Lecture 2: Overview of Financial Data and R

Brian Thelen
258 West Hall
bjthelen@umich.edu

Statistics 509 - Winter 2022

Outline of topics

- Data
- More R with financial data
  - Basics
  - Loading data
  - Plotting
- Financial data sets
  - Stock prices, returns, log-returns
  - Bonds: corporate and treasury
  - Indices: price and value-weighted
- Example Statistical Analyses

Financial data sets

- Will have a number of data sets used in class.
- Some of this will be from the WRDS (Wharton Research Data Service) – this is a wonderful resource available to U-M students and faculty.
  - There is a wide variety of carefully curated data sets.
  - These data come with 1 year delay and are for research purposes.
  - Good and online/real-time financial data is very expensive
- Yahoo Finance also has a number of indices

Basics I

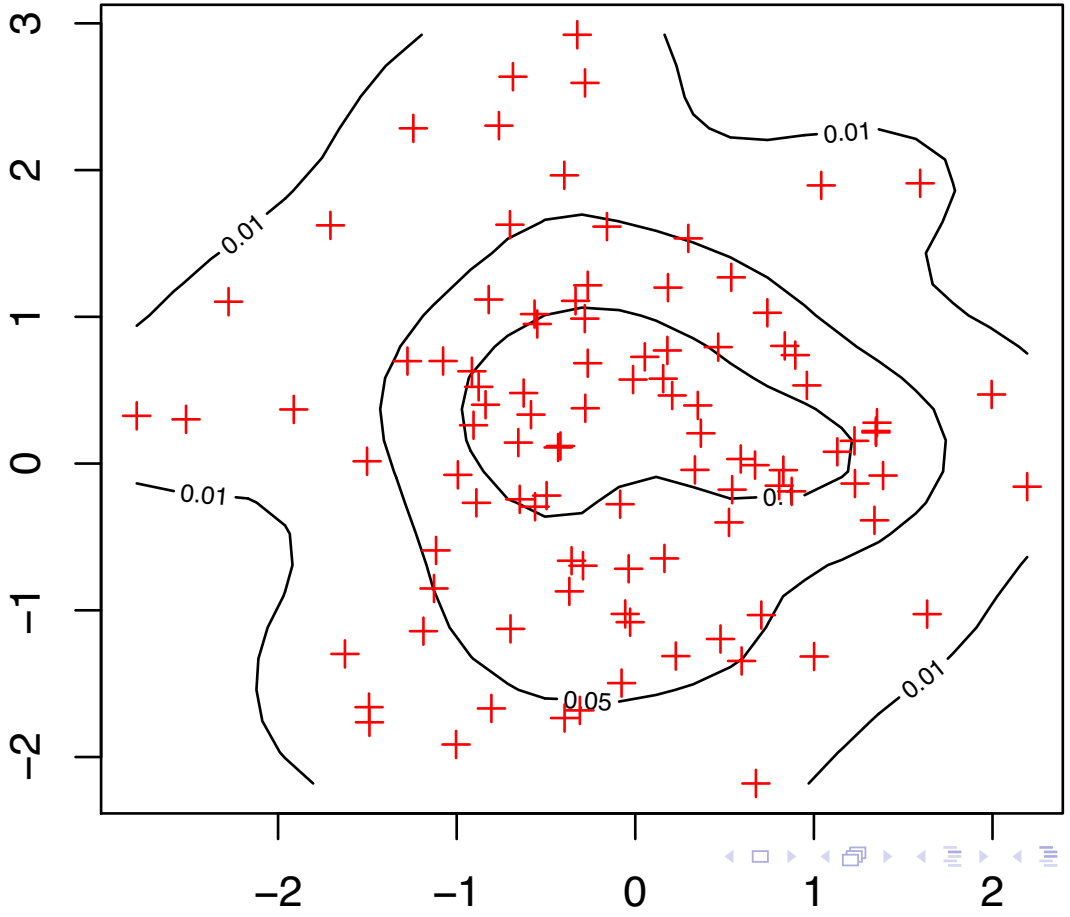
- Setting the working directory:

```
setwd("directory")
```

- Installing packages:

```
install.packages("MASS",repos="http://cran.us.r-project.org")
library("MASS")
set.seed(2016)
x = matrix(rnorm(200),nrow=100,ncol=2);
f = MASS::kde2d(x[,1],x[,2]); # or simply kde2d
contour(f, levels  =  c(0.01, 0.05, 0.1) )
points(x[,1],x[,2],col="red",pch=3)
```

Basics II



Loading data

Launch R or better Rstudio and load a file with daily prices of 4 stocks International Business Machines (IBM), Intel (INTC), Apple (AAPL), and Microsoft (MSFT):

```
dat = read.csv("stocks_ibm_intc_aapl_msft_1970_2015.csv",header = T)
nd = dim(dat)
names(dat)

[1] "PERMNO" "date"   "TICKER" "PRC"
```

The variable dat is a 38826×4 array. Here are the last few rows

```
tail(dat)

      PERMNO       date TICKER   PRC
38821  59328 12/23/2015   INTC 35.00
38822  59328 12/24/2015   INTC 34.98
38823  59328 12/28/2015   INTC 34.93
38824  59328 12/29/2015   INTC 35.44
38825  59328 12/30/2015   INTC 34.99
38826  59328 12/31/2015   INTC 34.45
```

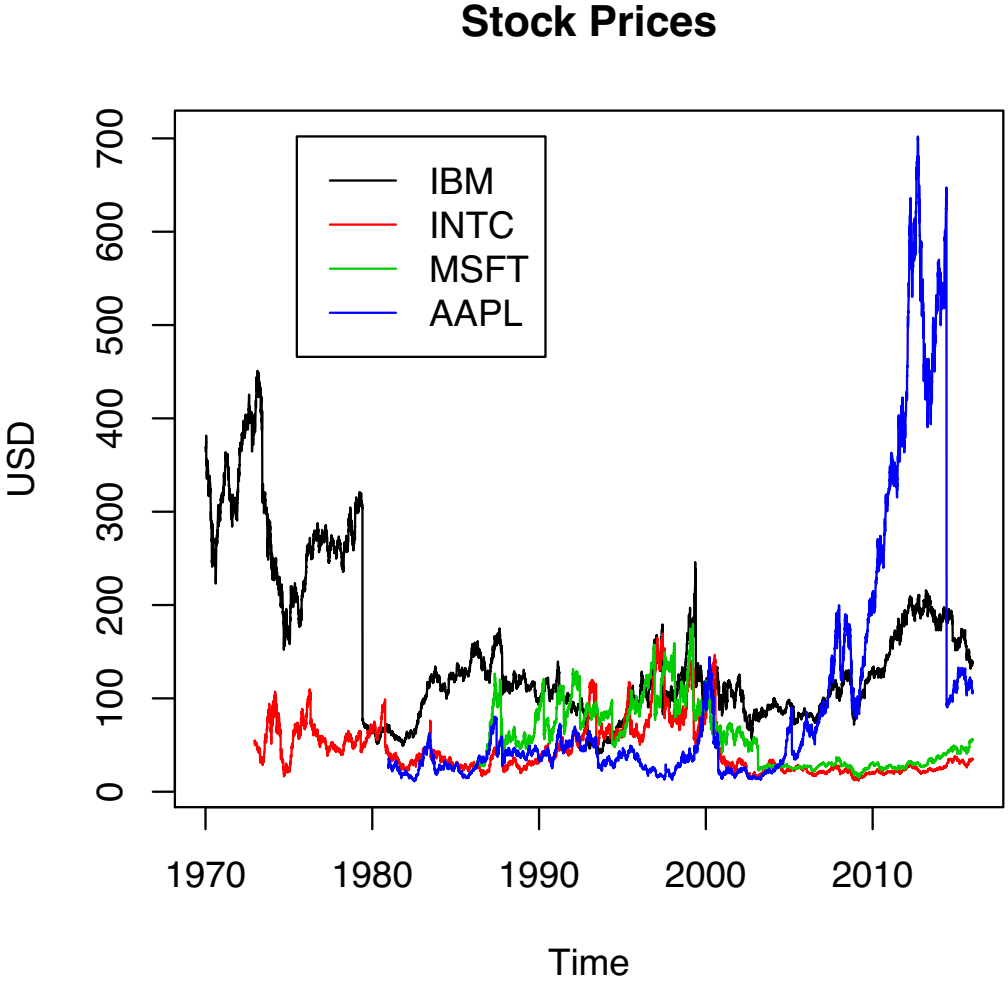
Converting the dates to Date structure and extracting stocks

```
t<-as.Date(dat$date,format="%m/%d/%Y")
stocks = list("IBM"=c(),"INTC"=c(),"MSFT"=c(),"AAPL"=c());
times  = stocks;
for (tick in c("IBM","INTC","MSFT","AAPL")){
  idx0 =  which(dat$TICKER==tick);
  idx1 =  which(is.na(dat$PRC[idx0])==FALSE);
  times[[tick]]    = t[idx0[idx1]];
  stocks[[tick]] = abs(dat$PRC[idx0[idx1]])
}
```

NOTE: Observe that we take the abs value of dat$PRC. This is because in the CRSP data base "negative" prices indicate that this was a bid/ask average and not a closing price.

Plot the stock prices on the same graph...

```
x0 = min(times$IBM,times$INTC,times$MSFT,times$AAPL,na.rm=T)
x1 = max(times$IBM,times$INTC,times$MSFT,times$AAPL,na.rm=T)
y0 = min(stocks$IBM,stocks$INTC,stocks$MSFT,stocks$AAPL,na.rm=T)
y1 = max(stocks$IBM,stocks$INTC,stocks$MSFT,stocks$AAPL,na.rm=T)
plot(times[[1]],abs(stocks[[1]]),type="l",col=1,xlim=c(x0,x1),
     ylim=c(y0,y1),xlab="Time",ylab="USD",main="Stock Prices")
for (i in c(2:4)){
  lines(times[[i]],abs(stocks[[i]]),type="l",col=i)
}
legend(2000,y1,legend=names(stocks),lty=1,col=c(1:4))
```


Stock Prices
IBM
INTC
MSFT
AAPL
USD
0
100
200
300
400
500
600
700
1970
1980
1990
2000
2010
Time

Comments on stock prices

- The huge drop in the price of IBM is due to a 4-for-1 stock split on May 31, 1979.
- From the WSJ (Wall Street Journal, Jun 9, 2015):
  Apple has taken investors on quite a ride since it instituted a "very unusual" 7-for-1 stock split exactly one year ago today. On June 6, 2014, Apple shares closed at $645.57 apiece. The following Monday, they opened at $92.70 each after the stock split took effect.
- Intel and Microsoft have had stock splits.
- Stock merges (aka reverse stock split) are also possible although less common.
- So long time series of stock prices should be treated with care.

Returns and Log Returns

Returns and Log Returns I

Suppose $\{X_t\}$ is the daily/weekly/monthly (closing) stock-price of an asset.

- **Return** also known as net return is defined as:

$$R_t := \frac{X_t}{X_{t-1}} - 1 \quad .$$

- Depending on the time units being days, weeks or months, we talk about daily, weekly or monthly returns, respectively.
- **Log-returns** are defined as

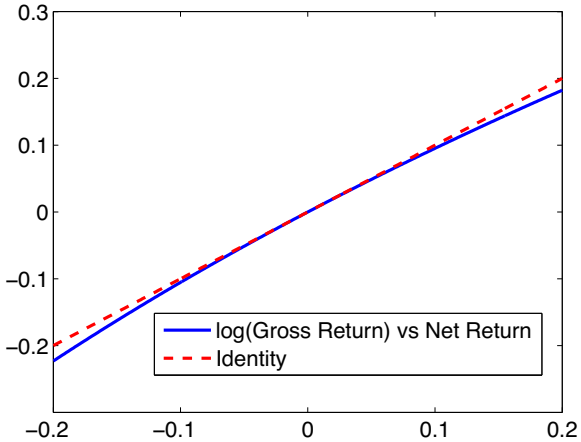
$$r_t := \lg \frac{X_t}{X_{t-1}} \quad .$$

Returns and Log Returns II

- By the Taylor expansion of the function $x \mapsto \log(1+x)$, we have $\log(1+x) = x + O(x^2)$, as $x \to 0$. Thus,

$$r_t = R_t + O(R_t^2) \quad .$$

Log Return vs Net Return



Formulas

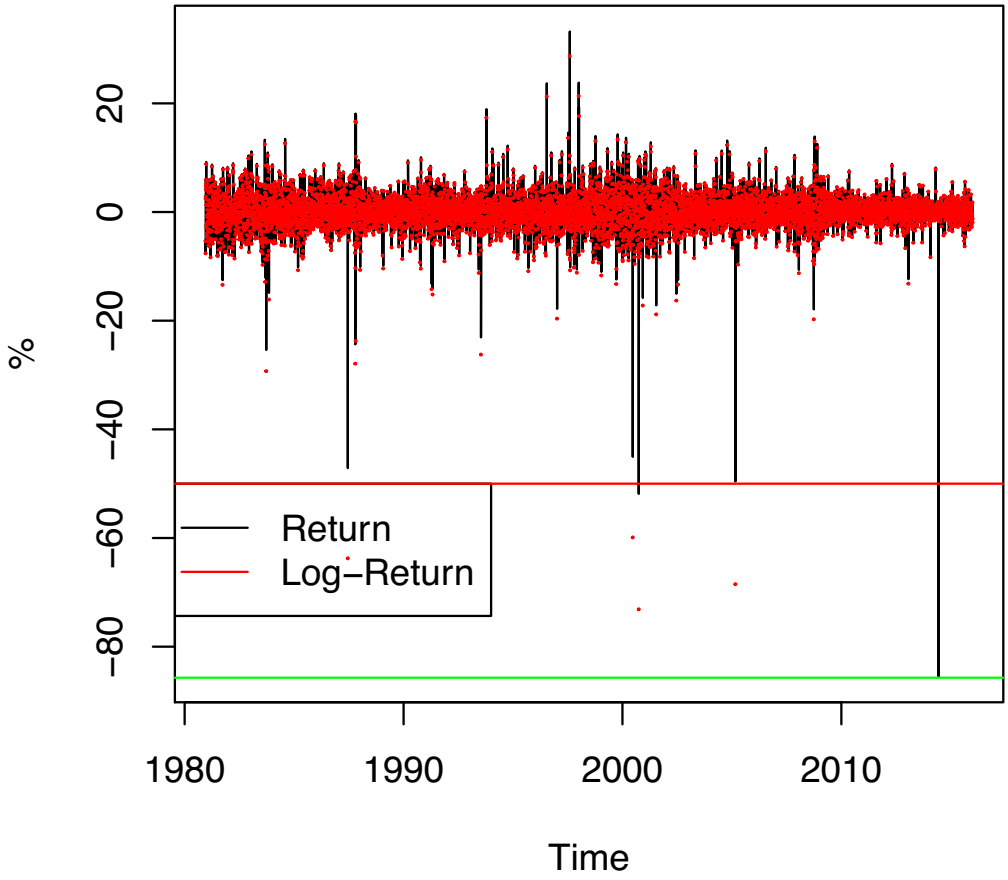
$$R_t = e^{r_t} - 1$$

Returns and Log Returns III

- **Units:** The returns have units – percent per unit time, e.g., 6% per year or 0.1% per day
  - which corresponds to

```
par(mfrow=c(1,1))
n=length(stocks$AAPL)
R.AAPL = stocks$AAPL[-1]/stocks$AAPL[-n] - 1
r.AAPL = diff(log(stocks$AAPL))
plot(times$AAPL[-1],100*R.AAPL,type="l",xlab="Time",ylab="%",col=1)
points(times$AAPL[-1],100*r.AAPL,col=2,cex=0.1)
legend(3050,-50,legend=c("Return","Log-Return"),lty=1,col=c(1:2))
abline(h=-50,col="red") # 2-for-1 split
abline(h=-100*(1-1/7),col="green") # 7-for-1 split
```

Returns and Log Returns IV



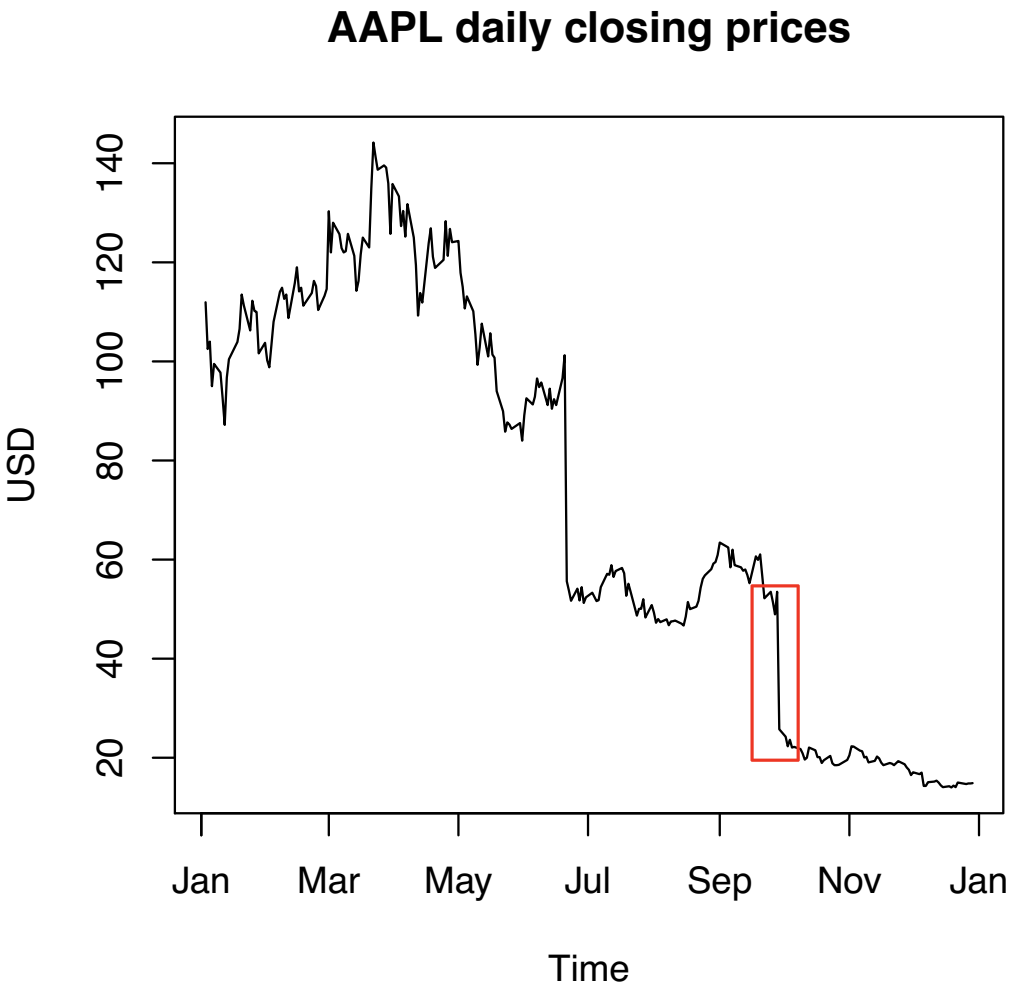
Looking more closely at 2000 I

Apple's split history:

Date	Split type
06/16/1987	2 for 1
06/21/2000	2 for 1
02/28/2005	2 for 1
06/09/2014	7 for 1

```
t = as.POSIXlt(times$AAPL)
plot(t[t$year=="100"],stocks$AAPL[t$year=="100"],xlab="Time",
     ylab="USD",main="AAPL daily closing prices",type="l")
```

Looking more closely at 2000 II



Apple in September 2000

- There was only one 2 for 1 split in 2000 on June 21st!
- There was a huge more than 50% drop in AAPL stock price on September 29th!
- For more details, see the CNN money article from September 29, 2000.

Caution: The returns could be quite heavy-tailed. And yes, you can lose $1/2$ of your portfolio on a single day if you are not diversified!

Multiperiod Returns

- k-**period return:**

$$R_t(k) := \frac{X_t}{X_{t-k}} - 1 =$$

$$= (R_{t+1})(R_{t-1+1}) \cdots - 1$$

- k-**period log return:**

$$r_t(k) := \log\left(\frac{X_t}{X_{t-k}}\right)$$

$$= \log\left(\frac{X_t}{X_{t-1}} \cdot \frac{X_{t-1}}{X_{t-2}} \cdots\right)$$

$$= r_{t+1} r_{t-1+1} \cdots r_{t-k+1}$$

- Log-returns are

Adjusting for dividends

Many stocks pay dividends several times a year. This should be reflected in the returns.

- Let D_t denote the dividend payed in period t (which could be 0).
- **Adjusted returns**

$$R_t := \frac{P_t + D_t}{P_{t-1}} - 1$$

- **Adjusted log returns**

$$r_t := \log \frac{P_t + D_t}{P_{t-1}}$$

Note:

Statistical Properties of the Returns

For this data turns out that

- The returns are essentially uncorrelated, i.e.,

$$\mathrm{Cov}(r_t, r_{t+h}) \approx 0$$

收益和亏损抵消了?

long range correlation can happen.

- The squares of the returns are significantly correlated!

$$\mathrm{Cov}(r_t^2, r_{t+h}^2)$$

波动率相关性较大

- The variability or volatility of the returns clusters, i.e., a period of high volatility persists for some time.
- The marginal distribution of the returns has significantly heavier tail than the Normal distribution.

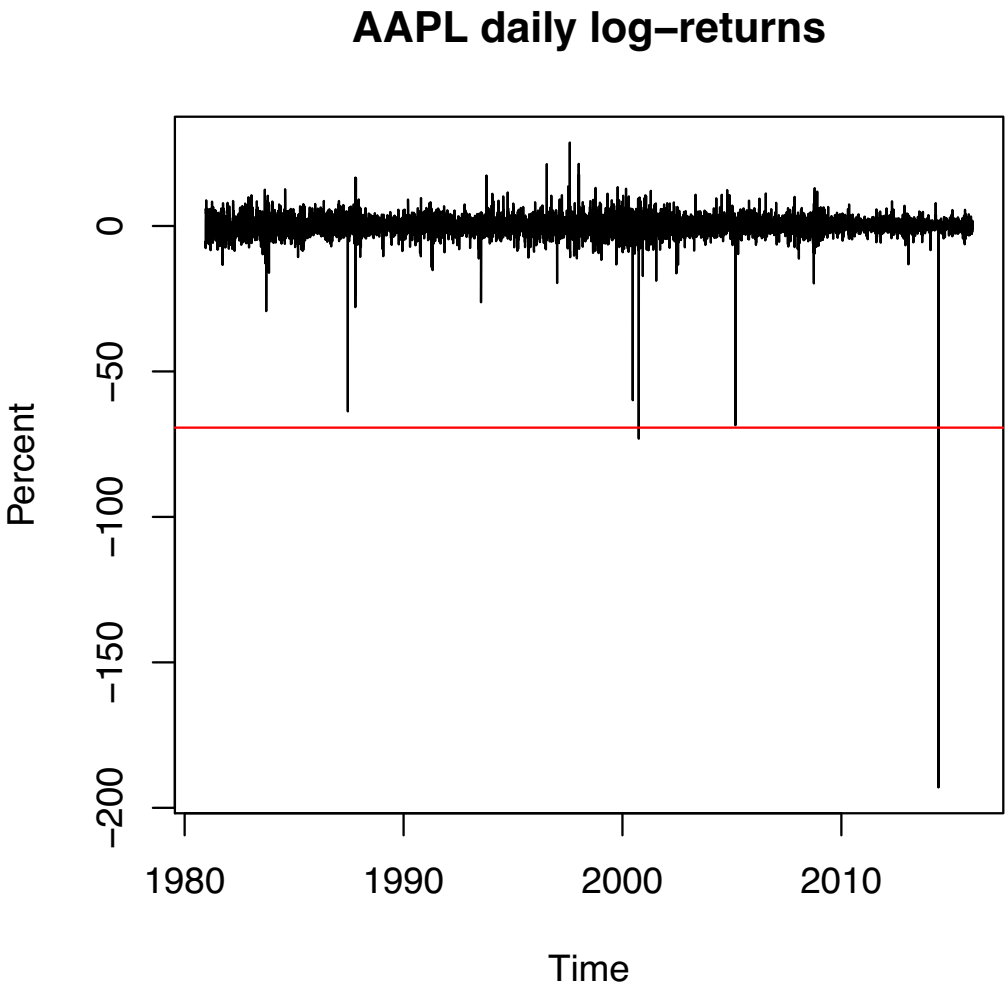
1. Above is important for applications in **Risk** and **Optimization**
2. There are limitations ....

Raw log-returns I

We shall illustrate these facts with some basic statistical analysis of the daily log returns of AAPL.

```
r.AAPL = diff(log(stocks$AAPL)) 做log差分
t.AAPL = times$AAPL[-1] # all but the first time
plot(t.AAPL,r.AAPL*100,xlab="Time",ylab="Percent",type="l",
     main="AAPL daily log-returns")
abline(h=log(0.5)*100,col="red")
```

Raw log-returns II



Note: Concepts covered in future lectures

- A time series $\{Y_t, t = 1, 2, \dots\}$ is stationary, if

$$(Y_t, t = 1, 2, \cdots) \stackrel{d}{=} (Y_{t+h}, t = 1, 2 \cdots)$$ 分布相同

  for all $h \in \mathbb{N}$.

- To a first approximation and in near-term the daily returns may be modeled as stationary time series.
- Their auto-correlation function (ACF) is 如果stationary则自相关系数和t无关

$$\gamma_r(k) = \text{Corr}(r_t, r_{t+k}) = \frac{\text{Cov}[r_t, r_{t+k}]}{\sigma^2} = \frac{\widehat{\text{Cov}}[r_t, r_{t+k}]}{\hat{\sigma}^2},$$

  where $\sigma^2 =$. $\text{VAR}[r_i] = \text{VAR}[r_n]$ 需要先移除极值

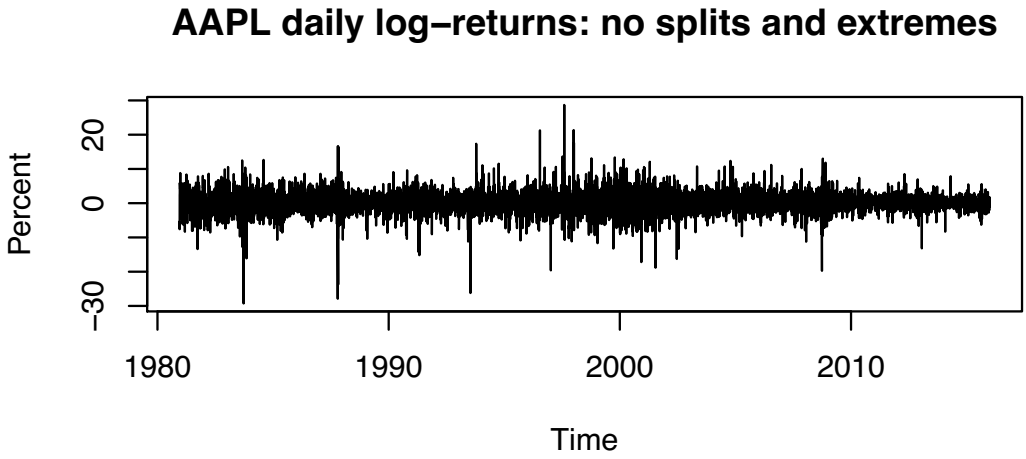
- Observe the `ACF` plots for the time-series $\{r_t\}$ and $\{r_t^2\}$ with the corresponding pointwise 95%-confidence bounds.

Correlation structure II

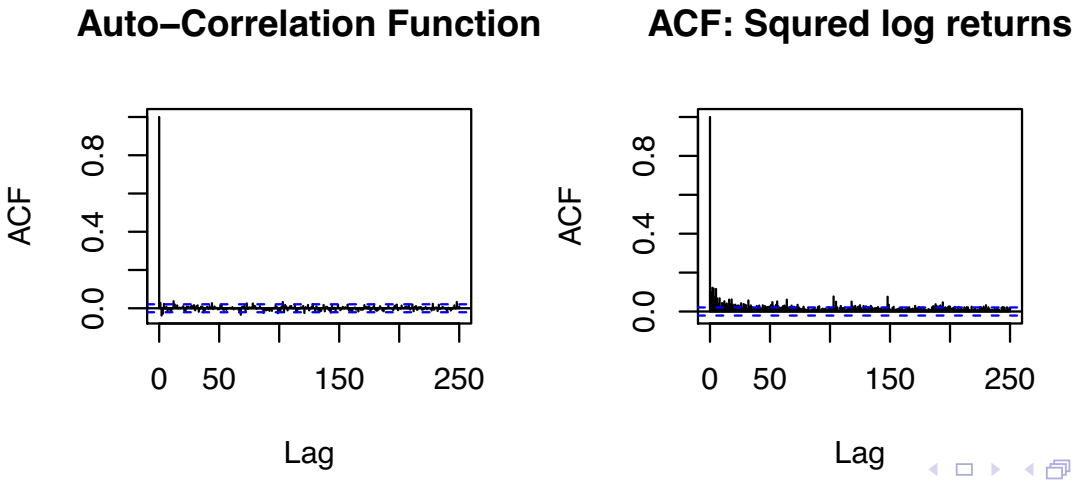
```
nf <- layout(matrix(c(1,1,2,3),2,2,byrow = TRUE), c(2,2), c(2,2), TRUE)
#layout.show(nf)
plot(t.AAPL[r.AAPL>-0.5],r.AAPL[r.AAPL>-0.5]*100,xlab="Time",ylab="Percent",typ
     main="AAPL daily log-returns: no splits and extremes")
acf(r.AAPL[r.AAPL>-0.5],#na.action = na.pass,
    main="Auto-Correlation Function",lag.max = 250)
acf(r.AAPL[r.AAPL>-0.5]^2,#na.action = na.pass,
    main="ACF: Squred log returns",lag.max = 250)
```

Correlation structure III

记得移除分红导致的极值



square return有明显的自相关性



Quantile-Quantile Plots I

也可以比较理想分布和真实分布的cdf,但不好找tail

- Let $Y_i,\ i = 1, \ldots, n$ be some data that want to model with a (cumulative) distribution F. The quantile-quantile plot is a good way of checking whether the model agrees with the data.
- The generalized quantile function of the distribution F is defined as:

$$F^{-1}(p) = \inf\{y\ :\ F(y) \geq p\}.$$

从概率得到分位值

- The QQ-plot is a plot of the quantiles $(F^{-1}(i/(n+1)), Y_{(i)}), i = 1, \ldots, n$, where

$$\frac{1}{n+1}\ Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}\ \frac{n}{n+1}$$ 分位值

are the order statistics of the data and the $\frac{i}{n+1}$ quantiles of the empirical data.

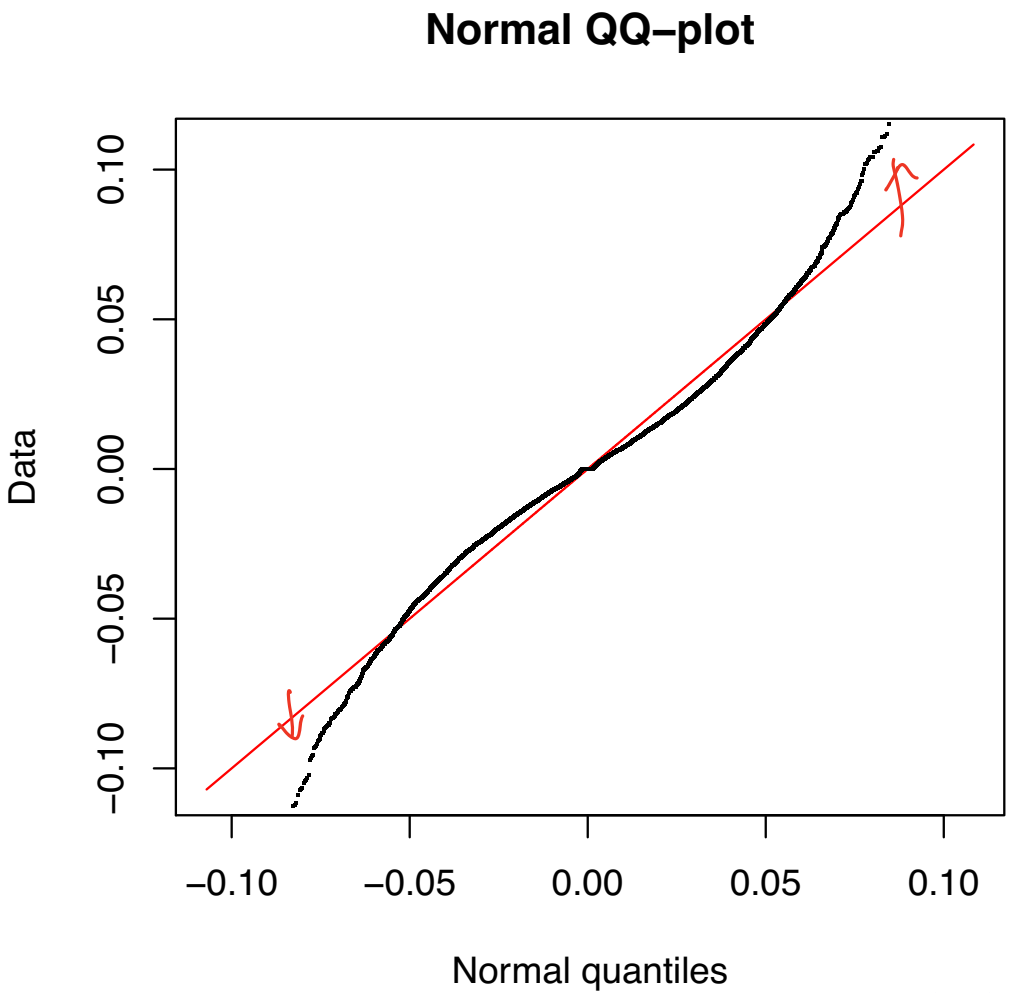
- If the Y_i's come from this model, then the plot will be approximately linear.

Quantile-Quantile Plots II

- If the right-side of the plot curves **up** away from the line, then the distribution of the data has heavier right tail than the model.
- Similarly, if the left-side curves **down**, then the left tail of the data is heavier.
- **Important:** If the x- and y-axes are flipped, the interpretation changes. Make sure you know which axis corresponds to the data and which to the model!

```
n = length(r.AAPL[r.AAPL>-0.5])
mu  = mean(r.AAPL[r.AAPL>-0.5])
sig = sd(r.AAPL[r.AAPL>-0.5])
Normal_quantiles = qnorm(c(1:n)/(n+1),mean=mu,sd=sig)
plot(Normal_quantiles,Normal_quantiles,col="red",type="l",
     xlab="Normal quantiles",
     ylab="Data",main="Normal QQ-plot");
points(Normal_quantiles,sort(r.AAPL[r.AAPL>-0.5]),pch=".",cex=0.8)
```

Quantile-Quantile Plots III



越直越heavy

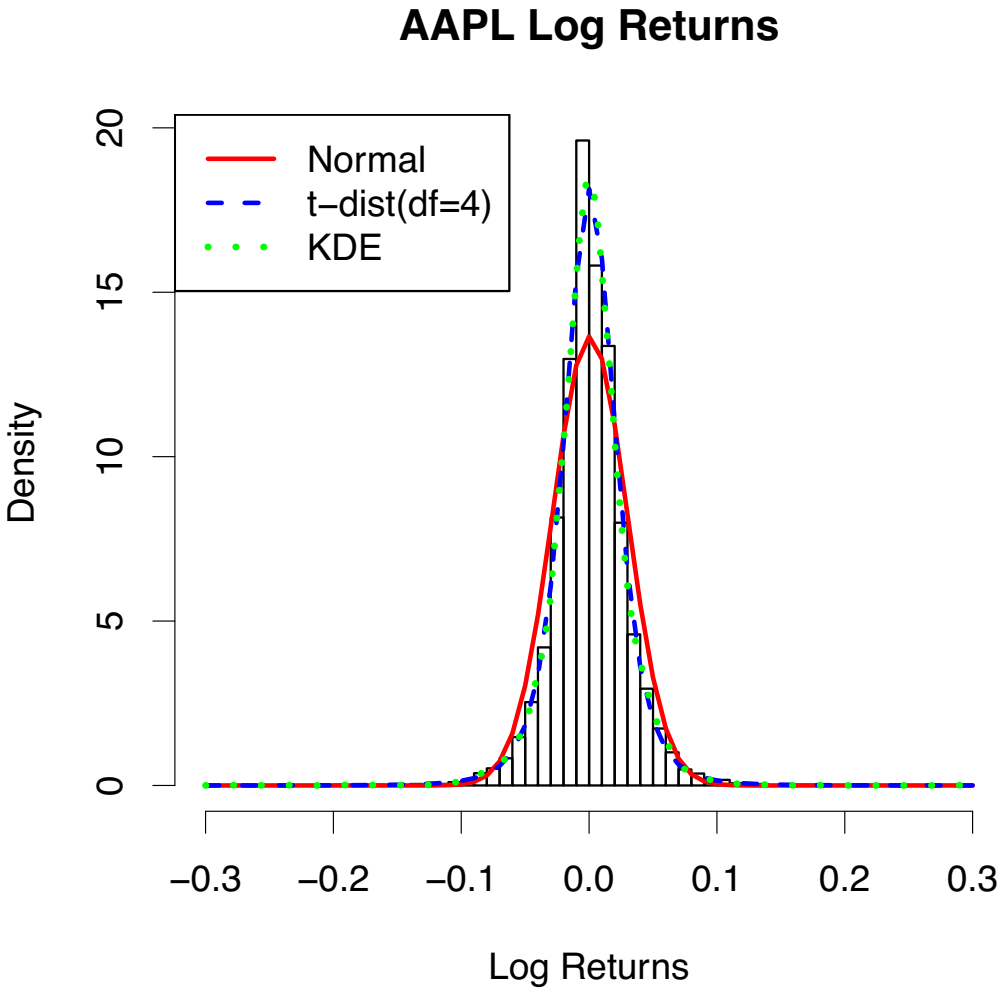
Histograms and KDEs I

- The histogram is a standard and quick visual display of the distribution.
- The kernel density estimators (KDE) offer an improvement.
- Both tend to mis-represent the tails. The transformed KDEs provide a good alternative.
- Here is a quick plot illustrating a Normal and t-distribution fit to the AAPL log returns

Histograms and KDEs II

```
#library("fGarch",quietly = T)
grid=seq(-0.3,0.3,0.01)
hist(r.AAPL[r.AAPL>-0.5],xlab="Log Returns",
    freq=FALSE,breaks=grid,main="AAPL Log Returns",lwd=0.5)
lines(grid,dnorm(grid,mean=mu,sd=sig),col="red",lty=1,lwd=2)
lines(grid,fGarch::dstd(grid,mean=mu,sd = sig,nu=4),
      col="blue",lty=2,lwd=2)
lines(density(r.AAPL,from=-0.3,to=0.3,n=128),
      col="green",lty=3,lwd=3)
legend("topleft",c("Normal","t-dist(df=4)","KDE"),
       col=c("red","blue","green"),lty=c(1,2,3),lwd=c(2,2,3))
```

Histograms and KDEs III

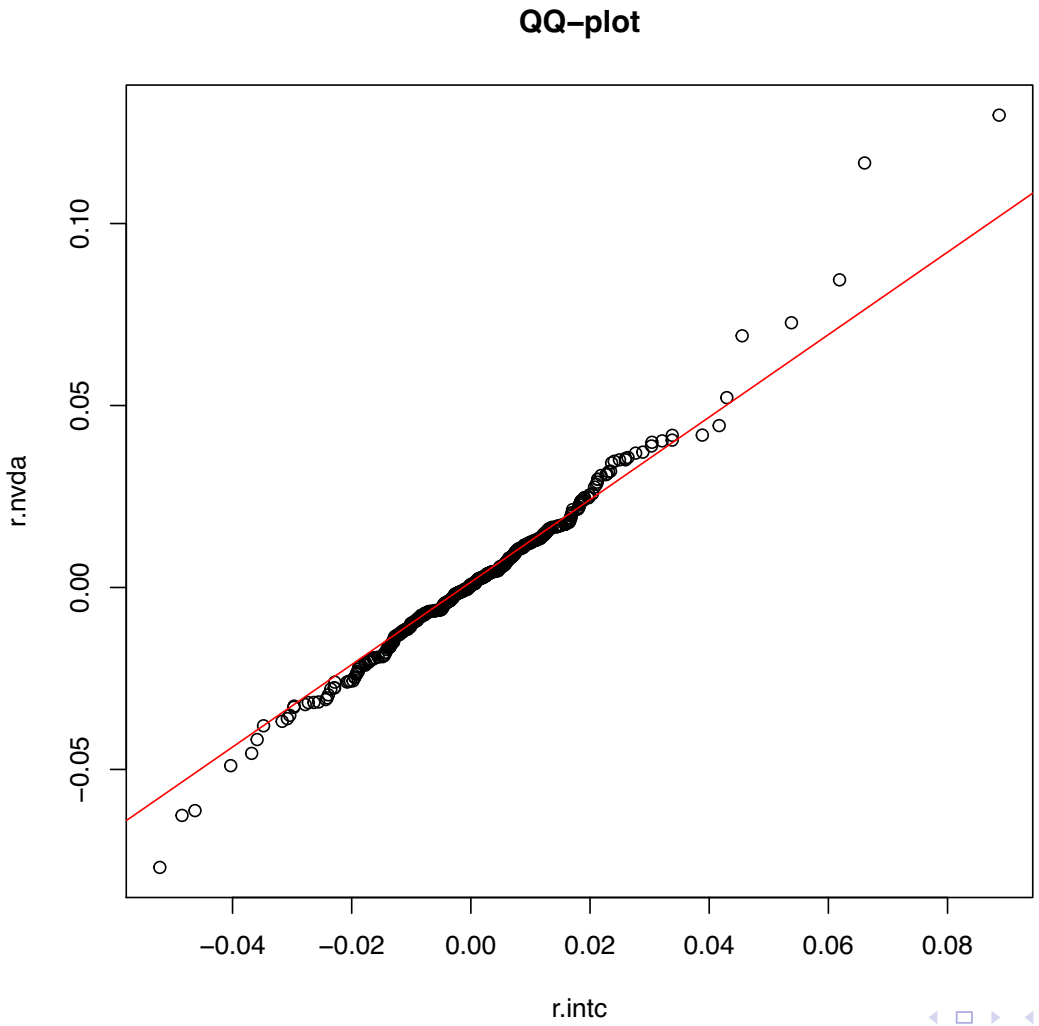


More QQ-plots I

We explore next the goodness of fit for Student t-distribution models through QQ-plots.

```
par(mfrow=c(2,2))
for (df in c(3,4,5,6)){
  n = length(r.AAPL[r.AAPL>-0.5])
  mu  = mean(r.AAPL[r.AAPL>-0.5])
  sig = sd(r.AAPL[r.AAPL>-0.5])
  t_quantiles = fGarch::qstd(c(1:n)/(n+1),mean=mu,sd=sig,nu=df)
  plot(t_quantiles,t_quantiles,type="l",col="red",
       ylab="Data",xlab="t-dist quantiles",
       main=paste("QQ-plot: t-dist(df=",df,")",sep=""))
  points(t_quantiles,sort(r.AAPL[r.AAPL>-0.5]),pch=".",cex=0.8)
}
```

More QQ-plots II



Stock indices

Stock Indices I

Many single-number summaries of the financial markets exist, known as indices. They are designed to quantify the economy, various sectors (real estate, energy, manufacturing, technology, etc.) or types of stocks (e.g. large-cap, small-cap, value, etc.). Some indices just keep track of the volatility (e.g. VIX).
Here are some of the more popular indices.

- **DJIA (Dow Jones Industrial Average):** This is a simple price average of 30 selected stocks, multiplied by a divisor.
- The divisor is updated when there are stock-splits so as to maintain the continuity of the index.
- One drawback of DJIA and other price average based indices is that the price of the stock does not reflect the overall capitalization of the company. DJIA不能体现市值

Stock Indices II

- **INDEXSP (S&P 500 – Standard & Poor's)** is a capitalization-weighted average of the 500 select stock prices. That is,

$$\texttt{INDEXSP} = \frac{1}{\mathrm{d}} \sum_{i=1}^{505} N_i \times P_i,$$

市值权重指数

  where P_i is the price and N_i the number of outstanding shares of stock i. The divisor d is proprietary and adjusted to maintain continuity. See, e.g., the Wikipedia article.

- It is updated every 15 seconds.
- Offers a good representation of the health of the American stockmarket and economy.
- **IXIC: NASDAQ Composite** is a index composed primarily of information technology companies. It is also capitalization-weighted avarage with some restrictions on the influence of individual components.

Stock Indices III

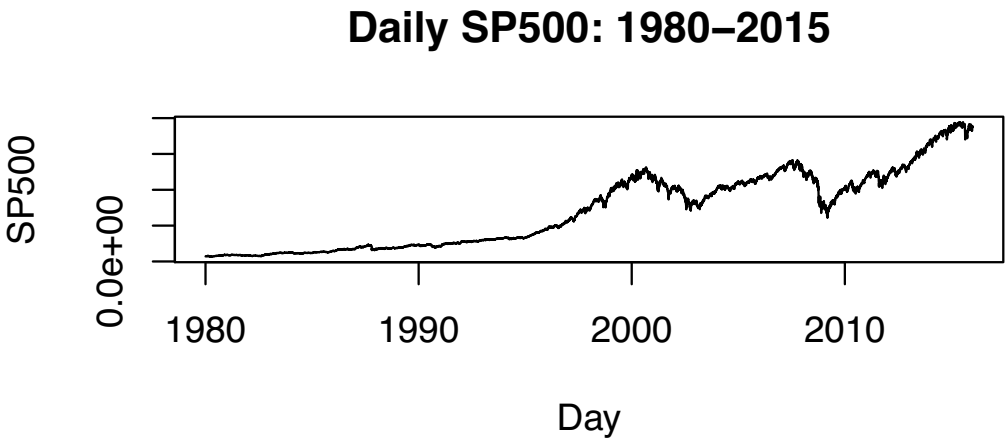
- **Other polular indices:** Nikkey 225 (Japan), FTSE 100 (UK), DAX (Germany), etc.

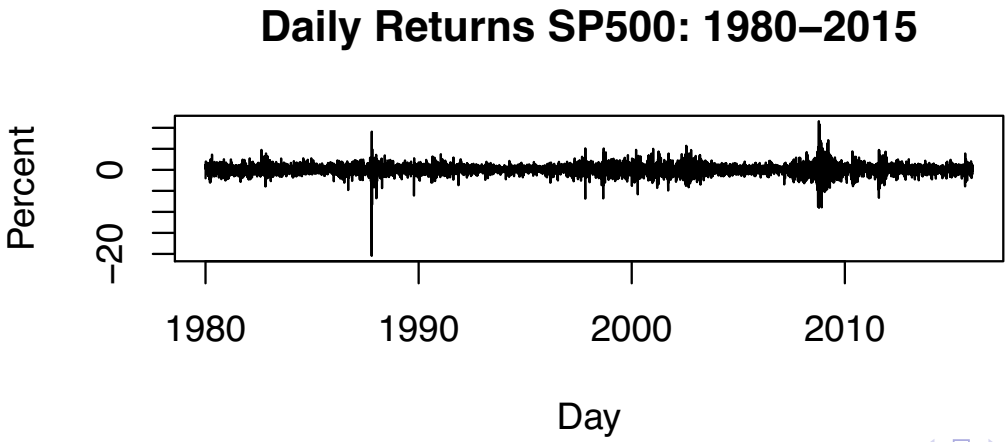
```
sp = read.csv("~/Dropbox/doc/courses/STATS_509_Fall_2016/Data/SP500_dai
names(sp)

 [1] "caldt"  "vwretd" "vwretx" "ewretd" "ewretx" "totval" "totcnt"
 [8] "usdval" "usdcnt" "spindx" "sprtrn"

t.sp = as.Date(x=as.character(sp$caldt),format="%Y%m%d")
# Getting the times
par(mfrow=c(2,1))
plot(t.sp,sp$usdval,type="l",xlab="Day",ylab="SP500"
        ,main="Daily SP500: 1980-2015")
plot(t.sp,sp$sprtrn*100,type="l",xlab="Day",ylab="Percent"
        ,main="Daily Returns SP500: 1980-2015")
```

Stock Indices IV





A list of a few US Stock Market crises

The large drops in the S&P 500 returns and also periods of high volatility may be associted with financial crises in the US economy.

基本假设:未来和以前走势相同

- **Black Monday** on October 19, 1987.
- This event was followed by the recession of the early 1990s.
- **Dot-com bubble** is the period of 1997–2000, which collapsed in 1999–2001. In this period many speculative Interent startups failed. Others suffered but survived. For example, the Amazon.com stock went from from $107 to $7 during the crash. It trades at $1,189 per share today (January 2, 2018).
- From the Wikipedia article on the early 2000s recession:

  From 2000 to 2001, the Federal Reserve, in a move to protect the economy from the overvalued stock market, made successive interest rate increases; while this may have initiated the readjustment, it is starkly contrasted with the severe, prolonged recession that would have occurred had the unsustainable growth continued unabated.

- **Sub-prime mortgage meltdown:** Contributed to a severe US recession in the period December 2007 – June 2009.