Multicollinearity

Hong Kong Baptist University

Fall 2021

Multicollinearity

Multicollinearity exists when two or more of the independent variables used in the model are moderately or highly correlated

If we have designed experiments, then the values of the independent variables are well controlled by us and we probably can avoid the problem of multicollinearity.

However, if we have only observational studies, then the values of the independent variables are uncontrolled (but we assumed that they are measured without error; regression with measurement error is an advanced topic in statistics) and so multicollinearity may be a problem.

Undesirable consequences of multicollinearity include (but not limited to) (i) the inverse of the matrix $\boldsymbol{X}'\boldsymbol{X}$ is highly sensitive to rounding errors in the calculation of the LSE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ and (ii) inflated standard errors.

Typical indicators of multicollinearity are:

1. High correlations between pairs of independent variables.
2. Insignificant t-test for all or nearly all the individual coefficients while the F-test for overall model adequacy is significant.
3. Opposite signs (from what is expected) in the estimated parameters.
4. Large changes in the parameter estimates resulted from deletion of a row or column of the $\boldsymbol{X}$ matrix. See Figure 1 for a geometric interpretation.

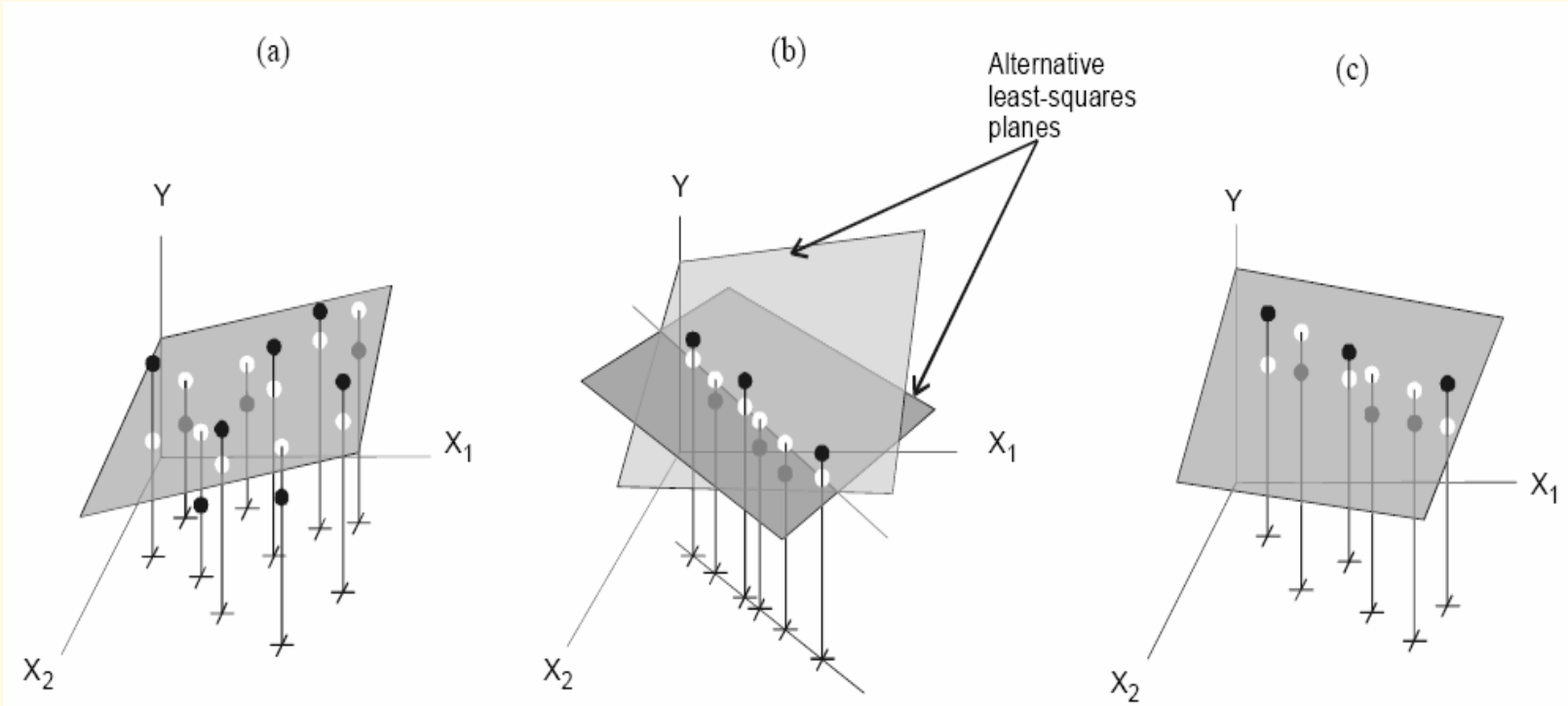

Figure 1:

(a) Low correlation between x_1 and x_2 — regression plane well supported.

(b) Perfect correlation between x_1 and x_2 — infinitely many least squares regression planes.

(c) High but not perfect correlation between x_1 and x_2 — regression plane not well supported; a small change in (x_1, x_2), or a deletion/addition of a vector may lead to a dramatic change in the parameter estimates.

Multicollinearity may happen without high correlation between one single pair of independent variables; it may be caused by that one independent variable can be expressed as a linear combination of all other independent variables.

In terms of matrices, it means that it is not necessarily two column vectors are linearly dependent (or nearly linearly dependent); the matrix is still singular (or close to singular) if one column can be expressed as (or almost as) a linear combination of other columns, and the latter is exactly the interpretation of a high R_i^2.

5. A *variance inflation factor* (VIF) for a β-parameter greater than 10, where

$$\mathrm{VIF}_i := \frac{1}{1 - R_i^2}, \qquad i = 1, 2, \ldots, k,$$

in which R_i^2 is the multiple coefficient of determination for the regression model:

$$x_i = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{i-1} x_{i-1} + \alpha_{i+1} x_{i+1} + \cdots + \alpha_k x_k + \varepsilon'.$$

By definition, a high R_i^2 means a high VIF_i, but why we transform R_i^2 to VIF_i?

One reason why the t-tests on the individual β-parameters suggest insignificance is that the standard errors of the least squares estimates $\hat{\beta}_i$ are inflated in the presence of multicollinearity. More precisely, when the response and the independent variables are standardised to have zero sample mean and unit sample variance, we have

$$s^2_{\hat{\beta}_i} = \frac{\hat{\sigma}^2}{n} \cdot \frac{1}{1 - R_i^2},$$

and hence the name "variance inflation factor."

Some software packages will report, equivalently, the *tolerance*, which is just the reciprocal of VIF. The rule of thumb is that the multicollinearity problem is severe when

$\boxed{R_i^2 > 0.9,}$ or equivalently $\boxed{\mathrm{VIF}_i > 10,}$ or equivalently $\boxed{\mathrm{Tolerance}_i < 0.1.}$

6. High *condition indices* associated with high *proportions of variance* in the eigensystem analysis of $\boldsymbol{X}'\boldsymbol{X}$. If there are one or more linear dependences in the data, then one or more eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ will be small. Recall from Linear Algebra that if $\boldsymbol{M}$ is a square matrix, the eigendecomposition is

$$\boldsymbol{M} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1},$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix containing eigenvalues and columns of $\boldsymbol{Q}$ are the corresponding eigenvectors. In particular, if $\boldsymbol{M}$ is symmetric, then all eigenvalues are real numbers. If $\boldsymbol{M}$ is nonsingular, then

$$\boldsymbol{M}^{-1} = \boldsymbol{Q}\boldsymbol{\Lambda}^{-1}\boldsymbol{Q}^{-1}.$$

Thus, if some eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ are small, the inverse $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ will be unstable and may suffer from serious rounding error.

We define

$$\text{condition index}_i := \sqrt{\frac{\lambda_{\max}}{|\lambda_i|}} = \frac{\sigma_{\max}}{\sigma_i},$$

where λ_i are eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ and $\lambda_{\max}$ is the maximum of the absolute values of eigenvalue, while σ_i are known as *singular values* of $\boldsymbol{X}$ and $\sigma_{\max}$ is the maximum singular value. (Not all matrices have real eigenvalues but all matrices have real, nonnegative singular values. That is the reason why we use singular values to define condition index, because the it can be defined for all matrices. For more details, check the topic "singular value decomposition", or "SVD" for short. Note that we borrow this concept from Linear Algebra, in which people would not restrict themselves to symmetric matrices. That is the reason why here even it is possible, we do not consider the ratios of eigenvalues but the ratios of singular values.)

The rule of thumb is that when a condition number (the largest conditional index, i.e. the ratio of the maximum singular value to the minimum singular value) is greater than 30, then we *may* have multicollinearity, but there is one further condition to check, namely, to check its *proportions of variance*, which tell us how influential that particular singular value is. To see its influence, we consider the fact (without proof) that

$$\mathrm{VIF}_j = \sum_{i=1}^{k+1} \frac{t_{ji}^2}{\lambda_i}, \qquad j = 1, \cdots, k+1$$

where t_{ji} is the j^{th} element in the eigenvector $\boldsymbol{t}_i$ corresponding to the eigenvalue λ_i. Thus, we have

$$\pi_{ij} = \frac{t_{ji}^2/\lambda_i}{\mathrm{VIF}_j} = \text{proportion of the variation of the } j^{\text{th}} \text{ VIF contributed by the } i^{\text{th}} \text{ eigenvalue.}$$

(Technical note: Consider the eigendecomposition of $\boldsymbol{X}'\boldsymbol{X}$ as $\boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^{-1}$, where $\boldsymbol{\Lambda}$ is a $(k+1)\times(k+1)$ diagonal matrix whose main diagonal elements λ_i are the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$, and column vectors $\boldsymbol{t}_i$ of $\boldsymbol{T}$ are eigenvectors. It is obvious that if $\boldsymbol{T}$ is multiplied by a scalar, then $\boldsymbol{T}^{-1}$ will be divided by the same scalar and so the decomposition remains valid, meaning that $\boldsymbol{T}$ is not unique. Here, we require that $\boldsymbol{T}$ is an orthogonal matrix, i.e. $\boldsymbol{T}' = \boldsymbol{T}^{-1}$, making it unique. We are allowed to do so because $\boldsymbol{X}'\boldsymbol{X}$ is symmetric. Hence, we are considering the unique decomposition

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}'$$

in this eigensystem analysis.) Now, the rule of thumb is:

> If, for some i, the condition index for the i^{th} eigenvalue is greater than 30 and two or more π_{ij} are greater than 0.5, then the multicollinearity is severe.

Remedies to multicollinearity

If we have the multicollinearity problem, what can we do? There are several solutions, described as follows.

1. Drop one or more correlated independent variables by stepwise-type regression.
2. If you are interested in estimation and prediction only, you may decide not to drop any of the independent variables. Confidence intervals for the mean estimation and individual response prediction generally remain unaffected as long as the values of the independent variables used to predict Y follow the same pattern of multicollinearity exhibited in the sample data, i.e. the values of independent variables are in the experimental region defined jointly by the values of observed $x_1, \ldots, x_k$.
3. Combine two or more independent variables into a single index.

4. Use ridge regression, which is explained in details below. The problem of multicollinearity is that the inverse $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is unstable. The *ridge regression* (also known as *Tikhonov regularisation* in mathematics) is a modification of least squares such that the ridge estimator $\hat{\boldsymbol{\beta}}^{\mathrm{R}}$ is the solution to

$$(\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*} + c\boldsymbol{I})\hat{\boldsymbol{\beta}}^{\mathrm{R}} = \boldsymbol{X}^{*\prime}\boldsymbol{Y}^{*}, \qquad c \geq 0,$$

where the independent variables and the response have been often centred and scaled by:

$$Y_j^{*} = \frac{1}{\sqrt{n-1}}\left(\frac{Y_j - \overline{Y}}{s_Y}\right), \quad x_{ij}^{*} = \frac{1}{\sqrt{n-1}}\left(\frac{x_{ij} - \overline{x}_i}{s_{x_i}}\right),$$

in which $s_Y = \sqrt{\sum_j (Y_j - \overline{Y})^2/(n-1)}$, $s_{x_i} = \sqrt{\sum_j (x_{ij} - \overline{x}_i)^2/(n-1)}$, so that $\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*}$ is the correlation matrix of $\boldsymbol{X}$, and $\boldsymbol{X}^{*\prime}\boldsymbol{Y}^{*}$ is the vector of correlation coefficients between Y and each x_i.

Note that the above transformation of the data leads to the *standardised model*. For ridge regression it is typical (e.g. in SAS) but not mathematically a must to consider the standardised model. (In SAS output the estimates are transformed back to become parameter estimates for the original model.)

When $c = 0$, it is the same as LSE.
When $c > 0$, $\hat{\boldsymbol{\beta}}^{\mathrm{R}}$ is biased, but let us consider the mean squared error (MSE): Denote $\boldsymbol{\mu}^{\mathrm{R}} = \mathbb{E}(\hat{\boldsymbol{\beta}}^{\mathrm{R}})$.

$$
\begin{aligned}
MSE(\hat{\boldsymbol{\beta}}^{\mathrm{R}}) &= \mathbb{E}\left\{(\hat{\boldsymbol{\beta}}^{\mathrm{R}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}^{\mathrm{R}} - \boldsymbol{\beta})\right\} \\
&= \mathbb{E}\left\{(\hat{\boldsymbol{\beta}}^{\mathrm{R}} - \boldsymbol{\mu}^{\mathrm{R}} + \boldsymbol{\mu}^{\mathrm{R}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}^{\mathrm{R}} - \boldsymbol{\mu}^{\mathrm{R}} + \boldsymbol{\mu}^{\mathrm{R}} - \boldsymbol{\beta})\right\} \\
&= \mathbb{E}\left\{(\hat{\boldsymbol{\beta}}^{\mathrm{R}} - \boldsymbol{\mu}^{\mathrm{R}})'(\hat{\boldsymbol{\beta}}^{\mathrm{R}} - \boldsymbol{\mu}^{\mathrm{R}})\right\} + (\boldsymbol{\mu}^{\mathrm{R}} - \boldsymbol{\beta})'(\boldsymbol{\mu}^{\mathrm{R}} - \boldsymbol{\beta}) \\
&\quad +2\mathbb{E}\left\{(\hat{\boldsymbol{\beta}}^{\mathrm{R}} - \boldsymbol{\mu}^{\mathrm{R}})'(\boldsymbol{\mu}^{\mathrm{R}} - \boldsymbol{\beta})\right\} \\
&= \sum \mathrm{var}(\hat{\beta}_i^{\mathrm{R}}) + \sum [\mathrm{bias}(\hat{\beta}_i)]^2 = \\
&= \sigma^2 \sum \frac{\lambda_i}{(\lambda_i + c)^2} + c^2 \boldsymbol{\beta}'(\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + c\boldsymbol{I})^{-2}\boldsymbol{\beta}.
\end{aligned}
$$

Thus, when c increases, the variance decreases but the bias increases.

If some λ_i is very small, the LSE $\hat{\boldsymbol{\beta}}$ (when $c = 0$) is unbiased but imprecise because of the term $1/\lambda_j$, whereas $\hat{\boldsymbol{\beta}}^{\mathrm{R}}$ is more precise but has a bias. In ridge regression, we would like to choose c such that the reduction in the variance is greater than the increase in the squared bias such that $MSE(\hat{\boldsymbol{\beta}}^{\mathrm{R}}) < MSE(\hat{\boldsymbol{\beta}})$.
Mathematics shows that there always exists some value $c > 0$ for which the inequality $MSE(\hat{\boldsymbol{\beta}}^{\mathrm{R}}) < MSE(\hat{\boldsymbol{\beta}})$ can really be achieved. The difficulty is that the optimal value of c varies from one application to another and hence is not a universal constant. And when we said "there exists", it means the proof of its existence is not by construction and hence the explicit expression (in terms of $\boldsymbol{X}$ and $\boldsymbol{Y}$) of the optimal c is unknown. Therefore, the determination of c will be a problem that we have to solve whenever we apply the ridge regression.

A commonly used method of determining c is based on the *ridge trace*, which is a plot of individual parameter estimates in $\hat{\boldsymbol{\beta}}^{\mathrm{R}}$ versus c for values of c usually in $[0, 1]$.
The estimates $\hat{\boldsymbol{\beta}}^{\mathrm{R}}$ may fluctuate widely as c changes slightly from 0, and some of them may even change signs. Gradually, however, these wide fluctuations cease and the magnitude of the regression coefficients tend to move slowly toward zero as c increases further.
At the same time VIF_i tend to fall rapidly as c changes from zero and become stable as c increases further.
We therefore examine the ridge trace and choose the smallest c where it is deemed that the parameter estimates first become stable and VIF values have become sufficiently small (remind you that a VIF value greater than 10 is an indicator of multicollinearity). Hopefully this will produce a set of estimates with smaller MSE than the MSE of the least squares estimators.

Example

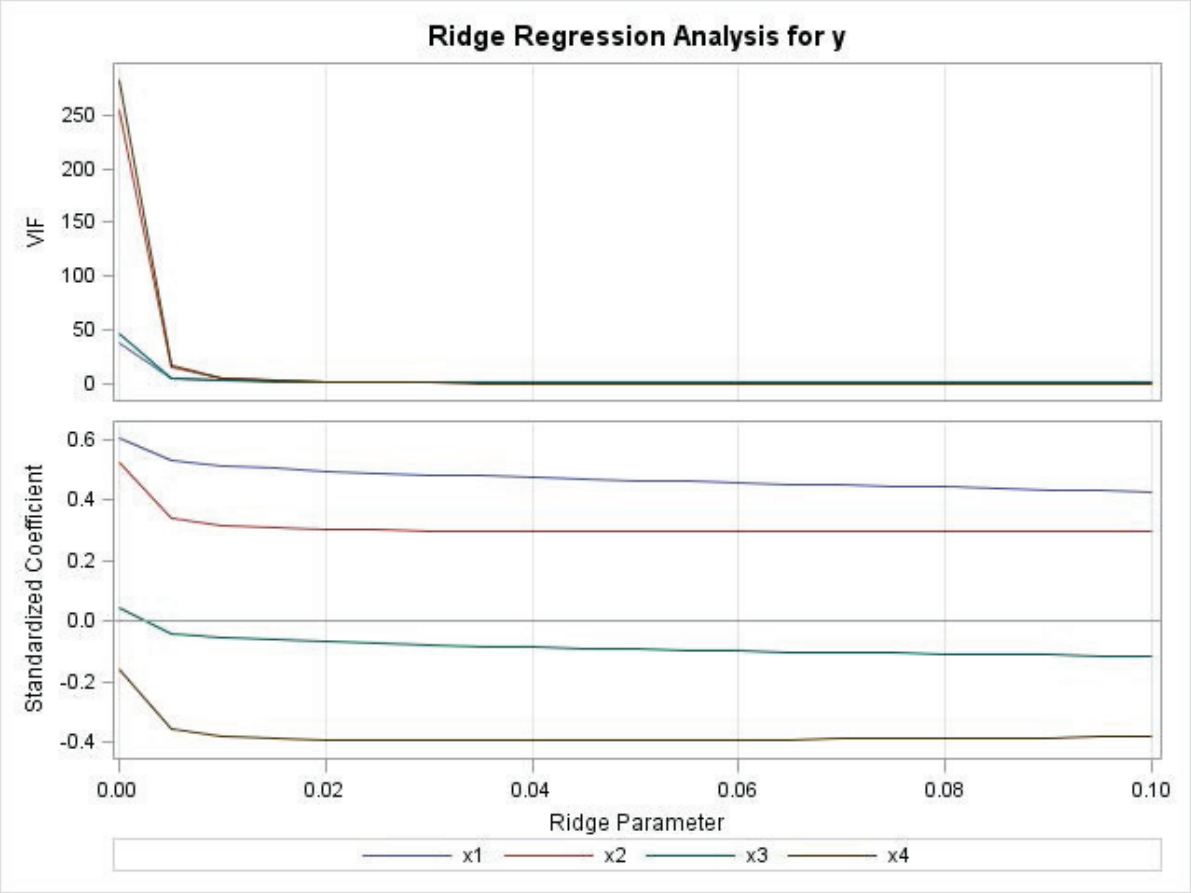

Figure: A plot of VIF vs ridge parameter and ridge trace

$c = 0.01$ is a good choice.

Two data-driven methods for the determination of c:

1. a fixed value
$$c = \frac{k \cdot \hat{\sigma}^2}{\hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\beta}}},$$
where k is the number of parameters, excluding the intercept, and $\hat{\boldsymbol{\beta}}$ the least squares estimators (in the above example this formula gives $c = 0.0131$);
2. an iterative procedure
$$c_0 = \frac{k \cdot \hat{\sigma}^2}{\hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\beta}}}, \text{ the fixed value as above,}$$
$$c_i = \frac{k \cdot \hat{\sigma}^2}{\hat{\boldsymbol{\beta}}^{\mathrm{R}}(c_{i-1})' \hat{\boldsymbol{\beta}}^{\mathrm{R}}(c_{i-1})}, \quad i \geq 1,$$
until the difference in $c_{i+1} - c_i$ is negligible, where $\hat{\boldsymbol{\beta}}^{\mathrm{R}}(c_{i-1})$ is the ridge estimator with $c = c_{i-1}$.

There is, however, **no guarantee that these methods are superior to the straightforward inspection of the ridge trace.**

A major limitation of ridge regression is that ordinary inference procedures are not applicable and exact distributional properties are not known.
Just for your interest: Ridge estimates can be obtained by the method of *penalised least squares* which, in this case, minimises

$$\sum_{j=1}^{n}\{Y_j^* - (\beta_1 x_{1j} + \cdots + \beta_k x_{kj})\}^2 + c\sum_{i=1}^{k}\beta_i^2,$$

Thus, for $c > 0$, the "best" coefficients generally will be smaller in absolute magnitude than the least squares because large absolute parameters lead to a large penalty. Because of this, the ridge estimators are examples of the *shrinkage estimators*.