Chapter 6

Statistical Package Dealing with MC

At present time, there exist more Statistical packages that can do all different Statistical analysis like SAS, BMDP, MINITAB, STYSTAT and STATA. In all these packages, the skill for making a program in the computer is not required because these packages depend on the sentences and commands which use to make any statistical analysis. Statistical packages are useful because they help for reducing the time and possibly eliminate mistakes in a statistical data analysis. When we surveyed some popular and commonly used statistical packages, we could not find any package that covers the recent methods developed to handle a specific problem like MC. Also for some users the statistical packages not available but the microsoft office which is working under the any version of microsoft windows is available.

In this chapter, we describe a new software package (Multicollinearity) that implements the recent methods on MC. Multicollinearity uses MS Excel as the front-end and MATLAB as the back-end. That is, the user interface is through MS Excel, where a menu item is added to the menu bar of MS Excel. Each command is communicated to MATLAB for execution in background and the answer is returned to the MS Excel worksheet, so that the user obtains the answer in MS Excel. This is how MATLAB is the back-end component of the software package. Multicollinearity will not work if MATLAB is not installed on the system.
Multicollinearity implements different methods like biased estimation, adding linear restriction to the linear regression model, making selection of variables or diagnostics when the CMLRM has MC. Also, any person having programming skills can modify it to add new methods related to MC or other problems of interest.

6.1 Excel-Matlab interaction

Multicollinearity.xll is an Excel add-in that acts as an interface between MS Excel and Matlab. This DLL is created using Visual C++ and uses Excel C API and Matlab C++ library to effectively allow Excel to act as a front-end for Matlab.

The only pre-requisite for using this add-in is that both MS Excel and Matlab 6.5 or later should be installed.

The following figure shows how the interaction between Excel and Matlab is achieved.

![Diagram showing the interaction between Excel and Matlab](image)

Figure 6.1: The interaction between Excel and Matlab
6.2 Installation of the package

The Excel software gives opportunity to add any item to the menu bar by using the command `add-in` available in the `Tools` command in the menu bar. We use this facility to add the `Multicollinearity` file. We do this by running the `add-in` as follows (see also Figure 6.2).

1. Copy `Multicollinearity` folder to some location on the hard disk (say C:).
2. Open excel.
3. Click on `Tools` $\rightarrow$ `Add-Ins`.
4. Click the `Browse` button.
5. Navigate to the folder created in step 1 (C:/multicollinearity if copied to C:).
7. The `add-in` is now ready to use.

Remark: As mentioned in the beginning, `Matlab` must be installed before adding multicollinearity to the menu bar of Excel. According to Figure 6.2, the new item is ready for executing the analysis of the interred data. The system is used as follows:

1. Select the add-in in Excel by going to `Tools` $\rightarrow$ `Add-Ins`.
2. Enter data in Excel.
3. Select a command from the `Multicollinearity` menu.
4. Select cell-ranges for variables and results (and in some cases, provide values).
5. The values in the specified cell-ranges are converted to Matlab variables by `multicollinearity.xll`.

142
6. Multicollinearity.xll invokes the appropriate Matlab script(s).

7. Results are retrieved from Matlab by multicollinearity.xll and displayed in the cells specified for results.

6.3 What is inside Multicollinearity?

In order to analyze any data that suffers from MC, Multicollinearity covers MC in the CMLR model as much as possible. Therefore, Multicollinearity contains the following items.

6.3.1 Diagnostics

In this item, methods are available for diagnosing the existence of MC in data and also for measuring the degree of MC. Most of these diagnostics are given in Chapter 1. The Diagnostics item has the following sub-items (see Figure 6.4).
Correlation matrix

We can decide whether the data suffers from MC from the off-diagonal elements of the correlation matrix. If the off-diagonal elements have large values, there is MC. If the off-diagonal elements are small in magnitude, it cannot be said that there is no MC, because some independent variable may be a linear combination of two or more other independent variables.

Variance Inflation Factor

See Chapter 1.

Determinant of the Correlation Matrix

Since the values of the correlation matrix $X'X$ belong to the closed interval $[0,1]$, $0 \leq |X'X| \leq 1$. If $|X'X| \approx 0$, then the independent variables suffer from MC and this MC will increase as $|X'X|$ approaches zero.

Condition Number

See Chapter 1.
Direct Effect

See Chapter 1.

Figure 6.4: The contains of the Diagnostics list

Executing Items in Diagnostics

As an example, suppose we want to execute the correlation matrix item. After pressing correlation matrix item, a new window will appear for the user to specify the data-range for independent variables. The second row result requires the position of the output. See Figure 6.5.
6.4 Fitting of Model

In this item, we fit the CMLR model under MC using different biased estimators. The menu item Fitting of Model shows six sub-items to select from. For every sub-item, the model has two options: with intercept (None) and without intercept (Standard). For example, see Figure 6.7 that shows the ordinary ridge regression window.

The window of this item contains the following description (see Figure 6.8)
Figure 6.6: The fit of model item and its contents

- **dependent variable**: We inter the dependent variable of the data.

- **independent variable**: We inter the independent variables of the data.

- **save**: This item gives an extra output with the fitted model, like MSE.

- **Type of Biased Parameter**: Here, we give an option for the type of biased parameter, like HKB (Eq.(2.19)).

**Remark**: Other items in *Fitting of Model* have a similar explanation.

### 6.5 Selection of Variables

In this item, different procedures for selection of variables in presence of MC are given.
6.5.1 Nested Estimate

See Chapter 5.

6.5.2 Ridge Regression

This item specifies ORR estimator for $\beta$ and then significant independent variables ($p$-value < 0.05) are selected.

6.5.3 Lasso

See Chapter 5. $k$ is the tuning parameter.

6.5.4 Elastic Net

The elastic net penalty (Zou and Hastie 2005) is a compromise of Lasso and ORR, and has the form

$$\sum_{i=1}^{p} (\alpha |\beta_i| + (1 - \alpha)\beta_i^2).$$  \hspace{1cm} (6.1)
Figure 6.8: The ordinary ridge regression window after inter the informations

The second term averages highly correlated features, while the first term obtains a sparse solution in the coefficients of these averaged features. The elastic net penalty can be used with any linear model, in particular for regression or classification (see Friedman et al., 2008).

$k_1$ and $k_2$ are the tuning parameters, and when $k_1 = 0$ we get Lasso solution.

Remark: The program in lasso and Elastic net are taken from the following web site www.cs.ubc.ca/ murphyk/Software/DAGlearn/lars.m

6.5.5 VIF procedure

See Chapter 5.
6.6 Restricted Model

In this item, Multicollinearity gives different estimators when linear restrictions are imposed on the true parameters $\beta$. We assume that the linear restrictions are satisfied (see Chapter 4). We have the following sub-items.

1. Restricted OLS : Chapter 4 Eq. 4.3.

2. Restricted Ridge Regression 1 : Chapter 4 Eq. 4.7.

3. Restricted Ridge regression 2 : Chapter 4 Eq. 4.5.

4. Restricted Liu estimator : Chapter 4 Eq. 4.6.

5. Restricted (k-d) estimator : Chapter 4 Eq. 4.11.

When a sub-item is selected in Restricted Model, the input window requires specification of Dependent variable, independent variables, $R$ matrix, $r$ vector, etc.

![Figure 6.9: Restricted(k – d) estimator window](image)

For the sub-item Restricted (k-d) estimator, the two biasing parameters Biasing parameter 1 and Biasing parameter 2 are $k$ and $d$ respectively.
6.7 The software package CD

The software package described in this chapter is provided on the accompanying CD. The CD also contains matlab version 6.5.1. The MS Excel is not provided on the CD.

The complete listing of the VC++ source code of the program is provided in appendix B if the executable provided in the CD is not compatible with the operating system or MS Excel version, the user may compile the source code in their system.