Linear Models

Fernando Racimo, 2023

The two sides of statistics

Probability Theory

"What can we say about the data generated by a given process?"

The two sides of statistics

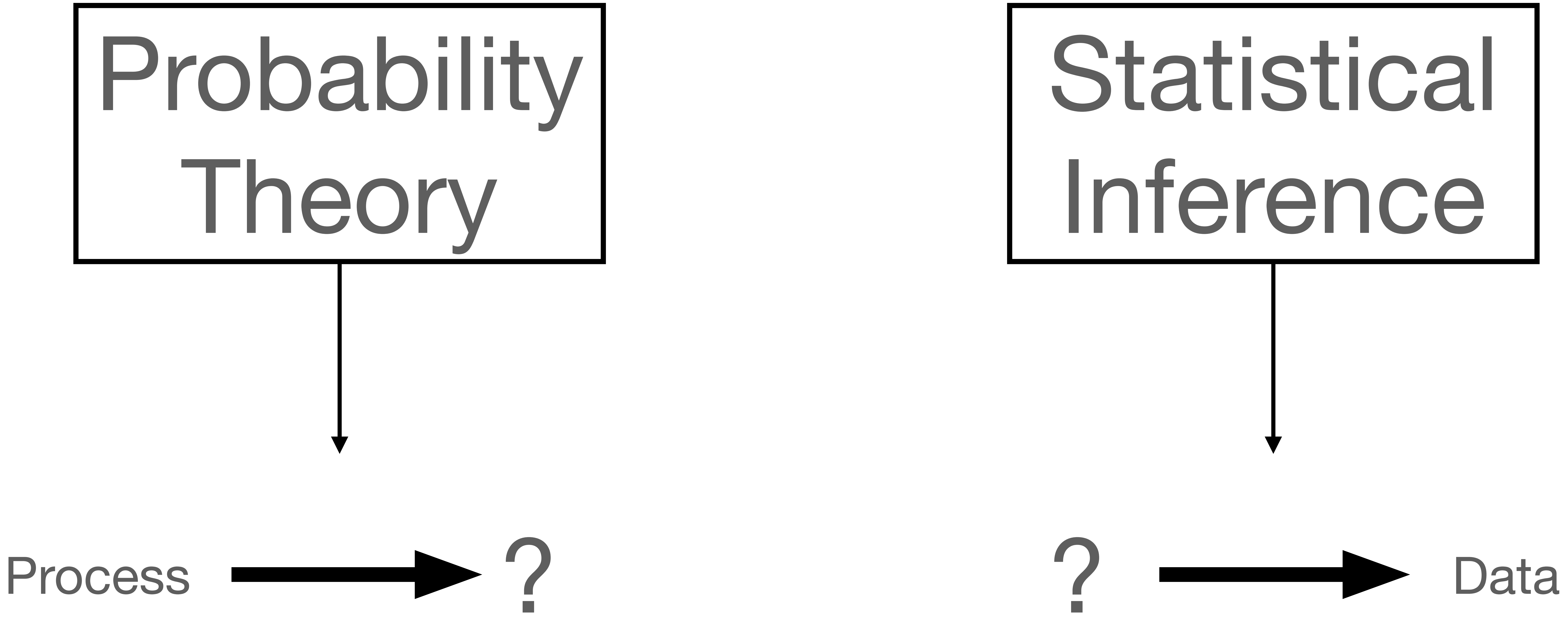
Probability Theory

“What can we say about the data generated by a given process?”

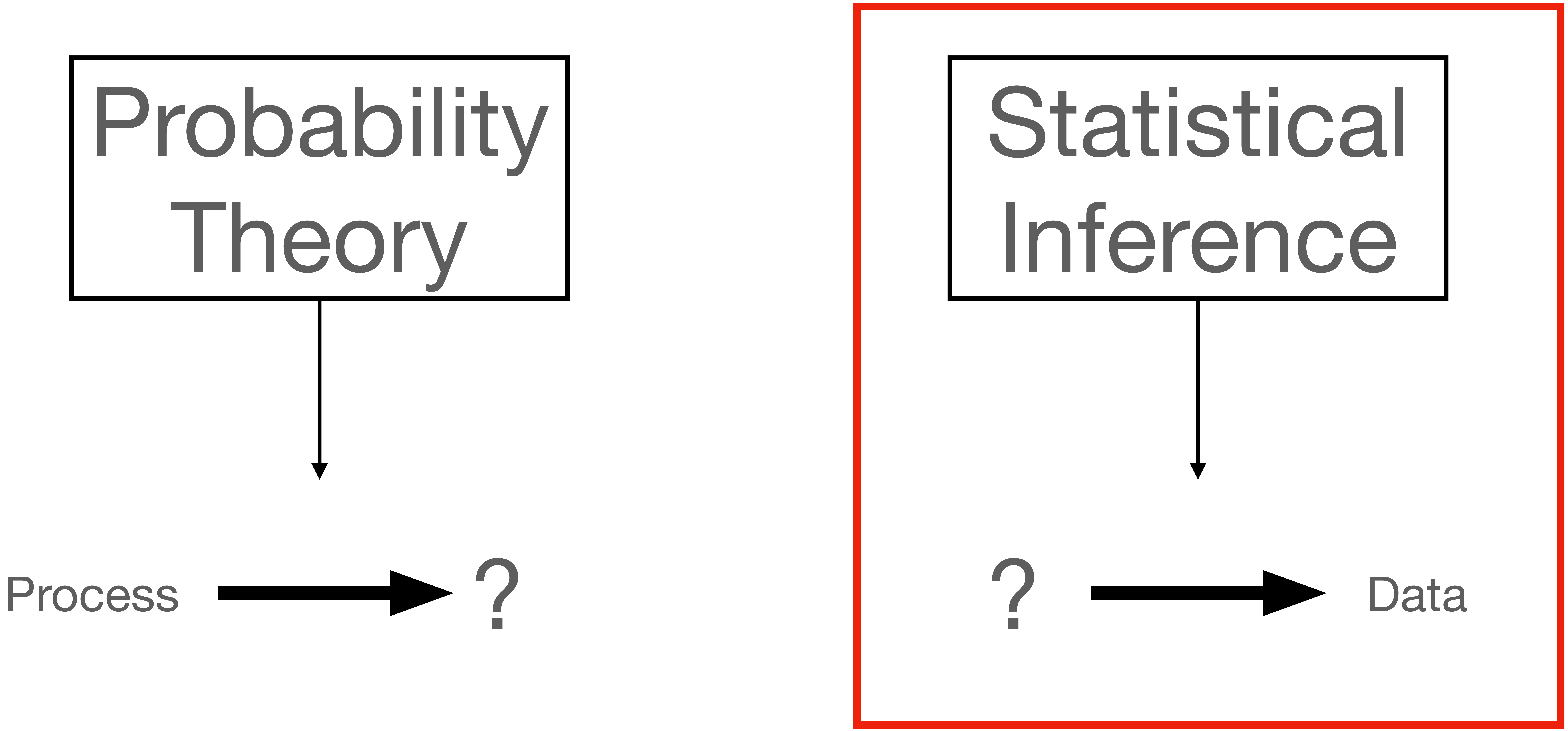
Statistical Inference

“What can we say about the process that generated a given data?”

The two sides of statistics



The two sides of statistics



Statistical inference: two “flavors”

Supervised learning: **today**

Predictor variables | Response variables

$$X \xrightarrow{?} Y$$

Statistical inference: two “flavors”

Supervised learning: **today**

Predictor variables   Response variables

$$X \xrightarrow{\quad ? \quad} Y$$

Unsupervised learning: **tomorrow**

Variables

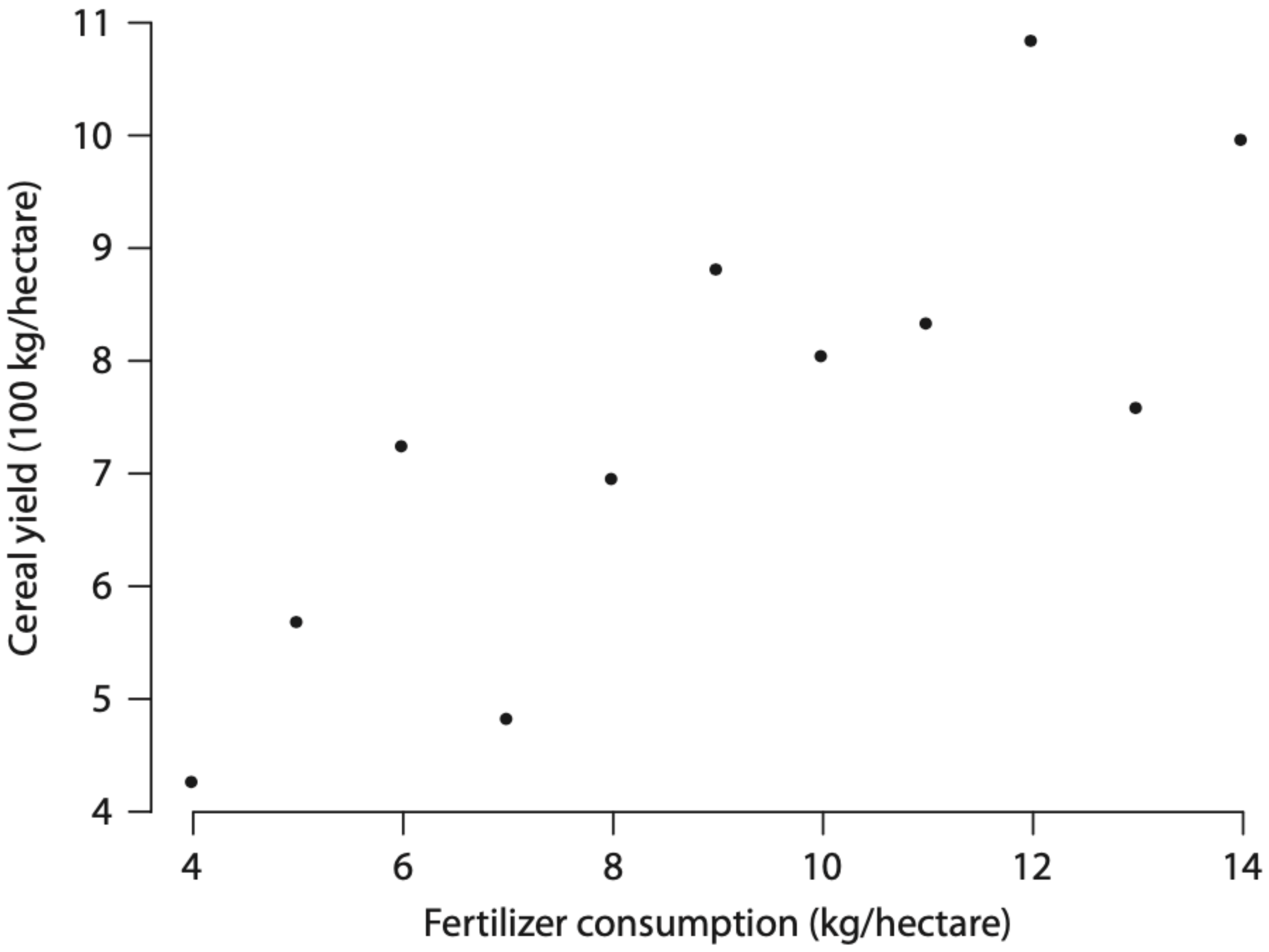
$$? \xrightarrow{\quad ? \quad} Y$$

Linear Regression

- Simplest model in supervised learning
- Jumping-off point for more complex models
- Many statistical models are extensions or generalizations of the linear model

Linear Regression

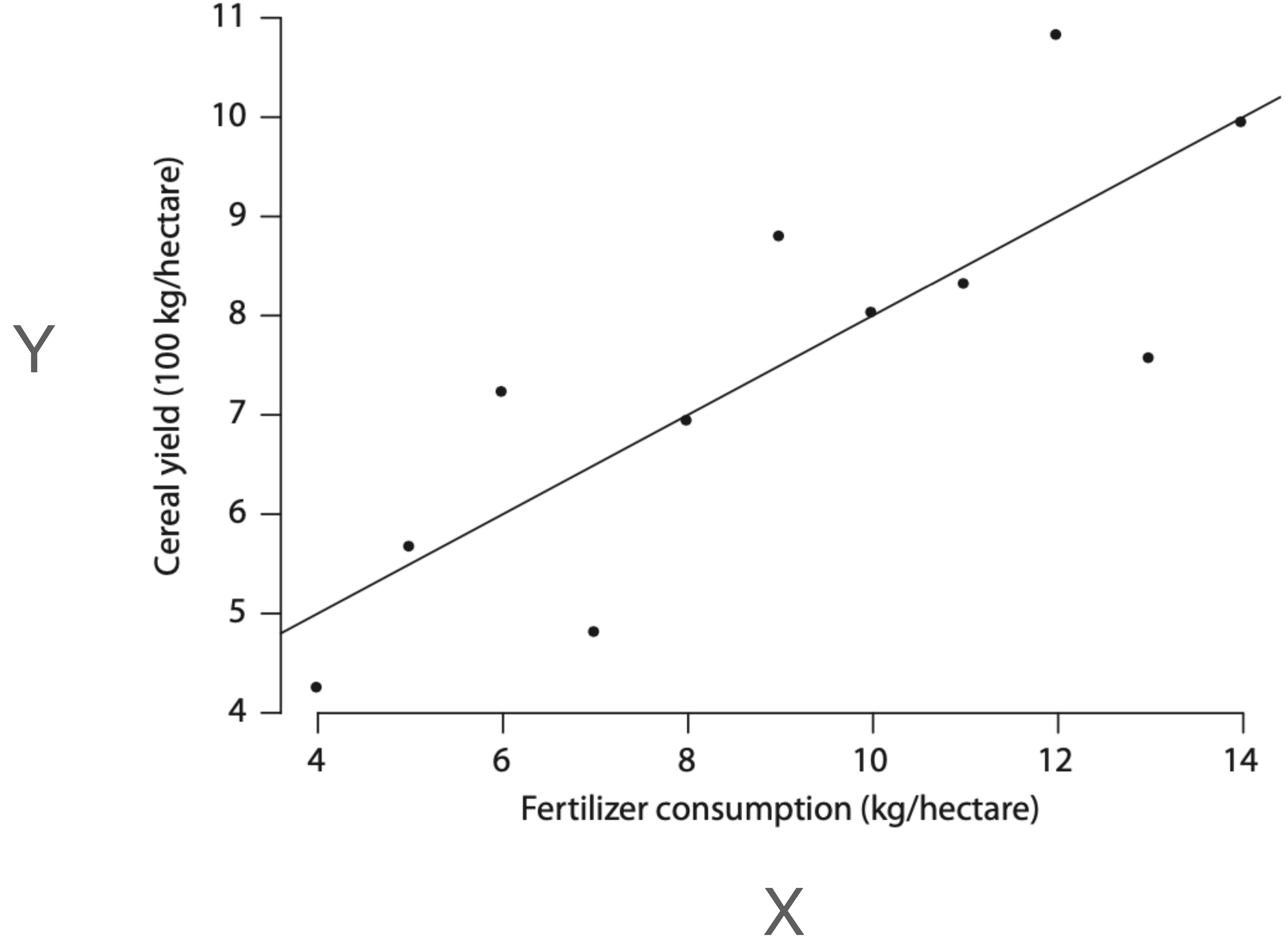
Y



X

Edge 2019

Linear Regression: “a line of best fit”

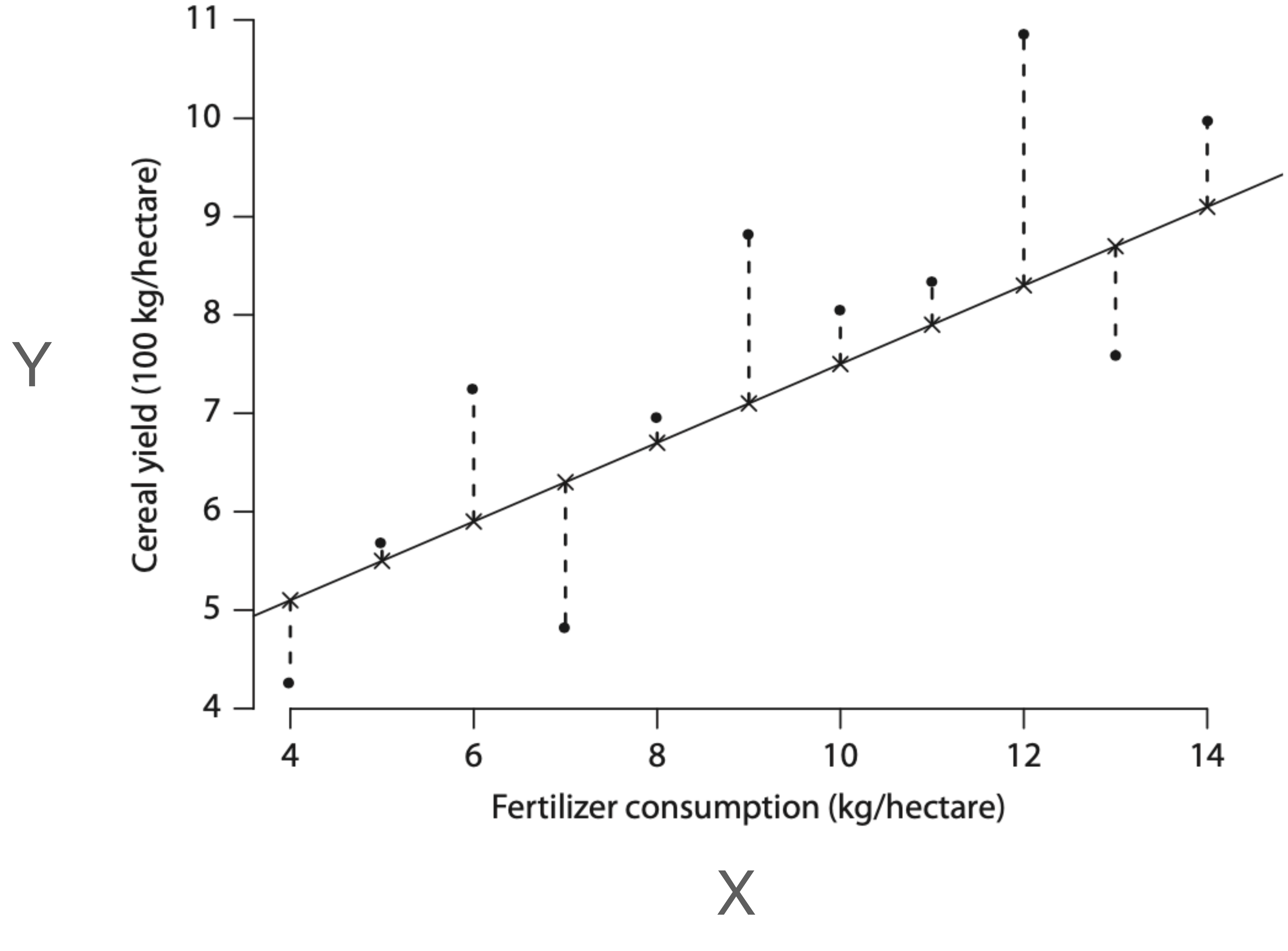


Edge 2019

Linear Regression: types of questions

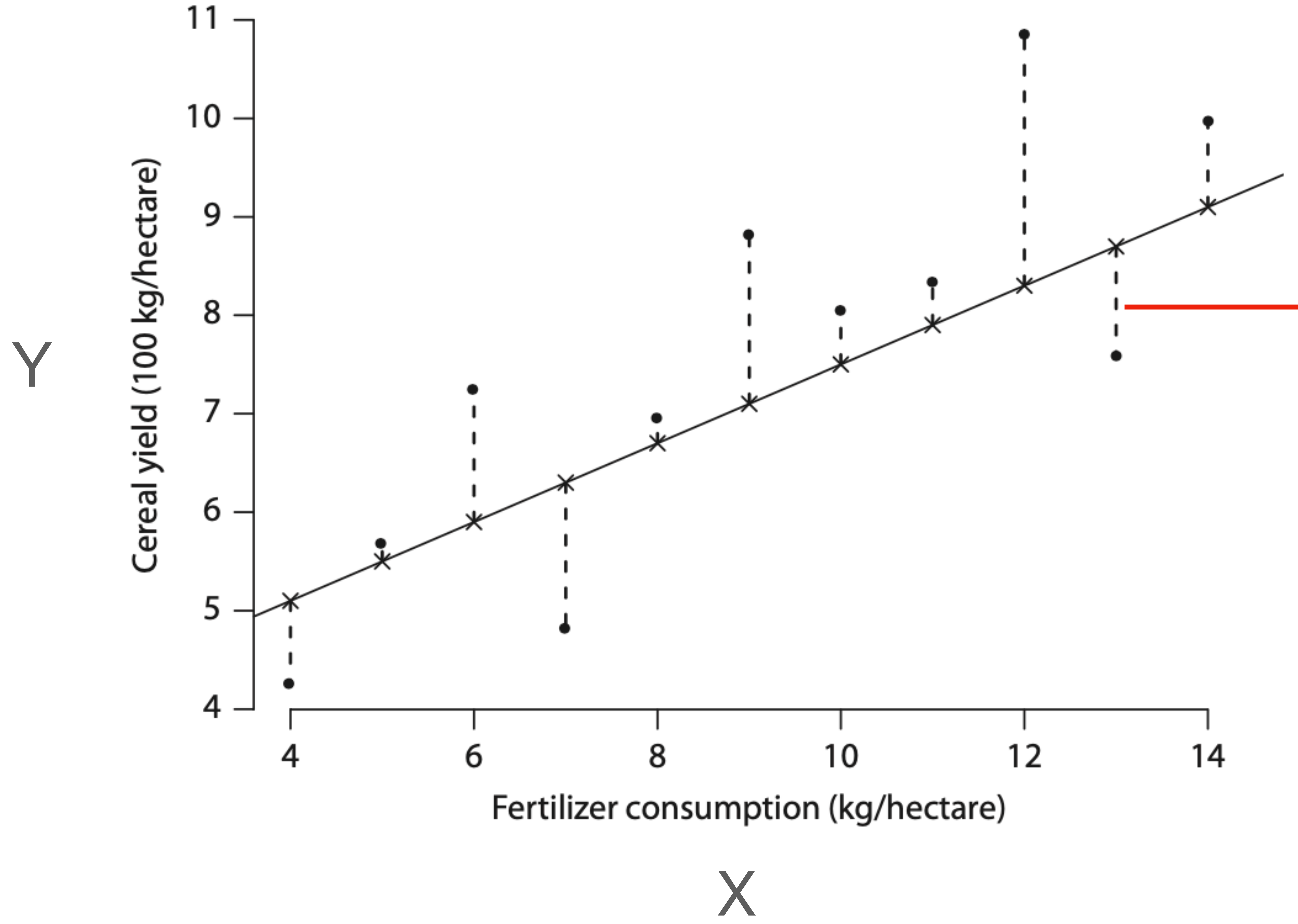
- Is there a relation between variable X and variable Y?
- How strong is the relation between variable X and variable Y?
- How well can we predict variable Y from the values of variable X?
- Is the relationship between variable X and Y linear?
- Which variables contribute to variable Y?

Linear Regression: “a line of best fit”



Edge 2019

Linear Regression: “a line of best fit”



Residual:
difference between a true data point and its prediction under the fitted model

Edge 2019

Simple linear regression

- **1** predictor variable (x)
- **1** response variable (y)

Simple linear regression

- **1** predictor variable (x)
- **1** response variable (y)

$$\mathbf{y} = \mathbf{f}(\mathbf{x})$$

Variable **y** is a **function** of **x**

Simple linear regression

- **1** predictor variable (x)
- **1** response variable (y)

$$\mathbf{y} = \mathbf{f}(\mathbf{x})$$

Variable **y** is a **function** of **x**

$$\mathbf{y} \approx \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{x}$$

Variable **y** is a **linear** function of variable **x**

Simple linear regression

- **1** predictor variable (x)
- **1** response variable (y)

$$\mathbf{y} = \mathbf{f}(\mathbf{x})$$

Variable **y** is a **function** of **x**

$$\mathbf{y} \approx \beta_0 + \beta_1 \mathbf{x}$$

Variable **y** is a **linear** function of variable **x**

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \epsilon$$

Variable **y** is a linear function of **x**, plus some **noise**

Simple linear regression

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \epsilon$$

Variable **y** is a linear function of **x**, plus some **noise**

Simple linear regression

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \epsilon$$

Variable **y** is a linear function of **x**, plus some **noise**

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Each value y_i has a specific predictor value x_i and noise value ϵ_i

Simple linear regression

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \epsilon$$

Variable **y** is a linear function of **x**, plus some **noise**

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Each value y_i has a specific predictor value x_i and noise value ϵ_i

Individual values are represented with subscripts

Simple linear regression

Vectors of values (variables) are represented in bold

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \epsilon$$

Variable **y** is a linear function of **x**, plus some **noise**

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Each value y_i has a specific predictor value x_i and noise value ϵ_i

Individual values are represented with subscripts

Parameters vs. Estimates

$$\mathbf{y} = \boxed{\beta_0} + \boxed{\beta_1}\mathbf{x} + \epsilon$$

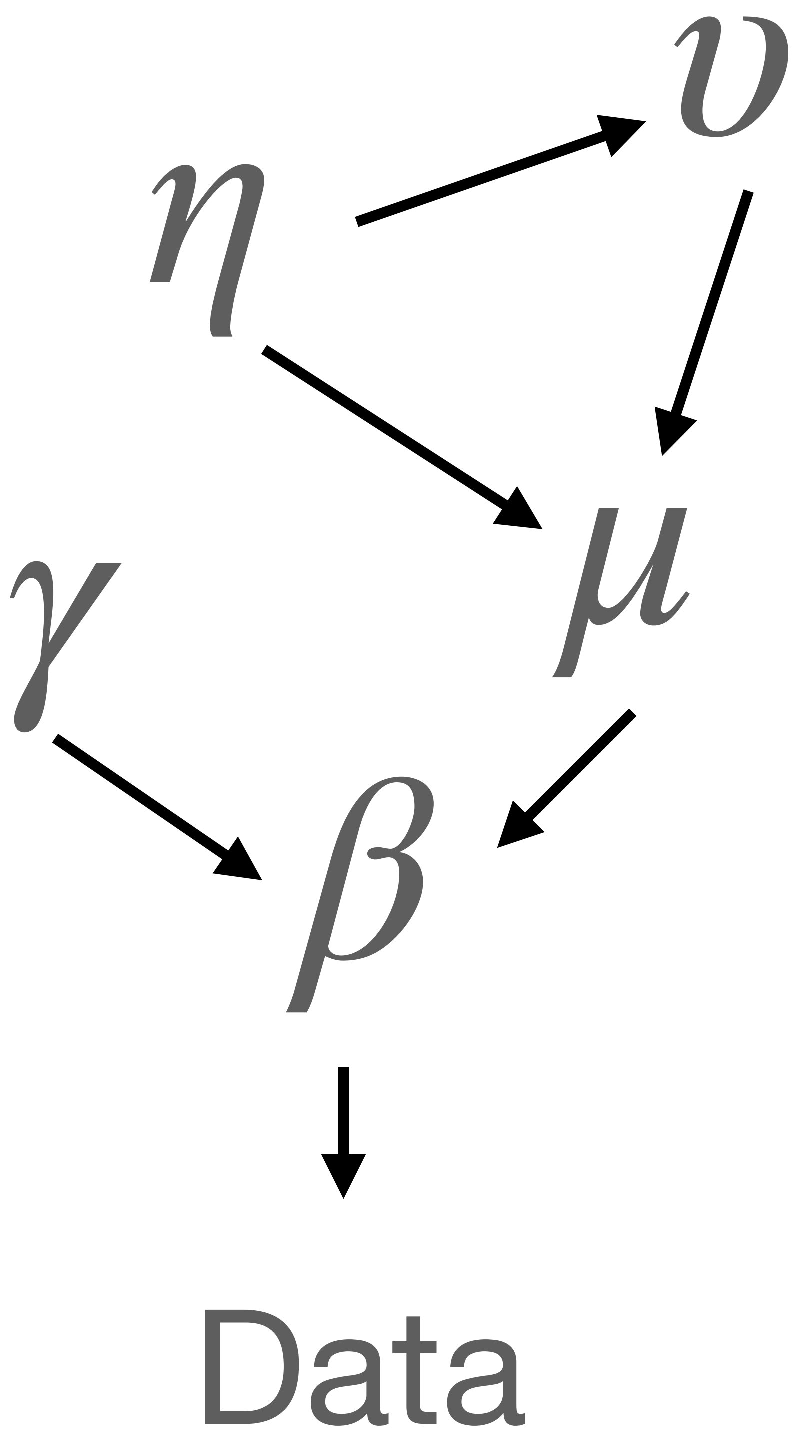
β_0 and β_1 are unknown parameters in our model: **we do not know their value**

$$\mathbf{y} = \boxed{\hat{\beta}_0} + \boxed{\hat{\beta}_1}\mathbf{x} + \epsilon$$

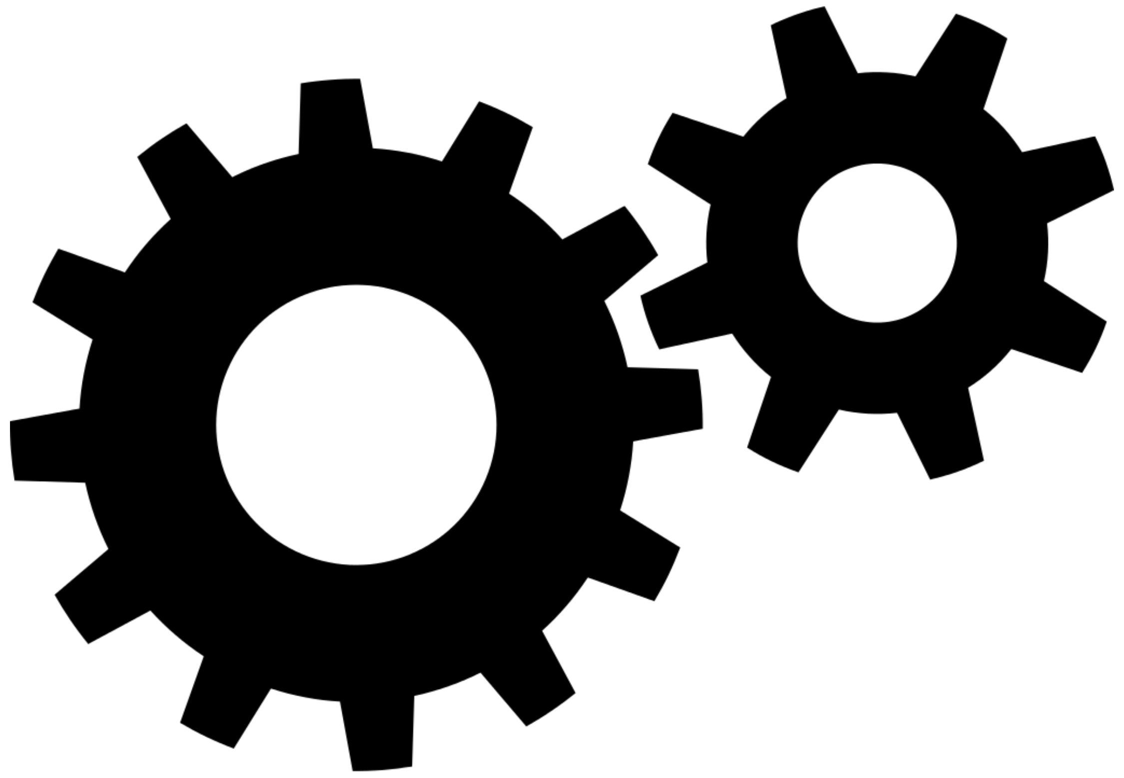
$\hat{\beta}_0$ and $\hat{\beta}_1$ are our **best estimates** of the above parameters

Model vs. Inference Method

Model



Inference method



Parameter estimates

$$\hat{\alpha} = 0.56$$
$$\hat{\beta} = 3.2$$
$$\hat{\mu} = -2$$
$$\cdots$$

Model vs. Inference Method

Model

Inference method

Linear regression

Ordinary least squares

Weighted least squares

Maximum likelihood

Ridge regression

…

Model vs. Inference Method

Model

Linear regression

Inference method

$\boxed{\text{Ordinary least squares}}$

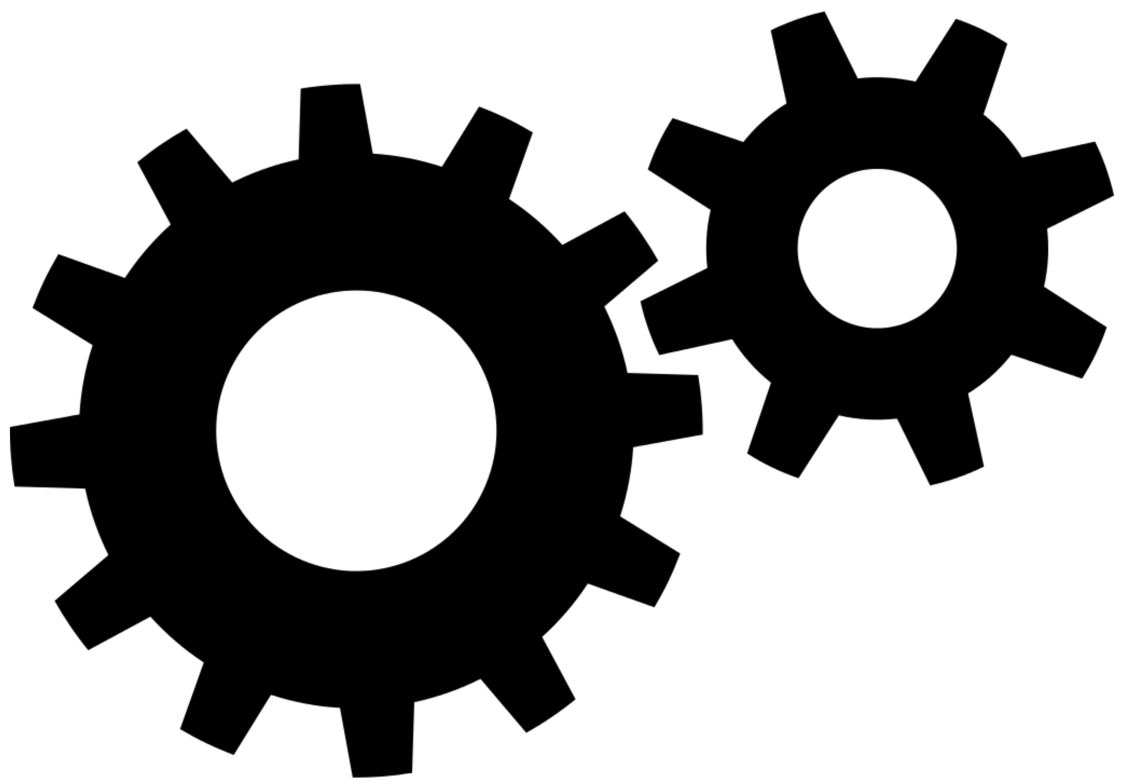
Weighted least squares

Maximum likelihood

Ridge regression

...

Ordinary least squares



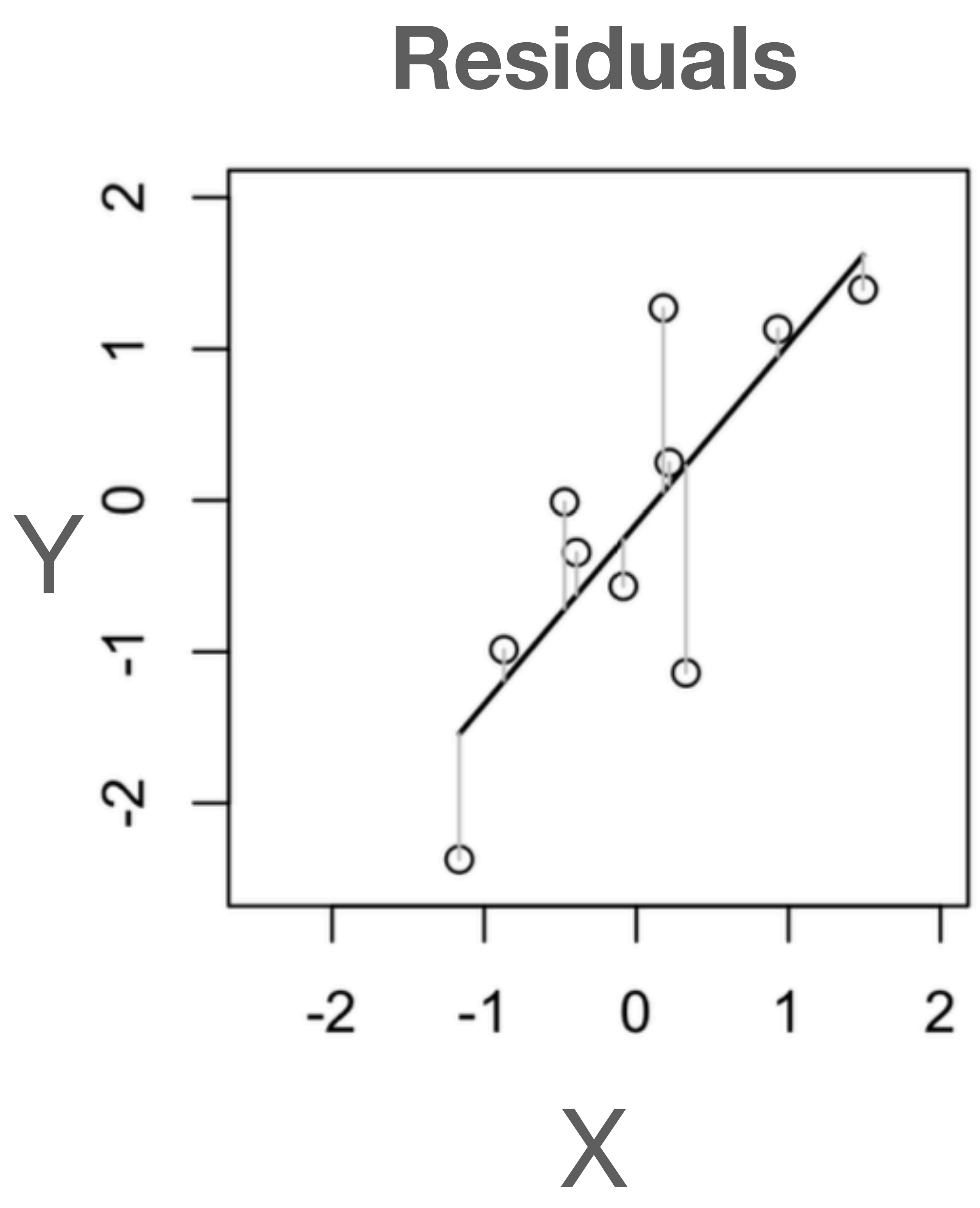
"Find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the parameters β_0 and β_1 by **minimizing the Sum of Squared Residuals**"

Ordinary least squares

"Find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the parameters β_0 and β_1 by **minimizing the Sum of Squared Residuals**"

?

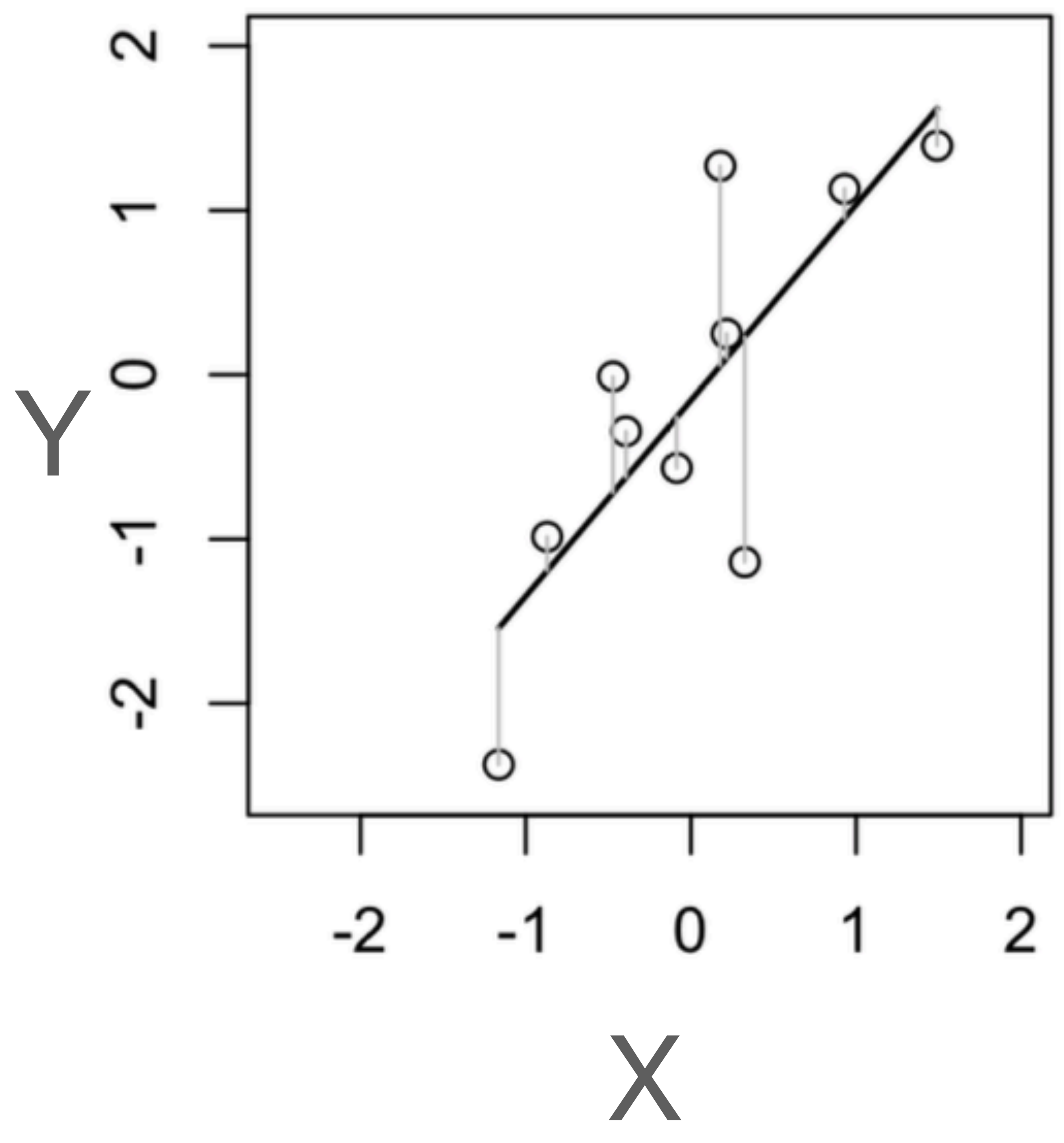
Ordinary least squares: what does this mean?

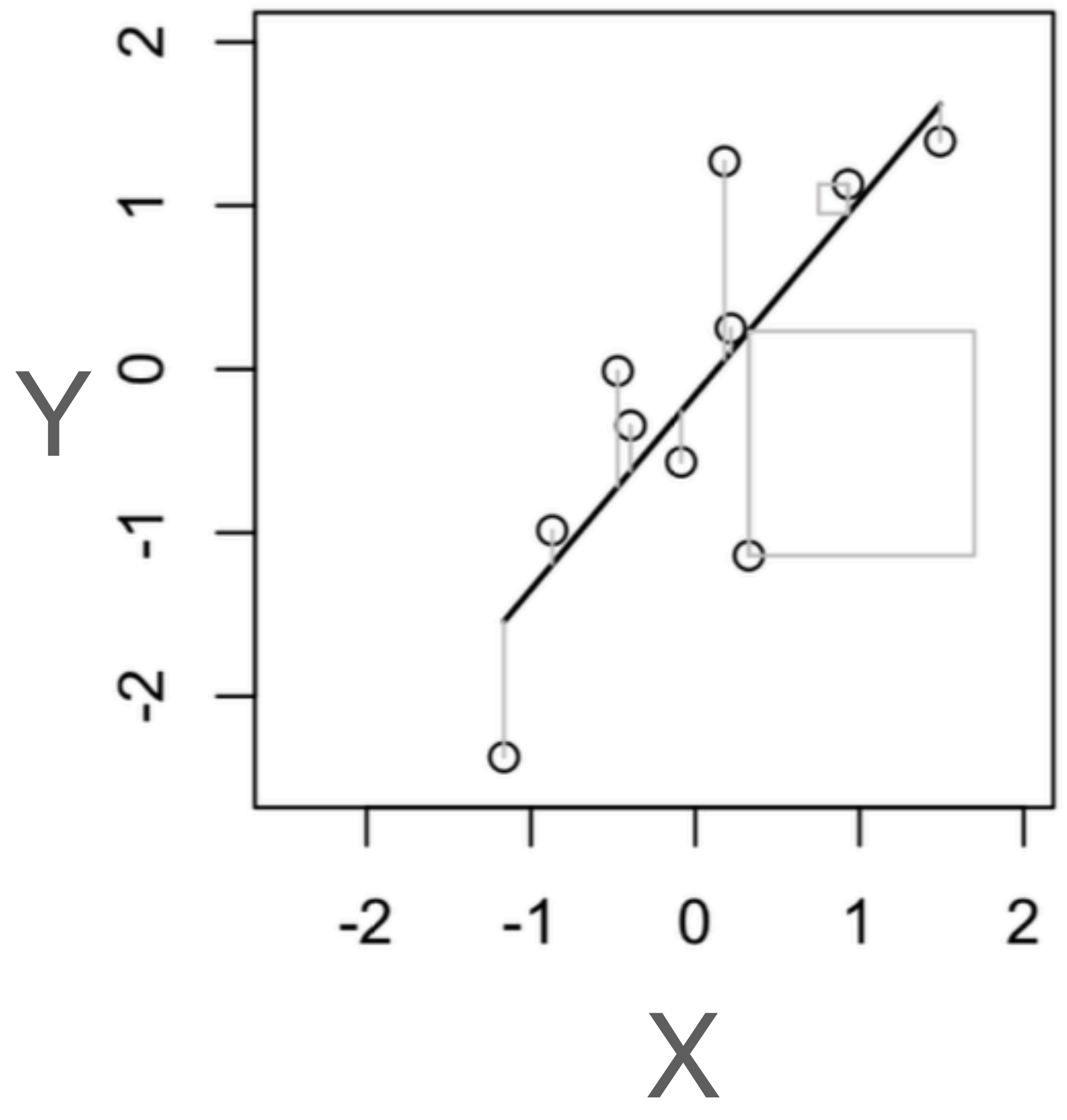


res = $y_i - \hat{y}$

Ordinary least squares: what does this mean?

Residuals

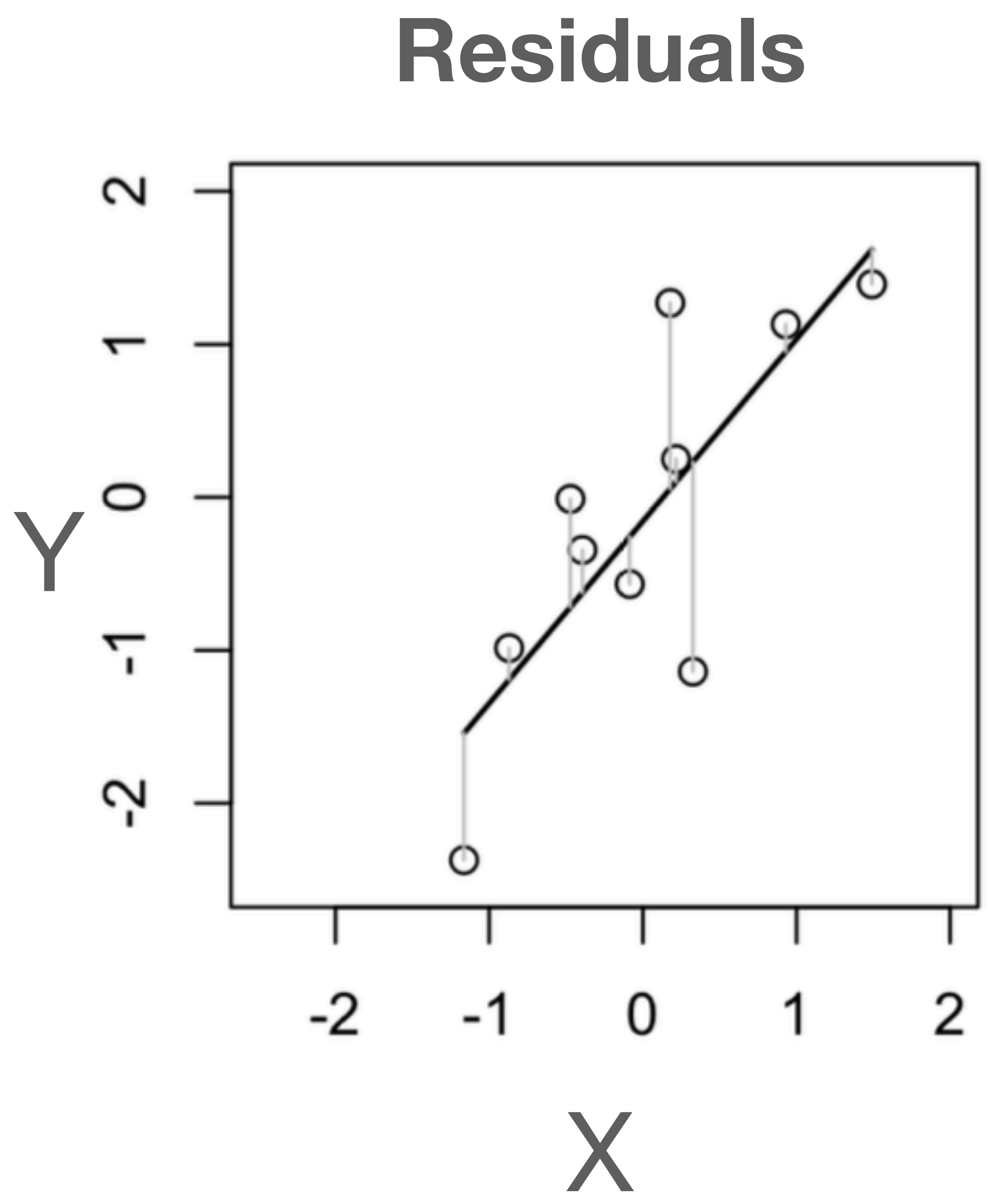


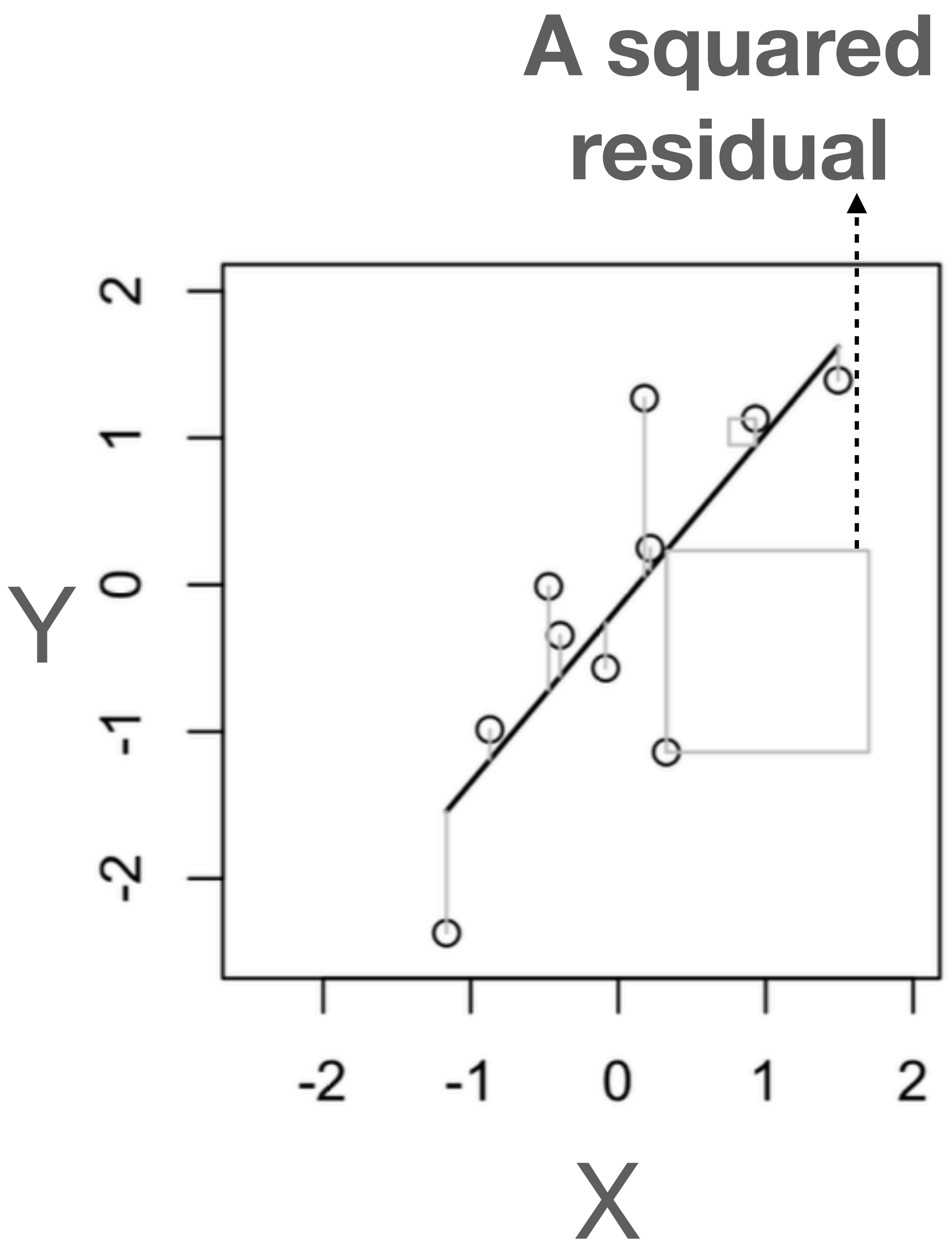


res = $y_i - \hat{y}$

$(y_i - \hat{y})^2$

Ordinary least squares: what does this mean?

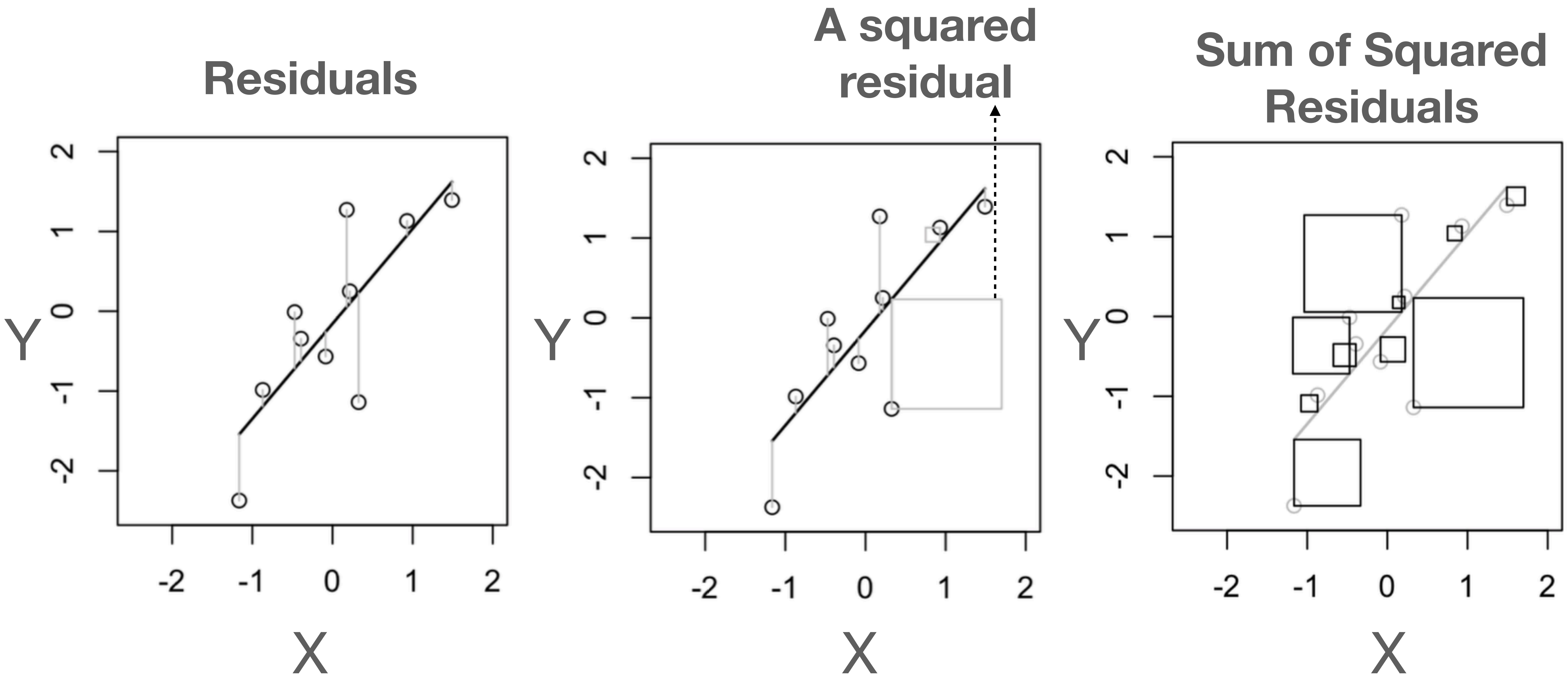




res = $y_i - \hat{y}$

$(y_i - \hat{y})^2$

Ordinary least squares: what does this mean?

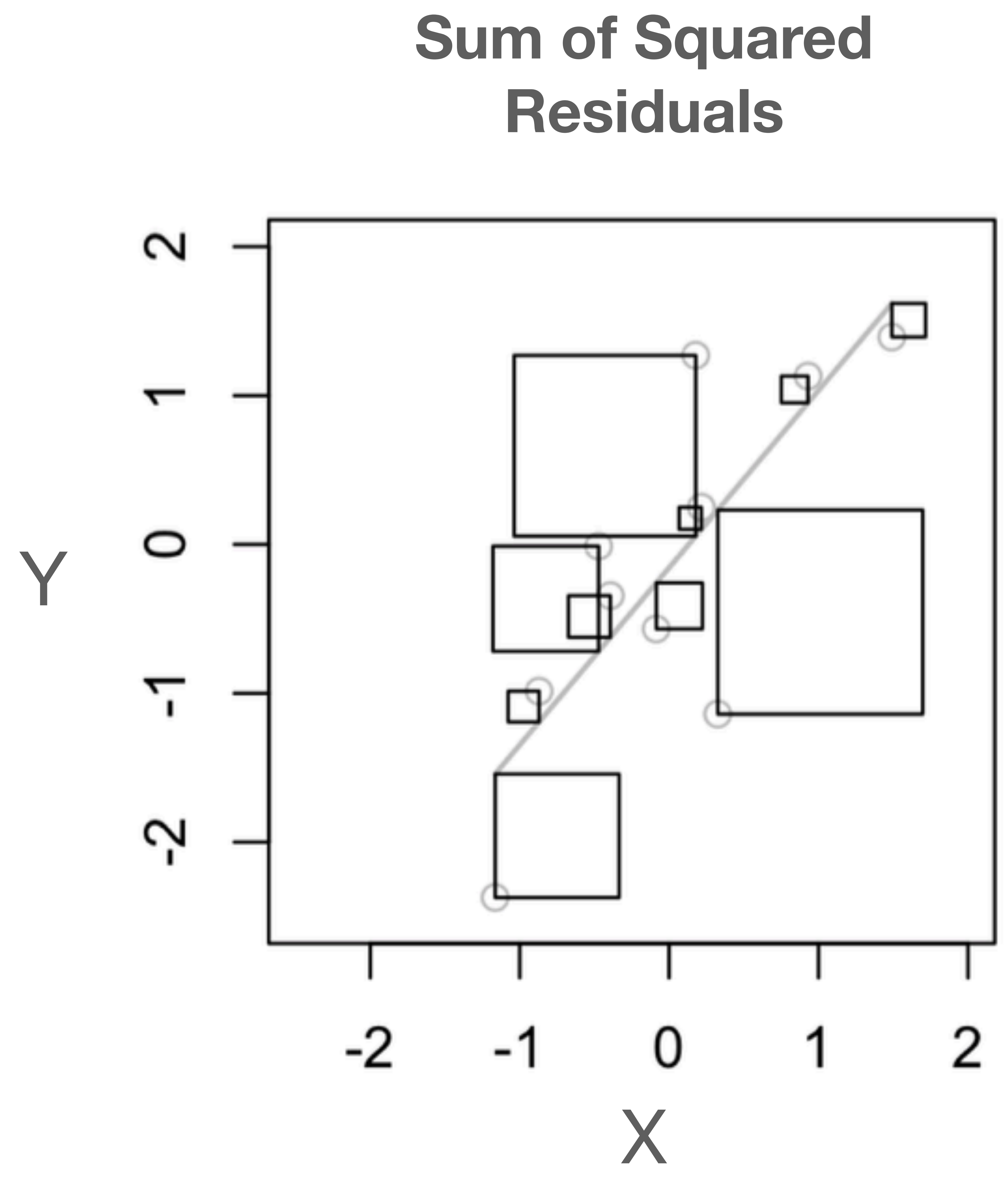


res = $y_i - \hat{y}$

$(y_i - \hat{y})^2$

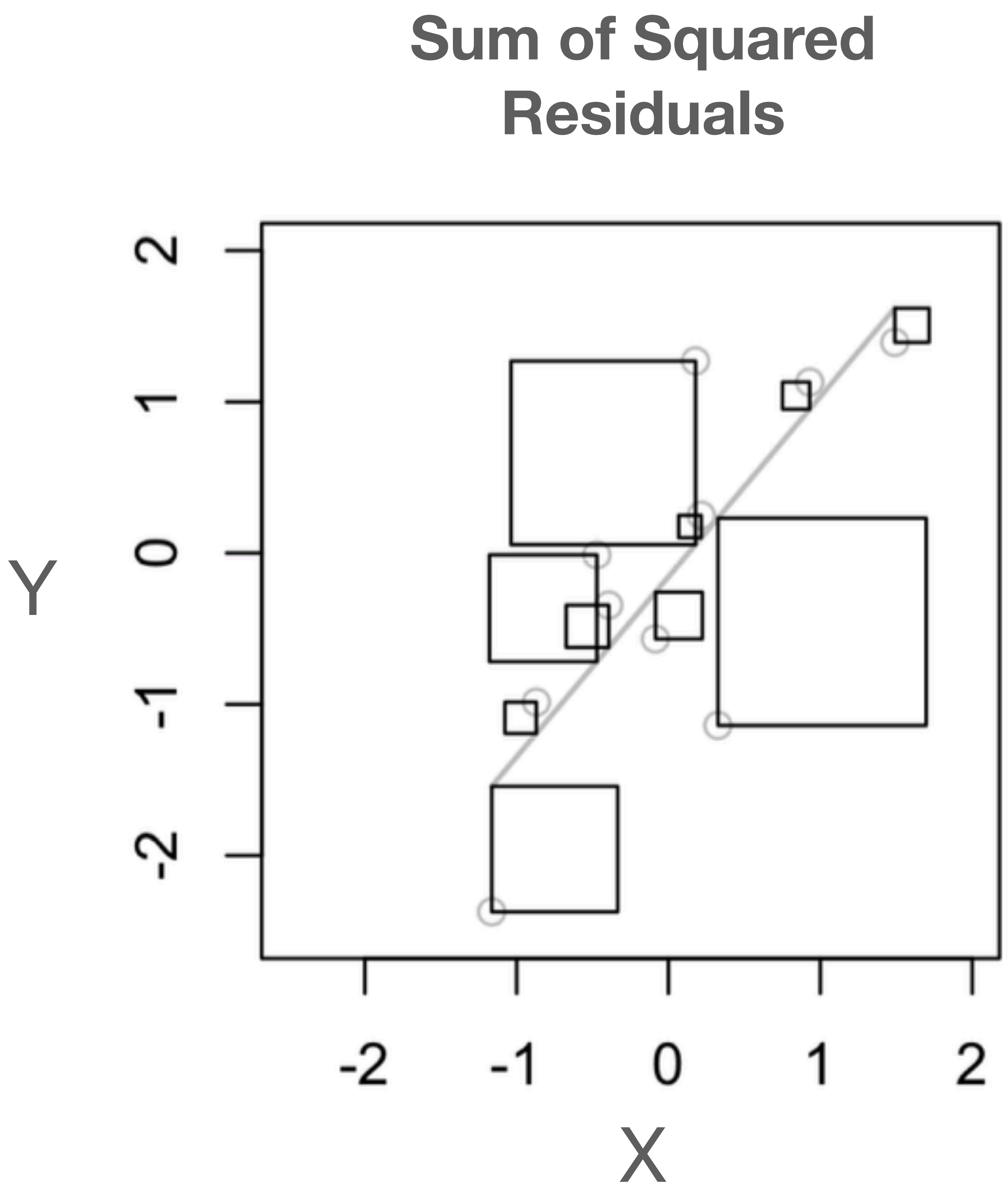
$SS_{res} = \sum_{i}^{n} (y_i - \hat{y})^2$

Ordinary least squares: what does this mean?



$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Ordinary least squares: what does this mean?

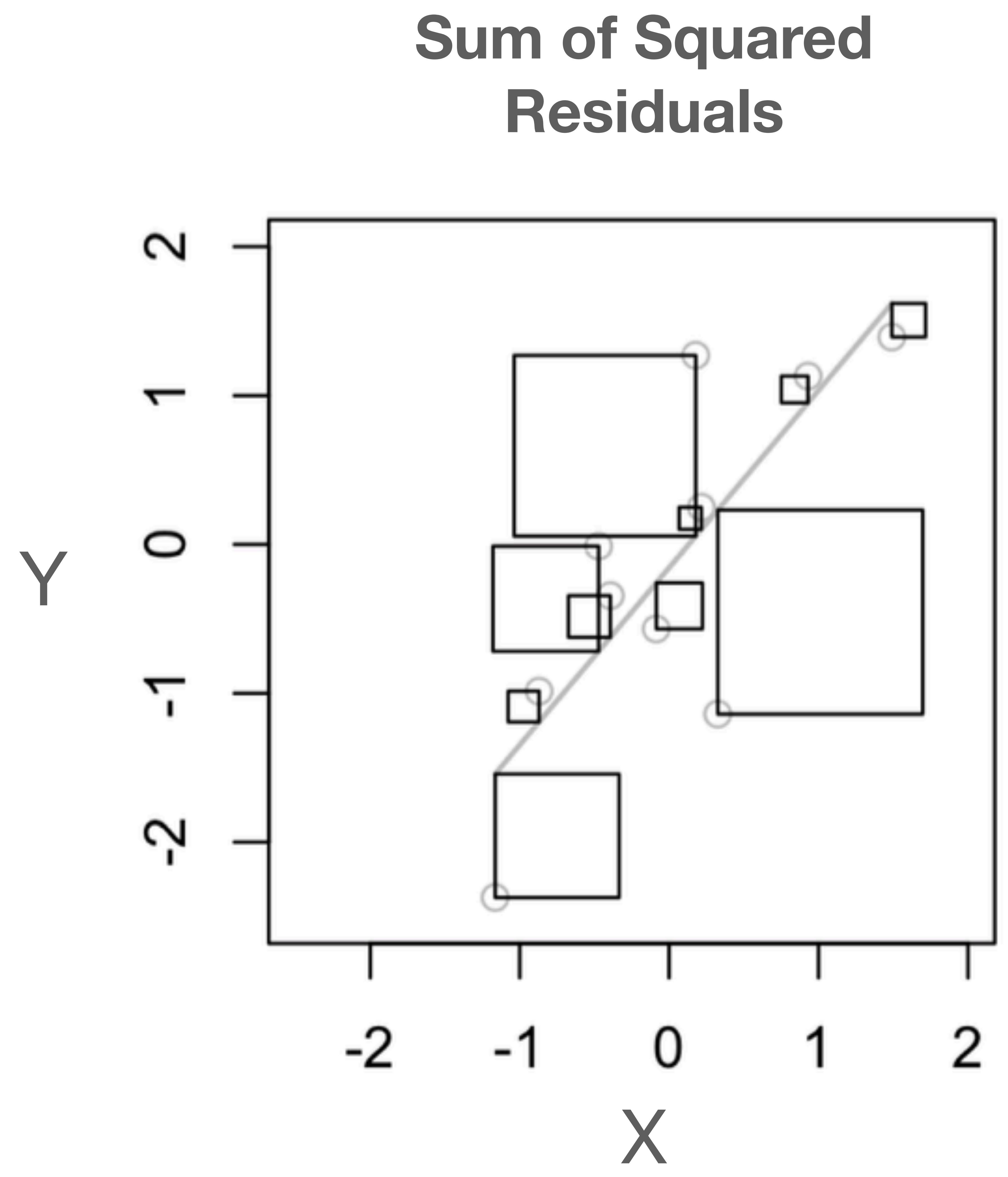


$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

but our estimate $\hat{y}_i$ is simply a linear function of x_i:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Ordinary least squares: what does this mean?



$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

but our estimate $\hat{y}_i$ is simply a linear function of x_i:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Ordinary least squares: what does this mean?

$$SS_{res} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The values of β_0 and β_1 that **minimize SSres** are:

Ordinary least squares: what does this mean?

$$SS_{res} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The values of β_0 and β_1 that **minimize SSres** are:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Ordinary least squares: what does this mean?

$$SS_{res} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The values of β_0 and β_1 that **minimize SSres** are:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Average of y   Average of x

Ordinary least squares: what does this mean?

$$SS_{res} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The values of β_0 and β_1 that **minimize SSres** are:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} [(x_i - \bar{x})^2]}$$

Average of y

Average of x

Ordinary least squares: what does this mean?

$$SS_{res} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The values of β_0 and β_1 that **minimize SSres** are:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} [(x_i - \bar{x})^2]}$$

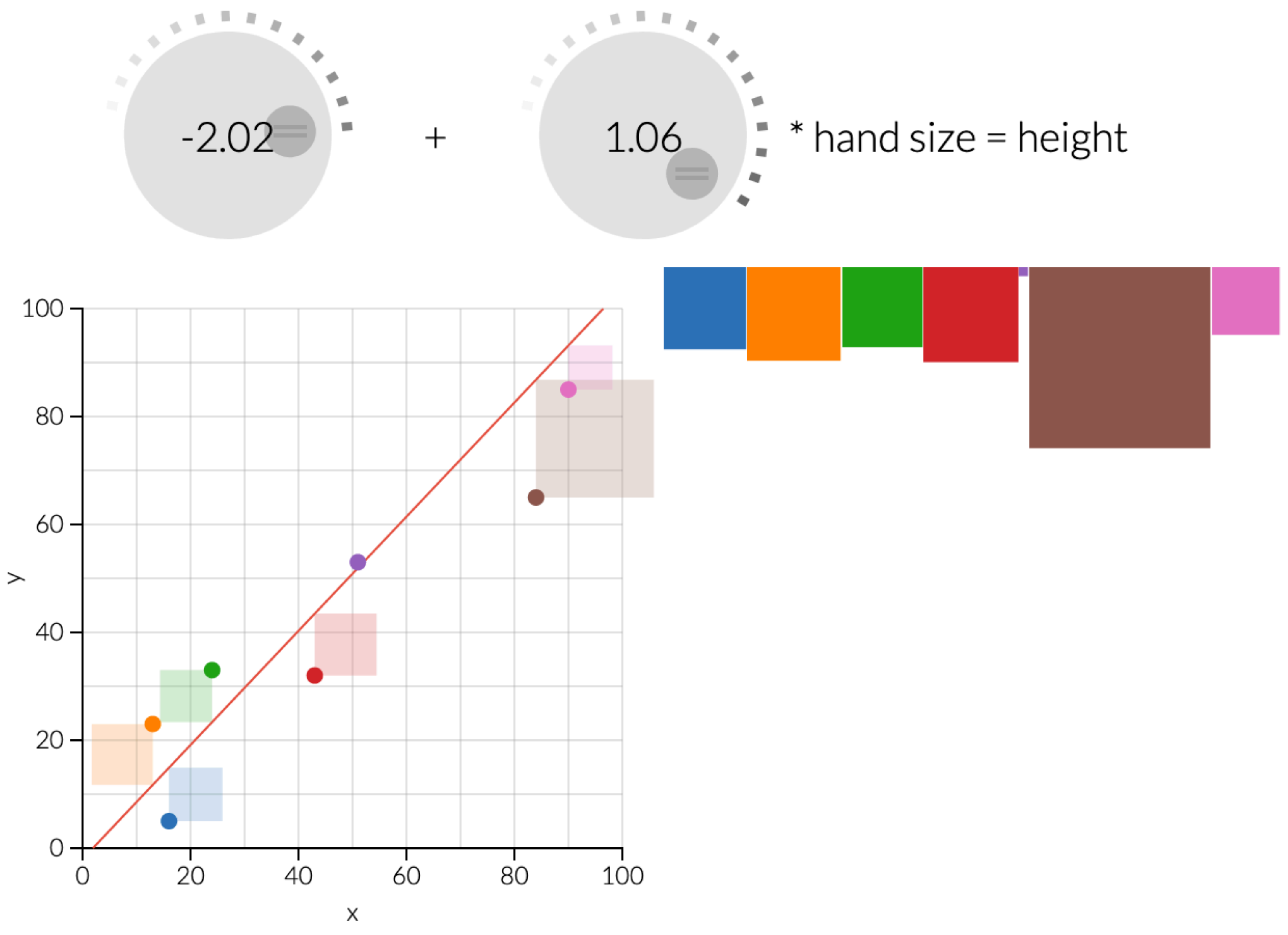
Average of y

Average of x

You'll have to trust me here… or not:

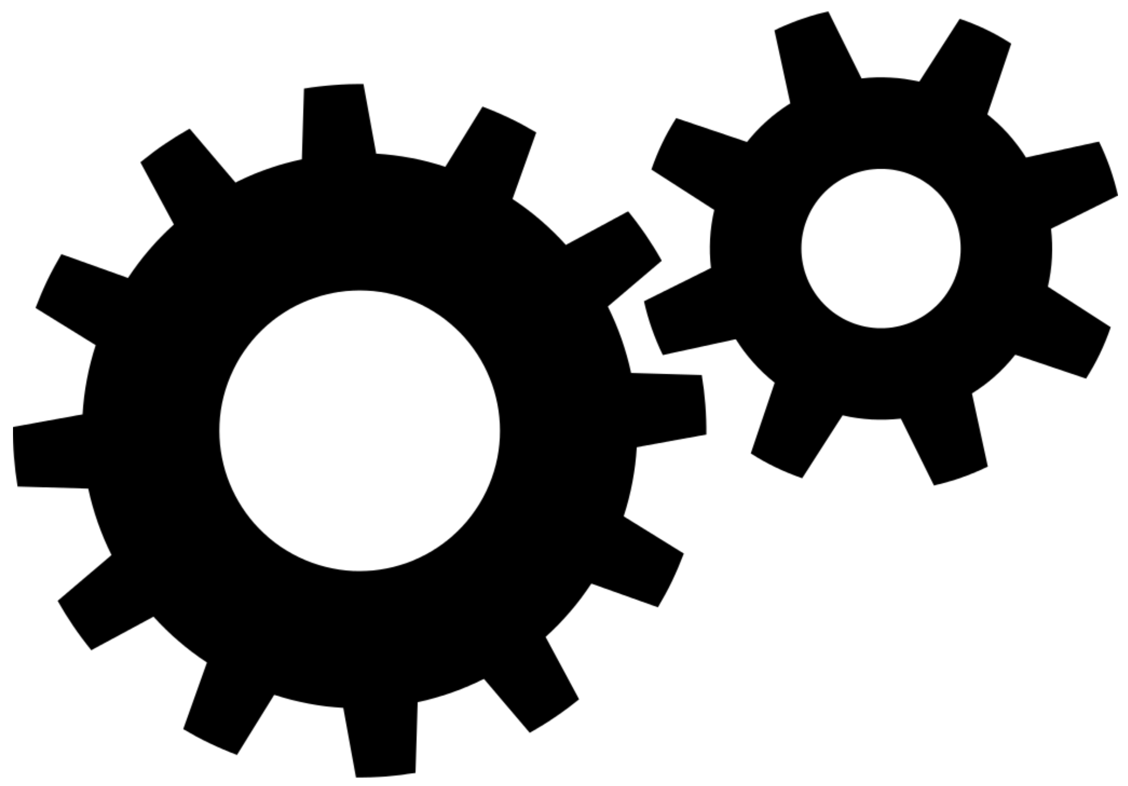
for proof, try taking the derivative of SSres and equalizing it to 0.

Ordinary least squares: interactive session



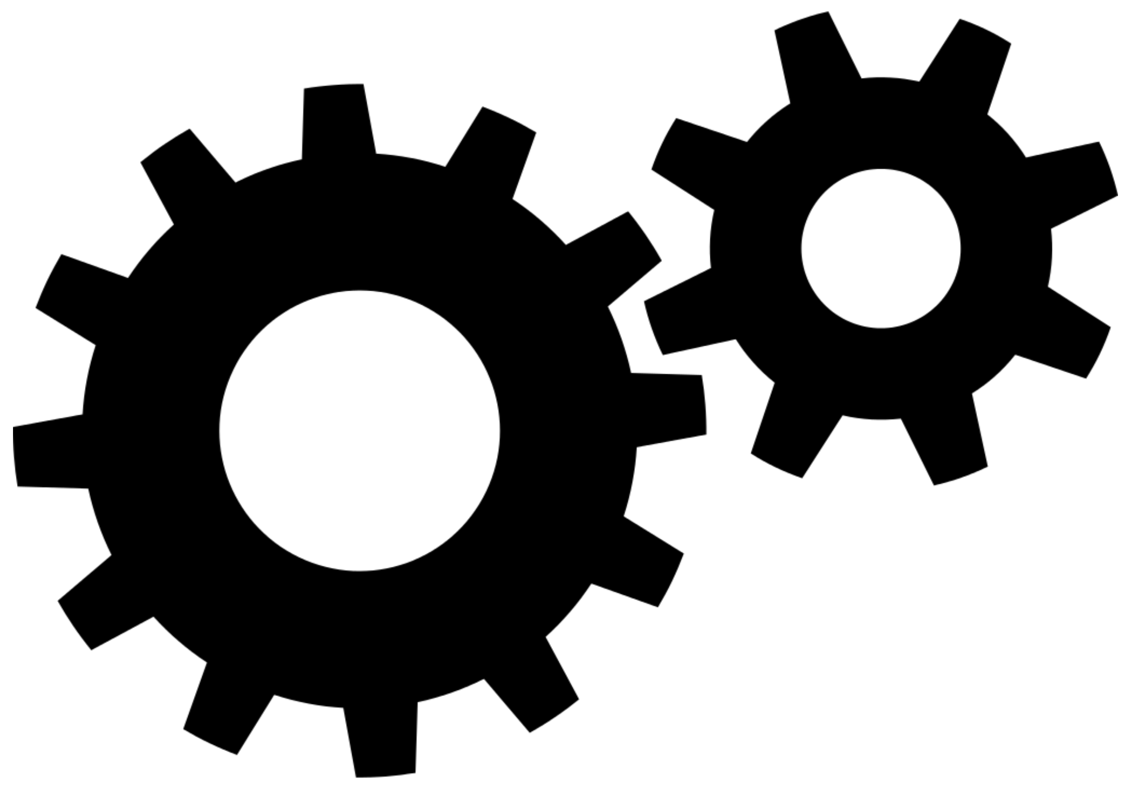
https://setosa.io/ev/ordinary-least-squares-regression/

Ordinary least squares



“Find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the parameters β_0 and β_1 by **minimizing the Sum of Squared Residuals**”

Ordinary least squares



"Find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the parameters β_0 and β_1 by **minimizing the Sum of Squared Residuals**"

Loss function

Model vs. Inference Method

Model

Linear regression

Inference method

Ordinary least squares

Weighted least squares

Maximum likelihood

Ridge regression

…

Each of these methods has a different loss function!

What does the slope mean?

$$\hat{\beta}_1 = ?$$

What does the slope mean?

$$\hat{\beta}_1 = ?$$

The **slope** measures the **covariance between X and Y**, as a proportion of the **variance of X**

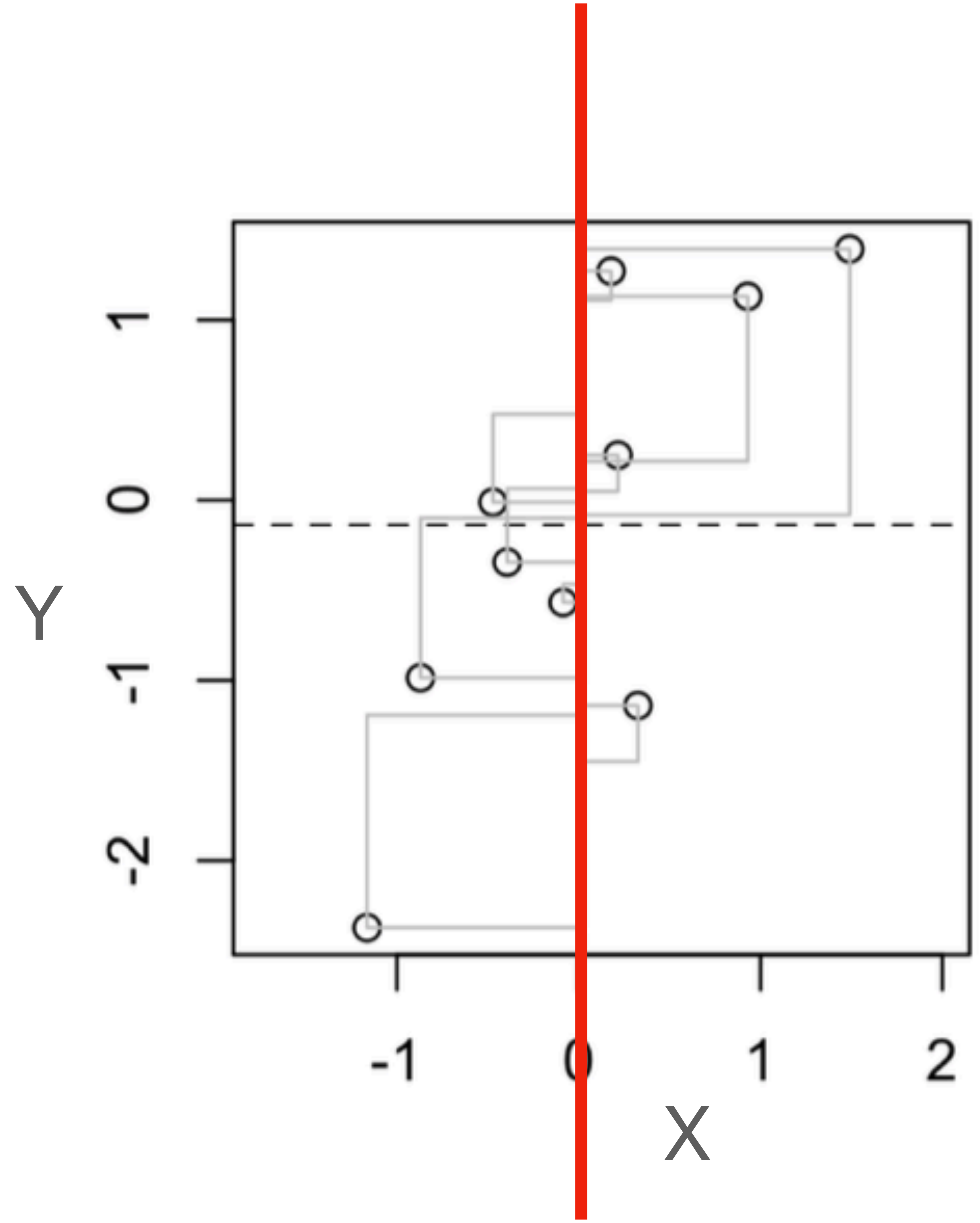
What does the slope mean?

$$\hat{\beta}_1 = ?$$

The **slope** measures the **covariance between X and Y**, as a proportion of the **variance of X**

Let's unpack this a bit…

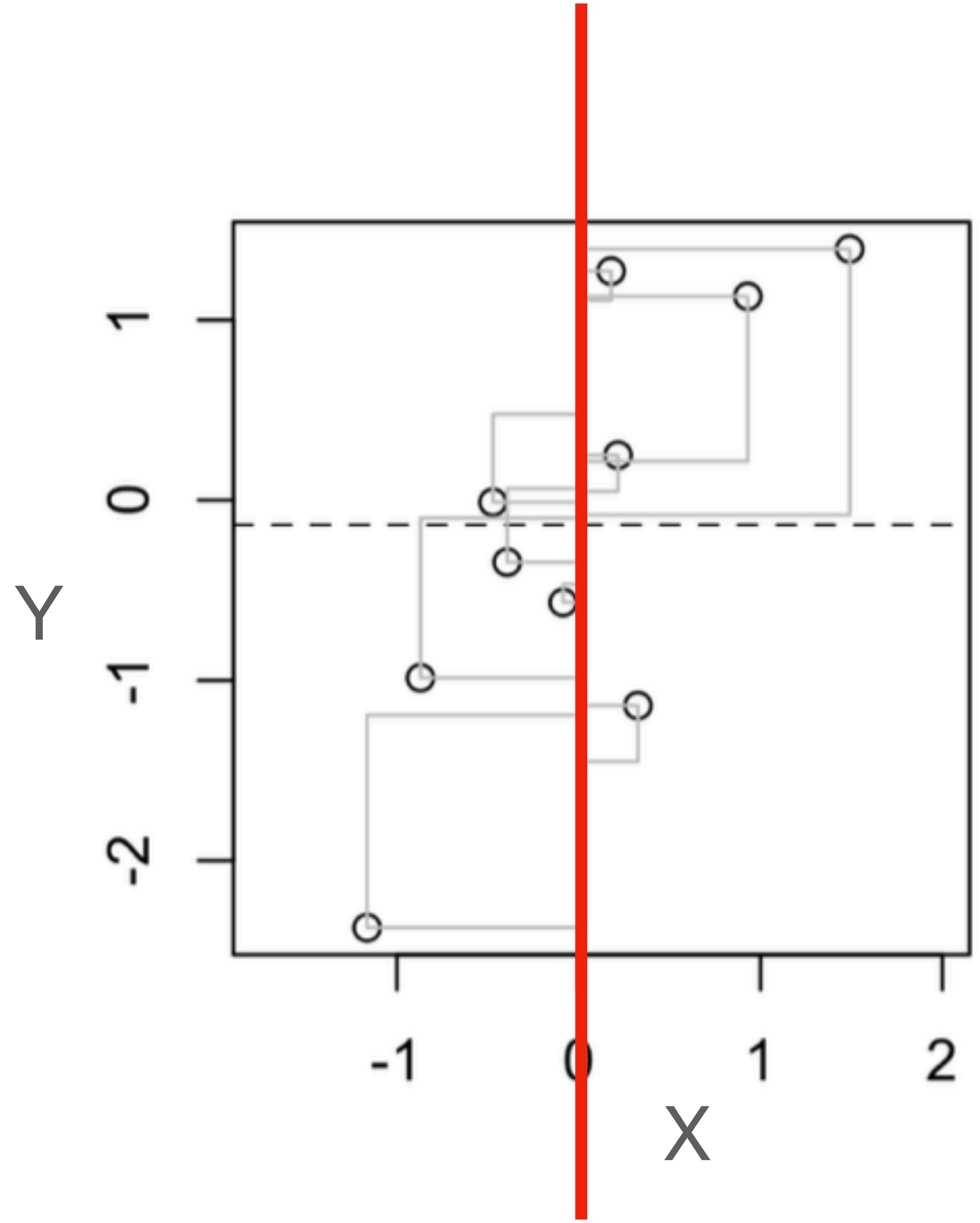
Measuring variation



Measuring variation

Sum of squares of X

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



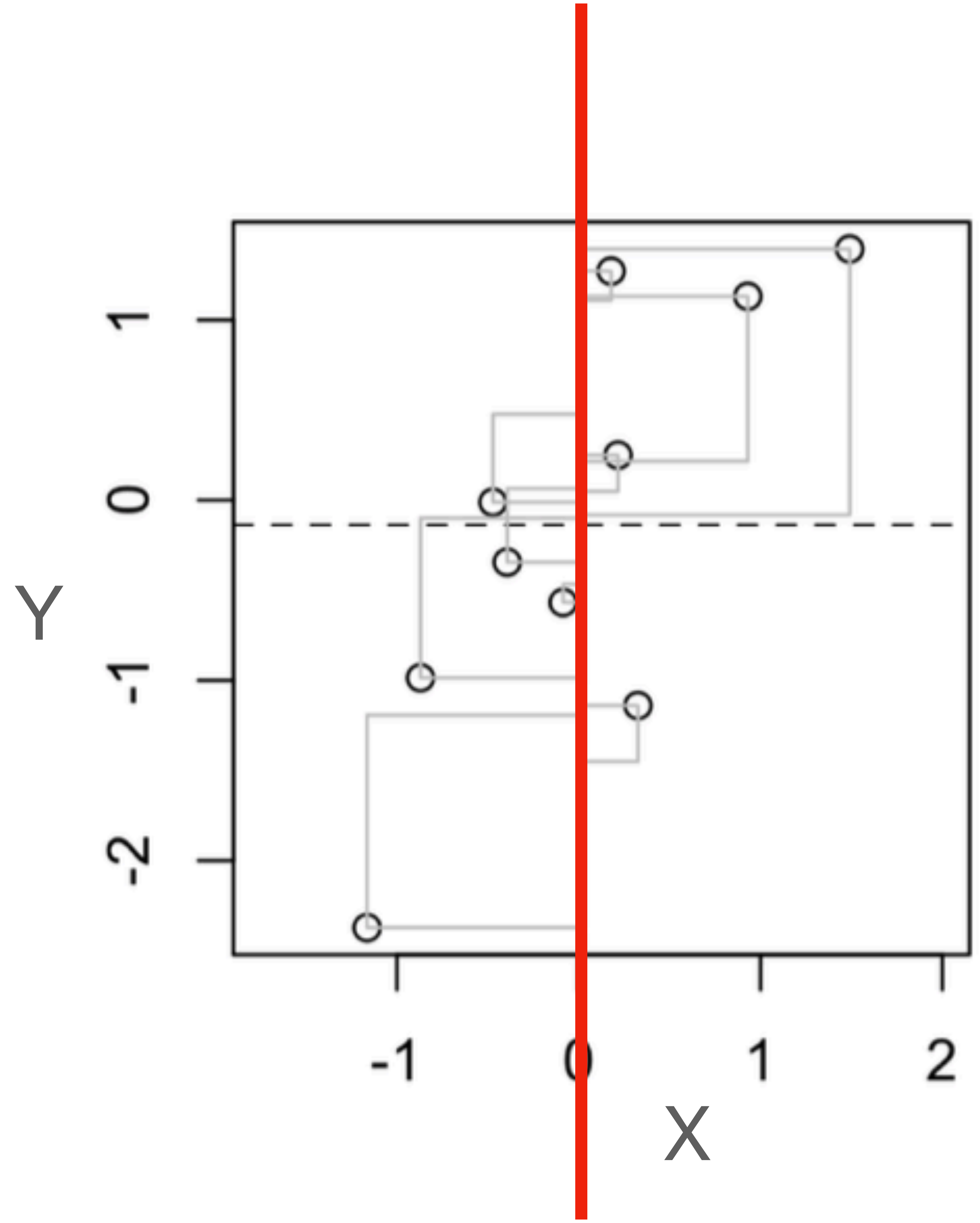
Measuring variation

Sum of squares of X

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

Sample variance of X:

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



Measuring variation

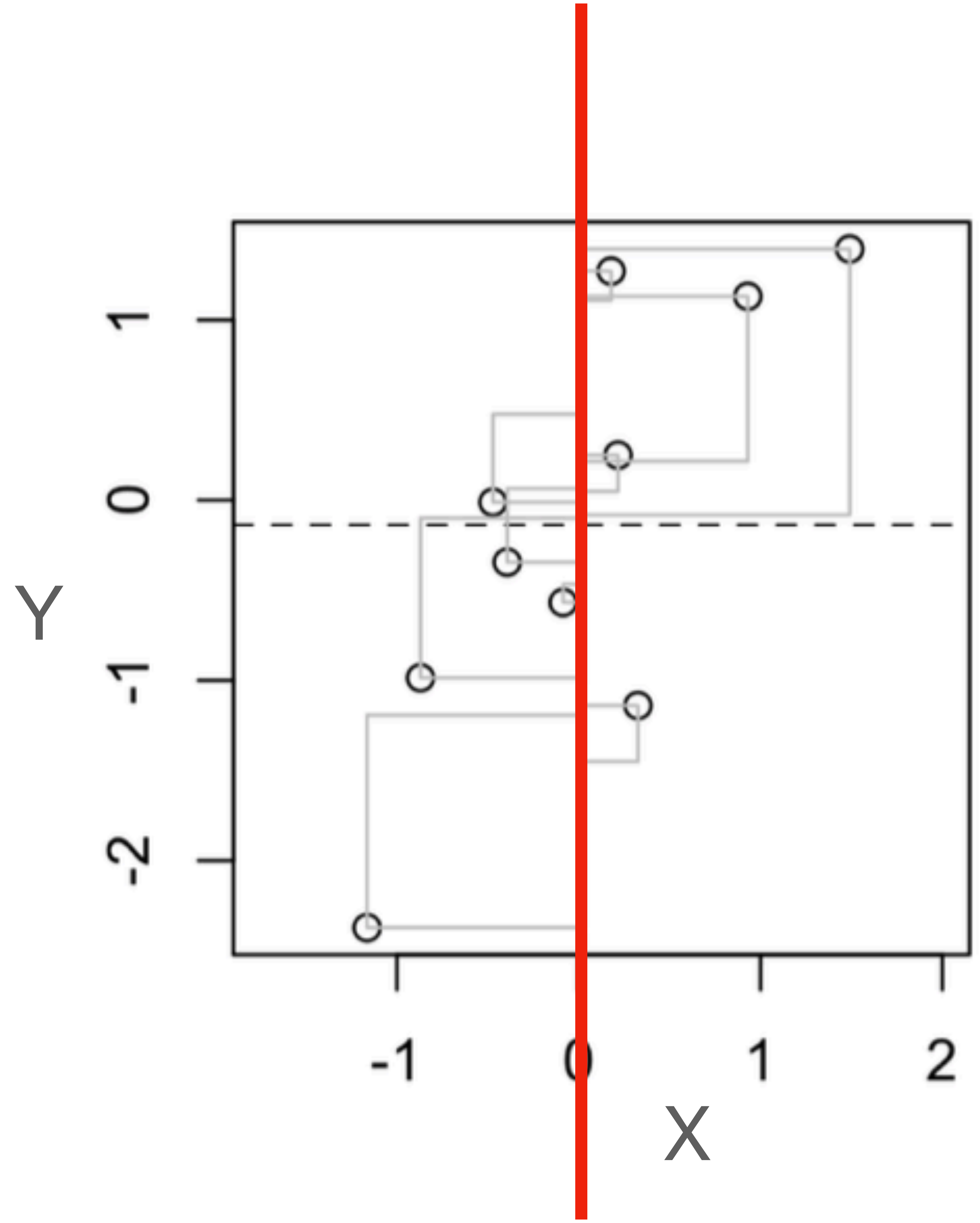
Sum of squares of X

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

Sample variance of X:

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

This constant is applied to the sum of squares to obtain an **unbiased estimate of the true variance** when we only have **finite samples**

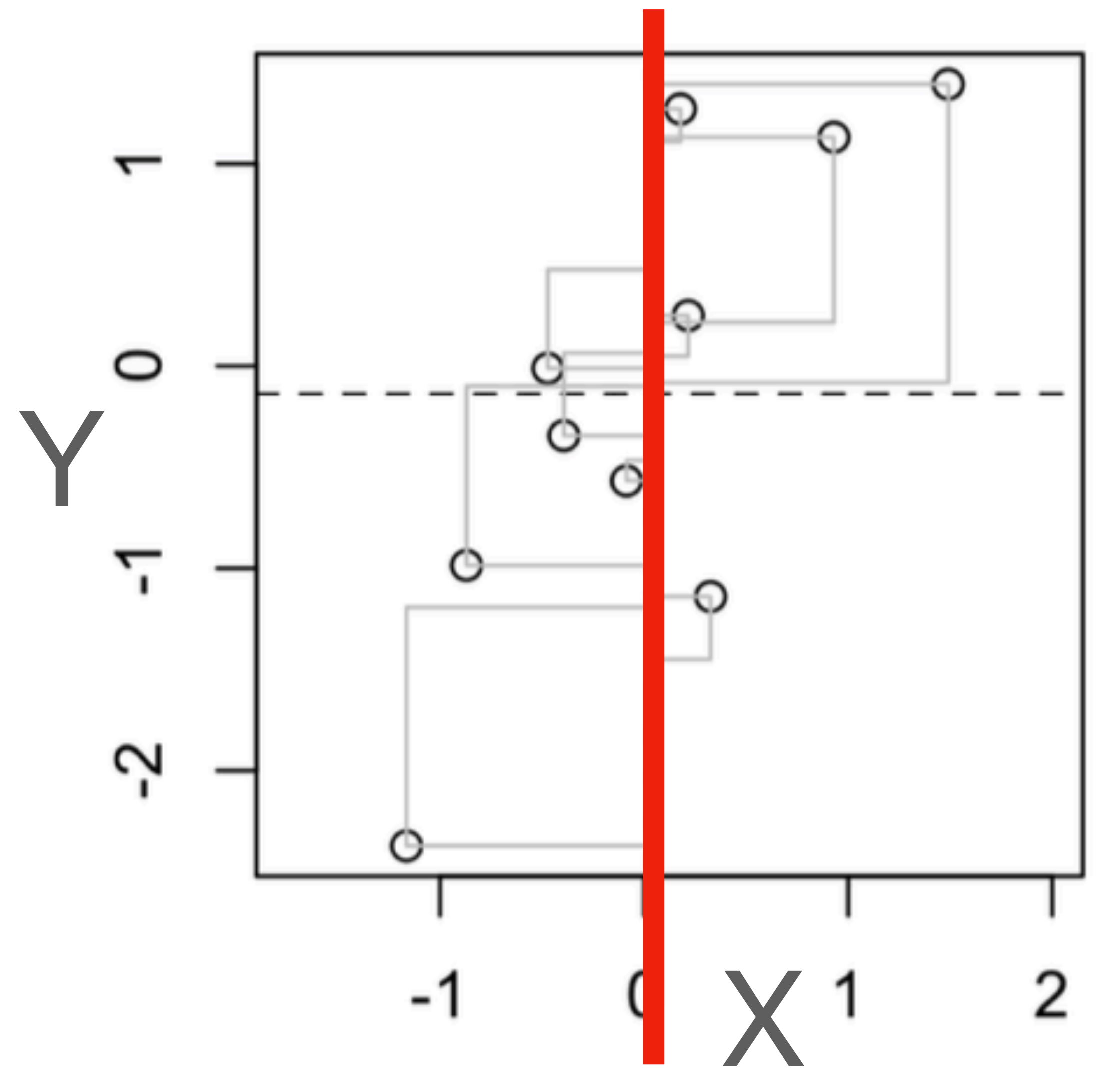


Measuring variation

Sample variance of X:

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

"Sum of squares of X"

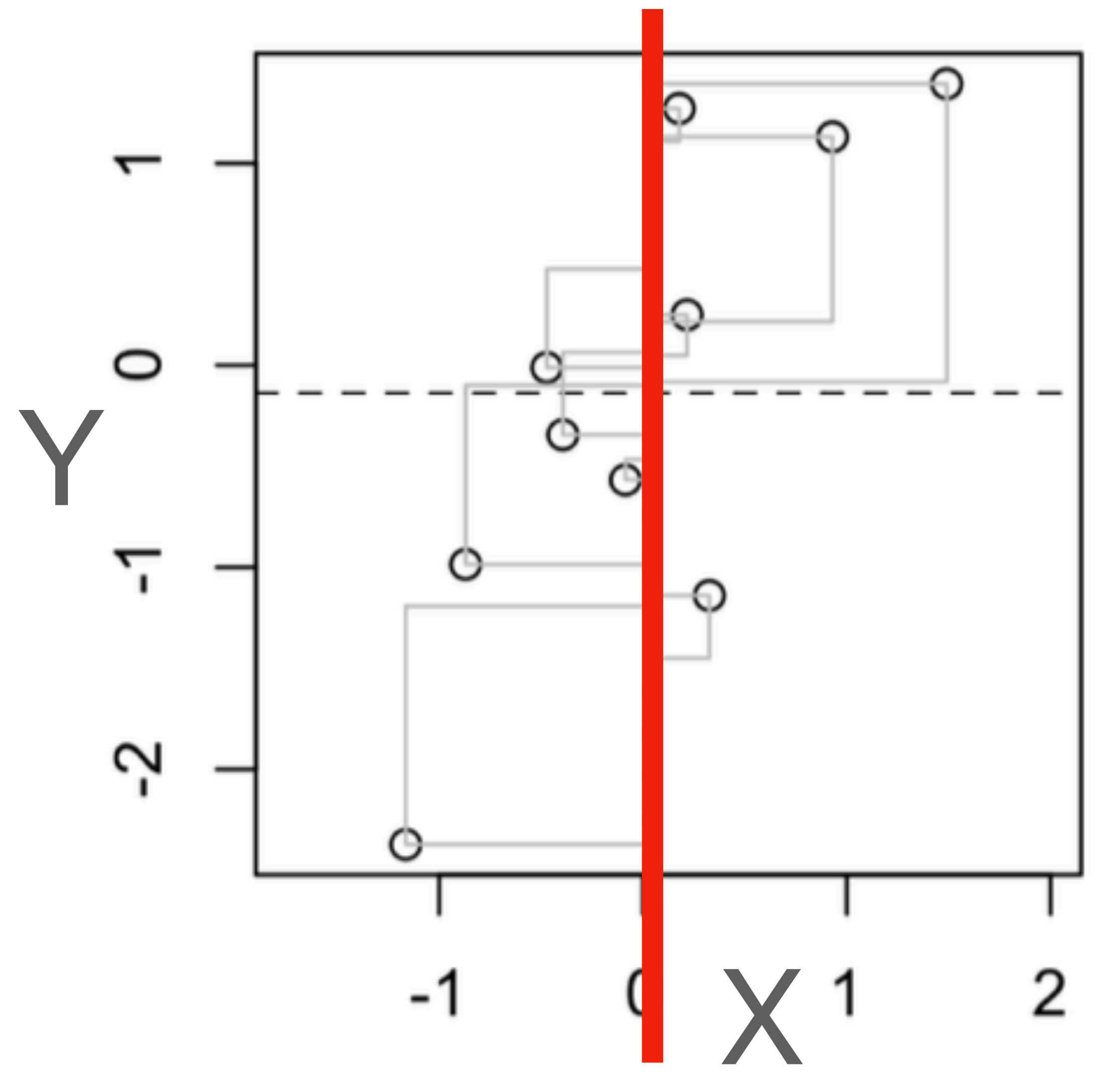


Measuring variation

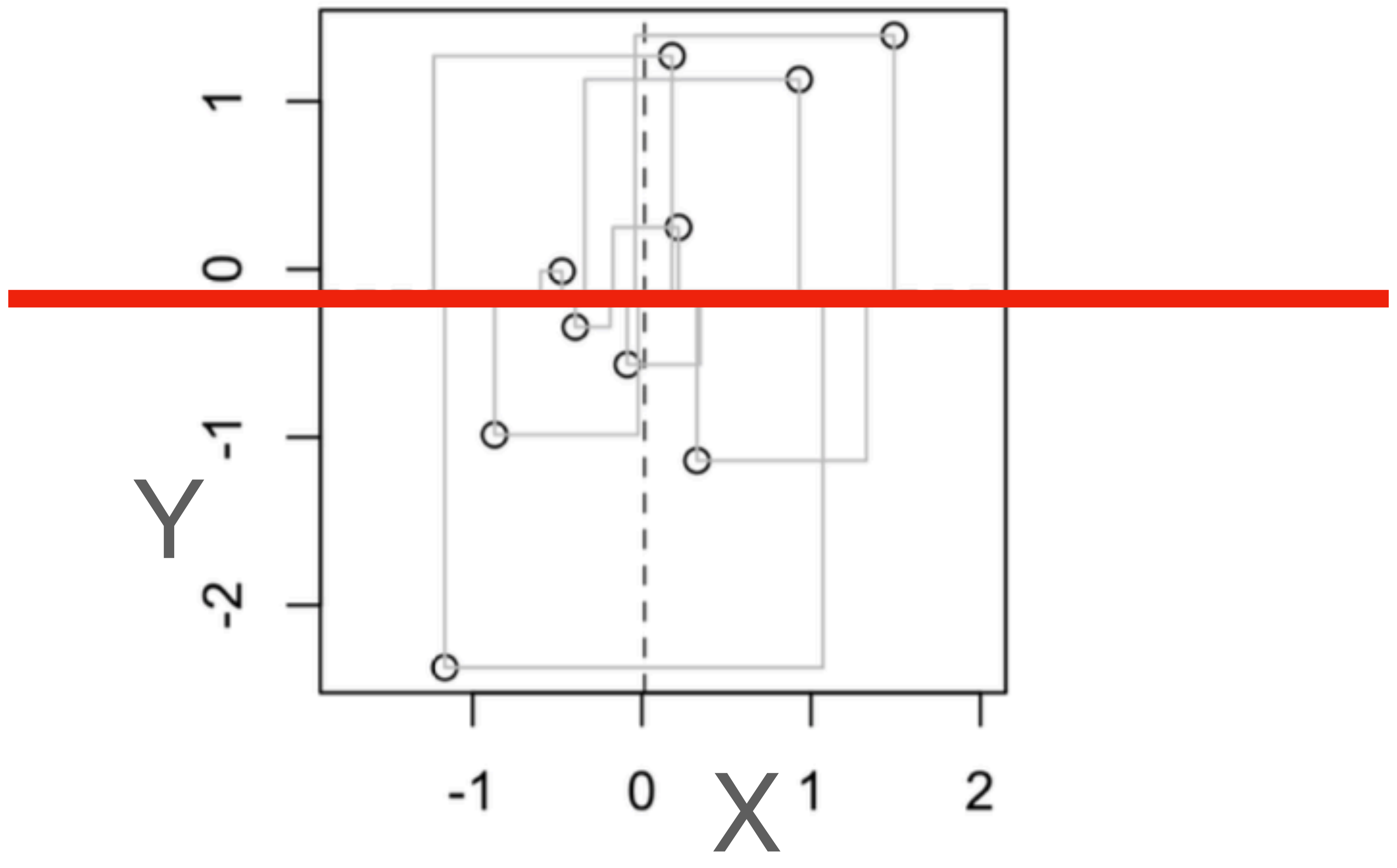
Sample variance of X:

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

“Sum of squares of X”



“Sum of squares of Y”



Measuring variation

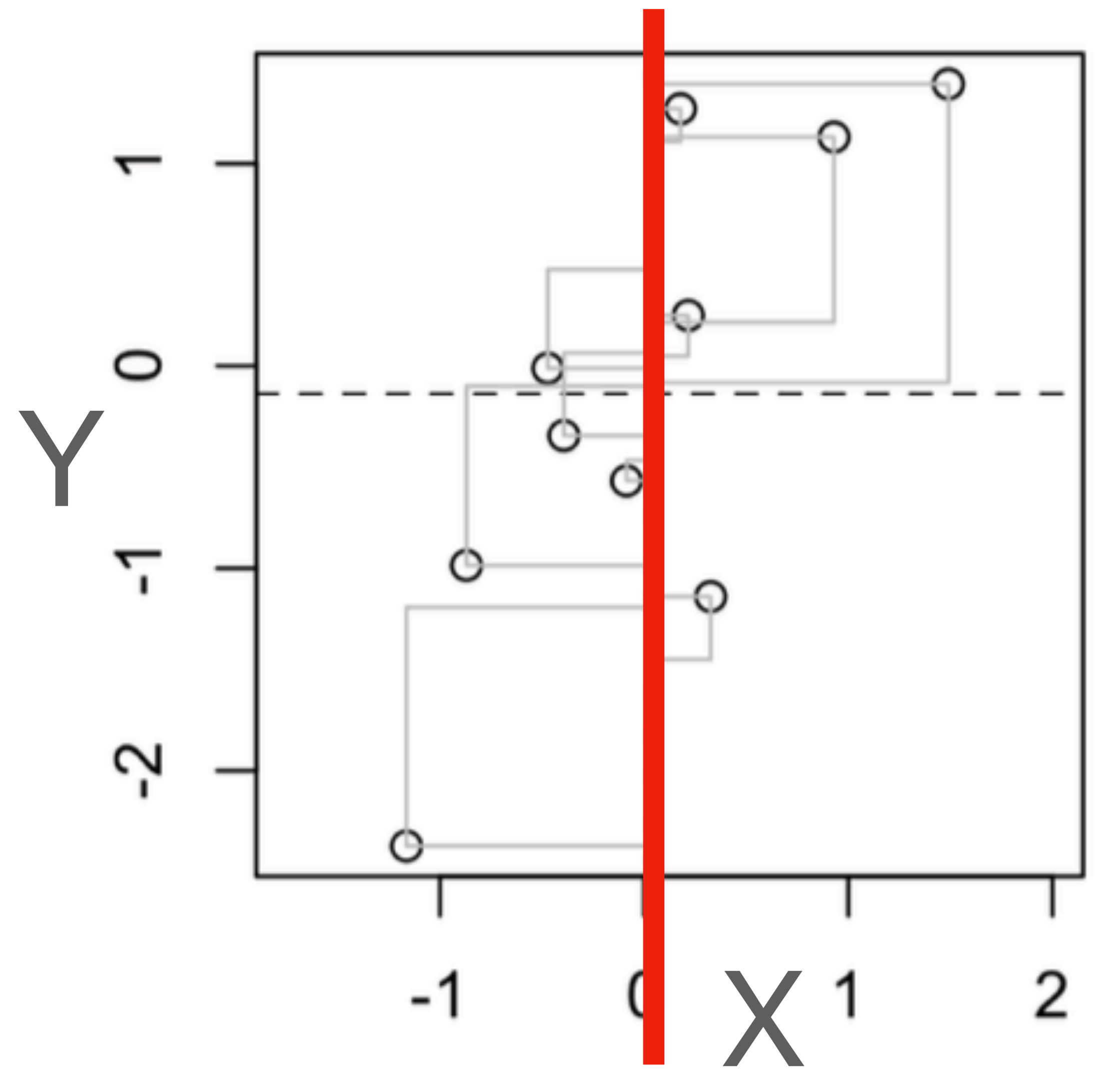
Sample variance of X:

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

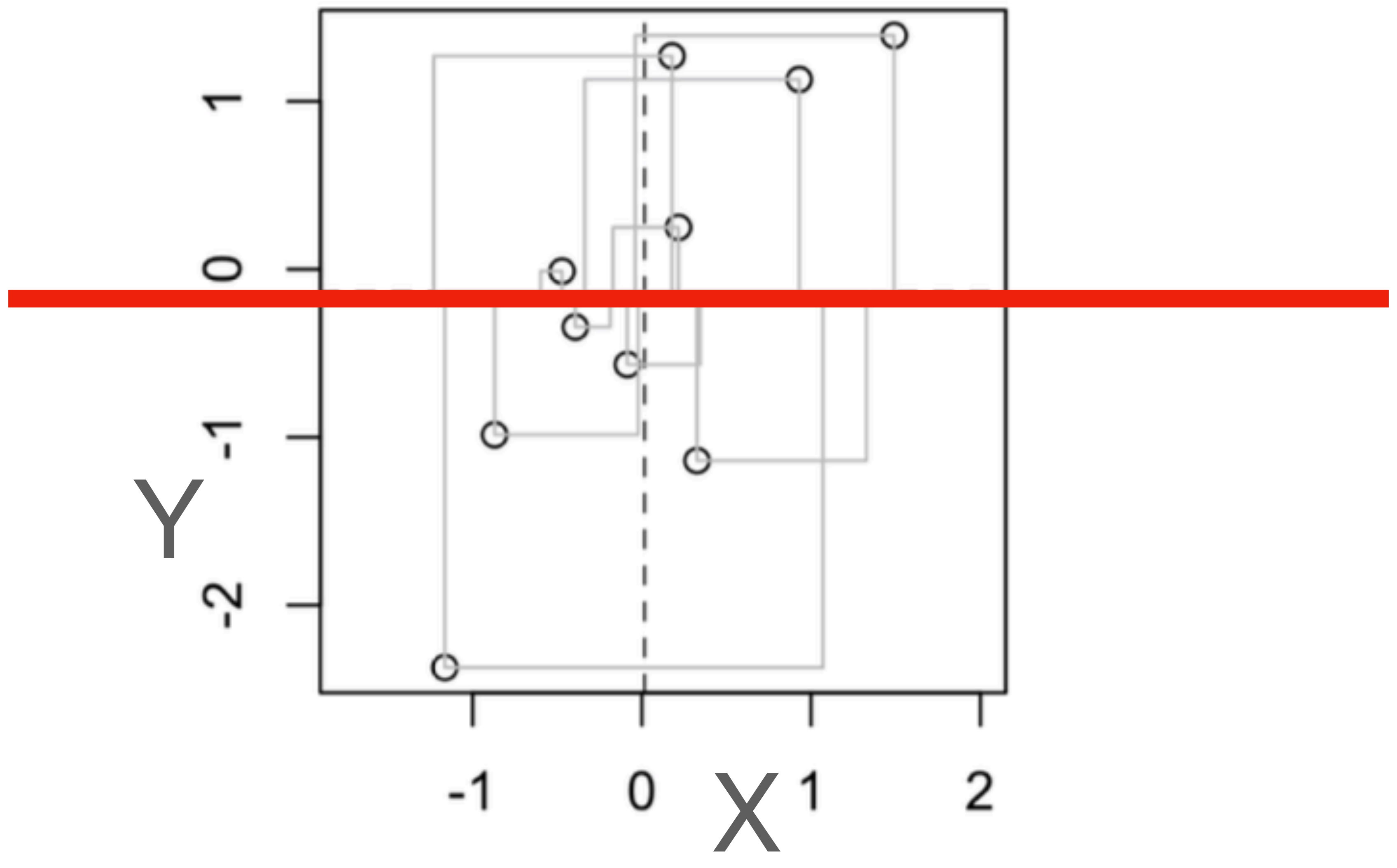
Sample variance of Y:

$$s_Y = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

“Sum of squares of X”



“Sum of squares of Y”

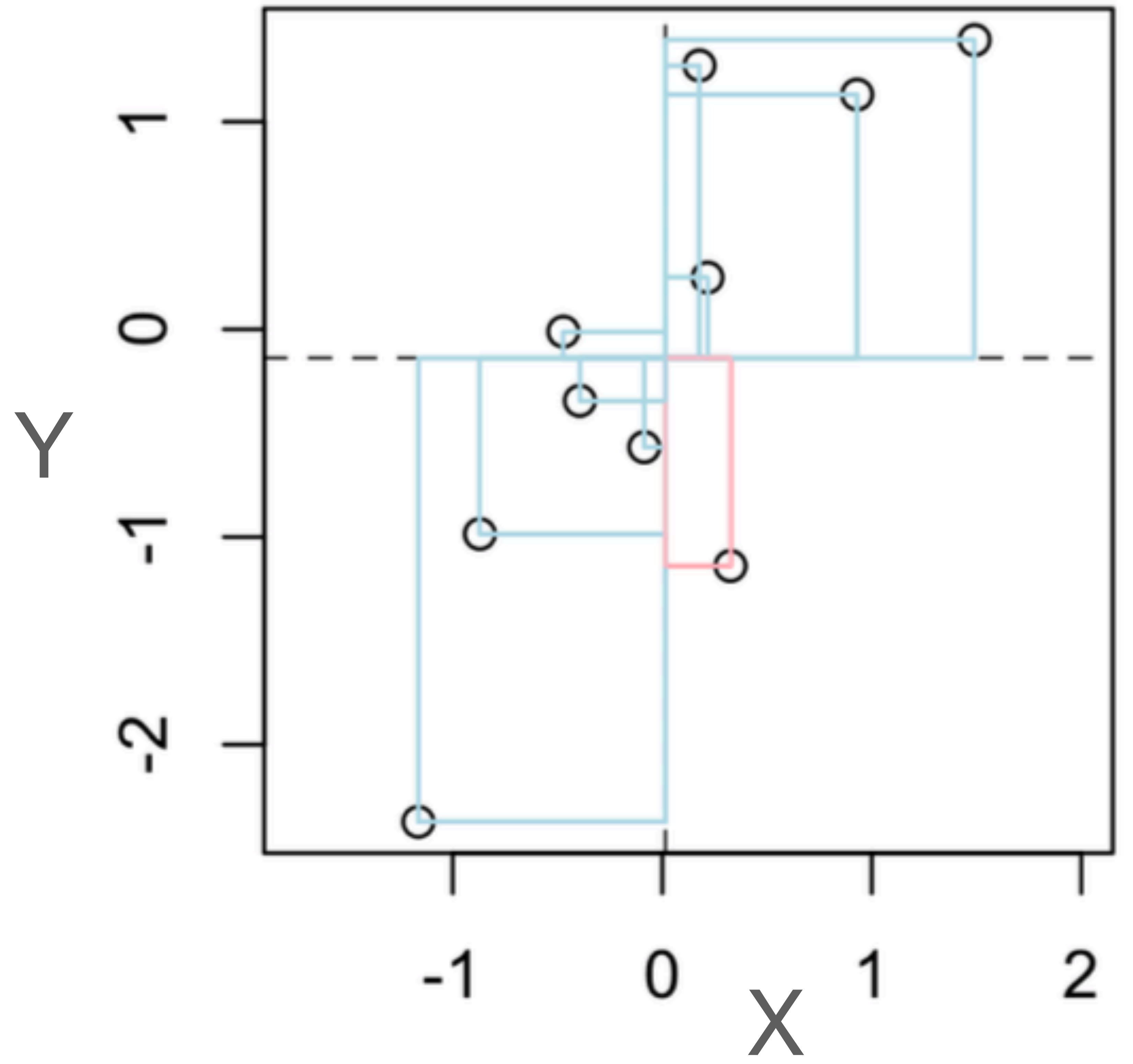


Measuring co-variation

Sample covariance of X and Y:

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

"Sum of XY rectangles"



The Slope

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} [(x_i - \bar{x})^2]}$$

Sample variance of X

The Slope

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}[(x_i - \overline{x})^2]}$$

Sample covariance between X and Y

Sample variance of X

The Slope

The constant n-1 is in both the numerator and denominator, so it cancels out

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} [(x_i - \bar{x})^2]}$$

Sample covariance between X and Y

Sample variance of X

The Slope

The constant n-1 is in both the numerator and denominator, so it cancels out

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} [(x_i - \bar{x})^2]}$$

Sample covariance between X and Y

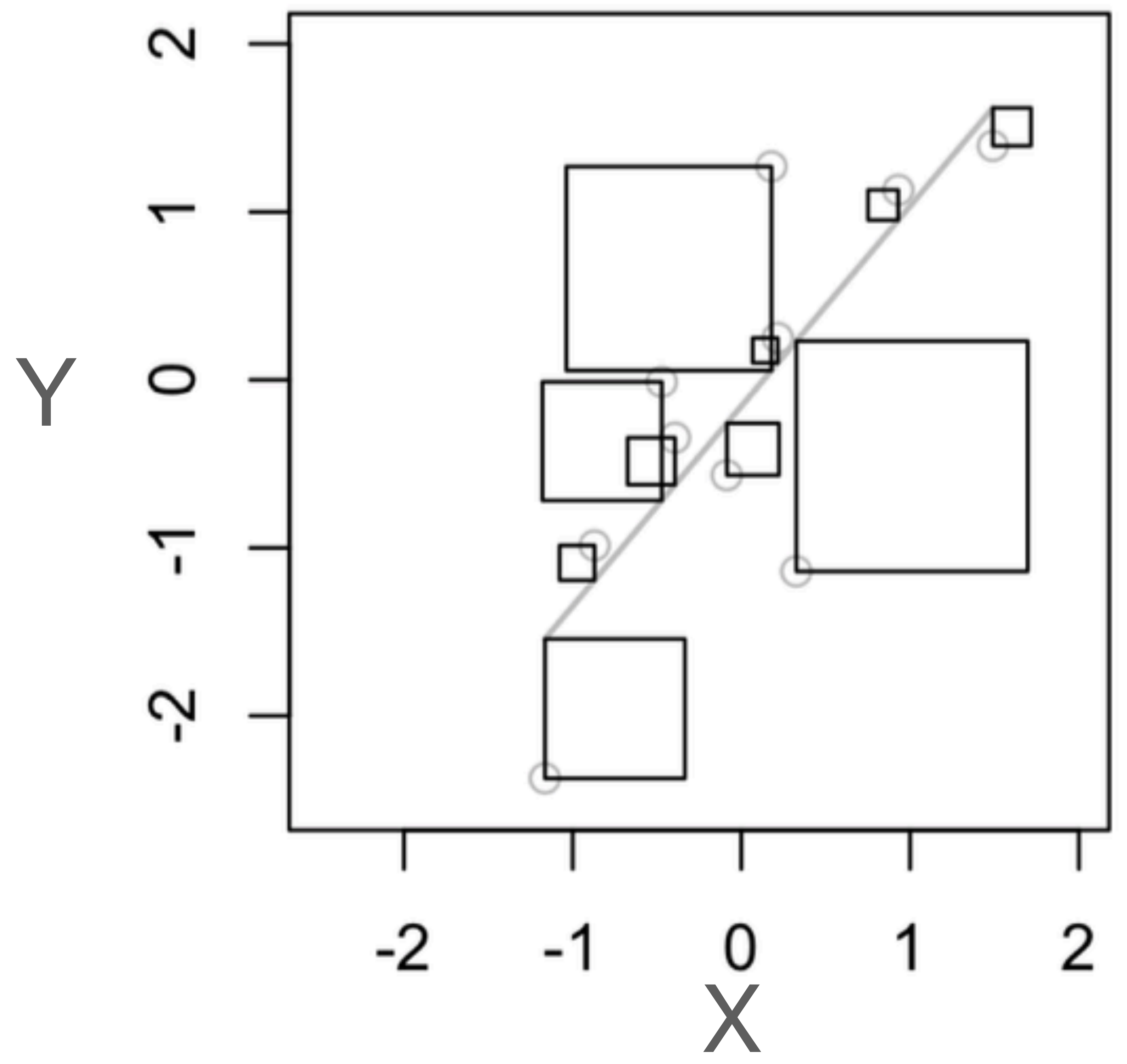
Sample variance of X

The **slope** measures the **covariance between X and Y**, as a proportion of the **variance of X**

How good is our model?

- No model is perfect, but some are more useful than others
- After fitting our model, we still have **unexplained variation in the dependent variable**: the sums of squares of the residuals (SS_{res}).

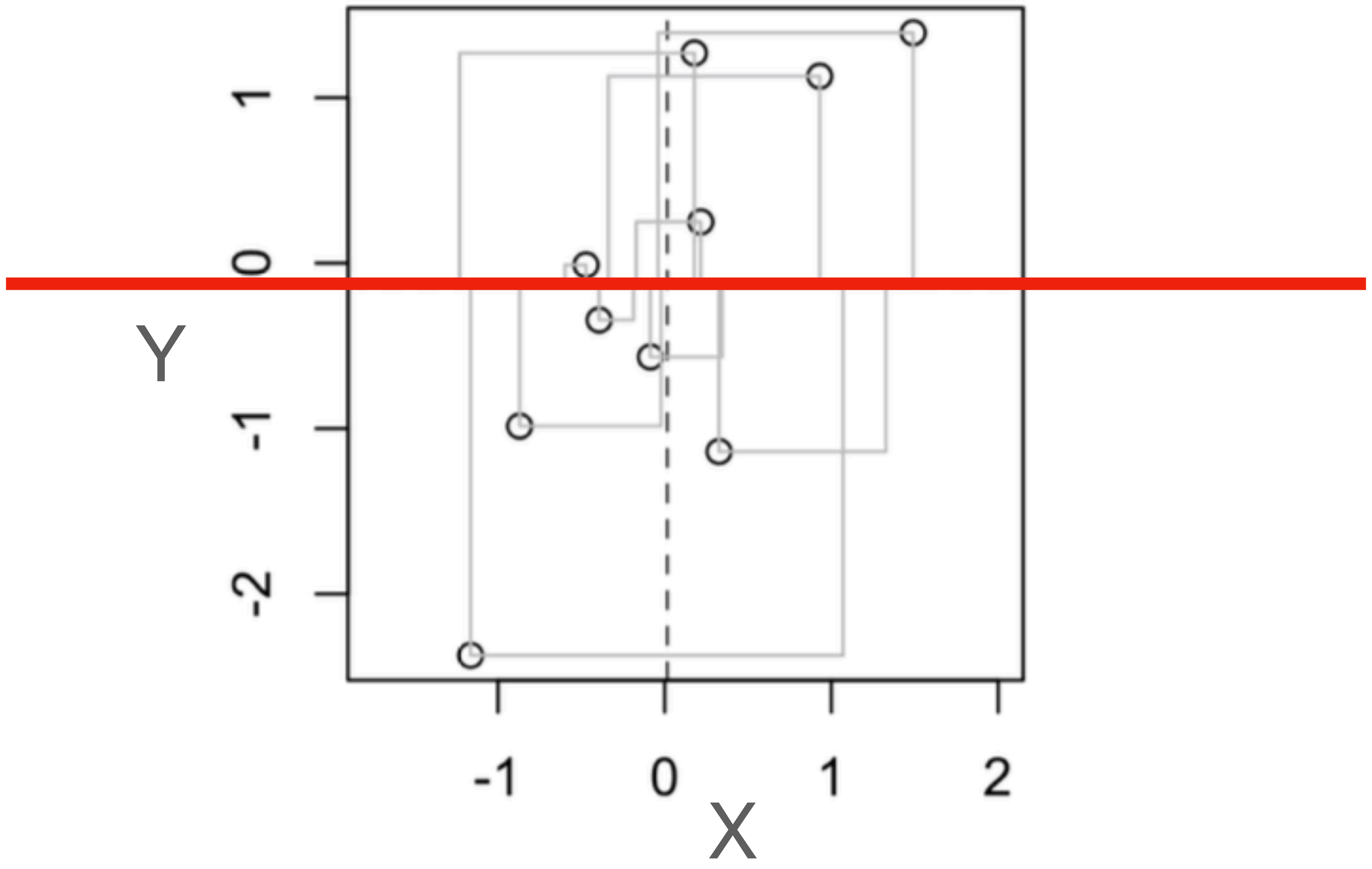
$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



How good is our model?

- We also have a measure of the **total variation in the dependent variable**: the sum of the squares of Y (SS_Y)

$$SS_Y = \sum_{i=1}^{n} (y_i - \overline{y})^2$$



How good is our model?

- A natural measure of fit: the **coefficient of determination (R^2)**

$$R^2 = 1 - \frac{SS_{res}}{SS_Y}$$

How good is our model?

- A natural measure of fit: the **coefficient of determination (R^2)**

$$R^2 = 1 - \frac{SS_{res}}{SS_Y}$$

How **bad** we are at explaining variation in Y

How good is our model?

- A natural measure of fit: the **coefficient of determination (R^2)**

$$R^2 = 1 - \frac{SS_{res}}{SS_Y}$$

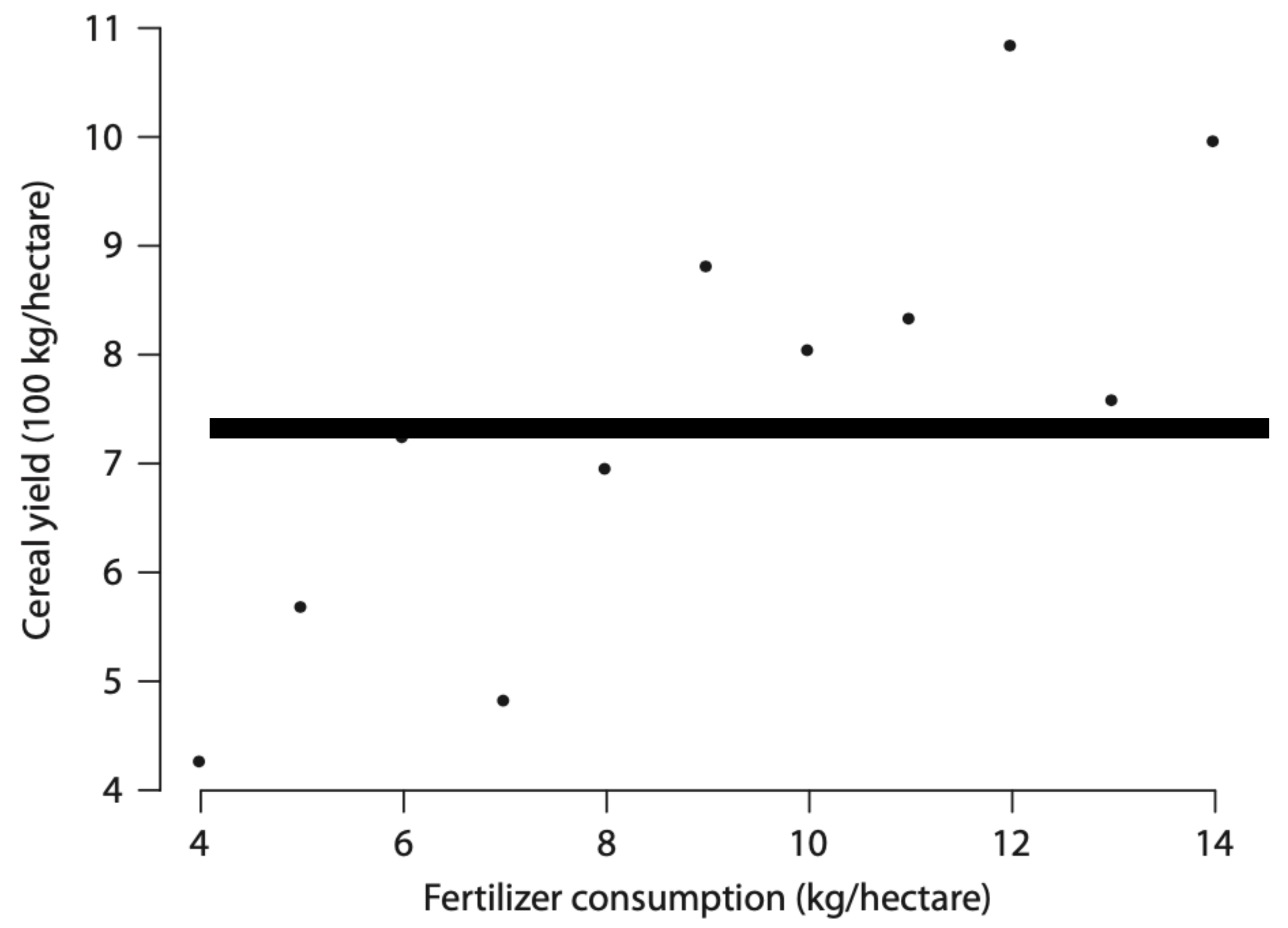
How **good** we are at explaining variation in Y

Hypothesis testing via a linear model

Hypothesis testing

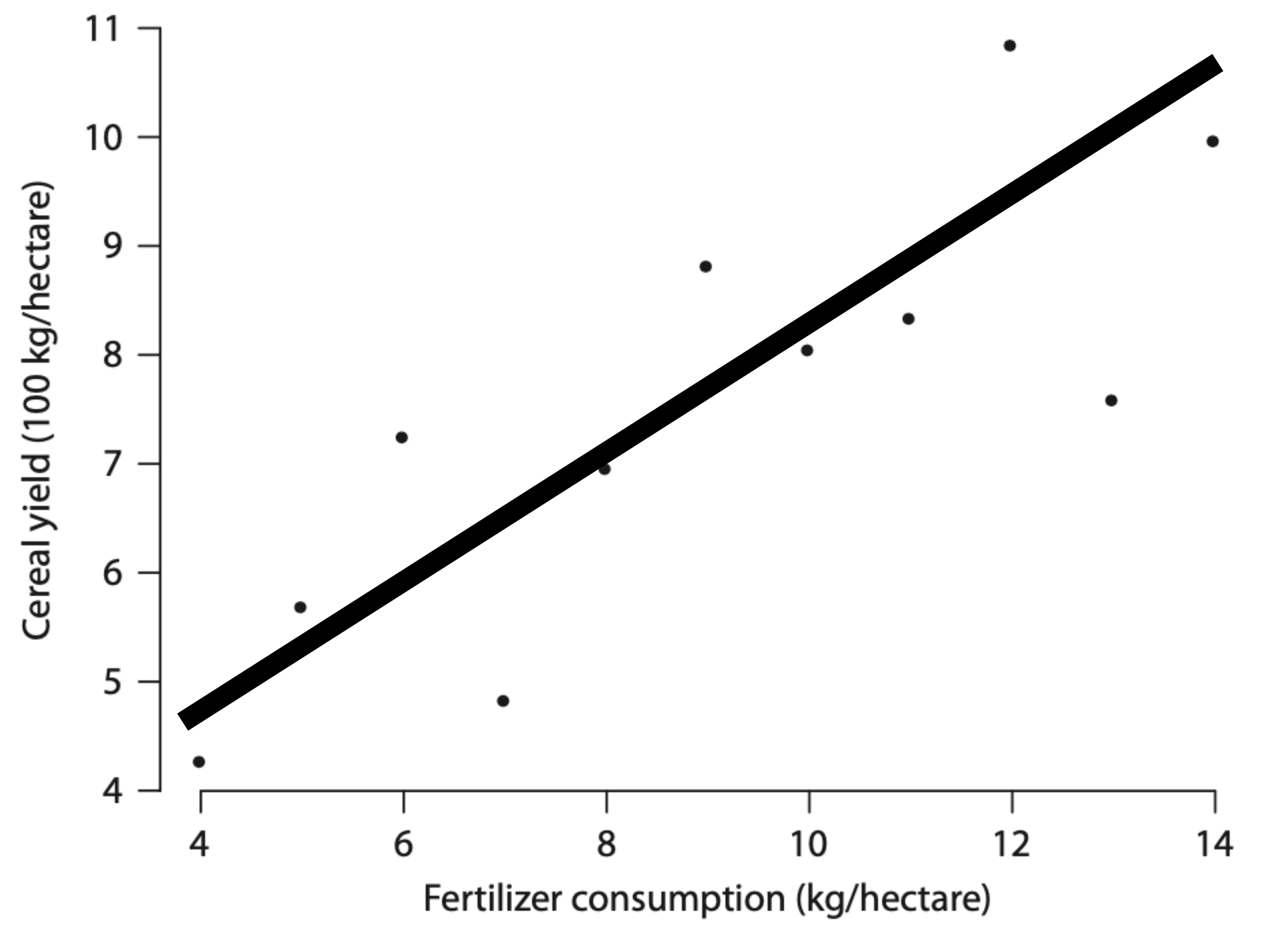
H_0: there is **no** relationship between fertilizer consumption and cereal yield

Null (simpler) hypothesis → Can we reject it?



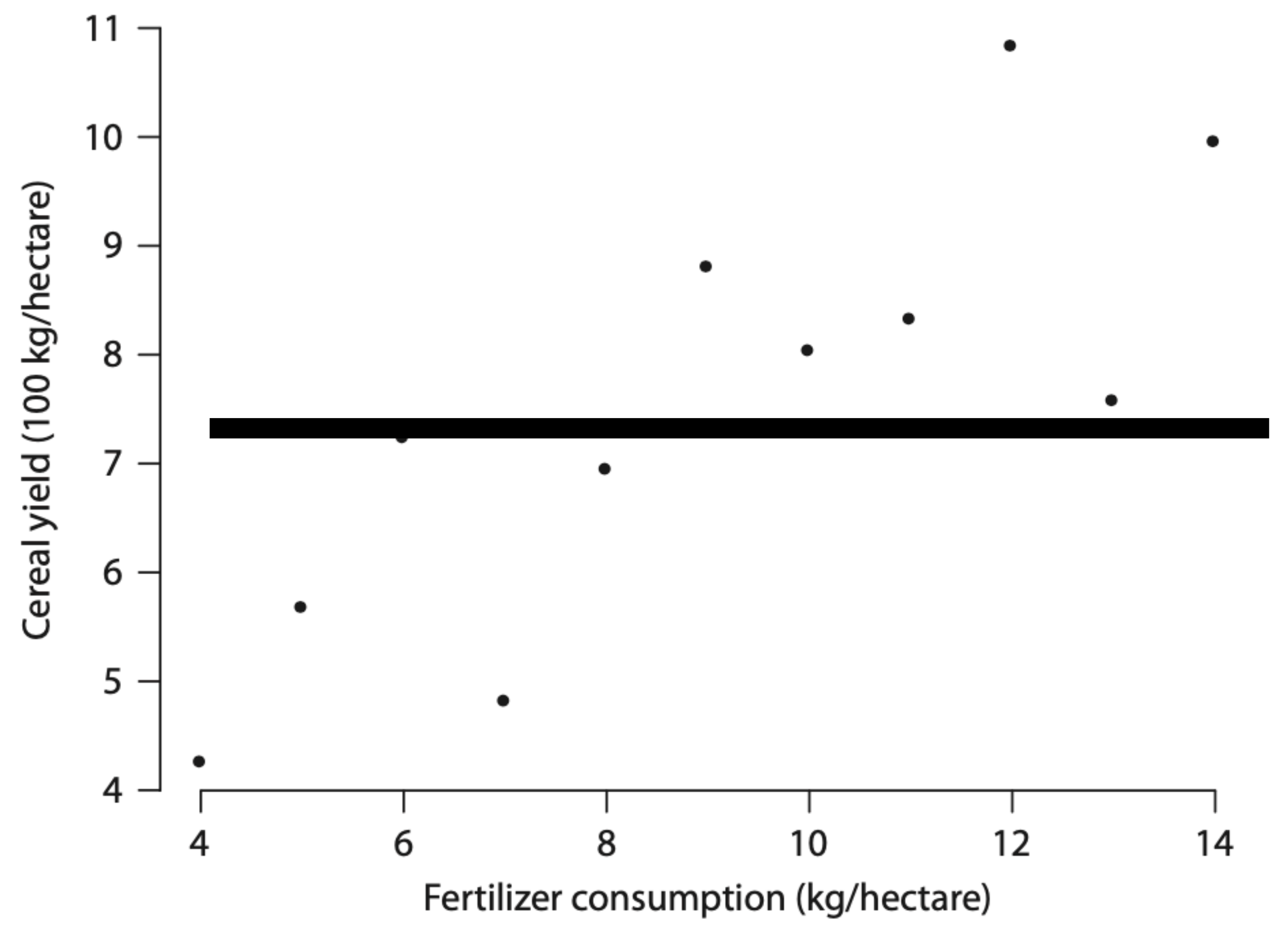
H_1: there is **some** relationship between fertilizer consumption and cereal yield

Alternative (more complicated) hypothesis

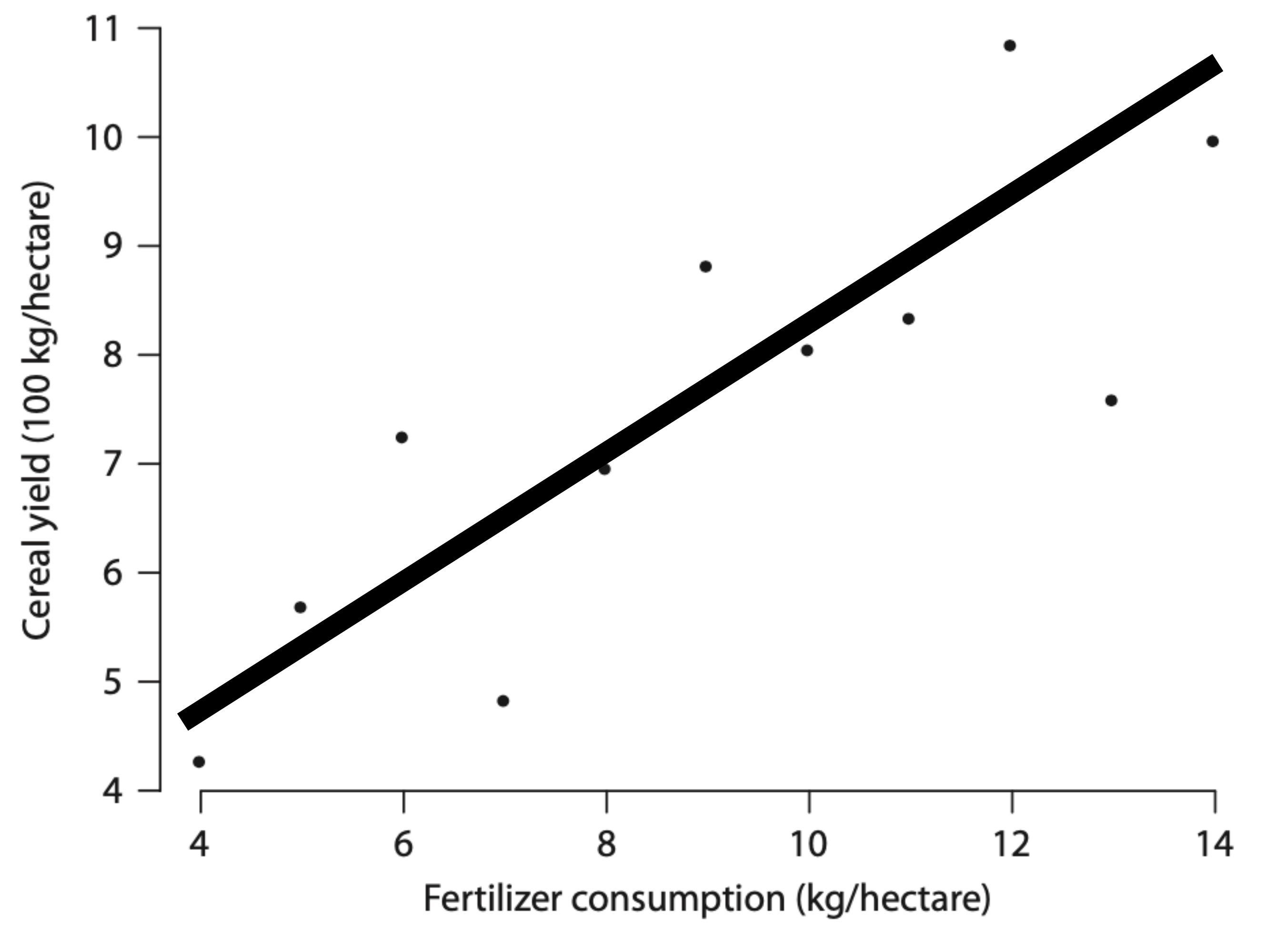


Hypothesis testing

$$\mathbf{y} = \beta_0 + \epsilon$$

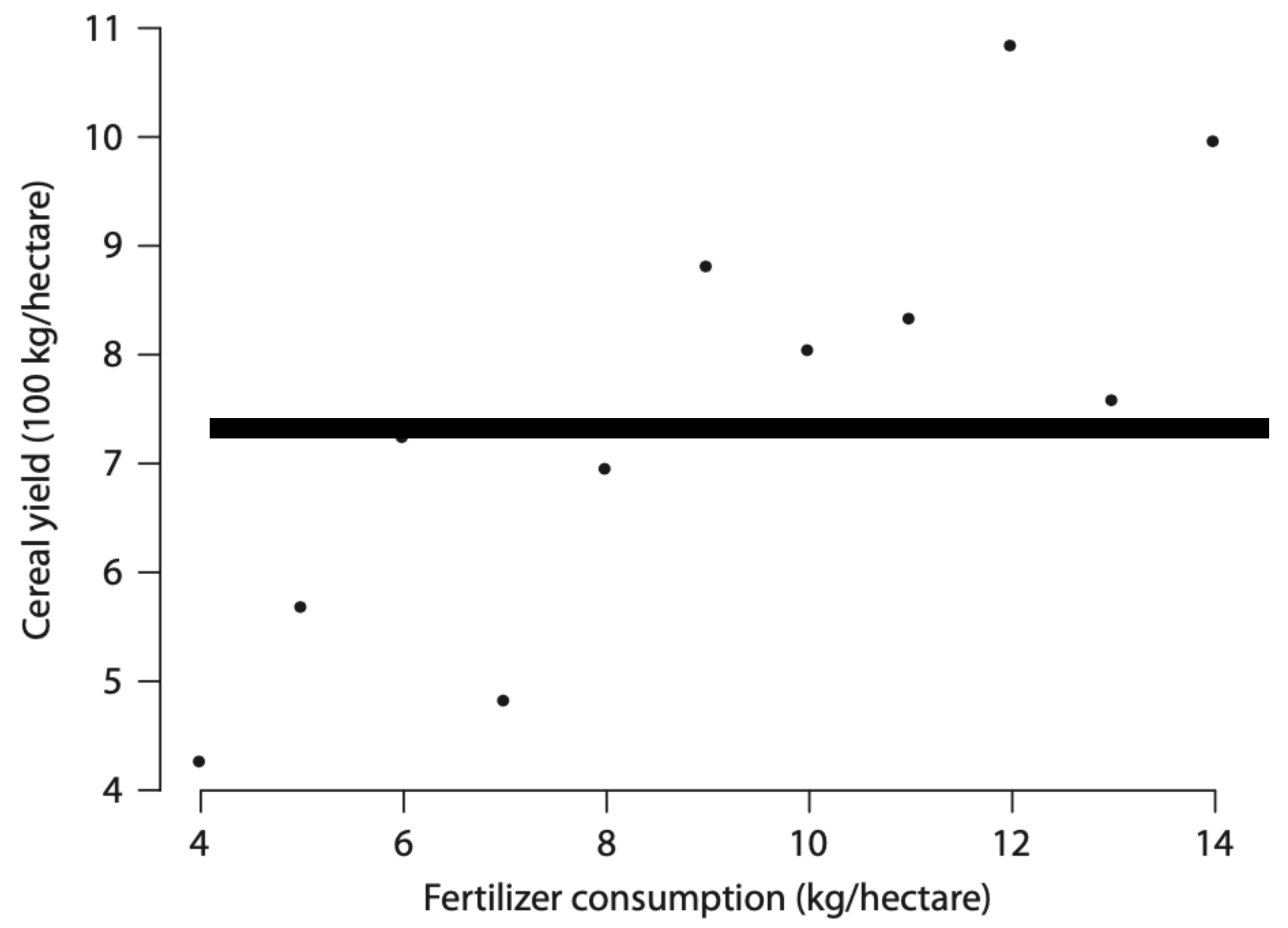


$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \epsilon$$

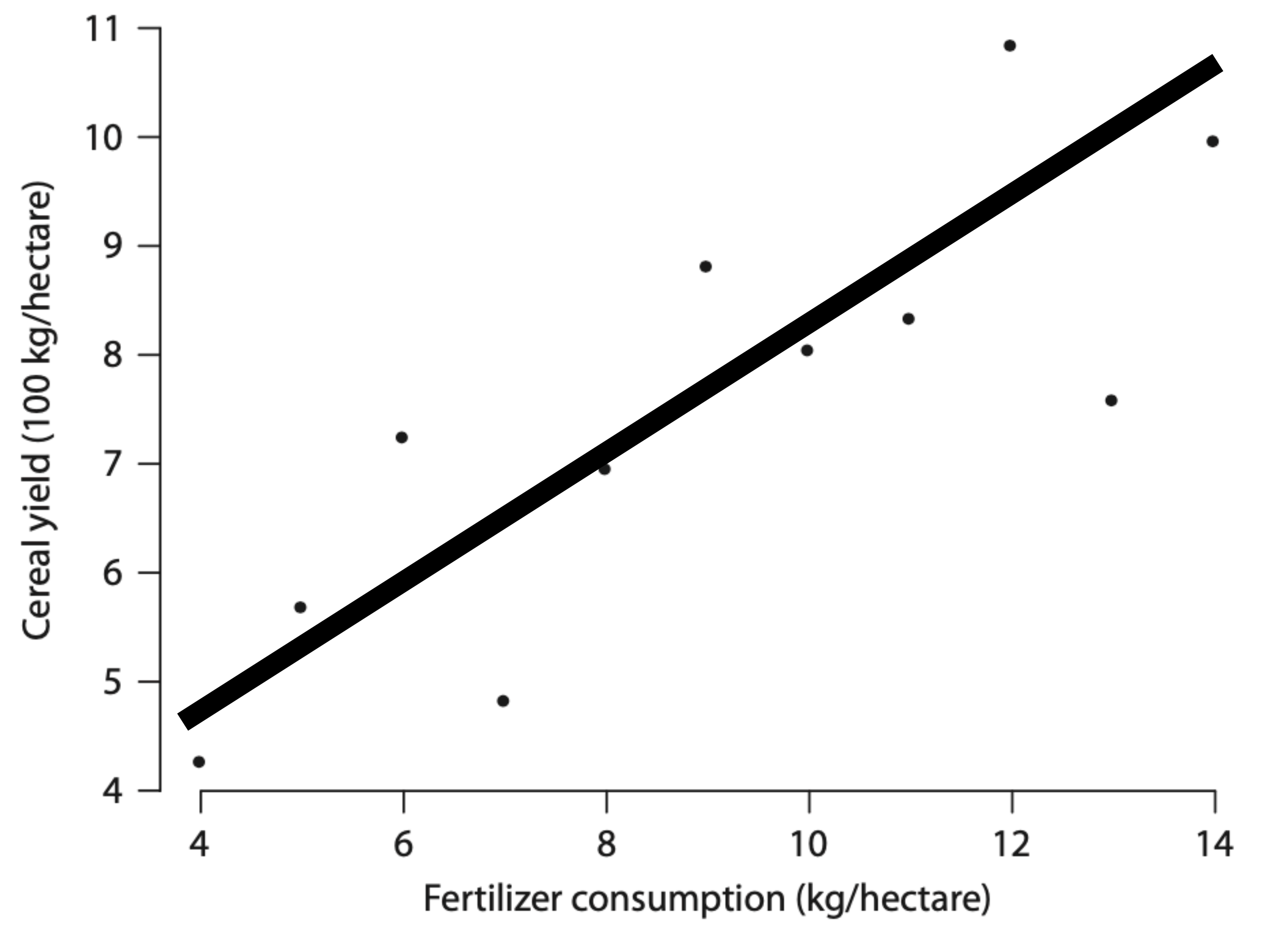


Hypothesis testing

$$\beta_1 = 0$$



$$\beta_1 \neq 0$$



Test statistic: *t*

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Test statistic: *t*

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Difference between our estimate $\hat{\beta}_1$ and 0

Test statistic: *t*

The bigger the difference, the bigger is *t*

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Difference between our estimate $\hat{\beta}_1$ and 0

Test statistic: *t*

The bigger the difference, the bigger is *t*

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Difference between our estimate $\hat{\beta}_1$ and 0

Standard Error: a measure of how inaccurate our estimate $\hat{\beta}_1$ is at estimating β_1

Test statistic: *t*

The bigger the difference, the bigger is *t*

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

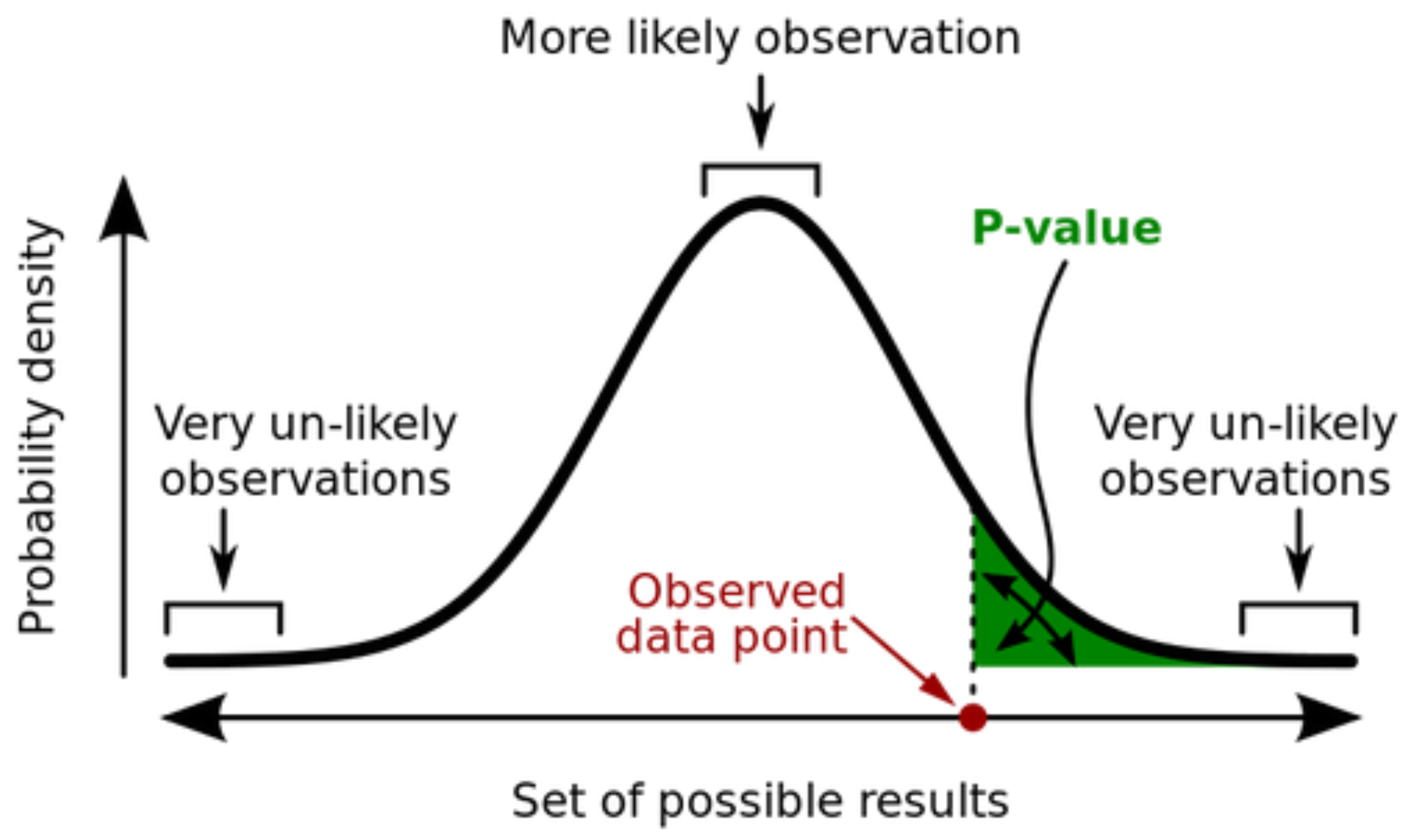
$\longrightarrow$ Difference between our estimate $\hat{\beta}_1$ and 0

$\longrightarrow$ Standard Error: a measure of how inaccurate our estimate $\hat{\beta}_1$ is at estimating β_1

The bigger the error, the smaller is *t*

Test statistic: *t*

Assuming $\beta_1 = 0$, then the t-statistic should follow a well-known distribution: the t-distribution.



If our estimate is **too unlikely** under this distribution, then we can **reject the hypothesis** $\beta_1 = 0$

Frequentist hypothesis testing

Assuming the null hypothesis is true…

The data is unlikely ($p < 0.05$)

The data is plausible ($p > 0.05$)

We reject the null hypothesis

We cannot reject the null hypothesis

Problems with hypothesis testing

- Who decides what is “too unlikely”? ($p < 0.05$ is an arbitrary cutoff)
- Could a “rejected” model still be a plausible explanation of the data?
- If we “reject” a null model, are we necessarily “accepting” the alternative?
- What if there are multiple alternative models?

Problems with hypothesis testing

- Who decides what is “too unlikely”? ($p < 0.05$ is an arbitrary cutoff)
- Could a “rejected” model still be a plausible explanation of the data?
- If we “reject” a null model, are we necessarily “accepting” the alternative?
- What if there are multiple alternative models?

Frequentist hypothesis testing

Assuming the null hypothesis is true…

The data is unlikely ($p < 0.05$)

We reject the null hypothesis

The data is plausible ($p > 0.05$)

We cannot reject the null hypothesis

Frequentist hypothesis testing

Assuming the null hypothesis is true…

The data is unlikely ($p < 0.05$)

We reject the null hypothesis

The data is plausible ($p > 0.05$)

We cannot reject the null hypothesis

Bayesian model choice

Given the data…

Model 1 has probability 23% of being true

Model 2 has probability 72% of being true

Model 2 is more likely than model 1

Frequentist hypothesis testing

Assuming the null hypothesis is true…

- The data is unlikely ($p < 0.05$) → We reject the null hypothesis
- The data is plausible ($p > 0.05$) → We cannot reject the null hypothesis

Bayesian model choice

Given the data…

- Model 1 has probability 23% of being true
- Model 2 has probability 72% of being true

→ Model 2 is more likely than model 1

To find out more: take the course "Advanced Topics in Data Analysis"!

- Fitting a simple linear model in R
- Interpreting a linear model in R

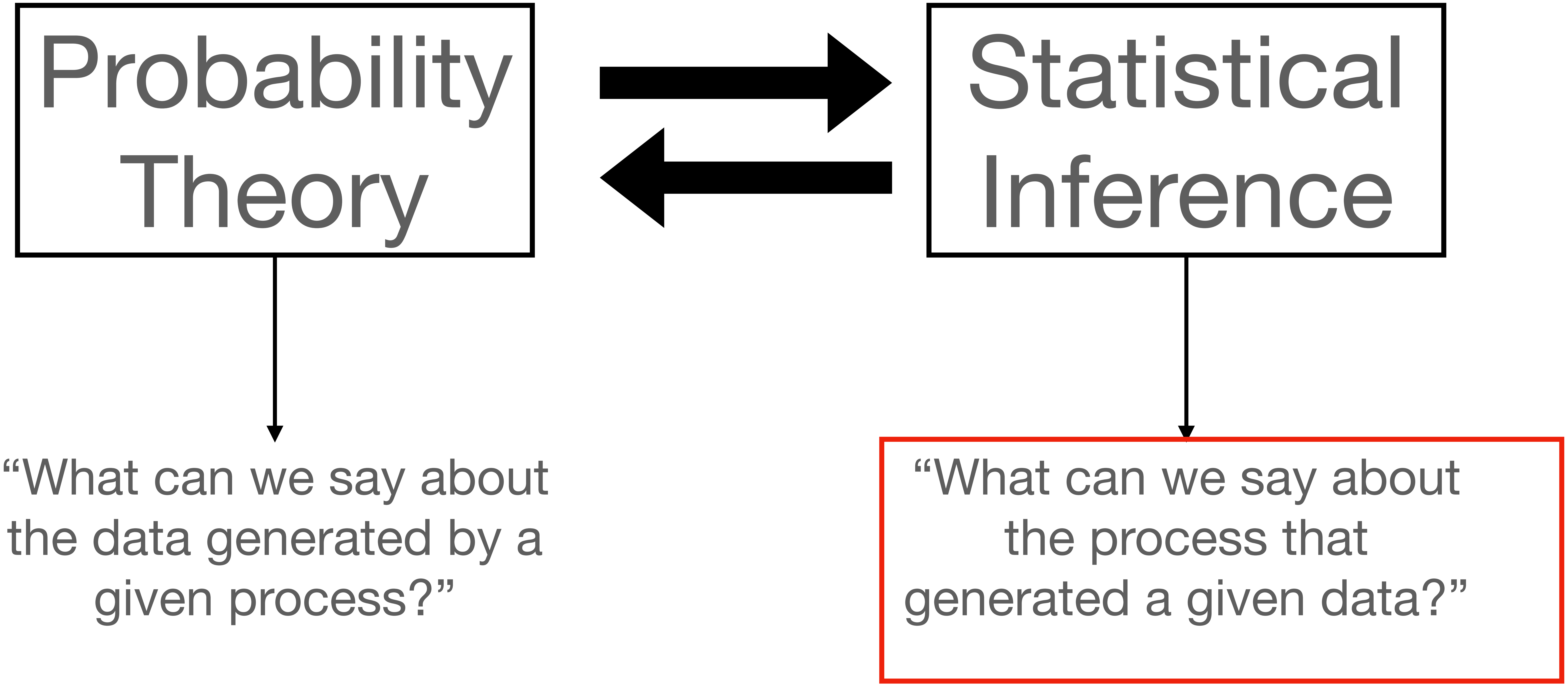
The two sides of statistics

Probability Theory ⇄ **Statistical Inference**

Probability Theory: "What can we say about the data generated by a given process?"

Statistical Inference: "What can we say about the process that generated a given data?"

The two sides of statistics



Linear Regression: a probabilistic interpretation

Question: What can we say about Y after we've fitted our model?

?

Linear Regression: a probabilistic interpretation

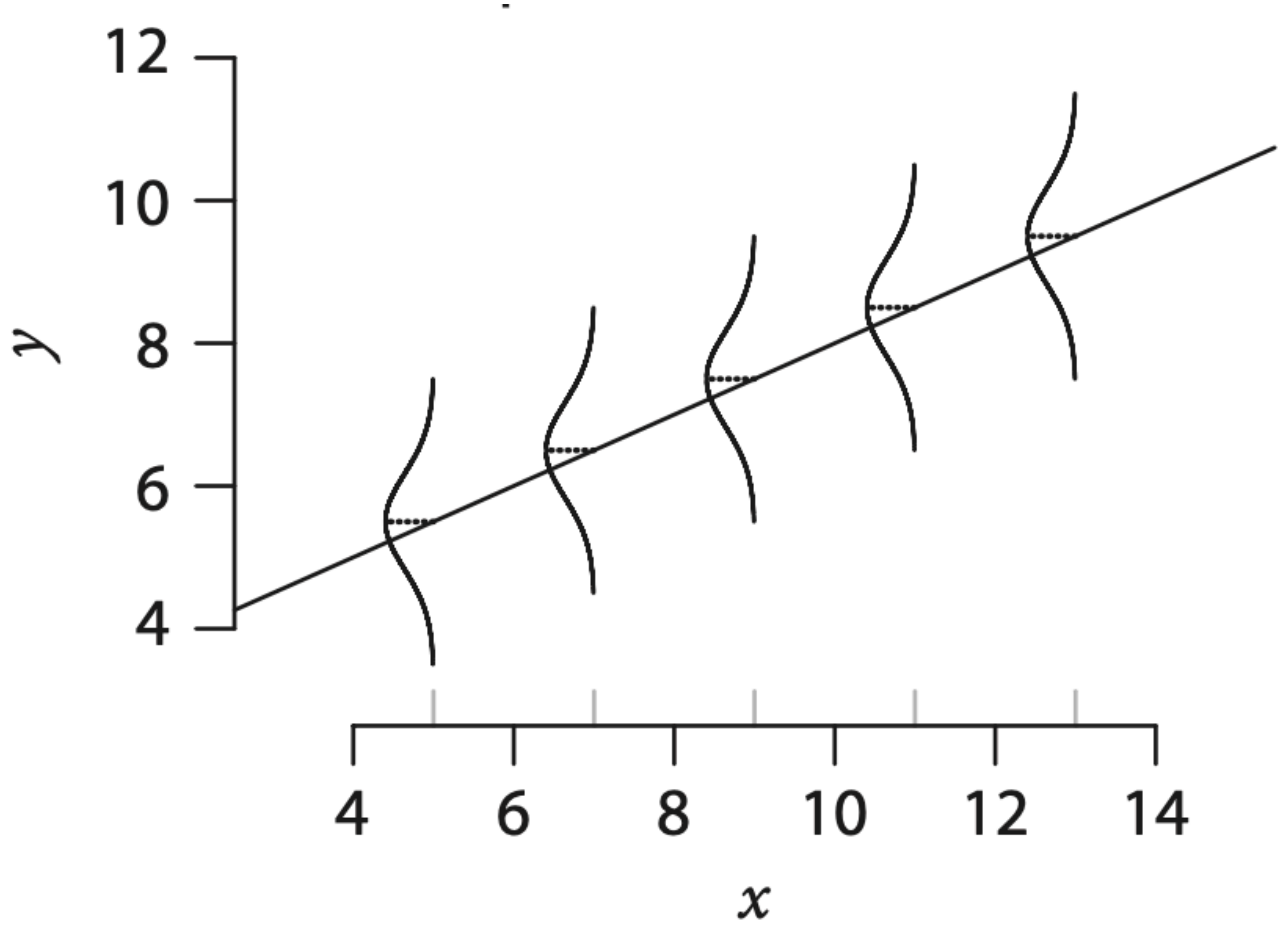
Question: What can we say about Y after we've fitted our model?

!

Answer: It depends on what we're willing to assume…

Linear Regression: a probabilistic interpretation

Let's treat the variable Y as a random variable with some (possibly unknown) distribution



Edge 2019

“Linearity” assumption

$$E[\epsilon \,|\, X = x] = 0$$

“The expected value of the error term is zero, regardless of the value of X”

“Linearity” assumption

$$E[\epsilon \,|\, X = x] = 0$$

“The expected value of the error term is zero, regardless of the value of X”

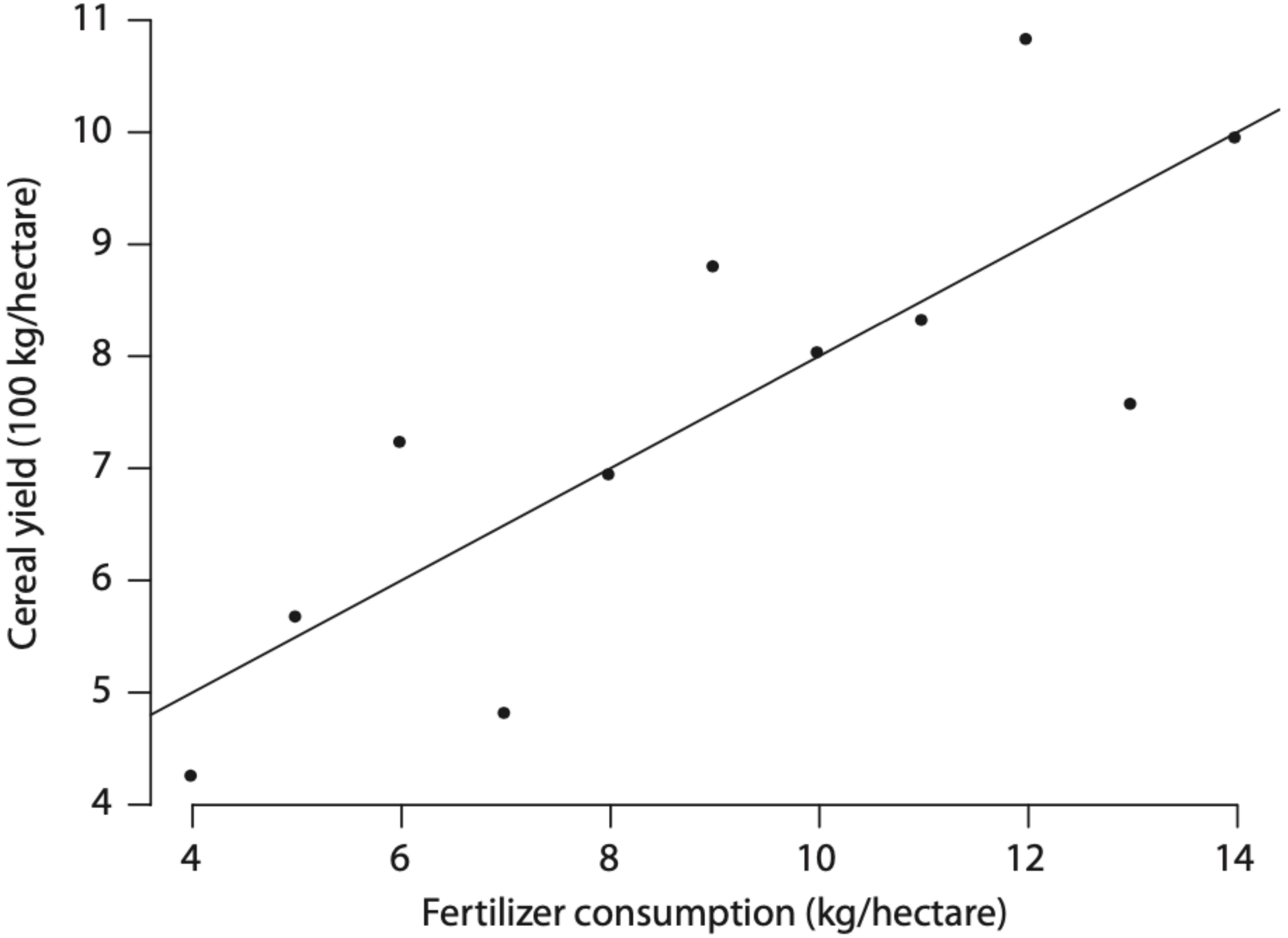
Then:

$$E[Y \,|\, X = x] = \beta_0 + \beta_1 x$$

“Given a value of X, the expected value of Y is a linear function of that value of X”

E[Y|X=x] is the best linear predictor for X=x

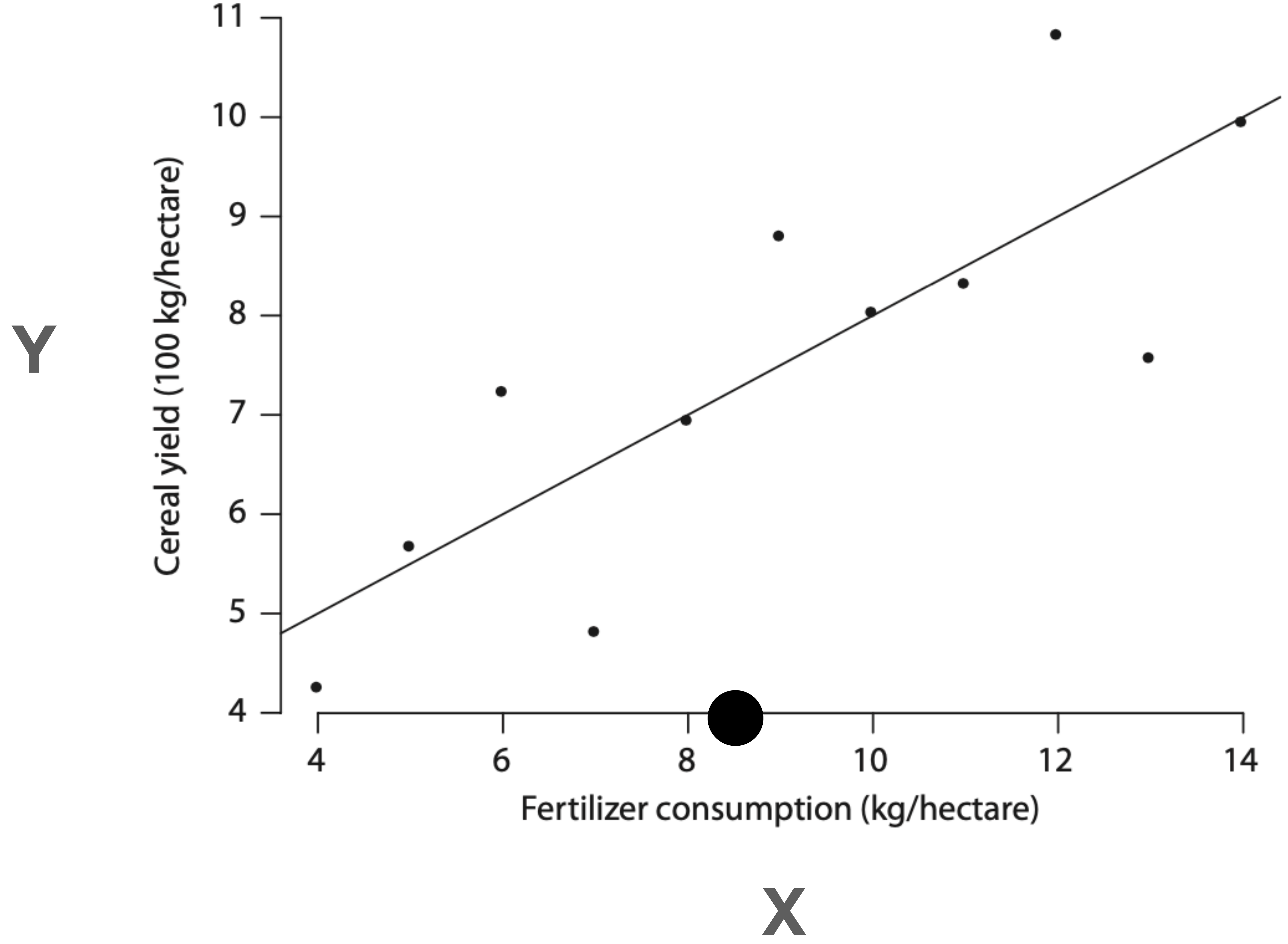
Y



X

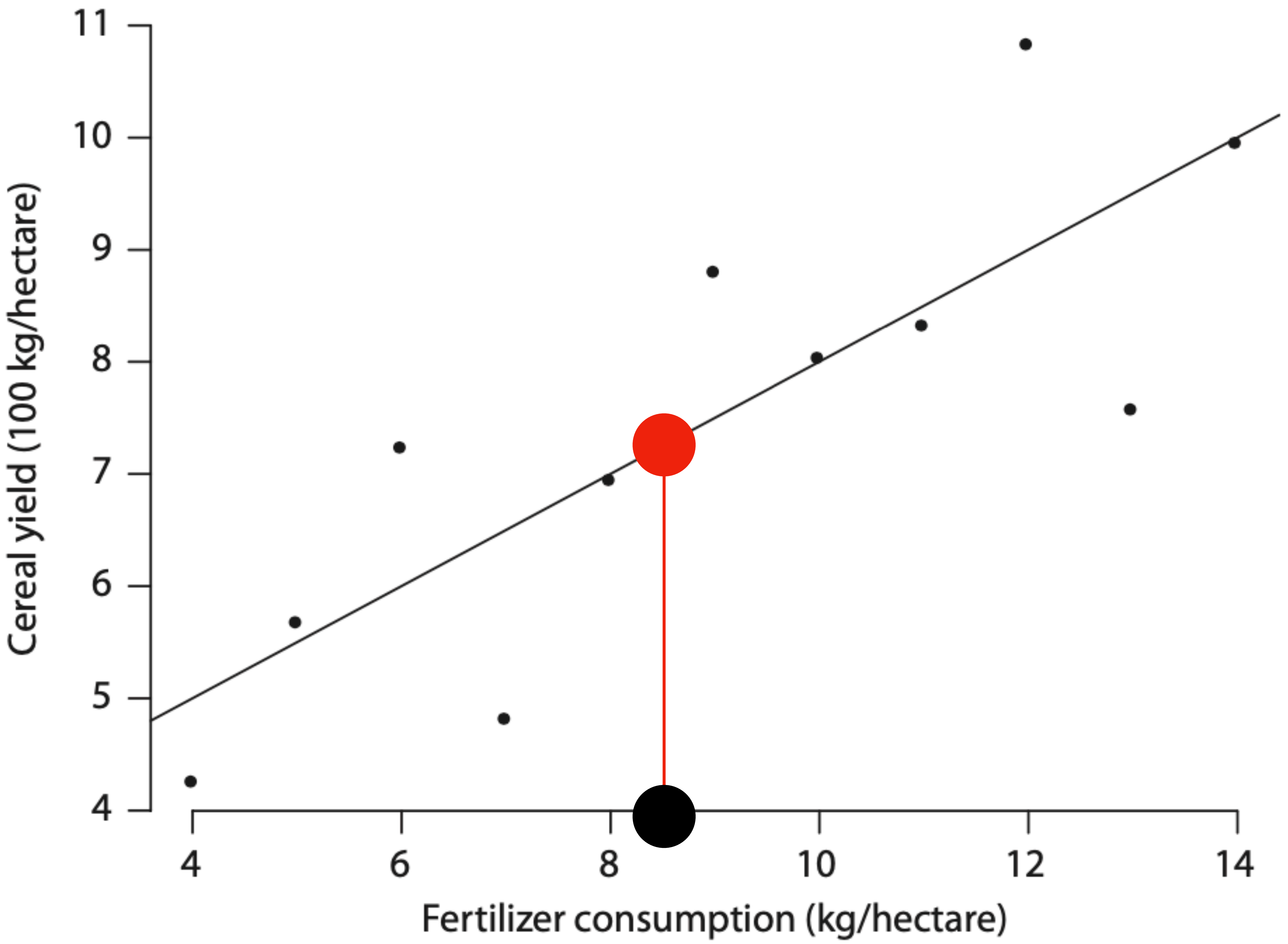
Edge 2019

E[Y|X=x] is the best linear predictor for X=x



Edge 2019

E[Y|X=x] is the best linear predictor for X=x

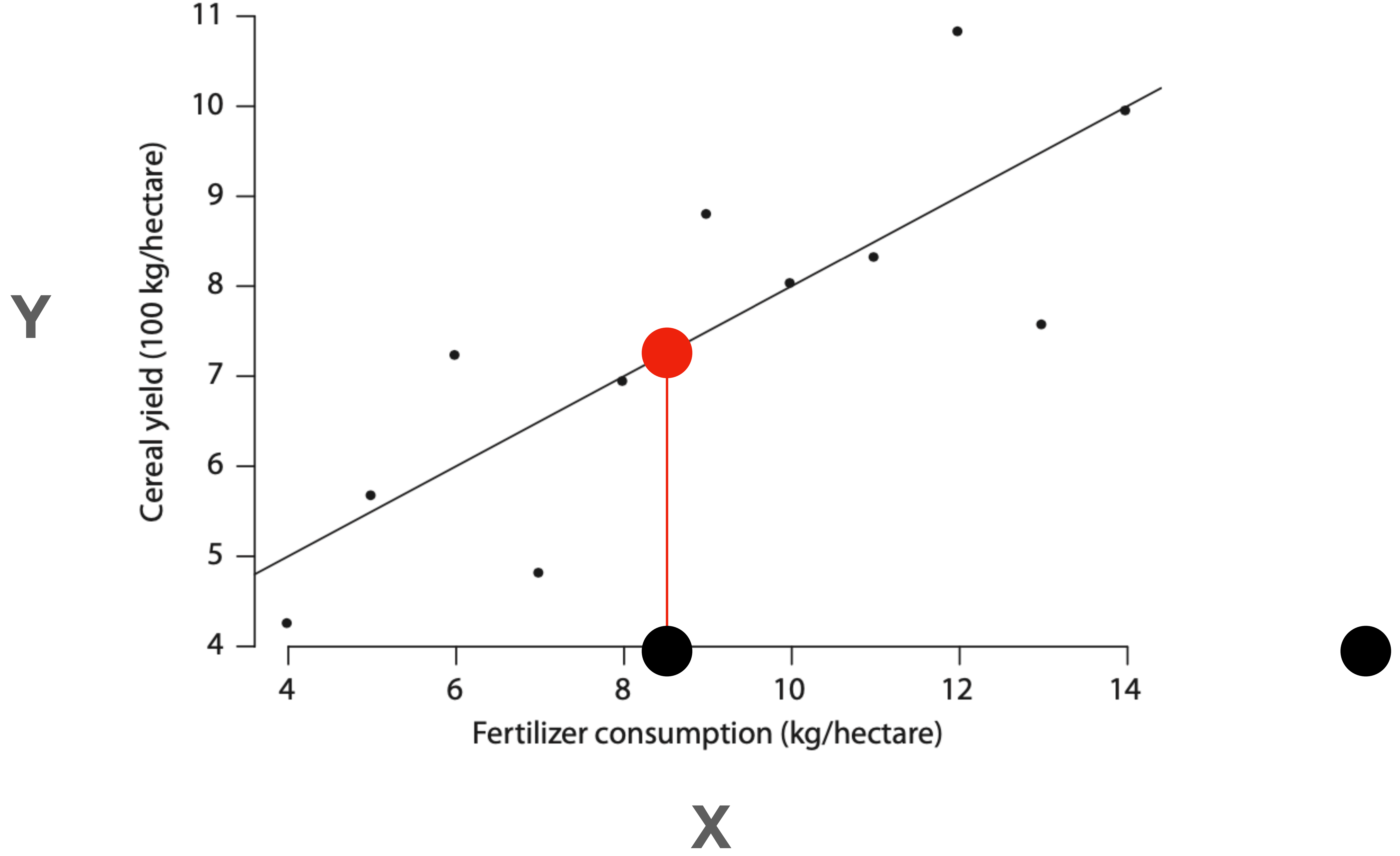


Y

X

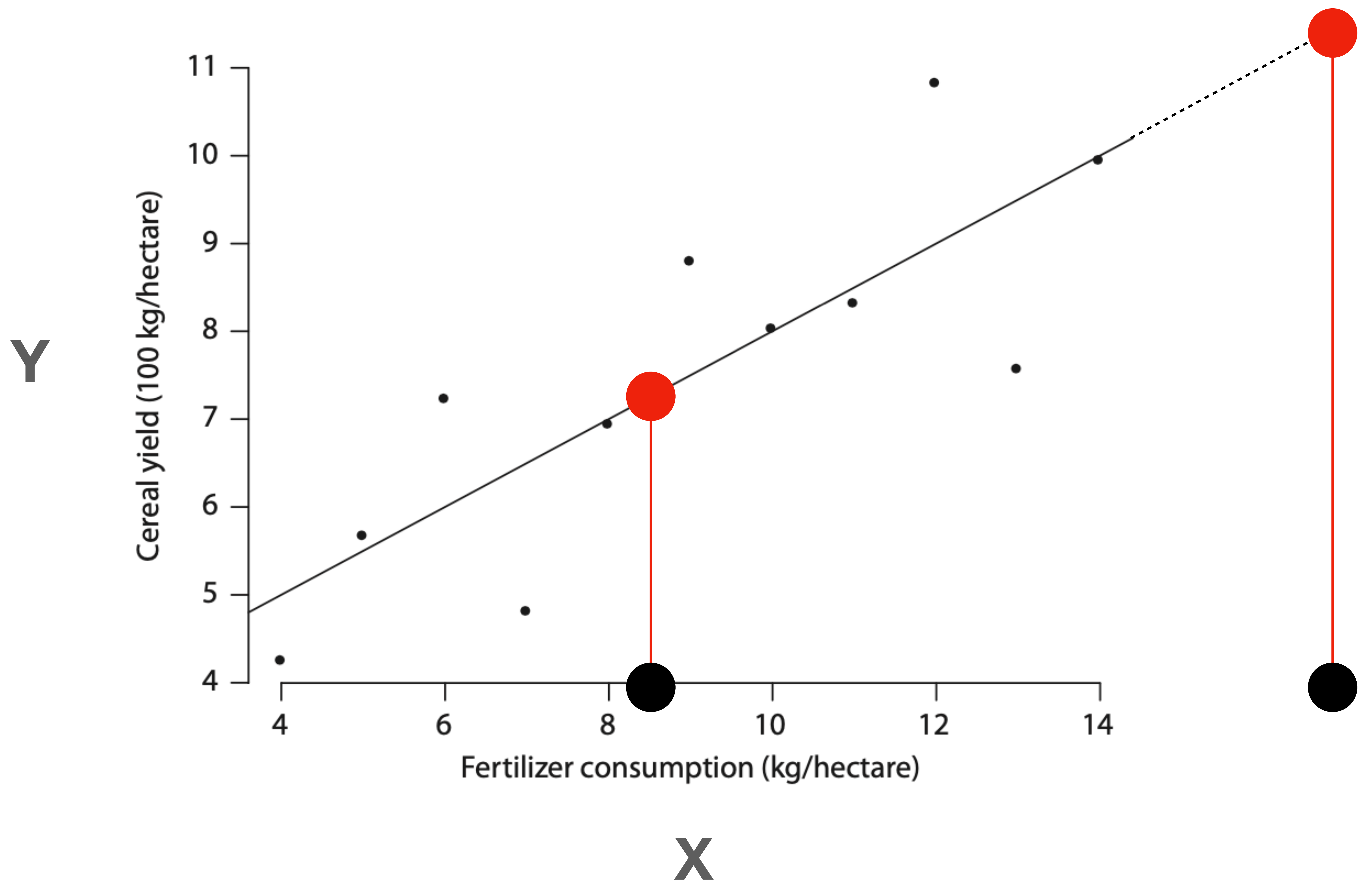
Edge 2019

E[Y|X=x] is the best linear predictor for X=x



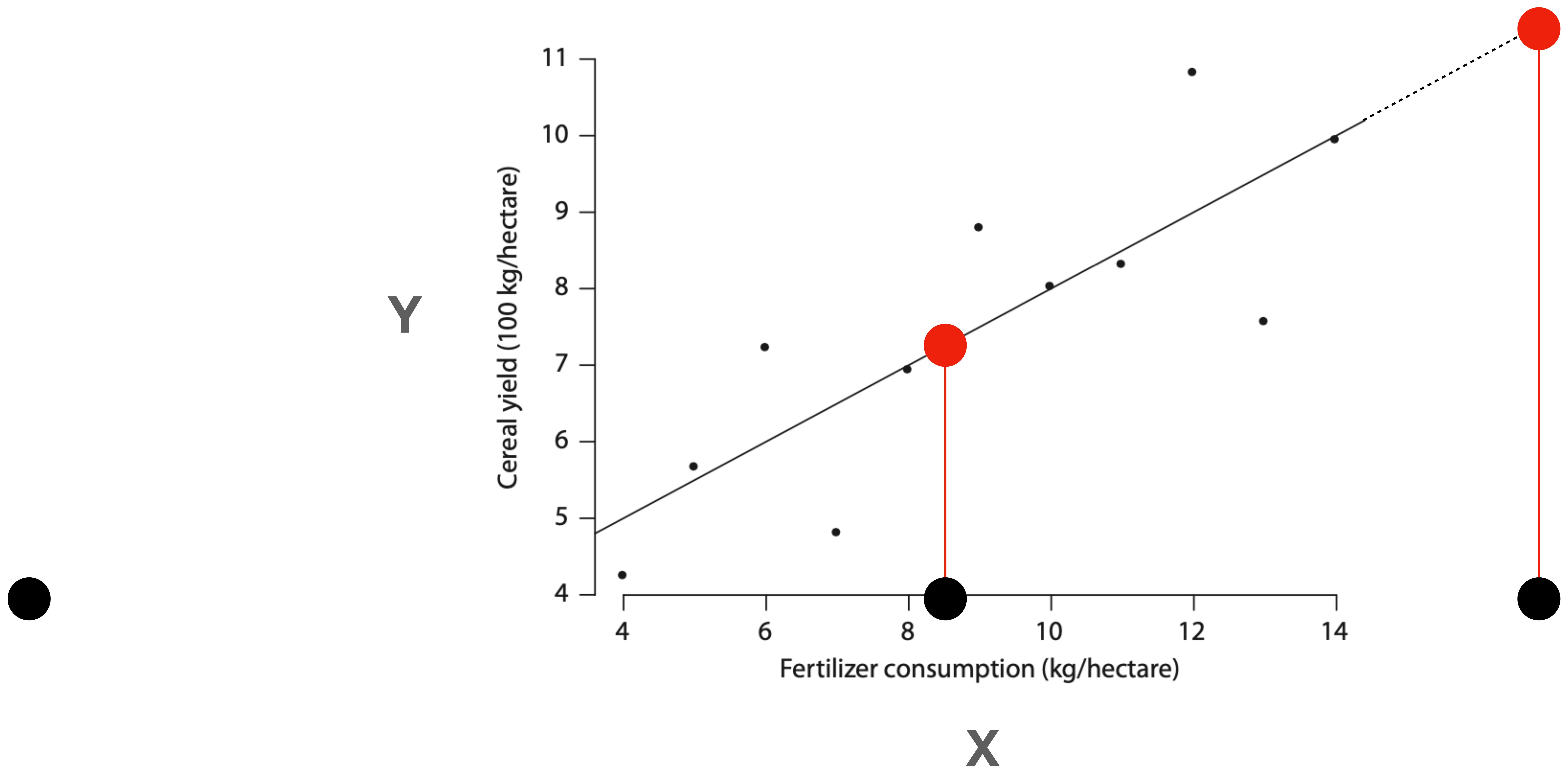
Edge 2019

$E[Y|X=x]$ is the best linear predictor for $X=x$



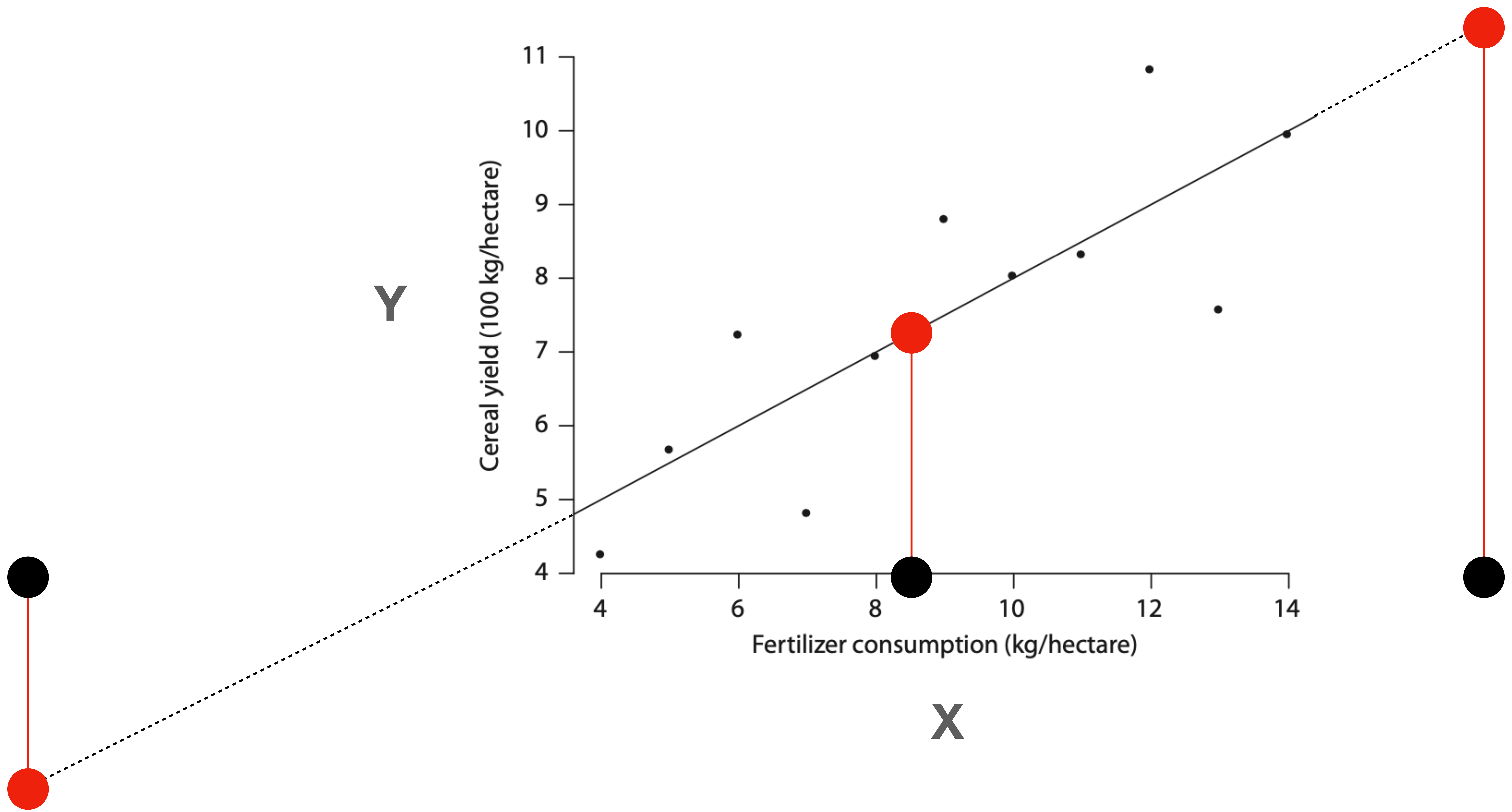
Edge 2019

E[Y|X=x] is the best linear predictor for X=x



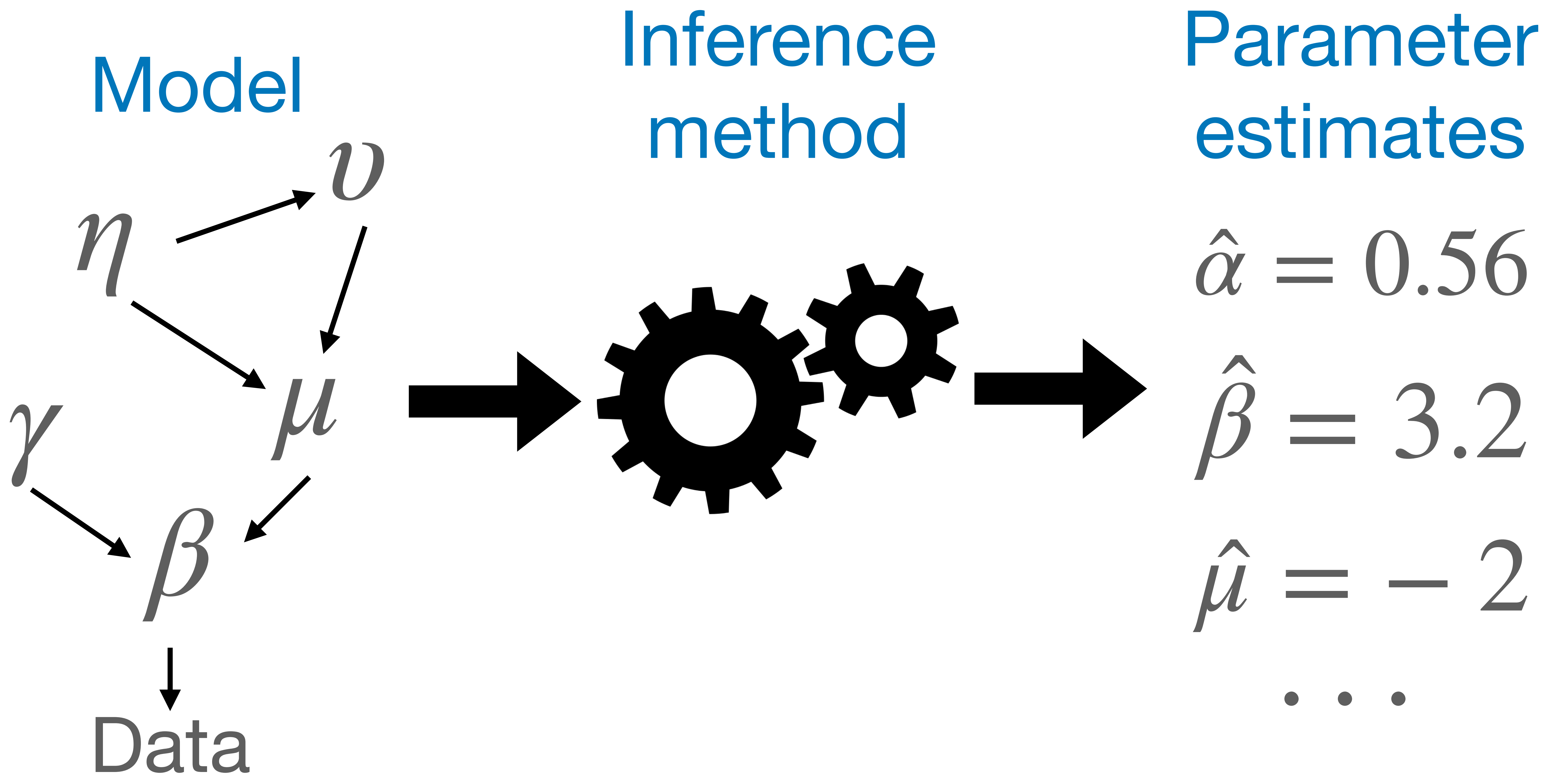
Edge 2019

E[Y|X=x] is the best linear predictor for X=x

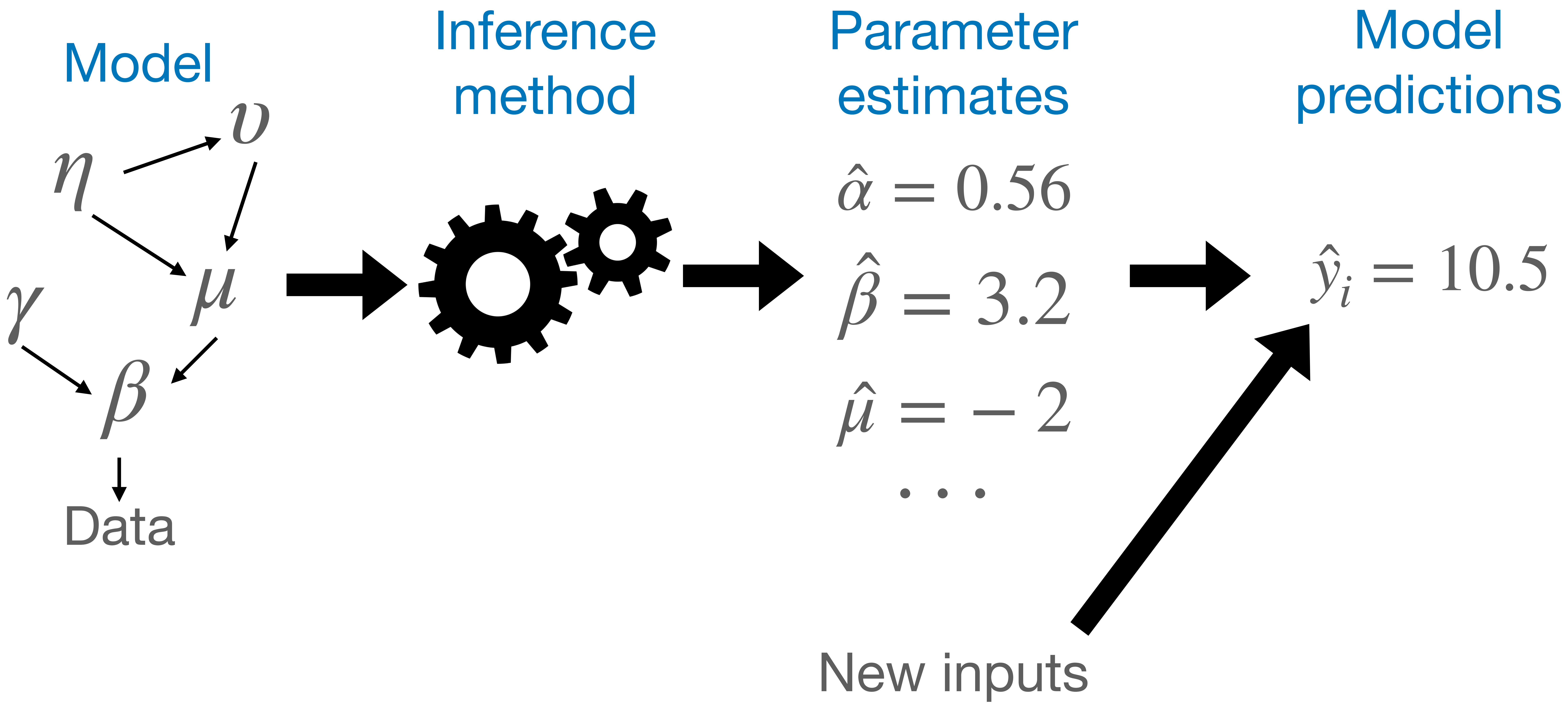


Edge 2019

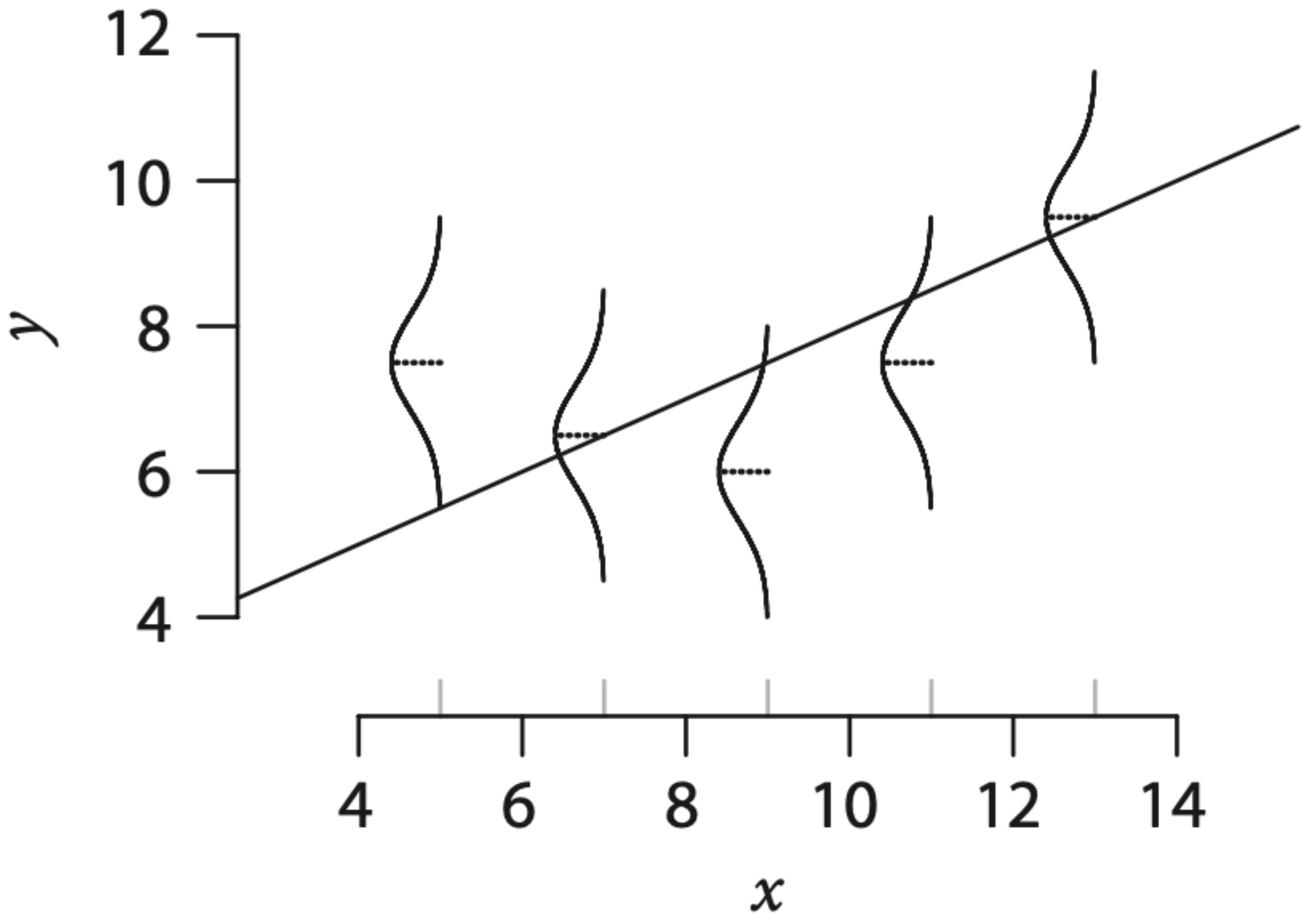
Inference vs. Prediction



Inference vs. Prediction



Linearity violated



“Homoscedasticity” assumption

$$Var[\epsilon \,|\, X = x] = \sigma_\epsilon^2$$

“The variance of the error is a constant, regardless of the value of X”

"Homoscedasticity" assumption

$$Var[\epsilon \,|\, X = x] = \sigma_\epsilon^2$$

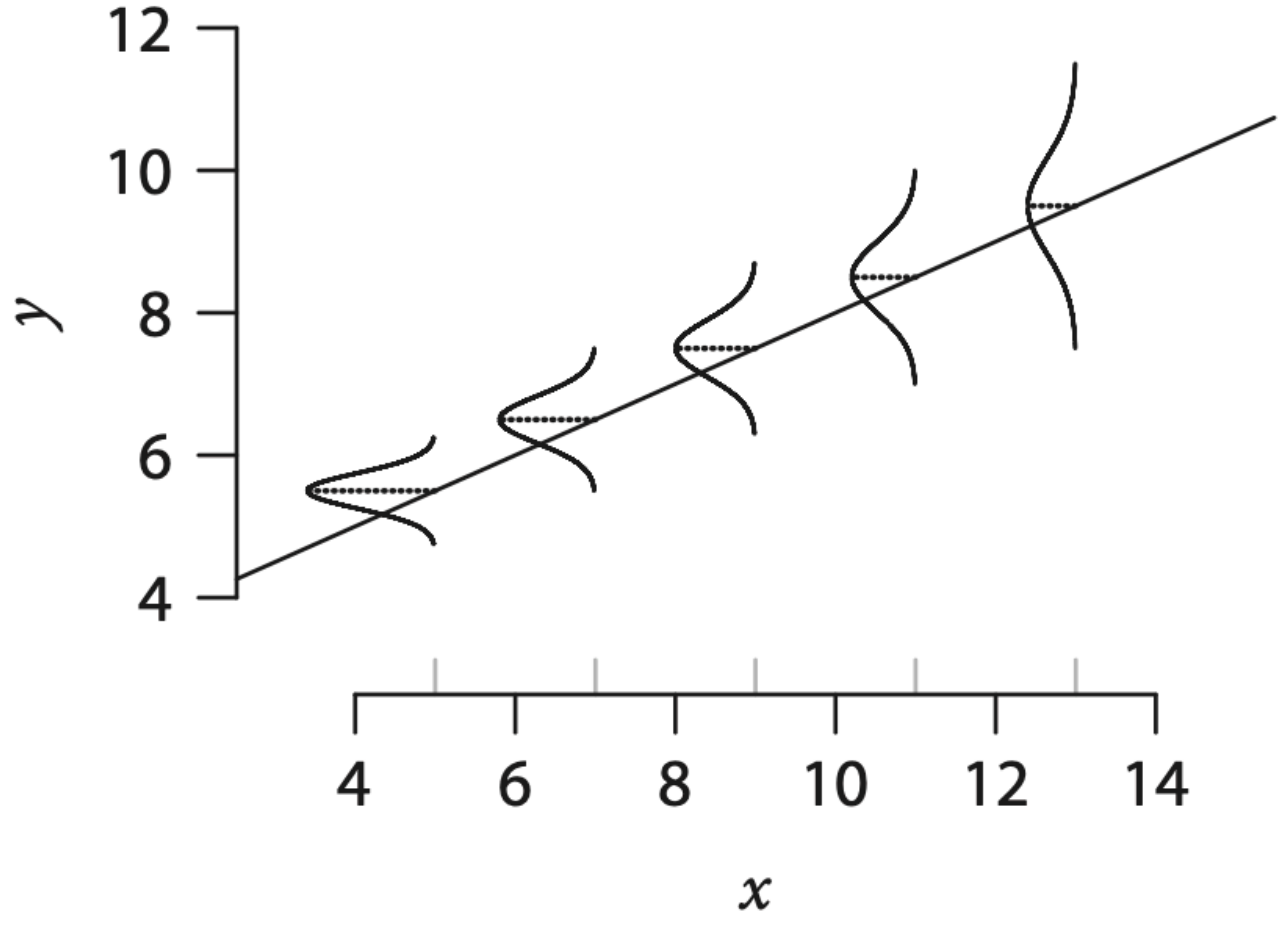
"The variance of the error is a constant, regardless of the value of X"

Then:

$$Var[Y \,|\, X = x] = \sigma_\epsilon^2$$

"The variance of Y is a constant (equal to the variance of the error), regardless of the value of X"

Homoscedasticity violated



“Normality” assumption

$$\epsilon_i \,|\, (X = x) \sim Normal(0, \sigma_\epsilon^2)$$

and all ϵ_i are independent of each other

“The error is normally distributed, regardless of the value of X”

“Normality” assumption

$$\epsilon_i \,|\, (X = x) \sim Normal(0, \sigma_\epsilon^2)$$

and all ϵ_i are independent of each other

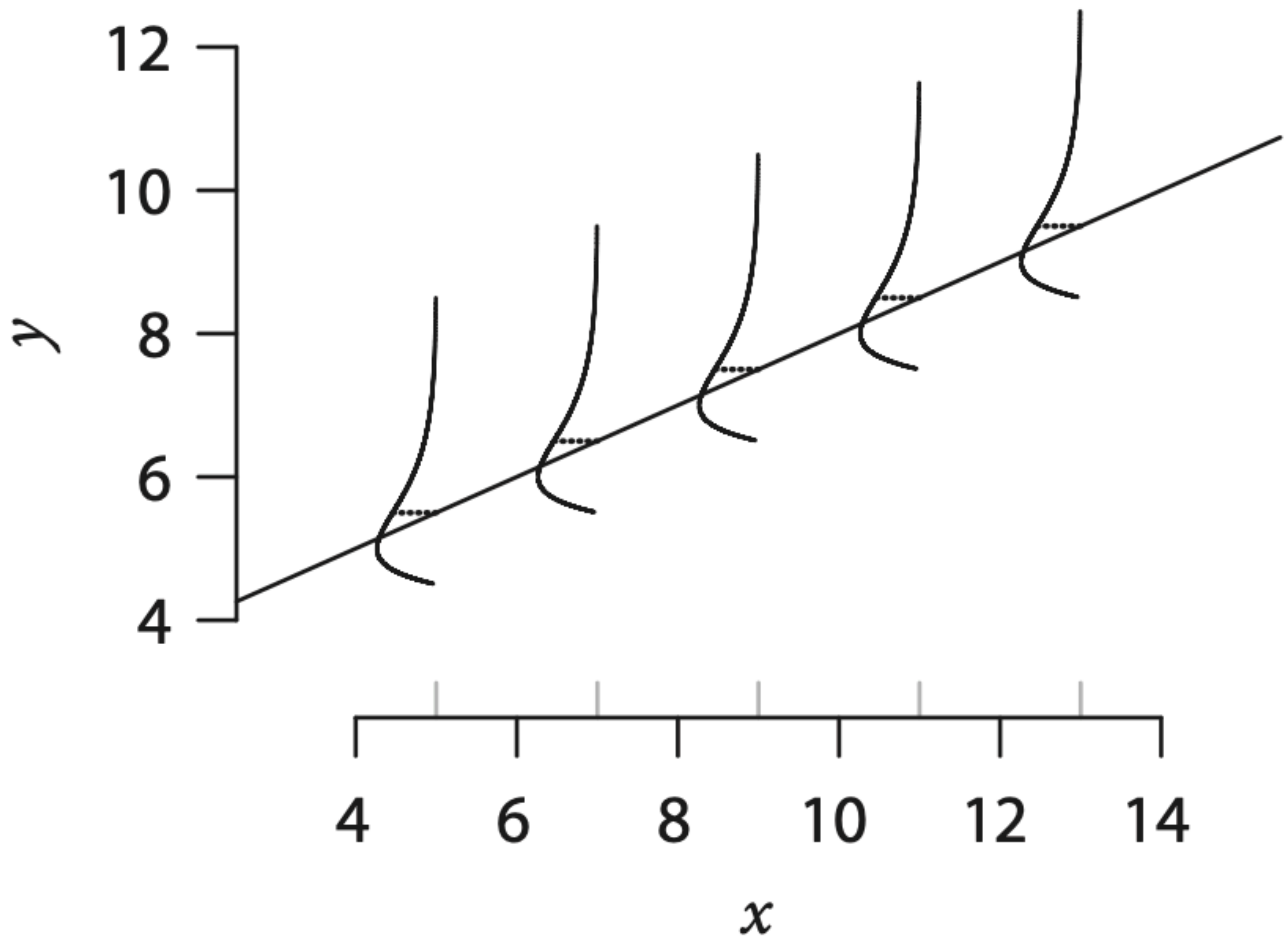
“The error is normally distributed, regardless of the value of X”

Then:

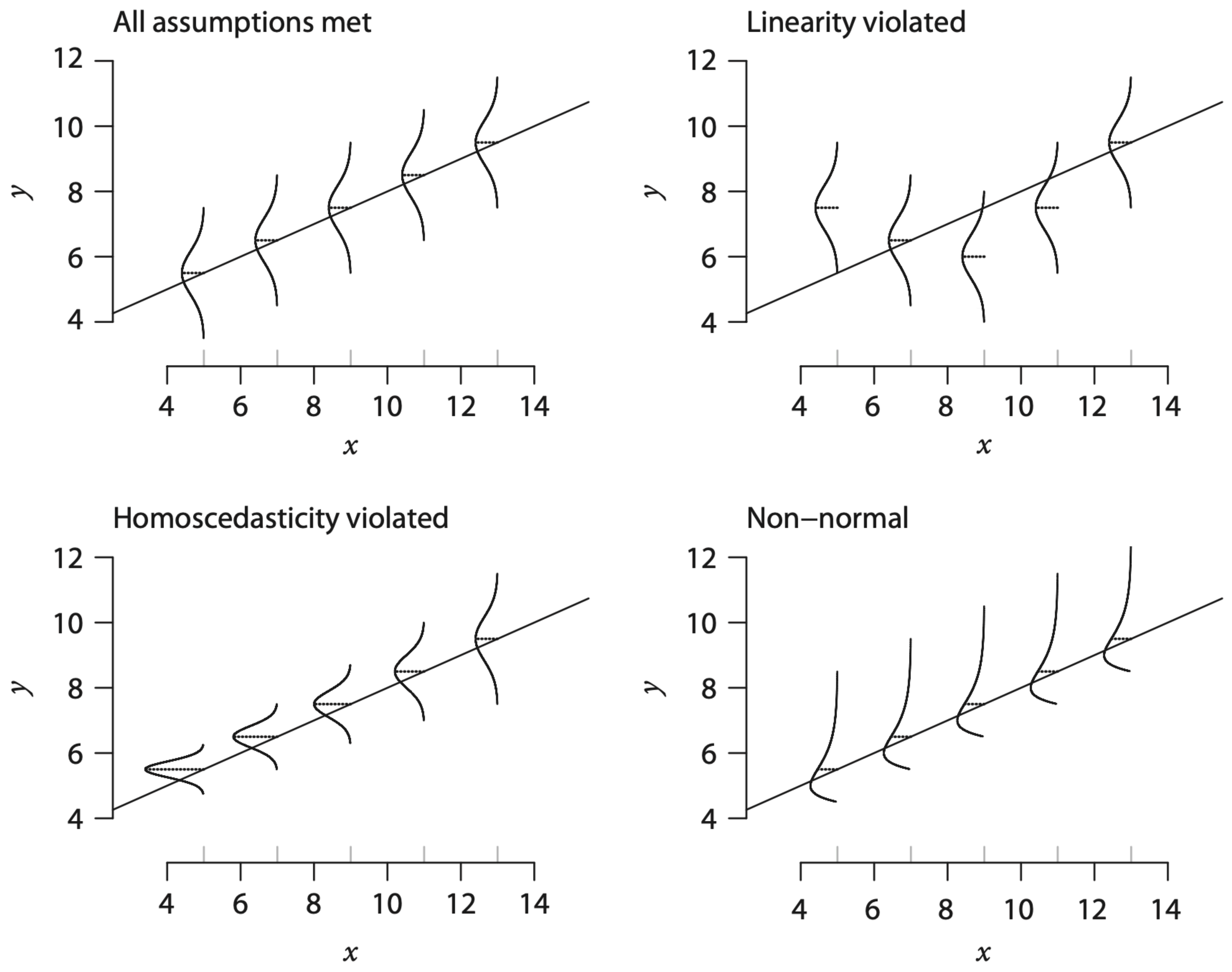
$$Y \,|\, (X = x) \sim Normal(\beta_0 + \beta_1 x, \sigma_\epsilon^2)$$

“Y is normally distributed, with mean equal to a linear function of the value of X”

Normality violated



Violations of assumptions



Edge 2019

That’s it for simple linear regression today, but:

A zoo of models…

Simple linear regression

- **1** predictor variable (x)
- **1** response variable (y)

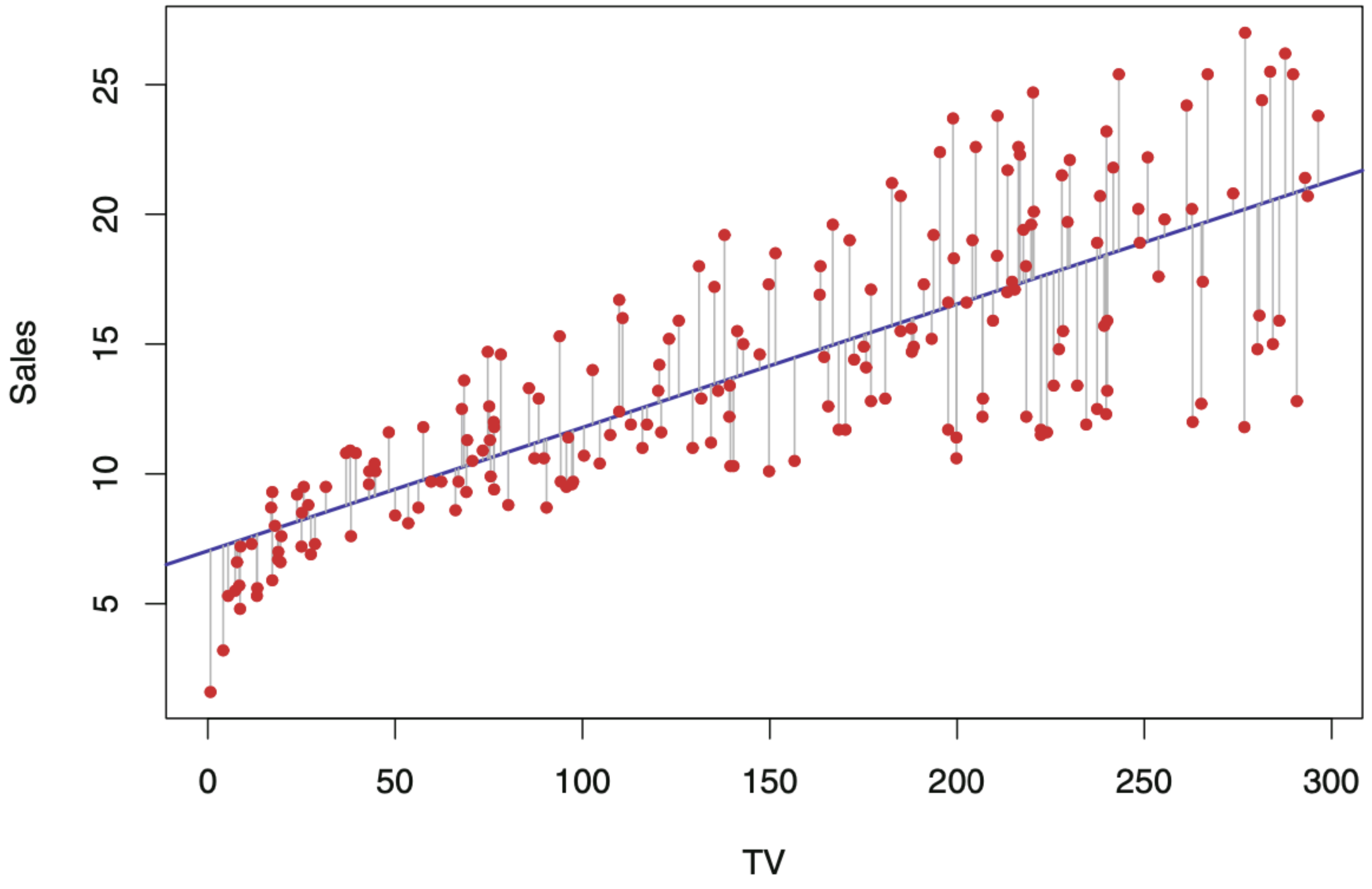
$$\mathbf{y} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{x} + \epsilon$$

Variable **y** is a linear function of **x**, plus some **noise**

James et al. 2014

Simple linear regression

- **1** predictor variable (x)
- **1** response variable (y)



James et al. 2014

Multiple linear regression

- **Several** predictor variables
- **1** response variable

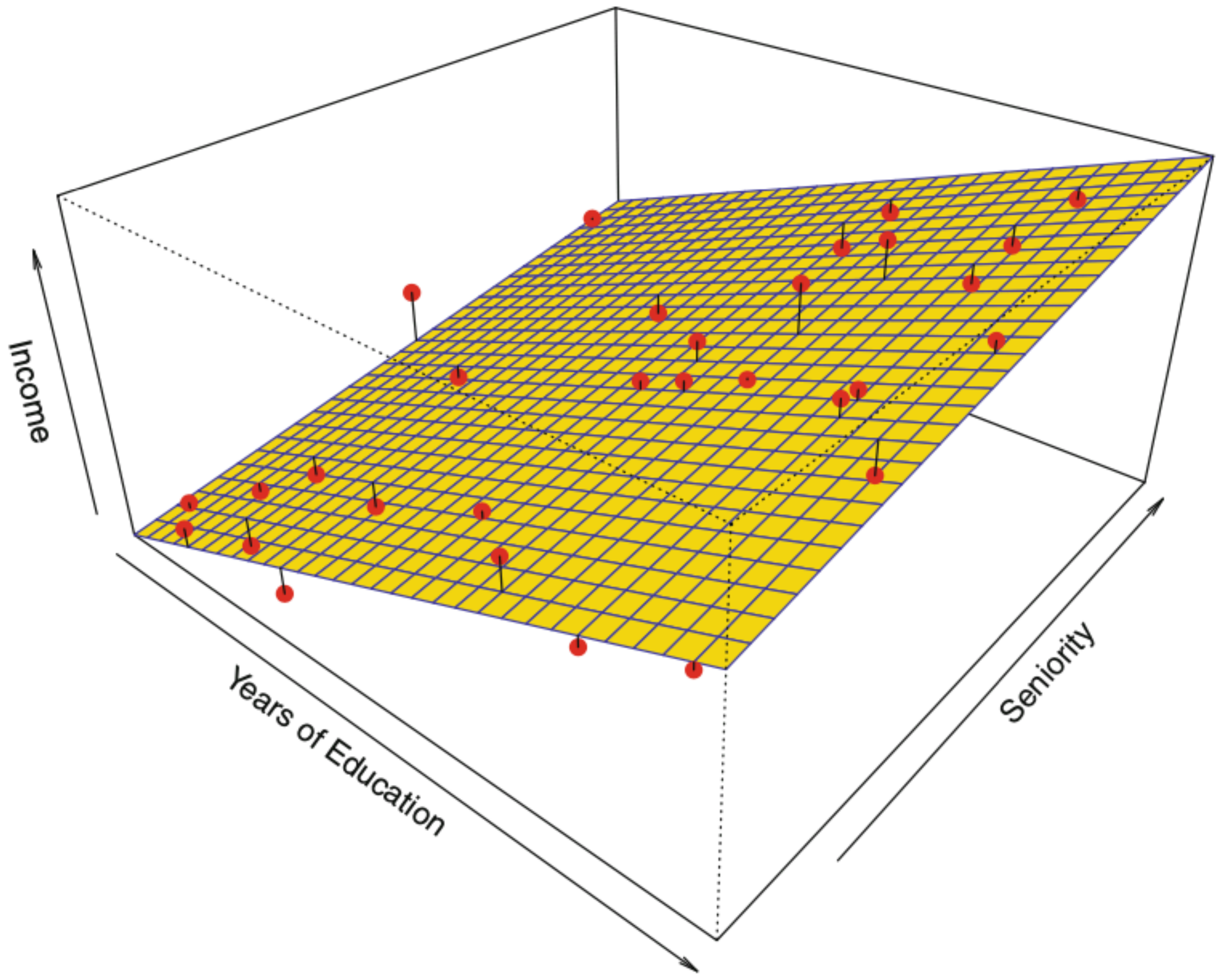
$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \ldots + \epsilon$$

Variable **y** is a linear function of **x_1, x_2, x_3, etc.** plus some **noise**

James et al. 2014

Multiple linear regression

- **Several** predictor variables
- **1** response variable



James et al. 2014

Expanding the model: interaction terms

$$\mathbf{y} = \beta_{\mathbf{0}} + \beta_{\mathbf{1}}\mathbf{x}_{\mathbf{1}} + \beta_{\mathbf{2}}\mathbf{x}_{\mathbf{2}} + \beta_{\mathbf{3}}\mathbf{x}_{\mathbf{1}}\mathbf{x}_{\mathbf{2}} + \ldots + \epsilon$$

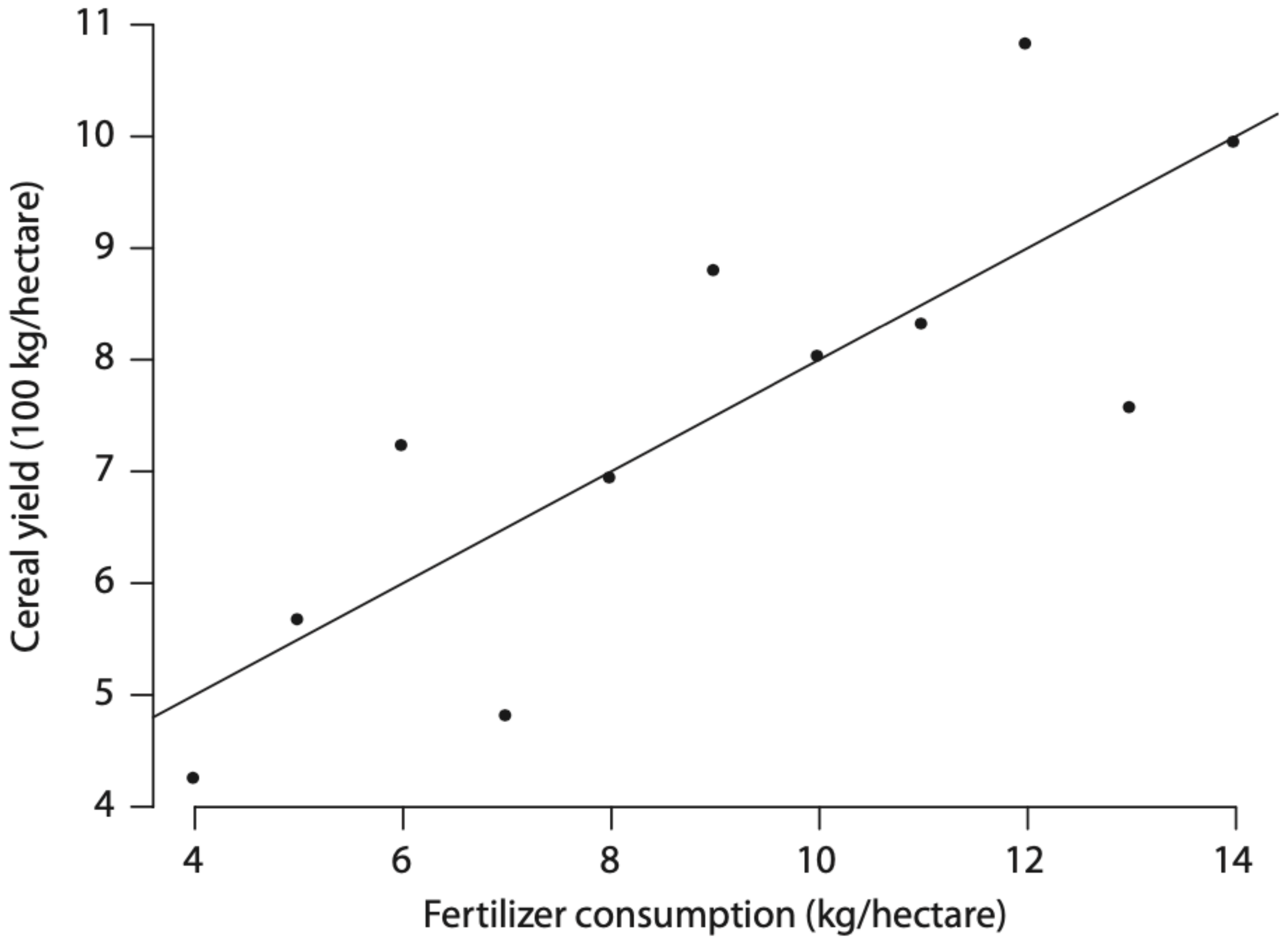
Expanding the model: interaction terms

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_1 \mathbf{x}_2 + \ldots + \epsilon$$

will be large when the combined behavior of x_1 and x_2 jointly influence the value of y

Multiple linear regression with categorical predictors (ANOVA)

Y

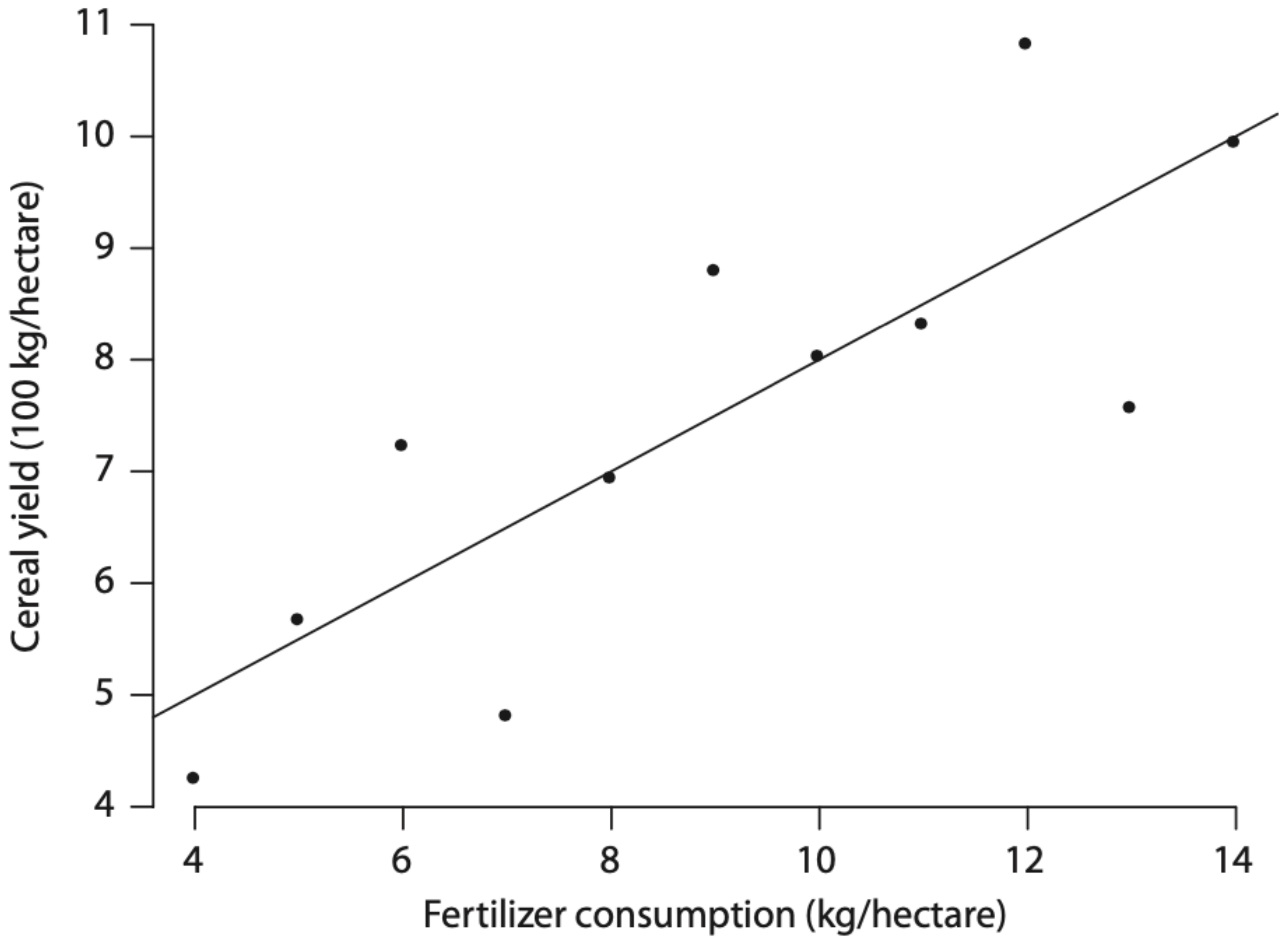


X

So far, we've assumed that the value of **X** is a continuous number

Multiple linear regression with categorical predictors (ANOVA)

Y

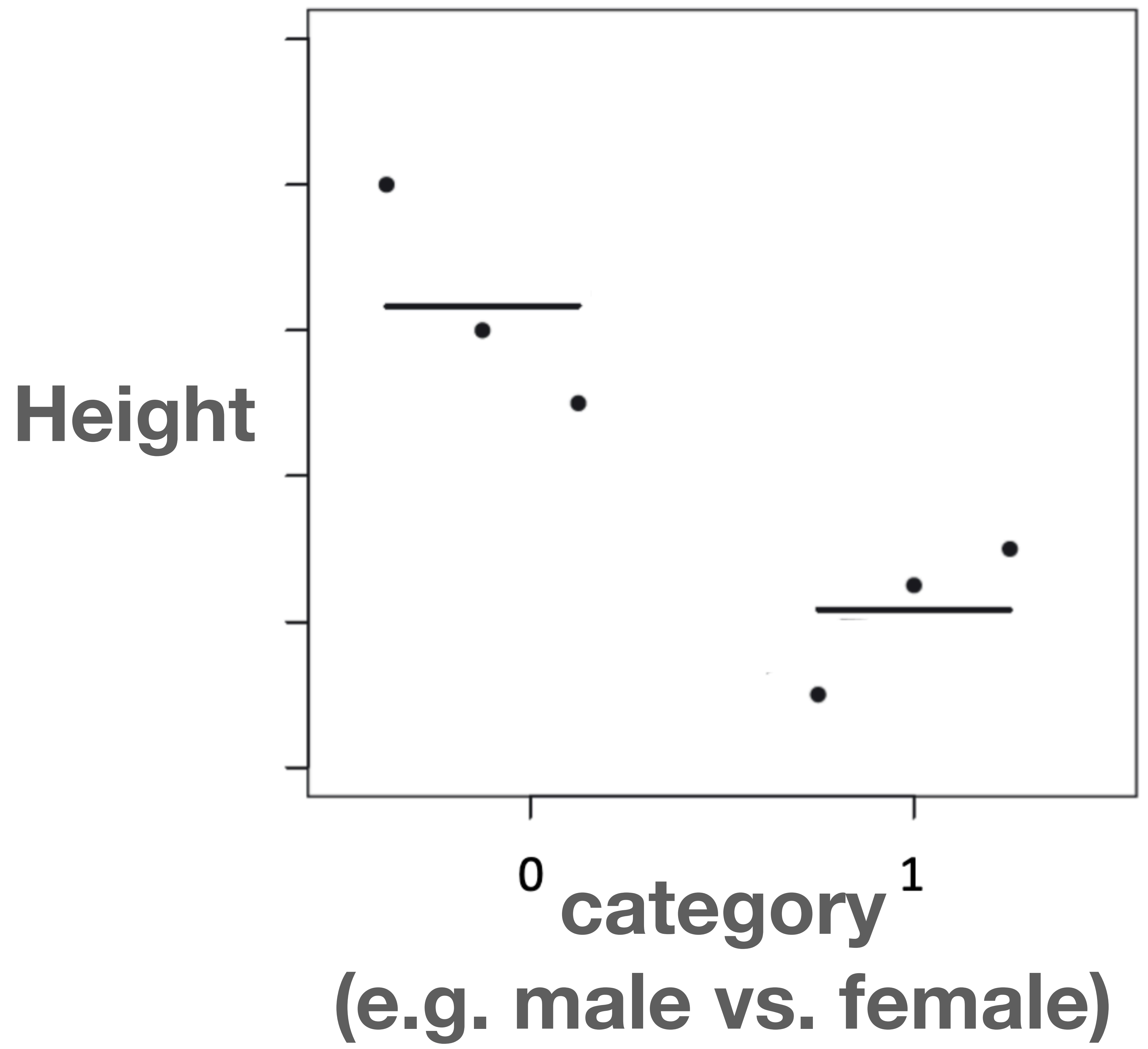


X

What if it was a **category** instead?

So far, we've assumed that the value of **X** is a continuous number

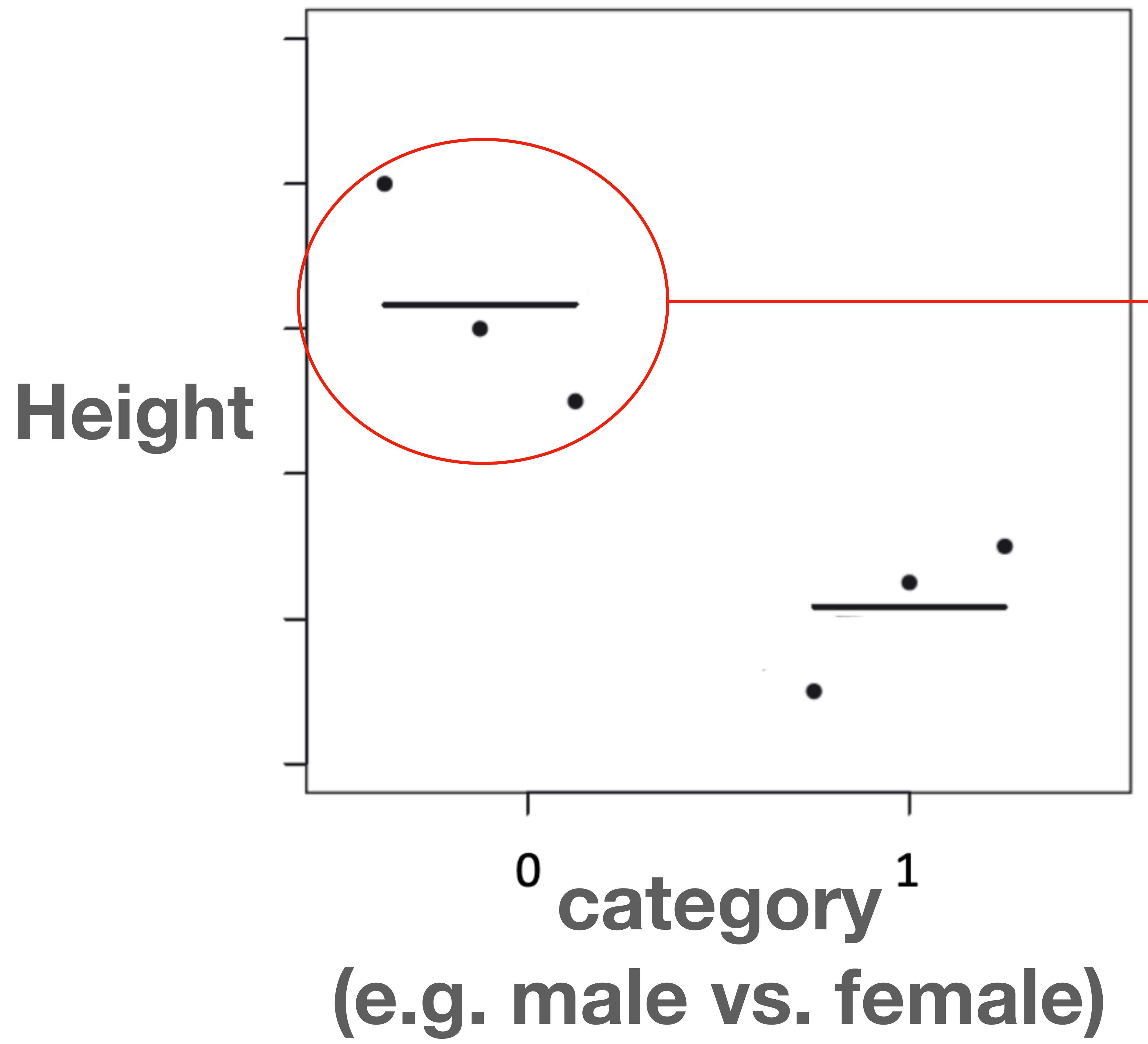
Multiple linear regression with categorical predictors (ANOVA)



$$\mathbf{y} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{x} + \epsilon$$

We can modify our original linear model such that **x** is now a **dummy variable.**

Multiple linear regression with categorical predictors (ANOVA)

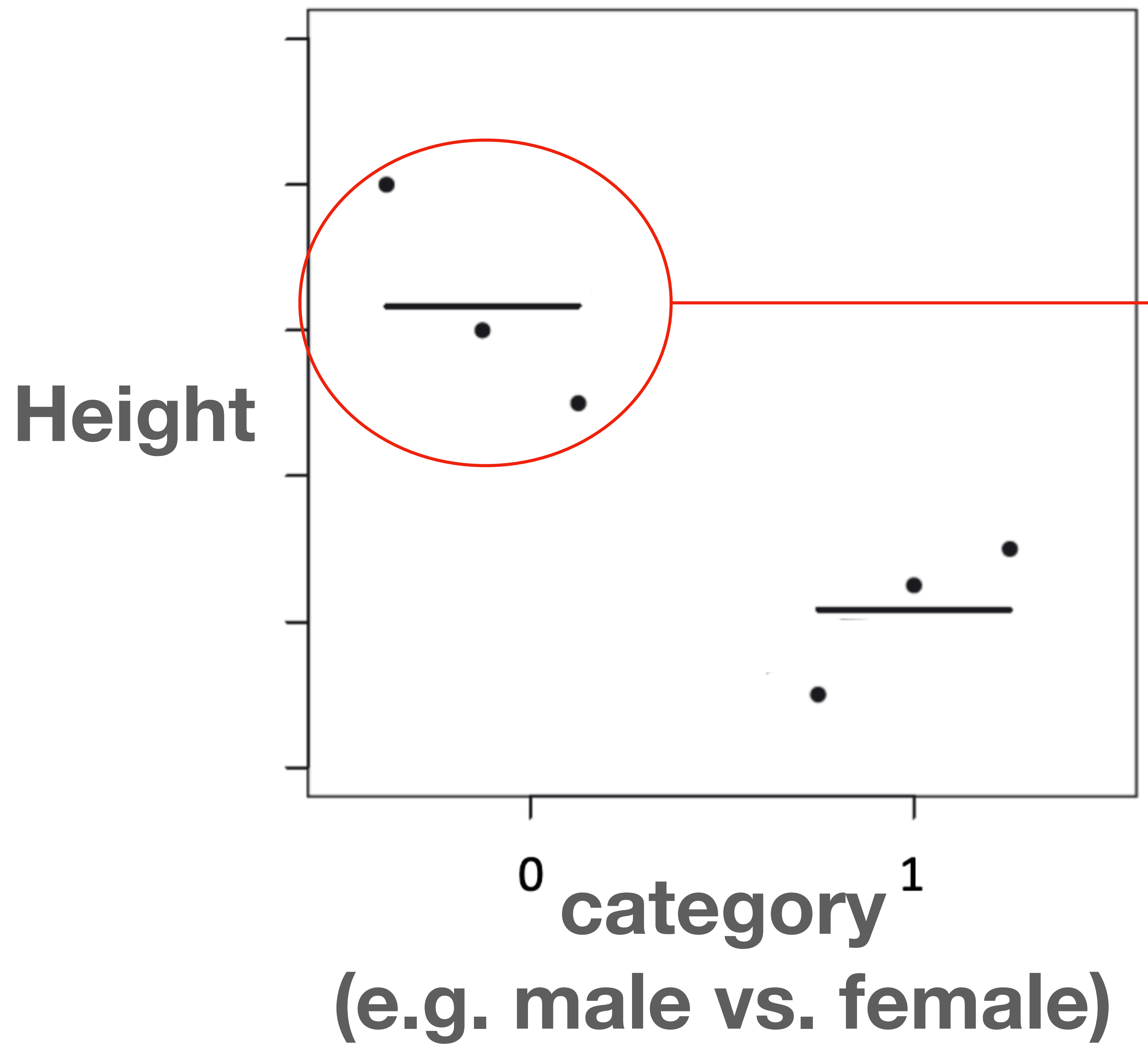


$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \epsilon$$

We can modify our original linear model such that **x** is now a **dummy variable.**

These observations get assigned x = 0

Multiple linear regression with categorical predictors (ANOVA)



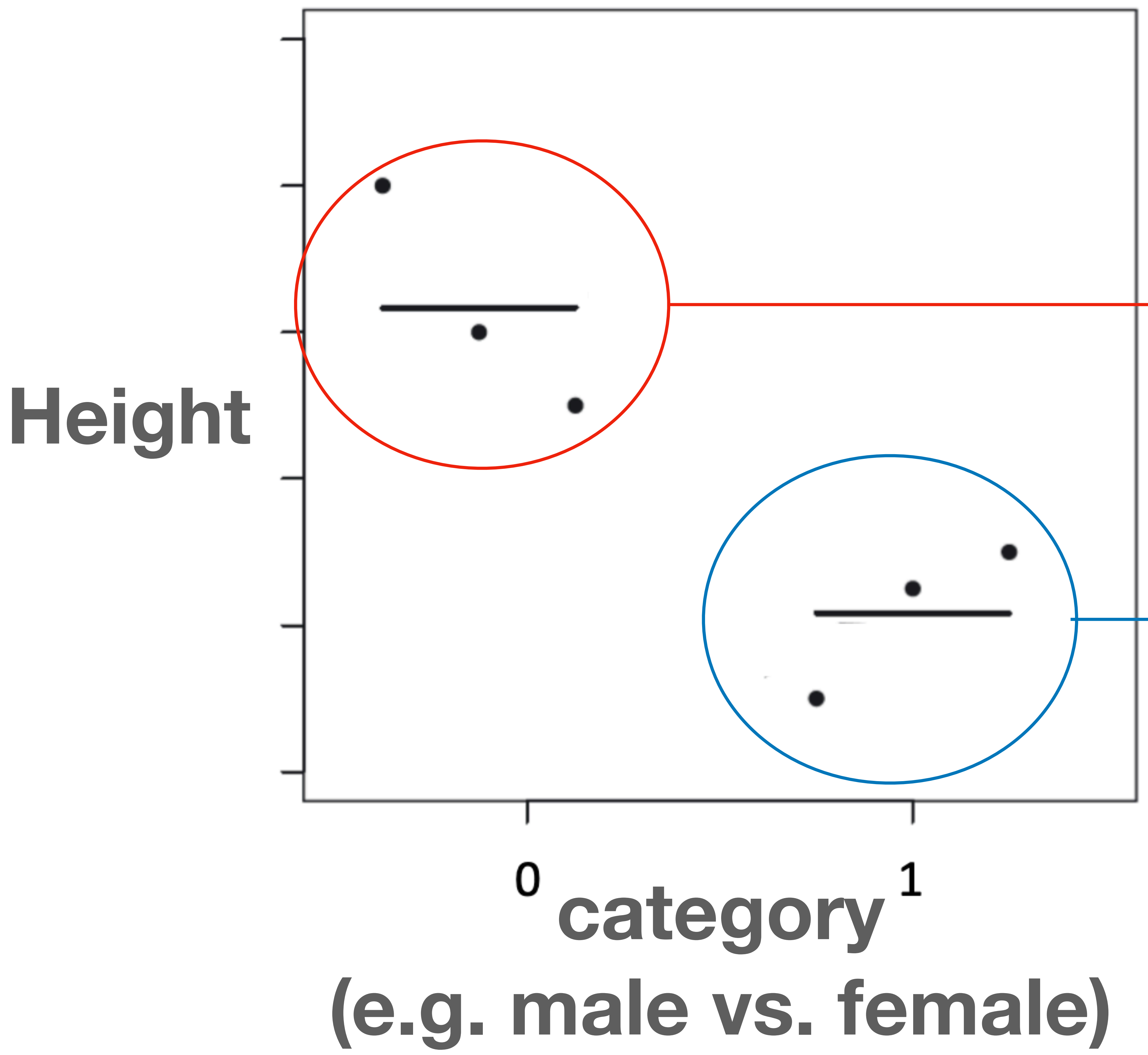
$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \epsilon$$

We can modify our original linear model such that **x** is now a **dummy variable.**

These observations get assigned x = 0

$$\mathbf{y} = \beta_0 + \epsilon$$

Multiple linear regression with categorical predictors (ANOVA)



$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \epsilon$$

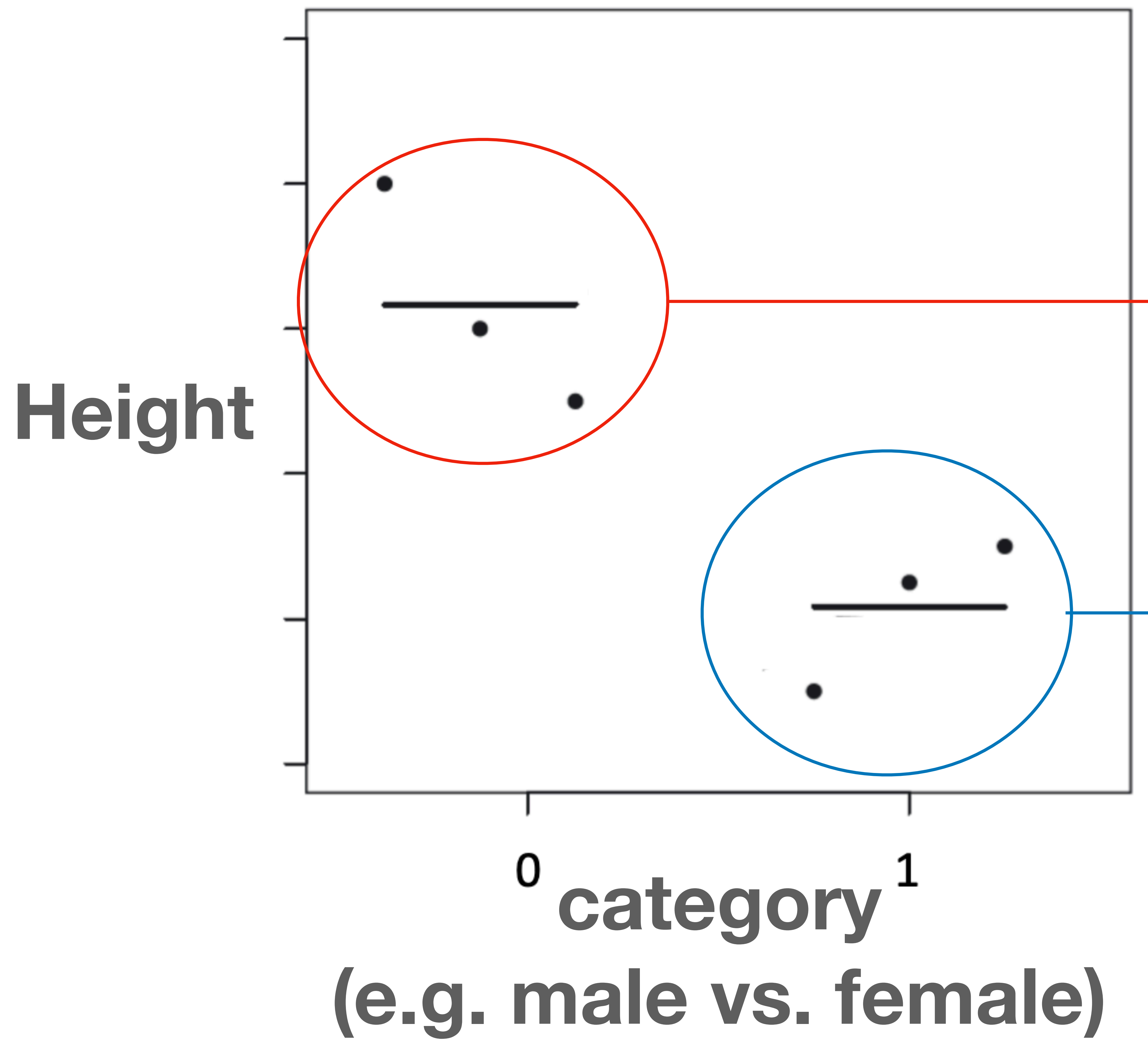
We can modify our original linear model such that **x** is now a **dummy variable.**

These observations get assigned x = 0

$$\mathbf{y} = \beta_0 + \epsilon$$

These observations get assigned x = 1

Multiple linear regression with categorical predictors (ANOVA)



$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \epsilon$$

We can modify our original linear model such that **x** is now a **dummy variable.**

These observations get assigned x = 0

$$\mathbf{y} = \beta_0 + \epsilon$$

These observations get assigned x = 1

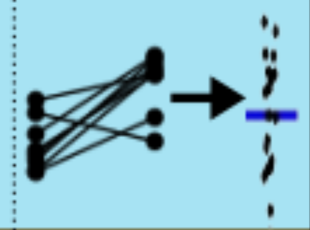
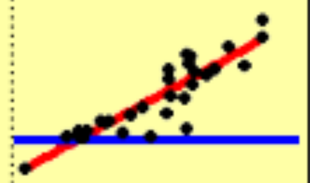
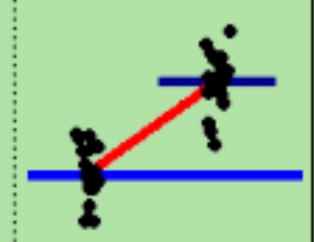
$$\mathbf{y} = \beta_0 + \beta_1 + \epsilon$$

A little secret…

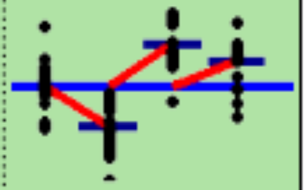
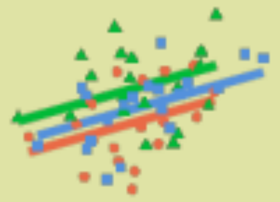
Common statistical tests are linear models

Last updated: 02 April, 2019

See worked examples and more details at the accompanying notebook: https://lindeloev.github.io/tests-as-linear

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: lm(y ~ 1 + x)	**y is independent of x** P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	lm(y ~ 1) lm(signed_rank(y) ~ 1)	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y_1, y_2, paired=TRUE) wilcox.test(y_1, y_2, paired=TRUE)	lm(y_2 - y_1 ~ 1) lm(signed_rank(y_2 - y_1) ~ 1)	✓ for N >14	One intercept predicts the pairwise **y_2-y_1** differences. - (Same, but it predicts the *signed rank* of **y_2-y_1**.)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y_1, y_2, var.equal=TRUE) t.test(y_1, y_2, var.equal=FALSE) wilcox.test(y_1, y_2)	lm(y ~ 1 + G_2)^A gls(y ~ 1 + G_2, weights=…^{B})^A lm(signed_rank(y) ~ 1 + G_2)^A	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	

A little secret…

Multiple regression: lm(y ~ 1 + x_1 + x_2 +...)	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	lm(y ~ 1 + G_2 + G_3 +...+ G_N)^A lm(rank(y) ~ 1 + G_2 + G_3 +...+ G_N)^A	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	
	P: One-way ANCOVA	aov(y ~ group + x)	lm(y ~ 1 + G_2 + G_3 +...+ G_N + x)^A	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	
	P: Two-way ANOVA	aov(y ~ group * sex)	lm(y ~ 1 + G_2 + G_3 + ... + G_N + S_2 + S_3+ ... + S_K + G_2*S_2+G_3*S_3+...+G_N*S_K)	✓	Interaction term: changing **sex** changes the **y ~ group** parameters. *Note: $G_{2\ to\ N}$ is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for $S_{2\ to\ K}$ for sex. The first line (with G_i) is main effect of group, the second (with S_i) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.*	*[Coming]*
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	**Equivalent log-linear model** glm(y ~ 1 + G_2 + G_3 + ... + G_N + S_2 + S_3+ ... + S_K + G_2*S_2+G_3*S_3+...+G_N*S_K, family=...)^A	✓	Interaction term: (Same as Two-way ANOVA.) *Note: Run glm using the following arguments:* `glm(model, family=poisson())` *As linear-model, the Chi-square test is $\log(y_i) = \log(N) + \log(\alpha_i) + \log(\beta_j) + \log(\alpha_i\beta_j)$ where α_i and β_j are proportions. See more info in the accompanying notebook.*	*Same as Two-way ANOVA*
	N: Goodness of fit	chisq.test(y)	glm(y ~ 1 + G_2 + G_3 +...+ G_N, family=...)^A	✓	(Same as One-way ANOVA and see Chi-Square note.)	*1W-ANOVA*

Visit https://lindeloev.github.io/tests-as-linear/ for more information

General multivariate regression

- **Several** predictor variables
- **Several** response variables

General multivariate regression

- **Several** predictor variables
- **Several** response variables

$$\mathbf{y}_1 = \beta_0 + \beta_{1,1}\mathbf{x}_1 + \beta_{2,1}\mathbf{x}_2 + \beta_{3,1}\mathbf{x}_3 + \ldots + \epsilon_1$$

Variable $\mathbf{y_1}$ is a linear function of $\mathbf{x_1}$**,** $\mathbf{x_2}$**,** $\mathbf{x_3}$**, etc.** plus some **noise**

General multivariate regression

- **Several** predictor variables
- **Several** response variables

$$\mathbf{y}_1 = \beta_0 + \beta_{1,1}\mathbf{x}_1 + \beta_{2,1}\mathbf{x}_2 + \beta_{3,1}\mathbf{x}_3 + \ldots + \epsilon_1$$

Variable $\mathbf{y_1}$ is a linear function of $\mathbf{x_1}$**,** $\mathbf{x_2}$**,** $\mathbf{x_3}$**, etc.** plus some **noise**

$$\mathbf{y}_2 = \beta_0 + \beta_{1,2}\mathbf{x}_1 + \beta_{2,2}\mathbf{x}_2 + \beta_{3,2}\mathbf{x}_3 + \ldots + \epsilon_2$$

Variable $\mathbf{y_2}$ is a linear function of $\mathbf{x_1}$**,** $\mathbf{x_2}$**,** $\mathbf{x_3}$**, etc.** plus some **noise**

General multivariate regression

- **Several** predictor variables
- **Several** response variables

$$\mathrm{y}_1 = \beta_0 + \beta_{1,1}\mathrm{x}_1 + \beta_{2,1}\mathrm{x}_2 + \beta_{3,1}\mathrm{x}_3 + \ldots + \epsilon_1$$

Variable $\mathbf{y_1}$ is a linear function of $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$, **etc.** plus some **noise**

$$\mathrm{y}_2 = \beta_0 + \beta_{1,2}\mathrm{x}_1 + \beta_{2,2}\mathrm{x}_2 + \beta_{3,2}\mathrm{x}_3 + \ldots + \epsilon_2$$

Variable $\mathbf{y_2}$ is a linear function of $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$, **etc.** plus some **noise**

$$\mathrm{y}_3 = \beta_0 + \beta_{1,3}\mathrm{x}_1 + \beta_{2,3}\mathrm{x}_2 + \beta_{3,3}\mathrm{x}_3 + \ldots + \epsilon_3$$

Variable $\mathbf{y_3}$ is a linear function of $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$, **etc.** plus some **noise**

Generalized linear models

Generalized linear models

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \ldots + \epsilon$$

Generalized linear models

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \ldots + \epsilon$$

y is continuous between -inf and inf

Generalized linear models

$$\mathbf{y} = \beta_{\mathbf{0}} + \beta_{\mathbf{1}}\mathbf{x_1} + \beta_{\mathbf{2}}\mathbf{x_2} + \beta_{\mathbf{3}}\mathbf{x_3} + \ldots + \epsilon$$

y is continuous between -inf and inf

Categorical

Probabilities

Counts

Data

Compositions

etc.

Generalized linear models

$$\mathbf{y} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{x_1} + \beta_\mathbf{2}\mathbf{x_2} + \beta_\mathbf{3}\mathbf{x_3} + \ldots + \epsilon$$

y is continuous between -inf and inf

Link function f(y)

Data

Categorical

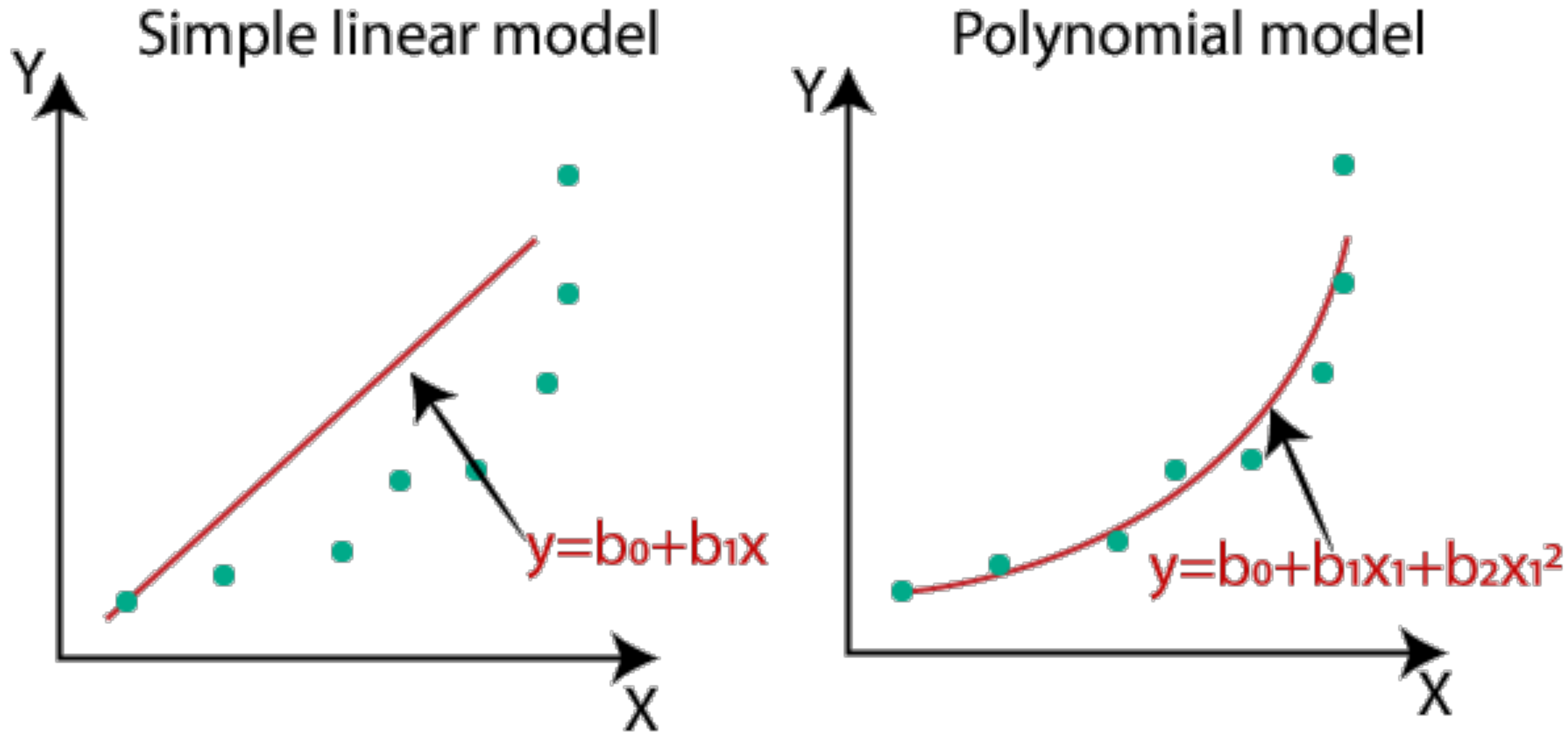
Probabilities

Counts

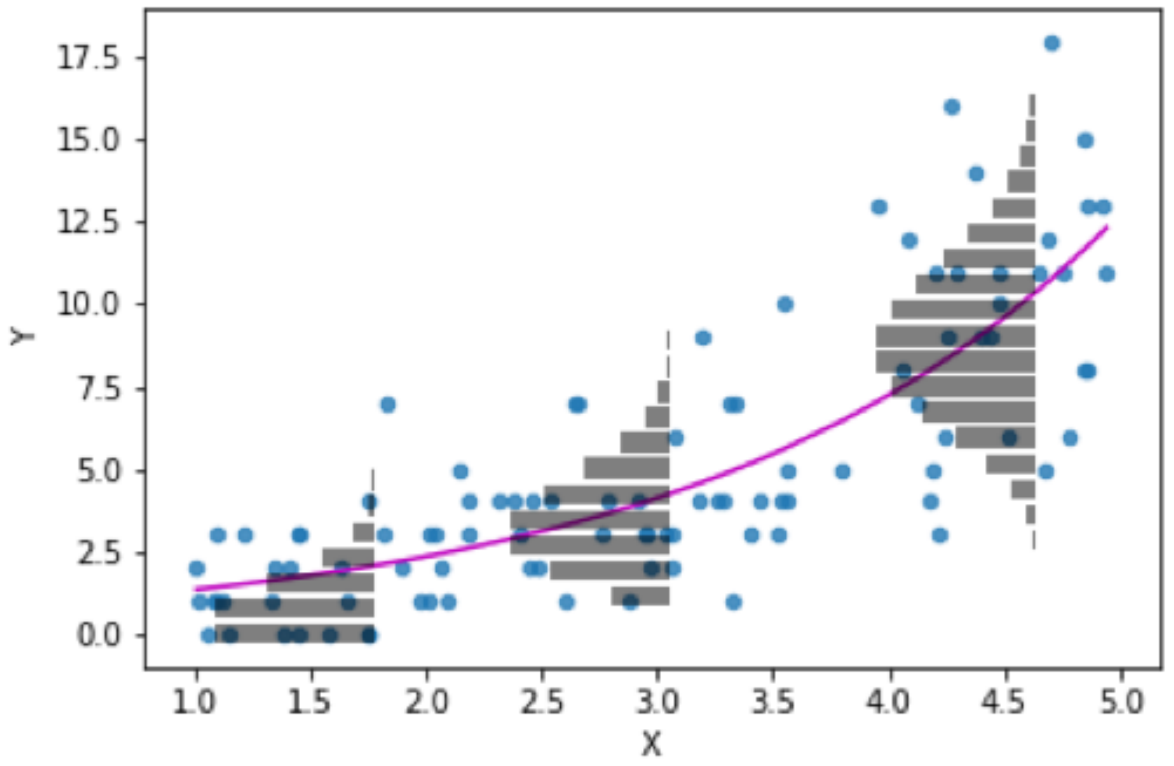
Compositions

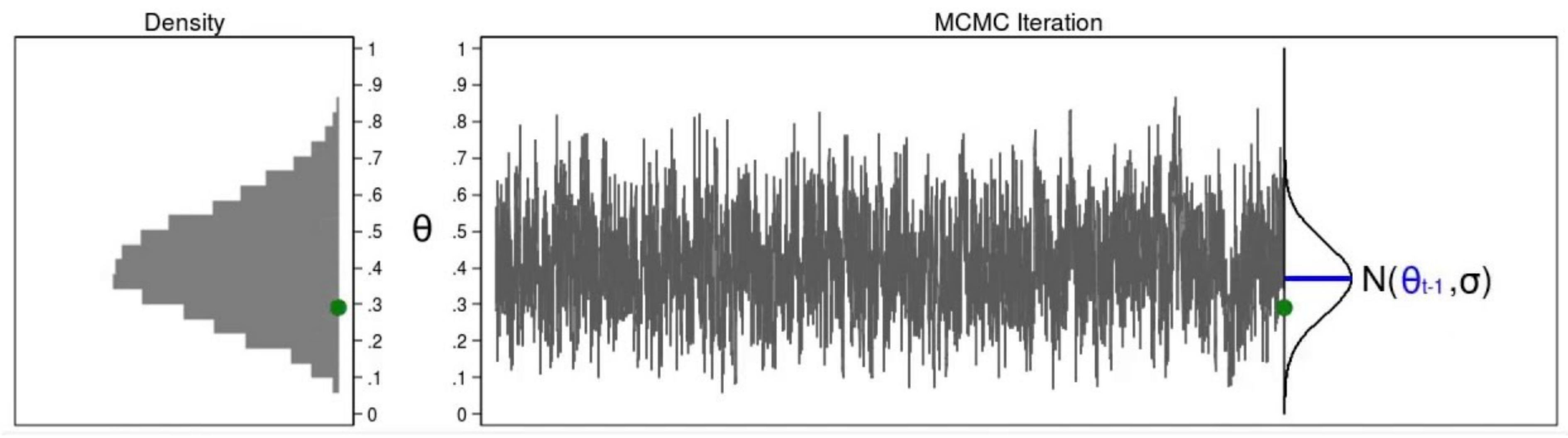
etc.

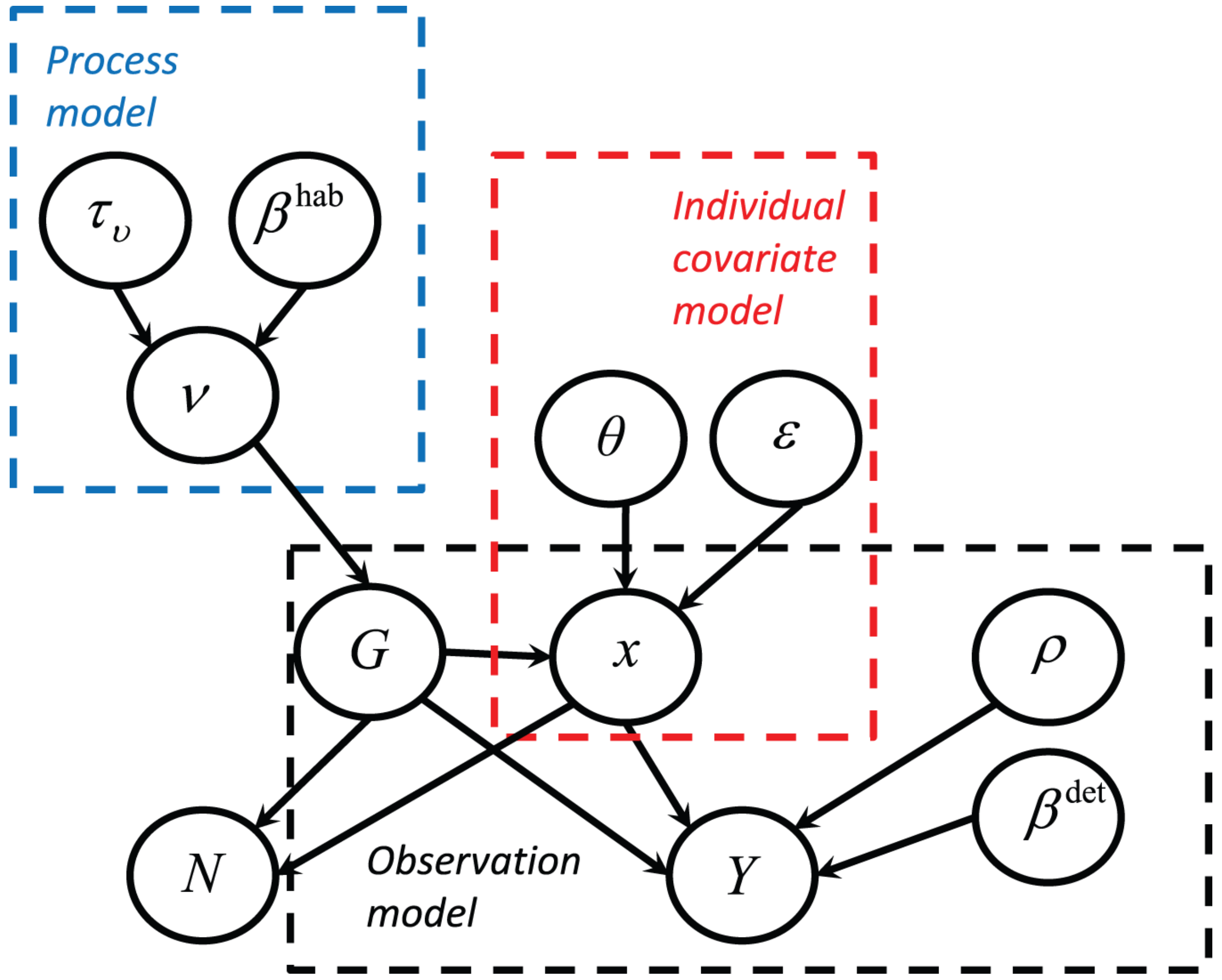
Non-linear models

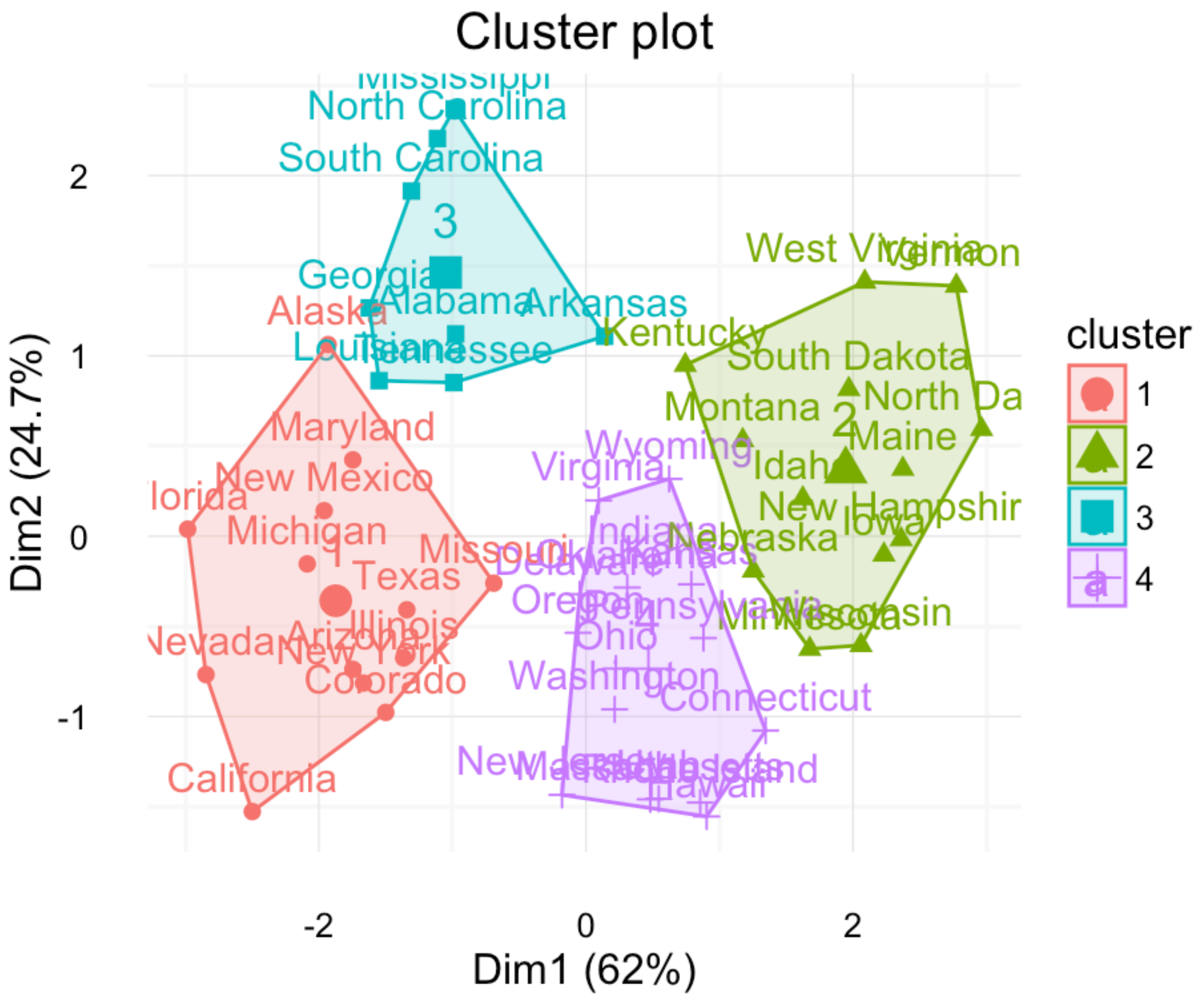


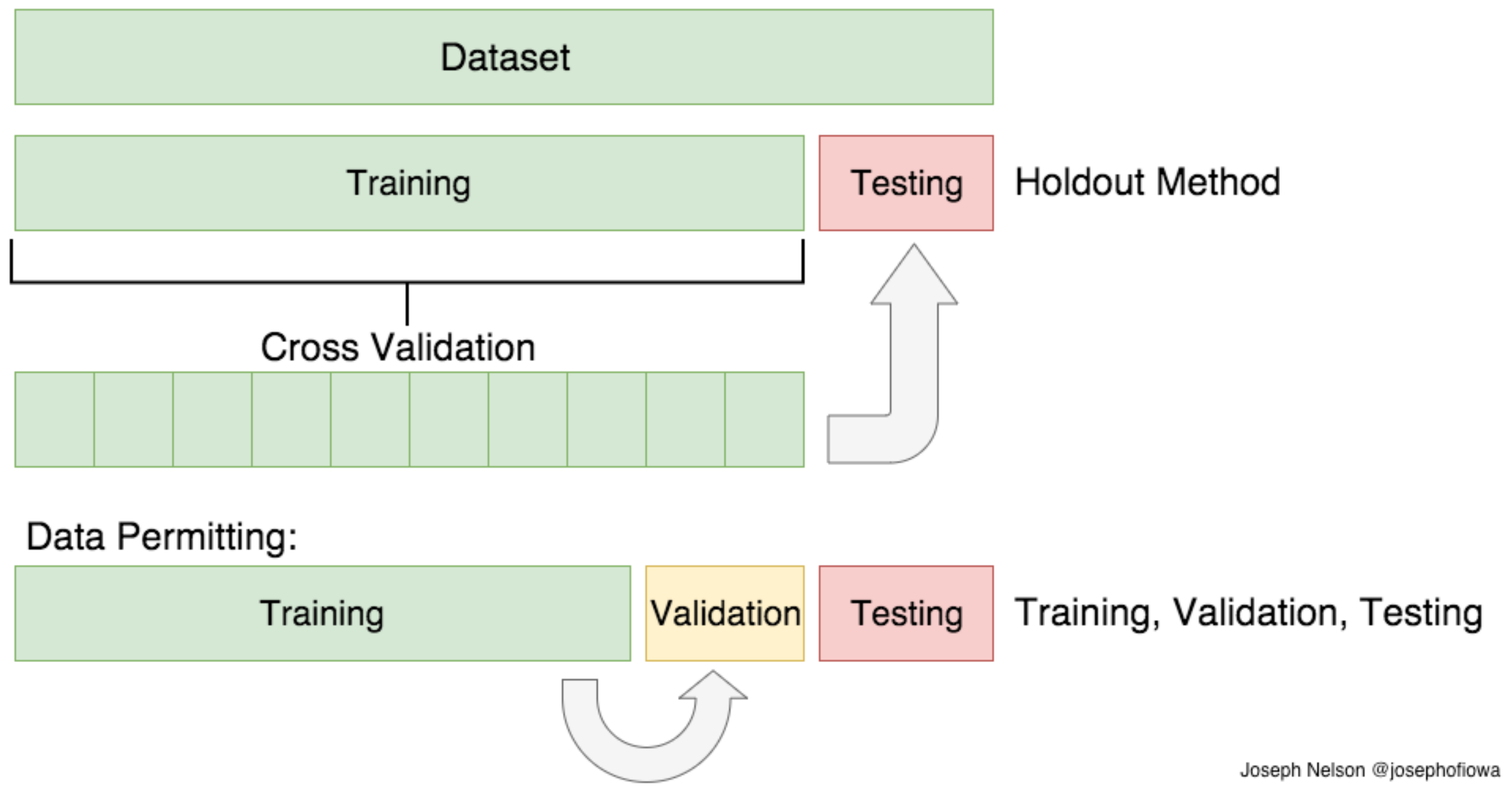
Advanced Topics in Data Analysis

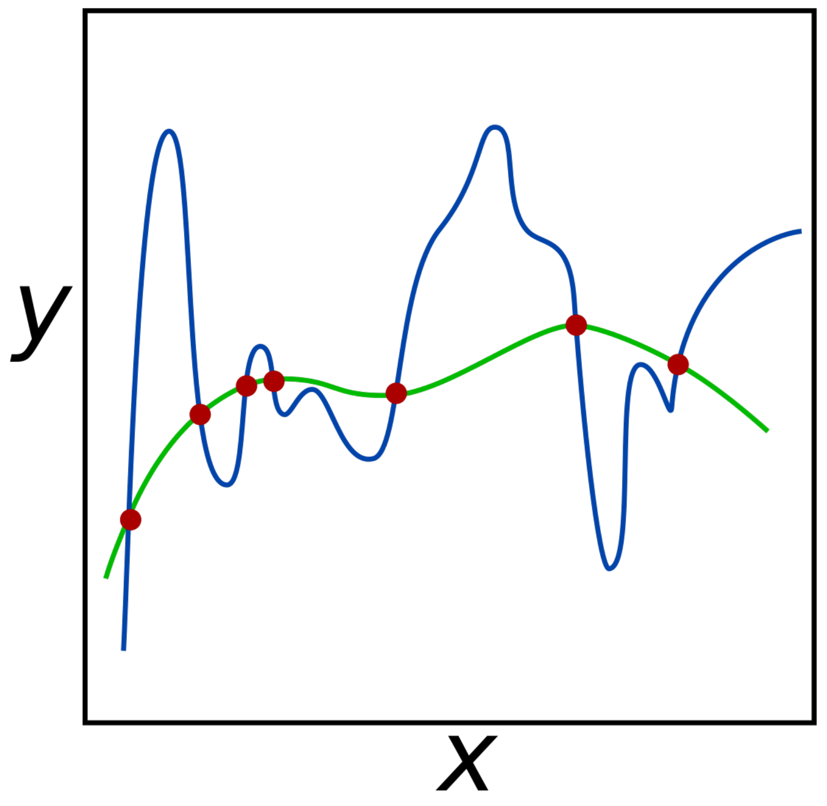












- Simulating a linear model in R
- Multivariate linear regression