Lesson 19: Least-Squares Regression

Daily Data Collection

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

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

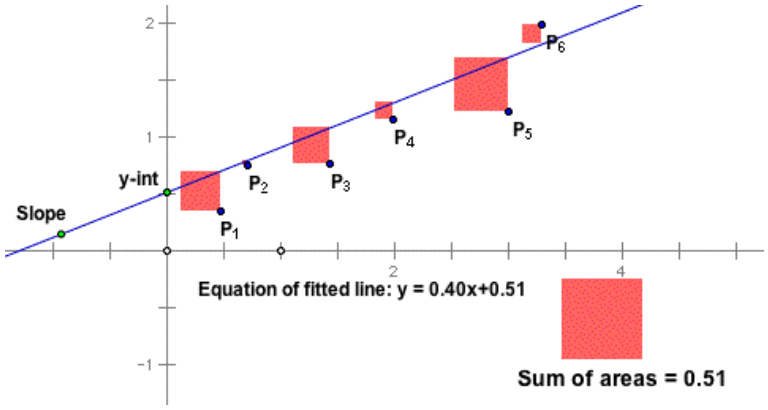
Least Squares Regression A.K.A. linear regression allows you to fit a line to a scatter diagram in order to be able to predict what the value of one variable will be based on the value of another variable.

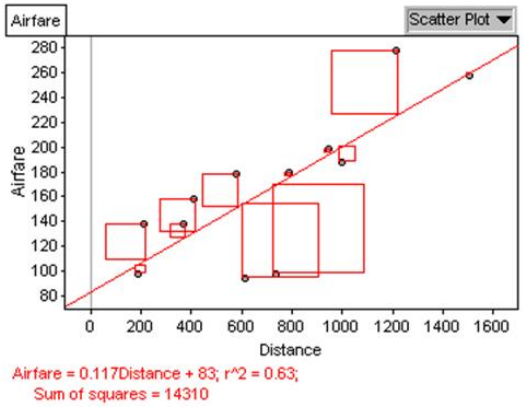
The fitted line is called the *line of best fit, linear regression line, or least squares regression line*, (LSRL) and has the form $\hat{y} = a + bx$ where:

a: y intercept

b: slope of the line Note: sometimes this is listed as $\hat{y} = b_0 + bx$

The way the line is fitted to the data is through a process called the *method of least squares*. The main idea behind this method is that the square of the **vertical** distance between each data point and the line is minimized.





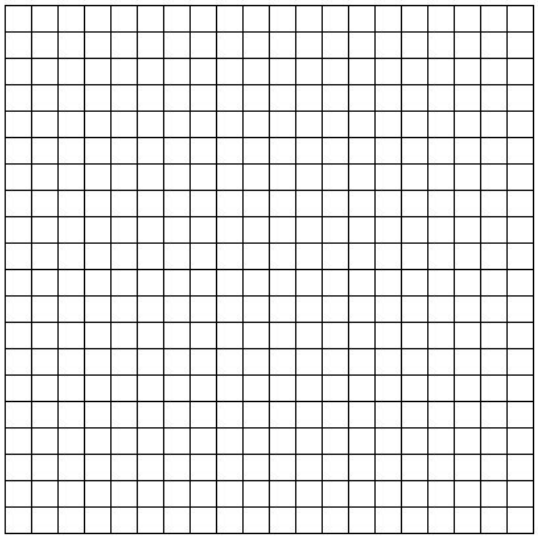
The least squares regression line is a mathematical model for the data that helps us **predict** values of the response (dependant) variable from the explanatory (independent) variable. Therefore, with regression, unlike with correlation, we must specify which is the response and which is the explanatory variable.

Class Data:

Create a scatterplot.
Explanatory Variable:

Response variable:

Use a spreadsheet!



Describe the Direction

Describe the Form

Describe the Strength, including the correlation

Write an equation for the regression line

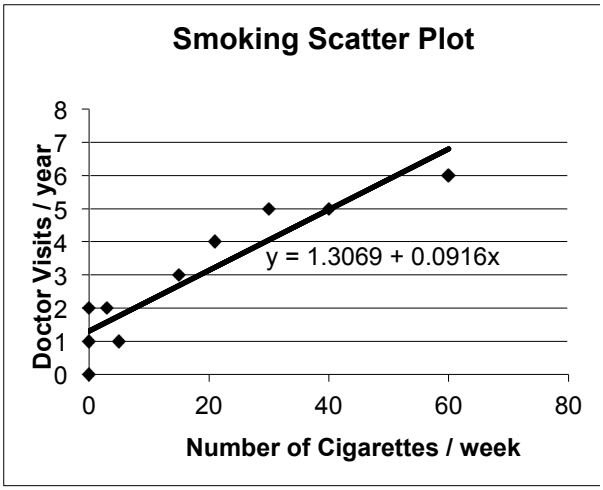
Describe the slope in the context of the situation

Conclusion/Analysis

Example:

# of Cigarettes Per Week	0	3	21	15	30	5	40	60	0	0
Number of doctor visits per year	1	2	4	3	5	1	5	6	2	0

We will use the calculator to find regression lines. The meaning of the line y = 1.3 + .09x is described below:



Slope is .09.
This means that as # of cigarettes per week increases by 1, the number of doctor visits per year will increase by .09.

The y-intercept is 1.3
This means that if a person smoked 0 cigarettes per week, then we would expect them to visit the doctor 1.3 times per year.

Y hat is a prediction of y based on x.
If x = 40 cigarettes per week, then we would expect 1.3+.09*40 = 4.9 doctor visits per year.

Extrapolation: using the regression line for x values beyond the data used to create the regression line. This is considered dangerous and is to be avoided.
Example the regression line was created for x values 0 to 60. Avoid using x values over 60.

CHECK YOUR UNDERSTANDING

Some data were collected on the weight of a male white laboratory rat for the first 25 weeks after its birth. A scatterplot of the weight (in grams) and time since birth (in weeks) shows a fairly strong, positive linear relationship. The linear regression equation $\widehat{weight} = 100 + 40(time)$ models the data fairly well.

1. What is the slope of the regression line? Explain what it means in context.
2. What's the *y* intercept? Explain what it means in context.
3. Predict the rat's weight after 16 weeks. Show your work.
4. Should you use this line to predict the rat's weight at age 2 years? Use the equation to make the prediction and think about the reasonableness of the result. (There are 454 grams in a pound.)

HW 19 – Section 3-2: 35, 36, 37, 39, 41 **Study for Quiz 6 over lessons 17-19, r11**