Section 8-2 Deviations, Residuals and the Correlation Coefficient

In Section 8.1 we analyzed data for y sit-ups that Moe could do x days after surgery.

x (days)	y (sit ups)	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$	$y - \bar{y}$	$(y - \bar{y})^2$
2	8	7.6	0.4	0.16		
4	10	11.8	-1.8	3.24		
6	19	16	3	9		
8	18	20.2	-2.2	4.84		
10	25	24.4	0.6	0.36		
Sums: 30	80		0	17.6		

We found the linear **regression equation** to be $\hat{y} = 2.1x + 3.4$

1. Suppose we want to estimate the number of sit-ups Moe could do, but we are not told the number of days. Our best estimate would be $\bar{y}$ (pronounced "y bar") , which is **the average of the y-values.**

2. Calculate y - $\bar{y}$ and add the column to the table above. Draw the horizontal line y = $\bar{y}$ on the scatterplot. Show the **deviation**, y - $\bar{y}$, of at least one point. (deviation is how far the data point is away from the y = $\bar{y}$ line)

When you **sum the squares of the deviations** this number is $SS_{dev} = \sum(y - \bar{y})^2$

3. Draw the regression line, $\hat{y} = 2.1x + 3.4$ on the scatterplot below. Show the **residual**, of at least one point. (residual is how far away the data point is from the regression line)

When you **sum the squares of the residuals**, this number is $SS_{res} = \sum(y - \hat{y})^2$

4. Calculate the **coefficient of determination**, $r^2 = \frac{SS_{dev} - SS_{res}}{SS_{dev}}$

This is the fraction of SS_{dev} that has been removed by the linear regression.

The number, **r**, is called the **correlation coefficient**. It indicates how *well* the best-fitting function fits the values. (use the positive square root if the slope of the line is positive, and use the negative square root if the slope is negative)

The closer the value of r is to 1 or -1, the stronger the fit of the data is to the function. r = 0 means no relationship.

Residuals

- Difference between y value of data point and y value of regression equation ($y - \hat{y}$)
- $\hat{y}$ = y hat
- See graph – discuss residual and square of residual (numerically and graphically)

Deviation

- Difference between y value of data point and y average ($y - \overline{y}$)
- $\overline{y}$ = y bar
- See graph – discuss deviation and square of deviation (numerically and graphically)

Basic definitions

- **Regression line:** the line that makes SS_{res} a minimum.

  (AKA – regression equation: y^ = ……)
- SS_{res}: sum of the squares of the residuals
- SS_{dev}: sum of the squares of the deviations
- Coefficient of determination: r^2
- $r^2 = \frac{SS_{dev} - SS_{res}}{SS_{dev}}$
- Correlation coefficient: r
  - (this is the # we will check to see which regression equation is the best fit)
  - See p. 328