Online Appendix for “How Wise Are Crowds? Insights from Retail Orders and Stock Returns”
by Eric K. Kelley and Paul C. Tetlock

1 Overview

The goal of this online appendix is to quantify the degree of error in measuring the persistence in retail order imbalances in the regression model in Table VIII.

2 Calibrating the True Persistence Process

Let a firm’s true persistence be given by $x_{it}$ and its measured persistence ($m_{it}$) be given by:

$$m_{it} = x_{it} + \varepsilon_{it} \quad (1)$$

For simplicity, we assume that measurement error ($\varepsilon_{it}$) is independent over time and across firms. We further assume that true persistence follows a first-order auto-regressive (AR(1)) process in which:

$$x_{it} = \rho x_{it-1} + \nu_{it} \quad (2)$$

where the innovation in persistence $\nu_{it}$ is independent over time and across firms.

To estimate the magnitude of measurement error in our autocorrelated flow regression, we need to know $\rho$, which is the AR(1) coefficient for true persistence. We can estimate $\rho$ indirectly using regressions of measured persistence on lagged measured persistence. The regression coefficient of $m_{it}$ on $m_{it-1}$ is:

$$\beta_1 = \frac{\text{Cov}(m_{it}, m_{it-1})}{\text{Var}(m_{it-1})} = \frac{\text{Cov}(x_{it}, x_{it-1})}{\text{Var}(m_{it-1})} = \frac{\rho \text{Var}(x_{it-1})}{\text{Var}(m_{it-1})} \quad (3)$$

The regression coefficient of $m_{it}$ on $m_{it-2}$ is:

$$\beta_2 = \frac{\text{Cov}(m_{it}, m_{it-2})}{\text{Var}(m_{it-2})} = \frac{\text{Cov}(x_{it}, x_{it-2})}{\text{Var}(m_{it-2})} = \rho^2 \frac{\text{Var}(x_{it-2})}{\text{Var}(m_{it-2})} \quad (4)$$

We assume that the $m_{it}$ and $x_{it}$ processes are homoskedastic, which allows us to solve for $\rho$.
using the ratio of the two coefficients above:

\[
\frac{\beta_2}{\beta_1} = \rho
\]  (5)

We can now estimate the variance of true persistence using:

\[
\frac{\beta_1}{\rho} Var(m) = Var(x)
\]  (6)

where \( Var(m) \) is the cross-sectional variance of measured persistence, which we can estimate. The variance in measurement error is simply the difference between the variances of measured and true persistence:

\[
Var(m) - Var(x) = Var(m) - \frac{\beta_1}{\rho} Var(m) = Var(\varepsilon)
\]  (7)

The expectation of true persistence is the same as the expectation of measured persistence:

\[
E(m) = E(x)
\]  (8)

Lastly, we can infer the variance in the innovation in persistence from the AR(1) coefficient and the variance in true persistence:

\[
(1 - \rho^2) Var(x) = Var(\nu)
\]  (9)

Using our estimates of \( Var(m) \), \( E(m) \), \( \beta_1 \), and \( \beta_2 \) (0.150², 0.0858, 0.132, 0.058), we obtain the following values:

\[
E(x) = 0.0885
\]  (10)

\[
\rho = 0.439
\]  (11)

\[
Var(x) = 0.00676
\]  (12)

\[
Var(\varepsilon) = 0.01574
\]  (13)

\[
Var(\nu) = 0.005454
\]  (14)
3 Imbalance and Return Processes

Each firm’s imbalances ($i_{it}$) follows an AR(1) process governed by the true persistence parameter ($x_{it}$):

$$i_{it} = x_{it-1}i_{it-1} + \delta_{it}$$  \hspace{1cm} (15)

where $\delta_{it}$ is the innovation in imbalances. We ignore the mean of imbalances, which is approximately zero.

A firm’s returns are governed by a simplified version of our regression model, where all coefficients are zero except the direct coefficient on imbalances and the interaction coefficient between imbalances and persistence. That is, we assume that the empirical regression model used in Table VIII is correctly specified, except that true persistence appears instead of measured persistence:

$$r_{ik} = \gamma_1i_{it} + \gamma_2x_{it}i_{it} + \eta_{it}$$  \hspace{1cm} (16)

where the $k$ subscript captures the fact that returns are measured over an unspecified interval after period $t$. The only other minor difference is that Table VIII uses quintiles, whereas the equation above uses a continuous variable for true persistence. The simulation procedure below accounts for this discrepancy.

4 Simulation Details

All of the simulation equations are described above. We assume that all random variables follow normal distributions. We set the initial values of the imbalance, persistence, and measured persistence variables equal to random numbers drawn from their steady state distributions. Returns in each period are determined by the regression equation above. The imbalance and true persistence variables evolve according to the laws of motion in Equations (15) and (2), while measured persistence satisfies Equation (1) in each period.

All simulation parameters are given in Equations (10) through (14), except that the variance in the innovation in imbalances is calibrated to match the empirical cross-sectional variance in imbalance, which is $0.527^2$. The variance in the disturbance term ($\eta_{it}$) in the return regression does not affect any of the regression coefficients so we set it close to zero to minimize the estimation error in the simulations.

The only two undetermined parameter are the return predictability coefficients on imbalance and its interaction with true persistence ($\gamma_1$ and $\gamma_2$). To determine these two coefficients,
we estimate a simplified version of the empirical regression equation in Table VIII:

\[ r_{ik} = b_1 i_{it} + b_2 Q(m_{it})i_{it} + e_{it} \]

(17)

where \( Q(m_{it}) \) is a variable equal to -2, -1, 0, +1, or +2 representing the quintiles of \( m_{it} \) in period \( t \). The two estimated \( b_1 \) and \( b_2 \) coefficients are the measured counterparts of \( \gamma_1 \) and \( \gamma_2 \). We run simulations using different \( \gamma_1 \) and \( \gamma_2 \) values until we obtain estimates of \( b_1 \) and \( b_2 \) that match our empirical estimates in in Table VIII. For example, these estimates are 0.00142 and 0.000205 at the one-week return horizon in Table VIII.

In each simulation, we use 2000 firms and 20 quarterly time periods, analogous to our data. We use Fama-MacBeth time series averages of the cross-sectional regression coefficients to estimate \( b_1 \) and \( b_2 \). We evaluate this time series average across 100 simulations to ensure that our estimates are precise. When we simulate the system using \( \gamma_1 \) and \( \gamma_2 \) values of 0.000816 and 0.00682, we obtain \( b_1 \) and \( b_2 \) values of 0.00142 and 0.000205, which matches the one-week return predictability estimates in Table VIII.

The last step is to use these \( \gamma_1 \) and \( \gamma_2 \) values to estimate the idealized counterpart of the model in Table VIII in which we replace measured persistence with true persistence:

\[ r_{ik} = c_1 i_{it} + c_2 Q(x_{it})i_{it} + e_{it} \]

(18)

The values of \( c_1 \) and \( c_2 \) are the true regression coefficients of interest. When we use \( \gamma_1 \) and \( \gamma_2 \) that produce (the one-week) \( b_1 \) and \( b_2 \) values of 0.00142 and 0.000205, the corresponding values of \( c_1 \) and \( c_2 \) are 0.00142 and 0.000373.

By comparing the \( b \) and \( c \) coefficients, we see that \( b_1 = c_1 \), whereas \( b_2 = (0.548)c_2 < c_2 \) because of measurement error in persistence. We therefore employ a correction factor of \( 1/0.548 = 1.824 \) to increase the magnitude of our interaction coefficients and their standard errors in the one-week specification. We use an analogous procedure to adjust the magnitude of the two-week return predictability coefficients on the interaction term, obtaining correction factors of 1.823 and 1.823, respectively. These results strongly suggest that the same correction factor applies to all horizons, meaning that measurement error has a proportional impact on the interaction coefficients.