Introduction to linear models

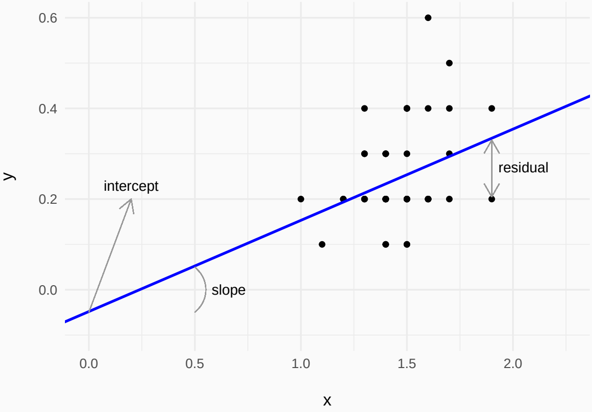
Francisco Rodríguez-Sánchez

https://frodriguezsanchez.net

Our unified regression framework (GLM)

$$y_i = a + bx_i + \varepsilon_i$$

$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$



Data

y = response variable

x = predictor

Parameters

a = intercept

b = slope

σ = residual variation

ε = residuals

What's the intercept?

Expected value of y when predictors (x) = 0

If x = 0:

- y = a + b*0

What's the intercept?

Expected value of y when predictors (x) = 0

If x = 0:

- y = a + b*0
- y = a

What's the slope?

How much y increases (or decreases) when x increases in 1 unit

If we have model

$y = 0.5 + 2 * x$

- If x = 10 -> $y = 0.5 + 2 * 10 = 20.5$

If x increases 1 unit, y increases 2 units

What's the slope?

How much y increases (or decreases) when x increases in 1 unit

If we have model

$y = 0.5 + 2 * x$

- If x = 10 -> $y = 0.5 + 2 * 10 = 20.5$
- If x = 11 -> $y = 0.5 + 2 * 11 = 22.5$

If x increases 1 unit, y increases 2 units

Slopes can be negative

If we have model

$y = 0.5 - 2 * x$

- If x = 10 -> $y = 0.5 - 2 * 10 = -19.5$

If x increases 1 unit, y decreases 2 units

Slopes can be negative

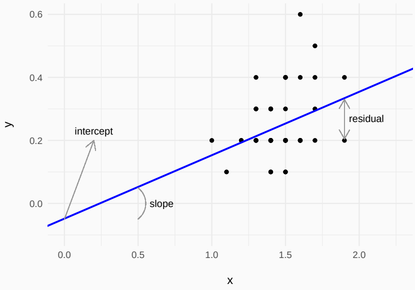
If we have model

$y = 0.5 - 2 * x$

- If x = 10 -> $y = 0.5 - 2 * 10 = -19.5$
- If x = 11 -> $y = 0.5 - 2 * 11 = -21.5$

If x increases 1 unit, y decreases 2 units

What are residuals?



How far points fall from the regression line

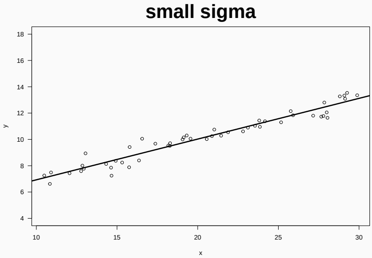
Difference between observed values and values predicted by model (regresssion line)

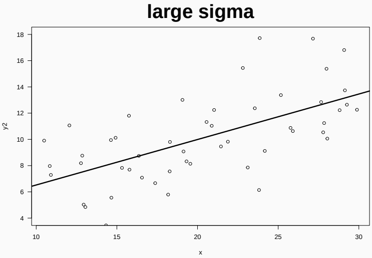
If sigma is large, residuals are larger

$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$

If sigma is larger:

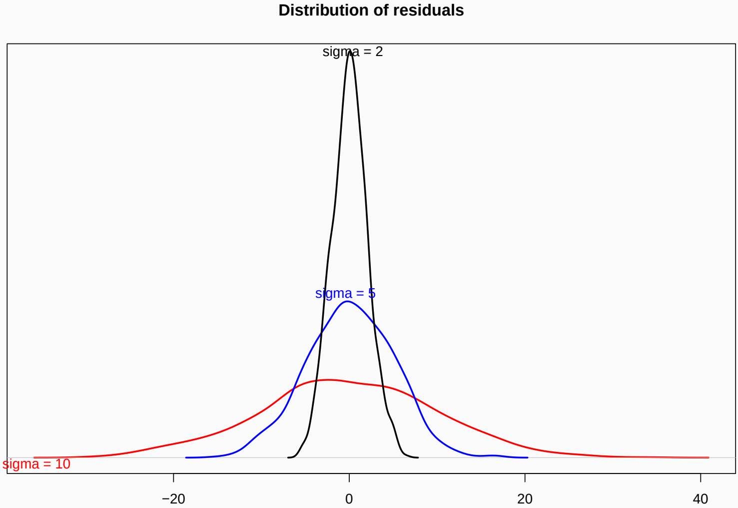
- points farther from regression line
- larger difference of observed - predicted values





Residual variation (sigma) is the Std. Dev. of residuals

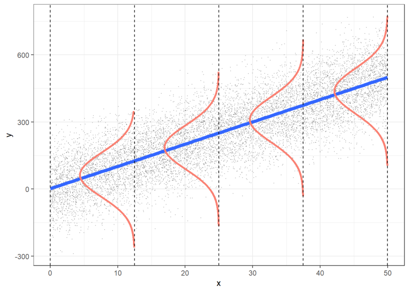
$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$



In a general linear model we assume residuals are

$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$

- Normal

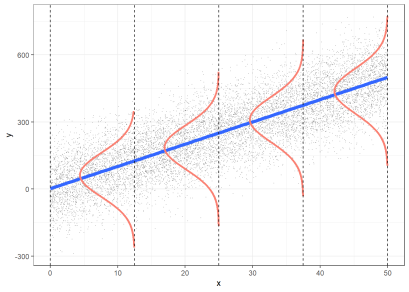


Roback & Legler 2021

In a general linear model we assume residuals are

$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$

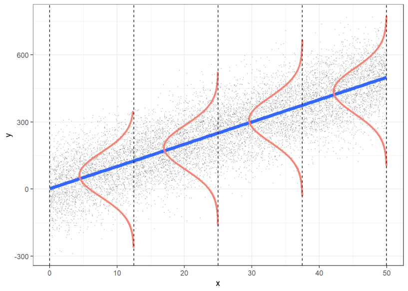
- Normal
- Centred on 0 (no bias)



In a general linear model we assume residuals are

$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$

- Normal
- Centred on 0 (no bias)
- Homogeneous variance (*homoscedasticity*)



Different ways to write same model

$$
\begin{aligned}
y_i &= a + bx_i + \varepsilon_i \\
\varepsilon_i &\sim N\left(0, \sigma^2\right)
\end{aligned}
$$

$$
\begin{aligned}
y_i &\sim N\left(\mu_i, \sigma^2\right) \\
\mu_i &= a + bx_i \\
\varepsilon_i &\sim N\left(0, \sigma^2\right)
\end{aligned}
$$

Quiz

https://pollev.com/franciscorod726