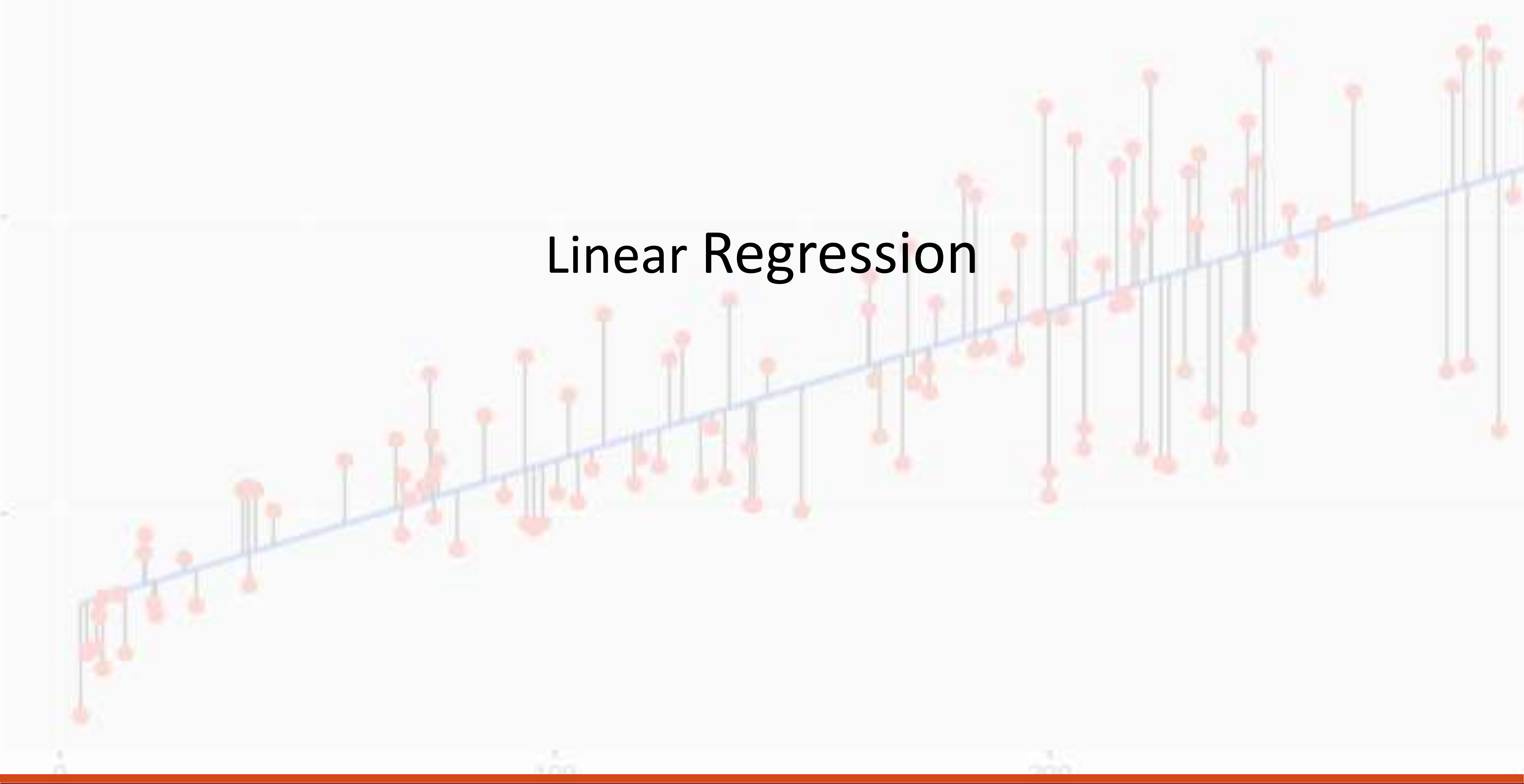
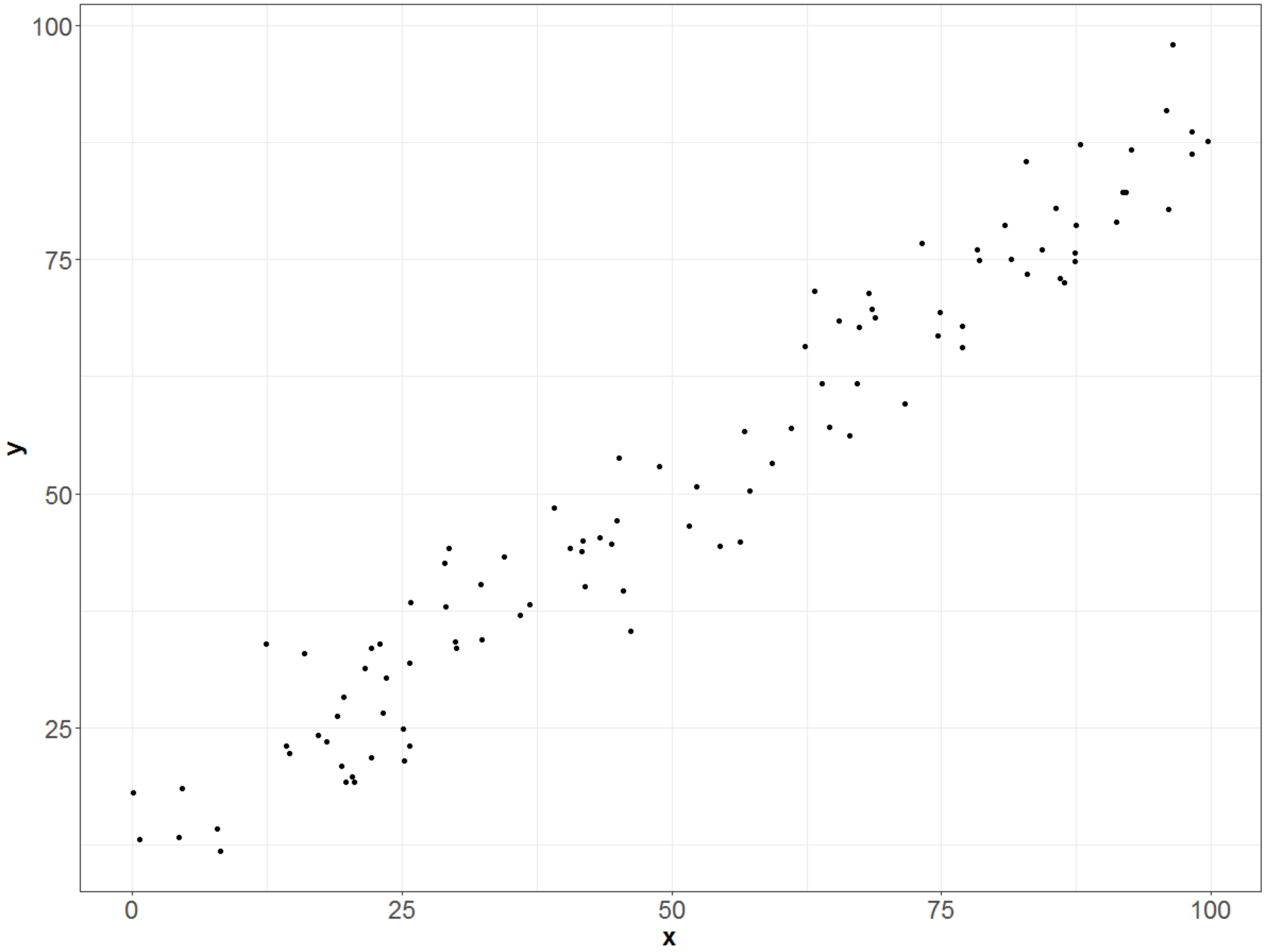
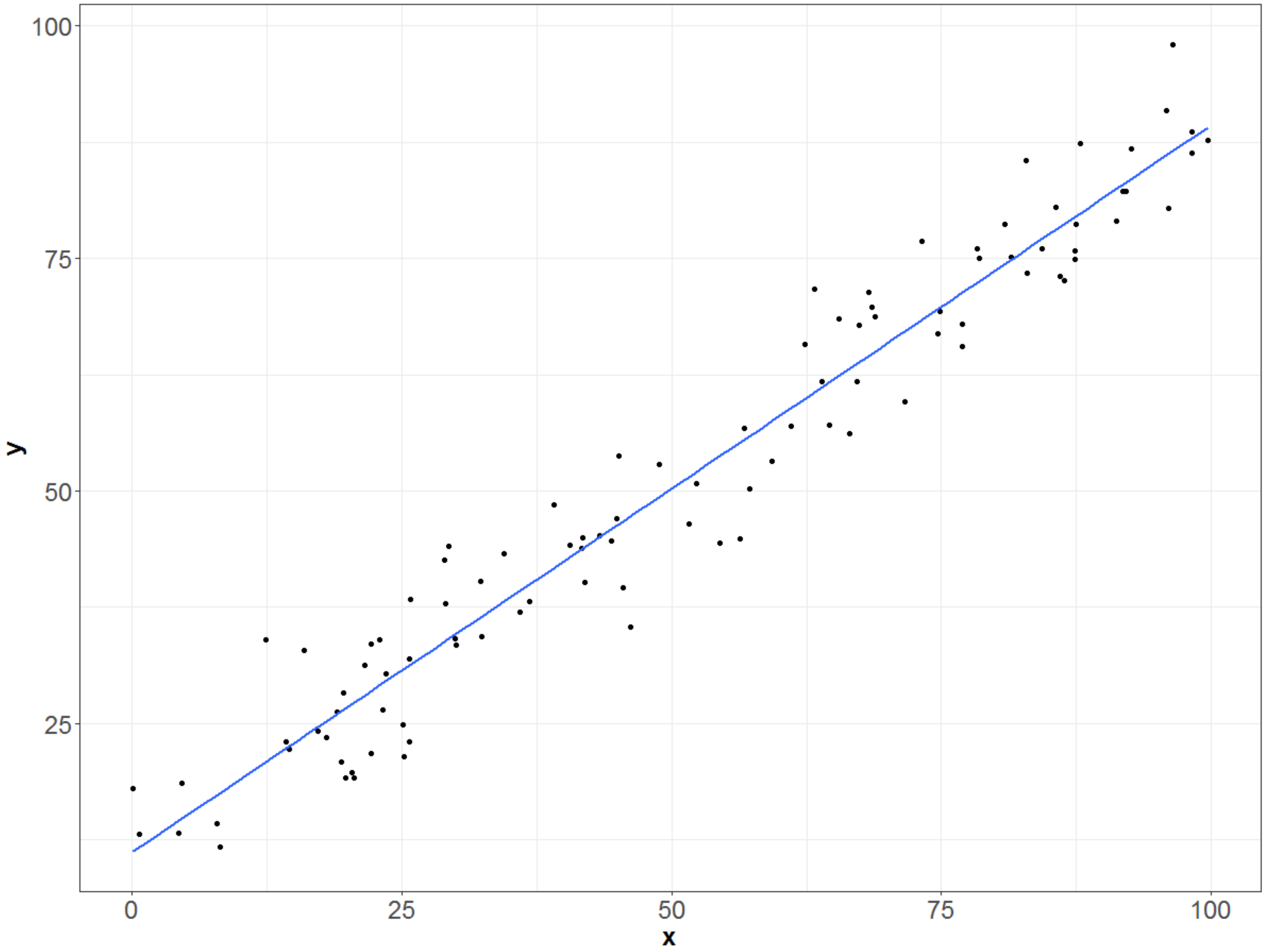
Linear Regression



Regression Analysis

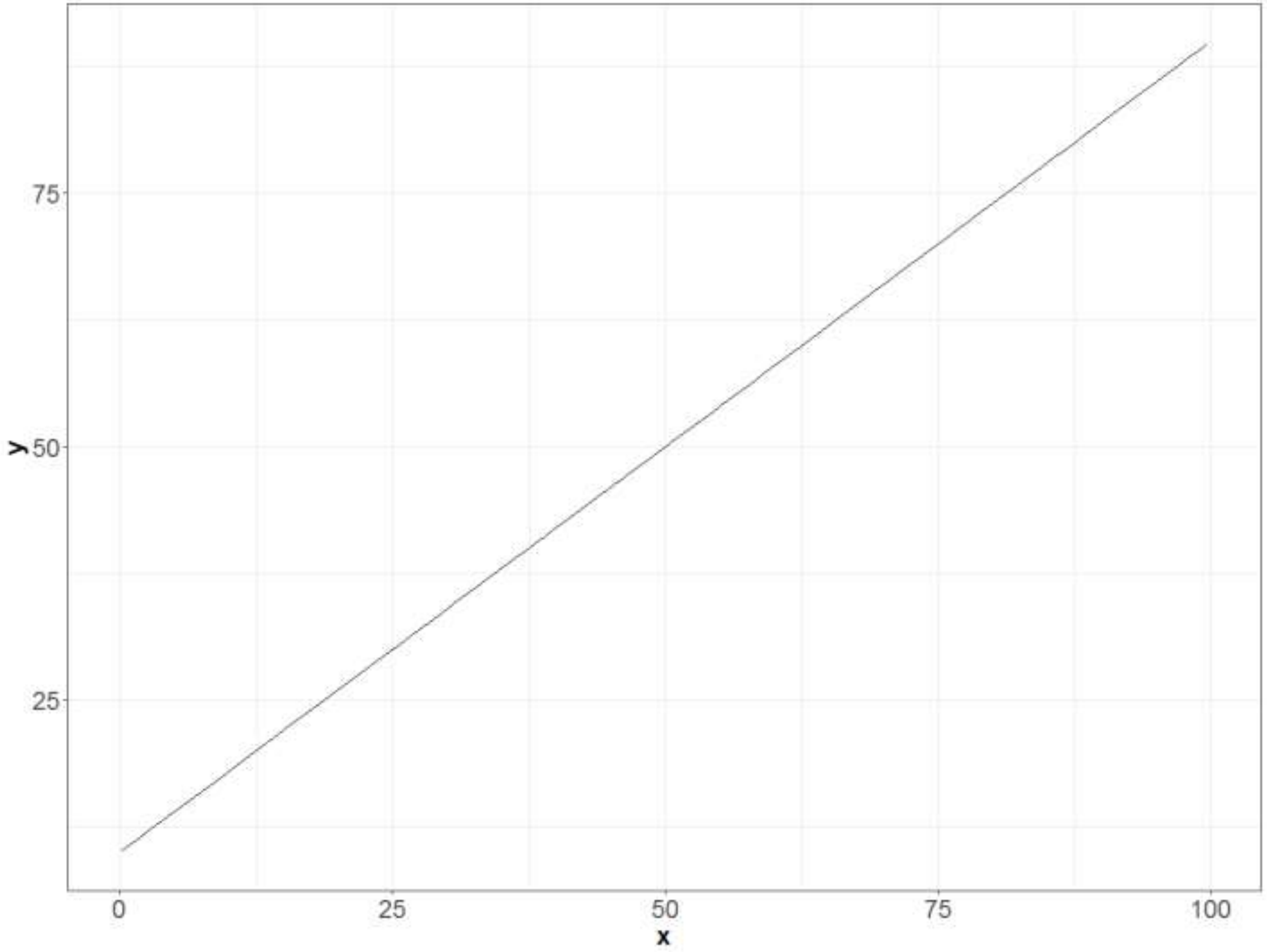


Regression Analysis = drawing the “best” line through data



The equation of a line

$$y = mx + b$$

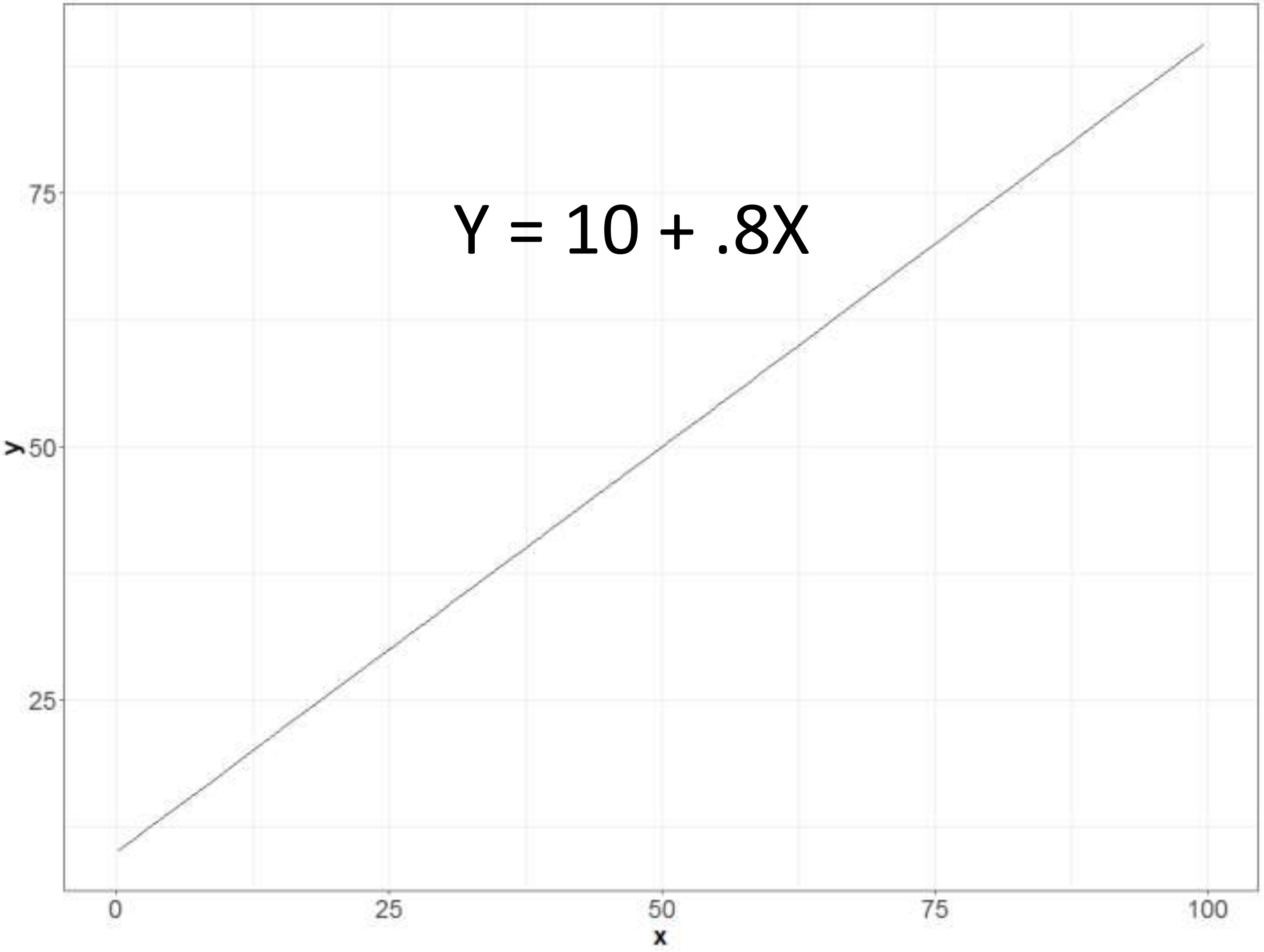


The equation of a line

Rearranging:

$$y = b + mx$$

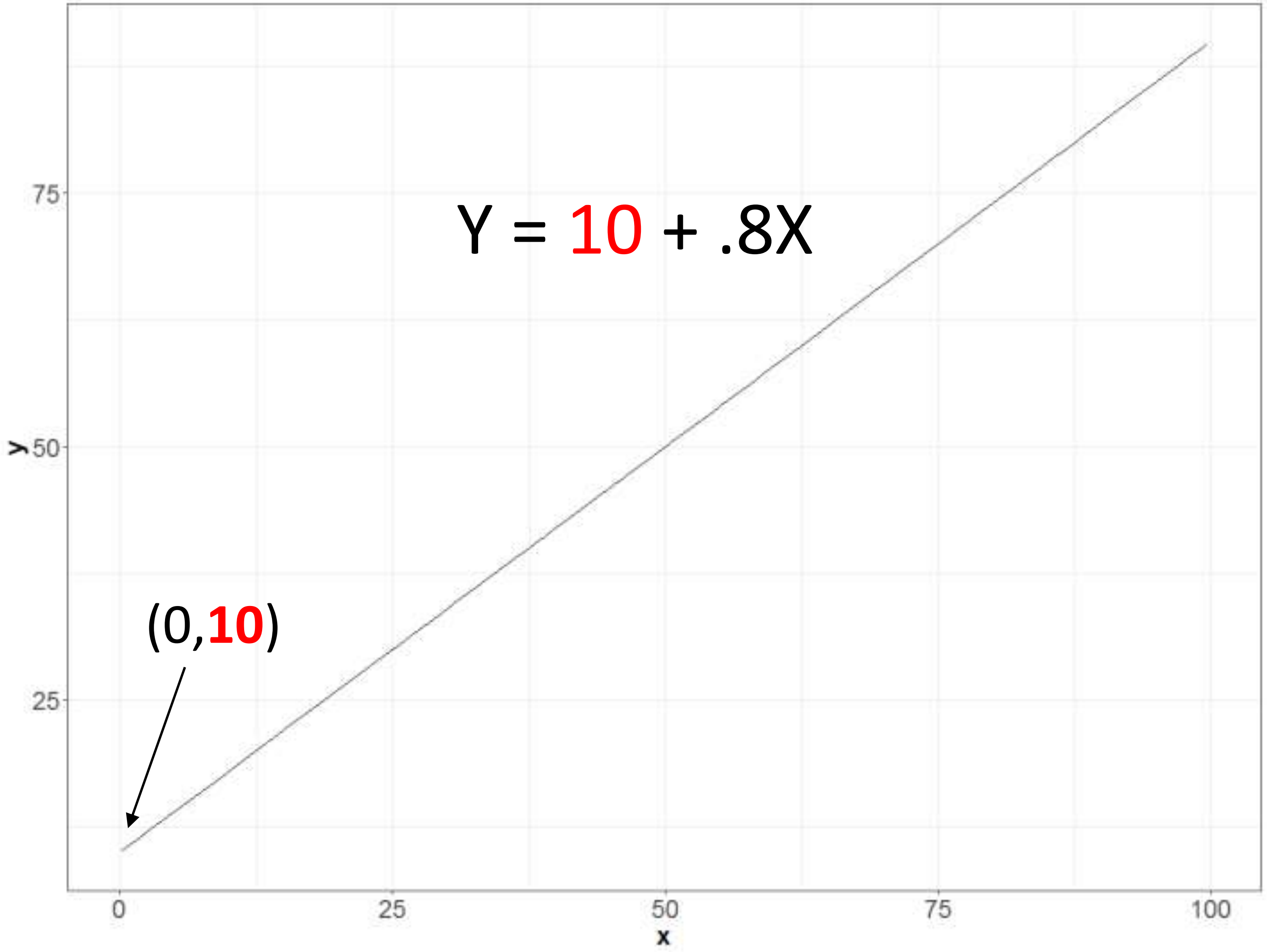
- b is the y-intercept
- m is the slope



The equation of a line

$$y = b + mx$$

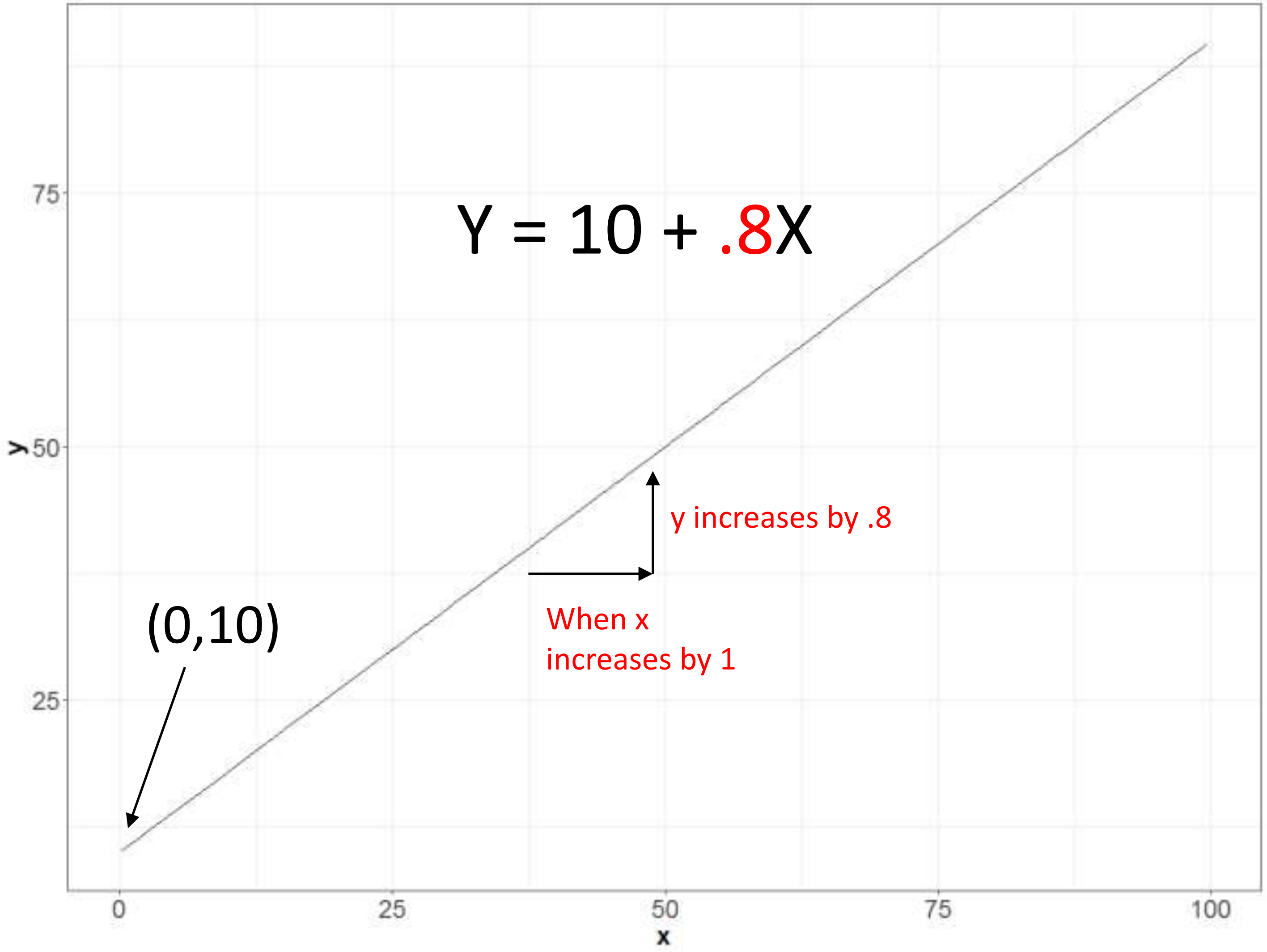
- **b** is the y-intercept
- m is the slope



The equation of a line

$$y = b + mx$$

- b is the y-intercept
- **m** is the slope



Linear Regression

Simple Linear Regression

- Our goal is to use the data to estimate the “best” line through the data- specifically the intercept and the slope.

- In linear regression, we typically use the symbol β for parameters (values of the slope and intercept)

- So we try to estimate the following β_0 and β_1:

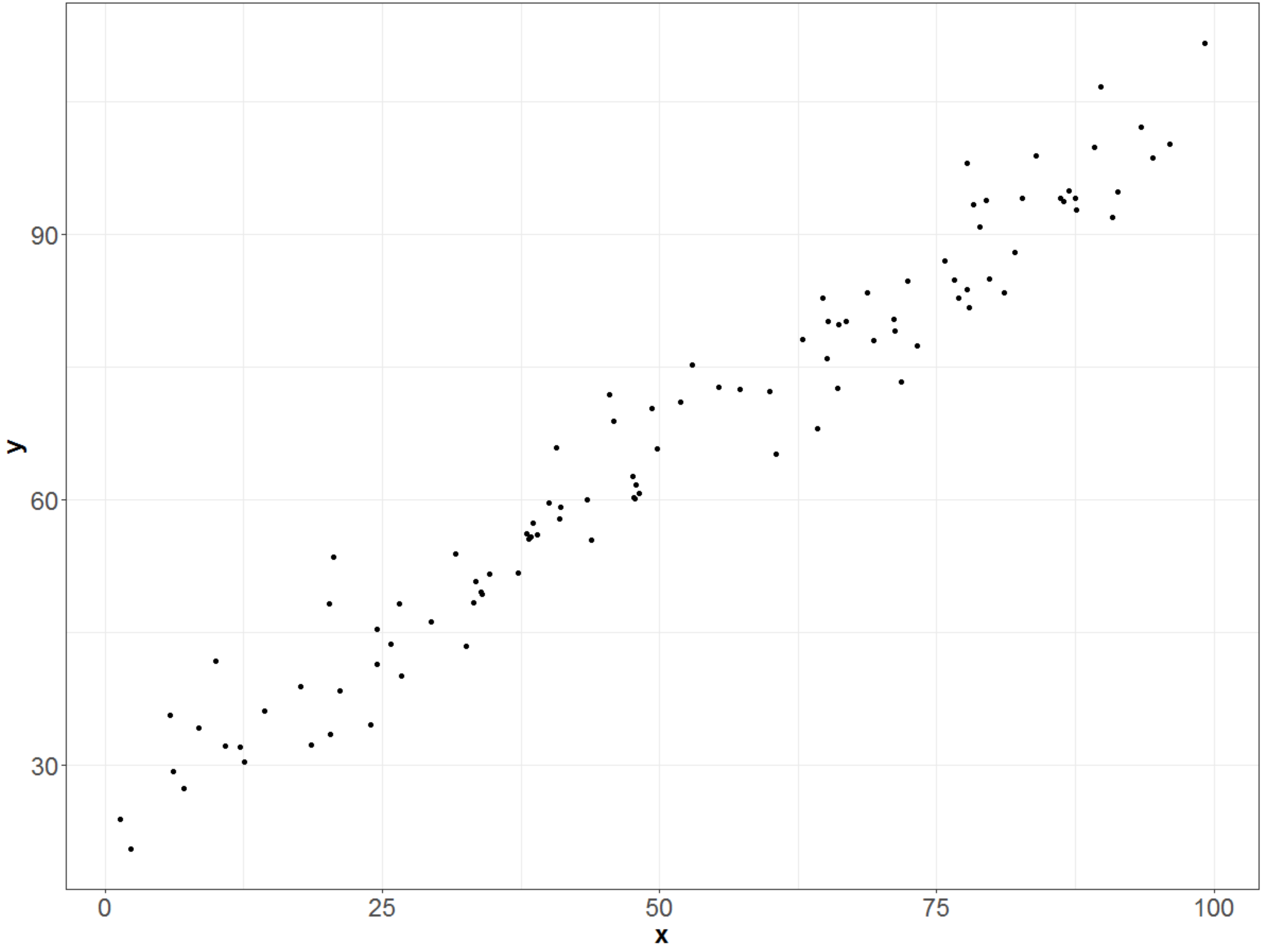
$$Y = \beta_0 + \beta_1 X$$

  - β_0 is the y intercept

  - β_1 is the slope

- However, we cannot find a slope and intercept that perfectly fits the data.

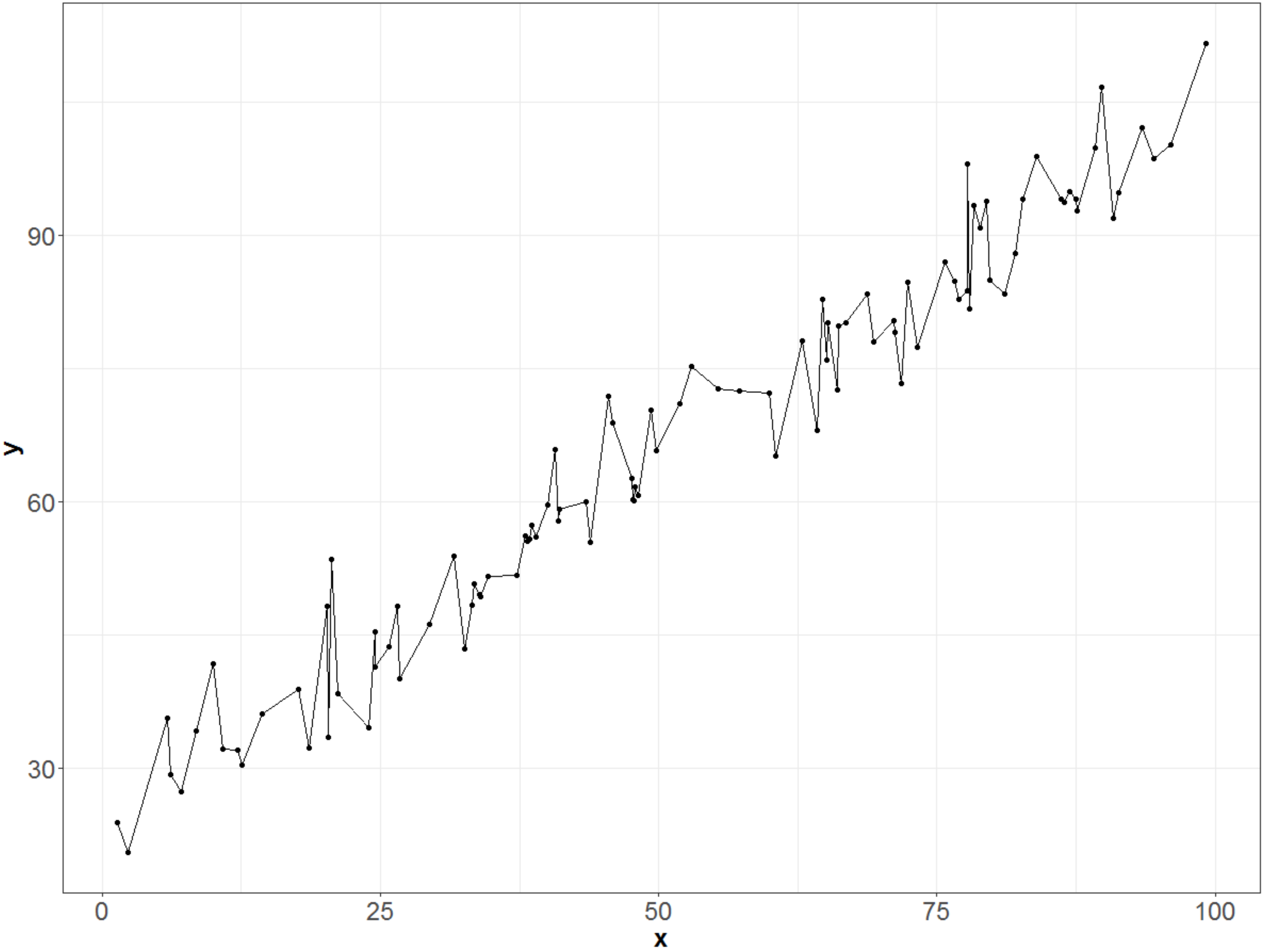
- For example, the slopes between any two observations are different.



- However, we cannot find a slope and intercept that perfectly fits the data.

- For example, the slopes between any two observations are different.

- So we do the best we can at drawing ONE line through the data…

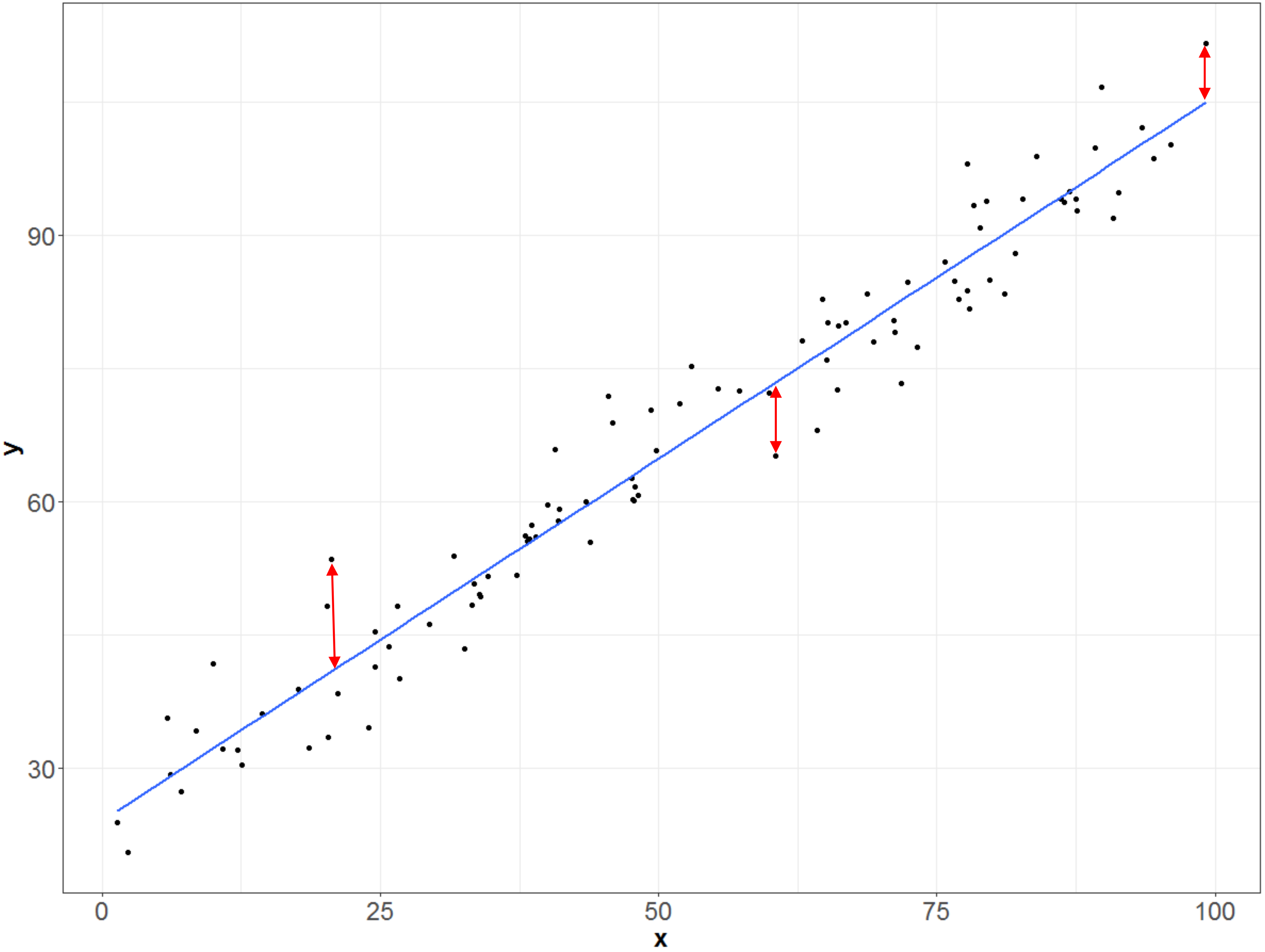


- However, we cannot find a slope and intercept that perfectly fits the data.

- For example, the slopes between any two observations are different.

- So we do the best we can at drawing ONE line through the data…

- …and the differences between the data and our line are called *residuals*



Linear Regression

Simple Linear Regression Model:

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

- β_0 is the y-intercept
- β_1 is the slope
- ε_i is the residual for observation *i=1,2,…N*
- X_i is the independent variable
- Y_i is the dependent variable

Which Variable do I Make Y, and Which Variable do I make X?

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

- You will need to carefully read the question to find key words that tell you which is the dependent variable (*Y*) and which is the independent variable *(X).*
- A variable that is being "predicted", "explained", "affected", "impacted", etc., is the dependent variable.
- On the other hand, the variable that does the predicting, explaining, affecting, etc. is the independent variable.

(Ex) A professor wants to know how well studying predicts test scores. What is the dependent and independent variable?

The dependent variable is **test scores**, the independent variable is **studying**.

What do β_0 and β_1 tell us?

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

- They describe the *relationship* between the independent and dependent variables.

(Ex) Suppose Y is annual sales, and X is customers.

- If the number of customers increases by 1, $\boldsymbol{\beta_1}$ (the slope coefficient) tells us how much annual sales will change.

- $\boldsymbol{\beta_0}$, the y-intercept coefficient, tells us the annual sales when X is exactly zero.
  - Although this can sometimes be interesting, we are usually more interested in the slope, β_1.

Practice

Suppose you estimate the following relationship between a manager's salary (in thousands) and their age (in years):

$$\widehat{salary} = 48.4 + 5.2\ Age$$

- What is the interpretation of the coefficient on Age?
- What is the interpretation of the y-intercept?

Linear Regression

Simple Linear Regression Model:

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

- Note that β_0 and β_1 are *population* parameters.
- The represent the real relationship between X and Y.
- But just like with hypothesis testing, we typically only have a *sample* of data, so we use this to estimate the population parameters.
- The estimates of β_0 and β_1 are typically denoted b_0 and b_1 .

Linear Regression

How do we define the "best" line through the data?

- Intuitively, we would want to make the residuals as small as possible.

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

- Solve for the residuals:

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

- Add up the residuals for every observation:

$$\sum_{i=1}^{N} \varepsilon_i = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_i)$$

- Note that some residuals are negative and some are positive

- So when we add them together, the negative values partially offset the positive values

- This is bad because it will underestimate the total distance between the errors and the line

- To fix this problem, we *square* the residuals so they're all positive

Linear Regression

How do we define the “best” line through the data?

$$\sum_{i=1}^{N} \varepsilon_i = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_i)$$

Linear Regression

How do we define the “best” line through the data?

$$\sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- This equation gives you the **sum of the squared residuals,** or **SSR**.
- Our goal is to minimize this value.
- Since Y_i and X_i are data that we've observed (i.e. they can't be changed), we can only adjust β_o and β_1 to achieve this goal.
- This is called the Least Squares Method of estimating β_o and β_1 .

Linear Regression

Estimation

- In practice, we use calculus to choose b_0 and b_1 to minimize the sum of the squared residuals.
- But for simple (one-variable) regression, there is an easy formula for the slope b_1:

$$\boxed{b_1 = \frac{Cov(X,Y)}{Var(x)}}$$

- After we compute the slope, we can use it to solve for the intercept with the following formula:

$$\boxed{b_0 = \bar{Y} - b_1\bar{X}}$$

  - $\bar{Y}$ is the mean of Y
  - $\bar{X}$ is the mean of X

Suppose you want to know how the number of customers near your store effects annual sales. You decide to use simple linear regression.

- Annual sales (in millions) is your dependent variable, Y.

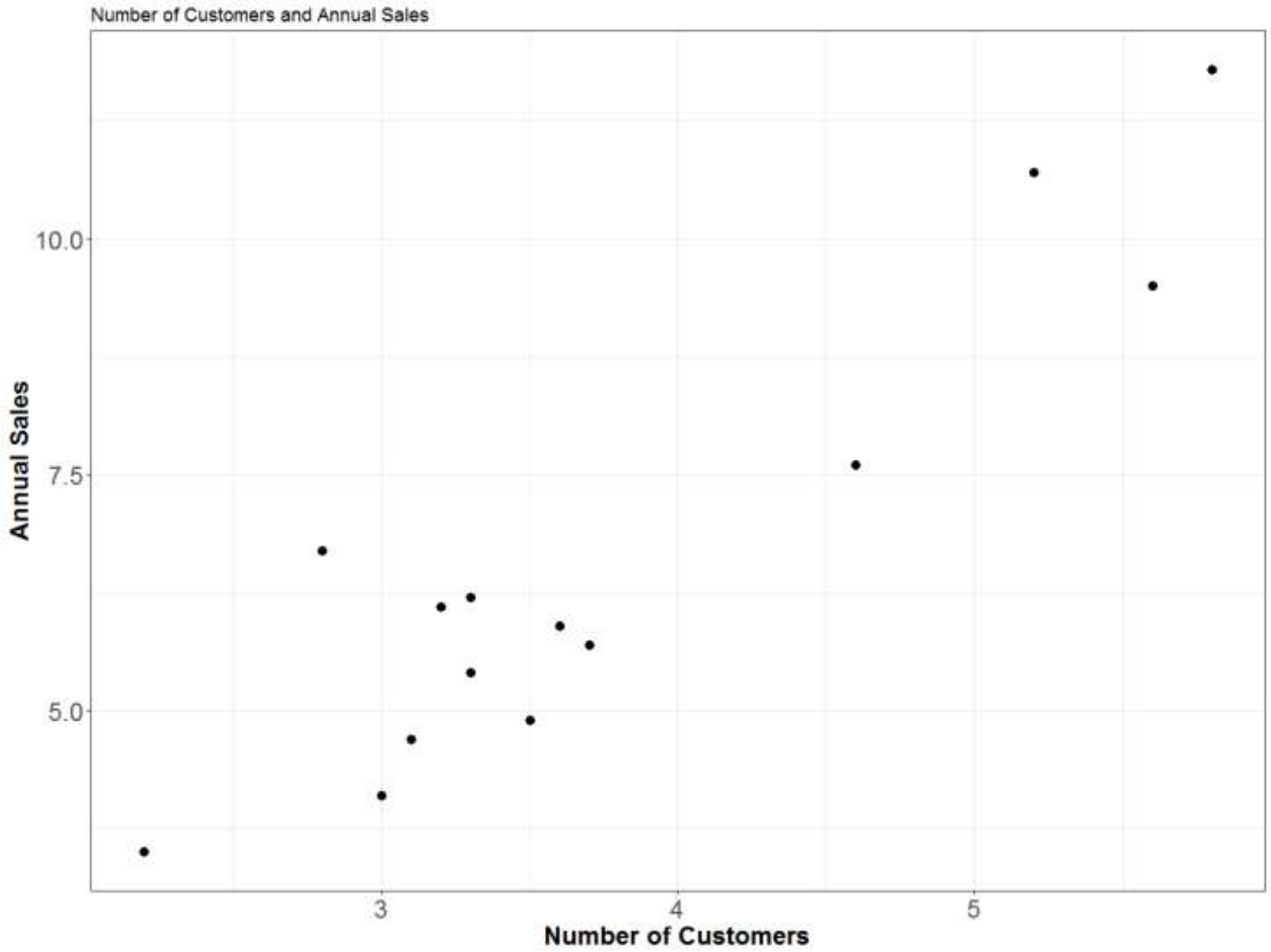
- Number of customers (in millions) is your independent variable, X.

Estimate b_1 and b_0

- $b_1 = 2.07$
- $b_0 = -1.21$

Our estimated regression line is:

$$\boxed{\widehat{Y}_i = -1.21 + 2.07X_i}$$



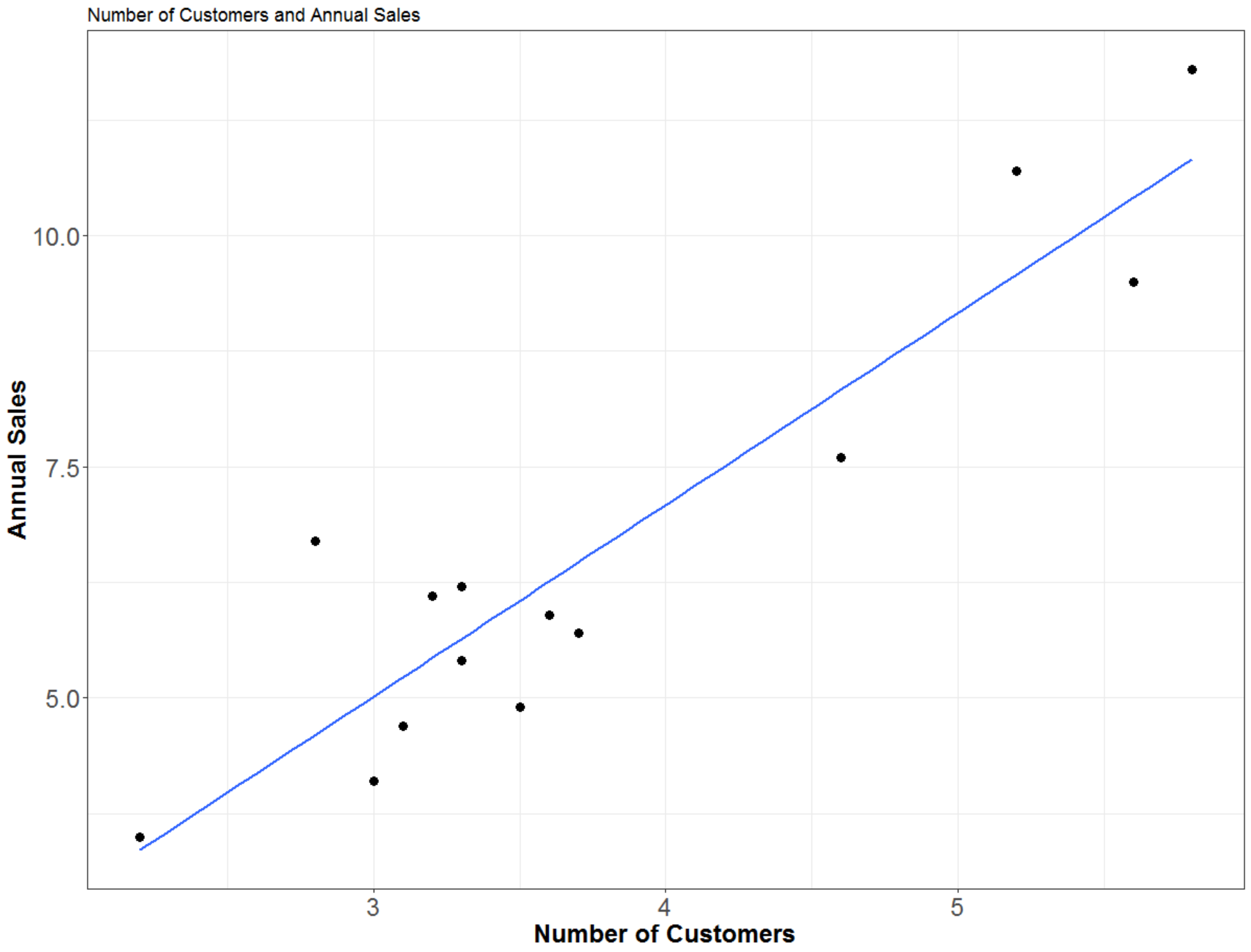
Estimate b_1 and b_0

- $b_1 = 2.07$

- $b_0 = -1.21$

Our estimated regression line is:

$$\boxed{\widehat{Y_i} = -1.21 + 2.07X_i}$$



Linear Regression

Interpretation

$$\boxed{\widehat{Y}_i = -1.21 + 2.07X_i}$$

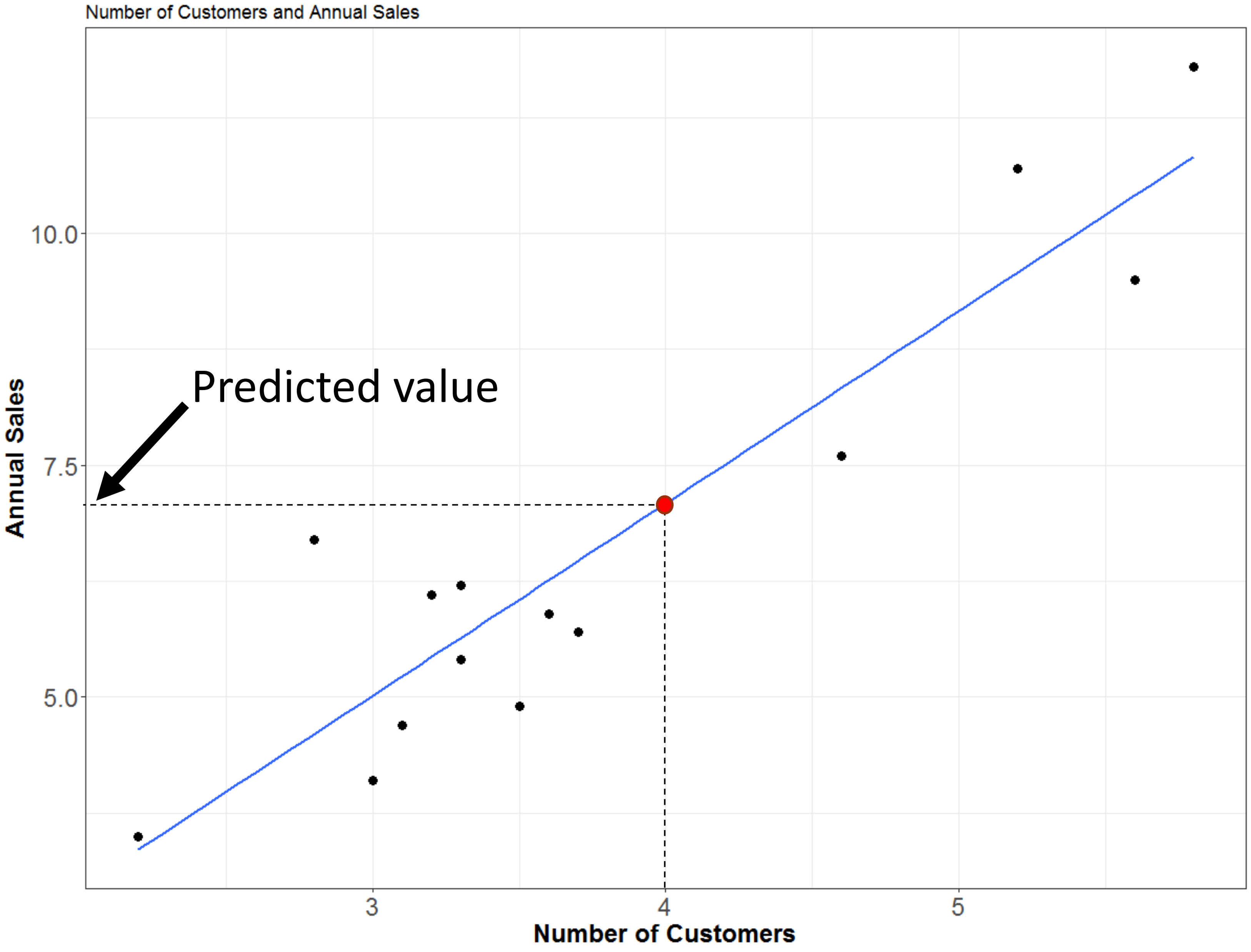
- If the number of customers increases by 1 million, the annual sales increase by 2.07 million.
- If the number of customers is zero, the **average** annual sales are -1.21 million.

Linear Regression

Predictions

$$\boxed{\widehat{Y}_i = -1.21 + 2.07X_i}$$

- What are the predicted sales if there are 4 million customers?


Number of Customers and Annual Sales
Predicted value
Annual Sales
10.0
7.5
5.0
3
4
5
Number of Customers

Linear Regression

Predictions

$$\boxed{\widehat{Y}_i = -1.21 + 2.07X_i}$$

- What are the predicted sales if there are 4 million customers?

$$\widehat{Y}_i = -1.21 + 2.07 * 4 \ = 7.07 \text{ million dollars}$$

- NOTE: Whenever possible, base your predictions off the *exact* coefficient estimates, rather than the rounded numbers.

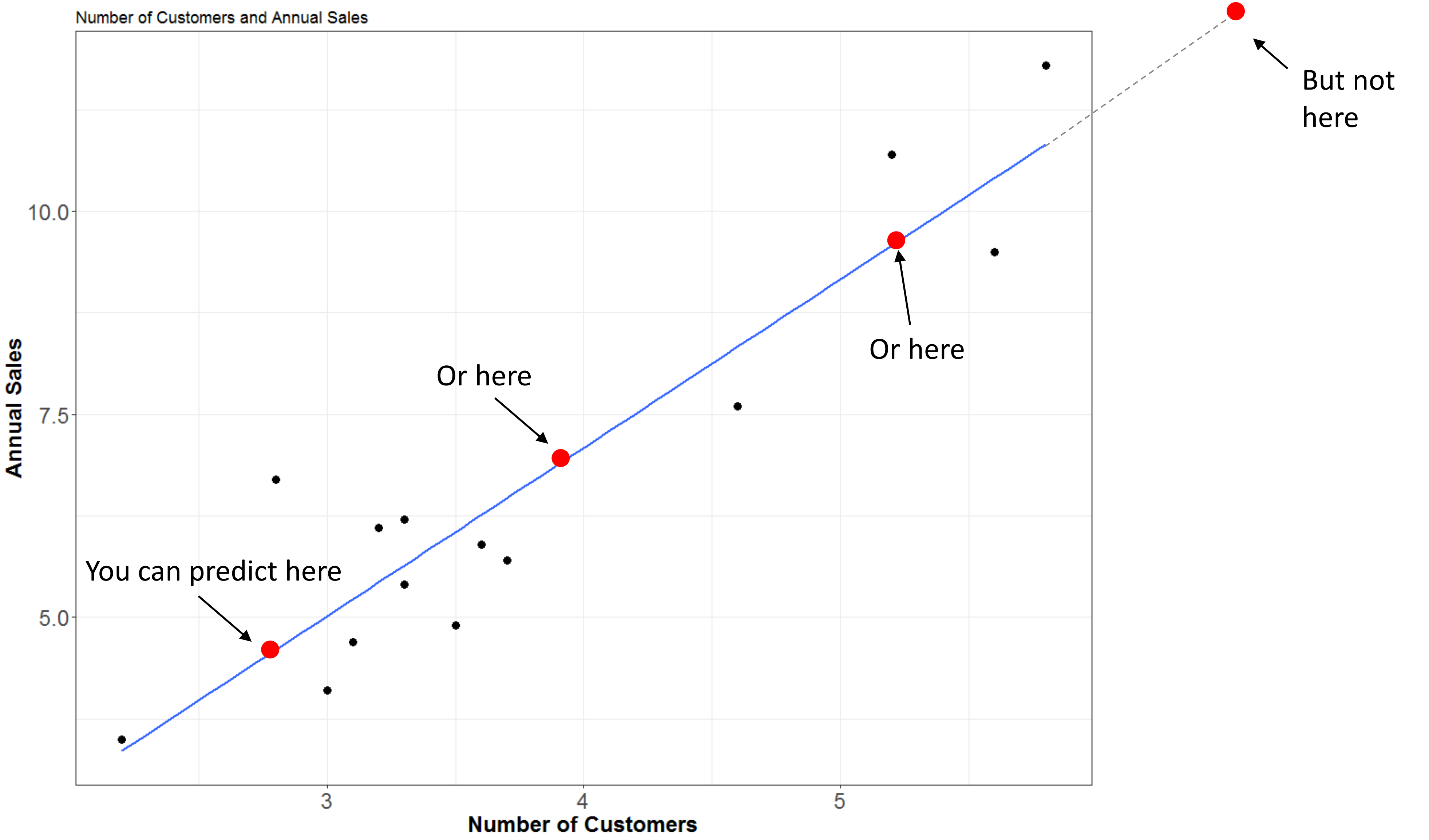
  - With the exact numbers in Excel, the prediction would be 7.09 million dollars. This is a slightly more accurate estimate.

Linear Regression

Your Predictions are Limited

- Only make predictions that are within the relevant range of your data
- In other words, you can predict *Y* for values of *X* that are between the smallest and the largest values of X in your data
- This is called Interpolation.
- Predicting values outside of your relevant range is called extrapolation, and should be avoided.

In other words…

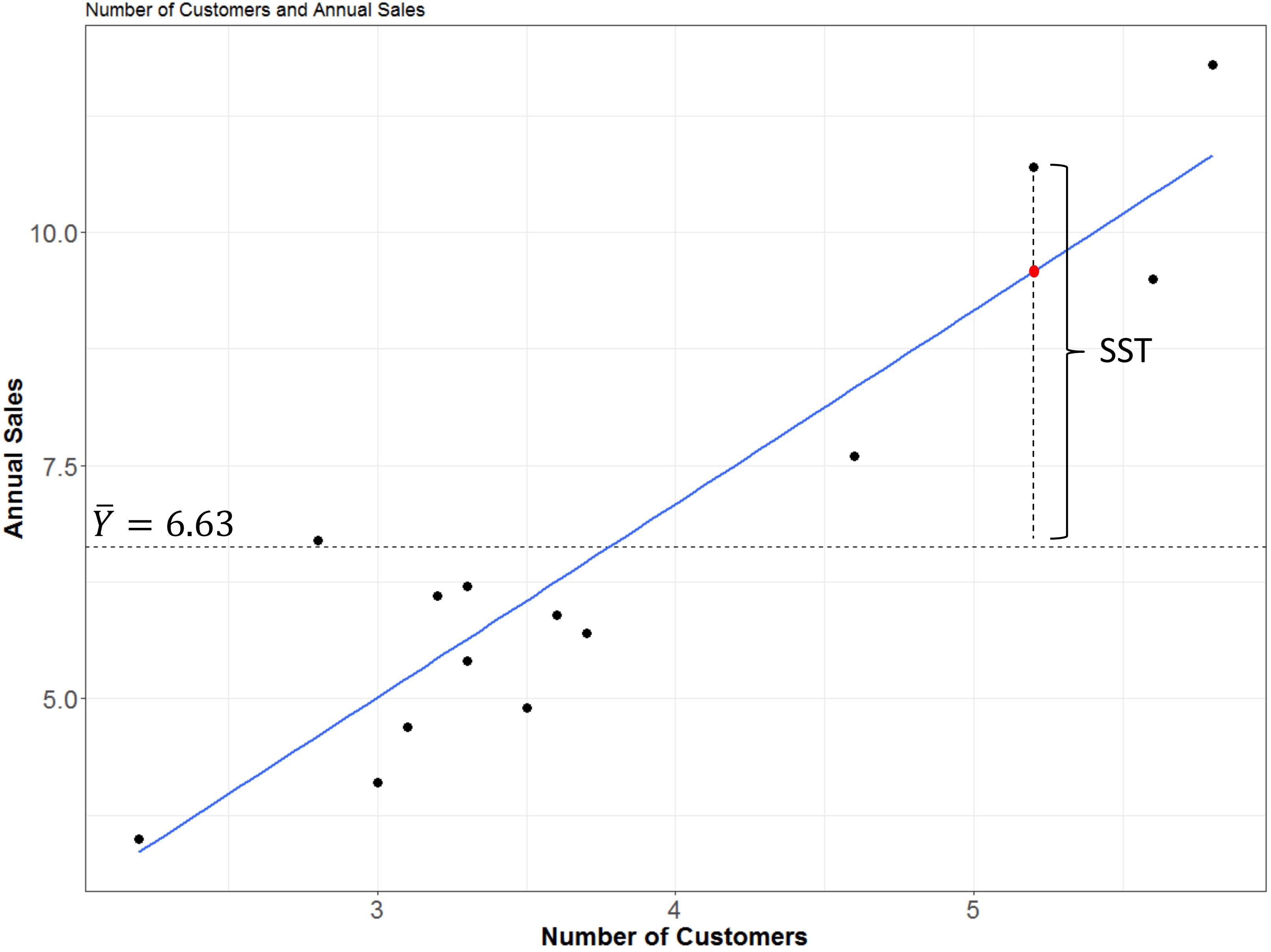


Measures of Variation

- Just like with ANOVA, it can be helpful to break the total variation in the data into 3 different groups

1. Variation of the observed data around the mean. This is the total sum of squares, or SST.

$$SST = \sum_{i=1}^{N} (\mathrm{Y_i} - \bar{Y})^2$$



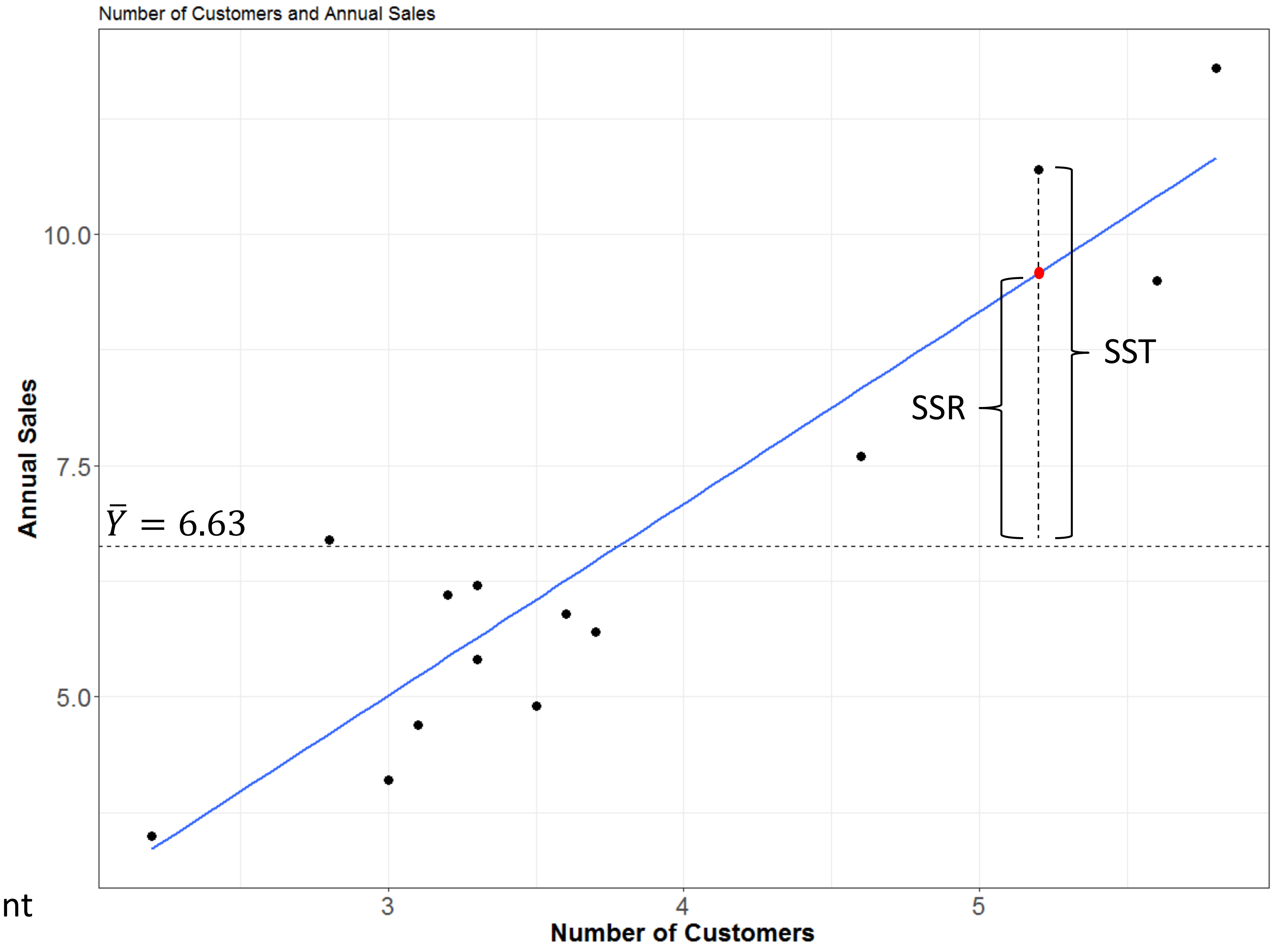
Measures of Variation

- Just like with ANOVA, it can be helpful to break the total variation in the data into 3 different groups

2. Variation of the predicted values around the mean. This is the sum of the squared residuals, or SSR.

$$SSR = \sum_{i=1}^{N} \left(\widehat{Y}_i - \bar{Y}\right)^2$$

- *Note*: the above formula is equivalent to the previous one given for SSR.

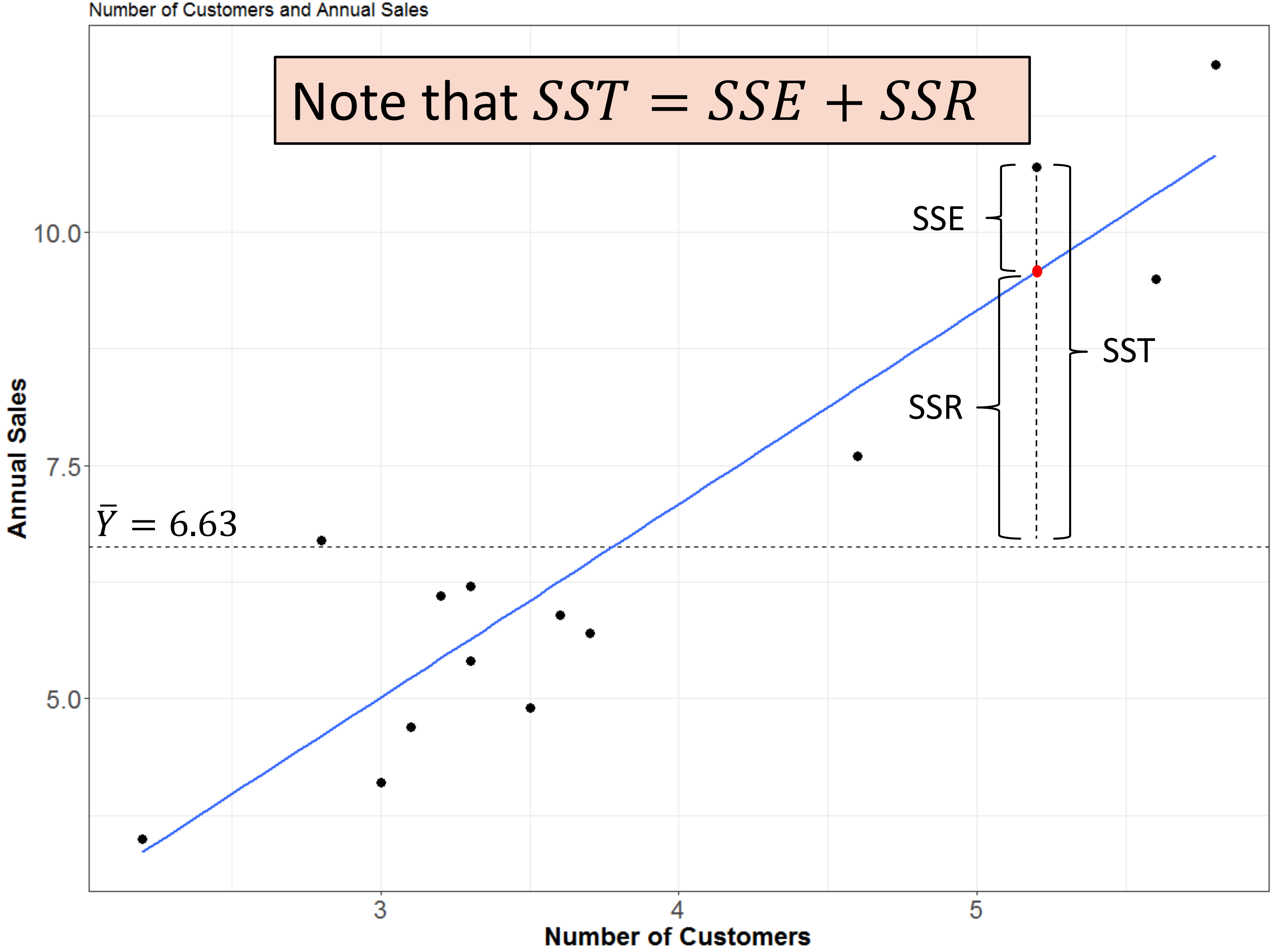


Measures of Variation

- Just like with ANOVA, it can be helpful to break the total variation in the data into 3 different groups

3. Variation of the observed values around the predicted values. This is the error sum of squares, or SSE.

$$SSE = \sum_{i=1}^{N} (Y_i - \widehat{Y}_i)^2$$



Linear Regression

Two ways to Evaluate a Model Using Variation

1. The coefficient of determination, R^2

$$R^2 = \frac{SSR}{SST}$$

- This measures the amount of variation in Y that is explained by X.

- A high R^2 means your independent variable, X, is a good predictor of Y.

(Ex) If your $R^2 = .90$, then your model explains 90% of the variation in Y. This is considered a very good fit, and should make relatively good predictions.

Linear Regression

Two ways to Evaluate a Model Using Variation

2. The standard error of the estimate, S_{xy}

$$S_{xy} = \sqrt{\frac{SSE}{n-2}}$$

- This is the standard deviation of observations around the prediction line.
- It tells you, on average, how far off a prediction will be.

(Ex) Say, for our previous example with annual sales and customers, we get a $S_{xy} = 1.5$. Then, on average, our predictions are off by 1.5 (million) dollars.