Chapter 8 Covariance and Correlation

In this chapter, we will look at one of **two** methods of analyzing the extent to which two random variables, say X and Y, are linearly related. These two methods are related to each other. The first method is called **correlation** and the second method is called **simple linear regression**. Thus, if we have two random variables X and Y, we can ask what is the correlation between X and Y, and we can also ask what is the linear regression of Y on X. The answers to the two questions are connected mathematically. You can discuss correlation or linear regression, but in fact you are discussing the same thing. They are two sides of the same coin.

In earlier chapters, we consider a single random variable. Associated with this random variable, X, was a function called a pdf, f(s). The pdf is always nonnegative and sums to 1. Now we have two random variables. These two random variables, X and Y, share a pdf. We can write this pdf as f(s,w). The same conditions hold for this joint pdf. It must be nonnegative and it must sum to zero when summed (or integrated) over both s and w. The theoretical or population means can be easily written as

$$E[X] = \int\int sf(s,w)dsdw \text{ and } E[Y] = \int\int wf(s,w)dsdw$$

and the theoretical or population variances can be written similarly as

$$\text{var}[X] = \int\int (s - E[X])^2 f(s,w)dsdw \quad \text{and} \quad \text{var}[Y] = \int\int (w - E[Y])^2 f(s,w)dsdw$$

Now these calculations may look scary, but they are not really so bad. The double integral signs tell us that we are no longer adding up along a single dimension. We are now adding up along two dimensions – one corresponding to X and one corresponding to Y. Intuitively, E[X] and E[Y] tell the position where *most* of the probability of X and Y is located. Similarly, the variances tell how much the probability of X and Y are *spread out* in the two directions from the point (E[X], E[Y]). Once again, these "moments about the mean" are simply numbers that crudely describe a surface, i.e. the bivariate pdf – which is of course difficult to do since the bivariate surface is infinitely many points and can be shaped in infinitely many ways.

The introduction of two random variables at the same time allows us to create a new measure of the linear relation between these two. The measure is called the correlation coefficient of X and Y. However, to get this measure we must first add one more interesting measure called the covariance. Remember that the variance is the expectation of the squared deviation of the random variable about its mean

$$\text{var}(X) = E[(X - E[X])^2]$$

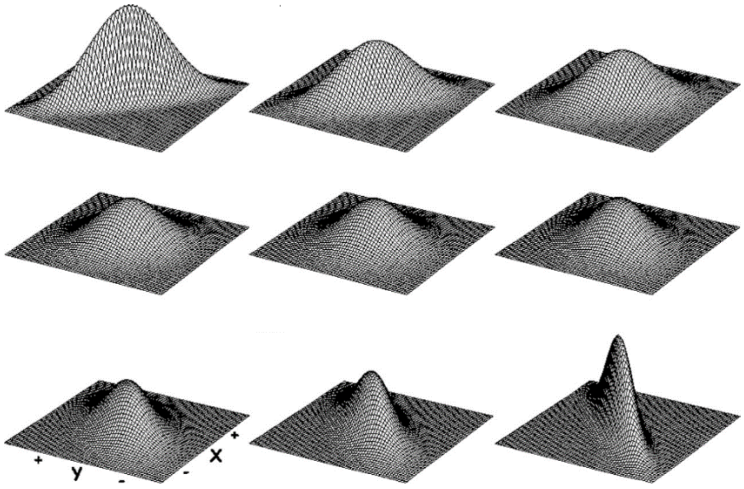
We can generalize this idea by writing the **covariance of X and Y** as

$$\text{cov}(X,Y) = E[(X - E[X])(Y - E[Y])] = \int\int (s - E[X])(w - E[Y]) f(s,w)dsdw$$

Notice how that if Y = X then the covariance of Y and X is simply the variance of X. The covariance of Y and X is measuring the extent to which the X and Y vary linearly with each other. That is, when one goes up, does the other go up, down, or no change, in a probabilistic or expectational way. Also, is this variation between them like a line – is it linear. You should be able to see that covariance is incredibly important in modeling real world phenomena.

How might X and Y co-vary? If they have no relation at all, then we would expect that the covariance would be zero. In fact, if X and Y are statistically independent, the covariance will be zero.[1] If Y goes down when X goes up, then the covariance will be negative (an inverse relation). If Y goes up when X goes up, then the covariance will be positive (a direct relation).

Below, we have a very clear set of graphs that show nine bivariate normal joint pdfs.[2] As you can see, each of these joint pdfs is an example of a joint pdf for random variables X and Y. They have been graphed assuming normality of the random variables X and Y. However, they are quite different. They progress from northwest to southeast as negative, zero, and positive covariance random variables, respectively. The volumes below



these pdfs measure the probability of X and Y being in a particular region. Note how that, in a probabilistic way, the top left and bottom right sub-graphs show a definite systematic and linear relation between X and Y. The top left is a negatively sloped probability relation, while the bottom right has a positive probability relation. However, the middle three graphs do not seem to favor a negative or positive relation between X and Y – they have a near zero covariance.

---

[1] The converse of this statement is not true in general, but will be true if the two random variables are normal random variables. Zero correlation implies statistical independence when the two variables are normally distributed.

[2] See http://advan.physiology.org/content/34/4/186

Now, let's turn to correlation. We define correlation by using three things – the covariance of X and Y, the variance of X and the variance of Y. This is a theoretical object we are defining. It is calculated using the pdf only – no data need be used. The correlation coefficient is defined as

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}}$$

It is not hard to show that this measure lies between -1 and 1. It is equal to zero if and only if the cov(X,Y) = 0. The correlation coefficient is an unknown number. It is related to the p,arameters in a joint pdf between X and Y, but make no mistake about it, it is still just a fixed number.

Now, how might we go about estimating this correlation coefficient? That is, if we have observed data on X and Y, how can we use this data to guess the unknown theoretical number, $\rho_{X,Y}$? And, is it possible to test whether the unknown correlation coefficient, $\rho_{X,Y}$, is equal to zero? These are two important questions – the first is a problem of estimation, while the second is a problem of hypothesis testing.

To estimate the correlation coefficient, we merely estimate the components that make up the correlation coefficient; namely, the covariance and the two variances. Here is how we estimate these

$$\hat{\text{cov}}(X,Y) = \sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y}) / (N-1)$$

$$\hat{\text{var}}(X) = \sum_{k=1}^{N}(x_k - \bar{x})^2 / (N-1)$$

$$\hat{\text{var}}(Y) = \sum_{k=1}^{N}(y_k - \bar{y})^2 / (N-1)$$

The estimated correlation coefficient can now be calculated as

$$\hat{\rho}_{X,Y} = \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{N}(x_k - \bar{x})^2 \sum_{k=1}^{N}(y_k - \bar{y})^2}}$$

Note how that N-1 cancels out in the numerator and denominator. It is important to realize that this is not the correlation coefficient. It is the *estimated* correlation coefficient. The real correlation coefficient is theoretical. It is unknown and it is unknowable. Even with millions of observations on X and Y we could not be sure of the real value of the correlation coefficient. The estimated coefficient uses a sample, and the error between the real and estimated coefficient is called the sampling error. There will always be some sampling error in our estimate. However, the way we have decided to estimate the correlation

coefficient will make this sampling error (theoretical – estimated) smaller in larger samples. This is why statisticians always want more data.

Next, testing the hypothesis H_o: $\rho_{X,Y} = 0$ is actually very complicated. The problem is that in order to carry out the steps of the Neyman-Pearson method, we need to know the pdf of the test statistic $\hat{\rho}_{X,Y}$ under the assumption that H_o is true. Research on this question began in the early 20th century. It was shown that if X and Y are bivariate normal random variables, then the following is true.

$$\hat{T} = \hat{\rho}_{X,Y}\sqrt{\frac{N-2}{1-\hat{\rho}_{X,Y}^2}}$$

has the symmetric Student's t-distribution with N-2 degrees of freedom where we assume H_o is true. It will be approximately true even if the variables are not bivariate normal.

We can now find the 0.05 critical values of this t-density by looking at a table of the t-distribution, shown on the next page. With N = 30 observations, the degrees of freedom are df = N-2 = 28 and the 0.05 critical values are therefore equal to -2.04841 and 2.04841. These two critical values correspond to -0.36 and 0.36 for $\hat{\rho}_{X,Y}$. Thus, if the measured value of $|\hat{\rho}_{X,Y}| > 0.36$ then we should reject the null hypothesis that $\rho_{X,Y} = 0$. This is how we often test if two variables have a statistically significant correlation between them.

Problems:

(P1) Why do we need two random variables to discuss correlation?

(P2) If the covariance between X and Y is positive, then the correlation must be positive. Explain

(P3) What is a "joint" pdf for random variables X and Y?

(P4) What *three* things are used when we estimate the correlation coefficient from data?

(P5) The correlation coefficient and estimated correlation coefficient are different. Explain.

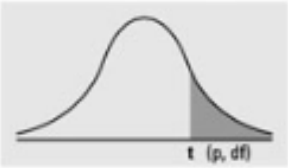
(P6) Suppose we have the following data

t:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18 = N
x:	1	3	6	2	4	2	4	6	5	2	3	3	1	0	8	3	2	5
y:	2	5	3	2	5	3	2	9	6	4	2	3	0	1	6	2	3	6

Now estimate the sample means of X and Y, the sample variances of X and Y, the sample covariance of X and Y, the sample correlation coefficient, and test the hypothesis

H_o: $\rho = 0$. You may use Excel.

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%