Chapter 4. Regression Diagnostics: Detection of Model Violations

Outline

4.1. The Standard Regression Assumptions

4.1 The Standard Regression Assumptions

Introduction

*We have stated the basic results that are used for **making inferences** about simple and multiple linear regression models in Chapters 2 and 3. The results are based on summary statistics that are computed from the data. In fitting a model to a given body of data, we would like to ensure that the fit is not **overly** determined by one or a few observations. The distribution theory, confidence intervals, and tests of hypotheses outlined in Chapters 2 and 3 are valid and have meaning **only if** the standard regression **assumptions** are satisfied. These assumptions are stated in this chapter.*

*When these assumptions are **violated**, the standard results quoted previously do not hold and an application of them may lead to **serious error**. We re-emphasize that the **prime** focus of this course is on the **detection** and correction of violations of the basic linear model assumptions as a means of achieving a thorough and informative analysis of the data. This chapter presents methods for **checking** these assumptions. We will rely **mainly** on **graphical** methods as opposed to applying **rigid** numerical rules to check for model violations.*

4.1 The Standard Regression Assumptions

Assumptions

*In the previous two chapters we have given the **least squares estimates** of the regression parameters and stated their properties. The properties of least squares estimators and the statistical analysis presented in Chapters 2 and 3 are based on the following **assumptions**:*

Assumption 1

1. **Assumptions about the form of the model**: The model that relates the response Y to the predictors $X_1, X_2, \cdots, X_p$ is assumed to be linear in the regression parameters $\beta_0, \beta_1, \cdots, \beta_p$, namely,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \tag{4.1}$$

which implies that the ith observation can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \ i = 1, 2, \cdots, n. \tag{4.2}$$

We refer to this as the *linearity* assumption. Checking the linearity assumption in simple regression is easy because the validity of this assumption can be determined by examining the scatter plot of Y versus X. A linear scatter plot ensures linearity. Checking the linearity in multiple regression is more difficult due to the high dimensionality of the data. Some graphs that can be used for checking the linearity assumption in multiple regression are given later in this chapter. When the linearity assumption does not hold, transformation of the data can sometimes lead to linearity.

4.1 The Standard Regression Assumptions

Assumptions

Assumption 2

2. **Assumptions about the errors**: The errors $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ in (4.2) are assumed to be *independently and identically distributed* (iid) normal random variables each with mean zero and a common variance σ^2. Note that this implies four assumptions:

   - The error $\varepsilon_i, i = 1, 2, \cdots, n$, has a normal distribution. We refer to this as the *normality assumption*. The normality assumption is not as easily validated especially when the values of the predictor variables are not replicated. The validity of the normality assumption can be assessed by examination of appropriate graphs of the residuals, as we describe later in this chapter.
   - The errors $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ have mean zero.
   - The errors $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ have the same (but unknown) variance σ^2. This is the *constant variance assumption*. It is also known by other names such as the *homogeneity* or the *homoscedasticity* assumption. When this assumption does not hold, the problem is called the *heterogeneity* or the *heteroscedasticity* problem.
   - The errors $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are independent of each other (their pairwise covariances are zero). We refer to this as the *independent-errors assumption*. When this assumption does not hold, we have the *autocorrelation* problem.

4.1 The Standard Regression Assumptions

Assumption 3

Assumptions

***Assumptions about the predictors**: There are **three** assumptions concerning the predictor variables:*

- The predictor variables $X_1, X_2, \cdots, X_p$ are nonrandom, that is, the values $x_{1j}, x_{2j}, \cdots, x_{nj}; j = 1, 2, \cdots, p$, are assumed fixed or selected in advance. This assumption is satisfied only when the experimenter can set the values of the predictor variables at predetermined levels. It is clear that under nonexperimental or observational situations this assumption will not be satisfied. The theoretical results that are presented in Chapters 2 and 3 will continue to hold, but their interpretation has to be modified. When the predictors are random variables, all inferences are conditional, conditioned on the observed data. It should be noted that this conditional aspect of the inference is consistent with the approach to data analysis presented in this course . Our main objective is to extract the maximum amount of information from the available data.
- The values $x_{1j}, x_{2j}, \cdots, x_{nj}; j = 1, 2, \cdots, p$, are measured without error. This assumption is hardly ever satisfied. The errors in measurement will affect the residual variance, the multiple correlation coefficient, and the individual estimates of the regression coefficients. The exact magnitude of the effects will depend on several factors, the most important of which are the standard deviation of the errors of measurement and the correlation structure among the errors. The effect of the measurement errors will be to increase the residual variance and reduce the magnitude of the observed multiple correlation coefficient. The effects of measurement errors on individual regression coefficients are more difficult to assess. The estimate of the regression coefficient for a variable is affected not only by its own measurement errors, but also by the measurement errors of other variables included in the equation.

  Correction for measurement errors on the estimated regression coefficients, even in the simplest case where all the measurement errors are uncorrelated, requires a knowledge of the ratio between the variances of the measurement errors for the variables and the variance of the random error. Since these quantities are seldom, if ever, known (particularly in the social sciences, where this problem is most acute), we can never hope to remove completely the effect of measurement errors from the estimated regression coefficients. If the measurement errors are not large compared to the random errors, the effect of measurement errors is slight. In interpreting the coefficients in such an analysis, this point should be remembered. Although there is some problem in the estimation of the regression coefficients when the variables are in error, the regression equation may still be used for prediction. However, the presence of errors in the predictors decreases the accuracy of predictions.

- The predictor variables $X_1, X_2, \cdots, X_p$ are assumed to be linearly independent of each other. This assumption is needed to guarantee the uniqueness of the least squares solution (the solution of the normal equations in Chapter 3). If this assumption is violated, the problem is referred to as the *collinearity* problem.

  so that the inverse of $\mathbf{X}'\mathbf{X}$ exists

The first two of the above assumptions about the predictors cannot be validated, so they do not play a major role in the analysis. However, they do influence the interpretation of the regression results.

4.1 The Standard Regression Assumptions

Assumptions

Assumption 4

4. **Assumptions about the observations**: All observations are equally reliable and have an approximately equal role in determining the regression results and in influencing conclusions.

A feature of the method of least squares is that small or minor violations of the underlying assumptions do not invalidate the inferences or conclusions drawn from the analysis in a major way. Gross violations of the model assumptions can, however, seriously distort conclusions. Consequently, it is important to investigate the structure of the residuals and the data pattern through graphs.

4.2. Various Types of Residuals

4.2 Various Types of Residuals

Introduction

*A **simple** and **effective** method for detecting model deficiencies in regression analysis is the examination of **residual** plots. Residual plots will point to **serious violations** in one or more of the standard assumptions when they exist. Of more importance, the analysis of residuals may lead to suggestions of structure or point to information in the data that might be **missed** or **overlooked** if the analysis is based only on **summary** statistics. These suggestions or cues can lead to a better understanding and possibly a better model of the process under study. A careful **graphical analysis** of residuals may often prove to be the most important part of the regression analysis.*

As we have seen in Chapters 2 and 3, when fitting the linear model in (4.1) to a set of data by least squares, we obtain the fitted values,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}, \quad i = 1, 2, \cdots, n, \tag{4.3}$$

and the corresponding *ordinary* least squares residuals,

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \cdots, n. \tag{4.4}$$

4.2 Various Types of Residuals

Projection Matrix

The fitted values in (4.3) can also be written in an alternative form as

$$\hat{y}_i = p_{i1}y_1 + p_{i2}y_2 + \cdots + p_{in}y_n, \quad i = 1, 2, \cdots, n, \tag{4.5}$$

where the p_{ij}'s are quantities that depend only on the values of the predictor variables (they do not involve the response variable). Equation (4.5) shows directly the relationship between the observed and predicted values. In simple regression, p_{ij} is given by

$$p_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum(x_i - \bar{x})^2}. \tag{4.6}$$

In multiple regression the p_{ij}'s are elements of a matrix known as the *hat* or *projection* matrix

Recall that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and so that $\hat{\mathbf{y}} = \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{P}}\mathbf{y}$.

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

4.2 Various Types of Residuals

Leverage

When $i = j$, p_{ii} is the ith diagonal element of the projection matrix $\mathbf{P}$. In simple regression,

$$p_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2} \,. \tag{4.7}$$

The value p_{ii} is called the *leverage* value for the ith observation because, as can be seen from (4.5), $\hat{y}_i$ is a weighted sum of all observations in Y and p_{ii} is the weight (leverage) given to y_i in determining the ith fitted value $\hat{y}_i$

Thus, we have n leverage values and they are denoted by

$$p_{11}, p_{22}, \cdots, p_{nn}. \tag{4.8}$$

The leverage values play an important role in regression analysis and we shall often encounter them.

The residual $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{P})\mathbf{y}$, then $\text{Cov}(\mathbf{e}) = \text{Cov}\big((\mathbf{I} - \mathbf{P})\mathbf{y}\big) = \text{Cov}\big((\mathbf{I} - \mathbf{P})\boldsymbol{\varepsilon}\big) = \sigma^2(\mathbf{I} - \mathbf{P})$. This implies that $\text{Var}(e_i) = \sigma^2(1 - p_{ii})$ for $i = 1, \cdots, n$

4.2 Various Types of Residuals

Standardized Residual

When the assumptions stated in Section 4.1 hold, the ordinary residuals, e_1, e_2, $\cdots$, e_n, defined in (4.4), will sum to zero, but they will not have the same variance because

$$\text{Var}(e_i) = \sigma^2(1 - p_{ii}), \tag{4.9}$$

where p_{ii} is the ith leverage value in (4.8), which depends on $x_{i1}, x_{i2}, \cdots, x_{ip}$. To overcome the problem of unequal variances, we standardize the ith residual e_i by dividing it by its standard deviation and obtain

$$z_i = \frac{e_i}{\sigma\sqrt{1 - p_{ii}}}. \tag{4.10}$$

This is called the ith *standardized residual* because it has mean zero and standard deviation 1. The standardized residuals depend on σ, the unknown standard deviation of ε. An unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n - p - 1} = \frac{\sum(y_i - \hat{y}_i)^2}{n - p - 1} = \frac{\text{SSE}}{n - p - 1}, \tag{4.11}$$

where SSE is the sum of squares of the residuals. The number $n - p - 1$ in the denominator of (4.11) is called the *degrees of freedom* (df). It is equal to the number of observations, n, minus the number of estimated regression coefficients, $p + 1$.

4.2 Various Types of Residuals

Studentized Residual

An alternative unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2_{(i)} = \frac{\text{SSE}_{(i)}}{(n-1)-p-1} = \frac{\text{SSE}_{(i)}}{n-p-2}, \tag{4.12}$$

where $\text{SSE}_{(i)}$ is the sum of squared residuals when we fit the model to the $n-1$ observations obtained by omitting the ith observation. Both $\hat{\sigma}^2$ and $\hat{\sigma}^2_{(i)}$ are unbiased estimates of σ^2.

Using $\hat{\sigma}$ as an estimate of σ in (4.10), we obtain

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-p_{ii}}}, \tag{4.13}$$

whereas using $\hat{\sigma}_{(i)}$ as an estimate of σ, we obtain

$$r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-p_{ii}}}. \tag{4.14}$$

The form of residual in (4.13) is called the *internally studentized residual*, and the residual in (4.14) is called the *externally studentized residual*, because e_i is not involved in (external to) $\hat{\sigma}_{(i)}$. For simplicity of terminology and presentation, however, we shall refer to the studentized residuals as the standardized residuals.

The standardized residuals do not sum to zero, but they all have the same variance. The externally standardized residuals follow a t-distribution with $n-p-2$ degrees of freedom, but the internally standardized residuals do not. However, with a moderately large sample, these residuals should approximately have a standard normal distribution. The residuals are not strictly independently distributed, but with a large number of observations, the lack of independence may be ignored.

$$r_i^* = r_i\sqrt{\frac{n-p-2}{n-p-1-r_i^2}},$$

They are (approximately) one-to-one related.

From now on, we only consider the internally studentized residuals and call them the standardized residuals

4.2 Various Types of Residuals

Behavior of Standardized Residuals

If the aforesaid assumptions of linear models are ALL CORRECT, we have

$r_1, \cdots, r_n$ are approximately **i.i.d.** standard normal random variables, i.e., $N(0,1)$

$r_1, \cdots, r_n$ are uncorrelated with $\mathbf{x}_1, \cdots, \mathbf{x}_n$

$r_1, \cdots, r_n$ are uncorrelated with $\hat{y}_1, \cdots, \hat{y}_n$

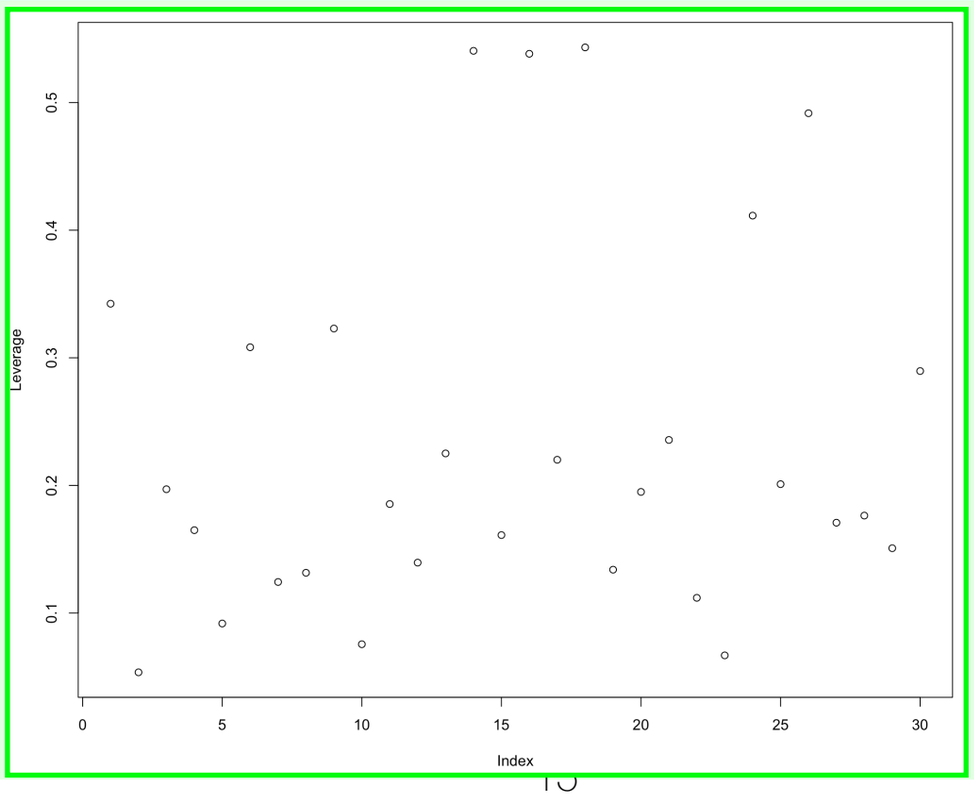
4.2 Various Types of Residuals

Example on Supervisor Dataset

```
> ##### Types of Residuals on Supervisor Dataset
> supervisor_dat<-read.table('data/P060.txt',header=TRUE)   ## read the data
> fmodel<-lm(Y~.,data=supervisor_dat)     ## fit a full model
> risd<-summary(fmodel)$residuals                  ## residuals
> risd
         1          2          3          4          5          6          7          8          9         10         11         12         13         14         15
 -8.1102953  1.6472337  1.0605589 -0.2268416  6.5462010 -10.9418499 -9.1484140  0.9029929 -7.5309862  7.8015424  6.0742817 11.5989723  9.4183197 -2.2140147  0.4506705
        16         17         18         19         20         21         22         23         24         25         26         27         28         29         30
 -3.5478519 -2.1501319  3.6026355 -3.0165587 -5.6201442  7.3967582  0.1809831 -10.6639999 -4.6247464  5.6828983 -1.8434727  2.8596385 -8.0453540  7.3394730  5.1215016
```

```
> X<-as.matrix(supervisor_dat[,-1])
> X<-cbind(rep(1,30),X)
> hat_mat<-X %*% solve(t(X) %*% X) %*% t(X)      ## the hat matrix P whose diagonals are the leverage values
> leverages <- diag(hat_mat)                     ## get the leverage values
> leverages
 [1] 0.34237207 0.05351803 0.19700315 0.16492711 0.09177912 0.30826724 0.12431711 0.13153467 0.32292639 0.07553199 0.18538862 0.13948385 0.22505820 0.54053420 0.16107694 0.53823675
[17] 0.22007830 0.54322942 0.13396133 0.19489768 0.23562337 0.11189120 0.06680849 0.41139884 0.20097670 0.49165320 0.17078409 0.17634399 0.15078197 0.28961597
```

```
plot(seq(1,30),leverages,xlab="Index",ylab="Leverage",lty=p)      ## scatter plot of the leverages
```



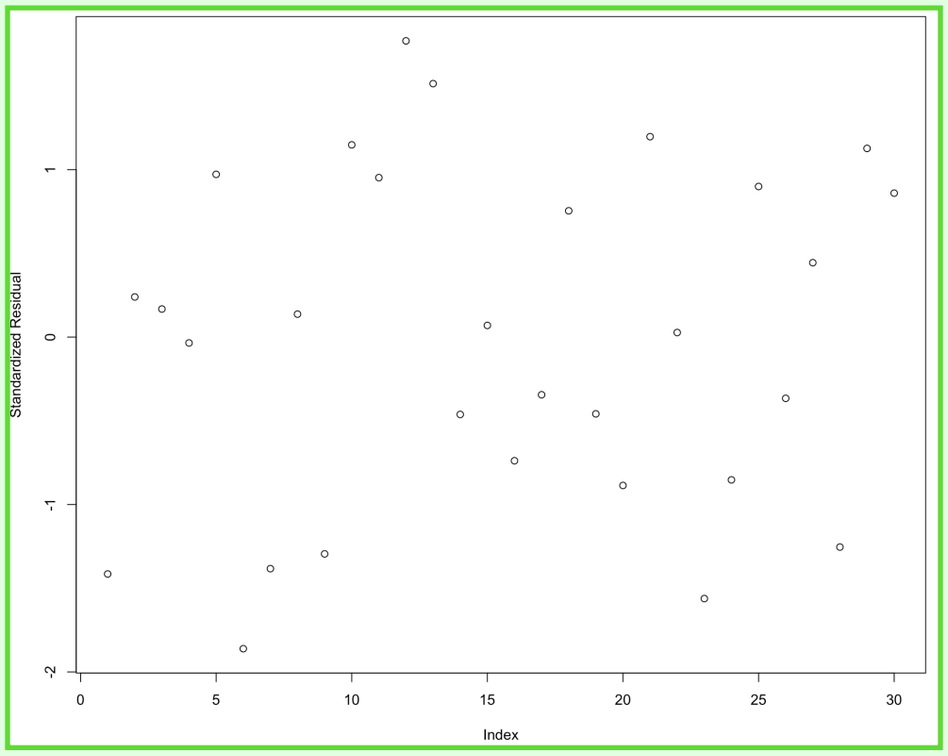
4.2 Various Types of Residuals

Example on Supervisor Dataset

```
> int_risd<-risd/(summary(fmodel)$sigma * sqrt(1-leverages))       ## internally studentized residual j
> int_risd
         1          2          3          4          5          6          7          8          9         10         11         12         13         14         15
-1.41498026  0.23955370  0.16744867 -0.03512080  0.97184596 -1.86133876 -1.38317210  0.13709194 -1.29490454  1.14799070  0.95218982  1.76906521  1.51371017 -0.46212316  0.06961486
        16         17         18         19         20         21         22         23         24         25         26         27         28         29         30
-0.73868563 -0.34446368  0.75418016 -0.45861365 -0.88618779  1.19699287  0.02717120 -1.56184734 -0.85286680  0.89948517 -0.36581416  0.44430497 -1.25422677  1.12683185  0.85971512
```

```
> #### obtain external studentized residual by the relation between int_risd and ext_risd, see the formula on lecture slides
> n<-30
> p<-6
> ext_risd <- int_risd * sqrt((n-p-2)/(n-p-1-int_risd^2))
> ext_risd
         1          2          3          4          5          6          7          8          9         10         11         12         13         14         15
-1.44835328  0.23458097  0.16386794 -0.03434974  0.97062209 -1.97526518 -1.41280382  0.13413337 -1.31529351  1.15637546  0.95017640  1.86145176  1.56019127 -0.45407837  0.06809185
        16         17         18         19         20         21         22         23         24         25         26         27         28         29         30
-0.73117411 -0.33776450  0.74689589 -0.45059801 -0.88189556  1.20894332  0.02657438 -1.61559196 -0.84763116  0.89560731 -0.35881868  0.43641573 -1.27088888  1.13380428  0.85466249
```

```
plot(seq(1,30),int_risd,xlab="Index",ylab="Standardized Residual",lty=p)      ## scatter plot of the standardized residuals
```



4.2 Various Types of Residuals

Example on Supervisor Dataset

```
> rstandard(fmodel)          ## obtain the standardized residual, i.e., the internal studentized residuals by a built-in function
         1          2          3          4          5          6          7          8          9         10         11         12         13         14         15
-1.41498026  0.23955370  0.16744867 -0.03512080  0.97184596 -1.86133876 -1.38317210  0.13709194 -1.29490454  1.14799070  0.95218982  1.76906521  1.51371017 -0.46212316  0.06961486
        16         17         18         19         20         21         22         23         24         25         26         27         28         29         30
-0.73868563 -0.34446368  0.75418016 -0.45861365 -0.88618779  1.19699287  0.02717120 -1.56184734 -0.85286680  0.89948517 -0.36581416  0.44430497 -1.25422677  1.12683185  0.85971512
```

```
> as.data.frame(hatvalues(fmodel))   ## obtain the leverages using a built-in function
   hatvalues(fmodel)
1          0.34237207
2          0.05351803
3          0.19700315
4          0.16492711
5          0.09177912
6          0.30826724
7          0.12431711
8          0.13153467
9          0.32292639
10         0.07553199
11         0.18538862
12         0.13948385
13         0.22505820
14         0.54053420
15         0.16107694
16         0.53823675
17         0.22007830
18         0.54322942
19         0.13396133
20         0.19489768
21         0.23562337
22         0.11189120
23         0.06680849
24         0.41139884
25         0.20097670
26         0.49165320
27         0.17078409
28         0.17634399
29         0.15078197
30         0.28961597
```

4.3. Graphical Methods: Before and After Fitting a Model

4.3 Graphical Methods: Before and After Fitting a Model

Introduction

Graphical *methods play an important role in data analysis. It is of particular* ***importance*** *in fitting linear models to data. As Chambers et al. (1983, p. 1) put it, "There is no single statistical tool that is as powerful as a well-chosen graph." Graphical methods can be regarded as* ***exploratory*** *tools. They are also an integral part of* ***confirmatory*** *analysis or* ***statistical*** *inference.*

Huber (1991, p. 121) says, "Eye-balling can give diagnostic insights no formal diagnostics will ever provide." One of the best examples that illustrates this is the Anscombe quartet, the four data sets given in Chapter 2 (Table 2.4). The four data sets are constructed by Anscombe in such a way that all pairs (Y, X) have ***identical*** *values of* ***descriptive*** *statistics (same correlation coefficients, same regression lines, same standard errors, etc.), yet their* ***pairwise scatter plots*** *(reproduced in Figure 4.1 for convenience) give completely different scatters.*

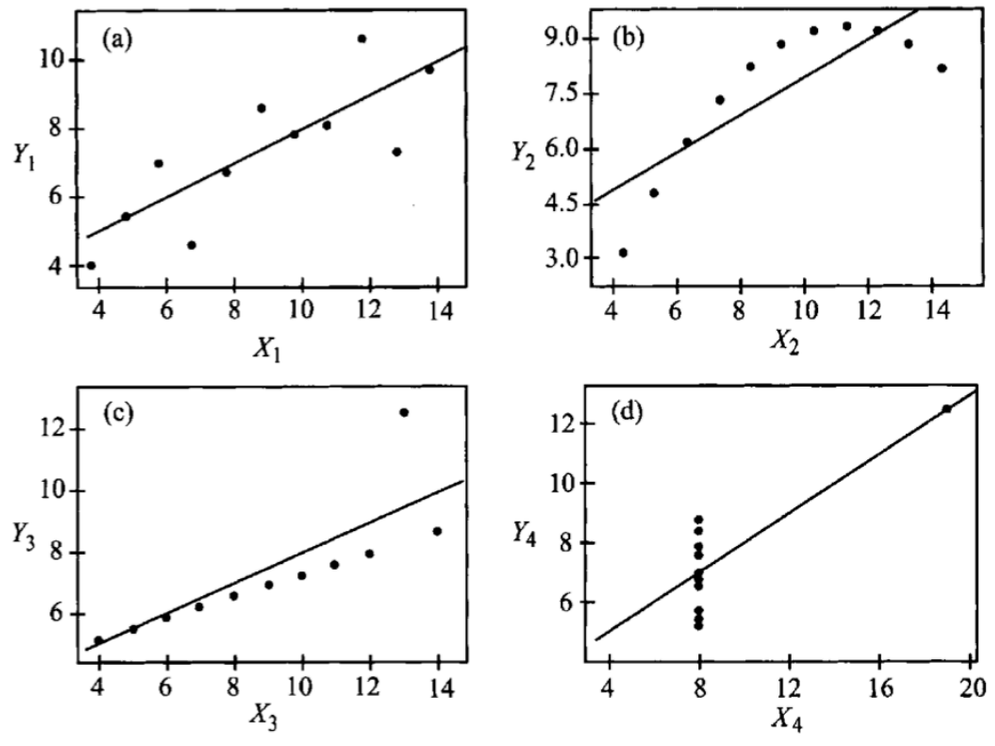


Figure 4.1 Plot of the data (X, Y) with the least squares fitted line for the Anscombe quartet.

More explanation on next slide

4.3 Graphical Methods: Before and After Fitting a Model

Introduction

The scatter plot in Figure 4.1(a) indicates that a linear model may be reasonable, whereas the one in Figure 4.1(b) suggests a (possibly linearizable) nonlinear model. Figure 4.1(c) shows that the data follow a linear model closely except for one point which is clearly off the line. This point may be an outlier, hence it should be examined before conclusions can be drawn from the data. Figure 4.1(d) indicates either a deficient experimental design or a bad sample. For the point at $X = 19$, the reader can verify that (a) the residual at this point is always zero (with a variance of zero) no matter how large or small its corresponding value of Y and (b) if the point is removed, the least squares estimates based on the remaining points are no longer unique (except the vertical line, any line that passes through the average of the remaining points is a least squares line!). Observations which unduly influence regression results are called *influential observations*. The point at $X = 19$ is therefore extremely influential because it alone determines both the intercept and the slope of the fitted line.

4.3 Graphical Methods: Before and After Fitting a Model

Introduction

We have used the scatter plot here as an exploratory tool, but one can also use graphical methods to complement numerical methods in a confirmatory analysis. Suppose we wish to test whether there is a positive correlation between Y and X or, equivalently, if Y and X can be fitted by a positively sloped regression line. The reader can verify that the correlation coefficients are the same in all four data sets [$\text{Cor}(Y, X) = 0.80$] and all four data sets also have the same regression line ($Y = 3 + 0.5\ X$) with the same standard errors of the coefficients. Thus, based on these numerical summaries, one would reach the erroneous conclusion that all four data sets can be described by the same model. The underlying assumption here is that the relationship between Y and X is linear and this assumption does not hold here, for example, for the data set in Figure 4.1(b). Hence the test is invalid. The test for linear relationship, like other statistical methods, is based on certain underlying assumptions. Thus conclusions based on these methods are valid only when the underlying assumptions hold. It is clear from the above example that if analyses were solely based on numerical results, wrong conclusions will be reached.

4.3 Graphical Methods: Before and After Fitting a Model

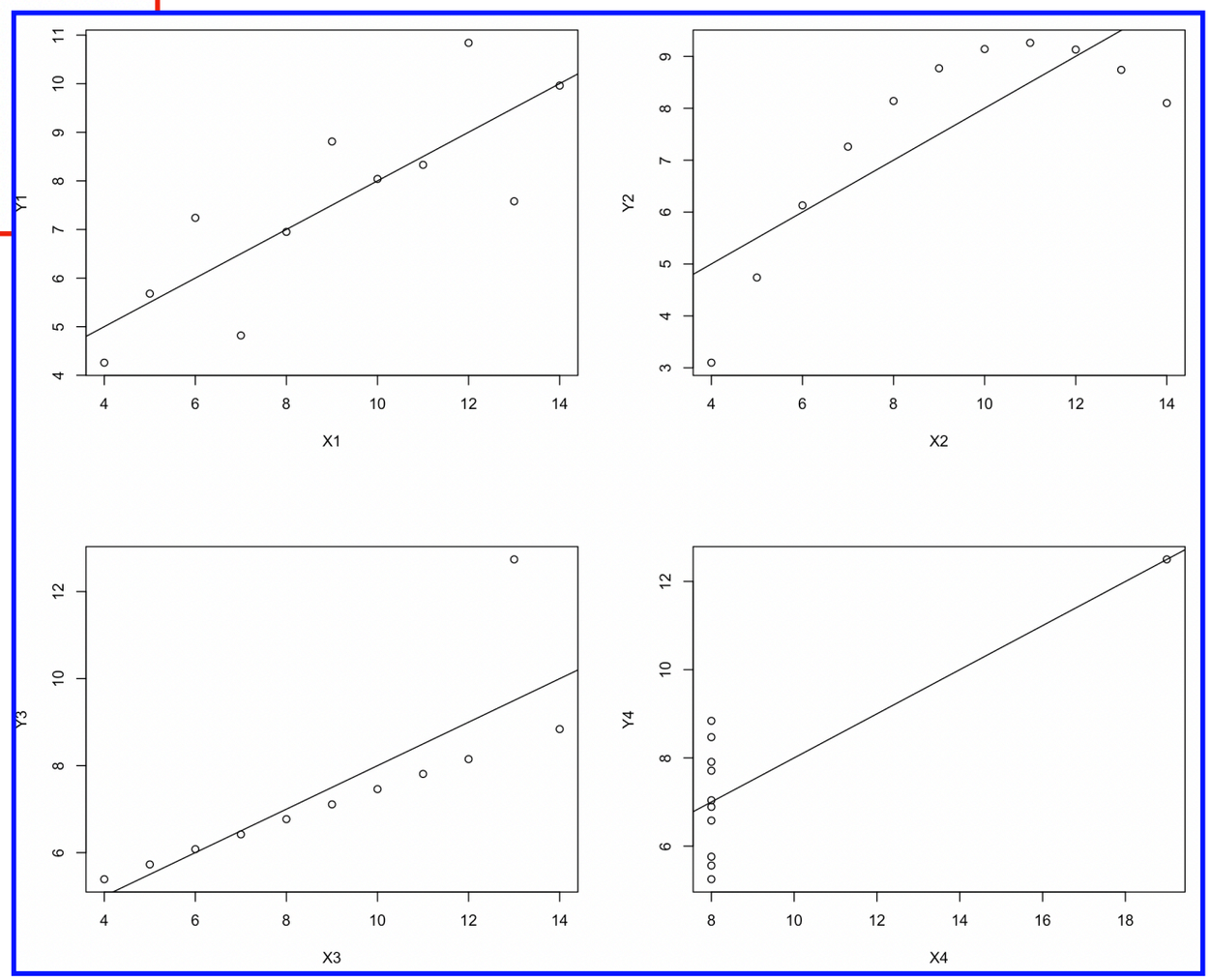
Use R to plot Anscombe Dataset

```
###### plots of Anscombe Quartet Dataset
anscombe<-read.table('data/P029b.txt',header=TRUE)   ## read the data
md1<-lm(Y1~X1,data=anscombe)
par(mfrow=c(2,2))
plot(anscombe$X1,anscombe$Y1,xlab="X1",ylab="Y1")
abline(md1)

md2<-lm(Y2~X2,data=anscombe)
plot(anscombe$X2,anscombe$Y2,xlab="X2",ylab="Y2")
abline(md2)

md3<-lm(Y3~X3,data=anscombe)
plot(anscombe$X3,anscombe$Y3,xlab="X3",ylab="Y3")
abline(md3)

md4<-lm(Y4~X4,data=anscombe)
plot(anscombe$X4,anscombe$Y4,xlab="X4",ylab="Y4")
abline(md4)
```



4.3 Graphical Methods: Before and After Fitting a Model

Introduction

Graphical methods can be useful in many ways. They can be used to:

1. Detect errors in the data (e.g., an outlying point may be a result of a typographical error)
2. Recognize patterns in the data (e.g., clusters, outliers, gaps, etc.)
3. Explore relationships among variables
4. Discover new phenomena
5. Confirm or negate assumptions
6. Assess the adequacy of a fitted model
7. Suggest remedial actions (e.g., transform the data, redesign the experiment, collect more data, etc.)
8. Enhance numerical analyses in general

This chapter presents some graphical displays useful in regression analysis. The graphical displays we discuss here can be classified into two (not mutually exclusive) classes:

- Graphs before fitting a model. These are useful, for example, in correcting errors in data and in selecting a model.
- Graphs after fitting a model. These are particularly useful for checking the assumptions and for assessing the goodness of the fit.

4.3 Graphical Methods: Before and After Fitting a Model

Graphs Before Fitting a Model

*The form of a model that represents the **relationship** between the response and predictor variables should be based on the theoretical background or the hypothesis to be tested. But if no prior information about the form of the model is available, the data may be used to **suggest** the model. The data should be examined thoroughly before a model is fitted. The graphs that one examines **before** fitting a model to the data serve as **exploratory** tools. Three possible groups of graphs are*

1. One-dimensional graphs
2. Two-dimensional graphs
3. Rotating plots

4.3 Graphical Methods: Before and After Fitting a Model

Graphs Before Fitting a Model

One-Dimensional Graphs

Data analysis usually begins with the examination of each variable in the study. The purpose is to have a general idea about the distribution of each individual variable. One of the following graphs may be used for examining a variable:

- Histogram
- Dot plot
- Box plot

The one-dimensional graphs serve two major functions. They indicate the distribution of a particular variable, whether the variable is symmetric or skewed. When a variable is very skewed, it should be transformed. For a highly skewed variable a logarithmic transformation is recommended. Univariate graphs provide guidance on the question as to whether one should work with the original or with the transformed variables.

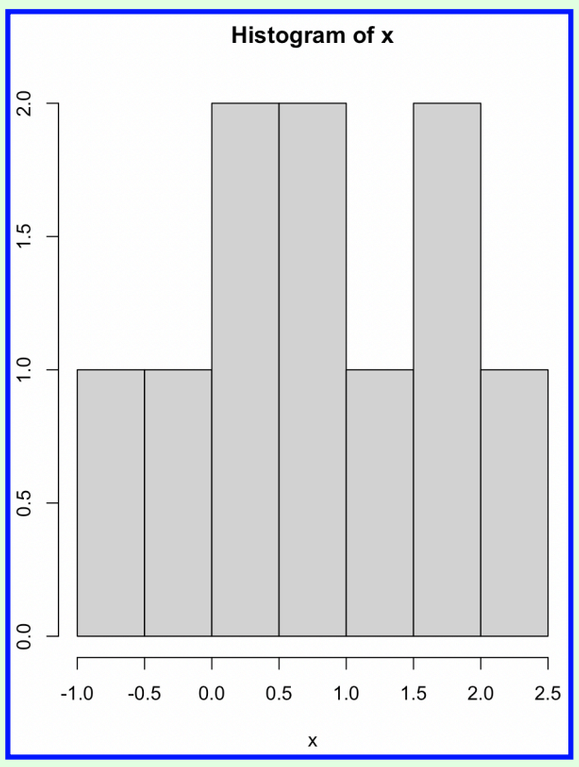
Univariate graphs also point out the presence of outliers in the variables. Outliers should be checked to see if they are due to transcription errors. No observation should be deleted at this stage. They should be noted as they may show up as troublesome points later.

4.3 Graphical Methods: Before and After Fitting a Model

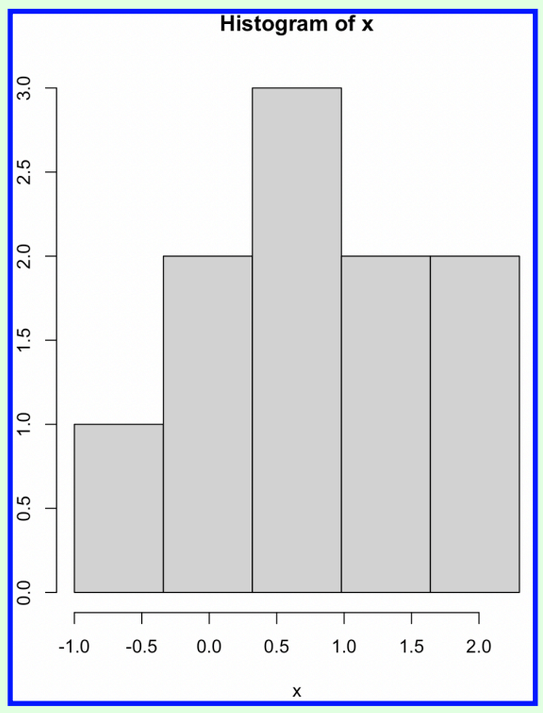
Graphs Before Fitting a Model

```
###################### histogram, dot plot and box plot
x<-c(970,612,1201,1003,666,1088,744,898,964,1135,983,1016,1029, 1058, 1085, 1122, 1022, 623, 1197, 883)    ## create the data
```

```
> ## histogram
> hist(x)
```




```
> hist(x,breaks = seq(min(x), max(x), length.out = 6))
```

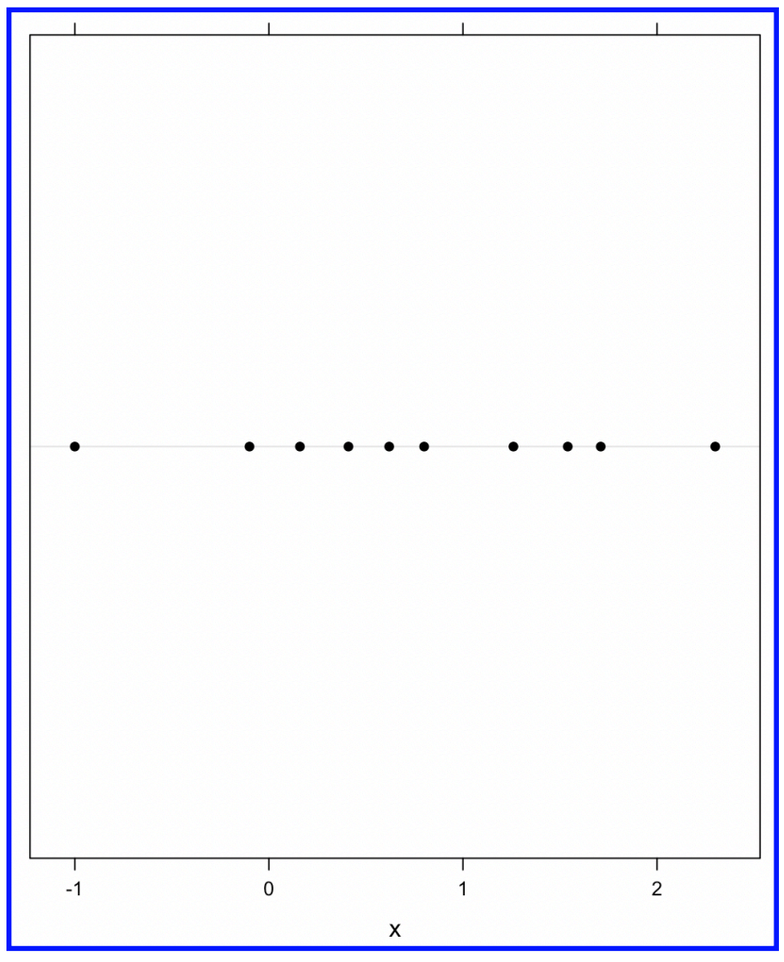


4.3 Graphical Methods: Before and After Fitting a Model

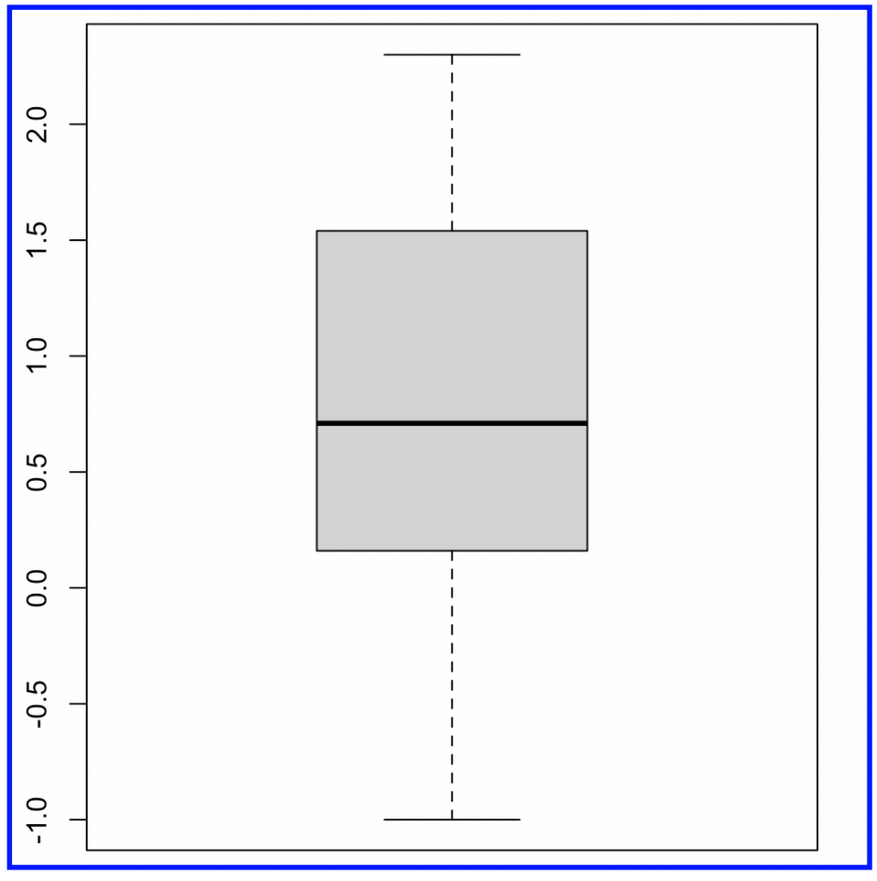
Graphs Before Fitting a Model

```
###################### histogram, dot plot and box plot
x<-c(970,612,1201,1003,666,1088,744,898,964,1135,983,1016,1029, 1058, 1085, 1122, 1022, 623, 1197, 883)    ## create the data
```

```
### dotplot
dotplot(x)
```




```
### boxplot
boxplot(x)
```



4.3 Graphical Methods: Before and After Fitting a Model

Graphs Before Fitting a Model

Two-Dimensional Graphs

*Ideally, when we have **multidimensional** data, we should examine a graph of the same dimension as that of the data. Obviously, this is feasible only when the number of variables is small. However, we can take the variables in pairs and look at the scatter plots of each variable versus each other variable in the data set. The purposes of these **pairwise scatter plots** are to explore the relationships between each pair of variables and to identify general patterns.*

*When the number of variables is **small**, it may be possible to arrange these **pairwise** scatter plots in a **matrix** format, sometimes referred to as the **draftsman's plot** or the **plot matrix**. Figure 4.2 is an example of a plot matrix for one response and two predictor variables. The pairwise scatter plots are given in the **upper triangular part** of the plot matrix. We can also arrange the corresponding correlation coefficients in a matrix. The corresponding correlation coefficients are given in the lower triangular part of the plot matrix. These arrangements facilitate the examination of the plots. The pairwise correlation coefficients should always be interpreted in conjunction with the corresponding scatter plots. The reason for this is **twofold**: (a) the correlation coefficient measures only linear relationships, and (b) the correlation coefficient is **nonrobust**, that is, its value can be **substantially** influenced by one or two observations in the data.*

Figure 4.2

4.3 Graphical Methods: Before and After Fitting a Model

Graphs Before Fitting a Model

Two-Dimensional Graphs

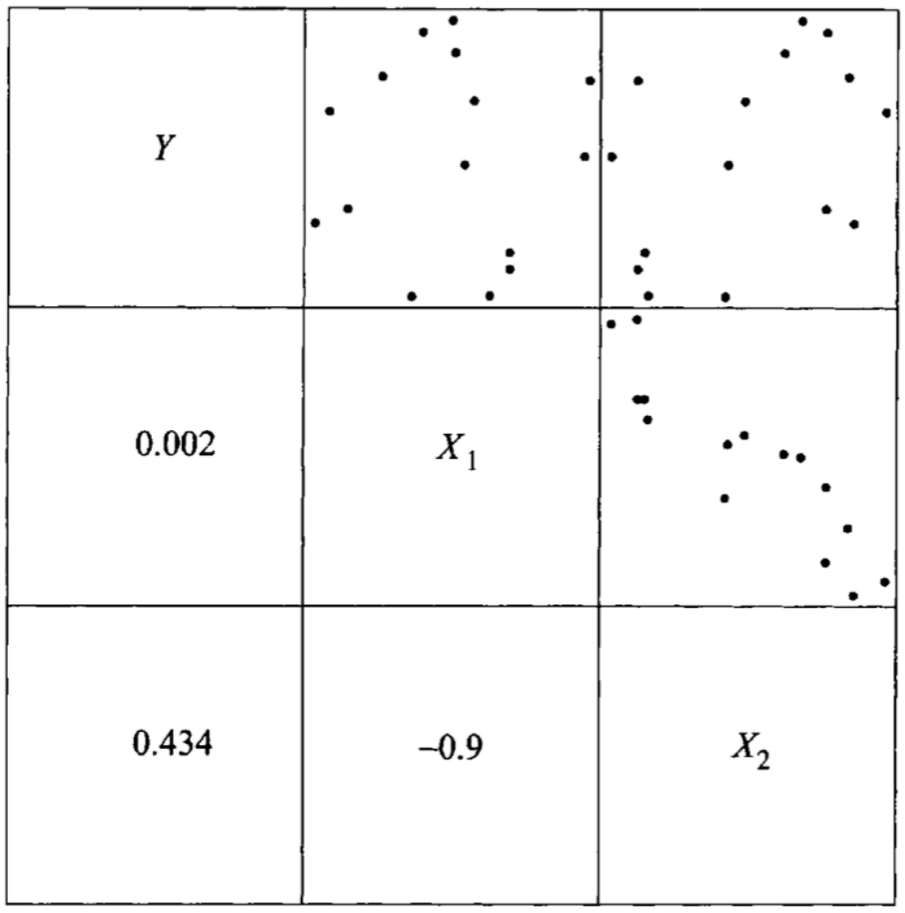


Figure 4.2 Plot matrix for Hamilton's data with the pairwise correlation coefficients.

*Ideally, What do we expect each of the graphs in the plot matrix to look like? In simple regression, the plot of Y versus X is expected to show a **linear** pattern. In multiple regression, however, the scatter plots of Y versus each predictor variable **may or may not** show linear patterns. Where the presence of a linear pattern is reassuring, the absence of such a pattern does not imply that our linear model is **incorrect**. An example is given in the next slide.*

Example

4.3 Graphical Methods: Before and After Fitting a Model

Two-Dimensional Graphs

Graphs Before Fitting a Model

Example: Hamilton's Data

Hamilton (1987) generates sets of data in such a way that Y depends on the predictor variables collectively but not individually. One such data set is given in Table 4.1. It can be seen from the plot matrix of this data (Figure 4.2) that no linear relationships exist in the plot of Y versus X_1 ($R^2 = 0$) and Y versus X_2 ($R^2 = 0.19$). Yet, when Y is regressed on X_1 and X_2 simultaneously, we obtain an almost perfect fit. The reader can verify that the following fitted equations are obtained:

$$\hat{Y} = 11.989 + 0.004X_1; \quad t\text{-Test} = 0.009; \quad R^2 = 0.0,$$
$$\hat{Y} = 10.632 + 0.195X_2; \quad t\text{-Test} = 1.74; \quad R^2 = 0.188,$$
$$\hat{Y} = -4.515 + 3.097X_1 + 1.032X_2; \quad F\text{-Test} = 39222; \quad R^2 = 1.0.$$

The first two equations indicate that Y is related to neither X_1 nor X_2 individually, yet X_1 and X_2 predict Y almost perfectly. Incidentally, the first equation produces a negative value for the adjusted R^2, $R_a^2 = -0.08$.

The scatter plots that should look linear in the plot matrix are the plots of Y versus each predictor variable after adjusting for all other predictor variables (i.e., taking the linear effects of all other predictor variables out).

The pairwise scatter plot of the predictors should show no linear pattern (ideally, we should see no discernible pattern, linear or otherwise) because the predictors are assumed to be linearly independent. In Hamilton's data, this assumption does not hold because there is a clear linear pattern in the scatter plot of X_1 versus X_2 (Figure 4.2). We should caution here that the absence of linear relationships in these scatter plots does not imply that the entire set of predictors are linearly independent. The linear relationship may involve more than two predictor variables. Pairwise scatter plots will fail to detect such a multivariate relationship.

Table 4.1 Hamilton's (1987) Data

Y	X_1	X_2	Y	X_1	X_2
12.37	2.23	9.66	12.86	3.04	7.71
12.66	2.57	8.94	10.84	3.26	5.11
12.00	3.87	4.40	11.20	3.39	5.05
11.93	3.10	6.64	11.56	2.35	8.51
11.06	3.39	4.91	10.83	2.76	6.59
13.03	2.83	8.52	12.63	3.90	4.90
13.13	3.02	8.04	12.46	3.16	6.96
11.44	2.14	9.05			

Figure 4.2

4.3 Graphical Methods: Before and After Fitting a Model

Graphs Before Fitting a Model

Rotating Plots

Recent advances in computer hardware and software have made it possible to plot data of three or more dimensions. The simplest of these plots is the three-dimensional rotating plot. The rotating plot is a scatter plot of three variables in which the points can be rotated in various directions so that the three-dimensional structure becomes apparent. Describing rotating plots in words does not do them justice. The real power of rotation can be felt only when one watches a rotating plot in motion on a computer screen. The motion can be stopped when one sees an interesting view of the data. For example, in the Hamilton data we have seen that X_1 and X_2 predict Y almost perfectly. This finding is confirmed in the rotating plot of Y against X_1 and X_2. When this plot is rotated, the points fall on an almost perfect plane. The plot is rotated until an interesting direction is found. Figure 4.3 shows one such direction, where the plane is viewed from an angle that makes the scatter of points seem to fall on a straight line.

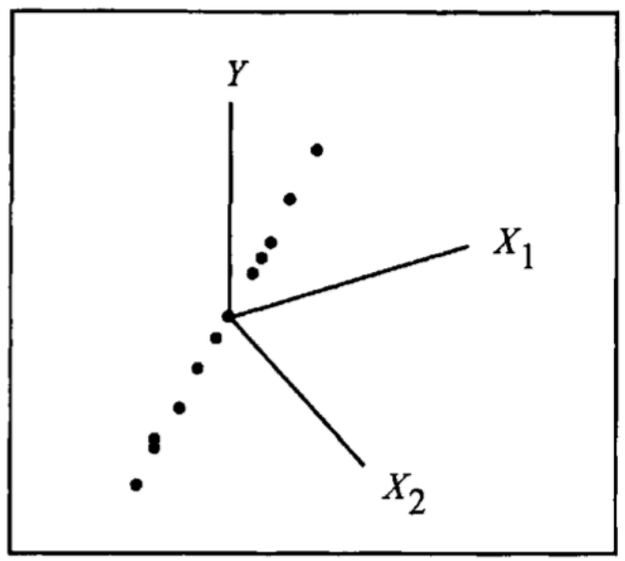


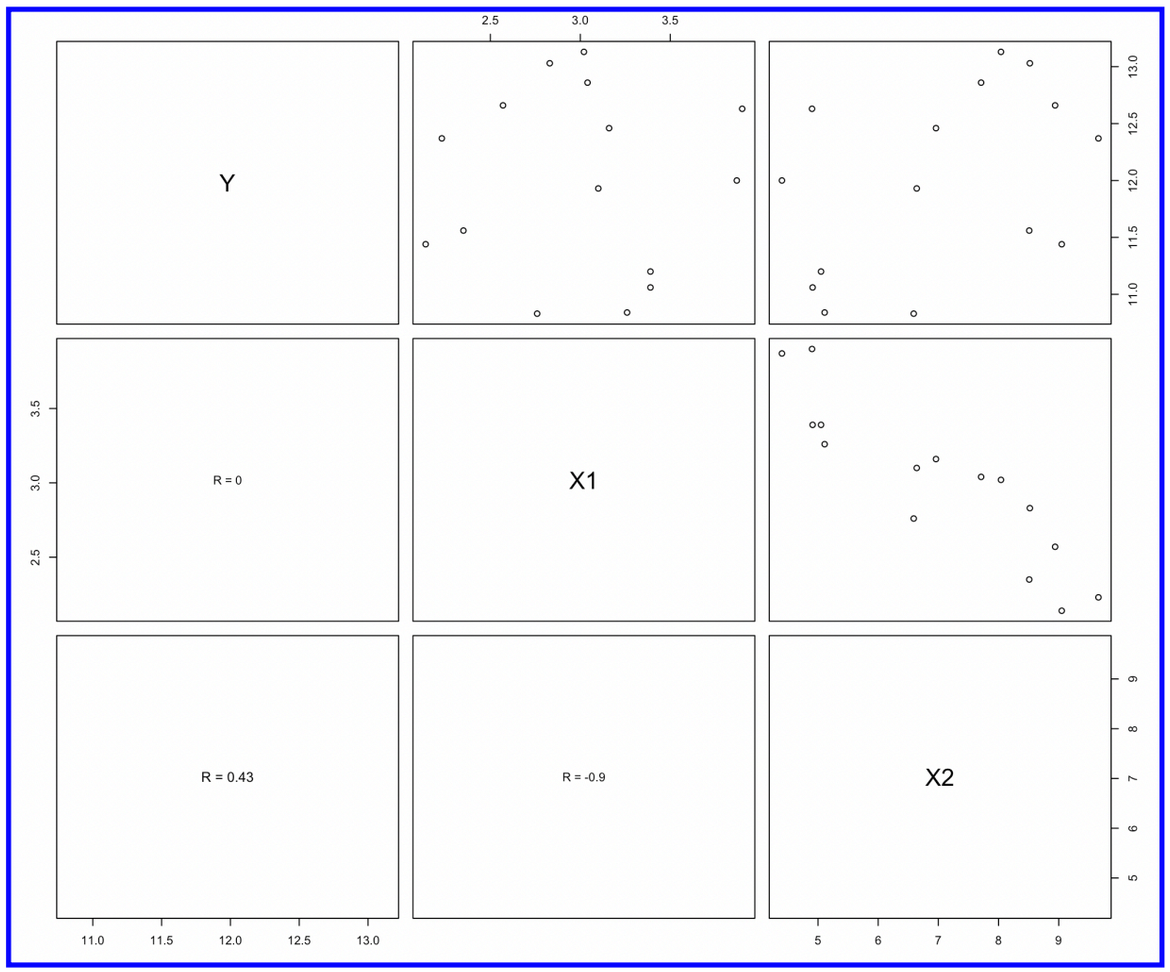
Figure 4.3 Rotating plot for Hamilton's data.

4.3 Graphical Methods: Before and After Fitting a Model

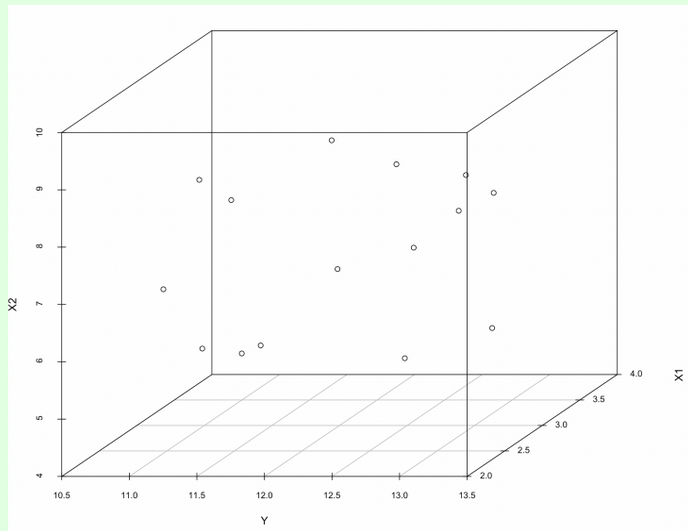
Use R on the Hamilton's Dataset

```
###### On Hamilton's Dataset
hamilton_dat<-read.table('data/P103.txt',header=TRUE)   ## read the data

# Correlation panel
panel.cor <- function(x, y){
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- round(cor(x, y), digits=2)
  txt <- paste0("R = ", r)
  cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}
pairs(hamilton_dat, lower.panel = panel.cor)          ## scatterplot matrix
```




```
par(mfrow=c(1,1))
library(scatterplot3d)
scatterplot3d(hamilton_dat)
```




```
#####  Rotatable 3d plot
library(rgl)
Y<-hamilton_dat$Y
X1<-hamilton_dat$X1
X2<-hamilton_dat$X2
plot3d(Y,X1,X2)
```

4.3 Graphical Methods: Before and After Fitting a Model

Graphs After Fitting a Model

*The graphs presented in the previous section are useful in **data checking** and the model **formulation** steps. The graphs after fitting a model to the data help in checking the **assumptions** and in assessing the **adequacy** of the fit of a given model. These graphs can be grouped into the following classes:*

1. Graphs for checking the linearity and normality assumptions
2. Graphs for the detection of outliers and influential observations
3. Diagnostic plots for the effect of variables

All these methods will be explained in the Section 4.4, 4.5 and 4.6.

4.4. Checking Linearity and Normality Assumptions

4.4 Checking Linearity and Normality Assumptions

Graphs After Fitting a Model

Residual versus Normal Score

*When the number of variables is **small**, the assumption of linearity can be checked by interactively and dynamically manipulating the plots discussed in the previous section. The task of checking the linearity assumption becomes **difficult** when the number of variables is **large**. However, one can check the linearity and normality assumptions by examining the **residuals** after fitting a given model to the data.*

*The following plots of the standardized **residuals** can be used to check the linearity and normality assumptions:*

*1. **Normal probability plot of the standardized residuals**: This is a plot of the ordered standardized residuals versus the so-called **normal scores**. The normal scores are what we would expect to obtain if we take a sample of size n from a standard **normal** distribution. If the residuals are normally distributed, the **ordered** residuals should be approximately the same as the **ordered** normal scores. Under normality assumption, this plot should resemble a (nearly) **straight line** with an intercept of zero and a slope of one (these are the mean and the standard deviation of the standardized residuals, respectively).*

More on next slide

4.4 Checking Linearity and Normality Assumptions

Graphs After Fitting a Model

After fitting the model, the standardized residuals (internally studentized residuals) are

Sample Quantiles

$$r_1, r_2, \cdots, r_n$$

Sort them from smallest to largest, called *order statistics*, and write them as

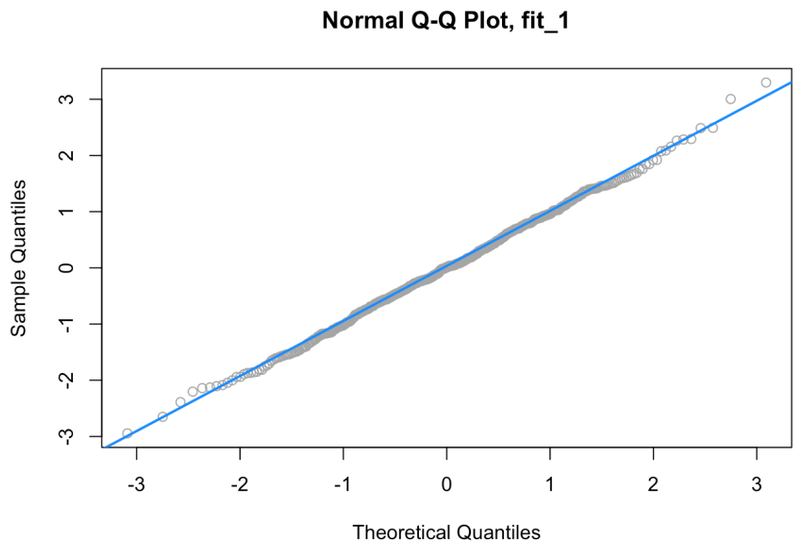
$$r_{(1)}, r_{(2)}, \cdots, r_{(n)}$$

Theoretical Quantiles

The normal scores are the z-score values $z_{(j-0.5)/n}$ for $j = 1, \cdots, n$. Basically, $z_{(j-0.5)/n}$ is the value that $\Phi(z_{(j-0.5)/n}) = \frac{j-0.5}{n}$ where $\Phi(\cdot)$ is the c.d.f. of $N(0,1)$.

The **normal Q-Q plot** (quantile-quantile plot) is the scatter plot of $(r_{(j)}, z_{(j-0.5)/n}), j = 1, \cdots, n$.

Namely the Q-Q plot but with standardised residuals.
(See Chapter 2 for Q-Q plot)



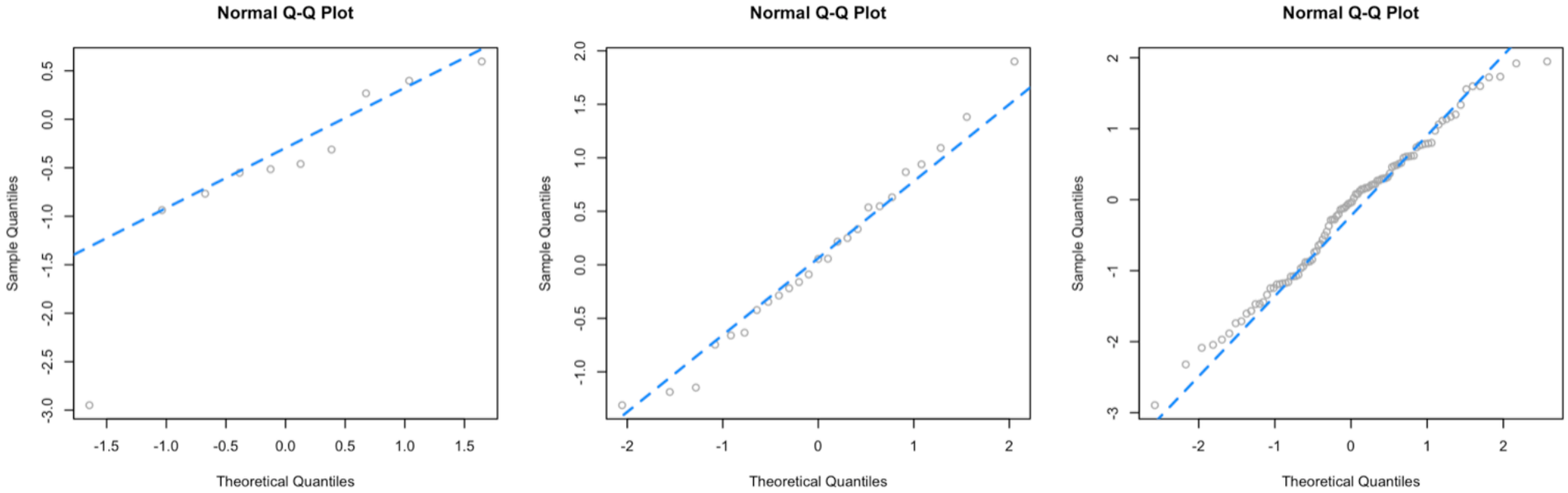
More on next slide

4.4 Checking Linearity and Normality Assumptions

Graphs After Fitting a Model

Residual versus Normal Score

In these three examples, the data **are** sampled from normal distributions. But the sample sizes are from small to large. This examples shows that the Q-Q plot sometimes is not stable when sample size is small.



Since this data **is** sampled from a normal distribution, these are all, by definition, good Q-Q plots. The points are “close to the line” and we would conclude that this data could have been sampled from a normal distribution. Notice in the first plot, one point is *somewhat* far from the line, but just one point, in combination with the small sample size, is not enough to make us worried. We see with the large sample size, all of the points are rather close to the line.

4.4 Checking Linearity and Normality Assumptions

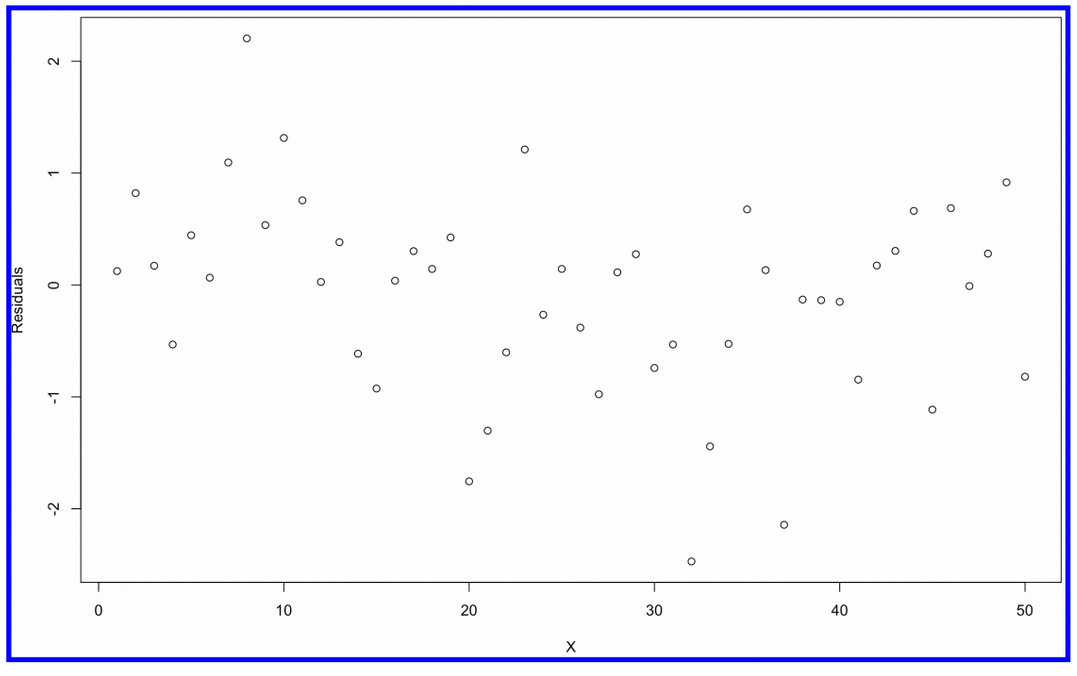
Graphs After Fitting a Model

Residual versus Predictor Variables

*2. **Scatter plots of the standardized residual against each of the predictor variables**: Under the standard assumptions, the standardized residuals are **uncorrelated** with each of the predictor variables. If the assumptions hold, this plot should be a random scatter of points. Any **discernible** pattern in this plot may indicate **violation** of some assumptions. If the linearity assumption does not hold, one may observe a plot like the one given in Figure 4.4(a). In this case a transformation of the Y and/or the particular predictor variable may be **necessary** to achieve linearity. A plot that looks like Figure 4.4(b) may indicate **heterogeneity of variance**. In this case a transformation of the data that stabilizes the variance may be needed.*

scatter plot of $(x_i, r_i), i = 1, \cdots, n$

good



bad

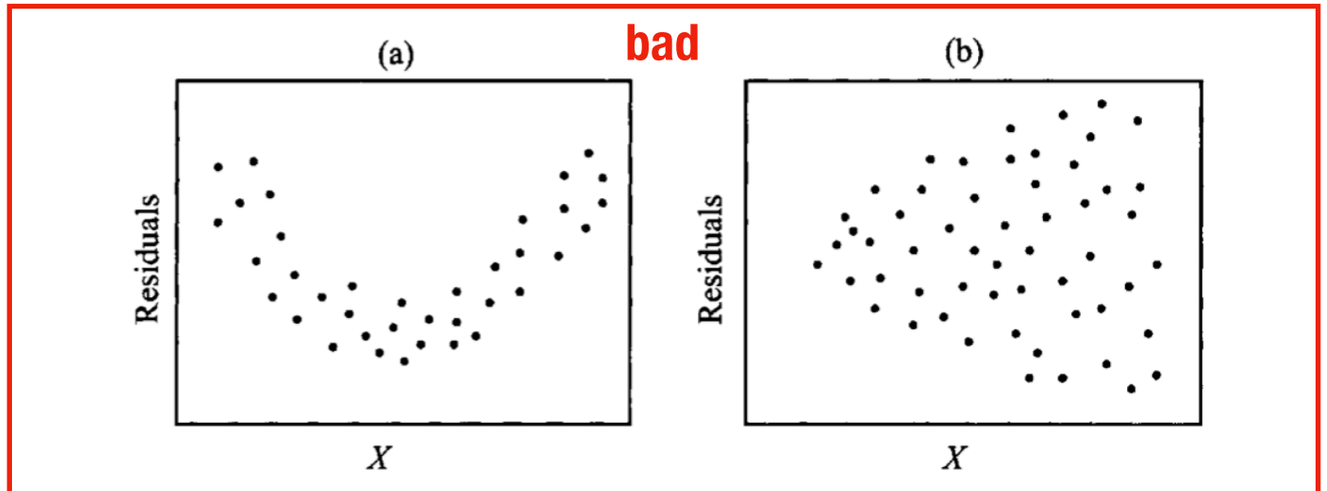


Figure 4.4 Two scatter plots of residuals versus X illustrating violations of model assumptions: (a) a pattern indicating nonlinearity and (b) a pattern indicating heterogeneity.

More on next slide

4.4 Checking Linearity and Normality Assumptions

Graphs After Fitting a Model

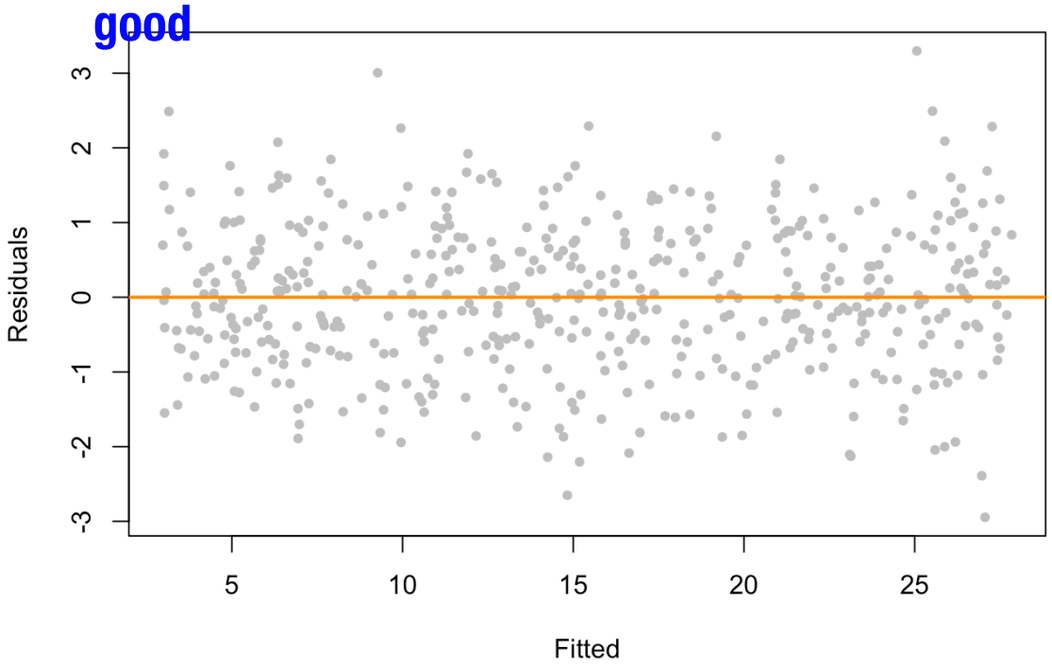
Residual versus fitted value

$$\text{scatter plot of } (\hat{y}_i, r_i), i = 1, \cdots, n$$

3. ***Scatter plot of the standardized residual versus the fitted values****: Under the standard assumptions, the* ***standardized*** *residuals are also uncorrelated with the fitted values; therefore, this plot should also be a* ***random*** *scatter of points. In simple regression,the plots of standardized residuals against X and against the fitted values are identical.*

More on next slide

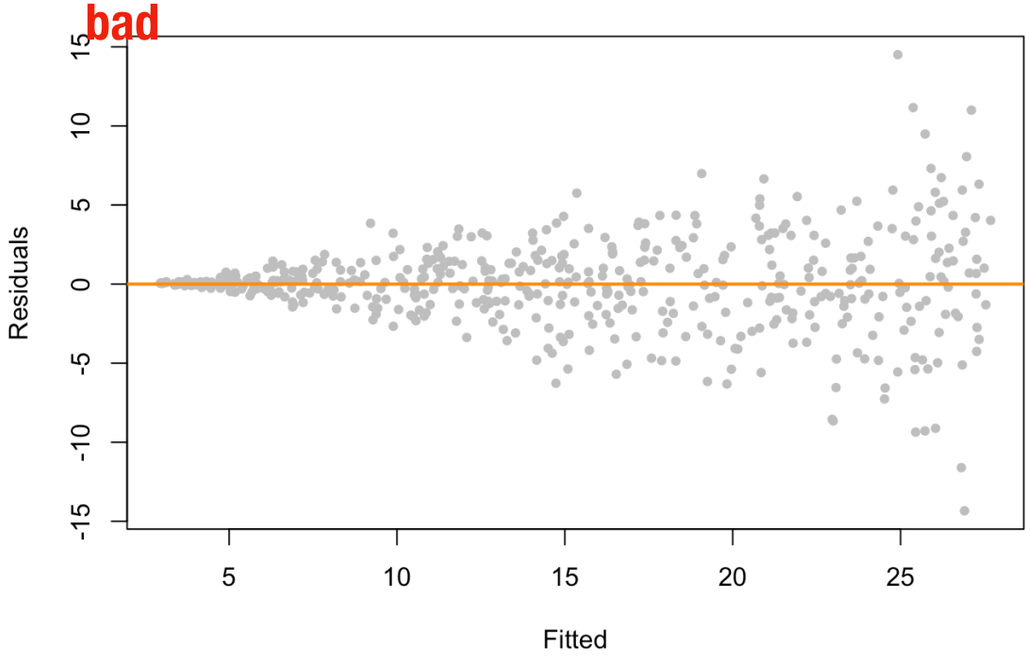
good



At any fitted value, the mean of the residuals should be roughly 0. If this is the case, the linearity assumption is valid. For this reason, we generally add a horizontal line at y=0 to emphasize this point.

At every fitted value, the spread of the residuals should be roughly the same. If this is the case, the constant variance assumption is valid.

bad

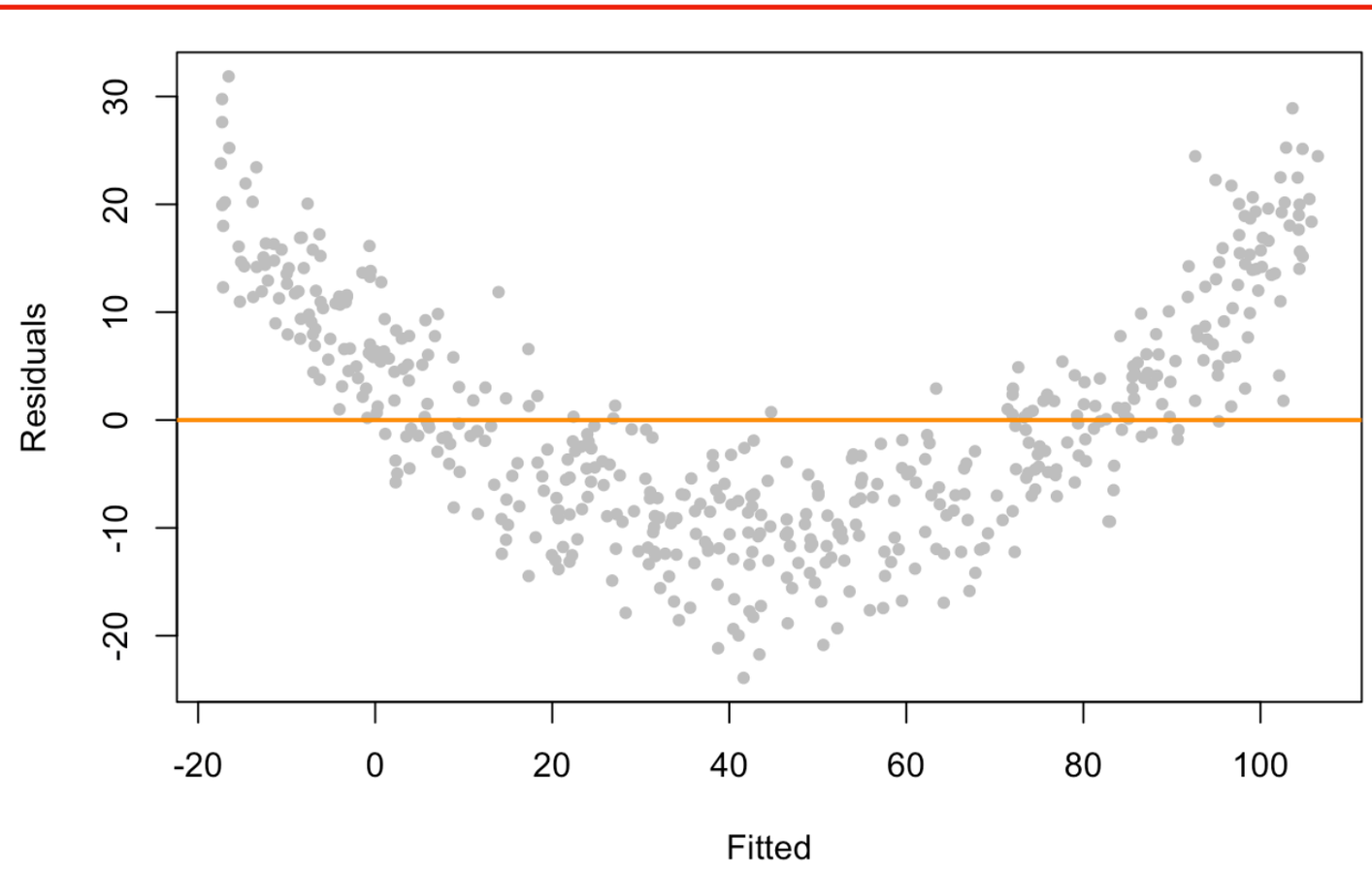


For any fitted value, the residuals seem roughly centered at 0. This is good! The linearity assumption is not violated. However, we also see very clearly, that for larger fitted values, the spread of the residuals is larger. This is bad! The constant variance assumption is violated here.

4.4 Checking Linearity and Normality Assumptions

Graphs After Fitting a Model

Residual versus fitted value



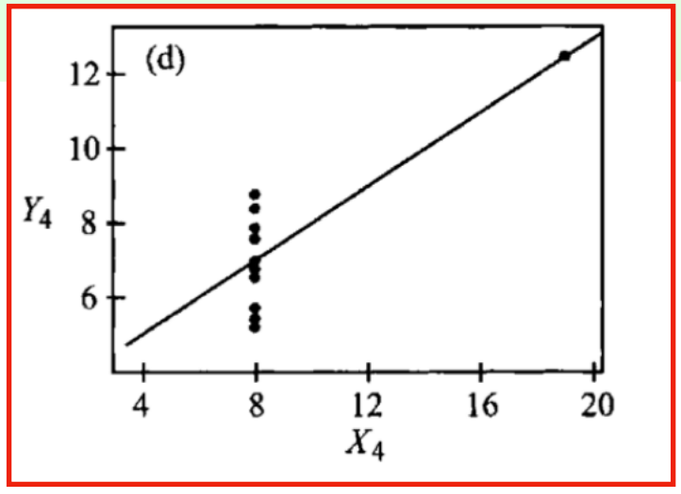
This time on the fitted versus residuals plot, for any fitted value, the spread of the residuals is about the same. However, they are not even close to centered at zero! At small and large fitted values the model is underestimating, while at medium fitted values, the model is overestimating. These are systematic errors, not random noise. So the constant variance assumption is met, but the linearity assumption is violated. The form of our model is simply wrong. We're trying to fit a line to a curve!

4.5. Leverage, Influence and Outliers

4.5 Leverage, Influence and Outliers

Influential Point

*In fitting a model to a given body of data, we would like to ensure that the fit is **not overly** determined by **one** or a **few** observations. Recall, for example, that in the Anscombe quartet data, the straight line for the data set in Figure 4.1(d) is determined **entirely** by one point. If the **extreme** point were to be removed, a very different line would result. When we have several variables, it is not possible to **detect** such a situation **graphically**. We would, however, like to know the **existence** of such points. It should be pointed out that looking at residuals in this case would be of no help, because the residual for this point is **zero**! The point is therefore not an outlier because it does not have a large residual, but it is a very **influential** point.*



*A point is an **influential point** if its **deletion**, **singly** or in **combination** with others (two or three), causes **substantial** changes in the fitted model (estimated coefficients, fitted values, t-Tests, etc.). Deletion of any point will in general cause changes in the fit. We are interested in **detecting** those points whose deletion cause large changes (i.e., they exercise undue influence). This point is illustrated by an example.*

Example on next slide

4.5 Leverage, Influence and Outliers

Example: New York Rivers Data

Consider the New York Rivers data described in Table 1.8 and given in Table 1.9. Let us fit a linear model relating the mean nitrogen concentration, Y, and the four predictor variables representing land use:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon. \tag{4.16}$$

Table 4.2 shows the regression coefficients and the t-Tests for testing the significance of the coefficients for three subsets of the data. The second column in Table 4.2 gives the regression results based on all 20 observations (rivers). The third column gives the results after deleting the Neversink River (number 4). The fourth column gives the results after deleting the Hackensack River (number 5).

Note the striking difference among the regression outputs of three data sets that differ from each other by only one observation! Observe, for example, the values of the t-Test for β_3. Based on all data, the test is insignificant, based on the data without the Neversink River, it is significantly negative, and based on the data without the Hackensack River, it is significantly positive. Only one observation can lead to substantially different results and conclusions! The Neversink and Hackensack Rivers are called influential observations because they influence the regression results substantially more than other observations in the data. Examining the raw data in Table 1.9, one can easily identify the Hackensack River because it has an unusually large value for X_3 (percentage of residential land) relative to the other values for X_3. The reason for this large value is that the Hackensack River is the only urban river in the data due to its geographic proximity to New York City with its high population density. The other rivers are in rural areas. Although the Neversink River is influential (as can be seen from Table 4.2), it is not obvious from the raw data that it is different from the other rivers in the data.

It is therefore important to identify influential observations if they exist in data. We describe methods for the detection of influential observations. Influential observations are usually outliers in either the response variable Y or the predictor variable (the X-space).

Table 1.8 Variables in Study of Water Pollution in New York Rivers

Variable	Definition
Y	Mean nitrogen concentration (mg/liter) based on samples taken at regular intervals during the spring, summer, and fall months
X_1	Agriculture: percentage of land area currently in agricultural use
X_2	Forest: percentage of forest land
X_3	Residential: percentage of land area in residential use
X_4	Commercial/Industrial: percentage of land area in either commercial or industrial use

Table 1.9 New York Rivers Data

Row	River	Y	X_1	X_2	X_3	X_4
1	Olean	1.10	26	63	1.2	0.29
2	Cassadaga	1.01	29	57	0.7	0.09
3	Oatka	1.90	54	26	1.8	0.58
4	Neversink	1.00	2	84	1.9	1.98
5	Hackensack	1.99	3	27	29.4	3.11
6	Wappinger	1.42	19	61	3.4	0.56
7	Fishkill	2.04	16	60	5.6	1.11
8	Honeoye	1.65	40	43	1.3	0.24
9	Susquehanna	1.01	28	62	1.1	0.15
10	Chenango	1.21	26	60	0.9	0.23
11	Tioughnioga	1.33	26	53	0.9	0.18
12	West Canada	0.75	15	75	0.7	0.16
13	East Canada	0.73	6	84	0.5	0.12
14	Saranac	0.80	3	81	0.8	0.35
15	Ausable	0.76	2	89	0.7	0.35
16	Black	0.87	6	82	0.5	0.15
17	Schoharie	0.80	22	70	0.9	0.22
18	Raquette	0.87	4	75	0.4	0.18
19	Oswegatchie	0.66	21	56	0.5	0.13
20	Cohocton	1.25	40	49	1.1	0.13

Table 4.2 New York Rivers Data: The t-Tests for the Individual Coefficients

	Observations Deleted		
Test	None	Neversink	Hackensack
t_0	1.40	1.21	2.08
t_1	0.39	0.92	0.25
t_2	−0.93	−0.74	−1.45
t_3	−0.21	−3.15	4.08
t_4	1.86	4.45	0.66

4.5 Leverage, Influence and Outliers

Use R for the previous example

```
> ##################  New York River's Dataset
> NYriver<-read.table('data/P010.txt',header=TRUE)   ## read the data
> NYriver<-NYriver[,-1]
> mod_full<-lm(Nitrogen~.,data=NYriver)
> summary(mod_full)

Call:
lm(formula = Nitrogen ~ ., data = NYriver)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49404 -0.13180  0.01951  0.08287  0.70480

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.722214   1.234082   1.396   0.1832
Agr          0.005809   0.015034   0.386   0.7046
Forest      -0.012968   0.013931  -0.931   0.3667
Rsdntial    -0.007227   0.033830  -0.214   0.8337
ComIndl      0.305028   0.163817   1.862   0.0823 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2649 on 15 degrees of freedom
Multiple R-squared:  0.7094,    Adjusted R-squared:  0.6319
F-statistic: 9.154 on 4 and 15 DF,  p-value: 0.0005963
```

```
> #### delete Neversink
> mod_par1<-lm(Nitrogen~.,data=NYriver[-4,])
> summary(mod_par1)

Call:
lm(formula = Nitrogen ~ ., data = NYriver[-4, ])

Residuals:
     Min       1Q   Median       3Q      Max
-0.36421 -0.11154  0.00406  0.12059  0.26538

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.099471   0.911636   1.206 0.247788
Agr          0.010137   0.010984   0.923 0.371705
Forest      -0.007589   0.010222  -0.742 0.470098
Rsdntial    -0.123793   0.039337  -3.147 0.007134 **
ComIndl      1.528956   0.343719   4.448 0.000551 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1925 on 14 degrees of freedom
Multiple R-squared:  0.8557,    Adjusted R-squared:  0.8145
F-statistic: 20.76 on 4 and 14 DF,  p-value: 9.087e-06
```

```
> #### delete Hackensack
> mod_par1<-lm(Nitrogen~.,data=NYriver[-5,])
> summary(mod_par1)

Call:
lm(formula = Nitrogen ~ ., data = NYriver[-5, ])

Residuals:
     Min       1Q   Median       3Q      Max
-0.40124 -0.09184  0.02912  0.10840  0.22493

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.626014   0.781091   2.082  0.05620 .
Agr          0.002352   0.009539   0.247  0.80881
Forest      -0.012760   0.008815  -1.448  0.16976
Rsdntial     0.181161   0.044390   4.081  0.00112 **
ComIndl      0.075618   0.113957   0.664  0.51775
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1676 on 14 degrees of freedom
Multiple R-squared:  0.864,     Adjusted R-squared:  0.8252
F-statistic: 22.24 on 4 and 14 DF,  p-value: 6.055e-06
```

4.5 Leverage, Influence and Outliers

Leverage

Recall the leverage of i-th data point is defined by p_{ii}, i.e., the i-th diagonal entry of $\mathbf{P}$, the **Projection or hat matrix**

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

The average leverage is $\frac{\sum_{i=1}^{n} p_{ii}}{n}$ → **In fact, this always equals** $\frac{p+1}{n}$

p is the number of predictors

The i-th data point has **high leverage** if

$$p_{ii} \geq 2\left(\frac{\sum_{i=1}^{n} p_{ii}}{n}\right)$$

4.5 Leverage, Influence and Outliers

Outliers

Outliers in the Response Variable

Outliers in the Response Variable

Observations with large standardized residuals are outliers in the response variable because they lie far from the fitted equation in the Y-direction. Since the standardized residuals are approximately normally distributed with mean zero and a standard deviation 1, points with standardized residuals larger than 2 or 3 standard deviations away from the mean (zero) are called *outliers*. Outliers may indicate a model failure for those points. They can be identified using formal testing procedures

or through appropriately chosen graphs of the residuals, the approach we adopt here. The pattern of the residuals is more important than their numeric values. Graphs of residuals will often expose gross model violations when they are present. Studying residual plots is one of the main tools in our analysis.

4.5 Leverage, Influence and Outliers

Outliers

Outliers in the Predictors

$$p_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{in the case of simple linear regression}$$

Outliers in the Predictors

Outliers can also occur in the predictor variables (the X-space). They can also affect the regression results. The leverage values p_{ii}, described earlier, can be used to measure outlyingness in the X-space. This can be seen from an examination of the formula for p_{ii} in the simple regression case given in (4.7), which shows that the farther a point is from $\bar{x}$, the larger the corresponding value of p_{ii}. This is also true in multiple regression. Therefore, p_{ii} can be used as a measure of outlyingness in the X-space because observations with large values of p_{ii} are outliers in the X-space (i.e., compared to other points in the space of the predictors). Observations that are outliers in the X-space [e.g., the point with the largest value of X_4 in Figure 4.1(d)] are known as *high-leverage* points to distinguish them from observations that are outliers in the response variable (those with large standardized residuals).

The leverage values possess several interesting properties

For example, they lie between 0 and 1 and their average value is $(p+1)/n$. Points with p_{ii} greater than $2(p+1)/n$ (twice the average value) are generally regarded as points with high leverage

4.5 Leverage, Influence and Outliers

Outliers

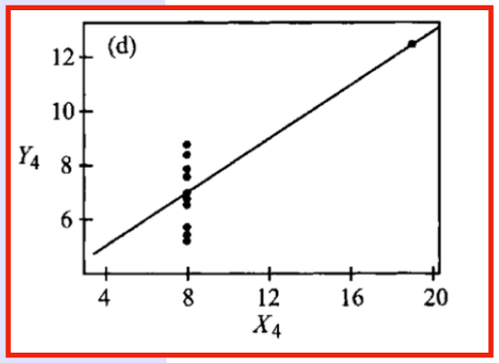
Masking and Swamping Problems

The standardized residuals provide valuable information for validating linearity and normality assumptions and for the identification of outliers. However, analyses that are based on residuals alone may fail to detect outliers and influential observations for the following reasons:

1. *The presence of high-leverage points*: The ordinary residuals, e_i, and leverage values, p_{ii}, are related by

$$p_{ii} + \frac{e_i^2}{\text{SSE}} \leq 1, \tag{4.17}$$

where SSE is the residual sum of squares. This inequality indicates that high-leverage points (points with large values of p_{ii}) tend to have small residuals. For example, the point at $X = 19$ in Figure 4.1(d) is extremely influential even though its residual is identically zero. Therefore, in addition to an examination of the standardized residuals for outliers, an examination of the leverage values is also recommended for the identification of troublesome points.



2. *The masking and swamping problems*: Masking occurs when the data contain outliers but we fail to detect them. This can happen because some of the outliers may be hidden by other outliers in the data. Swamping occurs when we wrongly declare some of the nonoutlying points as outliers. This can occur because outliers tend to pull the regression equation toward them, hence make other points lie far from the fitted equation. Thus, masking is a false negative decision whereas swamping is a false positive. An example of a data set in which masking and swamping problems are present is given below.

For the above reasons, additional measures of the influence of observations are needed. Before presenting these methods, we illustrate the above concepts using a real-life example.

4.5 Leverage, Influence and Outliers

Outliers

Example: New York Rivers Data

Consider the New York Rivers data, but now for illustrative purpose, let us consider fitting the simple regression model

$$Y = \beta_0 + \beta_4 X_4 + \varepsilon, \qquad (4.18)$$

relating the mean nitrogen concentration, Y, to the percentage of land area in either industrial or commercial use, X_4. The scatter plot of Y versus X_4 together with the corresponding least squares fitted line are given in Figure 4.5. The corresponding standardized residuals, r_i, and the leverage values, p_{ii}, are given in Table 4.3 and their respective index plots are shown in Figure 4.6. In the index plot of the standardized residuals all the residuals are small indicating that there are no outliers in the data. This is a wrong conclusion because there are two clear outliers in the data as can be seen in the scatter plot in Figure 4.5. Thus masking has occurred! Because of the relationship between leverage and residual in (4.17), the Hackensack River with its large value of $p_{ii} = 0.67$ has a small residual. While a small value of the residual is desirable, the reason for the small value of the residual here is not due to a good fit; it is due to the fact that observation 5 is a high-leverage point and, in collaboration with observation 4, they pull the regression line toward them.

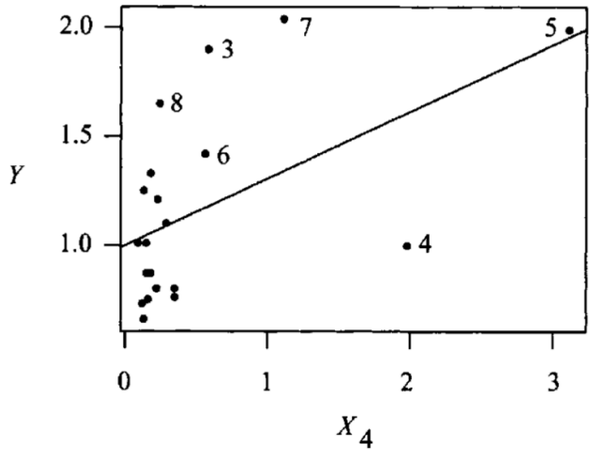


Figure 4.5 New York Rivers Data: Scatter plot of Y versus X_4.

Figure 4.6 and Table 4.3

4.5 Leverage, Influence and Outliers

Outliers

Example: New York Rivers Data

A commonly used cutoff value for p_{ii} is $2(p+1)/n = 0.2$. Accordingly, two points (Hackensack, $p_{ii} = 0.67$, and Neversink, $p_{ii} = 0.25$) that we have seen previously stand out in the scatter plot of points in Figure 4.5, are flagged as high-leverage points as can be seen in the index plot of p_{ii} in Figure 4.6(b), where the two points are far from the other points. This example shows clearly that looking solely at residual plots is inadequate.

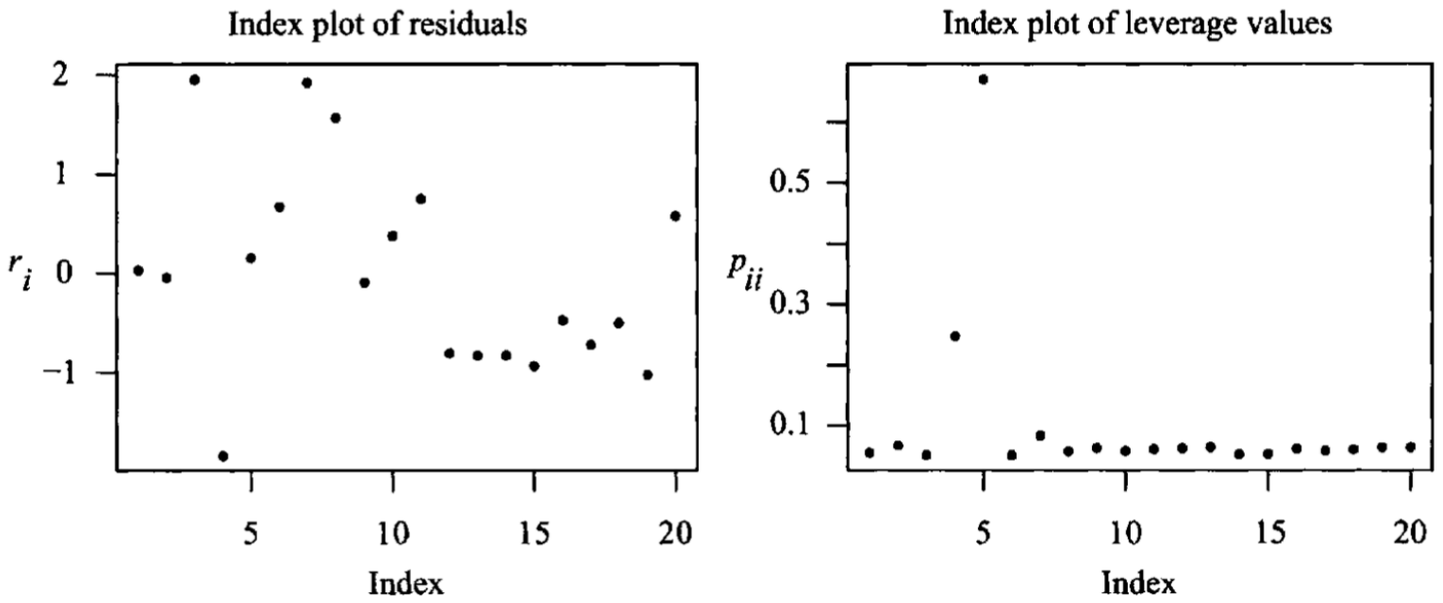


Figure 4.6 New York Rivers Data: Index plots of the standardized residuals, r_i, and the leverage values, p_{ii}.

Table 4.3

4.5 Leverage, Influence and Outliers

Outliers

Example: New York Rivers Data

Table 4.3 New York Rivers Data: Standardized Residuals, r_i, and Leverage Values, p_{ii}, from Fitting Model 4.18

Row	r_i	p_{ii}	Row	r_i	p_{ii}
1	0.03	0.05	11	0.75	0.06
2	−0.05	0.07	12	−0.81	0.06
3	1.95	0.05	13	−0.83	0.06
4	−1.85	0.25	14	−0.83	0.05
5	0.16	0.67	15	−0.94	0.05
6	0.67	0.05	16	−0.48	0.06
7	1.92	0.08	17	−0.72	0.06
8	1.57	0.06	18	−0.50	0.06
9	−0.10	0.06	19	−1.03	0.06
10	0.38	0.06	20	0.57	0.06

4.5 Leverage, Influence and Outliers

Summary

Influence: deletion of this observation substantially affects the LSE

There are two types of **outliers**:

- outlying in the response – *can be checked by standardized residual if* $r_i \geq 3$
- outlying in the predictor – *can be checked by leverage values if* $p_{ii} \geq \frac{2(p+1)}{n}$, *or in simple linear regression by* x_i *being far away from* $\bar{x}$

In the next section, we shall see that **standardized residual** and **leverage** collectively determines whether an observation is influential

4.6 Measures of Influence

4.6 Measures of Influence

Leave one Data point out

The influence of an observation may be measured by the effects it produces on the fit when it is omitted from the data in the fitting process. This deletion is almost always done one point at a time. Let $\hat{\beta}_{0(i)}, \hat{\beta}_{1(i)}, \cdots, \hat{\beta}_{p(i)}$ denote the regression coefficients obtained when the ith observation is deleted $(i = 1, 2, \cdots, n)$. Similarly, let $\hat{y}_{1(i)}, \hat{y}_{2(i)}, \cdots, \hat{y}_{n(i)}$, and $\hat{\sigma}^2_{(i)}$ be the predicted values and residual mean square when we drop the ith observation. Note that

$$\hat{y}_{m(i)} = \hat{\beta}_{0(i)} + \hat{\beta}_{1(i)} x_{m1} + \cdots + \hat{\beta}_{p(i)} x_{mp} \tag{4.19}$$

is the fitted value for observation m when the fitted equation is obtained with the ith observation deleted. Influence measures look at differences produced in quantities such as $\hat{\beta}_j - \hat{\beta}_{j(i)}$ or $\hat{y}_j - \hat{y}_{j(i)}$.

where $\hat{\beta}_1, \cdots, \hat{\beta}_p$ denotes the LSE based on ALL data points.

$\hat{\beta}_{0(i)}, \hat{\beta}_{1(i)}, \cdots, \hat{\beta}_{p(i)}$ are the LSE based on $n-1$ observations with the i-th observation deleted from the original n observations.

4.6 Measures of Influence

Cook's Distance

An influence measure proposed by Cook (1977) is widely used. *Cook's distance* measures the difference between the regression coefficients obtained from the full data and the regression coefficients obtained by deleting the ith observation, or equivalently, the difference between the fitted values obtained from the full data and the fitted values obtained by deleting the ith observation. Accordingly, Cook's distance measures the influence of the ith observation by

$$C_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{\hat{\sigma}^2(p+1)}, \quad i = 1, 2, \cdots, n. \tag{4.20}$$

$\hat{\sigma}^2$ is obtained from LSE using ALL observations

It can be shown that C_i can be expressed as

$$C_i = \frac{r_i^2}{p+1} \times \frac{p_{ii}}{1-p_{ii}}, \quad i = 1, 2, \cdots, n. \tag{4.21}$$

Thus, Cook's distance is a multiplicative function of two basic quantities. The first is the square of the standardized residual, r_i, defined in (4.13) and the second is the so-called *potential* function $p_{ii}/(1-p_{ii})$, where p_{ii} is the leverage of the ith observation introduced previously. If a point is influential, its deletion causes large changes and the value of C_i will be large. Therefore, a large value of C_i indicates that the point is influential. It has been suggested that points with C_i values greater than the 50% point of the F-distribution with $p+1$ and $n-p-1$ degrees of freedom be classified as influential points. A practical operational rule is to classify points with C_i values greater than 1 as being influential. Rather than using a rigid cutoff rule, we suggest that all C_i values be examined graphically. A dot plot or an index plot of C_i is a useful graphical device. When the C_i values are all about the same, no action need be taken. On the other hand, if there are data points with C_i values that stand out from the rest, these points should be flagged and examined. The model may then be refitted without the offending points to see the effect of these points.

potential function

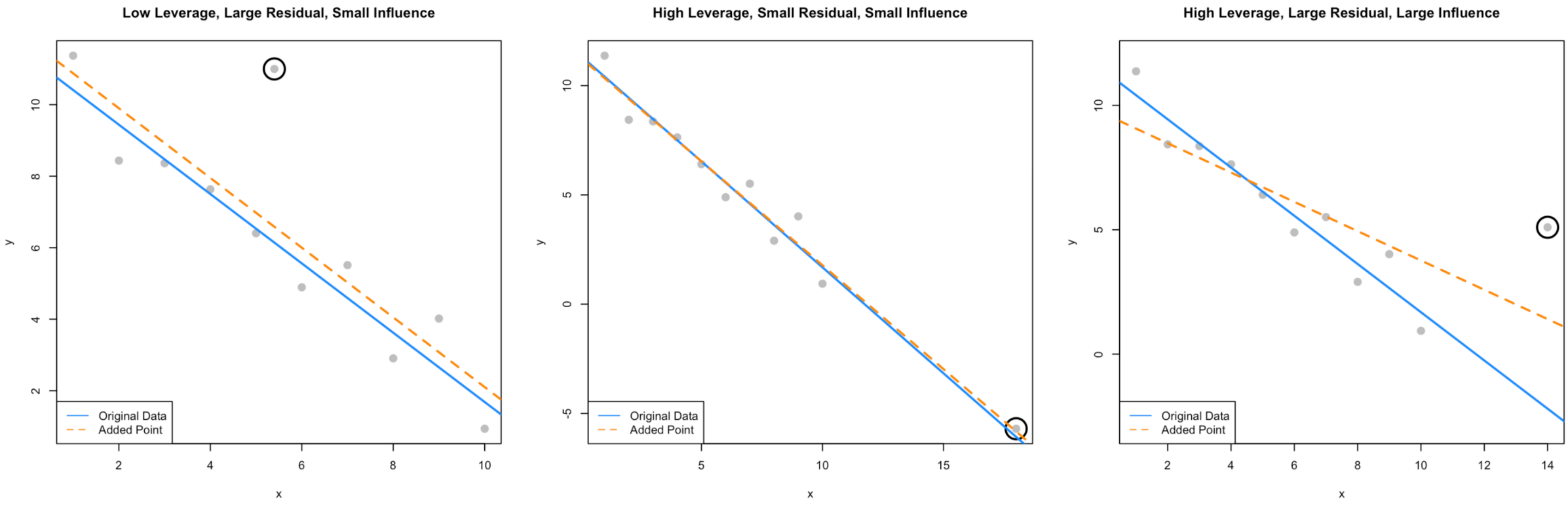
Example on next slide

4.6 Measures of Influence

Cook's Distance

These three examples illustrate the relationship between Cook's distance, residual and leverage.

For simple linear regression, x_i is far away from $\bar{x}$, then it has high leverage



The blue line is obtained without using the circled point; the orange line is obtained including the circled point

$$C_i = \frac{r_i^2}{p+1} \times \frac{p_{ii}}{1-p_{ii}},$$

4.6 Measures of Influence

Welsch and Kuh Measure

A measure similar to Cook's distance has been proposed by Welsch and Kuh (1977) and named DFITS. It is defined as

$$\text{DFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{p_{ii}}}, \quad i = 1, 2, \cdots, n. \tag{4.22}$$

Thus, DFITS_i is the scaled difference between the ith fitted value obtained from the full data and the ith fitted value obtained by deleting the ith observation. The difference is scaled by $\hat{\sigma}_{(i)}\sqrt{p_{ii}}$. It can be shown that DFITS_i can be written as

externally studentized residual

$$\text{DFITS}_i = r_i^* \sqrt{\frac{p_{ii}}{1 - p_{ii}}}, \quad i = 1, 2, \cdots, n, \tag{4.23}$$

where r_i^* is the standardized residual defined in (4.14). DFITS_i corresponds to $\sqrt{C_i}$ when the normalization is done by using $\hat{\sigma}_{(i)}$ instead of $\hat{\sigma}$. Points with $|\text{DFITS}_i|$ larger than $2\sqrt{(p+1)/(n-p-1)}$ are usually classified as influential points. Again, instead of having a strict cutoff value, we use the measure to sort out points of abnormally high influence relative to other points on a graph such as the index plot, the dot plot, or the box plot. There is not much to choose between C_i and DFITS_i – both give similar answers because they are functions of the residual and leverage values. Most computer software will give one or both of the measures, and it is sufficient to look at only one of them.

4.6 Measures of Influence

Hadi's Influence Measure

Hadi (1992) proposed a measure of the influence of the ith observation based on the fact that influential observations are outliers in either the response variable or in the predictors, or both. Accordingly, the influence of the ith observation can be measured by

$$H_i = \frac{p_{ii}}{1-p_{ii}} + \frac{p+1}{1-p_{ii}} \frac{d_i^2}{1-d_i^2}, \quad i = 1, 2, \cdots, n, \qquad (4.24)$$

where $d_i = e_i/\sqrt{\text{SSE}}$ is the so-called *normalized residual*. The first term on the right-hand side of (4.24) is the potential function which measures outlyingness in the X-space. The second term is a function of the residual, which measures outlyingness in the response variable. It can be seen that observations will have large values of H_i if they are outliers in the response and/or the predictor variables, that is, if they have large values of r_i, p_{ii}, or both. The measure H_i does not focus on a specific regression result, but it can be thought of as an overall general measure of influence which depicts observations that are influential on at least one regression result.

Note that C_i and DFITS_i are multiplicative functions of the residuals and leverage values, whereas H_i is an additive function. The influence measure H_i can best be examined graphically in the same way as Cook's distance and Welsch and Kuh measure.

4.6 Measures of Influence

Example

Example: New York Rivers Data

Consider again fitting the simple regression model in (4.18), which relates the mean nitrogen concentration, Y, to the percentage of land area in commercial/industrial use, X_4. The scatter plot of Y versus X_4 and the corresponding least squares regression are given in Figure 4.5. Observations 4 (the Neversink River) and 5 (the Hackensack River) are located far from the bulk of other data points in Figure 4.5. Also observations 7, 3, 8, and 6 are somewhat sparse in the upper-left region of the graph. The three influence measures discussed above which result from fitting model (4.18) are shown in Table 4.4, and the corresponding index plots are shown in Figure 4.7. No value of C_i exceeds its cutoff value of 1. However, the index plot of C_i in Figure 4.7(a) shows clearly that observation number 4 (Neversink) should be flagged as an influential observation. This observation also exceeds its DFITS_i cutoff value of $2\sqrt{(p+1)/(n-p-1)} = 2/3$. As can be seen from Figure 4.7, observation number 5 (Hackensack) was not flagged by C_i or by DFITS_i. This is due to the small value of the residual because of its high leverage and to the multiplicative nature of the measure.

Table 4.4 New York Rivers Data. Influence Measures from Fitting Model 4.18: Cook's Distance, C_i, Welsch and Kuh Measure, DFITS_i

Row	C_i	DFITS_i	Row	C_i	DFITS_i
1	0.00	0.01	11	0.02	0.19
2	0.00	−0.01	12	0.02	−0.21
3	0.10	0.49	13	0.02	−0.22
4	0.56	−1.14	14	0.02	−0.19
5	0.02	0.22	15	0.02	−0.22
6	0.01	0.15	16	0.01	−0.12
7	0.17	0.63	17	0.02	−0.18
8	0.07	0.40	18	0.01	−0.12
9	0.00	−0.02	19	0.04	−0.27
10	0.00	0.09	20	0.01	0.15

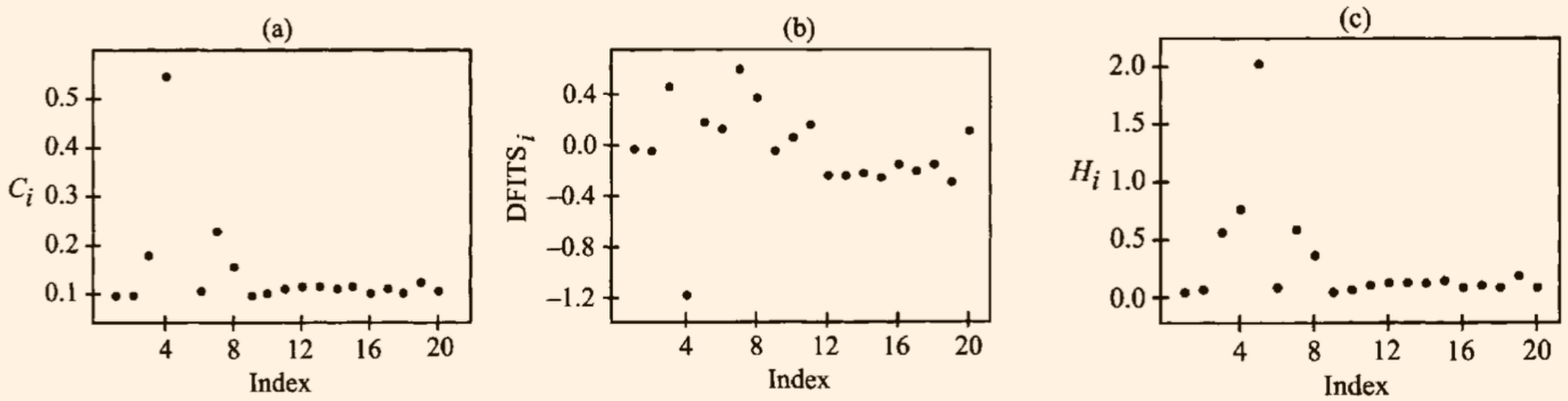


Figure 4.7 New York Rivers data: Index plots of influence measures: (a) Cook's distance, C_i, (b) Welsch and Kuh measure, DFITS_i, and (c) Hadi's influence measure H_i.

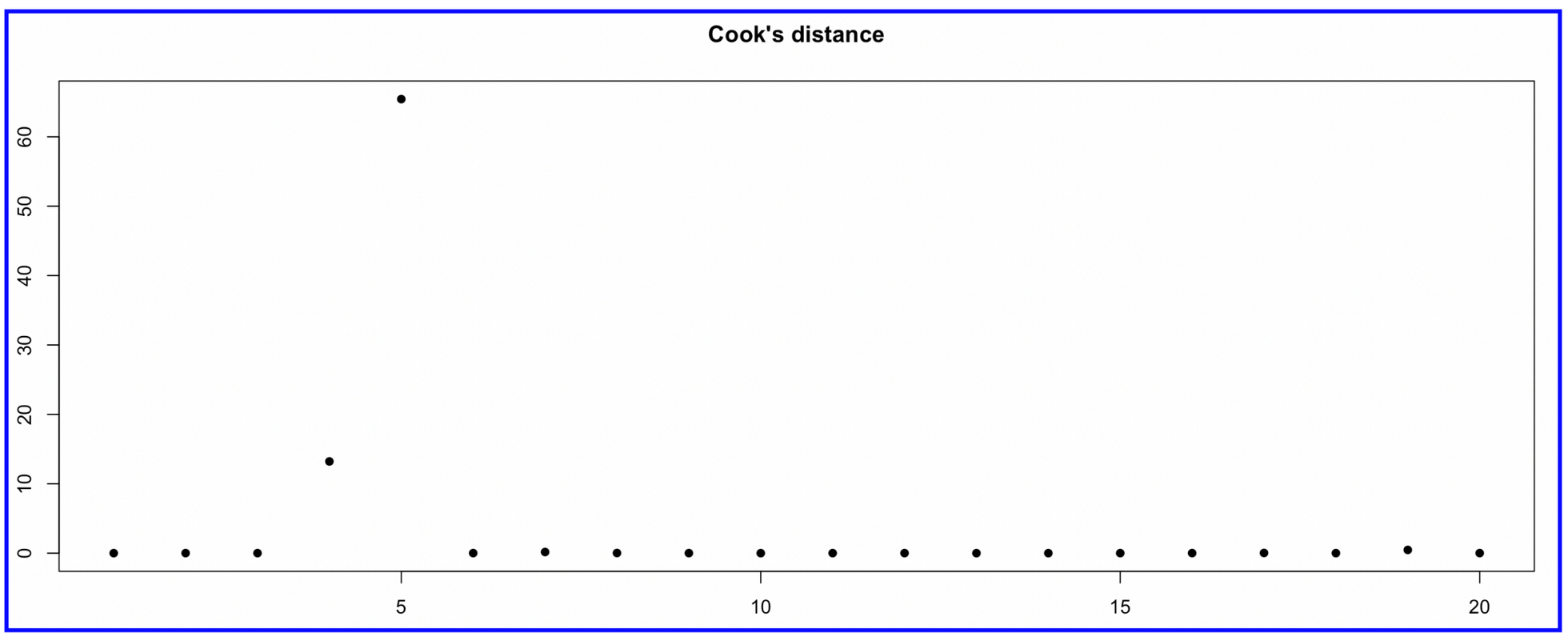
4.6 Measures of Influence

Use R for last example

```
> ####################  Influential measures
> ##################  New York River's Dataset   --- Cooks' Distnace
> NYriver<-read.table('data/P010.txt',header=TRUE)   ## read the data
> NYriver<-NYriver[,-1]
> mod_full<-lm(Nitrogen~.,data=NYriver)
> cooks.distance(mod_full)
          1           2           3           4           5           6           7           8           9          10          11          12
4.913989e-04 7.584870e-03 5.632525e-03 1.321955e+01 6.542632e+01 9.094491e-03 1.643046e-01 2.289146e-02 6.587261e-03 5.461864e-04 7.304071e-03 7.140521e-03
         13          14          15          16          17          18          19          20
5.871323e-04 5.940847e-05 5.354455e-03 9.563833e-03 2.855140e-02 4.443351e-03 4.681928e-01 9.524415e-03
```

using all predictors

```
> plot(1:n,cooks.distance(mod_full),pch=16,main="Cook's distance")
```



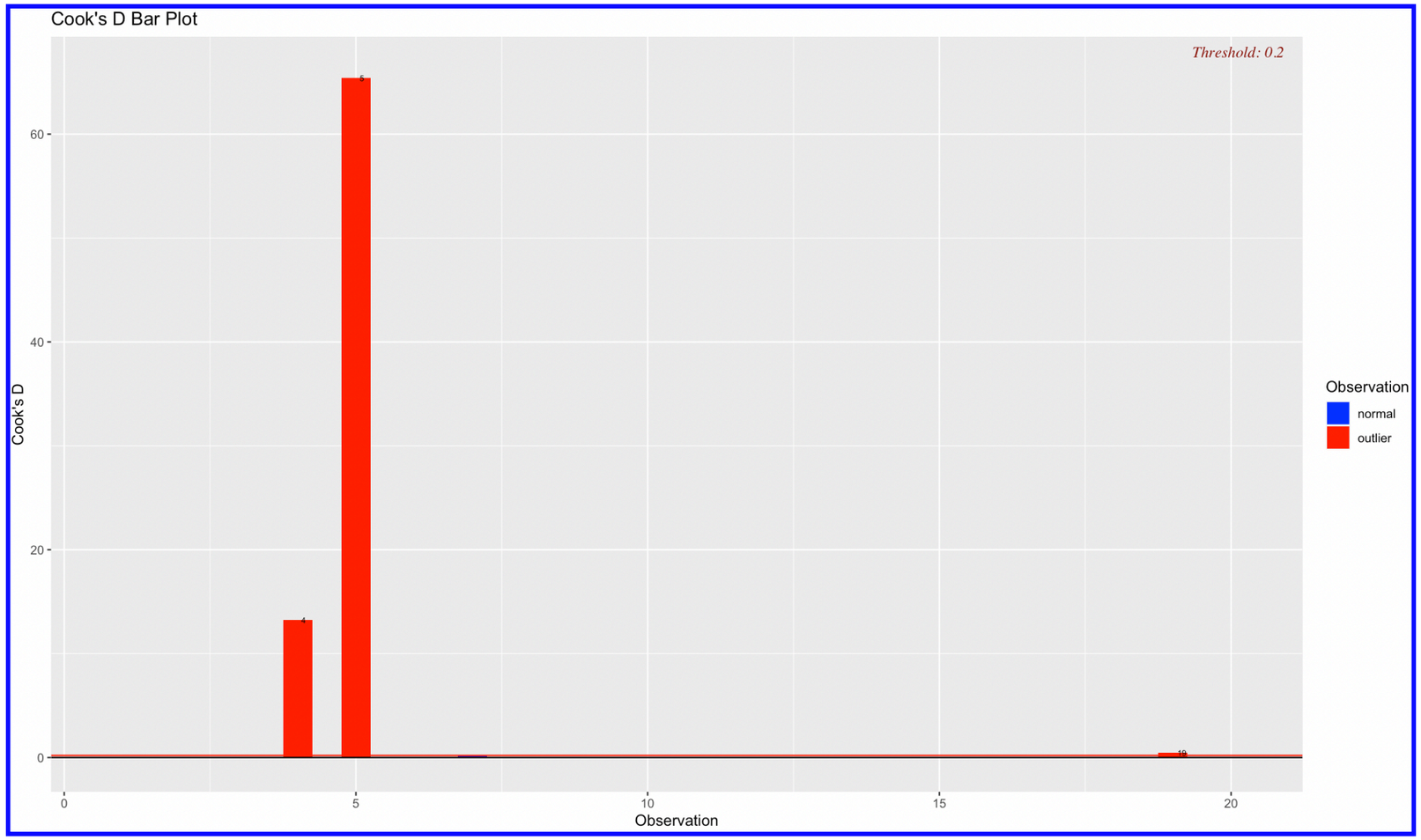
4.6 Measures of Influence

Use R for last example

Use a library to automatically draw the influence measure plot

```
library(olsrr)        ### use a library to draw the plot
NYriver<-read.table('data/P010.txt',header=TRUE)   ## read the data
NYriver<-NYriver[,-1]
mod_full<-lm(Nitrogen~.,data=NYriver)
ols_plot_cooksd_bar(mod_full)
```

using all predictors



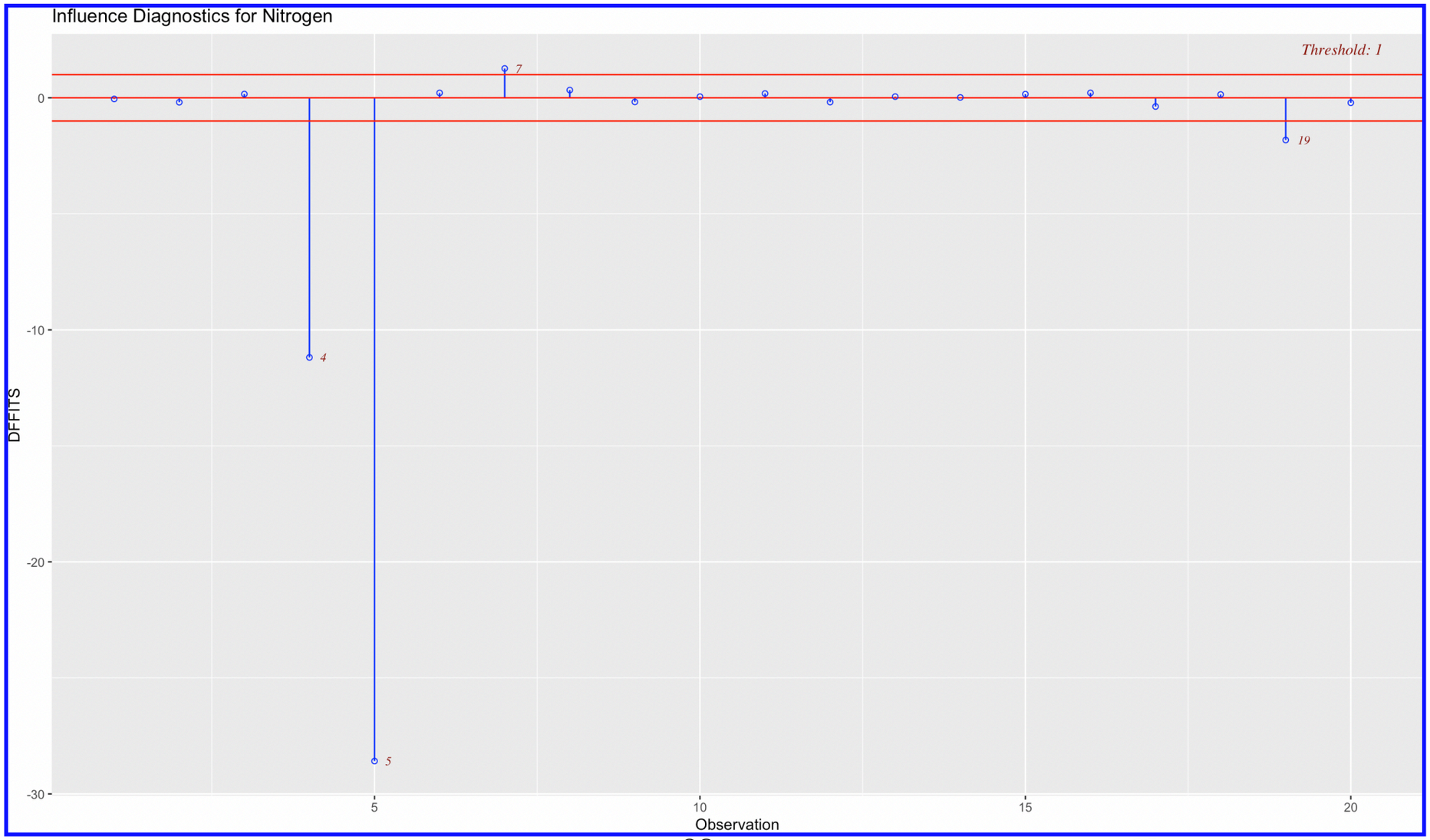
4.6 Measures of Influence

Use R for last example

Use a library to automatically draw the influence measure plot

```
library(olsrr)        ### use a library to draw the plot
NYriver<-read.table('data/P010.txt',header=TRUE)   ## read the data
NYriver<-NYriver[,-1]
mod_full<-lm(Nitrogen~.,data=NYriver)
ols_plot_dffits(mod_full)
```

using all predictors



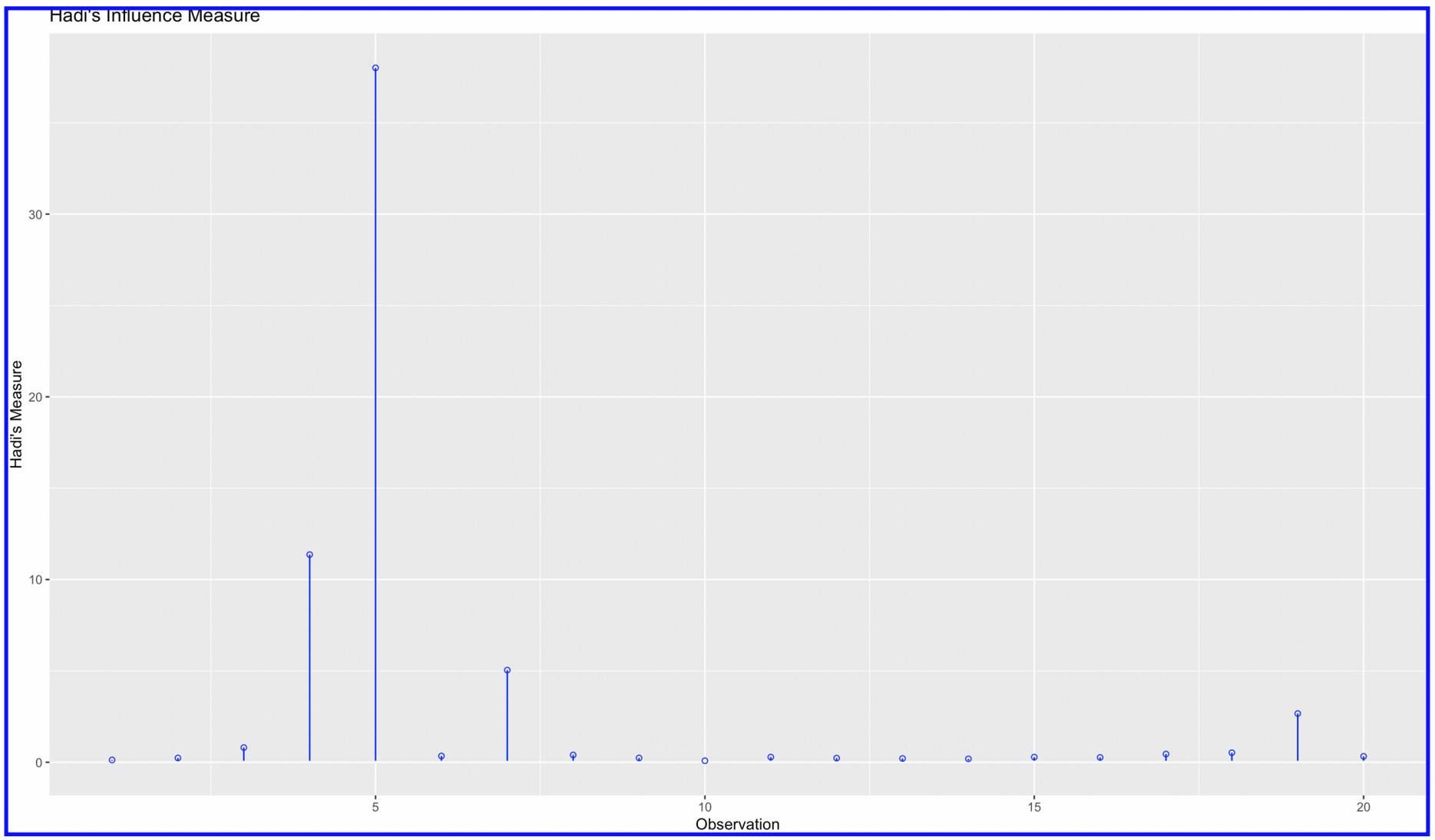
4.6 Measures of Influence

Use R for last example

Use a library to automatically draw the influence measure plot

```
library(olsrr)        ### use a library to draw the plot
NYriver<-read.table('data/P010.txt',header=TRUE)   ## read the data
NYriver<-NYriver[,-1]
mod_full<-lm(Nitrogen~.,data=NYriver)
ols_plot_hadi(mod_full)
```

using all predictors



4.6 Measures of Influence

Potential-Residual Plot

The formula for H_i in (4.24) suggests a simple graph to aid in classifying unusual observations as high-leverage points, outliers, or a combination of both. The graph is called the *potential-residual* (P-R) plot (Hadi, 1992) because it is the scatter plot of

Potential Function		Residual Function
$\frac{p_{ii}}{1-p_{ii}}$	versus	$\frac{p+1}{1-p_{ii}}\frac{d_i^2}{1-d_i^2}$.

$$d_i = \frac{e_i}{\sqrt{\text{SSE}}}$$

As an illustrative example, the P-R plot obtained from fitting model $Y = \beta_0 + \beta_4 X_4 + \varepsilon$ on New York Rivers Data is shown in Figure 4.8. Observation 5, which is a high-leverage point, is located by itself in the upper-left corner of the plot. Four outlying observations (3, 7, 4, and 8) are located in the lower-right area of the graph.

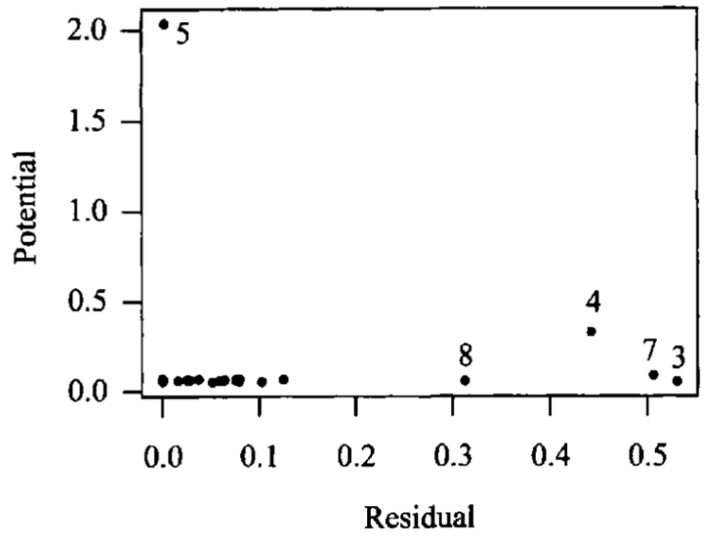
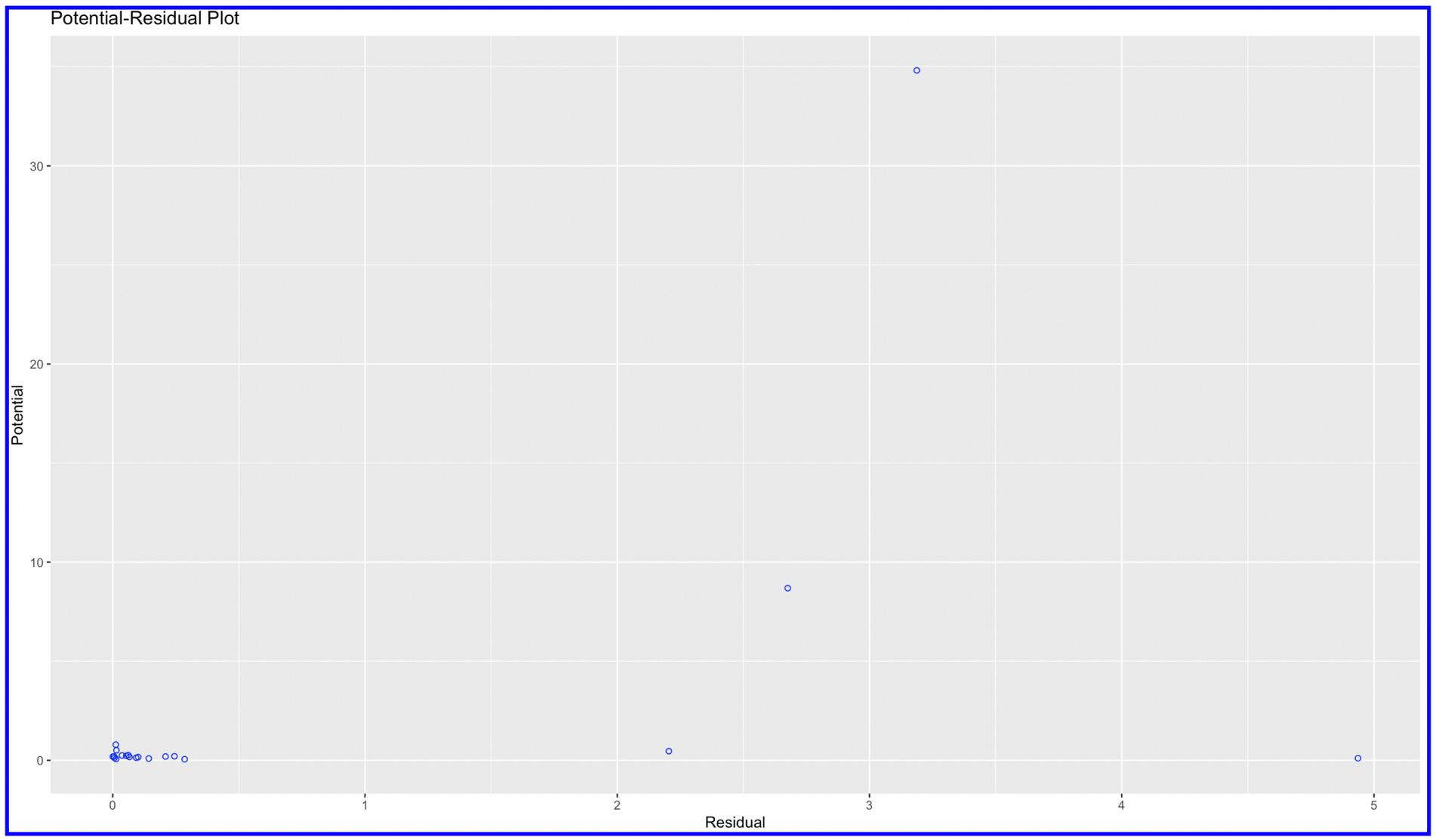


Figure 4.8 New York Rivers data: Potential-Residual plot.

4.6 Measures of Influence

Potential-Residual Plot

```
library(olsrr)        ### use a library to draw the plot
NYriver<-read.table('data/P010.txt',header=TRUE)   ## read the data
NYriver<-NYriver[,-1]
mod_full<-lm(Nitrogen~.,data=NYriver)
ols_plot_resid_pot(mod_full)
```

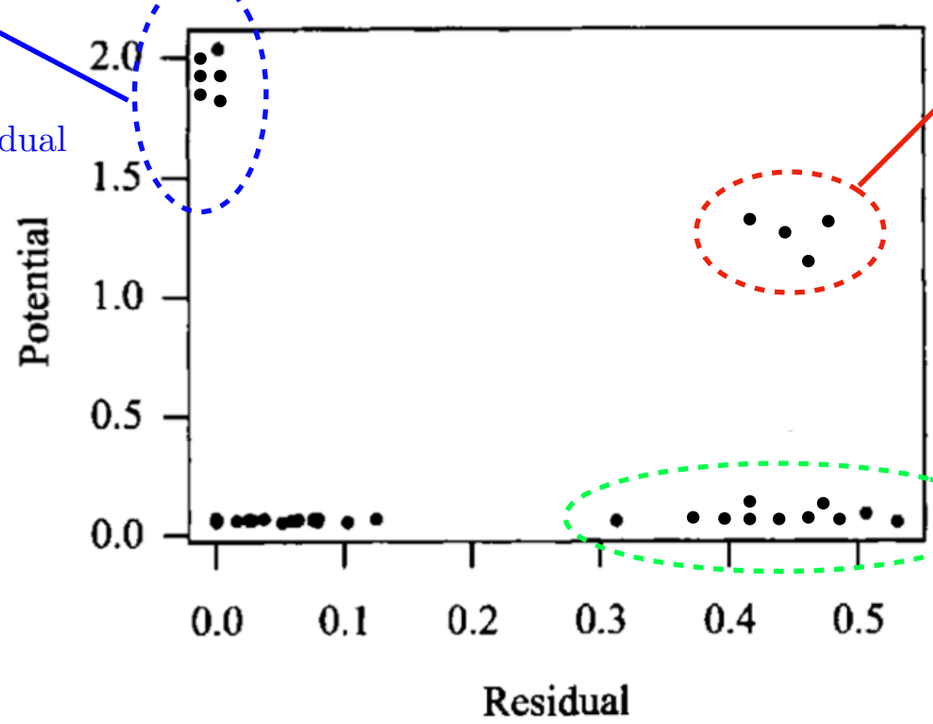


4.6 Measures of Influence

Potential-Residual Plot

large leverage and small residual
outlying in the X-space

whether they are influential depends on
the combined influence of leverage and residual



large leverage and large residual,
they are usually the influential points

small leverage and large residual
outlying in the Y-space, i.e., response

whether they are influential depends on
the combined influence of leverage and residual

It is clear now that some individual data points may be flagged as outliers, leverage points, or influential points. The main usefulness of the leverage and influence measures is that they give the analyst a complete picture of the role played by different points in the entire fitting process. Any point falling in one of these categories should be carefully examined for accuracy (gross error, transcription error), relevancy (whether it belongs to the data set), and special significance (abnormal condition, unique situation). Outliers should always be scrutinized carefully. Points with high leverage that are not influential do not cause problems. High-leverage points that are influential should be investigated because these points are outlying as far as the predictor variables are concerned and also influence the fit. To get an idea of the sensitivity of the analysis to these points, the model should be fitted without the offending points and the resulting coefficients exam

4.7 What to Do with the Outliers?

4.7 What to Do with the Outliers?

*Outliers and influential observations should not **routinely** be deleted or automatically **down-weighted** because they are **not necessarily** bad observations. On the contrary, if they are correct, they may be the most **informative** points in the data. For example, they may indicate that the data did not come from a normal population or that the model is not linear. To illustrate that outliers and influential observations can be the most informative points in the data, we use the exponential growth data described in the following example.*

*Figure 4.9 is the scatter plot of two variables, the size of a certain population, Y, and time, X. As can be seen from the scatter of points, the **majority** of the points resemble a linear relationship between population size and time as indicated by the straight line in Figure 4.9. According to this model the two points 22 and 23 in the upper-right corner are **outliers**. If these points, however, are **correct**, they are the only observations in the data set that indicate that the data follow a **nonlinear** (e.g., exponential) model, such as the one shown in the graph. Think of this as a population of bacteria which increases very slowly over a period of time. After a critical point in time, however, the population explodes.*

Figure 4.9

4.7 What to Do with the Outliers?

What to do with outliers and influential observations once they are identified? Because outliers and influential observations can be the most ***informative*** *observations in the data set, they should not be automatically* ***discarded*** *without justification. Instead, they should be examined to determine why they are* ***outlying*** *or* ***influential****. Based on this examination, appropriate corrective actions can then be taken. These corrective actions include:* ***correction of error*** *in the data,* ***deletion*** *or* ***down-weighing outliers****,* ***transforming*** *the data, considering a* ***different*** *model, and* ***redesigning*** *the experiment or the sample survey, collecting* ***more*** *data.*

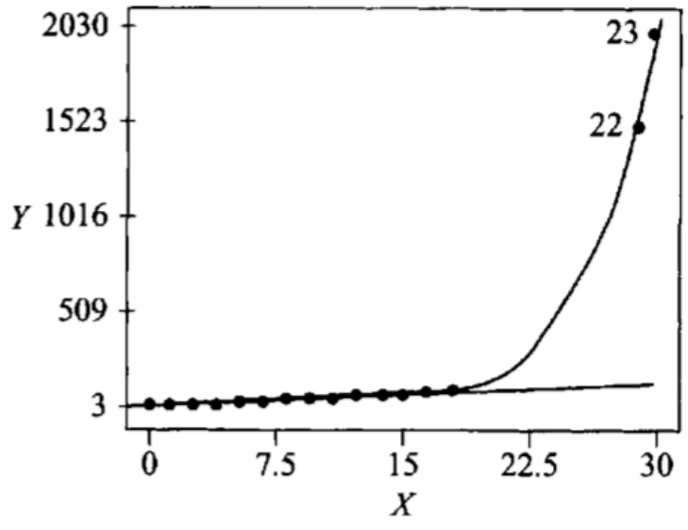


Figure 4.9 Scatter plot of population size, Y, versus time, X. The curve is obtained by fitting an exponential function to the full data. The straight line is the least squares line when observations 22 and 23 are are deleted.