Linear Regression

Professor Jie Peng, PhD

Department of Statistics

University of California, Davis

Model Building: Overview

Model-Building Steps

- Data collection and processing
- Exploratory data analysis
- Preliminary model investigation
- Model selection
- Model diagnostics and validation

Case Study: Surgical Unit

A hospital surgical unit was interested in predicting survival times of patients (in days, ascertained in a follow-up study) undergoing a particular type of liver operation. 108 such patients were randomly selected for this study. The following variables were measured for each patient: blood clotting score, prognostic index, enzyme function test score, liver function test score, age (in years), gender (male or female) and history of alcohol use (none, moderate or severe). We use half of the data to build the model (**training data**) and use the other half to perform model validation (**validation data**) later.

hold out data

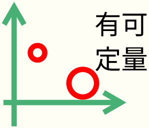
Model Building:

Exploratory Data Analysis

Exploratory Data Analysis

定量 quantity 定性 quanlity

- Type of each variable: quantitative or qualitative?
- Distribution of each variable: symmetric or skewed? outliers?
  - 定量 Quantitative: histogram, boxplot, summary statistics, etc.
  - 定性 Qualitative: pie chart, frequency table, etc.
- Relationships among variables:
  - scatter plot matrix, correlation matrix, side-by-side box plots
  - nonlinear pattern? clusters? outliers?

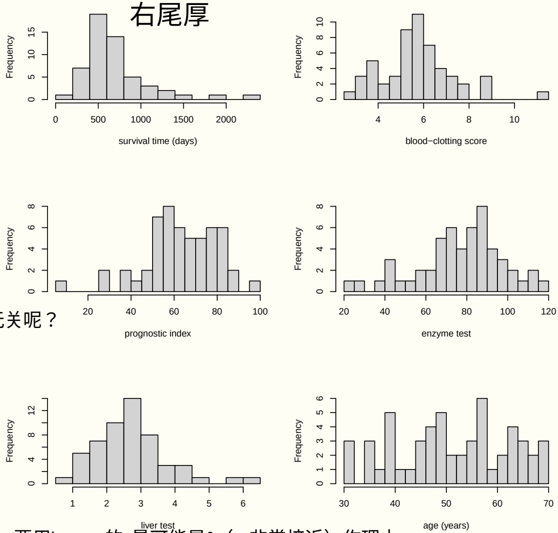


有可能
定量的被分为定性

Figure: Histograms of quantitative variables

right-skewed=可以用log(y)来fix
右尾厚

response variable



其实是用boxcox
搞一个合适的λ
使得trans(Y)~X
的residual=Y-hatY
最像normal

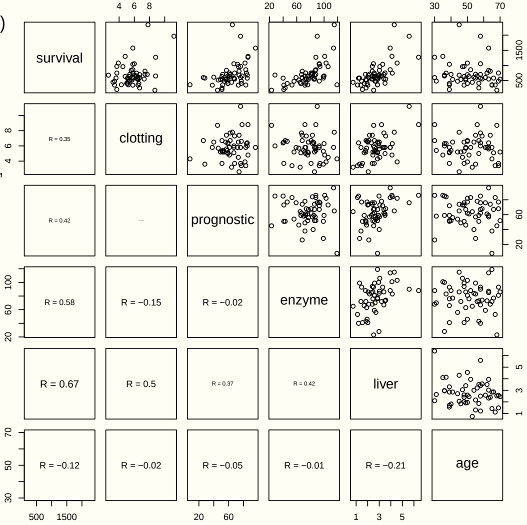
根据经验
right-skewed的Y
的λ常常接近0
就是
trans=log的时候
见PPTweek3P26

但是
这个事情为什么与X无关呢？
其实有关
当X不是normal
Y很自然不是normal
e也可以是normal
但根据经验，
往往log就可以了，
不过分析的时候，
logY~N，
不能作为做log的理由，要用boxcox的λ最可能是0（or非常接近）作理由

```
panel.cor <- function(x, y){
  #usr <- par("usr")
  #on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- round(cor(x, y, use="complete.obs"), 2)
  txt <- paste0("R = ", r)
  cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt,
    cex = cex.cor * r)
}

pairs(~weight+waist+hip
+height+age, data=diabetes,
 lower.panel = panel.cor)
```

Figure: Scatter plot matrix of quantitative variables



定性

Figure: Pie charts of qualitative variables

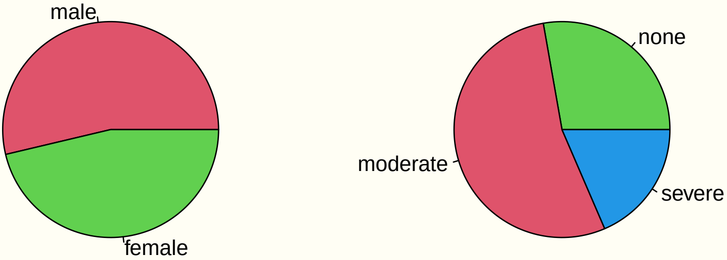
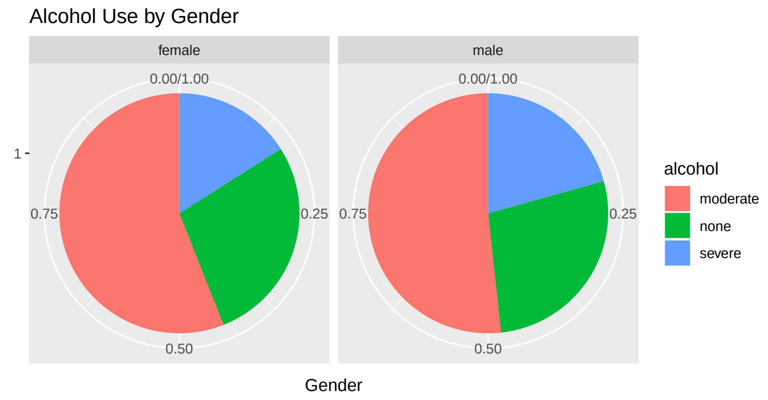
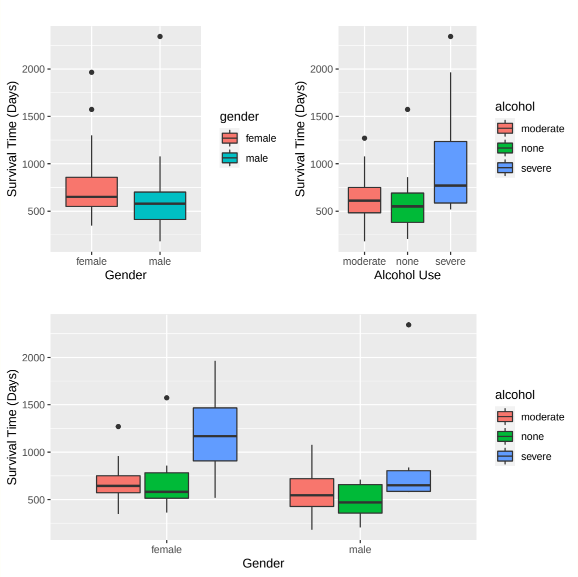


Figure: Side-by-side pie charts



file:///D:/t%E7%BB%9F%E8%AE%A1%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0/HW/HW3/HW3.html
有很棒的代码

Figure: Side-by-side box plots



Model Building:

Preliminary Fit

Preliminary Model Fitting

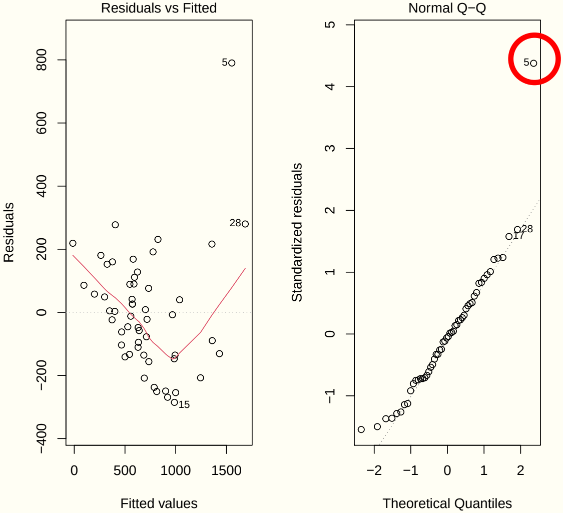
- Residual plots based on initial fits:
  - nonlinearity? departure from Normality? nonconstant error variance?

    ? log transformation
  - transformations needed?
  - interaction terms and/or high-order power terms?
- The goal is to decide on:
  - Functional forms in which variables should enter the model;
  - Potential pool of X variables to be considered in subsequent analysis;
- This process should be aided by prior knowledge and domain expertise if available.

Surgical Unit: First-Order Model

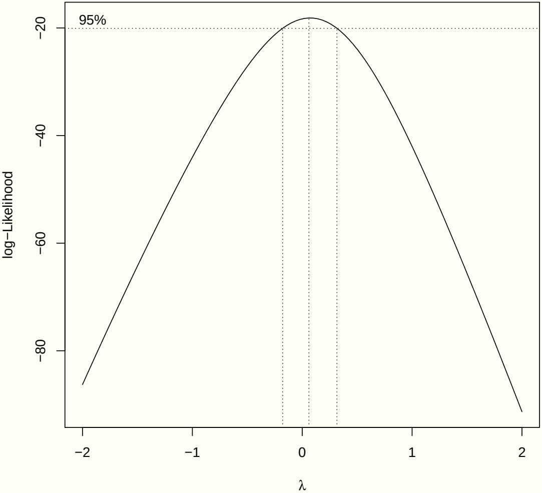
Fit a first-order model with survival time as response, and blood clotting score, prognostic index, enzyme function test score, liver function test score, age, gender (male or female) and history of alcohol use (none, moderate or severe) as X variables. Note that, gender and alcohol use should be treated as factors.

There appears to be non-linearity in regression relation. Residual Q-Q plot indicates outliers on the right tail.

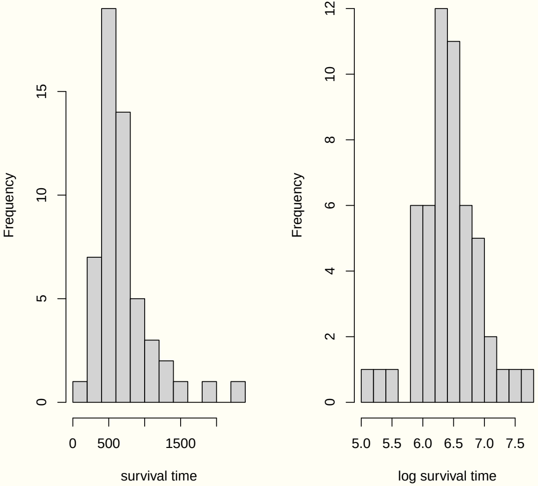


Box-cox procedure suggests logarithm transformation of the response variable.

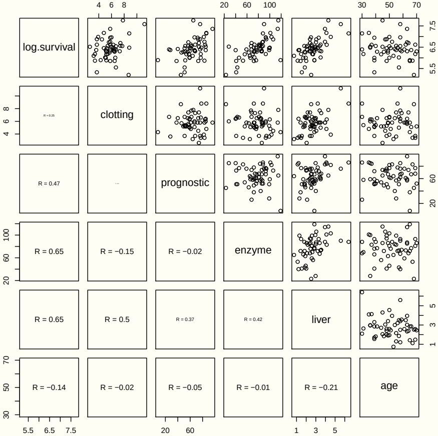
Linear Regression第一步
做个tranform of Y，
使得回归后的residual=Y-hatY
最像~N（0，..）

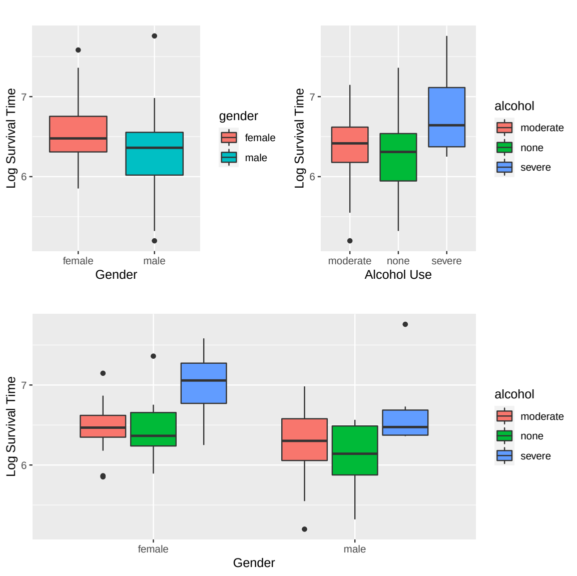


Surgical Unit: Log-Transformation

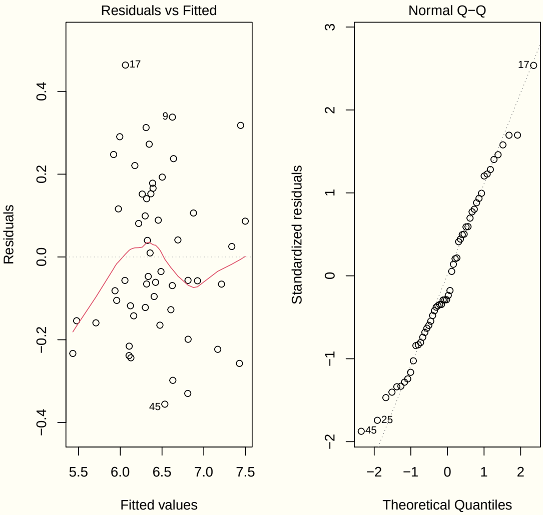


No obvious nonlinearity between log-survival-time and the quantitative *X* variables:




Log Survival Time
7
6
female
male
Gender
gender
female
male
Log Survival Time
7
6
moderate
none
severe
Alcohol Use
alcohol
moderate
none
severe
Log Survival Time
7
6
female
male
Gender
alcohol
moderate
none
severe

Fit the first-order model with log-survival-time as response: model assumptions appear to hold better.



Based on these preliminary fits, we decided to:

- use log-survival-time as the response variable;
- not include any interaction terms: this could be further examined by plotting residuals versus various interaction terms (e.g., those involving significant predictors).

Next, we should examine whether all predictors are needed or a subset of them is adequate in explaining log-survival-time $\Longrightarrow$ **model selection**

Bias-Variance Trade-off

Correct Models vs. Good Models

- Correct models are those that contain all important X variables $\Longrightarrow$ little model bias.
- However, a correct model is not necessarily a good model because it may include too many nuisance variables $\Longrightarrow$ large sampling variability and overfitting.
- A good model should contain all important X variables (correct: little bias), and at the same time it should have few nuisance variables (simple: small variability) $\Longrightarrow$ achieves *bias-variance trade-off.*

Example

$$Y = 1 + 2X_1 + 3X_2 + \epsilon$$

correct model=有显著变量(+垃圾变量)的model

- Any model contains (X_1, X_2) is a correct model, e.g., $\{X_1, x_2\}, \{X_1, X_2, X_1X_2\}$,v$\{X_1, X_2, X_1^2, X_2^2\}$, $\{X_1, X_2, X_3, X_4, X_5\}$.
    - These models have unbiased estimates.
    - However, some of them may have very large model variance such that the estimates behave erratically with even very small perturbation of the data.
- The models $\{X_1\}$ or $\{X_2\}$ both have an important X variable omitted and thus have substantial model bias.

In the following:

- Assume the response vector $\mathbf{Y}$ has $Var(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$.
- Let $\boldsymbol{\mu} = E(\mathbf{Y})$ denote the mean of the response vector.
- Let $\mathbb{M} = \mathbb{M}(X_1, \cdots, X_{p-1})$ denote an arbitrary model (**not necessarily** a correct model) and $\mathbf{X}$ denote its corresponding design matrix.
- Let $H(\mathbf{X}) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ be the hat matrix and $\hat{\mathbf{Y}} = \hat{\mathbf{Y}}(\mathbf{X}) = H(\mathbf{X})\mathbf{Y}$ be the fitted values vector.

Note that, $\mathbb{M}$ being a correct model means that there exists a vector $\boldsymbol{\beta}$ such that $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$.

Model Variance

- The (in-sample) variance of $\mathbb{M}$ is the overall variances of the fitted values:

$$Var_{in}(\mathbb{M}) := \sum_{i=1}^{n} Var(\hat{Y}_i) = Tr(Var(\hat{\mathbf{Y}})) = \sigma^2 Tr(H(\mathbf{X})) = p\sigma^2$$

- Therefore, larger models always have larger variances, whether they are correct or not.

Model Bias

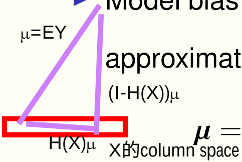
- The (in-sample) bias of $\mathbb{M}$ is the overall biases of the fitted values:

model不是correct model的时候，会有in sample bias

当我们没有include所有important的X的时候，会出现in sample bias

$$bias_{in}(\mathbb{M}) := \|E(\hat{\mathbf{Y}}) - E(\mathbf{Y})\|_2 = \|(H(\mathbf{X}) - \mathbf{I})\boldsymbol{\mu}\|_2$$

- Model bias depends on how well the column space $\langle \mathbf{X} \rangle$ approximates the mean response vector $\boldsymbol{\mu}$:



column space
C(X):=span(col(X))

span(col(X))的正交补空间

$$\boldsymbol{\mu} = E(\mathbf{Y}) = \boldsymbol{\mu}_X + \boldsymbol{\mu}_{X^\perp}, \quad \boldsymbol{\mu}_X \in \langle \mathbf{X} \rangle, \ \ \boldsymbol{\mu}_{X^\perp} \in \langle \mathbf{X} \rangle^\perp$$

symmetric 对称
idempotent 幂等

$$(H(\mathbf{X}) - \mathbf{I})\boldsymbol{\mu} = -\boldsymbol{\mu}_{X^\perp}, \quad bias_{in}^2(\mathbb{M}) = \boldsymbol{\mu}^T(\mathbf{I} - H(\mathbf{X}))\boldsymbol{\mu} = \|\boldsymbol{\mu}_{X^\perp}\|_2^2$$

- If $\mathbb{M}$ is a correct model, then $bias_{in}(\mathbb{M}) = 0$ because:

$$\boldsymbol{\mu} = E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \in \langle \mathbf{X} \rangle, \quad so, \quad \boldsymbol{\mu}_{X^\perp} = \mathbf{0}$$

Mean-Squared-Estiamtion-Error

有时候也叫"mse"，不是MSE=mean sum of squared error!

- Mean squared estimation error (msee) of $\hat{Y}_i$:

E[(θ - hatθ)^2]

"mean"

est.err

squared

=E[(hatYi - EhatYi + EhatYi - μi)^2]

=E[(hatYi-EhatYi)^2]+(EhatYi-μi)^2

+2(EhatYi-μi)E(hatYi-EhatYi)

=0

$$
\begin{aligned}
msee_i(\mathbb{M}) &:= E((\hat{Y}_i - \mu_i)^2) \\
&= Var(\hat{Y}_i) + (E(\hat{Y}_i) - \mu_i)^2
\end{aligned}
$$

bias(hatYi)^2

=E[(hatθ - Ehatθ + Ehatθ - θ)^2]

=E[(hatθ-Ehatθ)^2]+(Ehatθ-θ)^2+2(Ehatθ-θ)E(hatθ-Ehatθ)

=E[(hatθ-Ehatθ)^2]+(Ehatθ-θ)^2=Var(hatθ)+bias(hatθ)^2

- The (in-sample) msee of $\mathbb{M}$ equals model variance plus squared model bias:

对不同的hatYi，μi是不同的（由所有significant的Xi决定，这些Xi不一定在我们的model里面）

$$
\begin{aligned}
msee_{in}(\mathbb{M}) &:= \sum_{i=1}^{n} msee_i(\mathbb{M}) \\
&= Var_{in}(\mathbb{M}) + bias^2_{in}(\mathbb{M}) \\
&= p\sigma^2 + \|\boldsymbol{\mu}_{X^\perp}\|_2^2
\end{aligned}
$$

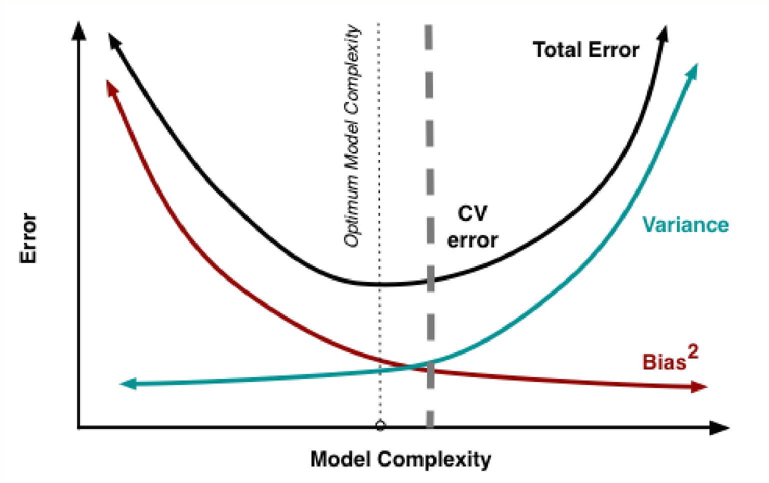
这里有tradeoff

minimize msee可以是我们的goal

p越大

线性空间dim越大
正交补空间dim越小
无法解释的μ越少

Bias-Variance Trade-off



$E(SSE)$ of a Model

- $SSE = \mathbf{e}^T\mathbf{e} = \mathbf{Y}^T(\mathbf{I} - H(\mathbf{X}))\mathbf{Y}$, is a measure of *goodness-of-fit* of the model to the **observed data Y**.
- $E(SSE)$ is affected by three factors: (i) model complexity p; (ii) error variance σ^2; (iii) and model bias $bias_{in}$.

$$
\begin{aligned}
E(SSE) &= E(Tr((\mathbf{I} - H(\mathbf{X}))\mathbf{Y}\mathbf{Y}^T)) = Tr((\mathbf{I} - H(\mathbf{X}))E(\mathbf{Y}\mathbf{Y}^T)) \\
&= Tr((\mathbf{I} - H(\mathbf{X}))(\sigma^2\mathbf{I} + \boldsymbol{\mu}\boldsymbol{\mu}^T)) \\
&= (n-p)\sigma^2 + \boldsymbol{\mu}^T(\mathbf{I} - H(\mathbf{X}))\boldsymbol{\mu} \\
&= (n-p)\sigma^2 + bias_{in}^2 \geq (n-p)\sigma^2
\end{aligned}
$$

p越大，E（SSE）一定越小，所以minimize E(SSE)不是我们的goal

- If $\mathbb{M}$ is a correct model, then $bias_{in}(\mathbb{M}) = 0$ and thus $E(SSE) = (n - p)\sigma^2$ and $E(MSE) = \sigma^2$.
- If $\mathbb{M}$ is an incorrect model, i.e., $\boldsymbol{\mu} = E(\mathbf{Y}) \notin \langle \mathbf{X} \rangle$, then $E(SSE) > (n - p)\sigma^2$ and $E(MSE) > \sigma^2$.

Summary

- Larger models have larger variances.
- Model bias depends on how well the column space of its design matrix approximates the mean response vector.
- For two correct models, the larger model has a smaller $E(SSE)$, but a larger variance and thus a larger overall mean-squared-estimation-error. So it tends to *overfit* the observed data.
- Incorrect models have larger $E(SSE)$ than correct models of the same size, so they tend to *underfit* the observed data.

Model Selection: Overview

Full Model vs. Candidate Model

- *Full model*: The model that contains all $P - 1$ potential X variables in the pool.
  - **Assume the full model is a correct model**.
- *Candidate model*: A model that contains a subset of $p - 1$ X variables with $1 \leq p \leq P$.
- The goal is to choose good model(s) (subset(s) of X variables) that balances bias and variance.

Key Components for Model Selection

- **Criterion to compare models**:
  - $R_a^2, C_p, AIC_p, BIC_p, Press_p$, etc.
- **Procedure to search for good model(s):**
  - *Best subset selection*: Exhaustive search; Applicable when the number of potential X variables is not too big ;
  - *Stepwise regression*: Greedy search; The number of potential X variables can be large;

Surgical Unit

If `clotting` (X_1), `prognostic` (X_2), `enzyme` (X_3), `liver` (X_4) form the potential pool of X variables, then there are 16 sub-models.

	选：								大	小	小	小	小
	p	intercept	X1	X2	X3	X4	sse	R^2	R^2_a	Cp	aic	bic	press
none model	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
full model	5	1	1	1	1	1	3.084	0.759*	0.739	5.000	-144.587	-134.642	4.069

cp值可以小于p的

The full model p
Cp≡p
by definition

不同criteria
不一定选中
同一个model

Model Selection: Criteria

Mallows' C_p Criterion

For full model:
Cp=(SSEp)/hatσ2-(n-2p)=(n-p)-(n-2p)=p

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

- n : sample size
- p: number of regression coefficients in the candidate model
- SSE_p: error sum of squares of the candidate model
- $\hat{\sigma}^2$: an unbiased estimator of the error variance σ^2:

need assumption
that full model is correct

$$\hat{\sigma}^2 = MSE_{\text{full model}} \quad \text{=SSEfull/(n-p)}$$

Mallows' C_p: Interpretation

Let $\mathbb{M} = \mathbb{M}(X_1, \cdots, X_{p-1})$ denote the candidate model, then

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

E[SSE(M)/hatσ^2]-(n-2p)
这里用了一个
非常不对的近似
E[SSE(M)/hatσ^2]
≈E(SSE(M))/σ^2
考试不可用这个

very rough

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

How to Use C_p?

- If a model has no bias, i.e., a correct model, 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.
- Look for models with (i) the C_p value not far above p and (ii) less X variables $\Longrightarrow$ small bias and small variance

AIC$_p$ and *BIC*$_p$ Criteria

在函数stepAIC( )里面叫做k

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

前项≈loglikelihood

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

n>8时
AIC会选更大的p
更注重拟合效果

- *Bayesian information criterion (BIC)*:

SSE

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

n>8时
BIC会选跟小的p
更注重模型正确性

- How to use: Look for models with small AIC (BIC)

当candidate models里有
correct model，BIC基本以概率1选中正确的model

AIC_p and BIC_p: Interpretation

- The first term: $n \log \frac{SSE_p}{n}$ reflects the *goodness-of-fit* of the model to the **observed data**:
  - decreases by adding more X variables into the model
- The second term, $2p$ for AIC and $(\log n)p$ for BIC, reflects model complexity:
  - 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 large SSE (underfitting, high bias).
- Overly complicated models may have a small SSE, but they have large model complexity (overfitting, high variance).
- By minimizing AIC (or BIC), we are trying to find a model that balances between model complexity and the goodness-of-fit.

衡量预测能力的判别标准

$Press_p$ Criterion

leave-one-out cross validation

Predicted residual sum of squares ($Press_p$):

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

第i个case 里面没有Xi 预测的Y=

- Y_i is the observed response of the *ith* case.
- $\widehat{Y}_{i(i)}$ is the 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.

hii=1可以取到
iff 列向量Pi在λ=0的地方都是0

0<= hii <= 1

hii=0可以取到
iff 列向量Pi在λ=1的地方都是0

$Press_p$: Calculation

hii = ei'Hei = ei'P' capitalλ Pei =Pi' capitalλ Pi
<= ||Pi||_2=1
又Pi' capitalλ Pi=sum j∈supp(λ) Pij^2 >= 0[not all λ=0，some =1]

Press$_p$ can be calculated without actually performing *n* regressions because the *deleted residual* for the *ith* case:

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

when fitting using all n cases

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

Hat matrix H cannot be trace zero！ so this inequation can never become equation！

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

>SSE

计算的时候用的是这个式子

Surgical Unit: Full Model

```
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

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


Analysis of Variance Table

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
```

- Full model has $P = 5$ and

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

- By definition, for the full model, $C_P = P = 5$
- Sample size $n = 54$, so for the full model:

$$AIC_P = 54 \log(3.0841/54) + 2 \times 5 = -144.5871 \text{ and}$$

$$BIC_P = 54 \log(3.0841/54) + \log(54) \times 5 = -134.6422$$

- $Press_p = 4.069$

influence

```
> 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
```

R code for Press_p

Model Selection: Stepwise Regression

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* Greedy Strategy
  - *Best subsets algorithms*: Not applicable when the pool of potential X variables is large.

Stepwise Regression Procedures

Stepwise可以从任何inital model开始
从none model（只有intercept term）开始=forward stepwise
从full model开始=backward stepwise

- Use “greedy” search strategies to examine a sequence of models by adding or deleting only one *X* variable according to a pre-specified criterion (e.g., *AIC*) at each search step.
- Could end up with a *local optimal model* rather than the global “best” model.
- Commonly used stepwise procedures: *forward stepwise*, *forward selection*, *backward stepwise* and *backward elimination*.

Forward Stepwise Procedure

Inputs:

- 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 terms 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 pool $\mathcal{X}$ that are not currently in the model. Examine the change of the criterion by adding each such variable into the current model. examine "what if Add one in"

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

(c) Choose the operation that improves the criterion the most and update the current model accordingly. Choose adding one in/droping one out iff AIC or BIC decreases and decreases most.

Repeat steps (a) – (c) until there is no operation that can improve the criterion anymore.

Forward Selection and Backward Elimination

- *Forward selection* is a simplified version of forward stepwise procedure by omitting the considerations of dropping a variable currently in the model at each step.

Forward selection不考虑在过程中drop

- *Backward elimination* is the opposite of the forward selection:

Backward elimination不考虑在过程中add

  - Start with a “big” initial model, e.g., the full model.
  - At each step, examine the change of the criterion by dropping a variable currently in the model.

- *Backward stepwise procedure*: opposite of forward stepwise.

Forward/Backward stepwise: 过程中既有add又有drop
区别在初始模型：F是从intercept term开始，B是从Full model开始

Stepwise Procedures: Comparisons

- Forward stepwise procedure often works better than forward selection when there is high multicollinearity among the potential X variables.
- Backward procedures are not good when the number of potential X variables is large. Particularly, they are not feasible when $P > n$, since then the full model can not be fitted.
- A commonly used alternative to forward stepwise procedure is to perform one pass of forward selection, followed by one pass of backward elimination.

没有stepBIC()【可以通过k=log(n)做到，这个时候就BIC】

`stepAIC()` Function in R library *MASS*

- `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`).

  full model 是 upper，不会选择比upper的X variables 之外的X变量了！

  lower model中的每一个X变量都总在model中
- `k=2` corresponds to *AIC* criterion; `k=log(n)` corresponds to *BIC* criterion.

Surgical Unit

initial model

full model

```
> fit.0=lm(log(survival)~1, data=data.o) ##initial model, only intercept
> step.aic=stepAIC(fit.0, scope=list(upper=~clotting+prognostic+enzyme+liver+age+gender
+alcohol.mod+alcohol.sev, lower=~1), direction="both", k=2, trace=FALSE)
```

~1就是总有intercept term

不要show过程

```
> step.aic$anova
Stepwise Model Path
Analysis of Deviance Table
Initial Model:
log(survival) ~ 1
Final Model:
log(survival) ~ enzyme + prognostic + alcohol.sev + clotting + gender + age
```

both就是

stepwise

k=2就是AIC

其实stepwise只有initial model之间的区别

F/B stepwise之间没有本质区别

```
Step              Df   Deviance Resid. Df Resid. Dev        AIC
1                                      53  12.804509  -75.71608
2       + enzyme   1 5.47078352        52   7.333726 -103.81102
3   + prognostic   1 3.02085553        51   4.312870 -130.47855
4  + alcohol.sev   1 1.47089284        50   2.841977 -151.00214
5     + clotting   1 0.66416961        49   2.177808 -163.37593
6       + gender   1 0.09659084        48   2.081217 -163.82569
7          + age   1 0.07688125        47   2.004335 -163.85826
```

initial model

maybe only this matters->this dummy v can be just "severe or not"

Model Validation

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.

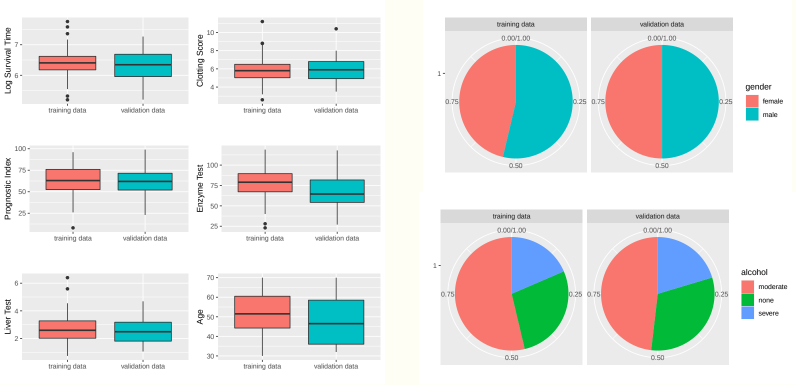
Training Data vs. Validation Data

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

- Training data should be sufficiently large so that a reliable model can be built from 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.

Surgical Unit: Training Data vs. Validation Data

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



Internal Validation by $Press_p$ and C_p

Press_p>SSE for sure:

Press_p =Σ_i [ei^2/(1-hii)^2] > Σei^2=SSE

- $Press_p$ is a measure of the predictive ability of the model: $Press_p$ not much larger than SSE_p means there is no severe over-fitting by the model.

no severe over-fitting

- $C_p \approx p$ indicates little bias in the model, whereas $C_p >> p$ indicates substantial model bias.

External Validation by Mean Squared Prediction Error

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

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.

- ► $MSPE_v$ is a measure of the predictive ability of the model.
- ► $MSPE_v$ is usually larger than SSE/n: $MSPE_v$ not much larger than SSE/n indicates no severe over-fitting 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: Model 1 External Validation

```
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
```

最大的mspe，估计误差最大

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 due to relatively large SE of this coefficient.
- Models 1 and 2 perform similarly in validation.

吝啬

- Based on the **principle of parsimony (“Occam’s Razor”)**, choose Model 1 as the final model and re-fit Model 1 on all data.

Surgical Unit: Model 1 Fitted on All Data

```
lm(formula = log(Y) ˜ X1 + X2 + X3 + X8, data = rbind(data.o,data.v))
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 ***
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

Analysis of Variance Table
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
```