Multicollinearity

Gujarati(2003): Chapter 10
The nature of multicollinearity

In general:
When there are some functional relationships among independent variables, that is

\[ \sum \lambda_i X_i = 0 \]

or

\[ \lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 + \ldots + \lambda_i X_i = 0 \]

Such as

\[ 1X_1 + 2X_2 = 0 \quad \Rightarrow \quad X_1 = -2X_2 \]

If multicollinearity is perfect, the regression coefficients of the \( X_i \) variables, \( \beta_i \)'s, are indeterminate and their standard errors, \( Se(\beta_i) \)'s, are infinite.
Example: 3-variable Case:

\[ Y = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\mu} \]

If \( x_3 = \lambda x_2 \),

\[
\hat{\beta}_2 = \frac{(\Sigma y x_2)(\Sigma x_3^2) - (\Sigma y x_3)(\Sigma x_2 x_3)}{(\Sigma x_2^2)(\Sigma x_3^2) - (\Sigma x_2 x_3)^2}
\]

\[
\hat{\beta}_2 = \frac{(\Sigma y x_2)(\lambda^2 \Sigma x_2^2) - (\lambda \Sigma y x_2)(\lambda \Sigma x_2 x_2)}{(\Sigma x_2^2)(\lambda^2 \Sigma x_2^2) - \lambda^2(\Sigma x_2 x_2)^2}
\]

\[
\hat{\beta}_2 = \frac{0}{0}
\]

Indeterminate

Similarly, if \( x_3 = \lambda x_2 \),

\[
\hat{\beta}_3 = \frac{(\Sigma y x_3)(\Sigma x_2^2) - (\Sigma y x_2)(\Sigma x_2 x_3)}{(\Sigma x_2^2)(\Sigma x_3^2) - (\Sigma x_2 x_3)^2}
\]

\[
\hat{\beta}_3 = \frac{(\lambda \Sigma y x_2)(\Sigma x_2^2) - (\Sigma y x_2)(\lambda \Sigma x_2 x_2)}{(\Sigma x_2^2)(\lambda^2 \Sigma x_2^2) - \lambda^2(\Sigma x_2 x_2)^2}
\]

\[
\hat{\beta}_3 = \frac{0}{0}
\]

Indeterminate
If multicollinearity is imperfect,

\[ x_3 = \lambda_2 x_2 + \nu \]

where \( \nu \) is a stochastic error

(or \( x_3 = \lambda_1 + \lambda_2 x_2 + \nu \))

Then the regression coefficients, although determinate, possess large standard errors, which means the coefficients can be estimated but with less accuracy.

\[
\hat{\beta}_2 = \frac{(\Sigma y x_2)(\lambda_2^2 \Sigma x_2^2 + \Sigma \nu^2) - (\lambda_2 \Sigma y x_2 + \Sigma y \nu)(\lambda_2 \Sigma x_2 x_2 + \Sigma x_2 \nu)}{(\Sigma x_2^2)(\lambda_2^2 \Sigma x_2^2 + \Sigma \nu^2) - (\lambda_2 \Sigma x_2 x_2 + \Sigma x_2 \nu)^2}
\]

\[ \neq 0 \]

(Why?)
Perfect vs. Less Than Perfect Multicollinearity

• Perfect multicollinearity
\[ \lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_k X_k = 0 \]
\[ X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_k}{\lambda_2} X_{ki} \]

• Less than perfect multicollinearity
\[ \lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_2 X_k + \nu_i = 0 \]
\[ X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_k}{\lambda_2} X_{ki} - \frac{1}{\lambda_2} \nu_i \]
(a) No collinearity

(b) Low collinearity

(c) Moderate collinearity

(d) High collinearity

(e) Very high collinearity
Sources of Multicollinearity

1. *The data collection method employed*, for example, sampling over a limited range of the values taken by the regressors in the population.

2. *Constraints on the model or in the population being sampled*. For example, in the regression of electricity consumption on income ($X_2$) and house size ($X_3$) there is a physical constraint in the population in that families with higher incomes generally have larger homes than families with lower incomes.

3. *Model specification*, for example, adding polynomial terms to a regression model, especially when the range of the $X$ variable is small.

4. *An overdetermined model*. This happens when the model has more explanatory variables than the number of observations. This could happen in medical research where there may be a small number of patients about whom information is collected on a large number of variables.
It can be shown that even if multicollinearity is very high, as in the case of near multicollinearity, the OLS estimators still retain the property of BLUE. Multicollinearity violates no regression assumptions. Unbiased, consistent estimates will occur, and their standard errors will be correctly estimated. The only effect of multicollinearity is to make it hard to get coefficient estimates with small standard error.
MULTICOLLINEARITY: MUCH ADO ABOUT NOTHING?

- First, it is true that even in the case of near multicollinearity the OLS estimators are unbiased. But unbiasedness is a multisample or repeated sampling property. But this says nothing about the properties of estimators in any given sample.
- Second, it is also true that collinearity does not destroy the property of minimum variance: But this does not mean that the variance of an OLS estimator will necessarily be small.
- Third, *multicollinearity is essentially a sample (regression) phenomenon in* the sense that even if the $X$ variables are not linearly related in the population, they may be so related in the particular sample at hand: In short, our sample may not be “rich” enough to accommodate all $X$ variables in the analysis.
Consequences of imperfect multicollinearity

1. Although the estimated coefficients are BLUE, OLS estimators have large variances and covariances, making the estimation with less accuracy.
2. The estimation confidence intervals tend to be much wider, leading to accept the “zero null hypothesis” more readily.
3. The t-statistics of coefficients tend to be statistically insignificant.
4. The $R^2$ can be very high.
5. The OLS estimators and their standard errors can be sensitive to small change in the data.
Large variance and covariance of OLS estimators

\[ \text{var}(\hat{\beta}_2) = \frac{\sigma_u^2}{\sum x_2^2 (1 - r_{23}^2)} = \frac{\sigma_u^2}{\sum x_2^2} VIF \]

Variance-inflating factor:

\[ VIF = \frac{1}{1 - r_{23}^2} \]

Higher pair-wise correlation \( \Rightarrow \) higher VIF \( \Rightarrow \) larger variance

where

\[ r_{23}^2 \leq OLS : X_2 = \alpha_1 + \alpha_2 X_3 + \nu \]

\[ r_{32}^2 \leq OLS : X_3 = \alpha_1' + \alpha_2' X_2 + \nu' \]
Large variance and covariance of OLS estimators

\[
\text{var} (\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} \left( \frac{1}{1 - R_j^2} \right)
\]

\(\hat{\beta}_j\) = (estimated) partial regression coefficient of regressor \(X_j\)

\(R_j^2 = R^2\) in the regression of \(X_j\) on the remaining \((k - 2)\) regressions

[Note: There are \((k - 1)\) regressors in the \(k\)-variable regression model.]

The inverse of the VIF is called \textbf{tolerance (TOL)}:

\[
\text{TOL}_j = \frac{1}{\text{VIF}_j} = (1 - R_j^2)
\]

As a rule of thumb, if the VIF of a variable exceeds 10, which will happen if \(R_j^2\) exceeds 0.90, that variable is said be highly collinear.

VIF > 10 \(\Rightarrow\) TOL < 0.1
**THE EFFECT OF INCREASING \( r_{23} \) ON \( \text{VAR}(\hat{\beta}_2) \) AND \( \text{COV}(\hat{\beta}_2, \hat{\beta}_3) \)**

<table>
<thead>
<tr>
<th>Value of ( r_{23} )</th>
<th>VIF</th>
<th>( \frac{\sigma^2}{\sum x_{2i}^2} = A )</th>
<th>( \frac{\text{var}(\hat{\beta}<em>2)(r</em>{23} \neq 0)}{\text{var}(\hat{\beta}<em>2)(r</em>{23} = 0)} )</th>
<th>( \text{cov}(\hat{\beta}_2, \hat{\beta}_3) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>1.00</td>
<td>( \frac{\sigma^2}{\sum x_{2i}^2} = A )</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>0.50</td>
<td>1.33</td>
<td>1.33 ( \times A )</td>
<td>1.33</td>
<td>0.67 ( \times B )</td>
</tr>
<tr>
<td>0.70</td>
<td>1.96</td>
<td>1.96 ( \times A )</td>
<td>1.96</td>
<td>1.37 ( \times B )</td>
</tr>
<tr>
<td>0.80</td>
<td>2.78</td>
<td>2.78 ( \times A )</td>
<td>2.78</td>
<td>2.22 ( \times B )</td>
</tr>
<tr>
<td>0.90</td>
<td>5.76</td>
<td>5.26 ( \times A )</td>
<td>5.26</td>
<td>4.73 ( \times B )</td>
</tr>
<tr>
<td>0.95</td>
<td>10.26</td>
<td>10.26 ( \times A )</td>
<td>10.26</td>
<td>9.74 ( \times B )</td>
</tr>
<tr>
<td>0.97</td>
<td>16.92</td>
<td>16.92 ( \times A )</td>
<td>16.92</td>
<td>16.41 ( \times B )</td>
</tr>
<tr>
<td>0.99</td>
<td>50.25</td>
<td>50.25 ( \times A )</td>
<td>50.25</td>
<td>49.75 ( \times B )</td>
</tr>
<tr>
<td>0.995</td>
<td>100.00</td>
<td>100.00 ( \times A )</td>
<td>100.00</td>
<td>99.50 ( \times B )</td>
</tr>
<tr>
<td>0.999</td>
<td>500.00</td>
<td>500.00 ( \times A )</td>
<td>500.00</td>
<td>499.50 ( \times B )</td>
</tr>
</tbody>
</table>

*Note: \( A = \frac{\sigma^2}{\sum x_{2i}^2} \)

\( B = \frac{-\sigma^2}{\sqrt{\sum x_{2i}^2 \sum x_{3i}^3}} \)

\( \times = \text{times} \)
Detecting Multicollinearity

1. High $R^2$ but few significant t ratios.

\[
\hat{Y}_i = 24.7747 + 0.9415X_{2i} - 0.0424X_{3i}
\]

\[
(6.7525) \quad (0.8229) \quad (0.0807)
\]

\[
t = (3.6690) \quad (1.1442) \quad (-0.5261)
\]

\[
R^2 = 0.9635 \quad \bar{R}^2 = 0.9531 \quad \text{df} = 7
\]
Detecting Multicollinearity

1. High $R^2$ but few significant t ratios.

**TABLE 10.3**
HYPOTHETICAL DATA ON $Y$, $X_2$, AND $X_3$

<table>
<thead>
<tr>
<th>$Y$</th>
<th>$X_2$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>16</td>
</tr>
</tbody>
</table>

**TABLE 10.4**
HYPOTHETICAL DATA ON $Y$, $X_2$, AND $X_3$

<table>
<thead>
<tr>
<th>$Y$</th>
<th>$X_2$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>16</td>
</tr>
</tbody>
</table>
Detecting Multicollinearity

2. High pair-wise correlations among regressors.

\[ Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \epsilon_i \]

**Suppose:** \[ X_{4i} = \lambda_2 X_{2i} + \lambda_3 X_{3i} \]

where \( \lambda_2 \) and \( \lambda_3 \) are constants, not both zero. Obviously, \( X_4 \) is an exact linear combination of \( X_2 \) and \( X_3 \), giving \( R_{24.23} = 1 \), the coefficient of determination in the regression of \( X_4 \) on \( X_2 \) and \( X_3 \).

\[
R^2_{4,23} = \frac{r_{42}^2 + r_{43}^2 - 2r_{42}r_{43}r_{23}}{1 - r_{23}^2}
\]

Since \( R_{24.23} = 1 \), we get

\[
1 = \frac{r_{42}^2 + r_{43}^2 - 2r_{42}r_{43}r_{23}}{1 - r_{23}^2}
\]

This is satisfied by \( r_{42} = 0.5 \), \( r_{43} = 0.5 \), and \( r_{23} = -0.5 \), which are not very high values.
Detecting Multicollinearity

3. Examination of partial correlations.

In the regression of Y on $X_2$, $X_3$, and $X_4$, a finding that $R^2_{1.234}$ is very high but $r^2_{12.34}$, $r^2_{13.24}$, and $r^2_{14.23}$ are comparatively low may suggest that the variables $X_2$, $X_3$, and $X_4$ are highly intercorrelated and that at least one of these variables is superfluous.

Although a study of the partial correlations may be useful, there is no guarantee that they will provide an infallible guide to multicollinearity, for it may happen that both $R^2$ and all the partial correlations are sufficiently high.
Detecting Multicollinearity

4. Auxiliary regressions.

Regress each $X_i$ on the remaining $X$ variables and compute the corresponding $R^2$, which we designate as $R_{2i}^2$; each one of these regressions is called an auxiliary regression.

Model: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$

Auxiliary Regressions:

$X_{2i} = \alpha_1 + \alpha_3 X_{3i} + \alpha_4 X_{4i} + u_{2i} \Rightarrow R_{2i}^2$

$X_{3i} = \gamma_1 + \gamma_2 X_{2i} + \gamma_4 X_{4i} + u_{3i} \Rightarrow R_{3i}^2$

$X_{4i} = \phi_1 + \phi_2 X_{2i} + \phi_3 X_{3i} + u_{3i} \Rightarrow R_{4i}^2$

Check the overall $F$, $F_i$, $i=2,3,4$, in each of these auxiliary regressions. If the computed $F$ exceeds the critical $F_i$ at the chosen level of significance, it is taken to mean that the particular $X_i$ is collinear with other $X$'s.

Klein's rule of thumb: multicollinearity may be a troublesome problem only if the $R^2$ obtained from an auxiliary regression is greater than the overall $R^2$. 
Detecting Multicollinearity

5. Eigenvalues and condition index.

\[ \Xi = (X'X) \] a \((m \times m)\) symmetric positive (semi)definite matrix

Eigenvalue of \(\Xi\) a resolution to

\[ |\Xi - \lambda I| = 0 \] solution gives \(m\) eigenvalues: \(\lambda_1, \lambda_1, \ldots, \lambda_m\)

\[ \Xi = \begin{bmatrix} 20 & 5 \\ 5 & 10 \end{bmatrix}, \quad \Xi - \lambda I = \begin{bmatrix} 20 - \lambda & 5 \\ 5 & 10 - \lambda \end{bmatrix} \]

\[ |\Xi - \lambda I| = (20 - \lambda)(10 - \lambda) - (5)(5) = \lambda^2 - 30\lambda + 175 = 0 \]

\[ \Rightarrow \lambda_1 = 15 + 5\sqrt{2}, \quad \lambda_2 = 15 - 5\sqrt{2} \]
Detecting Multicollinearity

5. Eigenvalues and condition index.

**condition number**

$k$ defined as

**condition index (CI) defined as**

$$k = \frac{\text{Maximum eigenvalue}}{\text{Minimum eigenvalue}}$$

$$\text{CI} = \sqrt{\frac{\text{Maximum eigenvalue}}{\text{Minimum eigenvalue}}} = \sqrt{k}$$

**Rule of thumb:**

1. $100 < k < 1000$ ($10 < \text{CI} < 30$) ➔ moderate to strong multicollinearity
2. $k > 1000$ ($\text{CI} > 30$) ➔ severe multicollinearity
Detecting Multicollinearity

6. Tolerance and variance inflation factor

Rule of thumb:

If the VIF of a variable exceeds 10, which will happen if $R^2_j$ exceeds 0.90, that variable is said be highly

or

The closer is TOL$_j$ to zero, the greater the degree of collinearity of that variable with the other regressors. On the other hand, the closer TOL$_j$ is to 1, the greater the evidence that $X_j$ is not collinear with the other regressors.
6. Tolerance and variance inflation factor

VIF is neither necessary nor sufficient to get high variances and high standard errors. Therefore, high multicollinearity, as measured by a high VIF, may not necessarily cause high standard errors. In all this discussion, the terms **high** and **low** are used in a relative sense.
Remedial Measures

1. Utilise a priori information

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \]
\[ = \beta_1 + \beta_2 X_2 + 0.1\beta_2 X_3 + u \]

*given* \( \beta_3 = 0.1\beta_2 \)

\[ = \beta_1 + \beta_2 (X_2 + 0.1X_3) + u \]
\[ = \beta_1 + \beta_2 Z + u \]

2. Combining cross-sectional and time-series data

3. Dropping a variable(s) and re-specify the regression

4. Transformation of variables:
   (i) First-difference form
   \[ \Delta Y = \beta_1 + \beta_2 \Delta X_2 + \beta_3 \Delta X_3 + u' \]
   \[ \frac{Y}{X_3} = \beta_1 \left( \frac{1}{X_3} \right) + \beta_2 \left( \frac{X_2}{X_3} \right) + \beta_3 + u' \]
   (ii) Ratio transformation

5. Additional or new data

6. Reducing collinearity in polynomial regression

\[ Y = \beta_1 + \beta_2 X_2^2 + \beta_3 X_3 + u' \]
\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3^2 + u' \]

7. Do nothing (if the objective is only for prediction)

M. Balçlar, EMU, ECON 530
Remedial Measures

7. Other remedies

Factor analysis and principal components
Collinear variables are reduced into a single variable (a combination of all collinear variables)

Ridge regression
A solution in case of perfect multicollinearity. We accept a (small) bias in the estimates
Since GDP and GNP are highly related

Other examples:  CPI <=> WPI;
CD rate <=> TB rate
M2 <=> M3
Example: Longley Data

Time series for the years 1947–1962:

\[ Y = \text{number of people employed, in thousands}; \]
\[ X_1 = \text{GNP implicit price deflator}; \]
\[ X_2 = \text{GNP, millions of dollars}; \]
\[ X_3 = \text{number of people unemployed in thousands}, \]
\[ X_4 = \text{number of people in the armed forces}, \]
\[ X_5 = \text{noninstitutionalized population over 14 years of age}; \]
\[ X_6 = \text{year, equal to 1 in 1947, 2 in 1948, and 16 in 1962}. \]
## Example: Longley Data

### Longley Data

<table>
<thead>
<tr>
<th>Observation</th>
<th>$y$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1947</td>
<td>60,323</td>
<td>830</td>
<td>234,289</td>
<td>2356</td>
<td>1590</td>
<td>107,608</td>
<td>1</td>
</tr>
<tr>
<td>1948</td>
<td>61,122</td>
<td>885</td>
<td>259,426</td>
<td>2325</td>
<td>1456</td>
<td>108,632</td>
<td>2</td>
</tr>
<tr>
<td>1949</td>
<td>60,171</td>
<td>882</td>
<td>258,054</td>
<td>3682</td>
<td>1616</td>
<td>109,773</td>
<td>3</td>
</tr>
<tr>
<td>1950</td>
<td>61,187</td>
<td>895</td>
<td>284,599</td>
<td>3351</td>
<td>1650</td>
<td>110,929</td>
<td>4</td>
</tr>
<tr>
<td>1951</td>
<td>63,221</td>
<td>962</td>
<td>328,975</td>
<td>2099</td>
<td>3099</td>
<td>112,075</td>
<td>5</td>
</tr>
<tr>
<td>1952</td>
<td>63,639</td>
<td>981</td>
<td>346,999</td>
<td>1932</td>
<td>3594</td>
<td>113,270</td>
<td>6</td>
</tr>
<tr>
<td>1953</td>
<td>64,989</td>
<td>990</td>
<td>365,385</td>
<td>1870</td>
<td>3547</td>
<td>115,094</td>
<td>7</td>
</tr>
<tr>
<td>1954</td>
<td>63,761</td>
<td>1000</td>
<td>363,112</td>
<td>3578</td>
<td>3350</td>
<td>116,219</td>
<td>8</td>
</tr>
<tr>
<td>1955</td>
<td>66,019</td>
<td>1012</td>
<td>397,469</td>
<td>2904</td>
<td>3048</td>
<td>117,388</td>
<td>9</td>
</tr>
<tr>
<td>1956</td>
<td>67,857</td>
<td>1046</td>
<td>419,180</td>
<td>2822</td>
<td>2857</td>
<td>118,734</td>
<td>10</td>
</tr>
<tr>
<td>1957</td>
<td>68,169</td>
<td>1084</td>
<td>442,769</td>
<td>2936</td>
<td>2798</td>
<td>120,445</td>
<td>11</td>
</tr>
<tr>
<td>1958</td>
<td>66,513</td>
<td>1108</td>
<td>444,546</td>
<td>4681</td>
<td>2637</td>
<td>121,950</td>
<td>12</td>
</tr>
<tr>
<td>1959</td>
<td>68,655</td>
<td>1126</td>
<td>482,704</td>
<td>3813</td>
<td>2552</td>
<td>123,366</td>
<td>13</td>
</tr>
<tr>
<td>1960</td>
<td>69,564</td>
<td>1142</td>
<td>502,601</td>
<td>3931</td>
<td>2514</td>
<td>125,368</td>
<td>14</td>
</tr>
<tr>
<td>1961</td>
<td>69,331</td>
<td>1157</td>
<td>518,173</td>
<td>4806</td>
<td>2572</td>
<td>127,852</td>
<td>15</td>
</tr>
<tr>
<td>1962</td>
<td>70,551</td>
<td>1169</td>
<td>554,894</td>
<td>4007</td>
<td>2827</td>
<td>130,081</td>
<td>16</td>
</tr>
</tbody>
</table>

M. Balçlar, EMU, ECON 503
# Example: Longley Data

**Dependent Variable:** Y  
**Method:** Least Squares  
**Date:** 12/18/07  **Time:** 13:03  
**Sample:** 1947-1962  
**Included observations:** 16

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>77270.12</td>
<td>22506.71</td>
<td>3.433204</td>
<td>0.0075</td>
</tr>
<tr>
<td>X1</td>
<td>15.06187</td>
<td>84.91493</td>
<td>0.177376</td>
<td>0.8631</td>
</tr>
<tr>
<td>X2</td>
<td>-0.035819</td>
<td>0.033491</td>
<td>-1.069516</td>
<td>0.3127</td>
</tr>
<tr>
<td>X3</td>
<td>-2.020230</td>
<td>0.488400</td>
<td>-4.136427</td>
<td>0.0025</td>
</tr>
<tr>
<td>X4</td>
<td>-1.033227</td>
<td>0.214274</td>
<td>-4.821985</td>
<td>0.0009</td>
</tr>
<tr>
<td>X5</td>
<td>-0.051104</td>
<td>0.226073</td>
<td>-0.226051</td>
<td>0.8262</td>
</tr>
<tr>
<td>X6</td>
<td>1829.151</td>
<td>455.4785</td>
<td>4.015890</td>
<td>0.0030</td>
</tr>
</tbody>
</table>

- **R-squared:** 0.995479  
- **Mean dependent var:** 65317.00  
- **S.D. dependent var:** 3511.968  
- **Akaike info criterion:** 14.57718  
- **Schwarz criterion:** 14.91519  
- **F-statistic:** 330.2853  
- **Prob(F-statistic):** 0.000000
Example: Longley Data
### Example: Longley Data

**Dependent Variable:** $X_1$  
**Method:** Least Squares  
**Date:** 12/18/07  
**Time:** 13:06  
**Sample:** 1947 1962  
**Included observations:** 16

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>204.4583</td>
<td>53.33698</td>
<td>3.833331</td>
<td>0.0033</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.000256</td>
<td>9.48E-05</td>
<td>2.700628</td>
<td>0.0223</td>
</tr>
<tr>
<td>$X_3$</td>
<td>0.003192</td>
<td>0.001513</td>
<td>2.109831</td>
<td>0.0611</td>
</tr>
<tr>
<td>$X_4$</td>
<td>0.000880</td>
<td>0.000748</td>
<td>1.176973</td>
<td>0.2665</td>
</tr>
<tr>
<td>$X_5$</td>
<td>-0.001755</td>
<td>0.000633</td>
<td>-2.771998</td>
<td>0.0197</td>
</tr>
<tr>
<td>$X_6$</td>
<td>-0.999219</td>
<td>1.666535</td>
<td>-0.599579</td>
<td>0.5621</td>
</tr>
</tbody>
</table>

| R-squared | 0.992622 | Mean dependent var | 101.6813 |
| Adjusted R-squared | 0.988933 | S.D. dependent var | 10.79155 |
| S.E. of regression | 1.135293 | Akaike info criterion | 3.371655 |
| Sum squared resid | 12.88890 | Schwarz criterion | 3.661376 |
| Log likelihood | -20.97324 | F-statistic | 269.0649 |
| Durbin-Watson stat | 1.870344 | Prob(F-statistic) | 0.000000 |
### Example: Longley Data

#### $R^2$ VALUES FROM THE AUXILIARY REGRESSIONS

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>$R^2$ value</th>
<th>Tolerance (TOL) = $1 - R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>0.9926</td>
<td>0.0074</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.9994</td>
<td>0.0006</td>
</tr>
<tr>
<td>$X_3$</td>
<td>0.9702</td>
<td>0.0298</td>
</tr>
<tr>
<td>$X_4$</td>
<td>0.7213</td>
<td>0.2787</td>
</tr>
<tr>
<td>$X_5$</td>
<td>0.9970</td>
<td>0.0030</td>
</tr>
<tr>
<td>$X_6$</td>
<td>0.9986</td>
<td>0.0014</td>
</tr>
</tbody>
</table>

M. Balciğlar, EMU, ECON 503
Example: Longley Data

Remedial Actions:

1. Express GNP not in nominal terms, but in real terms, \( \frac{X2}{X3} \)
2. Noninstitutional population over 14 years of age \( X5 \) grows over time because of natural population growth, it will be highly correlated with time, the variable \( X6 \) in our model. Therefore, instead of keeping both these variables, we will keep the variable \( X5 \) and drop \( X6 \).
3. There is no compelling reason to include \( X3 \), the number of people unemployed
Example: Longley Data

Dependent Variable: Y  
Method: Least Squares  
Date: 12/18/07  Time: 13:11  
Sample: 1947 1962  
Included observations: 16

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>65720.37</td>
<td>10624.81</td>
<td>6.185558</td>
<td>0.0000</td>
</tr>
<tr>
<td>X2/X1</td>
<td>9.736496</td>
<td>1.791552</td>
<td>5.434671</td>
<td>0.0002</td>
</tr>
<tr>
<td>X4</td>
<td>-0.687966</td>
<td>0.322238</td>
<td>-2.134965</td>
<td>0.0541</td>
</tr>
<tr>
<td>X5</td>
<td>-0.299537</td>
<td>0.141761</td>
<td>-2.112965</td>
<td>0.0562</td>
</tr>
</tbody>
</table>

R-squared 0.981404  Mean dependent var 65317.00  
Adjusted R-squared 0.976755  S.D. dependent var 3511.968  
S.E. of regression 535.4492  Akaike info criterion 15.61641  
Sum squared resid 3440470.  Schwarz criterion 15.80955  
Log likelihood -120.9313  F-statistic 211.0972  
Durbin-Watson stat 1.654069  Prob(F-statistic) 0.000000  

M. Balçtlar, EMU, ECON 503