Correlation and Regression (also known as the general problem of assessing association between variables) –

We start with the simplest case of just two variables and then extend this –

First of all, three distinctions must be made – these three distinctions with be covered in turn.

1) Descriptive nature and use of correlation and regression

2) Correlational model

3) Regression model

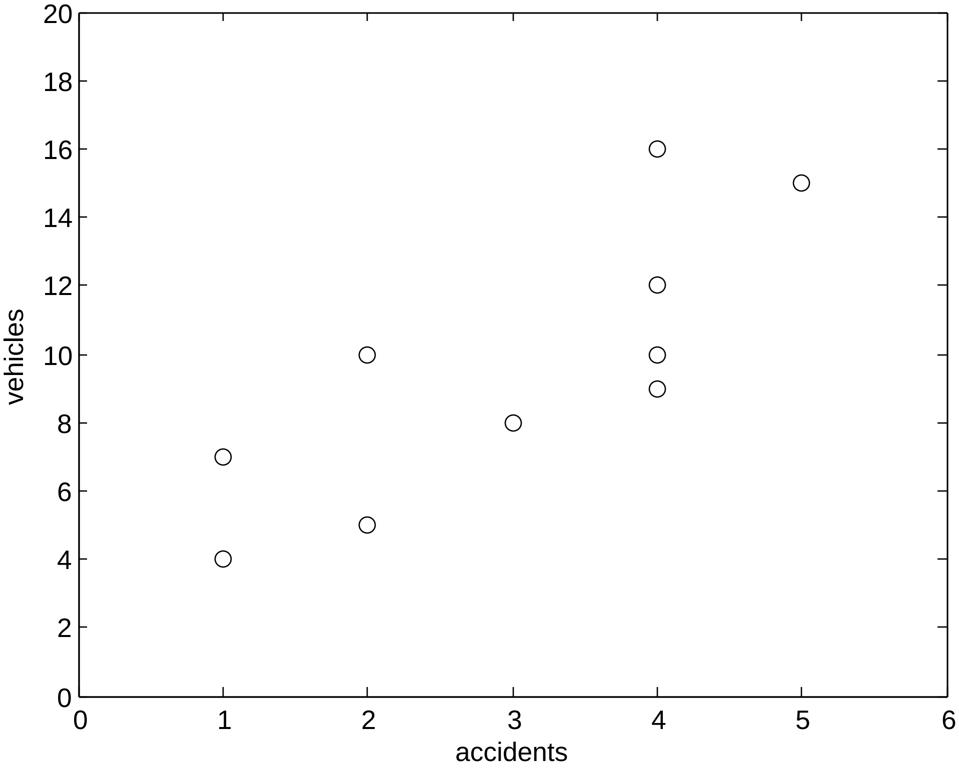
These last two are based on differing underlying assumptions about the populations that we observe.

Best to present the distinctions in terms of an example – we will carry through this example for a while.

Example: an investigation of the relationship between the number of licensed vehicles in a community and the number of accidents.

community	licensed vehicles (thousands)	accidents (hundreds)
1	4	1
2	10	4
3	15	5
4	12	4
5	8	3
6	16	4
7	5	2
8	7	1
9	9	4
10	10	2

Yogi Berra: you can see a lot by just looking

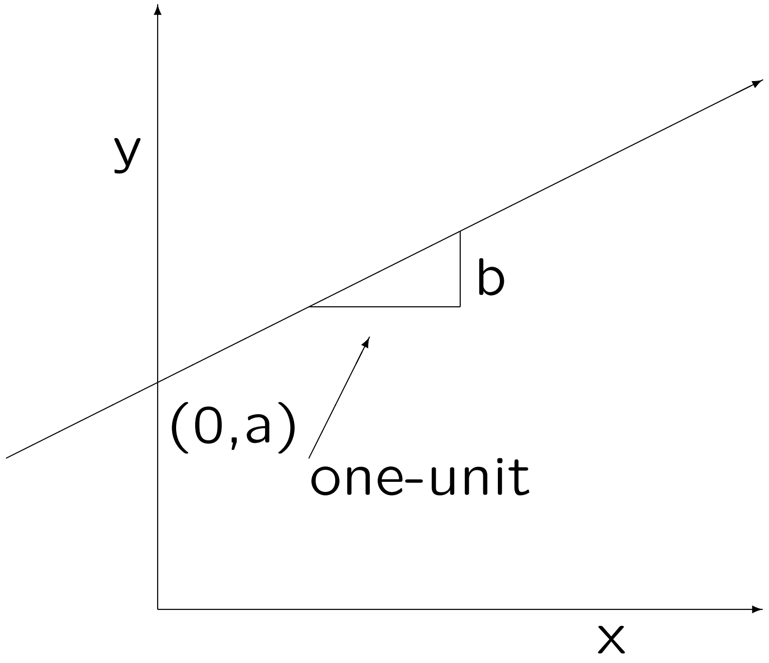

vehicles
0
2
4
6
8
10
12
14
16
18
20
0
1
2
3
4
5
6
accidents

1) the descriptive nature of this problem is to find the line that approximates this scatter of points the best – it is called the least-squares line because it is the line at which the squared deviations from the line are at a minimum.

2) regression – the x variable, i.e., accidents, is considered fixed and we observe a number of vehicles within each fixed value.

3) correlation – both sets of scores are random variables and we select the pair *together* from some bivariate population – one set is not fixed in advance.

We begin with the descriptive aspects of the problem.



For each unit on x, we go b units on y.

Slope-intercept form for the line: $y = bx + a$

Descriptive Statistics of Correlation and Regression:

Suppose I have two sets of scores that are paired (N pairs):

Set 1	Set 2
X_1	Y_1
X_2	Y_2
$\vdots$	$\vdots$
X_N	Y_N

For example, number of vehicles and accidents.

To make things somewhat more standard, suppose I normalize each of the variables to z-scores (so the z-scores have mean zero and standard deviation of one):

$Z_{X_i} = \frac{X_i - M_X}{S_X}$ and $Z_{Y_i} = \frac{Y_i - M_Y}{S_Y}$

where M_X and S_X (M_Y and S_Y) are the mean and standard deviation of the X (Y) scores

So, we have:

Set 1	Set 2
Z_{X_1}	Z_{Y_1}
Z_{X_2}	Z_{Y_2}
$\vdots$	$\vdots$
Z_{X_N}	Z_{Y_N}

What we are looking for is some linear function of the Z_{X_i} scores that will predict the Z_{Y_i} scores "well" −

In other words, we are looking for a good equation of the form:

$$Z'_{Y_i} = bZ_{X_i} + a$$

that produces a line in the scatterplot; Z'_{Y_i} is the predicted value for Z_{Y_i}

In addition, we would like to choose a and b so it is a "good" equation.

If I have an equation of the form $Z'_{Y_i} = bZ_{X_i} + a$, the term $Z'_{Y_i} - Z_{Y_i}$ is a “deviation” or “error” of what the equation says and what the actual observed quantity is for Z_{Y_i} (a vertical discrepancy in the scatterplot).

We will say that the a and b are the best coefficients in a least-squares sense if

$$\frac{\sum_{i=1}^{N}(Z'_{Y_i} - Z_{Y_i})^2}{N}$$

is at a minimum. If a and b are these coefficients, then the equation $Z'_{Y_i} = bZ_{X_i} + a$ is called the “least-squares line” for predicting Z_{Y_i} from Z_{X_i}.

The “correlation” between Z_{Y_i} and Z_{X_i} is defined as

$$r_{XY} = \frac{\sum_{i=1}^{N} Z_{X_i} Z_{Y_i}}{N}$$

I want to show that the best we can do is to let $b = r_{XY}, a = 0$. So, the least-squares line is $Z'_{Y_i} = r_{XY} Z_{X_i}$

First, show $a = 0$:

$\frac{\sum_{i=1}^{N}(Z'_{Y_i} - Z_{Y_i})^2}{N}$ has to be at a minimum.

$$\frac{\sum_{i=1}^{N}(Z'_{Y_i} - Z_{Y_i})^2}{N} =$$

$$\frac{\sum_{i=1}^{N}(bZ_{X_i} + a - Z_{Y_i})^2}{N} = \ldots =$$

$$\frac{\sum_{i=1}^{N}(bZ_{X_i} - Z_{Y_i})^2}{N} + a^2$$

which is a minimum when $a = 0$;

now,

$$\frac{\sum_{i=1}^{N}(bZ_{X_i}-Z_{Y_i})^2}{N} = b^2 - 2br_{XY} + 1;$$

suppose $b = r_{XY} + c$ for some value of c. Then, $b^2 - 2br_{XY} + 1 = (1 - r_{XY}) + c^2$, which is minimized when $c = 0$.

Note that

$$\frac{\sum_{i=1}^{N}(Z'_{Y_i}-Z_{Y_i})^2}{N} = 1 - r_{XY}^2$$

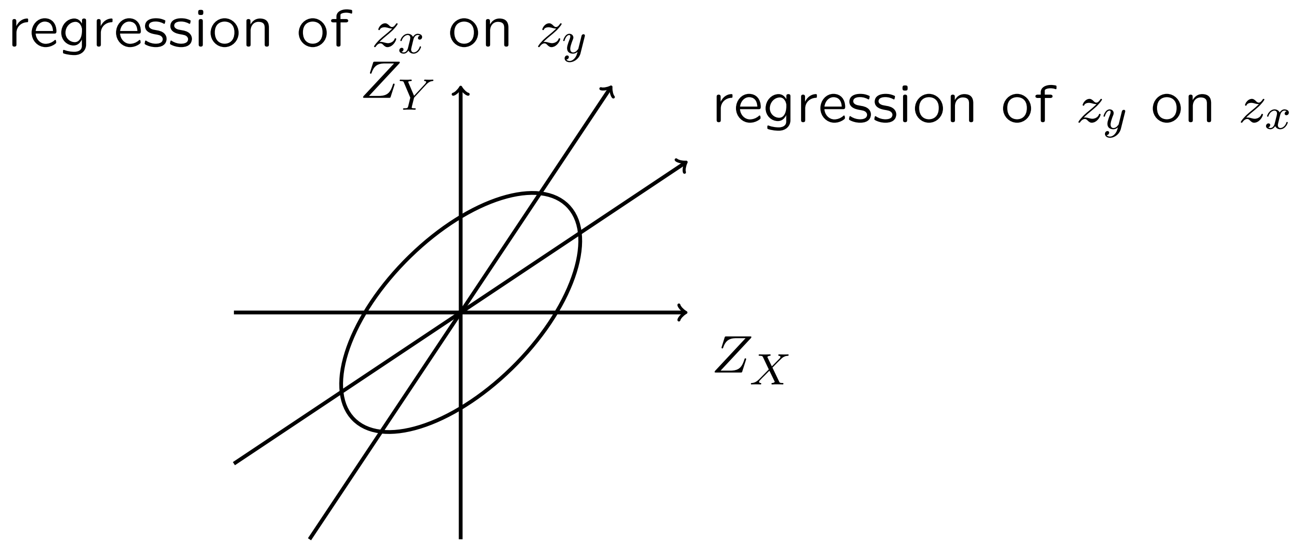
when $b = r_{XY}$ and $a = 0$. This is called the "sample variance of estimate for standard scores", and denote it by $S_{Z_Y \cdot Z_X}^2 (= 1 - r_{XY}^2)$.

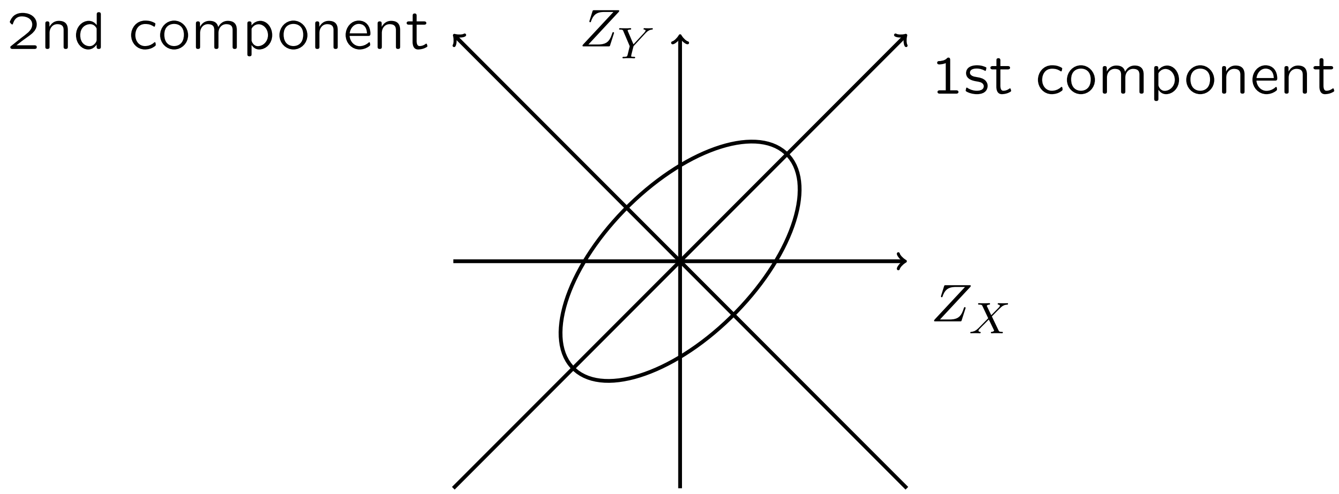
Or, $S_{Z_Y \cdot Z_X} (= \sqrt{1 - r_{XY}^2})$ is the "sample standard deviation of estimate for standardized scores" (also, called the "standard error of estimate for standardized scores") –

Because $\frac{\sum_{i=1}^{N}(Z^{'}_{Y_i} - Z_{Y_i})^2}{N} = 1 - r^2_{XY} \geq 0$, $1 \geq r^2_{XY}$ and $-1 \leq r_{XY} \leq 1$, i.e., the correlation lies between minus 1.0 and plus 1.0.

If $r_{XY} = 0$, then $Z^{'}_{Y_i}$ is 0.0, and we estimate Z_{Y_i} by the mean of the Z_{Y_i}'s; in other words, $r_{XY} = 0$ implies $\frac{\sum_{i=1}^{N}(Z^{'}_{Y_i} - Z_{Y_i})^2}{N} = \frac{\sum_{i=1}^{N}(Z_{Y_i})^2}{N} =$ 1, which is the original variance of the Z_{Y_i} scores (so no predictive advantage is achieved). One would expect to see a circular smear in the scatterplot of Z_{Y_i} against Z_{X_i}.

If $r_{XY} = 1$, then $\frac{\sum_{i=1}^{N}(Z^{'}_{Y_i} - Z_{Y_i})^2}{N} = 0$ and we have perfect prediction. All the scores in the scatterplot of Z_{Y_i} against Z_{X_i} lie on a line with slope of 1.0. If $r_{XY} = -1$, then $\frac{\sum_{i=1}^{N}(Z^{'}_{Y_i} - Z_{Y_i})^2}{N} =$ 0 and we again have perfect prediction. All the scores in the scatterplot of Z_{Y_i} against Z_{X_i} lie on a line with slope of -1.0.


regression of z_x on z_y
Z_Y
regression of z_y on z_x
Z_X


2nd component
Z_Y
1st component
Z_X

The quantity r^2_{XY} is called the "coefficient of determination" and indicates the percent of variance in Y and can be accounted for by a simple linear function of X

The sample variance of the predicted scores, $Z'_{Y_1}, \ldots, Z'_{Y_N}$ (or, $r_{XY}Z_{X_1}, \ldots, r_{XY}Z_{X_N}$) is

$(1/N)\sum_{i=1}^{N}(Z'_{Y_i})^2 - ((1/N)\sum_{i=1}^{N}(Z'_{Y_i}))^2$,

which reduces to r^2_{XY}.

Explained variance = (sample variance of predicted scores)/(total sample variance) = $r^2_{XY}/1.0$ = r^2_{XY}, the coefficient off determination. The unexplained variance is $1 - r^2_{XY}$, which is the value of the least-squares loss criterion.

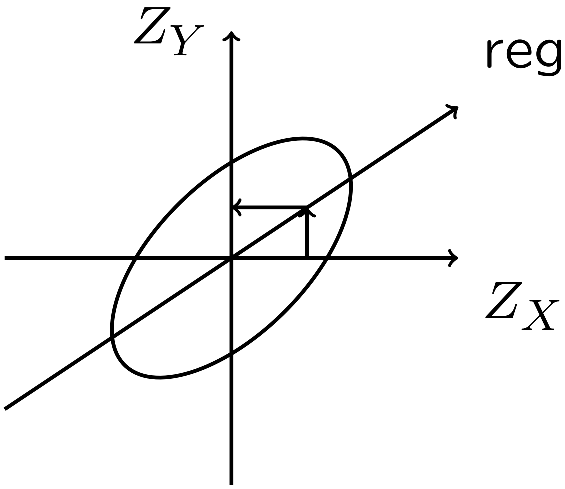
For our example of predicting the number of accidents from the number of vehicles, it will be shown (i.e., computed) later that $r_{XY} =$ .80; thus, $r^2_{XY} = .64$, so 64% of the variance in accidents can be accounted from by a simple linear function of the number of vehicles.

Regression toward the mean –

$Z_Y^{'} = r_{XY} Z_X$, but because $-1 \leq r_{XY} \leq +1$, the predicted score for $Z_Y^{'}$ will be closer toward its mean of zero than was Z_X toward its mean of zero.

For example, let Y be a child's height, and X be a father's height. The best estimate of the child's height is closer to the average height than was his father's (this is also true if reversed: the best estimate of a father's height is closer to the average height than was the child's – $Z_{X_i}^{'} = r_{XY} Z_{Y_i}$)

This was once considered to be some type of natural law called regression toward the mean (and we were all moving toward a "sea of mediocrity"). We now realize it to be a function of fallible measurement and imperfect relations rather than because of any linear rule.



regression of z_y on z_x

Beware of regression effects in such things as test scores, or in selecting extreme scores to form groups. One expects regression effects on repeat testing (and therefore we do not have a causal argument that what you did to the groups "caused" them to "get better") –

Raw Scores –

We have developed all of our techniques in terms of standardized scores but it is convenient to rephrase our equations in terms of raw scores ($Y^{'}$ is the (implicitly defined) predicted raw score on Y) –

$Z_Y^{'} = r_{XY} Z_X$ implies $\frac{Y^{'} - M_Y}{S_Y} = r_{XY}(\frac{X - M_X}{S_X})$

or, $Y^{'} = \frac{r_{XY} S_Y}{S_X}(X - M_X) + M_Y =$

$Y^{'} = \frac{r_{XY} S_Y}{S_X} X + (M_Y - \frac{r_{XY} S_Y}{S_X} M_X)$

These are the raw score forms of the regression equation for the prediction of Y from X

The sample regression coefficient is denoted by $b_{Y \cdot X}$ ($\equiv \frac{r_{XY} S_Y}{S_X}$)

$$\frac{\sum_{i=1}^{N}(Z'_{Y_i} - Z_{Y_i})^2}{N} = 1 - r^2_{XY} (\equiv S^2_{Z_Y \cdot Z_X}) =$$

$$\frac{\sum_{i=1}^{N}((\frac{Y'_i - M_Y}{S_Y}) - (\frac{Y_i - M_Y}{S_Y}))^2}{N} = 1 - r^2_{XY},$$

or

$$\frac{\sum_{i=1}^{N}(Y_i - Y'_i)^2}{N} = S^2_Y(1 - r^2_{XY})(\equiv S^2_{Y \cdot X}) -$$

So, the "variance" around the predicted values (rather than around the mean) and is called the "sample variance of estimate" for predicting Y from X; the square root ($S_{Y \cdot X} = S_Y\sqrt{(1 - r^2_{XY})}$) is called the "sample standard error (deviation) of estimate"

Computational form for the correlation:

$$r_{XY} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{(N\sum X^2 - (\sum X)^2)(N\sum Y^2 - (\sum Y)^2)}}$$

Computational form for the regression coefficient:

$$b_{Y \cdot X} = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$$

For our data on X (number of accidents) and Y (number of vehicles),

$\sum X = 30; \sum Y = 96; \sum X^2 = 108; \sum Y^2 = 1060; \sum XY = 328; N = 10$

Plugging into the computational formulas, $r_{XY} =$ .80 and $b_{Y \cdot X} =$ 2.22; thus, the least-squares line has the form:

$$Y' = 2.22(X - 3) + 9.6 = 2.22X + 2.94$$

Because $r_{XY}^2 =$ .64, 64% of the variance in Y is accounted for by a linear relation to X. Also, because $S_Y^2 =$ 13.86, the sample variance of estimation is $S_{Y \cdot X}^2 =$ 13.86(1 – .64) = 4.98; $S_{Y \cdot X} = \sqrt{4.98} =$ 2.23 and represents deviation within the X values across the regression line.

Because correlation is defined in terms of standard scores, a linear function of the original scores does not change the correlation. Thus, if $U_i = cX_i + g$ and $V_i = dY_i + h$, then $r_{UV} = \pm r_{XY}$ (positive if both c and d have the same signs, and negative if c and d have opposite signs)

The regression coefficient $b_{Y.X}$ changes with linear transformation of the original variables. Moral: be wary of any attempt to interpret the regression coefficient without a mention of the scale on which the variables are measured.

Remember that we have assumed that a linear rule is the “true” one. Thus, one might have a perfect curvilinear relationship (i.e., the X and Y pairs lie on a simple curved structure) but not have values on the correlation close to plus (or minus) one. If one changes to ranks and recomputes the correlation (now called the Spearman rank-correlation coefficient) and the curved structure is one where X and Y are monotonically related – as X goes up, so does Y – then a perfect Spearman correlation would ensue.

Final point: no assumptions about what kind of variables we were using (e.g., normally distributed). Our concern was only with descriptive statistics, and the best “linear” rule for predicting Y from X.

To make inferences, we need to make some stronger assumptions – we will cover the regression and correlation models in the population.

Doing it in R:

```
accidents = c(1,4,5,4,3,4,2,1,4,2)

comm.labels = c(’a’,’b’,’c’,’d’,’e’,’f’,’g’,’h’,’i’,’j’)

vehicles = c(4,10,15,12,8,16,5,7,9,10)

police = c(20,6,2,8,9,8,12,15,10,10)

community =

data.frame(comm.labels,accidents,vehicles,police)

community.model = lm(vehicles ~ accidents)

plot(vehicles ~ accidents)

abline(community.model)

coef(community.model)

summary(community.model)
```

To replicate this analysis in SYSTAT we have a file called community_data.syz ;

for MATLAB we have a file called community_data.dat

Demo follows –

Algebraic restrictions on correlations:

In any multiple variable context, it is possible to derive the algebraic restrictions present among any subset of the variables based on the correlations among all the variables.

The simplest case involves three variables, say X, Y, and W.

From the basic formula for the partial correlation between X and Y "holding W constant," an *algebraic* restriction is present on r_{XY} given the values of r_{XW} and r_{YW}:

$$r_{XW}r_{YW} - \sqrt{(1 - r_{XW}^2)(1 - r_{YW}^2)} \leq$$

$$r_{XY} \leq r_{XW}r_{YW} + \sqrt{(1 - r_{XW}^2)(1 - r_{YW}^2)} \ .$$

Note that this is not a probabilistic statement (that is, it is not a confidence interval); it says that no dataset exists where the correlation r_{XY} lies outside of the upper and lower bounds provided by $r_{XW}r_{YW} \pm \sqrt{(1 - r_{XW}^2)(1 - r_{YW}^2)}$.

As a numerical example, suppose X and Y refer to height and weight, respectively, and W is a measure of age. If, say, the correlations, r_{XW} and r_{YW} are both .8, then $.28 \leq r_{XY} \leq 1.00$.

In fact, if a high correlation value of .64 were observed for r_{XY}, should we be impressed by the magnitude of the association between X and Y?

Probably not;

if the partial correlation between X and Y "holding W constant" were computed with $r_{XY} = .64$, a value of zero would be obtained.

All of the observed high association between X and Y can be attributed to their association with the developmentally related variable.

Conversely, if X and Y are both uncorrelated with W (so $r_{XW} = r_{YW} = 0$), then no restrictions are placed on r_{XY};

the algebraic inequalities reduce to a triviality: $-1 \leq r_{XY} \leq +1$.

These very general restrictions on correlations have been known for a long time and appear, for example, in Yule's first edition (1911) of *An Introduction to the Theory of Statistics* under the title, "Conditions of Consistence Among Correlation Coefficients."

Also, see the chapter, "Fallacies in the Interpretation of Correlation Coefficients," in this same volume.