**Statistical Analysis**

**Report**

*Objective of this document is to outline dataset and its variables, data cleaning methods, statistical tests on the cleaned dataset, and the applied linear regression on the cleaned datasets*.

Table of Contents

[1 Dataset 2](#_Toc40784977)

[2 Data Cleaning 2](#_Toc40784978)

[3 STATISTICAL TESTS 3](#_Toc40784979)

[4 Linear Regression 4](#_Toc40784980)

# Dataset

|  |  |
| --- | --- |
| Variable name | Type |
| symbol | Independent |
| exchange | Independent |
| date | Date, timeseries |
| adjusted.close | Independent |
| option.symbol | Independent |
| expiration | Independent |
| strike | Independent |
| call.put | Independent |
| style | Independent |
| ask | Independent |
| bid | Independent |
| mean.price | Independent |
| iv | Dependent |
| volume | Independent |
| open.interest | Independent |
| stock.price.for.iv | Independent |
| delta | Independent |
| vega | Independent |
| gamma | Independent |
| theta | Independent |
| rho | Independent |

# Data Cleaning

Following data cleaning methods was applied:

1. For each variable
2. If remove NA’s
3. For categorical variable, keep it as it is.
4. For numerical variable remove zero
5. For numerical variable the difference between minimum and maximum value is too high or the deviation or the standard deviation is high, then take natural log of that numerical variable with log base 10.
6. The natural log with base 10 has been applied in the following variable
   1. Strike
   2. Ask
   3. Bid
   4. Mean price
   5. Volume
   6. Open interest

# STATISTICAL TESTS

* Since we have dependent variable as continuous data (IV) and independent variable as categorical data (CALL/PUT), the most appropriate test is ANOVA test (Analysis of Variance).
* As only two groups (i.e. CALL and PUT) are present, we will use one- way ANOVA.
* ANOVA is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact by comparing the means of the two samples.
* It checks for the between group variability(i.e. b/w CALL and PUT) and within group variability and calculates the F-statistic which is :

**F = (Between group variability) / (within group variability)**

* This F-statistic calculated here is compared with the F-critical value for making a conclusion. If value of the calculated F-statistic is more than the F-critical value (for a specific α/significance level), then we reject the null hypothesis.
* In our both cases i.e. UVXY and VIX, the value of F-statistic is higher than critical F-value which rejects the hypothesis that the IV FOR CALL is equal to IV FOR PUT. And thus, the difference is significant and we conclude that IV FOR CALL is greater than IV FOR PUT for both UVXY and VIX.
* Also, the magnitude by which the F-statistic differs from F-critical value is higher for UVXY than for VIX. This shows that the magnitude by which IV for UVXY CALL is higher than IV for UVXY PUT, is more than the magnitude by which IV for VIX CALL is higher than IV for VIX PUT.

# Linear Regression

The linear regression algorithm was applied to find how the other independent variables are related to IV in case of UVXY-CALL, UVXY-PUT, VIX-CALL and VIX-PUT respectively:

* Whole data was separated based on symbol i.e. UVXY and VIX and further, divided into CALL and PUT. So, whole data in total is divided into four categories: UVXY+CALL, UVXY+PUT, VIX+CALL and VIX+PUT.
* Taking one category at a time, the dependent variable is the IV and the variables (or features) are adjusted close, strike, ask, bid, mean price, volume, open interest, stock price for iv, delta, vega, gamma, theta and rho ( i.e. total 13 independent variables).
* The above data is given as input in the linear regression model and the output returns the coefficients for the independent variables for the equation relating IV with the independent variables i.e.

**IV = c1.(adjusted close) + c2.(strike) + c3.(ask) + … + c13.(rho)**

* The above mentioned steps are continued for each category.
* The coefficients calculated from the model for each category i.e. UVXY+CALL, UVXY+PUT, VIX+CALL, VIX+PUT are given below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | UVXY + CALL | UVXY + PUT | VIX + CALL | VIX + PUT |
| ADJUSTED CLOSE | -0.038604 | 0.164835 | 0.108645 | 0.070383 |
| STRIKE | 0.000743 | -0.127589 | 0.004192 | 0.01253 |
| ASK | 0.016617 | 0.048304 | 0.018192 | -0.0409 |
| BID | 0.019488 | 0.048362 | -0.02875 | 0.041148 |
| MEAN PRICE | 0.018052 | 0.048333 | -0.00528 | 0.000123 |
| VOLUME | -0.000058 | -0.000032 | -5.94E-07 | 5.64E-07 |
| OPEN INTEREST | 0.000009 | 0.000004 | 7.94E-07 | -1.70E-07 |
| STOCK PRICE FOR IV | 0.009472 | -0.088813 | -0.21553 | -0.13867 |
| DELTA | -0.602352 | -1.325529 | -0.55156 | -0.45448 |
| VEGA | -1.385037 | 6.744883 | -2.02124 | 1.134561 |
| GAMMA | -2.761601 | -1.764737 | -3.92468 | -3.11999 |
| THETA | -7.850976 | -12.31616 | -7.18048 | -7.35674 |
| RHO | -4.057269 | 3.083372 | 7.948769 | 2.837582 |

IF LOGARITHM OF VALUES IS CONSIDERED

* If instead, logarithmic values are used i.e. log(strike), log(ask), log(bid), log(mean price), log(volume), log(open interest) are used to form the equation, then the coefficients calculated are as below :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | UVXY + CALL | UVXY + PUT | VIX + CALL | VIX + PUT |
| ADJUSTED CLOSE | -0.051941 | 0.127164 | 0.102455 | 0.066086 |
| LOG(STRIKE) | -0.105732 | 0.041800 | 0.274708 | 1.060693 |
| LOG(ASK) | 0.389162 | 0.222199 | 1.046761 | -0.218421 |
| LOG(BID) | -0.023051 | -0.203407 | 0.201574 | -0.072561 |
| LOG(MEAN PRICE) | -0.777595 | -0.458429 | -1.497761 | 0.415369 |
| LOG(VOLUME) | -0.036673 | --0.080025 | 0.003463 | -0.002177 |
| LOG(OPEN INTEREST) | 0.049987 | 0.045318 | 0.031958 | 0.003088 |
| STOCK PRICE FOR IV | 0.038389 | -0.169102 | -0.206779 | -0.149043 |
| DELTA | 0.342661 | -1.641948 | -0.082706 | 0.044176 |
| VEGA | 5.962487 | 19.887762 | 2.257163 | -1.808340 |
| GAMMA | -3.889316 | -4.552150 | -4.230678 | -2.945199 |
| THETA | -7.009917 | -11.679799 | -7.780115 | -7.433843 |
| RHO | -5.291196 | 1.641072 | 5.510388 | 2.445019 |