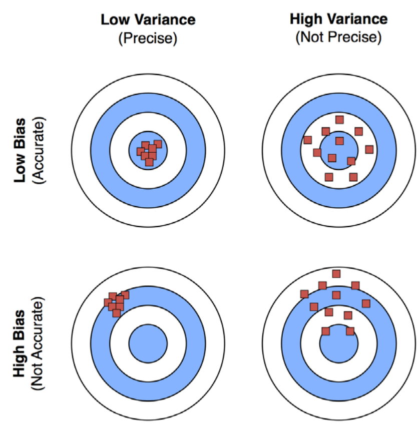
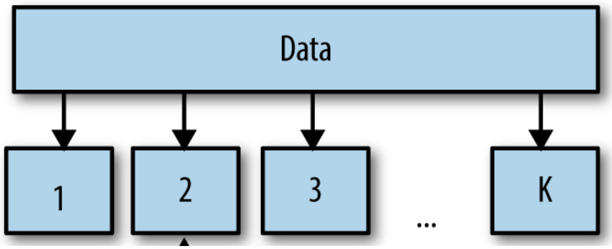
IOWA STATE UNIVERSITY

Department of Animal Science

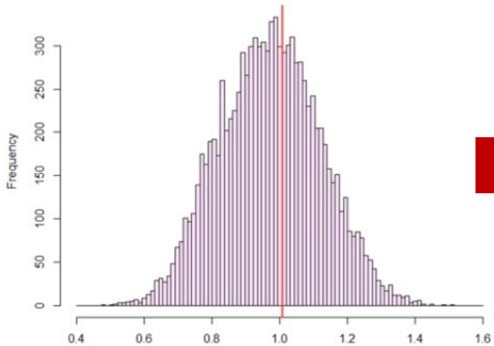
Cross Validation and Bootstrapping in Evaluating Linear Regression Models

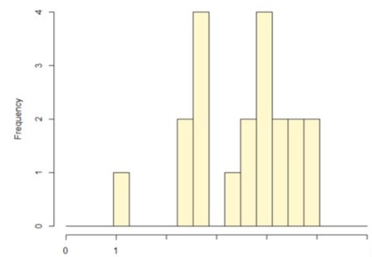




Ranga Appuhamy[1] & Luis Moraes[2]
[1]Iowa State University; [2]The Ohio State University

Modeling





Data

Model development

Model evaluation

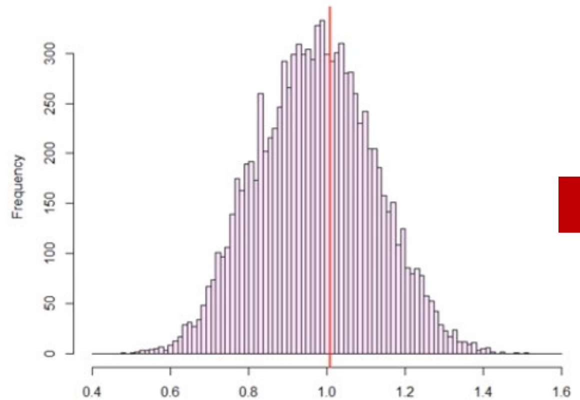
$$Error = Predicted\ value\ (\hat{Y})\ - Observed\ value\ (Y)$$

Predicted	Observed	Error	Absolute error	Squared error
10	12	-2	2	4
9	7	2	2	4
13	10	3	3	9
8	8	0	0	0
11	14	-3	3	9
			2.0 (MAE)	5.2 (MSPE)

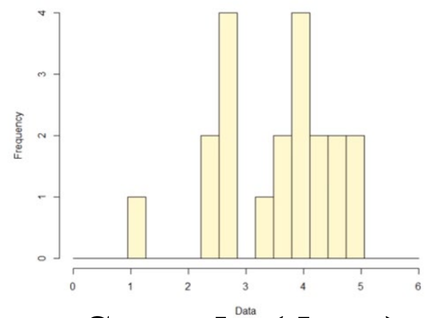
$$\mathrm{RMSPE} = \sqrt{\mathrm{MSPE}} = 2.28$$

Ranga Appuhamy
Department of Animal Science

Model Evaluation



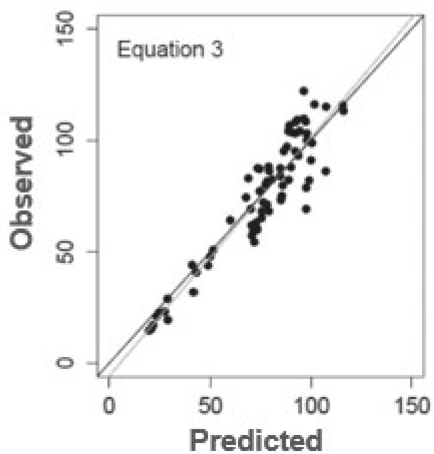
Parent population



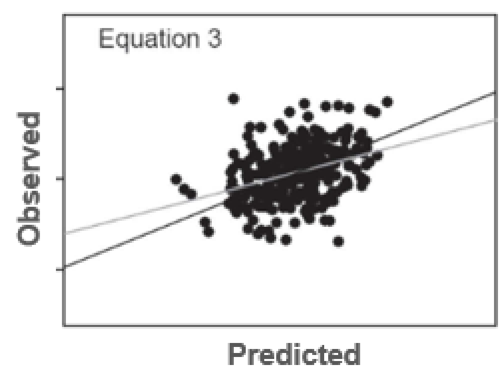
Sample (data)

Models

Using the data used to develop the model



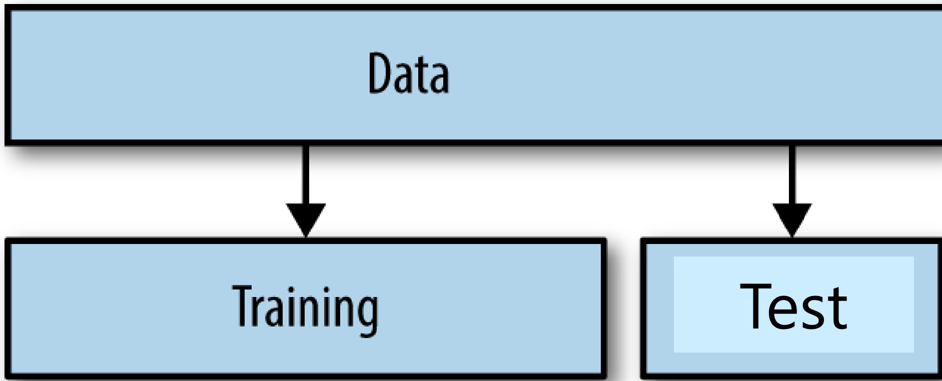
Independent dataset



IOWA STATE UNIVERSITY
Ranga Appuhamy
Department of Animal Science

Model Evaluation

Hold-out method



70 to 80% | 20 to 30%

$Y_i = \beta_0 + \beta_1 X_i$ ⇨ Prediction ⇨ Error (RMSPE)

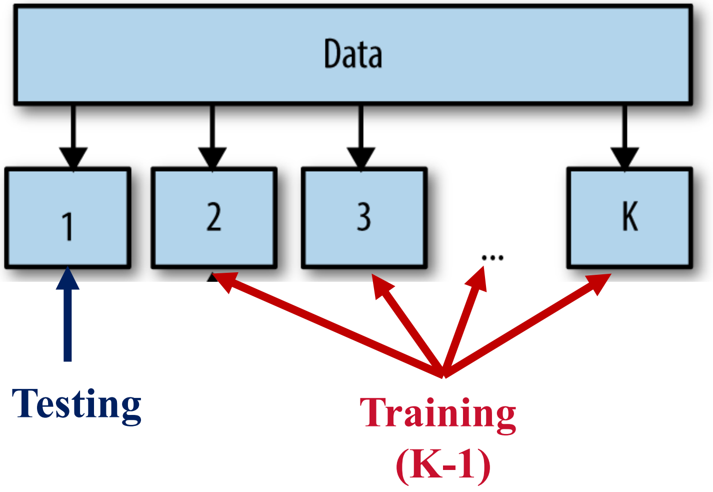
Disadvantages

- does not use all available data
- the error estimates highly depend on the split

Cross Validation

- Multiple rounds of data splitting
- Each data point has a chance of being validated against
  1. **K-fold cross validation**
  2. **Leave-one-out cross validation (LOOCV)**

K-fold cross validation



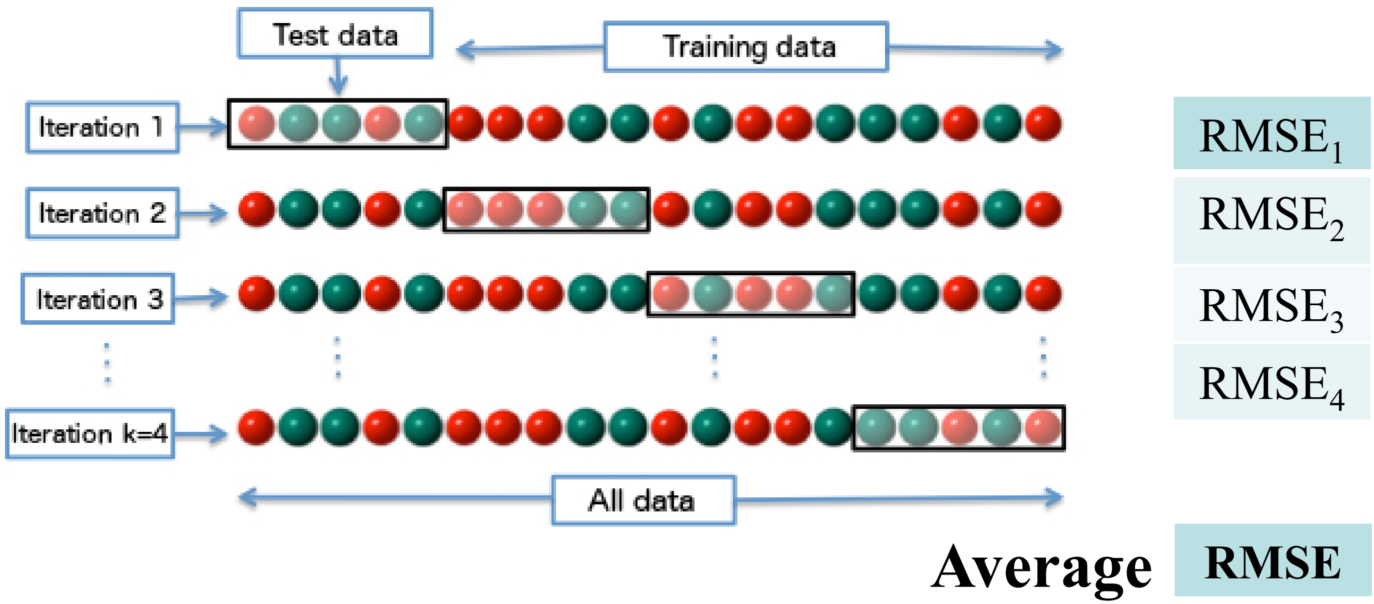
Ranga Appuhamy
Department of Animal Science

K-fold Cross validation (N=20 & K=4)

$$Y_i = \beta_0 + \beta_1 X_i$$

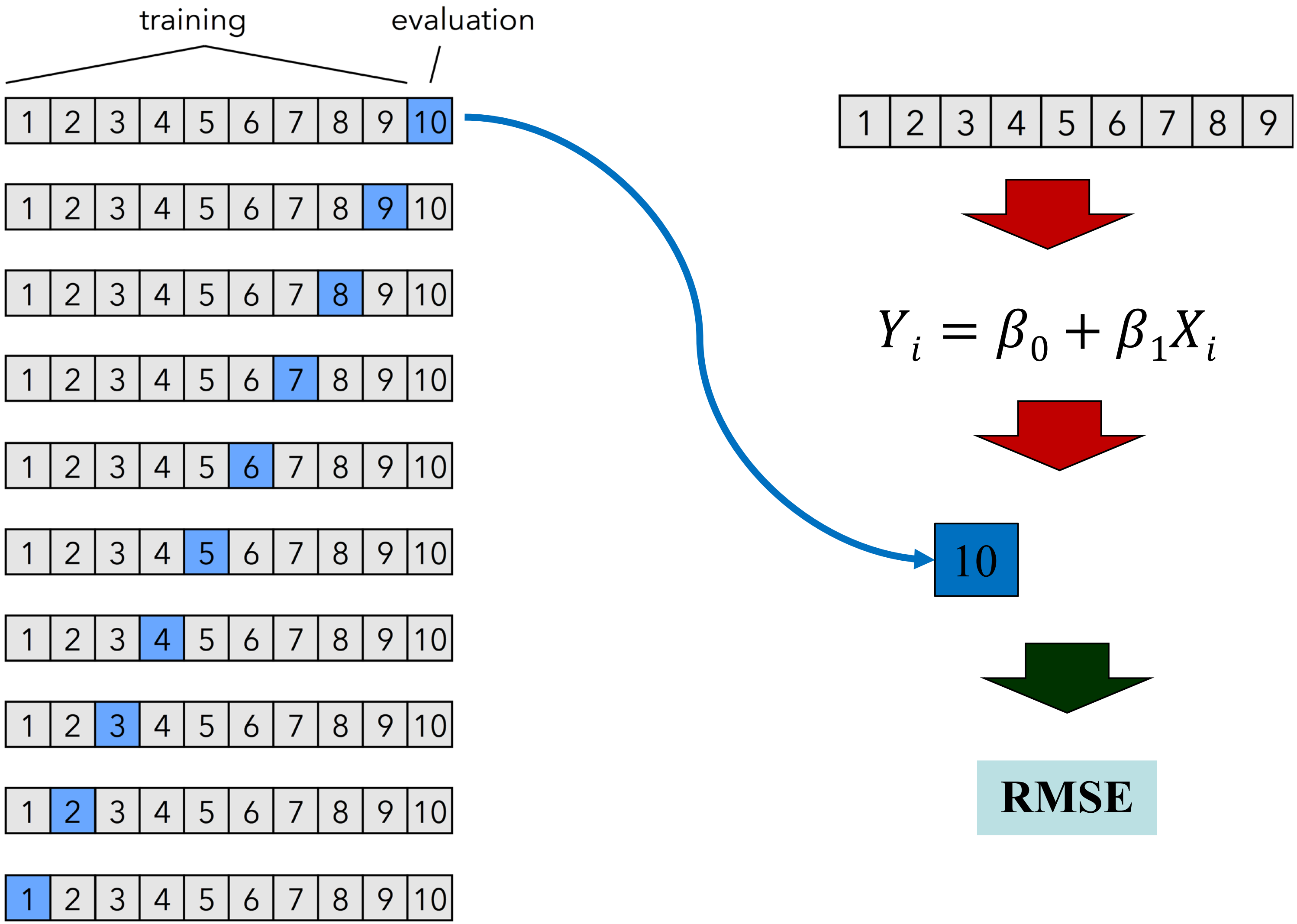
K-fold Cross validation (N=20 & K=4)

N = 20 & K = 4



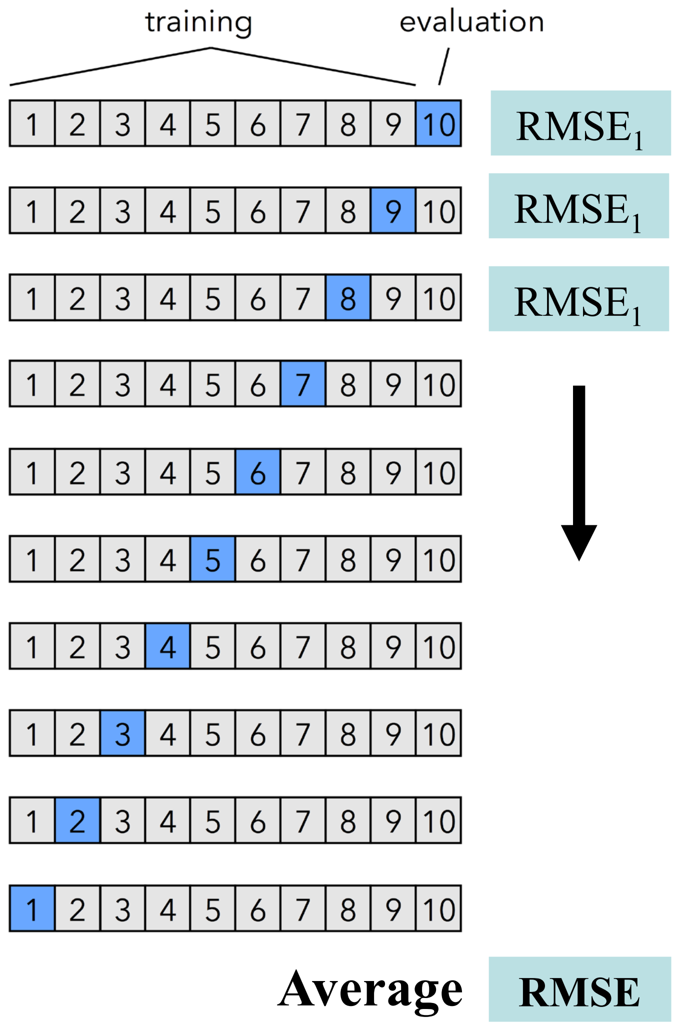
Ranga Appuhamy
Department of Animal Science

Leave-one-out Cross validation (LOOCV); K = N



Ranga Appuhamy
Department of Animal Science

Leave-one-out Cross validation (LOOCV)



IOWA STATE UNIVERSITY
Ranga Appuhamy
Department of Animal Science

Hold-out method

1 2 3 4 5 6 7 8 **9 10**

Model development | **Model evaluation**

1 × RMSPE

1 2 3 4 5 6 7 8 9 10

Model development

K-fold CV

1 < K < N

1 2 3 4 5 6 7 8 **9 10**

1 2 3 4 5 6 **7 8** 9 10

1 2 3 4 **5 6** 7 8 9 10

1 2 **3 4** 5 6 7 8 9 10

1 2 3 4 5 6 7 8 9 10

K × RMSPE

1 2 3 4 5 6 7 8 9 10

Model development

LOOCV

K = N

1 2 3 4 5 6 7 8 9 **10**

1 2 3 4 5 6 7 8 **9** 10

1 2 3 4 5 6 7 **8** 9 10

1 2 3 4 5 6 **7** 8 9 10

1 2 3 4 5 **6** 7 8 9 10

1 2 3 4 **5** 6 7 8 9 10

1 2 3 **4** 5 6 7 8 9 10

1 2 **3** 4 5 6 7 8 9 10

1 **2** 3 4 5 6 7 8 9 10

1 2 3 4 5 6 7 8 9 10

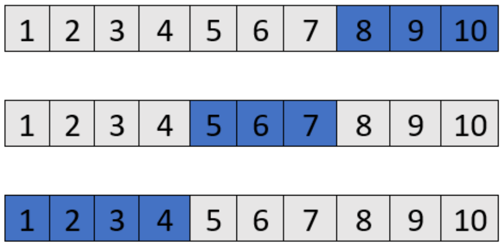
N × RMSPE

More information about model performance →

IOWA STATE UNIVERSITY
Ranga Appuhamy
Department of Animal Science

K-fold CV

K = 3

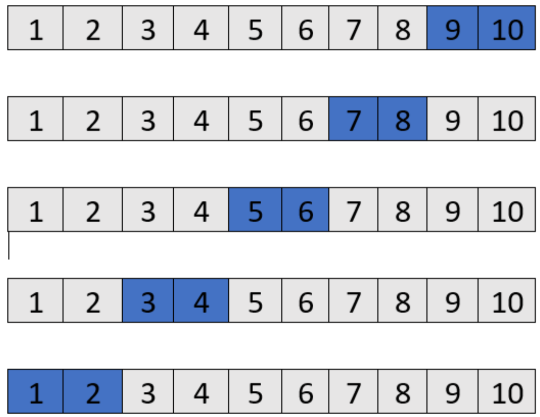


55% overlapping

(among training data)

K-fold CV

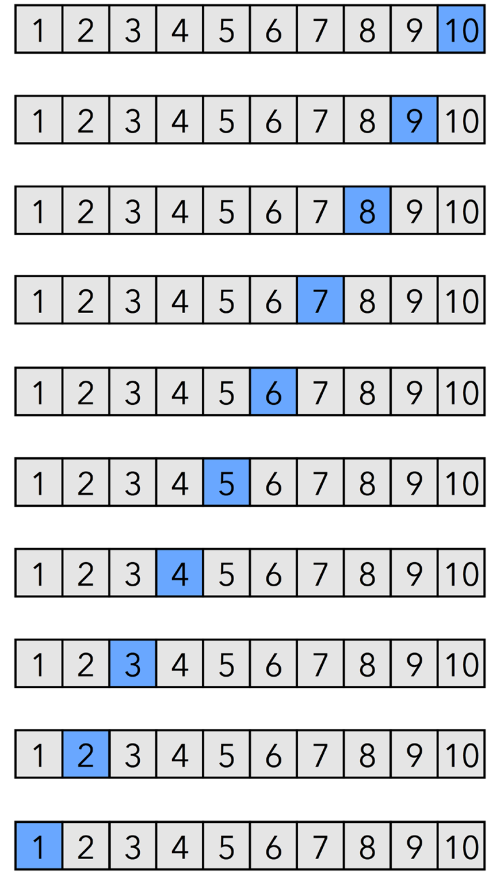
K = 5



75% overlapping

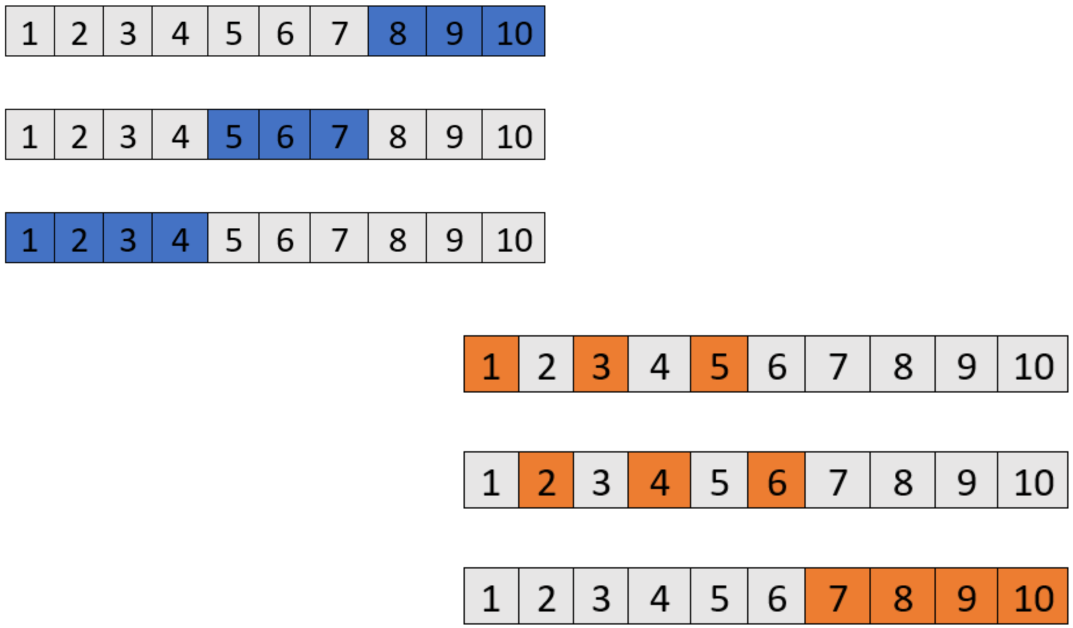
LOOCV

K = N

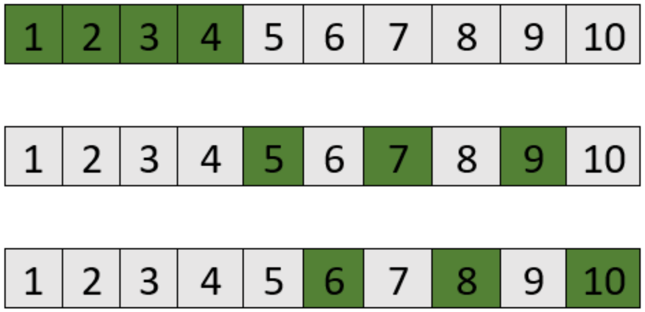


89% overlapping

Repeated K-fold cross validation



- K remains the same
- The overlapping remains the same
- Increased number of performance estimates (e.g., 3 vs. 9 RMSPE)



Ranga Appuhamy
Department of Animal Science

The caret package in R

<u>Two functions</u>

1. trainControl

```
SplitRule=trainControl(method="cv", number=10)
```

2. train

```
train(CH4~DMI, data=dat, trControl=SplitRule, method="lm")
```


Ranga Appuhamy
Department of Animal Science

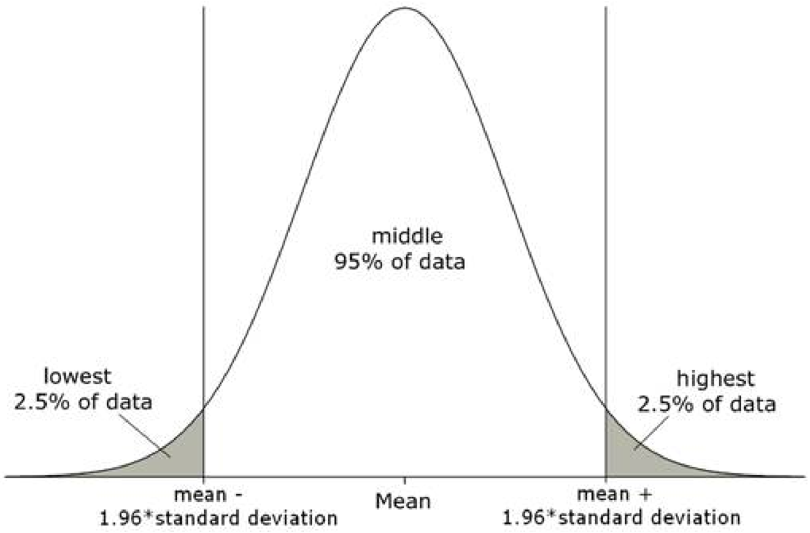
Bootstrapping

Main purpose

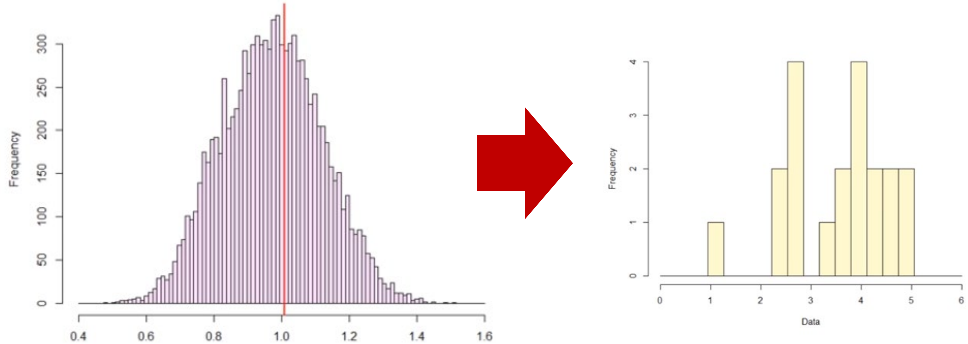
Determining standard errors and confidence intervals of

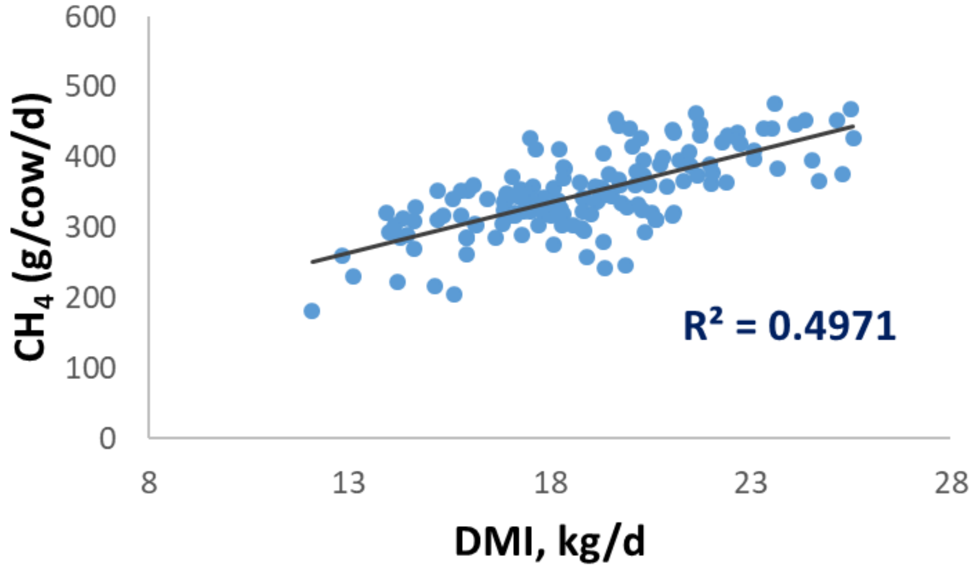
1, parameters in the model ($Y_i = \beta_0 + \beta_1 X_i$)

2, model performance estimates (RMSPE, MAE, R^2)



Ranga Appuhamy
Department of Animal Science



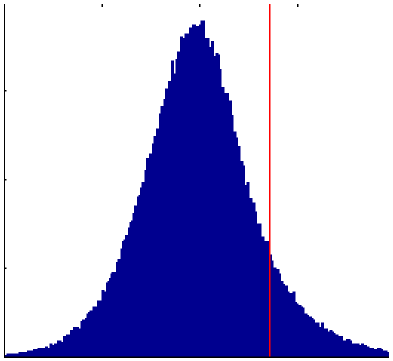


- The 95% confidence interval of R^2 value? (no standard method available)

Ranga Appuhamy
Department of Animal Science

Theoretically…………….

Parent population



Sampling (size = N)

Sample statistics

$X_1 = [x_1, x_2, x_3, \ldots\ldots x_N]$ → $(\hat{\theta}_1)$

$X_2 = [x_1, x_2, x_3, \ldots\ldots x_N]$ → $(\hat{\theta}_2)$

B = 1000

$X_B = [x_1, x_2, x_3, \ldots\ldots x_N]$ → $(\widehat{\theta_B})$

IOWA STATE UNIVERSITY
Ranga Appuhamy
Department of Animal Science

Central limit theorem

The distribution of a sample statistic approximates a normal distribution as the sample size (N) gets larger (usually >30), regardless of true population distribution shape.

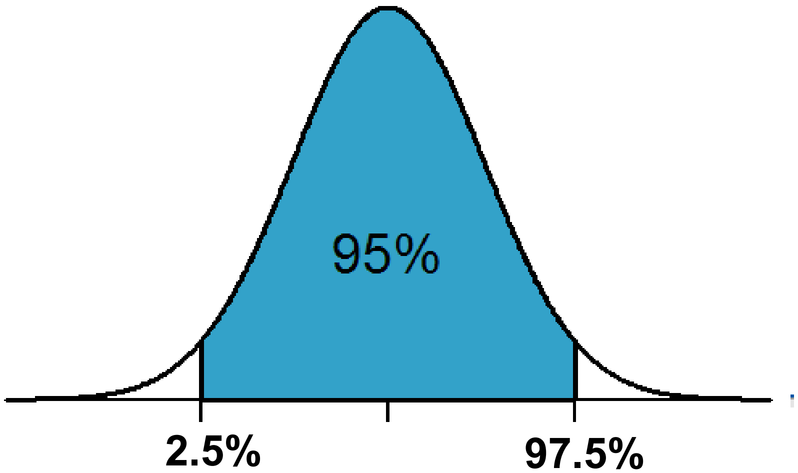
Sample statistics

$X_1 = [x_1, x_2, x_3, \ldots\ldots x_N]$ $\hat{\theta}_1$

$X_2 = [x_1, x_2, x_3, \ldots\ldots x_N]$ $\hat{\theta}_2$

B = 1000

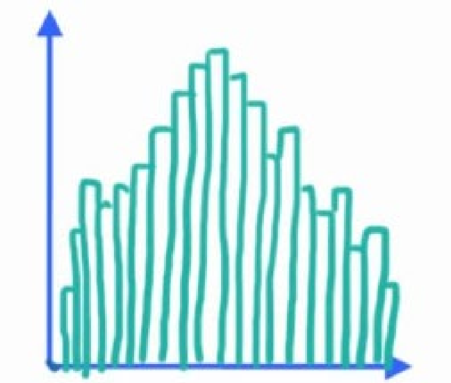
$X_B = [x_1, x_2, x_3, \ldots\ldots x_N]$ $\widehat{\theta_B}$



Sample distribution

Ranga Appuhamy
Department of Animal Science

Bootstrap Resampling



Parent population

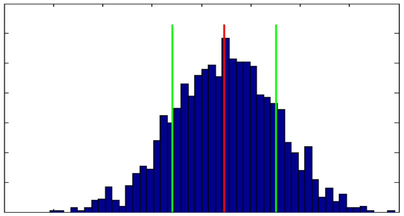
N

$X=[x_1, x_2, x_3, x_4, \ldots\ldots, x_N]$

The sample

I**<u>ndependent</u>** samples of **<u>size N</u>**

Bootstrap samples



Bootstrap distribution

SE and 95% CI

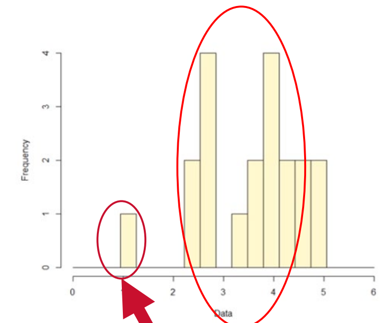
Ranga Appuhamy
Department of Animal Science

How do bootstrap samples become independent?

Bootstrap sampling is performed with replacement

IOWA STATE UNIVERSITY
Ranga Appuhamy
Department of Animal Science

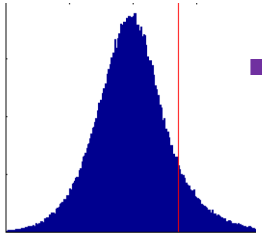
Size of bootstrap samples are as large as the original sample (N)



These observations could be less frequently sampled and thus create a bias in bootstrapping

https://www.youtube.com/watch?v=gcPIyeqymOU&t=429s

IOWA STATE UNIVERSITY
Ranga Appuhamy
Department of Animal Science



Parent population

Original sample $X = [x_1, x_2, x_3, x_4, x_5 \ldots x_N]$

Bootstrap samples

$X^{boot}{}_1 = [x_1, x_2, x_1, x_5, \ldots x_N]$

$X^{boot}{}_2 = [x_2, x_7, x_1, x_5, \ldots x_N]$

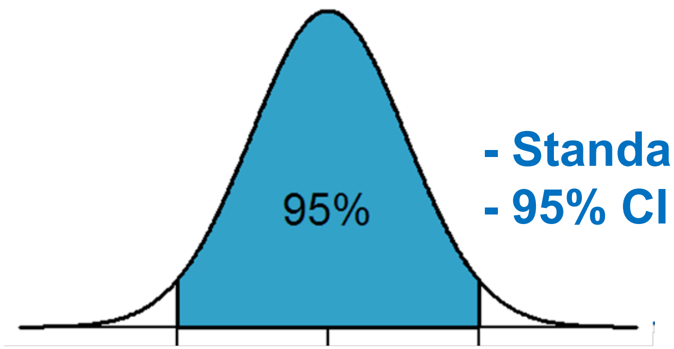
$X^{boot}{}_B = [x_5, x_9, x_3, x_9, \ldots x_N]$

θ_1^{boot} θ_2^{boot} …………………. θ_B^{boot}

If N is large enough

How many bootstrap samples (B)?

B = 1,000 to 10,000

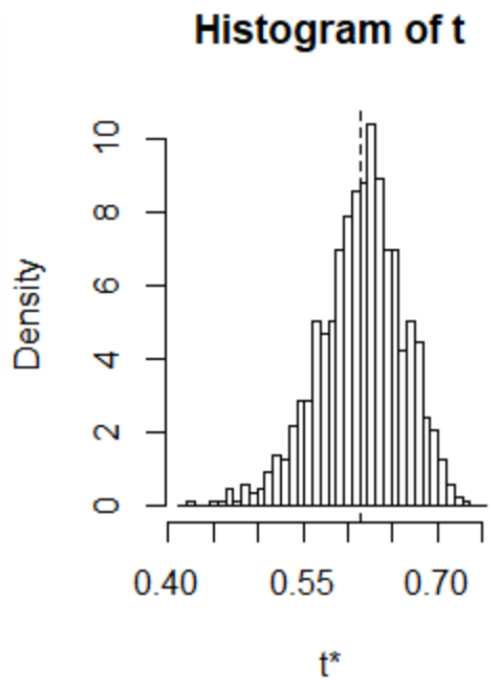


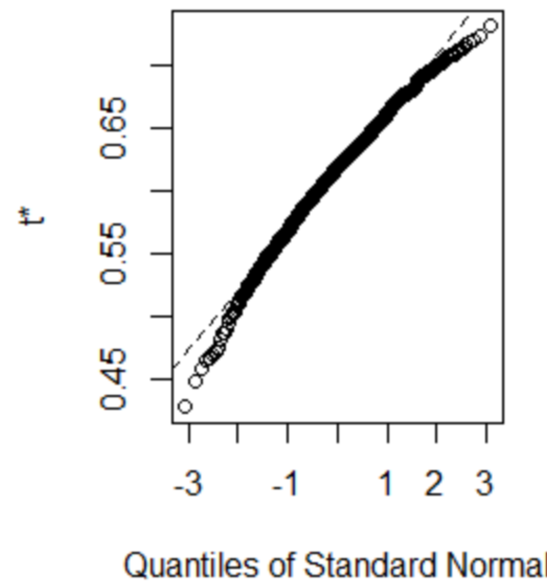
- Standard error
- 95% CI

Bootstrap distribution

IOWA STATE UNIVERSITY
Ranga Appuhamy
Department of Animal Science

Boot package in R





https://www.statmethods.net/advstats/bootstrapping.html

Ranga Appuhamy
Department of Animal Science

Questions?

Ranga Appuhamy
Department of Animal Science