Cubist Systematic Data Exercise

Saurabh Gokhale

February 6, 2020

Introduction

- We're given the time-series of SPY Close Prices and a signal
- Data spans from Jan 3, 2012 to 29 Aug, 2014
- Our goal is to analyze the predictive power of this signal and any viability or shortcomings

Methodology

High-level data cleaning steps:

1. Found no missing data for Signal and ClosePrice
2. Found couple of entries for weekends (non-trading days). Removed those entries
3. Found few outliers in Signal as well as ClosePrice

Outliers were identified using following methods:

1. Z-score (comparatively not good)
2. Modified Z-score (works better)
3. InterQuartile (IQR) Range (works better)

Outliers identified by IQR method were replaced by first the missing entries and then interpolation was performed to obtain clean data.

Model Selection

1. Scatter plot of Signal and ClosePrice showed a clear linear relationship, so simple linear regression model was used. Signal was used as X (independent variable) and ClosePrice was used a Y (dependent variable)
2. Linear Regression Model equation is as follows:

$$ClosePrice = 1.6529 + 40.6862 * Signal \quad (1)$$

3. P-value of the Signal coefficient (see the next slide on Model Summary) is 0.00, which is less than significance level of 0.05, indicating the Signal is statistically significant in predicting ClosePrice

Model Summary

OLS Regression Results

Dep. Variable:	ClosePrice	R-squared:	0.990
Model:	OLS	Adj. R-squared:	0.990
Method:	Least Squares	F-statistic:	6.789e+04
Date:	Thu, 06 Feb 2020	Prob (F-statistic):	0.00
Time:	21:42:33	Log-Likelihood:	-1439.1
No. Observations:	665	AIC:	2882.
Df Residuals:	663	BIC:	2891.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.6529	0.617	2.679	0.008	0.442	2.864
Signal	40.6862	0.156	260.549	0.000	40.380	40.993

Omnibus:	32.061	Durbin-Watson:	0.643
Prob(Omnibus):	0.000	Jarque-Bera (JB):	53.880
Skew:	0.356	Prob(JB):	2.00e-12
Kurtosis:	4.199	Cond. No.	31.7

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Model Assumption Verification

1. **Mean of residuals should be 0:** Satisfied!
2. **Residuals are homoscedastic:** Not satisfied, residuals exhibit heteroscedasticity (tested via Breusch-Pagan test and Residual vs Fitted plot)
3. **Normality of Residuals:** Not satisfied, residuals do not follow Gaussian distribution (tested via Jarque-Bera test and QQ-plot)
4. **Residuals are serially uncorrelated:** Not satisfied, since low positive serial autocorrelation exists (tested via ACF plot and Durbin-Watson test)
5. **Independent variable (Signal) and residuals are uncorrelated:** Not Satisifed, low positive correlation exists (tested via p-value at 5% significance level)

Conclusion/Recommendation:

1. Despite **Signal** being a statistical significant predictor of **ClosePrice**, some of the assumptions of the linear model are violated and thus the model cannot be used as it is for ClosePrice prediction
2. Necessary steps must be taken to ensure assumptions are satisfied (adding lagged variables, transforming the features, differencing of feature values, taking log, using Generalized Least Squares etc) under a new linear model