Correlations

(Part 1 of 3)

Correlations

Correlation coefficients are used to describe relationships among quantitative variables.

- The signs + or - indicates the direction of the relationship (positive or inverse).
- The magnitude indicates the strength of the relationship (ranging from 0 for no relationship to 1 for a perfectly predictable relationship)

Data

We will use the state.x77 dataset available in the base R installation. It provides data on the population, income, illeteracy rate, life expectancy, murder rate, and high school graduation rate for teh 50 US states in 1977.

Data Description

state.x77: matrix with 50 rows and 8 columns giving the following statistics in the respective columns.

Population: population estimate as of July 1, 1975 Income: per capita income (1974) Illiteracy: illiteracy (1970, percent of population) Life Exp: life expectancy in years (1969-71) Murder: murder and non-negligent manslaughter rate per 100,000 population (1976) HS Grad: percent high-school graduates (1970)

```
states<- state.x77[,1:6] # take a subset of 6 columns
View(states) # view the dataset

states<- state.x77[,1:6] # take a subset of 6 columns
dim(states)  # it has 50 rows, 6 columns
```

```
## [1] 50  6
```

```
library(psych)
describe(states) # see the Compative Statics -- mean, sd, median
```

```
##            vars  n    mean      sd  median trimmed     mad     min     max
## Population    1 50 4246.42 4464.49 2838.50 3384.28 2890.33  365.00 21198.0
## Income        2 50 4435.80  614.47 4519.00 4430.07  581.18 3098.00  6315.0
## Illiteracy    3 50    1.17    0.61    0.95    1.10    0.52    0.50     2.8
## Life Exp      4 50   70.88    1.34   70.67   70.92    1.54   67.96    73.6
## Murder        5 50    7.38    3.69    6.85    7.30    5.19    1.40    15.1
## HS Grad       6 50   53.11    8.08   53.25   53.34    8.60   37.80    67.3
##               range  skew kurtosis     se
## Population 20833.00  1.92     3.75 631.37
## Income      3217.00  0.20     0.24  86.90
## Illiteracy     2.30  0.82    -0.47   0.09
## Life Exp       5.64 -0.15    -0.67   0.19
## Murder        13.70  0.13    -1.21   0.52
## HS Grad       29.50 -0.32    -0.88   1.14
```

Types of correlations

The *Pearson product-moment correlation* assesses the degree of linear relationship between two quantitative variables.

Spearman's rank-order correlation coefficient assesses the degree of relationship between two rank-ordered variables.

In R,

- The `cor()` function produces all three correlation coefficients.
- The `cov()` function provides covariances.
- There are many options, but a simplified format for producing correlations is `cor(x, use= , method= )`.
- The default options are `use="everything"` and `method="pearson"`.

Suppose we need to create a variance-covariance table for the 6 variables in `states`.

```
cov(states)
```

```
##               Population      Income   Illiteracy     Life Exp      Murder
## Population 19931683.7588 571229.7796  292.8679592 -407.8424612 5663.523714
## Income       571229.7796 377573.3061 -163.7020408  280.6631837 -521.894286
## Illiteracy      292.8680   -163.7020    0.3715306   -0.4815122    1.581776
## Life Exp       -407.8425    280.6632   -0.4815122    1.8020204   -3.869480
## Murder         5663.5237   -521.8943    1.5817755   -3.8694804   13.627465
## HS Grad       -3551.5096   3076.7690   -3.2354694    6.3126849  -14.549616
##                 HS Grad
## Population -3551.509551
## Income      3076.768980
## Illiteracy    -3.235469
## Life Exp       6.312685
## Murder       -14.549616
## HS Grad       65.237894
```

Here, the diagonal elements give us the variance of each variable. The off-diagonal elements give us the covariance of a given pair of variables.

Suppose we need to create a correlation matrix for the 6 variables in `states`, using the *Pearson product-moment correlation* measure.

```
cor(states)
```

```
##             Population     Income Illiteracy    Life Exp     Murder
## Population  1.00000000  0.2082276  0.1076224 -0.06805195  0.3436428
## Income      0.20822756  1.0000000 -0.4370752  0.34025534 -0.2300776
## Illiteracy  0.10762237 -0.4370752  1.0000000 -0.58847793  0.7029752
## Life Exp   -0.06805195  0.3402553 -0.5884779  1.00000000 -0.7808458
## Murder      0.34364275 -0.2300776  0.7029752 -0.78084575  1.0000000
## HS Grad    -0.09848975  0.6199323 -0.6571886  0.58221620 -0.4879710
##                HS Grad
```

```
## Population -0.09848975
## Income      0.61993232
## Illiteracy -0.65718861
## Life Exp    0.58221620
## Murder     -0.48797102
## HS Grad     1.00000000
```

Suppose we need to create a correlation matrix for the 6 variables in `states`, using the *Spearman's rank-order correlation coefficient* measure.

```
cor(states, method="spearman")
```

```
##            Population     Income Illiteracy   Life Exp     Murder
## Population  1.0000000  0.1246098  0.3130496 -0.1040171  0.3457401
## Income      0.1246098  1.0000000 -0.3145948  0.3241050 -0.2174623
## Illiteracy  0.3130496 -0.3145948  1.0000000 -0.5553735  0.6723592
## Life Exp   -0.1040171  0.3241050 -0.5553735  1.0000000 -0.7802406
## Murder      0.3457401 -0.2174623  0.6723592 -0.7802406  1.0000000
## HS Grad    -0.3833649  0.5104809 -0.6545396  0.5239410 -0.4367330
##               HS Grad
## Population -0.3833649
## Income      0.5104809
## Illiteracy -0.6545396
## Life Exp    0.5239410
## Murder     -0.4367330
## HS Grad     1.0000000
```

Some observations:

- A strong positive correlation exists between income and high school graduation rate.

- A strong negative correlation exists between illiteracy rates and life expectancy.
- Notice that we get square matrices by default -- all variables are crossed with all other variables.

We can also produce nonsquare matrices, as shown in the following example.

```
x <- states[,c("Population", "Income", "Illiteracy", "HS Grad")]
y <- states[,c("Life Exp", "Murder")]
cor(x,y)
```

```
##               Life Exp     Murder
## Population -0.06805195  0.3436428
## Income      0.34025534 -0.2300776
## Illiteracy -0.58847793  0.7029752
## HS Grad     0.58221620 -0.4879710
```