Elementary Statistics – Simple Linear Regression Pt 1

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Kaitlyn Cook (https://www.smith.edu/people/kaitlyn-cook)

Plan for Today

- Simple Linear Regression
    - Why?
    - Interpretation

Warm Up: EDA for two Numerical Variables

Download the Housing_Data.csv from the course website (under the Examples tab).

What is the observational unit in this dataset? What are the variables?

Create a data visualization to summarize the relationship between the LivingSpace and Price variables. (Use Excel, GoogleSheets, or some other tool of your choosing.)

Warm Up: EDA for two Numerical Variables

Download the Housing_Data.csv from the course website (under the Examples tab).

What is the observational unit in this dataset? What are the variables?

Create a data visualization to summarize the relationship between the LivingSpace and Price variables. (Use Excel, GoogleSheets, or some other tool of your choosing.)

Calculate the Pearson Correlation between Price and LivingSpace.

Do you think there is a relationship? What type? Is it strong?

Big Picture

Over the last few days, we've seen visuals and single number summaries for describing the distribution of individual variables, as well as the *form*, *direction*, and *strength* of their relationship with one another.

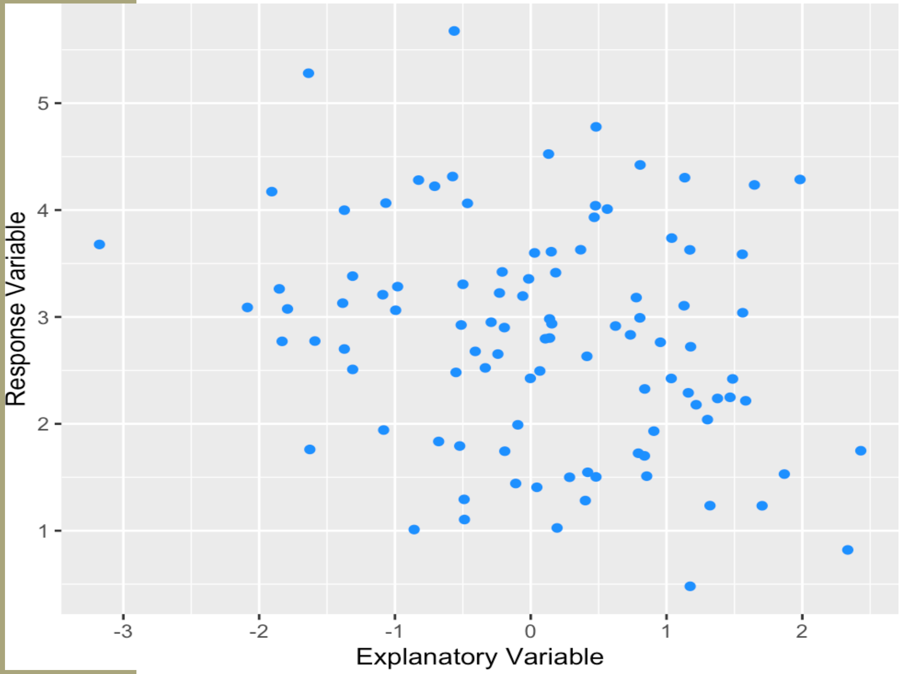
We can use statistical models to capture all of these features in a more precise and quantitative way!

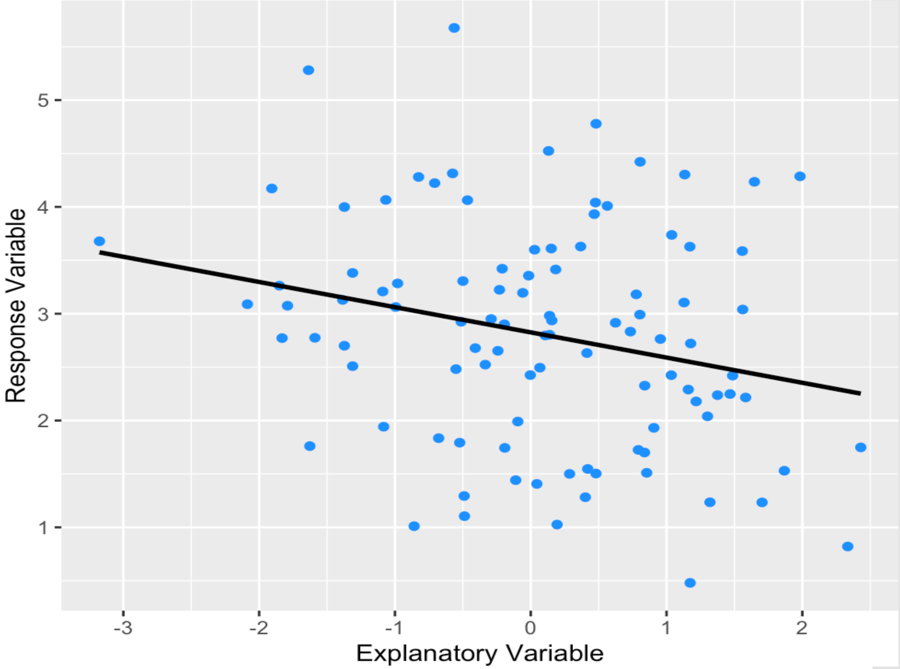
→ In this class, we'll focus on linear regression models, where we seek to model this relationship using a straight line

Big Picture

We can use statistical models to capture *form*, *direction*, and *strength* of relationships between variables in a more precise and quantitative way!

→ In this class, we'll focus on linear regression models, where we seek to model this relationship using a straight line





Motivating Example

Suppose that you manufacture button-down dress shirts for Jackson & Connor (a clothing brand) and want to better understand how their shirts should be sized.

Motivating Example

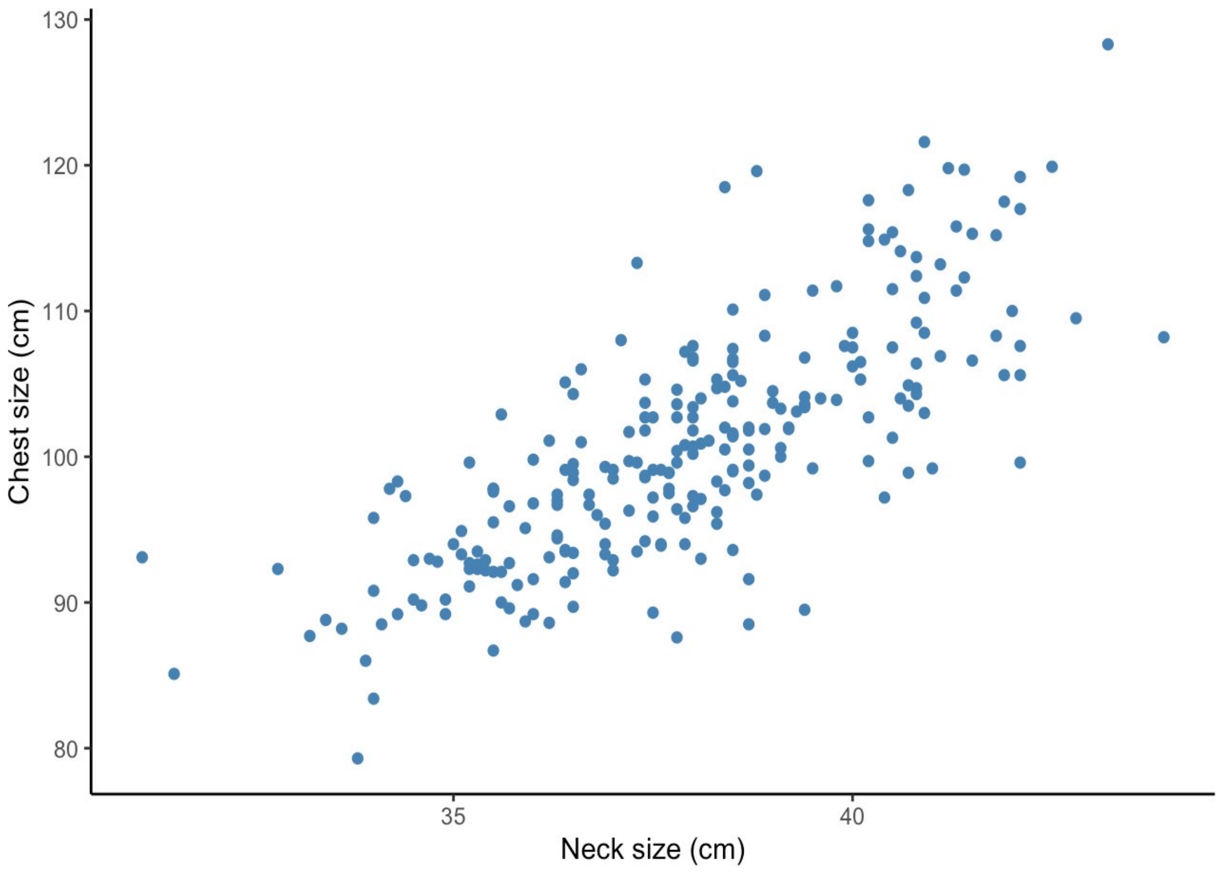
Suppose that you manufacture button-down dress shirts for Jackson & Connor (a clothing brand) and want to better understand how their shirts should be sized.

Then we may wish to build a regression model for the relationship between neck size (cm) and chest size (cm) in order to:

1. Explain and quantify the association between our variables of interest
   - How do an individual's neck and chest size relate with one another?
2. Model or predict the possible values of our response variable
   - Given that an individual has a neck size of 38 cm, what might we expect their chest size to be?

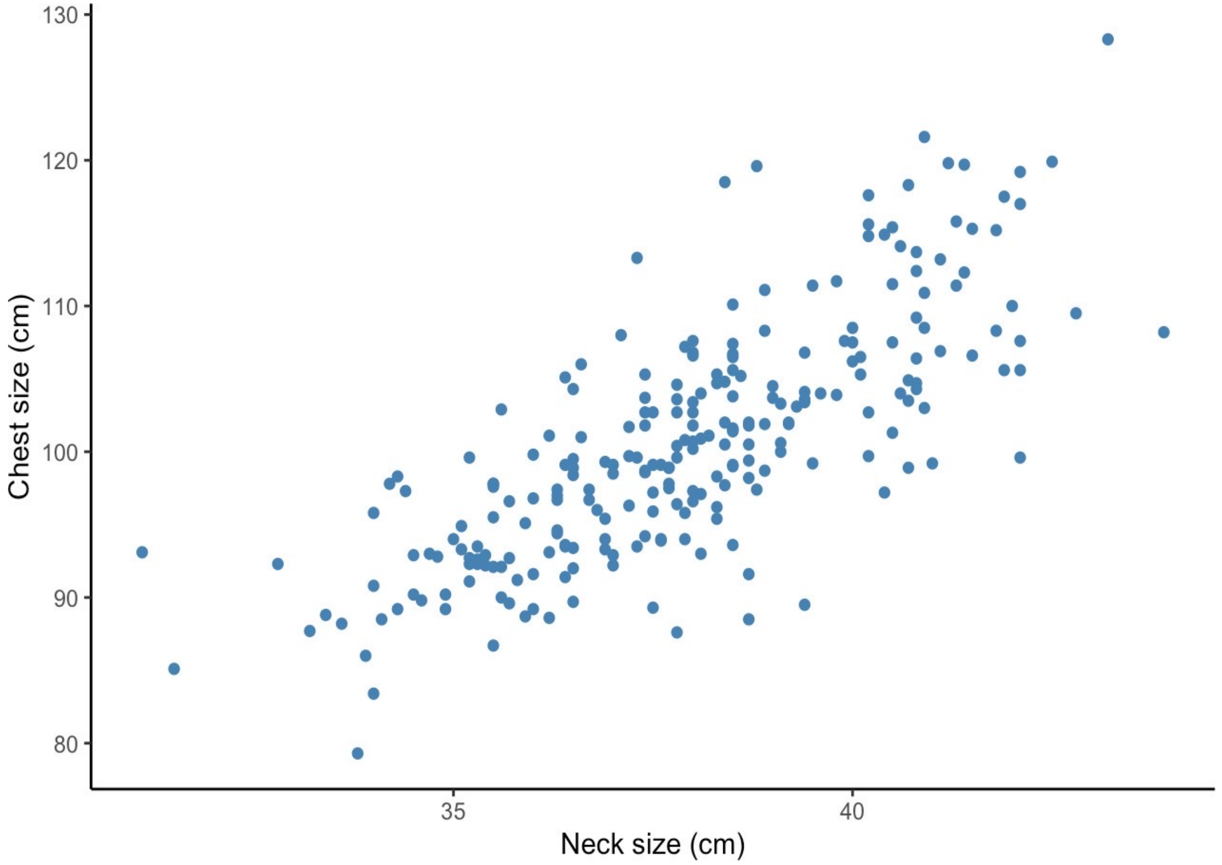
Motivating Example

The following scatterplot displays neck and chest size measurements for 251 individuals:



Motivating Example

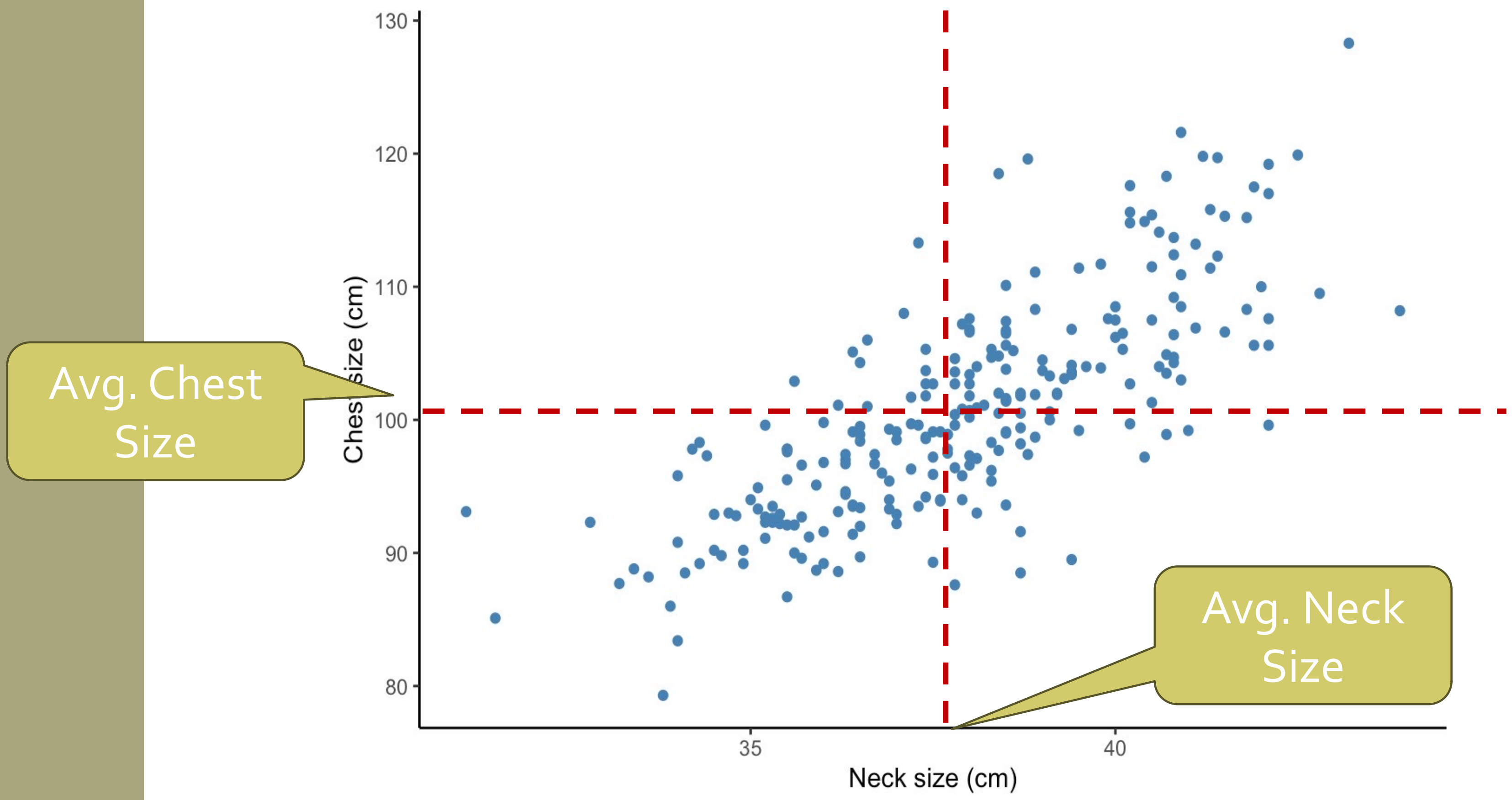
The following scatterplot displays neck and chest size measurements for 251 individuals:



If Jackson & Connor were to offer only one shirt size, they might tailor it to fit an individual of *average* neck size ($\bar{x}$) and *average* chest size ($\bar{y}$). . .

Motivating Example

The following scatterplot displays neck and chest size measurements for 251 individuals:

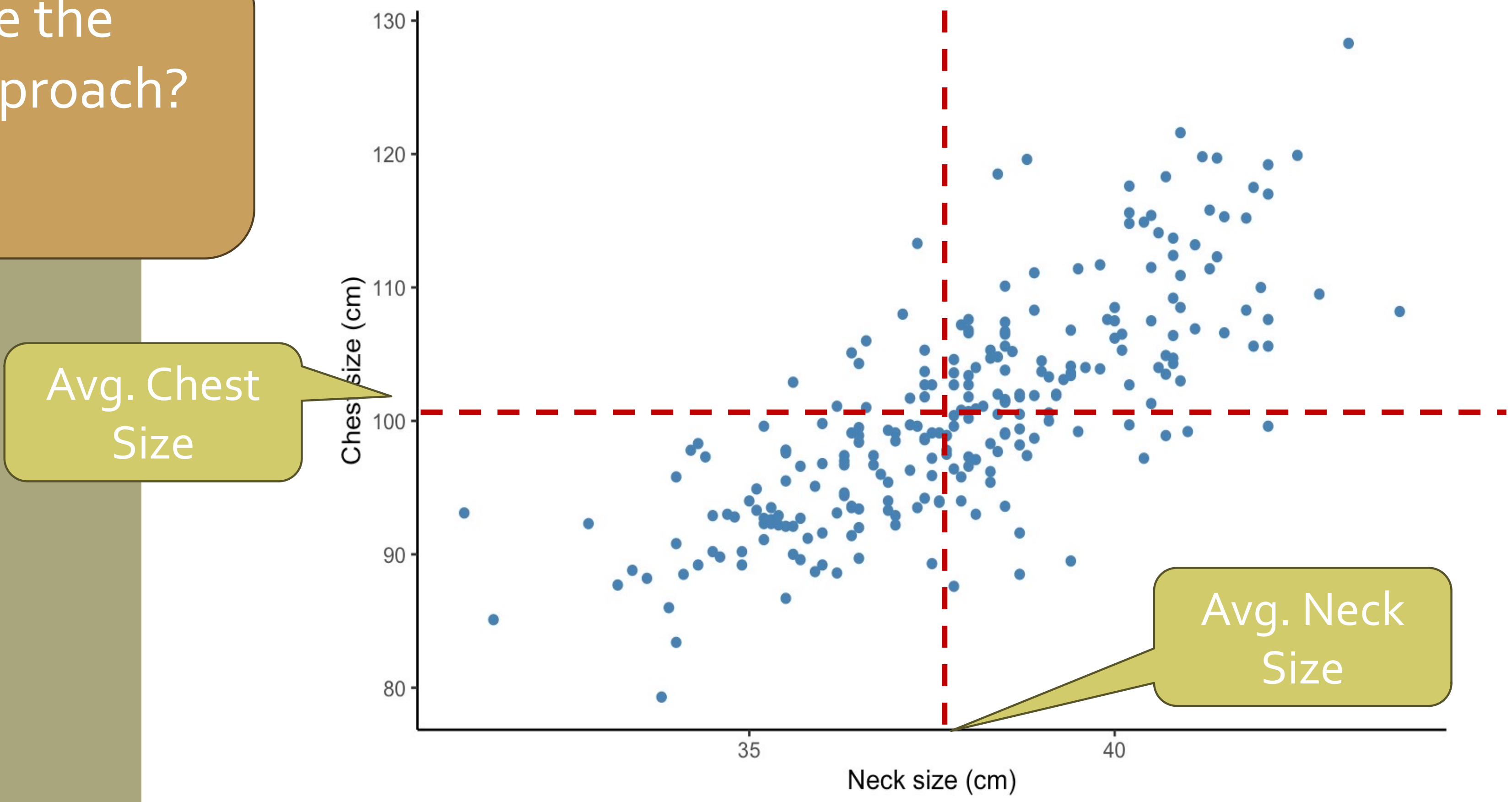


If Jackson & Connor were to offer only one shirt size, they might tailor it to fit an individual of *average* neck size ($\bar{x}$) and *average* chest size ($\bar{y}$). . .

The following scatterplot displays neck and chest size measurements for 251 individuals:

Does this seem like the best marketing approach? Why or why not?

Motivating Example

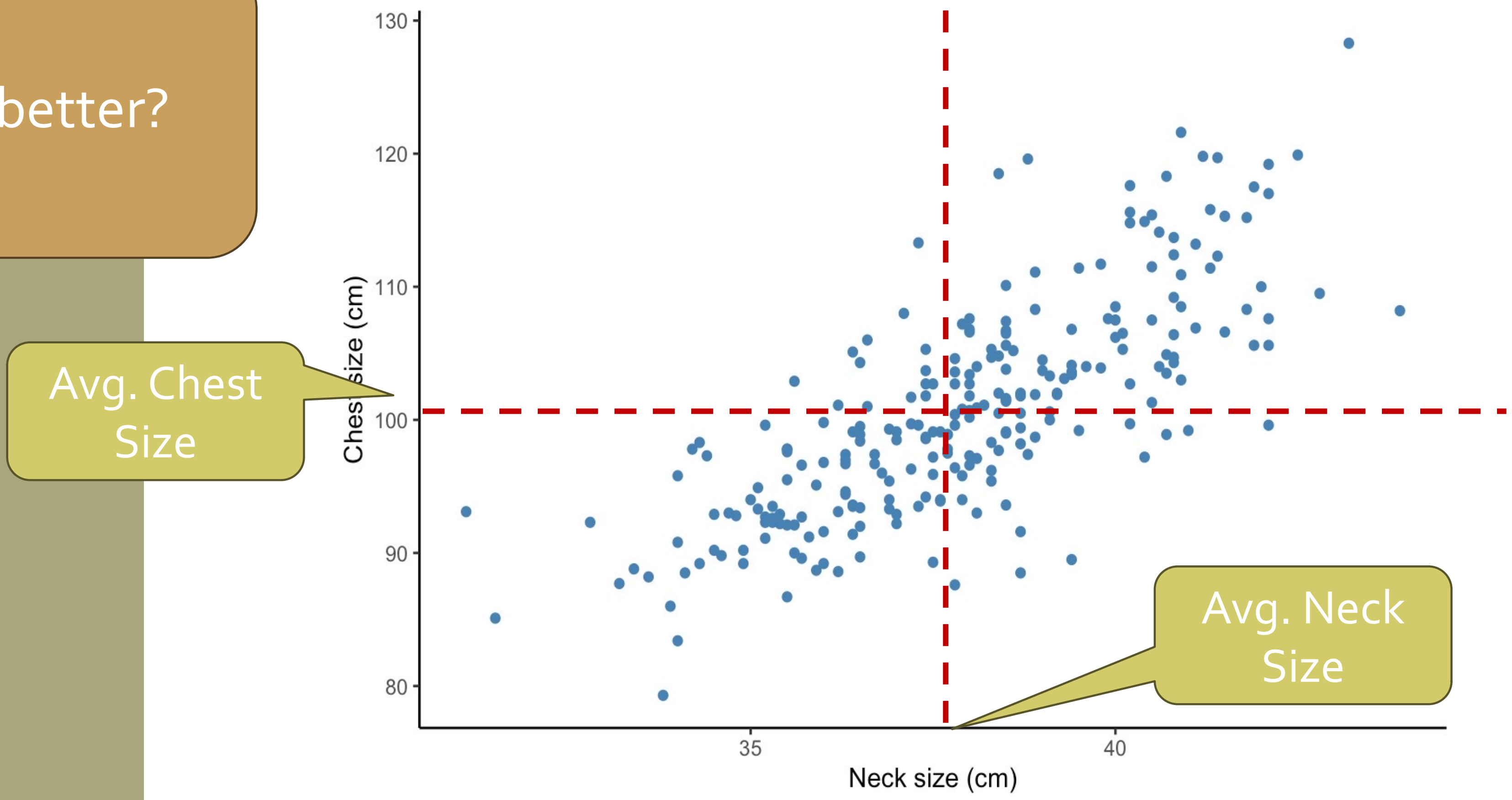


If Jackson & Connor were to offer only one shirt size, they might tailor it to fit an individual of *average* neck size ($\bar{x}$) and *average* chest size ($\bar{y}$). . .

The following scatterplot displays neck and chest size measurements for 251 individuals:

What might work better?

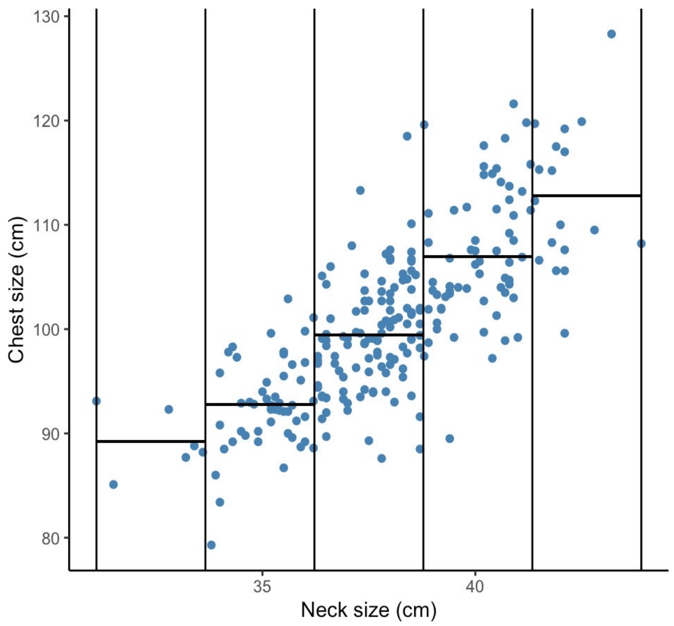
Motivating Example



If Jackson & Connor were to offer only one shirt size, they might tailor it to fit an individual of *average* neck size ($\bar{x}$) and *average* chest size ($\bar{y}$). . .

Motivating Example

In order to improve the fit of their shirts (and to allow them to be worn by more individuals!), Jackson & Connor might want to incorporate more information about how average chest sizes varies as a function of neck size:

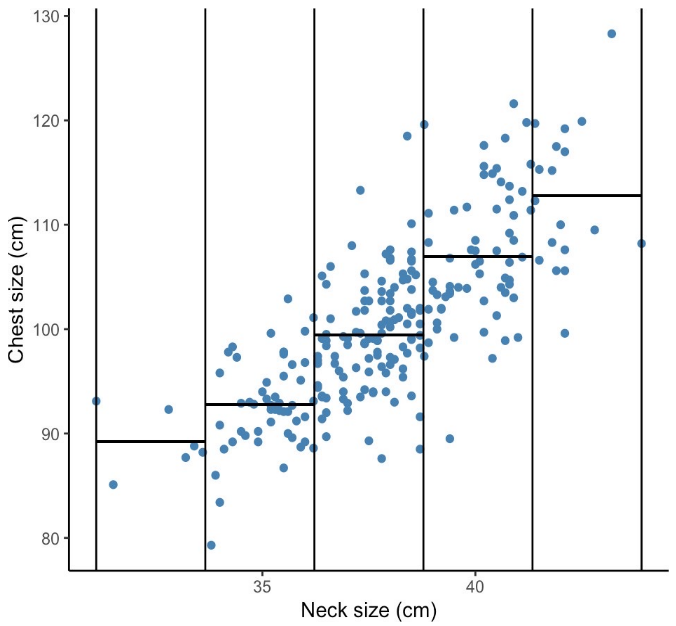


Five shirt sizes:
XS, S, M, L, XL

Is this a better approach? Why or why not?

Motivating Example

In order to improve the fit of their shirts (and to allow them to be worn by more individuals!), Jackson & Connor might want to incorporate more information about how average chest sizes varies as a function of neck size:

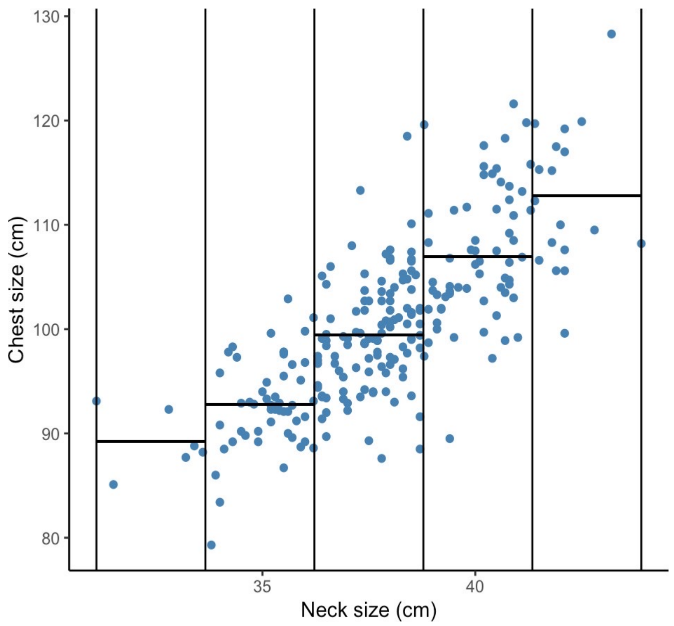


Five shirt sizes:
XS, S, M, L, XL

If you were to improve it, how could you?

Motivating Example

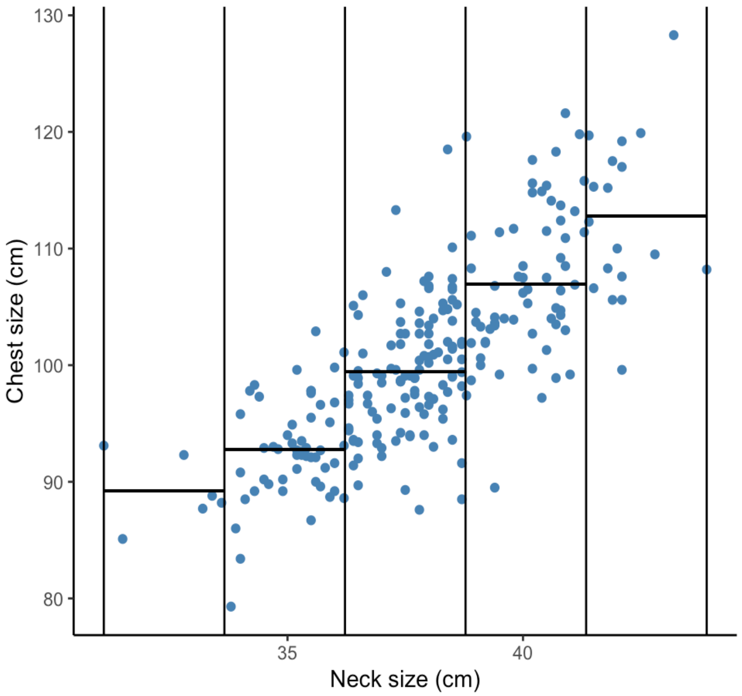
In order to improve the fit of their shirts (and to allow them to be worn by more individuals!), Jackson & Connor might want to incorporate more information about how average chest sizes varies as a function of neck size:



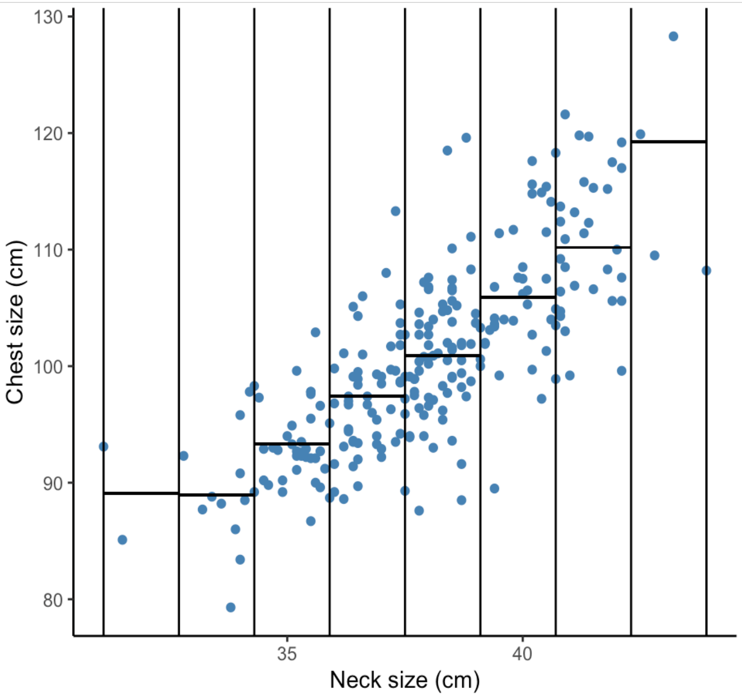
Five shirt sizes:
XS, S, M, L, XL

Motivating Example

In order to improve the fit of their shirts (and to allow them to be worn by more individuals!), Jackson & Connor might want to incorporate more information about how average chest sizes varies as a function of neck size:



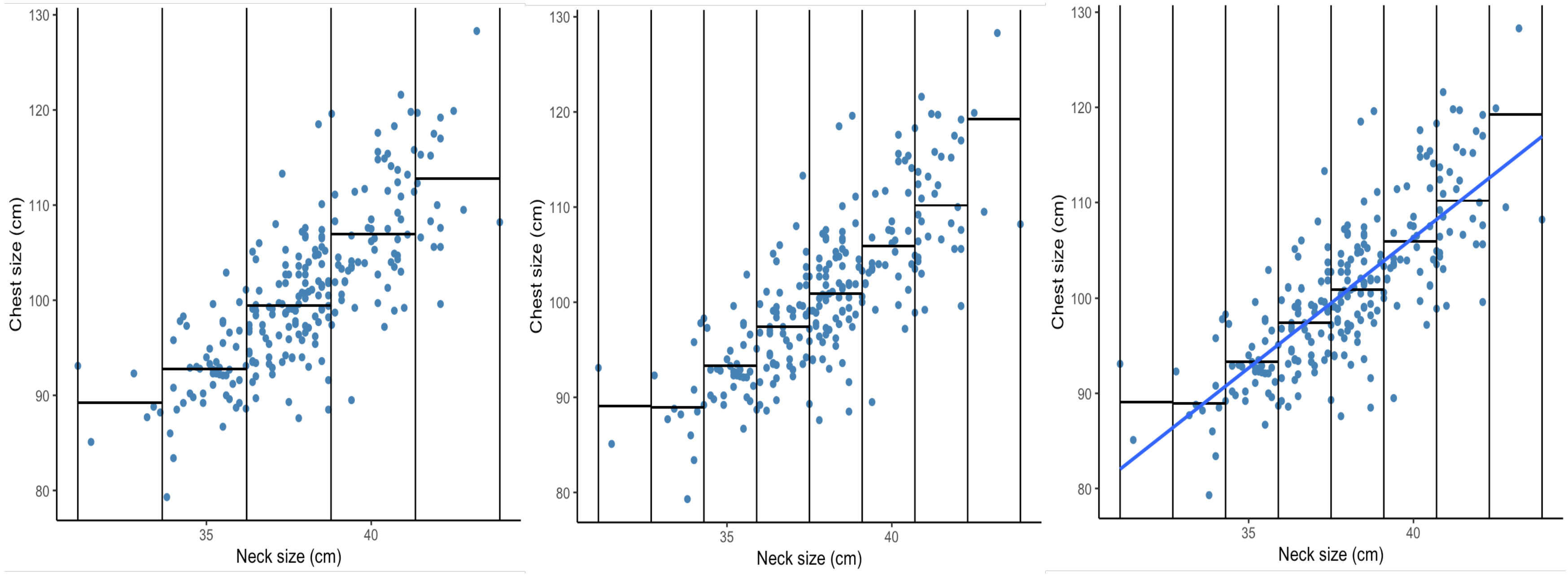
Five shirt sizes:
XS, S, M, L, XL



More sizes (this will fit more people)

Motivating Example

In order to improve the fit of their shirts (and to allow them to be worn by more individuals!), Jackson & Connor might want to incorporate more information about how average chest sizes varies as a function of neck size:



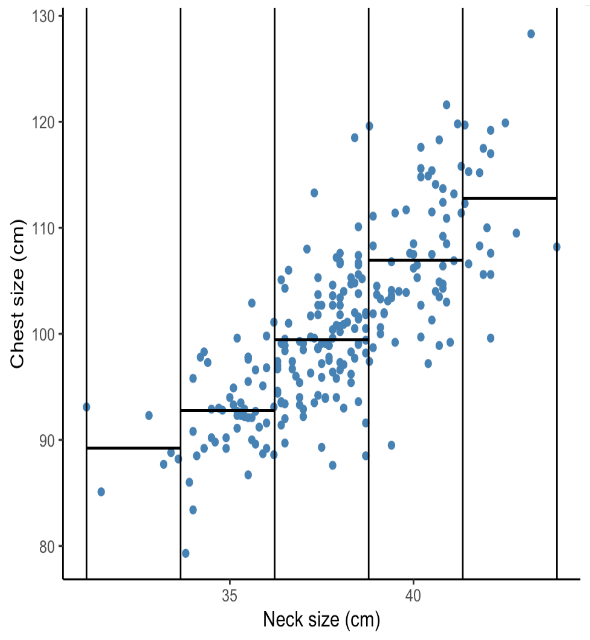
Five shirt sizes: XS, S, M, L, XL

More sizes (this will fit more people)

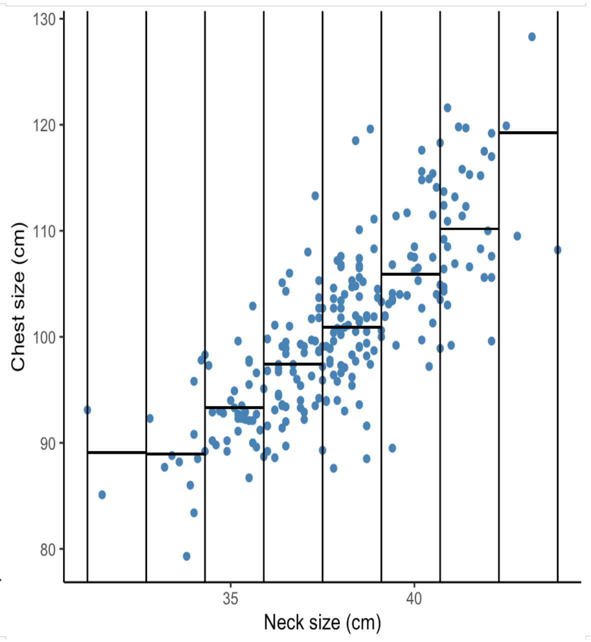
Fit a line for infinite sizes

Motivating Example

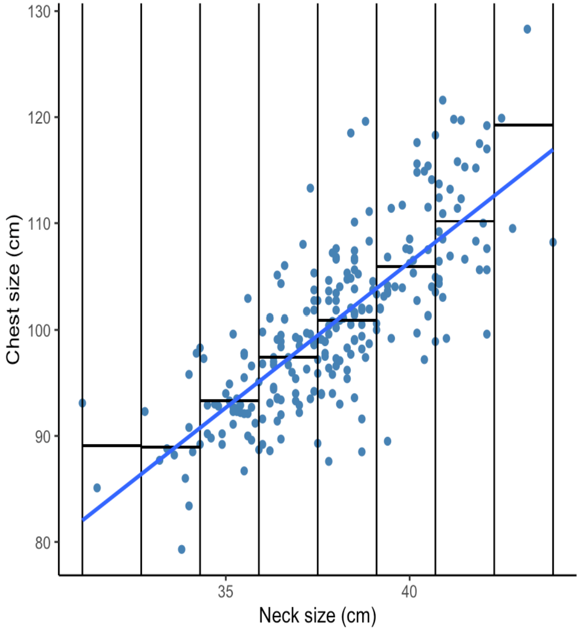
In order to improve the fit of their shirts (and to allow them to be worn by more individuals!), Jackson & Connor might want to incorporate more information about how average chest sizes varies as a function of neck size:



Five shirt sizes: XS, S, M, L, XL



More sizes (this will fit more people)



Fit a line for infinite sizes

Linear Regression!

Population Regression Line

Primary Motivation for Regression Modeling

There is a lot of variability in outcomes (Y) in the world, and we believe that at least some of this variability can be explained by other factors of interest (X)!

Population Regression Line

Primary Motivation for Regression Modeling

There is a lot of variability in outcomes (Y) in the world, and we believe that at least some of this variability can be explained by other factors of interest (X)!

We capture this intuition through the following model:

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

Population Regression Line

Primary Motivation for Regression Modeling

There is a lot of variability in outcomes (Y) in the world, and we believe that at least some of this variability can be explained by other factors of interest (X)!

We capture this intuition through the following model:

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

chest size

Population Regression Line

Primary Motivation for Regression Modeling

There is a lot of variability in outcomes (Y) in the world, and we believe that at least some of this variability can be explained by other factors of interest (X)!

We capture this intuition through the following model:

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{explained by neck size}} + \epsilon_i$$

chest size

Population Regression Line

Primary Motivation for Regression Modeling

There is a lot of variability in outcomes (Y) in the world, and we believe that at least some of this variability can be explained by other factors of interest (X)!

We capture this intuition through the following model:

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{explained by neck size}} + \epsilon_i$$

chest size

explained by everything else (error!)

Population Regression Line

Primary Motivation for Regression Modeling

There is a lot of variability in outcomes (Y) in the world, and we believe that at least some of this variability can be explained by other factors of interest (X)!

We capture this intuition through the following model:

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{explained by neck size}} + \epsilon_i$$

chest size

explained by neck size

explained by everything else (error!)

This line might not perfectly capture everyone in our population, but we assume that—on average—it provides an accurate picture.

Interpreting the Regression Line

Given a linear regression line where Y_i represents an individual's chest size (in cm) and X_i their neck size (in cm):

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

β_0: the intercept term captures the average chest size for the population of individuals who have a neck size of 0 cm

β_1: the slope term captures the expected (average) change in chest size associated with a one centimeter increase in neck size

Interpreting the Regression Line

Given a linear regression line where Y_i represents an individual's chest size (in cm) and X_i their neck size (in cm):

Practice: Suppose the regression is:

$$Y_i = 5 + 3X_i$$

β_0: the intercept term captures the average chest size for the population of individuals who have a neck size of 0 cm

β_1: the slope term captures the expected (average) change in chest size associated with a one centimeter increase in neck size

What is the average chest size for someone with a neck size of 0 cm?

How much would you expect chest size to change with a one cm increase in neck size?

What would the chest size for someone with a neck size of 15 cm be?

Interpreting the Regression Line

Given a linear regression line where Y_i represents an individual's chest size (in cm) and X_i their neck size (in cm):

Practice: Suppose the regression is:

$$Y_i = 2.5X_i$$

β_0: the intercept term captures the average chest size for the population of individuals who have a neck size of 0 cm

β_1: the slope term captures the expected (average) change in chest size associated with a one centimeter increase in neck size

What is the average chest size for someone with a neck size of 0 cm?

How much would you expect chest size to change with a one cm increase in neck size?

What would the chest size for someone with a neck size of 15 cm be?

Estimating the Regression Line

If β_0 and β_1 are population parameters, how do we estimate them?

Estimating the Regression Line

If β_0 and β_1 are population parameters, how do we estimate them?

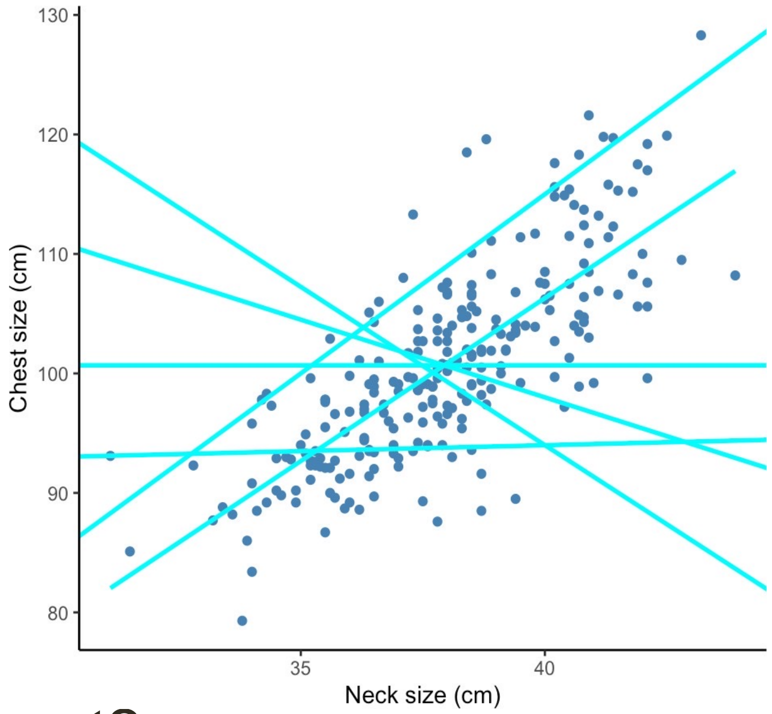
The key insight of linear regression modeling is that we can obtain summary statistics (i.e., “best guesses” or “estimates”’) for β_0 and β_1 by fitting a line to our observed data!

Estimating the Regression Line

If β_0 and β_1 are population parameters, how do we estimate them?

The key insight of linear regression modeling is that we can obtain summary statistics (i.e., "best guesses" or "estimates"') for β_0 and β_1 by fitting a line to our observed data!

However there are many options…



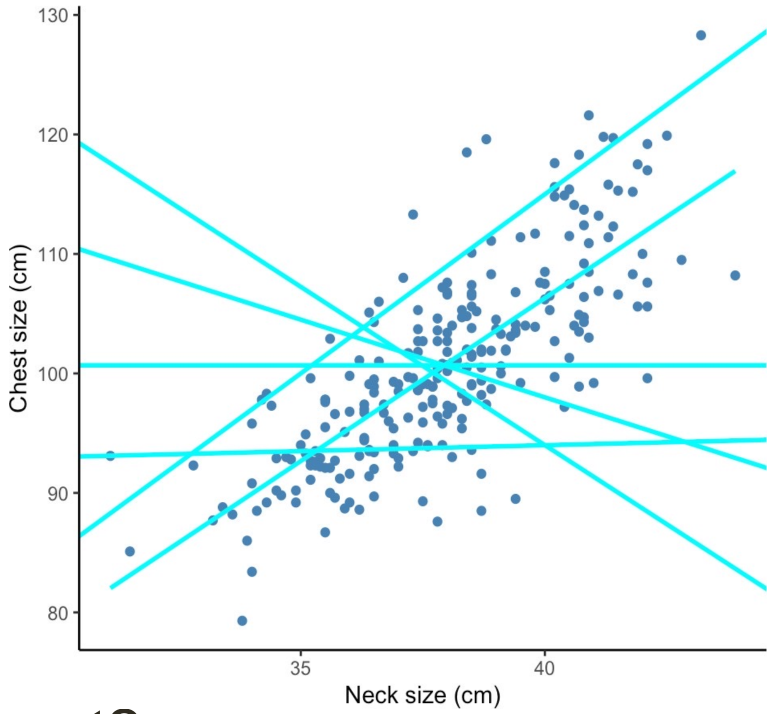
How do we choose the best?

Estimating the Regression Line

If β_0 and β_1 are population parameters, how do we estimate them?

The key insight of linear regression modeling is that we can obtain summary statistics (i.e., "best guesses" or "estimates") for β_0 and β_1 by fitting a line to our observed data!

However there are many options…



Ideas? Which lines look good to you in this chart? Which look bad? Why?

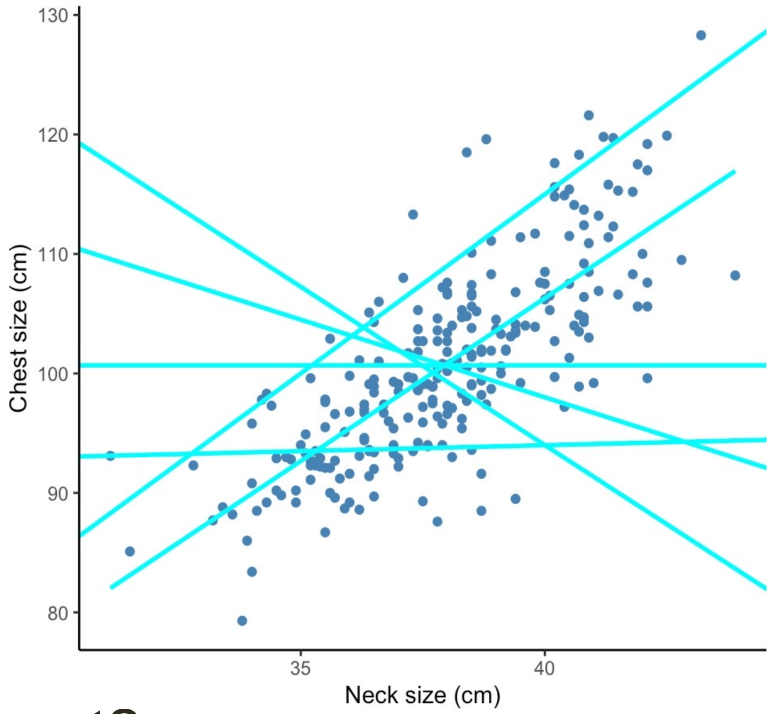
How do we choose the best?

Estimating the Regression Line

If β_0 and β_1 are population parameters, how do we estimate them?

The key insight of linear regression modeling is that we can obtain summary statistics (i.e., "best guesses" or "estimates") for β_0 and β_1 by fitting a line to our observed data!

However there are many options…



Ideas? Which lines look good to you in this chart? Which look bad? Why?

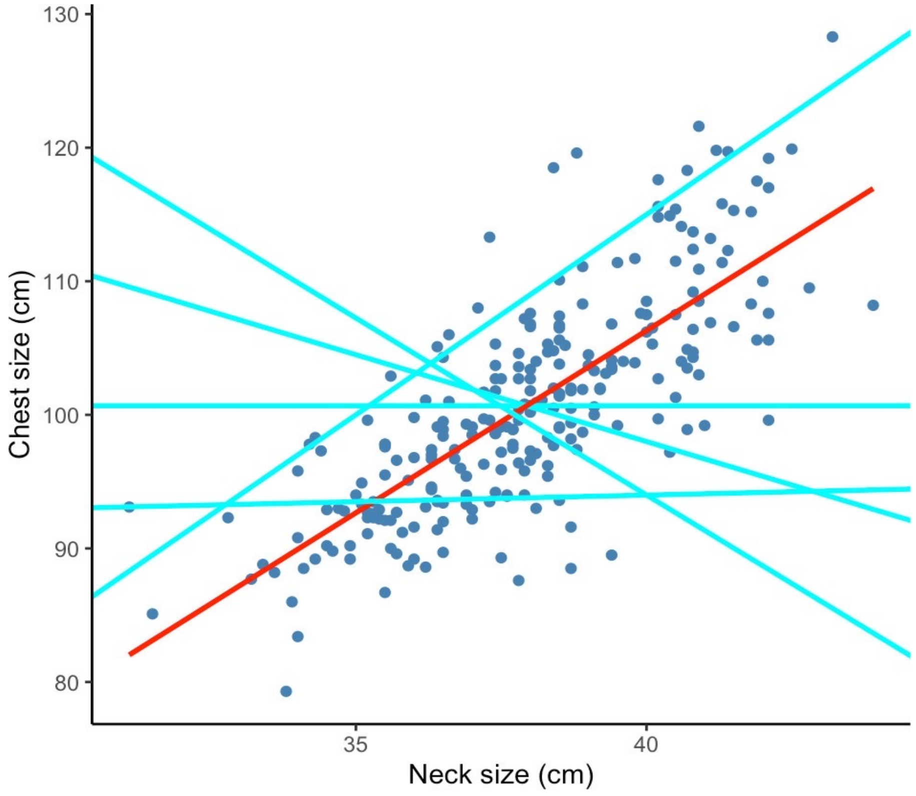
How do we choose the best?

Estimating the Regression Line

If β_0 and β_1 are population parameters, how do we estimate them?

The key insight of linear regression modeling is that we can obtain summary statistics (i.e., “best guesses” or “estimates”’) for β_0 and β_1 by fitting a line to our observed data!

We choose the line that minimizes left over or unexplained variance in Y.



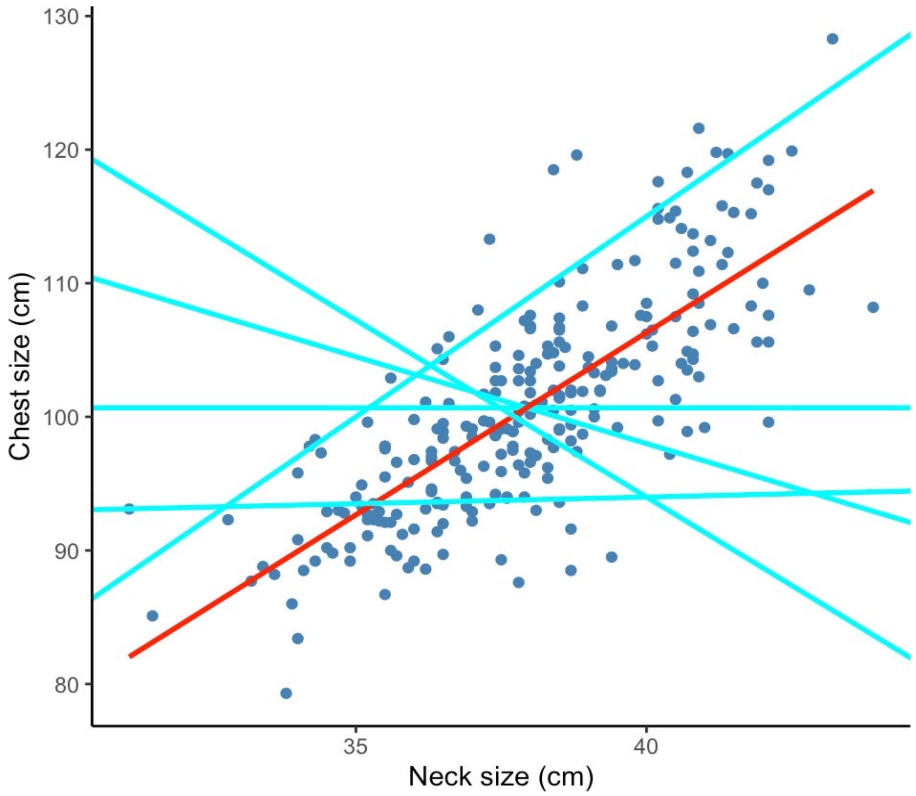
Estimating the Regression Line

If β_0 and β_1 are population parameters, how do we estimate them?

The key insight of linear regression modeling is that we can obtain summary statistics (i.e., “best guesses” or “estimates””) for β_0 and β_1 by fitting a line to our observed data!

We choose the line that minimizes left over or unexplained variance in Y.

This is called the least squares line.



Estimating the Regression Line

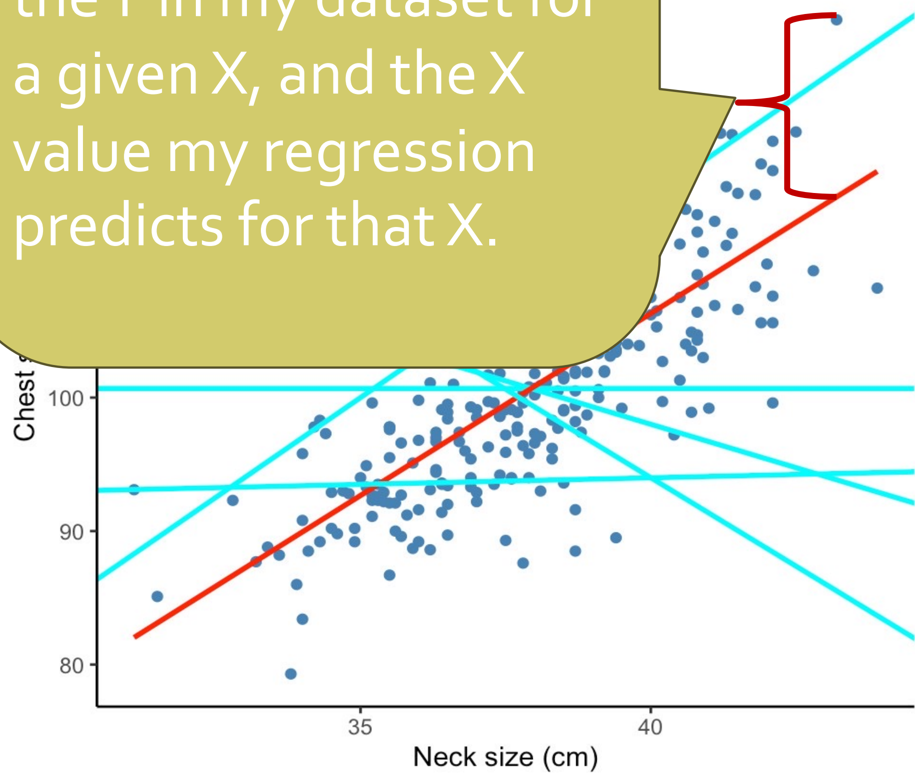
If β_0 and β_1 are population parameters, how do we estimate them?

The key insight of linear … hat we can obtain summary statistic … "estimates"") for β_0 and β_1 by fitting a … !

Error or ***residual*** is the difference between the Y in my dataset for a given X, and the X value my regression predicts for that X.

We choose the line that minimizes left over or unexplained variance in Y.

This is called the least squares line.



If β_0 and β_1 are population parameters, how do we estimate them?

The key insight of linear hat we can obtain summary statistic “estimates”’) for β_0 and β_1 by fitting a !

Estimating the Re Lir

The big idea behind least squares is that we *minimize* the value we get when we square and sum residuals for every observation in the data.

ne that er or ance

Error or ***residual*** is the difference between the Y in my dataset for a given X, and the X value my regression predicts for that X.

This is called the least squares line.