ECN 6578A, Économétrie des marchés financiers, Hiver 2020

Cours 11 : données de haute fréquence (cont.)

William McCausland

2020-03-23

Durées et comptes

Durées (durations)

- Durées ordinaires : durée entre transactions
- Price duration : durée entre deux transactions avec un écart de prix minimal de (mettons) 1/8 \$.
- Volume duration : durée entre deux transactions où la volume cumulative entre elles est plus grande qu'un seuil de (mettons) 25000.

Comptes (counts)

- Nombre de transactions par période de (mettons) 5 minutes.

Pourquoi?

- Les deux sont reliés aux questions de liquidité, price discovery, microstructure.

Données : ibm.txt (du 1 nov. 1990 au 31 janv 1991)

datetime	volume	bid	ask	price
90110134228	18800	105.3750	105.5000	105.3750
90110134236	400	105.3750	105.5000	105.3750
90110134236	100	105.3750	105.5000	105.3750
90110134237	1000	105.2500	105.6200	105.3750
90110134242	100	105.3750	105.5000	105.3750
90110134246	100	105.3750	105.5000	105.3750
90110134308	1000	105.1250	105.7500	105.5000
90110134310	2000	105.1250	105.7500	105.3750
90110134322	100	105.1250	105.7500	105.3750
90110134344	5000	105.1250	105.7500	105.5000

Taux d'incidence (taux de défaillance, hazard rate) I

- Soit T une variable aléatoire pour une durée.
- Exemples
  - temps entres transactions
  - durée de chômage
  - vie d'une pièce mécanique
- Fonction de répartition :

$$F(t) = \Pr[T \leq t]$$

- Fonction de densité :

$$f(t) = \frac{dF(t)}{dt}$$

- Fonction de survie :

$$S(t) = (1 - F(t)) = \Pr[T > t]$$

Taux d'incidence II

- Définition du taux d'incidence

$$h(t) \equiv \lim_{\Delta t \downarrow 0} \frac{\Pr[T \leq t + \Delta t | T > t]}{\Delta t} = \lim_{\Delta t \downarrow 0} \frac{\Pr[t < T \leq t + \Delta t]}{\Delta t \Pr[T > t]}$$

$$h(t) = \lim_{\Delta t \downarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t S(t)} = \frac{f(t)}{S(t)}.$$

- Taux d'incidence pour une variable aléatoire exponentielle

$$F(t) = 1 - e^{-\lambda t} \quad f(t) = \lambda e^{-\lambda_t} \quad S(t) = e^{-\lambda t}$$

$$h(t) = f(t)/S(t) = \lambda.$$

Distribution Weibull

- Fonction de répartition : $F(x|\alpha, \beta) = 1 - e^{-(x/\beta)^\alpha}$
- Fonction de densité : $f(x|\alpha, \beta) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-(x/\beta)^\alpha}$
- Taux d'incidence : $h(x|\alpha, \beta) = \frac{f(x|\alpha,\beta)}{1-F(x|\alpha,\beta)} = \frac{\alpha}{\beta^\alpha} x^{\alpha-1}$
- Le taux d'incidence est
  - constant, pour $\alpha = 1$
  - infini à $x = 0$ et décroissant, pour $\alpha < 1$
  - nul à $x = 0$ et croissant, pour $\alpha > 1$.

Propriété sans mémoire des variables aléatoires exponentielles

- En général :

$$\Pr[T \leq t+\tau | T > t] = \frac{\Pr[t < T \leq t+\tau]}{\Pr[T > t]} = \frac{F(t+\tau) - F(t)}{1 - F(t)}.$$

- Cas particulier exponentiel :

$$\Pr[T \leq t+\tau | T > t] = \frac{(1 - e^{-\lambda(t+\tau)}) - (1 - e^{-\lambda t})}{e^{-\lambda t}} = 1 - e^{-\lambda\tau}$$

- Caractère sans mémoire :

$$\Pr[T \leq t+\tau | T > t] = \Pr[T \leq \tau] = F(\tau).$$

Intuition pour les taux d'incidence

1. Autobus tous les 15 minutes, pile
2. Grand nombre d'autobus, grand cycle, 15 minutes en moyenne
3. Horaire incertain : autobus tous les 10 minutes ou tous les 20 minutes, chaque horaire avec probabilité $1/2$.
4. Pièce mécanique : (pièce ratée, usure)

Durées, comptes

- Approximation de première ordre :
  - Beaucoup d'acheteurs et de vendeurs potentiels
  - Chacun a une probabilité faible d'initier un échange à un moment donné
  - Les arrivés des investisseurs sont indépendantes
  - Selon ces hypothèses, le taux d'incidence est constant (théorie des files d'attentes)
  - Implication : mettons que le taux d'incidence est de λ échanges par heure
    - le nombre d'échanges dans x heures est $\mathrm{Po}(\lambda x)$ (Poisson)
    - la durée entre échanges est $\mathrm{Exp}(\lambda)$ (Exponentiel)
- Une réalité plus complexe :
  - Régularité diurne (plus d'échanges au début et à la fin des heures d'ouverture)
  - Échanges synchronisés par des nouvelles (moment annoncé ou non)
  - Transactions multiples
  - Une transaction importante comme une nouvelle en soi

Notation pour les heures de transaction est les durées

- t_i est l'heure de la i-ième transaction, en s après minuit.
- $\Delta t_i = t_i - t_{i-1}$, une durée en s.
- Le marché ouvre à 9h30, $9.5 \times 60 \times 60 = 34200s$ après minuit.
- Le marché ferme à 16h00, $16 \times 60 \times 60 = 57600s$ après minuit.
- On utilise l'application suivante pour créer une variable temporelle dans l'intervalle $[-1, 1]$:

$$\tau_i = \frac{2t_i - 34200 - 57600}{57600 - 34200}.$$

- Si $t_i = 57600$, $\tau_i = 1$ et si $t_i = 34200$, $\tau_i = -1$.

Une décomposition des durées

- On décompose $\Delta t_i = f(t_i)\Delta t_i^*$
  - $f(t_i)$ est une fonction déterministe pour capturer la régularité diurne.
  - Δt_i^* est une durée transformée.
  - Le processus Δt_i^* est plus plausiblement stationnaire que le processus Δt_i.
- La décomposition est additive en logs :

$$\ln \Delta t_i = \ln f(t_i) + \ln \Delta t_i^*.$$

- Nous trouverons une approximation en polynômes du type Legendre :

$$\ln f(t_i) = \sum_{k=0}^{K} \beta_k P_k(\tau_i)$$

- Une fois que la fonction $f(t_i)$ est estimée, on calcule

$$\Delta t_i^* = \Delta t_i / f(t_i).$$

Polynômes de Legendre

- Site Wikipédia « polynôme de Legendre »
- Un système de polynômes sur l'intervalle $[-1, 1]$.
- L'orthogonalité est utile pour la régression linéaire.
- Utile pour l'approximation d'une fonction sur un intervalle
- Les premiers :

$$P_0(x) = 1$$

$$P_1(x) = x$$

$$P_2(x) = (3x^2 - 1)/2$$

$$P_3(x) = (5x^3 - 3x)/2.$$

Estimation de $\ln f(\cdot)$ par régression linéaire

- La transformation en logs, encore :

$$\ln \Delta t_i = \ln f(t_i) + \ln \Delta t_i^*.$$

- Idéalement, le processus $\ln \Delta t_i^*$ ne dépend pas de t_i.
- Si $\ln f(t_i)$ capture bien la dépendance de $\ln \Delta t_i$ de t_i,

$$E[\ln \Delta t_i^* | \ln f(t_i)] = 0.$$

- On estime les coefficients de la régression linéaire suivante :

$$\ln \Delta t_i = \sum_{k=0}^{K} \beta_k P_k(\tau_i) + \ln \Delta t_i^*.$$

- On observe t_i, $i = 1, \ldots, N$, on calcule τ_i, $P_k(\tau_i)$, $\ln \Delta t_i$.
- N est le nombre d'observations et K est le nombre de termes de l'expansion.

Code en R préliminaire

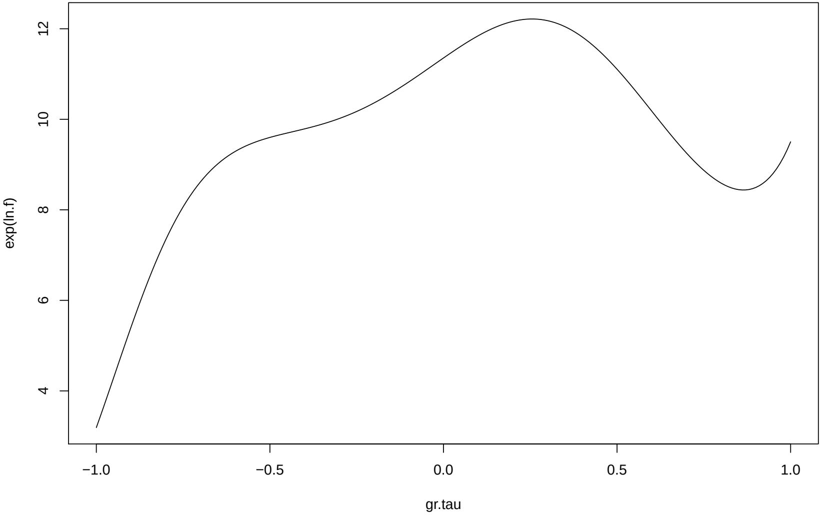
```
t = read.table('ibm.txt')
n = nrow(t)
colnames(t) = c('datetime','volume','bid','ask','price')
t$time = t$datetime %% 100000
t$date = t$datetime %/% 100000
t$dur = pmax(0, c(0,diff(t$time)))
t = t[t$dur >= 0,]
t = t[t$time <= 57600,]
t$dur = pmax(0.5, t$dur)
tau = (2*t$time-(34200 + 57600))/(57600-34200)
L1 = tau
L2 = 0.5*(3*tau^2-1)
L3 = 0.5*(5*tau^3-3*tau)
L4 = 0.125*(35*tau^4-30*tau^2+3)
L5 = 0.125*(63*tau^5-70*tau^3+15*tau)
lndur = log(t$dur)
lm.diurnal = lm(lndur ~ L1 + L2 + L3 + L4 + L5)
```

Construction de $f(t_i)$

```
gr.tau = seq(-1,1,by=0.01)
gr.L1 = gr.tau
gr.L2 = 0.5*(3*gr.tau^2-1)
gr.L3 = 0.5*(5*gr.tau^3-3*gr.tau)
gr.L4 = 0.125*(35*gr.tau^4-30*gr.tau^2+3)
gr.L5 = 0.125*(63*gr.tau^5-70*gr.tau^3+15*gr.tau)
beta = lm.diurnal$coefficients
ln.f = beta[1] + beta[2]*gr.L1 + beta[3]*gr.L2 + beta[4]*gr
t$ln.f = beta[1] + beta[2]*L1 + beta[3]*L2 + beta[4]*L3 + b
t$dur.et = t$dur / exp(t$ln.f)
```

Cycle diurne estimé

```
plot(gr.tau, exp(ln.f), type='l')
```



Un modèle pour les durées transformées

- Le modèle ACD (Autoregressive Conditional Duration) capture l'aspect dynamique des durées transformées $x_i \equiv \Delta t_i^*$.

$$x_i = \psi_i \epsilon_i$$

$$\psi_i = \omega + \sum_{j=1}^{r} \gamma_j x_{i-j} + \sum_{j=1}^{s} \omega_j \psi_{i-j}.$$

  où ϵ_i est un aléa tel que $E[\epsilon_i] = 1$, comme un exponentiel où un Weibull.
- Rappel : l'exponentiel est sans mémoire, avec un taux de défaillance constant

Modèle Probit ordonné pour les changements de prix : variables pertinentes

- t_i est l'heure du i-ième échange
- $P^*_{t_i}$ est le prix virtuel d'un actif à t_i
- y^*_i est le changement de prix virtuel entre t_{i-1} et t_i
- P_{t_i} est le prix observé d'un actif à t_i.
- y_i est le changement de prix observé entre t_{i-1} et t_i.
- x_i et w_i sont des vecteurs de variables explicatives

Un modèle pour les changements de prix virtuels

- Le prix virtuel n'est pas un « vrai » prix latent comme avant
- Le modèle pour les changements de prix est

$$y_i^* = x_i\beta + \epsilon_i,$$

où

  - x_i est un vecteur $1 \times k$ de variables explicatives,
  - β est un vecteur $k \times 1$ de coefficients,
  - $E[\epsilon_i|x_i] = 0$, $\mathrm{Cov}[\epsilon_i, \epsilon_j]|x_i, x_j] = 0$.
  - $\mathrm{Var}[\epsilon_\mathrm{i}|\mathrm{x_i}, \mathrm{w_i}] = \sigma_\mathrm{i}^2 = \mathrm{g}(\mathrm{w_i})$,
  - w_i est un vecteur de variables explicatives,
  - Souvent, $\epsilon_i|x_i, w_i \sim N(\mu, \sigma_i^2)$.

Une transformation en changements de prix observés

- On suppose que y_i prend des valeurs discrètes :

$$y_i \in \{s_1, \ldots, s_k\}$$

- Exemple: $k = 7$, $y_i \in \{-3/8, -1/4, -1/8, 0, 1/8, 1/4, 3/8\}$
- La relation entre y_i et y_i^* est

$$y_i = \begin{cases} s_1 & -\infty < y_i^* \leq \alpha_1 \\ s_2 & \alpha_1 < y_i^* \leq \alpha_2 \\ \vdots & \\ s_k & \alpha_{k-1} < y_i^* < \infty. \end{cases}$$

- $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ est un vecteur de paramètres inconnus.

Élimination des y_i^*, cas gaussien

Dans le cas où $\epsilon_i | x_i, w_i$ est gaussien,

$$\Pr[y_i = s_j | x_i, w_i] = \begin{cases} \Phi\left(\frac{\alpha_1 - x_i\beta}{\sigma_i(w_i)}\right) & j = 1 \\ \Phi\left(\frac{\alpha_j - x_i\beta}{\sigma_i(w_i)}\right) - \Phi\left(\frac{\alpha_{j-1} - x_i\beta}{\sigma_i(w_i)}\right) & j = 2, \ldots, k-1 \\ 1 - \Phi\left(\frac{\alpha_{k-1} - x_i\beta}{\sigma_i(w_i)}\right) & j = k \end{cases}$$

- Notes
    - Les valeurs de α_i influent l'asymétrie, l'aplatissement conditionnelle
    - La variabilité de σ_i influe l'aplatissement inconditionnelle

Des variables dépendantes prometteures

- x_i peut inclure
  - la durée transformée Δt_i^*
  - des valeurs retardées y_{i-m}, $m = 1, 2, \ldots$
  - dernier rendement 5-minutes de l'index S&P 500
  - la volume du dernier échange
  - index cours acheteur-cours vendeur IBS_{i-m}, $m = 1, 2, \ldots$ (1 si P_{i-m} est plus près du cours vendeur, -1 si plus près du cours acheteur, 0 sinon)
- w_i peut inclure
  - Δt_i^*
  - écart cours acheteur-cours vendeur AB_i

Résultats empiriques I

- Données : IBM, $n = 206794$, transaction avec volume > 100
- Papier : Hauseman et al. (1992)
- Estimation des α_i

α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
-4.67	-4.16	-3.11	-1.34	1.33	3.13	4.21	4.73

- Remarques :
  - asymétrie négative, pas fort
  - aplatissment élevé

Résultats empiriques II

- Moyenne conditionnelle $x_i\beta$

x_{ik}	coeff
Δt_i^* (100s)	-0.12
y_{i-1} (ticks)	-1.01
y_{i-2}	-0.53
y_{i-3}	-0.21
IBS_{i-1}	-1.14
IBS_{i-2}	-0.37
IBS_{i-3}	-0.1

- Remarques
  - renversements des changements de prix

Résultats empiriques III

- Variance conditionnelle $\sigma_i^2(w_i) = 1 + \gamma_1^2 \Delta t_i^* + \gamma_2^2 \mathrm{AB}_{i-1}$

w_{ik}	coeff
1	1
Δt_i	0.40^2
AB_{i-1}	0.52^2

- Remarques
  - La constante 1 est une normalisation
  - Δt_i^* indique et le passage du temps et la tranquilité (résultat anticipé ambigue)
  - écart cours acheteur-cours vendeur comme prédicteur de volatilité (attention : interpretation causale douteuse)