

Lecture 19: Diagnostics: correlation matrix, $|R|$, partial regressions and VIFs

Econometrics ■ Unit II: Multicollinearity and heteroskedasticity

Kapil Kumar

Department of Statistics
Chaudhary Charan Singh University, Meerut

B.Sc. (Hons) Statistics, Semester V
Academic Session 2026–2027

Previous lecture

Lecture 18: OLS estimates too large on average; practical consequences; detection by $|X'X|$.

Today

- Tools for handling the problem: diagnostics
 - Inspection of the correlation matrix
 - Determinant of the correlation matrix
 - Measure based on partial regressions
 - Variance inflation factors

Large off-diagonal entries r_{ij} of the correlation matrix indicate near-dependence between pairs of variables. However, a dependence involving three or more variables need not produce any large pairwise correlation.

Example 3.1 (Pairwise correlations can miss a dependence)

Let w_1, w_2 be standardised variables with correlation 0, and let $w_3 = (w_1 + w_2)/\sqrt{2}$. Then

$$r_{13} = w'_1 w_3 = \frac{1}{\sqrt{2}}(w'_1 w_1 + w'_1 w_2) = \frac{1}{\sqrt{2}}(1 + 0) = 0.707,$$

and similarly $r_{23} = 0.707$, while $r_{12} = 0$. No correlation is close to 1, yet $w_1 + w_2 - \sqrt{2} w_3 = 0$ exactly.

Let $D = |X'X|$ with standardised data, so $X'X$ is the correlation matrix. Then

$$0 \leq D \leq 1.$$

Reason. The eigenvalues λ_j are ≥ 0 , so $D = \prod \lambda_j \geq 0$. Their sum is the trace, $\sum \lambda_j = \text{tr}(X'X) = k$ (all diagonal entries are 1). By the inequality between the geometric and arithmetic means,

$$D^{1/k} = \left(\prod \lambda_j \right)^{1/k} \leq \frac{1}{k} \sum \lambda_j = 1.$$

- $D = 0$: exact linear dependence.
- $D = 1$: the columns are orthonormal (no multicollinearity).
- D close to 0: high degree of multicollinearity.

For $k = 2$, $D = 1 - r_{12}^2$.

Advantages over $|X'X|$: it is bounded and not affected by the spread of the variables.

Limitation: it does not tell how many dependencies there are, nor which variables are involved.

Let R^2 be the coefficient of determination of the full model. Drop x_1 , regress y on the remaining $k - 1$ variables and record R_1^2 . Repeat for each variable to get $R_1^2, \dots, R_k^2$, and let

$$R_L^2 = \max(R_1^2, R_2^2, \dots, R_k^2).$$

- If multicollinearity is present, dropping one variable hardly reduces the fit, because the others carry the same information.
- So R_L^2 is close to R^2 , and $R^2 - R_L^2$ close to 0 indicates a high degree of multicollinearity.

Limitations. It does not tell how many relations exist or which variables are involved. A small $R^2 - R_L^2$ can also arise from a poorly specified model.

By Result 3.2, the j th diagonal element of $C = (X'X)^{-1}$ is $C_{jj} = 1/(1 - R_j^2)$. If x_j were orthogonal to all other regressors, $R_j^2 = 0$ and $C_{jj} = 1$. So C_{jj} is the factor by which $\text{Var}(b_j)$ is inflated by the near-dependence.

Definition 3.3

The **variance inflation factor** of the j th explanatory variable is

$$\text{VIF}_j = \frac{1}{1 - R_j^2}.$$

One or more large VIFs indicate multicollinearity. In practice $\text{VIF}_j > 5$ or $\text{VIF}_j > 10$ is taken to mean that β_j is poorly estimated.

The confidence interval for β_j is $b_j \pm \hat{\sigma} \sqrt{C_{jj}} t_{\alpha/2, n-k-1}$, with length

$$L_j = 2\hat{\sigma} \sqrt{C_{jj}} t_{\alpha/2, n-k-1}.$$

If the regressors were orthogonal ($X'X = I$, so $C_{jj} = 1$), with the same sample size and the same spread, the length would be $L^* = 2\hat{\sigma} t_{\alpha/2, n-k-1}$.
Dividing,

$$\frac{L_j}{L^*} = \sqrt{C_{jj}} = \sqrt{\text{VIF}_j}.$$

So $\sqrt{\text{VIF}_j}$ measures how much longer the confidence interval is because of multicollinearity.

Example 3.2

If $R_j^2 = 0.96$, then $VIF_j = 1/0.04 = 25$. The variance of b_j is 25 times, and its standard error and confidence-interval length $\sqrt{25} = 5$ times, what they would be with orthogonal regressors.

Limitations. VIFs do not show how many dependencies there are. The cut-off 5 or 10 is a rule of thumb.

Remember

$$VIF_j = \frac{1}{1 - R_j^2}, \quad \text{Var}(b_j) = \sigma^2 VIF_j \text{ (standardised data).}$$

Practice

- Exercises 3.3 and 3.5 (Exercises and Solutions booklet).

Next lecture

Lecture 20: Diagnostics: condition indices and variance-decomposition proportions.