Correlation

Alan T. Arnholt
Department of Mathematical Sciences
Appalachian State University
arnholt@math.appstate.edu

Spring 2006 R Notes

*Copyright © 2006 Alan T. Arnholt
Correlation

Overview of Correlation

The R Script
The **correlation coefficient**, denoted by $r$, measures the strength and direction of the linear relationship between two numeric variables $X$ and $Y$ and is defined by

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_X} \right) \left( \frac{y_i - \bar{y}}{s_Y} \right)$$  \hspace{1cm} (1)
Correlation Properties

- The value for $r$ will always be between $-1$ and $+1$. 
Correlation Properties

- The value for $r$ will always be between $-1$ and $+1$.
- When $r$ is close to $+1$, it indicates a strong positive linear relationship. That is, when $x$ increases so does $y$ and vice versa.
Correlation Properties

- The value for $r$ will always be between $-1$ and $+1$.
- When $r$ is close to $+1$, it indicates a strong positive linear relationship. That is, when $x$ increases so does $y$ and vice versa.
- When the value of $r$ is close to $-1$, it indicates a strong negative linear relationship. Values of $r$ close to zero indicate weak linear relationships.
Correlation Properties

- The value for $r$ will always be between $-1$ and $+1$.
- When $r$ is close to $+1$, it indicates a strong positive linear relationship. That is, when $x$ increases so does $y$ and vice versa.
- When the value of $r$ is close to $-1$, it indicates a strong negative linear relationship. Values of $r$ close to zero indicate weak linear relationships.
- To compute the correlation between two numeric vectors with R, one may use the function `cor(x,y)`.
Use the data frame `Correlat` from the BSDA package to:

1. Create a scatterplot of $Y$ versus $X$. 
Use the data frame `Correlat` from the BSDA package to:

1. Create a scatterplot of \( Y \) versus \( X \).
2. Compute the sample correlation coefficient \( r \) using the definition.
Use the data frame Correlat from the BSDA package to:

1. Create a scatterplot of $Y$ versus $X$.
2. Compute the sample correlation coefficient $r$ using the definition.
3. Compute the sample correlation coefficient $r$ using the R function cor().

```r
> library(BSDA)
> attach(Correlat)
> plot(X,Y,col="blue",main="Scatterplot")
```
Scatterplot of Y Versus X
Using the Formula

```r
> m.x <- mean(X)
> m.y <- mean(Y)
> s.x <- sd(X)
> s.y <- sd(Y)
> Z.x <- (X-m.x)/s.x
> Z.y <- (Y-m.y)/s.y
> ZxZy <- Z.x*Z.y
> r <- (1/(length(X)-1))*sum(ZxZy)
> r
[1] -0.813717
> cor(X,Y)
[1] -0.813717
```
Code Continued

```r
> c(m.x, m.y, s.x, s.y)
> stuff <- cbind(X,Y,Z.x,Z.y,ZxZy)
> stuff
        X      Y   Z.x   Z.y  ZxZy
[1,]  42.0  75.0 -0.575  0.323 -0.186
[2,]  61.0  49.0  0.296 -1.059 -0.314
[3,]  12.0  95.0 -1.949  1.385 -2.702
[4,]  71.0  64.0  0.755 -1.059 -0.314
[5,]  52.0  83.0 -1.164  0.748 -0.087
[6,]  48.0  84.0 -0.299  0.801 -0.240
[7,]  74.0  38.0  0.892 -1.643 -1.466
[8,]  65.0  58.0  0.479 -0.580 -0.278
[9,]  53.0  81.0 -0.070  0.641 -0.045
[10,]  63.0  47.0  0.388 -1.165 -0.452
[11,]  55.0  78.0  0.021  0.482  0.010
[12,]  94.0  51.0  1.809 -0.952 -1.723
[13,]  19.0  93.0 -1.629  1.279 -2.084
```
Correlation and Causation

- A positive correlation between two variables means that large values of one variable tend to be associated with large values in the other variable.
Correlation and Causation

- A positive correlation between two variables means that large values of one variable tend to be associated with large values in the other variable.

- **This does not necessarily mean** that the large values of the first variable caused the large values of the other variable.
A positive correlation between two variables means that large values of one variable tend to be associated with large values in the other variable.

This does not necessarily mean that the large values of the first variable caused the large values of the other variable.

Correlation measures the linear association between two variables, not the causal effect.
Link to the R Script

- Go to my web page *Script for Correlation*
- Homework: problems 2.20, 2.21, 2.23, 2.27, 2.28, 2.30 and 2.31
- See me if you need help!