Stat 206: Linear Models

Lecture 13

Nov. 16, 2015

Key Components for Model Selection

- A criterion to compare models.
  - $R_p^2, R_{a,p}^2, C_p, AIC_p, BIC_p, Press_p$, etc.
- A procedure to search through the model space to find a good model according to a pre-specified criterion.
  - Stepwise regression: Greedy search algorithm; Applicable even if the number of potential X variables is large.

Notations and assumptions.

- Denote the total number of potential X variables by $P - 1$ and assume that all important X variables are included in this pool.
- *Full model*: The model that contains all potential X variables.
  - The full model is a correct model.
  - It is often used to provide an unbiased estimator for the error variance. In that case, it is also needed that the sample size n is sufficiently large, e.g., n/P at least 6.
- *Candidate model*: A model that contains a subset of $p - 1$ X variables with $1 \leq p \leq P$.

R_p^2 Criterion

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

- The R_p^2 criterion favors models with high coefficient of multiple determination.
- The subscript p indicates the number of regression coefficients in the model.
- The R_p^2 criterion is equivalent to using the error sum of squares SSE_p as a criterion where models with small SSE_p are considered “good”. (Note $SSTO$ is the same for all models.)

The R_p^2 criterion is to choose **the** model that maximizes R^2.

- R^2 is always maximized by .
- **The R_p^2 criterion should be used to find the point starting from where adding more X variables only leads to a marginal increase in R^2.**

The R_p^2 criterion is **not intended** to choose **the** model that maximizes R^2.

- R^2 is always maximized by the full model.
- **The R_p^2 criterion should be used to find the point starting from where adding more X variables only leads to a marginal increase in R^2.**

$R^2_{a,p}$ Criterion

Unlike R^2_p, the adjusted coefficient of multiple determination $R^2_{a,p}$ takes into account of   in the model:

$$R^2_{a,p} = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SSTO} = 1 - \frac{MSE_p}{MSTO}.$$

- R^2_a could   when additional X variables are added into the model since the increase in R^2 may be too small to offset the loss of   .
- The $R^2_{a,p}$ criterion is equivalent to using the mean squared error MSE_p as a criterion where models with small MSE_p are considered "good".

$R^2_{a,p}$ Criterion

Unlike R^2_p, the adjusted coefficient of multiple determination $R^2_{a,p}$ takes into account of the number of parameters in the model:

$$R^2_{a,p} = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SSTO} = 1 - \frac{MSE_p}{MSTO}.$$

- R^2_a could decrease when additional X variables are added into the model since the increase in R^2 may be too small to offset the loss of degrees of freedom.
- The $R^2_{a,p}$ criterion is equivalent to using the mean squared error MSE_p as a criterion where models with small MSE_p are considered “good”.

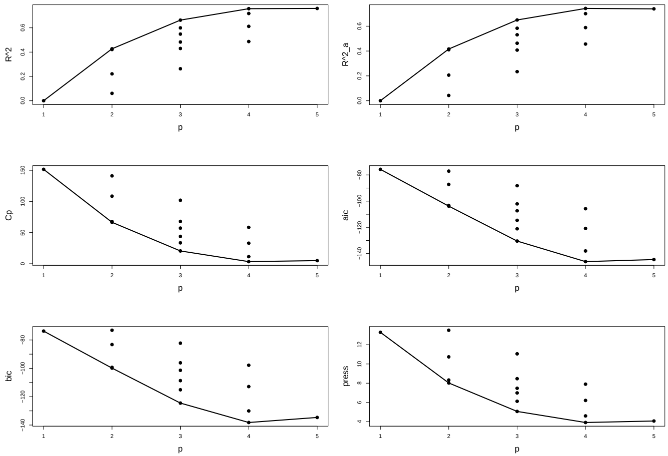
Surgical Unit: Model Selection Criteria

For illustration purposes, only consider X_1, X_2, X_3, X_4 (`clotting`, `prognostic`, `enzyme`, `liver`) as the potential X variables. There are 16 sub-models.

p	intercept	X1	X2	X3	X4	sse	R^2	R^2_a	Cp	aic	bic	press
1	1	0	0	0	0	12.805	0.000	0.000	151.569	-75.716	-73.727	13.292
2	1	0	0	1	0	7.334	0.427	0.416	66.518	-103.811	-99.833	8.329
2	1	0	0	0	1	7.408	0.421	0.410	67.696	-103.268	-99.290	8.024
2	1	0	1	0	0	9.974	0.221	0.206	108.469	-87.205	-83.227	10.738
2	1	1	0	0	0	12.028	0.061	0.043	141.093	-77.096	-73.118	13.508
3	1	0	1	1	0	4.313	0.663	0.650	20.523	-130.479	-124.512	5.066
3	1	0	0	1	1	5.132	0.599	0.583	33.536	-121.089	-115.122	6.123
3	1	1	0	1	0	5.783	0.548	0.531	43.873	-114.644	-108.677	6.989
3	1	0	1	0	1	6.620	0.483	0.463	57.175	-107.342	-101.375	7.474
3	1	1	0	0	1	7.299	0.430	0.408	67.961	-102.070	-96.103	8.472
3	1	1	1	0	0	9.437	0.263	0.234	101.937	-88.194	-82.227	11.055
4	1	1	1	1	0	3.109	0.757	0.743*	3.388*	-146.161*	-138.205*	3.914*
4	1	0	1	1	1	3.615	0.718	0.701	11.434	-138.011	-130.055	4.598
4	1	1	0	1	1	4.970	0.612	0.589	32.960	-120.823	-112.867	6.209
4	1	1	1	0	1	6.568	0.487	0.456	58.358	-105.763	-97.807	7.902
5	1	1	1	1	1	3.084	0.759*	0.739	5.000	-144.587	-134.642	4.069

Within each subset size, models are sorted in ascending *SSE* so that the first one has the smallest *SSE*. Consequently, within each subset size, $R_p^2, R_{a,p}^2$ are from the largest to the smallest and C_p, BIC_p, AIC_p are from the smallest to the largest. $Press_p$ may not be monotone with *SSE*.

Figure: Surgical Unit: Model Selection Criteria Plots



- $\max(R_p^2)$ (eq. $\max(R_{a,p}^2)$) denotes the largest R^2 (eq. R_a^2) among models with p regression coefficients.
- The $\max(R_p^2)$ versus p plot is necessarily .
- From the plot, there is little increase in both $\max(R_p^2)$ and $\max(R_{a,p}^2)$ going from $p =$ to $p =$.
    - Best model with 3 X variables ($p = 4$) contains X_1, X_2, X_3 with $R^2 = 0.757, R_a^2 = 0.743$.
    - Best model with 4 X variables ($p = 5$) contains X_1, X_2, X_3, X_4 with $R^2 = 0.759, R_a^2 = 0.739$.
    - $\max(R_{a,p}^2)$ decreases when p increases from 4 to 5.
- By both R_p^2 and $R_{a,p}^2$ criteria, the model with is favored among the 16 models being considered.
- Although X_3 and X_4 correlate most highly with the response variable ($r = 0.65$), they do not both appear in the "best" model. This means that X_1, X_2, X_3 contain much of the information in X_4 in terms of explaining Y.

- $\max(R_p^2)$ (eq. $\max(R_{a,p}^2)$) denotes the largest R^2 (eq. R_a^2) among models with p regression coefficients.
- The $\max(R_p^2)$ versus p plot is necessarily nondecreasing.
- From the plot, there is little increase in both $\max(R_p^2)$ and $\max(R_{a,p}^2)$ going from $p = 4$ to $p = 5$.
  - Best model with 3 X variables ($p = 4$) contains X_1, X_2, X_3 with $R^2 = 0.757, R_a^2 = 0.743$.
  - Best model with 4 X variables ($p = 5$) contains X_1, X_2, X_3, X_4 with $R^2 = 0.759, R_a^2 = 0.739$.
  - $\max(R_{a,p}^2)$ decreases when p increases from 4 to 5.
- By both R_p^2 and $R_{a,p}^2$ criteria, the model with X_1, X_2, X_3 is favored among the 16 models being considered.
- Although X_3 and X_4 correlate most highly with the response variable ($r = 0.65$), they do not both appear in the "best" model. This means that X_1, X_2, X_3 contain much of the information in X_4 in terms of explaining Y.

Mallows' C_p Criterion

Mallows' C_p for a model with p regression coefficients:

$$C_p := \frac{SSE_p}{\hat{\sigma}^2} - (n - 2p).$$

- n : sample size (constant across models).
- SSE_p: error sum of squares of the candidate model.
- $\hat{\sigma}^2$: an unbiased estimator of the error variance σ^2. E.g.,

$$\hat{\sigma}^2 = MSE_{\text{full model}} = MSE(X_1, \cdots, X_{P-1}).$$

  - $\hat{\sigma}^2$ is unbiased due to the assumption that the full model contains all important X variables so that

    .
  - C_p of the full model is always .

Mallows' C_p Criterion

Mallows' C_p for a model with p regression coefficients:

$$C_p := \frac{SSE_p}{\hat{\sigma}^2} - (n - 2p).$$

- n : sample size (constant across models).
- SSE_p: error sum of squares of the candidate model.
- $\hat{\sigma}^2$: an unbiased estimator of the error variance σ^2. E.g.,

$$\hat{\sigma}^2 = MSE_{\text{full model}} = MSE(X_1, \cdots, X_{P-1}).$$

  - $\hat{\sigma}^2$ is unbiased due to the assumption that the full model contains all important X variables so that it has no model bias.
  - C_p of the full model is always P.

Mallows' C_p as an Estimator of msee

Let $\mathbb{M} = \mathbb{M}(X_1, \cdots, X_{p-1})$ be a model. Then:

$$
\begin{aligned}
E(C_p(\mathbb{M})) &\approx \frac{E(SSE(\mathbb{M}))}{\sigma^2} - (n - 2p) \\
&= \frac{(n-p)\sigma^2 + \|bias_{in}(\mathbb{M})\|_2^2}{\sigma^2} - (n - 2p) \\
&= \frac{p\sigma^2 + \|bias_{in}(\mathbb{M})\|_2^2}{\sigma^2} \\
&= \frac{Var_{in}(\mathbb{M}) + \|bias_{in}(\mathbb{M})\|_2^2}{\sigma^2} = \frac{msee_{in}(\mathbb{M})}{\sigma^2}.
\end{aligned}
$$

So C_p can be viewed as an estimator of the overall (in-sample) mean-squared-estimation-error divided by the error variance.

How to Use C_p?

- If a model has no (in-sample) bias, i.e., $bias_{in}(\mathbb{M}) = \mathbf{0}$, then $E(C_p)$. Otherwise $E(C_p)$ tends to be than p.
- When C_p is plotted against p, then models with will tend to fall near the diagonal line $C_p = p$.
- On the other hand, models with will tend to fall considerably above this line.
- **We should look for models with (i) the C_p value not far above p and (ii) small C_p value.** Such models have bias and number of X variables (thus model variance).
    - Surgical unit. The model with X_1, X_2, X_3 has $C_p = 3.38 < p = 4$, indicating little or no bias. Its C_p value is also the smallest among all models being considered.

How to Use C_p?

- If a model has no (in-sample) bias, i.e., $bias_{in}(\mathbb{M}) = \mathbf{0}$, then $E(C_p) \approx p$. Otherwise $E(C_p)$ tends to be larger than p.
- When C_p is plotted against p, then models with little bias will tend to fall near the diagonal line $C_p = p$.
- On the other hand, models with substantial bias will tend to fall considerably above this line.
- **We should look for models with (i) the C_p value not far above p and (ii) small C_p value.** Such models have small bias and small number of X variables (thus less model variance).
    - Surgical unit. The model with X_1, X_2, X_3 has $C_p = 3.38 < p = 4$, indicating little or no bias. Its C_p value is also the smallest among all models being considered.

AIC_p and BIC_p Criteria

- *Akaike's information criterion (AIC)*:

$$AIC_p = n \log \frac{SSE_p}{n} + 2p.$$

- *Bayesian information criterion (BIC)*:

$$BIC_p = n \log \frac{SSE_p}{n} + (\log n)p.$$

- **We should look for models with small AIC (BIC).**
  - Surgical unit. The model with X_1, X_2, X_3 has the smallest AIC and BIC among the models being considered.

Both AIC and BIC consist of two terms.

- The first term: $n \log \frac{SSE_p}{n}$ reflects the of the model to the observed data.
  - It by adding more X variables into the model.
- The second term, $2p$ for AIC and $(\log n)p$ for BIC, reflects .
  - It by adding more X variables into the model.
  - If $n \geq 8$, then $\log n > 2$ and BIC puts more penalty on model complexity and tends to choose models than AIC.
- Overly simplified models have model complexity (p), but they tend to have a SSE (underfitting). On the other hand, overly complicated models may have a SSE (overfitting), but they have model complexity (p).
- **By minimizing AIC (or BIC), we are trying to find a model that balances between model complexity and the goodness-of-fit of the model to the observed data**.

Both AIC and BIC consist of two terms.

- The first term: $n \log \frac{SSE_p}{n}$ reflects the *goodness-of-fit* of the model to the observed data.
  - It decreases by adding more X variables into the model.
- The second term, $2p$ for AIC and $(\log n)p$ for BIC, reflects model complexity.
  - It increases by adding more X variables into the model.
  - If $n \geq 8$, then $\log n > 2$ and BIC puts more penalty on model complexity and tends to choose smaller models than AIC.
- Overly simplified models have small model complexity (p), but they tend to have a large SSE (underfitting). On the other hand, overly complicated models may have a small SSE (overfitting), but they have large model complexity (p).
- **By minimizing AIC (or BIC), we are trying to find a model that balances between model complexity and the goodness-of-fit of the model to the observed data**.

$Press_p$ Criterion

$Press_p$ (predicted residual sum of squares) is a measure on how well a model performs in terms of prediction.

- $Press_p$ is the sum of the squared prediction errors:

$$Press_p = \sum_{i=1}^{n} (Y_i - \widehat{Y}_{i(i)})^2.$$

  - Y_i is the observed response of the *ith* case.
  - $\widehat{Y}_{i(i)}$ is a predicted value for the ith case obtained by fitting the model only using $n-1$ cases excluding case i.

- $Press_p$ is also known as *leave-one-out-cross-validation (LOOCV)*.

- **Models with small $Press_p$ are considered good in terms of predictive ability.**
  - Surgical unit: the model with X_1, X_2, X_3 has $Press_p = 3.914$ which is the smallest among all models being considered here.

Calculate $Press_p$

$Press_p$ can be calculated without actually performing n regressions.

- The *deleted residual* for the *ith* case :

$$d_i := Y_i - \widehat{Y}_{i(i)} = \frac{e_i}{1 - h_{ii}}, \quad i = 1, \cdots, n.$$

  where $e_i = Y_i - \widehat{Y}_i$ is the residual of the *ith* case and h_{ii} is the *ith* diagonal element of the hat matrix $\mathbf{H}$, both from the regression fit using all n cases.

- So

$$Press_p = \sum_{i=1}^{n} \frac{(Y_i - \widehat{Y}_i)^2}{(1 - h_{ii})^2}.$$

Derive the Deleted Residuals

- Define $\tilde{\mathbf{Y}}$ by replacing the ith element of the response vector $\mathbf{Y}$ with the leave-i-out predicted value $\hat{Y}_{i(i)}$ of the ith case:

$$\tilde{\mathbf{Y}} = (Y_1, \cdots, Y_{i-1}, \hat{Y}_{i(i)}, Y_{i+1}, \cdots, Y_n)^T.$$

- Let $\hat{\beta}_{(i)}$ be the leave-i-out LS fitted regression coefficients. Then $\hat{\beta}_{(i)}$ is also the LS fitted regression coefficients by using $\tilde{\mathbf{Y}}$ as the response vector, i.e. $\hat{\beta}_{(i)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\tilde{\mathbf{Y}}$. *Why?*
- So the leave-i-out fitted values for the n cases are:

$$\hat{\mathbf{Y}}_{(i)} = \mathbf{X}\hat{\beta}_{(i)} = H\tilde{\mathbf{Y}} = H(\mathbf{d}_{(i)}+\mathbf{Y}), \quad \mathbf{d}_{(i)} = \tilde{\mathbf{Y}}-\mathbf{Y} = (0, \cdots, -d_i, \cdots, 0)^T.$$

- Subtract the ith element from Y_i on both sides:

$$d_i = h_{ii}d_i + e_i \Longrightarrow d_i = \frac{e_i}{1-h_{ii}}.$$

Surgical Unit: Full Model X_1, X_2, X_3, X_4

```
> fit.f =lm(log(Y)~X1+X2+X3+X4, data=data.o)
> summary(fit.f)
Call:
lm(formula = log(Y) ~ X1 + X2 + X3 + X4, data = data.o)
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.851933   0.266263  14.467  < 2e-16 ***
X1          0.083739   0.028834   2.904  0.00551 **
X2          0.012671   0.002315   5.474 1.50e-06 ***
X3          0.015627   0.002100   7.440 1.38e-09 ***
X4          0.032056   0.051466   0.623  0.53627
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1
Residual standard error: 0.2509 on 49 degrees of freedom
Multiple R-squared: 0.7591,     Adjusted R-squared: 0.7395
F-statistic: 38.61 on 4 and 49 DF,  p-value: 1.398e-14
> anova(fit.f)
Analysis of Variance Table

Response: log(Y)
Df Sum Sq Mean Sq  F value    Pr(>F)
X1         1 0.7770  0.7770  12.3443 0.0009618 ***
X2         1 2.5904  2.5904  41.1565 5.341e-08 ***
X3         1 6.3286  6.3286 100.5490 1.838e-13 ***
X4         1 0.0244  0.0244   0.3879 0.5362698
Residuals 49 3.0841  0.0629
```

Surgical Unit: Full Model

- Full model has $P = 5$ and

  $$SSE = 3.0841, \;\; MSE = 0.0629, \;\; R^2 = 0.7591, \;\; R_a^2 = 0.7395.$$

- By definition, for the full model, $C_P =$.
- Sample size $n = 54$, so for the full model:
  $AIC_P =$
  and $BIC_P =$
- $Press_p = 4.069.$

```
> e.f=fit.f$residuals  ## residuals
> h.f=influence(fit.f)$hat  ## diagonals of hat matrix
> press.f= sum(e.f^2/(1-h.f)^2)  ## calculate press
```

Surgical Unit: Full Model

- Full model has $P = 5$ and

  $SSE = 3.0841, \;\; MSE = 0.0629, \;\; R^2 = 0.7591, \;\; R_a^2 = 0.7395.$

- By definition, for the full model, $C_P = P = 5$.
- Sample size $n = 54$, so for the full model:
  $AIC_P = 54\log(3.0841/54) + 2 \times 5 = -144.5871$ and
  $BIC_P = 54\log(3.0841/54) + \log(54) \times 5 = -134.6422$.
- $Press_p = 4.069$.

```
> e.f=fit.f$residuals  ## residuals
> h.f=influence(fit.f)$hat  ## diagonals of hat matrix
> press.f= sum(e.f^2/(1-h.f)^2)  ## calculate press
```

Surgical Unit: Null Model

The *null model* refers to the model with no X variable and only the intercept term:

$$Y_i = \beta_0 + \epsilon_i, \quad i = 1, \cdots, n.$$

- $SSE =$, $R^2 =$, $R_a^2 =$, $p =$.
- Use $MSE = 0.0629$ from the "full model", $C_p =$ Here, $C_p =$, indicating model bias.
- $AIC_p = 54\log(12.80451/54) + 2 \times 1 = -75.71608$ and $BIC_p = 54\log(12.80451/54) + \log(54) \times 1 = -73.72709$.

What is $Press_p$ of the null model?

Surgical Unit: Null Model

The *null model* refers to the model with no X variable and only the intercept term:

$$Y_i = \beta_0 + \epsilon_i, \quad i = 1, \cdots, n.$$

- $SSE = SSTO = 12.80451, \quad R^2 = 0, \quad R_a^2 = 0, \quad p = 1.$
- Use $MSE = 0.0629$ from the "full model",

$$C_p = \frac{12.80451}{0.0629} - (54 - 2 \times 1) = 151.5693.$$

  Here, $C_p = 151 >> p = 1$, indicating severe model bias.
- $AIC_p = 54 \log(12.80451/54) + 2 \times 1 = -75.71608$ and $BIC_p = 54 \log(12.80451/54) + \log(54) \times 1 = -73.72709.$

What is $Press_p$ of the null model? For null model: $\hat{Y}_i \equiv \bar{Y}$ and $h_{ii} \equiv \frac{1}{n}$, so

$$Press_p = \sum_{i=1}^{n} \frac{(Y_i - \overline{Y})^2}{(1 - \frac{1}{n})^2} = (\frac{n}{n-1})^2 \times SSTO.$$

Model Search Procedures

- The number of possible models, 2^{P-1}, grows very fast with the number potential X variables $P - 1$.
- Evaluating every possible model can be computationally infeasible even for moderate P.
- A variety of search procedures have been developed to efficiently search for the "best" model(s) in the model space.
  - Stepwise regression procedures
  - Best subsets algorithms: Not applicable when the pool of potential X variables is large, say more than 30.

Stepwise Regression Procedures

These are procedures that search for the "best" subset in a **sequential** manner.

- They are applicable to situations with a large number of potential X variables.
- They rely on "greedy" search strategies by developing a sequence of models, at each step adding or deleting only one X variable according to a pre-specified criterion (e.g., AIC).
- Stepwise procedures could end up with a *suboptimal model* rather than the global "best" model.
- Commonly used stepwise procedures include: *forward stepwise*, *forward selection*, *backward stepwise* and *backward elimination*.

Forward Stepwise Procedure

We need to specify the following.

- A model selection criterion, e.g., *AIC*.
- An initial model $\mathbb{M}_0$, usually a small model, e.g., the null-model with no X variable.
- The pool of potential X variables $\mathcal{X}$.
- The set of X variables that will always be in the model $\mathcal{X}_0$, e.g., the intercept term.

Starting from the initial model $\mathbb{M}_0$, at each step:

(a) Consider the X variables in the potential pool $\mathcal{X}$ that are not currently in the model. Examine the change in the criterion by adding each such variable into the current model.

(b) Consider the X variables that are already in the model but not in the set $\mathcal{X}_0$. Examine the change in the criterion by dropping each such variable out of the current model.

- Choose the operation that improves the criterion the most and update the current model accordingly. Repeat steps (a) and (b) for the updated model.
- If there is no operation that can improve the criterion anymore, then stop the search procedure and return the current model as the selected model.

Forward Selection and Backward Elimination

- Forward selection is a simplified version of forward stepwise procedure, omitting the considerations of dropping a variable currently in the model at each step.
- Backward elimination is the opposite of the forward selection.
  - It starts with a “big” initial model, e.g., the full model.
  - At each step, it examines the change of the criterion by dropping a variable currently in the model.
  - It then chooses the operation that improves the criterion the most to update the current model. It stops if no operation is able to improve the criterion anymore.
- Backward stepwise procedure. *Guess what is it?*

`stepAIC ()` Function

We can use the `stepAIC()` function in the `MASS` library to perform various stepwise regression.

- `direction=''both''` corresponds to forward stepwise procedure or backward stepwise procedure (depending on the initial model); `direction=''forward''` corresponds to froward selection; `direction=''backward''` corresponds to backward elimination.
- The option `scope` specifies the potential pool of X variables (`upper`) and the X variables that should always be included in the model (`lower`).
- `k=2` corresponds to *AIC* criterion; `k=log(n)` corresponds to *BIC* criterion.

Surgical Unit: Forward Stepwise

Start with the null-model.

```
> library(MASS)
> fit.0 =lm(log(Y)~1, data=data.o)  ## initial model: null-model with only intercept term
> step.0=stepAIC(fit.0,scope=list(upper=~X1+X2+X3+X4+X5+X6+X7+X8, lower=~1), direction="both", k=2)
Start:  AIC=-75.72
log(Y) ~ 1
Df Sum of Sq     RSS      AIC
+ X3    1    5.4708  7.3337 -103.811
+ X4    1    5.3967  7.4079 -103.268
+ X2    1    2.8303  9.9742  -87.205
+ X8    1    1.7808 11.0238  -81.802
+ X1    1    0.7770 12.0275  -77.096
+ X6    1    0.6889 12.1156  -76.703
<none>              12.8045  -75.716
+ X5    1    0.2694 12.5351  -74.864
+ X7    1    0.2067 12.5978  -74.595

Step:  AIC=-103.81
log(Y) ~ X3
Df Sum of Sq     RSS      AIC
+ X2    1    3.0209  4.3129 -130.479
+ X4    1    2.2018  5.1319 -121.089
+ X1    1    1.5512  5.7825 -114.644
+ X8    1    1.1386  6.1951 -110.922
<none>               7.3337 -103.811
+ X6    1    0.2582  7.0755 -103.747
+ X5    1    0.2390  7.0947 -103.600
+ X7    1    0.0659  7.2679 -102.298
- X3    1    5.4708 12.8045  -75.716
```

Surgical Unit: Forward Stepwise (Cont'd)

```
Step:  AIC=-130.48
log(Y) ~ X3 + X2

Df Sum of Sq     RSS       AIC
+ X8    1     1.4709 2.8420 -151.002
+ X1    1     1.2044 3.1085 -146.161
+ X4    1     0.6979 3.6150 -138.011
+ X7    1     0.2280 4.0849 -131.412
+ X5    1     0.1648 4.1481 -130.583
<none>               4.3129 -130.479
+ X6    1     0.0822 4.2306 -129.518
- X2    1     3.0209 7.3337 -103.811
- X3    1     5.6613 9.9742  -87.205

Step:  AIC=-151
log(Y) ~ X3 + X2 + X8
Df Sum of Sq     RSS       AIC
+ X1    1     0.6642 2.1778 -163.376
+ X4    1     0.4658 2.3761 -158.669
+ X6    1     0.1372 2.7048 -151.674
<none>               2.8420 -151.002
+ X5    1     0.0709 2.7711 -150.367
+ X7    1     0.0241 2.8179 -149.462
- X8    1     1.4709 4.3129 -130.479
- X2    1     3.3531 6.1951 -110.922
- X3    1     4.9403 7.7823  -98.605
```

Surgical Unit: Forward Stepwise (Cont'd)

```
Step:  AIC=-163.38
log(Y) ~ X3 + X2 + X8 + X1

Df Sum of Sq     RSS      AIC
+ X6    1     0.0966 2.0812 -163.826
<none>               2.1778 -163.376
+ X5    1     0.0760 2.1018 -163.293
+ X4    1     0.0415 2.1363 -162.415
+ X7    1     0.0224 2.1554 -161.935
- X1    1     0.6642 2.8420 -151.002
- X8    1     0.9307 3.1085 -146.161
- X2    1     2.9891 5.1670 -118.722
- X3    1     5.4459 7.6237  -97.717

Step:  AIC=-163.83
log(Y) ~ X3 + X2 + X8 + X1 + X6
Df Sum of Sq     RSS     AIC
+ X5    1     0.0769 2.0043 -163.86
<none>               2.0812 -163.83
- X6    1     0.0966 2.1778 -163.38
+ X7    1     0.0219 2.0593 -162.40
+ X4    1     0.0163 2.0649 -162.25
- X1    1     0.6236 2.7048 -151.67
- X8    1     0.9754 3.0567 -145.07
- X2    1     2.8287 4.9099 -119.48
- X3    1     5.0742 7.1554  -99.14
```

Surgical Unit: Forward Stepwise (Cont'd)

```
Step:  AIC=-163.86
log(Y) ~ X3 + X2 + X8 + X1 + X6 + X5
Df Sum of Sq     RSS      AIC
<none>                2.0043 -163.858
- X5     1     0.0769 2.0812 -163.826
- X6     1     0.0975 2.1018 -163.293
+ X7     1     0.0326 1.9718 -162.743
+ X4     1     0.0022 2.0021 -161.919
- X1     1     0.6284 2.6327 -151.133
- X8     1     0.9011 2.9054 -145.810
- X2     1     2.7644 4.7688 -119.052
- X3     1     5.0752 7.0795  -97.716

> step.0$anova
Stepwise Model Path
Analysis of Deviance Table
Initial Model:
log(Y) ~ 1
Final Model:
log(Y) ~ X3 + X2 + X8 + X1 + X6 + X5
Step Df   Deviance Resid. Df Resid. Dev        AIC
1                         53  12.804509  -75.71608
2 + X3  1 5.47078352       52   7.333726 -103.81102
3 + X2  1 3.02085553       51   4.312870 -130.47855
4 + X8  1 1.47089284       50   2.841977 -151.00214
5 + X1  1 0.66416961       49   2.177808 -163.37593
6 + X6  1 0.09659084       48   2.081217 -163.82569
7 + X5  1 0.07688125       47   2.004335 -163.85826
```

Surgical Unit: Forward Stepwise (Cont'd)

- The selected model is $X_1, X_2, X_3, X_5, X_6, X_8$ ($p = 7$) with $AIC_p = -163.858$.
- In this case, the forward selection procedure also selects the same model.

```
## forward selection
> step.0.f=stepAIC(fit.0,  scope=list(upper=~X1+X2+X3+X4+X5+X6+X7+X8, lower=~1),
+ direction="forward", k=2)
```

Surgical Unit: Backward Elimination

Start with the full model with all eight predictors.

```
> fit.f =lm(log(Y)~., data=data.o)
> step.b=stepAIC(fit.f, scope= list(upper=~., lower=~1), direction="backward", k=2)
Start:  AIC=-160.78
log(Y) ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8

Df Sum of Sq    RSS     AIC
- X4     1   0.00126 1.9718 -162.74
- X7     1   0.03159 2.0021 -161.92
- X5     1   0.07359 2.0441 -160.80
<none>               1.9705 -160.78
- X6     1   0.08403 2.0545 -160.52
- X1     1   0.31845 2.2890 -154.69
- X8     1   0.84489 2.8154 -143.51
- X2     1   2.09285 4.0634 -123.70
- X3     1   2.98863 4.9591 -112.94
```

Surgical Unit: Backward Elimination (Cont'd)

```
Step:  AIC=-162.74
log(Y) ~ X1 + X2 + X3 + X5 + X6 + X7 + X8
Df Sum of Sq    RSS      AIC
- X7     1    0.0326 2.0043 -163.858
<none>               1.9718 -162.743
- X5     1    0.0876 2.0593 -162.396
- X6     1    0.0969 2.0687 -162.152
- X1     1    0.6269 2.5987 -149.835
- X8     1    0.8438 2.8156 -145.506
- X2     1    2.6755 4.6473 -118.446
- X3     1    5.0934 7.0652  -95.825
Step:  AIC=-163.86
log(Y) ~ X1 + X2 + X3 + X5 + X6 + X8
Df Sum of Sq    RSS      AIC
<none>               2.0043 -163.858
- X5     1    0.0769 2.0812 -163.826
- X6     1    0.0975 2.1018 -163.293
- X1     1    0.6284 2.6327 -151.133
- X8     1    0.9011 2.9054 -145.810
- X2     1    2.7644 4.7688 -119.052
- X3     1    5.0752 7.0795  -97.716

## backward stepwise
> step.bs=stepAIC(fit.f, scope= list(upper=~., lower=~1),
+ direction="both", k=2)
```

Again the model $X_1, X_2, X_3, X_5, X_6, X_8$ is selected. Backward stepwise also selects the same model.

Stepwise Procedures: Comments

- Forward stepwise procedure often works better than forward selection when there is .
- Backward procedures are not good when the number of potential X variables, $P - 1$, is . Particularly, they are not feasible when P n, since then the full model can not be fitted.
- A potential disadvantage of forward procedures is the MSE and thus the standard errors of the LS estimators tend to be in the initial steps since important X variables are likely to be omitted in those steps.
- Another commonly used strategy is to perform one pass of forward selection followed by one pass of backward elimination.

Stepwise Procedures: Comments

- Forward stepwise procedure often works better than forward selection when there is high multicollinearity.
- Backward procedures are not good when the number of potential X variables, $P - 1$, is large. Particularly, they are not feasible when $P > n$, since then the full model can not be fitted.
- A potential disadvantage of forward procedures is the MSE and thus the standard errors of the LS estimators tend to be overestimated in the initial steps since important X variables are likely to be omitted in those steps. This will affect testing-based criteria.
- Another commonly used strategy is to perform one pass of forward selection followed by one pass of backward elimination.

Model Building and Selection: Comments

- For the sake of interpretability, it is often appropriate to select all the indicator variables corresponding to a qualitative variable as a group (i.e., to be in or out of the model simultaneously), even if a subset containing only part of them is “better” according to the model selection criterion.
- **Hierarchical principle**: If higher-order terms (e.g., interactions, powers) are selected, it is often appropriate to include the related lower-order terms as well.
- When the number of potential interactions is large, we could
    - use *aprior* knowledge to decide on the most likely interaction terms to be included.
    - first fit a first-order model and then plot the residuals against interaction terms to decide which interactions should be included. These plots may be limited to interactions that involve important X variables based on the first-order model.

Model Validation

- Internal validation: Check validity using **the same data** used to fit the model.
- External validation: Check validity using **new data** – either newly collected or a holdout sample *(i.e., cross-validation)*. *Which of the two validations is more trustworthy?*
- Compare results with theoretical expectations, previous results, and simulation results.

Internal Validation

We can use $Press_p$ and C_p to conduct internal validation of candidate models.

- $Press_p$ **is always**   **than** SSE_p as

$$|d_i| = |Y_i - \widehat{Y}_{i(i)}| = |\frac{Y_i - \widehat{Y}_i}{1 - h_{ii}}| \geq |Y_i - \widehat{Y}_i| = |e_i|, \quad i = 1, \cdots, n.$$

  - $Press_p/n$ can be viewed as an estimator of the *(out-of-sample) mean squared prediction error*:

$$mspe := E((\hat{y} - y)^2).$$

  - It is a measure for the   of the model.
  - $Press_p$ not much larger than SSE_p means there is   by the model.
- $C_p \approx p$ indicates   in the model, whereas $C_p >> p$ indicates   model bias.

Internal Validation

We can use $Press_p$ and C_p to conduct internal validation of candidate models.

- $Press_p$ **is always larger than** SSE_p as

$$|d_i| = |Y_i - \widehat{Y}_{i(i)}| = |\frac{Y_i - \widehat{Y}_i}{1 - h_{ii}}| \geq |Y_i - \widehat{Y}_i| = |e_i|, \quad i = 1, \cdots, n.$$

  - $Press_p/n$ can be viewed as an estimator of the *(out-of-sample) mean squared prediction error*:

$$mspe := E((\hat{y} - y)^2).$$

  - It is a measure for the predictive ability of the model.
  - $Press_p$ not much larger than SSE_p means there is no severe overfitting by the model.
- $C_p \approx p$ indicates little bias in the model, whereas $C_p >> p$ indicates substantial model bias.

Cross-Validation

When sample size is sufficiently large, we can split the data into two sets, a *training data* used to develop the model and a *validation data* used to check model validity.

- Validation data is used to check consistency of the fitted parameters and predictive ability.
- Training data should be sufficiently large (e.g., n/P at least 6) so that a reliable model can be developed based on it. Sometimes, the validation data will have to be smaller.
- Once a final model has been validated and chosen, it is a common practice to use the entire data set to re-fit the final model.

Mean Squared Prediction Error

Mean squared prediction error (MSPE) on the validation data:

$$MSPE_v = \frac{\sum_{j=1}^{m}(Y_j - \widehat{Y}_j)^2}{m},$$

where m is the sample size of the validation data, Y_j is the *jth* observation in the validation data, and $\widehat{Y}_j$ is the predicted value of the *jth* case in the validation data based on the model fitted on the training data.

- $MSPE_v$ can be viewed as an estimator of the *(out-of-sample) mean squared prediction error* and thus a measure for the predictive ability of the model.
- $MSPE_v$ is usually than SSE/n, since the model is fitted on the training data and thus it naturally would fit the training data than it fits the validation data.
- If $MSPE_v$ is not much larger than SSE/n, then there is by the model.

Mean Squared Prediction Error

Mean squared prediction error (MSPE) on the validation data:

$$MSPE_v = \frac{\sum_{j=1}^{m}(Y_j - \widehat{Y_j})^2}{m},$$

where m is the sample size of the validation data, Y_j is the *jth* observation in the validation data, and $\widehat{Y_j}$ is the predicted value of the *jth* case in the validation data based on the model fitted on the training data.

- $MSPE_v$ can be viewed as an estimator of the *(out-of-sample) mean squared prediction error* and thus a measure for the predictive ability of the model.
- $MSPE_v$ is usually larger than SSE/n, since the model is fitted on the training data and thus it naturally would fit the training data better than it fits the validation data.
- If $MSPE_v$ is not much larger than SSE/n, then there is no severe overfitting by the model.

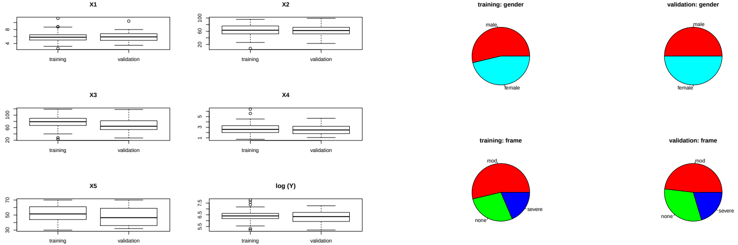
Surgical Unit: Internal Validation

Three "best" models according to various criteria.

- By BIC_p and $Press_p$: Model 1, $\log Y \sim X_1, X_2, X_3, X_8$.
  - $p = 5, SSE_p = 2.178, C_p = 5.734, Press_p = 2.736$.
- By C_p: Model 2, $\log Y \sim X_1, X_2, X_3, X_6, X_8$.
  - $p = 6, SSE_p = 2.081, C_p = 5.528, Press_p = 2.782$.
- By $R^2_{a,p}$ and AIC_p: Model 3, $\log Y \sim X_1, X_2, X_3, X_5, X_6, X_8$.
  - $p = 7, SSE_p = 2.004, C_p = 5.772, Press_p = 2.771$.
- For all three models, $Press_p$ and SSE_p are reasonably close and $C_p \approx p$, supporting their validity.

Surgical Unit: External Validation

Figure: Distributions of variables in training ($n = 54$) and validation ($n = 54$) sets.



No big difference in how variables are distributed in these two sets.

All three models have:

- Consistency in parameter estimation: same sign and similar magnitude between the two sets of estimated coefficients and their standard errors.
- $MSPE_v$ based on the validation data is not much larger than SSE/n and Press/n based on the training data.

Surgical Unit: Model 1 External Validation

```
## fit model 1 on training and validation sets
> fit1=lm(log(Y)~X1+X2+X3+X8, data=data.o)
> fit1.v=lm(log(Y)~X1+X2+X3+X8, data=data.v)

## get estimates and statistics
> est1=cbind(summary(fit1)$coefficients[,1:2], summary(fit1.v)$coefficients[,1:2])
> sse1=c(anova(fit1)["Residuals",2],anova(fit1.v)["Residuals",2])
> mse1=c(anova(fit1)["Residuals",3],anova(fit1.v)["Residuals",3])
> Rs1=c(summary(fit1)$adj.r.squared, summary(fit1.v)$adj.r.squared)
> press1= c(sum(fit1$residuals^2/(1-influence(fit1)$hat)^2),
+ sum(fit1.v$residuals^2/(1-influence(fit1.v)$hat)^2))

## MSPE on validation set
> newdata=data.v[,1:8] ## validation cases for prediction
> mspe1=c('NA', mean((predict.lm(fit1, newdata)-log(data.v$Y))^2))

## display results
> temp=cbind(sse1, mse1, Rs1, press1, press1/n,mspr1)
> rownames(temp)=c("Training", "Validation")
> colnames(temp)=c("sse","mse","R2_a","press","press/n", "mspe")
> round(est1,3)
> round(temp,3)
```

Surgical Unit: Model 1 External Validation (Cont'd)

```
Training              Validation
Estimate Std. Error Estimate Std. Error
(Intercept)     3.853      0.193     3.635      0.289
X1              0.073      0.019     0.096      0.032
X2              0.014      0.002     0.016      0.002
X3              0.015      0.001     0.016      0.002
X8              0.353      0.077     0.186      0.096

                sse   mse  R2_a press press/n  mspe
Training   2.178 0.044 0.816 2.736  0.051   --
Validation 3.794 0.077 0.682  --     --     0.077
```

Surgical Unit: Model 2 External Validation

```
Training              Validation
Estimate Std. Error Estimate Std. Error
(Intercept)     3.867      0.191    3.614      0.291
X1              0.071      0.019    0.100      0.032
X2              0.014      0.002    0.016      0.002
X3              0.015      0.001    0.015      0.002
X6              0.087      0.058    0.073      0.079
X8              0.363      0.077    0.189      0.097

             sse   mse  R2_a press press/n  mspe
Training   2.081 0.043 0.821 2.782   0.052    --
Validation 3.728 0.078 0.682  --       --   0.076
```

Surgical Unit: Model 3 External Validation

	Training		Validation	
	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	4.054	0.235	3.470	0.347
X1	0.072	0.019	0.099	0.032
X2	0.014	0.002	0.016	0.002
X3	0.015	0.001	0.016	0.002
X5	-0.003	0.003	0.003	0.003
X6	0.087	0.058	0.073	0.079
X8	0.351	0.076	0.193	0.097

	sse	mse	R2_a	press	press/n	mspe
Training	2.004	0.043	0.823	2.771	0.051	--
Validation	3.681	0.078	0.679	--	--	0.079

Surgical Unit: Choice of Final Model

- $MSPE_v$ of the three models have similar values, indicating that they have similar predictive ability.
- Model 3 has one estimated regression coefficient changing sign from training data to validation data, probably due to relatively large SE of this coefficient. So it is eliminated from further consideration.
- Models 1 and 2 perform similarly in validation. Based on the *principle of parsimony*, we choose Model 1 as the final model.
- Fit Model 1 on all data ($n = 108$):

$$\begin{aligned}\log & (\text{Survial Time}) = 3.76 + 0.084 \times \text{clotting score} \\ + & \; 0.015 \times \text{ prognostic index} + 0.016 \times \text{enzyme score} \\ + & \; 0.265 \times I(\text{severe use of alcohol}).\end{aligned}$$

Surgical Unit: Final Model Fitted on All Data

```
Call:
lm(formula = log(Y) ~ X1 + X2 + X3 + X8, data = rbind(data.o,
data.v))
Residuals:
Min       1Q   Median       3Q      Max
-0.60369 -0.15201  0.00977  0.13175  0.57726
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.756276   0.162825  23.069  < 2e-16 ***
X1          0.083744   0.016781   4.990 2.46e-06 ***
X2          0.014988   0.001409  10.641  < 2e-16 ***
X3          0.015690   0.001134  13.839  < 2e-16 ***
X8          0.265096   0.060045   4.415 2.50e-05 ***
---
Signif. codes:  0 ?***?0.001 ?**?0.01 ?*?0.05 ?.?0.1 ? ?1
Residual standard error: 0.2446 on 103 degrees of freedom
Multiple R-squared: 0.7642,     Adjusted R-squared: 0.755
F-statistic: 83.45 on 4 and 103 DF,  p-value: < 2.2e-16
> anova(fit1.all)
Analysis of Variance Table

Response: log(Y)
Df  Sum Sq Mean Sq F value    Pr(>F)
X1          1  1.0809  1.0809  18.064 4.703e-05 ***
X2          1  6.5415  6.5415 109.322 < 2.2e-16 ***
X3          1 11.1859 11.1859 186.940 < 2.2e-16 ***
X8          1  1.1663  1.1663  19.492 2.498e-05 ***
Residuals 103  6.1632  0.0598
---
Signif. codes:  0 ?***?0.001 ?**?0.01 ?*?0.05 ?.?0.1 ? ?1
```