Model (variable) selection

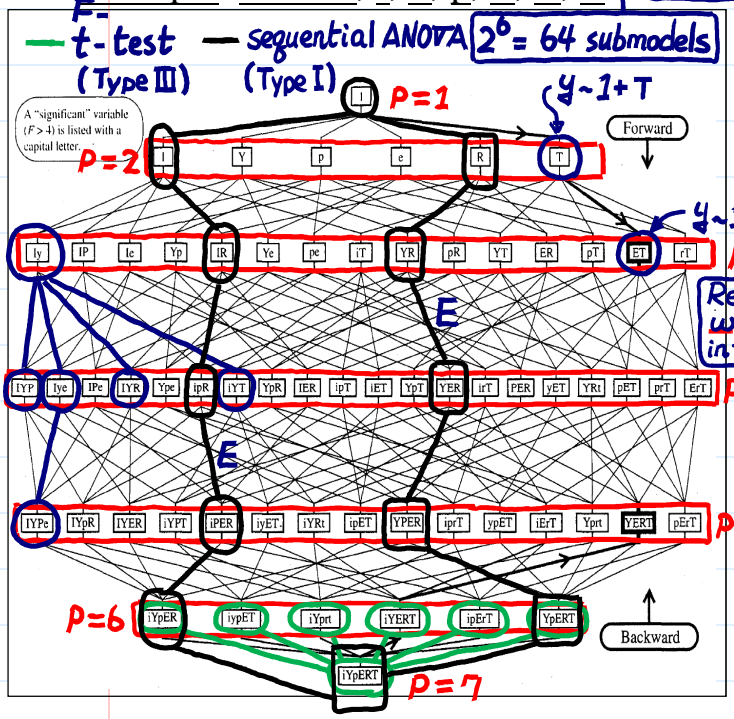
Recall: ① Lack-of-fit (underfitting) ② overfitting ← t-test, F-test

- Q: what is a "model selection" problem?
  - consider the *full model*: (effect (term)) (iid $N(0,\sigma^2)$)

$$y=\beta_0+\beta_1 g_1(x_1,\ldots,x_m)+\beta_2 g_2(x_1,\ldots,x_m)+\cdots+\beta_{k-1} g_{k-1}(x_1,\ldots,x_m)+\epsilon$$

For $1\le i\le k-1$, should the term $\beta_i g_i$ be included in the *final fitted model*?

Example: 6 terms, I, Y, p, E, R, T (model space)



- (sub-)model: a subset set of all $k-1$ terms, e.g., $y\sim 1+g_1+g_2$, $y\sim 1+g_2+g_4+g_5+g_{k-1}$; always include intercept in these submodels: $\{1, g_1, g_2\}$, $\{1, g_2, g_4, g_5, g_{k-1}\}$, …
  - p = # of parameters in a sub-model (not include σ^2)
  - # of different sub-models = 2^{k-1}
- hierarchical structure of all sub-models (see graph) (nested model)
- objective of model selection: select a "best" sub-model (or some good ones)

Recall ω & Ω in testing

Q: what is a good sub-model? We usually hope a good model to have
- high R^2 ➢ $\hat\sigma\approx\sigma$ (lack of fit)
- not too many terms
- terms with significant t-tests

cf.

- Q: why bother to select a best subset of all terms? ← Why not just use the full model?
  - simplicity: principle of Occam's Razor, removal of redundant terms results in a simpler model ← overfitting: $\mathrm{Var}(\hat Y)\uparrow$ (LNp.6-7); may be easier to interpret $\hat\beta$
  - unnecessary terms will cost d.f. and add noise to the estimation of other quantities ⇒ less precise test/C.I. and tend to increase the standard error ← $\hat\sigma^2=\frac{RSS}{n-p}$ ↓
  - collinearity reduction: collinearity is caused by having too many terms trying to do same job → significant t-test becomes insignificant ($(X^TX)^{-1}_{ii}$ may become large) (cf.)
  - save cost : if model is used for prediction, can save time and/or money by not measuring redundant terms

∵ model selection is sensitive to outliers & influential points

- preliminary steps before performing variable selection
  - identify outliers and influential points --- may exclude them temporarily
  - add any terms, transformations, or (linear) combinations of the predictors or extra predictors that seem appropriate (y, x_i's)
- two types of variable selection procedures: *testing-based* and *criterion-based*
- testing-based procedure (impossible to perform if $k\ge n$ (supersaturated))
  - **Recall**: (large → insig, small → sig.)
    - the p-value of t-test is an index of effect significance/importance (∵ collinearity)
    - cannot simultaneously remove terms with insignificant p-values
  - backward elimination: (1) start with full model (all terms); (2) eliminate the term with the largest p-value greater than "α-to-remove" preset value; (3) refit the model and go to step (2); (4) stop when all p-values < α-to-remove

① no need to be 0.05 ② can be, say 0.15~0.2 for prediction (because of multiple testing)

control the complexity of the chosen model

check graph in LNp.9-1

forward selections: (1) start with no terms in the model ($y \sim 1$);
(2) For terms not in the model, check their p-values if they are added to the model. Add the term with the smallest p-value less than "α-to-enter" preset value;
(3) refit the model and go to step (2); (4) stop when all the p-values > α-to-enter

control the complexity of the chosen model → say, 0.15 ~ 0.2 ← avoid stop too soon

check graph in LNp.9-1

stepwise regression: a combination of forward and backward and there are several variations on exactly how this is done. Roughly speaking, at any step, it can (1) select a new term, according to "α-to-enter", or (2) remove a term from model, according to "α-to-remove", or (3) stop ← usually suggest α-to-enter ≤ α-to-remove to avoid infinite loop.

check graph in LNp.9-1

α-to-remove | α-to-enter

- drawbacks:
  - may miss "optimal" model because of its "one-at-a-time" adding/dropping
  - α-values (α-to-enter and α-to-remove) should not be treated too literally: because of multiple testing occurring

no need to interpret as the significance level

  - removal of less significant terms tends to increase the significance of the remaining terms ⇒ may lead to overstate the importance of the remaining terms

$\hat{\beta}$ might change & $\hat{\sigma}^2 = \frac{RSS\uparrow}{n-p\uparrow}$

e.g., In the example in LNp 9-1, if R & T have strong collinearity, after T is included, it might be very hard to explore submodels containing R. The procedure may only search part of the model space.

  - The procedure is not directly linked to final objectives of regression, such as prediction or interpretation. It's only based on statistical significance of testing in its selection.

physical significance

  - for prediction purpose, testing-based procedure tends to pick smaller models than desired

- criterion-based procedure (k: # of all parameters, including intercept; ← but not σ^2
  p: # of parameters in a sub-model; m_p: a sub-model with p parameters):
  - pick a criterion for judging the worth of a sub-model, consider all possible sub-models and pick *those* with best values of the criterion

criterion ?
$m_p \to f(m_p)$
$\in$ model space
$\arg \min_{m_p} f(m_p)$ / $\arg \max_{m_p} f(m_p)$

  - # of all possible sub-models=2^{k-1} ⇒ if k is large, computation may be too expensive, clever algorithm like "branch-and-bound" method can avoid it

or stepwise search

  - adjusted R^2 (denoted by R_a^2) criterion: for a sub-model m_p,

RSS: from m_p
TSS: from null model $y \sim 1$

    - $R^2 = 1 - (RSS/TSS)$: not good, adding terms always increase R^2
    - $R_a^2 = 1 - \{[RSS/(n-p)]/[TSS/(n-1)]\} = 1 - [(n-1)/(n-p)](1-R^2) = 1 - \left(\hat{\sigma}^2_{m_p} / \hat{\sigma}^2_{\text{null}}\right)$

same for all m_p's

$y \sim 1$

    - will only increase when a term has some value
    - larger R_a^2 is better [notice the connection between R_a^2 and $\hat{\sigma}^2_{m_p}$]

$R_a^2 \uparrow \Leftrightarrow \hat{\sigma}^2_{m_p} \downarrow$

  - PRESS (Predicted REsidual Sum of Square) criterion

PRESS = $\sum_i \hat{\varepsilon}^2_{(i)}$, where $\hat{\varepsilon}_{(i)}$ are non-standardized jacknife residuals

y_i ↔ similar ? $\hat{y}_{(i)}$

    - smaller value of PRESS is better
    - more expensive computation than R_a^2
    - tends to pick bigger models (⇒ may be desirable for prediction purpose)

For prediction, tend to prefer more effects

entropy in information theory

Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC)

substitute MLE

AIC = −2(maximized log-likelihood) + 2 p
BIC = −2(maximized log-likelihood) + $\log(n)$ p

usually $\log(n) > 2$

for linear model, −2(maximized log-likelihood) = $n \log(RSS_{m_p}/n)$ + constant

LNp.4-9

- smaller value of AIC or BIC is better
- get a balance between model fit and model size: BIC penalizes larger models more heavily than AIC ⇒ BIC tends to prefer smaller models

➢ Mallow's C_p statistics: MSE of prediction, $E(\hat{\theta}-\theta)^2 = Var(\hat{\theta}) + Bias(\hat{\theta})^2$

$(1/\sigma^2)\,\Sigma_i\, E[\,\hat{y}_i - E(y_i)\,]^2$ (from m_p)

would be a good criterion, which can be estimated by:

$\frac{1}{\sigma^2}\Sigma_i MSE(\hat{y}_i) = E\left(\frac{RSS_{m_p}}{\sigma^2} + 2p - n\right)$ → $C_p = RSS_{m_p}/\hat{\sigma}^2_{\text{full}} + 2p - n$,

unknown, use the full model ($\sigma^2_{Full} \to \sigma^2$, $X\beta_{Full} \to E(y)$) (Recall. pure-error estimate in test for lack-of-fit)

Note they are same for all m_p's

where $\hat{\sigma}^2_{\text{full}}$ estimated from the model with all terms (full model) and RSS_{m_p} is obtained from a sub-model m_p

$\Sigma_i MSE(\hat{y}_i) = E\|H_1 Y - X\beta_{Full}\|^2 = E\|H_1(X\beta_F+\varepsilon) - X\beta_F\|^2$
$= E\|-(I-H_1)X\beta_F + H_1\varepsilon\|^2 = p\sigma^2_F + \|(I-H_1)X\beta_F\|^2$

$E(RSS_{m_p}) = E\|(I-H_1)Y\|^2 = E\|(I-H_1)X\beta_F + (I-H_1)\varepsilon\|^2$
$= (n-p)\sigma^2_F + \|(I-H_1)X\beta_F\|^2$

H_1: hat matrix under m_p

- cheap to compute
- closely related to R_a^2 and AIC, BIC
- under full model: $RSS_{\{\text{full model}\}} = (n-k)\hat{\sigma}^2_{\text{full}}$, so $C_k = k$ for full model
- for sub-models that fit: $E(RSS_{m_p}) = (n-p)\sigma^2$, so $C_p \approx p$, i.e., C_p close to p implies the sub-model fits

if $\hat{\sigma}^2_{\text{full}} \approx \hat{\sigma}^2_{m_p} \approx \sigma^2$, then $C_p \approx \frac{(n-p)\hat{\sigma}^2_{\text{full}}}{\hat{\sigma}^2_{\text{full}}} + 2p - n = p$

$C_p = (n-k) + 2k - n$

- for sub-models that do not fit: $E(RSS_{m_p}) >> (n-p)\sigma^2$ and $C_p >> p$
- it's usual to plot C_p against p. Models with small p and C_p around the C_p=p line or less than p are desirable

cf. different from other criteria in punishment term

➢ Note: C_p, R_a^2, AIC, BIC all trade-off fit in terms of RSS_{m_p} against complexity (p) ⇒ we prefer models with smaller RSS_{m_p} and smaller p; however, $RSS_{m_p}\downarrow$ as $p\uparrow$

- ambiguity about the best model is possible. When several candidate models exist:

➢ check if models make similar predictions? if yes, can make decision on the basis of cost; if no, do not pick one model arbitrary. Report a range of models.

➢ interpretations qualitatively similar? if not, avoid strong conclusion and report a range of models

➢ examine which has the best diagnostics (LNp.1-12)

can sumarize consistent patterns in these models to get conclusion with stronger level of evidence

These criteria make sense from the statistical perspective. However, they may not make sense from the physical perspective. (Note. These criteria do not use domain knowledge. Under such knowledge, some m_p's might be preferred than the others.

- Notes:

➢ terms not in final model can still be correlated with the response ⇒ not to say they are unrelated to the response;

⇒ better to say they provide no additional explanatory effect beyond those terms included in final model

e.g. ∵ collinearity

an optimization

cor(X_1, X_2) ≈ 1
true model: $y \sim X_1$
model selection might choose $y \sim X_2$

➢ It's important to keep in mind that model selection should not be divorced from the underlying purpose of investigation

➢ automatic variable selection are not guaranteed to be consistent with your goals. Use these methods as a guide only.

can put weights on m_p's

➢ these methods do not consider the natural hierarchy in some models: For example, in polynomial model, higher-order terms (such as x_1^2, x_2^2, x_1x_2) should be considered only when corresponding lower-order terms (such as x_1 and x_2) have been included in the model ⇒ not all sub-models are candidate models

❖ **Reading**: Faraway (1st ed.), chapter 8; W, 10.2, 10.3, 10.4 ❖ **Further reading**: D&S, chapter 15