Handout 3

Model Selection

Model building/selection is an indispensable part of any statistical analysis of data. We will briefly discuss a few methods which are useful in regression and time series modeling. We will write these down in the regression context, but these same procedures also work for time series modeling. The goal of any model selection methods is to choose an appropriate one from a given class of candidate models.

We have three independent variables (X_1=income, X_2=persons, X_3=area) and the dependent (or the response) variable is the electric bill (Y). We will look at all possible regression models.

Models to be considered:

i) no predictor variable: $Y = \beta_0 + \varepsilon$

ii) regression with one predictor variable:

$Y = \beta_0 + \beta_1 X_1 + \varepsilon$,

$Y = \beta_0 + \beta_2 X_2 + \varepsilon$,

$Y = \beta_0 + \beta_3 X_3 + \varepsilon$,

iii) regression with two predictor variables:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$,

$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$,

$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$.

iv) regression with three predictor variables: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$.

The goal is to select the most appropriate of these $2^3 = 8$ models. For any model with $p-1$ predictor variables, the number of regression parameters (beta parameters) estimated is equal to p and we will denote the residual sum of squares by SSE_p. Note that

$df(SSE_p) = n - p$,

$MSE_p = SSE_p/(n-p)$,

$R^2_p = SSR_p/SSTO = 1 - SSE_p/SSTO$,

$R^2_{adj,p} = 1 - MSE_p/MSTO = 1 - \frac{n-1}{n-p}(1 - R^2_p)$.

As more variables are included in the model, SSE_p will decrease and R^2_p will increase. So neither SSE_p nor R^2_p can be used to select an appropriate model. One may decide to select the model for which MSE_p is the smallest (or equivalently $R^2_{adj,p}$ is the largest). However, MSE_p (or $R^2_{adj,p}$) criterion is not necessarily appropriate always.

Here are some other well known model selection criteria. For any of these methods, model selection is done by selecting the model at which the criterion function is the smallest. [note: "log" denotes the natural logarithm]

$$\text{Akaike's final prediction error} : FPE_p = \frac{n+p}{n-p} SSE_p,$$

$$\text{Mallows' } C_p = \frac{SSE_p}{MSE_{p\,\max}} - (n-2p),$$

where $MSE_{p\,\max}$ is the mean square error of the largest model under consideration (here $MSE_{p\,\max} =$ 18339),

$$\begin{aligned} \text{Akaike's information criterion} : AIC_p &= n\log(SSE_p) - n\log(n) + 2p, \\ \text{Corrected AIC criterion} : AICC_p &= n\log(SSE_p) - n\log(n) + 2p + 2p(p+1)/(n-p-1) \\ &= AIC_p + 2p(p+1)/(n-p-1), \\ \text{Bayesian information criterion} : BIC_p &= n\log(SSE_p) - n\log(n) + [\log(n)]p. \end{aligned}$$

Variables in model	p	SSE_p	R_p^2	MSE_p	$R^2_{adj,p}$	FPE_p	C_p	AIC_p	$AICC_p$	BIC_p
None	1	3701668	0	112172	0	3926020	169.25	396.3	396.5	397.9
X_1	2	1109145	.700	34661	.691	1247796	30.48	357.4	357.7	360.4
X_2	2	2797872	.244	87433	.221	3147588	122.56	388.8	389.2	391.9
X_3	2	669609	.819	20925	.813	753300	6.41	340.2	340.6	343.2
X_1, X_2	3	578508	.844	18662	.834	690494	3.55	337.2	338.0	341.8
X_1, X_3	3	616131	.834	19875	.823	735375	5.60	339.4	340.2	343.9
X_2, X_3	3	555753	.850	17928	.840	663336	2.30	335.9	336.7	340.4
X_1, X_2, X_3	4	550163	.851	18339	.837	696882	4.00	337.5	338.9	343.6

According to all the six criteria, the most appropriate model seems to be

$\hat{Y} = -255.95 + 42.03X_2 + .40429X_3$, $R^2 = .850$, $R^2_{adj} = .840$.

Remarks

1. In time series, the last three criteria (i.e., AIC, $AICC$ and BIC) are usually employed for model selection.

2. It is important to point out that $\hat{\sigma}_p^2 = SSE_p/n$ is known as the maximum likelihood estimate of σ^2. And the criteria functions can be re-written as

$$\begin{aligned} AIC_p &= n\log(\hat{\sigma}_p^2) + 2p, \ BIC_p = n\log(\hat{\sigma}_p^2) + [\log(n)]p, \\ AICC_p &= n\log(\hat{\sigma}_p^2) + 2p + 2p(p+1)/(n-p-1) \\ &= AIC_p + 2p(p+1)/(n-p-1). \end{aligned}$$

3. For any of the criteria $AIC, AICC$ and BIC, the term $n\log(n)$ does not depend on p and hence it plays no role in selecting the appropriate model. For this reason, in some packages criteria are written without the $n\log(n)$ term:

$$\begin{aligned} AIC_p &= n\log(SSE_p) + 2p,\ BIC_p = n\log(SSE_p) + [\log(n)]p, \\ AICC_p &= n\log(SSE_p) + 2p + 2p(p+1)/(n-p-1) \\ &= AIC_p + 2p(p+1)/(n-p-1). \end{aligned}$$