# Chapter 11
# Model Building III – Remedial Measures

Instructor: Li C.X.

# Outline

- Unequal Error Variance---Weighted Least Squares
- Multicollinearity---Ridge Regression
- Influential Cases---Robust Regression
- Nonparametric Regression
  - --Lowess Method and Regression Trees
- Evaluating Precision---Bootstrapping

# Unequal Error Variances – Weighted Least Squares (WLS)

- Case 1 – Error Variances known exactly (VERY rare)
- Case 2 – Error Variances unknown
  - Estimating unknown variance

# WLS –Known Variances

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad \varepsilon_i \sim N\left(0, \sigma_i^2\right) \quad i = 1,...,n \qquad \sigma\left\{\varepsilon_i, \varepsilon_j\right\} = 0 \quad \forall\, i \neq j$$

$$\Rightarrow \boldsymbol{\sigma}^2\left\{\boldsymbol{\varepsilon}\right\} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Maximum Likelihood Estimation:

$$L\left(\boldsymbol{\beta}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2\sigma_i^2}\left(Y_i - \beta_0 - \beta_1 X_{i1} - ... - \beta_{p-1} X_{i,p-1}\right)^2\right] \text{ setting: } w_i = \frac{1}{\sigma_i^2}$$

$$\Rightarrow \quad L\left(\boldsymbol{\beta}\right) = \left[\prod_{i=1}^{n} \sqrt{\frac{w_i}{2\pi}}\right] \exp\left[-\frac{1}{2}\sum_{i=1}^{n} w_i \left(Y_i - \beta_0 - \beta_1 X_{i1} - ... - \beta_{p-1} X_{i,p-1}\right)^2\right]$$

To maximize $L\left(\boldsymbol{\beta}\right)$ , we need to minimize $Q_w = \sum_{i=1}^{n} w_i \left(Y_i - \beta_0 - \beta_1 X_{i1} - ... - \beta_{p-1} X_{i,p-1}\right)^2$

Note that values with smaller $\sigma_i^2$ have larger weights $w_i$ in the weighted least squares criterion.

# WLS –Known Variances

Easiest to set up in matrix form, where: $\mathbf{W} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix} = \mathbf{W}'$

$$\boldsymbol{\sigma}^2\left\{\mathbf{Y}\right\} = \boldsymbol{\sigma}^2\left\{\boldsymbol{\varepsilon}\right\} = \mathbf{W}^{-1}$$

Normal Equations: $\left(\mathbf{X'WX}\right)\mathbf{b}_w = \mathbf{X'WY}$

$$\Rightarrow \quad \mathbf{b}_w = \left(\mathbf{X'WX}\right)^{-1}\mathbf{X'WY} = \mathbf{AY} \qquad \mathbf{A} = \left(\mathbf{X'WX}\right)^{-1}\mathbf{X'W}$$

$$\Rightarrow \quad \mathbf{E}\left\{\mathbf{b_w}\right\} = \mathbf{AE}\left\{\mathbf{Y}\right\} = \mathbf{AX}\boldsymbol{\beta} = \left(\mathbf{X'WX}\right)^{-1}\mathbf{X'WX}\boldsymbol{\beta} = \boldsymbol{\beta}$$

$$\Rightarrow \quad \boldsymbol{\sigma}^2\left\{\mathbf{b_w}\right\} = \mathbf{A}\boldsymbol{\sigma}^2\left\{\mathbf{Y}\right\}\mathbf{A'} = \left(\mathbf{X'WX}\right)^{-1}\mathbf{X'WW}^{-1}\mathbf{WX}\left(\mathbf{X'WX}\right)^{-1} = \left(\mathbf{X'WX}\right)^{-1}$$

# Unknown Error Variances

In reality, one rarely known the variances .

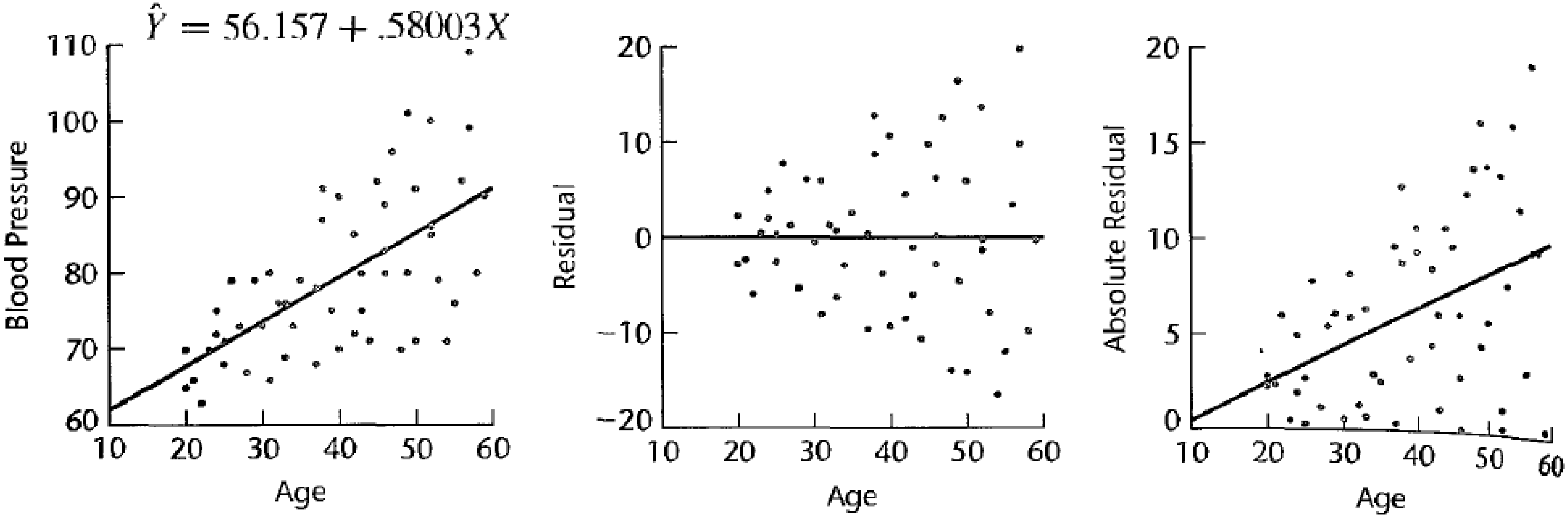
- Estimation of Variance Function or Standard Deviation Function
  - Variance (or Standard Deviation) is related to one or more predictors, and relation can be modeled (see Breusch-Pagan Test in Chapter 3)
- Use of Replicates or Near Replicates
  - When replicates or near replicates are available, use the sample variance of the replicates as the estimate for the variances.
  - In observational studies, replicate observations often are not available.

# Estimated Variances

- Use squared residuals (estimated variances) or absolute residuals (estimated standard deviations) from OLS to model their levels as functions of predictor variables
  - Plot Residuals, squared residuals and absolute residuals versus fitted values and predictors
  - Using model building techniques to obtain a model for either the variances or standard deviations as functions of the $X^s$
  - Fit estimated WLS with the estimated weights (1/variances)
  - Iterate until regression coefficients are stable (iteratively re-weighted least squares)

$$\hat{w}_i = \frac{1}{\hat{v}_i} = \frac{1}{\left(\hat{s}_i\right)^2} \qquad \mathbf{b}_{\hat{w}} = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{W}}\mathbf{Y} \qquad MSE_{\hat{w}} = \frac{\left(\mathbf{Y} - \mathbf{X}\mathbf{b}_{\hat{w}}\right)'\hat{\mathbf{W}}\left(\mathbf{Y} - \mathbf{X}\mathbf{b}_{\hat{w}}\right)}{n-p}$$

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Subject | Age | Diastolic Blood Pressure | | | | |
| $i$ | $X_i$ | $Y_i$ | $e_i$ | $\lvert e_i \rvert$ | $\hat{s}_i$ | $w_i$ |
| 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 |



Step1: do a regular OLS

$$\hat{Y} = \underset{(3.994)}{56.157} + \underset{(.09695)}{.58003X}$$

Step2: regress absolute residuals or squared residuals on X

$$\hat{s} = -1.54946 + .198172X$$

Step3: use the fitted values to get the weights

$$\hat{s}_1 = -1.54946 + .198172(27) = 3.801$$

$$w_1 = \frac{1}{(\hat{s}_1)^2} = \frac{1}{(3.801)^2} = .0692$$

Step4: use weighted least squares

$$\hat{Y} = 55.566 + \underset{(0.0792)}{.59634X}$$

# Multicollinearity – Remedial Measures - I

- Prediction – Multicollinearity not an issue if new cases have similar multicollinearity among predictors
- Inference of the coefficients
- Linear Combinations of Predictors can be generated that are uncorrelated (Principal Components)
  - Good: Multicollinearity gone
  - Bad: New variables may not have "physical" interpretation

# Multicollinearity – Ridge Regression

- Mean Square Error of an Estimator = Variance + Bias$^2$
- Goal: Add Bias to estimator to reduce MSE(Estimator)
- Makes use of Standard Regression Model with Correlation Transformation

$$MSE\left\{b^R\right\} = E\left\{\left(b^R - \beta\right)^2\right\} = \sigma^2\left\{b^R\right\} + \left(E\left\{b^R\right\} - \beta\right)^2 = \text{Variance + (Bias)}^2$$

Correlation Transformation / Standardized Regression Model:

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

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

$$\Rightarrow \quad \mathbf{r}_{XX}\mathbf{b} = \mathbf{r}_{YX}$$

# Ridge Regression

- OLS:

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{Y}$$

- Transformed by correlation transformation:

$$\mathbf{r}_{XX}\mathbf{b} = \mathbf{r}_{YX}$$

- Ridge Estimator: for a constant $c \geq 0$,

$$(\mathbf{r}_{XX} + c\mathbf{I})\mathbf{b}^R = \mathbf{r}_{YX}$$

- $c = 0$, OLS
- $c > 0$, biased, but much more stable.

Choose small *c* such that the estimators stabilize and VIF$^s$ get small.

- Ridge estimator

$$(\mathbf{r}_{XX} + c\mathbf{I})\mathbf{b}^R = \mathbf{r}_{YX}$$

$$\mathbf{b}^R = \left(\mathbf{r}_{XX} + c\mathbf{I}\right)^{-1}\mathbf{r}_{YX} = \begin{bmatrix} b_1^R \\ \vdots \\ b_{p-1}^R \end{bmatrix}$$

- It’s equivalent to minimize

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

# Choice of $c$

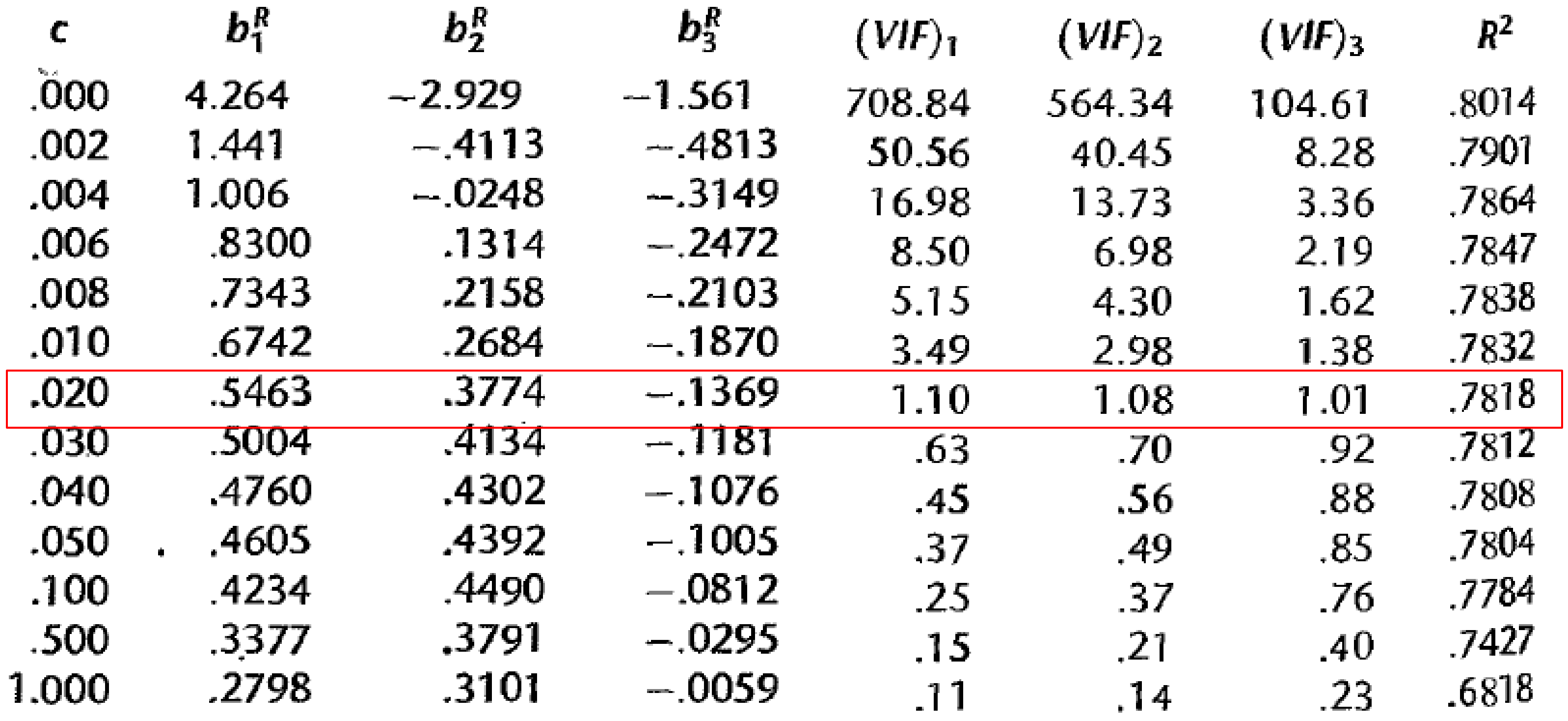
- Ridge trace ($0 \leq c \leq 1$) and the variance inflation factors

$$\mathbf{b}^R = \left(\mathbf{r}_{XX} + c\mathbf{I}\right)^{-1} \mathbf{r}_{YX}$$

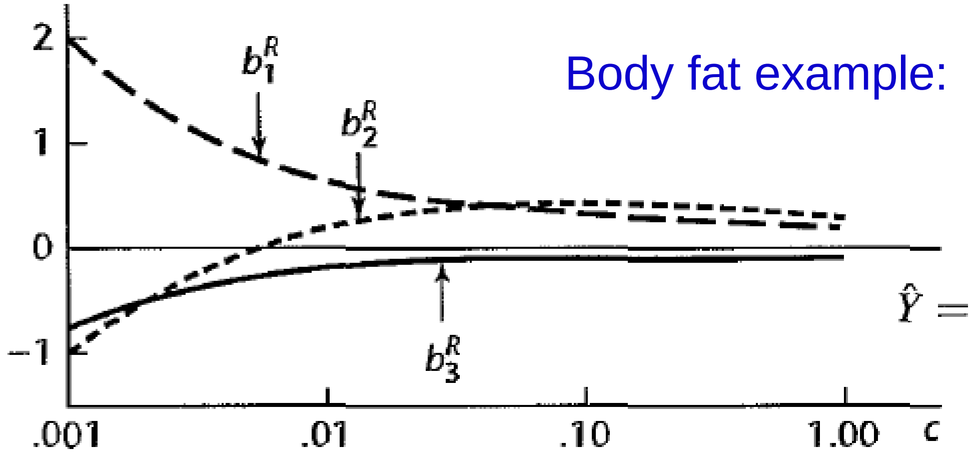
- Choose the $c$ where the Ridge trace starts to become stable and $\mathrm{VIF}_k$(c) has become sucffiently small.
- Recall: VIF value measure how large is the variance of b.
- Since

$$\sigma^2\{(\mathbf{r}_{XX} + c\mathbf{I})^{-1}\mathbf{r}_{YX}\} = (\mathbf{r}_{XX} + c\mathbf{I})^{-1}\mathbf{r}_{XX}(\mathbf{r}_{XX} + c\mathbf{I})^{-1}$$

$VIF_k(c)$ is the $k$-th diagonal element of $(\mathbf{r}_{XX} + c\mathbf{I})^{-1}\mathbf{r}_{XX}(\mathbf{r}_{XX} + c\mathbf{I})^{-1}$



| $c$ | $b_1^R$ | $b_2^R$ | $b_3^R$ | $(VIF)_1$ | $(VIF)_2$ | $(VIF)_3$ | $R^2$ |
|---|---|---|---|---|---|---|---|
| .000 | 4.264 | −2.929 | −1.561 | 708.84 | 564.34 | 104.61 | .8014 |
| .002 | 1.441 | −.4113 | −.4813 | 50.56 | 40.45 | 8.28 | .7901 |
| .004 | 1.006 | −.0248 | −.3149 | 16.98 | 13.73 | 3.36 | .7864 |
| .006 | .8300 | .1314 | −.2472 | 8.50 | 6.98 | 2.19 | .7847 |
| .008 | .7343 | .2158 | −.2103 | 5.15 | 4.30 | 1.62 | .7838 |
| .010 | .6742 | .2684 | −.1870 | 3.49 | 2.98 | 1.38 | .7832 |
| .020 | .5463 | .3774 | −.1369 | 1.10 | 1.08 | 1.01 | .7818 |
| .030 | .5004 | .4134 | −.1181 | .63 | .70 | .92 | .7812 |
| .040 | .4760 | .4302 | −.1076 | .45 | .56 | .88 | .7808 |
| .050 | .4605 | .4392 | −.1005 | .37 | .49 | .85 | .7804 |
| .100 | .4234 | .4490 | −.0812 | .25 | .37 | .76 | .7784 |
| .500 | .3377 | .3791 | −.0295 | .15 | .21 | .40 | .7427 |
| 1.000 | .2798 | .3101 | −.0059 | .11 | .14 | .23 | .6818 |



Body fat example:  take c=0.02

$$\hat{Y}^* = .5463X_1^* + .3774X_2^* - .1369X_3^*$$

$$\hat{Y} = -7.3978 + .5553X_1 + .3681X_2 - .1917X_3$$

# Robust Regression for Influential Cases

- Influential cases, after having been ruled out as recording errors can be reduced in terms of their impacts on the regression model
- Two commonly applied Robust Regression Models:
  - Least Absolute Residuals (LAR) or Least Absolute Deviation (LAD) Regression – Choose the coefficients that minimize sum of absolute deviations (as opposed to squared deviations in OLS). No closed form solutions, must use specialized programs (can be run in R)
  - IRLS Robust Regression – Uses Iteratively Re-weighted Least Squares where weights are based on how much each case is an outlier (lower weights for larger outliers)

- Robust to outlying and influential cases.
- LAD (Least Absolute Deviation) regression. To minimize

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

- LMS (Least Median of Squares) regression. To minimize

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

# IRLS Robust Regression

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

# IRLS – Robust Regression

Applying Procedure for Huber wight function

(1.345 is tuning parameter to achieve high efficiency to normal model)

1. Huber weight function:

$$w = \begin{cases} 1 & |u| \le 1.345 \\ \dfrac{1.345}{|u|} & |u| > 1.345 \end{cases}$$

2. Use initial residuals from OLS fit after scaling by

Median Absolute Deviation (Robust estimate of $\sigma$):

$$u_i = \frac{e_i}{\hat{\sigma}_R} = \frac{e_i}{MAD}$$

$$\boxed{\begin{array}{l} \varepsilon \sim N(0,\sigma^2) \Rightarrow \text{median}\,|\,\varepsilon - E(\varepsilon)\,| = \Phi^{-1}(0.75)\sigma \\ \Phi^{-1}(0.75) = 0.6745 \end{array}}$$

$$MAD = \frac{1}{0.6745}\,\text{median}\left\{\left|e_i - \text{median}\left\{e_i\right\}\right|\right\}$$

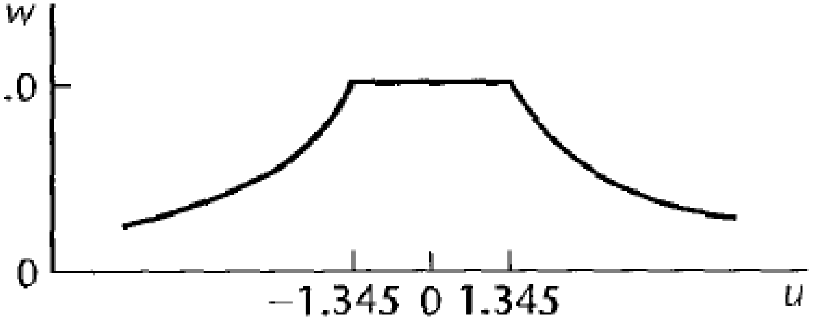
3. Iterate to Convergence

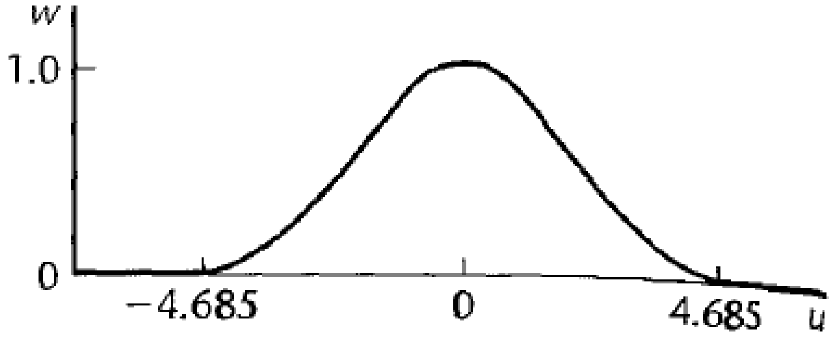
Huber weight function:

$$w = \begin{cases} 1 & |u| \leq 1.345 \\ \dfrac{1.345}{|u|} & |u| > 1.345 \end{cases}$$

Bisquare weight Function:

$$w = \begin{cases} \left[1 - \left(\dfrac{u}{4.685}\right)^2\right]^2 & |u| \leq 4.685 \\ 0 & |u| \leq 4.685 \end{cases}$$





# Nonparametric Regression

- Locally Weighted Regressions (Lowess)
  - Works best with few predictors, and when data have been transformed to normality with constant error variances, and predictors have been scaled by their standard deviation (sqrt(MSE) or MAD when outliers are present)
  - Applies WLS with weights as distances from the individual points in a neighborhood to the target
- Regression Trees
  - X-space is broken down into multiple sub-spaces (1-step at a time), and each region is modeled by the mean response
  - Each step is chosen to minimize the within region sum of squares

# Lowess Method

Two predictor variables, fitted value at $(X_{h1}, X_{h2})$.

- Distance Measure.

$$d_i = [(X_{i1} - X_{h1})^2 + (X_{i2} - X_{h2})^2]^{1/2}. \tag{7}$$

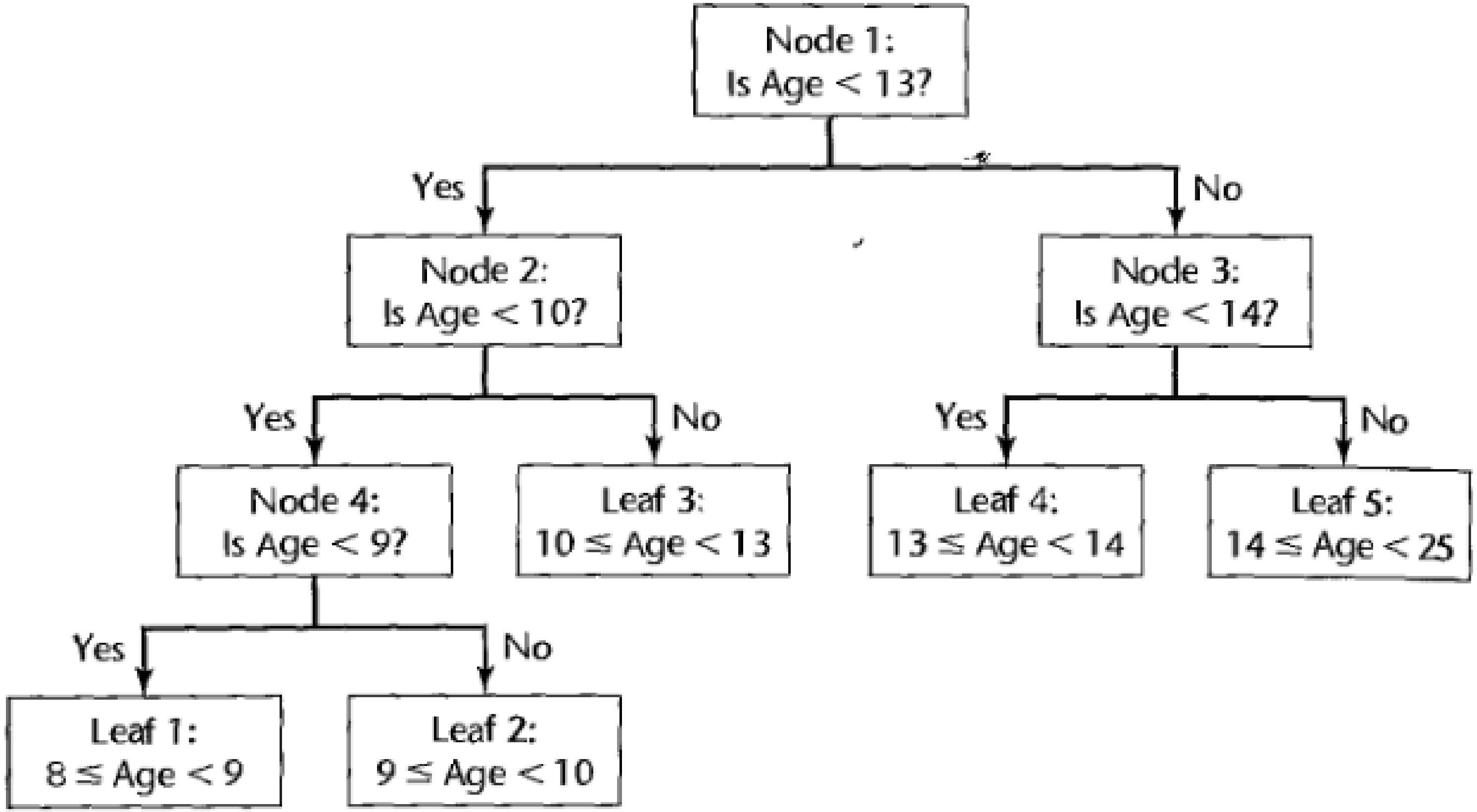
- Proportion of the data $q$ that are nearest to $(X_{h1}, X_{h2})$. Larger $q$ leads to smoother fit, but may increase the bias. Usually between .4 and .6.
- Weight Function.

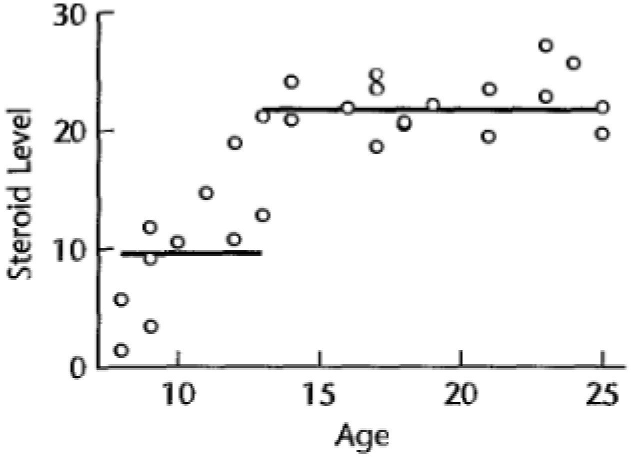
$$w_i = \begin{cases} \left[1 - (\frac{d_i}{d_q})^3\right]^3 & d_i < d_q. \\ 0 & d_i \geq d_q. \end{cases} \tag{8}$$

# Regression Trees

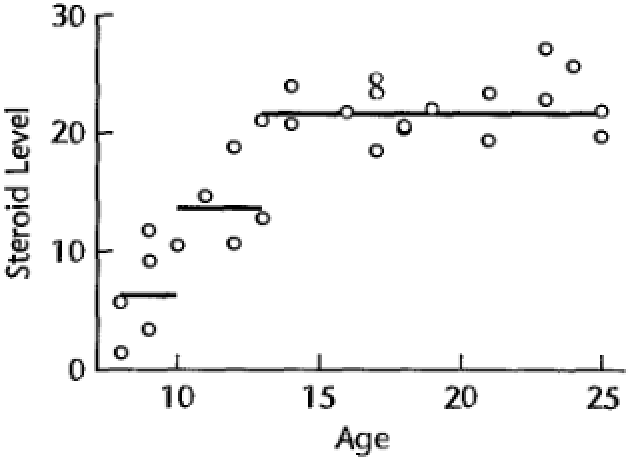
- A powerful nonparametric regression method.
- Can handle multiple predictors.
- Easy to calculate and require virtually no assumptions.
- Achieved by partitioning the covariates.
- Key quantities.
  - Number of regions r .
  - Split points between the regions.

- Take a single predictor as an example.

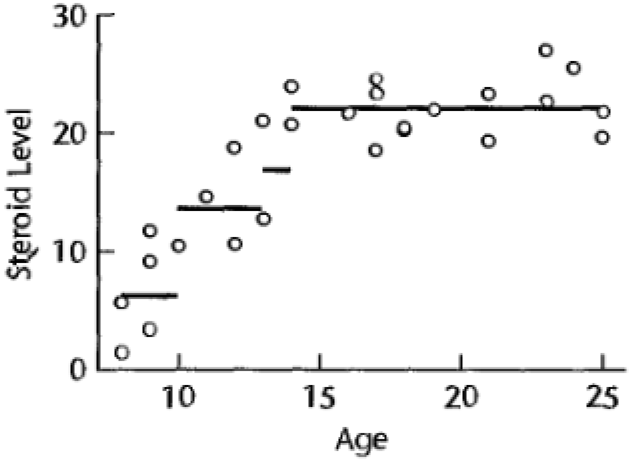



30
20
10
0
Steroid Level
8
10
15
20
25
Age

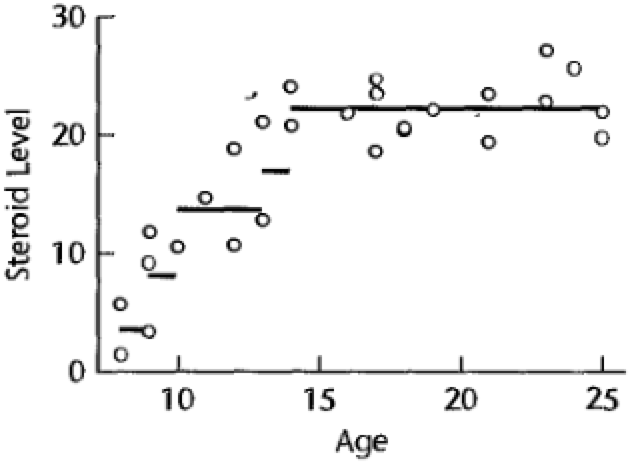
(a)


30
20
10
0
Steroid Level
8
10
15
20
25
Age

(b)


30
20
10
0
Steroid Level
8
10
15
20
25
Age

(c)


30
20
10
0
Steroid Level
8
10
15
20
25
Age

(d)

- Divide $X$ into two regions: $R_{21}$ and $R_{22}$.
- The optimal split point $X_s$ would be to minimize the SSE.

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

where

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

# Bootstrapping to Estimate Precision

- Computationally intensive methods used to estimate precision of estimators in non-standard situations
- Based on re-sampling (with replacement) from observed samples, re-computing estimates repeatedly. Once original sample is observed, and quantity of interest estimated, many samples of size n (selected with replacement) are obtained, each sample providing a new estimate.
- The standard deviation of the new sample quantities represents an estimate of the standard error

- Suppose we want to evaluate the precision of an estimated regression coefficient $b_1$.
- Fix B as the number of bootstrap samples to be generated. Say B = 500.
- For each k = 1,…,B, sample n observations with replacement from the original n observations.
- Fit a linear regression model using the bootstrap sample which leads to coefficient $b_1(k)$ .
- The sample standard deviation of $\{b_1 (1), b_1(2) ,…, b_1(B)\}$ is a measure of the precision of $b_1$.

# Two Basic Approaches

- Fixed X Sampling (Model is good fit, constant variance, predictor variables are fixed, i.e. controlled experiment)
  - Fit the regression, obtain all fitted values and residuals
  - Keeping the corresponding X-level(s) and fitted values, re-sample the n residuals (with replacement)
  - Add the bootstrapped residuals to the fitted values, and re-fit the regression (repeat process many times)

$$Y_i^* = \hat{Y}_i + e_i^*.$$

Then regress Y* values on the original X variables.

# Two Basic Approaches

- Random X Sampling (Not sure of adequacy of model fit, variance, random predictor variables)
  - After fitting regression, and estimating quantities of interest, sample n cases (with replacement) and re-estimate quantities of interest with "new" datasets (repeat many times)

# R code

```
bp = read.table('blood-pressure.txt')
X  = bp[,1];Y  = bp[,2]

###step1: do a regular OLS
lmfit = lm(Y~X)
plot(X,Y);  abline(lmfit)

resi = residuals(lmfit)
par(mfrow=c(1,2))
plot(X,resi);   abline(0,0)
plot(X,abs(resi))
```

# R code

```
###step2: regress absolute residuals on X
vfit = lm(abs(resi)~X)
abline(vfit)
###step3: use the fitted values to get the weights
wlist = vfit$fitted^(-2)
###step4: use weighted least squares
wlfit = lm(Y~X, weights=wlist)
###compare the two fit…
summary(lmfit)
summary(wlfit)
```

# R code

```
###########Ridge Regression
dat = read.table('fat.txt');  dat = scale(dat)
X1 = dat[,1];  X2 = dat[,2];  X3 = dat[,3];  Y = dat[,4]
library(MASS)
X = cbind(dat[,1:3])
lamlist = seq(from=0, to=1, by=0.01)
fit = lm.ridge(Y~X1+X2+X3-1, lambda=lamlist)

###########Robust regression
math = read.table('math.txt',header=F)
fit.robust = rlm(Y~X1+X2+X3+X4+X5)
```

```
####### regression trees
install.packages('tree');   library(tree)
data = read.table('steroid.txt',header=F)
names(data)<- c('steroid','age')
tree.control (dim(data)[1], minsize=5)
fit = tree(steroid~age, data=data)
plot(fit);   text(fit)

fit = tree(steroid~age, data=data, minsize=5)
cv.tree(fit, prune.tree)
fit2  = prune.tree(fit, k=4)
plot(fit2);  text(fit2)
```

# Homework

- P472 11.6(a), (d),(e)