Bachelor 2015

Alisa Dammer

January 14, 2015

1. Introduction:
    - 1.1 Motivation
    - 1.2 Overview of the existing approaches, models and programs
    - 1.3 Sentiment data analysis
2. The model
    - 2.0 Explanation of the Thesis of the Bachelor
        - 2.0.1 Theoretical background and the problem statement
        - 2.0.2 The program broken into several steps
            - 2.0.2.1 Data preparation
            - 2.0.2.2 Model fine-tuning
            - 2.0.2.3 Checking the hypothesis for specified limitations
    - 2.1 Data preparation (correlation matrices and further filtering)
    - 2.2 Model specification and limitations
        - 2.2.1 Specification of the quality and quantity of the parameters
        - 2.2.2 Iterative fine-tuning of the model
        - 2.2.3 Error check
    - 2.3 Second data processing (sentimental analysis for the found indexes)
    - 2.4 Integration of the sub-results into final result
    - 2.5 Testing of the result
3. Conclusion
    - 3.1 The interpretation of the results
    - 3.2 Bottle neck and other problems of the model
    - 3.3 The potential and further use
    - 3.4 Possible improvements
4. References

1 Introduction

1.1 Motivation

Big Data
Data mining
Sentiment data analysis

1.2 An overview of the existing approaches, models and programs

classical econometric - regressions
Among the classical econometric approaches two main directions can be called: linear and non-linear regressions. First of all we will consider linear regressions and discuss their advantages and disadvantages.
According to the most common definition, in statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and explanatory variable denoted X [1]. There are one-variable regressions called simple linear regressions and multiple regressions, which shows the dependencies of the dependent variable from more than one explanatory variables. The regression equation looks following way:

$$Y = a + bX + \varepsilon$$

where Y - dependent variable, X - explanatory variable, a - the intercept of regression line, b - the slope of the regression line, ε - the error term.
The simple linear regression is based on 6 strict assumptions:

1. Because of the linear form neither parameter a, nor parameter b may have higher power than one.

2. The independent variable X is not random.

3. The variance is constant. Which means that the expected error value is 0.

4. The variance of ε is constant for all observations.

5. Errors are not correlated.

6. The error's distribution is normal.

On one hand the simple linear regression has several disadvantages: the form of the equation (here number of the predicting variables is concerned) and underlying assumptions. The dependency from only one variable can not describe fully all the relationship on the market. Even if you consider an equations system, consisting from linear regressions with different variables, it will lead to unnecessary complexity and not realistic or complete dependencies between variables. The second disadvantage is the assumptions - they are unrealistic in modern

world and are appropriate only for very simplified model of he world.
The big advantage of the simple linear regression on the other hand is it's estimators: ordinary lest squares, generalized least squares, instrumental variables, maximum likelihood (ML) and various ML techniques - these are the most known and not too complex approaches to estimate the regression. There also many other different techniques, but we won't discuss them here. An example of a simple linear regression could be dependency between the harvest volume and amount of rain during the season. Let's say Y = amount of harvest (potato for example) in kg and X is amount of the precipitation in mm
Than the more rain will drop during the season the bigger amount of potato will be harvested at the end. The slope of the regression line is positive. And the intercept is also more than 0, because you can't dig out less, than you have planted. So, the regression will look like:

$$amount\,of\,harvest = |a| + |b| * amount\,of\,rain\epsilon$$

Since it is likely to be impossible to build a proper prediction for the variable only with one predicting variable, multiple linear regression was invented. As you can see from the name, multiple linear regression describes relationship between more than one explanatory variables and dependent variable. The form of the equation looks as following:

$$Y = a + b_1X_1 + ... + b_nX_n + \epsilon,\ where\ n \in N$$

b_i - the "contribution" of the i-th variable to the regression, a -the intercept of the regression line, Y - dependant variable, X - explanatory variables.
Unlike simple linear regression, multiple linear regression does not have the form-disadvantage. However, the linear nature of the regression provides the model with several assumptions:

1. All b_i have the power of 1 - linear dependency between Y and all Xi-th.

2. The residuals are distributed normally. (Residuals - the difference between predicted value and observed value).

3. Uncertain number of the explanatory variables X - there is no technique to choose exact number of the variables that will optimally predict the value Y.

4. Completely "substitutive" variables X - unnecessary big number of the variables without increasing the accuracy of the prediction.

5. Complex form of the variables (X can have a higher-order polynomial form) leads to reduction of transparency and wrong results.

The advantage of the multiple linear regression is that it can be tested several ways. First, you can test whether the regression has the linear character by using F-test or ANOVA table. Second, to check whether the model is good enough you

can use the same tests as for simple linear regression: determination coefficient R^2. You can also run several hypothesis-test for each explanatory variable (b_i coefficient is equal to 0 - the variable is insignificant).
As an example of the multiple linear regression we can take previous example and extend it. The amount of the precipitation is not the only factor for the harvest to grow. Our Y variable also depends on amount of sunny days (X_2) and average temperature during the season (X_3). The higher temperature, the better will be the harvest (in reality this is a non linear dependency; most likely the dependency here takes the form of ($c - d * \sqrt{x}$): the harvest will be big, if the temperature is between certain degrees). The higher number of the sunny days the less rainy days, but higher the temperature. This is an example where two explanatory variables are not complete substitutes, but they have negative dependency with each other, so, their dependency will be explicitly included into the regression. The general form of the equation will take following form:

$$harvest = |a| + |b_1| * rain + |b_2| * sun + |b_3| * temperature + |b_{12}| * sun * rain + \epsilon$$

So the multiple linear regression is better for real purposes (real market analysis) in comparison to simple linear regression, but it still contains restrictions that do not allow high accuracy prediction.
Unfortunately linear dependencies between all the parameters and variables in the equation is a rare thing in the real life. That is why nonlinear regressions took place in statistical analysis. The general form of the nonlinear regression looks as following:

$$Y = f(\beta, x'_i) + \epsilon'$$

where f() - nonlinear function, x'_i - vector of predictors, ϵ' - vector of parameters, $\epsilon_i \sim N(0, \sigma^2)$.
Nonlinear regressions can be divided into two main classes:

1. Nonlinear dependencies between explanatory variables (x_i) and dependent variable (Y), linear character of the parameters (β).

2. Linear dependencies between explanatory variables (x_i) and dependent variable (Y), nonlinear character or the parameters (β)

The first class is relatively easy to work with: you can use variables substitution and transform the original equation to the linear form; than use the same techniques as for the linear multiple regression and find estimates for the parameters. For example, you can use least squares method with additional steps. To use LS method you will first need to change the form of the variables to the linear form: $z_i = f^{-1}(x_i)$, where z is a new variable in a linear form. The equation will then take following form: $y = a + b_1 z_1 + ... b_n z_n$. As the next step your find the estimates using LS method. However, the estimates will be biased towards original dependencies, because we found them for variables z_i, which are the transformation of the original variables [non linear j]. So you need to do the last transformation of the estimates for original variables x_i.
The second class is more difficult to use. Because of the nonlinear nature of

the parameters you can not use normal estimation techniques. On one hand you can use the ordinary optimum finding algorithms, but you will run into different difficulties because of the character of the dependencies. There are exist different approaches to get the estimates for the regressions: Gauss-Newton method [nonlinear m], Levenberg-Marquardt method [nonlinear n] (we will not concentrate on these approaches in our work, for further information you can read sources from reference-list). You can also iteratively try to use LS method (iterative transformations), but with each iteration the estimates will get additional error, which can result in immense accuracy decrease.
As an example for a nonlinear regression we can take the earlier described dependency between harvest volume, amount of rain, number of sunny days and temperature.

$$harvest = a+b_1*ln(sun)+b_2*(c-d*(rain-e)^2)+b_3*f(temperature)+b_{12}*g(sun,rain)$$

where $f(n) = \begin{cases} 0,\ x < 10 \\ \alpha x,\ x \in [10,30] and \alpha > 1 \\ -\alpha x,\ x > 30 \end{cases}$; c, d and e > 0, b_i - reggresion coefficients here they are first-order parameters.
As you can see,the normal approach here is is almost impossible to use, moreover, here we see different form of dependencies at one time (normally people use uniform functions like polynomials, logarithms etc.) But this equation can be transformed into linear form via several steps of the variables substitution and transformations, then the LS will be found and the process of "backwards transformation" will be started, until we get the estimates for the original variables.

neural networks

The biggest disadvantage of the classical regression models on one hand is increasing complexity when using too many variables and parameters. Also, they are not flexible enough for a quickly changing and constantly growing and changing financial markets. On the other hand, when using too few parameters and variables the model looses a lot of information and flexibility. Too overcome these problems neural networks are used to make predictions and analyse the increasing amount of data. According to Wikipedia:"In machine learning, artificial neural networks (ANNs) are a family of statistical learning algorithms inspired by biological neural networks (the central nervous systems of animals, in particular the brain) and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which can compute values from inputs, and are capable of machine learning as well as pattern recognition thanks to their adaptive nature" [ANN a]. The ANNs ae used in different spheres where pattern recognition can be used. Only the use of NNs on financial markets is relevant for this work, so we will only consider appropriate forms of the networks and learning-mechanisms.
The basic structure of the ANN consists of 3 layers: input, hidden and output layer. As follows from the name, the bottom layer is a input layer, that sends all the data (in our case prices, time series) to the next level via "synopses".

Synopses in ANNs are the dependencies between input units and the units of the hidden layer (a weight-function). For example,

$$y_j = \sum_{i \in I} w_{ij} x_i$$

where y_j - an unit in the hidden layer, w_{ij} -weight for input unit i in jth hidden unit (influence if the input node on the hidden node), and x_i - input unit.
Next dependeny-level is the transformation the hidden layers into output. (This dependency might look the same as the previous one). As an example of the model following form can be used:

$$z = f(y)$$

where z - an output unit, f - a function that transforms all hidden units y_j into the output (y_j is equals to the sum of the weighted input units, see above), here y is a vector of y_j, $y = (y_1, ..., y_n)$.
Not only predefined functions are important in building an ANN, but also the way the network will be trained. The advantage of the ANNs is the possibility to adapt (search for some sub-optimal transformation function $f^*(y)$). In literature you will find the definition of "cost function". Minimum of the cost function (normally this function depends on the observations or input data,for example, the coefficient of the determination can be used as a cost function) will signal, that our Network has reached some local optimum. (it is hard to find the global optimum and moreover to prove, that found one is indeed the global optimum). The learning process can be different, depending on the goal. However, the main steps are the same:

1. The whole data set is divided into several chunks: training set, testing set and validation set.

2. The training data set is properly processed before the training (learning) phase starts. The data needs to be cleared out of noise and additional shifts for the network to get the pattern properly.

3. After the training on the train set, the cost function will be compared with a desirable state. The parameters will be fine-tuned if need.

4. The network with fine-tuned parameters will be tested on the test-set. If needed this and previous steps will be done again.

5. The "trained" network will be validated on the validation test.

ANNs have several advantages and disadvantages, that make them not that easy to use. The main advantages are the flexibility and ability to work with immense amount of the data. By flexibility here we mean, that the network is able to recognize the pattern and its development, if there any change over time. By immense amount of data here we mean, that the model can contain incredible amount of input nodes, as well as several hidden layers, each containing huge number of nodes. Such structure allows to recognize different patterns at once and model the behaviour of the dependant variables very accurate.

On the other hand, the disadvantages are the run time, no opportunity to debug the network on the fly and initial parameters. Let's start with the initial parameters problem. The network should be given some assumptions about the data set, its behaviour and possible dependencies. These assumptions result in an initial weight matrices or parameters for the initial dependency functions. If the parameters were poorly estimated and contain serious errors, the network might even show the opposite behaviour to optimum finding. The structure of the NN must be defined before the training. Depending on the initial assumptions and the goal of the training, the structure may vary from Deep Boltzmann Machine (only visible and hidden units: one hidden layer - containing up to 10 nodes in total) to incredibly huge perceptron (5 hidden layers, each containing more than starting from 300 nodes). The bigger and complexer is the structure, the more parameters should be estimated in the beginning, the more time will the NN need to be trained. Another aspect of the complexity problem is the way the NN is going to be trained. Most famous approaches are bottom-up and top-down techniques. The learning method can be incorrectly used or wrongly assigned to the network, which will result in low accuracy prediction and pattern recognition. The second problem - the run time depends on the structure of the ANN, amount of the parameters and the size of the training data set. It can only increase with adding new parameters to the model. The third problem is that you won't know whether the model is good until it is validated. There is no possibility to stop the training process and debug the model. The process of the training is automatic for the network. If you will interrupt current training session and fine-tune the parameters manually, that will mean the new training session without any progress saved for the model.

However the the complexity of the ANNs is compensated through their ability to recognize the patters in changing world, their ability to adapt to changing market structure and new data. Unfortunately this is a complete overkill for this bachelor, and I am not really sure how to use it for my hypothesis anyhow, mostly because of the no possibility to debug it. Most time will be spend on waiting for the network to finish the training and testing stages. It can be later used as a combination of my hypothesis and "advanced forecasting in the next work

Generally speaking, all "self-developing" models can be referred to machine learning. According to Wikipedia:"Machine learning is a scientific discipline that explores the construction and study of algorithms that can learn from data. Such algorithms operate by building a model based on inputs and using that to make predictions or decisions, rather than following only explicitly programmed instructions" [machine learning a]. Machine learning defines several steps for ANNs for instance. The way you can teach your AN is first defined in machine learning field. Also the rules for different data sets are defined in this scientific sphere. Building a regression can also be viewed as a machine learning problem, especially if the regression is built for changing environment. Machine learning

can be used as for short- and midterm predictions and for strategical decision support (trend prediction and long-term prediction have shown better results in some papers). For this bachelor learning rule (supervised or unsupervised), data sets partitioning are important.
Not only models develop to increase the accuracy of prediction, but also tools and software for data collecting and analysing. The quality and amount of data increases exponentially and old approaches becoming unusable.
relatively new trend in data analysis is the sentimental data analysis. The big role for stock exchange plays the psychological component: the way investors predict and react on certain events, the way they collect and share information. Social media has become a source of important information for decision making in all short-, mid- and long-term. According to Wikipedia:"Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials" [sentiment a]. The sentimental analysis combines different approaches and models to detect the mood, key words (proper information) and the insensitivity of the discussion (if more than one source is analysed) or the how the object is important for the person. The sentiment data can not only give a prediction for financial tools on the markets, but also give a hint about the components of the whole behaviour. We know, that the price of the share, for example, reflects the real value of the firm, the business reputation (what people expect of the firm), noise and external extreme factors (not ordinary shifts, gaps and jumps). The sentimental data analysis combined with classical fundamental and technical analysis can result in high accuracy prediction for all time periods: also for portfolio strategies and simple trading.

2 The model

2.1 Explanation of the Thesis of the Bachelor

The Idea of the work is to combine classical technical analysis and sentimental analysis and use them indirectly. The first part is to choose the most appropriate regression for the dependant variable. The result of the first step is the names of the explanatory variables, we will need them for further work. The second step is to get predictions for each explanatory variable using sentimental data analysis. For each of the variables in the found optimum we will search for text data and analyse it for predictions, buzz (how frequent will be discussed the topic) and mood. The third step is to use predicted values of the explanatory variables in the original regression model to get prediction for the dependant variable. At the same time mood from the sentimental data and buzz will fine-tune the parameters in the original equation. As the result we will get new time series for testing (predictions from text analysis) and fine-tuned parameters. In simple words, the sentimental data analysis is used for fine-tuning the parameters and including psychological component to the strict regression model. To validate the final equation test set will be used: the predicted values of the dependant variables will be compared with the actual value. If the difference will stay within 75% of accuracy, the approach for fine-tuning the parameters of the equation can be considered efficient.
The whole methodology can be broken into two main algorithms, each algorithm can be broken into several steps and additional forward and backward connection-steps between the algorithms:

1. The data in form of the time-series for most appropriate indexes will be collected. The data will be divided into three chunks: training set, test set and validation set.

2. For the chosen equation and chosen dependant variable (regression model) iteratively the best combination of the explanatory variables will be chosen. The parameters estimated based on the training set.

3. For found dependant variables text sources containing relevant information will be collected and analysed. As the result we get the values of the explanatory variables in time-series form, buzz in numeric form and mood in form of $(-1)^n$, where n=0, if the mood is positive and n=1, if the mood is negative.

4. Based on the new time series the parameters of the model will be re-estimated and fine-tuned by mood and buzz (buzz in this case is time-dependant weight modifier as well as the mood).

5. New form of the regression will be again tested on the test set, and if needed, the parameters will be fine-tuned again.

6. After testing, we will get the forecast for certain period for the dependant variable and compare it with the actual value from the validation set.

7. Based on the difference between actual and predicted value the approach can be considered good or bad.

In simple words, the goal of this bachelor is to test the ability of the sentimental data analysis to make classical regression models more flexible and time-responsive.

2.2 Data preparation

Here independent of the main index (not chosen yet) based on correlation-matrices certain amount of indexes will be taken for further consideration. For example every index with correlation-vector bigger or equal to +-0.3
Some period of the time-series (for all indexes) will be left as test-sample. (let's say about a month-period).
All time series will be cleaned from the noise using following technics: mooving average, extracting a trend line on high frequency long-term data,

2.3 Model specification and limitations

First, the equation or the model description will be given, then the number for the parameters will be set (with all necessary explanations).
Second, the automatic optimum finding for given limitations will be maid. (python iterative optimum-finding test). At the end another target function, built from the found optimum will be tested (error check).
Third, texts for the chosen indexes will be analysed. The forecast, probability and maybe buzz will be the results of the sub-chapter.
Forth, the results of the previous stage will be united into one forecast and probability for the main index (index on the left side of the main equation).
Finally, the results of the model will be compared with the actual results that we left as the test-sample.

3 conclusion

3.1 Interpretation of the results

Here the difference of the actual value (test-sample) and our forecasts will be discussed and explained.

3.2 Problems

Here the biggest difficulties will be stated and also discussed the influence of the strong limitations.

3.3 Potential

Here most possible use and advantage of the model will be discussed.

3.4 Improvements

In order to be more useful the program can take several improvements...
Theoretical
Data
Technical

4 References

1. Wikipedia: linear regeression http://en.wikipedia.org/wiki/Linear_regression