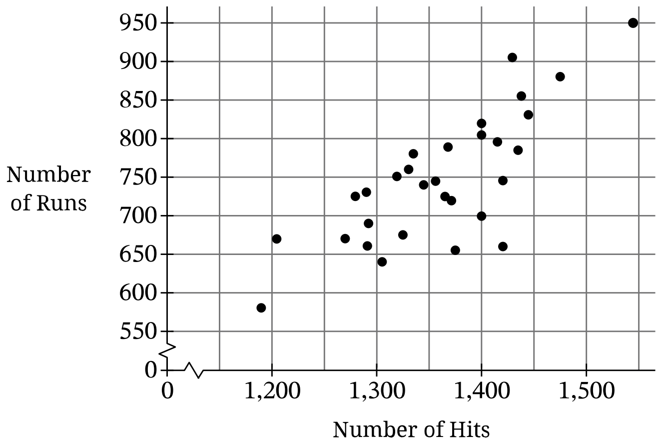
The following information applies to parts A and B.

6. To win a game, a baseball team needs to score runs. To score runs, players need to get on base. The primary way players get on base is by hitting the ball. Figure 1 shows the relationship between the number of hits and the number of runs for 30 randomly selected professional baseball teams.

Figure 1. Scatterplot of Number of Runs by Number of Hits



Part A

Consider the scatterplot in Figure 1.

i. Describe the relationship between the number of hits and the number of runs for these baseball teams, in context.

ii. Based on the data from the sample of 30 teams, an equation of the least-squares regression line model for predicting the number of runs from the number of hits is as follows.

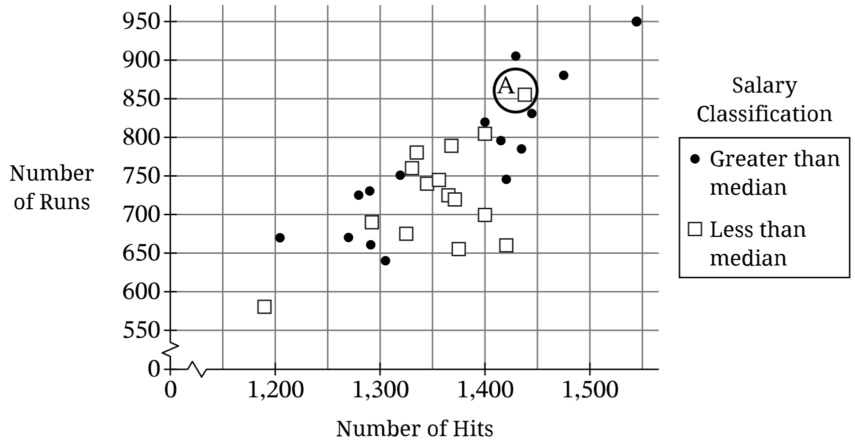
$$\text{Predicted number of runs} = -372.2 + 0.823\,(\text{number of hits})$$

Using the given regression equation, calculate the point estimate for the predicted number of runs a team would score if they were to achieve their goal of 1,250 hits. Show your work.

Part B

Each team has a total salary equal to the sum of the salaries of all players on the team. The median total salary for the sample of 30 teams is $160 million. In Figure 2, points with dots represent teams with a total salary greater than the median, and points with squares represent teams with a total salary less than the median.

Figure 2. Scatterplot of Number of Runs by Number of Hits and Salary Classification



Consider the scatterplot in Figure 2.

i. Compare the team represented by the point that is circled and labeled A with the other teams that have the same total salary classification.

ii. For each salary classification, consider the linear relationship between the number of hits and the number of runs. For teams with a total salary greater than the median, is the strength of the linear relationship stronger than, weaker than, or similar to the strength of the linear relationship for teams with a total salary less than the median? Explain.

Part C

Different types of intervals can be used when working with linear regression models. One type of interval is a confidence interval for the slope of the least-squares regression line. Two other types of intervals in the context of this problem are as follows:

- A confidence interval that estimates the mean number of runs for all teams with a specific number of hits.
- A prediction interval that predicts the number of runs for a single team with a specific number of hits.

All three intervals use the same basic formula:

$$\text{Point Estimate} \pm t^{*}(\text{standard error}),$$

where critical value t^* comes from a t-distribution that has $n-2$ degrees of freedom. The same critical value is used when finding each of these intervals with the same level of confidence.

Consider the point estimate for the predicted number of runs found in part A (ii) for the team whose goal is to achieve 1,250 hits next year. Recall that this value was calculated using the linear regression model based on the data from the sample of 30 teams.

i. What is the critical value for a 95% confidence level that would be used for the confidence interval for the mean number of runs as well as for the prediction interval for the number of runs? Indicate the answer to two decimal places.

ii. The standard error for the confidence interval for the mean number of runs is 17.48. Assuming the conditions for inference are met, calculate the 95% confidence interval for the mean number of runs for all teams with 1,250 hits. Show your work.

iii. The standard error for the prediction interval for the number of runs is 56.78. Assuming the conditions for inference are met, calculate the 95% prediction interval for the number of runs for a single team with 1,250 hits. Show your work.

Part D

i. Would a distribution of sample means be expected to have more variability or less variability than a distribution of individual observations? Explain.

ii. The standard error used in the confidence interval that estimates the mean number of runs for all teams with x hits can be found using the following formula.

$$\sqrt{s^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

The standard error used in the prediction interval that predicts the number of runs for a single team with x hits can be found using the following formula.

$$\sqrt{s^2+s^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

In both standard error formulas, s is the standard deviation of the residuals, n is the sample size, and $\bar{x}$ is the sample mean number of hits. Based on the answer from part D (i) and the standard error formulas for the confidence interval and the prediction interval, explain why the prediction interval calculated in part C (iii) is wider than the confidence interval calculated in part C (ii).

STOP

END OF EXAM