STA 674

Regression Analysis And Design Of Experiments

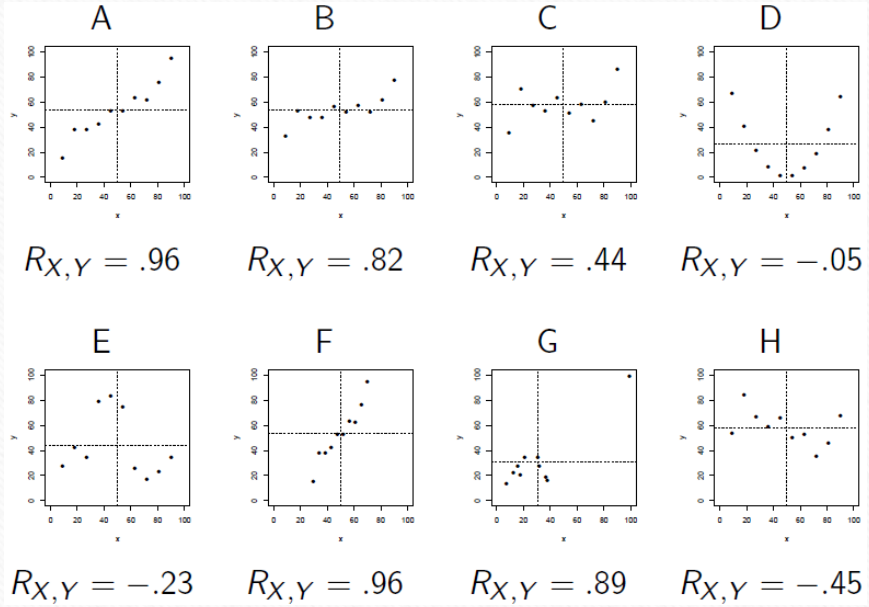
Measuring Association between Two Variables – Lecture 5

STA 674, RADOE:
Measuring Association between Two Variables

- Interpretation of Correlation
  - Last time – magnitude
  - This time – cautions (warnings?)

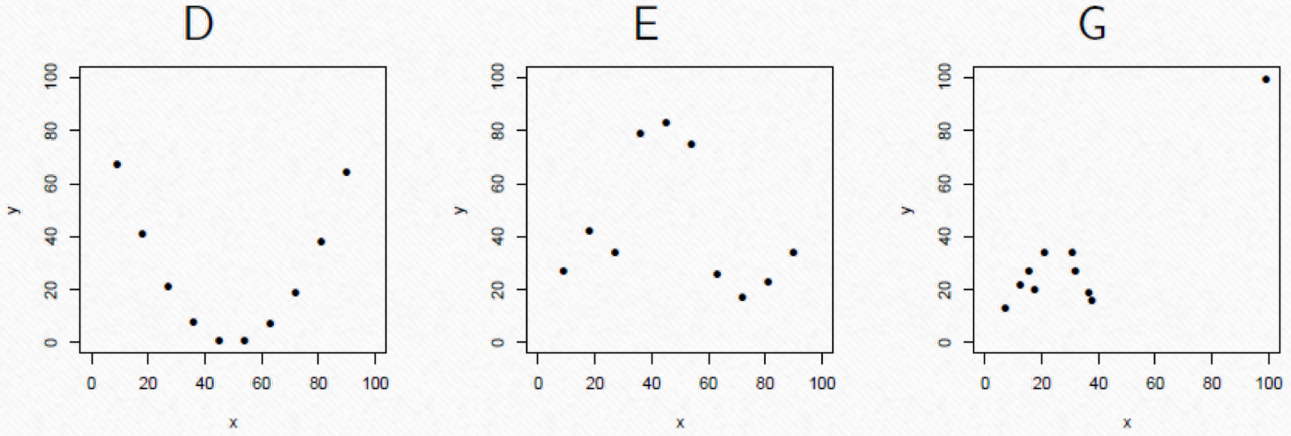
STA 674: Measuring Association between Two Variables Correlation

- The correlation, $R_{X,Y} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{(n-1)s_X s_Y}$, always lies between -1 and 1, inclusive.



STA 674: Measuring Association between Two Variables Correlation

- Caution #1: Recall that correlation is a measure of the strength of linear association: $R_{X,Y}$ is *not* an appropriate measure of the strength of association if the relationship between X and Y is *not linear*.
- From the exercise:
- Moral: **Plot your data!**



$R_{X,Y} = -.05$ $R_{X,Y} = -.23$ $R_{X,Y} = .89$

STA 674: Measuring Association between Two Variables Correlation

- Caution #2: Correlation does not imply causation: There are many reasons that two variables may be strongly correlated. The fact that $R_{X,Y}$ is close to -1 or 1 does not mean that changes in X cause changes in Y or vice versa.
- Examples:
  - Temperatures in Lexington and Johannesburg will have a strong negative correlation over one year.
  - Total daily ice cream sales in the U.S. are highly correlated with the number of drowning deaths.
  - As the number of pirates has decreased, there has been an increase in the number of cell phone towers. Therefore, pirates were destroying the cell phone towers.
- Moral: **Use your brain!**

STA 674: Measuring Association between Two Variables Correlation

- Caution #3: When analyzing grouped data, correlations within individual groups may be obscured or even reversed when looking at data aggregated over all groups (Simpson's paradox).
- Example:

X	Y
0.3	1.8
0.7	2.3
0.7	2.2
0.9	2.6
1	2.2
1.3	0.8
1.7	1.3
1.7	1
1.9	1.5
2	1.4

$$R = -0.65$$

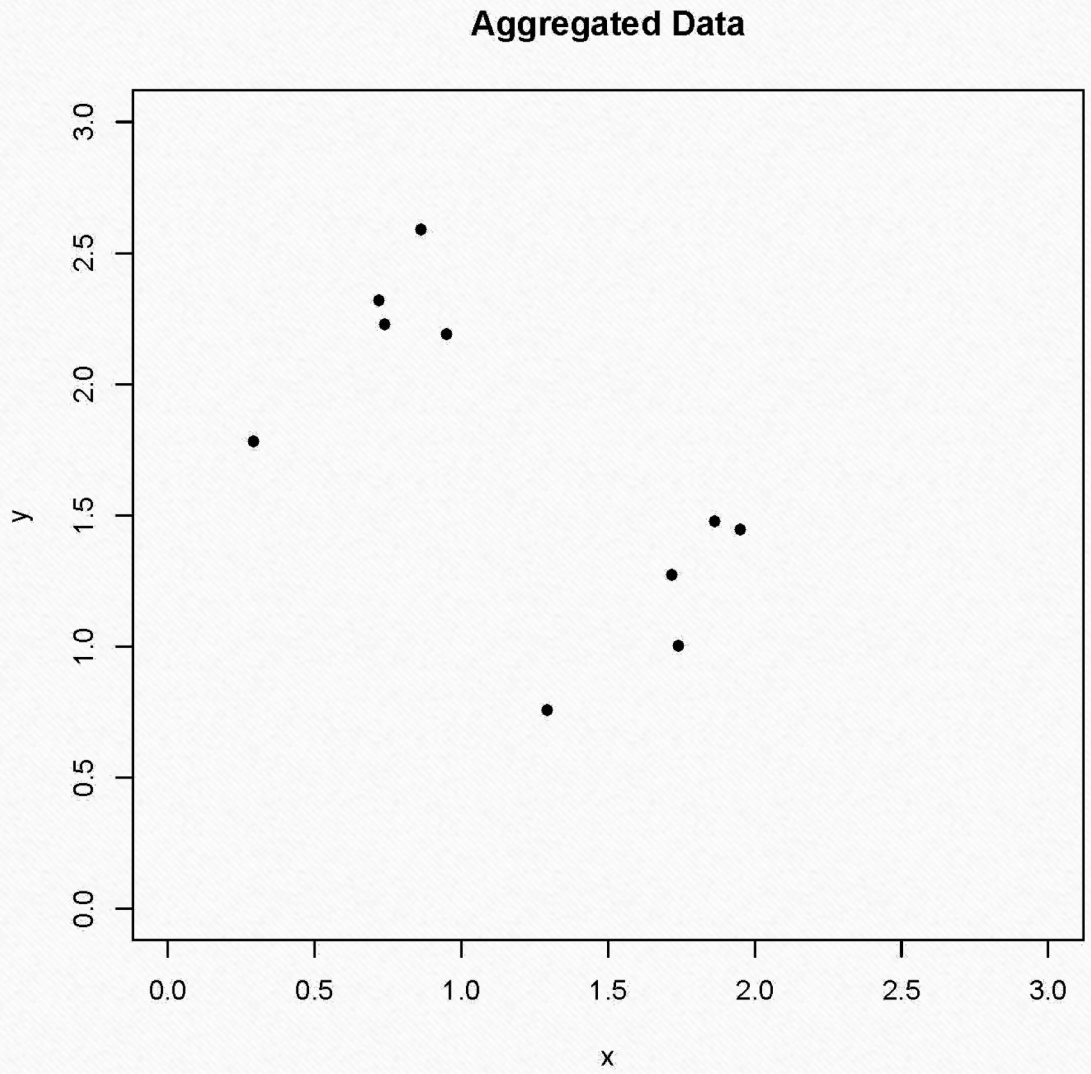
STA 674: Measuring Association between Two Variables Correlation

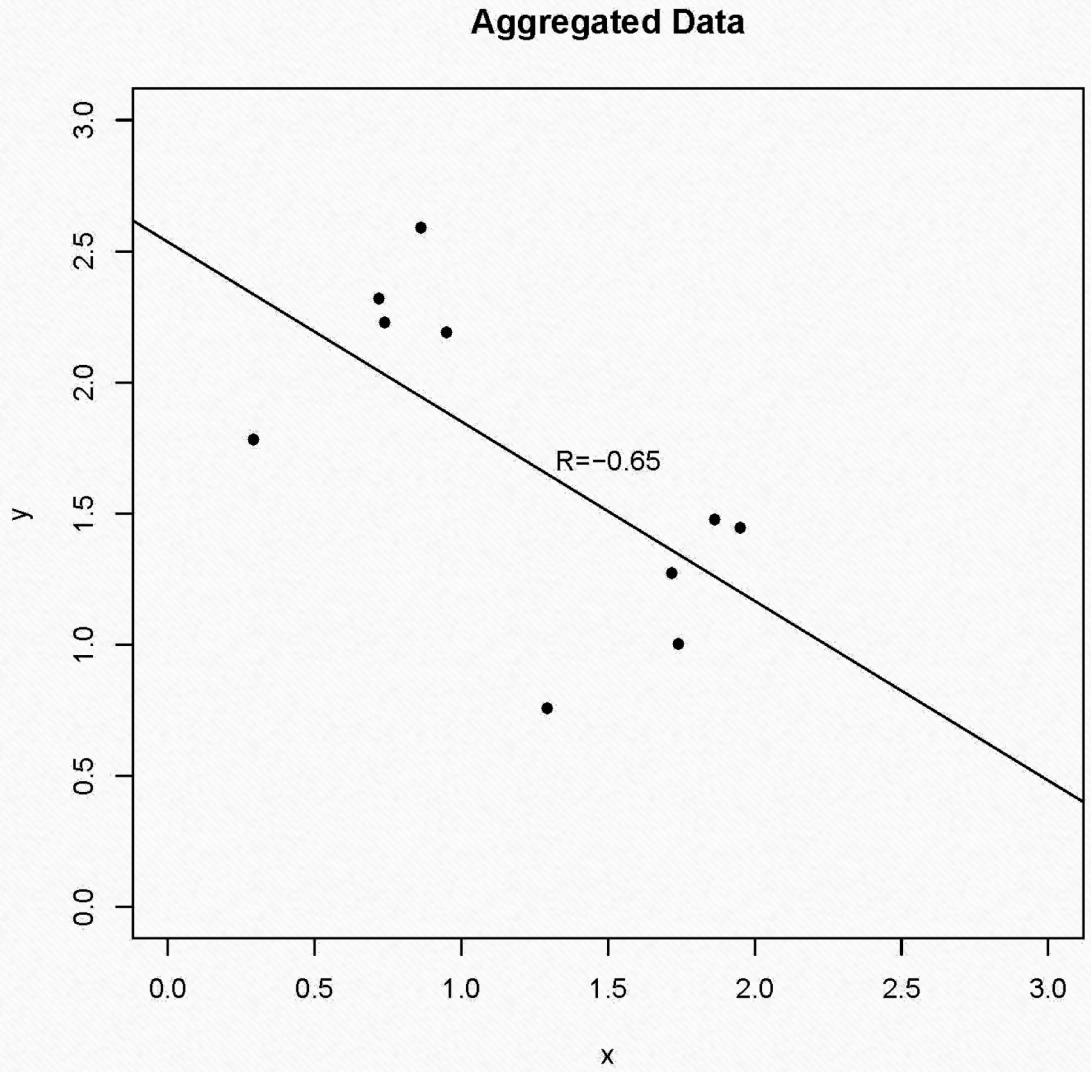
- Caution #3: When analyzing grouped data, correlations within individual groups may be obscured or even reversed when looking at data aggregated over all groups (Simpson's paradox).
- Example:

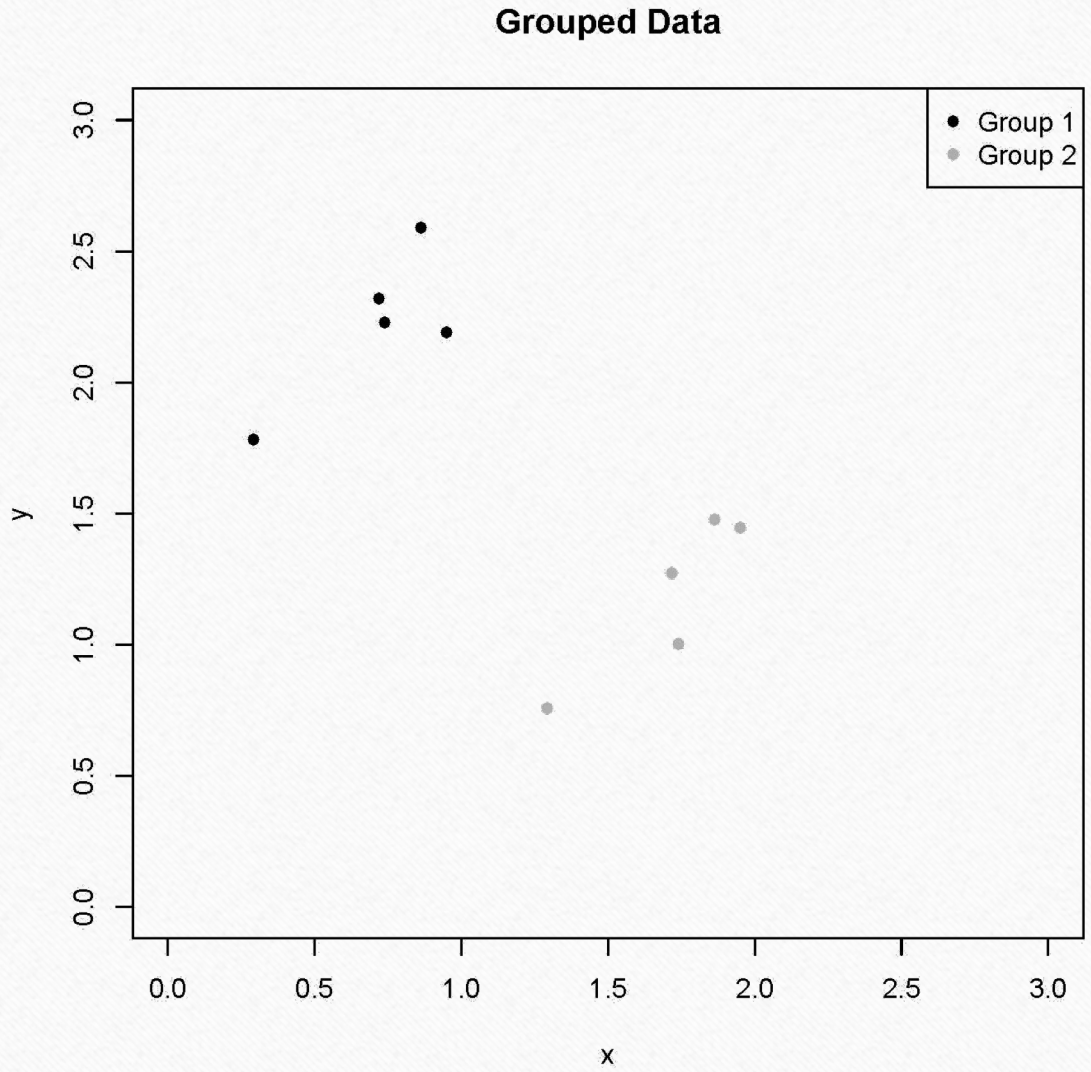
X	Y
0.3	1.8
0.7	2.3
0.7	2.2
0.9	2.6
1	2.2
1.3	0.8
1.7	1.3
1.7	1
1.9	1.5
2	1.4

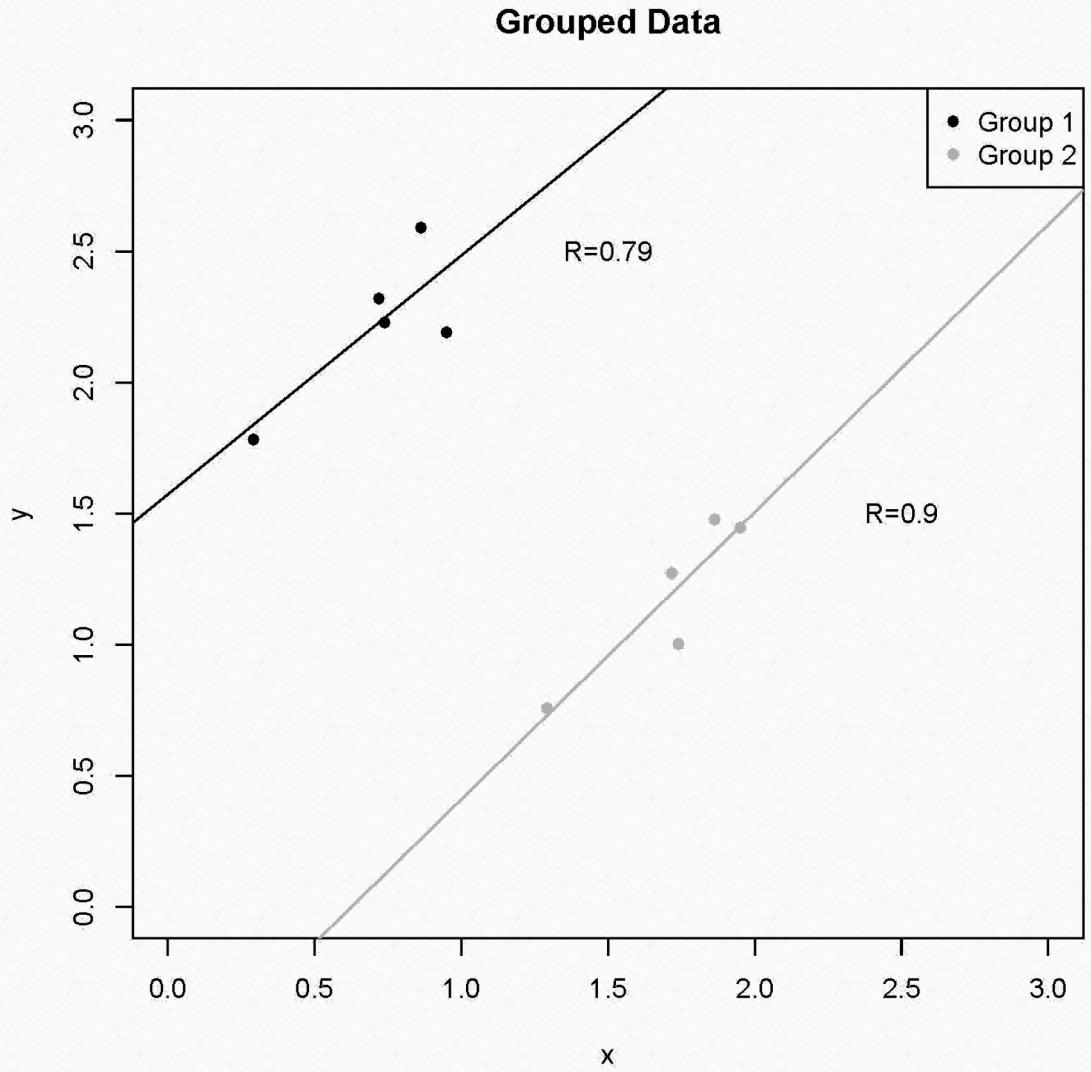
$R = 0.79$

$R = 0.90$


Aggregated Data
y
x
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Aggregated Data
R=−0.65
y
x
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Grouped Data
Group 1
Group 2
y
x
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Grouped Data
Group 1
Group 2
R=0.79
R=0.9
y
x
0.0
0.5
1.0
1.5
2.0
2.5
3.0

STA 674, RADOE:
Measuring Association between Two Variables

- Correlation
  - What next? Fitting Simple Linear Regression Models