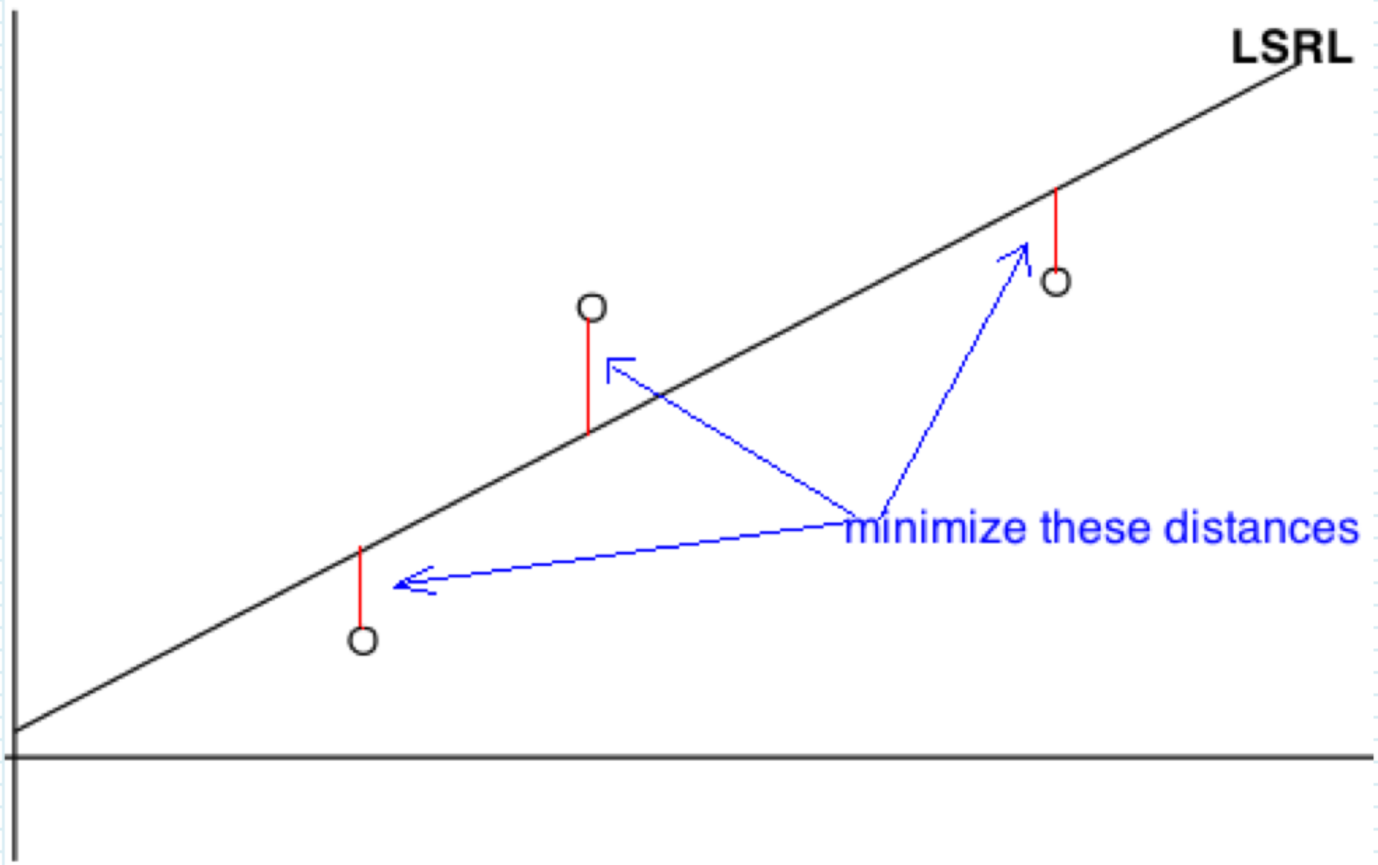
Least Squares Regression Line

- A **regression line** is a line that describes the relationship between the explanatory variable x and the response variable y.
- Regression lines can be used to predict a value for y given a value of x.
- The **least squares regression line** (or LSRL) is a mathematical model used to represent data that has a linear relationship.

- We want a regression line that makes the vertical distances of the points in a scatter plot from the line as small as possible.



- The least squares regression line formula is $\hat{y} = a + bx$
- The slope, *b*, is calculated with the formula

$$b = r\left(\frac{s_y}{s_x}\right)$$

- And the *y*-intercept is $a = \bar{y} - b\bar{x}$

- Notice that the formula for slope is

$$b = r\left(\frac{s_y}{s_x}\right)$$

- This means that a change in one standard deviation in x corresponds to a change of r standard deviations in y.
- In other words, we can say that on average, for each unit increase in x, then is an increase (or decrease if slope is negative) of $|b|$ units in y.

5.3 LSRL

Example:

1. Suppose we want to know if there is an association between the number of spaces a property is from GO and the cost of the property in a monopoly game.

Compute the LSRL for the data.

$\hat{y} = 67.28 + 6.78x$

Property	Spaces from GO	Cost
Mediterranean Avenue	1	60
Baltic Avenue	3	60
Reading Railroad	5	200
Oriental Avenue	6	100
Vermont Avenue	8	100
Connecticut Avenue	9	120
St. Charles Place	11	140
Electric Company	12	150
States Avenue	13	140
Virginia Avenue	14	160
Penn Railroad	15	200
St. James Place	16	180
Tennessee Avenue	18	180
New York Avenue	19	200
Kentucky Avenue	21	220
Indiana Avenue	23	220
Illinois Avenue	24	240
B & O Railroad	25	200
Atlantic Avenue	26	260
Ventnor Avenue	27	260
Water Works	28	150
Marvin Gardens	29	280
Pacific Avenue	31	300
North Carolina Avenue	32	300
Pennsylvania Avenue	34	320
Short Line Railroad	35	200
Park Place	37	350
Boardwalk	39	400

- The LSRL can be used to predict values of y given values of x.
- We need to be careful when predicting. When we are estimating y based on values of x that are much larger or much smaller than the rest of the data, this is called **extrapolation.**

Example:

2. Use the LSRL found in example 1 to predict the cost of a property that is 50 spaces from GO.

$$x = 50$$

$$\hat{y} = 67.28 + 6.78x$$

$$\hat{y}(50) = 67.28 + 6.78(50)$$

$$\hat{y}(50) = 406.51$$

- The square of the correlation (r), is called the **coefficient of determination.**
- It is the fraction of the variation in the values of y that is explained by the regression line and the explanatory variable.
- When asked to interpret r^2 we say, "approximately $r^2(100)\%$ of the variation in y is explained by the LSRL of y on x."

- Facts about the coefficient of determination:
  - The coefficient of determination is obtained by squaring the value of the correlation coefficient.
  - The symbol used is r^2
  - Note that $0 \leq r^2 \leq 1$
  - r^2 values close to 1 would imply that the model is explaining most of the variation in the dependent variable and may be a very useful model.
  - r^2 values close to 0 would imply that the model is explaining little of the variation in the dependent variable and may not be a useful model.

Example:

3. The following 9 observations compare the Quetelet index, x (a measure of body build) and dietary energy density, y.

x	221	228	223	211	231	215	224	233	268
y	.67	.86	.78	.54	.91	.44	.9	.94	.93

outlier?

a. Make a scatter-plot of the data.

b. Compute the LSRL. $\hat{y} = -4.266 + .0225x$ ← after removed pt (268, .93)

c. Provide an interpretation of the slope of this line in the context of these data. For every one increase in the Quetelet index, there is an increase of 0.0225 units in dietary energy density.

d. Find the correlation coefficient for the relationship. Interpret this number. $r = .9131$

e. Find the coefficient of determination for the relationship. Interpret this number. $r^2 = .8337$