Linear Regression

Learning Objectives

- Matrix Formulation
- Multiple Linear Regression
- Model Assumptions

Matrix Formulation

Matrix Version of Model

$$\underbrace{Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i}$$

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

$$\underset{1\times 1}{Y_i} = \underset{1\times 2}{X_i^{T}}\ \underset{2\times 1}{\beta} + \underset{1\times 1}{\varepsilon_i}$$

- Y_i: Outcome Variable
- $\boldsymbol{X}_i = (1, X_i)^{\mathrm{T}}$: Predictors

  2×1
- $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$: Coefficients
- ϵ_i: error term

Data Matrix Formulation

$X_i = (1 \quad x_i)^T$

For n data points

$$\boldsymbol{Y} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$n \times 1 \quad n \times 2 \; 2 \times 1 \quad n \times 1 \quad = \quad n \times 1$

- $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^{\mathrm{T}}$: Outcome Variable
- $\boldsymbol{X} = (X_1, \cdots, X_n)^{\mathrm{T}}$: Predictors ($X_1$: $p \times 1$; $\boldsymbol{X}$: $p \times n$)
- $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$: Coefficients
- $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^{\mathrm{T}}$: Error terms

Least Squares Formula

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

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

Estimates

$$\hat{\beta} = (\overset{2\times n}{\boldsymbol{X}^{T}}\,\overset{n\times 2}{\boldsymbol{X}})^{-1}\,\overset{2\times n}{\boldsymbol{X}^{T}}\,\overset{n\times 1}{\boldsymbol{Y}} \qquad 2\times 1$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \bar{Y} - \hat{\beta}_1 \bar{X} \\ \dfrac{\Sigma (Y_i - \bar{Y})(X_i - \bar{X})}{\Sigma (X_i - \bar{X})^2} \end{pmatrix}$$

Multiple Linear Regression

MLR

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

Continuous Variable

$p+1 << n$

100 99

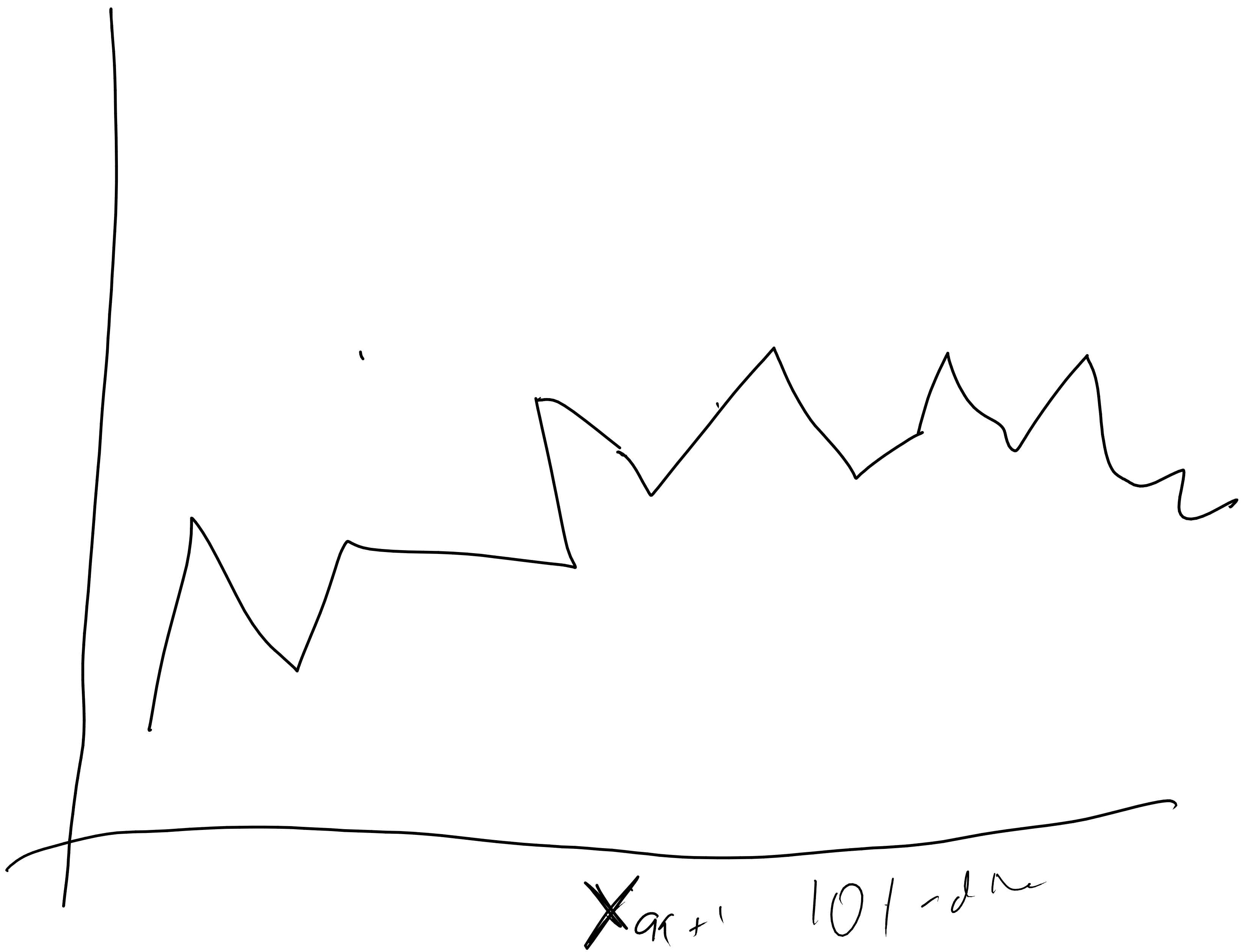
To fit an additional continuous random variable to the model, we will only need to add it to the model:

overfitting

body mass wingspan Beak size

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

As X_1 increase by 1 unit, Y will increase/decrease by an average of β_1 units, adjusting for Beak size


101

Categorical Variable

Species = $\begin{cases} 1 & \text{seagull} \\ 2 & \text{condor} \\ 3 & \text{duck} \end{cases}$

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

$$Y = \beta_0 + \beta_1 (\text{wins}) + \beta_2 D_2 + \beta_3 D_3$$

Species

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

reference category seagull

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

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

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

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

$$X = \begin{pmatrix} 1 \\ \text{Wingspan} \\ \text{Are they condor?} \\ \text{Are they duck?} \end{pmatrix}$$

$$\hat{\beta} = \Big(\underset{4\times n}{X}\ \underset{n\times 4}{X^T}\Big)^{-1} \Big(\underset{4\times n}{X}\ \underset{n\times 1}{y}\Big) = \underset{4\times 1}{\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix}}$$

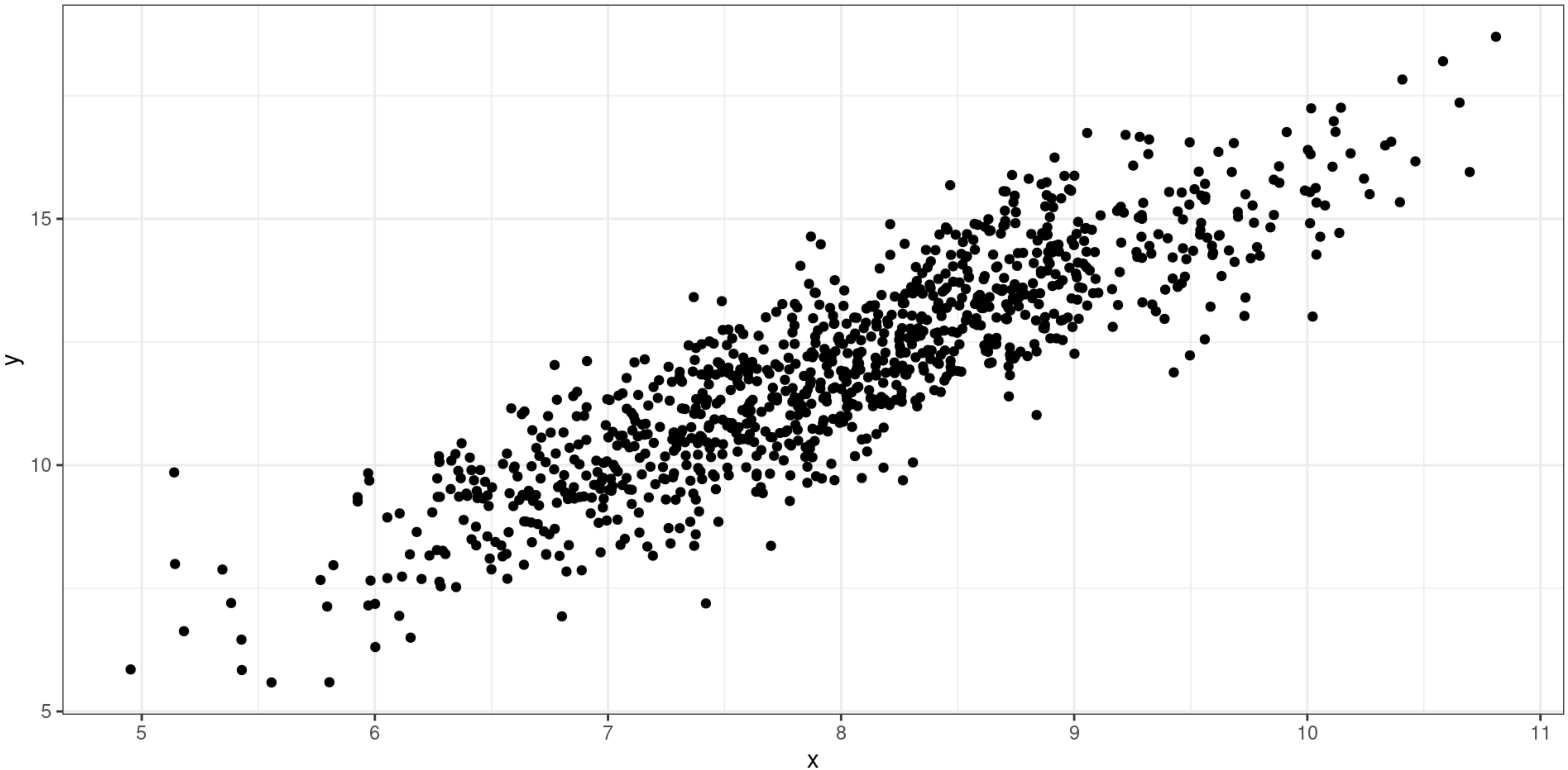
Model Assumptions

Model

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

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

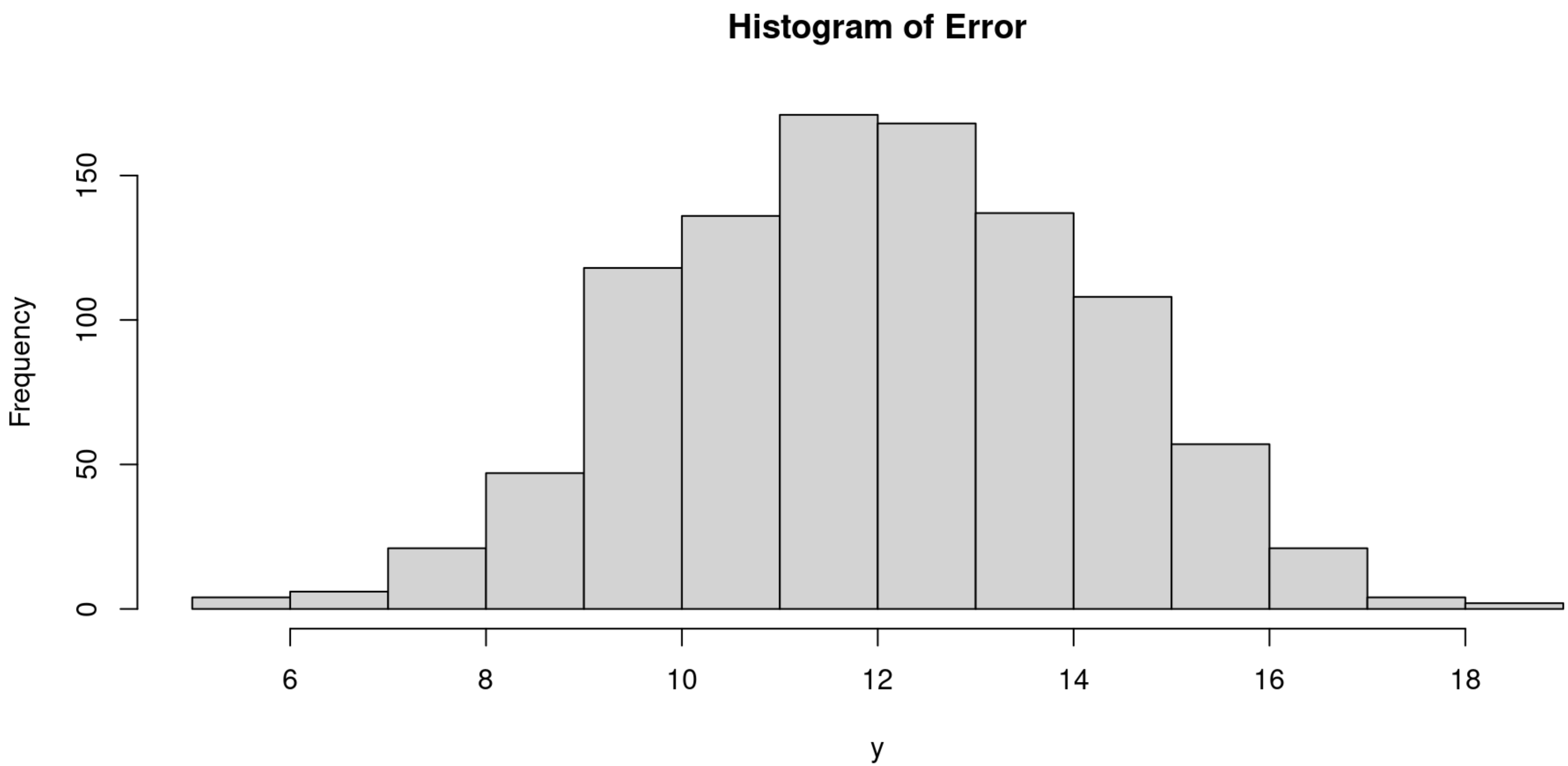
Model Scatter Plot



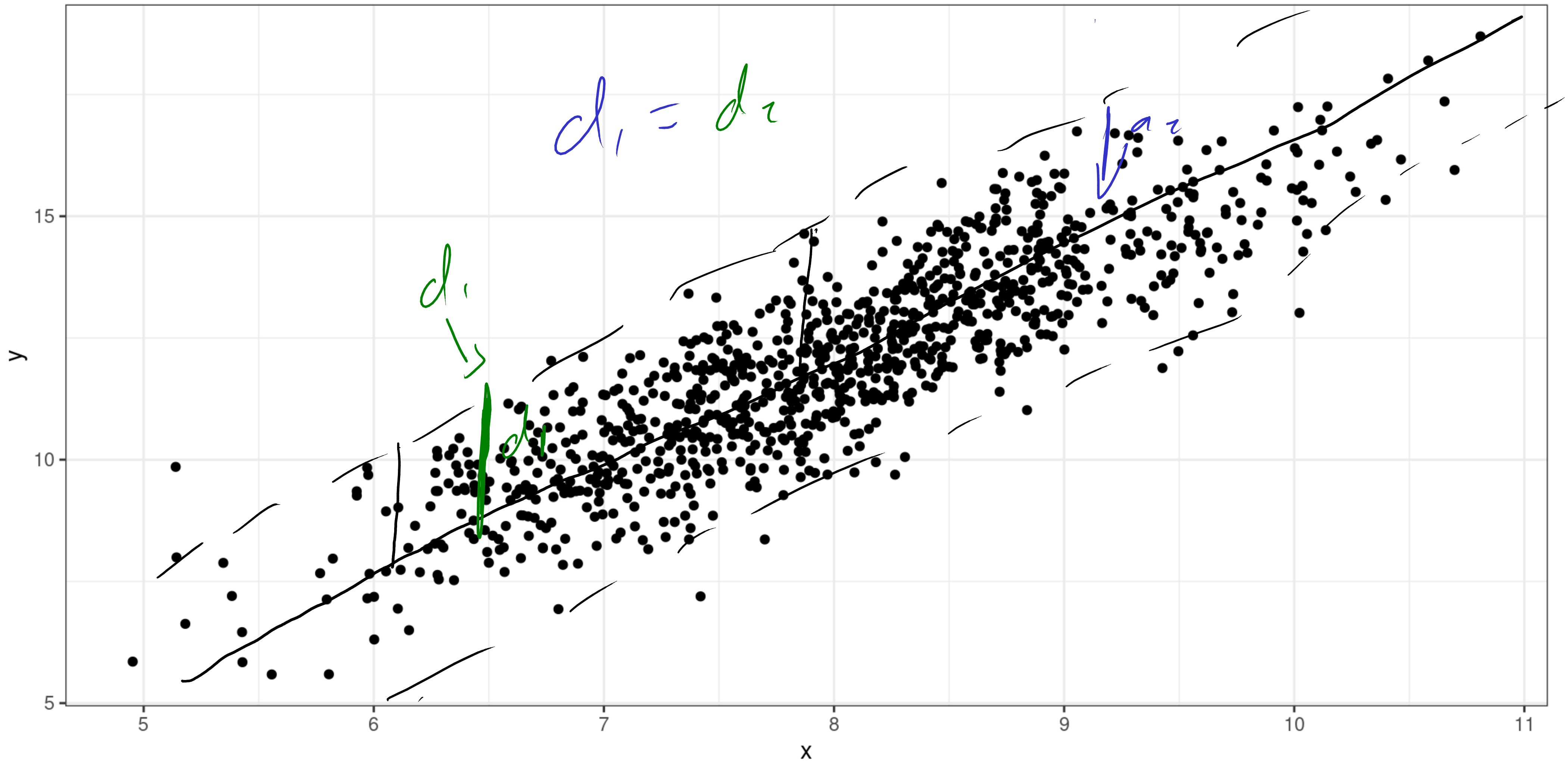
Model Assumptions

- Errors are normally distributed $r_i = y_i - \hat{y}$
- Constant Variance
- Linearity
- Independence
- No outliers

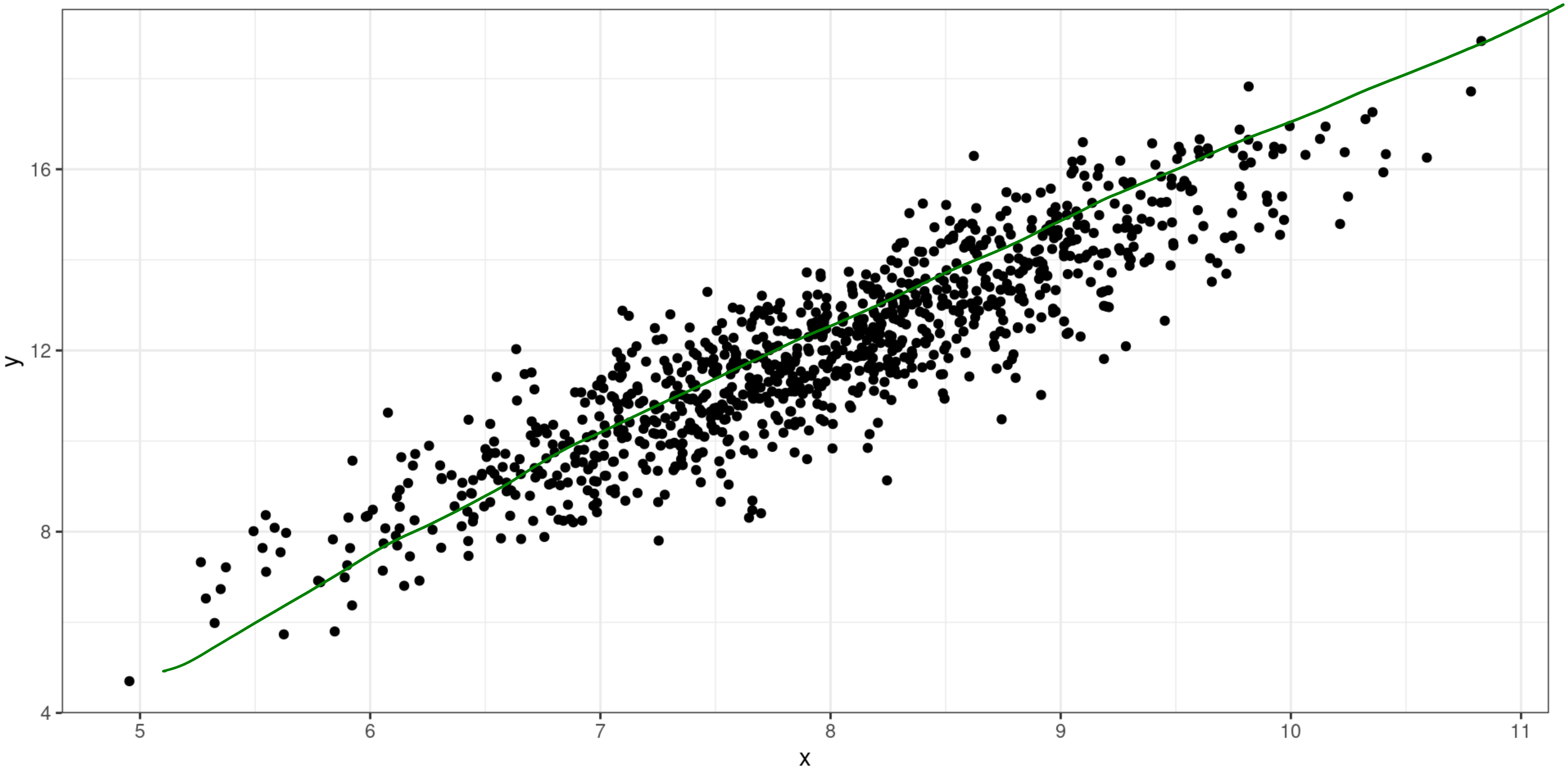
Errors Normally Distributed



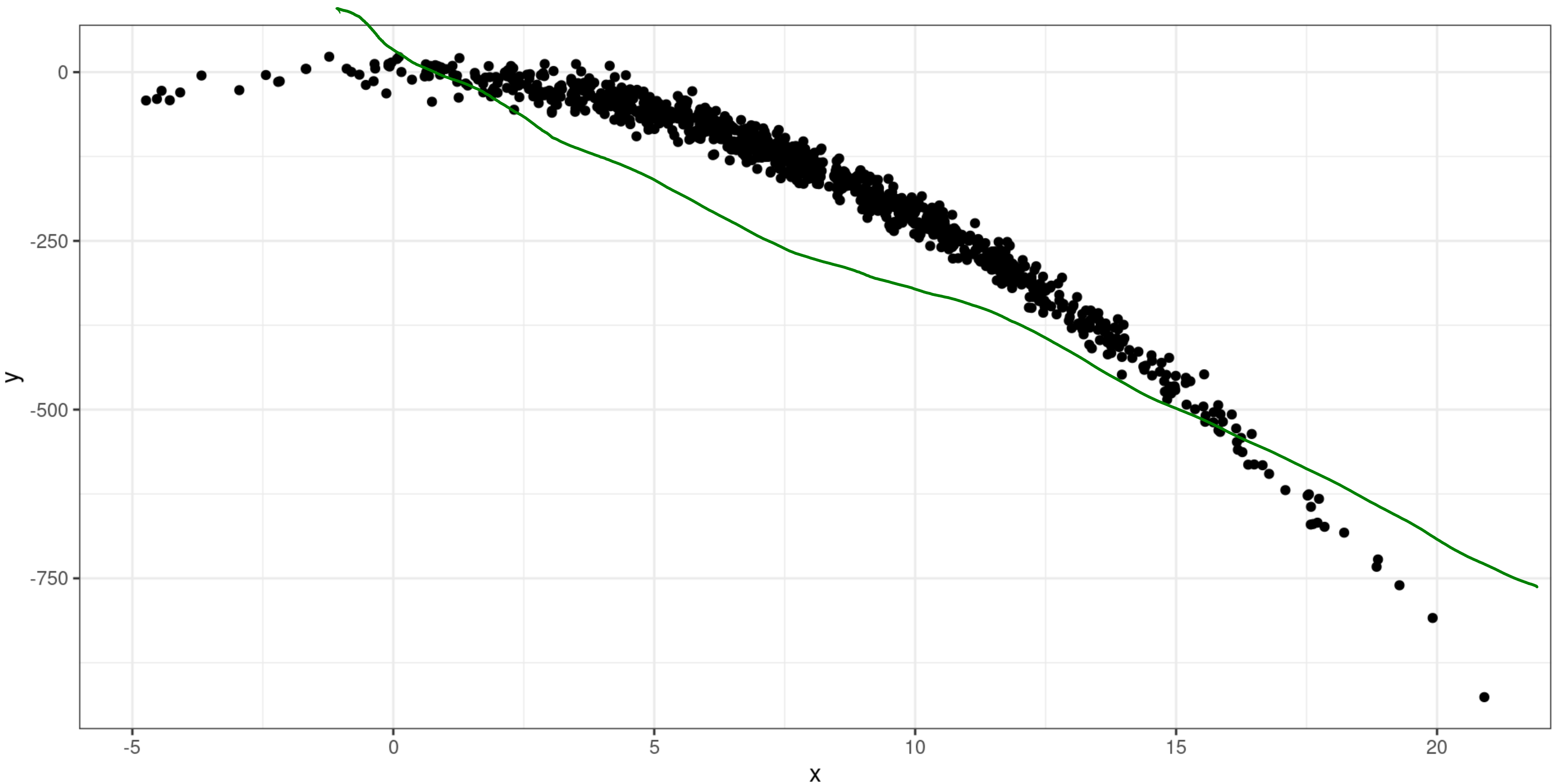
Constant Variance



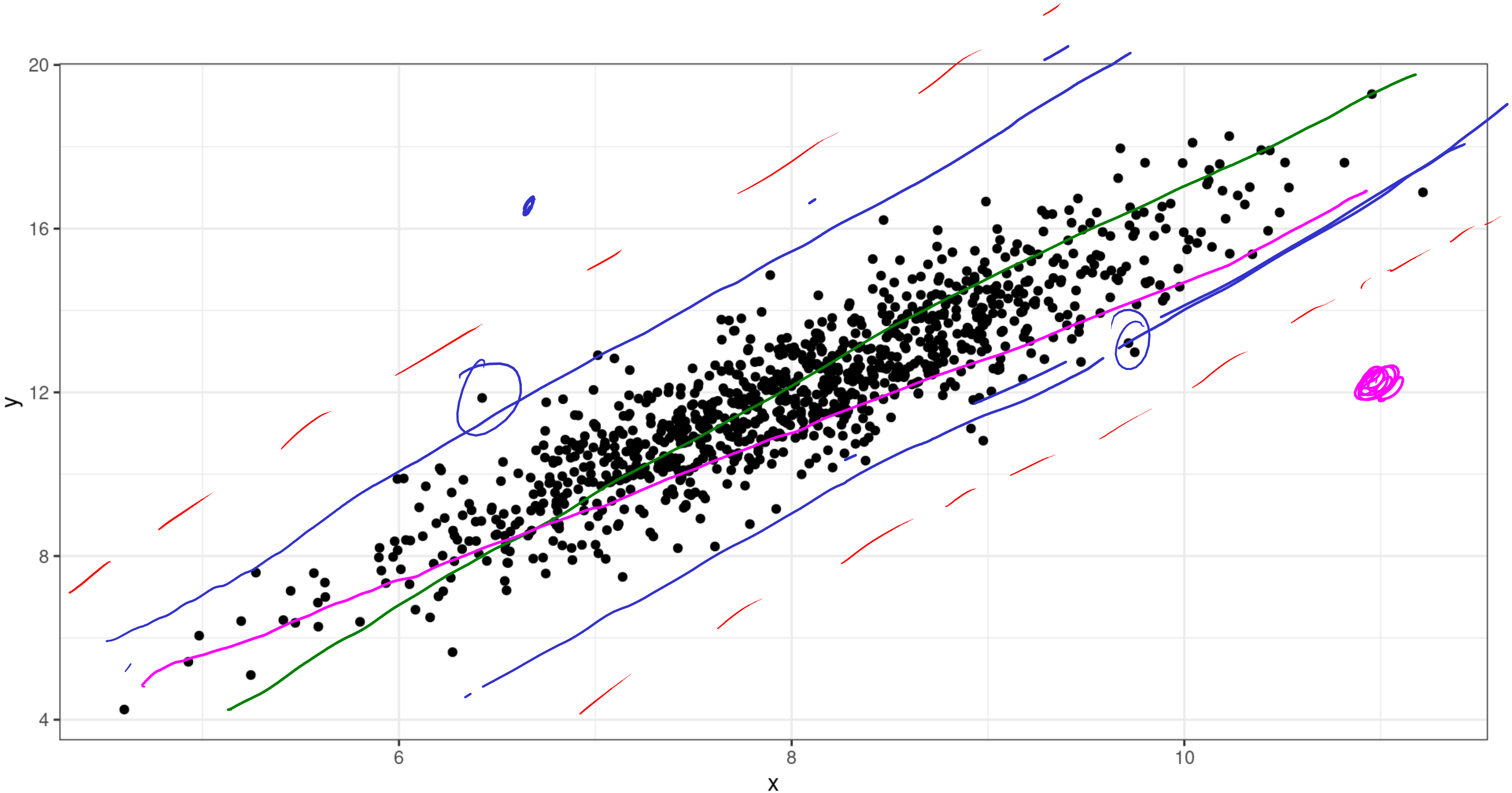
Linearity



Linearity



No Outliers



Residual Analysis

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

Residuals = $Y_i - \hat{Y}_i$

Standardized

Studentized

Hat values

Cook's Distance