Topic 9:
Assumption Violation:
Multicollinearity

Major Topics:
Terminology
Cases
Significance of Problem
Detection
Remedies

Fig. 9.1 Our Enhanced Roadmap This enhancement of our Roadmap shows that we are now checking the assumptions about the variance of the disturbance term. The focus in the chapter is the multicollinearity assumption: no exact linear relationship between two or more independent variables.
Multicollinearity
Introduction

- Recall that OLS estimators are
  \[ \hat{\beta} = (X'X)^{-1} X'Y \]
  - Key component: inverse of sum of squares and cross-products matrix, \((X'X)^{-1}\)
  - Matrix is a function of all the independent variables
    - Finding inverse is not trivial
    - Cannot find inverse if two or more columns (variables) or linearly related

Multicollinearity
Introduction (Continued)

- Suppose there are three variables: \(X_1, X_2, X_3\)
  - The variables are linearly related if
    \[ X_1 = \alpha_1 X_2 + \alpha_2 X_3 \]
  - Example:
    GDP = C + I + G + X
Multicollinearity
Terminology

• Interchangeable terms
  – Multicollinearity
  – Collinearity
  – Ill-conditioning
    • Primarily used by numerical analysts

Multicollinearity
Terminology (Continued)

• Orthogonal vs. Non-orthogonal Data
  – *Orthogonal* refers to lack of linear relationship between data
    • Usual case for designed experiments
  – *Non-orthogonal* data typical of economic data which are not collected by experiments
Multicollinearity

Cases

• Three cases can be distinguished
  – Perfect Multicollinearity
    • Perfect linear relationship among the variables
    • One or more variables are redundant
    • Holds for all observations in the dataset
    • Not really typical of economic data
    – Usually introduced into a problem by accident

Multicollinearity

Cases (Continued)

• Three cases can be distinguished (Continued)
  – Near-Perfect Multicollinearity
    • Almost perfect linear relationship
    • Typical of economic data
  – No Multicollinearity
    • No linear relationship at all
    • Not really typical of economic data
    – Some degree of multicollinearity is usually present in the data
Multicollinearity
Significance of Problem

• Perfect Multicollinearity
  – Cannot invert \((X'X)^{-1}\) matrix
    • Strictly a computational problem
    • Hence reference to numerical analysts
    • Economics *per se* not involved
    • Non-invertible matrix is called *Singular* or *Ill-conditioned*
  • Some packages give warning such as:
    “MATRIX IS SINGULAR”
Multicollinearity
Significance of Problem (Continued)

• Perfect Multicollinearity (Continued)
  – If inverse cannot be found, cannot find parameter estimates
  – Whole estimation process breaks down
    • Cannot find OLS or BLU Estimates

Multicollinearity
Significance of Problem (Continued)

• Near-perfect Multicollinearity
  – Can invert matrix and calculate estimates
  – Problem, however, is that estimates are not “stable”
    • inverted matrix is “inflated”
    • Results may be counterintuitive
      – Signs may change
      – Magnitudes may change
Multicollinearity
Significance of Problem (Continued)

• Near-perfect Multicollinearity (Continued)
  – More important problem
    • Variances of estimators “blow-up”
      – Recall that
        \[
        \sigma^2_{\hat{\beta}_j} = \sigma^2 \left( X' X \right)^{-1}
        \]
      – If \((X'X)^{-1}\) is inflated, then variances are inflated

Multicollinearity
Significance of Problem (Continued)

• Let \(VIF_j\) be variance inflation factor for parameter \(j, j = 1, 2, \ldots, p\)
  – Amount by which variance is inflated due to multicollinearity
  – Write variance for \(j\) as
    \[
    \sigma^2_{\hat{\beta}_j} = \sigma^2 f \left( (X'X)^{-1} \right) VIF_j
    \]
  • If \(VIF_j = 1\) then no inflation
    – No multicollinearity
  • Expect \(\text{VIF}_j \geq 1\)
Multicollinearity
Significance of Problem (Continued)

• Near-perfect Multicollinearity (Continued)
  – Variance inflation
    • If variances are inflated, then t-statistics are too small
      – Recall that
        \[ t_{c_j} = \hat{\beta}_j / \hat{s}_j \]
      • Will not reject null hypothesis more often than we should
        – Lead to believe coefficient is statistically zero when it is not
        – Make incorrect decisions

Multicollinearity
Detection

• Multicollinearity is not a **present/absent** problem
  – It is a matter of degree
    • **None** to **near to perfect**
      – Must check for the severity of multicollinearity, not presence or absence
        – Always assume it is present in economic data
Multicollinearity Detection (Continued)

- Look for a contradiction between the t-statistics and $R^2$
  - t-statistics deflated because of negatively biased $s^2$
    - Typical for most economic data
    - All t-statistics would be insignificant
  - $R^2$ inflated since it is a function of inflated parameter vector
    - $R^2$ would be too large

Multicollinearity Detection (Continued)

- Check Pearson correlation matrix
  - Correlations would be near 1 if linear relationship exists between \textit{two} variables
    - Reason: correlations are pair-wise
  - Problem
    - If more than two variables are involved in linear relationship, then correlations may be low despite high degree of multicollinearity
Multicollinearity Detection (Continued)

- Check Pearson correlation matrix (Continued)
  - Check correlation matrix
    - If correlations are high, then have multicollinearity between two variables
    - If correlations are low, then do other checks

Example
- Generated two independent random N(0,1) variables - $X_2$, $X_3$
- Generated $X_1 = 0.5X_2 + 0.2X_3$

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>0.58</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td>-0.17</td>
<td>-0.03</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Correlation Matrix

Shows No Linear Relationship
Multicollinearity Detection (Continued)

• Another possibility: look at linear relationships between each explanatory variable and all the others
  - Let $R^2_j$ be the squared multiple correlation between $x_j$ and all other variables in the model
  - That is: regress $X_j$ on the other variables and save $R^2$
  - Omit the constant

Multicollinearity Detection (Continued)

• Another possibility (Continued)
  - Define Tolerance
    \[ TOL_j = 1 - R^2_j \]
  - If no strong linear relationship exists, then $R^2_j = 0$ and $TOL_j = 1$
  - If a strong relationship exists, then $R^2_j = 1$ and $TOL_j = 0
Multicollinearity Detection (Continued)

• Another possibility (Continued)
  – Range For $TOL_j$

$$0 < TOL_j < 1$$

<table>
<thead>
<tr>
<th>Perfect Linear Relationship</th>
<th>No Linear Relationship</th>
</tr>
</thead>
</table>

– Can also define variance inflation factor

$$VIF_j = \frac{1}{TOL_j}$$

• No linear relationship implies that $VIF_j = 1$
  – Large values of $VIF_j$ indicate high degree of multicollinearity
Multicollinearity Detection (Continued)

• Another possibility (Continued)
  – Rule-of-thumb
    • Maximum value of $VIF_j$ greater than 10 indicates presence of high degree of multicollinearity

Multicollinearity Remedies

• Several possibilities
  – Drop one variable from model
    • One variable is redundant so get rid of it
  – Increase sample size
    • Multicollinearity is a data problem so correct by getting more data
    • May not be feasible
  – Transform variables
    • Logs or first differences
    • Change time series frequency
Multicollinearity Remedies (Continued)

• Several Possibilities (Continued)
  – Try another estimation technique
    • We do not discuss others but several are available

Multicollinearity Knowledge Checks

• What is Multicollinearity?
• What is a “linear relationship?”
• Describe the three degrees of multicollinearity?
• What is Tolerance?
• What is the Variance Inflation Factor (VIF) and how is it related to Tolerance?
• Describe some clues that multicollinearity is a problem?
• Is multicollinearity a prevalent problem in econometrics?