6.7: Sample Correlation and Regression

Descriptive Theory

Recall the basic model of statistics: we have a population of objects of interest, and we have various measurements (variables) that we make on these objects. We select objects from the population and record the variables for the objects in the sample; these become our data. Our first discussion is from a purely descriptive point of view. That is, we do not assume that the data are generated by an underlying probability distribution. But as always, remember that the data themselves define a probability distribution, namely the *empirical distribution* that assigns equal probability to each data point.

Suppose that x and y are real-valued variables for a population, and that $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$ is an observed sample of size n from (x, y). We will let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ denote the sample from x and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ the sample from y. In this section, we are interested in statistics that are *measures of association* between the $\boldsymbol{x}$ and $\boldsymbol{y}$, and in finding the line (or other curve) that best fits the data.

Recall that the sample means are

$$m(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad m(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{6.7.1}$$

and the sample variances are

$$s^2(\boldsymbol{x}) = \frac{1}{n-1}\sum_{i=1}^{n} [x_i - m(\boldsymbol{x})]^2, \quad s^2(\boldsymbol{y}) = \frac{1}{n-1}\sum_{i=1}^{n} [y_i - m(\boldsymbol{y})]^2 \tag{6.7.2}$$

Scatterplots

Often, the first step in *exploratory data analysis* is to draw a graph of the points; this is called a *scatterplot* an can give a visual sense of the statistical realtionship between the variables.

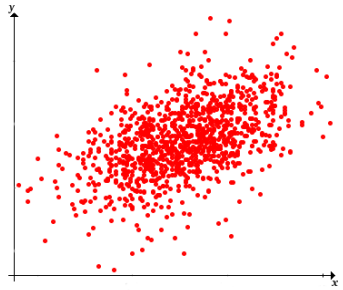

Figure 6.7.1: A scatterplot

In particular, we are interested in whether the cloud of points seems to show a linear trend or whether some nonlinear curve might fit the cloud of points. We are interested in the extent to which one variable x can be used to predict the other variable y.

Defintions

Our next goal is to define statistics that measure the association between the x and y data.

The *sample covariance* is defined to be

$$s(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n-1}\sum_{i=1}^{n} [x_i - m(\boldsymbol{x})][y_i - m(\boldsymbol{y})] \tag{6.7.3}$$

Assuming that the data vectors are not constant, so that the standard deviations are positive, the *sample correlation* is defined to be

$$r(\boldsymbol{x}, \boldsymbol{y}) = \frac{s(\boldsymbol{x}, \boldsymbol{y})}{s(\boldsymbol{x}) s(\boldsymbol{y})} \tag{6.7.4}$$

Note that the sample covariance is an average of the product of the deviations of the x and y data from their means. Thus, the physical unit of the sample covariance is the product of the units of x and y. Correlation is a standardized version of covariance. In particular, correlation is dimensionless (has no physical units), since the covariance in the numerator and the product of the standard devations in the denominator have the same units (the product of the units of x and y). Note also that covariance and correlation have the same *sign*: positive, negative, or zero. In the first case, the data $\boldsymbol{x}$ and $\boldsymbol{y}$ are said to be *positively correlated*; in the second case $\boldsymbol{x}$ and $\boldsymbol{y}$ are said to be *negatively correlated*; and in the third case $\boldsymbol{x}$ and $\boldsymbol{y}$ are said to be *uncorrelated*

To see that the sample covariance is a measure of association, recall first that the point $(m(\boldsymbol{x}), m(\boldsymbol{y}))$ is a measure of the center of the bivariate data. Indeed, if each point is the location of a unit mass, then $(m(\boldsymbol{x}), m(\boldsymbol{y}))$ is the *center of mass* as defined in physics. Horizontal and vertical lines through this center point divide the plane into four quadrants. The product deviation $[x_i - m(\boldsymbol{x})][y_i - m(\boldsymbol{y})]$ is positive in the first and third quadrants and negative in the second and fourth quadrants. After we study linear regression below, we will have a much deeper sense of what covariance measures.

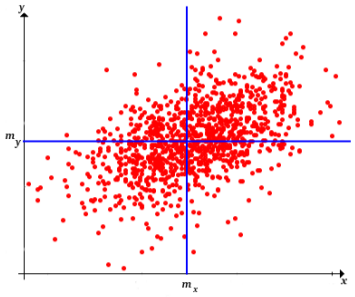

Figure 6.7.2: Scatterplot with means

You may be perplexed that we average the product deviations by dividing by $n-1$ rather than n. The best explanation is that in the probability model discussed below, the sample covariance is an unbiased estimator of the distribution covariance. However, the mode of averaging can also be understood in terms of *degrees of freedom*, as was done for sample variance. Initially, we have $2n$ degrees of freedom in the bivariate data. We lose two by computing the sample means $m(\boldsymbol{x})$ and $m(\boldsymbol{y})$. Of the remaining $2n-2$ degrees of freedom, we lose $n-1$ by computing the product deviations. Thus, we are left with $n-1$ degrees of freedom total. As is typical in statistics, we average not by dividing by the number of terms in the sum but rather by the number of degrees of freedom in those terms. However, from a purely descriptive point of view, it would also be reasonable to divide by n.

Recall that there is a natural probability distribution associated with the data, namely the empirical distribution that gives probability $\frac{1}{n}$ to each data point (x_i, y_i). (Thus, if these points are distinct this is the discrete uniform distribution on the data.) The sample means are simply the expected values of this bivariate distribution, and except for a constant multiple (dividing by $n-1$ rather than n), the sample variances are simply the variances of this bivarite distribution. Similarly, except for a constant multiple (again dividing by $n-1$ rather than n), the sample covariance is the covariance of the bivariate distribution and the sample correlation is the correlation of the bivariate distribution. All of the following results in our discussion of descriptive statistics are actually special cases of more general results for probability distributions.

Properties of Covariance

The next few exercises establish some essential properties of sample covariance. As usual, bold symbols denote samples of a fixed size n from the corresponding population variables (that is, vectors of length n), while symbols in regular type denote real numbers. Our first result is a formula for sample covariance that is sometimes better than the definition for computational purposes. To state the result succinctly, let $\boldsymbol{xy} = (x_1 y_1, x_2 y_2, \ldots, x_n y_n)$ denote the sample from the product variable xy.

The sample covariance can be computed as follows:

$$s(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n-1} \sum_{i=1}^{n} x_i y_i - \frac{n}{n-1} m(\boldsymbol{x}) m(\boldsymbol{y}) = \frac{n}{n-1} [m(\boldsymbol{xy}) - m(\boldsymbol{x}) m(\boldsymbol{y})] \tag{6.7.5}$$

Proof

Note that

$$\sum_{i=1}^{n} [(x_i - m(\boldsymbol{x})][y_i - m(\boldsymbol{y})] = \sum_{i=1}^{n} [x_i y_i - x_i m(\boldsymbol{y}) - y_i m(\boldsymbol{x}) + m(\boldsymbol{x}) m(\boldsymbol{y})] \tag{6.7.6}$$

$$= \sum_{i=1}^{n} x_i y_i - m(\boldsymbol{y}) \sum_{i=1}^{n} x_i - m(\boldsymbol{x}) \sum_{i=1}^{n} y_i + n m(\boldsymbol{x}) m(\boldsymbol{y}) \tag{6.7.7}$$

$$= \sum_{i=1}^{n} x_i y_i - n m(\boldsymbol{y}) m(\boldsymbol{x}) - n m(\boldsymbol{x}) m(\boldsymbol{y}) + n m(\boldsymbol{x}) m(\boldsymbol{y}) \tag{6.7.8}$$

$$= \sum_{i=1}^{n} x_i y_i - n m(\boldsymbol{x}) m(\boldsymbol{y}) \tag{6.7.9}$$

The following theorem gives another formula for the sample covariance, one that does not require the computation of intermediate statistics.

The sample covariance can be computed as follows:

$$s(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)(y_i - y_j) \tag{6.7.10}$$

Proof

Note that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)(y_i - y_j) = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} [x_i - m(\boldsymbol{x}) + m(\boldsymbol{x}) - x_j][y_i - m(\boldsymbol{y}) + m(\boldsymbol{y}) - y_j] \tag{6.7.11}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} ([(x_i - m(\boldsymbol{x})][y_i - m(\boldsymbol{y})] + [x_i - m(\boldsymbol{x})][m(\boldsymbol{y}) - y_j] + [m(\boldsymbol{x}) - x_j][y_i - m(\boldsymbol{y})] + [m(\boldsymbol{x}) - x_j][m(\boldsymbol{y}) - y_j]) \tag{6.7.12}$$

We compute the sums term by term. The first is

$$n \sum_{i=1}^{n} [x_i - m(\boldsymbol{x})][y_i - m(\boldsymbol{y})] \tag{6.7.13}$$

The second two sums are 0. The last sum is

$$n\sum_{j=1}^{n}[m(\boldsymbol{x})-x_j][m(\boldsymbol{y})-y_j]=n\sum_{i=1}^{n}[x_i-m(\boldsymbol{x})][y_i-m(\boldsymbol{y})] \tag{6.7.14}$$

Dividing the entire sum by $2n(n-1)$ results in $\operatorname{cov}(\boldsymbol{x},\boldsymbol{y})$.

As the name suggests, sample covariance generalizes sample variance.

$s(\boldsymbol{x},\boldsymbol{x})=s^2(\boldsymbol{x})$.

In light of the previous theorem, we can now see that the first computational formula and the second computational formula above generalize the computational formulas for sample variance. Clearly, sample covariance is *symmetric*.

$s(\boldsymbol{x},\boldsymbol{y})=s(\boldsymbol{y},\boldsymbol{x})$.

Sample covariance is linear in the first argument with the second argument fixed.

If $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ are data vectors from population variables x, y, and z, respectively, and if c is a constant, then

1. $s(\boldsymbol{x}+\boldsymbol{y},\boldsymbol{z})=s(\boldsymbol{x},\boldsymbol{z})+s(\boldsymbol{y},\boldsymbol{z})$
2. $s(c\boldsymbol{x},\boldsymbol{y})=cs(\boldsymbol{x},\boldsymbol{y})$

Proof

1. Recall that $m(\boldsymbol{x}+\boldsymbol{y})=m(\boldsymbol{x})+m(\boldsymbol{y})$. Hence

$$\begin{aligned} s(\boldsymbol{x}+\boldsymbol{y},\boldsymbol{z}) &= \frac{1}{n-1}\sum_{i=1}^{n}[x_i+y_i-m(\boldsymbol{x}+\boldsymbol{y})][z_i-m(\boldsymbol{z})] && (6.7.15)\\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left([x_i-m(\boldsymbol{x})]+[y_i-m(\boldsymbol{y})]\right)[z_i-m(\boldsymbol{z})] && (6.7.16)\\ &= \frac{1}{n-1}\sum_{i=1}^{n}[x_i-m(\boldsymbol{x})][z_i-m(\boldsymbol{z})]+\frac{1}{n-1}\sum_{i=1}^{n}[y_i-m(\boldsymbol{y})][z_i-m(\boldsymbol{z})] && (6.7.17)\\ &= s(\boldsymbol{x},\boldsymbol{z})+s(\boldsymbol{y},\boldsymbol{z}) && (6.7.18)\end{aligned}$$

2. Recall that $m(c\boldsymbol{x})=cm(\boldsymbol{x})$. Hence

$$\begin{aligned} s(c\boldsymbol{x},\boldsymbol{y}) &= \frac{1}{n-1}\sum_{i=1}^{n}[cx_i-m(c\boldsymbol{x})][y_i-m(\boldsymbol{y})] && (6.7.19)\\ &= \frac{1}{n-1}\sum_{i=1}^{n}[cx_i-cm(\boldsymbol{x})][y_i-m(\boldsymbol{y})]=cs(\boldsymbol{x},\boldsymbol{y}) && (6.7.20)\end{aligned}$$

By symmetry, sample covariance is also linear in the second argument with the first argument fixed, and hence is *bilinear*. The general version of the bilinear property is given in the following theorem:

Suppose that $\boldsymbol{x}_i$ is a data vector from a population variable x_i for $i\in\{1,2,\ldots,k\}$ and that $\boldsymbol{y}_j$ is a data vector from a population variable y_j for $j\in\{1,2,\ldots,l\}$. Suppose also that $a_1,\ a_2,\ldots,\ a_k$ and $b_1,\ b_2,\ldots,b_l$ are constants. Then

$$s\left(\sum_{i=1}^{k}a_i\,\boldsymbol{x}_i,\sum_{j=1}^{l}b_j\,\boldsymbol{y}_j\right)=\sum_{i=1}^{k}\sum_{j=1}^{l}a_i\,b_j\,s(\boldsymbol{x}_i,\boldsymbol{y}_j) \tag{6.7.21}$$

A special case of the bilinear property provides a nice way to compute the sample variance of a sum.

$s^2(\boldsymbol{x}+\boldsymbol{y})=s^2(\boldsymbol{x})+2s(\boldsymbol{x},\boldsymbol{y})+s^2(\boldsymbol{y})$.

Proof

From the preceding results,

$$\begin{aligned} s^2(\boldsymbol{x}+\boldsymbol{y}) &= s(\boldsymbol{x}+\boldsymbol{y},\boldsymbol{x}+\boldsymbol{y})=s(\boldsymbol{x},\boldsymbol{x})+s(\boldsymbol{x},\boldsymbol{y})+s(\boldsymbol{y},\boldsymbol{x})+s(\boldsymbol{y},\boldsymbol{y}) && (6.7.22)\\ &= s^2(\boldsymbol{x})+2s(\boldsymbol{x},\boldsymbol{y})+s^2(\boldsymbol{y}) && (6.7.23)\end{aligned}$$

The generalization of this result to sums of three or more vectors is completely straightforward: namely, the sample variance of a sum is the sum of all of the pairwise sample covariances. Note that the sample variance of a sum can be greater than, less than, or equal to the sum of the sample variances, depending on the sign and magnitude of the pure covariance term. In particular, if the vectors are pairwise uncorrelated, then the variance of the sum is the sum of the variances.

If $\boldsymbol{c}$ is a constant data set then $s(\boldsymbol{x},\boldsymbol{c})=0$.

Proof

This follows directly from the definition. If $c_i=c$ for each i, then $m(\boldsymbol{c})=c$ and hence $c_i-m(\boldsymbol{c})=0$ for each i.

Combining the result in the last exercise with the bilinear property, we see that covariance is unchanged if constants are added to the data sets. That is, if $\boldsymbol{c}$ and $\boldsymbol{d}$ are constant vectors then $s(\boldsymbol{x}+\boldsymbol{c},\boldsymbol{y}+\boldsymbol{d})=s(\boldsymbol{x},\boldsymbol{y})$.

Properties of Correlation

A few simple properties of correlation are given next. Most of these follow easily from the corresponding properties of covariance. First, recall that the standard scores of x_i and y_i are, respectively,

$$u_i = \frac{x_i - m(\boldsymbol{x})}{s(\boldsymbol{x})}, \quad v_i = \frac{y_i - m(\boldsymbol{y})}{s(\boldsymbol{y})} \tag{6.7.24}$$

The standard scores from a data set are dimensionless quantities that have mean 0 and variance 1.

The correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$ is the covariance of their standard scores $\boldsymbol{u}$ and $\boldsymbol{v}$. That is, $r(\boldsymbol{x}, \boldsymbol{y}) = s(\boldsymbol{u}, \boldsymbol{v})$.

Proof

In vector notation, note that

$$\boldsymbol{u} = \frac{1}{s(\boldsymbol{x})}[\boldsymbol{x} - m(\boldsymbol{x})], \quad \boldsymbol{v} = \frac{1}{s(\boldsymbol{y})}[\boldsymbol{y} - m(\boldsymbol{y})] \tag{6.7.25}$$

Hence the result follows immediatedly from properties of covariance:

$$s(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{s(\boldsymbol{x}) s(\boldsymbol{y})} s(\boldsymbol{x}, \boldsymbol{y}) = r(\boldsymbol{x}, \boldsymbol{y}) \tag{6.7.26}$$

Correlation is symmetric.

$r(\boldsymbol{x}, \boldsymbol{y}) = r(\boldsymbol{y}, \boldsymbol{x})$.

Unlike covariance, correlation is unaffected by multiplying one of the data sets by a positive constant (recall that this can always be thought of as a change of scale in the underlying variable). On the other hand, muliplying a data set by a negative constant changes the sign of the correlation.

If $c \neq 0$ is a constant then

1. $r(c\boldsymbol{x}, \boldsymbol{y}) = r(\boldsymbol{x}, \boldsymbol{y})$ if $c > 0$
2. $r(c\boldsymbol{x}, \boldsymbol{y}) = -r(\boldsymbol{x}, \boldsymbol{y})$ if $c < 0$

Proof

By definition and from the scaling property of covariance,

$$r(c\boldsymbol{x}, \boldsymbol{y}) = \frac{s(c\boldsymbol{x}, \boldsymbol{y})}{s(c\boldsymbol{x}) s(\boldsymbol{y})} = \frac{cs(\boldsymbol{x}, \boldsymbol{y})}{|c| \, s(\boldsymbol{x}) s(\boldsymbol{y})} = \frac{c}{|c|} r(\boldsymbol{x}, \boldsymbol{y}) \tag{6.7.27}$$

and of course, $c / |c| = 1$ if $c > 0$ and $c / |c| = -1$ if $c < 0$.

Like covariance, correlation is unaffected by adding constants to the data sets. Adding a constant to a data set often corresponds to a *change of location*.

If $\boldsymbol{c}$ and $\boldsymbol{d}$ are constant vectors then $r(\boldsymbol{x} + \boldsymbol{c}, \boldsymbol{y} + \boldsymbol{d}) = r(\boldsymbol{x}, \boldsymbol{y})$.

Proof

This result follows directly from the corresponding properties of covariance and standard deviation:

$$r(\boldsymbol{x} + \boldsymbol{c}, \boldsymbol{y} + \boldsymbol{d}) = \frac{s(\boldsymbol{x} + \boldsymbol{c}, \boldsymbol{y} + \boldsymbol{d})}{s(\boldsymbol{x} + \boldsymbol{c}) s(\boldsymbol{y} + \boldsymbol{d})} = \frac{s(\boldsymbol{x}, \boldsymbol{y})}{s(\boldsymbol{x}) s(\boldsymbol{y})} = r(\boldsymbol{x}, \boldsymbol{y}) \tag{6.7.28}$$

The last couple of properties reinforce the fact that correlation is a standardized measure of association that is not affected by changing the units of measurement. In the first Challenger data set, for example, the variables of interest are temperature at time of launch (in degrees Fahrenheit) and O-ring erosion (in millimeters). The correlation between these variables is of critical importance. If we were to measure temperature in degrees Celsius and O-ring erosion in inches, the correlation between the two variables would be unchanged.

The most important properties of correlation arise from studying the line that best fits the data, our next topic.

Linear Regression

We are interested in finding the line $y = a + bx$ that best fits the sample points $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$. This is a basic and important problem in many areas of mathematics, not just statistics. We think of x as the *predictor variable* and y as the *response variable*. Thus, the term *best* means that we want to find the line (that is, find the coefficients a and b) that minimizes the average of the squared errors between the actual y values in our data and the predicted y values:

$$\text{mse}(a, b) = \frac{1}{n - 1} \sum_{i=1}^{n} [y_i - (a + b \, x_i)]^2 \tag{6.7.29}$$

Note that the minimizing value of (a, b) would be the same if the function were simply the sum of the squared errors, of if we averaged by dividing by n rather than $n - 1$, or if we used the square root of any of these functions. Of course that actual *minimum value* of the function would be different if we changed the function, but again, not the point (a, b) where the minimum occurs. Our particular choice of mse as the error function is best for statistical purposes. Finding (a, b) that minimize mse is a standard problem in calculus.

The graph of mse is a paraboloid opening upward. The function mse is minimized when

$$b(\boldsymbol{x}, \boldsymbol{y}) = \frac{s(\boldsymbol{x}, \boldsymbol{y})}{s^2(\boldsymbol{x})} \tag{6.7.30}$$

$$a(\boldsymbol{x}, \boldsymbol{y}) = m(\boldsymbol{y}) - b(\boldsymbol{x}, \boldsymbol{y}) m(\boldsymbol{x}) = m(\boldsymbol{y}) - \frac{s(\boldsymbol{x}, \boldsymbol{y})}{s^2(\boldsymbol{x})} m(\boldsymbol{x}) \tag{6.7.31}$$

Proof

We can tell from the algebraic form of mse that the graph is a paraboloid opening upward. To find the unique point that minimizes mse, note that

$$\frac{\partial}{\partial a}\text{mse}(a,b) = \frac{1}{n-1}\sum 2[y_i-(a+bx_i)](-1) = \frac{2}{n-1}[-\sum_{i=1}^n y_i + na + b\sum_{i=1}^n x_i] \tag{6.7.32}$$

$$\frac{\partial}{\partial b}\text{mse}(a,b) = \frac{1}{n-1}\sum 2[y_i-(a+bx_i)](-x_i) = \frac{2}{n-1}[-\sum_{i=1}^n x_i y_i + a\sum_{i=1}^n x_i + b\sum_{i=1}^n x_i^2] \tag{6.7.33}$$

Solving $\frac{\partial}{\partial a}\text{mse}(a,b) = 0$, gives $a = m(\boldsymbol{y}) - bm(\boldsymbol{x})$. Substituting this into $\frac{\partial}{\partial b}\text{mse}(a,b) = 0$ and solving for b gives

$$b = \frac{n[m(\boldsymbol{xy}) - m(\boldsymbol{x})m(\boldsymbol{y})]}{n[m(\boldsymbol{x}^2) - m^2(\boldsymbol{x})]} \tag{6.7.34}$$

Dividing the numerator and denominator in the last expression by $n-1$ and using the computational formula above, we see that $b = s(\boldsymbol{x},\boldsymbol{y})/s^2(\boldsymbol{x})$.

Of course, the optimal values of a and b are *statistics*, that is, functions of the data. Thus the *sample regression line* is

$$y = m(\boldsymbol{y}) + \frac{s(\boldsymbol{x},\boldsymbol{y})}{s^2(\boldsymbol{x})}[x - m(\boldsymbol{x})] \tag{6.7.35}$$

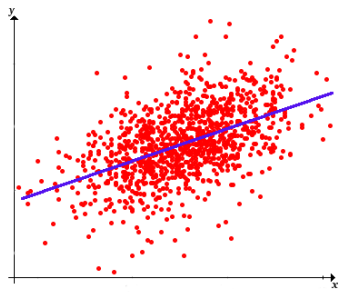

Figure 6.7.3: Scatterplot with regression line

Note that the regression line passes through the point $(m(\boldsymbol{x}), m(\boldsymbol{y}))$, the center of the sample of points.

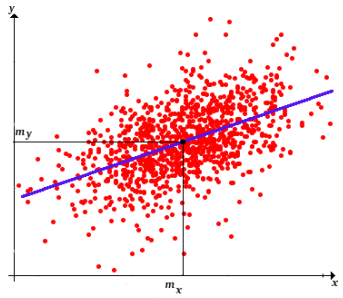

Figure 6.7.4: The regression line passes through the center

The minimum mean square error is

$$\text{mse}\left[a(\boldsymbol{x},\boldsymbol{y}), b(\boldsymbol{x},\boldsymbol{y})\right] = s(\boldsymbol{y})^2\left[1 - r^2(\boldsymbol{x},\boldsymbol{y})\right] \tag{6.7.36}$$

Proof

This follows from substituting $a(\boldsymbol{x},\boldsymbol{y})$ $b(\boldsymbol{x},\boldsymbol{y})$ into mse and simplifying.

Sample correlation and covariance satisfy the following properties.

1. $-1 \le r(\boldsymbol{x},\boldsymbol{y}) \le 1$
2. $-s(\boldsymbol{x})s(\boldsymbol{y}) \le s(\boldsymbol{x},\boldsymbol{y}) \le s(\boldsymbol{x})s(\boldsymbol{y})$
3. $r(\boldsymbol{x},\boldsymbol{y}) = -1$ if and only if the sample points lie on a line with negative slope.
4. $r(\boldsymbol{x},\boldsymbol{y}) = 1$ if and only if the sample points lie on a line with positive slope.

Proof

Note that $\text{mse} \ge 0$ and hence from the previous theorem, we must have $r^2(\boldsymbol{x},\boldsymbol{y}) \le 1$. This is equivalent to part (a), which in turn, from the definition of sample correlation, is equivalent to part (b). For parts (c) and (d), note that $\text{mse}(a,b) = 0$ if and only if $y_i = a + bx_i$ for each i, and moreover, $b(\boldsymbol{x},\boldsymbol{y})$ has the same sign as $r(\boldsymbol{x},\boldsymbol{y})$.

Thus, we now see in a deeper way that the sample covariance and correlation measure the degree of linearity of the sample points. Recall from our discussion of measures of center and spread that the constant a that minimizes

$$\text{mse}(a) = \frac{1}{n-1}\sum_{i=1}^n (y_i - a)^2 \tag{6.7.37}$$

is the sample mean $m(\boldsymbol{y})$, and the minimum value of the mean square error is the sample variance $s^2(\boldsymbol{y})$. Thus, the difference between this value of the mean square error and the one above, namely $s^2(\boldsymbol{y})r^2(\boldsymbol{x},\boldsymbol{y})$ is the reduction in the variability of the y data when the linear term in x is added to the predictor. The fractional reduction is $r^2(\boldsymbol{x},\boldsymbol{y})$,

and hence this statistics is called the (sample) *coefficient of determination*. Note that if the data vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are uncorrelated, then x has no value as a predictor of y; the regression line in this case is the horizontal line $y = m(\boldsymbol{y})$ and the mean square error is $s^2(\boldsymbol{y})$.

The choice of predictor and response variables is important.

The sample regression line with predictor variable x and response variable y is not the same as the sample regression line with predictor variable y and response variable x, except in the extreme case $r(\boldsymbol{x}, \boldsymbol{y}) = \pm 1$ where the sample points all lie on a line.

Residuals

The difference between the actual y value of a data point and the value predicted by the regression line is called the *residual* of that data point. Thus, the residual corresponding to (x_i, y_i) is $d_i = y_i - \hat{y}_i$ where $\hat{y}_i$ is the regression line at x_i:

$$\hat{y}_i = m(\boldsymbol{y}) + \frac{s(\boldsymbol{x}, \boldsymbol{y})}{s(\boldsymbol{x})^2}[x_i - m(\boldsymbol{x})] \tag{6.7.38}$$

Note that the predicted value $\hat{y}_i$ and the residual d_i are *statistics*, that is, functions of the data $(\boldsymbol{x}, \boldsymbol{y})$, but we are suppressing this in the notation for simplicity.

The residuals sum to 0: $\sum_{i=1}^n d_i = 0$.

Proof

This follows from the definition, and is a restatement of the fact that the regression line passes through the center of the data set $(m(\boldsymbol{x}), m(\boldsymbol{y}))$.

Various plots of the residuals can help one understand the relationship between the x and y data. Some of the more common are given in the following definition:

Residual plots

1. A plot of (i, d_i) for $i \in \{1, 2, \ldots, n\}$, that is, a plot of indices versus residuals.
2. A plot of (x_i, d_i) for $i \in \{1, 2, \ldots, n\}$, that is, a plot of x values versus residuals.
3. A plot of (d_i, y_i) for $i \in \{1, 2, \ldots, n\}$, that is, a plot of residuals versus actual y values.
4. A plot of $(d_i, \hat{y}_i)$ for $i \in \{1, 2, \ldots, n\}$, that is a plot of residuals versus predicted y values.
5. A histogram of the residuals $(d_1, d_2, \ldots, d_n)$.

Sums of Squares

For our next discussion, we will re-interpret the minimum mean square error formulat above. Here are the new definitions:

Sums of squares

1. $\text{sst}(\boldsymbol{y}) = \sum_{i=1}^n [y_i - m(\boldsymbol{y})]^2$ is the *total sum of squares*.
2. $\text{ssr}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^n [\hat{y}_i - m(\boldsymbol{y})]^2$ is the *regression sum of squares*
3. $\text{sse}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ is the *error sum of squares*.

Note that $\text{sst}(\boldsymbol{y})$ is simply $n - 1$ times the variance $s^2(\boldsymbol{y})$ and is the total of the sums of the squares of the deviations of the y values from the mean of the y values. Similarly, $\text{sse}(\boldsymbol{x}, \boldsymbol{y})$ is simply $n - 1$ times the minimum mean square error given above. Of course, $\text{sst}(\boldsymbol{y})$ has $n - 1$ degrees of freedom, while $\text{sse}(\boldsymbol{x}, \boldsymbol{y})$ has $n - 2$ degrees of freedom and $\text{ssr}(\boldsymbol{x}, \boldsymbol{y})$ a single degree of freedom. The total sum of squares is the sum of the regression sum of squares and the error sum of squares:

The sums of squares are related as follows:

1. $\text{ssr}(\boldsymbol{x}, \boldsymbol{y}) = r^2(\boldsymbol{x}, \boldsymbol{y})\text{sst}(\boldsymbol{y})$
2. $\text{sst}(\boldsymbol{y}) = \text{ssr}(\boldsymbol{x}, \boldsymbol{y}) + \text{sse}(\boldsymbol{x}, \boldsymbol{y})$

Proof

By definition of sst and r, we see that $r^2(\boldsymbol{x}, \boldsymbol{y})\text{sst}(\boldsymbol{y}) = s^2(\boldsymbol{x}, \boldsymbol{y}) / s^2(\boldsymbol{x})$. But from the regression equation,

$$[\hat{y}_i - m(\boldsymbol{y})]^2 = \frac{s^2(\boldsymbol{x}, \boldsymbol{y})}{s^4(\boldsymbol{x})}[x_i - m(\boldsymbol{x})]^2 \tag{6.7.39}$$

Summing over i gives

$$\text{ssr}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^n [\hat{y}_i - m(\boldsymbol{y})]^2 = \frac{s^2(\boldsymbol{x}, \boldsymbol{y})}{s^2(\boldsymbol{x})} \tag{6.7.40}$$

Hence $\text{ssr}(\boldsymbol{x}, \boldsymbol{y}) = r^2(\boldsymbol{x}, \boldsymbol{y})\text{sst}(\boldsymbol{y})$. Finally, multiplying the result above by $n - 1$ gives $\text{sse}(\boldsymbol{x}, \boldsymbol{y}) = \text{sst}(\boldsymbol{y}) - r^2(\boldsymbol{x}, \boldsymbol{y})\text{sst}(\boldsymbol{y}) = \text{sst}(\boldsymbol{y}) - \text{ssr}(\boldsymbol{x}, \boldsymbol{y})$.

Note that $r^2(\boldsymbol{x}, \boldsymbol{y}) = \text{ssr}(\boldsymbol{x}, \boldsymbol{y}) / \text{sst}(\boldsymbol{y})$, so once again, $r^2(\boldsymbol{x}, \boldsymbol{y})$ is the coefficient of determination—the proportion of the variability in the y data explained by the x data. We can average sse by dividing by its degrees of freedom and then take the square root to obtain a standard error:

The *standard error of estimate* is

$$\text{se}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{\text{sse}(\boldsymbol{x}, \boldsymbol{y})}{n - 2}} \tag{6.7.41}$$

This really is a *standard* error in the same sense as a *standard* deviation. It's an average of the errors of sorts, but in the root mean square sense.

Finally, it's important to note that linear regression is a much more powerful idea than might first appear, and in fact the term *linear* can be a bit misleading. By applying various transformations to y or x or both, we can fit a variety of two-parameter curves to the given data $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$. Some of the most common transformations are explored in the exercises below.

Probability Theory

We continue our discussion of sample covariance, correlation, and regression but now from the more interesting point of view that the variables are random. Specifically, suppose that we have a basic random experiment, and that X and Y are real-valued random variables for the experiment. Equivalently, (X, Y) is a random vector taking values in $\mathbb{R}^2$. Let $\mu = \mathbb{E}(X)$ and $\nu = \mathbb{E}(Y)$ denote the distribution means, $\sigma^2 = \text{var}(X)$ and $\tau^2 = \text{var}(Y)$ the distribution variances, and let $\delta = \text{cov}(X, Y)$ denote the distribution covariance, so that the distribution correlation is

$$\rho = \text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\text{sd}(X)\,\text{sd}(Y)} = \frac{\delta}{\sigma\,\tau} \tag{6.7.42}$$

We will also need some higher order moments. Let $\sigma_4 = \mathbb{E}\left[(X - \mu)^4\right]$, $\tau_4 = \mathbb{E}\left[(Y - \nu)^4\right]$, and $\delta_2 = \mathbb{E}\left[(X - \mu)^2 (Y - \nu)^2\right]$. Naturally, we assume that all of these moments are finite.

Now suppose that we run the basic experiment n times. This creates a compound experiment with a sequence of independent random vectors $((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$ each with the same distribution as (X, Y). In statistical terms, this is a random sample of size n from the distribution of (X, Y). The statistics discussed in previous section are well defined but now they are all random variables. We use the notation established previously, except that we use our usual convention of denoting random variables with capital letters. Of course, the deterministic properties and relations established above still hold. Note that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the distribution of X and $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ is a random sample of size n from the distribution of Y. The main purpose of this subsection is to study the relationship between various statistics from $\boldsymbol{X}$ and $\boldsymbol{Y}$, and to study statistics that are natural estimators of the distribution covariance and correlation.

The Sample Means

Recall that the sample means are

$$M(\boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^n X_i, \quad M(\boldsymbol{Y}) = \frac{1}{n} \sum_{i=1}^n Y_i \tag{6.7.43}$$

From the sections on the law of large numbers and the central limit theorem, we know a great deal about the distributions of $M(\boldsymbol{X})$ and $M(\boldsymbol{Y})$ *individually*. But we need to know more about the *joint distribution*.

The covariance and correlation between $M(\boldsymbol{X})$ and $M(\boldsymbol{Y})$ are

1. $\text{cov}[M(\boldsymbol{X}), M(\boldsymbol{Y})] = \delta / n$
2. $\text{cor}[(M(\boldsymbol{X}), M(\boldsymbol{Y})] = \rho$

Proof

Part (a) follows from the bilinearity of the covariance operator:

$$\text{cov}\left(\frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{j=1}^n Y_j\right) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \text{cov}(X_i, Y_j) \tag{6.7.44}$$

By independence, the terms in the last sum are 0 if $i \neq j$. For $i = j$ the terms are $\text{cov}(X, Y) = \delta$. There are n such terms so $\text{cov}[M(\boldsymbol{X}), M(\boldsymbol{Y})] = \delta / n$. For part (b), recall that $\text{var}[M(\boldsymbol{X})] = \sigma^2 / n$ and $\text{var}[M(\boldsymbol{Y})] = \tau^2 / n$. Hence

$$\text{cor}[M(\boldsymbol{X}), M(\boldsymbol{Y})] = \frac{\delta / n}{(\sigma / \sqrt{n})(\tau / \sqrt{n})} = \frac{\delta}{\sigma \tau} = \rho \tag{6.7.45}$$

Note that the correlation between the sample means is the same as the correlation of the underlying sampling distribution. In particular, the correlation does not depend on the sample size n.

The Sample Variances

Recall that special versions of the sample variances, in the unlikely event that the distribution means are known, are

$$W^2(\boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2, \quad W^2(\boldsymbol{Y}) = \frac{1}{n} \sum_{i=1}^n (Y_i - \nu)^2 \tag{6.7.46}$$

Once again, we have studied these statistics individually, so our emphasis now is on the joint distribution.

The covariance and correlation between $W^2(\boldsymbol{X})$ and $W^2(\boldsymbol{Y})$ are

1. $\text{cov}[W^2(\boldsymbol{X}), W^2(\boldsymbol{Y})] = (\delta_2 - \sigma^2 \tau^2) / n$
2. $\text{cor}[W^2(\boldsymbol{X}), W^2(\boldsymbol{Y})] = (\delta_2 - \sigma^2 \tau^2) / \sqrt{(\sigma_4 - \sigma^4)(\tau_4 - \tau^4)}$

Proof

For part (a), we use the bilinearity of the covariance operator to obtain

$$\text{cov}[W^2(\boldsymbol{X}), W^2(\boldsymbol{Y})] = \text{cov}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2, \frac{1}{n} \sum_{j=1}^n (Y_j - \nu)^2\right) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \text{cov}[(X_i - \mu)^2, (Y_j - \nu)^2] \tag{6.7.47}$$

By independence, the terms in the last sum are 0 when $i \neq j$. When $i = j$ the terms are

$$\text{cov}[(X - \mu)^2 (Y - \nu)^2] = \mathbb{E}[(X - \mu)^2 (Y - \nu)^2] - \mathbb{E}[(X - \mu)^2] \mathbb{E}[(Y - \nu)^2] = \delta_2 - \sigma^2 \tau^2 \tag{6.7.48}$$

There are n such terms, so $\text{cov}[W^2(\boldsymbol{X}), W^2(\boldsymbol{Y})] = (\delta_2 - \sigma^2 \tau^2) / n$. Part (b) follows from part (a) and the variances of $W^2(\boldsymbol{X})$ and $W^2(\boldsymbol{Y})$ from the section on Sample Variance.

Note that the correlation does not dependend on the sample size n. Next, recall that the standard versions of the sample variances are

$$S^2(\boldsymbol{X}) = \frac{1}{n-1}\sum_{i=1}^n [X_i - M(\boldsymbol{X})]^2, \quad S^2(\boldsymbol{Y}) = \frac{1}{n-1}\sum_{i=1}^n [Y_i - M(\boldsymbol{Y})]^2 \tag{6.7.49}$$

The covariance and correlation of the sample variances are

1. $\cov[S^2(\boldsymbol{X}), S^2(\boldsymbol{Y})] = (\delta_2 - \sigma^2\tau^2)/n + 2\delta^2/[n(n-1)]$
2. $\cor[S^2(\boldsymbol{X}), S^2(\boldsymbol{Y})] = [(n-1)(\delta_2 - \sigma^2\tau^2) + 2\delta^2]/\sqrt{[(n-1)\sigma_4 - (n-3)\sigma^4][(n-1)\tau_4 - (n-3)\tau^4]}$

Proof

Recall that

$$S^2(\boldsymbol{X}) = \frac{1}{2n(n-1)}\sum_{i=1}^n\sum_{j=1}^n (X_i - X_j)^2, \quad S^2(\boldsymbol{Y}) = \frac{1}{2n(n-1)}\sum_{k=1}^n\sum_{l=1}^n (Y_k - Y_l)^2 \tag{6.7.50}$$

Hence using the bilinearity of the covariance operator we have

$$\cov[S^2(\boldsymbol{X}), S^2(\boldsymbol{Y})] = \frac{1}{4n^2(n-1)^2}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{l=1}^n \cov[(X_i - X_j)^2, (Y_k - Y_l)^2] \tag{6.7.51}$$

We compute the covariances in this sum by considering disjoint cases:

- $\cov[(X_i - X_j)^2, (Y_k - Y_l)^2] = 0$ if $i = j$ or if $k = l$, and there are $2n^3 - n^2$ such terms.
- $\cov[(X_i - X_j)^2, (Y_k - Y_l)^2] = 0$ by independence if i, j, k, l are distinct, and there are $n(n-1)(n-2)(n-3)$ such terms.
- $\cov[(X_i - X_j)^2, (Y_k - Y_l)^2] = 2\delta_2 - 2\sigma^2\tau^2 + 4\delta^2$ if $i \neq j$ and $\{k, l\} = \{i, j\}$, and there are $2n(n-1)$ such terms.
- $\cov[(X_i - X_j)^2, (Y_k - Y_l)^2] = \delta_2 - \sigma^2\tau^2$ if $i \neq j$, $k \neq l$, and $\#(\{i, j\} \cap \{k, l\}) = 1$, and there are $4n(n-1)(n-2)$ such terms.

Substituting and simplifying gives the result in (a). For (b), we use the definition of correlation and the formulas for $\var[S^2(\boldsymbol{X})]$ and $\var[S^2(\boldsymbol{Y})]$ from the section on the sample variance.

Asymptotically, the correlation between the sample variances is the same as the correlation between the special sample variances given above:

$$\cor\left[S^2(\boldsymbol{X}), S^2(\boldsymbol{Y})\right] \to \frac{\delta_2 - \sigma^2\tau^2}{\sqrt{(\sigma_4 - \sigma^4)(\tau_4 - \tau^4)}} \text{ as } n \to \infty \tag{6.7.52}$$

Sample Covariance

Suppose first that the distribution means μ and ν are known. As noted earlier, this is almost always an unrealistic assumption, but is still a good place to start because the analysis is very simple and the results we obtain will be useful below. A natural estimator of the distsribution covariance $\delta = \cov(X, Y)$ in this case is the *special sample covariance*

$$W(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)(Y_i - \nu) \tag{6.7.53}$$

Note that the special sample covariance generalizes the special sample variance: $W(\boldsymbol{X}, \boldsymbol{X}) = W^2(\boldsymbol{X})$.

$W(\boldsymbol{X}, \boldsymbol{Y})$ is the sample mean for a random sample of size n from the distribution of $(X - \mu)(Y - \nu)$ and satisfies the following properties:

1. $\mathbb{E}[W(\boldsymbol{X}, \boldsymbol{Y})] = \delta$
2. $\var[W(\boldsymbol{X}, \boldsymbol{Y})] = \frac{1}{n}(\delta_2 - \delta^2)$
3. $W(\boldsymbol{X}, \boldsymbol{Y}) \to \delta$ as $n \to \infty$ with probability 1

Proof

These results follow directly from the section on the Law of Large Numbers. For part (b), note that

$$\var[(X - \mu)(Y - \nu)] = \mathbb{E}[(X - \mu)^2(Y - \nu)^2] - \left(\mathbb{E}[(X - \mu)(Y - \nu)]\right)^2 = \delta_2 - \delta^2 \tag{6.7.54}$$

As an estimator of δ, part (a) means that $W(\boldsymbol{X}, \boldsymbol{Y})$ is *unbiased* and part (b) means that $W(\boldsymbol{X}, \boldsymbol{Y})$ is *consistent*.

Consider now the more realistic assumption that the distribution means μ and ν are unknown. A natural approach in this case is to average $[(X_i - M(\boldsymbol{X})][Y_i - M(\boldsymbol{Y})]$ over $i \in \{1, 2, \ldots, n\}$. But rather than dividing by n in our average, we should divide by whatever constant gives an unbiased estimator of δ. As shown in the next theorem, this constant turns out to be $n - 1$, leading to the standard *sample covariance*:

$$S(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n-1}\sum_{i=1}^n [X_i - M(\boldsymbol{X})][Y_i - M(\boldsymbol{Y})] \tag{6.7.55}$$

$\mathbb{E}[S(\boldsymbol{X}, \boldsymbol{Y})] = \delta$.

Proof

Expanding as above we have,

$$\sum_{i=1}^n [X_i - M(\boldsymbol{X})][Y_i - M(\boldsymbol{Y})] = \sum_{i=1}^n X_i Y_i - nM(\boldsymbol{X})M(\boldsymbol{Y}) \tag{6.7.56}$$

But $\mathbb{E}(X_i Y_i) = \cov(X_i, Y_i) + \mathbb{E}(X_i)\mathbb{E}(Y_i) = \delta + \mu\nu$. Similarly, from the covariance of the sample means and the unbiased property, $\mathbb{E}[M(\boldsymbol{X})M(\boldsymbol{Y})] = \cov[M(\boldsymbol{X}), M(\boldsymbol{Y})] + \mathbb{E}[M(\boldsymbol{X})]\mathbb{E}[M(\boldsymbol{Y})] = \delta/n + \mu\nu$. So taking expected values in the displayed equation above gives

$$\mathbb{E}\left(\sum_{i=1}^n [X_i - M(\boldsymbol{X})][Y_i - M(\boldsymbol{Y})]\right) = n(\delta + \mu\nu) - n(\delta/n + \mu\nu) = (n-1)\delta \tag{6.7.57}$$

$S(\boldsymbol{X}, \boldsymbol{Y}) \to \delta$ as $n \to \infty$ with probability 1.

Proof

Once again, we have

$$S(\boldsymbol{X}, \boldsymbol{Y}) = \frac{n}{n-1}[M(\boldsymbol{XY}) - M(\boldsymbol{X})M(\boldsymbol{Y})] \tag{6.7.58}$$

where $M(\boldsymbol{XY})$ denotes the sample mean for the sample of the products $(X_1Y_1, X_2Y_2, \ldots, X_nY_n)$. By the strong law of large numbers, $M(\boldsymbol{X}) \to \mu$ as $n \to \infty$, $M(\boldsymbol{Y}) \to \nu$ as $n \to \infty$, and $M(\boldsymbol{XY}) \to \mathbb{E}(XY) = \delta + \mu\nu$ as $n \to \infty$, each with probability 1. So the result follows by letting $n \to \infty$ in the displayed equation.

Of courese, the *sample correlation* is

$$R(\boldsymbol{X}, \boldsymbol{Y}) = \frac{S(\boldsymbol{X}, \boldsymbol{Y})}{S(\boldsymbol{X})\, S(\boldsymbol{Y})} \tag{6.7.59}$$

Since the sample correlation $R(\boldsymbol{X}, \boldsymbol{Y})$ is a nonlinear function of the sample covariance and sample standard deviations, it will not in general be an unbiased estimator of the distribution correlation ρ. In most cases, it would be difficult to even compute the mean and variance of $R(\boldsymbol{X}, \boldsymbol{Y})$. Nonetheless, we can show convergence of the sample correlation to the distribution correlation.

$R(\boldsymbol{X}, \boldsymbol{Y}) \to \rho$ as $n \to \infty$ with probability 1.

Proof

This follows immediately from the strong law of large numbers and previous results. From the result above $S(\boldsymbol{X}, \boldsymbol{Y}) \to \delta$ as $n \to \infty$, and from the section on the sample variance, $S(\boldsymbol{X}) \to \sigma$ as $n \to \infty$ and $S(\boldsymbol{Y}) \to \tau$ as $n \to \infty$, each with probability 1. Hence $R(\boldsymbol{X}, \boldsymbol{Y}) \to \delta / \sigma\tau = \rho$ as $n \to \infty$ with probability 1.

Our next theorem gives a formuala for the variance of the sample covariance, not to be confused with the covariance of the sample variances given above!

The variance of the sample covariance is

$$\operatorname{var}[S(\boldsymbol{X}, \boldsymbol{Y})] = \frac{1}{n}\left(\delta_2 + \frac{1}{n-1}\sigma^2\,\tau^2 - \frac{n-2}{n-1}\delta^2\right) \tag{6.7.60}$$

Proof

Recall first that

$$S(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{2\,n\,(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)(Y_i - Y_j) \tag{6.7.61}$$

Hence using the bilinearity of the covariance operator we have

$$\operatorname{var}[S(\boldsymbol{X}), \boldsymbol{Y})] = \frac{1}{4n^2(n-1)^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} \operatorname{cov}[(X_i - X_j)(Y_i - Y_j), (X_k - X_l)(Y_k - Y_l)] \tag{6.7.62}$$

We compute the covariances in this sum by considering disjoint cases:

- $\operatorname{cov}[(X_i - X_j)(Y_i - Y_j), (X_k - X_l)(Y_k - Y_l)] = 0$ if $i = j$ or if $k = l$, and there are $2n^3 - n^2$ such terms.
- $\operatorname{cov}[(X_i - X_j)(Y_i - Y_j), (X_k - X_l)(Y_k - Y_l)] = 0$ if i, j, k, l are distinct, and there are $n(n-1)(n-2)(n-3)$ such terms.
- $\operatorname{cov}[(X_i - X_j)(Y_i - Y_j), (X_k - X_l)(Y_k - Y_l)] = 2\,\delta_2 + 2\sigma^2\tau^2$ if $i \neq j$ and $\{k, l\} = \{i, j\}$, and there are $2n(n-1)$ such terms.
- $\operatorname{cov}[(X_i - X_j)(Y_i - Y_j), (X_k - X_l)(Y_k - Y_l)] = \delta_2 - \delta^2$ if $i \neq j$, $k \neq l$, and $\#(\{i, j\} \cap \{k, l\}) = 1$, and there are $4n(n-1)(n-2)$ such terms.

Substituting and simplifying gives the result

It's not surprising that the variance of the standard sample covariance (where we don't know the distribution means) is greater than the variance of the special sample covariance (where we do know the distribution means).

$\operatorname{var}[S(\boldsymbol{X}, \boldsymbol{Y})] > \operatorname{var}[W(\boldsymbol{X}, \boldsymbol{Y})]$

Proof

From results above, and some simple algebra,

$$\operatorname{var}[S(\boldsymbol{X}, \boldsymbol{Y})] - \operatorname{var}[W(\boldsymbol{X}, \boldsymbol{Y})] = \frac{1}{n(n-1)}(\delta^2 + \sigma^2\tau^2) > 0 \tag{6.7.63}$$

But note that the difference goes to 0 as $n \to \infty$.

$\operatorname{var}[S(\boldsymbol{X}, \boldsymbol{Y})] \to 0$ as $n \to \infty$. Thus, the sample covariance is a *consistent* estimator of the distribution covariance.

Regression

In our first discussion above, we studied regression from a deterministic, descriptive point of view. The results obtained applied only to the sample. Statistically more interesting and deeper questions arise when the data come from a random experiment, and we try to draw inferences about the underlying distribution from the sample regression. There are two models that commonly arise. One is where the response variable is random, but the predictor variable is deterministic. The other is the model we consider here, where the predictor variable and the response variable are both random, so that the data form a random sample from a bivariate distribution.

Thus, suppose again that we have a basic random vector (X, Y) for an experiment. Recall that in the section on (distribution) correlation and regression, we showed that the best linear predictor of Y given X, in the sense of minimizing mean square error, is the random variable

$$L(Y \mid X) = \mathbb{E}(Y) + \frac{\operatorname{cov}(X, Y)}{\operatorname{var}(X)}[X - \mathbb{E}(X)] = \nu + \frac{\delta}{\sigma^2}(X - \mu) \tag{6.7.64}$$

so that the *distribution regression line* is given by

$$y = L(Y \mid X = x) = \nu + \frac{\delta}{\sigma^2}(x - \mu) \tag{6.7.65}$$

Moreover, the (minimum) value of the mean square error is $\mathbb{E}\{[Y - L(Y \mid X)]\} = \text{var}(Y)[1 - \text{cor}^2(X, Y)] = r^2(1 - \rho^2)$.

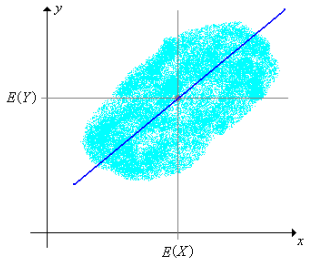

Figure 6.7.5: The distribution regression line

Of course, in real applications, we are unlikely to know the distribution parameters μ, ν, σ^2, and δ. If we want to estimate the distribution regression line, a natural approach would be to consider a random sample $((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$ from the distribution of (X, Y) and compute the sample regression line. Of course, the results are exactly the same as in the discussion above, except that all of the relevant quantities are random variables. The sample regression line is

$$y = M(\boldsymbol{Y}) + \frac{S(\boldsymbol{X}, \boldsymbol{Y})}{S^2(\boldsymbol{X})}[x - M(\boldsymbol{X})] \tag{6.7.66}$$

The mean square error is $S^2(\boldsymbol{Y})[1 - R^2(\boldsymbol{X}, \boldsymbol{Y})]$ and the coefficient of determination is $R^2(\boldsymbol{X}, \boldsymbol{Y})$.

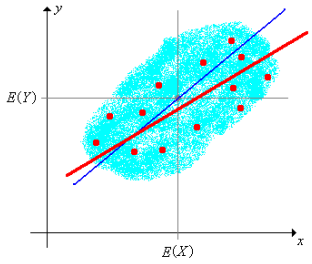

Figure 6.7.6: The distribution and sample regression lines

The fact that the sample regression line and mean square error are completely analogous to the distribution regression line and mean square error is mathematically elegant and reassuring. Again, the coefficients of the sample regression line can be viewed as estimators of the respective coefficients in the distribution regression line.

The coefficients of the sample regression line converge to the coefficients of the distribution regression line with probability 1.

1. $\frac{S(\boldsymbol{X}, \boldsymbol{Y})}{S^2(\boldsymbol{X})} \to \frac{\delta}{\sigma^2}$ as $n \to \infty$
2. $M(\boldsymbol{Y}) - \frac{S(\boldsymbol{X}, \boldsymbol{Y})}{S^2(\boldsymbol{X})} M(\boldsymbol{X}) \to \nu - \frac{\delta}{\sigma^2}\mu$ as $n \to \infty$

Proof

This follows from the strong law of large numbers and previous results. with probability 1, $S(\boldsymbol{X}, \boldsymbol{Y}) \to \delta$ as $n \to \infty$, $S^2(\boldsymbol{X}) \to \sigma^2$ as $n \to \infty$, $M(\boldsymbol{X}) \to \mu$ as $n \to \infty$, and $M(\boldsymbol{Y}) \to \nu$ as $n \to \infty$.

Of course, if the linear relationship between X and Y is not strong, as measured by the sample correlation, then transformation applied to one or both variables may help. Again, some typical transformations are explored in the exercises below.

Exercises

Basic Properties

Suppose that x and y are population variables, and $\boldsymbol{x}$ and $\boldsymbol{y}$ samples of size n from x and y respectively. Suppose also that $m(\boldsymbol{x}) = 3$, $m(\boldsymbol{y}) = -1$, $s^2(\boldsymbol{x}) = 4$, $s^2(\boldsymbol{y}) = 9$, and $s(\boldsymbol{x}, \boldsymbol{y}) = 5$. Find each of the following:

1. $r(\boldsymbol{x}, \boldsymbol{y})$
2. $m(2\boldsymbol{x} + 3\boldsymbol{y})$
3. $s^2(2\boldsymbol{x} + 3\boldsymbol{y})$
4. $s(2\boldsymbol{x} + 3\boldsymbol{y} - 1, 4\boldsymbol{x} + 2\boldsymbol{y} - 3)$

Suppose that x is the temperature (in degrees Fahrenheit) and y the resistance (in ohms) for a certain type of electronic component after 10 hours of operation. For a sample of 30 components, $m(\boldsymbol{x}) = 113$, $s(\boldsymbol{x}) = 18$, $m(\boldsymbol{y}) = 100$, $s(\boldsymbol{y}) = 10$, $r(\boldsymbol{x}, \boldsymbol{y}) = 0.6$.

1. Classify x and y by type and level of measurement.
2. Find the sample covariance.
3. Find the equation of the regression line.

Suppose now that temperature is converted to degrees Celsius (the transformation is $\frac{5}{9}(x - 32)$).

4. Find the sample means.

5. Find the sample standard deviations.
6. Find the sample covariance and correlation.
7. Find the equation of the regression line.

Answer

1. continuous, interval
2. $m = 45^\circ$, $s = 10^\circ$

Suppose that x is the length and y the width (in inches) of a leaf in a certain type of plant. For a sample of 50 leaves $m(\boldsymbol{x}) = 10$, $s(\boldsymbol{x}) = 2$, $m(\boldsymbol{y}) = 4$, $s(\boldsymbol{y}) = 1$, and $r(\boldsymbol{x}, \boldsymbol{y}) = 0.8$.

1. Classify x and y by type and level of measurement.
2. Find the sample covariance.
3. Find the equation of the regression line with x as the predictor variable and y as the response variable.

Suppose now that x and y are converted to inches (0.3937 inches per centimeter).

4. Find the sample means.
5. Find the sample standard deviations.
6. Find the sample covariance and correlation.
7. Find the equation of the regression line.

Answer

1. continuous, ratio
2. $m = 25.4$, $s = 5.08$

Scatterplot Exercises

Click in the interactive scatterplot, in various places, and watch how the means, standard deviations, correlation, and regression line change.

Click in the interactive scatterplot to define 20 points and try to come as close as possible to each of the following sample correlations:

1. 0
2. 0.5
3. -0.5
4. 0.7
5. -0.7
6. 0.9
7. -0.9.

Click in the interactive scatterplot to define 20 points. Try to generate a scatterplot in which the regression line has

1. slope 1, intercept 1
2. slope 3, intercept 0
3. slope -2, intercept 1

Simulation Exercises

Run the bivariate uniform experiment 2000 times in each of the following cases. Compare the sample means to the distribution means, the sample standard deviations to the distribution standard deviations, the sample correlation to the distribution correlation, and the sample regression line to the distribution regression line.

1. The uniform distribution on the square
2. The uniform distribution on the triangle.
3. The uniform distribution on the circle.

Run the bivariate normal experiment 2000 times for various values of the distribution standard deviations and the distribution correlation. Compare the sample means to the distribution means, the sample standard deviations to the distribution standard deviations, the sample correlation to the distribution correlation, and the sample regression line to the distribution regression line.

Transformations

Consider the function $y = a + bx^2$.

1. Sketch the graph for some representative values of a and b.
2. Note that y is a linear function of x^2, with intercept a and slope b.
3. Hence, to fit this curve to sample data, simply apply the standard regression procedure to the data from the variables x^2 and y.

Consider the function $y = \frac{1}{a+bx}$.

1. Sketch the graph for some representative values of a and b.
2. Note that $\frac{1}{y}$ is a linear function of x, with intercept a and slope b.
3. Hence, to fit this curve to our sample data, simply apply the standard regression procedure to the data from the variables x and $\frac{1}{y}$.

Consider the function $y = \frac{x}{a+bx}$.

1. Sketch the graph for some representative values of a and b.
2. Note that $\frac{1}{y}$ is a linear function of $\frac{1}{x}$, with intercept b and slope a.

3. Hence, to fit this curve to sample data, simply apply the standard regression procedure to the data from the variables $\frac{1}{x}$ and $\frac{1}{y}$.
4. Note again that the names of the intercept and slope are reversed from the standard formulas.

Consider the function $y = ae^{bx}$.

1. Sketch the graph for some representative values of a and b.
2. Note that $\ln(y)$ is a linear function of x, with intercept $\ln(a)$ and slope b.
3. Hence, to fit this curve to sample data, simply apply the standard regression procedure to the data from the variables x and $\ln(y)$.
4. After solving for the intercept $\ln(a)$, recover the statistic $a = e^{\ln(a)}$.

Consider the function $y = ax^b$.

1. Sketch the graph for some representative values of a and b.
2. Note that $\ln(y)$ is a linear function of $\ln(x)$, with intercept $\ln(a)$ and slope b.
3. Hence, to fit this curve to sample data, simply apply the standard regression procedure to the data from the variables $\ln(x)$ and $\ln(y)$.
4. After solving for the intercept $\ln(a)$, recover the statistic $a = e^{\ln(a)}$.

Computational Exercises

All statistical software packages will perform regression analysis. In addition to the regression line, most packages will typically report the coefficient of determination $r^2(\boldsymbol{x}, \boldsymbol{y})$, the sums of squares $\text{sst}(\boldsymbol{y})$, $\text{ssr}(\boldsymbol{x}, \boldsymbol{y})$, $\text{sse}(\boldsymbol{x}, \boldsymbol{y})$, and the standard error of estimate $\text{se}(\boldsymbol{x}, \boldsymbol{y})$. Most packages will also draw the scatterplot, with the regression line superimposed, and will draw the various graphs of residuals discussed above. Many packages also provide easy ways to transform the data. Thus, there is very little reason to perform the computations by hand, except with a small data set to master the definitions and formulas. In the following problem, do the computations and draw the graphs with minimal technological aids.

Suppose that x is the number of math courses completed and y the number of science courses completed for a student at Enormous State University (ESU). A sample of 10 ESU students gives the following data: $((1,1),(3,3),(6,4),(2,1),(8,5),(2,2),(4,3),(6,4),(4,3),(4,4))$

1. Classify x and y by type and level of measurement.
2. Sketch the scatterplot.

Construct a table with rows corresponding to cases and columns corresponding to i, x_i, y_i, $x_i - m(\boldsymbol{x})$, $y_i - m(\boldsymbol{y})$, $[x_i - m(\boldsymbol{x})]^2$, $[y_i - m(\boldsymbol{y})]^2$, $[x_i - m(\boldsymbol{x})][y_i - m(\boldsymbol{y})]$, $\hat{y}_i$, $\hat{y}_i - m(\boldsymbol{y})$, $[\hat{y}_i - m(\boldsymbol{y})]^2$, $y_i - \hat{y}_i$, and $(y_i - \hat{y}_i)^2$. Add a rows at the bottom for totals and means. Use precision arithmetic.

3. Complete the first 8 columns.
4. Find the sample correlation and the coefficient of determination.
5. Find the sample regression equation.
6. Complete the table.
7. Verify the identities for the sums of squares.

Answer

i	x_i	y_i	$x_i - m(\boldsymbol{x})$	$y_i - m(\boldsymbol{y})$	$[x_i - m(\boldsymbol{x})]^2$	$[y_i - m(\boldsymbol{y})]^2$	$[x_i - m(\boldsymbol{x})][y_i - m(\boldsymbol{y})]$	$\hat{y}_i$	$\hat{y}_i - m(\boldsymbol{y})$	$[\hat{y}_i - m(\boldsymbol{y})]^2$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	1	1	-3	-2	9	4	6	$9/7$	$-12/7$	$144/49$	$-2/7$	$4/49$
2	3	3	-1	0	1	0	0	$17/7$	$-4/7$	$16/49$	$4/7$	$16/49$
3	6	4	2	1	4	1	2	$29/7$	$8/7$	$64/49$	$-1/7$	$1/49$
4	2	1	-2	-2	4	4	4	$13/7$	$-8/7$	$64/49$	$-6/7$	$36/49$
5	8	5	4	2	16	4	8	$37/7$	$16/7$	$256/49$	$-2/7$	$4/49$
6	2	2	-2	-1	4	1	2	$13/7$	$-8/7$	$64/49$	$1/7$	$1/49$
7	4	3	0	0	0	0	0	3	0	0	0	0
8	6	4	2	1	4	1	2	$29/7$	$8/7$	$64/49$	$-1/7$	$1/49$
9	4	3	0	0	0	0	0	3	0	0	0	0
10	4	4	0	1	0	1	0	3	0	0	1	1
Total	40	30	0	0	42	16	24	30	0	$96/7$	0	$16/7$
Mean	4	3	0	0	$14/3$	$16/9$	$8/3$	3	0	$96/7$	0	$2/7$

1. discrete, ratio
4. $r = 2\sqrt{3/14} \approx 0.926$, $r^2 = 6/7$
5. $y = 3 + \frac{4}{7}(x - 4)$
7. $16 = 96/7 + 16/7$

The following two exercise should help you review some of the probability topics in this section.

Suppose that (X, Y) has a continuous distribution with probability density function $f(x, y) = 15x^2 y$ for $0 \le x \le y \le 1$. Find each of the following:

1. $\mu = \mathbb{E}(X)$ and $\nu = \mathbb{E}(Y)$
2. $\sigma^2 = \text{var}(X)$ and $\tau^2 = \text{var}(Y)$
3. $\sigma_3 = \mathbb{E}\left[(X - \mu)^3\right]$ and $\tau_3 = \mathbb{E}\left[(Y - \nu)^3\right]$
4. $\sigma_4 = \mathbb{E}\left[(X - \mu)^4\right]$ and $\tau_4 = \mathbb{E}\left[(Y - \nu)^4\right]$
5. $\delta = \text{cov}(X, Y)$, $\rho = \text{cor}(X, Y)$, and $\delta_2 = \mathbb{E}\left[(X - \mu)^2 (Y - \nu)^2\right]$

6. $L(Y \mid X)$ and $L(X \mid Y)$

Answer

1. $5/8$, $5/6$
2. $17/448$, $5/252$
3. $-5/1792$, $-5/1512$
4. $305/86\,016$, $5/3024$
5. $5/336$, $\sqrt{5/17}$, $1/768$
6. $L(Y \mid X) = \frac{10}{17} + \frac{20}{51} X$, $L(X \mid Y) = \frac{3}{4} Y$

Suppose now that $((X_1, Y_1), (X_2, Y_2), \ldots (X_9, Y_9))$ is a random sample of size 9 from the distribution in the previous exercise. Find each of the following:

1. $\mathbb{E}[M(\boldsymbol{X})]$ and $\text{var}[M(\boldsymbol{X})]$
2. $\mathbb{E}[M(\boldsymbol{Y})]$ and $\text{var}[M(\boldsymbol{Y})]$
3. $\text{cov}[M(\boldsymbol{X}), M(\boldsymbol{Y})]$ and $\text{cor}[M(\boldsymbol{X}), M(\boldsymbol{Y})]$
4. $\mathbb{E}[W^2(\boldsymbol{X})]$ and $\text{var}[W^2(\boldsymbol{X})]$
5. $\mathbb{E}[W^2(\boldsymbol{Y})]$ and $\text{var}[W^2(\boldsymbol{Y})]$
6. $\mathbb{E}[S^2(\boldsymbol{X})]$ and $\text{var}[S^2(\boldsymbol{X})]$
7. $\mathbb{E}[S^2(\boldsymbol{Y})]$ and $\text{var}[S^2(\boldsymbol{Y})]$
8. $\mathbb{E}[W(\boldsymbol{X}, \boldsymbol{Y})]$ and $\text{var}[W(\boldsymbol{X}, \boldsymbol{Y})]$
9. $\mathbb{E}[S(\boldsymbol{X}, \boldsymbol{Y})]$ and $\text{var}[S(\boldsymbol{X}, \boldsymbol{Y})]$

Answer

1. $5/8$, $17/4032$
2. $5/6$, $5/2268$
3. $5/3024$ $\sqrt{5/17}$
4. $17/448$, $317/1\,354\,752$
5. $5/252$, $5/35\,721$
6. $17/448$, $5935/21\,676\,032$
7. $5/252$, $115/762\,048$
8. $5/336$, $61/508\,032$
9. $5/336$, $181/1\,354\,752$

Data Analysis Exercises

Use statistical software for the following problems.

Consider the height variables in Pearson's height data.

1. Classify the variables by type and level of measurement.
2. Compute the correlation coefficient and the coefficient of determination
3. Compute the least squares regression line, with the height of the father as the predictor variable and the height of the son as the response variable.
4. Draw the scatterplot and the regression line together.
5. Predict the height of a son whose father is 68 inches tall.
6. Compute the regression line if the heights are converted to centimeters (there are 2.54 centimeters per inch).

Answer

1. Continuous, ratio
2. $r = 0.501$, $r^2 = 0.251$
3. $y = 33.893 + 0.514x$
5. 68.85
6. $y = 86.088 + 0.514x$

Consider the petal length, petal width, and species variables in Fisher's iris data.

1. Classify the variables by type and level of measurement.
2. Compute the correlation between petal length and petal width.
3. Compute the correlation between petal length and petal width by species.

Answer

1. Species: discrete, nominal; petal length and width: continuous ratio
2. 0.9559
3. Setosa: 0.3316, Verginica: 0.3496, Versicolor: 0.6162

Consider the number of candies and net weight variables in the M&M data.

1. Classify the variable by type and level of measurement.
2. Compute the correlation coefficient and the coefficient of determination.
3. Compute the least squares regression line with number of candies as the predictor variable and net weight as the response variable.
4. Draw the scatterplot and the regression line in part (b) together.
5. Predict the net weight of a bag of M&Ms with 56 candies.
6. Naively, one might expect a much stronger correlation between the number of candies and the net weight in a bag of M&Ms. What is another source of variability in net weight?

Answer

1. Number of candies: discrete, ratio; net weight: continuous, ratio
2. $r = 0.793$, $r^2 = 0.629$
3. $y = 20.278 + 0.507x$
5. 48.657
6. Variability in the weight of individual candies.

Consider the response rate and total SAT score variables in the SAT by state data set.

1. Classify the variables by type and level of measurement.
2. Compute the correlation coefficient and the coefficient of determination.
3. Compute the least squares regression line with response rate as the predictor variable and SAT score as the response variable.
4. Draw the scatterplot and regression line together.
5. Give a possible explanation for the negative correlation.

Answer

1. Response rate: continuous, ratio. SAT score could probably be considered either discrete or continuous, but is only at the interval level of measurement, since the smallest possible scores is 400 (200 each on the verbal and math portions).
2. $r = -0.849$, $r^2 = 0.721$
3. $y = 1141.5 - 2.1x$
5. States with low response rate may be states for which the SAT is optional. In that case, the students who take the test are the better, college-bound students. Conversely, states with high response rates may be states for which the SAT is mandatory. In that case, all students including the weaker, non-college-bound students take the test.

Consider the verbal and math SAT scores (for all students) in the SAT by year data set.

1. Classify the variables by type and level of measurement.
2. Compute the correlation coefficient and the coefficient of determination.
3. Compute the least squares regression line.
4. Draw the scatterplot and regression line together.

Answer

1. Continuous perhaps, but only at the interval level of measurement because the smallest possible score on each part is 200.
2. $r = 0.614$, $r^2 = 0.377$
3. $y = 321.5 + 0.3\,x$

Consider the temperature and erosion variables in the first data set in the Challenger data.

1. Classify the variables by type and level of measurement.
2. Compute the correlation coefficient and the coefficient of determination.
3. Compute the least squares regression line.
4. Draw the scatter plot and the regression line together.
5. Predict the O-ring erosion with a temperature of 31° F.
6. Is the prediction in part (c) meaningful? Explain.
7. Find the regression line if temperature is converted to degrees Celsius. Recall that the conversion is $\frac{5}{9}(x - 32)$.

Answer

1. temperature: continuous, interval; erosion: continuous ratio
2. $r = -0.555$, $r^2 = 0.308$
3. $y = 106.8 - 1.414x$
5. 62.9.
6. This estimate is problematic, because 31° is far outside of the range of the sample data.
7. $y = 61.54 - 2.545x$