Statistics 10 Lecture 21 (Chapter 12.1)

1. The regression Line

Given a set of data, find the line that best summarizes the relationship between two quantitative variables X and Y. The method of least squares (or minimizing the deviations in the vertical direction from the line) is used because it is simple. A line can be found from just knowing the means and standard deviations of X and Y and their correlation r.

Note: Y is the dependent or response or outcome variable in this context and X is the independent or predictor or explanatory variable. In correlation, it doesn't matter which one is designated as X and which one Y, but in regression, it matters.

2. The regression line

A. Scatter diagrams again

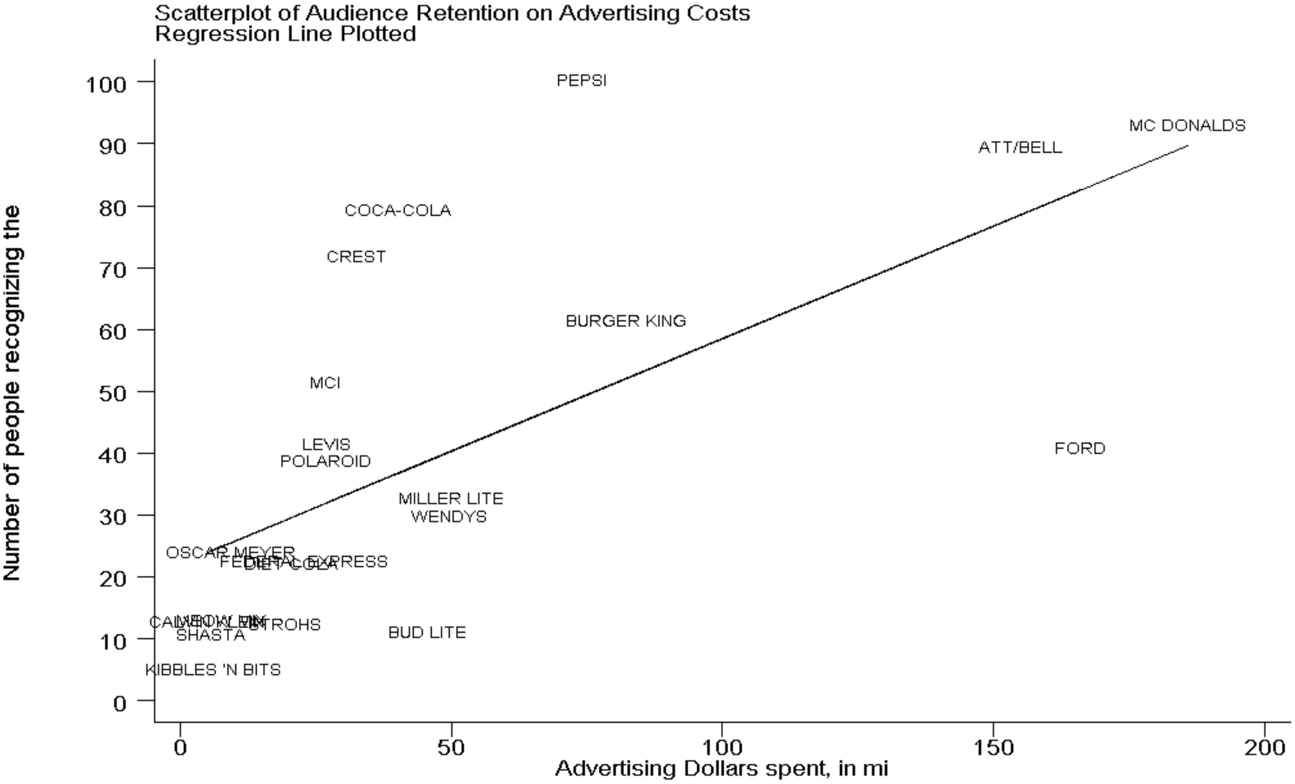
The X values are plotted along the horizontal axis, the Y values along the vertical. X is the explanatory variable, Y is the response variable. In Chapter 12 we would like to describe an overall pattern of a scatter diagram by fitting a straight line to the scatter diagram. Here is some information on the amount of money companies spend on advertising (in millions) and the number of people (in millions) who can recognize the brand. Which one would you consider X and which one would you consider Y?

Company Name	Spending	Recognition
MILLER LITE	50.1	32.1
PEPSI	74.1	99.6
STROHS	19.3	11.7
FEDERAL EXPRESS	22.9	21.9
BURGER KING	82.4	60.8
COCA-COLA	40.1	78.6
MC DONALDS	185.9	92.4
MCI	26.9	50.7
DIET COLA	20.4	21.4
FORD	166.2	40.1
LEVIS	27	40.8
BUD LITE	45.6	10.4
ATT/BELL	154.9	88.9
CALVIN KLEIN	5	12
WENDYS	49.7	29.2
POLAROID	26.9	38
SHASTA	5.7	10
MEOW MIX	7.6	12.3
OSCAR MEYER	9.2	23.4
CREST	32.4	71.1
KIBBLES 'N BITS	6.1	4.4

This is what data looks like. This is why a graphical summary (scatter diagram) or a numerical summary (mean, median, standard deviation) is more useful. We only have 20 companies here. There are far more companies out there, but it would be difficult to examine a huge list.

The correlation for advertising dollars spent and number of people recognizing the brand name is .6511

The mean for recognition is 40.4667 and the standard deviation is 30.18061
The mean for spending is 50.4 and the standard deviation is 54.10918



B. Things to think about when looking a regression line on a scatter diagram

(1) It's formula is:

y = (slope*x) + intercept OR y = bx + a OR y = mx + b

No straight line will pass exactly through all of the points. A fitted line comes as close as possible to all of the points simultaneously. The assumption being made here is that y is dependent on x or y is the response, dependent, or outcome variable, x is the explanatory, independent or predictor variable.

(2) Calculating the slope (b) of the line

the formula is

b = r * ((SD of Y)/(SD of X)) (see page 204)

The slope b measures the average observed change in Y when X changes by one unit. It is thought of as a rate of change. For our example on advertising dollars spent and the number recognizing the company name:

b = .6511 * 30.18061/54.10918 = .36317

Note that the correlation, r, is determining the sign of the slope b. If r were equal to 1 or negative 1, the X and Y variables would be changing at a similar rate. In this example, y is changing at a slower rate than x.

(3) Calculating the intercept: a (page 205)

a = (average of the y variable) - b(average of the x variable) or 22.163=40.46667-(.36317*50.4)

This is the value of y when x=0. And note that the line then always crosses through the point represented by the means of x and of y. This is a check if you are calculating a regression with a hand calculator and not a computer.

Note: you must have calculated slope before you can apply this formula to calculate intercept.

(4) Put it all together in a regression equation

y = (slope*x)+ intercept so y = .36317(ad dollars spent in millions) + 22.16269

3. Using the Regression Line

Prediction: The prediction equation for the number recognizing the company is:

Number Recognizing in Millions = 22.16269 + (.3631741)(Ad Dollars Spent in Millions)

Interpretation: slope tells you how much change on average to expect in Y if X is changing. Intercept tells you what Y would be if X were equal to zero (sometimes this is nonsense). Most applications of regression are interested in slope. In our example, for each million dollars spent on advertising, we expect about .36 million or about 360,000 people to recognize the company name. If a company spent 100 million on advertising then we would expect 36 million people to recognize the company. For the intercept, if a company spent nothing, that is X=0, the model predicts that there would be 22.16269 million or 22,162,690 people will recognize the company name.

Extrapolation. If someone were interested in the number recognizing the company name if the advertising dollars were to rise to 300 million, this is using a regression line to predict values outside of the range that we have. These predictions are not usually accurate.