Chapter 9

Model Selection and Validation

Instructor: Li C.X.

Outline

- Model-building process
- Criteria for model selection
- Search procedures for model selection
  - Best subsets algorithm
  - Stepwise, forward,…
- Model validation

9.1 Overview of model-building process

- Data Collection and preparation
- Reduction of explanatory or predictor variables (for exploratory observational studies)
- Model refinement and selection (This class!)
- Model validation

Data Collection Strategies

- Controlled Experiments – Subjects (Experimental Units) assigned to X-levels by Experimenter
  - Purely Controlled Experiments – Researcher only uses predictors that were assigned to units
  - Controlled Experiments with Covariates – Researcher has information (additional predictors) associated with units
- Observational Studies – Subjects (Units) have X-levels associated with them (not assigned by researcher)
  - Confirmatory Studies – New (primary) predictor(s) believed to be associated with Y, controlling for (control) predictor(s), known to be associated with Y
  - Exploratory Studies – Set of potential predictors believed that some or all are associated with Y

Reduction of Explanatory Variables

- Controlled Experiments
  - Purely Controlled Experiments – Rarely any need or desire to reduce number of explanatory variables
  - Controlled Experiments with Covariates – Remove any covariates that do not reduce the error variance
- Observational Studies
  - Confirmatory Studies – Must keep in all control variables to compare with previous research, should keep all primary variables as well
  - Exploratory Studies – Often have many potential predictors (and polynomials and interactions). Want to fit parsimonious model that explains much of the variation in Y, while keeping model as basic as possible.

Trouble in model selection

- Form any set of p predictors, 2^p different linear regression models can be constructed.
- Search in that space is exponentially difficult.
- Greedy strategies are typically utilized.
- Is this the only way?

9.2 Surgical unit example

- Surgical unit wants to predict survival in patients undergoing a specific liver operation
- Random sample of 108 patients
- Y is post-operation survival time
- Predictor variables:
  - $X1$: blood clotting score
  - $X2$: prognostic index
  - $X3$: enzyme function score
  - $X4$: liver function score
  - $X5$:age
  - $X6$:indicator for gender
  - $X7$ and $X8$: indicator for alcohol use

Survival Time as Response

- Often skewed with a few long-lived times
- In this case, we observe all survival times
- Times can be censored if the study were prior to some subjects' deaths
  - Survival analysis techniques could be used
- Use only first 54 of the 108 patients, and 4 predictors X1~X4 in the following analysis
- Transformation of survival times will be investigated using Box-Cox transformation
  - $Y' = \ln(Y)$
- $2^4 = 16$ models

```
alldat = read.table('surgical.txt')
dat0 = alldat[1:54,c(1:4, 9)]
names(dat0) = c('X1','X2','X3','X4','Y')
library(MASS)
fit = lm(Y~X1+X2+X3+X4,data=dat0)
bxcx = boxcox(fit)
```

9.3 Model Selection Criteria

- In order to select between models, some score must be given to each model.
- The likelihood of the data under each model is not sufficient because the likelihood of the data can always be improved by adding more parameters
- Accordingly some penalty that is a function of the complexity of the model must be included in the selection procedure.
- There are several choices for how to do this
  - Explicit penalization of the number of parameters in the model (AIC,BIC, etc.)
  - Implicit penalization through cross validation
  - Bayesian regularization (putting certain prior distribution on each model).

Model Selection Criteria

- Six Criteria

$$R_p^2, R_{a,p}^2, C_p, AIC_p, BIC_p(SBC_p), PRESS_p$$

Two distinct questions

- What is the appropriate subset size?
  - adjusted R^2 or MSE, C_p, PRESS, AIC, SBC
- What is the best model for a fixed size?
  - R^2

R^2 and adjusted R^2 Criterion

$p = \#$ of parameters in current model

R_p^2 or SSE_p criterion

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

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

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

Mallows' *Cp* Criterion

- Squared error for estimating μ_i

$$
\begin{aligned}
(\hat{Y}_i - \mu_i)^2 &= (\hat{Y}_i - \mathsf{E}(\hat{Y}_i) + \mathsf{E}(\hat{Y}_i) - \mu_i)^2 \\
&= (\mathsf{E}(\hat{Y}_i) - \mu_i)^2 + (\hat{Y}_i - \mathsf{E}(\hat{Y}_i))^2 + [E(\hat{Y}_i) - \mu_i][\hat{Y}_i - E(\hat{Y}_i)] \\
&= \mathsf{Bias}^2 + (\hat{Y}_i - \mathsf{E}(\hat{Y}_i))^2 + [E(\hat{Y}_i) - \mu_i][\hat{Y}_i - E(\hat{Y}_i)]
\end{aligned}
$$

- Mean value is $(\mathsf{E}(\hat{Y}_i) - \mu_i)^2 + \sigma^2(\hat{Y}_i)$
- Total mean value is $\sum(\mathsf{E}(\hat{Y}_i) - \mu_i)^2 + \sum \sigma^2(\hat{Y}_i)$
- Cp criterion compares total mean squared error with σ^2

$$
\begin{aligned}
\Gamma_p &= \frac{\sum(\mathsf{E}(\hat{Y}_i) - \mu_i)^2 + \sum \sigma^2(\hat{Y}_i)}{\sigma^2} \\
&= \frac{\sum \mathsf{Bias}^2 + \sum \mathsf{Var(prediction)}}{\mathsf{Var(error)}}
\end{aligned}
$$

Mallows' *Cp* Criterion

- Consider current model with $p-1$ predictors
  - Can show $E(\mathsf{SSE}_p) = \sum (E(\hat{Y}_i) - \mu_i)^2 + (n-p)\sigma^2$

$\mathrm{E}\mathbf{X} := \boldsymbol{\mu}$ *and* $\mathrm{Var}(\mathbf{X}) := \boldsymbol{\Sigma}$*, then*

$$\mathrm{E}\left(\mathbf{X}'\mathbf{A}\mathbf{X}\right) = \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}.$$

Proof: $\underset{n\times 1}{\mathbf{Y}} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \quad \mathbf{E}(\mathbf{Y}) = \boldsymbol{\mu}$

$$\underset{n\times 1}{\hat{\mathbf{Y}}} = \underset{n\times p}{\mathbf{X}}\ \underset{p\times 1}{\mathbf{b}} = \underset{n\times n}{\mathbf{H}}\ \mathbf{Y}, \quad \mathbf{E}(\hat{\mathbf{Y}}) = \mathbf{H}\boldsymbol{\mu}$$

$$SSE_p = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}$$

$$E(SSE_p) = E\left\{\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}\right\} = \mathrm{tr}\left[(\mathbf{I}-\mathbf{H})\right]\sigma^2 + \boldsymbol{\mu}'(\mathbf{I}-\mathbf{H})\boldsymbol{\mu}$$

$$= (n-p)\sigma^2 + \left[(\mathbf{I}-\mathbf{H})\boldsymbol{\mu}\right]'\left[(\mathbf{I}-\mathbf{H})\boldsymbol{\mu}\right]$$

$$= (n-p)\sigma^2 + \left[\boldsymbol{\mu} - \mathbf{E}(\hat{\mathbf{Y}})\right]'\left[\boldsymbol{\mu} - \mathbf{E}(\hat{\mathbf{Y}})\right]$$

$$= (n-p)\sigma^2 + \sum_{i=1}^{n} [E(\hat{Y}_i) - \mu_i]^2$$

Mallows' *Cp* Criterion

$$\Gamma_p = \frac{\sum (E(\hat{Y}_i) - \mu_i)^2 + \sum \sigma^2(\hat{Y}_i)}{\sigma^2} = \frac{\sum \text{Bias}^2 + \sum \text{Var(prediction)}}{\text{Var(error)}}$$

- Estimate σ^2 from the full model ($P-1$ predictors in total)

$$\hat{\sigma}^2 = \text{MSE}(X_1, X_2, ..., X_{P-1}) = \text{MSE}_P$$

- Consider current model with $p-1$ predictors

$$E(\text{SSE}_p) = \sum (E(\hat{Y}_i) - \mu_i)^2 + (n-p)\sigma^2$$

  - Estimate the bias part

$$\sum (E(\hat{Y}_i) - \mu_i)^2 \text{ by } \text{SSE}_p \text{ - } (n-p) \text{ MSE}_P$$

  - Variance part

$$\boldsymbol{\sigma}^2\left\{\hat{\mathbf{Y}}\right\} = \boldsymbol{\sigma}^2\left\{\mathbf{HY}\right\} = \mathbf{H}\boldsymbol{\sigma}^2\left\{\mathbf{Y}\right\}\mathbf{H}' = \boldsymbol{\sigma}^2\mathbf{H}$$

$$\sum \sigma^2(\hat{Y}_i) = \text{Trace}\{\boldsymbol{\sigma}^2(\hat{\mathbf{Y}})\} = \sigma^2 \text{Trace}\{\mathbf{H}\} = p\sigma^2$$

Mallows' *Cp* Criterion

- Putting it together, Γp is estimated by

$$\begin{aligned} C_p &= \frac{(\mathrm{SSE}_p - (n-p)\mathrm{MSE}_P) + p\mathrm{MSE}_P}{\mathrm{MSE}_P} \\ &= \frac{\mathrm{SSE}_p}{\mathrm{MSE}(X_1, X_2, \ldots, X_{P-1})} - (n-2p) \end{aligned}$$

$$\begin{aligned} \Gamma_p &= \frac{\sum (\mathrm{E}(\hat{Y}_i) - \mu_i)^2 + \sum \sigma^2(\hat{Y}_i)}{\sigma^2} \\ &= \frac{\sum \mathrm{Bias}^2 + \sum \mathrm{Var(prediction)}}{\mathrm{Var(error)}} \end{aligned}$$

- A good model has no bias

$$\Gamma_p = \frac{0 + p\sigma^2}{\sigma^2} = p; \quad E(C_p) \approx p;$$

- A bad model is biased

$$\Gamma_p > \frac{0 + p\sigma^2}{\sigma^2} = p; \quad E(C_p) > p;$$

AIC and SBC(BIC) Criteria

$$\ln L_p(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu_i),$$

where $\mu_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_{p-1} X_{p-1,i}$

$$\ln L_p(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2} - \frac{n}{2}\ln\left(\frac{SSE_p}{n}\right)$$

- AIC(Akaike's information criterion) and SBC(BIC) criterion are based on minimizing $-2\log(\text{likelihood})$ plus a penalty.

$$AIC_p = n\ln\left(\frac{SSE_p}{n}\right) + 2p$$

$$SBC_p = n\ln\left(\frac{SSE_p}{n}\right) + \left[\ln(n)\right]p$$

- AIC and BIC can be used to compare non-nested models

PRESSp Criterion

- Looks at the **PRE**diction **S**um of **S**quares which quantifies how well the fitted values can predict the observed responses
- For each case i, predict Y_i using model generated from other $n - 1$ cases

$$PRESS_p = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{i(i)} \right)^2$$

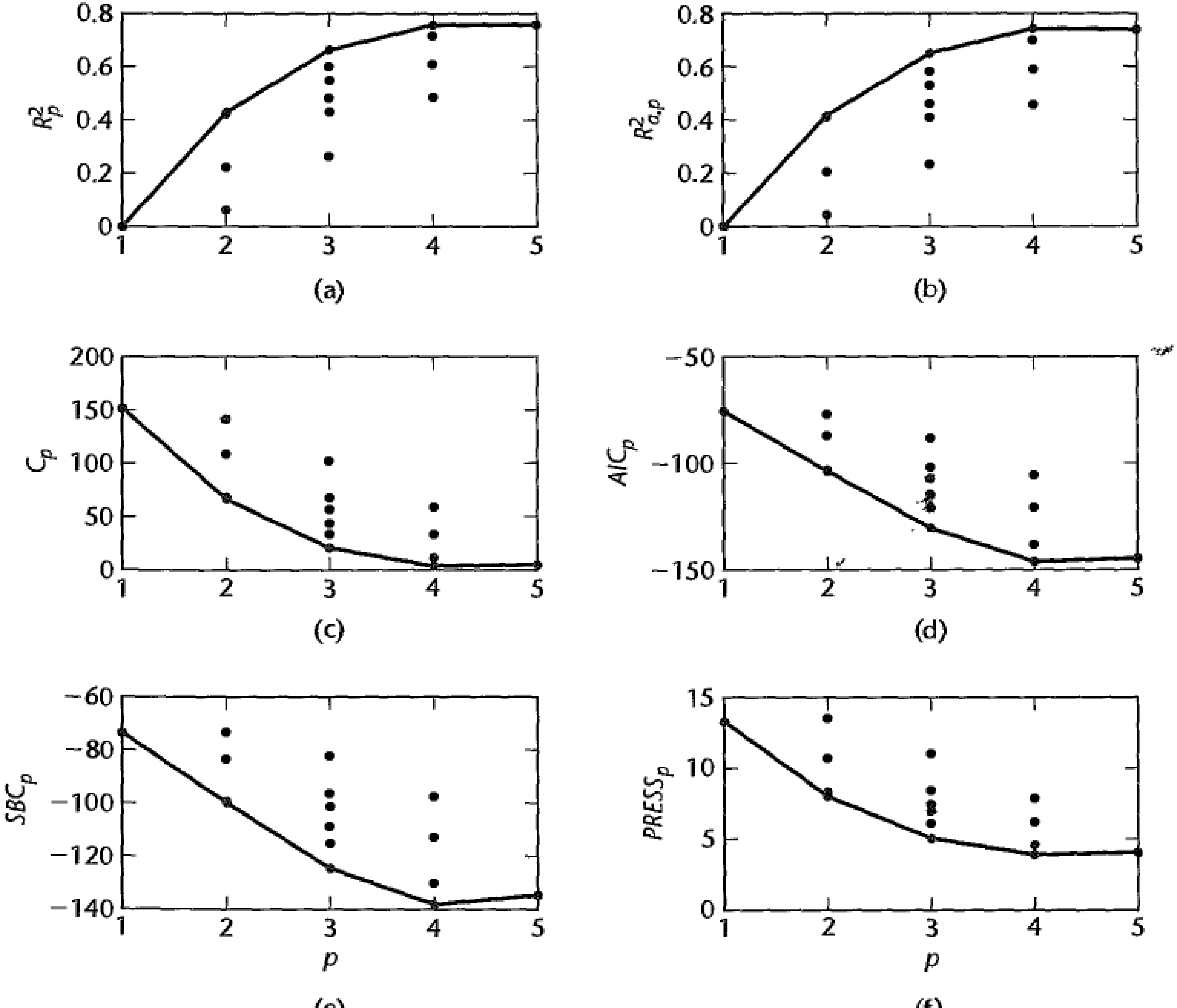
$\hat{Y}_{i(i)} \equiv$ fitted value for i^{th} case when it was not used in fitting model

- It's leave-one-out cross validation
- Can calculate this in one fit (Chapter 10)

Surgical unit example

- 16 models

X Variables in Model	(1) p	(2) SSE_p	(3) R^2_p	(4) $R^2_{a,p}$	(5) C_p	(6) AIC_p	(7) SBC_p	(8) $PRESS_p$
None	1	12.808	0.000	0.000	151.498	−75.703	−73.714	13.296
X_1	2	12.031	0.061	0.043	141.164	−77.079	−73.101	13.512
X_2	2	9.979	0.221	0.206	108.556	−87.178	−83.200	10.744
X_3	2	7.332	0.428	0.417	66.489	−103.827	−99.849	8.327
X_4	2	7.409	0.422	0.410	67.715	−103.262	−99.284	8.025
X_1, X_2	3	9.443	0.263	0.234	102.031	−88.162	−82.195	11.062
X_1, X_3	3	5.781	0.549	0.531	43.852	−114.658	−108.691	6.988
X_1, X_4	3	7.299	0.430	0.408	67.972	−102.067	−96.100	8.472
X_2, X_3	3	4.312	0.663	0.650	20.520	−130.483	−124.516	5.065
X_2, X_4	3	6.622	0.483	0.463	57.215	−107.324	−101.357	7.476
X_3, X_4	3	5.130	0.599	0.584	33.504	−121.113	−115.146	6.121
X_1, X_2, X_3	4	3.109	0.757	$\boxed{0.743}$	$\boxed{3.391}$	$\boxed{-146.161}$	$\boxed{-138.205}$	$\boxed{3.914}$
X_1, X_2, X_4	4	6.570	0.487	0.456	58.392	−105.748	−97.792	7.903
X_1, X_3, X_4	4	4.968	0.612	0.589	32.932	−120.844	−112.888	6.207
X_2, X_3, X_4	4	3.614	0.718	0.701	11.424	−138.023	−130.067	4.597
X_1, X_2, X_3, X_4	5	3.084	0.759	0.740	5.000	−144.590	−134.645	4.069



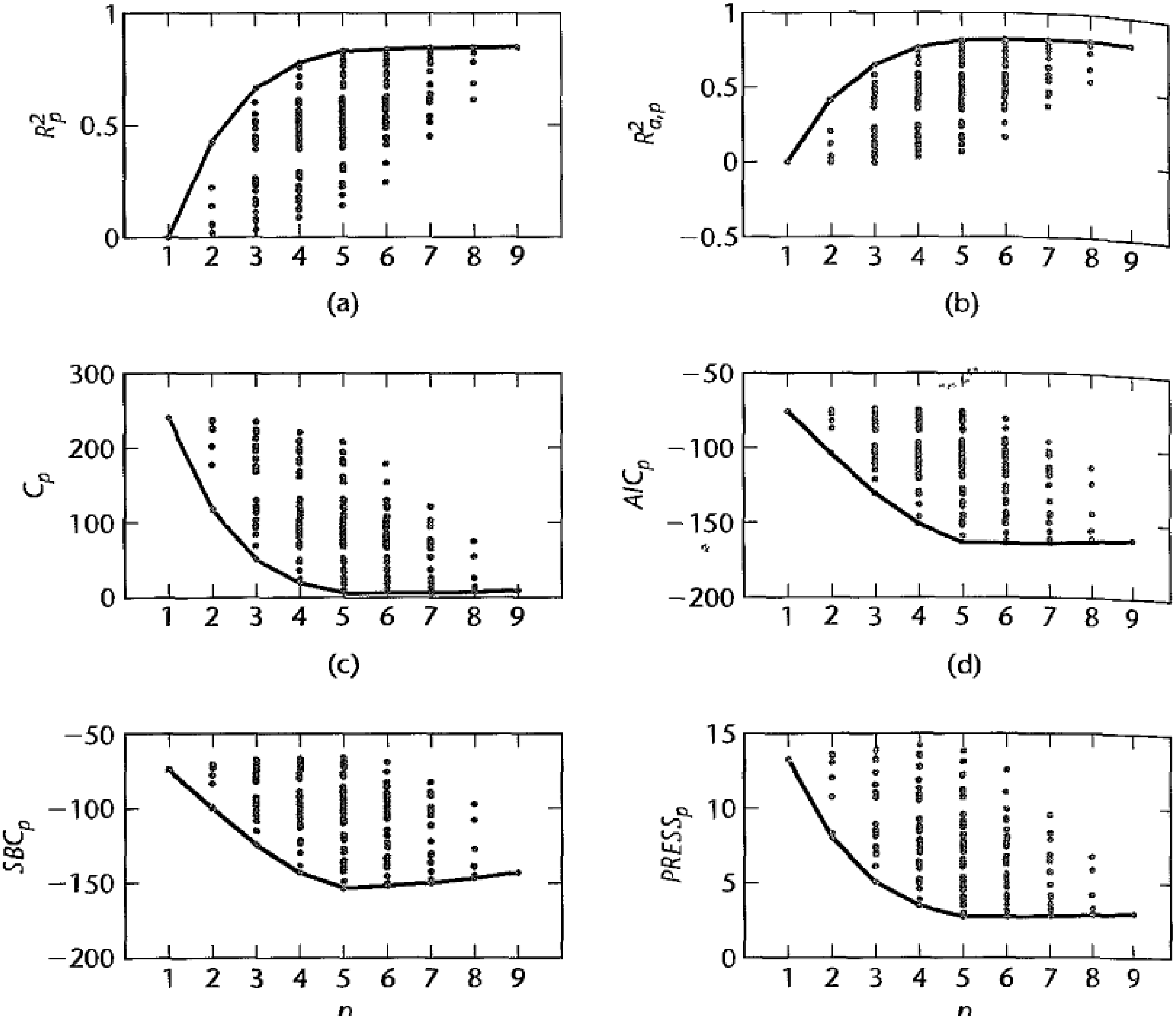
R_p^2
$R_{a,p}^2$
C_p
AIC_p
SBC_p
$PRESS_p$
p
p
(a)
(b)
(c)
(d)
(e)
(f)

9.4 Automatic search procedures for model selection

- Automated Procedures and all possible regressions:
  - “Best” subsets algorithm
  - Backward Elimination (Top down approach)
  - Forward Selection (Bottom up approach)
  - Stepwise Regression (Combines Forward/Backward)

Best subset search

- Consider all the possible subset. For each of the model, evaluate the criteria.
- Time-saving algorithms have been developed, which require the calculation of only a small fraction of all possible models.
- Still, if $P > 30$, it requires excessive computer time.
- Several regression models can be identified as "good" for final consideration, depending on which criteria we use.



1
0.5
0
R_p^2
1 2 3 4 5 6 7 8 9
(a)
1
0.5
0
−0.5
$R_{a,p}^2$
1 2 3 4 5 6 7 8 9
(b)
300
200
100
0
C_p
1 2 3 4 5 6 7 8 9
(c)
−50
−100
−150
−200
AIC_p
1 2 3 4 5 6 7 8 9
(d)
−50
−100
−150
−200
SBC_p
1 2 3 4 5 6 7 8 9
p
15
10
5
0
$PRESS_p$
1 2 3 4 5 6 7 8 9
p

Best subsets for surgical unit example

p	(1) SSE_p	(2) R^2_p	(3) $R^2_{a,p}$	(4) C_p	(5) AIC_p	(6) SBC_p	(7) $PRESS_p$
1	12.808	0.000	0.000	240.452	−75.703	−73.714	13.296
2	7.332	0.428	0.417	117.409	−103.827	−99.849	8.025
3	4.312	0.663	0.650	50.472	−130.483	−124.516	5.065
4	2.843	0.778	0.765	18.914	−150.985	−143.029	3.469
5	2.179	0.830	0.816	5.751	−163.351	<u>−153.406</u>	<u>2.738</u>
6	2.082	0.837	0.821	<u>5.541</u>	−163.805	−151.871	2.739
7	2.005	0.843	<u>0.823</u>	5.787	<u>−163.834</u>	−149.911	2.772
8	1.972	0.846	<u>0.823</u>	7.029	−162.736	−146.824	2.809
9	<u>1.971</u>	<u>0.846</u>	0.819	9.000	−160.771	−142.870	2.931

Backward Elimination

- Select a significance level to stay in the model (e.g. SLS=0.20, generally .05 is too low, causing too many variables to be removed)
- Start with all the variables. Fit the full model with all possible predictors
- Consider the predictor with lowest t-statistic (highest P-value).
  - If P > SLS, remove the predictor and fit model without this variable (must re-fit model here because partial regression coefficients change)
  - If $P \leq$ SLS, stop and keep current model
- Continue until all predictors have P-values below SLS
- R uses model based criteria: AIC, SBC instead

Forward Selection

- Choose a significance level to enter the model (e.g. SLE=0.20, generally .05 is too low, causing too few variables to be entered)
- Start with no variables
- Add one variable with highest *t* or F-value (only if p-value < SLE)
- Add the next variable with highest partial F-value given the previous variables in the model (only if p-value <SLE
- Continue until no new predictors have $P \leq$ SLE
- Note: R uses model based criteria: AIC, SBC instead

Stepwise Regression

- Select SLS and SLE (SLE<SLS)
- Starts like Forward Selection (Bottom up process)
- New variables must have $P \leq$ SLE to enter
- Re-tests all "old variables" that have already been entered, must have $P \leq$ SLS to stay in model
- Continues until no new variables can be entered and no old variables need to be removed
- Note: R uses model based criteria: AIC, SBC instead

R code

- full = lm(y~x1+x2+x3+x4+x5+x6+x7+x8,data=dat)
- null = lm(y~1, data=dat)

- Forward Stepwise Regression:
  - step(null, scope=list(upper=full, lower=null), direction='both')
- Forward Regression:
  - step(null, scope=list(upper=full, lower=null), direction='forward')
- Backward Elimination:
  - step(full, scope=list(upper=full, lower=null), direction='backward')

- If the number of variable is not large, it is best to fit all the possible models, and choose the one with the smallest AIC, BIC, Cp, PRESS.
- If the number of variable is too large, then, using stepwise forward regression is recommended.
- If $p > n$, then direct regularization techniques are needed, such as LASSO, SCAD, etc.

9.6 Model Validation

- When we have a lot of data, we would like to see how well a model fit on one set of data (training sample) compares to one fit on a new set of data (validation sample), and how the training model fits the new data.
- Training set should have at least 6-10 times as many observations than potential predictors
- Mean Square Prediction Error when training model is applied to validation sample:

$$MSPR = \frac{\sum_{i=1}^{n^*}\left(Y_i^V - \hat{Y}_i^V\right)^2}{n^*} \qquad \hat{Y}_i^V = b_0^T + b_1^T X_{i1}^V + \ldots + b_{p-1}^T X_{i,p-1}^V$$

R code

```
#####surgical example
alldat = read.table('surgical.txt')
names(alldat) = c(paste("X",1:8,sep=""), 'Y', 'logY')
dat0 = alldat[1:54,c(1:4, 9)];  dat = alldat[1:54,c(1:4, 10)]
X=alldat[1:54,1:4]
names(dat0) = c('X1','X2','X3','X4','Y')
names(dat) = c('X1','X2','X3','X4','logY')
fit = lm(logY~.,data=dat)
summary(fit)
plot(fit$fitted, fit$residuals)
qqnorm(fit$residuals)
```

```
#####Stepwise  selection for surgical example
full = lm(logY~.,data=alldat[1:54,c(1:8, 10)])
null = lm(logY~1, data=alldat[1:54,c(1:8, 10)])
step(null, scope=list(upper=full, lower=null), direction='both', trace=TRUE)
step(null, scope=list(upper=full, lower=null), direction='forward', trace=TRUE)
step(full, scope=list(upper=full, lower=null), direction='backward', trace=TRUE)
#The default criteria in "step" fuction is AIC
step(null, scope=list(upper=full, lower=null), direction='both', trace=TRUE,
k=log(54))  #set k=log(n), the criteria changed to be BIC

####add or drop one variable
add1 (null, ~X1+X2+X3+X4+X5+X6+X7+X8) #The default is AIC criteria
drop1 (full)
add1 (null, ~X1+X2+X3+X4+X5+X6+X7+X8,test='F')
drop1 (full,test='F')
```

```
#####best subset for surgical example
####using package "bestglm"
library(bestglm)
fit1 = bestglm(alldat[1:54,c(1:8, 10)],IC='LOOCV')
#LOOCV means leave-one-out cross validation.
# The criteria for is LOOCV is MSPE(Mean Square Prediction Error)= PRESSp/n
fit1$Subsets;    fit1$Subsets$LOOCV*54

fit2 = bestglm(alldat[1:54,c(1:8, 10)],IC='AIC')
fit2$Subsets

fit3 = bestglm(alldat[1:54,c(1:8, 10)],IC='BIC')
fit3$Subsets
```

```
#####best subset using package "leaps"
library(leaps)
leaps10<-regsubsets(logY~.,data=alldat[1:54,c(1:8, 10)],nbest=10)
#nbest means the max number of optimal model for each size
summary(leaps10)
# plot a table of models showing variables in each model.
# models are ordered by the selection statistic.
plot(leaps10,scale="r2");  plot(leaps10,scale="adjr2")
plot(leaps10,scale="Cp"); plot(leaps10,scale="bic")

leaps1<-regsubsets(logY~.,data=alldat[1:54,c(1:8, 10)],nbest=1)
summary(leaps1);    plot(leaps1,scale="adjr2");
plot(leaps10,scale="Cp");   plot(leaps10,scale="bic")
```

Homework

- P377

9.11   9.18   9.22 (b) (c)first half part