MTH 224O, Spring 2024

Instructor: Bahman Angoshtari

Lecture 23

Sections 6.5: regression

0. Overview of the lecture

- We then learn about **regression**. It is one of the most common statistical models that is used to identify relationships between various quantities.

In [1]:
```
# Setting parameters of the Jupyter notebook
# This cell is only usefull if your are using Jupyter
sc = 0.75
options(repr.plot.width=16*sc,
        repr.plot.height=6*sc,
        repr.plot.pointsize = 20, # Text height in pt
        repr.plot.bg        = 'white',
        repr.plot.antialias = 'gray',
        #nice medium-res DPI
        repr.plot.res       = 300,
        #jpeg quality bumped from default
        repr.plot.quality   = 90,
        #vector font family
        repr.plot.family    = 'serif', # Vector font family. 'sans', 'serif'
        "getSymbols.warning4.0"=FALSE)
```

1. Regression

- Regression is one of the most widely used statistical models.
- Suppose we have a sample (X_1, Y_1), (X_2, Y_2), ... , (X_N, Y_N) of two random variables.
- We would like to explain (or predict) the Y-values based on the X-values. In this case, we say Y is the **response** variable and X is the **explanatory** or **predictor** variable.

- **Line regression model:**

  $$Y_n = \beta_0 + \beta_1 X_n + \varepsilon_n; \quad \text{for } n = 1, 2, \ldots, N,$$

  where β_0 and β_1 are unknown parameters (or coefficients) and ε_1, ... , ε_N are random variables representing error.

- The following assumptions are usually made for a line regression model
  1. $\mathbb{E}(\varepsilon_i) = 0$
  2. $\varepsilon_1, \ldots, \varepsilon_N$ are independent
  3. $\mathrm{Var}(\varepsilon_n) = \sigma_\varepsilon^2$, for all n
  4. ε_n are normally distributed

- The conditions above are summarized by the notation $\varepsilon_n \overset{i.i.d.}{\sim} N(0, \sigma_\varepsilon^2)$
- Here, i.i.d. stands for independent and identically distributed
- We may also say that $\{\varepsilon_n\}_{n=1}^N$ is a normal (or Gaussian) white noise

2. Least-squares (LS) estimation of the coefficients

- Our first goal is to "fit" the line regression model (to the sample $\{(X_n, Y_n)\}_{n=1}^N$). That is, to find "good estimates" of the unknown coefficients β_0 and β_1

- There are three main approaches to fit statistical models:

  a. Least-squares (LS) estimation

  b. Maximum likelihood estimation (MLE)

  c. Bayesian estimation

- We use the simplest method, namely LS. In this method, the estimate of the parameter are the minimizers of the the sum of the squared errors

  $$f(b_0, b_1) = \sum_{n=1}^N \Big[Y_n - (b_0 + b_1 X_n)\Big]^2$$

- We denote the estimate of parameters β_0 and β_1 by $\widehat{\beta}_0$ and $\widehat{\beta}_1$, respectively. Then, the estimates are given by

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \underset{b_0, b_1}{\arg\min}\ f(b_0, b_1)$$

- The function $f(b_0, b_1)$ is shown to be a convex function. Therefore, its global minimizer $(\widehat{\beta}_0, \widehat{\beta}_1)$ is the uniques solution of the system of equation

$$\begin{cases} \dfrac{\partial f}{\partial b_0}(b_0, b_1) = \dfrac{\partial f}{\partial b_0}\left(\sum_{n=1}^{N}\left[Y_n - (b_0 + b_1 X_n)\right]^2\right) = 0 \\ \dfrac{\partial f}{\partial b_1}(b_0, b_1) = \dfrac{\partial f}{\partial b_1}\left(\sum_{n=1}^{N}\left[Y_n - (b_0 + b_1 X_n)\right]^2\right) = 0 \end{cases}$$

$$\Longrightarrow \quad \begin{cases} \sum_{n=1}^{N} Y_n - N\, b_0 - b_1 \sum_{n=1}^{N} X_n = 0 \\ \sum_{n=1}^{N} X_n Y_n - b_0 \sum_{n=1}^{N} X_n - b_1 \sum_{n=1}^{N} X_n^2 = 0 \end{cases}$$

- Solving this system of linear equations yields

$$\widehat{\beta}_1 = \frac{\sum_{n=1}^{N}(Y_n - \bar{Y})(X_n - \bar{X})}{\sum_{n=1}^{N}(X_n - \bar{X})^2}, \quad \text{and} \quad \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}$$

where we have defined $\bar{X} = \frac{\sum_{n=1}^{N} X_n}{N}$ and $\bar{Y} = \frac{\sum_{n=1}^{N} Y_n}{N}$

- The formulas for $\widehat{\beta}_1$ is usually expressed in the following form

$$\widehat{\beta}_1 = \frac{S_{XY}}{S_X^2}$$

- Here, the sample variance S_X^2 and sample covariance S_{XY} are

$$S_X^2 = \frac{\sum_{n=1}^{N}(X_n - \bar{X})^2}{N-1}, \text{ and } S_{XY} = \frac{\sum_{n=1}^{N}(Y_n - \bar{Y})(X_n - \bar{X})}{N-1}$$

- The least square line is given by

$$\hat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X = \bar{Y} + \widehat{\beta}_1 (X - \bar{X}) = \bar{Y} + \frac{S_{XY}}{S_X^2}(X - \bar{X})$$

- $\hat{Y}_n = \widehat{\beta}_0 + \widehat{\beta}_1 X_n$ is called the "fitted value" of Y (at X_n)

- $\hat{\varepsilon}_n = Y_n - \hat{Y}_n$, $n = 1, \ldots, N$, are called the "residuals". They are estimates of the error $\varepsilon_n = Y_n - \beta_0 - \beta_1 X_n$.

- Note the difference between residuals and errors!

Summary of line regression:

- You are given a sample $\{(X_n, Y_n)\}_{n=1}^N$.

- Calculate $\bar{X}$, $\bar{Y}$, S_X^2, S_{XY}:

$$\bar{X} = \frac{\sum_{n=1}^N X_n}{N}, \quad \bar{Y} = \frac{\sum_{n=1}^N Y_n}{N}$$
$$S_X^2 = \frac{\sum_{n=1}^N (X_n - \bar{X})^2}{N-1}, \quad S_{XY} = \frac{\sum_{n=1}^N (Y_n - \bar{Y})(X_n - \bar{X})}{N-1}$$

- The least square line is given by $\hat{Y} = \bar{Y} + \frac{S_{XY}}{S_X^2}(X - \bar{X})$

Example 1:

- Let us see how to fit a line regression in R.

- We want to see if there is a linear relationship between government interest rate and corporate interest rate.

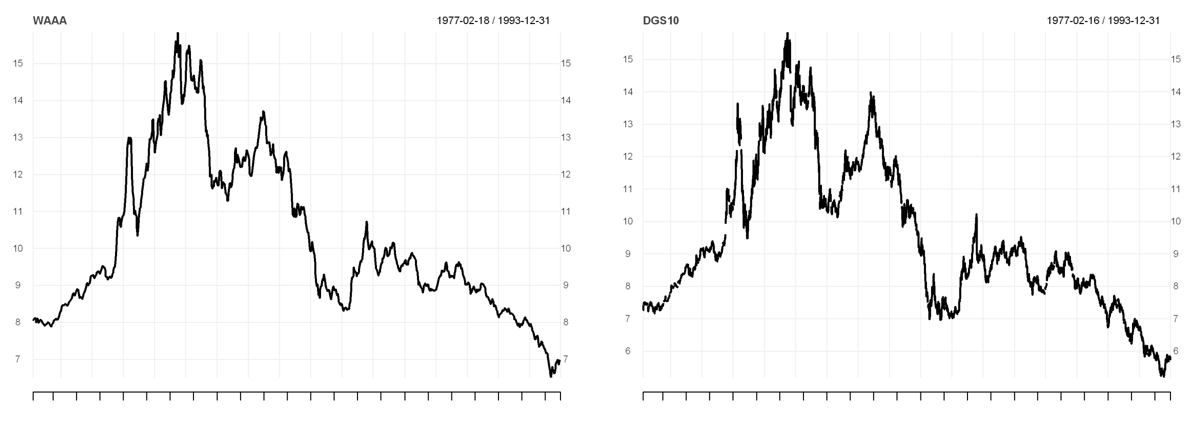
- Let us obtain historical 10-year Treasury constant maturity rate (symbol `DGS10`) and Moody's seasoned corporate AAA yields (symbol `AAA`), both of which are available from FRED.

In [2]:
```
library("quantmod")
getSymbols(c("WAAA","DGS10"), src="FRED")
WAAA = WAAA["1977-02-16/1993-12-31"]
DGS10 = DGS10["1977-02-16/1993-12-31"]
```

```
Loading required package: xts

Loading required package: zoo


Attaching package: 'zoo'


The following objects are masked from 'package:base':

    as.Date, as.Date.numeric


Loading required package: TTR

Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
```

'WAAA' · 'DGS10'

In [3]:
```
par(mfrow=c(1,2))
par(mar=c(3,3,3,3))
Mytheme = chart_theme()
Mytheme$col$line.col = "black"
chart_Series(WAAA, theme = Mytheme)
chart_Series(DGS10, theme = Mytheme)
```



- Note that the series have different frequencies. `WAAA` is weekly, while `DGS10` is daily.

In [4]:
```
dim(WAAA)
dim(DGS10)
```

881 · 1

4403 · 1

- We need to match the two first. The following code merge the two time series to have weekly frequency.

In [5]:
```
dat = merge(WAAA,DGS10)
dat = na.locf(dat) # filling NA's with "last observation carried forward" ru
dat = dat[index(WAAA)]
dim(dat)
head(dat)
```

881 · 2

```
           WAAA DGS10
1977-02-18 8.04  7.41
1977-02-25 8.08  7.48
1977-03-04 8.10  7.48
1977-03-11 8.12  7.44
1977-03-18 8.09  7.44
1977-03-25 8.00  7.48
```

- With the sample $\{(\mathrm{WAAA}_n, \mathrm{DGS10}_n)\}_{n=1}^{881}$ at hand, let us fit the line regression

  $$\Delta\mathrm{WAAA}_n = \beta_0 + \beta_1 \Delta\mathrm{DGS10}_n + \varepsilon_n$$

- Here, $\Delta\mathrm{WAAA}_n = \mathrm{WAAA}_n - \mathrm{WAAA}_{n-1}$ is the weekly corporate rate change. Similarly, $\Delta\mathrm{DGS10}_n = \mathrm{DGS10}_n - \mathrm{DGS10}_{n-1}$ is the weekly treasury rate change.

In [6]:
```
# Calculating the differences
AAA_dif = diff(as.vector(dat[,"WAAA"]))
DGS10_dif = diff(as.vector(dat[,"DGS10"]))
```

- Next, we fit the line regression model, using the R function `lm()`, and output the summary.

- The fitted line is

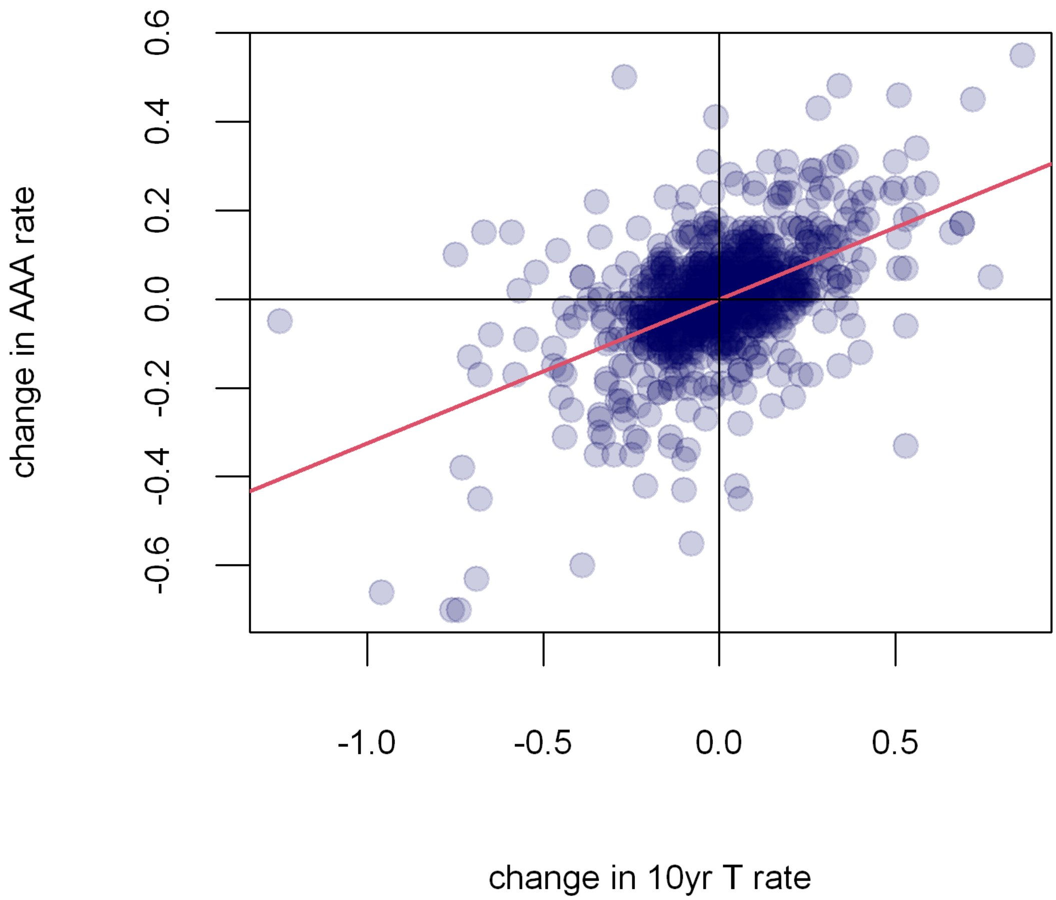
$$\Delta\mathrm{WAAA}_n = -0.000669 + 0.3232680\,\Delta\mathrm{DGS10}_n + \varepsilon_n$$

In [7]:
```
# Fitting a linear model
fit1 = lm(AAA_dif ~ DGS10_dif)
summary(fit1)
```

```
Call:
lm(formula = AAA_dif ~ DGS10_dif)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.52346 -0.04873  0.00139  0.05104  0.58799 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.0006694  0.0037824  -0.177     0.86    
DGS10_dif    0.3233945  0.0182126  17.757   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1122 on 878 degrees of freedom
Multiple R-squared:  0.2642,	Adjusted R-squared:  0.2634 
F-statistic: 315.3 on 1 and 878 DF,  p-value: < 2.2e-16
```

- We can then produce a simple scatter plot of the sample along with the regression line.

In [8]:
```
options(repr.plot.width=6, repr.plot.height=6)
plot(DGS10_dif,AAA_dif,xlab="change in 10yr T rate",
     ylab="change in AAA rate",pch=19,col=rgb(0,0,100,50,maxColorValue=255))
abline(fit1, col=2, lwd=2)
abline(h=0,v=0)
```



3. Course Conclusion

- MTH224 is an introductory course. We covered many fundamental topics in probability, and some topics in statistics.
  - Probability: sample spaces, events, definition of probability functions, conditional probability and the corresponding rules, discrete random variables, discrete joint distribution, expectation and conditional expectations, common discrete distributions (binomial, Poisson, ...), continuous distributions and their density functions, exponential and normal distributions, central limit theorem.
  - Statistics: common sample statistics (sample mean, variance, quartiles), common data visualization techniques (histogram, boxplots, scatter plots), point estimation and MLE, line regression.

- You now have a foothold for learning more topics in probability and statistics. Here are some suggestions:
    - MTH524 (Intro. to Prob.) and MTH525 (Intro. to Math. Stat.): You will gain in-depth knowledge on theory of probability and statistics. Plan to take multivariate calculus before taking these.
    - MTH542 (Statistical Analysis): An application oriented class on statistical methods. You can take this after MTH224.
    - MTH533 and MTH534 (Real Analysis): Real analysis covers measure and integration theory, which are essential for a proper understanding of probability theory. For those of you who want to know more, and willing to pay the cost!