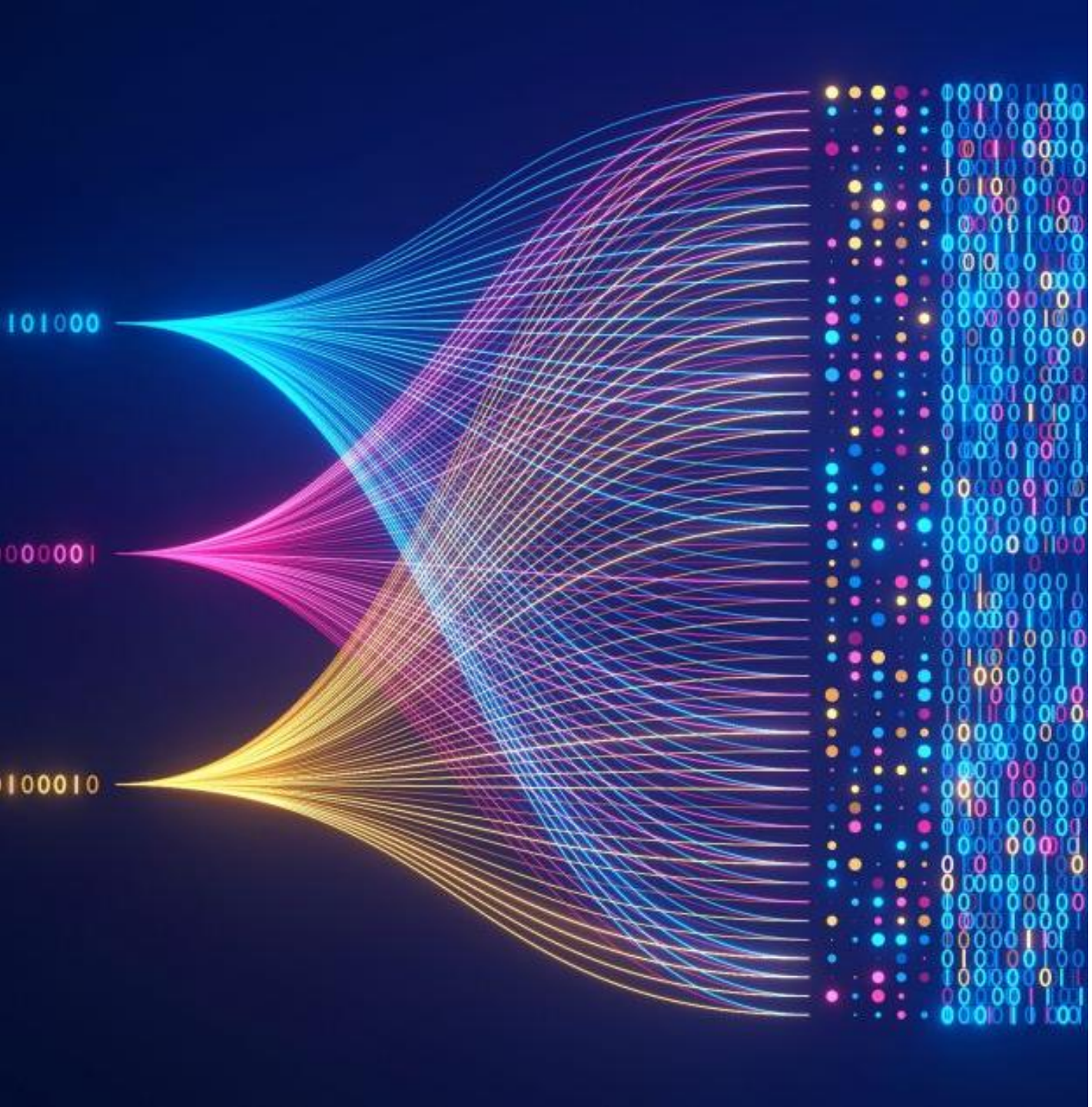


Machine Learning with R

Hamidreza Bolhasani

PhD, AI/ML Data Scientist

March 2024

DataBio
we give life to the data!

RStudio®

Table of contents

Covariance

Variance

Gives information of a single variable

$$S_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Covariance

Gives information on the degree to which two variables vary together.

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- When X ↑ and Y ↑ cov (x,y) = pos.
- When X ↓ and Y ↑ cov (x,y) = neg.
- When no constant relationship: cov (x,y) = 0

Covariance Example

	High variance data				Low variance data		
Subject	**x**	**y**	**x error * y error**		**x**	**y**	**X error * y error**
1	101	100	2500		54	53	9
2	81	80	900		53	52	4
3	61	60	100		52	51	1
4	51	50	0		51	50	0
5	41	40	100		50	49	1
6	21	20	900		49	48	4
7	1	0	2500		48	47	9
Mean	51	50			51	50	
Sum of x error * y error :			7000		**Sum of x error * y error :**		28
Covariance:			**1166.67**		**Covariance:**		**4.67**

Correlation & Regression

Correlation

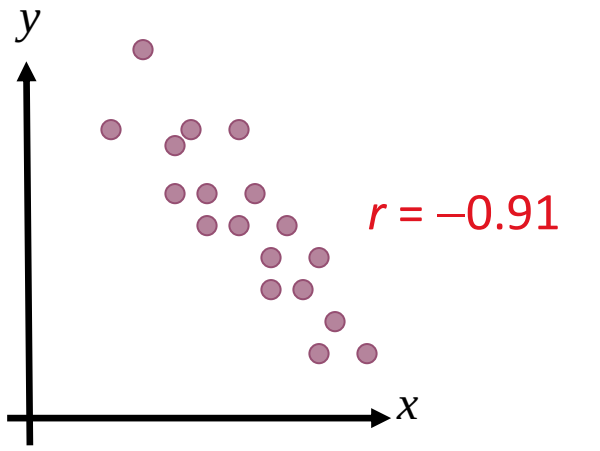
- Is there any relationship between 2 variables (x,y)?
- X is independent (Explanatory) and Y is dependent (Response)
- Correlation ≠ Causation

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\left(N\sum X^2 - \left(\sum X\right)^2\right)\left(N\sum Y^2 - \left(\sum Y\right)^2\right)}}$$

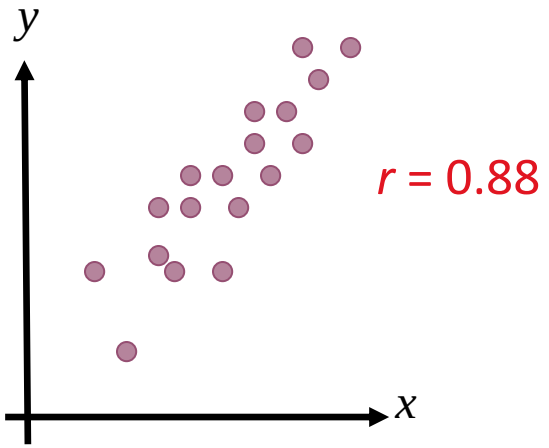
Regression

How well a certain independent variable predict dependent variable?

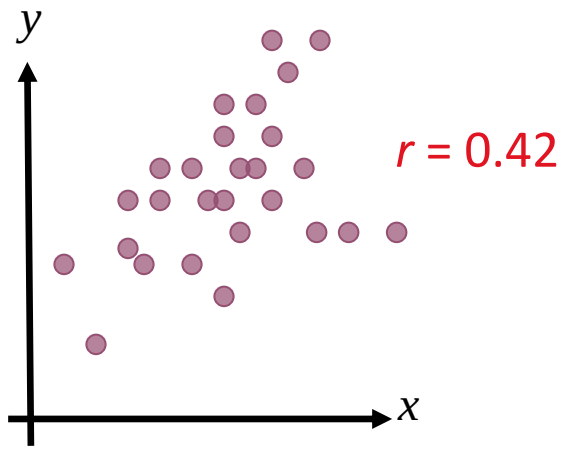
Correlation in Scatter Diagrams



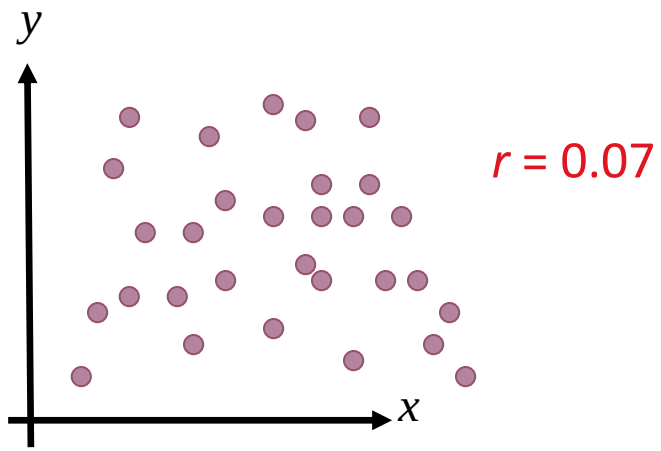
Strong negative correlation



Strong positive correlation



Weak positive correlation



Nonlinear Correlation

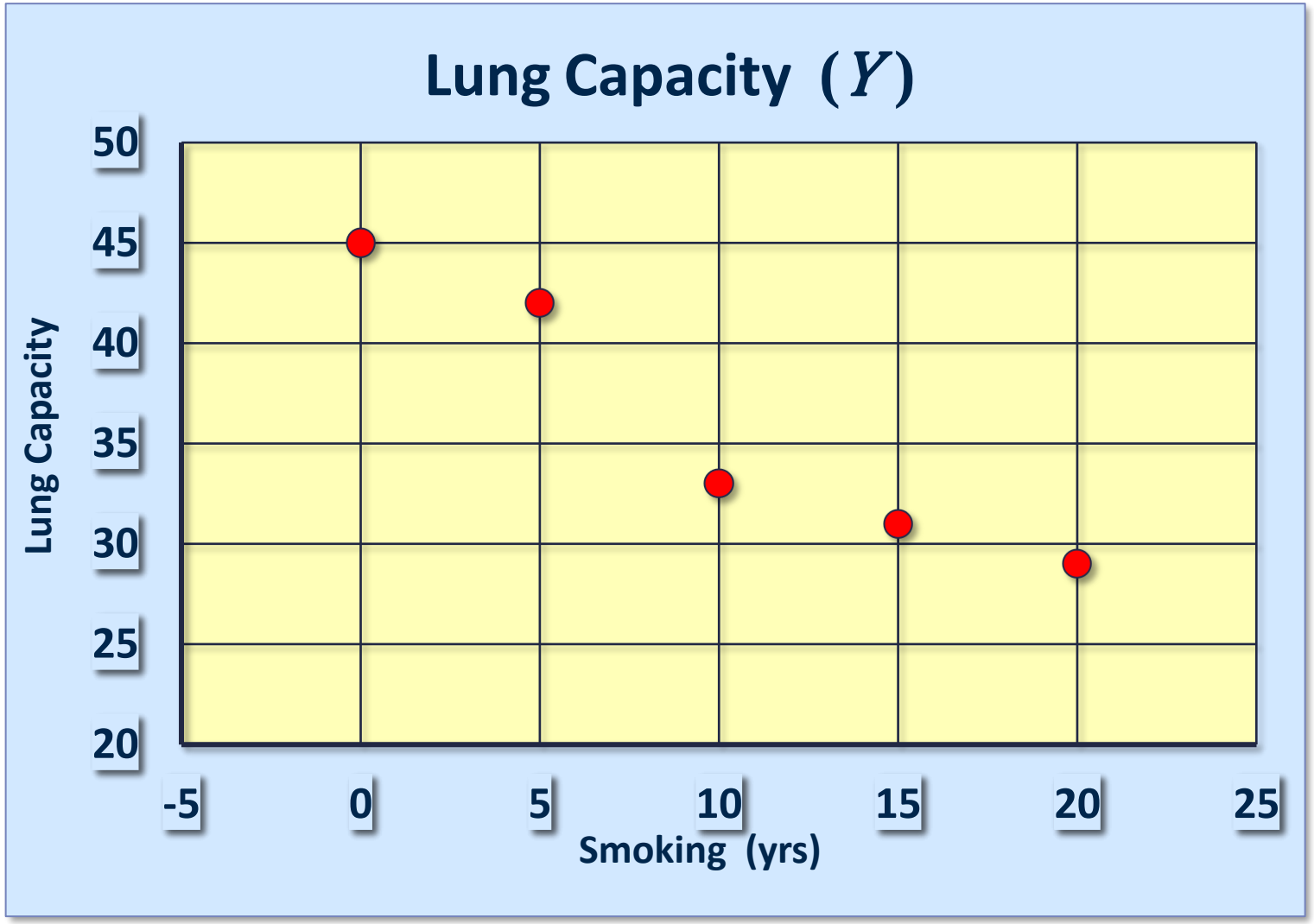
Regression Example

Smoking vs Lung Capacity

N	Cigarettes (X)	Lung Capacity (Y)
1	0	45
2	5	42
3	10	33
4	15	31
5	20	29

Example Analysis

Smoking vs Lung Capacity



$$S_{xy} = \frac{1}{4}(-215) = -53.75$$

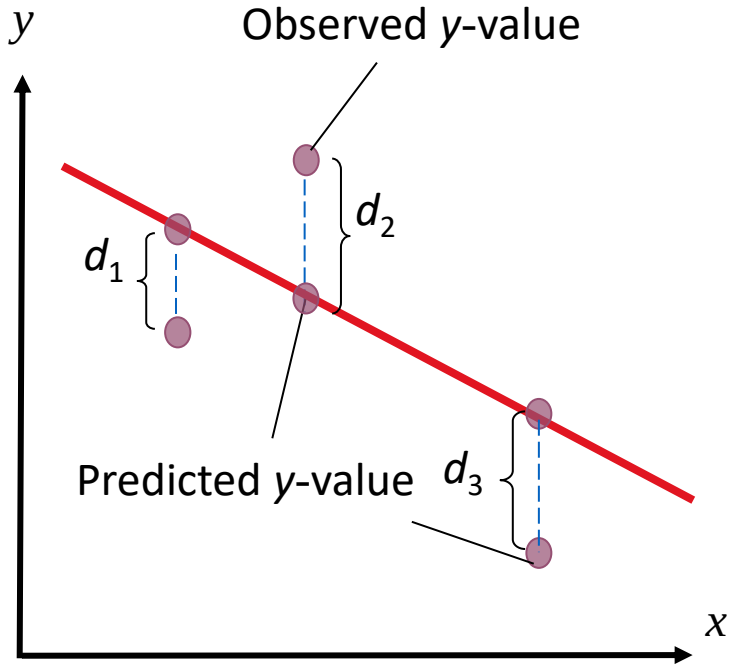
When smoking is above its group means, lung capacity tends to be below its group mean.

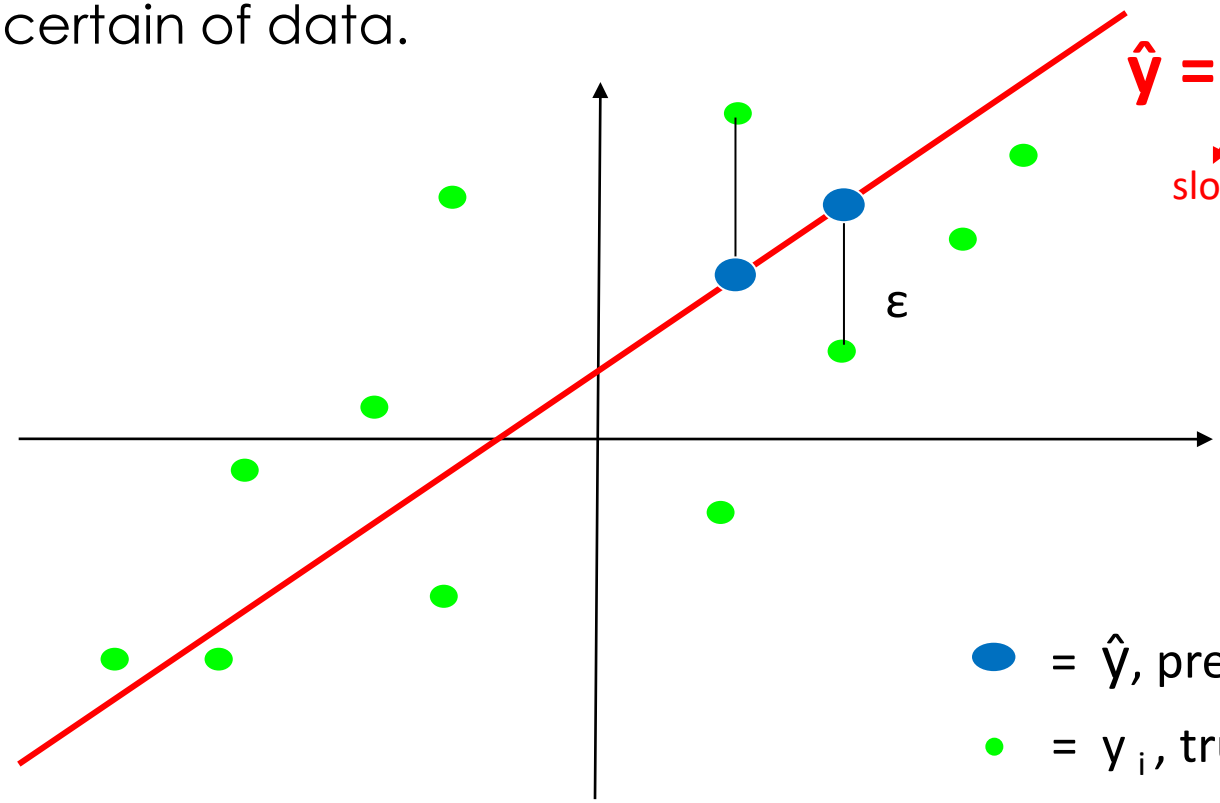
$$r_{xy} = -0.96$$

Greater smoking exposure implies greater likelihood of lung damage.

Regression

- The process of predicting variable Y using variable X.
- Tells us how values in Y changes as a function of changes in value X.
- Calculates the "best-fit" line for a certain of data.





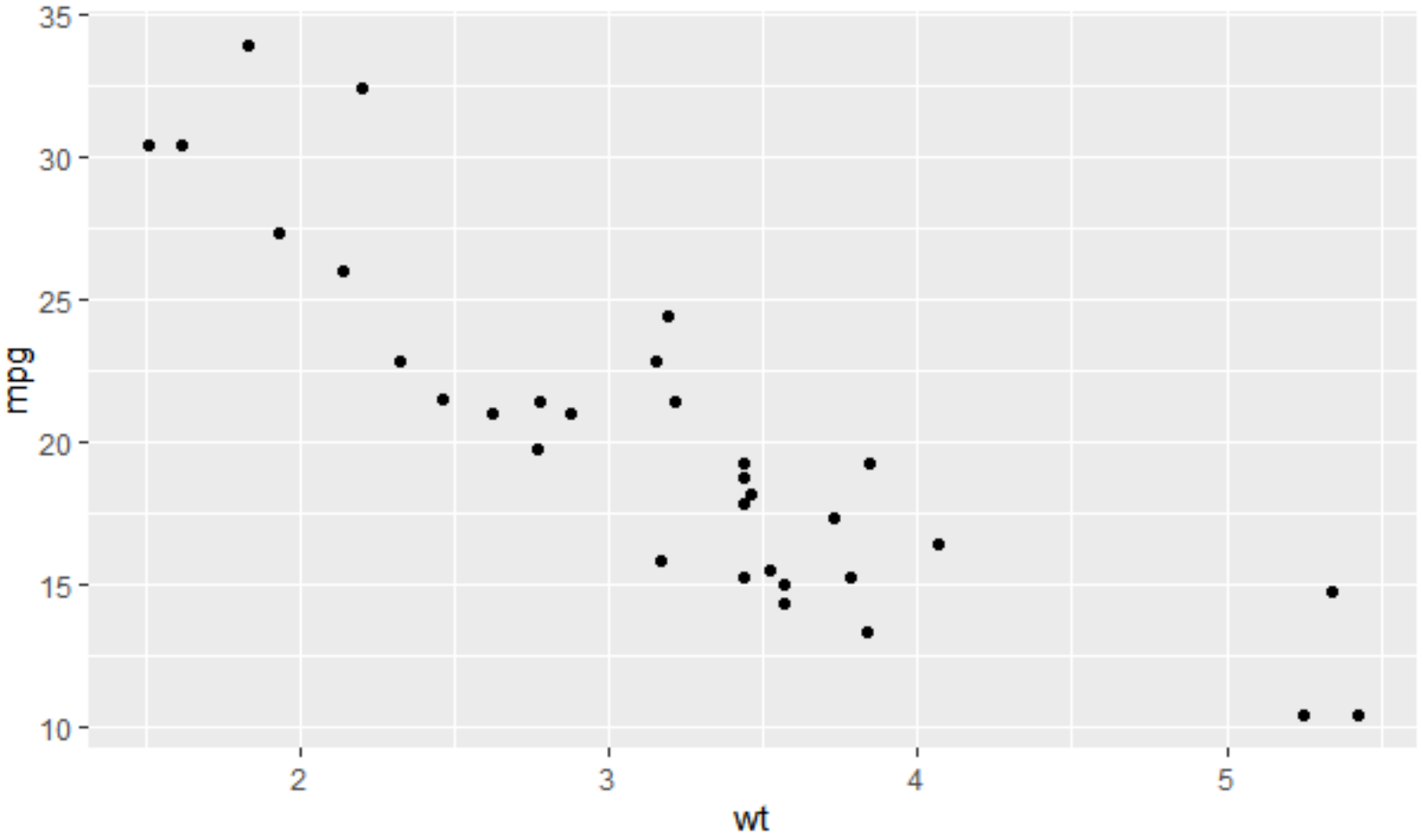
= ŷ, predicted value

= y_i, true value

ε = residual error

Regression: Case Study in R

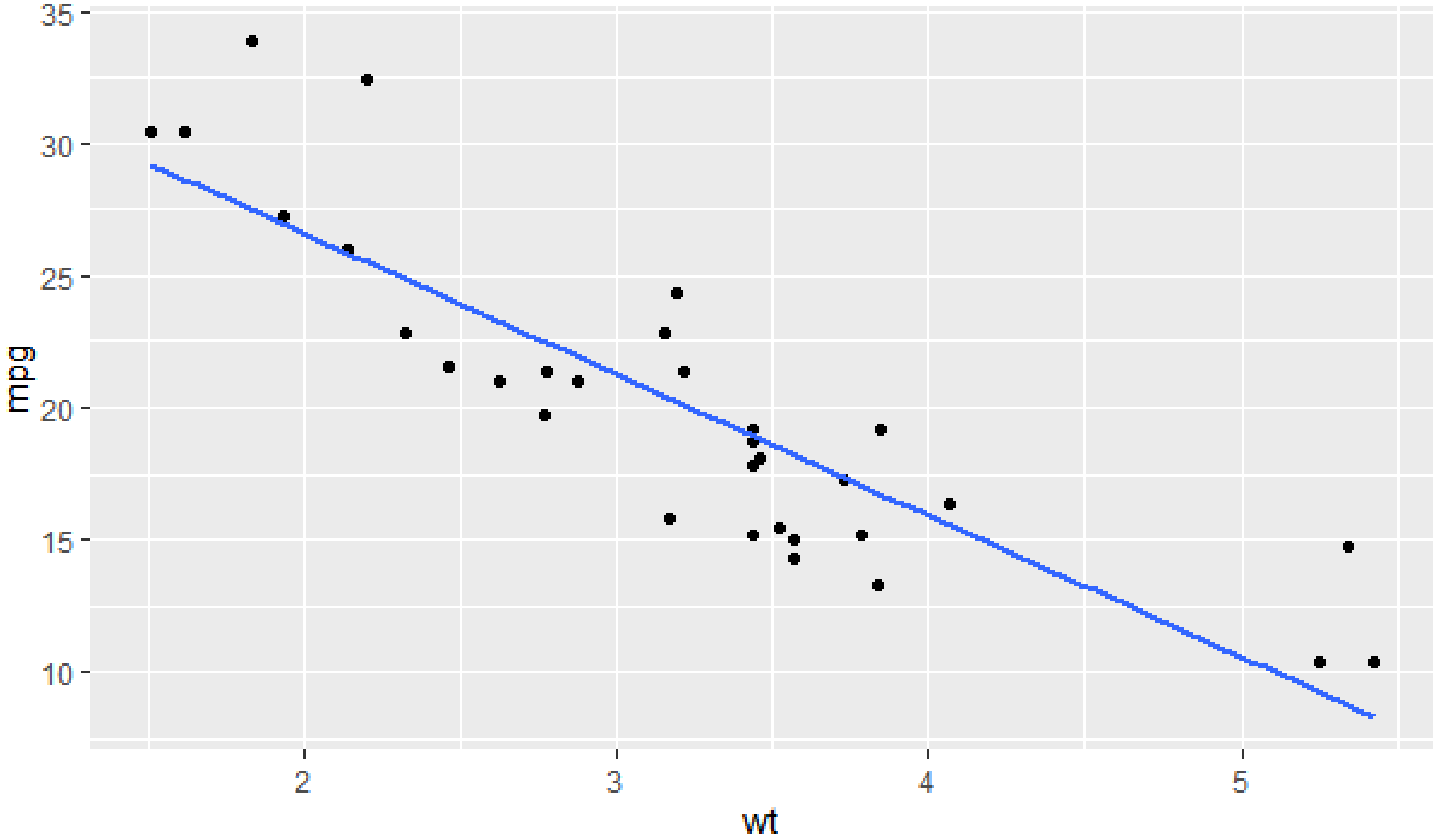
```
library(ggplot2)
ggplot(data=mtcars, aes(x=wt, y=mpg))+geom_point()
```



- Data = mtcars
- wt: weight
- Mpg: miles per gallon
- S(x,y) = -5.11
- r(x,y) = -0.86

Regression: Case Study in R

```
ggplot(data=mtcars, aes(x=wt, y=mpg))+geom_point()+geom_smooth(method="lm",se=FALSE)
lm(data=mtcars, mpg ~ wt)
```



- Intercept = 37
- Slope = -5.3

Thanks!

Hamidreza Bolhasani
bolhasani@gmail.com
March 2024