DATA SCIENCE WITH R

DATA SCIENCE WITH R

REGRESSION ANALYSIS

Overview

Simple Linear Regression

Multiple Linear Regression

Regression Assumptions

Implementation in SAS

Regression

SIMPLE LINEAR REGRESSION

- ✓ **Concepts - OLS**
- ✓ How to Run
- ✓ Interpret Results

Simple Linear Regression

The simplest case we are trying to find is the relationship between baby birth weight and gestation period

Simple Linear Regression

The simplest case we are trying to find is the relationship between baby birth weight and gestation period

Mathematically:

Birthweight = *f* (gestation weeks)

Simple Linear Regression

The simplest case we are trying to find is the relationship between baby birth weight and gestation period

Mathematically:

Birthweight = *f* (gestation weeks)

where *f* is the functional form that we are trying to determine

Simple Linear Regression

The simplest case we are trying to find is the relationship between baby birth weight and gestation period

Mathematically:

Birthweight = *f* (gestation weeks)

where *f* is the functional form that we are trying to determine

We are currently reviewing a linear regression model

Simple Linear Regression

The simplest case we are trying to find is the relationship between baby birth weight and gestation period

Mathematically:

Birthweight = *f* (gestation weeks)

where *f* is the functional form that we are trying to determine

We are currently reviewing a linear regression model –

A linear relationship between two variables is essentially a straight line relationship

Simple Linear Regression

What is the mathematical equation that denotes a linear (straight line) relationship between two variables, x and y?

$$y = mx + c$$

Simple Linear Regression

What is the mathematical equation that denotes a linear (straight line) relationship between two variables, x and y?

$$y = mx + c$$

Where, **m = Slope, c = Intercept**

Simple Linear Regression

What is the mathematical equation that denotes a linear (straight line) relationship between two variables, x and y?

$$y = mx + c$$

Where, **m = Slope, c = Intercept**

Slope is the rate of change of Y when X changes, or the magnitude of impact of changes in X on Y

- What if B = 0? Then Y is a constant so there is no relationship between Y and X, because, however much X changes, Y does not change

Simple Linear Regression

What is the mathematical equation that denotes a linear (straight line) relationship between two variables, x and y?

$$y = mx + c$$

Where, **m = Slope, c = Intercept**

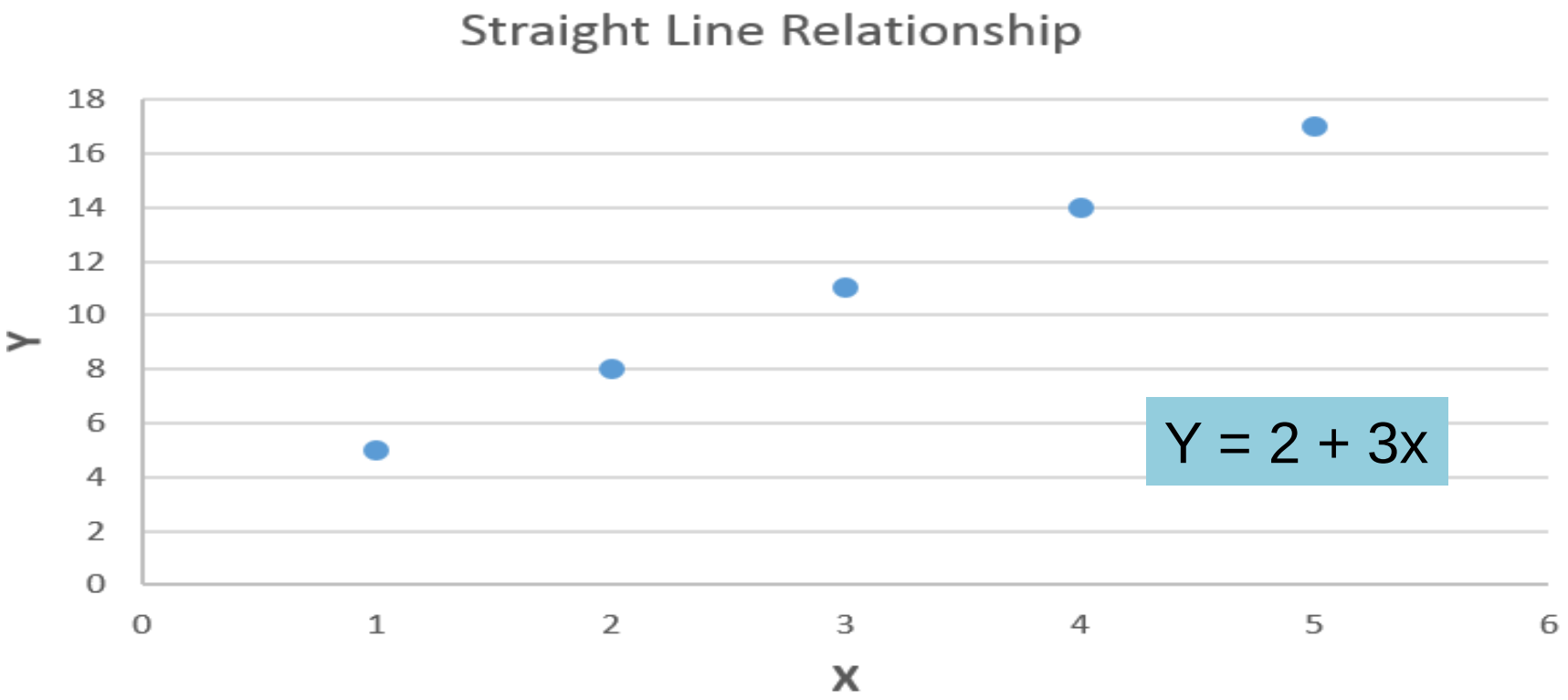
Slope is the rate of change of Y when X changes, or the magnitude of impact of changes in X on Y

- What if B = 0? Then Y is a constant so there is no relationship between Y and X, because, however much X changes, Y does not change

Intercept is the value of Y when X = 0.

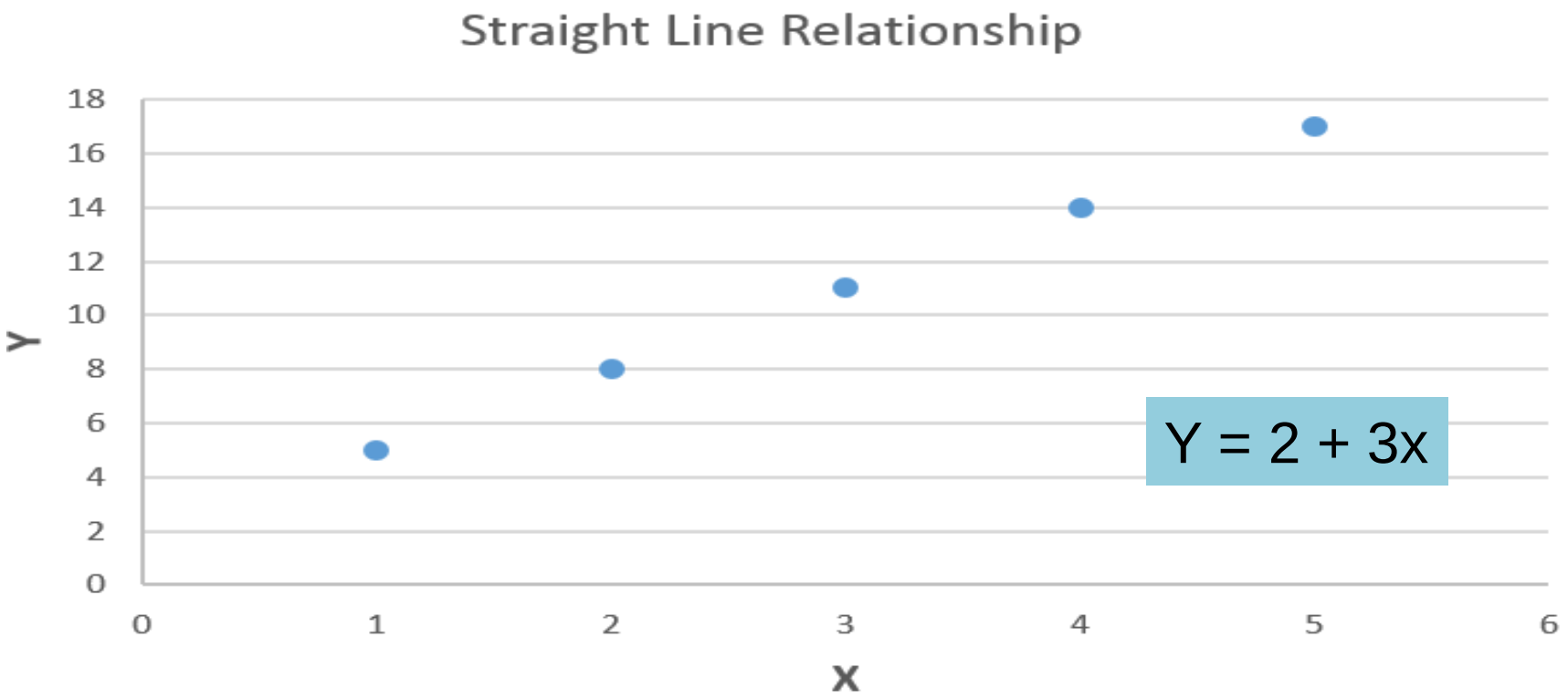
- What if A = 0? Then the line passes through the origin, and Y is directly proportional to X

Simple Linear Regression



Intercept = 2
Slope = 3

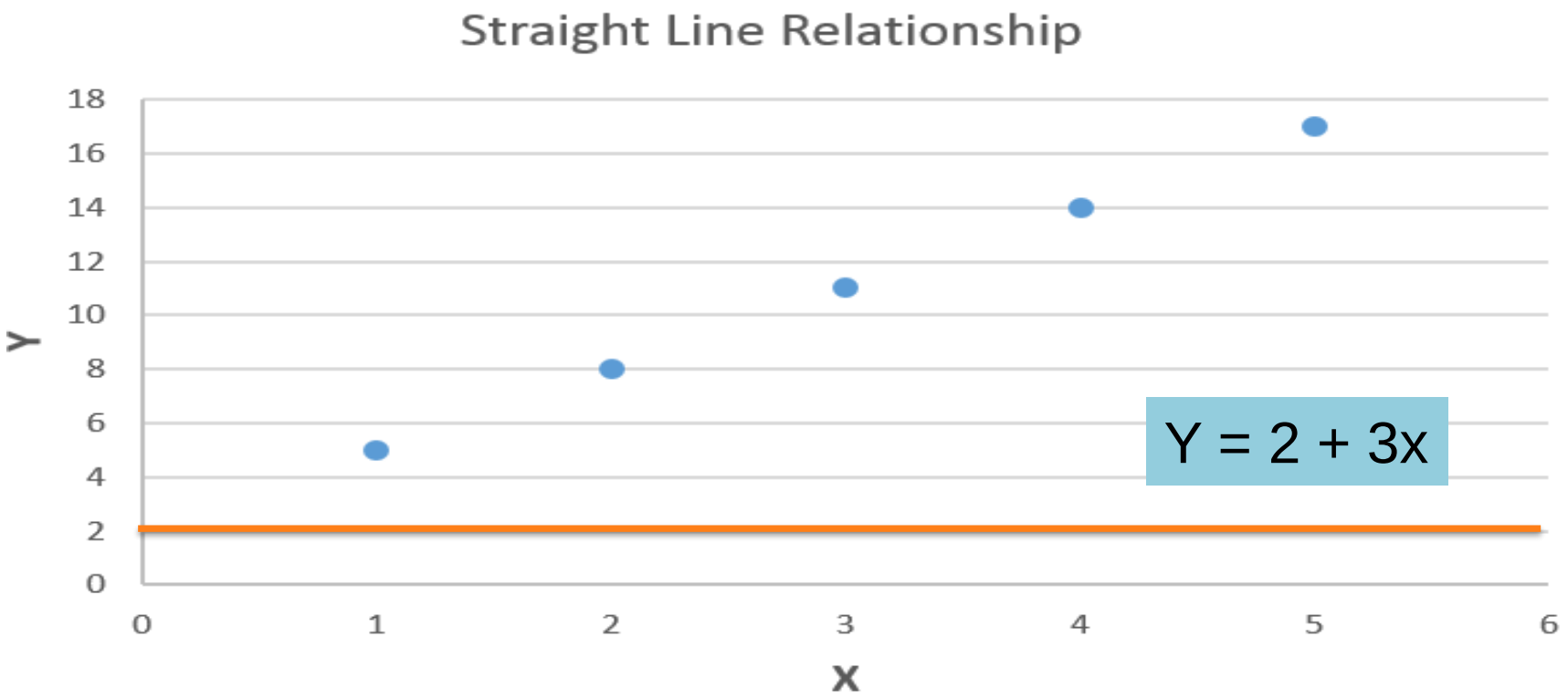
Simple Linear Regression



Intercept = 2
Slope = 3

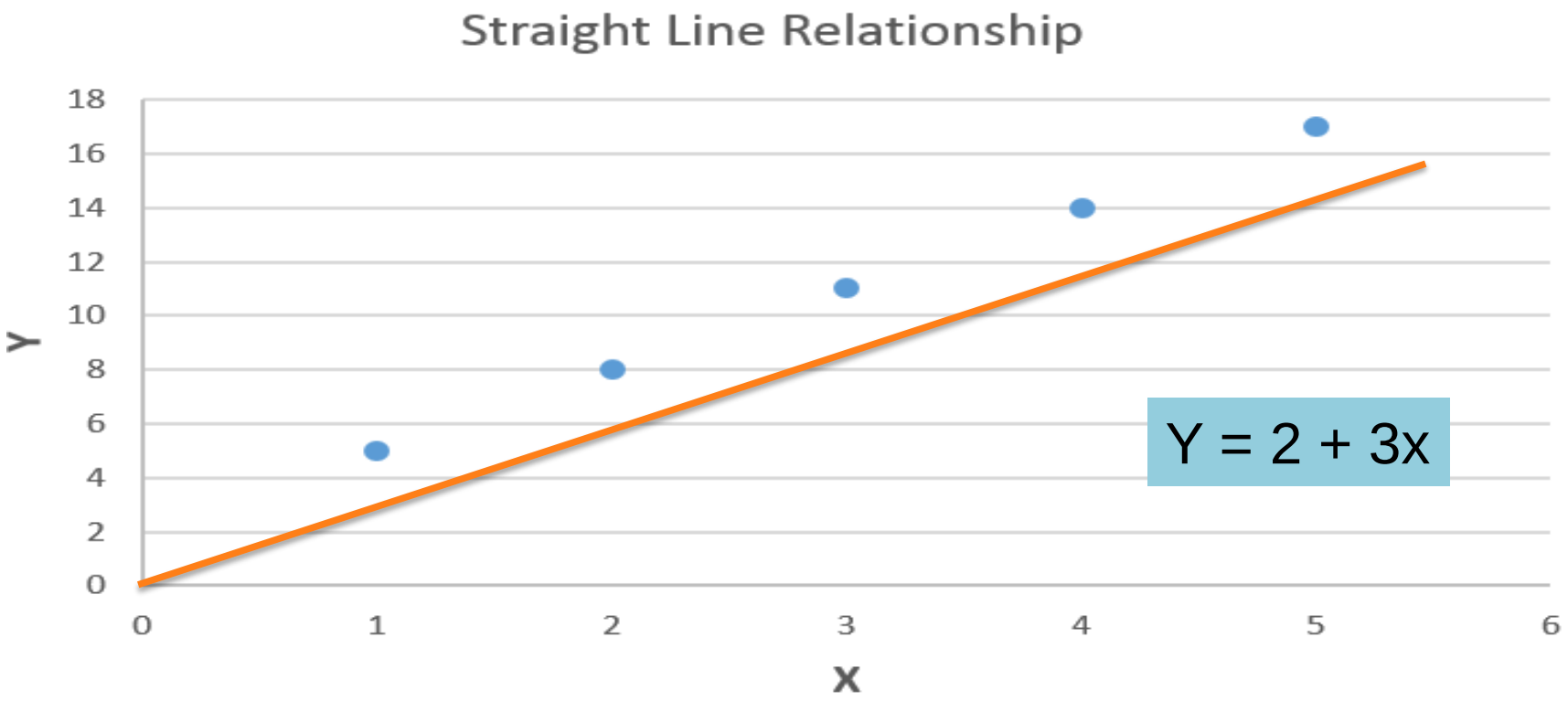
For a unit change in X, Y Changes by a constant amount (3)

Simple Linear Regression



If m = 0?

Simple Linear Regression



If c = 0?

Simple Linear Regression

In a linear regression model, since we are looking at a straight line relationship, the relationship is usually shown as

Simple Linear Regression

In a linear regression model, since we are looking at a straight line relationship, the relationship is usually shown as

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

where, β_0 = Intercept

β_1 = Slope

e = Error

Simple Linear Regression

In a linear regression model, since we are looking at a straight line relationship, the relationship is usually shown as

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

where, β_0 = Intercept

β_1 = Slope

e = Error

Therefore, in order to understand the relationship between X and Y, we need to figure out what the values of the BETAs are

Simple Linear Regression

TERMINOLOGY

Simple Linear Regression

TERMINOLOGY

Dependent Variable: Y: Predicted Variables: *The variable whose behavior we hypothesize can be explained or influenced by other factors*

Simple Linear Regression

TERMINOLOGY

Dependent Variable: Y: Predicted Variables: *The variable whose behavior we hypothesize can be explained or influenced by other factors*

Independent Variable (s): X(s): Predictor(s): *The factor(s) that we hypothesize influence the dependent variable*

Simple Linear Regression

TERMINOLOGY

Dependent Variable: Y: Predicted Variables: *The variable whose behavior we hypothesize can be explained or influenced by other factors*

Independent Variable (s): X(s): Predictor(s): *The factor(s) that we hypothesize influence the dependent variable*

Beta Coefficient(s): *The estimate of magnitude of impact of changes in the predictor(s) on the predicted variable*

Simple Linear Regression

TERMINOLOGY

Dependent Variable: Y: Predicted Variables: *The variable whose behavior we hypothesize can be explained or influenced by other factors*

Independent Variable (s): X(s): Predictor(s): *The factor(s) that we hypothesize influence the dependent variable*

Beta Coefficient(s): *The estimate of magnitude of impact of changes in the predictor(s) on the predicted variable*

Error:e:u: *The impact of the unobserved variables on the dependent variable, usually calculated as the difference between the predicted value of Y given the estimated regression function and the actual value of Y*

Simple Linear Regression

In the birthweight example, we believe:

$$Birthweight = \beta_0 + \beta_1 * Gestation\ Period + e$$

Simple Linear Regression

In the birthweight example, we believe:

$$Birthweight = \beta_0 + \beta_1 * Gestation\ Period + e$$

We now need to estimate what the beta coefficients values are, from the data available to us, that will best capture the relationship between Birthweight and Gestation Period

Ordinary Least Squares Regression

Ordinary Least Squares Regression

The Ordinary Least Squares Regression (OLS) technique estimates coefficients on the variables hypothesized to have an impact on the variable of interest by identifying the line that minimizes the sum of squared differences between points on the estimated line and the actual values of the independent variable

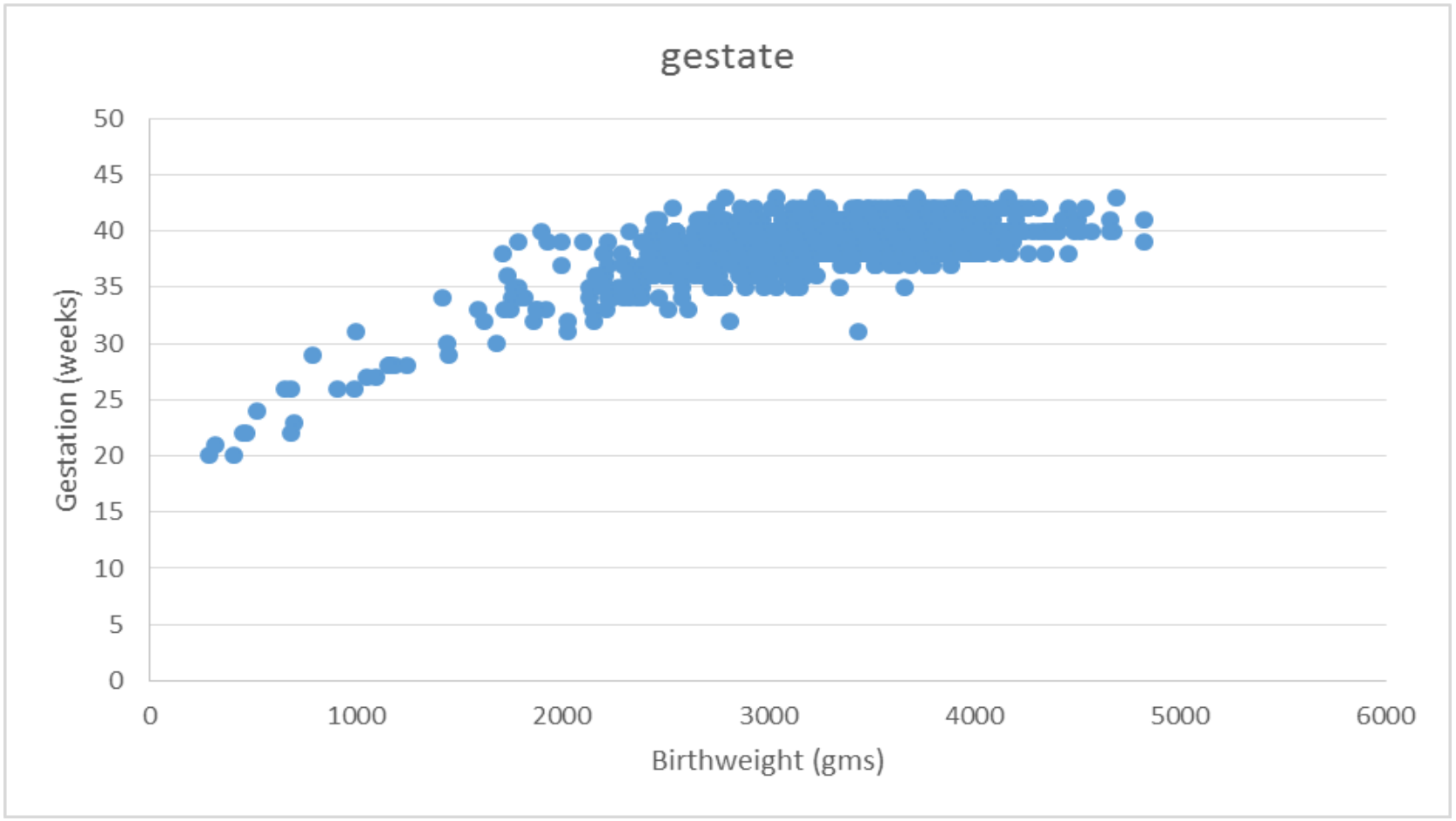
Ordinary Least Squares Regression

The Ordinary Least Squares Regression (OLS) technique estimates coefficients on the variables hypothesized to have an impact on the variable of interest by identifying the line that minimizes the sum of squared differences between points on the estimated line and the actual values of the independent variable

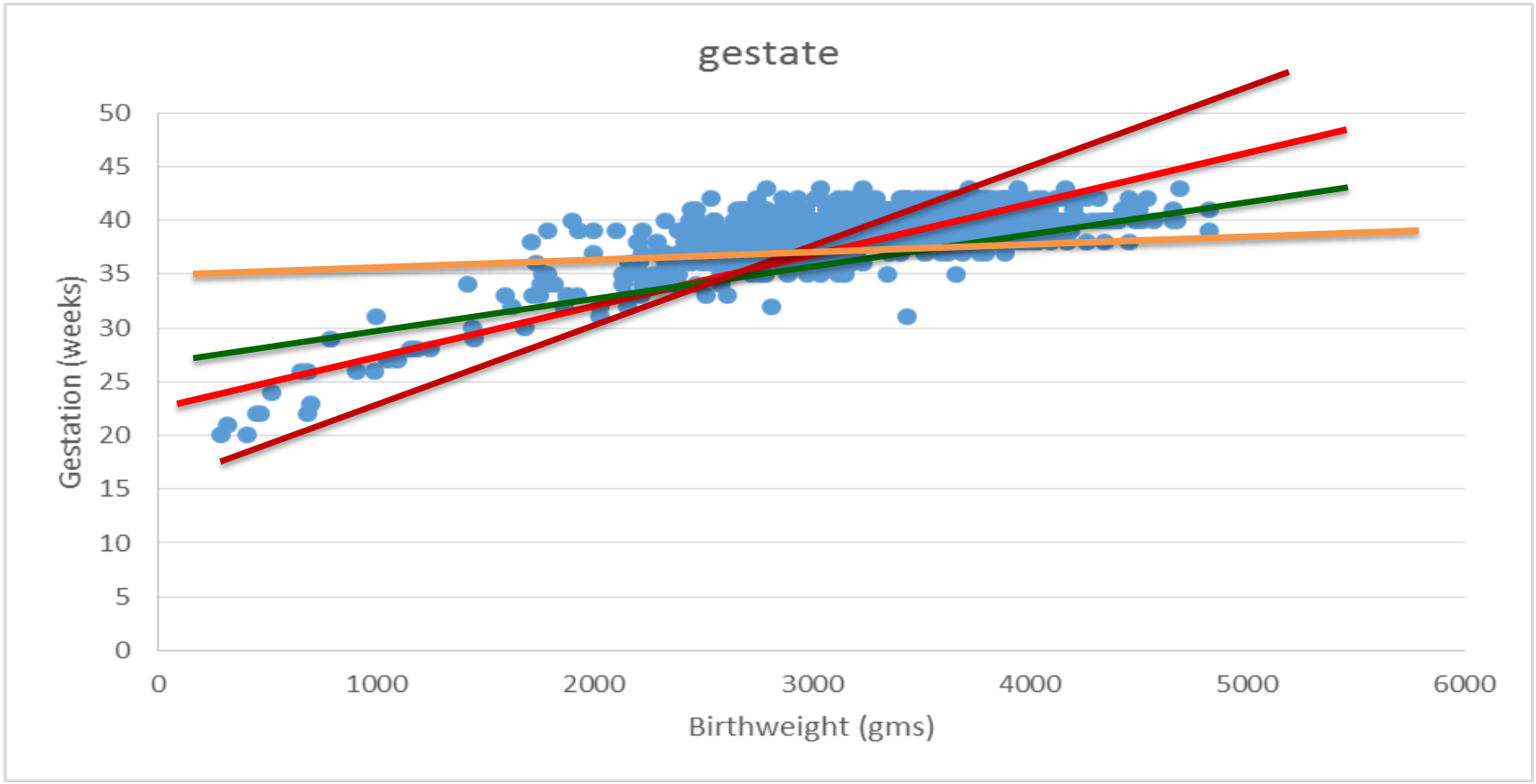
- **Coefficients**: Betas
- **Minimizes**: Least
- **Sum of Squared Differences:** Square of residuals
- **Estimated Line:** Regression Line
- **Actual Values:** Values in data set

OLS Regression

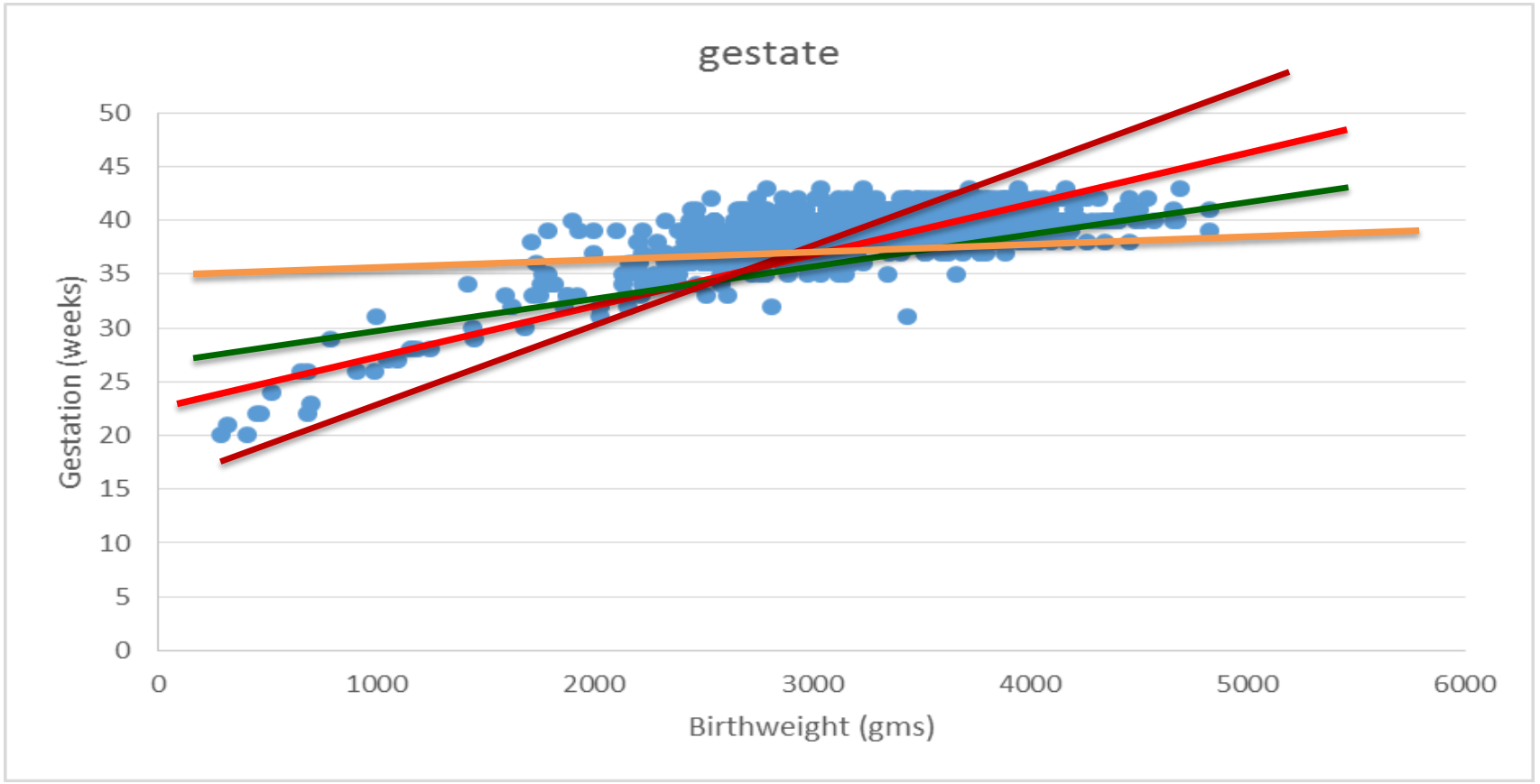
A scatter plot of values of birth weight and gestation period



OLS Regression



OLS Regression



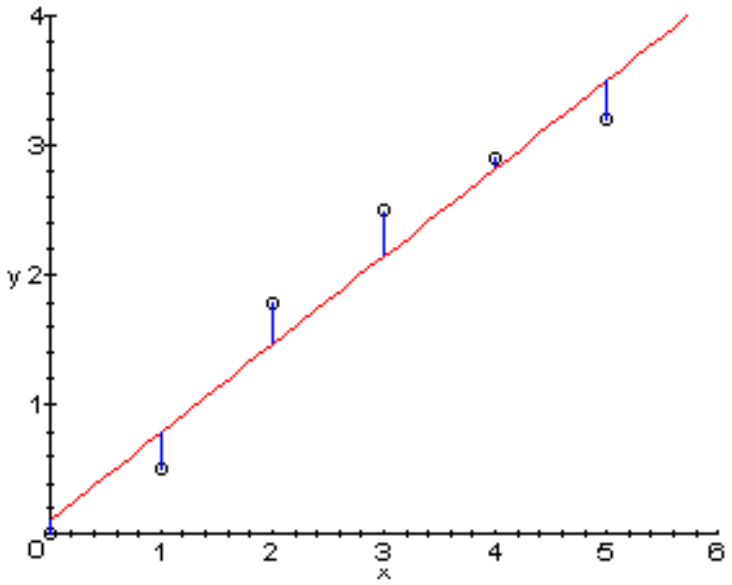
- Is there a linear relationship?
- Would it be possible to fit a straight line through these points?
- How many straight lines?

OLS Regression

Clearly we can fit many straight lines that each will cover some of the points

Is there a straight line that can hit all points?

One way of choosing a line among all possible lines is to identify the line that would explain most variation in Y -
In other words, have least total error



OLS Regression

OLS Estimates

The Ordinary Least Squares regression find that line by looking at the residuals (or the difference between the points on each line and actual Y) and minimizing the sum of their squares

Why sum of squares?

Positive and Negative Differences

Mathematically, minimize $$Q = \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_i)^2$$

OLS Regression

Using differential calculus, we will get

$$b_o = \frac{\Sigma X_i^2 \Sigma Y_i - \Sigma X_i \Sigma X_i Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

$$b_1 = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

OLS Regression

Using differential calculus, we will get

$$b_o = \frac{\Sigma X_i^2 \Sigma Y_i - \Sigma X_i \Sigma X_i Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

$$b_1 = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

These estimates are called the **Ordinary Least Squares** estimates

OLS Regression

Using differential calculus, we will get

$$b_o = \frac{\Sigma X_i^2 \Sigma Y_i - \Sigma X_i \Sigma X_i Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

$$b_1 = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

These estimates are called the **Ordinary Least Squares** estimates

We can be sure that given the data, the Ordinary Least Squares estimate line minimizes errors more than any other line that we choose

OLS Regression

Once we estimate the coefficients, we have an equation like this:

Birthweight **=** **Intercept estimate** **+** **Beta Coeff*** **Gestation**

OLS Regression

Once we estimate the coefficients, we have an equation like this:

Birthweight **=** **Intercept estimate** **+** **Beta Coeff*** **Gestation**

Remember, this is the best fitted line, but this line will not cover every single point on the scatter plot

To Be Continued

Regression Analysis

Simple Linear Regression

THANK YOU