## Machine Learning for Official Statistics and SDGs

# Statistical learning:
# *You've seen this before*

UNITED NATIONS SIAP
Statistical Institute for Asia and the Pacific

# Who is this course for?

# WHO IS THIS COURSE FOR?

- This course is for (present or future) Data Scientists

# Who is this course for?

- This course is for (present or future) Data Scientists
- This course uses both statistical and computational concepts

# Who is this course for?

- This course is for (present or future) Data Scientists
- This course uses both statistical and computational concepts
- This course uses many applied examples and a very progressive approach

# WHO IS THIS COURSE FOR?

- This course is for (present or future) Data Scientists
- This course uses both statistical and computational concepts
- This course uses many applied examples and a very progressive approach

*Data Scientist: "Person who is better at statistics than any software engineer and better at software than any statistician."*

J. Wills (2012)

# What Is Statistical Learning?

# What Is Statistical Learning?

*“ Statistical learning refers to a vast set of tools for understanding data”*

Gareth James, Daniela Witten, Trevor Hastie , Robert Tibshirani (2021)

# What Is Statistical Learning?

*" Statistical learning refers to a vast set of tools for understanding data"*

Gareth James, Daniela Witten, Trevor Hastie , Robert Tibshirani (2021)

- Involves building statistical models

# WHAT IS STATISTICAL LEARNING?

*" Statistical learning refers to a vast set of tools for understanding data"*

Gareth James, Daniela Witten, Trevor Hastie , Robert Tibshirani (2021)

- Involves building statistical models
- Goals are estimation or prediction

# What Is Statistical Learning?

Two main learning problems:

# WHAT IS STATISTICAL LEARNING?

Two main learning problems:

- We observe **both** the outcome $y$ and regressors (also called features) $x$s

# What Is Statistical Learning?

Two main learning problems:

- We observe **both** the outcome $y$ and regressors (also called features) $x$s

$\hookrightarrow$ **Supervised** learning

# What Is Statistical Learning?

Two main learning problems:

- We observe **both** the outcome $y$ and regressors (also called features) $x$s

$\hookrightarrow$ **Supervised** learning

*Most of the examples and applications are supervised learning*

# WHAT IS STATISTICAL LEARNING?

Two main learning problems:

- We observe **both** the outcome $y$ and regressors (also called features) $x$s

$\hookrightarrow$ **Supervised** learning

*Most of the examples and applications are supervised learning*

- We **do not** observe the outcome $y$ but **only** several $x$s

# What Is Statistical Learning?

Two main learning problems:

- We observe **both** the outcome $y$ and regressors (also called features) $x$s

$\hookrightarrow$ **Supervised** learning

*Most of the examples and applications are supervised learning*

- We **do not** observe the outcome $y$ but **only** several $x$s

$\hookrightarrow$ **Unsupervised** learning (or *cluster analysis*)

# What Is Statistical Learning?

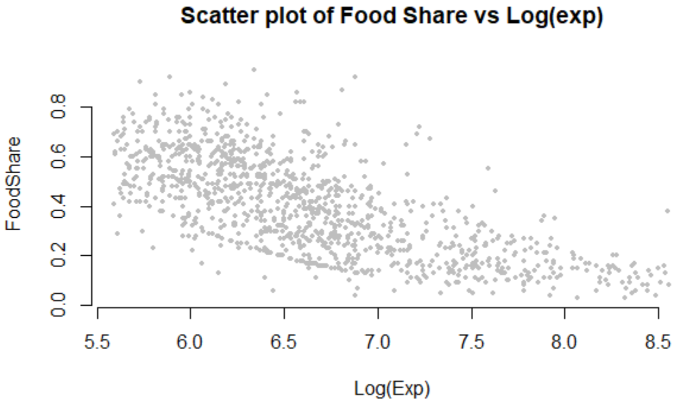
Two main learning problems:

- ► We observe **both** the outcome $y$ and regressors (also called features) $x$s
- $\hookrightarrow$ **Supervised** learning

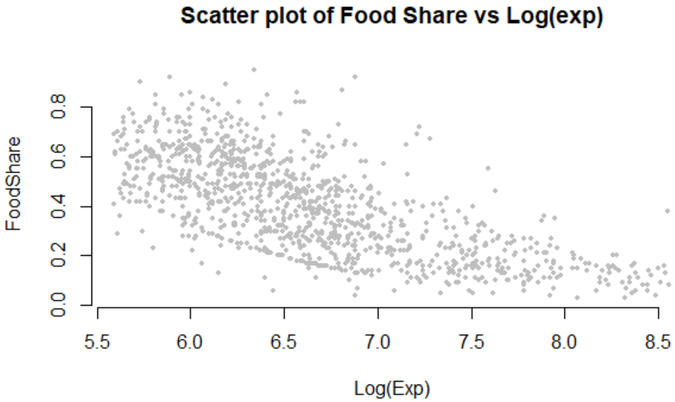
  *Most of the examples and applications are supervised learning*
- ► We **do not** observe the outcome $y$ but **only** several $x$s
- $\hookrightarrow$ **Unsupervised** learning (or *cluster analysis*)

  *More complex models we'll see at the end*

# Statistical learning on an example

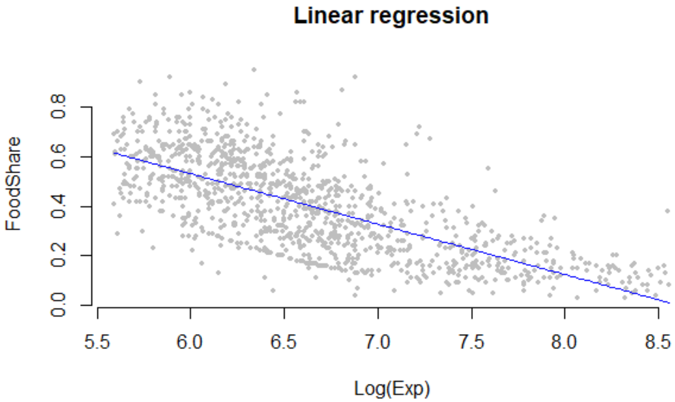


# Statistical learning on an example

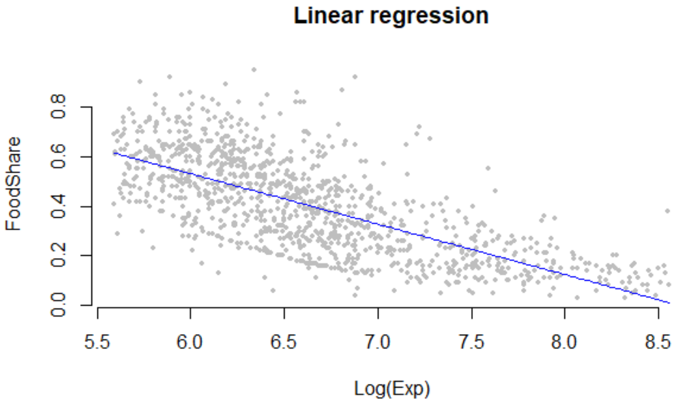



We may be interested in the relationship between the two variables

# UNDERSTANDING = ESTIMATE $f(\cdot)$



# UNDERSTANDING = ESTIMATE $f(\cdot)$




$f(\cdot)$ is the regression line

# WHY ESTIMATING $f(\cdot)$?

- Inference

# WHY ESTIMATING $f(\cdot)$?

- Inference

  Understand the nature of the relationship between $X$ and $Y$

# WHY ESTIMATING $f(\cdot)$?

- Inference

  Understand the nature of the relationship between $X$ and $Y$
  Identify "important" variables to understand $Y$

# WHY ESTIMATING $f(\cdot)$?

- Inference
  - Understand the nature of the relationship between $X$ and $Y$
  - Identify "important" variables to understand $Y$
- Prediction

# WHY ESTIMATING $f(\cdot)$?

- Inference

  Understand the nature of the relationship between $X$ and $Y$
  Identify "important" variables to understand $Y$
- Prediction

  Predict $y$ for any new $x$ using $f(\cdot)$

# Why estimating $f(\cdot)$?

- Inference

  Understand the nature of the relationship between $X$ and $Y$
  Identify "important" variables to understand $Y$
- Prediction

  Predict $y$ for any new $x$ using $f(\cdot)$
- In practice we must estimate $f(\cdot)$ using a model:

# WHY ESTIMATING $f(\cdot)$?

- Inference

  Understand the nature of the relationship between $X$ and $Y$
  Identify "important" variables to understand $Y$
- Prediction

  Predict $y$ for any new $x$ using $f(\cdot)$
- In practice we must estimate $f(\cdot)$ using a model:

$$y = f(x) + \varepsilon$$

# WHY ESTIMATING $f(\cdot)$?

- Inference

  Understand the nature of the relationship between $X$ and $Y$
  Identify "important" variables to understand $Y$
- Prediction

  Predict $y$ for any new $x$ using $f(\cdot)$
- In practice we must estimate $f(\cdot)$ using a model:

$$y = f(x) + \varepsilon$$

We denote by $\widehat{f(\cdot)}$ the estimate of $f(\cdot)$

# How to estimate $f(\cdot)$?

- Parametric methods

# How to estimate $f(\cdot)$?

- Parametric methods

  Specify a form for $f(\cdot)$, for example linear:

# How to estimate $f(\cdot)$?

- Parametric methods

  Specify a form for $f(\cdot)$, for example linear:

$$y = \beta_0 + \beta_1 x_+ \varepsilon$$

# How to estimate $f(\cdot)$?

- Parametric methods

  Specify a form for $f(\cdot)$, for example linear:

$$y = \beta_0 + \beta_1 x_+ \varepsilon$$

  - The goal is to find the line that is **minimizing** the distance to the observed points $(x_i, y_i)$. The distance is computed as the Mean Square Error (MSE):

$$MSE(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

# HOW TO ESTIMATE $f(\cdot)$?

- Parametric methods

  Specify a form for $f(\cdot)$, for example linear:

$$y = \beta_0 + \beta_1 x_+ \varepsilon$$

  - The goal is to find the line that is **minimizing** the distance to the observed points $(x_i, y_i)$. The distance is computed as the Mean Square Error (MSE):

$$MSE(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2$$

  - The regression line, defined by $\beta_0$ and $\beta_1$, is simply the solution of:

$$Min_{\ (\beta_0, \beta_1)}\ MSE(\beta_0, \beta_1)$$

# How to estimate $f(\cdot)$?

- Parametric methods

  Specify a form for $f(\cdot)$, for example linear:

$$y = \beta_0 + \beta_1 x_+ \varepsilon$$

  - The goal is to find the line that is **minimizing** the distance to the observed points $(x_i, y_i)$. The distance is computed as the Mean Square Error (MSE):

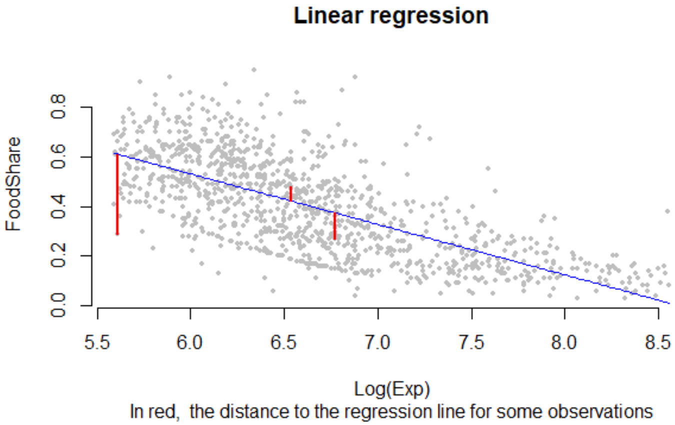
$$MSE(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2$$

  - The regression line, defined by $\beta_0$ and $\beta_1$, is simply the solution of:

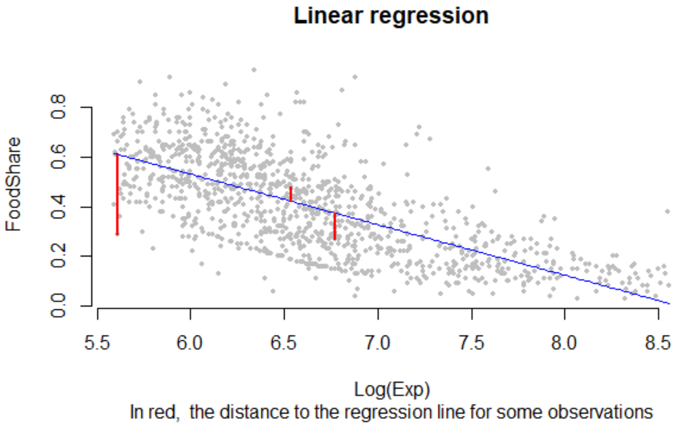
$$Min_{\ (\beta_0, \beta_1)}\ MSE(\beta_0, \beta_1)$$

- The MSE it is the *cost function* used to estimate $(\widehat{\beta}_0, \widehat{\beta}_1)$

# HOW TO ESTIMATE $f(\cdot)$: IN PRACTICE



# HOW TO ESTIMATE $f(\cdot)$: IN PRACTICE




The regression line is found by minimizing the sum of all distances or **MSE**

# RESULTS: $\widehat{f(\cdot)}$

From the result and the estimated parameters $(\widehat{\beta}_0, \widehat{\beta}_1)$, we see that there is a relation, and that it is decreasing.

| | Estimate | Std. Error | t value | $Pr(> \|t\|)$ |
|---|---|---|---|---|
| (Intercept) | 1.75 | 0.04 | 41.09 | 0 |
| ltexp | $-0.20^{***}$ | 0.01 | -31.84 | 0 |

The quality of the adjustment may be measured by the $R^2 = 0.478$

# What did we learn?

- Is the line fitting the data?

# What did we learn?

- Is the line fitting the data?

  One may always find a relation between two variables

# What did we learn?

- Is the line fitting the data?

  One may always find a relation between two variables

  Statistical test help asses the significance of a variable

# What did we learn?

- Is the line fitting the data?
  - One may always find a relation between two variables
  - Statistical test help asses the significance of a variable
- How good is the linear adjustment: $R^2$

# What did we learn?

- Is the line fitting the data?

  One may always find a relation between two variables
  
  Statistical test help asses the significance of a variable
- How good is the linear adjustment: $R^2$

$$R^2 = \frac{TSS - RSS}{TSS}$$

# What did we learn?

- Is the line fitting the data?

  One may always find a relation between two variables
  Statistical test help asses the significance of a variable

- How good is the linear adjustment: $R^2$

$$R^2 = \frac{TSS - RSS}{TSS}$$

with: $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ and $RSS = \sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2$

# WHAT DID WE LEARN?

- Is the line fitting the data?

  One may always find a relation between two variables
  Statistical test help asses the significance of a variable

- How good is the linear adjustment: $R^2$

$$R^2 = \frac{TSS - RSS}{TSS}$$

with: $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ and $RSS = \sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2$

- $R^2$ is a very popular measure. The closer to 1, the better.

# What did we learn?

- Is the line fitting the data?

  One may always find a relation between two variables
  Statistical test help asses the significance of a variable

- How good is the linear adjustment: $R^2$

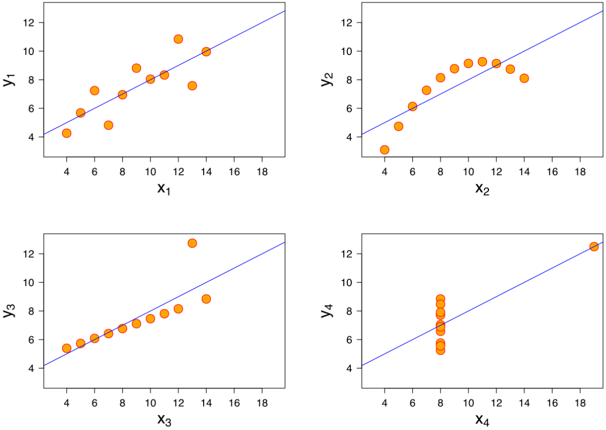
$$R^2 = \frac{TSS - RSS}{TSS}$$

  with: $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ and $RSS = \sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2$

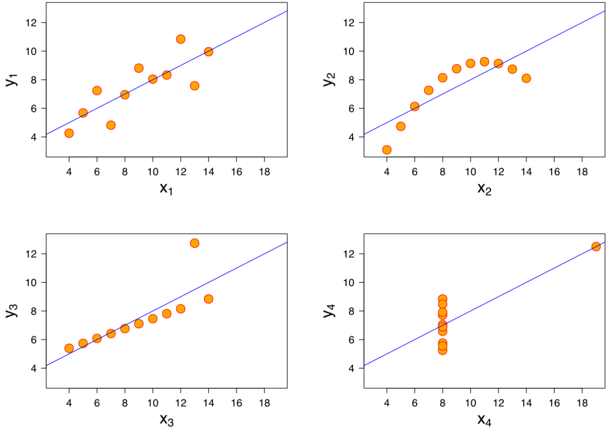
- $R^2$ is a very popular measure. The closer to 1, the better.
- $R^2$ can be very misleading

# Beware of $R^2$: Anscombe Quartet (1973)

# Beware of $R^2$: Anscombe Quartet (1973)



# Beware of $R^2$: Anscombe Quartet (1973)




In all these data sets the $R^2$ is 0.67

# Practice of Statistical Learning

*" ...make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding."*

F. J. Anscombe (1973)

# PRACTICE OF STATISTICAL LEARNING

*" ...make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding."*

F. J. Anscombe (1973)

See also the datasaurus

# Practice of Statistical Learning

*" ...make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding."*

F. J. Anscombe (1973)

See also the datasaurus

- Practice of Statistical learning can be challenging

# Practice of Statistical Learning

*" ...make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding."*

F. J. Anscombe (1973)

See also the datasaurus

- Practice of Statistical learning can be challenging

$\hookrightarrow$ Need to compute good indicators

# Practice of Statistical Learning

*" ...make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding."*

F. J. Anscombe (1973)

See also the datasaurus

- Practice of Statistical learning can be challenging

$\hookrightarrow$ Need to compute good indicators

$\hookrightarrow$ Need to understand the indicators computed

# Practice of Statistical Learning

*" ...make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding."*

F. J. Anscombe (1973)

See also the datasaurus

- Practice of Statistical learning can be challenging

$\hookrightarrow$ Need to compute good indicators

$\hookrightarrow$ Need to understand the indicators computed

$\hookrightarrow$ Need to go beyond linearity

Machine Learning for Official Statistics and SDGs

# Statistical learning:
# *Beyond linearity*

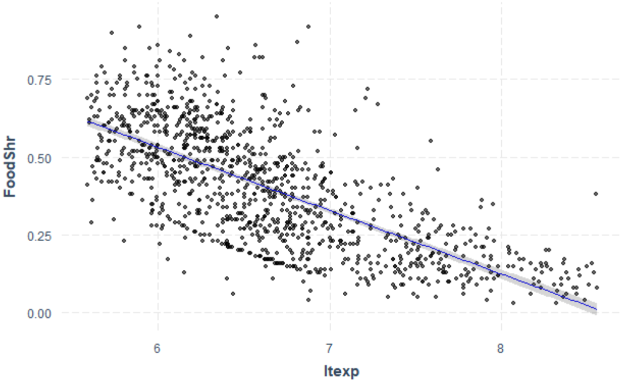
UNITED NATIONS
SIAP
Statistical Institute for Asia and the Pacific

# Beyond linearity

# Beyond linearity

- A linear model may be unadapted or too simple

$$y = \beta_0 + \beta_1 x + \varepsilon$$

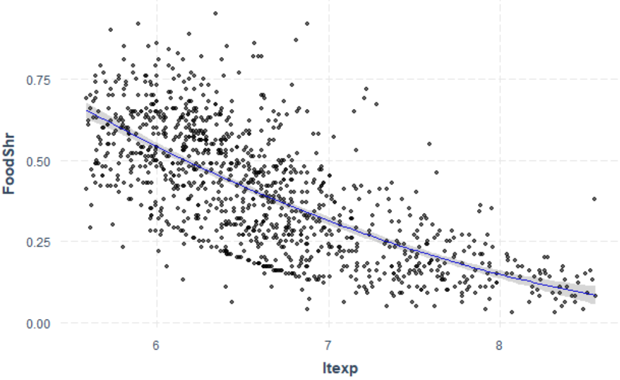



The fit (measured by $R^2$) is: $R^2$= 0.478

# Beyond linearity

- A Polynomial model may be better adapted: **Quadratic** model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

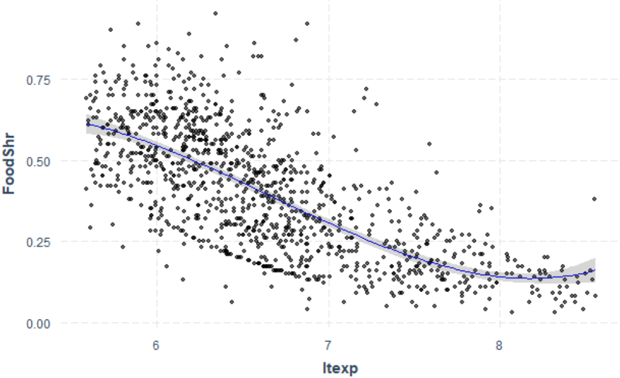



Do we have a better fit? $R^2 = 0.484$

# Beyond linearity

- Polynomial may be better adapted: **Cubic** model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$




Do we have a better fit? $R^2 = 0.490$

# IN PRACTICE

- Polynomial models may be useful

# In practice

- Polynomial models may be useful

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon$$

# In practice

- Polynomial models may be useful

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon$$

How to choose the degree $p$ ?

# IN PRACTICE

- Polynomial models may be useful

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon$$

How to choose the degree $p$ ?

Collinearity of $x^p$ and $x^q$ for $p \neq q$?

# Parametric vs nonparametric models

- Linear and polynomial models are determined by parameters $(\beta_0, \beta_2, \cdots, \beta_p)$

# Parametric vs nonparametric models

- Linear and polynomial models are determined by parameters $(\beta_0, \beta_2, \cdots, \beta_p)$
- Other methods more flexible

# Parametric vs nonparametric models

- Linear and polynomial models are determined by parameters $(\beta_0, \beta_2, \cdots, \beta_p)$
- Other methods more flexible
- Nearest neighbors (or k-NN)

# Parametric vs nonparametric models

- Linear and polynomial models are determined by parameters $(\beta_0, \beta_2, \cdots, \beta_p)$
- Other methods more flexible
- Nearest neighbors (or k-NN)

  The goal is to estimate $f(\cdot)$ not $\beta_s$!

# Parametric vs nonparametric models

- Linear and polynomial models are determined by parameters $(\beta_0, \beta_2, \cdots, \beta_p)$
- Other methods more flexible
- Nearest neighbors (or k-NN)

  The goal is to estimate $f(\cdot)$ not $\beta_s$!

  Similar in spirit to moving average

# Parametric vs nonparametric models

- Linear and polynomial models are determined by parameters $(\beta_0, \beta_2, \cdots, \beta_p)$
- Other methods more flexible
- Nearest neighbors (or k-NN)

  The goal is to estimate $f(\cdot)$ not $\beta_s$!
  Similar in spirit to moving average

$$\widehat{f}(x_i) = \frac{1}{k} \sum_{\substack{j \in \{k-nearest \\ neighbours\ of\ x_i\}}} y_j$$

# Parametric vs nonparametric models

- Linear and polynomial models are determined by parameters $(\beta_0, \beta_2, \cdots, \beta_p)$
- Other methods more flexible
- Nearest neighbors (or k-NN)

  The goal is to estimate $f(\cdot)$ not $\beta_s$!
  Similar in spirit to moving average

$$\widehat{f}(x_i) = \frac{1}{k} \sum_{\substack{j \in \{k-nearest \\ neighbours\ of\ x_i\}}} y_j$$

$k$ is the number of neighbors of $x_i$ taken into account in the estimation.

# Parametric vs nonparametric models

- Linear and polynomial models are determined by parameters $(\beta_0, \beta_2, \cdots, \beta_p)$
- Other methods more flexible
- Nearest neighbors (or k-NN)

  The goal is to estimate $f(\cdot)$ not $\beta_s$!
  Similar in spirit to moving average

$$\widehat{f}(x_i) = \frac{1}{k} \sum_{\substack{j \in \{k-nearest \\ neighbours\ of\ x_i\}}} y_j$$

  $k$ is the number of neighbors of $x_i$ taken into account in the estimation.

  - The method follows a very general idea:

# Parametric vs nonparametric models

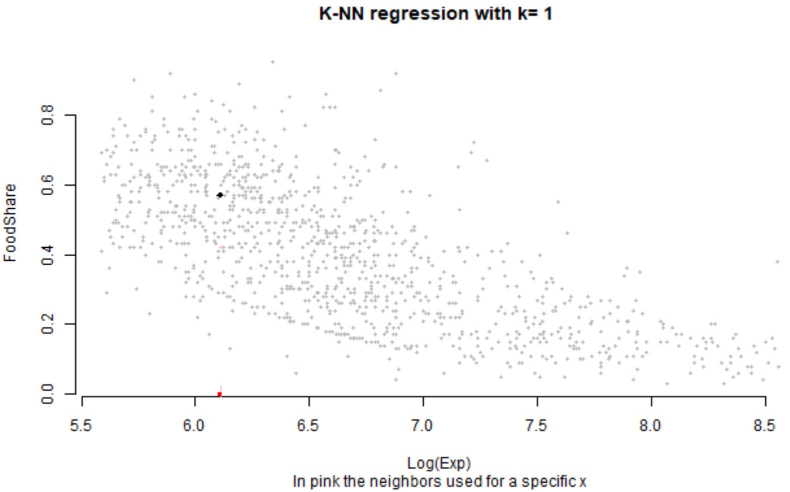
- Linear and polynomial models are determined by parameters $(\beta_0, \beta_2, \cdots, \beta_p)$
- Other methods more flexible
- Nearest neighbors (or k-NN)

  The goal is to estimate $f(\cdot)$ not $\beta_s$!
  Similar in spirit to moving average

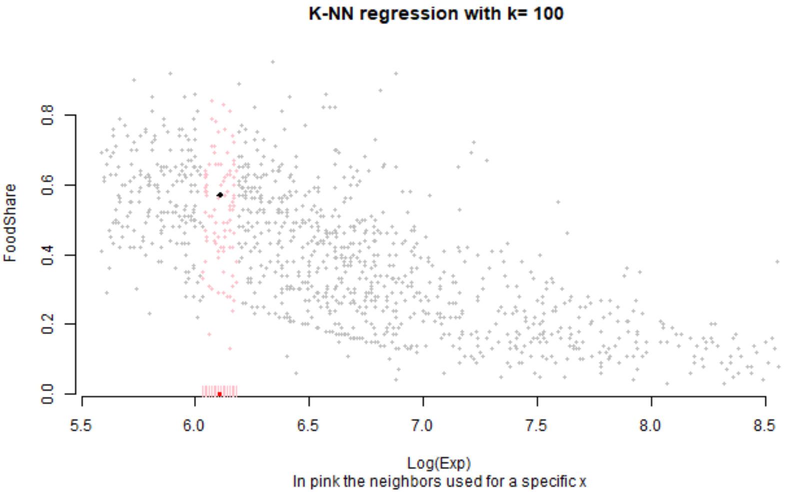
$$\widehat{f}(x_i) = \frac{1}{k} \sum_{\substack{j \in \{k-nearest \\ neighbours\ of\ x_i\}}} y_j$$

  $k$ is the number of neighbors of $x_i$ taken into account in the estimation.
  - The method follows a very general idea:
    "*Observations close in the x dimension should be close in the y dimension*"

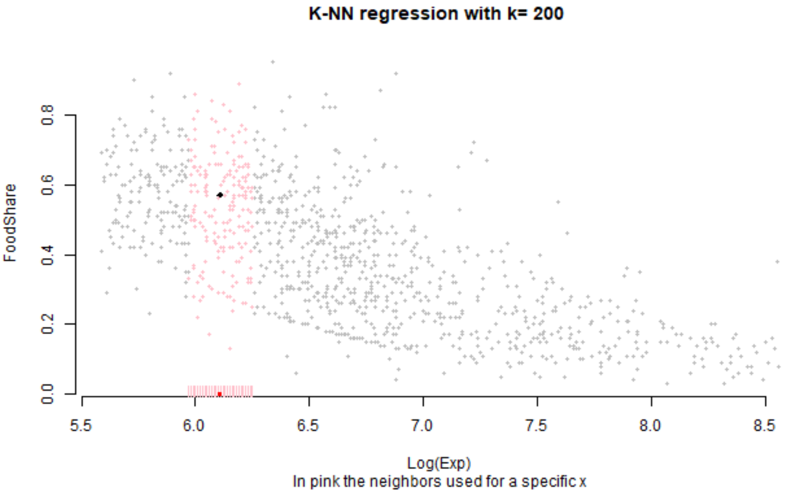
# K-NN IN PRACTICE



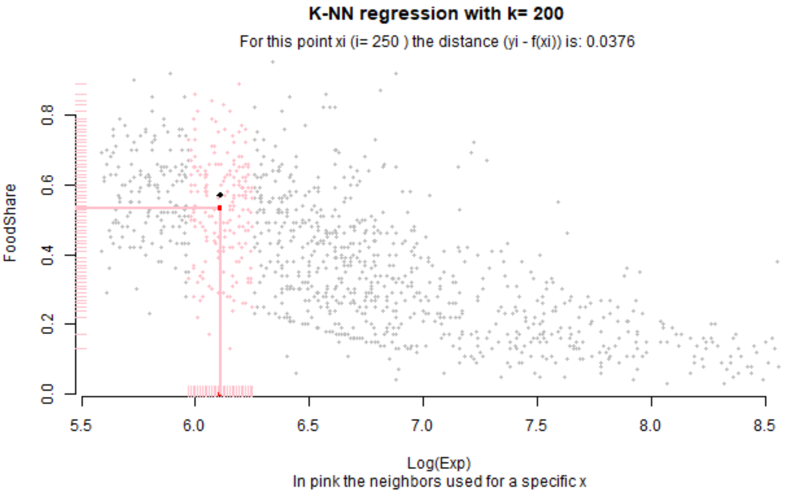
# K-NN in practice



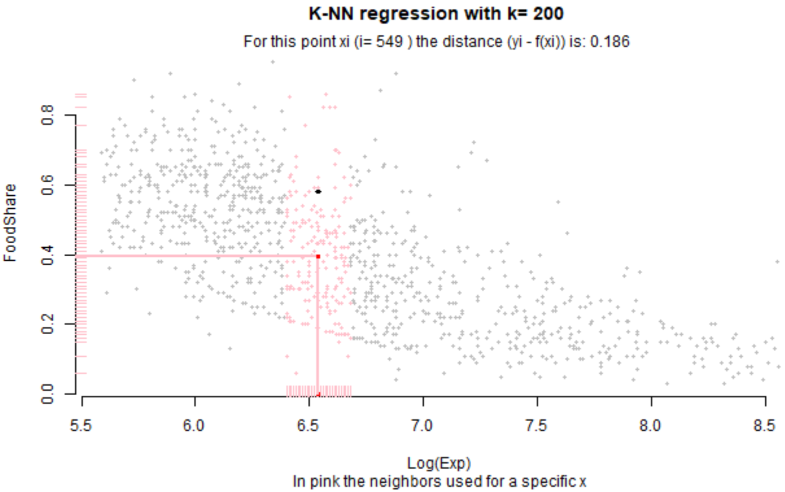
# K-NN in practice



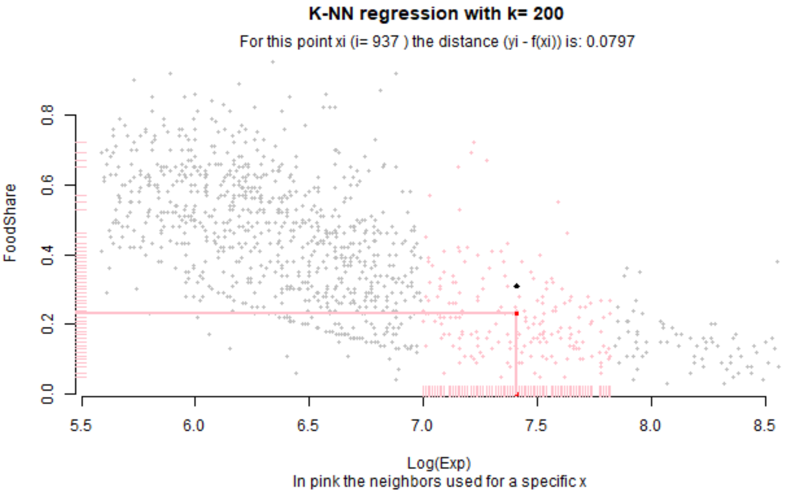
# K-NN IN PRACTICE



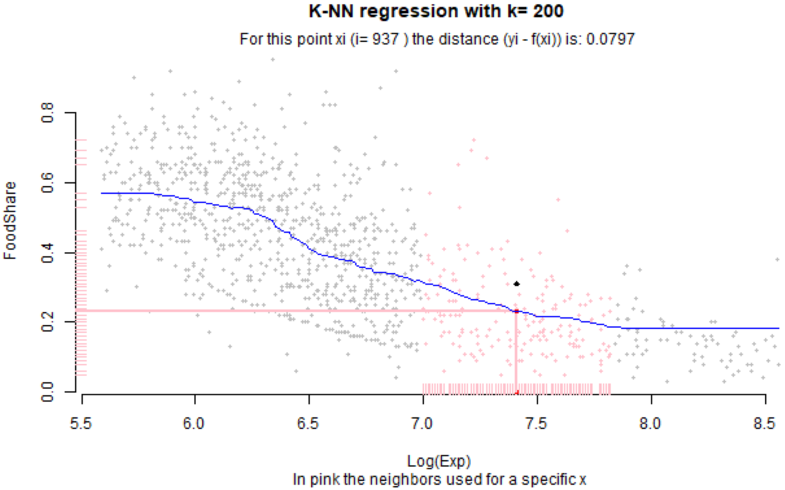
# K-NN in practice



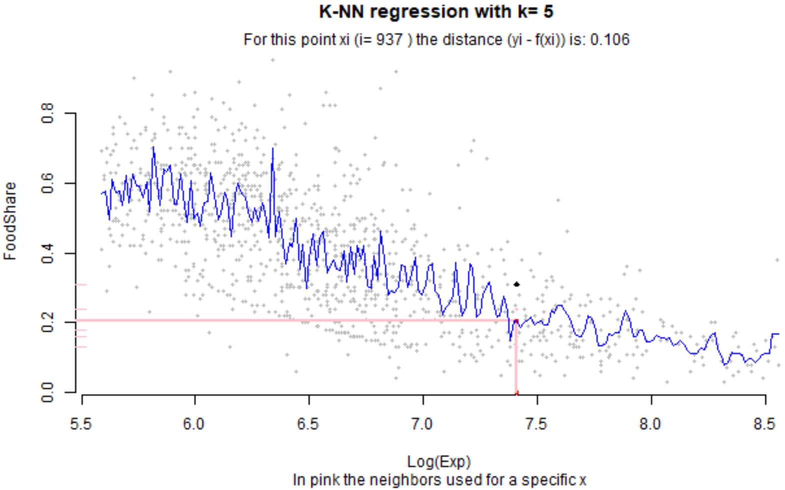
# K-NN IN PRACTICE



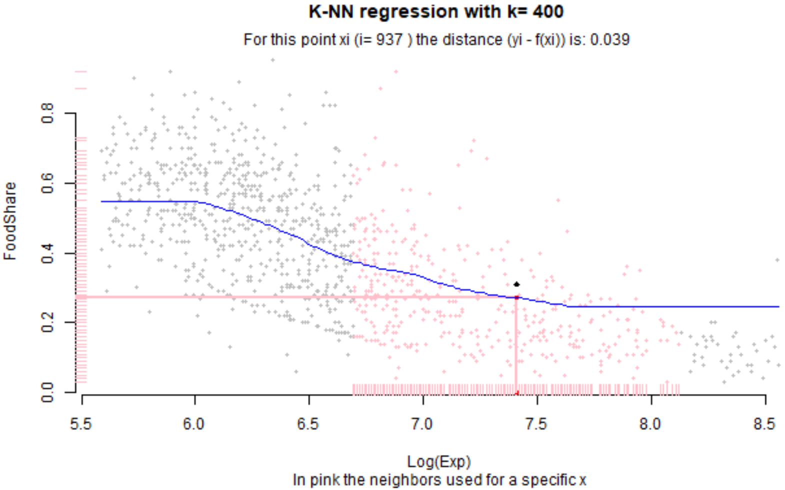
# K-NN in practice



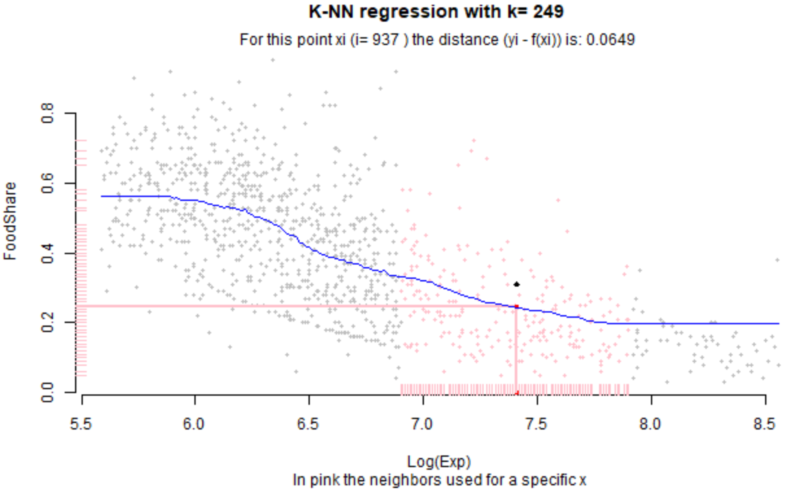
# K-NN in practice: Choosing k



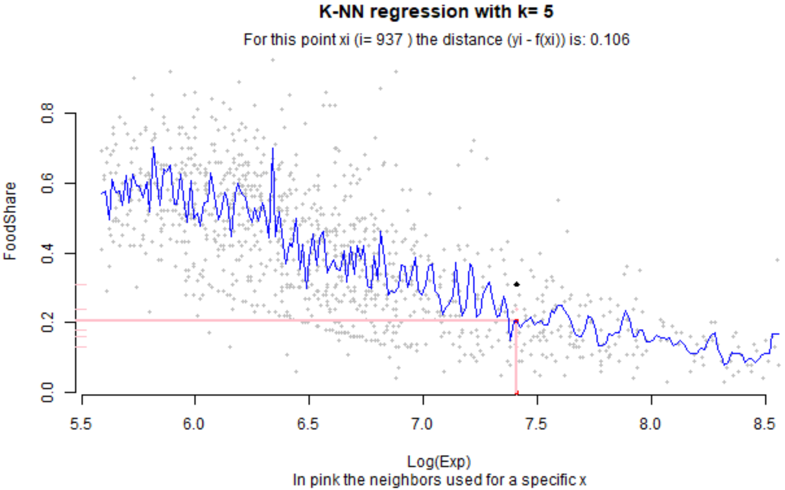
# K-NN in practice: Choosing k



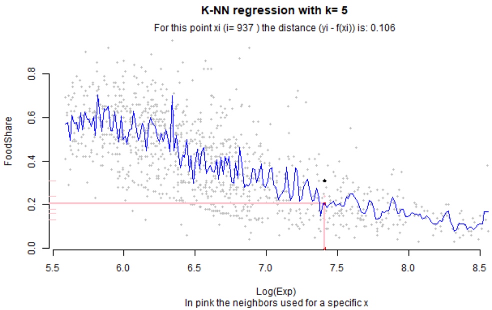
# K-NN IN PRACTICE: CHOOSING K



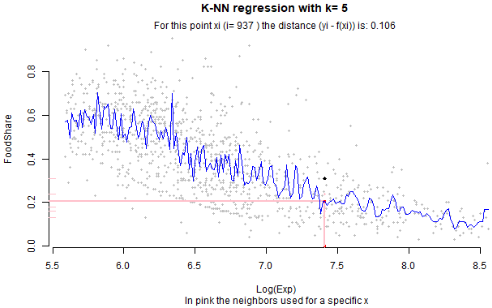
# K-NN in practice: Choosing k



# K-NN in practice: Overfitting

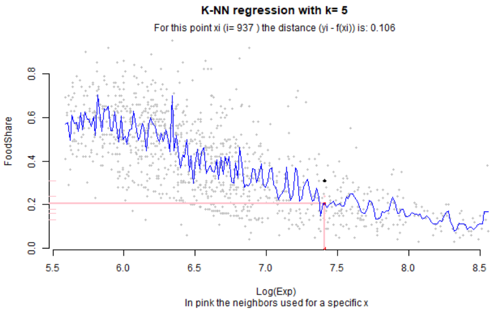


# K-NN in practice: Overfitting




Overfitting has many consequences

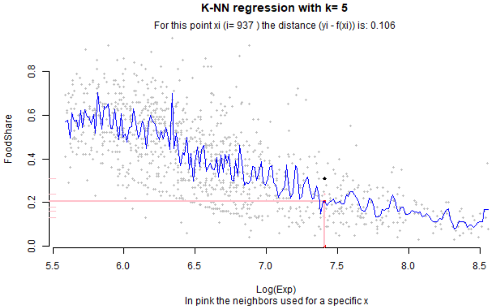
# K-NN in practice: Overfitting




Overfitting has many consequences

- The estimated curve follows the data too closely

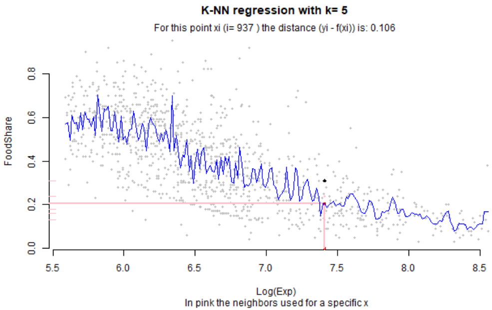
# K-NN in practice: Overfitting




Overfitting has many consequences

- The estimated curve follows the data too closely
- The estimated curve follows the **errors** too closely

# K-NN in practice: Overfitting




Overfitting has many consequences

- The estimated curve follows the data too closely
- The estimated curve follows the **errors** too closely
- The estimated function will not provide good estimates on **new observations**

Machine Learning for Official Statistics and SDGs

# Statistical learning: *vs* Machine Learning

UNITED NATIONS SIAP
Statistical Institute for Asia and the Pacific

# What learning means?

The goal is to estimate $f(\cdot)$

- Many statistical learning methods are relevant and useful to estimate $f(\cdot)$

# What learning means?

The goal is to estimate $f(\cdot)$

- Many statistical learning methods are relevant and useful to estimate $f(\cdot)$
- $f(\cdot)$ can take continuous (regression) or discrete values (classification)

# What learning means?

The goal is to estimate $f(\cdot)$

- Many statistical learning methods are relevant and useful to estimate $f(\cdot)$
- $f(\cdot)$ can take continuous (regression) or discrete values (classification)
- If predicting is the goal, one may focus on prediction accuracy

# WHAT LEARNING MEANS?

The goal is to estimate $f(\cdot)$

- Many statistical learning methods are relevant and useful to estimate $f(\cdot)$
- $f(\cdot)$ can take continuous (regression) or discrete values (classification)
- If predicting is the goal, one may focus on prediction accuracy

  $\hookrightarrow$ this is the goal of **Machine Learning**

# What learning means?

The goal is to estimate $f(\cdot)$

- Many statistical learning methods are relevant and useful to estimate $f(\cdot)$
- $f(\cdot)$ can take continuous (regression) or discrete values (classification)
- If predicting is the goal, one may focus on prediction accuracy
  $\hookrightarrow$ this is the goal of **Machine Learning**
- If the goal is to formalize a model, one may focus on testing statistical properties

# What learning means?

The goal is to estimate $f(\cdot)$

- Many statistical learning methods are relevant and useful to estimate $f(\cdot)$
- $f(\cdot)$ can take continuous (regression) or discrete values (classification)
- If predicting is the goal, one may focus on prediction accuracy

  $\hookrightarrow$ this is the goal of **Machine Learning**
- If the goal is to formalize a model, one may focus on testing statistical properties

  $\hookrightarrow$ this is the goal of **Statistical Learning**

# What learning means?

The goal is to estimate $f(\cdot)$

- Many statistical learning methods are relevant and useful to estimate $f(\cdot)$
- $f(\cdot)$ can take continuous (regression) or discrete values (classification)
- If predicting is the goal, one may focus on prediction accuracy

  $\hookrightarrow$ this is the goal of **Machine Learning**
- If the goal is to formalize a model, one may focus on testing statistical properties

  $\hookrightarrow$ this is the goal of **Statistical Learning**
- In practice we'll use both tools to "*understand the data*"

# What learning means?

The classical approach

- So far, we have estimated $f(\cdot)$ on the whole data set

# What learning means?

The classical approach

- So far, we have estimated $f(\cdot)$ on the whole data set

# What learning means?

## The classical approach

- So far, we have estimated $f(\cdot)$ on the whole data set


- We have estimated $f(\cdot)$ by $\widehat{f}(\cdot)$ and minimized some cost function such as the $MSE = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{f}(x_i)\right)^2$

# What learning means?

The classical approach

- So far, we have estimated $f(\cdot)$ on the whole data set


- We have estimated $f(\cdot)$ by $\widehat{f}(\cdot)$ and minimized some cost function such as the $MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}(x_i)\right)^2$
- The data serve both for estimating $\widehat{f}(\cdot)$ and computing the prediction error

# What learning means?

A different approach: *resampling*

- Our goal is evaluate the prediction accuracy of $\widehat{f}(\cdot)$ on a new, **unseen**, data set

# What learning means?

A different approach: *resampling*

- Our goal is evaluate the prediction accuracy of $\widehat{f}(\cdot)$ on a new, **unseen**, data set
- Since we may not have **unseen** data, we will construct one

# WHAT LEARNING MEANS?

A different approach: *resampling*

- Our goal is evaluate the prediction accuracy of $\widehat{f}(\cdot)$ on a new, **unseen**, data set
- Since we may not have **unseen** data, we will construct one

# What learning means?

A different approach: *resampling*

- Our goal is evaluate the prediction accuracy of $\widehat{f}(\cdot)$ on a new, **unseen**, data set
- Since we may not have **unseen** data, we will construct one


- And the true value of $y_i$ will be available to compute MSE of the prediction

# Why different sets?

Predicting using predictions capability

- When estimating $f(\cdot)$ on the whole data set, over-fitting may occur

# Why different sets?

Predicting using predictions capability

- When estimating $f(\cdot)$ on the whole data set, over-fitting may occur
- The validation set provides a good way to evaluate the prediction capabilities of a model and the prediction error on a new data set

# Why different sets?

Predicting using predictions capability

- When estimating $f(\cdot)$ on the whole data set, over-fitting may occur
- The validation set provides a good way to evaluate the prediction capabilities of a model and the prediction error on a new data set

# Why different sets?

Predicting using predictions capability

- When estimating $f(\cdot)$ on the whole data set, over-fitting may occur
- The validation set provides a good way to evaluate the prediction capabilities of a model and the prediction error on a new data set


- Prediction accuracy (using $\widehat{f}(\cdot)$) is then evaluated on the validation set **only**

# Constructing training & validation sets

In practice, the validation set is not a block

# Constructing training & validation sets

In practice, the validation set is not a block


- The validation set is constructed from a randomly drawn observations.

# Constructing training & validation sets

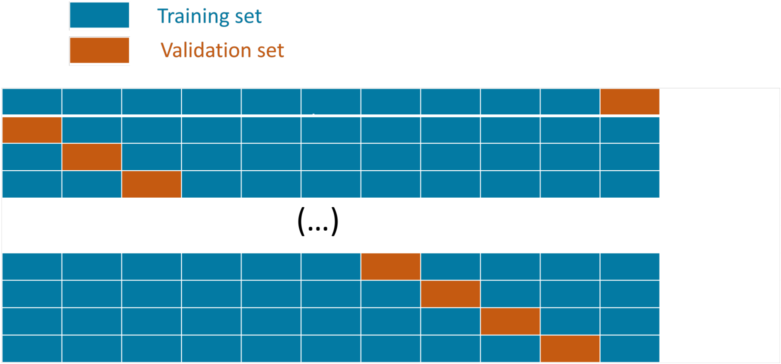
In practice, the validation set is not a block


- The validation set is constructed from a randomly drawn observations.

# *Many* DIFFERENT SETS!

Using resampling methods to estimate the error on the prediction

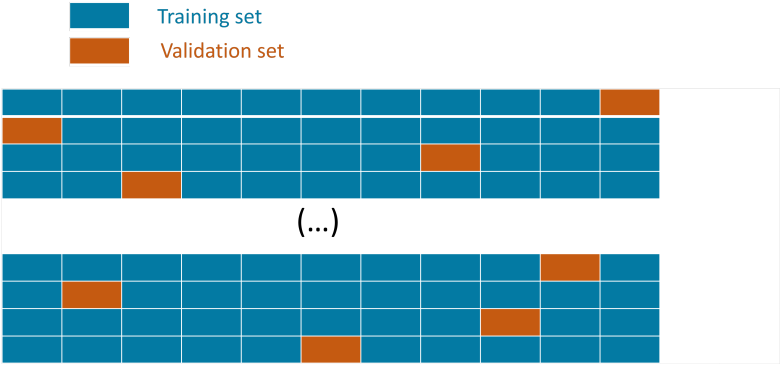
- Cross validation is used to select $m$-(training-validation) sets from the original data set (here again randomly)



# *Many* DIFFERENT SETS!

Using resampling methods to estimate the error on the prediction

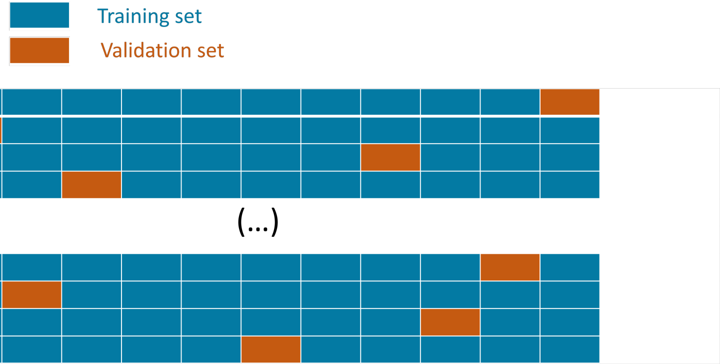
- Cross validation is used to select $m$-(training-validation) sets from the original data set (here again randomly)



# *Many* DIFFERENT SETS!

Using resampling methods to estimate the error on the prediction

- Cross validation is used to select *m*-(training-validation) sets from the original data set (here again randomly)



# *Many* DIFFERENT SETS!

*m-fold* Cross-Validation estimates the average prediction error on $m$ different(training-validation) sets

# *Many* DIFFERENT SETS!

*m-fold* Cross-Validation estimates the average prediction error on $m$ different(training-validation) sets

- For each training-validation) set $i$, one can compute the $MSE_i$ since the true $y_i$ is known on the validation set!

# *Many* DIFFERENT SETS!

*m-fold* Cross-Validation estimates the average prediction error on $m$ different(training-validation) sets

- For each training-validation) set $i$, one can compute the $MSE_i$ since the true $y_i$ is known on the validation set!
- Cross Validation error is then:

$$CV_{(m)} = \frac{1}{m} \sum_{i=1}^{m} MSE_i$$

# *Many* DIFFERENT SETS!

*m-fold* Cross-Validation estimates the average prediction error on $m$ different(training-validation) sets

- For each training-validation) set $i$, one can compute the $MSE_i$ since the true $y_i$ is known on the validation set!
- Cross Validation error is then:

$$CV_{(m)} = \frac{1}{m} \sum_{i=1}^{m} MSE_i$$

- $CV_{(m)}$ is a good estimate of the prediction error of the model

# *Many* DIFFERENT SETS!

*m-fold* Cross-Validation estimates the average prediction error on $m$ different(training-validation) sets

- For each training-validation) set $i$, one can compute the $MSE_i$ since the true $y_i$ is known on the validation set!
- Cross Validation error is then:

$$CV_{(m)} = \frac{1}{m} \sum_{i=1}^{m} MSE_i$$

- $CV_{(m)}$ is a good estimate of the prediction error of the model
- $CV_{(m)}$ can serve to select and compare models (example: select $k$ in k-NN regression)

# *Many* DIFFERENT SETS!

*m-fold* Cross-Validation estimates the average prediction error on $m$ different(training-validation) sets

- For each training-validation) set $i$, one can compute the $MSE_i$ since the true $y_i$ is known on the validation set!
- Cross Validation error is then:

$$CV_{(m)} = \frac{1}{m} \sum_{i=1}^{m} MSE_i$$

- $CV_{(m)}$ is a good estimate of the prediction error of the model
- $CV_{(m)}$ can serve to select and compare models (example: select $k$ in k-NN regression)
- In practice $m \in 5, \cdots, 10$ shows good performances

# Tasks for Machine Learning

Machine Learning involves several tasks, some are time consuming

# Tasks for Machine Learning

Machine Learning involves several tasks, some are time consuming

- Data collection (not treated here)

# Tasks for Machine Learning

Machine Learning involves several tasks, some are time consuming

- Data collection (not treated here)
- Data organization (not treated here)

# Tasks for Machine Learning

Machine Learning involves several tasks, some are time consuming

- Data collection (not treated here)
- Data organization (not treated here)
- Data cleaning (not treated here)

# Tasks for Machine Learning

Machine Learning involves several tasks, some are time consuming

- Data collection (not treated here)
- Data organization (not treated here)
- Data cleaning (not treated here)
- Data visualization

# Tasks for Machine Learning

Machine Learning involves several tasks, some are time consuming

- Data collection (not treated here)
- Data organization (not treated here)
- Data cleaning (not treated here)
- Data visualization
- Data analysis $\leftarrow$ this is the core of this course

# Wrap-up

# Wrap-up

- To understand the data, we can go beyond linearity using polynomial or nonparametric model ($k$-NN)

# Wrap-up

- To understand the data, we can go beyond linearity using polynomial or nonparametric model ($k$-NN)
- More complex models allow for more accuracy, but introduce variance in the estimation

# Wrap-up

- To understand the data, we can go beyond linearity using polynomial or nonparametric model ($k$-NN)
- More complex models allow for more accuracy, but introduce variance in the estimation
- There is a unavoidable **bias-variance** trade-off

# Wrap-up

- To understand the data, we can go beyond linearity using polynomial or nonparametric model ($k$-NN)
- More complex models allow for more accuracy, but introduce variance in the estimation
- There is a unavoidable **bias-variance** trade-off
- Theory helps understanding but not choosing the right model

# Wrap-up

- To understand the data, we can go beyond linearity using polynomial or nonparametric model ($k$-NN)
- More complex models allow for more accuracy, but introduce variance in the estimation
- There is a unavoidable **bias-variance** trade-off
- Theory helps understanding but not choosing the right model
- The train + validation sets approach is central in machine learning