Risk Prediction Modelling

Introduction to Data Science for Health and Social Care

(Week 9 – 23/11/2022)

Kevin Tsang

THE UNIVERSITY *of* EDINBURGH | usher institute

Risk Prediction Modelling

- Why Use Risk Prediction Modelling
- Training-Validation-Testing
- Real-World Example: Death From Injury
  - Model: Logistic Regression
  - Performance Metric
- Summary

Why Use Risk Prediction Modelling

- Can help **guide** decisions
- Make data-driven decisions
- Can predict future events based on historic patterns

Training-Validation-Testing

Training	Validation	Testing
Fit model to the data.	Assess the model fitted to the training data. Refine model where needed and retrain the model if needed.	Evaluate the model on completely unseen data. Ideally, models are tested only once, and no changes are made at this stage.

Steps to Build and Test Model

Build:

1. Pick predictor variables and target variable
2. Train model using training set
3. Evaluate and optimise model using validation set

Test:

4. Test model with unseen testing set

Someone who is **70 years-old** has just died from injury,
what is the probability that their death resulted from a fall?

Real-World Example: Death From Injury

Total Patients:	88,670
Years:	2012 – 2021
Location:	Scotland
Source:	Public Health Scotland

Reference: https://www.opendata.nhs.scot/dataset/unintentional-injuries

Steps to Build and Test Model

Build:

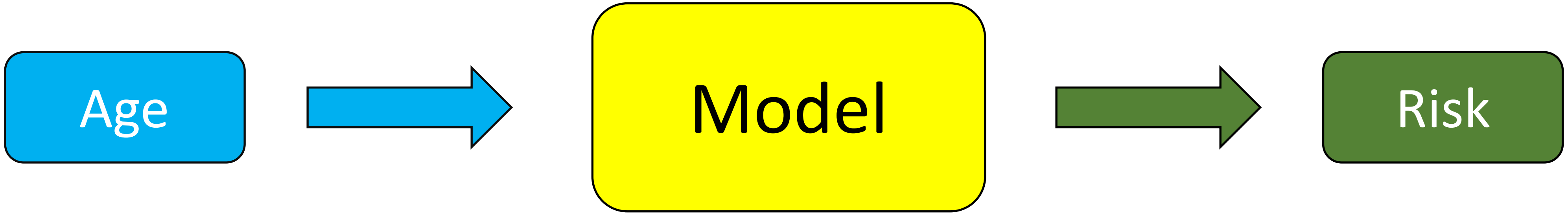
1. Pick predictor variables and target variable
2. Train model using training set
3. Evaluate and optimise model using validation set

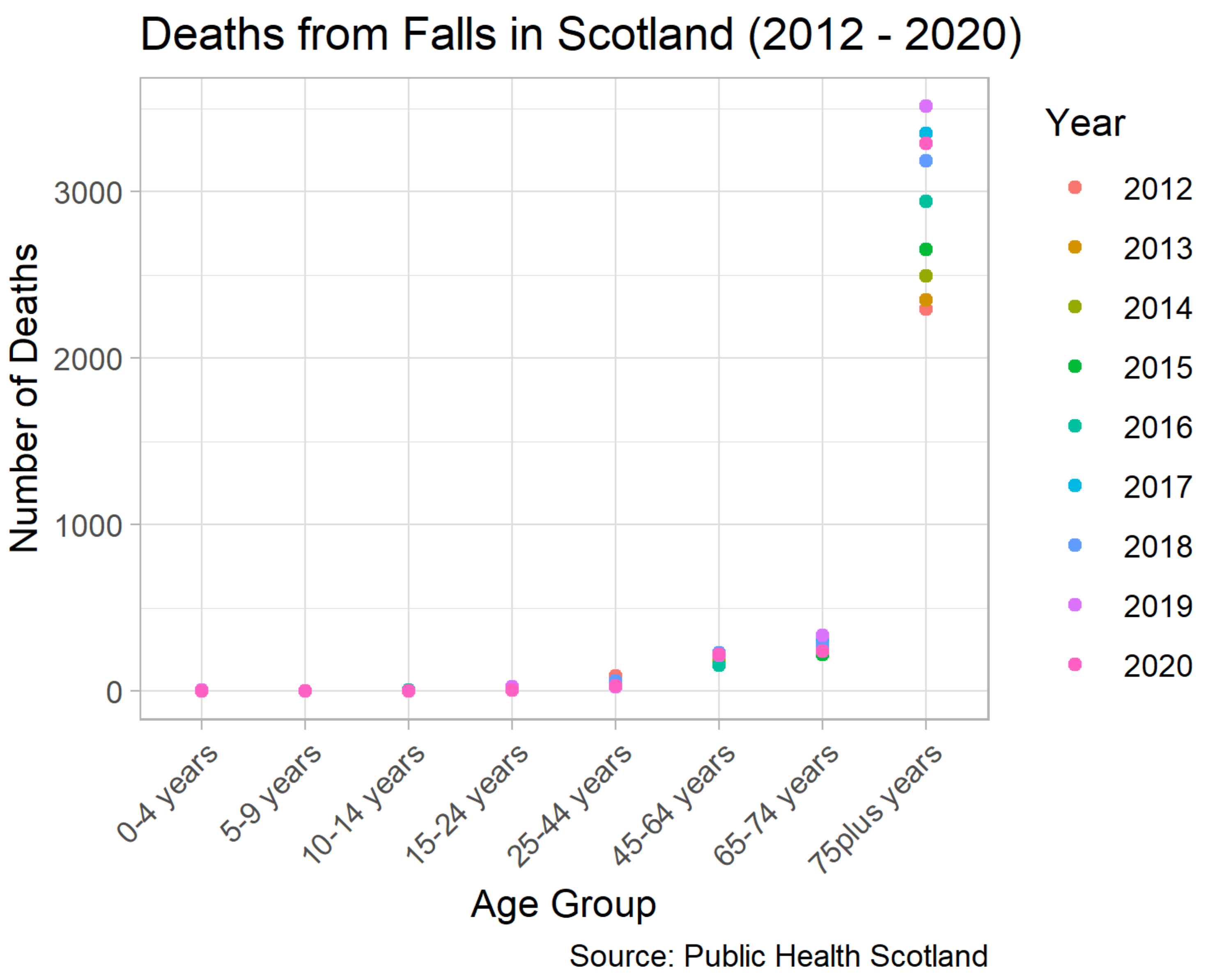
Test:

4. Test model with unseen testing set

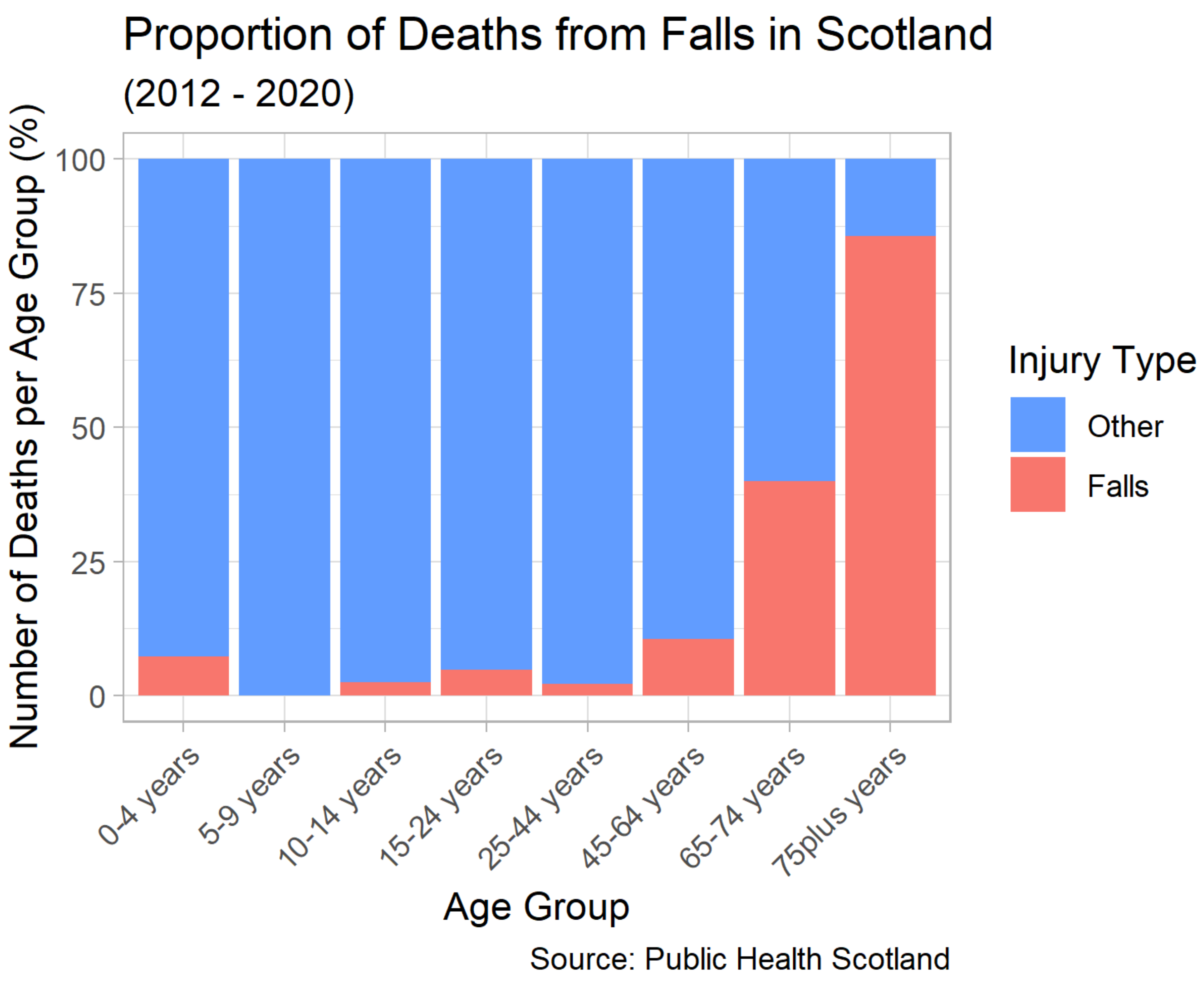
A Simple Model

- Predictor variables (input): Age Range
- Target variable (output): Risk of Death from Fall
- Model: Logistic Regression





Deaths from Falls in Scotland (2012 - 2020)
Number of Deaths
0
1000
2000
3000
0-4 years
5-9 years
10-14 years
15-24 years
25-44 years
45-64 years
65-74 years
75plus years
Age Group
Year
2012
2013
2014
2015
2016
2017
2018
2019
2020
Source: Public Health Scotland



Proportion of Deaths from Falls in Scotland
(2012 - 2020)
Number of Deaths per Age Group (%)
100
75
50
25
0
0-4 years
5-9 years
10-14 years
15-24 years
25-44 years
45-64 years
65-74 years
75plus years
Age Group
Injury Type
Other
Falls
Source: Public Health Scotland

Steps to Build and Test Model

Build:

1. Pick predictor variables and target variable
2. Train model using training set
3. Evaluate and optimise model using validation set

Test:

4. Test model with unseen testing set

Logistic Regression

Target ~ Predictors

(~) can be read as "modelled by"

In our case:

Risk of death by fall ~ age

Logistic Regression (Equations)

Target = Sigmoid(Intercept + Coefficient 1 × Predictor 1 + Coefficient 2 × Predictor 2 + …)

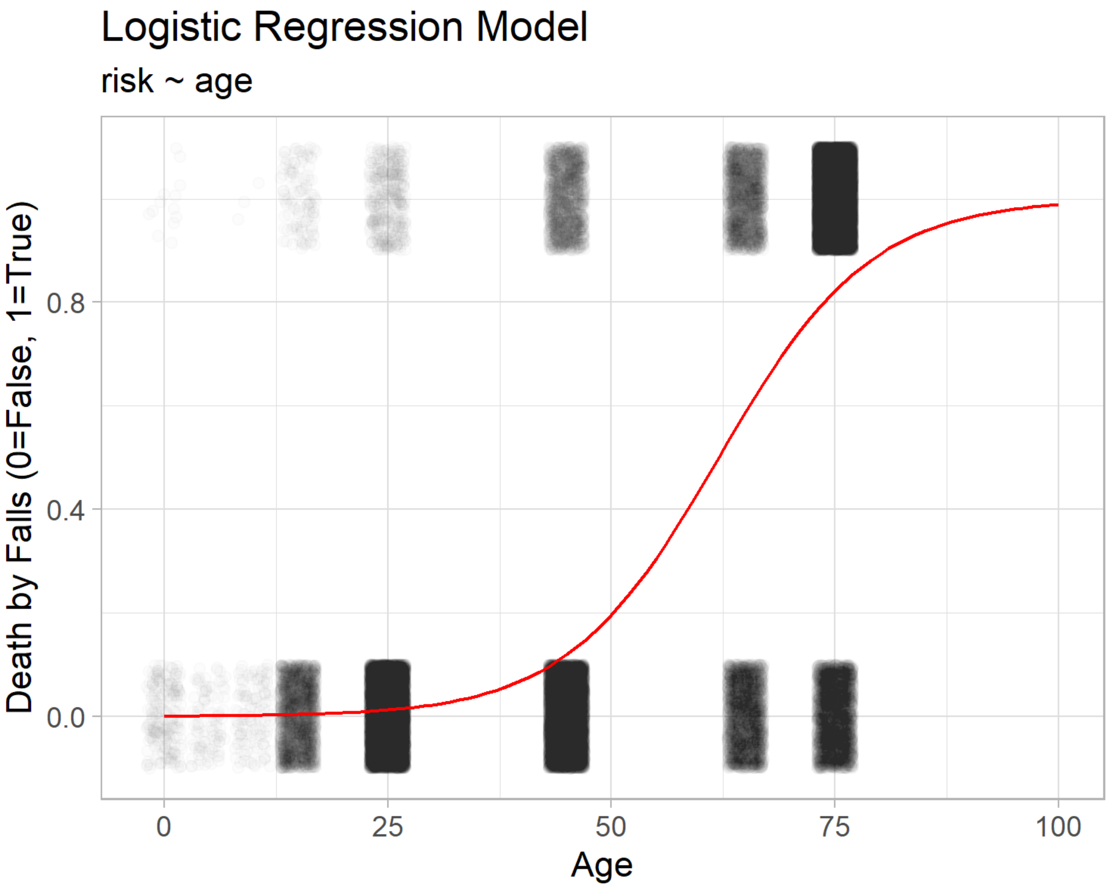
Sigmoid(x) = $\frac{1}{1+\exp(-x)}$, this is a S-shaped curve.

When we **fit** or **train** a model to the data, the Intercept and Coefficients are appropriately chosen

In our case:

Risk of death by fall = Sigmoid(Intercept + Coefficient 1 × age)

Logistic Regression Fit



Our Model:

Risk of death by fall =
Sigmoid(Intercept + Coefficient 1 × age)

Fitted Model:

Risk of death by fall =
Sigmoid(**-7.29** + **0.12** × age)

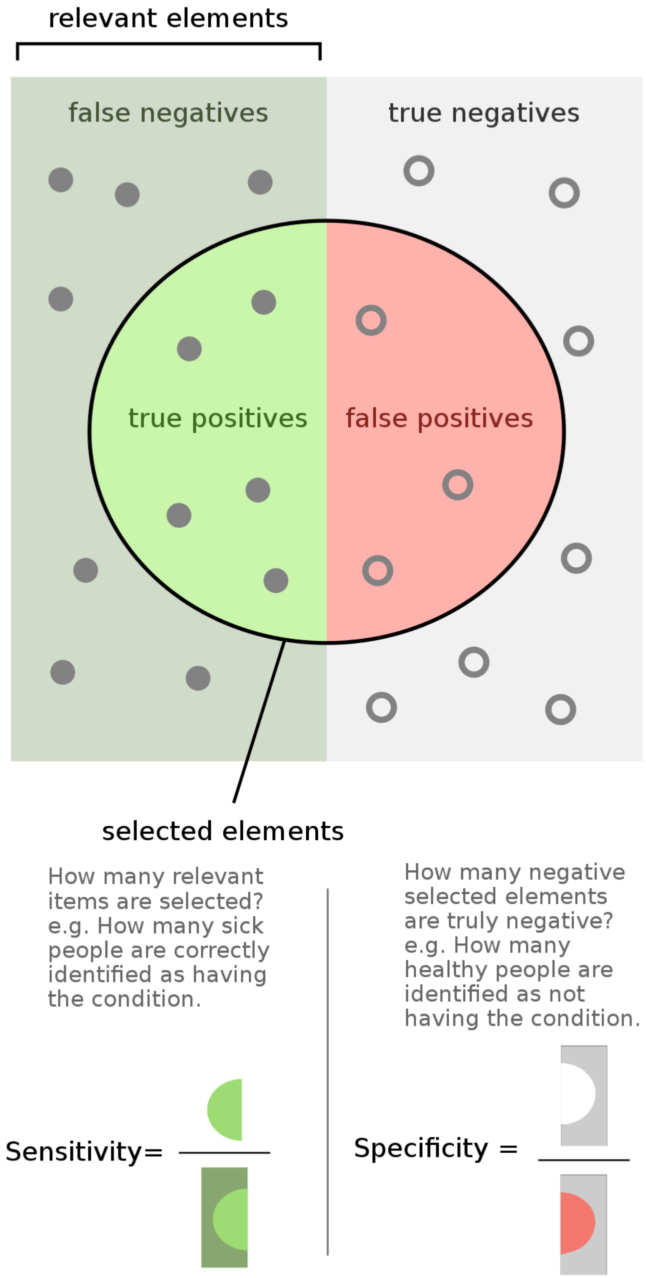
Steps to Build and Test Model

Build:

1. Pick predictor variables and target variable
2. Train model using training set
3. Evaluate and optimise model using validation set

Test:

4. Test model with unseen testing set



Source: https://en.wikipedia.org/wiki/
Sensitivity_and_specificity

Performance Metric

Sensitivity: the measure of how many <u>positive</u> outcomes were correctly identified

Number of predicted deaths by falls ÷ actual deaths by falls

Specificity: the measure of how many <u>negative</u> results were correctly identified

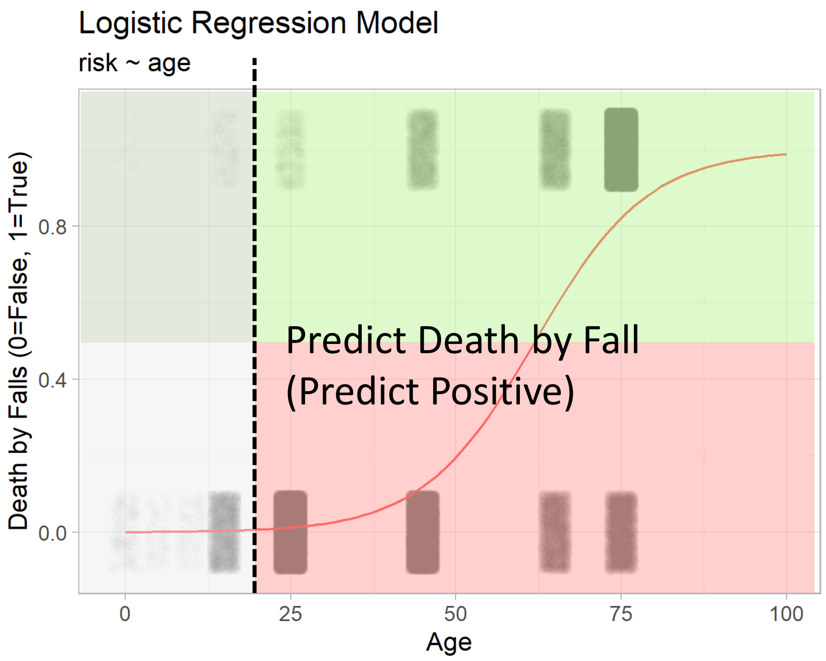
Number of predicted deaths by other injuries ÷ actual deaths by other injuries

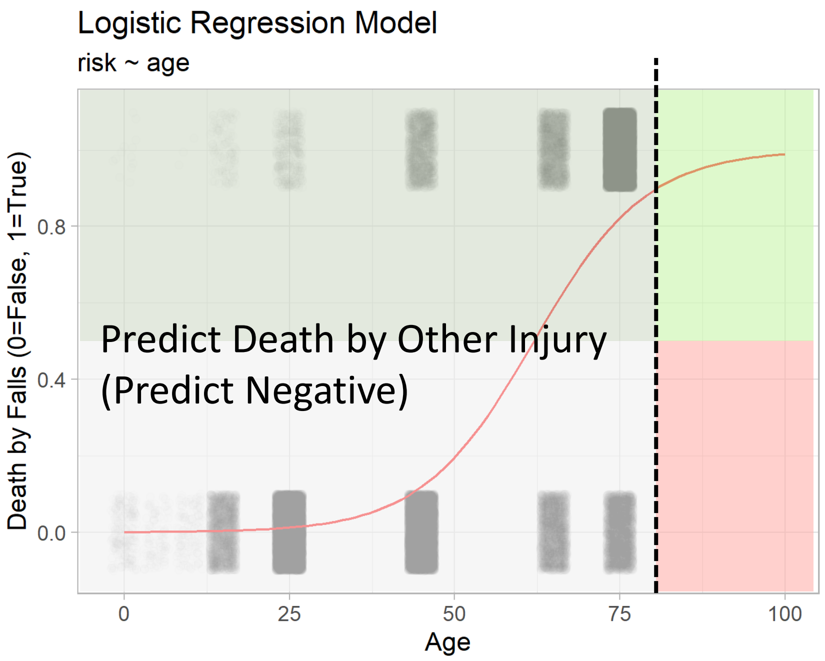
Confusion Matrix

		Ground truth	
		Positive (death by fall injury)	Negative (death by other injury)
Prediction	Positive (fall injury predicted)	True Positive (predicted fall and died by fall)	False Positive (predicted fall but died by other injury)
	Negative (other injury predicted)	False Negative (predicted other injury but died by fall)	True Negative (predicted other injury and died by other injury)

Sensitivity = $\frac{TP}{TP + FN}$ Specificity = $\frac{TN}{FP + TN}$

Thresholds and Decision Boundaries





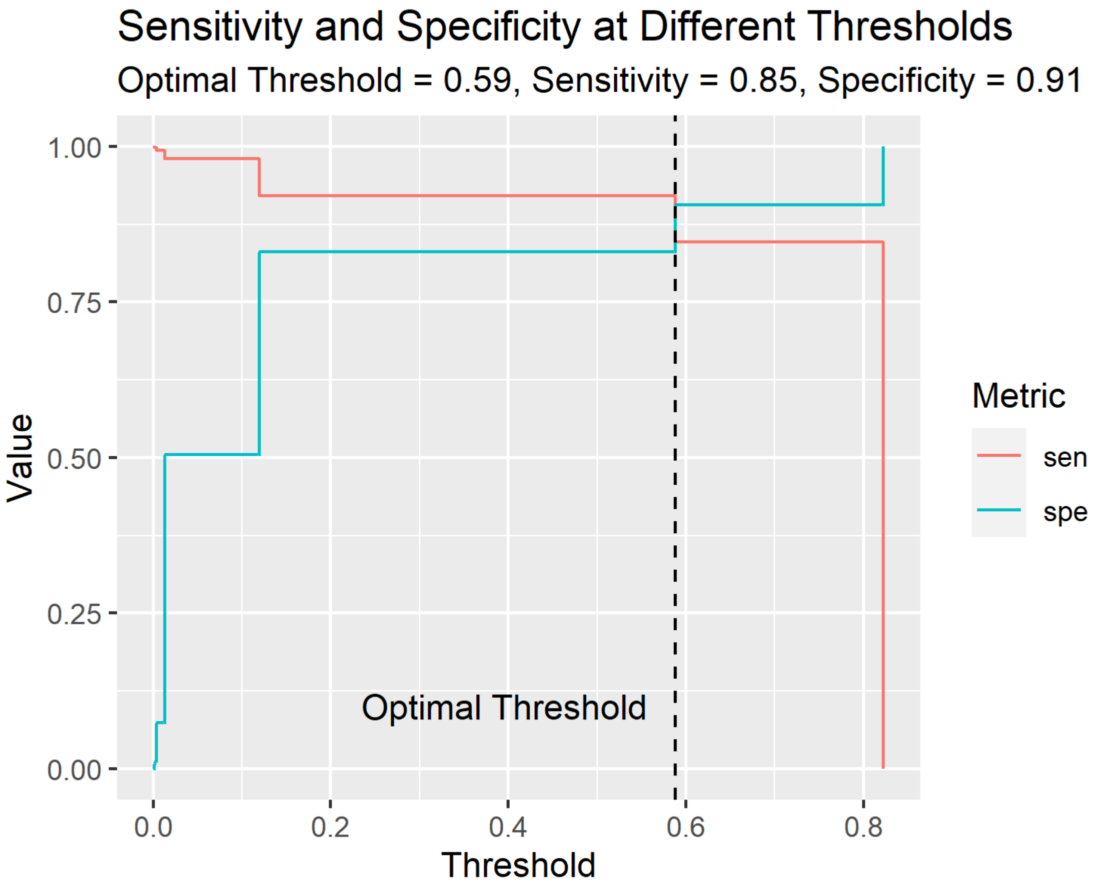
High Sensitivity (True Positive Rate)
Low Specificity (True Negative Rate)

Low Sensitivity (True Positive Rate)
High Specificity (True Negative Rate)

Pick The “Best” Threshold

There is no single “best” threshold

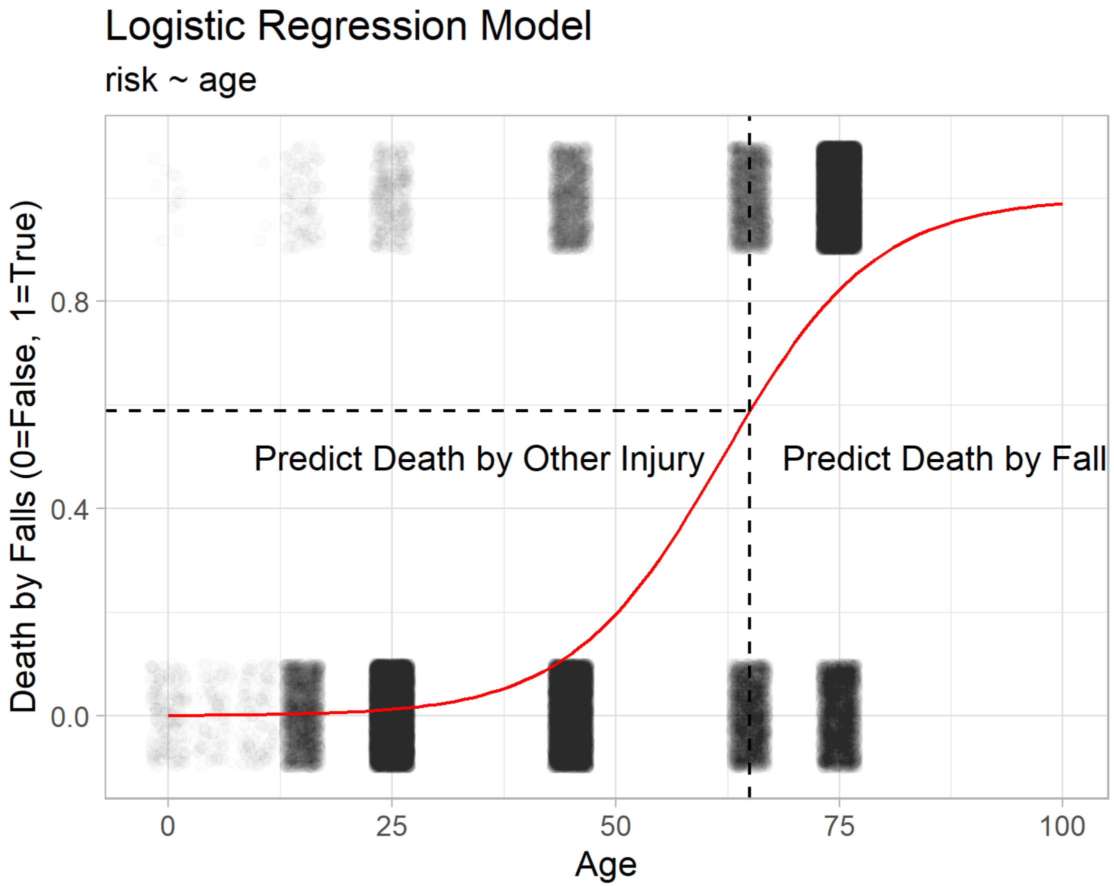
… using the balance of sensitivity and specificity is one method to define an optimal threshold



Optimised Model

Using the optimal threshold 0.59, is equivalent to an age threshold of **65 years-old**.

Model interpretation: Death by other injury is more likely when the patient is younger than **65 years-old**



Steps to Build and Test Model

Build:

1. Pick predictor variables and target variable
2. Train model using training set
3. Evaluate and optimise model using validation set

Test:

4. Test model with unseen testing set

Test On Unseen Data (Year 2021)

Model: If the patient is or older than 65 years-old, then predict death by fall injury. Otherwise, predict death by other injury.

Sensitivity = 0.93

Specificity = 0.85

		Ground truth	
		Positive (death by fall injury)	Negative (death by other injury)
Prediction	Positive (fall injury predicted) (Age ≥ 65 years-old)	**3652** True Positive	**992** False Positive
	Negative (other injury predicted) (Age < 65 years-old)	**256** False Negative	**5786** True Negative

$$\text{Sensitivity} = \frac{3652}{3652 + 256} \qquad \text{Specificity} = \frac{5786}{992 + 5786}$$

Someone who is **70 years-old** has just died from injury,
what is the probability that their death resulted from a fall?

According to Our Model

$$\mathbf{75\%} = \frac{1}{1+\exp(-(-7.29 + 0.12 \times 70 \text{ years old}))}$$

Using the model, the patient would be predicted to have died from a fall injury.

Summary

- Risk prediction models can help guide decisions using data
- Models are trained by appropriately choosing parameters
- Model performance can be assessed using sensitivity and specificity