

Lecture 1

Introduction

STAT 8020 Statistical Methods II
August 20, 2020

Whitney Huang
Clemson University



Who is the instructor?
Class Policies / Schedule
Tell us about yourself
Simple Linear Regression
SLR Parameter Estimation
Residual Analysis

1.1

Notes

Who is the instructor?



Who is the instructor?
Class Policies / Schedule
Tell us about yourself
Simple Linear Regression
SLR Parameter Estimation
Residual Analysis

1.2

Notes

Who am I?

- **Second year** Assistant Professor of Applied Statistics and Data Science
- Born in Laramie, Wyoming, grew up in Taiwan



- With a B.S. in Mechanical Engineering, switched to Statistics in graduate school
- Got a Ph.D. (Statistics) in 2017 at Purdue University.



Who is the instructor?
Class Policies / Schedule
Tell us about yourself
Simple Linear Regression
SLR Parameter Estimation
Residual Analysis

1.3

Notes

How to reach me?

- **Email:** wkhuang@clermson.edu
- **Office:** O-221 Martin Hall
- **Office Hours:** TR 11:00am – 12:00pm and by appointment

CLEMSON
UNIVERSITY

Who is the instructor?

Class Policies / Schedule

Tell us about yourself

Simple Linear Regression

SLR Parameter Estimation

Residual Analysis

1.4

Notes

Class Policies / Schedule

CLEMSON
UNIVERSITY

Who is the instructor?

Class Policies / Schedule

Tell us about yourself

Simple Linear Regression

SLR Parameter Estimation

Residual Analysis

1.5

Notes

Logistics

- We will meet TR 12:30pm – 1:45pm via Zoom
- There will be [three online exams](#) and a (comprehensive) online final. The (tentative) dates for the three exams are:
 - **Exam I:** Sept. 24, Thursday
 - **Exam II:** Oct. 20, Tuesday
 - **Exam II:** Nov. 12, Tuesday
 - The **Final Exam** will be given on Wednesday, Dec. 7, 3:00 pm -5:30 pm.
- No classes on [Nov. 3 \(Fall Break\)](#) & [26 \(Thanksgiving\)](#)

CLEMSON
UNIVERSITY

Who is the instructor?

Class Policies / Schedule

Tell us about yourself

Simple Linear Regression

SLR Parameter Estimation

Residual Analysis

1.6

Notes

Class Website

CANVAS and my teaching website (link:
https://whitneyhuang83.github.io/STAT8020/Fall2020/stat8020_2020Fall1.html)

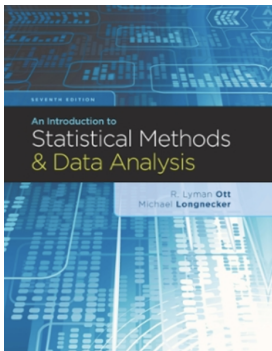
- Course syllabus [\[Link\]](#) / Announcements
- Lecture slides/notes
- Exam schedule
- Data sets
- R code



Notes

Recommended Textbook

An Introduction to Statistical Methods and Data Analysis, 6th Edition. Lyman Ott and Micheal T. Longnecker, Duxbury, 2010; ISBN-13: 978-1305269477



Notes

Evaluation

● Grade Distribution:	Exam I:	25%
	Exam II	25%
	Exam III	25%
	Final Exam	25%

● Letter Grade:	>= 90.00	A
	88.00 ~ 89.99	A-
	85.00 ~ 87.99	B+
	80.00 ~ 84.99	B
	78.00 ~ 79.99	B-
	75.00 ~ 77.99	C+
	70.00 ~ 74.99	C
	68.00 ~ 69.99	C-
	<= 67.99	F



Notes

Tentative Topics and Dates

Part I: Regression Analysis (August 20 – September 24)

- Review of Simple Linear Regression
- Multiple Linear Regression: Statistical Inference; Model Selection and Diagnostics
- Regression Models with Quantitative and Qualitative Predictors
- Nonlinear and Non-parametric Regression

Part II: Categorical Data Analysis (September 29 – October 20)

- Review of Inference for Proportions and Contingency Tables
- Relative Risk and Odds Ratio
- Logistic Regression and Poisson Regression



Notes

Tentative Topics and Dates cont'd

Part III: Experimental Design (October 22 – November 12)

- Introduction to Experimental Design: Principles and Techniques
- Completely randomized Designs, Block Designs, Latin Square Designs, Nested and Split-Plot Designs
- Computer experiments

Part IV: Multivariate, Spatial and Time Series Analysis (November 17 – December 3)

- Discriminate Analysis, Principle Components Analysis, and Cluster Analysis
- Basic of time series and spatial data analysis



Notes

Computing

We will use software to perform statistical analyses. The recommended software for this course are JASP and R/Rstudio

● JASP

- a **free/open-source** graphical program for statistical analysis
- available at <https://jasp-stats.org/>

● R / Rstudio

- a **free/open-source** programming language for statistical analysis
- available at <https://www.r-project.org/> (R); <https://rstudio.com/> (Rstudio)

You are welcome to use a different package (e.g. SAS, JMP, SPSS, Minitab) if you prefer



Notes

Tell us about yourself

CLEMSON
UNIVERSITY

Who is the instructor?
Class Policies / Schedule
Tell us about yourself
Simple Linear Regression
SLR Parameter Estimation
Residual Analysis

1.13

Notes

Tell us about yourself

- Your name
- Degree program
- Your background in Statistics/Computing

CLEMSON
UNIVERSITY

Who is the instructor?
Class Policies / Schedule
Tell us about yourself
Simple Linear Regression
SLR Parameter Estimation
Residual Analysis

1.14

Notes

Review of Simple Linear Regression

CLEMSON
UNIVERSITY

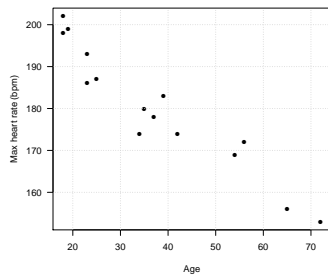
Who is the instructor?
Class Policies / Schedule
Tell us about yourself
Simple Linear Regression
SLR Parameter Estimation
Residual Analysis

1.15

Notes

What is Regression Analysis?

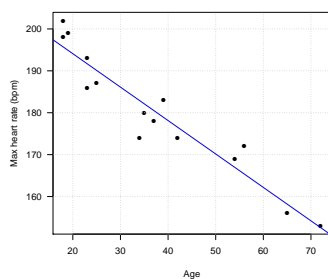
Regression analysis: A set of statistical procedures for estimating the relationship between **response variable** and **predictor variable(s)**



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

Notes

Scatterplot: Is Linear Trend Reasonable?



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

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

Notes

Simple Linear Regression (SLR)

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

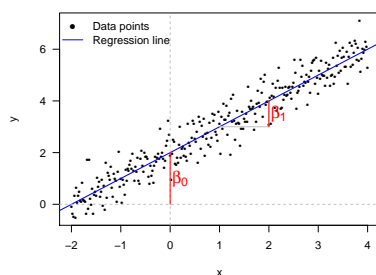
- In SLR we **assume** there is a **linear relationship** between X and Y :

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

- We need to estimate β_0 (intercept) and β_1 (slope)
- 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)

Notes

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



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

CLEMSON UNIVERSITY

Who is the instructor?

Class Policies / Schedule

Tell us about yourself!

Simple Linear Regression

SLR Parameter Estimation

Residual Analysis

1.19

Notes

Assumptions about the Random Error ε

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

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

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 mean curve** whereas σ^2 measures the magnitude of the **variation** around the regression curve

CLEMSON UNIVERSITY

Who is the instructor?

Class Policies / Schedule

Tell us about yourself!

Simple Linear Regression

SLR Parameter Estimation

Residual Analysis

1.20

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*:

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

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

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

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

We also need to **estimate** σ^2

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

CLEMSON UNIVERSITY

Who is the instructor?

Class Policies / Schedule

Tell us about yourself!

Simple Linear Regression

SLR Parameter Estimation

Residual Analysis

1.21

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 ε_i

CLEMSON UNIVERSITY

Who is the instructor?

Class Policies / Schedule

Tell us about yourself

Simple Linear Regression

SLR Parameter Estimation

Residual Analysis

1.22

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:

`MaxHeartRate = 220 – Age.`

Suppose we have 15 people of varying ages are tested for their maximum heart rate (bpm) (link to the "dataset": [whitneyhuang83.github.io/STAT8010/Data/maxHeartRate.csv](https://github.com/whitneyhuang83/STAT8010/Data/maxHeartRate.csv))

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

CLEMSON UNIVERSITY

Who is the instructor?

Class Policies / Schedule

Tell us about yourself

Simple Linear Regression

SLR Parameter Estimation

Residual Analysis

1.23

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

CLEMSON UNIVERSITY

Who is the instructor?

Class Policies / Schedule

Tell us about yourself

Simple Linear Regression

SLR Parameter Estimation

Residual Analysis

1.24

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	56	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	-18.33	-14.33	4.67	-19.33	1.67	-9.33
	21.73	5.73	6.73	-0.27	-24.27	-11.27	-8.27	-8.27	-27.27	18.73	12.73	-6.27	17.73	2.73	-2.27
	-420.18	-82.18	-83.04	0.62	-671.38	-187.78	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	205.44	21.78	373.78	2.78	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	179.94	180.53

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


CLEMSON
UNIVERSITY

Who is the instructor?
Class Policies / Schedule
Tell us about yourself
Simple Linear Regression
SLR Parameter Estimation
Residual Analysis

1.25

Notes

Let's Double Check

Output from  (R 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
```

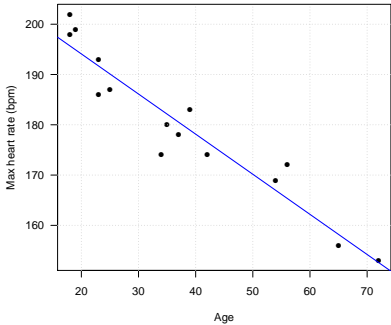
CLEMSON
UNIVERSITY

Who is the instructor?
Class Policies / Schedule
Tell us about yourself
Simple Linear Regression
SLR Parameter Estimation
Residual Analysis

1.26

Notes

Linear Regression Fit



Max heart rate (bpm)

Age

CLEMSON
UNIVERSITY

Who is the instructor?
Class Policies / Schedule
Tell us about yourself
Simple Linear Regression
SLR Parameter Estimation
Residual Analysis

1.27

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

Notes

Residuals

- The residuals are the differences between the observed and fitted values:

$$e_i = Y_i - \hat{Y}_i,$$

where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

- e_i is NOT the error term $\varepsilon_i = Y_i - E[Y_i]$
- Residuals are very useful in assessing the appropriateness of the assumptions on ε_i . Recall
 - $E[\varepsilon_i] = 0$
 - $Var[\varepsilon_i] = \sigma^2$
 - $Cov[\varepsilon_i, \varepsilon_j] = 0, \quad i \neq j$

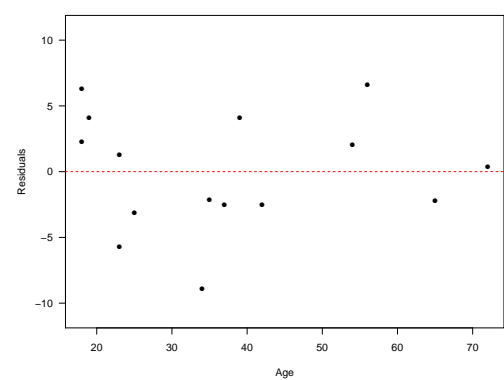
CLEMSON
UNIVERSITY

Who is the instructor?
Class Policies / Schedule
Tell us about yourself
Simple Linear Regression
SLR Parameter Estimation
Residual Analysis

1.28

Notes

Maximum Heart Rate vs. Age Residual Plot: ε vs. X



CLEMSON
UNIVERSITY

Who is the instructor?
Class Policies / Schedule
Tell us about yourself
Simple Linear Regression
SLR Parameter Estimation
Residual Analysis

1.29

Notes

Interpreting Residual Plots

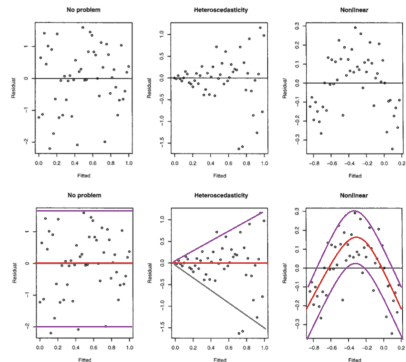


Figure: Figure courtesy of Faraway's Linear Models with R (2005, p. 59).

CLEMSON
UNIVERSITY

Who is the instructor?
Class Policies / Schedule
Tell us about yourself
Simple Linear Regression
SLR Parameter Estimation
Residual Analysis

1.30

Notes

Summary

In this lecture, we reviewed

- 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

- ➊ More on residual analysis
- ➋ Normal Error Regression Model and statistical inference for β_0 , β_1 , and σ^2
- ➌ Prediction



Who is the instructor?
Class Policies / Schedule
Tell us about yourself
Simple Linear Regression
SLR Parameter Estimation
Residual Analysis

Notes

Notes

Notes
