Lesson 21: Coefficient of Determination r^2

Daily Data Collection

Select two topics you think are correlated, make a hypothesis,

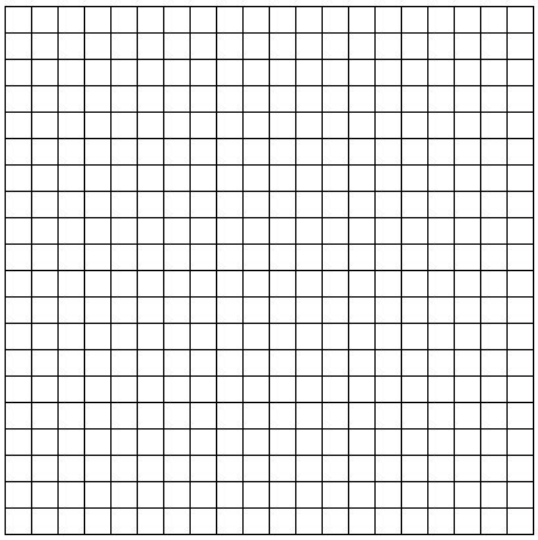
and run a test to see if your assumptions were true.

Class Data:

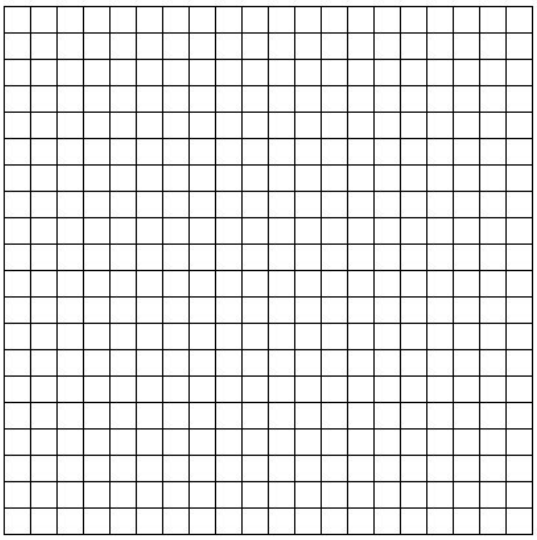
Explanatory Variable:

Response variable:

Create a scatterplot.



Create a residual plot.



Describe the Direction, Form, and Strength

Write an equation for the regression line

Describe the slope in the context of the situation

Find the residual value for your own data point

Describe the meaning of the r^2 value:

R^2 Coefficient of Determination *- How good is our prediction?*

The strength of a prediction which uses the LSRL depends on how close the data points are to the regression line. The mathematical approach to describing this strength is via the **coefficient of determination.** The coefficient of determination gives us *the proportion of variation in the values of y that is explained by least-squares regression of y on x*. The coefficient of determination turns out to be r^2 (correlation coefficient squared).

Whenever you use the regression line for prediction, also include r^2 as a measure of how successful the regression is in explaining the response.

In our example r^2 =0.8569. This means that over 85% of the variation in doctor visits per year can be explained by the linear relationship it has with cigarettes smoked per week.

AP Tip: Generalized way to describe r^2:

"_____% of the variation in _____ is accounted for by the regression line."
r^2 response variable (y)

Lurking variable:
A variable not shown in the regression equation that affects both x and y. Many times, there is no association between x and y, but it appears there is an association because they are both affected by the third variable (the lurking variable). This is called a nonsense correlation.

Example: studies show a strong correlation between the number of cars owned (x) and the length of life (y) of the owner.

Remember: Correlation does NOT imply causation. AKA Association does NOT imply causation.

CHECK YOUR UNDERSTANDING

Multiple choice: Select the best answer.

1. For the least-squares regression of fat gain on NEA, $r^2 = 0.606$. Which of the following gives a correct interpretation of this value in context?

(a) 60.6% of the points lie on the least-squares regression line.
(b) 60.6% of the fat gain values are accounted for by the least-squares line.
(c) 60.6% of the variation in fat gain is accounted for by the least-squares line.
(d) 77.8% of the variation in fat gain is accounted for by the least-squares line.

2. A recent study discovered that the correlation between the age at which an infant first speaks and the child's score on an IQ test given upon entering elementary school is −0.68. A scatterplot of the data shows a linear form. Which of the following statements about this finding is correct?

(a) Infants who speak at very early ages will have higher IQ scores by the beginning of elementary school than those who begin to speak later.
(b) 68% of the variation in IQ test scores is explained by the least-squares regression of age at first spoken word and IQ score.
(c) Encouraging infants to speak before they are ready can have a detrimental effect later in life, as evidenced by their lower IQ scores.
(d) There is a moderately strong, negative linear relationship between age at first spoken word and later IQ test score for the individuals in this study.

Facts about least-squares regression

Fact 1. The distinction between explanatory and response variables is essential in regression. Least-squares regression looks at the distances of the data points from the line only in the *y* direction. If we reverse the roles of the two variables, we get different least-squares regression line.

Fact 2. There is a close connection between correlation and the slope of the LSRL.
The slope is : $\boxed{b = r\frac{s_y}{s_x}}$

This equation says that along the regression line, a change in one standard deviation in x corresponds to a change of *r* standard deviations in *y*.

Fact 3. The LSRL always passes through the point $(\bar{x}, \bar{y})$

Fact 4. The correlation *r* describes the strength of a straight-line relationship. In the regression setting, this description takes a specific form: the square of the correlation, r^2, is the fraction of the variation in the values of *y* that is explained by the least-squares regression of *y* on x.

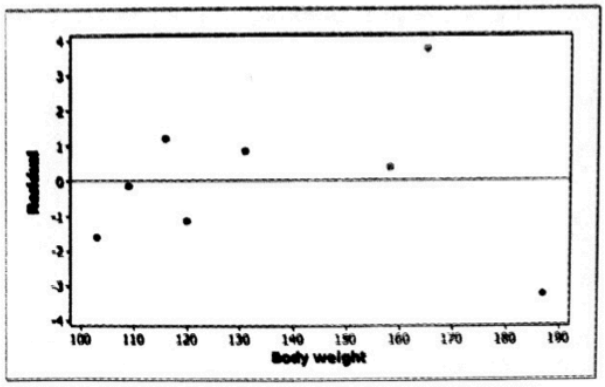
Fact 5. R-sq has no units

SEE = Standard Error of Estimate

- The estimate is y hat
- The error is the residual
- Standard means Standard Deviation
- So SSE is the standard deviation of the residuals.
- The Letter S = tells this on the computer printout (near R-sq)
- On the calculator: Run the AP program, select 4 Lin Reg. then run "residual scatterplot"

CHECK YOUR UNDERSTANDING

In the Check Your Understanding on page 171, we asked you to perform least-squares regression on the familiar hiker data shown in the table below. The graph shown is a residual plot for the least-squares regression of pack weight on body weight for the 8 hikers.



Body weight (lb):	120	187	109	103	131	165	158	116
Backpack weight (lb):	26	30	26	24	29	35	31	28

1. The residual plot does not show a random scatter. Describe the pattern you see.
2. For this regression, $s = 2.27$. Interpret this value in context.

HW 21 – Section 3-2: 48, 49, 50, 57, 58