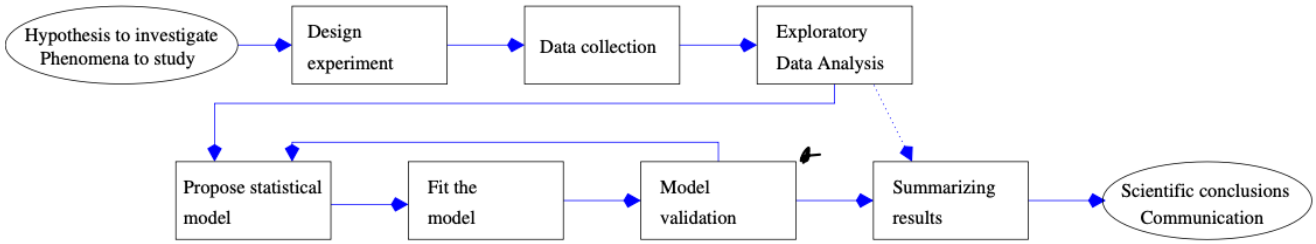
STA121 Summary

Lecture 1 - Resampling and model validation

Learning goal: Understand the conceptual idea of statistical modeling and workflow

The usefulness of a model determines if we should proceed after model validation or move back to propose an alternative statistical model.



There are three types of model qualities:

1) Underfitting: the model is too simplistic and doesn't fit all the necessary signal. Such models usually have a low-variance but high-bias tradeoff. One of these models is for example a constant line.
2) Good Fitting/Robust: the model is not underfitting nor overfitting. It captures the needed signal and does not capture the noise. Such models have a low-variance and low-bias tradeoff. It is the best model to aim for when selecting a model. Good for fitting the existing data but also prediction data.
3) Overfitting: the model has residuals with low variance, but it is more complex than it should be, it is overcomplicated, it fits too precisely (fits noise too). Such models have a high-variance and very low-bias tradeoff.
   If we remove on point from the set the model would change a lot (See slide 21). Such models don't generalize to new and unobserved cases. (e.g. interpolation that connects lines through all the points).

An optimal model is somewhere in between a simple and a complex model.

Bias: describes how well a model matches the training set. A model with high bias (underfitting models) won't match the data closely, a model with low bias (robust models or overfitting ones) will match the data set very closely.

Variance: describes how much a model changes when you train it using different portions of your data set. A model with high variance (overfitting ones) will have the flexibility to match any data set that is provided to it, potentially resulting in dramatically different models each time.

Training data: used to fit the model. Training error: measure of how well a model fits the training data.

Testing data: used to calculate the performance of the fitted model on new data. Testing error: measure of how well a model fits and generalizes to new and unseen data.

If we look at one training set and one testing set only it is very computationally efficient, but it is not good for measuring the quality of the model on new unobserved data.

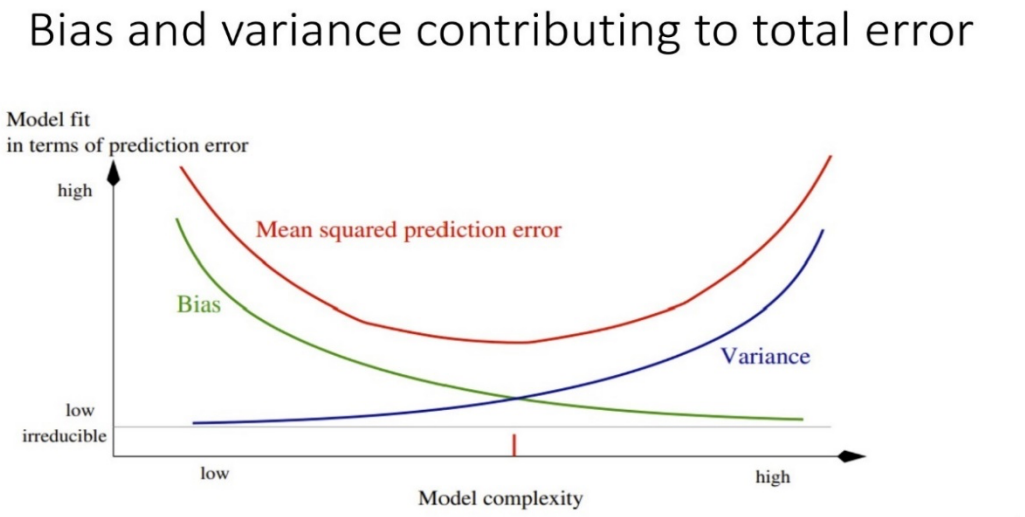
Using a criteria like the mean squared prediction error, we can find “optimal models”.

It’s nearly improbable to have a model with high training error and low testing error.

Aim for model with low training error and low testing error -> good fit.

A model with low training error and high testing error resembles high variance and indicates that model is overfitted to training data and is not able to generalize on new unseen data.

A model with high training error and high testing error resembles high bias and indicates that the model is underfitted to training data and is not able to capture the structure of the data.



The variance (blue line) increases since we do not capture the model but the noise instead (the model is too complex to be optimal). The gray line indicates the irreducible error (epsilon).

Mean Squared Prediction Error: measures how close a regression line is to a set of data points. If we minimize it too much (means the residuals are minimized: the distance between the observed points and the fitted points of the estimated model are smallest as possible), we might come up with an overfitting model (which is not what we want).

Best model is the one with the lowest mean squared error in the test set.

The mean squared error is calculated by adding the squared bias of the model we want to estimate to the variance of the model we want to estimate + the variance.

The goal should not be to aim for the model with the lowest prediction error because that model is probably overfitting and over complicated (costs a lot computationally wise and fits noise of the set too). Aim for a balanced model, as mentioned before. It is a decomposition in positive terms consisting of squared bias, variance and variance of the noise. The noise is irreducible and cannot be reduced by a different model. The choice of the model can though lead to smaller or larger squared bias and variance term but it is not possible to simultaneously reduce both arbitrarily.

Here is the formula:

$$\begin{aligned}\mathrm{E}\big((Y-\widehat{g}(\boldsymbol{x}))^2\big) &= \big(\mathrm{E}(\widehat{g}(\boldsymbol{x}))-g(\boldsymbol{x})\big)^2+\mathrm{E}\big((\widehat{g}(\boldsymbol{x})-\mathrm{E}(\widehat{g}(\boldsymbol{x}))^2\big)+\mathrm{E}\big((Y-g(\boldsymbol{x}))^2\big)\\ &= \mathrm{bias}\big(\widehat{g}(\boldsymbol{x})\big)^2+\mathrm{Var}\big(\widehat{g}(\boldsymbol{x})\big)+\sigma^2,\end{aligned}$$

As the variance estimate we use s^2 = 1/n-1 etc. because the one with 1/n is biased. Remember:

Variance can only be non-negative

Learning goal: Understand how resampling can help us in understanding variability of estimators

Sampling: process of selecting a subset of individuals from a population in order to make inference about the whole population.

Random sampling: sampling technique where each individual has the same probability of being chosen.

Statistical sampling: process of generating "new" data based on observations or other existing data and possibly on some additional statistical assumptions.

Bootstrapping: resampling technique that involves taking repeated samples from a sample taken from a population repeatedly with replacement.

By repeatedly sampling from a data, we create many artificial samples from which estimates can be computed (e.g. mean, variance). By continuously resampling we can construct a histogram of the obtained estimates and we obtain the empirical density distribution of the estimated parameter. This distribution can be used to derive empirical confidence intervals for the estimator of interest.

To calculate the sum of the squared residuals of a linear model in R: sum(fit$residuals^2). Substitute fit with the name of the variable associated with the linear model.

Learning goal: Understand how prediction error helps us understand how well a prediction model fits the data

Prediction error: measure of how well a model predicts a variable. It is calculated as difference between actual value and the value of the predicted observation fitted by the model.

By comparing the actual predicted values with the ones predicted by the model, we can understand how close the prediction of the model is to the actual data and how good the estimation on unobserved data is. If the model is able to accurately predict the variable, the prediction error will be small. On the other hand, if the model is not able to accurately predict the variable, the prediction error will be large. If the prediction error is small, it suggests that the model is able to accurately capture the underlying relationships in the data and is a good fit for the data. If the prediction error is large, it suggests that the model is not able to accurately capture the underlying relationships in the data and is a poor fit for the data.

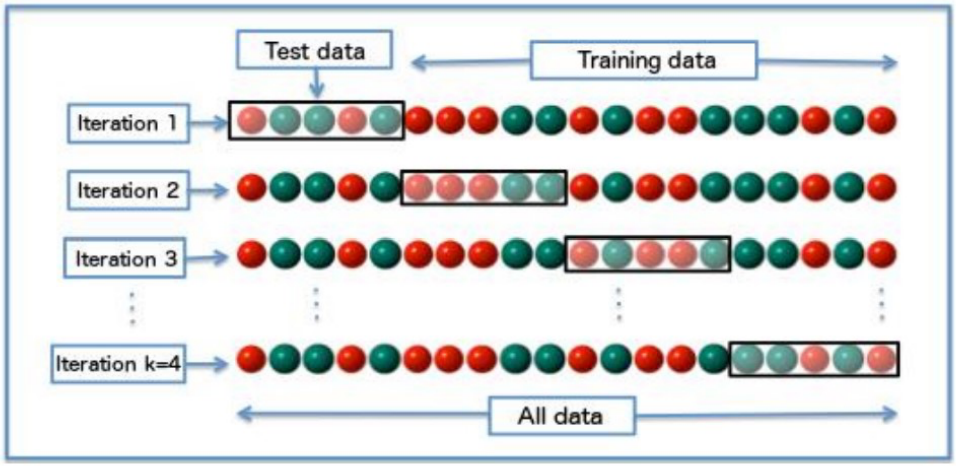
Learning goal: Explain and apply cross-validation, k-fold cross validation

Data is divided into three sets: a training, a validation set and a test set. The training set and validation set are part of the cross validation set, used for fitting the model and usually corresponds to 80% of the total data, whereas the test set to the remaining 20%.

The test set is set aside and not used in the fitting and validation procedure. After the best model (with lowest mean squared prediction error) has been selected, we take the test set to report the quality of this model.

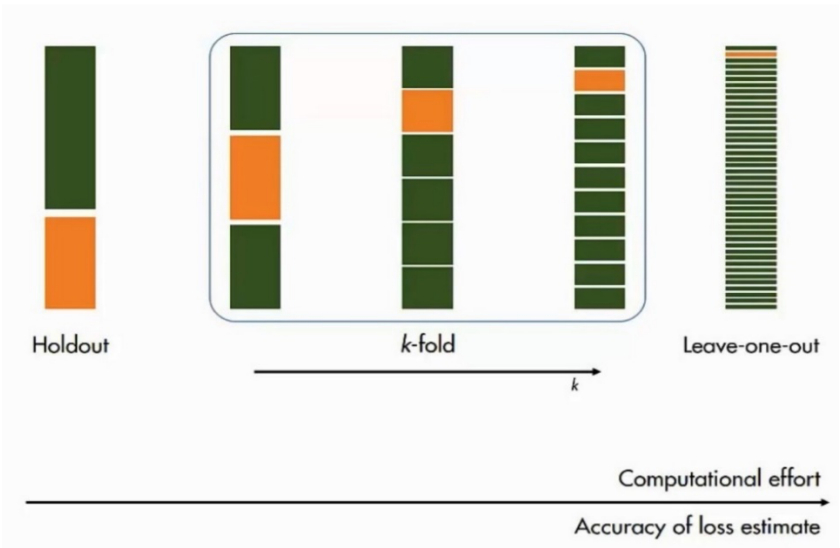
If we look at one model only it is computationally efficient, but we don't know how exact we are in our measure of the quality of the model -> less accurate. Cross validation can be used to compare models and choose the best one for given data.

The validation and training sets are always different at each iteration. For each iteration we compute a new model always in the same way and then we select the best model out of the ones (with lowest testing error) we calculated with cross validation and compare it with the test set. An observation is in the validation data at least one time and k-1 in the training data, but it cannot be simultaneously in the training set and in the validation set.



Cross validation is often used in nonparametric smoothing, where the fitting paradigm is presented under the bias-variance tradeoff. The tradeoff refers to the problem of simultaneously minimizing different sources of error (bias and variance).

Methods of Cross validation



Holdout cross validation: we just have one training set and one validation set, it is less accurate but the computational effort is minimized. We fit the model once.

K-fold cross validation: we fit the model k times. There are multiple subsets, one is the validation set and the rest n-1 is in the training set, we perform this operation for k different validation sets and at the end we compute k different models. 1/k is the percentage of data we use as the validation set (e.g. if we have a 3-fold cross validation each time we use 33.3% of the data as vaidation set and the

remaining 66.6% as the training set). This method of cross validation is usually used in complex models (neural networks, etc.). After selecting the best model out of all the computed models, which is the one with the smallest mean squared prediction error, we take the test set and calculate the testing error of the model against it. If we have n data points, n-fold-cross validation is the same as leave-one-out cross validation. In k-fold cross validation, each observation is used k-1 times in a training set, and exactly one time in a validation set. Be careful not to test the model by re-using the data that has been used for fitting the model.

Cross validation can be used for both model selection (selecting best model based on existing data) and model validation (Evaluating performance of model on unseen data).

Be careful not to test the model by re-using the data that has been used for fitting the model.

Learning goal: Understand the leave-one-out statistics and leave-one-out cross-validation and its applications

Leave-one-out cross validation: It is the most expensive computationally but also the most accurate out of the tree methods (gives the most reliable result). We compute models n times and n-1 sets are in the training set and one is the validation set. One observation at a time is omitted, and the remaining observations are used to fit the model and compare the prediction against the omitted observation. LOOCV is form of k-fold where k is the size of the dataset.

Applications of LOOCV: estimate the predictive performance of a model on unseen data, compare any two models, or parameter tuning for complex models (process of selecting best parameters for a complex model).

Leave-one-out statistics: Summary statistics based on quantities from omitting observations. Leave-one-out cross validation refers to entire process of cross-validation, while leave-one-out statistics only to the statistics that can be computed with the process of cross-validation.

Learning goal: Explain and apply Bootstrap

Statistical Sampling is used when it is too expensive or impossible to measure the value of a parameter of a population. In Statistical Sampling we sample a population, measure a statistic of its sample, and then use this statistic to say something about the corresponding parameter of the population. The computational cost of sampling is cheap and thus we can draw a large number of samples.

Bootstrapping is a method that involves sampling repeatedly with replacement (one observation can be choosen multiple times in the same sample) from a sample of a population to estimate its parameters. It is a resampling technique to approximate distributions based on an empirical sample. It is a tool of model validation.

Some people might never be sampled (first sample doesn't contain them or never picked in second sample) others will be sampled multiple times. We sample from a sample. The randomly chosen sample looks quite the population it came from so we can make assumptions about the distribution of that population (e.g. normal, binomial).

In bootstrapping, the size of the bootstrapped sample is usually the same as the size of the original sample we picked from a population. The sample size is fix.

<u>Bootstrap summary</u>

The principle: resample with replacement from a sample, use that to approximate the sampling distribution.

Then derive the estimates of standard errors and confidence intervals for complex estimators of complex parameters of the distribution.

Then check the stability of the results (if we have a too wide range or strange results it might mean that the sample size is too small)

Bootstrapping is (under some conditions) asymptotically consistent, meaning that it does not provide general finite-sample (-> strange results with small sample size).

Make sure that we sample from the wide format and not the long format (because we want to sample in groups according to object of interest).

Types of Bootstrapping:

1) Non-parametric bootstrapping: does not make any assumptions about the underlying distribution of the data. This method can be used to estimate the variability of any statistic.
2) Parametric bootstrapping: assumes that the data comes from a known distribution with unknown parameters. Should yield more accurate confidence intervals if the parametric model is a good description of the data.
   For small sample size n, the nonparametric bootstrap (sampling bootstrap) may perform poorly, it is not sensitive to model misspecification. In parametric bootstrapping, instead of sampling directly from the original data set, new samples are generated from a probability distribution that is estimated from the original data. This distribution is typically chosen to be a parametric distribution, such as a normal distribution, that best fits the original data.

Questions on Lecture 1:

a) We have n data points. In this case, n-fold cross validation is the same as leave-one-out cross-validation -> True.
b) We are considering 4 models: M1, M2, M3, M4. Mean squared errors on the training sets for these models are: T1 = 120, T2 = 135, T3 = 40, T4 = 145. Mean squared error on the validation sets are V1 = 300, V2 = 150, V3 = 400, V4 = 200. Which model would you choose? M2 because it has the smallest validation error.
c) In k-fold cross validation, each observation is used k-1 times in a training set, and exactly one time in a validation set (No test set here, only the training set and the validation set) -> True.
d) In a simple subsampling CV, there is no guarantee that each observation will only be used once, but one observation cannot be included in both train and validation (test) sets at the same time -> True.

e) When doing a k-fold cross validation, each data point will be in the test set exactly once, no matter what k is -> TRUE. (Considering the test set as the validation set in classic ML literature), then yes, since each fold is used k -1 times for fitting the model, and only once for computing the prediction metric (for example (RMSPE).

Learning goal: Explain motivation for PCA

One of the main motivations for PCA is to reduce the complexity of the data while retaining as much of the original information as possible. In many applications, the data has many features and

dimensions, and it is often difficult to make sense of the data or to find relationships in it. By reducing the dimensionality of the data, PCA makes it possible to visualize and understand the data in a simpler and more interpretable form, while still capturing the essence of the data in a few principal components, which explain most of the variation in the data. Dimensionality reduction has multiple benefits: first of all we can better visualize the data, and with fewer dimensions we can make a better generalization about the data. PCA also helps minimizing the residuals (squared distance) and maximizing the variance (squared distance).

Learning goal: Explain PCA (including eigendecomposition*)
Principal Component Analysis (PCA) is a technique for dimensionality reduction. PCA expresses the data on a new coordinate basis. The new basis vectors, called Principal Components, are linear combinations of the original variables. They are chosen in such a way, that the first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. It is achieved by choosing as the PCs the eigenvectors of the empirical covariance matrix of the data. PC1 corresponds to the eigenvector with the largest eigenvalue, PC2 to the eigenvector with the second largest eigenvalue and so on. Because PCs are ordered by the amount of the variability explained, one can represent the data using only a subset of PCs without losing a lot of information, which allows to reduce the number of dimensions, either for further analysis, visualization, or data compression.
Dimensions are called PC1, PC2 and PC3 instead of x, y and z. There might exist “better” dimensions than the original ones to represent the data. It might be possible to use only a subset of dimensions (and skip/discard the ones that explain the least amount of variance/information about the data).

The method we will see in this lecture is called Feature extraction.

Feature extraction: we create” new” independent variables where each new independent variable is a combination of the old independent variables. We drop the least important/significant variables/dimensions. Because these new independent variables are based on the old ones, we are still keeping the most valuable parts of our old variables, even when we drop one or more of these new variables.
The principal components are orthogonal to each other. Because they are orthogonal to one another, they are statistically independent of one another. To obtain principal components perpendicular to each other data has to be centered (center = true).

The PCA transformation (rotation) ensures that the horizontal axis PC1 has the most variation. The vertical axis PC2 has the second-most and so on. Rotation consists of switching from a vector basis to another one.

While applying the PCA algorithm, if we get all eigenvectors the same, then the algorithm won't be able to select the Principal Components because in such cases, which will make PCA perform badly.

We should not use categorical variables for a PCA analysis, only quantitative variables.

It is better to have a highly correlated set of variables to perform a PCA rather than a low correlated set of variables (Cov/Correlation matrix).

When we perform PCA manually, the pca with function in R is calculated on the realisations of X1 and X2 while the Eigenvalues of Σ are from the theoretical distribution of X1 and X2. As we increase n, the two converge.

In PCA, variables are often scaled (i.e., standardized). Data standardization is recommended when variables are measured in different scales (e.g. kilometers vs centimeters). If we have different scales the PCA would be wrong. One of the variables may dominate all others due to its huge variance and thus PCA may not achieve its goals.

The goal is to make the variables comparable. Variables are scaled to have standard deviation and variance of 1 and mean of zero. With rescaling, PCA translates to an eigenvalue decomposition of the empirical correlation matrix. We want to ensure that all variables have the same weights. Otherwise, the PCA result would be swamped by the axes with large digits in variance.
In the original dataset after scaling the variance stays the same as the one described by pca.
When we have a gridded space-time data we often talk about empirical orthogonal function (EOF) decomposition instead of PCA.
The main difference between PCA and EOF is notational in nature.
PCA requirements/assumptions:
-In the data there are multiple variables that should be measured at the continuous level (not just 0,1 alternated).
-In the data there is a linear relationship between all variables (high correlation). The reason for this is that PCA is based on Pearson correlation coefficients.
-In the data there is sample adequacy, which means that for PCA to produce a reliable result, large enough sample sizes are required.
-The data should be suitable for data reduction. You need to have adequate correlations between the variables in order for variables to be reduced to a smaller number of components.
-There should not be outliers in the data. With outliers PCA results would not be reliable.

In a normal PCA we have more samples than dimensions. In a data where we have more dimensions than samples/observations we would instead perform a decomposition of the empirical variance-covariance matrix based on a singular value decomposition (SVD), instead of an eigen-decomposition.

Learning goal: Assess how many principal components are needed

To select the number of components we should keep we use North's rule of thumb.

The maximum number of principal components in PCA has to be smaller equal than the number of features/variables.

How many dimensions/features should we keep? Method 1: proportion of variance explained (e.g. 75%), method 2: scree plot: identifying the point where adding a new feature as a significant drop in variance explained relative to the previous feature.

Method 3: North's rule of thumb: truncate the expansion when the sampling error becomes larger than the distance between two eigenvalues. The uncertainty of the eigenvalues due to measurement error can be approximated by sqrt(2/n_of_variables). This method is more precise than the first 2 ones. If we have a very small n (number of observations) and p (number of variables) north's rule doesn't work very well.

Learning goal: Interpret principal component scores and loadings

If the scores are large, it means that the data point is located far away from the origin in the new space, and therefore has high values for the corresponding principal component. If the loading is large, it means that the variable has a strong correlation with the corresponding principal component. Positive loadings indicate that high values of the variable correspond to high scores of the principal component, while negative loadings indicate that high values of the variable correspond to low scores of the principal component.

Questions on Lecture 2:

a) PCA can be used for projecting and visualizing data in lower dimensions -> True
b) All principal components are orthogonal to each other -> True
c) Maximum number of principal components <= number of features -> True
d) What will happen when eigenvalues are roughly equal? PCA will perform badly
e) PCA performs best when the eigenvalues of the covariance matrix are roughly equal -> FALSE. PCA performs best when we have only a few eigenvalues with high values. This means that all the variability of our data will be mostly explained with only a few PCA dimensions.
f) Principal components are the main modes of variability of a spatio-temporal dataset and depend on time -> FALSE, it would be EOF, principal components do not depend on time, they can be calculated regardless of time and result is same. EOFs can account for temporal dependencies in the data, since they are calculated based on the covariance or correlation between different time steps.

Learning goal: Describe different clustering methods

The idea of clustering is to group n observations into k homogeneous classes where k is possibly also unknown (but usually assumed to be much smaller than n). Each observation is assigned to one of the k different clusters. The clustering assignment is based on some measure of similarity, with observations within a group displaying high similarity (or equivalently, low dissimilarity). We so maximize the similarity within each group.
Clustering reduces the complexity of the data and facilitates its interpretation. Clustering helps finding patterns, features or even details in the data and groups data with similar patterns together.
In order to decide which groups to merge, dissimilarity is used: three properties:

1) Symmetry: d(a,b) = d(b,a)
2) Positivity: d(a,b) >= 0
3) d(a,b) increases as the dissimilarity between a and b increases.

Dissimilar points are placed in different clusters, whereas similar points in the same cluster (or one very similar). The Euclidean distance can be used as the dissimilarity between single observations. The observations (or features) should be standardized before clustering to avoid having too much weight in some dimensions.

Hierarchical Clustering: In hierarchical clustering algorithms we build the clusters from bottom up. We start with each observation from its own cluster and then iteratively combine the two clusters that are most similar into a single cluster.

1)

Three different methods of hierarchical clustering (to define dissimilarity):

1) Complete linkage (furthest-neighbor linkage): the distance between clusters is the maximum distance between their members. In complete linkage for calculating the distance we compute the maximal difference between the two clusters, and then we group clusters together with the smallest of the calculated distances.
   It generally yields more compact clusters than the Ward linkage method.
2) Single linkage (nearest-neighbor linkage): we group together clusters with the smallest dissimilarity. This method may yield “stretched out” clusters that are less descriptive. The distance between two clusters is the minimum distance between their members.
3) Ward Linkage/Method: we group together pair of clusters that lead to minimum increase in total within-cluster variance after merging them together. Can be seen as a middle distance/average linkage clustering method. The distance between two clusters is how much the sum of squares will increase when we merge two clusters together.

At each step of clustering, we want to combine together two clusters, that are the closest to each other (no matter, if we use single linkage, complete linkage, Ward).

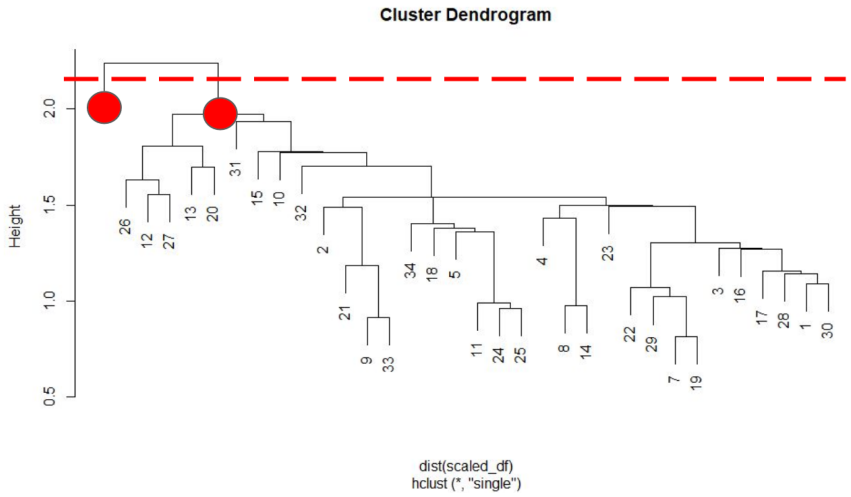
Dendrogram: graph that depicts the set of nested clusters resulting at each step of aggregation. The leaves of the dendrogram of a hierarchical clustering represent singletons (single data points, observations) whereas the root group contains the whole dataset (not interesting, at root there is not clustering (all in one group)). Each node represents a group. Each internal node has two children, representing the groups that were merged to form it.
Chaining: method of incorporating observations between clusters into existing clusters instead of initiating a new cluster.

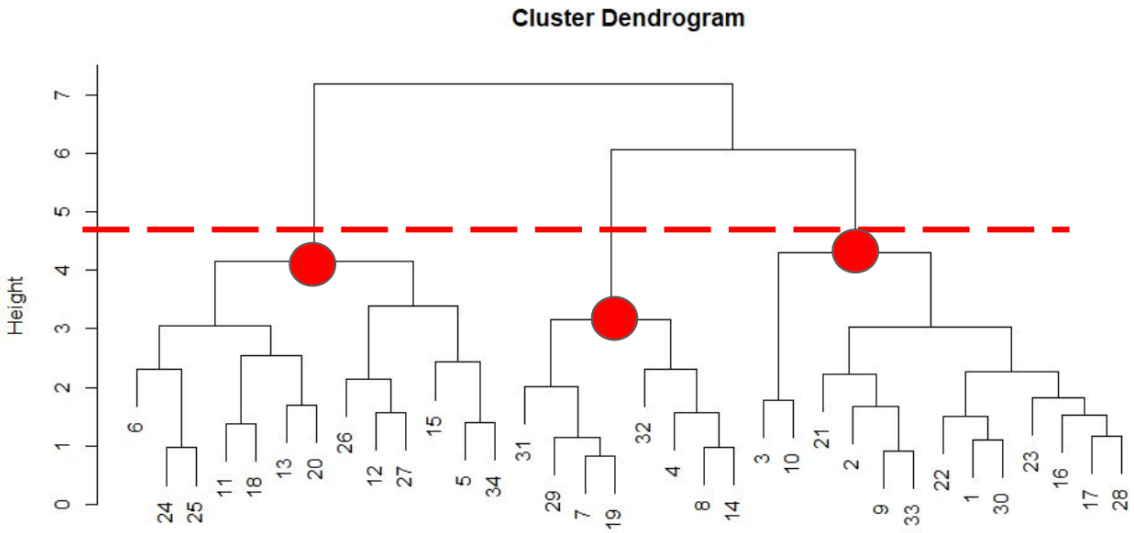
Pruning at a lower height in clustering continues to create very tiny clusters (often singletons) at the end.

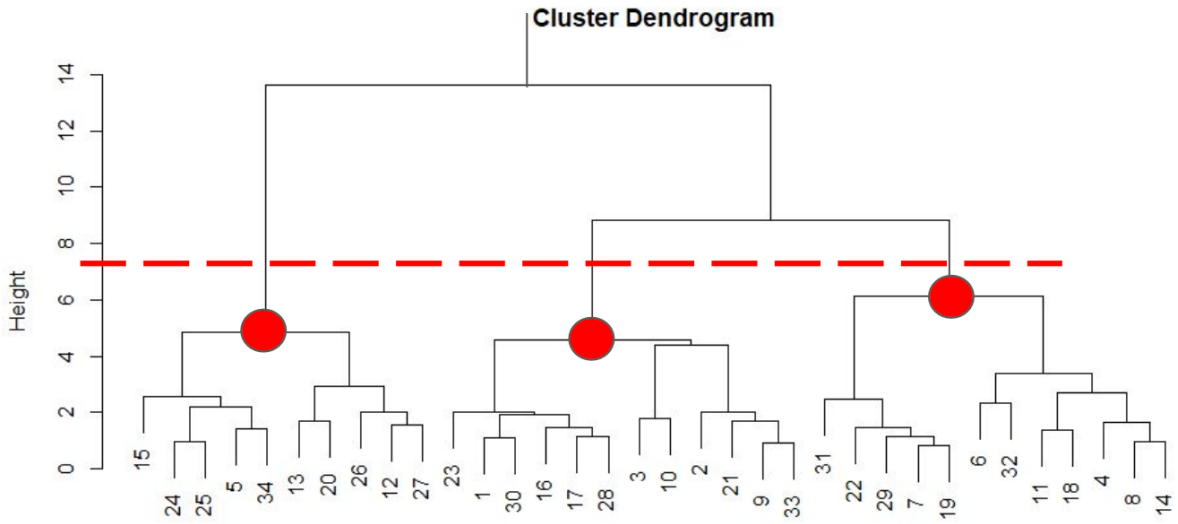
Where to cut:

In a dendrogram, locate the levels where a large distance exists between two consecutive fusions, and cut the tree there. Then at each fusion level, compute silhouettes measuring the intensity of the link of the objects to their groups, and choose the level where the within-group mean intensity is the highest (largest “silhouette width”), i.e. choose a line/level for which the clusters have the highest (average) silhouette value.

In a **dendogram**, locate the levels where a large distance exists between two consecutive fusions, and cut the tree there.







2) K-means algorithm: clustering method (not hierarchical) which has an a priori fixed number of clusters. The clusters are here represented by a mean vector (representing the "center" of the cluster).

   Step-by-step algorithm of the k-means clustering method:

   1) Start with k cluster centers
   2) Assign observations to the nearest cluster center
   3) Recompute the k cluster centers (mean of observations assigned to each cluster)
   4) If centers are the same as before then stop, otherwise go to step 2.

Assignment step (k means algorithm): Each observation is assigned to the cluster whose mean yields the smallest within-cluster sum of squares. Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean.

Update step (k means algorithm): The centroids or means of the observations in the new clusters are calculated with the formula of the arithmetic mean. Since the arithmetic mean is a least-squares estimator, this also minimizes the within-cluster sum of squares.

Due to the assignment step (assignment to the "nearest" group), the k-means algorithm tends to yield cluster of similar size (area of volume, not observation number). The clusters are typically convex in shape even when the true shape is not.

It can happen that we do not obtain the same clustering so as a solution we can start from random centers and perform clustering on them, then repeat it a certain number of times and see which one appears the most often-> it is the one that shows the real structure of the data. With clustering we obtain local optimum but not global optimum.

For hierarchical clustering the number of clusters can be assessed by inspecting the dendrogram or more formally by looking at the height of clustering heights or their difference. For k-means we do not have a dendrogram and different approaches are required -> Silhouette plots

The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. The silhouette score of i can take a value from -1 to 1, where when it is close to one it indicates that the observation is far away from the neighboring clusters and it belongs to the right cluster. If it is close to zero it indicates that the observation is on or very close to the decision boundary between two neighboring classes (it's not clear to which cluster the observation belongs to). Lastly, if it is close to -1 it means that the observation belongs to the wrong cluster.
Increasing the number of clusters decreases the quality of the individual clusters (to consider if i use the k-means clustering method).

3) Model-Based Clustering: consider the data coming from a distribution that is a mixture of two or more clusters. Each cluster is described by a density and has an associated probability or 'weight' in the mixture. It uses a soft assignment, where each data point has a probability of belonging to each cluster. Two steps:
Estimation step: find the expected value of the full data log-likelihood given the observed data and current parameter estimates.
Maximization step: maximize the expression over current parameter values to find the next parameter estimates.

We assign observations not to center with closest distance but with highest likelihood. Then we compute estimates of gaussian distribution

Clustering is an unsupervised learning algorithm: the class labels of training data are unknown. Given a set of measurement, the aim is to establish the existence of classes, clusters and structure in the data.
Methods like cross validation (which is a supervised method) don't work to measure the quality of clustering. In a supervised method the labels of the class of observations in the training data are known. The new data is classified based on the training set.
Supervised algorithms predict values that could be compared with correct answers.
Likelihood: data is fixed (n, sample size is fixed), we look at the probability for discrete variables and at the density for continuous variables.
Probability corresponds to finding the chance of something given a sample distribution of the data, while on the other hand, Likelihood refers to finding the best distribution of the data given a particular value of some feature or some situation in the data. The peak in the likelihood function corresponds to the real parameter of p (probability) (or another unknown parameter), or it is very close to it.

Questions on Lecture 3:

a) For two runs (starting with different centers) of k-mean clustering we will always get the same clustering results -> False, since data points might be assigned to a different center.
b) It is possible that the assignment of observations to clusters does not change between successive iterations in K-means -> True, at the end the algorithm converges, and the assignment does not change.
c) Clustering is an unsupervised method -> True
d) K-means algorithm automatically chooses an optimal number of clusters -> False, the number of k clusters is pre-determined.
e) When we use hierarchical clustering to split the given data set into three classes, the exact composition of the three classes depends on the specific linkage method used -> True.

PCA and clustering: no assumption on the data's distribution, unsupervised algorithm, no prediction, just describe and find patterns in data. PCA gives the 'why' to the clustering's result.

The combination of PCA and clustering can be useful in situations where you have a large number of features (i.e., variables) in your data, and you want to reduce the dimensionality of the data before performing clustering.

Learning goal: Understand the basic setup of a classification problem

1) Divide data into training set and testing set: The training set is used to train the classifier, while the test set is used to evaluate the performance of the classifier.
2) Train the classifier
3) Test the classifier

The goal is to, given a certain data that follows a distribution, to determine to which of the groups a realization belongs to, i.e. to classify the observation.

Compared to the case of clustering, we now know the group attribute of all observations, i.e., we have data from different groups or populations. The goal is to find a "separation" rule that separates the space of the variables or observations according to the different groups.

Classification is the process of classifying the data with the help of class labels. Classification is a supervised learning method (class labels are known), whereas clustering is an unsupervised learning method (no need to know the group/classes). In case of classification methos a training sample is provided, while in case of clustering the training data is not provided.

Supervised method: datasets that are supervised are designed to train or "supervise" algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over time. Examples of unsupervised methods: classification, regression.

Classification attempts to allocate new objects to predefined groups. Discrimination attempts to separate distinct sets of objects. In classification, the goal is to predict the class label of a given input data. A classifier is trained on a labeled dataset, where each input data point is associated with a class label. The classifier can then be used to predict the class label of new, unseen data points.

Discrimination is the process that defines how we separate one group from the other, Classification when we have data and have to say to which class the observations belong to.

When the groups have been decided, a new point (observation) is assigned to the group with the highest (parametric) density at that point. To select the best number of cuts/groups, cross validation can be used.

There exist different dimensionality reduction methods to solve classification problems:

- Linear discrimination analysis (LDA): used when the variance of the two groups is the same. In case of Gaussian bivariate densities (with equal variances) the parameters we have to find can be found by maximizing the likelihood of the density or by minimizing the error of the wrong classification. LDA fails if the discriminatory information is not in the mean but in the variance of the data. The goal of LDA is to maximize the distance between clusters.

- Quadratic discrimination analysis (QDA): used in case of unequal variances.

In terms of dimensionality reduction, LDA and QDA aim to find a linear combination of the input features that best separates the different classes in the data. This is done by maximizing the between-class variance and minimizing the within-class variance of the data. The result is a new feature space with fewer dimensions that still captures the most important information for classifying the data.

In terms of classification, LDA and QDA can be used as a classifier to predict the class label of a given input data.

If we do not have parametric densities for the individual groups, a standard likelihood approach cannot be used for classification. An alternative is to search a plane/direction to which the data is divided best into subgroups. We so want to find a direction to maximize the sum of squares between groups -> Fisher's Linear Discrimination Rule. It reminds of the PCA method but in PCA we just want to maximize the variance, in LDA we don't really care about the variance, we want to maximize the component axes for the class-separation.

Assumptions of linear discrimination (lda):

1) Equal variance between groups
2) Normal distribution on each group

Assumptions of quadratic discrimination (qda):

1) Normal distribution on each group

These assumptions are fulfilled with prior information (Bayesian discrimination). The normality assumption is hard to achieve in a multivariate set.

Under non-normality:

- LDA works better than QDA

- LDA tolerates few asymmetries and is robust under symmetric distributions.

- QDA deteriorates with higher number of predictors

Rules of thumb:

- if n1, n2 < 25, p < 6 and sigma1 == sigma1 then use LDA

- if sigma1!=sigma2 and p > 6, then use QDA with large n1 and n2

Classification Trees: breaks down the multivariate data into successive univariate rules that best separate the data. At each consecutive step the algorithm separates a subset of all individuals into two groups (according to a threshold on which we do the cut-off), augmenting the purity within each group, and thus creating a new node in the tree. The separation is based on a single variable that best discriminates between one group and all others. Classification trees are preferred over LDA and QDA since they have no assumption. Advantages of classification trees: very easy to understand, explain and visualize. Disadvantages: non-robust, new observations might change the cuts/classes division. To solve this we could use another method that allows to resample and train the model multiple times -> Bagging.

Bagging: abbreviation of Bootstrap aggregation. It tries to reduce the prediction variance by artificially increasing the training data. A bootstrap approach resamples from the original (training) dataset and creates a set of additional training sets of equal sizes. To each of the samples a simple classification algorithm is applied, and a separate prediction model using each training set is built. We then take the average of the resulting prediction for the classification.
While the predictive power of the method is not improved /we use the same type of algorithm for the original and the bootstrapped test sets), the variance in the prediction is reduced. Bagging typically prevent overfitting.
To compare a set of predicted class labels with observed ones, it is possible to display the confusion matrix: it is a matrix with ijth entry giving the number of predicted class i for the observed class j. A good classification yields a almost diagonal confusion matrix. If there is a strong predictor, bagging trees will use this strong predictor in the top split. Hence, the predictions from the bagged trees will be highly correlated. Averaging many highly correlated trees does not lead to as large of a reduction is variance as averaging many uncorrelated quantities.*

Boosting: based on the idea that it is possible to construct a strong classifier from many week classifiers. It calculates the output using several different models and then averages the result using a weighted average approach. The key is to wisely select the weak classifiers and to appropriately weight their result. A weak classifier is a classifier that performs slightly better than random guessing. The weak classifiers are trained on different subsets of the data and their predictions are combined in a way that gives more weight to the instances that are misclassified by the previous classifiers.

Random forest: most often used method out of the tree methods. It similar to bagging and decorrelates the trees: before fitting each new tree or split a random subset of predictors/features is chosen. For each tree we choose approximately predictors. At each split in the tree, the algorithm is not allowed to consider a majority of the available predictors. Individual trees are built based on a subset of features and observations. *Random forests overcome this problem by forcing each split to consider only a subset of the predictors. This approach prevents strongly correlated trees that may occur if there are very prominent predictors. Once a forest consists of a (determined) number of trees, the classification for a new one is given by the majority vote based on the classification of each tree.

Difference between bagging and boosting: The methods are built independently for Bagging, whereas boosting tries to add new models that do well where previous models fail. Bagging it’s best to avoid overfitting. Both methods make the decision by averaging the models but it is an equally

weighted average for Bagging, and a weighted average for Boosting. Both try to reduce the variance of the method and provide a higher stability, but only Boosting tries to reduce the bias, but boosting can also increase over-fitting. In boosting each classifier is trained on data, taking into account the previous classifiers' success. After each training step, the weights of the data points are redistributed. Misclassified data increases its weights to emphasize the most difficult cases.

In summary:

Classification tree is a decision tree that is used for classification.

Bagging is a technique used to improve the performance of a classifier by creating multiple versions of the training set and taking the average of the predictions.

Boosting is a technique used to improve the performance of a classifier by combining the predictions of multiple weak classifiers.

Random Forest is an ensemble method that combines the predictions of multiple decision trees which are grown on different subsets of the data.

Questions on Lecture 4:

a) LDA will fail, if the discriminatory information is not in the mean but in the variance of the data -> True
b) LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class -> True
c) In Random Forest, individual trees are built based on a subset of features and observations -> True
d) You have one very strong predictor in your data. Would you rather use bagging or random forest to predict your data? Random Forest, because in the collection of bagged trees, most or all of the trees will use this strong predictor in the top split. Hence the predictions from the bagged trees will be highly correlated. Averaging many highly correlated trees does not lead to as large of a reduction in variance as averaging many uncorrelated quantities. Random forests overcome this problem by forcing each split to consider only a subset of the predictors.

e) Under the hypothesis of Gaussian distributions, there is no theoretical difference between a linear discriminant analysis and a quadratic discriminant analysis -> FALSE. The theoretical differences rely on how the classification rules are defined in LDA and QDA (linear and quadratic discrimination rules). In the scenario where the variances among the different groups are equal, the classification rules will be approximate equals (the quadratic term in the QDA will be near 0).

f) Bagging does not improve predictive power of the method, but typically it avoids overfitting -> True: Bagging, by itself, does not improve the predictive power of a model because it simply trains multiple instances of the same model on different subsets of the data, and then combines the predictions. This can help to reduce overfitting, but it doesn't necessarily improve the ability of the model to make accurate predictions on new data.

g) Clustering a dataset into two groups will return the same result as applying a discriminant analysis with two classes -> FALSE. Clustering techniques assign labels to different unlabeled observations, while discriminant analysis gives a classification rule for given labeled observations. The results could match if for example the two classes are related to the best

separation in the sense of maximizing the between variability and minimizing the intra variability of the clusters, and if the discrimination rule is relatively simple, but this in general is more the exception than the rule

h) In random forest, the predictors get chosen at random for every tree. How are predictors chosen in Bagging? With a single classification tree. We choose a threshold and variable for which the split would be the cleanest.
Is boostrapping necessary in bagging? Yes, If you didn't do the bootstrapping, the method would not get more robust for outliers.

Learning goal: Understand the concept of the least squares criterion

The least squares criterion is a method for determining the line of best fit that minimizes the sum of the squares of the differences between the fitted values (calculated by the model) and the actual values (i.e., the values in the data set). This criterion is based on the assumption that the errors are normally distributed and independent from each other. The residual is equal to the distance between the observed value and the predicted/fitted value (value estimated by the model). The best line is the one that minimized the sum of the squared residuals (by using the least squares criterion).

Learning goal: Understand the statistical model of linear regression

Linear Regression is sensitive to outliers.

Sample data is used to find the estimated regression line.

Y = beta0 + beta1*x + e -> equation of the regression line. Beta0 is the intercept of the line with the y-axis and beta1 the slope of the regression line.

4 assumptions (to be checked with diagnostic plots): residuals have linear patterns, residuals are normally distributed, they have equal variance (homoscedasticity), check influential points (outliers).

We don't have to test for the normality of the data, but we have to for the normality of the residuals.

Linear regression is sensitive to outliers. Estimate/std.error = t.value.

The multiple linear regression model with p predictors is given by

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \\ &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad\qquad i = 1, \ldots, n, \quad n > p, \end{aligned}$$

where

- Y_i: dependent variable, modeling the observation, data y_i,
- $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$: free variables, predictors,
- $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top$: parameter vector (unknown),
- ε_i: error (unknown), with symmetric distribution and $\mathrm{E}(\varepsilon_i) = 0$.

Vector of true Betas is unknown.

We calculate the sum of errors squared so the errors only have positive values and won't cancel each other out. The goal is to find a vector of estimates such that the sum of squared errors is minimized.

The hat matrix, H, is also known as projection matrix because it projects the vector of observations y onto the vector of prediction y hat. This matrix determines the fitted/predicted values since y hat = H*y.

The vector of the estimated coefficients can be calculated with the formula: B_Hat = $(X^TX)^{-1}X^Ty$ -> least squares solution

Var(B_Hat) = $\sigma^2(X^TX)^{-1}$. σ^2 can be substituted by s^2 if sample variance is given.

Error terms have a different distribution that residuals because an error term represents the way observed data differs from the actual population, whereas a residual represents the way observed data differs from sample population data.

The variance of the vector of estimated Betas is unknown, so we need to estimate it with a t-distribution. We use the t-distribution and not the normal distribution since the variance is unknown (we work with estimated variance) and it looks closer to the variance of a t-distribution rather than a normal distribution.
Sample variance estimation (1/n-1 etc.) vs Regression variance estimation (1/n-p-1, because we estimate p parameters and not just one):
When we estimate the variance for a sample, it makes sense only when we have at least 2 points.
When we estimate the variance in linear regression, it makes sense only if we have more points than the parameters (p+1), otherwise, the sum of squared residuals would be 0.

Learning goal: Articulate assumptions for multiple linear regression, understand why we need to check the assumptions of the model, check the assumptions using diagnostic plots

Assumptions of linear models:

1) Linear relationship between the response and predictor variables (equal spreadness).
2) Normality: Error terms are normally distributed
3) Equal variance between the error terms (homoscedasticity) and zero mean
4) The errors are uncorrelated (no or little multicollinearity). To check it we need extra info: time sequence in which the data was collected (not checked in this course).
5) Independency of Observations:
   violated if there are:
   a) Repeated measures: several observations from the same individual (all other parameters kept fixed).
      Problems with repeated measures:
      1) Observations of the same (random) individual are more correlated than observations of different individuals.
      2) Parameters of each individual or level i might be slightly different.
   b) Hierarchical models: observations belong to subpopulations.
   c) Longitudinal settings: observations follow the same individual or subject over time. How do the observations change over time? Longitudinal data is a data that consists of repeated measurements of several individuals over time.

The observations in a dataset are not independent if the groups of observations are more correlated to each other than to the other observations.

Why is violating independency assumptions an issue? you run the risk that all of your results will be wrong.

Rule of thumb: to check independence, plot residuals against any time variables present (e.g., order of observation), any spatial variables present, and any variables used in the technique (e.g., factors, regressors). A pattern that is not random suggests lack of independence.

In R 4 diagnostic plots are returned:

- Residual vs Fitted values: to check for Linear relationship between residuals.

- Normal Q-Q Plot: to check for normality.

- Scale Location Plot: to check for homoscedasticity.

- Residuals vs Leverage Plot: to check for outliers/influential points (high-leverage points).

We assume that the rank of the matrix X (with dimensions n x (p+1)) containing the rows x_i^T has rank p+1.

T-value = Estimate/Std. Error

The error term is iid and normally distributed with mean 0 and variance o^2_i

As part of assessing the adequacy of the linear model, we should check if the error distribution is adequate and check if there is any evidence against the iid assumption. To do it we can use the 4 diagnostic plots.

Leverage: measure that indicates how much an observation influences its prediction. The closer a predictor is to the (hypothetical) center of the predictors, the smaller the leverage is. A high leverage is typically seen as twice the average of all leverages. The leverages in the Hat Matrix should ideally be roughly equal, we don't want leverages which are much higher than the others. If we have some very small ones, that would not bother us.

On a general level, a model fit is a partition of variances. The variability in the data is separated into variability components of the model and remaining variability (error). Naturally, we would like the remaining error variance to be as small as possible, and the variance explained by the model to be as high as possible.

If the error is just "somewhat" non-Gaussian it is often permissible to proceed by assuming Gaussianity.

Learning goal: Translate research questions involving slope parameters into the appropriate hypotheses for testing

The null hypothesis (H0) represents the assumption that there is no relationship between the independent variable and the dependent variable. The alternative hypothesis (Ha) represents the assumption that there is a relationship between the independent variable and the dependent variable.

Here are a few examples of research questions involving slope parameters and their corresponding hypotheses:

Research Question: "Is there a significant relationship between variable x1 and variable y?"

H0: The slope of variable x1 is equal to 0 (i.e., there is no relationship between x1 and y)

Ha: The slope of variable x1 is not equal to 0 (i.e., there is a relationship between x1 and y)

Learning goal: Understand the leave-one-out statistics and leave-one-out cross-validation and its applications for linear regression (cont. lecture 1)

Examples: DFFIT (difference in fit), DFFITS (difference in fits), COVRATIO (measures the influence of a single observation on the precision of the estimates). Another statistic is DFBETAS: measures how much a regression coefficient beta hat changes, in standard deviation units if an ith observation was deleted. The higher the worse (in abs. value). DFBETAS stays for difference in Betas. Cook's distance: consider location and response variable in measuring influence. The higher the worse.

Leave-one-out cross-validation (LOOCV) in linear regression is a technique to access the optimum model by omitting one observation at a time and fitting the linear model with the remaining observations. There is no need to refit the model for each omitted observation i. In regression, we can compute LOOCV statistics directly from the hat matrix (no need to refit the model multiple times). We also would like the LOOCV to be invariant to rotations, e.g. if we would like to apply PCA to our predictors. That's the issue GCV is addressing.

Learning goal: Use ANOVA to compare models

ANOVA is used to gain information about the relationship between the dependent and independent variables (i.e., to see if there is a difference between groups). If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1. ANOVA is a method for comparing models.

One way ANOVA is used to investigate if there are at least two mean values which show a significant difference. One way ANOVA has one independent variable, whereas two-way one has two independent variables.

ANOVA Table:

Source	Sums of squares	Degrees of freedom	Mean squares	F-value
Model	$\mathrm{SS}_{\mathrm{model}} = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2$	$\mathrm{df}_{\mathrm{model}} = p$	$\mathrm{MS}_{\mathrm{model}} = \frac{\mathrm{SS}_{\mathrm{model}}}{\mathrm{df}_{\mathrm{model}}}$	$F_{\mathrm{obs}} = \frac{\mathrm{MS}_{\mathrm{model}}}{\mathrm{MS}_{\mathrm{error}}}$
Error	$\mathrm{SS}_{\mathrm{error}} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$	$\mathrm{df}_{\mathrm{error}} = n - p - 1$	$\mathrm{MS}_{\mathrm{error}} = \frac{\mathrm{SS}_{\mathrm{error}}}{\mathrm{df}_{\mathrm{error}}}$	
Total	$\mathrm{SS}_{\mathrm{total}} = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$\mathrm{df}_{\mathrm{total}} = n - 1$		

To use ANOVA to compare models, you would first fit each model to the data, and then calculate the following statistics:

Total Sum of Squares (TSS): The total variation in the dependent variable.

Model Sum of Squares (SSM): The variation in the dependent variable that can be explained by the model.

Error Sum of Squares (SSE): The variation in the dependent variable that cannot be explained by the model.

Degrees of freedom (df): The number of observations in the data set minus the number of parameters in the model.

Then, you can use these statistics to calculate the following:

Mean Square Error (MSE) = SSE/df

Mean Square Model (MSM) = SSM/df

F-Statistic = MSM/MSE

Source	df	SS	Mean squares	F-value
Model	2	96	48	24
Error	12	24	2	
Total	14	120		

It's important to note that ANOVA assumes that the error terms across the different models are normally and independently distributed with equal variances (homoscedasticity). If these assumptions are not met, other statistical tests like likelihood ratio test or AIC should be considered.

```
> anova(fit1,fit2)
refitting model(s) with ML (instead of REML)
Data: jspr
Models:
fit2: math ~ social + raven + (1 | school)
fit1: math ~ social + gender + raven + (1 | school)
     npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
fit2   12 3556.6 3608.7 -1766.3   3532.6
fit1   13 3558.6 3615.0 -1766.3   3532.6 0.036  1     0.8496
```

From the R output here and in the solution, we can find that the difference between “fit1” and “fit2” is not significant; “fit2” performs significantly the best among “fit1-4”. So, we conclude that “gender” does not contribute to the model, and “fit2” is the best choice.

Learning goal: Know the types of research questions that can be answered using the linear regression and ANOVA

Is there a significant difference in the mean of variable Y between groups A, B, and C?

How much of the variation in variable Y can be explained by variable X?

Case in which the errors are not independent or identically distributed: If the variances vary for the observations, the weighted least squares (WLS) approach is applied; if the variances are not independent, the generalized least-squares (GLS) approach is used.

Multivariate regression is regression with multiple dependent variables. Multiple regression has instead multiple predictor variables.

A F-test is used to check if the ratio of two scaled sum of squares reflects different sources of variability. If the data values are independent and normally distributed with a common variance, and the model does not explain the data significantly better than just by taking the mean of the data, the ratio of two scaled sums of squares would follow an F distribution.

Another method for comparing models is Information criterion: it balances the goodness of fit of the estimated models with its complexity, measured by the number of parameters. In maximum likelihood estimation, the larger the likelihood function, the better the model is. Maximizing the likelihood may lead to overfitting.

The probability of a continuous variable to be equal some value is 0 because the area under the curve if we just consider that given value is zero. In reality continuous measurements are always rounded by some epsilon value.
AIC and BIC: the smaller, the better the model. IN BIC we penalize for the sample size n because otherwise one tends to choose overcomplex models when n is large.

To check if at least one of the predictors is useful in predicting the response, check the F-value.

To see if all predictors help to explain the dependent variable or only a subset if predictors is enough use step function for model selection (comparing AIC, BIC).

To know how well the model fits the data have a look at the adjuster R-squared statistic.

To know what response value we should predict and how accurate is out prediction given a set of predictors values, use predict() with a confidence interval.

To check if there are influential points use influential.points().

Complicated models with a lot of parameters are worse for prediction than simple models with just a few parameters.

A leverage observation is an observation which if it has an unusual value, it does not affect the estimates of the regression coefficients, on the other hand an influential observation has an unusual value but affects the estimates of the regression coefficients.

Linear regression and discriminant analysis: predictive methods, supervised, assumption that data is normally distributed.

Questions on Lecture 5:

a) In a linear regression problem, we are using “R-squared” to measure goodness-of-fit. We add a feature in linear regression model and retrain the same model -> Individually R-squared cannot tell about variance importance. (If R squared decreases, the variable is not significant).
b) We applied linear regression to model y~ x + z. We got the p-value 0.00001 corresponding to the beta_z coefficient. We can conclude that changing z by 1 will have a large influence on the value of y -> False, it suggests that there is a very strong association between the independent variable z and the dependent variable y. However, a low p-value does not necessarily indicate that changing z by 1 will have a large influence on the value of y. The magnitude of the effect of z on y is represented by the beta_z coefficient, which measures the change in y for a one-unit change in z.
c) Linear regression is sensitive to outliers -> True.

d) We would like to model the concentration of carbon dioxide in Earth's atmosphere using monthly mean carbondioxide levels measured at Mauna Loa Observatory. Would the assumptions of the linear regression be fulfilled? (It is not obvious without seeing the data, but what would you expect with this kind of data.) -> Probably not since the assumption of equal variance of error terms and normal distribution would not be fulfilled with time-series data.
e) We would like to model the happiness level of people depending on the time spent on doing sports. The activity of 30 people was measured, for each of them we got 3 data points collected in 3 weeks. Would the assumptions of the linear model be fulfilled? No, repeated measurements from individuals are correlated.
f) We would like to model the probability of failure of a construction element depending on the temperature. Would the assumptions of the linear model be fulfilled? No, since we need probability, we use logistic regression.
g) We consider the model $y = \beta o + \beta 1x + e$. Let $[-0.01, 1.5]$ be the 95%-confidence interval for $\beta 1$. In this case, a t-Test with significance level 1% rejects the null hypothesis $H0 : \beta 1 = 0$ -> False, 99% confidence interval includes the 95% confidence interval in itself, so if 95% does not reject it, neither will the 99% confidence interval.

f) Give two examples, where the assumptions of linear regression are violated. One example is when we take measurements from the same subject through time. This violates the assumption of independence between errors since for example, measurements taken close in time will tend to be similar to obs. more distant in time. A similar scenario lies when we have spatially located observations. Nearer observations will tend to be similar to more distant observations.

g) What alternative models can be used in those situations? How do they address the violation of the linear models' assumptions? For both scenarios we could extend the linear regression with a mixed effect model, now addressing the correlation of the errors by including a covariance structure in the model.

Learning goal: Explain the difference between random effect models and ANOVA

ANOVA assumes that the observations are independent and identically distributed (iid) and that the errors are normally distributed with constant variance.

Random effects instead assume that the observations are correlated and that the errors are normally distributed with constant variance.

ANOVA is used to test the difference in the mean between two or more groups, whereas random models are used to analyze nested data or data with hierarchical structure.

<u>Lecture 6 – Mixed Models</u>

Example: we would like to model the happiness level of people depending on the time spent doing sports. For each of the measured people we cannot say that the assumptions of the linear model are fulfilled, because an observation from the same person might not be independent. The independency of observations in a dataset cannot be seen through plots but usually if we consider a dataset with repeated measurements of the same observation we have a relation between the different experiments and thus the observations are not linearly independent.

If we create very similar data calculating the parameters results in a based and wrong result.

Mixed effects models are simply an extension of regression models that take into account the impact group membership has on an outcome of interest. They are used to model the relationship between a dependent variable and one or more independent variables when the observations are not independent, but rather are clustered or nested within certain groups. These models allow to account for the correlation between observations and to estimate the variance of the random effects, which improves the estimation of the coefficients and allows to test hypotheses about the random effects.

Learning goal: Explain random and fixed effects

Fixed effects are constant for all observations, while random effects vary. If an effect is assumed to be a realized value of a random variable, it is called a random effect.

For random effects, the general interest lies on the variable itself, not on the comparison of specific values of this variable. We are much more interested in modeling variances rather than individual levels.

Random effects are useful to compute the standard error of fixed effect.

Mixed effects model = random effect + fixed effect

$$Y_{ij} = \underbrace{\mathbf{X}_{ij}\boldsymbol{\beta}}_{\text{fixed effects}} + \underbrace{\mathbf{Z}_{ij}\boldsymbol{\alpha}_i}_{\text{random effects}} + \varepsilon_{ij}, \quad \boldsymbol{\alpha}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

X and Z are known matrices of appropriate size (here row vectors).

Fixed effects are the one we are interested in, random effects are effects which we are not interested in but have to be included in the calculation because they influence the model.

If then we are interested in an other parameter we fix that one and the one that was fix before becomes the random effect.

A fixed effects model is a model with only fixed effects, a random effects model is a model with random effects only. A mixed effects model consists of both fixed effects and random effects.

Learning goal: Use of random effects models

Random effect models allow to account for correlations or a correlation structure in the data. If there is a correlation it is implicitly captured in the model. The correlation can be seen if we have a look at the expected value, variance and covariances of the observations.

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{with} \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

Random effect model

In the Random Effect Model the vector of alphas and the error terms are independent to each other, so a covariance that involves one of the two or both of them is equal 0. In general, Cov(X, X) = Var(X)

Random effect models can be used when the independency assumption for linear models are not fulfilled, for example with:

- Repeated measures: we have several observations from the same individual (all other parameters kept fixed),
- Hierarchical models: observations belong to subpopulations (animals in stables in farms in regions . . .),
- Longitudinal settings: observations follow the same individual or subject over time

The estimation in a mixed model is not straightforward (easy) to perform, because unlike in the classical linear model framework, no closed form solutions exist.

Main approaches to estimation of model parameters:

- Maximum likelihood (ML) estimation: finds parameter theta for which an expression has the highest likelihood. Beneficial in comparing different 8nested) models but yields less stable variance components than REML.

- Restricted maximum likelihood (REML) estimation. This approach is computationally demanding and not often used in practice. However, even ML estimation is quite elaborate and based on profile likelihood steps.

- Markov Chain Monte Carlo (MCMC) approaches

When comparing arbitrary fixed effects models and mixed effects models, ML method is required. If the fixed effects terms are the same in all methods, either REML or ML is fine, but REML is usually preferred. The interpretation of the results is usually difficult when both fixed effects and random effects are changing. Hence, it is recommended to change only one or the other at a time.

A likelihood ratio test (LRT) is delicate when testing if one variance component is zero because the typical chi-squared distribution does not hold. Chi-squared distribution is used to describe the distribution of a sum of squared random variables.

Mixed models using lme4

Formula	Model
`(1 | group)`	random intercept within *group*
`(x | group)`	random slope for *x* and intercept within *group*, with correlation between intercept and slope
`(0 + x | group)`	random slope for *x* within *group*, no variation in intercept
`(1 | group1) + (1 | group2)`	random intercepts for two (crossed or nested) grouping factors
`(x || group)` or `(1 | group) + (0 + x | group)`	uncorrelated random intercept and random slope for *x* within *group*

Z|i: i specified the variable for which a random component is considered.

In mixed effect models there is no automatic p-value. The reason is that degrees of freedom are not easy to calculate in mixed models, a possible solution is to use bootstrapping.

Pooled observations are observations considered to belong all (or partially in partial pooling) to the same group/pool (and so model). No pooling observations are observations that do not share any similarities and are so entirely independent.

When we use mixed models we usually don't see any violation of the assumptions of the linear regression through the diagnostic plots but that doesn't mean that there is no linear relation in the data. Always check for repeated measurements, hierarchical models, longitudinal settings and check description of data/exercise.

A mixed model is a model that has a reduced estimation error, estimates fewer parameter and avoids problems of multiple comparisons. Mixed models not only account for the correlations among observations in the same cluster, they give you an estimate of that correlation.

Linear mixed models are often use when we have longitudinal data. The two random effects that are most common in this situation are a random intercept for each individual and a random slope for time for each individual. This model ignores within-subject correlation. As a sequence, the variability of coefficients is underestimated because correlated data contain less information that independent data. It is important to note that in cases where we do not have a balanced design (different number of observations at the same time points for each individual), a mixed model will lead to different estimates than a linear model. In general, by taking into account the correlation structure, a shrinkage of the estimates towards the mean can be observed.

Null hypothesis when comparing linear model H0: "the smaller model is sufficient".

Questions on Lecture 6:

a) The researcher is interested in the outcomes of one patient. He collects data about his performance on the test; he measures his performance multiple times. The researcher would like to build a model predicting the patient's performance, depending on the brain activity level. For the sake of this question, we ignore time and possible improvement of test results in time. In this case, we should use a mixed model because we have repeated measurements -> False, no more correlated groups of observations within our observations. But we model only this one patient. There are no subgroups of data here, we just collected data with one people so we cannot predict other patients with only one patient.
b) There was research done on the variables influencing salaries. There were 100 participants of age 35-45 in the study, from which many socio-behavioral variables were collected. The study took 5 years, and data from each participant was collected five times (i.e. once per year). A linear regression was conducted to
model the data. That was not a correct choice, as data coming from one individual might be strongly correlated. Because of that, the confidence intervals for the coefficients in the linear regression are probably wider, than they would be, when computed using the mixed model -> False, they are narrower if we omit the correlation between observations, rather than when considering the correlation.
c) Is the following assertion true? We cant interpret the REML by itself but when comparing two models, a higher REML indicates a better fit, right?
Yes, but only if we only change random effects. If fixed effects differ between the models, we would rather compare models using AIC or BIC, and then we need to fit models using ML, and not REML.

Learning goal: Describe the motivation for non-parametric regression

Nonparametric regression methods are useful when the relationship between the independent and dependent variables is not well understood, or when the data does not meet the assumptions of parametric methods (e.g. independency assumption is violated with repeated measurements). Nonparametric regression methods do not make any assumptions about the form of the relationship.

In the non-parametric regression, we don't know the general form of the function. We look for this function in a different way than by looking for parameters. In Linear regression we look for beta coefficients, in non-parametric we look for some function g(x) that describes the y data points as good as possible. Non-parametric methods are not useful for testing the effect of a predictor on the outcome (response) variable, since they are useful for understanding the distribution and not to study outcomes, like linear regression does. Non-parametric methods are not linear regression methods. The term "nonparametric" is not meant to imply that such models completely lack parameters, but rather that the number and nature of the parameters are flexible and not fixed in advance.

Many non-parametric techniques exist, and most can be organized into one of the following three approaches:

- Local estimation approaches (kernel smoothers, local polynomials)

- Penalized approaches (splines)

- Locally adaptive approaches (wavelets)

We extend the linear model to:

$$Y_i = g(x_i) + \varepsilon_i, \;\; \varepsilon_i \overset{\text{iid}}{\sim} (0, \sigma^2)$$

g is a twice differentiable function

Learning goal: Describe kernel approaches, penalized approaches, locally adaptive approaches

1) Local estimation approaches
a) Kernel Approach/Smoothers:

Definition 7.1. A kernel is a function $k(x)$ that

- is symmetric (around zero) and non-negative, i.e., $k(u) = k(-u)$ and $k(u) \geq 0$;
- is normed, i.e., $\int k(u)du = 1$;
- has a bounded second moment, i.e., $\int u^2 k(u)du < \infty$;
- is square integrable, i.e., $\int k(u)^2 du < \infty$.

Usually, one assumes that $\int u^2 k(u)du = 1$. A kernel is a density but not vice-versa.

If we want to know value of a point, we pick a window and calculate the mean of all the points within the window -> Box kernel approach: The box kernel assigns the same weight to

all points within a fixed-size rectangular region around the target point, and a weight of zero to all points outside this region. The size of the rectangular region is determined by the width and height of the kernel, which are specified by the user. For points outside of window we cannot compute the estimate. In practice we most often use the Gaussian Kernel, which assigns higher weights to points that are closer to the target point and lower weights to points that are farther away (assumption: data follows Gaussian distribution). With box kernel we can have no prediction for some points (if there are too far away or outside the box). Gaussian Kernel is described for whole domain -> we always have an estimation. In practice we could use any probability density that is symmetric as kernel. Lambda in kernel estimator determines how large is the kernel. If window is too large, we have influence also from observations far away, model will smoothen out too much and model would be just mean of observations as fitted model (horizontal line, strongly underfitted). If window is too small, model is overfitted (we just connect the dots, have no smoothing and many jumps and breaks). It determines the smoothness of the model. The key difficulty of kernel estimators is the choice of the bandwidth parameter lambda. The bandwidth determines the amount of smoothing and roughly said lambda that goes to infinity results in the mean. A sensible and careful choice of the bandwidth parameter balances bias and variance. Small bandwidth values result in a small amount of bias but large variance whereas for large bandwidth the relation is inverted. Choosing the smoothing/bandwidth parameter of non-parametric regressions is the classic example of cross-validation

Kernel estimator:

- the larger the weights, the closer xi is to the point x.

- the weights are equal in [x – c, x + c], otherwise 0.

- widths of the kernel can be adjusted using lambda

There exist different kernel functions: uniform, triangle, quartic, gaussian, cosine, etc. All of them satisfy the properties of kernel functions (see image above).

Kernel estimator (formula):

$$\widehat{g}(x) = \frac{1}{n}\sum_{i}\frac{1}{\lambda}k\left(\frac{x_i - x}{\lambda}\right)Y_i.$$

Ideas:
- the larger the weights, the closer x_i to the point x.

- the weights are equal in $[x - c, x + c]$, otherwise 0

- one can adjust the widths of the kernel using λ

Nadaraya-Watson estimator: type of kernel estimator for locally weighted average, using a kernel as a weighting function. Gives higher weights to points closer to observation we want to predict.

Formula:

$$\hat{g}(x) = \frac{\sum_i \frac{1}{\lambda} k\left(\frac{x_i - x}{\lambda}\right) Y_i}{\sum_i \frac{1}{\lambda} k\left(\frac{x_i - x}{\lambda}\right)}$$

→ weighted sum

Sum of weights

Lambda doesn't determine the length of the window (interval). Divide value by lambda and check if it is in the original kernel.

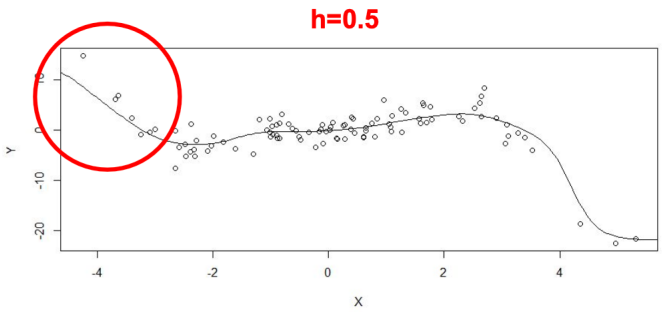
We want to choose the right smoothing so that the sum of squared residuals for the model is the smallest possible.

Problem with kernel approaches: if we are close to start or end of data, we get influence only from one side but not from the other and depends on how data points are distributed (Boundary issues) and to a lesser extend from bias by the design points {xi}. The kernel estimator fits a local mean at each point x and thus cannot even estimate a line without a bias.

Kernel smoothers are also used when estimating densities.

- Should we pick a wider or a narrower window to get a better fit?

A) Wider window
B) Narrower window



b) Local polynomials: another type of kernel approach but we fit model for window by using value for one given point instead of whole model. We can decide what type of polynomial the fit is (based on cross validation for example). We fit a polynomial function to a subset of data rather than the entire dataset. This subset of data is chosen based on the location of the point at which the estimate is desired and is typically chosen to be a small neighborhood around that point (according to weights, which are given by Nadaraya-Watson estimator). By fitting the polynomial function to a localized subset of the data, local polynomials can produce more accurate estimates near the edges of the data.
Each point has its window and based on points within this window, we fit a model with these points and then calculate prediction for a new data point with model we fitted. We have a moving window. We have a new model for each data point.
In the context of local polynomials, the smoothing parameter controls the degree of the polynomial that is fit to the localized subset of data.

A higher degree polynomial will have more flexibility and will be able to fit more complex patterns in the data. However, it will also be more prone to overfitting, which means it will perform well on the training data but not generalize well to new data. On the other hand, a lower degree polynomial will have less flexibility and will be less prone to overfitting, but it will also be less able to fit complex patterns in the data. We can use cross validation to choose best smoothing parameter.

Exam question: What is a smoothing parameter? it is a parameter that balances bias and variance. Typically is represented with a λ parameter, where a small value means that the fit ”follows” the data. On the other hand, large λ will lead to a linear regression fit.

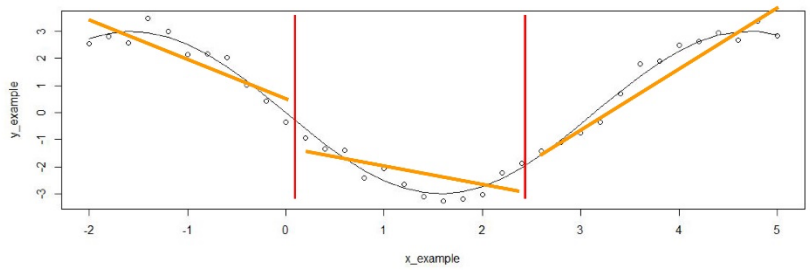
LOESS: locally estimated scatterplot smoothing
LOWESS: locally weighted scatterplot smoothing
In R, both LOESS and LOWESS implement local polynomials. Loess() has more features but is slower.

2) Penalized approaches:
Splines: model is built in pieces (there are fixed points where model changes), a spline is a piecewise function, where each segment is a polynomial. Splines are meant to be continuous and have continuous derivatives. Smoothing splines are designed to balance fit and smoothness. The smoothing spline is found by formulating a basis for the splines and then applying a ridge regression approach. In a ridge regression the model is penalized for a high number of coefficients, because too much variance is being explained (overfitting). The idea of splines is to Fit different regression line (low order polynomial) on each segment divided up by these knots or division points.
In Splines estimator there is a penalty: smoother functions yield smaller penalties. The smoothness can be measured with the second derivative. We force the function to be continuous (first derivative and second of the connected functions in point are same so connected functions are smooth). We mostly use cubic splines (splines with poly degree of 3). Penalizes for too many parameters or functions that are not smooth enough.
Splines vs Local polynomials: we want to fit regression line for whole segment at once and not for each considered point separately.
A spline is a piecewise function, where each segment is polynomial. Splines are meant to be continuous and have continuous derivatives. Smoothing splines are designed to balance fit and smoothness.

$$\widehat{g}(x) = \underset{g}{\operatorname{argmin}} \sum_{i=1}^{n} \big(y_i - g(x_i)\big)^2 + \lambda J(g),$$

where λ is a smoothing parameter and J is a penalty function, i.e., smoother functions yield smaller penalties.

Smoothness can be quantified with derivatives, in fact the penalty function is the integral of the second derivative of $g(x)^2$. A sweet spot (balance) has to be found between the smoothness of the penalty function and smoothness of model regarding SSR. First parameter is to minimize sum of squared residuals and second is penalty function multiplied by the smoothing parameter lambda. Not all the points are knots, splines pass through knots.



From this model we could add restrictions, such as continuity, to make it smoother. A variable is computed for each knot-region and they are then inserted into the model.

In splines to represent the fit we use a set of basis, one easy to understand is called Truncated power series basis:

$$\{1, x, x^2, x^3, (x-\xi_1)_+^3, \ldots, (x-\xi_k)_+^3\}, \text{ where } x_+ = \max(x, 0).$$

Greek letters represent knots (index represents number of knot: first knot, second, etc.).

Smoothing splines and wavelets are particularly suitable for extensions in two dimensions.

3) Locally adaptive approaches: We want window size to change and have more precision in some areas and less in other areas (adaptation of window size only allowed here, for other approaches it is fixed). Modeled with wavelets. The key feature of wavelets is that they are "locally adaptive" which means they can be adjusted to the local features of a signal or image. Wavelets can be thought of as a set of functions that can be scaled and translated to match the characteristics of the signal or image being analyzed.

Learning goal: Know the strengths and weaknesses of the various methods introduced

Kernel smoothers: + simple and computationally cheap, - sensitive to choice of bandwidth parameter, suffer from boundary issues.

Local polynomials: + flexible, produce more accurate estimates near the edges of the data than Kernel smoothers. – sensitive to the choice of bandwidth parameter.

Penalized approaches (splines): + flexible, can handle both smooth and non-smooth functions and both univariate and multivariate data. – sensible to the choice of knots, can be computationally expensive

Locally adaptive approaches (such as wavelets) : + Can adapt to the local features of a signal or image, Can provide a multi-resolution representation of a signal or image. - Computationally intensive

Questions on Lecture 7:

a) How are splines different from local polynomials? In splines we want to fit the regression line for the whole segment at once, not for each considered point separately.
b) Every kernel is a density, but not every density is a kernel -> True, not all density functions are symmetric, but all Kernels are.
c) What scientific questions can be answered with a non-parametric model? Find a question that can be answered with a non-parametric model, and another one that can be answered with a linear model. -> In a non-parametric model, we state a flexible non-pre-specified relationship between -for example- two variables. This flexibility leads to a useful tool for

discovering relationships and building prediction models. A question for a Parametric linear model could be those over the relevance of a parameter of the model, let say the slope β1, such as 'Does an increase in one unit of X changes the response value?'. On the other hand, a non-parametric regression approach would be more on the side of the behavior type of questions.

Aggregated dataset: when each of the rows of the dataset correspond to specific profiles (specific combinations of the variables, e.g. one value for pressure and one gender), rather than being related to each individual. Aggregated datasets have the information of how many observations correspond to that specific profile (n_i in our toy example). Aggregated datasets correspond to the typical dataset-style for binomial glm.

Linear model for the log odds:

$$g(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x.$$

The probabilities are linked to one or several predictors using the logit function:

$$\operatorname{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

Learning goal: Explain the difference between the linear and logistic regression

Linear -> parametric regression, logistic -> nonparametric (no assumption about distribution of data).

Linear -> assumes normal or gaussian distribution of dependent variable. Logistic -> assumes binomial distribution of the dependent variable.

Linear regression: supervised regression model. Logistic: supervised classification model.

Linear based on LSQ estimation whereas logistic is based on ML estimation.

Logistic regression is a special case of generalized linear models (GLMs), which extend ordinary linear regression to situations where the response variable -as in this scenario- is binomial. This allow us to a more natural treatment of this type of data, providing predictions that lives in the domain of a probability [0,1], which cannot be guaranteed when using a linear regression. Logistic regression is not used to predict the outcome of a variable, but rather to study the outcome probabilities. When we do a linear regression, we estimate both the mean and the variance from the data. In contrast, with logistic regression, we estimate the mean from the data, and the variance is derived from the mean.
The link function g(.) provides the relationship between the predictor and the mean of the distribution function. It describes how the mean response (E(y) = mu), is linked to the covariates through the linear predictors (n = g(mu)). In poisson regression, the link function is log(), whereas the inverse of the link function is exp. The inverse of the link function ($g^{-1}(.)$) is the logistic function and thus a regression problem based on this function is called logistic regression.

For a change predictor variable unit (x) -> predictor variable unit (x) + 1, odds -> odds * e^beta_1 (coefficient of x)

The probit function is the inverse of the cdf of a standard normal distribution.

Learning goal: Define Generalized linear model

Generalized linear models generalize linear regression for all the distributions that belong to the exponential family.

A GLM consists of three components:

1) Distribution of Yi (independent conditional on the predictors). E.g. binomial in logistic regression, poisson in Poisson regression.
2) A linear function of predictors $x_i^T B$ (B: vector of coefficients)
3) A function linking E(Yi) = ui and the predictors (link function) $g(ui) = x_i^T B$

The inverse of the link function is useful to calculate the relationship between the parameters and the coefficients. If we are interested in the probability, we use the inverse of the log function.

Linear regression is also a generalized linear model: The distribution of Yi is the normal distribution. There is a linear function of predictors $x_i^T B$ and the link function of linear regression models is the identity function, it maps every element in a set to itself (in other words, it directly predicts the outcome). $Y \sim N_n(XB, o^2 I)$.

In generalized linear regression: we do not look at additive errors but only the distribution of observations (parameters of distribution and probabilities, e.g. binomial). The predictors are used additively for transformed parameters of this distribution. The sum of squared residuals is not minimized (result would be biased since the distances depend on the values of x, the variance would not be same for a different x). Exact distributions of estimators cannot be derived and therefore confidence intervals are typically not exact. Assumptions of generalized linear models: linear relationship and normally distributed error terms.

Deviance: quantity describing the quality of a fit and is a generalization of the classical sum of squared residuals (SSR). More specifically, instead of the SSR, the difference between the log-likelihoods of two models is used. The larger the deviance, the poorer the fit to the data. The model with the perfect fit is called saturated model, and it is used as benchmark model for the deviance. Model deviances are not interpreted directly, but compared with each other.

Residual deviance: saturated model – fitted model. Measure of how well the data is explained by the fitted model.

Null model: corresponds to the null hypothesis and for the simplest cases consists of only one parameter (intercept beta0).

Null deviance: saturated model – null model. Measure of how well the data is explained by the null model, which is the model that assumes that the response variable is independent of the predictor variables.

Fisher's scoring algorithm is a derivative of Newton's method for solving maximum likelihood problems numerically. The smaller the number of Fisher scoring iterations, the faster the model converged (the less it took to fit the model).

Overdispersion: occurs when errors (residuals) are more variable than expected from the theorized distribution.

In generalized linear models we do not predict the outcomes, but the outcome probabilities. We do not use type = "response" since even on the response level, the prediction does not exactly yield the estimation of Y.

In R: if log_dose is 1.96, an increase of one unit in dose increases the odds of dying by exp(1.96) = 7 times.

Learning goal: Checking if a distribution belongs to the exponential family of distributions

A distribution falls into the exponential family if its distribution function can be written as:

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

, where the parameter of interest tetha = g(mu) depends on the expected value of y. The parameter Ø models dispersion (variance).

g(mu) is then a natural link function for the considered distribution.

To find the exponential family of a distribution calculate $e^{\log(\text{density function of distribution})}$ and bring it to the formula of the exponential family of distributions.

Table 8.1: Some members of the exponential family and the natural link function.

Distribution	Link function	
Gaussian	Identity:	$g(\mu) = \mu$
Binomial	Logit:	$g(\mu) = \log \frac{\mu}{1-\mu}$
Poisson	Log:	$g(\mu) = \log \mu$
Gamma	Inverse:	$g(\mu) = 1/\mu$

The logistic regression can be used to classify a set of observations into two (or possibly more) groups. For example, we predict the probabilities of belonging to a certain class given covariates. For two groups, we may classify the observations in one of the groups if probability is below ½ and vice versa (e.g. LD50).

Learning goal: Understand some similarities and differences between LDA and logistic regression

Both are supervised classification techniques. Logistic regression relies on ML estimation, while LDA relies on Least-squares. Logistic regression can adopt more complex boundaries between groups. They both have assumptions, one of the critical assumptions of logistic regression is that the relationship between the logit (aka log-odds) of the outcome and each continuous independent variable is linear. LDA assumes that the data is normally distributed. LDA makes the assumption that the classes have equal covariance matrices, while logistic regression does not make this assumption. In LDA we don't have other predictors, logistic regression is broader (can have more variables). LDA has stronger assumptions, is simpler and we can work with very few data points but applications are more limited than with logistic regression.

Questions on Lecture 8:

a) The logistic regression can be used to classify a set of observations into two groups, similarly as the linear discriminant analysis. Hence the assumptions for those two methods are the same -> False, logistic regression doesn't assume equal variance between groups and does not assume data that is gaussian-distributed.
b) Logistic regression is fitted using the Least Square Error method -> False, using ML.
c) Logistic regression is a supervised algorithm -> True, we have labeled data with correct answers.

Learning goal: Describe survival data, the roles played by censoring, survival and hazard functions

In Survival Analysis we model observations representing the time until the occurrence of a certain event. Survival data describes the duration from a starting event until an end event (e.g. time until a volcano erupts, life of a tiger, etc.).

Continuous distributions, which are defined for non-negative numbers (since time can't take negative values) are appropriate for modeling survival data. From the distributions we know so far the most appropriate one is the exponential distribution; another function is the Weibull distribution. The exponential function is a special case of the Weibull distribution. The exponential distribution should be preferred over the Weibull when there is a constant failure rate.

Censoring: information on time to outcome event is not available for some participants to the study.

In many cases, the end of the study occurs before every individual in the data set has died or experienced the event of interest. The individuals that are still alive and observed at the end of the observation time are called censored. All of these situations are called "right-censoring", the most common censoring mechanism in survival data. An alternative is left censoring, exemplified by when a disease has been diagnosed but the exact time when the disease first began is unknown. Censoring is assumed to be independent (non-informative) of the survival time or any other relevant information of the individual/case. Kaplan-Meier estimator or Cox proportional hazards model are used to account for censorship in the data.

The survivor function describes the surviving time, it calculates the probability that an event happened after time t.
The hazard function (or hazard rate) is the probability that if the event in question has not already occurred, it will occur in the next time interval, divided by the length of that interval. It is not a density, it is written as conditional probability and when we integrate it may integrate to more than 1.
The cumulative hazard function is the integral of the hazard function. It is the cumulative risk of death up to any specified time given that death has not occurred until then. The cumulative hazard is not a cdf and can therefore exceed one.

Definition 9.1. For a survival time T with pdf $f_T(t)$ and cdf $F_T(t)$ we define the following functions.

1. The survivor function

$$S(t) = \mathrm{P}(T \geq t).$$

2. The hazard function or hazard rate

$$h(t) = \lim_{\Delta t \searrow 0} \frac{1}{\Delta t} P\{t \leq T < t + \Delta t \mid T \geq t\}.$$

3. The cumulative hazard function

$$H(t) = \int_0^t h(u)du.$$

Survivor function S(t) = 1 - F(T) (cdf), and gives probability that patient survives until time t.
The pdf is the derivative of the cdf. The cdf is used to be sure hat the predicted probabilities lie between 0 and 1.
The idea of the proportional hazards model is that the hazard function can be split into a baseline hazard component and covariate effects component. The hazard function of individual I with covariates xi is:

$$h_i(t; \boldsymbol{x}_i) = h_0(t) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}),$$

, where h0(t) represents the baseline hazard rate.
The Cox proportional hazards model makes two assumptions: (1) survival curves for different strata must have hazard functions that are proportional over the time t and (2) the relationship between the log hazard and each covariate is linear, which can be verified with residual plots.

Property 9.1. *For a continues survival time T, we have for $t \geq 0$:*

1. $h(t) = \dfrac{f(t)}{S(t)}$,
2. $S(t) = \exp\big(-H(t)\big)$.

➔ H(t) = -log(S(t))

One common (theoretical) distribution of survival times is the Weilbull distribution. The density for $t > 0$ is given by:

$$f_T(t) = \frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1} \exp\left(-\left(\frac{t}{\beta}\right)^{\alpha}\right)$$

with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$. Note that for $\alpha = 1$ we have the classical exponential density with rate $\lambda = \frac{1}{\beta}$

For t = 1, the Weibull distribution is the exponential distribution. For 0 < k < 1, the Weibull distribution is strictly decreasing. For t > 1, the Weibull distribution has exactly one maximum point.
The Kaplan-Meier Estimator is a non-parametric method used on survival data which is used to estimate the survival function S(t). The Kaplan-Meier curve illustrates the survival function, it's a step function illustrating the cumulative survival probability over time. The curve is horizontal over periods where no event occurs, then drops vertically corresponding to a change in the survival function at each time an event occurs. The steeper the curve, the worse the survival experience. The Kaplan-Meier estimator is a non-parametric method for survival analysis.
A typical goal when analyzing survival data is to compare survival rates under different treatments (such as treatments with different drugs) and testing if one of them is better. A simple non-parametric test to be used in such a situation is the log-rank test. There are two major problems associated with the application of the log-rank test: first, it can be used only to compare categories of one categorical covariate, but not for several (possibly continuous) covariates. Second, no quantification of a covariate effect is possible, only a difference can be stated.
Binomial vs Hypergeometric distribution: they both describe the number of times an event occurs in a fixed number of trials but in binomial distribution the probability is the same for every trial, whereas in hypergeometric each trial changes the probability for each subsequent trial because there is no replacement. Approximation of the binomial distribution is the normal distribution.
Non-parametric methods are useful in the analysis of a single sample of survival data, or comparison of two or more groups of survival times. Cox regression, also known as proportional hazards regression, is a statistical method used to analyze the relationship between a set of predictor variables and the time it takes for an event of interest to occur (i.e. time-to-event data or survival data). The proportional hazards assumption states that the effect of a predictor variable on the hazard rate is constant over time. Cox regression can be used to test whether this assumption holds for a given set of data, by comparing the fit of a model that assumes proportionality to one that allows the effects of predictors to vary over time. If the two models have similar fit, then the proportional hazards assumption is likely to hold. In cox regression methods the demographic, physiological and life-style variables are recorded and used in the model.
Concordance: fraction of all pairs of subjects whose predicted survival times are correctly ordered among all subjects than can actually be ordered.
As a general rule, if the hazard plots cross, the Cox proportional hazards model is not appropriate. It is not the hazard function we are looking at with these plots, but residuals.
By looking at the hazard function instead of the survivor function, one can quantify the effects of covariates ad make statements such as "the risk in group 1 is beta times higher than the risk in group 2".

Disadvantage of logrank: we can say the survival experience of the groups is different or equal (with or without medicament) but we cannot include demographic information or by how much it is different, to look at this we look at cox proportional hazards

- $n_{1,k} :=$ number of individuals in Group 1 at risk at (or shortly before) $t_{(k)}$
- $n_{2,k} :=$ number of individuals in Group 2 at risk at (or shortly before) $t_{(k)}$

- $d_{1,k}$:= number of events in Group 1 at $t_{(k)}$
- $d_{2,k}$:= number of events in Group 2 at $t_{(k)}$.

The setting can be summarized by the following 2×2 table:

Group	# of events	# surviving beyond $t_{(k)}$	# at risk at $t_{(k)}$
1	$d_{1,k}$	$n_{1,k} - d_{1,k}$	$n_{1,k}$
2	$d_{2,k}$	$n_{2,k} - d_{2,k}$	$n_{2,k}$
Total	d_k	$n_k - d_k$	n_k

The null hypothesis of the log-rank test is

$$H_0 : S_1(t) = S_2(t),$$

and the corresponding alternative hypothesis is

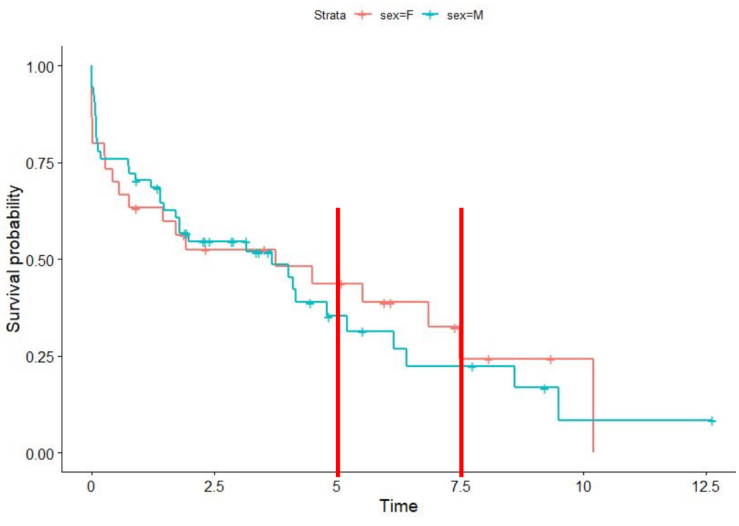
$$H_1 : S_1(t) \neq S_2(t).$$

- Who seems to have a better survival experience between 5 and 7.5 years?

less sleep = better surv. experience

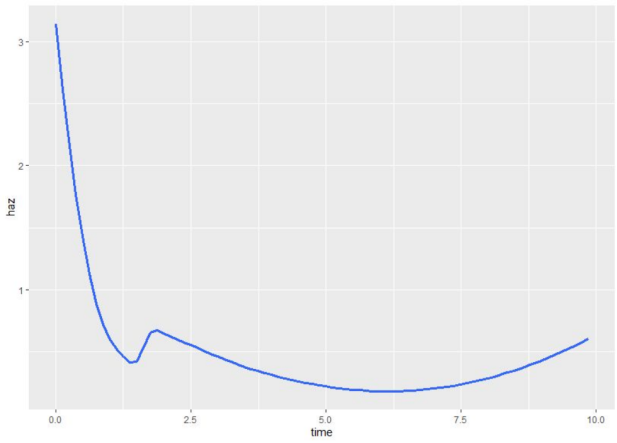
sleepiness has impact on surv-experience

- A) Red
- B) Turquoise



sleepiness → worse survival experience

- Given the estimation of the hazard function, which distribution would suit better?



- A) Exponential → assume constant failure rate → special case of Weibull
- B) Weibull
- C) no idea

- If we would like to report how the risk/hazard changes for a one-year increase in age, we should look at:
- A) Red
- B) Blue

```
Call:
coxph(formula = surv ~ age, data = df)

       coef exp(coef) se(coef)     z     p
age 0.02146   1.02170  0.02327 0.922 0.356

Likelihood ratio test=0.86  on 1 df, p=0.3537
n= 84, number of events= 55
```

Questions on Lecture 9:

a) The heights of the steps in the Kaplan Meier plot are all equal for a considered data set (even if we have some censored data) -> False.

b) The mathematical formula of Kaplan-Meier curve reflects the 'conditional probabilities'. They are conditioned on being at risk (alive and not censored) at each event time -> True.

$$h_k = \mathrm{P}\left(T \in [t_{(k-1)}, t_{(k)}) \mid T \geq t_{(k-1)}\right), \tag{9.10}$$

i.e., the conditional probability of death in interval k, given that interval k has been reached.

c) When checking statistically the assumptions for Cox regression, the p-value below the assumed significance level means, the assumptions are not fulfilled -> True.

d) There is no intercept in the Cox model -> True, we only look at parameters.

e) If censored observations tend to have worse prognosis than those remaining in the study: Observed survival > True survival -> True.

f) Hazard function is a density -> False, e.g., for the exponential distribution, the hazard function is constant from 0 to infinity, meaning that it may not integrate to 1.

Learning goal: Identify important features on a time series plot

A time series plot is a graph that displays how a variable changes over time. Important features of a time series plot are:

- Seasonality: Repeated patterns that occur at regular intervals, such as daily, weekly, or yearly.
- Stationarity: a Time Series is said to be stationary if its statistical properties such as mean, variance remain constant over time -> autocovariance is constant. Weak stationarity: time series has a constant mean and a constant variance over time, but its autocovariance function (ACF) depends on time. The mean and the autocovariance function only depend on the time lag between observations and not on the specific time point. Strong stationarity: time series has a constant mean, variance, and autocovariance function. not only the mean and the autocovariance function, but also the probability distribution of the time series is constant over time and does not depend on the specific time point or the time lag between observations.

Learning goal: Identify and interpret an AR(p), MA(q)

Autoregressive model: the current observation depends on the previous one. An autoregressive (AR) model predicts future behavior based on past behavior. It's used for forecasting when there is some correlation between values in a time series and the values that precede and succeed them. If we use all past observations to compute the future one it could lead to overfitting, we will end up with a more complex model than needed. For simplicity we start by assuming that the time series has zero mean. The term autoregressive refers to the fact that the model is based on the idea that the value of a variable at a given point in time is related to the values of the same variable at previous points in time. AR(p) is a model with p lags (a lag refers to the number of time periods that have passed between the observation of a variable and the observation of the same variable at an earlier point in time).

An AR(p) model calculates the difference in value between the current time period and the past p time periods. Specifically, the value of the variable at the current time period (Y(t)) is modeled as a linear combination (dependent relationship) of the past p values of the variable (Y(t-1), Y(t-2), ..., Y(t-p)) plus an error term (e(t)). In other words, the model takes the past p values of the variable as input and uses them to predict the value of the variable at the current time period.

Moving average model: model where the observation at current time depends on a weighted average of white noise components (error terms). Moving average means that the model exploits the relationship between the residual error and the observations. In an MA(q) model, the variable of interest is modeled as a function of the past q errors or residuals (e(t-1), e(t-2), ..., e(t-q)) plus some constant term (mu) which is the mean of the series.

Autoregressive Integrated Moving Average (ARIMA) model: combines AR(p) and MA(q) together.

Learning goal: Interpret PACF, ACF (used to find best number of lags)

The PACF plot is used to identify the number of lags to include in a autoregressive (AR) model, while the ACF plot is used to identify the number of lags to include in an moving average (MA) model.

For an AR(p) model, the PACF vanishes after p lags. For an MA(q) model, the ACF vanishes after q lags. Hence, the PACF and ACF can be used to select an appropriate model. ACF is used to identify

patterns that repeat at a certain interval, while PACF is used to identify direct relationships between a variable and its lags. When the ACF plot shows a strong correlation at a certain lag value, it means that there is a pattern in the data that repeats at that interval. When the PACF plot shows a strong correlation at a certain lag value, it means that there is a direct relationship between the variable and its lag value, after controlling for the effect of the intermediate lags.

In the case of the PACF plot, if the plot shows a strong correlation at a certain lag value, and then drops off abruptly, it suggests that an AR model with that lag value may be appropriate. In the case of the ACF plot, if the plot shows a strong correlation at a certain lag value and then decays exponentially, it suggests that an AR model with that lag value may be appropriate.

Learning goal: Distinguish AR terms and MA terms from exploring an ACF and PACF

In an ACF (Autocorrelation Function) plot, if there is a significant correlation at lag k, it may indicate that there is an MA term of order k in the underlying time series model.

In a PACF (Partial Autocorrelation Function) plot, if there is a significant correlation at lag k, it may indicate that there is an MA term of order k in the underlying time series model.

If there is a significant correlation at a lag k in both the ACF and PACF plots, it may indicate that an ARIMA model is appropriate.

In the context of time series, prediction of observed values is called smoothing of future (unobserved) values or a forecast.

Seasonal Trend Decomposition (STL): An algorithm to divide up a time series into three components: the trend, seasonality and remainder.

$$\mathrm{E}(Y_{t+n} \mid Y_t) = \phi^n Y_t$$

$$\mathrm{Var}(Y_{t+n} \mid Y_t) = \sigma^2 \sum_{i=0}^{n-1} (\phi^{2i})$$

Questions on Lecture 10:

a) Autocovariance measures linear dependence between two points on the same series observed at different times -> True.
b) It is a necessary condition for weakly stationary time series that the autocovariance function depends on s and t only through their difference |s-t| (where t and s are moments in time) -> True.
c) Autocovariance function for weakly stationary time series does not depend on the location of point at a particular time -> True.
d) Adjacent observations in time series data are independent and identically distributed -> False, they are dependent and identically distributed.
e) What is an AR(1) model and what are the implications of using such a model compared to an iid setting? It is a type of autoregressive models with a single lag. An AR(1) model assumes that the current value of a variable depends on its previous value, whereas an iid setting assumes that each observation is independent of all others. IId setting does not make any

assumption about stationarity, but an AR(1) model assumes that the mean and the variance of the time series are constant over time. AR can be used for forecasting, iid cannot.

f) Explain what covariance function in spacial process is and how it is estimated: the covariance function encodes the behaviors between observations that are spatially located. This will represent how strong they covariate. Nearer distances are expected to be more similar than more distant observations. To estimate the covariance function, we should look for a suitable model that truthfully represents these behaviors. Then, one approach is the traditional ML framework, where we could proceed with a profile likelihood approach, updating iteratively the fixed-mean part and the covariance structure.

Lecture 11 – Spatial Statistics

We can expect similarity between points that are close to each other and dissimilarity between those that are far away from each other.

Interpolation: prediction of an unobserved quantity in the domain.

Gaussian spatial process: for every variable in the spatial domain, that variable is a normal random variable.

Sample surface: realization of a spatial process, set of data measured at locations s1, …, sn in domain D. It's the subset of a potentially infinite number of measurements.

$$\{z(\boldsymbol{s}_i) : \boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \in \mathcal{D}\},$$

We observer the sample surface at a finite number of locations in the domain.

Learning goal: Define stationarity of a process, additive decompositions of a process

Isotropic spatial process: properties (mean, variance, covariance) are the same in all directions. Covariance only depends on distance and not on direction.

Anisotropic spatial process: properties (mean, variance, covariance) vary on direction. Covariance depends on distance and direction.

Second-order stationarity (Weak stationarity): constant mean and variance and a covariance structure that is translation invariant (covariance depends on the distance between the points and not on their specific location, if you shift all the points in the same direction or same distance, the covariance between the points will remain the same.).

Intrinsic stationarity: similar to second-order stationarity, but it takes into account the underlying structure of the process. A process is said to be intrinsically stationary if it is stationary after a certain transformation or change of coordinates. In other words, it is a process that appears non-stationary when observed in one set of coordinates but becomes stationary when observed in another set of coordinates.

For example, a spatial process that is anisotropic, meaning that the properties of the process vary depending on the direction, can be made isotropic, or stationary, by rotating the coordinates system. After the rotation, the properties of the process will not depend on the direction anymore, and the process will be considered intrinsic stationary.

Stationarity is a property of the process and not of the data, so it can't be tested or proven from the data directly.

Additive decomposition:

$$Y(\boldsymbol{s}) = \mu(\boldsymbol{s}) + Z(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}), \qquad \boldsymbol{s} \in \mathcal{D},$$

- $\mu(\cdot) = \mathrm{E}(Y(\cdot))$ is the deterministic mean structure, called large-scale variation;
- $Z(\cdot)$ is a zero-mean, stationary process. The process $Y(\cdot)$ is called smooth small-scale variation;
- $\varepsilon(\cdot)$ is a zero-mean white-noise process, independent of $Z(\cdot)$ and is considered as measurement error. We often assume $\varepsilon(\boldsymbol{s}) \sim \mathcal{N}(0, \sigma^2)$.

The large-scale variation $\mu(\cdot)$ is also called trend or drift.

Additive decomposition is used to decompose a spatial variable into two or more components that represent different spatial patterns or scales. The trend component represents the underlying spatial pattern of the variable.

Variogram and semivariogram: function that expresses the spatial or temporal dependency structure of a stationary process. It is used in geostatistics and spatial statistics to describe the spatial or temporal variability of a variable, such as temperature or precipitation. It is so used to describe the similarity between points at different locations. A variogram is a measure of the variance of the difference between two random variables. It is defined as the expected value of the squared differences between values at two locations, divided by two. A semivariogram is similar to a variogram, but it is defined as the expected value of the squared differences between values at two locations, divided by one (half the variogram).

Learning goal: Explain a covariance function

A covariance function (also known as a variogram or a semivariogram) is a mathematical function that describes the degree of similarity or dependence between values of a variable at different locations.

Exponential covariance function:

$$c(\boldsymbol{h}; \boldsymbol{\theta}) = \theta_1 \exp(-\|\boldsymbol{h}\|/\theta_2), \qquad \theta_1 > 0, \quad \theta_2 > 0.$$

, theta1 is called partial sill and theta2 is called range (distance at which covariance vanishes).

Covariance function needs to be a positive (semi-)definite function.

For covariance functions that only asymptotically reach zero, we often use the term practical range, which is defined as the distance at which the covariance function is at 5% of the practical sill (we cover 95% of the points that influence our point).

Variograms should not be estimated for lags bigger than half the diameter of the domain.

All lags h should be chosen such that at least thirty pairs should be included

Stationarity is an assumption made by the statistician and it cannot be proven or rejected with the data itself — at most we could show that we cannot reject the hypothesis

Learning goal: Define Kriging

If we wish to make predictions at a location not included in the observed locations, for the purpose of establishing a pollution map for example, all measures will have to be taken into account in computing the predicted value. Of course, their contributions will be weighted by the strength of their correlation with the location of interest. Kriging is a minimum mean squared error method of spatial prediction.
The process of kriging consists of the following steps:
Estimate the experimental variogram or the covariance function of the variable based on the sampled data.
Select an appropriate theoretical model for the covariance function based on the shape of the experimental variogram.
Solve the system of equations that relates the values of the variable at the sampled locations to the value of the variable at the unsampled location.
Estimate the uncertainty of the interpolated values by calculating the kriging variance.

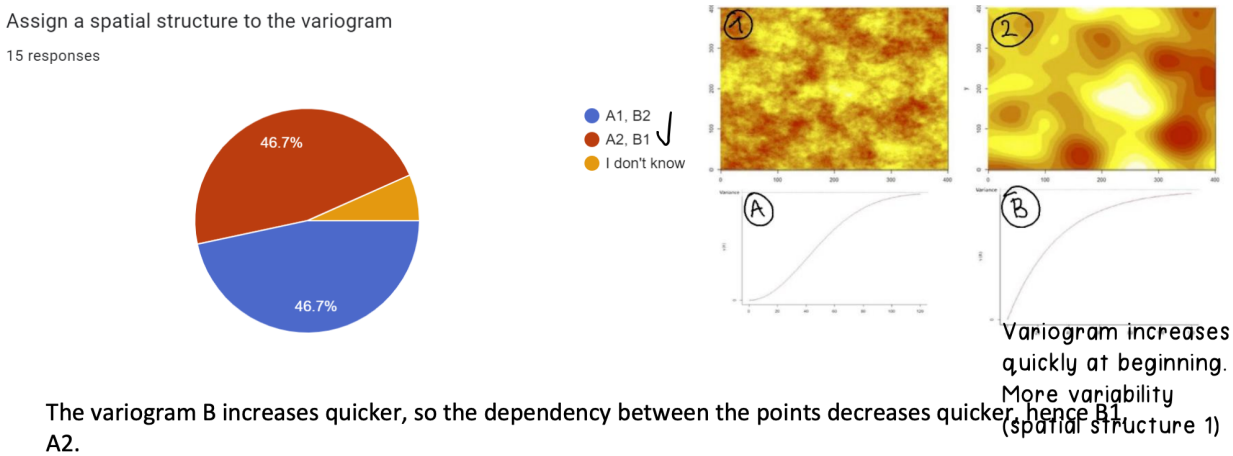
The result of kriging is an estimate of the value of the variable at the unsampled location.
Assumptions of kriging: spatial process is stationary.
Pure nugget model: a constant value for all distances greater than zero

A nugget allows the variogram to assume a non-zero value for two observations having a distance of zero.

Nugget effect: the presence of a nugget means that any two observations sampled arbitrarily closely will not necessarily have the same value. The nugget effect can be thought of as an extra variance component that is added to the spatial variance component. If we see changes (variability) between closely spaced data we should add nugget effect to the model.

Questions on Lecture 11:

a) It is possible that the covariance function is discontinuous -> Yes, because of nugget effect. The nugget effect can make the covariance function discontinuous at zero lag distance, as it is represented by a constant term that is added to the covariance function. This means that the covariance function will not be smooth and continuous at zero lag distance, but will jump abruptly from zero to the value of the nugget variance.



The variogram B increases quicker, so the dependency between the points decreases quicker, hence B1, A2.

b)

Learning goal: Explain the motivation for applying extreme value theory

Extreme value theory (EVT) is a branch of statistics that deals with the modeling and analysis of rare or extreme events. The motivation for applying EVT comes from the fact that many natural and man-made systems exhibit extreme behavior, such as floods, droughts, hurricanes, earthquakes, financial crises, and many others. These extreme events can have significant consequences, such as loss of life, damage to property, and disruption of social and economic activities. Therefore, understanding and predicting the likelihood of such events is of great importance for risk management and mitigation. It's not about finding optimal but rather values when looking at risk.

The statistical theory of extreme values (maximum, minimum or exceedances of a large threshold) is called extreme value theory (EVT). IN EVT we focus on extreme and rare events.

Learning goal: Explain the modeling the maxima approach

The approach models the distribution of the maxima of a sample of data, rather than the distribution of the entire sample. This allows for the estimation of the probability of extreme events that are more likely to occur than those estimated using the traditional method of modeling the entire sample. MTM can be applied to both continuous and discrete data and can be used to estimate the parameters of a variety of probability distributions, such as the Generalized Extreme Value distribution.

In practice we hardly ever have replicates of an entire series to estimate the maximum. To circumvent this lack, one subdivides an entire sequence into several shorter sub-sequences, called blocks. From each of these blocks, we take the maxima, then we move to the next time period(block) and look at distribution of maximas and estimate parameters of GEV.

Disadvantages of the block maxima? We could get some values and miss some values (lose them). We lose some high-value observations and those we got might not even be the most extreme ones. It has a severe reduction of data, since in many applications we reduce the dataset to annual maxima. Solution: take values above some threshold and define those as extreme enough. For modeling the minima we would lose even more data.

Learning goal: Describe the generalized extreme value distribution (GEV)

The GEV distribution is defined by three parameters: the scale parameter (sigma), the location parameter (mu), and the shape parameter (xi). The scale parameter determines the spread of the

distribution, the location parameter determines the location of the distribution, and the shape parameter determines the shape of the tail of the distribution.

Gumbel: $\xi = 0$ exponential tail

Fréchet: $\xi > 0$ so called „fat tail“

Weibull: $\xi < 0$ upper finite endpoint

Q: To which of the three distribution does the distribution of maximum of a uniform random variables converge?

Weibull because it has the upper bound (when we get close to the border it converges). It is not smoothly continuing to infinity. 19

Learning goal: Explain Modeling Peaks over Thresholds approach

we look at values about certain threshold and consider those as maximas.

The POT approach involves two steps:

1) Selecting a threshold value, above which events are considered extreme.
2) Modeling the distribution of the exceedances (events above the threshold) using a probability distribution such as the Generalized Pareto Distribution (GPD).

The GPD is a probability distribution that is often used to model the distribution of extreme events. It has two parameters, the scale parameter (sigma) and the shape parameter (xi), which determine the spread and the shape of the tail of the distribution respectively. The shape parameter also allows to distinguish between the three different types of tail behavior: light tail (shape = 0, corresponds to an exponential distribution), intermediate tail (shape < 0) and heavy tail (shape > 0).

Threshold Selection (not straightforward):

1) Mean residual life plot: expresses the mean of the excess over a threshold as a function of the threshold. Choose threshold before CI starts increasing.
2) Parameter estimates against threshold plot: select the lowest threshold that yields roughly the same parameter estimates as any higher threshold

GEV vs GPD: Which one is better?

- A priori, each method needs to set one arbitrary value: the time interval duration in the GEV method and the threshold in the GPD method
- GPD method requires the determination of only two parameters, compared with the three parameters needed in the GEV method.
- It is recommended using both methods simultaneously to check their mutual consistency

Learning goal: Explain Return Level Plots

Return level is the level that is expected to be exceeded by the process on average once in T-years. It is that extreme so that it exceeds only one time in t years.

Learning goal: Give some examples of application of the EVT

Finance: maximal daily lost. Hydrology: protection against flood, maximal flow expected once in 100 years. Meteorology: extreme winds, heavy precipitation events.

Questions on Lecture 12:

a) We draw n independent and identically distributed random variables $X\_1, \ldots, X\_n$ from a distribution $F(x)$. The distribution of the maximum M_n converges to a normal distribution (based on Central Limit Theorem) -> False, converges to a GEV.
b) We are interested in the distribution of the minimum temperatures on a research station in Antarctica. I have daily temperature data from the last 20 years. To find parameters of GEV distribution (using e.g. function fevd() in R), I can: Work directly with the measured temperatures and work with temperatures*(-1), to look at the distribution of the maxima.

Learning goal: Understand the concept of (machine) learning

Learning is seen as an automatic process, involving a loss function as well as a rule for updating the model. Machine learning refers to building models that machines learn and are then able to imitate human behavior.

Learning goal: Describe the concept of a neutral network (including input layer, hidden layers, output layer, neuron, activation function, weights, forward propagation, backward propagation, error function, hyperparameters of the network)

Neural networks are motivated by the desire to mimic single biological neurons and simple combinations thereof, called neural nets. A single neuron is modeled by a function that receives many inputs of varying strength with output the result of a transform function. The latter is often used in the context of a threshold, the neuron triggers as soon as a certain critical value is exceeded. Such an activation function outputs a binary response (off/on), depending on the strength of the input. Hence a neuron can be seen as a function f : R → 0, 1. The weights are the parameters that need to be learned. If several neurons are used sequentially, the output of one is used as input of the next. Usually, they are arranged in a set of neurons named layers.

Loss is a measure of how well the network's predictions match the true values. Goal: minimize it. A loss function measures the loss. The goal of training a neural network is to find the set of weights that minimize the loss function.

Activation function should be chosen in such way that we see influence of derivative and we can still adjust weights

The activation functions are necessary to prevent linearity. Without them, the data would pass through the nodes and layers of the network only going through linear functions (a*x+b).

First layers in a NNtrain simple structures and last ones train more complex structures.

Forward propagation is the first step in evaluating a neural network. It involves passing the input data through the network, starting at the input layer and working through each subsequent layer until reaching the output layer.

Backpropagation is the second step in evaluating a neural network. It involves starting at the end of the network, with the loss function, and then working backwards through the network to compute the gradient of the loss with respect to each weight. The gradient is then used to update the weights in a way that reduces the loss. This process is repeated for multiple epochs, with the weights being adjusted at each epoch based on the gradient, until the loss reaches a minimum and the model's predictions are as accurate as possible. In backpropagation, the number of iterations to perform is determined by the number of epochs and the batch size. These hyperparameters determine the performance of the model.

Backpropagation is a supervised learning algorithm because it uses labeled training data to adjust the weights and biases of a neural network.

Dropout: process of dropping random layers out.

Questions on Lecture 13:

a) For effective training of a neural network, the network should have at least 5-10 times as many weights as there are training samples -> False, we would have too many weights and too few data.
b) The backpropagation learning algorithm is based on the gradient- descent method -> True
c) Based on the input data (e.g. images of digits and their labels), a neural network can learn weights, the number of layers, the activation function - all the parameters required to have a trained neural network -> False, Neural Networks can learn weights but what belongs to the structure of the NN must be defined beforehands.

Neural networks are fitted using maximum likelihood -> False, we do not use likelihood at all, We just compute output and adjust weights according to what we want as result.

<u>Exam questions</u>

- What is an estimator? What is an estimate?

Consider a real-valued statistic $T_n = h(X_{1:n})$, based on a random sample $X_{1:n}$ from a distribution with probability mass or density function $f(x;\theta)$ where θ is an unknown scalar parameter. If the random variable T_n is computed to make inference about θ, then it is called an estimator. We may simply write T rather than T_n if the sample size n is not important. The particular value $t = h(x_{1:n})$ that an estimator takes for a realization $x_{1:n}$ of the random sample $X_{1:n}$ is called an estimate.

- TRUE/FALSE Questions

a) Correlation between two PC dimensions depends on their corresponding eigenvalues. FALSE. the principal component technique always provides a set of non-correlated PC dimensions, regardless of their eigenvalues.
b) Consider a time series AR(1). In this case, the autocorrelation function vanishes after 1 lag. FALSE. in an AR (1) we will see an exponential decrement of the ACF values, and only one spike at lag 1 and then 0 for their PACF.

c) Each hazard function is a density. FALSE. hazard functions do not have to integrate to 1.
d) Each kernel is a density. TRUE. a kernel has to be normed and non-negative among other properties.

e) We consider a logistic regression: glm(y ∼ x). In this model, if the coefficient for x was estimated to be 2, then increasing x by 1, would cause y to increase by a factor of 2. FALSE. if we are under a logistic regression, then an increase on one unit in x will have associated an increase by the factor of 2 to the logit of y (log of the odds of y).

f) We are clustering points A, B, C, and D. It is impossible, that single linkage clustering would result in clusters (A, B) and (C, D), and complete linkage clustering would result in clusters (A, C) and (B, D). TRUE. In the first step all observations are one single clusters, and there is no difference between different clustering approaches, therefore at least (A, B) or (C, D) should be on the complete linkage clustering.

g) Let T be a random variable representing survival time. Let the hazard function be h(t) = 3. Derive the density of the distribution of T. What is the name of this distribution? we know that f(t) = h(t) ∗ S(t) which is equal to h(t) ∗ exp(−H(t)). Since h(t) is constant = 3, then H(t) is 3 ∗ t + c, and since H(t) has to be 0 when t = 0 (integration over a point) then c = 0. Therefore f(t) = 3 ∗ exp(−3t) , which has the expression of an exponential distribution with λ = 3.

h) We are considering a dosage of a insecticide. We would like to create a model, which would allow us to predict the percentage of killed insects. A linear regression is a good model to predict the percentage of insects killed. FALSE. We should use a GLM approach, since it guarantees the predictions between [0,1].

i) There was research done on the variables influencing salaries. There were 100 participants of age 35-45 in the study, from which many socio-behavioral variables were collected. The study took 5 years, and data from each participant was collected five times (i.e. once per year). A linear regression was conducted to model the data. That was not a correct choice, as data coming from one individual might be strongly correlated. Because of that, the confidence intervals for the coefficients in the linear regression are probably larger, than they would be, when computed using the mixed model. FALSE (last statement). If we omit the correlation between observations, then the confidence intervals will be narrower than when considering the correlation.

j) We would like to apply PCA to reduce the number of dimensions in our data. When the eigenvalues of the covariance matrix of the data are roughly equal, we lose the least amount of information when using only a few first principal components to represent the data. FALSE. it is the opposite, we want to have just a few large eigenvalues in comparison to the others. This translates into a few PC's explaining a high percentage of the total variability.

k) Bootstrapping can be used only for normally distributed data. FALSE. We do not state a distribution assumption when using bootstrapping.

l) ML vs REML:
REML: unbiased estimates of variance. ML is biased for the estimation of variance components. What are the advantages of REML vs ML? Under what circumstances may REML be preferred over ML (or vice versa) when fitting a mixed effects model? for small sample sizes REML is preferred. However, likelihood ratio tests for REML require the same fixed effects specification in both models. So, to compare models with different fixed effects (a common scenario) with an LR test, ML must be used. To just compute model use REML but if we want to compare models between each other is better to use ML.

m) What is the pure nugget effect? Pure nugget effect: no spatial correlation, just variability. Variogram is constant, no matters how close observations are since there is no spatial correlation.

n) What is the fundamental assumption difference between a linear discriminant analysis and a quadratic discriminant analysis?
In LDA, the assumption is that the data in each class is normally distributed with the same covariance matrix. In QDA, the assumption is that the data in each class is normally distributed with different covariance matrices.

o) We are simulating a stationary time series AR(2). In this case, the ACF vanishes after 2 lags. TRUE -> An AR(2) model is a model where the current value of the series at time t is a linear combination of two previous values of the series and a white noise term.

p) What kind of statistics belong to within-sample validation and to out-of-sample validation?
Within-sample validation: training error, Out-of-sample validation: testing error

q) In beginning of clustering the centers can be chosen at random from the range of data points (between min and max of the data). There could be different solutions when a data point has the same distance to several centers, one option is, that it gets assigned to the first center from the list/vector of centers.

r) Empirical orthogonal functions (EOF) are a decomposition of a spatio-temporal dataset in therms of orthogonal functions. The calculation is similar to performing a principal component analysis (PCA). -> TRUE, they both involve in finding eigenvectors of the covariance matrix. The main difference between EOFs and PCA is that EOFs are specifically designed to analyze spatio-temporal datasets, while PCA is a more general method that can be applied to any type of data. EOF is used to identify patterns in a dataset of spatial and/or temporal fields, such as weather patterns or ocean currents. EOF is similar to PCA in that it also transforms a set of correlated variables into a set of uncorrelated variables (empirical orthogonal functions) that explain the most variance in the data.

s) Multiple Linear regression vs Factor Analysis: Multiple linear regression is used to predict the value of a dependent variable based on one or more independent variables, while factor analysis is used to identify patterns and structure in a set of variables. Linear regression assumes that the relationship between the variables is linear and that the errors are normally distributed and independent.

Factor analysis assumes that the observed variables are a linear combination of a smaller number of latent variables or factors.

t) Covariance Function with uncorrelated data has covariance around zero, and variogram too-> False, variogram does not have covariance around zero because it looks at level of variance between the points.

u) Lower AIC scores are better, and AIC penalizes models that use more parameter. Lower BIC also better -> True.