**Chapter 2: Overview of Statistical Learning**

2.1 Introduction to regression models

* The standard model for regression is , where is the ***reducible* error** (the error in prediction can be potentially reduced by choosing a different model) and is the ***irreducible error*** (it originates from e.g. measurement errors or inherent stochasticity and cannot be reduced by changing the model).
* A good will allow us to make predictions of the value of at new points of .
* The ideal function will give the expectation value of at :

This ideal function is called the ***regression function***. This ideal function is optimal is the optimal predictor of in the sense that it will minimize the ***sum of squared errors***.

* Typically, there will be very few, if any, data points that correspond to a certain value of (for example, there may be no values for the output data at exactly ). In that case the expectation value cannot be determined.
* In this case, is estimated based on the points in the ***neighbourhood*** of :

Where is some neighbourhood around . This is called ***Nearest Neighbour or Local averaging***.

2.2 Dimensionality and Structured Models

* Nearest neighbour averaging becomes more problematic as the number of dimensions increases: For a high number of dimensions, the width of the neighbourhood for nearest neighbour averaging needs to increase to capture the same number of data points.
* As a result, nearest neighbour averaging becomes less local for as the number of dimensions increases. This is called the ***Curse of Dimensionality***.
* The curse of dimensionality can be circumvented by using ***structural models*** that do not depend local properties and nearest neighbour averaging. An example of a structural model is a linear model for which each of the parameters can be found by fitting it to all datapoints. As a result, these linear models do not depend on any local properties or nearest neighbour averaging.
* Structural models that are less ***flexible*** (for example, linear vs. quadratic models) typically have lower ***interpretability***, although they may provide a better fit to the data.
* It is important to choose a model that is not too flexible to prevent ***overfitting*** of the data.

2.3 Model Selection and Bias-Variance Tradeoff

* To see how well a model performs, it can be fitted to a set of ***training data***:

However, this can be ***biased towards more overfit models***. Instead, the value should be evaluated on a set of ***test data***:

* The performance of models with different flexibilities can be assessed by plotting against model flexibility. The value of will typically have a minimum value for a certain flexibility of the model, while will keep decreasing as the model becomes more flexible.
* The choice for a certain flexibility of the model is subject to a ***Bias-Variance Tradeoff***: the variance in predictions of a model will increase as the model becomes more flexible (the model will become more sensitive to the peculiar characteristics of the training dataset). In return, the bias will decrease as the model becomes more flexible.

2.4 Classification

* In classification, the response variable is ***qualitative*** and the aim is to build a classifier that can assign a class label from the set of labels for future observations of .
* An ideal classifier is the ***Bayes Optimal Classifier*** that minimizes the probability of misclassification.
* If there are elements in the set of class labels then the ***Conditional Class Probabilities*** are:

The Bayes Optimal Classifier at is given by:

In other words, the Bayes Optimal Classifier assigns the observation to the ***most class with the highest probability***.

* In classification, nearest neighbour averaging can be used as in logistic regression (take the conditional probabilities in the neighbourhood of ). In that case, the curse of dimensionality still applies. However, the curse of dimensionality has less impact on than on .
* The performance of the classifier is typically measured using the ***Misclassification Rate***:

In words: ***the error is the average number of mistakes***. This error is the smallest for the Bayes Classifier.

* ***Support Vector Machines*** build structured models for **.**

**Chapter 3: Linear Regression**

3.1 Simple Linear Regression

* Linear regression is a simple approach to ***Supervised Learning*** (it is based on the assumption that the dependence of on is linear).
* Simplest case of linear regression is a model with a single predictor:

When the model coefficients have been estimated, the output can be predicted by:

* The best values for the parameters can be estimated by ***least squares***.
* The r***esidual*** (which is the discrepancy of the actual outcome with the predicted outcome) for the prediction of based on the value of is given by:

And the ***Residual Sum of Squares*** is:

* The unique line with values for and that minimizes the RSS is the ***least squares fit***.
* The accuracy of the estimates for and can be determined using the ***Standard Error***. The standard Error will ***increase when the variance () increases*** and the Standard Error will ***decrease when the spread of x-values around their mean increases***.
* The standard Errors can be used to define ***Confidence Intervals*** (For example, an interval that has a 95% probability to contain the true slope value).

3.2 Hypothesis testing and Confidence Intervals

* The standard error can also be used for performing ***Hypothesis Testing***. The most common Hypothesis test involves testing the Null Hypothesis:

***Null Hypothesis*** (): There is no relationship between and .

***Alternative Hypothesis*** (): There is some relationship between and .

* The Null Hypothesis is tested by computing a ***t-statistic:***

If the Null Hypothesis is true, this will ***have a t-distribution with degrees of freedom***.

* The ***p-value*** is the probability to get ***at least*** the obtained value for .
* ***Hypothesis Testing and using Confidence Tests are equivalent***: if one of the two shows that there is a relation between and then so will the other and vice versa. However, the Confidence Test also tells you how big the effect is.
* The accuracy of the model can be determined by computing the ***Residual Squared Error***:

And by computing the  ***value*** (the describes the fraction of the variance in the data that is explained by the model). , where is the correlation between and .

3.3 Multiple Linear Regression

* Multiple Linear Regression is regression with more than one predictor:
* How to interpret the different predictors now:

-***If the predictors are not correlated data***, each predictor can be interpreted separately.

-***Usually, the different predictors are correlated,***and this makes their interpretation much more complex. In addition, ***the variance on all coefficients can increase*** when the predictors are correlated. This is because the relation of a predictor to a certain variable is less strong in that case (the predictors of different variables can be swapped without significantly changing the result). In this case ***Claims od Causality*** should be avoided.

* The ***Least Squares Estimates*** can be determined in much the same way as for a simple linear model ().
* Instead of a line, the function will now be a ***Hyperplane***. Therefore, in this case the Least Squares Estimates minimize the distance between each data point and its closest point on this hyperplane.
* For multiple Linear Regression, the standard error, t-statistic and p-value will tell you whether one variable affects the outcome ***in the presence of the other variables***. Although a variable may not have a significant affect in the presence of other variables, it could have a significant effect on its own.
* ***The correlation between the variables*** will tell you something about whether the information on a certain variable significantly improves the prediction of the model given that you have information on the other variables. In other words, data on a variable may have become redundant in the presence of another variable when they are strongly correlated.

3.4 Some Important Questions

* Which of the predictors is useful for the prediction of the response? And how accurate is the prediction?
* To determine if any of the predictors is useful to predict the response, the ***-statistic*** is used:

Here, the  ***is the Total Sum of Squares***, which is the sum of the residuals when no model (the predictor of the data is the mean) is used. is the Residual Sum of Squares, which are the residuals that remain when using our predictive model. is the number of parameters in the model (in total parameters are fitted to the data, the accounts for the intercept) and is the sample size.

* In the case that there is no effect of the predictors (the null hypothesis), the ***-statistic will follow an -distribution with degrees of freedom ()*.**
* The values for the -distribution can be looked up in tables. ***A large -statistic indicates that there is a strong effect of the predictors on the response prediction.***
* To decide whether a variable is important for the prediction of a linear regression model, ***All subsets or best subsets regression*** can be used.
* In all subsets regression, the least squares fits for all possible subsets of variables are made and the best subset is chosen based on a balance between training error and model size.
* The problem with all subsets regression is that the number of models that need to be evaluated grows exponentially with the number of parameters in the model.
* Two commonly used approaches for all subsets regression:

***Forward Selection***

Starts with the ***null model*** (a model with only the intercept and no predictors). Then each of the variables is added one at a time to the null model and simple linear regressions are performed. ***The variable that results in the lowest is added to the null model***. This is repeated for two-variable models, three variable models etc., until some stopping rule is reached (such as ***all the remaining variables have a p-value above some threshold***).

***Backward Selection***

Start with a model that contains ***all the variables***. Then remove the variable that has the largest p-value (the one that is the least statistically significant) from the model. Repeat this process for a model with variables, variables etc., until some stopping rule is reached (such as ***all the remaining variables have a p-value below some threshold***).

* How to deal with ***Qualitative Variables*** instead of quantitative variables: incorporate a ***dummy variable*** in the model (there will always be one fewer dummy variable than the number of levels in the set of qualitative variables).

Such that the equation of the model becomes:

In this case ­will tell you the effect of belonging to class A versus the ***baseline***, which in this case corresponds to belonging to class B. The value of the intercept () represents the average value for for class B. The value of will represent the strength of belonging to class A over belonging to class B and the p-value of will tell you if there is any significant difference in the expected value of depending on whether something belongs to class A vs class B.

3.5 Extensions of the Linear Model

* Make models that include ***Interactions*** and ***Nonlinearity*** instead of relying on the assumption that the effect of the different parameters are additive.
* To account for interactions, you include ***Product terms*** between the different parameters:

Instead of an independent coefficient for the variable , the coefficient for is now modified to be ***dependent on the second variable*** . Now the value and p-value of the coefficient  ***gives you information about the strength and significance of the interaction***.

* Sometimes the interaction terms between variables are very significant (low p-value), but their main effects do not have strong significance (that is, there is a significant contribution of the interaction between and on the response variable , but the contribution of either or alone does not). In that case, it is custom to follow the ***Hierarchy Principle***:

*If we include an interaction in a model we should also include the main effects, even if the p-values associated with their coefficients are not significant.*

The reason that this is done is because it is usually difficult to interpret the model when you do not include the main effects. Specifically, ***the*** ***interaction terms also contain main effects if there are no main effect terms added to the model***.

* Models can be modified in the following way to include ***interactions between a*** ***quantitative and a qualitative variable*** in the following way:

In this case where there is ***no interaction*** between the classes and the variables, ***only the intercept*** will be different depending on whether the datapoint belongs to class A or class B.

In this case where ***there is an interaction*** between the classes and the variables, ***both the intercept and the slope*** can be different depending on whether the datapoint belongs to class A or class B.

* The same approach can be used to accommodate ***polynomials*** in the model:

In this case the model is still called a ***linear model*** (and so we still do linear regression), because the model is still linear in the coefficients.

**Chapter 4: Classification**

4.1 Introduction to Classification Problems

* Qualitative variables have values from an ***unordered set***.
* The task of the classification function is to take the values from and predict their qualitative values such that it assigns them to a specific class in the set.
* Typically, the major interest will be to ***estimate the*** ***probabilities that belongs to each category***.
* Can linear regression be used for a classification task?
* In the case of a binary outcome, linear regression can do quite well when implemented as follows:

You can then simply perform a linear regression of on and classify as **yes** when  .

* In addition, in this example of a binary outcome, in the population we have , which would be a further justification for the use of linear regression.
* In this case linear regression is equivalent to ***Linear Discriminant Analysis***.
* However, linear regression can produce probabilities that are ***Lower than zero and larger than one***. For this reason, ***Logistic Regression*** is more appropriate.
* When the variable can be classified into ***more than two categories*** then assigning values to the categories does not work, as it suggests an ***ordering*** between them. For example, when modelling the response as:

This implies that the difference between A and B is the same as between B and C, but not the same as the difference between A and C.

* For this case of a variable that can be classified into more than two categories, ***Linear Regression is not appropriate***.
* Instead, ***Multiclass Logistic Regression*** or ***Discriminant Analysis*** are more appropriate.

4.2 Logistic Regression

* ***Logistic Regression*** equation takes the following form:

This equation will always take on values ***between 0 and 1***, as is required for a probability.

* The equation can be rewritten to:

This shows that the equation for logistic regression still entails a ***linear model, but with the probabilities modelled on a non-linear scale***.

* To estimate the parameters when the different observations are independent of each other, we use ***maximum likelihood***:

This likelihood gives the probability of having the number of zeros and ones that we see in the data. For example, if the series of zeros and ones look like 1101001, then the above equation would calculate , where is the probability of having a one and is the probability of having a zero. The parameters and are chosen such that ***the likelihood of the observed data is maximized***.

* From fitting a logistic regression model, you can get the ***Z-statistic*** of the values that were found for the parameters. The Z-statistic shows a sort of standardized value for the parameters. For example, the value of the slope of the model will depend on the units that are used, and the Z-statistic corrects for that. In that way it gives a better view of how strong the relation is between the variables and the response.

4.3 Multivariate Logistic Regression

* Logistic regression with multiple variables takes the form:

Or, to have it in a form that is linear with respect to the variables:

* This can be used in a similar way as for linear regression to estimate the variables. Remember that some of these variables can be strongly correlated, which can make it more difficult to interpret their effect on the response (it can change depending on which other variables are present).

4.4 Logistic Regression - Case-Control Sampling and Multiclass

* ***Case Control Sampling*** is a popular tool in epidemiology because it allows to correct for a difference in the sampled cases compared to the actual number of cases in a population (For example, the prevalence of a disease may be 35% in the sampled population, while it is 5% in the actual population).
* With logistic regression and case control, you can still accurately estimate the regression parameters of interest (these are the slopes of the different variables , excluding the intercept ).
* The estimated value of will initially be incorrect, but can be corrected after fitting the logistic model using this transformation:

Where is the prevalence of cases in the actual population and is the prevalence of cases in the sampled population.

* Thus, with case control sampling you can estimate the parameters ***without requiring a large sample population*** (which is particularly a problem if the occurrence of the event is rare) and ***without requiring a large amount of time*** to determine how often the event occurs. Instead, you only need to take a small sample population of the controls (cases where the event did not happen) and compare it to a sample population where the event did happen.
* An example is the click through rate for ads. If you take a random subset of subjects that have been exposed to ads, you will have a large amount of 0’s (people that do not click) and only a few 1’s (people that do click). ***In case control, you do not use all this data to fit your model, but only take a sample of the controls, which will still give accurate estimates for the main parameters of the model when you make the fit.***
* When using case control, ***there is a trade-off between control:case ratio and the variance in the estimate of the parameters.*** When the ratio of control:case ratio reaches a value of around 5:1 to 6:1, the variance starts to flatten out, so at this threshold increasing the number of controls relative to the number of cases no longer significantly increases the accuracy if the estimated parameters.
* Logistic regression can be easily ***generalized to a situation where there are more than two classes***. One version used in R has the following symmetric form:

Where is one of the classes in the total set of classes. Here, you sum over the different linear functions for each class and you weigh the exponential term for class against this sum. This multiclass logistic regression is also known as ***Multinomial Regression***.

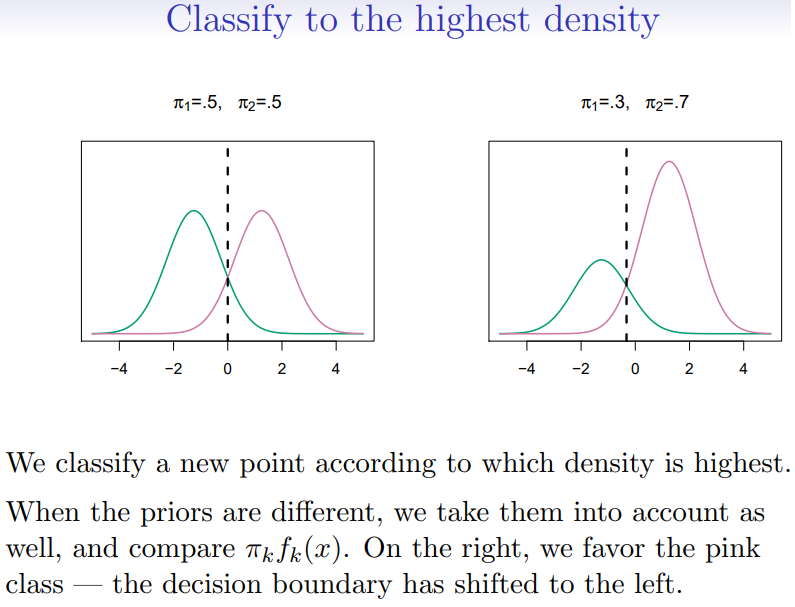
4.5 Discriminant Analysis

* In ***Discriminant Analysis*** the approach is to model the distribution of in each of the classes separately and then use ***Bayes theorem*** to find
* When each class is modelled using ***Normal/Gaussian distributions***, this leads to ***Linear or Quadratic Discriminant Analysis***.
* ***Bayes Theorem*** states:

For discriminant analysis, this is written as:

Here,is the (***Probability) Density for in class k***. is the ***Marginal or Prior probability for class k.***

* In discriminant analysis, points are basically classified ***according to which density is highest***. The ***decision boundary*** is the point at which the densities for the different classes intersect.

******

* Discriminant analysis provides an ***advantage over Logistic Regression when the classes are well-separated***, in which case the parameter estimates from logistic regression can be unstable (coefficients can go to infinity when classes are perfectly separated). Linear discriminant analysis does not suffer from this problem.
* The linear discriminant model is also more stable than logistic regression when ***the sample size is small and the distribution of the predictors is approximately normal in each of the classes***.
* When there are ***more than 2 classes***, linear discriminant analysis gives a nice low-dimension view of the data.

4.6 Gaussian Discriminant Analysis – One Variable

* The Gaussian probability density function is:

Where is the distribution mean and is the variance (in class ). The part that ***makes the probability dependent on is in the exponential***.

* Whether the ***variance is the same in all classes*** (that is, , where is the variance in each class) ***will determine if discriminant analysis gives linear or quadratic functions***.
* Inserting the equation for a Gaussian function (assuming that ) into ***Bayes theorem*** gives:
* To classify an observation to a class, you initially don´t need to evaluate the probabilities, ***you simply need to see which of them is the largest***. By taking the logs and discarding the terms that do not depend on , you can see that this is equivalent to assigning to the class with the largest ***Discriminant Score***:

Where is a linear function of .

* To get the parameters of the Gaussian distribution (such as and ) from real data, you will need to estimate them from your dataset. These estimates are calculated as follows:

When the variance in all classes is the same (), you can use:

Here, is the number of observations of class and is the total number of observations.

4.7 Gaussian Discriminant Analysis-Many variables

* When there is a ***correlation*** between the different variables of the model, the Gaussian probability density function ***will become stretched*** as compared to a model where the variables are uncorrelated (in that case, the Gaussian probability density function looks like a Bell function).
* To extend Linear discriminant analysis from one variable to many variables, you need to use the ***Covariance Matrix*** in the formula for that was set-up for the single variable linear discriminant analysis.
* If you would know the true probability densities for the different classes, you would be able to determine the ***Bayes Decision Boundaries***. These are the decision boundaries that would give the ***fewest misclassification errors, among all possible classifiers***.
* Linear discriminant analysis ***classifies datapoints to the closest centroid***, which means that if there are classes, linear discriminant analysis can be viewed in exactly a dimensional plot.
* If the number of variables becomes ***very large*** (if you have variables, the covariance matrix will have size by ), you need to make other modifications to use discriminant analysis.
* If you have estimates for , then these estimates can be turned into estimates for the ***Class Probabilities***:

So classifying to the largest is the same as classifying to the class for which is largest.

* This also shows that linear discriminant analysis does not only give us the class to which an observation was classified, it also gives us ***the probability*** ***that observation belongs to class*** .
* The misclassification rate can be assessed using a ***Confusion Matrix***.
* The misclassification rate should be compared to the ***Null Rate*** to see how well the classification actually is. The null rate corresponds to the misclassification rate that is obtained when you ***always classify to the prior*** (that is, you classify to the largest class).
* The ***ROC Curve*** can be used to simultaneously plot both the ***False positive and the True positive rates***.
* The ***Area Under Curve (AUC)*** of the ROC curve to summarize how well the classification is performed (higher AUC is better). It tells you how close you are to a true positive rate of 1 and a false positive rate of 0.

4.8 Quadratic Discriminant Analysis and Naïve Bayes

* When the densities are Gaussian densities with ***the same covariance matrix*** , then the quadratic terms ***cancel*** and this results in ***Linear Discriminant Analysis***.
* When the densities are Gaussian densities with ***different covariance matrices*** for each class, then the quadratic terms ***do not cancel*** and you get ***Quadratic Discriminant Analysis***.
* So far, the equation for discriminant analysis:

Has been applied for Gaussian functions, but it is general enough to also apply to different estimates of densities.

* If you have a large number of features (which will result in that you have to estimate a very large covariance matrix ), then one simplification that can be done is to ***assume that the different variables are conditionally independent in each of the classes***, which results in that in each class the density becomes the multiplicative sum of the densities of the different variables:

This will result in the covariance matrices becoming diagonal and you will only need to estimate parameters instead of parameters. This is called the ***Naïve Bayes Classifier***. Notice that in this case the ***covariance matrices can still be different for each class, but the assumption is that they are diagonal***.

* The naïve Bayes classifier is likely to result in a flattened and biased estimates for the probabilities, but since for classification we only need to know the largest probability to classify an observation (which allows you to tolerate some bias), the naïve Bayes is still useful.
* Naïve Bayes can be used with ***mixed features*** (both quantitative and qualitative features). For the quantitative features you can then use Gaussian functions, while for the qualitative features you can use histograms or mass density functions.
* Quadratic Discriminant Analysis is most attractive ***when the number of features/variables are small***, otherwise you will have to estimate large covariance matrices and things can break down. When the number of ***features/variables does become very large***, naïve Bayes becomes attractive (even over Linear Discriminant Analysis).
* For a classification problem with two classes, ***Linear Discriminant Analysis and Logistic Regression take on the same form***:

But the difference is in ***how the parameters are estimated***:

* Logistic regression uses the ***conditional likelihood*** , which is know as ***Discriminative Learning***.
* Linear Discriminant Analysis uses the ***Full Likelihood*** based on *,* which is known as***Generative Learning***.
* Logistic regression can also fit quadratic boundaries like Quadratic Discriminant Analysis by explicitly adding quadratic terms in the model.

**Chapter 5: Resampling Methods**

5.1 Cross-Validation

* Resampling is used when you want to you want to test the predictions of a model that has been built based on data, but you do not have a new or large test dataset to test it on.
* This section will discuss two resampling methods: ***Cross-validation and Bootstrap****.*
* The ***training error*** is the error that the model has on the training data set. This error will decrease as you increase the fit on the data (also when you overfit).
* The ***test error*** is the error the model has on a set of data it has never seen before. This error will likely increase as you overfit your model on the training data.
* Usually, the ***training error can dramatically underestimate the test error***.
* Some methods use a ***mathematical adjustment*** to the training error to estimate the test error. These methods include ***CP statistic, AIC and BIC***.
* Other methods estimate the test error by ***holding out a subset of data from the training set*** and then applying the model to that subset of data that was excluded from the training set. These methods are called ***validation or cross-validation***.
* For the ***validation*** set approach:
* The available set of data is randomly divided into two parts, a ***training*** set and a ***validation or hold-out set.***
* The model is first fit on the training set and then used to predict the responses for the observations in the validation set.
* The validation set error gives an estimate of the training error of the model, which will be the Mean Squared Error for a quantitative model and the misclassification rate for a qualitative model.
* The validation set approach can ***lead to a lot of variability*** in the test error when the process is repeated for different splitting’s of the data, which is a consequence of splitting the data randomly in two (the error can change depending on how the data was split). Often, the model flexibility (number of parameters) that gives the lowest training error for a given split of the data is more or less preserved, while the actual level of the test error can vary significantly for different splits of the data.
* Another drawback of the validation set approach is that you ***lose a lot of training data*** (you lose half of the training data that goes to estimating the test error).
* It will also tend to ***overestimate the test error***, since the test error is for a model trained on a training dataset of size will be lower than on a test dataset of size .

5.2 K-fold Cross-Validation

* K-fold Cross-validation is ***a very important technique that is widely used.***
* K-fold cross-validation is applying the ***Cross-Validation method K-times***, each time using a different part of the data as the training and as validation set.
* For K-fold Cross validation, the dataset is separated into ***K roughly equal sized parts***. Typically, the best choices for the number of folds is about 5 or 10.
* For example, if you use , you will have 4 training sets and 1 validation data set. You then fit the model on the cumulative of the 4 training sets (you cluster the 4 training sets together) and then use the validation dataset to determine the test error of the fitted model. This procedure is repeated while rotating which of the 5 blocks of data will be the validation data set. The ***Cross-Validation error*** is the error from all 5 parts of the validation datasets added together.
* If you let the parts of the data be , where denotes the indices of the observations in part . Part contains observations, which will equal if the number of observations is a multiple of . Then, mathematically the cross-validation error can be expressed as follows:

Where is the cross validation error and is the Mean Squared Error of using part of the data as the validation dataset. This is for ***Quantitative Datasets***. For ***Qualitative Datasets***, this translated into:

Where is the misclassification rate. Other than that, the procedure for K-fold Cross-Validation is ***completely equivalent when using quantitative or qualitative datasets.***

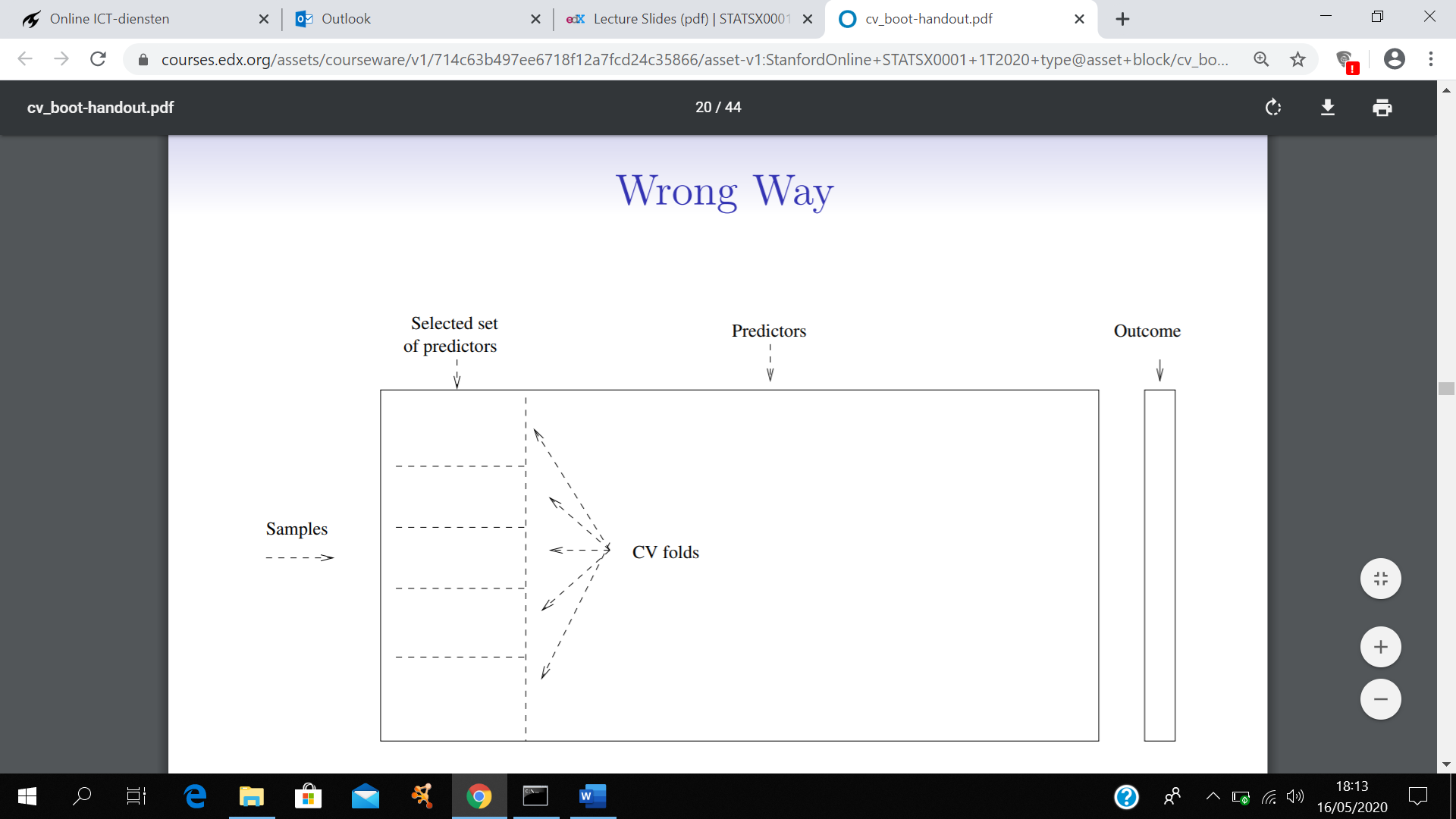
* A special case is ***Leave One Out Cross-Validation***, which is when the number of folds is equal to the number of observations , such that ***each single observation*** will take turns to be the validation data set.
* A nice thing about leave one out cross-validation is that under some circumstances, you do not need to make a new fit of the data for each fold , but ***only need to fit the model once*** using the entire dataset as training set.
* A downside of leave one out cross validation is that the ***estimates from each fold are highly correlated*** (all the different training sets look very much like each other) and as a result their average can have a ***high variance (but low bias)***. This is why is most often a better choice for the number of folds. Choosing is also making a bias-variance trade-off and these values of usually give a good variance vs. bias ratio.
* In other words: the ***bias is minimized for*** as is done in leave one out cross validation, but the variance in the estimates is high. Using is usually a good compromise on the bias-variance trade-off.
* For the success of cross validation, it is crucial that ***there is no overlap between the folds used for training and the fold used for validation.***

5.3 Cross-Validation: the Wrong and the Right Way

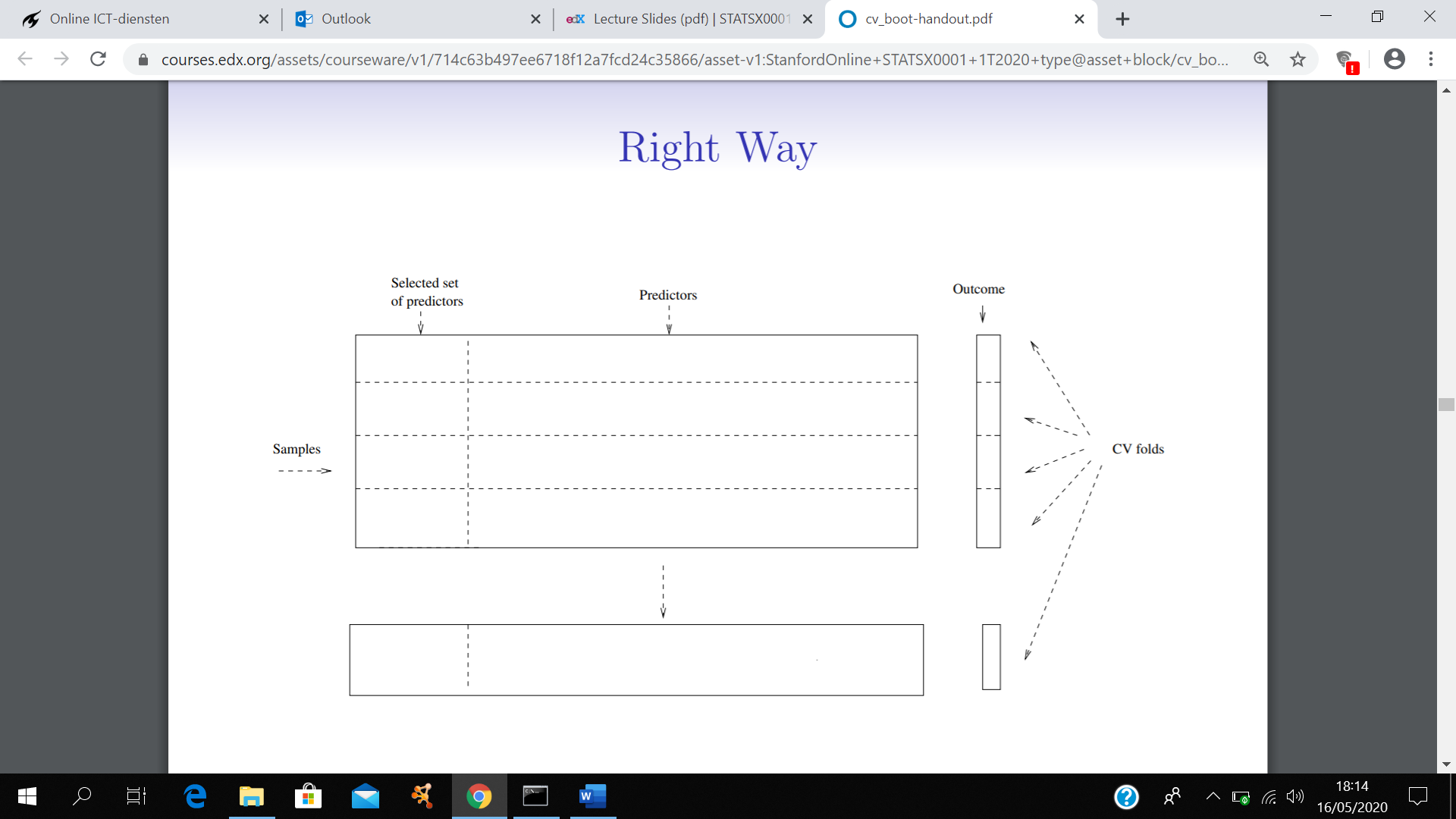
* When you want to apply cross validation to a dataset where you have two steps:

1. Filter the number of predictors (since you may have many) to the top predictors that have the best correlation with the class labels
2. Apply a classifier such as logistic regression, to only these predictors.

In this case it is important to ***Involve both steps in the cross-validation process and not just step 2!***



* This is because, in step 1 the procedure has already seen the labels of the training data, which is ***a form of training that must be included in the cross-validation***. Doing it this way will cause the classification error to be ***unrealistically low*** and can severely overestimate the predictive power of the model.
* The right way to do cross validation in this case is to ***create the folds before you filter or fit the data***. After making the folds, you can filter the number of predictors you have to the best correlated ones in the training dataset. The best correlated predictors can be different for each fold of the training dataset, which is a factor of variability that must be taken into account during cross validation.



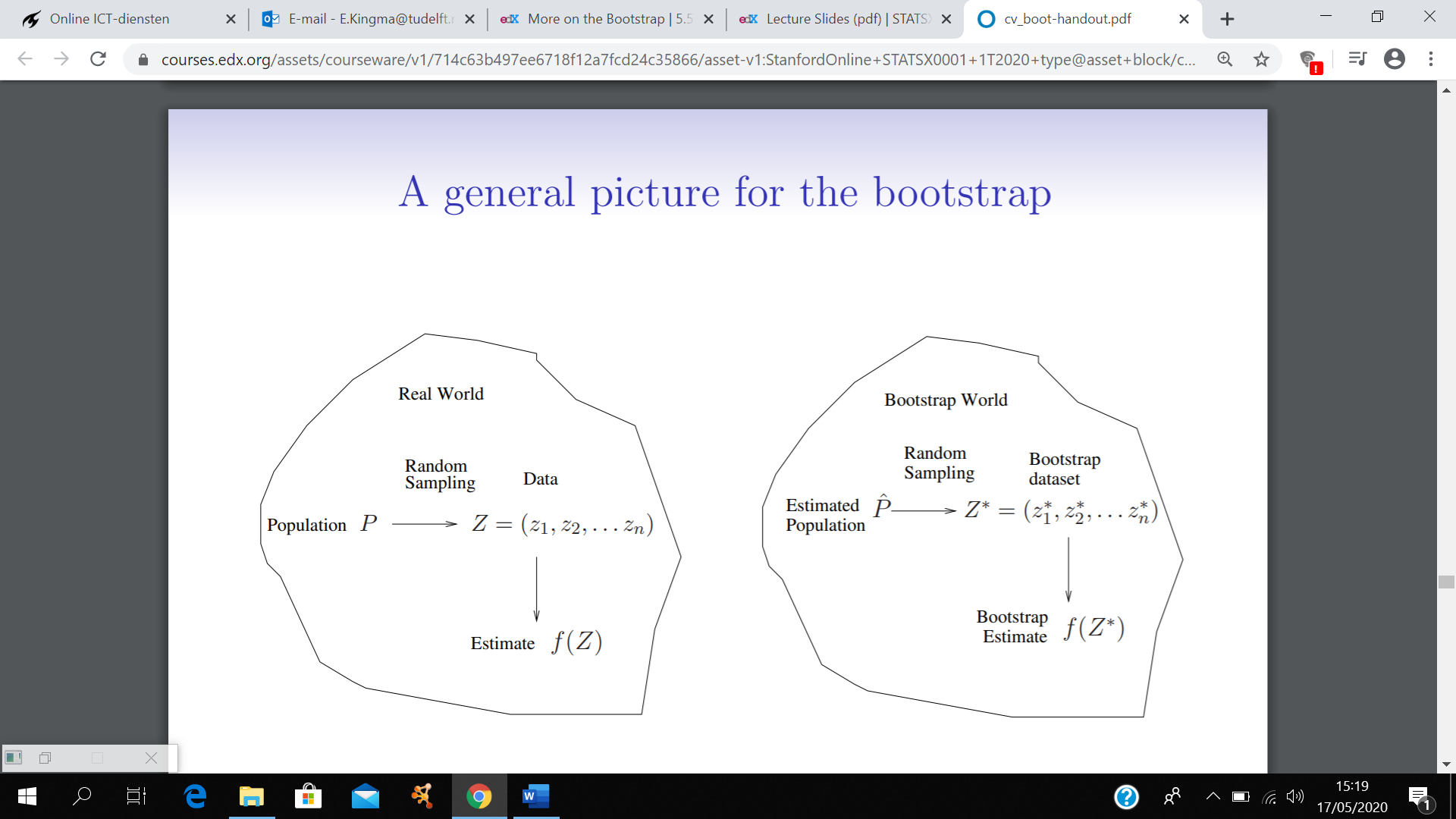
5.4 The Bootstrap

* The ***Bootstrap*** is a tool that can be used to ***quantify the uncertainty*** associated with a given estimator or statistical learning method.
* The Bootstrap can be used to ***estimate population quantities*** when you cannot sample from the population since you do not know what the population looks like (this is usually the case in real world problems).
* Instead, you will typically only have access to a limited number of data (observations or realisations that came from that population), but you do not know what population they came from.
* Bootstrapping allows you to ***mimic the process of obtaining new data sets*** from the population ***without generating additional samples***.
* Instead of obtaining independent datasets from the population, you obtain distinct datasets by repeatedly ***sampling from the original dataset with replacement***.
* Each of these distinct datasets that are obtained by sampling with replacement from the original dataset are ***the same size as the original dataset***.
* These bootstrapped datasets can then be used to estimate population quantities, such as variances and covariances.
* The standard error of the bootstrapped estimates of these quantities is given by:

Where is the number of times the boostrap is repeated (meaning that you have bootstrapped datasets), is the estimate of parameter obtained from the -ed bootstrapped dataset and   is the average of the parameter estimates  obtained from the different bootstrapped datasets.

5.5 More on the Bootstrap

* In the bootstrap you use the ***Empirical Distribution Function*** instead of the actual population and use that to determine the uncertainty in an estimated variable.



* The population is ***the training sample itself*** in which there is a probability of (with the sample size) for each of the training points to become an observation.
* A problem for the bootstrap when using it for predictive models is that the samples in the training data set are usually ***correlated to each other*** (which is something that also gives the model predictive power), while when sampling from the population for bootstrapping you assume that ***they are independent of each other***.
* So to use bootstrapping, you must figure out which part of the data are independent of each other.
* In time series data where the different observations are not independent, people use the ***Block Bootstrap***
* In the block bootstrap, the data is divided up into blocks and it is assumed that the ***data between different blocks are independent***. You then sample from these data blocks with replacement as if they were independent data points. This results in a set of new timeseries where the different data blocks are shuffled (and each block may appear more than once in the new series).
* The ***block size*** is determined by what is a reasonable assumption for the datapoints are uncorrelated between the different blocks. For example, it could be a reasonable assumption that after a lag time of three datapoints, the new datapoints no longer depend on what happened in the past.
* The main use of the bootstrap is to estimate the error of an estimate (such as ). Another use is to determine the ***confidence intervals of an estimate***.
* When the confidence interval is obtained from bootstrapping, it is called the ***Bootstrap Percentile*** confidence interval.
* To estimate ***Prediction Error*** from the bootstrap, one possibility would be to use each bootstrapped dataset as a training set and then use the original training set as the validation set. The problem with this is that ***the datapoints in the bootstrapped dataset will also appear in the validation dataset***. This will cause the prediction error to be ***biased downward*** when this method would be used.
* Doing it the other way around (using the training dataset to train the model and the bootstrapped datasets as validation sets) ***will create an even worse bias***. This is probably because in that case ***all*** the datapoints in the validation sets have already appeared in the training sets.
* In fact, using bootstrapping to estimate the prediction error will only lead to something that is about as good as ***cross-validation***,but much ***more complicated***. So in the end, Cross-validation would be the recommended method to use.

**Chapter 6: Linear Model Selection and Regularization**

6.1 Introduction and Best Subset Selection

* Linear models have advantages in terms of ***interpretability*** and often show ***goof predictive performance***.
* Previously, linear models were made using ordinary ***Least Squares Fitting***, but there are some ***possible alternative procedures*** that can be used to improve the simple linear model.
* These improvements to the linear model refer to improvements in both ***interpretability*** (by reducing the umber of coefficients and variables in the model) and its ***ability to make accurate predictions***.
* Alternatives to least squares may be considered when you want to
* to ***improve accuracy***, particularly when the number of features is larger than the number of samples (). In that case there is no well-defined solution to fitting the model using least squares.
* to ***increase interpretability*** by reducing the number of (irrelevant) features in the model. This relates to using methods for ***feature selection***.
* The three classes of techniques that can be used to find the most useful features in a model are ***subset selection, shrinkage*** and ***dimension reduction***. Although these are presented in the context of linear regression here, they can also be applied to other models, such as ***logistic regression***.
* For ***Best Subset Selection*** you start with the ***null model*** which contains no predictors (only the ***intercept***)and sequentially add the best predictors:

For

1. Fit all models that contain exactly predictors
2. From the number of models, pick the model that gives the best results and call it . In this case, the best model is defined as the model that results in the lowest Residual Sum of Squares (= highest ).

Then select the single best model from the set using cross-validated prediction error, (AIC), BIC or adjusted . ***It is very important to select the best model based on the test error and not the training error!*** The training error will go down in any case as you increase the number of predictors, so you have to use the test error to determine which predictors are useful.

6.2 Stepwise Selection

* For models other than linear regression, the role that the RSS plays in least squares regression is played by ***the deviance*** for a broader class of models (such as logistic models).
* The deviance equals negative two times the log-likelihood.
* In the case of least squares the RSS and the deviance are equivalent. The deviance is simply a generalization of the RSS.
* Best subset selection does not work well when you have a model with a large number of predictors (usually the limit is set to 30-40 predictors):
* The ***computational cost*** for doing best subset selection with many predictors becomes very large (if you have predictors you will have to fit models to the data.
* When is large best subset selection also suffers from ***statistical problems***. Because of the large number of models that are being evaluated there is a large chance of selecting a model that looks good on the training data but does not have a good predictive power. The large search space of models can lead to ***overfitting and high variance of the coefficient estimates***.
* For these reasons, ***stepwise selection*** methods, which evaluate a more restricted number of models, are good alternatives for best subset selection when the number of predictors is large.
* In stepwise selection you will need to ***evaluate models instead of the***  models that you need to evaluate in best subset selection.
* The benefit of stepwise selection over best subset selection ***seems counter-intuitive***, but it can actually be better to look only at a subset of all possible models rather than all models when is large. It is not always best to do a full search because you ***pay a price in variance***.
* Stepwise selection can be used as ***Forward Stepwise Selection*** or ***Backwards Stepwise Selection***.
* Forward stepwise selection ***starts with a model containing no predictors*** (the null model )and then adds predictors one at a time until all predictors are in the model. At each step, the variable that gives the greatest ***additional*** improvement to the fit is added to the model. The ***big difference with best subsets selection*** is that you do not evaluate all models with predictors at each step, but only models. So the search is much more restricted than for best subset selection.
* Forward stepwise selection is implemented as follows:

Start with the null model which contains no predictors (only the intercept).

For :

1. Consider all models that augment the predictor in with one additional predictor.
2. Choose the best among these models and call it . Here, best is defined as having the smallest RSS or equivalently the largest .

Select the single best model from the models using cross-validated prediction error, (AIC), BIC or adjusted .

* The resulting models are ***nested***, which is not the case when you do best subset selection.
* For a given model size forward stepwise selection will have an ***RSS that is above the RSS of best subset selection***, since you do not evaluate all possible models. This is related to that forward stepwise selection ***is not guaranteed*** to find the best possible model out of all models that are evaluated in best subset selection.
* However, although best subset selection might find the model that ***performs the best on the training*** data does not mean that that model will also be better in the ***context of test data***.
* The discrepancy between best subset selection and forward stepwise selection arises when the ***variables are correlated***. When the variables are uncorrelated, the variables selected by the two methods are exactly the same.

6.3 Backward Stepwise Selection

* In ***Backward Stepwise Selection*** you start with a model with predictors and ***remove the predictors*** one at a time until you end up with the null model that contains only the intercept.
* As for forward stepwise selection, backward stepwise selection does not guarantee that the ***best model containing a subset of the predictors will be found***.
* Backward stepwise selection is implemented as follows:

Start with the ***full model*** which contains all predictors.

For :

1. Consider all models that contain all but one of the predictors in , for a total of predictors.
2. Select the best among these models and call it . Here, best is defined as having the smallest RSS or equivalently the largest .

Select the single best model from the models using cross-validated prediction error, (AIC), BIC or adjusted .

* In total, both backward stepwise selection and forward stepwise selection require the evaluation of models (as opposed to models that need to be evaluated for best subset selection.
* Backward stepwise selection requires that ***the number of samples is larger than the number of variables*** (a least squares model can only be fitted if the number of samples is larger than the number of variables ). In contrast, forward stepwise selection can be used ***even when*** , making it the only method that can be used when is very large.

6.4 Estimating Test Error

* One approach for ***indirectly*** ***estimating the test error*** of the models is to determine the training error and then adjusting it.
* The idea behind this adjustment is to ***account for the bias due to overfitting*** of the training error to get an estimate of the test error that is based on the training error.
* An alternative approach is to ***directly estimate the test error*** using methods discussed in chapter 5 (such as the validation set or cross-validation approach).
* , AIC, BIC (Bayesian Information Criterion) and adjusted ***adjust the training error*** to give an estimate of the test error. To select a model that has good predictive power, you want ***, AIC and BIC to be as small as possible*** and the ***adjusted to be as large as possible***.
* , which is ***Mallow’s***  is defined as:

Where is the total number of parameters used (Which includes the intercept) and is an estimate of the variance in the error associated with each response measurement. is the same for all models being compared, and it is calculated using the model containing all predictors by taking the mean square residual to calculate . As a result, ***the calculation of gives problems***when the number of predictors is larger than the number of samples , because when the error will become and there will be no adjustment of the RSS to correct for the number of parameters when calculating . ***The calculation of is restricted to cases where*** .

* The AIC criterion, which is ***Akaike Information Criterion***, is defined for a large class of models fit by maximum likelihood.

Where is the maximized value of the likelihood function for the estimated model. AIC is a quantity that ***can be calculated for many different types of models*** (For example, not just linear models, but also logistic models). For the case of a linear model:. As a result, AIC and Mallow’s are proportional to each other when you have a linear model and choosing your model based on either of them is equivalent. For other types of models, AIC is a good approach.

* The BIC criterion, which is the ***Bayesian Information Criterion***, is defined as follows:

Comparing this to the equation for AIC, you see that the only difference is that the term in the AIC has been replaced by in the BIC. When then , so the BIC will put ***a larger penalty on a large model with more parameters*** when you have more than 7 observations than the AIC. In other words, ***when the number of observations is larger than 7, BIC will tend to choose smaller models than AIC***.

* For a least squares model with variables, the adjusted statistic is calculated as follows:

Where TSS is the Total Sum of Squares. Maximizing the adjusted is equivalent to minimizing . Unlike the statistic, the adjusted statistic ***puts a penalty on the inclusion of unnecessary variables in the model***. The advantage of using the adjusted statistic is that it does not require the estimate , and this allows you to use it ***even when*** , which is not the case for , AIC and BIC. However, adjusted cannot really be generalized to other types of models (so you cannot use it for, for example, logistic regression).

6.5 Validation and Cross-Validation

* Validation and Cross-Validation can be used to choose the best model size from the different methods (best subset, forward stepwise and backward stepwise) by ***using the training data set to pick the best model of size to get the models*** and the ***use the test data set to choose which of these model sizes gives the best prediction error.***
* A big advantage of using this approach is that ***it does not require an estimate of*** . That is important because when , it can be very difficult to get an estimate of . That is because you cannot use the full model (that is, a full model that contains all predictor ) in that case as it would saturate the model and give an error of 0. Thus, in that case you would have to ***arbitrarily*** pick a subset of the predictors to include in the model, but you do not know whether these are good variables or noise variables.
* Another advantage is that ***it does not require you to know the number of parameters in the model .*** Although for a linear model it may be clear what the number of parameters is, the value of may not at all be clear or easy to determine when you use other types of models.
* To pick a good model size after evaluating the prediction error using k-fold cross validation, you can use ***one-standard-error rule***. In that case, you calculate the standard error in the prediction error as a function of model size from the different predictions errors you get from the k different folds of the cross validation. For example, if you do 10-fold cross-validation you can calculate the standard error of the mean of the 10 numbers that you get from the cross-validation. ***Then you choose the simplest model (the model that contains the least predictors) that comes within one standard error of the minimum of the prediction error.***
* The idea behind the one-standard-error rule is that we cannot really tell the different model sizes apart based on the prediction error when they fall within one-standard-error of each other. In other words, the variation in the prediction error is too large to tell them apart. ***In that case we favor the smallest model that still falls within one-standard-error of the minimum, because it is easier to interpret***.

6.6 Shrinkage Methods and Ridge Regression

* Forward stepwise selection, backward stepwise selection and best subset selection all use ***least squares to fit the coefficients.***
* Another method discussed here is ***Shrinkage***, which involves two techniques: ***Ridge Regression and LASSO***.
* ***Ridge Regression*** is a modification of least square fitting where the coefficients are being pushed towards 0 by giving a penalty (called a ***shrinkage penalty***) for large coefficient values. Mathematically, the least squares fitting procedure estimates the parameters by minimizing:

While in ridge regression the coefficient estimates are the values that minimize:

Where is the ***tuning parameter***.

* The result is that in ridge regression there is a ***trade-off between the fit (RSS) and the size of the coefficients***, as it is costly to have many large coefficient values.
* The tuning parameter is used to control the relative impact of these to factor and ***selection of a good value for is critical***. The value of is chosen using ***cross-validation***.
* When the result will be the same as least squares fitting, but the larger is, the stronger the effect of doing ridge regression will be.
* The effect of ridge regression is that some of the parameters in the model will ***shrink towards 0*** and therefore it will cause these shrunken parameters.
* One useful representation of the effect ridge regression has had on the estimated coefficient values is the **-Norm**:

When the -Norm is closer to 0, it means that the effect of ridge regression has been strong and has shrunken most coefficients towards 0, while if the -Norm is larger it means that the effect of ridge regression is weaker and the coefficient estimates have largely been obtained through least squares fitting.

* An important thing to pay attention to is that ***Ridge Regression is not scale invariant*** (while least squares fitting is scale invariant) and therefore ***it matters a lot whether you scale the variables or not***. This means that if you multiply or divide a variable by a constant, this will significantly affect the outcome of ridge regression. This also means that ***the units matter*** (for example, meters vs. kilometers will give different results).
* Because of this it is ***very important to standardize the variables before doing ridge regression!*** Standardizing the variables is done by:

This formula takes the variable or feature and divides it by the standard deviation of that feature over all variables. This will make the standard deviation of all the features equal 1.

* Although ridge regression causes some of the coefficients to shrink to 0, it does not actually select any of the variables by making some of the coefficient exactly 0 (they approach 0, but they never fully reach it).

6.7 The Lasso

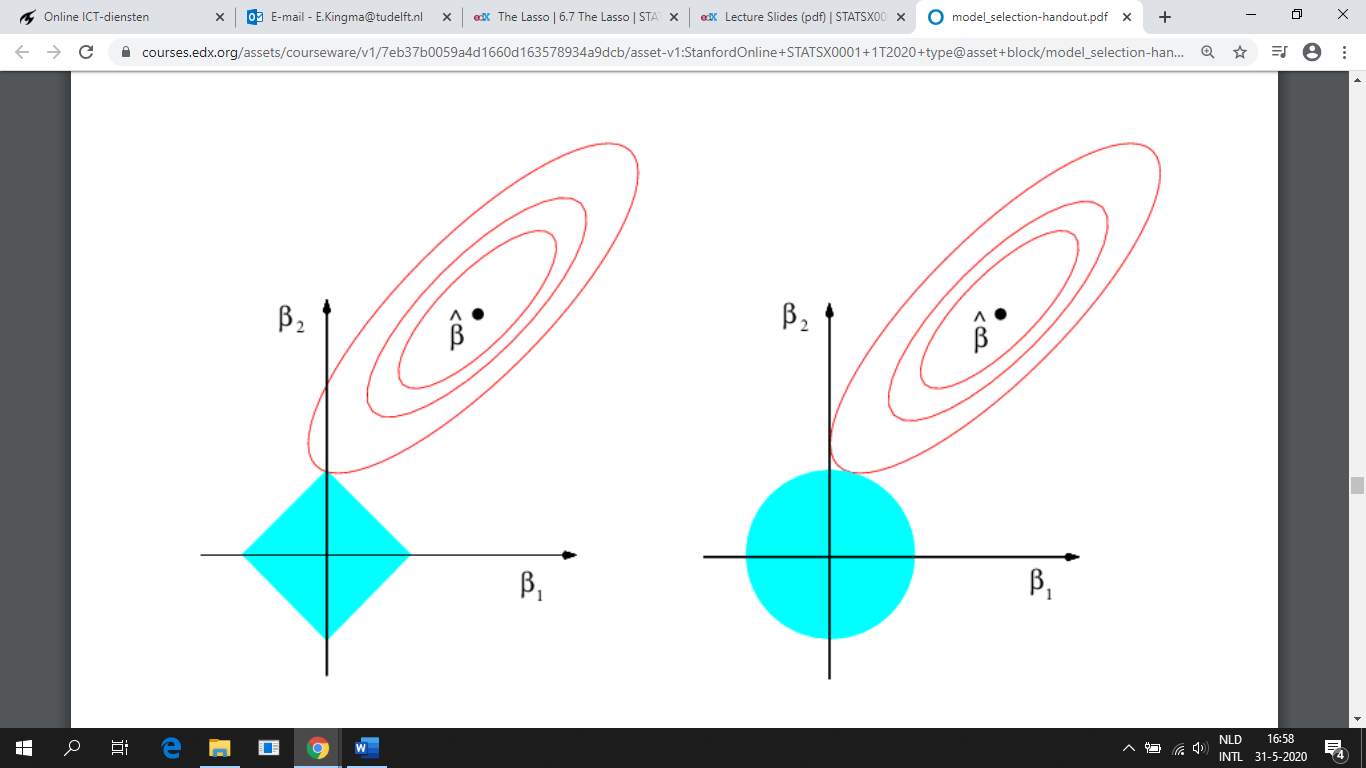
* The ***drawback of ridge regression*** is that it does not actually select variables by setting their coefficient to 0.
* ***The Lasso*** is an alternative to ridge regression that overcomes this drawback. The lasso coefficients,   minimize the quantity:

Now the penalty on the coefficient size is the ***absolute value*** of the coefficients instead of the sum of the squares as was the case with ridge regression.

* The penalty from the lasso is called the penalty (because the penalty is based on the absolute values) instead of the penalty for ridge regression. The norm of a coefficient vector is given by .
* As in ridge regression, the lasso makes the coefficient shrink towards 0, but ***it can also set the coefficient values exactly to 0 when is large enough.***
* This is referred to as ***sparse***: the lasso yield sparse models, that is, models that involve only a subset of the variables.
* As in ridge regression, the value of is chosen by cross-validation.
* An intuitive way to see why the lasso makes coefficients go to 0 while ridge regression does not is to consider it as an optimization problem with a limit in budget. This would mean that for the lasso the problem becomes

While for ridge regression the minimization problem would be:

The difference between the sum of the absolute values in the lasso and the sum of the squared values in ridge regression means that ***the constraint region of the lasso is diamond shaped, while the constraint region for the lasso is a circle.***



**Left: RSS contour plot for the lasso. Right: RSS contour plot for ridge regression.** In this case the model contains two coefficients, and **.** The red circular lines are contour plots of the RSS value, the dot denoted by is the point at which the RSS is minimal, the farther the red lines go out, the higher the value for the RSS on that contour. The blue region is the region of coefficient values that satisfy the constraint . Note that for the lasso, the lowest value of the RSS (that is, the most optimal fit) that satisfies the constraint lies on the y-axis, where is zero. For ridge regression, this point is not on the y-or x-axis, and as a result ridge regression does not set any of the coefficients to 0 during optimization. This difference is caused by the different shape of the constraint (blue) regions of the lasso and ridge regression.

* ***When you have a dense (or non-sparse) situation***, the lasso and ridge regression will give very similar results since the model depends on all the variables (and so none of the coefficients have to be set to 0).
* ***When the situation is sparse***, the lasso will give significantly better results than ridge regression.
* The lasso will give better results when the ***model is quite sparse***, while ridge regression will give better results when the ***model is quite dense***. So depending on whether the model is sparse or dense it is better to do the lasso or ridge regression.
* Usually, you do not know beforehand if the model is sparse or dense. In that case you apply both methods and select the model coming out of each method using cross-validation and then compare the cross-validated error for the two methods to determine which one works better.

6.8 Tuning Parameter Selection

* For the lasso and ridge regression we ***need to determine a value for***  and the solution of both methods depends strongly on the value of .
* ***Cross-validation*** is a good method for choosing .
* You ***cannot use other methods as AIC, BIC and***  because these methods depend on the number of parameters that are used in the model. In this case, the number of parameters is unclear because although all parameters may be in the model, some of them might shrink by using ridge regression, which makes it unclear how many parameters are actually being used to fit the model. In other words, in this case does not only depend on the number of parameters that you have used, but it also depends on how you fit them.
* The procedure for using cross-validation is the same as you do for subset selection: you subdivide the data into parts, you fit the model to parts and apply ridge regression and then you record the error on the last part (the part of data that was not in the subsets) and record the error. You do this times with each time a new subset of the data playing the role of the validation part. Then you add up all the error values on the different validation sets and you can make a cross-validation curve as a function of .

6.9 Dimension Reduction Methods

* In ***dimension reduction methods*** you first ***transform the variables*** (by making linear combinations of the original predictors ) and then fit a least squares model using these transformed variables.
* This is called dimension reduction because you use the original predictors to fit a new model containing predictors, where (which you can do as will be a linear combination of the different predictors).
* So let represent the ***linear combinations*** of the original predictors:

For some constants . Then you can fit the following linear regression model using ordinary least squares:

Here, are the regression coefficients and the are the predictors.

* If the constants are chosen wisely, this dimension reduction method can ***outperform the original least squares*** that you would have had when you had used the unprocessed predictors.
* This data reduction method is the same as doing ordinary least squares fitting, but ***with a constraint on the coefficients*** . The summation term in the equation for the fitting using the new predictors can be rewritten as:

Where:

This shows that dimension reduction is the same as ordinary least squares fitting, but where the estimated coefficients are constrained because they must satisfy the above equation.

* This method can result in a model that has ***both lower bias and lower variance*** relative to the model that you would have gotten if you had done least squares on the original variables. To achieve this ***it is crucial that*** .

6.10 Principal Components Regression and Partial Least Squares