# MATH 3307
# Lesson 19

# Regression Lines

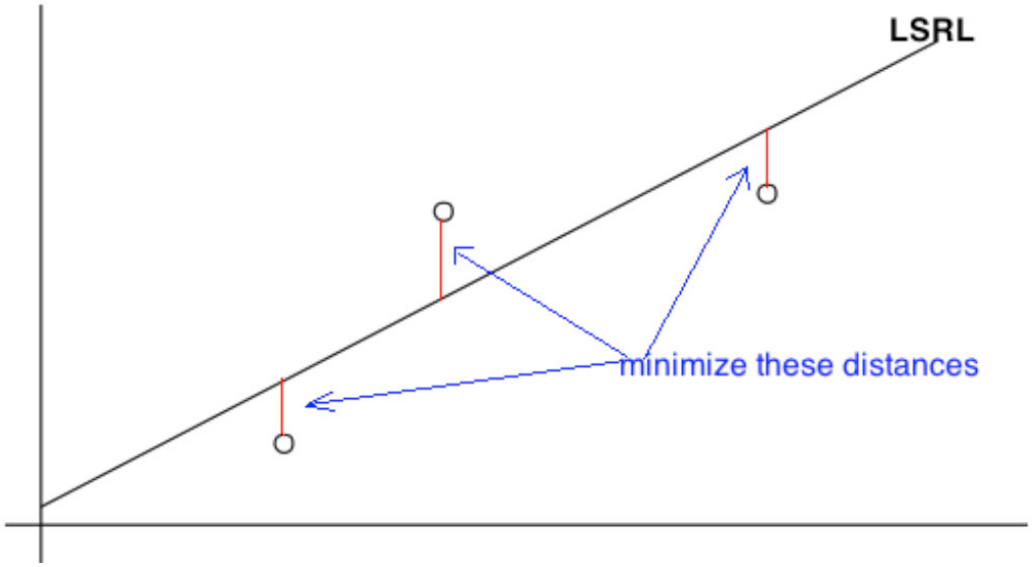
A **regression line** is a line that describes the relationship between the explanatory variable *x* and the response variable *y*.

Regression lines can be used to predict a value for *y* given a value of *x*.

This is a straight line the crosses through the center field of your scatter plot. (hits as many points as possible, equal number of missed points above and below)

# Least Squares Regression Lines (LSRL)

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.




Note: To calculate this by hand, you are going to use optimization techniques from Calculus to minimize the distance between a point (x,y) from your scatter plot, and the line, y = mx + b by minimizing the distance formula:

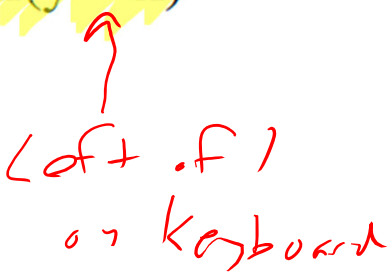
$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

# Calculating a Least Squares Regression Line

The least squares regression line formula is $\hat{y} = a + bx$

The slope, $b$ is calculated using $b = r\left(\frac{s_y}{s_x}\right)$ and the $y$-intercept is $a = \bar{y} - b\bar{x}$.

To calculate the values of $a$ and b for the regression line with R-Studio, we use the command lm(y ~ x)



# Example:

Using the Monopoly Problem, Calculate the Regression Line:

optimal

regline=lm(cost~spaces)

regline

```
> plot(spaces,cost)
> regline=lm(cost~spaces)
> regline

Call:
lm(formula = cost ~ spaces)

Coefficients:
(Intercept)       spaces
     67.283        6.784
```

y-int → 67.283 ; 6.784 ← slope

<<This will give you the information about the linear equation>>

$y = mx + b$

$\hat{y} = 6.784x + 67.283$ or $\hat{y} = 67.283 + 6.784x$

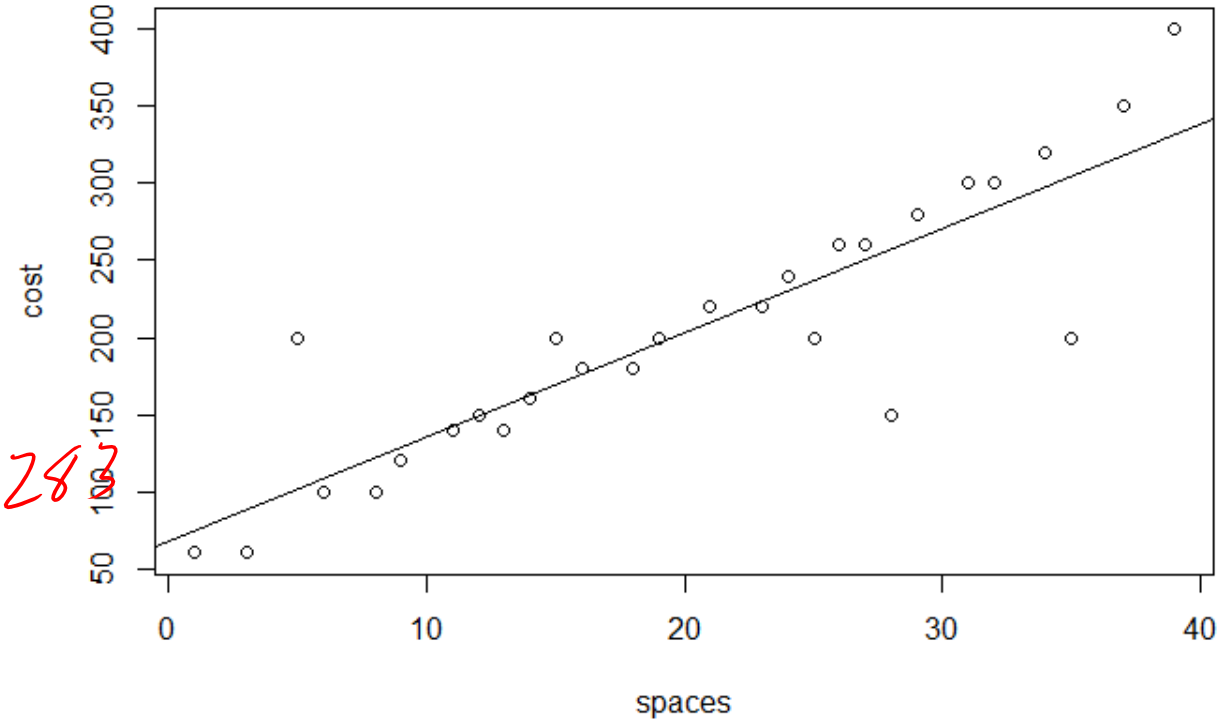
# Viewing the Scatterplot and the Regression Line

Note that I assigned a name to the lm command, this is not required unless you wish to use it again. We will use it again to plot the regression line on top of the scatterplot. The command is abline.

```
abline(regline)
```

```
y=ax+b
a=6.784462068
b=67.28274214
r²=.7708376438
r=.8779736009
```

$y = 6.784x + 67.28$



# Making Predictions:

The LSRL can be used to predict values of *y* given values of *x*.

Let's use our model to predict the cost of a property 50 spaces from GO

$\hat{y} = 6.784x + 67.283$ → Plug in for x

$\hat{y} = 6.784(50) + 67.283$

```
6.784*50+67.283
[] 406.483
```

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.**

Our given x-values were 0 to 39. The farther away from 39 that you predict, the more inaccuracy is present

# Interpreting the Slope

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*. This means that on average, for each unit increase in *x*, then is an increase (or decrease if slope is negative) of |*b*| units in *y*.

Interpret the meaning of the slope for the Monopoly example

There is a [slope value] [increase/decrease] in [y-unit] for every one increase in your [x-unit]
m = 6.784
There is a $6.78 increase in cost for every increase of 1 space from GO.

# Interpreting the Slope

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*. This means that on average, for each unit increase in *x*, then is an increase (or decrease if slope is negative) of |*b*| units in *y*.

Interpret the meaning of the slope for the Monopoly example:

For every increase of 1 space from go, there is an increase of $6.79 of cost.

# Coefficient of Determination

The square of the correlation ($r$), $r^2$ 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 $x$."

This tells how accurate the measurement is based on the regression line.

# Facts about the coefficient of determination:

1. The coefficient of determination is obtained by squaring the value of the correlation coefficient.

2. The symbol used is $r^2$

3. Note that $0 \leq r^2 \leq 1$

4. $r^2$ 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.*

5. $r^2$ 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.*

# Interpret $r^2$ for the Monopoly problem

```
> cor(spaces,cost)
[1] 0.8779736
> cor(spaces,cost)^2
[1] 0.7708376
```

$r = .8779$

Correlation Coefficient

$r^2 = .7708$

Coefficient of Determination

Interpretation:
77% of our variation in cost is explained by the LSRL

Popper 14: 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 |

1. Compute the LSRL

   a. y = -.8985x + .0073  b. y = .0073x - .8985
   c. y = .0073x + .8985  d. y = .8985x - .0073

2. Find the Correlation Coefficient

   a. .6579  b. .7325  c. .9231  d. .0607

3. Find the coefficient of determination

   a. .8111  b. .7834  c. .0023  d. .4328

```
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
   -0.89846      0.00733
```

$\hat{y} = .0073x - .8985$

```
> cor(x,y)
[1] 0.6578901
> cor(x,y)^2
[1] 0.4328194
```

Interpret the (4) slope of the graph, and the (5) coefficient of determination

# To Copy into RStudio

```
assign("x",c(221,228,223,211,231,215,224,233,268))

assign("y",c(.67,.86,.78,.54,.91,.44,.9,.94,.93))
```

# Popper 14, Continued

4. Interpretation of the slope

a. For every 0.00733 units up move 1 unit to the right.

b. There is an increase of 0.00733 unit of dietary energy density for every unit increase of body build.

c. There is a decrease of 0.00733 unit of dietary energy density for every unit increase of body build.

d. The slope is an abstract measure with no interpretation possible

5. Interpretation of the coefficient of determination

a. Since this measure is positive, there is a positive relationship between the variables.

b. Since this measure is close to 0.5, the LSRL is a reasonable approximation of the data.

c. Roughly, 43% of the variation in dietary energy density is explained by the LSRL

d. Roughly, 43% of the data points lie on or near the LSRL