

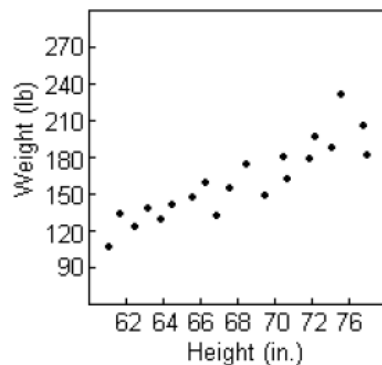
Statistics
Test 4 Review

Major concepts to remember:

- Association and Correlation
- Association and/or correlation do NOT imply causation
- Correlation coefficient r ranges from -1 (perfectly linear with negative slope) to 0 (no correlation) to 1 (perfectly linear with positive slope)
- Positive correlation means that an increase in explanatory variable predicts an increase in response variable; negative correlation means that an increase in explanatory variable predicts a decrease in response variable
- Correlation coefficient r is unitless—it does not change if all the x 's or y 's are multiplied by a constant and/or increased by a constant. It does not change if the x 's and y 's switch places.
- The Least Squares Regression line is a model—it does not describe individual behavior, but rather a general trend.
- The slope of the Least Squares Regression line = $r \frac{s_y}{s_x}$. This is because r is the slope of the least squares regression line of the z-scores.
- The coefficient of determination: r^2 . It's the portion of the variation of the response variable that is explained by the variation in the explanatory variable. When $r^2 = 1$, the variation in the response variable is explained 100% by the explanatory variable, which means response variable can be predicted exactly. When $r^2 = .5$, then half of the variation in the response variable can be explained, but the other half is still unknown. And when $r^2 = 0$, the variation in the response variable is completely unpredictable.
- Residuals: observed values – predicted values. They should appear random, without pattern. Negative residual means that the observed value is lower than predicted; positive residual means the observed value is higher than predicted. If there is a pattern to the residuals, then a different (non-linear) model is probably better. In that case, then you might be able to interpolate (if r^2 is sufficiently high), but you absolutely should not extrapolate.

Problems:

1. Data collected from students in Statistics classes included their heights (in inches) and weights (in pounds). For the students' heights and weights, the correlation is 0.636. Suppose the variable weight is recorded in kilograms rather than in pounds. What will be the correlation?



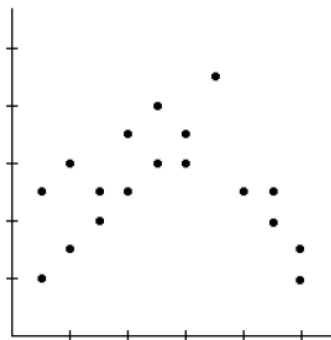
2. The paired data below consist of the test scores of 6 randomly selected students and the number of hours they studied for the test. What is the line of best fit that predicts the scores from the number of hours that they studied? What is the correlation coefficient?

Hours	Score
5	64
10	86
4	69
6	86
10	59
9	87

3. A science instructor assigns a group of students to investigate the linear relationship between the pH of the water of a river and its water's hardness (measured in grains). Some students wrote these conclusions: "My correlation of -0.94 shows that there is almost no association between pH of the water and water's hardness." Is the interpretation of the correlation appropriate?
4. Find the Least Squares Regression Line:

\bar{x}	s_x	\bar{y}	s_y	r	$\hat{y} = b_0 + b_1x$
40	20	8	11	8	$\hat{y} = ?$

5. Tell what the residual plot indicates about the appropriateness of the linear model that was fit to the data.



6. A random sample of records of electricity usage of homes in the month of July gives the amount of electricity used and size (in square feet) of 135 homes. A regression was done to predict the amount of electricity used (in kilowatt-hours) from size. The residuals plot indicated that a linear model is appropriate. The model is:

$$\widehat{usage} = 1204 + 0.6 \text{ size}.$$

How much electricity would you predict would be used in a house that is 2273 square feet?

7. Suppose the correlation between SAT Verbal scores and Math scores is 0.5 and that these scores are normally distributed. If a student's Verbal score places her at the 97.5th percentile, at what percentile would you predict her Math score to be?

8. The correlation coefficient between high school grade point average (GPA) and college GPA is 0.560. For a student with a high school GPA that is 2.5 standard deviations above the mean, would we expect that student to have a college GPA that is above or below the mean?
9. Which statement about residuals plot is true?
- A curved pattern indicates nonlinear association between the variables.
 - A pattern of increasing spread indicates the predicted values become less reliable as the explanatory variable increases.
 - Randomness in the residuals indicates the model will predict accurately.
10. The relationship between the number of games won by a minor league baseball team and the average attendance at their home games is analyzed. A regression to predict the average attendance from the number of games won has an $r^2 = 32.0\%$. The residuals plot indicated that a linear model is appropriate. Write a sentence summarizing what the coefficient of determination r^2 says about this regression. What is the correlation between average attendance and number of games won?
11. A data set has per capita numbers of cigarettes smoked (sold) by 43 states and the District of Columbia in 1960 together with death rates per thousand population from lung cancer. A positive association is observed, and the residuals plot indicates that a linear model is appropriate. The following statistics are calculated:
- x = per capita numbers of cigarettes sold
 y = death rates per thousand from lung cancer
 $\bar{x} = 24.91$
 $\bar{y} = 19.65$
 $s_x = 5.51$
 $s_y = 4.18$
 $r^2 = 0.486$
- How good is this model?
 - Find the Least Squared Regression line, and tell what the slope of the line means in this particular context.
 - Find the residuals for Texas (who had 23.57 cigarettes sold per capita and 20.74 deaths per thousand from lung cancer) and Nevada (who had 42.40 cigarettes sold per capita and 23.03 deaths per thousand from lung cancer).