Lecture 2 | 05.03.2024

# Regression and classification

# Regression ...

- Historically, an accidental word invented by Francis Galton (1822 – 1911) because the heights of sons, while following the tendency of their parents (tall parents have tall sons, small parents small sons), tend to return – "regress" – towards the mediocrity/median/average (population stability).
- Nowadays, "regression" is understood as a technique for fitting functional relationships (not necessarily linear, nor parametric) to data (regardless of whether the "slope" is less or greater than 1).
- Mathematically, it describes the relationship between one or more 'input' variable(s) $\boldsymbol{X} \in \mathbb{R}^p$ and an 'output' variable $Y \in \mathbb{R}$. It gives us an equation to predict values (resp. specific characteristics) for the unknown 'output' variable, by plugging in the observed values of the 'input' variables.
- Generally, it is functional relation ship of the form

$$Y = f(\boldsymbol{X}) + error$$

for some well-specified (but somehow still unknown) function $f$ (model) and some unobserved random noise (error, fluctuation respectively).
$\hookrightarrow$ systematic vs. non-systematic part of $Y$...

# Regression model

## General/generic model formulation

$$Y = f(\boldsymbol{X}) + \varepsilon$$

- $Y \in \mathbb{R}$ is a random variable, the covariate of interest (dependent variable)
- $X \in \mathbb{R}$ (or $\boldsymbol{X} = (X_1, \dots, X_p)^\top \in \mathbb{R}^p$) is a random variable (or a random vector respectively) which represents the explanatory information
- $f(\cdot)$ is a measurable regression function from the domain of $X$ (or $\mathbb{X}$ respectively) to the domain of $Y$ – the systematic part
- $\varepsilon$ represents the irreducible error – even if we observe a given realization of $X$ and we know $f(\cdot)$ apriori, there will be some uncertainty left, because $Y$ is a random variable – the non-systematic part
- instead of "predicting" one specific value of $Y$ with the regression function $f(\cdot)$ and the observed realization "$X = x$" we would like to rather estimate some (useful) characteristic of the whole distribution of possible values for $Y$ when (conditionally on) "$X = x$"

# Principal roles of the regression

Regression models and data smoothing techniques (e.g., moving averages, weighted averages, splines, parametric smoothing, Whittaker-Henderson) are essentially very similar but there is at least one principal and crucial difference – while the data smoothing techniques just smooth the empirical data the regression methods goes beyond as they try to learn important facts about the unknown population – the theoretical model behind the data.

# Principal roles of the regression

Regression models and data smoothing techniques (e.g., moving averages, weighted averages, splines, parametric smoothing, Whittaker-Henderson) are essentially very similar but there is at least one principal and crucial difference – while the data smoothing techniques just smooth the empirical data the regression methods goes beyond as they try to learn important facts about the unknown population – the theoretical model behind the data.

- **Goal #1**
  with a good choice of the model (i.e., the regression function $f(\cdot)$) we can use the information contained in $\boldsymbol{X}$ (the explanatory variable) to say something relevant about $Y$ (the dependent variable) **But why do we want to do so?**

- **Goal #2**
  if the set of the explanatory variables is relatively very rich, it can be useful to say which components of $\boldsymbol{X} = (X_1, \dots, X_p)^\top \in \mathbb{R}^p$ are relevant (play a role) in the relationship between $Y$ and $\boldsymbol{X}$ **Why to select if we can use all?**

- **Goal #3**
  once we know which information in $\boldsymbol{X} = (X_1, \dots, X_p)^\top$ has an impact on the values of $Y$ it is often of interest to quantify this effect – to evaluate how a specific component of $\boldsymbol{X}$ affects the value of $Y$ **Why is this useful in practice?**

# General regression setup

- Generic random vector $(Y, \boldsymbol{X}^\top)^\top$ with some joint distribution $F_{Y,\boldsymbol{X}}(y, \boldsymbol{x})$
- Generic (population) model: $Y = f(\boldsymbol{X}) + \varepsilon$

# General regression setup

- Generic random vector $(Y, \boldsymbol{X}^\top)^\top$ with some joint distribution $F_{Y,\boldsymbol{X}}(y, \boldsymbol{x})$
- Generic (population) model: $Y = f(\boldsymbol{X}) + \varepsilon$

- Random sample from the population: $\{(Y_i, \boldsymbol{X}_i);\ i = 1, \dots, n\}$ for $n \in \mathbb{N}$
- Empirical/data model: $Y_i = f(\boldsymbol{X}_i) + \varepsilon_i$ for $i = 1, \dots, n$

# General regression setup

- Generic random vector $(Y, \boldsymbol{X}^\top)^\top$ with some joint distribution $F_{Y,\boldsymbol{X}}(y, \boldsymbol{x})$
- Generic (population) model: $Y = f(\boldsymbol{X}) + \varepsilon$

- Random sample from the population: $\{(Y_i, \boldsymbol{X}_i);\ i = 1, \dots, n\}$ for $n \in \mathbb{N}$
- Empirical/data model: $Y_i = f(\boldsymbol{X}_i) + \varepsilon_i$ for $i = 1, \dots, n$

- What is known: dependent observations $Y_i$ and explanatory variable(s) $\boldsymbol{X}_i$
- What is unknown: random errors $\varepsilon_i$ and the regression function $f(\cdot)$

# General regression setup

- Generic random vector $(Y, \boldsymbol{X}^\top)^\top$ with some joint distribution $F_{Y,\boldsymbol{X}}(y, \boldsymbol{x})$
- Generic (population) model: $Y = f(\boldsymbol{X}) + \varepsilon$

- Random sample from the population: $\{(Y_i, \boldsymbol{X}_i);\ i = 1, \dots, n\}$ for $n \in \mathbb{N}$
- Empirical/data model: $Y_i = f(\boldsymbol{X}_i) + \varepsilon_i$ for $i = 1, \dots, n$

- What is known: dependent observations $Y_i$ and explanatory variable(s) $\boldsymbol{X}_i$
- What is unknown: random errors $\varepsilon_i$ and the regression function $f(\cdot)$

- **Typical assumptions:**
  - the error terms (unobserved fluctuations or disturbances respectively) have a zero mean and some finite (typically unknown) variance $\sigma^2 > 0$
  - the unknown regression function $f(\cdot)$ is expected to belong to some well specified class of functions

# General regression setup

- Generic random vector $(Y, \boldsymbol{X}^\top)^\top$ with some joint distribution $F_{Y,\boldsymbol{X}}(y, \boldsymbol{x})$
- Generic (population) model: $Y = f(\boldsymbol{X}) + \varepsilon$

- Random sample from the population: $\{(Y_i, \boldsymbol{X}_i);\ i = 1, \dots, n\}$ for $n \in \mathbb{N}$
- Empirical/data model: $Y_i = f(\boldsymbol{X}_i) + \varepsilon_i$ for $i = 1, \dots, n$

- What is known: dependent observations $Y_i$ and explanatory variable(s) $\boldsymbol{X}_i$
- What is unknown: random errors $\varepsilon_i$ and the regression function $f(\cdot)$

- **Typical assumptions:**
  - the error terms (unobserved fluctuations or disturbances respectively) have a zero mean and some finite (typically unknown) variance $\sigma^2 > 0$
  - the unknown regression function $f(\cdot)$ is expected to belong to some well specified class of functions
  - and possibly others... (depending on the specific model formulations)

# Conditional distribution of $Y$

- for different values of the independent variable $X$ the possible values of the dependent variable $Y$ may have different distribution
  $\Longrightarrow$ conditional distribution of $Y$ given "$X = x$"
  *(an analogy to a $K \in \mathbb{N}$ sample problem, for $K \longrightarrow \infty$)*
- infinitely many characteristics can be used to characterize the (conditional) distribution of $Y$ (given $X$)... Which are good/ideal ones?
- the answer usually depends on the criterion we choose to measure the quality of the model/fit – the so-called "goodness-of-fit" criterion

# Conditional distribution of $Y$

- for different values of the independent variable $X$ the possible values of the dependent variable $Y$ may have different distribution
  $\Longrightarrow$ conditional distribution of $Y$ given "$X = x$"
  *(an analogy to a $K \in \mathbb{N}$ sample problem, for $K \longrightarrow \infty$)*
- infinitely many characteristics can be used to characterize the (conditional) distribution of $Y$ (given $X$)... Which are good/ideal ones?
- the answer usually depends on the criterion we choose to measure the quality of the model/fit – the so-called "goodness-of-fit" criterion

  - **Mean squared error** (as a theoretical functional)
  $$min_f E[Y - f(X)]^2$$
  - **Least squares** (as an empirical counterpart)
  $$min_f \frac{1}{n} \sum_{i=1}^{n} [Y_i - f(X_i)]^2$$

$\hookrightarrow$ where both minimization problems are taken with respect to some well-defined class of regression functions $f$ (note the analogy between the theoretical mean and its empirical estimate – the average)

# Estimation of the regression function

- **Two sample problem**
  if $X$ only takes two values (e.g., $X = \pm 1$), the observations (random sample) $(Y_i, X_i)$ for $i = 1, \dots, n$ can be split into two parts – values of Y for which $X_i = -1$ and the values of $Y$ for which $X = 1$ – and a simple average is calculated in both groups

- **Multiple samples**
  if $X$ takes finitely many different values (e.g., $X$ is a categorical variable with $K \in \mathbb{N}$ levels), the random sample $(Y_i, X_i)$ for $i = 1, \dots, n$ can be split into $K$ disjoint groups and, again, simple averages can be calculated for each of $K$ groups

- **Continuous explanatory variable**
  if $X$ is a continuous variable (taking infinitely/uncountable many values), the sample can not be split into all possible groups – for very many "$X = x$" there will be simply no observations of $Y$ available
  $\Longrightarrow$ **borrowing power from the neighbors**

# From local techniques to parametric ones (or vice versa?)

- **Nonparametric regression techniques**
  - the conditional distribution of $Y$ given $X = x$ estimated locally for $x \in \mathbb{R}$
  - very flexible technique, adapts to any functional form of $f(\cdot)$
  - the number of unknown parameters to be estimated is large ($\rightarrow \infty$)
  - the amount of flexibility is an important aspect to control for

- **Parametric regression techniques**
  - a limited class of functions is used, the class depends on some parameters
  - the number of unknown parameter is relatively small (and fixed)
  - the flexibility of the model is determined by the analytical form of $f(\cdot)$
  - in many cases straightforward and relatively simple interpretation

# From local techniques to parametric ones (or vice versa?)

- **Nonparametric regression techniques**
  - the conditional distribution of $Y$ given $X = x$ estimated locally for $x \in \mathbb{R}$
  - very flexible technique, adapts to any functional form of $f(\cdot)$
  - the number of unknown parameters to be estimated is large ($\to \infty$)
  - the amount of flexibility is an important aspect to control for

- **Parametric regression techniques**
  - a limited class of functions is used, the class depends on some parameters
  - the number of unknown parameter is relatively small (and fixed)
  - the flexibility of the model is determined by the analytical form of $f(\cdot)$
  - in many cases straightforward and relatively simple interpretation

- **Semi-parametric regression techniques**
  - a bridge between parametric and non-parametric methods
  - the idea is to select positive properties from both
  - negative properties are, however, inherited accordingly
  - still very popular in practical applications and theoretical developments

# Some trade-offs to keep in mind

- **Mathematics:** parsimonious models vs. "black-box" algorithms
  *(transparent models are tractable by mathematical theory)*
- **Probability:** bias vs. variability of the estimate
  *(small bias means better accuracy, large variance means high uncertainty)*
- **Utilization:** prediction purposes vs. explanation of the relationship
  *(different models are build depending on the primary purpose)*
- **Computation:** computational tractability and time efficiency
  *(machine limitations in algorithmic computations does not allow for an arbitrary model)*
- **Interpretation:** simple models are easy to interpret but less accurate
  *complex models are very difficult (impossible) to reasonable explain*

# Some trade-offs to keep in mind

- **Mathematics:** parsimonious models vs. "black-box" algorithms
  *(transparent models are tractable by mathematical theory)*
- **Probability:** bias vs. variability of the estimate
  *(small bias means better accuracy, large variance means high uncertainty)*
- **Utilization:** prediction purposes vs. explanation of the relationship
  *(different models are build depending on the primary purpose)*
- **Computation:** computational tractability and time efficiency
  *(machine limitations in algorithmic computations does not allow for an arbitrary model)*
- **Interpretation:** simple models are easy to interpret but less accurate
  *complex models are very difficult (impossible) to reasonable explain*

*"All models are wrong, but some are useful!"*

George E. P. Box (1919 – 2013)

# Model accuracy

Let's assume that for the (generic) model $Y = f(X) + \varepsilon$ we obtained the estimated $\hat{f}(\cdot)$ based on the random sample $\{(Y_i, X_i);\ i = 1, \dots, n\}$

How to access the model quality (its accuracy) quantitatively?

- Using the "training data" $\{(Y_i, X_i);\ i = 1, \dots, n\}$
- Using a fresh "testing data" $\{(Y_i, X_i);\ i = n + 1, \dots, N\}$

# Model accuracy

Let's assume that for the (generic) model $Y = f(X) + \varepsilon$ we obtained the estimated $\hat{f}(\cdot)$ based on the random sample $\{(Y_i, X_i);\ i = 1, \dots, n\}$

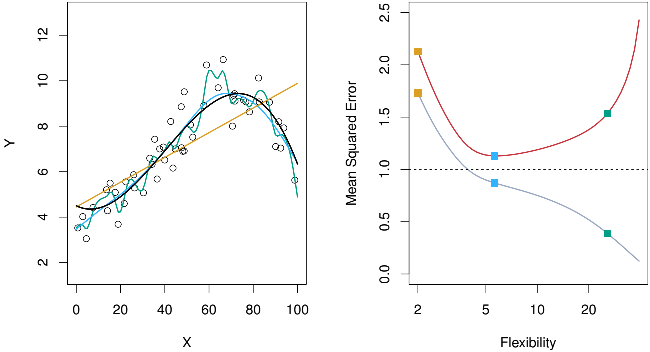
How to access the model quality (its accuracy) quantitatively?

- Using the "training data" $\{(Y_i, X_i);\ i = 1, \dots, n\}$
- Using a fresh "testing data" $\{(Y_i, X_i);\ i = n + 1, \dots, N\}$

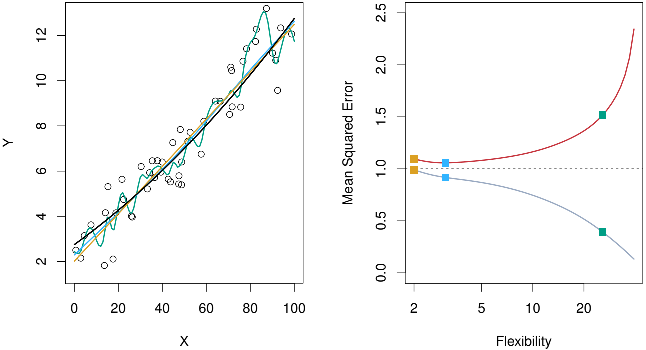
How to access the model quality (its accuracy) qualitatively?

- Using mathematical/stochastic theory and various statistical tools
- Using expert knowledge, previous experience, common sense, etc.

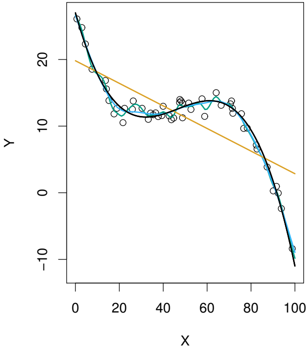
# Model prediction error – Example I

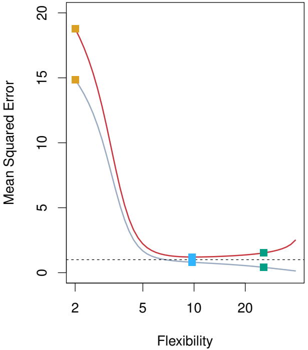


# Model prediction error – Example II



# Model prediction error – Example III

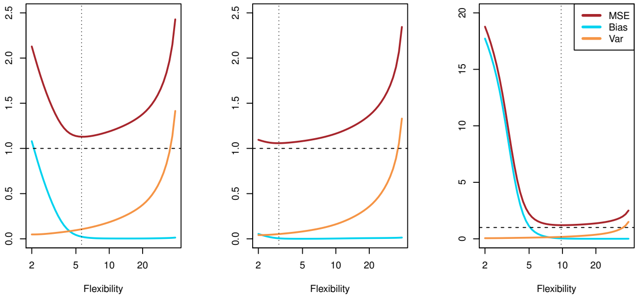


# Bias-variance Trade-Off

**Mean Squared Error (MSE):**

$$
\begin{aligned}
E[Y-\hat{f}(X)]^2 &= E[(f(X)+\varepsilon - E\hat{f}(X)) - (\hat{f}(X) - f(X))]^2 \\
&= E[\hat{f}(X) - E\hat{f}(X)]^2 + \Big(E\hat{f}(X) - f(X)\Big)^2 + E\varepsilon^2 \\
&= Var\ \hat{f}(X) + \Big(Bias\ \hat{f}(X)\Big)^2 + Var\ \varepsilon
\end{aligned}
$$



# Optimal model

- again, there are many different approaches to say which model is a good one (optimal one, useful or practical one, ...)
- in terms of the bias-variance trade-off the optimal model is the one that minimizes the mean squared error criterion
- the minimization of the mean squared criterion results in the minimization of the expected square of the error term
- in applications, instead of the theoretical (generic) error term $\varepsilon$ we work with the empirical residual terms (residuals respectively)
- instead of minimizing the MSE criterion, we minimize the sum of the squared residuals (i.e., empirical estimate for MSE)

# More general: Regression vs. Classification

- what is the nature of the input variable(s) $\boldsymbol{X} \in \mathbb{R}^p$?
- what is the nature of the output variable $Y \in \mathbb{R}$?

# More general: Regression vs. Classification

- what is the nature of the input variable(s) $\boldsymbol{X} \in \mathbb{R}^p$?
- what is the nature of the output variable $Y \in \mathbb{R}$?

- ordinary linear regression model
- analysis of variance
- classification
- contingency table

# Vague motivation of the classification problem

- if the response variable $Y$ is qualitative and the explanatory variables $\boldsymbol{X} = (X_1, \dots, X_p)^\top$ are continuous/discrete/mixed we are (typically) dealing with a classification problem

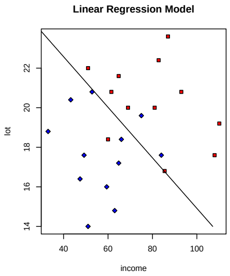
- the goal is to use the information in $\boldsymbol{X}$ to assign a classification label for $Y$ (to decide into which category it belongs)

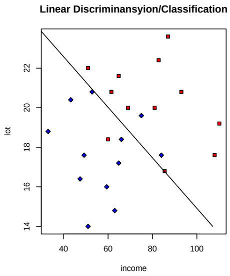
- the "goodness-of-fit" in classification problems is (commonly) measured by a missclassification error rate $\sum_{i=1}^{n} \mathbb{I}_{\{Y_i \notin \widehat{C}(\boldsymbol{X}_i)\}}$

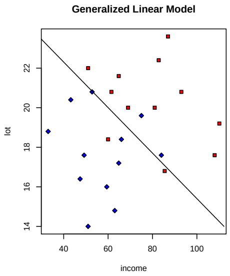
- the value $\widehat{C}(\boldsymbol{X}_i) \in \{1, \dots K\}$ is the assigned classification label which typically maximize the corresponding posterior probability
  $\hookrightarrow$ so called the Bayes classification rule

# Classification vs. regression problem

- Mowers data: $\{(Y_i, X_{i1}, X_{i2})^\top;\ i = 1, \ldots, 24\}$ Model: $Y_i \sim X_{i1} + X_{i2}$




Model 1

Model 2

Model 3

$$E[Y|X_1, X_2] = \alpha + \beta_1 X_1 + \beta_2 X_2$$

$$C(X_1, X_2) = \begin{cases} +1 & \text{if } \beta_1 X_1 + \beta_2 X_2 > \frac{\mu_1+\mu_2}{2} \\ -1 & \text{if } \beta_1 X_1 + \beta_2 X_2 < \frac{\mu_1+\mu_2}{2} \end{cases}$$

$$\log \frac{P[Y=1|X_1, X_2]}{1-P[Y=1|X_1, X_2]} = \alpha + \beta_1 X_1 + \beta_2 X_2$$

# Generalized regression models

- considering the model $Y = f(X) + \varepsilon$ and the support of the dependent variable $Y$ which is limited/bounded/finite, it is not reasonable to assume, for instance, linear/unbounded/continuous function $f(\cdot)$ in the model...
- on the other hand, recall that model expressed as $Y = f(X) + \varepsilon$ and $E[Y|X = x] = f(x)$ are, actually (under some mild assumptions), two equivalent (linear regression) model formulations
- even discrete distribution of $Y$ can be well-specified by some continuous characteristic – e.g., some probability parameter $p \in (0, 1)$
- how to mathematically formalize a regression model in such situations?
  $\hookrightarrow$ generalized (linear) regression models $g(E[Y = 1|X = x]) = f(x)$
- what are the analogies with the regression model $Y = f(X) + \varepsilon$?