12.2.2: Residuals

When we overlay the regression equation on a scatterplot, most of the time, the points do not lie on the line itself. The vertical distance between the actual value of y and the predicted value of $\hat{y}$ is called the **residual**. The numeric value of the residual is found by subtracting the predicted value of y from the actual value of y: $y - \hat{y}$. When we find the line of best fit using least squares regression, this finds the regression equation with the smallest sum of the residuals $\sum y - \hat{y}$.

When your residual is positive, then your data point is above the regression line, when the residual is negative, your data point is below the regression line. If you were to find the residuals for all the sample points and add them up you would get zero. The expected value of the residuals will always be zero. The regression equation is found so that there is just as much distance for the residuals above the line as there is below the line.

Find the residual for the point $(15, 80)$ for the exam data.

Hours Studied for Exam 20 16 20 18 17 16 15 17 15 16 15 17 16 17 14 Grade on Exam 89 72 93 84 81 75 70 82 69 83 80 83 81 84 76

Solution

Figure 12-15 is a scatterplot with the regression equation $\hat{y} = 26.742 + 3.216346x$ from the exam data.

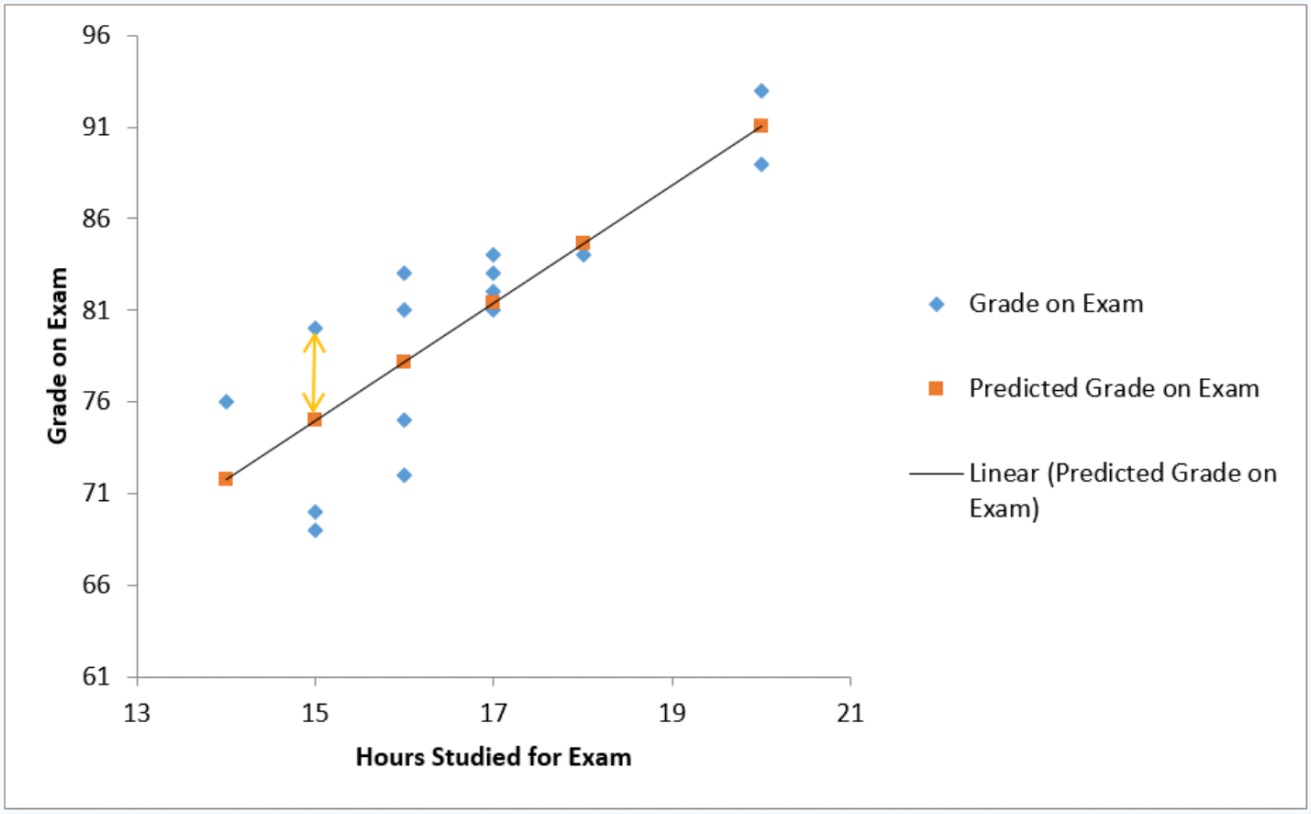


Figure 12-15: Scatterplot showing the linear prediction of grades and the distance from point (15, 80) to the predicted grade for $x = 15$.

The blue diamonds represent the sample data points. The orange squares are the predicted y for each value of x. If we connect the orange squares, we get the linear regression equation. The vertical distance between each data point and the regression equation is called the residual. The numeric value can be found by subtracting the observed y with its corresponding predicted value, $y - \hat{y}$. We use e_i to represent the i^{th} residual where $e_i = y_i - \hat{y}_i$. The residual for the point $(15, 80)$ is drawn on the scatterplot vertically as a yellow double-sided arrow to visually show the size of the residual.

If you were to predict a student's exam grade when they studied 15 hours, you would get a predicted grade of $\hat{y} = 26.742 + 3.216346 \cdot 15 = 74.9865$ The residual for the point $(15, 80)$ then would be $y - \hat{y} = 80 - 74.9865 = 5.0135$. This is the length of the vertical yellow arrow connecting the point $(15, 80)$ to the point $(15, 74.9865)$

Standard Error of Estimate

The standard deviation of the residuals is called the **standard error of estimate** or s. Some texts will use a subscript s_e or s_{est} to distinguish the different standard deviations from one another. When all of your data points line up in a perfectly straight line, $s = 0$ since none of your points deviate from the regression line. As your data points get more scattered away from a regression line, s gets larger. When you are analyzing a regression model, you want s to be as small as possible.

Standard Error of Estimate

$$s_{est} = s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{MSE}$$

The standard error of estimate is the standard deviation of the residuals. The standard error of estimate measures the deviation in the vertical distance from data points to the regression equation. The units of s are the same as the units of y.

Use the exam data to find the standard error of estimate.

Solution

To find the $\sum (y_i - \hat{y}_i)^2$ you would need to find the residual for every data point, square the residuals, and sum them up. This is a lot of math. Recall the regression ANOVA table found earlier. The MSE = 15.4912.

Source	SS	*df*	MS	F
Regression	430.3471154	1	430.3471154	27.780
Error	201.3862	13	15.4912	
Total	631.7333	14		

The mean square error is the variance of the residuals, if we take the square root of the MSE we find the standard deviation of the residuals, which is the standard error of estimate.

$s = \sqrt{MSE} = \sqrt{15.4912} = 3.9359$

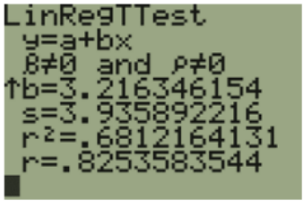
You can also use the technology to find s.

Excel:

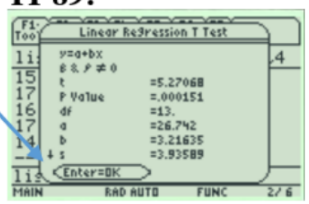
Regression Statistics	
Multiple R	0.825358
R Square	0.681216
Adjusted R Square	0.656695
Standard Error	3.935892
Observations	15

s = standard error

TI-84:



TI-89:



SPSS:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.825[a]	.681	.657	3.93589