# Homework 2 of *Introductory Econometrics 1 - Linear Models with R*

Keio University

Due: Friday 2023/12/01 in class

The dataset *SRE.csv* is typically used in the field of statistics for financial high frequency data. It consists of all the information related to (intraday) transactions from the asset Sempra Energy traded on the SP 500 on the date 2016/01/04. The six variables are:

- *Price* corresponds to the price of the transaction (in US dollars).
- *Priceop* corresponds to the price (in US dollars) of the best opposite ask/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* corresponds to the transaction time (in second).
- *Volume* corresponds to the volume (in number of shares) of the transaction.
- *Depth* corresponds to the depth (in number of shares) in the limit order book.

For any variable $V \in \{\text{Price}, \text{Priceop}, \text{Trade}, \text{Time}, \text{Volume}, \text{Depth}\}$, we have $n = 15127$ (intraday) observations denoted $V_i$.

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 lagged differences 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. $\Delta V = (1, 2)$. Bear in mind that the size of $V$ is $n$ and $\Delta V$ 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 fit the following linear model:

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

# 1 Pre-process of the data

1. Create $\Delta\text{Price}, \Delta\text{Trade}, \Delta\text{SSpread}, \Delta(S\Delta\text{Time}), \Delta\text{SVolume}, \Delta\text{SDepth}$. *Hint: you will most likely find the function diff useful.* As the variable $\Delta(S\Delta\text{Time})$ is of size $n-2$ whereas $\Delta\text{Price}, \Delta\text{Trade}, \Delta\text{SSpread}, \Delta\text{SVolume}, \Delta\text{SDepth}$ are of size $n-1$, remove the first observation of $\Delta\text{Price}, \Delta\text{Trade}, \Delta\text{SSpread}, \Delta\text{SVolume}, \Delta\text{SDepth}$ so that all the variables in (1) are now of the same size $n-2$.

# 2 Analysis

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

2. Report the $R^2$ 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 model 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).