Title *

Tianyu Du †

Sunday 2nd February, 2020

Contents

*Compile Date: 17:26 Sunday 2nd February, 2020

†tianyu.du@mail.utoronto.ca

Table 2:

1 Missing Data

Because the focus is forecasting crude oil returns, which captures the difference between two consecutive prices, The missing data problem can be crucial here.

This paper calculates crude oil returns on one particular day t by taking the difference in logged prices at t and the previous trading day:

$$r_t := \ln(p_t) - \ln(p_{t-\Delta}) \tag{1.1}$$

where $t - \Delta$ is the last trading day before day t. As mentioned before, the time gap between two observed prices are not even. For instance, if t is a Monday, then r_t computes the crude oil return between the close price on Monday and the close price on Friday (with $\Delta = 3$). The value of Δ can be even larger with holidays.

Day of the week	Num. Days.	Num. Trading Days	Δ=1	2	3	4	5
Monday	1031	927	0	0	883	33	11
Tuesday	1030	1018	921	0	0	97	0
Wednesday	1030	1022	1011	5	0	0	6
Thursday	1030	1002	994	8	0	0	0
Friday	1030	986	969	17	0	0	0
Saturday	1030	0	0	0	0	0	0
Sunday	1030	0	0	0	0	0	0
Total	7211	4955	3895	30	883	130	17

Table 1: The values of Δ used to calculate returns. This table only include trading days, but the first day with price observation in this dataset was dropped because it did not have a previous trading day, so return on this day cannot by computed using our definition.

As mentioned before, the oil price dataset does not have any prices over weekends. The table below summarizes dates that are most frequently associated with a missing data over the span of 20 years.

2 Day of the Week Effect

The work by Gibbons and Hess examined returns on stocks from S&P 500, Dow Jones 30, and Treasury Bills. They found strong negative mean returns on Monday compared with other weekdays. The seasonality persisted even after market adjustment measures, such as using mean-adjusted returns instead, were taken (Gibbons & Hess, 1981). Analysis in this paper suggests a similar daily seasonality presents in crude oil returns as well. Panels in the figure below demonstrate the empirical distributions of returns on each day of the week. We can see that Monday and Wednesday have relatively larger variances, which again matches Gibbons and Hess' observations.

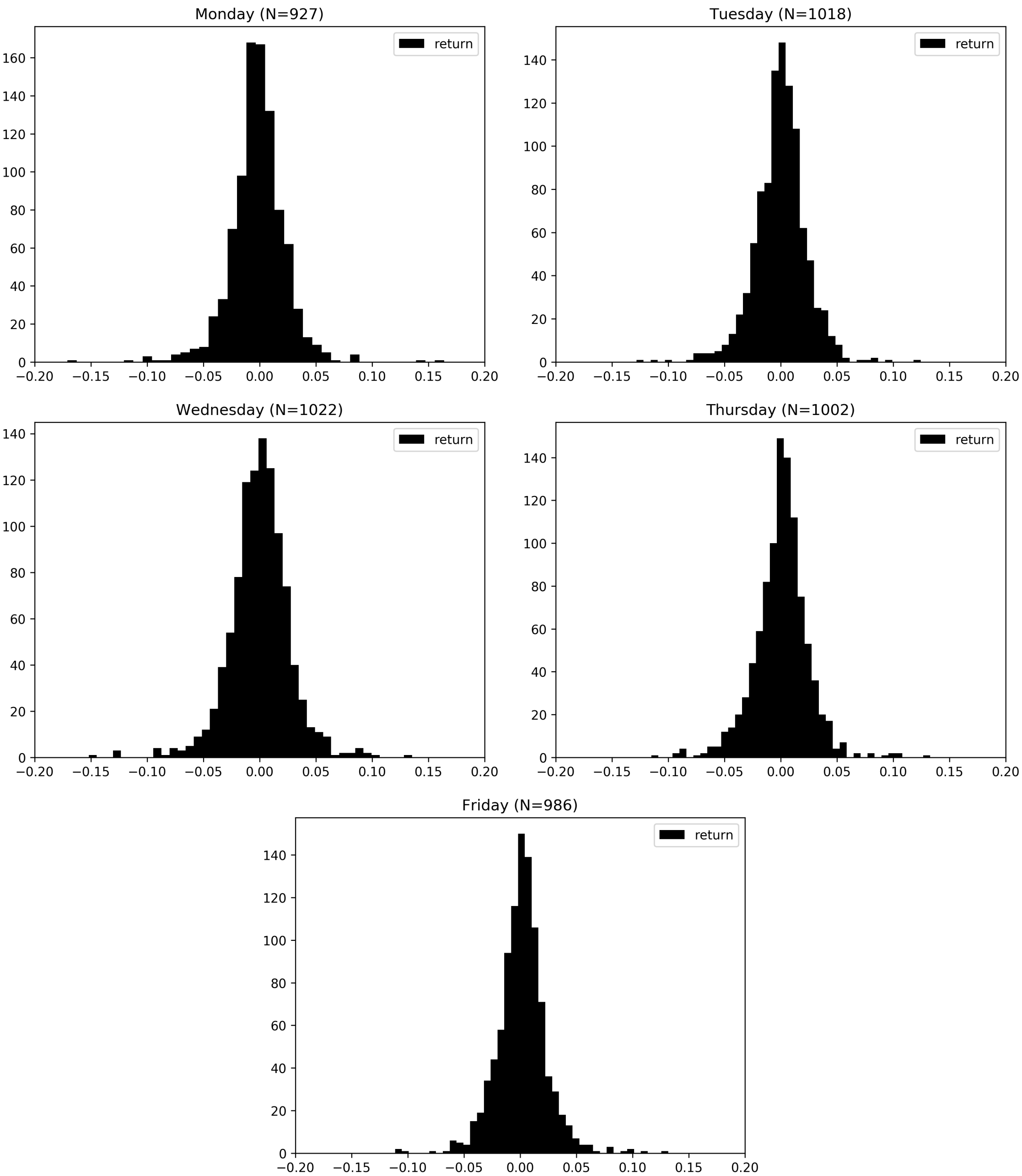

Figure 1: Crude oil returns on each weekday. Weekend data are not available in the daily dataset provided by EIA. The range of y-axis in all five histograms are from -0.2 to 0.2. Ns in parentheses denote the number of observations. See appendix for distributions of crude oil prices.

The two tables below provide summary statistics for prices and returns on each day. It turns out that Monday is the only weekday with a mean return significantly less than zero.

Day of the week	Num. Obs.	Mean	Std.	3^{rd} Moment
Monday	927	62.072	26.493	7081.163
Tuesday	1019	61.828	26.317	6895.638
Wednesday	1022	61.810	26.398	7049.810
Thursday	1002	62.005	26.431	6955.555
Friday	986	62.079	26.247	6676.566
Total	4956			

Table 3: Summary statistics of crude oil prices on each day of week

Day of the week	Num. Obs.	Mean (P-Value)	Std.	3^{rd} Moment
Monday	927	**-0.002 (0.049)**	0.025	-0.0000019
Tuesday	1018	-0.000 (0.900)	0.023	-0.0000031
Wednesday	1022	0.000 (0.884)	0.027	-0.0000054
Thursday	1002	0.001 (0.361)	0.024	-0.0000006
Friday	986	**0.002 (0.0311)**	0.023	0.0000021
Total	4955			

Table 4: Summary statistics of crude oil returns on each day of week. The first day (January 1, 2000) of the oil price dataset was Saturday, and the observation on the following Monday (January 3) was missing. Hence, the return on Tuesday (January 4) could not be computed because it was the first trading day in this dataset, and there are only 1018 Tuesday in the dataset of returns. A value of -0.000 indicates a negative value with magnitude less than 0.0005. P-values are calculated in a two-tailed t-test with $\mu_0 = 0$. Bold fonts indicate statistically significance at level $\alpha = 0.05$.

2.1 Kolmogorov-Smirnov test for Distributional Similarities

Smirnov developed a non-parametric method of testing the equality between two continuous distributions, with CDFs $F(x)$ and $G(x)$ respectively, (Smirnov, 1939). Refer to Hodges' work for a detailed review on the Kolmogorov-Smirnov test (Hodges, 1958). I am using the two-tailed version of Kolmogorov-Smirnov test to check whether distributions of two different days are similar. Given two datasets, take returns on Monday and Tuesday for example, the null hypothesis says those two datasets are drawn from the same distribution, and the alternative says they are from different distributions [1]. Firstly, the Kolmogorov–Smirnov test constructs the empirical CDFs $F_{Mon,927}(x)$ and $F_{Tue,1018}(x)$ from the dataset. Then, the Kolmogorov–Smirnov statistic measures the maximum discrepancy between two distribution functions, which is

$$D := \sup_x |F_{Mon,927}(x) - F_{Tue,1018}(x)| \in [0,1] \tag{2.1}$$

[1] Different alternative hypotheses can be used in Kolmogorov–Smirnov test: i) $H_1 : F(x) \geq G(x)$, ii) $H_1 : F(x) \leq G(x)$, and iii) $H_1 : F(x) \neq G(x)$. This paper is using the third (two-tailed) alternative hypothesis.

A smaller D-statistic implies stronger distributional similarity between two distributions. For instance, when $F_{Mon,927}(x)$ and $F_{Tue,1018}(x)$ are exactly the same, the D-statistic is zero. In contrast, let $X = 0$ and $Y = 1$ be two deterministic random variables, in this case, $D_{X,Y} = 1$.
At a significance level of α, H_0 is rejected if

$$D > \sqrt{-\frac{1}{2}\ln\frac{\alpha}{2}}\sqrt{\frac{n+m}{nm}} \tag{2.2}$$

where m and n denote sizes of two datasets.

D-Statistic (P-Value)	Monday	Tuesday	Wednesday	Thursday	Friday
Monday	0.000 (1.000)	**0.061 (0.048)**	**0.065 (0.030)**	**0.092 (0.001)**	**0.092 (0.001)**
Tuesday		0.000 (1.000)	**0.044 (0.260)**	0.036 (0.505)	0.044 (0.264)
Wednesday			0.000 (1.000)	0.053 (0.114)	**0.073 (0.009)**
Thursday				0.000 (1.000)	0.025 (0.900)
Friday					0.000 (1.000)

Table 5: The Kolmogorov-Smirnov D-Statistic for all pairs of distributions. Bold font indicates the null hypothesis is rejected at a significance level of 0.05, which implies discrepancy in distributions.

The table above presents the Kolmogorov-Smirnov D-Statistic for distributions of every pairs of days. At a significance level of 0.05, we can see that Monday follows a distribution significantly different from distributions other days follow. Because the dataset does not contain weekend data, the return on Monday is always computed using the difference between log prices on Monday and the previous Friday (Thursday if Friday is not a trading day and so on). Therefore, returns associated with Mondays pick the weekend effect. In fact, the distribution of returns on Monday (over weekend) is the only one with negative mean among distributions of all five days.

References

Gibbons, M. R., & Hess, P. (1981). Day of the Week Effects and Asset Returns. *The Journal of Business*, *54*(4), 579. doi: 10.1086/296147

Hodges, J. L. (1958). The significance probability of the smirnov two-sample test. *Arkiv för Matematik*, *3*(5), 469–486. doi: 10.1007/bf02589501

Smirnov, N. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin Moscow University*, *2*, 3–16.

3 Appendix