## 3 So far: Conditional Mean:

$$
\begin{aligned}
r_t &= E\left(r_t | I_{t-1}\right) + \varepsilon_t \\
&= \mu\left(X_{t-1} : \beta\right) + \varepsilon_t
\end{aligned}
$$

- where $E\left(\varepsilon_t | X_{t-1}\right) = 0$
- where the expected return function $\mu\left(X_{t-1} : \beta\right)$ is usually a linear function.
- and $X_{t-1}$ is the information set (or a sufficient statistic). $X_{t-1}$ might include lagged $r_t$, and other conditioning variables.
- The variation in expected returns is order of magnitudes smaller that the variation in $\varepsilon_t$.

- Thus far, we have assumed that $Var_t(\varepsilon_t) = Var(\varepsilon_t) = \sigma^2$ is a constant.
- This is not a realistic assumption in finance.
- We will consider models of the variance of $\varepsilon_t$ where $Var_t(\varepsilon_t) \neq Var(\varepsilon_t)$.
- Before we do that, we will briefly introduce another method of estimation, maximum likelihood.

## 4 Maximum Likelihood Estimation

- Suppose we have observed a sample of size T on some random variable $Y_t$. Let $f_{Y_1,...,Y_T}(y_1, y_2, ..., y_T|\theta)$ denote the joint density of $Y_1, ..., Y_T$.
- The density depends on unknown parameters $\theta$ to be estimated.
- We view the density as a function of $\theta$, given the sample data.
- Example. Supose T i.i.d observations drawn from a $N(\mu, \sigma^2)$ distribution. Then, $\theta = (\mu, \sigma^2)$. The joint density is:

$$f_{Y_1,...,Y_T}(y_1, y_2, ..., y_T|\theta) = \Pi_{t=1}^{T} f_{Y_t}(y_t|\theta)$$

- We can also define the log-likelihood:

$$\ell(...|\theta) = \log f_{Y_1,...,Y_T}(y_1, y_2, ..., y_T|\theta) = \Sigma_{t=1}^{T} \log f_{Y_t}(y_t|\theta)$$

$$= \Sigma_{t=1}^{T} \log \left\{ \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma^2}} e^{\frac{-(y_t-\mu)^2}{2\sigma^2}} \right\}$$

$$= -\frac{T}{2}\Sigma_{t=1}^{T} \log(2\pi) - \frac{T}{2}\Sigma_{t=1}^{T} \log \sigma^2 - \Sigma_{t=1}^{T} \frac{(y_t-\mu)^2}{2\sigma^2}$$

- Note: We can generalize this expression: The mean and the variance might be time-varying.
- We want to find the value of $\theta$ that makes the sample likelihood as large as possible. This value is called the maximum likelihood estimate of $\theta$.

- The properties of $\hat{\theta}^{MLE}$ are, in general, well-understood.
- We have the convenient simple result:

$$\sqrt{T}\left(\hat{\theta}-\theta\right)\rightarrow^{d} N\left(0,\Im^{-1}T^{-1}\right)$$

- where $\Im$ is called the information matrix and can be estimated in one of two ways. The first way involves the Hessian:

$$\hat{\Im}_1 = -T^{-1}\frac{\partial^2 \ell}{\partial \theta^2}|_{\hat{\theta}}$$

- The second method is:

$$\begin{aligned}\hat{\Im}_2 &= T^{-1}\Sigma_{t=1}^{T} h\left(\hat{\theta},Y_t\right) h\left(\hat{\theta},Y\right)' \\ h\left(\hat{\theta},Y_t\right) &= \frac{\partial \log f_{Y_t(y_t|\theta)}}{\partial \theta}|_{\hat{\theta}}\end{aligned}$$

- In finite samples, the two methods might give you different estimates. Which one to choose?
- If the two measures differ, it might mean that the model is misspecified (White (1982)).

- Likelihood Ratio Test:
  - Estimate the likelihood $\ell(...|\theta)$ without restrictions, obtaining $\hat{\theta}^u$ and $\ell\left(...|\hat{\theta}^u\right)$.
  - Estimate the likelihood $\ell(...|\theta)$ with some restrictions, obtaining $\hat{\theta}^r$ and $\ell\left(...|\hat{\theta}^r\right)$.
  - It must be the case that $\ell\left(...|\hat{\theta}^u\right) > \ell\left(...|\hat{\theta}^r\right)$ (better fit with unrestricted model).
  - Then, we can prove that:

$$2\left(\ell\left(...|\hat{\theta}^u\right) - \ell\left(...|\hat{\theta}^r\right)\right) \sim .\chi^2 (\text{number of restrictions})$$

  - Very useful test.
  - Nested tests

- The estimation of the ARCH(1) (or any finite order) model is easy if we assume that $z_t$ is iid $N(0,1)$. Then, the conditional density of $r_t$ is

$$f_{r_t}(r_t|x_{t-1}...) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\sigma_t^2}}e^{\frac{-\left(y_t-c-\phi x_{t-1}\right)^2}{2\sigma_t^2}}$$

- where $\sigma_t^2 = V(r_t|x_{t-1}...) = \kappa + \alpha\left(y_t - c - \phi x_{t-1}\right)^2 = \kappa + \alpha\varepsilon_{t-1}^2$.

- We can maximize the likelihood with respect to $c$, $\phi$, $\kappa$, $\alpha$. The function to maximize is:

$$-\frac{T}{2}\Sigma_{t=1}^{T}\log\sigma_t^2 - \Sigma_{t=1}^{T}\frac{\left(y_t-\mu\right)^2}{2\sigma_t^2}$$

- We can make other assumptions about the density of $z_t$, but we have to make some assumptions.

- Q: Suppose, we are not willing to make any distributional assumptions about $z_t$. Can we still consistently estimate the parameters?
- If we can write:

$$
\begin{aligned}
E\left(z_t | x_{t-1}...\right) &= 0 \\
E\left(z_t | x_{t-1}...\right) &= 1
\end{aligned}
$$

then, maximizing the normal log-likelihood function will still provide consistent estimator of the parameters (Bollerslev and Woolridge (1992), Glosten et al. (1993)).
- But we have to adjust the standard errors (constent, but not efficient estimator).
- The intuition will come when we introduce GMM.
- This is called Quasi-Maximum Likelihood Estimation.
- Testing for ARCH (LR test): Impose the restrictions that the $\alpha' s$ are equal to zero.

Other methods for modelling volatility:

- French, Schwert and Stambaugh (1987): The idea is to use higher frequency data to estimate the variance as:
$$\sigma_t^2 = \frac{1}{k}\sum_{d=1}^{k} \varepsilon_{t+d}^2$$
  - where $\varepsilon_t$ are measured in days, and we estimate monthly variance.
  - This produces a monthly sequence $\left\{\hat{\sigma}_t^2\right\}$ of estimated variances.
  - There is nothing wrong with this scheme.
- Another method: AR model for volatility:
$$|\varepsilon_t| = \eta + \gamma\,|\varepsilon_{t-1}| + \upsilon_t$$
  - where the $\varepsilon_t$ are estimated from a first step procedure.
  - There is nothing wrong with this method. It provides another model for stochastic volatility.
- Since we don't observe true volatility, we can't really say which method is the best at capturing it.
- Long memory fractionally integrated processes

- GARCH in Mean or GARCH-M models (Engle, Lilien, and Robins (1987)):

$$\begin{aligned} r_t &= a + b x_{t-1} + c\sigma_t^2 + \varepsilon_t \\ \varepsilon_t &= z_t \sigma_t \\ \sigma_t^2 &= \kappa + \beta(L)\,\sigma_{t-1}^2 + \alpha(L)\,\varepsilon_{t-1}^2 \end{aligned}$$

- The difference from the previous models is that the volatility enters also in the mean of the return.
- This is exactly what Merton's (1973, 1980) ICAPM produces–risk-return tradeoff.
- It must be the case that $b > 0$.
- The GARCH-M is estimated with ML or QML.
- The evidence on the risk-return tradeoff is not good.
- French, Schwert and Stambaugh conduct similar tests, but their method is a two-step procedure (inefficient, and potentially problematic.)

- Cutting edge: MIDAS estimators–Mixed Data Sampling estimators (Ghysels, Santa-Clara, Valkanov (2002a,b))
- Idea: Use data at different frequencies to estimate the risk-return tradeoff

$$R_{t+1} = \alpha + \beta \left( \sum_{d=1}^{D} w_d r_{t-d}^2 \right) + \varepsilon_t$$

- where $R_t$ is at monthly frequency and $r_t$ is at daily.
- The weights $w_d$ sum up to one.
- Given that vol is persistent, there might be many weights to estimate, which would result in inefficient estimators.
- Hence, we parameterize $w_d(\theta)$ and estimate the shape of the weights.
- There are several advantages:
  - Higher frequency data, i.e. better estimates of vol.
  - Joint estimation of $\theta$, $\alpha$, $\beta$
  - Flexibility of weights
  - Easy to implement other variables, asymmetries.
  - Estimation: NLS. Standard LS theory applies.

- Stochastic volatility models:

$$
\begin{aligned}
r_t &= a + b r_{t-1} + \varepsilon_t \\
\varepsilon_t &= z_t \sigma_t \\
\sigma_t &= \kappa + \beta \sigma_{t-1} + v_t
\end{aligned}
$$

- The difference here is that the shocks that govern the volatilty are not necessarily $\varepsilon_t^{2\prime} s$.
- This is really a discretization of a continuous-time model, where the mean and the variance follow two OU processes.
- Stochastic vol models can be estimated by MLE or other methods.