Allianz

High-dimensional statistics and Machine learning with applications to Insurance

November 2022
Masaryk University, Brno

Ivana Milović, MAS PhD

Introducing Myself

Ivana Milović, MAS PhD

Non-Life Pricing Actuary (SME)
ivana.milovic@allianz.at

Prior experience

- Uniqa Insurance Group – Non-Life Pricing Actuary (Motor)
- Lecturer - University of Vienna
- Prae and Post-Doc Researcher - Department of Statistics,
  University of Vienna

Education

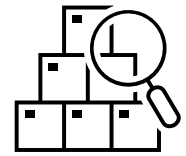
- PhD in Statistics (Univ. of Vienna, 2016)
- Master of Advanced Studies in Mathematics (Univ. of Cambridge, 2011)
- BSc in Mathematics and Computer Science (Univ. of Belgrade, 2010)

What is pricing?

Allianz

What is pricing?

“Pricing is the way that a company decides prices for its products or services, or the prices decided” – Cambridge dictionary

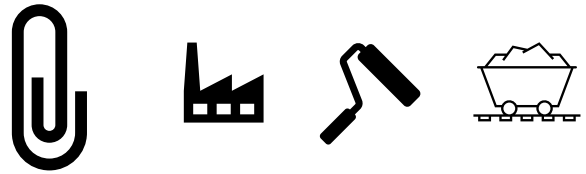


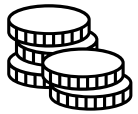
€

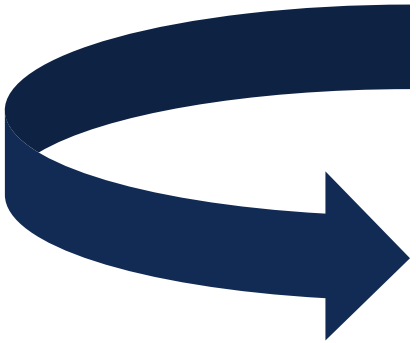
Why do we need statistics and mathematical modelling for pricing in insurance?

Classical industry example: Selling paperclips

- Known operating costs (rent, maintenance, salaries, marketing, etc.)
- Known production costs (materials, etc.)
- Known profit margin







Known price of a paperclip

Fully deterministic!

Why do we need statistics and mathematical modelling for pricing in insurance?

Classical insurance example: Selling a policy

- Known operating costs (rent, maintenance, salaries, marketing, etc.)
- Unknown claim costs (claim occurrence and severity are random events)
- Known profit margin

Unknown price of a policy

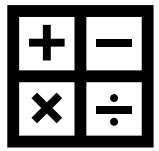
Not deterministic!

Why do we need statistics and mathematical modelling for pricing in insurance?

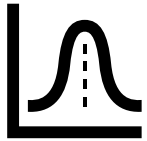
If the cost of policy is random, how do we estimate it?

There are two ways:

- Based on the historical data/expert judgement (simplistic approach)



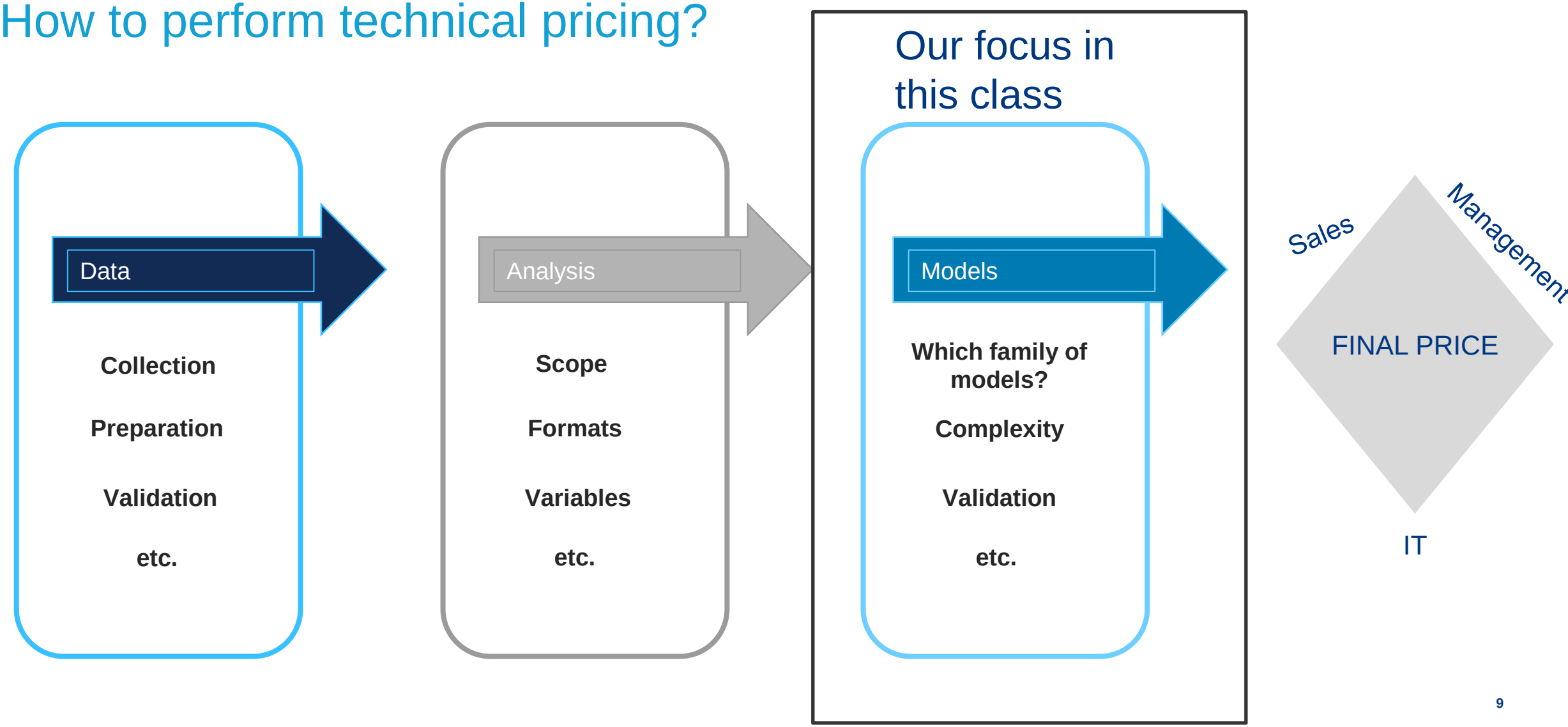
- Fitting statistical models to historical data -> technical pricing.



What are the goals of technical pricing?

- To provide the best estimate for the expected cost of an insurance policy -> fair price
- Help us predict future losses and to better assess the portfolio and segment performance
- Know which are the technically unprofitable and profitable segments -> Identify business opportunities

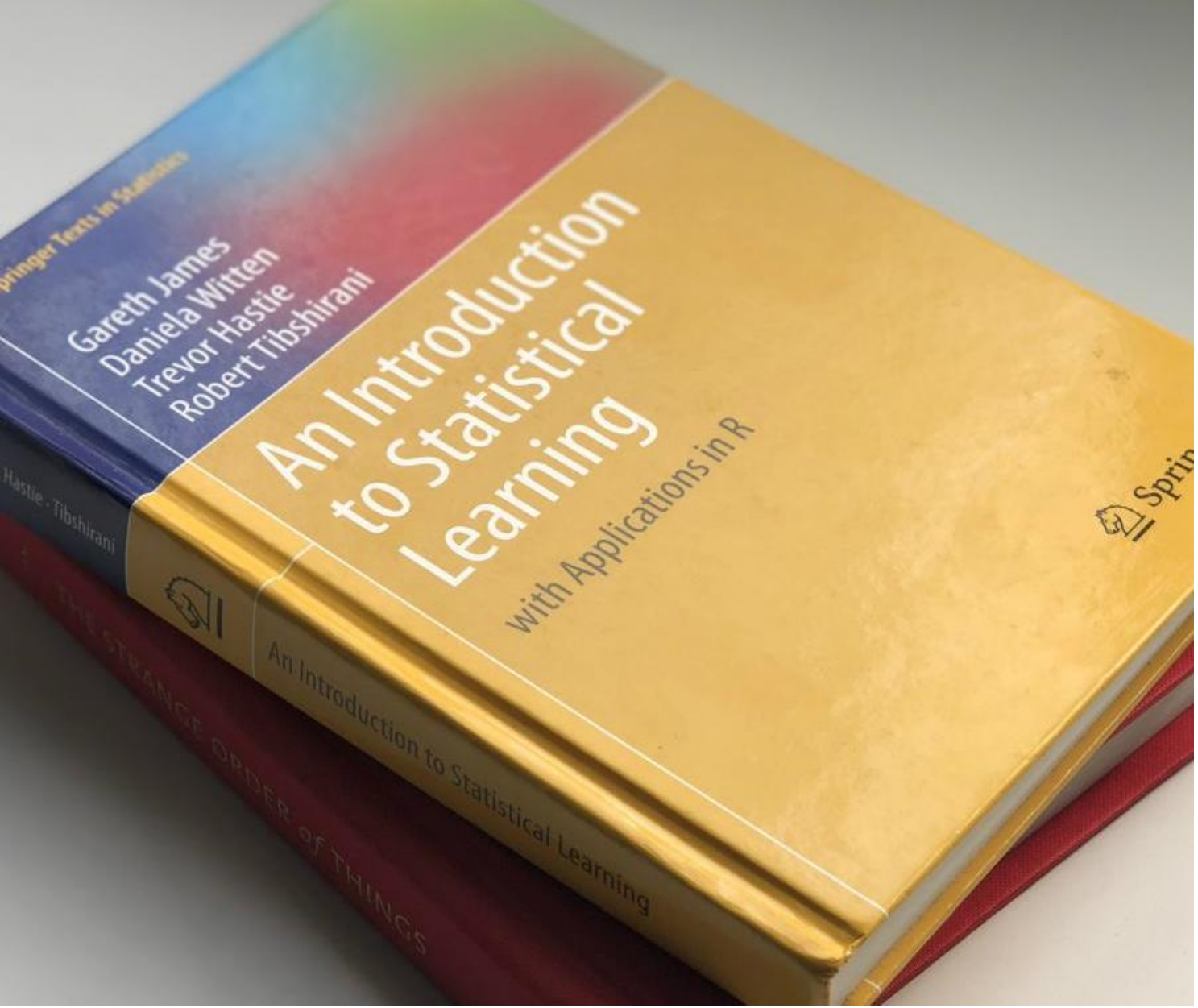
How to perform technical pricing?



Content

Topics

- Model assessment and selection
- Cross validation, AIC, BIC
- Linear Models
- PCR, Regularization methods
- Generalized Linear models
- Pricing process
- Machine Learning in Insurance



Allianz

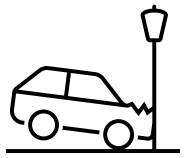
Let’s get started

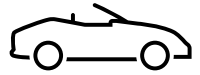
Introduction

Introduction

- Let Y be a quantitative response and $X = (X_1, \dots, X_p)$ be a set of regressors and suppose: $Y = f(X) + \epsilon$, for some **fixed (but unknown) function f** .
- ϵ has mean 0 and is **independent** of X. Often, we assume normality.
- Note: X can be fixed or random

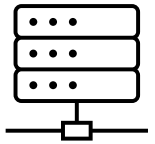
Example: Y is the number of claims and X are the characteristics of a driver and his car





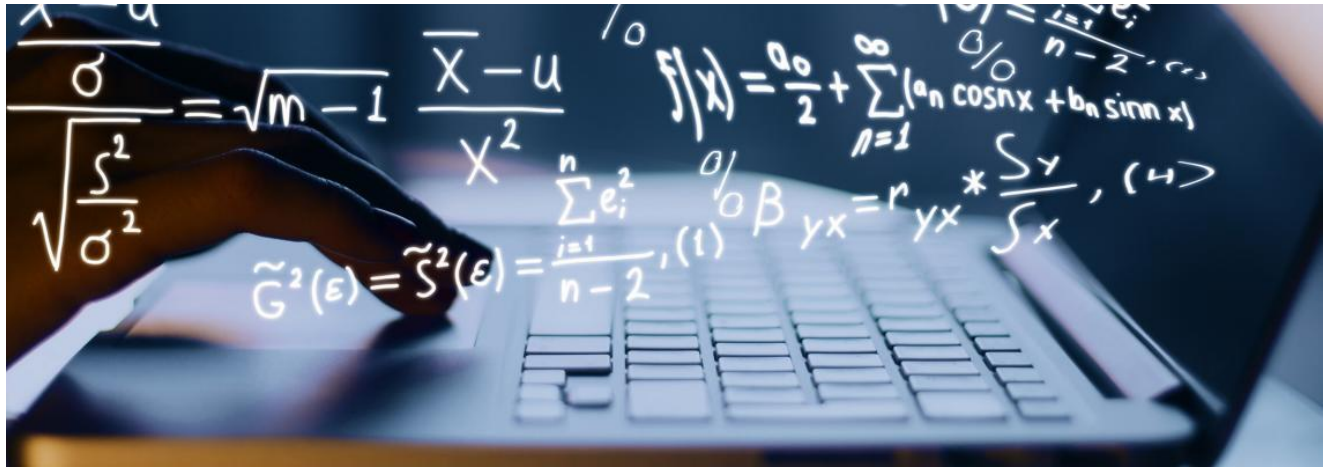
Introduction

Statistical learning is a set of approaches for estimating f by $\hat{f}$ from the data.



Estimation goals can be:

- Prediction
- Inference



This Photo CC BY-SA

Allianz

Introduction

Prediction: $\hat{Y} = \hat{f}(X)$, for some estimate $\hat{f}$.

If prediction is our only goal and we do not have interest in the form of f, then many modern techniques give good results: random forests, gradient boosting trees, etc.

Example: predicting prices on the stock exchange. Here the interpretation is not important, as long as, the results are good.

Allianz

Introduction

- The accuracy of $\hat{Y}$ depends on two quantities:
  - reducible error – coming from approximating f by $\hat{f}$
  - irreducible error – the error coming from ϵ
- We measure the accuracy by the **expected prediction error**
- $E(Y-\hat{Y})^2 = \underbrace{E(f(X)-\hat{f}(X))^2}_{\text{reducible}} + \underbrace{Var(\epsilon)}_{\text{irreducible}}$
- Goal: to find a method that has a small reducible error

Allianz

Introduction

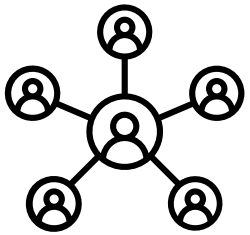
- Note that

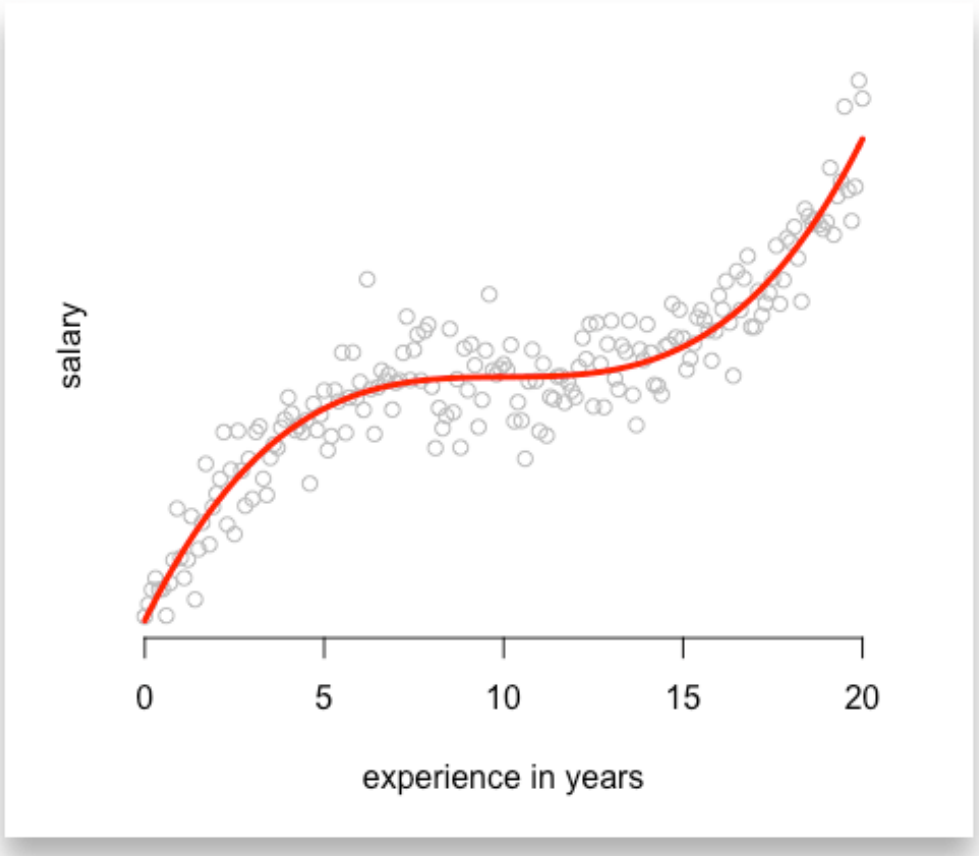
$$\begin{aligned}\mathbb{E}(Y-\hat{Y})^2 &= \mathbb{E}(f(X)+\epsilon-\hat{f}(X))^2 \\ &= \mathbb{E}(f(X)-\hat{f}(X))^2+\mathbb{E}(\epsilon^2) \\ &+ 2\mathbb{E}\left[(f(X)-\hat{f}(X))\epsilon\right] \\ &= \mathbb{E}(f(X)-\hat{f}(X))^2 + \mathrm{Var}(\epsilon)\end{aligned}$$

Introduction

Inference: we want to also understand the form of f, i.e. the relationship between Y and $X = (X_1, \dots, X_p)$.

- Is f linear or more complex?
- Which regressors are associated with Y?
- What is their relationship?





Choice of Model

Allianz

Choice of Model

We may choose our model based on what we are more interested in: prediction or inference

Example:

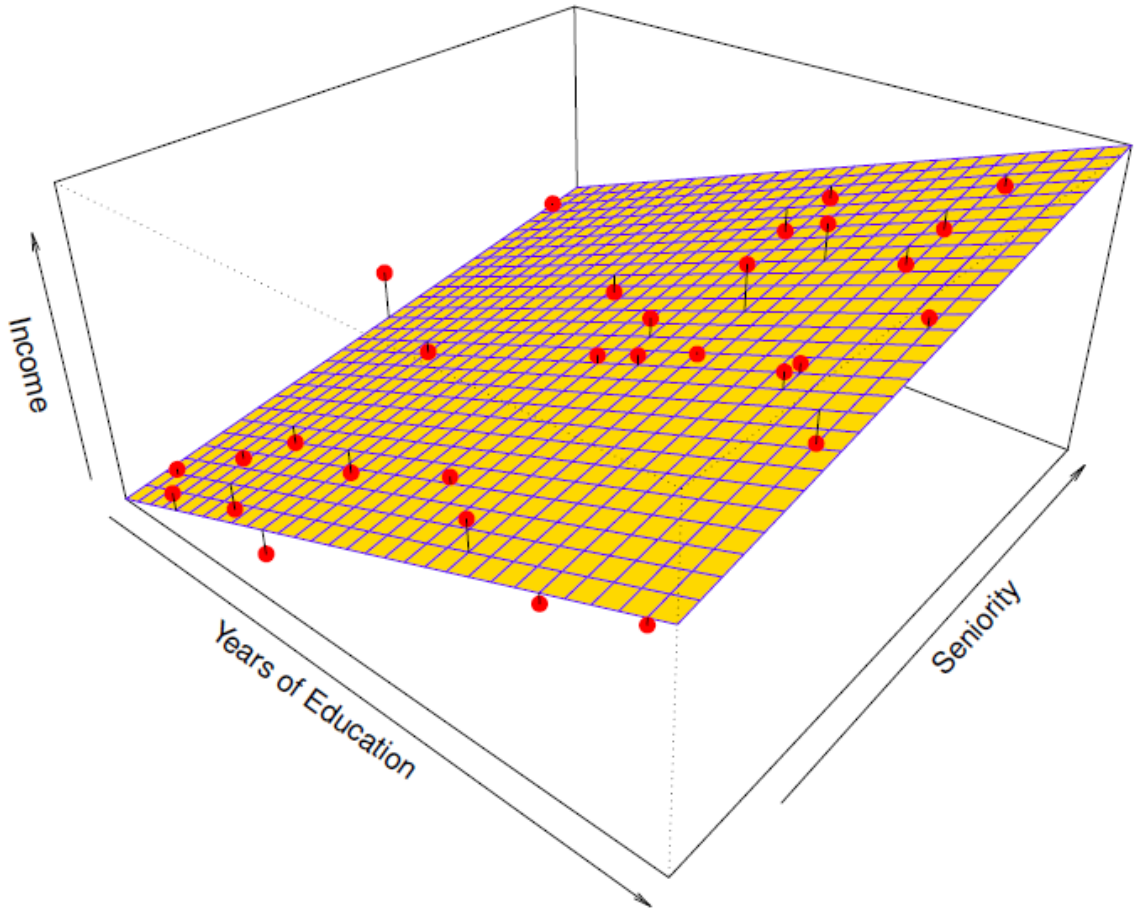
- **Parametric** models like linear models and GLMs: **simple and interpretable**, but not always very accurate
- **Non-parametric** models like splines, GBM, random forests: better predictions but **much less interpretable**

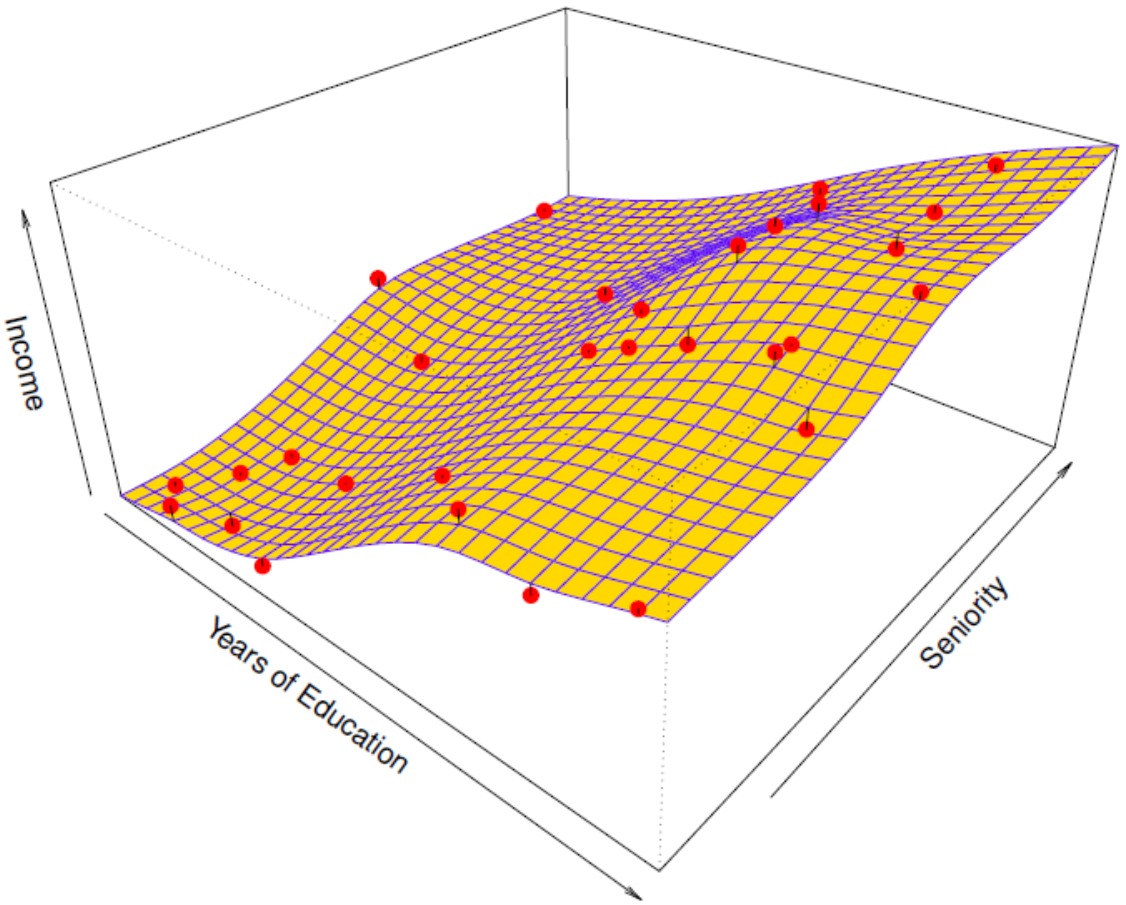
Factors like sample size, computational power, etc. also play a significant role in decision making.

EXAMPLE: LINEAR REGRESSION VS. SPLINES

Allianz

Choice of Model





Interpretability?

Controversies

Allianz

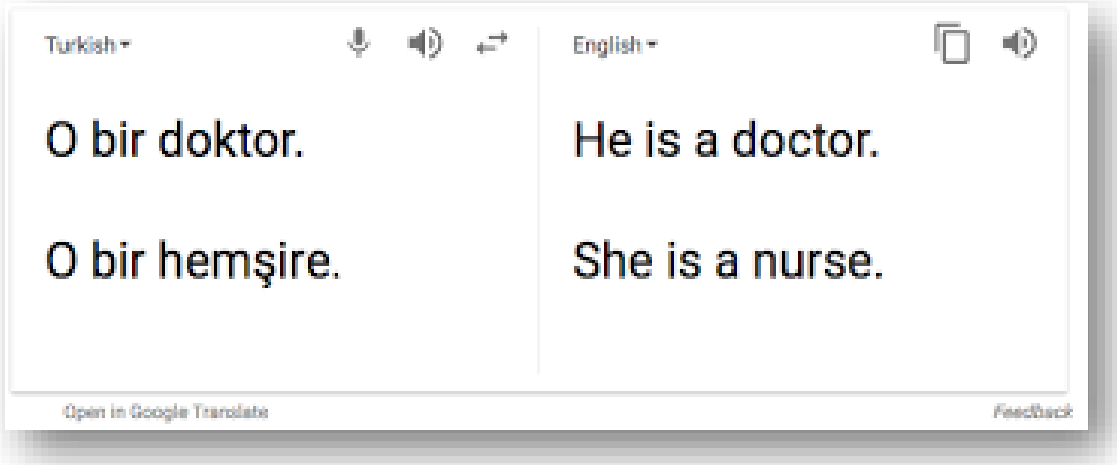
Machine learning controversy

Many machine learning techniques offer **fully automatized** routines for calculating prices, insurance premiums, etc. or clustering data into different segments (for example: brands or regions)

But if the interpretability is missing, many problems might occur

Machine learning controversy

- Certain companies have sparked controversy as ethnic, gender or ‘unethical’ variables slipped into their models, often because data bias was not corrected



Poland: Banks obliged to explain their credit decisions

By Panoptykon Foundation

Owing to the initiative of the Polish EDRi member Panoptykon, bank clients in Poland will have the right to receive an explanation of the assessment of their creditworthiness. The initiative proposed and fought for amendments in the Polish banking law, and resulted in an even higher standard than the one envisioned in the General Data Protection Regulation (GDPR).

Machine learning controversy

- Certain companies have sparked controversy as ethnic, gender or ‘unethical’ variables slipped into their models, often because data bias was not corrected

In 2019, Facebook was found to be in contravention of the U.S. constitution, by allowing its advertisers to deliberately target adverts according to gender, race, and religion, all of which are protected classes under the country’s legal system.

Job adverts for roles in nursing or secretarial work were suggested primarily to women, whereas job ads for janitors and taxi drivers had been shown to a higher number of men, in particular **men from minority backgrounds**.

The algorithm learned that ads for real estate were likely to attain better engagement stats when shown to white people, resulting in them **no longer being shown to other minority groups**.

Allianz

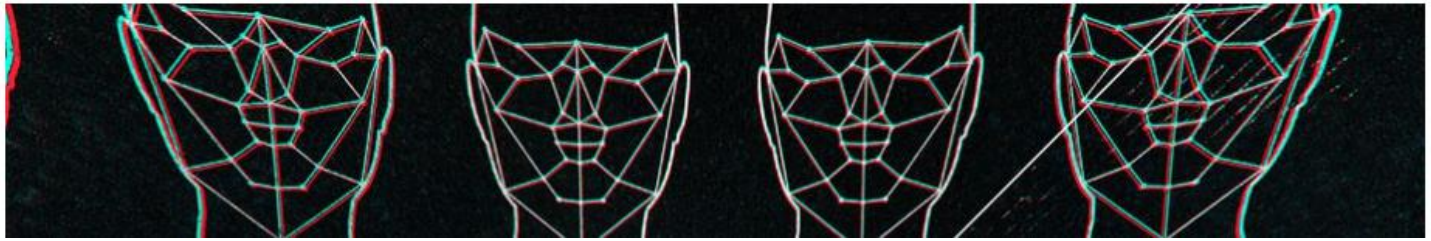
Machine learning controversy

Gender and racial bias found in Amazon's facial recognition technology (again)

Research shows that Amazon's tech has a harder time identifying gender in darker-skinned and female faces

By James Vincent | Jan 25, 2019, 9:45am EST

SHARE

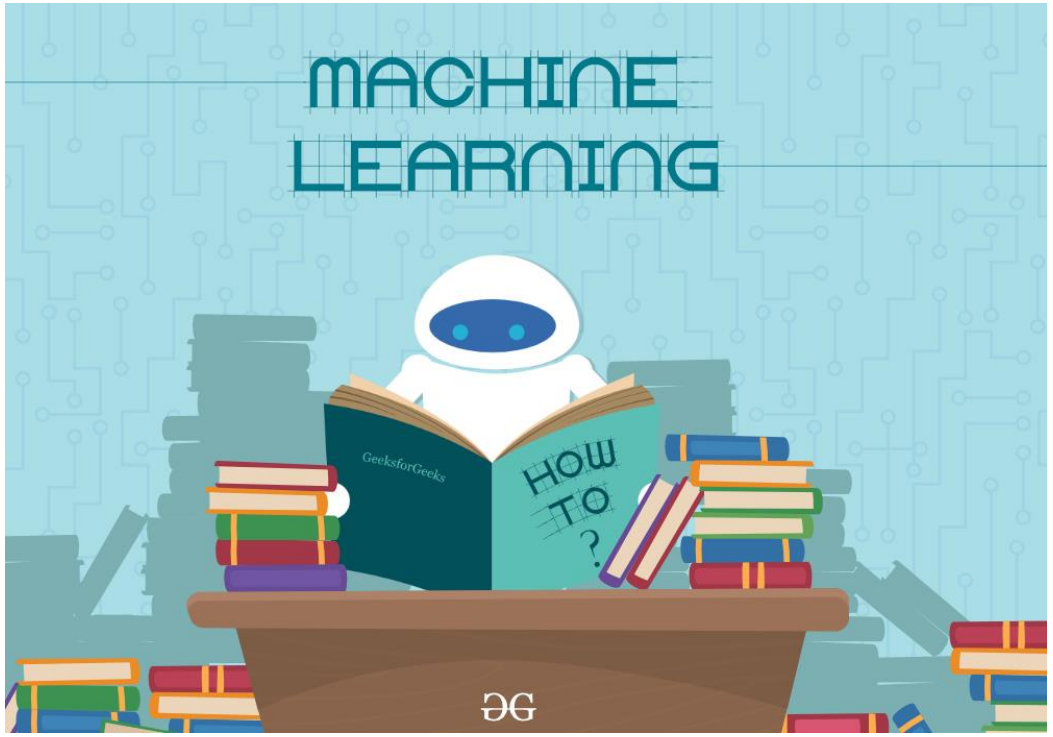


Machine learning controversy

What about the insurance industry?

- **Current standard: GLM models**
- Can Machine learning replace them?

Later on that!



Assessing model accuracy

Allianz

Assessing Model Accuracy

- No model dominates all other models over all possible data sets. We need to decide which model is most suitable **based on the data set given**
- The prediction error $E(Y - \hat{f}(X))^2$ can be estimated by the mean-squared error (**MSE**)

  $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(X_i))^2$

  given a sample $(X_i, Y_i)_{i=1}^{n}$.
- Here X_i denotes a p –vector of regressors for the i-th data point

Allianz

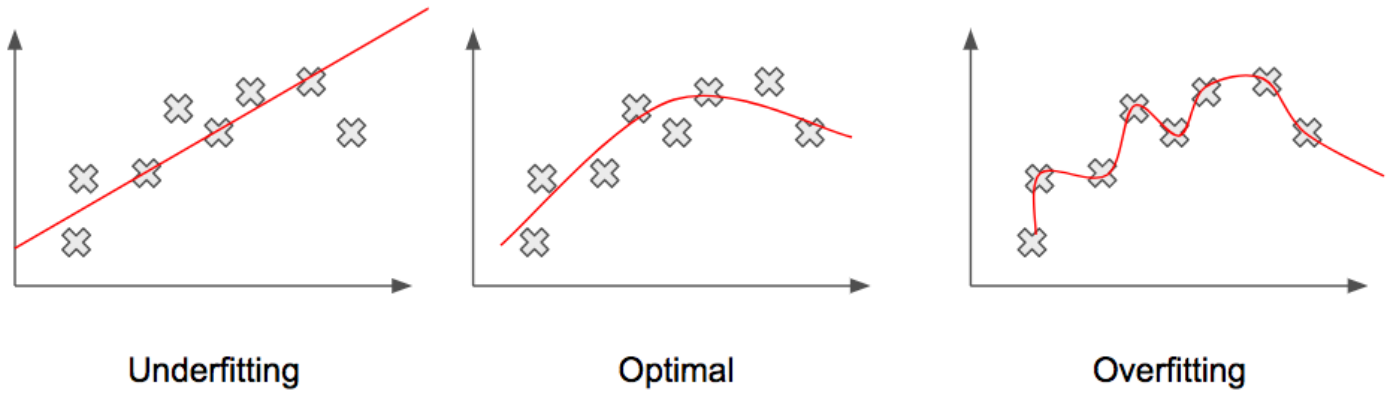
Assessing Model Accuracy

- But we do not want to predict the model accuracy on the data we already observed!
  $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(X_i))^2$ is, actually, an **in-sample (training) MSE.**

- We want our model to perform well on the **future data**,
- For a new (unseen) observation (X_0, Y_0), it should hold that $\hat{f}(X_0) \approx Y_0$.
- In general, when considering all new data points: $\underbrace{Average}_{(X_0,Y_0)}(Y_0 - \hat{f}(X_0))^2$ should be small. This is an **out-of-sample (testing) MSE**

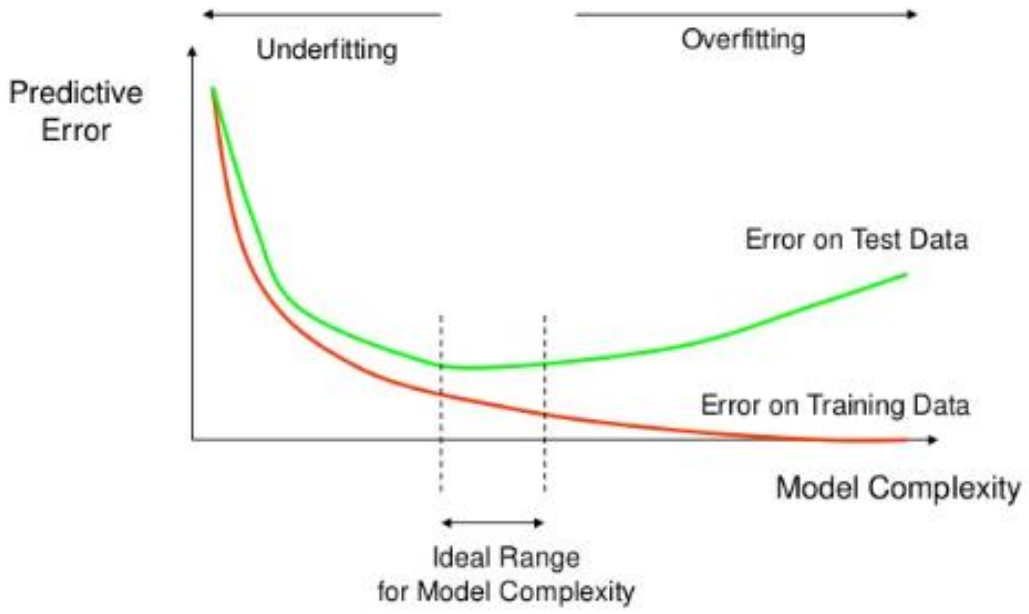
Assessing Model Accuracy

There is no guarantee that a model with a small training MSE will also have a small testing MSE. This leads to concepts of underfitting and overfitting.



Allianz

Assessing Model Accuracy



As the model complexity increases, the training error gets smaller, but the testing error increases.

Underfitting: the model is too simple and performs badly on the training data, and consequently on the testing data

Overfitting: the training data is modelled too well, because non-existing patterns in the data are found (coming from the noise). Therefore, the performance on the future data is poor.

Bias-variance trade-off

- Let X_0 be fixed. Note that the **test MSE** can be written as

$$E(Y_0 - \hat{f}(X_0))^2 = \underbrace{\left(Bias\left(\hat{f}\,(X_0)\right)\right)^2 + Var\left(\hat{f}(X_0)\right)}_{\text{reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{irreducible}}.$$

- Bias: Error introduced by approximating f by $\hat{f}$
- Variance: how much $\hat{f}$ changes if we use different data sets for training

Bias-variance trade-off (additional)

Allianz

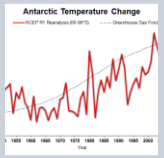
$$
\begin{aligned}
\mathbb{E}[(Y_0 - \hat{f})^2] &= \mathbb{E}[(Y_0 - f + f - \hat{f})^2] \\
&= \mathbb{E}[(Y_0 - f)^2] + \mathbb{E}[(f - \hat{f})^2] + 2\mathbb{E}[(f - \hat{f})(Y_0 - f)] \\
&= E[(f + \epsilon - f)^2] + \mathbb{E}[(f - \hat{f})^2] + 2\mathbb{E}[fY_0 - f^2 - \hat{f}Y_0 + \hat{f}f] \\
&= \mathbb{E}[\epsilon^2] + \mathbb{E}[(f - \hat{f})^2] + 2(f^2 - f^2 - f\mathbb{E}[\hat{f}] + f\mathbb{E}[\hat{f}]) \\
&= \sigma^2 + \mathbb{E}[(f - \hat{f})^2] + 0.
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[(f - \hat{f})^2] &= \mathbb{E}[(f - \mathbb{E}[\hat{f}] + \mathbb{E}[\hat{f}] - \hat{f})^2] \\
&= \mathbb{E}\left[f - \mathbb{E}[\hat{f}]\right]^2 + \mathbb{E}\left[\hat{f} - \mathbb{E}[\hat{f}]\right]^2 \\
&= \left[f - \mathbb{E}[\hat{f}]\right]^2 + \mathbb{E}\left[\hat{f} - \mathbb{E}[\hat{f}]\right]^2 \\
&= (Bias[\hat{f}])^2 + \mathrm{Var}[\hat{f}],
\end{aligned}
$$

Bias-variance trade-off

Allianz

Easy to find a method with low bias and high variance, just use a curve that connects all the points

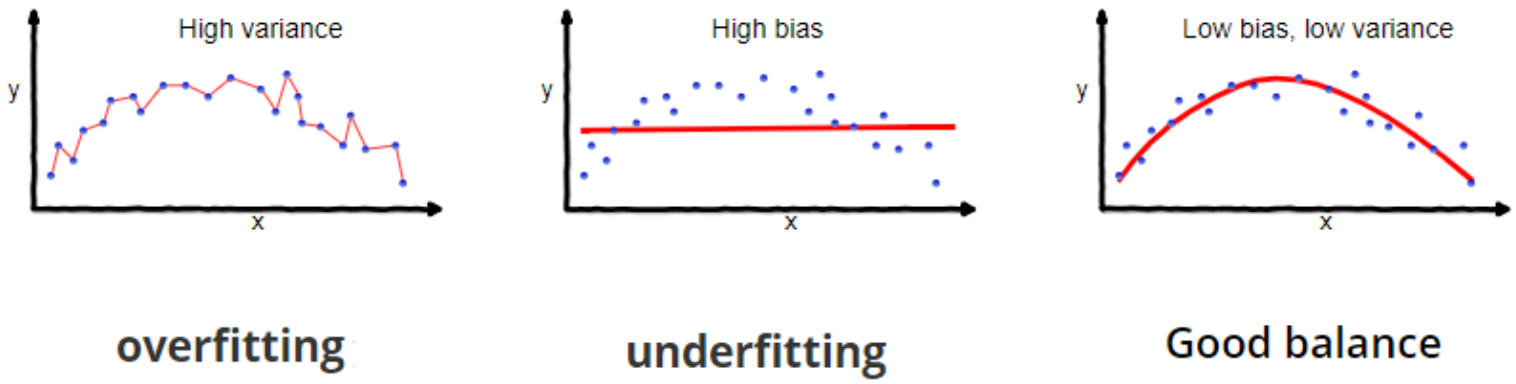


Easy to find a method with low variance and high bias, just take a flat line through the data

But, we want a method that simultaneously has low bias and low variance.

Bias-variance trade-off

Example:



Allianz

Test MSE Estimation

But in real-life situations it is **not possible to compute the test MSE**, because f is unknown, so we need to estimate it.

Remember: the **test MSE** equals:

$$E(Y_0 - \hat{f}(X_0))^2$$

The estimation be done in the following ways:

- Cross-Validation: directly estimating test MSE by using **resampling**
- Indirect way of estimating test error: **adjust the training error** by a penalty term which takes the model dimension into account, i.e. test MSE=train MSE +penalty term

Cross-validation

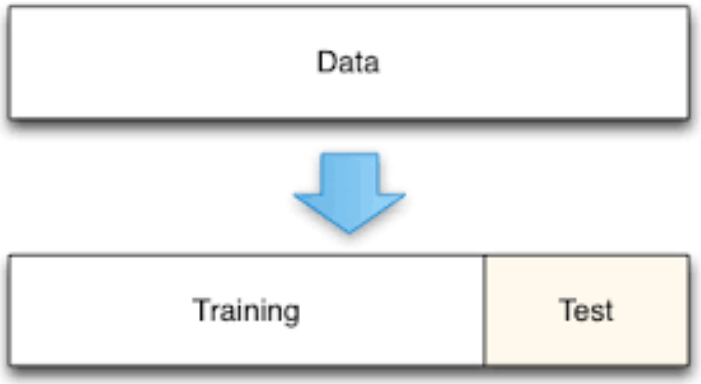
Cross-Validation

- Used to estimate the test MSE, for a given statistical model
- It tells us how our model performs on **unseen data**
- When comparing several competing models, the one with the **smallest** cross-validation error (CV) is preferred.
- It can also be used for selecting tuning parameters for a chosen model (Ridge, Lasso, etc.)

Cross-Validation

There are 3 ways in which CV can be done:

1. **Validation set approach**: divide the data **randomly** into two data sets: training and testing. Usually an 80-20% split is done. The model is then fitted using the training set and the prediction error $\frac{1}{m}\sum_{i=1}^{m}\left(Y_i - \widehat{Y}_i\right)^2$ is calculated on the testing data



Cross-Validation

Example:

- The model trained on 80% of the data gives the following prediction: $\hat{Y} = 2X$.
- The test data is:

Y	X
5	2
9	5
10	4

- CV equals: $\frac{1}{3}[(5-4)^2+(9-10)^2+(10-8)^2] = \frac{6}{3} = 2$

Cross-Validation

Allianz

Drawbacks:

- CV error can be extremely variable, depending on how the data was split
- Only a subset of the data was used for training, this introduces a lot of bias so we might overestimate the testing error

2. **Leave-one-out cross-validation (LOOCV):** Dataset with n sample points is split into $n - 1$ data points, on which model training is done and the testing is done on the remaining one data point. This is then repeated n times, so that each point gets to be in the training and the validation data set. The prediction errors are then averaged out.

Allianz

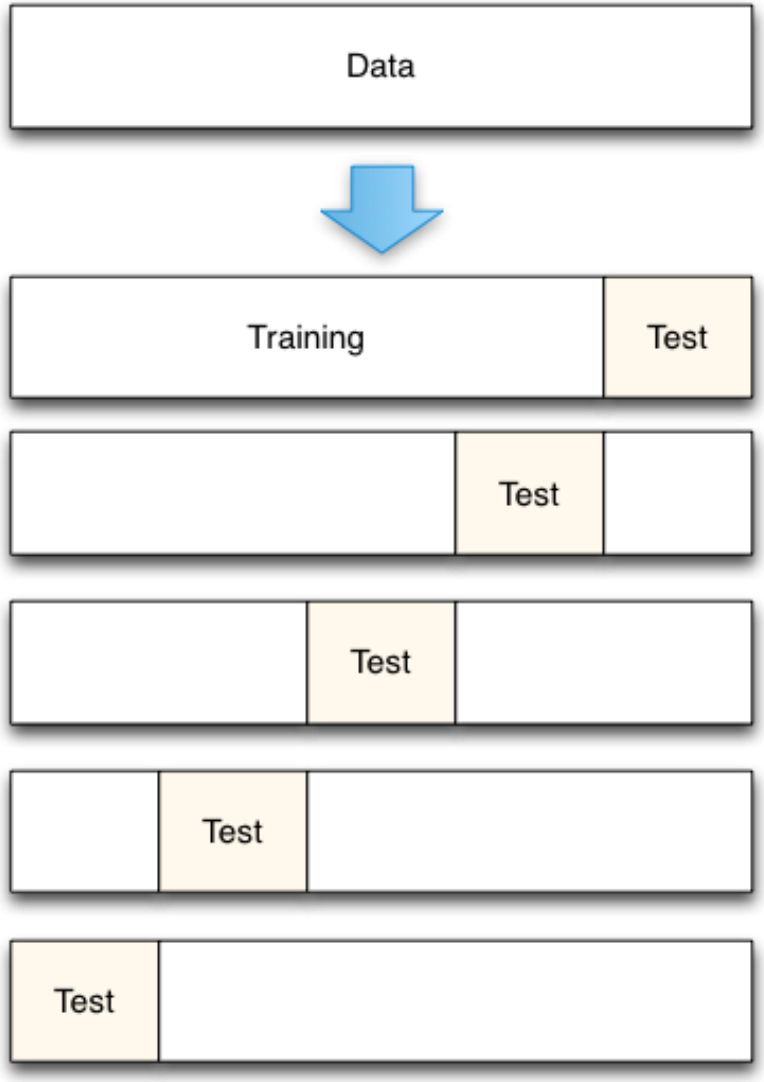
Cross-Validation

- Now there is no randomness in data splits, and there is much less bias compared to the previous method, because $n - 1$ points are used for training
- Problem: we have to fit the model $\boldsymbol{n}$ **times**. Computationally extensive.

Allianz

Cross-Validation

3. K-fold cross-validation: Randomly divide the data set into k parts of (approximately) equal size. Then train the model on $k-1$ parts and test on the remaining part. Repeat k times and average out the testing error.



Allianz

Cross-Validation

- How big should k be?
- Experience shows that $\boldsymbol{k = 5}$ **or** $\boldsymbol{k = 10}$ show best results.
- We fit the model only k times
- The bias remains small, because we fit on almost all data and variability of the CV estimate gets smaller compared to LOOCV, because the outputs for each fit are less correlated
- This method corrects the disadvantages of the previous two.

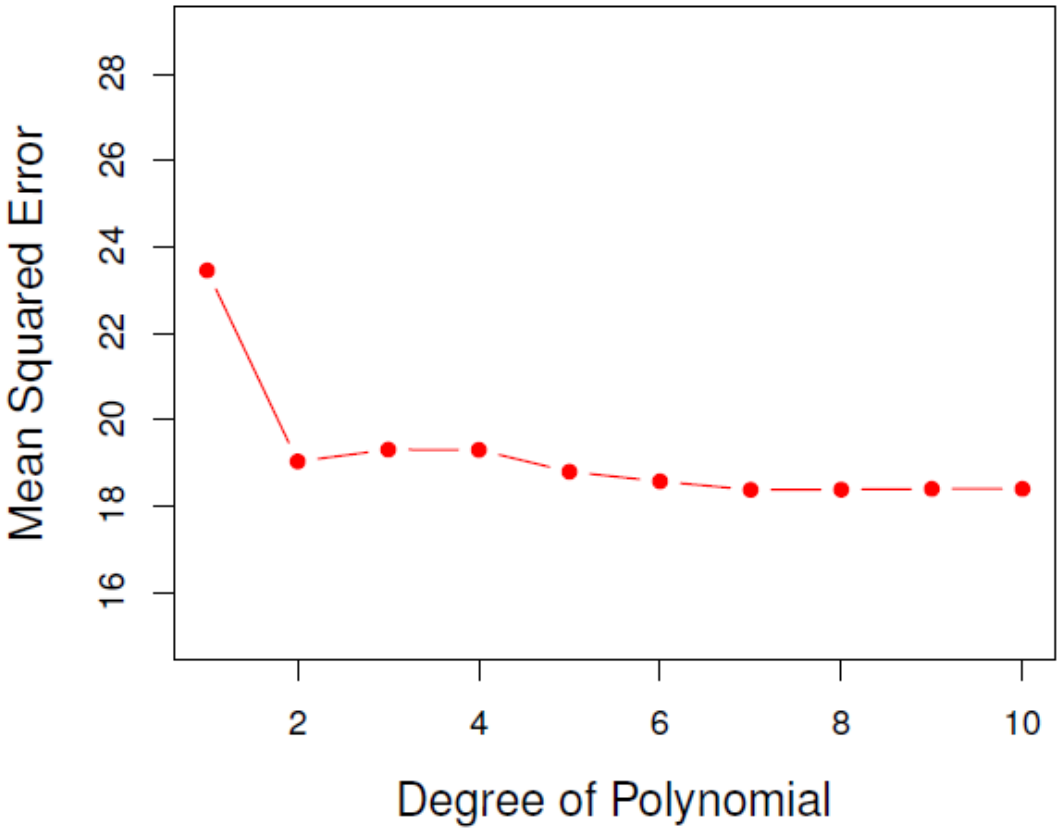
Cross-validation example

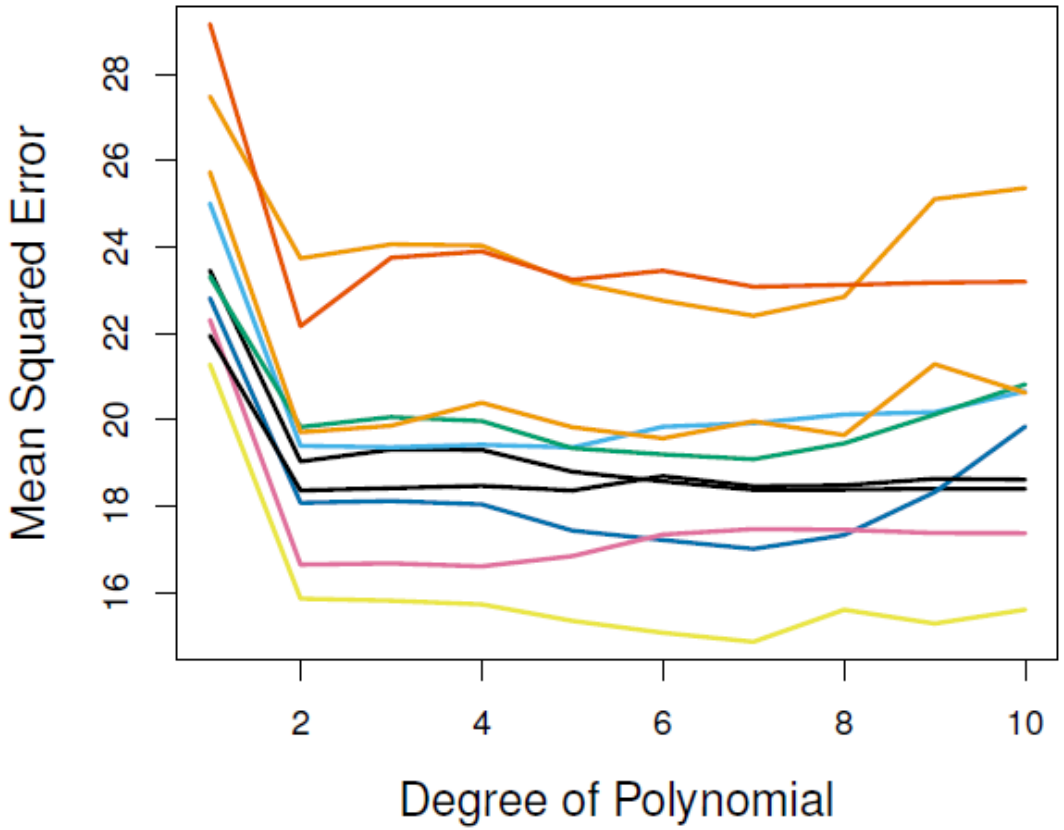
- Response variable **mpg – miles per gallon**
- Polynomial regression is performed with the regressor **horsepower**. But which degree to take?
- Cross-validation can give us an answer

Cross-validation example

Allianz

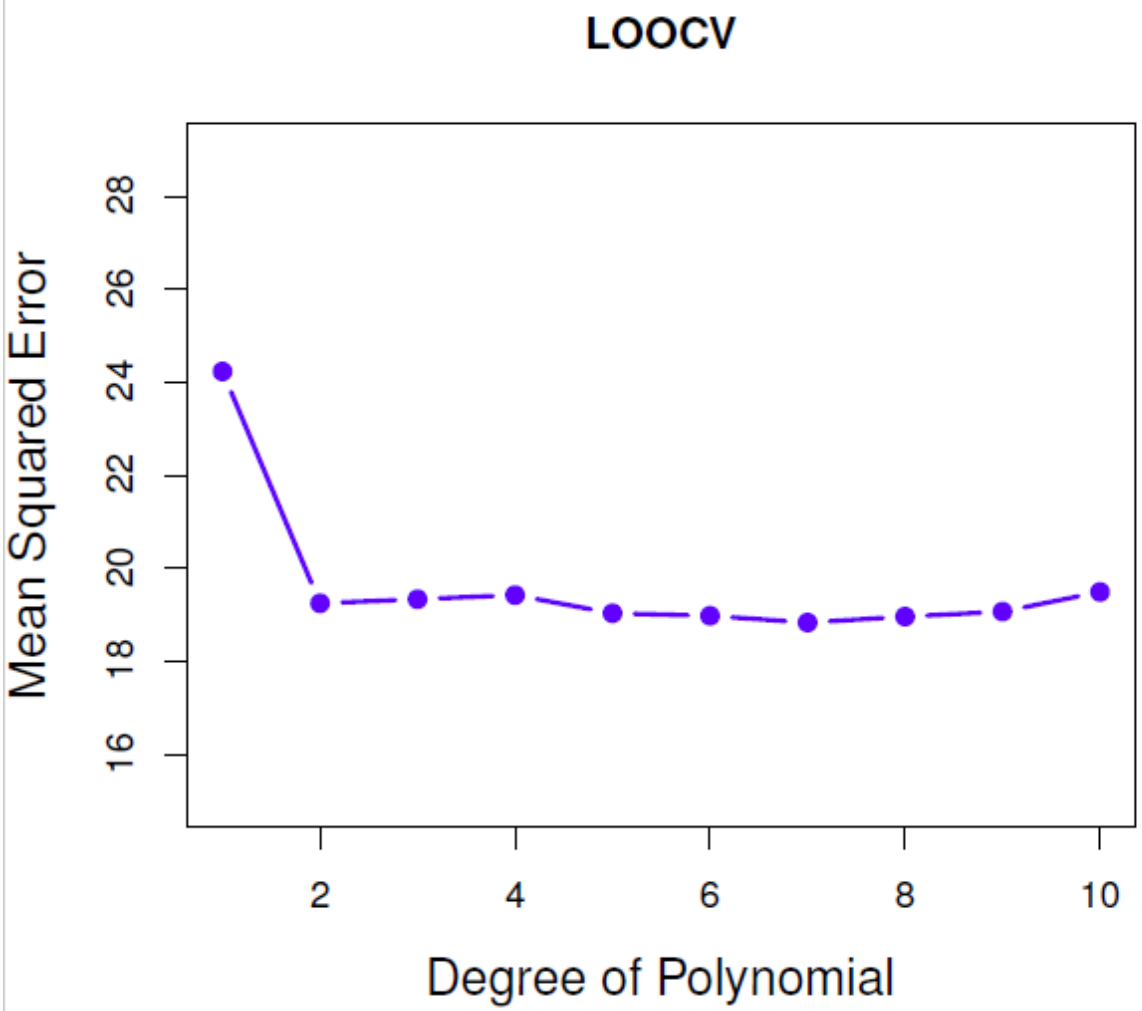
Validation set approach

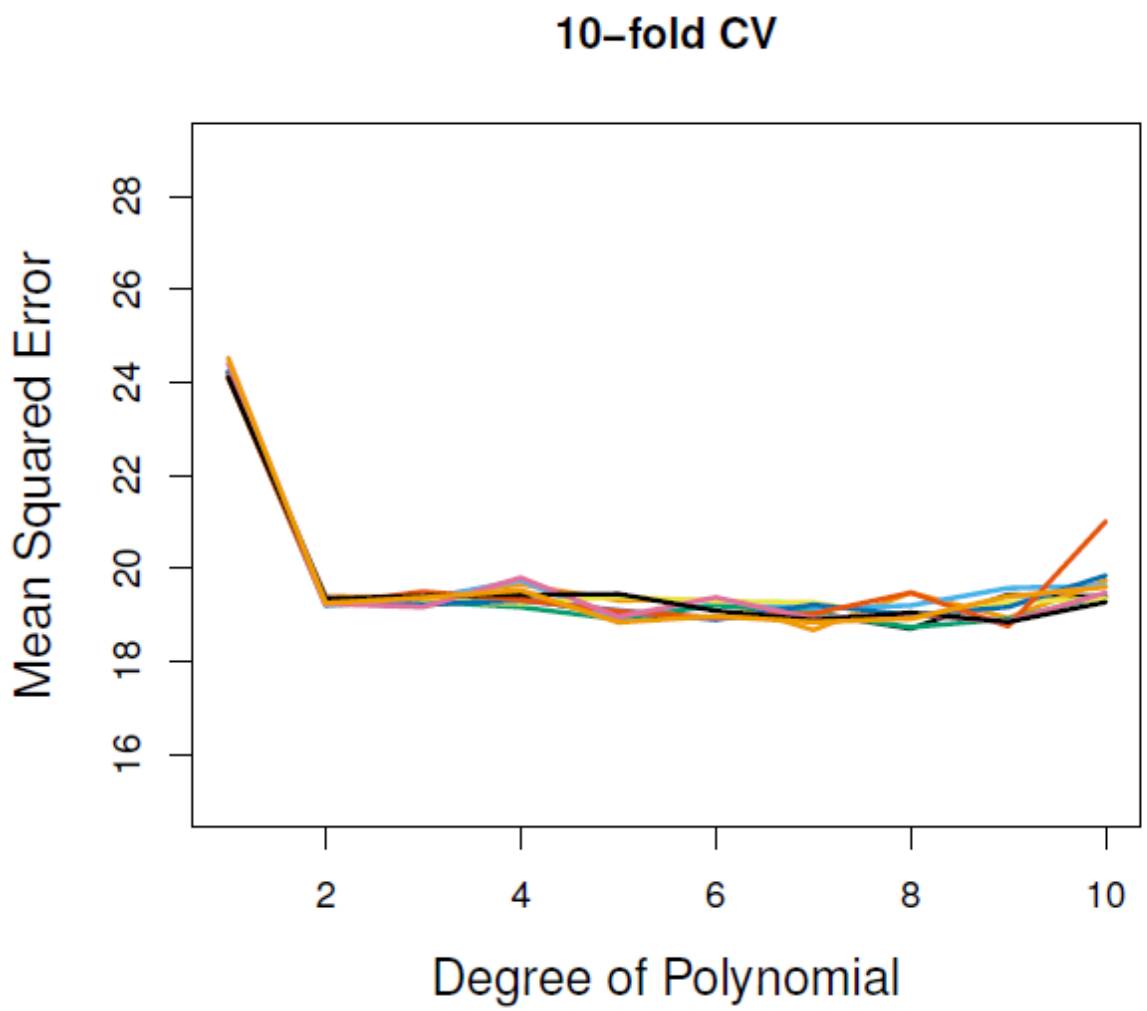




Cross-validation example

Allianz





Adjust the training error: AIC, BIC, etc.

Other way of estimating the test MSE error is by **adjusting the training MSE.**

Allianz

AIC, BIC, etc

- AIC (Akaike Information Criterion) is an estimator for an out-of-sample prediction error and thereby for the relative quality of a statistical model for a given set of data.
- Given a collection of models, AIC estimates the quality of each model. Thus, AIC provides a means for model selection.
- Akaike extends the concept of the maximum likelihood estimation to the case where the number of parameters p is also unknown. A penalty is introduced, depending on p. **So, a parameter is added to the model, only if it leads to a significant improvement in the fit.**

AIC, BIC, etc

- Let $f(y|\theta)$ be a candidate model for estimating Y, for $\theta \in R^p$. For example: $f(y|\theta)$ is the density of $N(X\theta, I)$
- Let $\hat{\theta} = \hat{\theta}(Y)$ be the MLE estimator, given the data $Y \in R^n$.
- Then, $\boldsymbol{AIC = -2logf\left(Y\middle|\widehat{\theta}\right) + 2p}$ is the estimate of the test MSE
- Model with the smallest AIC is chosen

Allianz

AIC, BIC, etc.

BIC (Bayesian Information Criterion) is a similar method to AIC.

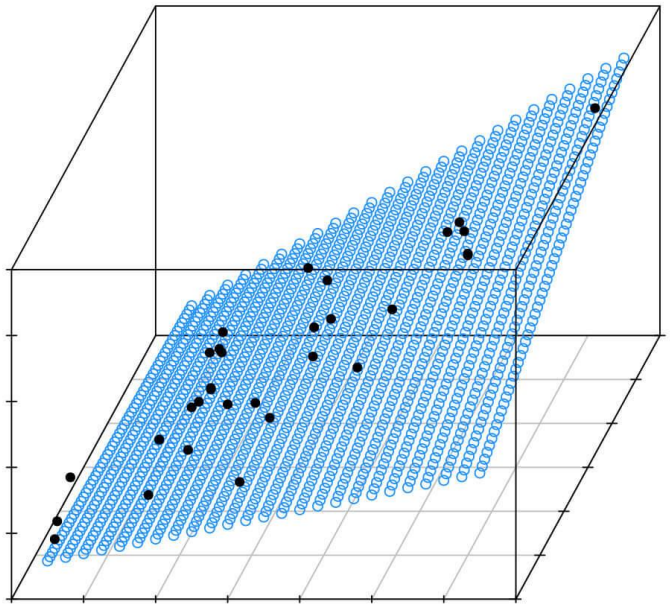
- The model with the smallest $\boldsymbol{BIC = -2logf\left(Y\middle|\widehat{\theta}\right) + \mathbf{p}\,\mathbf{log}(\mathbf{n})}$ is chosen.
- Since the penalty term here is larger, sparser models are selected than with AIC.
- In the linear regression model with normal errors: AIC and BIC have the following forms:

$$\boldsymbol{AIC = n\,\mathbf{log}(\,MSE) + 2p} \text{ and } \boldsymbol{BIC = n\,log(MSE) + plog(n)}$$

Types of Models

Allianz

Linear Models



Allianz

Model selection and regularization

- **Linear models** (and generalized linear models: GLMs), though simple, turn out to be surprisingly competitive in real-world problems, compare to more complex models
- Reason for that lies in their simplicity and interpretability
- GLMs are the standard in the insurance business and most of the results for linear models can be naturally generalized

More tomorrow!

- But what is their prediction accuracy and what happens when the number of parameters p **is large compared to the sample size** n?

Model selection and regularization

Allianz

- Let us focus on linear models, for demonstration
- Assume that: $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, for some $\beta \in R^p$

  $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma I$.

  Also, $\mathrm{Y} \in R^n$ and $\mathrm{X} \in R^{n \times p}$.

- OLS estimator $\hat{\beta} = (X'X)^{-1}X'Y$ is well-defined for $n \geq p$ and it is unbiased. Therefore, the estimates $\hat{Y} = X\hat{\beta}$ are unbiased.
- For $p > n,$ OLS is not even defined. Therefore, we have to come up with some other estimators.

Allianz

Model selection and regularization

But what about the **variance** of these estimates?

- If $n \gg p,$ the variance is usually small, and our estimates are accurate

- But if two or more variables are **highly correlated**, this could lead to high variance and therefore unstable estimates. This happens, because $\det(X'X)$ is almost 0 and the matrix inversion becomes very unstable

Model selection and regularization

Allianz

Example of (potentially) **highly-correlated variables** in Motor Insurance

Vehicle age and contract age

Population density and regional segmentation variables

Example of (potentially) **highly-correlated variables** in SME Insurance

Turnover and number of employees

Allianz

Model selection and regularization

- Also, if n is not much larger than p, the estimates can get very unstable.
- Example: if all regressors are i.i.d. N(0,1) the variance of the predictions equals $\sigma \frac{p}{n-p-1}$.
- This is problematic for p large compared to n.

Allianz

Model selection and regularization

Alternatives to OLS in linear regression:

- Subset selection (best subset and stepwise)
- Dimension reduction (PCA, for example)
- Shrinkage methods (Ridge, Lasso, etc.)

Subset Selection

Subset Selection

Allianz

1. **Best subset selection**: for a linear model with p predictors do

- Let M_0 be the null model with zero regressors, i.e. sample mean of Y is used as a predictor
- For $k = 1,2, \dots, p$
  1. Fit all $\binom{p}{k}$ models that contain exactly k predictors
  2. Pick the best among these $\binom{p}{k}$ models and call it M_k. I.e., choose the model with the largest R^2.
  - Select the best model from $M_0, M_1, \dots, M_p$ using cross-validation, AIC, BIC, etc.
  - Note: here you cannot use R^2 because then the largest model would always be chosen.

https://en.wikipedia.org/wiki/Coefficient_of_determination

Allianz

Subset Selection

- This method is conceptually very simple to understand
- Problem? Too many models to fit!
- How many? 2^p models to fit.
- For example: for $p = 40$, there are 1 073 741 824 models to fit!
- So, we need another solution.

Allianz

Subset Selection

2. Stepwise selection
   - ➢ Forward
   - ➢ Backward

Forwards stepwise selection

- Computationally efficient alternative to the best subset selection
- Here we begin with the null model and add predictors **one at the time** until we get the full model (or some stopping rule is applied)
- Then we choose among these models using cross-validation, AIC, BIC, etc.

Allianz

Subset Selection

More formally:

Forwards stepwise selection: for a linear model with p predictors do

- Let M_0 be the null model with zero regressors, i.e. sample mean of Y is used as a predictor
- For $k = 0,1, \dots, p-1$
  1. Consider all $p-k$ models that add one additional predictor to the model M_k
  2. Pick the best among these $p-k$ models and call it M_{k+1}. I.e. choose the model with the largest R^2.
  - Select the best model from $M_0, M_1, \dots, M_p$ using cross-validation, AIC, BIC, etc.
  - Note: here you cannot use R^2 because then the largest model would always be chosen.

Allianz

Subset Selection

- Here we fit only $1 + \sum_{k=0}^{p-1}(p-k) = 1 + \frac{p\,(p+1)}{2}$ models
- For example: for $p = 40$, there are **466** models to fit. Much better than before.
- This procedure works well in practice, but now there is no guarantee that we will select the best method overall

Backwards stepwise selection:

Similar: here you start with the full model and delete regressors one at the time

Allianz

Example: Prostate cancer

- The data come from a study that examined the correlation between the level of prostate specific antigen (response variable) and a number of clinical measures (regressors) in men who were about to receive a radical prostatectomy.
- It is data frame with 97 rows and 9 columns.

Example: Prostate cancer

Allianz

These data come from a study that examined the correlation between the level of prostate specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. It is data frame with 97 rows and 9 columns.

Usage

```
data(Prostate)
```

Format

The data frame has the following components:

`lcavol`
: log(cancer volume)

`lweight`
: log(prostate weight)

`age`
: age

`lbph`
: log(benign prostatic hyperplasia amount)

`svi`
: seminal vesicle invasion

`lcp`
: log(capsular penetration)

`gleason`
: Gleason score

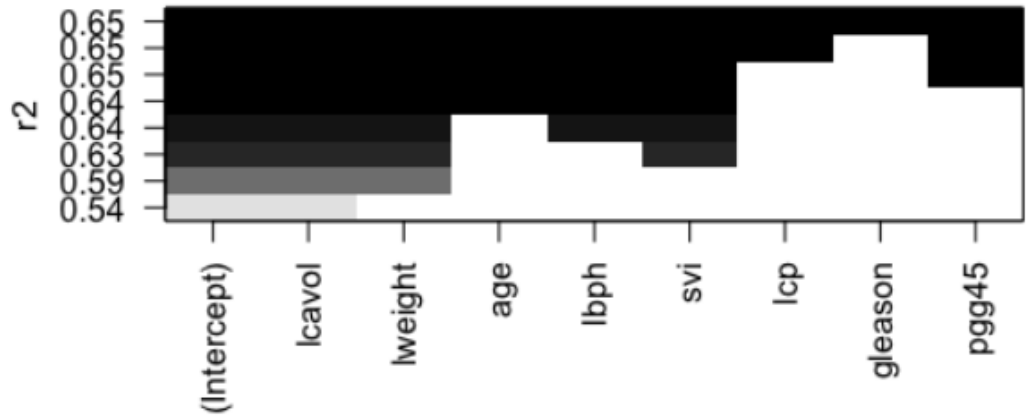
`pgg45`
: percentage Gleason scores 4 or 5

`lpsa`
: log(prostate specific antigen)

Allianz

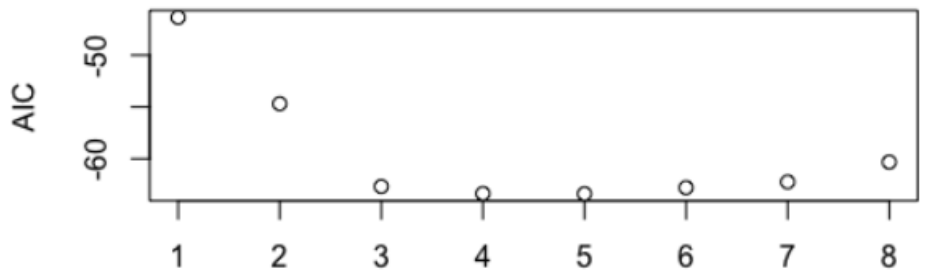
Example: Prostate cancer

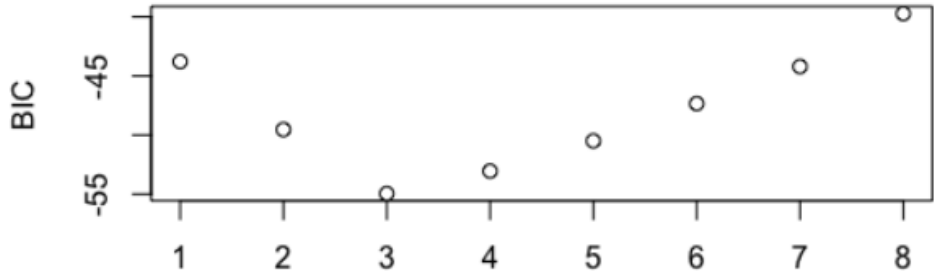
R Package **Leaps** is used to select the best model (based on R^2) of each size



Example: Prostate cancer

- Then AIC and BIC are calculated for each of these models, based on the formula for linear regression with normal errors.






```
> v<-leaps.out$which[which.min(AIC),] #which variables are chosen T/F
> names(X)[v] #gives us the names of those variables
[1] "lcavol"  "lweight" "age"     "lbph"    "svi"
> v<-leaps.out$which[which.min(BIC),] #which variables are chosen T/F
> names(X)[v] #gives us the names of those variables
[1] "lcavol"  "lweight" "svi"
```

Summary for today

Allianz

Summary

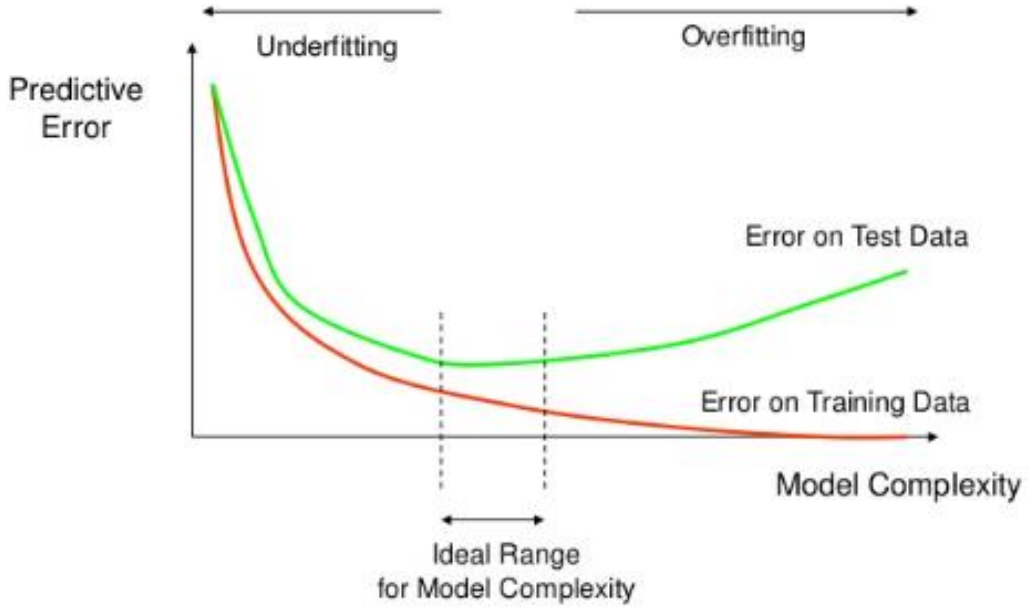
- We assess the model quality by its **prediction error**

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(X_i))^2$$

  given a sample $(X_i, Y_i)_{i=1}^{n}$.
- But this is only one part of it – **training (in-sample) error**
- It is necessary to estimate this error for new (unseen) data – **testing (out-of-sample) error**

Allianz

Summary



A model (and its complexity) should be chosen based on these two prediction errors:

Allianz

Summary

- The training error we can estimate from the sample directly
- There are two types of methods for estimating the testing error
  1. Cross-validation: based on **resampling**
  2. AIC, BIC, etc.: based on **testing error ≈ training error + dimension penalty**

Allianz

Summary

Linear models: simple but widely-used because of its simplicity and interpretability

OLS well-defined for $n \geq p$

But they perform badly if

- p is large compared to n
- some of the regressors are highly correlated

Allianz

Summary

Some methods to reduce the number of parameters:

1. Best subset selection: **all submodels** are considered, but this is computationally infeasible
2. Stepwise-regression: regressors are added **one at the time.** Once a regressor is chosen, it stays

Preview

We are still to see:

- Some other methods that do model selection for linear models
- How to deal with correlations
- How to deal with $p > n$ case?

Thank you!