Measures of Correlation

In linear regression, there is a dependent variable Y and an independent regressor variable X. We think of Y as being approximately a function of X. We want to know if there is a relation of the form

$$E(Y \mid X = x) = a + bx$$

$$\shortparallel$$

$$\sum_y y \, P(Y = y \mid X = x)$$

For example $\quad Y = a + bX + \varepsilon$
where $\varepsilon \sim N(0, \sigma^2)$

or even $\quad Y = a + bX$

In these cases $\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, a + bX)$
$= b \, \mathrm{Var}(X)$

$$b = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}.$$

We want estimates of $\mathrm{Cov}(X, Y)$.

The covariance of X and Y is

$$Cov(X,Y) = E[(X-E(X))(Y-E(Y))]$$
$$= E[XY] - E[X]E[Y]$$

The correlation of X and Y is defined by

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

we have

① $-1 \leq Corr(X,Y) \leq +1$

② If X and Y are independent, $Cov(X,Y)=0$
but $Cov(X,Y)=0$ does not imply that X and Y are independent.

③ If $Y = aX + b$, then

$$Corr(X,Y) = \frac{b\,Var(X)}{\sqrt{Var(X)\cdot b^2 Var(X)}}$$

$$= \frac{b}{|b|}$$

$$= \begin{cases} 1 & \text{if } b>0 \\ 0 & \text{if } b=0 \\ -1 & \text{if } b<0 \end{cases}$$

We want measures of correlation for samples $\vec{x} = (x_1, \ldots, x_n)$ and $\vec{y} = (y_1, \ldots, y_n)$.

~~The~~ Three such measures are in use.

① Pearson's Correlation Coefficient

Let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

Pearson Corr. Coeff. =

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

The code for calculating r

Cor (<Data ~~(x, y)~~>, ~~method=pears~~ method="pearson")

or Cor (<Data>)

② Spearman's correlation coefficient

First we define the ranks of the variables.

The rank of X_i is j if X_i is the jth smallest of $X_1, X_2, \ldots, X_n$.

The rank of Y_i is j if Y_i is the jth smallest of $Y_1, \ldots, Y_n$

Spearman's correlation coefficient r_s is Pearson's correlation coefficient applied to

$$(\text{rank}(X_1), \text{rank}(X_2), \ldots, \text{rank}(X_n))$$

$$(\text{rank}(Y_1), \text{rank}(Y_2), \ldots, \text{rank}(Y_n))$$

The R code for calculating r_s is

```
Cor(<Data>, method="spearman")
```

Fact

If the x_i's are unique (no repetitions) and the y_i's are themselves unique

then

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where $d_i = \text{rank}(Y_i) - \text{rank}(X_i)$

3 Kendall's correlation coefficient

We consider ~~pairs~~ vectors of ~~obser~~ observations

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \quad i = 1, \ldots, n$$

A pair of vectors $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$ $\begin{pmatrix} x_j \\ y_j \end{pmatrix}$

is <u>concordant</u> ~~if~~ ~~$x_i < x_j$~~

if either $x_i < x_j$ and $y_i < y_j$

or $x_i > x_j$ and $y_i > y_j$

A pair of vectors is discordant if

either $x_i < x_j$ and $y_i > y_j$

or $x_i > x_j$ and $y_i < y_j$

Let n_c be the number of concordant pairs
and n_d be the number of discordant pairs.

Note: $n_c + n_d = \binom{n}{2}$

Kendall's correlation coefficient

τ "tau"

is

$$\tau = \frac{n_c - n_d}{\binom{n}{2}} = \frac{n_c - n_d}{\frac{n(n-1)}{2}}$$

"Kendall's τ can be calculated from the ranks.

We can calculate it in R using

Cor (<Data>, method = "kendall")

Inference on correlation coefficients

Pearson

We need an assumption about the joint distribution of X and Y.
We assume (X, Y) is bivariate normal distributed with parameters $\mu_x, \sigma_x, \mu_y, \sigma_y, \rho$ where ρ is the correlation.
We have ~~$E(Y \mid X=x) = \mu_y + \rho \frac{\sigma_x}{\sigma_y}$~~

$$E(Y \mid X=x) = \mu_y + \rho \frac{\sigma_y}{\sigma_x}(x-\mu_x)$$

Theorem

Under $H_0: \rho = 0$ the statistic

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

is t-distributed with $n-2$ degrees of freedom

where r is Pearson's correlation
n is number of observations.

In R cor.test(Data, method="pearson")