Data modeling: CSCI E-106

Applied Linear Statistical Models

Chapter 11 – Building the Regression Model III: Remedial Measures

Overview

remedial measures may need to be taken:

- a regression model is not appropriate
- several cases are very influential

previous: transformations to linearize the regression relation

- the error distributions more nearly normal
- make the variances of the error terms more nearly equal

Remedial Measures

In this chapter- remedial measure to deal with:

- unequal error variances
- a high degree of multicollinearity
- influential observations

two methods for nonparametric regression:

- lowess
- regression trees

bootstrapping: for evaluating the precision of the complex estimators

Weighted Least Squares (WLS)

- Chap. 3,6: transformation of Y - reducing or eliminating unequal variances
- difficulty: may create an inappropriate regression relationship
- weighted least squares
  - when an appropriate regression relationship has been found but the variances of the error terms are unequal

Weighted Least Squares (WLS), cont'd

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \qquad i = 1, \cdots, n$$

- parameters: $\beta_0, \beta_1, \cdots, \beta_{p-1}$
- known constants: $X_{i,1}, \cdots, X_{i,p-1}$
- $\varepsilon_i \sim \mathrm{N}(0, \sigma_i^2)$, independently normally distributed

$$\Rightarrow \underset{n \times n}{\boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\}} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

- $b = (X'X)^{-1}X'Y \Rightarrow$ unbiased, consistent but don't have the minimum variance

Weighted Least Squares (WLS), cont'd

- Consider: Observations with small variances provide more reliable information about the regression function than those with large variances.
- Error Variances Know ($w_i = \frac{1}{\sigma_i^2}$) $\Rightarrow$ **UNREALISTIC!**
- Error Variances Know up to Proportionality Constant ($w_i = k\frac{1}{\sigma_i^2}$)
- Error Variances Unknown: estimation of variance function or standard deviation function ($w_i = \frac{1}{(\hat{s}_i)^2}$, $w_i = \frac{1}{\hat{v}_i}$)

WLS – Error Variances Known

Methods of maximum likelihood to obtain estimators:

- $\varepsilon_i \overset{indep.}{\sim} N(0, \sigma_i^2)$ and $w_i = \frac{1}{\sigma_i^2}$

$$\Rightarrow L(\beta) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma_i^2)^{1/2}} \exp\left[-\frac{1}{2\sigma_i^2}\left(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1}\right)^2\right]$$

$$= \left[\prod_{i=1}^{n} \left(\frac{w_i}{2\pi}\right)^{1/2}\right] \exp\left[-\frac{1}{2}\sum_{i=1}^{n} w_i \left(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1}\right)^2\right]$$

$$\Rightarrow \boldsymbol{b} = \arg\max_{\boldsymbol{\beta}} L(\beta)$$

$$\Leftrightarrow \boldsymbol{b} = \arg\min_{\boldsymbol{\beta}} Q_w \quad \left(Q_w = \sum_{i=1}^{n} w_i \left(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1}\right)^2\right)$$

- Called *weighted least squares criterion* (min Q_w)
- $w_i = 1 \Rightarrow$ OLS

WLS – Error Variances Known, cont'd

- $w_i = \frac{1}{\sigma_i^2}$: reflects the amount of information contained in the observation Y_i
- $var\, Y_i$ large $\Rightarrow$ w_i less

$$\underset{n\times n}{\boldsymbol{W}} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}$$

- The normal equations: $(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})\boldsymbol{b}_w = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}$
- The weighted least squares (MLE) for β:

$$\boldsymbol{b}_w = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}$$

$$\Rightarrow \underset{p\times p}{\sigma^2\{\boldsymbol{b}_w\}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1} \quad (\because \boldsymbol{\sigma}^2\{\boldsymbol{Y}\} = \boldsymbol{W}^{-1})$$

WLS – Error Variances Known, cont'd

- $\boldsymbol{b}_w$: unbiased, consistent, have minimum variance among unbiased linear estimators
- When the weighted are known, $\boldsymbol{b}_w$ generally exhibits less variability than $\boldsymbol{b}$.

WLS - Error Variances Known Up To Proportionality Constant

$$w_i = k\frac{1}{\sigma_i^2}$$

$$\boldsymbol{b}_w = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y} \quad \text{(unaffected)}$$

$$\Rightarrow \underset{p\times p}{\sigma^2\{\boldsymbol{b}_w\}} = k(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}$$

$$k\text{: unknwon} \Rightarrow \underset{p\times p}{s^2\{\boldsymbol{b}_w\}} = MSE_w(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}$$

$$\text{where } MSE_w = \frac{\sum w_i(Y_i - \hat{Y}_i)^2}{n-p} = \frac{\sum w_i e_i^2}{n-p}$$

- MSE_w: an estimator of the proportionality constant k

WLS - Error Variances Unknown

- One rarely has knowledge of the variances $\sigma_i^2 \Rightarrow$ estimate of the variances

$$\sigma_i^2 = E\left(\varepsilon_i^2\right) - \left(E(\varepsilon_i)\right)^2$$

- Estimator of σ_i^2 : the squared residual e_i^2
- Estimator of $\sigma_i = \left|\sqrt{\sigma_i^2}\right|$: the absolute residuals $|e_i|$

WLS - Error Variances Unknown, cont'd

the estimation process:

1. Fit the regression model by OLS and analyze e_i
2. Estimate the variance function or the standard deviation function by regressing e_i^2 or $|e_i|$ on the appropriate predictor(s).
3. $\hat{s}_i$ for standard deviation function, $\hat{v}_i$ for the variance function: to obtain w_i
4. obtain b_w by using w_i

(Sometimes, iterated for several times to reach stabilize to stabilize the estimated regression coefficients ⇒ iteratively reweighted least squares)

WLS - Error Variances Unknown, cont'd

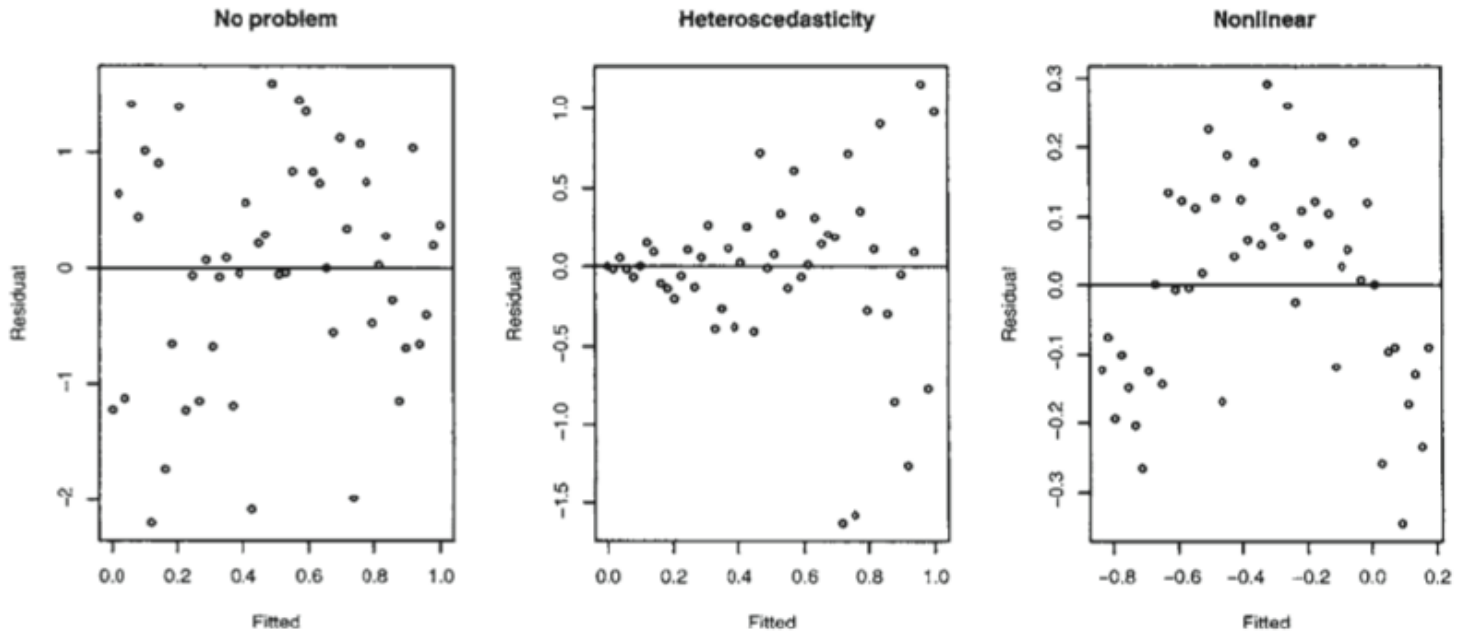


Figure 4.1 *Residuals vs. fitted plots—the first suggests no change to the current model while the second shows nonconstant variance and the third indicates some nonlinearity, which should prompt some change in the structural form of the model.*

- Reference: figure from Faraway's Linear Models with R (2005, p. 59)

WLS - Error Variances Unknown, cont'd

Some possible variance and standard deviation functions:

1. plot e_i vs. X_1 : a megaphone shape $\Rightarrow |e_i|$ regresses on X_1
2. plot e_i vs. $\hat{Y}$: a megaphone shape $\Rightarrow |e_i|$ regresses on $\hat{Y}$
3. plot e_i^2 *vs.* X_3 : upward tendency $\Rightarrow e_i^2$ regresses on X_3
4. plot e_i vs. X_2 : increase rapidly with X_2 up to a point and then increases more slowly $\Rightarrow |e_i|$ regresses on X_2 , X_2^2

$\Rightarrow w_i = \frac{1}{(\hat{s}_i)^2}$: $\hat{s}_i$ is fitted value from standard deviation function

$\Rightarrow w_i = \frac{1}{\hat{v}_i}$: $\hat{v}_i$ is fitted value from variance function

$\Rightarrow$ the weight matrix W $\Rightarrow b_w = (X'WX)^{-1} X'WY$

WLS - Error Variances Unknown, cont'd

Using of Replicated or Near Replicates

- Replicate observations are made at each combination of levels of *X*s
- If the number of replications (or near replications) is large, w_i may be obtained directly from the sample variances of *Y* observations at each combination of levels on *X* variables
- each case in a replicate group receives the same weight with this method
- Confidence interval: (similar but approximate)

$$b_k \pm t(1 - \alpha/2; n - p)s\{b_k\}$$

WLS - Error Variances Unknown, cont'd

Using of Ordinary Least Squares with Unequal Error Variances:

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

$$\Rightarrow \boldsymbol{\sigma}^2\{\boldsymbol{b}\} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{\sigma}^2\{\varepsilon\}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$\Rightarrow \boldsymbol{S}^2\{\boldsymbol{b}\} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{S}_0\{\varepsilon\}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

where $\boldsymbol{S}_0 = \text{diag}\{e_1^2, e_2^2, \ldots, e_n^2\}$

Blood Pressure Example

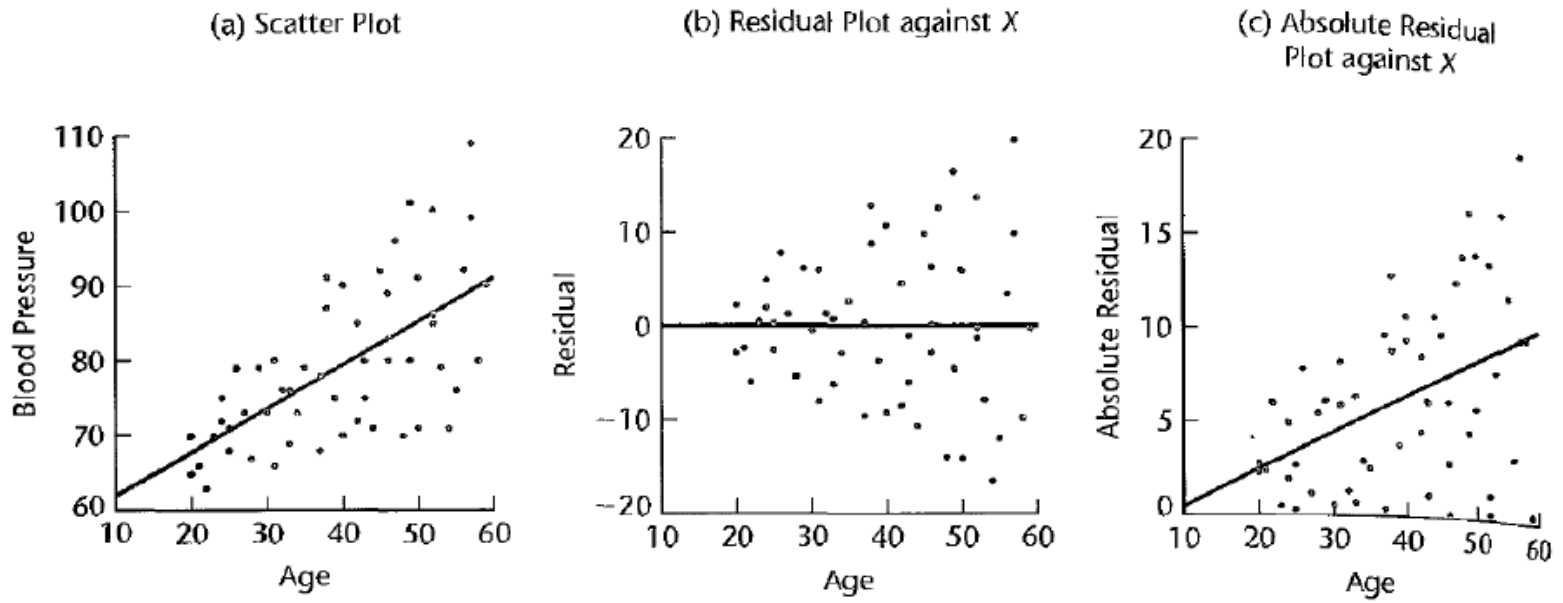
- 54 subjects: 20-60 women

TABLE 11.1 Weighted Least Squares—Blood Pressure Example.

Subject i	(1) Age X_i	(2) Diastolic Blood Pressure Y_i	(3) e_i	(4) $\lvert e_i \rvert$	(5) $\hat{s}_i$	(6) w_i
1	27	73	1.18	1.18	3.801	.06921
2	21	66	−2.34	2.34	2.612	.14656
3	22	63	−5.92	5.92	2.810	.12662
...	...	...	...	...	...	...
52	52	100	13.68	13.68	8.756	.01304
53	58	80	−9.80	9.80	9.944	.01011
54	57	109	19.78	19.78	9.746	.01053

Blood Pressure Example, cont'd

FIGURE 11.1 Diagnostic Plots Detecting Unequal Error Variances—Blood Pressure Example.



- linear regression function by unweighted least squares:

$$\hat{Y} = \underset{(3.994)}{56.157} + \underset{(0.09695)}{0.58003}X$$

Blood Pressure Example, cont'd

- $|e_i|$ regresses on X: $\hat{s} = -1.54946 + 0.198172X$

$$\hat{s}_1 = 3.801 \Rightarrow w_1 = 1/(\hat{s}_1)^2 = 0.0692$$

- WLS: $\hat{Y} = 55.566 + 0.59634X$
- C.I. for β_1: $\alpha = 0.05$; $s\{b_{w1}\} = 0.07924$

$$0.59643 \pm \underbrace{t(0.975; 52)}_{=2.007}(0.07924) \Rightarrow 0.437 \leq \beta_1 \leq 0.755$$

Blood Pressure Example, cont'd

1. unequal variances: heteroscedasticity and equal variances: homoscedasticity
2. R^2 does not have a clear-cut meaning for WLS
3. WLS may be view as OLS of transformed variables:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon, \quad \boldsymbol{E}\{\varepsilon\} = \boldsymbol{0}, \boldsymbol{\sigma}^2\{\varepsilon\} = \boldsymbol{W}^{-1}$$

$$\Rightarrow \underbrace{\boldsymbol{W}^{1/2}\boldsymbol{Y}}_{\boldsymbol{Y}_w} = \underbrace{\boldsymbol{W}^{1/2}\boldsymbol{X}\beta}_{\boldsymbol{X}_w\beta} + \underbrace{\boldsymbol{W}^{1/2}\varepsilon}_{\varepsilon_w}$$

$$\Rightarrow \boldsymbol{E}\{\varepsilon_w\} = \boldsymbol{0}; \quad \boldsymbol{\sigma}^2\{\varepsilon_w\} = \boldsymbol{I};$$

$$\boldsymbol{b}_w = (\boldsymbol{X}'_w\boldsymbol{X}_w)^{-1}\boldsymbol{X}'_w\boldsymbol{Y}_w = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}$$

Blood Pressure Example, cont'd

The weighted least squares normal equations:

$$\sum w_i Y_i = b_{w0} \sum w_i + b_{w1} \sum w_i X_i$$
$$\sum w_i X_i Y_i = b_{w0} \sum w_i X_i + b_{w1} \sum w_i X_i^2$$

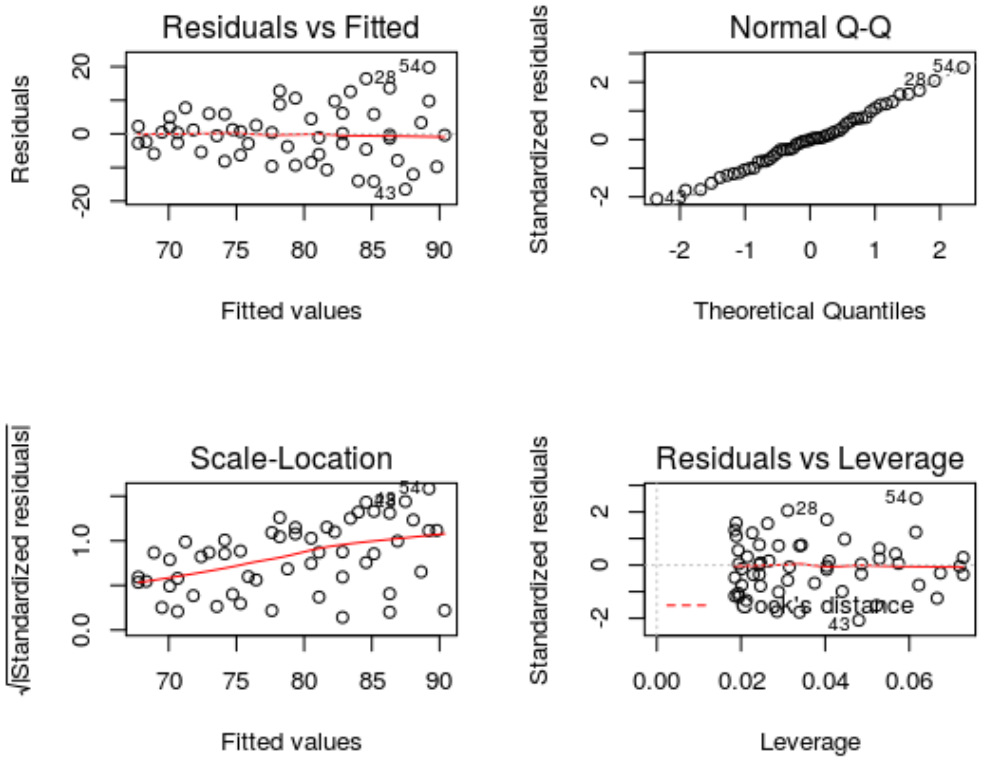
The weighted least squares estimators b_{w0} and b_{w1} in (11.9) are:

$$b_{w1} = \frac{\sum w_i X_i Y_i - \frac{\sum w_i X_i \sum w_i Y_i}{\sum w_i}}{\sum w_i X_i^2 - \frac{(\sum w_i X_i)^2}{\sum w_i}}$$
$$b_{w0} = \frac{\sum w_i Y_i - b_{w1} \sum w_i X_i}{\sum w_i}$$

Blood Pressure Example – R Codes

```
g<-lm(Y~X, data=Blood.Pressure)
par(mfrow=c(2,2))
plot(g)
ei<-g$residuals
abs.ei<-abs(ei)
g1<-lm(abs.ei~Blood.Pressure$X)
summary(g1)
```

```
Call:
lm(formula = abs.ei ~ Blood.Pressure$X)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7639 -2.7882 -0.1587  3.0757 10.0350

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.54948    2.18692  -0.709  0.48179
Blood.Pressure$X  0.19817    0.05309   3.733  0.00047 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.461 on 52 degrees of freedom
Multiple R-squared:  0.2113,  Adjusted R-squared:  0.1962
F-statistic: 13.93 on 1 and 52 DF,  p-value: 0.0004705
```



Blood Pressure Example – R Codes, cont'd

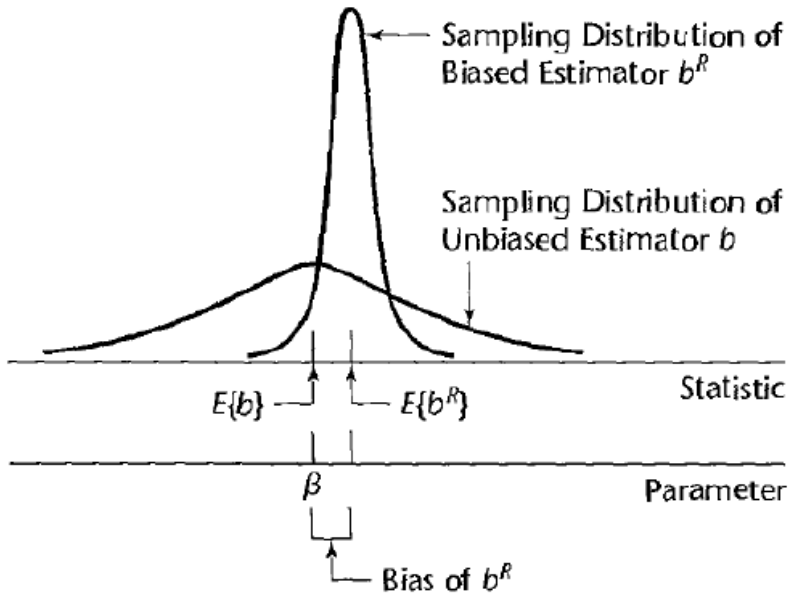
```
 > s<-g1$fitted.values
> s[1:3]
      1       2       3
3.801175 2.612141 2.810313
> wi=1/(s^2)
> g2<-lm(Y~X,weights = wi, data=Blood.Pressure)
> summary(g2)
Call:
lm(formula = Y ~ X, data = Blood.Pressure, weights = wi)
Weighted Residuals:
    Min      1Q  Median      3Q     Max
-2.0230 -0.9939 -0.0327  0.9250  2.2008
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.56577    2.52092  22.042  < 2e-16 ***
X            0.59634    0.07924   7.526 7.19e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.213 on 52 degrees of freedom
Multiple R-squared:  0.5214,  Adjusted R-squared:  0.5122
F-statistic: 56.64 on 1 and 52 DF,  p-value: 7.187e-10
```

Ridge Regression

- Ridge regression: a method of overcoming serious multicollinearity problem by modifying the method of least squares.
- allowing biased estimators of the regression functions
- When an estimator has only a small bias and is substantially more precise than an unbiased estimator, it may well be the preferred estimator since it will have a larger probability of being close to the true parameter value.

Ridge Regression, cont'd

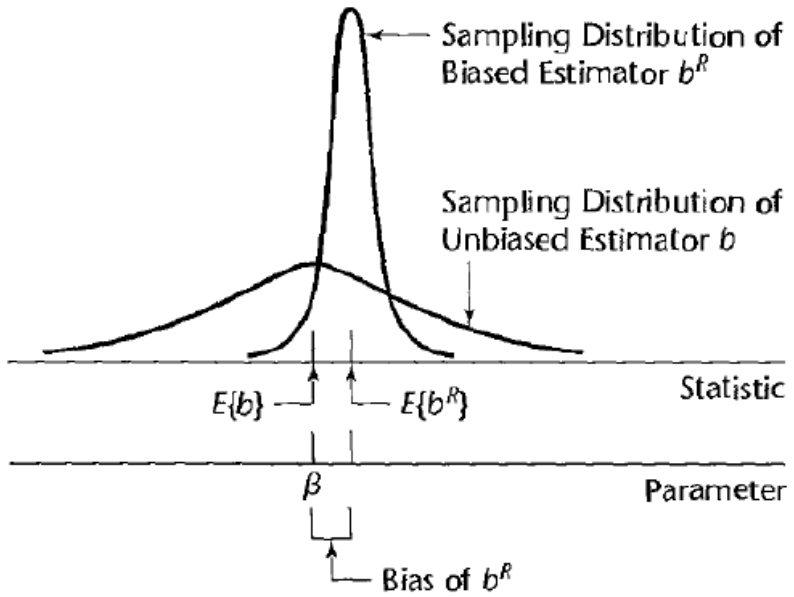
FIGURE 11.2
Biased Estimator with Small Variance May Be Preferable to Unbiased Estimator with Large Variance.



- Estimator b is unbiased but imprecise
- Estimator b^R is much more precise but has a small bias
  $\Rightarrow$ The probability that b^R falls near β is much greater than that for b

Ridge Regression, cont'd

FIGURE 11.2
Biased Estimator with Small Variance May Be Preferable to Unbiased Estimator with Large Variance.



- Estimator b is unbiased but imprecise
- Estimator b^R is much more precise but has a small bias
  $\Rightarrow$ The probability that b^R falls near β is much greater than that for b

Ridge Regression, cont'd

A measure of the combined effect of bias and sampling variation:

- The mean squared error (MSE):

$$E\{b^R - \beta\}^2 = \sigma^2\{b^R\} + (E\{b^R\} - \beta)^2$$

- Transformed variables:

$$Y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i - \bar{Y}}{s_Y}\right), \quad X_{ik}^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_{ik} - \bar{X}_k}{s_k}\right)$$

$$Y_i^* = \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \cdots \beta_{p-1}^* X_{i,p-1}^* + \varepsilon_i^*$$

$$\Rightarrow \boldsymbol{r}_{XX}\boldsymbol{b} = \boldsymbol{r}_{YX}$$

Ridge Regression, cont'd

- The ridge standardized regression estimators are obtained by:

$$(\boldsymbol{r}_{XX} + c\boldsymbol{I})\boldsymbol{b}^R = \boldsymbol{r}_{YX}$$
$$\Rightarrow \boldsymbol{b}^R = (\boldsymbol{r}_{XX} + c\boldsymbol{I})^{-1}\boldsymbol{r}_{YX}$$

- Biasing constant $c \geq 0$: reflects the amount of bias
- $c = 0 \Rightarrow$ OLS; $c > 0 \Rightarrow$ bias but more stable (less variable) than OLS estimators

Ridge Regression, cont'd

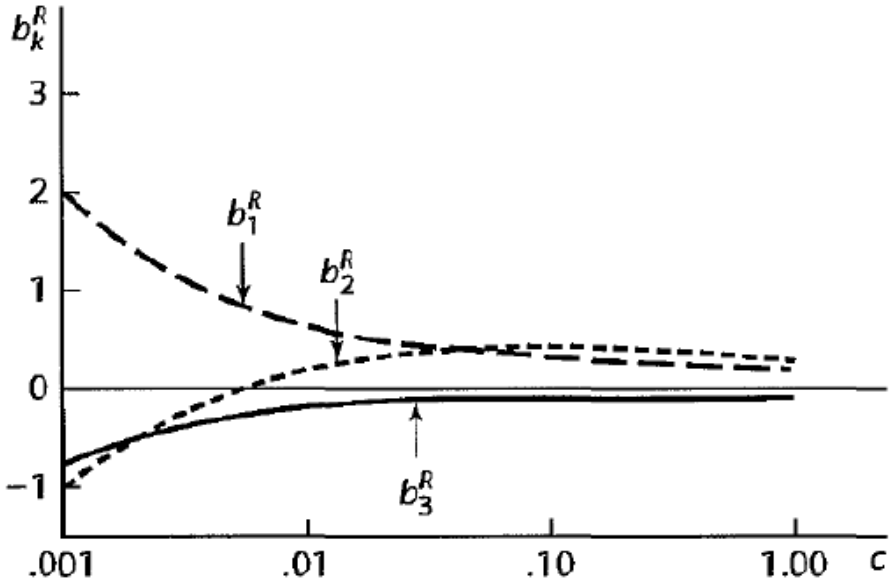
Choice of Biasing Constant *c*:

- Squared bias: $E\{b^R - \beta\}^2 = \sigma^2\{b^R\} + (E\{b^R\} - \beta)^2$
  - bias component $(E\{b^R\} - \beta)$ of $E\{b^R - \beta\}^2 \uparrow$ as $c \uparrow$ while the variance $\sigma^2\{b^R\}$ component becomes smaller
  - Always ∃ a *c* s.t. b^R has a smaller total MSE that *b*
- difficulty:
  - the optimum value of *c* varies from one application to another
  - and is unknown
- Using Ridge trace and $(VIF)_k$

Ridge Regression, cont'd

- Ridge trace:a simultaneous plot of the values of the $(p-1)$ b^R for different c; $0 \leq c \leq 1$

FIGURE 11.3 Ridge Trace of Estimated Standardized Regression Coefficients—Body Fat Example with Three Predictor Variables.



- Choose c: the Ridge trace starts to become stable and VIF has become sufficiently small.

Ridge Regression, cont'd

- The normal equation for the ridge estimators:

$$
\begin{aligned}
(1+c)b_1^R + r_{12}b_2^R + \cdots + r_{1,p-1}b_{p-1}^R &= r_{Y1} \\
r_{21}b_1^R + (1+c)b_2^R + \cdots + r_{2,p-1}b_{p-1}^R &= r_{Y2} \\
&\vdots \\
r_{p-1,1}b_1^R + r_{p-1,2}b_2^R + \cdots + (1+c)b_{p-1}^R &= r_{Y,p-1}
\end{aligned}
$$

- *VIF* for b_k^R are defined analogously to those for OLS b_k: measures how large is the variance of b_k^R relative to what the variance would be if Xs were uncorrelated.

Ridge Regression, cont'd

- *VIF* for b_k^R: the diagonal elements of the following matrix

$$(\boldsymbol{r}_{XX} + c\boldsymbol{I})^{-1}\boldsymbol{r}_{XX}(\boldsymbol{r}_{XX} + c\boldsymbol{I})^{-1}$$

- Ridge regression estimates can be obtained by the method of *penalized least squares*.

$$Q = \sum_{i=1}^{n}\left[Y_i^* - (\beta_1^* X_{i1}^* + \cdots + \beta_{p-1}^* X_{i,p-1}^*)\right]^2 + c\left[\sum_{j=1}^{p-1}(\beta_j^*)^2\right]$$

  sometimes referred to as shrinkage esimators

- Limitation: ridge regression is that ordinary inference procedures are not applicable and exact distribution properties are not known. (bootstrapping)

Ridge Regression, cont'd

TABLE 11.2 Ridge Estimated Standardized Regression Coefficients for Different Biasing Constants c—Body Fat Example with Three Predictor Variables.

c	b_1^R	b_2^R	b_3^R
.000	4.264	−2.929	−1.561
.002	1.441	−.4113	−.4813
.004	1.006	−.0248	−.3149
.006	.8300	.1314	−.2472
.008	.7343	.2158	−.2103
.010	.6742	.2684	−.1870
.020	.5463	.3774	−.1369
.030	.5004	.4134	−.1181
.040	.4760	.4302	−.1076
.050	.4605	.4392	−.1005
.100	.4234	.4490	−.0812
.500	.3377	.3791	−.0295
1.000	.2798	.3101	−.0059

(a) Table 11.2

TABLE 11.3 *VIF* Values for Regression Coefficients and R^2 for Different Biasing Constants c—Body Fat Example with Three Predictor Variables.

c	$(VIF)_1$	$(VIF)_2$	$(VIF)_3$	R^2
.000	708.84	564.34	104.61	.8014
.002	50.56	40.45	8.28	.7901
.004	16.98	13.73	3.36	.7864
.006	8.50	6.98	2.19	.7847
.008	5.15	4.30	1.62	.7838
.010	3.49	2.98	1.38	.7832
.020	1.10	1.08	1.01	.7818
.030	.63	.70	.92	.7812
.040	.45	.56	.88	.7808
.050	.37	.49	.85	.7804
.100	.25	.37	.76	.7784
.500	.15	.21	.40	.7427
1.000	.11	.14	.23	.6818

(b) Table 11.23

Ridge Regression, cont'd

$$c = 0.02 \Rightarrow \hat{Y}^* = 0.5463X_1^* + 0.3774X_2^* - 0.1369X_3^*$$

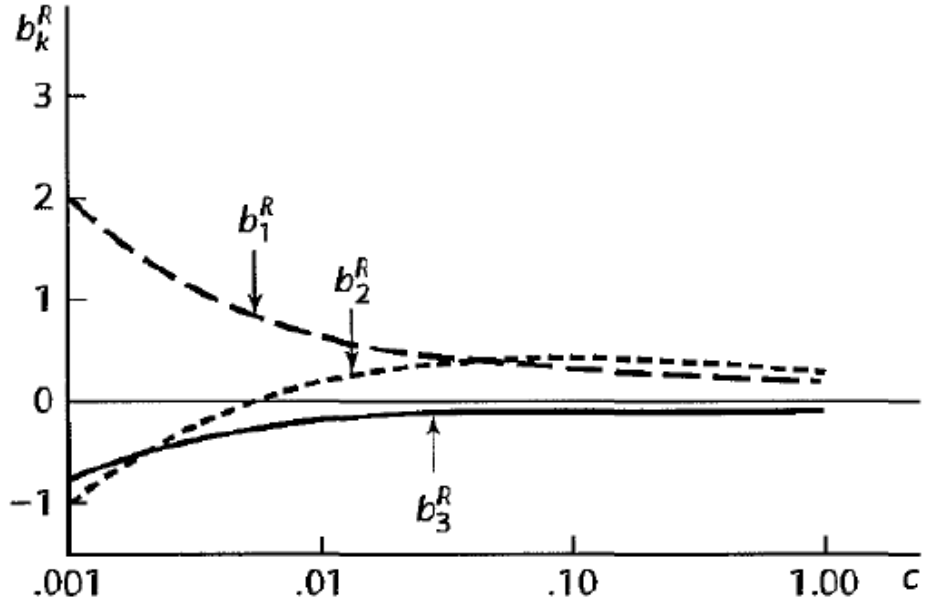


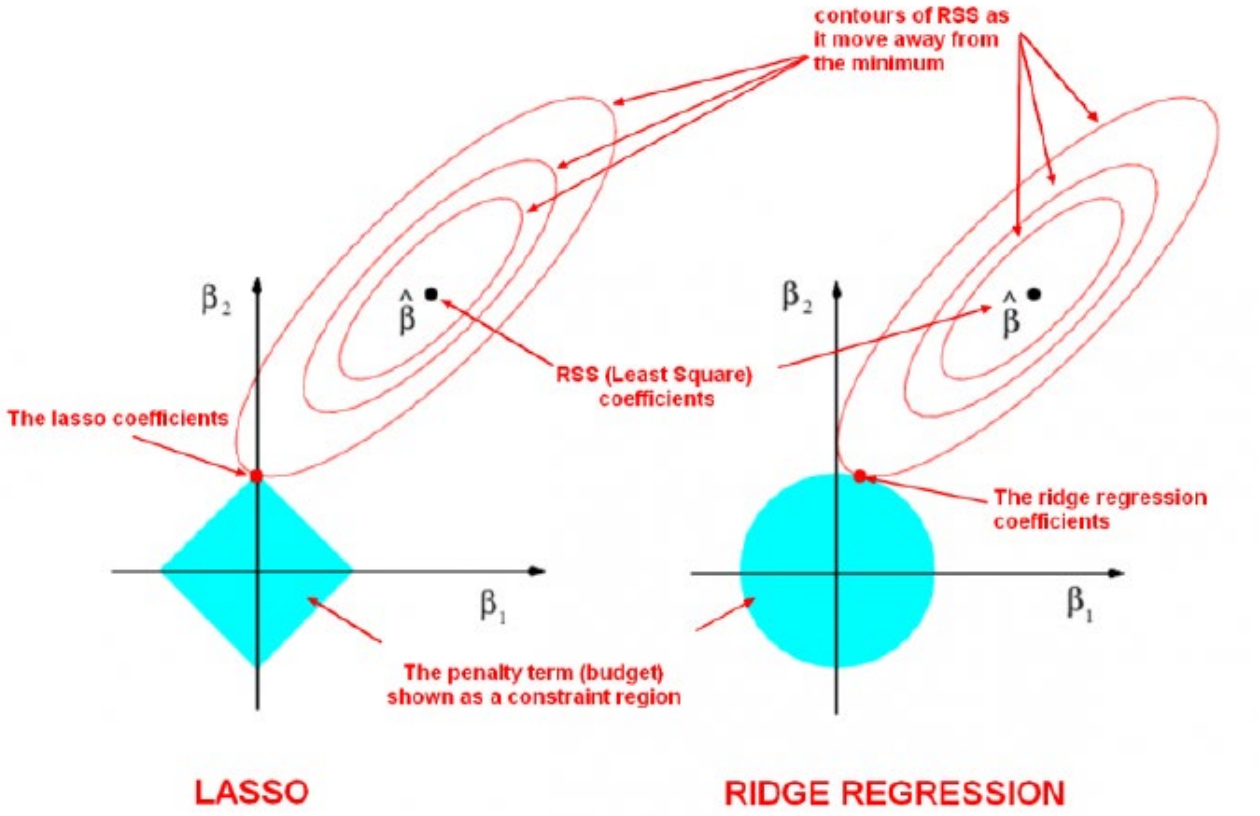
FIGURE 11.3 Ridge Trace of Estimated Standardized Regression Coefficients—Body Fat Example with Three Predictor Variables.

Ridge Regression, cont'd

LASSO: Least absolute shrinkage and selection operator

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p-1} X_{ij}\beta_j)^2 + \lambda \sum_{j=0}^{p-1} |\beta_j|$$

Ridge Regression, cont'd



Robust Regression

- To examine whether an outlying case is the result of a recording error, breakdown of a measurement instrument, or the like.
  - Erroneous data can be corrected ⇒ corrected
  - Erroneous data cannot be corrected ⇒ discarded

- Many time: impossible to tell for certain whether the observations for an outlying case are erroneous ⇒ not be discarded
- If an outlying influential case is not clearly erroneous ⇒ to examine the adequacy of the model

Robust Regression, cont'd

- Numerous robust regression procedures have been developed.
- LAR or LAD regression: Least absolute residuals or least absolute deviations

minimum $L1$ norm regression

Criterion:

$$\min_{\beta} L_1 = \min_{\beta} \left\{ \sum_{i=1}^{n} |Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} x_{i,p-1})| \right\}$$

- LAR: less emphasis on outlying observations than does the method of least squares
- Linear programming method: to obtain the LAR estimators
- LAR method: the residuals will not sum to zero

Robust Regression, cont'd

- LMS regression: Least median of squares

$$\text{median}\left\{[Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} x_{i,p-1})]^2\right\}$$

- IRLS robust regression: Iteratively reweighted least squares method

Iteratively reweighted least squares Regression (IRLS)

IRLS ⇒ Weighted Regression

- Outlying cases that have large residuals ⇒ given smaller weights.
- The weights are revised as each iteration yields new residuals until the estimation process stabilizes.

A summary of the steps follows:

1. Choose a weight function for weighting the cases.
2. Obtain starting weights for all cases.
3. Use ,the starting weights in weighted least squares and obtain the residuals from the fitted regression function.
4. Use the residuals in step 3 to obtain revised weights.
5. Continue the iteration until convergence is obtained.

IRLS, cont'd

Weighted Functions:

$$\text{Huber: } w = \begin{cases} 1 & |u| \le 1.345 \\ \dfrac{1.345}{|u|} & |u| > 1.345 \end{cases} \qquad \textbf{(11.44)}$$

$$\text{Bisquare: } w = \begin{cases} \left[1 - \left(\dfrac{u}{4.685}\right)^2\right]^2 & |u| \le 4.685 \\ 0 & |u| > 4.685 \end{cases} \qquad \textbf{(11.45)}$$

$$u_i = \frac{e_i}{MAD}$$

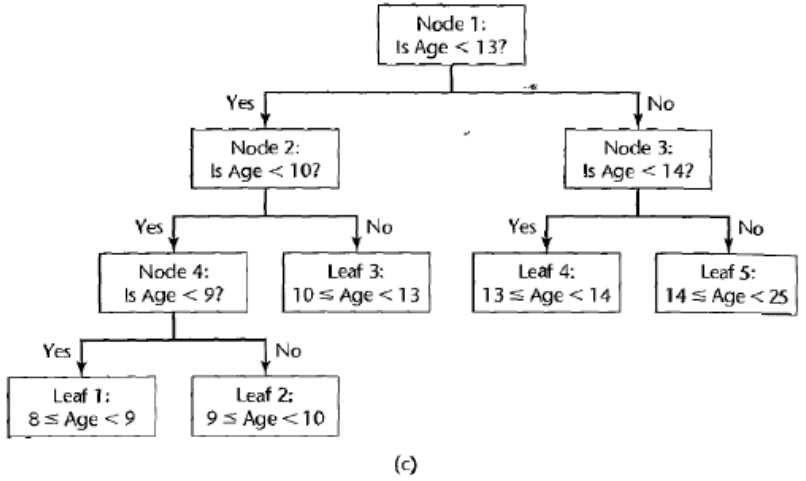
$$MAD = \frac{1}{.6745}\,\text{median}\{|e_i - \text{median}\{e_i\}|\}$$

Regression Tree

- Powerful, conceptually simple
- method of nonparametric regression
  1. the range of the predictor is partitioned into segments
  2. within each segment the estimated regression fit is given by the mean of the response in the segment

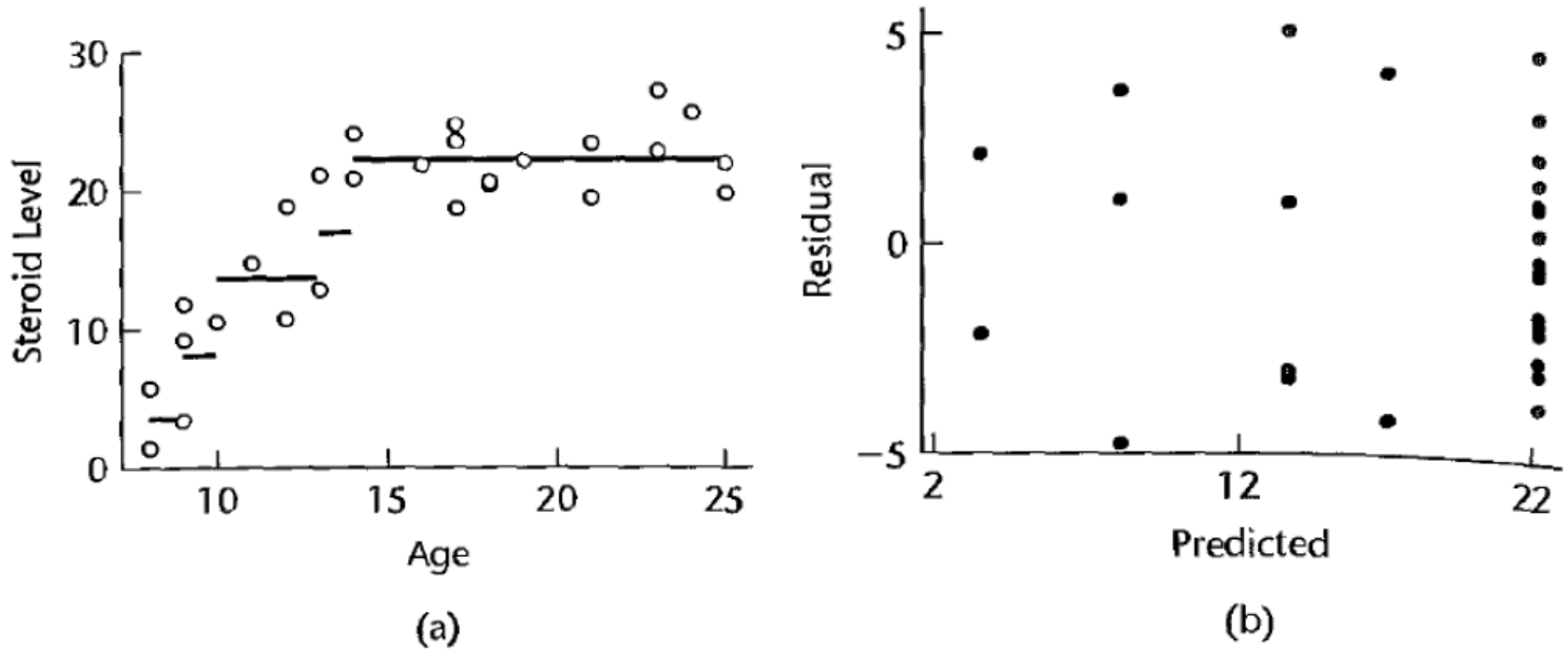
TABLE 11.8 Data Set and 5-Region Regression Tree Fit—Steroid Level Example.

(1) Case i	(2) Steroid Level Y_i	(3) Age X_i	(4) Region Number k	(5) Region R_{Sk}	(6) Fitted Value $\bar{Y}_{R_{Sk}}$
1	27.1	23	1	$8 \le X < 9$	3.550
2	22.1	19	2	$9 \le X < 10$	8.133
3	21.9	25	3	$10 \le X < 13$	13.675
...	...	...	4	$13 \le X < 14$	16.950
25	12.8	13	5	$14 \le X < 25$	22.200
26	20.8	14			
27	20.6	18			

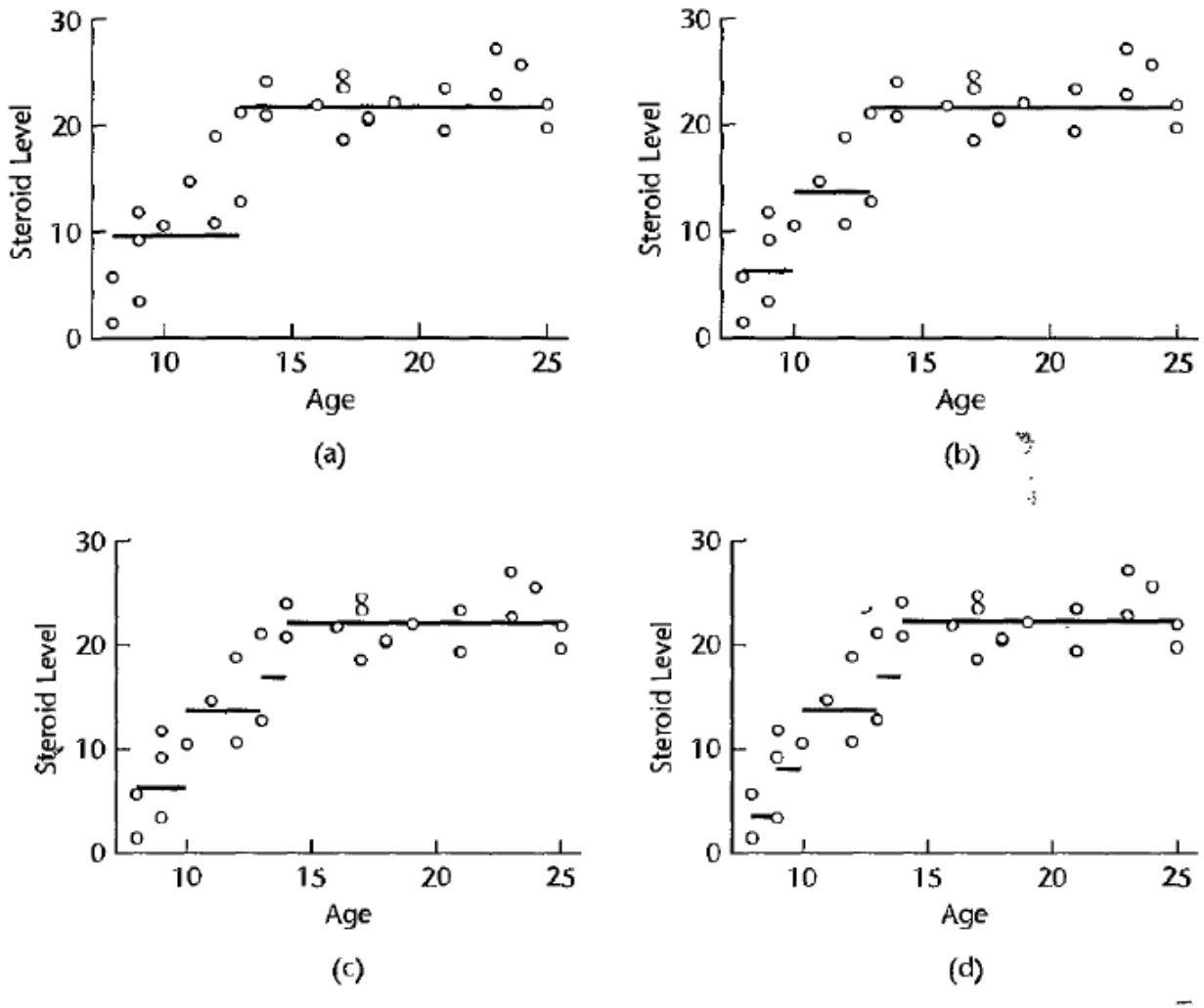


(c)

Regression Tree, cont'd



Regression Tree, cont'd



Regression Tree, cont'd

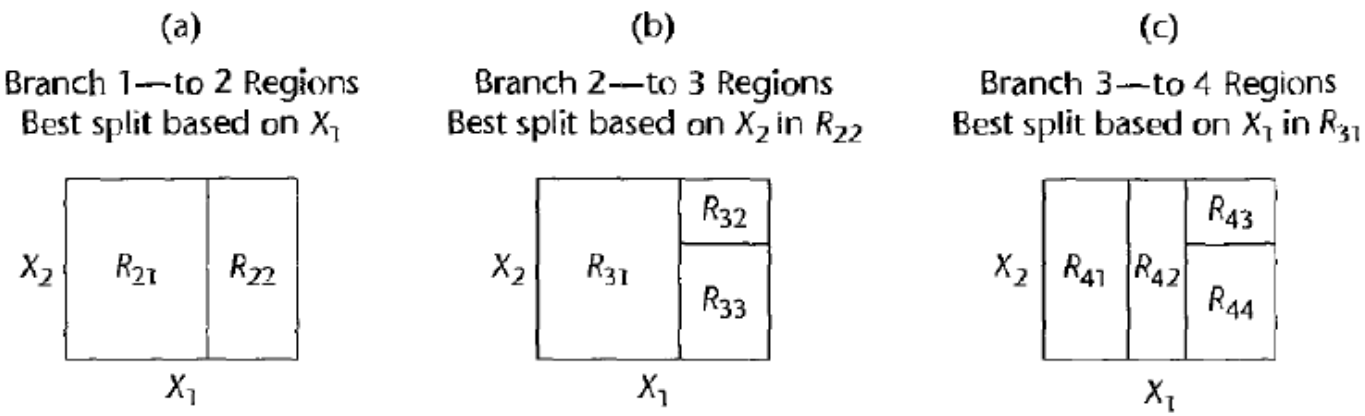
- $r = 2$: the best point is choose to

$$\min SSE = SSE(R_{21}) + SSE(R_{22})$$

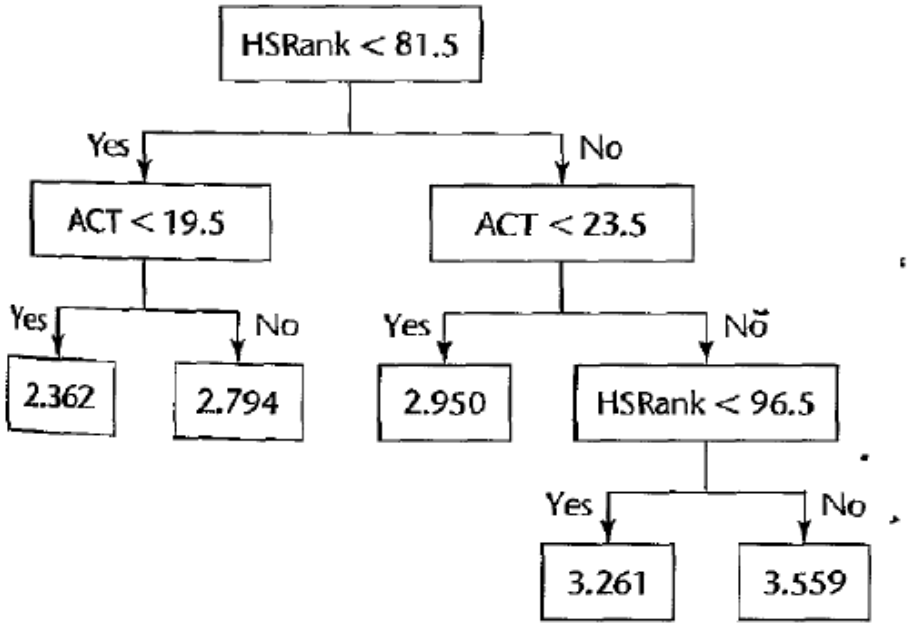
$$SSE(R_{rj}) = \sum (Y_i - \bar{Y}_{R_{jk}})^2$$

- $SSE(R_{rj})$: the sum of squared residuals in region R_{rj}

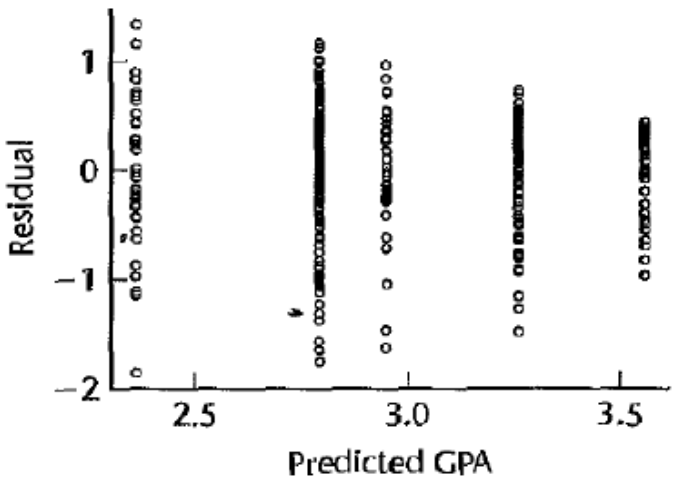
FIGURE 11.11 Regression Tree Growth—Two-Predictor Example.



Regression Tree, cont'd



(c)

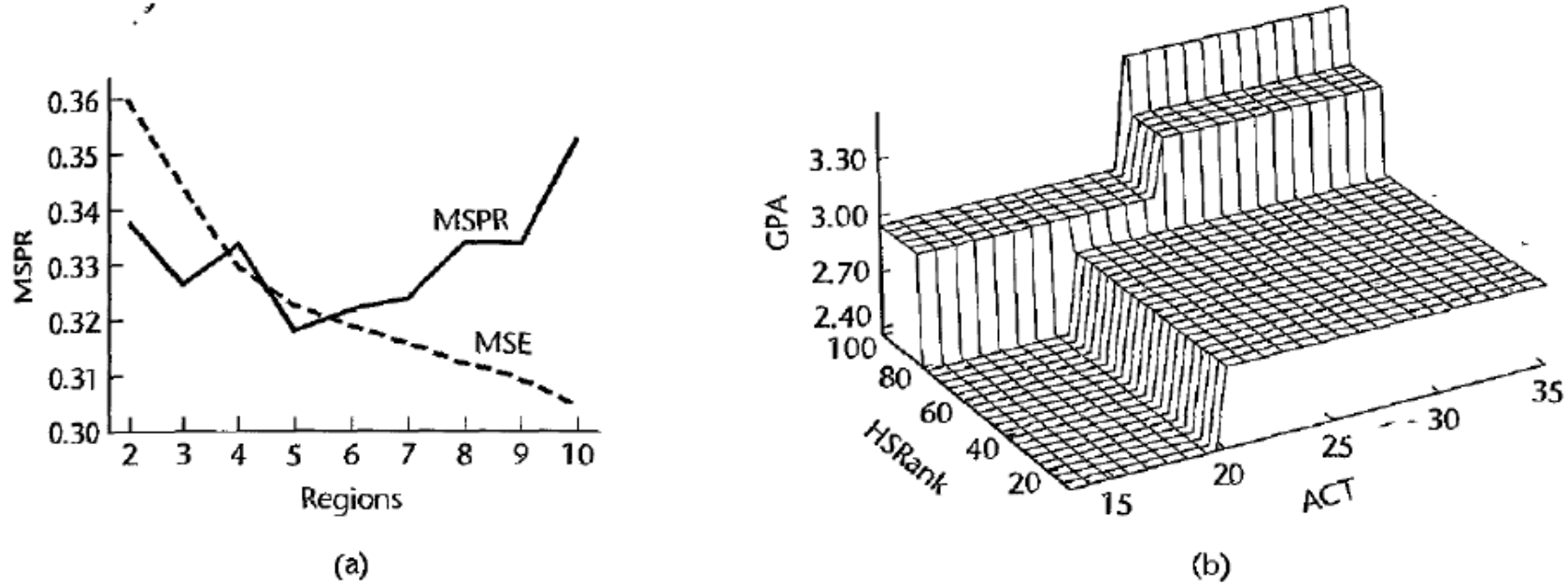


(d)

Regression Tree, cont'd

Determining the number of Regions R

FIGURE 11.12 S-Plus Regression Tree Results—University Admissions Example.



Bagging on Decision Trees

By averaging a collection of bootstrap samples from the training data set, we can dramatically reduce the model variance of trees, leading to improved prediction. This is called bagging.

- Bagging = Bootstrap Aggregating
- For a given B, generate B bootstrapped training data sets.
- For each test observation, we record the class predicted by each of B decision trees, and choose the majority vote.
- The overall prediction is the most commonly occurring class(majority vote) among the B predictions.

Random Forest

- The idea for bagging of decision trees is to average many noisy trees and hence to reduce the model variance.
- However, since each tree generated in bagging is based on identical features, the expectation value of the averages is the same as the expectation of any one of them. This means the bias will not be improved.
- Random forests is a modification of bagging that builds a large collection of de-correleated trees, and then averages them.

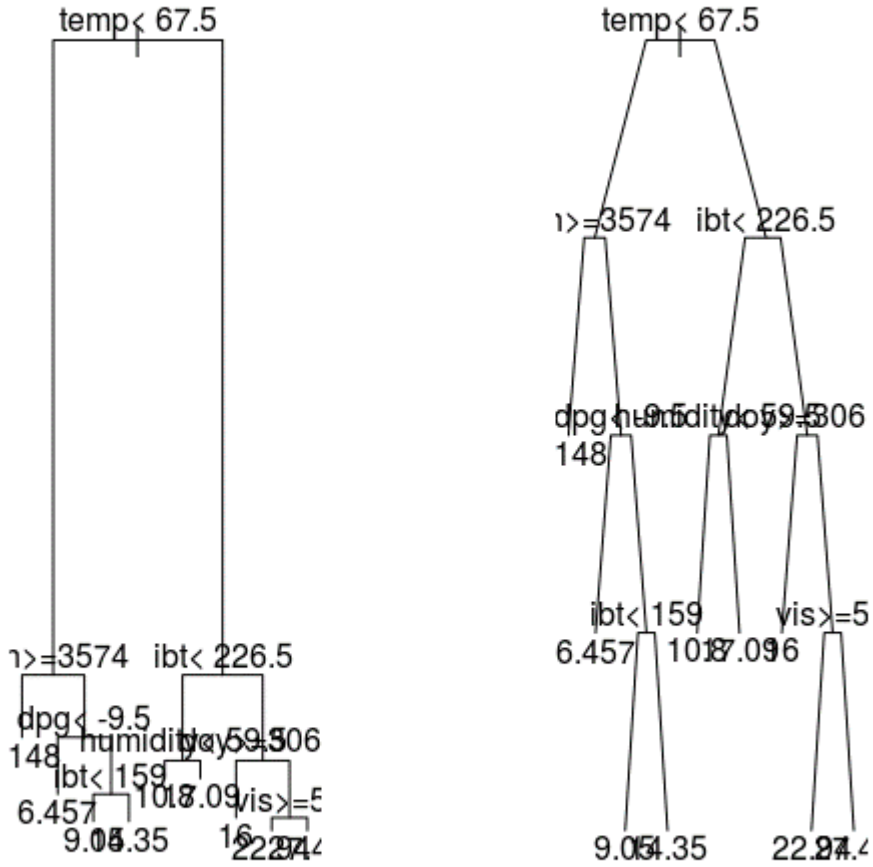
Example

```
data(ozone, package = faraway)
library(rpart)
tmod<-rpart(O3~.,ozone)
tmod
n= 330

node), split, n, deviance, yval
      * denotes terminal node
 1) root 330 21115.4100 11.775760
   2) temp< 67.5 214  4114.3040  7.425234
     4) ibh>=3573.5 108   689.6296  5.148148 *
     5) ibh< 3573.5 106  2294.1230  9.745283
      10) dpg< -9.5 35   362.6857  6.457143 *
      11) dpg>=-9.5 71  1366.4790 11.366200
        22) ibt< 159 40   287.9000  9.050000 *
        23) ibt>=159 31   587.0968 14.354840 *
   3) temp>=67.5 116  5478.4400 19.801720
     6) ibt< 226.5 55  1276.8360 15.945450
      12) humidity< 59.5 10   167.6000 10.800000 *
      13) humidity>=59.5 45   785.6444 17.088890 *
     7) ibt>=226.5 61  2646.2620 23.278690
      14) doy>=306.5 8   398.0000 16.000000 *
      15) doy< 306.5 53  1760.4530 24.377360
        30) vis>=55 36  1149.8890 22.944440 *
        31) vis< 55 17   380.1176 27.411760 *
```

We see that the first split (nodes 2 and 3) is on temperature: 214 observations have temperatures less than 67.5 with a mean response value of 7.4, whereas 116 observations have temperatures greater than 67.5 with a mean response value of 20. The total RSS has been reduced from 21,115 to 4114 + 5478 = 9592.

Example, cont’d

```
> text(tmod)
> par(mfrow=c(1,2))
> plot(tmod)
> text(tmod)
> plot(tmod,compress=T,uniform=T,branch=0.4)
>text(tmod)

> library(randomForest)
> fmod<-randomForest(O3~.,ozone)
> plot(fmod,main="")
```



Bootstrapping

- In many nonstandard situations, (ex: nonconstant error variances estimated by iteratively reweighted least squares), standard methods for evaluating the precision may not be available or may only be approximately applicable when the sample size is large
- Bootstrapping: provide estimates of the precision of sample estimated for these complex cases

Bootstrapping, cont'd

general procedure: obtain b_1 by some procedure and wish to evaluate the precision of b_1 by bootstrap method

- The bootstrap method calls for the selection from the observed sample data of a random sample of size *n* with replacement.
- the bootstrap sample may contain some duplicate data from the original sample and omit some other data in the original sample
- calculates the estimated regression coefficients from the bootstrap sample $\Rightarrow b_1^*$
- repeated a large number of times
- The estimated standard deviation of all of $b_1^* \Rightarrow s\{b_1^*\}$ an estimate of the variability of the sampling distribution of b_1

Bootstrapping, cont'd

Bootstrap Sampling: two basic ways

- When the regression function being fitted is a good model for the data, the error terms have constant variance, and fixed *X* sampling is appropriate.
  - e_i from the original fitting are regarded as the sample data to be sampled with replacement
  - After a bootstrap sample with *n*:

$$
\begin{aligned}
& e_1^*, \ldots, e_n^* \\
\Rightarrow & Y_i^* = \hat{Y}_i + e_i^* \\
& Y^* \text{ regression on } X\text{s} \Rightarrow b_1^*
\end{aligned}
$$

Bootstrapping, cont'd

- When there is some doubt about the adequacy of the regression function being fitted, the error variance are not constant, and/or *random X sampling* is appropriate.
  - (X_i ,Y_i) in the original sample are considered to be the data to be sampled with replacement
  - After a bootstrap sample with *n*:

$$n \text{ pairs: } (X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$$

$$Y^* \text{ regression on } X^*\text{s} \Rightarrow b_1^*$$

Bootstrapping, cont'd

- 200-500 bootstrap samples are adequate
- One can observe the variability of the bootstrap estimates by $s^*\{b_1^*\}$ as the number of bootstrap samples is increased.

Bootstrap Confidence Intervals

- From the bootstrap distribution of b_1^*, find the $\alpha/2$ and $1-\alpha/2$ quantiles $b_1^*(\alpha/2)$ and $b_1^*(1-\alpha/2)$
- Percentiles from b_1:

$$b_1^*(\alpha/2) \leq \beta_1 \leq b_1^*(1-\alpha/2)$$

- reflection method: require a larger number of bootstrap samples

$$2b_1 - b_1^*(\alpha/2) \leq \beta_1 \leq 2b_1 - b_1^*(\alpha/2)$$