REVOLUTION ANALYTICS

Introduction to Revolution R Enterprise Module 5: Estimating Correlation & Variance/Covariance Matrices

Covariance

Covariance measures the impact of the influence one variable may have over another. For example, consider two variables, warm days and ice cream:

- Positive covariance implies that warm days are accompanied by greater purchases of ice cream, and colder days are likewise accompanied by lesser purchases of ice cream.

Alternately, consider two new variables, warm days and coats:

- Negative covariance implies that warm days are accompanied by lesser purchases of coats, and colder days are accompanied by greater purchases of coats.

Variance-Covariance

The variance-covariance matrix computes the covariance between every combination of elements in a random vector.

- As a result, the variance-covariance matrix extends covariance to multiple dimensions.

Cross-Product Matrix

Additionally, a cross-product matrix is a matrix of the form

$$X'X$$

where X represents an arbitrary set of raw or standardized variables. More generally, this is a matrix of the form

$$X'WX$$

where W is a diagonal weighting matrix.

rxCovCor

The rxCovCor function in ScaleR calculates the covariance, correlation, or sum of squares/cross product matrix for a set of variables in an XDF file (given that the data is stored on Hadoop). However, it is usually simpler to use one of the following convenience functions instead:

- rxCov: Use rxCov to return the covariance matrix
- rxCor: Use rxCor to return the correlation matrix
- rxSSCP: Use rxSSCP to return the augmented cross-product matrix
  - First add a column of 1's (if no weights are specified) or a column equaling the square root of the weights to the data matrix, and then compute the cross-product.

Example: Covariance Matrix

Let's compute the Pearson's correlation matrix using the rxCor command, for variables age, balance, and housing. Since housing is a factor variable, we will use the transforms argument to change it into a logical variable to be used in the creation of the correlation matrix:

```
infile <- file.path("data", "BankXDF.xdf")
BankDS <- RxXdfData(file = infile)

houseCor <- rxCor(formula = ~age + balance + noHousing, data = BankDS, transforms = list(noHousir
    "no"))
```

```
## Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds
## Rows Read: 10000, Total Rows Processed: 20000, Total Chunk Time: 0.005 seconds
## Rows Read: 10000, Total Rows Processed: 30000, Total Chunk Time: 0.004 seconds
## Rows Read: 10000, Total Rows Processed: 40000, Total Chunk Time: 0.005 seconds
## Rows Read: 5211, Total Rows Processed: 45211, Total Chunk Time: 0.002 seconds
## Computation time: 0.040 seconds.
```

Example: Covariance Matrix

Considering the result,

```
houseCor
```

```
##                age balance noHousing
## age        1.00000 0.09778   0.18551
## balance    0.09778 1.00000   0.06877
## noHousing 0.18551 0.06877   1.00000
```

We can see that all of the covariance computations listed in the above matrix are positive - therefore, while not all covariances may be extremely strong, they are nevertheless all positively correlated.

Example: Covariance Matrix

Considering age and balance:

- Both variables share a positive covariance - older ages tend to occur with higher account balances, and vice versa, however the covariance is small.
- Both age and noHousing share a positive covariance - older ages tend to occur with no housing, and vice versa, with a slightly larger covariance than with age and balance.
- Both balance and noHousing share a positive covariance - higher account balances tend to occur with no housing, and vice versa, with the lowest covariance. In this case, the covariance is so slight that it is more likely that there is little correlation between the variables.

Exercise: Covariance Matrix

Compute the covariance matrix, again using rxCov, this time with variables age, duration, and y (term subscription). Remember to transform the factor variable y such that the variable reflects a logical value, as in the example above. Interpret the results.

Exercise: Solution

Setting up the computation,

```
yCor <- rxCor(formula=~age + duration + yesy, data = BankDS,
                  transforms = list(
                    yesy = y == "yes"))
```

Considering the result,

```
yCor
```

Exercise: Solution

Interpreting the variables:

- Both age and duration have negative covariance, indicating that older ages tend to be on the advertising call for less time, and vice versa. However, the covariance is very close to zero, indicating that no correlation exists between the variables.
- Both age and yesy have positive covariance, indicating that older ages tend to have lower rejections for term subscriptions, and vice versa. Again, however, the covariance is very close to zero, indicating that no correlation exists between the variables.
- Both duration and yesy have positive covariance, indicating that higher advertising call times tend to have lower rejections for term subscriptions, and vice versa.

Recap

Let's review some of the concepts covered in this module:

- What is covariance?
  - What is a covariance matrix?
  - What does positive covariance imply?
- How do you interpret covariance?

REVOLUTION ANALYTICS

Thank you

Revolution Analytics is the leading commercial provider of software and support for the popular open source R statistics language.

www.revolutionanalytics.com, 1.855.GET.REVO, Twitter: @Revo-lutionR