Robert Hennings 12.10.2020

**Handout Moving Averages in Python**

**Zielsetzung und Idee**

Idee des Moving Averages, oder auch anders bezeichnet als Rolling-,Running- oder Moving Mean (MM), ist es Zeitreihen zu analysieren und kurzzeitige Änderungen sowie Trends abbilden zu können. Je nach Wahl der Zeitreihenanalysezeiträume soll der Moving Average die Daten glätten. Es sind praktisch betrachtet nichts anderes als Durchschnitte des Datensatzes über ein Zeitintervall.

**Verschiedene Berechnungsmöglichkeiten der gleitenden Durchschnitte**

Simple Grundidee der Umsetzung:

Es wird ein fixer Zeitraum bestimmt, z.B. 60 Tage, innerhalb dieser 60 Tage wird dann der Durchschnitt berechnet, es werden wieder die nächsten 60 Tage bestimmt und dort erneut der Durchschnitt berechnet und so weiter. Der jeweils ältere, weiter zurückliegende Zeitraum wird dann herausgelassen, es wird weitergezogen zum nächsten Intervall.

Gleitende Durchschnitte bilden oft das Rückgrat für deutlich komplexere Berechnungen in Form von Algorithmen, wie z.B. bei dem Autoregressive Integrated Moving Average (ARIMA), welcher die gleitenden Durchschnitte für Zeitreihenvoraussagen nutzt.

**Berechnungen:**

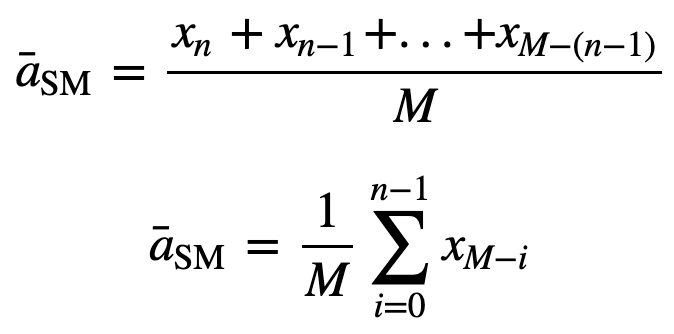
1. SMA (Simple Moving Averages)

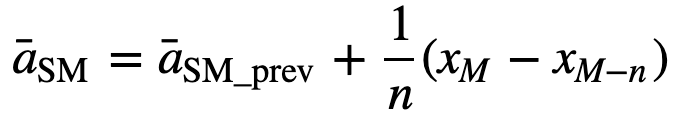
Ein Bild, das Zeichnung enthält.

Automatisch generierte BeschreibungBenutzt wird ein gleitendes Fenster, innerhalb welchem der Durchschnitt über den enthaltenen Datensatz berechnet wird, als gleich gewichtetem Durchschnitt über die n Daten.

Beispiel:

Gleichgewichteter Durchschnitt für N Daten mit dem Datenfenster M:



Da ja vorher der Durchschnitt für das Fenster M berechnet wurde, kann dies auch weiter für jedes weitere verwendet werden:

Die jeweils ältesten Beobachtungen warden herausgelassen und durch die neuen ersetzt.

**Implementierung als Python Code:**

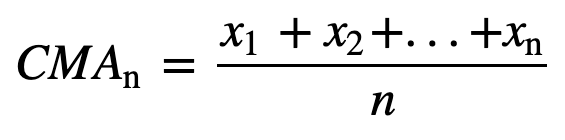
1. CMA (Cumulative Moving Average)

Anders als die vorangegangene Variante wird hier jede vorherige Beobachtung miteinbezogen und nicht herausgezogen um durch neue zu ersetzen. Daher ist dies aber auch keine sonderlich gute Methode um Daten zu analysieren, da alle Daten bis zum aktuellen Tag gemittelt werden (equally weighted average of the sequence of n values):

Ein Bild, das Zeichnung enthält.

Automatisch generierte Beschreibung

Bis hin zum aktuellen Tag, zur aktuellen Beobachtung:



Nachfolgend als:

Ein Bild, das Objekt, Uhr enthält.

Automatisch generierte Beschreibung

**Implementierung als Python Code:**

1. EMA (Exponentially Moving Average):

EMA gewichtet die aktuelleren Beobachtungen im Durchschnitt mehr, als die deutlich weiter zurückliegenden und wird daher auch als Standard herangezogen. Es wird mehr auf die aktuellen Preisänderungen reagiert, um den aktuellen Trend besser abbilden zu können.

Die Reaktion des EMAs ist direkt proportional zum Datensatz.

**Implementierung als Python Code:**

1. ARIMA Modell (Autoregressive Integrated Moving Average)

Diese Art des gleitenden Durchschnitts wird für Zeitreihenprognosen benutzt, als Fortführung des ARMA Modell. Es handelt sich um eine leistungsstarke Modellklasse, die den autoregressiven Teil und den gleitenden Mittelwertbeitrag des ARMA-Modells um die Differenzierung und Integration zur Trendbeseitigung und Herstellung der Stationarität erweitert. Anders als beim ARMA Modell können die zu analysierenden Zeitreihen hier auch Trendverläufe enthalten. Ganz typisch wären als Beispiel Verkaufszahlen, die Trends und Saisonalitäten (wiederkehrende Muster) beinhalten.

Es ist eine Generalisierung von den einfacheren Autoregressive Moving Averages und fügt den Part der Integration hinzu. Kernpunkte des Modells sind:

AR: Autoregression, Das Modell benutzt die abhängige Beziehung zwischen einer Beobachtung und einer Anzahl an weiterzurückliegenden Beobachtungen. Erstellt Vorhersagen auf Basis der zurückliegenden Ereignisse mithilfe von linearen Beschreibungen. Beobachtungen zum Zeitpunkt werden nur auf die vergangenen Beobachtungen zurückgeführt.

Allgemeine mathematische Darstellungsweise AR(p):



P bestimmt die Ordnung des Modells.

I: Integrated, Die Verwendung einer differenzierten Betrachtung von Rohbeobachtungen um die Zeitreihe stationär zu machen (Abziehen einer Beobachtung von einer anderen aus einem vorangegangen Zeitschritt/ Zeitintervall). Erste Ableitung der Zeitreihe.

MA: Moving Average, Das Modell nutzt die Abhängigkeit zwischen einer Beobachtung und einem Residual Error von einem MA der auf weiterzurückliegende Beobachtungen angewendet wird. MA-Prozess geht davon aus, dass Ereignisse zu einem bestimmten Zeitpunkt vom Rauschverhalten aktueller und früherer Zeitpunkte abhängen. Konkret bedeutet das, dass nicht vorangegangene Ereignisse die Vorhersagen bestimmen, sondern die vorangegangenen Schätz- oder Vorhersagefehler in die Berechnung des nächsten Werts einer Zeitreihe einfließen. Beobachtungen werden jedoch nicht auf die Beobachtungen , sondern auch auf den nicht-beobachteten Fehler  der vergangenen Zeitperioden zurückgeführt, der ebenso einen Einfluss auf Deine zukünftigen Beobachtungen ausübt.

Allgemeine Darstellung von MA(q):



Die Kombination von AR und MA Modellen liefert das ARMA(p,q) Modell:



Herleitung des ARIMA(p,d,q) Modells aus dem ARMA(p,q) Modells:

Nach einer zusätzlichen Differenzierung der Zeitreihe und einer Integration nach Anwenden des Modells, spricht man von ARIMA-Modellen. Sie werden eingesetzt, wenn die Filterung eines Trends benötigt wird. Der Parameter d des ARIMA[p,d,q]-Modells bestimmt die Zahl der Differenzierungsschritte:

* Zuerst wird die Zeitreihe d-mal abgeleitet, bis sie stationär ist.
* Dann wird ein geeignetes ARMA[p,q]-Modell an die resultierende Serie angepasst.
* Zuletzt müssen die geschätzten Voraussagen d-mal integriert werden.

**Wrap Up AR-AM Prozesse**

MA-Prozesse sind besser darin, Abhängigkeiten in Deinen Beobachtungen zu modellieren, die kurz auftreten und sich dann wieder auflösen, während AR-Prozesse zeitliche Abhängigkeiten modellieren, die von Dauer über die Zeitreihe auftreten. Um herauszufinden, ob ein AR oder MA-Prozess vorliegt, hilft ein genauerer Blick auf die Autokorrelationsfunktion der Zeitreihe.

**Festlegung AR-AM Prozess**

Autokorrelationen können über ein Korrelogram der Zeitreihe mit sich selber bzw. ihren lags, also ihren verzögerten Wirkungen, beschrieben werden. Stellen wir eine Autokorrelation über mehrere lags fest, handelt es sich um einen autoregressiven Prozess, der Abhängigkeiten über die gesamte Zeitreihe berücksichtigen sollte. Stellen wir nur Abhängigkeiten für eine kurze Dauer fest, handelt es sich eher um einen MA-Prozess. Falls AR-Prozesse vorliegen, stellt sich die Frage, ob es sich um AR-Prozesse höherer Ordnung handelt, also ob es sich um Abhängigkeiten handelt, die zwar über die gesamte Zeitreihe, aber nicht nur von jeweiligen vergangenen Zeitpunkt, sondern auch von Zeitpunkten in der ferneren Vergangenheit abhängen. Dafür hilft ein Blick auf die partielle Autokorrelationsfunktion. Diese untersucht die Abhängigkeiten in der Zeitreihe, wenn die Abhängigkeit zum Zeitpunkt t-1 aufgelöst wird. Beobachten wir dann weitere Abhängigkeiten über den Zeitpunkt t-1 hinaus, solltest Du AR-Prozesse höherer Ordnung modellieren.

**Basis: Box-Jenkins Prozess**

Das ARIMA-Modell basiert in Teilen auf dem sogenannten Box-Jenkins-Ansatz. Unter anderem behandelt der Box-Jenkins-Ansatz die Festlegung der Ordnung des Modells (Spezifikation), die Schätzung der Parameter des Modells und die Überprüfung der Güte des Modells (Diagnose oder [Validierung](https://www.bigdata-insider.de/was-ist-validierung-a-884683/)). Der Ansatz geht auf eine Arbeit von George E. P. Box und Gwilym M. Jenkins aus dem Jahr 1970 zurück. In der Arbeit gehen die beiden Wissenschaftler von einem stochastischen Prozess zur Modellierung von Zeitreihen aus. Dieser stochastische Ansatz erlaubt für viele reale Zeitreihen beispielsweise aus der Ökonomie eine realitätsnähere Betrachtung als andere Ansätze. Box und Jenkins fordern eine sparsame Parametrisierung des verwendeten Modells.

**Spezifikationen eines Modells**

Aufgabe der Spezifikation eines Modells ist es, die grundlegenden Parameter und die Ordnung des Modells zu bestimmen. Anschließend folgt die Schätzung der Parameter. Die Validierung ermöglicht es, die Eignung des Modells zu beurteilen. Hierfür lassen sich verschiedene Kriterien heranziehen. Erfüllt ein Modell die Anforderungen nicht, kann es durch erneutes Spezifizieren und Schätzen optimiert werden. Ist ein optimales Modell gefunden, ist es für Zeitreihenanalysen und Vorhersagen nutzbar.

Jeder dieser drei einzelnen Modellbestandteile wird im Gesamtmodell durch seinen eigenen Parameter repräsentiert, daher wird das Modell oft wie folgt geschrieben: ARIMA(p,d,q)

P: Die Anzahl an Lag Beobachtungen im Modell, auch bezeichnet als die Lag order

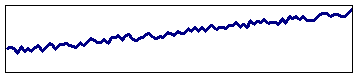
D: Die Anzahl der Differenzierungen der Rohbeobachtungen, auch bezeichnet als der Grad der Differenzierung

Q: Die Größe des MA Fensters, auch bezeichnet als die Order of MA

**Erzielung von Stationarität**

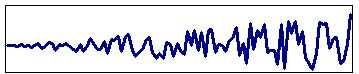
Die AR- und MA-Prozesse erfordern eigentlich stationäre Zeitreihen. Stationär bedeutet, dass sich die Randbedingungen einer Zeitreihe nicht verändern. Die zugrundeliegende Verteilungsfunktion der Zeitreihenwerte muss zeitlich konstant sein. Mittelwert und Varianz sind zu jeder Zeit gleich und folgen beispielsweise keinem Trend. Das Besondere am ARIMA-Modell im Vergleich zum ARMA-Modell ist, dass es durch eine zusätzliche Differenzierung und Integration Trends herausfiltern kann und durch diese Trendbeseitigung die geforderte Stationarität herstellt. Mit dem ARIMA-Modell lassen sich daher auch Zeitreihen analysieren und beschreiben, die einem Trend folgen.

Beispiel zur Veranschaulichung:



Hier nimmt der Mittelwert im Zeitverlauf zu, Zeitreihen mit (nicht nur linearem) Trend können mit dem ARIMA Modell unter Umständen erfolgreich beschrieben werden.

Im Gegensatz dazu ist folgender Fall nicht gut verwendbar, da die Varianz mit der Zeit zunimmt.



Zeitreihen mit veränderlicher Varianz und veränderlicher höherer [Momente](http://www.reiter1.com/Glossar/Glossar_Detailliert_M04.htm#Momente) können mit der ARIMA Methode nicht beschrieben werden.

**ARIMA Schritt für Schritt**

Schritt 1:

In einem Schritt für Schritt Verfahren wird hier einmal beispielhaft die Trendbeseitigung gezeigt:

Hat der Trend die Form eines Polynoms n-ter Ordnung:



dann lässt er sich einfach durch n-faches Differenzieren beseitigen

Aus Sicht des ARIMA Modells ist die Originalmessreihe folglich integriert (Integrated).

ARIMA Differenzierung

Nach 2facher Differenzierung (Abziehen jeweils benachbarter Werte) einer Reihe mit offenbar quadratischem Trend, erhält man eine Reihe, die offenbar keinen Trend mehr enthält. (Rauschen wurde der Übersicht halber weggelassen).

Saisonale Schwankungen (Periodizität) sind eine weitere Verletzung von Stationarität.

Sie lassen sich dadurch beseitigen, indem man im ersten Differezierungsschritt nicht jeweils benachbarte Werte voneinander abzieht, sondern zB. den 6. vom 1., den 7. vom 2., den 8. vom 3., usw. (In diesem Beispiel besteht die Periodendauer aus 5 Messwerten)

Anschließend kann -falls notwendig- wieder "normal", also zwischen jeweils benachbarten Werten differenziert werden.

Saisonale Schwankungen lassen sich aber auch durch die autoregressive Komponente (AR) beschreiben, welche im nächsten Schritt beschrieben wird.

Waren im konkreten Fall beispielsweise 2 Differenzierungen zur Erreichung von Stationarität notwendig, dann muss man zur Vorhersage bezüglich der Originalreihe erst wieder 2 mal integrieren.

Formal wird dieser Fall als ARIMA(p,d,q) mit d=2, also ARIMA(p,2,q) bezeichnet.

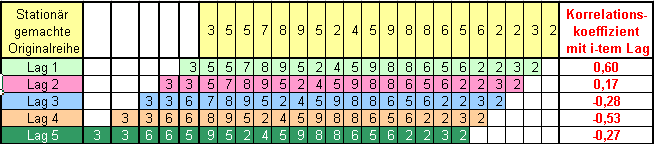
Schritt 2:

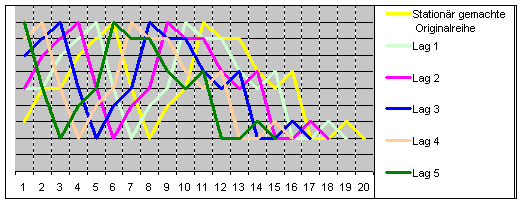
Ergebnis dieses Schrittes ist eine Gleichung der Form

,

der n-te Wert hängt also von einer Reihe vorausgegangener Werte ab. (Rauschen wurde hier weggelassen)

Um die Koeffizienten an-i zu ermitteln wird zunächst der [Korrelationskoeffizient](http://www.reiter1.com/Glossar/Korrelationskoeffizient_Produkt_Moment.html) zwischen der stationär gemachten Messreihe und der um i Messwerte verschobenen stationär gemachten Messreihe (sogenanntes "i-tes Lag") berechnet.





In der rechten Spalte der Tabelle (rot) stehen die Korrelationskoeffizienten zwischen der stationär gemachten Originalreihe und ihrem 1. bis 5. Lag.

Es ist nicht auszuschliessen, dass es unter den noch höheren Lags einige mit ebenfalls bedeutsamen Korrelationskoeffizienten gibt.

Bei der Berechnung der Korrelationskoeffizienten wird nicht zyklisch gerechnet (wie bei der [Autokorrelation](http://www.reiter1.com/Glossar/Autokorrelation.html)), sondern es werden nur übereinanderstehende Werte verwendet. Das bedeutet, dass die Anzahl Wertepaare für höhere Lags geringer wird.

Folgende Tabelle zeigt die Berechnungen der Signifikanz der Korrelationskoeffizienten.

Das genaue Vorgehen hierzu ist unter der Rubrik [Z-transformation](http://www.reiter1.com/Glossar/Z_Transformation_Fisher.html) beschrieben.

Ein Bild, das Uhr, Zeichnung enthält.

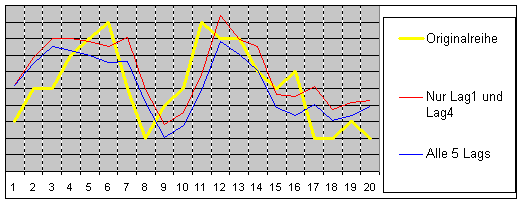
Automatisch generierte Beschreibung

Die Tabelle zeigt 5 einzeln und unabhängig durchgeführte Tests. Zur hier auftretenden Problematik siehe [Multiples Testen](http://www.reiter1.com/Glossar/Multiples_Testen.html) und [Alpha Inflation](http://www.reiter1.com/Glossar/Glossar_Detailliert_A01.htm#Alpha%20Inflation). Wir könnten hier an dieser Stelle entscheiden, dass der 1. und 4. Lag zur Modellierung ausreichen.

Genausogut könnten wir auch alle 5 Komponenten im weiteren Modell hinzunehmen.

Beide Fälle sind in folgender Grafik dargestellt.

Man sieht, dass die Hinzunahme der Lags 2,3 und 5 nicht unbedingt das bessere Modell ergibt.



Die Berechnung erfolgte so, dass die Summe der quadrierten Korreletionskoeffizienten der jeweils verwendeten Lags zu Eins normiert und gewichtet worden ist.

Die bisher ermittelten Modellgleichungen der beiden Modelle lauten:

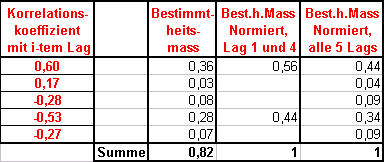
ARIMA(4,2,q):     bzw.

ARIMA(5,2,q):   

Hier ist 5.2 der Mittelwert der Originalreihe.

Die Werte der anderen Vorfaktoren ergeben sich aus den normierten [Bestimmtkeitsmassen](http://www.reiter1.com/Glossar/Bestimmtheitsmass.htm) (=quadrierte Korrelationskoeffizienten) der Lags, wobei die Vorzeichen von den Korrelationskoeffizienten übernommen wurden.

Folgend Tabelle veranschaulicht den Rechengang:



Es ist zu bedenken, dass die [Signifikanz](http://www.reiter1.com/Glossar/Signifikanz.html)werte in der Tabelle keine Verbindung mit einem mehr oder weniger guten Modell haben. Sie bedeuten lediglich, dass die Korrelationskoeffizienten "nicht bloss Zufall" sind.

Außerdem wurde hier nicht berechnet, wie Lag 4 direkt mit der stationär gemachten Originalreihe korreliert, da der hier berechnete Korrelationskoeffizient alle Einflüsse der Lags 1, 2, 3 und 4 beinhaltet.

Diese Art Korrelation heißt partielle Autokorrelation und wird hier nicht behandelt.

Es gibt spezielle Signifikanztests, die auf Autokorrelation testen.

* [Durbin h-Statistik](http://www.reiter1.com/Glossar/Glossar_Detailliert_D01.htm#Durbin%20h-Statistik): Testet die Autokorrelation der Zeitreihenwerte mit dem ersten Lag.
* [Durbin Watson Test:](http://www.reiter1.com/Glossar/Glossar_Detailliert_D01.htm#Durbin%20Watson%20Test) Testet die Autokorrelation der [Residuen](http://www.reiter1.com/Glossar/Glossar_Detailliert_R01.htm#Residuen) der Zeitreihenwerte mit dem ersten Lag.
* Testet also auf Autokorrelation der Fehler --> Schritt 3.
* [Ljung-Box Test](http://www.reiter1.com/Glossar/Glossar_Detailliert_L01.htm#Ljung-Box%20Test): Testet alle Lags auf einmal.

Schritt 3:

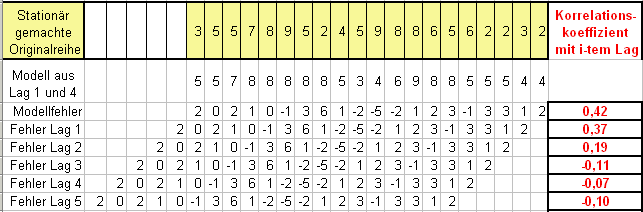
Unter "Fehler" ist hier zufälliger statistischer Einfluss zu verstehen, denn eine stationäre Zeitreihe besteht  aus Werten, die entsprechend der zugrundegelegten Verteilungsfunktion um einen zeitlich konstanten Mittelwert streuen.

Ergebnis dieses Schrittes ist eine Gleichung der Form

.

Die autoregressive Komponente des vorhergehenden Schrittes 2 wird also mit gewichteten Fehlern vorangehender Werte korrigiert.

Folgende Tabelle enthält in der obersten Zeile die stationär gemachte Originalreihe aus Beispiel 2, in der 2. Zeile das AR Modell aus Schritt 2, dann den Fehler des Modells aus Schritt 2, und schliesslich die ersten 5 Lags des Fehlers (also die Wertereihe des Modellfehlers um 1,2,3,4 und 5 Positionen verschoben).



Ohne explizite Rechnung ist bereits erkennbar, dass keiner der  Korrelationskoeffizienten [signifikant](http://www.reiter1.com/Glossar/Signifikanz.html) ist, ja sogar jeder relativ klein ist. Das deutet stark darauf hin, dass der Fehler des in Schritt 2 gewonnenen Modells fast nur aus zufälligem  ([normalverteiltem](http://www.reiter1.com/Glossar/Normalverteilung.html)) Rauschen besteht.

Das bedeutet konkret:

Der n+1 -te Messwert wird durch keine Zufallskomponente irgendeines vorhergehenden Wertes n, n-1, ...n-s beeinflusst

Die Fehler korrelieren nicht einmal mit den Werten selbst (0.20)

Es gibt in der vorliegenden Reihe keine Fehlerfortpflanzung.

Das bisher entwickelte Modell lautet demnach ARIMA(4,2,0)

4: Der autoregressive Teil des Modells (AR) greift bis auf den 4. Lag zurück

2: Die Originalreihe musste 2 Mal differenziert werden, um stationär zu werden.

0: Der Moving Average Teil (MA) greift auf keinen Lag zurück.

Im Folgenden seien zum allgemeinen Verständnis bildhaft ein paar "schöne" Autokorrelationsfunktionen und partielle Autokorrelationsfunktionen sowie die dazugehörende Nomenklatur dargestellt.

Die Säulen stellen Korrelationswerte dar.

Bei Autokorrelationsfunktionen, ACF, handelt es sich um Funktionen wie bisher beschrieben, d.h., es werden alle Einflüsse berrücksichtigt. In dem obigen Beispiel 2 wurde zwar entschieden, nur Lag 1 und 4 für das zu erstellende Modell zu verwenden, trotzdem sind dort die eventuellen Einflüsse der Lags 2 und 3 mit enthalten, denn Lag 4 kann ja von Lag 3 abhängen, und Lag 3 von Lag 2, und dieses wiederum von Lag 1; alternativ könnte Lag 4 aber auch direkt von Lag 1 abhängen und nicht von Lag 2 und 3, wieder alternativ könnte Lag 4 von allen Lags 1,2 und 3 abhängen; die bisher beschriebene Vorgehensweise zur Bildung der Autokorrelationsfunktion kann diese Fälle grundsätzlich nicht unterscheiden (ob Lags direkt voneinander abhängen oder über dazwischenliegende Lags).

Aus diesem Grund verwendet man Partielle Autokorrelationsfunktionen, PACF: Dort berechnet man z.B.den direkten Einfluss des Lags 4 auf die originale Messreihe und rechnet die Einflüsse der Lags 1,2 und 3 auf Lag 4 heraus.

Die blosse visuelle Analyse der beiden Funktionen ACF(pdq) und PACF(pdq) erlaubt in vielen Fällen bereits richtungsweisende Aussagen.

Allerdings erfordert bereits die Erstellung der beiden Funktionen schon spezielle Statistiksoftware.

Stehen die Parameter fest, so wird eine lineare Regression konstruiert auf Basis der angegeben Werte, um eine gewisse Statik zu erzielen (Herausrechnen von Trends und Saisoneffekten, die die Regression negativ beeinflussen).

Wird ein Wert von 0 für einen Parameter verwendet, so fällt dieser aus dem Modell und wird nicht berücksichtigt. Das Modell vereinfacht sich daher unter Umständen zu ARMA, AR, I, oder MA Modellen.

**Wichtige Annahme für das ARIMA Modell**

Als Annahme, dass ein ARIMA Modell für eine Zeitreihe konstruiert werden soll gilt, dass der Grundprozess ebenfalls ein ARIMA Prozess ist.

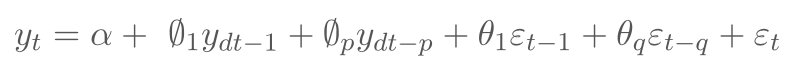
Dieser Prozess muss in den Rohdaten zunächst einmal nachgewiesen werden, um im nächsten Schritt das Modell zu erstellen.

Stationarität der Zeitreihe, eine Zeitreihe hat zu jeder Zeit den gleichen Mittelwert und die gleiche Varianz; Trends und unterschiedliche saisonale Effekte liegen in stationären Zeitreihen nicht (mehr) vor. Mit einfachen Worten könnte man auch sagen, Deine Zeitreihe ist gleichmäßig. In der Realität sind viele Zeitreihen aber nicht gleichmäßig, sondern verfolgen Trends oder Saisonalitäten, die unterschiedliche Mittelwerte und Varianzen über die Zeit zeigen. Zum Beispiel ist in der Regel die Varianz in den Verkäufen eines Unternehmens nicht regelmäßig und folgt einem gewissen Trend. Um nicht-stationäre oder ungleichmäßige Zeitreihen korrekt zu bestimmen, musst Du diese stationär machen.

**Wrap Up ARIMA Modell**

In ARIMA-Prozessen werden Trends in Zeitreihen über Differenzierung integriert und dadurch stationär. Das heißt, der Mittelwert Deiner Beobachtungen wird konstant, indem Dein Outcome, bspw. Deine Verkäufe  zum Zeitpunkt  von  zum Zeitpunkt  subtrahiert werden. Es kann genauso sein, dass die Varianz Deiner Beobachtungen nicht stationär ist, sodass  mithilfe des natürlichen Logarithmus stationär transformiert werden kann. Genauso sind natürliche, andere Transformationen der Zeitreihe möglich, um stationäre Zeitreihen sicherzustellen.

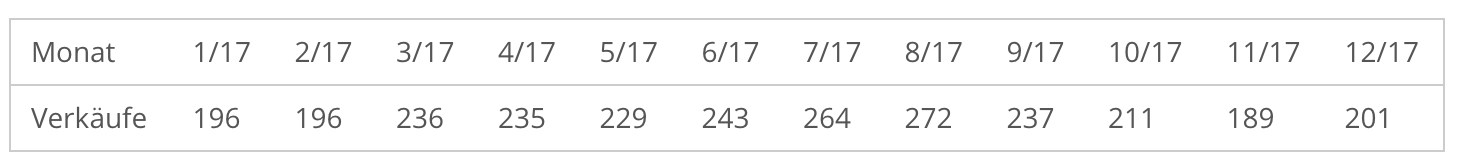
Kombiniert werden AR und AM Prozesse wie folgt:



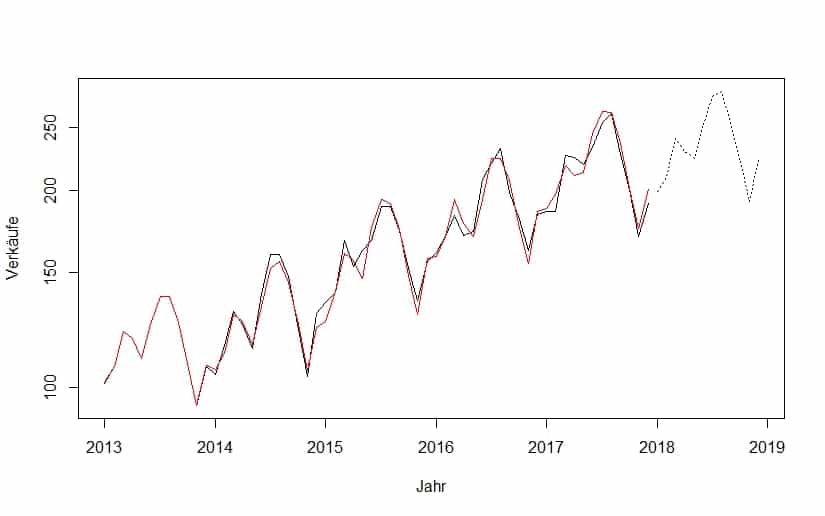
yd sind dabei Deine d-fach differenzierten Beobachtungen, die einem AR-Prozess mit p Ordnungen und einem MA-Prozess mit q Ordnungen folgen können. Die Parameter d, also die Ordnung der notwendigen Integration bzw. Differenzierung, sowie p und q müssen im ARIMA (p,d,q)-Modell spezifiziert werden.

**Einfaches Intro Beispiel**

Analyse der Verkäufe in einem Online-Shop seit Gründung vor 5 Jahren im Jahre 2013 und es ist festzustellen, dass das Geschäft mit jedem Jahr besser läuft. Ein Ausschnitt der monatlichen Verkaufszahlen im Jahre 2017 sieht folgendermaßen aus:

****

Der Verlauf der Verkäufe zeigt jedes Jahr zwei Verkaufsspitzen. Im Frühjahr und im Sommer gibt es relativ stark steigende Verkäufe, die daraufhin wieder genauso stark fallen. Nun soll herausgefunden werden, wie sich die Verkäufe im kommenden Jahr entwickeln werden. Erkennbar in der Zeitreihe ist, dass die Verkäufe keinem zufälligen Verlauf folgen, sondern einem klaren Trend mit saisonalen Effekten folgen. Einen solchen Verlauf kannst man gut mithilfe eines ARIMA-Modells beschreiben. Die Autokorrelationsfunktion bestätigt, dass im Frühjahr und im Sommer jeweils ein starker Monat existiert, nach dem die Verkäufe wieder abfallen. Die Verkäufe folgen keinem AR-Prozess, sondern einem MA(1)-Prozess. Man berücksichtigt den Trend in den Verkäufen, indem die Daten zur ersten Ordnung differenziert werden. Damit sind alle Parameter des Modells bestimmt. Das Modell folgt keinem AR-Prozess, sondern einem MA(1)-Prozess, den man zur ersten Ordnung integriert. Man bestimmt also ein ARIMA(0,1,1)-Modell. Die ungleiche Varianz wird mit einer Log-Transformation der Daten behoben, sodass sich der Verlauf einer stationären Zeitreihe korrekt schätzen lässt. Der untenstehende Graph der Verkäufe zeigt die tatsächlichen Verkäufe der letzten 5 Jahre zwischen 2013 und 2018 (schwarze Linie). Die Verkäufe lassen sich über das ARIMA(0,1,1) gut spezifizieren. Die Schätzung der Verkäufe (rote Linie) folgt relativ gut den tatsächlichen Verkäufen. Nun werden die Verkäufe mithilfe des ARIMA(0,1,1)-Modells für das kommende Jahr geschätzt und man sieht (gestrichelte Linie), dass sich der Aufwärtstrend in den Verkäufen fortsetzen und auch die saisonalen Effekte im Frühjahr und Sommer mit Wahrscheinlichkeit (nicht dargestellt) wiederholen werden.



**Überleitung zu (G)ARCH Modellen**

Aufgrund der Volatilität in den Verkäufen könnte man nun nicht mehr den Verlauf der Verkäufe beschreiben, sondern das Ziel könnte eher sein, die Varianz in den Verkäufen besser zu verstehen und vorhersagen zu wollen. Dann kommen sogenannte (G)ARCH-Modelle in Betrachtung.

**ARIMA Prozess**

**Anwendungsgebiete des ARIMA Prozesses**

* Auswertung von medizinischen Daten, wie sie beispielsweise bei EEG-Messungen auftreten
* Auswertung von Finanzdaten beispielsweise von Börsenkursen
* Auswertung von Daten eines Callcenters beispielsweise Bearbeitungsdauer eines Support-Vorgangs
* Auswertung von Wetterdaten beispielsweise Temperatur- oder Windvorhersagen
* Auswertung von sozialwissenschaftlichen Daten und Zeitreihen beispielsweise die Messung des Bevölkerungswachstums

**Implementierung als Python Code:**

Als Basisdatensatz dienen Shampooverkaufszahlen über einen 3 Jahreszeitraum, monatliche Zahlen (36 Beobachtungen). Als Attribut gibt es lediglich die Sales Zahlen pro Monat.

Bei der Verwendung wurde der Zeitraum ab 1900 gewählt zur Vereinfachung, wobei auch jeder andere Zeitraum als Index verwendet werden kann. Es fand demnach eine simple Festlegung auf einen einfachen Startzeitraum statt.

Nachdem der Startpunkt für den AR Teil ausgewählt wurde (Lag 5 nach der Autokorrelation), soll nun das ARIMA Modell aus der Library statsmodels gebildet werden:

Zentrale Befehle/ Vorgehensweise:

1. Modellparameter definieren mit ARIMA(p,d,q)
2. Modell wird mit den Trainingsdaten gefittet mit der fit() Funktion
3. Vorhersagen für einen entsprechenden Zeitraum können mit der predict() Funktion gemacht werden

Das ARIMA Modell wird zunächst auf den gesamten Shampoo Datensatz gefittet, anschließend die Fehler der Residuen betrachtet.

Als erster Ansatz: ARIMA(5,1,0), Lag auf den Wert 5 für den AR Teil, eine Differenzierungsorder von 1 um die Zeitreihe stationär zu machen und einen Anteil von MA von 0. Zu beachten gilt, dass normalerweise der Fit Datensatz nicht 100% des gesamten Datensatzes entsprechen sollte, sondern nur 70-80%!

Die Schwierigkeit liegt im Folgenden bei der korrekten Wahl der einzelnen Parameter p,d,q.

Die optimalen Werte zu erhalten, wird oft mittel des Box-Jenkins Ansatzes erreicht.

Dieser Ansatz nutzt Zeitreihenanalysen, um die besten Parameter ausfindig zu machen.

Vorgehensweise des Modells:

1. Modell Identifikation: Plotten und Statistikübersichten nutzen um Trends, Saisoneffekte, Autoregressionen festzustellen, um eine Vorstellung zu erlangen.
2. Parameterschätzung: Mit Fitting Methoden die besten Koeffizienten des Regressionsmodells finden.
3. Modell Überprüfung: Plotten und Statistikübersichten nutzen um die Fehler der Residuen zu bestimmen um den Umfang/ Typ der zeitlichen Struktur zu erfassen, der nicht vom Modell berücksichtig wird.

Diese Schritte werden solange wiederholt, bis die besten Parameter für z.B. ein Fitting Level gefunden sind. Das Modell kann auf moderat große Datensätze effizient angepasst werden, dabei hilfreich kann die Grid Search sein, die eine eigene Methode darstellt.

**Quellen**

1. <https://www.datacamp.com/community/tutorials/moving-averages-in-pandas>
2. <https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/>
3. <https://www.bigdata-insider.de/was-ist-das-arima-modell-a-914956/>
4. <https://www.statistik-nachhilfe.de/ratgeber/statistik/induktive-statistik/statistische-modellbildung-und-weitere-methoden/zeitreihenanalyse/arima-modelle>
5. <http://www.statistics4u.com/fundstat_germ/cc_timeser_arima.html>
6. <http://www.reiter1.com/Glossar/ARIMA.htm>