Math 2311

Bekki George – bekki@math.uh.edu

Office Hours: MW 11am to 12:45pm in 639 PGH

Online Thursdays 4-5:30pm

And by appointment

Class webpage: http://www.math.uh.edu/~bekki/Math2311.html

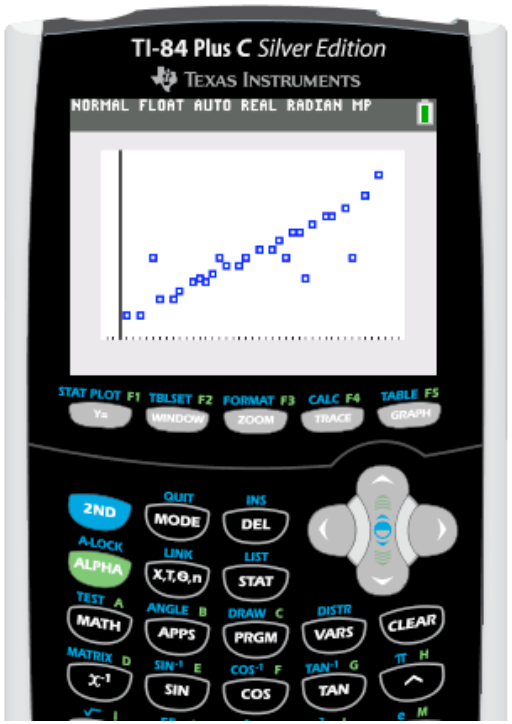
Let's continue our look at Bivariate Data with this example:
Suppose we want to know if there is an association between the number of spaces a property is from GO and the cost of the property in a monopoly game. The data is below:

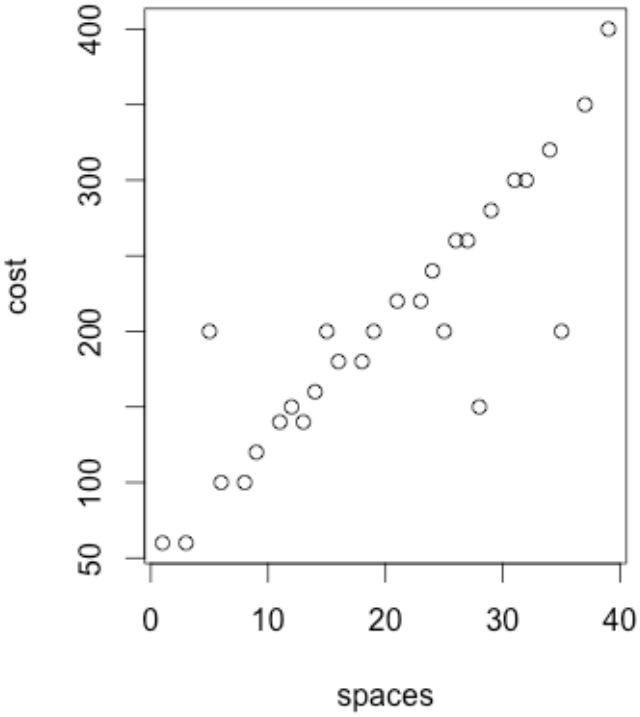
Property	**Spaces from GO**	**Cost**
Mediterranean Avenue	1	60
Baltic Avenue	3	60
Reading Railroad	5	200
Oriental Avenue	6	100
Vermont Avenue	8	100
Connecticut Avenue	9	120
St. Charles Place	11	140
Electric Company	12	150
States Avenue	13	140
Virginia Avenue	14	160
Penn Railroad	15	200
St. James Place	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

Which variable is explanatory and which is response?

Enter your data and create a scatterplot.





What can be said of the scatterplot?

The **correlation coefficient** measures the strength and direction of the linear relationship between two quantitative variables. The formula to find *r* is:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

The point $(\bar{x}, \bar{y})$ is:

The values of s_x and s_y are the individual standard deviations of *x* and *y* respectively.
n represents the number of data pieces.

Facts about Correlation:

1. Positive *r* indicates positive association and negative *r* indicates negative association between variables.
2. *r* is always between –1 and 1.
3. The closer |*r*| is to 1, the stronger the association. A weak association will have an *r* value close to 0.
4. Correlation is strongly influenced by outliers.

Using the Monopoly data, we can find the correlation coefficient with R using the command

cor(spaces, cost).
> cor(spaces,cost)
[1] 0.8779736

With the TI-83/84 calculator, we need to take a few more steps:

1. We must make sure the diagnostics is turned on:
   Do this by clicking 2ND – CATALOG and scroll down to Diagnostics
2. Choose STAT – CALC then 4:LinReg(ax+b)
3. Make sure your Xlist is L1 and Ylist is L2 and select Calculate:

What does this value reveal about our data from the Monopoly example?

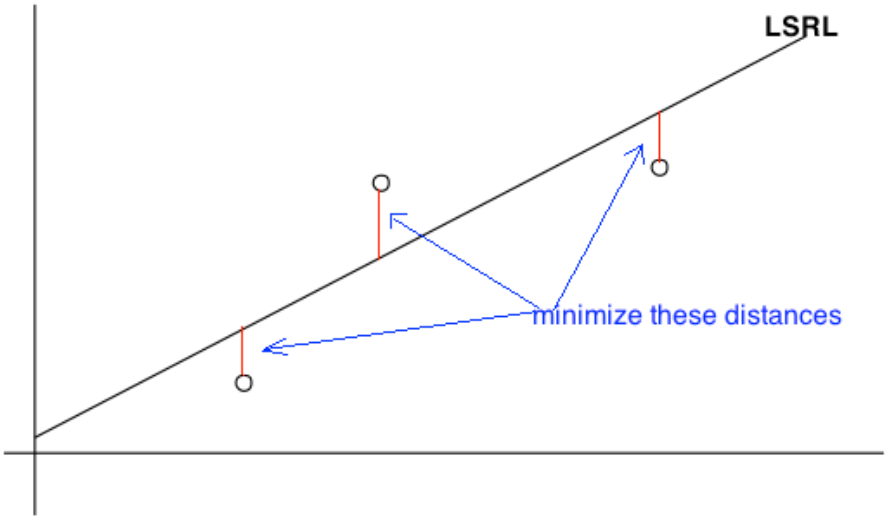
Popper 10

1. If your computed correlation coefficient r = +1.2, then you have better than a perfect positive correlation.

2. You should expect that there is a positive correlation between the age of your computer and its resale value.

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

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

For the Monopoly problem from section 5.1, we have:
> regline= lm(cost~spaces)
> regline

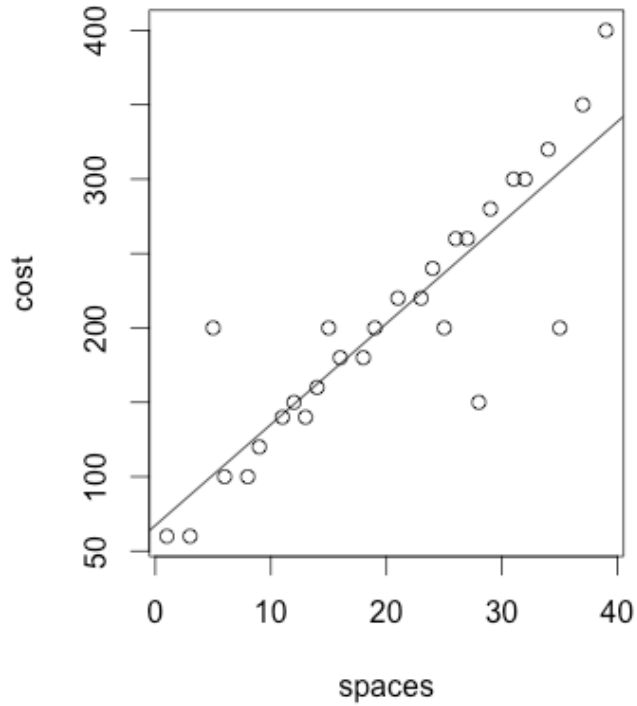
Call:
lm(formula = cost ~ spaces)

Coefficients:
(Intercept) spaces
67.283 6.784

So, the regression line is $y = 6.794x + 67.283$.

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)

Now we can see how well the model fits the graph.



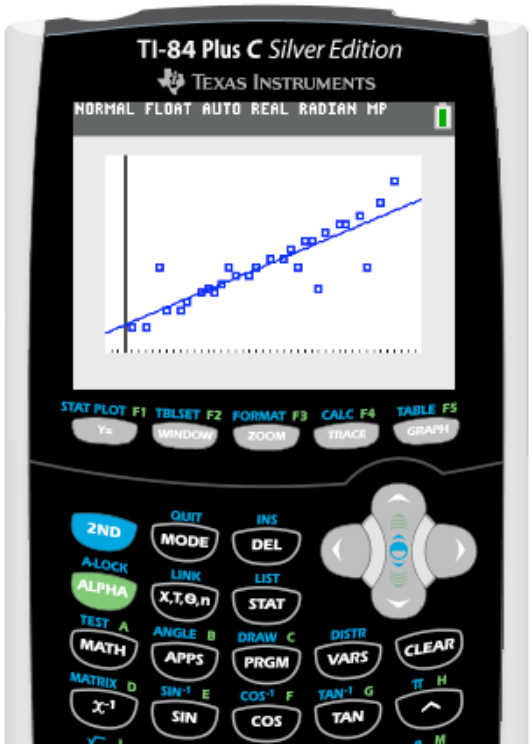
With the TI-83/84 we will follow some of the steps from section 5.2 with one difference. When we choose STAT – CALC.

If you are using a TI-84 Plus C, you will enter Y1 where it says Stor RegEQ on the LinReg screen

With the other TI-83/84 version, we will choose LinReg(ax+b) L1, L2, Y1

You select Y1 from VARS – Y-VARS

Now go to graph and graph the function. You may need to choose ZoomStat again.



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.

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.

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.

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."

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 $0 \le r^2 \le 1$
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^2values 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.

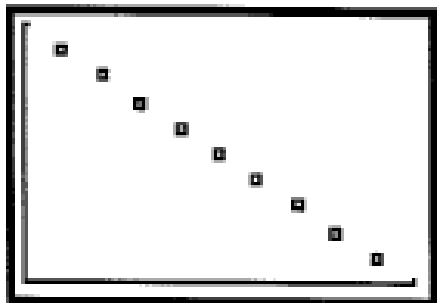
Any questions on homework or quizzes??

Popper 10

3. A least-squares regression line was fitted to the weights (in pounds) versus age (in months) of a group of many young children. The equation of the line is $\hat{y} = 16.6 + 0.65t$. Predict the weight of the child at 20 months.

4. A wildlife biologist is interested in the relationship between the number of chirps per minute for crickets (y) and temperature. Based on the collected data, the least-squares regression line is $\hat{y} = 10.53 + 3.41x$, where x is the number of degrees Fahrenheit by which the temperature exceeds 50 °. Which of the following best describes the meaning of the slope of the least-squares regression line?

Match the following scatter plots with the appropriate correlations from the list:

5. 

6.