AMA3602
Applied Linear Models
Department of Applied Mathematics

Lecturer : Dr. Catherine Liu

Contact: 2766 6931 (O); Office Venue: TU830

April 2024

Chapter 5

Part I Multicollinearity and Ridge Regression

References:

Chapter 9: Multicollinearity

Montgomery, D.C., Peck, E.A., & Vining, G.G. (2012) *Introduction to Linear Regression Analysis*. 5th Ed. Wiley.

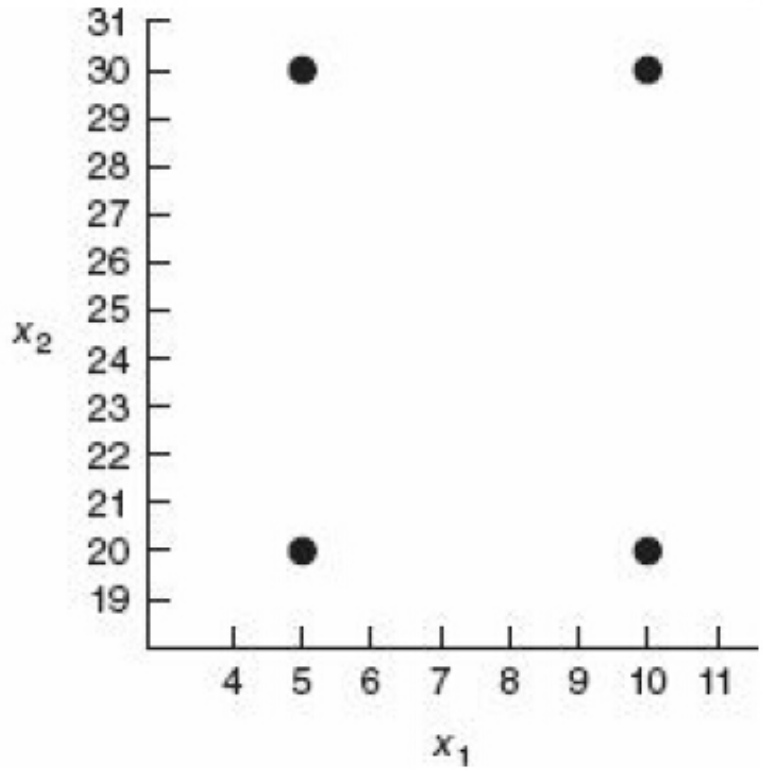
Outline

- Multicollinearity: measures, consequences and solutions
- Ridge regression
- Principle component regression

Multicollinearity

- Multicollinearity is said to exist when there are near-linear dependencies among the regressors.
- The regressors are the columns of the $\boldsymbol{X}$ matrix, so clearly an exact linear dependence would result in a singular $\boldsymbol{X}'\boldsymbol{X}$

x_1	x_2
5	20
10	20
5	30
10	30
5	20
10	20
5	30
10	30



Definition and Source of Multicollinearity

- Model: $\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$, $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 \boldsymbol{I}_n)$
- Muliticollinearity: near-linear dependence of $\boldsymbol{X} = [X_1, \cdots, X_p]$, i.e.there is a set of constants $t_1, \cdots, t_p$, not all zero, such that

$$\sum_{j=1}^{p} t_j X_j \approx 0$$

- primary sources:
  - ⋆ The data collection method employed;
  - ⋆ Constraints on the model or in the population.
  - ⋆ Model specifiction
  - ⋆ An overdefined model

Effects of multicollinearity on LSE

- Unit normal scaling: centered, scaled to unit length: $z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$, where $s_j^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$, $i = 1, \cdots, n, j = 1, \cdots, k$;

  $r_{ij} = \frac{\sum_{u=1}^{n} (x_{ui} - \bar{x}_i)(x_{uj} - \bar{x}_i)}{(S_{ii} S_{jj})^{1/2}} = \frac{S_{ij}}{(S_{ii} S_{jj})^{1/2}}$, where $S_{jj} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$

  $\boldsymbol{X}'\boldsymbol{X} = (r_{ij})_{p \times p}$: a correlation matrix between regressors.
- Unit scaled: $\boldsymbol{y} = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, the least-squares normal equations are:

  $(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$, i.e. $\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$

  The inverse of $(\boldsymbol{X}'\boldsymbol{X})$: $C \equiv (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix}$.

  OLS estimator of $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \frac{r_{1y} - r_{12} r_{2y}}{1-r_{12}^2} \\ \frac{r_{2y} - r_{12} r_{1y}}{1-r_{12}^2} \end{bmatrix}$.
- Effects:
  - Strong multicollinearity between x_1 and x_2 results in large variances and covariances for the LSE of the regression coefficients;
    $|r_{12}| \to 1, \mathrm{Var}(\hat{\beta}_j) = C_{jj}\sigma^2 \to \infty$ and $\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = C_{12}\sigma^2 \to \pm\infty$.
  - Multicollinearity also tends to produce LSE $\hat{\beta}_j$ that are too large in absolute value;
  - When strong multicollinearity is present, LS will generally produce poor estimates of the individual model parameters.

Multicollinearity diagnostics

Techniques for detecting multicollinearity:

- Examination of the correlation matrix
- VIF
- Eigensystem analysis of $\boldsymbol{X}'\boldsymbol{X}$
- Others

Examination of the correlation matrix

- A very simple measure of multicollinearity is inspection of the off-diagonal elements r_{ij}in $\boldsymbol{X}'\boldsymbol{X}$, where $\boldsymbol{X}$ is centered and scaled to unit length. If regressors x_i and x_j are **nearly linearly dependent**, then $|r_{ij}|$ will be **near unity**.
- To illustrate this procedure, consider the acetylene data. Following table presents data concerning the percentage of conversion of n-heptane to acetylene and three explanatory variables. These are typical chemical process data for which a full quadratic response surface in all three regressors is often considered to be an appropriate tentative model.

i	p	T	H	C
1	49.00	1300	7.50	0.01
2	50.20	1300	9.00	0.01
3	50.50	1300	11.00	0.01
4	48.50	1300	13.50	0.01
5	47.50	1300	17.00	0.01
6	44.50	1300	23.00	0.01
7	28.00	1200	5.30	0.04
8	31.50	1200	7.50	0.04
9	34.50	1200	11.00	0.03
10	35.00	1200	13.50	0.03
11	38.00	1200	17.00	0.03
12	38.50	1200	23.00	0.04
13	15.00	1100	5.30	0.08
14	17.00	1100	7.50	0.10
15	20.50	1100	11.00	0.09
16	29.50	1100	17.00	0.09

p: percentage of conversion, T: reactor temperature, H: ratio of H_2 to n-Heptane, C: contact time.

Examination of the correlation matrix

- Full model (Acetylene Data):
  $p = \gamma_0 + \gamma_1 T + \gamma_2 H + \gamma_3 C + \gamma_{12} TH + \gamma_{13} TC + \gamma_{23} HC + \gamma_{11} T^2 + \gamma_{22} H^2 + \gamma_{33} C^2$,
  p: percentage of conversion, T: reactor temperature, H: ratio of H_2 to n-Heptane, C: contact time.
- The least-squares fit:

$$\hat{p} = 35.897 + 4.019T + 2.781H - 8.031C - 6.457TH$$
$$-26.982TC - 3.768HC - 12.54T^2 - 0.973H^2 - 11.594C^2$$

- The correlation matrix:

$$\boldsymbol{x'x} = \begin{bmatrix} 1.000 & 0.224 & -0.958 & -0.132 & 0.443 & 0.205 & -0.271 & 0.031 & -0.577 \\ & 1.000 & -0.240 & 0.039 & 0.192 & -0.023 & -0.148 & 0.498 & -0.224 \\ & & 1.000 & 0.194 & -0.661 & -0.274 & 0.501 & -0.018 & 0.765 \\ & & & 1.000 & -0.265 & -0.975 & 0.246 & 0.398 & 0.274 \\ & & & & 1.000 & 0.323 & -0.972 & 0.126 & -0.972 \\ & & & & & 1.000 & -0.279 & -0.374 & 0.358 \\ & & & & & & 1.000 & -0.124 & 0.874 \\ & & & & & & & 1.000 & -0.158 \\ & & \text{Symmetric} & & & & & & 1.000 \end{bmatrix}$$

- The correlation matrix reveals the high correlation between reactor temperature and contact time since $r_{13} = -0.958$. Furthermore, there are other large correlation coefficients between TC and HC, TC and T^2, T^2 and C^2.
- Examining the simple correlations r_{ij} between the regressors is helpful in detecting near-linear dependence between pairs of regressors. However, inspection of the r_{ij} is not sufficient for detecting anything more complex than pairwise multicollinearity.

VIF

- Variance inflation factor (VIF):

$$VIF_j = C_{jj} = (1 - R_j^2)^{-1}$$

one or more large VIFs indicate multicollinearity.

Practical experience indicates that if any of the VIFs exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity.

- Since the variance of the jth regression coefficients is $C_{jj}\sigma^2$, we can view C_{jj} as the factor by which the variance of $\hat{\beta}_j$ is increased due to near-linear dependences among the regressors.
- It also can be interpreted in other perspective.

Eigensystem analysis of $\boldsymbol{X}'\boldsymbol{X}$

- Assume $\lambda_1, \cdots, \lambda_p$ are the eigenvalues of the positive definite matrix $\boldsymbol{X}'\boldsymbol{X}$. One or more small eigenvalues imply that there are near-linear dependences among the columns of $\boldsymbol{X}$.
- The **condition number** of $\boldsymbol{X}'\boldsymbol{X}$:

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$$

  - ⋆ No serious multicollinearity ($\kappa \leq 100$)
  - ⋆ Moderate to strong multicollinearity ($100 \leq \kappa \leq 1000$)
  - ⋆ Severe multicollinearity ($\kappa \geq 1000$)

- The **condition indices** of $\boldsymbol{X}'\boldsymbol{X}$:

$$\kappa_j = \frac{\lambda_{\max}}{\lambda_j}, j = 1, \cdots, p.$$

  The number of condition indices that are large (say ≥ 1000) is a useful measure of the number of near-linear dependences in $\boldsymbol{X}'\boldsymbol{X}$.

Other methods

- The **determinant** of $\boldsymbol{X}'\boldsymbol{X}$ can be used as an index of multicollinearity. The degree of multicollinearity becomes more severe as $|\boldsymbol{X}'\boldsymbol{X}|$ approaches zero.
- The **signs** and **magnitudes** of the regression coefficients will sometimes provide an indication that multicollinearity is present.
- If adding or removing a regressor produces large changes in the estimates of the regression coefficients, multicollinearity is indicated.

Methods for dealing with multicollinearity

- Collecting additional data
  - in a manner to break up the multicollinearity in the existing data
  - economic constraints; sampling process not available; model/population constraints
- Model respecification
  - to redefine the regressors to preserve information whilst reduce the ill-conditioning
  - to eliminate variables but may damage the predictive power of the model
  - No assurance that the final model will exhibit any lesser degree of multicollinearity
- Estimation methods other than least squares
  - to combat the problem induced by multicollinearity
  - Ridge regression
  - Principle component regression

Ridge regression

- **Ridge regression:** $(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})\hat{\boldsymbol{\beta}}_R = \boldsymbol{X}'\boldsymbol{y}$.
  Solution:

$$\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

where constant k is the biasing parameter. How to determine the value of k?

  - ⋆ Ridge trace: a plot of the elements of $\hat{\boldsymbol{\beta}}_R$ versus k for values of k usually in the interval $0 - 1$.

If multicollinearity is severe, the instability in the regression coefficients will be obvious from ridge trace. As k is increased, some of the ridge estimates will vary dramatically, though the ridge estimate will stabilize at some value of k.

The objective is to select a reasonably small value of k at which the ridge estimates are stable. Hopefully this will produce a set of estimates with smaller MSE than the least-squares estimates.