Chapter 11
Building the Regression Model II:
Remedial Measures (補救措施)

許湘伶

Applied Linear Regression Models
(Kutner, Nachtsheim, Neter, Li)
remedial measures may need to be taken:

- a regression model is not appropriate
- several cases are very influential

previous: transformations to linearize the regression relation

- the error distributions more nearly normal
- make the variances of the error terms more nearly equal
remedial measures

In this chapter- remedial measure to deal with:

- unequal error variances
- a high degree of multicollinearity
- influential observations

two methods for nonparametric regression:

- lowess
- regression trees

bootstrapping: for evaluating the precision of the complex estimators
Weighted Least Squares (WLS)

- Chap. 3,6: transformation of $Y$ - reducing or eliminating unequal variances
- difficulty: may create an inappropriate regression relationship
- weighted least squares: when an appropriate regression relationship has been found but the variances of the error terms are unequal
Weighted Least Squares (WLS) (cont.)

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \quad i = 1, \ldots, n \]

- parameters: \( \beta_0, \ldots, \beta_{p-1} \)
- known constants: \( X_{i1}, \ldots, X_{i,p-1} \)
- \( \epsilon_i \sim \text{iid} \sim N(0, \sigma_i^2) \)

\[
\begin{bmatrix}
\sigma_1^2 & 0 & \cdots & 0 \\
0 & \sigma_2^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma_n^2
\end{bmatrix}
\]

\[ b = (X'X)^{-1}X'Y \Rightarrow \text{unbias, consistent but don't have minimum variance} \]
Weighted Least Squares (WLS) (cont.)

- Consider: Observations with small variances provide more reliable information about the regression function than those with large variances.

- Error Variances Know \( w_i = \frac{1}{\sigma_i^2} \) (unrealistic 不切實際的)

- Error Variances Know up to Proportionality Constant \( w_i = k \frac{1}{\sigma_i^2} \)

- Error Variances Unknown: estimation of variance function or standard deviation function \( w_i = \frac{1}{(\hat{s}_i)^2}, \quad w_i = \frac{1}{\hat{\nu}_i} \)
WLS-error variances known

Methods of maximum likelihood to obtain estimators:

- \( \varepsilon_i \overset{\text{indep.}}{\sim} N(0, \sigma_i^2) \) and \( w_i = \frac{1}{\sigma_i^2} \)

\[
L(\beta) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma_i^2)^{1/2}} \exp \left[ -\frac{1}{2\sigma_i^2} \left( Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1} \right)^2 \right]
\]

\[
= \left[ \prod_{i=1}^{n} \left( \frac{w_i}{2\pi} \right)^{1/2} \right] \exp \left[ -\frac{1}{2} \sum_{i=1}^{n} w_i \left( Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1} \right)^2 \right]
\]

\[
\Rightarrow \quad b = \arg \max_{\beta} L(\beta)
\]

\[
\iff b = \arg \min_{\beta} Q_w \quad (Q_w = \sum_{i=1}^{n} w_i (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1})^2)
\]

- Called \textit{weighted least squares criterion} (min \( Q_w \))
- \( w_i = 1 \Rightarrow \text{OLS} \)
WLS-error variances known (cont.)

- $w_i = \frac{1}{\sigma_i^2}$: reflects the amount of information contained in the observation $Y_i$
- $\text{var} Y_i$ large $\Rightarrow w_i$ less

$$
\begin{bmatrix}
w_1 & 0 & \cdots & 0 \\
0 & w_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & w_n
\end{bmatrix}
$$

- The normal equations: $(X'WX)b_w = X'WY$
- The weighted least squares (MLE) for $\beta$:

$$
b_w = (X'WX)^{-1}X'WY
\Rightarrow \sigma^2\{b_w\} = (X'WX)^{-1} \quad \therefore \sigma^2\{Y\} = W^{-1}
$$
WLS-error variances known (cont.)

- $b_w$: unbiased, consistent, have minimum variance among unbiased linear estimators

- When the weighted are known, $b_w$ generally exhibits less variability than $b$. 
WLS-error variances known up to proportionality constant

\[ w_i = k \frac{1}{\sigma^2_i} \]

\[ b_w = (X'WX)^{-1}X'WY \quad \text{(unaffected)} \]

\[ \Rightarrow \sigma^2 \{b_w\} = k(X'WX)^{-1} \]

\[ k: \text{unknown} \Rightarrow s^2 \{b_w\} = MSE_w(X'WX)^{-1} \]

where \( MSE_w = \frac{\sum w_i(Y_i - \hat{Y}_i)^2}{n - p} = \frac{\sum w_ie_i^2}{n - p} \)

- \( MSE_w \): an estimator of the proportionality constant \( k \)
WLS-error variances unknown

Estimation of Variance Function or Standard Deviation Function

- One rarely has knowledge of the variances $\sigma_i^2$. $\Rightarrow$ estimate of the variances

$$\sigma_i^2 = E\{\varepsilon_i^2\} - (E\{\varepsilon_i\})^2 = E\{\varepsilon_i^2\}$$

- Estimator of $\sigma_i^2$: the squared residual $e_i^2$
- Estimator of $\sigma_i = |\sqrt{\sigma_i^2}|$: the absolute residual $|e_i|$
WLS-error variances unknown (cont.)

the estimation process:

1. Fit the regression model by OLS and analyze $e_i$
2. Estimate the variance function or the standard deviation function by regressing $e_i^2$ or $|e_i|$ on the appropriate predictor(s).
3. $\hat{s}_i$ for standard deviation function, $\hat{\nu}_i$ for the variance function: to obtain $w_i$
4. obtain $b_w$ by using $w_i$

(Sometimes, iterated for several times to reach stabilize⇒ iteratively reweighted least squares)
WLS-error variances unknown (cont.)

Figure 4.1 Residuals vs. fitted plots—the first suggests no change to the current model while the second shows nonconstant variance and the third indicates some nonlinearity, which should prompt some change in the structural form of the model.

Reference: figure from Faraway’s Linear Models with R (2005, p. 59)
Some possible variance and standard deviation functions:
1. plot $e_i$ vs. $X_1$: a megaphone shape ⇒ $|e_i|$ regresses on $X_1$
2. plot $e_i$ vs. $\hat{Y}$: a megaphone shape ⇒ $|e_i|$ regresses on $\hat{Y}$
3. plot $e_i^2$ vs. $X_3$: upward tendency ⇒ $e_i^2$ regresses on $X_3$
4. plot $e_i$ vs. $X_2$: increase rapidly with $X_2$ and increases more slowly ⇒ $|e_i|$ regresses on $X_2, X_2^2$

⇒ $w_i = \frac{1}{(\hat{s}_i)^2}$ : $\hat{s}_i$-fitted value from standard function

$w_i = \frac{1}{\hat{\nu}_i}$ : $\hat{\nu}_i$-fitted value from variance function

⇒ the weight matrix $W$ ⇒ $b_w = (X'WX)^{-1}X'WY$
WLS-error variances unknown (cont.)

Using of Replicated or Near Replicates

- Replicate observations are made at each combination of levels of $X$s
- If the number of replications (or near replications) is large, $w_i$ may be obtained directly from the sample variances of $Y$ observations at each combination of levels on $X$ variables
- Each case in a replicate group receives the same weight with this method
- Confidence interval: (similar but approximate)

$$b_k \pm t(1 - \alpha/2; n - p)s\{b_k\} \quad (6.50)$$
Using of Ordinary Least Squares with Unequal Error Variances:

\[ b = (X'X)^{-1}X'Y \]

\[ \Rightarrow \sigma^2\{b\} = (X'X)^{-1}(X'\sigma^2\{\varepsilon\}X)(X'X)^{-1} \]

\[ \Rightarrow S^2\{b\} = (X'X)^{-1}(X'S_0\{\varepsilon\}X)(X'X)^{-1} \]

where \( S_0 = \text{diag}\{e_1^2, e_2^2, \ldots, e_n^2\} \)
blood pressure example

- 54 subjects: 20-60 women

| Subject | Age | Y | \( e_i \) | \( |e_i| \) | \( s_i \) | \( w_i \) |
|---------|-----|---|----------|-------|-------|-------|
| 1       | 27  | -73 | 1.18    | 1.18  | 3.801 | .06921 |
| 2       | 21  | 66  | -2.34   | 2.34  | 2.612 | .14656 |
| 3       | 22  | 63  | -5.92   | 5.92  | 2.810 | .12662 |
| ...     | ... | ... | ...     | ...   | ...   | ...   |
| 52      | 52  | 100 | 13.68   | 13.68 | 8.756 | .01304 |
| 53      | 58  | 80  | -9.80   | 9.80  | 9.944 | .01011 |
| 54      | 57  | 109 | 19.78   | 19.78 | 9.746 | .01053 |
blood pressure example (cont.)

linear regression function by unweighted least squares:

\[ \hat{Y} = 56.157 + 0.58003X \]

\[ (3.994) \quad (0.09695) \]
| $e_i|$ regresses on $X$: $\hat{s} = -1.54946 + 0.198172X$

$\hat{s}_1 = 3.801 \Rightarrow w_1 = 1/(\hat{s}_1)^2 = 0.0692$

WLS: $\hat{Y} = 55.566 + 0.59634X$

C.I. for $\beta_1$: $\alpha = 0.05$; $s\{b_{w1}\} = 0.07924$

$0.59643 \pm t(0.975; 52)(0.07924) \Rightarrow 0.437 \leq \beta_1 \leq 0.755$

$= 2.007$
blood pressure example (cont.)

- unequal variances: heteroscedasticity (異質性); ⇔ equal variances: homoscedasticity (等變異性)
- $R^2$ does not have a clear-cut meaning for WLS
- WLS may be view as OLS of transformed variables:

\[
Y = X\beta + \varepsilon, \quad E\{\varepsilon\} = 0, \sigma^2\{\varepsilon\} = W^{-1}
\]

\[
\Rightarrow W^{1/2} Y = W^{1/2} X\beta + W^{1/2}\varepsilon
\]

\[
\Rightarrow E\{\varepsilon_w\} = 0; \quad \sigma^2\{\varepsilon_w\} = I;
\]

\[
b_w = (X'_w X_w)^{-1} X'_w Y_w = (X'WX)^{-1} X'Wy
\]
and the weighted least squares estimators $b_{w0}$ and $b_{w1}$ in (11.9) are:

$$b_{w1} = \frac{\sum w_i X_i Y_i - \frac{\sum w_i X_i \sum w_i Y_i}{\sum w_i}}{\sum w_i X_i^2 - \frac{(\sum w_i X_i)^2}{\sum w_i}}$$  \hspace{1cm} (11.26a)$$

$$b_{w0} = \frac{\sum w_i Y_i - b_1 \sum w_i X_i}{\sum w_i}$$  \hspace{1cm} (11.26b)$$
Ridge Regression

- **ridge regression**: a method of overcoming serious multicollinearity problem by modifying the method of least squares.
- allowing *biased estimators* of the regression functions
- When an estimator has only a small bias and is substantially more precise than an unbiased estimator, it may well be the preferred estimator since it will have a larger probability of being close to the true parameter value.
The mean squared error (MSE):

\[ E \{ \hat{b}^R - \beta \}^2 = \sigma^2 \{ \hat{b}^R \} + (E \{ \hat{b}^R \} - \beta)^2 \]
Ridge Regression (脊迴歸) (cont.)

- Transformed variables:
  
  \[ Y_i^* = \frac{1}{\sqrt{n-1}} \left( \frac{Y_i - \bar{Y}}{s_Y^2} \right), \quad X_{ik}^* = \frac{1}{\sqrt{n-1}} \left( \frac{X_{ik} - \bar{X}_k}{s_k^2} \right) \]

  \[ \Rightarrow r_{XX} b = r_{YX} \]

- The ridge standardized regression estimators are obtained by:
  
  \[ (r_{XX} + cI)b^R = r_{YX} \]
  
  \[ \Rightarrow b^R = (r_{XX} + cI)^{-1} r_{YX} \]

- \( c \geq 0 \): reflects the amount of bias
- \( c = 0 \Rightarrow \text{OLS}; \ c > 0 \Rightarrow \text{bias but more stable (less variable)} \)
Choice of Biasing Constant $c$:

- bias of $E \{ b^R - \beta \}^2 \uparrow$ as $c \uparrow$ while the variance component becomes smaller

- difficulty: the optimum value of $c$ varies from one application to another and is unknown

- Using Ridge trace and $(VIF)_k(c)$
Ridge Regression (脊迴歸) (cont.)

- **Ridge trace**: a simultaneous plot of the values of the $(p - 1) b^R$; $0 \leq c \leq 1$

![Ridge Trace of Estimated Standardized Regression Coefficients—Body Fat Example with Three Predictor Variables.](image)

- Choose $c$: the Ridge trace starts to become stable and VIF has become sufficiently small.
Ridge Regression (脊迴歸) (cont.)

- The normal equation for the ridge estimators:

\[
\begin{align*}
(1 - c)b_1^R + r_{12}b_2^R + \cdots + r_{1,p-1}b_{p-1}^R &= rY_1 \\
 r_{21}b_1^R + (1 + c)b_2^R + \cdots + r_{2,p-1}b_{p-1}^R &= rY_2 \\
&\vdots \\
r_{p-1,1}b_1^R + r_{p-1,2}b_2^R + \cdots + (1 - c)b_{p-1}^R &= rY_{p-1}
\end{align*}
\]

- VIF for \(b_k^R\) are defined analogously to those for OLS \(b_k\): measures how large is the variance of \(b_k^R\) relative to what the variance would be if \(Xs\) were uncorrelated.
11.2 Ridge Regression

Ridge Regression (脊迴歸) (cont.)

- **VIF** for $b_k^R$: the diagonal elements of the following matrix

$$ (r_{XX} + cI)^{-1} r_{XX} (r_{XX} + cI)^{-1} $$

- Ridge regression estimates can be obtained by the method of **penalized least squares**.

$$ Q = \sum_{i=1}^{n} \left[ Y_i^* - (\beta_1^* X_{i1}^* + \cdots + \beta_{p-1}^* X_{i,p-1}^*) \right]^2 + c \left[ \sum_{j=1}^{p-1} (\beta_j^*)^2 \right] $$

sometimes referred to as shrinkage estimators

- **Limitation**: ridge regression is that ordinary inference procedures are not applicable and exact distribution properties are not known. (bootstrapping)
### TABLE 11.2 Ridge Estimated Standardized Regression Coefficients for Different Biasing Constants $c$—Body Fat Example with Three Predictor Variables.

<table>
<thead>
<tr>
<th>$c$</th>
<th>$b_1^R$</th>
<th>$b_2^R$</th>
<th>$b_3^R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.000</td>
<td>4.264</td>
<td>-2.929</td>
<td>-1.561</td>
</tr>
<tr>
<td>.002</td>
<td>1.441</td>
<td>-4.113</td>
<td>-4.813</td>
</tr>
<tr>
<td>.004</td>
<td>1.006</td>
<td>-.0248</td>
<td>-3.149</td>
</tr>
<tr>
<td>.006</td>
<td>.8300</td>
<td>.1314</td>
<td>-.2472</td>
</tr>
<tr>
<td>.008</td>
<td>.7343</td>
<td>.2158</td>
<td>-.2103</td>
</tr>
<tr>
<td>.010</td>
<td>.6742</td>
<td>.2684</td>
<td>-.1870</td>
</tr>
<tr>
<td>.020</td>
<td>.5463</td>
<td>.3774</td>
<td>-.1369</td>
</tr>
<tr>
<td>.030</td>
<td>.5004</td>
<td>.4134</td>
<td>-.1181</td>
</tr>
<tr>
<td>.040</td>
<td>.4760</td>
<td>.4302</td>
<td>-.1076</td>
</tr>
<tr>
<td>.050</td>
<td>.4605</td>
<td>.4392</td>
<td>-.1005</td>
</tr>
<tr>
<td>.100</td>
<td>.4234</td>
<td>.4490</td>
<td>-.0812</td>
</tr>
<tr>
<td>.500</td>
<td>.3377</td>
<td>.3791</td>
<td>-.0295</td>
</tr>
<tr>
<td>1.000</td>
<td>.2798</td>
<td>.3101</td>
<td>-.0059</td>
</tr>
</tbody>
</table>

### TABLE 11.3 VIF Values for Regression Coefficients and $R^2$ for Different Biasing Constants $c$—Body Fat Example with Three Predictor Variables.

<table>
<thead>
<tr>
<th>$c$</th>
<th>$(VIF)_1$</th>
<th>$(VIF)_2$</th>
<th>$(VIF)_3$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.000</td>
<td>708.84</td>
<td>564.34</td>
<td>104.61</td>
<td>.8014</td>
</tr>
<tr>
<td>.002</td>
<td>50.56</td>
<td>40.45</td>
<td>8.28</td>
<td>.7901</td>
</tr>
<tr>
<td>.004</td>
<td>16.98</td>
<td>13.73</td>
<td>3.36</td>
<td>.7864</td>
</tr>
<tr>
<td>.006</td>
<td>8.50</td>
<td>6.98</td>
<td>2.19</td>
<td>.7847</td>
</tr>
<tr>
<td>.008</td>
<td>5.15</td>
<td>4.30</td>
<td>1.62</td>
<td>.7838</td>
</tr>
<tr>
<td>.010</td>
<td>3.49</td>
<td>2.98</td>
<td>1.38</td>
<td>.7832</td>
</tr>
<tr>
<td>.020</td>
<td>1.10</td>
<td>1.08</td>
<td>1.01</td>
<td>.7818</td>
</tr>
<tr>
<td>.030</td>
<td>.63</td>
<td>.70</td>
<td>.92</td>
<td>.7812</td>
</tr>
<tr>
<td>.040</td>
<td>.45</td>
<td>.56</td>
<td>.88</td>
<td>.7808</td>
</tr>
<tr>
<td>.050</td>
<td>.37</td>
<td>.49</td>
<td>.85</td>
<td>.7804</td>
</tr>
<tr>
<td>.100</td>
<td>.25</td>
<td>.37</td>
<td>.76</td>
<td>.7784</td>
</tr>
<tr>
<td>.500</td>
<td>.15</td>
<td>.21</td>
<td>.40</td>
<td>.7427</td>
</tr>
<tr>
<td>1.000</td>
<td>.11</td>
<td>.14</td>
<td>.23</td>
<td>.6818</td>
</tr>
</tbody>
</table>

(a) Table 11.2

(b) Table 11.23
\( c = 0.02 \Rightarrow \hat{Y}^* = 0.5463X_1^* + 0.3774X_2^* - 0.1369X_3^* \)
Bootstrapping

- In many nonstandard situations, (ex: nonconstant error variances estimated by iteratively reweighted least squares), standard methods for evaluating the precision may not be available or may only be approximately applicable when the sample size is large.

- Bootstrapping: provide estimates of the precision of sample estimated for these complex cases.
Bootstrapping (cont.)

**general procedure**: obtain $b_1$ by some procedure and wish to evaluate the precision of $b_1$ by bootstrap method

- The bootstrap method calls for (需要) the selection from the observed sample data of a random sample of size $n$ with replacement.

- the bootstrap sample may contain some duplicate (複製的) data from the original sample and omit some other data in the original sample

- calculates the estimated regression coefficients from the bootstrap sample $\Rightarrow b_1^*$

- repeated a large number of times

- The estimated standard deviation of all of $b_1^* \Rightarrow s^*\{b_1^*\}$ an estimate of the variability of the sampling distribution of $b_1$
Bootstrap Sampling: two basic ways

- When the regression function being fitted is a good model for the data, the error terms have constant variance, and fixed $X$ sampling is appropriate.
  - $e_i$ form the original fitting are regarded as the sample data to be sampled with replacement
  - After a bootstrap sample with $n$:
    
    $e_1^*, \ldots, e_n^*$
    
    $\Rightarrow Y_i^* = \hat{Y}_i + e_i^*$
    
    $Y^*$ regression on $X$s $\Rightarrow b_1^*$
When there is some doubt about the adequacy of the regression function being fitted, the error variance are not constant, and/or random X sampling is appropriate.

- $(X_i, Y_i)$ in the original sample are considered to be the data to be sampled with replacement.
- After a bootstrap sample with $n$:

\[
n \text{pairs: } (X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)
\]

$Y^*$ regression on $X^*$s $\Rightarrow b_1^*$
200-500 bootstrap samples are adequate

One can observe the variability of the bootstrap estimates by $s^*\{b_1^*\}$ as the number of bootstrap samples is increased.

**Bootstrap Confidence Intervals**

- From the bootstrap distribution of $b_1^*$, find the $\alpha/2$ and $1 - \alpha/2$ quantiles $b_1^*(\alpha/2)$ and $b_1^*(1 - \alpha/2)$
- Percentiles from $b_1$:
  
  $b_1^*(\alpha/2) \leq \beta_1 \leq b_1^*(1 - \alpha/2)$

- **reflection method**: require a larger number of bootstrap samples
  
  $2b_1 - b_1^*(\alpha/2) \leq \beta_1 \leq 2b_1 - b_1^*(\alpha/2)$
Other remedial measures

- Section 11.3 Robust regression
  - Least absolute residuals (LAR)-minimum $L_1$-norma regression
- Section 11.4 Nonparametric regression
  - Lowess Method
  - Regression Trees