Lecture 2

Simple Linear Regression I

Reading: Chapter 11

STAT 8020 Statistical Methods II
August 23, 2019

Whitney Huang
Clemson University

Agenda

Announcements

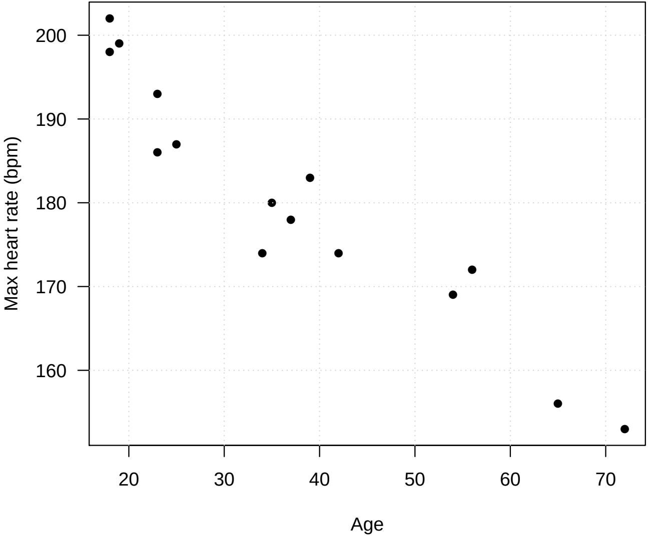
- Syllabus and lecture notes are in CANVAS and my personal website (link: https://whitneyhuang83.github.io/stat8020_2019Fall.html)

- **Academic Continuity Statement** is added in the updated syllabus (link: https://whitneyhuang83.github.io/STAT8010_Syllabus_2019_Fall.pdf)

- Please talk to me if you would like to share your data set to be used for this class

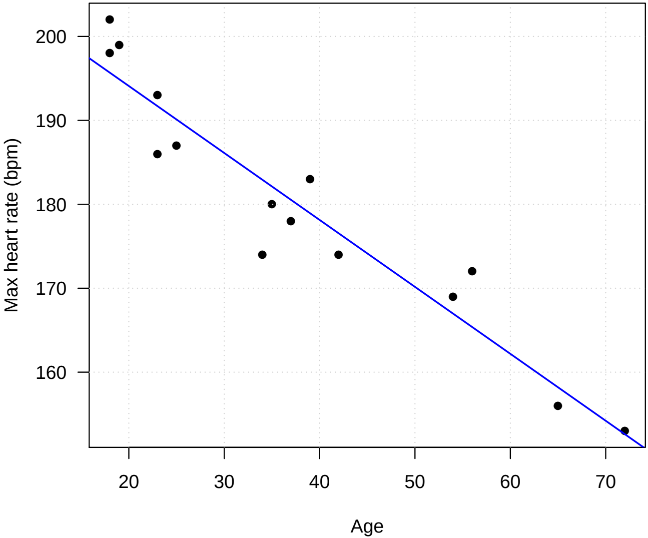
What is Regression Analysis?

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



Simple linear regression

Scatterplot: Is Linear Trend Reasonable?



Simple Linear Regression (SLR)

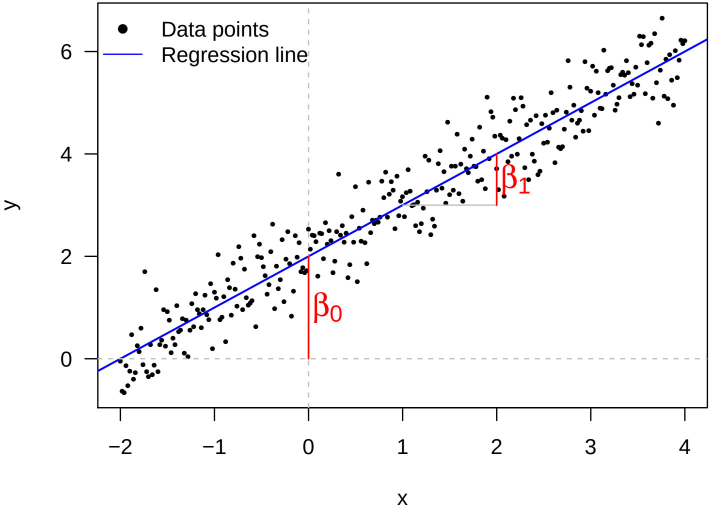
Y: dependent (response) variable; X: independent (predictor) variable

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

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- 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
- Yet we need to quantify our uncertainty regarding the linear relationship

Regression equation: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$



Assumptions about 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

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*:

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

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

- $\hat{\beta}_{1,\mathrm{LS}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2}$
- $\hat{\beta}_{0,\mathrm{LS}}$ = $\bar{Y} - \hat{\beta}_{1,\mathrm{LS}}\bar{X}$

We also need to **estimate** σ^2

- $\hat{\sigma}^2_{\mathrm{LS}} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$, where $\hat{Y}_i = \hat{\beta}_{0,\mathrm{LS}} + \hat{\beta}_{1,\mathrm{LS}} X_i$

Properties of Least Squares Estimates

- **Gauss-Markov** theorem states that in a linear regression these least squares estimators

  1. **Are unbiased**, i.e.,
     - $E[\hat{\beta}_{1,\mathsf{LS}}] = \beta_1$; $E[\hat{\beta}_{0,\mathsf{LS}}] = \beta_0$
     - $E[\hat{\sigma}^2_{\mathsf{LS}}] = \sigma^2$
  2. Have **minimum variance** among all unbiased linear estimators

Note that we do not make any distributional assumption on ε_i

Example: Maximum Heart Rate vs. Age

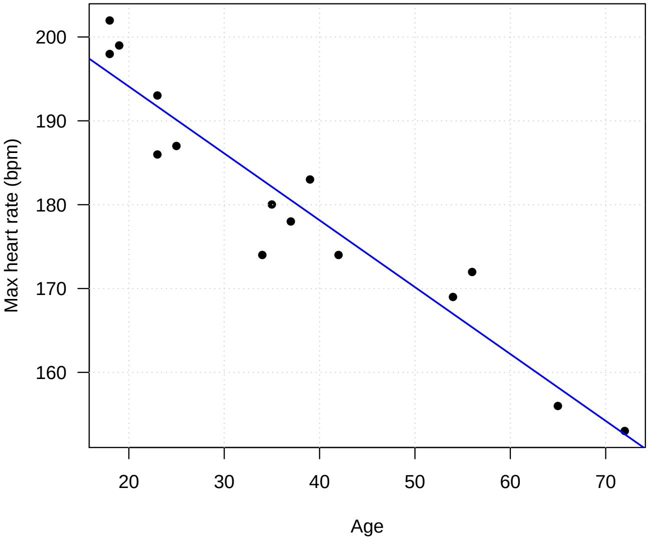
The maximum heart rate `MaxHeartRate` of a person is often said to be related to age `Age` by the equation:

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

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

Linear Regression Fit



Question: Is linear relationship between max heart rate and age reasonable? ⇒ Residual Analysis

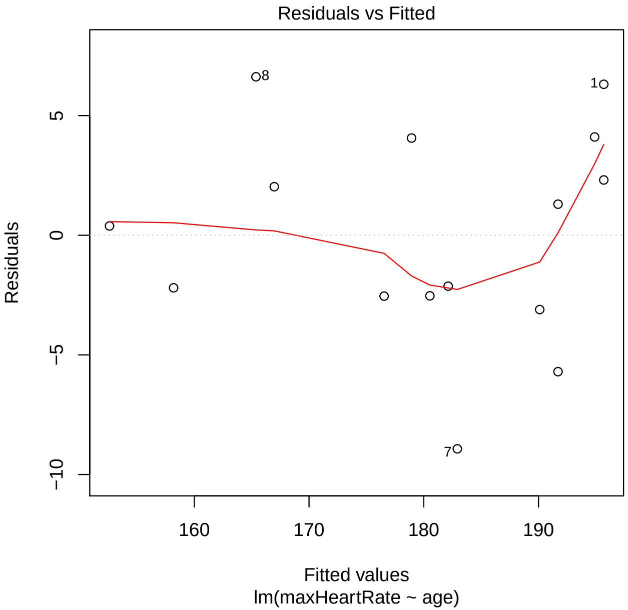
Residuals

- The residuals are the differences between the observed and fitted values:
$$e_i = Y_i - \hat{Y}_i,$$
where $\hat{Y}_i = \hat{\beta}_{0,\text{LS}} + \hat{\beta}_{1,\text{LS}} X_i$

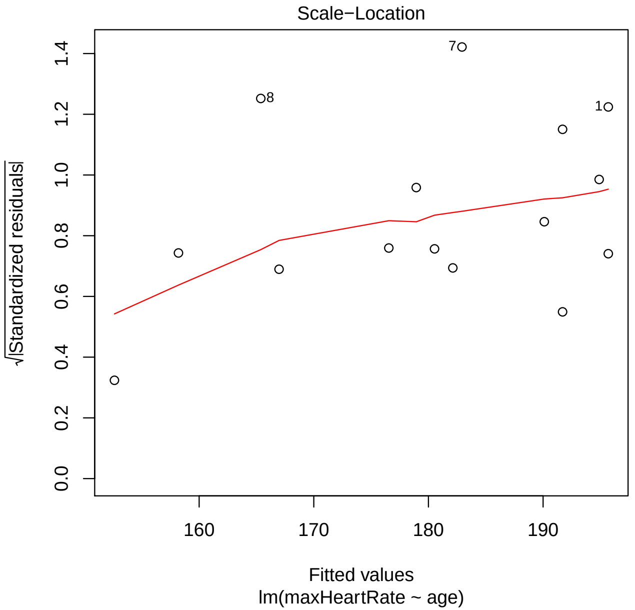
- e_i is NOT the error term $\varepsilon_i = Y_i - \mathrm{E}[Y_i]$

- Residuals are very useful in assessing the appropriateness of the assumptions on ε_i. Recall
  - $\mathrm{E}[\varepsilon_i] = 0$
  - $\mathrm{Var}[\varepsilon_i] = \sigma^2$
  - $\mathrm{Cov}[\varepsilon_i, \varepsilon_j] = 0, \quad i \neq j$

Residual Analysis



Residual Analysis



Summary

In this lecture, we learned

- Simple Linear Regression: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- Method of Least Square for parameter estimation
- Residual analysis to check model assumptions

Next time we will talk about

1. More on residual analysis
2. Normal Error Regression Model and statistical inference for β_0, β_1, and σ^2
3. Prediction