Financial Data: Modeling and Analysis

Ka Wai Tsang

December 19-20, 2015

Outline

Univariate time series

- A time series $\{x_1, x_2, x_3, \ldots\}$ is said to be weakly stationary (or covariance stationary) if Ex_t and $\mathrm{Cov}(x_t, x_{t+k})$ do not depend on $t \geq 1$.
- For a weakly stationary sequence x_t, $\mu := Ex_t$ is called its mean and $\gamma_h := \mathrm{Cov}(x_t, x_{t+h})$, $h \geq 0$, is called the autocovariance function. Note that the correlation coefficient $\mathrm{Corr}(x_t, x_{t+h})$ at lag h also does not depend on t and is given by $\rho_h := \gamma_h/\gamma_0, h \geq 0$, which is called the autocorrelation function (ACF) of $\{x_t\}$.
- A sequence $\{x_t\}$ is said to be stationary (or strictly stationary) if the joint distribution of $(x_{t+h_1}, \ldots, x_{t+h_m})$ does not depend on $t \geq 1$ for every $m \geq 1$ and $0 \leq h_1 < \cdots < h_m$.

Tests of independence

- Estimate γ_h and ρ_h by

$$\hat{\gamma}_h = \frac{1}{n-h}\sum_{t=1}^{n-h}(x_{t+h}-\bar{x})(x_t-\bar{x}), \quad \hat{\gamma}_h/\hat{\gamma}_0$$

- Suppose x_t are i.i.d. and $Ex_t^2 < \infty$. Then $\rho_h = 0$ and $\hat{\rho}_h$ is asymptotically $N(0, 1/n)$ as $n \to \infty$ for any fixed $h \geq 1$, and $\hat{\rho}_1, \ldots, \hat{\rho}_m$ are asymptotically independent.
- To test the null hypothesis $\rho_1 = \cdots = \rho_m = 0$, we use
  - Box-Pierce statistic $Q^*(m) = n\sum_{h=1}^m \hat{\rho}_h^2$, or
  - Ljung-Box statistic $Q(m) = n(n+2)\sum_{h=1}^m \hat{\rho}_h^2/(n-h)$.

  Both are asymptotically χ_m^2 as $n \to \infty$ when the x_t are i.i.d.
- For a moderate sample size m, $Q(m)$ is better approximated by χ_m^2 than $Q^*(m)$. The R function `Box.test` can be used to compute the Box-Pierce or Ljung-Box statistic for the null hypothesis.

Moving average model

- (Wold decompostion): If $\sum_{h=0}^{\infty} |\gamma_h| < \infty$, then a weakly stationary sequence $\{x_t\}$ can be expressed as

$$x_t = \mu + u_t + \sum_{j=1}^{\infty} \psi_j u_{t-j},$$

in which u_t are uncorrelated random variables with $Eu_t = 0$ and $\text{Var}(u_t) = \sigma^2$.

- For a given positive integer q, we consider a moving average model $MA(q)$,

$$x_t = \mu + u_t + \sum_{j=1}^{q} \psi_j u_{t-j}$$

Autoregressive model

- Let B denote the backshift operator defined by $Bx_t = x_{t-1}$, then the Wold decomposition can be rewritten as $x_t = \mu + \psi(B)u_t$, where $\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \ldots$.
- By replacing x_t by $x_t - \mu$, we have

$$u_t = \frac{1}{\psi(B)} x_t = \phi(B)x_t,$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots$ is the Taylor expansion of $1/\psi(B)$, which is valid if the zeros of $\psi(z)$ are outside the unit circle.

- Approximating the power series $\phi(B)$ by $1 - \sum_{i=1}^{p} \phi_i B^i$ yields a autoregressive model $AR(p)$

$$x_t = \mu + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + u_t, \ E(u_t) = 0.$$

Stationary and ARMA model

- The stationarity condition for $AR(p)$ is: the zeros of $\Phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ are outside the unit circle.
- We can combine $AR(p)$ and $MA(q)$ to form a ARMA(p,q) model:

$$x_t = \mu + \sum_{i=1}^{p} \phi_i x_{t-i} + u_t + \sum_{j=1}^{q} \psi_j u_{t-j}.$$

Forecasting in ARMA models

Assume $x_0 = \cdots = x_{-p+1} = 0 = u_0 = \cdots = u_{-q+1}$.

- One-step-ahead forecast:

$$\hat{x}_{t+1|t} = \mu + \sum_{i=1}^{p} \phi_i x_{t+1-i} + \sum_{j=1}^{q} \psi_j u_{t+1-j}.$$

- h-step-ahead forecast:

$$\hat{x}_{t+h|t} = \mu + \sum_{i=1}^{p} \phi_i x_{t+h-i|t} + \sum_{j=1}^{q} \psi_j u_{t+h-j|t}.$$

Parameter estimation and order determination

- Consider the $ARMA(p,q)$ model with $u_t \sim N(0,\sigma^2)$. Let $\boldsymbol{\theta} = (\mu, \phi_1, \ldots, \phi_p, \psi_1, \ldots, \psi_q)^T$. The log-likelihood function is given by

$$\ell(\boldsymbol{\theta}, \sigma) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(x_t - \mu - \sum_{i=1}^{p}\phi_i x_{t-i} - \sum_{j=1}^{q}\psi_j u_{t-j})^2,$$

  in which $u_1, \ldots, u_{n-1}$ can be computed recursively for a given $\boldsymbol{\theta}$ under the initial condition
  $x_0 = \cdots = x_{-p+1} = 0 = u_0 = \cdots = u_{-q+1}$.

- The order (p,q) can be chosen by
  - $AIC(d) = -2\ell(\hat{\boldsymbol{\theta}}, \hat{\sigma}) + 2d$,
  - $BIC(d) = -2\ell(\hat{\boldsymbol{\theta}}, \hat{\sigma}) + d\log n$,

  where n is the sample size, $d = p + q + 1$.

Unit-root nonstationary

- A time series y_t is said to be unit-root nonstationary if it is nonstationary but $\Delta y_t = y_t - y_{t-1}$ is weakly stationary.
- Spurious regression: Assume $y_t = \alpha + \beta x_t + \varepsilon_t$. OLS estimator $\hat{\alpha}$ and $\hat{\beta}$ may be non-consistent estimate when x_t and y_t are unit-root non-stationary (Granger and Newbold (1974)).

Tests for stationarity

- Consider an AR(p) model for $\{x_t\}$ with $\Phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$.
- If $\Phi(1) = 0$, it is unit-root non-stationary. Rewrite x_t as $x_t = \mu + \beta_1 x_{t-1} - \sum_{j=1}^{p-1} \beta_{j+1} \Delta x_{t-j} + u_t$, where $\beta_j = \sum_{i=j}^{p} \phi_i$ and $\Delta x_t = x_t - x_{t-1}$. By definition, $\Phi(1) = 0$ if and only if $\beta_1 = 1$. Therefore, we can test if it is unit-root non-stationary by testing the null hypothesis $H_0 : \beta_1 = 1$.
  - augmented Dickey-Fuller (ADF) test: assumes u_t's are i.i.d. The `R` function is `adf.test`.
  - Phillips-Perron (PP) test: u_t's can be correlated. The `R` function is `PP.test`.
- KPSS Test for stationarity: The `R` function is `kpss.test`.

Macroeconomic time series modeling

- Suppose we want to understand the relation among interest rates, industrial production index and consumer price index to predict if the Federal Reserve will adjust the current interest rates.
  - Industrial production index (IPI) measures the amount of output from the manufacturing, mining, electric and gas industries.
  - Consumer price index (CPI) measures changes in the price level of a market basket of consumer goods and services purchased by households.
- One of the most successful and flexible models for analysis of multivariate time series is the vector autoregression (VAR) method. Let $\mathbf{y}_t \in \mathbb{R}^q$ be the vector to be forecast, we assume

$$\mathbf{y}_t = \mathbf{a}_0 + \sum_{s=1}^{l} \mathbf{A}_s^T \mathbf{y}_{t-s} + \boldsymbol{\epsilon}_t,$$

where $\mathbf{a}_0 \in \mathbb{R}^q$ and $\mathbf{A}_s \in \mathbb{R}^{q \times q}$ are parameter matrices to be estimated. The errors $\boldsymbol{\epsilon}_t$ is $\mathcal{F}_{t-1}$-measurable (i.e. depends on the information set $\mathcal{F}_{t-1}$ up to time $t-1$).

Price puzzle

- Bernanke et al. (2005) fitted a VAR with three variables, IPI, CPI, and the federal funds rate, with $l = 5$ lags. They tested the model by giving a positive impulse to the federal funds rate and computed their impulse response functions.
- They observed that CPI increased after the impulse, rather than a decrease as standard economic theory would predict. They think that the puzzle was because there was information related to those three variables that had not been involved in the VAR model.
- They introduce factor-augmented VAR (FAVAR) models to solve the puzzle.

Limitation of standard VAR

- A major difficulty with VAR is that it has too many parameters relative to the size of the training sample. Because of possible structural changes over long periods of time, the length of the time series (and therefore the number of observations) used to fit a VAR model is not large, especially when there are only 12 observations per year for monthly data.)
- Bernanke et al. (2005) commented that "to conserve degrees of freedom, standard VARs rarely employ more than six to eight variables. This small number of variables is unlikely to span the information sets used by actual central banks, who are known to follow literally hundreds of data series."
- Recall that the number of parameters is of $O(q^2)$.

VAR models with exogenous variables (VARX)

- Let $\mathbf{Y}_t = [\mathbf{y}_t, \ldots, \mathbf{y}_{t-n+1}]^T \in \mathbb{R}^{n \times q}$ be the responses matrix and we assume it is related to its past values by

$$\mathbf{Y}_t = \mathbf{1}\mathbf{a}_0^T + \sum_{s=1}^{l} \mathbf{Y}_{t-s}\mathbf{A}_s + \mathbf{X}_{t-1}\mathbf{B} + \mathbf{E}_t.$$

  When $\mathbf{B} = \mathbf{0}$, then it is the VAR model.

- The number p of variables in $\mathbf{X}$ can be larger than the number n of samples. We want to reduce the dimension of $\mathbf{X}$ to a much smaller value r and rewrite the model as

$$\mathbf{Y}_t = \mathbf{1}\mathbf{a}_0^T + \sum_{s=1}^{l} \mathbf{Y}_{t-s}\mathbf{A}_s + \mathbf{F}_{t-1}\mathbf{B}_f + \mathbf{E}_t,$$

  where $\mathbf{F}_{t-1} \in \mathbb{R}^{n \times r}$ is an unknown factor matrix that is constructed from the input matrix $\mathbf{X}_{t-1} \in \mathbb{R}^{n \times p}$.

Factor Augmented Autoregression (FAAR)

- Stock and Watson (2002) first propose the idea of using principal components from a large number of predictors in a VAR model. The principal components estimator solves the least squares problem $\min_{\mathbf{F},\mathbf{\Lambda},\mathbf{F}^T\mathbf{F}=\mathbf{I}_{\hat{r}}} V_{\hat{r}}(\mathbf{F},\mathbf{\Lambda})$, where

$$V_{\hat{r}}(\mathbf{F},\mathbf{\Lambda}) = \|\mathbf{X}_{t-1} - \mathbf{F}\mathbf{\Lambda}^T\|_F^2/(np),$$

  $\mathbf{F} =\in \mathbb{R}^{n\times\hat{r}}$ and $\mathbf{\Lambda} \in \mathbb{R}^{p\times\hat{r}}$.

- Bai and Ng (2002) propose an information criterion to choose $\hat{r} = \min_h \mathrm{IC}_{BN}(h)$ such that $\hat{r}$ is a consistent estimator of the true number r of underlying factors, where

$$\mathrm{IC}_{BN}(h) = np \log V_h(\hat{\mathbf{F}}_h, \hat{\mathbf{\Lambda}}_h) + hc(n+p)\log\left(\frac{np}{n+p}\right)$$

  and $(\hat{\mathbf{F}}_h, \hat{\mathbf{\Lambda}}_h)$ is the minimizer of $V_h(\mathbf{F},\mathbf{\Lambda})$.

K-fold cross validation to choose c

Given a data set $\mathbf{X}$ and a particular c, we do the followings.

1. The original sample $\mathbf{X}$ is randomly partitioned into K equal sized subsamples. Of the K subsamples, a single subsample, say the i-th subsample, is retained as the validation data for testing the model, denote as $\mathbf{X}_{test}(i)$, and the remaining $k-1$ subsamples are used as training data, , denote as $\mathbf{X}_{train}(i)$.
2. For $i = 1, \ldots, K$,
    1. Estimate $\hat{r}(i, c)$ based on $\mathbf{X}_{train}(i)$ and the given c.
    2. Compute $V_{\hat{r}(i,c)}(\hat{\mathbf{F}}, \hat{\mathbf{\Lambda}})$ based on $\mathbf{X}_{test}(i)$
3. Output $CV(c) = \frac{1}{K} \sum_{i=1}^{K} V_{\hat{r}(i,c)}(\hat{\mathbf{F}}, \hat{\mathbf{\Lambda}})$

We choose c such that $CV(c)$ is minimum.

Reduce Rank Regression

- The solution of reduced-rank regression is

$$\mathbf{B}(h) = \arg \min_{\mathbf{B} \in \mathbb{R}^{p \times q}, \mathrm{r}(\mathbf{B}) \leq h} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2$$

- $\mathbf{B}(h) = \widehat{\mathbf{C}}\widehat{\mathbf{A}}$,

$$\widehat{\mathbf{C}} = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}(\mathbf{b}_1, \ldots, \mathbf{b}_r), \quad , \widehat{\mathbf{A}} = (\mathbf{b}_1, \ldots, \mathbf{b}_r)^T,$$

  where $\mathbf{S}_{xx} = n^{-1}\sum_{k=1}^{n} \mathbf{x}_k\mathbf{x}_k^T$, $\mathbf{S}_{yy} = n^{-1}\sum_{k=1}^{n} \mathbf{y}_k\mathbf{y}_k^T$, $\mathbf{S}_{xy} = n^{-1}\sum_{k=1}^{n} \mathbf{x}_k\mathbf{y}_k^T$, $\mathbf{S}_{yx} = \mathbf{S}_{xy}$ and $(\mathbf{b}_1, \ldots, \mathbf{b}_r)$ are the first r singular vectors of $\mathbf{H} = \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$.

- Let $\mathbf{P}_r = (\mathbf{b}_1, \ldots, \mathbf{b}_r)(\mathbf{b}_1, \ldots, \mathbf{b}_r)^T$. Note that

$$\mathbf{B}(h) = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{P}_r = \widehat{\mathbf{B}}^{OLS}\mathbf{P}_r.$$

Forecasting using targeted predictors

- When most of the columns in **X** are irrelevant to the response **Y**, the first few principal components of **X** may not be a good predictors of **Y**.
- Bai and Ng (2008) forest Consumer Price Index (CPI) based on 132 market variables. They find improvements at all forecast horizons over the FAAR models by estimating the factors using fewer but informative predictors. Their findings suggest that variable selection should be done before applying FAAR.
- They apply hard and soft thresholding to select informative predictors. In particular, for soft thresholding methods, they apply the Lasso, the elastic net and the least angle regression (LARs).

Group Lasso

- Consider the minimization problem

$$\min_{\mathbf{B}} \|\mathbf{Y} - \mathbf{XB}\|_F^2 + \lambda \|\mathbf{B}\|_{2,1},$$

where $\|\mathbf{B}\|_{2,1} = \sum_{i=1}^{p} \|\mathbf{b}_i^T\|_2$ is the sum of the 2-norm of the rows of $\mathbf{B}$. It is a kind of group lasso problem (Yuan and Lin, 2007).

- Bunea, She and Wegkamp (2012) introduce rank-constrained group lasso (RCGL) to solve the problem

$$\hat{\mathbf{B}}_{rcgl,l} = \min_{r(\mathbf{B}) \leq l} \|\mathbf{Y} - \mathbf{XB}\|_F^2 + \lambda \|\mathbf{B}\|_{2,1}.$$

They also propose Rank Selection Criterion (RSC) to choose rank:

$$\hat{r}^{RSC} = \min_{k} \left\{ \min_{B, r(B)=k} \{\|\mathbf{Y} - \mathbf{XB}\|_F^2 + \mu k\} \right\}$$

, which is equivalent to the number of singular values of $\mathbf{P}_X\mathbf{Y} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{Y}$ that exceed $\mu^{1/2}$.

Group Orthogonal Greedy Algorithm

First initialize with $\mathbf{U}^0 = [\mathbf{U}_v^0]_{v=1}^q = \mathbf{Y}$, $\widehat{I}_0 = \emptyset$, and empty matrices $\mathbf{Q}_0$ and $\mathbf{R}_0$. For $k = 1$ to m do:

1. Choose $\hat{i}_k = \arg\min_{1 \le i \le p} \left(\min_{\beta \in \mathbb{R}^q} \| \mathbf{U}^{k-1} - \mathbf{X}_i \beta^T \|_F^2 \right)$.
2. Update $\widehat{I}_k = \widehat{I}_{k-1} \bigcup \{\hat{i}_k\}$ and compute the QR decomposition

$$\mathbf{X}_{\widehat{I}_k} = [\mathbf{X}_{\widehat{I}_{k-1}} \mathbf{X}_{\hat{i}_k}] = \begin{bmatrix} \mathbf{Q}_{k-1} & \mathbf{q}_k \end{bmatrix} \begin{bmatrix} \mathbf{R}_{k-1} & \vdots \\ 0 \cdots 0 & r_k \end{bmatrix} = \mathbf{Q}_k \mathbf{R}_k.$$

3. Update $\mathbf{U}^k = \mathbf{U}^{k-1} - \mathbf{q}_k \beta_k^T$, where $\beta_k^T = \mathbf{q}_k^T \mathbf{U}^{k-1}$.
4. End for, with $\hat{i}_k$th row of $\widehat{\mathbf{B}} \in \mathbb{R}^{p \times q}$ equal to the kth row of $\mathbf{R}_m^{-1} [\beta_1 \cdots \beta_m]^T$ and the other rows equal to $\mathbf{0}^T$.

Factor construction based on reduced-rank regression after group OGA

- An extension of the HDIC for univariate regression to remove irrelevant variables along the group OGA path.

$$\text{HDIC}(k) = n \log\left(\|\mathbf{U}^k\|_F^2\right) + k w_n \log(p_n),$$

in which different criteria correspond to different choices of w_n.

- The solution of reduced-rank regression is

$$\mathbf{B}(h) = \arg \min_{\mathbf{B} \in \mathbb{R}^{\hat{k}_n \times q}, \mathrm{r}(\mathbf{B}) \le h} \|\mathbf{Y} - \mathbf{X}_S \mathbf{B}\|_F^2$$

after obtaining the selected subset of $\hat{k}_n$ predictors by group OGA.

- The solution $\mathbf{B}(h) = \mathbf{B}_1 \mathbf{B}_2^T$ is a product of 2 matrices of rank h, $\mathbf{B}_1 \in \mathbb{R}^{\hat{k}_n \times h}$ and $\mathbf{B}_2 \in \mathbb{R}^{q \times h}$. We construct factors $\mathbf{F} = \mathbf{X}_S \mathbf{B}_1$.

Rank selection criterion based on reduced-rank regression

- Information criterion

$$\mathrm{IC}_{LT}(h) = nq \log \hat{\sigma}^2(h) + hc(n+q) \log\left(\frac{nq}{n+q}\right)$$

to choose the rank $\hat{r} = \arg\min_h \mathrm{IC}(h)$, where

$$\hat{\sigma}^2(h) = \|\mathbf{Y} - \mathbf{X}_S \mathbf{B}(h)\|_F^2/(nq).$$

- A natural modification of cross-validation for time series data to choose c is the accumulated predictive error (APE) criterion introduced by Rissanen (1986); see also Wei (1992). The idea is to choose c from a grid of values to minimize

$$\mathrm{APE}(c) = \sum_{t=m_0+1}^{n} \|\mathbf{y}_t - \mathbf{B}_{t-1}^T(\hat{r}(c))\mathbf{x}_t^S\|^2,$$

where the transpose of $\mathbf{x}_t^S$ is the tth row of $\mathbf{X}_S$, $\mathbf{B}_s(h)$ is the rank-h estimate based on $\{(\mathbf{x}_i^S, \mathbf{y}_i) : i \le s\}$ and m_0 is the starting sample size for which $\mathbf{B}_{m_0}(\hat{r}(c))$ is uniquely defined for all c belonging to the grid.

Cointegration vectors

- A nonzero $k \times 1$ vector $\mathbf{b}$ is called a cointegration vector of a unit-root nonstationary time series $\mathbf{y}_t$ of length k if $\mathbf{b}^T \mathbf{y}_t$ is weakly stationary.
- A multivariate time series is said to be cointegrated if all its components are unit-root nonstationary and there exists a cointegration vector. If the linear space of cointegrating vectors (with 0 adjoined) has dimension $r > 0$, then the time series is said to be cointegrated with order r.

Cointegration vectors

- Recall a AR(p) model for $\mathbf{y}_t = (y_{t,1}, \ldots, y_{t,k})^T$ can be written as

$$\Delta \mathbf{y}_t = \boldsymbol{\mu} + (\mathbf{B}_1 - \mathbf{I})\mathbf{y}_{t-1} - \sum_{j=1}^{p-1} \mathbf{B}_{j+1} \Delta \mathbf{y}_{t-j} + \boldsymbol{\varepsilon}_t,$$

  where $\mathbf{B}_j = \sum_{i=j}^{p} \boldsymbol{\Phi}_i$. Letting $\boldsymbol{\Phi}(z) = \mathbf{I} - \boldsymbol{\Phi}_1 z - \cdots - \boldsymbol{\Phi}_p z^p$ and assume that $\det(\boldsymbol{\Phi}(z))$ has all zeros at 1 or outside the unit circle.

- Let $\boldsymbol{\Pi} = \mathbf{B}_1 - \mathbf{I}$. If $\text{rank}(\boldsymbol{\Pi}) = k$, then $\mathbf{y}_t$ is stationary; If $\text{rank}(\boldsymbol{\Pi}) = 0$, then $\Delta \mathbf{y}_t$ is stationary.
- If $\text{rank}(\boldsymbol{\Pi}) = r$, $0 < r < 1$, then there exists $k \times r$ matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^T$, $\text{rank}(\boldsymbol{\alpha}) = \text{rank}(\boldsymbol{\beta}) = r$.

Cointegration vectors

- The column vectors of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r)$ have the property that $\boldsymbol{\beta}_i^T \mathbf{y}_t$ is weakly stationary and are therefore cointegrating vectors. An economic interpretation of a cointegrated multivariate time series $\mathbf{y}_t$ is that its components have some common trends that result in $\boldsymbol{\beta}_i^T \mathbf{y}_t$ having long-run equilibrium for $1 \leq i \leq r$ even though the individual components $\mathbf{y}_{ti}$ are nonstationary and have variances diverging to ∞. In particular, if $\beta_{ji} \neq 0$, then linear regression of y_{tj} on the other components of $\mathbf{y}_t$ would not be spurious even though $\mathbf{y}_t$ is unit-root nonstationary.
- Johansen's test for the number of cointegration vectors: testing $H_0 : \text{rank}(\boldsymbol{\Pi}) \leq r$. The R function is `ca.jo`.

GARCH

- Recall an AR(p) for $\{x_t\}$ is

$$x_t = \mu + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + u_t, \ E(u_t) = 0.$$

- GARCH(1,1):

$$u_t = \sigma_t \varepsilon_t, \ \sigma_t^2 = \omega + \beta \sigma_{t-1}^2 + \alpha u_{t-1}^2, \ E(\varepsilon_t) = 0, \ \mathbf{Var}(\varepsilon_t) = 1.$$

  For GARCH(h,k), $\sigma_t^2 = \omega + \sum_{i=1}^{h} \beta_i \sigma_{t-i}^2 + \sum_{j=1}^{k} \alpha u_{t-j}^2$.

- Let $\boldsymbol{\theta} = (\omega, \alpha, \beta)^T$. Assuming that $\varepsilon_t \sim N(0,1)$, the log-likelihood function is given by

$$\ell(\boldsymbol{\theta}) = \frac{1}{2} \sum_{t=1}^{n} \log \sigma_t^2 - \sum_{t=1}^{n} \frac{u_t^2}{2\sigma_t^2} - \frac{n}{2} \log(2\pi),$$

  in which the σ_t^2 can be computed recursively with the initial value σ_0^2 and $\boldsymbol{\theta}_0$ are given.

MLE $\hat{\theta}$

- The MLE $\hat{\boldsymbol{\theta}}$ is the solution of

$$0 = \nabla \ell(\boldsymbol{\theta}) = -\frac{1}{2} \sum_{t=1}^{n} \left(\frac{1}{\sigma_t^2} - \frac{u_t^2}{\sigma_t^4} \right) \nabla \sigma_t^2,$$

  in which $\nabla \sigma_t^2$ can be evaluated recursively by

$$\nabla \sigma_t^2 = (1, u_{t-1}^2, \sigma_{t-1}^2)^T + \beta \nabla \sigma_{t-1}^2.$$

- For non-Gaussian, we use the same $\ell(\boldsymbol{\theta})$ for Gaussian as quasi-MLE (QMLE). Jeanthean (1998) proves the QMLE to be consistent under the main assumption that the considered multivariate process is strictly stationary and ergodic.
  - A stationary time series is ergodic if sample moments converge in probability to population moments.

BFGS

To solve $\nabla \ell(\boldsymbol{\theta}) = \mathbf{0}$ by BFGS (Quasi-Newton Method)

- Pick an initial guess $\boldsymbol{\theta}_0$, an approximate Hessian matrix $\mathbf{B}_0$.
- Solve $\mathbf{p}_k$: $\mathbf{B}_k \mathbf{p}_k = -\nabla \ell(\boldsymbol{\theta}_k)$.
- Choose α_k and update $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_k \mathbf{p}_k$.
- Set $\mathbf{s}_k = \alpha_k \mathbf{p}_k$, $\mathbf{y}_k = \nabla \ell(\boldsymbol{\theta}_{k+1}) - \nabla \ell(\boldsymbol{\theta}_k)$, updates

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{\mathbf{y}_k \mathbf{y}_k^T}{\mathbf{y_k}^T \mathbf{s}_k} - \frac{\mathbf{B}_k \mathbf{s}_k \mathbf{s}_k^T \mathbf{B}_k}{\mathbf{s}_k^T \mathbf{B}_k \mathbf{s}_k}$$

.

Multivariate GARCH models

- Extent the univariate GARCH(1,1) to multivariate (of dimension q)

$$\mathbf{u}_t = \mathbf{\Sigma}_t \varepsilon_t, \ \mathbf{\Sigma}_t = \mathbf{C}\mathbf{C}^T + \mathbf{A}\mathbf{u}_{t-1}\mathbf{u}_{t-1}^T\mathbf{A}^T + \mathbf{B}\mathbf{\Sigma}_{t-1}\mathbf{B}^T,$$

  where $\mathbf{C}, \mathbf{A}, \mathbf{B} \in \mathbb{R}^{q\times q}$ and $\mathbf{C}$ is a lower triangular matrix.

- Let $\boldsymbol{\theta} = (\mathbf{C}, \mathbf{A}, \mathbf{B})$, the log-likelihood function is

$$\ell(\boldsymbol{\theta}) = \frac{nq}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n}|\mathbf{\Sigma}_t| - \frac{1}{2}\sum_{t=1}^{n}\mathbf{u}_t^T\mathbf{\Sigma}_t^{-1}\mathbf{u}_t.$$

Exponentially weighted moving averages

- Moving average:

$$\widehat{\Sigma}_t = \sum_{i=1}^{k} \alpha_i \mathbf{u}_{t-i} \mathbf{u}_{t-i}^T, \ \sum_{i=1}^{k} \alpha_i = 1$$

- For $k = \infty$, take $\alpha_i = (1-\lambda)/\lambda^{i-1}$ for some $0 < \lambda < 1$, then the exponentially weighted moving average (EWMA) model is

$$\widehat{\Sigma}_t = \lambda \widehat{\Sigma}_{t-1} + (1-\lambda) \mathbf{u}_{t-1} \mathbf{u}_{t-1}^T.$$

No-Arbitrage assumption

- An arbitrage is a strategy with no cost making a profit in the future without risks.
- Main Assumption: There are no arbitrage opportunities.
- Law of one price: Suppose that there are no arbitrage. If two self-financing strategies have the same payoffs at time T, then they must have the same value at any time $t \leq T$.
    - A self-financing is a strategy that doesn't require injection of funds after the initial date and from which no cash withdrawn.

One period binomial model

- Assume there is one stock S with initial value $S_0 > 0$ fixed and a constant interest rate r.
- We suppose that the stock price at $T > 0$ can only take the two values $S_T = uS_0$ with probability p and $S_T = dS_0$ with probability $1 - p$, with $d < u$.
- Suppose there is an option with payoff (at T) $H = H_u$ if $S_T = uS_0$ and $H = H_d$ if $S_T = dS_0$. What should be the value V_0 of this option?

Pricing under no-arbitrage condition

- Consider a replicating portfolio $V = aS + b$, where

$$a = \frac{H_u - H_d}{S_0(u-d)}, \; b = e^{-rT}\frac{uH_d - dH_u}{(u-d)}.$$

- At time T, the payoff of $V = H$. Therefore, they the should have same value at time 0.

$$\begin{aligned} V_0 = aS_0 + b &= \frac{H_u - H_d}{(u-d)} + e^{-rT}\frac{uH_d - dH_u}{(u-d)} \\ &= e^{-rT}\left(\frac{e^{rT}-d}{u-d}H_u + (1 - \frac{e^{rT}-d}{u-d})H_d\right) \\ &= e^{-rT}(qH_u + (1-q)H_d), \end{aligned}$$

  where $q = (e^{rT} - d)/(u-d)$.

- The one-step binomial tree is arbitrage-free if and only if $d < e^{rT} < u$.

Stochastic interest rates and short-rate models

- To price interest rate derivatives that are contingent on future interest rates, stochastic models of interest rate dynamics are prescribed and pricing is based on arbitrage-free arguments similar to those in the Black-Scholes theory.
- A major class of stochastic models in the literature consists of diffusion processes for the short rate r_t,

$$dr_t = m(t, r_t)dt + s(t, r_t)dw_t,$$

where w_t is the Brownian motion driving the process with $dw_t \sim N(0, dt)$.

- Let $\mu(t, r) = m(t, r)/r$, $\sigma(t, r) = s(t, r)/r$, so $dr_t/r_t = \mu(t, r_t)dt + \sigma(t, r_t)dw_t$. it can be shown that

$$\lambda(t, r) = \frac{\mu(t, r) - r}{\sigma(t, r)}$$

is the same for all derivatives dependent on r_t (e.g., bonds with the same maturities). $\lambda(t, r)$ is called the market price of risk.

Stochastic interest rates and short-rate models

- In Financial Econometric, interest rates are assumed to fluctuate around some level (mean-reverting property).
- Based on mean reversion, Vasicek (1977) introduce a short rate model

$$dr_t = \kappa(\theta - r_t)dt + \sigma dw_t,$$

where θ represents the mean level that short rate r_t will evolve around in the long run, κ represents the speed of reversion, σ represents instantaneous volatility.

- Various modifications of this model have been developed,
  - Cox, Ingersoll, and Ross (1985): $dr_t = \kappa(\theta - r_t)dt + \sigma\sqrt{r_t}dw_t$, which ensure $r_t > 0$ when $2\kappa\theta > \sigma^2$.
  - Hull and White (1990): $dr_t = (\theta_t - \kappa r_t)dt + \sigma dw_t$.

$P(t, T)$ under different r_t models

Under Vasicek model, Cox, Ingersol, and Ross (CIR) model and Hull and White model,

- the price $P(t, T)$ of a zero-coupon bond with face value 1 given by

$$P(t, T) = \alpha(t, T)e^{-\beta(t,T)r_t}.$$

- The spot rate $R(t, T)$ is an affine function (linear function with a constant term) of the short rate r_t:

$$R(t, T) = -\frac{\log \alpha(t, T)}{T - t} + r_t \frac{\beta(t, T)}{T - t}.$$

- The instantaneous forward rate is also an affine function of r_T:

$$f(t, T) = -\frac{\partial}{\partial T} \log P(t, T) = -\frac{\partial}{\partial T} \log \alpha(t, T) + r_t \frac{\partial}{\partial T} \beta(t, T).$$

$P(t, T)$ under different r_t models

- Vasicek model, $\beta(t, T) = (1 - e^{-\kappa(T-t)})/\kappa$,

$$\alpha(t, T) = \exp\left\{\left(\theta - \frac{\sigma^2}{2\kappa^2}\right)(\beta(t, T) - (T - t)) - \frac{\sigma^2}{4\kappa}\beta^2(t, T)\right\}.$$

- CIR model, letting $h = \sqrt{\kappa^2 + 2\sigma^2}$,

$$\alpha(t, T) = \left(\frac{2he^{(\kappa+h)(T-t)/2}}{2h + (\kappa + h)(e^{(T-t)h} - 1)}\right)^{2\kappa\theta/\sigma^2},$$

$$\beta(t, T) = \frac{2(e^{(T-t)/h} - 1)}{2h + (\kappa + h)(e^{(T-t)h} - 1)}.$$

- Hull-White model, $\beta(t, T) = (1 - e^{-\kappa(T-t)})/\kappa$,

$$\log \alpha(t, T) = \log \frac{P(0, T)}{P(0, t)} - \beta(t, T)\frac{\partial}{\partial t} \log P(0, t) - \frac{\sigma^2(1 - e^{-2\kappa t})}{4\kappa}\beta^2(t, T).$$

Multifactor affine yield models

- The Vasicek, CIR, and Hull and White models are one-factor models, which only rely on one source of randomness dw_t in the short rate r_t. In these models, the yields for different maturities are perfectly correlated.
- A simple way to introduce additional sources of randomness is to replace dw_t by $d\mathbf{w}_t$, where $\mathbf{w}_t$ is d-dimensional Brownian motion,
  - $d\mathbf{x}_t = \mathbf{B}\mathbf{x}_t dt + \mathbf{\Sigma}^{1/2} d\mathbf{w}_t$.
  - $r_t = \mu + \boldsymbol{\theta}^T \mathbf{x}_t$.

Multifactor affine yield models: An example (Sect. 10.4.4)

- Consider the two-factor model: $r_t = x_t + y_t$

$$dx_t = -ax_t dt + \sigma dw_t^{(1)}, \quad dy_t = -by_t dt + \tilde{\sigma} dw_t^{(2)},$$

where $a, b, \sigma, \tilde{\sigma}$ are positive constants and $(w_t^{(1)}, w_t^{(2)})$ is a two-dimensional Brownian motion under risk neutral probability measure.

- Then the price of a zero-coupon bond maturing at time T can be expressed as

$$P(t,T) = \exp\left\{-\frac{1-e^{-a(T-t)}}{a}x_t - \frac{1-e^{-b(T-t)}}{b}y_t + \frac{1}{2}\alpha(t,T)\right\},$$

where

$$\alpha(t,T) = \frac{\sigma^2}{a^2}\left(T-t+\frac{2}{a}e^{-a(T-t)} - \frac{1}{2a}e^{-2a(T-t)} - \frac{3}{2a}\right) + \frac{\tilde{\sigma}^2}{b^2}\left(T-t+\frac{2}{b}e^{-b(T-t)} - \frac{1}{2b}e^{-2b(T-t)} - \frac{3}{2b}\right).$$

- Empirical results on PCA of yield curves suggest that usually $d = 2$ or 3 suffices for the number of factors.

Stochastic volatility models

- Assume σ_t follows a stochastic process:

$$dS_t = (r - q)S_t dt + \sqrt{v_t} S_t dW_t^1$$
$$dv_t = \alpha(S_t, v_t, t)dt + \beta(S_t, v_t, t)\sqrt{v_t} dW_t^2,$$

  with $\text{Cov}(dW_t^1, dW_t^2) = \rho dt$

- Hull and White (1987): $\alpha(S_t, v_t, t) = \alpha(v^* - v_t)$, $\beta(S_t, v_t, t) = \beta v_t^{\xi}$ and $\rho = 0$. For this SV model, they have shown that the price of a European option is given by $\int_0^\infty b(\omega)g(\omega)d\omega$, where $b(\omega)$ is the Black-Scholes price in which σ is replaced by ω, and ω is the average variance rate during the life of the option, which is a random variable with density function g determined by the stochastic dynamics for v_t. They have used this representation of the option price to develop closed-form approximations to the model price.

Introduction to credit scoring

- Credit scoring systems have been developed for virtually all types of credit analysis, from consumer credit to commercial loans. The idea is to pre-identify certain key factors that determine the probability of default (PD) and combine or weight them into a quantitative score.
- Thomas (2000) gives a survey of credit scoring for consumer loans. To build a credit scorecard, one uses a training sample that consists of $(\mathbf{x}_i, I_i)$, $1 \leq i \leq n$, from n loans, where $\mathbf{x}_i$ a d-dimensional vector of predictor variables (attributes) of the ith loan, and I_i is a binary response variable defined by $I_i = 1$ (or 0) if the ith loan defaults (or stays clean) within the last 12 or 18 or 24 months after at least one year since the loan was offered.

Introduction to credit scoring

- Regarding $(\mathbf{x}_1, I_1), \cdots, (\mathbf{x}_n, I_n)$ as independent realizations of $(\mathbf{x}, I)$, the statistical problem is to estimate $P(I = 1|\mathbf{x} = \mathbf{x})$ and the methods commonly used in credit scoring include
  - linear or quadratic discriminant analysis
  - logistic regression
- These methods estimate $P(I = 1|\mathbf{x} = \mathbf{x})$ by some function of $\widehat{\beta}^T\mathbf{x}$, in which $\widehat{\beta}$ is estimated from the training sample. The linear combination $\widehat{\beta}^T\mathbf{x}$ therefore provides the "score" of an obligor with attribute vector $\mathbf{x}$.

An example of discriminant analysis for default prediction

The first work to use discriminant analsis for default prediction was reported by Altman (1968). He considered a discriminant function, labeled as the Z-score, has the form

$$d(\mathbf{x}) = 0.012x_1 + 0.014x_2 + 0.033x_3 + 0.006x_4 + 0.999x_5,$$

in which

- x_1 is the ratio of working capital and total assets,
- x_2 is the ratio of retained earnings and total asset,
- x_3 is the ratio of earnings before interest and taxes and total assets,
- x_4 is the ratio of a corporate market value of equity and book value of total debt,
- x_5 is the ratio of corporate sales and total assets.

In Altman (1968), a lower threshold of 1.81 is reported below which all firms in the study defaulted and an upper threshold of 2.67 above which all firms survived.

k-means algorithm

Given a set of observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we want to separate them into k groups.

- Pick k points, that are far from each other, as centroids.
- For each point, place it in the cluster whose current centroid it is nearest.
- After all points are assigned, update the locations of centroids of the k clusters.
- Reassign all points to their closest centroid.
- Repeat until the centroids are stabilized.

How to select k?

- Try different k, looking at the change in the average distance to centroid as k increases.
- Average falls rapidly until right k, then changes little

BFR Algorithm

- BFR (Bradley-Fayyad-Reina) is a variant of k-means designed to handle very large data sets
- Assumes that clusters are normally distributed around a centroid in a Euclidean space
- Most points from previous memory loads are summarized by simple statistics
- 3 sets of points:
  - Discard set (DS): Points close enough to a centroid to be summarized
  - Compression set (CS): Groups of points that are close together but not close to any existing centroid
  - Retained set (RS): Isolated points waiting to be assigned to a compression set

BFR Algorithm

- The discard set (or compression set) is summarized by:
  - The number of points, N
  - The vector SUM, whose ith component is the sum of the coordinates of the points in the ith dimension
  - The vector $SUMSQ$: ith component is the sum of squares of coordinates in ith dimension
- For each DS or CS, the centroid can be calculated as $SUM(i)/N$, and the variance in dimension i is $SUMSQ(i)/N - (SUM(i)/N)^2$

BFR Algorithm

- Starting from k centroids. Adding points and assign them to different clusters as in the original k-means.
- Points that are closed to each other can be summerized and discarded to DS (if includes the centroid) or CS.
- When a point is added to a DS or CS, adjust the corresponding statistics N, SUM and $SUMSQ$.
- After all points are assigned to different clusters, update the locations of centroids of the k clusters.
- Reassign all points, DS's and CS's to their closest centroid.
- Repeat until the centroids are stabilized. If this is the last round, merge all CS's and all RS points into their nearest cluster.

Linear discriminant analysis

If a sample of K classes $\{(x_i, I_i)\}_{i=1}^n$ are observed, where x_i is a $d-$dimensional vector of predictor variables, I_i is its corresponding class index that $I_i \in \{1, \ldots, K\}$.

- Assume the class-conditional density of X in class $I = k$ is $f_k(x)$.
- The prior probability of class k is $\{\pi_k\}_{k=1}^n$, and $\sum_{k=1}^K \pi_k = 1$.
- Apply Bayes theorem, we have

$$Pr(I = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^K f_l(x)\pi_l}$$

- By the Bayes rule of 0-1 loss,

$$\begin{aligned} \hat{I}(x) &= \arg\max_k Pr(I = k|X = x) \\ &= \arg\max_k f_k(x)\pi_k \end{aligned}$$

Linear discriminant analysis

- Suppose each class density is multivaraite Gaussian

$$f_k(x) = (2\pi)^{-d/2}|\Sigma_k|^{-1/2}e^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}$$

- Linear discriminant analysis(LDA) arises in the special case where the classes are assumed to have a common covariance matrix $\Sigma_k = \Sigma, \forall k$.
- In this case,

$$\begin{aligned}\hat{I}(x) &= \arg\max_k f_k(x)\pi_k \\ &= \arg\max_k \Big[-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k) + \log(\pi_k)\Big]\end{aligned}$$

Linear discriminant analysis

In practice we need to estimate the parameters for Gaussian density from training sample:

- $\hat{\pi}_k = N_k/N$, where N_k is the number of class-k observations,
- $\hat{\mu}_k = \sum_{I_i=k} x_i/N_k$,
- $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{I_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T/(N-K)$.

Quadratic discriminant analysis(QDA)

- The Σ_k are not assumed to be equal, estimate them separately for $k = 1, \ldots, K$.
- *Quadratic discriminant function*:

$$\delta_k(x) = \frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log(\pi_k),$$

and the classification rule is the same as in LDA:
$\hat{I}(x) = \arg\max_k \delta_k(x)$

- QDA fits the data better than LDA, but each Σ_k is a $d \times d$ matrix hence the number of parameters to be estimated is much larger.

Logistic regression and generalized models

- Whereas LDA is based on the assumption that there are two populations $(\mathbf{x}, 0)$ and $(\mathbf{x}, 1)$ corresponding to $I = 0, 1$, logistic regression treats $P(I = 1|\mathbf{x}) = E(I|\mathbf{x})$ as a regression function.
- Instead of the usual linear regression function $E(I|\mathbf{x} = \mathbf{x}) = \beta^T \mathbf{x}$, which is unbounded if $\mathbf{x}$ does not have bounded support, logistic regression uses the *logit* transformation $g(p(\mathbf{x}))$ of $p(\mathbf{x}) = P(I = 1|\mathbf{x} = \mathbf{x})$ so that $g(p(\mathbf{x})) = \beta^T \mathbf{x}$.
- The logit function $g(p) = \text{logit}(p)$ defined by

$$g(p) = \log[p/(1-p)], \ 0 < p < 1$$

is the canonical link in generalized linear models, of which logistic regression is a special case.

Generalized linear models

Generalized linear models that make the following assumptions on the relationship between the response variable y_t to the covariate $\mathbf{x}_t$:

(A1) y_t has density function

$$f(y;\theta,\phi) = \exp\left\{ \left[y\theta - b(\theta) \right]/\phi + c(y,\phi) \right\}.$$

(A2) $h(\theta) = \boldsymbol{\beta}^T \mathbf{x}$ for some given smooth increasing function h and unknown parameter $\boldsymbol{\beta}$.

(A3) $(\mathbf{x}_t, y_t)$, $1 \leq t \leq n$, are independent.

Generalized linear models

The parametric family of density functions (**??**) is called an *exponential family* with *canonical parameter* θ and *dispersion parameter* $\phi > 0$. It includes

- normal $N(\theta, \sigma^2)$ family with $b(\theta) = \theta^2/2$, $\phi = \sigma^2$, and $c(y, \phi) = -\{y^2/\sigma^2 + \log(2\pi\sigma^2)\}/2$,
- Poisson density $f(y; \lambda) = e^{-\lambda}\lambda^y/y!$ $(y = 0, 1, 2, \dots)$ with $\theta = \log \lambda$, $b(\theta) = e^\theta$, and $\phi = 1$.

The mean and variance of the distribution of y_t in (**??**) are given by

$$\text{mean} = b'(\theta), \qquad \text{variance} = \phi b''(\theta).$$

Likelihood functions for generalized linear models

- Given the observed $\{(\mathbf{x}_i, y_i) : 1 \leq i \leq n\}$, the log-likelihood associated with (A1)-(A3) is

$$\ell(\boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} \left\{ [y_i h^{-1}(\boldsymbol{\beta}^T \mathbf{x}_i) - b(h^{-1}(\boldsymbol{\beta}^T \mathbf{x}_i))]/\phi + c(y_i, \phi) \right\}.$$

- In particular, for logistic regression, the log-likelihood function is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \boldsymbol{\beta}^T \mathbf{x}_i - \log[1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i)] \right\}.$$

- To find the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}$, we can apply numerical optimization methods (e.g. iteratively reweighted least squares, see p.96 of Lai and Xing (2008)). The following software packages can be used to fit generalized linear models
R : `glm(formula, family, data)` and
MATLAB : `glmfit(X, y, distribution)`.

Pearson residuals

- Let $\widehat{\beta}$ and $\widehat{\theta}$ be the parameter estimates in the generalized linear model, the *Pearson residuals* are

$$e_i^P = (y_i - \widehat{\mu}_i)/\sqrt{\widehat{v}_i}, \quad \text{where } \widehat{\mu}_i = b'(\widehat{\theta}_i), g(\widehat{\theta}_i) = \widehat{\beta}^T \mathbf{x}_i, \widehat{v}_i = \widehat{\phi} b''(\widehat{\theta}_i).$$

- For logistic regression, the Pearson residuals are

$$e^P(\mathbf{x}_j) = \frac{y_j - m_j \widehat{p}_j}{\sqrt{m_j \widehat{p}_j (1 - \widehat{p}_j)}}, \ 1 \leq j \leq J,$$

where y_j is the sum of output with the same covariate $\mathbf{x}_j$, $\widehat{p}_j = \widehat{p}(\mathbf{x}_j)$ is the estimated probability at covariate $\mathbf{x}_j$ by the fitted logistic regression model and m_j is the number of observations having the same $\mathbf{x}_j$ so that $\sum_{j=1}^{J} m_j = n$.

Pearson χ^2 test

- Note that the sum of squares of the Pearson residuals is the usual chi-square statistic

$$TS = \sum_j \frac{(y_j - m_j \widehat{p}_j)^2}{m_j \widehat{p}_j (1 - \widehat{p}_j)} \sim \chi^2$$

for testing the null hypothesis that the data are generated by the logistic regression model. Under H_0 (the model is correct), TS has an approximately χ^2-distribution with $J - d$ degrees of freedom, where d is the dimensionality of $\boldsymbol{\beta}$, assuming that $\min_{1 \leq j \leq J} m_j \geq 5$.

Generalized likelihood ratio statistic

- Let Θ_q be a q-dimensional subspace of the parameter space (e.g. $\{\beta_i = 0,\ i = q+1, q+2, \ldots\}$) and Θ_p, $p > q$, be a p-dimensional subspace of the parameter space that contains Θ_q (e.g. $\{\beta_i = 0,\ i = p+1., +2, \ldots\}$)
- To test $H_0 : \boldsymbol{\beta} \in \Theta_q$, the generalized likelihood ratio (GLR) statistic is

$$\Lambda(\Theta_q, \Theta_p) = 2\left\{\sup_{\boldsymbol{\beta}\in\Theta_p} \ell_n(\boldsymbol{\beta}) - \sup_{\boldsymbol{\beta}\in\Theta_q} \ell_n(\boldsymbol{\beta})\right\},$$

which has a limiting χ^2_{p-q} distribution under H_0.

Deviance χ^2-test

- Instead of Pearson χ^2-test, an alternative test is the GLR test that uses the GLR statistic

$$\Lambda = 2\sum_{j=1}^{J}\left\{y_j \log\left(\frac{y_j}{m_j\widehat{p}_j}\right) + (m_j - y_j)\log\left(\frac{m_j - y_j}{m_j(1-\widehat{p}_j)}\right)\right\}$$

  which is also called the *deviance*. Under H_0, Λ has an approximately χ^2-distribution with $J - d$ degrees of freedom; see Section 2.4.2 of Lai and Xing (2008).

- The *deviance residuals* are the signed square roots of the summands of Λ:

$$e^D(\mathbf{x}_j) = \text{sign}(y_j - m_j\widehat{p}_j)\left\{y_j \log\left(\frac{y_j}{m_j\widehat{p}_j}\right) + (m_j - y_j)\log\left(\frac{m_j - y_j}{m_j(1-\widehat{p}_j)}\right)\right\}^{1/2}.$$

Model selection for logistic regression

- AIC: $-2\ell(\hat{\beta}) + 2k$, where k is the number of variables in the model.
- BIC: $-2\ell(\hat{\beta}) + k(\log n)/n$, where n is the number independent observations.
- Forward stepwise
- Backward elimination
- Lasso

Partial F-tests

- Forward stepwise

STEP 1 Start from $\Theta_q = \emptyset$

STEP 2 Pick $j \in \Theta_q^c$ such that $\Lambda(\Theta_q, \Theta_q \cup \{j\})$ is largest. Denote it as j^*

STEP 3 If $\Lambda(\Theta_q, \Theta_q \cup \{j^*\}) < \chi^2_{1,1-\alpha}$, stop and output Θ_q; Otherwise, $\Theta_q := \Theta_q \cup \{j^*\}$ and repeat STEP 2 and 3.

- Backward elimination

STEP 1 Start from $\Theta_p = \{1, \ldots, p\}$

STEP 2 Pick $j \in \Theta_p$ such that $\Lambda(\Theta_p - \{j\}, \Theta_p)$ is smallest. Denote it as j^*.

STEP 3 If $\Lambda(\Theta_p - \{j^*\}, \Theta_p) > \chi^2_{1,1-\alpha}$, stop and output Θ_p; Otherwise, $\Theta_p := \Theta_p - \{j^*\}$ and repeat STEP 2 and 3.

Introduction to Survival Analysis

- Let τ be the failure time of an individual from a homogeneous population. The **survival function** of τ is

$$S(t) = P(\tau > t) = 1 - F(t),$$

where $F(s) = P(\tau \leq s)$.

- If τ is absolutely continuous, the **probability density function** of τ is $f(t) = -dS(t)/dt$, and the **hazard** (or **intensity**) function is

$$\lambda(t) = \lim_{h \to 0+} \frac{P(t < \tau \leq t + h \mid \tau > t)}{h} = f(t)/S(t).$$

- Integrating $\lambda(t)$ with respect to t yields

$$S(t) = \exp\left[-\int_0^t \lambda(u)du\right] = \exp[-\Lambda(t)],$$

where $\Lambda(t) = \int_0^t \lambda(s)ds$ is called the **cumulative hazard function**.

Nonparametric estimate $S(t)$

- Based on n independent observations τ_i sampled from F, the nonparametric maximum likelihood estimate of F is the empirical distribution function $\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\tau_i \leq t\}}$
- τ_i may not be fully observable because of censoring. Some subjects (or firms) may not fail during the observation period. The data on these individuals are said to be *right-censored*.

Lifetables

Altman (1989) and others have developed mortality tables for loans and bonds by using methods in actuarial science. Partition time into disjoint intervals $I_1 = (0, t_1], I_2 = (t_1, t_2]$, etc.

- A life table summarizes the mortality results of a large cohort of n subjects as follows:

  $n_j =$ number of subjects alive at the beginning of I_j,

  $d_j =$ number of deaths during I_j,

  $\ell_j =$ (number lost to follow-up during I_j)/2.

- It estimates $p_j = P(\text{dying during } I_j \mid \text{alive at the beginning of } I_j)$ by $\widehat{p}_j = d_j/(n_j - \ell_j)$ so that $(1 - \widehat{p}_1) \cdots (1 - \widehat{p}_k)$ can be used to estimate $P(\text{alive at time } t_k) = (1 - p_1) \cdots (1 - p_k)$.

- The actuarial (life-table) estimate of $P(\tau > t_k)$ is the product

$$\widehat{S}_{LT}(t_k) = \prod_{j=1}^{k} (1 - \widehat{p}_j) = \prod_{j=1}^{k} \left(1 - \frac{d_j}{n_j - \ell_j}\right).$$

Delta method

- If $\hat{\theta} \sim N(\theta, \sigma^2)$ asymptotically with small σ and g is a continuously differentiable function in some neighborhood of the θ, then $g(\hat{\theta}) - g(\theta)$ is asymptotically normal with mean 0 and variance $(g'(\hat{\theta}))^2\sigma^2$.
- Using delta method, the variance of $\widehat{S}_{LT}(t_k)$ is approximately given by

$$\widehat{S}_{LT}(t_k)^2 \sum_{j=1}^{k} \frac{d_j}{(n_j - \ell_j - d_j)(n_j - \ell_j)}$$

Product-limit (Kaplan-Meier) estimate of survival function

- Let c_i be the length of the observation period, $T_i = \min(\tau_i, c_i)$ and $\delta_i = I_{\{\tau_i \le c_i\}}$ be the censoring indicator that indicates whether T_i is an actual failure time or is censored.. The observations, therefore, are $(T_i, \delta_i), i = 1, \cdots, n$. Subject i is "at risk" at time s if $T_i \ge s$ (i.e., has not failed and not been lost to follow-up prior to s). Let

$$Y(s) = \sum_{i=1}^{n} I_{\{T_i \ge s\}}, \ N(s) = \sum_{i=1}^{n} I_{\{T_i \le s, \ \delta_i = 1\}}.$$

- Product-limit (Kaplan-Meier) estimate:

$$\widehat{S}_{KM}(t) = \prod_{s<t} \left(1 - \frac{\Delta N(s)}{Y(s)}\right) = \prod_{T_i<t} \left(1 - \frac{\Delta N(T_i)}{Y(T_i)}\right),$$

where we use the notation $0/0 = 0$. Using delta method, the variance of $\widehat{S}_{KM}(t)$ is approximately given by

$$\widehat{S}_{KM}(t)^2 \sum_{T_i<t} \frac{\Delta N(T_i)}{(Y(T_i) - \Delta N(T_i))(Y(T_i))}$$

Nelson-Aalen estimator of the cumulative hazard function

- Note that $Y(s)$ is the risk set size and $\Delta N(s) = N(s) - N(s-)$ is the number of observed deaths at time s, and that $\Delta N(s)/Y(s)$ is the analog of $\widehat{p}_j$.
- *Nelson-Aalen estimator*:

$$\widehat{\Lambda}_{NA}(t) = \sum_{s<t} \frac{\Delta N(s)}{Y(s)} = \int_0^t \frac{I_{\{Y(s)>0\}}}{Y(s)} dN(s)$$

  for the censored data $(T_i, \delta_i), 1 \leq i \leq n$.
- Using delta method, the variance of $\widehat{\Lambda}_{NA}(t)$ is approximately given by

$$\sum_{T_i<t} \frac{\Delta N(T_i)}{(Y(T_i) - \Delta N(T_i))(Y(T_i))}$$

Hazard regression: parametric approach

- We have focused so far on the estimation of the survival distribution of a failure time τ. In applications, one often wants to use a model for τ to predict future failures from a vector $\mathbf{x}(t)$ of predictors based on current and past observations; $\mathbf{x}(t)$ is called a time-varying (or time-dependent) covariate.
- When $\mathbf{x}(t) = \mathbf{x}$ does not depend on t, it is called a time-independent (or baseline) covariate. In practice, some predictors may be time-independent while other components of $\mathbf{x}(t)$ may be time-varying.
- Since prediction of future default from $\mathbf{x}(t)$ is relevant only if $\tau > t$ (i.e., if default has not occurred at or before t), one is interested in modeling the conditional distribution of τ given $\tau > t$, e.g., by relating the hazard function $\lambda(t)$ to $\mathbf{x}(t)$.

Hazard regression: parametric approach

- *Proportional hazards* (or *Cox regression*) model

$$\lambda(t) = \lambda_0(t) \exp(\boldsymbol{\beta}^T \mathbf{x}(t)),$$

  in which $\lambda_0(t)$ is called the baseline hazard function.

- The parameters are $\boldsymbol{\beta}$ and $\lambda_0(\cdot)$. Since the function λ_0 is an infinite-dimensional parameter and there are only n observations, one way to overcome difficulties caused by the infinite-dimensional parameter λ_0 is to specify it up to a parameter vector $\boldsymbol{\theta}$ so that $(\boldsymbol{\beta}, \boldsymbol{\theta})$ can be estimated by maximum likelihood.

- Commonly used parametric models for $\lambda_0(t)$ are: Exponential density $\theta e^{-\theta t}$ and Weibull density $\theta t^{\theta-1} \exp(-t^\theta)$ (with $\theta > 0$).

Hazard regression: parametric approach

- Assume that the covariates are time-invariant. $f_i = \lambda_i S_i$ is the density function of the τ_i, $\log S_i(t) = -\Lambda_i(t)$ and $\Lambda_i(t) = \Lambda_0(t)\exp(\boldsymbol{\beta}^T \mathbf{x}_i)$, where $\Lambda_i(t)$ is the cumulative hazard function of τ_i and $\Lambda_0(t) = \int_0^t \lambda_0(u)du$, therefore the log-likelihood function can be written as

$$\begin{aligned}\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \sum_{i=1}^{n} \{\delta_i \log f_i(T_i) + (1-\delta_i)\log S_i(T_i)\} \\ &= \sum_{i=1}^{n} \left\{\delta_i[\log \lambda_0(T_i;\boldsymbol{\theta}) + \boldsymbol{\beta}^T \mathbf{x}_i] - \Lambda_0(T_i;\boldsymbol{\theta}) e^{\boldsymbol{\beta}^T \mathbf{x}_i})\right\}\end{aligned}$$

Semiparametric approach and partial likelihood

- Cox (1972, 1975) introduced a semiparametric method to estimate the finite-dimensional parameter β in the presence of an infinite-dimensional nuisance parameter $\lambda_0(\cdot)$
- It is semiparametric in the sense of being nonparametric in λ_0 but parametric in β in terms of $\beta^T \mathbf{x}$.
- Order the observed censored failure times as $\tau_{(1)} < \cdots < \tau_{(m)}$, with $m \leq n$. Noting that with probability 1 there is only one failure at $\tau_{(j)}$ because the failure time distributions have density functions. Let $R_{(j)} = \{i : T_i \geq \tau_{(j)}\}$ denote the risk set at $\tau_{(j)}$. Then

$$P\left\{(j) \text{ fails at } \tau_{(j)} | R_{(j)}, \text{ one failure at } \tau_{(j)}\right\}$$
$$= \exp\left(\beta^T \mathbf{x}_{(j)}\right) / \sum_{i \in R_{(j)}} \exp\left(\beta^T \mathbf{x}_{(i)}\right).$$

Semiparametric approach and partial likelihood

The partial likelihood is

$$\prod_{j=1}^{m} P\left\{(j) \text{ fails at } \tau_{(j)} | R_{(j)}, \text{ one failure at } \tau_{(j)}\right\}.$$

Cox's regression estimator $\widehat{\boldsymbol{\beta}}$ is the maximizer of the partial log-likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{j=1}^{m} \left\{ \boldsymbol{\beta}^T \mathbf{x}_{(j)} - \log \left(\sum_{i \in R_{(j)}} \exp(\boldsymbol{\beta}^T \mathbf{x}_{(i)}) \right) \right\},$$

or equivalently, the solution of $\frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = 0$.

Properties of the partial likelihood estimate

- Making use of martingale theory, Cox's regression estimator $\boldsymbol{\beta}$ can be shown to satisfy the usual asymptotic properties of maximum likelihood estimates even though partial likelihood is used. In particular, it can be shown that as $n \to \infty$,

$$(-\ddot{\ell}(\boldsymbol{\beta}))^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \text{ has a limiting standard normal distribution,}$$

where $\ddot{I}(\boldsymbol{\beta})$ is the Hessian matrix of second partial derivatives $(\partial^2/\partial\beta_k\partial\beta_h)\ell(\boldsymbol{\beta})$.

Properties of the partial likelihood estimate

- One can perform usual likelihood inference, treating the partial likelihood as a likelihood function and apply likelihood-based selection of covariates similar to that for generalized linear models.
- Moreover, even though $\widehat{\beta}$ is based on partial likelihood, it has been shown to be asymptotically efficient. That is, $\widehat{\beta}$ has the smallest variance among all other competing estimators for β, at least, when the sample size is large.
- Computation of $\widehat{\beta}$ and related likelihood inference can be carried out by the R functions `survfit` and `coxph` in library `survival`.

Estimation of baseline hazard

- When there are no covariates, $\Lambda = \Lambda_0$ can be estimated by the Nelson-Aalen estimator. Note that $\widehat{\Lambda}_{NA}(t)$ has jumps only at uncensored observations and that $Y(s)$ is the sum of 1's over the risk set $\{i : T_i \geq s\}$ at s.
- When τ_i has hazard function $\exp(\boldsymbol{\beta}^T \mathbf{x}_i(s))\lambda_0(s)$, we modify $Y(s)$ to

$$Y(s) = \sum_{i \in R_{(j)}} \exp(\boldsymbol{\beta}^T \mathbf{x}_i) \quad \text{at } s = \tau_{(j)},$$

which is the Breslow estimator of Λ_0 in the proportional hazards regression model.

Partial likelihoods when ties are present

- For $d_j \geq 1$, let D_j be the set of all individuals who fail at time $\tau_{(j)}$.
- Breslow (1974) partial likelihood

$$L(\beta) = \prod_{j=1}^{m} \frac{\exp(\sum_{i \in D_j} \beta^T \mathbf{x}_{(i)})}{\left(\sum_{i \in R_{(j)}} \exp(\beta^T \mathbf{x}_{(i)})\right)^{d_j}}$$

- This likelihood considers each of the d_j events at a given time as distinct, constructs their contribution to the likelihood function, and obtains the contribution to the likelihood by multiplying over all events at time $\tau_{(j)}$.

Partial likelihoods when ties are present

- Efron (1977) partial likelihood

$$L(\beta) = \prod_{j=1}^{m} \frac{\exp(\sum_{i \in D_j} \beta^T \mathbf{x}_{(i)})}{\prod_{k=1}^{d_j} \left(\sum_{i \in R_{(j)}} \exp(\beta^T \mathbf{x}_{(i)}) - \frac{k-1}{d_j} \sum_{i \in D_j} \exp(\beta^T \mathbf{x}_{(i)}) \right)}$$

- This likelihood is closer to the correct partial likelihood than Breslow's likelihood.
- When the number of ties is small, Efron's and Breslow's likelihoods are quite close, and works quite well.
- Breslow's likelihood is the default for many statistical packages, but R uses Efron's partial likelihood.

Intensity-based models for pricing default risk

- Besides the important roles in the statistical analysis of failure-time data, intensity modeling has also become a standard approach to pricing corporate bonds and other credit-sensitive securities since the 1990s.
- Intensity models provide an elegant way to combine term-structure modeling for default-free bonds with the default risk and recovery features of corporate bonds.

Intensity-based models for pricing default risk

- A *Cox process* is often used to model the default intensity of the bond's issuer. The default intensity λ_t is assumed to be governed by an exogenous stochastic process $\mathbf{x}_t, t \geq 0$, so that $\lambda_t = \lambda(\mathbf{x}_t)$ and the stochastic dynamics of λ_t are specified through $\mathbf{x}_t$.
- Let τ be the default time. $\Lambda(\tau)$ is distributed as an exponential random variable ϵ_1 with mean 1 that is independent of $\{\mathbf{x}_s, s \geq 0\}$, where $\Lambda(t) = \int_0^t \lambda(\mathbf{x}_s) ds$. Hence we can use $\lambda(\mathbf{x}_s)$ to generate τ by

$$\tau = \inf \left\{ t : \int_0^t \lambda(\mathbf{x}_s) ds \geq \epsilon_1 \right\}.$$

Price of a zero-coupon defaultable bond with zero recovery

- All the stochastic processes are defined on the same probability space in which the probability measure Q is a risk-neutral (or martingale) measure associated with arbitrage-free pricing. Expectations taken under this measure will be denoted by E^Q, while E is used to denote expectation under the real-world (or physical) measure P.
- Consider a zero-coupon bond, with maturity date T and par value 1, issued by a firm at time t. Assume that there is a short-rate process $r(\mathbf{x}_s)$ under the risk-neutral measure Q such that the default-free bond price is given by

$$p(t,T) = E^Q\left\{\exp\left(-\int_t^T r(\mathbf{x}_s)ds\right)\right\},$$

Price of a zero-coupon defaultable bond with zero recovery

Assuming $\lambda(\mathbf{x}_s)$ to be the intensity process for the default time τ of the firm and assuming zero recovery at default, the price of the defaultable bond at time t given $\tau > t$ is

$$\begin{aligned}
\pi(t,T) &= E^Q\left\{ I_{\{\tau>T\}} \exp\left(-\int_t^T r(\mathbf{x}_s)ds \right) \mid \tau > t \right\} \\
&= E^Q\left\{ E^Q\left[\cdots | \tau > t, \mathbf{x}_s, s \le T \right] \right\} \\
&= E^Q\left\{ \exp\left(-\int_t^T r(\mathbf{x}_s)ds \right) E^Q\left[I_{\{\tau>T\}} | \tau > t, \mathbf{x}_s, s \le T \right] \right\} \\
&= E^Q\left\{ \exp\left(-\int_0^T (r+\lambda)(\mathbf{x}_s)ds \right) \right\}.
\end{aligned}$$

The short rate in the default-free bond pricing formula is replaced by the default-adjusted short rate $(r+\lambda)(\mathbf{x}_s)$. Moreover, closed-form expressions are available for $\pi(t,T)$ if one uses affine models such as the CIR model for both r and λ.

Forms of recovery

- The first type of recovery assumptions is *recovery of market value*. This measures the change in market prices before and after the default, representing the loss in the bond's value associated with the default.
- Consider a defaultable bond $\pi(t, T)$ with face value 1 at T. The bond is said to have a fractional recovery of market value of δ at default time τ if the amount recovered in the event of default is equal to

$$h(\tau) = \delta\pi(\tau-, T) \text{ for } \tau \leq T,$$

where $\pi(\tau-, T)$ is the value of the claim just prior to default and $\delta \in [0, 1)$.

- With this recovery assumption, the price at t of the defaultable bond if there has not been a default up to date t is

$$\pi(t, T) = E^Q\left[\exp\left(-\int_t^T (r + (1-\delta)\lambda)(\mathbf{x}_s)ds\right)\right]$$

Forms of recovery

- A variant of recovery of market value is *recovery of face value* ($\delta(\tau) = \delta$), which is the closest to legal practice in the sense that debt with the same priority is assigned a fractional recovery depending on the outstanding notional amount but not on maturity or coupon. It is also the measure typically used in rating-agency studies.
- The third type of recovery assumptions is *recovery of treasury* ($\delta(\tau) = \delta p(\tau, T)$), under which the corporate bond in default is replaced with a treasury bond with the same maturity but a reduced payment. Unlike recovery of face value, it tries to correct for the fact that amounts of principal with long maturity should be discounted more than principal payments with short maturity.

Insuring against a corporate bonds credit loss with CDS

- A *credit default swap* (CDS) is a contract between a buyer, who obtains the right to sell bonds issued by a company, and a seller (of the CDS contract) who agrees to buy the bonds for their face values when a credit event occurs.
- The face value of a coupon-bearing bond is the principal that the issuer repays at maturity if it does not default. The total face value of the bonds is the notional principal of the CDS.
- The buyer of the CDS makes periodic payments to the seller until the expiration date T of the CDS or until a credit event occurs.

CDS spread and valuation

The total amount paid per year by the buyer, as a percentage of the notional principal per annum, is called the *CDS spread*. Suppose the reference bond has coupon dates $t_1, \cdots, t_M = T$ and the protection buyer pays a constant premium c at these dates prior to τ, the time of the credit event. Assume that the face value of the bond is 1 and that the protection seller pays $1 - \delta$ in the case of a credit event. Then the value v_b of the premium leg of the swap is

$$\begin{aligned} v_b &= cE^Q \left\{ \sum_{i=1}^{M} \exp\left(- \int_0^{t_i} r_s ds \right) I_{\{\tau > t_i\}} \right\} \\ &= cE^Q \left\{ \sum_{i=1}^{M} \exp\left(- \int_0^{t_i} (r_s + \lambda_s) ds \right) \right\} = c \sum_{i=1}^{M} \pi(0, t_i), \end{aligned}$$

where r_s is the short rate process, λ_s is the default intensity process and $\pi(0, t_i)$ is the price of the defaultable bond with zero recovery and maturity t_i.

CDS spread and valuation

- Similarly the value v_s of the default leg of the swap, to be paid by the protection seller, is

$$v_s = (1-\delta)E^Q\left\{\exp\left(-\int_0^\tau r_s ds\right) I_{\{\tau \le T\}}\right\}$$
$$= (1-\delta)\int_0^T E^Q\left\{\exp\left(-\int_0^\tau r_s ds\right)\right\} f^Q(t)dt,$$

  where f^Q is the density function of τ under the risk-neutral measure Q, assuming that the short rate process and the default intensity process are independent under Q.

- Let $\lambda^Q = f^Q/S^Q$ be the hazard function of τ under Q, it then follows that

$$v_s = (1-\delta)\int_0^T p(0,t)\lambda^Q(t)S^Q(t)dt,$$

  where $p(0,t)$ is the price of the default-free bond.

CDS spread and valuation

Note that under this independence assumption on the short rate and default intensity processes, we can further simplify $\pi(0, t_i)$ to

$$\pi(0, t_i) = E^Q\left[\exp\left(-\int_0^{t_i} r_s ds\right)\right] E^Q\left[\exp\left(-\int_0^{t_i} \lambda_s ds\right)\right] = p(0, t_i) S^Q(t_i).$$

In the absence of arbitrage, the premium c_{DS} of the default swap is the premium c for which $v_b = v_s$. Therefore, combining the formula for v_b and v_s yields

$$c_{DS} = (1 - \delta) \int_0^T p(0, t) S^Q(t) \lambda^Q(t) dt \Big/ \sum_{i=1}^{M} p(0, t_i) S^Q(t_i).$$