

# Complex model to complex data

— the statistical approach

**Feng Li**

`<feng.li@cufe.edu.cn>`

**School of Statistics and Mathematics  
Central University of Finance and Economics**

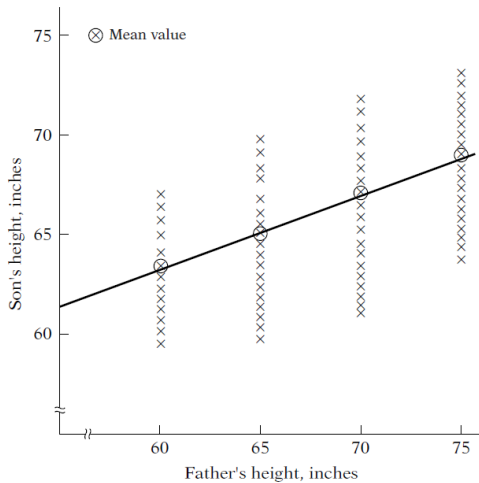
## Have you ever thought about...

- Why the weather forecast is not accurate sometimes?  
(rain or not  $\Leftrightarrow$  cloudy, humidity, historical data)
- How does the email filter know whether a mail is a spam or not?  
(spam or not  $\Leftrightarrow$  sender, keywords)
- Can we predict the next financial crisis?  
(Next crisis time  $\Leftrightarrow$  when was last time, stock prices, exchange rate, unemployment rate)
- ...

**Statistical modeling** is trying to formalize (**model**) the relationships among the variables (**data**).

## Toy examples

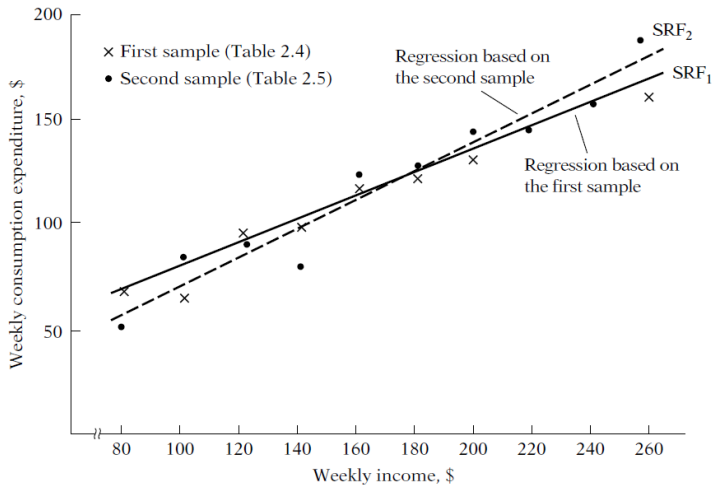
### ↪ Father's height vs son's height



Gujarati, D. N. (2003). Basic Econometrics. 4th.

# Toy examples

## ↪ Family income and consumption

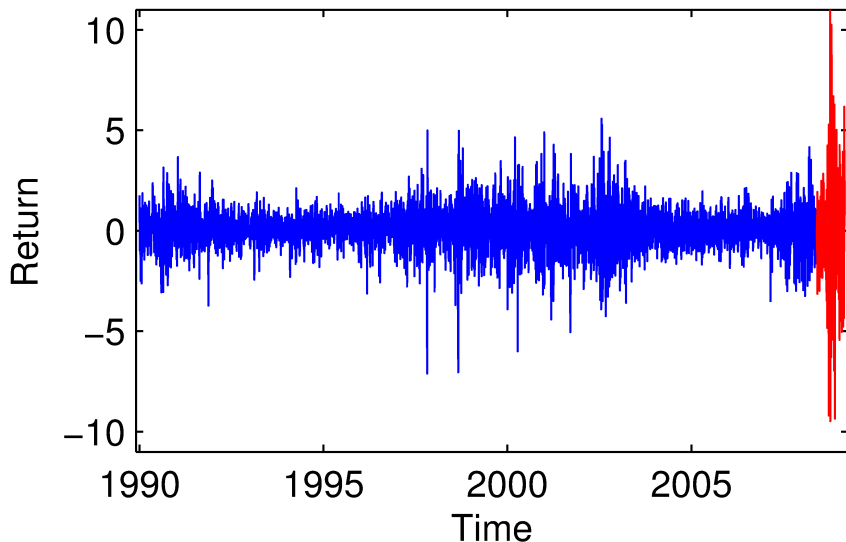


Gujarati, D. N. (2003). Basic Econometrics. 4th.

- Models are simple.
- Works well at most situations.
- Easy to imagine and implement.
- It takes less than 1 second to have the result with a laptop.

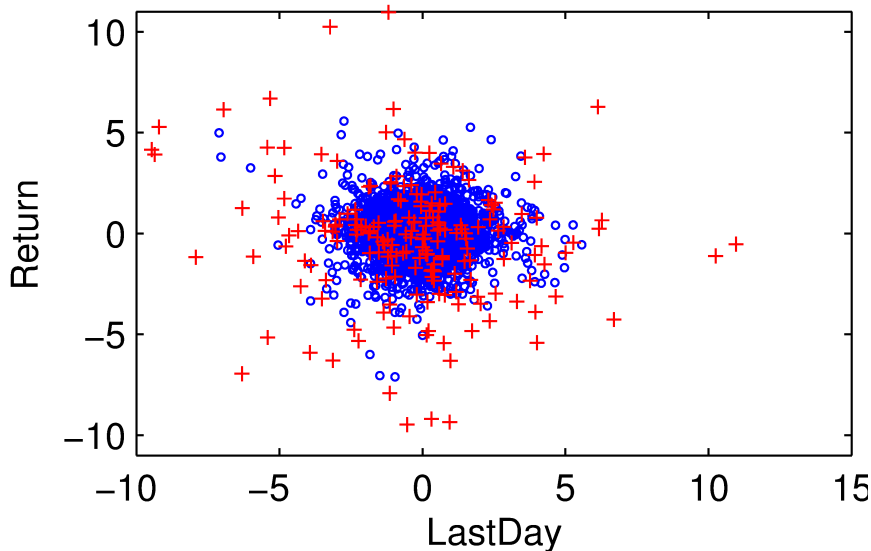
## The typical data in finance

↪ Daily stock market returns



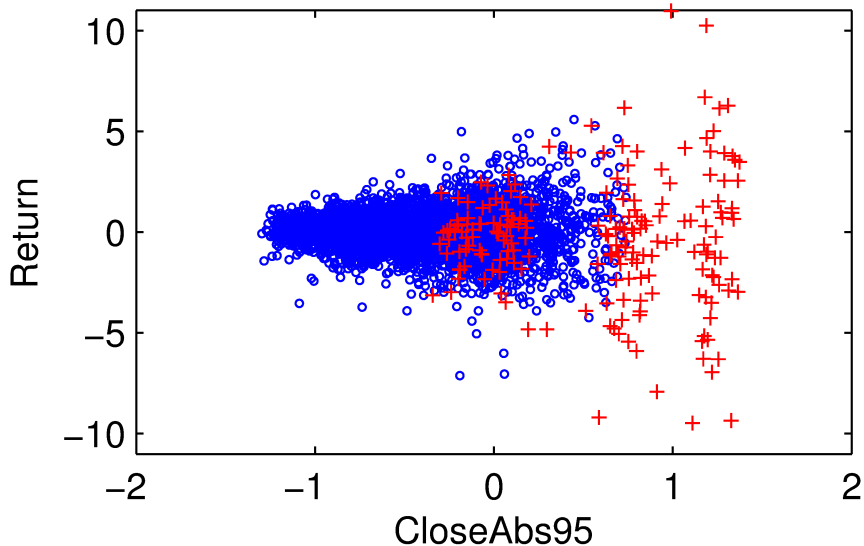
## The typical data in finance

↳ Daily stock market returns, a closer look



## The typical data in finance

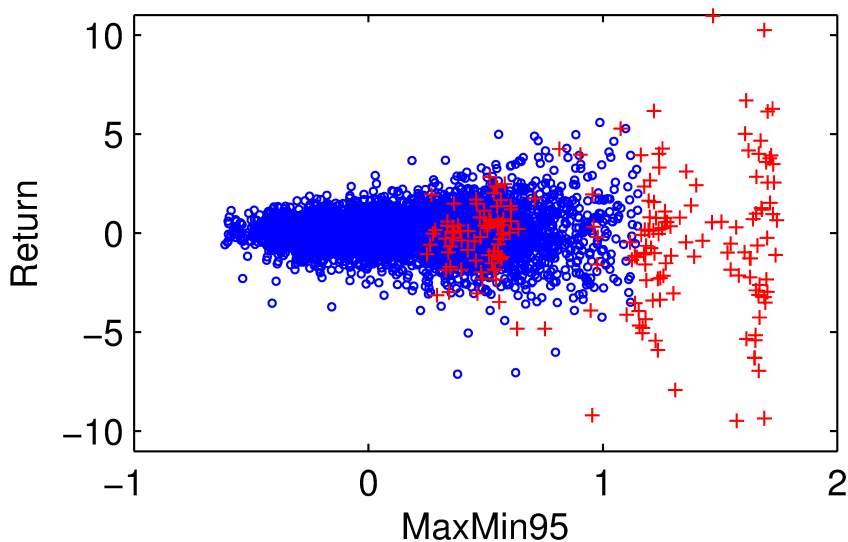
↪ Daily stock market returns, a closer look





## The typical data in finance

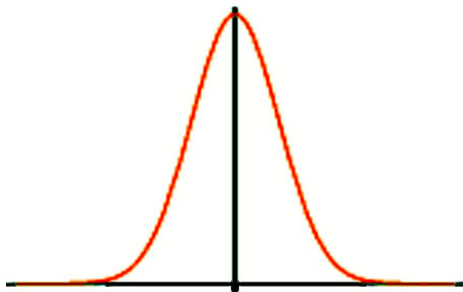
↳ Daily stock market returns, a closer look



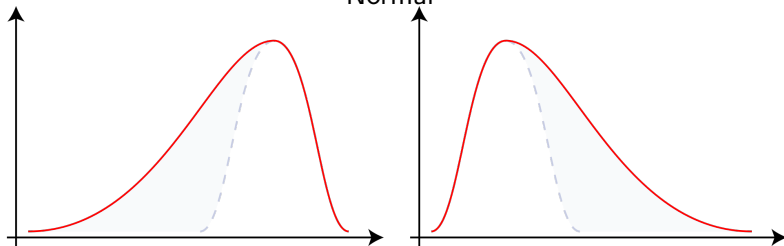
## What can we find?

- This does not look like as **normal** (think about the mean and variation)!
- How do we describe it in the language of statistics?
  - We use **mean** and **variance** (*standard deviation*) to describe normality.
  - We use **skewness**, and **kurtosis** (*degrees of freedom*) to detect the abnormal events.

## Normal and not normal



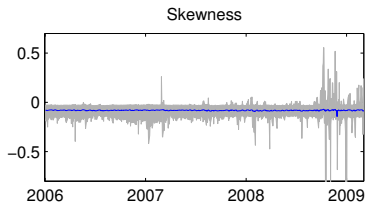
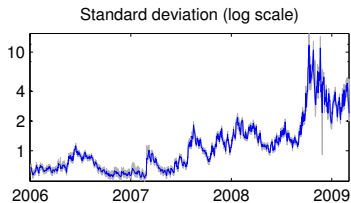
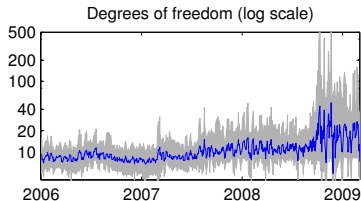
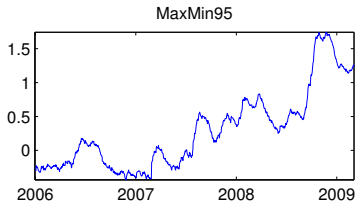
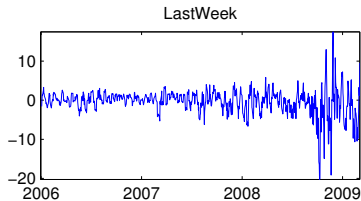
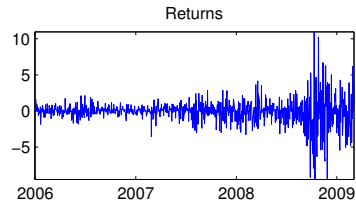
Normal



Negative Skew

Positive Skew

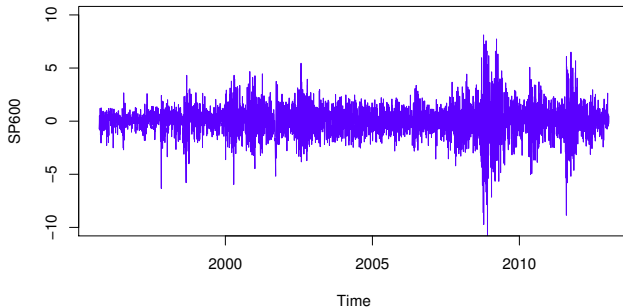
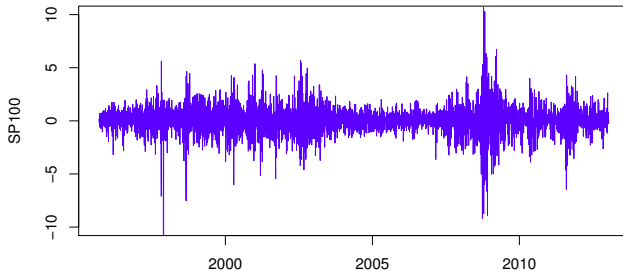
# Detecting the financial crisis



## What if we insist using the normal model?

- The model will be misspecified.
- The conclusion based on that model can lead to a wrong decision.
- But people still do it anyway!
  - The normal model is simpler anyway.
  - We eventually do not know that we are wrong.
  - The computational tools are still feasible for everyone to use.
    - There was no ready-to-use computer software to use for this model.
    - The model takes a night to estimate with a cluster.

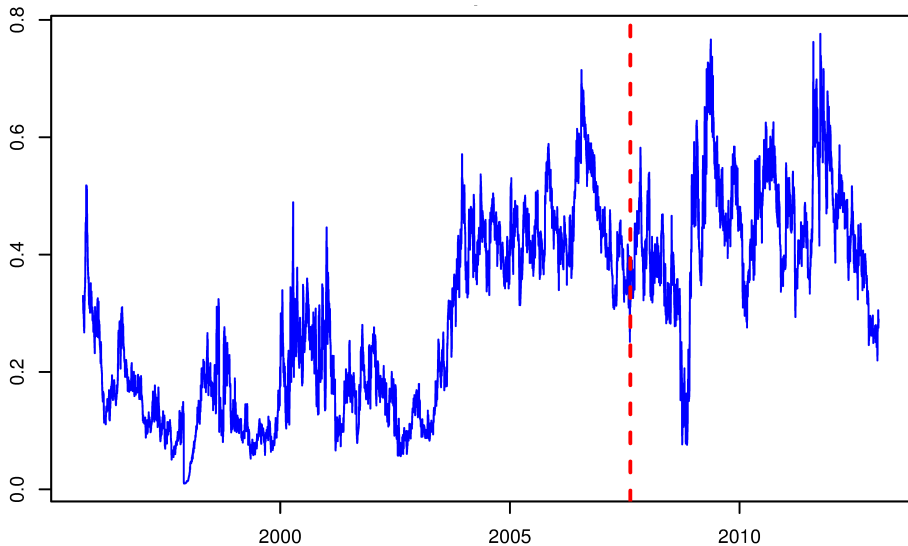
## A more complicated situation



## Our interests

- The S&P100 index includes the largest and most established companies in the U.S.
- The S&P600 index covers the small capitalization companies which present the possibility of greater capital appreciation, but at greater risk
- We are not only interested to detect the extreme events of a single stock, but also the co-movement of a two or more stocks.
  - What will happen to S&P100 if S&P600 collapses?
  - We call this as **tail-dependence**—the dependence only happens when extreme events happen.
  - What are the underlying factors that are connected to this dependence?

## The dependence on the tail





# Knowing the elephant

## ↳ The trend of statistical modeling

- In the 1950s, linear regression model was considered as very advanced which is now the standard course content for university students.
- The data are much more complicated nowadays we meet.
  - Numerical, categorical, brain image...
  - A few observations to millions by millions.
  - Very high-dimensional data are not rare anymore.

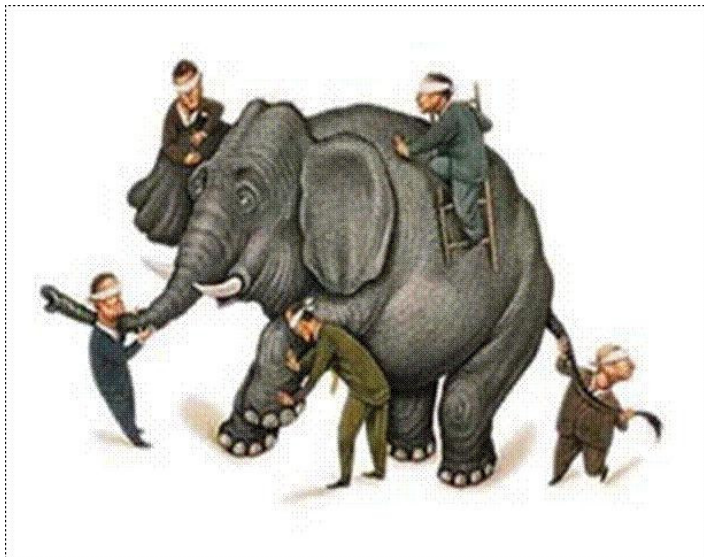
# Knowing the elephant

## ↳ The common procedure statistical modeling

- Data collection
- Model estimation
- Model evaluation
- Model comparison
- Prediction (if needed)

## Knowing the elephant

↪ Can we have a model that is big like an elephant?



## Knowing the elephant

- Sophisticated models are essential for such situations.
- In principle, the complicated model should be able to capture more complicated data features.
- Estimating such model is not easy.
- There is huge space to explore.
  - The computer is still not fast enough.
  - Techniques like parallel computing should be used to speed up the computation.
  - Statistics with big data is the new challenge.