STAT 2593

Lecture 038 - The Simple Linear Regression Model

Dylan Spicker

The Simple Linear Regression Model

Learning Objectives

1. Describe the simple linear regression model and the constituent components.

2. Understand the normal assumptions for the simple linear regression model.

Regression Models

- We have discussed models which compare two different populations against one another.

Regression Models

- We have discussed models which compare two different populations against one another.
- It is often of interest to consider the relationship between two different variables

Regression Models

- We have discussed models which compare two different populations against one another.
- It is often of interest to consider the relationship between two different variables
  - The effect of treatment on a health outcome.

Regression Models

- We have discussed models which compare two different populations against one another.
- It is often of interest to consider the relationship between two different variables
  - The effect of treatment on a health outcome.
  - Housing factors influencing the cost of a home.

Regression Models

- We have discussed models which compare two different populations against one another.
- It is often of interest to consider the relationship between two different variables
  - The effect of treatment on a health outcome.
  - Housing factors influencing the cost of a home.
  - Material treatments to influence its durability.

Regression Models

- We have discussed models which compare two different populations against one another.
- It is often of interest to consider the relationship between two different variables
  - The effect of treatment on a health outcome.
  - Housing factors influencing the cost of a home.
  - Material treatments to influence its durability.
- When we are interested in describing this relationship directly, we typically use **regression models**.

The Deterministic Linear Model

- The simplest **deterministic** relationship between two variables is a straight line.

The Deterministic Linear Model

- The simplest **deterministic** relationship between two variables is a straight line.
  - We can write $y = \beta_0 + \beta_1 x$ to describe any linear relationship between y and x.

The Deterministic Linear Model

- The simplest **deterministic** relationship between two variables is a straight line.
  - We can write $y = \beta_0 + \beta_1 x$ to describe any linear relationship between y and x.
  - y is considered the **response** or **dependent** variable.

The Deterministic Linear Model

- The simplest **deterministic** relationship between two variables is a straight line.
  - We can write $y = \beta_0 + \beta_1 x$ to describe any linear relationship between y and x.
  - y is considered the **response** or **dependent** variable.
  - x is considered the **predictor** or **independent** variable.

The Deterministic Linear Model

- The simplest **deterministic** relationship between two variables is a straight line.
  - We can write $y = \beta_0 + \beta_1 x$ to describe any linear relationship between y and x.
  - y is considered the **response** or **dependent** variable.
  - x is considered the **predictor** or **independent** variable.
  - β_0 is the intercept for the line.

The Deterministic Linear Model

- The simplest **deterministic** relationship between two variables is a straight line.
  - We can write $y = \beta_0 + \beta_1 x$ to describe any linear relationship between y and x.
  - y is considered the **response** or **dependent** variable.
  - x is considered the **predictor** or **independent** variable.
  - β_0 is the intercept for the line.
  - β_1 is the slope for the line.

The Deterministic Linear Model

- The simplest **deterministic** relationship between two variables is a straight line.
  - We can write $y = \beta_0 + \beta_1 x$ to describe any linear relationship between y and x.
  - y is considered the **response** or **dependent** variable.
  - x is considered the **predictor** or **independent** variable.
  - β_0 is the intercept for the line.
  - β_1 is the slope for the line.
- The **simple linear regression model** takes this, and makes it probabilistic.

The Linear Regression Model

- Instead of considering y and x fixed, we take them to be random quantities.

The Linear Regression Model

- Instead of considering y and x fixed, we take them to be random quantities.
  - We write that

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

The Linear Regression Model

- Instead of considering y and x fixed, we take them to be random quantities.
  - We write that
$$Y = \beta_0 + \beta_1 X + \epsilon.$$
  - Y, X, β_0, and β_1 are as they were in the deterministic model.

The Linear Regression Model

- Instead of considering y and x fixed, we take them to be random quantities.
  - We write that
    $$Y = \beta_0 + \beta_1 X + \epsilon.$$
  - Y, X, β_0, and β_1 are as they were in the deterministic model.
  - ϵ is the noise term, which makes these quantities random.

The Linear Regression Model

- Instead of considering y and x fixed, we take them to be random quantities.
  - We write that
  $$Y = \beta_0 + \beta_1 X + \epsilon.$$
  - Y, X, β_0, and β_1 are as they were in the deterministic model.
  - ϵ is the noise term, which makes these quantities random.
- Note that $E[Y|X = x] = \beta_0 + \beta_1 x$.

The Linear Regression Model

- Instead of considering y and x fixed, we take them to be random quantities.
  - We write that
    $$Y = \beta_0 + \beta_1 X + \epsilon.$$
  - Y, X, β_0, and β_1 are as they were in the deterministic model.
  - ϵ is the noise term, which makes these quantities random.
- Note that $E[Y|X = x] = \beta_0 + \beta_1 x$.
  - The deterministic model is taken as a model for **mean of** Y.

The Linear Regression Model

- Instead of considering y and x fixed, we take them to be random quantities.
  - We write that
  $$Y = \beta_0 + \beta_1 X + \epsilon.$$
  - Y, X, β_0, and β_1 are as they were in the deterministic model.
  - ϵ is the noise term, which makes these quantities random.
- Note that $E[Y|X = x] = \beta_0 + \beta_1 x$.
  - The deterministic model is taken as a model for **mean of** Y.
  - The randomness comes from ϵ.

Normality Assumptions

- Often we will assume that $\epsilon \sim N(0, \sigma^2)$.

Normality Assumptions

- Often we will assume that $\epsilon \sim N(0, \sigma^2)$.
  - It is not strictly necessary to do this for the techniques to function well.

Normality Assumptions

- Often we will assume that $\epsilon \sim N(0, \sigma^2)$.
  - It is not strictly necessary to do this for the techniques to function well.
  - If this assumption is made then we have that $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$.

Normality Assumptions

- Often we will assume that $\epsilon \sim N(0, \sigma^2)$.
  - It is not strictly necessary to do this for the techniques to function well.
  - If this assumption is made then we have that $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$.
  - It is common, and unrestrictive, to assume that $E[\epsilon] = 0$.

Normality Assumptions

- Often we will assume that $\epsilon \sim N(0, \sigma^2)$.
  - It is not strictly necessary to do this for the techniques to function well.
  - If this assumption is made then we have that $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$.
  - It is common, and unrestrictive, to assume that $E[\epsilon] = 0$.
  - We typically assume that $\epsilon \perp X$.

Normality Assumptions

- Often we will assume that $\epsilon \sim N(0, \sigma^2)$.
  - It is not strictly necessary to do this for the techniques to function well.
  - If this assumption is made then we have that $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$.
  - It is common, and unrestrictive, to assume that $E[\epsilon] = 0$.
  - We typically assume that $\epsilon \perp X$.
- If we have values for x, and estimates of β_0 and β_1, then we can predict values of Y.

Normality Assumptions

- Often we will assume that $\epsilon \sim N(0, \sigma^2)$.
  - It is not strictly necessary to do this for the techniques to function well.
  - If this assumption is made then we have that $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$.
  - It is common, and unrestrictive, to assume that $E[\epsilon] = 0$.
  - We typically assume that $\epsilon \perp X$.
- If we have values for x, and estimates of β_0 and β_1, then we can predict values of Y.
  - If we assume normality, we can also predict intervals around these predictions.

Scatterplots

- A scatterplot is perhaps the most common type of plot

Scatterplots

- A scatterplot is perhaps the most common type of plot
- For bivariate data, we simply plot each datapoint at the corresponding location on the (x, y) plane.

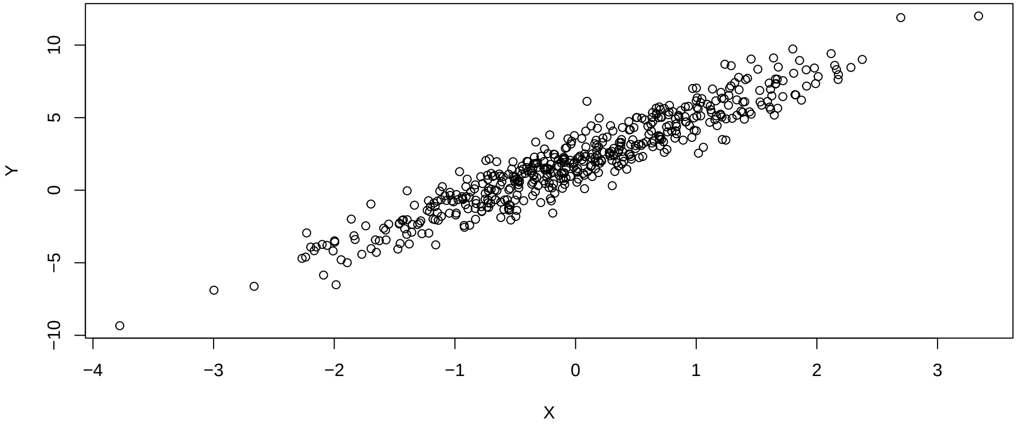
Scatterplots

- A scatterplot is perhaps the most common type of plot
- For bivariate data, we simply plot each datapoint at the corresponding location on the (x, y) plane.
- Scatterplots are useful for determining the relationship between two different variables, and in particular, assessing whether a specified relationship looks reasonable.

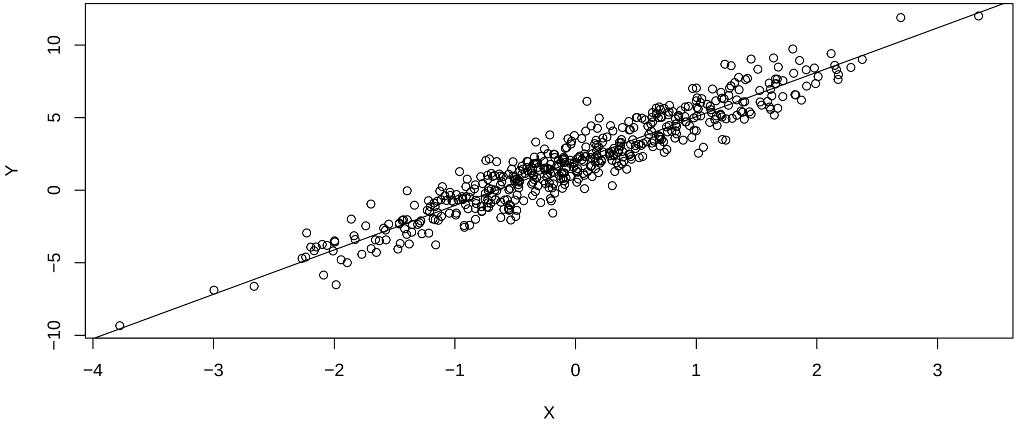
Scatterplots

- A scatterplot is perhaps the most common type of plot
- For bivariate data, we simply plot each datapoint at the corresponding location on the (x, y) plane.
- Scatterplots are useful for determining the relationship between two different variables, and in particular, assessing whether a specified relationship looks reasonable.
- In our case: does it seem like a straight line would fit the data well?

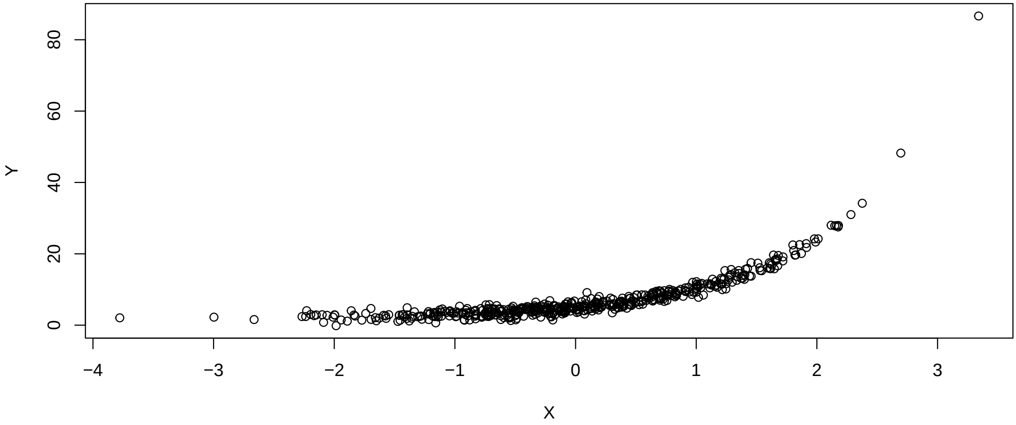
Examples



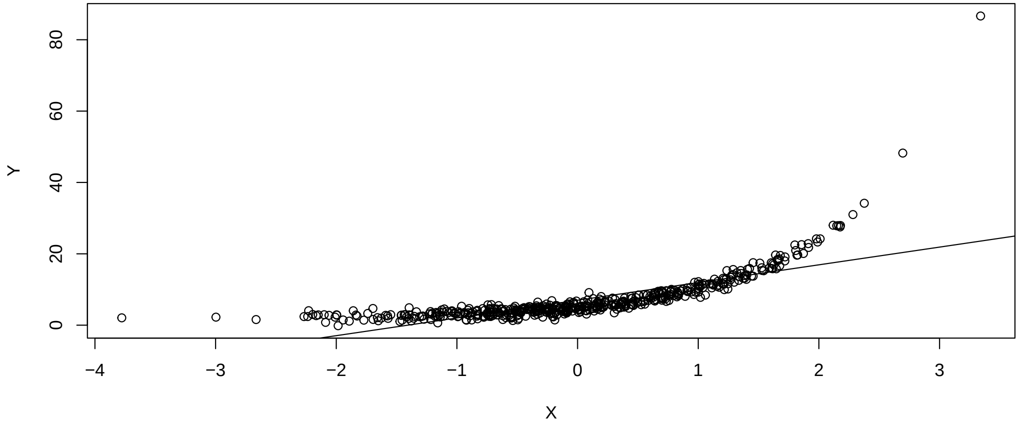
Examples



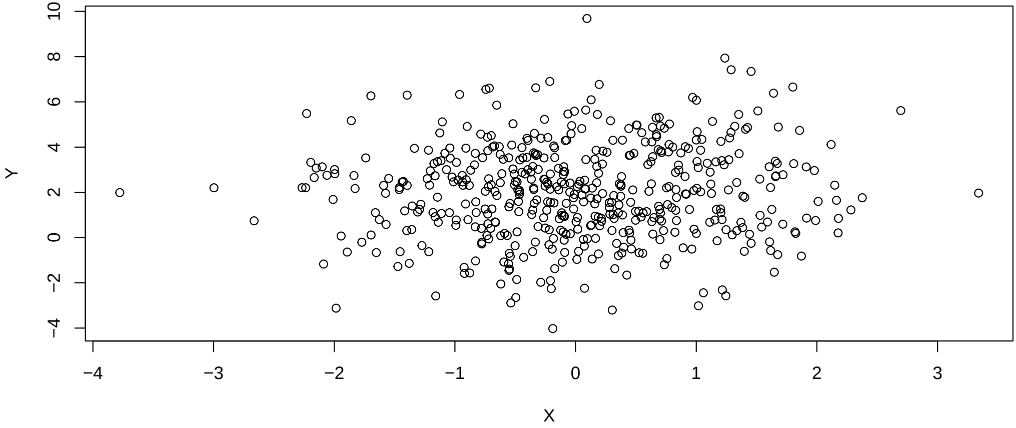
Examples



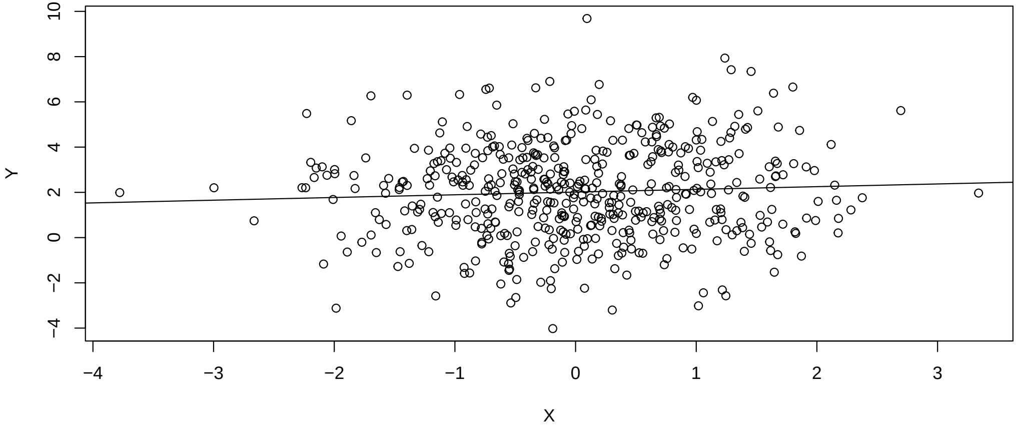
Examples



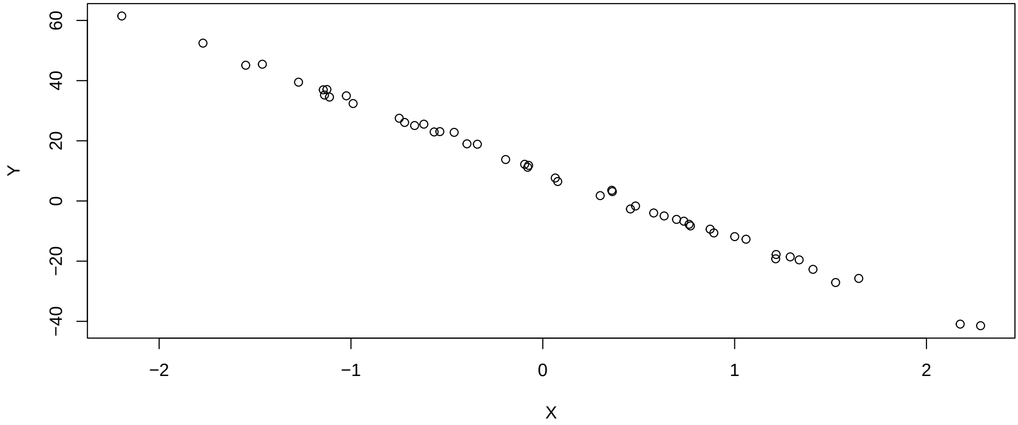
Examples



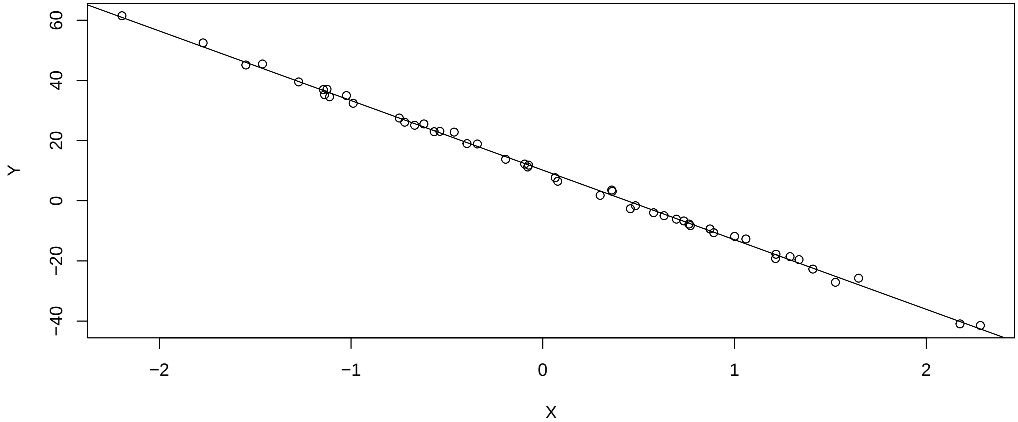
Examples



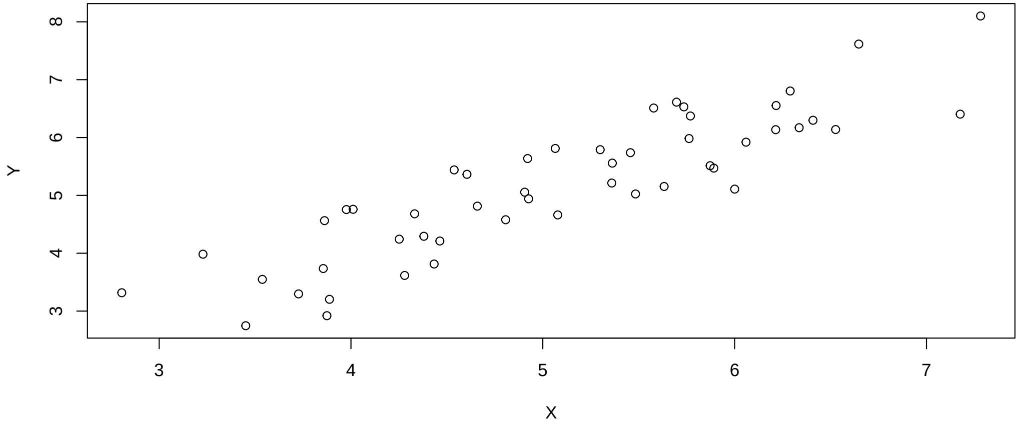
Examples



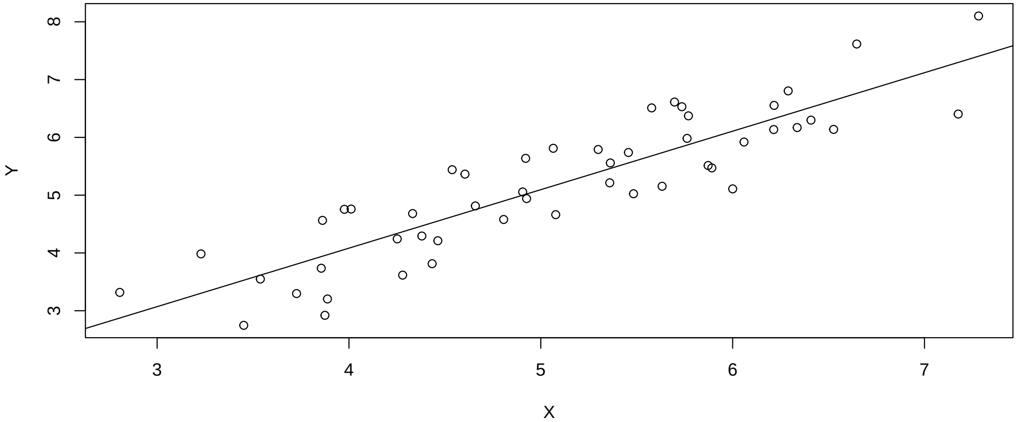
Examples



Examples



Examples



Summary

- Linear relationships are the simplest relationships to capture dependency between two variables.
- The linear regression model adds randomness through the use of an error term.
- We can use scatterplots to assess the linearity of a relationship (and perhaps to find transformations that make it more linear!).