Chapter 4.1

Scatter plots and correlation

Julian Chan

Department of Mathematics
Weber State University

September 15, 2011
In statistics we are ultimately interested in collecting data for one of two reasons:
In statistics we are ultimately interested in collecting data for one of two reasons:

1. To predict an outcome of an event.
In statistics we are ultimately interested in collecting data for one of two reasons:

1. To predict an outcome of an event
2. To determine which set of response variables affects to response variable.
What is a natural starting point to determine if a relationship exists between two variables?
What is a natural starting point to determine if a relationship exists between two variables?

Draw a picture!
1. What is a natural starting point to determine if a relationship exists between two variables?

2. Draw a picture!

3. In particular we will draw a scatter diagram. Each individual is represented by a point on the scatter diagram. The explanatory variable is plotted on the horizontal axis and the response variable on the vertical axis.
A doctor wanted to determine whether a relation exists between a male's age and his HDL cholesterol. From a simple random sample he obtained the following data.
Examples

Table: Age and HDL

<table>
<thead>
<tr>
<th>Age</th>
<th>HDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>57</td>
</tr>
<tr>
<td>42</td>
<td>54</td>
</tr>
<tr>
<td>46</td>
<td>34</td>
</tr>
<tr>
<td>32</td>
<td>56</td>
</tr>
<tr>
<td>55</td>
<td>35</td>
</tr>
<tr>
<td>52</td>
<td>40</td>
</tr>
<tr>
<td>61</td>
<td>42</td>
</tr>
<tr>
<td>61</td>
<td>38</td>
</tr>
<tr>
<td>26</td>
<td>47</td>
</tr>
<tr>
<td>38</td>
<td>44</td>
</tr>
<tr>
<td>66</td>
<td>62</td>
</tr>
<tr>
<td>52</td>
<td>48</td>
</tr>
</tbody>
</table>
Examples

1. We now provide a scatter plot of the data:
We now provide a scatter plot of the data:

Scatter plot of Age Vs. HDL
The correlation coefficient for the scatter plot above is \( r = -0.2 \).
The American black bear is one of eight bear species in the world. It is the smallest North American bear and the most common bear species. In 1969 Dr. Pelton initiated a long term study of the population. One aspect of the study was to develop a model that could be used to predict a bear's weight. One variable thought to be related to the bear's weight is the length of the bear.
The American black bear is one of eight bear species in the world. It is the smallest North American bear and the most common bear species. In 1969 Dr. Pelton initiated a long term study of the population. One aspect of the study was to develop a model that could be used to predict a bear's weight. One variable thought to be related to the bear's weight is the length of the bear.

What is the response and explanatory variable?
The American black bear is one of eight bear species in the world. It is the smallest North American bear and the most common bear species. In 1969 Dr. Pelton initiated a long term study of the population. One aspect of the study was to develop a model that could be used to predict a bear's weight. One variable thought to be related to the bear's weight is the length of the bear.

What is the response and explanatory variable?

we collect the following data.
### Table: Age and HDL

<table>
<thead>
<tr>
<th>Length(cm)</th>
<th>Weighth(Kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>139</td>
<td>60</td>
</tr>
<tr>
<td>138</td>
<td>65</td>
</tr>
<tr>
<td>139</td>
<td>68</td>
</tr>
<tr>
<td>120.5</td>
<td>50</td>
</tr>
<tr>
<td>148</td>
<td>85</td>
</tr>
<tr>
<td>141</td>
<td>100</td>
</tr>
<tr>
<td>141</td>
<td>90</td>
</tr>
<tr>
<td>150</td>
<td>120</td>
</tr>
<tr>
<td>166</td>
<td>155</td>
</tr>
<tr>
<td>151.5</td>
<td>144</td>
</tr>
<tr>
<td>129.5</td>
<td>72</td>
</tr>
<tr>
<td>150</td>
<td>148</td>
</tr>
</tbody>
</table>
We now provide a scatter plot of the data:
We now provide a scatter plot of the data:

**Length Vs. Weight**

- **Weight**
  - 0
  - 20
  - 40
  - 60
  - 80
  - 100
  - 120
  - 140
  - 160
  - 180

- **Length**
  - 110
  - 120
  - 130
  - 140
  - 150
  - 160
  - 170

Column B
The correlation coefficient for the scatter plot above is $r = .85$. 
There are two fundamental differences in the two scatter plots of the data!
There are two fundamental differences in the two scatter plots of the data!

One has a slight downward trend and the other has a general upward trend!
A downward trend is what statisticians call a negative association.

Two variables that are linearly related are said to be negatively associated when above average values of one variable are associated to above average values of the other variable.
Examples

1. A downward trend is what statisticians call a negative association.

2. Two variables that are linearly related are said to be negatively associated when above below average values of one variable are associated to above average values of the other variable.
A upward trend is what statisticians call a positive association.
Examples

1. A upward trend is what statisticians call a positive association.

2. Two variables that are linearly related are said to be positively associated when above average values of one variable are associated to above average values of the other variable.
What did you notice about $r$ in the first example?
What did you notice about $r$ in the first example?

- It was negative!
$1$ What did you notice about $r$ in the first example?

$2$ It was negative!

$3$ What does this reflect?
Examples

1. What did you notice about $r$ in the first example?
2. It was negative!
3. What does this reflect?
4. Negative correlation!
Examples

1. What did you notice about $r$ in the second example?
What did you notice about $r$ in the second example?

It was positive!
Examples

1. What did you notice about $r$ in the second example?
2. It was positive!
3. What does this reflect?
What did you notice about $r$ in the second example?

It was positive!

What does this reflect?

Positive correlation!
This suggests that $r$ which is called the linear correlation coefficient is useful.
This suggests that $r$ which is called the linear correlation coefficient is useful.

How can one compute this number?
This suggests that $r$ which is called the linear correlation coefficient is useful.

How can one compute this number?

$$r = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y} / (n - 1)$$
Properties of the linear correlation coefficient.
Introduction

Examples

1. Properties of the linear correlation coefficient.
2. \(-1 \leq r \leq 1\).
Properties of the linear correlation coefficient.

- $-1 \leq r \leq 1$.

- If $r = 1$ then there is perfect positive correlation.

- The closer $r$ is to 0 is more indication of NO LINEAR relation between the variables.

- $r$ is not resistant meaning that an observation which does not follow the general pattern of the data can change the value of $r$!
Examples

1. Properties of the linear correlation coefficient.
2. $-1 \leq r \leq 1$.
3. If $r = 1$ then there is perfect positive correlation.
4. If $r = -1$ then there is perfect negative correlation.
Properties of the linear correlation coefficient.

- $-1 \leq r \leq 1$.

- If $r = 1$ then there is perfect positive correlation.

- If $r = -1$ then there is perfect negative correlation.

- The closer $r$ is to 0 is more indication of NO LINEAR relation between the variables.
Examples

1. Properties of the linear correlation coefficient.
2. $-1 \leq r \leq 1$.
3. If $r = 1$ then there is perfect positive correlation.
4. If $r = -1$ then there is perfect negative correlation.
5. The closer $r$ is to 0 is more indication of NO LINEAR relation between the variables.
6. $r$ is not resistant meaning that an observation which does not follow the general pattern of the data can change the value of $r$!
Examples

1. You tell me. Positive, negative, or no association.
Examples

**Scatter Plot Example - Strong Negative Correlation**

$y = -1.223x + 83.967$

$R^2 = 0.8941$
Examples

Scatter Plot of State’s Average Temperature and Tax Rates
We practice computing the linear correlation coefficient with the data below.
**Table: Age and HDL**

<table>
<thead>
<tr>
<th>Length(cm)</th>
<th>Weight(Kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>139</td>
<td>60</td>
</tr>
<tr>
<td>138</td>
<td>65</td>
</tr>
<tr>
<td>139</td>
<td>68</td>
</tr>
<tr>
<td>120.5</td>
<td>50</td>
</tr>
</tbody>
</table>
We find that $\tilde{X}_L = 134.13$ and $\tilde{Y}_W = 60.75$. We compute sample standard deviation and find that $S_x = 9.1$, and $S_y = 7.89$. We find that the correlation coefficient is $r = 0.9$. 
We find that $\bar{X}_L = 134.13$ and $\bar{Y}_W = 60.75$.

We compute sample standard deviation and find that $S_x = 9.1$, and $S_y = 7.89$. 

1. We find that $\bar{X}_L = 134.13$ and $\bar{Y}_W = 60.75$

2. We compute sample standard deviation and find that $S_x = 9.1$, and $S_y = 7.89$. 

1 We find that $\bar{X}_L = 134.13$ and $\bar{Y}_W = 60.75$
2 We compute sample standard deviation and find that $S_x = 9.1$, and $S_y = 7.89$.
3 We find that the correlation coefficient is $r = .9$. 
To determine if linear correlation exists between an explanatory variable and a response variable. We compute the absolute value of $r$ and see if it is greater than the appropriate value in appendix A table 2.
To determine if linear correlation exists between an explanatory variable and a response variable. We compute the absolute value of $r$ and see if it is greater than the appropriate value in appendix A table 2.

For the length and weight of bears we found above using a smaller data set that $r = .9$
Lurking variables can cause a higher or lower correlation coefficient!
Lurking variables can cause a higher or lower correlation coefficient!

A researcher wants to know if drinking cola decreases Bone density in women. The data shows that there is negative correlation suggestion evidence for this claim. However, there are a number of lurking variables such as other consumption or health factors! This data was also collected over time therefore it could be the case that naturally these womens done density has decreased with age!