Statistical Modeling, Inference, and Forecasting

Hui Chen

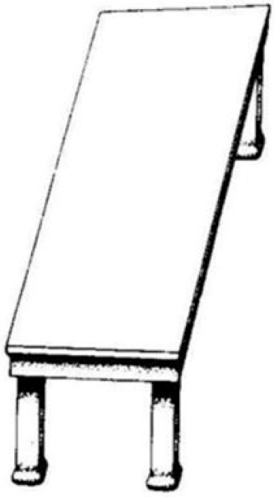
MIT Sloan

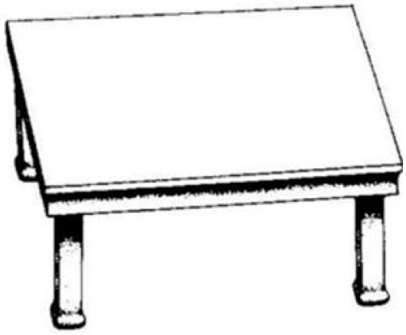
15.450, Spring 2023

Outline

Technical Analysis

Eye-Ball Econometrics





Link

1. To time the market, we first need to form a view of the market by forecasting future returns.
2. Let's construct a statistical model to predict (monthly) excess returns on the market index:

$$R^e_{t+1} = b_0 + b_1 x_t + \varepsilon_{t+1}$$

   - ↪ Signal x_t: price-dividend ratio
   - ↪ Estimation: $\hat{b}_0, \hat{b}_1$
   - ↪ Inference: Is the signal useful ($b_1 \neq 0$)?
   - ↪ Predicted return:

$$E_t[R^e_{t+1}] = \hat{b}_0 + \hat{b}_1 x_t$$

3. Build the portfolio based on our predictions. Let the fraction of the portfolio invested in the market index be:

$$\omega(x_t) = kE_t[R^e_{t+1}]$$

   Q: Does such an approach make economic sense?

Outline

Statistical Learning

- We want to learn the (conditional) distribution of y from the observed data.
  - ↪ Stock returns, volatility, corporate earnings, default rates, GDP growth, inflation ...
  - ↪ Usually an intermediate step. The ultimate purpose is to make investments, manage risk ...

- The true (and unknown) data-generating process (DGP):

$$y = f(\mathbf{x}) + \varepsilon$$

  - ↪ y: response (dependent variable)
  - ↪ $\mathbf{x}$: features (independent variables, covariates)
  - ↪ ε: additional randomness; independent of $\mathbf{x}$ and has zero mean.

- Our hope is to find an $\hat{f}$ that is "close" to f.
  - ↪ By learning from a sample of examples: $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$.
  - ↪ This is referred to as "supervised learning."
  - ↪ What does "being close" mean?
  - ↪ Two key challenges

- True DGP (f):

$$R^e_{j,t} = \alpha_j + \beta_j R^e_{m,t} + \gamma_j R_{hml,t} + \delta_j R_{smb,t} + \varepsilon_{j,t}$$

- Fitted model ($\hat{f}$):

$$R^e_{j,t} = \hat{\alpha}_j + \hat{\beta}_j R^e_{m,t} + \hat{\varepsilon}_{j,t}$$

What is a good model?

Prediction: $\hat{y} = \widehat{f}(\mathbf{x})$

Prediction error: $y - \hat{y}$

- Mean-Squared Error (MSE): A common criterion to measure a model's performance.

$$E\left[(y - \hat{y})^2\right]$$

  - ↪ Expectation is taken with respect to the true distribution, which is usually unknown.
  - ↪ With a given sample:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}(\mathbf{x}_i)\right)^2$$

  - ↪ MSE is a specific kind of loss function, which expresses a particular view on the losses of different types of prediction errors.
  - ↪ Many alternative loss functions. How to choose? Financial decision making should play a role.

How to model and estimate $\widehat{f}$

- Parametric methods
  1. Select a functional form (including assumptions about ε).
     Example:
     $$f(\mathbf{x}_i) = \theta_0 + \theta_1 x_{i,1} + \cdots + \theta_p x_{i,p}, \quad E[\varepsilon_i] = 0, E[\mathbf{x}_i \varepsilon_i] = 0$$
  2. Train the model (estimate θ) using a sample of data (training set).
  3. Ideally, evaluate the fitted model out of sample (using test set).

- Non-parametric methods

- Theory-driven vs. data-driven view
  - ↪ Chris Anderson, "The End of Theory." Wired, 2008.
  - ↪ Tim Hardford, "Big data: are we making a big mistake?" Financial Times, 2014.

- Key concepts we will discuss later:
  - ↪ In-sample vs. out-of-sample
  - ↪ Bias-variance trade-off
  - ↪ Model and feature selection

Outline

Estimators

Estimator

Let θ be the parameter vector, and $\psi = h(\theta)$ a function of the model parameters. An *estimator* of ψ is a function of the observed sample, $\delta(\mathbf{x})$.

Random sample

A sample of n observations, $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, is a *random sample* if the n observations are drawn independently from the same population.

- The random sample $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ is said to be independent, identically distributed (IID).

Example

Estimating the Sharpe ratio

Consider a sample $\mathbf{x}$ of excess returns:

$$R_t^e = \mu + \sigma\varepsilon_t, \quad t = 1, ..., T$$

where ε_t are $IID\,\mathcal{N}(0,1)$ random variables, and $\sigma > 0$. μ and σ are two unknown parameters. How to estimate the Sharpe ratio, $\psi = \mu/\sigma$?

- An intuitive estimator of ψ is

$$\delta(\mathbf{x}) = \frac{\widehat{\mu}}{\widehat{\sigma}}$$

where

$$\widehat{\mu} = \frac{1}{T}\sum_{t=1}^{T} R_t^e \quad \text{and} \quad \widehat{\sigma} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(R_t^e - \widehat{\mu}\right)^2}$$

- Another estimator: $\delta(\mathbf{x}) \equiv 0.2$. Valid but not useful.

Estimators

- The estimator is a random variable! The realized value of the estimator depends on the sample, which is random.
- A single realization of the estimator is called an *estimate*, or a *point estimate*.
- What makes a good estimator?
  - ↪ Typically, we want estimators that are informative, i.e., realized values of the estimator of $h(\theta)$ are close to the true value of $h(\theta)$.
  - ↪ Again, the proper criterion depends on what we use the estimator for.
- To facilitate decision making, typically augment point estimates with further information, such as standard errors (discussed later).

Properties of Estimators

- Which is a better estimator of the Sharpe ratio,

$$\delta(\mathbf{x}) = \frac{\widehat{\mu}}{\widehat{\sigma}} \quad \text{or} \quad \delta(\mathbf{x}) = 0.2\,?$$

- Want estimators with useful properties.

Consistency

An estimator $\delta(\mathbf{x})$ of $\psi = h(\theta)$ is consistent if, for all true values of the parameter θ_0, as the sample size increases to infinity the estimator converges to $\psi(\theta_0)$ (in probability).

Unbiasedness

An estimator $\delta(\mathbf{x})$ of $\psi = h(\theta)$ is unbiased if, given the true value of θ, the conditional expectation of $\delta(\mathbf{x})$ equals ψ.

- We generally prefer consistent estimators. Bias may be impossible to avoid. We just don't want it to be too large.

Method of Moments Estimator (MM)

- Consider a sample of independent and identically distributed (IID) observations drawn from the distribution family with density $p(x|\theta_0)$ (a random sample),

$$\mathbf{x} = (x_1, \ldots, x_n)$$

- Want to estimate the N-dimensional parameter vector θ_0.
- Consider a vector of N functions $f_j(x, \theta)$ ("moments").
- Suppose we know that

$$\begin{aligned} &\mathrm{E}[f_1(x,\theta_0)] = \cdots = \mathrm{E}[f_N(x,\theta_0)] = 0, && \text{if } \theta = \theta_0 \\ &\textstyle\sum_{j=1}^{N} \left(\mathrm{E}[f_j(x,\theta)]\right)^2 > 0, && \text{if } \theta \neq \theta_0 \end{aligned} \qquad \text{(Identification)}$$

MM Estimator

The method-of-moments estimator $\widehat{\theta}$ of parameter θ_0 is the solution to

$$\widehat{\mathrm{E}}[f(x,\widehat{\theta})] \equiv \frac{1}{n}\sum_{i=1}^{n} f(x_i,\widehat{\theta}) = 0$$

Example

MM, Mean-Variance

- Suppose we have a sample from a distribution with mean μ_0 and variance σ_0^2.
- To estimate the parameter vector $\theta_0 = (\mu_0, \sigma_0)'$, $\sigma_0 \geq 0$, choose the functions $f_j(x, \theta)$, $j = 1, 2$:

$$f_1(x_t, \theta) = x_t - \mu$$
$$f_2(x_t, \theta) = (x_t - \mu)^2 - \sigma^2$$

- Verify that $\mathrm{E}[f(x_t, \theta_0)] = 0$.
- Verify that if $\theta \neq \theta_0$, then $\mathrm{E}[f(x_t, \theta)] \neq 0$.
- Parameter estimates:

Maximum Likelihood Estimator (MLE)

Likelihood function

Description of how observed data, $\mathbf{x} = (x_1, ..., x_n)$, depends on the model parameters, θ. This is given by the probability distribution of the data conditional on the model parameters.

$$p(\mathbf{x}|\theta)$$

MLE

The estimator $\delta(\mathbf{x})$ of the parameter vector θ defined by

$$\delta(\mathbf{x}) = \arg\max_{\theta} \; p(\mathbf{x}|\theta) = \arg\max_{\theta} \ln p(\mathbf{x}|\theta)$$

is called the Maximum Likelihood Estimator (MLE).

- Typically solved numerically. In R, use the function 'mle' in the 'stats4' package.

Maximum Likelihood

- MLE has an intuitive interpretation as the parameter value that maximizes the likelihood of the observed sample.
- Under mild regularity conditions, MLE is a consistent estimator.
- In general MLE is not unbiased.

Example

MLE, Gaussian distribution

- IID Gaussian observations, mean μ, variance σ^2. Parameter vector $\theta = (\mu, \sigma^2)'$.
- The log likelihood for the sample $x_1, ..., x_T$ is

$$\ln p(\mathbf{x}|\theta) = \ln \prod_{t=1}^{T} p(x_t|\theta) = \sum_{t=1}^{T} \ln p(x_t|\theta) = \sum_{t=1}^{T} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x_t - \mu)^2}{2\sigma^2}$$

- MLE:

$$\widehat{\theta} = \arg\max_{\theta} \ln p(\mathbf{x}|\theta)$$

- First-order optimality condition:

- MLE:

Example

MLE: exponential distribution

- IID sample x_t, $t = 1, ..., T$ from an exponential distribution

$$p(x|\theta) = \theta e^{-\theta x}, \quad \theta > 0$$

- Likelihood function

- First-order optimality condition:

- MLE:

Outline

Example: S&P GSCI Index

- Model daily changes in S&P GSCI index.
- The S&P GSCI index is a composite commodity index, maintained by S&P.
  "The S&P GSCI® provides investors with a reliable and publicly available benchmark for investment performance in the commodity markets. The index is designed to be tradable, readily accessible to market participants, and cost efficient to implement. The S&P GSCI is widely recognized as the leading measure of general commodity price movements and inflation in the world economy."

Source: Standard & Poor's.

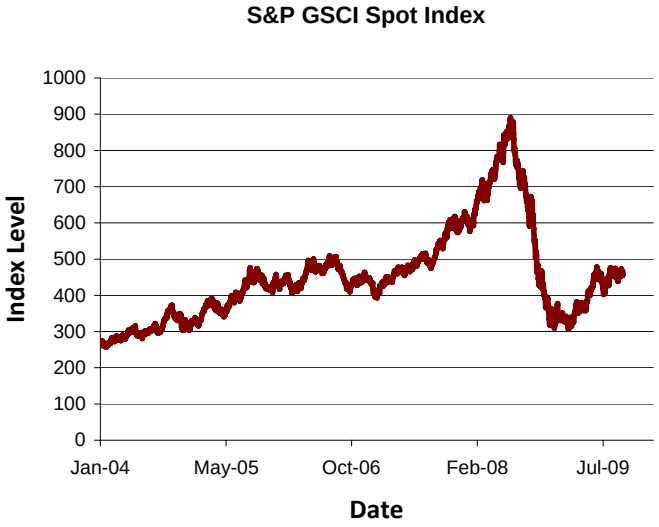
- Changes in daily spot index levels:

$$z_t = \ln \frac{P_t}{P_{t-1}}$$

- Consider the sample period from 09/02/1999 to 09/23/2009.
- De-mean the series of daily changes, define

$$x_t = z_t - \frac{1}{T} \sum_{t=1}^{T} z_t$$

Example: S&P GSCI Index



Model of the Shocks

- Model shocks as

$$x_t = \sigma \varepsilon_t$$

  where ε_t has zero mean and unit variance.

- Model ε_t using the Student's t distribution

$$p(\varepsilon_t|\nu) = \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)\sqrt{\pi(\nu-2)}} \left(1 + \frac{\varepsilon_t^2}{\nu-2}\right)^{-(\nu+1)/2}, \quad \nu > 2$$

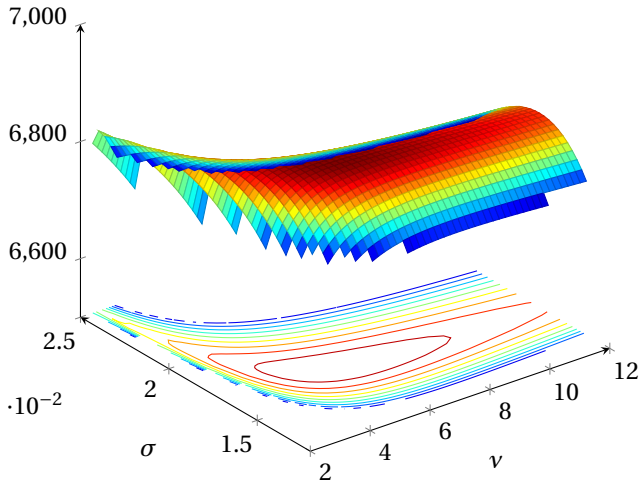
- $\sqrt{\nu/(\nu-2)}\varepsilon_t$ have the Student's t distribution with ν degrees of freedom.
- Γ is the Gamma function, $\Gamma(x) = \int_0^\infty z^{x-1} e^{-z}\, dz$.

MLE for Student's t Distribution

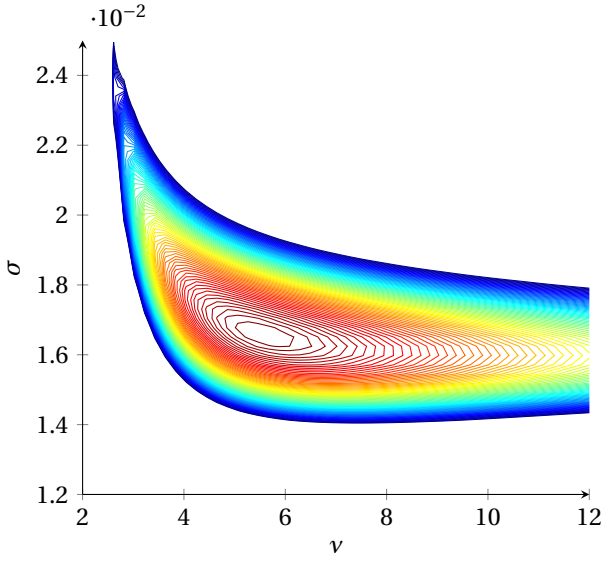
- Log-likelihood function

$$\ln p(\mathbf{x}|\theta) = \sum_{t=1}^{T} \ln\left(\frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)\sqrt{\pi(\nu-2)}}\right) - \frac{\nu+1}{2}\ln\left(1+\frac{x_t^2}{(\nu-2)\sigma^2}\right) - \ln(\sigma)$$

Log Likelihood



Level Curves for Log-Likelihood



Why should we care?

Example: S&P GSCI Index

- MLE estimate of ν is 5.51.
- A "5-σ" daily event on average occurs once every 3.74 years.
- What if we had assumed Gaussian distribution for ε_t?
  - ↪ A "5-σ" daily event once every 6,922 years.
- Much heavier tails under the Student's t distribution!
- Q: Which one is the better model?

Summary

- The framework for going from statistical modeling to financial decision making
  - ↪ parametric vs. non-parametric
  - ↪ MSE, loss function
  - ↪ Training set vs. test set
- Estimators
  - ↪ Unbiasedness, consistency
  - ↪ MM, MLE