Multiple
Linear Regression

The Multiple Linear Regression Model

Multiple linear regression model

- employs at least two regressor variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + \varepsilon_i ,$$

$$i = 1, 2, \ldots , n$$

where:

Y_i :value of the response variable in the ith trial ;
$\beta_0,\beta_1,\beta_2,\ldots,\beta_k$:the parameters of the model ;
X_{1i} :value of the 1st regressor variable in the ith trial ;
X_{2i} :value of the 2nd regressor variable in the ith trial ;

.
.
.

X_{ki} :value of the kth regressor variable in the ith trial ;
ε_i :random error term in the ith trial.

Example:

Consider the dependent variable *selling price of house* (PRICE) as a function of the independent variable *floor space* (FLR) and another independent variable, say *garage size* (GAR), which can also be used to predict PRICE.

$$\text{PRICE} = \beta_0 + \beta_1 \text{ FLR} + \beta_2 \text{ GAR} + \varepsilon \, .$$

Assumptions of the multiple linear regression model:

1. For the ith trial, the error component εi has a Normal distribution with mean 0 and variance σ^2.

2. The error components in any pair of trials, say the ith and the jth, are independent.

3. The terms β_0, β_1, β_2, …, and βk in the model are parameters whose values are typically unknown and must therefore be estimated from the sample data.

4. The *k* regressor variables X_1, X_2, …, X_k are considered to be known constants that are fixed or pre-chosen.

Remark: As much as possible, values of the regressor variables should not be highly correlated to avoid model-fitting problems.

The consequences of these assumptions are:

1. The regression function corresponding to the multiple linear regression model is

   $$E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + . . . + \beta_k X_{ki} .$$

2. The variance of the probability distribution of Y for a given value of X is σ^2.
3. The observed value of Y in the ith trial is larger or smaller than its mean by the amount ε_i, the value of the error component in the ith trial.
4. The outcome in any trial is neither affected by nor affects the error term in any other trial.

The Regression Coefficients

Consider the estimated model given by

$$\text{PRICE} = 33.705 + 0.01694\ \text{FLR} + 4.5048\ \text{GAR}\ .$$

- the value 33.705 represents the estimated mean selling price of the house when both floor space and garage size are set to 0 ;
- 0.01694 represents the estimated increase in the mean selling price when FLR is increased by 1 square foot while holding the value of GAR fixed ;
- 4.5048 represents the estimated increase in the mean selling price when GAR is increased by 1 unit while holding the value of FLR fixed.

Note: The value 33.705 is meaningless because the regression model is relevant only for the range of values of both FLR and GAR. The values of FLR range from 596 to 2261 and the values of GAR from 0 to 2; therefore, we do not interpret the value 33.705.

In general, for the *estimated multiple linear regression model* of the form

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} ,$$

- b_0 is the estimated mean of Y when the values of all regressor variables X_1, X_2, X_3, …, X_k are set to 0 ;

- b_1 is the estimated increase or decrease in the mean of Y for every unit increase in the value of X_1 holding the values of X_2, X_3, … and X_k fixed ;

- b_2 is the estimated increase or decrease in the mean of Y for every unit increase in the value of X_2 holding the values of X_1, X_3, … and X_k fixed ;

  .

  .

  .

- b_k is the estimated increase or decrease in the mean of Y for every unit increase in the value of X_k holding the values of X_1, X_2, … and X_{k-1} fixed .

Note: The value of b_0 is relevant only if zero is within the range of values of all regressor variables.

The Analysis of Variance (ANOVA)

Analysis of Variance method

- decomposes the variation in the response variable Y into explained and unexplained variation

Explained variability = amount of variability in the response variable that can be attributed to the set of k regressor variables

Unexplained variability = variability in the response variable attributed to error

For the selling price data,

Table 2.1. Analysis of variance corresponding to the *Selling Price* data

Effect	Sum of Squares	df	Mean Squares	F-value	p-level
Regression	2429.729	2	1214.864	18.29069	.000018
Residual	1527.656	23	66.420		
Total	3957.385				

The value 2429.729 represents the amount of variation inselling price (PRICE) explained by the variables FLR and GAR. The value 1527.656 represents the unexplained variation in PRICE. And the value 3957.385 represents the total variation in PRICE, which is simply the sum of the explained and unexplained parts.

In general, the ANOVA table is given by

Effect	Sum of Squares	df	Mean Squares	F-value
Regression	SSR	k	MSR = SSR/ k	F = MSR/MSE
Error	SSE	n-(k+1)	MSE = SSE/(n-k-1)	
Total	TSS	n-1		

where:

SSR = sum of squares due to regression; represents the amount of variability in the response variable that can be explained by the set of k regressor variables in the model ;

SSE = sum of squares due to error; represents the amount of variability in the response variable that can no longer be explained by the linear relationship of Y with the set of k regressor variables ;

TSS = total sum of squares; measures the total variability in the response variable ;

k=number of regressor variables ;

MSR=mean squares due to regression ;

MSE=mean squares due to error ;

n=number of observations/trials .

The Coefficient of Multiple Determination (R^2)

- a general measure of the goodness-of-fit of the model to the sample data
- denoted by R^2
- computed as:

$$R2 = \frac{SSR}{TSS} .$$

- ■ is normally expressed as a percentage
- ■ is interpreted as the amount of variability in the response variable that can be explained by the set of *k* regressor variables

In the *selling price* example, R^2 is given by

$$R^2 = \frac{2429.729}{3957.385} = .6140 \ .$$

Thus, we say that 61.40% of the variability in selling price of house can be explained by the variability in both floor space and garage size.

The Adjusted R^2

- has the same interpretation as the R^2 but is used for comparing the goodness-of-fit of different regression models with varying numbers of regressor variables

$$\text{adjusted } R^2 = 1 - \frac{SSE/(n-k-1)}{TSS/(n-1)} .$$

In the *selling price* example, the adjusted R^2 is given by

$$\text{adjusted } R^2 = 1 - \frac{1527.656 / 23}{3957.385 / 25} = .5804$$

An estimated regression model with a higher adjusted R^2 is said to have better fit than models having lower adjusted R^2 values.

The F – Test

The F-test tests whether:

Ho: *all parameters (except β_0) are equal to zero*

vs.

Ha: *there is at least one parameter which is not equal to zero* .

Referring to the ANOVA table in page 37, the F-statistic tests whether:

Ho: $\beta_1 = \beta_2 = 0$

vs.

Ha: *at least one β differs from zero* .

The F-statistic of 18.291, with a corresponding p-value of 0.000018, leads us to reject the null hypothesis at a level of significance α = .05. Thus, we can say that *either* FLR *or* GAR or *both* have a significant *linear relationship* with PRICE.

In general, the F-test tests the hypothesis

Ho: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$

vs.

Ha: *at least one* β *differs from zero .*

If the null hypothesis is not rejected,

- conclude that no regressor variable has a significant linear relationship with the response variable

If the null hypothesis is rejected,

- conclude that there is at least one regressor variable that has a significant linear relationship with the response variable

- proceed to the individual *t*-tests to identify which among the regressor variables can explain a significant amount of variability in the response variable

The *t* – Tests

Table 2.2. Results of the t-Tests Performed for the Regression Model

$\text{PRICE} = \beta0 + \beta1\text{FLR} + \beta2\text{GAR} + \varepsilon$

		BETA	St. Err. Of BETA	B	St. Err. Of B	t(23)	p-level
①	Intercept			33.70503	4.2296	7.9689	.00000
②	FLR	.61678	.141685	.01694	.0039	4.3532	.00023
③	GAR	.29425	.141685	4.50481	2.1691	2.0768	.04917

The *t*-tests above test three different hypotheses:

① H_o: $\beta_0 = 0$ vs. H_a: $\beta_0 \neq 0$

② H_o: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$

③ H_o: $\beta_2 = 0$ vs. H_a: $\beta_2 \neq 0$.

For each of the tests above, using a level of significance of .05, the corresponding p-value tells us to

① *Reject* H_o: $\beta_0 = 0$

② *Reject* H_o: $\beta_1 = 0$

③ *Reject* H_o: $\beta_2 = 0$.

- Tests ② and ③ above imply that the variables FLR and GAR have a significant contribution in predicting the mean selling price of houses.

- Test ① implies that the Y-intercept is significantly different from zero.

Thus, the estimated regression model can be written as

$$\text{PRICE}_i = 33.70503 + 0.01694\ \text{FLR}_i + 4.50481\ \text{GAR}_i\ .$$

In general, the individual *t*-tests test the hypotheses

$$H_o: \quad \beta_0 = 0 \qquad vs. \qquad H_a: \quad \beta_0 \neq 0$$
$$H_o: \quad \beta_1 = 0 \qquad vs. \qquad H_a: \quad \beta_1 \neq 0$$
$$H_o: \quad \beta_2 = 0 \qquad vs. \qquad H_a: \quad \beta_2 \neq 0$$

.

.

.

$$H_o: \quad \beta_k = 0 \qquad vs. \qquad H_a: \quad \beta_k \neq 0 \ .$$

- If any of these tests leads to the acceptance of the null hypothesis, then we say that the corresponding regressor variable does not have a significant contribution in predicting the mean of the response variable

- If any of the tests above leads to the rejection of the null hypothesis, then we say that the corresponding regressor variable has a significant contribution in predicting the mean of the dependent variable.

Regression Diagnostics and Residual Analysis

The regression diagnostics applied in the simple linear regression model is the same as that for the multiple linear regression model. In addition, we test for the linear dependencies among the regressor variables.

In summary, we check for:

- linearity of each regressor variable with the response variable
- constancy of error variance
- outlying observations
- Normality of the error terms
- multicollinearity

Linearity

Scatterplots can be made for the response against each of the regressor variables.

Figure 2.1 shows that there exists a positive relationship between FLR and PRICE; that the points can be summarized by a straight line.

Figure 2.2 suggests that there is a slight positive relationship between GAR and PRICE. Likewise, the points can be summarized by a straight line.

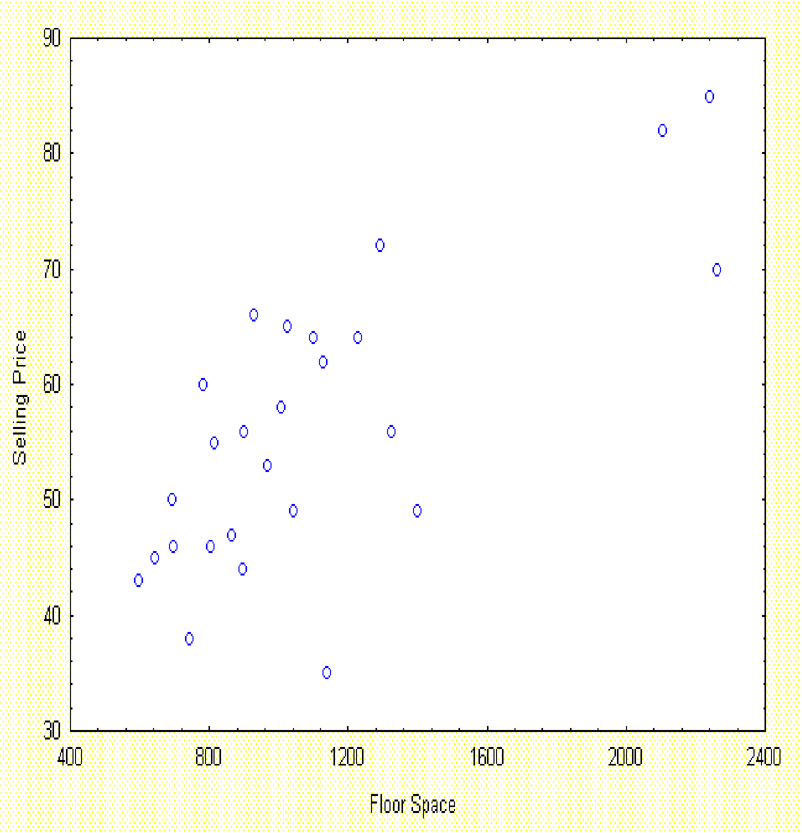

Figure 2.1. PRICE vs FLR

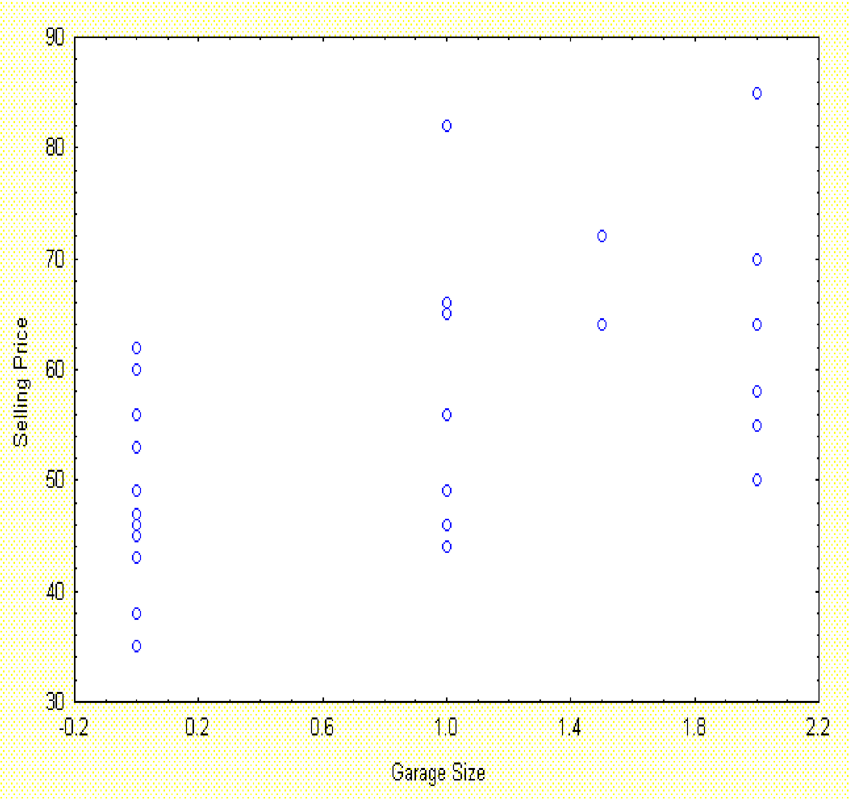

Figure 2.2. PRICE vs GAR

Note:

Because we are looking only at the relationship of one of the regressor variables with the response variable, any departure from a straight line does not imply that the multiple linear regression model is not appropriate for the data.

If any of the regressor variables do not seem to be linearly related with the response variable, transform either the response variable or the regressor variable

Possible transformations

$\log Y$ and/or $\log X$

$\sqrt{Y}$ and/or $\sqrt{X}$

$\sqrt[3]{Y}$ and/or $\sqrt[3]{X}$

${}^{1}\!/\!{}_{Y}$ and/or ${}^{1}\!/\!{}_{X}$

Constancy of Error Variance

For the multiple linear regression model, the plot of the residuals versus the predicted values should form a horizontal band around zero which should show neither an increasing nor a decreasing trend.

The figure below gives the residual plot for the *selling price* data. This plot shows that the residuals are randomly scattered around the mean value of zero.

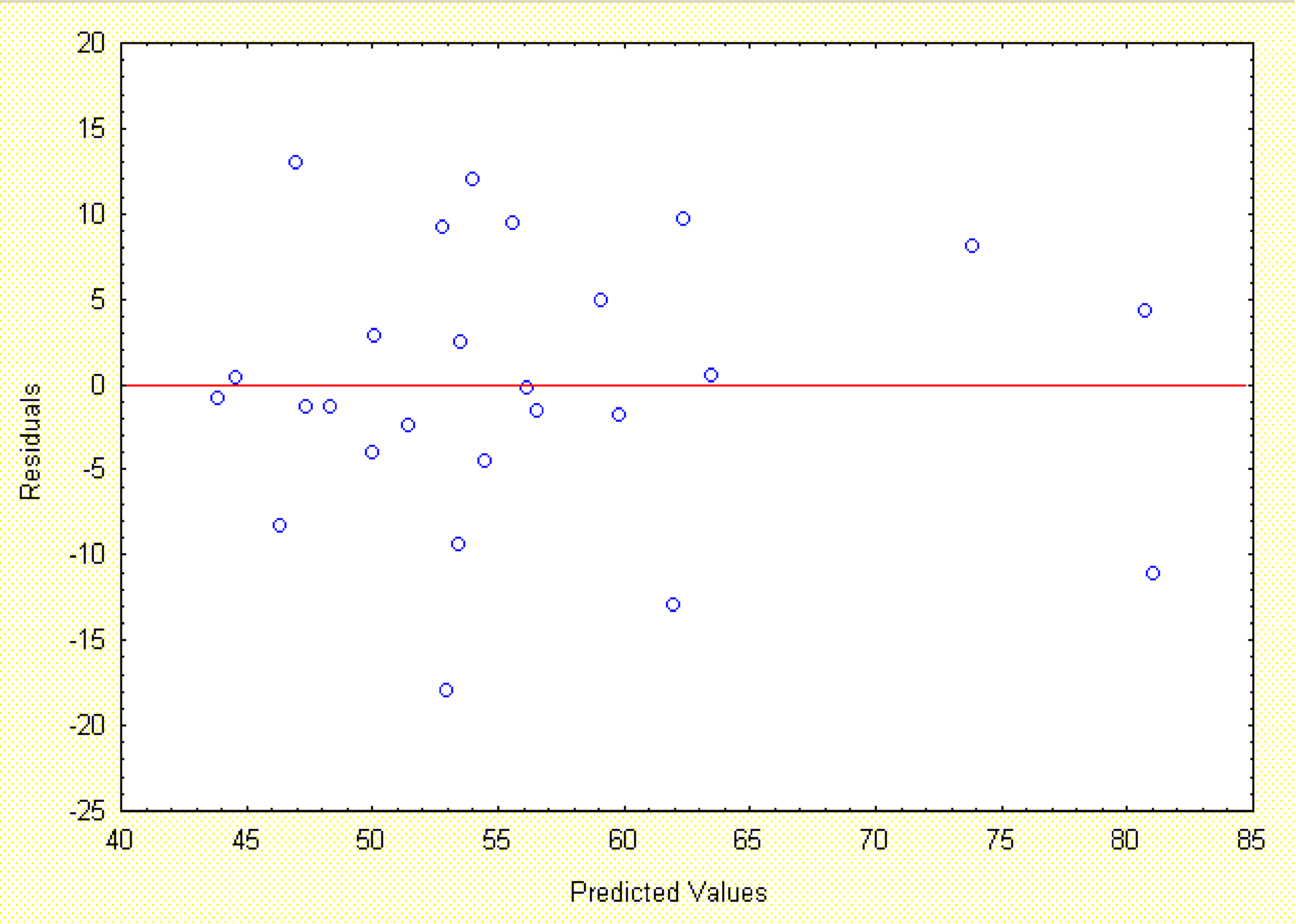

Figure 2.3. Scatterplot of the residuals vs. the predicted values

Normality of the Error Terms

A frequency histogram can be constructed to assess the Normality of the residuals. The shape of the histogram should approximate that of a Normal distribution.

The figure below gives the frequency histogram of the residuals for the *selling price* data. Superimposed on this histogram is a plot of a Normal distribution with a mean of zero and a variance equal to 66.420, the mean square error. The plot suggests that the distribution of the residuals tends to follow the Normal curve.

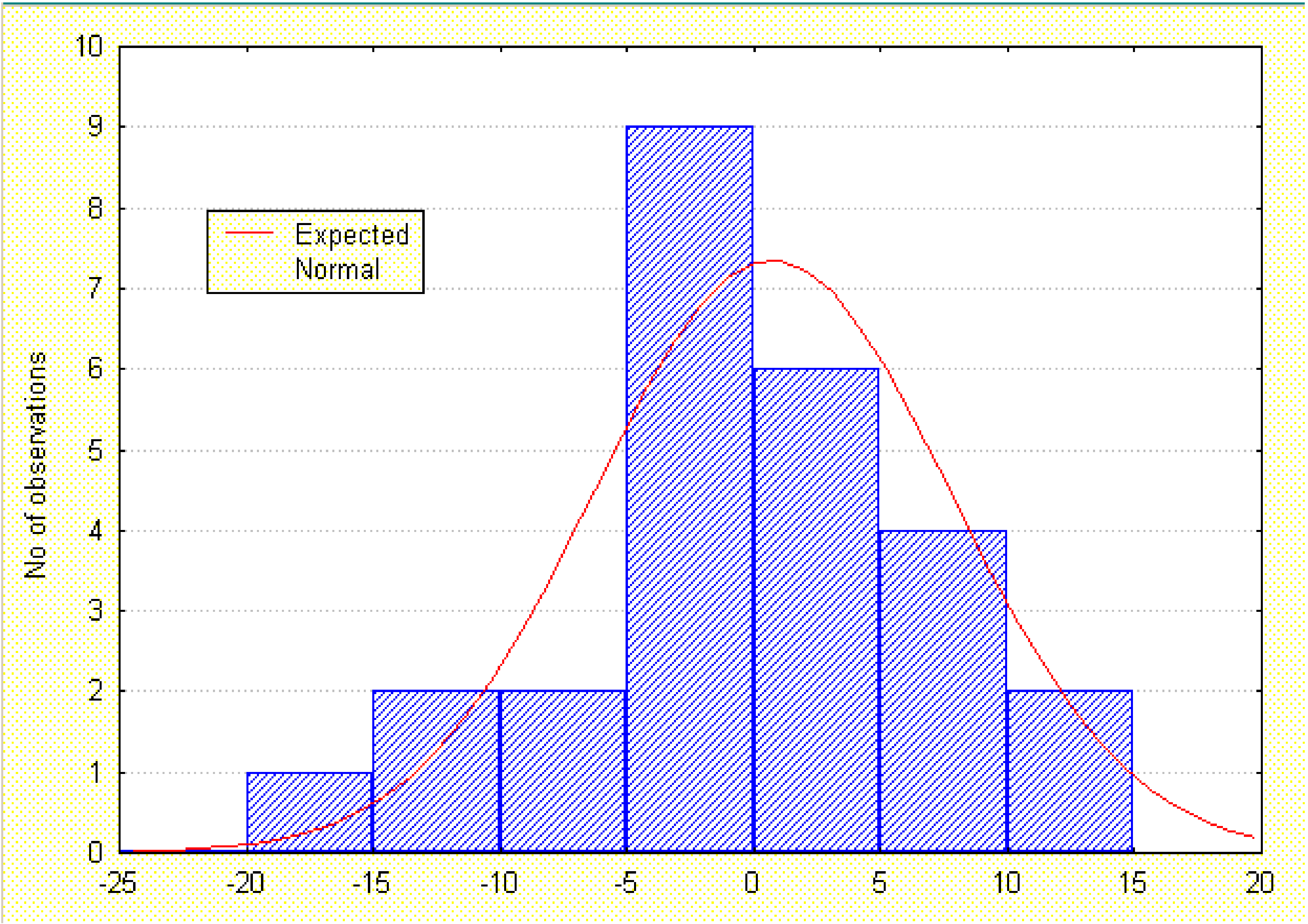

Figure 2.4. Frequency histogram of the residuals

Another graph that can be used to detect non-normality of the residuals is the Normal probability plot. The interpretation of the Normal probability plot is the same as in the simple linear regression case.

Below is the Normal probability plot of the residuals for the *selling price* data. The plot shows that the residuals follow the expected percentiles of the Normal distribution. Thus, we say that the residuals follow a Normal distribution.

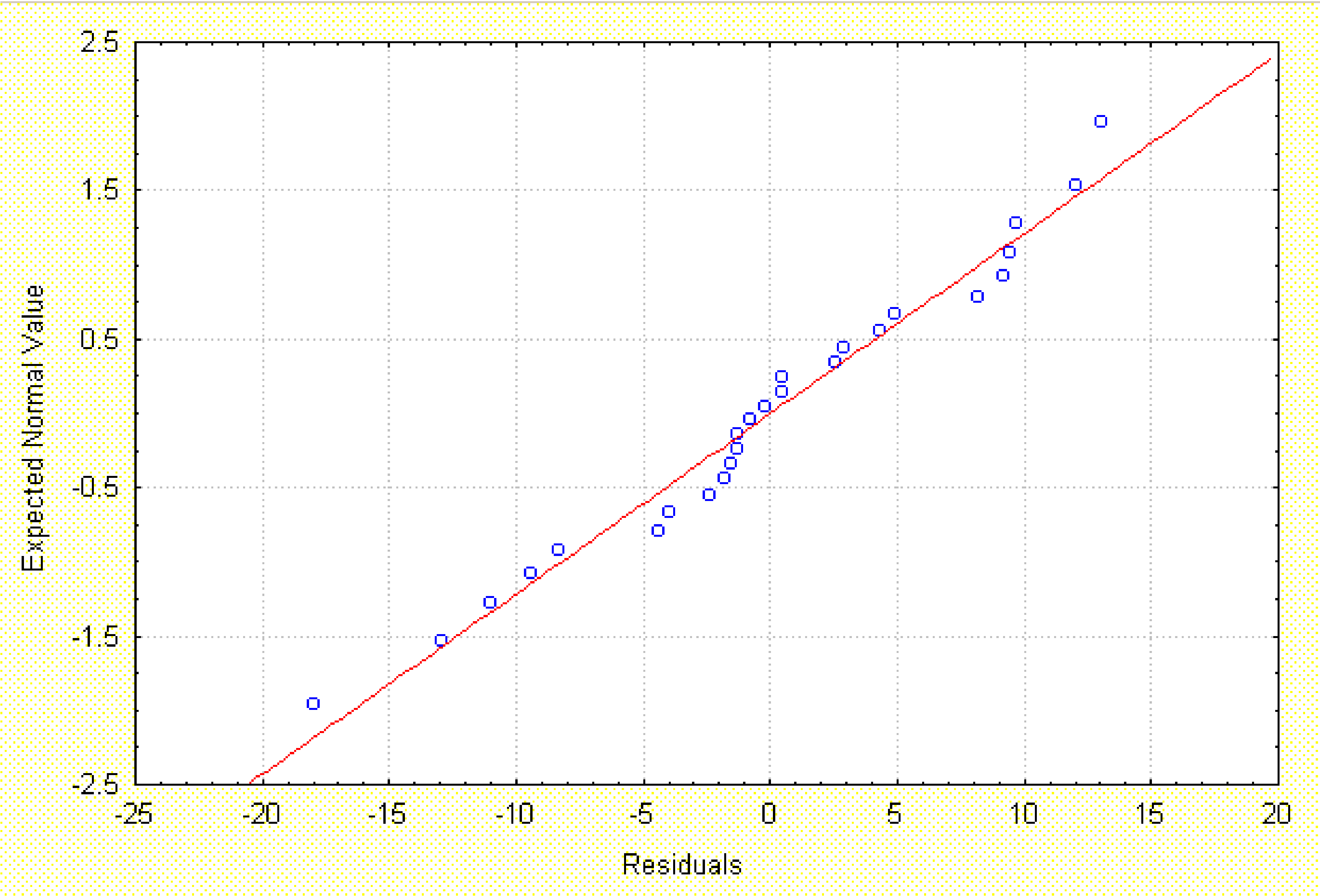

Figure 2.5. Normal probability plot of the residuals

Outlying Observations

Outlying observations can be identified using either the residual plot or the standardized residual plot. The interpretation of these plots is the same as in the simple linear regression case. Below is the standardized residual plot for the selling price data.

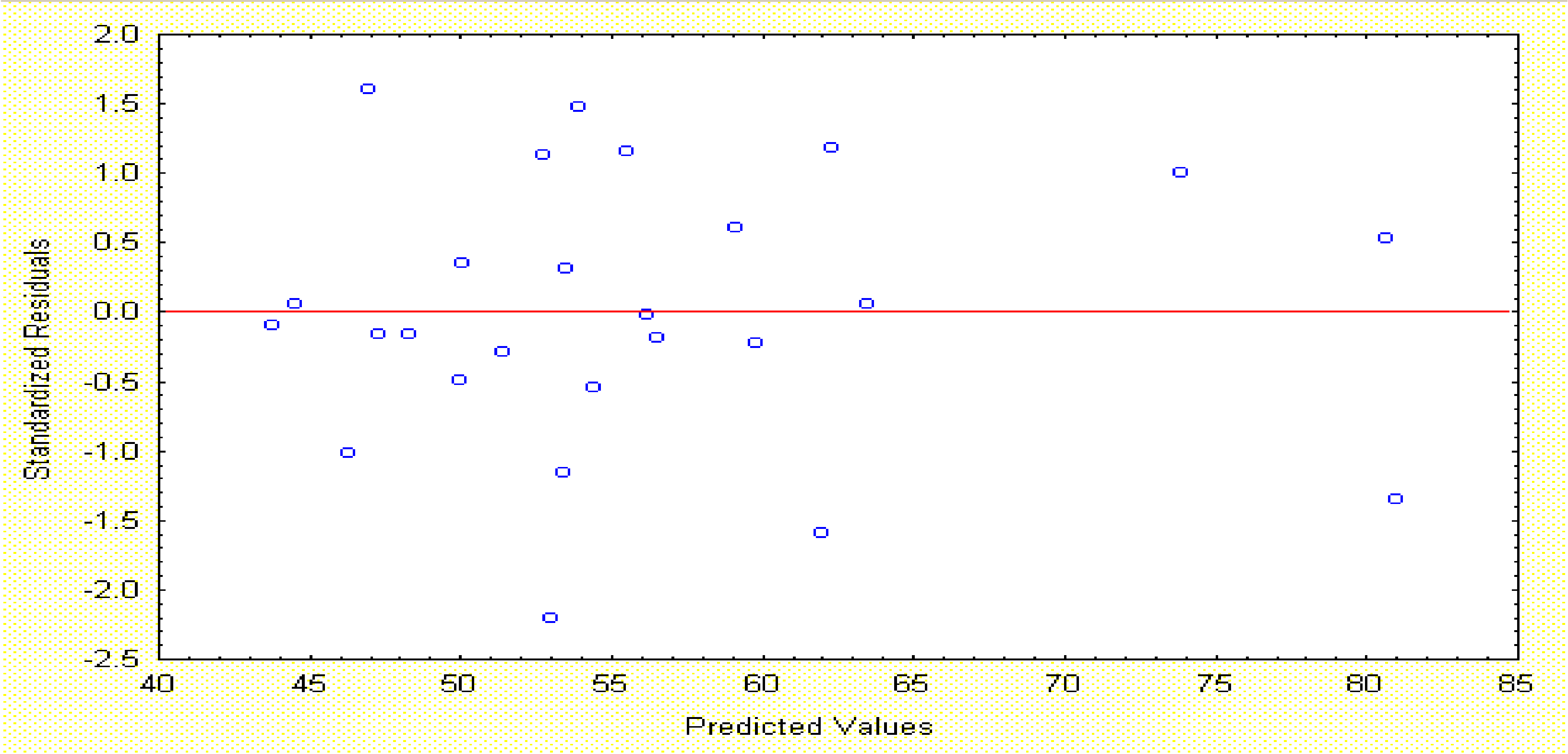

Figure 2.6. Standardized residual plot for the *selling price* data

The distribution of the points in Figures 2.3 and 2.6 are the same. For both plots, the values are clustered together suggesting that there is no outlying observation in the data set.

Prototype Residual Plots:

The figures below are some examples of residual plots that might be encountered in practice.

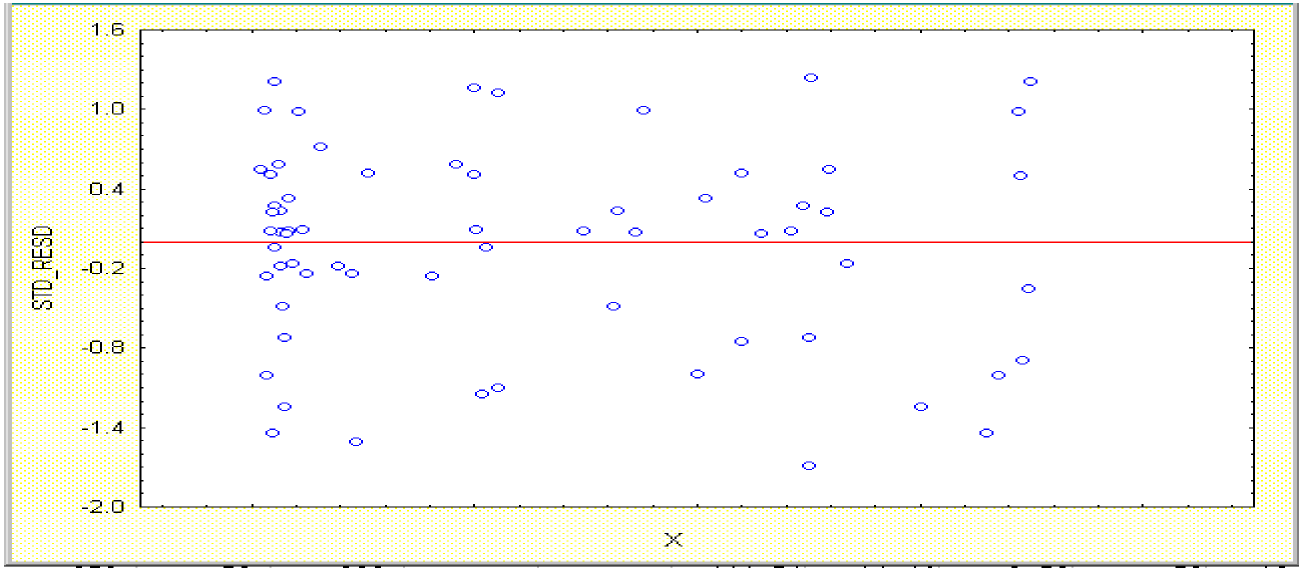

Figure 2.7

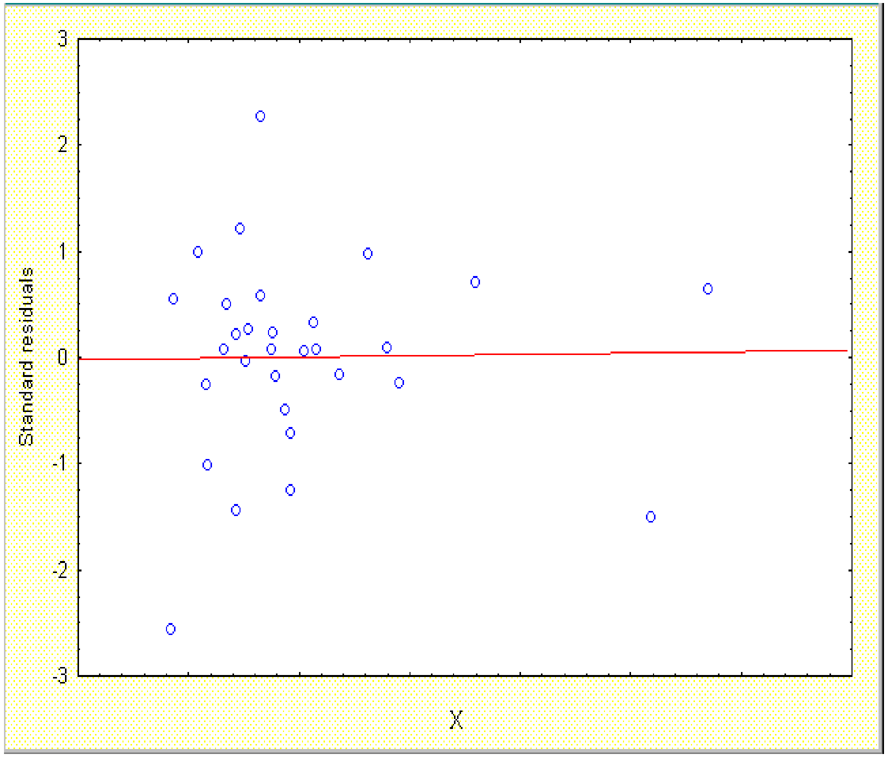

Figure 2.8a

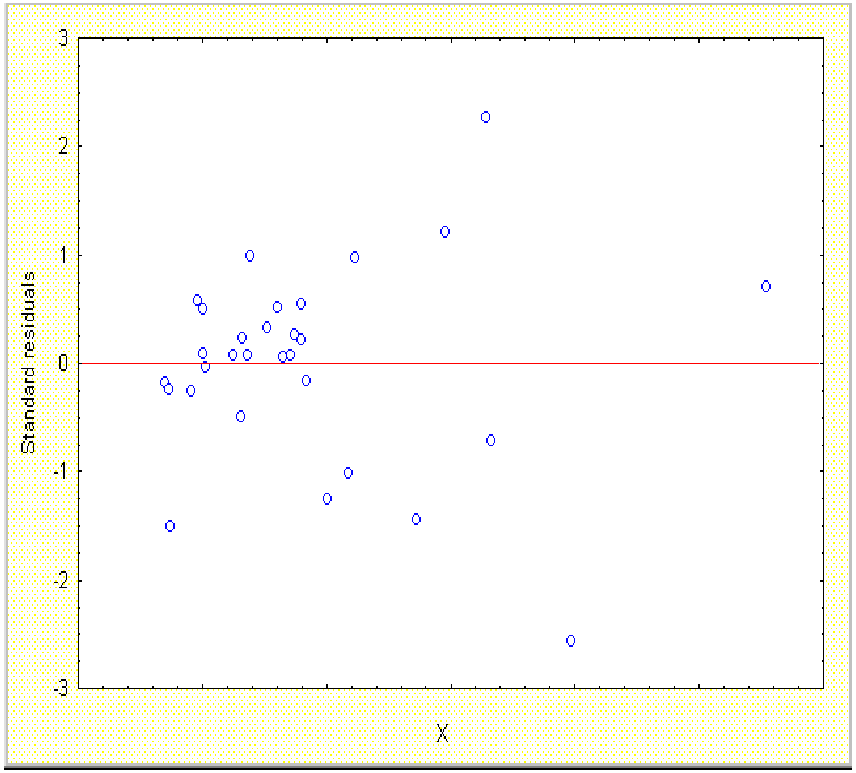

Figure 2.8b

Figure 2.7 shows a residual plot that follows no systematic pattern, suggesting no depatures from the assumptions of the model.

Figures 2.8a and 2.8b show residual plots that exhibit nonconstancy of variance. Figure 2.8a exhibits a funnel-shaped graph opening to the right. This suggests that the residual values increase as the predicted values increase. Likewise, Figure 2.8b exhibits a funnel-shaped graph opening to the left. This suggests that the residual values decrease as the predicted values increase.

Multicollinearity

Multicollinearity refers to the presence of linear dependencies among the regressor variables. This occurs when one or more of the regressor variables can be expressed as a linear combination of the other regressor variables.

One measure that is used to detect the presence of multicollinearity is the tolerance value. One regressor variable is regressed on the remaining regressor variables and the coefficient of multiple determination (R^2) is computed. The tolerance value is obtained by subtracting the R^2 from 1. Below are the tolerance values computed for the *selling price* data.

Table 2.3. Tolerance values computed for the *selling price* data

	Tolerance	**R–square**	**Partial Correlation**	**Semipartial Correlation**
FLR	.836065	.163935	.672111	.563967
GAR	.836065	.163935	.397382	.269053

A tolerance value smaller than 0.1 indicates the presence of multicollinearity. The tolerance values given in the table above are all greater than 0.1, indicating that serious multicollinearity is not present.

Some remedial measures that can be used to treat the problem of multicollinearity include:

1. deletion of one or more regressor variables in the model
2. the use of ridge regression
3. the use of principal components regression

Variable Selection

- Not all variables hypothesized to affect the response variable may have a significant contribution in predicting the mean value of the response variable. Thus model selection procedures are needed to choose the "best" set of regressor variables.

Some Common Variable Selection Procedures

- Backward Selection Procedure

- Forward Selection Procedure

Variable Selection

Backward Selection

- starts with all predictor variables in the model
- removes the regresssors sequentially, depending on whether or not the regression coefficient associated with the regressor variable is significantly different from zero

Table 2.4 gives the ANOVA for fitting a regression model containing the regressor variables FLR, GAR and BDR, number of bedrooms.

Table 2.4. Analysis of variance table for the regression model containing the regressor variables FLR, GAR and BDR

Effect	Sum of Squares	df	Mean Squares	F-value	p-level
Regression	2476.559	3	825.5197	12.26440	.000063
Residual	1480.826	22	67.3103		
Total	3957.385				

The table below gives the summary of the backward selection procedure for fitting a regression model with FLR, GAR and BDR as regressors.

Table 2.5. Summary output of the backward selection procedure for the regression model containing the regressor variables FLR, GAR and BDR

	BETA	St. Err. of BETA	B	St. Err. of B	t	p-level
Intercept			35.29369	4.66440	7.56661	.00000
BDR	-.14768	.177053	-1.42314	1.70617	-.83411	.00023
FLR	.71894	.188001	.01975	.005164	3.82415	.00092
GAR	.28825	.142813	4.41294	2.18639	2.01837	.05591

	BETA	St. Err. of BETA	B	St. Err. of B	t	p-level
Intercept			35.69431	4.96181	7.19380	.00000
BDR	-.16568	.188273	-1.59662	1.81430	-.88002	.38795
FLR	.84781	.188273	.02329	.005171	4.5031	.00016

	BETA	St. Err. of BETA	B	St. Err. of B	t	p-level
Intercept			33.91683	4.51075	7.51912	.00000
FLR	.73592	.138205	.02021	.003796	5.32486	.00002

Forward Selection

- starts with one regressor variable entered into the equation

- adds variables sequentially depending on an entry criterion (F-to-enter).

The table below gives the summary of the forward selection procedure for fitting a regression model with FLR, GAR and BDR as regressors.

Table 2.6. Summary output of the forward selection procedure for the regression model with FLR, GAR and BDR

	Step +in / - out	**Multiple R-square**	**R-square change**	**F – to enter/rem**	**p-level**	**Variables included**
FLR	1	.541584	.541584	28.35417	.000021	1
GAR	2	.613973	.072390	4.31307	.049177	2

- Thank you.