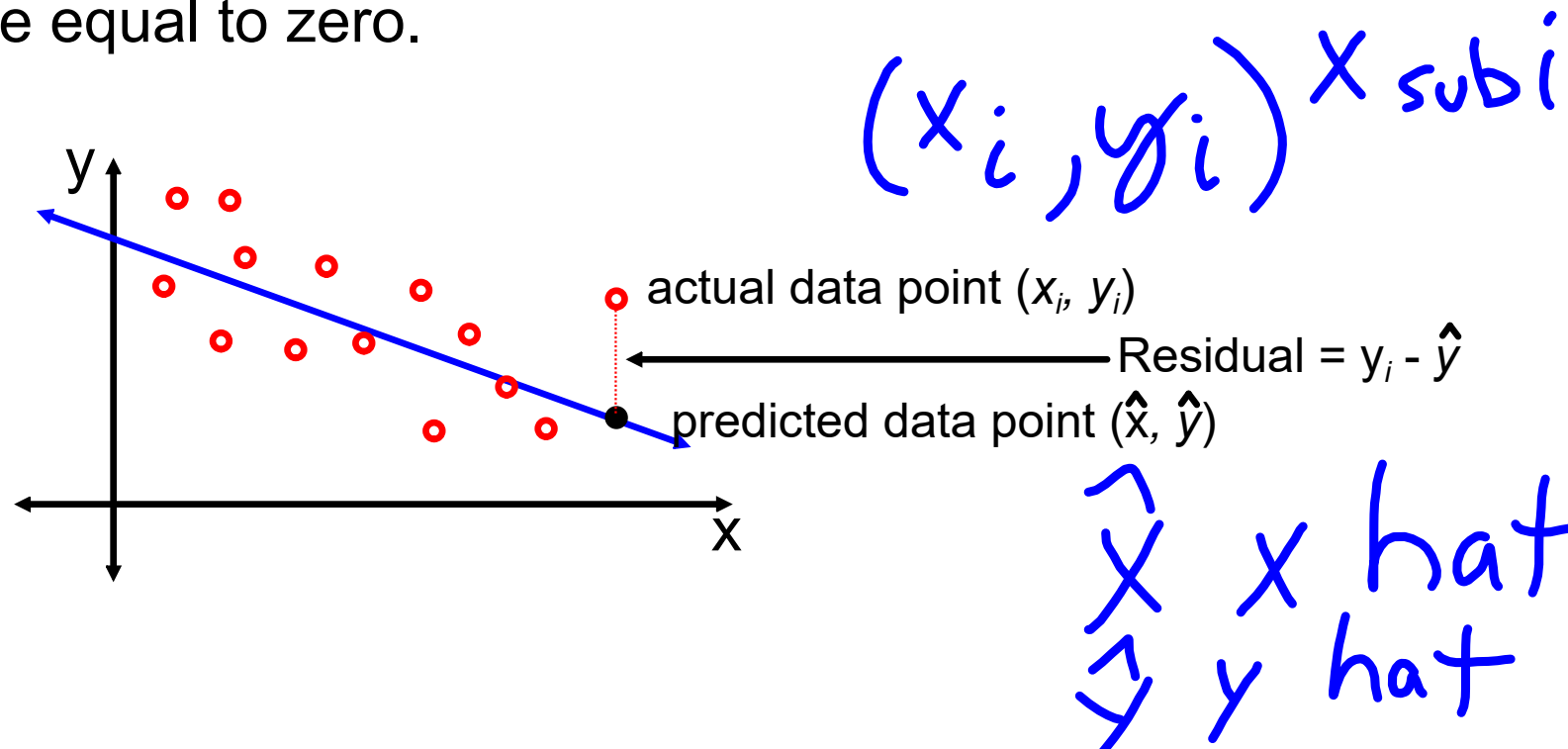


Exploring Bivariate Numerical Data: Part 2

- Topics: Residuals and Least-Square Lines
- Objective: Students will be able to interpret residual points and interpret slope and y-intercepts of linear models
- Standards: AP Stats: DAT-1 (EU), DAT-1.E (LO), DAT-1.E.1 (EK)

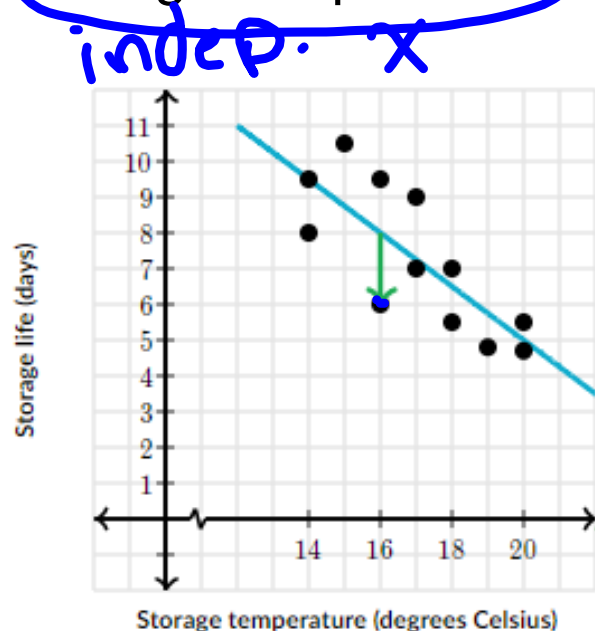
What is a Residual?

Definition: In regression analysis, the difference between the observed value of the dependent variable (y) and the predicted value (\hat{y}) is called the residual (e). Each data point has one residual. Residual = Observed value - Predicted value. $e = y - \hat{y}$ Both the sum and the mean of the residuals are equal to zero.



Calculating and Interpreting Residuals

Example 1: Cadan tracked the storage life of bunches of bananas in his store and the storage temperature of the area where he displayed them. An approximate least-squares regression line was used to predict the storage life from a given storage temperature.



dependent (y)

Interpret the residual for the bunch indicated in the scatterplot above.

- ~~1.~~ This bunch's storage temperature was 2°C warmer than predicted based on storage life.
- ~~2.~~ This bunch's storage temperature was 2°C cooler than predicted based on storage life.
3. This bunch's storage life was 2 more days than predicted based on the storage temperature.
- 4.** This bunch's storage life was 2 fewer days than predicted based on the storage temperature.

Calculating and Interpreting Residuals

Example 2:

Alexander uses cupric chloride to etch circuit boards. He recorded the room temperature, in $^{\circ}\text{C}$, and the etching rate, in $\frac{\mu\text{m}}{\text{min}}$, of the cupric chloride.

After plotting his results, Alexander noticed that the relationship between the two variables was fairly linear, so he used the data to calculate the following least squares regression equation for predicting the etching rate from the room temperature:

$$\hat{y} = 2 + \frac{1}{5}x$$

What is the residual if the room temperature was 25°C and the cupric chloride had an etching rate of $5 \frac{\mu\text{m}}{\text{min}}$?

$\frac{\mu\text{m}}{\text{min}}$

Actual

$$5 - 7 = -2$$

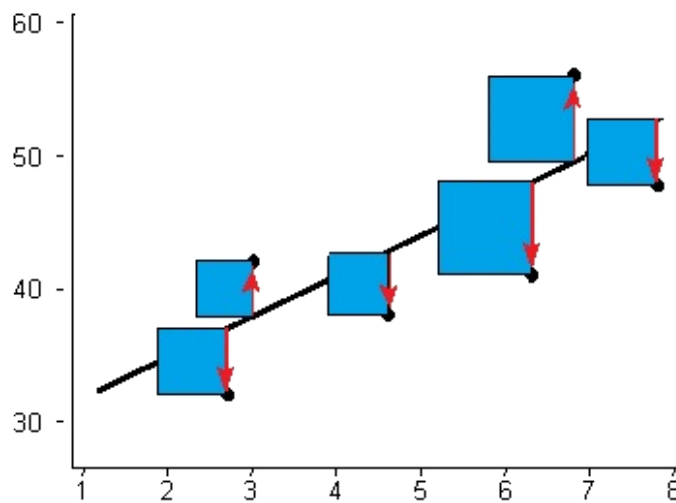
predicted

$$y = 2 + \frac{1}{5}(25)$$

$$2 + 5$$

Least-Square Regression Lines

Definition: The Least Squares Regression Line is the line that makes the vertical distance from the data points to the regression line as small as possible. It's called a "least squares" because the best line of fit is one that minimizes the variance (the sum of squares of the errors).



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

$$y = mx + b$$

$$\left(\begin{array}{c} 1 \\ 4 \end{array} , \begin{array}{c} 3 \\ -5 \end{array} \right) \quad \frac{y_2 - y_1}{x_2 - x_1}$$

$$y - 3 = -2(x - 1)$$

Least-Square Regression Lines

How to calculate the Least-Square Line:

$$\hat{y} = \overset{\text{y-intercept}}{a} + bx$$

$$b = r \frac{S_y}{S_x}$$

S_x

S_y

	mean	standard deviation	
$x = \text{creek temperature } (^{\circ}\text{C})$	$\bar{x} = 10.2$	$s_x = 2.8$	Sample standard deviation for x
$y = \text{number of flatworms}$	$\bar{y} = 37.6$	$s_y = 30.8$	Sample standard deviation for y
		$r = -0.98$	Correlation coefficient

Mean of x

Mean of y

Least-Square Regression Lines

Example1: A limnologist takes samples from a creek on several days and counts the numbers of flatworms in each sample. The limnologist wants to look at the relationship between the temperature of the creek and the number of flatworms in the sample. The data show a linear pattern with the summary statistics shown below:

	mean	standard deviation
$x = \text{creek temperature } (^{\circ}\text{C})$	$\bar{x} = 10.2$	$s_x = 2.8$
$y = \text{number of flatworms}$	$\bar{y} = 37.6$	$s_y = 30.8$
	$r = -0.98$	

$$b = r \frac{S_y}{S_x}$$

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

Find the equation of the least-squares regression line for predicting the number of flatworms from the creek temperature. *Round to the nearest hundredth.*

$$\hat{y} = \boxed{} + \boxed{}x$$

Least-Square Regression Lines

$$\hat{y} = a + bx \quad b = r \frac{S_y}{S_x} = -0.98 \left(\frac{30.8}{2.8} \right)$$

$$37.6 = a + (-10.786)(10.2)$$

$$37.6 = a - 109.956$$

$$-10.786$$

	mean	standard deviation
$x = \text{creek temperature } (^{\circ}\text{C})$	$\bar{x} = 10.2$	$s_x = 2.8$
$y = \text{number of flatworms}$	$\bar{y} = 37.6$	$s_y = 30.8$
	$r = -0.98$	

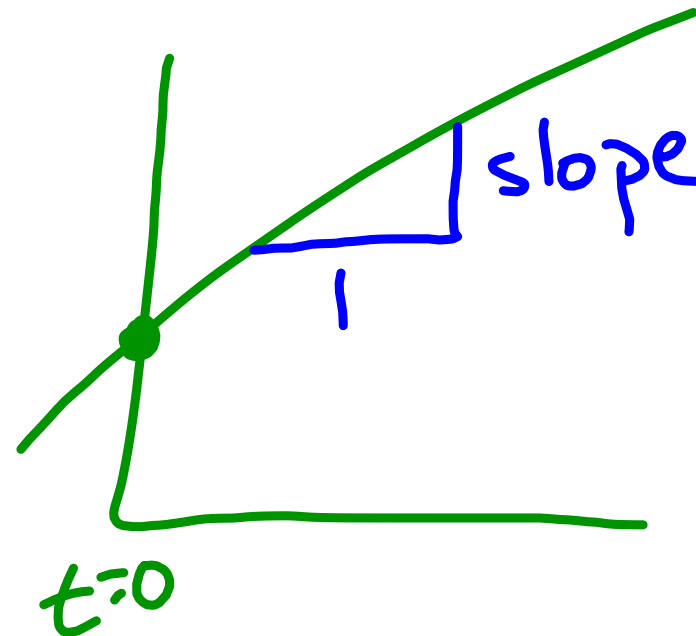
Find the equation of the least-squares regression line for predicting the number of flatworms from the creek temperature. *Round to the nearest hundredth.*

$$\hat{y} = 147.56 - 10.78x$$

Interpreting Slope & y-Intercept for Linear Models

Definition: Slope - describes rate of change in a function

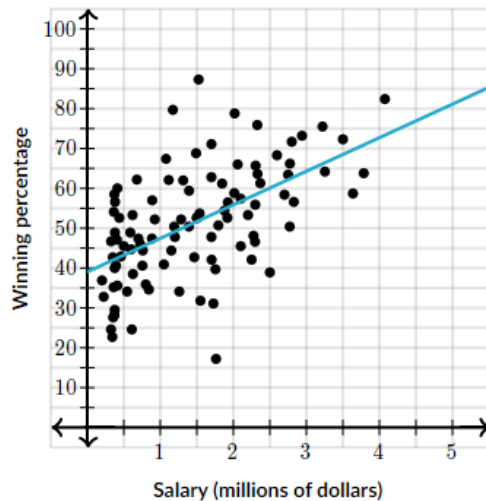
Definition: y-intercept - describes the initial (beginning) value



Interpreting Slope & y-Intercept for Linear Models

Example 1: Abigail gathered data on different schools' winning percentages and the average yearly salary of their head coaches (in millions of dollars) in the years 2000-2011. She then created the following scatterplot and regression line.

- The fitted line has a slope of 8.42.
- The fitted y-intercept is 40
- What is the best interpretation of this slope and y-intercept?

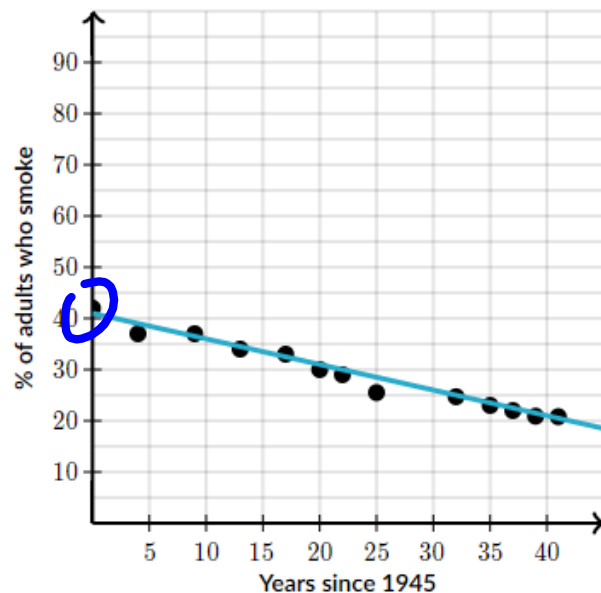


- y-int: if paid \$0 million winning % = 40%
- Slope: for every additional \$million up increase by 8.42

Interpreting Slope & y-Intercept for Linear Models

Example 2: The scatterplot and regression line below show the relationship between the percentage of American adults who smoke and years since 1945.

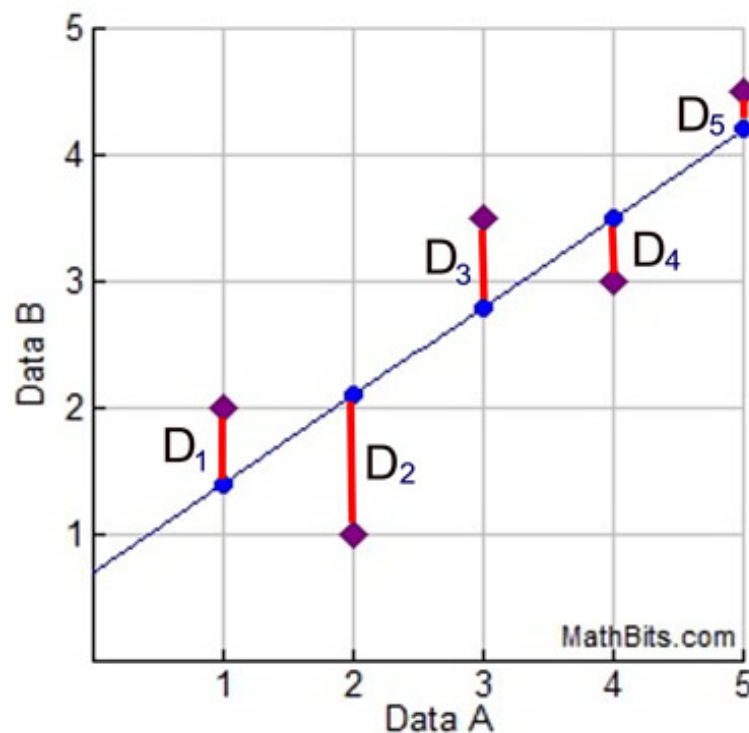
- The fitted line has a y-intercept of 41.
- What is the best interpretation of this y-intercept?



-In 1945, 41% of adults smoked.

SS_Residual Points

Definition: A residual is the vertical distance between a data point and the regression line. Each data point has one residual. They are positive if they are above the regression line and negative if they are below the regression line. If the regression line actually passes through the point, the residual at that point is zero.



◆ Scatter Plot Points:

$\{(1,2), (2,1), (3,3\frac{1}{2}), (4,3), (5,4)\}$

● Regression Points

$\{(1,1.4), (2,2.1), (3,2.8), (4,3.5), (5,4.2)\}$

The Red Line Segments:

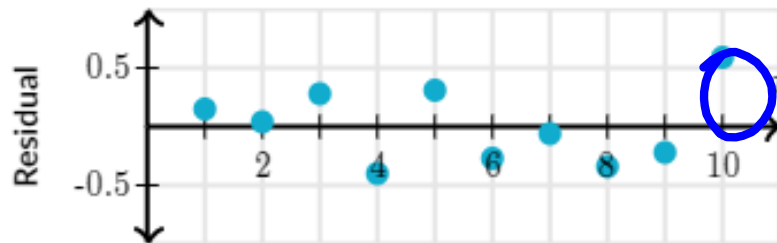
The red line segments represent the distances between the y-values of the actual scatter plot points, and the y-values of the regression equation at those points.

The lengths of the red line segments are called RESIDUALS.



SS_Residual Points

Example: The graph displays a residual plot that was constructed after running a least-squares regression on a set of bivariate numerical data (x,y) .

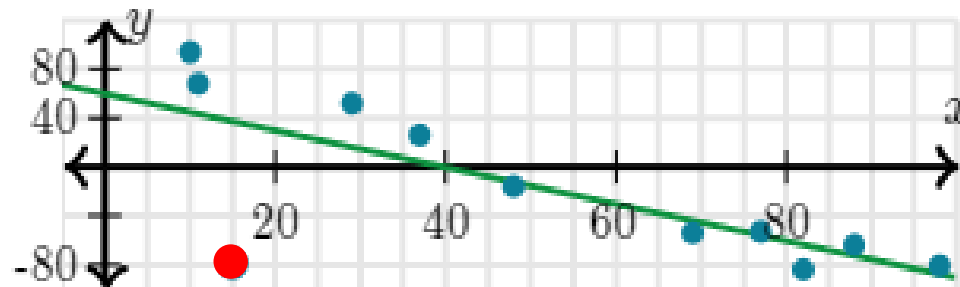



What can you conclude from this graph?

- This is a good model, because all of the residuals are close to the line $y=0$.
- The least squares regression equation overestimates y more often than it underestimates y .
- When $x=5$, the least squares regression equation underestimates y .

SS_Influential Points

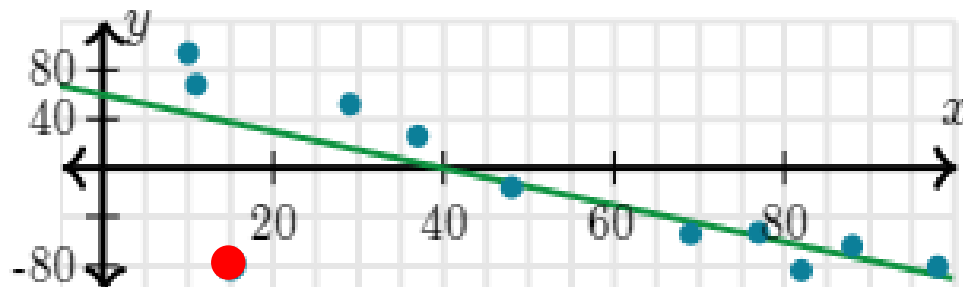
Definition: An influential point is an outlier that greatly affects the slope of the regression line.



Think of an influential point as an anchor  pulling the regression line down (or up).

What would happen if you remove the anchor?

SS_Influential Points

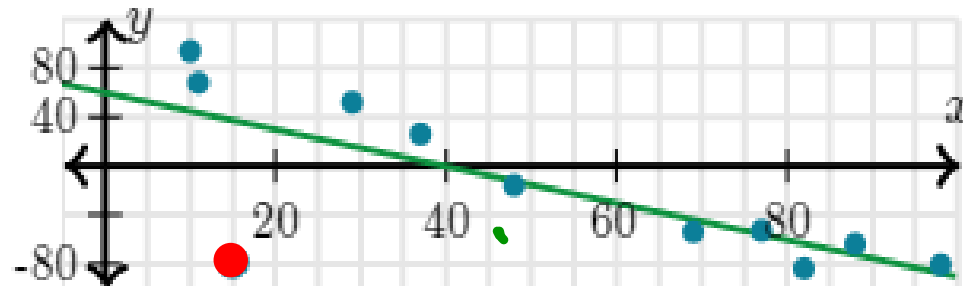


Other Terms to Know:

- Correlation coefficient (r)
- Coefficient of determination (r^2)
- Slope of the least-square regression line
- y-intercept of the least-square regression line

SS_Influential Points

Example: The scatterplot below displays a set of bivariate data along with its least-squares regression line.



Consider removing the outlier (15,-79) and calculating a new least-squares regression line.

What effect(s) would removing the outlier have?

- The coefficient of determination r^2 would increase. *yes*
- The correlation coefficient (r) would get closer to -1. *yes*
- The slope of the least-square regression line would increase. *NO*

Displaying and Comparing Quantitative Data

You should be working on the following skills:

1. Calculating and interpreting residuals
2. Calculating the equation of the least-square regression line
3. Interpreting slope and y-intercept for linear models
4. Residual points
5. Influential points

Attachments

Ztable.pdf