**Linear Regression**

**Author: Yusuf Tatlier**

1. **Simple Linear Regression**

**Goal**

Explaining variation in the dependent variable/response , by the variation in the independent variable/explanatory variable/covariate/regressor .

*Assumptions*:

1. (Data Generating Process)
2. The observations are fixed numbers, repeating the experiment gives the same numbers
3. and iid (3 combined assumptions)
4. are unknown fixed numbers

Remarks: Trend of is given by part , while the variance is given by .

**Approach**

The model will be used and so the coefficients and are estimated by and . This is done by minimizing the sum of the squared error terms, which is . This is done by finding the parameters for which the partial derivatives are zero:

This gives:

It is good to know that for standardized variables, , while doesn’t change.

**Coefficient of determination and F-test**

The following holds:

As cross-terms drop out, this leads to:

By naming all three terms, this can be written as:

Here:

* SST: Sum Squared Total
* SSE: Sum Squared Explained
* SSR: Sum Squared Residuals

A measure of how well the fit is, is the coefficient of determination :

Gain in can be used to compare two nested models. More specifically if we look at two nested models indexed by 1 and 2, with and number of coefficients respectively where (i.e. model 1 is the super model) and examine the scaled gain in the SSR (so ), this quantity can be shown to be equivalent with the F-test. Assuming that the model is fitted on a dataset with datapoints, the following relationship holds:

It needs to be noted here that the deviance measure which we will later discuss in context of GLMs coincides with the SSR. So in statistical tools the deviance (function) can be used to calculate this quantity.

**Properties of parameters:**

From , it can be obtained that , which means that

While from , it can be obtained that , which means that

This implies that the residuals have degrees of freedom

It can be shown that *b* has the form , as it is a linear combination of normally distributed random variables, is itself normally distributed.

It can be shown that:

Furthermore it can be derived that and .

**Multiple Linear regression**

**Goal:** Explaining variation in the dependent variable/response , by the variation in the independent variables/explanatory variables/covariates/regressors .

*Assumptions*:

1. (Data Generating Process)

Or in matrix form , where all components of this equations are vectors or matrices that contain the set of equations for all .

1. The observations in the matrix are fixed numbers, repeating the experiment gives the same numbers.
2. and iid (3 combined assumptions).
3. are unknown fixed numbers.

It can be writen that:

The first term is the partial effect, while the second term is the indirect effect of the regression. In interpreting coefficients in linear regression, we use a ‘ceteris paribus’ approach in which we assume all other coefficients/effects are frozen (so don’t look at indirect effects).

**Approach:** The model is used and so the vector is estimated by the vector . This is done by minimizing the sum of the squared error terms, which is . This is done by finding the parameters for which the partial derivatives are zero:

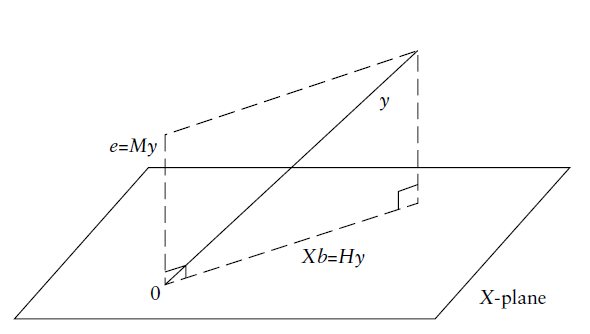
This gives:

It is good to know that for a standardizes variables, , while doesn’t change.

The OLS estimate is BLUE, which means that it is the most efficient linear unbiased estimator for .

**Geometric interpretation:**

There are two matrices *M* and *H* for which the following situation holds:



**Properties of parameters:**

It holds that

So it holds that . Additionally it can be shown that . As is normally distributed, by the above relationship it further holds that .

**Tests**

As it holds that in which is element in , it holds that that

and .

The t-test can be used to test the null hypothesis that vs. the alternative hypothesis that this is not the case.

Testing multiple factors at once can be done by the F-test.

**Omitted variable bias**

(i). Suppose that the true data generating process is given , while the model is used. Now it can be seen that

Note that contains an indirect effect in the form of , so leaving out variables doesn’t mean that the effect of the variable automatically disappears. However it can be shown that

So omission of variables that are:

(i). Determinants of

(ii). Correlated with an independent variable

leads to biased estimates that have a smaller variance however.

Another way to look at it is that . The bias thus occurs when is correlated with and the omitted variable is significant.

**Too many variables**

Now suppose that the data is generated by the process , while the model is employed. It is important to note that, although this model neglects that , it is not wrongly specified as it satisfies the most OLS assumptions, the variance of will now increase and has the form:

It is important to note that model complexity increases variance of estimators, as they are more strongly based on the sample set (with possible overfitting as result).

**Multicollinearity**

It can be shown that the variance of the estimator of the ’th factor is given by

Here is the sample variance of the’th factor and is the coefficient of determination of regression of ’th variable on the other variables. Here is called the Variance Inflation Factor (VIF). The case is called perfect multicollinearity and the OLS estimate doesn’t exist anymore as the matrix *X* is not invertible. In milder cases, when is larger, the following possible cures exist:

(i). Increase the sample size

(ii). Drop out variables (There is a trade-off with omitted variable bias however)

**Partial effects**

As seen in this section, in practice the used predictors themselves can have a significant correlation, from which it can be difficult to assess the actual effect of a certain regressor and the response variable. Leaving out variables will in this case lead to an omitted variable bias. So what should we do?

In such a case the variables can be cleaned for indirect effects before regression or determining correlations.

Suppose that there are two regressors . The partial effect between and can be determined by performing the following steps:

* Regress on and determine the differences between the realizations and model predictions .
* Regress on and determine the differences between the realizations and model predictions .
* Regressing on (or by taking the correlation between the two, depending on the goal).

So the idea here is to look at filtering out unexplained variation in the variables before testing for effects.

**Model construction**

In constructing models, different metrics can be employed:

(i). Information criteria: AIC[[1]](#footnote-1), BIC

(ii). Out of sample prediction: RMSE

(iii). Significance: t-tests, F-tests

(iv). Model Fit: and adjusted

In model selection a stepwise forward or backward method can be used. In the first case one starts with an empty model and adds variables until all terms are significant, while in the later one starts with the full model and variables are dropped until all terms are significant.

**Testing model assumptions**

**Correct model specification**

The underlying data doesn’t have to come from a linear data generating process. This can be tested by simply looking at the plots or the residuals, although more complex relationships can’t be recognized easily. Non-linear terms can be added to the regression and it can be tested whether the estimated factors significantly differ from zero.

**Normality**

The t-tests on the estimated factors make an assumption that the residuals are normally distributed. In case this is not the case, the estimated factors can still be unbiased but the t-test results will be unreliable. In case of a large sample size, the distribution of the variables will still converge to the normal distribution and the t-test can still be used to make inferences upon. There are different tests for normality, like the Jarque-Bera test and the Shapiro-Wilk test.

**Heteroskedasticity**

In case of heteroskedasticity, it holds that that:

So the variance of the residuals is not the same over the entire dataset. This implies that

In this case it holds that:

In case the residuals are not homoskedastic, the estimated factors are still unbiased but the t-tests will be unreliable. The OLS estimator will not be the most efficient estimator in this case. It is not straightforward how to solve the heteroskedasticity problem as it is often not known how the standard deviation of the residuals depends on the regressors. In case this relationship is unknown, a correction can be applied. Suppose for example that it holds that , in this case it will hold that . It can be shown that by making the transformations

The following transformed data generating process is obtained:

This transformed process is homoskedastic, with the efficient OLS estimator:

This estimator makes sense as data with a lower variance has a higher weight in determining the estimate, as there is smaller uncertainty in these observations.

There are different tests in order to test on heteroskedasticity, like the Breusch-Pagan and the Goldfeldt-Quandt tests, which have as null hypothesis that the residuals are homoskedastic.

**Autocorrelation**

Autocorrelation means that has non-zero diagonal elements, so that there is a correlation between different residuals. Autocorrelation has the same negative consequences as heteroskedasticity, which means the OLS estimate is unbiased but not the most efficient estimator. A possible test for autocorrelation is the Durbin-Watson test.

**Endogeneity**

**Cause of endogeneity**

Endogeneity occurs when X is correlated with the error term, so , implying that . Explanatory variables in the OLS should be exogeneous, so

The violation occurs due to violation in the assumption that the explanatory variables are non-stochastic. This has as implications that:

* There is a controlled experiment in which the regressors would stay constant. In economics controlled experiments are rare however, X variables are the consequence of an economic process and therefore stochastic.
* *b* is a consistent estimator of β

In case the non-stochasticity assumption is violated, the OLS is not useful under endogeneity.

Real world meaning of endogeneity is that you get the causation wrong. An example is the omitted variable bias which was discussed. In this case the error term will compensate for significant variables that were omitted and correlated with other regressors.

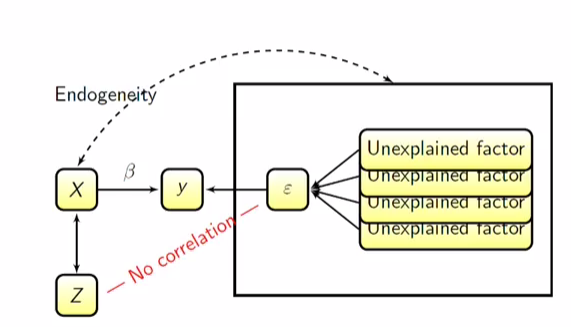
*Example of endogeneity*: Regression ‘student score’ on ‘college attendance’, while omitting ‘student motivation’. Often endogeneity occurs by omitting drivers of certain behavior.

Noisy dependent variables lead to endogeneity, leading to ‘stretching’ and to lower estimates of OLS factors.

**Correcting for endogeneity**

How can one correct for endogeneity, without needing data on the omitted data? New variables called instruments need to be introduced. Instruments have to satisfy:

* + - and are correlated
    - is uncorrelated with the error term



Using this information one can obtain new estimators for , which is known as the 2SLS estimator. This can be obtained in two steps:

1. Regress on *,* so fitting a model  *,* which will give the OLS estimate , with a fit of , in which is the projection of on the linear space spanned by (see figure …)
2. Now regress on , giving the 2SLS estimator

In order to use this method note that:

* By the Gauss-Markov theorem the OLS estimator will have a smaller variance than the 2SLS estimator. So the latter should be used only in case of clear endogeneity
* One needs to choose a well-suited instrument (for each endogeneous variable!), often this is based on expert opinion. In the ‘student score’ example, an instrument for the ‘college attendance’ variable is for example the travel time until the university as this is helps explain ‘college attendance’ but should not be significantly related to the ‘student score’.

**Other Regression Approaches**

This section discusses other approaches to regression, more on this can be found in the R tutorials. GLMs and generalizations are discussed in a separate document

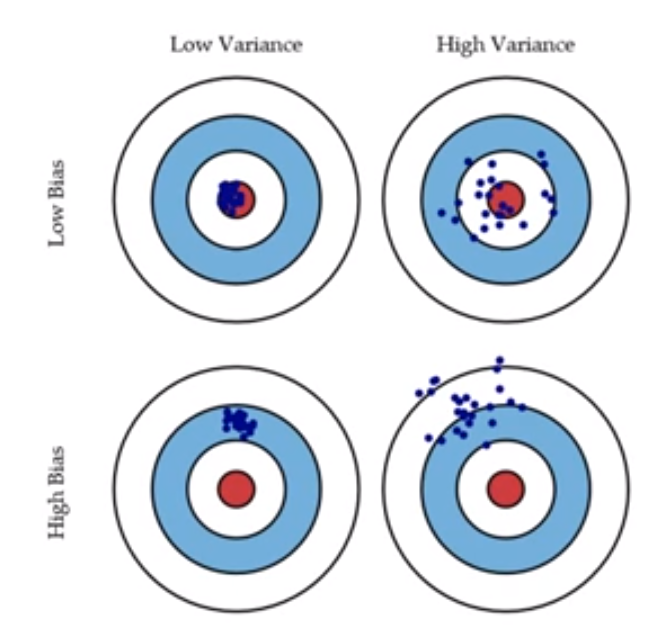
**MSE optimization**

The OLS method provides BLUE (Best Linear Unbiased Estimators) in case the assumptions are satisfied. One could consider estimators that are biased but have potentially a lower variance (less uncertainty). In this manner one could aim to minimize the Mean Squared Error (MSE). The MSE measures the deviation of the estimate from the real parameter:

To give an example, for consider the sample mean estimator , then

One sees that the sample size is important, if the bias remains non-zero regardless of the sample size one should go for an unbiased estimator if the sample size is sufficiently large.

The figure below is quite popular and is used in order to showcase the bias variance trade-off one has in estimators.



It turns out the OLS estimate is not the most efficient one in many cases. There are other estimators for which the opposite holds, so that are biased but more efficient estimators, like it will be seen for shrinkage methods.

**Shrinkage methods (Ridge, Lasso)**

Shrinkage methods have the advantage that they tend to be less customized by the training set and therefore lead to an estimator with a lower variance and possibly a lower MSE. This has as advantage that they are less susceptible to overfitting.

In OLS variables are either included or not, which makes it a discrete process. In case of shrinkage methods all variables are included, but subjected to a shrinkage procedure. Depending on the type of method the coefficients will either be present regardless of the amount of shrinkage or even become zero. Two shrinkage methods will be discussed, being ridge and lasso regression.

**Ridge regression**

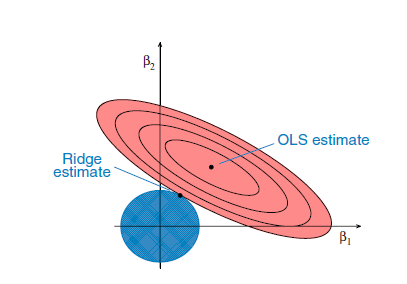
In case of the ridge regression the sum of the squared factors next to the sum of the squared residuals is included and the following minimization needs to be performed:

Here is the vector of all coefficients for and are the explanatory variables. The parameter λ determines the weight of the shrinkage term in the total. will often be a linear function in our case as we examine linear regressions.

Equivalent to this optimization is the optimization

With an additional requirement of the following type:

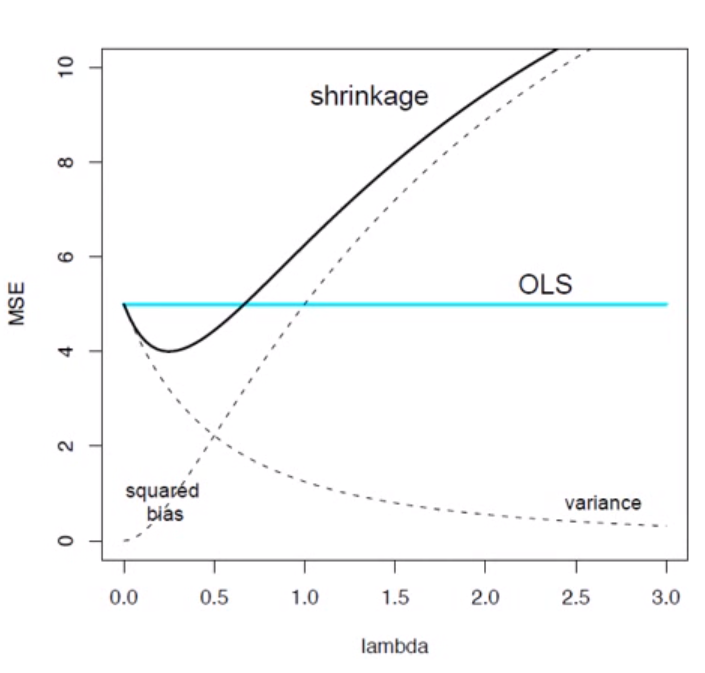
Making a separation between these two quantities gives a better understanding of the optimal coefficients that are returned by this procedure. From the below figure one can see that the second quantity (\*\*) geometrically represents a circle, increasing the shrinkage factor λ here will lead to a decrease in the coefficients in order to offset the increase in the total quantity. Movements arising in the coefficients will however affect (\*) and can increase this quantity. The optimal coefficients will therefore balance both quantities and lie somewhere in between the contours of (\*\*) and the OLS estimate.



There is an analytical solution for the ridge regression problem and is given by

Here is the matrix with observations of the independent variables and is the vector of the observations of the dependent variable. Note that this estimator is similar to the OLS estimator, but with additional terms on the diagonal. As , the estimated factor converges to the OLS estimate, while as , the estimate converges to zero.

Another way to view it is that by taking , now the estimates become biased in case , but for the right choice of the shrinkage parameter the MSE of the estimator is smaller as can be seen in the figure below.



One of the strengths of the ridge regression is that is still exists in case of perfect multicollinearity.

Choosing is not so straightforward, packages will often provide an optimal level based on performance on the training set. A visual rule of thumb is to choose the coefficient for which the estimators stabilize.

**Lasso regression**

This regression is similar to ridge regression, in which the sum of absolute coefficients, also called the norm, is included in the quantity that needs to be minimized:

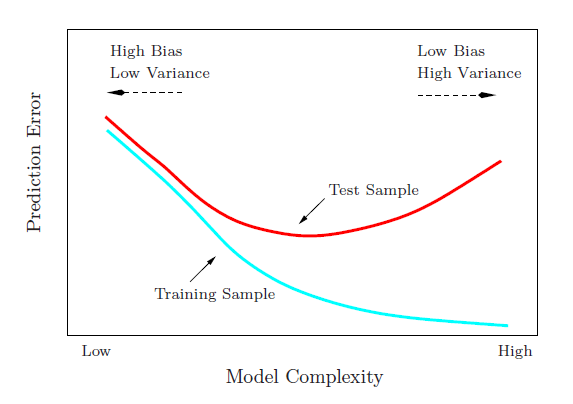
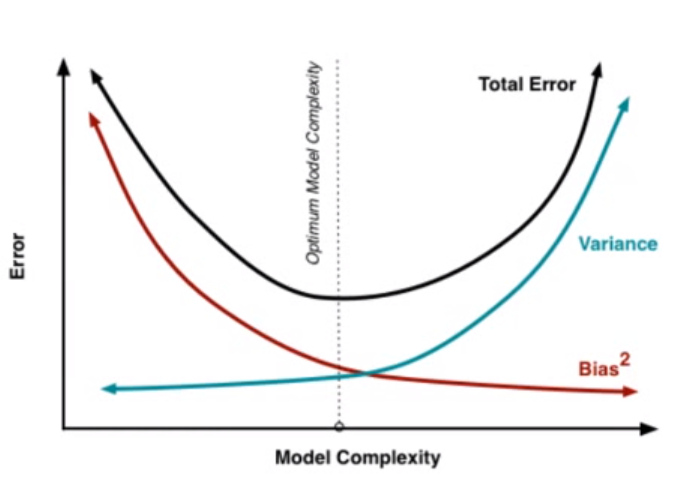
Just like in case of the ridge regression, the shrinkage term can be taken separately, in which case there is an additional requirement of the following form:

The contours from this second equation will follow a diamond shape as opposed to the circles in case of the ridge regression. Also as a result the big difference between ridge and lasso regression is that in the first case all coefficients are considered in all cases, while in the latter case factors can become zero. Another difference is that ridge regression has an analytical solution, while lasso doesn’t have one. The first feature might lead to lasso regression being preferred as it is more helpful in subset selection.

**Regression in Machine Learning**

A last topic is how regression is applied in machine learning and how it differs from the application in statistics. In statistics one is more interested in statistics and mechanics of the method. Model selection is based on these considerations. In Machine Learning approaches one is more pragmatic in setting up a model in that the model is chosen based on performance in a training set. The data is separated in at least a training and test set and the performance is then evaluated on the test set. An important aim is to avoid overfitting, as can be seen below a high model complexity can lead to a better performance on the training set and yield worse results on the test sample due to too much customization on the training sample.

The pragmatic philosophy of machine learning enables one to make use of a large(r) scala of methods as long as one has the necessary computational power.



**Time Series**

1. **Introduction and examples of important time series**

**Definition**

Cross-sectional data was considered thus far, but now time series data will be examined, which is a stochastic process indexed by time. The goal is to examine time series processes that can be used to model real world phenomena.

**Examples of time series**

* *White noise process:*

,

Note that as the are iid, the process at each time is independent from the other times. Practically all time series have a white noise component and it will also be seen that white noise generates various time series. Going forward will be used in order to denote white noise terms.

In financial mathematics a white noise term can for example be seen as new information entering the market causing fluctuations in the price.

* *Random walk with a drift:*

A random walk is basically a cumulative white noise process:

A random walk is actually a special case of an AR(1) process.

* *Moving Average process (MA(q)):*

A moving average process with drift of order has the form:

A moving average process is basically a linear combination of past white noise terms. In a financial mathematics context this can be seen as information entering the market and having an impact for a certain time period.

* *Autoregressive process (AR(p)):*

An autoregressive process of order p with drift is given by:

An AR(p) process is basically a linear combination of past values plus a white noise term. In a financial mathematics context a value process can for example be modeled as the value in the past (few) day(s) plus a white noise term representing new information in the market.

* *Autoregressive Moving Average process (ARMA(p,q)):*

The ARMA(p,q) process is a combination of the AR(p) and MA(q) processes and therefore has the form:

This process can be seen as a regression of the present outcome on past outcomes with the sum of the other terms as the residual.

* *Autoregressive Integrated Moving Average process (ARIMA(p,d,q)):*
* *Seasonal Autoregressive Integrated Moving Average process (SARIMA(p,d,q)):*

1. **Important concepts**

**Mean and auto-covariance**

One can determine the mean and (auto-)variance of time series, but note here that these can be time dependent. It will be seen that the special case that these quantities are time independent are important.

The mean function is defined in the following way:

The auto-covariance function is defined as:

Note here that .

The auto-correlation function is defined as:

For time series often the notation is used in case the auto-covariance function doesn’t depend on a specific time index but only on a time difference. This implies that the auto-covariance of the time series remains constant over time.

In case that for all the process at different times is independent. A special case is the so called white noise process, for which past values hold no predictive power for the future and the process is random. For a successful time series model, the residual needs to be a white noise process as otherwise not all information has been distilled from the dependent variable.

**Partial autocorrelation**

As explained for cross-sectional data, it can be difficult to determine the impact of a variable on the response because of indirect effects caused by correlations between regressors. This is in specific an issue for time series due to autocorrelation. Therefore the notion of partial autocorrelation is important in which indirect effects are partialled out. This follows the same steps/ideas as explained for cross-sectional data.

**Stationarity, consequences and order of integration**

In the analysis of time series, a systematic change in the behavior of the time series (like properties, trend etc.) is undesired, as otherwise properties derived in a certain moment will not hold at a different moment in time. In case no systematic changes occur, i.e. the series is ‘similar’ at different times, the time series is said to be stationary. There are levels/forms levels of stationarity:

* Strong or strict stationarity: The joint distribution of the time series remains unchanged, i.e. for all (integer) it will hold for the time series:
* Weak stationarity:
  + The mean of the time series is constant and does not have a time dependence
  + , i.e. the autocorrelation function only depends on the time difference and not on the time instances.

In case one non-stationary time series variable is regressed on another, one can infer non-existing relationships even if the processes are completely uncorrelated, this is known as spurious regression. In order to avoid this problem the time series that is used in the regression needs to be stationary.

Often time series will be non-stationary and the time series need to be differenced in order to obtain a stationary time series. A time series is said to be integrated of order , written as if differences need to be applied in order to obtain a (weakly) stationary time series. In many cases will be 1 or 2. Over differencing may introduce artificial dependence.

**Backward and difference operator**

The backshift operator moves the state of the time series with one time unit, i.e. is defined as:

and

The difference operator gives the difference of two consecutive states of a time series:

Note that it holds that:

Similarly it can be shown that:

**Polynomial backshift operator for MA(q) and AR(p)**

One can express the MA(q) and AR(p) process using a polynomial operator containing the backshift operator.

For an MA(q) process one can write:

So it holds that:

Here define the lag polynomial operator as . The corresponding characteristic polynomial is defined as .

For the AR(p) process one can write:

So that it holds:

Where the lag polynomial operator can be defined as . The corresponding characteristic polynomial is defined as .

**Linear process and invertibility**

is said to be a linear process if it holds that:

and

It can be seen that the auto-covariance of this process is given by:

Given a stochastic process and a white noise process , the process is called invertible in case that it holds that:

while converges.

One can look at the invertibility of the MA(q) and AR(p) process using the polynomial operators on the backshift operators as was derived earlier. For example for a MA(1) process one can write , where in specific for it holds that:

For convergence it is required that . Similarly it can be shown that for a general it holds that:

The that are not so straightforward to determine. Inverting a process yields a process. Similarly it can be shown that inverting a process yields a process, for some :

The key point here is that in case of stationarity the MA and AR processes can be inverted into each other (there is a duality). This can be useful in deriving quantities, let’s for example look at the process as a process:

**Stationarity, characteristic polynomial and invertibility**

The characteristic polynomial and invertibility concepts can be used in order to prove that a time series is stationary using the following results:

* A time series is stationary if and only if the roots of its characteristic polynomial all lie outside the unit circle.
* A time series is stationary if and only if it is invertible.

**Inverting ARMA(p,q) process**

For an ARMA(p,q) process it is possible to write:

By inverting the operators and it is possible to express the mixed ARMA process as a MA or AR process.

**ARIMA(p,d,q) process**

Given a time series that is not stationary, it is said to be ARIMA(p,d,q) if is ARMA(p,q) where:

So is .

1. **Stationarity for introduced classes of time series**

* *White noise process*

It is easy to see that a white noise process is stationary as the process at different times is iid.

* *Random walk with a drift*

Note that a random walk is not stationary, as it has a drift and the variance is increasing and blows up:

The first order difference results in a stationary series however:

This process has a stable mean and variance and is stationary.

It will be sees later that the random walk that was examined is a special case of an AR(1) process.

**MA(q) process**

*MA(1) process*

An MA(1) process has the following form:

This has as mean and variance:

The autocovariance for is only non-zero for :

Note that a MA(1) process is stationary for all .

*MA(q) process*

The *MA(q)* process has a constant mean:

Assuming that it can be seen that the autocovariance is only non-zero in case that as there is otherwise no overlap between the white noise terms. Assuming this condition holds, the following can be obtained:

For example for :

Note that a MA(1) process is stationary for all choices of .

Note additionally that using the backshift operator one can write an MA(q) process as:

Here is the polynomial operator acting on the white noise terms .

**AR(1) process**

*AR(1) process*

The AR(1) process has the following form:

By writing this out further:

It can be seen that this process is generated by white noise with exponential weights. It holds that:

Note that as becomes large and it will hold that .

In order to calculate the autocovariance, note that there is an overlap in the series starting from the backwards. All these terms in this region have a contribution with weights :

In specific it holds for the variance that:

Note that as becomes large and this auto-covariance goes to .

Also from the above calculations, there is a distinction between three situations:

Under these situations the process is stationary, unit-root and explosive respectively.

*AR(p) process*

Deriving the moments for a general AR(p) process requires more work and becomes messy.

It can be shown that the AR(p) process is stable if all roots of the characteristic polynomial lie outside the unit circle. More generally if for AR(p) process there are roots inside the unit circle, the time series is .

*ARMA(p,q) process*

For the ARMA(p,q) process it holds that it is stationary if the AR(p) part is stationary.

**AR(p) estimation: OLS and** **Yule-Walker estimation**

AR(p) model coefficients can be estimated by using simply an OLS approach in which the different lags are treated as different variables.

Another option is to use the Yule Walker equations. It can be shown that the autocorrelation of an AR(P) process satisfies:

Note that and .

As the Yule-Walker equations can be written out for different , the set of Yule-Walker equations can be expressed in matrix form , where is vector of autocorrelations, is the matrix of all lagged autocorrelations an is the vector of all coefficients. As these are unknown they need to be estimated. As is invertible and therefore the autocorrelations can be derived first in order to estimate by using that .

MA(q) estimation is less straightforward and we will not discuss it.

1. **Seasonality in time series**

**SARIMA**

SARIMA stands for seasonal ARIMA and is intended for correlations that only occur over larger lags. It does this by basically introducing a separate ARIMA relationship for the seasonal component, i.e. on top of the (non-seasonal) ARIMA parameters (p,d,q) SARIMA also introduces the parameters (P,D,Q,s) for the seasonal relationship. Here the first three parameters have the same meaning while s is the seasonality lag. This in total leads to the following relationship using characteristic polynomials for the non-seasonal and seasonal part:

It is not trivial how the time series is constructed once the parameters have been estimated. We will therefore look at an example:

Look at the example where . This results in the following relationship:

We can then rewrite this as:

From this example we can see that SARIMA follows the same logical structure as

This relationship indicates that we will see an autoregressive impact at 1, 12 but also at 13 due to the interaction/cross-term in the two components.

1. **Smoothing**

We have discussed time series analysis most generally using the SARIMA framework that we have built stepwise. We will now look at a different approach of forecasting time series, namely using exponential smoothing. This provides a different approach in dealing with time series, which can be used as a back-up for the SARIMA framework but is also a practical approach in itself.

**Heuristics and level, trend, seasonality**

Given time series realizations until time , future values can be predicted using simple heuristics. The most simple heuristic that uses this data is the naïve heuristic which assumes the last value provides the most accurate prediction, i.e. . Similarly one can use a seasonal naïve heuristic looking at a lag of steps. One can use more data simultaneously by for example looking at historical averages over a certain window which gives .

In making use of heuristics it can be useful to separate the level (often corresponding with a historic average), trend (improvement or decline over time) and seasonality (behavior depending on season, periodicity) of the series for modeling purposes.

**Exponential Smoothing**

We will look at different exponential smoothing methods.

* Simple Exponential Smoothing (SES): This models the time series based on previous predictions and new information using weights that decay (geometrically) with higher lags:

Here is the weight given to the latest realization, also called the level smoothing parameter.

* Double Exponential Smoothing: Notice that the SES doesn’t include a trend. Generalizing the SES by adding a trend leads to the Double Exponential Smoothing (DES) method. Here the prediction is the sum of both the level and the trend:

In which the level and trend are given respectively by (note that next to the level smoothing parameter we now have an additional trend smoothing parameter ):

and

* Triple Exponential Smoothing: The DES approach can be further generalized by adding seasonality in order to get to the Triple Exponential Smoothing (TES) method.

Depending on additional or multiplicative seasonality the time series will be modeled differently. Therefore one first needs to examine the time series for the type of seasonality present. One can make distinction based on the following rules of thumb, here the periodicity of the seasonality is *m*:

* + Additional seasonality will have a stable amplitude and leads to the following relationship:
  + Multiplicative seasonality will have a non-stable (increasing or decreasing) amplitude and leads to the following relationship:

The components are constructed as follows (in the given order):

Note here that the level is de-seasonalized.

**Main references**

This document was developed as reference for courses, separate R markdown files were set up (these can also be found on GitHub) to implement parts of the theory as discussed here.

The document can contain errors, please feel free to notify me on this point.

Mainly the below resources were used. Also some helpful pictures from the mentioned sources were used:

* Econometric methods with applications in business and economics, C Heij et al, 2004
* Elements of Statistical Learning (2nd edition), T Hastie et al, 2016
* Principles of econometrics with R, C Colonescu, 2016
* Linear models with R, JJ Faraway, 2015
* Machine Learning Foundations: A case study approach, University of Washington (Coursera)
* Practical time series analysis, The State University of New York (Coursera)
* Time series analysis and its applications with R examples (Third edition), R Shumway and D Stoffer

1. Can be done in R using the function step(). [↑](#footnote-ref-1)