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

Quiz 5

- Load the dataset `diamond` in the library `UsingR`.
  - Problem 1: What is the Pearson correlation coefficient between diamond weight and price?
  - Problem 2: Do you suggest to use Spearman correlation coefficient here?

Quiz 5

- Solutions

```
# Load data
library(UsingR)
data("diamond")

# Problem 1
cor(diamond$price, diamond$carat)

# Problem 2
plot(price ~ carat, data = diamond)
# Stong linear correlation pattern, no need
   to compute the Spearman correlation
```

Learning Goals

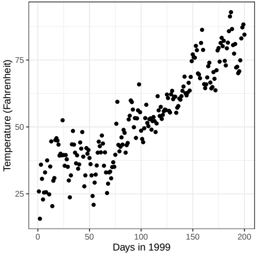
- Graphical summaries for two continuous data
  - Trend line/curve

Continuous Data

- Examples
  - The weight of a person in lbs
  - The height of a person in inches
  - The total STAT3355 score of a student
  - The average weekly learning hours of a student for STAT3355
- Are large/small values of one variable related to large/small values of the other variable?
- Estimation: What is the expected change in one variable for a one-unit change in the other variable?
- Prediction: For a (new) sample, could the prediction of one variable be made from the known value of the other variable?

Linear Model

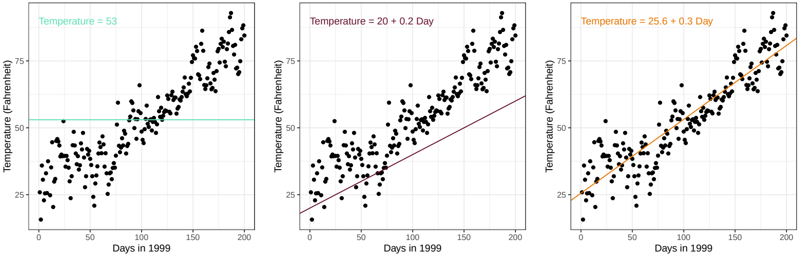
- Make inference and prediction via "curve fitting"
- Maximum daily temperatures in New York City in the first 200 days in 1999



- Questions
  - How to summarize the data using a few words?
  - How to summarize the data using a few numbers?
  - How to predict the temperature at day 201?

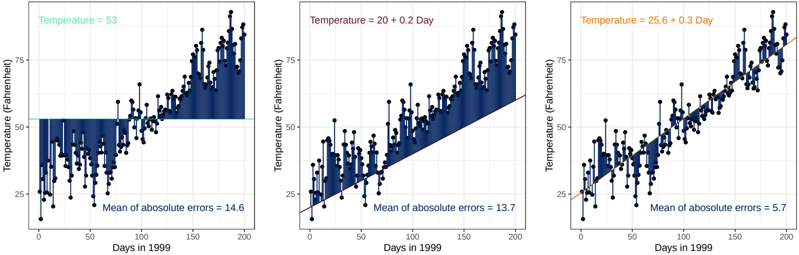
Linear Model

- Answers
  - Linear trend
  - The intercept and slope of the straight line: $y = a + bx$
  - Temperature at day 201 $= a + b \times 201$
- What are the best choices of a and b, given the data?



Linear Model

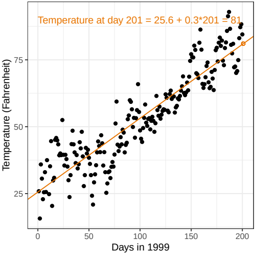
- Criterion
  - The minimum sum of residuals between points and the line

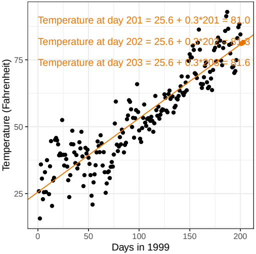


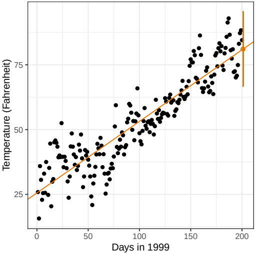
- Interpretation
  - The temperature of the first day is 25.6
  - The temperature increase 0.3 in average each day

Linear Model

- Prediction
  - Predict one day: The temperature at day 201 is 81
  - Predict multiple day: At day $202, 203, \dots$
  - Predict the range based on the sum of residuals: $[67, 96]$



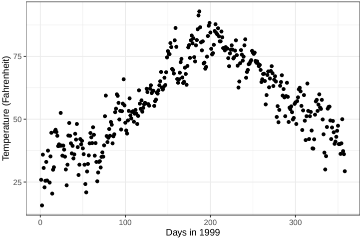


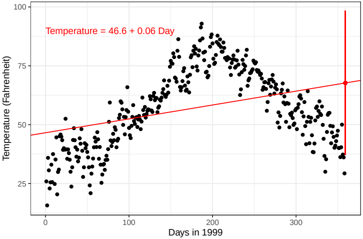


- Is it a good prediction model?
  - The temperature at Christmas is $25.6 + 0.3 \times 359 = 133$

Linear Model

- Need to collect more data to make reliable prediction

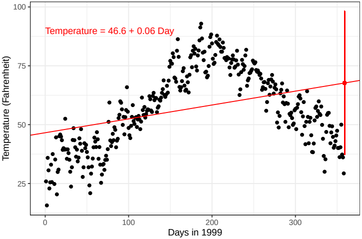


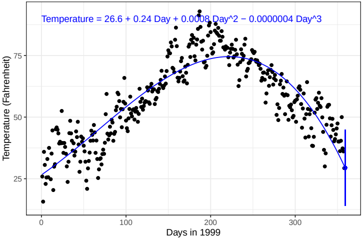


- Interpretation
  - The temperature of the first day is 46.6
  - The temperature increase 0.06 in average each day
- Prediction
  - Predict one day: The temperature at Christmas is 68
  - Predict the range based on the sum of residuals: $[37, 98]$

Polynomial Model

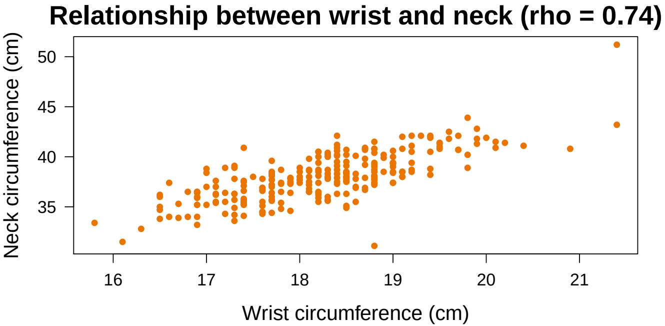
- Need to choose a better model: $y = a + bx + cx^2 + dx^3$





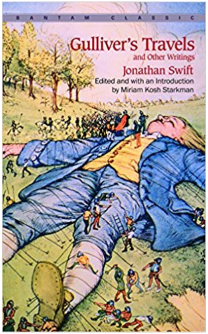
- Interpretation
  - The temperature of the first day is 26.6
  - No linear explanation
- Prediction
  - Predict one day: The temperature at Christmas is 29
  - Predict the range based on the sum of residuals: $[14, 45]$

Scatter Plot



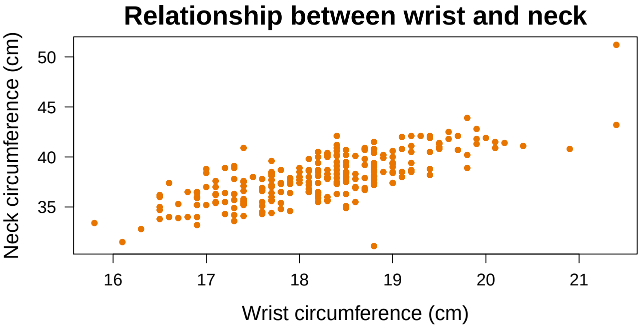
- Displays two continuous data in a Cartesian plane
  - x axis represents the value of a continuous data
  - y axis represents the value of the other continuous data
  - Each point corresponds to a sample

Gulliver's Travels



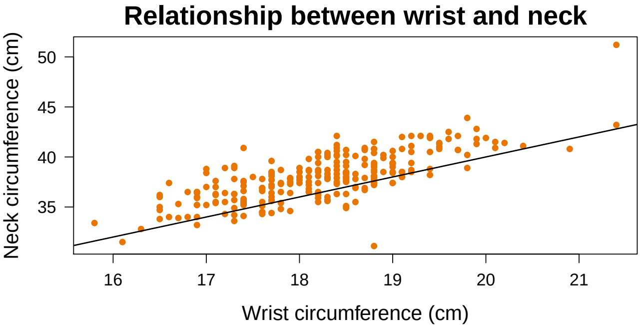
Then they measured my right thumb, and desired no more; for by a mathematical computation, that twice round the thumb is once round the wrist, and so on to the neck and the waist, and by the help of my old shirt, which I displayed on the ground before them for a pattern, they fitted me exactly.

Trend Line



- Use the function `abline(a = 0, b = 2)` to draw a line

Trend Line



- Use the function `abline(a = 0, b = 2)` to draw a line
- Is it good enough?

Trend Line

- Denote two univariate continuous data by

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

- Denote a line that summarizes their relationship by

$$y = a + bx$$

- For each $x = x_i$, the observed value of y is y_i, while the expected value is $a + bx_i$
- Define

$$\text{Residual} = \text{Observed} - \text{Expected} = y_i - (a + bx_i)$$

Least Squares Regression Line

- A line that minimizes the sum of squared residuals

$$\mathsf{SSR} = \mathsf{SSR}(a,b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

- Find out the optimum values of a and b that minimize SSR
- As $\mathsf{SSR}(a,b)$ is convex multivariate function, the optimum solution lies at gradient zero

$$\begin{cases} \frac{\partial}{\partial a}\mathsf{SSR}(a,b) = 0 \\ \frac{\partial}{\partial b}\mathsf{SSR}(a,b) = 0 \end{cases}$$

- Let's solve the simultaneous equations!

Least Squares Regression Line

- Simplify SSR

$$
\begin{aligned}
\text{SSR} &= \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \\
&= \sum_{i=1}^{n} \left[y_i^2 + (a + bx_i)^2 - 2y_i(a + bx_i)\right] \\
&= \sum_{i=1}^{n} \left(y_i^2 + a^2 + b^2x_i^2 + 2abx_i - 2ay_i - 2bx_iy_i\right) \\
&= \sum_{i=1}^{n} y_i^2 + na^2 + b^2\sum_{i=1}^{n} x_i^2 + 2ab\sum_{i=1}^{n} x_i - 2a\sum_{i=1}^{n} y_i - 2b\sum_{i=1}^{n} x_iy_i \\
&= \sum_{i=1}^{n} y_i^2 + na^2 + b^2\sum_{i=1}^{n} x_i^2 + 2n\bar{x}ab - 2n\bar{y}a - 2b\sum_{i=1}^{n} x_iy_i
\end{aligned}
$$

Least Squares Regression Line

- Derive $\frac{\partial}{\partial a}\mathsf{SSR}$

$$\begin{aligned}\frac{\partial}{\partial a}\mathsf{SSR}(a,b) &= \frac{\partial}{\partial a}\left(\sum_{i=1}^{n} y_i^2 + na^2 + b^2\sum_{i=1}^{n} x_i^2 + 2n\bar{x}ab - 2n\bar{y}a - 2b\sum_{i=1}^{n} x_i y_i\right)\\ &= 2na + 2n\bar{x}b - 2n\bar{y}\end{aligned}$$

- Derive $\frac{\partial}{\partial b}\mathsf{SSR}$

$$\begin{aligned}\frac{\partial}{\partial b}\mathsf{SSR}(a,b) &= \frac{\partial}{\partial b}\left(\sum_{i=1}^{n} y_i^2 + na^2 + b^2\sum_{i=1}^{n} x_i^2 + 2n\bar{x}ab - 2n\bar{y}a - 2b\sum_{i=1}^{n} x_i y_i\right)\\ &= 2b\sum_{i=1}^{n} x_i^2 + 2n\bar{x}a - 2\sum_{i=1}^{n} x_i y_i\end{aligned}$$

Least Squares Regression Line

- Solve

$$\begin{cases} \frac{\partial}{\partial a}\mathsf{SSR}(a,b) = 2na + 2n\bar{x}b - 2n\bar{y} = 0 \\ \frac{\partial}{\partial b}\mathsf{SSR}(a,b) = 2b\sum_{i=1}^{n} x_i^2 + 2n\bar{x}a - 2\sum_{i=1}^{n} x_i y_i = 0 \end{cases}$$

$$\begin{cases} a + \bar{x}b - \bar{y} = 0 \\ b\sum_{i=1}^{n} x_i^2 + n\bar{x}a - \sum_{i=1}^{n} x_i y_i = 0 \end{cases}$$

- Obtain

$$\begin{cases} b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ a = \bar{y} - b\bar{x} \end{cases}$$

Least Squares Regression Line

- The sample covariance

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

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ are the sample means

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n} (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y}) \\
&= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y} - \sum_{i=1}^{n} \bar{x} y_i + \sum_{i=1}^{n} \bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i - \bar{x} \sum_{i=1}^{n} y_i + n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y} n \bar{x} - \bar{x} n \bar{y} + n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}
\end{aligned}$$

Least Squares Regression Line

- $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$
- $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$
- The mathematical solution:

$$\begin{cases} \frac{\partial \mathsf{SSR}(a,b)}{\partial a} = 0 \\ \frac{\partial \mathsf{SSR}(a,b)}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ a = \bar{y} - b\bar{x} \end{cases}$$

- The slope

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Least Squares Regression Line

- The slope

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

  - The Pearson correlation coefficient $\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$
  - The sample variances $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ and $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

- The slope is the correlation scaled to fit the scale of the problem
- $\boldsymbol{x}$ and $\boldsymbol{y}$ is not interchangeable

Least Squares Regression Line

- Least Squares Regression Line

$$y = a + bx$$

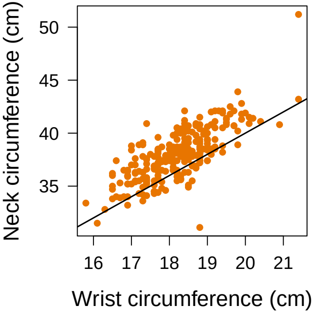
- Interpretations
  - Intercept a: The expected value of y if the value of x is zero, or the baseline response
  - Slope b: The expected change in y for a one unit change in x, or the effect of x on y

Least Squares Regression Line

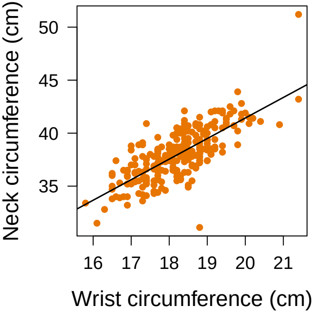
- Implementation in R
  - $\boldsymbol{x}$ and $\boldsymbol{y}$ are two numeric vectors
    - `m <- lm(`$y \sim x$`)`
  - $\boldsymbol{x}$ and $\boldsymbol{y}$ are two numeric variables in a data frame $\boldsymbol{D}$
    - `m <- lm(y_name ~ x_name, data = `D`, subset = )`
  - Obtain the intercept and slope:
    - `coef(m)`
    - `summary(m)`
  - Add a trend line to the existing plot:
    - `abline(a = coef(m)[1], b = coef(m)[2])`
    - `abline(coef(m))`
    - `abline(m)`

Least Squares Regression Line

- Which one is better?



$y = 2x$



$y = 2.64 + 1.94x$

Your Turn

- Use the dataset `fat` in the package `UsingR` to make a plot
  - Draw a scatter plot, where x axis represents the wrist circumference (cm) and y axis represents the neck 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 vertical dashed line to indicate the mean of wrist and add a red horizontal dashed line to indicate the mean of neck
  - Add a blue solid line to represent the least squares regression line

Your Turn

- Solutions

```
library(UsingR)

# Load data
data("fat")

# Draw a scatter plot for neck against wrist
plot(neck ~ wrist, data = fat, pch = 16,
    xlab = "Wrist circumference (cm)", ylab =
     "Neck circumference (cm)", main = "
    Relationship between neck and wrist")
```

Your Turn

- Solutions

```
# Add lines to indicate their means
abline(h = mean(fat$neck), col = "red", lty
   = 2)
abline(v = mean(fat$wrist), col = "red", lty
    = 2)

# Fit a regression line and plot it
m <- lm(neck ~ wrist, data = fat)
abline(a = coef(m)[1], b = coef(m)[2], col =
    "blue", lwd = 1.5)
```

After-class Reading

- *Using R for Introductory Statistics (1st Ed.)* by John Verzani
- Chapter 3 Bivariate data
  - Section 3.4 Simple linear regression
    - Subsection 3.4.2 Finding the regression coefficients using `lm()`
    - Subsection 3.4.4 Interacting with a scatterplot
    - Subsection 3.4.7 Trend lines

After-class Reading

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