2.10. Model selection – Criteria and Search Procedures

Criteria for Model selection

- Adj- R^2
- Mallow's C_p
- t test p values
- AIC and BIC

Automatic Search Procedures

- Best Subset
- Stepwise
- Forward, Backward or both

↑ commonly seen in practice.

So far by talking about t test, anova., R^2 and adj-R^2, we have a taste that choosing what predictors to be left in the model is not always obvious.

- When we use different criterias, it might lead to different candidates.

Another difficulty might be when we have a large number of predictors, there are too many candidate models. For ex, 10 ~~predictors~~ parameters → $2^{10} = 1024$ possible regression models.

so some automatic search proc procedures are needed to help us.

- Best subset would fit all possible models but could be a huge computational burden

- Stepwise is simpler but might "miss" the best model. since it's just adding / dropping predictors in one-way, and stops when it meets the criteria.

1. Selection Criterias

(i) $\text{adj-}R^2 = 1 - \left(\frac{n-1}{n-p}\right) \cdot \frac{SSE}{SST}$

takes a balame of ~~gain in SSE~~ decrease in SSE and decrease in $df = n-p$ when adding more predictors

(ii) Mallow's Cp.

The Mean Square Error for $\hat{Y}_i$ is defined as (← dist. version, not sample version.)

$$E(\hat{Y}_i - \mu_i)^2$$

where $\mu_i = E(Y_i)$

The bias of $\hat{Y}_i - \mu_i$ can be broken into

$$\hat{Y}_i - \mu_i = (E(\hat{Y}_i) - \mu_i) + (\hat{Y}_i - E(\hat{Y}_i))$$

$(E(\hat{Y}_i) - \mu_i)$: when the model is incorrect, there is a bias

$(\hat{Y}_i - E(\hat{Y}_i))$: $\hat{Y}_i$ based on a sample data has a bias from $E(\hat{Y}_i)$

Then it can be shown that

$$E(\hat{Y}_i - \mu_i)^2 = [E(\hat{Y}_i) - \mu_i]^2 + Var(\hat{Y}_i)$$

And the total mean square error for all observations is

$$\sum_{i=1}^{n} [E(\hat{Y}_i) - \mu_i]^2 + \sum_{i=1}^{n} Var(\hat{Y}_i)$$

Defined the "standardized" version as

$$P_p = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\left[E(Y_i) - \mu_i\right]^2 + \sum_{i=1}^{n} Var(Y_i)\right]$$

[OPTIONAL]

$$SSE = \Sigma (Y_i - \hat{Y}_i)^2$$

$$= (\vec{Y} - \vec{X}\vec{b})^T (\vec{Y} - \vec{X}\vec{b})$$

$$= \vec{Y}^T \underbrace{(I_n - \vec{X}(\vec{X}^T\vec{X})^{-1}\vec{X}^T)}_{\vec{A}} \vec{Y}$$

By using a result for quadratic forms:

$$E(\vec{Y}^T\vec{A}\vec{Y}) = E(\vec{Y}^T)\vec{A}E(\vec{Y}) + tr(\Sigma A)$$

where $\Sigma = Var(\vec{Y})$

$$E(SSE) = E(\vec{Y}^T)(I_n - \vec{X}(\vec{X}^T\vec{X})^{-1}\vec{X}^T)E(\vec{Y})$$

$$+ tr[I_n - \vec{X}(\vec{X}^T\vec{X})^{-1}\vec{X}^T]\sigma^2$$

$$= \sum_{i=1}^{n} (E(y_i) - E(\hat{y}_i))^2 + \sigma^2 (n - \underline{tr[(\vec{X}^T\vec{X})(\vec{X}^T\vec{X})^{-1}]})$$

$= tr(I_p)$

$$= \sum_{i=1}^{n} (E(\hat{y}_i) - \mu_i)^2 + \sigma^2 (n-p)$$

$= p$

$$Var(\hat{y}_i) = cov(\hat{y}_i, \hat{y}_i) = cov\left(\sum_{j=1}^{n} h_{ij} y_j, \sum_{k=1}^{n} h_{ik} y_k\right)$$

$$= \sum_{j=1}^{n} \sum_{k=1}^{n} h_{ij} h_{ik} cov(y_j, y_k)$$

$$= \sum_{j=1}^{n} h_{ij}^2 Var(y_j)$$

$$= \sigma^2 \sum_{j=1}^{n} h_{ij}^2 = \sigma^2 h_{ii}$$

$\sum_{j=1}^{n} h_{ij}^2 = h_{ii}$

because $HH = H$

$$\sum_{i=1}^{n} Var(\hat{y}_i) = \sigma^2 \Sigma h_{ii}$$

$$= \sigma^2 tr(H)$$

$$= \sigma^2 p$$

$tr(H) =$

$tr(X(X^TX)^{-1}X^T)$

$= tr(X^TX(X^TX)^{-1})$

$= tr(I_p)$

$= p$

Therefore,

$$P_p = \frac{1}{\sigma^2}\left[E(SSE) - (n-p)\sigma^2 + p\sigma^2\right]$$

$$= \frac{E(SSE)}{\sigma^2} - (n-2p)$$

when the fitted model is accurate, $E(SSE) = (n-p)\sigma^2$

then $P_p = P$ □

- In practice, we have a total of $(p-1)$ predictors (p parameters) and Mallow's C_p is calculated for each model with $k-1 \leq p-1$ predictors. (k parameters):

$$C_p = \frac{SSE\_\text{fitted model with}(k-1)\text{predictors}}{MSE\_\text{full model with}(p-1)\text{predictors}} - (n-2k)$$

  - If the model with the chosen $k-1$ predictors is "good" model, AKA. MSE_{k-1} is close to MSE_{p-1} C_p should be around k.

- In practice, since C_p is a sample estimate, it could be less than p. $k+1$

  When $C_p < 1$, it's doubtful, might suggest overfitting

- The commonly used $C_p = p$ when including all possible predictors, so it can not be used to evaluate the full model, only the model with less than $p-1$ predictors.

- It's recommended to use Mallow's C_p to choose some initial candidates models.

(iii) use t test p-values: less common.

(iv) AIC and BIC – Note: Textbook gives a version with modified constants, what we introduced here is more common.

- Akaike's Information Criterion
- Bayesian Information Criterion (or Schwartz's Bayesian SBC)

They both measure the ~~likelihood of the model~~ amount of information LOST by a given model:

the less information a model loses, the higher the quality of that model

(a) $AIC = 2K - 2\ln(\hat{L})$

K: number of parameters

$\hat{L}$: maximum likelihood function

- Recall: For MLR, the likelihood function

$$L(\beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_i - \mu_i(\beta))^2}{2\sigma^2}\right]$$

To maximize the likelihood function, it's equivalent to maximize

$$\ell(\beta, \sigma) = \ln(L) = -n\log\sigma - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (X\beta))^2$$

and $\hat{\beta}_{MLE} = \hat{\beta}_{OLS}$, $\hat{\sigma}^2_{MLE} = \frac{SSE}{n}$, $\ln(\hat{L})$

Note: $\ln(\hat{L})$ is negative

the maximized value is the minimized

$-\ln(\hat{L})$, which is positive.

In fact, $-2\ln(\hat{L}) \sim \chi^2_{df=n-p}(a)$ non-central chi-square.

- In summary, a smaller value of $-2\ln(\hat{L})$ indicates a better goodness of fit:

$$\to -\ln(\hat{L}) = \cancel{n\log\hat{\sigma}^2}\ n\ln\hat{\sigma}^2 + n\ln 2\pi + \frac{1}{\hat{\sigma}^2}SSE$$

where $\hat{\sigma}^2 = \frac{SSE}{n}$

- AIC consider the penalty of adding predictors.:

$$AIC = -2\ln(\hat{L}) + 2K$$

↑ penalty coefficient on number of parameters

- with the same number of predictors,

smallest AIC has the biggest likelihood.

- with the same likelihood,

AIC would choose (smaller) the model with less predictors.

(b) Through complicated calculation, Gideon E. Schwarz gave a Bayesian argument to give a heavier penalty on the number of parameter, under the consideration of sample size

$$BIC = -2\ln(\hat{L}) + \underbrace{\ln n}_{\text{penalty coefficient on number of parameters}} \cdot K$$

- AIC and BIC theoritically have different set-up:

  AIC is based on likelihood function, BIC is estimating a function of a posterior probability of a model being true. Both criteria are based on various assumptions and asymptotic approximations.

- In practice, the only difference is in the size of the penalty. BIC penalizes model complexity more heavily.

- AIC is better when we have false negative finding more

  BIC is better when we have false positive finding more

II. Automatic Search Procedures.

No matter which criteria we choose, there are two most common automatic search procedures:

(a) Best Subsets:

It fits all 2^p models and print the best fitting model(s) with one predictor, two predictors, three predictors and so on (the one with only intercept) is normally ignored)

— Good to give a list of candidates

Ex. $y \sim x_1 + x_2 + x_3 + x_4$

Number of Predictors	adj-R^2	Mallows Cp	x_1	x_2	x_3	x_4
1	67.5	138.7				X
1	64.5	142.5		X		
2	97.4	2.7	X	X		
2	97.7	5.5	X			X
3	97.6	3.0	X	X		X
3	97.6	3.0	X	X	X	
4	97.4	3.0	X	X	X	X

adj-R^2 choice

mallow Cp choice

(b) Stepwise – less seen in practice.

Basic idea: adding or removing variables based on t test or anova test ~~result~~. p-value

Before starting: α_E: to-enter significant level

α_R: to-remove significant level

Procedure: Using $Y \sim X_1 + X_2 + X_3$ as example.

(i) Fit SLR: $Y \sim X_1$, $Y \sim X_2$, $Y \sim X_3$

If none of the pvalue $\leq \alpha_E$ → pick $Y \sim$ Null → stop

---

For those with p-value $\leq \alpha_E$

choose the one with smallest p-value.

i.e. X_1

(i) Fit: $Y \sim X_1 + X_2$, $Y \sim X_1 + X_3$

If none of the 2nd pvalue $\leq \alpha_E$ → pick $Y \sim X_1$ → stop

---

For the 2nd predictor with p-value $\leq \alpha_E$

pick the 2nd predictor with the smallest p-value to enter the model.

p-value of $X_1 > \alpha_R$ → Remove X_1 → $Y \sim X_2$

then check p-value of X_1

↓ pvalue of $X_1 \leq \alpha_R$

$Y \sim X_1 + X_2$

(c) StepAIC. StepBIC

- R comes with stepAIC with options "both", "forward", "Backward"
- Python doesn't have build-in automatic search. (yet)

$$y \sim X_1 + X_2 + X_3$$

- Backforward Starting with $y \sim X_1 + X_2 + X_3$

(i) Drop one:

X_3 :	$y \sim X_1 + X_2$	63.8	lowest
X_2 :	$y \sim X_1 + X_3$	65.8	
X_1 :	$y \sim X_2 + X_3$	75.8	

(ii) start with $y \sim X_1 + X_2$

Drop one:

X_1 :	$y \sim X_2$	73.1
X_2 :	$y \sim X_1$	78.4

Stop. pick $y \sim X_1 + X_2$

- Forward starting with $y \sim$ Null and add one.:

(i) $y \sim x_1$ 73.1

$y \sim x_2$ 78.4

$y \sim x_3$ 70.1

(ii) $y \sim x_2 + x_3$ ~~75.8~~

$y \sim x_1 + x_3$ 65.8

stop: pick $y \sim x_1 + x_3$

- Direction = both.: each step would add back the one deleted from previous step to see if it increase/decreasing AIC/BIC.

(i) starting with $y \sim x_1 + x_2 + x_3 + x_4$ AIC = 56.28

$-x_1$: $y \sim x_2 + x_3 + x_4$ 54.44

$-x_2$: $y \sim x_1 + x_3 + x_4$ 54.88

$-x_3$: $y \sim x_1 + x_2 + x_4$ 56.28

$-x_4$: $y \sim x_1 + x_2 + x_3$ 58.22

$-x_4$: $y \sim x_2 + x_3$ 49.10

(ii) Starting: $y \sim x_2 + x_3$

$+x_1$: $y \sim x_1 + x_2 + x_3$

$-x_2$: $y \sim x_3$

$-x_3$: $y \sim x_2$

$+x_4$: $y \sim x_2 + x_3 + x_4$

(ii) starting with $y \sim x_2 + x_3 + x_4$ ... until find the lowest AIC.

$+x_1$: $y \sim x_1 + x_2 + x_3 + x_4$ 56.28 add it back

$-x_2$: $y \sim x_3 + x_4$ 52.11

$-x_3$: $y \sim x_2 + x_4$ 53.49