Lecture 24

Correlation and Regression Analysis

Text: Chapter 11

STAT 8010 Statistical Methods I
November 19, 2020

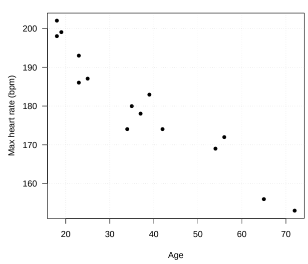
Whitney Huang
Clemson University

Notes

Motivated Example: Maximum Heart Rate vs. Age

Suppose we have 15 people of varying ages are tested for their maximum heart rate (bpm):

Age	18	23	25	35	65	54	34	56	72	19	23	42	18	39	37
MaxHeartRate	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178

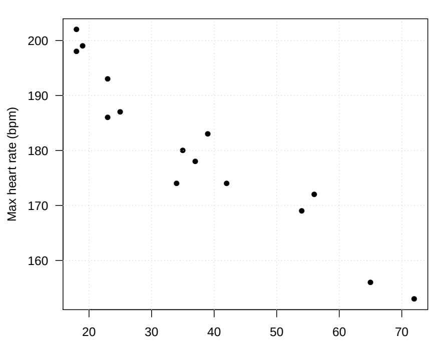


Question: How to describe the relationship between maximum heart rate and age?

Notes

Scatterplot

A scatterplot is a useful tool to graphically display the relationship between two numerical variables. Each dot on the scatterplot represents one observation from the data



Notes

Scatterplot Cont'd

Typical questions we want to ask for a scatterplot:

- the form of relationship between two variables e.g. linear, quadratic, $\cdots$
- the strength of the relationship between two variables e.g. weak, moderate, strong
- the direction of the relationship between two variables e.g positive, negative

In the next few slides we will learn how to quantify the strength and direction of the linear relationship between two variables

Notes

Variance, Covariance, and Correlation

- **Recall:** Variance is a measure of the variability of one quantitative variable
- Covariance is a measure of how much two quantitative random variables change together
- The sign of the covariance shows the direction in the linear relationship between the variables
- The normalized version of the covariance, the correlation shows both the the direction and the strength of the linear relation

Notes

Correlation: Pearson Correlation Coefficient (ρ)

- We use ρ to denote the population correlation and r to denote the sample correlation
- The value of the correlation is between -1 and 1
- The strength of the linear relation:
  - If $\rho = 1$ (-1): the two variables have a perfect positive (negative) linear relationship
  - If $0.7 < |\rho| < 1$: we say the two variables have a strong linear relationship
  - If $0.3 < |\rho| < 0.7$: we say the two variables have a moderate linear relationship
  - If $0 < |\rho| < 0.3$: we say the two variables have a weak linear relationship
  - If $\rho = 0$: we say the two variables have no linear relationship

Notes

Scatterplot & Pearson Correlation Coefficient

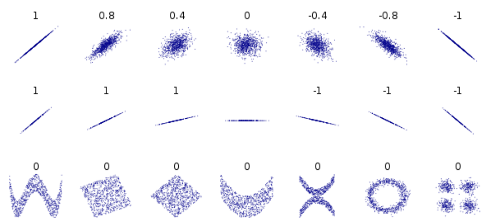


Figure: Image courtesy of Wikipedia at https://en.wikipedia.org/wiki/Correlation_and_dependence

Notes

Covariance and Correlation

- **Recall:** Variance
  - Sample variance: $s_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$
  - Population variance: $\sigma_X^2 = \mathrm{E}[(X - \mu_X)^2]$
- Covariance
  - Sample covariance: $s_{X,Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$
  - Population covariance: $\sigma_{X,Y} = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)]$
- Correlation
  - Sample correlation: $r_{X,Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$ or $\frac{s_{X,Y}}{s_X s_Y}$
  - Population correlation: $\rho_{X,Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\mathrm{E}[(X-\mu_X)^2]}\sqrt{\mathrm{E}[(Y-\mu_Y)^2]}}$ or $\frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$

Notes

A Toy Example

You wonder how sleep affects productivity. You take a sample of 4 of your friends and measure last night's sleep and today's productivity in hours. Here are the results:

Sleep (X)	Productivity (Y)
2	4
4	12
6	14
10	10

Calculate the means, variances, and standard deviations of each variable and the correlation coefficient of these two variables

Notes

Toy Example Cont'd

Solution.

Let X denote last night's sleep in hours and Y denote today's productivity in hours

- $\bar{X} = \frac{2+4+6+10}{4} = 5.5, \qquad \bar{Y} = \frac{4+12+14+10}{4} = 10$
- $s_X^2 = \frac{(2-5.5)^2+(4-5.5)^2+(6-5.5)^2+(10-5.5)^2}{4-1} = \frac{35}{3}$
  $s_Y^2 = \frac{(4-10)^2+(12-10)^2+(14-10)^2+(10-10)^2}{4-1} = \frac{56}{3}$
- $s_X = \sqrt{s_X^2} = \sqrt{\frac{35}{3}}, \qquad s_Y = \sqrt{s_Y^2} = \sqrt{\frac{56}{3}}$
- $r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y}$
  $s_{X,Y} = \frac{(2-5.5)(4-10)+(4-5.5)(12-10)+(6-5.5)(14-10)+(10-5.5)(10-10)}{3}$
  $= \frac{20}{3} \Rightarrow r_{X,Y} = \frac{\frac{20}{3}}{\sqrt{\frac{35}{3}}\sqrt{\frac{56}{3}}} = \frac{20}{\sqrt{35 \times 56}} = 0.4518$

Notes

Inference/Hypothesis Test on ρ

1. $H_0 : \rho = 0$ vs. $H_a : \rho \neq 0$
2. Test statistic: $t^* = r\sqrt{\frac{n-2}{1-r^2}}$
3. Under H_0: $t^* \sim t_{df=n-2}$

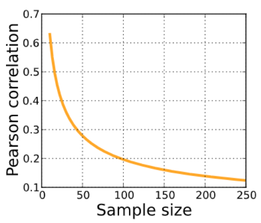
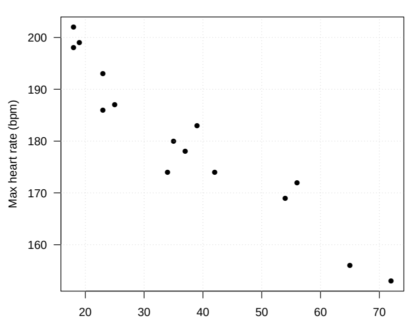


Figure: Image courtesy of Wikipedia

Notes

Maximum Heart Rate Example Revisited

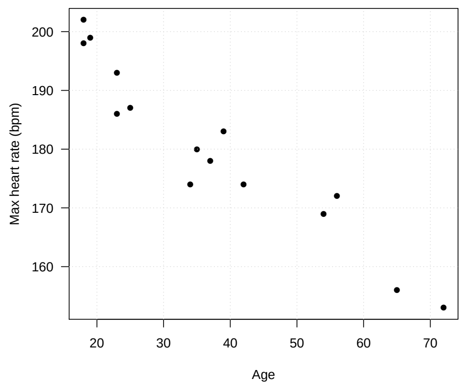


We may want to predict maximum heart rate for an individual based on his/her age $\Rightarrow$ Regression Analysis

Notes

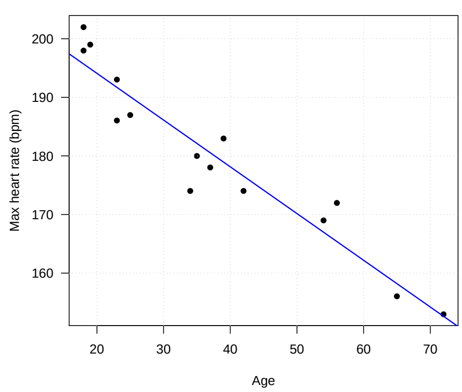
What is Regression Analysis?

Regression analysis: A set of statistical procedures for estimating the relationship between response variable and predictor variable(s)



Notes

Scatterplot: Is Linear Trend Reasonable?



Notes

Simple Linear Regression (SLR)

Y: dependent (response) variable; X: independent (predictor) variable

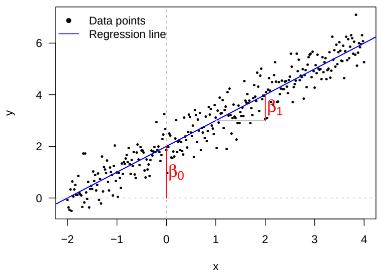
- In SLR we **assume** there is a **linear relationship** between X and Y:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- We will need to estimate β_0 (intercept) and β_1 (slope)
- Then we can use the estimated regression equation to
  - make predictions
  - study the relationship between response and predictor
  - control the response

Notes

Regression equation: $Y = \beta_0 + \beta_1 X$



- β_0: $\mathrm{E}[Y]$ when $X = 0$
- β_1: $\mathrm{E}[\Delta Y]$ when X increases by 1

Notes

Assumptions about the Random Error ε

In order to estimate β_0 and β_1, we make the following assumptions about ε

- $\mathrm{E}[\varepsilon_i] = 0$
- $\mathrm{Var}[\varepsilon_i] = \sigma^2$
- $\mathrm{Cov}[\varepsilon_i, \varepsilon_j] = 0, \quad i \neq j$

Therefore, we have

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 X_i, \text{ and}$$
$$\mathrm{Var}[Y_i] = \sigma^2$$

> The regression line $\beta_0 + \beta_1 X$ represents the **conditional expectation curve** whereas σ^2 measures the magnitude of the **variation** around the regression curve

Notes

Estimation: Method of Least Square

For the given observations $(x_i, y_i)_{i=1}^n$, choose β_0 and β_1 to minimize the *sum of squared errors*:

$$\mathrm{L}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Solving the above minimization problem requires some knowledge from Calculus....

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

We also need to **estimate** σ^2

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}, \text{ where } \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Notes

Properties of Least Squares Estimates

- **Gauss-Markov** theorem states that in a linear regression these least squares estimators
  1. **Are unbiased**, i.e.,
     - $\mathrm{E}[\hat{\beta}_1] = \beta_1$; $\mathrm{E}[\hat{\beta}_0] = \beta_0$
     - $\mathrm{E}[\hat{\sigma}^2] = \sigma^2$
  2. Have **minimum variance** among all unbiased linear estimators

Note that we do not make any distributional assumption on ε_i

Notes

Example: Maximum Heart Rate vs. Age

The maximum heart rate `MaxHeartRate` of a person is often said to be related to age `Age` by the equation:

$$\text{MaxHeartRate} = 220 - \text{Age}.$$

Suppose we have 15 people of varying ages are tested for their maximum heart rate (bpm) (link to the "dataset": `http://whitneyhuang83.github.io/maxHeartRate.csv`)

1. Compute the estimates for the regression coefficients
2. Compute the fitted values
3. Compute the estimate for σ

Notes

Estimate the Parameters β_1, β_0, and σ^2

Y_i and X_i are the Maximum Heart Rate and Age of the i[th] individual

- To obtain $\hat{\beta}_1$
  1. Compute $\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$, $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$
  2. Compute $Y_i - \bar{Y}$, $X_i - \bar{X}$, and $(X_i - \bar{X})^2$ for each observation
  3. Compute $\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})$ divived by $\sum_i^n (X_i - \bar{X})^2$
- $\hat{\beta}_0$: Compute $\bar{Y} - \hat{\beta}_1 \bar{X}$
- $\hat{\sigma}^2$
  1. Compute the fitted values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \cdots, n$
  2. Compute the **residuals** $e_i = Y_i - \hat{Y}_i, \quad i = 1, \cdots, n$
  3. Compute the **residual sum of squares (RSS)** $= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ and divided by $n-2$ (why?)

Notes