Chapter 5

Model Adequacy Checking

In chapters 2, 3 and 4 we assumed that the regression model was a valid model for the data. We will start by examining whether the model under the consideration is indeed valid. We will then check the validity of assumptions we made previously.

5.1 Evaluating the Fit

How to evaluate the proposed model?

For this, we may look at the numerical regression output such as least-squares estimates, t-test statistic, F statistic or R^2. But, we usually cannot detect departures from underlying assumptions by examination of the standard summary statistics. Anscombe's four constructed data sets illustrate the point that looking only at the numerical regression output may lead to very misleading conclusions about the data and lead to adopting the wrong model.

Example of valid and invalid regression models: Anscombe's four data sets

Observation	X1	X2	X3	X4	Y1	Y2	Y3	Y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.10	5.39	12.50
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

Source: Anscombe, F. (1973) Graphs in statistical analysis, *The American Statistician* **27, 17–21.**

When a regression model is fitted to data sets 1, 2, 3 and 4, in each case the fitted line is

$$\hat{Y} = 3 + .5X.$$

SAS Output

```
------------------------------------------------------------------------------
                        Dependent Variable: Y1

                         Analysis of Variance

                                   Sum of          Mean
Source                     DF      Squares        Square   F Value   Pr > F

Model                       1     27.51000      27.51000     17.99   0.0022
Error                       9     13.76269       1.52919
Corrected Total            10     41.27269

          Root MSE              1.23660    R-Square     0.6665
          Dependent Mean        7.50091    Adj R-Sq     0.6295
          Coeff Var            16.48605

                         Parameter Estimates

                      Parameter       Standard
  Variable     DF      Estimate          Error    t Value    Pr > |t|

  Intercept     1       3.00009        1.12475       2.67      0.0257
  X1            1       0.50009        0.11791       4.24      0.0022

------------------------------------------------------------------------------
                        Dependent Variable: Y2

                         Analysis of Variance

                                   Sum of          Mean
Source                     DF      Squares        Square   F Value   Pr > F

Model                       1     27.50000      27.50000     17.97   0.0022
Error                       9     13.77629       1.53070
Corrected Total            10     41.27629

          Root MSE              1.23721    R-Square     0.6662
          Dependent Mean        7.50091    Adj R-Sq     0.6292
          Coeff Var            16.49419

                         Parameter Estimates

                      Parameter       Standard
  Variable     DF      Estimate          Error    t Value    Pr > |t|
```

Intercept	1	3.00091	1.12530	2.67	0.0258
X2	1	0.50000	0.11796	4.24	0.0022

Dependent Variable: Y3

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27.47001	27.47001	17.97	0.0022
Error	9	13.75619	1.52847		
Corrected Total	10	41.22620			

Root MSE	1.23631	R-Square	0.6663
Dependent Mean	7.50000	Adj R-Sq	0.6292
Coeff Var	16.48415		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.00245	1.12448	2.67	0.0256
X3	1	0.49973	0.11788	4.24	0.0022

Dependent Variable: Y4

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27.49000	27.49000	18.00	0.0022
Error	9	13.74249	1.52694		
Corrected Total	10	41.23249			

Root MSE	1.23570	R-Square	0.6667
Dependent Mean	7.50091	Adj R-Sq	0.6297
Coeff Var	16.47394		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.00173	1.12392	2.67	0.0256
X4	1	0.49991	0.11782	4.24	0.0022

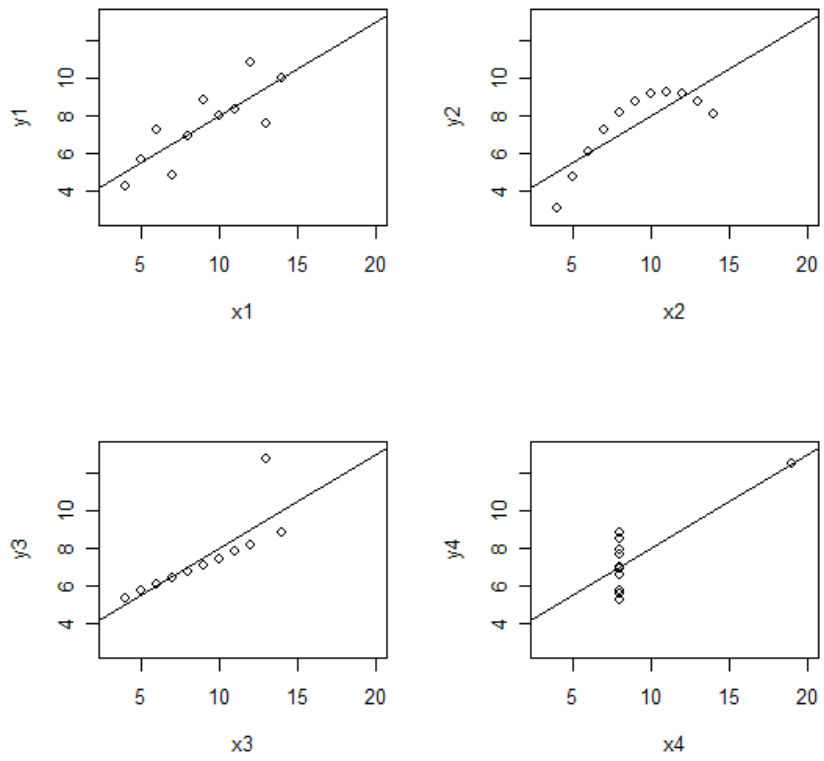

Figure 5.1: Plots of Anscombe's four data sets.

The regression output for four different data sets is almost identical in every respect. However, from Figure 5.1 it is obvious that a straight line regression model is appropriate only for Data set 1. On the other hand, the data in Data set 2 seem to have a curved line relationship rather than a straight line. The third data set has an extreme outlier that should be investigated and the fourth data set has a point that solely determines the slope of the regression line. This example demonstrates that the numerical regression output should always be supplemented by an analysis to ensure that an appropriate model has been fitted to the data. In this case it is sufficient to look at the scatter plots in Figure 5.1 to determine whether an appropriate model has been fit. However, when we consider situations in which there is more than one predictor, we shall need some additional tools in order to check the appropriateness of the fitted model.

5.2 Effects of violations in assumptions

- Linearity: Violations may occur when a proposed model may be inadequate (for example, the regression might contain some curvature) or when a proposed model may be appropriate for most of the data, but contamination from one or several outliers from different populations may render it inapplicable to the entire set. Both of these violations can cause the least squares estimates to give misleading answers to the questions of interest. Estimated mean and predictions can be biased (they systematically under or overestimate the intended quantity) and tests and CIs may inaccurately reflect uncertainty. The severity of the consequences depends on the severity of the violation.

- Constant variance: Although the least squares estimates are still unbiased even if the variance is nonconstant, the standard errors inaccurately describe the uncertainty in the estimates. Tests and confidence intervals can be misleading.

- Normality: Estimates of the coefficients and their standard errors are robust to nonnormal distribution. Although the tests and confidence intervals originate from normal distributions, the consequences of violating this assumption are usually minor. The only situation of substantial concern is when the distribution has long tails (outliers are present) and sample size is moderate to small. However, if prediction intervals are used, departures from normality become important. This is because the prediction intervals are based directly on the normality of the population distribution whereas confidence intervals are based on the normality of the sampling distributions of the estimates which may be approximately normal for large sample size even when the population distribution is not.

- Independence: Lack of independence causes no bias in least squares estimates of the coefficients, but standard errors are seriously affected.

5.3 Analysis of Residuals

One tool we will use to validate a regression model is one or more plots of residuals or studentized residuals. Such plots enable us to assess visually whether an appropriate model has been fitted to the data no matter how many predictors are used.

Residuals

Let us recall the definition of residuals. We have previously defined the residuals as

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n.$$

Since the residuals are measure of the variability in the observations not explained by the regression model, any departures from the assumptions on the errors should show up in

the residuals. Residual plot enable us to assess visually whether the assumptions are being violated and point to what should be done to overcome these violations.

- **Properties of the residuals**

  Recall that the residuals are given by

  $$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}.$$

  It can be shown that

  $$\mathrm{E}[\mathbf{e}] = \mathbf{0},\ \mathrm{Var}(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\sigma^2.$$

  Thus, the variance of the ith residual is

  $$\mathrm{Var}(e_i) = (1 - h_{ii})\sigma^2$$

  and the covariance between the residuals e_i and e_j is

  $$\mathrm{Cov}(e_i, e_j) = -h_{ij}\sigma^2,$$

  where h_{ii} is the ith diagonal element of the hat matrix $\mathbf{H}$ and h_{ij} is the (i, j)th element of the hat matrix.

  For a simple linear regression, it can be shown that

  $$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}, \quad h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}.$$

- **Residual plot**: Plot of e_i versus $\hat{y}_i$ (or x_i).

  One way to check whether a correct model has been fit is to plot residuals versus $\hat{y}$ (or x) and look for patterns. If no pattern is found then this indicates that the model provides an adequate summary of the data. If a pattern is found then the shape of the pattern provides information on the function of x that is missing from the model.

  See figure 4.3 on p.132 in textbook.

  Example (Anscombe's four data sets)

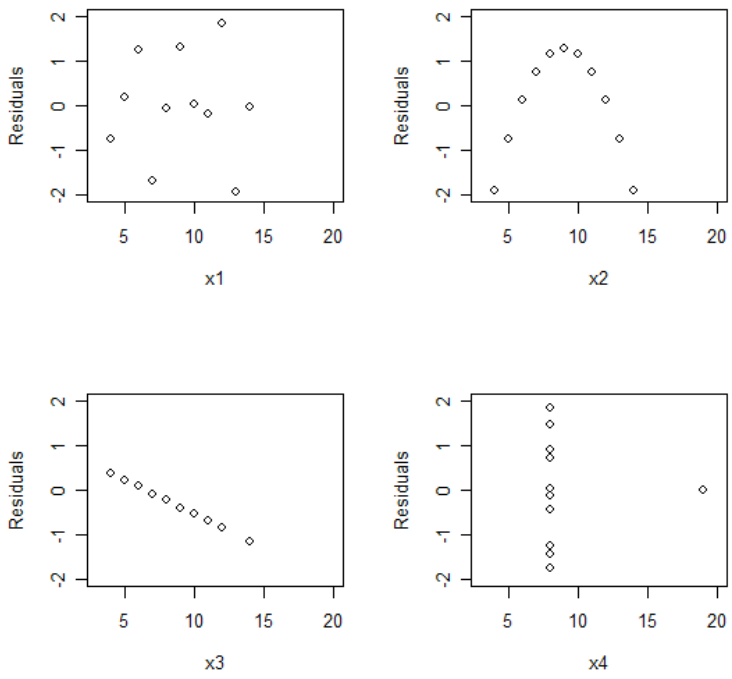

Figure 5.2: Residual plots for Anscombe's four data sets.

Figure 5.2 provides plots of the residuals versus x for each of Anscombe's four data sets. There is no discernible pattern in the plot of the residuals from Data set 1. This indicates that an appropriate model has been fitted to the data.

Studentized residuals

Notice that the residual does not have the same variance. To overcome the problem of the residuals having different variances, we will work with scaled residuals.

Studentized residuals:

$$r_i = \frac{e_i}{\text{se}(e_i)}, \quad i = 1, \ldots, n,$$

with $\text{se}(e_i) = \sqrt{MS_{Res}(1 - h_{ii})}$ the estimated standard error of the ith residual. Remark that r_i does not exactly follow a t distribution.

It is generally more informative to look at plots of studentized residuals rather than residual plots since the residual plots will have non-constant variance even if the errors have constant variance. The other advantage of studentized residuals is that they are helpful in finding outliers. Note that for moderately large n, the studentized residuals $r_i, i = 1, \ldots, n$ behaves

like independent observations from $N(0,1)$ if the model assumptions are correct. Thus, if the assumptions are not violated, then most of the studentized residuals r_i fall between -2 and 2 and the plot of r_i versus $\hat{y}_i$ (or x_i) produces a random pattern.

Other residuals

- Standardized residuals:
$$d_i = \frac{e_i}{\sqrt{MS_{Res}}}, \quad i = 1, \ldots, n.$$
The standardized residuals have mean zero and approximately unit variance. Consequently, a large standardized residual potentially indicates an outlier.

- PRESS residuals:
The following prediction errors are usually called PRESS residuals.
$$e_{(i)} = y_i - \hat{y}_{(i)} = \frac{e_i}{1 - h_{ii}}, \quad i = 1, \ldots, n,$$
where $\hat{y}_{(i)}$ is the fitted value of the ith response based on all observations except the ith one. If the ith observation is deleted, then $\hat{y}_{(i)}$ cannot be influenced by that observation, so the resulting residual should be likely to indicate the presence of the outlier.

- R-student:
$$t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1 - h_{ii})}}, \quad i = 1, \ldots, n,$$
where $S^2_{(i)} = \frac{1}{n-p-1} \sum_{j \neq i}^{n} (y_j - \hat{y}_{j(i)})^2$ with $\hat{y}_{j(i)}$ the fitted value of the jth response based on all observations except the ith one. Note that $S^2_{(i)}$ is an unbiased estimate of σ^2.

Read pp. 123–127 in textbook for the details.

Graphical methods for model assumption checks

1) Linearity

A graphical way to determine whether the proposed regression model is a valid model is to plot residuals versus $\hat{y}_i$ (or x_i) or plot the studentized residuals versus $\hat{y}_i$ (or x_i) and look for patterns.

See if there is any curvature. If the curve band or a nonlinear pattern show up then either higher order terms in x (e.g., polynomial terms) or a transformation should be considered.

See figure 4.3 on p.130 in textbook or Figure 5.3 below.

2) Normality

The assumption of normal errors is needed in small samples for the validity of t-distribution based hypothesis tests and confidence intervals and for all sample sizes for prediction intervals.

A common way to assess normality of the errors is to look at a normal probability plot or a normal Q-Q plot of the studentized residuals (r_i). A normal Q-Q plot of the studentized residuals is obtained by plotting the ordered studentized residuals on the vertical axis against the expected order statistics from a standard normal distribution on the horizontal axes. If the resulting plot produces points close to a straight line then the data are consistent with that from a normal distribution. On the other hand, departures from linearity provide evidence of non-normality.

Note that we will also use the Shapiro-Wilk test for normality test.

See if the points closely fall onto a straight line.

See figure 4.1 on p.130 in textbook.

3) Constant variance (Homogeneity)

Look at a plot of studentized residuals (r_i) versus the fitted values ($\hat{y}_i$).

See if scatter increases with fitted values. When the variance is found to be non-constant, we can consider transformations and generalized least squares/weighted least squares.

See figure 4.3 on p.130 in textbook or Figure 5.3 below.

An effective plot to diagnose non-constant error variance is a plot of $\sqrt{|r_i|}$ versus $\hat{y}_i$.

4) Independence

If the data are collected over time, we need to examine whether the data are correlated over time. When the independence is violated, we can consider correlations using time series models or generalized least squares.

See figure 4.6 at p.134 in textbook.

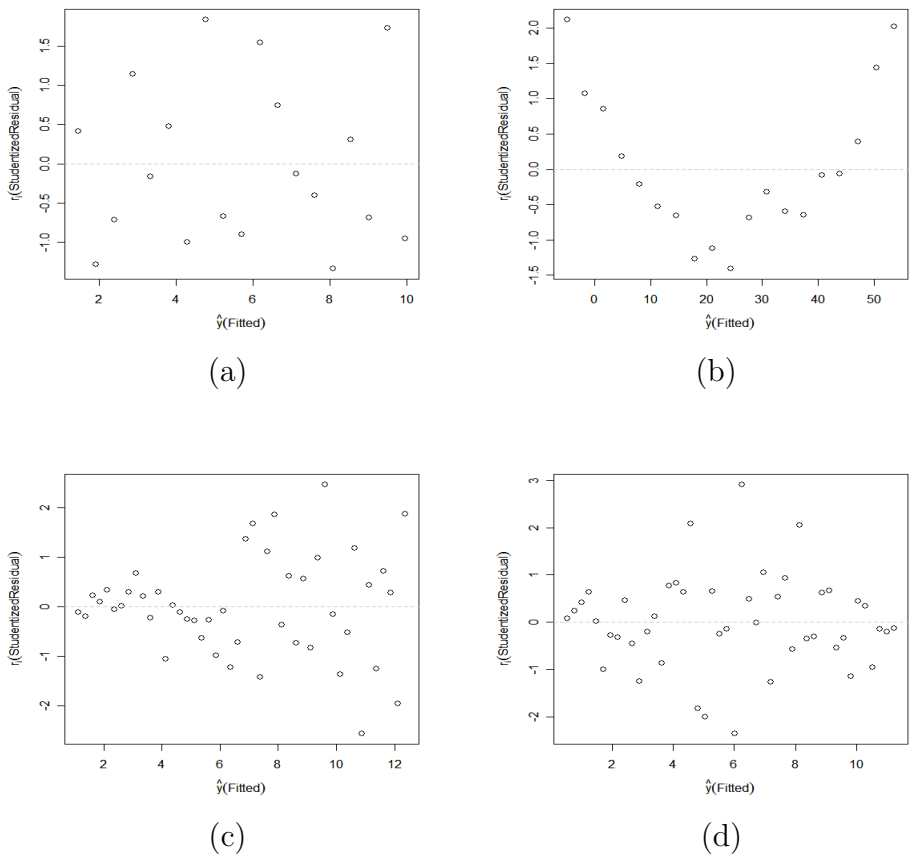

Figure 5.3: Patterns of plots of r_i versus $\hat{y}_i$: (a) ideal; (b) curvature; (c) funnel (heterogeneous variance); (d) double bow (heterogeneous variance).

Examples

- Rocket propellant data

  SAS Program:

```
proc reg data=rocket;
model Strength = Age;
plot r.*(pred. Age); /* plot of residuals vs fitted values(or X) */
plot student.*pred.; /* plot of studentized residuals vs fitted values */
plot student.*nqq.;  /* normal Q-Q plot of studentized residuals */
output out=out1 p=pred r=resid student=student;
run;
data out2;
set out1;
sqrtstudent= sqrt(abs(student));
```

```
run;
proc gplot data=out2;
plot sqrtstudent*pred; /* plot of |studentized residuals|^.5 vs fitted values */
run;
proc univariate data=out1 normaltest; /* normality test and normal Q-Q plot */
var student;
qqplot student / normal(mu=0 sigma=1) square;
run;
```

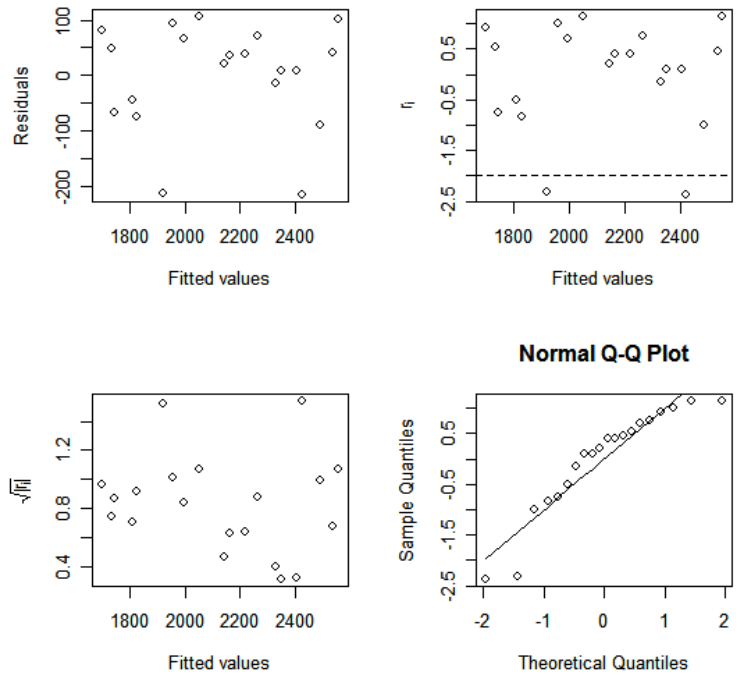

Figure 5.4: Plots of residual analysis for the rocket propellant data.

Tests for Normality

Test		--Statistic---		-----p Value------
Shapiro-Wilk	W	0.877153	Pr < W	0.0157
Kolmogorov-Smirnov	D	0.187144	Pr > D	0.0660
Cramer-von Mises	W-Sq	0.118989	Pr > W-Sq	0.0592
Anderson-Darling	A-Sq	0.798432	Pr > A-Sq	0.0333

Normality assumption is suspected. This indicates that there may be some problems with the normality assumption or that there may be outliers in the data.

- Delivery time data

  SAS Program:

```
option ls=75 ps=80;
title 'Delivery Time Data';
proc reg data=delivery;
model Time = Cases Distance;
plot student.*(pred. Cases Distance);
plot student.*nqq.;
output out=out1 r=resid p=pred student=student;
run;
proc univariate data=out1 normaltest;
var student;
qqplot student / normal(mu=0 sigma=1) square;
run;
```

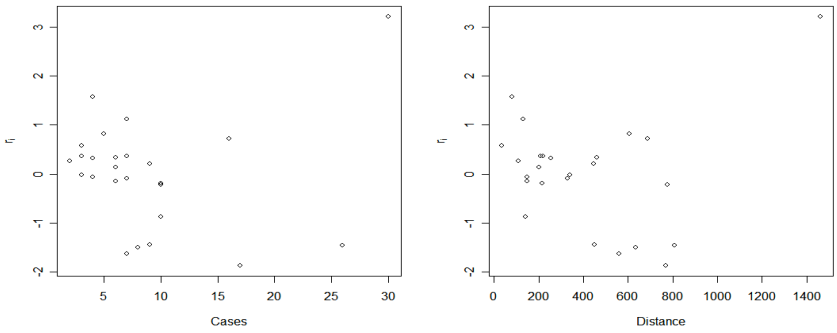

Figure 5.5: Plots of studentized residuals against each predictor for the delivery time data.

Neither of these plots reveals any clear indication of a problem with misspecification of the predictor (implying the need for either a transformation on the predictor or higher order terms in cases and/or distance), although the moderately large residual associated with point 9 is apparent on both plots.

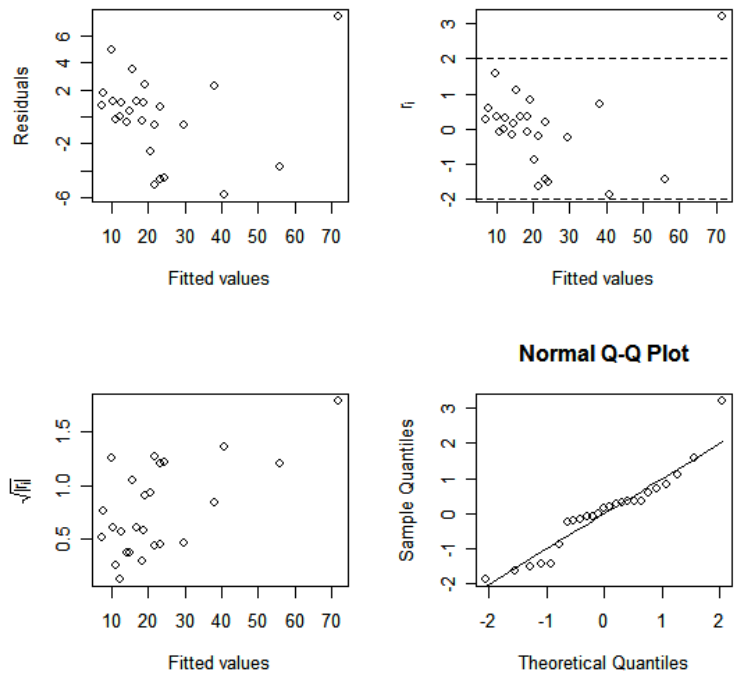

Figure 5.6: Plots of residual analysis for the delivery time data.

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.922851	Pr < W	0.0595
Kolmogorov-Smirnov	D	0.173145	Pr > D	0.0510
Cramer-von Mises	W-Sq	0.126832	Pr > W-Sq	0.0462
Anderson-Darling	A-Sq	0.744467	Pr > A-Sq	0.0464

It does seem to be a slight tendency for the model to underpredict short delivery times and overpredict long delivery times. The counterproductive influence of outliers is suspected. This indicates that the detection of outliers and/or transformation are considered.

- Hardwood Data

  SAS Program:

```
proc reg data=hardwood;
model Strength = Concentration;
plot Strength*Concetration;
plot student.*(pred. Concentration);
plot student.*nqq.;
run;
proc reg data=hardwood;
model Strength = Concentration Concentration2;
plot Strength*Concetration;
plot student.*pred.;
plot student.*nqq.;
output out=out1 p=pred r=resid student=student;
run;
data out2;
set out1;
sqrtstudent=sqrt(abs(student));
run;
proc gplot data=out2;
plot sqrtstudent*pred;
run;
proc univariate data=out1 normaltest;
var student;
qqplot student / normal(mu=0 sigma=1) square;
run;
```

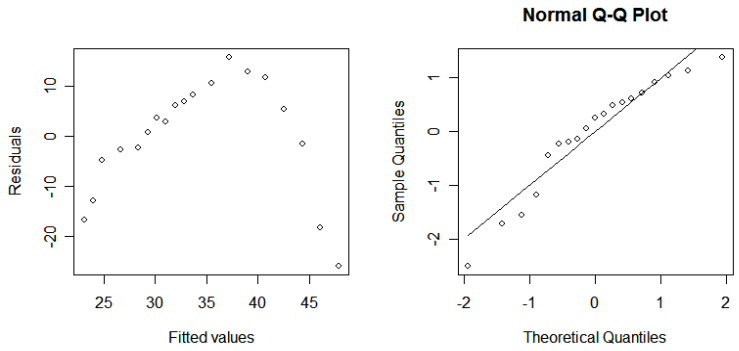

Figure 5.7: Plots of residual analysis of the simple linear regression model for the hardwood data.

There is clear curvature pattern in the residual plot and so adding a quadratic term to the model will be helpful.

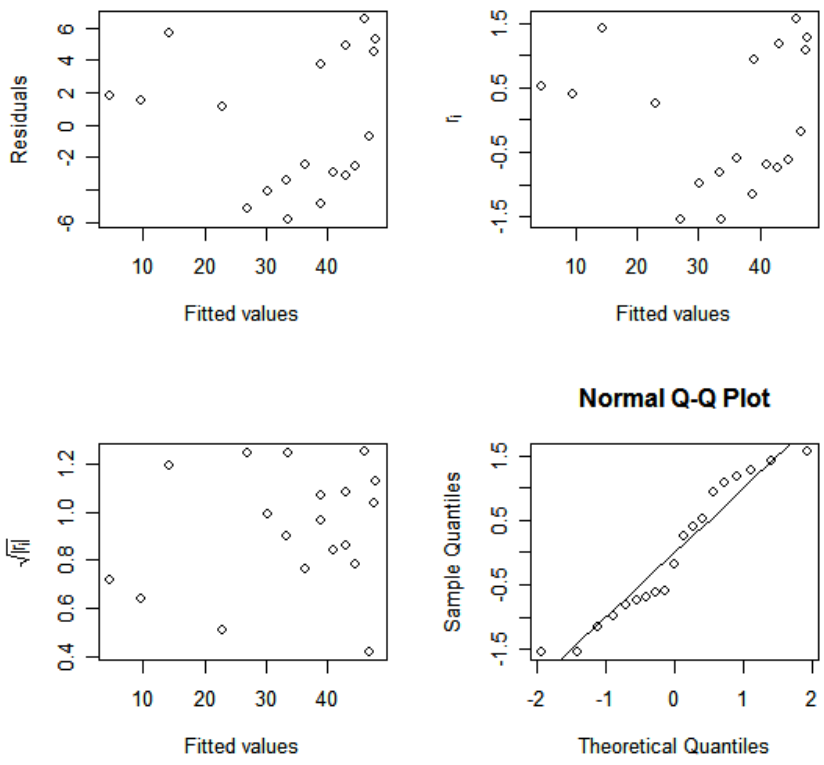

Figure 5.8: Plots of esidual analysis of the polynomial regression model for the hardwood data.

Tests for Normality

Test		--Statistic---		-----p Value------
Shapiro-Wilk	W	0.922993	Pr < W	0.1286
Kolmogorov-Smirnov	D	0.183638	Pr > D	0.0898
Cramer-von Mises	W-Sq	0.089797	Pr > W-Sq	0.1455
Anderson-Darling	A-Sq	0.538461	Pr > A-Sq	0.1485

There is no clear pattern. All the assumptions are met.

- Nutritional requirement data

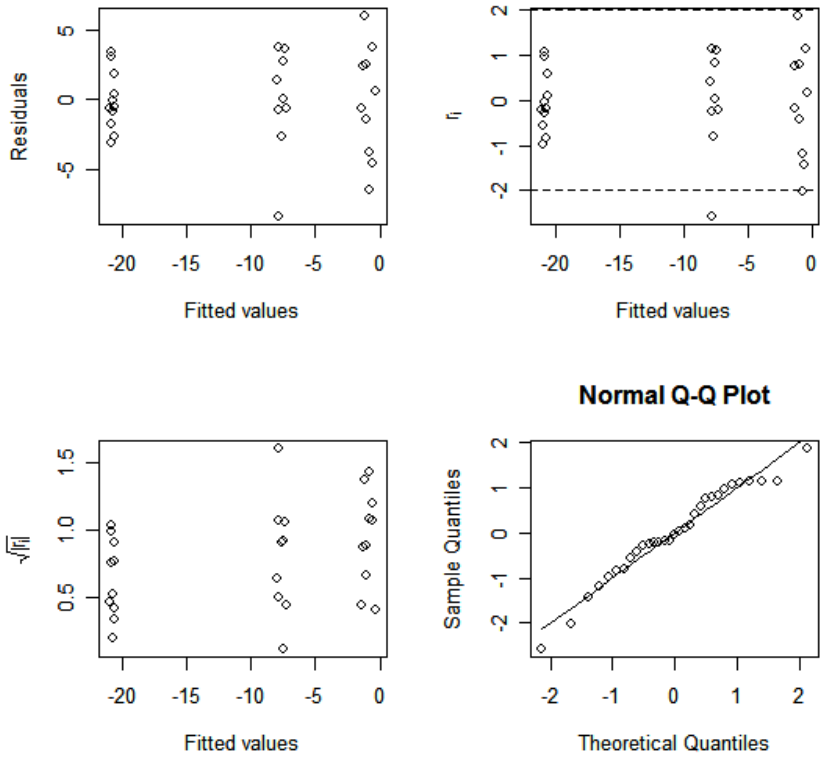

Figure 5.9: Plots of residual analysis for the nutritional requirement data.

Tests for Normality

Test		--Statistic---		-----p Value------
Shapiro-Wilk	W	0.9645	Pr < W	0.3817
Kolmogorov-Smirnov	D	0.104308	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.055487	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.386262	Pr > A-Sq	>0.2500

All the assumptions are met.

Detection of Outliers

Outliers: Data points which are not typical of the rest of the data. Data points with large residuals are outliers.

Look at a plot of studentized residuals versus the fitted values. Then, look for extreme values (outliers). An outlier is a point whose studentized residual falls outside the interval from -2 to 2. We shall follow the common practice of labeling points as outliers in small to moderate size data sets if the studentized residual for the point fall outsize of the interval from -2 to 2. In very large data sets, we shall change this rule to -4 to 4. (Otherwise, many points will be flagged as potential outliers.)

Examples

- Rocket propellant data.

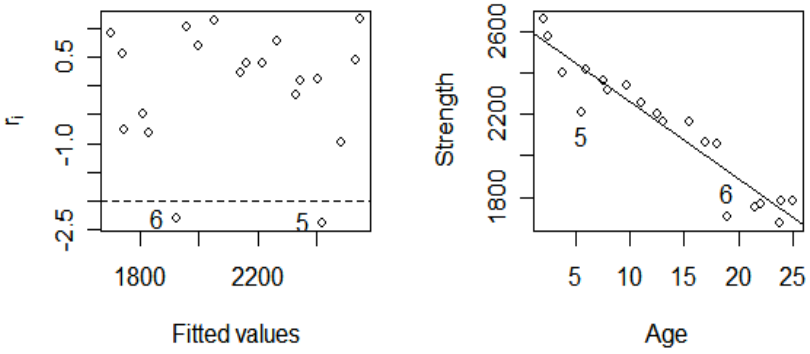

Figure 5.10: A plot of the studentized residuals against the fitted values for the rocket propellant data.

From the above plot, two rocket propellants whose ages are 5.5 and 19 weeks are suspected as outliers.

- Delivery time data.

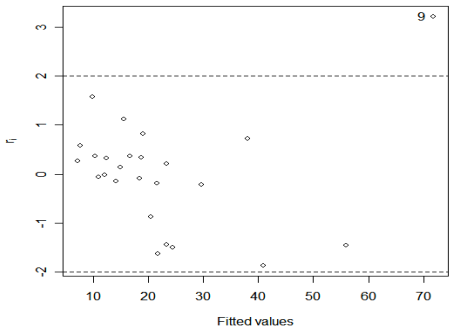

Figure 5.11: A plot of the studentized residuals against the fitted values for the delivery time data.

There is one outlier in the data. We know that the studentized residual for observation 9 is large ($r_9 = 3.2138$).

- Nutritional requirement data.

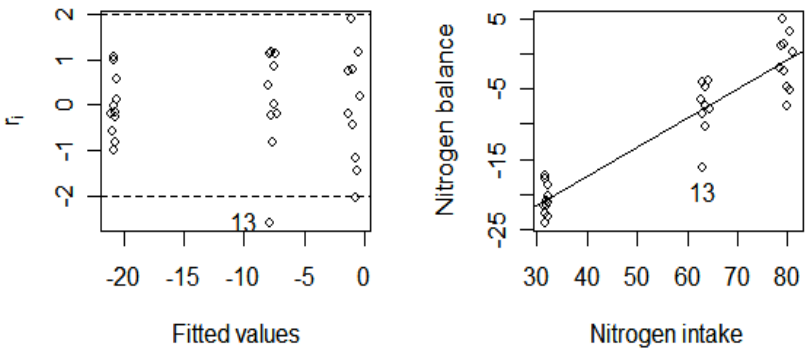

Figure 5.12: A plot of the studentized residuals against the fitted values for the nutritional requirement data.

Recommendations for handling outliers

- Outliers, when detected, should be thoroughly investigated before any action is taken.
- Legitimate deletion of invalid observations is sensible, but it should not be done routinely just because they do not fit the model.

- When a point is deleted from an analysis, it should be noted and justified.
- Outliers may occur for several reasons. These include transcription errors, coming from a different group, or not fitted by the model under consideration.
- Outliers often point out an important feature of the problem not considered before. They are very informative in this way. They may point to an alternative model on which the points are not an outlier. In this case it is then worth considering fitting an alternative model.
- If there is no reason to delete outliers, then robust methods (nonparametric statistics or robust regression) have to be considered to reduce the effects of outliers.

  Read section 4.4 on pp.142–144 in textbook. Refer to Chapter 12 in textbook for robust regression.

Added variable plots

The residual plot (a plot of residuals versus a predictor) is useful in determining whether a curvature effect for that predictor is needed in the model. A limitation of these plots is that they may not completely show the correct or complete marginal effect of a predictor, given the other predictors in the model.

Added variable plots (called partial regression plots, partial leverage plots or adjusted variable plots) enable us to visually assess the effect of each predictor on the response variable, having adjusted for the effect of other predictors. This plot can be very useful in evaluating whether we have specified the relationship between the response variable and the predictors correctly. In added-variable plot, the response variable Y and the predictor X_j are both regressed against the other preidctors in the model and the residuals obtained for each regression.

Recall the multiple linear regression model in Chapter 3

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon.$$

Then, regress Y on $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$; i.e., fit the model

$$Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{j-1} X_{j-1} + \alpha_{j+1} X_{j+1} + \cdots + \alpha_k X_k + u,$$

and obtain the fitted values and residuals:

$$\hat{y}_i(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k) = \hat{\alpha}_0 + \hat{\alpha}_1 x_{i1} + \cdots + \hat{\alpha}_{j-1} x_{i,j-1} + \hat{\alpha}_{j+1} x_{i,j+1} + \cdots + \hat{\alpha}_k x_{ik},$$
$$e_i(Y|X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k) = y_i - \hat{y}_i(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k), \quad i = 1, \ldots, n.$$

Now regress X_j on $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$. That is, fit the model

$$X_j = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_{j-1} X_{j-1} + \gamma_{j+1} X_{j+1} + \cdots + \gamma_k X_k + v.$$

Then, obtain the fitted values and the residuals:

$$\hat{x}_{ij} \equiv \hat{x}_{ij}(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k) = \hat{\gamma}_0 + \hat{\gamma}_1 x_{i1} + \cdots + \hat{\gamma}_{j-1} x_{i,j-1} + \hat{\gamma}_{j+1} x_{i,j+1} + \cdots + \hat{\gamma}_k x_{ik},$$
$$e_i(X_j | X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k) = x_{ij} - \hat{x}_{ij}, \quad i = 1, \ldots, n.$$

The added-variable plot for a predictor X_j is obtained by plotting on the vertical axis the residuals $e_i(Y|X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k)$ against on the horizontal axis the residuals $e_i(X_j|X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k)$. Notice that the effects due to $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$ are removed from both axes in the added-variable plot for predictor X_j.

If the predictor X_j enters the model linearly, then the added-variable plot should fall along a straight line with a nonzero slope. If the added-variable plot shows a curvilinear band, then higher order terms in X_j or a transformation may be helpful. When X_j is a candidate variable being considered for inclusion in the model, a horizontal band on the added-variable plot indicates that there is no additional useful information in X_j for predicting Y.

The plots in Figure 5.13 show the situations with curvilinear band in X_1 and no additional useful information in X_2.

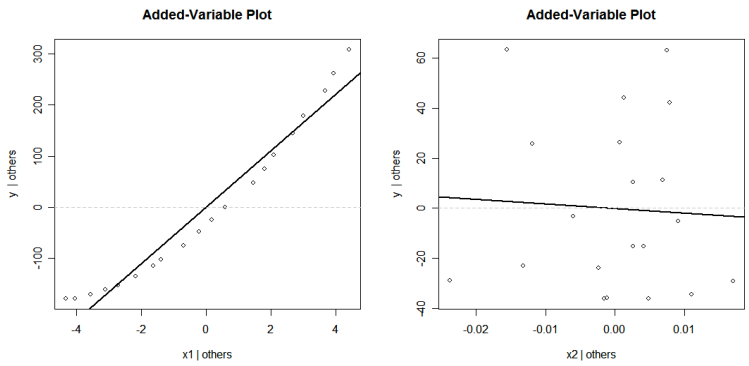

Figure 5.13: Examples of added-variable plots.

Example

Revisit the delivery time data.

SAS Program:

```
proc reg;
model Time = Cases Distance / partial; /* added-variable plots */
run;
```

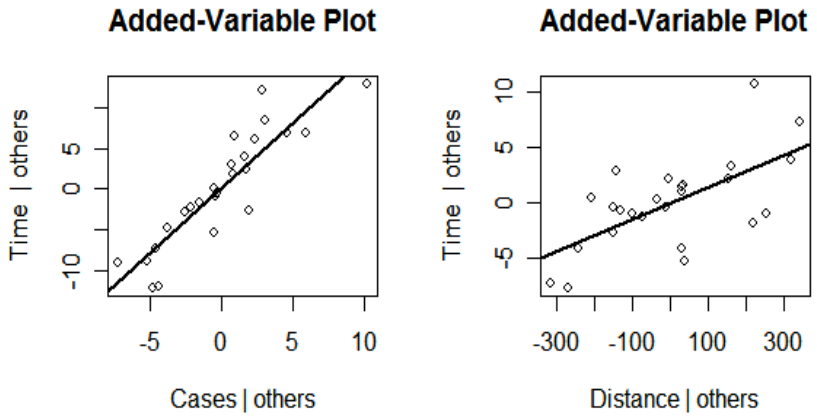

Figure 5.14: Added-variable plots for the delivery time data.

The added-variable plots show that the predictors Cases and Distances are clearly evident in the model.

- Mathematical justification for added-variable plots

  We will use matrix notation for this. For $j = 1, \ldots, k$, let

$$\mathbf{x}_j = \begin{pmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{n,j} \end{pmatrix}, \quad \mathbf{X}_{(j)} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,j-1} & x_{1,j+1} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2,j-1} & x_{2,j+1} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,j-1} & x_{n,j+1} & \cdots & x_{nk} \end{pmatrix}$$

  and

$$\mathbf{e}_{\mathbf{y}|\mathbf{X}_{(j)}} = (e_1(Y|X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k), \ldots, e_n(Y|X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k))^T,$$
$$\mathbf{e}_{\mathbf{x}_j|\mathbf{X}_{(j)}} = (e_1(X_j|X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k), \ldots, e_n(X_j|X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k))^T.$$

  Then, consider the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_{(j)}\boldsymbol{\beta}_{(j)} + \beta_j \mathbf{x}_j + \boldsymbol{\epsilon}.$$

  Premultiplying by $\mathbf{I} - \mathbf{H}_{(j)}$ results in

$$(\mathbf{I} - \mathbf{H}_{(j)})\mathbf{y} = \beta_j(\mathbf{I} - \mathbf{H}_{(j)})\mathbf{x}_j + (\mathbf{I} - \mathbf{H}_{(j)})\boldsymbol{\epsilon}$$

  since $(\mathbf{I} - \mathbf{H}_{(j)})\mathbf{X}_{(j)} = \mathbf{0}$. Consequently,

$$\mathbf{e}_{\mathbf{y}|\mathbf{X}_{(j)}} = \beta_j \mathbf{e}_{\mathbf{x}_j|\mathbf{X}_{(j)}} + \boldsymbol{\epsilon}^*$$

  where $\boldsymbol{\epsilon}^* = (\mathbf{I} - \mathbf{H}_{(j)})\boldsymbol{\epsilon}$.

  Notice that the slope of this line will be the regression coefficient of X_j in the multiple linear regression model.

5.4 Lack of fit of the regression model

The lack-of-fit test is a formal statistical test for testing linearity *when we have replicate observations on the response variable for at least one level of the predictor*. The procedure assumes that the normality, independence, and constant variance requirements are met and that only the straight line of the relationship is in doubt.

The lack-of-fit test is testing the hypotheses

$$H_0 : \mathrm{E}[Y|X_1 = x_1, \ldots, X_k = x_k] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \quad \text{versus} \quad H_1 : \text{ not } H_0.$$

Suppose that we have n_i observations on the response at the ith level of the predictor $x_i, i = 1, \ldots, m$. Let y_{ij} denote the jth observation on the response at x_i, $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$. Then, $n = \sum_{i=1}^{m} n_i$. The test procedure involves partitioning the residual sum of squares into two components, say

$$SS_{Res} = SS_{PE} + SS_{LOF},$$

where $SS_{PE} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ is the sum of squares due to pure error and $SS_{LOF} = \sum_{i=1}^{m} n_i (\bar{y}_i - \hat{y}_i)^2$ is the sum of squares due to lack of fit.

Under the assumptions of normality, homogeneity and independence, note that

$$\frac{SS_{PE}}{\sigma^2} \sim \chi^2_{n-m}, \quad \frac{SS_{LOF}}{\sigma^2} \sim \chi^2_{m-p}.$$

Thus, the test statistic for lack of fit is

$$F_0 = \frac{SS_{LOF}/(m-p)}{SS_{PE}/(n-m)} \sim F_{m-p,n-m}$$

when H_0 is true. Therefore, we conclude that the regression function is not linear if $F_0 > F_{\alpha,m-p,n-m}$.

Example

Recall the nutritional requirement data in Chapter 1.

SAS Program

```
option ls=75 ps=80;
title 'Nutritional Requirement Data';
data nutrition;
infile 'nutrition.dat';
input kcal ni niq balance;
```

```
run;
proc rsreg;
model balance = ni / lackfit covar=1; /* Lack-of-fit test */
run;
```

Output

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	22	221.619516	10.073614	0.71	0.7501
Pure Error	7	99.625000	14.232143		
Total Error	29	321.244516	11.077397		