

16. SIMPLE LINEAR REGRESSION I

Often, we wish to study the relationship between two variables:

- GMAT Scores and Grade Point Averages of First-Year Business School Students.
- Current GDP, Unemployment in 2 Years.
- Fat Content of Ice Cream, and its Sales.
- Height, Salary.

Purpose:

- (1) To describe and understand how the variables are related. For example, Starbucks can use their understanding of the relationship between daily coffee consumption and the consumer's age (as well as other demographic factors) to help them decide when and where to open a new store.
- (2) To forecast a new observation. If we know the relationship between current GDP and future unemployment, then we can try to forecast the unemployment.
- (3) To adjust or control a process. If we know how fat content influences ice cream sales, we can adjust our fat content to (hopefully) improve sales.

- The first thing to do is to make a **scatterplot**: a graphical display of each data point using two axes to represent the two variables.

If one variable is seen as causing or influencing the other, it is called X and defines the horizontal axis.

Some common names for X are:

Predictor Variable, Explanatory Variable, Independent Variable, Exogenous Variable, Regressor, Factor.

The variable that might respond or be influenced is called Y and defines the vertical axis. Some common names for Y are:

Response Variable, Dependent Variable, Endogenous Variable.

Our data take the form $(x_1, y_1), \dots, (x_n, y_n)$, where each X value is paired with its corresponding Y value.

Even if the scatterplot shows a relationship between the variables, this does not in any way prove that X *causes* Y. For example, it would be ludicrous to argue that business students with A averages got them *because* they did well on the GMAT.

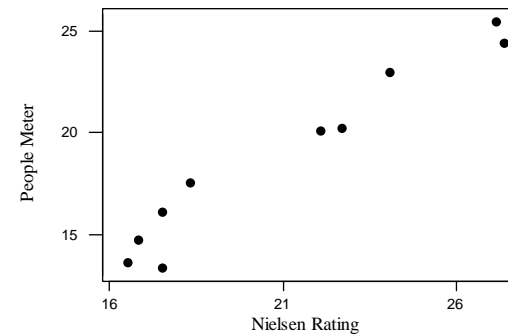
Sometimes (but not always) our scatterplot will show a **linear relationship**: The points seem to be bunched randomly around a straight line. Ideally, the amount of scatter does not depend on X and there are no extreme outliers. In this case, the methods to be studied here can be safely applied.

Eg: Consider the ratings for the top 10 TV shows from Oct 10 to Oct 23, 1986. The ratings were obtained by the Nielsen Index (based on a written diary of TV shows watched in a household), and People Meters (which record the information automatically).

One rating point represents 1% of the 97.7 Million adults aged 25-54.

	<u>Nielsen Index</u>	<u>People Meter</u>
The Cosby Show	27.4	24.4
Family Ties	27.2	25.5
Cheers	24.1	23.0
Moonlighting	22.7	20.2
Night Court	22.1	20.1
Growing Pains	18.3	17.5
Who's The Boss	17.5	16.1
Family Ties (Special)	17.5	13.3
Murder She Wrote	16.8	14.7
60 Minutes	16.5	13.6

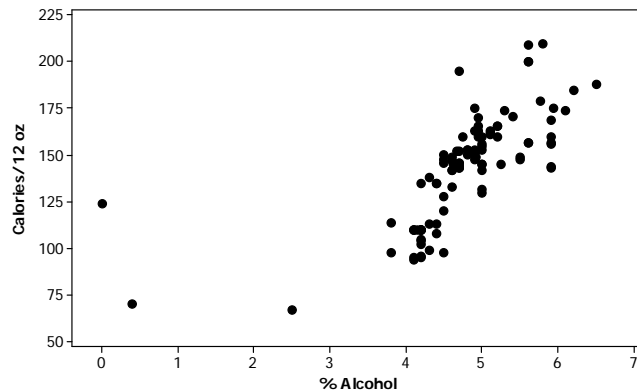
Scatterplot of People Meter vs. Nielsen Ratings



The scatterplot shows a strong positive linear association, but not a perfect relationship. Note also that People Meter ratings seem to be systematically less than the Nielsen Index.

Eg: Consider the calories per 12-ounce serving and the percentage of alcohol for 101 US domestic beers.

US Domestic Beers: Calories vs. % Alcohol



Overall, we see a positive linear association but there are at least two strong outliers. To identify them in Minitab, hold the cursor over the data point.

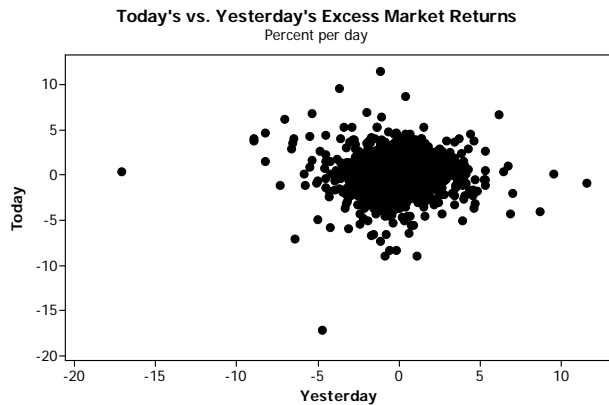
It is also possible that a scatterplot will show **no relationship**, i.e., the points seem randomly distributed, with no particular tendency to move up or down as X changes.

Eg: Does the Stock Market have "momentum"? That is, is the Market likely to keep going up today because it went up yesterday? To study this, consider the daily excess returns on the Market (see Handout 1), and make a scatterplot of Today's return (Y) against Yesterday's return (X).

“I don’t know what the S&P 500 will return, and I am the chairman of the index committee which runs it.”

--- David Blitzer, chief economist at Standard & Poor’s Corp.

Clearly, stock returns are not very predictable.



Top 4 outliers:
Oct 19-Oct 20, 1987 (Crash),
Oct 13-Oct 14, 2008 (Surge).

The scatterplot shows virtually no relationship between X and Y. This, of course, would be consistent with the Efficient Market Hypothesis, i.e., that daily prices are a random walk.

If you want, you can view "no relationship" as being the same as "linear relationship with a slope of zero". But if the slope is zero, today's return is useless for predicting tomorrow's. That is, Y does not depend in any meaningful way on X.

Another possibility is that the scatterplot shows a **nonlinear relationship**: The points are bunched around a curve rather than a straight line. Since there are so many different kinds of curves that can be drawn, the analysis is more complex. For example, consider the height of children (Y) and their ages (X).