Homework 3 of *INTRODUCTION TO ECONOMETRICS/ESSENTIALS OF REGRESSION ANALYSIS USING R/INTRODUCTORY ECONOMETRICS (GPP)*

Yoann Potiron
potiron@keio.jp
Keio University

Due: Wednesday 2025/07/16 in class

The dataset *SRE.csv* is financial high frequency data. By high-frequency data, we mean all the information recorded during a day of trading. It focuses on the asset Sempra Energy traded on the SP 500 on the date 2016/01/04. There are $n = 15127$ observations. The six variables are:

- *Price* is the price of the transaction in US dollars.
- *Priceop* is the price in US dollars of the best opposite ask or bid.
- *Trade* is the trade indicator, i.e. equal to 1 if the transaction is buyer initiated or equal to -1 if the transaction is seller initiated.
- *Time* is the transaction time in seconds.
- *Volume* is the volume of the transaction, in number of shares.
- *Depth* corresponds to the depth in the limit order book, in number of shares.

When using financial data, the price is nonstationary. If we make a linear regression directly on a nonstationary dataset, the model will not perform well, and we can expect a very low proportion of variance explained. Thus, the data requires a pre-process.

For any variable $V \in \{\text{Price}, \text{Priceop}, \text{Trade}, \text{Time}, \text{Volume}, \text{Depth}\}$, we denote its ith observation as V_i. We define the positive variable $\text{Spread}_i := \mid \text{Price}_i - \text{Priceop}_i \mid$ and the signed spread as $\text{SSpread}_i := \text{Trade}_i \times \text{Spread}_i$. We also define $\text{SVolume}_i := \text{Trade}_i \times \text{Volume}_i$ and $\text{SDepth}_i := \text{Trade}_i \times \text{Depth}_i$.

For any variable $V = (V_1, \cdots, V_n)$, we define the difference of V as $\Delta V := (V_2 - V_1, \cdots, V_n - V_{n-1})$. For example, if $V = (1, 2, 4)$, then $\Delta V = (2 - 1, 4 - 2)$, i.e. $\Delta V = (1, 2)$. The number of observations of V is n, while the number of observations of ΔV is $n - 1$. Finally, we define $\text{S}\Delta\text{Time}_i := \text{Trade}_{i+1} \times \Delta\text{Time}_i$ for $i = 1, \cdots, n - 1$.

We aim to use the following linear regression

$$\begin{aligned} \Delta\text{Price} &= \theta_0 + \theta_1 \Delta\text{Trade} + \theta_2 \Delta\text{SSpread} + \theta_3 \Delta(S\Delta\text{Time}) + \theta_4 \Delta\text{SVolume} \\ &\quad + \theta_5 \Delta\text{SDepth}. \end{aligned} \tag{1}$$

1. As in Homework 2, we also consider the linear regression incorporating *Time* as a sixth explaining variable, i.e.

$$\begin{aligned} \Delta\text{Price} &= \theta_0 + \theta_1 \Delta\text{Trade} + \theta_2 \Delta\text{SSpread} + \theta_3 \Delta(S\Delta\text{Time}) + \theta_4 \Delta\text{SVolume} \\ &\quad + \theta_5 \Delta\text{SDepth} + \theta_6 \text{Time}. \end{aligned} \tag{2}$$

   Implement an hypothesis test to compare models and see whether the fit of (2) significantly improves the fit of (1). Interpret the results of the test statistics.

2. Fit two new linear regressions, i.e.

$$\begin{aligned} \Delta\text{Price} &= \theta_0 + \theta_1 \Delta\text{Trade} + \theta_2 \Delta\text{SSpread} + \theta_3 \Delta(S\Delta\text{Time}) + \theta_4 \Delta\text{SVolume} \\ &\quad + \theta_5 \Delta\text{SDepth} + \theta_6 (\Delta\text{SSpread})^2 \end{aligned} \tag{3}$$

   and

$$\begin{aligned} \Delta\text{Price} &= \theta_0 + \theta_1 \Delta\text{Trade} + \theta_2 \Delta\text{SSpread} + \theta_3 \Delta(S\Delta\text{Time}) + \theta_4 \Delta\text{SVolume} \\ &\quad + \theta_5 \Delta\text{SDepth} + \theta_6 (\Delta\text{SSpread})^2 + \theta_7 (\Delta\text{SSpread})^3. \end{aligned} \tag{4}$$

Provide the summary tables, give an interpretation and implement an hypothesis test to compare regressions to see whether one or both regressions improve the fit of (1) or not.

3. Make one prediction of the explained variable ΔPrice based on the linear regression (1) when the explaining values are equal to ΔTrade = 2, ΔSSpread = 0, $\Delta(S\Delta\text{Time}) = 0.1$, ΔSVolume = 100, ΔSDepth = 100. Provide two different confidence intervals for each individual explaining variable, i.e. the confidence interval for "prediction of a future value" and "prediction of the mean response". Interpret about the results.

4. Do you think that the variance of the errors is constant over time? Illustrate with one or several plots and interpret the results. Implement a test of constant variance of the residuals in (1) that you find suitable.

5. Do you think that the the errors follow a normal distribution? Illustrate with one or several plots and interpret the results. Implement a test for normality of the errors in (1).

6. Do you think that the the errors are correlated with each over? Illustrate with one or several plots and interpret the results. Implement a test for correlation of errors in (1).

7. Do you think that the the linear model is a good model? Illustrate with one or several plots and interpret the results.

8. Discuss the implications of the results from questions 4 to 7 regarding the validity of the model, predictions and risk.