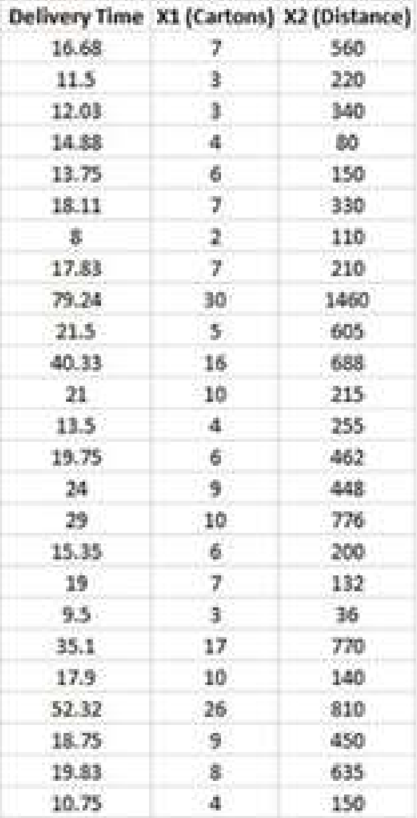
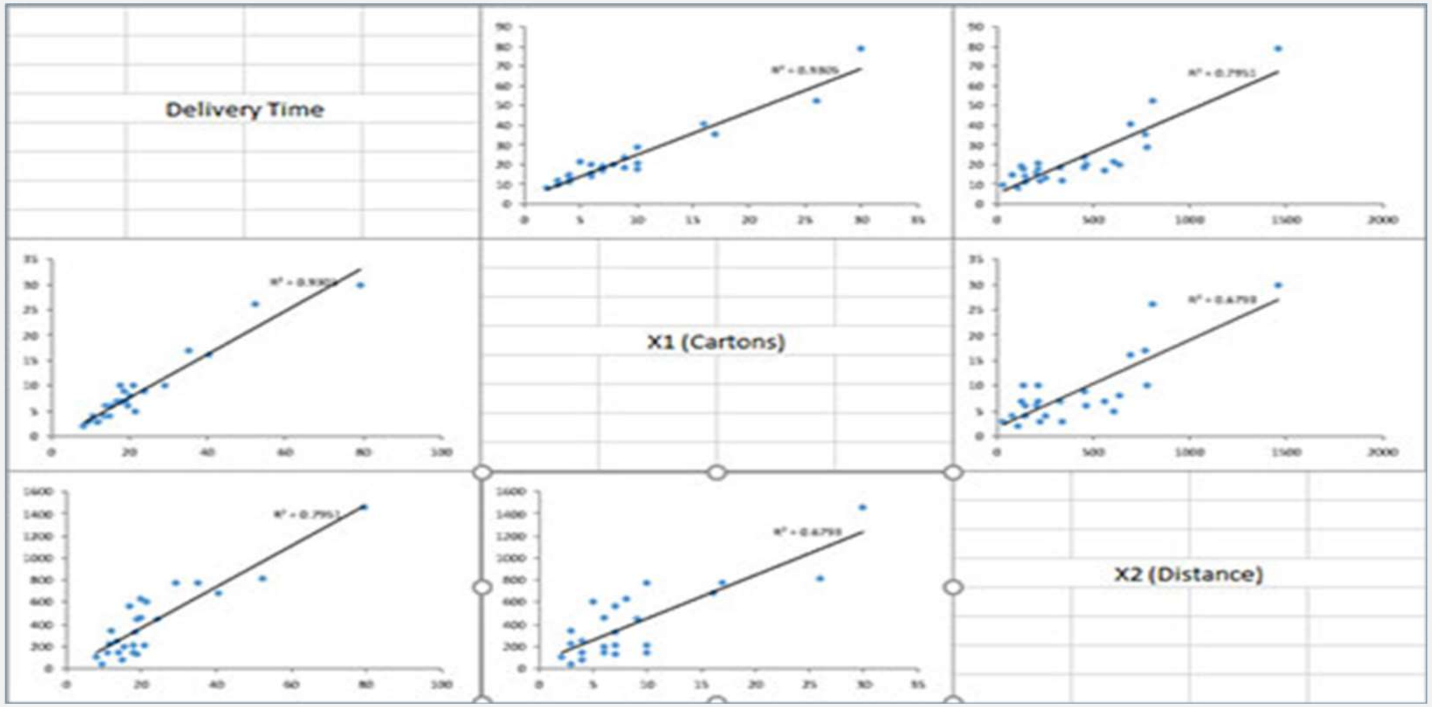
Assumption 1: Linearity

Example: A soft drink bottler is trying to predict delivery times for a driver. He has collected data on the delivery time, the number of cartons delivered and the distance the driver walked. He wants to see the relationship between these three variables. We will use the scatter plot matrix to do this.



Delivery Time	X1 (Cartons)	X2 (Distance)
16.68	7	560
11.5	3	220
12.03	3	340
14.88	4	80
13.75	6	150
18.11	7	330
8	2	110
17.83	7	210
79.24	30	1460
21.5	5	605
40.33	16	688
21	10	215
13.5	4	255
19.75	6	462
24	9	448
29	10	776
15.35	6	200
19	7	132
9.5	3	36
35.1	17	770
17.9	10	140
52.32	26	810
18.75	9	450
19.83	8	635
10.75	4	150



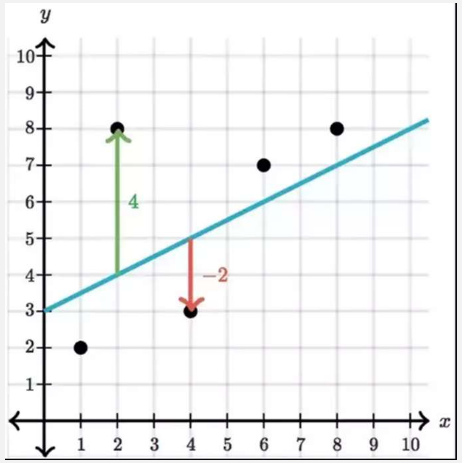
Residuals

- The residual, or error, of the regression model is the difference between the y value and the predicted value. Each data point has one residual.

Residual = Observed value - Predicted value

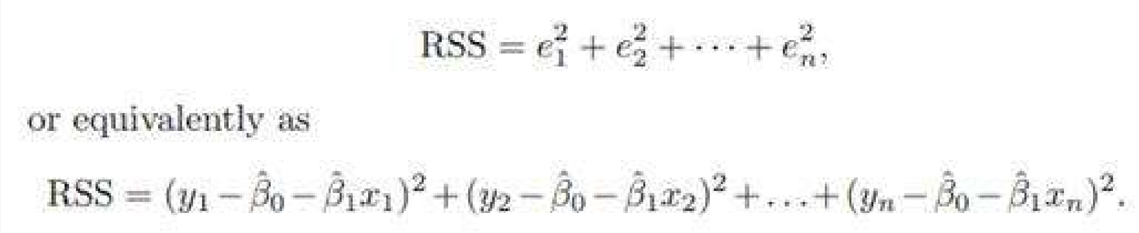
$$\text{Residual} = y - \hat{y}$$

- Because a linear regression model is not always appropriate for the data, you should assess the appropriateness of the model by defining residuals and examining residual plots.
- Both the sum and the mean of the residuals are equal to zero.



Residuals

- We define the residual sum of squares (RSS) as

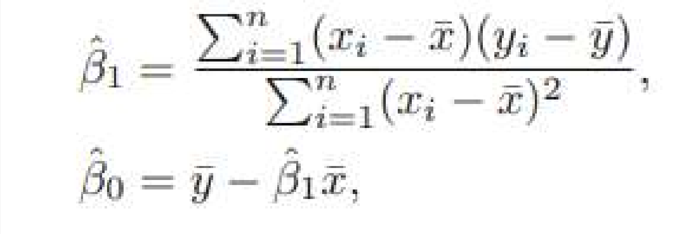


$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\mathrm{RSS} = (y_1 - \hat\beta_0 - \hat\beta_1 x_1)^2 + (y_2 - \hat\beta_0 - \hat\beta_1 x_2)^2 + \ldots + (y_n - \hat\beta_0 - \hat\beta_1 x_n)^2.$$

- The least squares approach chooses βˆ0 and βˆ1 to minimize the RSS.



$$\hat\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar x)(y_i - \bar y)}{\sum_{i=1}^n (x_i - \bar x)^2},$$

$$\hat\beta_0 = \bar y - \hat\beta_1 \bar x,$$

Residuals

Using Residuals to Test the Assumptions of the Regression Model

- One of the major uses of residual analysis is to test some of the assumptions underlying regression. The following are the assumptions of simple regression analysis.

  1. The model is linear.(Assumption 01)

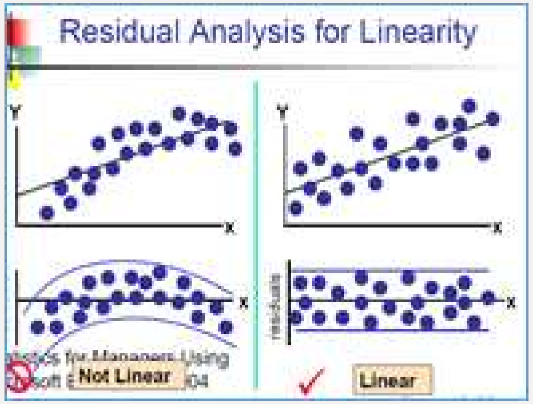
  2. The error terms have constant variances.(Assumption 03)

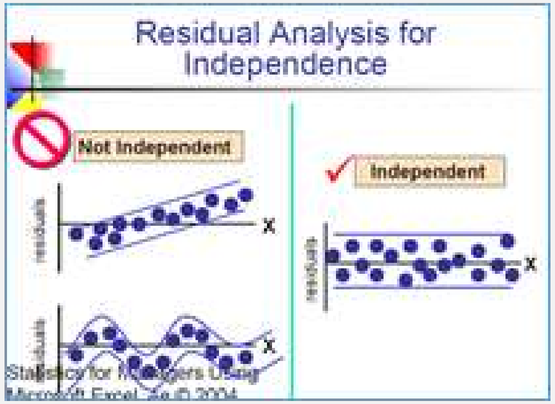
  3. The error terms are independent.

  4. The error terms are normally distributed.(Assumption 06)

Residuals Plot

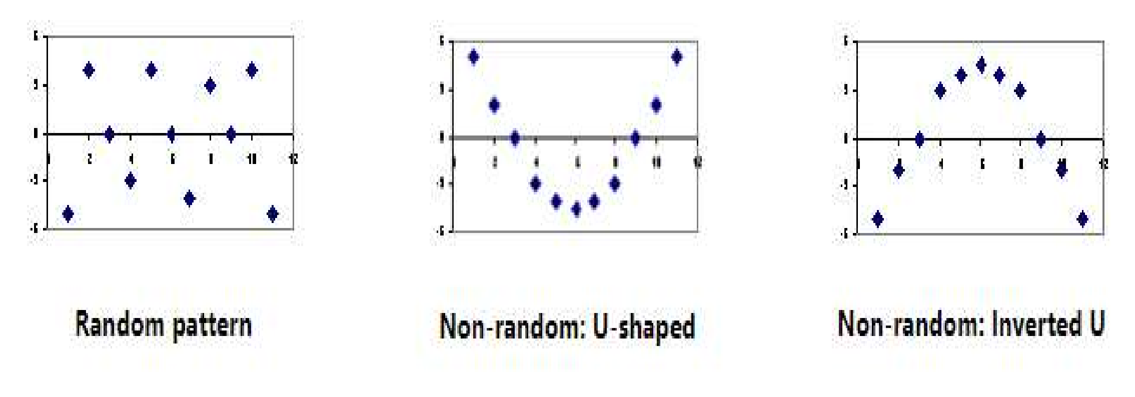
- A particular method for studying the behavior of residuals is the residual plot.
- The residual plot is a type of graph in which the residuals for a particular regression model are plotted along with their associated value of x as an ordered pair(x, Residual).
- A residual plot is a graph in which residuals are on the vertical axis and the independent variable is on the horizontal axis.
- If the dots are randomly dispersed around the horizontal axis then a linear regression model is appropriate for the data; otherwise, choose a non-linear model.





Residuals Plot

Following example shows few patterns in residual plots.



Random pattern

Non-random: U-shaped

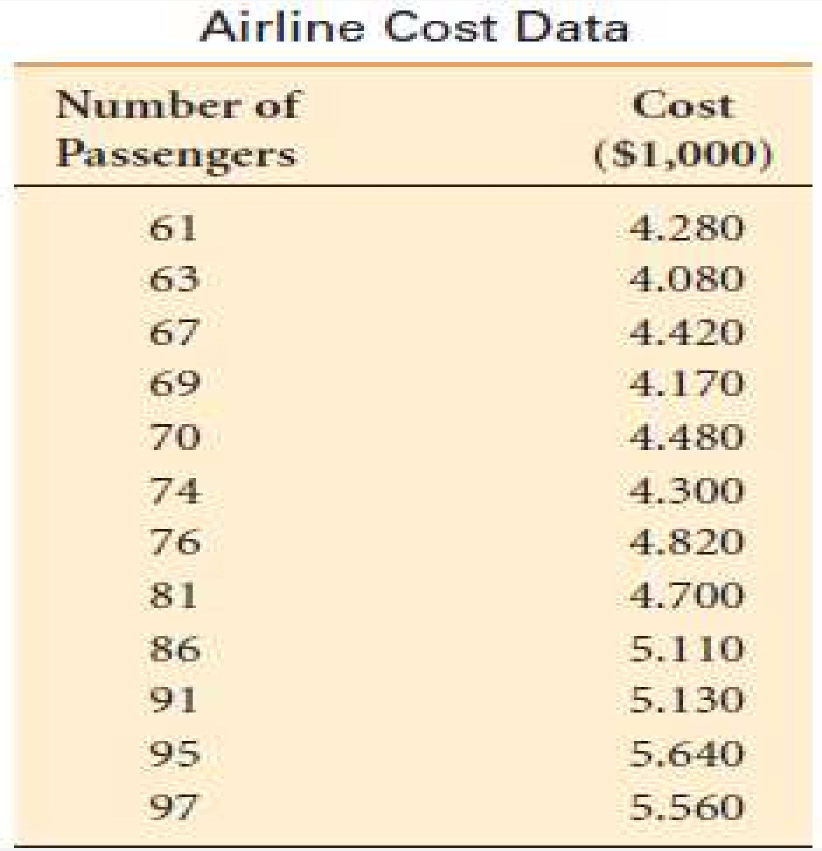
Non-random: Inverted U

- In first case, dots are randomly dispersed. So linear regression model is preferred.

- In Second and third case, dots are non-randomly dispersed and suggests that a non-linear regression method is preferred

Problem 1:

Suppose a study is conducted using only Boeing 737s traveling 500 miles on comparable routes during the same season of the year. Can the number of passengers predict the cost of flying such routes? Suppose the data displayed in Table are the costs and associated number of passengers for twelve 500-mile commercial airline flights using Boeing 737s during the same season of the year. Use these data to develop a regression model to predict cost by number of passengers. Analyze the residuals linearity by using graphic diagnostics.



Airline Cost Data

Number of Passengers	Cost ($1,000)
61	4.280
63	4.080
67	4.420
69	4.170
70	4.480
74	4.300
76	4.820
81	4.700
86	5.110
91	5.130
95	5.640
97	5.560

Problem 1 Solution:

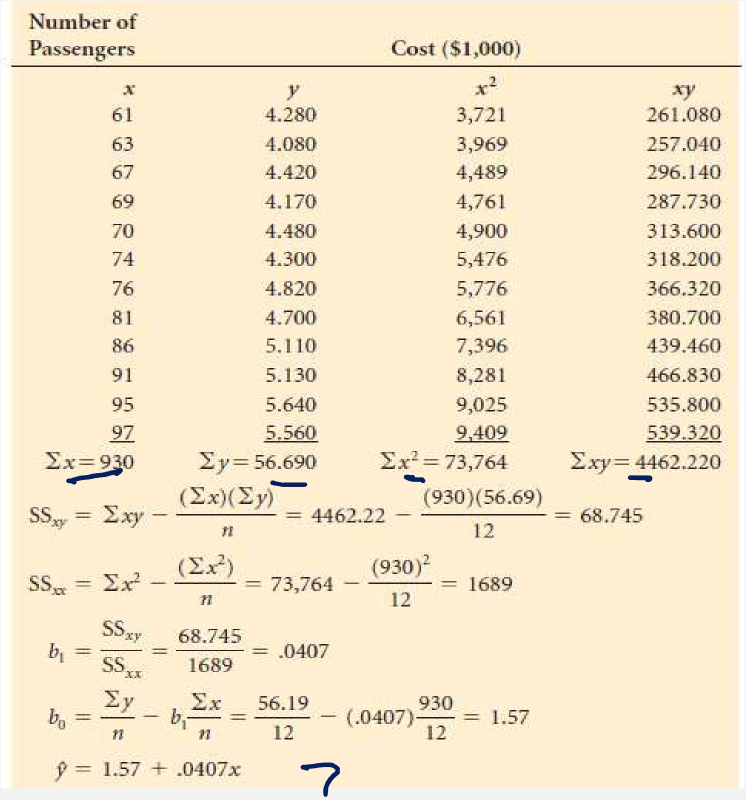
N.B

$\Sigma y = Na + b\Sigma x \longrightarrow (i)$

$\Sigma xy = a\Sigma x + \Sigma x^2 \longrightarrow (ii)$

Assumption 01: Linearity

The first step in simple regression analysis is to construct a scatter plot. The scatter plot gives some idea of how well a regression line fits the data.



Number of Passengers	Cost ($1,000)		
x	y	x^2	xy
61	4.280	3,721	261.080
63	4.080	3,969	257.040
67	4.420	4,489	296.140
69	4.170	4,761	287.730
70	4.480	4,900	313.600
74	4.300	5,476	318.200
76	4.820	5,776	366.320
81	4.700	6,561	380.700
86	5.110	7,396	439.460
91	5.130	8,281	466.830
95	5.640	9,025	535.800
97	5.560	9,409	539.320
$\Sigma x = 930$	$\Sigma y = 56.690$	$\Sigma x^2 = 73{,}764$	$\Sigma xy = 4462.220$

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 4462.22 - \frac{(930)(56.69)}{12} = 68.745$$

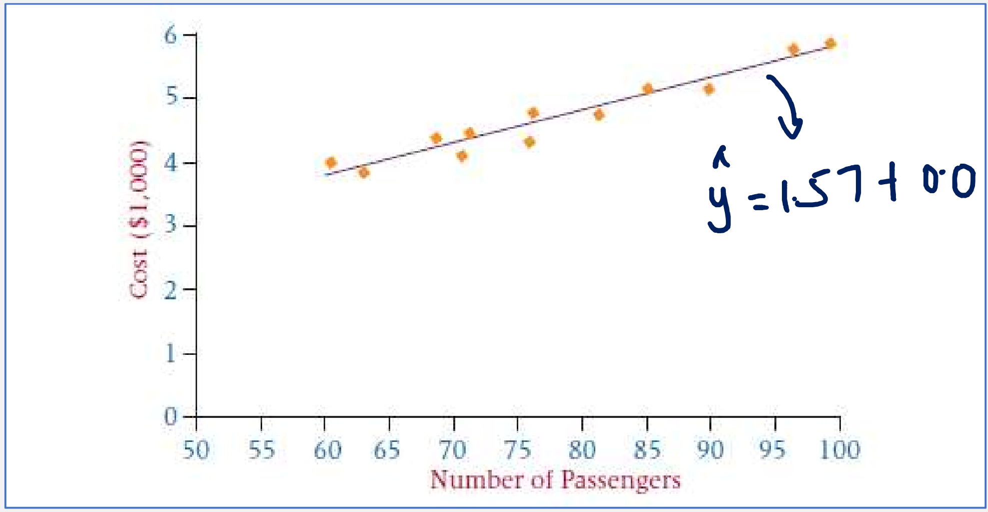
$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x^2)}{n} = 73{,}764 - \frac{(930)^2}{12} = 1689$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{68.745}{1689} = .0407$$

$$b_0 = \frac{\Sigma y}{n} - b_1\frac{\Sigma x}{n} = \frac{56.19}{12} - (.0407)\frac{930}{12} = 1.57$$

$$\hat{y} = 1.57 + .0407x$$

Superimposing the line representing the least squares equation for this problem on the scatter plot indicates how well the regression line fits the data points.

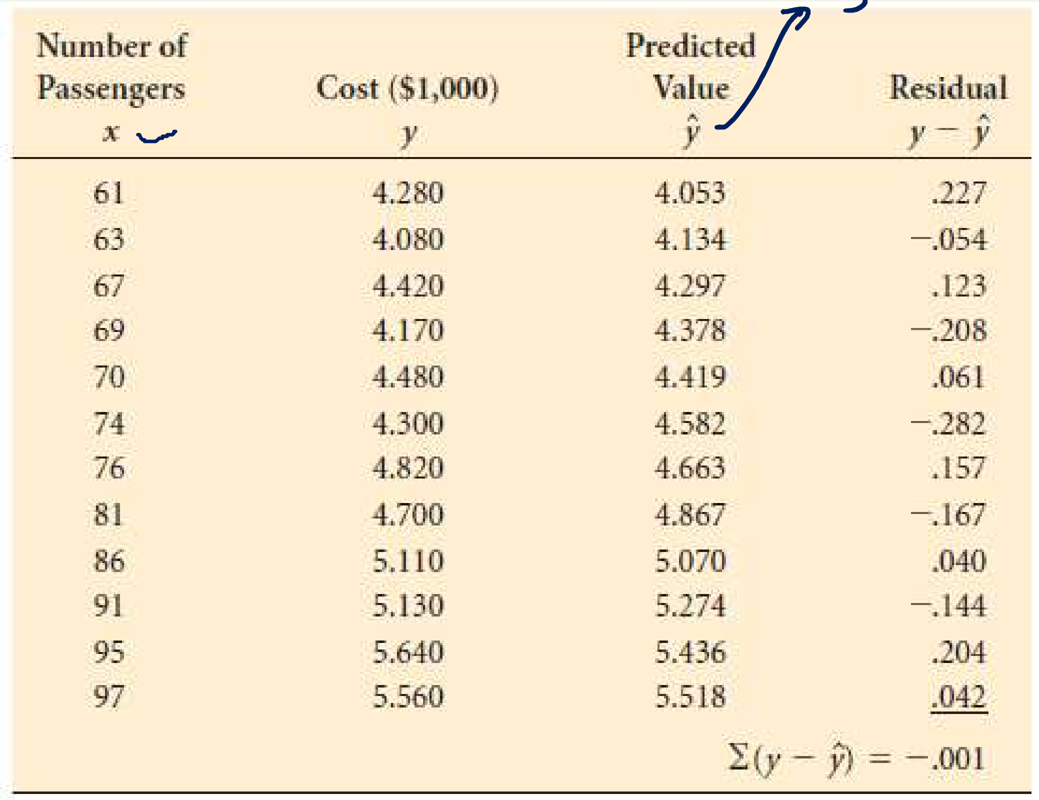


$\hat{y} = 1.57 + 0.0407x$

Problem 1 Solution:

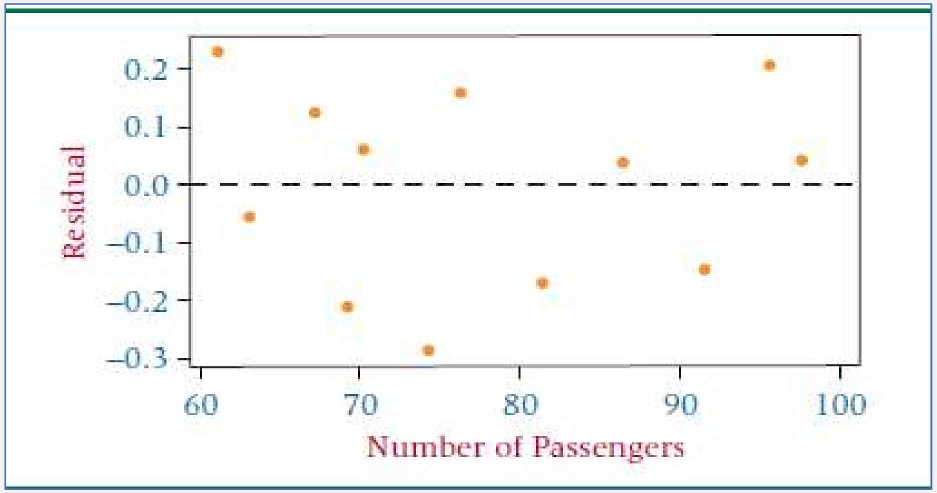
Assumption 02 Residual Analysis

$\hat{y} = 1.57 + 0.0407x$



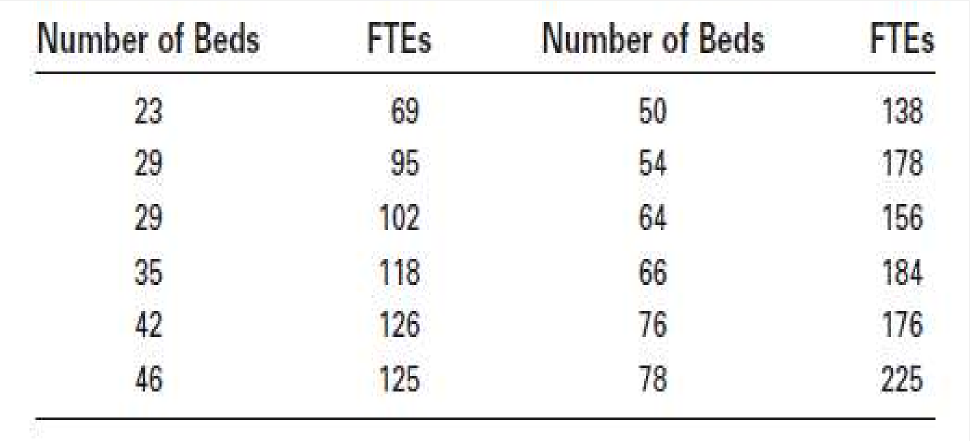
Number of Passengers x	Cost ($1,000) y	Predicted Value $\hat{y}$	Residual $y - \hat{y}$
61	4.280	4.053	.227
63	4.080	4.134	−.054
67	4.420	4.297	.123
69	4.170	4.378	−.208
70	4.480	4.419	.061
74	4.300	4.582	−.282
76	4.820	4.663	.157
81	4.700	4.867	−.167
86	5.110	5.070	.040
91	5.130	5.274	−.144
95	5.640	5.436	.204
97	5.560	5.518	.042
			$\Sigma(y - \hat{y}) = -.001$

Note that the sum of the residuals is approximately zero. Except for rounding error, the sum of the residuals is *always zero.* Residuals are usually plotted against the *x*-axis, which reveals a view of the residuals as *x* increases.



Problem 2:

A specialist in hospital administration stated that the number of FTEs (full-time employees) in a hospital can be estimated by counting the number of beds in the hospital (a common measure of hospital size). A healthcare business researcher decided to develop a regression model in an attempt to predict the number of FTEs of a hospital by the number of beds. She surveyed 12 hospitals and obtained the following data. The data are presented in sequence, according to the number of beds:



Number of Beds	FTEs	Number of Beds	FTEs
23	69	50	138
29	95	54	178
29	102	64	156
35	118	66	184
42	126	76	176
46	125	78	225

Compute the residuals for Demonstration Problem in which a regression model was developed to predict the number of full-time equivalent workers (FTEs) by the number of beds in a hospital. Analyze the residuals by using graphic diagnostics.