Homework 2 of *INTRODUCTION TO ECONOMETRICS/ESSENTIALS OF REGRESSION ANALYSIS USING R/INTRODUCTORY ECONOMETRICS (GPP)*

Yoann Potiron
Keio University

Due: Wednesday 2025/06/18 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 Pre-process of the data

1. Create $\Delta\text{Price}, \Delta\text{Trade}, \Delta\text{SSpread}, \Delta(S\Delta\text{Time}), \Delta\text{SVolume}, \Delta\text{SDepth}.$ by using the function diff carefully. As the variable $\Delta(S\Delta\text{Time})$ has $n-2$ observations whereas $\Delta\text{Price}, \Delta\text{Trade}, \Delta\text{SSpread}, \Delta\text{SVolume}, \Delta\text{SDepth}$ have $n-1$ observations, remove the first observation of $\Delta\text{Price}, \Delta\text{Trade}, \Delta\text{SSpread}, \Delta\text{SVolume}\Delta\text{SDepth}$. By doing this, all the variables in (1) have now $n-2$ observations.

2 Analysis

1. Fit the linear regression (1). Provide a summary table and a written explanation of the results.

2. Report the proportion of variance explained R^2 obtained for each individual regression:

$$\Delta\text{Price} = \theta_0 + \theta_1 \text{V}, \tag{2}$$

where $V \in \{\Delta\text{Trade}, \Delta\text{SSpread}, \Delta(S\Delta\text{Time}), \Delta\text{SVolume}, \Delta\text{SDepth}\}$. Give a written interpretation of the results.

3. Fit a new linear regression incorporating *Time* as a sixth regressor

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

As *Time* is of size n, remove its two first observations in the regression. Provide a summary table, and write an interpretation comparing the results of regression (1) and regression (3).