

**Recall:** Correlation coefficient (linear relationship between  $x$ ,  $y$ ).

**OMIT:** Hypothesis testing for  $\rho$  in Sect. 10–3.

**Equation of Regression Line:** (p. 469)

$$y' = a + bx$$

where

$y'$ : estimated (predicted) value of  $y$

$a$ : y-intercept

$b$ : slope (amount by which  $y'$  changes if  $x$  increases by 1 unit).

1

3

## Regression

**Text:** Sect. 10–4

If a scatter plot suggests a linear relationship exists between 2 variables ( $x$  and  $y$ ), we then want to determine the equation of the **regression line** to describe the relationship between  $x$  and  $y$ .

The regression line is the **line of best fit** (p. 469): it is the line such that the sum of the squares of the errors in the  $y$  direction from each point to the line is as small as possible.

The line of best fit is also called the **least squares regression line**.

We don't know  $a$ ,  $b$ : need to estimate them from data to give us the line of best fit.

2

4

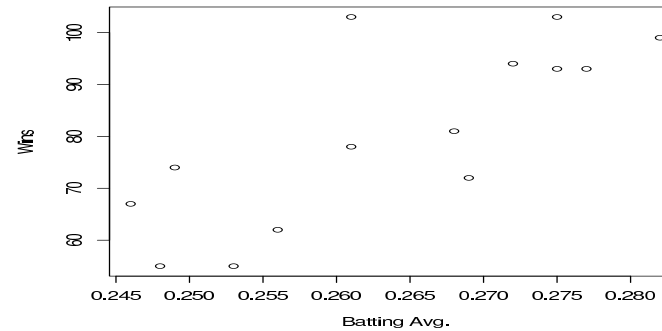
**Calculation of  $a$  and  $b$  (p. 470)**

$$b = \frac{SS_{xy}}{SS_{xx}}$$

$$a = \bar{y} - b\bar{x}$$

**NOTE:** The text writes these equations for  $a$  and  $b$  differently, but they will give the same answers to what is written above.

Add line of best fit to plot:



**EX. 10.1** (cont'd): Find the line of best fit that predicts wins from batting average.

avg. (x): 0.275 0.261 ... 0.261  
 wins (y): 103 78 ... 103

**Prediction and Residuals**

We can use our regression line to predict what happens at a particular  $x$  value.

**EX 10.1** (cont'd). Predict the number of wins for a team with a 0.266 batting average.

**EX 10.1** (cont'd). Find the residual for the number of wins the NY Yankees had in 2002.

**WARNING:** Be careful of **extrapolation**: making predictions beyond the range of the data (p. 473).

9

11

**Residual:** (observed response) – (predicted response), or

$$e = y - y'$$

**RULE:** If we find the residual  $e$  corresponding to each observation  $y$ , then

$$\sum e = 0$$

10