Chapter 3 Other Issues in Multiple regression (Part 4)

1 Overview

Besides the above two issues (dependent random errors, unequal variance), there are other issues about modelling remedies including

1. Nonlinearity
2. The distribution is not normal
3. collinearity

2 Non-normal distribution: Transformation of the data

1. If X and Y are jointly normally distributed, then their relationship must be linear! Otherwise, it is possible that their relation is nonlinear. One idea is to transform each variable to be approximately normal.
2. By transformation, some nonlinear model can be transformed to linear model. for example,
$$Y_t = \beta_0 X_1^{\beta_1} \exp(\beta_0 + \beta_2 X_2^2)\varepsilon$$
3. For linear regression model, the best transformation is to maximize the R^2.

Suppose Z is a positive random variable. the Box-Cox transformation is

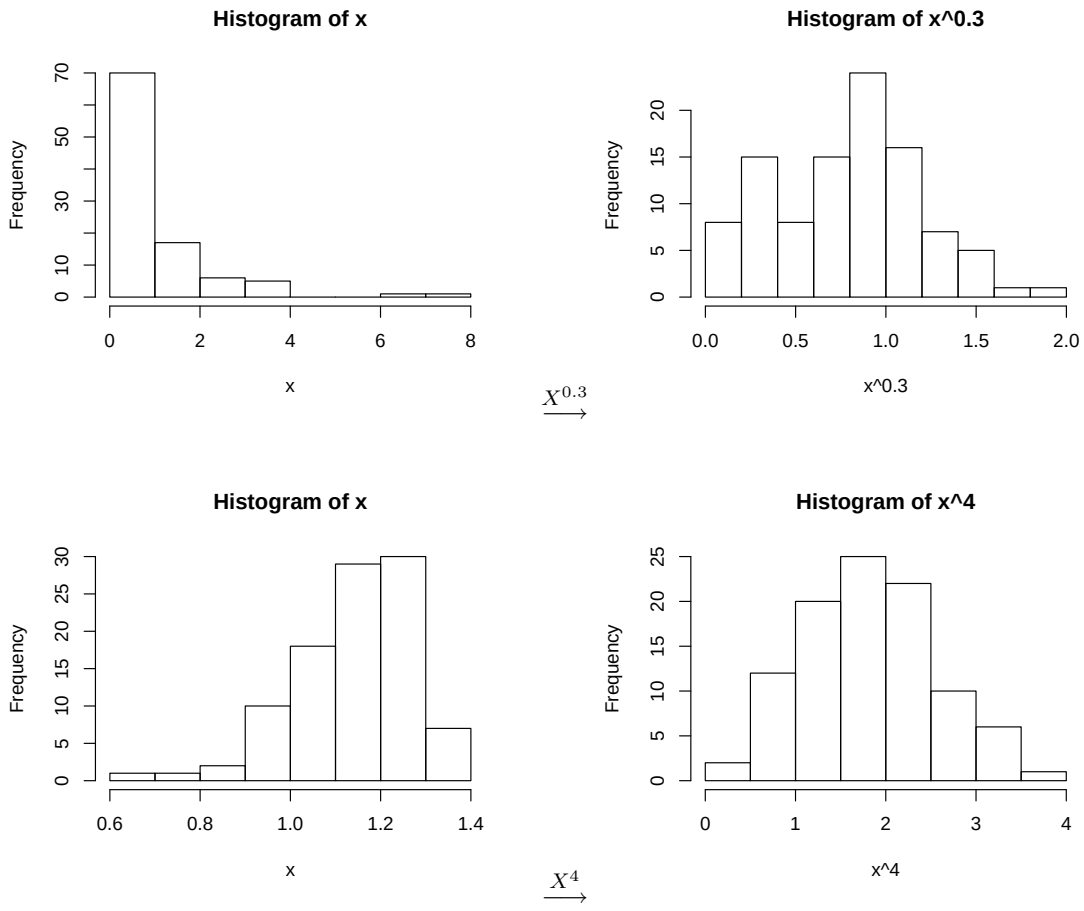
$$\tilde{Z} = \frac{Z^\lambda - 1}{\lambda} = \begin{cases} (Z^\lambda - 1)/\lambda, & \text{if } \lambda > 0 \\ \log(Z), & \text{if } \lambda = 0 \end{cases}$$

where $\lambda \geq 0$, or simply seaking

$$\tilde{Z} = Z^\lambda, \qquad \tilde{Z} = \log(Z).$$

The value of λ are selected to serve the above purpose.

We then applied the linear regression model to the transformed data (instead of the original data).



3 Collinearity: Ride regression

Note that the basic requirement for the Least squares (LSE estimation of a linear regression is

$$(\mathbf{X}'\mathbf{X})^{-1}\text{exists.}$$

There are two reasons that the inverse does not exits. (1) $p > n$ and (2) collinearity. The "badly conditioned linear regression problems" (Hoerl and Kennard, 1970) has long been an important problem in statistics and computer science. The problem is more serious in high dimensional data as most of the genetics data are. The technique of ridge regression (RR) is one of the most popular and best performing (Frank and Friedman, 1993) alternatives to the ordinary least squares (LSE) methods.

A simple way to guarantee the invertibility is adding a diagonal matrix to $\mathbf{X}'\mathbf{X}$, i.e. $\mathbf{X}'\mathbf{X} + \lambda I$, where I is a $(p+1) \times (p+1)$ identity matrix. The ridge regression estimator is

then

$$b_r = (\mathbf{X}'\mathbf{X} + \lambda I)^{-1}\mathbf{X}'\mathbf{Y}$$

where $\lambda > 0$ is a parameter needs to be chosen (HOW? any idea). To make the notation clear, denote the LSE estimator by

$$b_{LSE} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

(providing the inverse exists)

Expectation of b_r

$$\begin{aligned} Eb_r &= E\{(\mathbf{X}'\mathbf{X} + \lambda I)^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{E})\} \\ &= (\mathbf{X}'\mathbf{X} + \lambda I)^{-1}\mathbf{X}'\mathbf{X}\beta = \beta - \lambda(\mathbf{X}'\mathbf{X} + \lambda I)^{-1}\beta \end{aligned}$$

It is not unbiased. The bias is

$$bias(b_r) = Eb_r - \beta = \lambda(\mathbf{X}'\mathbf{X} + \lambda I)^{-1}\beta$$

Recall that

$$Eb_{LSE} = \beta.$$

which is unbiased

Variance-covariance matrix of b_r[1]: If $Var(\mathbf{E}) = \sigma^2 I_n$, then

$$Var(b_r) = (\mathbf{X}'\mathbf{X} + \lambda I)^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda I)^{-1}\sigma^2$$

Recall that

$$Var(b_{LSE}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$$

Estimation Error[2] b_r : $E||b_r - \beta||^2$, also called mean squared error MSE of the estimator, which is different from our MSE of the model.

$$\begin{aligned} E||b_r - \beta||^2 &= E(b_r - \beta)'(b_r - \beta) \\ &= E(b_r - Eb_r + Eb_r - \beta)'(b_r - Eb_r + Eb_r - \beta) \\ &= E||b_r - Eb_r||^2 + 2E\{(Eb_r - \beta)'(b_r - Eb_r)\} + ||Eb_r - \beta||^2 \\ &= E||b_r - Eb_r||^2 + ||Eb_r - \beta||^2 \\ &= tr(Variance) + ||bias||^2 \end{aligned}$$

The LSE has no bias but with a bigger variance than the ridge regression estimator. People proved that we can always find a λ such that

$$MSE(b_r) < MSE(b_{LSE}).$$

In other words, ridge regression can improve the estiamtion of β.

[1]This part will not be included in the examination
[2]This part will not be included in the examination

3.1 An alternative way of understanding ridge regression

The motivation of ridge regression is very simple, but it has good performance. Another way to understand it is that we dont expect an estimator with too large β. Thus, we penalize the value of β. Recall the LSE estimation is to minimize

$$\sum_{i=1}^{n}(Y_i - X_i\beta)^2$$

where $X_i = (1, X_{i1}, ..., X_{ip})$. or

$$\min_{\beta}\sum_{i=1}^{n}(Y_i - X_i\beta)^2$$

To penalize the value of β, we can consider estimate β by minimizing

$$\sum_{i=1}^{n}(Y_i - X_i\beta)^2 + \lambda||\beta||^2.$$

or

$$\min_{\beta}\{\sum_{i=1}^{n}(Y_i - X_i\beta)^2 + \lambda||\beta||^2\}.$$

It is not difficult to prove that to solution of β to the above problem is

$$b_r = (\mathbf{X}'\mathbf{X} + \lambda I)^{-1}\mathbf{X}'Y.$$

Note that with larger λ, the penalty on β tends to be stronger; the solution of β will be smaller.

3.2 Selection of λ via CV

For different λ, we have different ridge regression estimator for the model.

We select a large range for possible λ: $[0, c]$. For each fixed λ in $[0, c]$, consider the CV as follows. For each j,

$$b_r^j(\lambda) = (\sum_{i\neq j} X_i'X_i + \lambda I)^{-1}\sum_{i\neq j} X_i'Y_i.$$

The prediction error for (X_j, Y_j) is

$$err^j(\lambda) = (Y_j - X_j b_r^j(\lambda))^2$$

The CV value is then

$$CV(\lambda) = n^{-1}\sum_{j=1}^{n} err^j(\lambda)$$

The best λ is the minimum point of $CV(\lambda)$.

Example 3.1 (Near Infra-red Calibration for Protein, Fearn (1983)) **(dataA)**. In the data, Y is protein percentage with 6 explanatory variables $X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_6)$, which are log(1/reflectance) values at six wavelengths.

The LSE estimated model is

$$y = 29.372 - 0.1692\mathbf{x}_1 - 0.1536\mathbf{x}_2 + 0.5333\mathbf{x}_3 - 0.1362\mathbf{x}_4 - 0.008\mathbf{x}_5 - 0.0615\mathbf{x}_6$$

Using CV, the selected λ is 4.4. The estimated model is

$$y = 1.8843 + 0.0515\mathbf{x}_1 - 0.22783\mathbf{x}_2 + 0.4726\mathbf{x}_3 - 0.28769\mathbf{x}_4 + 0.0058\mathbf{x}_5 - 0.0154\mathbf{x}_6$$

To check the models, a new experiment was done and the data were collected **(dataB)**

The prediction errors for the new data set are respectively: Least square estimation: 0.09397779; Ridge regression: 0.07783629.

R code for the calculation **(code)**

Example 3.2 (cell classification based on gene) For the leukemia gene expression data (**(training data)**. There are 38 cells with 250 genes (selected from about 7000 genes). they are from two types of cells.

To check the models, a new experiment was done and the data were collected (**(testing set)**.

The prediction errors for the new data set based on Ridge regression: 3.676901e-08. (with ridge parameter $\lambda = 0.05$) From figure 1, we can see that we can have a very accurate classification for the new data.

R code for the calculation **(code)**

3.3 Other selection of λ

Suppose we can obtain the least square estimator $\hat{\beta}$ and estimator of $\hat{\sigma}^2$. then

$$\lambda = \frac{(p+1)\hat{\sigma}^2}{||b_{LSE}||^2}$$

If we cannot get the least square estimator, we can use a ridge regression with very small λ. And get a similar value of λ.

3.4 Extension of ridge regression

The bridge regression proposed by Frank and Friedman (1993) can be written as

$$\min_{\beta} \left\{ n^{-1}(Y - X\beta)'(Y - X\beta) + \lambda \sum_{k=1}^{p} |\beta_k|^{\gamma} \right\}, \tag{1}$$

where $\gamma > 0$. If $\gamma = 2$, it is the ridge regression; if $\gamma = 1$, it is an equivalent of the Lasso proposed by Tibishrani (1996).

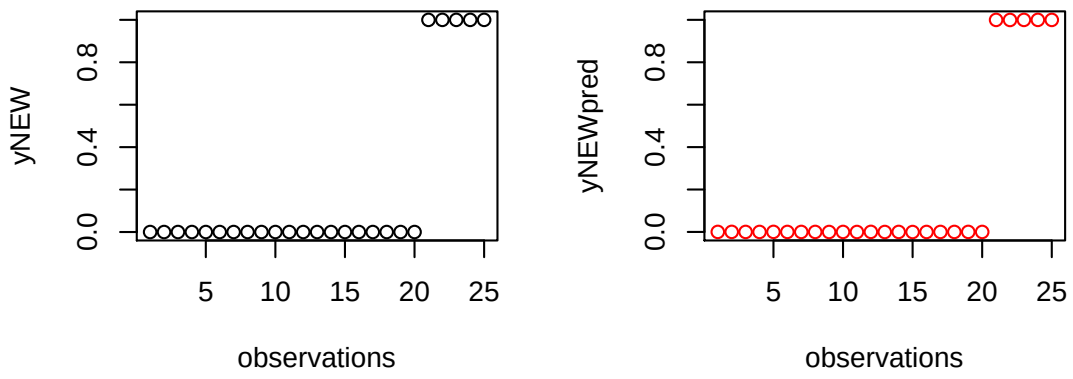


Figure 1: The left panel is the true classification of the 13 cells; the right panel is predicted classification

3.5 Lasso: Least absolute shrinkage and selection operator

If we estimate β by

$$\min_{\beta} \left\{ n^{-1}(Y - X\beta)'(Y - X\beta) + \lambda \sum_{k=1}^{p} |\beta_k| \right\}, \tag{2}$$

the estimation procedure is called Lasso. Lasso simultaneously accomplish model estimation and variable selection.

4 Nonlinearity 1: polynomial models

After we estimate a model, for example,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

we plot the residuals against each predictor, to check whether there is nonlinear patterns. If there is nonlinearity, we can consider polynomial models

$$\begin{aligned} Y_i \quad = \quad & \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} \\ & + \beta_{11} X_{i1}^2 + \beta_{22} X_{i2}^2 + \beta_{33} X_{i3}^2 \\ & + \beta_{12} X_{i1} X_{i2} + \beta_{13} X_{i1} X_{i3} + \beta_{23} X_{i2} X_{i3} \\ & + \varepsilon_i \end{aligned}$$

estimate and refine this model.

We plot the residuals of the model against each predictor, to check whether there is nonlinear patterns. If yes, we consider higher order polynomial models, with terms $X_{i1}^3, ...$

Until there is no nonlinearity in the residuals.

Example 4.1 *for* **data** *with* $Y, X1, X2$. *we have the analysis; see* **code**

5 Nonlinearity 2: Other nonlinear models

Suppose Y_i is the response and $\mathbf{X}_i = (X_{i1}, ..., X_{ip})$ are the predictors. A nonlinear model can be written as

$$Y_i = f(\mathbf{X}_i, \gamma) + \varepsilon_i$$

where $f(.)$ is a specified function, γ is unknown parameters. Examples are

- Linear regression model

$$\begin{aligned} Y_i &= \gamma_0 + \gamma_1 X_{i1} + ... + \gamma_p X_{ip} + \varepsilon_i \\ &= (1, \mathbf{X}_i)'\gamma \end{aligned}$$

  where $\gamma = (\gamma_0, ..., \gamma_p)'$ are parameters and X_i are known constant, ε_i are IID N(0, σ^2).

- Exponential regression models.

$$Y_i = \gamma_0 \exp(\gamma_1 X_i) + \varepsilon_i$$

  where $\gamma = (\gamma_0, \gamma_2)'$ are parameters and X_i are known constant, ε_i are IID N(0, σ^2).

- Logistic regression models

$$Y_i = \frac{\lambda_0}{1 + \gamma_1 \exp(\gamma_2 X_i)} + \varepsilon_i$$

  where $\gamma = (\gamma_0, \gamma_2)'$ are parameters and X_i are known constant, ε_i are IID N(0, σ^2).

5.1 How to select the model

1. By the knowledge of the background; 2. By the plotting

5.2 Least Square Estimation of Nonlinear regression

Estimate the parameters by minimizing the sum of squares of errors

$$Q(\gamma) = \sum_{i=1}^{n} [Y_i - f(\mathbf{X}_i, \gamma)]^2$$

Note that

$$\frac{\partial Q}{\partial \gamma_k} = \sum_{i=1}^{n} -2[Y_i - f(\mathbf{X}_i, \gamma)] \frac{\partial f(\mathbf{X}_i, \gamma)}{\partial \gamma_k}$$

Letting the partial derivatives to be 0, we have the following normal equations

$$\sum_{i=1}^{n} Y_i \frac{\partial f(\mathbf{X}_i, \gamma)}{\partial \gamma_k} - \sum_{i=1}^{n} f(\mathbf{X}_i, \gamma) \frac{\partial f(\mathbf{X}_i, \gamma)}{\partial \gamma_k} = 0, \ k = 0, ..., p$$

Suppose the solution to the N.E. are

$$\mathbf{g} = \begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_p \end{pmatrix}$$

[There is no simple way to find the solution, but numerical methods are available, for example, the Gaussian-Newton methods.]

5.3 R code

package nls2 can be used,

```
library('nls2')
nls(formula, data, start=list(para1 = value1, para2 = value2, ..))
```

Example 5.1 *The yield of a chemical process depends on the temperature X_1 and the pressure X_2 according to the following model*

$$Y_i = \gamma_0 (X_{i1})^{\gamma_1} (X_{i2})^{\gamma_2} + \varepsilon_i$$

The data is observed and available at **(data)**

The estimated model is **(code)**

$$\hat{Y} = 10.08 * X_1^{0.4987} * X_2^{0.3020}$$

5.4 Inference about the coefficients

All the methods are still applicable, including

- Test of $H_0 : \beta_k = 0$ using $t-$test
- Test of $H_0 : \beta_k = 0$ using $F-$test
- Test of reduced model against Full model using F-test, (use $SSE = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$, but not SSR, to calculate F).
- The degree of freedom can be calculated based on the number of normal equations as before.