Chapter 5

Other Methods for Handling MC

In this chapter, we introduce different methods for handling MC. In Section 1, we introduce a new Liu-type estimator which covers ordinary least squares estimator, ordinary ridge regression estimator, Liu estimator, \((k - d)\) class estimator, principal components regression (PCR) estimator, \((r - d)\) class estimator and \((r - k)\) class estimator as special cases. The proposed estimator is superior by scalar mean squared error criterion under some conditions. A numerical example illustrates theoretical results. In Section 2, two procedures are introduced to remove MC: one when variables can be selected and the other when all variables must be included.

The relation between the correlation matrix \((R)\) and VIF is used for detecting MC (see Chapter 1) and for selecting variables. The first procedure does not fit regression of one explanatory variable on the others for computing the VIF. The second procedure replaces an explanatory variable by the residual from its regression on other explanatory variables. As a result, the final model is without MC. The performance of these procedures is compared with some recently developed methods. Three examples illustrate these procedures.
5.1 Principal components regression

The principal components regression (PCR) reduces the effect of MC. The concept of PCR can be interpreted in two different ways: first as a biased estimator of regression coefficients (see Montgomery, 2003) and second as a special case of RLS (see Groß, 2003). PCR uses eigenvalues and eigenvectors for providing specific information on the nature of MC and, at the same time, reduces the effect of MC by selecting fewer principal components (PC) in the model.

5.1.1 PCR as a biased estimator

In the canonical form of CMLR, \( Y = Z\alpha + \epsilon, \Lambda = \text{diag}\{\lambda_1, \cdots, \lambda_p\}\) is the diagonal matrix of eigenvalues of \( X'X \), \( T \) is the matrix whose columns are eigenvectors associated with \( \lambda_1, \cdots, \lambda_p \), \( \alpha = T'\beta \) and \( Z = XT \) is the \( n \times p \) matrix of principal components (PC). The OLS estimator of \( \alpha \) is

\[
\hat{\alpha} = (Z'Z)^{-1}Z'Y = (T'X'XT)^{-1}Z'Y = \Lambda^{-1}Z'Y.
\]

Since \( \text{Var}(\hat{\alpha}) = \sigma^2\Lambda^{-1} \), and \( \lambda_i \) is the variance of the \( i \)th PC, PCR has a possibility of removing impreciseness when \( \lambda_i \approx 0 \).

We find the PCR estimator by assuming that the independent variables are arranged in order of decreasing eigenvalues, \( \lambda_1 \geq \lambda_2 \cdots \geq \lambda_p > 0 \). Suppose \( p - s \) eigenvalues are approximately zero. Then, the PCR estimator is obtained by removing \( p - s \) PC’s and applying the least squares method to the remaining PC’s

\[
\hat{\alpha}_r = \sum_{i=1}^{s} \lambda_i^{-1}Z'Y. \tag{5.1}
\]

5.1.2 PCR as restricted least squares estimator

PCR can be considered as a special case of RLS as follows.
Using spectral decomposition, $X'X$ can be written as
\[
X'X = (U_1, U_2) \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix} \begin{pmatrix} U_1' \\ U_2' \end{pmatrix}, \tag{5.2}
\]
where $\Lambda_1 = \text{diag}\{\lambda_1, \cdots, \lambda_s\}$ and $\Lambda_2 = \text{diag}\{\lambda_{s+1}, \cdots, \lambda_p\}$ are $s \times s$ and $(p - s \times (p - s))$ diagonal matrices such that elements of $\Lambda_1$ are the $s$ largest eigenvalues of $X'X$ and elements of $\Lambda_{p-s}$ are the remaining eigenvalues of $X'X$. The matrix $T = (T_1, T_{p-s})$ is an orthogonal matrix with $T_1$ containing the first $s$ columns and $T_2$ containing the remaining $p - s$ columns of $T$. According to Groß(2003), “The $p$ columns of $TA = (A_1^{t_1}, \cdots, A_p^{t_p})$ generate the whole $\mathbb{R}^p$. Since fairly small eigenvalues of $X'X$ are responsible for the impreciseness of the least squares estimator, it seems to be a good strategy to prevent the estimate to move in directions $\lambda_i t_i$ with corresponding fairly small $\lambda_i$. One such strategy is offered by the restricted least squares estimator for $\beta \in \mathbb{R}^p$ by restricting $\beta$ to lie in the subspace generated by the columns $\lambda_1 t_1, \cdots, \lambda_r t_r$. Now we give the following definition.

**Definition 5.1** Let $A$ be an $m \times n$ matrix. The column space of $A$ is the set of all $m \times 1$ vectors $y$ satisfying $y = Ax$ for some $n \times 1$ vector $x$. The column space of $A$ is denoted by $C(A)$.

Using Definition 5.1 column space $C(T_1)$ of the matrix $T_1$ is generated by the columns $\lambda_1 t_1, \cdots, \lambda_r t_r$. Therefore, we restrict the true parameter $\beta$ to this space $C(T_1)$, that is, we assume that $\beta$ can be written as $\beta = T_1x$ for some $x$. This is satisfied if and only if $T_1'T_1\beta = \beta$ (Groß, 2003).

Since $T = (T_1, T_2)$ is orthogonal, $T_1'T_1 = I_s$, $T_2'T_2 = I_{p-s}$ and $T_1'T_2 = 0$ as well as $T_1'T_1 + T_2'T_2 = I_p$.

Now, $T_1'T_1\beta = \beta$ is equivalent to $T_2'\beta = 0$ (since $I_p - T_1'T_1 = T_2'T_2$). Therefore, we apply the concept of restricted least squares estimator with restriction
$T_1^2 \beta = 0$ to obtain the following estimator

$$
\hat{\beta}_r^* = \hat{\beta} - (X'X)^{-1}T_2[T_2' (X'X)^{-1}T_2]^{-1}T_2^2 \hat{\beta}.
$$

(5.3)

Now, we prove that RLS $\hat{\beta}_r^*$ is nothing but PCR $\hat{\beta}_r = T_1 (T_1' X'X T_1)^{-1} T_1' X'Y$.

**Theorem 5.1** Under the CMLR model with linear restrictions $T_2^2 \beta = 0$,

$$
\hat{\beta}_r^* = \hat{\beta} - (X'X)^{-1}T_2[T_2' (X'X)^{-1}T_2]^{-1}T_2^2 \hat{\beta}
= T_1 \Lambda_1^{-1} T_1' X'Y
= \hat{\beta}_r.
$$

(5.4)

**Proof:** See Groß (2003) for proof.

We have

$$(X'X)^{-1} = T_1 \Lambda_1^{-1} T_1' + T_2 \Lambda_2^{-1} T_2'.
$$

Multiply by $T_2'$ from left and $T_2$ from right and take inverse to get

$$
[T_2' (X'X)^{-1} T_2]^{-1} = \Lambda_2.
$$

Again, multiply both sides by $(X'X)^{-1}T_2$ from left and $(X'X)^{-1}T_2'$ from right to get

$$(X'X)^{-1} T_2[T_2' (X'X)^{-1} T_2]^{-1}T_2^2 (X'X)^{-1} = T_2 \Lambda_2^{-1} T_2'.
$$

But $(X'X)^{-1} = T_1 \Lambda^{-1} T_1' + T_2 \Lambda_1^{-1} T_2'$. Then

$$(X'X)^{-1} - (X'X)^{-1} T_2[T_2' (X'X)^{-1} T_2]^{-1}T_2^2 (X'X)^{-1} = T_1 \Lambda_1^{-1} T_1',
$$

where $\Lambda_1 = T_1' X'X T_1$. Therefore, we obtain

$$
\hat{\beta}_r^* = \hat{\beta} - (X'X)^{-1}T_2[V_2' (X'X)^{-1} T_2]^{-1} T_2^2 \hat{\beta}
$$

111
\[ (X'X)^{-1} - (X'X)^{-1}T_2T_2'(X'X)^{-1}T_2 ]^{-1}T_2'(X'X)^{-1}] X'Y \]
\[ = T_1A_1^{-1}I_1 \]
\[ = T_1(T_1'X'XV_1) - 1T_1'X'Y \]
\[ = \hat{\beta}_r. \]

### 5.2 Modified Liu-type estimator based on \((r - k)\) class estimator

The researchers try to improve the PCR estimator in order to give more precision in the estimation of the true parameter \(\beta\) in presence of MC.

For this purpose, Baye and Parker (1984) introduced a new estimator based on ridge estimation and principal components regression. They proposed a generalized estimator that combines the techniques of ORR and PCR as follows.

\[ \hat{\beta}_r(k) = V_r(V_r'X'T_r + kI_r)^{-1}V_r'X'Y, \quad k \geq 0. \]

This is known as the \((r - k)\) class estimator. The \((r - k)\) class estimator includes PCR, ORR and OLS estimators as special cases. By using \(\text{mse}\) as a criterion, they show that there exists \(k > 0\) such that \(\text{mse}(\hat{\beta}_r(k)) < \text{mse}(\hat{\beta}_r)\) for all \(0 < r \leq p\).

Kaçiranlar and Sakallioglu (2001) introduced the following estimator based on Liu estimator and PCR:

\[ \hat{\beta}_r(d) = V_r(V_r'X'T_r + I_r)^{-1}(V_r'X'Y + dV_r'\hat{\beta}_r). \quad 0 < d < 1. \]

\(\hat{\beta}_r(d)\) is known as the \((r - d)\) class estimator. The \((r - d)\) class estimator is also a general estimator and includes OLS, PCR and Liu estimators as special cases. There exists \(0 < d < 1\) such that \(\text{mse}(\hat{\beta}_r(d)) < \text{mse}(\hat{\beta}_r)\).

As we mention in Chapter 4, Sakallioglu and Kaçiranlar (2008) introduced the
(\(k - d\)) class estimator as a biased estimator based on ridge regression.

\[
\hat{\beta}(k, d) = (X'X + I)^{-1}(X'Y + d\hat{\beta}_k), k > 0, -\infty < d < \infty.
\]

Now, we introduce a new estimator for \(\beta\) as follows.

\[
\hat{\beta}_r(k, d) = V_r(V_r'X'XV_r + I_r)^{-1}(V_r'X'Y + dV_r'\hat{\beta}_r(k)),
\]

\(-\infty < d < \infty\) and \(k > 0\). We call this estimator the \((r - (k - d))\) class estimator. \(\hat{\beta}_r(k, d)\) generalizes OLS, ORR, LE, \((k - d)\), PCR, \((r - k)\) and \((r - d)\) estimators:

\[
\hat{\beta}_p(0, 1) = \hat{\beta}, \\
\hat{\beta}_p(k, 1 - k) = \hat{\beta}(k), \\
\hat{\beta}_p(0, d) = \hat{\beta}(d), \\
\hat{\beta}_p(k, d) = \hat{\beta}(k, d), \\
\hat{\beta}_r(0, 1) = \hat{\beta}_r, \\
\hat{\beta}_r(0, d) = \hat{\beta}_r(k), \\
\hat{\beta}_r(k, 1 - k) = \hat{\beta}_r(d).
\]

### 5.3 Comparison of estimators

In order to compare the performance of \((r - (k - d))\) estimator with others, we use MMSE and mse as criteria.

#### 5.3.1 Comparison between \((r - (k - d))\) class estimator and \((r - d)\) class estimator

MMSE and mse for \(\hat{\beta}_r(k, d)\) and \(\hat{\beta}_r(d)\) are

\[
\text{MMSE}(\hat{\beta}_r(k, d)) = \sigma^2 T_r S^{-1}_r(1)(I_r + dS^{-1}_r(k))T_r' ST_r(I_r + dS^{-1}_r(k))S^{-1}_r(1)T_r' + (T_r S^{-1}_r(1)(I_r - dS^{-1}_r(k))T_r' ST_r T_r' + T_{p-r} T_{p-r} )\beta'
\]
\[ (T_r S_{r}^{-1}(1)(I_r - dT_r^T S_r S_{r}^{-1}(k))T_r^T + T_{p-r} T_{p-r}^T). \quad (5.8) \]

\[
\text{MMSE}(\hat{\beta}_r(d)) = \sigma^2 T_r S_{r}^{-1}(1)T_r^T S_{r}^{-1} T_r S_r(d) T_r^T S_{r}^{-1} T_r^T S_{r}^{-1}(1)T_r^T \\
+ [(d - 1)T_r S_{r}^{-1}(1)T_r^T - T_{p-r} T_{p-r}^T] \beta \beta^t \\
[(d - 1)T_r S_{r}^{-1}(1)T_r^T - T_{p-r} T_{p-r}^T], \quad (5.9)\]

where \( S_{r}^{-1}(1) = (\Lambda_r + I_r) \) and \( S_{r}^{-1}(d) = (\Lambda_r + dI_r) \).

Also,

\[
\text{mse}(\hat{\beta}_r(k, d)) = \sum_{i=1}^{r} \frac{\sigma^2 \lambda_i (\lambda_i + k + d)^2 + (\lambda_i + k - d\lambda_i)^2 \alpha_i^2}{(\lambda_i + k)^2(\lambda_i + 1)^2} \\
+ \sum_{i=p-r}^{p} \alpha_i^2. \quad (5.10)\]

When \( k = 0 \) in (5.10), we get mse of the \((r - d)\) class estimator

\[
\text{mse}(\hat{\beta}_r(d)) = \text{mse}(\hat{\beta}_r(0, d)) = \sum_{i=1}^{r} \frac{\sigma^2 (\lambda_i + d)^2 + \lambda_i(1 - d)^2 \alpha_i^2}{\lambda_i(\lambda_i + 1)^2} \\
+ \sum_{i=p-r}^{p} \alpha_i^2. \quad (5.11)\]

Minimizing \( \text{mse}(\hat{\beta}_r(k, d)) \) with respect to \( d \), we get

\[
d_{\text{opt}} = \frac{\sum_{i=1}^{r} \lambda_i(\alpha_i^2 - \sigma^2)/(\lambda_i + k)(\lambda_i + 1)^2}{\sum_{i=1}^{r} \lambda_i(\lambda_i \alpha_i^2 + \sigma^2)/(\lambda_i + k)^2(\lambda_i + 1)^2}. \]

Fix \( k \), so,

\[
\text{mse}(\hat{\beta}_r(k, d)) - \text{mse}(\hat{\beta}_r(d)) = d^2 \sum_{i=1}^{r} \frac{(\lambda_i \alpha_i^2 + \sigma^2)(\lambda_i^2 - (\lambda_i + k)^2)}{\lambda_i(\lambda_i + k)^2(\lambda_i + 1)^2} + \\
2dk \sum_{i=1}^{r} \frac{(\alpha_i^2 - \sigma^2)}{(\lambda_i + k)(\lambda_i + 1)^2}. \]

Thus we have the following theorem.

114
Theorem 5.2 Let

\[ d^* = \frac{2k \sum_{i=1}^{r} \frac{(\sigma_i^2 - \sigma^2)}{(\lambda_i + k)(\lambda_i + 1)^2}} {\sum_{i=1}^{r} \frac{((\lambda_i + k)^2 - \lambda_i^2)(\lambda_i \alpha_i^2 + \sigma^2)}{\lambda_i(\lambda_i + 1)^2(\lambda_i + k)^2}}. \]

Then

a) When \( \sum_{i=1}^{r} \frac{(\sigma_i^2 - \sigma^2)}{(\lambda_i + k)(\lambda_i + 1)^2} > 0 \) then:

1) \( \text{mse}(\hat{\beta}_r(k, d)) > \text{mse}(\hat{\beta}_r(d)) \) for \( 0 < d < d^* \)

2) \( \text{mse}(\hat{\beta}_r(k, d)) < \text{mse}(\hat{\beta}_r(d)) \) for \( d^* < d \) or \( d < 0 \)

b) When \( \sum_{i=1}^{r} \frac{(\sigma_i^2 - \sigma^2)}{(\lambda_i + k)(\lambda_i + 1)^2} < 0 \) then:

1) \( \text{mse}(\hat{\beta}_r(k, d)) > \text{mse}(\hat{\beta}_r(d)) \) for \( d^* < d < 0 \)

2) \( \text{mse}(\hat{\beta}_r(k, d)) < \text{mse}(\hat{\beta}_r(d)) \) for \( 0 < d < d^* \).

5.3.2 Comparison between \( (r - (k - d)) \) class estimator and \( (r - k) \) class estimator

MMSE and mse of \( (r - k) \) class estimator are

\[
\text{MMSE}(\hat{\beta}_r(k)) = \sigma^2 T_r S_r^{-1}(k) \Lambda_r S_r^{-1}(k) T'_r + [T_r S_r^{-1}(k) \Lambda_r T'_r - I_p] \\
\beta \beta'[T_r S_r^{-1}(k) \Lambda_r T'_r - I_p]. \quad (5.12)
\]

When \( d = 1 - k \) in (5.10), we get mse of \( (r - k) \) class estimator

\[
\text{mse}(\hat{\beta}_r(k)) = \text{mse}(\hat{\beta}_r(k, 1 - k))
\]

115
\[ \text{MMSE}(\hat{\beta}_r) = \sigma^2 T_r S_r^{-1} T_r^* + [T_r S_r^{-1} \Lambda_r T_r^* - I_p] \hat{\beta}^\prime \]
\[ [T_r S_r^{-1} \Lambda_r T_r^* - I_p], \quad (5.14) \]

\[ \text{mse}(\hat{\beta}_r) = \text{mse}(\hat{\beta}_r(0, 1)) \]
\[ = \sum_{i=1}^r \frac{\sigma^2}{\lambda_i} + \sum_{i=p-r}^p \alpha_i^2. \quad (5.15) \]

Baye and Parker (1984) showed that \( \text{mse}(\hat{\beta}_r(k)) < \text{mse}(\hat{\beta}_r) \) for \( k > \frac{\sigma^2}{\max \alpha_i^2} > 0. \)

But, when \( d = 1 - k \), \( \text{mse}(\hat{\beta}_r(k, 1 - k)) = \text{mse}(\hat{\beta}_r(k)). \) For this, we have the following theorem.
Theorem 5.4  For $k > \frac{\sigma^2}{\max \sigma_i^2} > 0$, there exists $d < 1 - k$ such that $\text{mse}(\hat{\beta}_r(k, d))$ is smaller than $\text{mse}(\hat{\beta}_r)$.

5.3.4  An example

We illustrate the results of this Section using a dataset on Portland cement given in earlier Chapters. The model includes the intercept term. The matrix $X'X$ has eigenvalues $\lambda_1 = 211.367, \lambda_2 = 77.236, \lambda_3 = 28.459, \lambda_4 = 10.267$ and $\lambda_5 = 0.0349$. The condition number of $X'X$ is $\kappa = \lambda_{\text{max}}/\lambda_{\text{min}} = 6056.37$ and so $X$ is quite "ill-Conditioned". The least squares estimator of the regression coefficients is:

$$\hat{\beta} = (X'X)^{-1}X'Y$$
$$= [\hat{\beta}_0 \; \hat{\beta}_1 \; \hat{\beta}_2 \; \hat{\beta}_3 \; \hat{\beta}_4]'$$
$$= [62.4052 \; 1.5511 \; 0.5102 \; 0.1019 \; -0.1441]'$$.

The standardization is accomplished by transforming the linear model $Y = X\beta + \epsilon$ to $Y_s = X_s\beta_s + \epsilon$. The corresponding least squares estimator is:

$$\hat{\beta}_s = (X_s'X_s)^{-1}X_s'Y_s = [0.6065 \; 0.5277 \; 0.0434 \; -0.1603]'$$.

Since there are thirteen observations and four parameters in the standardized data, we obtain the following

$$\sigma^2_s = \frac{(Y_s - X_s\hat{\beta}_s)'(Y_s - X_s\hat{\beta}_s)}{n-p} = 0.00196$$.

The eigenvalues of $X_s'X_s$ are $2.2357, 1.57606, 0.18661, 0.00162$. The $4 \times 4$ matrix $V$ is the matrix of eigenvectors, $\Lambda$ is a $4 \times 4$ diagonal matrix of eigenvalues of $X_s'X_s$ so that $X_s'X_s = V\Lambda V'$. Then, $Z = X_sV$ and $\alpha = V'\beta_s$ so that $Y_s = X_s\beta_s + \epsilon = X_sVV'\beta_s + \epsilon = Z\alpha + \epsilon$. In orthogonal coordinates, the least squares estimator of the regression coefficients are:
\[ \hat{\alpha} = \Lambda^{-1} Z' Y_s = [0.65696 - 0.00831 0.3028 0.388]' \]

Numerical results are summarized in Table 1 to compare the new proposed estimator with OLS, \((r-k)\) estimator, PCR estimator and \((r-d)\) estimator. Values of \(d_{opt}, d^*\) and \(\text{mse}\) are obtained by replacing theoretical expressions of all unknown model parameters by their OLS estimates.

1. Using \(k_{HK} = \hat{\sigma}^2 / \sum_{i=1}^{r} \hat{\alpha}_i^2 = 0.00374\), we get \(d_{opt} = 0.9284\) and \(d^* = 0.84\). Now, Table 5.1 shows that \(\hat{\alpha}_r(d)\) has a smaller \(\text{mse}\) than \(\hat{\alpha}_r(k, d)\) for \(0 < d < d^*\), where

\[ \sum_{i=1}^{r} \frac{(\alpha_i^2 - \sigma^2)}{\lambda_i + k} \frac{1}{(\lambda_i + 1)^2} > 0. \]

For example,

\[ \text{mse}(\hat{\alpha}_r(d = 0.1)) = 0.1623 < \text{mse}(\hat{\alpha}_r(k = 0.00374, d = 0.1)) = 0.2385. \]

This matches the theoretical result in Theorem 5.2 part a(1). Also, \(\hat{\alpha}_r(k, d)\) has a smaller \(\text{mse}\) than \(\hat{\alpha}_r(d)\) for \(d > d^*\). For example,

\[ \text{mse}(\hat{\alpha}_r(k = 0.00374, d = 0.95)) = 0.161 < \text{mse}(\hat{\alpha}_r(d = 0.95)) = 0.163. \]

This matches the theoretical result in Theorem 5.2 part a(2).

2. Comparison between \(\text{mse}(\hat{\alpha}_r(k = 0.00374, d_{opt} = 0.928) = 0.161\) and \(\text{mse}(\hat{\alpha}_r(k = 0.00374)) = 0.163\) shows that \(\hat{\alpha}_r(k, d)\) has a smaller \(\text{mse}\) than \(\hat{\alpha}_r(k)\), and this also matches the theoretical result of Theorem 5.3.

3. Let \(k = 0.005\) and \(d = 0.95 < 1 - 0.005\). Comparison between \(\text{mse}(\hat{\alpha}_r(k = 0.005, d = 0.95)) = 0.162\) and \(\text{mse}(\hat{\alpha}_r = 0.1632\) shows that \(\hat{\alpha}_r(k, d)\) has a smaller \(\text{mse}\) than \(\hat{\alpha}_r\), and this also matches the theoretical result of Theorem 5.4.
Table 5.1: Values of estimates and mse for the estimators

<table>
<thead>
<tr>
<th></th>
<th>mse</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{\alpha})</td>
<td>1.22</td>
</tr>
<tr>
<td>(d = 0.1)</td>
<td></td>
</tr>
<tr>
<td>(\hat{\alpha}_r(d))</td>
<td>0.162</td>
</tr>
<tr>
<td>(\hat{\alpha}_r(k, d))</td>
<td>0.24</td>
</tr>
<tr>
<td>(d^* = 0.84)</td>
<td></td>
</tr>
<tr>
<td>(\hat{\alpha}_r(d))</td>
<td>0.163</td>
</tr>
<tr>
<td>(\hat{\alpha}_r(k, d))</td>
<td>0.162</td>
</tr>
<tr>
<td>(d_{opt} = 0.928)</td>
<td></td>
</tr>
<tr>
<td>(\hat{\alpha}_r(k))</td>
<td>0.163</td>
</tr>
<tr>
<td>(\hat{\alpha}_r(k, d))</td>
<td>0.153</td>
</tr>
<tr>
<td>(\hat{\alpha}_r)</td>
<td>0.165</td>
</tr>
<tr>
<td>(\hat{\alpha}_r(k, d))</td>
<td>0.161</td>
</tr>
</tbody>
</table>

The plot of \(\text{mse}(\hat{\beta}_r(k, d))\) and \(\text{mse}(\hat{\beta}_r(d))\) vs \(d\) on the interval \([0, 1]\) when \(k\) is fixed at \(k = 0.00374\) is presented in Figure 5.1. This figure indicates that \(\text{mse}(\hat{\beta}_r(k, d))\) decreases as \(d\) increases and, for large value of \(d\), \(\hat{\beta}_r(k, d)\) dominates \(\hat{\beta}_r(d)\). On the other hand, \(\text{mse}(\hat{\beta}_r(d))\) increases slowly as \(d\) increases and \(\hat{\beta}_r(d)\) dominates \(\hat{\beta}_r(k, d)\) for large values of \(d\).

The plot of \(\text{mse}(\hat{\beta}_r(k, d))\) and \(\text{mse}(\hat{\beta}_r(k))\) vs \(k\) on the interval \([0, 1]\) when \(d\) is fixed at \(d = 0.9284\) is presented in Figure 5.2. This figure indicates that both \(\text{mse}(\hat{\beta}_r(k, d))\) and \(\text{mse}(\hat{\beta}_r(k, d))\) increase as \(k\) increases. The estimator \(\hat{\beta}_r(k, d)\) dominates \(\hat{\beta}_r(k)\) when \(k > 0.10\). The plot of \(\text{mse}(\hat{\beta}_r(k, d))\) and \(\text{mse}(\hat{\beta}_r)\) vs \(k\) on the interval \([0, 0.05]\) when \(d\) is fixed at \(d = 0.928\) is presented in Figure 5.3. Note that both estimators dominate each other for some values of \(k\). For small values of \(k\), the proposed estimator dominates PCR, and for large values of \(k\), PCR dominates the proposed estimator.
5.4 Two other strategies for removing MC

Selection of variables is a problem in regression models when observations are collected on a large number of variables and the goal is to obtain a model with only important explanatory variables. There are several reasons for this. For example, regression models with a reasonable number of explanatory variables are easy to analyze and understand. Also, the presence of mutual relations among the explanatory variables is not likely to add much to the predictive power of the model while substantially increasing the sampling variation of the regression coefficients (see Neter et al., 1983).

MC is more often a "feature" of data rather than of the model. Several techniques have been proposed for detecting MC. One of these techniques is the variance inflation factor, abbreviated as VIF (Marquardt, 1970). VIF measures the increase in the variance of a regression estimate due to MC. Redundant variables can be dropped from the model to alleviate this problem. Most of the variable selection methods fall in one of the following two categories.

1. Methods not dealing with MC.

Figure 5.1: Plot of $\text{mse}(\hat{\beta}_r(k, d))$ and $\text{mse}(\hat{\beta}_r(d))$ vs $d$ when $k$ is fixed at $k = 0.00374$
2. Methods dealing with MC.

Stepwise regression methods (forward, backward, and stepwise) belong to the first category (Montgomery et al. 2003). These methods cannot detect MC or identify variable(s) that cause MC. Principal component regression (Shahar and Gonzalo, 1994), Lasso (Tibshirani, 1996) and Elastic net (Zou and Hastie, 2005) belong to the second category.
The purpose of this section is to introduce two new procedures of removing MC, depending on whether selection of variables is possible or not. We use the relation between VIF and the diagonal elements of the inverse of R and remove explanatory variables having VIF exceeding 10. When selection of variables is possible, we remove insignificant variables from the resulting model. The final model obtained in this way will have no MC and no insignificant variables. Otherwise, the final model will be without MC but may include some insignificant variables.

### 5.5 Background of the procedures

VIF is the coefficient of multiple determination in the regression of $x_i$ on the other $X$ variables. When the model (1.1) is fitted by least squares, the variances of the estimates $\hat{\beta}_1, \ldots, \hat{\beta}_p$ are

$$\text{Var} \left( \hat{\beta}_i \right) = \text{VIF}_i \left( \frac{\sigma^2}{S_{ii}} \right), \ i = 1, 2, \ldots, p,$$

where

$$S_{ii} = \sum_{u=1}^{n} \left( x_{iu} - \bar{x}_i \right)^2$$

is the usual corrected sum of squares of the column $x_i$.

If a column $x_i$ is orthogonal to all other columns of $X$, then $\text{VIF}_i = 1$. Thus, $\text{VIF}_i$ is a measure of inflation in $\sigma^2 / S_{ii}$ due to the relationship of other columns of $X$ with the column $x_i$. The $\text{VIF}_i$ can be defined specifically in the following way. Suppose that $R_i^2$ is the coefficient of determination obtained when $x_i$ is regressed on the remaining predictors $x_j$ with $j \neq i$. Then

$$\text{VIF}_i = \frac{1}{1 - R_i^2}. \quad (5.16)$$

If $R_i^2 = 0$, then $\text{VIF}_i$ will be 1. As $R_i^2$ approaches 1, $\text{VIF}_i$ will approach
infinity. Marquardt (1980) suggested that a VIF greater than 10 indicates the presence of strong MC.

MC in the linear regression model (1.1) is usually resolved by dropping redundant variables. That is, by avoiding redundant variables in the regression model unlike stepwise regression methods.

This section has two goals. First, use VIF as a criterion for selection of variables (if it is permissible) without fitting the model using inverse of the correlation matrix. Then fit the model and remove the insignificant variables. Second, use VIF as an indicator of MC (if explanatory variables cannot be removed) and use residual of the explanatory variable having highest VIF in place of that variable. In both cases, MC is removed from the model. Details of these procedures are described below.

5.5.1 Case one

In this case, we assume that selecting variables is possible. Write \( R^{-1} = ((a_{ij})) \), to obtain

\[
\text{VIF}_i = a_{ii}.
\]

Since we are interested in VIF as an indicator of MC, it is sufficient to invert the correlation matrix \( R \) and use \( a_{ii} \) corresponding to each explanatory variable. Then it is not necessary to fit the regression of \( x_i \) on other explanatory variables.

The following algorithm obtains a model that has no MC and no insignificant explanatory variables.

Algorithm 1

1. Compute the correlation matrix \( R \) for explanatory variables.

2. Compute \( R^{-1} \).
3. Remove the explanatory variable that has highest $a_{ii} > 10$. Let this explanatory variable be $x_i$. Otherwise, go to Step 5.

4. Remove $x_i$ from the model and go to step (1).

5. We have removed MC from the model. Now we can identify insignificant variable(s) using the $t - test$.

6. Remove insignificant explanatory variable(s) (that is, variables with $p-value > 0.05$).

7. Now the model is without MC and without insignificant variables.

5.5.2 Case two

If selection of variables in the model is not possible, the following approach is suggested to circumvent MC (see Douglas et al. 1985). Suppose $a_{ii}$ is highest and $a_{ii} > 10$, then we regress $X_1$ on other explanatory variables and compute the residuals. These residuals are the part of $x_1$ not explained by other explanatory variables and thus are orthogonal to others.

The algorithm in this case is as follows.

**Algorithm 2**

1. Compute the correlation matrix $R$ for the explanatory variables.

2. Compute $R^{-1}$.

3. Identify the highest $a_{ii}$. If it exceeds 10, then go to Step 4; otherwise, go to Step 8.

4. Regress the explanatory variable having highest $a_{ii} > 10$ on all other explanatory variables. Let us denote it by $x_i$.

5. Compute the residuals $e_i$. 

124
6. Use \( e_i \) instead of \( x_i \) in the model.

7. Goto step (1).

8. This model is without MC. Model is fitted using the explanatory variables identified by this algorithm.

5.6 Three examples

Three examples given below illustrate the method of this section.

5.6.1 Hald cement data (Draper and Smith, 1998)

Case one:

The R matrix is given by:

\[
R = \begin{pmatrix}
1 & 0.2286 & -0.8241 & -0.2454 \\
0.2286 & 1 & -0.1392 & -0.9730 \\
-0.8241 & -0.1392 & 1 & 0.0295 \\
-0.2454 & -0.9730 & 0.0295 & 1 \\
\end{pmatrix}
\]

Hence, \( R^{-1} \) is

\[
R^{-1} = \begin{pmatrix}
38.4962 & 94.1197 & 41.8841 & 99.7858 \\
94.1197 & 254.4232 & 105.0914 & 267.5394 \\
41.8841 & 105.0914 & 46.8684 & 111.1451 \\
99.7858 & 267.5394 & 111.1451 & 282.5129 \\
\end{pmatrix}
\]

the highest value among \( a_{ii} \) is \( a_{44} = 282.5129 \), implying that \( x_4 \) is redundant (see Figure 5.4).

By computing \( R^{-1} \) for the explanatory variables not including \( x_4 \) we have
obtained

\[
R^{-1}\big|_{x_4} = \begin{pmatrix}
3.2511 & -0.3774 & 2.6268 \\
-0.3774 & 1.0636 & -0.1629 \\
2.6268 & -0.1629 & 3.1421
\end{pmatrix}.
\]

In this case, \( a_{ii} \) in \( R^{-1}\big|_{x_4} \) indicates that we have removed MC.

Fit a multiple regression model for all the explanatory variables except \( x_4 \) to remove insignificant variables from the final model (see Table 5.6.1). The relevant statistics for the final model are shown in Table 5.3.

Table 5.2: Relevant statistics using all explanatory variables except \( x_4 \) for Hald cement data

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Parameter estimate</th>
<th>t-statistic</th>
<th>p-value</th>
<th>( a_{ii} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>48.194</td>
<td>12.32</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>( x_1 )</td>
<td>1.6959</td>
<td>8.29</td>
<td>0.000</td>
<td>3.3</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>0.65691</td>
<td>14.85</td>
<td>0.000</td>
<td>1.1</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>0.2500</td>
<td>1.35</td>
<td>0.209</td>
<td>3.3</td>
</tr>
<tr>
<td>( x_4 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Sq = 98.2\%     R-Sq (adj) = 97.2\%

Table 5.2 has no \( a_{ii} \) greater than 10. That is, we have removed MC. Also, Table indicates that we remove the explanatory variable \( x_3 \) since \( p-value > 0.05 \). From Table 5.2, the final model is without MC and without insignificant explanatory variables.

Table 5.3: The relevant statistics using all explanatory variables except \( x_4 \) and \( x_3 \) for Hald cement data

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Parameter estimate</th>
<th>t-statistic</th>
<th>p-value</th>
<th>( a_{ii} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>52.577</td>
<td>23.00</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>( x_1 )</td>
<td>1.4683</td>
<td>12.10</td>
<td>0.000</td>
<td>1.1</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>0.66225</td>
<td>14.44</td>
<td>0.000</td>
<td>1.1</td>
</tr>
<tr>
<td>( x_4 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Sq = 97.9\%     R-Sq (adj) = 97.4\%

126
Case two

In this case, we include all the explanatory variables in the model. According to Step 3, regress $x_4$ on others explanatory variables. Compute the residual of $x_4$ ($e_{x_4}$)

$$
e_{x_4} = \begin{pmatrix}
1.0379 \\
-0.6836 \\
-2.3264 \\
-0.7925 \\
0.5227 \\
-0.3169 \\
0.1658 \\
0.5504 \\
-1.1281 \\
0.4439 \\
0.7463 \\
0.8882 \\
0.8924
\end{pmatrix}.
$$

Use $e_{x_4}$ instead of $x_4$ as an explanatory variable with others. Computing $R^{-1}$ for $x_1, x_2, x_3$ and $e_{x_4}$, one obtains

$$
R^{-1} = \begin{pmatrix}
3.2511 & -0.3774 & 2.6268 & -0.0000 \\
-0.3774 & 1.0636 & -0.1629 & -0.0000 \\
2.6268 & -0.1629 & 3.1421 & -0.0000 \\
-0.0000 & -0.0000 & -0.0000 & 1.0000
\end{pmatrix}.
$$

Since $R^{-1}$ has no $a_{ii} > 10$, we have removed MC. Then we fit the model (see Table).

5.6.2 Steam plant data (Draper and Smith, 1998)

In this data, $x_1$ is the response variable and $x_2 - x_{10}$ are the explanatory variables (see Draper and Smith, 1998, p.46).
Table 5.4: The relevant statistics using $x_1, x_2, x_3$ and $e_{x_4}$

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Parameter estimate</th>
<th>$a_{ii}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>48.194</td>
<td></td>
</tr>
<tr>
<td>$x_1$</td>
<td>1.6959</td>
<td>3.3</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.65691</td>
<td>1.1</td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.2500</td>
<td>3.3</td>
</tr>
<tr>
<td>$e_{x_4}$</td>
<td>-0.1441</td>
<td>1.0</td>
</tr>
<tr>
<td>R-Sq</td>
<td>98.2%</td>
<td></td>
</tr>
<tr>
<td>R-Sq (adj)</td>
<td>97.2%</td>
<td></td>
</tr>
</tbody>
</table>

Case one

The diagonal elements ($a_{ii}$) of $R^{-1}$ are as follows.

$$a_{ii} = \begin{pmatrix}
15.7466 \\
20.1371 \\
126.6256 \\
1.8366 \\
4.4119 \\
4.6950 \\
6.0674 \\
107.5909 \\
2.3850
\end{pmatrix}$$

Note that, $a_{33}=126.6256$ is very large. That is, $x_3$ does not have independent information in addition to other explanatory variables (see Figures 5.5 and 5.6). Therefore, we remove $x_3$ from the model. If we recompute $R_{|x_3}$ and calculate
Since \( a_{22} = 16.4214 \) is greater than 10, we remove \( x_2 \) from the model and recompute \( R^{-1}|_{x_3,x_2} \).

\[
R^{-1}|_{x_3,x_2} = \begin{pmatrix}
14.6328 \\
16.4214 \\
1.8242 \\
3.8567 \\
4.6945 \\
5.0564 \\
1.8676 \\
2.2562
\end{pmatrix}
\]

At this step, we can fit the model (without \( x_3 \) and \( x_2 \)) to remove the insignificant explanatory variables. Table 5.5 shows the final model for the Steam plan data. Nine explanatory variables in the original model explain 87.4\% of the variation in the dependent variable. Table 5.5 shows the model with only two explanatory variables that explain 83.8\% of the variation in the dependent variable. The second model has reduced the number of explanatory variables from nine to two (that is, by 77.77\%) at a loss of 3.6\% in the explained variation.

**Case two**

Like case two in example 5.6.1, Table 5.6 shows the final model without MC with all explanatory variables as follows.
Table 5.5: The relevant statistics by using the remained variables after removing the MC and insignificant explanatory variables for the Steam plant data

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Parameter estimate</th>
<th>t-statistic</th>
<th>p-value</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>9.188</td>
<td>8.49</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>$x_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.19771</td>
<td>4.40</td>
<td>0.000</td>
<td>1.0</td>
</tr>
<tr>
<td>$x_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_5$</td>
<td>-0.071998</td>
<td>-9.17</td>
<td>0.000</td>
<td>1.0</td>
</tr>
<tr>
<td>$x_6$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_7$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_8$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_9$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_{10}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R-Sq = 85.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R-Sq (adj) = 83.8%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.6.3 Residential load survey data (see Wadsworth, 1998)

From Table 5.7, working of the new procedures is clear. Therefore, results for the two cases are given directly in Tables 5.8-5.9. This example also shows the performance of the new procedures for removing MC whether selection of variables is allowed or not.

In this example, $a_{ii} = 191.3, 180.7, 4.0, 11.5$ and $1.1$. Therefore, using steps 1-7 in case one, Table 5.8 explains the final model (see Figure 5.7).

5.7 Comparison of the new procedures with some selection of variables procedures and biased estimators

Since several procedures deal with MC, we need to compare the performance of the new procedures with others. Since we are dealing with different procedures in the same model, the performance will be compared for the final model. Therefore,
Table 5.6: The relevant statistics by using the explanatory variables with Residuals of some explanatory variables for the Steam plant data

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Parameter estimate</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.58</td>
<td></td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.389</td>
<td>3.8</td>
</tr>
<tr>
<td>$e_{x_3}$</td>
<td>1.11</td>
<td>1</td>
</tr>
<tr>
<td>$e_{x_4}$</td>
<td>-3.62</td>
<td>1</td>
</tr>
<tr>
<td>$x_5$</td>
<td>0.144</td>
<td>1.6</td>
</tr>
<tr>
<td>$x_6$</td>
<td>0.159</td>
<td>2.7</td>
</tr>
<tr>
<td>$x_7$</td>
<td>-0.0113</td>
<td>4.7</td>
</tr>
<tr>
<td>$x_8$</td>
<td>-0.0869</td>
<td>5</td>
</tr>
<tr>
<td>$x_9$</td>
<td>-0.00651</td>
<td>1.8</td>
</tr>
<tr>
<td>$x_{10}$</td>
<td>-0.215</td>
<td>2.2</td>
</tr>
</tbody>
</table>

R-Sq = 92.11%  
R-Sq (adj) = 87.4%

Table 5.7: The relevant statistics using all the explanatory variables for the Residential Load Survey Data

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Parameter estimate</th>
<th>t-statistic</th>
<th>p-value</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.0403</td>
<td>-0.16</td>
<td>0.874</td>
<td></td>
</tr>
<tr>
<td>$x_1$</td>
<td>-2.498</td>
<td>-2.43</td>
<td>0.020</td>
<td>191.3</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.26940</td>
<td>3.01</td>
<td>0.005</td>
<td>180.7</td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.41418</td>
<td>10.56</td>
<td>0.000</td>
<td>4.0</td>
</tr>
<tr>
<td>$x_4$</td>
<td>0.37955</td>
<td>4.24</td>
<td>0.000</td>
<td>11.5</td>
</tr>
<tr>
<td>$x_5$</td>
<td>0.03027</td>
<td>1.31</td>
<td>0.200</td>
<td>1.1</td>
</tr>
</tbody>
</table>

R-Sq = 97.9%  
R-Sq (adj) = 96.6%

we use $(\hat{R}^2) = R^2_{Adjusted}$,

$$\hat{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1} \quad (5.17)$$

as a measure for goodness of fit.
Table 5.8: The relevant statistics using the remaining variables after removing MC and insignificant explanatory variables for the Residential Load Survey Data.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Parameter estimate</th>
<th>t-statistic</th>
<th>p-value</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.2631</td>
<td>1.33</td>
<td>0.191</td>
<td></td>
</tr>
<tr>
<td>$x_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.45385</td>
<td>13.37</td>
<td>0.000</td>
<td>2.3</td>
</tr>
<tr>
<td>$x_4$</td>
<td>0.50289</td>
<td>10.92</td>
<td>0.000</td>
<td>2.3</td>
</tr>
<tr>
<td>$x_5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Sq = 97.0%  R-Sq (adj) = 96.8%

Table 5.9: The relevant statistics using the explanatory variables with Residuals of some explanatory variables for the Residential Load Survey Data.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Parameter estimate</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.0818</td>
<td></td>
</tr>
<tr>
<td>$\epsilon_{x_1}$</td>
<td>-2.498</td>
<td>1</td>
</tr>
<tr>
<td>$\epsilon_{x_2}$</td>
<td>0.06173</td>
<td>1</td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.45053</td>
<td>2.3</td>
</tr>
<tr>
<td>$x_4$</td>
<td>0.51403</td>
<td>2.4</td>
</tr>
<tr>
<td>$x_5$</td>
<td>0.04251</td>
<td>1</td>
</tr>
</tbody>
</table>

R-Sq = 97.9%  R-Sq (adj) = 97.6%

5.7.1 Variable selection methods

Least absolute shrinkage and selection operator (Lasso)

Tibshirani (1996) introduced the Lasso. The Lasso minimizes the residual sum of squares subject to the sum of the absolute values of the coefficients being less than a constant $s$. If the data are standardized to have mean 0, the Lasso estimate is defined as follows:

$$ \hat{\beta}_{Lasso} = \arg \min_{\beta} (Y - X\beta)'(Y - X\beta) \ \text{subject to} \ \sum |\beta_j| \leq s. \quad (5.18) $$

The solution of (5.18) can be obtained by standard quadratic programming with linear inequality constraints. The use of Least Angle regression (LARS) algorithm reduces the computational burden (Efron et al., 2004).
The nested estimate procedure

Feng-Jenq Lin (2008) introduced the nested estimate procedure. The procedure concept is based completely on the least squares method. He tried to solve MC in linear regression model. He estimated the different parameters of explanatory variables individually and sequentially at each iteration. The final model avoided MC and insignificant variables.

Table 5.10 shows the final model for the nested estimate procedure, the new procedure (case one) and Lasso for the examples (5.6.1-5.6.3). From this Table, the following facts can be seen:

1. The Lasso technique selected more explanatory variables than the new procedure. But, $R^2_{\text{Adjusted}}$ is still less than the $R^2_{\text{Adjusted}}$ for the new procedure. Therefore, it is better to deal with the new procedure.

2. The nested estimate procedure and the new procedure have the same number of variables, but the variables selected by the new procedure have high $R^2_{\text{Adjusted}}$ compared with the nested estimate procedure except for example 5.6.2 where the nested estimate explains better than the new estimate, but slightly.

5.7.2 Some biased estimators

In order to compare the performance of case two, where all the explanatory variables are included in the final model, we introduce some biased estimators like ridge regression estimator and Liu estimator.

The biased estimation is another technique that deals with MC. In biased estimation we add a constant such that it reduces the variance and does not give high bias. Heorl and Kennard introduced the ridge regression estimator as follows

$$\hat{\beta}(k) = (X'X + kI)^{-1} X'Y, \quad 0 < k < 1$$
Table 5.10: The final model for the Lasso, the nested estimate procedure and the new procedure case one (New1) for all the examples

<table>
<thead>
<tr>
<th>Example</th>
<th>Procedure</th>
<th>Final model</th>
<th>R-Sq (adj)%</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.6.1</td>
<td>Lasso</td>
<td>71.6704+1.4507(x_1)+0.4158(x_2) [-0.2365(x_3) | 73.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nested</td>
<td>107+1.35 (x_1)+0.738 (x_4) | 94.91</td>
<td></td>
</tr>
<tr>
<td></td>
<td>New1</td>
<td>48.194+1.4683 (x_1)+0.6625 (x_2) | 97.4</td>
<td></td>
</tr>
<tr>
<td>5.6.2</td>
<td>Lasso</td>
<td>-0.7830(-0.7884) (x_1)+2.6726 (x_2) (+0.0492) (x_3)+0.2941 (x_4)+0.0151 (x_6) -0.0285 (x_7)+0.9957 (x_9) | 57.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nested</td>
<td>9.49(+0.756) (x_2)+0.079 (x_8) | 84.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>New1</td>
<td>9.188(+0.1977) (x_6)-0.072 (x_8) | 83.8</td>
<td></td>
</tr>
<tr>
<td>5.6.3</td>
<td>Lasso</td>
<td>-0.2128(+0.0623) (x_2)+0.3943 (x_3) (+0.3328) (x_4)+0.0336 (x_5) | 96.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nested</td>
<td>-1.99(+0.254) (x_2)+0.101 (x_3) | 91.31</td>
<td></td>
</tr>
<tr>
<td></td>
<td>New1</td>
<td>0.26(+0.454) (x_3)+0.5 (x_4) | 96.8</td>
<td></td>
</tr>
</tbody>
</table>

Since \(k\) is small, ORR is unstable. Therefore, Liu combined ORR with Stein estimator to use advantages of ORR and Stein estimator as follows

\[ \hat{\beta}(d) = (X'X+I)^{-1}(X'Y+d\hat{\beta}) \], \quad 0 < d < 1.

Table 5.11 gives a summary of the performance of three methods for above three examples. Note that, the new method is the best in two of three examples.
Table 5.11: $R^2$ Adjusted for Ridge, Liu and case two of the new procedure (New2) for all the examples

<table>
<thead>
<tr>
<th>Example</th>
<th>Procedure</th>
<th>R-Sq (adj)%</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>ridge</td>
<td>96.87</td>
</tr>
<tr>
<td></td>
<td>Liu</td>
<td>86.8072</td>
</tr>
<tr>
<td></td>
<td>New2</td>
<td>97.2</td>
</tr>
<tr>
<td>3.2</td>
<td>Ridge</td>
<td>85.2874</td>
</tr>
<tr>
<td></td>
<td>Liu</td>
<td>83.8934</td>
</tr>
<tr>
<td></td>
<td>New2</td>
<td>87.4</td>
</tr>
<tr>
<td>3.3</td>
<td>Ridge</td>
<td>97.4442</td>
</tr>
<tr>
<td></td>
<td>Liu</td>
<td>95.8795</td>
</tr>
<tr>
<td></td>
<td>New2</td>
<td>96.8</td>
</tr>
</tbody>
</table>
Figure 5.5: Predicted against observed values for Steam data
Figure 5.6: Predicted against observed values for Steam data
Figure 5.7: Predicted against observed values for Residential data