STAT 3355
Introduction to Data Analysis

Lecture 06: Summaries for Univariate Data I

Created by: Qiwei Li
Assistant Professor of Statistics
Presented by: Octavious Smiley
Assistant Professor of Instruction

Department of Mathematical Sciences
The University of Texas at Dallas

Learning Goals

- Graphical representation for two continuous data
  - Scatter plot
- Numerical summaries for two continuous data
  - Pearson correlation coefficient
  - Spearman's rank correlation coefficient

Continuous Data

- Unlikely for multiple samples to share the same value
- Data type
  - Integer (if the number of unique values is large)
  - Numeric data
- Examples
  - The height of person in cm
  - The weight of a person in lb
  - The age of a person in year

Paired Continuous Data

- Denote two univariate continuous data by

$$\boldsymbol{x} = [x_1, \ldots, x_i, \ldots, x_n], \text{ where } x_i \in \mathbb{R}$$
$$\boldsymbol{y} = [y_1, \ldots, y_i, \ldots, y_n], \text{ where } y_i \in \mathbb{R}$$

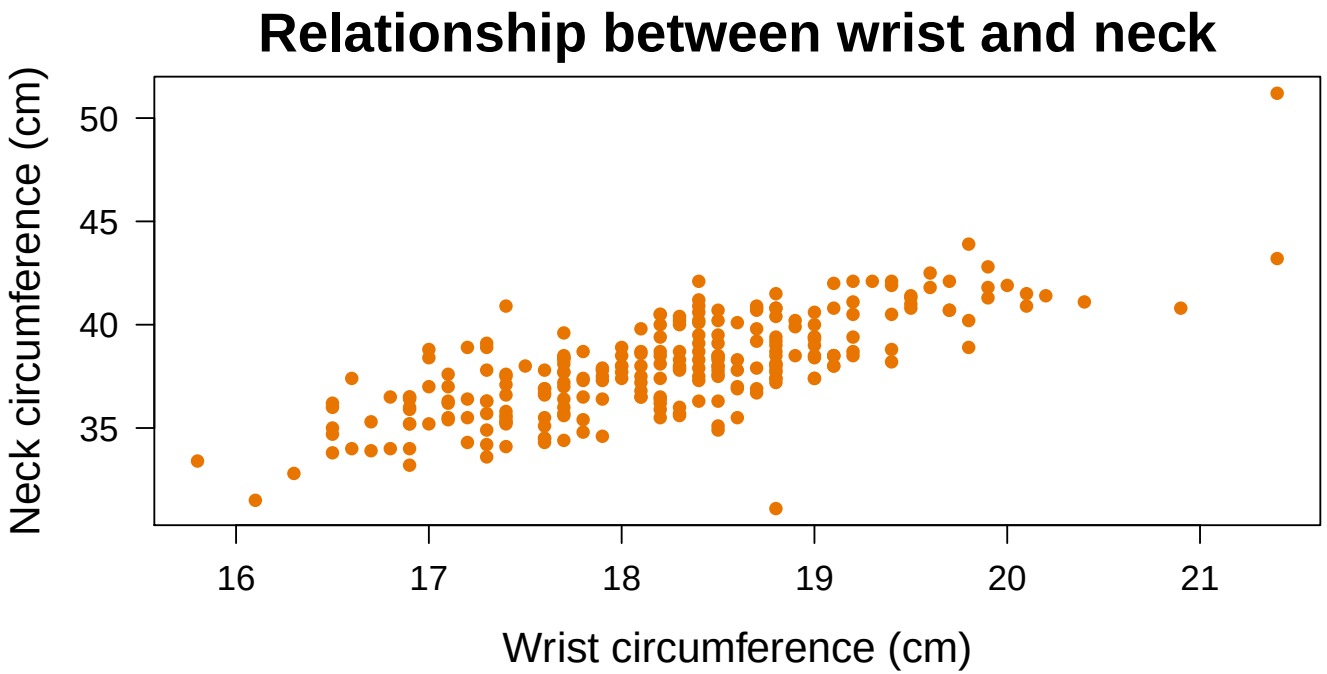
- The data under consideration is naturally paired off

$$(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$$

- Each pair of data is a point on a Cartesian plane
  - One x axis and one y axis, which are perpendicular to each other
  - The origin $(0, 0)$ is the reference point

Scatter Plot

- Displays two continuous data in a Cartesian plane
  - One axis represents the value of a continuous data
  - The other axis represents the value of the other continuous data
  - Each point corresponds to a sample



Scatter Plot

- Implementation in R
  - $\boldsymbol{x}$ and $\boldsymbol{y}$ are two numeric vectors
    - `plot(`x`, `y`)`
    - `plot(`$y \sim x$`)`
  - $\boldsymbol{x}$ and $\boldsymbol{y}$ are two numeric variables in a data frame $\boldsymbol{D}$
    - `plot(`D`$x_name, `D`$y_name)`
    - `plot(y_name` $\sim$ `x_name, data =` D`, subset = )`

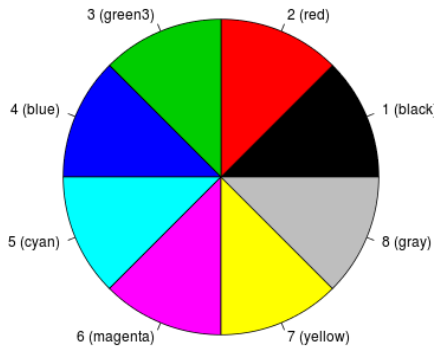
Scatter Plot

- Examples

```
library(UsingR)

# Load data
data("fat")

# Draw a scatter plot for neck against wrist
plot(fat$wrist, fat$neck)
plot(neck ~ wrist, data = fat)

# Draw a scatter plot for height against
    wrist
plot(height ~ wrist, data = fat)
plot(height ~ wrist, data = fat, subset =
    height > 50)
```

The Function `plot()`

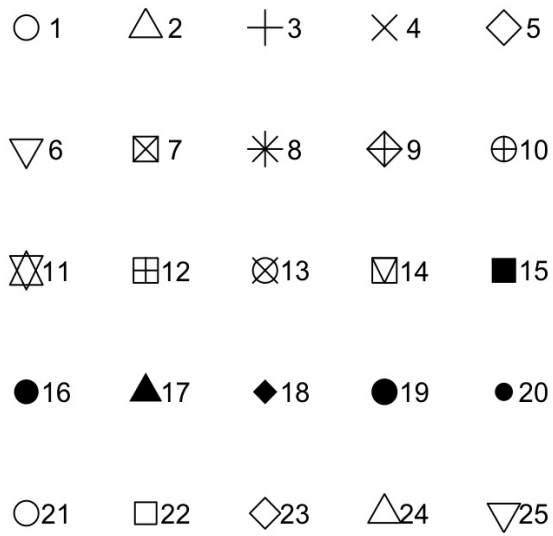
- Important arguments controlling the point looks (i.e. aesthetic)
  - `col`: A color value/vector for point(s)
    - Basic eight colors



    - Colors in R
    - `rgb(red = , green = , blue = , alpha =)`, e.g.
      UTD orange: `rgb(red = 223 / 255, green = 117 / 255, blue = 0 / 255)` or "#e87500"
      UTD green: `rgb(red = 18 / 255, green = 71 / 255, blue = 52 / 255)` or "#154734"

The Function `plot()`

- Important arguments controlling the point looks (i.e. aesthetic)
  - `pch`: A numeric shape value/vector for point(s)
    - Basic shapes



  - `cex`: A numeric size value/vector for point(s)

The Function `plot()`

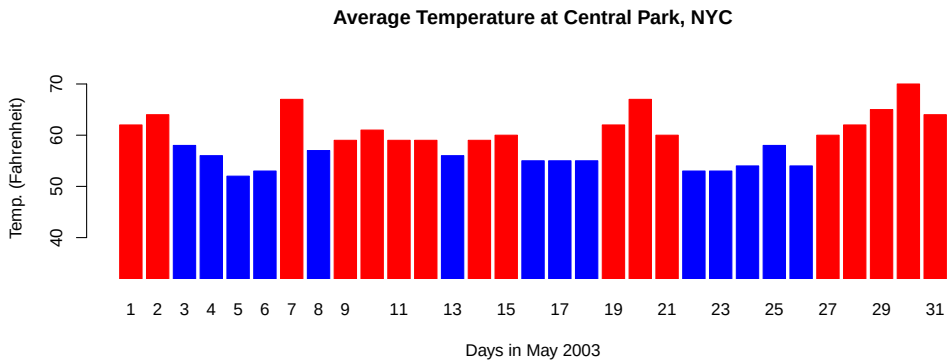
- Important arguments controlling the Cartesian plane
  - `xlim` or `ylim`: A numerical vector of two values indicating the limits for x and y axis, respectively
  - `asp`: A numerical value for the y/x aspect ratio
  - `axes`: A logical value for displaying both x and y axes or not
- Draw an empty Cartesian plane by `plot(NULL, xlim = c(xmin, xmax), ylim = c(ymin, ymax))`
  - Use `par()` to set graphical parameters
  - A square plotting region: `par(pty = "s")`
  - A maximal plotting region (default): `par(pty = "m")`

The Function `plot()`

- Important arguments controlling the labels
  - `main`: A string of title
  - `xlab` and `ylab`: A string of label for x axis name and y axis name, respectively
  - `las`: A numeric value of $\{0, 1, 2, 3\}$ for the orientation of axis tick labels
- Important arguments controlling the label size
  - `cex.main`: A numeric value for the title size
  - `cex.lab`: A numeric value for the size of axis labels
  - `cex.axis`: A numeric value for the size of x axis label

Your Turn

- We have learned how to use `barplot()` to make a bar plot of the average daily temperature in the dataset `central.park` in the package `UsingR`



  - Draw a scatter plot, where x axis represents the day and y axis represents the average temperature
  - Make the plot more informative by adding 1) a title, 2) label for x axis, 3) label for y axis
  - Try your favorite point look by changing 1) point color, 2) point shape, 3) point size

Your Turn

- Examples

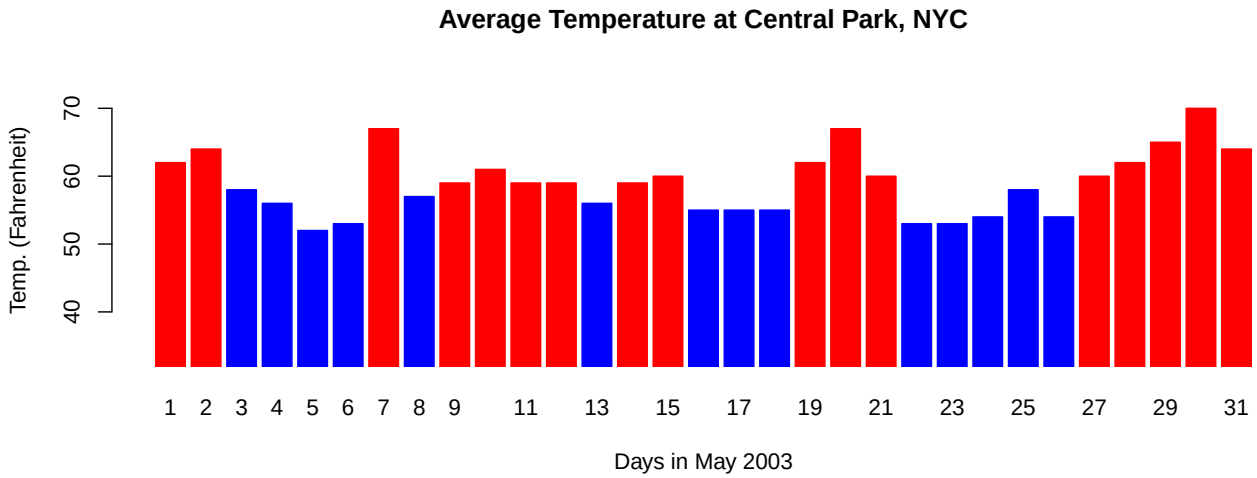
```
# Load data
data("central.parks")

# Draw a scatter plot
plot(AVG ~ DY, data = central.park)

# Add a title and axis labels
plot(AVG ~ DY, data = central.park, xlab = "
    Days in May 2003", ylab = "Temp. (
    fahrenheit)", main = "Average Temperature
     at Central Park, NYC")
```

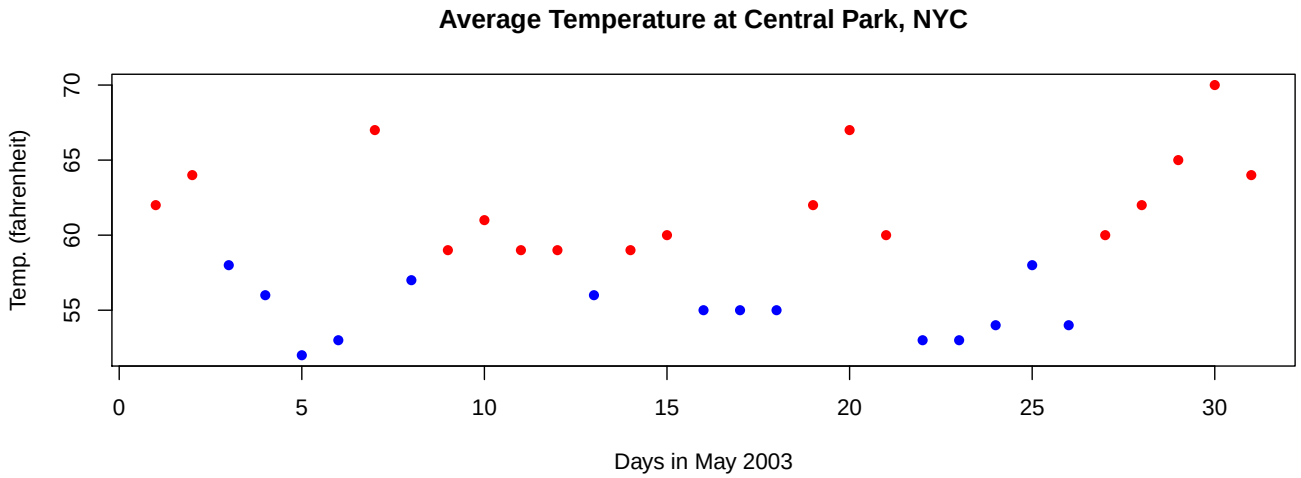
Your Turn

- Examples

```
# Change the point look
plot(AVG ~ DY, data = central.park, xlab = "
    Days in May 2003", ylab = "Temp. (
    fahrenheit)", main = "Average Temperature
     at Central Park, NYC", pch = 16, col = "
    #e87500")

# Make color indicate if the daily
    temperature was greater the average
plot(AVG ~ DY, data = central.park, xlab = "
    Days in May 2003", ylab = "Temp. (
    fahrenheit)", main = "Average Temperature
     at Central Park, NYC", pch = 16, col =
    ((AVG < mean(AVG)) + 1) * 2)
```

Your Turn


Average Temperature at Central Park, NYC
Temp. (Fahrenheit)
40
50
60
70
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31
Days in May 2003


Average Temperature at Central Park, NYC
Temp. (fahrenheit)
55
60
65
70
0
5
10
15
20
25
30
Days in May 2003

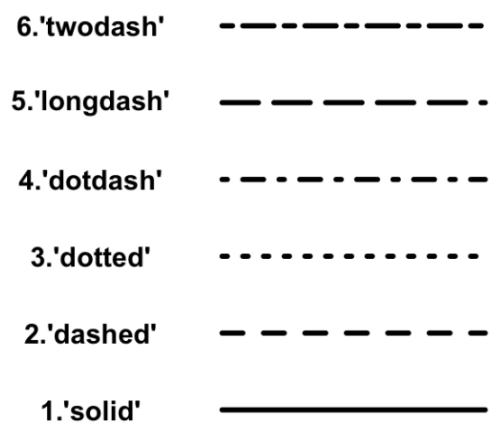
The Function `plot()`

- Important arguments controlling the type of plot
  - `type`: A character to control the type of plot

Type	Value
"p"	Points only
"l" and "c"	A line that connects each point sequentially
"b" and "o"	Both points and the line
"h"	"Histogram" like vertical lines
"s" and "S"	Stair steps

The Function `plot()`

- Important arguments controlling the line looks
  - `col`: A color value of line
  - `lty`: A numeric value of line type



  - `lwd`: A numeric value of line width

Other Related Plotting Functions

- `points(`x`,` y`)`: Add points to a graphic, where x and y are coordinate vectors of points to plot
- `col`, `pch`, and `cex` are important arguments for the function `points()`

The Function points()

- Examples

```
# Load data
data("central.park")

# Draw an empty figure
plot(NULL, xlim = c(min(central.park$DY),
    max(central.park$DY)), ylim = c(min(
    central.park$AVG), max(central.park$AVG))
    , xlab = "Days in May 2003", ylab = "Temp
    . (fahrenheit)")

# Add points
points(central.park$DY, central.park$AVG,
    pch = 16)
```

Other Related Plotting Functions

- `lines(`x`, `y`)`: Add line segments to a graphic, where x and y are coordinate vectors of points to join
- `abline(a = , b = )`, `abline(h = )`, or `abline(v = )`: Add a line to a graphic,
  - `a` and `b`: Intercept and slope
  - `h`: a number or a numeric vector of y value(s) for horizontal line(s)
  - `v`: a number or a numeric vector of x value(s) for vertical line(s)
- `col`, `lty` and `lwd` are also important arguments for the functions `lines()` and `abline()`

The Function lines() and abline()

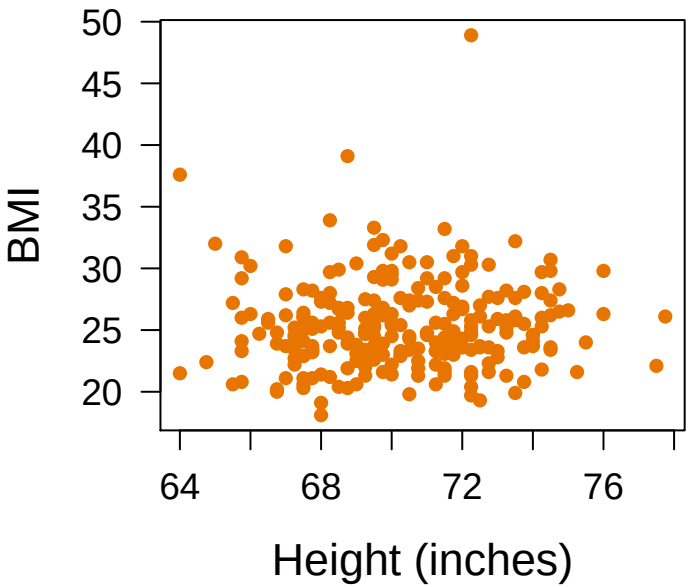
- Examples

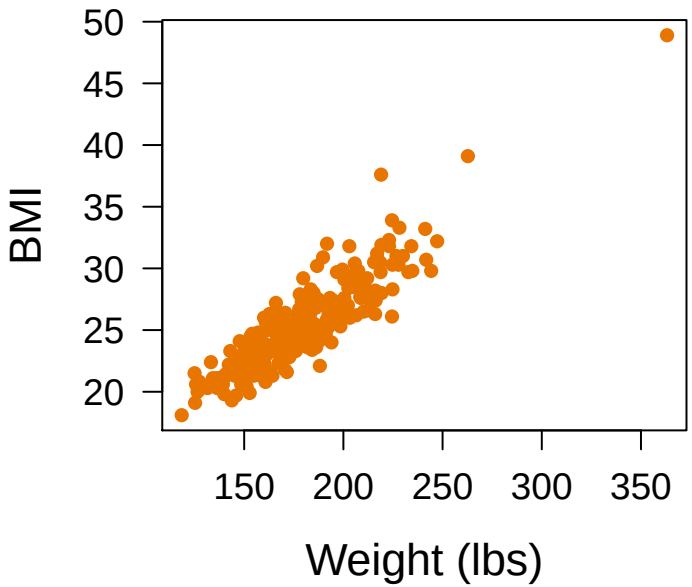
```
# Draw an empty figure
plot(NULL, xlim = c(min(central.park$DY),
    max(central.park$DY)), ylim = c(min(
    central.park$AVG), max(central.park$AVG))
    , xlab = "Days in May 2003", ylab = "Temp
    . (fahrenheit)")

# Add line segements
lines(central.park$DY, central.park$AVG)

# Add a horizontal line indicating the
    average of daily average tempratures
abline(h = mean(central.park$AVG), lty = 2)
```

Measures of Correlation

- Is BMI more linearly correlated with weight than height?

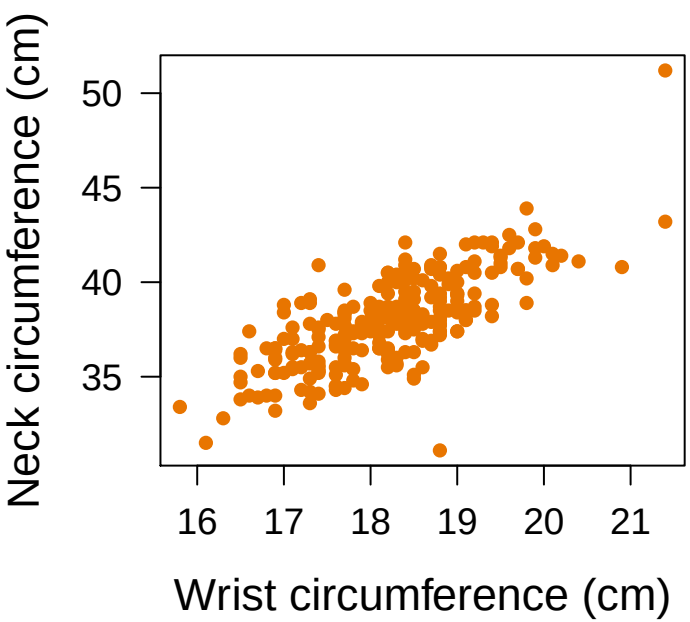
$$\text{BMI} = \frac{\text{weight in kg}}{\text{height in m}^2} = 703\frac{\text{weight in lb}}{\text{height in inch}^2}$$

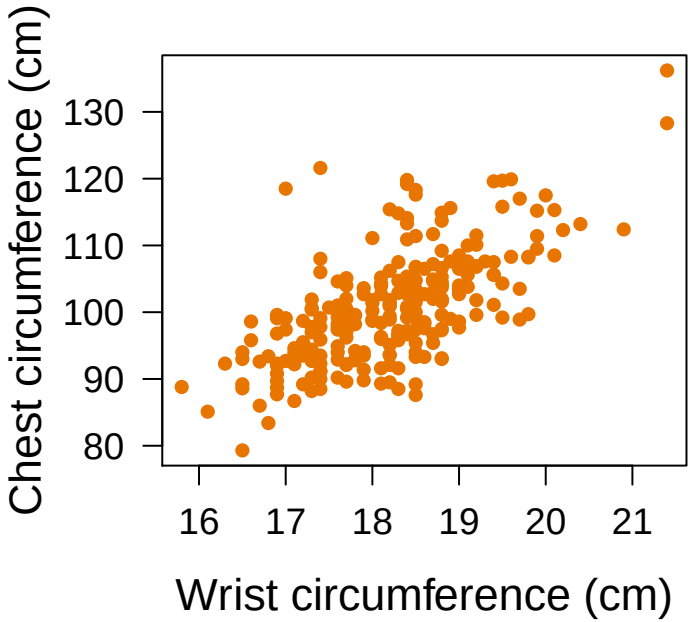




Measures of Correlation

- Is wrist circumference more correlated with neck circumference than chest circumference?





The Sample Covariance Variance

- Denote two univariate continuous data by

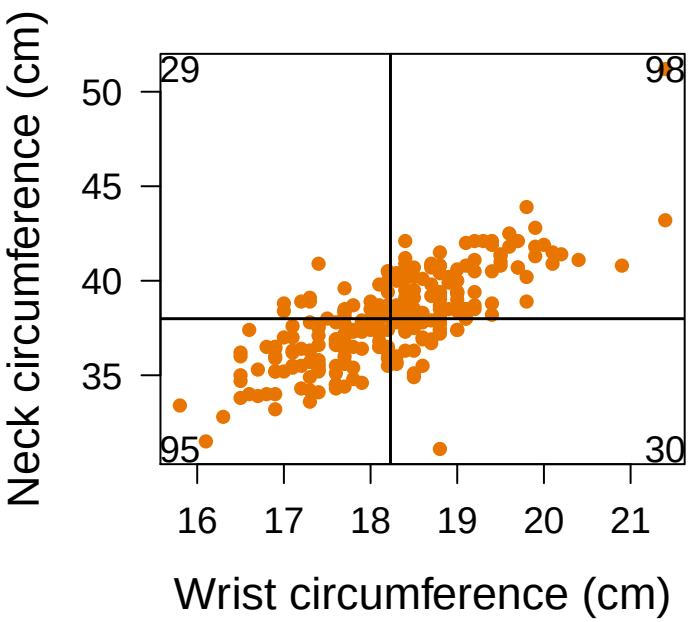
$$\boldsymbol{x} = [x_1, \ldots, x_i, \ldots, x_n], \text{ where } x_i \in \mathbb{R}$$
$$\boldsymbol{y} = [y_1, \ldots, y_i, \ldots, y_n], \text{ where } y_i \in \mathbb{R}$$

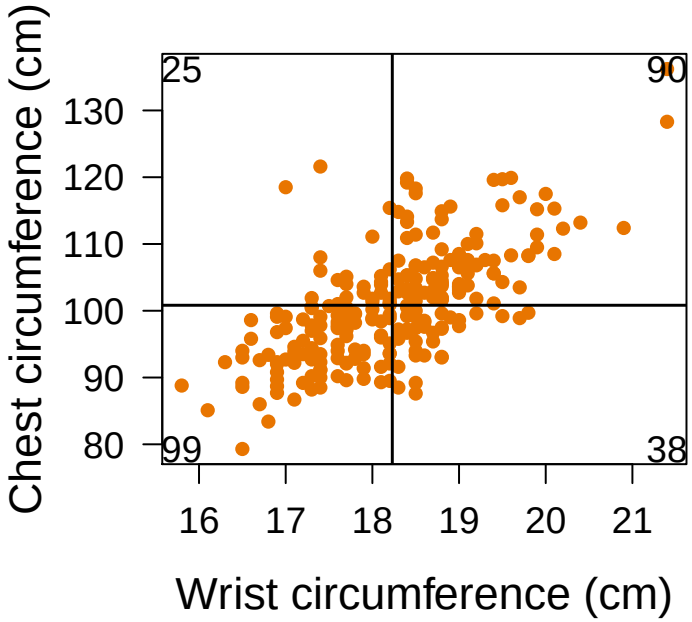
- The sample covariance is

$$\text{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- Interpretation
  - Large absolute value corresponds to large covariance
  - $\text{Cov}(\boldsymbol{x}, \boldsymbol{x}) = s_x^2$ and $\text{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \text{Cov}(\boldsymbol{y}, \boldsymbol{x})$
  - Isolate points into four regions determined by their values comparing with the means

The Sample Covariance Variance

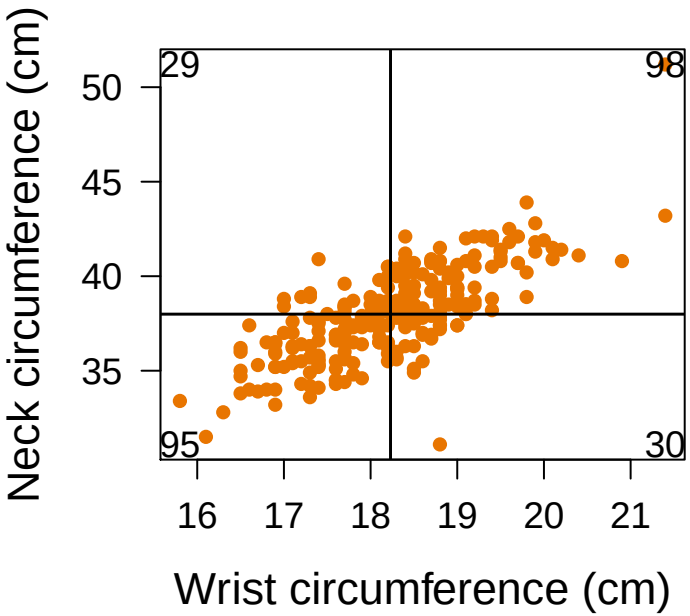


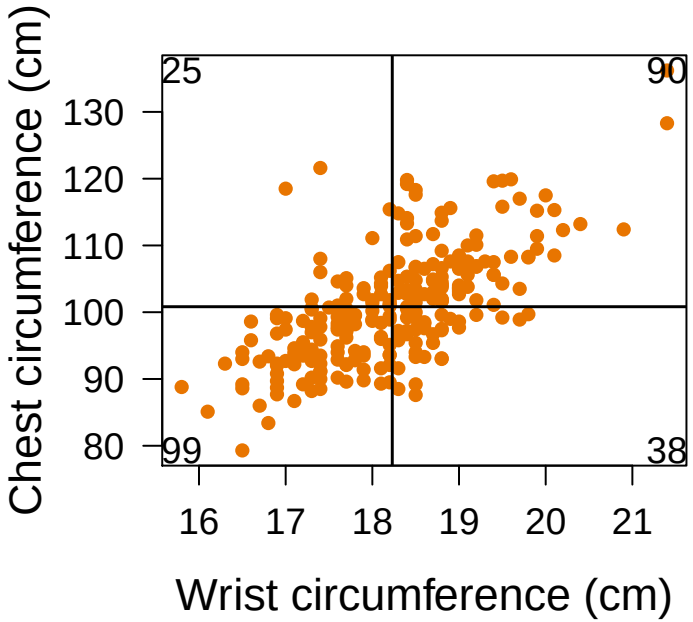


The Sample Covariance Variance

- Implementation in R
  - $\boldsymbol{x}$ and $\boldsymbol{y}$ are two numeric vectors
    - `cov(`x`, `y`)`
  - $\boldsymbol{x}$ and $\boldsymbol{y}$ are two numeric variables in a data frame $\boldsymbol{D}$
    - `cov(`D`$x_name, `D`$y_name)`

The Sample Covariance Variance



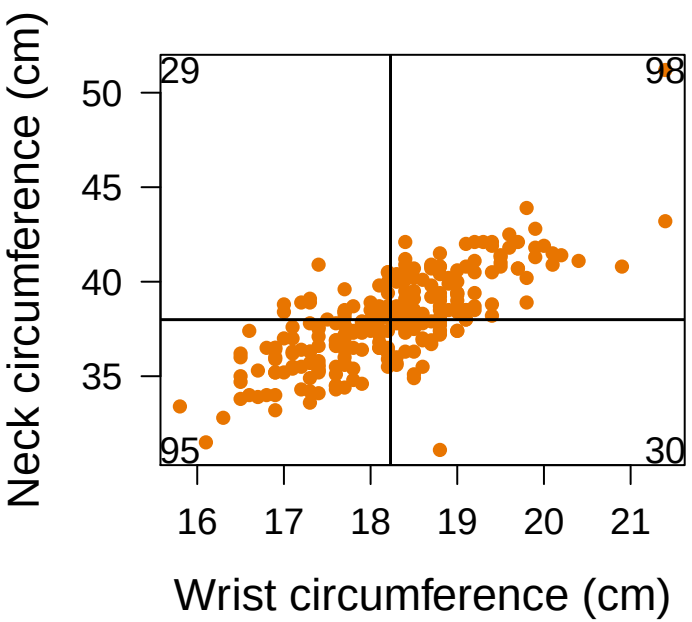


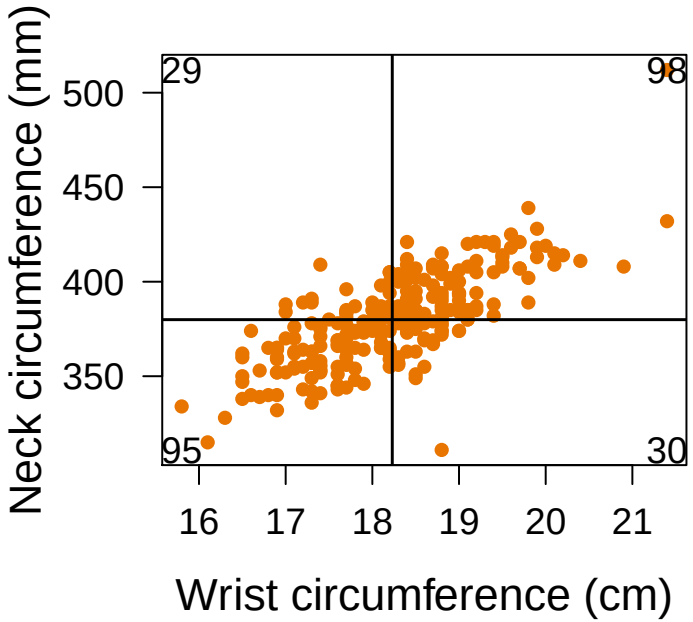
$\text{Cov}(\text{Wrist}, \text{Neck}) = 1.69$

$\text{Cov}(\text{Wrist}, \text{Chest}) = 5.20$

- $s^2_{\text{Neck}} = 5.91$
- $s^2_{\text{Chest}} = 71.07$

The Sample Covariance Variance





$\text{Cov}(\text{Wrist}, \text{Neck (cm)}) = 1.69$ $\text{Cov}(\text{Wrist}, \text{Neck (mm)}) = 16.90$

The Pearson Correlation Coefficient

- Denote two univariate continuous data by

$$\boldsymbol{x} = [x_1, \ldots, x_i, \ldots, x_n], \text{ where } x_i \in \mathbb{R}$$
$$\boldsymbol{y} = [y_1, \ldots, y_i, \ldots, y_n], \text{ where } y_i \in \mathbb{R}$$

- The Pearson correlation coefficient is

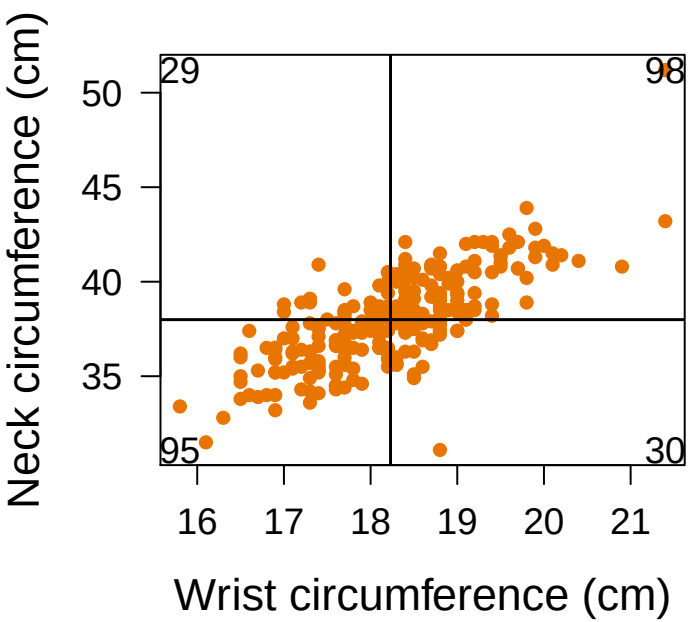
$$\rho_{x,y} = \frac{\text{Cov}(\boldsymbol{x}, \boldsymbol{y})}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- Interpretation
  - The Cauchy-Schwarz inequality: $\left(\sum_{i=1}^n v_i u_i\right)^2 \leq \sum_{i=1}^n v_i^2 \sum_{i=1}^n u_i^2$
  - $-1 \leq \rho_{x,y} \leq 1$
  - Negative, no, and positive linear correlation

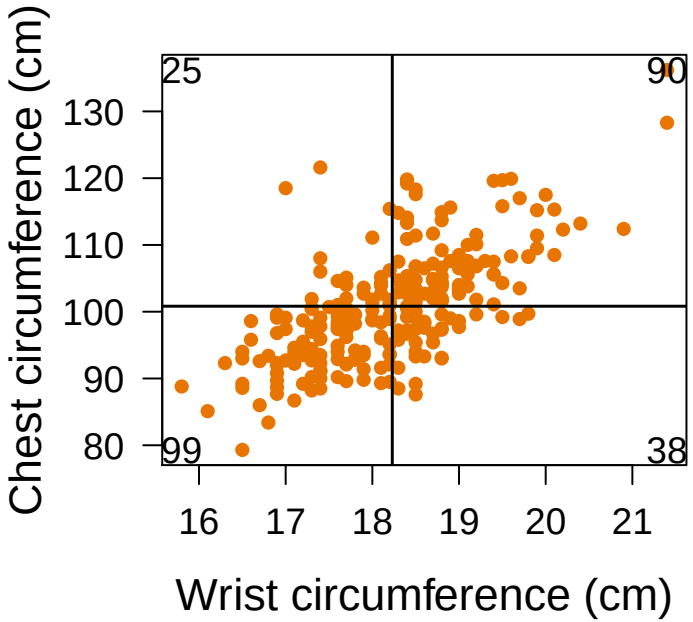
The Pearson Correlation Coefficient

- Implementation in R
  - $\boldsymbol{x}$ and $\boldsymbol{y}$ are two numeric vectors
    - `cor(`x`, `y`, method = "pearson")`
  - $\boldsymbol{x}$ and $\boldsymbol{y}$ are two numeric variables in a data frame $\boldsymbol{D}$
    - `cor(`D`$x_name, `D`$y_name, method = "pearson")`

The Pearson Correlation Coefficient

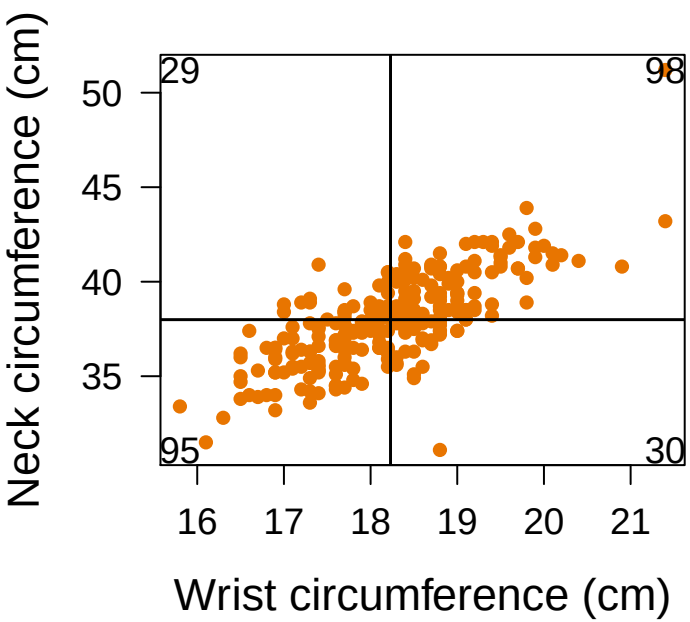


$\rho_{\text{Wrist,Neck}} = 0.75$

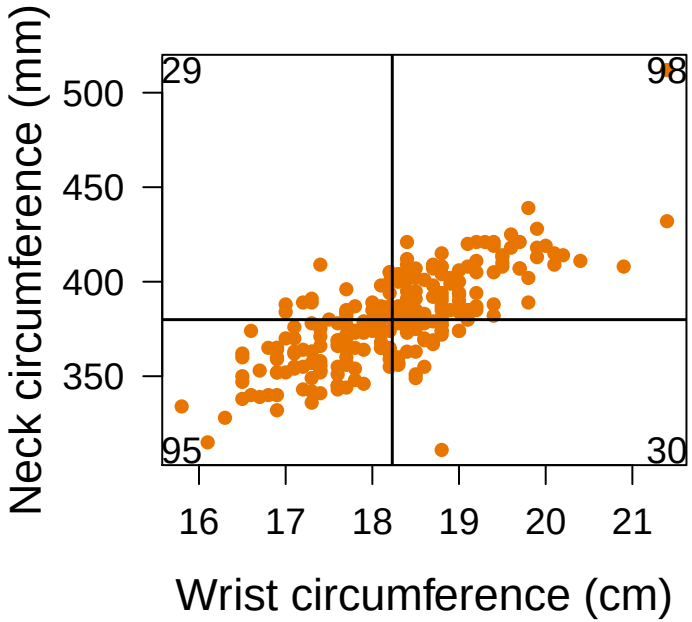


$\rho_{\text{Wrist,Chest}} = 0.66$

The Pearson Correlation Coefficient



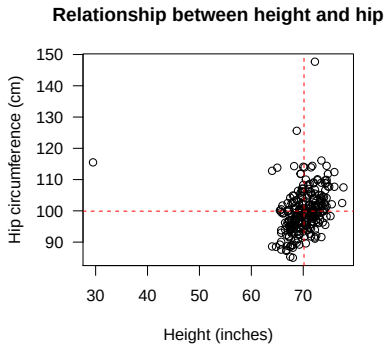
$\rho_{\text{Wrist,Neck (cm)}} = 0.75$



$\rho_{\text{Wrist,Neck (mm)}} = 0.75$

Your Turn

- Reproduce the following scatter plot using the dataset `fat` in the package `UsingR`



  - Draw a scatter plot, where x axis represents the height (inches) and y axis represents the hip circumference (cm)
  - Make the plot more informative by adding 1) a title, 2) label for x axis, 3) label for y axis
  - Add a red horizontal line to indicate the mean of hip and add a red vertical line to indicate the mean of height
  - Calculate the Pearson correlation coefficient

Your Turn

- Solutions

```
# Load data
data("fat")

# Draw a scatter plot
plot(hip ~ height, data = fat)

# Add a title and axis labels
plot(hip ~ height, data = fat, las = 1, xlab
     = "Height (inches)", ylab = "Hip
    circumference (cm)", main = "Relationship
     between height and hip")
```

Your Turn

- Solutions

```
# Add a red horizontal dashed line to
    indicate the mean of hip
abline(h = mean(fat$hip), col = 2, lty = 2)

# Add a red horizontal dashed line to
    indicate the mean of height
abline(v = mean(fat$height), col = 2, lty =
    2)

# Calculate the correlation coefficient
cor(fat$hip, fat$height)
```

The Spearman's Rank Correlation Coefficient

- Denote two univariate continuous data by

$$\boldsymbol{x} = [x_1, \ldots, x_i, \ldots, x_n], \text{ where } x_i \in \mathbb{R}$$
$$\boldsymbol{y} = [y_1, \ldots, y_i, \ldots, y_n], \text{ where } y_i \in \mathbb{R}$$

  and their rank (in ascending order) vectors are

$$\boldsymbol{r} = [r_1, \ldots, r_i, \ldots, r_n], \text{ where } 1 \leq r_i \leq n$$
$$\boldsymbol{s} = [s_1, \ldots, s_i, \ldots, s_n], \text{ where } 1 \leq s_i \leq n$$

- The Spearman's rank correlation coefficient is

$$\rho_{x,y}^{\text{spearman}} = \rho_{r,s} = \frac{\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2}}$$

The Spearman's Rank Correlation Coefficient

- The Spearman's rank correlation coefficient is

$$\rho_{x,y}^{\text{spearman}} = \rho_{r,s} = \frac{\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2}}$$

- Interpretation
  - If all ranks are distinct integers, then

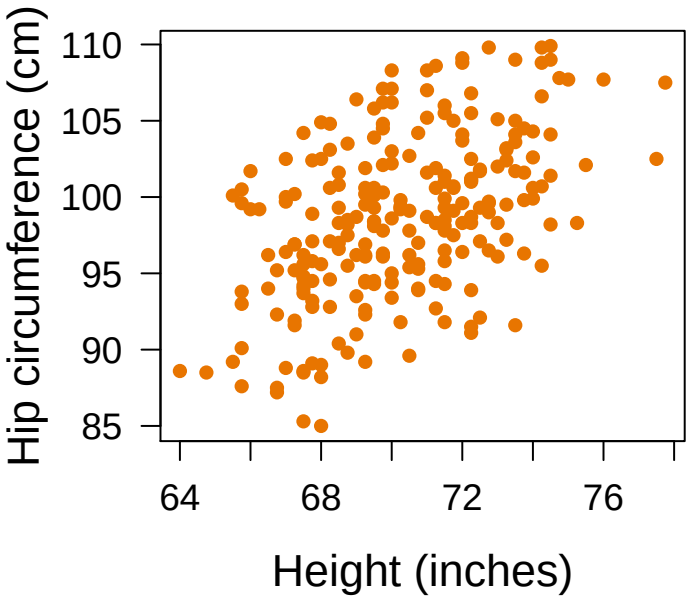
$$\rho_{x,y}^{\text{spearman}} = 1 - \frac{6\sum_{i=1}^{n}(r_i - s_i)^2}{n(n^2-1)}$$

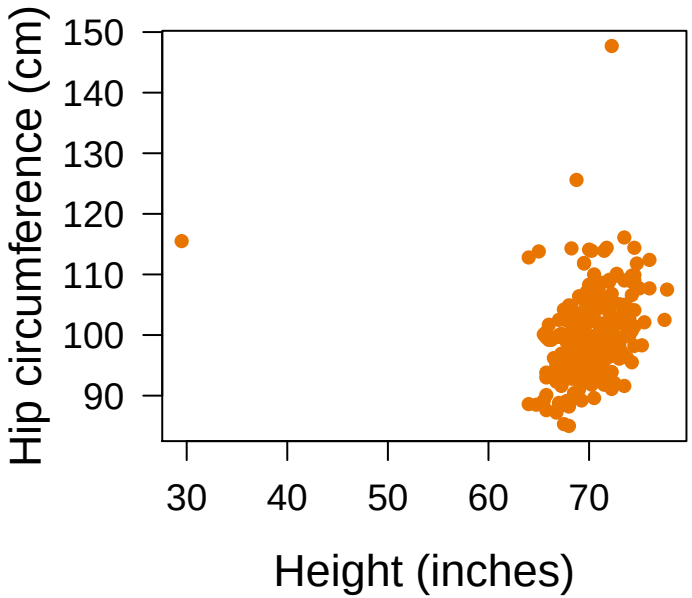
  - The sign indicates the direction of association (not linear correlation)
  - Resistant to the outliers

The Spearman's Rank Correlation Coefficient

- Implementation in R
  - $\boldsymbol{x}$ and $\boldsymbol{y}$ are two numeric vectors
    - `cor(`x`, `y`, method = "spearman")`
    - `cor(rank(`x`), rank(`y`), method = "pearson")`
  - $\boldsymbol{x}$ and $\boldsymbol{y}$ are two numeric variables in a data frame $\boldsymbol{D}$
    - `cor(`D`$x_name, `D`$y_name, method = "spearman")`
    - `cor(rank(`D`$x_name), rank(`D`$y_name), method = "pearson")`

The Spearman's Rank Correlation Coefficient





$\rho_{\text{Height,Hip}} = 0.50,\ \rho^{\text{Spearman}}_{\text{Height,Hip}} = 0.48$ $\rho_{\text{Height,Hip}} = 0.17,\ \rho^{\text{Spearman}}_{\text{Height,Hip}} = 0.42$

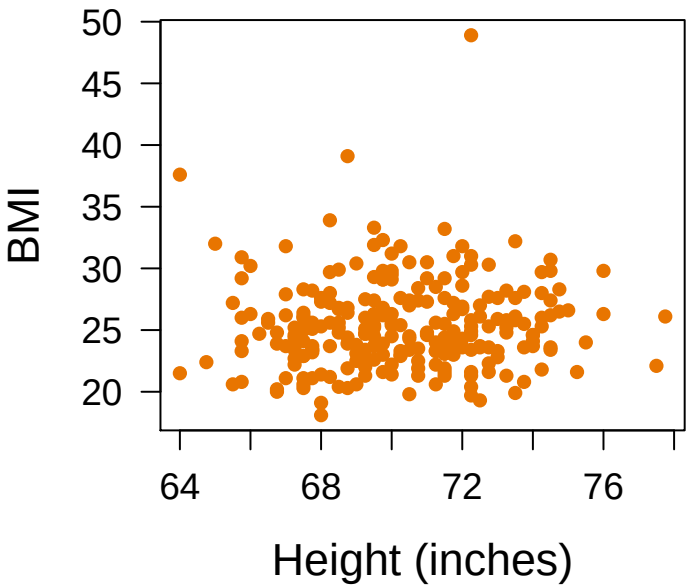
Discussion

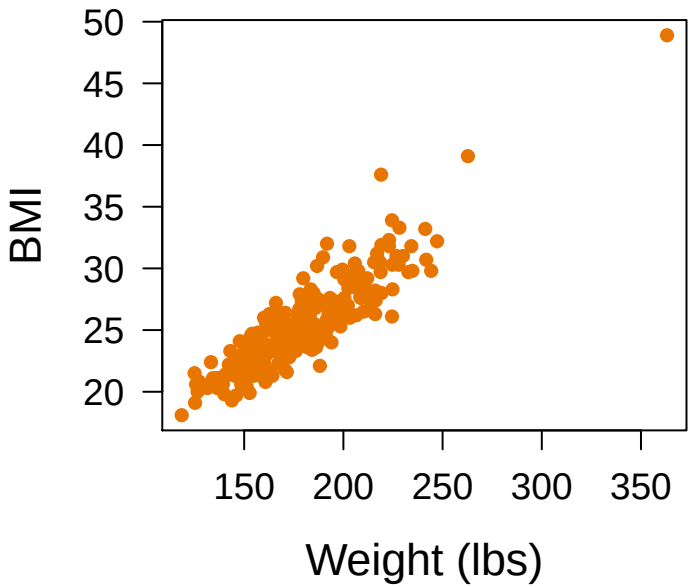
- Correlation is not casusation!
- Two variables may be influenced by a confounding variable

Simpson's Paradox

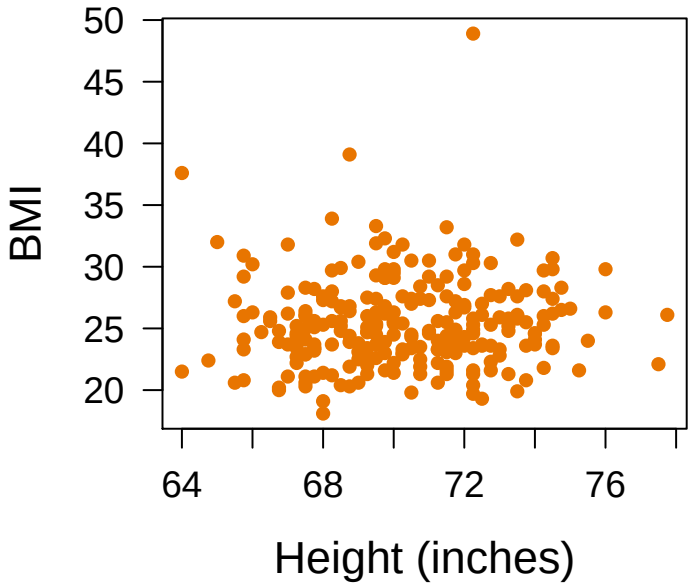
- Is BMI more linearly correlated with weight than height?

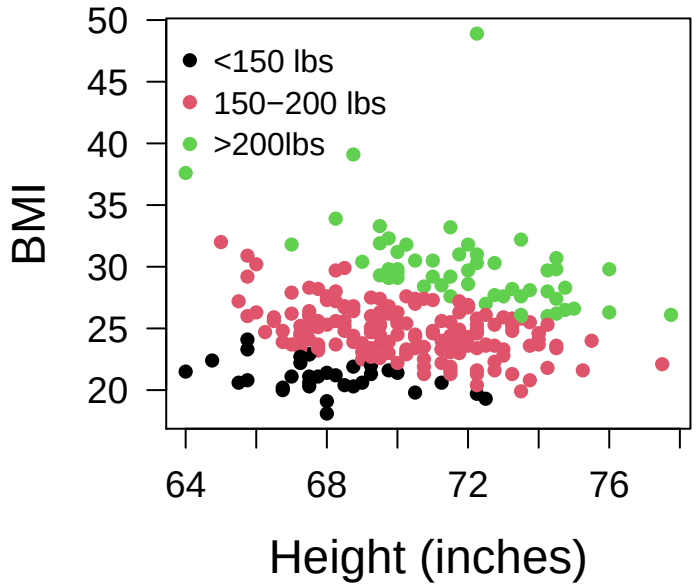
$$\text{BMI} = \frac{\text{weight in kg}}{\text{height in m}^2} = 703\frac{\text{weight in lb}}{\text{height in inch}^2}$$





Simpson's Paradox





$\rho_{\text{BMI,Ht}} = -0.02,\ \rho^{\text{Spearman}}_{\text{BMI,Ht}} = 0.07$ $\rho_{\text{BMI,Ht}} = -0.46,\ \rho^{\text{Spearman}}_{\text{BMI,Ht}} = -0.39$

- BMI $\propto$ weight/height
- A trend appears in several different groups of data but disappears or reverses when these groups are combined.

After-class Reading

- *Using R for Introductory Statistics (1st Ed.)* by John Verzani
- Chapter 3 Bivariate data
  - Section 3.3 Relationships in numeric data
    - Subsection 3.3.1 Using scatterplots to investigate relationships
    - Subsection 3.3.2 The correlation between two variables

After-class Reading

- *Using R for Introductory Statistics (2nd Ed.)* by John Verzani
- Chapter 3 Bivariate data
  - Section 3.3 Paired data