On *p* vs *n*
(# predictors vs # examples)

ML 4 Health, Supplementary

Suppose we have the following data

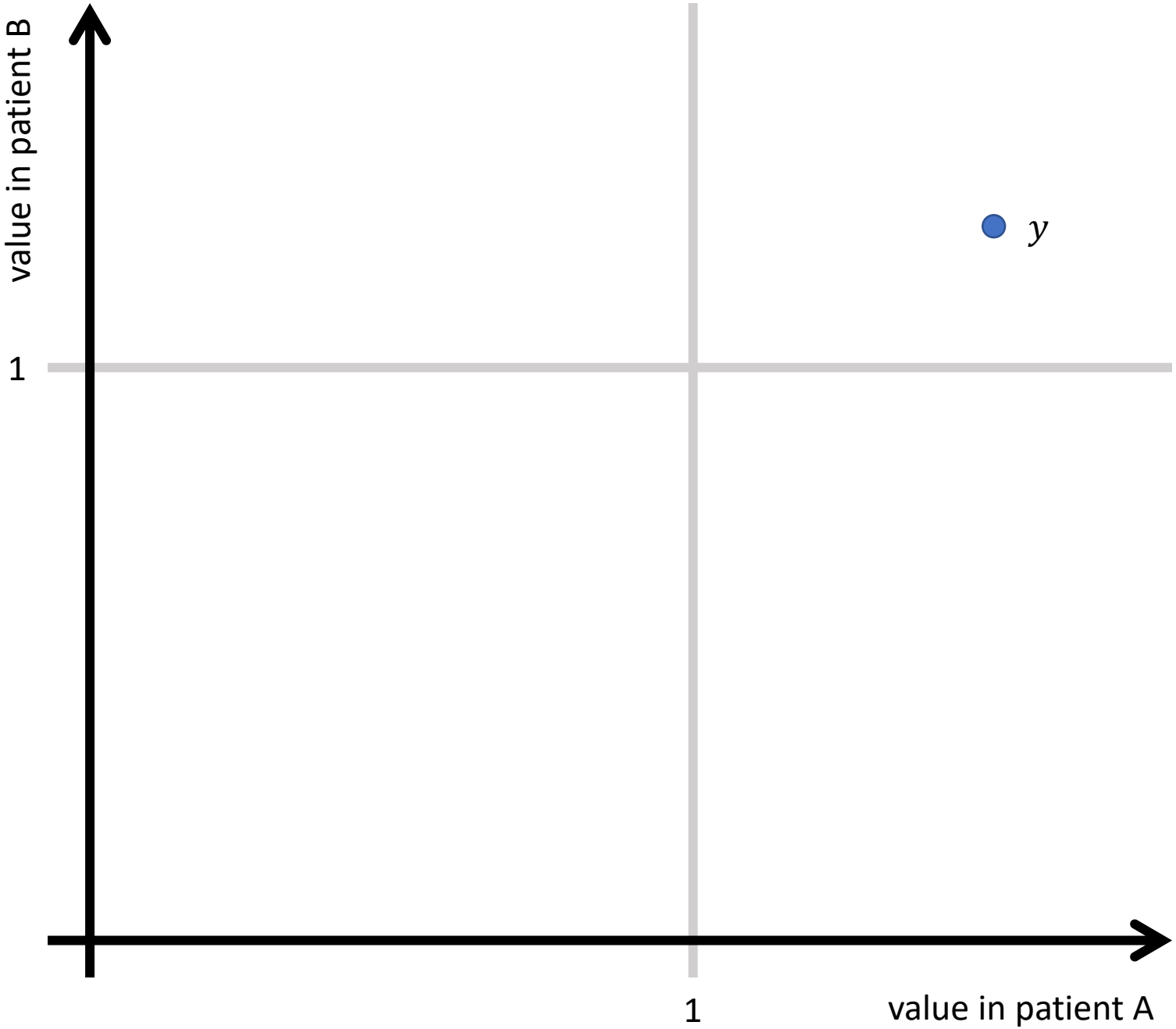
Patient	Predictor 1 (numeric)	Predictor 2 (numeric)	Outcome (numeric)
A	.5	.75	1.5
B	1	.75	1.25

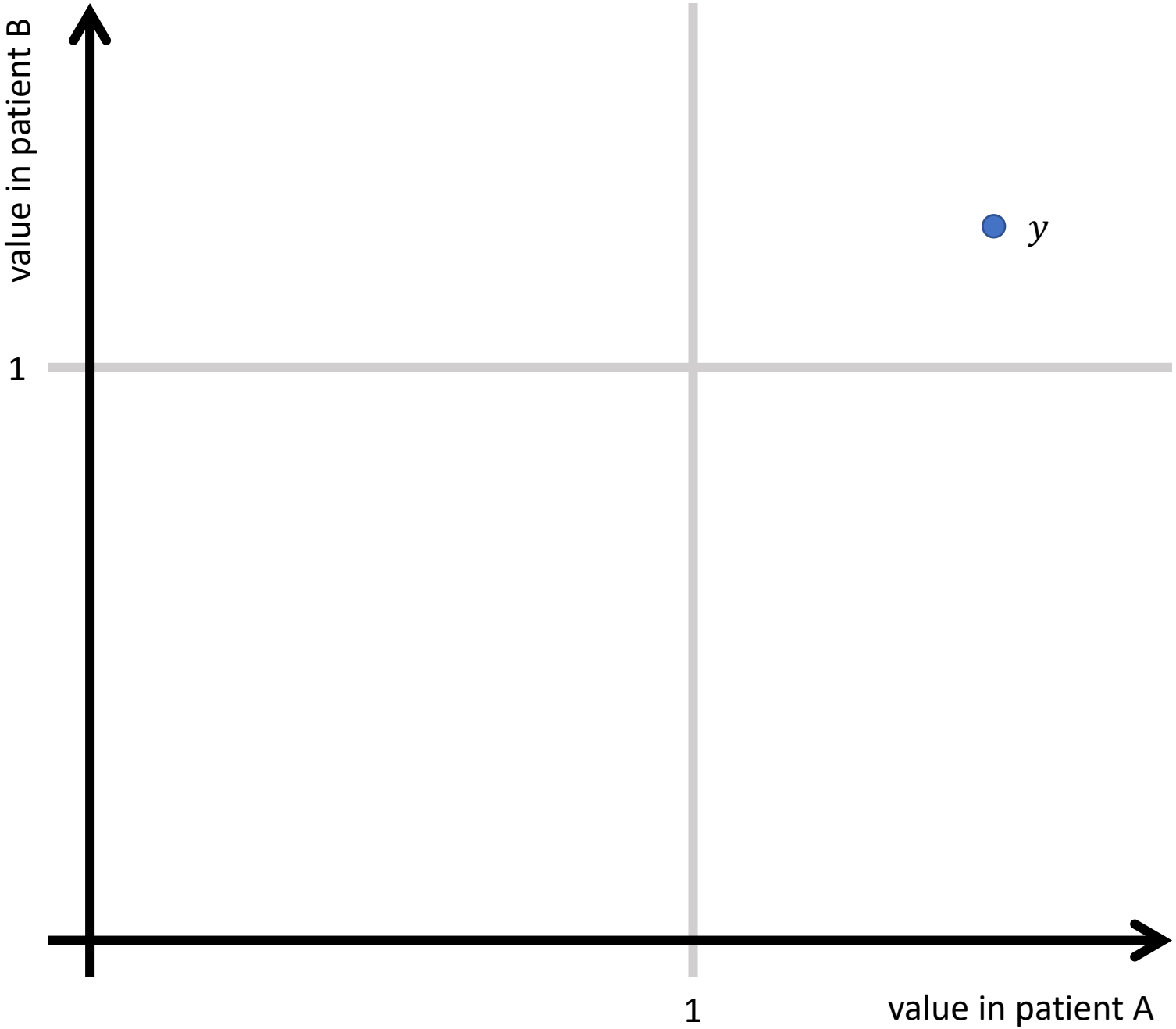
> Goal: find the linear equation that best predicts the outcome

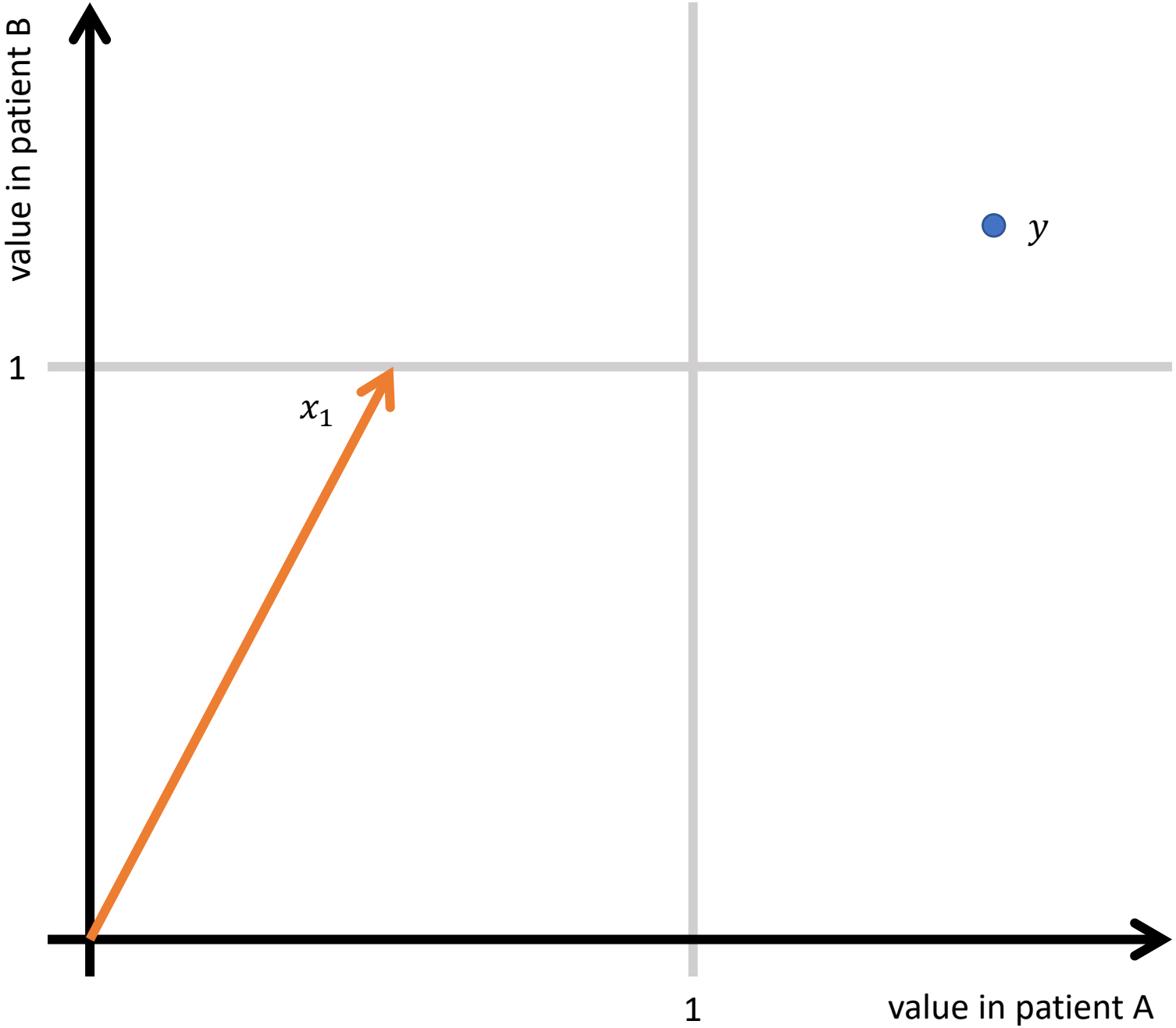
Patient	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{y}$
A	.5	.75	1.5
B	1	.75	1.25

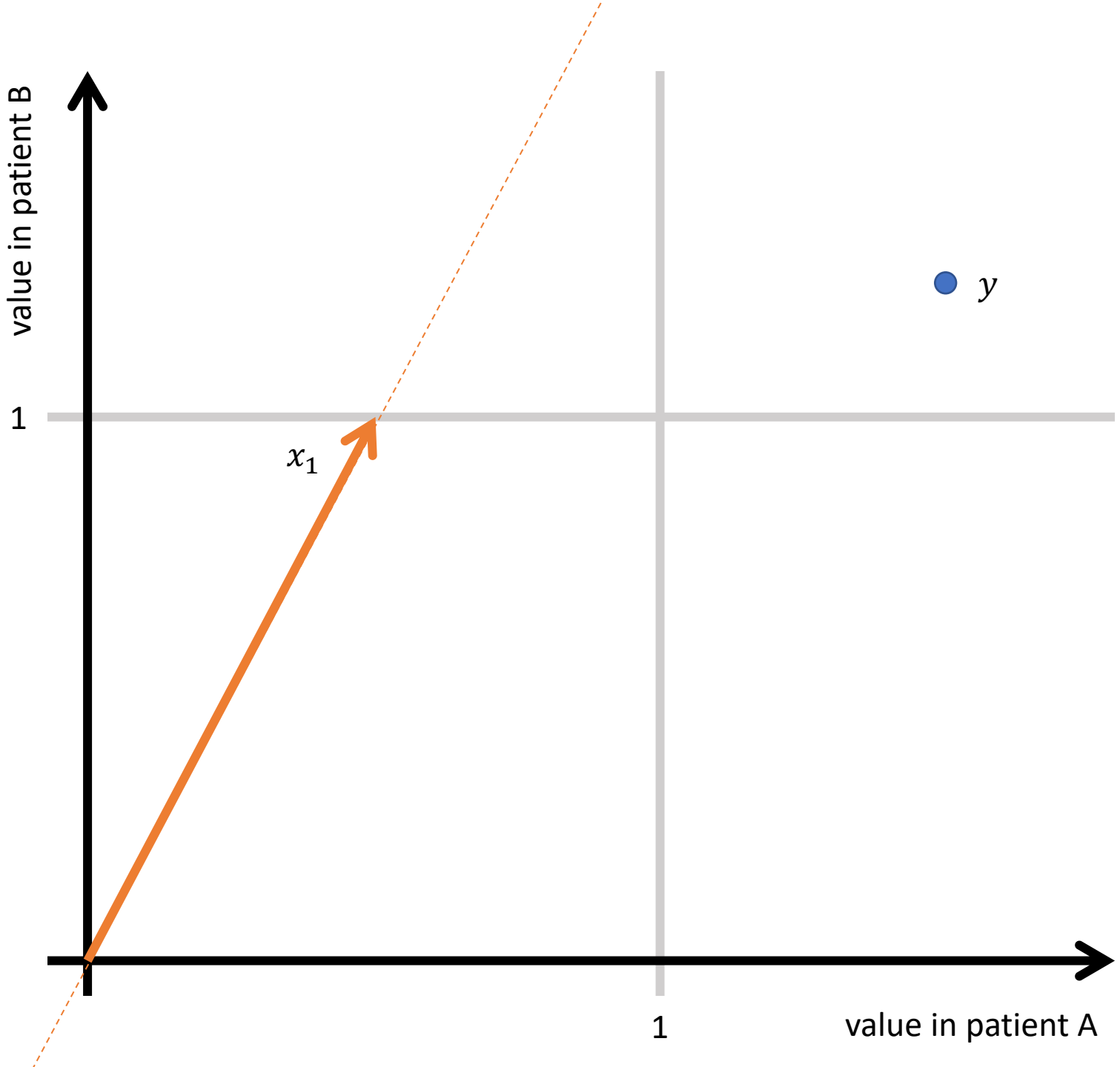
> Goal: find the linear equation that best predicts the outcome

$$b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 = \boldsymbol{y}$$


value in patient B
1
y
1
value in patient A


value in patient B
1
y
1
value in patient A


value in patient B
1
x_1
y
1
value in patient A



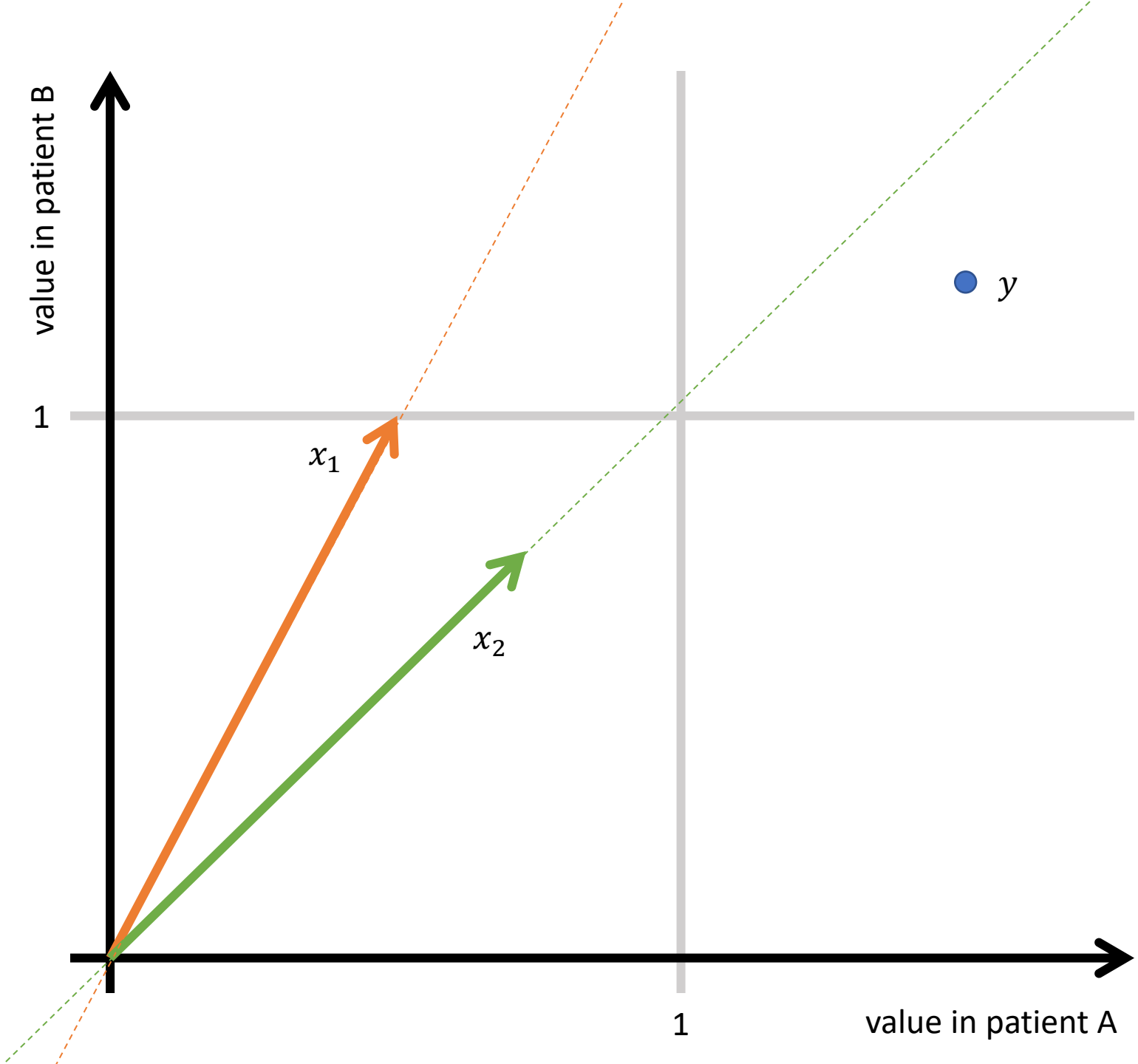
Goal: Predict y with x_1 only

From a graphical perspective, our goal is to get as close as possible to y, but we can only move in the x_1 direction

---

$p < n$

More patients than predictors.



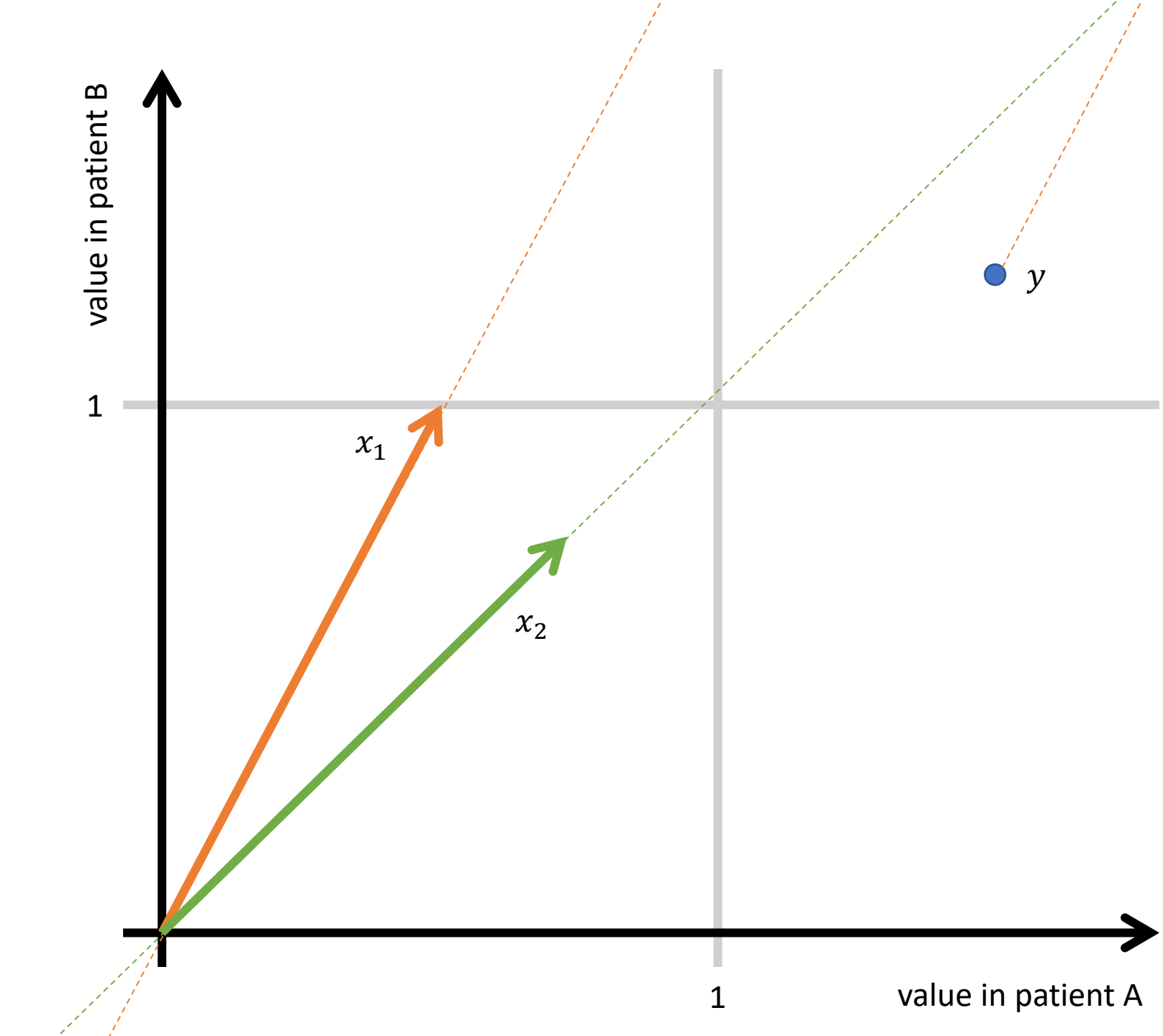
Goal: Predict y with x_1, x_2

From a graphical perspective, our goal is to get as close as possible to y. We can now move in both the x_1 direction and the x_2 direction.

$p = n$

Can always* predict perfectly on training set

*assuming linearly independent predictors



Goal: Predict y with x_1, x_2

From a graphical perspective, our goal is to get as close as possible to y. We can now move in both the x_1 direction and the x_2 direction.

$p = n$

Can always* predict perfectly on training set

*assuming linearly independent predictors