Linear Regression

Learning Objectives

- Matrix Formulation
- Multiple Linear Regression
- Model Assumptions

Matrix Formulation

Matrix Version of Model

$$Y_i = \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i$$

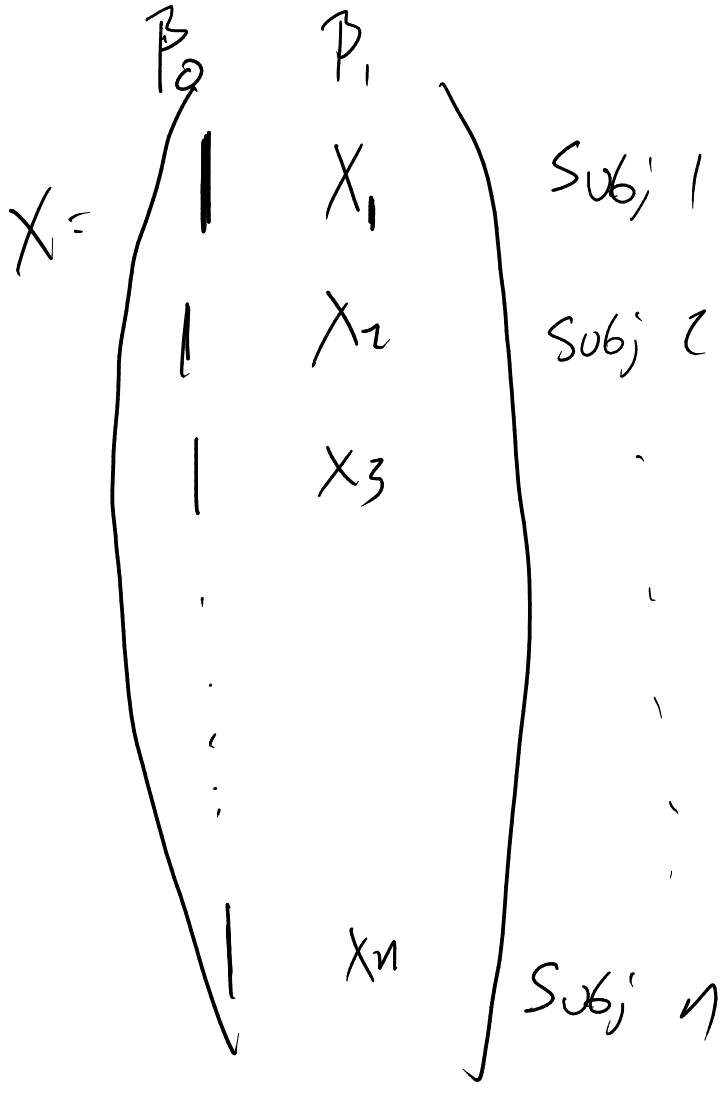
- Y_i: Outcome Variable
- $\boldsymbol{X}_i = (1, X_i)^{\mathrm{T}}$: Predictors
- $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$: Coefficients
- ϵ_i: error term

Data Matrix Formulation

For n data points $Y_i = \beta_0 + \beta_1 X_i$

$$\underset{n\times 1}{\boldsymbol{Y}} = \underset{n\times 2}{\boldsymbol{X}^{\mathrm{T}}} \underset{2\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}$$

- $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^{\mathrm{T}}$: Outcome Variable
- $\underset{2\times n}{\boldsymbol{X}} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^{\mathrm{T}}$: Predictors
- $\underset{2\times 1}{\boldsymbol{\beta}} = (\beta_0, \beta_1)^{\mathrm{T}}$: Coefficients
- $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^{\mathrm{T}}$: Error terms



Least Squares Formula

$$(Y - X^{\mathrm{T}}\beta)^{\mathrm{T}}(Y - X^{\mathrm{T}}\beta)$$

$$\Updownarrow$$

$$\sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2$$

Estimates

$$\underset{n\times 2}{X} = \begin{pmatrix} 1 & x_1 \\ 1 & \vdots \\ 1 & x_n \end{pmatrix} \qquad X^T = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & \cdots & & & x_n \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$$

$2\times n \quad n\times 2 \quad 2\times n \quad n\times 1$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

2×1

Multiple Linear Regression

MLR

Multivariable linear regression models are used when more than one explanatory variable is used to explain the outcome of interest.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

(X_1: wing, X_2: beak)

As wingspan increases by 1 unit, body mass will increase/decrease by an average of β_1 units, after adjusting for beak size

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}) \right)^2$$

with x_{1i} and x_{2i} being

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & & \\ 1 & x_{1n} & x_{2n} \end{pmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

3×1 $\quad$ $3 \times n$ $n \times 3$ $\quad$ $3 \times n$ $n \times 1$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$$

Continuous Variable

To fit an additional continuous random variable to the model, we will only need to add it to the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Categorical Variable

$$X_2 = \begin{cases} 1 & \text{Seagull} \\ 2 & \text{condor} \\ 3 & \text{duck} \end{cases}$$

A categorical variable can be included in a model, but a reference category must be specified.

$$y^{\text{body mass}} = \beta_0 + \beta_1 X_1^{\text{wingspan}} + \beta_2 D_2 + \beta_3 D_3$$

$$\hat{y}_{\text{duck}} - \hat{y}_{\text{condor}} = \hat{\beta}_0 + \hat{\beta}_1(7) + \hat{\beta}_3 - \left(\hat{\beta}_0 + \hat{\beta}_1(5) + \hat{\beta}_2\right)$$

$$\hat{\beta}_1(2) + \hat{\beta}_3 - \hat{\beta}_2$$

$$\hat{\beta}_3 - \hat{\beta}_2$$

$$D_2 = \begin{cases} 1 & \text{Condor} \\ 0 & \text{o.w.} \end{cases} \qquad D_3 = \begin{cases} 1 & \text{duck} \\ 0 & \text{o.w} \end{cases}$$

reference category = Seagull

On average the body mass of a ~~condor~~ duck is greater/

lesser than a seagull by 87 units, adjusting for wingspan

Fitting a model with categorical variables

To fit a model with categorical variables, we must utilize dummy (binary) variables that indicate which category is being referenced. We use $C - 1$ dummy variables where C indicates the number of categories. When coded correctly, each category will be represented by a combination of dummy variables.

Example

If we have 4 categories, we will need 3 dummy variables:

	Cat 1	**Cat 2**	**Cat 3**	**Cat 4**
Dummy 1	1	0	0	0
Dummy 2	0	1	0	0
Dummy 3	0	0	1	0

Which one is the reference category?

Cat 4

Matrix Notation

$$Y = \boldsymbol{\beta}^T \boldsymbol{X}$$

- $\boldsymbol{\beta}$: a column vector of regression coefficients
- $\boldsymbol{X}$: a column vector of predictor variables

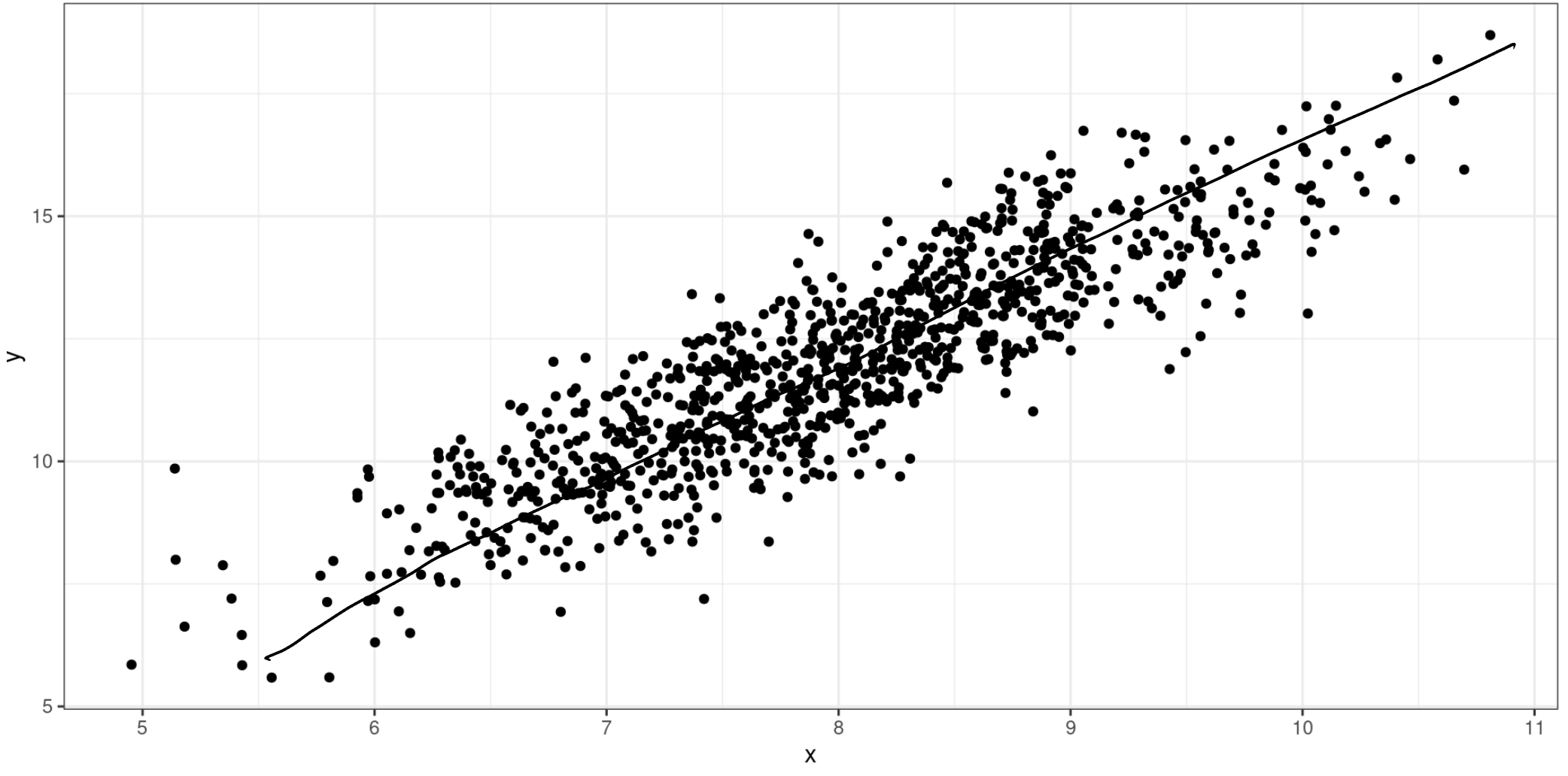
Model Assumptions

Model

$$Y = \boldsymbol{\beta}^T \boldsymbol{X}$$

- $\epsilon \sim N(0, \sigma^2)$

Model Scatter Plot

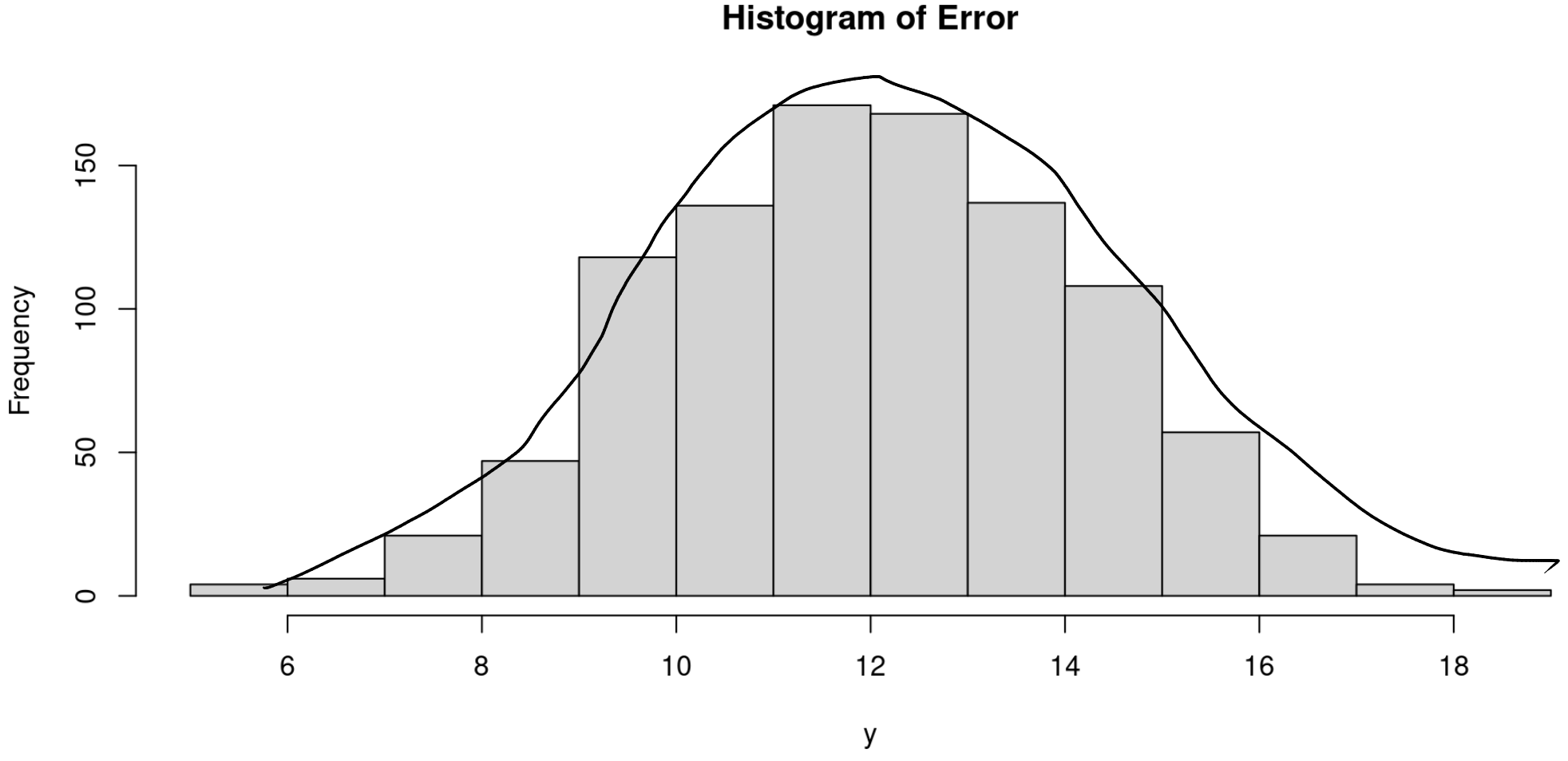


(X_1, X_2, X_3)

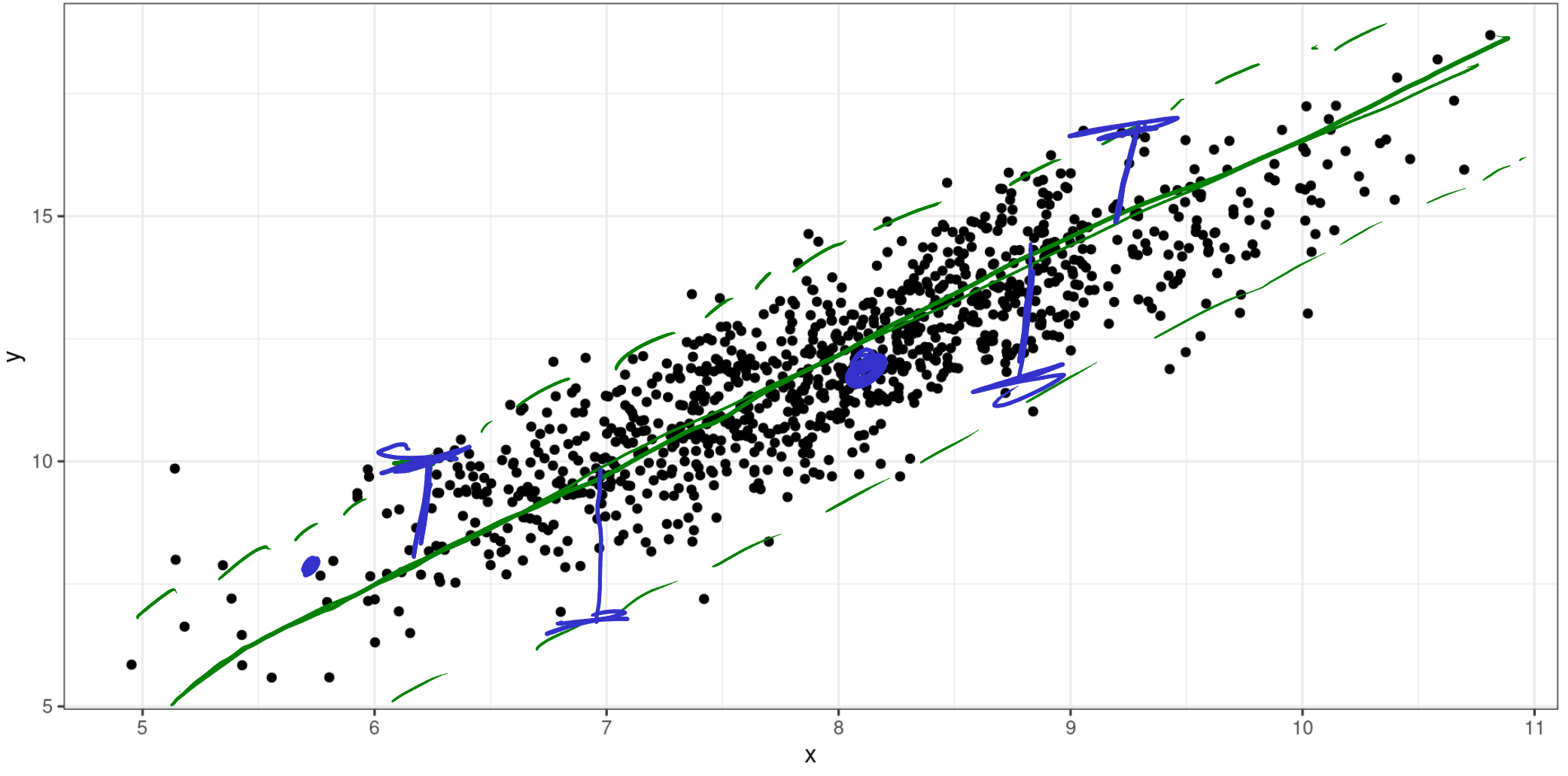
Model Assumptions

- Errors are normally distributed
- Constant Variance
- Linearity
- Independence
- No outliers

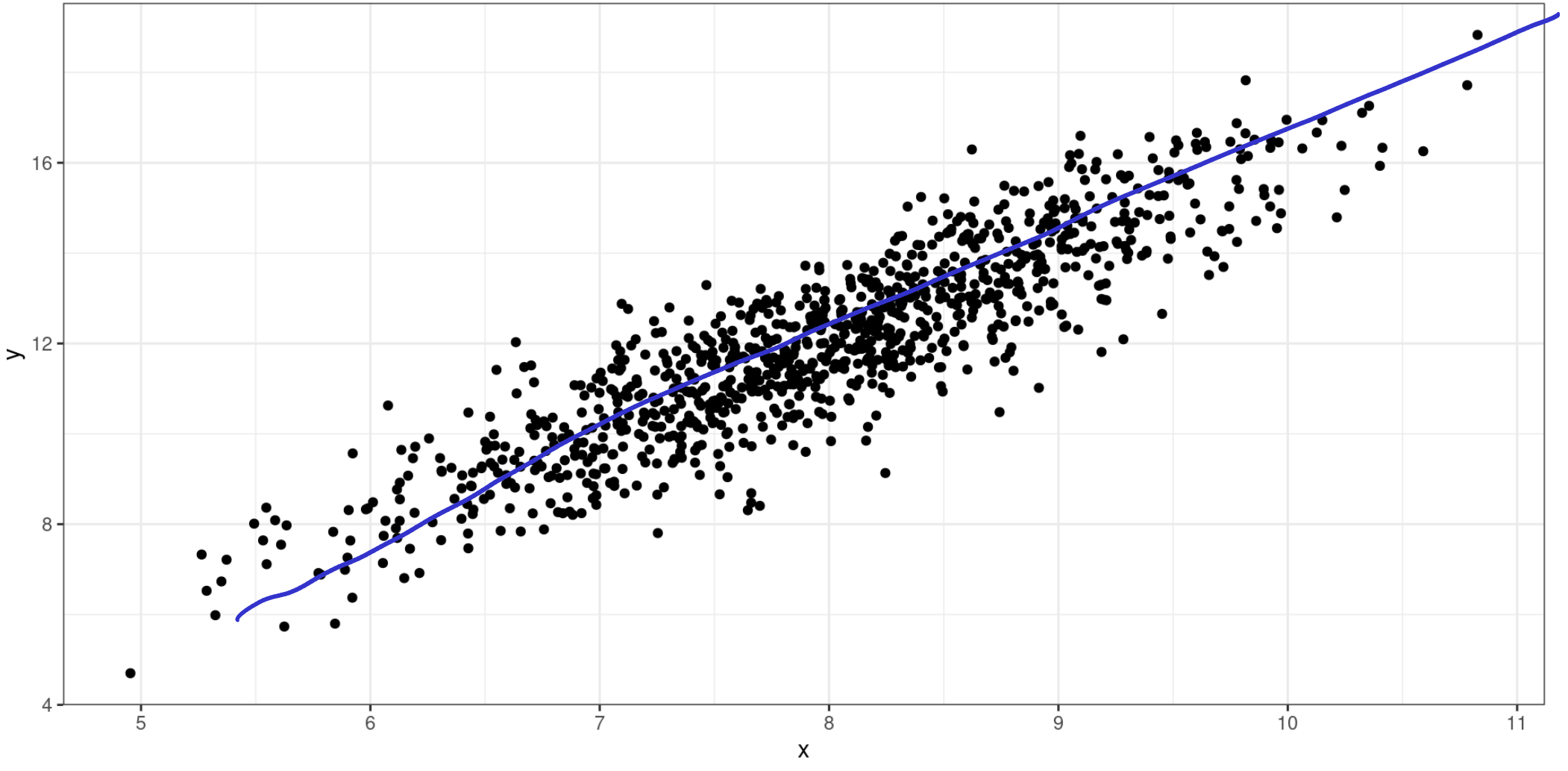
Errors Normally Distributed



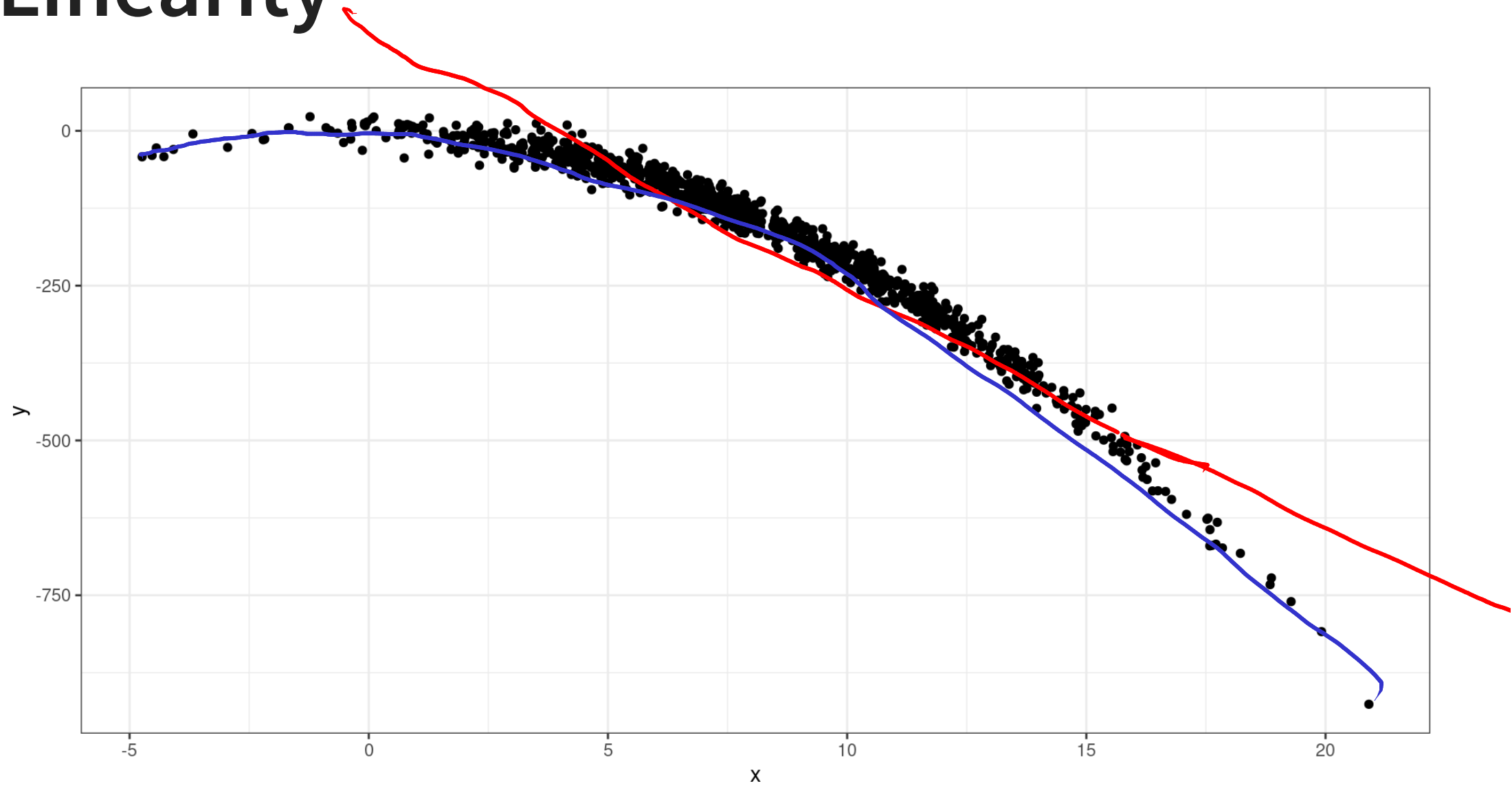
Constant Variance



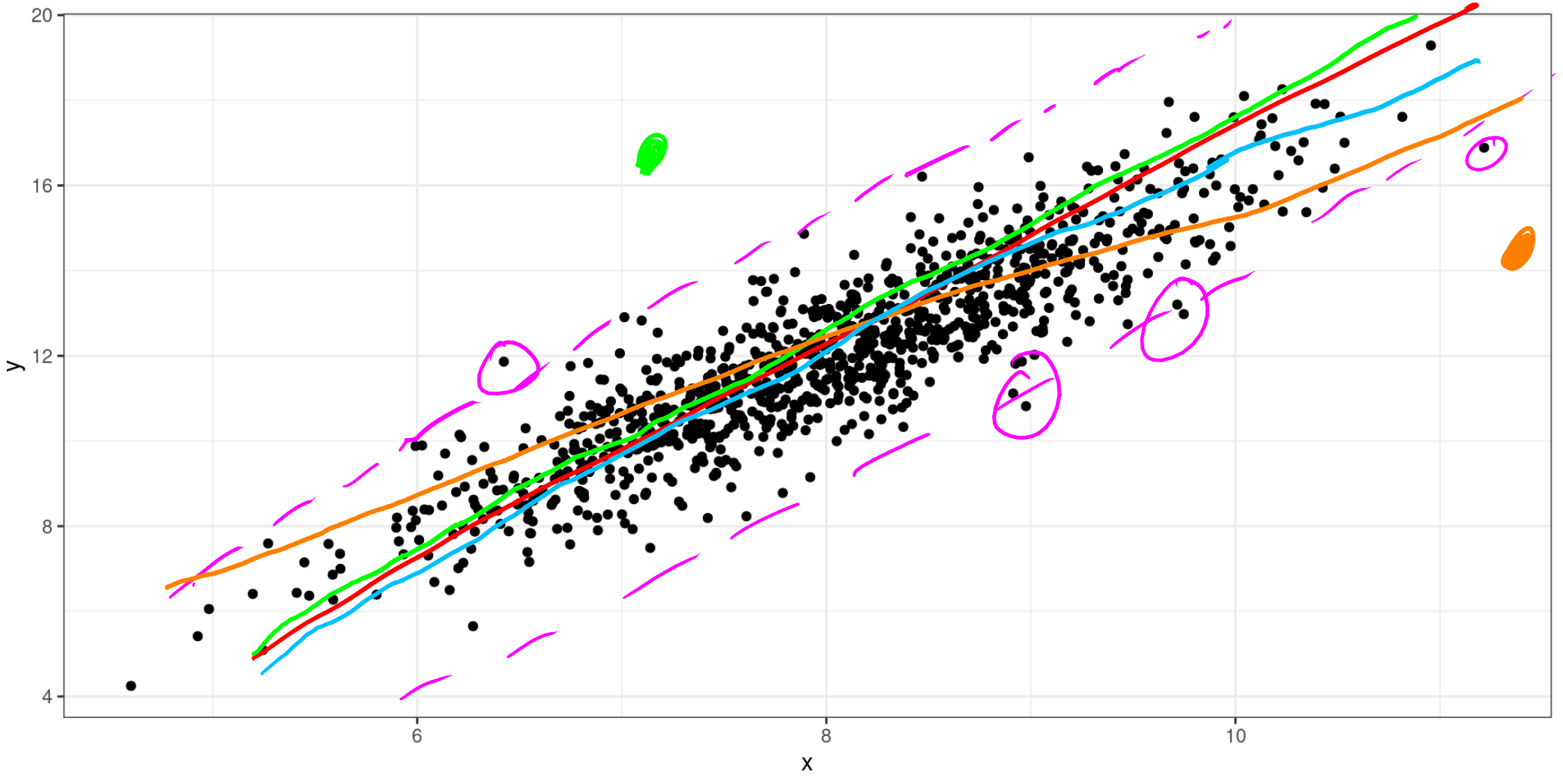
Linearity



Linearity



No Outliers



Residual Analysis

A residual analysis is used to assess the validity of the assumptions.

$$r_i = y_i - \hat{y}_i$$

Studentized

Standardized

hat values

Cook's Distance