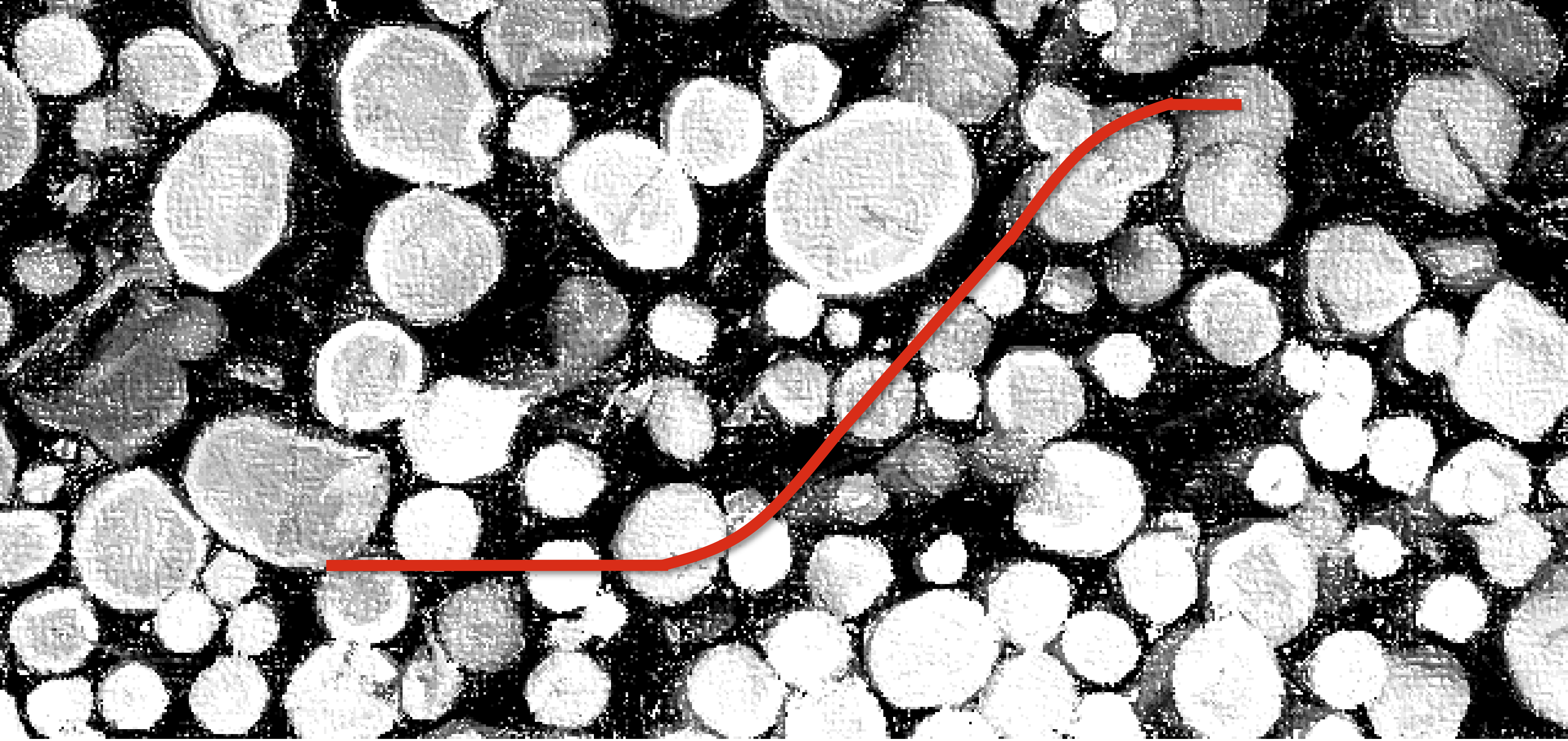


Intro to Risk Prediction Using Logistic Regression Modelling

Andrew Jones | andrew.jones40@nhs.net | The Strategy Unit

Potentially helpful resources

Script (and dataset) to accompany these slides

https://bit.ly/2TV6z9c .

Click "Clone or Download" to download a zip file. Note: won't work in Internet Explorer.

IDRE data analysis examples (see logistic regression!):
https://stats.idre.ucla.edu/other/dae/

Interpreting odds ratios in logistic regression:

https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/

An Introduction to Statistical Learning

Widely recommended textbook. Chapter 4 on logistic regression.
http://faculty.marshall.usc.edu/gareth-james/ISL/

Feature Engineering and Selection (chapter on performance metrics)

https://bookdown.org/max/FES/measuring-performance.html

Outline

Risk prediction: definition and examples

Our task for today

Why logistic regression

How does it work?

Interpreting outputs

Performance measures

Quick demo of how it works in R

Risk

The possibility of something bad happening

- Cambridge English Dictionary

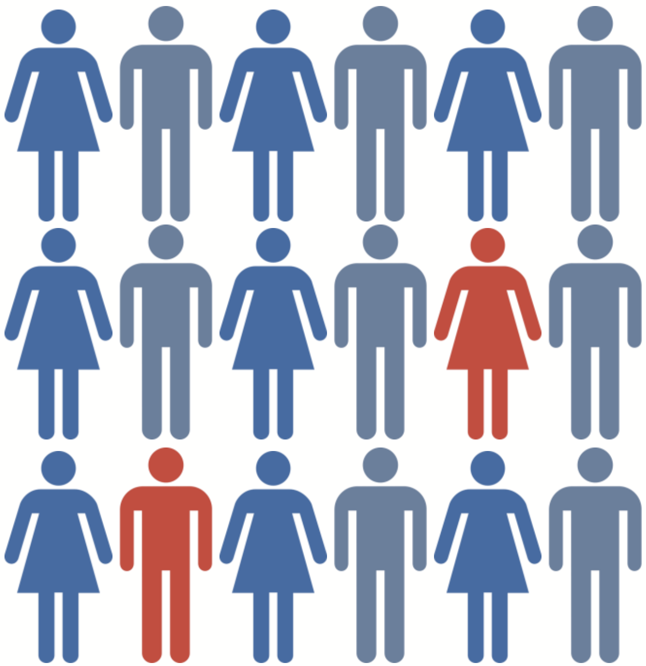
Risk Prediction

(or % chance) (?)

Predicting the probability of something (bad) happening

based on one or more known factors

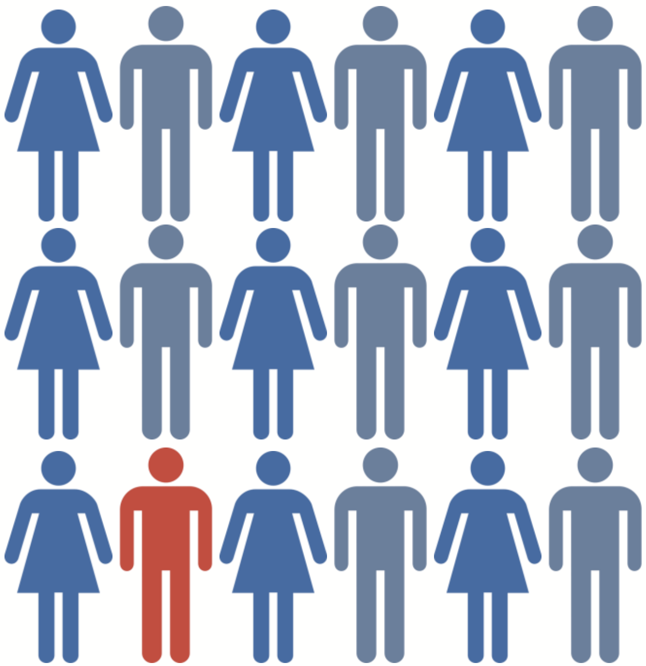
Uses of risk prediction



GP practice pop.

Greatest risk of emergency admission

Uses of risk prediction



Local authority popn.

Our task – Assist Trust X

Reduce LoS for Mental Health

Predict stay > 60 days

We have been given a dataset:

Predictors *Response (or outcome)*

id	sex	age	diag	…	long_stay
1	F	72	dement		1
2	M	22	neurosis		0
…	…	…	…		…
250	F	40	mood		0

Binary


PREDICTION
hot water
EXPLANATION

Explanatory models

We wish to quantify the relationship between predictors and response

Association vs. Causation

Causal inference methods

Remit of statistics

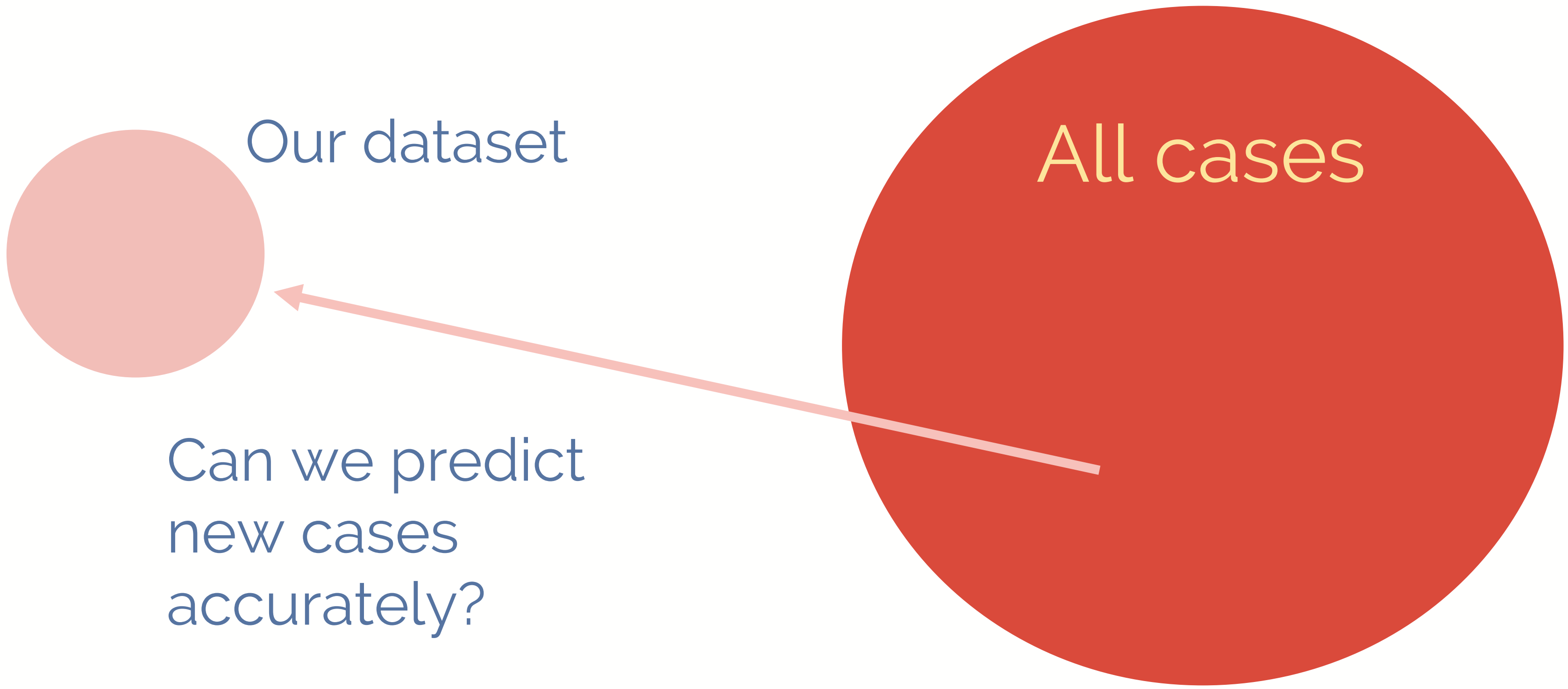
Predictive Models

Optimise prediction accuracy

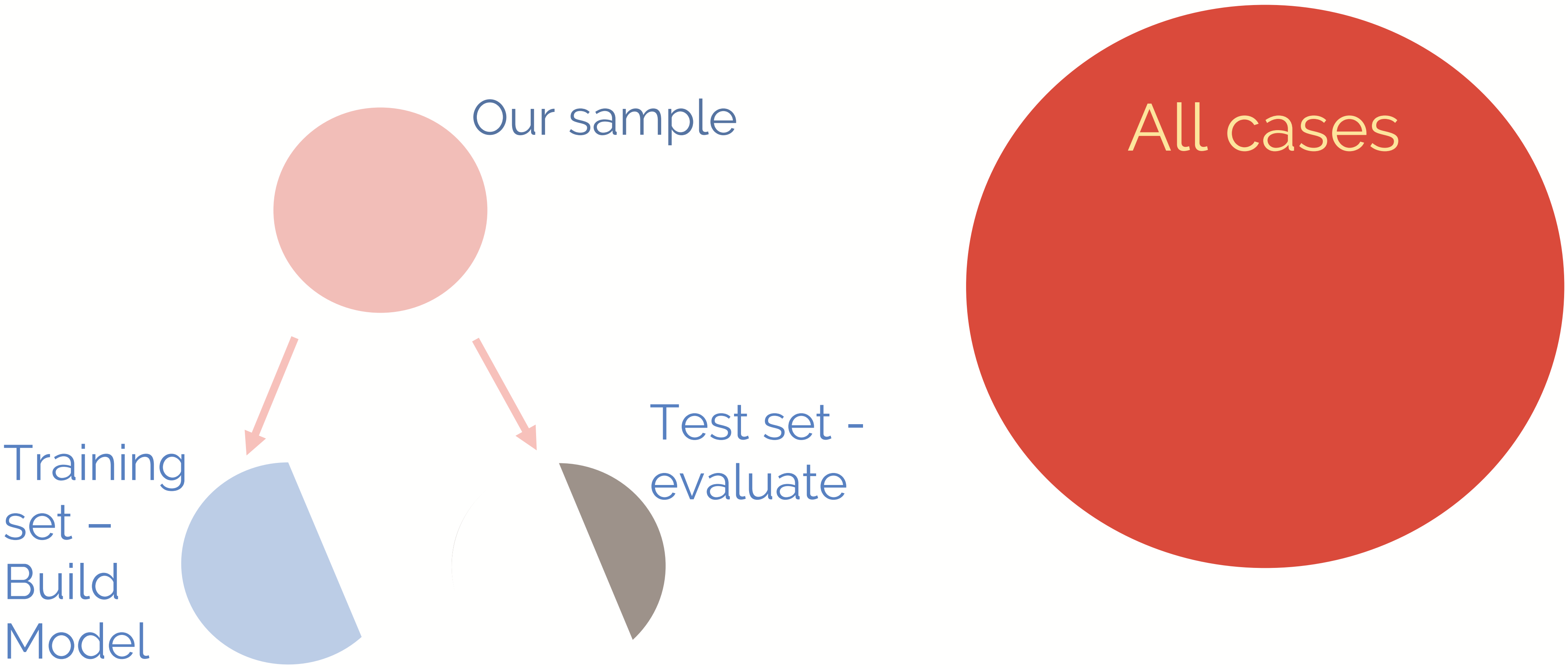
(often at the expense of interpretability)

Now largely the realm of Machine Learning

Train and Test data



Train and Test data



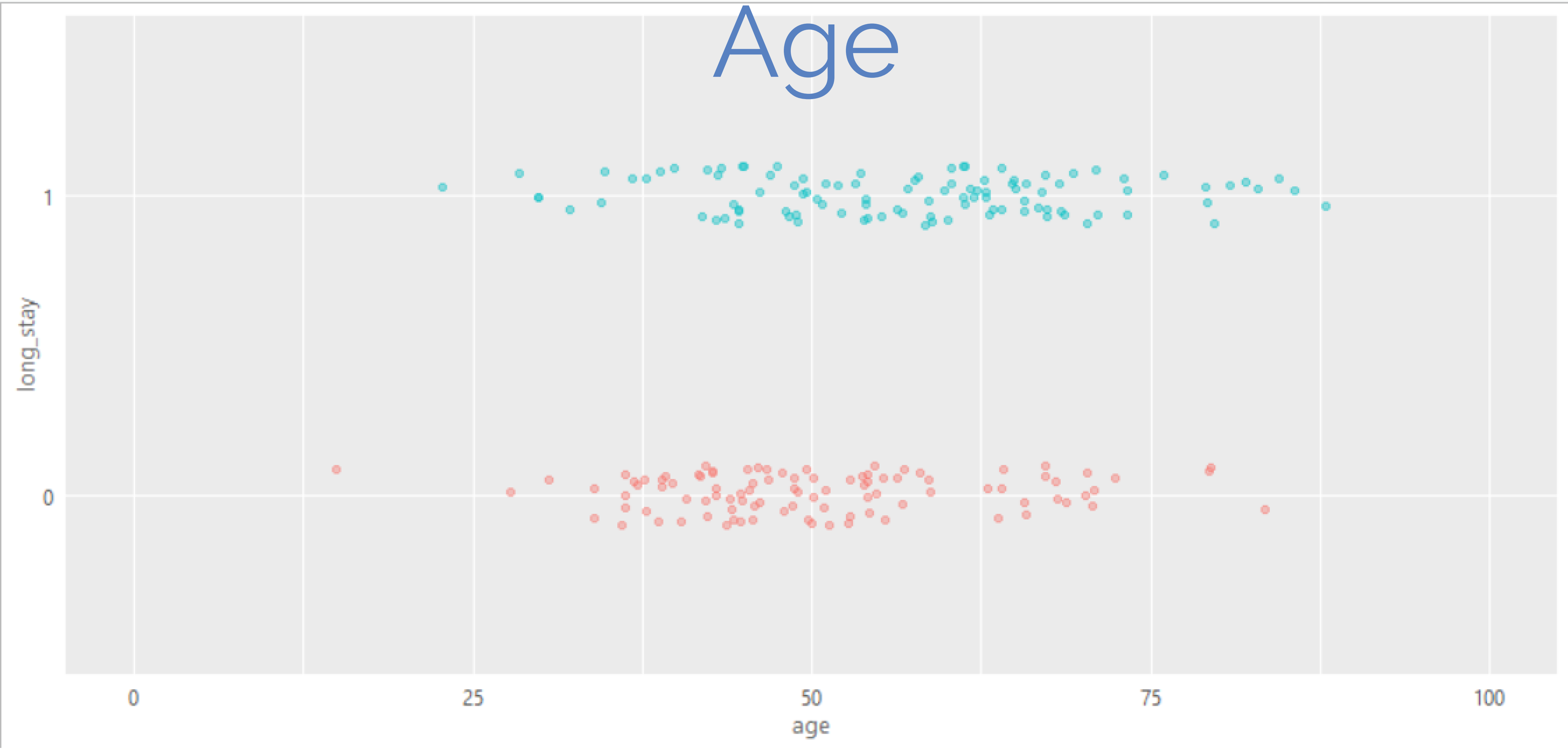
Before modelling: Know your data

Tabulate/ visualise response and predictor distributions and relationships

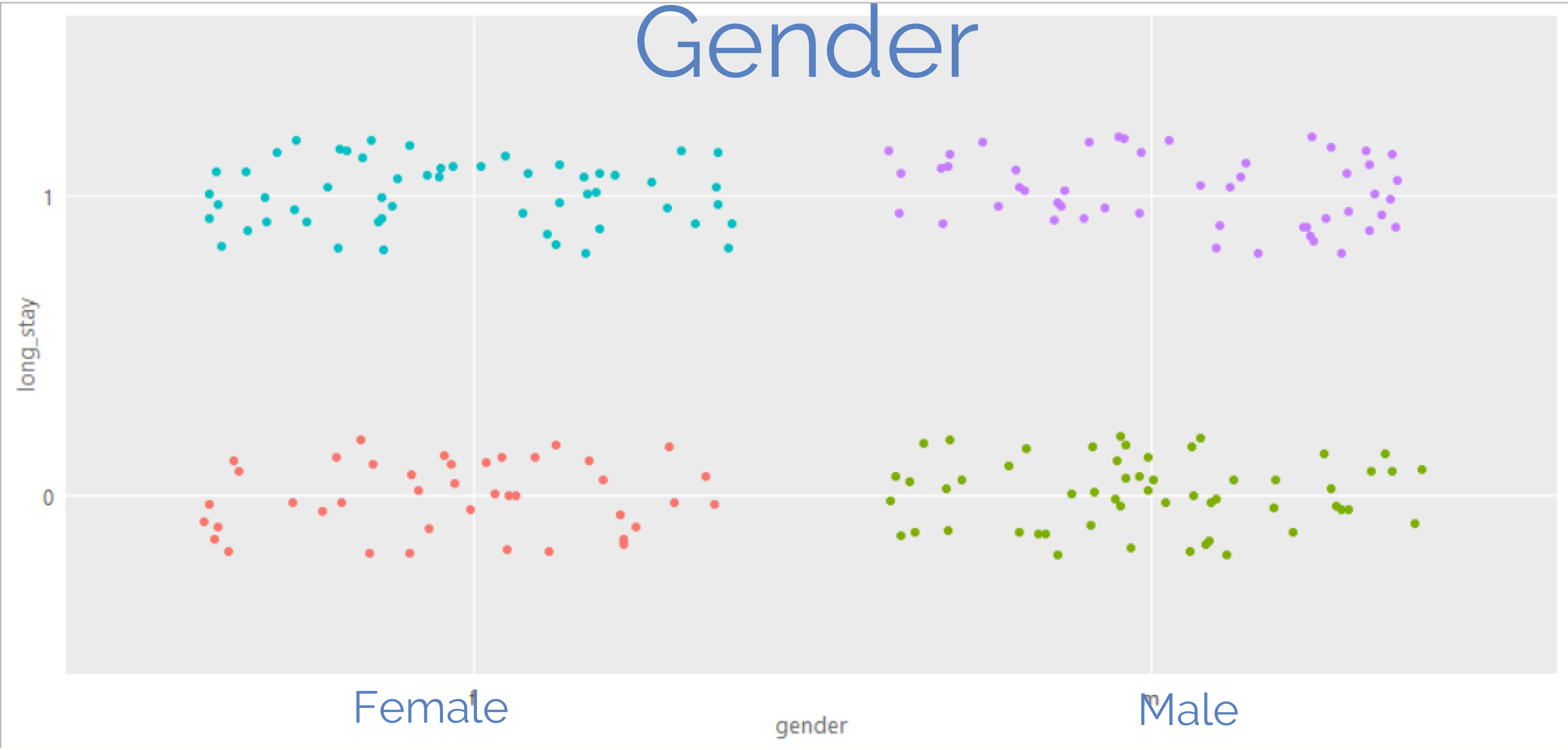
May influence the class of model you use.

	(long_stay)	(long_stay)
	0	1
(number of patients)	97	104

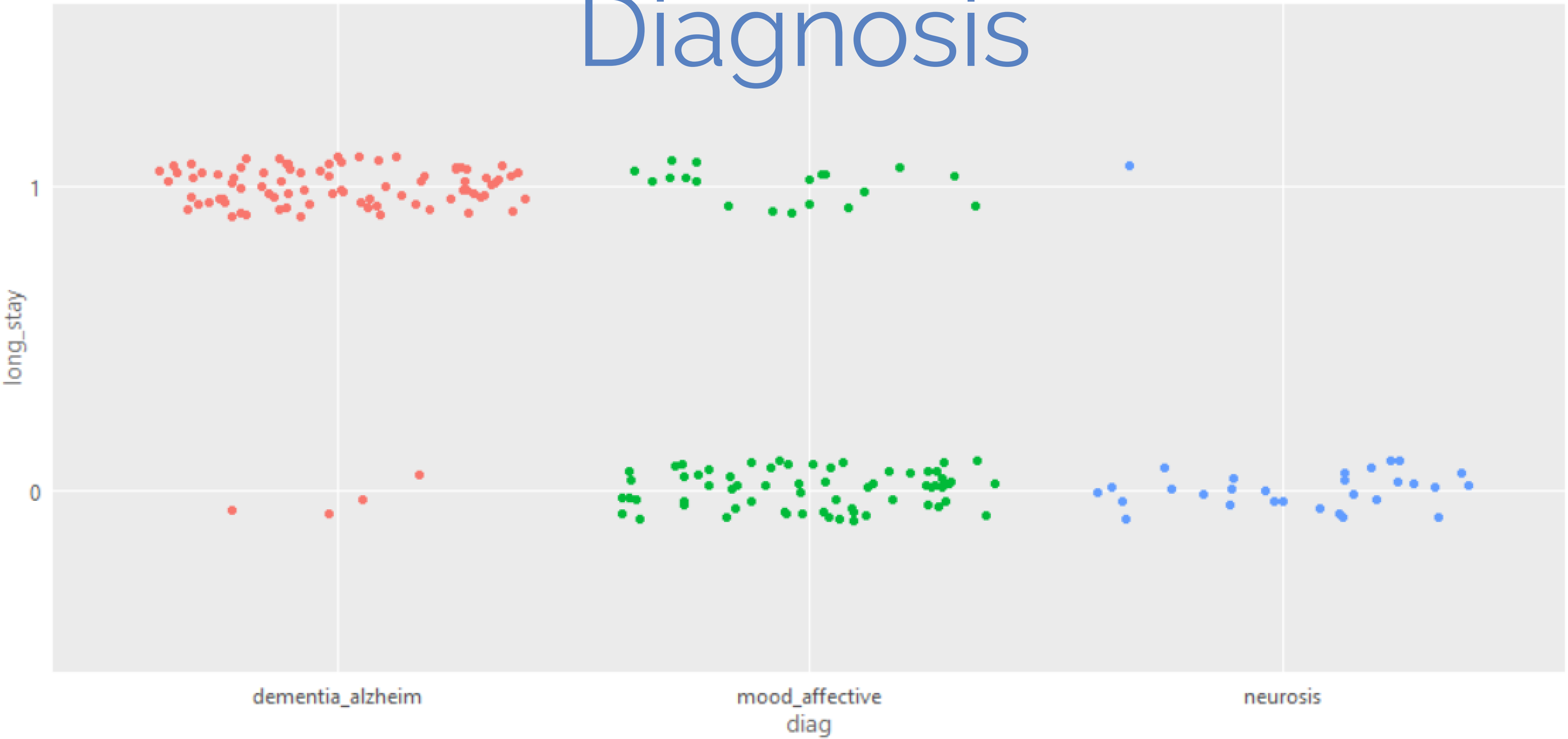
Age



Gender



Diagnosis


1
0
long_stay
dementia_alzheim
mood_affective
neurosis
diag

What do we want from a model?

binary/ qualitative

Predict the risk (probability) of long stay

based on patient characteristics

It's a "classification" problem

Which model class?

Random forest

Support vector machine

Neural network

Linear discriminant analysis

Logistic regression

Gradient boosting machine

Why logistic regression?

1. Well established (many books written*)
2. Remains powerful and easy to execute
3. Interpretable

* Some are even accessible (see resources on slide 2)

How does logisitic regression work?

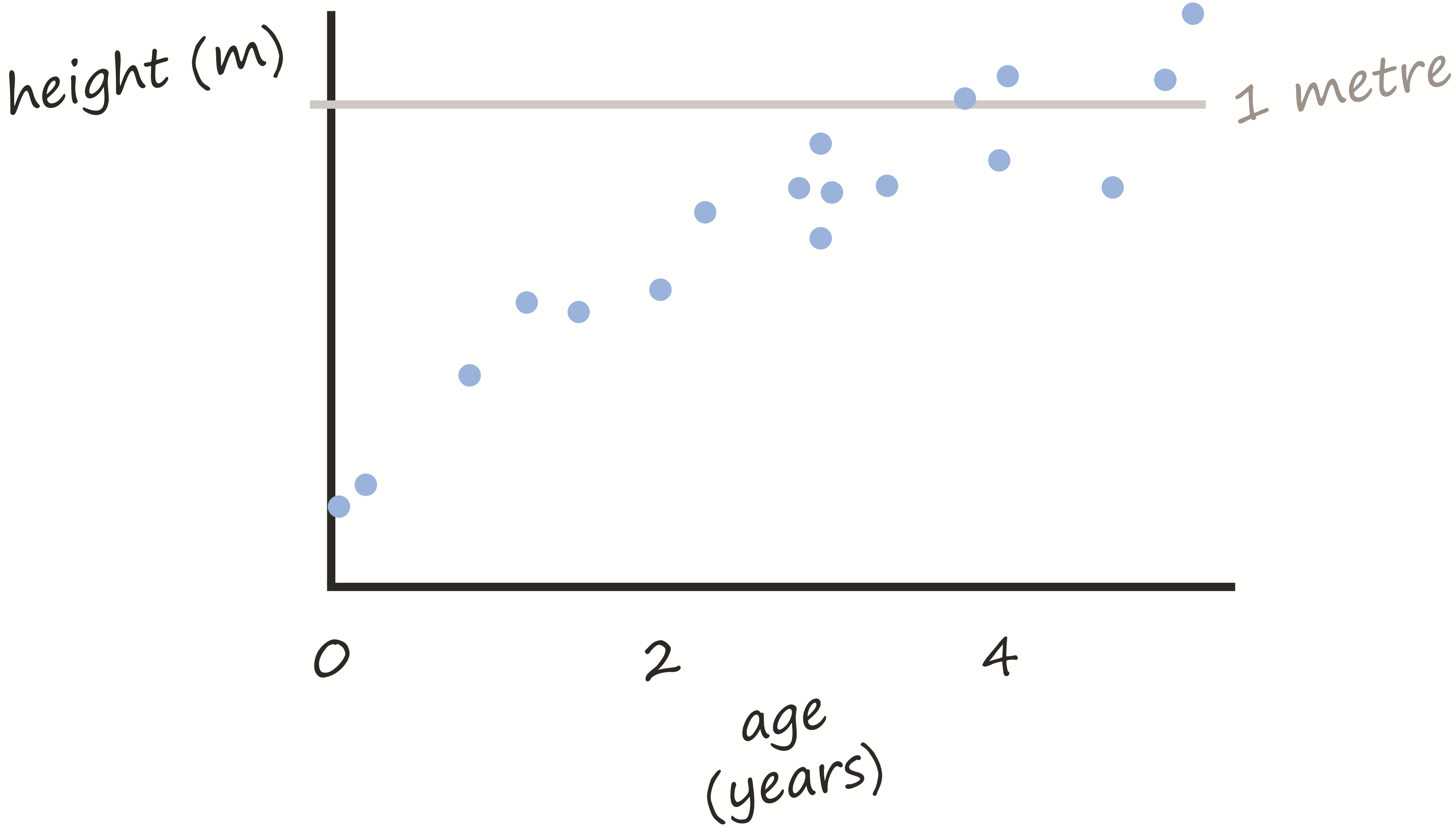
age vs. height

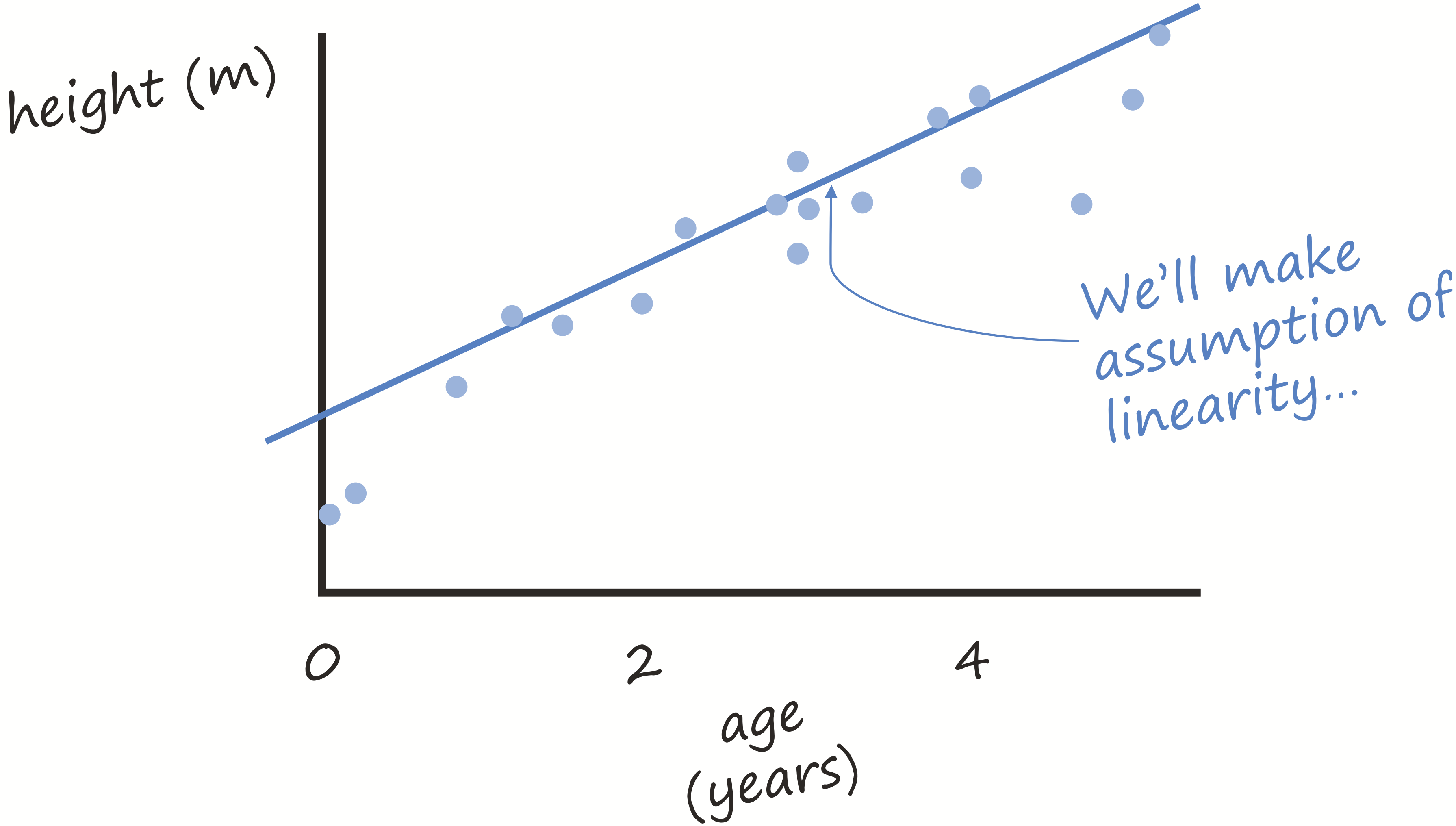
for young children

Given age, predict height

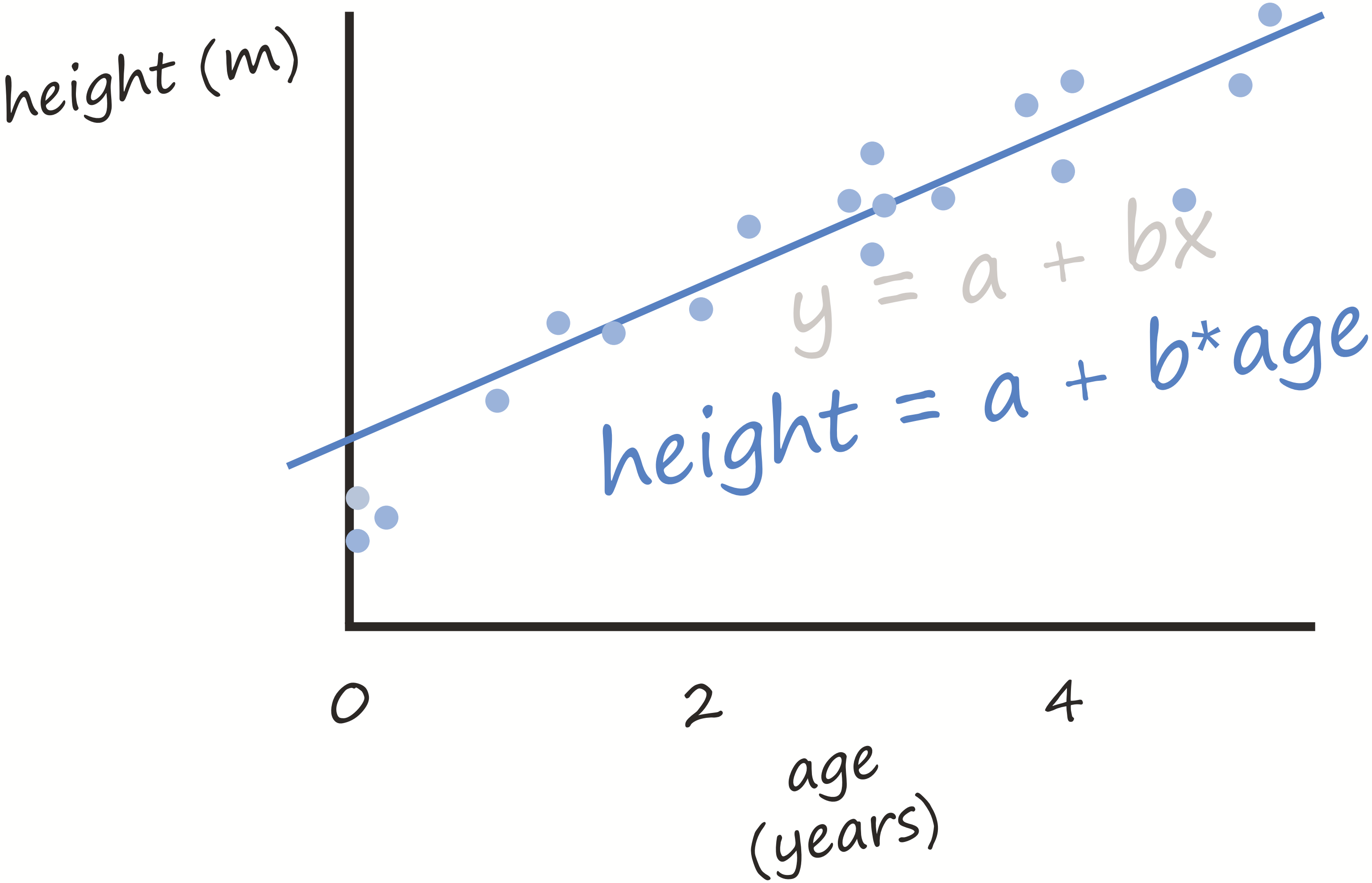
(Build a statistical model)

$$\underbrace{\text{age}}_{\text{Predictor}} \text{ vs. } \underbrace{\text{height}}_{\text{Response}}$$

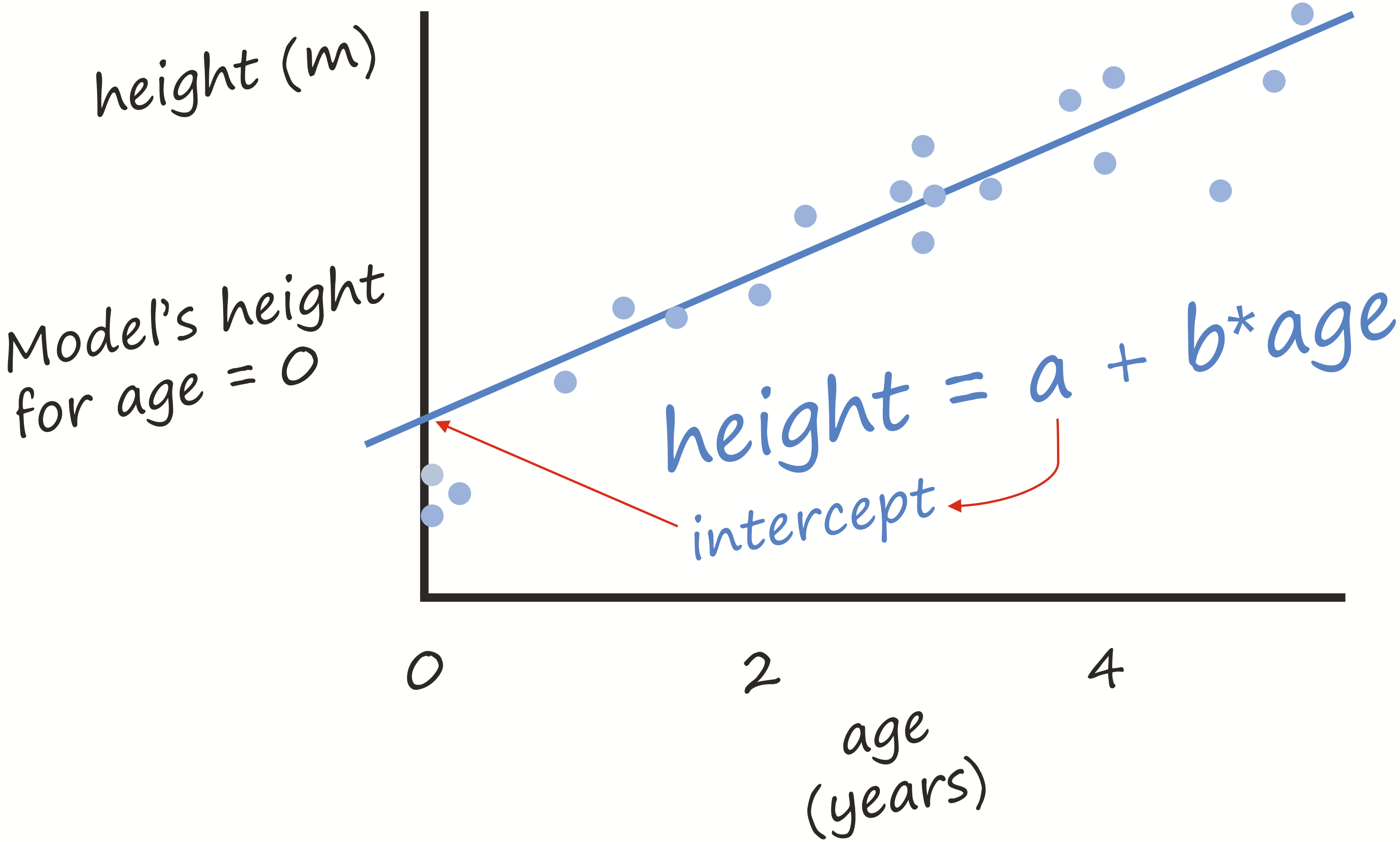

height (m)
1 metre
0
2
4
age
(years)


height (m)
0
2
4
age (years)
We'll make assumption of linearity...

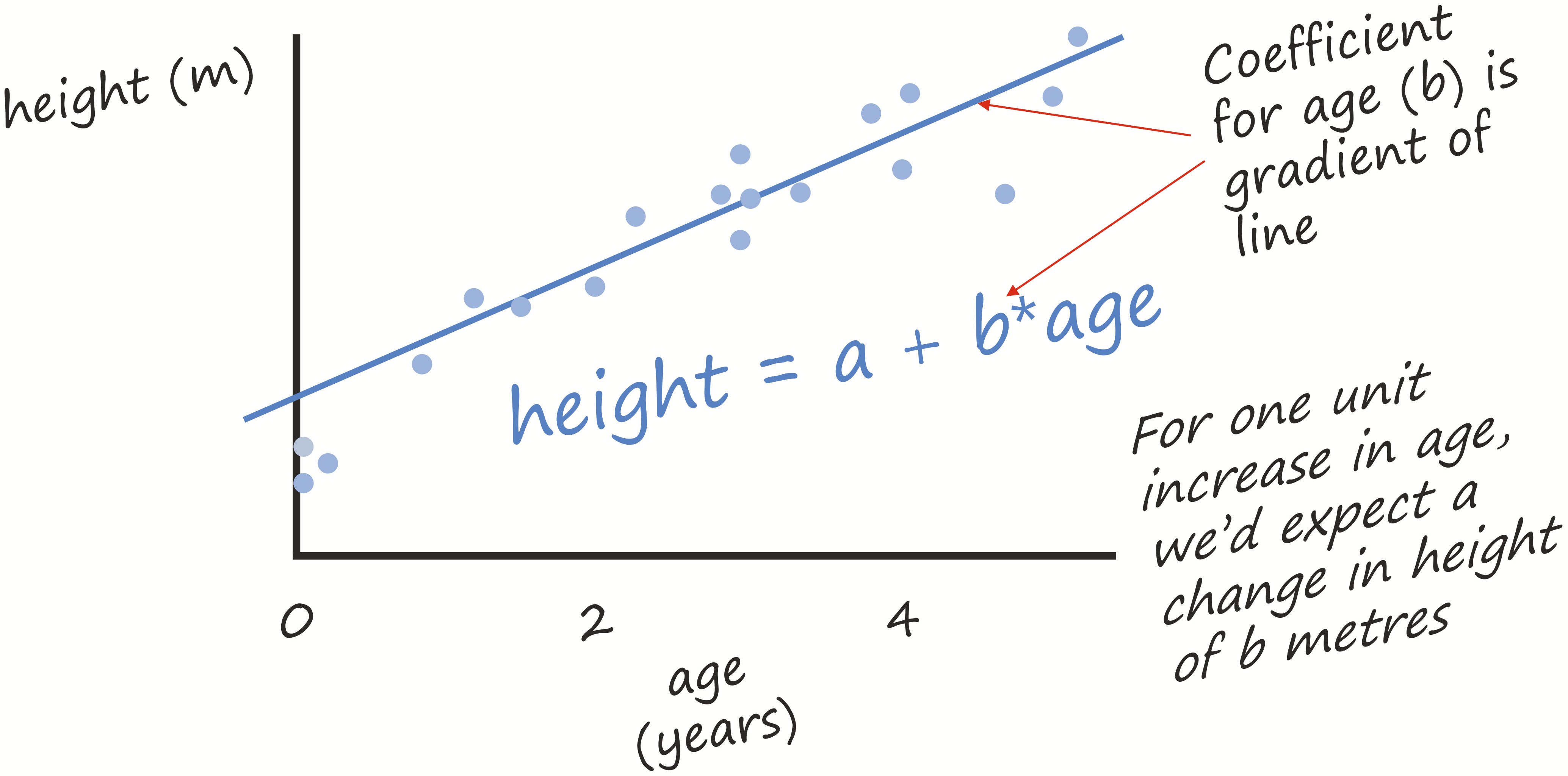
The linear model


height (m)
y = a + bx
height = a + b*age
0
2
4
age
(years)

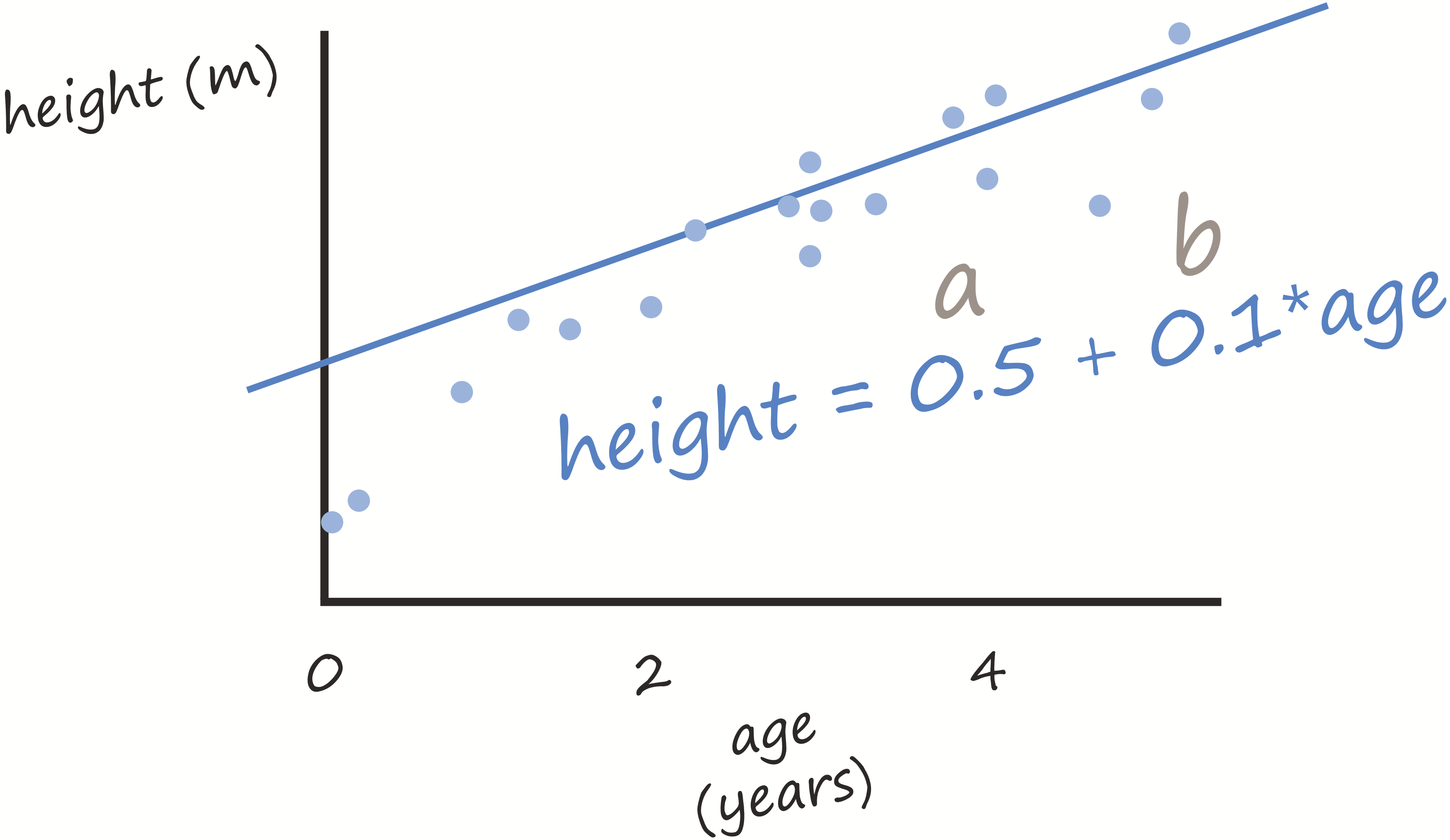
Model: age vs. height



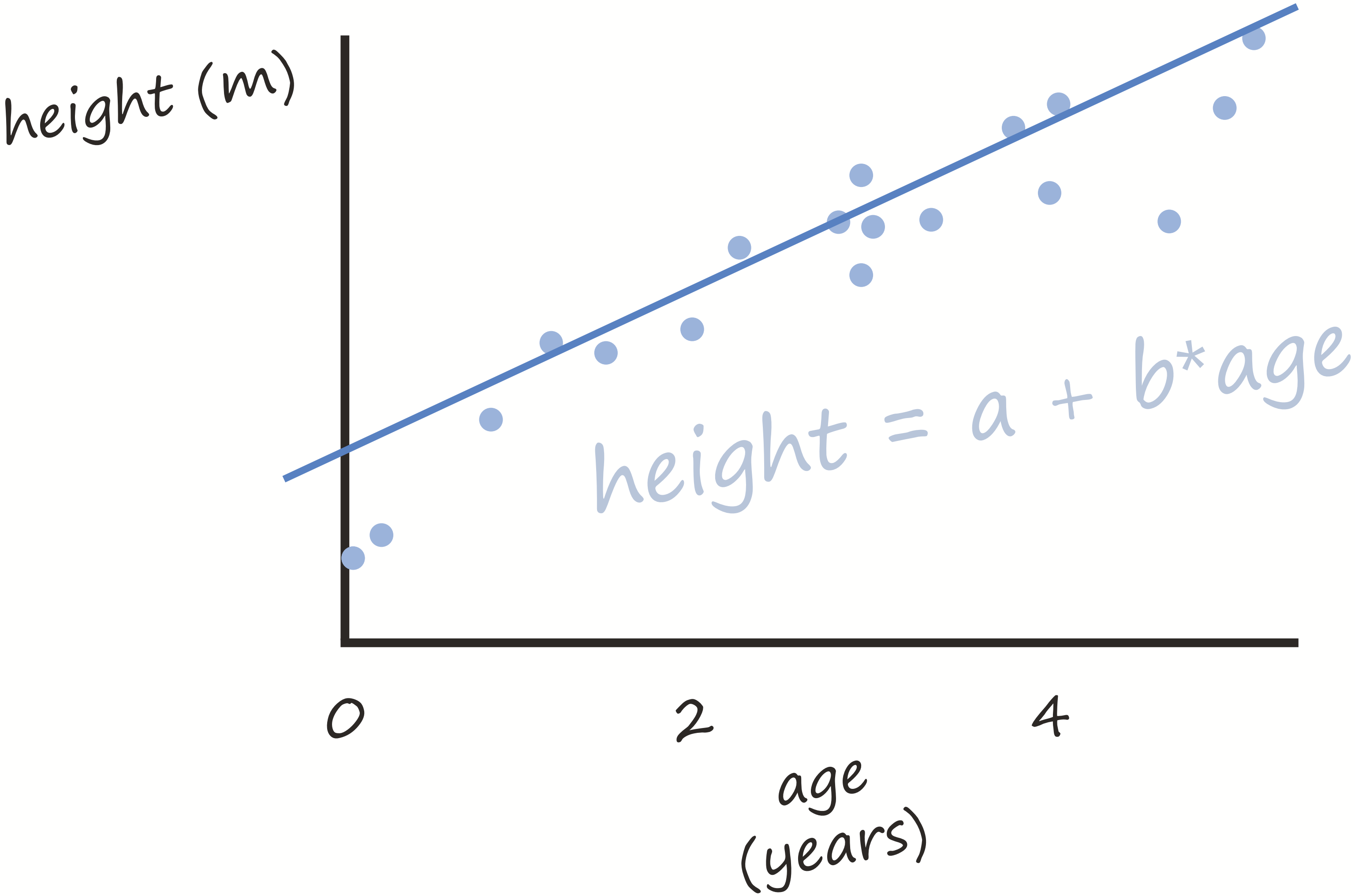
Model: age vs. height



Model: age vs. height

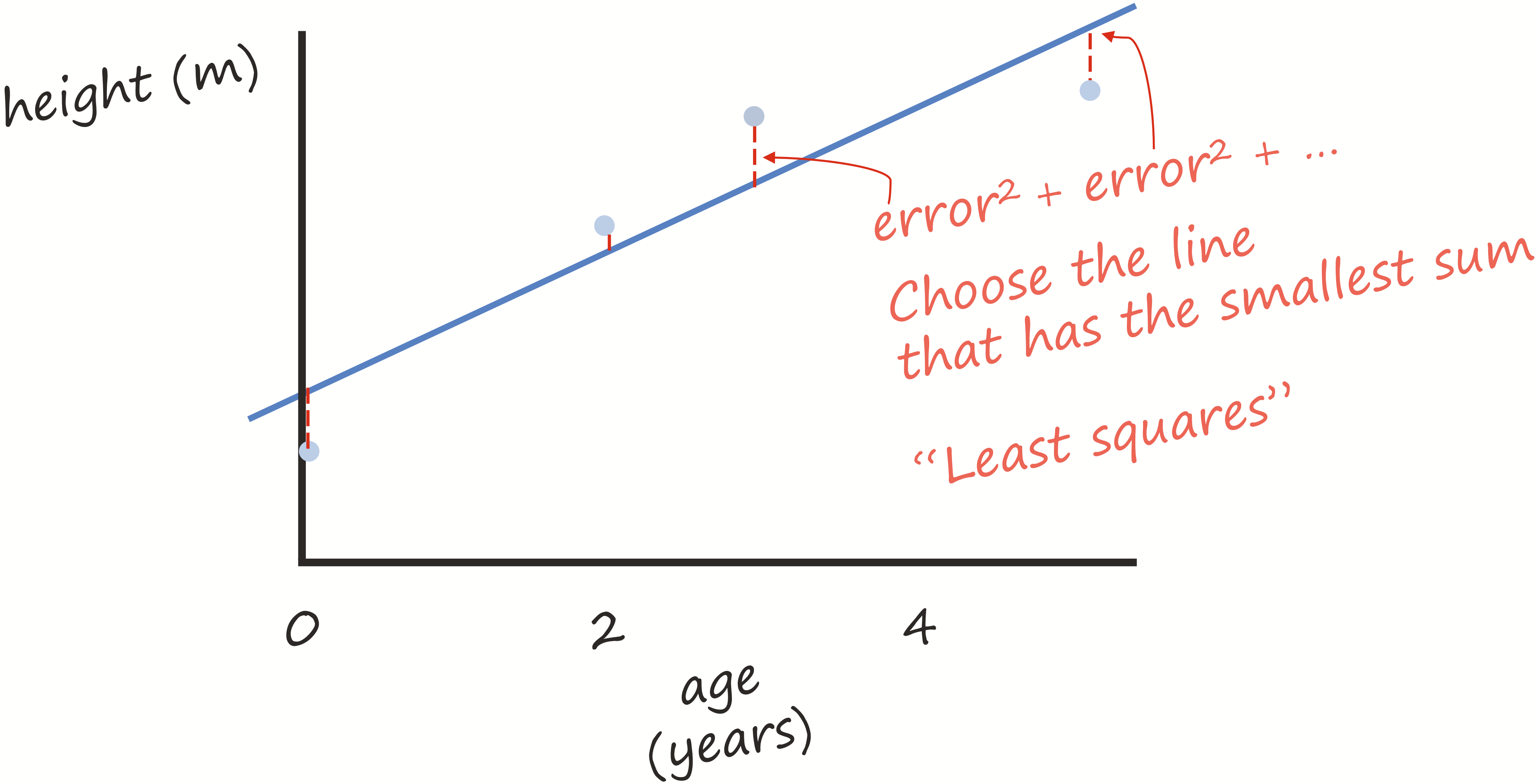


Model: age vs. height

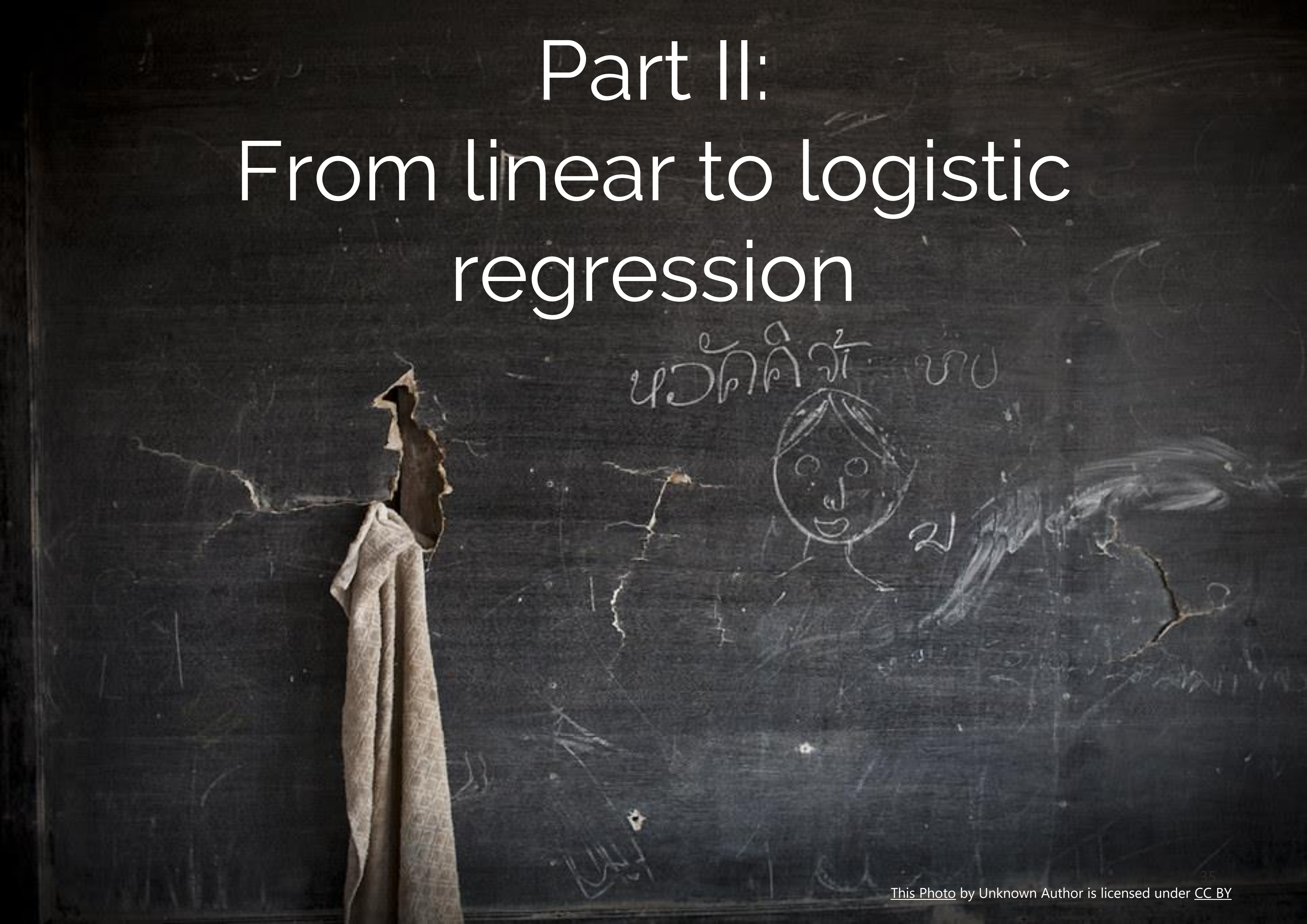


How do we choose which coefficients produce best fit?

Model: age vs. height



Part II: From linear to logistic regression



Prompted by a different question:

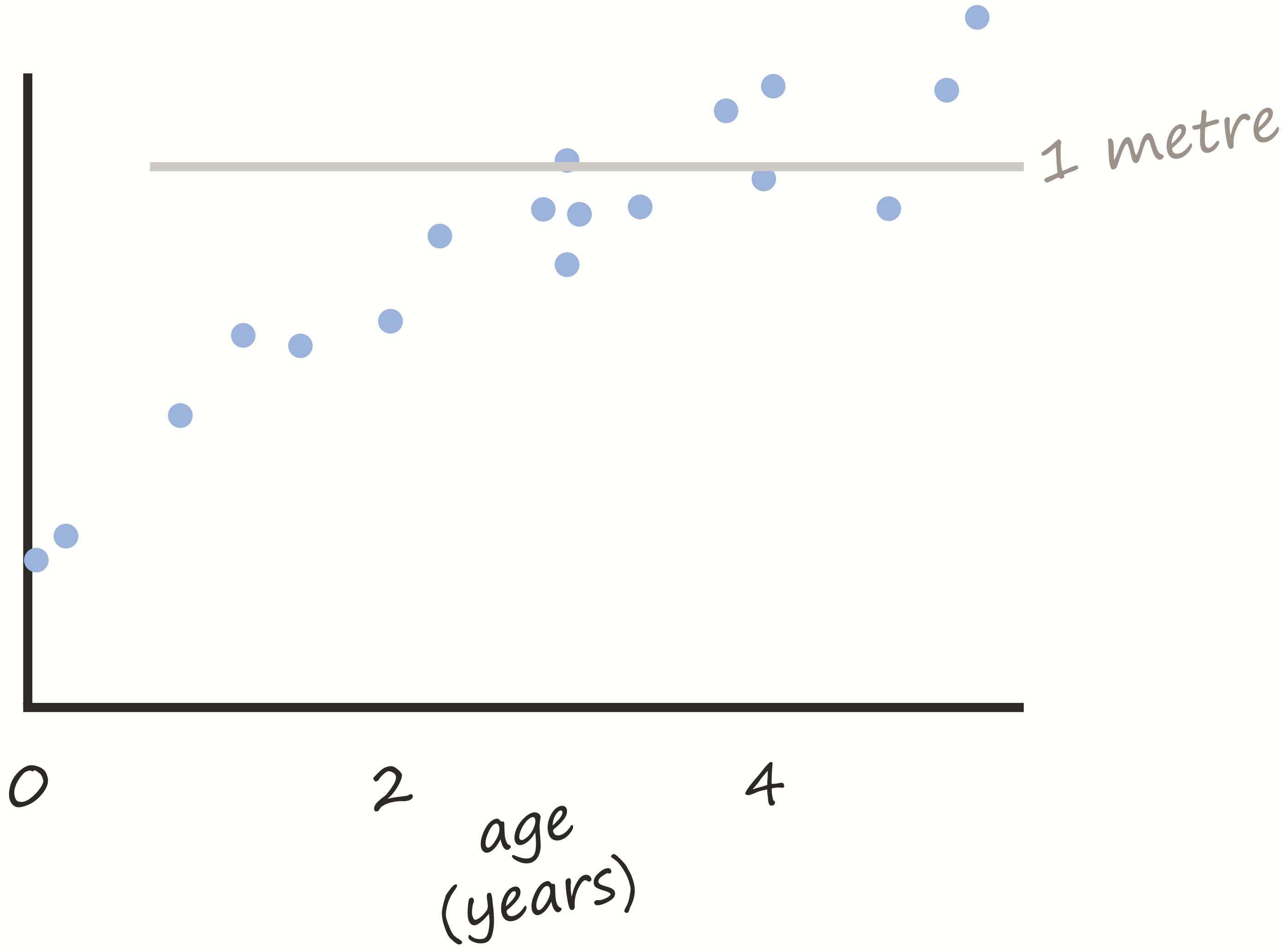
Given age, predict height

↓

Given age, what is the probability that a child is > 1m tall?

Given age, what is the probability that a child is > 1m tall?

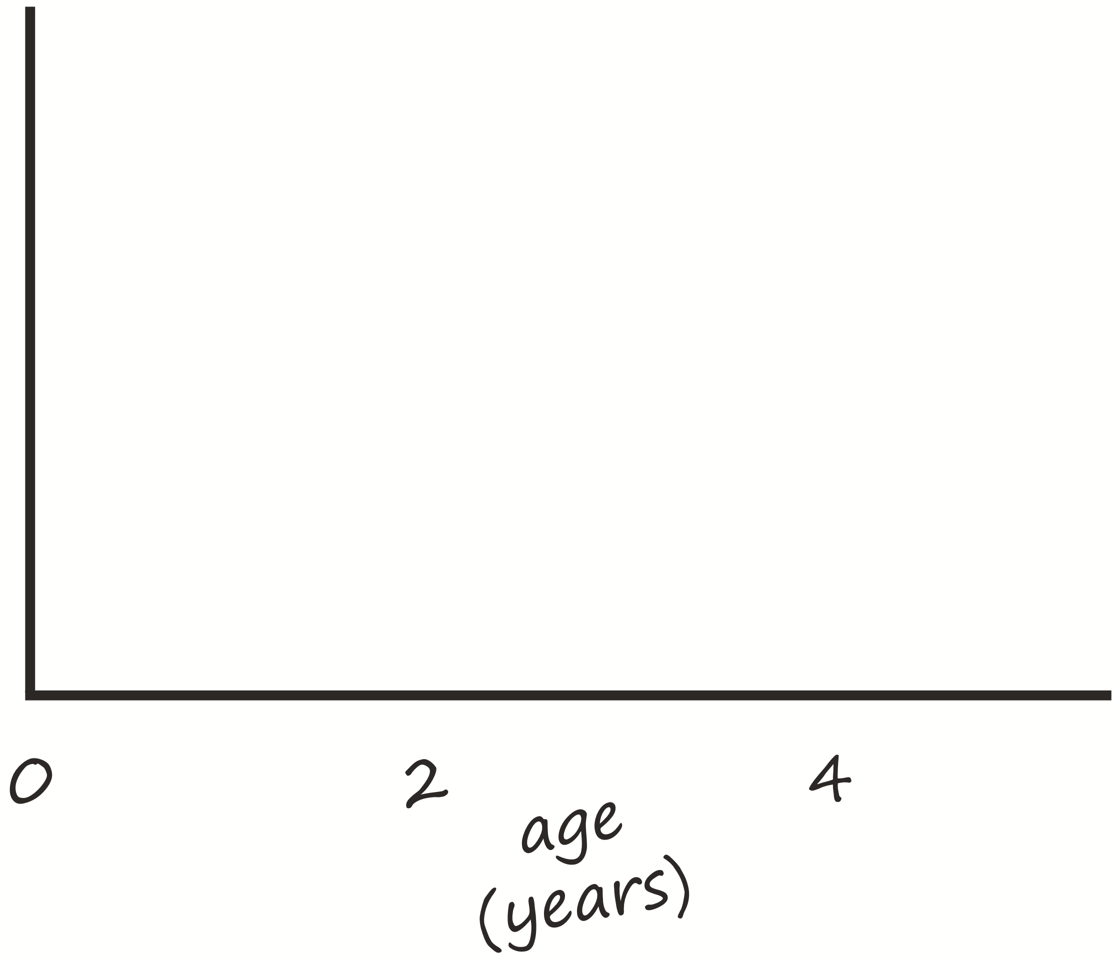
We must change our response variable:

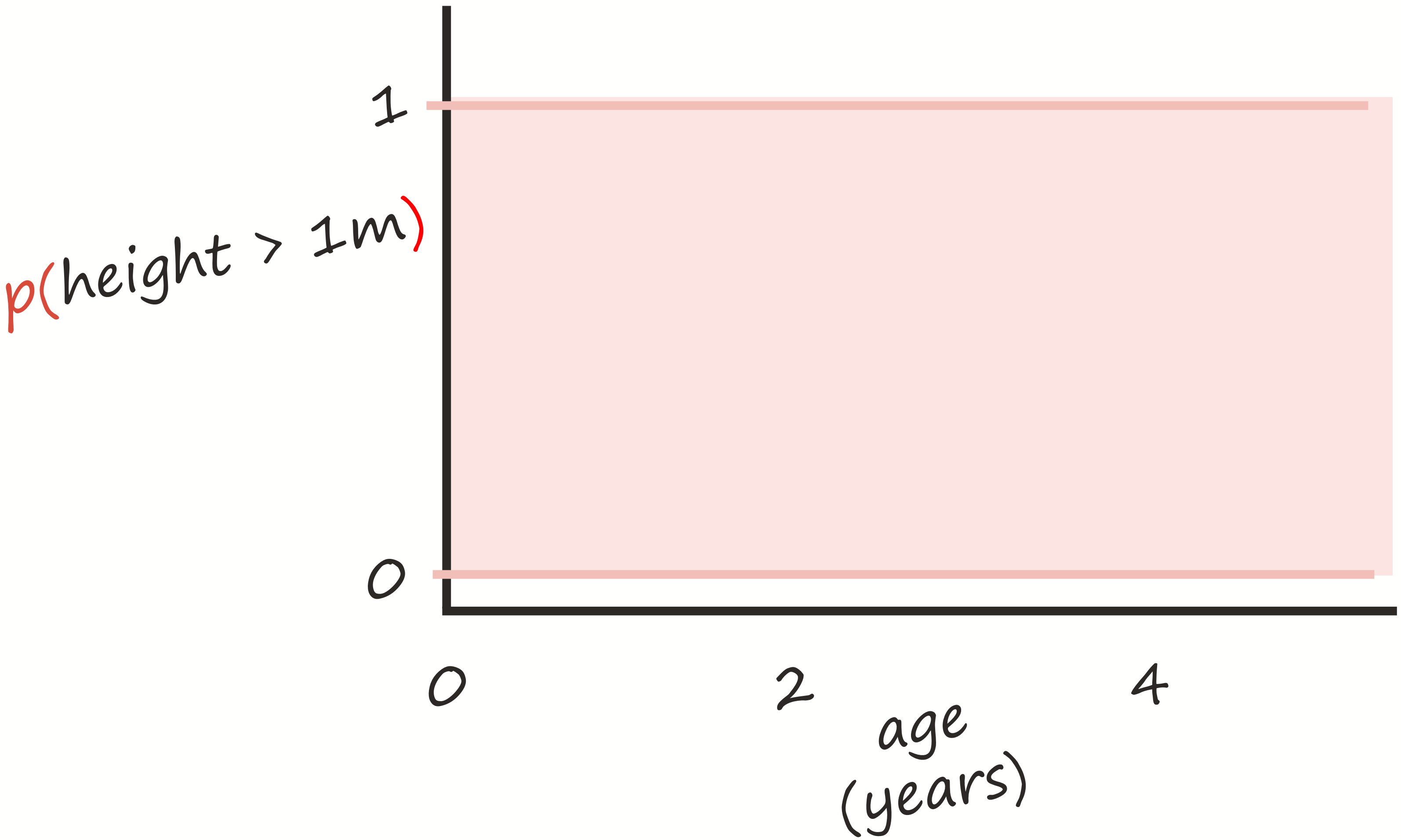


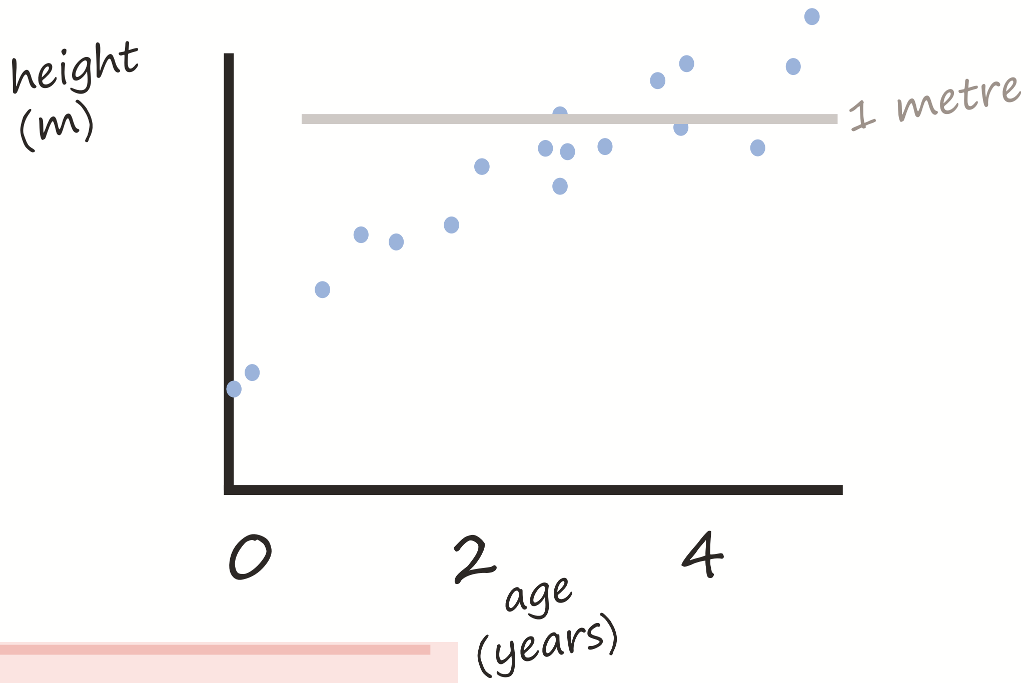
We must change our response variable:

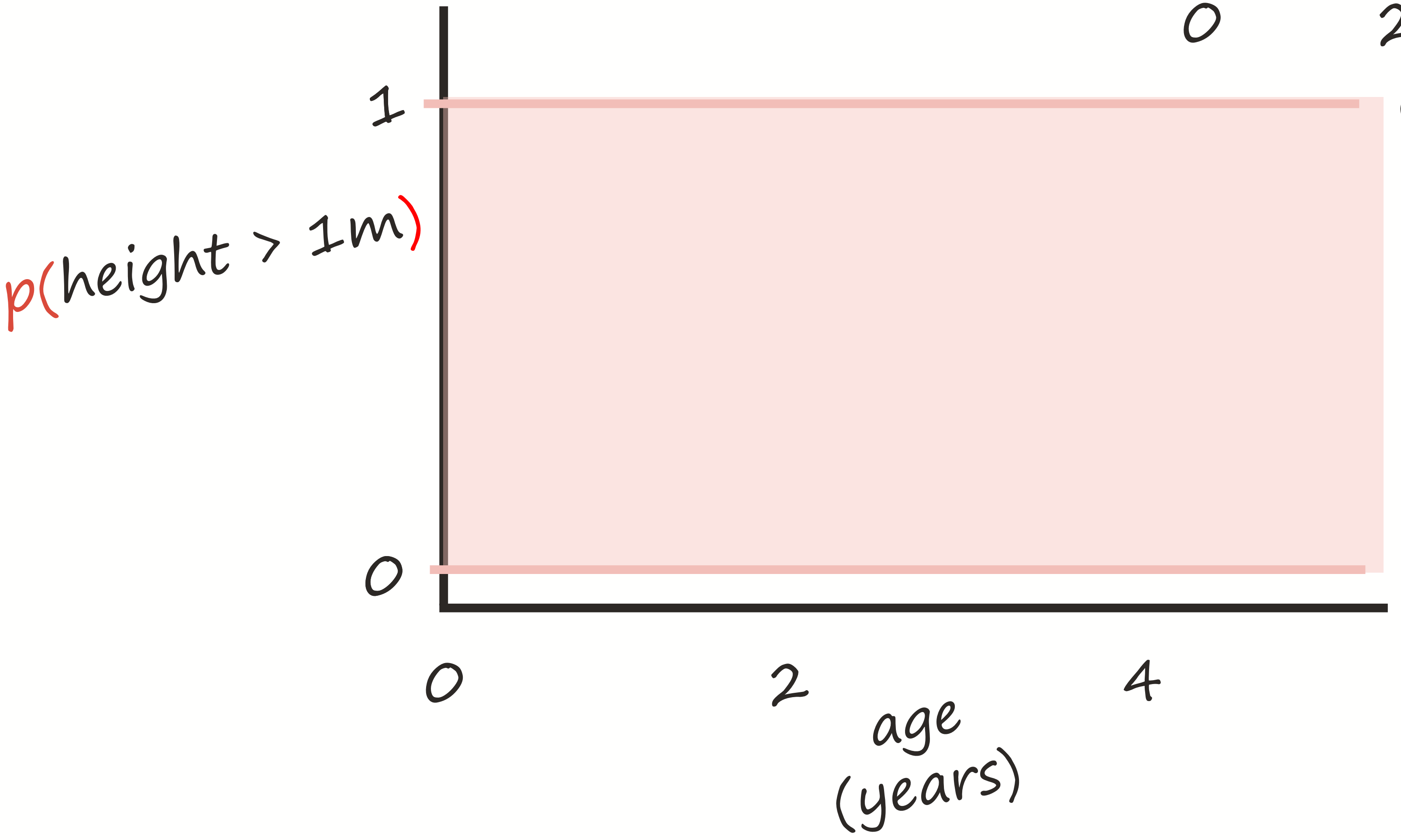
~~height (m)~~

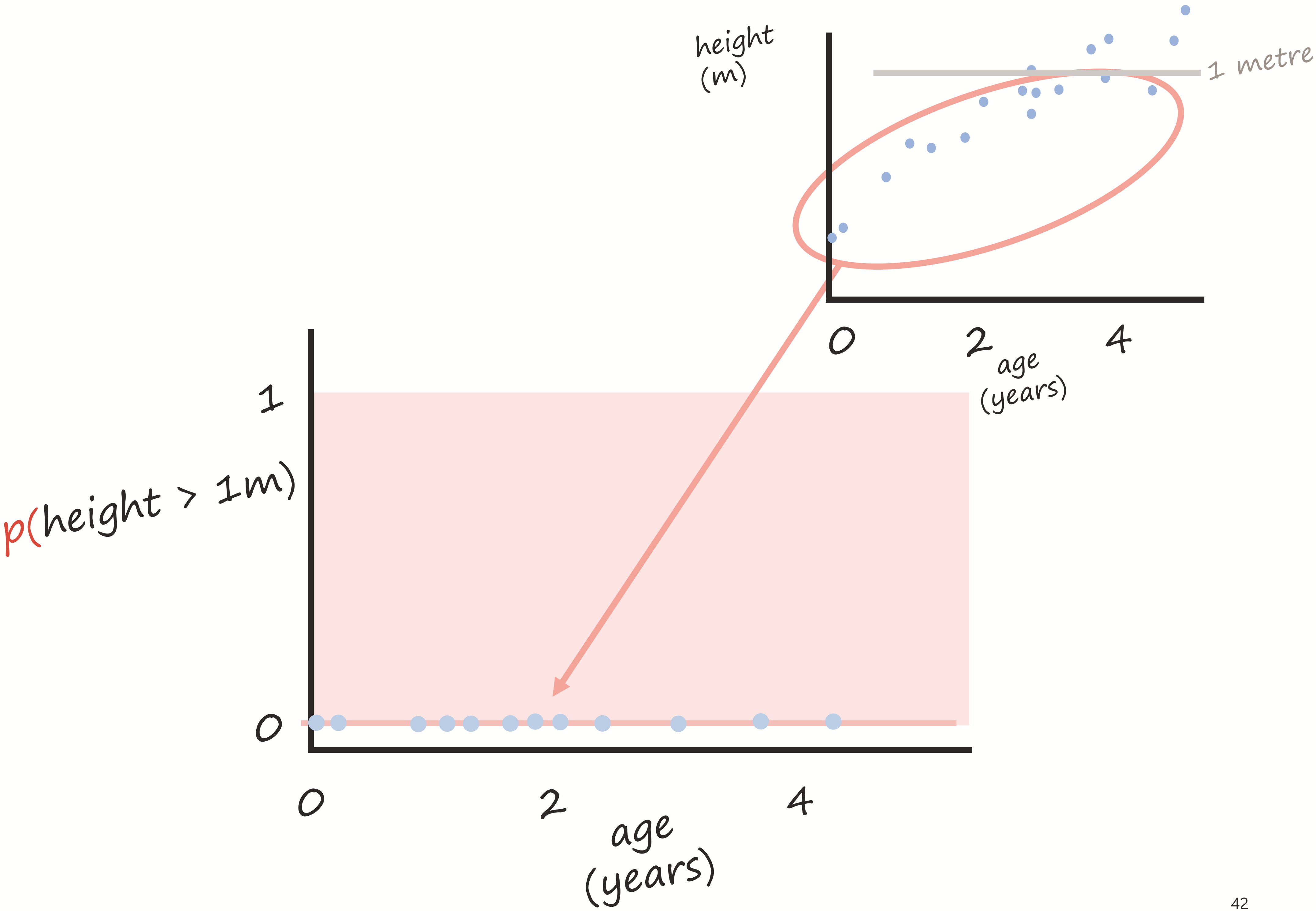
Probability height > 1m

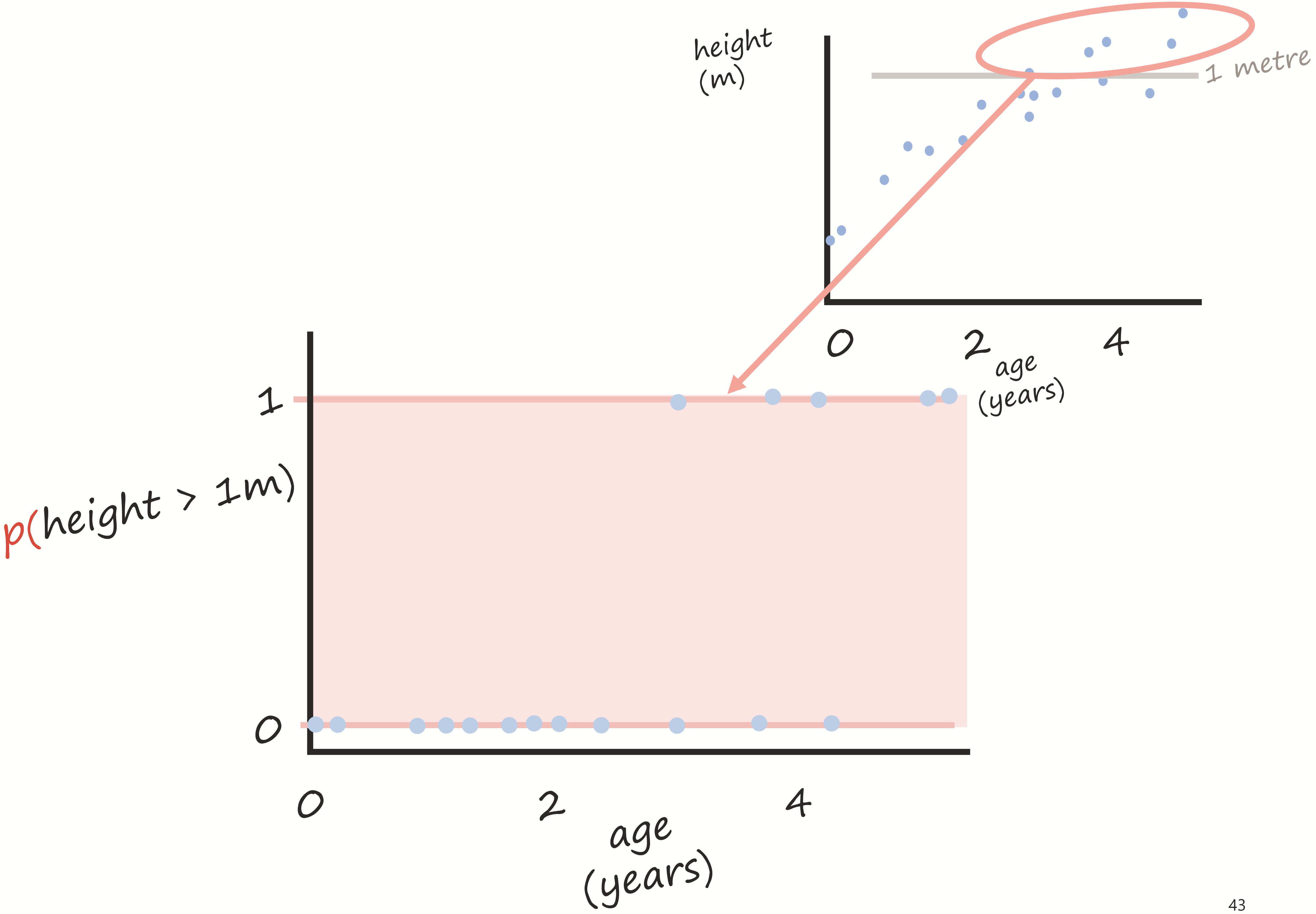



1
0
p(height > 1m)
0
2
4
age
(years)

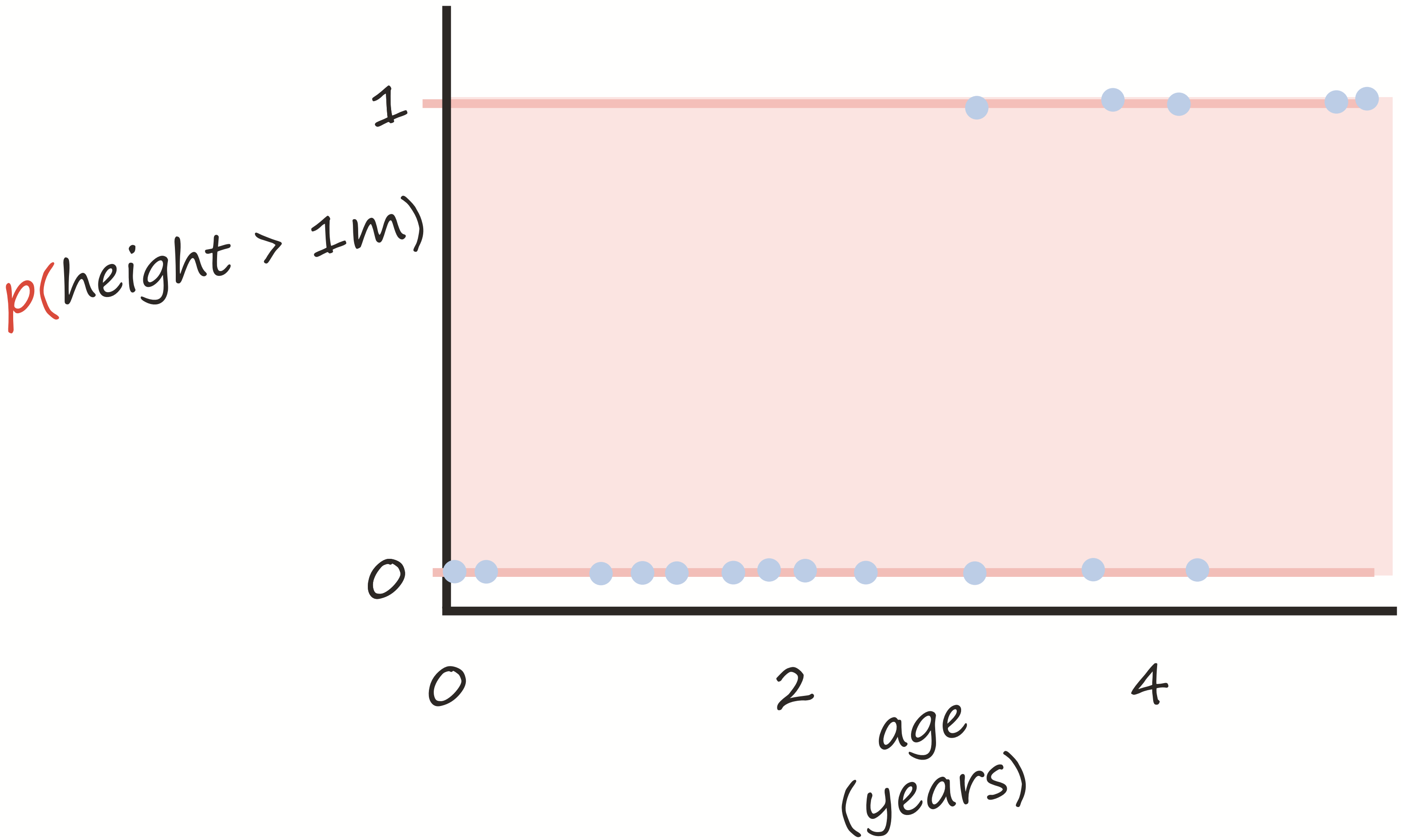

height
(m)
1 metre
0
2
4
age
(years)

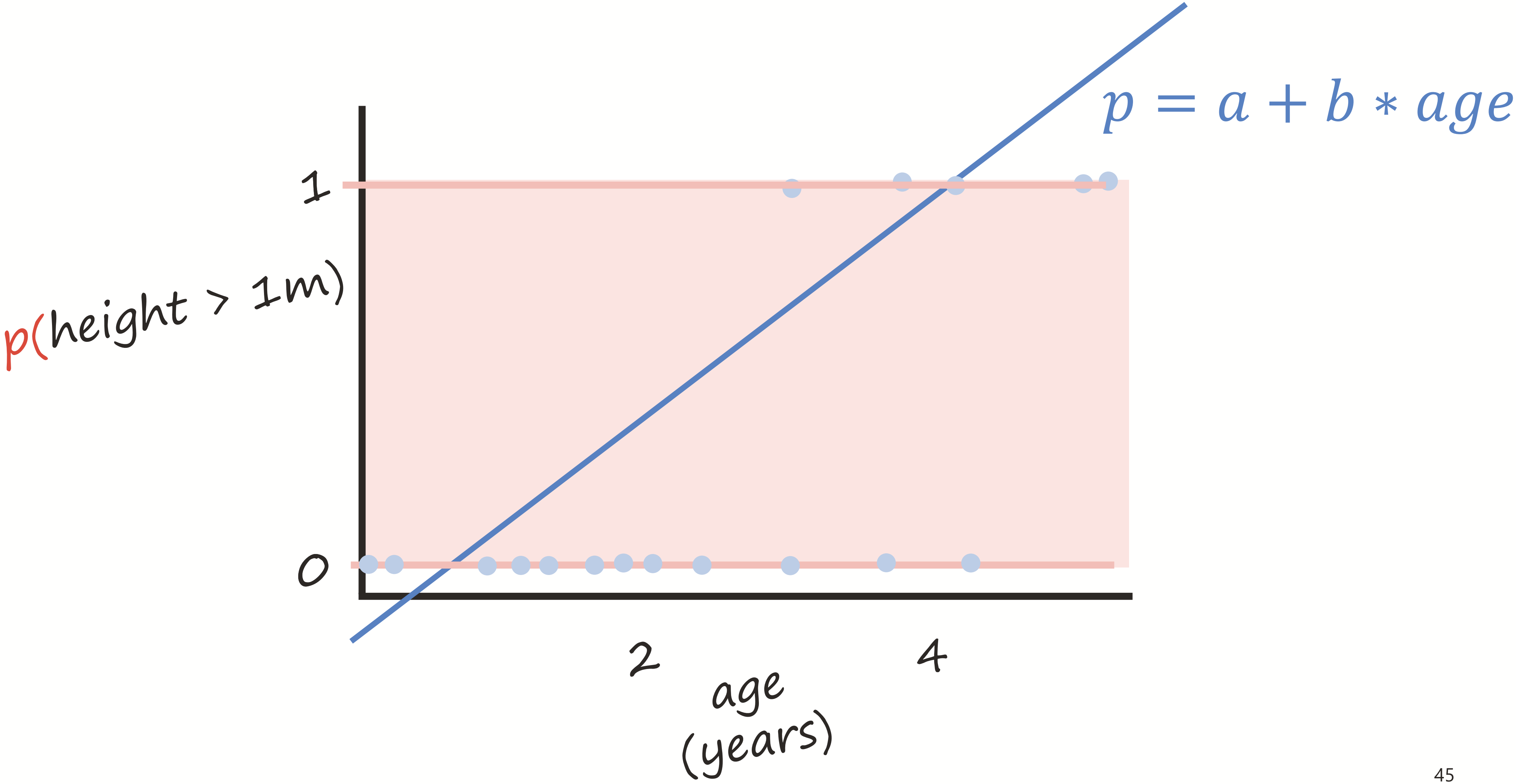

p(height > 1m)
1
0
0
2
4
age
(years)


height (m)
1 metre
0
2
4
age (years)
p(height > 1m)
1
0
0
2
4
age (years)

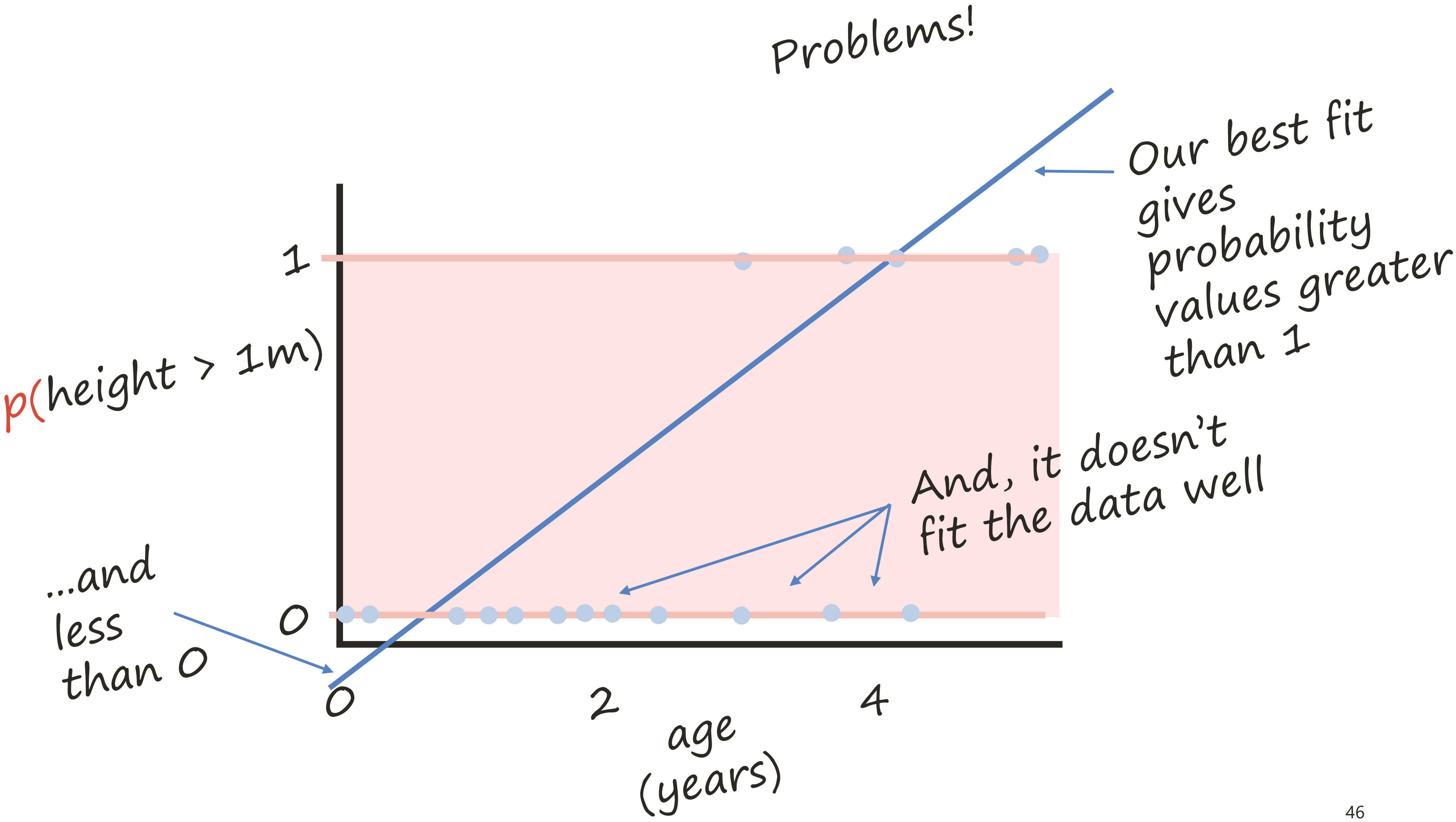

height
(m)
1 metre
0
2
4
age
(years)
1
0
p(height > 1m)
0
2
4
age
(years)

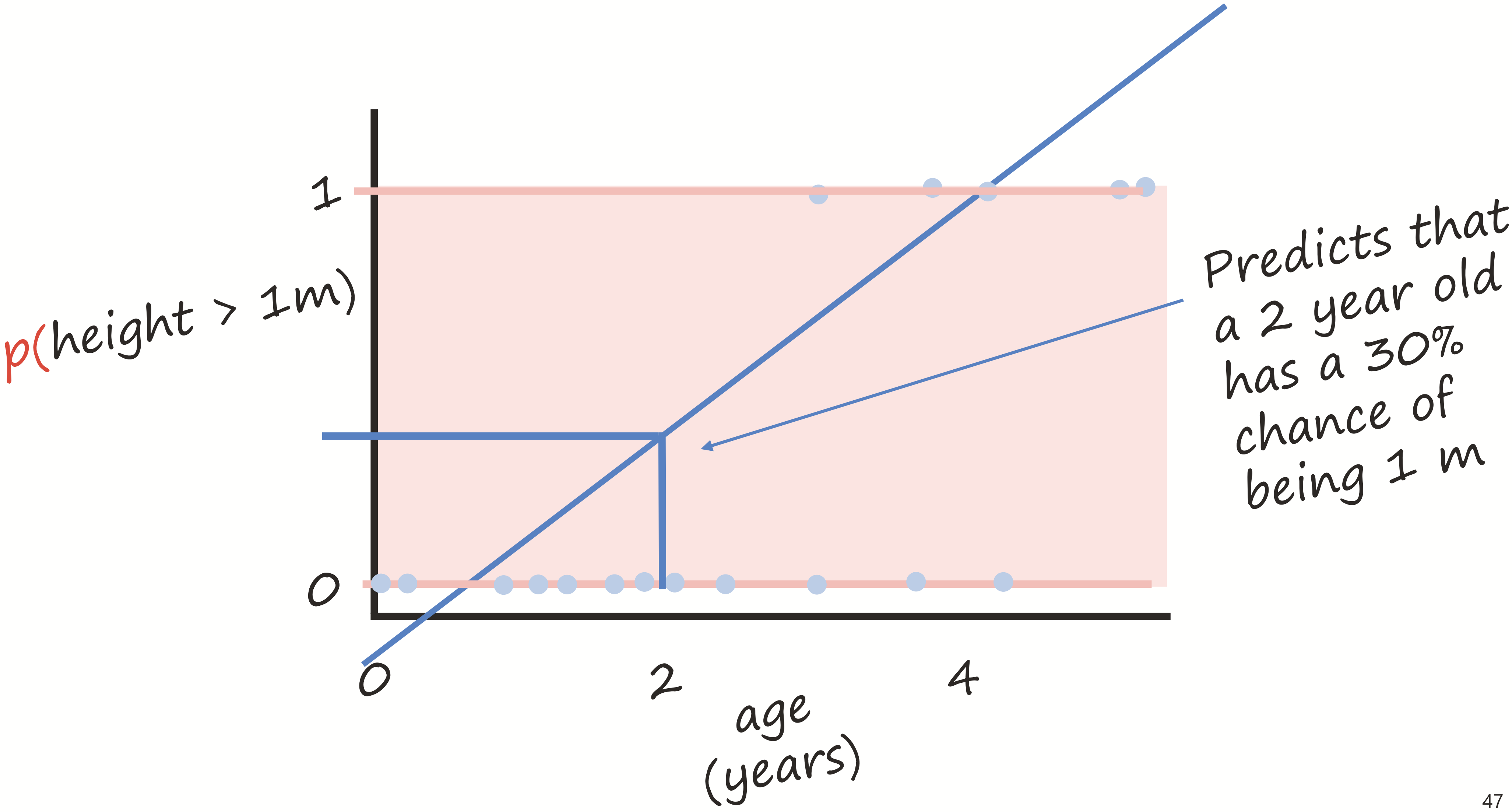
Will a linear model do?





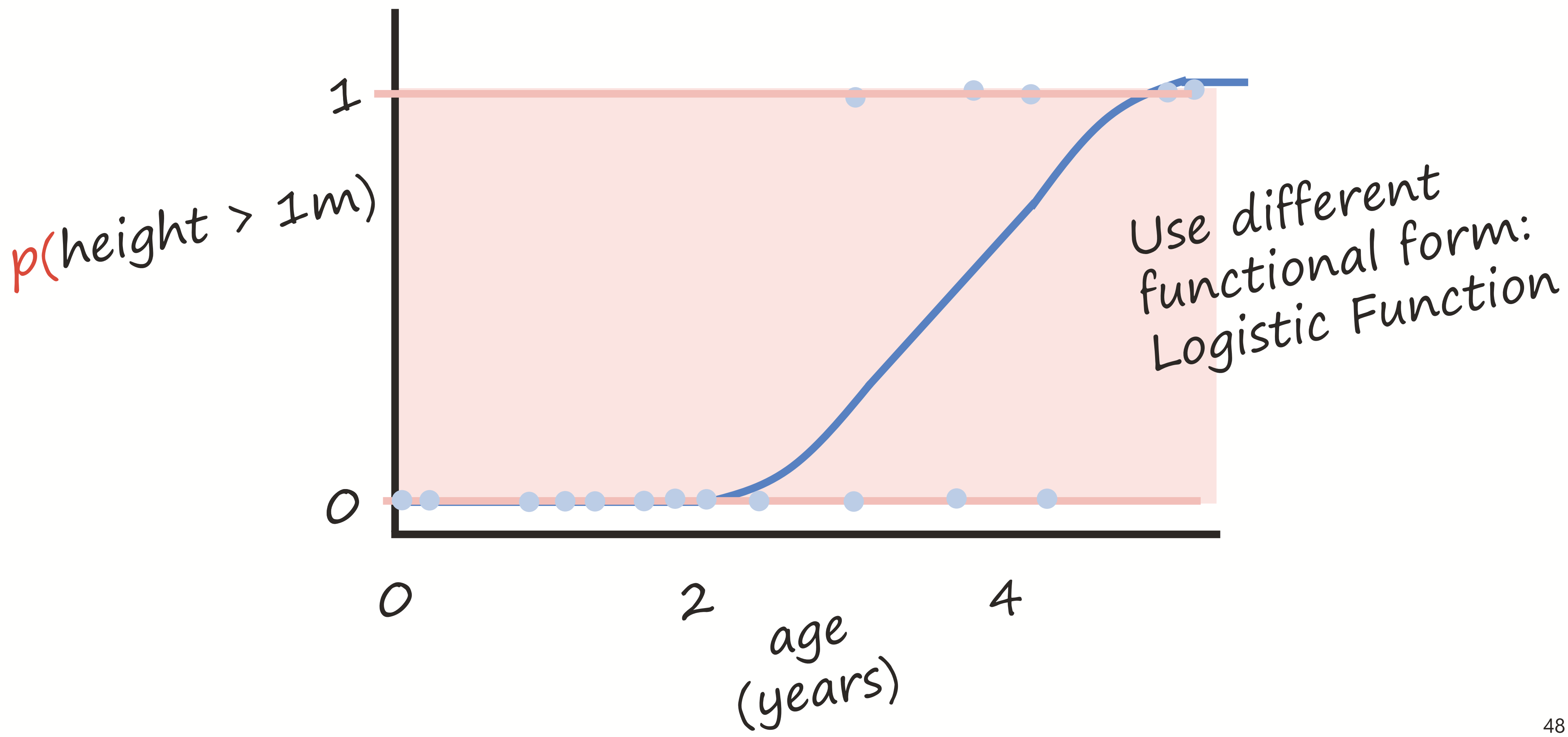
$p = a + b * age$
1
0
p(height > 1m)
2
4
age
(years)

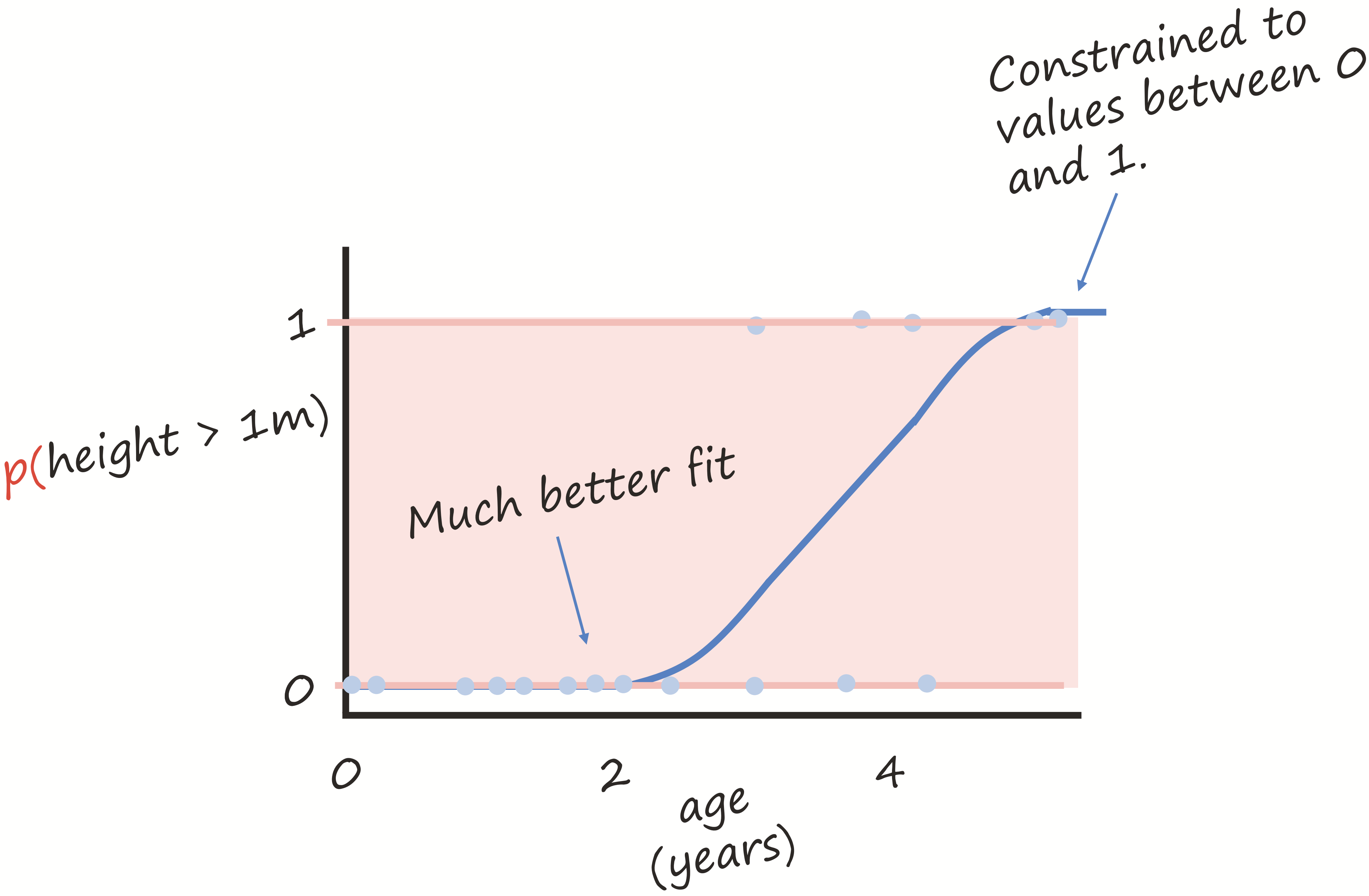

Problems!
Our best fit gives probability values greater than 1
p(height > 1m)
1
0
...and less than 0
And, it doesn't fit the data well
0
2
4
age (years)

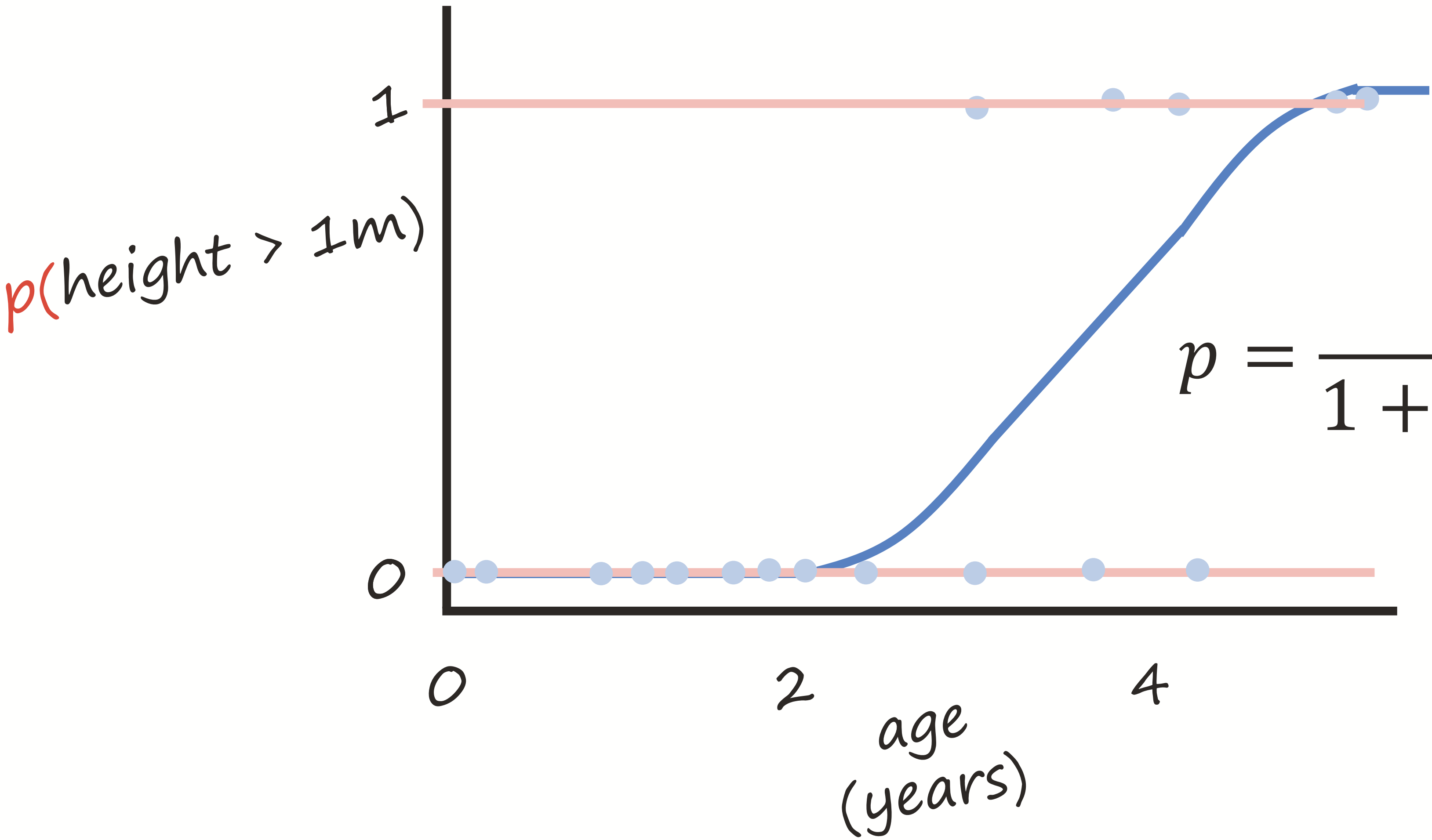


p(height > 1m)
1
0
0
2
4
age (years)
Predicts that a 2 year old has a 30% chance of being 1 m

Another strategy:




Constrained to values between 0 and 1.
p(height > 1m)
1
0
Much better fit
0
2
4
age (years)



$$p = \frac{1}{1 + e^{-\underbrace{(a+b*age)}_{\text{familiar form}}}}$$

$$y = a + b * x$$

$$= a + b * age$$

$$p = \frac{1}{1 + e^{-(a+b*age)}}$$

$$\log\left(\frac{p}{1-p}\right) = a + b * age$$

$$p = \frac{1}{1 + e^{-(a+b*age)}}$$

$$\overbrace{log\left(\underbrace{\frac{p}{1-p}}_{\text{odds}}\right)}^{\text{Logit function}} = a + b * age$$

a and b
in metres

$$height = 0.3 + 0.2 * age$$

$$log\left(\frac{p}{1-p}\right) = a + b * age$$

$$height = a + b * age$$

$$log\left(\frac{p}{1-p}\right) = a + b * age$$

a and b
in
"log odds"

Summary of maths

Binary outcome -> model p(outcome)

Linear model won't do!

Logit transformation:

$$log\left(\frac{p}{1-p}\right) = a + b * age$$

a, b … coefficients will be log odds

Building our model

Let's return to our MH question and dataset...

Our task – Help Trust X

Reduce Mental Health LoS

Predict stay > 60 days ?

Model formula

	id	gender	age	diag	…	long_stay
	1	F	72	dement		1

```
long_stay ~ age + sex + diag
```

Model formula

$$\underbrace{\texttt{long\_stay}}_{\textit{Response}} \sim \underbrace{\texttt{age + sex + diag}}_{\textit{Predictors}}$$

Build model on “training” data

Pseudo code (see script (pg. 2) for true code)

formula →

```
model(
    long_stay ~ age + sex + diag,
    type = “logistic regression”,
    data = training_sample
) -> our_model
```


PREDICTION
hot water
EXPLANATION

Explanatory models

```
summary(our_model)
```

summary(our_model)

Coefficient	Estimate	Std.err	z-value	Pr(>|z|)
(Intercept)	-5.19	2.45	-2.12	0.034
age	0.16	0.05	3.17	0.001
sex_m	0.18	0.82	0.22	0.824
diag_mood	-5.85	1.25	-4.68	<0.001
diag_neurosis	-25.10	-	-0.01	0.991

Dummy variables

Coefficient	Estimate
(Intercept)	-5.19
age	0.16
sex_m	0.18
diag_mood	-5.85
diag_neurosis	-25.10

Dummy variable

For categorical variable with 2 levels

From original dataset:

id	sex
1	F
2	M
3	F

Transformed to dummy variable:

sex_m
0
1
0

reference level is female

(male =0)

Dummy variable

For categorical variable with 3 levels

From original dataset:

id	diag
1	mood
2	neurosis
3	dementia

Dummy for mood

diag_mood
1
0
0

Dummy for neurosis

diag_neuros
0
1
0

Dummy variable

For categorical variable with 3 levels

From original dataset:

id	diag	diag_mood	diag_neuros
1	mood	1	0
2	neurosis	0	1
3	dementia	0	0

reference level is dementia (others =0)

`summary(our_model)`

Coefficient	Estimate	Std.err	z-value	Pr(>|z|)
(Intercept)	-5.19	2.45	-2.12	0.034
age	0.16	0.05	3.17	0.001
gender_m	0.18	0.82	0.22	0.824
diag_mood	-5.85	1.25	-4.68	<0.001
diag_neurosis	-25.10	-	-0.01	0.991

`summary(our_model)`

Coefficient	Estimate	Std.err	z-value	Pr(>|z|)
(Intercept)	-5.19	2.45	-2.12	0.034
age	0.16	0.05	3.17	0.001
gender_m	0.18	0.82	0.22	0.824
diag_mood	-5.85	1.25	-4.68	<0.001
diag_neurosis	-25.10	-	-0.01	0.991

Coefficient estimates are log odds

`summary(our_model)`

Coefficient	Estimate	Std.err	z-value	Pr(>|z|)
(Intercept)	-5.19	2.45	-2.12	0.034
age	0.16	0.05	3.17	0.001
gender_m	0.18	0.82	0.22	0.824
diag_mood	-5.85	1.25	-4.68	<0.001
diag_neurosis	-25.10	-	-0.01	0.991

Coefficient estimates are log odds → *exp(…) to get the odds (ratios)*

summary(our_model)

Coefficient	exp(estimate)	Std.err	z-value	Pr(>|z|)
(Intercept)	0.005	2.45	-2.12	0.034
age	1.17	0.05	3.17	0.001
gender_m	1.20	0.82	0.22	0.824
diag_mood	0.002	1.25	-4.68	<0.001
diag_neurosis	<0.001	-	-0.01	0.991

odds (ratios)

Coefficients as odds (ratios)

For continuous variables (e.g. age)

$\exp(0.16) = 1.17$

```
long_stay ~ … + 0.16*age + …
```

We can say that for a one unit (year) increase in age the <u>odds</u> of a long stay increase:

- by a factor of 1.17
- or by 17%

Same thing

Holding all other variables constant!

Coefficients as odds (ratios)

For categorical variables (e.g. gender)

exp(0.18) = 1.20

```
long_stay ~ … + 0.18*gender_m +…
```

We can say that males are 1.2 times more likely to have a long stay than the reference level (females).

Holding all other variables constant!

Known as odds ratios

Coefficients as odds (ratios)

For diagnosis

exp(-5.85) = 0.002

```
long_stay ~ … -5.85*diag_mood +…
```

We can say that those admitted with mood affective disorders are 500 times less likely to have a long stay than the reference level (dementia).

It's possible to change the reference level

Known as <u>odds ratios</u>

Holding all other variables constant!


PREDICTION
hot water
EXPLANATION

How do we predict?

Test data set:

id	…	long_stay
81	…	1
82	…	0
83	…	1

How do we predict?

```
predict(our_model, data = test_set)
```

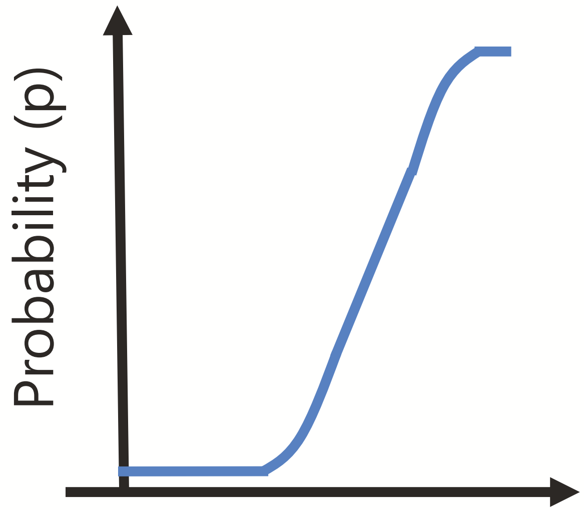
id	…	long_stay
81	…	1
82	…	0
83	…	1

Test data set:

How do we predict?

```
predict(our_model, data = test_set)
```

id	...	long_stay	.prob
81	...	1	0.61
82	...	0	0.45
83	...	1	0.47



Model will give us probabilities of stay > 60 days

How do we predict?

```
predict(our_model, data = test_set)
```

id	…	long_stay	.prob	predicted
81	…	1	0.61	1
82	…	0	0.45	0
83	…	1	0.47	0

By default, we use a threshold probability of 0.5 – everything ≥ 0.5 is predicted as long stay

So, how might we report model performance to Trust X?

(long_stay)

Observed

0	1
24	25

The confusion matrix

	Observed	
Predicted	0	1
0	20	7
1	4	18

On the diagonal we have correct predictions

Related performance metrics

Accuracy

The proportion of correctly predicted observations

Predicted	Observed 0	Observed 1
0	20	7
1	4	18

$$\frac{\text{Correctly classified: } 20 + 18}{\text{Total observations:}} = \frac{38}{49} = 0.78$$

Caution

Accuracy can be misleading (esp. with class imbalance)

	Observed	
	0	1
	24	2

Model predicts all zeros:

$$\frac{\text{Correctly classified: } 24}{\text{Total observations: } 26} = 0.92$$

Sensitivity

The proportion of long stay patients correctly predicted in the long stay group

How sensitive is our model to any long stay "signal" in the data?

Sensitivity

The proportion of long stay patients correctly predicted in the long stay group

	Observed	
Predicted	0	1
0	20	7
1	4	18

$$\frac{\text{Correctly classified as long\_stay:}}{\text{Total number of long\_stays:}} \quad \frac{18}{25} = 0.72$$

Specificity

The proportion of short stay patients correctly predicted in the short stay group

How well does our model avoid mis-classifying short stay patients as long stays?

Specificity

The proportion of short stay patients correctly predicted in the short stay group

	Observed	
Predicted	0	1
0	20	7
1	4	18

Correctly classified as short_stay:
Total number of short_stays:

$$\frac{20}{24} = 0.83$$

Adjusting Sensitivity/Specificity

id	…	long_stay	.prob	predicted
81	…	1	0.61	1
82	…	0	0.45	0
83	…	1	0.47	0

By default, a probability threshold ≥ 0.5
is used to predict a long stay

Adjusting Sensitivity/Specificity

Correctly classified as long_stay: 2

Total number of long_stays: 2

$$\frac{2}{2}$$

id	...	long_stay	.prob	predicted
81	...	1	0.61	1
82	...	0	0.45	1
83	...	1	0.47	1

But if we lower the threshold to 0.4...
we can increase our model sensitivity

Adjusting Sensitivity/Specificity

Correctly classified as short_stay: 0

Total number of short_stays: 1

$$\frac{0}{1}$$

id	…	long_stay	.prob	predicted
81	…	1	0.61	1
82	…	0	0.45	1
83	…	1	0.47	1

But if we lower the threshold to 0.4…
we can increase our model sensitivity
(this is often at the cost of specificity)

Sensitivity / Specificity

Whether this trade-off is worthwhile depends on context.

The ROC plot

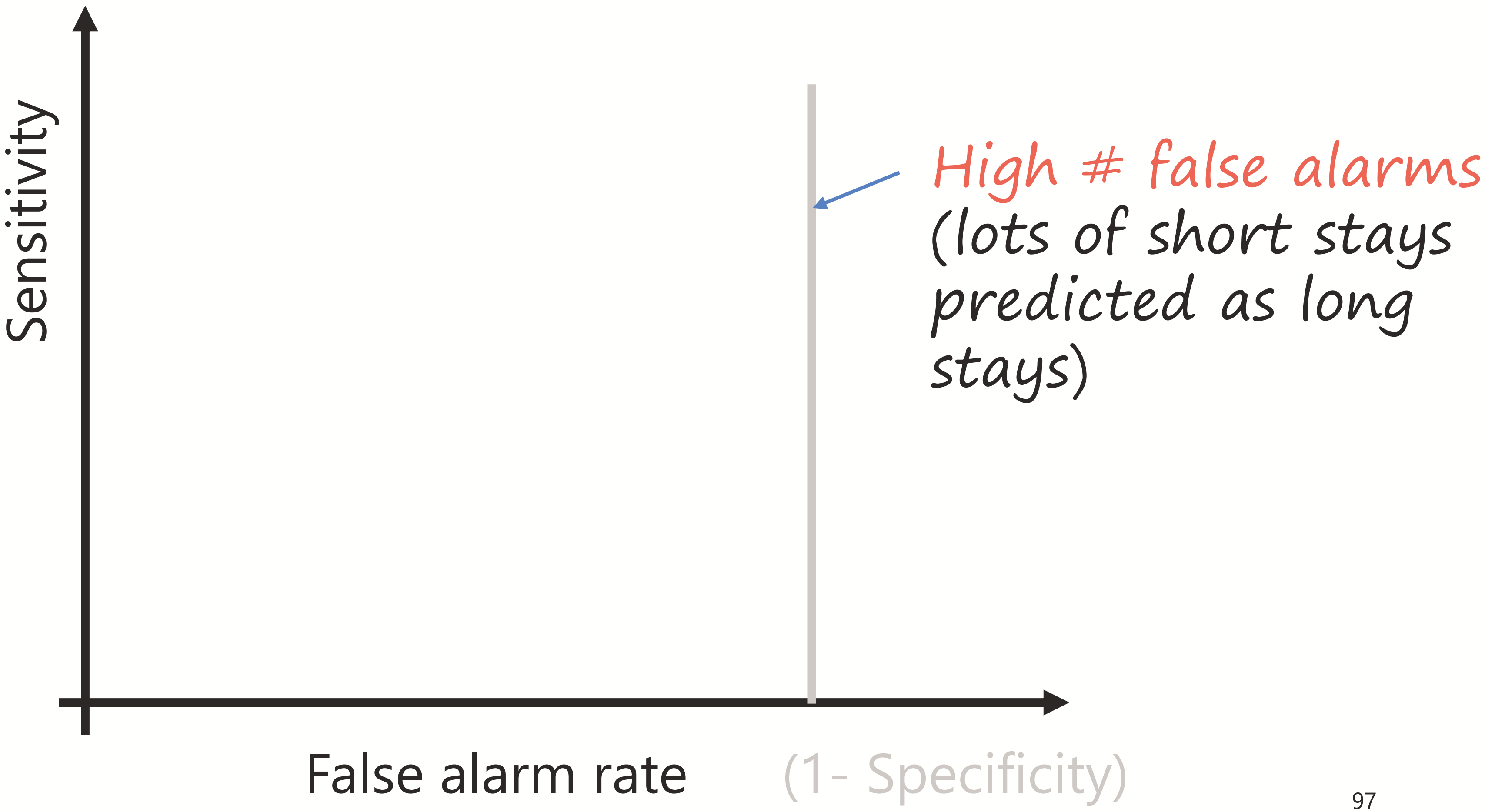
(Receiver operating characteristics)

The ROC plot

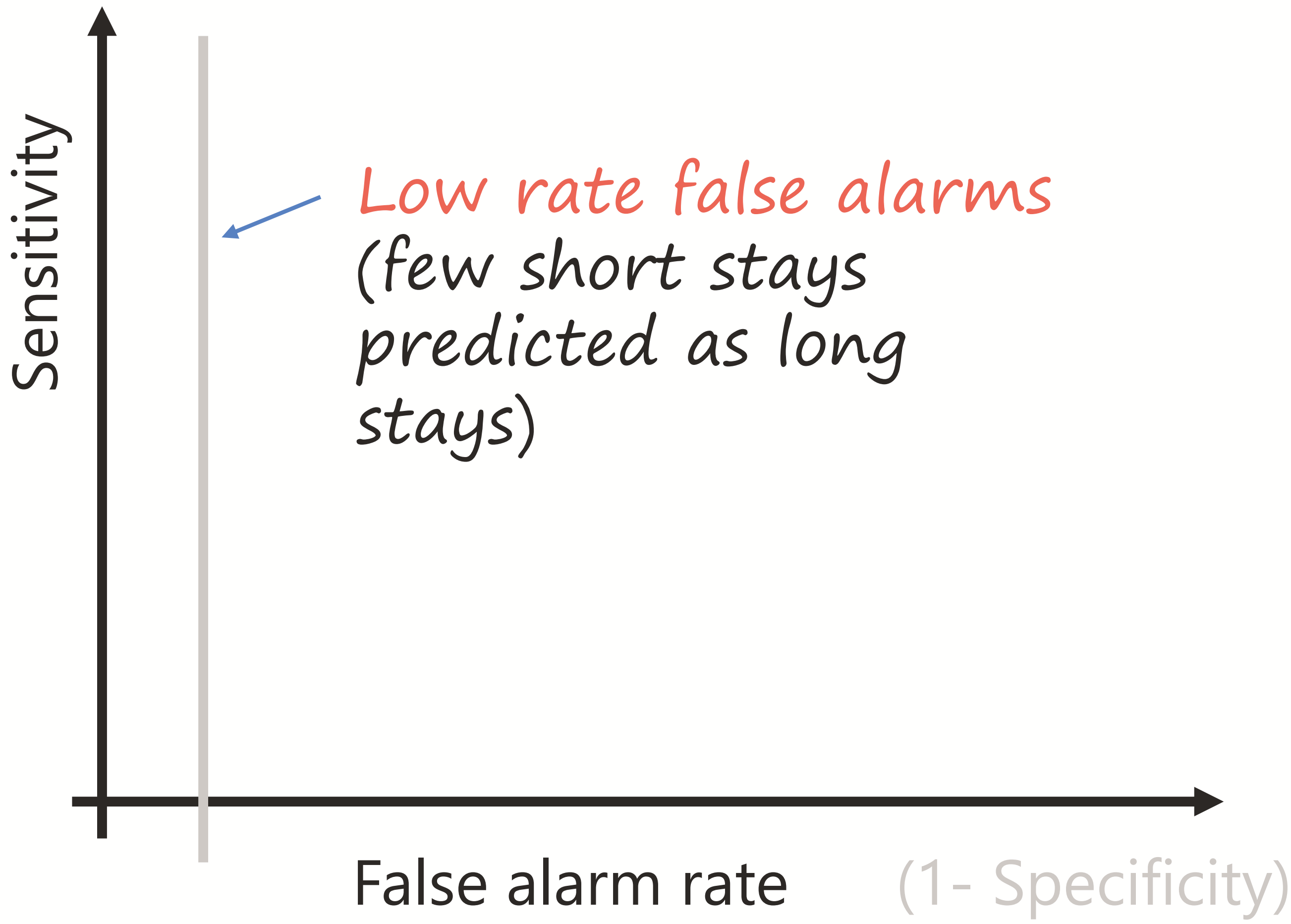
Considers <u>all thresholds</u> and tracks changes in sensitivity and specificity

1. Determine most appropriate threshold
2. Assessment without having to determine best threshold

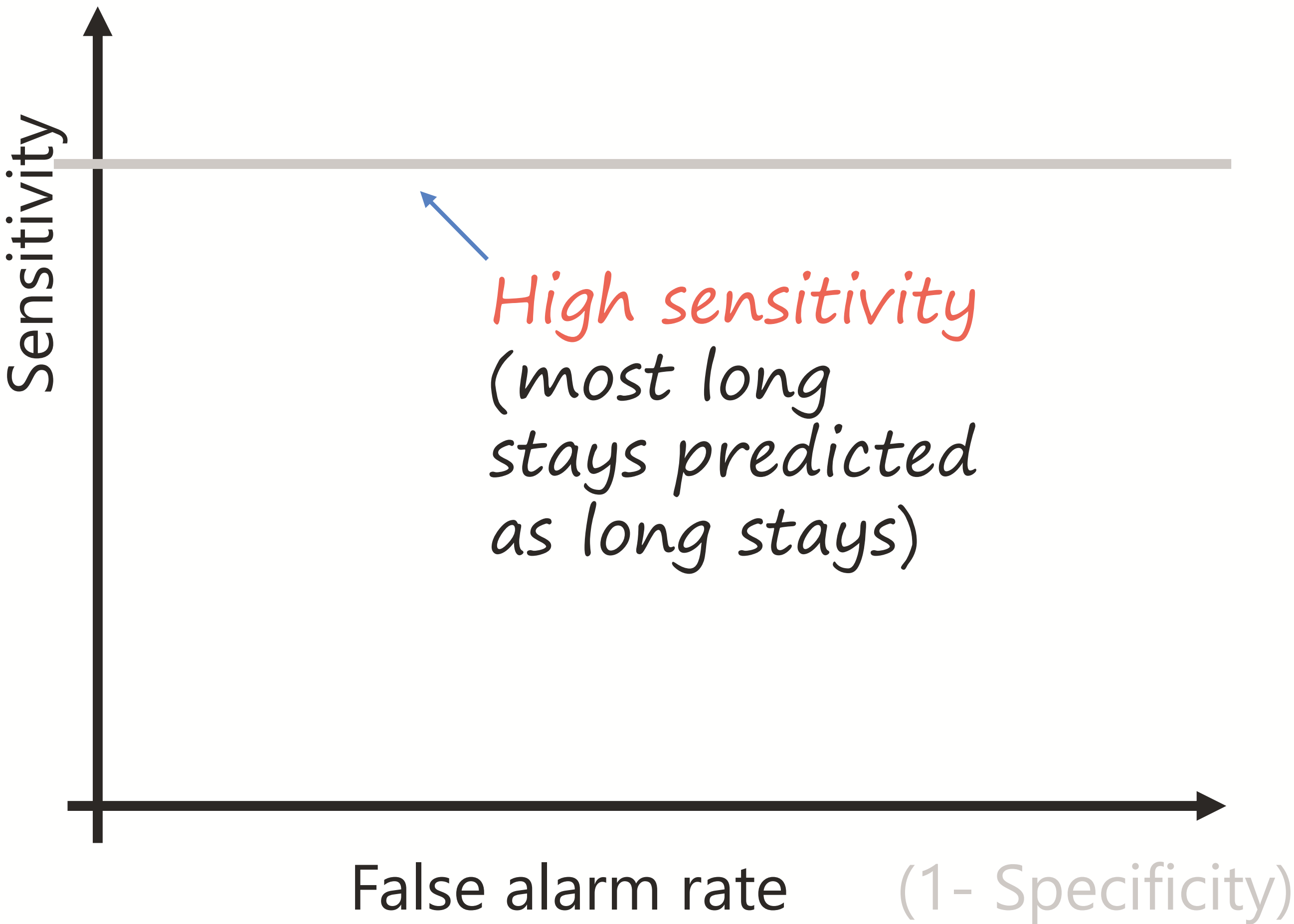
The ROC plot


Sensitivity
High # false alarms
(lots of short stays predicted as long stays)
False alarm rate
(1- Specificity)

The ROC plot

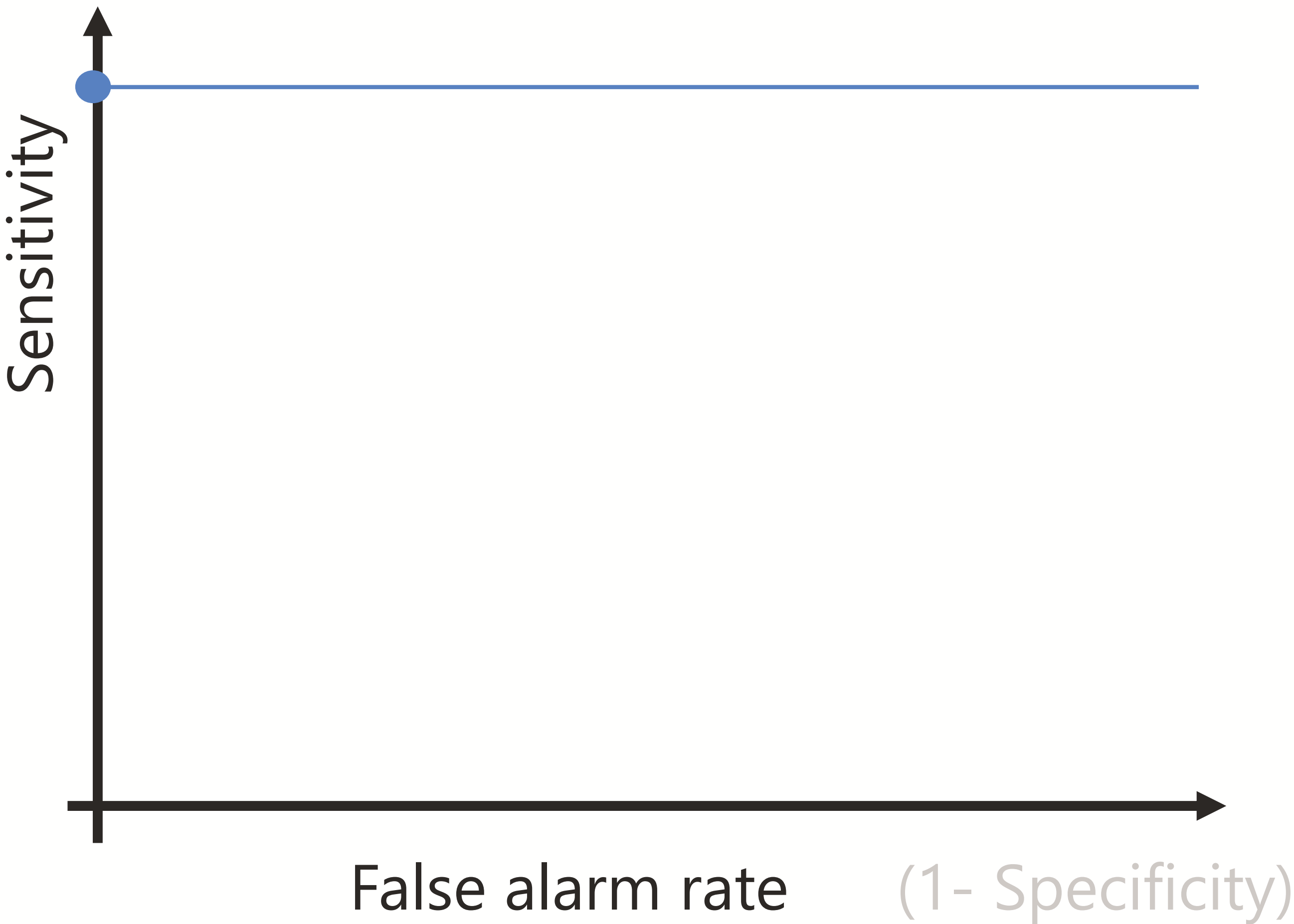

Sensitivity
Low rate false alarms
(few short stays predicted as long stays)
False alarm rate
(1- Specificity)

The ROC plot


Sensitivity
High sensitivity
(most long stays predicted as long stays)
False alarm rate
(1- Specificity)

The ROC curve

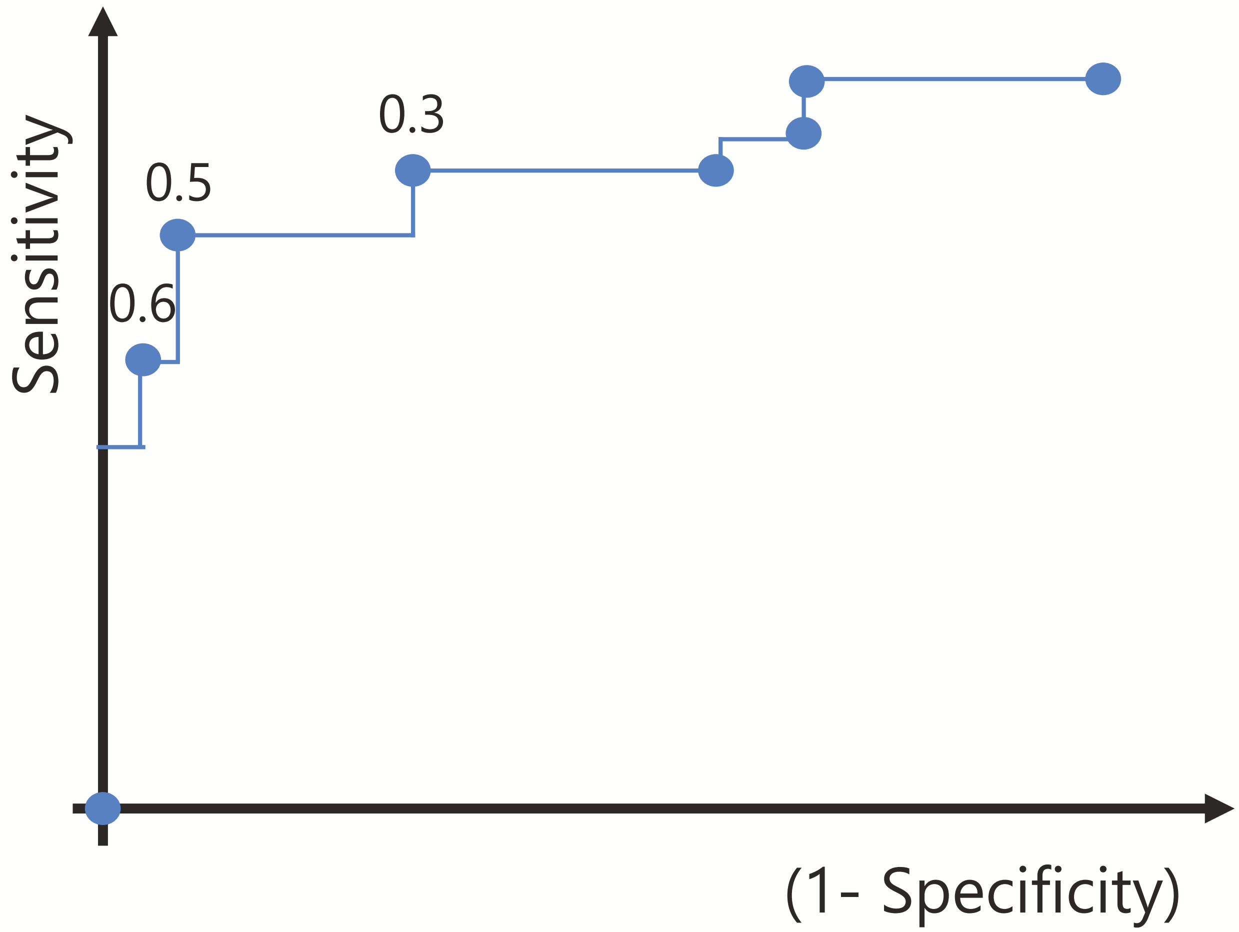
A "perfect" model



long_stay	.prob	predicted
1	0.61	1
0	0.45	1
1	0.47	1

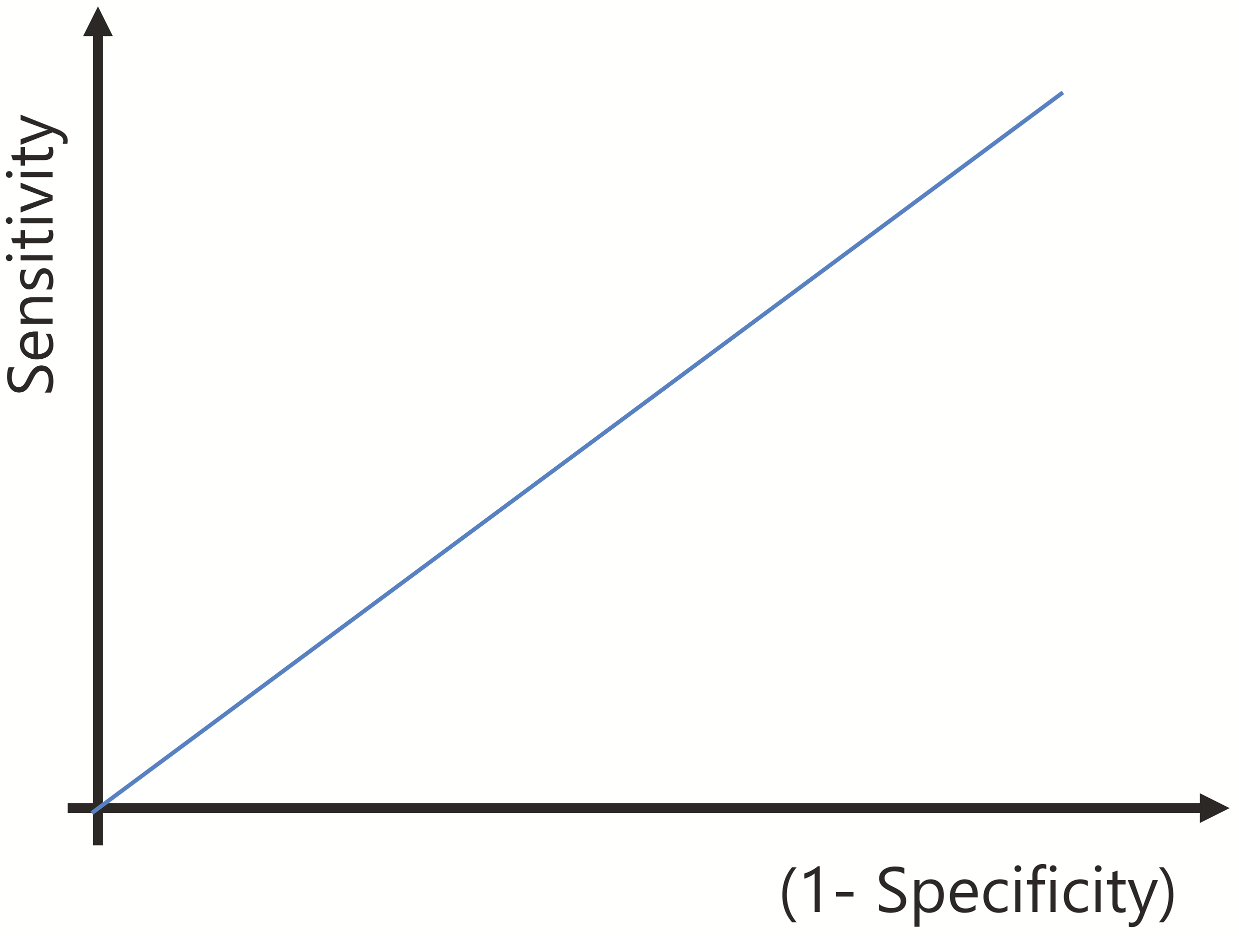
1. Observe how a particular threshold changes predictions
2. from predictions -> sensitivity/specificity metrics
3. Plot metrics on ROC plot

The ROC curve



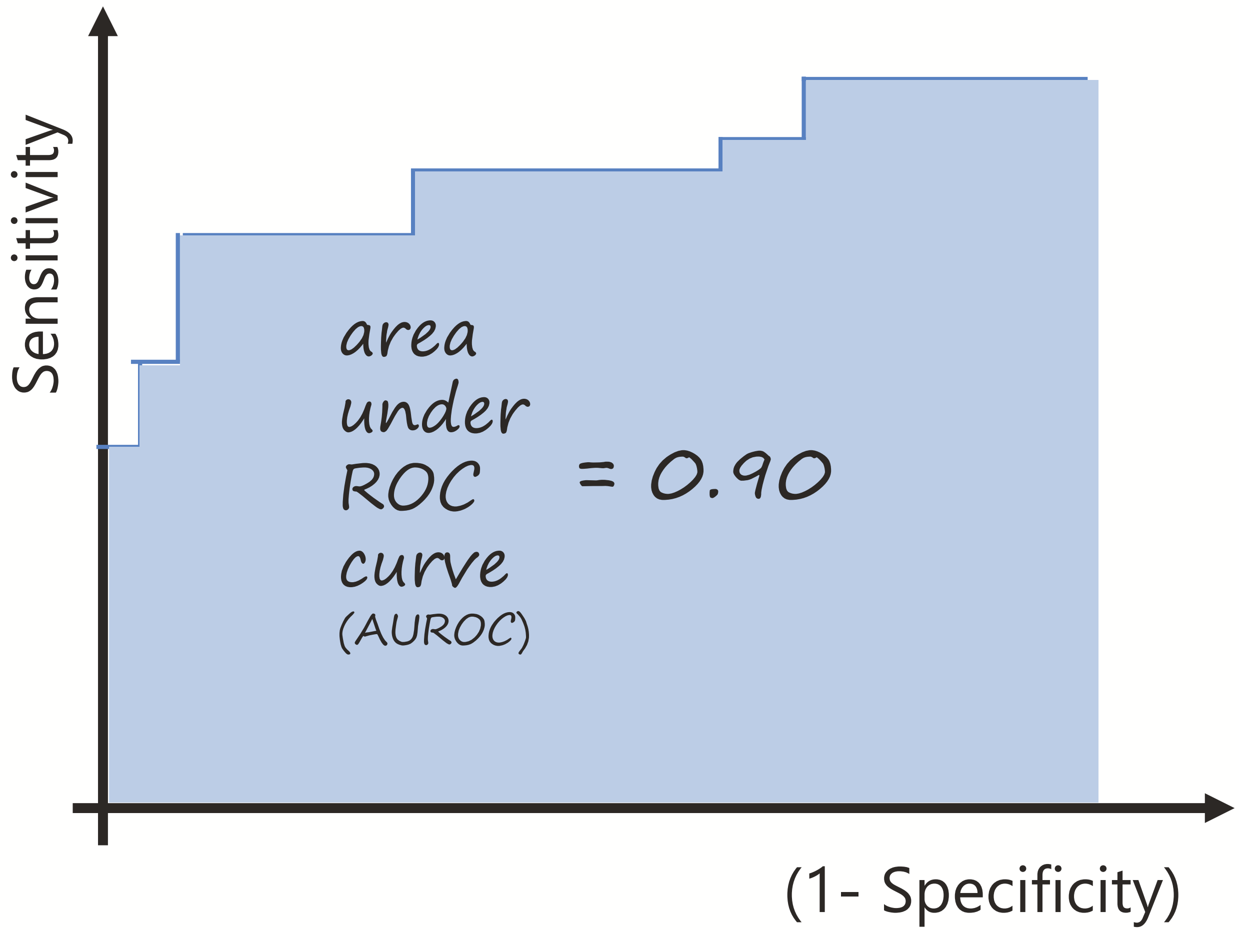
The ROC curve

A model that does no better than chance



Area under ROC curve

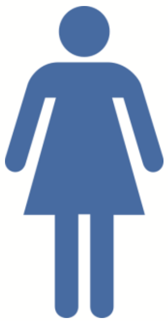
We can summarise curve in one metric



Our Report to Trust X

Accuracy, Sensitivity / Specificity....

but also specific examples:



Female, 36 years old, Mood (affective) disorder

Observed LoS: 41 days

Model probability of long stay: **0.4%** [0, 2]

End