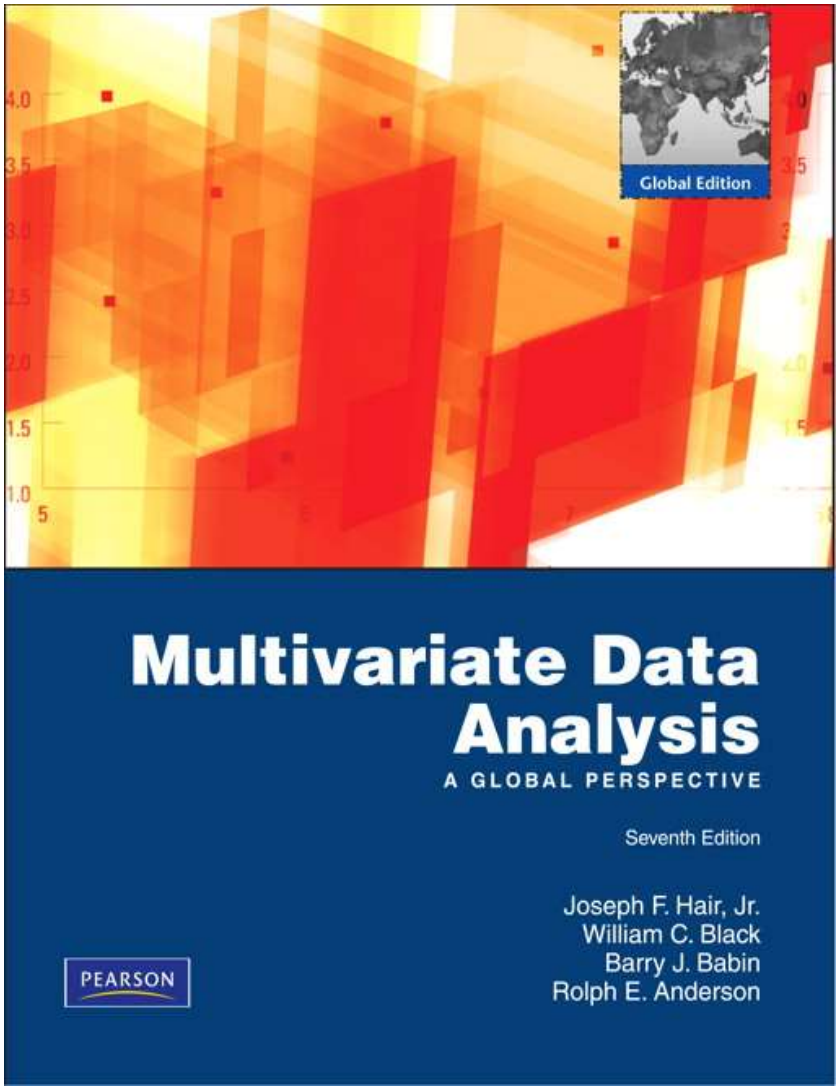
Chapter 4
Simple and Multiple Regression



Use different regression model

When a simple linear regression model is not appropriate and transformations are not helpful, then we will need to look for alternative regression model, such as the multiple regression model that we will discuss in the next section.

Chapter 4
Simple and Multiple Regression

LEARNING OBJECTIVES

Upon completing this chapter, you should be able to do the following:

- Determine when regression analysis is the appropriate statistical tool in analyzing a problem.
- Understand how regression helps us make predictions using the least squares concept.
- Use dummy variables with an understanding of their interpretation.
- Be aware of the assumptions underlying regression analysis and how to assess them.

Chapter 4
Simple and Multiple Regression

LEARNING OBJECTIVES continued . . .

Upon completing this chapter, you should be able to do the following:

- Select an estimation technique and explain the difference between stepwise and simultaneous regression.
- Interpret the results of regression.
- Apply the diagnostic procedures necessary to assess "influential" observations.

Multiple Regression and Correlation Analysis

When a simple linear regression model is not appropriate and transformations are not helpful, then we might need to describe the relationship using the multiple regression. In this section we will learn about the multiple regression and correlation analysis.

Multiple linear regression

- So far we have been dealing with a single predictor and a single response – simple linear regression
- When we have multiple predictors we talk about multiple linear regression
- Multiple linear regression is just as easy to perform as simple linear regression
- But there are a few more problems that can occur
- And we have to think much more about the process of model selection
- Model selection is possibly one of the hardest problems in statistics

Multiple Regression Defined

Multiple regression analysis . . . is a statistical technique that can be used to analyze the relationship between a single dependent (criterion) variable and several independent (predictor) variables.

Multiple Regression

$$Y' = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + e$$

Y' = Dependent Variable = # of credit cards

b_0 = intercept (constant) = constant number of credit cards independent of family size and income.

b_1 = change in # of credit cards associated with a unit change in family size (regression coefficient).

b_2 = change in # of credit cards associated with a unit change in income (regression coefficient).

X_1 = family size

X_2 = income

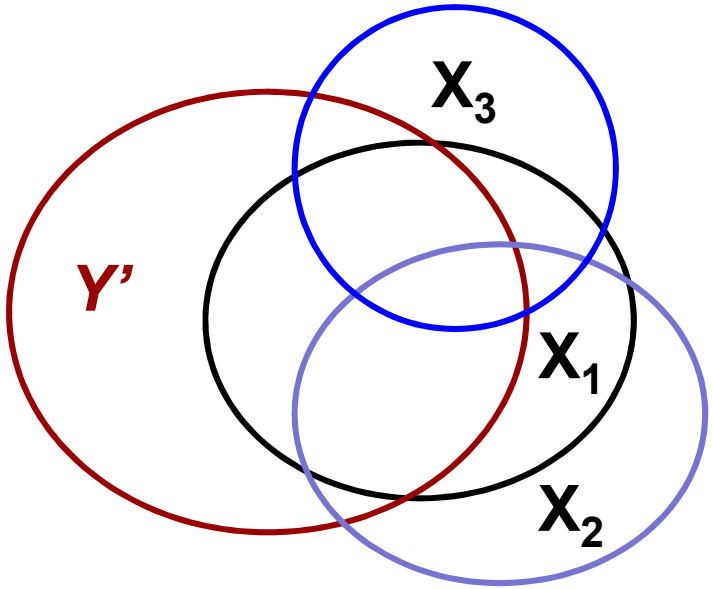
e = prediction error (residual)

Variate

Variate (Y') = $X_1b_1 + X_2b_2 + \ldots + X_nb_n$

A variate value (Y') is calculated for each respondent.

The Y' value is a linear combination of the entire set of variables that best achieves the statistical objective.



A probability model for multiple linear regression

- Now instead of one predictor we have p. Our model becomes

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}, \varepsilon_i \; iid \; N\left(0, \sigma^2\right)$$

- Note our model still has an intercept, and, it makes the same assumptions about the distribution of the residuals
- If we rewrite this in matrix form we write

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim MVN\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right)$$

- This looks just the same as before, but in fact some of the definitions have changed slightly

Matrix form of regression model for multiple linear regression

- Let $\boldsymbol{y}$ be a (n by 1) vector of responses, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$
- Let $\boldsymbol{\varepsilon}$ be a (n by 1) vector of errors, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$
- These are the same but now we

let β be a (p+1 by 1) vector of coefficients $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)^T$

and $\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$ $\boldsymbol{X}$ is called the design matrix

This form is convenient because the least squares solutions are found in exactly the same way as before $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$

Changes to the Regression ANOVA and Regression Table

- Our hypothesis for the ANOVA changes to

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

- Of course the alternative is now

$$H_1: \beta_i \neq 0 \text{ for some } i = 1, \ldots, p$$

- And a small p-value is evidence against the null hypothesis, i.e. evidence that the regression is significant
- There is now a t-test associated with each regression coefficient, with small P-values implying that the specific predictors are important in predicting the response

MULTIPLE LINEAR REGRESSION MODEL

The **statistical model for multiple linear regression** is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

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

The **mean response** μ_y is a linear function of the explanatory variables:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

The **deviations** ϵ_i are independent and Normally distributed with mean 0 and standard deviation σ. That is, they are an SRS from the $N(0, \sigma)$ distribution.

The parameters of the model are β_0, β_1, β_2, . . . , β_p, and σ.

Multiple Regression Analysis

The general multiple regression with k independent variables is estimated by:

$$Y' = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

- The least squares criterion is used to develop this equation.
- Because determining b_1, b_2, *etc*. is very tedious, a software package such as Excel or MINITAB is recommended.

Regression Models

Probabilistic Multiple Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon$$

Y = the value of the dependent (response) variable

β_0 = the regression constant

β_1 = the partial regression coefficient of independent variable 1

β_2 = the partial regression coefficient of independent variable 2

β_k = the partial regression coefficient of independent variable k

k = the number of independent variables

ε = the error of prediction

Estimated Regression Model

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \cdots + b_k X_k$$

where: $\hat{Y}$ = predicted value of Y

b_0 = estimate of regression constant

b_1 = estimate of regression coefficient 1

b_2 = estimate of regression coefficient 2

b_3 = estimate of regression coefficient 3

b_k = estimate of regression coefficient k

k = number of independent variables

Multiple Regression Analysis

For two independent variables, the general form of the multiple regression equation is:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

- X_1 and X_2 are the independent variables.
- a is the Y-intercept.
- b_1 is the net change in Y for each unit change in X_1 holding X_2 constant. It is called a partial regression coefficient, a net regression coefficient, or just a regression coefficient.

Multiple Regression Model with Two Independent Variables (First-Order)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Population Model

where: β_0 = the regression constant

β_1 = the partial regression coefficient for independent variable 1

β_2 = the partial regression coefficient for independent variable 2

ε = the error of prediction

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

Estimated Model

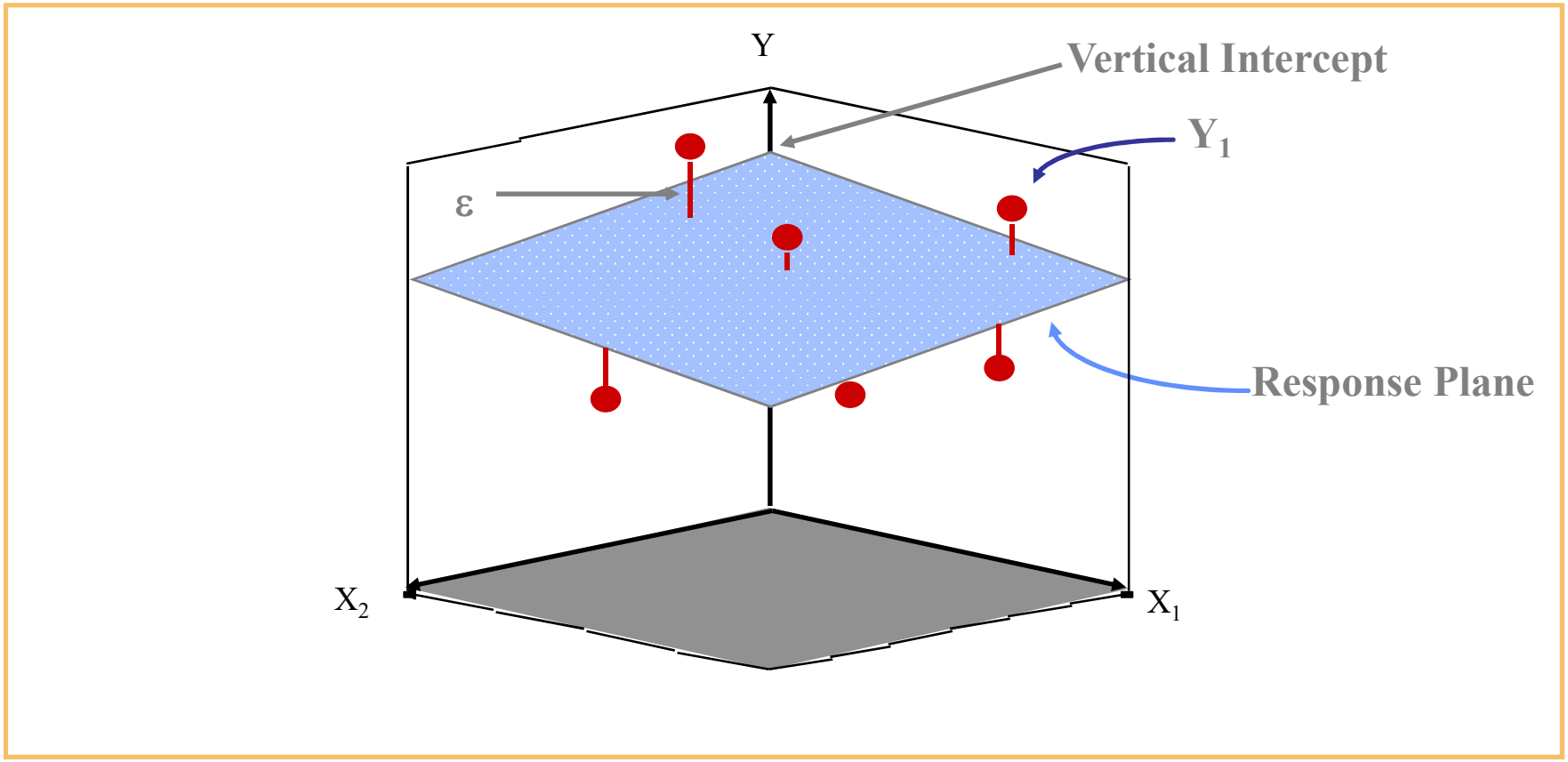
where: $\hat{Y}$ = predicted value of Y

b_0 = estimate of regression constant

b_1 = estimate of regression coefficient 1

b_2 = estimate of regression coefficient 2

Response Plane for First-Order Two-Predictor Multiple Regression Model



Least Squares Equations for k = 2

$$b_0 n + b_1 \sum X_1 + b_2 \sum X_2 = \sum Y$$

$$b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 = \sum X_1 Y$$

$$b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 = \sum X_2 Y$$

Multiple Regression Decision Process

Stage 1: Objectives of Multiple Regression

Stage 2: Research Design of Multiple Regression

Stage 3: Assumptions in Multiple Regression Analysis

Stage 4: Estimating the Regression Model and Assessing Overall Fit

Stage 5: Interpreting the Regression Variate

Stage 6: Validation of the Results

Stage 1: Objectives of Multiple Regression

In selecting suitable applications of multiple regression, the researcher must consider three primary issues:

1. the appropriateness of the research problem,
2. specification of a statistical relationship, and
3. selection of the dependent and independent variables.

Selection of Dependent and Independent Variables

The researcher should always consider three issues that can affect any decision about variables:

- The theory that supports using the variables,
- Measurement error, especially in the dependent variable, and
- Specification error.

Measurement Error in Regression

Measurement error that is problematic can be addressed through either of two approaches:

- Summated scales, or
- Structural equation modeling procedures.

Rules of Thumb 4–1

Meeting Multiple Regression Objectives

- Only structural equation modeling (SEM) can directly accommodate measurement error, but using summated scales can mitigate it when using multiple regression.
- When in doubt, include potentially irrelevant variables (as they can only confuse interpretation) rather than possibly omitting a relevant variable (which can bias all regression estimates).

Stage 2: Research Design of a Multiple Regression Analysis

Issues to consider . . .

- Sample size,
- Unique elements of the dependence relationship – can use dummy variables as independents, and
- Nature of independent variables – can be both fixed and random.

Rules of Thumb 4–2

Sample Size Considerations

- Simple regression can be effective with a sample size of 20, but maintaining power at .80 in multiple regression requires a minimum sample of 50 and preferably 100 observations for most research situations.
- The minimum ratio of observations to variables is 5 to 1, but the preferred ratio is 15 or 20 to 1, and this should increase when stepwise estimation is used.
- Maximizing the degrees of freedom improves generalizability and addresses both model parsimony and sample size concerns.

Rules of Thumb 4–3

Variable Transformations

- Nonmetric variables can only be included in a regression analysis by creating dummy variables.
- Dummy variables can only be interpreted in relation to their reference category.
- Adding an additional polynomial term represents another inflection point in the curvilinear relationship.
- Quadratic and cubic polynomials are generally sufficient to represent most curvilinear relationships.
- Assessing the significance of a polynomial or interaction term is accomplished by evaluating incremental R^2, not the significance of individual coefficients, due to high multicollinearity.

What happens, when the regression model is not appropriate?

There are many remedial action that one can take to correct the fact that a regression model is not appropriate. We will describe two types of remedial actions:

1. Transformations of variables.

2. Use different regression model.

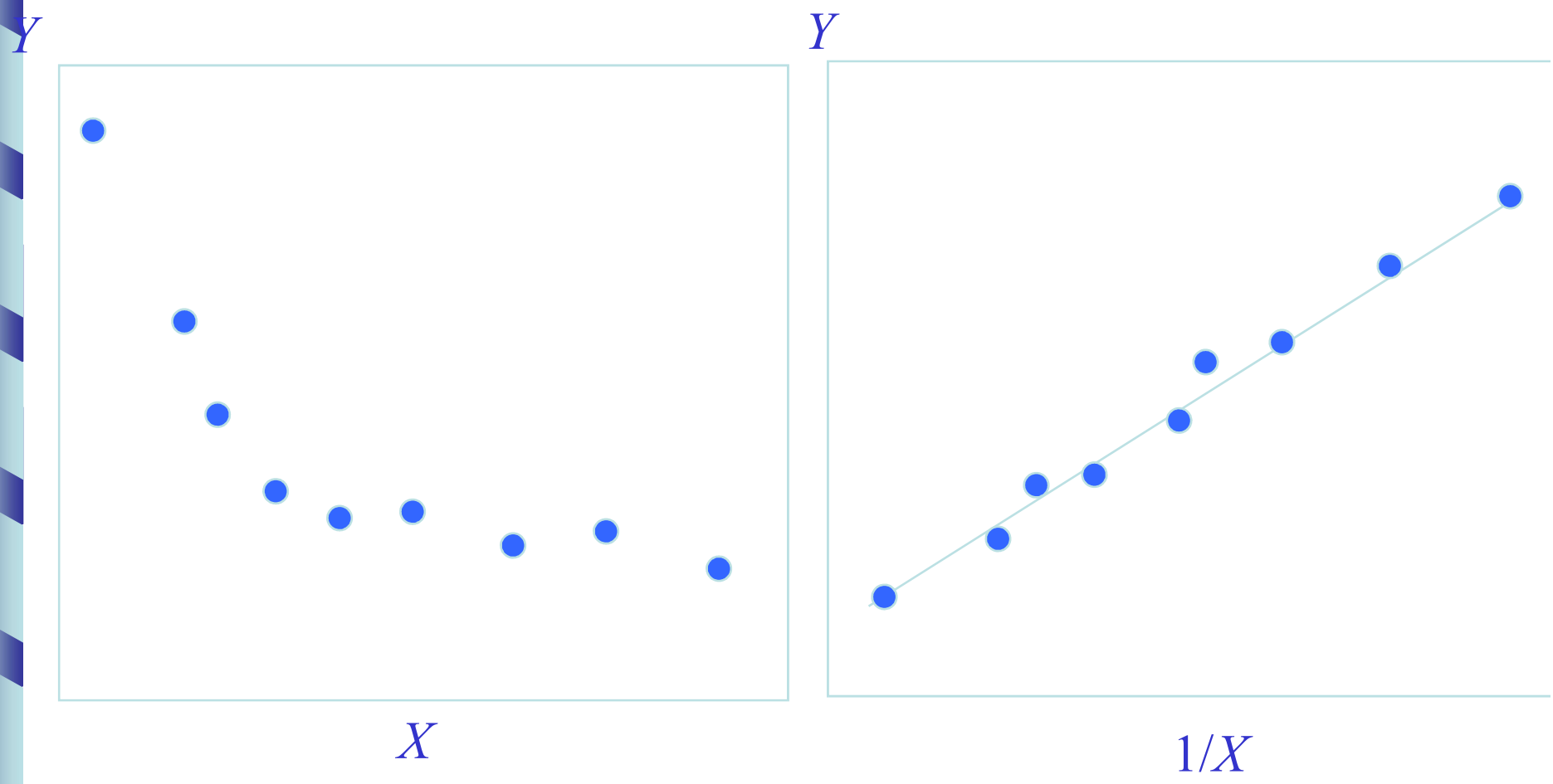
Transformations of variables

When a regression model is not linear, it might be possible to linearize it by transforming the independent variable.

However, if lack of linearity and non-constant error variance are both found to be present, then a transformation of dependent variable may be helpful. Otherwise, both the independent variable and the dependent variable may be transformed.

Transformation: logarithmic, square root or reciprocal.

Reciprocal transformation of the independent variable



Various forms of transformations of variables

$$Y = a + b\,\frac{1}{X}$$

$$Y = a + b\,\sqrt{X}$$

$$Y = a + b\,(\ln(X))$$

$$\ln(Y) = a + b\,(\ln(X)) \qquad etc\ .$$

Note: the last transformation is useful when $Y = \beta\alpha^{X}$.

Models or Prediction Equations

- Some examples of various possible relationships:

Linear: $\hat{y} = b_0 + b_1 x$

Quadratic: $\hat{y} = (a + bx)^2$

Exponential: $\hat{y} = a(b^x)$

Logarithmic: $\hat{y} = a \log_b x$

Reciprocal: $\hat{y} = \dfrac{1}{a + bx}$

Note: What would a scatter diagram look like to suggest each relationship?

Nonlinear Regression Models: Model Transformation

$$\hat{y} = ab^x$$

$$\Rightarrow \log(\hat{y}) = \log(a) + x\log(b)$$

$$\Rightarrow \hat{y}' = a' + b'x$$

where:

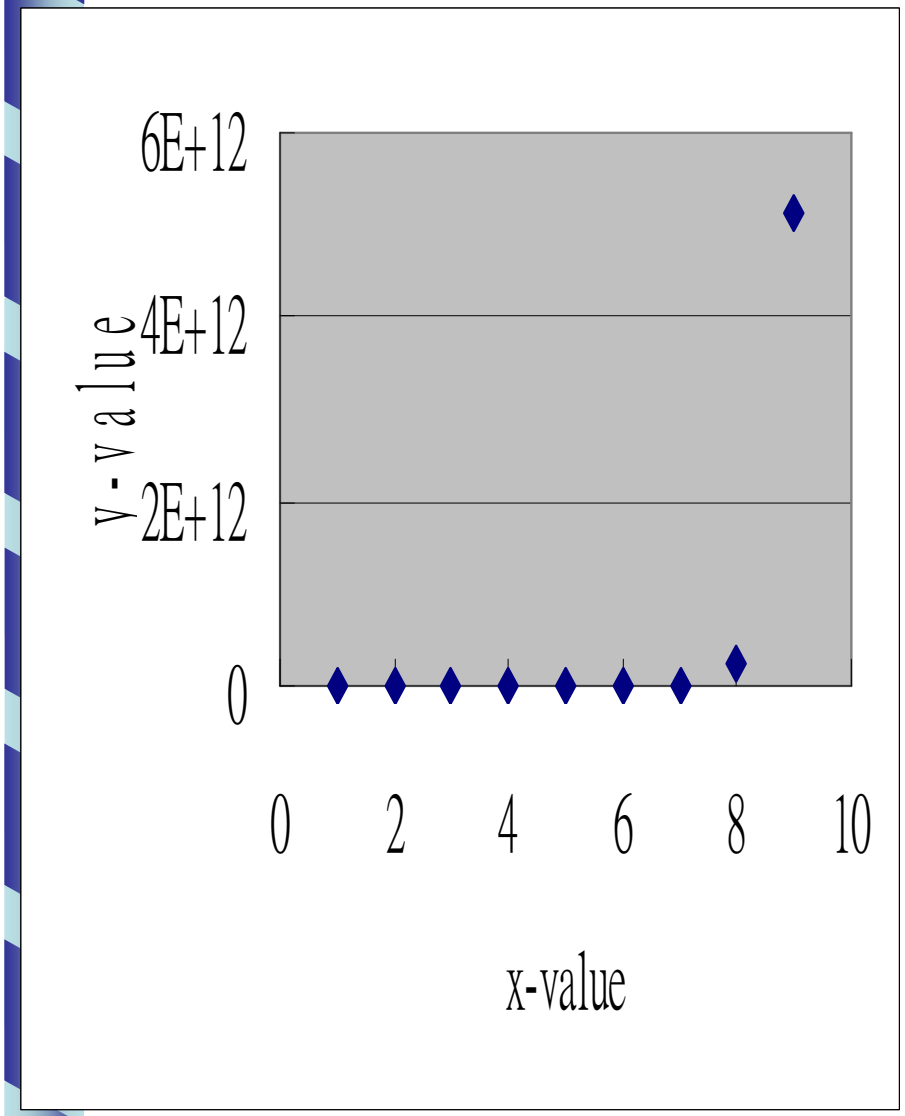
$$\hat{y}' = \log(\hat{y})$$

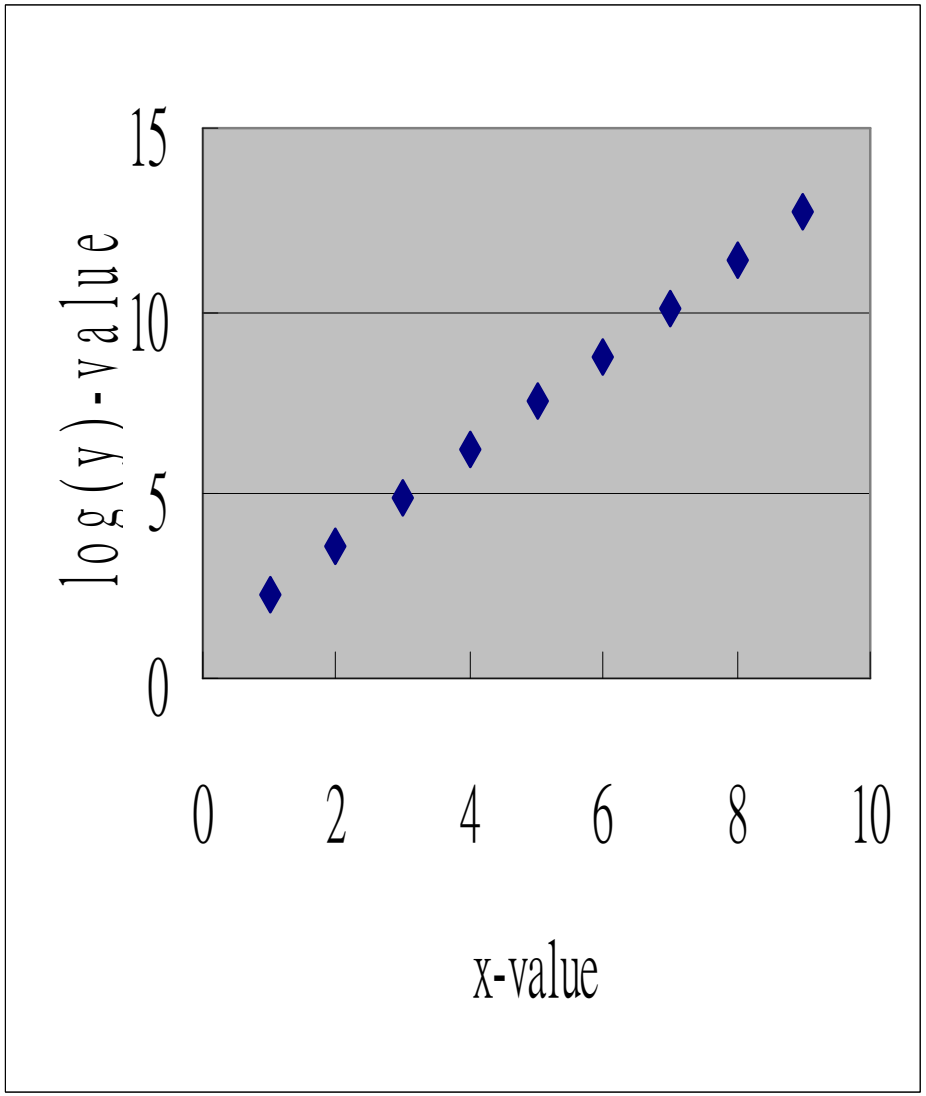
$$a' = \log(a)$$

$$b' = \log(b)$$

Hence we map $\hat{y}'$ vs. x

Corresponding Scatter Plot





Nonlinear Regression Models: Model Transformation

$$\hat{y} = (a + bx)^2$$

$$\Rightarrow \sqrt{\hat{y}} = a + bx$$

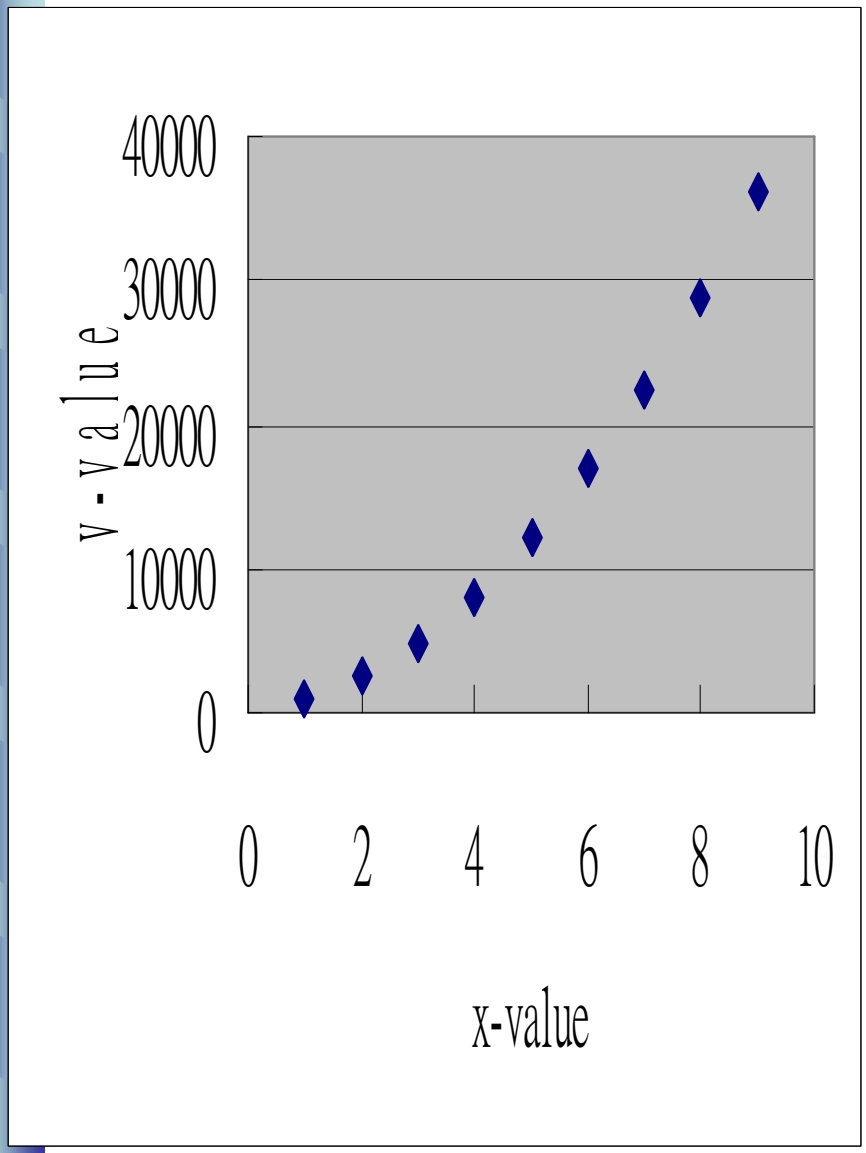
$$\Rightarrow \hat{y}' = a + bx$$

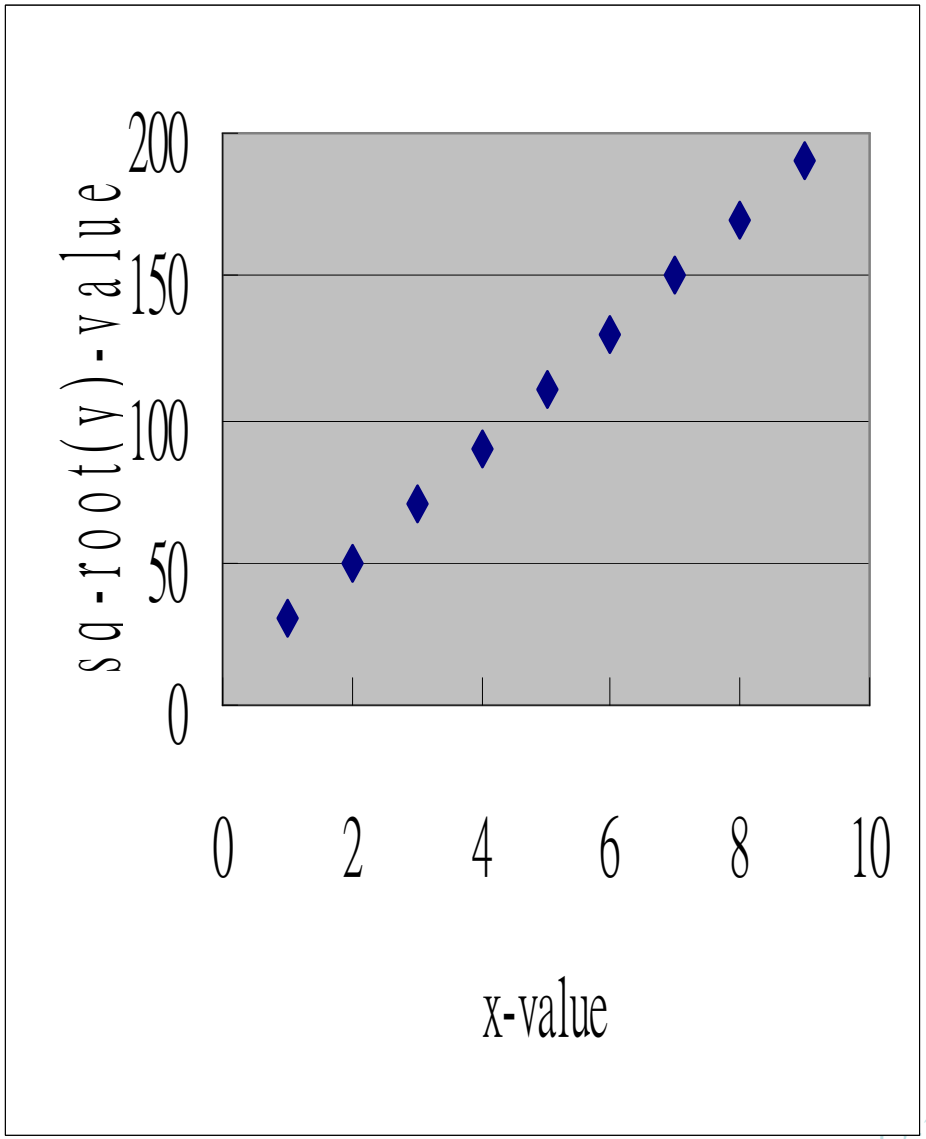
where:

$$\hat{y}' = \sqrt{\hat{y}}$$

Hence we map $\hat{y}'$ vs. x

Corresponding Scatter Plot





Nonlinear Regression Models: Model Transformation

$$\hat{y} = \frac{1}{a + bx}$$

$$\Rightarrow \frac{1}{\hat{y}} = a + bx$$

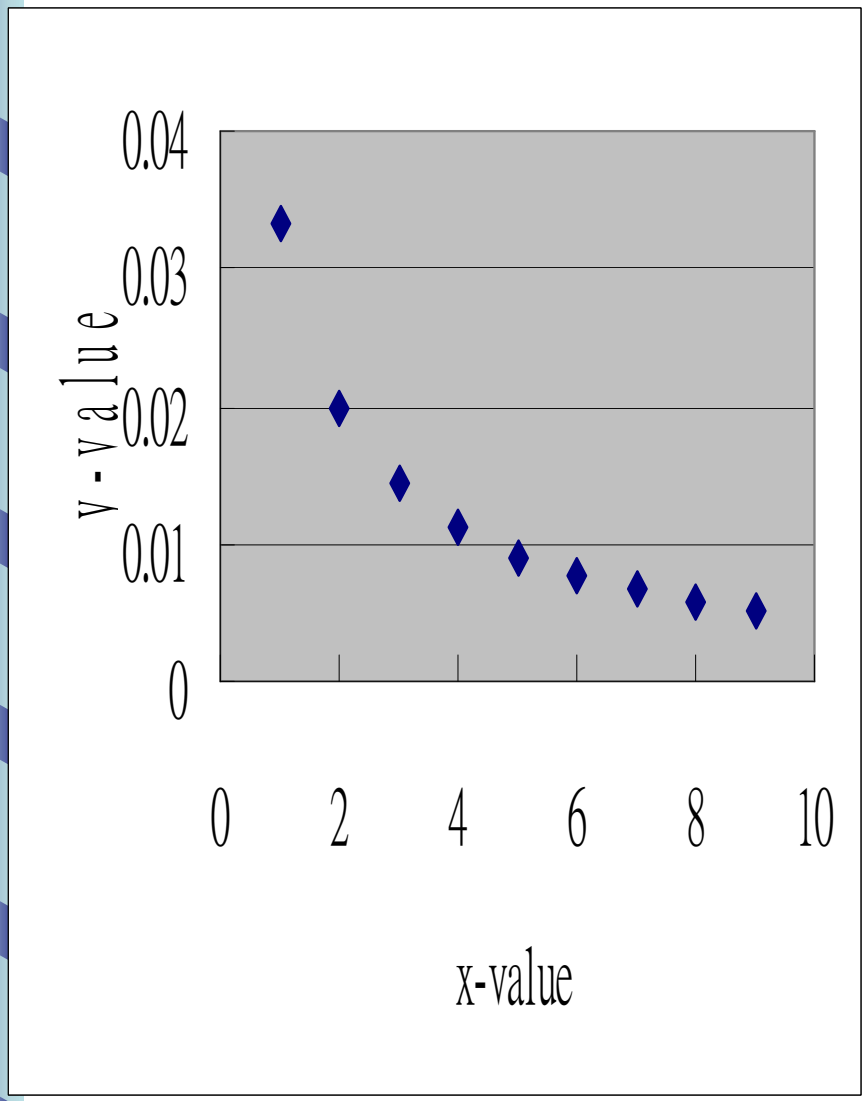
$$\Rightarrow \hat{y}' = a + bx$$

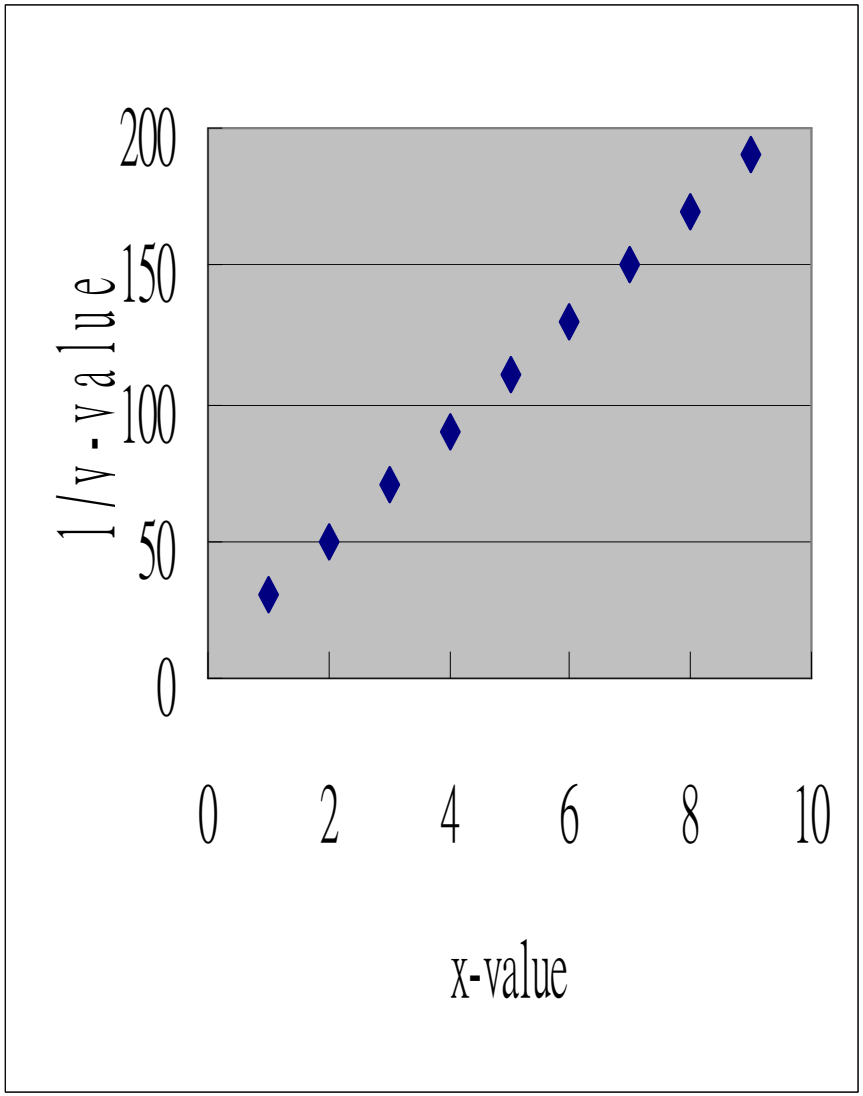
where :

$$\hat{y}' = \frac{1}{\hat{y}}$$

Hence we map $\hat{y}'$ vs. x

Corresponding Scatter Plot





Stage 3: Assumptions in Multiple Regression Analysis

- Linearity of the phenomenon measured.
- Constant variance of the error terms.
- Independence of the error terms.
- Normality of the error term distribution.

Rules of Thumb 4–4

Assessing Statistical Assumptions

- Testing assumptions must be done not only for each dependent and independent variable, but for the variate as well.
- Graphical analyses (i.e., partial regression plots, residual plots and normal probability plots) are the most widely used methods of assessing assumptions for the variate.
- Remedies for problems found in the variate must be accomplished by modifying one or more independent variables as described in Chapter 2.

Checking model assumptions

Assumptions Underlying Linear Regression

For each value of *X*, there is a group of *Y* values, and these *Y* values are *normally distributed*.

1. The *means* of these normal distributions of *Y* values all lie on the straight line of regression.
2. The *standard deviations* of these normal distributions are equal.
3. The *Y* values are statistically independent. This means that in the selection of a sample, the *Y* values chosen for a particular *X* value do not depend on the *Y* values for any other *X* values.

Evaluation of a Regression Model

GOALS

Next we will learn about how to find a regression model that is suitable for a given data set. Next we will investigate what one needs to after finding a model. When a regression model is fitted in practice, we will need to check if the model is appropriate or suitable, using the data themselves for guidance. A primary tool for studying the appropriateness of a regression model is residual analysis.

Residual Analysis

In the regression model, we made some assumptions. Now, we will investigate the cases when the assumptions do not hold. Here are some cases that we will investigate:

1. The regression function is not linear.
2. The distribution of Y do not have constant variances at all level of X (i.e. the error terms do not have constant variances).
3. The distributions of Y are not normal (i.e. the error terms are not normally distributed).
4. The error terms are not independent.

The linear model:

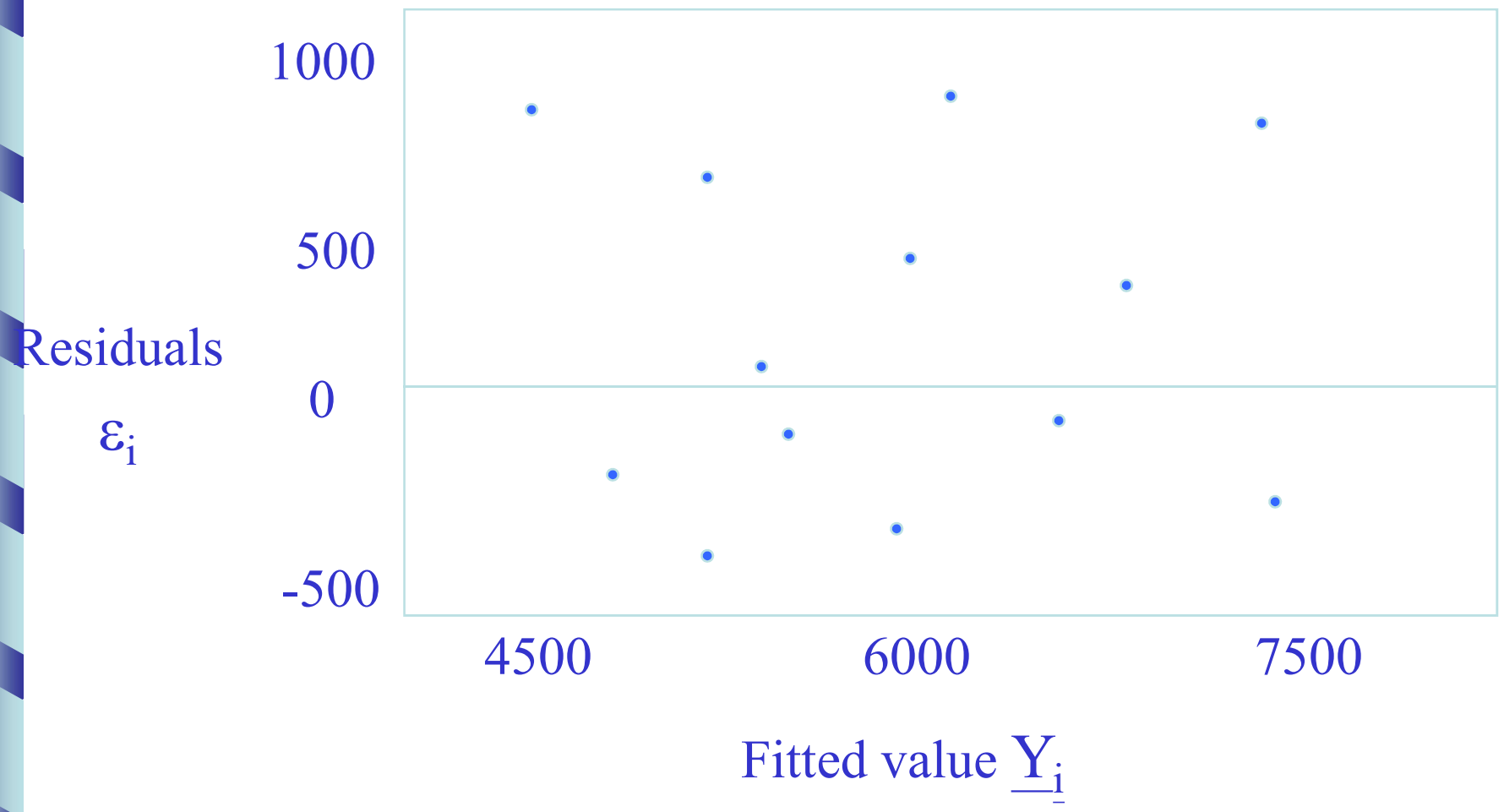
Let's look at the linear regression model used:

$Y_i = a + b\,X_i + \varepsilon_i$ where ε_i is the error term.

A residual plot is a very useful method to indicate a solution - plot the residual against the fitted value as scatter plot :

$\varepsilon_i\,(= Y_i - E[Y_i])$ v.s. $\underline{Y_i}$.

Residual Plot



Analysis of Residuals

A residual is the difference between the actual value of Y and the predicted value Y'.

- Residuals should be approximately normally distributed. Histograms and stem-and-leaf charts are useful in checking this requirement.
- A plot of the residuals and their corresponding Y' values is used for showing that there are no trends or patterns in the residuals.

Check for Normality

- Recall that we when proposed the regression model we made an assumption.
- Namely, the errors are normally distributed with mean zero and variance sigma squares
- This means, like in ANOVA, we need to check whether the residuals are normally distributed and whether the variances are equal amongst residuals
- We've seen how to check for normality – a norplot.

Diagnostics for lack of Normality

In this case, many method can be used: the normal probability plot of the residuals, a stem and leaf plot and the chi-square test all can be used to get this information.

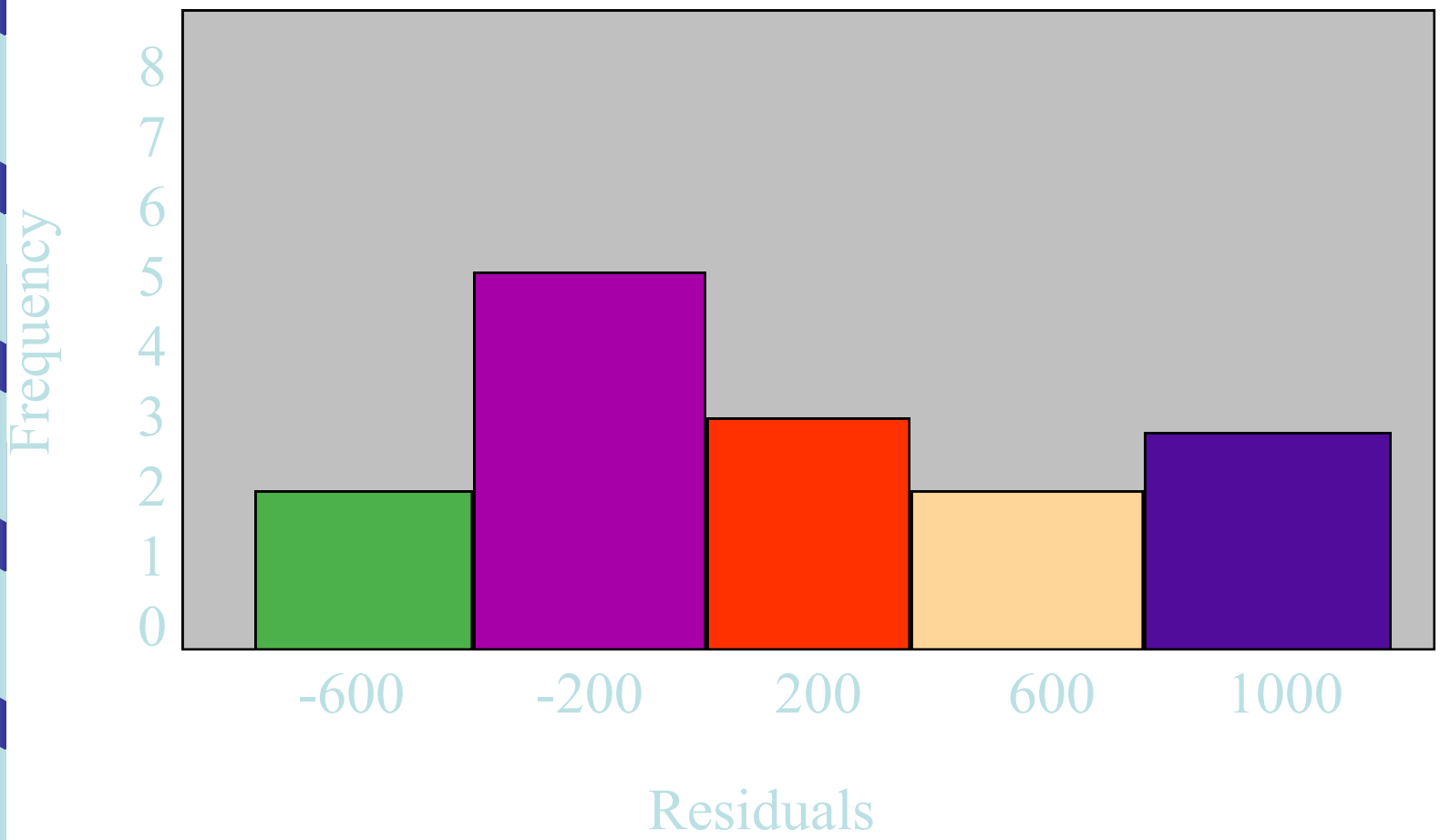
We will leave this till after we did the chi-square test.

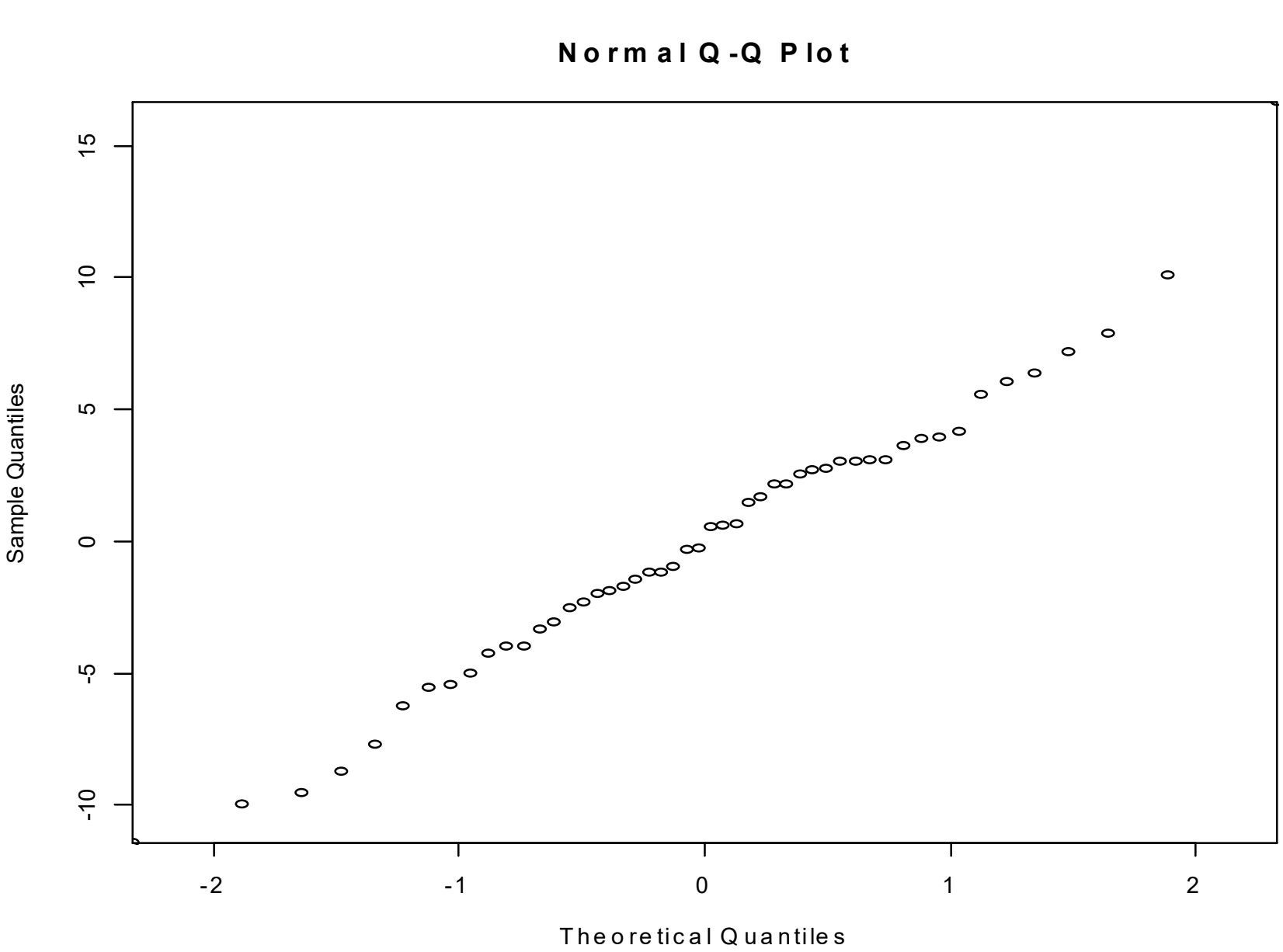
The normal probability plot map the error terms v.s. the expected value under normality -

$$Expected \quad value = z(\frac{i-0.5}{n})\sqrt{MSE}$$

Or the norplot: x v.s. z_x (where $\boldsymbol{x} = \mu + z_x\sigma$)

Histograms of Residuals





Our norplot is follows a straight line reasonably well, therefore we might think our assumption of normality is satisfied.

Equality of variance

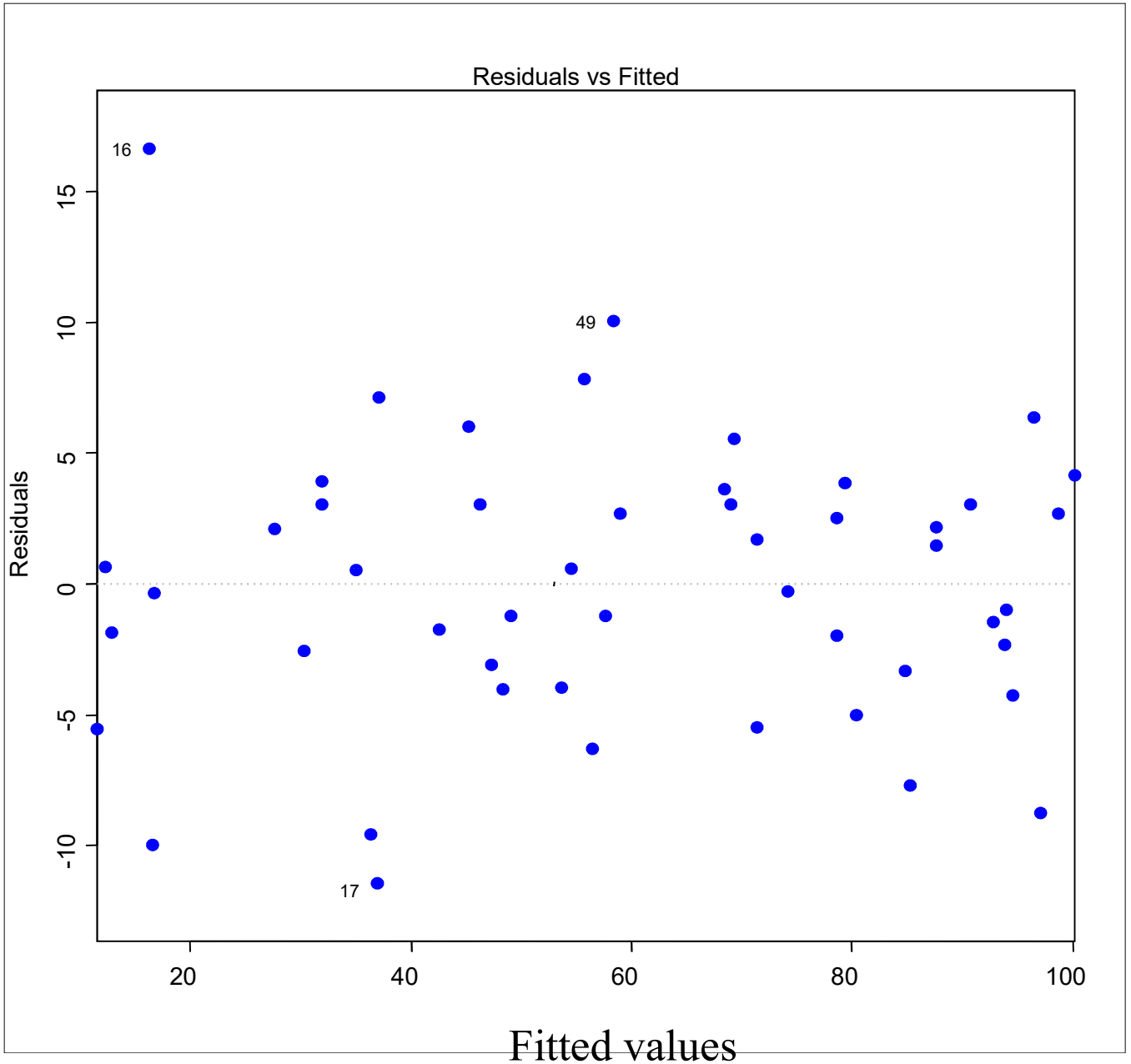
- It is possible to have normality without having equality of variance, i.e. in some situations we fit the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \text{ iid. } N(0, \sigma_i^2)$$

- However, we did not fit this model to this set of data, we assumed that we had equal variances for every error, i.e.

$$\sigma_i = \sigma, \forall i$$

  - We check this assumption, as before, with a residuals plot
  - This time however, we will generally see less patterning, because the data are not grouped


Residuals vs Fitted
Residuals
Fitted values
16
49
17
15
10
5
0
-5
-10
20
40
60
80
100

Interpreting residuals plots

- What do we look for in a residuals plot?
  1. Extreme residuals – our estimated standard deviation of the residuals is σ (=5.6 in this case)
  2. More negative residuals than positive or vice versa
  3. Strong patterns or trends in the residuals
- What do these features mean?
- If we have extreme residuals then there are a number of reasons why
  1. An outlier or a data entry error
  2. Poor model fit
  3. Possible high leverage points elsewhere
- If there is a disproportionate ratio of positive to negative residuals then we may have
  1. A skewed response variable
  2. Poor model fit

Interpreting residuals plots

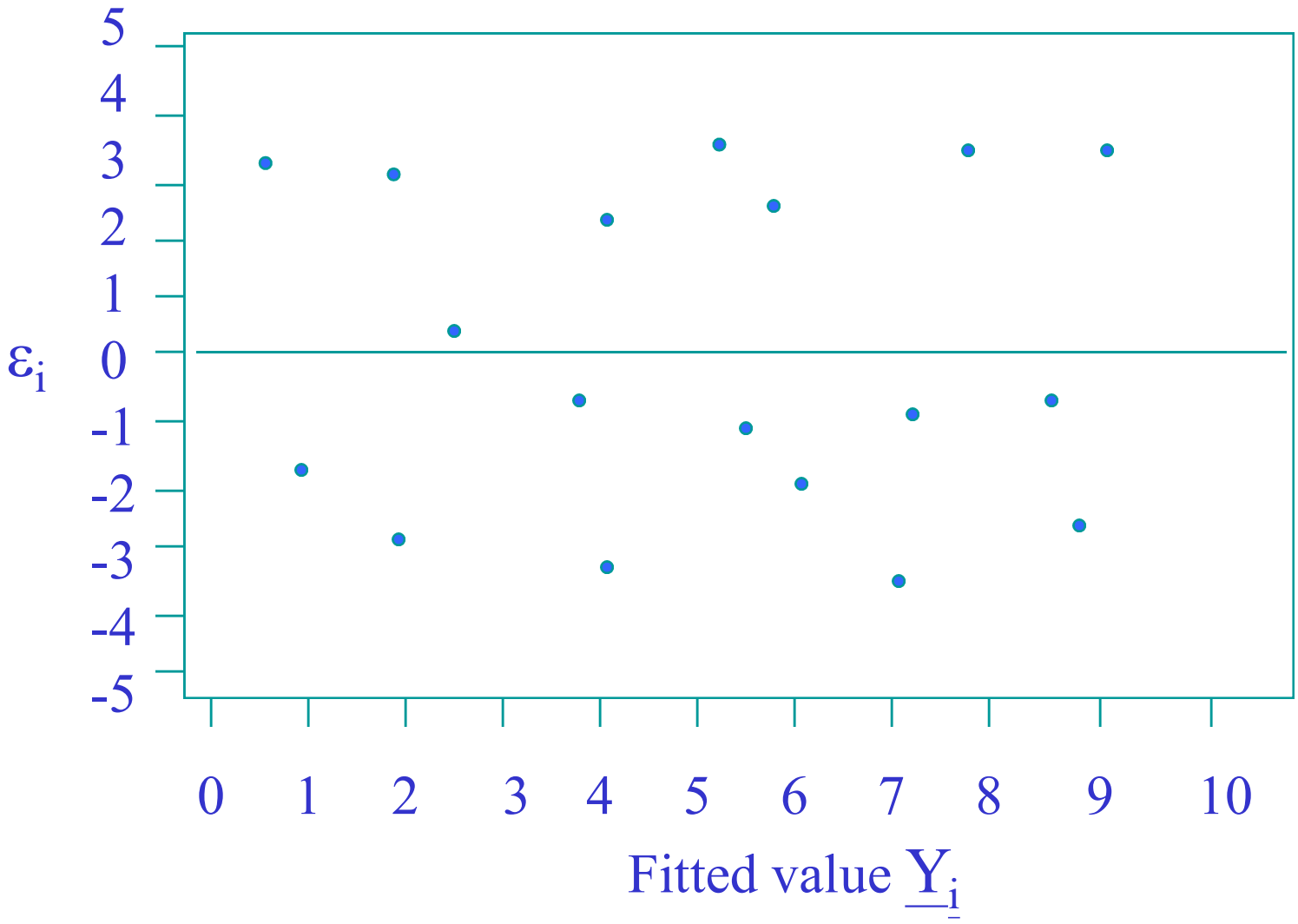
- If we have strong patterns in the residuals plot then this can mean a number of things
  1. The equality of variance assumption has been violated – this is usually shown by a funnel shape in the plot
  2. The simple linear model did not explain the trend in the data, i.e. there is some trend that still exists in the data which might require the addition of extra model terms – this is more likely in multivariate regression
  3. The data require transformation before a linear model is appropriate

Diagnostics for non-linearity of regression function

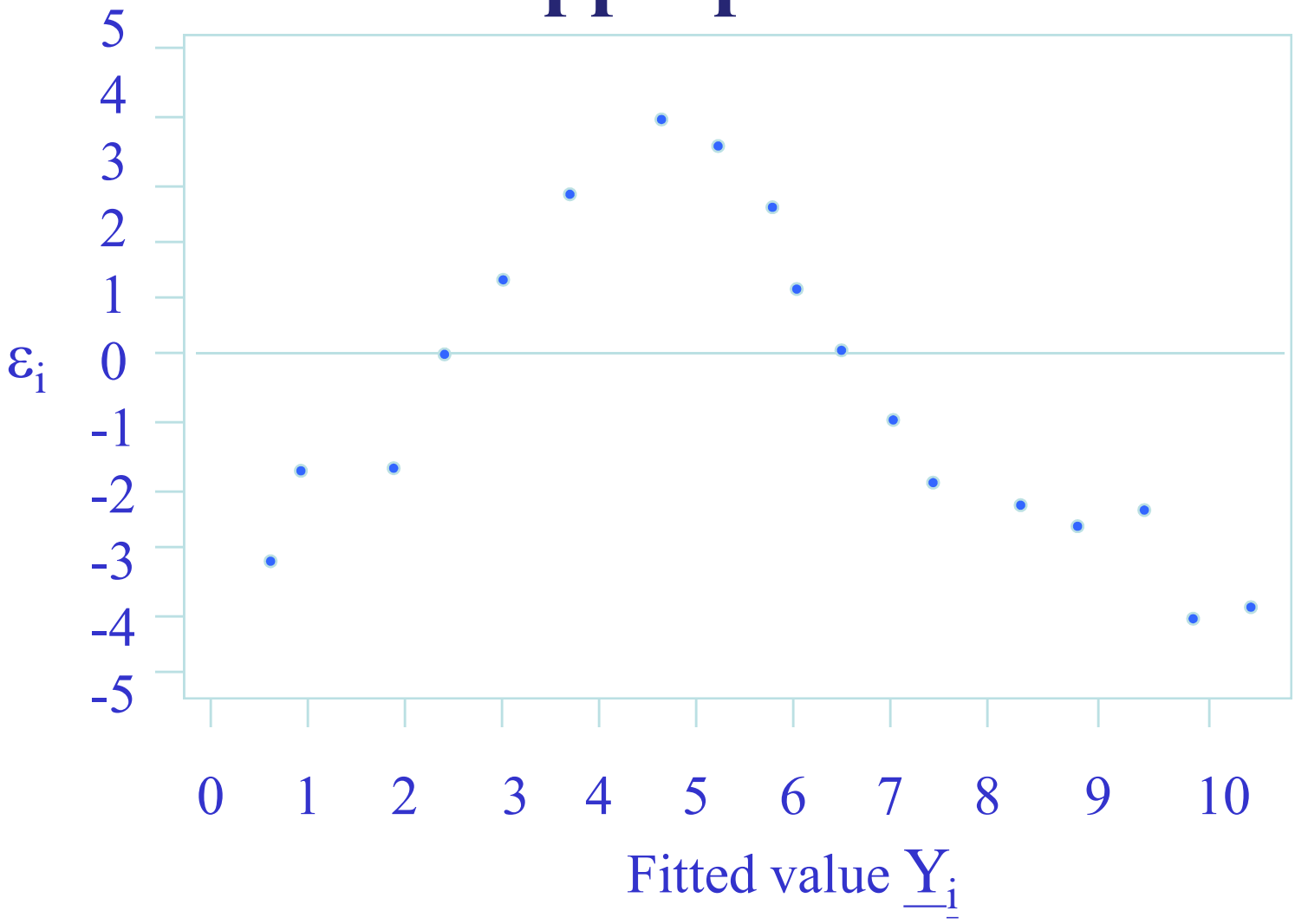
In this case we can use the residual plot to get an indication of this:

- A systematic pattern in the plot of the residuals indicates a non-linear regression.
- The absence of any systematic pattern in the points suggests that a linear regression is appropriate.

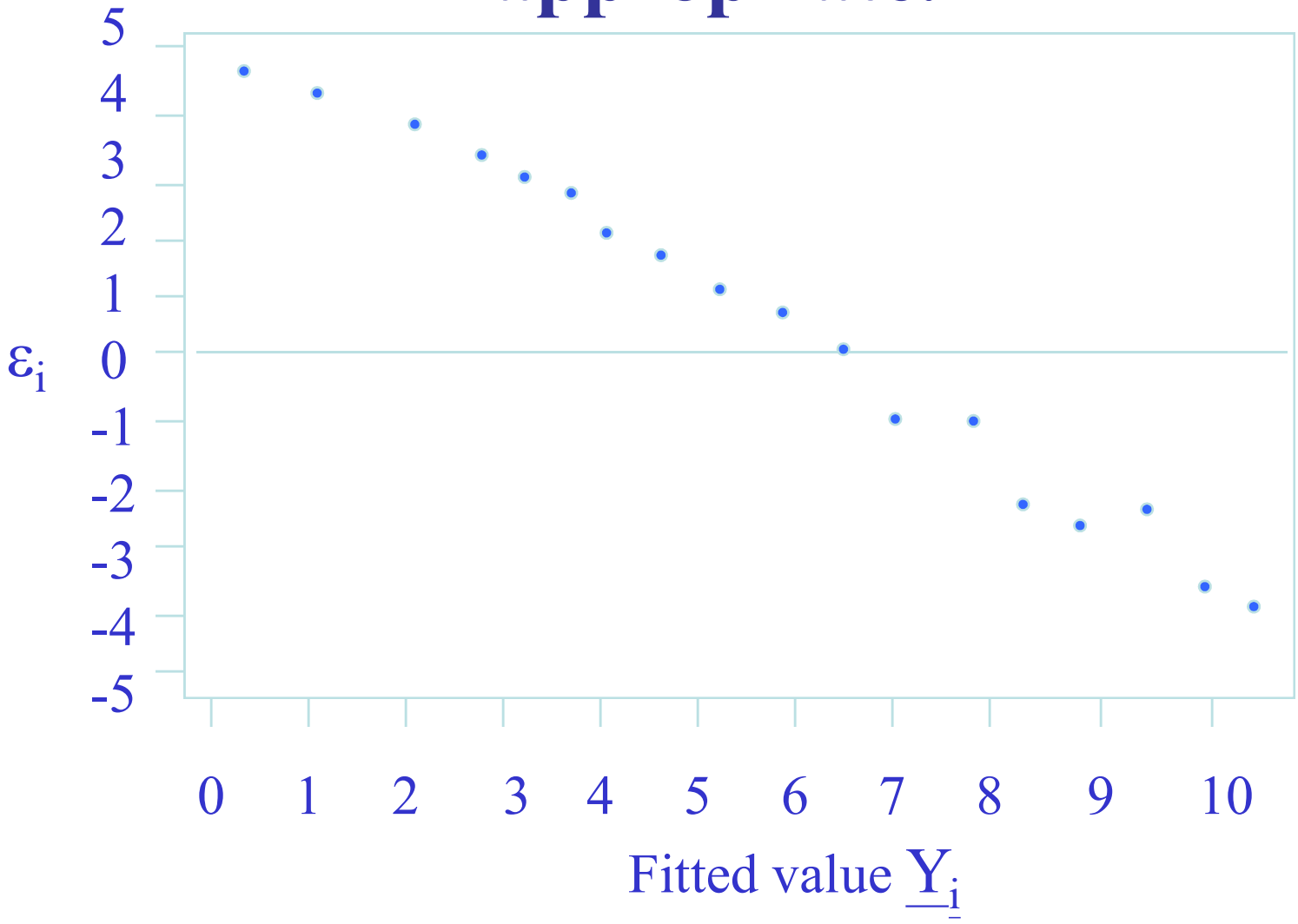
An illustration where the residual plot indicates a linear regression is appropriate.

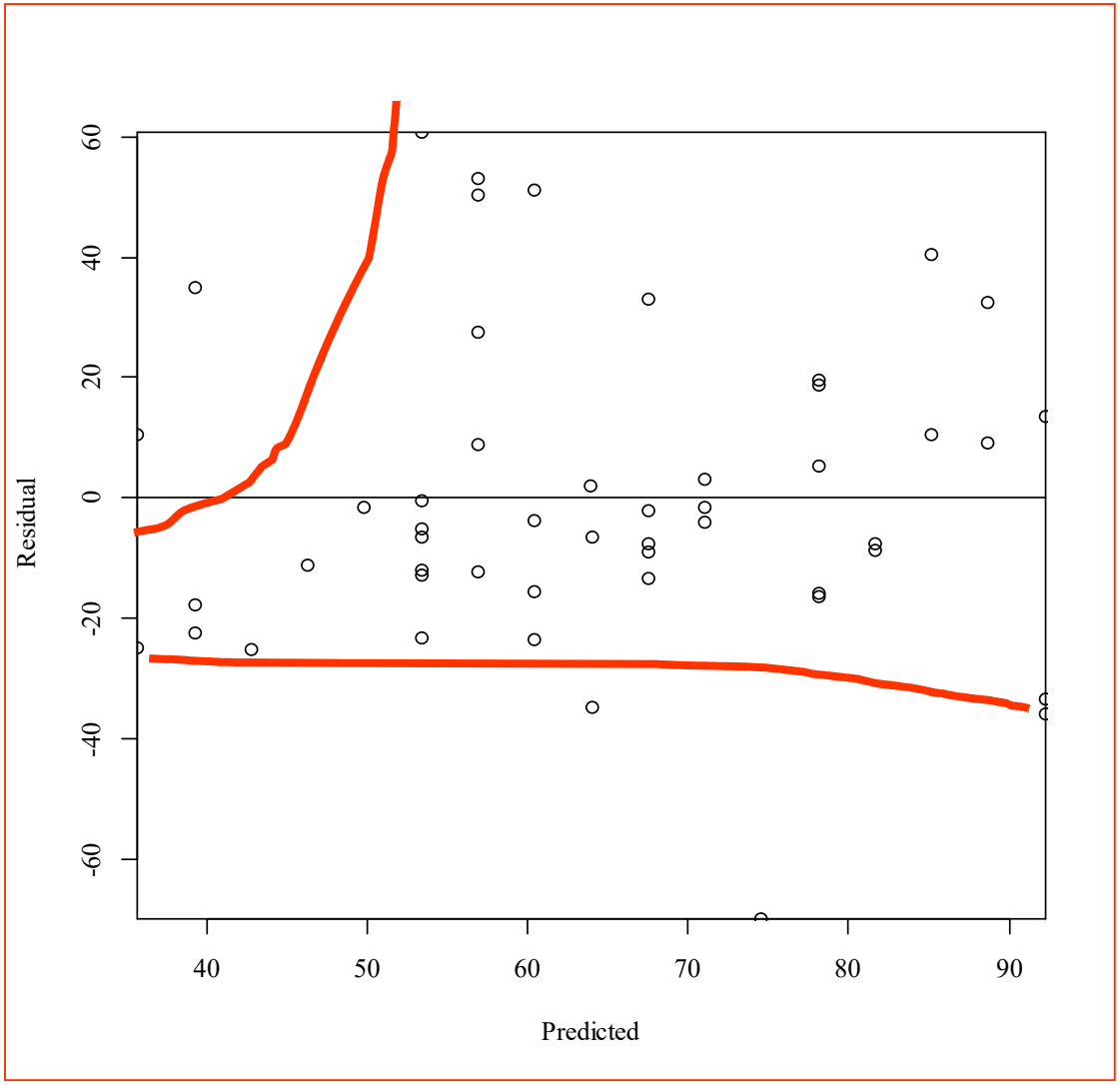


An illustration where the residual plot indicates a linear regression is not appropriate.



Another illustration where the residual plot indicates a linear regression is not appropriate.

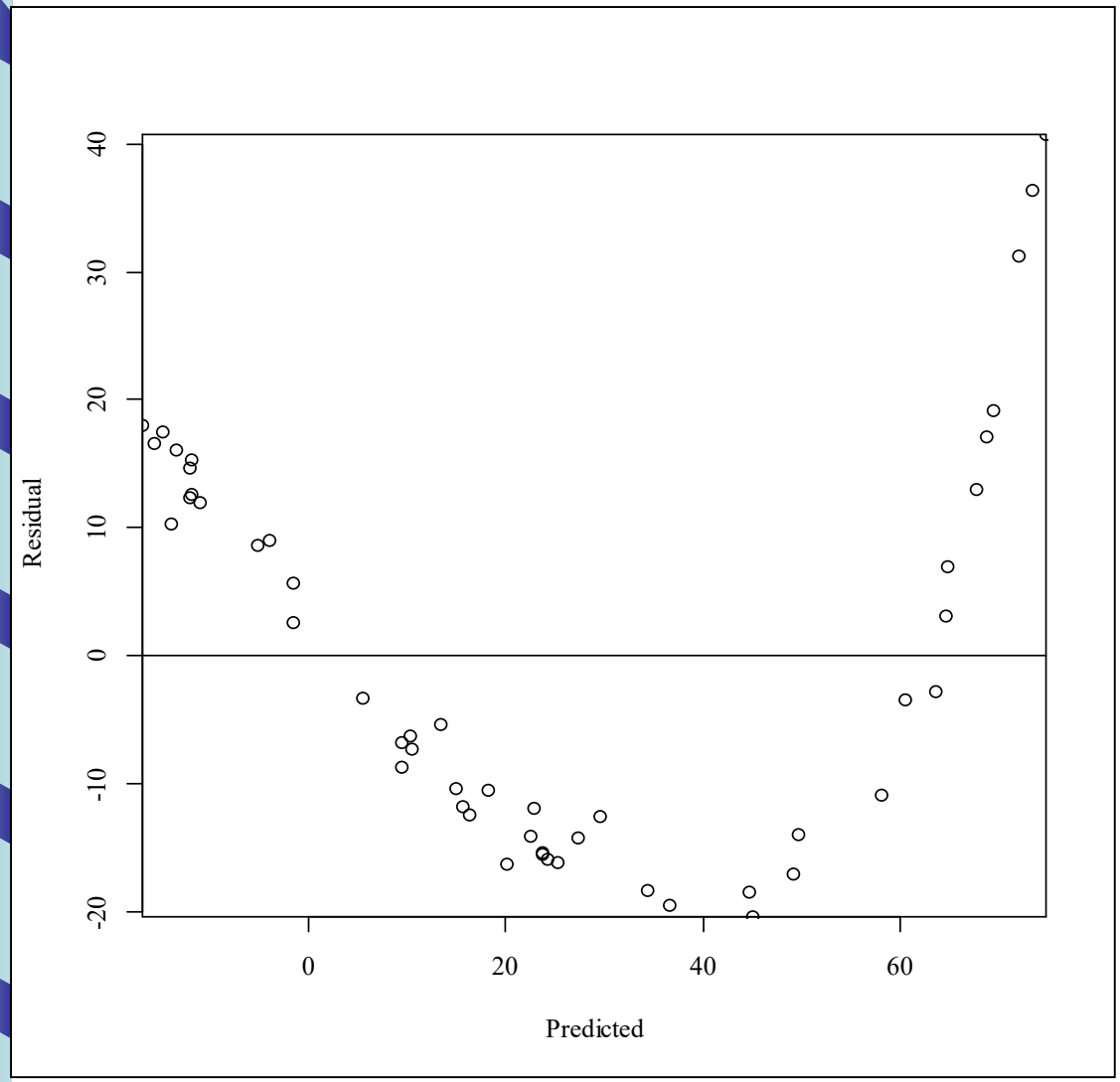




This funnel effect is evidence of "non-homogeneity of variance"

Usually we can get around it by transforming the data or fitting a different model

It is never valid to proceed from this point to the interpretation of the regression coefficients

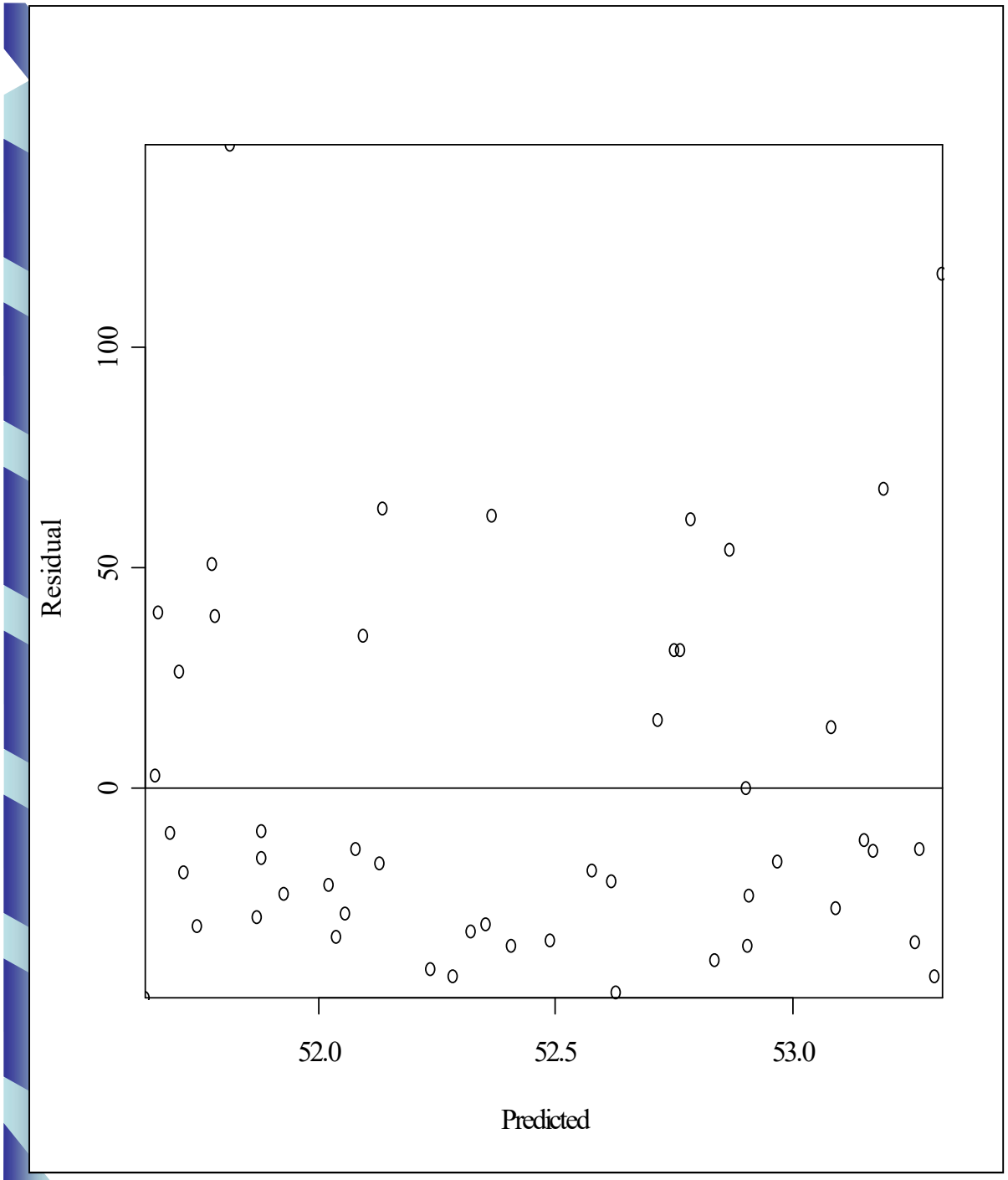


We usually see this type of effect when the real trend is really non-linear

The actual model here is y=exp(5x)

This is definitely a non-linear model

Taking logs would cure this problem – why?



Here we have more negative residuals than positive

The extreme residuals are all positive as well

This says the errors are skewed

This violates our assumption of normality; i.e. this might not be a normal distribution

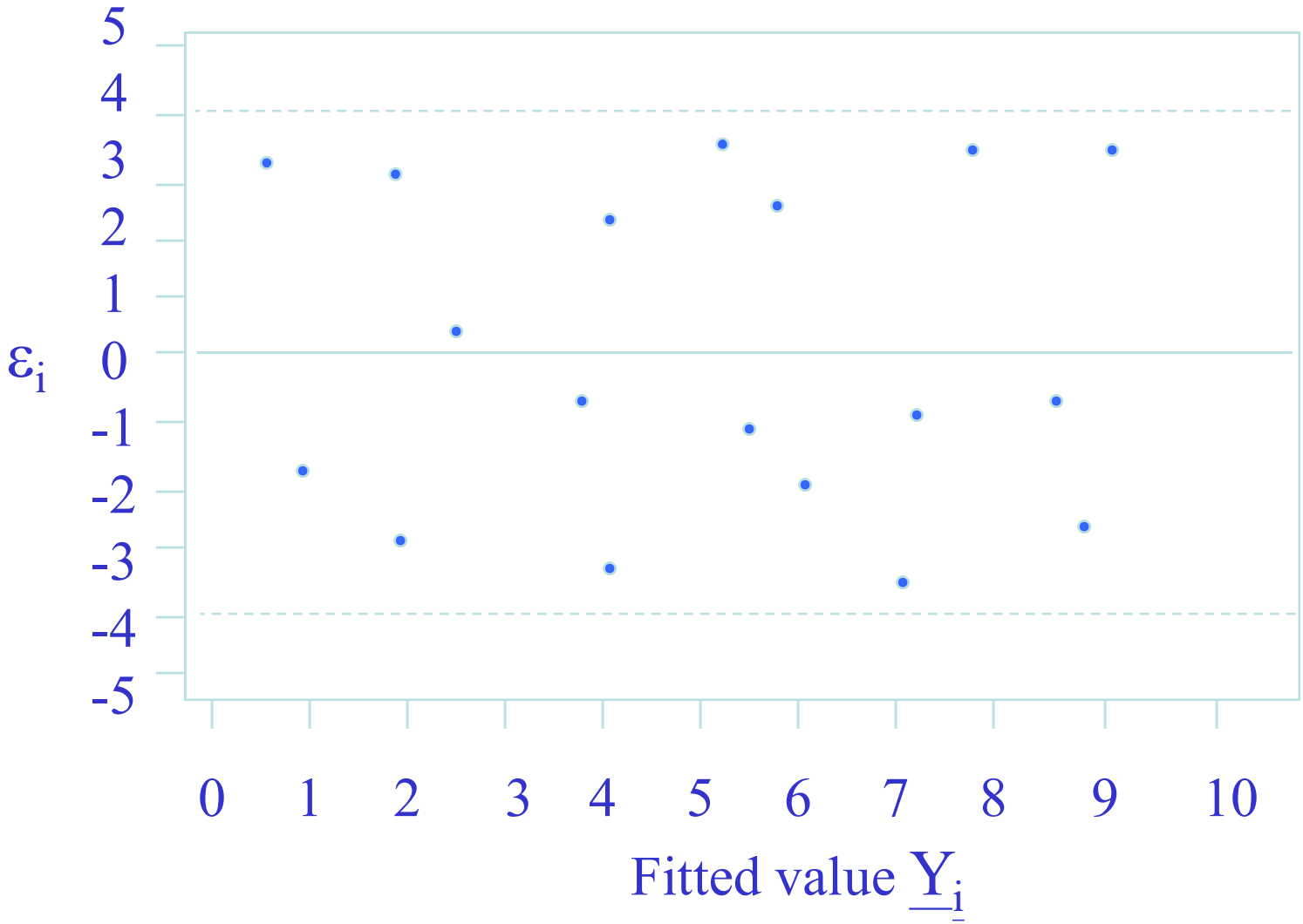
Diagnostics for non-constancy of error variances

Again, in this case, the residual plot will provides information as to weather or not the error terms have constant variance:

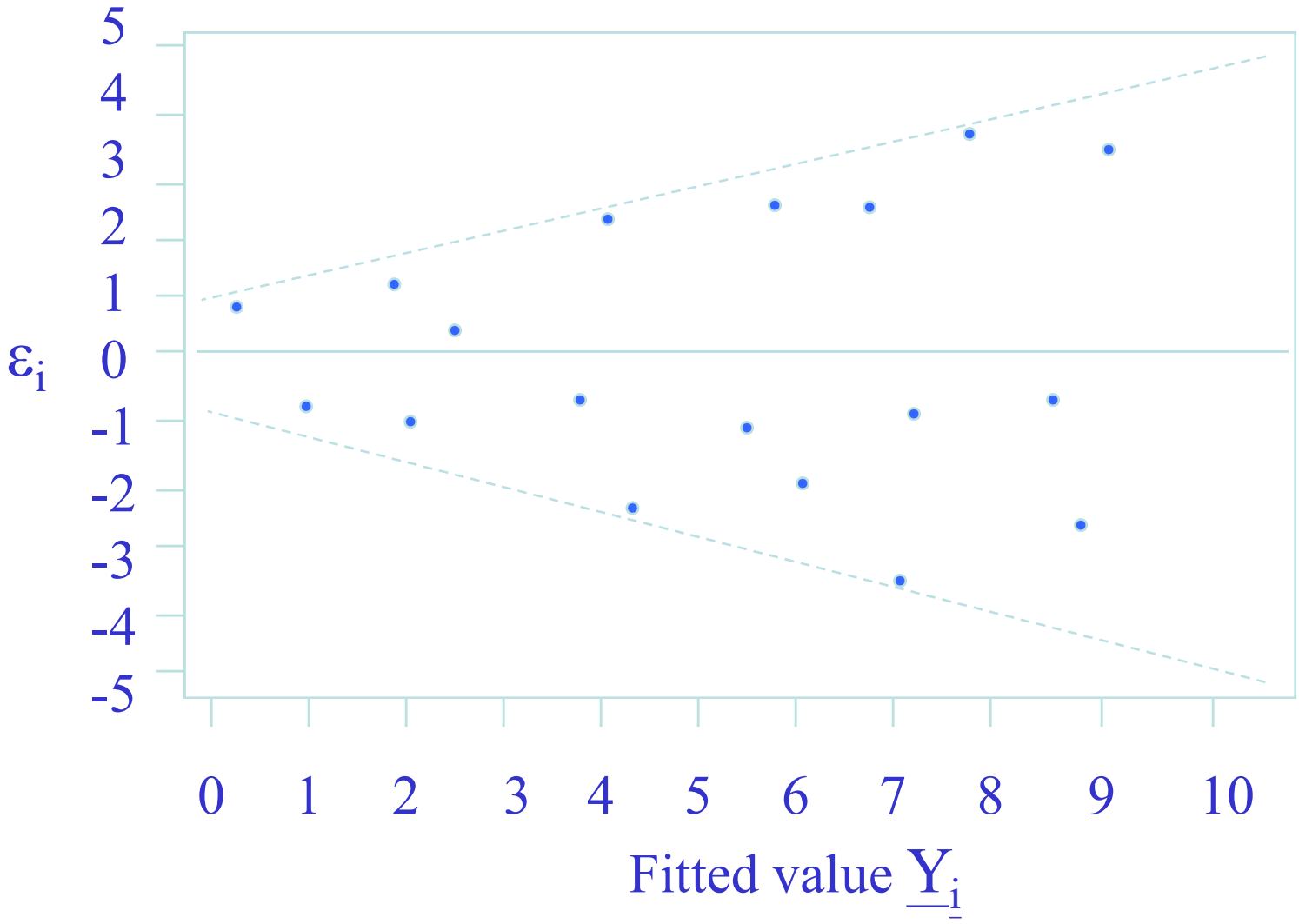
- If the error term variance is constant, then the plot of the residuals should fall within a horizontal band around the center.

- Else, the plot is suggesting that a linear regression is appropriate.

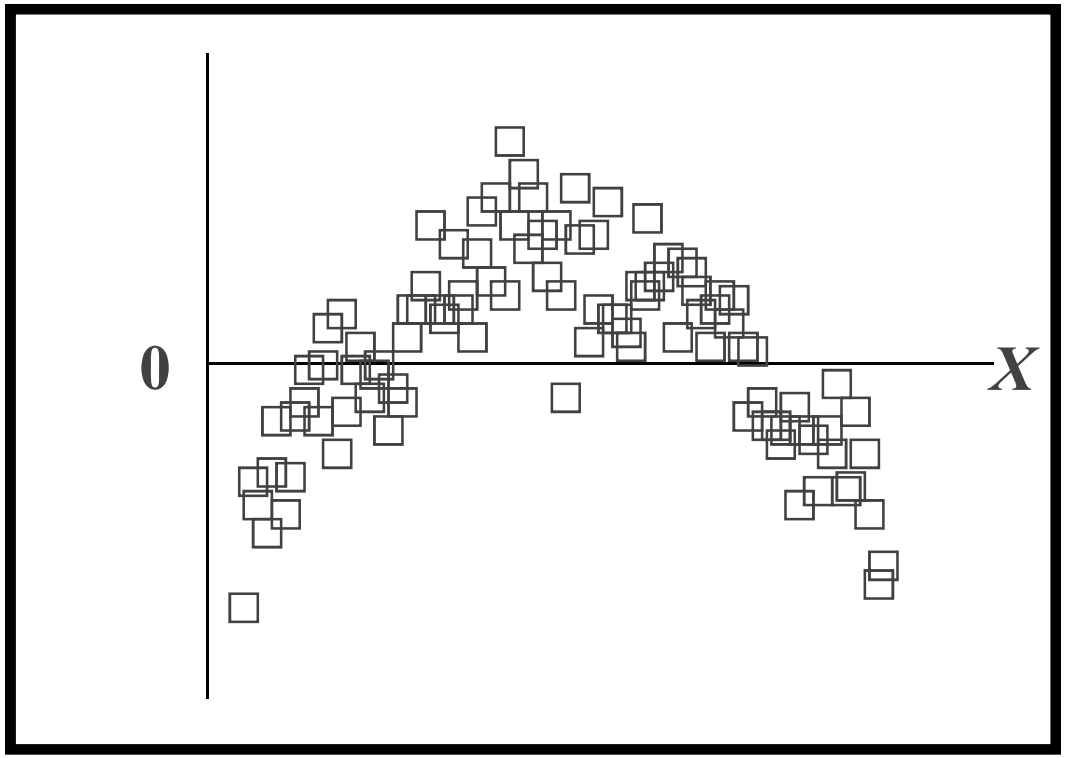
An illustration where the residual plot indicates the error term variance is constant.



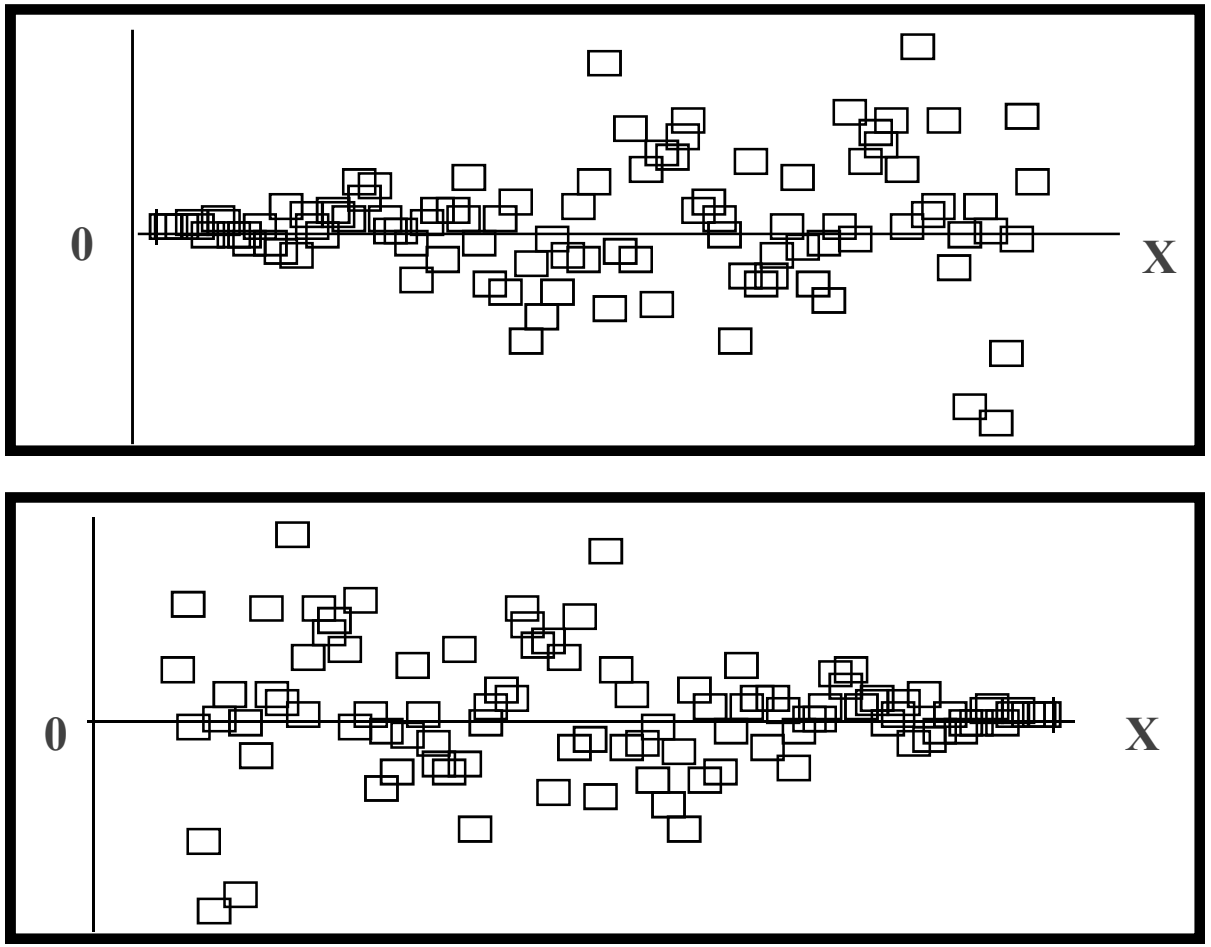
An illustration where the residual plot indicates the error term variance is not constant.



Nonlinear Residual Plot



Nonconstant Error Variance

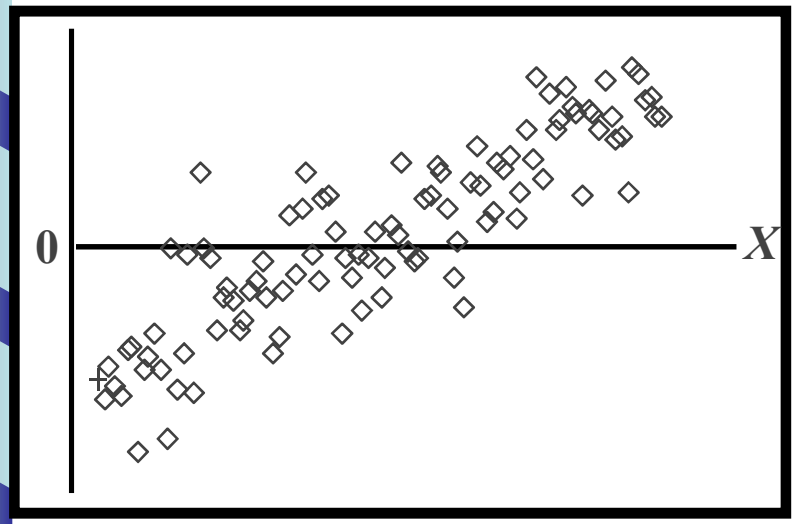


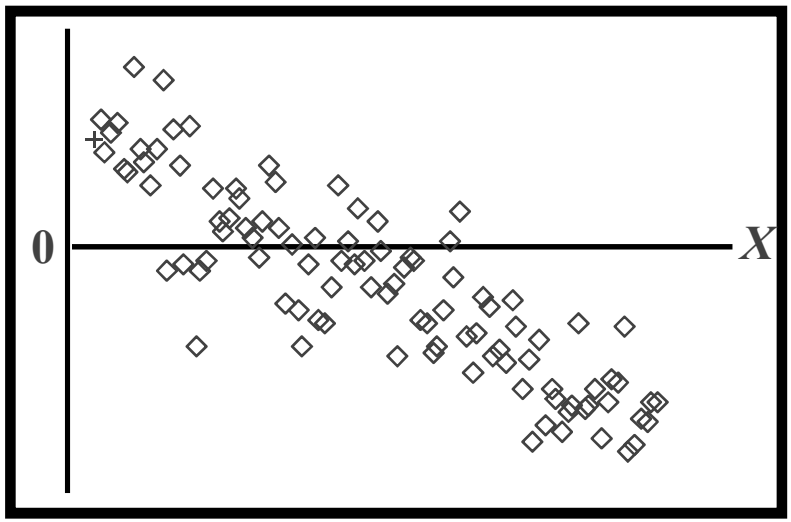
Diagnostics for lack of Independence

In many case the error terms are autocorrelated, when obtained in a time sequence (i.e. increase (or +/-) in this month implies increase in next month (or +/-)), which contradicts the assumption that error terms are assumed to be independent.

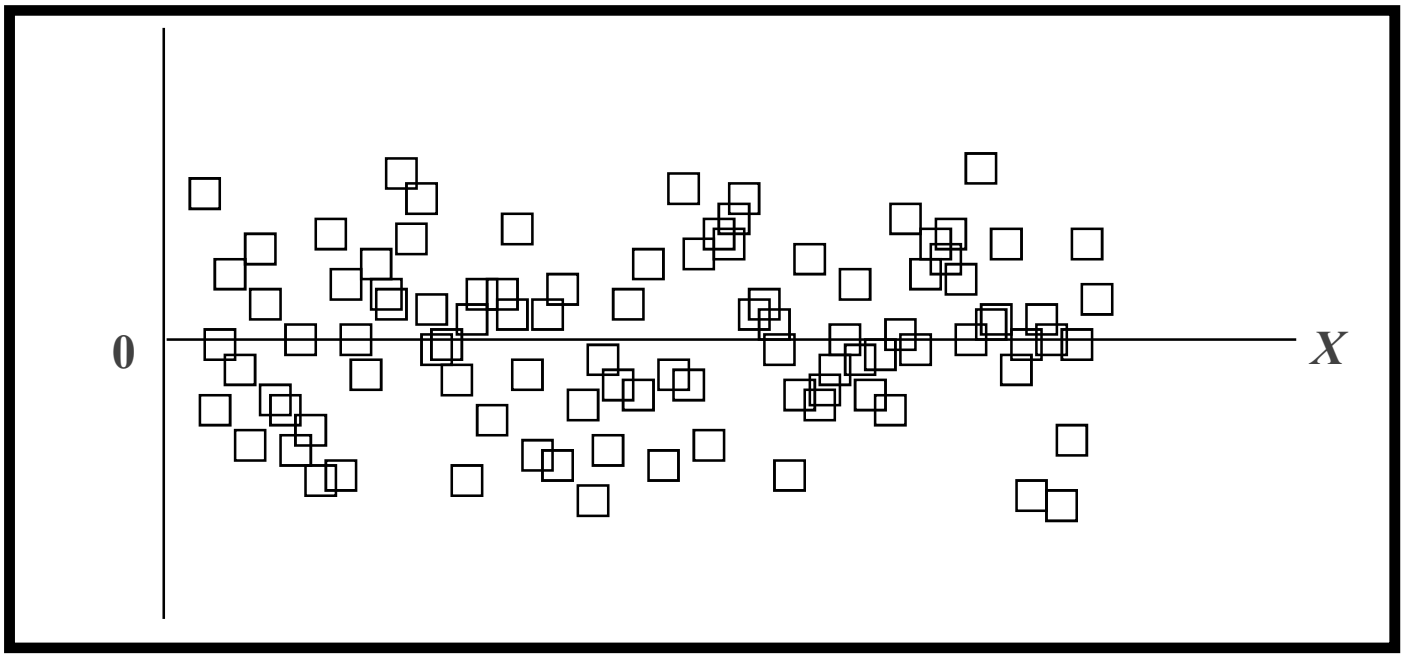
A plot of the residual against the time order (see later) or a formal test for independent of the error terms based on the residual can be used to get this information.

Graphs of Nonindependent Error Terms





Healthy Residual Plot



Stage 4: Estimating the Regression Model and Assessing Overall Model Fit

In Stage 4, the researcher must accomplish three basic tasks:

1. Select a method for specifying the regression model to be estimated,
2. Assess the statistical significance of the overall model in predicting the dependent variable, and
3. Determine whether any of the observations exert an undue influence on the results.

The ANOVA Table

- The ANOVA table reports the variation in the dependent variable. The variation is divided into two components.
- The Explained Variation is that accounted for by the set of independent variable.
- The Unexplained or Random Variation is not accounted for by the independent variables.

Real Estate Data

Observation	Market Price ($1,000) Y	Square Feet X_1	Age (Years) X_2	Observation	Market Price ($1,000) Y	Square Feet X_1	Age (Years) X_2
1	63.0	1,605	35	13	79.7	2,121	14
2	5.1	2,489	45	14	84.5	2,485	9
3	69.9	1,553	20	15	96.0	2,300	19
4	76.8	2,404	32	16	109.5	2,714	4
5	73.9	1,884	25	17	102.5	2,463	5
6	77.9	1,558	14	18	121.0	3,076	7
7	74.9	1,748	8	19	104.9	3,048	3
8	78.0	3,105	10	20	128.0	3,267	6
9	79.0	1,682	28	21	129.0	3,069	10
10	63.4	2,470	30	22	117.9	4,765	11
11	79.5	1,820	2	23	140.0	4,540	8
12	83.9	2,143	6				

MINITAB Output for the Real Estate Example

The regression equation is
Price = 57.4 + 0.0177 Sq.Feet - 0.666 Age

Predictor	Coef	StDev	T	P
Constant	57.35	10.01	5.73	0.000
Sq.Feet	0.017718	0.003146	5.63	0.000
Age	-0.6663	0.2280	-2.92	0.008

S = 11.96 R-Sq = 74.1% R-Sq(adj) = 71.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	8189.7	4094.9	28.63	0.000
Residual Error	20	2861.0	143.1		
Total	22	11050.7			

Evaluating the Multiple Regression Model

$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$

H_a: At least one of the regression coefficients is $\neq 0$

Testing the Overall Model

$H_0: \beta_1 = 0$ $\quad$ $H_0: \beta_3 = 0$

$H_a: \beta_1 \neq 0$ $\quad$ $H_a: \beta_3 \neq 0$

$\vdots$

$H_0: \beta_2 = 0$ $\quad$ $H_0: \beta_k = 0$

$H_a: \beta_2 \neq 0$ $\quad$ $H_a: \beta_k \neq 0$

Significance Tests for Individual Regression Coefficients

Testing the Overall Model for the Real Estate Example

$H_0: \beta_1 = \beta_2 = 0$

H_a: At least one of the regression coefficients is $\neq 0$

$F_{.01,2,20} = 5.85$

$F_{Cal} = 28.63 > 5.85$, reject H_0.

$$MSR = \frac{SSR}{k} \qquad MSE = \frac{SSE}{n-k-1} \qquad F = \frac{MSR}{MSE}$$

ANOVA

	df	SS	MS	F	p
Regression	2	8189.723	4094.86	28.63	.000
Residual (Error)	20	2861.017	143.1		
Total	22	11050.74			

Significance Test of the Regression Coefficients for the Real Estate Example

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

$H_0: \beta_2 = 0$

$H_a: \beta_2 \neq 0$

$t_{.025,20} = 2.086$

$t_{Cal} = 5.63 > 2.086$, reject H_0.

	Coefficients	Std Dev	t Stat	p
x_1 (Sq.Feet)	0.0177	0.003146	5.63	.000
x_2 (Age)	-0.666	0.2280	-2.92	.008

Predicting the Price of Home

$$\hat{Y} = 57.4 + 0.0177\,X_1 - 0.666\,X_2$$

For $X_1 = 2500$ and $X_2 = 12$,

$$\hat{Y} = 57.4 + 0.0177(2500) - 0.666(12)$$

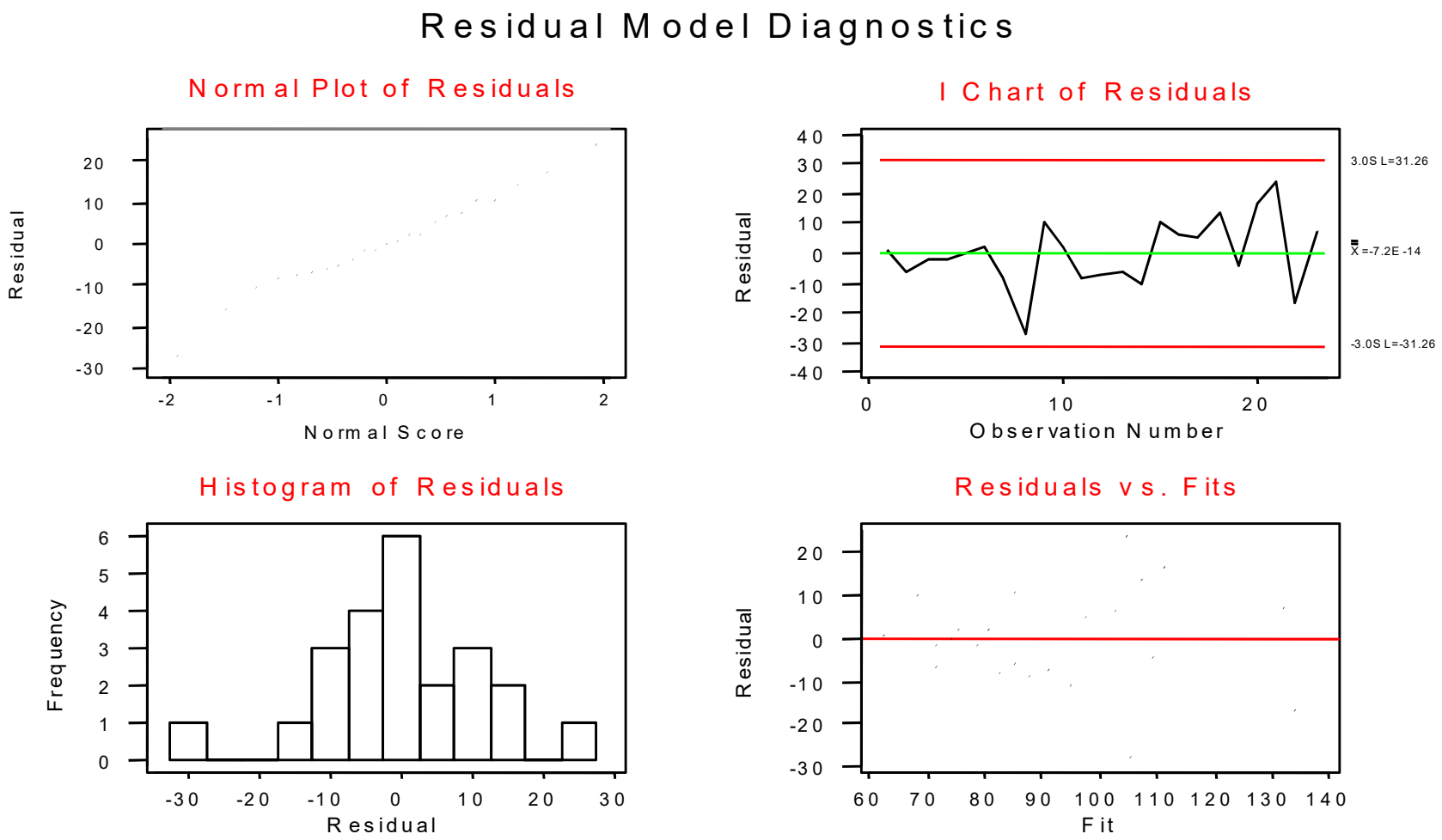
$$= 93.658 \text{ thousand dollars}$$

Residuals and Sum of Squares Error for the Real Estate Example

Observation	Y	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$	*Observation*	Y	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
1	43.0	42.466	0.534	0.285	13	59.7	65.602	-5.902	34.832
2	45.1	51.465	-6.365	40.517	14	64.5	75.383	-10.883	118.438
3	49.9	51.540	-1.640	2.689	15	76.0	65.442	10.558	111.479
4	56.8	58.622	-1.822	3.319	16	89.5	82.772	6.728	45.265
5	53.9	54.073	-0.173	0.030	17	82.5	77.659	4.841	23.440
6	57.9	55.627	2.273	5.168	18	101.0	87.187	13.813	190.799
7	54.9	62.991	-8.091	65.466	19	84.9	89.356	-4.456	19.858
8	58.0	85.702	-27.702	767.388	20	108.0	91.237	16.763	280.982
9	59.0	48.495	10.505	110.360	21	109.0	85.064	23.936	572.936
10	63.4	61.124	2.276	5.181	22	97.9	114.447	-16.547	273.815
11	59.5	68.265	-8.765	76.823	23	120.0	112.460	7.540	56.854
12	63.9	71.322	-7.422	55.092					

SSE 2861.017

Some Residual Diagnostics for the Real Estate Problem



Multiple Standard Error of Estimate

The multiple standard error of estimate is a measure of the effectiveness of the regression equation.

- It is measured in the same units as the dependent variable.
- It is difficult to determine what is a large value and what is a small value of the standard error.

Multiple Standard Error of Estimate

- The formula is:

$$s_e = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - (k + 1)}}$$

SSE and Standard Error of the Estimate

ANOVA

	df	SS	MS	F	P
Regression	2	8189.7	4094.9	28.63	.000
Residual (Error)	20	2861.0	143.1		
Total	22	11050.7			

SSE

$$S_e = \sqrt{\frac{SSE}{n-k-1}}$$

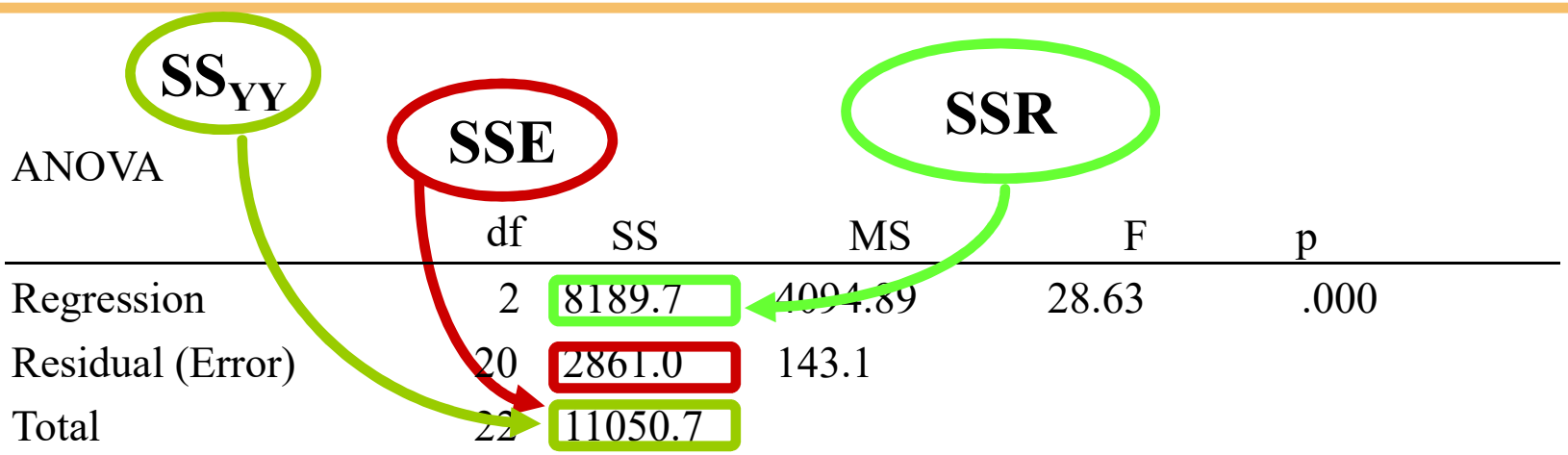
$$= \sqrt{\frac{2861}{23-2-1}}$$

$$= 11.96$$

where: n = number of observations

k = number of independent variables

Coefficient of Multiple Determination (R^2)



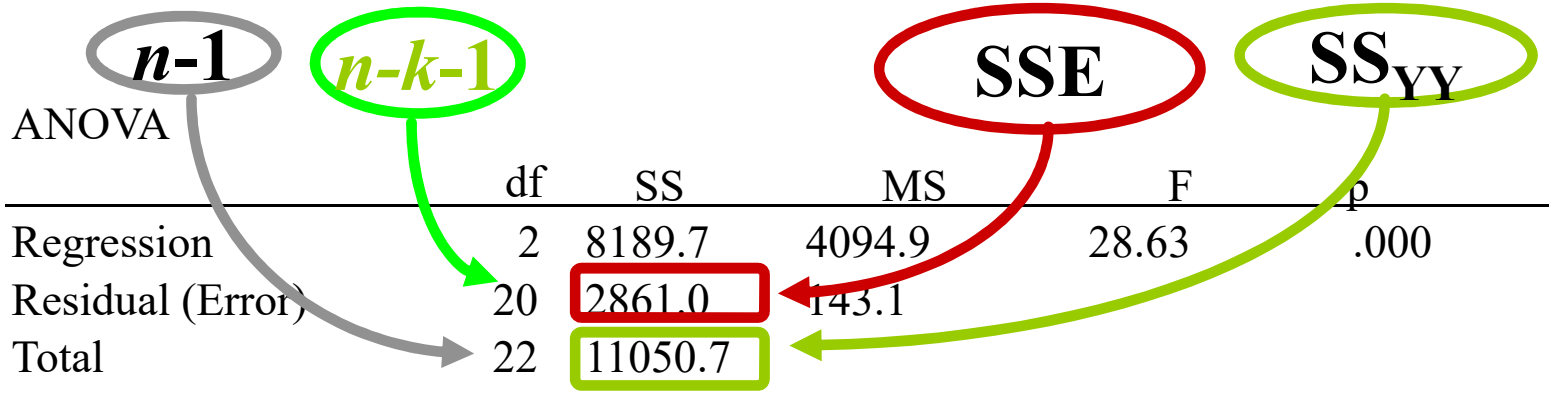
ANOVA

	df	SS	MS	F	p
Regression	2	8189.7	4094.89	28.63	.000
Residual (Error)	20	2861.0	143.1		
Total	22	11050.7			

$$R^2 = \frac{SSR}{SSY} = \frac{8189.723}{11050.74} = .741$$

$$R^2 = 1 - \frac{SSE}{SSY} = 1 - \frac{2861.017}{11050.74} = .741$$

Adjusted R^2



$$adj.R^2 = 1 - \frac{\frac{SSE}{n-k-1}}{\frac{SSY}{n-1}} = 1 - \frac{\frac{2861.017}{23-2-1}}{\frac{11050.74}{23-1}} = 1 - .285 = .715$$

THE SQUARED MULTIPLE CORRELATION

The statistic

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2}$$

is the proportion of the variation of the response variable y that is explained by the explanatory variables $x_1, x_2, \ldots, x_p$ in a multiple linear regression.

Global Test

- The global test is used to investigate whether any of the independent variables have significant coefficients. The hypotheses are:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$H_1 : \text{Not all } \beta \text{ s equal } 0$$

Global Test *continued*

- The test statistic is the *F* distribution with *k* (number of independent variables) and *n-(k+1)* degrees of freedom, where *n* is the sample size.

$$F = \frac{SSR \ / k}{SSE \ / n - (k + 1)}$$

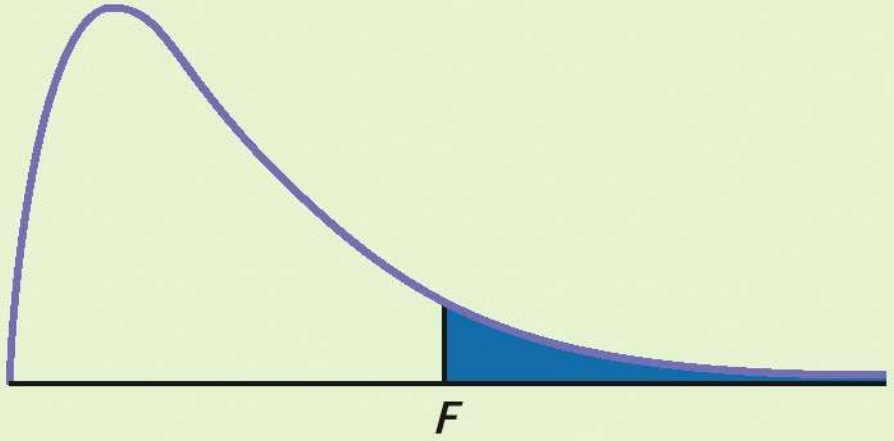
ANALYSIS OF VARIANCE *F* TEST

In the multiple regression model, the hypothesis

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

is tested by the analysis of variance F statistic

$$F = \frac{\text{MSR}}{\text{MSE}}$$



The P-value is the probability that a random variable having the $F(p, n - p - 1)$ distribution is greater than or equal to the calculated value of the F statistic.

F TEST FOR A COLLECTION OF REGRESSION COEFFICIENTS

In the multiple regression model with p explanatory variables, the hypothesis

H_0: q specific explanatory variables all have zero coefficients

is tested by an F statistic. The degrees of freedom are q and $n - p - 1$. The P-value is the probability that a random variable having the $F(q, n - p - 1)$ distribution is greater than or equal to the calculated value of the F statistic.

Test for Individual Variables

- The variables that have zero regression coefficients are usually dropped from the analysis
- This test is used to determine which independent variables have nonzero regression coefficients..
- The test statistic is the *t* distribution with *n*-(*k*+1) degrees of freedom.

$$t = \frac{b_i - 0}{S_{b_i}}$$

EXAMPLE

A market researcher for Super Dollar Super Markets is studying the yearly amount families of four or more spend on food. Three independent variables are thought to be related to yearly food expenditures (Food). Those variables are: total family income (Income) in $00, size of family (Size), and whether the family has children in college (College).

Example *continued*

Note the following regarding the regression equation:

- The variable college is called a dummy or indicator variable. It can take only one of two possible outcomes. That is a child is a college student or not.
- Other examples of dummy variables include gender, the part is acceptable or unacceptable, the voter will or will not vote for the incumbent governor.
- We usually code one value of the dummy variable as "1" and the other "0".

EXAMPLE *continued*

Family	Food	Income	Size	Student
1	3900	376	4	0
2	5300	515	5	1
3	4300	516	4	0
4	4900	468	5	0
5	6400	538	6	1
6	7300	626	7	1
7	4900	543	5	0
8	5300	437	4	0
9	6100	608	5	1
10	6400	513	6	1
11	7400	493	6	1
12	5800	563	5	0

EXAMPLE *continued*

- Use a computer software package, such as MINITAB or Excel, to develop a correlation matrix.
- From the analysis provided by MINITAB, write out the regression equation:

$$\hat{Y} = 954 + 1.09X_1 + 748X_2 + 565X_3$$

- What food expenditure would you estimate for a family of 4, with no college students, and an income of $50,000 (which is input as 500)?

Example *continued*

The regression equation is
Food = 954 + 1.09 Income + 748 Size + 565 Student

Predictor	Coef	SE Coef	T	P
Constant	954	1581	0.60	0.563
Income	1.092	3.153	0.35	0.738
Size	748.4	303.0	2.47	0.039
Student	564.5	495.1	1.14	0.287

S = 572.7 R-Sq = 80.4% R-Sq(adj) = 73.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	10762903	3587634	10.94	0.003
Residual Error	8	2623764	327970		
Total	11	13386667			

Example *continued*

From the regression output we note:

- Each additional $100 dollars of income per year will increase the amount spent on food by $109 per year.
- The coefficient of determination is 80.4 percent. This means that more than 80 percent of the variation in the amount spent on food is accounted for by the variables income, family size, and student.
- An additional family member will increase the amount spent per year on food by $748.
- A family with a college student will spend $565 more per year on food than those without a college student.

EXAMPLE *continued*

The estimated food expenditure for a family of 4 with a \$500 (that is \$50,000) income and no college student is \$4,491.

$$\hat{\mathbf{Y}} = 954 + 1.09(500) + 748(4) + 565\ (0)$$
$$= 4491$$

Example *continued*

Conduct a global test of hypothesis to determine if any of the regression coefficients are not zero.

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad H_1 : \text{at least one } \beta \neq 0$

- H_0 is rejected if $F>4.07$.
- From the MINITAB output, the computed value of F is 10.94.
- Decision: H_0 is rejected. Not all the regression coefficients are zero

EXAMPLE *continued*

- Conduct an individual test to determine which coefficients are not zero. This is the hypotheses for the independent variable family size.

$$H_0: \beta_2 = 0 \qquad H_1: \beta_2 \neq 0$$

- From the MINITAB output, the only significant variable is FAMILY (family size) using the *p*-values. The other variables can be omitted from the model.
- Thus, using the 5% level of significance, reject H_0 if the *p*-value<.05

EXAMPLE *continued*

- We rerun the analysis using only the significant independent family size.
- The new regression equation is:

$$Y' = 340 + 1031X_2$$

- The coefficient of determination is 76.8 percent. We dropped two independent variables, and the R-square term was reduced by only 3.6 percent.

Example *continued*

Regression Analysis: Food versus Size

```
The regression equation is
Food = 340 + 1031 Size

Predictor         Coef     SE Coef           T         P
Constant         339.7       940.7        0.36     0.726
Size            1031.0       179.4        5.75     0.000

S = 557.7       R-Sq = 76.8%     R-Sq(adj) = 74.4%

Analysis of Variance

Source             DF           SS           MS          F
  P
Regression          1     10275977     10275977      33.03
  0.000
Residual Error     10      3110690       311069
Total              11     13386667
```

Example *continued*

- The correlation matrix is as follows:

	Food	Income	Size
Income	0.587		
Size	0.876	0.609	
Student	0.773	0.491	0.743

- The strongest correlation between the dependent variable and an independent variable is between family size and amount spent on food.
- None of the correlations among the independent variables should cause problems. All are between –.70 and .70.

Variable Selection Approaches

- Confirmatory (Simultaneous)
- Sequential Search Methods:
  - ✓ Stepwise (variables not removed once included in regression equation).
  - ✓ Forward Inclusion & Backward Elimination.
  - ✓ Hierarchical.
- Combinatorial (All-Possible-Subsets)

Description of HBAT Primary Database Variables

Variable	Description	Variable Type
Data Warehouse Classification Variables		
X1	Customer Type	nonmetric
X2	Industry Type	nonmetric
X3	Firm Size	nonmetric
X4	Region	nonmetric
X5	Distribution System	nonmetric
Performance Perceptions Variables		
X6	Product Quality	metric
X7	E-Commerce Activities/Website	metric
X8	Technical Support	metric
X9	Complaint Resolution	metric
X10	Advertising	metric
X11	Product Line	metric
X12	Salesforce Image	metric
X13	Competitive Pricing	metric
X14	Warranty & Claims	metric
X15	New Products	metric
X16	Ordering & Billing	metric
X17	Price Flexibility	metric
X18	Delivery Speed	metric
Outcome/Relationship Measures		
X19	Satisfaction	metric
X20	Likelihood of Recommendation	metric
X21	Likelihood of Future Purchase	metric
X22	Current Purchase/Usage Level	metric
X23	Consider Strategic Alliance/Partnership in Future	nonmetric

Rules of Thumb 4–5

Estimation Techniques

- No matter which estimation technique is chosen, theory must be a guiding factor in evaluating the final regression model because:

1) Confirmatory Specification, the only method to allow direct testing of a pre-specified model, is also the most complex from the perspectives of specification error, model parsimony and achieving maximum predictive accuracy.
2) Sequential search (e.g., stepwise), while maximizing predictive accuracy, represents a completely "automated" approach to model estimation, leaving the researcher almost no control over the final model specification

Rules of Thumb 4–5

Estimation Techniques

3) Combinatorial estimation, while considering all possible models, still removes control from the researcher in terms of final model specification even though the researcher can view the set of roughly equivalent models in terms of predictive accuracy.

- No single method is "Best" and the prudent strategy is to use a combination of approaches to capitalize on the strengths of each to reflect the theoretical basis of the research question.

Regression Coefficient Questions

Three questions about the statistical significance of any regression coefficient:

1) Was statistical significance established?
2) How does the sample size come into play?
3) Does it have practical significance in addition to statistical significance?

Rules of Thumb 4–6

Statistical Significance and Influential Observations

- Always ensure practical significance when using large sample sizes, as the model results and regression coefficients could be deemed irrelevant even when statistically significant due just to the statistical power arising from large sample sizes.
- Use the adjusted R^2 as your measure of overall model predictive accuracy.
- Statistical significance is required for a relationship to have validity, but statistical significance without theoretical support does not support validity.
- While outliers may be easily identifiable, the other forms of influential observations requiring more specialized diagnostic methods can be equal to or even more impactful on the results.

Types of Influential Observations

- Leverage points are observations that are distinct from the remaining observations based on their independent variable values.
- Influential observations are the broadest category, including all observations that have a disproportionate effect on the regression results. Influential observations potentially include outliers and leverage points but may include other observations as well.

Types of Influential Observations

Influential observations . . . include all observations that have a disproportionate effect on the regression results. There are three basic types based upon the nature of their impact on the regression results:

- Outliers are observations that have large residual values and can be identified only with respect to a specific regression model.

Corrective Actions for Influentials

Influentials, outliers, and leverage points are based on one of four conditions, each of which has a specific course of corrective action:

1. An error in observations or data entry – remedy by correcting the data or deleting the case,
2. A valid but exceptional observation that is explainable by an extraordinary situation – remedy by deletion of the case unless variables reflecting the extraordinary situation are included in the regression equation,

Corrective Actions for Influentials

3. An exceptional observation with no likely explanation – presents a special problem because there is no reason for deleting the case, but its inclusion cannot be justified either, suggesting analyses with and without the observations to make a complete assessment, and
4. An ordinary observation in its individual characteristics but exceptional in its combination of characteristics – indicates modifications to the conceptual basis of the regression model and should be retained.

Assessing Multicollinearity

The researcher's task is to . . .

- Assess the degree of multicollinearity,
- Determine its impact on the results, and
- Apply the necessary remedies if needed.

Multiple Regression and Correlation Assumptions

- The independent variables and the dependent variable have a linear relationship.
- The dependent variable must be continuous and at least interval-scale. The variation in $(Y-\hat{Y})$ or residual must be the same for all values of Y. When this is the case, we say the difference exhibits homoscedasticity.
- The residuals should follow the normal distributed with mean 0.
- Successive values of the dependent variable must be uncorrelated. Violation of this assumption is called autocorrelation.

Correlation Matrix

A correlation matrix is used to show all possible simple correlation coefficients among the variables.

- The matrix is useful for locating correlated independent variables.
- It shows how strongly each independent variable is correlated with the dependent variable.
- The formula is:

$$r_{jY} = \frac{S_{X_j Y}}{\sqrt{S_{X_j X_j} S_{YY}}}$$

An example of correlation matrix

This is an example of a correlation matrix for 2 independent variables:

$$\begin{array}{c} \\ X_1 \\ X_2 \\ Y \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & Y \end{array} \\ \begin{pmatrix} 1 & r_{12} & r_{1Y} \\ r_{21} & 1 & r_{2Y} \\ r_{Y1} & r_{Y2} & 1 \end{pmatrix} \end{array} = \begin{pmatrix} 1 & 0.12 & 0.96 \\ 0.12 & 1 & 0.25 \\ 0.96 & 0.25 & 1 \end{pmatrix}$$

Multicollinearity

Condition that occurs when two or more of the independent variables of a multiple regression model are highly correlated

- Difficult to interpret the estimates of the regression coefficients
- Inordinately small t values for the regression coefficients
- Standard deviations of regression coefficients are overestimated
- Sign of predictor variable's coefficient opposite of what expected

Correlations among Oil Production Predictor Variables

	Energy Consumption	Nuclear	Coal	Dry Gas	Fuel Rate
Energy Consumption	1	0.856	0.791	0.057	0.791
Nuclear	0.856	1	0.952	-0.404	0.972
Coal	0.791	0.952	1	-0.448	0.968
Dry Gas	0.057	-0.404	-0.448	1	-0.423
Fuel Rate	0.796	0.972	0.968	-0.423	1

Multicollinearity Diagnostics

- Variance Inflation Factor (VIF) – measures how much the variance of the regression coefficients is inflated by multicollinearity problems. If VIF equals 0, there is no correlation between the independent measures. A VIF measure of 1 is an indication of some association between predictor variables, but generally not enough to cause problems. A maximum acceptable VIF value would be 10; anything higher would indicate a problem with multicollinearity.

.

Multicollinearity Diagnostics

- Tolerance – the amount of variance in an independent variable that is not explained by the other independent variables. If the other variables explain a lot of the variance of a particular independent variable we have a problem with multicollinearity. Thus, small values for tolerance indicate problems of multicollinearity. The minimum cutoff value for tolerance is typically .10. That is, the tolerance value must be smaller than .10 to indicate a problem of multicollinearity.

Interpretation of Regression Results

- Coefficient of Determination
- Regression Coefficients
  (Unstandardized – bivariate)
- Beta Coefficients (Standardized)
- Variables Entered
- Multicollinearity ??

Rules of Thumb 4–7

Interpreting the Regression Variate

- Interpret the impact of each independent variable relative to the other variables in the model, as model respecification can have a profound effect on the remaining variables:
  - ✓ Use beta weights when comparing relative importance among independent variables.
  - ✓ Regression coefficients describe changes in the dependent variable, but can be difficult in comparing across independent variables if the response formats vary.
- Multicollinearity may be considered "good" when it reveals a suppressor effect, but generally it is viewed as harmful since increases in multicollinearity:
  - ✓ reduce the overall R^2 that can be achieved,
  - ✓ confound estimation of the regression coefficients, and
  - ✓ negatively affect the statistical significance tests of coefficients.

Rules of Thumb 4–7 continued . . .

Interpreting the Regression Variate

- Generally accepted levels of multicollinearity (tolerance values up to .10, corresponding to a VIF of 10) almost always indicate problems with multicollinearity, but these problems may be seen at much lower levels of collinearity or multicollinearity.
  - ✓ Bivariate correlations of .70 or higher may result in problems, and even lower correlations may be problematic if they are higher than the correlations between the dependent and independent variables.
  - ✓ Values much lower than the suggested thresholds (VIF values of even 3 to 5) may result in interpretation or estimation problems, particularly when the relationships with the dependent variable are weaker.

Residuals Plots

- Histogram of standardized residuals – enables you to determine if the errors are normally distributed.
- Normal probability plot – enables you to determine if the errors are normally distributed. It compares the observed (sample) standardized residuals against the expected standardized residuals from a normal distribution.
- Scatter Plot of residuals – can be used to test regression assumptions. It compares the standardized predicted values of the dependent variable against the standardized residuals from the regression equation. If the plot exhibits a random pattern then this indicates no identifiable violations of the assumptions underlying regression analysis.

Stage 6: Validation of the Results

- Additional or Split Samples
- Comparing Regression Models
- Forecasting with the Model

Some considerations needed when applying the regression analysis (1)

Range of model: when we apply a regression model outside the range of observations forecasting etc.), the linear model might become inappropriate.
Considerable judgment is needed, when we doing this, a linear model might become a curvilinear regression when we extended the range too far.

Some considerations needed when applying the regression analysis (2)

Continuation of causal conditions: When the causal conditions have changed, the fitted model may no longer be appropriate. For example, the entry of a major new competitor may invalidate the past relation between a company's sales (Y) and the amount of advertising (X).

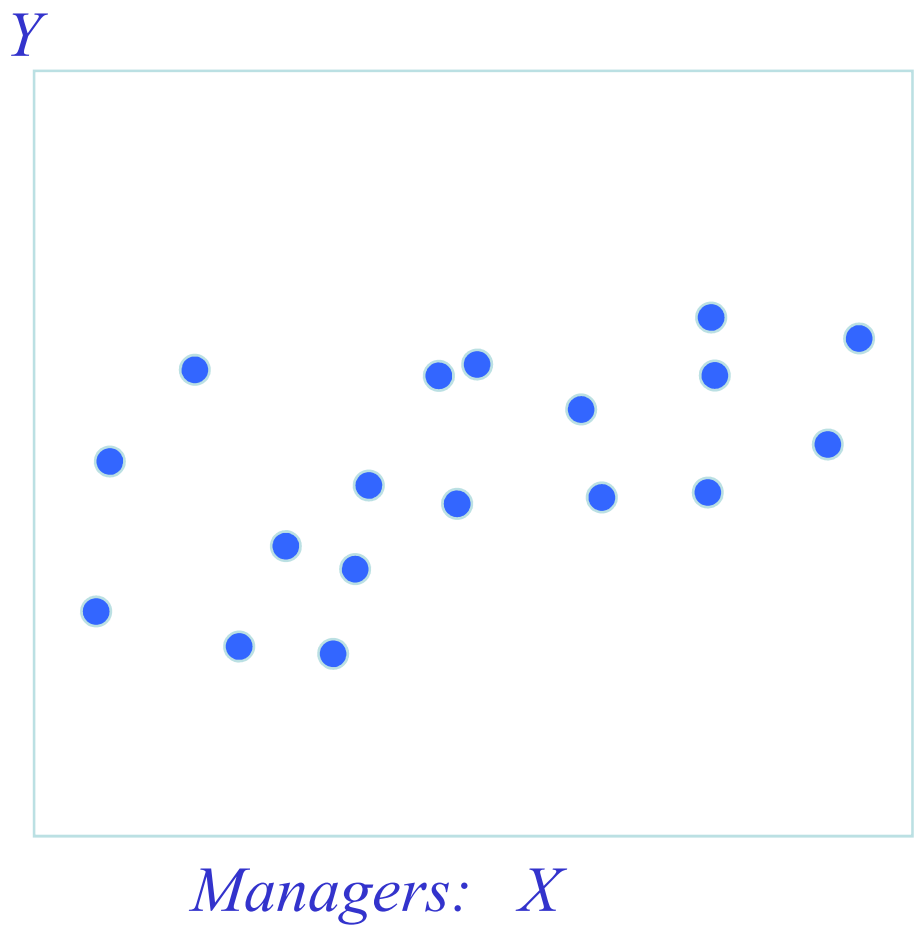
Some considerations needed when applying the regression analysis (3)

Causality: Presence of a regression relation does not implies a cause-and-effect. For example, reading ability (Y) was regressed on the shoe size (X) for a sample of an elementary school students and a positive regression was observed: what can we conclude???

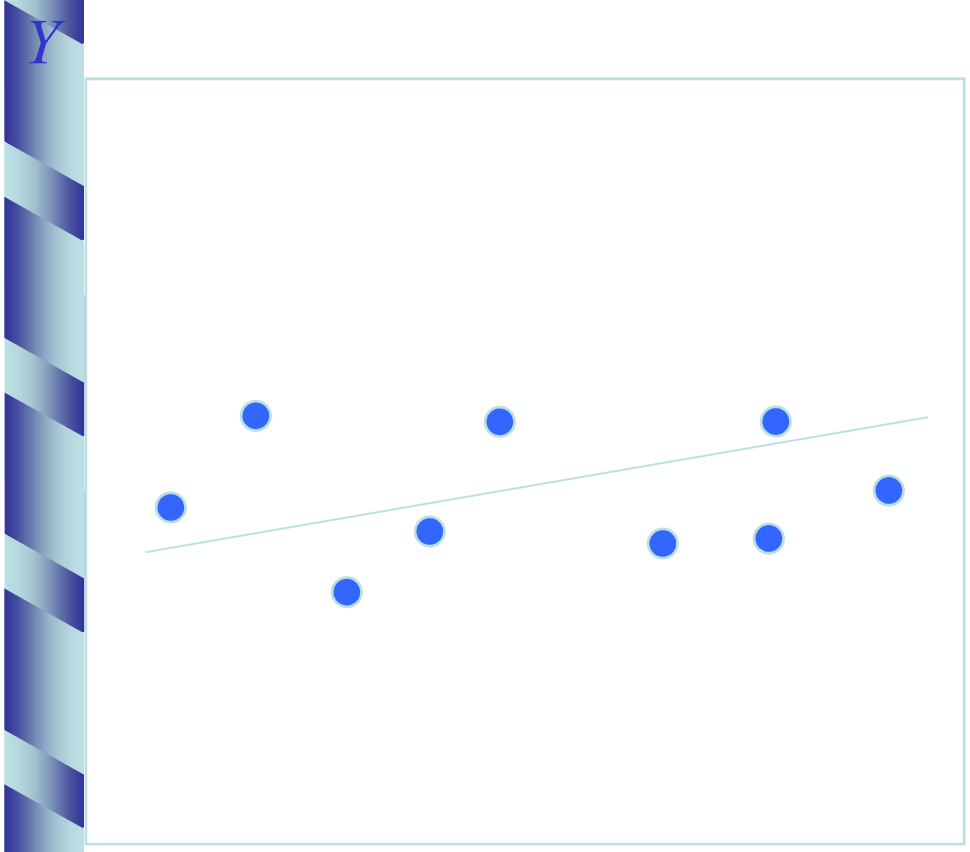
Some considerations needed when applying the regression analysis (4)

The omission of an intervening variable can sometimes also hide a relation between two variables. For example, a scatter plot of years of education (X) and salary (Y) of middle managers shows no regression relation, see next page:

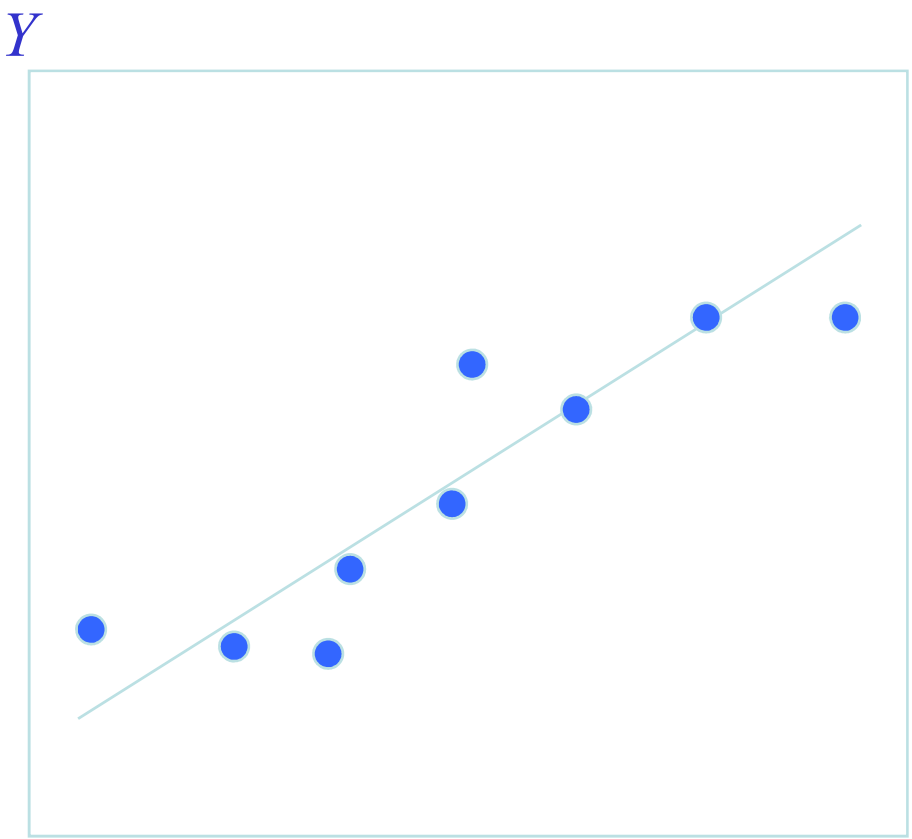
An example



Example - continue: hidden relations



Sales Manager: X

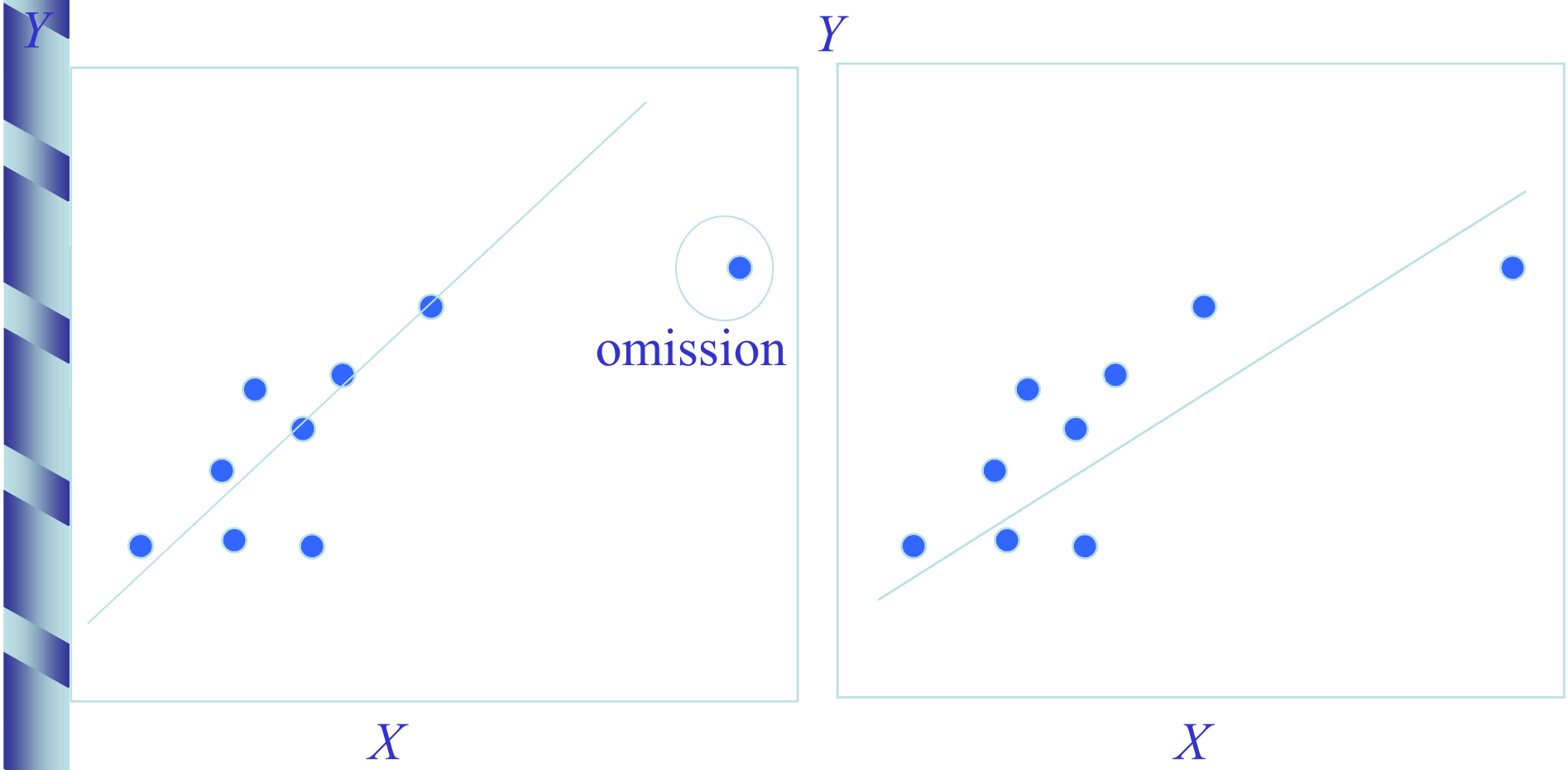


Other: X

Some considerations needed when applying the regression analysis (5)

Outlying observation: The outlying observations may have a substantial influence on the fitted regression function, where the inferences drawn from the regression study can be drastically differs on whether the outlying observation is included or not. Some preprocess of the data set is needed here.

An example of an influential observation



End of The Section