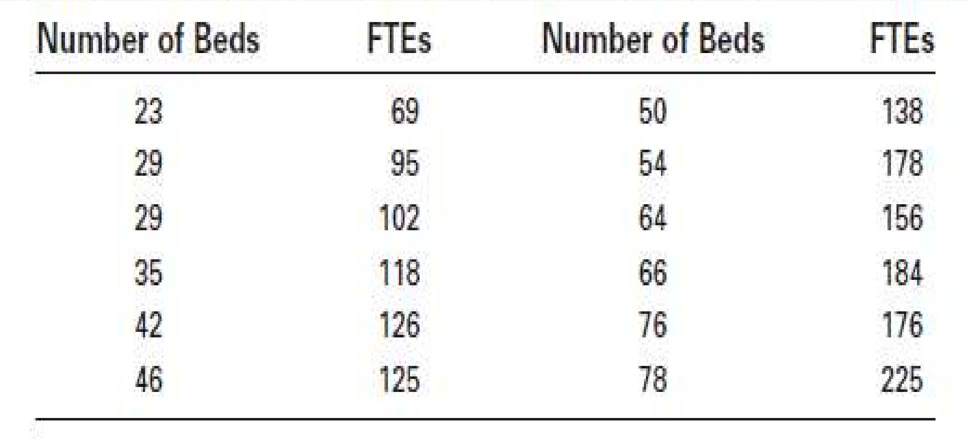
Problem 2:

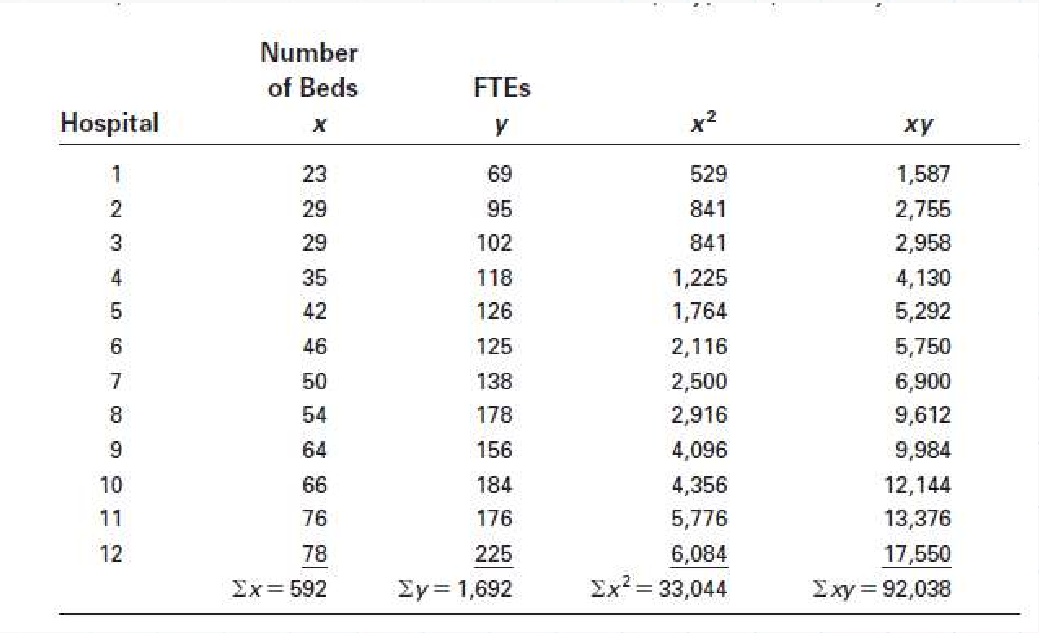
A specialist in hospital administration stated that the number of FTEs (full-time employees) in a hospital can be estimated by counting the number of beds in the hospital (a common measure of hospital size). A healthcare business researcher decided to develop a regression model in an attempt to predict 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:



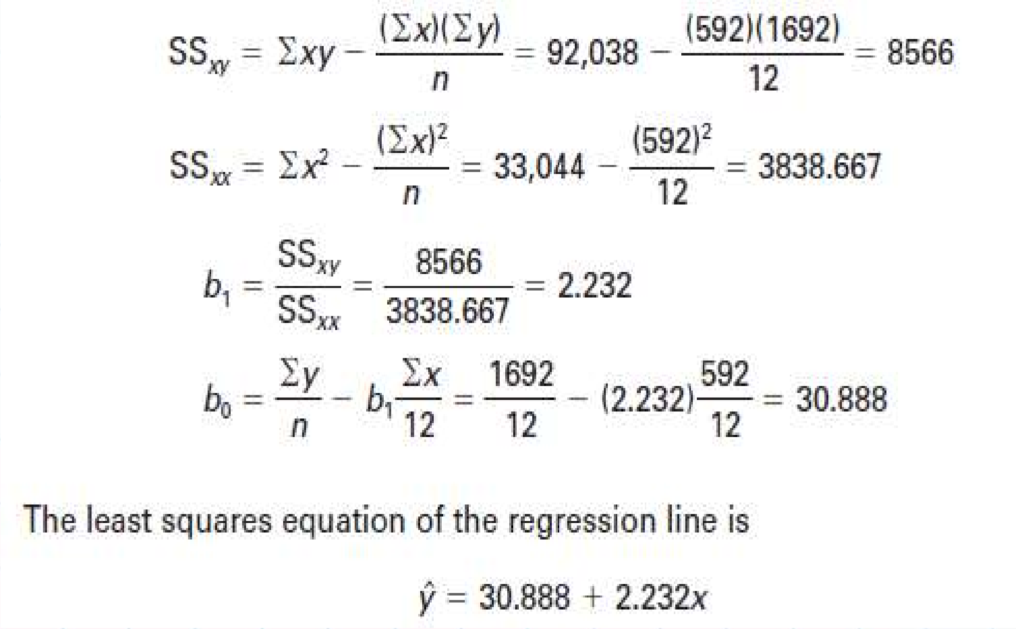
Number of Beds	FTEs	Number of Beds	FTEs
23	69	50	138
29	95	54	178
29	102	64	156
35	118	66	184
42	126	76	176
46	125	78	225

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.

Problem 2 Solutions:



Hospital	Number of Beds x	FTEs y	x^2	xy
1	23	69	529	1,587
2	29	95	841	2,755
3	29	102	841	2,958
4	35	118	1,225	4,130
5	42	126	1,764	5,292
6	46	125	2,116	5,750
7	50	138	2,500	6,900
8	54	178	2,916	9,612
9	64	156	4,096	9,984
10	66	184	4,356	12,144
11	76	176	5,776	13,376
12	78	225	6,084	17,550
	$\Sigma x = 592$	$\Sigma y = 1{,}692$	$\Sigma x^2 = 33{,}044$	$\Sigma xy = 92{,}038$



$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 92{,}038 - \frac{(592)(1692)}{12} = 8566$$

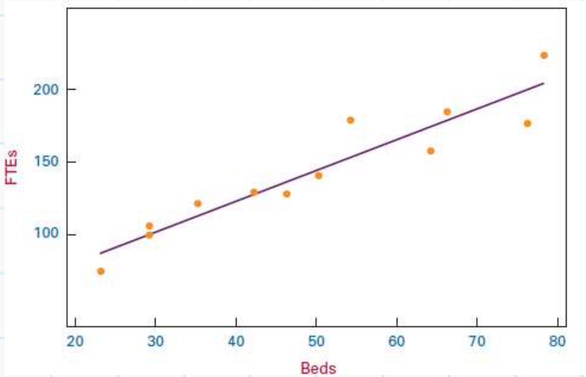
$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 33{,}044 - \frac{(592)^2}{12} = 3838.667$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{8566}{3838.667} = 2.232$$

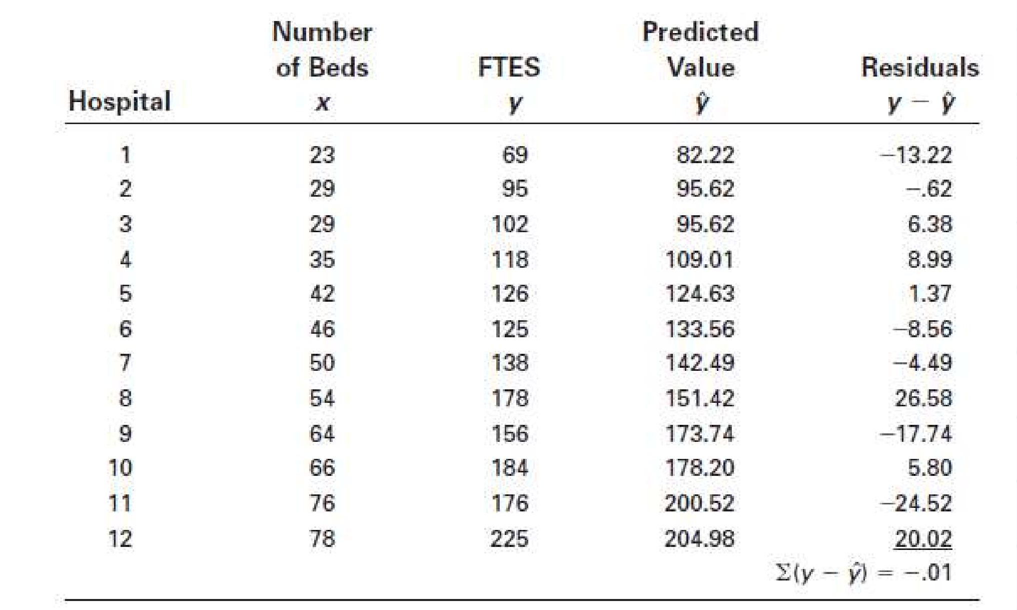
$$b_0 = \frac{\Sigma y}{n} - b_1\frac{\Sigma x}{12} = \frac{1692}{12} - (2.232)\frac{592}{12} = 30.888$$

The least squares equation of the regression line is

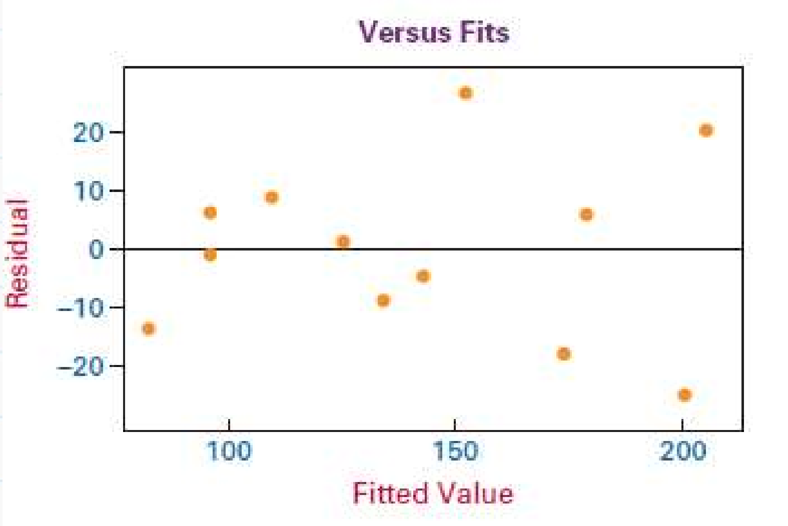
$$\hat{y} = 30.888 + 2.232x$$



Problem 2 Solutions:



Hospital	Number of Beds x	FTES y	Predicted Value $\hat{y}$	Residuals $y - \hat{y}$
1	23	69	82.22	−13.22
2	29	95	95.62	−.62
3	29	102	95.62	6.38
4	35	118	109.01	8.99
5	42	126	124.63	1.37
6	46	125	133.56	−8.56
7	50	138	142.49	−4.49
8	54	178	151.42	26.58
9	64	156	173.74	−17.74
10	66	184	178.20	5.80
11	76	176	200.52	−24.52
12	78	225	204.98	20.02
				$\Sigma(y - \hat{y}) = -.01$



Note that the regression model fits these particular data well for hospitals 2 and 5, as indicated by residuals of -.62 and 1.37 FTEs, respectively. For hospitals 1, 8, 9, 11, and 12, the residuals are relatively large, indicating that the regression model does not fit the data for these hospitals well.

SSE and Standard Error of the Estimate

- Residuals represent errors of estimation for individual points.
- With large samples of data, residual computations become laborious.
- Even with computers, a researcher sometimes has difficulty working through pages of residuals in an effort to understand the error of the regression model.
- An alternative way of examining the error of the model is the standard error of the estimate, which provides a single measurement of the regression error.
- Because the sum of the residuals is zero, attempting to determine the total amount of error by summing the residuals is fruitless.

- The total of the residuals squared column is called the sum of squares of error (SSE).

$$SSE = \sum (y - \hat{y})^2$$

- Computational formula for SSE:

$$SSE = \sum y^2 - b_0 \sum y - b_1 \sum xy$$

SSE and Standard Error of the Estimate (Accessing the accuracy)

In theory, infinitely many lines can be fit to a sample of points. Line of best fit is for which the SSE is the smallest. For this reason, the regression process is called least squares regression. A value closer to 0 indicates that the model has a smaller random error component, and that the fit will be more useful for prediction.

The standard error of the estimate

- A more useful measurement of error is the standard error of the estimate.
- The standard error of the estimate, denoted se, is a standard deviation of the error of the regression model and has more practical use than SSE.
- An assumption underlying regression analysis is that the error terms are approximately normally distributed with a mean of zero. With this information and by the empirical rule, approximately 68% of the residuals should be within 1se and 95% should be within 2se. This property makes the standard error of the estimate a useful tool in estimating how accurately a regression model is fitting the data.
- The standard error of the estimate is a standard deviation of error. The standard error of the regression provides the absolute measure of the typical distance that the data points fall from the regression line.

SSE and Standard Error of the Estimate

The standard error of the estimate is computed by dividing SSE by the degrees of freedom of error for the model and taking the square root. The standard error of the estimate follows.

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

where

n = number of observations

k = number of independent variables

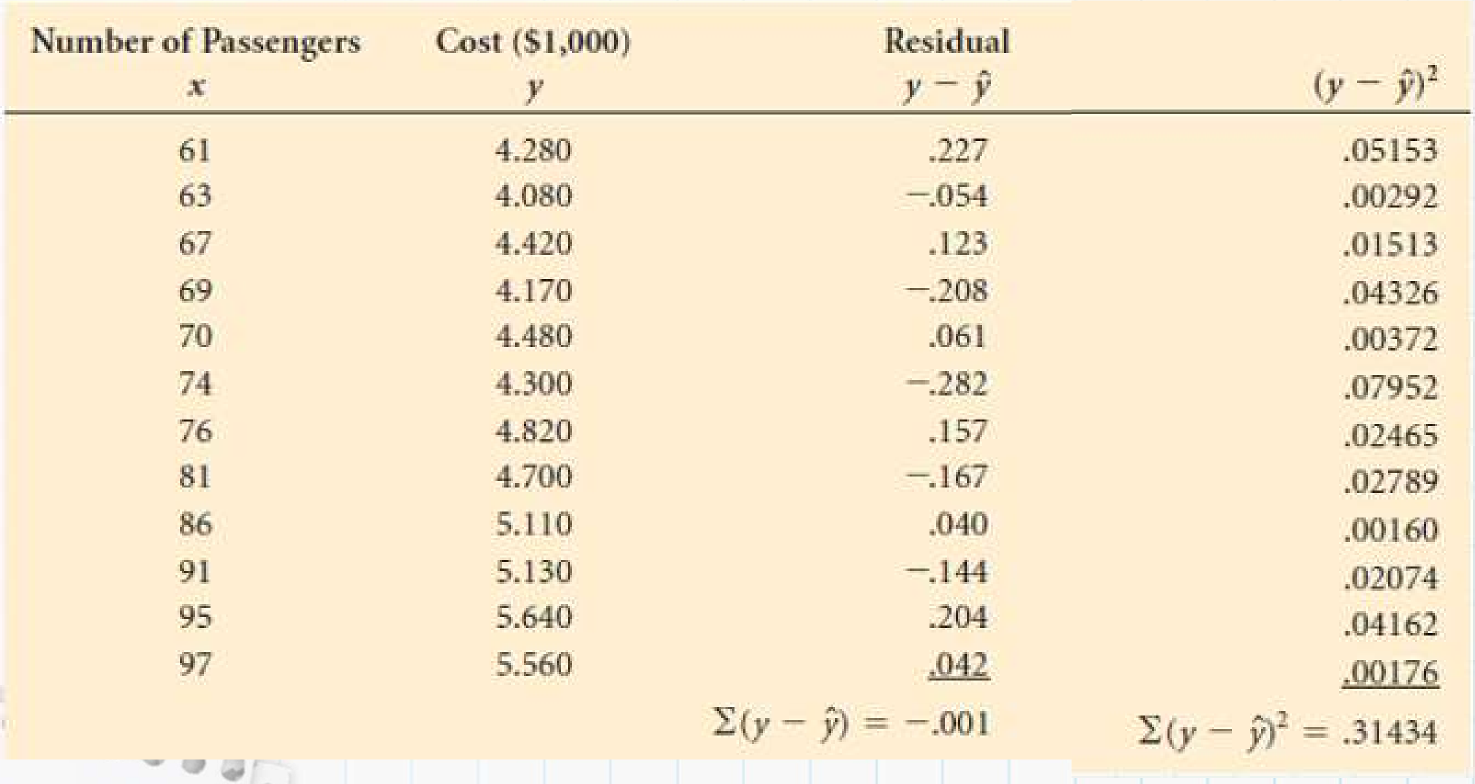
SSE/(n-k-1)= mean squared errors or (MSE).

n – k – 1 =degrees of freedom

SSE and Standard Error of the Estimate

Problem 1: (Slide 37)

Determining SSE and Standard Error of the Estimate for the Airline Cost Example 01

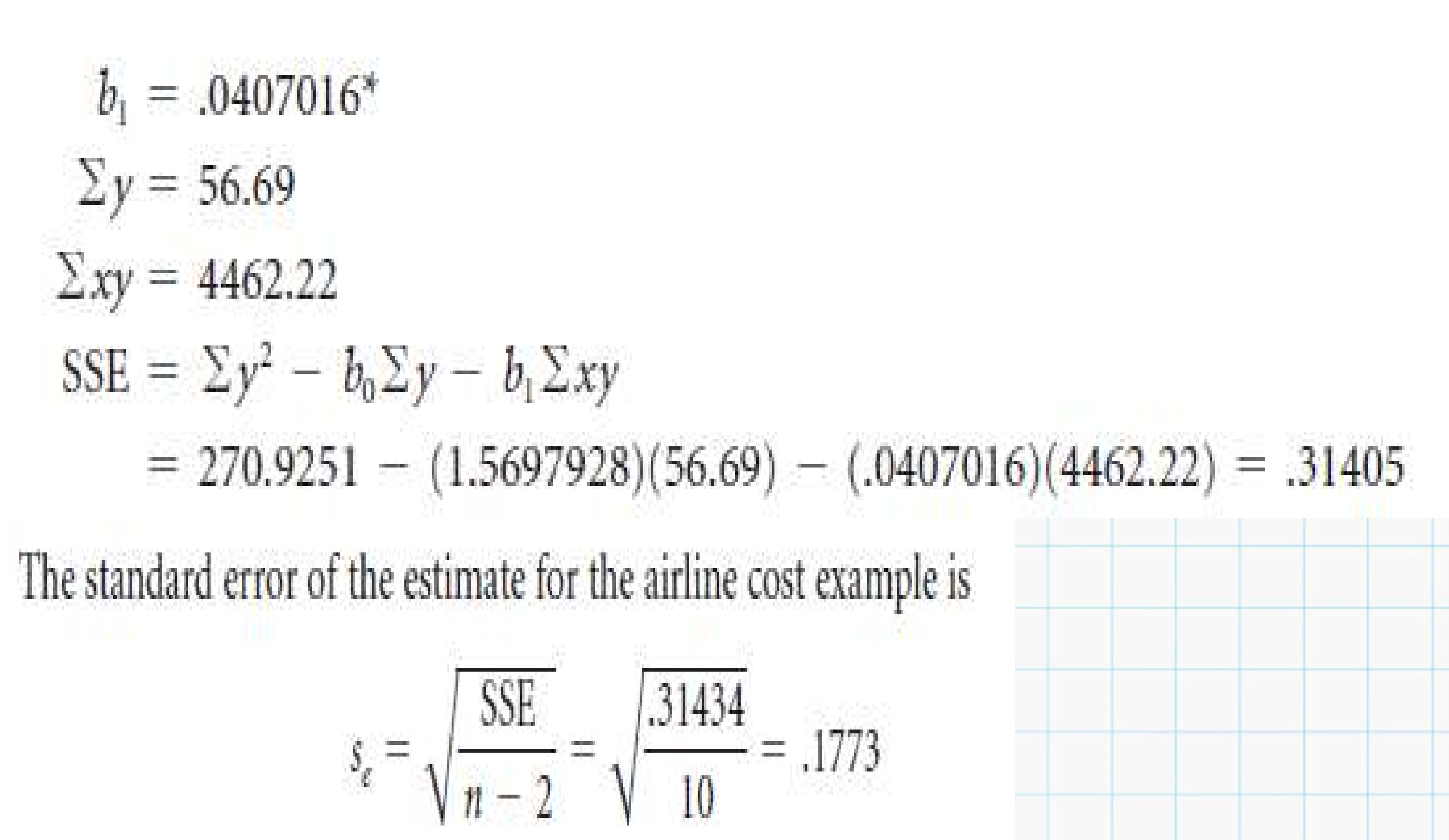


Number of Passengers x	Cost ($1,000) y	Residual $y-\hat{y}$	$(y-\hat{y})^2$
61	4.280	.227	.05153
63	4.080	−.054	.00292
67	4.420	.123	.01513
69	4.170	−.208	.04326
70	4.480	.061	.00372
74	4.300	−.282	.07952
76	4.820	.157	.02465
81	4.700	−.167	.02789
86	5.110	.040	.00160
91	5.130	−.144	.02074
95	5.640	.204	.04162
97	5.560	.042	.00176
		$\Sigma(y-\hat{y}) = -.001$	$\Sigma(y-\hat{y})^2 = .31434$

SSE and Standard Error of the Estimate

Problem 1: Another Method

Determining SSE and Standard Error of the Estimate for the Airline Cost Example 01



$$b_1 = .0407016^*$$

$$\Sigma y = 56.69$$

$$\Sigma xy = 4462.22$$

$$\begin{aligned} \text{SSE} &= \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy \\ &= 270.9251 - (1.5697928)(56.69) - (.0407016)(4462.22) = .31405 \end{aligned}$$

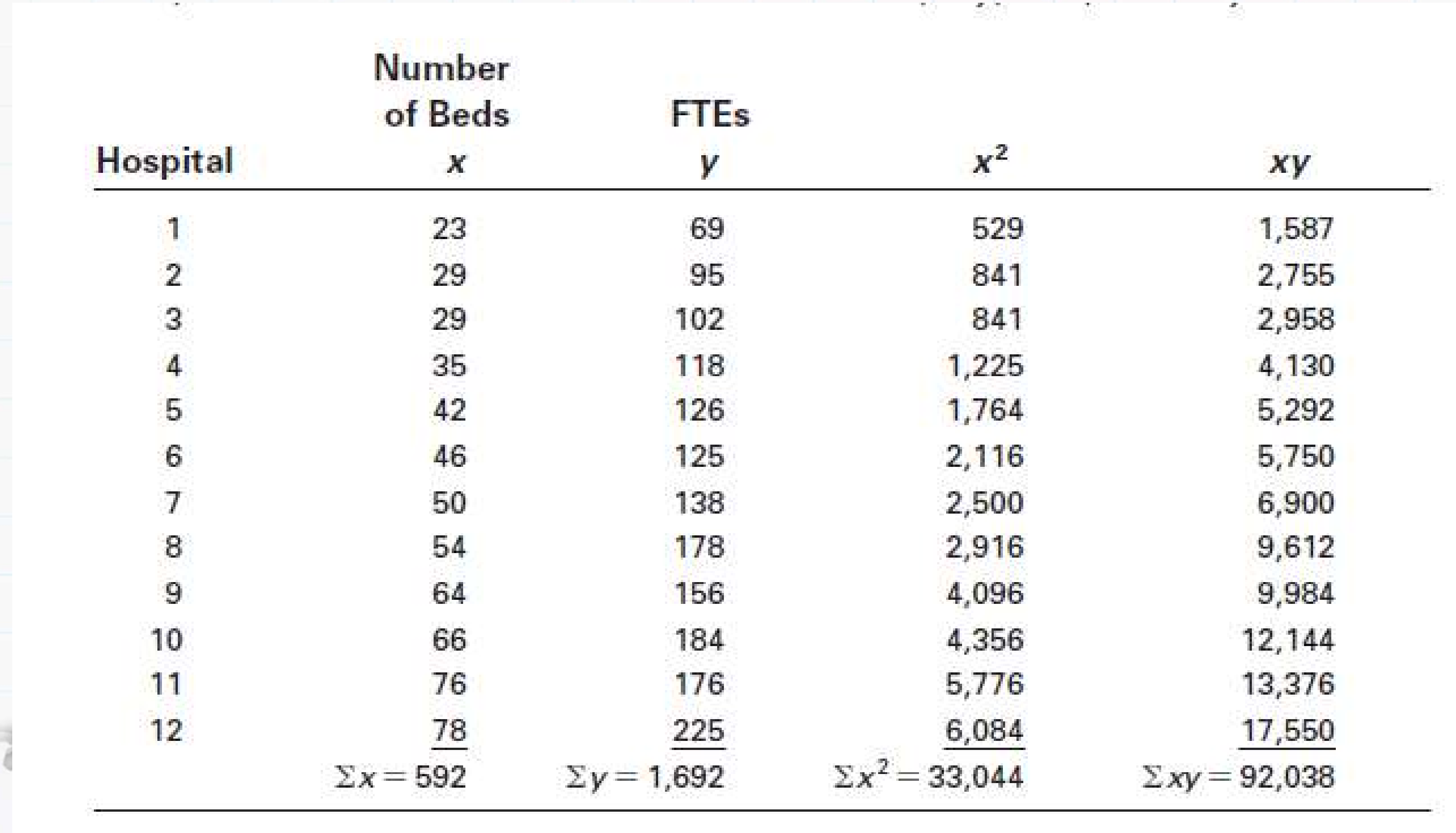
The standard error of the estimate for the airline cost example is

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{.31434}{10}} = .1773$$

SSE and Standard Error of the Estimate

Problem 2: (Slide No. 40)

Determining SSE and Standard Error of the Estimate for the Hospital Example 02

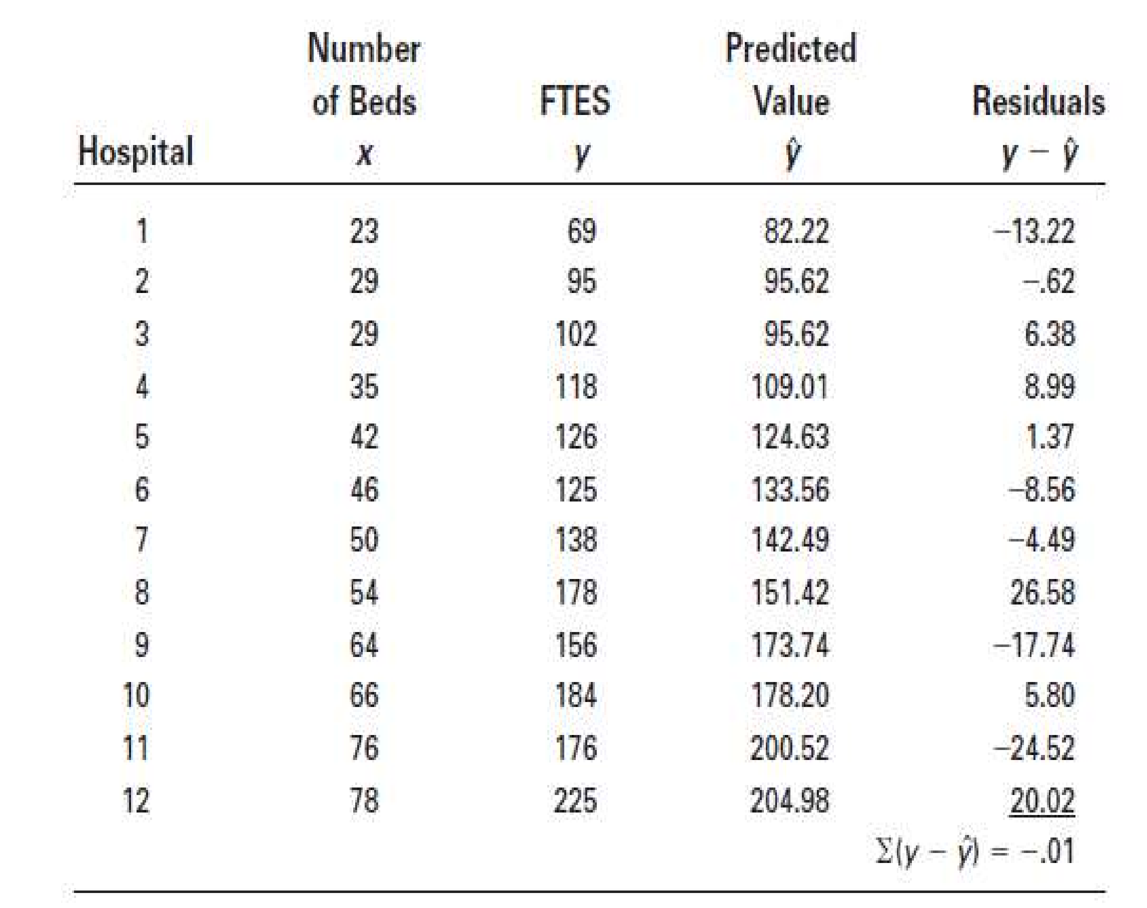


Hospital	Number of Beds x	FTEs y	x^2	xy
1	23	69	529	1,587
2	29	95	841	2,755
3	29	102	841	2,958
4	35	118	1,225	4,130
5	42	126	1,764	5,292
6	46	125	2,116	5,750
7	50	138	2,500	6,900
8	54	178	2,916	9,612
9	64	156	4,096	9,984
10	66	184	4,356	12,144
11	76	176	5,776	13,376
12	78	225	6,084	17,550
	$\Sigma x = 592$	$\Sigma y = 1{,}692$	$\Sigma x^2 = 33{,}044$	$\Sigma xy = 92{,}038$

SSE and Standard Error of the Estimate

Problem 2: (Slide No. 40)

Determining SSE and Standard Error of the Estimate for the Hospital Example 02

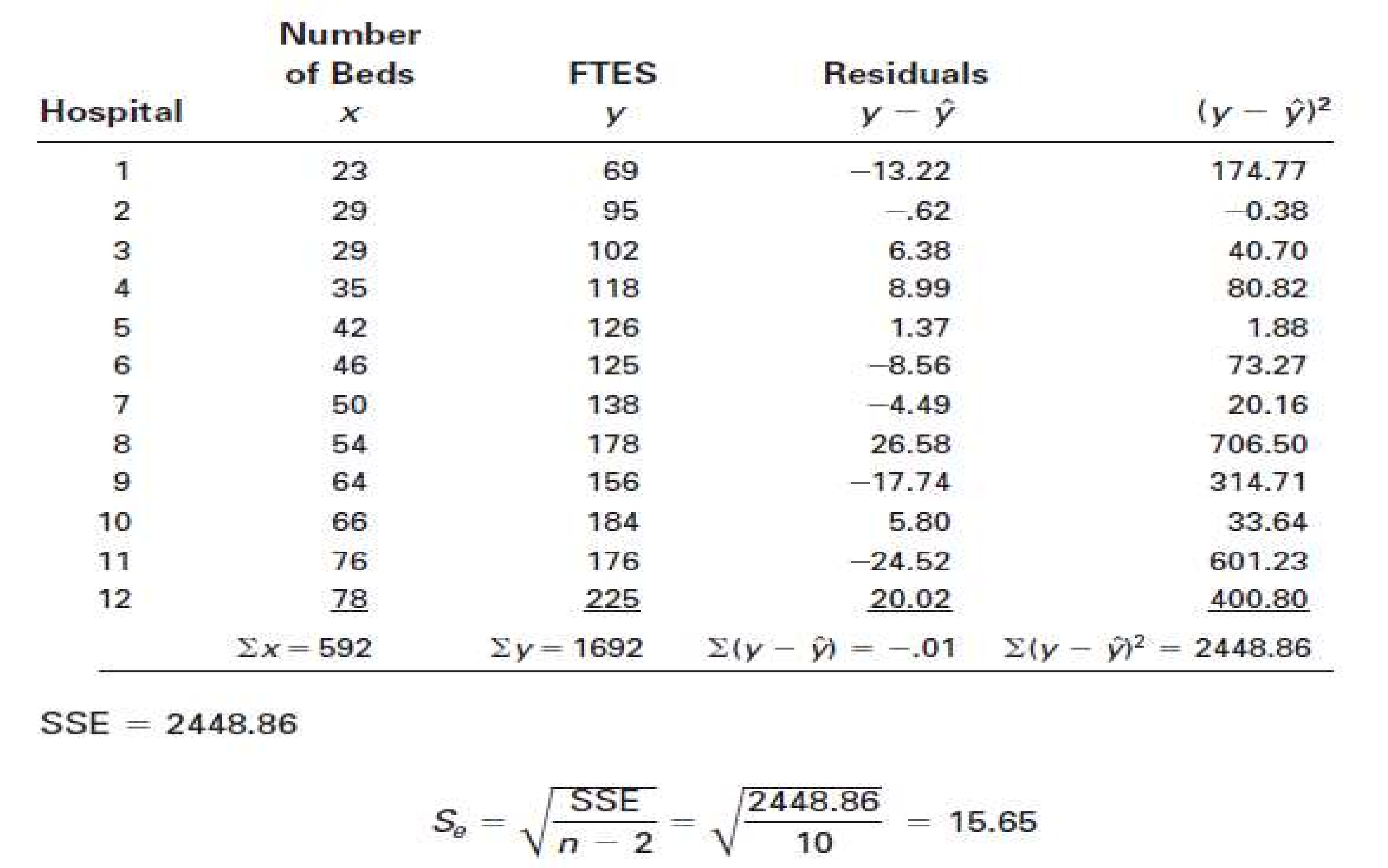


Hospital	Number of Beds x	FTES y	Predicted Value $\hat{y}$	Residuals $y - \hat{y}$
1	23	69	82.22	−13.22
2	29	95	95.62	−.62
3	29	102	95.62	6.38
4	35	118	109.01	8.99
5	42	126	124.63	1.37
6	46	125	133.56	−8.56
7	50	138	142.49	−4.49
8	54	178	151.42	26.58
9	64	156	173.74	−17.74
10	66	184	178.20	5.80
11	76	176	200.52	−24.52
12	78	225	204.98	20.02
				$\Sigma(y - \hat{y}) = -.01$

SSE and Standard Error of the Estimate

Problem 2: (Slide No. 40)

Determining SSE and Standard Error of the Estimate for the Hospital Example 02



Hospital	Number of Beds x	FTES y	Residuals $y - \hat{y}$	$(y - \hat{y})^2$
1	23	69	−13.22	174.77
2	29	95	−.62	−0.38
3	29	102	6.38	40.70
4	35	118	8.99	80.82
5	42	126	1.37	1.88
6	46	125	−8.56	73.27
7	50	138	−4.49	20.16
8	54	178	26.58	706.50
9	64	156	−17.74	314.71
10	66	184	5.80	33.64
11	76	176	−24.52	601.23
12	78	225	20.02	400.80
	$\Sigma x = 592$	$\Sigma y = 1692$	$\Sigma(y - \hat{y}) = -.01$	$\Sigma(y - \hat{y})^2 = 2448.86$

SSE = 2448.86

$$S_e = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{2448.86}{10}} = 15.65$$

Practice Problems

Q.02 Check where a linear regression model is appropriate for the following data.

x	60	70	80	85	95
y (Actual Value)	70	65	70	95	85
y^ (Predicted Value)	65.41	71.84	78.28	81.50	87.94

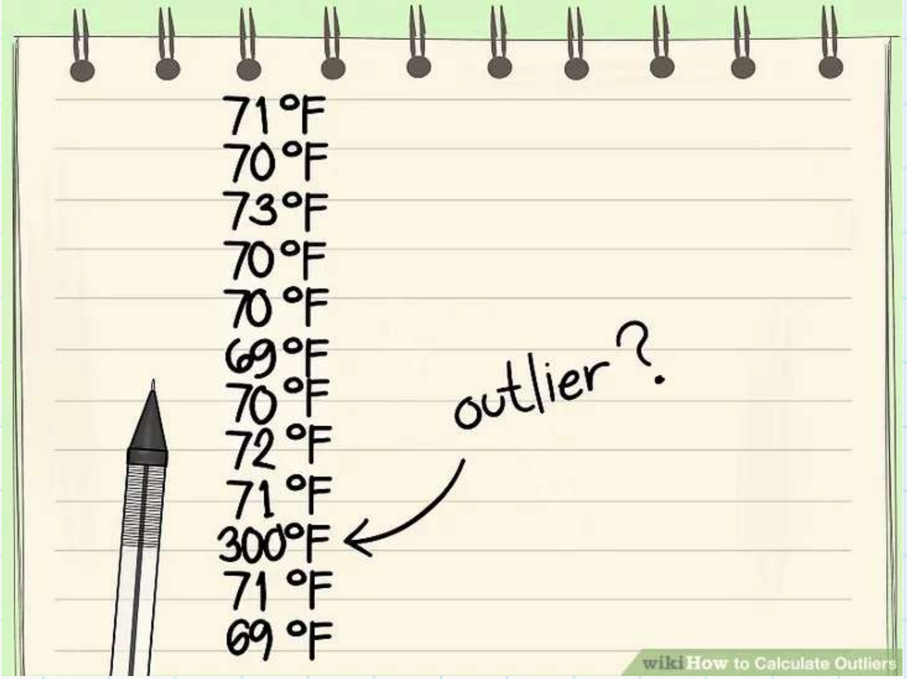
Q.03 The equation of a regression line is and the data are as follows:

x	57	11	12	19	25
y	47	38	32	24	22

Solve for the residuals and graph a residual plot. Do these data seem to violate any of the assumptions of regression?

Determine SSE and SE of the Estimate.

Assumption : Outliers / Influential Cases



Assumption : Outliers / Influential Cases

- An outlier is a data point which is very far, somehow, from the rest of the data.
- An outlier is an observation of data that does not fit the rest of the data. It is sometimes called an extreme value.
- When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. Outlier is a value that lies in a data series on its extremes, which is either very small or large and thus can affect the overall observation made from the data series. Outliers are also termed as extremes because they lie on the either end of a data series.
- Outliers are usually treated as abnormal values that can affect the overall observation due to its very high or low extreme values and hence should be discarded from the data series.

Assumption : Outliers / Influential Cases

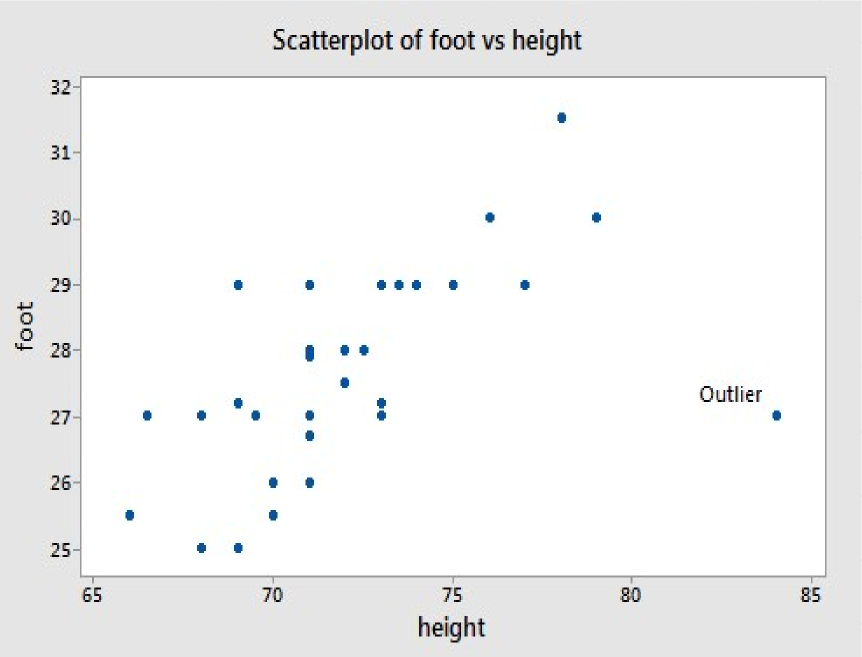
- Most common causes of outliers on a data set:
  1. Data entry errors (human errors)
  2. Measurement errors (instrument errors)
  3. Experimental errors (data extraction or experiment planning/executing errors)
  4. Intentional (dummy outliers made to test detection methods)
  5. Data processing errors (data manipulation or data set unintended mutations)
  6. Sampling errors (extracting or mixing data from wrong or various sources)
  7. Natural (not an error, novelties in data)

- Univariate outliers can simply be identified by considering the distributions of individual variables say by using boxplots. Multivariate outliers can be detected from residual scatterplots.

Assumption : Outliers / Influential Cases

Example:

Let us consider a dataset where y = foot length (cm) and x = height (in) for n = 33 male students in a class. A scatterplot of the male foot length and height data shows one point labeled as an outlier.

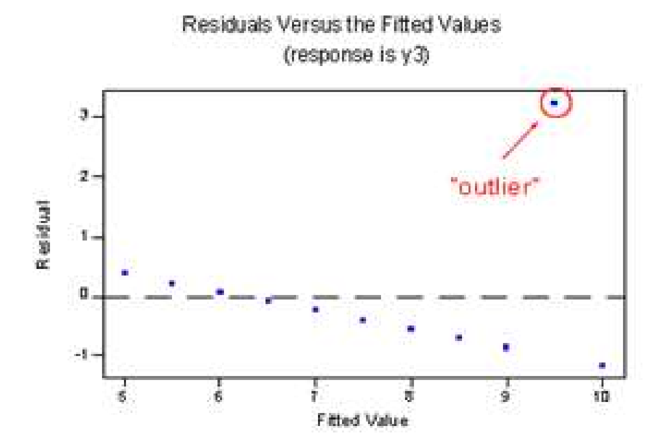


- There is a clear outlier with values (x_i , y_i) = (84, 27).
- If that data point is *deleted* from the dataset, the estimated equation, using the other 32 data points, is y^i=0.253+0.384xi.
- For the deleted observation, x_i = 84, so
  ^yi=0.253+0.384(84)=32.5093
- The (unstandardized) deleted residual is

  di= 27−32.5093=−5.5093

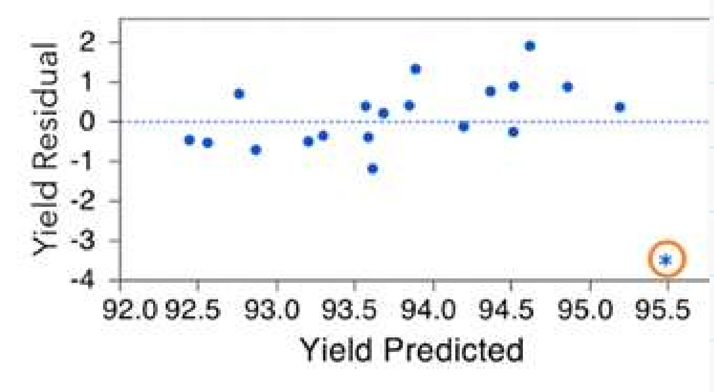
Assumption : Outliers / Influential Cases

Residual Plots to deduct Multivariate Outliers:



he outlier is clearly apparent in this residual plot. Image: PSU.edu

Another example is the regression model for **Yield** as a function of **Concentration** is significant, but note that the line of fit appears to be tilted towards the outlier. We can see the effect of this outlier in the residual by predicted plot. The center line of zero does not appear to pass through the points.



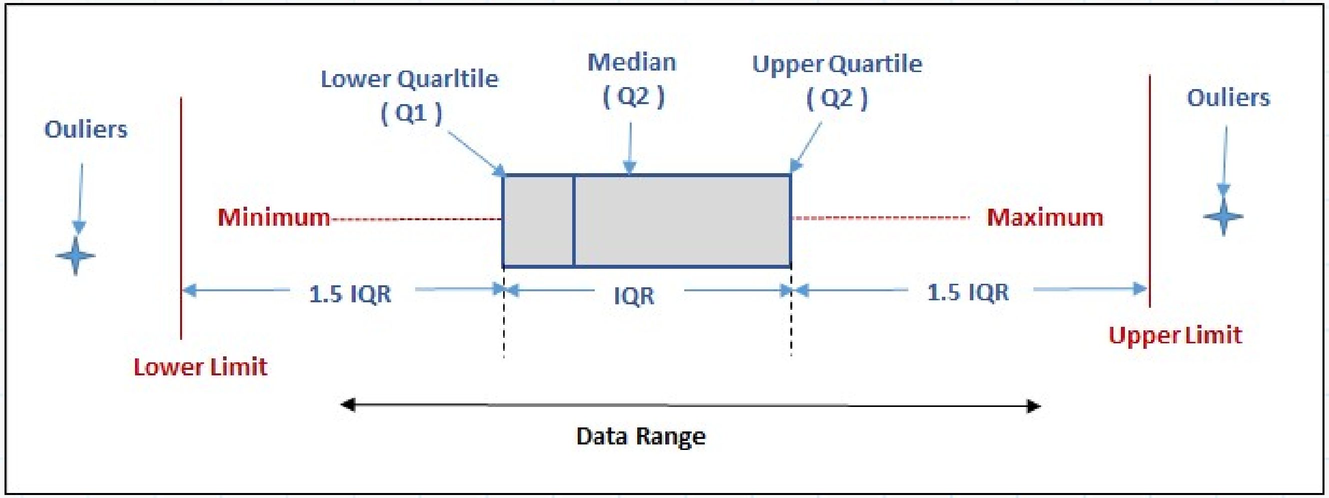
Assumption : Outliers / Influential Cases

Box Plot Diagram to identify Outliers Lower

- Box plot diagram also termed as Whisker's plot is a graphical method typically depicted by quartiles and inter quartiles that helps in defining the upper limit and lower limit beyond which any data lying will be considered as outliers.
- The very purpose of this diagram is to identify outliers and discard it from the data series before making any further observation so that the conclusion made from the study gives more accurate results not influenced by any extremes or abnormal values.
- Box plots can be used as an initial screening tool for outliers as they provide a graphical depiction of data distribution and extreme values

Assumption : Outliers / Influential Cases

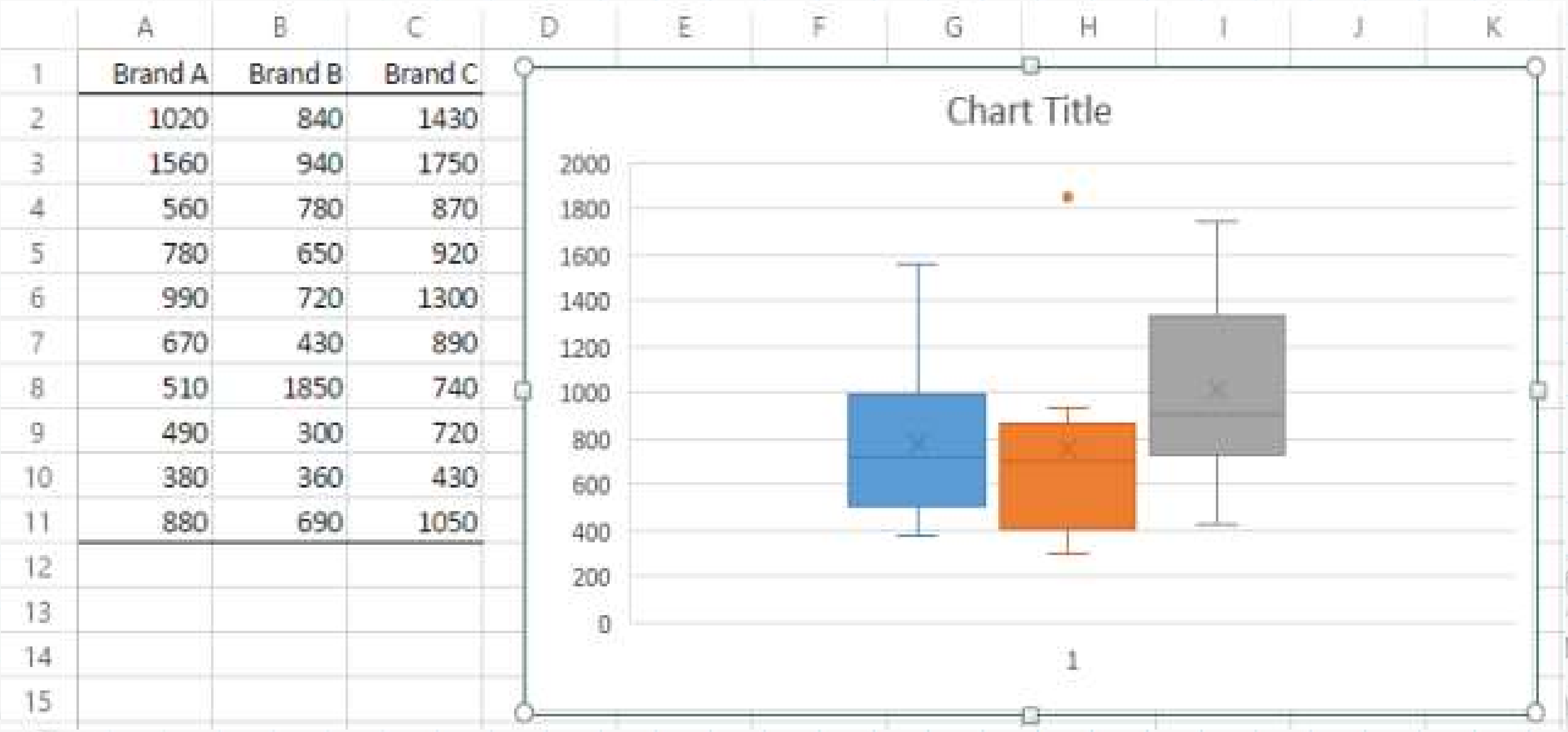
Box Plot Diagram to identify Outliers Lower



Where

- n be the number of data values in the data set.
- The Median (Q2) is the middle value of the data set.
- The Lower quartile (Q1) is the median of the lower half of the data set=$\frac{1}{4}(n+1)th\ term$
- The Upper quartile (Q3) is the median of the upper half of the data set=$\frac{3}{4}(n+1)th\ term$
- The Interquartile range (IQR) is the spread of the middle 50% of the data values.
- Interquartile Range (IQR) = Upper Quartile (Q3) – Lower Quartile (Q1) = Q3 – Q1
- Lower Limit = Q1 – 1.5 IQR.
- Upper Limit = Q3 + 1.5 IQR

Assumption : Outliers / Influential Cases



	A	B	C
1	Brand A	Brand B	Brand C
2	1020	840	1430
3	1560	940	1750
4	560	780	870
5	780	650	920
6	990	720	1300
7	670	430	890
8	510	1850	740
9	490	300	720
10	380	360	430
11	880	690	1050
12			
13			
14			
15			

Assumption : Autocorrelation

- Autocorrelation refers to the degree of correlation between the values of the same variables across different observations in the data.
- The concept of autocorrelation is most often discussed in the context of time series data in which observations occur at different points in time (e.g., air temperature measured on different days of the month).
- For example, one might expect the air temperature on the 1st day of the month to be more similar to the temperature on the 2nd day compared to the 31st day. If the temperature values that occurred closer together in time are, in fact, more similar than the temperature values that occurred farther apart in time, the data would be auto correlated.
- In a regression analysis, autocorrelation of the regression residuals can also occur if the model is incorrectly specified.
- For example, if you are attempting to model a simple linear relationship but the observed relationship is non-linear (i.e., it follows a curved or U-shaped function), then the residuals will be auto correlated.

Assumption : Autocorrelation

How to Detect Autocorrelation

- A common method of testing for autocorrelation is the Durbin-Watson test.
- The Durbin-Watson tests produces a test statistic that ranges from 0 to 4.
- Values close to 2 (the middle of the range) suggest less autocorrelation, and values closer to 0 or 4 indicate greater positive or negative autocorrelation respectively.

Coefficient of Determination – R^2 Statistics (Accessing the accuracy)

- The coefficient of determination is the square of the coefficient of correlation.
- Or The coefficient of determination is the proportion of variability of the dependent variable (y) accounted for or explained by the independent variable (x).
- The coefficient of determination ranges from 0 to 1. An r^2 of zero means that the predictor accounts for none of the variability of the dependent variable and that there is no regression prediction of y by x.
- An r^2 of 1 means perfect prediction of y by x and that 100% of the variability of y is accounted for by x.
- Of course, most r^2 values are between the extremes.
- The researcher must interpret whether a particular r^2 is high or low, depending on the use of the model and the context within which the model was developed.

Coefficient of Determination – R² Statistics

$$\boxed{r^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{\sum y^2 - \frac{(\sum y)^2}{n}}}$$

- The coefficient of multiple determination (R^2) is analogous to the coefficient of determination (r^2).

- R^2 represents the proportion of variation of the dependent variable, y, accounted for by the independent variables in the regression model.

- As with r^2, the range of possible values for R^2 is from 0 to 1.

Coefficient of Determination – R^2 Statistics

- R-squared is always between 0 and 100%:
- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.
- In general, the higher the R-squared, the better the model fits your data.
- Of course, it is desirable for R2 to be high, indicating the strong predictability of a regression model.
- The coefficient of multiple determination can be calculated by the following formula..

Coefficient of Determination – R² Statistics

$$R^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

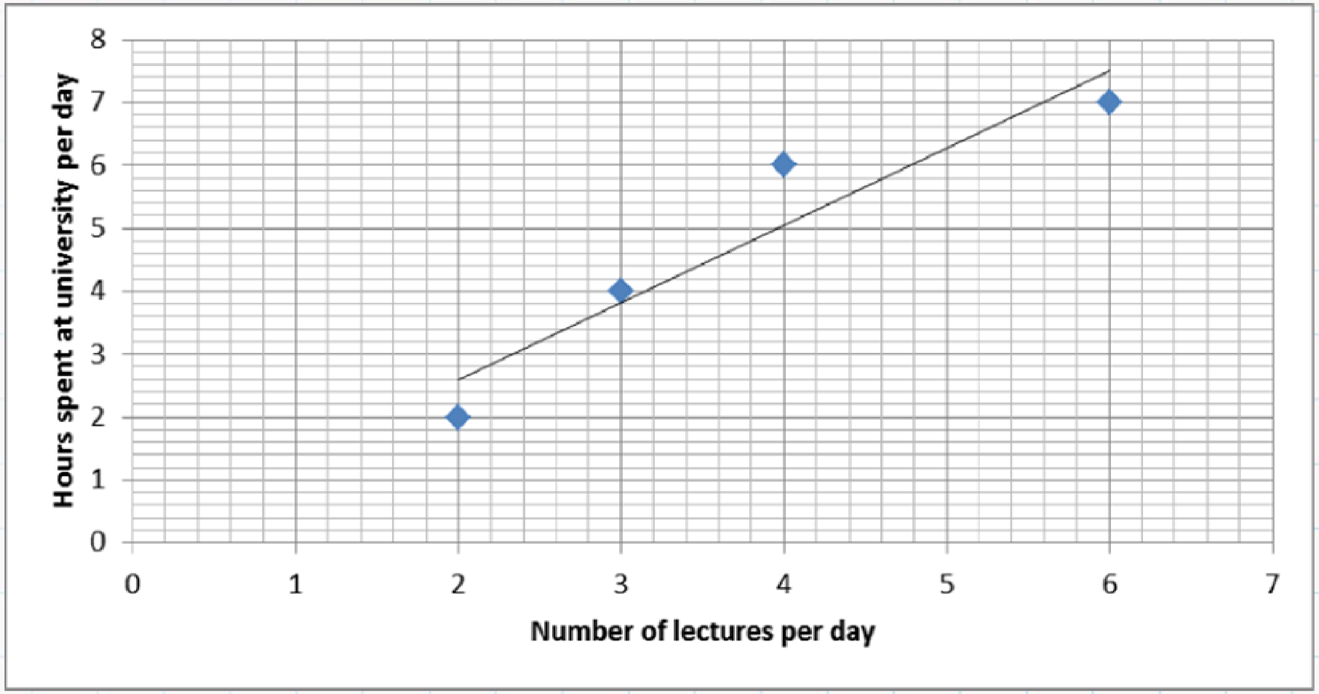
Key Limitations of R-squared:

- R-squared cannot determine whether the coefficient estimates and predictions are biased, which is why you must assess the residual plots.

- R-squared does not indicate whether a regression model is adequate. You can have a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data!

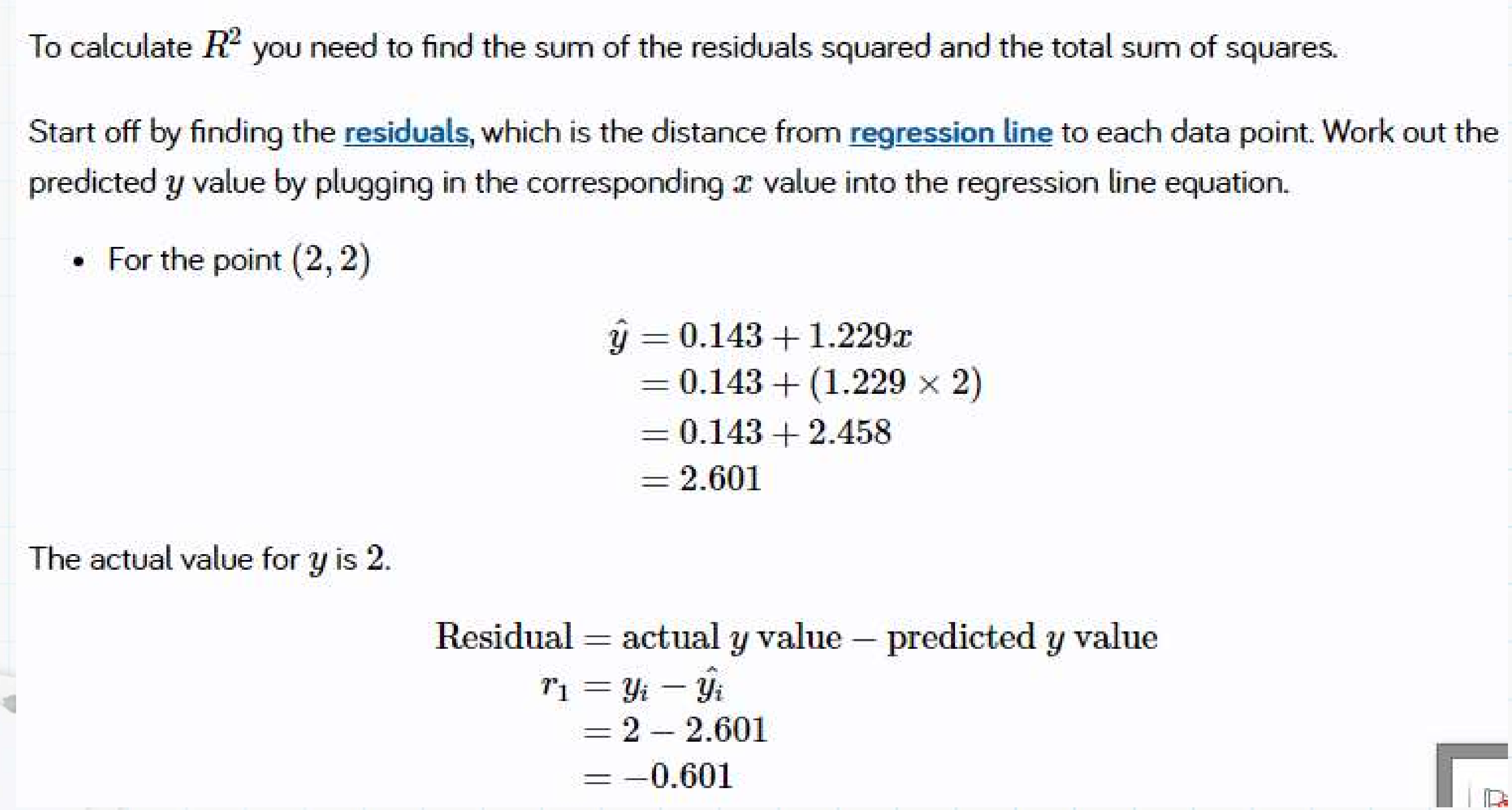
Problems

Below is a graph showing how the number lectures per day affects the number of hours spent at university per day. The equation of the regression line is drawn on the graph and it has equation y^=0.143+1.229x. Calculate R^2.



Problems

Below is a graph showing how the number lectures per day affects the number of hours spent at university per day. The equation of the regression line is drawn on the graph and it has equation y^=0.143+1.229x. Calculate R^2.



To calculate R^2 you need to find the sum of the residuals squared and the total sum of squares.

Start off by finding the residuals, which is the distance from regression line to each data point. Work out the predicted y value by plugging in the corresponding x value into the regression line equation.

- For the point $(2, 2)$

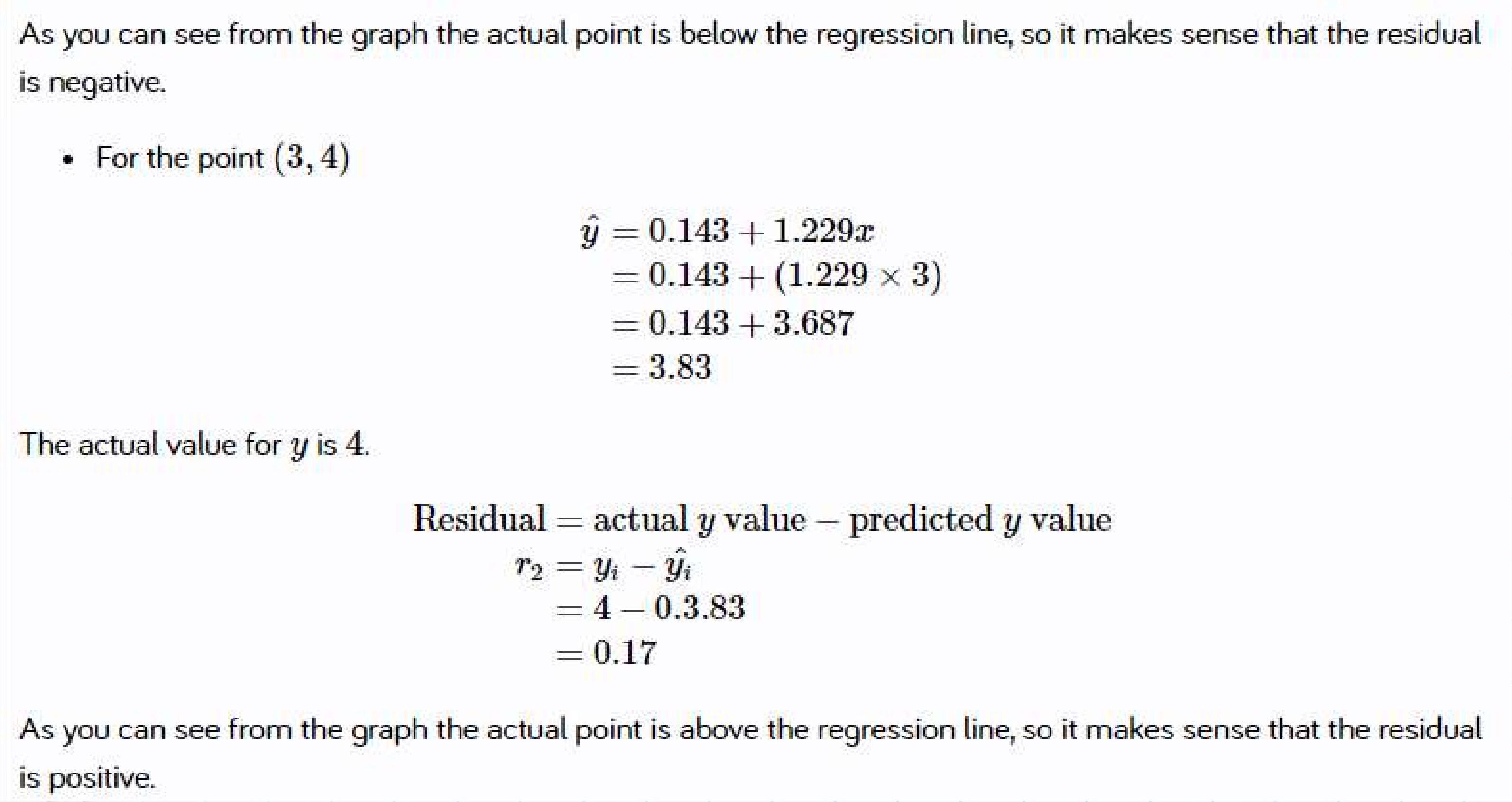
$$\begin{aligned}\hat{y} &= 0.143 + 1.229x\\ &= 0.143 + (1.229 \times 2)\\ &= 0.143 + 2.458\\ &= 2.601\end{aligned}$$

The actual value for y is 2.

$$\begin{aligned}\text{Residual} &= \text{actual } y \text{ value} - \text{predicted } y \text{ value}\\ r_1 &= y_i - \hat{y_i}\\ &= 2 - 2.601\\ &= -0.601\end{aligned}$$

Problems

Below is a graph showing how the number lectures per day affects the number of hours spent at university per day. The equation of the regression line is drawn on the graph and it has equation y^=0.143+1.229x. Calculate R^2.



As you can see from the graph the actual point is below the regression line, so it makes sense that the residual is negative.

- For the point $(3, 4)$

$$\begin{aligned}\hat{y} &= 0.143 + 1.229x \\ &= 0.143 + (1.229 \times 3) \\ &= 0.143 + 3.687 \\ &= 3.83\end{aligned}$$

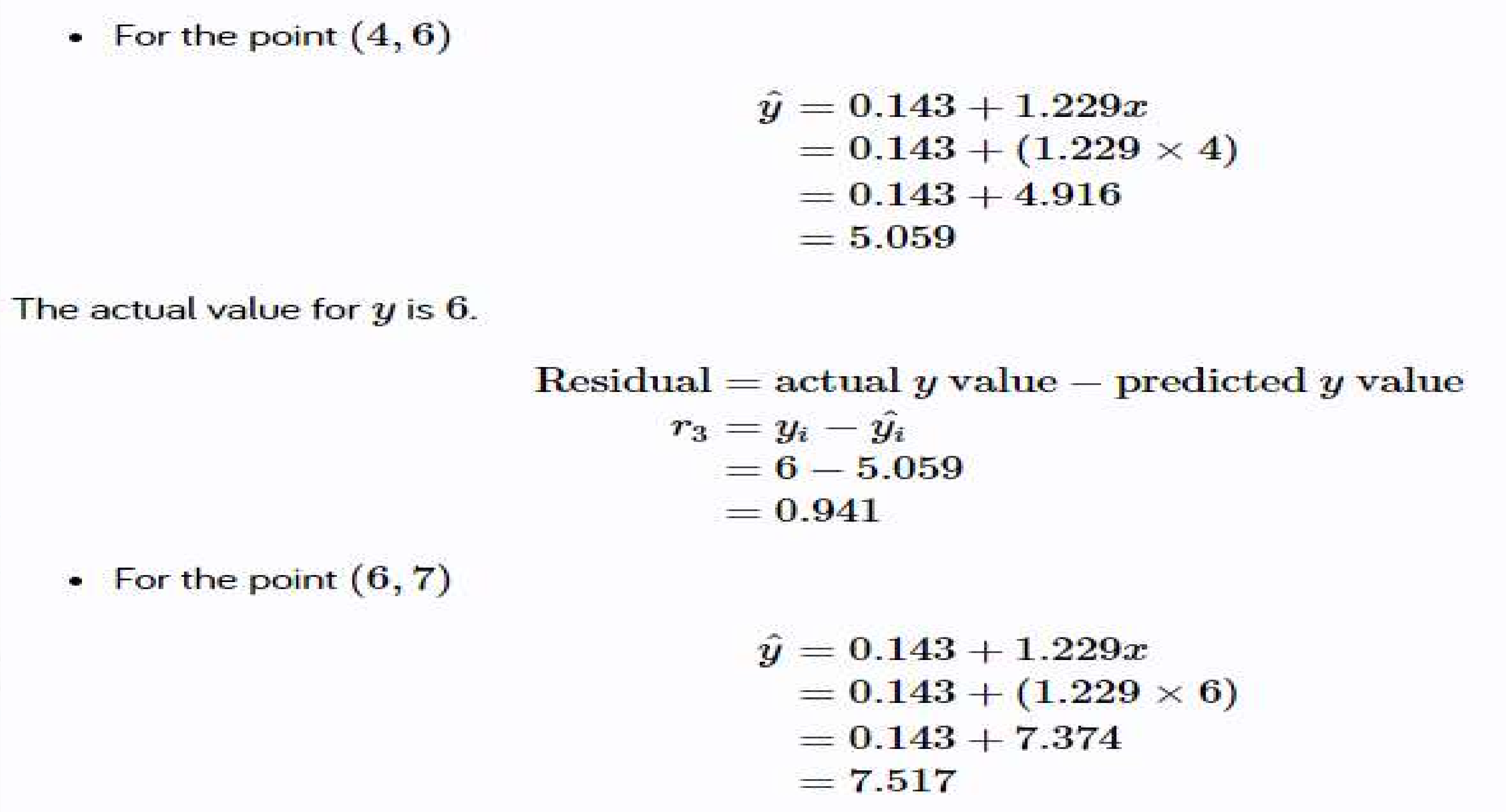
The actual value for y is 4.

$$\begin{aligned}\text{Residual} &= \text{actual } y \text{ value} - \text{predicted } y \text{ value} \\ r_2 &= y_i - \hat{y}_i \\ &= 4 - 0.3.83 \\ &= 0.17\end{aligned}$$

As you can see from the graph the actual point is above the regression line, so it makes sense that the residual is positive.

Problems

Below is a graph showing how the number lectures per day affects the number of hours spent at university per day. The equation of the regression line is drawn on the graph and it has equation y^=0.143+1.229x. Calculate R^2.



- For the point $(4, 6)$

$$\begin{aligned}\hat{y} &= 0.143 + 1.229x \\ &= 0.143 + (1.229 \times 4) \\ &= 0.143 + 4.916 \\ &= 5.059\end{aligned}$$

The actual value for y is 6.

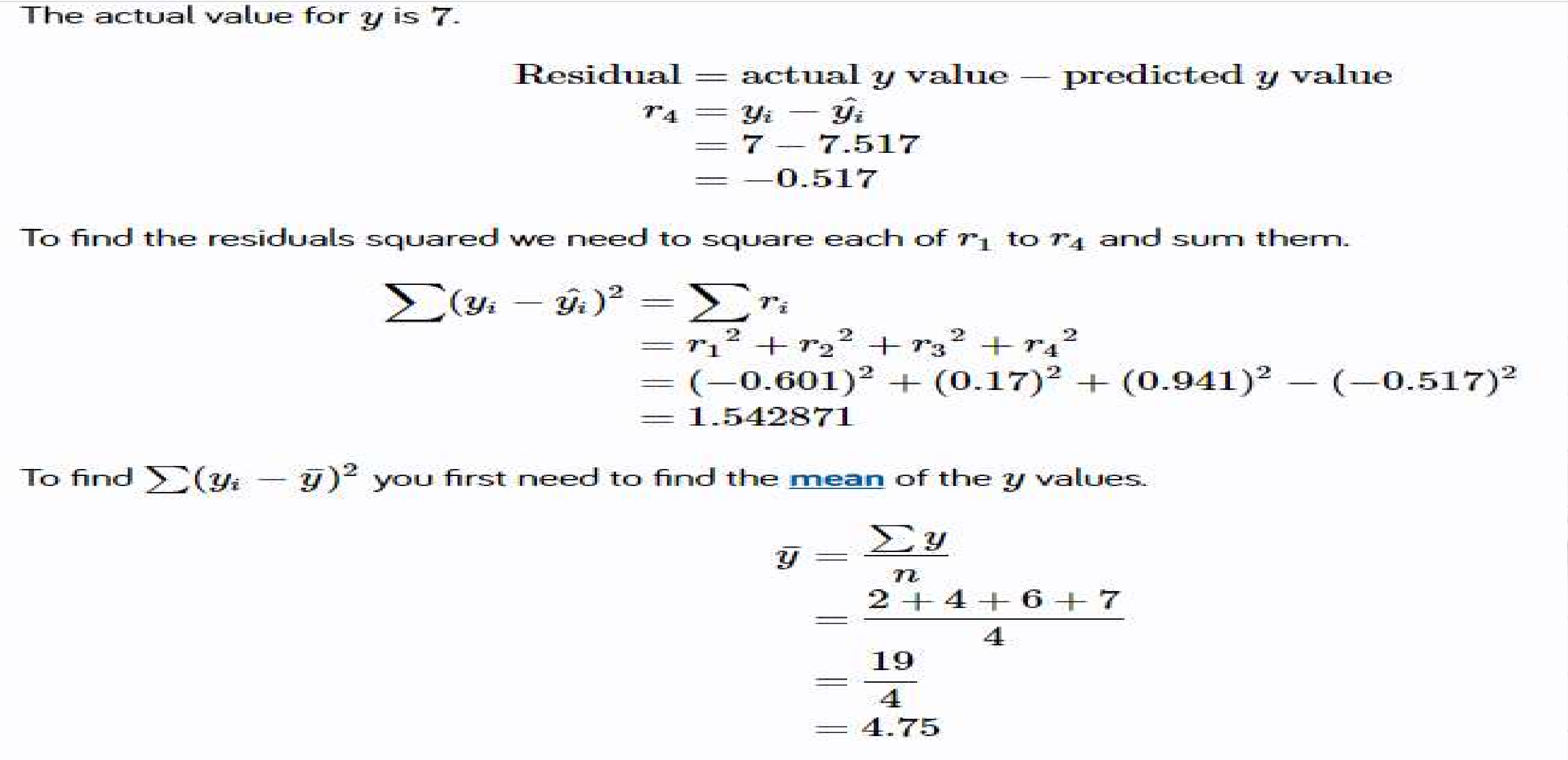
$$\begin{aligned}\text{Residual} &= \text{actual } y \text{ value} - \text{predicted } y \text{ value} \\ r_3 &= y_i - \hat{y_i} \\ &= 6 - 5.059 \\ &= 0.941\end{aligned}$$

- For the point $(6, 7)$

$$\begin{aligned}\hat{y} &= 0.143 + 1.229x \\ &= 0.143 + (1.229 \times 6) \\ &= 0.143 + 7.374 \\ &= 7.517\end{aligned}$$

Problems

Below is a graph showing how the number lectures per day affects the number of hours spent at university per day. The equation of the regression line is drawn on the graph and it has equation y^=0.143+1.229x. Calculate R^2.



The actual value for y is 7.

$$\begin{aligned}\text{Residual} &= \text{actual } y \text{ value} - \text{predicted } y \text{ value}\\ r_4 &= y_i - \hat{y_i}\\ &= 7 - 7.517\\ &= -0.517\end{aligned}$$

To find the residuals squared we need to square each of r_1 to r_4 and sum them.

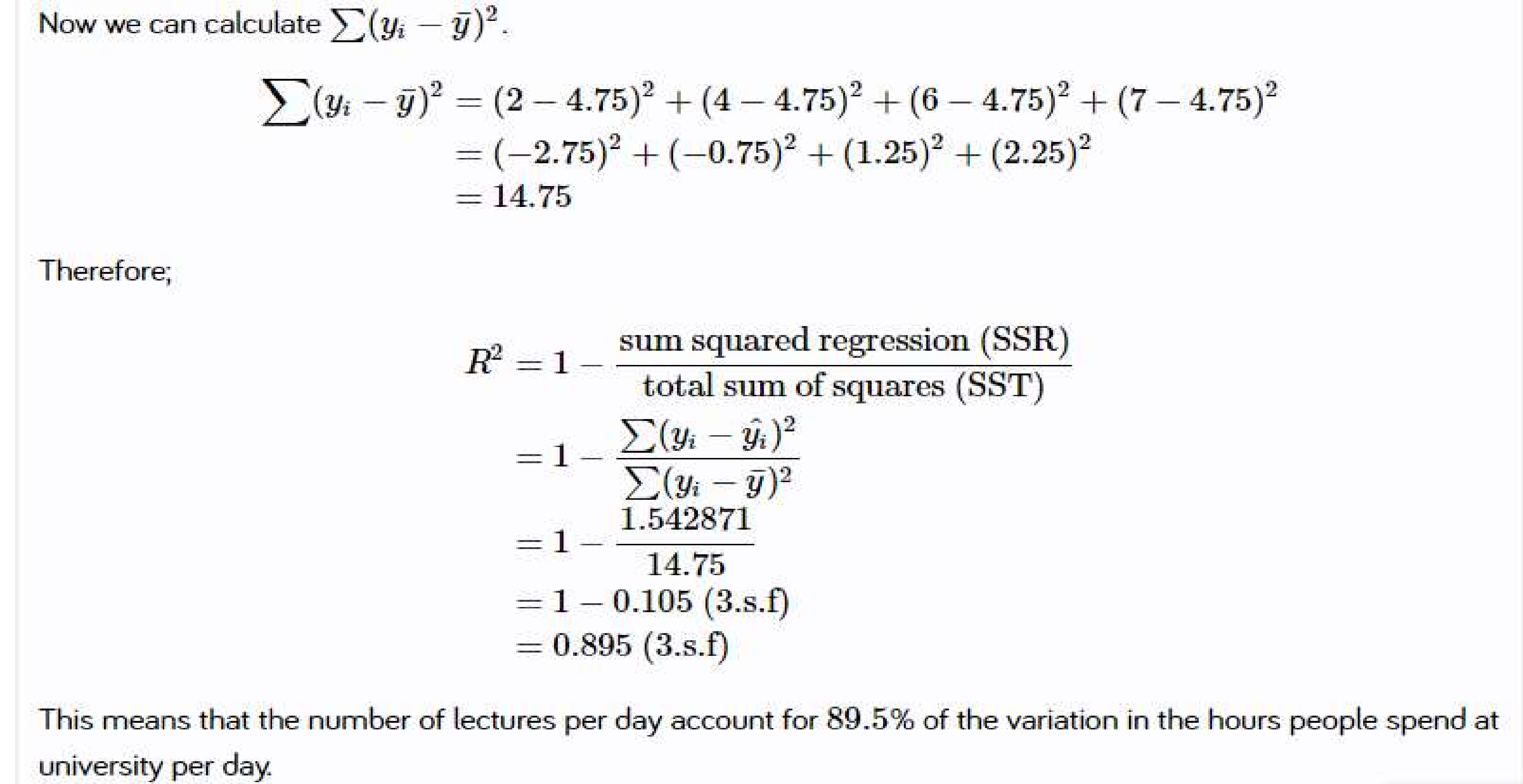
$$\begin{aligned}\sum(y_i - \hat{y_i})^2 &= \sum r_i\\ &= {r_1}^2 + {r_2}^2 + {r_3}^2 + {r_4}^2\\ &= (-0.601)^2 + (0.17)^2 + (0.941)^2 - (-0.517)^2\\ &= 1.542871\end{aligned}$$

To find $\sum(y_i - \bar{y})^2$ you first need to find the mean of the y values.

$$\begin{aligned}\bar{y} &= \frac{\sum y}{n}\\ &= \frac{2+4+6+7}{4}\\ &= \frac{19}{4}\\ &= 4.75\end{aligned}$$

Problems

Below is a graph showing how the number lectures per day affects the number of hours spent at university per day. The equation of the regression line is drawn on the graph and it has equation y^=0.143+1.229x. Calculate R^2.



Now we can calculate $\sum(y_i - \bar{y})^2$.

$$\begin{aligned}\sum(y_i - \bar{y})^2 &= (2-4.75)^2 + (4-4.75)^2 + (6-4.75)^2 + (7-4.75)^2 \\ &= (-2.75)^2 + (-0.75)^2 + (1.25)^2 + (2.25)^2 \\ &= 14.75\end{aligned}$$

Therefore;

$$\begin{aligned}R^2 &= 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}} \\ &= 1 - \frac{\sum(y_i - \hat{y_i})^2}{\sum(y_i - \bar{y})^2} \\ &= 1 - \frac{1.542871}{14.75} \\ &= 1 - 0.105 \text{ (3.s.f)} \\ &= 0.895 \text{ (3.s.f)}\end{aligned}$$

This means that the number of lectures per day account for 89.5% of the variation in the hours people spend at university per day.

Adjusted R^2

- The adjusted R-squared is a modified version of R-squared for the number of predictors in a model.
- R2 assumes that every single variable explains the variation in the dependent variable.
- The adjusted R2 tells you the percentage of variation explained by only the independent variables that actually affect the dependent variable.
- The value of R Squared never decreases.
- Adding new independent variables will result in an increased value of R Squared.
- This is a major flow as R Squared will suggest that adding new variables irrespective of whether they are really significant or not, will increase the value.
- For example, the person’s Name for predicting the Salary, the value of R squared will increase suggesting that the model is better.
- This is where Adjusted R Squared comes to the rescue.

Adjusted R²

- Compared to R Squared which can only increase, Adjusted R Squared has the capability to decrease with the addition of less significant variables, thus resulting in a more reliable and accurate evaluation.

$$\sqrt{SSE/(n-k-1)} = SE$$

$$Adj\ \mathrm{R}^2 = 1 - \frac{\left[SSE/(n-k-1) \right]}{SS_{yy}/(n-1)} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1} \right]$$

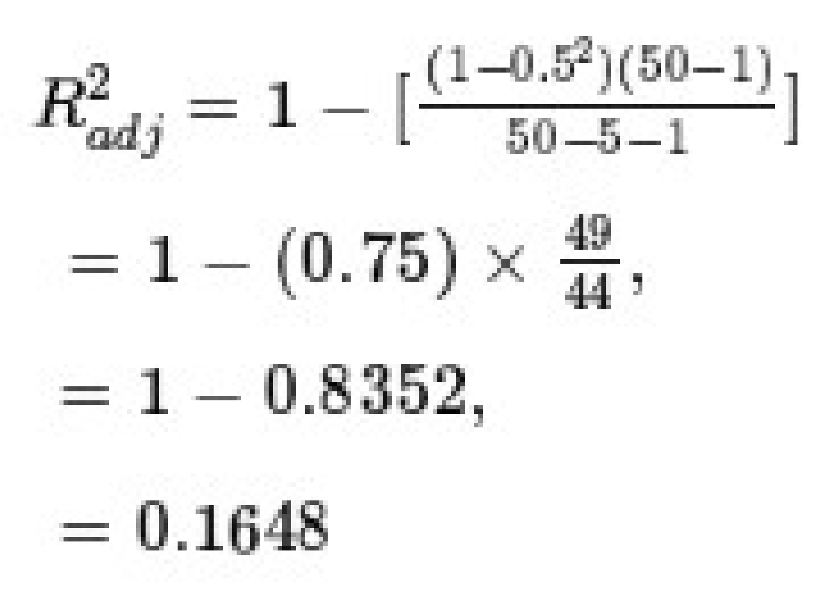
Note:

- Adjusted R-squared value always be less than or equal to r-squared value.
- Adjusted R-square should be used while selecting important predictors (independent variables) for the regression model.
- If you add more and more useless variables to a model, adjusted R-squared will decrease.
- If you add more useful variables, adjusted R-squared will increase.

Problems

A fund has a sample R-squared value close to $\overset{r^2}{\underline{0.5}}$ and it is doubtlessly offering higher risk adjusted returns with the sample size of $\underset{n}{\underline{50}}$ for $\underset{k}{5}$ predictors. Find Adjusted R square value.

Solution: .
Sample size = 50 Number of predictor = 5 Sample R - square = 0.5.Substitute the qualities in the equation



$$R^2_{adj} = 1 - [\frac{(1-0.5^2)(50-1)}{50-5-1}]$$

$$= 1 - (0.75) \times \frac{49}{44},$$

$$= 1 - 0.8352,$$

$$= 0.1648$$