Lecture 8: High-Frequency Financial Data (HFFD) & Market Microstructure

Outline

8.1 Features of HFFD

8.2 Models for Price Changes: Ordered probit and ADS Models

8.3 Models for Inter-Trade Times (Duration Models): ACD Model

8.4 Realized Volatility

8.1 Features of HFFD

Market microstructure: Why is it important?

- Important in market design & operation, e.g. to compare different markets (NYSE vs NASDAQ)
- To study price discovery, liquidity, volatility, etc.
- To understand costs of trading
- Important in learning the consequences of institutional arrangements on observed processes, e.g.
  - nonsynchronous trading
  - bid-ask bounce
  - impact of changes in tick size, after-hour trading, etc.
  - impact of daily price limits (many foreign markets)

We will look at some of these specialized features of HFFD next: **called F1–F8**.

Def: HFFD is data for which observations are collected with time intervals of 24 hours or less (intraday).

Some examples:

- Transaction (or tick-by-tick) data
- 5-minute returns in FX
- 1-minute returns on index futures and cash market

F1: Nonsynchronous trading

Key implication: may induce serial correlations even when the underlying returns are iid.

Example 1. *Stocks A and B are independent, but is A traded more frequently than B, if impactful news arrives near closing time, this will affect A more than B (A traded more often), and will only affect B the next day. This effect will show up as a lag 1 correlation between A and B, yet they are independent...*

Example 2 (Lo & MacKinlay (1990))**.** *Consider the following simplified setup:*

- *log returns,* $\{r_t\} \sim iid(\mu, \sigma^2)$
- *For each time index* t, $P(no\ trade) = \pi$
- *Cannot observe* r_t *if there is no trade*

A simple model for the observed log return series, r_t^o*, is given by:*

$$r_t^o = \sum_{i=0}^{\ell} r_{t-i}$$

where ℓ *is the largest integer such that no trading occurred during period* $t-1, \ldots, t-\ell$*.*

$$r_t^o = \begin{cases} 0, & \ell = \emptyset, \quad \Leftrightarrow \ no\ trade\ at\ t \\ r_t, & \ell = 0, \quad \Leftrightarrow \ trade\ at\ t\ and\ t-1 \\ r_t + r_{t-1}, & \ell = 1, \quad \Leftrightarrow \ trade\ at\ t\ and\ t-2\ but\ not\ at\ t-1 \\ \vdots & \end{cases}$$

so that (using Law of Total Probability) the distribution of r_t^o*, is given by:*

$$r_t^o = \begin{cases} 0, & with\ prob\ \pi \\ r_t, & with\ prob\ (1-\pi)^2 \\ r_t + r_{t-1}, & with\ prob\ (1-\pi)^2\pi \\ \vdots & \\ \sum_{i=0}^{k} r_{t-i}, & with\ prob\ (1-\pi)^2\pi^k \\ \vdots & \end{cases}$$

Using the resulting geometric sums (and some tricks), not hard to show:

$$Cov(r_t^o, r_{t-h}^o) = \begin{cases} \sigma^2 + \frac{2\pi\mu^2}{1-\pi}, & h = 0 \\ -\mu^2\pi^h, & h \geq 1 \end{cases}$$

Thus: $\{r_t^o\}$ *is negatively correlated, even though* $\{r_t\} \sim iid$*...*

F2: Bid-ask bounce

Key implication: may induce serial correlations even when the underlying returns are iid.

Background: Market makers facilitate transactions in some markets; make money by selling (P_a) at a higher price than they buy (P_b). The difference:

$$0 < S = P_a - P_b = \textbf{ bid-ask spread}.$$

The existence of S, although small in magnitude, has several important consequences in time series of asset returns. One of the most notable is that bid and ask quotes introduce negative lag-1 serial correlation in observed price changes, the so-called **bid-ask bounce**.

Example 3 (Roll (1984))**.** *Assume true (unobserved) market price of an asset $P_t^* = S/2$, is unchanged, so that $P_t^* - P_{t-1}^* = 0$.*

Because of market "friction", we observe the noisy version P_t, where:

$$P_t = P_t^* + \frac{S}{2} I_t, \qquad \{I_t\} \sim \textit{IID binary r.v.s with } P(I_t = +1) = \frac{1}{2} = P(I_t = -1)$$

which implies the following time series model for price changes (returns)

$$\Delta P_t = (1 - B)P_t = P_t - P_{t-1} = (I_t - I_{t-1})\frac{S}{2}$$

Since $\mathbb{E}(I_t) = 0$ and $\mathbb{V}(I_t) = 1$, we can show $\mathbb{E}(\Delta P_t) = 0$ and

$$Cov(\Delta P_t, \Delta P_{t-h}) = \begin{cases} S^2/2, & h = 0 \\ -S^2/4, & h = 1 \\ 0, & h \geq 2 \end{cases}$$

so that $\Delta P_t \sim$ MA(1) with ACF

$$\rho(h) = \begin{cases} -0.5, & h = 1 \\ 0, & h \geq 2 \end{cases}$$

This result continues to hold if P_t^* follows a random walk, so that $P_t^* - P_{t-1}^* \sim$ white noise.

Example 4 (Tsay IAFD (2013) Fig. 6-1–6.2)**.** *See Figure 1. Fig. 6.1 gives the trading prices and log returns of the intraday Caterpillar stock over 4 days in January 2010 ($n = 3895$), in 30-s intervals. Fig. 6.2 shows the sample ACF: we see a significant spike of $\hat{\rho}_1 = -0.052$.*

Figure 1: Tsay IAFD (2013) Fig. 6-1–6.2.

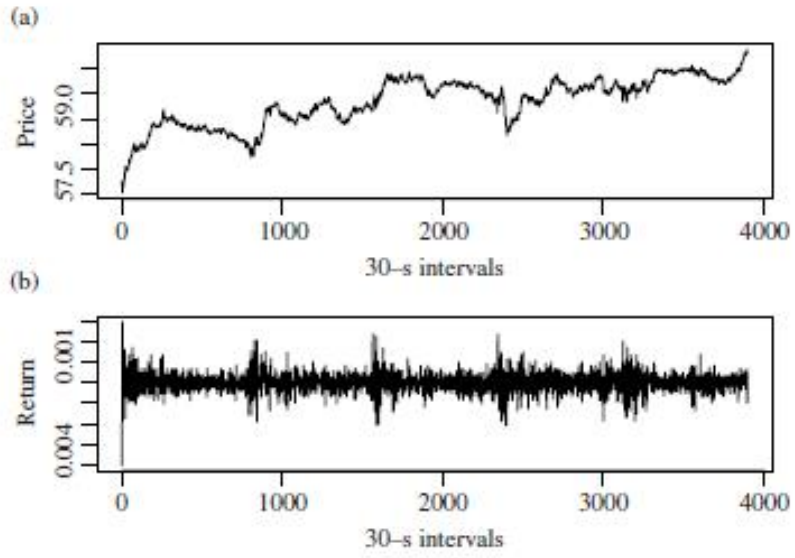


Figure 6.1. Intraday trading prices and log returns of Caterpillar stock from January 4 to January 8, 2010: (a) prices and (b) log returns both in 30-s interval.

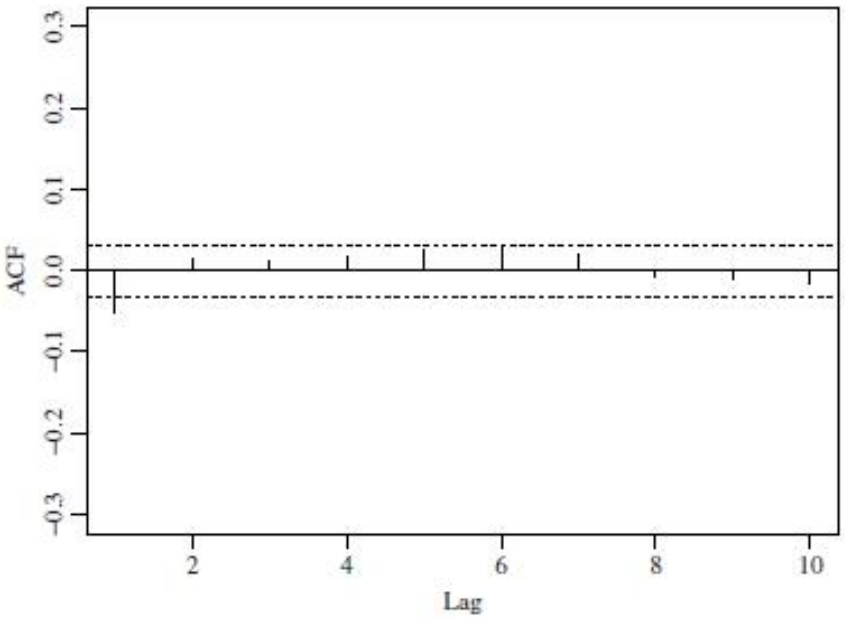


Figure 6.2. Sample ACF of the intraday 30-s long returns for Caterpillar stock from January 4 to January 8, 2010. There are 3895 returns.

F3-F8: Empirical characteristics of HFFD

Let: t_i be the calendar time (measured in seconds from midnight) at which the i-th trade of an asset takes place. Most common features of HFFD:

F3: High-dimensional data. The collection of variables associated with t_i are called *transactions data* (e.g., price, volume, quotes, etc.); typified by large sample sizes; need multi-dimensional dynamic models capable of handling many variables at once.

F4: Irregular time intervals. E.g., stock trades do not form an equally spaced time series (occur at random times). Time duration between trades is important, may contain useful information about market microstructure (e.g., trading intensity).

F5: Serial correlation, long-range dependence, leptokurtic/heavy tails. Apart from returns which are typically WN, most HFFD is autocorrelated, with short as well as long-range dependence. Also need to allow for heavier-than-Gaussian tails.

F6: Discrete and mixed values. E.g., price in multiples of tick size or of 1 cent; means we can treat series as being discrete. A lot of HFFD also consists of mixed continuous and discrete values.

F7: Positive values. Most HFFD is restricted to positive values, calling for specialized models/techniques.

F8: Strong intraday periodicites. E.g., diurnal pattern on NYSE, transactions are heavier at start and close of trading hours, thinner during lunch hour (U-shaped intensity). The effects are mirrored in time durations between transactions, and in stock volatility.

Definitions/Notation

- Δt_i = time change from $(i-1)$-th to i-th trade (measured in seconds from midnight)
- y_i = price change from $(i-1)$-th to i-th trade
- Let F_{i-1} be the information set available at the $(i-1)$-th transaction, and $\boldsymbol{x}_i \in F_{i-1}$, $\boldsymbol{z}_i \in F_{i-1}$, and $\boldsymbol{w}_i \in F_{i-1}$ be different sets of vectors of covariates.

Example 5 (Tsay IAFD (2013), Fig. 6.3–6.4)**.** *See Figure 2–3. Fig. 6.3 gives transaction-by-transaction data of J&J stock: 418,855 intraday price changes over 10 days in October 2010.*

```
R demonstration (Fig. 6.3 and Table 6.2)
> da=read.table("taq-jnj-t-oct4t152010.txt",header=T)
> head(da)
    date hour minute second price volume
20101004    6     25     15 61.75    100
20101004    8     33     19 61.56    100
20101004    8     41      9 61.56    100
20101004    8     48     50 61.60    100
20101004    8     48     55 61.60    100
> source("hfchg.R") ### R script to compute price change
> m1=hfchg(da)
number of trading days: 10
> names(m1)
[1] "pchange" "duration" "size"
> par(mfcol=c(2,1)); idx=c(410000:418854)
> plot(m1$pchange,type='l',ylab='change') #plot(idx,m1$pchange[idx],type='l',ylab='pch')
> hist(m1$pchange, nclass=400, xlim=c(-0.04,0.04)) ### May use xlim=c(-0.06,0.06)
> source("hfntra.R") # R script to tabulate number of transactions in a given
  time interval (measured in minutes).
> m1=hfntra(da,5)
> names(m1)
[1] "ntrad"
```

Main points from Fig. 6.3 and Table 6.2:

- *The histogram indicates most transactions (73%) are without price change, and about 26% result in a price change that is less than or equal to 1 cent.*
- *Only 0.83% of transactions were associated with a price change between (1, 2] cents.*
- *Only about 0.26% of the transactions resulted in price changes of 2 cents or more.*
- *The empirical distribution of price changes is approximately symmetric about zero.*

Fig. 6.4 is a time plot of the first 3000 durations (in seconds) between successive trades. Fig. 6.5(a) shows the number of transactions in 5-min contiguous intervals (the time gaps between trading days are ignored). Fig. 6.5(a) shows the corresponding sample ACF of the series in Fig. 6.5(a). Main points:

- *Fig. 6.4 confirms that trading did not occur at equally spaced intervals and there some zero durations (multiple trades in a second).*
- *The time plot in Fig 6.5(a) exhibits roughly a cyclical U-shaped pattern with 10 cycles.*
- *The ACF in Fig 6.5(b) shows a clear diurnal pattern in trading intensity seasonal with periodicity 78 (there are* ≈ 78 *5-min periods in the length of a trading day).*

Figure 2: Tsay IAFD (2013) Fig. 6.3 and Table 6.2.

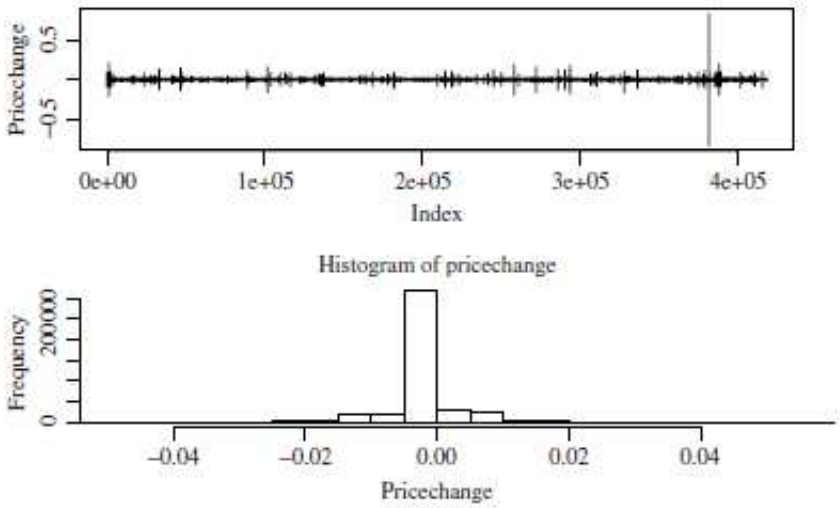


Figure 6.3. Time plot and histogram of intraday price changes in consecutive trades for Johnson and Johnson stock from October 4 to October 15, 2010. Only transactions occurred in the normal trading hours are used. There are 418,855 price changes in 10 trading days.

TABLE 6.2. Frequencies of Price Change in Consecutive Trades for Johnson and Johnson Stock From October 4 to October 15, 2010

Cents	< -2	$[-2,-1)$	$[-1,0)$	0	$(0,1]$	$(1,2]$	> 2
Counts	540	1,794	55,325	304,067	54,860	1,711	558
Percentage	0.128	0.428	13.209	72.595	13.098	0.408	0.132

[a]Only Transactions occurred in the normal trading hours are used. Total number of price changes is 418,855.

Figure 3: Tsay IAFD (2013) Fig 6.4 & 6.5.

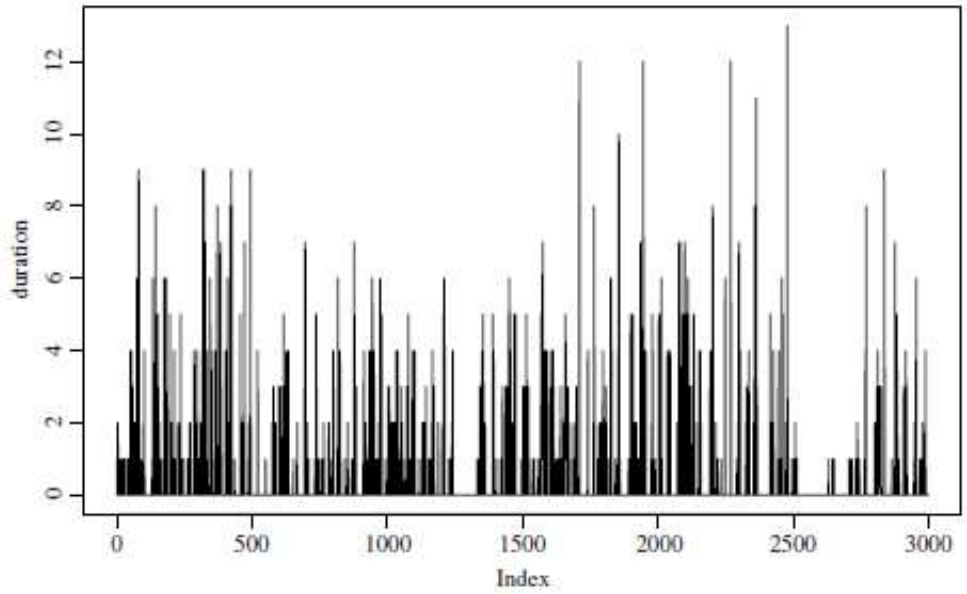


Figure 6.4. Time plot of durations, measured in seconds, between consecutive trades for Johnson and Johnson stock from October 4 to October 15, 2010. Only the first 3000 durations in the normal trading hours are shown.

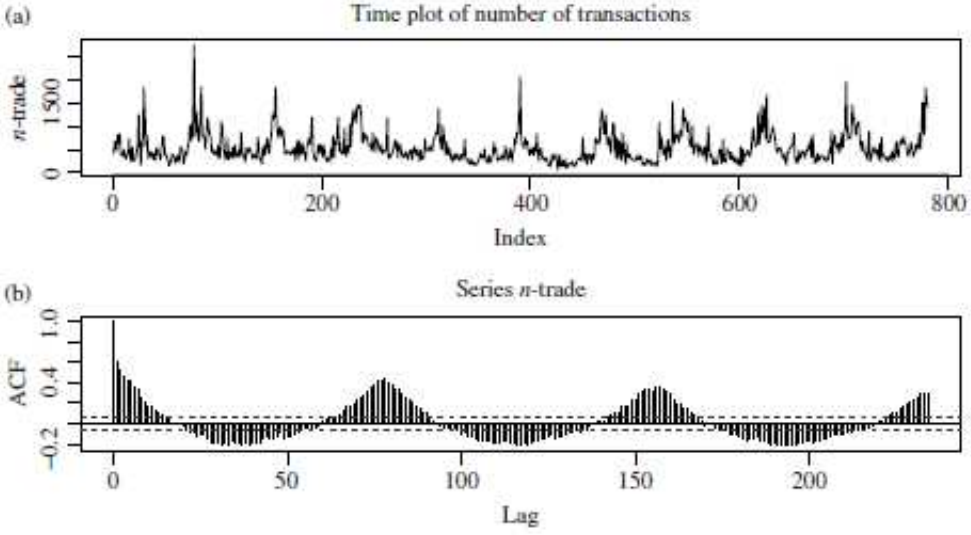


Figure 6.5. Johnson and Johnson intraday transactions data from October 4 to October 15, 2010: (a) the number of transactions in 5-min time intervals and (b) sample ACF of the series in part (a).

8.2 Models for Price Changes

We consider models for y_i and Δt_i, both individually and together, focusing on two models that use explanatory variables to study the intraday price movements.

Ordered Probit Model: Hauseman et al. (1992)

The first model assumes:

- y_t^* is continuous and follows the regression model

$$y_t^* = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \qquad \epsilon_i|\boldsymbol{x}_i \sim (0, \sigma_i^2)$$

- A conditional Gaussian assumption is made where the conditional variance is a positive function of another set of explanatory variables, $\boldsymbol{w}_i$, parametrized by $\boldsymbol{\theta}$

$$\mathbb{V}(\epsilon_i|\boldsymbol{x}_i) = \sigma_i^2 = \sigma_i^2(\boldsymbol{w}_i) = g(\boldsymbol{\theta}, \boldsymbol{w}_i) > 0$$

- This implies $\epsilon_i|(\boldsymbol{x}_i, \boldsymbol{w}_i) \sim N(0, \sigma_i^2)$ and therefore $y_i|(\boldsymbol{x}_i, \boldsymbol{w}_i) \sim N(\boldsymbol{x}_i\boldsymbol{\beta}, \sigma_i^2)$
- Relationship between y_i^* and y_i is that y_i is a condensation of y_i^* into k ordered categories

$$y_i = s_j \quad \Longleftrightarrow \quad \alpha_{j-1} < y_i^* < \alpha_j, \qquad j = 1, \ldots, k$$

where

$$-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_{k-1} = \alpha_k = \infty$$

Under the conditional Gaussian assumption we have:

$$\begin{aligned} P(y_i = s_j|\boldsymbol{x}_i, \boldsymbol{w}_i) &= P(\alpha_{j-1} < \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i \leq \alpha_j|x_i, w_i) \\ &= \begin{cases} \Phi\left(\frac{\alpha_1 - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i}\right), & \text{if } j = 1 \\ \Phi\left(\frac{\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i}\right) - \Phi\left(\frac{\alpha_{j-1} - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i}\right), & \text{if } j = 2, \ldots, k-1 \\ 1 - \Phi\left(\frac{\alpha_{k-1} - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i}\right), & \text{if } j = k \end{cases} \end{aligned}$$

The entire set of parameters, $\{\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}\}$, is estimated by maximizing the likelihood.

Example 6 (Tsay IAFD, Ex 6.1)**.** *Consider the intraday price changes of Caterpillar on 4 January 2010 (37,715 price changes), using following 7 categories:*

category	1	2	3	4	5	6	7
cents	$(-\infty,-2)$	$[-2,-1)$	$[-1,0)$	0	$(0,1]$	$(1,2]$	$(2,\infty)$
% of obs	0.61	1.70	15.2	64.98	15.04	1.83	0.66

To focus on dynamic dependence of intraday price changes, define dummy variables for ℓ lagged price changes being in category j (only 12 such dummies are needed):

$$y_{\ell,j} = \begin{cases} 1, & \textit{if price change } \ell \textit{ steps back } = s_j \\ 0, & \textit{otherwise} \end{cases}, \qquad \textit{for } j = 2,\ldots,7 \textit{ and } \ell = 1,2.$$

Also use the observed price changes from up to 3 lags back ($y_{i-\ell}$), and the lag-2 transaction volume divided by 100 (v_{i-2}), so that the regression model entertained here is:

$$\boldsymbol{x}_i\boldsymbol{\beta} = \beta_1 v_{i-2} + \sum_{\ell=1}^{3}\beta_{1+\ell}y_{i-\ell} + \sum_{j=2}^{7}(\gamma_{1,j}y_{1,j} + \gamma_{2,j}y_{2,j}), \qquad \sigma_i^2 = \textit{constant}$$

Some remarks:

- ***polr*** *allows for weighted regression to handle heteroscedasticity (option "weights"), but cannot simultaneously estimate $\boldsymbol{\beta}$ and σ_i^2 (which defaults to 1).*
- *Estimates of boundary parameters α_j not symmetric with respect to zero.*
- *Model implies that*

$$P(y_i^* \le s_j|\boldsymbol{x}_i, \boldsymbol{w}_i) = \Phi(\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta})$$

```
R demonstration (Fig. 6.8 and Table 6.4)
> da=read.table("taq-cat-t-jan042010.txt",header=T)
> head(da)
      date hour minute second price size
1 20100104    9     30      0 57.65 3910
.....
6 20100104    9     30      1 57.65  462
> vol=da$size/100
> da1=read.table("taq-cat-cpch-jan042010.txt")
> cpch=da1[,1] % category of price change
> pch=da1[,2] % price change
> cf=as.factor(cpch) % create categories in R
> length(cf)
[1] 37715
> y=cf[4:37715]
> y1=cf[3:37714] % create indicator variables for lag-1 cpch
> y2=cf[2:37713] % create indicator variables for lag-2 cpch
> vol=vol[2:37716]
> v2=vol[2:37713] % create lag-2 volume
```

```
> cp1=pch[3:37714] % select lagged price changes
> cp2=pch[2:37713]; cp3=pch[1:37712]
> library(MASS) % load package
> m1=polr(y~v2+cp1+cp2+cp3+y1+y2,method="probit")
> summary(m1)
Call:
polr(formula = y ~ v2 + cp1 + cp2 + cp3 + y1 + y2, method = "probit")
Coefficients:
        Value Std. Error t value
v2   -0.003765 0.0009453 -3.983
cp1  -7.836883 1.4613047 -5.363
cp2 -10.864394 1.5306456 -7.098
cp3 -12.283682 0.7710955 -15.930
y12  -0.274407 0.0923566 -2.971
y13  -0.742792 0.0908854 -8.173
y14  -1.330665 0.0963540 -13.810
y15  -1.858199 0.1042257 -17.829
y16  -2.261587 0.1218013 -18.568
y17  -2.493321 0.1563177 -15.950
y22  -0.098542 0.0935908 -1.053
y23  -0.307034 0.0923725 -3.324
y24  -0.531115 0.0980150 -5.419
y25  -0.744706 0.1062435 -7.009
y26  -0.932655 0.1238918 -7.528
y27  -0.858858 0.1596219 -5.381
Intercepts:
Value Std.Error t value
1|2 -4.5941 0.1459 -31.4803
2|3 -4.0170 0.1445 -27.7989
3|4 -2.8599 0.1438 -19.8926
4|5 -0.8528 0.1435 -5.9437
5|6 0.2868 0.1434 1.9996
6|7 0.8882 0.1435 6.1883
Residual Deviance: 74802.56
AIC: 74846.56
> names(m1)
[1] "coefficients" "zeta" "deviance" "fitted.values"
[5] "lev" "terms" "df.residual" "edf"
[9] "n" "nobs" "call" "method"
[13] "convergence" "niter" "lp" "model"
[17] "contrasts" "xlevels"
> yhat=m1$fitted.values
> print(yhat[1:5,],digits=3)
    1        2        3       4     5      6      7
1 1.11e-03 0.005420 0.08605 0.660 0.2134 0.0266 0.007696
2 1.55e-02 0.041461 0.27883 0.608 0.0535 0.0028 0.000444
3 8.99e-06 0.000094 0.00522 0.287 0.4311 0.1605 0.116298
4 1.87e-04 0.001251 0.03267 0.539 0.3343 0.0658 0.027144
5 6.41e-04 0.003470 0.06457 0.630 0.2527 0.0365 0.011836
```

Figure 4: Tsay IAFD (2013) Table 6.4 & Fig 6.8.

TABLE 6.4. Estimation Results of an Ordered Probit Model for the Intraday Price Changes of Caterpillar Stock on January 4, 2010 with 37,716 Transactions[a]

(a)	Boundary Partitions of the Probit Model					
Parameter	α_1	α_2	α_3	α_4	α_5	α_6
Estimate	−4.594	−4.017	−2.860	−0.853	0.287	0.888
t	−31.48	−27.80	−19.89	−5.944	2.000	6.188

(b)	Equation Prameters of Probit Model (Estimates are Negative)							
Parameter	β_1	β_2	β_3	β_4	$\gamma_{1,2}$	$\gamma_{1,3}$	$\gamma_{1,4}$	$\gamma_{1,5}$
Estimate	0.004	7.837	10.86	12.28	0.274	0.743	1.331	1.858
t	3.983	5.363	7.098	15.93	2.971	8.173	13.81	17.83
Parameter	$\gamma_{1,6}$	$\gamma_{1,7}$	$\gamma_{2,2}$	$\gamma_{2,3}$	$\gamma_{2,4}$	$\gamma_{2,5}$	$\gamma_{2,6}$	$\gamma_{2,7}$
Estimate	2.262	2.493	0.099	0.307	0.531	0.745	0.933	0.859
t	18.57	15.95	1.053	3.324	5.419	7.009	7.528	5.381

[a]The model is in Equation (6.19) and t denotes t-ratio.

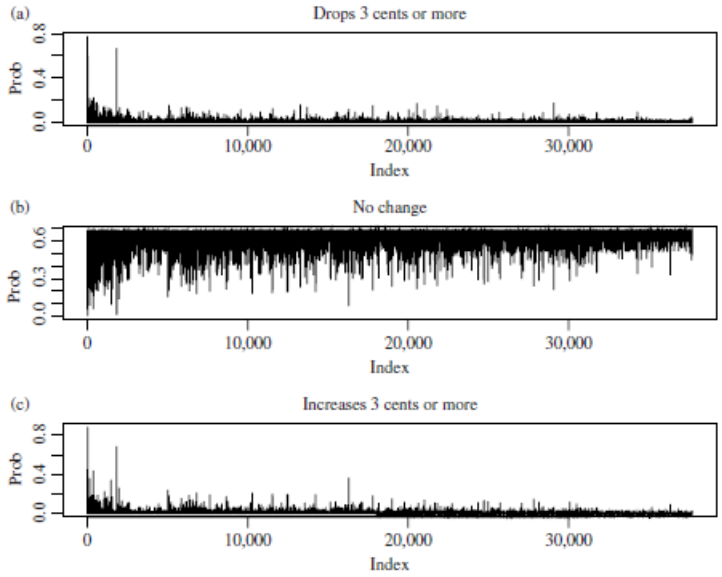


Figure 6.8. Time plots of fitted probabilities of price change for Caterpillar stock on January 4, 2010: the three plots are for (a) decrease 2 cents or more, (b) no change, and (c) increase 2 cents or more.

ADS Decomposition Model: McCulloch & Tsay (2000) and Rydberg & Shephard (2003)

Under the ADS model, the price change at i-th transaction is written as a product of 3 terms:

$$y_i = A_i D_i S_i$$

The evolution of price change can then be partitioned as

$$P(y_i|F_{i-1}) = P(S_i|D_i, A_i, F_{i-1})P(D_i|A_i, F_{i-1})P(A_i|F_{i-1})$$

where:

- The term $A_i \sim \text{Bernoulli}(p_i)$ is a binary indicator of price change with success probability modeled via logistic regression:

$$A_i = \begin{cases} 1, & y_i \neq 0, \\ 0, & y_i = 0 \end{cases}, \qquad p_i = P(A_i = 1) = \frac{\exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}\}}{1 + \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}\}}$$

- The term $D_i|(A_i = 1) \sim 2\text{Bernoulli}(\delta_i) - 1$ is a binary indicator of direction of price change, also modeled via logistic regression:

$$D_i|(A_i = 1) = \begin{cases} 1, & y_i > 0, \\ -1, & y_i < 0 \end{cases}, \qquad \delta_i = P(D_i = 1|A_i = 1) = \frac{\exp\{\boldsymbol{z}_i^T \boldsymbol{\gamma}\}}{1 + \exp\{\boldsymbol{z}_i^T \boldsymbol{\gamma}\}}$$

- The term $S_i \geq 0$ (conditional on D_i and $A_i = 1$) is the absolute value of the size of the price change, modeled via geometrics (≥ 1) with logistic regression varying rates (for $j = u, d$):

$$S_i|(D_i, A_i = 1) \sim \begin{cases} \text{Geometric}(\lambda_{u,i}), & D_i = 1,\ A_i = 1, \\ \text{Geometric}(\lambda_{d,i}), & D_i = -1,\ A_i = 1, \end{cases}, \qquad \lambda_{j,i} = \frac{\exp\{\boldsymbol{w}_i^T \boldsymbol{\theta}_j\}}{1 + \exp\{\boldsymbol{w}_i^T \boldsymbol{\theta}_j\}}$$

The resulting log-likelihood function

$$\log\{P(y_1, \ldots, y_n|F_0)\} = \sum_{i=1}^{n} \log\{P(y_i|F_{i-1})\}$$

is then maximized over the parameter vectors: $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}_u, \boldsymbol{\theta}_d)$.

Example 7 (Tsay IAFD, Ex 6.2)**.** *Model the 37,715 intraday price changes of Caterpillar stock during the normal trading hours of 4 January 2010. A simple model uses the following covariates:*

- $\boldsymbol{x}_i = A_{i-1}$*: the action indicator of the previous trade*
- $\boldsymbol{z}_i = D_{i-1}$*: the direction indicator of the previous trade*
- $\boldsymbol{w}_i = S_{i-1}$*: the size of the previous trade*

The respective logits are therefore:

$$\begin{aligned}
\log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 A_{i-1} \\
\log\left(\frac{\delta_i}{1-\delta_i}\right) &= \gamma_0 + \gamma_1 D_{i-1} \\
\log\left(\frac{\lambda_{u,i}}{1-\lambda_{u,i}}\right) &= \theta_{u,0} + \theta_{u,1} S_{i-1} \\
\log\left(\frac{\lambda_{d,i}}{1-\lambda_{d,i}}\right) &= \theta_{d,0} + \theta_{d,1} S_{i-1}
\end{aligned}$$

```
R demonstration (Table 6.5)
> da=read.table("taq-cat-cpch-jan042010.txt")
> dim(da)
[1] 37715 2
> pch=da[,2] % create Ai, Di, and Si and their lagged variables
> idx=c(1:37715)[pch > 0]
> jdx=c(1:37715)[pch < 0]
> A=rep(0,37715); A[idx]=1; A[jdx]=1
> D=rep(0,37715); D[idx]=1; D[jdx]=-1
> S=abs(da[,1]-4)
> Ai=A[2:37715]; Aim1=A[1:37714]
> Di=D[2:37715]; Dim1=D[1:37714]
> Si=S[2:37715]; Sim1=S[1:37714]
> m1=glm(Ai~Aim1,family="binomial")
> summary(m1)
Call: glm(formula = Ai ~ Aim1, family = "binomial")
Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -1.07342     0.01466   -73.22    <2e-16 ***
Aim1         1.18316     0.02277    51.95    <2e-16 ***
---
Residual deviance: 46085 on 37712 degrees of freedom
AIC: 46089
> di=Di[Ai==1]
> dim1=Dim1[Ai==1]
> di=(di+abs(di))/2 % transform di to binary
> m2=glm(di~dim1,family="binomial")
> summary(m2)
```

```
Call: glm(formula = di ~ dim1, family = "binomial")
Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -0.009755    0.018994   -0.514     0.608
dim1        -1.241364    0.028731  -43.207    <2e-16 ***
```

This leads to the estimates on Table 6.5 (see Fig. 5).

Figure 5: Tsay IAFD (2013) Table 6.5

TABLE 6.5. Parameter Estimates of the ADS Model in Equation (6.29) for Caterpillar Stock Traded on January 4, 2010

Parameter	β_0	β_1	γ_0	γ_1
Estimate	−1.073	1.183	−0.010	−1.241
standard error	0.015	0.023	0.019	0.029
Parameter	$\theta_{u,0}$	$\theta_{u,1}$	$\theta_{d,0}$	$\theta_{d,1}$
Estimate	1.649	−0.297	1.534	−0.162
standard error	0.041	0.035	0.039	0.037

We obtain the following results:

- *Probability of price change at i-th transaction:*

$$\begin{aligned} P(A_i = 1|A_{i-1} = 0) &= 0.255 = P(\textit{price change at } i \textit{ given there } \textbf{\textit{wasn't}} \textit{ one at } i-1) \\ P(A_i = 1|A_{i-1} = 1) &= 0.527 = P(\textit{price change at } i \textit{ given there } \textbf{\textit{was}} \textit{ one at } i-1) \end{aligned}$$

- *Probability of price increase at i-th transaction:*

$P(D_i = 1|A_{i-1} = 0) = 0.500 = P(\textit{price increase at } i \textit{ given there } \textbf{\textit{wasn't a price change}} \textit{ at } i-1)$

$P(D_i = 1|A_i = 1, D_{i-1} = 1) = 0.223 = P(\textit{price increase at } i \textit{ given there } \textbf{\textit{was a price increase}} \textit{ at } i-1)$

$P(D_i = 1|A_i = 1, D_{i-1} = -1) = 0.774 = P(\textit{price increase at } i \textit{ given there } \textbf{\textit{was a price decrease}} \textit{ at } i-1$

The fact that $0.223 < 0.774$ *supports the* ***bid-ask bounce*** *effect, and* ***price reversals*** *in HFFD.*

- *Size of price change at i-th transaction:*

$$S_i|(D_i = 1) = \textit{size of price increase at } i \sim \textit{Geometric}(\lambda_{u,i} = 1.649 - 0.297 S_{i-1})$$

Since $\mathbb{E}[\textit{Geometric}(\lambda)] = \lambda^{-1}$*, the prob of a large* S_i *increases with* S_{i-1}*.*

8.3 Models for Inter-Trade Times (Duration Models)

Duration models are concerned with **time intervals (durations) between trades** Δt_i (longer durations indicate lack of trading activities, which in turn signify a period of no new information).

Main Model: autoregressive conditional duration (ACD) proposed by Engle & Russell (1998):

- uses a GARCH-style model to capture the volatility of the conditional mean of the durations;
- extended by Zhang et al (2001) to account for nonlinearity and structural breaks;
- implemented in R package `ACDm` (function "acdFit");
- assumes the data has been deseasonalized...

Deseasonalization step

For $i = 1, \ldots, n$, let

$$x_i = \text{ adjusted duration } = \text{ deseasonalized } \Delta t_i = \frac{\Delta t_i}{d(t_i)}$$

where $d(t_i)$ is the estimated (deterministic) diurnal cyclical component of Δt_i.

- A common way to model cyclical component is via regression splines (after logging to ensure a positive solution):

$$\log[d(t)] = g(t), \qquad g(t) = \beta_0 + \sum_{j=1}^{J} \beta_j g_j(t)$$

  The appropriate formulation of the basis functions $g_j(\cdot)$ can be very dataset-specific. Typical functional forms are:

$$g_j(t) = \left(\frac{t - a_j}{b_j}\right)^2, \qquad \text{and} \qquad g_j(t) = \begin{cases} \left(\frac{t-a_j}{b_j}\right)^2, & \text{if } t \in (\alpha_j, \beta_j) \\ 0, & \text{otherwise} \end{cases}$$

  for some specified values of $\{a_j, b_j, \alpha_j, \beta_j\}$.

- The coefficients $\{\beta_j\}$ are fitted by least squares minimization:

$$\sum_i [z_i - g(t_i)]^2, \qquad z_i = \log(\Delta t_i)$$

  In practice obtain the $\{x_i\}$ by exponentiating the residuals from this fit:

$$x_i = \exp\{\hat{\varepsilon}_i\}, \qquad \hat{\varepsilon}_i = z_i - \widehat{g}(t_i)$$

- The default method in `ACDm` function "diurnalAdj" implements cubic splines, but several other methods for obtaining the $\{x_i\}$ are available.

ACD model step

The ACD model uses the GARCH idea to study the dynamic structure of x_i via

$$\mu_i = \mathbb{E}(x_i|F_{i-1})$$

- The basic ACD(r, s) model is defined as:

$$\begin{aligned} x_i &= \mu_i \epsilon_i \\ \{\epsilon_i\} &\sim \text{iid with } \mathbb{E}(\epsilon_i) = 1 \text{ and } \epsilon_i > 0 \\ \mu_i &= \omega + \sum_{j=1}^{r} \alpha_j x_{i-j} + \sum_{j=1}^{s} \beta_j \mu_{i-j} \end{aligned}$$

- Defining $\eta_i = x_i - \mu_i$, which can be shown to be white noise, model can be written as:

$$x_i = \sum_{j=1}^{\max(r,s)} (\alpha_j + \beta_j) x_{i-j} + \sum_{j=1}^{s} \beta_j \eta_{i-j} + \eta_i$$

  (where $\alpha_j = 0$ and $\beta_j = 0$ for $j > r$ and $j > s$, respectively).

- Can show that (under stationarity):

$$\mathbb{E}(x_i) = \frac{\omega}{1 - \sum_{j=1}^{m} (\alpha_j + \beta_j)}, \qquad m = \max(r, s)$$

  (Both numerator and denominator must be positive, since we need $\mathbb{E}(x_i) > 0$.)

- The $\{\epsilon_i\}$ must be modeled with a positive-valued distribution; common versions are:

  EACD: ACD with $\{\epsilon_i\}$ iid standard exponential;
  GACD: ACD with $\{\epsilon_i\}$ iid standard gamma (generalized);
  WACD: ACD with $\{\epsilon_i\}$ iid standard weibull;

Estimation step

This is via the usual MLE since the likelihood of the data vector $\boldsymbol{x}_n$ can be written as:

$$f(\boldsymbol{x}_n|\boldsymbol{\theta}) = \left\{ \prod_{i=m+1}^{n} f(x_i|F_{i-1}, \boldsymbol{\theta}) \right\} f(\boldsymbol{x}_m|\boldsymbol{\theta})$$

The marginal distribution of the first m observations, $f(\boldsymbol{x}_m|\boldsymbol{\theta})$, is usually ignored (implies QMLE).

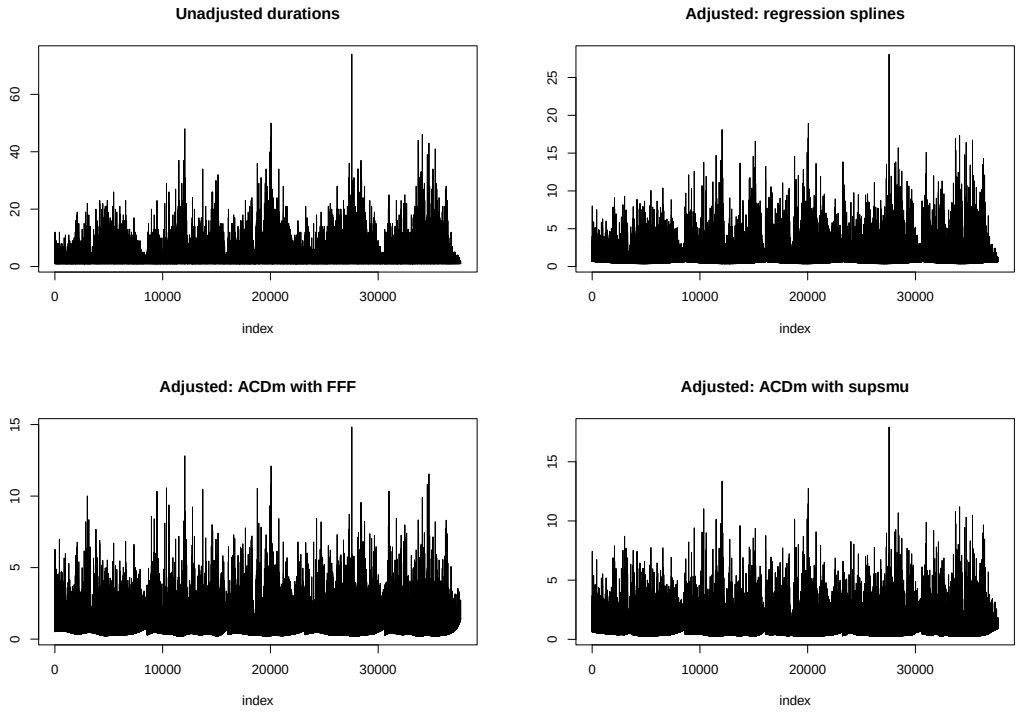
Example 8 (Tsay IAFD, Ex 6.3)**.** *Model the 37,674 positive durations of CAT stock during the 5 trading days from 4 to 8 January, 2010. We convert the data to seconds.*

- *Reasoning that diurnal trading follows approx a quadratic trend peaking at noon, and the pattern repeats daily, we use regression splines to deseasonalize with the functions:*

$$g_1(t) = \frac{t-a}{b}, \qquad \textit{and} \qquad g_2(t) = g_1^2(t),$$

where $a = 43,200$ *denotes 12:00 noon, and* $b = 23,400$ *is the number of trading hours (both measured in seconds). The resulting adjusted durations can be seen in Fig 6. The "FFF" method appears to be the best at detrending; the actual trends can be seen in Fig 7.*

Figure 6: Raw and adjusted durations for CAT data.



- *The ACFs in Fig 8 exhibit substantial autocorrelation; less so for "FFF" and "supsmu". Based on these results we proceed with an ACM model for the FFF adjusted durations.*

Figure 7: Diurnal cycles for CAT durations estimated by "FFF".

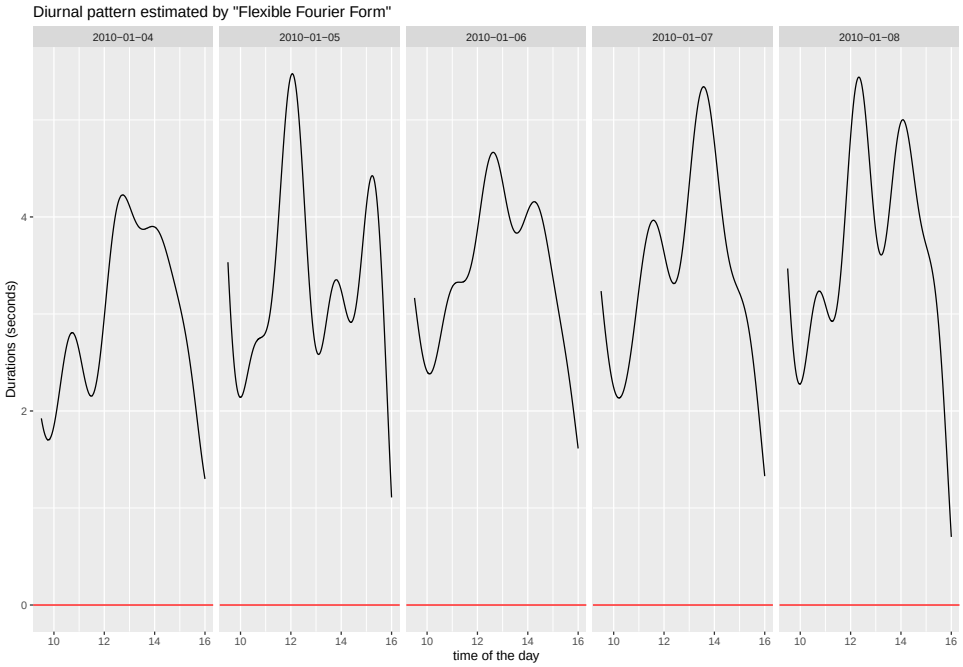
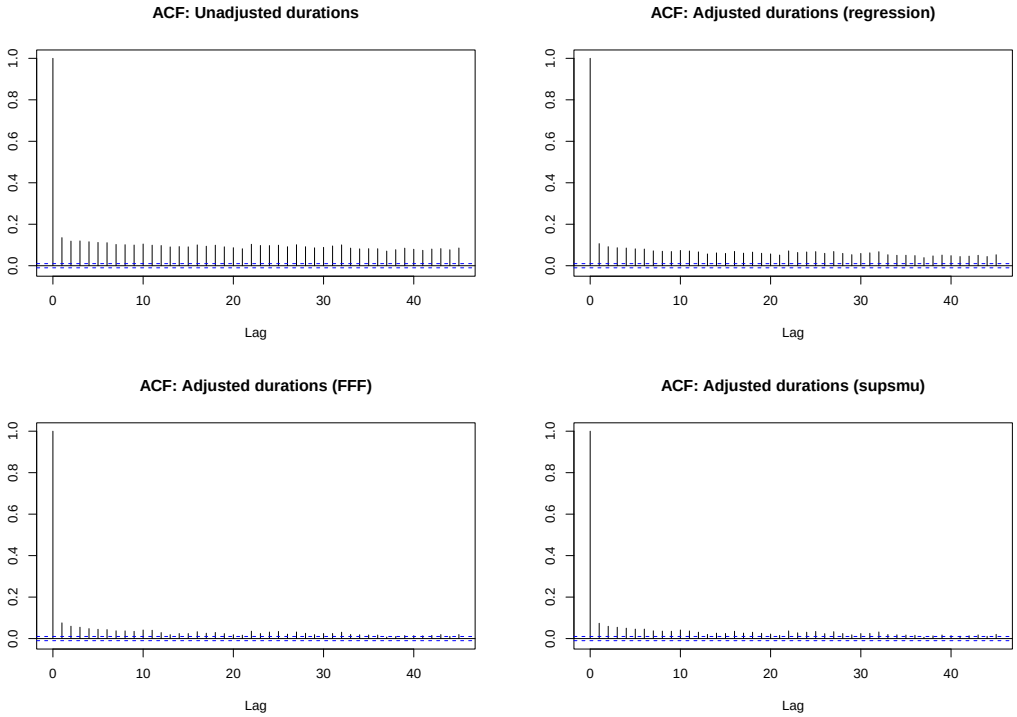


Figure 8: ACFs for unadjusted and adjusted CAT durations.



- *Fitting an EACD(1,1) and a WACD(2,1) models leads to the following goodness of fit summaries:*

```
Goodness of fit:       EACD(1,1)     WACD(2,1)
AIC                 74694.515031   71607.34172
BIC                 74720.125446   71650.02574
MSE                     0.922094       0.92184
Q(5) res p-value        0.051237       0.30222
Q(10) res p-value       0.046969       0.41184
Q(15) res p-value       0.046969       0.41184
Q(5) res^2 p-value      0.886437       0.61705
Q(10) res^2 p-value     0.209134       0.11299
Q(15) res^2 p-value     0.209134       0.11299
```

  The WACD has substantially lower AIC/BIC, and also easily passes tests for serial correlation in the residuals and their squares.

- *The fitted model is:*

```
Parameter estimate:
          Coef      SE PV
omega   0.0296 0.00298  0
alpha1  0.0688 0.00529  0
alpha2 -0.0267 0.00562  0
beta1   0.9291 0.00483  0
gamma   1.2425 0.00454  0

Note: The p-value for the distribution parameter gamma is from the
      2-tailed test H0: gamma = 1.

The fixed/unfree mean distribution parameter: theta: 0.9170524
```

  From this output we see that:

$$\begin{aligned} x_i &= \mu_i \epsilon_i \\ \{\epsilon_i\} &\sim \textit{iid std. Weibull}(\gamma) \\ \mu_i &= 0.0296 + 0.0688 x_{i-1} - 0.0267 x_{i-2} + 0.9291 \mu_{i-1} \end{aligned}$$

  The pdf of a std. Weibull(γ) is:

$$f(\epsilon) = \theta\gamma\epsilon^{\gamma-1}\exp\{-\theta\epsilon^{\gamma}\}, \qquad \theta = [\Gamma(1+1/\gamma)]^{\gamma}$$

  and in this case we have the estimates: $\gamma = 1.2425$ and $\theta = 0.9170524$.

- *Fig 9 shows the ACF of residuals and their squares for the WACD model. The bottom panel displays the estimates of the conditional mean series $\{\mu_i\}$ superimposed on the $\{x_i\}$.*

Figure 9: ACF of residuals and their squares for WACD model fitted to CAT durations. The bottom panel displays the data and estimated conditional mean.