Chapter 3: Other Issues in Multiple regression (Part 1)

1 Model (variable) selection

The difficulty with model selection: for p predictors, there are 2^p different candidate models. When we have many predictors (with many possible interactions), it can be difficult to find a good model. Model selection tries to simplify this task.

Suppose we have P predictors $X_1, ..., X_P$, but the true models only depends on a subset of $X_1, ..., X_P$. In other words in model

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_P X_P + \varepsilon$$

some of the coefficients are zeros. We need to find those predictors with nonzero coefficients. we call the set of predictors with nonzero coefficients "**best subset**", all the predictors in the "best subset" **important variables**

Criteria: Statistical test; some indices of the model; predictability (Distinction between predictive and explanatory research.)

Example 1.1 (Surgical Unit example) X_1 : blood clotting score; X_2 : Prognostic index; X_3 : enzyme function test score X_4 : liver function test score; X_5 : age in year; X_6 : indicator of gender (0=mail, 1=femail); X_7, X_8 indicator for alcohol use; Y: survival time.

If we only consider the first 4 predictors, we have the following calculation for the possible models

variables selected	p'	SSE	R^2	R_a^2	C_p	AIC	SBC (BIC)	PRESS (CV)
None	1	12.808	0	0	151.4	-75.7	-73.7	13.3
X1	2	12.0	0.06	0.043	141	-77	-73	13.5
X2	2	9.98	0.21	0.21	108.5	-87.17	-83.2	10.74
X3	2	7.3	0.428	0.417	66.49	-103.8	-99.84	8.32
X4	2	7.4	0.422	0.410	67.715	-103.26	-99.28	8.025
X1, X2	3	9.44	0.26	0.23	7102.037	-88.16	-82.19	11.06
X1, X3	3	5.71	0.549	0.531	43.85	-114.65	-108.69	6.98
X1, X4	3	7.29	0.43	0.408	67.97	-102.067	-96.1	8.472
X2, X3	3	4.312	0.663	0.65	20.52	-130.48	-124.5	5.065
X2, X4	3	6.62	0.483	0.463	57.21	-107.32	-101.357	7.476
X3, X4	3	5.13	0.6	0.58	33.5	-121.1	-115.146	6.12
X1, X2, X3	4	3.109	0.757	0.743	3.391	-146.161	-138.2	3.91
X1, X2, X4	4	6.57	0.487	0.456	58.39	-105.74	-97.79	7.9
X1, X3, X4	4	4.9	0.61	0.589	32.93	-120.8	-112.88	6.2
X2, X3, X4	4	3.6	0.718	0.7	11.42	-138.023	-130.067	4.597
X1, X2, X3, X4	5	3.08	0.759	0.74	5.00	-144.59	-134.65	4.07

where p' is the number of coefficients included in the model.

2 R^2 and R_a^2 Criterion

1. R^2 : can be used for models with the same number of parameters/coefficients.

2. R_a^2 : can be used for models with Different number of parameters/coefficients.

We need to choose a model with the biggest R_a^2.

In the above example, model with X_1, X_2, X_3 is selected by this criterion.

3 Mallows' C_p Criterion

Suppose we select p predictors, $p \leq P$ and try a model with the selected predictors. denote its SSE by $SSE_{p'}$. The criterion is

$$C_p = \frac{SSE_{p'}}{MSE(X_1, ..., X_P)} - (n - 2p')$$

where p' is the number of coefficients including intercept (if there is).

Criterion: We seek to identify subsets of X for which (1) the C_p values is small and (2) the C_p vale is near p'.

- If a selected model includes all the important variables (But with some other unimportant variables), the model is still correct. Then we have

  $$E\{SSE_{p'}\} = (n - p')\sigma^2$$

  On the other hand

  $$E\{MSE(X_1, ..., X_P)\} = \sigma^2$$

  Roughly speaking, we have

  $$C_p \approx n - p' - (n - 2p') = p'$$

  Question: are the estimators still unbiased?

- If a selected model does not include all the important variables, the model is wrong. Then

  $$SSE_p >> SSE_P$$

  $$C_p >> n - p' - (n - 2p') = p'$$

  Question: are the estimators still unbiased?

In the above example, model with X_1, X_2, X_3 is selected by this criterion.

4 Akaike's information criterion (AIC)

We cannot use SSE alone for the selection. As p' increases, $SSE_{p'}$ decreases. AIC try to balance the number of parameters and $SSE_{p'}$.

$$AIC_p = \log(\frac{SSE_{p'}}{n}) + \frac{2p'}{n}$$

or

$$AIC_p = n\log(\frac{SSE_{p'}}{n}) + 2p'$$

In the above example, model with X_1, X_2, X_3 is selected by this criterion.

5 Schwarz' Bayesian criterion (BIC or SBC)

Theoretically, people find that AIC does not give a right number of variables. Schwarz proposed the BIC

$$BIC_p = \log(\frac{SSE_{p'}}{n}) + \log(n)\frac{p'}{n}$$

or

$$BIC_p = n\log(\frac{SSE_{p'}}{n}) + \log(n)p'$$

BIC gives bigger penalty to the number of parameters

In the above example, model with X_1, X_2, X_3 is selected by this criterion.

6 Prediction sum of squares (PRESS) or Cross-validation criterion (CV)

A better model should have better prediction. Most of the time, we dont have a data for us to predict. A simple way is to partition the data to two parts: training samples (set) and prediction set (or validation set). Use training set to estimate the model and prediction set to check the predictability. A simple case that each time, the prediction set has one sample in turn. There are many partitions. Using all the partitions is the idea of cross-validation (CV). The idea was proposed by M. Stone (1974).

If we use 1 observation for validation and the other n-1 for model estimation, it is the leave-one-observation-out cross-validation

If we use m observations for validation and the other n-m for model estimation, it is the leave-m-observation-out cross-validation.

We need to select variables from $X_1, ..., X_p$ to be included in the model. There are many candidate variables. For example,

$$\begin{aligned}
&model\ 1: && Y = a_0 + a_1X_1 + \varepsilon \\
&model\ 2: && Y = b_0 + b_1X_1 + b_2X_4 + \varepsilon \\
&model\ 3: && Y = c_0 + c_1X_2 + \varepsilon \\
&\quad\ \cdots
\end{aligned}$$

Suppose we have n samples. For each i = 1, ..., n, we use data (Y_1, X_1), ..., (Y_{i-1}, X_{i-1}), (Y_{i+1}, X_{i+1}), ...(Y_n, X_n), where $X_i = (X_{i1}, ..., X_{iP})$, to estimate the models. the estimated models are, say,

$$\begin{aligned}
&model\ 1: \quad Y = \hat{a}_0^i + \hat{a}_1^i X_{i1} \\
&model\ 2: \quad Y = \hat{b}_0^i + \hat{b}_1^i X_{i1} + \hat{b}_2^i X_{i4} \\
&model\ 3: \quad Y = \hat{c}_0^i + \hat{c}_1^i X_{i2} \\
&\qquad \cdots
\end{aligned}$$

The prediction errors for (Y_i, X_i) are respectively

$$\begin{aligned}
&model\ 1: \quad err_1(i) = \{Y_i - \hat{a}_0^i - \hat{a}_1^i X_{i,1}\}^2 \\
&model\ 2: \quad err_2(i) = \{Y_i - \hat{b}_0^i - \hat{b}_1^i X_{i,1} - \hat{b}_2^i X_{i,4}\}^2 \\
&model\ 3: \quad err_3(i) = \{Y_i - \hat{c}_0^i - \hat{c}_1^i X_{i,2}\}^2 \\
&\qquad \cdots
\end{aligned}$$

The overall prediction errors (also called Cross-validation value) are respectively then

$$\begin{aligned}
&model\ 1: \quad CV_1 = n^{-1} \sum_{i=1}^{n} err_1(i) \\
&model\ 2: \quad CV_2 = n^{-1} \sum_{i=1}^{n} err_2(i) \\
&model\ 3: \quad CV_3 = n^{-1} \sum_{i=1}^{n} err_3(i) \\
&\qquad \cdots
\end{aligned}$$

The model with the smallest CV value is the model we prefer.

Example 6.1 For the same data above **(data)** Our candidate models are

$$\begin{array}{ll} \text{model 0} & Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon \\ \text{model 1} & Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \\ \text{model 2} & Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 X_5 + \varepsilon \\ \text{model 3} & Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon \\ \text{model 4} & Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon \\ \text{model 5} & Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon \end{array}$$

The CV values for the above model are respectivly

$$CV(\text{model 0}) = 0.3633548, CV(\text{model 1}) = 0.333161, CV(\text{model 2}) = 1.216745,$$

$$CV(\text{model 3}) = 0.3922781, CV(\text{model 4}) = 1.400237, CV(\text{model 5}) = 0.4589498$$

Thus model 1 is selected (and variable X_5 is deleted)

R-code for the calculation

K-fold cross-validation In K-fold cross-validation, the original sample is partitioned into K subsamples. Of the K subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $K - 1$ subsamples are used as training data. The cross-validation process is then repeated K times (the folds), with each of the K subsamples used exactly once as the validation data. The K results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used.

7 Searching for the "best subset"

- Forward selection: starting with no variables in the model, trying out the variables one by one and including them if they are 'statistically significant' or can increase the predictability.

- Backward elimination: starting with all candidate variables and testing them one by one for statistical significance, deleting any that are not significant or can increase the predictability.
- Stepwise: a combination of the above, testing at each stage for variables to be included or excluded.

8 R code

```
step(object, direction = c("both", "backward", "forward"), steps = 1000, k =
??)
```

where k can be any positive values, but $k = 2$ for AIC, and $k = \log(n)$ for BIC (SBC)

Example 8.1 For the first example above with **data**, the selected model variables are

$$\text{Based on BIC:} \quad X1 + X2 + X3 + X5 + X6 + X8$$

or

$$\text{Based on BIC:} \quad X1 + X2 + X3 + X8$$

(code)