Model Selection / Variable Selection,

Discrepancy Function Approach

Limitations of Hypothesis Testing

- only two (nested) models can be tested at a time
- "burden of proof" is on larger model

goal of discrepancy function approach

- Choose the "best" linear regression model from among a candidate class of models

- We consider the best model that which provides the most accurate estimation of $E(Y_i) = \mu_i$ at the input levels $\underline{x}_1, \dots, \underline{x}_n$. (most accurate prediction of $\underline{Y}_i$ at input levels.)

- The two sources of model error we need to consider are model misspecification (bias) and parameter estimation (variance). (bias/variance tradeoff) truth : $E\{Y_i\} = \mu_i$

approximating model: $\hat{Y}_i = \underline{x}_i' \hat{\underline{\beta}}$ (perhaps only including a subset of the available predictor variables)

squared error of estimation: (discrepancy function)

$$\Delta(\underline{\mu}, \hat{\underline{\beta}}) = \sum_{i=1}^{n} (\underline{x}_i' \hat{\underline{\beta}} - \mu_i)^2$$

$$= \sum_{i=1}^{n} (\hat{Y}_i - \mu_i)^2$$

where

$\hat{Y}_1, \ldots, \hat{Y}_n$ are the fitted values for a candidate model.

Note that

$\Delta(\mu, \hat{\underline{\beta}})$ is not a statistic (not observable)

$\Delta(\underline{\mu}, \hat{\underline{\beta}})$ is not a parameter (depends on sample)

Instead, we define the expected discrepancy as

$$\Delta = \sum_{i=1}^{n} E\{(\hat{Y}_i - \mu_i)^2\}$$ (mean squared error of estimation)

The candidate model for which Δ is minimum is considered the "best".

Demonstration:

- Consider the problem of deciding between two models:

$$(F): \quad E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

$$(R): \quad E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Our goal is not determining whether $\beta_4 = \beta_5 = 0$ or not, our goal is determining whether $\Delta(R) < \Delta(F)$ or not.

- Suppose the true model is

$$E(Y) = 7 + X_1 + X_2 + \frac{1}{2} X_3 + \frac{1}{10} X_4 + \frac{1}{10} X_5$$

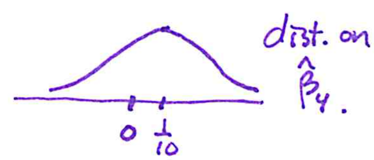
Then (F) is correctly specified, but (R) is not.

(i.e., $H_0: \beta_4 = \beta_5 = 0$ is not true)

- Data is available from a 2^5 design, ~~with~~ $(n=32)$ with error variance $\sigma^2 = 4$.

It can be shown that $\Delta(R) < \Delta(F)$. why?

A model coefficient near zero is better estimated as zero than by an estimate computed from the data.



(model selection depends on the data available)

④

Write

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

i.e.,

$$MSE = \text{variance} + (\text{bias})^2$$

So,

$$\Delta(M) = \sum_{i=1}^{n} Var(\hat{Y}_i) + \sum_{i=1}^{n}\left(E(\hat{Y}_i)-\mu_i\right)^2$$

(estimation variance) (approximation bias)

<u>Fact:</u> $\sum_{i=1}^{n} Var(\hat{Y}_i) = p\sigma^2$, where $\dim(\underline{\hat{\beta}}) = p$

Write $\underline{\underline{\Delta(M) = \sigma^2(p+\delta)}}$

In a model selection problem, the "best" model is the one for which $\underline{p+\delta \text{ is minimum.}}$

Model Selection Criterion

select model M from the candidate class

for which $p + \hat{\delta}$ is minimized.

Fact: Let M_F denote the largest candidate model.

Let M denote a candidate model with p regression parameters.

We must assume that M_F is correctly specified.

Then

$$C_p = \left(\frac{SSE(M)}{MSE(M_F)} - n \right) + 2p$$

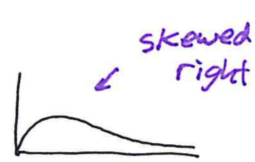
(Mallows' conceptual predictive statistic)

is approximately unbiased for $p + \delta$.

Aside: $\hat{\delta} = (P_F - P)(F^* - 1)$ is an estimate

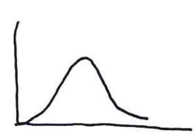
for the model bias δ. Then $C_p = p + \hat{\delta}$.

example: goal is to predict survival in patients undergoing a particular type of liver operation



response y is survival time in days.

model $\log(y)$ as distributed Normal



$(Y \sim LOGN(\mu, \sigma^2))$

(6)

input variables for surgical unit example

X_1 - blood clotting score

X_2 - prognostic index

X_3 - enzyme test

X_4 - liver test

X_5 - age

continuous predictor variables

X_6 - sex $\begin{pmatrix} 0 = \text{male} \\ 1 = \text{female} \end{pmatrix}$

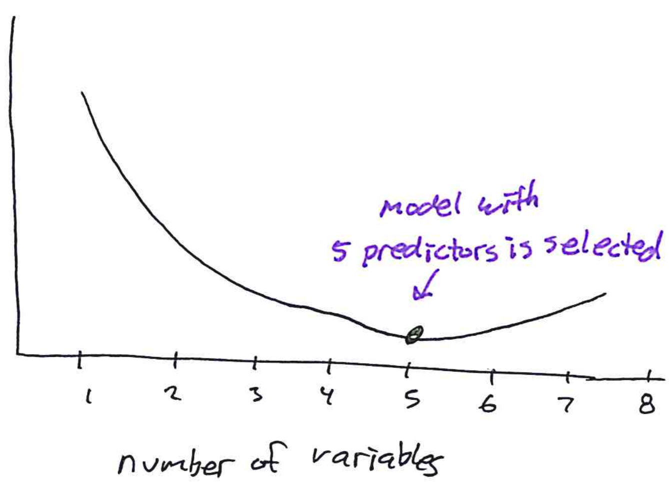
X_7, X_8 - alcohol use

$$(X_7, X_8) = \begin{cases} (0,0) \text{ if none} \\ (1,0) \text{ if moderate} \\ \cancel{0, \text{ if none}} \\ (0,1), \text{ if heavy} \end{cases}$$

categorical predictor variables

candidate class: $2^8 = 256$ different models

R : leaps package, full.mod = regsubsets(log.Y ~ .) include all predictors



models with the same number of parameters are ordered based on SSE.

			C_p
M_6 :	X_1, X_2, X_3, X_6, X_8	selected model	5.54
M :	X_1, X_2, X_3, X_8	closest models	5.75
M_{56} :	$X_1, X_2, X_3, X_5, X_6, X_8$		5.78

How does C_p model selection compare to hypothesis testing?

Consider the problem of testing M_F against M_R.

Let F^* denote the general linear test statistic.

Then

$$\underline{C_p(M_R) < C_p(M_F)} \iff \underline{F^* < 2}$$

~~[illegible]~~ approximately

For testing one parameter, we get the rule:

Accept M_R iff $|t^*| < 1.414$

Accept M_F iff $|t^*| > 1.414$ (p-value < .15)

The rule for adding predictors to a model is less stringent for discrepancy based model selection than the rule for hypothesis testing.

M: X_1, X_2, X_3, X_8

anova (M, M6) : $F^* = 2.23$ ($p = .1418$)

C_p rule decides in favor of adding X_6 to the model.

anova (M6, M56) : $F^* = 1.80$ ($p = .1862$)

C_p rule decides against adding X_5 to the model