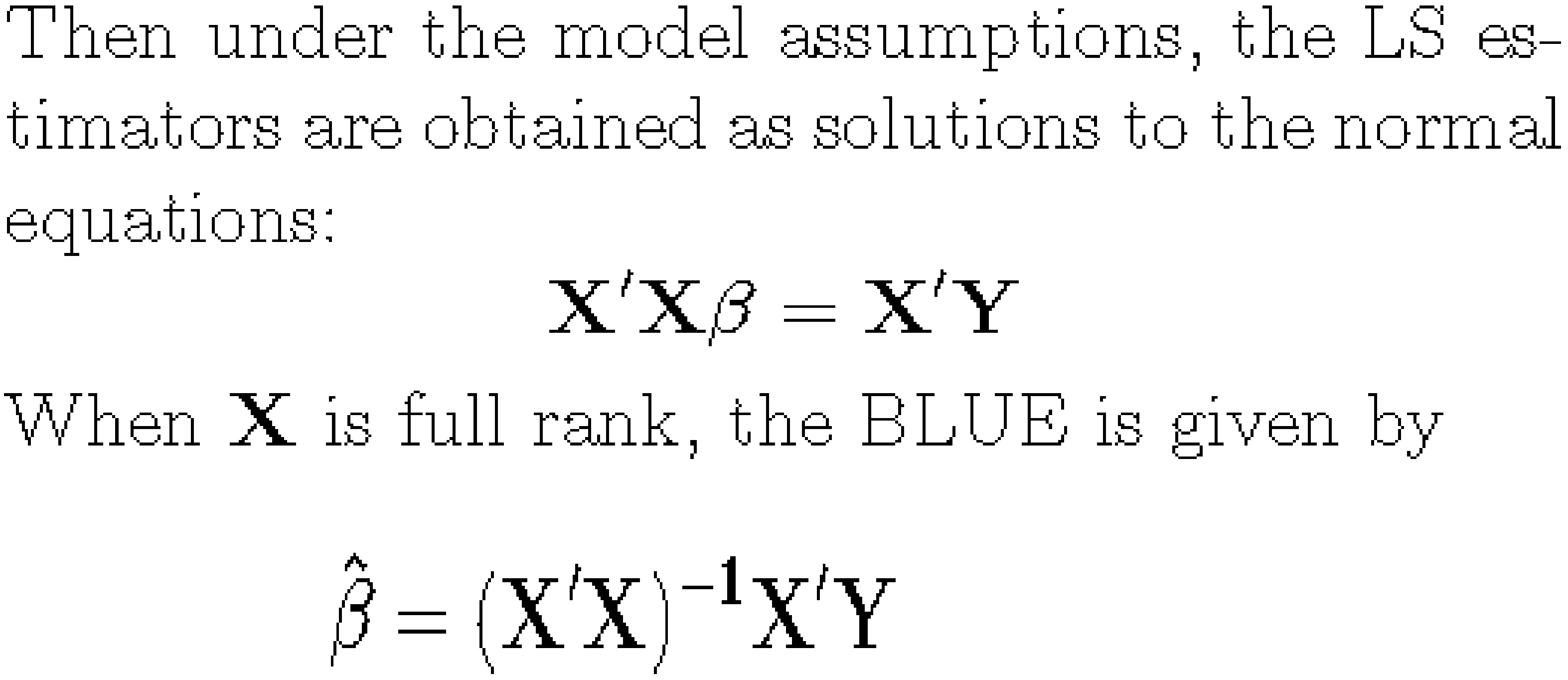
Linear Models

Let $(Y_i, X_{i,1}, \cdots, X_{i,p})$, $i = 1, \cdots, n$, be a random sample. It is often convenient to use the corresponding matrix formulations for the model:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$



Then under the model assumptions, the LS estimators are obtained as solutions to the normal equations:

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}$$

When $\mathbf{X}$ is full rank, the BLUE is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

OLS estimators, BLUE in the class of linear unbiased estimators.

- Model Validation
  - Linearity
  - Independence
  - Normality
  - Homoscedasticity
- Influential Points
  - Robust Statistics
- Multicollinearity
  - Ridge Regression
  - Principal Components
- Model Selection

```
> x <- cbind(x1,x2)
> Y2 <- Y
> Y
[1] 350.3001 203.9001 202.2531 202.7426 177.9737
```

```
> fit1 <-lm(Y~x1+x2)
> coef(fit1)
(Intercept)          x1          x2
 -0.9442516   3.2201178   6.0579185
```

```
> Y2 <- Y

> Y2
[1]   350.3001   203.9001   202.2531   202.7426 17797.3689
> fit2 <-lm(Y2~x1+x2)
coef(fit2)

(Intercept)          x1          x2
 19118.9442   -678.1958   -353.2782
```

Robust Regression

The goals of robust regression are:

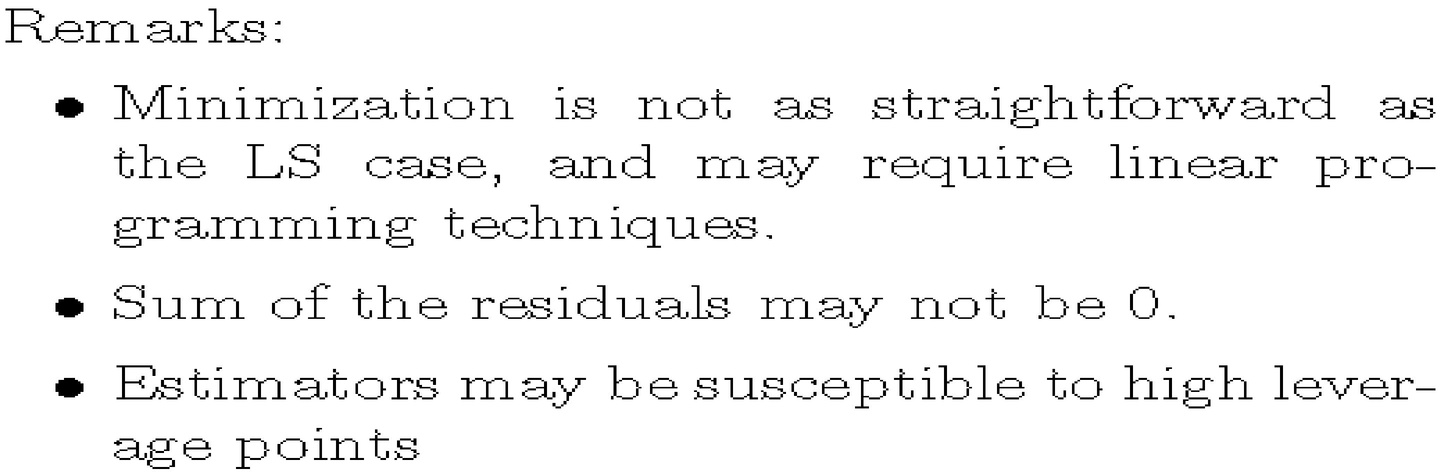
- To perform as well as the OLS when the latter works.
- To perform better than the OLS when the latter fails.
- Not complex to compute or understand.

$$\epsilon_i(\beta) = Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - X_{ip}$$

Least Absolute Deviation (L1) Regression

- Find estimators which minimize:

$$\sum_{i}^{n} \mid \epsilon_i(\beta) \mid$$

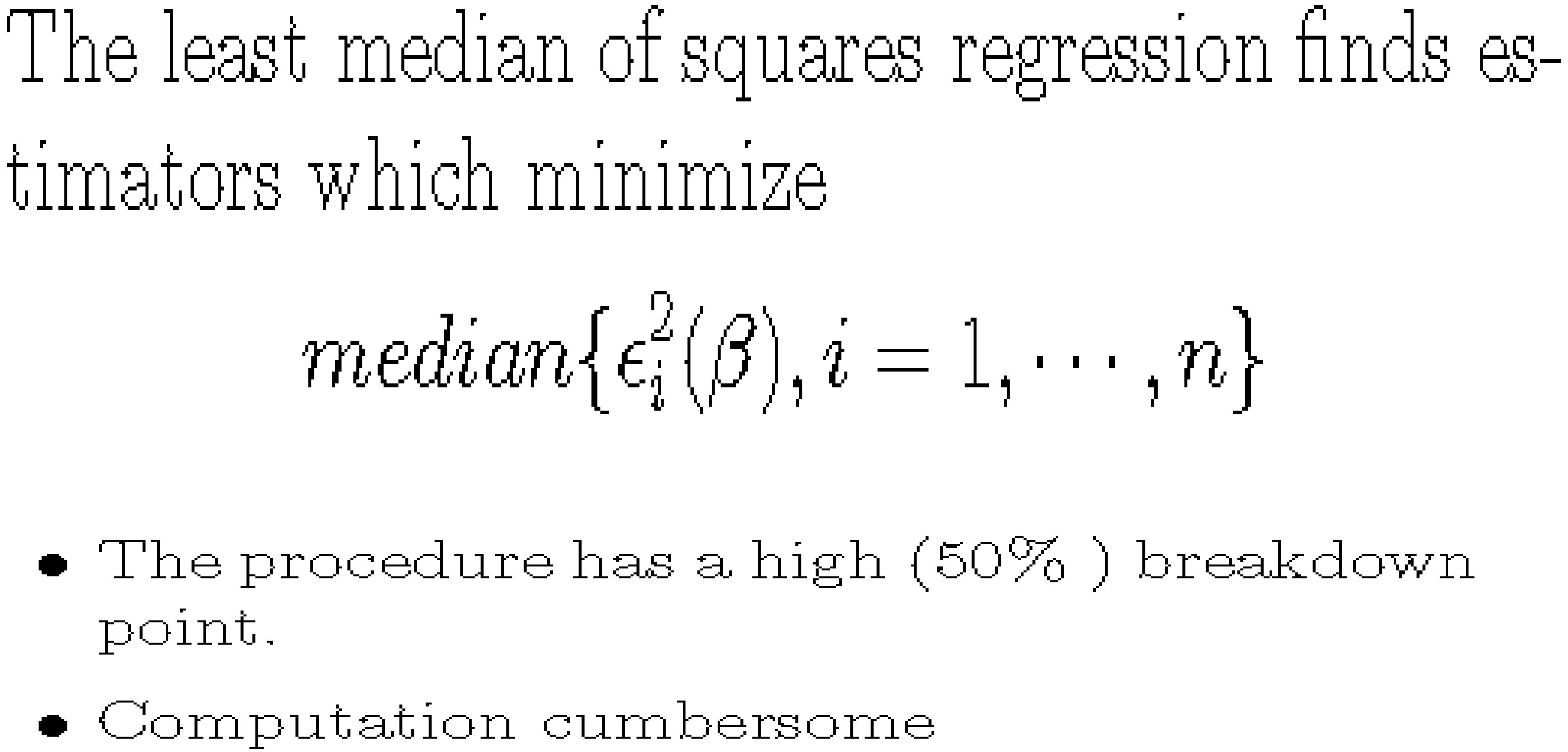


Remarks:

- Minimization is not as straightforward as the LS case, and may require linear programming techniques.
- Sum of the residuals may not be 0.
- Estimators may be susceptible to high leverage points

:

Least Median of Squares Regression



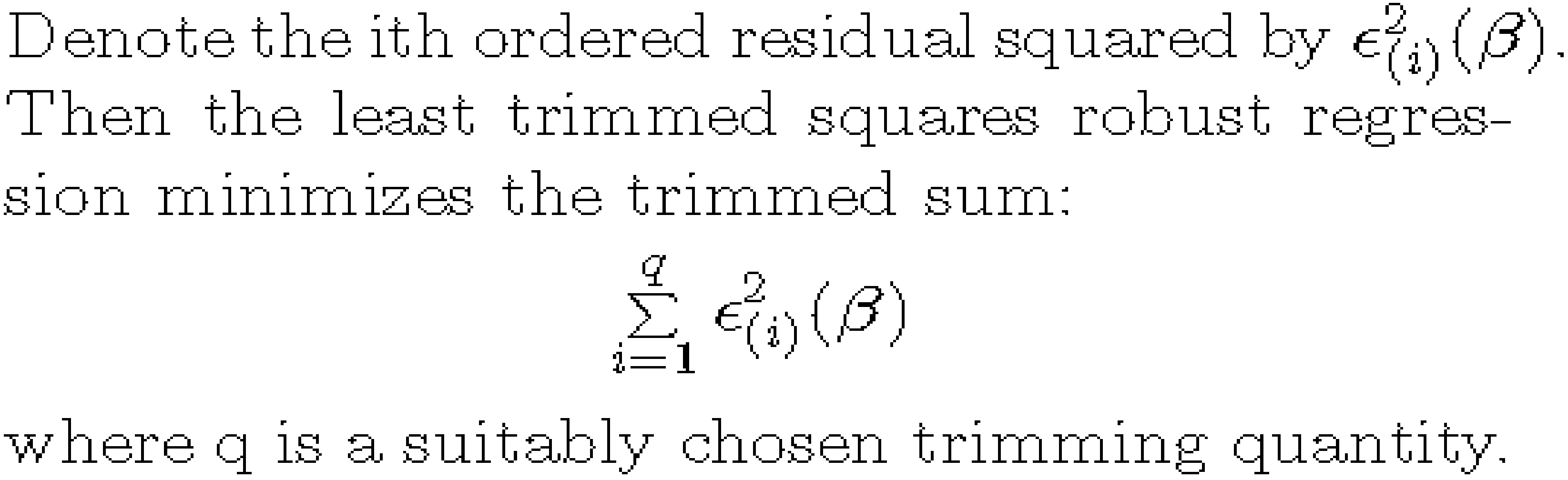
The least median of squares regression finds estimators which minimize

$$median\{\epsilon_i^2(\beta), i = 1, \cdots, n\}$$

- The procedure has a high (50%) breakdown point.
- Computation cumbersome

:

Least Trimmed Squares of Regression



Denote the ith ordered residual squared by $\epsilon^2_{(i)}(\beta)$. Then the least trimmed squares robust regression minimizes the trimmed sum:

$$\sum_{i=1}^{q} \epsilon^2_{(i)}(\beta)$$

where q is a suitably chosen trimming quantity.

Remarks

- Relatively high breakdown point, but < 50%.
- Calculation is complex, and uses random algorithms to get approximate solutions.

M-Estimates of Robust Regression

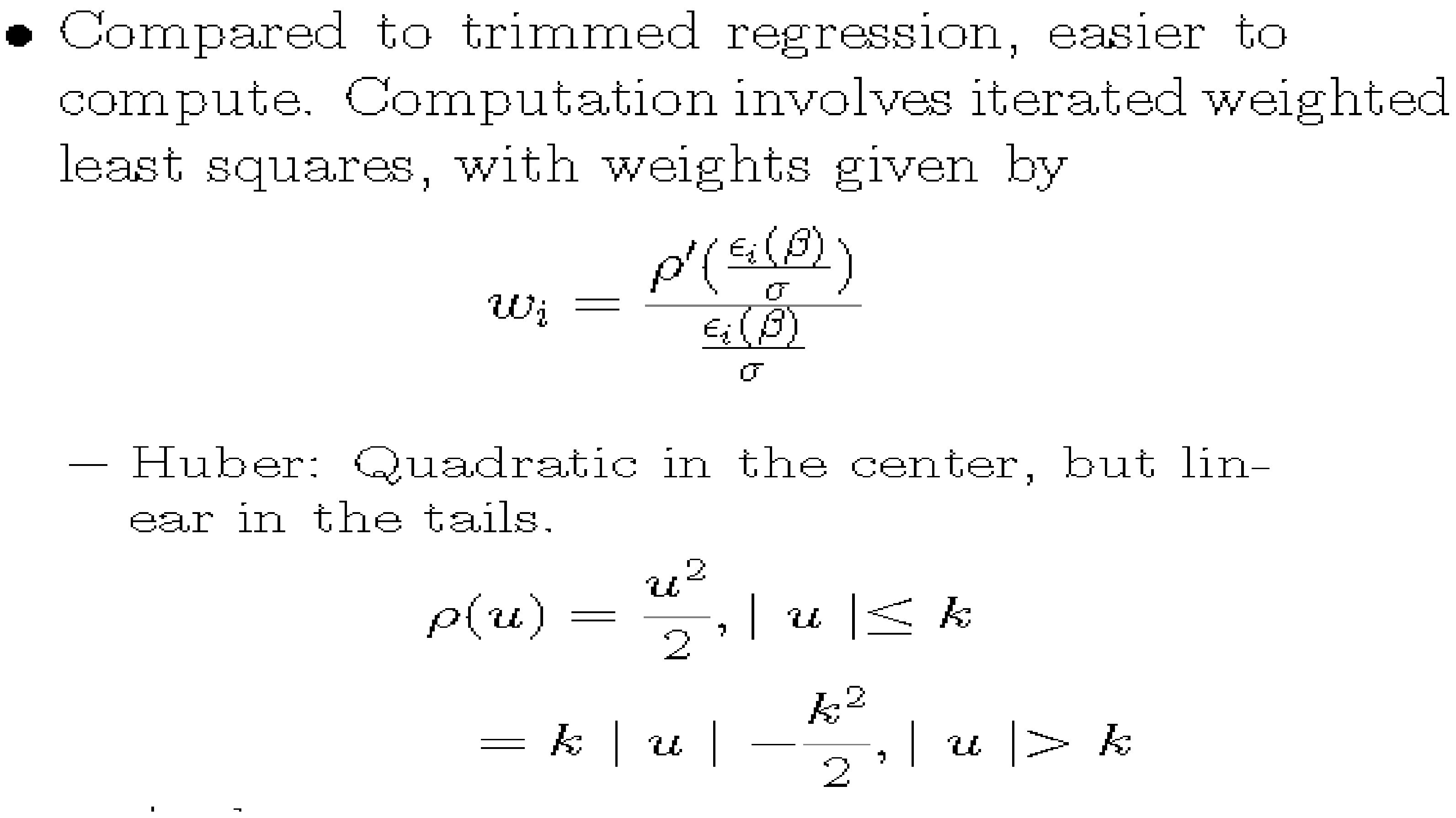
Given an objective function $\rho()$, M-estimates of robust regression estimates are obtained minimizing

$$\sum_{i=1}^{n} \rho(\frac{\epsilon_i(\beta)}{\sigma})$$

Remarks

- When $\rho(x) = x^2$, we get the OLS, whereas $\rho(x) =| x |$, gives L1 regression estimates.
- The procedure protects against Y outliers, but may be sensitive to leverage points in $\mathbf{X}$.

M-Estimates of Robust Regression



- Compared to trimmed regression, easier to compute. Computation involves iterated weighted least squares, with weights given by

$$w_i = \frac{\rho'\left(\frac{\epsilon_i(\beta)}{\sigma}\right)}{\frac{\epsilon_i(\beta)}{\sigma}}$$

  - Huber: Quadratic in the center, but linear in the tails.

$$\rho(u) = \frac{u^2}{2}, \mid u \mid \leq k$$

$$= k \mid u \mid - \frac{k^2}{2}, \mid u \mid > k$$

```
> x <- cbind(x1,x2)
> Y2 <- Y
> Y
[1] 350.3001 203.9001 202.2531 202.7426 177.9737

> Y2[5] <- Y[5] *100
> Y2
[1]   350.3001   203.9001   202.2531   202.7426 17797.3689
```

```
> fit1 <-lm(Y~x1+x2)
> fit2 <-lm(Y2~x1+x2)
> coef(fit1)
(Intercept)          x1          x2
 -0.9442516   3.2201178   6.0579185

> coef(fit2)
(Intercept)          x1          x2
 19118.9442   -678.1958   -353.2782
```

```
> fit1.lms <-lmsreg(x,Y)
> fit2.lms <-lmsreg(x,Y2)
> coef(fit1.lms)
(Intercept)          x1          x2
 -12.153527    4.888467    6.227675
> coef(fit2.lms)
(Intercept)          x1          x2
  -5.794811    3.426881    6.144069
 > fit1.rreg <- rlm(x,Y)
> fit2.rreg <- rlm(x,Y2)
> coef(fit1.rreg)
      x1       x2 
3.106378 6.047713 
> coef(fit2.rreg)
      x1       x2 
3.306668 6.042112 
```

R functions

- library(MASS)
- help(lqs)
- lqs(x,y,method=“lts”,”lqs”,”lms”,”S”)
- lmsreg()
- ltsreg()
- huber(); rlm()

Multicollinearity

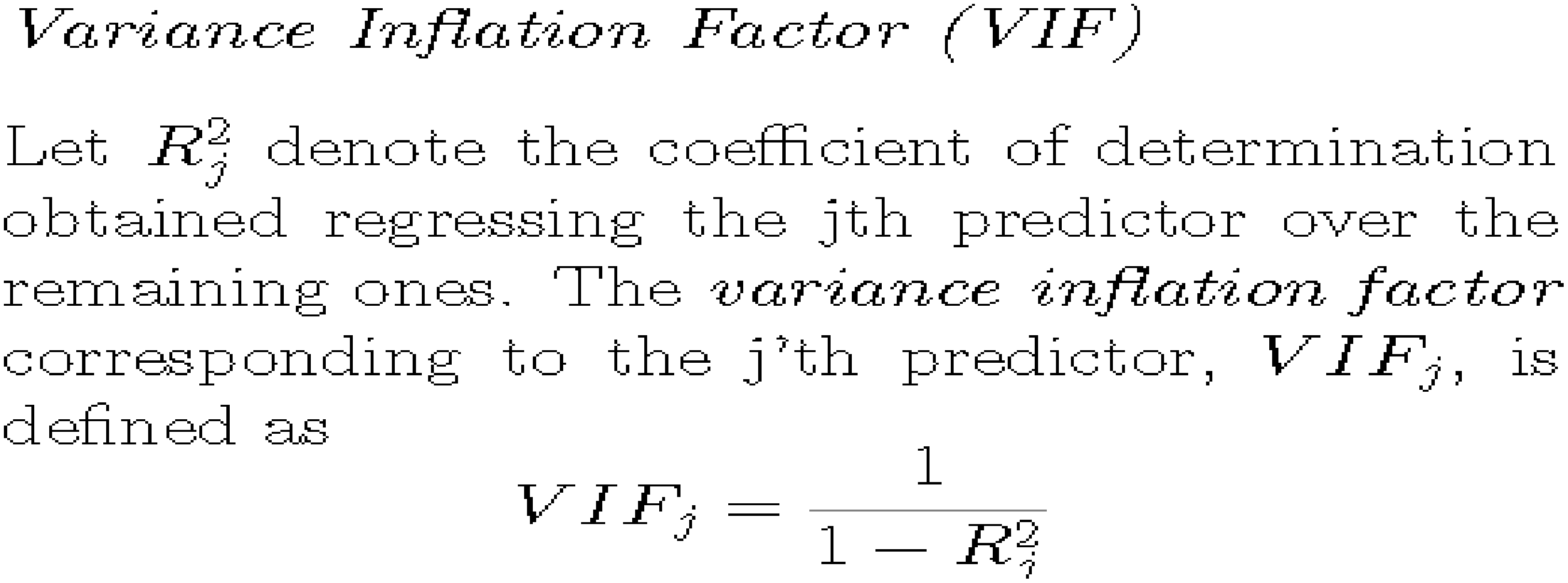
- Presence of a high degree of correlation among several independent variables.

Effects?

Multicollinearity: Effects

- The variance of the least squares estimators may be inflated. This will, of course, result in lack of significance or large confidence intervals.
- The least squares estimators my have the wrong signs.

Detection of Multicollinearity

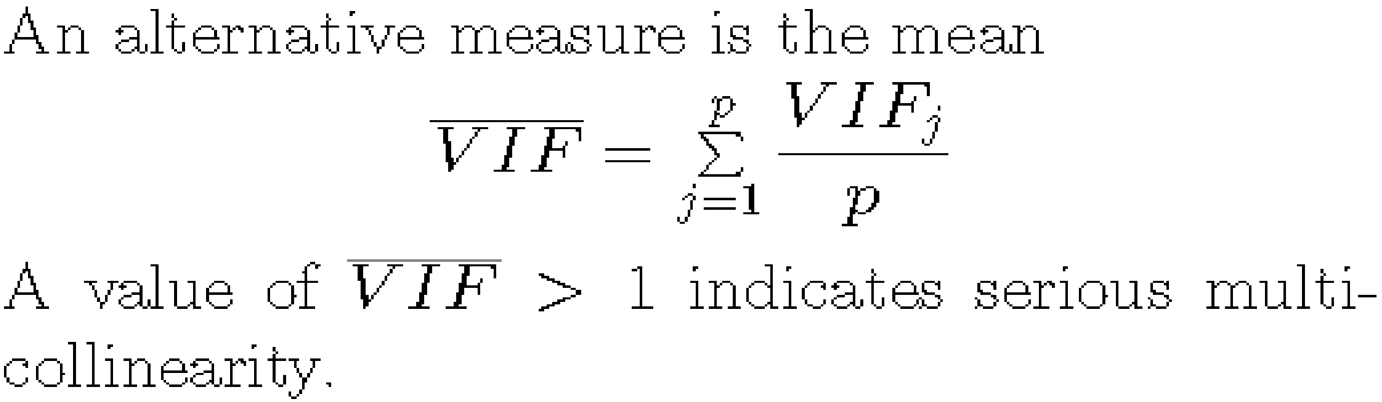


Variance Inflation Factor (VIF)

Let R_j^2 denote the coefficient of determination obtained regressing the jth predictor over the remaining ones. The ***variance inflation factor*** corresponding to the j’th predictor, VIF_j, is defined as

$$VIF_j = \frac{1}{1 - R_j^2}$$

Rule of thumb: VIF > 10 , multicollinearity



An alternative measure is the mean

$$\overline{VIF} = \sum_{j=1}^{p} \frac{VIF_j}{p}$$

A value of $\overline{VIF} > 1$ indicates serious multicollinearity.

Detection of Multicollinearity (cont’d)

Condition Number

Let $\mathbf{R}_{xx}$ denote the correlation matrix of X.

- Condition number:
  - Ratio of largest to smallest eigenvalues of $\mathbf{R}_{xx}$.
    - When exact collinearity, all eigenvalues will be 0.
    - A condition number > 30, multicollinearity
    - Values > 1000 generally imply serious collinearity.

Ridge Regression

Goal: Obtain estimators with some bias, but with smaller variance than an unbiased estimator.

A combined measure of bias and variance:

$$E[\tilde{\beta} - \beta]^2 = var(\tilde{\beta}) + [E(\tilde{\beta}) - \beta]^2$$

Find an estimator which minimizes the MSE.

Define the correlation transformations:

$$X_j' = \frac{X_j - \bar{X}}{\sqrt{n-1}Sx}$$

and

$$Y' = \frac{Y - \bar{Y}}{\sqrt{n-1}Sy}$$

Then the LS normal equations may be written as

$$\mathbf{R_{xx}b} = \mathbf{R_{yx}}$$

where $\mathbf{R_{yx}}$ is the correlation vector of Y and each of the explanatory variables.

$$(\mathbf{R_{xx}} - \mathbf{cI})\mathbf{b^R} = \mathbf{R_{yx}}$$

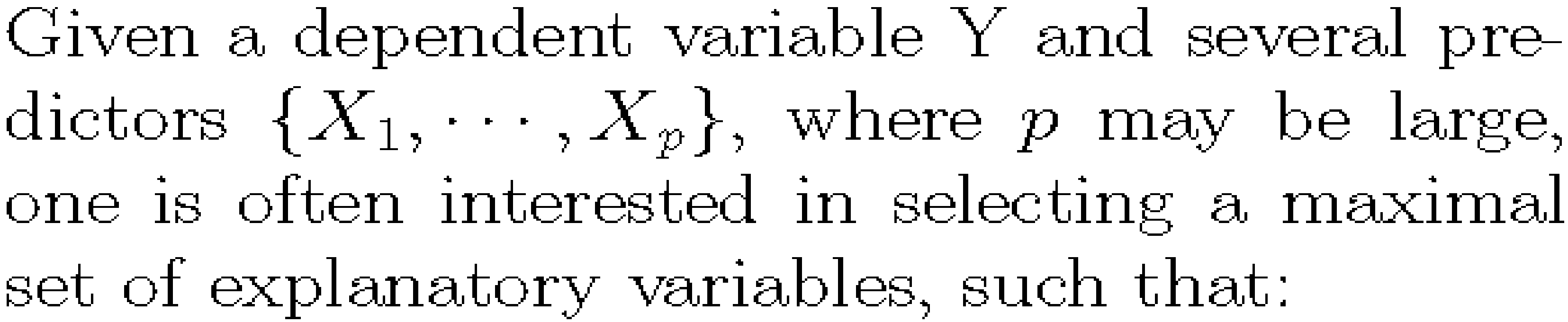
where $\mathbf{b^R}$ are the standardized ridge regression coefficient

$$\mathbf{b^R} = (\mathbf{R_{xx}} - \mathbf{cI})^{-1}\mathbf{R_{yx}}.$$

Principal Component Regression

- A common technique of obtaining uncorrelated predictors
- Involves transforming the independent variables using eigenvectors of R_{xx}.
- New variables, called principal components, are linear combinations of the original X's.
- Idea: Regress Y on the k useful principal components.
- Major drawback: Interpretation of the results not straightforward.

Model Selection



Given a dependent variable Y and several predictors $\{X_1, \cdots, X_p\}$, where p may be large, one is often interested in selecting a maximal set of explanatory variables, such that:

- The selected variables provide maximal predictive value

- All redundant variables are excluded from the model

- The size of the included set of variables permits the estimation of parameters with reasonable precision.

Why a parsimonious model?

- When number of variables is large relative to the number of observation, estimates of associated variables may be unstable or imprecise.

- Interpretation of models with too many predictors may be complex

- Too may variables may involve cumbersome computational efforts.

Criteria for Model Selection

i) Residual Sum of Squares (RSS)

Given alternative models, one which has the smallest RSS is chosen.

ii) Coefficient of Determination (R^2)

iii) The Adj-R^2 Criterion

Adjusted coefficient of determination

$$Adj - R^2 = 1 - \frac{n-1}{n-k-1}\frac{SSE_k}{SST}$$

Adj R^2 increases iff MSE_k decreases, where

$$MSE_k = \frac{SSE_k}{n-k-1}$$

Criteria for Model Selection

iv) The F Statistic Criterion

For a given set of coefficients, the F statistic corresponding to $H_o : \beta_k = 0$ is evaluted, and the one with the largest value and exceeding a threshold limit is taken.

v) The C_p Criterion (Mallow's C_p)

Let

$$C_p = \frac{SSE_k}{MSE_p} - (n - 2(k+1))$$

where SSE_k is SSE based on a subset having k predictors.

Criteria for Model Selection (cont'd)

- Note that

  $$E[C_k] = k + 1$$

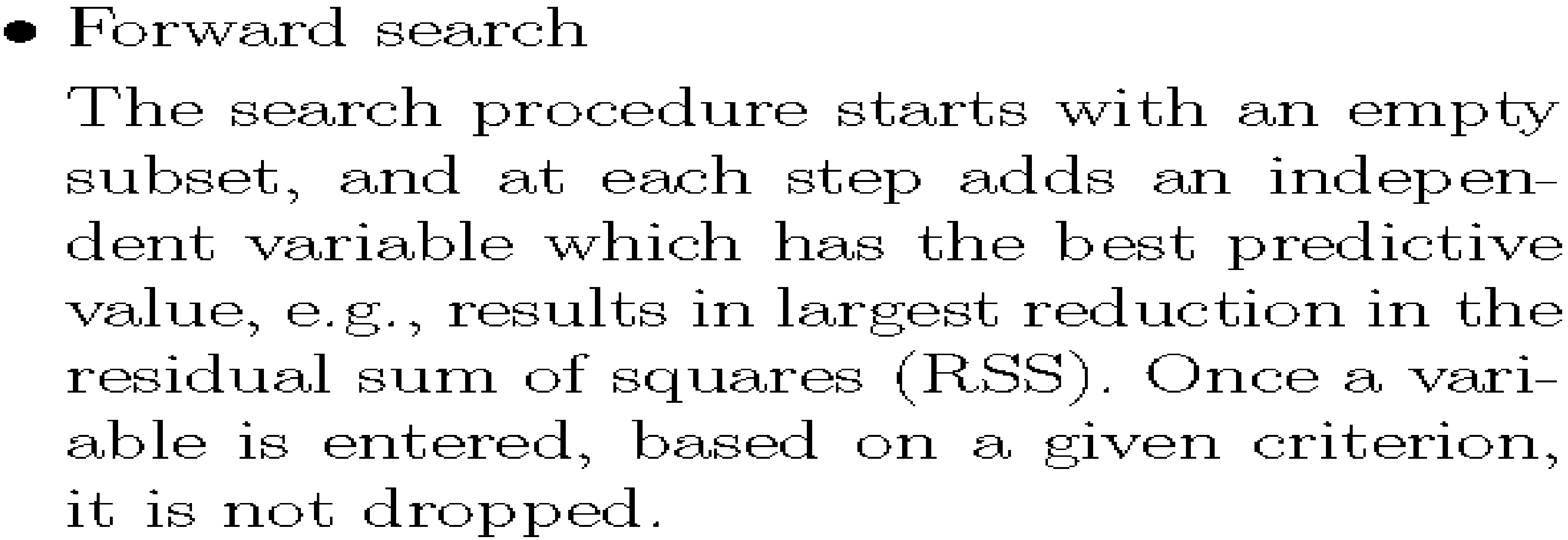
  Hence, if for some k, $C_k > k + 1$, bias, due to incompletely specified model.

  Conversely, $C_k < k+1$, overspecified model.

One would then identify a set of X's such that

- C_k is small
- $C_k \approx k + 1$
- Plot C_k vs $k + 1$, and choose the optimal set.

Stepwise Regression



- Forward search

  The search procedure starts with an empty subset, and at each step adds an independent variable which has the best predictive value, e.g., results in largest reduction in the residual sum of squares (RSS). Once a variable is entered, based on a given criterion, it is not dropped.

- Backward elimination

  Begin with a model containing all potential explanatory variables.

  At each step drop explanatory variable with least predictive value.

  Approach is computationally more cumbersome than forward method.

Stepwise Regression (cont’d)

- Efroymson's method

  Similar to *forward search* approach, except that when a new variable is entered, partial correlations are considered to see if any of the variables in the model should now be dropped.

- Exhaustive search

  Searches the optimal model by looking at all possible subsets and then looking at the criteria corresponding to the best subset.

  When p is large, this approach may not be appropriate

The Akaike Information Criterion (AIC) is a more general technique used in model selection.

R functions:

```
> library(help="MASS")
> help(stepAIC)
```

Statistical Issues with Stepwise Model Selection Procedures

- Stepwise regression should only be used for exploratory purposes or for purposes of pure prediction
- It should not be used for theory testing
  - Hypothesis must be pre-specified
- Nominal significance level used at each step is subject to inflation
  - Hard to adjust for multiplicity
- Automated fitting may lead to over-fitting
  - Generalization across data unreliable
  - R^2 estimates too high
  - Confidence intervals too narrow
- Affected by multicollinearity
- Dummy variables are usually treated individually
  - No obvious approach to add/remove sets of dummy variables

Reading Assignment

- See references given in: http://www.stata.com/support/faqs/stat/stepwise.html

Transformations

- Transforming models to achieve linearity

$$Y = exp\{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X + p\}\epsilon$$

- **Transforming predictors to improve fit**

$$E[Y] = \beta_o + \beta_1 X'$$

where

$$X' = \begin{cases} ln(X), & \alpha = 0 \\ X^{\alpha}, & \alpha \neq 0 \end{cases}$$

Box-Tidwell transformations

Transformations (cont'd)

- *Transforming Y to improve fit*

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

Box-Cox transformation

In R, boxcox(); library(MASS)

- *Transforming Y to achieve normality*

Transformations (cont'd)

- *Transforming Y to correct heteroscedasticity*

$$f(Y) = f(\mu) + f'(\mu)(Y - \mu) + (Y - \mu)^2 R_n$$

Then

$$var[f(Y)] \approx [f'(\mu)]^2 var(Y)$$

If Y is binomial, may use arcsin transformation.

For Poisson: Sqrt

Nonlinear Regression Models

If the functional form is known, nonlinear models may be fit. Let

$$Y = f(X, \beta) + \epsilon$$

- *Exponential models* . Such models may be used, for example, to relate concentration Y to elapsed time X:

$$Y_j = \beta_0 + \beta_1 e^{\beta_2 X_j} + \epsilon_j$$

- *General logistic model.* In population studies, such models may be used to relate number of species Y to time (X).

$$Y_j = \beta_0 / (1 + \beta_1 (exp\{\beta_2 X_j\}) + \epsilon_j$$

Nonlinear Regression Models

Estimation of the parameters involves applying least squares criteria, i.e., minimizing, with respect of the parameters, the sum of squares:

$$\sum_{j=1}^{n}(Y_j - f(X_j, \beta))^2$$

Solutions are often obtained by numerical methods.

- *The Gauss-Newton Method*
- *Method of steepest descent*
- *Marquardt algorithm*

Reading Assignment

Nonlinear Regression Models

A biased estimator of the error variance σ^2 is given by

$$\hat{\sigma}^2 = \sum_{j=1}^{2} (Y_j - f(x_j, \hat{\beta}))^2/(n-p)$$

It can be shown that $\hat{\beta}$ is approximately multivariate normal with mean $\hat{\beta}$ and variance covariance matrix $\sigma^2(D'D)^{-1}$, where D is a matrix of partial derivatives. Inference may be performed based on the Wald-type statistics:

$$T_k = (\hat{\beta}_k - \beta_k)/SE(\hat{\beta}_k).$$

Example 1. Orange[1:4,]

```
Grouped Data: circumference ~ age | Tree
  Tree  age circumference
1    1  118            30
2    1  484            58
3    1  664            87
4        1 1004            115
5        …..
```

➢R

➢nlsfit _ nls(circumference~A/(1+exp(-1*(age-B)/C)),
                        data=Orange, start=list(A=150,B=600,C=400))

```
> summary(nlsfit)

Parameters:
    Value Std. Error t value
A 192.718    20.2602 9.51214
B 728.912   107.3770 6.78836
C 353.650    81.5052 4.33898

Residual standard error: 23.3721 on 32 degrees of freedom
```

Scatterplot Smoothers

- Let Y=*f*(x) +e, with functional form of *f* unknown

Given data: (x_1,y_1), ….., (x_n,y_n), approximate *f*(x), by

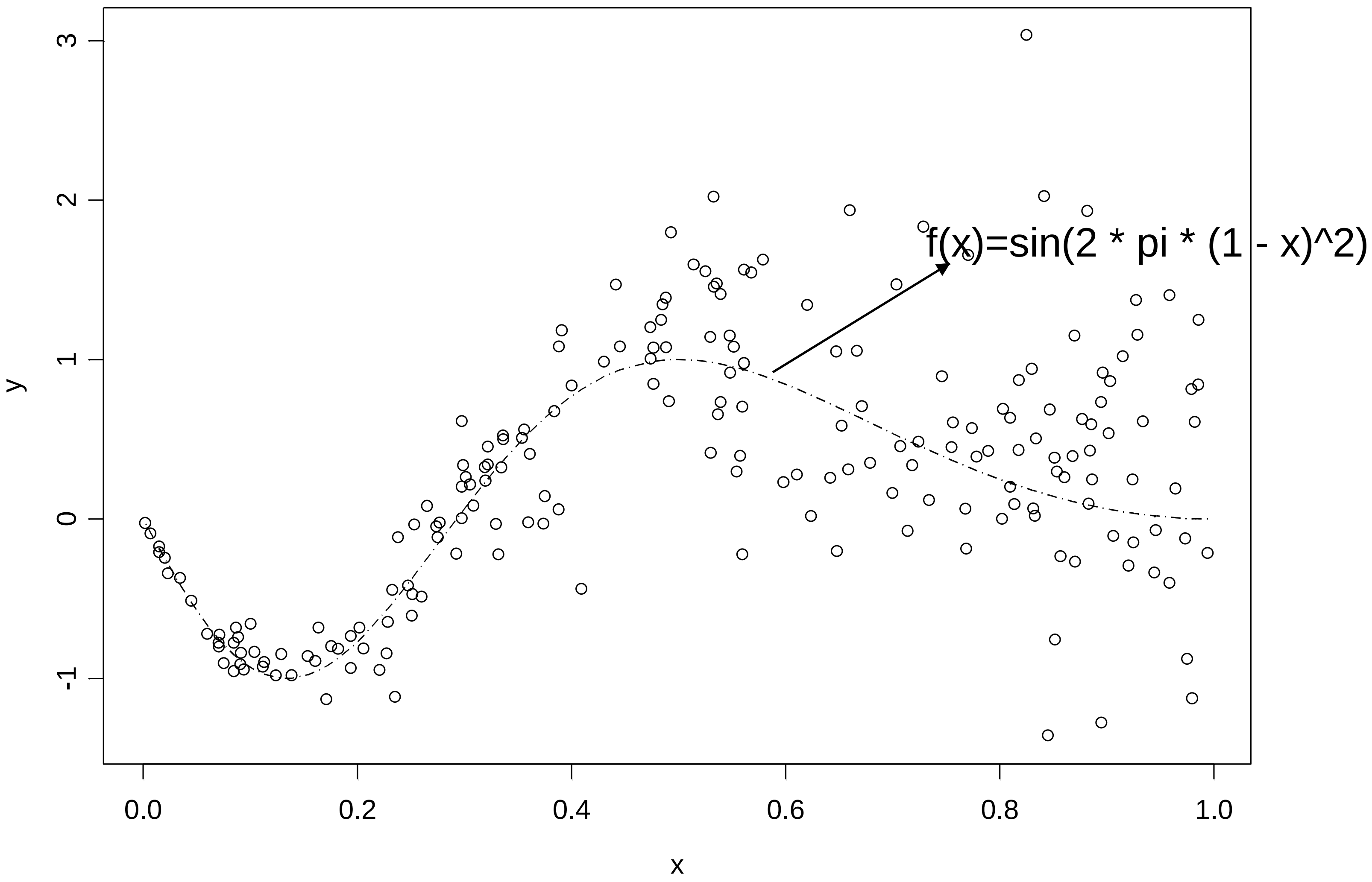
$$\hat{f}(x_i) = \sum_{j=1}^{n} w_{ij} y_j$$

Example 2: Suppose *f*(x) = sin(2 * pi * (1 - x)^2)

Generate

200 data points

```
x <- runif(200)
e <- rnorm(200)
x <- sort(x)
fx <- sin(2 * pi * (1 - x)^2)
y <- fx + x * e
```


f(x)=sin(2 * pi * (1 - x)^2)
3
2
1
0
-1
y
0.0
0.2
0.4
0.6
0.8
1.0
x

Scatterplot Smoothers

Kerenel-type Scatterplot Smoothers

Approximate f(x) by :

$$\hat{f}(x_i) = \sum_{j=1}^{n} w_{ij} y_j$$

where

$$w_{ij} = \frac{K(x_i - x_j)/b}{\Sigma_j K(x_i - x_j)/b}$$

- The kernel function satifies:
  - $K(t) \geq 0 \quad \forall t$
  - $\int K(t) = 1$
  - $K(-t) = K(t), \forall t$

- The kernel function is chosen such that values of Y_j with X_j close to x get larger weights compared to those that are some distance away.

Common choices:

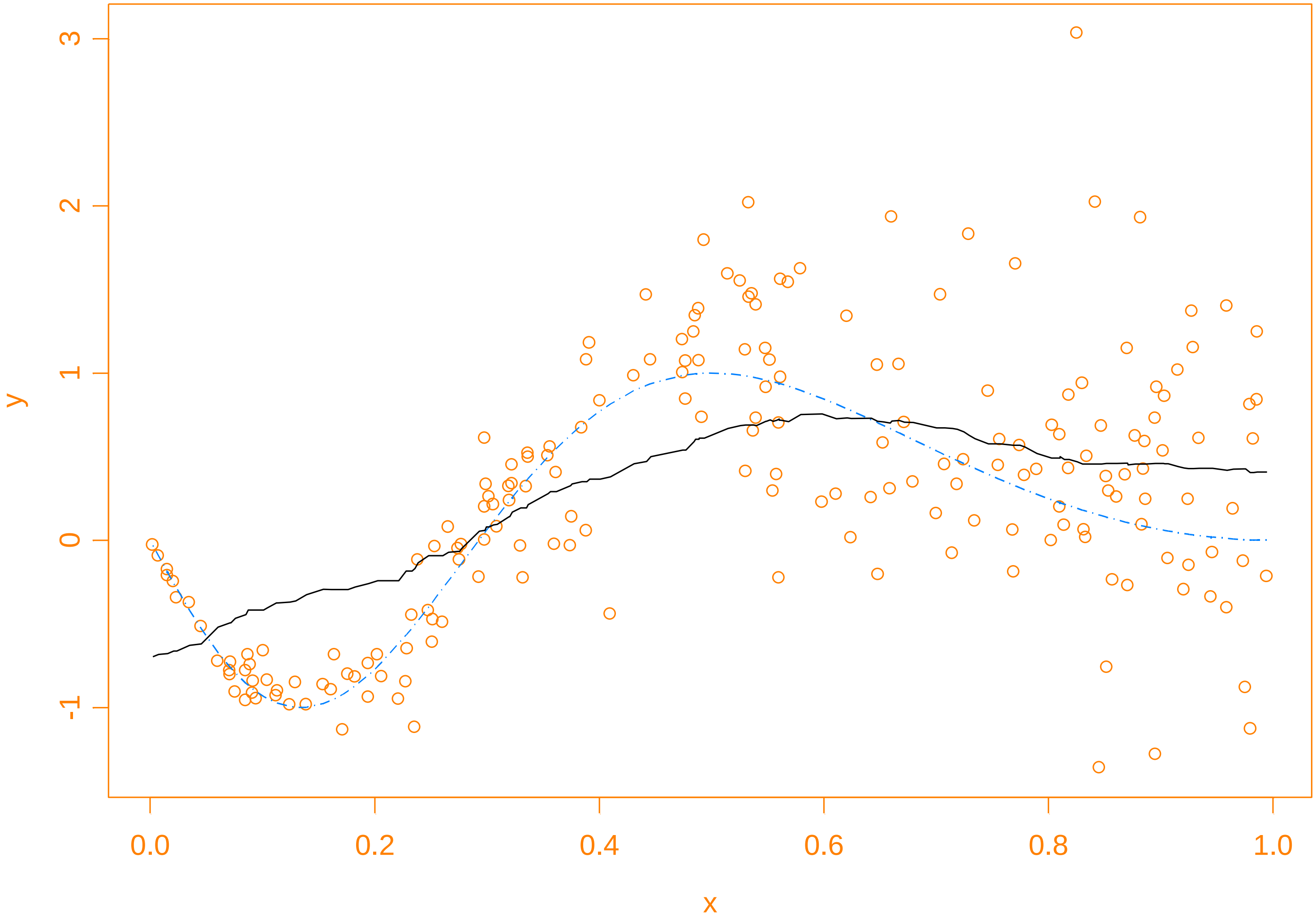
  - Box: $K(t) = 1$, for $\mid t \mid \leq 0.5$, and 0, otherwise.
  - Normal

$$K(t) = exp\{-t^2/(2\sigma^2)\}/(2\sqrt{\pi}\sigma)$$

- The bandwidth b must be estimated from the data. Large values of b result in biased but smoothed plots, while smaller values give less smooth, buth highly variable curve estimates. The optimal choice, therefore, balance bias and variance.

R functions

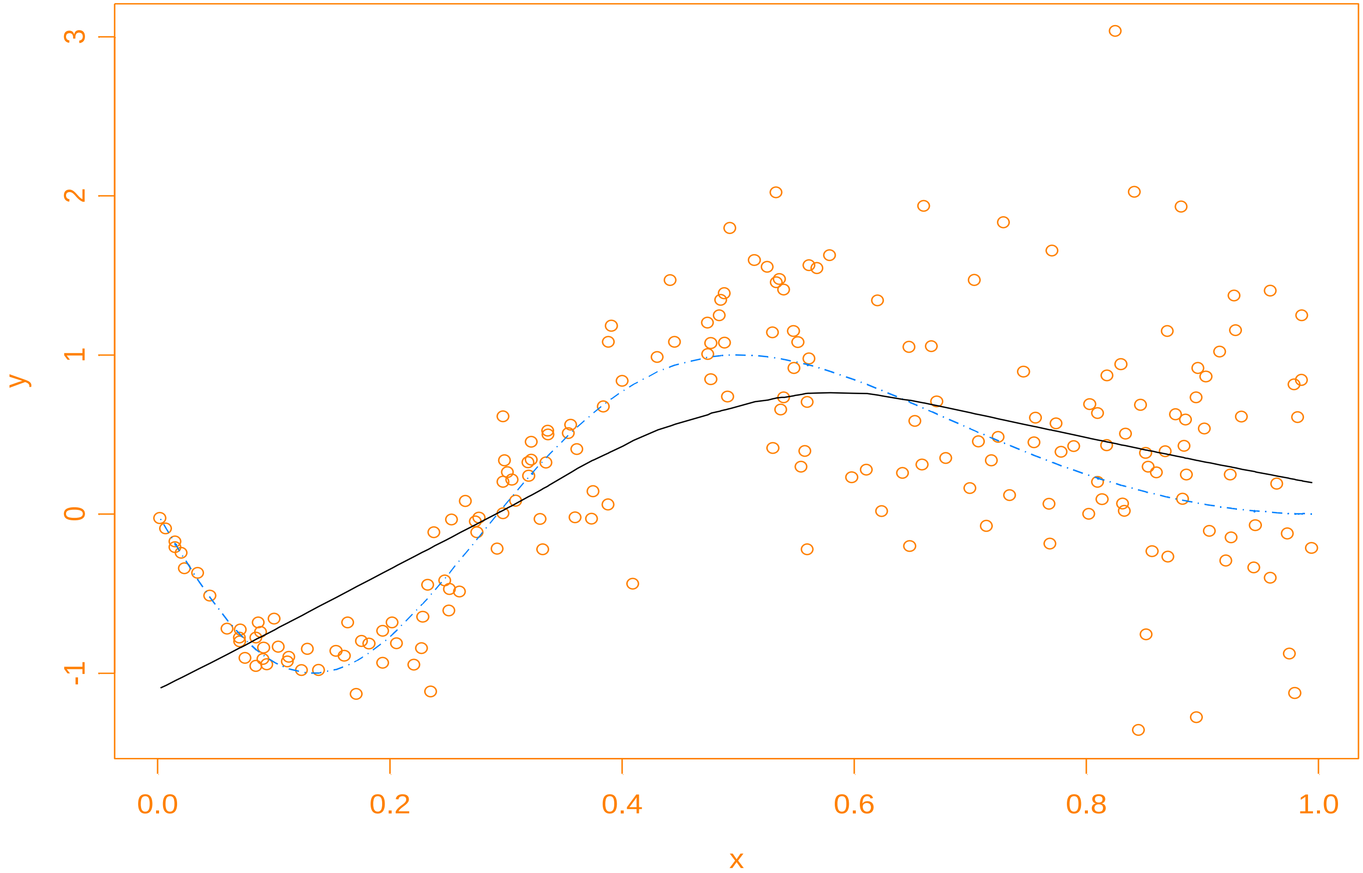
```
library(KernSmooth)
```



Kernel Smooth: Solid Line

Locally weighted regression smoothing

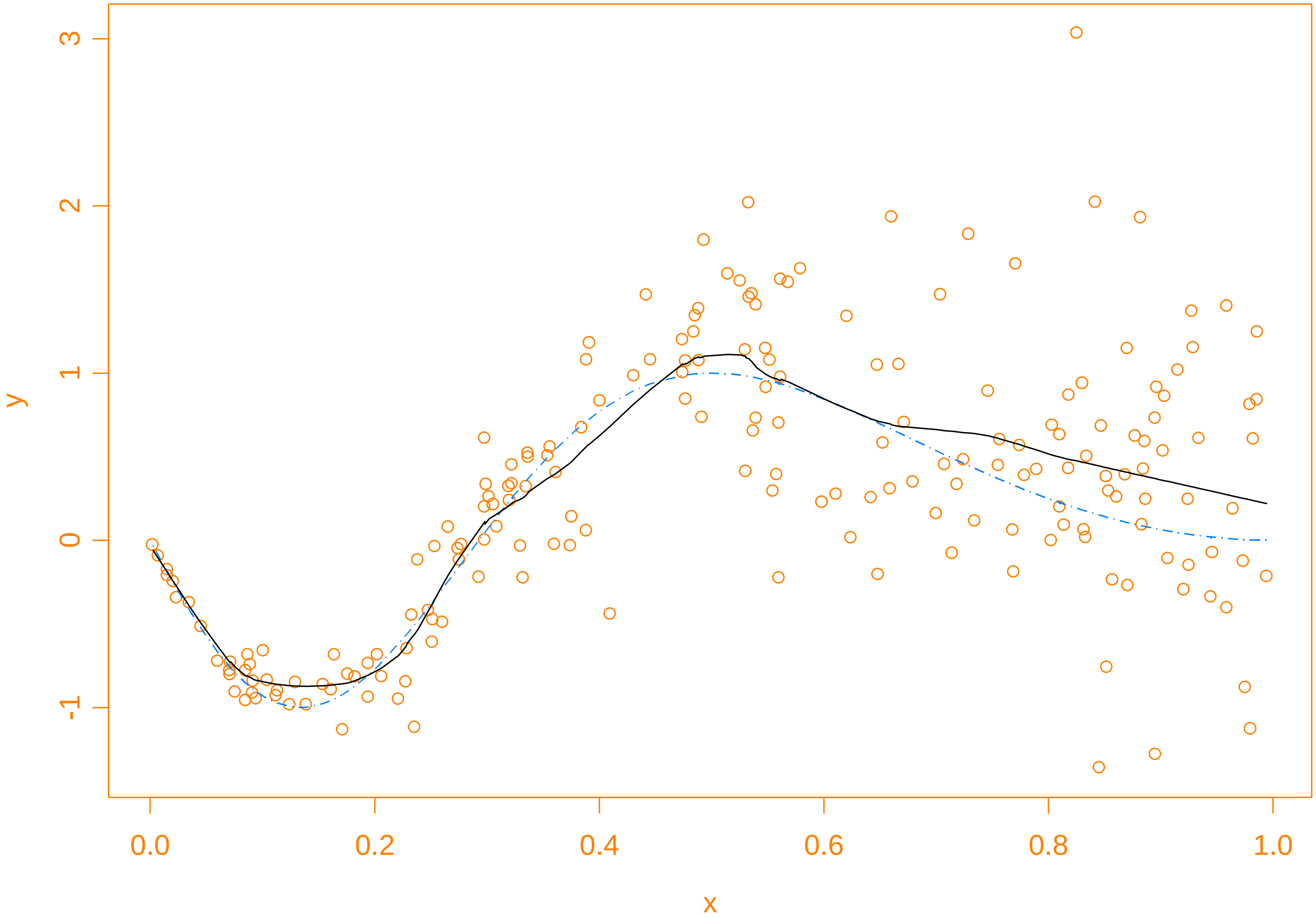
- WLS fit of y are computed based on the k nearest neighbors of x.
- Typically, a fixed span, i.e., the fraction of nearest neighbors, is kept constant for the entire range of x.
- The choice of the span is again determined with a view to balancing between variance and bias.



Lowess: Solid Line

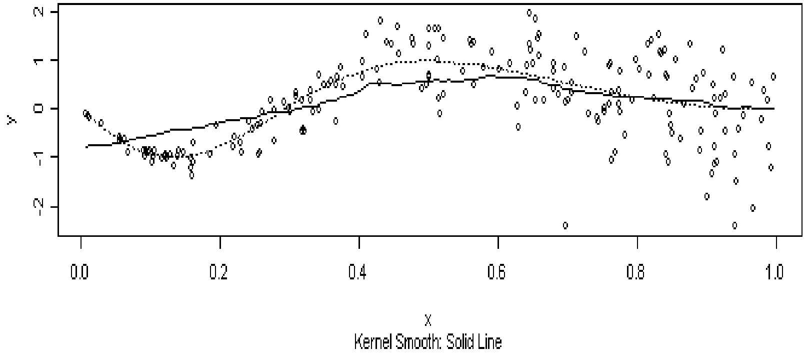
Supersmoother

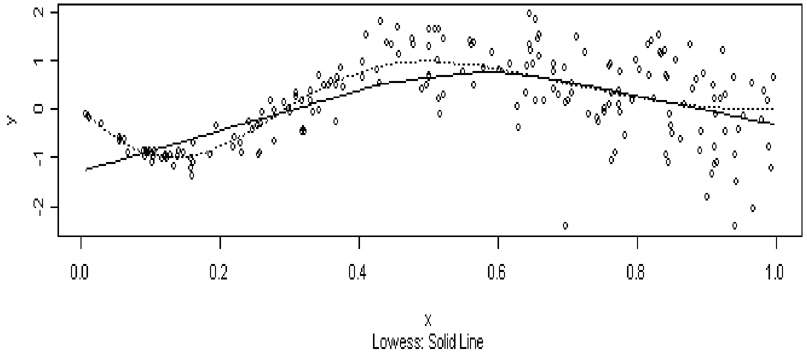
- Employs variable bandwidth (span), which is selected based on the data.
- Cross-validation technique is employed for the optimal bandwidth corresponding to each data point.

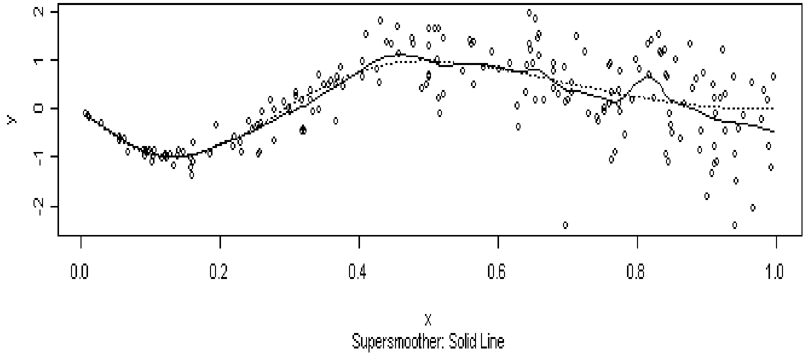


Supersmoother: Solid Line

```
R Function
Smoother <-function()
{
            set.seed(15)
            x <- runif(200)
            e <- rnorm(200)
            x <- sort(x)
            fx <- sin(2 * pi * (1 - x)^2)
            y <- fx + x * e
            par(mfrow = c(3, 1))
            plot(x, y)
            lines(x, fx, lty = 3)
            lines(ksmooth(x, y))
            title(sub = "Kernel Smooth: Solid
Line")
            plot(x, y)
            lines(x, fx, lty = 3)
            lines(lowess(x, y))
            title(sub = "Lowess: Solid Line")
            plot(x, y)
            lines(x, fx, lty = 3)
            lines(supsmu(x, y))
            title(sub = "Supersmoother: Solid
Line")
}
```







Problem Set 5

Reading Assignment: Ramsey & Schafer, Chapters 9, 10, 11, 12.

1. Using two or more suitable model selection criteria, select an optimal model to predict *'bwt' birth weight in grams* using the following set of predictors:

   'age' mother's age in years
   'lwt' mother's weight in pounds at last menstrual period
   'race' mother's race ('1' = white, '0' = other)
   'smoke' smoking status during pregnancy
   'ptl' number of previous premature labours
   'ht' history of hypertension
   'ui' presence of uterine irritability
   'ftv' number of physician visits during the first trimester

2. For the data set 'stackloss' in R, consider the multiple linear regression model of “stack loss” on the other explanatory variables

i) Investigate whether there is any multicollinearity, and suggest remedial measures if appropriate.

ii) Suppose the value of stack.loss[20] was changed from 14 to 1450, and those of Water.Temp[13] from 18 to 180, and Acid.Conc. [13] from 82 to 1.

a) Fit a multiple linear regression model on the new data.

b) Identify influential points using DFFITS, DFBETAS, Studentized Deleted Residuals and Cook’s D

c) Compare the estimates of the regression coefficients obtained before and after the above changes for each of the following:

- OLS
- Least median of squares regression
- Least trimmed squares robust regression
- M-estimates of regression with Huber weights