Least Squares for Predicting Continuous Responses

Dr Rebecca Barter

Labeled data

Outcome of interest: Healthcare expenses in next year

Age	Sex	Weight	Diabetes	Healthare expenses
54	M	132	N	844
76	F	155	Y	5,467
49	M	166	Y	8,089
39	F	129	N	103
47	M	177	N	6,591
70	F	192	N	4,300

Visualizing relationships between response and predictors

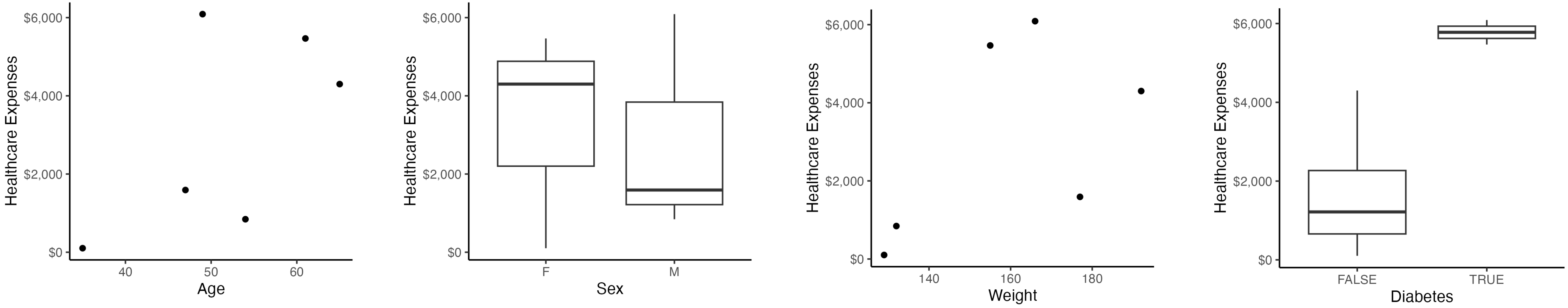
Outcome of interest: Healthcare expenses in next year

(a) Age

(b) Sex

(c) Weight

(d) Diabetes



(New) Unlabeled data

Outcome of interest: Healthcare expenses in next year

Age	Sex	Weight	Diabetes	Healthcare expenses
44	M	165	N	?
69	F	161	Y	?
78	M	170	N	?
66	M	191	N	?

Labeled versus Unlabeled data

Labeled data (**Training** data)

Age	Sex	Wt	Diab	Health exp
54	M	132	N	844
76	F	155	Y	5,467
49	M	166	Y	8,089
39	F	129	N	103
47	M	177	N	6,591
70	F	192	N	4,300

Unlabeled data

Age	Sex	Wt	Diab
44	M	165	N
69	F	161	Y
78	M	170	N
66	M	191	N

Prediction

Health exp
4,697
5,776
3,681
6,225

$$\widehat{\text{healthcare expenses}} = -10{,}612 - 44\ \text{age} + 1{,}405\ \text{sex} + 96\ \text{weight} + 3{,}968\ \text{diabetes}$$

Least Squares (LS) for continuous responses

$$\widehat{\text{healthcare expenses}} = -10{,}612 - 44\,\text{age} + 1{,}405\,\text{sex} + 96\,\text{weight} + 3{,}968\,\text{diabetes}$$

(Predicted) Response

(Predicted) Outcome

Predictor variables

Predictors

Covariates

Features

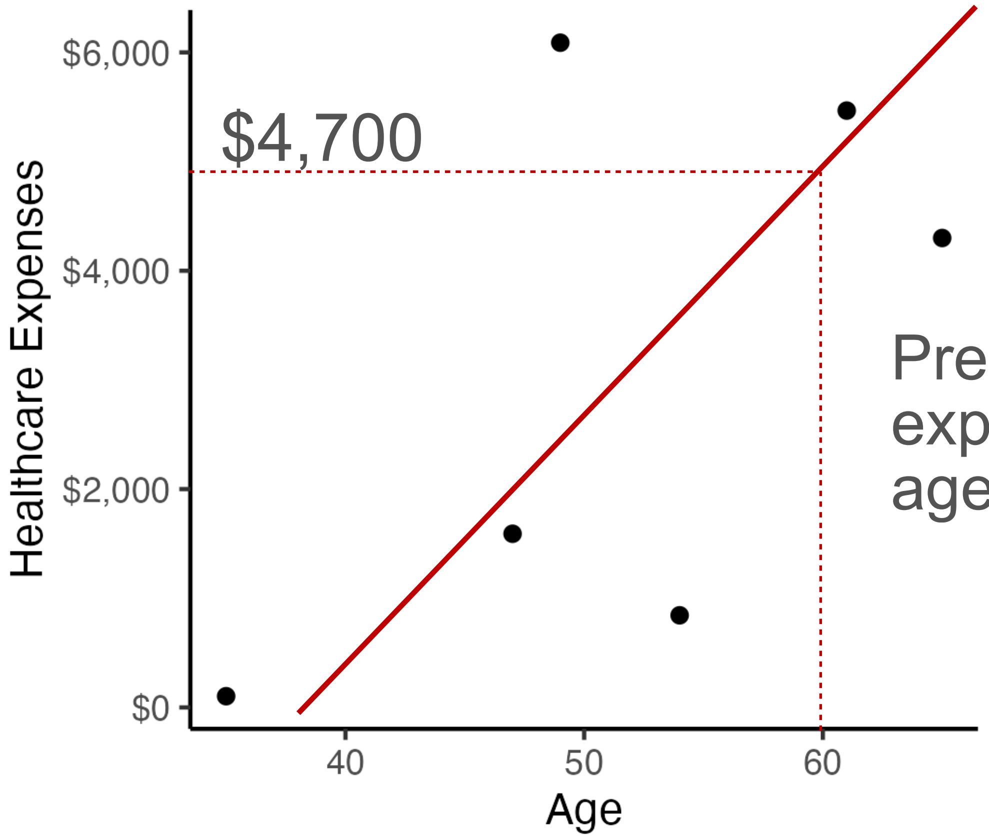
Notes:

LS prediction problems are sometimes called **linear regression** problems

LS can only be used to generate predictions of **continuous responses**

Generating predictions from a linear fit

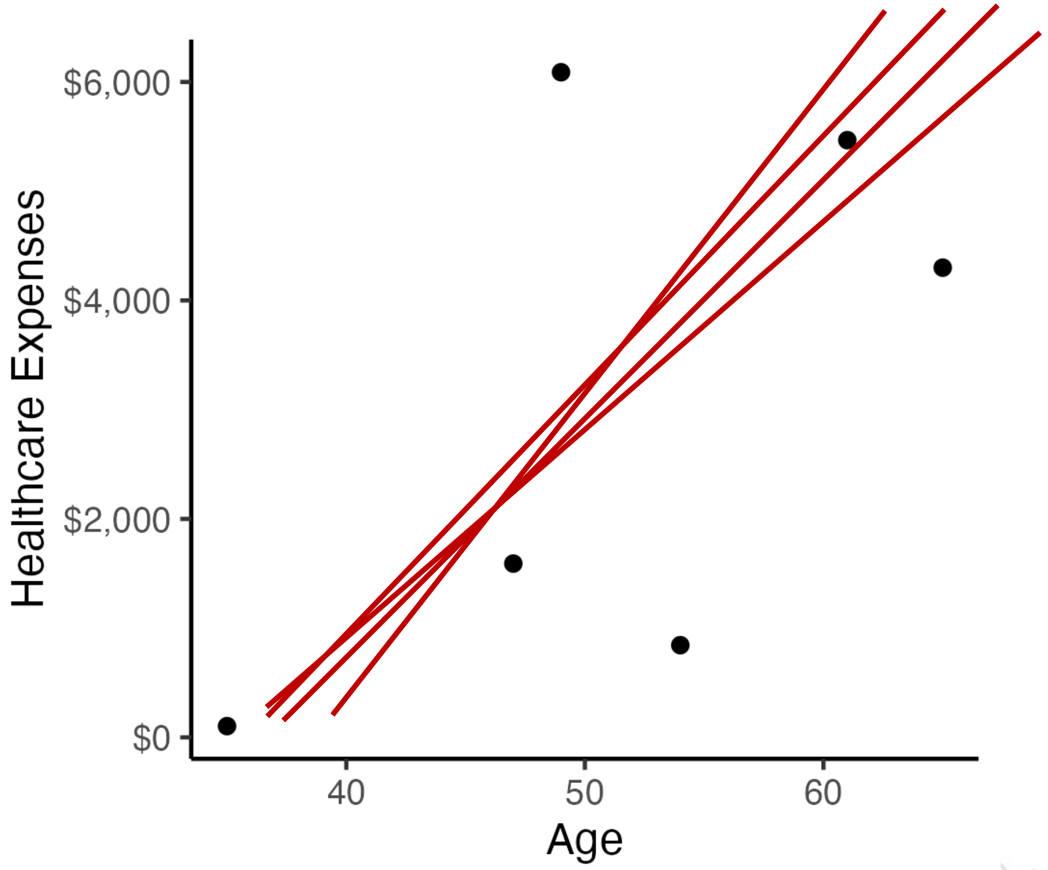
Consider a simplified linear predictive model: $\widehat{\text{hlthx}} = b_0 + b_1 \text{ age}$



Predicted healthcare expenses for someone aged 60

Choosing a linear fit

There are many possible linear fits to choose from



$$\widehat{\text{hlthx}} = -3500 + 162 \text{ age}$$
$$\widehat{\text{hlthx}} = -4100 + 149 \text{ age}$$
$$\widehat{\text{hlthx}} = -4290 + 142 \text{ age}$$
$$\widehat{\text{hlthx}} = -4400 + 135 \text{ age}$$

Least Squares (LS) for continuous responses

The **LS** fit is the one whose squared distance between the “observed” and “predicted” response is minimized

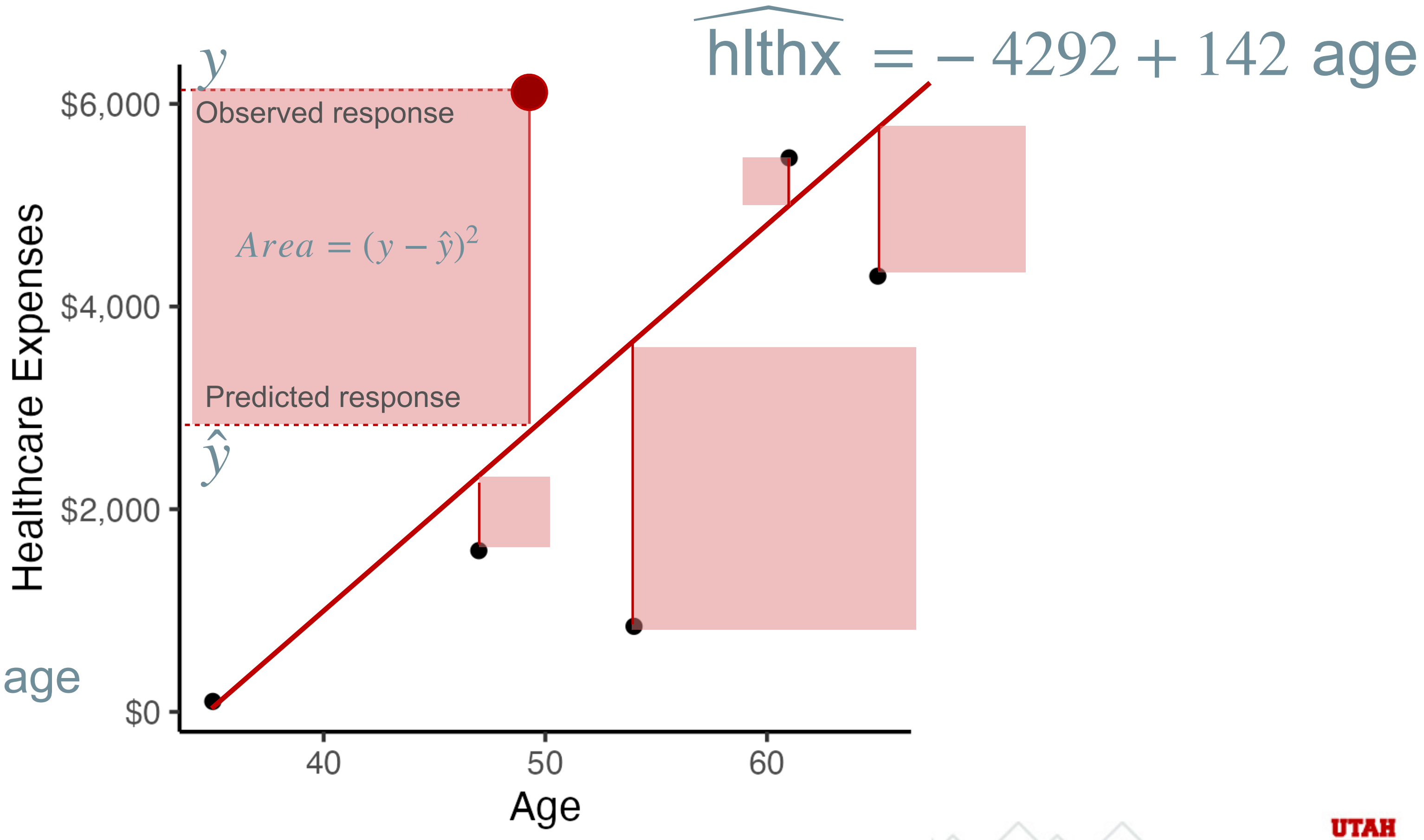
LS minimizes the

Mean Squared Error (MSE)

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

Actual observed response (hlthx)

$\widehat{\text{hlthx}} = -4292 + 142 \text{ age}$



Least Squares (LS) for continuous responses

$$\widehat{\text{healthcare expenses}} = -10{,}612 - 44\ \text{age} + 1{,}405\ \text{sex} + 96\ \text{weight} + 3{,}968\ \text{diabetes}$$

Age	Sex	Weight	Diabetes
44	M	165	N
69	F	161	Y
78	M	170	N
66	M	191	N

Predicted Healthcare exp
?
?
?
?

Least Squares (LS) for continuous responses

$$\widehat{\text{healthcare expenses}} = -10{,}612 - 44\ \text{age} + 1{,}405\ \text{sex} + 96\ \text{weight} + 3{,}968\ \text{diabetes}$$

Age	Sex	Weight	Diabetes
44	M	165	N
69	F	161	Y
78	M	170	N
66	M	191	N

Predicted Healthcare exp
4697
?
?
?

Least Squares (LS) for continuous responses

$$\widehat{\text{healthcare expenses}} = -10{,}612 - 44\ \text{age} + 1{,}405\ \text{sex} + 96\ \text{weight} + 3{,}968\ \text{diabetes}$$

Age	Sex	Weight	Diabetes
44	M	165	N
69	F	161	Y
78	M	170	N
66	M	191	N

Predicted Healthcare exp
4697
?
?
?

Least Squares (LS) for continuous responses

$$\widehat{\text{healthcare expenses}} = -10{,}612 - 44\ \text{age} + 1{,}405\ \text{sex} + 96\ \text{weight} + 3{,}968\ \text{diabetes}$$

Age	Sex	Weight	Diabetes
44	M	165	N
69	F	161	Y
78	M	170	N
66	M	191	N

Predicted Healthcare exp
4697
5776
?
?

Least Squares (LS) for continuous responses

$$\widehat{\text{healthcare expenses}} = -10{,}612 - 44\ \text{age} + 1{,}405\ \text{sex} + 96\ \text{weight} + 3{,}968\ \text{diabetes}$$

Age	Sex	Weight	Diabetes
44	M	165	N
69	F	161	Y
78	M	170	N
66	M	191	N

Predicted Healthcare exp
4697
5776
?
?

Least Squares (LS) for continuous responses

$$\widehat{\text{healthcare expenses}} = -10{,}612 - 44\ \text{age} + 1{,}405\ \text{sex} + 96\ \text{weight} + 3{,}968\ \text{diabetes}$$

Age	Sex	Weight	Diabetes
44	M	165	N
69	F	161	Y
78	M	170	N
66	M	191	N

Predicted Healthcare exp
4697
5776
3681
?

Least Squares (LS) for continuous responses

$$\widehat{\text{healthcare expenses}} = -10{,}612 - 44\ \text{age} + 1{,}405\ \text{sex} + 96\ \text{weight} + 3{,}968\ \text{diabetes}$$

Age	Sex	Weight	Diabetes
44	M	165	N
69	F	161	Y
78	M	170	N
66	M	191	N

Predicted Healthcare exp
4697
5776
3681
?

Least Squares (LS) for continuous responses

$$\widehat{\text{healthcare expenses}} = -10{,}612 - 44\ \text{age} + 1{,}405\ \text{sex} + 96\ \text{weight} + 3{,}968\ \text{diabetes}$$

Age	Sex	Weight	Diabetes
44	M	165	N
69	F	161	Y
78	M	170	N
66	M	191	N

Predicted Healthcare exp
4697
5776
3681
6225

Evaluating continuous response predictions

Dr Rebecca Barter

Evaluating predictions

Age	Sex	Weight	Diabetes
44	M	165	N
69	F	161	Y
78	M	170	N
66	M	191	N

Predicted Healthcare exp
4697
5776
3681
6225

How do we know whether these predicted healthcare expense values are accurate?

We need to compare them with the *observed* numbers.

But we haven't yet observed the *actual* numbers for these people…

Training an algorithm and generating a prediction

Labeled data

Age	Sex	Wt	Diab	Health exp
54	M	132	N	844
76	F	155	Y	5,467
49	M	166	Y	8,089
39	F	129	N	103
47	M	177	N	6,591
70	F	192	N	4,300

We **know** the **observed** healthcare expenses for the labeled data

Unlabeled data

Age	Sex	Wt	Diab
44	M	165	N
69	F	161	Y
78	M	170	N
66	M	191	N

Prediction

Pred Health exp
4697
5776
3681
6225

We **do not know** the **observed** healthcare expenses for the unlabeled data

$$\widehat{\text{healthcare expenses}} = -10{,}612 - 44\ \text{age} + 1{,}405\ \text{sex} + 96\ \text{weight} + 3{,}968\ \text{diabetes}$$

We need to **evaluate** our predictions using **labeled data**

Evaluating predictions

We need to evaluate our predictions using **labeled data**

Labeled data

Compare obs vs pred

Age	Sex	Wt	Diab	Health exp	Pred Helath exp
54	M	132	N	844	1,089
76	F	155	Y	5,467	4,892
49	M	166	Y	8,089	8,541
39	F	129	N	103	56
47	M	177	N	6,591	5,717
70	F	192	N	4,300	4,740

$$\widehat{\text{healthcare expenses}} = -\,10{,}612 - 44\ \text{age} + 1{,}405\ \text{sex} + 96\ \text{weight} + 3{,}968\ \text{diabetes}$$

Problem: The algorithm was *“trained”* using these labeled data

The algorithm may be better able to predict these responses than it would for data it was not trained on!

We should evaluate algorithms using data that reflects data we will be applying the algorithm to!

Training and testing sets

Since the only labeled data is usually the data we have, we need to **split** our data into training and testing sets

Labeled data

Age	Sex	Wt	Diab	Health exp
54	M	132	N	844
76	F	155	Y	5,467
49	M	166	Y	8,089
39	F	129	N	103
47	M	177	N	6,591
70	F	192	N	4,300

Training set (~70%)

Age	Sex	Wt	Diab	Health exp
54	M	132	N	844
76	F	155	Y	5,467
39	F	129	N	103
47	M	177	N	6,591

Train algorithm

$$\widehat{\text{health exp}} = -10{,}612 - 44\ \text{age} + 1{,}405\ \text{sex} + 96\ \text{weight} + 3{,}968\ \text{diabetes}$$

Test set (~30%)

Age	Sex	Wt	Diab	Health exp
49	M	166	Y	8,089
70	F	192	N	4,300

Evaluate algorithm

Pred Health exp
8,541
4,740

Training, **validation**, and test set

When we are fitting many algorithms, we often use the test/validation set performance to choose the best one

This means that our evaluations are no longer independent of our “final” algorithm

In practice, people will often split their data three ways into **training** (~60%), **validation** (~20%) and **test** (~20%) sets

For this course, we will just use a training and test set

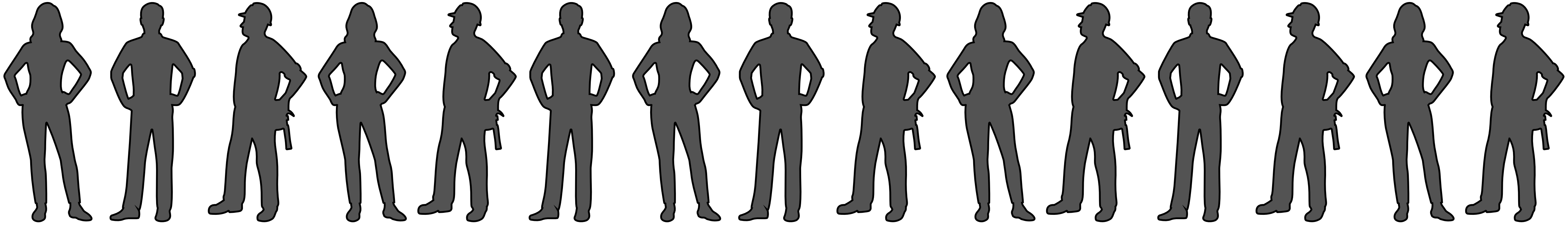
How to split?

Your test set should resemble the data that you will be applying your algorithm to

How to split? **Random split**

Your test set should resemble the data that you will be applying your algorithm to

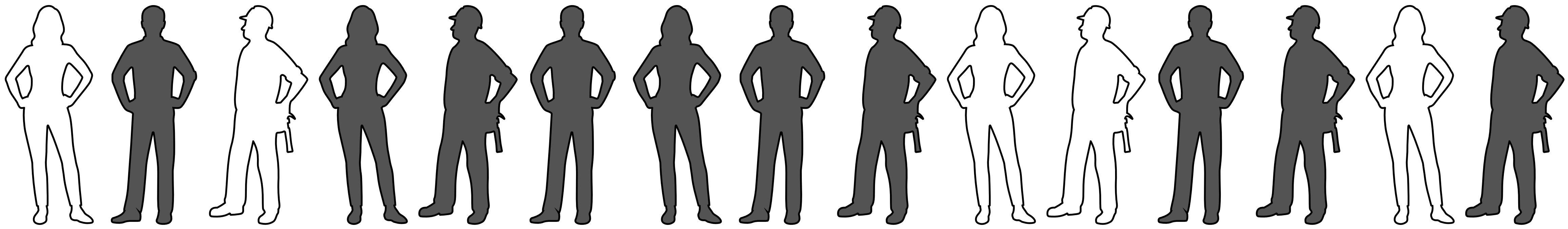
If you will be applying your algorithm to similar but **equivalent** people/units, then you should use a **random split** (i.e., a random set of 70% of the data are the training set and the other 30% of the data are the test set)



How to split? **Random split**

Your test set should resemble the data that you will be applying your algorithm to

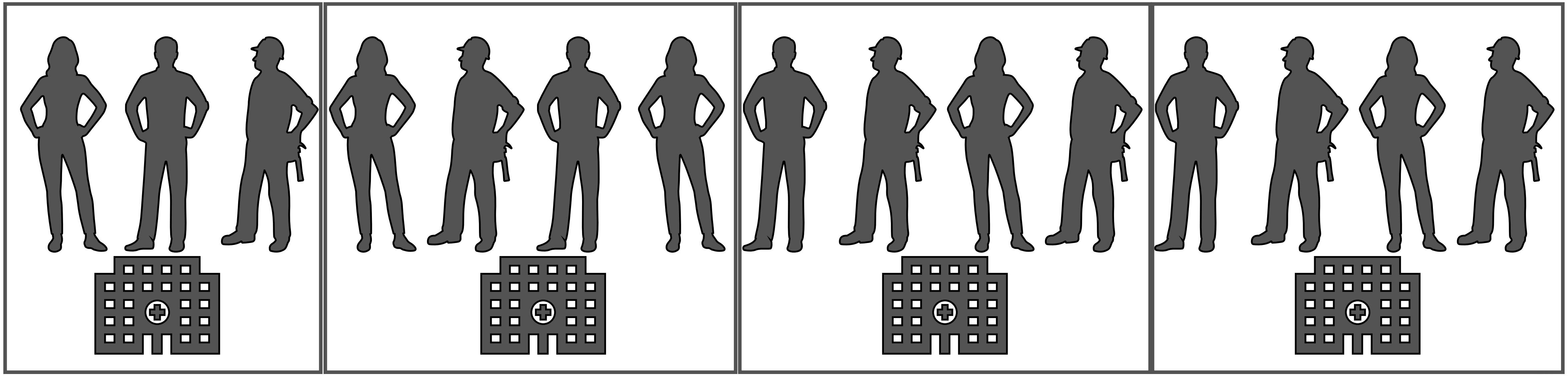
If you will be applying your algorithm to similar but **equivalent** people/units, then you should use a **random split** (i.e., a random set of 70% of the data are the training set and the other 30% of the data are the test set)



How to split? **Grouped split**

Your test set should resemble the data that you will be applying your algorithm to

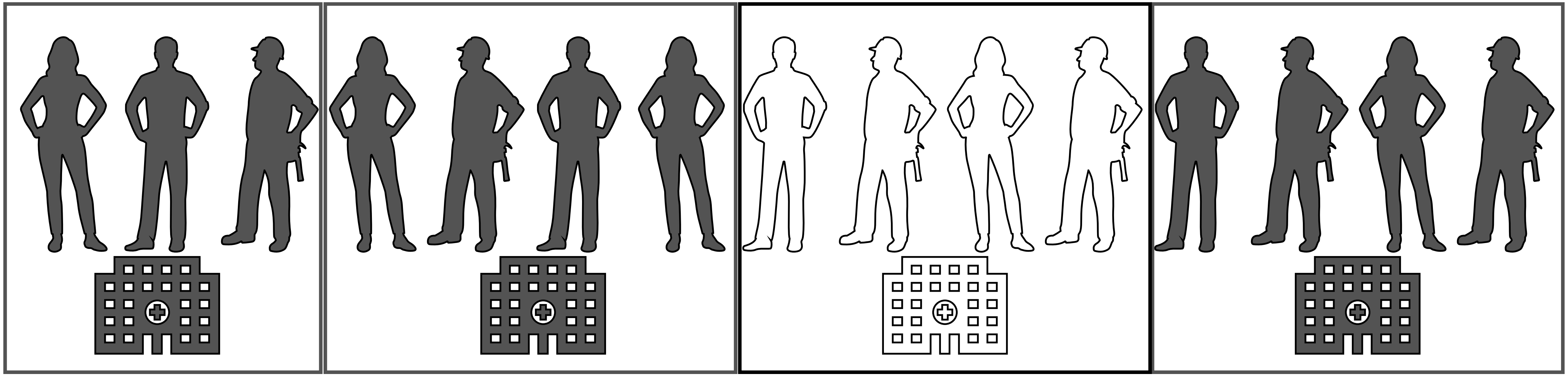
If your data comes from a collection of hospitals and you will be applying your algorithm to **new hospitals**, you should use a **grouped split** (e.g., 70% of the *hospitals* are the training set, and 30% are the test set)



How to split? **Grouped split**

Your test set should resemble the data that you will be applying your algorithm to

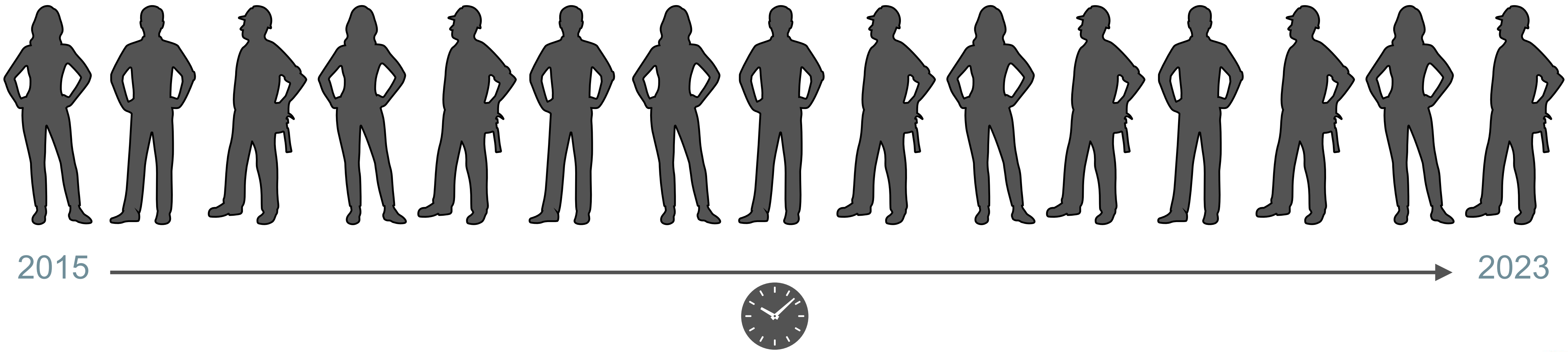
If your data comes from a collection of hospitals and you will be applying your algorithm to **new hospitals**, you should use a **grouped split** (e.g., 70% of the *hospitals* are the training set, and 30% are the test set)



How to split? **Time-based split**

Your test set should resemble the data that you will be applying your algorithm to

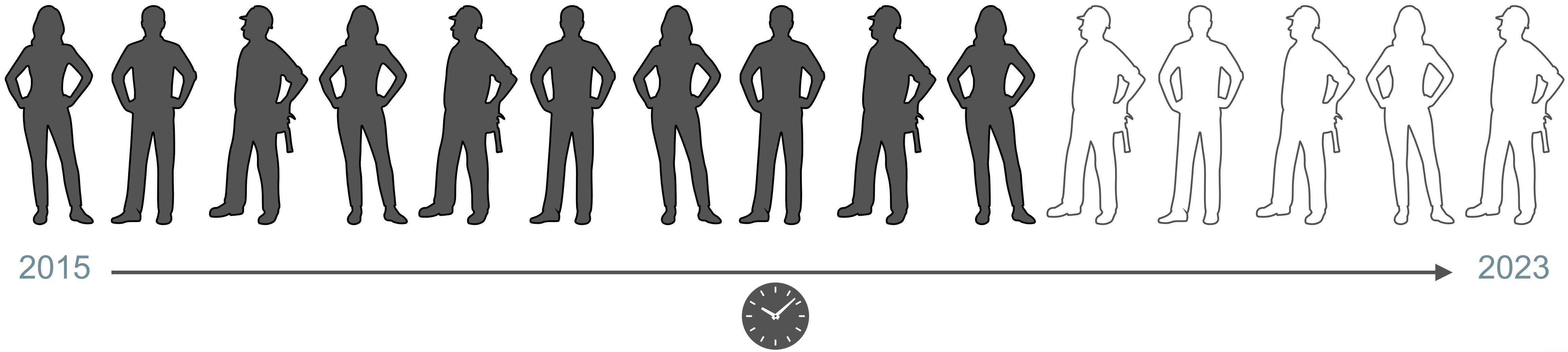
If you will be applying your algorithm to the same people/units, but in the **future**, you should use a **time-based split** (i.e., the earliest 70% of the data is the training set and the final 30% of the data are the test set)



How to split? **Time-based split**

Your test set should resemble the data that you will be applying your algorithm to

If you will be applying your algorithm to the same people/units, but in the **future**, you should use a **time-based split** (i.e., the earliest 70% of the data is the training set and the final 30% of the data are the test set)



Quantifying predictive performance (continuous)

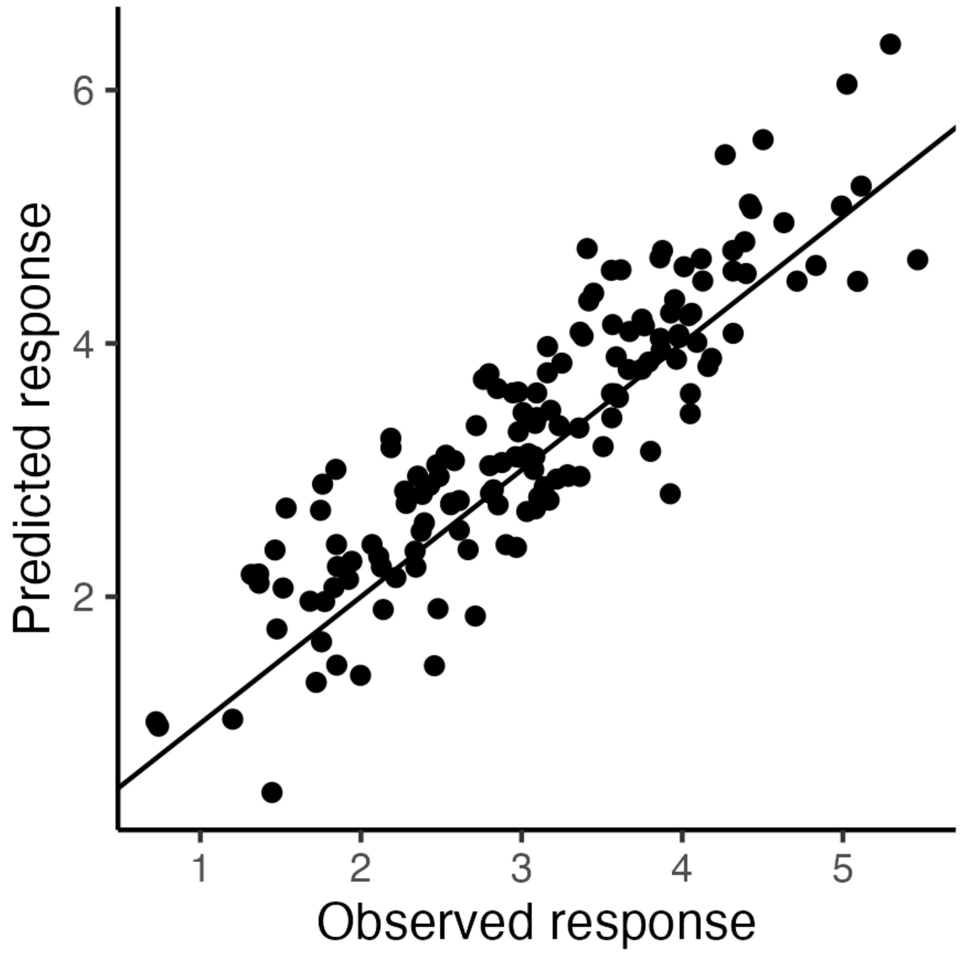
Test set predictions:

y Health exp		$\hat{y}$ Pred Health exp
8,089	⟷	8,541
4,300	⟷	4,740

Measures of predictive performance for continuous responses

Correlation (ρ)	R^2	(r)MSE
$\rho = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}$	ρ^2	$MSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ $rMSE = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

Visualizing predictive performance (continuous)



$$\rho = 0.78$$

$$R^2 = \rho^2 = 0.61$$

$$rMSE = 0.55$$