Applied Data Analytics

Statistics — Measures for bivariate data

Ordinary Least Squares (OLS): Intuition

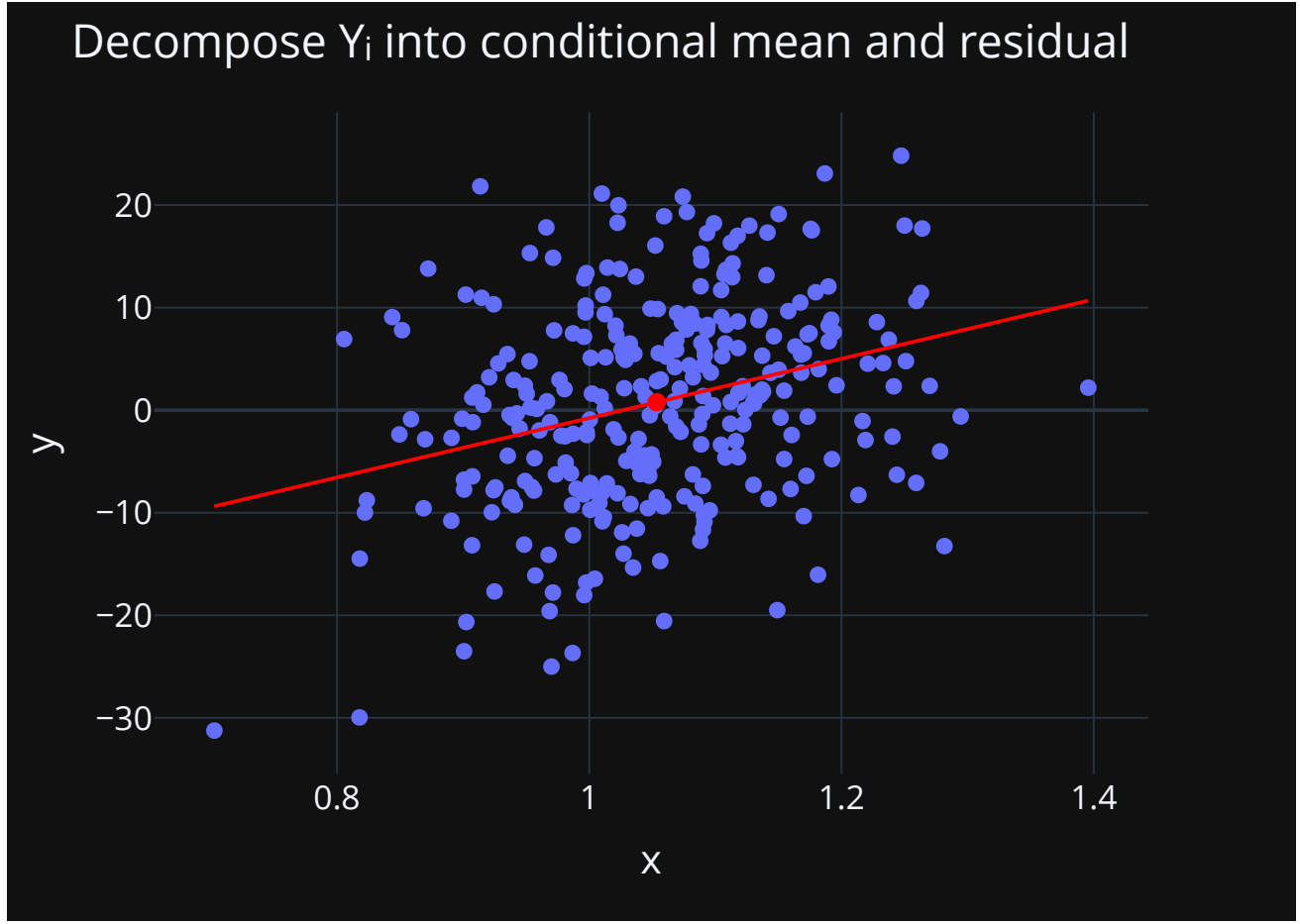
Hans-Martin von Gaudecker

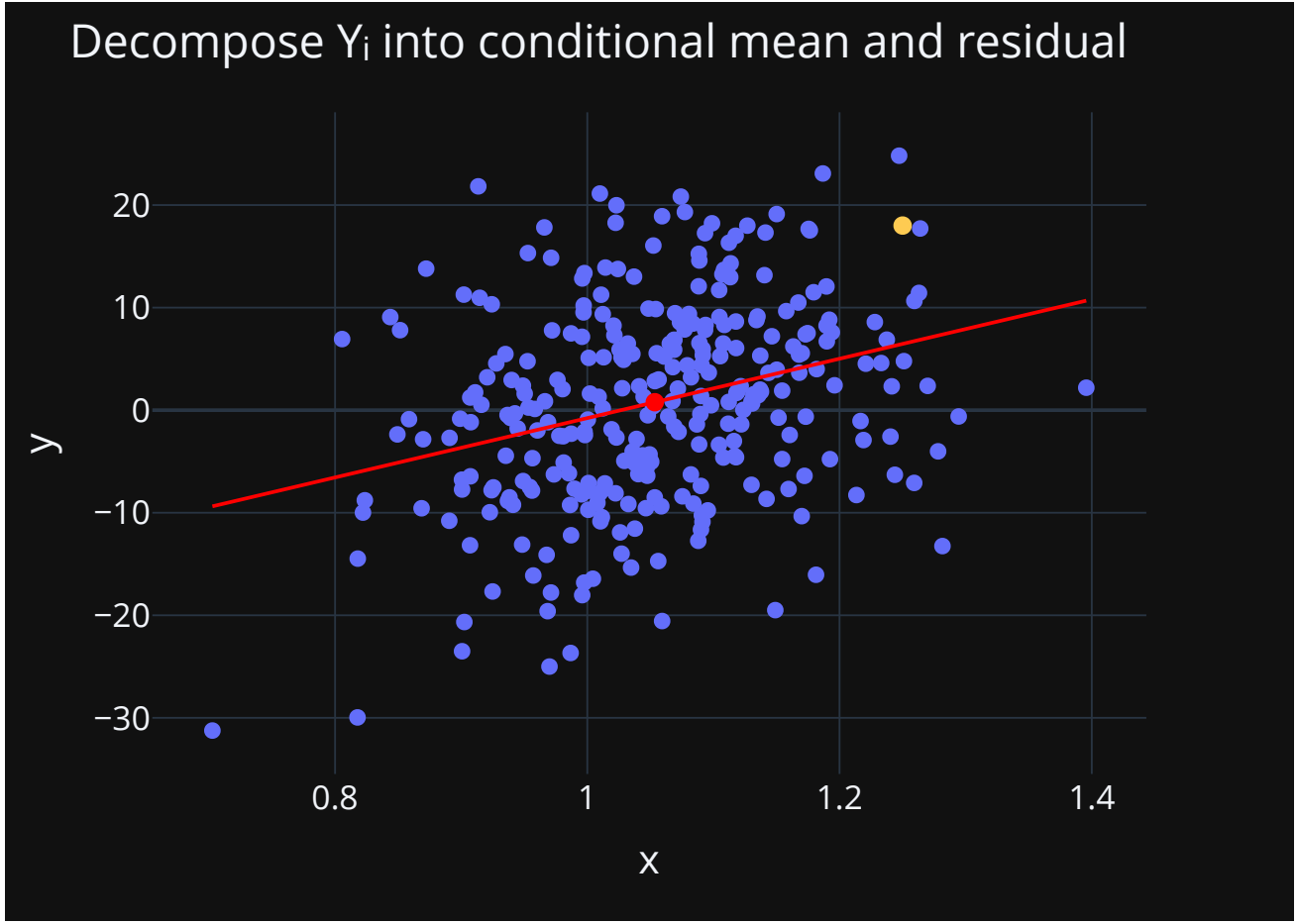
Linear conditional mean function

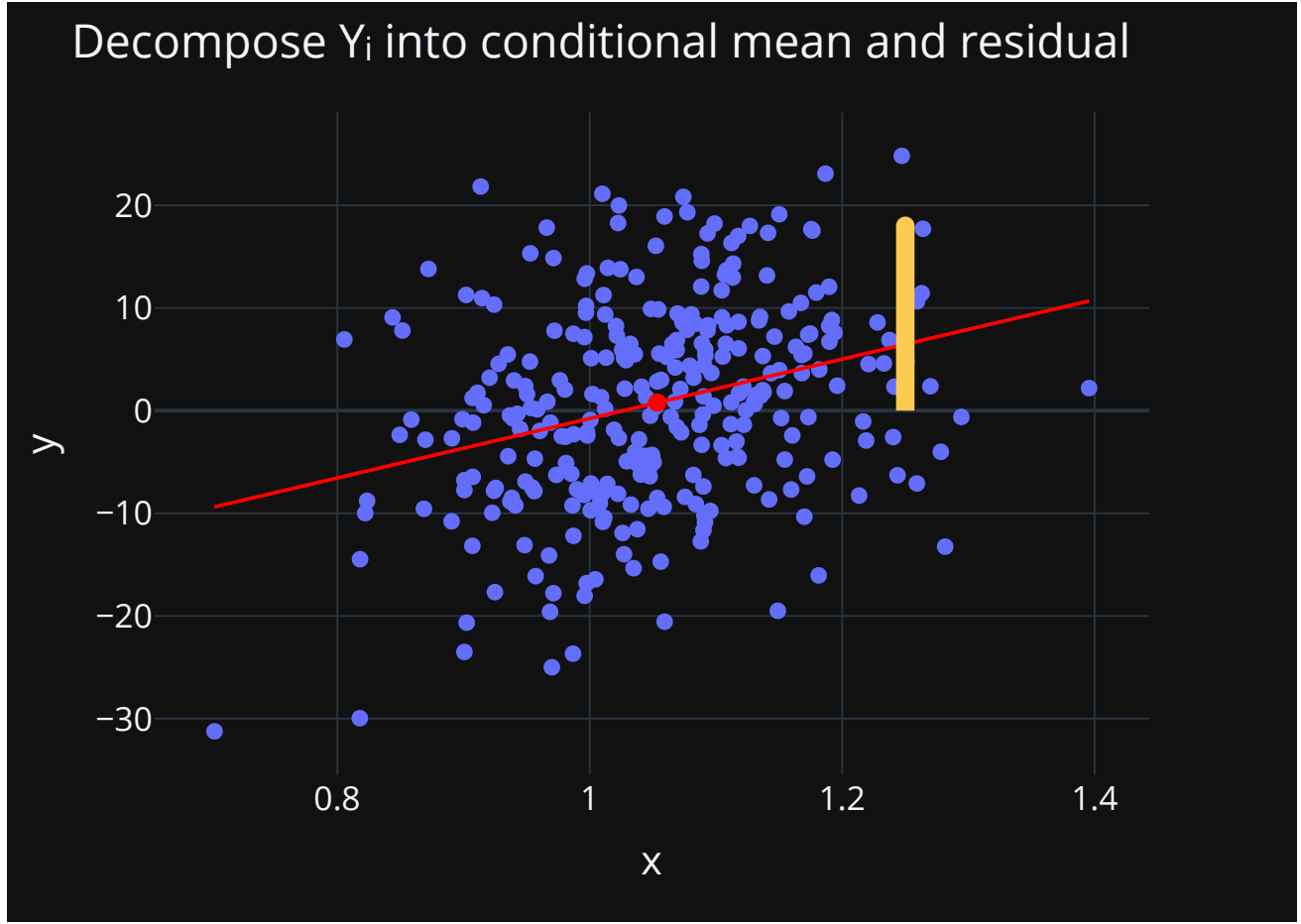
- Given that I know X, what is the mean value of Y?
- Linear relationship:

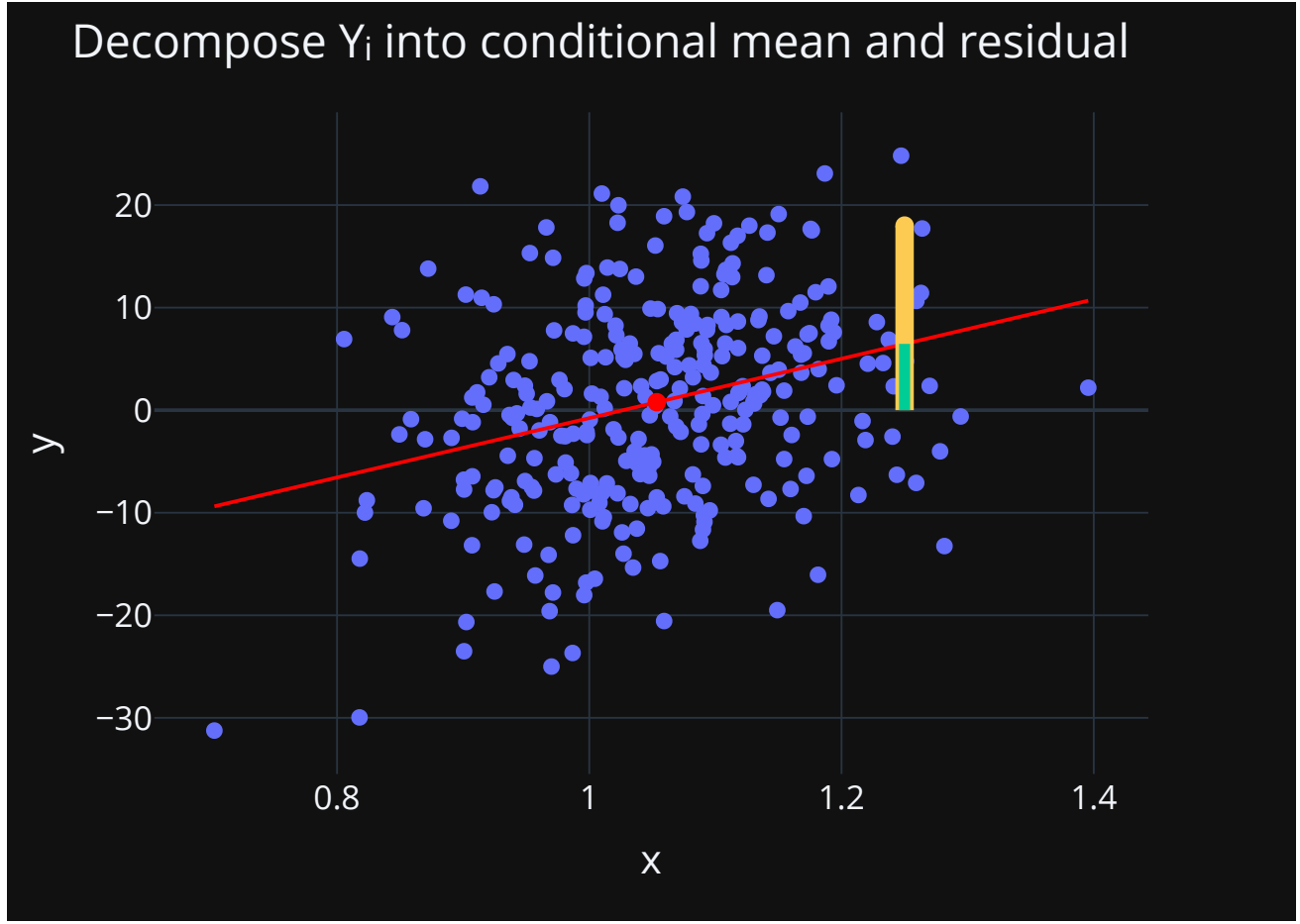
$$Y_i = \beta_0 + \beta_1 \cdot X_i + U_i$$

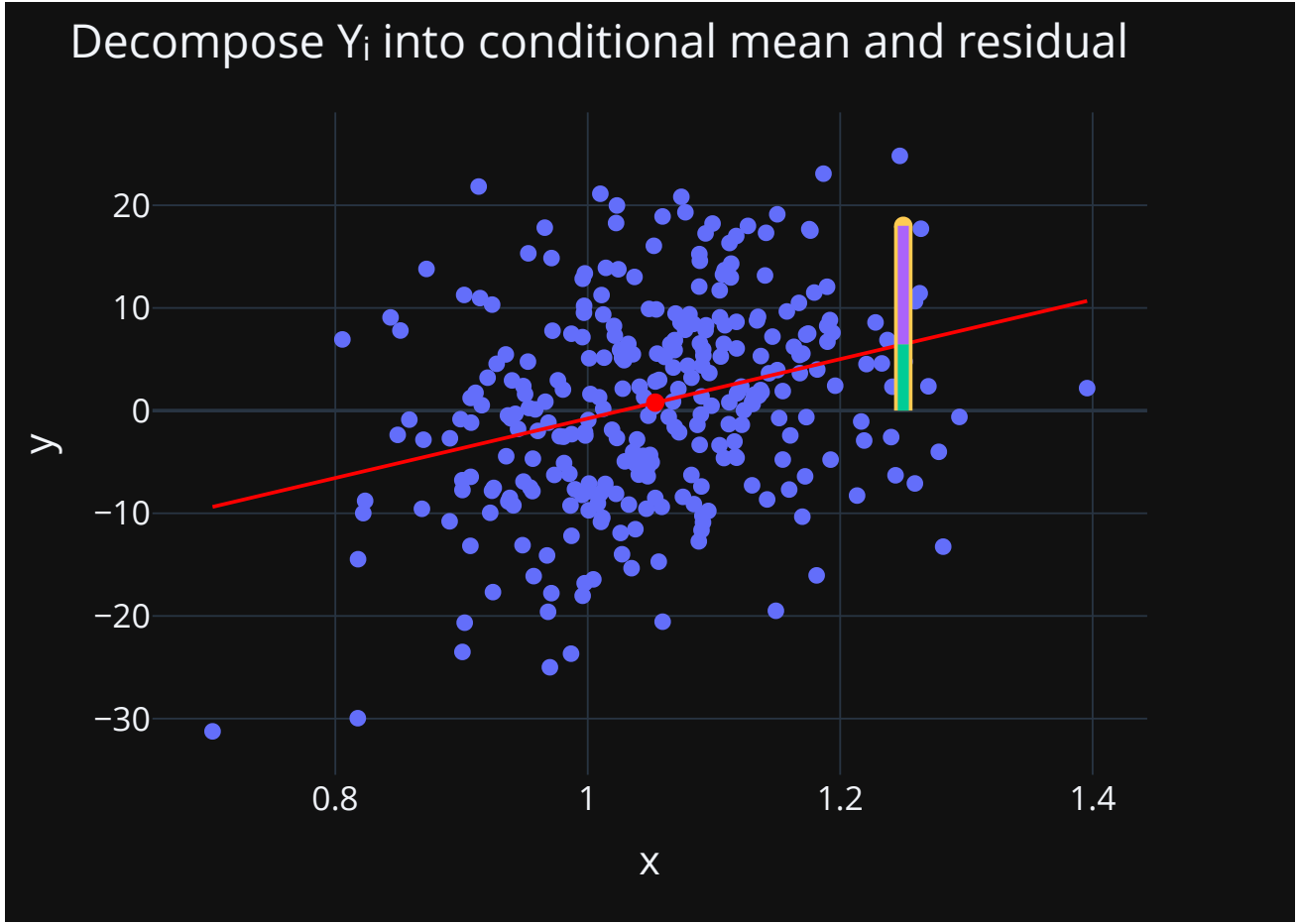
- Divides each Y_i into two parts:
  - Systematic part (conditional mean): $\beta_0 + \beta_1 \cdot X_i$
  - Residual / unobserved part: U_i

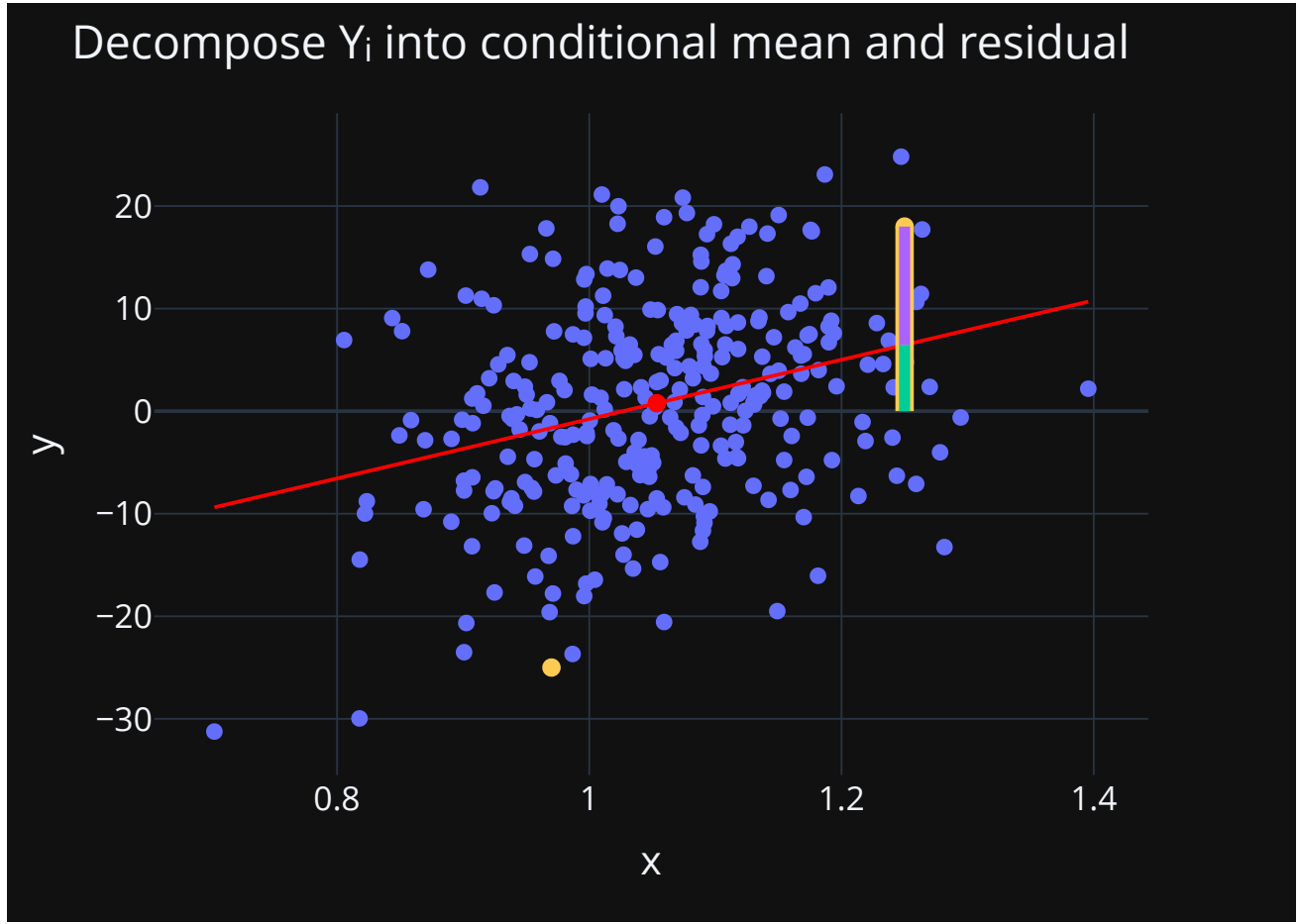

Decompose Y_i into conditional mean and residual
20
10
0
−10
−20
−30
y
0.8
1
1.2
1.4
x

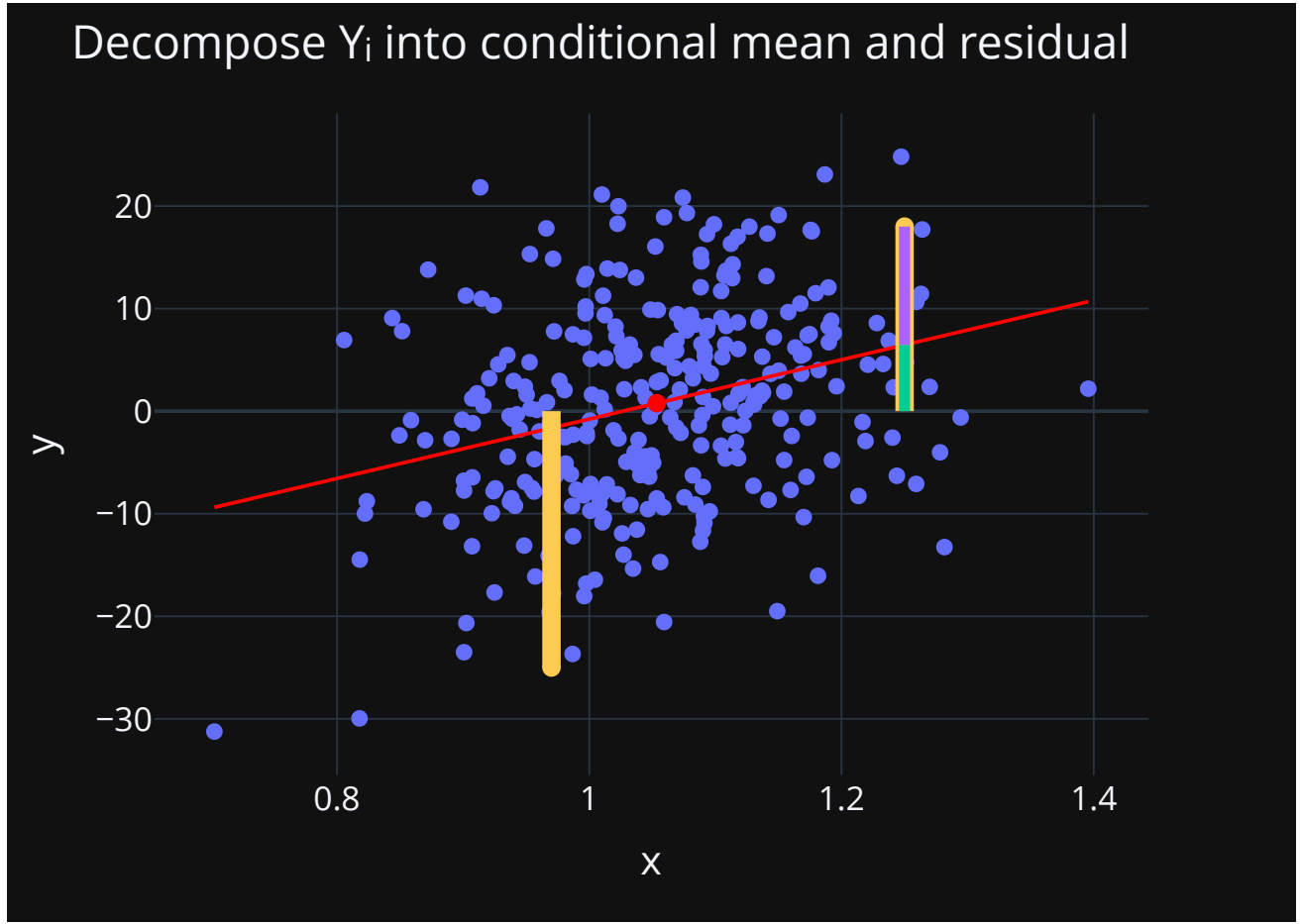

Decompose Yᵢ into conditional mean and residual
20
10
0
−10
−20
−30
y
0.8
1
1.2
1.4
x

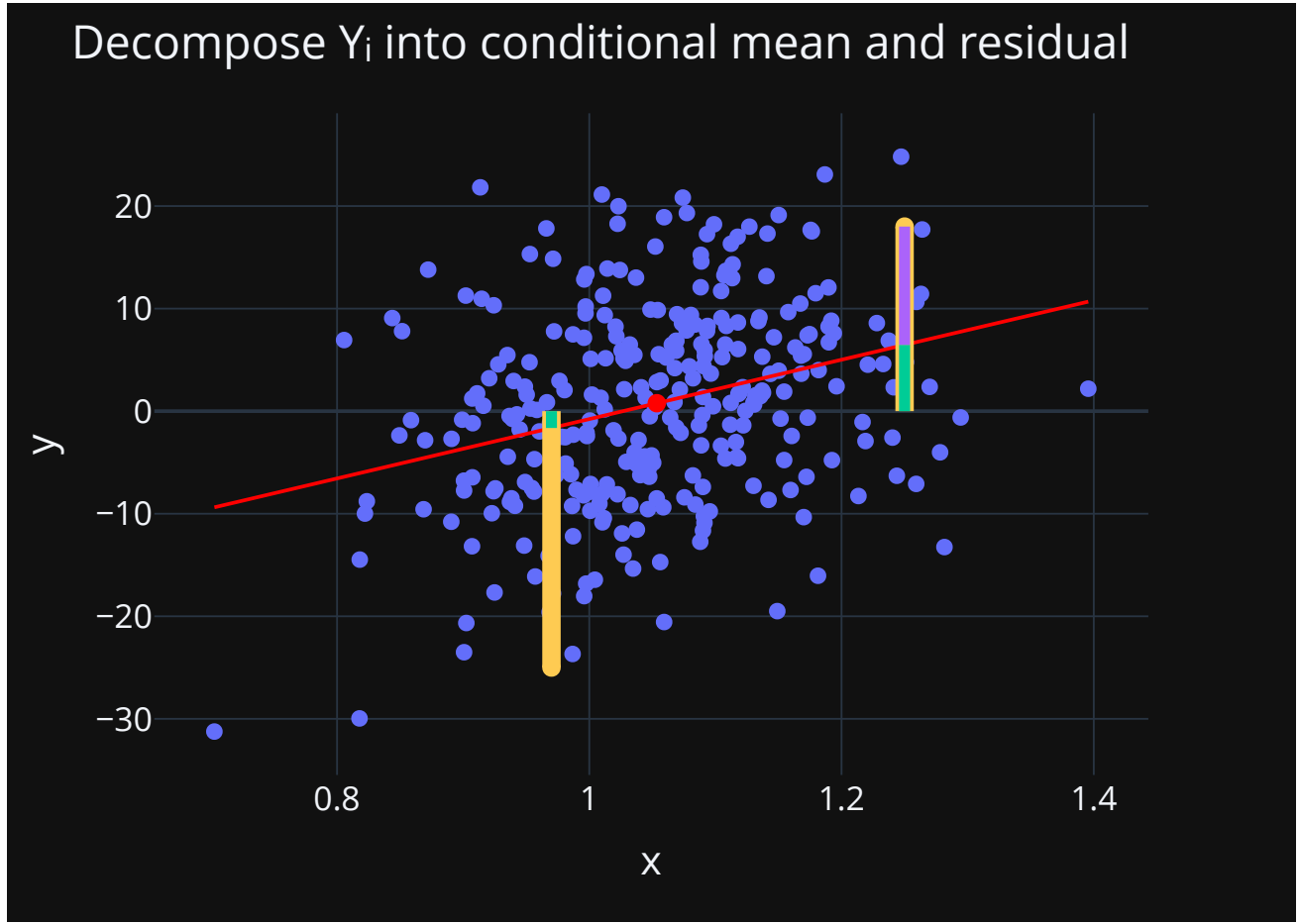

Decompose Yᵢ into conditional mean and residual
20
10
0
−10
−20
−30
y
0.8
1
1.2
1.4
x

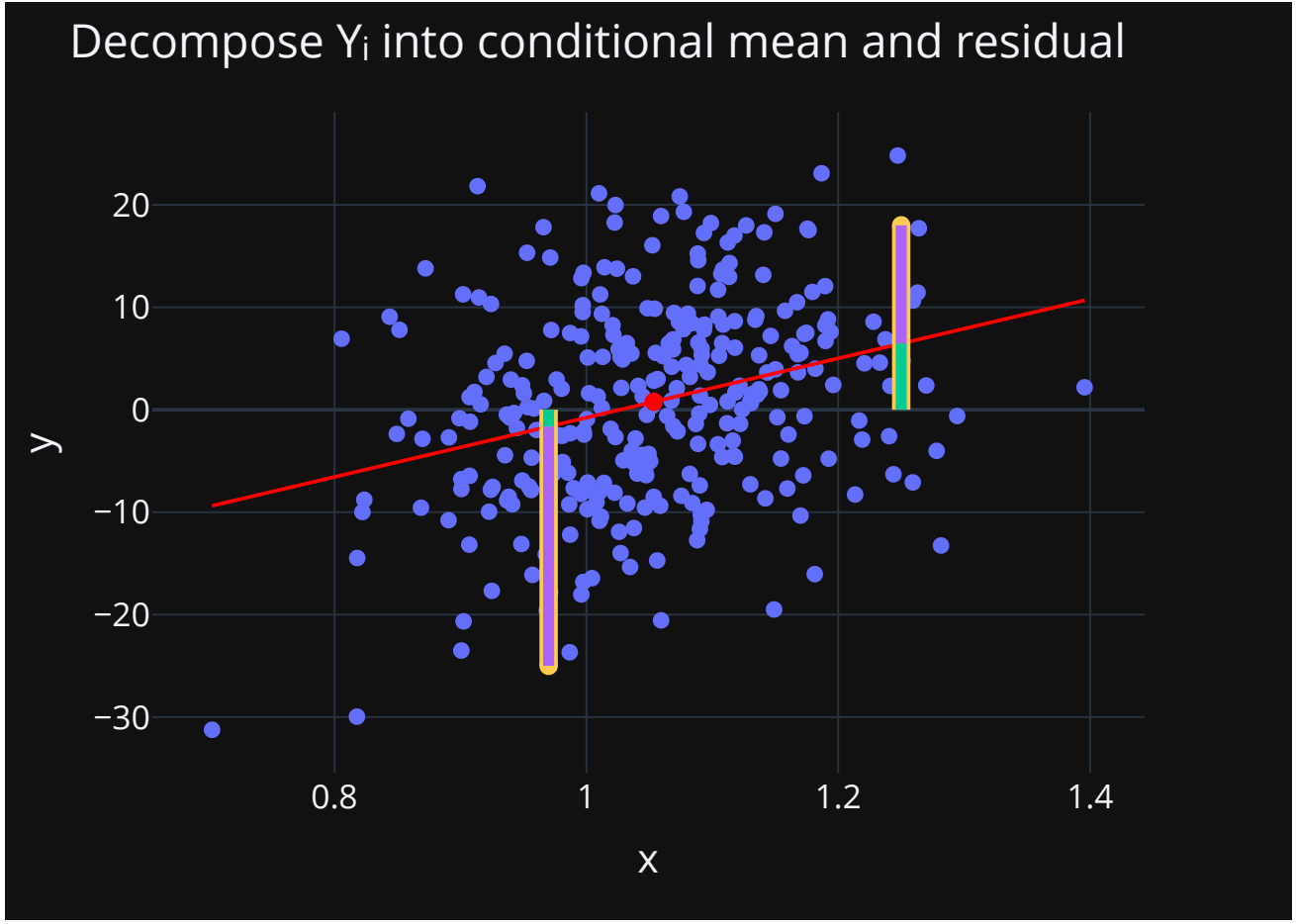

Decompose Yᵢ into conditional mean and residual
20
10
0
−10
−20
−30
y
0.8
1
1.2
1.4
x


Decompose Yᵢ into conditional mean and residual
20
10
0
−10
−20
−30
y
0.8
1
1.2
1.4
x


Decompose Yᵢ into conditional mean and residual
20
10
0
−10
−20
−30
y
0.8
1
1.2
1.4
x


Decompose Yᵢ into conditional mean and residual
20
10
0
−10
−20
−30
y
0.8
1
1.2
1.4
x


Decompose Y_i into conditional mean and residual
20
10
0
−10
−20
−30
y
0.8
1
1.2
1.4
x


Decompose Yᵢ into conditional mean and residual
20
10
0
−10
−20
−30
y
0.8
1
1.2
1.4
x

OLS: Setup

Rewrite

$$Y_i = \beta_0 + \beta_1 \cdot X_i + U_i$$

as

$$U_i = Y_i - \beta_0 - \beta_1 \cdot X_i$$

and pick β_0, β_1 as to minimize a total distance of the U_i

OLS: Setup

"pick β_0, β_1 as to minimize a total distance of the U_i":

- Total: Sum
- Criterion: Squared distances

$$
\begin{aligned}
(\hat{\beta}_0, \hat{\beta}_1) &= \arg\min_{b_0, b_1} \sum_{i=1}^{n} U_i^2 \\
&= \arg\min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2
\end{aligned}
$$

OLS: Why that setup?

- Among straight lines $a + b \cdot X$, OLS has the smallest in-sample sum of squared vertical residuals, by construction.
- Squaring gives larger residuals more weight in this criterion.
- Statistical optimality for a population requires additional assumptions about how the data were generated.
- Those assumptions and results are covered in later statistics and econometrics courses.

OLS: Hugely important

- Easily extends to multiple X, non-linear relationships, etc.
- Basis for huge amount of other statistical methods
- Will only scratch the surface here
- Large part(s) of econometrics course(s) will be spent on this model