Model Building Process,
Model Selection Guideline and Criteria

The Model Building Process

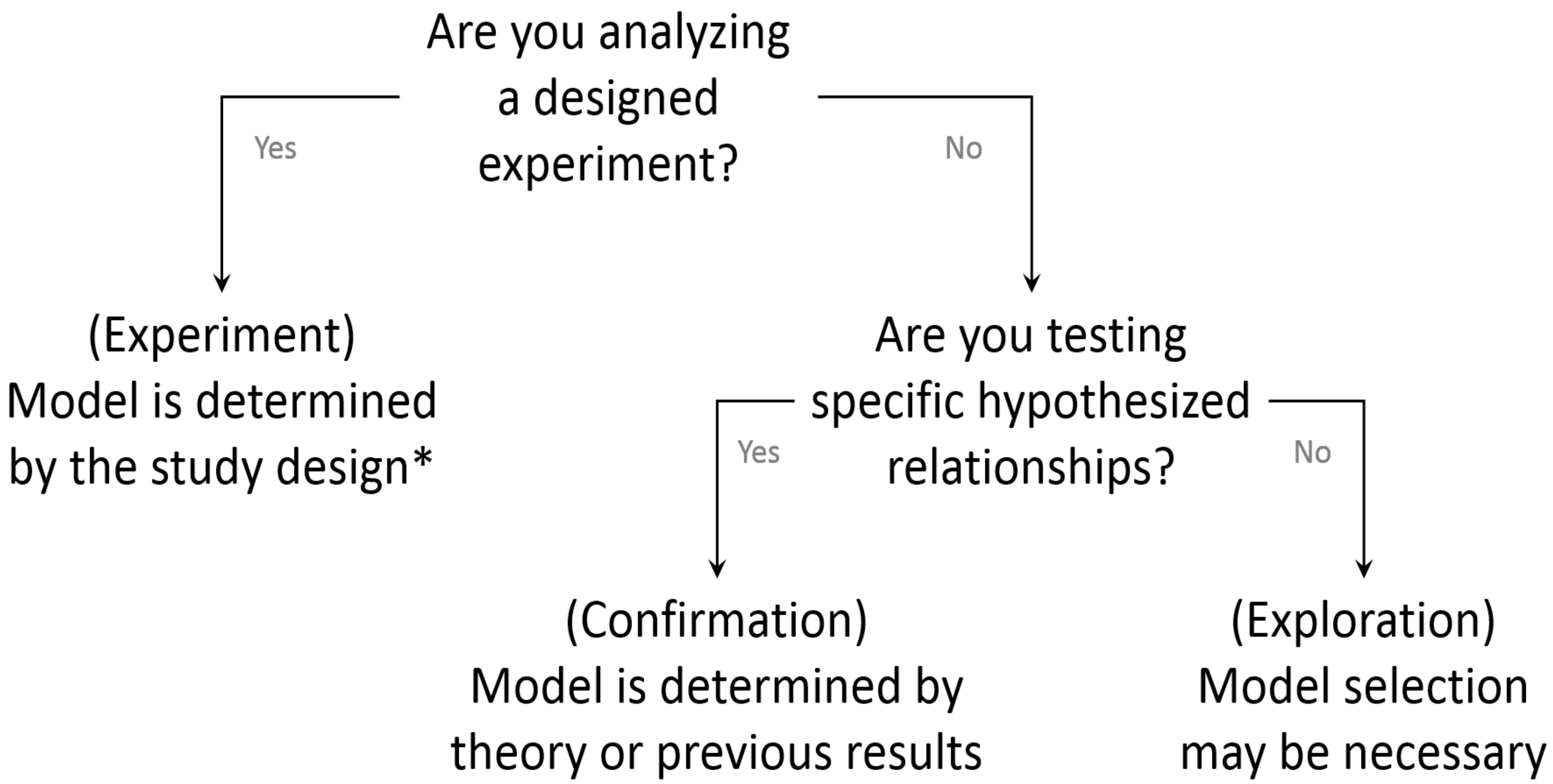
1. Planning and Data Collection:
    - Identify research questions and objectives
    - Plan data collection (decide sample unit, variable, sample size)
    - Collect data
    - Clean data (check for errors and organize database)

2. Model Exploration:
    - Use graphical screening and bivariate modeling to explore data
    - Identify relationships and potential outliers
    - Recognize possible interactions, especially for qualitative variables
    - Discuss possible sources of multicollinearity or other issues
    - Use various methods including histograms, scatterplots, contingency tables, boxplots, pairwise correlations, SLR results, residual plots, and diagnostics for individual predictors

3. Model Selection:
    - Fit various regression models
    - Compare results to identify best models that align with study objectives
    - Reduce explanatory variables depending on the nature of the study

4. Model Validation:
    - Compare model predictions against theoretical expectations
    - Check model's predictive ability with cross-validation

Model Selection Depends on the Nature of Study



Are you analyzing a designed experiment?

Yes

(Experiment)
Model is determined by the study design*

No

Are you testing specific hypothesized relationships?

Yes

(Confirmation)
Model is determined by theory or previous results

No

(Exploration)
Model selection may be necessary

* Model selection on *covariates* may be helpful.

The Nature of Study

I. Controlled experiment

- This study type involves controlling the levels of explanatory variables and assigning a treatment to each experimental unit to observe its response. In controlled experiments, the explanatory variables are often called factors or control variables. In controlled experiments, the explanatory variables are often called **factors** or **control variables**.
- For instance, an experiment that examines the impact of graphic presentation size (X1) and analysis time (X2) on accuracy (Y). A treatment consists of a specific combination of size and time.
- $Y{\sim}X1 + X2$

II. Controlled Experiments with covariates

- In this study, **uncontrolled variables or covariates** are included to reduce error variance.
- For example, in the previous experiment, gender (X3) and years of experience (X4) are measured as uncontrolled variables from each unit.
- $Y{\sim}X1 + X2 + \mathrm{X3} + \mathrm{X4}$

The Nature of Study

III. Confirmatory observational study

- This type of study is intended to test hypotheses based on observational data, not experimentation.
- The explanatory variables are called **primary variables**, and the variables included to reflect existing knowledge are called **control variables** (or known risk factors in epidemiology).
- In this study, the control variables are not controlled, but they reflect the known influence.
- For instance, in an observational study of the effect of vitamin E supplements (X1) on a certain type of cancer (Y), known risk factors such as age (X2), gender (X3), and race (X4) would be included as control variables, while the amount of vitamin E supplements taken daily would be the primary explanatory variable.
- $Y \sim X1 + X2 + \mathrm{X3} + \mathrm{X4}$

The Nature of Study

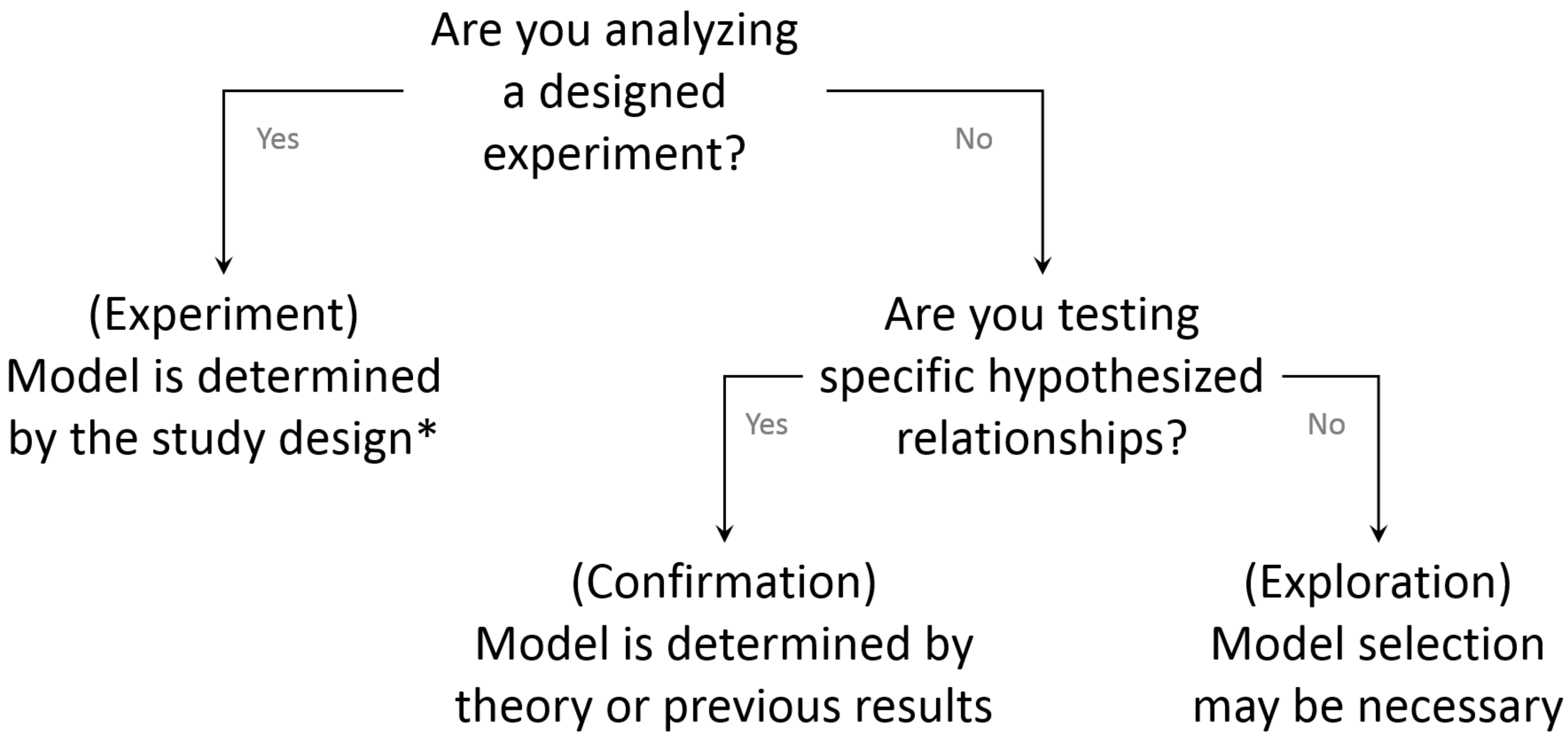
IV. Exploratory observational study

- This study is often used in social, behavioral, health science, management, and other fields when conducting controlled experiments is not possible, or when adequate knowledge for confirmatory observational studies is lacking.
- In this type of study, explanatory variables that are not directly measurable could be involved in any available theoretical model. Variables that could be conceivably related to the response variable are studied.
- The number of cases collected for an exploratory observational regression study depends on the size of the pool of variables. A general rule of thumb suggests that there should **be at least 6 to 10 cases for every variable in the pool.**
- $\boldsymbol{Y \sim X1 + X2 + X3 + X1^2 + X1X2 + X1X3}$

Model Selection Guideline on Reduction of Predictors

- In a controlled experiment, the reduction of explanatory variables is usually not an essential issue.
- In controlled experiments with covariates, some reduction of the covariates may take place.
- In a confirmatory observational study, no reduction of primary explanatory variables should generally take place. Even the controlled variables should be retained for comparison with earlier studies.
- In an exploratory observational study, many variables are frequently highly inter-correlated. The main goal is to determine the functional form, interactions, and reduce the variables accordingly.

When is Model Selection Needed?



Are you analyzing a designed experiment?

(Experiment)
Model is determined by the study design*

Are you testing specific hypothesized relationships?

(Confirmation)
Model is determined by theory or previous results

(Exploration)
Model selection may be necessary

* Model selection on *covariates* may be helpful.

Methods of Model Selection

1. Selection by design (experiments)

- One or a few specific models based on the design of the experiment

2. Interest/previous knowledge/expert opinion (confirmation, covariates)

   Selection informed by study objectives or previous experience

3. Best subsets algorithms

- identify the “best” model with a subset of $p-1$ predictors, according to some criterion

4. Stepwise algorithms

- construct the model by adding or removing variables one at a time and monitoring changes in a criterion

Some comments on model selection

Many criteria have been proposed to help identify the “best” subset of predictor variables

- Each has benefits and drawbacks
- In some cases, they may lead to different conclusions

In general, you should think of model selection criteria as tools that provide insights about your regression problem, not as magical oracles. Model building is about choices, determined by *you*.

Case Study: Surgical Unit Example

A hospital surgical unit was interested in predicting survival in patient undergoing a particular type of liver operation. A random number of 108 patients was available for analysis, but we only study (n=)54. For each patient record, the following information was extracted (data: surgery.csv):

Potential predictors include,

- Blood clotting score (X_1, blood)
- A prognostic index (X_2, prog)
- Enzyme function test (X_3, enz)
- Liver function test (X_4, liver)

The response variable is survival time in days (Y, surv)

We skip the model exploration process in this topic.

Check out the MLR diagnostic procedure case for the process of transforming the regression function

Current model $\ln(Y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_4 liver + \epsilon$

Should we delete some predictors?

We now proceed with the model selection process.

Criteria for Model Selection

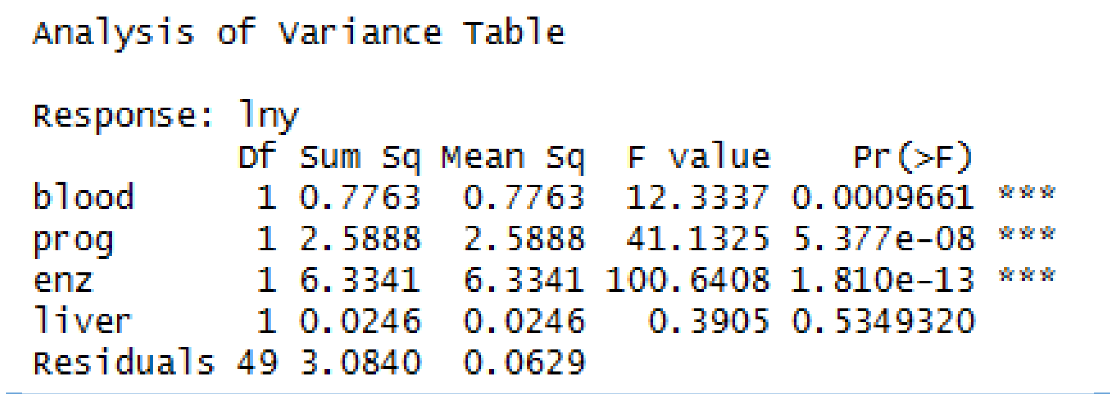
- R^2
- Adjusted R^2 (MSE)
- Mallows' C_p
- AIC
- SBC
- $PRESS$

Model selection: R_p^2 or SSE_p criterion

We will assume that the number of observations (n) exceeds the maximum number of potential parameters (P): $n > P$

$\boldsymbol{R_p^2}$: The multiple determination for p parameters or $p-1$ predictors

$$\boldsymbol{R_p^2} = 1 - \frac{\boldsymbol{SSE_p}}{SSTO}$$



```
Analysis of Variance Table

Response: lny
          Df Sum Sq Mean Sq  F value    Pr(>F)
blood      1 0.7763  0.7763  12.3337 0.0009661 ***
prog       1 2.5888  2.5888  41.1325 5.377e-08 ***
enz        1 6.3341  6.3341 100.6408 1.810e-13 ***
liver      1 0.0246  0.0246   0.3905 0.5349320
Residuals 49 3.0840  0.0629
```

SSTO= 12.8078

Model selection: $R_{a,p}^2$ or MSE_p criterion

The R_p^2 criterion is not intended to identify the subsets since it never decreases.

$$\boldsymbol{R_{a,p}^2} = 1 - \left(\frac{n-1}{n-p}\right)\frac{\boldsymbol{SSE_p}}{SSTO} = 1 - \frac{MSE_p}{SSTO/(n-1)}$$

Example:

$$\ln(Y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_4 liver + \epsilon$$

$$\boldsymbol{R_{a,p}^2} = 1 - \left(\frac{n-1}{n-p}\right)\frac{\boldsymbol{SSE_p}}{SSTO}$$

$$= 1 - \left(\frac{54-1}{54-5}\right)\frac{\boldsymbol{3.084}}{12.8078} = 0.7396$$

$$\ln(Y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + +\epsilon$$

$$\boldsymbol{R_{a,p}^2} = 1 - \left(\frac{n-1}{n-p}\right)\frac{\boldsymbol{SSE_p}}{SSTO}$$

$$= 1 - \left(\frac{54-1}{54-4}\right)\frac{\boldsymbol{3.084 + 0.0246}}{12.8078} = 0.743$$

Model selection: Mallows' C_p criterion

The squared error of the ith fitted value: $(\hat{Y}_i - \mu_i)^2$

The mean(expected) squared error of the ith fitted value : $E(\hat{Y}_i - \mu_i)^2 = (E\{\hat{Y}_i\} - \mu_i)^2 + \sigma^2\{\hat{Y}_i\}$

The total mean squared error: $\Sigma[E(\hat{Y}_i - \mu_i)^2] = \Sigma(E\{\hat{Y}_i\} - \mu_i)^2 + \Sigma\sigma^2\{\hat{Y}_i\}$

The total mean squared error divided by the error variance (σ^2):

$$\Gamma_p = \frac{1}{\sigma^2}[\Sigma(E\{\hat{Y}_i\} - \mu_i)^2 + \Sigma\sigma^2\{\hat{Y}_i\}]$$

Which can then be estimated by C_p

$$\boldsymbol{C_p = \frac{SSE_p}{MSE_{full}} - (n - 2P)}$$

Comments:

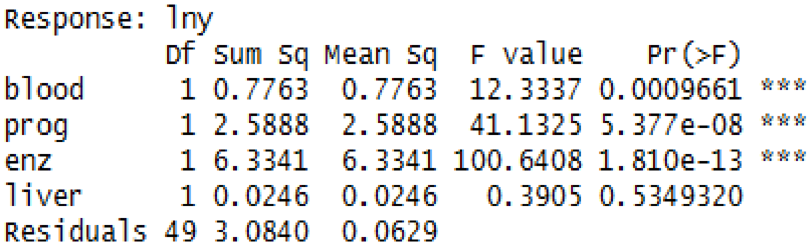
- **When there is no bias in the model with p- 1 predictors and $E\{\hat{Y}_i\} = \mu$** $\quad C_p \approx \mathrm{P}$
- **Model is better when C_p is : 1) small and 2) near p**
  - ➢ **It may sometimes occur that the regression model based on a subset of X variables with a small C_p but large bias.**
  - ➢ **One may prefer a model based on a somewhat more X with a slightly larger C_p but smaller bias.**

Model selection: Mallows' C_p criterion (should be small and near p)

$$C_p = \frac{SSE_p}{MSE_{full}} - (n - 2P)$$

For example,

$$\ln(Y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_4 liver + \epsilon$$



```
Response: lny
          Df Sum Sq Mean Sq  F value    Pr(>F)    
blood      1 0.7763  0.7763  12.3337 0.0009661 ***
prog       1 2.5888  2.5888  41.1325 5.377e-08 ***
enz        1 6.3341  6.3341 100.6408 1.810e-13 ***
liver      1 0.0246  0.0246   0.3905 0.5349320    
Residuals 49 3.0840  0.0629                       
```

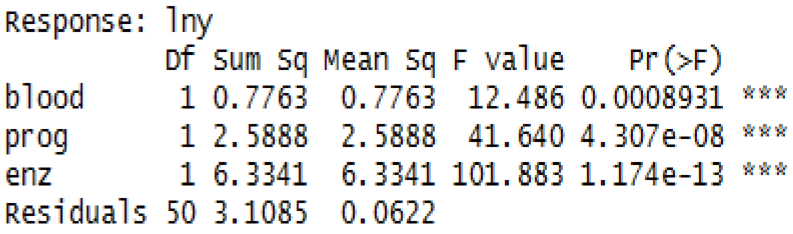
$$C_p = \frac{SSE_p}{MSE_{full}} - (n - 2p)$$

$$= \frac{SSE_{full}}{MSE_{full}} - (n - 2 * 5)$$

$$= n - p - (n - 2p) = p = 5$$

C_p of the full model is exactly p.

$$\ln(Y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \epsilon$$



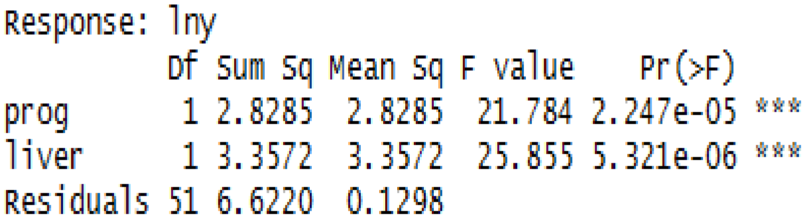
```
Response: lny
          Df Sum Sq Mean Sq F value    Pr(>F)    
blood      1 0.7763  0.7763  12.486 0.0008931 ***
prog       1 2.5888  2.5888  41.640 4.307e-08 ***
enz        1 6.3341  6.3341 101.883 1.174e-13 ***
Residuals 50 3.1085  0.0622                      
```

$$C_p = \frac{SSE_p}{MSE_{full}} - (n - 2p)$$

$$= \frac{3.1085}{0.0629} - (54 - 2 * 4) = 3.4$$

C_p is close to p=4: indicating little or no bias in this model.

$$\ln(Y) = \beta_0 + \beta_2 prog + \beta_4 liver + \epsilon$$



```
Response: lny
          Df Sum Sq Mean Sq F value    Pr(>F)    
prog       1 2.8285  2.8285  21.784 2.247e-05 ***
liver      1 3.3572  3.3572  25.855 5.321e-06 ***
Residuals 51 6.6220  0.1298                      
```

$$C_p = \frac{SSE_p}{MSE_{full}} - (n - 2p)$$

$$= \frac{6.622}{.0629} - (54 - 2 * 3) = 57.28$$

C_p is larger than in the second model. Plus, it is biased because Cp is much larger than p(=3) in this case.

Model selection: $\mathrm{AIC_p}$and $\mathrm{SBC_p}$ criteria

Akaike's information criterion

$$AIC_p = n * ln\left(SSE_p\right) - n * ln\,(n) + 2p$$

Schwarz's Bayesian criterion

Aka Bayesian information criterion (BIC)

$$SBC_p = n * ln\left(SSE_p\right) - n * ln\,(n) + [ln(n)] * p$$

Comments:

- Both methods based on the Maximum Likelihood method.
  - The model does a good job explaining the ***current*** data. But there is chance of overfitting for the future data.
  - Can be used to compare candidate models with different error distributions which ***may not be Normal.***
  - ***Do not*** assume any form of nesting, i.e., the p predictors are a subset of the full model. But all models need to be trained on the same data.
- The better the model, the smaller AIC_p or SBC_p is.
- AIC_p and SBC_p differ in the way they penalize for model complexity.
  - The AIC_p penalizes for the number of parameters in the model, while the SBC_p penalizes for both the number of parameters and the number of observations in the model.
  - In general, AIC_p is more suitable for small datasets, while SBC_p is more suitable for large datasets.
  - If $n \geq 8$, the penalty for SBC_p is larger than that for AIC_p; hence the SBC_p tends to favor simpler models
- AIC_p and C_p will tend to pick the same model.
- If the true model is a candidate,
  - AIC_p and C_p will tend to pick more complex models than the truth
  - SBC_p will tend to pick the true model more often

Model selection: AIC$_p$and SBC$_p$ criteria

Akaike's information criterion $\boldsymbol{AIC_p = n * ln\left(SSE_p\right) - n * ln\left(n\right) + 2p}$

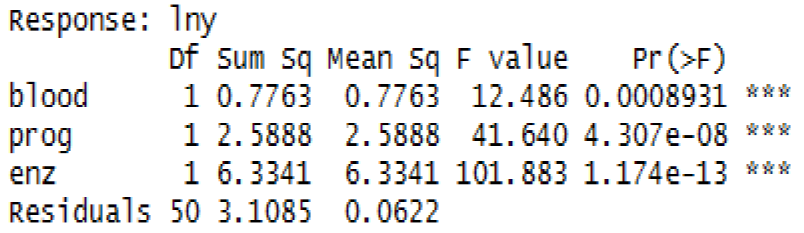
Schwarz's Bayesian criterion $\boldsymbol{SBC_p = n * ln\left(SSE_p\right) - n * ln\left(n\right) + [ln(n)] * p}$

Example

$$\ln(Y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \epsilon$$

$AIC_4 =$ $54 * \ln(3.1085) - 54 * \ln(54) + 2(4)$ = -146.162

SB$C_4 =$ $54 * \ln(3.1085) - 54 * \ln(54) + ln(54)$*4 = -138.206

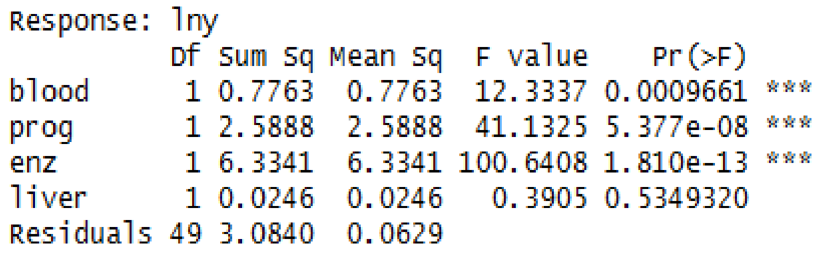


```
Response: lny
          Df Sum Sq Mean Sq F value    Pr(>F)    
blood      1 0.7763  0.7763  12.486 0.0008931 ***
prog       1 2.5888  2.5888  41.640 4.307e-08 ***
enz        1 6.3341  6.3341 101.883 1.174e-13 ***
Residuals 50 3.1085  0.0622                      
```

$$\ln(Y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_4 liver + \epsilon$$

$AIC_5 =$ $54 * \ln(3.084) - 54 * \ln(54) + 2(5)$ = -144.59

SB$C_5 =$ $54 * \ln(3.084) - 54 * \ln(54) + ln(54)$*5 = -134.645



```
Response: lny
          Df Sum Sq Mean Sq  F value    Pr(>F)    
blood      1 0.7763  0.7763  12.3337 0.0009661 ***
prog       1 2.5888  2.5888  41.1325 5.377e-08 ***
enz        1 6.3341  6.3341 100.6408 1.810e-13 ***
liver      1 0.0246  0.0246   0.3905 0.5349320    
Residuals 49 3.0840  0.0629                       
```

Model selection: $PRESS_p$ criterion

- The Prediction Sum of Squares (PRESS) criterion measures the effectiveness of using the fitted values from a subset model to predict the observed response.
- It differs from the Sum of Squares Error (SSE) in that each fitted value is obtained by **excluding the ith observation from the dataset**, and the model is estimated using the remaining **n-1 observations**, this predicted value is denoted by $\hat{\boldsymbol{Y}}_{\boldsymbol{i(i)}}$.
- PRESS is also referred to as "leave-one-out-cross-validation."

$$PRESS_p = \Sigma\left(Y_i - \hat{Y}_{i(i)}\right)^2 \qquad SSE_p = \Sigma\left(Y_i - \hat{Y}_i\right)^2$$

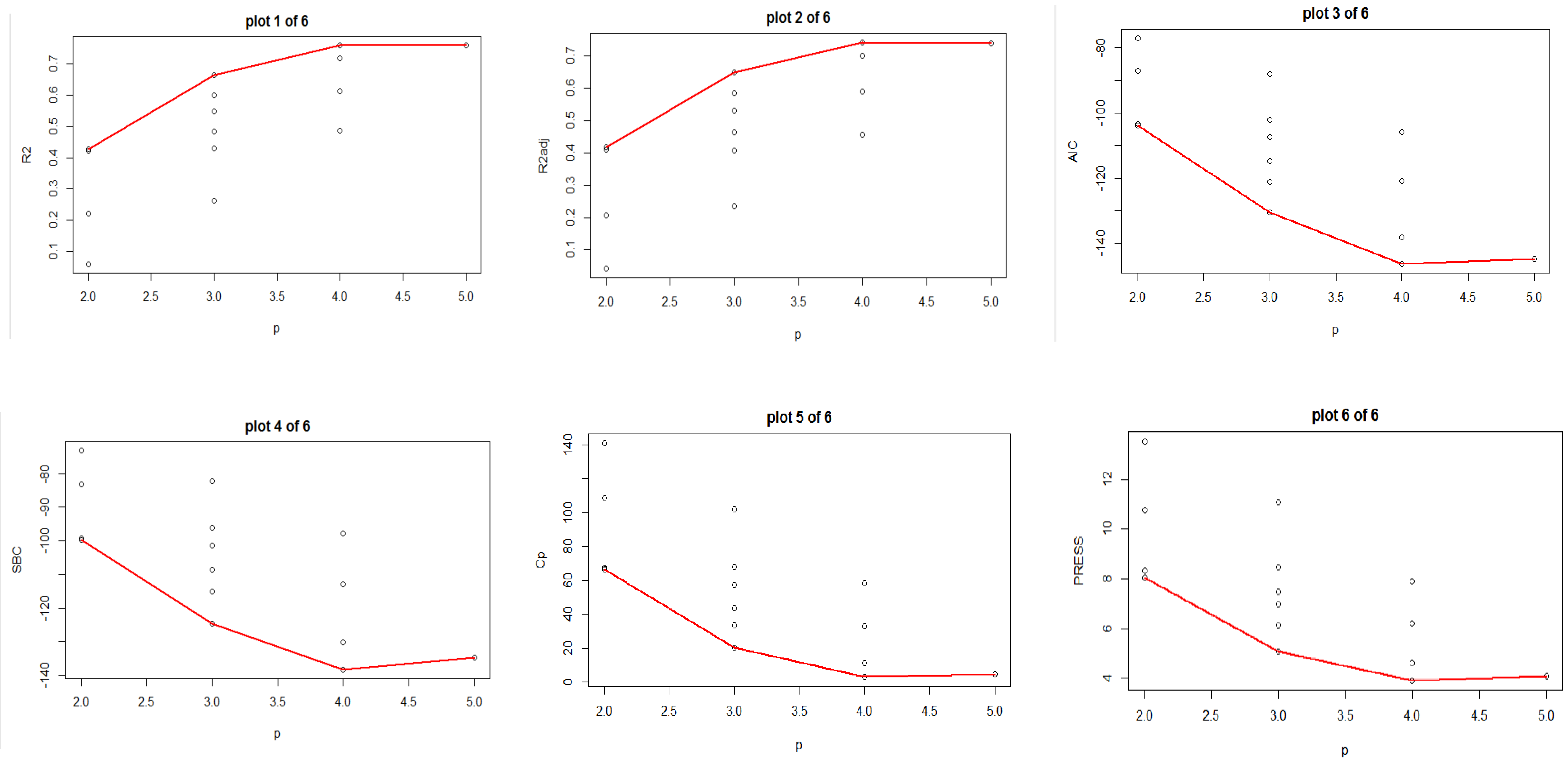
- Models with small $PRESS_p$ values are considered good.
- It is not necessary to refit the model n times, $PRESS$ can be calculated using the information in the Hat matrix

$$PRESS_p = \Sigma\left(\frac{Y_i - \hat{Y}_i}{1 - H_{ii}}\right)^2$$, where H_{ii} is the ith diagonal element of the Hat matrix.

- When the purpose of multiple linear regression (MLR) is to make predictions, it is recommended to use the PRESS criterion for model selection, since it is a **measure of the predictive accuracy of the model**, which is what **matters most.**

Plot of variables selection criteria-Surgical Unit Example

```
library(ALSM)
plotmodel.s(sur[,1:4], sur$lny)
```



Plot of variables selection

- Plots of variable selection criteria show the criteria for each possible subset of variables.
- There are six criteria used in the plots.
- The subset with the optimal criterion can be chosen based on the plots.
- Note that the plots do not tell you exactly which variables are selected,

only the number of variables in the best subset.

- For example, subsets x1, x2, x3 and x1, x2, x4 both have the same number of variables,

but they are different subsets. More on this will be introduced in the next topic.