RESEARCH METHODS FOR CLINICAL INVESTIGATORS
Session 7:

Correlation & Linear Regression

Veronica G. Brown, PhD, MSPH

Objectives

At the end of the presentation, the audience will be able to:

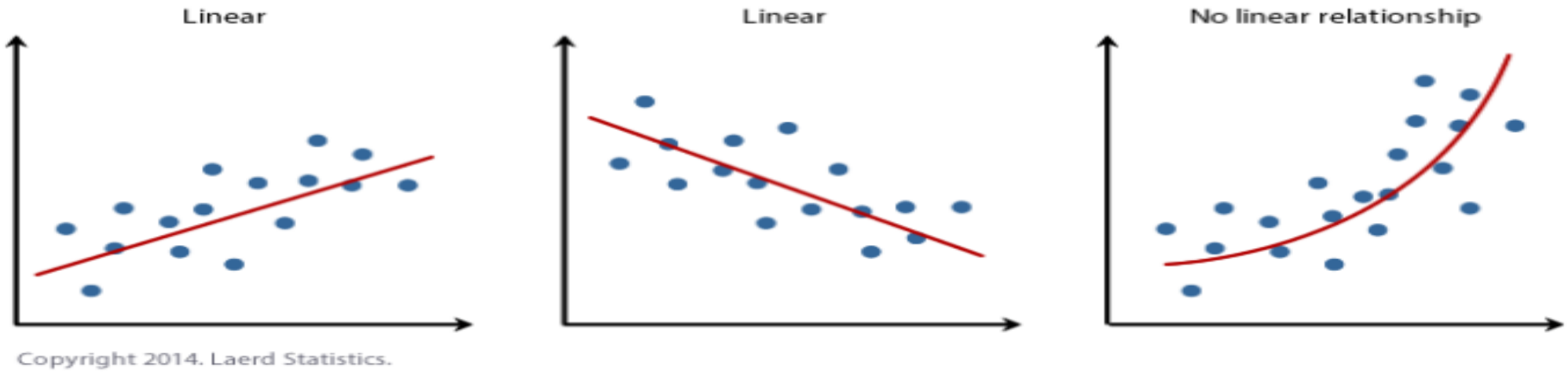
- Understand the concept of a linear relationship
- Understand the concept of correlation
- Understand and explain regression
  - Independent vs. Dependent Variable
  - Slope and intercept

Linear Relationship

- What does it mean?

  Any relationship bet

  ween two variables that creates a line



A scatterplot to check for linear relationship. Image courtesy: Laerd Statistics

1.) Laerd Statistics (2014)
2.) Baker, Charlotte (2014). Introduction to Epidemiology. Virginia Tech University

Correlation

- The linear association between two variables
  - Comparisons
    - Two continuous (numerical) variables
    - A continuous variable and a categorical variable
  - Examine the linear relationship between the two variables

Correlation cont'd

- Pearson's r
  - Correlation coefficient
  - Ranges from -1 to +1

- Correlation Strength
  - 0-0.19: Very weak
  - 0.20-0.39: Weak
  - 0.4-0.59: Moderate
  - 0.60-0.79: Strong
  - 0.80-1: Very Strong

Baker, Charlotte (2014). Introduction to Epidemiology. Virginia Tech University

Correlation cont'd

Correlation Matrix

	Hours spent studying	Exam score	IQ score	Hours spent sleeping	School rating
Hours spent studying	1.00	0.82	0.48	-0.22	0.36
Exam score	0.82	1.00	0.33	-0.04	0.23
IQ score	0.08	0.33	1.00	0.06	0.02
Hours spent sleeping	-0.22	-0.04	0.06	1.00	0.12
School rating	0.36	0.23	0.02	0.12	1.00

1.) https://www.statology.org/how-to-read-a-correlation-matrix/
2.) Motulsky, Harvey. (2014). *Intuitive Biostatistics: Nonmathematical Guide to Statistical Thinking*. Oxford University Press

Linear Regression

- Examines the relationship between two or more variables in a model
  - Independent variables added to a model can be continuous (numerical) or categorical

- Outcome of interest: Continuous
  - Similar to ANOVA

Motulsky, Harvey. (2014). *Intuitive Biostatistics: Nonmathematical Guide to Statistical Thinking*. Oxford University Press

Linear Regression cont'd

- Determines the specific influence of the independent variable on the outcome

Ex. Influence of Age on Weight

What is the relationship between age and weight

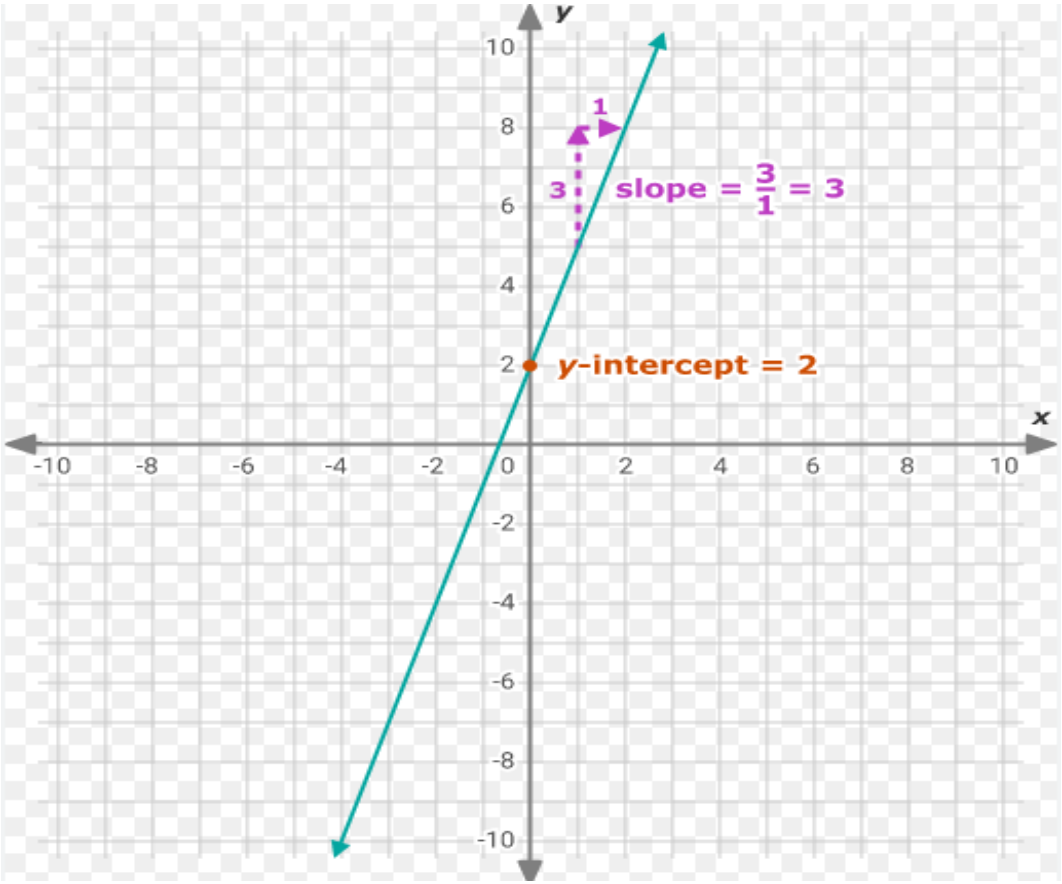
Baker, Charlotte (2014). Introduction to Epidemiology. Virginia Tech University

Linear Regression cont'd

- Slope & y-intercept

Slope: Describes the steepness of the line

Y-intercept: Point where the line crosses the y-axis



1.) Motulsky, Harvey. (2014). *Intuitive Biostatistics: Nonmathematical Guide to Statistical Thinking.* Oxford University Press
2.) https://www.ixl.com/math/lessons/slope-intercept-form

Linear Regression cont'd

- Slope & y-intercept

$$Y = \beta_0 + \beta_1 x$$

β_1= slope, X is the independent variable

β_0= y-intercept, Y is the dependent variable (outcome)

Linear Regression cont'd

- Age and Weight
  - Age (Independent Variable)
  - Weight (Dependent/Outcome Variable)

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-76.10604	538.56978	-0.14	0.8879
AGE	REPORTED AGE IN YEARS	1	11.93485	9.60989	1.24	0.2172

Baker, Charlotte (2014). Introduction to Epidemiology. Virginia Tech University

Linear Regression cont'd

- Age
  - Parameter Estimate: 11.9
  - Std. Error: 9.6
  - P-value: 0.22

- Interpretation

For every 1 year increase in age, there is an increase of approximately 12 pounds in weight

**P-value is not statistically significant

Baker, Charlotte (2014). Introduction to Epidemiology. Virginia Tech University

Linear Regression cont'd

- Multiple Variables can be implemented into a linear model
  - Multiple Linear Regression

Ex. Influence of Age and Sex on Weight

What is the relationship between age and sex on weight?

Linear Regression cont'd

- Age and Weight
  - Age & Sex (Independent Variables)
  - Weight (Dependent/Outcome Variable)

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-724.34408	733.80043	-0.99	0.3260
AGE	REPORTED AGE IN YEARS	1	9.04482	9.83323	0.92	0.3599
SEX	RESPONDENTS SEX	1	497.24959	383.84537	1.30	0.1982

Baker, Charlotte (2014). Introduction to Epidemiology. Virginia Tech University

Linear Regression cont'd

- Age
  - Parameter Estimate: 9.04
  - Std. Error: 9.8
  - P-value: 0.36

- Sex
  - Parameter Estimate: 497.2
  - Std. Error: 383.8
  - P-value: 0.20

**P-values are not statistically significant

Baker, Charlotte (2014). Introduction to Epidemiology. Virginia Tech University

Linear Regression cont'd

- Entire model is not significant

- Interpretation

For every 1 year increase in age, there is an increase of approximately 9 pounds in weight

**P-values are not statistically significant

1.) Baker, Charlotte (2014). Introduction to Epidemiology. Virginia Tech University
2.) Motulsky, Harvey. (2014). *Intuitive Biostatistics: Nonmathematical Guide to Statistical Thinking*. Oxford University Press

Linear Regression cont'd

- Multiple Variables can be implemented into a linear model
  - Multiple Linear Regression

Ex. Influence of several variables on Weight

What is the relationship between these variables and weight?

Age	Smoking Status
Sex	Sleep Quality
General Health	Education Level

Linear Regression cont'd

- Independent Variables
- Weight (Dependent/Outcome Variable)

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	-1058.20116	1037.99673	-1.02	0.3106
AGE	REPORTED AGE IN YEARS	1	7.00190	10.24668	0.68	0.4961
SEX	RESPONDENTS SEX	1	438.87772	392.02099	1.12	0.2658
GENHLTH	GENERAL HEALTH	1	-239.64731	162.91181	-1.47	0.1447
_SMOKER3	COMPUTED SMOKING STATUS	1	54.72797	161.29796	0.34	0.7352
QLREST2	HOW MANY DAYS DID YOU GET ENOUGH SLEEP IN PAST 30 DAYS	1	2.16899	4.97255	0.44	0.6637
_EDUCAG	COMPUTED LEVEL OF EDUCATION COMPLETED CATEGORIES	1	289.59263	154.54648	1.87	0.0641

Baker, Charlotte (2014). Introduction to Epidemiology. Virginia Tech University

Linear Regression cont'd

- Age
  - Parameter Estimate: 7.0
  - Std. Error: 10.2
  - P-value: 0.50

- Sex
  - Parameter Estimate: 438.9
  - Std. Error: 392.0
  - P-value: 0.27

**P-values are not statistically significant

Baker, Charlotte (2014). Introduction to Epidemiology. Virginia Tech University

Linear Regression cont'd

- General Health
  - Parameter Estimate: -239.6
  - Std. Error: 162.9
  - P-value: 0.14

- Smoking Status
  - Parameter Estimate: 54.7
  - Std. Error: 161.3
  - P-value: 0.74

**P-values are not statistically significant

Baker, Charlotte (2014). Introduction to Epidemiology. Virginia Tech University

Linear Regression cont'd

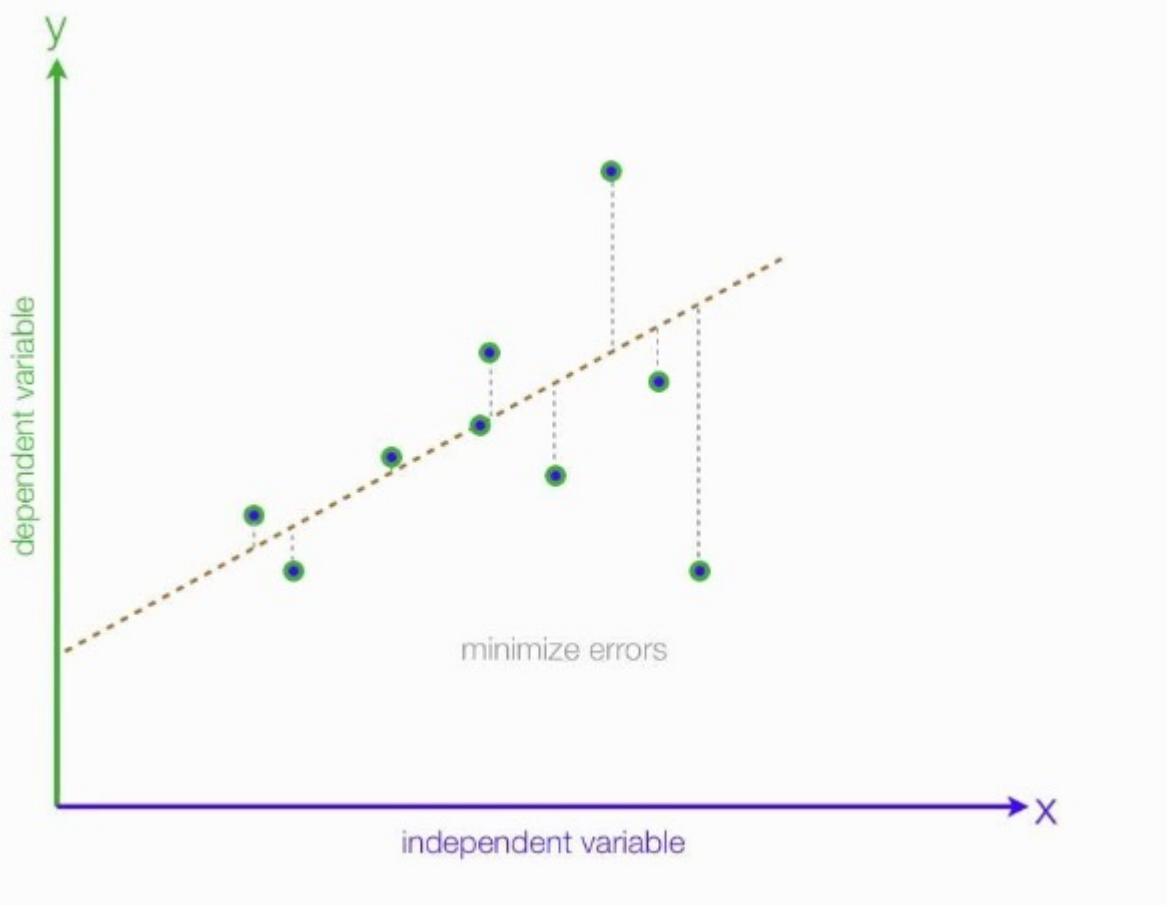
- Sleep Quality
  - Parameter Estimate: 2.17
  - Std. Error: 4.97
  - P-value: 0.66
- Education
  - Parameter Estimate: 289.6
  - Std. Error: 154.5
  - P-value: 0.06

**P-values are not statistically significant

Baker, Charlotte (2014). Introduction to Epidemiology. Virginia Tech University

Summary

- Linear Relationships
  - Correlation
  - Linear Regression



Gordis, Leon. (2018). *Epidemiology*. Saunders Elsevier.