

# Lecture 36

## Simple Linear Regression

STAT 8010 Statistical Methods I  
November 20, 2019

Whitney Huang  
Clemson University

Simple Linear Regression

CLEMSON UNIVERSITY

What is regression analysis

Simple Linear Regression (SLR)

Parameter Estimation in SLR

36.1

Notes

---

---

---

---

---

---

---

### Agenda

- 1 What is regression analysis
- 2 Simple Linear Regression (SLR)
- 3 Parameter Estimation in SLR

Simple Linear Regression

CLEMSON UNIVERSITY

What is regression analysis

Simple Linear Regression (SLR)

Parameter Estimation in SLR

36.2

Notes

---

---

---

---

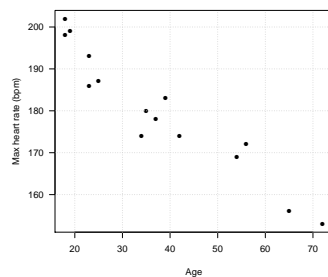
---

---

---

### What is Regression Analysis?

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



We will focus on **simple linear regression** in the next few lectures

Simple Linear Regression

CLEMSON UNIVERSITY

What is regression analysis

Simple Linear Regression (SLR)

Parameter Estimation in SLR

36.3

Notes

---

---

---

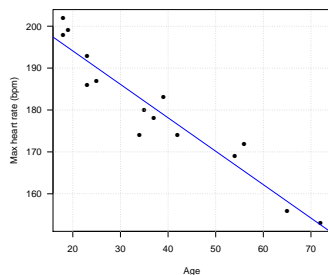
---

---

---

---

## Scatterplot: Is Linear Trend Reasonable?



The relationship appears to be linear. What about the **strength** of this linear relationship?

```
> cov(age, maxHeartRate)  > cor(age, maxHeartRate)
[1] -243.9524               [1] -0.9534656
```

Simple Linear Regression

CLEMSON UNIVERSITY

What is regression analysis

Simple Linear Regression (SLR)

Parameter Estimation in SLR

36.4

## 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  $\beta_0$  (intercept) and  $\beta_1$  (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 **estimation uncertainty** regarding the linear relationship (will talk about this next time)

Simple Linear Regression

CLEMSON UNIVERSITY

What is regression analysis

Simple Linear Regression (SLR)

Parameter Estimation in SLR

36.5

## Notes

---

---

---

---

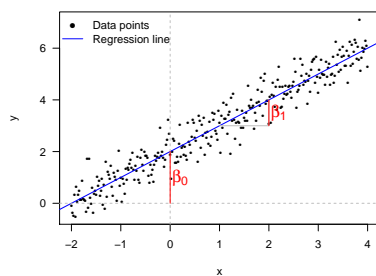
---

---

---

---

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



- $\beta_0$ :  $E[Y]$  when  $X = 0$
- $\beta_1$ :  $E[\Delta Y]$  when  $X$  increases by 1

Simple Linear Regression

CLEMSON UNIVERSITY

What is regression analysis

Simple Linear Regression (SLR)

Parameter Estimation in SLR

36.6

## Notes

---

---

---

---

---

---

---

---

## Assumptions about the Random Error $\varepsilon$

In order to estimate  $\beta_0$  and  $\beta_1$ , we make the following assumptions about  $\varepsilon$

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

Therefore, we have

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

The regression line  $\beta_0 + \beta_1 X$  represents the **conditional expectation curve** whereas  $\sigma^2$  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  $\beta_0$  and  $\beta_1$  to minimize the *sum of squared errors*:

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

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**  $\sigma^2$

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



## Notes

---

---

---

---

---

---

---

---

## Properties of Least Squares Estimates

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

1. **Are unbiased**, i.e.,

- $E[\hat{\beta}_1] = \beta_1; E[\hat{\beta}_0] = \beta_0$
- $E[\hat{\sigma}^2] = \sigma^2$

2. Have **minimum variance** among all unbiased linear estimators

Note that we do not make any distributional assumption on  $\varepsilon_i$



## Notes

---

---

---

---

---

---

---

---

Example: Maximum Heart Rate vs. Age

The maximum heart rate  $\text{MaxHeartRate}$  of a person is often said to be related to age  $\text{Age}$  by the equation:

MaxHeartRate = 220 – 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  $\sigma$

Simple Linear Regression

CLEMSON UNIVERSITY

What is regression analysis

Simple Linear Regression (SLR)

Parameter Estimation in SLR

36.10

Notes

Estimate the Parameters  $\beta_1$ ,  $\beta_0$ , and  $\sigma^2$

$Y_i$  and  $X_i$  are the Maximum Heart Rate and Age of the  $i^{\text{th}}$  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  $\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$  divided by  $\sum_{i=1}^n (X_i - \bar{X})^2$
- $\hat{\beta}_0$ : Compute  $\bar{Y} - \hat{\beta}_1 \bar{X}$
- $\hat{\sigma}^2$ 
  - 1 Compute the fitted values:  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ ,  $i = 1, \dots, n$
  - 2 Compute the **residuals**  $e_i = Y_i - \hat{Y}_i$ ,  $i = 1, \dots, n$
  - 3 Compute the **residual sum of squares (RSS)** =  $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$  and divided by  $n - 2$  (why?)

Simple Linear Regression

CLEMSON UNIVERSITY

What is regression analysis

Simple Linear Regression (SLR)

Parameter Estimation in SLR

36.11

Notes

Let’s Do the Calculations

$$\bar{X} = \sum_{i=1}^{15} \frac{18 + 23 + \dots + 39 + 37}{15} = 37.33$$
$$\bar{Y} = \sum_{i=1}^{15} \frac{202 + 186 + \dots + 183 + 178}{15} = 180.27$$

X	18	23	25	35	65	54	34	66	72	19	23	42	18	39	37
Y	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178
	-19.33	-14.33	-12.33	-2.33	-27.67	-16.67	-3.33	-16.67	-34.67	-16.33	-14.33	-4.67	-19.33	-1.67	-9.33
	21.73	6.73	6.73	-0.27	-24.27	-11.27	-6.27	-6.27	-27.27	18.73	12.73	-6.27	17.73	2.73	-2.27
	-420.18	-42.18	-83.04	0.82	-671.38	-167.78	20.89	-154.31	-945.24	-343.44	-182.51	-29.24	-342.84	4.56	0.76
	373.78	205.44	152.11	5.44	765.44	277.78	11.11	348.44	1201.78	336.11	295.44	21.78	373.78	5.78	0.11
	195.69	191.70	190.11	182.13	158.20	166.97	182.93	165.38	152.61	194.89	191.70	176.54	195.69	178.94	180.53

- $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = -0.7977$
- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 210.0485$
- $\hat{\sigma}^2 = \frac{\sum_{i=1}^{15} (Y_i - \hat{Y}_i)^2}{13} = 20.9563 \Rightarrow \hat{\sigma} = 4.5778$

Simple Linear Regression

CLEMSON UNIVERSITY

What is regression analysis



Simple Linear Regression (SLR)

Parameter Estimation in SLR

36.12

Notes

Let's Double Check

Output from  (  Studio)

```
> fit <- lm(MaxHeartRate ~ Age)
> summary(fit)


Call:
lm(formula = MaxHeartRate ~ Age)

Residuals:
    Min       1Q   Median       3Q      Max
-8.9258 -2.5383  0.3879  3.1867  6.6242

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 210.04846   2.86694   73.27  < 2e-16 ***
Age         -0.79773    0.06996  -11.40 3.85e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.578 on 13 degrees of freedom
Multiple R-squared:  0.9091,    Adjusted R-squared:  0.9021
F-statistic: 130 on 1 and 13 DF,  p-value: 3.848e-08
```

Simple Linear Regression



What is regression analysis

Simple Linear Regression (SLR)

Parameter Estimation in SLR

36.13

Notes

---

---

---

---

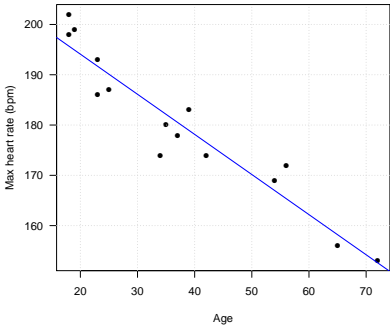
---

---

---


---

Linear Regression Fit



**Question:** Is linear relationship between max heart rate and age reasonable? ⇒ [Residual Analysis](#)

Simple Linear Regression



What is regression analysis

Simple Linear Regression (SLR)

Parameter Estimation in SLR

36.14

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---