**Theory**

**Regression model:**

<https://webfocusinfocenter.informationbuilders.com/wfappent/TLs/TL_rstat/source/LinearRegression41.htm>

Regression analysis is the method of using observations (data records) **to quantify** the relationship between a target variable (a field in the record set), also referred to as a dependent variable, and a set of independent variables, also referred to as a covariate.

For example, regression analysis can be used to determine whether the dollar value of grocery shopping baskets (the target variable) is different for male and female shoppers (gender being the independent variable). The regression equation estimates a coefficient for each gender that corresponds to the difference in value.

 know only the values of the independent variables in order to be able to make predictions about the value of the dependent variable. For example, when the coefficients for male and female shoppers are known, you can make precise revenue estimates for different distributions of shoppers. Specifically, you can predict the revenues for 80/20 females/males versus 60/40 females/males.

Example

*y* = β0 + β1 *x* + *error*

* Y = dependent variable (price of wine at the auction)
* X = independent variable (#years to sell the wine)
* β0 and β1 are parameters that are unknown and will be estimated by the equation
* β0 is a constant that defines where the linear trend line intercepts the Y-axis
* β1  is a constant that represents the rate of change in the dependent variable as a function of changes in the independent variable (in this case it says how much a single year influence the wine cost)
* Error representsthe unexplained variation in the target variable. It is treated as a random variable that picks up all the variation in Y that is not explained by X.

Regression analysis is used for:

Regression is used most frequently for prediction. Credit scoring applications use input variables (data collected on the applicant) to predict the likelihood that the applicant will repay the loan. In budgeting and finance, regression is used to estimate the relationship between profit and cost, which later can be used in applications to generate what if scenarios.

How Does Multiple Linear Regression Work?

Multiple linear regression enables you to add additional variables to improve the predictive power of the regression equation. On a very intuitive level, the producer of the wine matters (chateau). Thus, including the chateau as another independent variable is likely to increase the predictive power of the equation

y = β0 + β1x1 + β2x2 + ... + βnxn + error

Price = - 4744 + 1.03 \* WR + 295.84 \* T - 2.09 \* HR + 8.06 \* TSV + Chateau

Independent variables:

WR/HR/TSV (all numeric variables)

Chateau (categorical value: this is the name of the chateau where the wine was produced. A categorical value can assume only some fixed values)

Output from a linear regression:

Distribution of the Residuals**.**

 This shows the distribution of the residuals. The residuals are the difference between the prices in the training data set and the predicted prices by this model. A negative residual is an overestimate and a positive residual is an underestimate. Ideally, you should see a symmetrical distribution with a median near zero. In this case, the median of -9.47 is very close to zero.

Coefficients:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5191.3914 1405.3794 -3.694 0.000820 \*\*\*

WRAIN 0.9171 0.3502 2.619 0.013373 \*

Degrees\_in\_C 329.9655 83.0036 3.975 0.000375 \*\*\*

HRAIN -1.8965 0.4753 -3.990 0.000360 \*\*\*

Time\_Since\_Vintage 7.3211 13.6187 0.538 0.594587

ChateauCos d'Estournel -336.9234 130.7021 -2.578 0.014753 \*

ChateauLafite -70.7367 128.0801 -0.552 0.584590

ChateauLatour -27.4525 136.1005 -0.202 0.841422

ChateauMontrose -365.6623 140.1642 -2.609 0.013698 \*

ChateauPichon Lalande -283.3606 120.9652 -2.342 0.025538 \*

---

Coefficient Name:

Column 1 displays the names of the coefficients. Notice that for categorical variables, all values except the reference value are listed. For example, five out of six chateaus are listed.

Estimate:

 These are the estimated values for the coefficients. Except for the reference chateau, notice the separate coefficient for each unique value of the categorical variable. The displayed coefficients are not standardized, for example, they are measured in their natural units, and thus cannot be compared with one another to determine which one is more influential in the model. Their natural units can be measured on different scales, as are temperature and rain. (in ARIMA we can compare them with each other because we use autoregression, so the coefficients are about the same independent values, thus exogenous variable are the action unit that we don’t need to compare with the weights of the lags)

Standard Error:

These are the standard errors of the coefficients. They can be used to construct the lower and upper bounds for the coefficient. An example is **Coefficient ± Standard Error**, which provides an indication where the value may fall if another sample data set is used. The standard error is also used to test whether the parameter is significantly different from 0. If a coefficient is significantly different from 0, then it has impact on the dependent variable. (t-value)

T-value:

The t-value is the ratio of the regression coefficient β to its standard error (t = coefficient ÷ standard error). The t statistic tests the hypothesis that a population regression coefficient is 0. If a coefficient is different from zero, then it has a genuine effect on the dependent variable. However, a coefficient may be different from zero, but if the difference is due to random variation, then it has no impact on the dependent variable. In this example, Time\_Since\_Vintage is different from zero due to random variation and thus has no real impact on the dependent variable. The t-values are used to determine the P values (see below).

Pr(>|t|):

The P value indicates whether the independent variable has statistically significant predictive capability. It essentially shows the probability of the coefficient being attributed to random variation. The lower the probability, the more significant the impact of the coefficient. For example, there is less than a 1.3% chance that the WRAIN impact is due to random variation. The P value is automatically calculated by R by comparing the t-value against the Student's T distribution table. As a rule, a P value of less than 5% indicates significance. In theory, the P value for the constant could be used to determine whether the constant could be removed from the model.

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The significance codes indicate how certain we can be that the coefficient has an impact on the dependent variable. For example, a significance level of 0.01 indicates that there is less than a 0.1% chance that the coefficient might be equal to 0 and thus be insignificant. Stated differently, we can be 99.9% sure that it is significant. The significance codes (shown by asterisks) are intended for quickly ranking the significance of each variable.

Es:

pvalue>t for ChateauLatour: 0.841422 🡪 there is less than 1% (‘ ‘ = 0.1)chance that the coeff might be equal to zero and so insignificance (99% is influencing at that t-value).

**TO WORK WITH ARIMA, we need to know about the next stuff here.**

<https://www.quantstart.com/articles/Autoregressive-Moving-Average-ARMA-p-q-Models-for-Time-Series-Analysis-Part-1/>

<https://www.quantstart.com/articles/Autoregressive-Moving-Average-ARMA-p-q-Models-for-Time-Series-Analysis-Part-2/>

<https://www.quantstart.com/articles/Autoregressive-Moving-Average-ARMA-p-q-Models-for-Time-Series-Analysis-Part-3/>

**Autoregression (serial correlation)**

Autoregression is a time series model that uses observations from previous time steps as input to a regression equation to predict the value at the next time step.

It is a very simple idea that can result in accurate forecasts on a range of time series problems.

ARIMA is an autoregression model.

X(t+1) = b0 + b1\*X(t-1) + b2\*X(t-2)

In this case, b1 and b2 are coefficients for lags in the same series

the statsmodels library provides an autoregression model that automatically selects an appropriate lag value using statistical tests and trains a linear regression model. It is provided in the AR class.

**Autocorrelation**

Autocorrelation is a mathematical representation of the degree of similarity between a given [time series](https://www.investopedia.com/terms/t/timeseries.asp) and a lagged version of itself over successive time intervals. It is the same as calculating the correlation between two different time series, except autocorrelation uses the same time series twice: once in its original form and once lagged one or more time periods.

When computing autocorrelation, the resulting output can range from 1 to negative 1, in line with the traditional correlation statistic. An autocorrelation of +1 represents a perfect positive correlation, while an autocorrelation of negative 1 represents a perfect negative correlation.

For autoregression models:

The stronger the correlation between the output variable and a specific lagged variable, the more weight the autoregression model can put on that variable when modeling.

Example:

Autocorrelation can show if there is a momentum factor associated with a stock. For example, if investors know that a stock has a historically high positive autocorrelation value and they witness it making sizable gains over the past several days, then they might reasonably expect the movements over the upcoming several days (the leading time series) to match those of the lagging time series and to move upward.

**Statistical stationarity:**A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. Most statistical forecasting methods are based on the assumption that the time series can be rendered approximately stationary (i.e., "stationarized") through the use of mathematical transformations. A stationarized series is relatively easy to predict: you simply predict that its statistical properties will be the same in the future as they have been in the past. The predictions for the stationarized series can then be "untransformed," by reversing whatever mathematical transformations were previously used, to obtain predictions for the original series. (The details are normally taken care of by your software.) Thus, finding the sequence of transformations needed to stationarize a time series often provides important clues in the search for an appropriate forecasting model.  Stationarizing a time series through differencing (where needed) is an important part of the process of fitting an ARIMA model. Another reason for trying to stationarize a time series is to be able to obtain meaningful sample statistics such as means, variances, and correlations with other variables. Such statistics are useful as descriptors of future behavior only if the series is stationary. For example, if the series is consistently increasing over time, the sample mean and variance will grow with the size of the sample, and they will always underestimate the mean and variance in future periods. And if the mean and variance of a series are not well-defined, then neither are its correlations with other variables. **For this reason you should be cautious about trying to extrapolate regression models fitted to nonstationary data.**

**Moving Average (MA)**

A moving average (MA) is a widely used indicator in technical analysis that helps smooth out values filtering out the “noise” from random short-term fluctuations.

Moving average is a trend-following, or [lagging](https://www.investopedia.com/terms/l/laggingindicator.asp), indicator because it is based on past prices. The most common applications of moving averages are:

* to identify the trend direction
* to determine support and resistance levels.

**When are you happy with a model?**

When the remaining time seeries lacks any serial correlation.

**Residual Error Series**

The *residual error series* or *residuals*, xt, is a time series of the difference between an observed value and a predicted value, from a time series model, at a particular time t.

This means that each element of the serially uncorrelated residual series is an independent realisation from some probability distribution. That is, the residuals themselves are independent and identically distributed (i.i.d.).

If we are to begin creating time series models that explain away any serial correlation, it seems natural to begin with a process that produces independent random variables from some distribution. This directly leads on to the concept of (discrete) white noise.

Consider a time series {wt:t=1,...n}. If the elements of the series, wi, are independent and identically distributed (i.i.d.), with a mean of zero, variance σ2 and no serial correlation (i.e. Cor(wi,wj)≠0,∀i≠j) then we say that the time series is discrete white noise (DWN) (no error).

**Random Walk**

A *random walk* is a time series model xt such that xt=xt−1+wt, where wt is a discrete white noise series.

**Heteroskedasticity (Potrebbe non servire)**

In statistics, heteroskedasticity (or heteroscedasticity) happens when the standard errors of a variable, monitored over a specific amount of time, are non-constant.

**Akaike Information Criterion:**

If we take the [likelihood function](https://en.wikipedia.org/wiki/Likelihood_function) for a statistical model, which has k parameters, and L [maximizes the likelihood](https://en.wikipedia.org/wiki/Maximum_likelihood), then the Akaike Information Criterion is given by:  given by:

AIC=−2log(L)+2k

The preferred model, from a selection of models, has the minimum AIC of the group. You can see that the AIC grows as the number of parameters, k, increases, but is reduced if the negative [log-likelihood](https://en.wikipedia.org/wiki/Likelihood_function#Log-likelihood) increases. Essentially it penalizes models that are overfit.

We are going to be creating AR, MA and ARMA models of varying orders and one way to choose the "best" model fit a dataset is to use the AIC.

**Backward Shift Operator**

The backward shift operator or lag operator, B, takes a time series element as an argument and returns the element one-time unit previously: . Repeated application of the operator allows us to step back n times: 

**Autoregressive Model of order p**

A time series model, {xt}, is an autoregressive model of order p, AR(p), if:

|  |
| --- |
|  |

Where {wt} is white noise and αi∈R, with αp≠0 for a p-order autoregressive process.

If we consider the Backward Shift Operator, B then we can rewrite the above as a function θ of B:

|  |
| --- |
|  |

the autogressive model is an extension of the random walk, so this makes sense

### **Stationarity for Autoregressive**

One of the most important aspects of the AR(p) model is that it is not always stationary. Indeed, the stationarity of a model depends upon the parameters.

In order to determine whether an AR(p) process is stationary or not we need to solve the *characteristic equation*. The characteristic equation is simply the autoregressive model, written in backward shift form, set to zero:

|  |
| --- |
|  |

We solve this equation for B. In order for the particular autoregressive process to be stationary we need all of the absolute values of the roots of this equation to exceed unity. This is an extremely useful property and allows us to quickly calculate whether an AR(p) process is stationary or not.

Let's consider a few examples to make this idea concrete:

* **Random Walk** - The AR(1) process with α1=1 has the characteristic equation θ=1−B. Clearly this has root B=1 and as such is *not* stationary.
* **AR(1)** - If we choose α1=14 we get xt=14xt−1+wt. This gives us a characteristic equation of 1−14B=0, which has a root B=4>1 and so this particular AR(1) process is stationary.
* **AR(2)** - If we set α1=α2=12 then we get xt=12xt−1+12xt−2+wt. Its characteristic equation becomes −12(B−1)(B+2)=0, which gives two roots of B=1,−2. Since this has a unit root it is a non-stationary series. However, other AR (2) series can be stationary.

**Augmented Dickey-Fuller test:**

<https://machinelearningmastery.com/time-series-data-stationary-python/>

We can use ADF as method to see if a series is stationary.  It can provide a quick check and confirmatory evidence that your time series is stationary or non-stationary.

This test is a type of statistical test called Unit Root test. ADF uses an autoregressive model and optimizes an information criterion across multiple different lag values. The null hypothesis of the test is that the time series can be represented by a unit root, that it is not stationary (has some time-dependent structure). The alternate hypothesis (rejecting the null hypothesis) is that the time series is stationary.

* **Null Hypothesis (H0):** If satisfied, it suggests the time series has a unit root, meaning it is non-stationary. It has some time dependent structure.
* **Alternate Hypothesis (H1):** The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary. It does not have time-dependent structure.

We interpret this result using the p-value from the test. A p-value below a threshold (such as 5% or 1%) suggests we reject the null hypothesis (stationary), otherwise a p-value above the threshold suggests we fail to reject the null hypothesis (non-stationary).

p-value > 0.05: Fail to reject the null hypothesis (H0), the data has a unit root and is non-stationary.

p-value <= 0.05: Reject the null hypothesis (H0), the data does not have a unit root and is stationary.

**Moving Average Model of order q**

A Moving Average model is similar to an Autoregressive model, except that instead of being a linear combination of past time series values, it is a linear combination of the past white noise terms. Intuitively, this means that the MA model sees such random white noise "shocks" directly at each current value of the model.

A time series model, {xt}, is a moving average model of order q, MA(q), if:

|  |
| --- |
|  |

Where {wt} is [white noise](https://www.quantstart.com/articles/White-Noise-and-Random-Walks-in-Time-Series-Analysis) with E(wt)=0 and variance σ2.

If we consider the Backward Shift Operator, B then we can rewrite the above as a function ϕ of B:

|  |
| --- |
|  |

In our ARIMA models we use q=0, that means we are not considering the white noise for previous values, only for the current lag.

**Autoregressive (integrated) Moving Average (AR(I)MA):**

ARMA(p,q):

* p is the number of lags considered in the model to make a new prediction
* q

ARIMA(p,d,q) is an extension of ARMA(p,q) that allows to approximate time series in a stationary series. This is done by defining a series made by the differences of the original.

If d=1:

Differencing 🡪 deltaXt = xt- xt-1;

**Choosing the Best ARMA(p,q) Model**

In order to determine which order p,q of the ARMA model is appropriate for a series, we need to use the AIC (or BIC) across a subset of values for p,q, and then apply the Ljung-Box test to determine if a good fit has been achieved, for particular values of p,q. in our case p={3,5,7,9}

To show this method we are going to firstly simulate a particular ARMA(p,q) process. We will then loop over all pairwise values of p∈{0,1,2,3,4} and q∈{0,1,2,3,4} and calculate the AIC. We will select the model with the lowest AIC and then run a Ljung-Box test on the residuals to determine if we have achieved a good fit.

* Pick model with the lowest AIC (to see best in the subset)
* ~~Ljung-Box test on the residuals (diff series) to see if we have achieved a good fit~~
* Consider we have validation too for this part

We’ll return on Ljung Box test if needed.

Notice that the p-value must be greater than 0.05, this means the residuals *are* independent at the 95% level and thus an ARMA(p,q) model provides a good model fit.

**~~Ljung-Box test~~**~~(NO)~~

The Ljung-Box test is a classical hypothesis test that is designed to test whether a set of autocorrelations of a fitted time series model differ significantly from zero. The test does *not* test each individual lag for randomness, but rather tests the randomness over a group of lags.

We define the null hypothesis H0 as: The time series data at each lag are [i.i.d.](https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables), that is, the correlations between the population series values are zero.

We define the alternate hypothesis Ha as: The time series data are not i.i.d. and possess serial correlation. We calculate the following [test statistic](https://en.wikipedia.org/wiki/Test_statistic), Q:

|  |
| --- |
|  |

Where n is the length of the time series sample, ρ^k is the sample autocorrelation at lag k and h is the number of lags under the test.

The decision rule as to whether to reject the null hypothesis H0 is to check whether , for a [chi-squared distribution](https://en.wikipedia.org/wiki/Chi-squared_distribution) with h degrees of freedom at the 100(1−α)th percentile.

**Using PACF for defining p,q for the ARIMA model**

<https://machinelearningmastery.com/gentle-introduction-autocorrelation-partial-autocorrelation/>

Autocorrelation and partial autocorrelation plots are heavily used in time series analysis and forecasting.

**ACF**

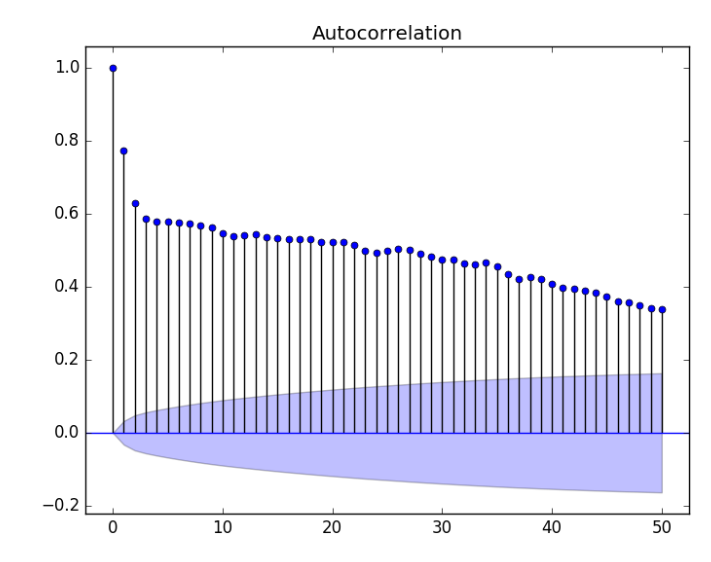
These are plots that graphically summarize the strength of a relationship with an observation in a time series with observations at prior time step.

A plot of the autocorrelation of a time series by lag is called the **A**uto**C**orrelation **F**unction, or the acronym ACF. This plot is sometimes called a correlogram or an autocorrelation plot.

2D plot showing the lag value along the x-axis and the correlation on the y-axis between -1 and 1.

Confidence intervals are drawn as a cone. By default, this is set to a 95% confidence interval, suggesting that correlation values outside of this code are very likely a correlation and not random values.

We can limit the number of lags on the x-axis to 50 to make the plot easier to read.



**PACF**

The partial autocorrelation at lag k is the correlation that results after removing the effect of any correlations due to the terms at shorter lags.

We know that the ACF describes the autocorrelation between an observation and another observation at a prior time step that includes direct and indirect dependence information.

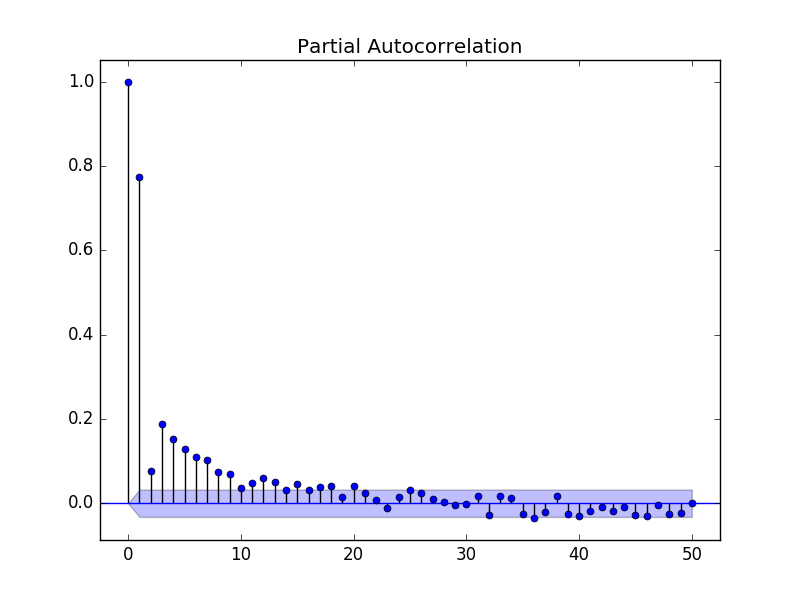
This means we would expect the ACF for the AR(k) time series to be strong to a lag of k and the inertia of that relationship would carry on to subsequent lag values, trailing off at some point as the effect was weakened.

The autocorrelation for an observation and an observation at a prior time step is comprised of both the direct correlation and indirect correlations. These indirect correlations are a linear function of the correlation of the observation, with observations at intervening time steps.

It is these indirect correlations that the partial autocorrelation function seeks to remove.

We know that the PACF only describes the direct relationship between an observation and its lag. This would suggest that there would be no correlation for lag values beyond *k*.

This is exactly the expectation of the ACF and PACF plots for an AR(k) process.



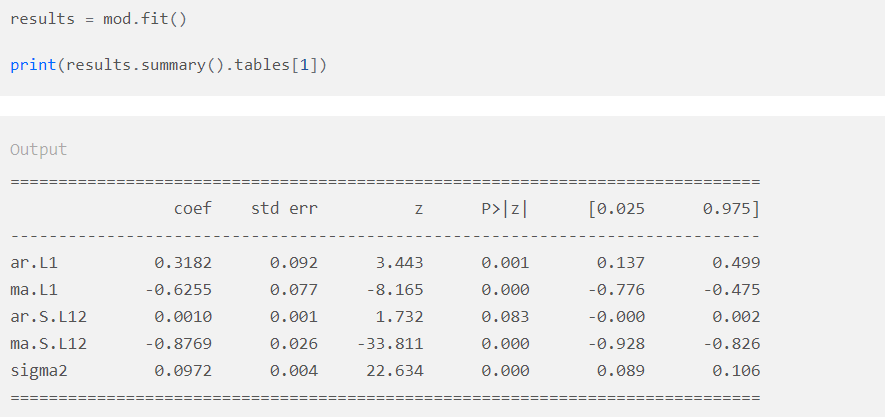
In the project we use PACF to find a subset of #lags to be used in the ARIMA model

ARIMA(p,1,0) where p = #lags

p we picked: [3,4,5,6,7,8,9]

Understand ARIMA result (summary)

<https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-arima-in-python-3>



The coef column shows the weight (i.e. importance) of each feature and how each one impacts the time series.

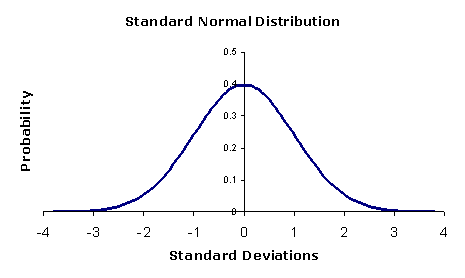
What is Z?

The P>|z| column informs us of the significance of each feature weight. Here, each weight has a p-value lower or close to 0.05, so it is reasonable to retain all of them in our mode.

|  |
| --- |
| What is a Z score What is a p-value |

Most statistical tests begin by identifying a null hypothesis. The null hypothesis for pattern analysis tools essentially states that there is no spatial pattern among the features, or among the values associated with the features, in the study area -- said another way: the expected pattern is just one of the many possible versions of complete spatial randomness. The Z score is a test of statistical significance that helps you decide whether or not to reject the null hypothesis. The p-value is the probability that you have falsely rejected the null hypothesis.

Z scores are measures of standard deviation. For example, if a tool returns a Z score of +2.5 it is interpreted as "+2.5 standard deviations away from the mean". P-values are probabilities. Both statistics are associated with the standard normal distribution. This distribution relates standard deviations with probabilities and allows significance and confidence to be attached to Z scores and p-values.



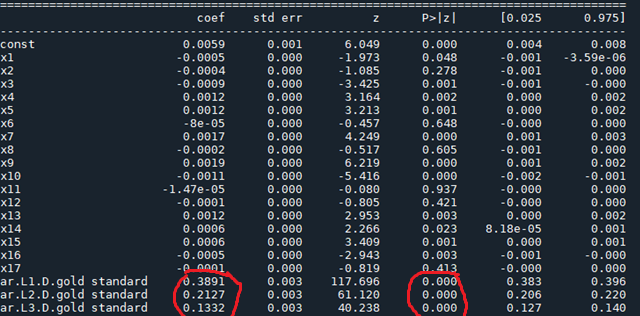
Very high or a very low (negative) Z scores, associated with very small p-values, are found in the tails of the normal distribution. When you perform a feature pattern analysis and it yields small p-values and either a very high or a very low (negative) Z score, this indicates it is very UNLIKELY that the observed pattern is some version of the theoretical spatial random pattern represented by your null hypothesis.

In order to reject the null hypothesis, you must make a subjective judgment regarding the degree of risk you are willing to accept for being wrong. This degree of risk is often given in terms of critical values and/or confidence levels.

To give an example: the critical Z score values when using a 95% confidence level are -1.96 and +1.96 standard deviations. The p-value associated with a 95% confidence level is 0.05. If your Z score is between -1.96 and +1.96, your p-value will be larger than 0.05, and you cannot reject your null hypothsis; the pattern exhibited is a pattern that could very likely be one version of a random pattern. If the Z score falls outside that range (for example -2.5 or +5.4), the pattern exhibited is probably too unusual to be just another version of random chance and the p-value will be small to reflect this. In this case, it is possible to reject the null hypothesis and proceed with figuring out what might be causing the statistically significant spatial pattern.

A key idea here is that the values in the middle of the normal distribution (Z scores like 0.19 or -1.2, for example), represent the expected outcome (the norm ...generally uninteresting). When the absolute value of the Z score is large (in the tails of the normal distribution) and the probabilities are small, you are seeing something unusual and generally very interesting. For the [Hot Spot Analysis](http://resources.esri.com/help/9.3/arcgisengine/java/gp_toolref/spatial_statistics_tools/hot_spot_analysis_getis_ord_gi_star_spatial_statistics_.htm) tool, for example, "unusual" means either a statistically significant hot spot or a statistically significant cold spot.

In our Example we had ARIMA(3,1,0) + exogen (see later)



Lags ar1 ar2, ar3 have all p>|Z| < p-value (0.05) so they are significative.

ARIMARESULTS.arparam returns the weight (coeff) of the lags , so we can avoid to apply ARIMA again on validation test .

**AR(I)MAX:**



It extends AR(I)MA model by including the linear effect that one or more exogenous series has on the [stationary](https://www.mathworks.com/help/econ/stationary-stochastic-process.html#btdw3g1) response series yt.

Exogenous series: Action units series

**Coefficent extractions:**

For the Aus I’ve reassigned the correct index set on the cvs in the series

