MAS 170 Elementary Statistics Spring 2020
Notes on Regression, Chapters 8--12

The purpose of this note is to supplement the material on linear regression in the textbook.

Regression

Given two lists of numbers X,Y of the same length and a linear function y=L(x)=ax+b (a and b are constants), the ***rms error for the line L*** is the rms of the list Y-L(X). (For example, if A is an entry in the list X and B is the corresponding entry in the list Y, then the corresponding entry in the list Y-L(X) is B-(aA+b).) We interpret the rms error for a line as a measure of how well L performs as a tool for predicting the Y value that corresponds to a given X value. Given an X value A, the value B=L(A) has a "prediction error" equal to B-L(A). The rms of the list of prediction errors is low if L is a good prediction tool and high if L is a poor prediction tool. We interpret the rms error for a line L as the size of the typical error that we would observe if we used L to predict an unknown Y value for a given X value.

Among all possible lines, there is one single line with the least possible rms error, called the ***regression line*** for Y on X. This means that the regression line is the best possible line to use as a prediction tool to guess an unknown corresponding Y value for a given X value. The regression line passes through the point (AVE(X),AVE(Y)), called the ***point of averages*** for the lists X,Y, and has slope r*SD(Y)/SD(X), where r is the number

r = AVE([X-AVE(X)/SD(X)]*[Y-AVE(Y)/SD(Y)])

called the ***correlation coefficient*** for the pair of lists X,Y. The rms error for the regression line is

rms error for regression line = SD(Y) * SQRT(1-r^2).

The list (X,Y) of pairs of points from the lists X,Y is called the ***scatter diagram*** of the lists X,Y. The scatter diagram is said to show ***linear association*** if the points are scattered roughly symmetrically about a linear axis. The size of the correlation coefficient determines the tightness or looseness of scattering of points (X,Y) about the regression line. For |r| near 1, points in the scatter diagram are tightly clustered about the regression line. As |r| approaches 0, the scatter diagram is more and more loosely clustered about the regression line.

A ***thin vertical strip*** in the scatter diagram is a set of points (X,Y) whose X-coordinate lies in a range from A-e to A+e, where A is a number within the range of entries of the list X and e is a small increment. If the various sets of Y values in thin vertical strips (for various values of A) has a similar SD as A varies, the scatter diagram is said to be ***homoscedastic***. Otherwise, the scatter diagram is said to be ***heteroscedastic***.

Let's write REGR(A) for the y-coordinate of a point on the regression line for the x-value A. Here's a formula.

REGR(A) = r * [A-AVE(X)/SD(X)] * SD(Y) + AVE(Y)

A ***residual*** is another name for a vertical error Y-REGR(X) for a data point (X,Y). The ***graph of residuals*** is the set of points (X,Y-REGR(X)). You can see linear association and homoscedasticity (or the lack of either one) very plainly in the graph of residuals: data lists X,Y have a linear association if the graph of residuals is roughly symmetric across the X-axis; and the X,Y data are homoscedastic if the graph of residuals has fairly even densities and vertical ranges in thin vertical strips across the horizontal range of X.

For the sake of comparison, here are three lines and their rms errors. All three lines pass through the point of averages. The SD line is defined to have positive slope if r>0 and negative slope if r<0.

line	slope	rms error
regression line	r*SD(Y)/SD(X)	SD(Y) * SQRT(1-r^2)
SD line	+/- SD(Y)/SD(X)	SD(Y) * SQRT(2(1-|r|))
average line	0	SD(Y)

Facts about the correlation coefficient r

* The value of r is in the range −1 to 1. The value of $|r|$ is 1 precisely when all the points (X,Y) on the scatter diagram lie on a straight line.

* The correlation coefficient is unitless. The value of r does not change if you rescale X or Y.

* It is tempting to summarize data using averages, but this has a misleading effect on correlation. Suppose that the points in a scatter diagram are subdivided into subsets, and that each subset is replaced by a single point

  (average of the X values in the subset, average of the Y values in the subset).

  The correlation for this smaller data set of average points is usually higher (in absolute value) than the correlation for the original scatter diagram. This phenomenon is called ***ecological correlation***.

* A strong correlation provides no evidence of a causal link between the variables X and Y.

The regression effect and regression fallacy

Start with an X-value A, somewhere above or below the value AVE(X). Use the regression line for Y on X to predict the average Y value for data points whose X value is near A. Call this prediction B. Now use the regression line for X on Y to predict the average X value for data points whose Y value is near B. Call this prediction C. It will always turn out that C is between AVE(X) and A. This is due simply to the fact that, for real data, the value of $|r|$ is less than 1. This is called the ***regression effect***. Any attempt to explain why C is between AVE(X) and A by any other reason than simply the fact that $|r|<1$ is called a ***regression fallacy***.