Part 1: Classic Outcome Variables in Clinical Research

Review of Linear Regression

Course: CRP245

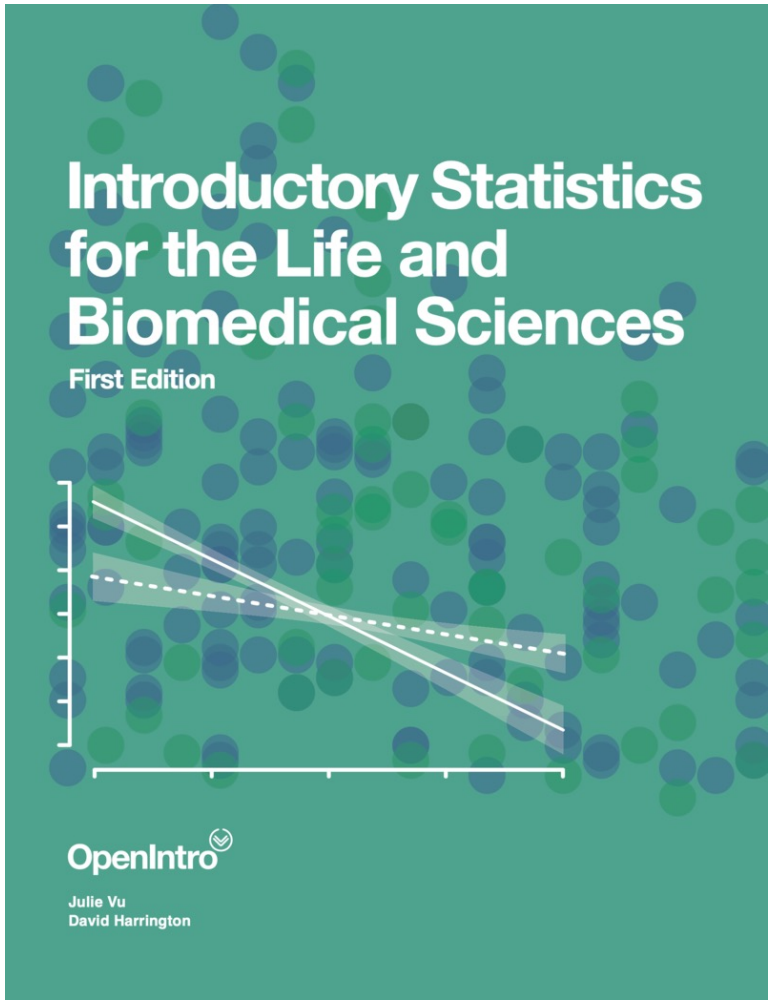
Duke University School of Medicine

CRTP CLINICAL RESEARCH TRAINING PROGRAM

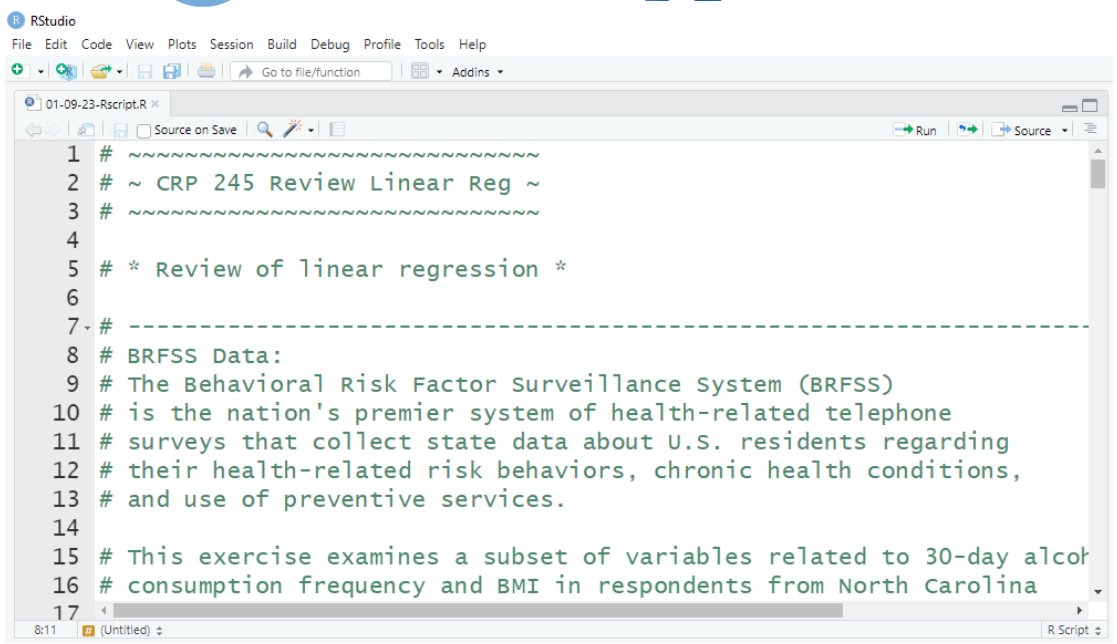
Recommended Reading for This Session

- Introductory Statistics for the Life and Biomedical Sciences

- Go to this URL (https://leanpub.com/biostat). You can click on the 'read free sample' button, and it will download a full PDF of the book
- Read Chapter 6 (Simple Linear Regression), pages 290-316. Most of this should be a review of our discussion of regression in 241.

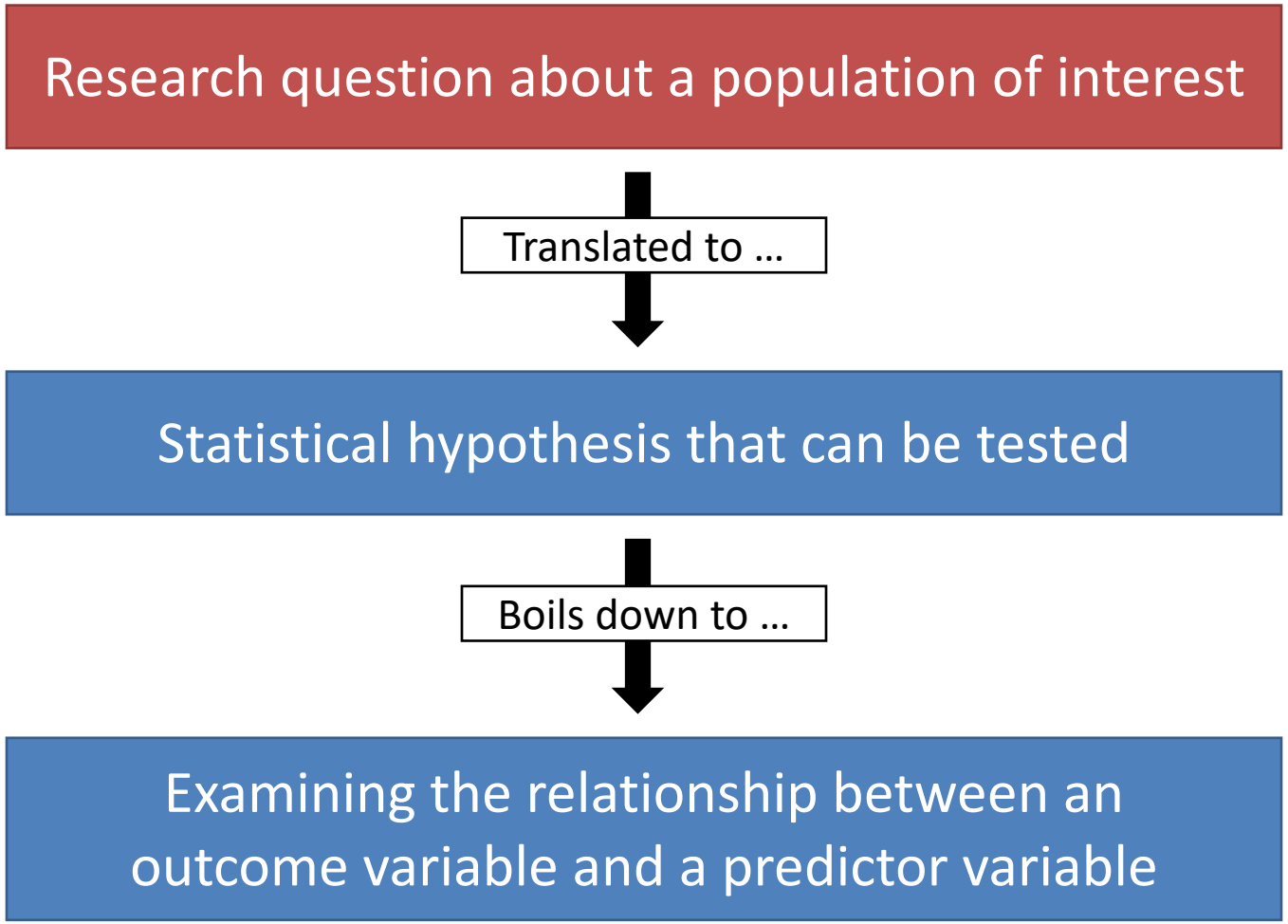


Other resources associated with this session



RScript file in Teams

Clinical Research Paradigm (Simplified)



Last Term We Discussed …

Can pick a statistical approach by answering the questions:

Is the outcome a continuous or categorical measure?

&

Are the covariate(s) of interest continuous or categorical measures?

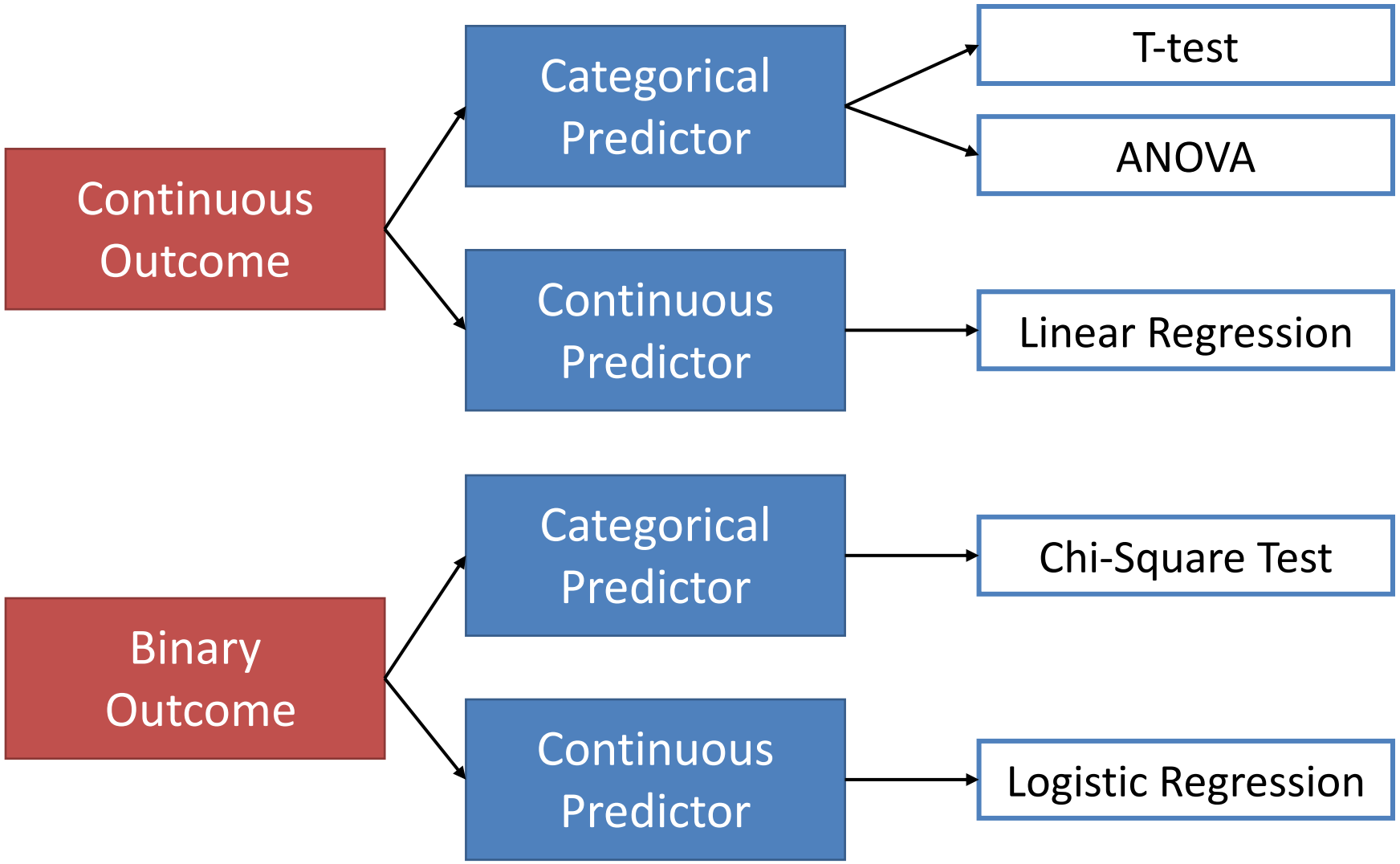
Answering the questions listed above can narrow the potential choices of statistical approaches.

- Then use the following questions to refine the choices to pick the most appropriate approach!

Are there any features of the data set that should be considered when selecting the statistical approach?

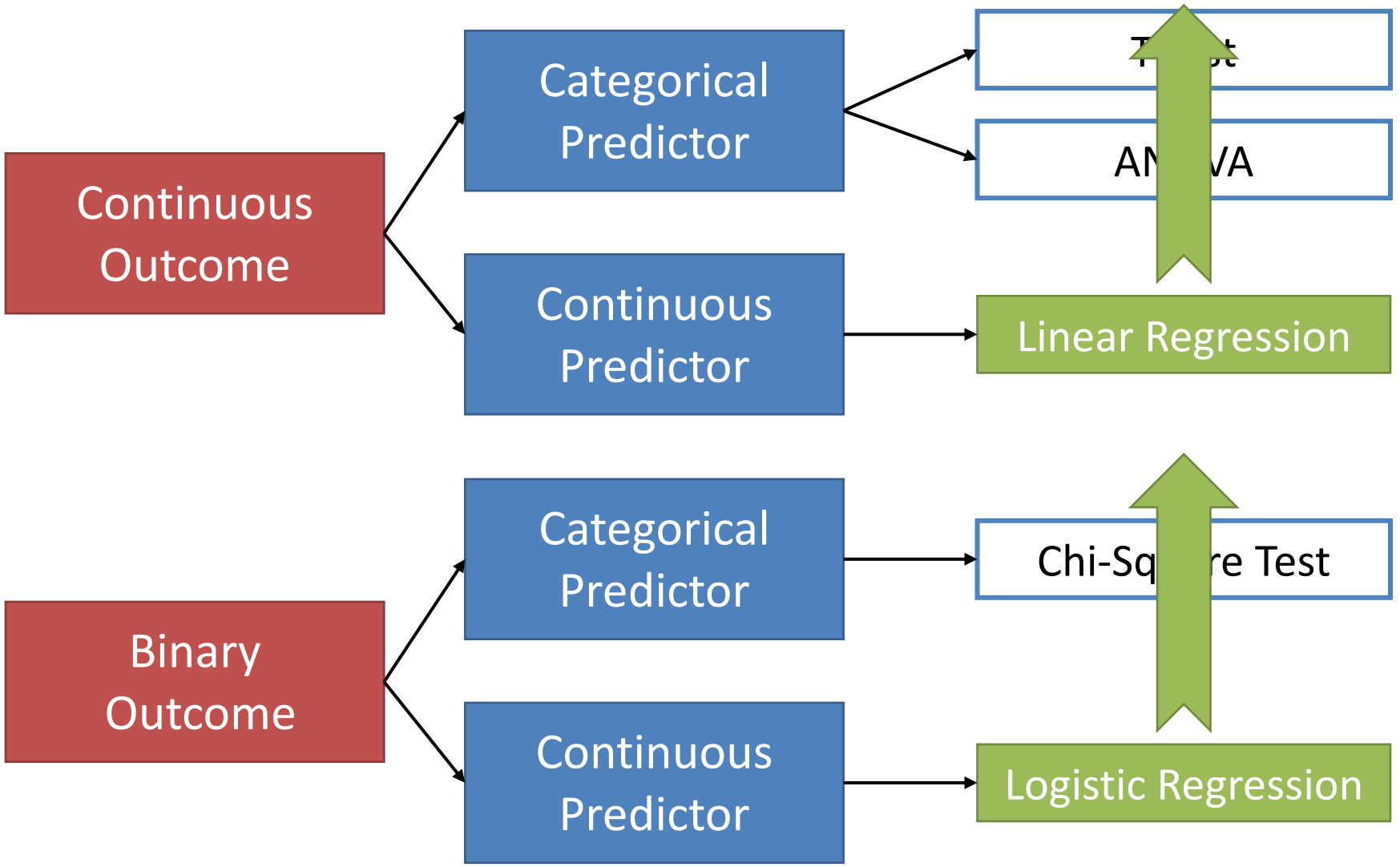
Are there any study design features that should be considered when selecting the statistical approach?

Last Semester We Discussed …



* And the nonparametric equivalents!

Last Semester We Discussed …



* And the non-parametric equivalents!

Regression is a Unifying Framework!

- Can use regression models to …

> Examine the relationship between an outcome variable and a predictor variable

- The Catch:
  - There is a different regression model for different types of outcome variables.

- No Worries!
  - Need to become familiar with the regression models available for the outcome variables typically encountered in clinical research.
  - Welcome to CRP 245 Part 1!

Classic Outcomes in Clinical Research

- **Continuous** * Review *
  - Example: Forced expiratory volume (liters)
  - Regression Model: Linear
- **Binary** * Review *
  - Example: Rejection status 30 days after lung transplant
  - Regression Model: Logistic
- **Time-to-Event** * Review *
  - Example: Time to graft failure after lung transplant
  - Regression Model: Cox Proportional Hazards
- **Count**
  - Example: Number of respiratory hospitalizations
  - Regression Model: Poisson or Negative Binomial

Goal of Part 1

- For each outcome type, understand:
  - How to visualize the data
  - How to analyze (i.e., model) the data
    - Specifically, what do the regression coefficients estimate?

- We will not discuss these topics until later in the course …
  - How different covariates work together in a model
  - How to check model assumptions
  - How to assess model fit and compare different models
  - How to pick the covariates to go into a model
  - We will get to these in Part 2 & Part 3

Review of Linear Regression for Continuous Outcomes

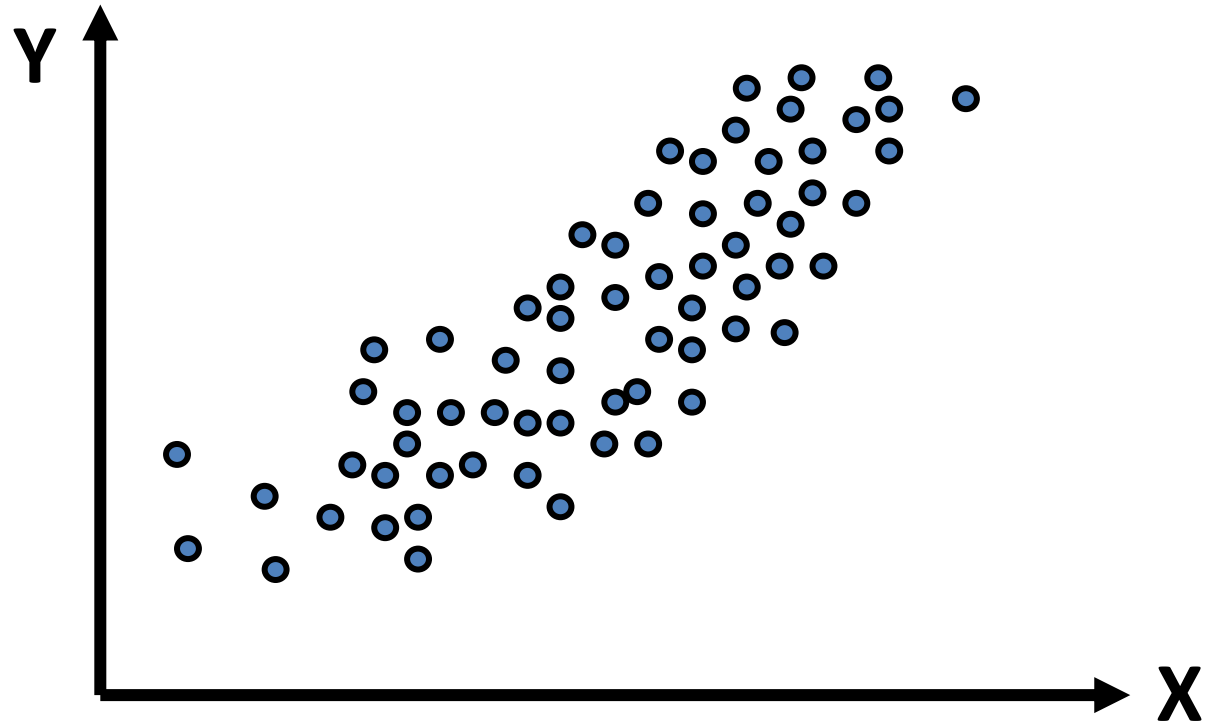
- Linear regression can be used to examine the **relationship** between a **continuous response** (Y) variable and **predictor variable** (X).
  - In clinical research, the term “relationship” is often used interchangeably with “association” or “correlation.”

- Linear regression was defined initially for continuous predictor variables.

- However, it can also be used with categorical predictor variables.

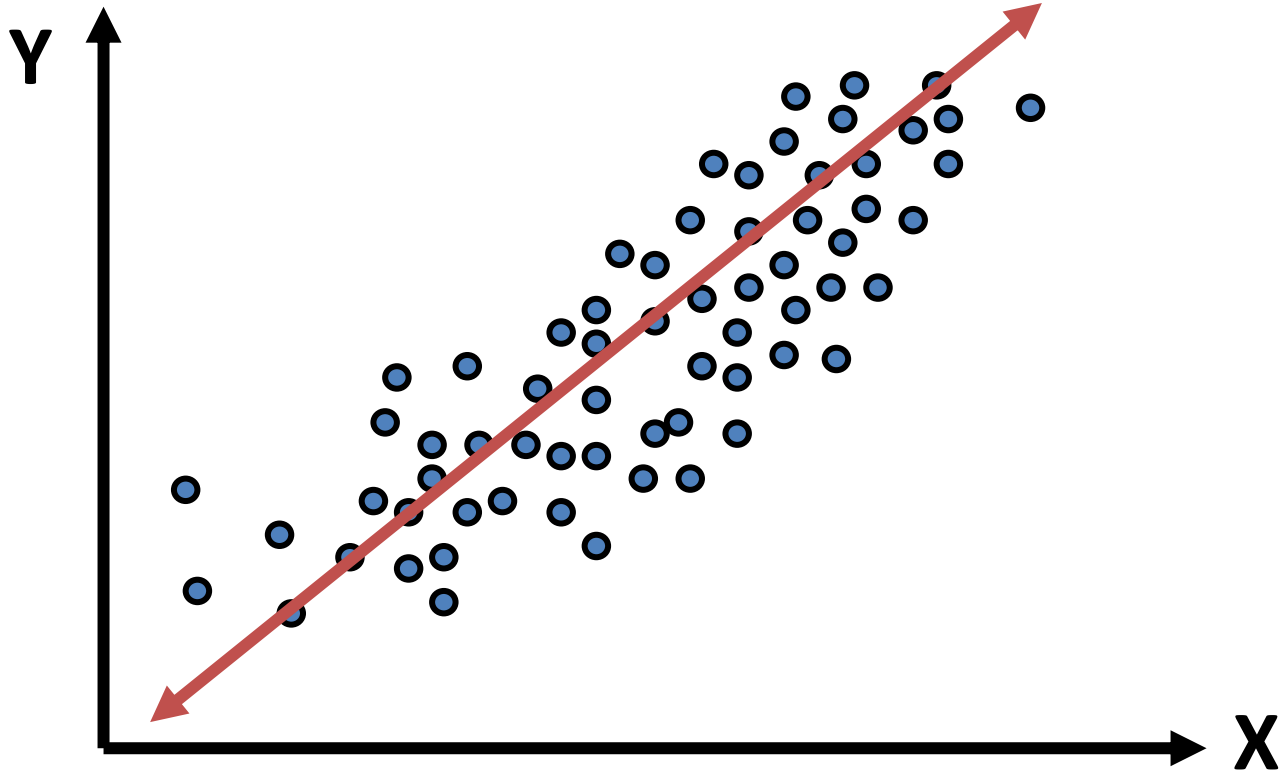
Visualizing the Data

- The relationship between X and Y can be visualized by creating a **scatterplot** of the observations in the data.
  - Plots the explanatory variable on the x-axis and the response variable on the y-axis as coordinate pairs.

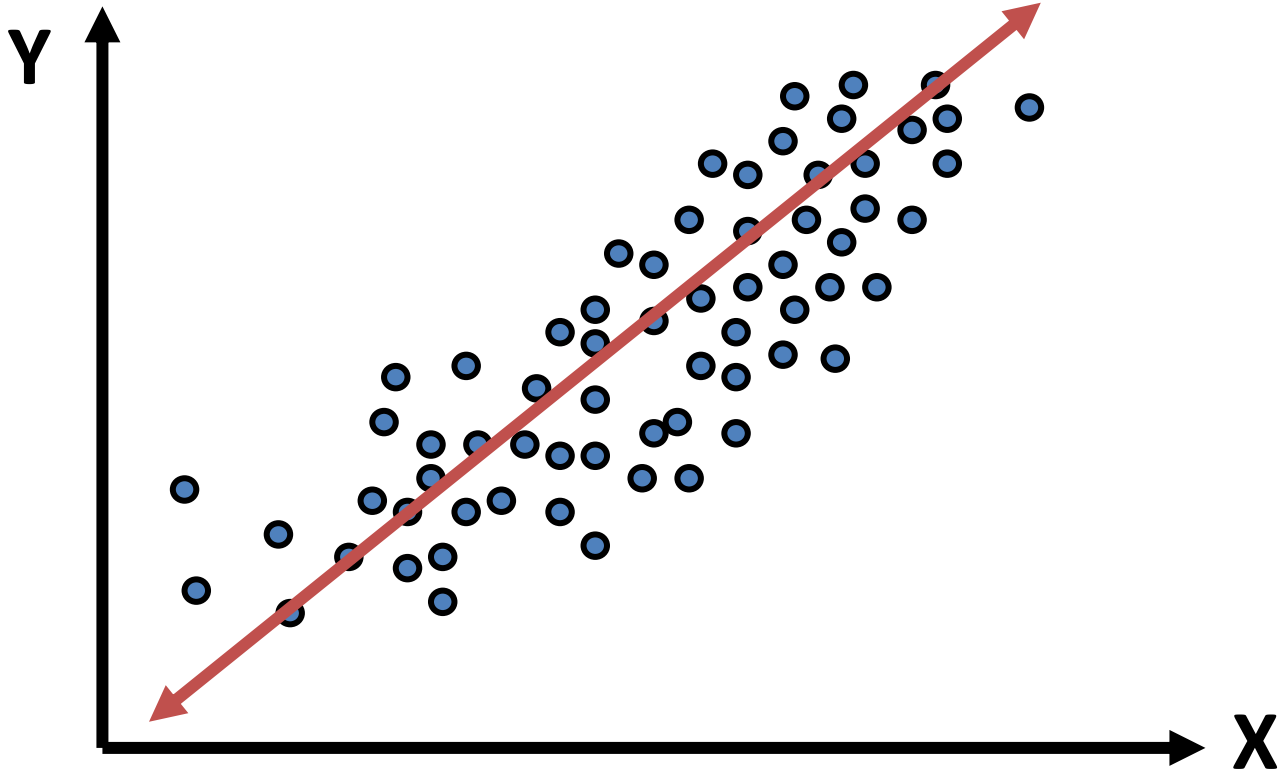


Visualizing the Data

- The regression model is estimated by fitting a line to the data. That is, regression finds the "**line of best fit**" by finding the line that is closest to all data points on average.



What Should You Look for in the Scatter Plot?



- Before fitting a linear regression model for Y as a function of X, you should look for evidence of ….
  - A **linear trend** in the data.
  - A **constant variance** of Y across the range of X.
- Linearity violations are common in clinical research.
  - We will discuss handling those in regression modeling in Part 2 of the course.

Linear Regression for Continuous Outcomes

- Models the linear relationship between continuous Y (outcome) and continuous X (predictor) by fitting a line to the data that estimates Y as a function of X.

- That is, there is an equation that represents the TRUE linear relationship between X and Y in the population.
  - This is the Regression **Model**.

Regression Model: $Y = \beta_0 + \beta_1 \cdot X$

- Then, that equation is estimated using sample data of the (X, Y) pairs from the population.
  - This is the **Fitted** Regression **Line**.

Fitted Regression Line: $\hat{Y} = b_0 + b_1 \cdot X$

Linear Regression for Continuous Outcome

- More specifically, the linear regression (LR) model estimates the **mean value of Y** for all subjects in the population **for a given value of X**.

- That is, the LR model can estimate Y's expected value (i.e., mean value) for all observed values of X in the population.

$$\text{Regression Model: } \mu(Y \mid X) = \beta_0 + \beta_1 \cdot X$$

- So LR is still estimating/comparing means like the t-test and ANOVA; it is doing it continuously!

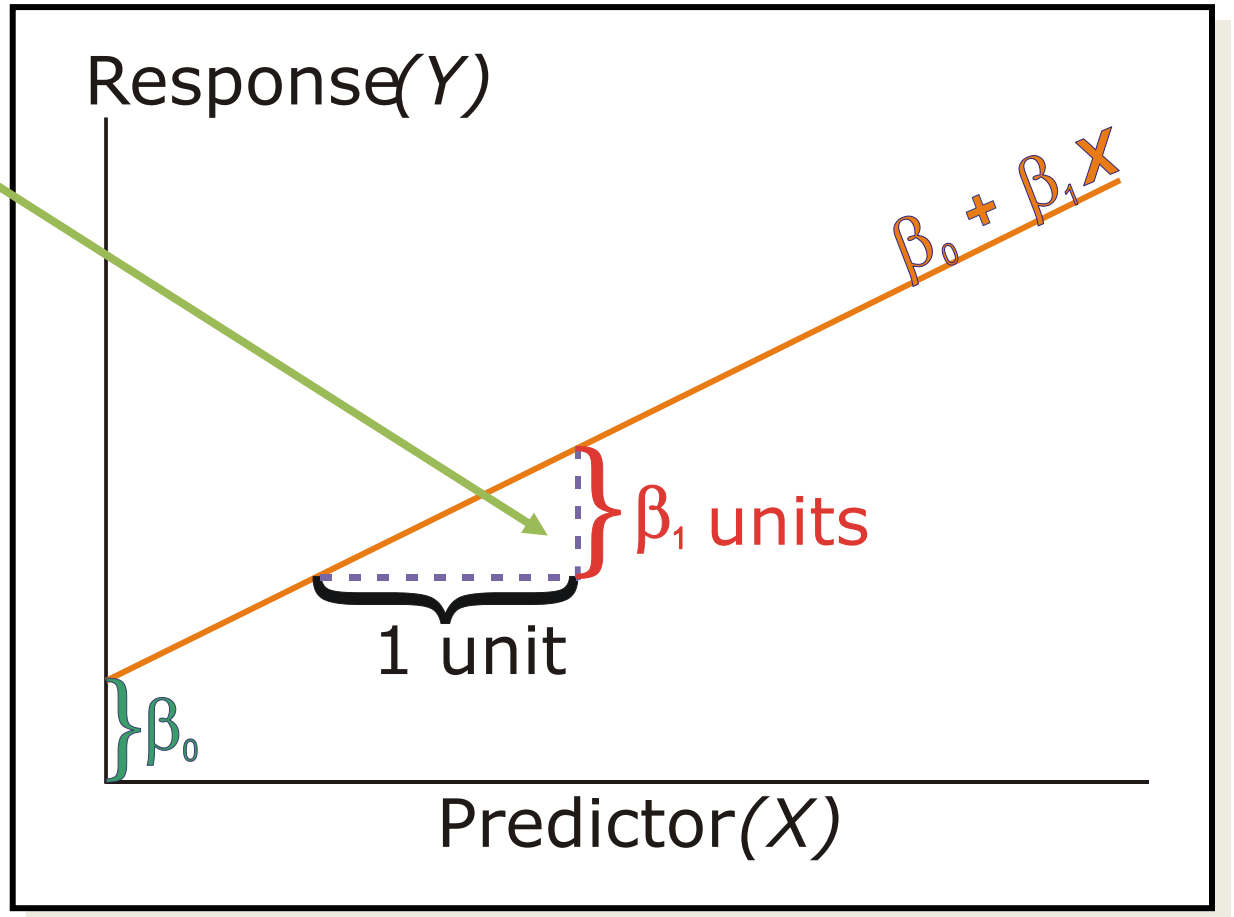
What do the LR Model Parameters Estimate?

Regression Model: $Y = \beta_0 + \beta_1 \cdot X$

β_1 is the **SLOPE** of the LR model.

It quantifies the **EXPECTED CHANGE** in the **MEAN of Y** for a 1 UNIT CHANGE in X.

That is, what is the difference in the mean of Y for subjects whose $X = x_0$ vs. $X = x_0+1$?

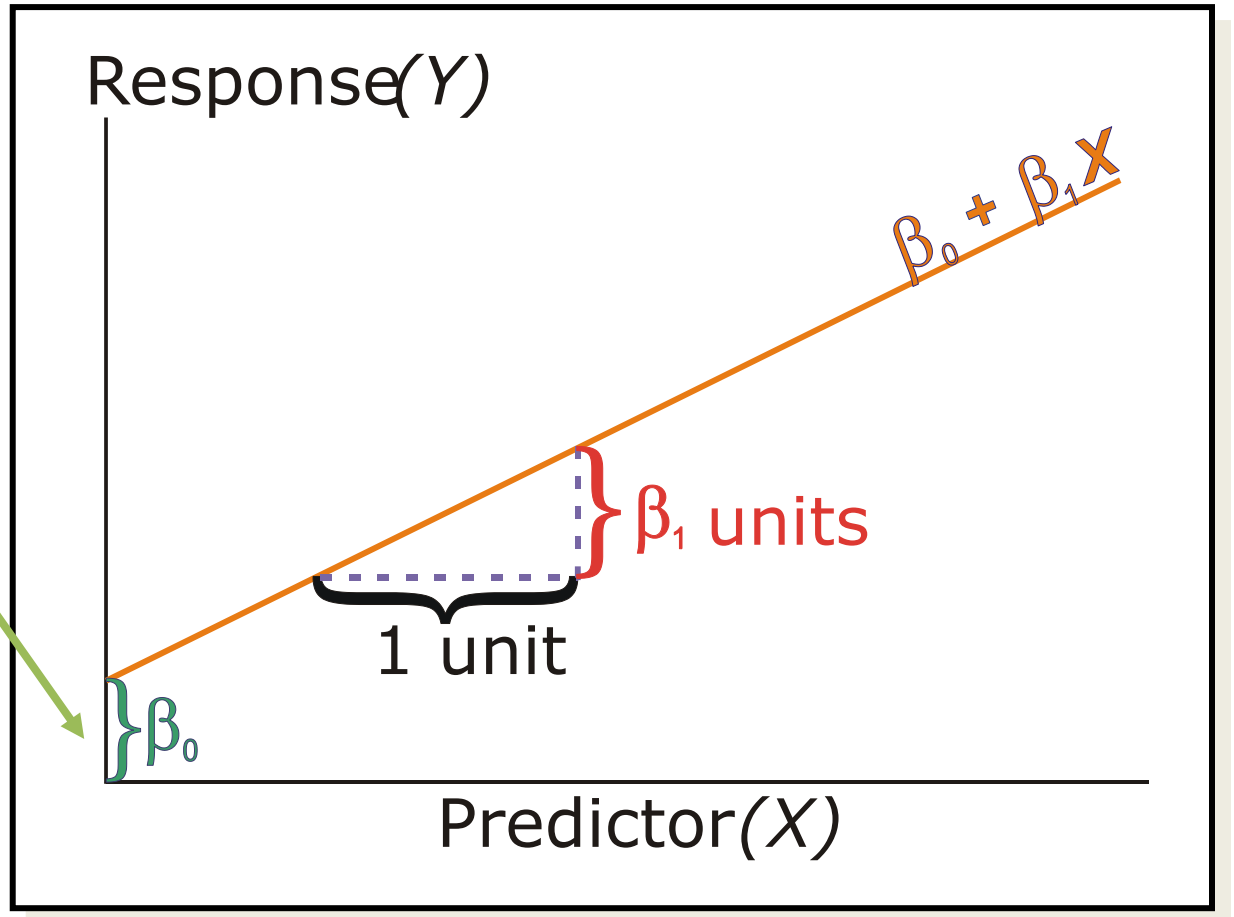


What do the LR Model Parameters Estimate?

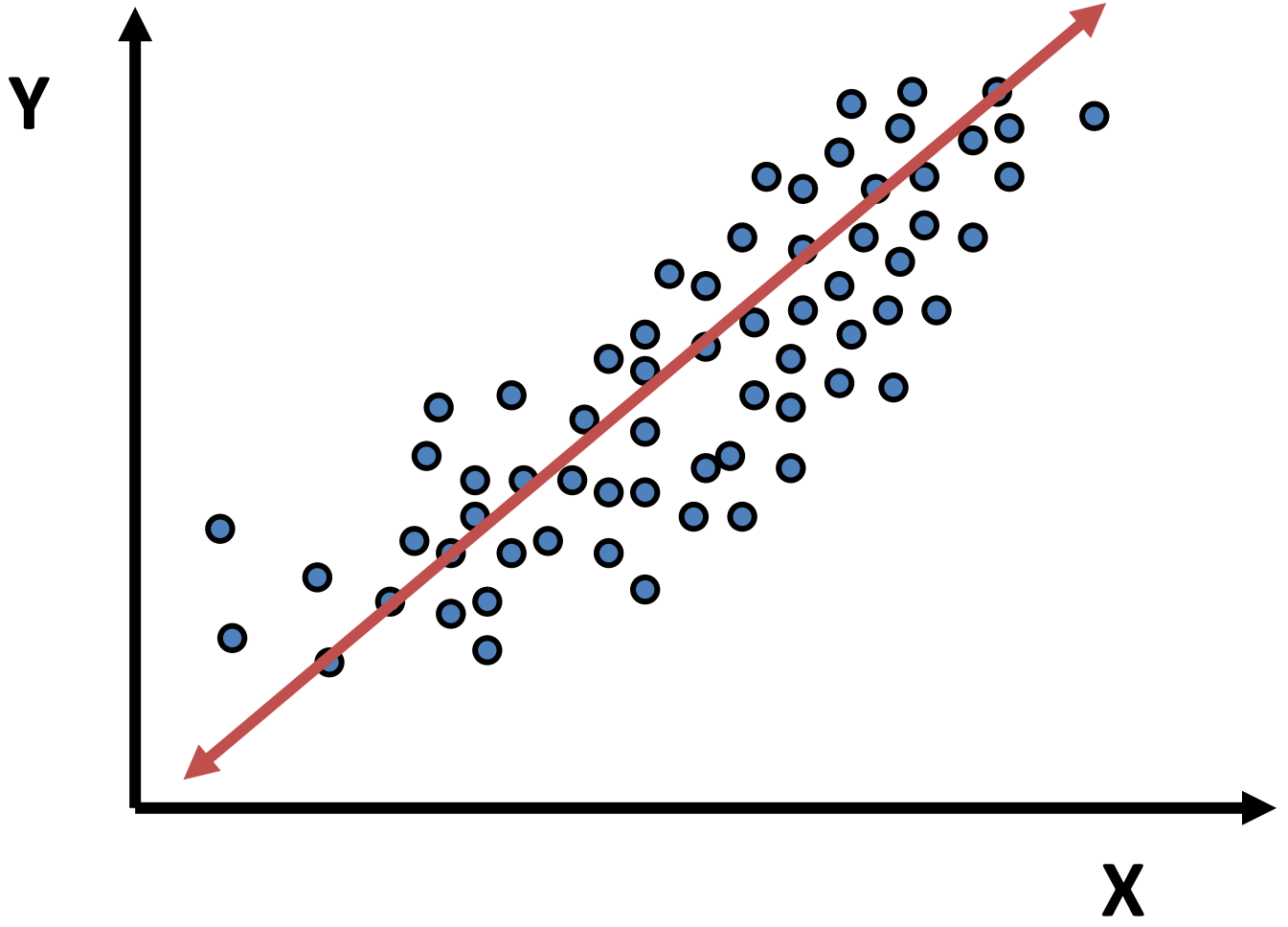
Regression Model: $Y = \beta_0 + \beta_1 \cdot X$

β_0 is the **INTERCEPT** of the LR model.

It quantifies the **EXPECTED MEAN of Y** when $X = 0$ (may not be meaningful if X=0 is not a plausible value in the population).



What do the LR Model Parameters Estimate?

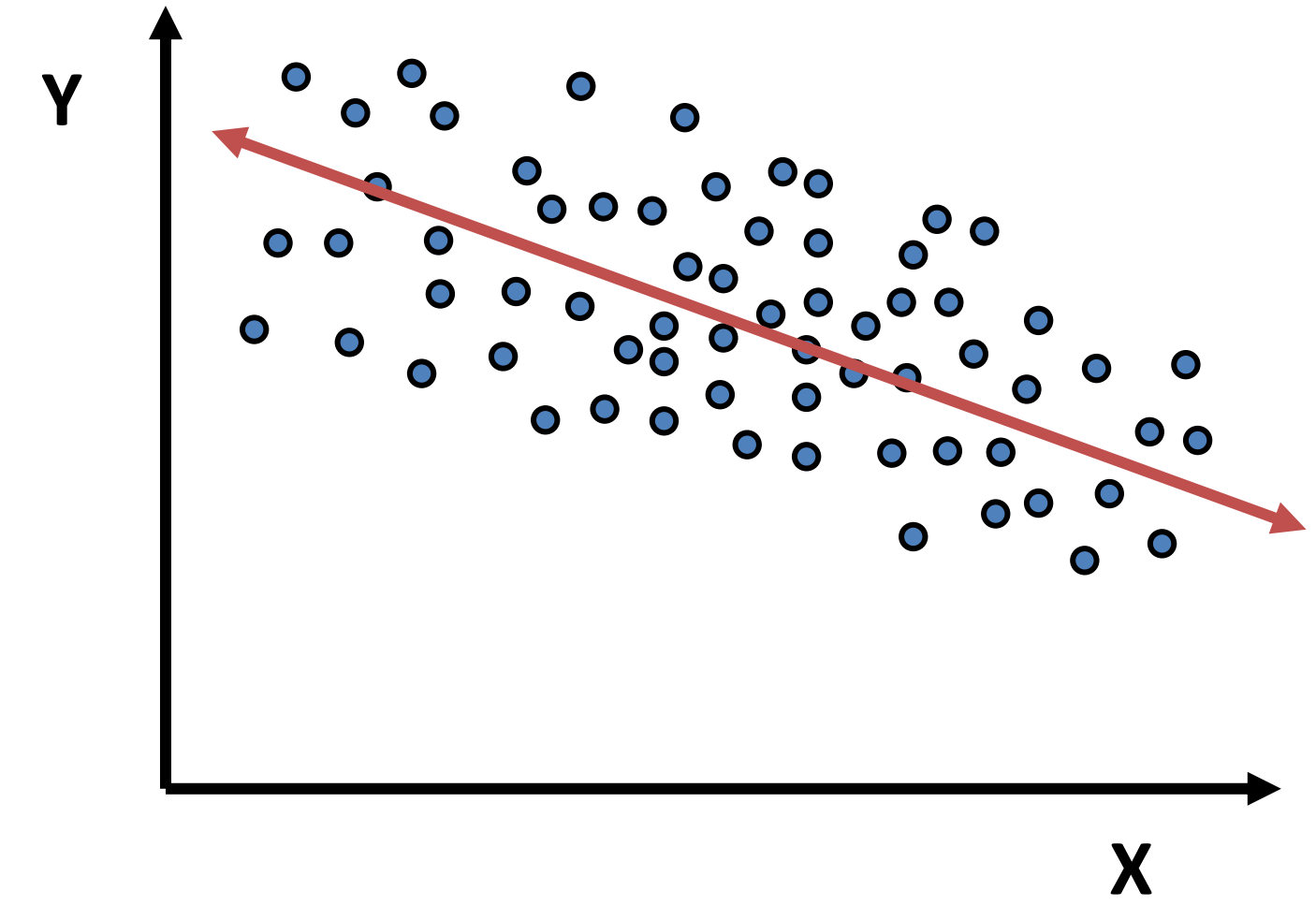


$\beta_1 > 0$

As X increases, so does the mean value of Y.

That is, X and Y are linearly correlated.

What do the LR Model Parameters Estimate?

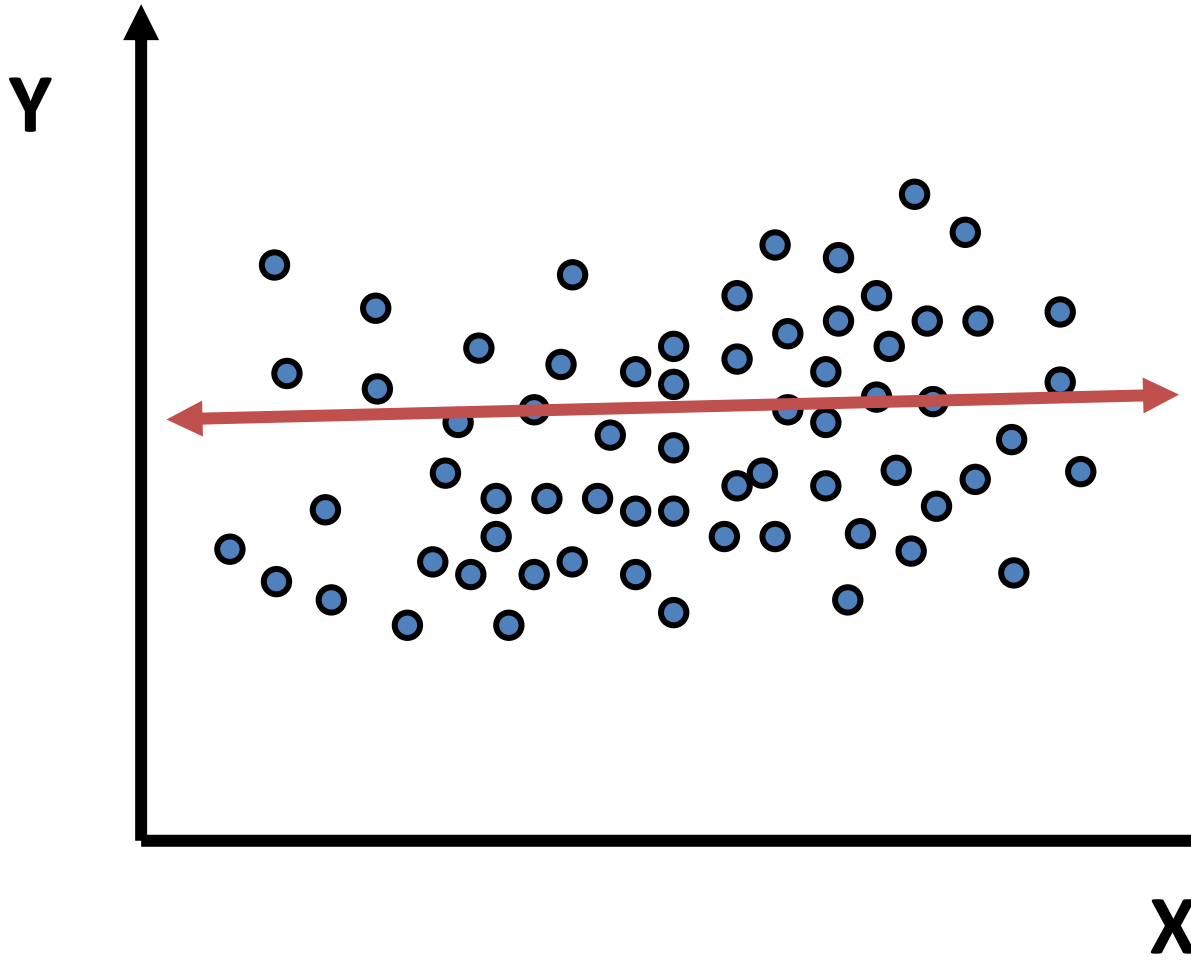


$\beta_1 < 0$

As X increases, the mean value of Y decreases.

That is, X and Y are linearly correlated.

What do the LR Model Parameters Estimate?



$\beta_1 \approx 0$

As X increases, the mean value of Y stays constant.

That is, X and Y are NOT linearly correlated.

Testing for Association using Linear Regression

Use a linear regression model to quantify the **direction** and **magnitude** of the linear correlation between two variables in a data set. Use the SLOPE estimate to test whether the correlation is statistically different from zero.

$H_o: \beta_1 = 0$	vs.	$H_1: \beta_1 \neq 0$
Outcome and Predictor are NOT linearly correlated.		Outcome and Predictor are linearly correlated.
Outcome and Predictor are NOT associated.		Outcome and Predictor are associated.

Significance Level

$\alpha = 0.05$

Testing for Association using Linear Regression

$H_o: \beta_1 = 0$ **vs.** $H_1: \beta_1 \neq 0$

- Should report the p-value for this test, the point estimate of β1, and its 95% confidence interval.
  - Confidence interval gives a range of plausible values of β_1.

- Can use the following R functions to obtain the LR modeling results:

Function	Description
plot()	scatterplot of X vs. Y
lm()	fits a linear regression model of Y as function of X
summary()	creates regression table with regression coefficients point estimates and p-value
confint()	computes the confidence interval for the point estimates

Testing for Association using Linear Regression

- R Code and Output:

```
> plot(MyData$X, MyData$Y)
> fit <- lm(Y ~ X, data=MyData)
> summary(fit)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)      ###         ###     ###      ###
          X      ###         ###     ###      ###
> confint(fit)
                2.5 %     97.5 %
(Intercept)       ###        ###
          X       ###        ###
```

Testing for Association using Linear Regression

- R Code and Output :

```
> plot(MyData$X, MyData$Y)
> fit <- lm(Y ~ X, data=MyData)
> summary(fit)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)       ###        ###     ###      ###
          X       ###        ###     ###      ###
> confint(fit)
                2.5 %     97.5 %
(Intercept)       ###        ###
          X       ###        ###
```

Use plot() function to create a scatterplot of the outcome and predictor variables
- X = Name of predictor variable in the data set
- Y = Name of outcome variable in the data set
- List X first (so it is plotted on the x-axis)!

Testing for Association using Linear Regression

- R Code and Output :

```
> plot(MyData$X, MyData$Y)
> fit <- lm(Y ~ X, data=MyData)
> summary(fit)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)       ###        ###     ###      ###
          X       ###        ###     ###      ###
> confint(fit)
                2.5 %     97.5 %
(Intercept)       ###        ###
          X       ###        ###
```

Use lm() function to fit linear regression model; The input is …
- Y = Name of outcome variable in the data set
- X = Name of predictor variable in the data set
- MyData = Name of data set

*** **Order of X and Y Matters!** ***

Testing for Association using Linear Regression

- R Code and Output :

```
> plot(MyData$X, MyData$Y)
> fit <- lm(Y ~ X, data=MyData)
> summary(fit)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)      ###         ###     ###      ###
          X      ###         ###     ###      ###
> confint(fit)
                2.5 %     97.5 %
(Intercept)       ###        ###
          X       ###        ###
```

Use the ~ to communicate to R that Y should be modeled “as function of” X.

Testing for Association using Linear Regression

- R Code and Output :

```
> plot(MyData$X, MyData$Y)
> fit <- lm(Y ~ X, data=MyData)
> summary(fit)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)       ###        ###     ###      ###
X                 ###        ###     ###      ###
> confint(fit)
                2.5 %     97.5 %
(Intercept)       ###        ###
X                 ###        ###
```

The output of interest is …
- P-value for association test between Y and X
- Point estimate of slope coefficient for X
- Confidence interval slope coefficient for X

Testing for Association using Linear Regression

- R Code and Output :

```
> plot(MyData$X, MyData$Y)
> fit <- lm(Y ~ X, data=MyData)
> summary(fit)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)       ###        ###     ###      ###
X                 ###        ###     ###      ###
> confint(fit)
                2.5 %     97.5 %
(Intercept)       ###        ###
X                 ###        ###
```

The output of interest is …
- P-value for association test between Y and X
- Point estimate of slope coefficient for X
- Confidence interval slope coefficient for X

Testing for Association using Linear Regression

- R Code and Output :

```
> plot(MyData$X, MyData$Y)
> fit <- lm(Y ~ X, data=MyData)
> summary(fit)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)       ###        ###     ###      ###
X                 ###        ###     ###      ###
> confint(fit)
                2.5 %     97.5 %
(Intercept)       ###        ###
X                 ###        ###
```

The output of interest is …
- P-value for association test between Y and X
- Point estimate of slope coefficient for X
- Confidence interval slope coefficient for X

Testing for Association using Linear Regression

- R Code and Output :

```
> plot(MyData$X, MyData$Y)
> fit <- lm(Y ~ X, data=MyData)
> summary(fit)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)       ###        ###     ###      ###
X                 ###        ###     ###      ###
> confint(fit)
                2.5 %     97.5 %
(Intercept)       ###        ###
X                 ###        ###
```

The output of interest is …
- P-value for association test between Y and X
- Point estimate of slope coefficient for X
- Confidence interval slope coefficient for X

Using Linear Regression Model for Estimation

- How could the LR model estimate the mean value of Y in the population when X is 10?

- The regression model represents the truth:

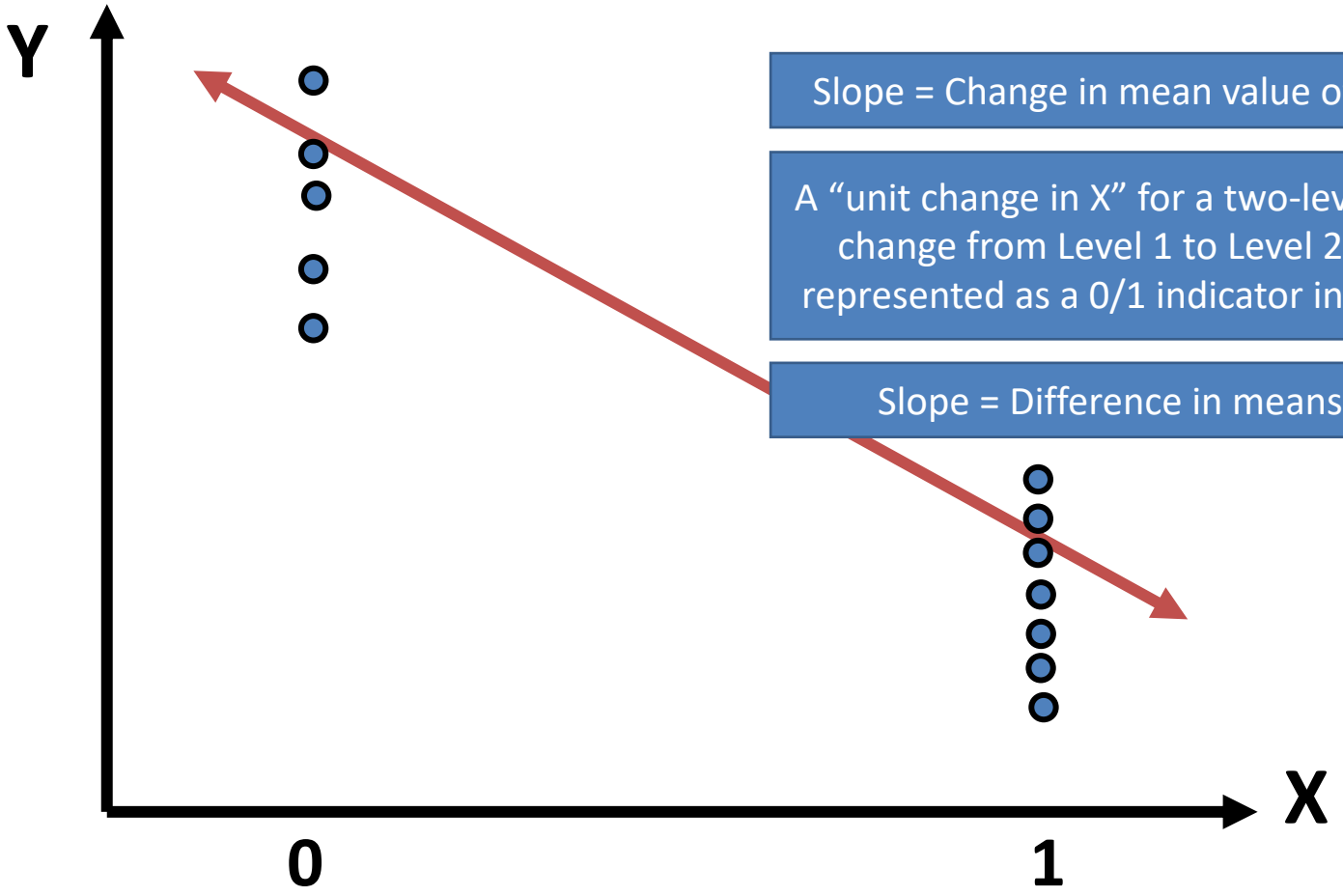
Regression Model: {Mean of Y when X=10} = $\beta_0 + 10 \cdot \beta_1$

- The fitted regression line represents the estimate based on the sample data:

Regression Line: {Mean of Y when X=10} = $b_0 + 10 \cdot b_1$

Does Any of this Change if X is Binary?

- No! Mechanics are all the same!
  - But the slope coefficient β1 has a special interpretation.



Slope = Change in mean value of Y for a unit change in X.

A "unit change in X" for a two-level categorical variable is a change from Level 1 to Level 2 because the variable is represented as a 0/1 indicator in the regression modeling.

Slope = Difference in means when X = 1 vs. X = 0.

Behavioral Risk Factor Surveillance System

Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

BRFSS™

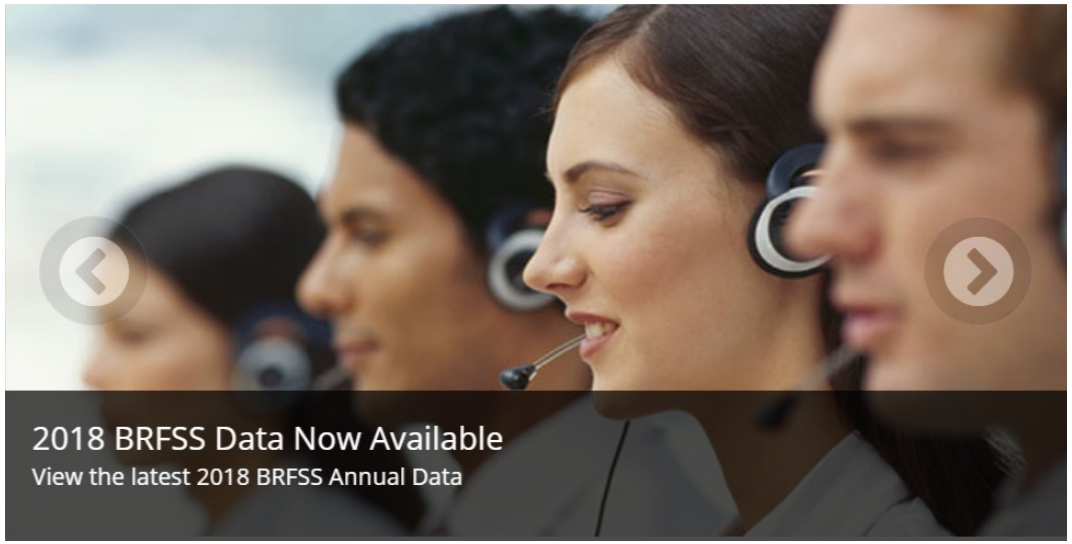


Image source: https://www.cdc.gov/brfss/index.html

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.

Suppose we want to examine the relationship between body mass index (BMI) and the frequency of alcoholic drink consumption (# of days in the last 30days where at least one alcoholic drink was consumed) among low-consumption adults living in NC.

Exercise: BRFSS Data

Question 1: Is there evidence of a relationship between BMI and frequency of drinking? Create a scatterplot to address this question.

Question 2: Is there evidence of a relationship between BMI and frequency of drinking at a significance level of 0.05? Address this question by fitting a linear regression model.

Question 3: Consider the model fit in Question 2. What does the data suggest about the magnitude of the difference in BMI for a 10-day drinker vs. a 20-day drinker in NC?

Exercise: BRFSS Data

Question 4: Consider the model fit in Question 2. Using this model, estimate the mean BMI among all 25-day drinkers in NC.

Question 5: Consider the model fit in Question 2. Can we predict the BMI for a single 25-day drinker in NC using this model? Is this point estimate different from the point estimated in Question 4? Is the precision of this point estimate different?

Exercise: BRFSS Data

Question 6: Consider the model fit in Question 2. Is using this model to estimate the mean BMI among all 25-day drinkers in the US reasonable?

Question 7: Consider the model fit in Question 2. Is using this model to estimate the mean BMI among all 45-day drinkers in NC reasonable?

Questions?

Duke University School of Medicine

CRTP CLINICAL RESEARCH TRAINING PROGRAM