Regression Analysis

Chapter 9: Analysis of Collinear Data

Kyusang Yu and Sunghwan Kim

Department of Applied Statistics

Konkuk University

Spring 2019

Interpretation of a regression coefficient

- The regression coefficient is interpreted as measuring the change in the response variable when the corresponding predictor variable is increased by one unit and all other predictor variables are held constant.
- This interpretation may not be valid if there are strong linear relationships among the predictor variables.
  - It may be impossible to change one variable while holding all others constant.
  - There may be no information about the result of this kind of manipulation in the estimation data.
  - Such interpretation may be impossible because of the structure of the predictors.

Extreme case

- Consider the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \tag{1}$$

- Suppose $X_1 = 2X_2$. Then, the model (1) becomes

$$Y = \beta_0 + (2\beta_1 + \beta_2) X_2 + \beta_3 X_3 + \epsilon$$

- As X_2 increases by one unit, the mean of Y increases by $(2\beta_1 + \beta_2)$ rather than β_2.
- Also, we can't increase X_2 while X_1 is held constant.
- Hence, the usual interpretation of coefficients is not valid.

Multicollinearity

- Predictors are said to be ***orthogonal*** when there is a complete absence of linear relationship among the predictors.
- Multicollinearity (M)
  - **Definition**: The predictors are strongly interrelated, or they have severe non-orthogonality.
  - M is more complex than the pairwise correlation between predictors.
  - M is not a modeling error but a condition of deficient data.

Consequences of multicollinearity

- It is impossible to estimate the unique effects of individual variables.
- The estimated values of the coefficients are very sensitive to slight changes in the data.
- The estimators of coefficients have large sampling errors, which affect both inference and forecasting.
- Multicollinearity is extremely difficult to detect.

Theoretical reason

- Least squares estimator:

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

- The predictors are said to be *orthogonal* if the sample version of $Cov(X_j, X_k) = 0$ for $j \neq k$.
- Then, $\mathbf{X}'\mathbf{X}$ is of full rank and thus, nonsingular and LSE can be defined uniquely as (2).
- Otherwise, $\mathbf{X}'\mathbf{X}$ is singular and LSE cannot be defined as (2).
- If the predictors are strongly interrelated then $\mathbf{X}'\mathbf{X}$ is nearly singular and LSE cannot be computed in practice as (2).

Focus

- Questions:
  - How does M affect statistical inference and forecasting?
  - How can M be detected?
  - What can be done to resolve the difficulties associated with M?
- These questions cannot be answered separately, but simultaneously.

Example: Equal Educational Opportunity (EEO)

- Objective: evaluate the effect of school inputs on achievement.
- Response Variable: Achievement (ACHV)
- Predictors: Aspects of the school environment (SCHOOL), Student's home environments (FAM), Influence of student's peer group (PEER)
- See the data taken for 70 schools selected at random (Tables 9.1-9.2, pp. 236-237).

Model for EEO Data

- Adjustment for the two basic variables (achievement and school) by using the regression model (3).

$$ACHV = \beta_0 + \beta_1 FAM + \beta_2 PEER + \beta_3 SCHOOL + \epsilon \quad (3)$$

$$ACHV - \beta_1 FAM - \beta_2 PEER = \beta_0 + \beta_3 SCHOOL + \epsilon \quad (4)$$

  - The left-hand side of (4) is an adjusted achievement index.
  - The representation (4) is in the form of a regression of the adjusted achievement score on the SCHOOL variable and is used only for the sake of interpretation.

Result for EEO Data (1)

Variable	Coefficient	s.e.	t-Test	p-value
Constant	-0.070	0.251	-0.28	0.7810
FAM	1.101	1.411	0.78	0.4378
PEER	2.322	1.481	1.57	0.1218
SCHOOL	-2.281	2.220	-1.03	0.3080
$n = 70$	$R^2 = 0.206$	$R^2_a = 0.170$	$\hat{\sigma} = 2.07$	$df = 66$

Table: EEO Data: Regression Results

- About 20% of the variation in achievement score is accounted for by the three predictors jointly ($R^2 = 0.206$).

Result for EEO Data (2)

- FAM, PEER and SCHOOL are valid predictor variables if they are taken together.
  - The F-value $= 5.72$ is significant at significance level 0.01.
  - The fact that $R^2 = 0.206$ does not serve as a counterexample.
- The individual t-values are not significant.
  - Any one predictor may be deleted from the model provided the other two are retained.

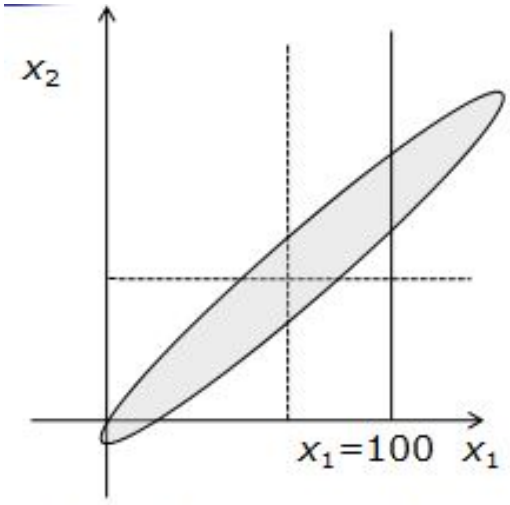
Result for EEO Data (3)

- This is typical of a situation where extreme M is present.
  - Predictors are so highly correlated that each one may serve as a proxy for the others.
  - Pairwise scatterplots and correlations confirm this (Figure 9.2, p. 239).
  - All the observations lie close to the straight line.

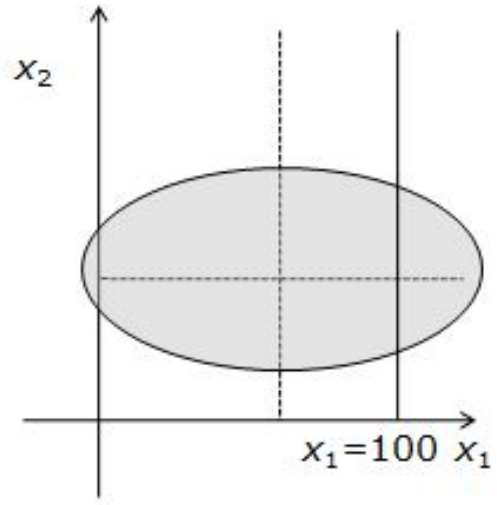
Multicollinearity for EEO

- It is not unreasonable to conclude that there are not three variables but in fact only one.
- Possibilities:
  - The sample data are deficient.
  - The interrelationships are an inherent characteristic of the process.

Data combinations and the interpretation of coefficients



Multicollinear case: One can not check the effect of x_2 holding x_1 fixed. For example, when x_1=100, most values of x_2 are above the average.



Non-multicollinear case: One can check the effect of x_2 holding x_1 fixed. For example, when x_1=100, about half of values of x_2 are above the average and the others, below the average.

Deficient sample (1)

- It can be improved with additional observations.
- In the scatterplot of FAM vs SCHOOL, no schools with values in the upper left or lower right regions.
- Hence, it is impossible to determine the individual effects of FAM and SCHOOL on ACHV.
- Assume that there were some observations on the upper left region.
- Then, it would be possible to compare average ACHV for low and high values of SCHOOL when FAM is held constant.

Deficient sample (2)

- In Table 9.4 (p. 240), + represents a value above the average and − represents a value below average.
- The large correlation suggests that only combinations 1 and 8 are represented in the data.
- Suppose that combinations 1 and 2 alone exist.
- The results would give only a partial answer, namely, an evaluation of the school-achievement relationship when FAM and PEER are both above average.

Remedy of deficient data

- Collect additional data on some of the other combination.
- It is often not possible to collect more data because of constraints on budgets, time, and staff.
- In experimental study, one can find a nice design which allows the 8 combinations in Table 9.4.
- In observational study, usually one cannot design the sampling scheme.
- It is fairly difficult to ensure that a balanced sample will be obtained in observational study.

Inherent characteristic

- FAM, PEER and SCHOOL exist in the population only as data combinations 1 and 8 of Table 9.4.
- It is impossible to estimate the individual effects of these variables on achievement.
- One may have to search for underlying causes that may explain the interrelationships of the predictors.

Summary of effects on inference

- When sample data are deficient,
  - it can be improved with additional data.
  - But, there may exist constraints on budgets, time and staff.
- When interrelationships among the variables are an inherent characteristic of the process, one would search for underlying causes that may explain the interrelationships of the predictor variables.

Effects of collinearity on forecasting

- Forecasts of the response variable are produced by using future values of the predictor variables .
- Future values of the predictors must be known or forecasted from other data and models.
- In our discussion, it is assumed that they are given.

Example: French economy data

- Aggregate data concerning import activity in the French economy
  - IMPORT = Imports
  - DOPROD = Domestic production
  - STOCK = Stock formation
  - CONSUM = Domestic consumption
- Model

$$IMPORT = \beta_0 + \beta_1 DOPROD + \beta_2 STOCK + \beta_3 CONSUM + \epsilon$$

Result for French economy data

Variable	Coefficient	s.e.	t-Test	p-value
Constant	-19.725	4.125	-4.78	0.0003
DOPROD	0.032	0.187	0.17	0.8656
STOCK	0.414	0.322	1.29	0.2195
CONSUM	0.243	0.285	0.85	0.4093
$n = 18$	$R^2 = 0.973$	$R_a^2 = 0.967$	$\hat{\sigma} = 2.258$	df=14

Table: Import data (1949-1966): Regression Results

- Multicollinearity appears to be present ($R^2 = 0.973$ and all t-values are small).

Other model

- Index plot of the standardized residuals shows a distinctive pattern.
  - They decreases until index is equal to 12 and increases after 12.
  - Index = 12 means 1960.
- This pattern suggests that the model is not well specified.
- European Common Market began operations in 1960, and it may cause changes in import-export relations.

Other model

- The effect of M is confounded with the effect of the structural change in 1960.
- Consider the 11 years 1949-1959.
  - Results are summarized in Table 9.7 (p. 243).
  - The residual plot is now satisfactory (Figure 9.4).

1949-1959

Variable	Coefficient	s.e.	t-Test	p-value
Constant	-10.128	1.212	-8.36	< 0.0001
DOPROD	-0.051	0.070	-0.73	0.4883
STOCK	0.587	0.095	6.20	0.0004
CONSUM	0.287	0.102	2.81	0.0263
$n = 11$	$R^2 = 0.992$	$R_a^2 = 0.988$	$\hat{\sigma} = 0.4889$	df=7

Table: Import data (1949-1959): Regression Results

Result

- Negative value for coefficient of DOPROD:
  - It is contrary to prior expectation.
  - We believe that if STOCK and CONSUM were held fixed, an increase in DOPROD would cause an increase in IMPORT.
  - It's not significant.
  - Correlation between CONSUM and DOPROD is 0.997.
  - CONSUM = 6.259 + 0.686·DOPROD

Naive forecast

- Forecast the change in IMPORT corresponding to an increase in DOPROD of 10 units while holding STOCK and CONSUM:
  - $\text{IMPORT}_{1960} = \text{IMPORT}_{1959} - 0.051(10) = \text{IMPORT}_{1959} - 0.51$.
  - This does not seem to correspond to our prior knowledge.

A better forecast

- $\text{CONSUM} = 6.259 + 0.686 \times \text{DOPROD}$
- $\text{IMPORT}_{1960} = \text{IMPORT}_{1959} - 0.51 + 0.287 \times 6.87 = \text{IMPORT}_{1959} + 1.46$
- This is a more satisfying and probably a better forecast since it corresponds to our prior knowledge.
- The case where DOPROD increases alone corresponds to a change in the basic structure of the data that were used to estimate the model parameters and cannot be expected to produce meaningful forecasts.

Problems with multicollinearity

- The two examples demonstrate that multicollinear data can seriously limit the use of regression analysis for inference and forecasting.
- Extreme care is required when M is suspected.
- There are various remedies for M.

Detection of multicollinearity

- Multicolliearity is associated with unstable estimated regression coefficient.
- This situation results from the presence of strong linear relationship among the predictor variables.

Indications of multicollinearity

- Large changes in the estimated coefficients when a variable is added or deleted.
- Large changes in the coefficients when a data point is altered or dropped.
- Signs of the estimated coefficients do not conform to prior expectations: IMPORT data
- Large SE (small t-value): EEO data
- Large correlation coefficients between predictors: IMPORT and EEO datasets

Source of multicollinearity

- It may be more subtle than a simple relationship between two variables.
- Example:
  - The effects of advertising expenditures, promotion expenditures and sales on the aggregate sales of a firm in period
  - A = advertising expenditures
  - P = promotion expenditures
  - E = sales expenditure

Example: Advertising Data (AD)

- The effects of advertising expenditures, promotion expenditures and sales on the aggregate sales of a firm in period t.

Variable	Explanation
A_t	Advertising expenditures
P_t	Promotion expenditures
E_t	Sales expense

Linear Model for AD

- Regression result

$$S_t = \beta_0 + \beta_1 A_t + \beta_2 P_t + \beta_3 E_t + \beta_4 A_{t-1} + \beta_5 P_{t-1} + \epsilon_t$$

- A_{t-1} and P_{t-1} are the lagged one-year variable

Variable	Coefficient	s.e.	t-Test	p-value
Constant	-14.194	18.715	-0.76	0.4592
A_t	5.361	4.028	1.33	0.2019
P_t	8.372	3.586	2.33	0.0329
E_t	22.521	2.142	10.51	< 0.0001
A_{t-1}	3.855	3.578	1.08	0.2973
P_{t-1}	4.125	3.895	1.06	0.3053
$n = 22$	$R^2 = 0.917$	$R_a^2 = 0.891$	$\hat{\sigma} = 1.320$	$df = 16$

Pairwise correlation coefficient

- The pairwise correlation coefficients between the predictor variables are small.
- Multicollinearity may exist under pairwise uncorrelatedness.

	A_t	P_t	E_t	A_{t-1}	P_{t-1}
A_t	1.000				
P_t	−0.357	1.000			
E_t	0.129	0.063	1.000		
A_{t-1}	−0.140	−0.316	−0.166	1.000	
P_{t-1}	−0.496	−0.296	0.208	−0.358	1.000

Stability of estimates

- Dropping the contemporaneous advertising variable A from the model
  - Coefficient of P_t: $8.27 \rightarrow 3.70$
  - Coefficient of P_{t-1} and A_{t-1}: changing signs
  - Coefficient of E_t is stable.
  - R^2 does not change much.
- There is some type of relationship involving the contemporaneous and lagged values of the advertising and promotions variables.

A cause of multicollinearity

- Regression of A_t on P_{t-1}, A_{t-1} and P_t:

$$\hat{A}_t = 4.63 - 0.87P_t - 0.86A_{t-1} - 0.95P_{t-1}$$

  - $R^2 = 0.973$
  - Upon investigation, there was an approximate rule:

$$A_t + P_t + A_{t-1} + P_{t-1} = 5.$$

- The relationship is the cause of the multicollinearity.

Investigation of multicolliearity

- Examining the value of R^2 that results from regressing each of the predictor variables against all the others.
- Relationship between the predictor variables can be judged by examining a quantity (VIF).

VIF

- VIF(variance inflation factors)

$$VIF_j = \frac{1}{1 - R_j^2}, \quad j = 1, 2, \ldots, p$$

- R_j^2 is the coefficient of determination that results when the predictor X_j is regressed against all other predictors.

	R_j^2 be close to 1	$R_j^2 = 0$
VIF_j	large	1
X_j	Presence of linear relationship in the predictor variables	Absence of any linear relationship between the predictor variables

Usage of VIF

- Deviation of VIF_j value from 1 indicates departure from orthogonality and tendency toward collinearity.
- If $VIF_j = 10$, the squared error in the LSE of j-th variable is 10 times as large as it would be if the predictors were orthogonal.
- Suggesting that a VIF_j in excess of 10 is an indication that multicollinearity may be causing problems in estimation.

Table 9.12

- EEO
  - All three variables are strongly intercorrelated.
- IMPORT
  - DOPROD and CONSUM are strongly correlated but are not correlated with STOCK.
- AD
  - E_t is not correlated with the remaining predictors.
  - There is a strong linear relationship among other four variables.