turquie.

DEDICATION TO ATATÜRK'S CENTENNIAL

Holding the torch that was lit by Atatürk in the hope of advancing our Country to a modern level of civilization, we celebrate the one hundredth anniversary of his birth. We know that we can only achieve this level in the fields of science and technology that are the wealth of humanity by being productive and creative. As we thus proceed, we are conscious that, in the words of Atatürk, "the truest guide" is knowledge and science.

As members of the Faculty of Science at the University of Ankara we are making every effort to carry out scientific research, as well as to educate and train technicians, scientists, and graduates at every level. As long as we keep in our minds what Atatürk created for his Country, we can never be satisfied with what we have been able to achieve. Yet, the longing for truth, beauty, and a sense of responsibility toward our fellow human beings that he kindled within us gives us strength to strive for even more basic and meaningful service in the future.

From this year forward, we wish and aspire toward surpassing our past efforts, and with each coming year, to serve in greater measure the field of universal science and our own nation.

The Effects Of Multicollinearity –A Geometric View–

Fikri AKDENİZ and Fikri ÖZTÜRK

University of Ankara, Faculty of Science, Department of Mathematics, Ankara, TURKEY.

(Received April 24, 1981; accepted July 15, 1981)

SUMMARY

This paper considers the effects of multicollinearity on the multiple coefficient of determination and on the estimated regression coefficients

1. INTRODUCTION

Consider the equation

$$\underline{Y} = \underline{1}\,\beta_o + Z\,\underline{\delta} + \underline{u} \qquad (1)$$

where $\underline{Y}$ is a nxl vector of observations on the dependent variable, $\underline{1}$ is a nxl vector of unit elements, $Z = (\underline{Z}_1, \underline{Z}_1, \ldots, \underline{Z}_{p-1})$ is nx (p–1) matrix of observations on (p–1) nonstochastic regressors, $\underline{\delta}$ is a (p–1) xl vector of unknown slope coefficients, β_o is an unknown constant, and $\underline{u}$ is a nxl vector of disturbances. It is assumed that $E(\underline{u}) = \underline{0}$, $E(\underline{u}\underline{u}') = \sigma^2 I$, $0 < \sigma^2 < \infty$, $X = (\underline{1}, Z)$ is a matrix of fixed elements, the rank of X is p. Let $\underline{\beta} = (\beta_o, \underline{\delta})'$. The objective is to estimate β. The best linear unbiased estimate of β is

$$\hat{\beta} = (X'X)^{-1} X'Y \qquad (2)$$

with

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}. \qquad (3)$$

If the Z's are higly multicollinear, the variance of $\underline{\hat{\beta}}$ tends to become large, and little confidence can be placed in $\underline{\hat{\beta}}$ as an estimate of $\underline{\beta}$.

In this paper, we use the definition of collinearity given by Silvey [9] ,in which collinearity is said to exist if there are one or more linear relationship between the predictor variables for each observation. Of course, in practice collinearity is said to exists when the linear relationships hold only approximately. (see Mason, Gunst and Webster [7]. This problem has been discussed throughly in the excellent paper by Farrar and Glauber [1], who gave precise statistical procedures for detecting and localizing sources of collinearity within a given data set.

This paper considers the effects of multicollinearity on the multiple coefficient of determination from geometric viewpoint. Furthermore, the limiting cases are investigated of the estimated regression coefficients.

2. EFFECTS OF MULTICOLLINEARITY ON R^2

The interpretation of the value of the multiple coefficient of determination, R^2, is effected by multicollinearity. First, consider the estimated regression parameters and the partial correlation coefficients using the perpendicular projection operators to find the value $\hat{\underline{\beta}}$ that minimize Q, we write Q as

$$Q = \| \underline{Y} - X \underline{\beta} \|^2$$

and notice that Q is the squared distance of $\underline{Y}$ from [X] the subspace of Euclidian n-space E^n spanned by the columns of X. Minimizing Q corresponds then to finding the point in [X] closest to $\underline{Y}$. When X has full column rank, then $X (X'X)^{-1} X'$ and $I - X (X'X)^{-1} X'$ are perpendicular projection operators onto the [X] and $[X]^{\perp}$ respectively, where $[X]^{\perp}$ is the orthogonal complement of [X], such that $R^n = [X] \oplus [X]^{\perp}$.

Now, let us assume that $r_{ij(p-3)}$ the partial correlation coefficient of Z_i and Z_j adjüsted for the remaining (p–3) regressor variables is defined as

$$r_{ij(p-3)} = z_i z_j / \| z_i \| . \| z_j \| \quad (4)$$

where $z_i = (I - B (B'B)^{-1} B') \underline{Z}_i$, $z_j = (I - B (B'B)^{-1} B') \underline{Z}_j$ and $B = (\underline{1}, \underline{Z}_1, \underline{Z}_2, \ldots, \underline{Z}_{i-1}, \underline{Z}_{i+1}, \ldots, \underline{Z}_{j-1}, \underline{Z}_{j+1}, \ldots, \underline{Z}_{p-1})$. Applying the appropriate complementary perpendicular projection operators to $\underline{Z}_i$ and $\underline{Z}_j$, which yields the partial correlation coefficient $r_{ij(p-3)}$ as a direction cosine between the vectors z_i and z_j.

If one lets $B = \underline{1}$, then the partial correlation coefficient between the vectors $\underline{Z}_i$ and $\underline{Z}_j$ becomes the simple correlation coefficient, that is

$$r_{ij} = (Z_i - \underline{1}\,\bar{Z}_i)'\,(Z_j - \underline{1}\,\bar{Z}_j)\,/\,\|\,Z_i - \underline{1}\,\bar{Z}_i\,\|\,.\,\|\,Z_j - \underline{1}\,\bar{Z}_j\,\|. \qquad (5)$$

where $\bar{Z}_i = 1/n \sum_{k=1}^{n} Z_{k_i}$ and $\bar{Z}_j = 1/n \sum_{k=1}^{n} Z_{k_j}$. Thus, the simple correlation is the direction cosine between the perpendicular projections in $[\underline{1}]^{\perp}$. Where the perpendicular projection operator is

$$A = I_n - 1/n\,\underline{1}\,\underline{1}'. \qquad (6)$$

Multiplying the model (1) by matrix A, we get

$$AY = A\,\underline{1}\,\beta_o + A\,Z\,\underline{\delta} + A\underline{u} \qquad (7)$$

or

$$\underline{Y} - \underline{1}\,\bar{Y} = (Z_1 - \underline{1}\,\bar{Z}_1, \ldots, Z_{p-1} - \underline{1}\,\bar{Z}_{p-1})\,\underline{\delta} + A\,\underline{u}.$$

If we wish to use the following deviations

$\tilde{\underline{Y}} = \underline{Y} - \underline{1}\,\bar{Y}$, $\tilde{Z}_i = Z_i - \underline{1}\,\bar{Z}_i$, $i = 1,2,\ldots,p-1$ and $\tilde{\underline{u}} = A\,\underline{u}$, we can write equation (8) as follows:

$$\tilde{Y} = (\tilde{Z}_1, \ldots, \tilde{Z}_{p-1})\,\underline{\delta} + \tilde{\underline{u}}$$

or

$$\tilde{\underline{Y}} = \tilde{Z}\,\underline{\delta} + \tilde{\underline{u}}. \qquad (9)$$

Let us define the following simple functions of the normalized vectors:

$$\underline{Y}^* = 1/\sqrt{n-1}.\ \tilde{\underline{Y}}/S_Y = \tilde{\underline{Y}}/\|\tilde{\underline{Y}}\|,\ Z^*_i = \tilde{Z}_i/\sqrt{n-1}.\ S_{z_i} = \tilde{Z}_i/\ \|Z_i\|$$

$i = 1,\ldots,p-1$ where $S^2_Y = 1/_{(n-1)} \sum_{k=1}^{n} (y_i - \bar{Y})^2$ and $S^2_{z_i} = 1/(n-1) \sum_{k=1}^{n} (Z_{k_i} - \bar{Z}_i)^2$, $i = 1,\ldots,p-1$. Hence, we have

$$\underline{Y}^* = \delta_1\,S_{z_1}/S_Y\,Z^*_1 + \ldots + \delta_{p-1}\,S_{z_{p-1}}/S_Y.\,Z^*_{p-1} + \underline{u}^*$$

or

$$\hat{\underline{Y}}^* = Z^*\,\hat{\underline{\delta}}^*. \qquad (10)$$

where

$$\hat{\delta}_i.\,S_{z_i}/S_Y = \hat{\delta}^*_i,\ i = 1,2,\ldots,p-1 \text{ and } Z_i^{*\prime}\,Z^*_j = r_{ij}.$$

On the other hand, let us consider the following matrices

$$R^* = (\underline{Y}^*, Z^*_1, \ldots, Z^*_{p-1})' \, (\underline{Y}^*, Z^*_1, \ldots, Z^*_{p-1}) \qquad (11)$$

and

$$R_{11} = Z^{*\prime} . \, Z^*, \qquad (12)$$

where R^* is the (symmetric) correlation matrix and R_{11} is the submatrix of R^* formed by deleting the first row and first column of R^*. Furthermore, we get

$$\hat{\delta}^*_i = (-1)^{i-1} \, | R_{1\,(i+1)} | / | R_{11} | \qquad i = 1, \ldots, p-1 \qquad (13)$$

and

$$R^2 = 1 - | R | / | R_{11} | . \qquad (14)$$

(see for example Johnson [4]).

Now, consider the multiple coefficient of determination, R^2, that is,

$$R^2 = \sum_{j=1}^{n} (\hat{y}_j - \bar{Y})^2 / \sum_{j=1}^{n} (y_j - \bar{Y})^2 = \| \hat{\underline{Y}} - \underline{1} \, \bar{Y} \|^2 / \| \underline{Y} - \underline{1} \, \bar{Y} \|^2 .$$

Then, R^2 can be expressed as:

$$R^2 = \| \hat{\underline{Y}} - \underline{1} \, \bar{Y} \|^2 / \| \underline{Y} - \underline{1} \, \bar{Y} \|^2 = \| A \, \hat{\underline{Y}} \|^2 / \| A \, \underline{Y} \|^2 = \| \hat{\tilde{\underline{Y}}} \|^2 / \| \tilde{\underline{Y}} \|^2 = (\cos \theta)^2 . \qquad (15)$$

(see Figure 1)

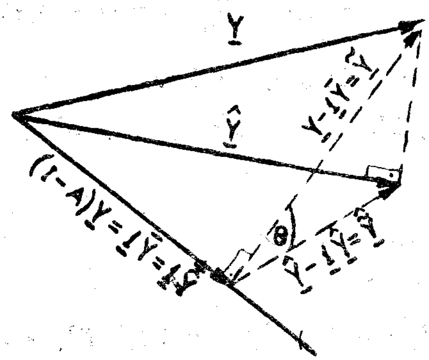

Fig 1

If there exist a multicollinearity among the vectors $Z_1, Z_2, \ldots, Z_K$, $K \leq p-1$, then $Z_1, Z_2, \ldots, Z_K$ are approximately near the hyperplane. From this explanation, we observe that $R^2 z_i$ is close to 1, where $R^2 z_i$ $(i \leq K)$ is the squared multiple correlation of Z_i and other independent variables $Z_1, Z_2, \ldots, Z_{i+1}, Z_{i\ 1}, \ldots, Z_{p-1}$. Thus, this implies that Marquardt's [6] variance inflation factor $(1-R^2 z_i)^{-1}$ is very large. Where $(1-R^2 z_i)^{-1}$ is the i^{th} diagonal element of R_{11}^{-1}.

3. TWO REGRESSOR VARIABLES

When $p-1 = 2$, then the expression for the coefficient of determination with two independent variables is

$$R^2 = (r^2_{y1} + r^2_{y2} - 2\, r_{y1}\, r_{y2}\, r_{12})/(1 - r^2_{12}). \qquad (16)$$

Fox and Cooney [2] and Fox [3] are concerned numerically the effects of multicollinearity for the case of two independent variables on the coefficient of determination for any particular values of r_{y1} and r_{y2}. For the case of two independent variables, the multiple correlation coefficient of determination is shown graphically as a function of the pairwise correlation coefficient and the intercorrelation of the independent variables by Weber and Monarchi [10], see Figure 2.

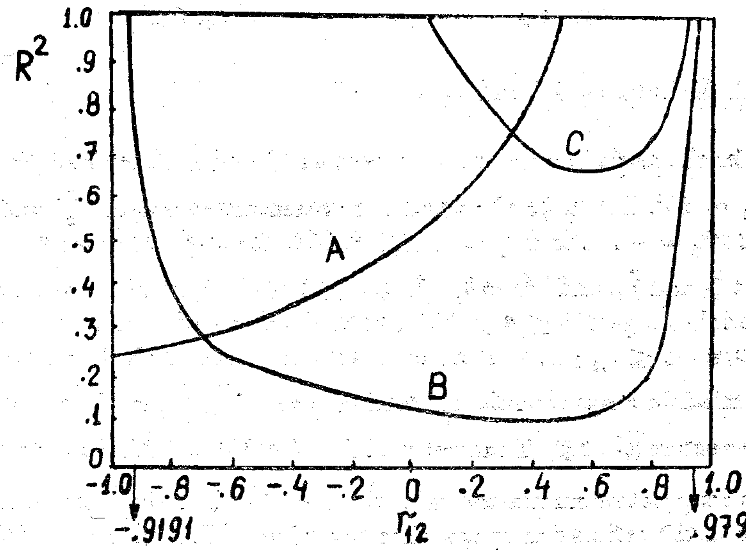

Fig 2

A: $r_{y1} = 0.5$ $r_{y2} = -0.5$

B: $r_{y1} = 0.1$ $r_{y2} = 0.3$

C: $r_{y1} = 0.8$ $r_{y2} = 0.6$

In this section, we would like to investigate from a geometric view-point, the variation of R^2, over the range [0, 1]. Note that R^2 is bounded over the range [0, 1]. For the special case of two explanatory variables, geometric represantations are shown in Figures 3A and 3B.

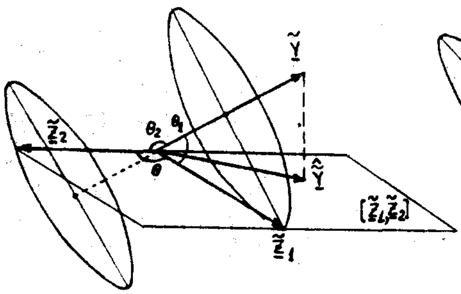

Fig 3 A

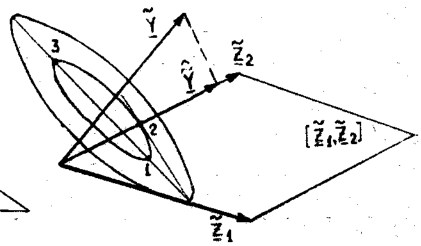

Fig 3 B

CASE A: $r_{y1} = 0.5$ and $r_{y2} = -0.5$

Let θ_2 be the angle between the two vectors $\underline{\tilde{Y}}$ and $\tilde{Z}_1$. Then $\cos\theta_1 = 1/2$ and $\theta_1 = 60°$. Let θ_2 be the angle between the two vectors $\underline{\tilde{Y}}$ and $\tilde{Z}_2$. Then $\cos\theta_2 = -1/2$ and $\theta_2 = 120°$. Let θ be the angle between the two vectors $\tilde{Z}_1$ and $\tilde{Z}_2$, and hence $\theta_2 - \theta_1 \leq \theta \leq \theta_2 + \theta_1$. When θ is varied over its possible range $60° \leq \theta \leq 180°$; then the value of r_{12} is restricted to the range $-1 \leq r_{12} \leq 0.5$. If r_{12} attains its minimum at the lower limit of its permissible range then the estimation space. $[\tilde{Z}_1, \tilde{Z}_2]$ is a line spanned by the vectors $\tilde{Z}_1$ or $\tilde{Z}_2$. Thus, we get $R^2 = (\cos\theta_1)^2 = 0.25$. Furthermore, r_{12} has its greatest numerical value for $\theta = 60°$. For this case, since $\underline{\tilde{Y}}$ will be in the estimation space, then the value of R^2 becomes 1. While r_{12} varies over its possible range, $-1 \leq r_{12} \leq 0.5$, R^2 varies over the range [0.25, 1]. (see Figure 3A)

CASE B: $r_{y1} = 0.1$ and $r_{y2} = 0.3$

For this case, we have θ_1 and θ_2 equal to 84° 16′ and 72° 33′ respectively. Since θ varied over the range [11° 43′, 156° 49′], so r_{12} varies over the range [−0.9191, 0.9791].

If we start from the upper limit of the range, namely $r_{12} = 0.9791$ and $\theta = 11° 44'$, $\tilde{Z}_2$ will be in the position 1. (see Figure 3B)

Since $\underline{Y}$ is in the estimation space (plane), then $R^2 = 1$. For $r_{12} = -0.9191$ or $\theta = 156° 49'$, $\tilde{Z}_2$ will be in position 3, and R^2 will be equal to 1. As $\tilde{Z}_2$ moves from position 1 to position 2, R^2 will decrease, and when it moves from position 2 to position 3, R^2 will increase.

When R^2 reaches its minimum value at position 2, the regression coefficient of a variable, which has a weak correlation with $\tilde{\underline{Y}}$, will vanish and change sign.

4. EFFECTS OF MULTICOLLINEARITY ON THE ESTIMATED REGRESSION COEFFICIENTS

Let us assume that variables Z_1 and Z_2 are intercorrelated but that all other explanatory variables have zero correlations with each other and with Z_1 and Z_2. Then, from equation (13), we have

$$\hat{\delta}^*_1 = (r_{y1} - r_{y2}\, r_{12})/(1 - r^2_{12}) \tag{17}$$

and

$$\hat{\delta}_2^* = (r_{y2} - r_{y1}\, r_{12})/(1 - r^2_{12}) \tag{18}$$

Let us rewrite the expression for $\hat{\delta}^*_1$ as (see Klein and Nakamura [5]

$$\begin{aligned}\hat{\delta}^*_1 &= (r_{y1} - r_{y2})/(1 + r_{12})(1 - r_{12}) + r_{y2}/(1 + r_{12}) \qquad (19)\\ &= (\underline{Y}^{*\prime} \underline{Z}^*_1 - \underline{Y}^{*\prime} \underline{Z}^*_2)/(1 + r_{12})(\underline{Z}^{*\prime}_1 \underline{Z}^*_1 - \underline{Z}^{*\prime}_1 \underline{Z}^*_2) + r_{y2}/(1 + r_{12})\\ &= \underline{Y}^{*\prime}(\underline{Z}^*_1 - \underline{Z}^*_2)/(1 + r_{12})\underline{Z}^{*\prime}_1(\underline{Z}^*_1 - \underline{Z}^*_2) + r_{y2}/(1 + r_{12}).\end{aligned}$$

Divide the numerator and denominator of the first term on the right by $\| \underline{Z}^*_1 - \underline{Z}^*_2 \|$. We have

$$\hat{\delta}^*_1 = r_{y(1-2)}/(1 + r_{12})\, r_{1(1-2)} + r_{y2}/(1 + r_{12}) \qquad (20)$$

Similarly, we obtain

$$\hat{\delta}^*_2 = - r_{y(1-2)}/(1 + r_{12})\, r_{1(1-2)} + r_{y1}/(1 + r_{12}) \qquad (21)$$

where $r_{y(1-2)}$ is the correlation between Y^* and $Z^*_1 - Z^*_2$, and $r_{1(1-2)}$ is the correlation between Z^*_1 and $Z^*_1 - Z^*_2$.

When $r_{12} \to 1$, the measure of the angle between the vectors $\underline{Z}^*_1$ and $\underline{Z}^*_2$ approaches zero. Then, the measure of the angle between the vectors $\underline{Z}^*_1$, $\underline{Z}^*_1 - \underline{Z}^*_2$ approaches 90°, thus $r_{1(1-2)}$ necessarily approaches zero as $r_{12} \to 1$.

In equations (20) and (21), the numerator, $r_{y(1-2)}$ may approaches to either zero or a nonzero value depending on the given vector $\underline{Y}^*$ and the direction in wich $\underline{Z}^*_1$ approaches $\underline{Z}^*_2$.

If $r_{y1} = r_{y2}$, then the value of r_{12} is resticted to the range $[2r^2_{y1} - 1, 1]$ by the condition $R^2 \leq 1$. Hence, from equations (20) and (21), we get the limiting values of $\hat{\delta}_1^*$ and $\hat{\delta}^*_2$ as

$$\lim_{r_{12} \to 1} \hat{\delta}^*_1 = \lim_{r_{12} \to 1} \hat{\delta}^*_2 = r_{y1}/2 = r_{y2}/2.$$

(see for example Sastry [8])

When $r_{y1} = r_{y2}$, the perpendicular projection, $\hat{Y}^*$ of Y^* can be written as a linear combination of the vectors $\underline{Z}^*_1$ and $\underline{Z}^*_2$ (see Figure 4A) that is, $\underline{\hat{Y}}^* = \hat{\delta}^*_1 \underline{Z}^*_1 + \hat{\delta}^*_2 \underline{Z}^*_2$. In the limiting case, we get

$$\lim_{r_{12} \to 1} \hat{\delta}^*_1 = \lim_{r_{12} \to 1} \hat{\delta}^*_2 = r_{y1}/2 = r_{y2}/2.$$

When $r_{y1} \neq r_{y2}$ (but $r_{y1} \to r_{y2}$), while r_{12} approaches its maximum positive value over the possible range, in other words, while the vector $\underline{Z}^*_2$ moves from position 2 to position 1, (see Figure 4B) then absolute value of each of the estimated regression coefficients, δ^*_1 and δ^*_2, will be large and in opposite signs.

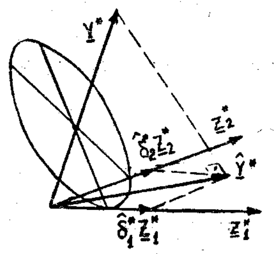

Fig 4 A

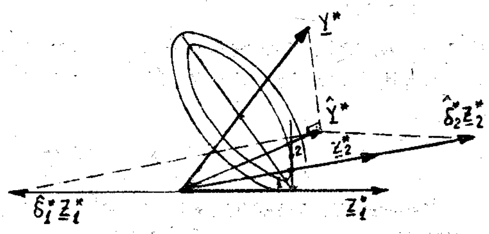

Fig 4 B

5. CONCLUSIONS

This paper has attempted to disclose to the users of regression procedures a better understanding of the limitations of ordinary least squares when multicollinearity is present in the data. Hence, geometric representations are helpful in revealing the sources of collinearity. The results of this paper show that, in general, the multiple coefficient of determination is not monotonic function of the multicollinearity for fixed pairwise correlations.

REFERENCES

[1] **Farrar, D.E. and Glauber, R.F.,** "Multicolliearity in Regression Analysis: The Problem Revisited" Review of Economics and Statistics, Vol. 49, (1967), 92–107.

[2] **Fox, K.A. and Conney, J.F.,** "Effects of Intercorrelations Upon Multiple Correlation and Regression Measures" U.S. Dept. of Agriculture, Agricultural Marketing Service, Washington D.C., 1954.

[3] **Fox, K.A.,** *Intermediate Economic Statistics*. New York, John Wiley and Sons, 1967.

[4] **Johson, J.,** *Econometric Methods*, 2 nd Edition. New York, McGraw-Hill, 1972.

[5] **Klein, L.R. and Nakamura, M.,** "Singulartiy in the Equation Systems of Econometrics: Some Aspects of the Problem of Multicollinearity. "International Economic Review, Vol. 3, (1962), 274–299.

[6] **Marquardt, D.W.,** "Generalized Inverses, Biased Linear Estimation and Nonlinear Estimation" Technometrics, Vol. 12, (1970), 591–612.

[7] **Mason, R.L., Gunst, R.F. and Webster, J.T.,** "Regression Analysis and Problems of Multicollinearity" Communications in Statistics, Vol. 4, (1975), 277–292.

[8] **Sastry, M.V.R.,** "Some Limits in the Theory of Multicollinearity" American Statistician, Vol. 24, (1970), 39–40.

[9] **Silvey, S.D.**, "Multicollinearity and Imprecise Estimation", J.R. Statist. Soc., B 31, (1969), 539–552.

[10] **Weber, J.E. and Monarchi, D.E.**, "Graphical Representaion of the Effects of Multicollinearity", Decision Sciences, Vol. 8, (1977), 534–547.

ÖZET

Bu çalışmada iç ilişkinin çoklu belirleyicilik katsayısı ve kestirilmiş regresyon katsayıları üzerindeki etkileri araştırılmıştır. Dik izdüşümler yardımıyla kısmi korelasyon katsayıları arasındaki bağıntılar incelenmiştir.

Prix de l'abonnement aunnel

Turquie: 15 TL; Etranger: 30 TL.
Prix de ce numèro: 5 TL (pour la vente en Turquie).
Prière de s'adresser pour l'abonnement à: Fen Fakültesi Dekanlığı Ankara, Turquie.

Ankara Üniversitesi Basımevi. Ankara - 1982