3.2 Regression

Now that we have a scatterplot to visualize our data correlation, we are going to describe the data further using **least-squares regression** by fitting a line to the data.

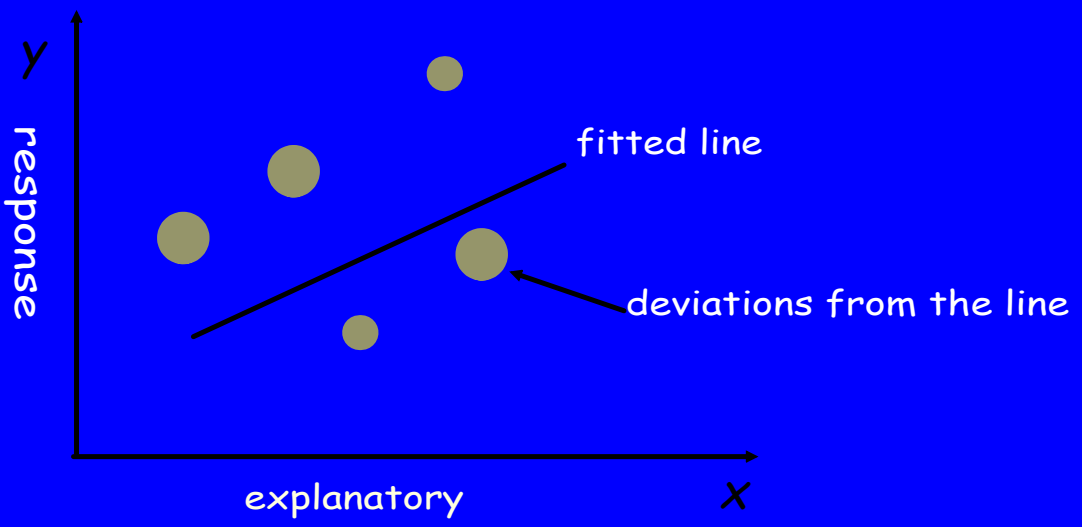
Regression can be used to make predictions. Take a value for x and substitute it into the equation then you'll get the value for $\hat{y}$ we call this value y-hat! It represents the predicted value.

EXTRAPOLATING - using the regression line to predict a value OUTSIDE the range of the data (you cannot do this)

INTERPOLATION - using the regression line to predict a value inside the range of data (you can do this)

The LSRL is the line that makes the sum of the squares of the deviations of the data points from the line, in the vertical direction, as small as possible.

it makes prediction errors very small.



The line is defined by the slope and y-intercept: $\hat{y}$ = a + bx where a is the y-intercept and b is the slope.

The slope of the line describes the rate of change in the response variable y as the explanatory variable x changes.

We can calculate a and b using formulas from information we are given:

$$b = r\frac{S_y}{S_x} \qquad a = \bar{y} - b\bar{x}$$

Procedure for regression:

- Determine explanatory and response variables - make a scatterplot to determine linearity
- Find the correlation coefficient r and the equation of the least-squares regression line. Use the calculator to obtain r,a and b. Write the equation in y hat = a + bx form using the context of the problem.
- Draw the regression line on the plot choose 2 or 3 values for x and solve for the corresponding y-hat values: plot the points and connect with a line

- Interpret the slope, y-intercept, r and r^2 in the context of the problem.

- Use the equation to make predictions.

Note: let's discuss r vs r^2 for a moment!

r: represents the strength of the line as we saw last week.

r^2: is the fraction of the data VARIATION accounted for in the model. It helps to determine whether the LSRL should be used for predictions. How successful is the regression in explaining the response?

Use this data for the next slide:

Low fat and low carb:

here are carb contents in grams and fat content in grams for

9 different types of hamburgers at McDonalds

Type	Carbs (g)	Fat (g)
Hamburger	31	9
Cheeseburger	33	12
Double Cheeseburger	34	23
Quarter Pounder	37	19
Quarter Lber w Cheese	40	26
Double Q Lber w Cheese	40	42
Big Mac	45	29
Big N tasty	37	24
Big N tasty with cheese	38	28

Calculator procedure for regression:

1. Put the data into two lists. Graph. Turn stat plot on and hit zoom 9. Check to see that the graph is approximately linear with no extreme outliers.

2. Use LinReg(a+bx) to interpret r and r^2

3. Interpret r and r^2 IN THE CONTEXT OF THE PROBLEM.

r - positive or negative, weak or strong

r^2- percent of the variation of __(%)_____ is explained by the approximate linear relationship between __expl__ & __resp__.

4. Use Stat-> 'calc' ->
LinReg(a+bx) L1, L2, Y1(VARS->Y-VARS->Function) to graph the LSRL.

5. Use the equation to make any predictions. Choose one value within the range of x-values and one outside the range of x-values. Which do you have more confidence in and why?

37. **Gas mileage** We expect a car's highway gas mileage to be related to its city gas mileage. Data for all 1198 vehicles in the government's *2008 Fuel Economy Guide* give the regression line predicted highway mpg = 4.62 + 1.109 (city mpg).

(a) What's the slope of this line? Interpret this value in context.

(b) What's the intercept? Explain why the value of the intercept is not statistically meaningful.

(c) Find the predicted highway mileage for a car that gets 16 miles per gallon in the city. Do the same for a car with city mileage 28 mpg.