Time Series Analysis Of Stock Market

Regression Analysis For Prediction Of Real Estate And Oil Price

Presented by :
Bhavyaraj Singh (N064)

Table of Contents

01.
Time Series Analysis Of Stock Market

AMAZON

Time Series Analysis

- A time series is a set of observations based on values taken by a variable at regular time intervals.

- Time series forecasting is used to predict future values based on previously observed values and one of the best tools for trend analysis and future prediction.It is recorded at regular time intervals, and the order of these data points is important. Therefore, any predictive model based on time series data will have time as an independent variable. The output of a model would be the predicted value or classification at a specific time.

- **ARIMA** stands for Auto Regressive Integrated Moving Average.

- Before working with non-stationary data, ARIMA converts it to stationary data. One of the most widely used models for predicting linear time series data is this one.The ARIMA model has been widely utilized in banking and economics since it is recognized to be reliable, efficient, and capable of predicting short-term share market movements.

Stock Price Prediction

- A Stock or share (also known as a company's "equity") is a financial instrument that represents ownership in a company
- In investing, technical analysis relies heavily on Time Series data for statistical inferences and predictions.
- We will be using time series analysis to fetch the data for AMAZON (AMZN)
- The base for technical analysis is the assumption that while forecasting stock data is that it has some relation with its historical values.
- The Random Walk Theory contends this belief. It states that changes in a share's successive price does not depend upon its previous value. But the theory also states, all price changes conform to a particular probability distribution, which can be used to forecast stock prices.

AMAZON STOCK

STOCK MARKET AMAZON.R

Source on Save　Run　Source

```r
install.packages('quantmod')
install.packages("tseries")
install.packages('forecast', dependencies = TRUE)

library(quantmod)
getSymbols("AMZN")
summary(AMZN)
chartSeries(AMZN,subset='last 6 months',type=1)
addBBands()
library(tseries,quietly = T)

adf.test(na.omit(AMZN$AMZN.Adjusted))

ret_AMZN<- 100*diff(log(AMZN$AMZN.Adjust[2274:2457]))
summary(ret_AMZN)

library(forecast,quietly = T)

AMZN_ret_train <- ret_AMZN[1:(0.9*length(ret_AMZN))]

AMZN_ret_test <- ret_AMZN[(0.9*length(ret_AMZN)+1):length(ret_AMZN)]

fit <- Arima(AMZN_ret_train, order = c(2,0,2))

preds <- predict(fit,n.ahead=(length(ret_AMZN)-(0.9*length(ret_AMZN))))$pred
preds

test_forecast <- forecast(fit,h=25)
test_forecast

plot(test_forecast,main="Arima forecast for AMAZON")
```

AMAZON :OHLC

	AMZN.Open	AMZN.High	AMZN.Low	AMZN.Close	AMZN.Volume	AMZN.Adjusted
2021-08-09	3343.61	3354.88	3328.52	3341.87	2148200	3341.87
2021-08-10	3345.01	3358.00	3315.00	3320.68	2412600	3320.68
2021-08-11	3331.45	3337.70	3277.79	3292.11	2947200	3292.11
2021-08-12	3290.00	3314.51	3269.67	3303.50	2314100	3303.50
2021-08-13	3305.67	3306.07	3283.00	3293.97	2052800	3293.97
2021-08-16	3283.00	3300.00	3211.13	3298.99	3319700	3298.99
2021-08-17	3277.50	3280.49	3225.68	3241.96	3387900	3241.96
2021-08-18	3241.99	3254.10	3200.00	3201.22	2804300	3201.22
2021-08-19	3194.02	3233.00	3182.46	3187.75	3782900	3187.75
2021-08-20	3203.87	3207.81	3175.76	3199.95	3341200	3199.95
2021-08-23	3211.90	3280.90	3210.01	3265.87	3268100	3265.87
2021-08-24	3280.00	3315.49	3274.58	3305.78	2551800	3305.78
2021-08-25	3309.87	3321.00	3286.15	3299.18	1680300	3299.18
2021-08-26	3299.00	3332.00	3296.00	3316.00	2098800	3316.00
2021-08-27	3333.23	3352.32	3313.75	3349.63	2391300	3349.63
2021-08-30	3357.43	3445.00	3355.22	3421.57	3192200	3421.57
2021-08-31	3424.80	3472.58	3395.59	3470.79	4356400	3470.79
2021-09-01	3496.40	3527.00	3475.24	3479.00	3629900	3479.00
2021-09-02	3494.76	3511.96	3455.00	3463.12	2923700	3463.12
2021-09-03	3452.00	3482.67	3436.44	3478.05	2575700	3478.05

Summary (AMZN)

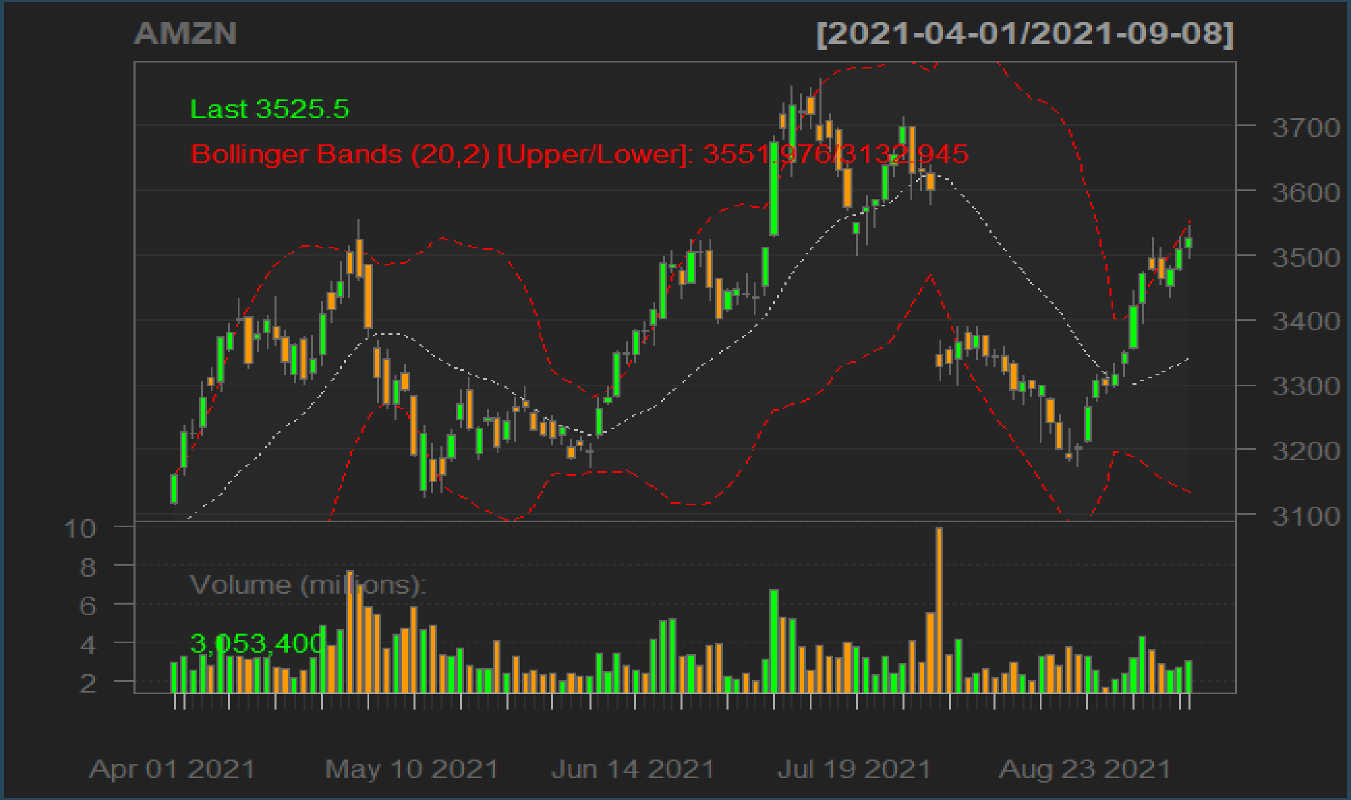
Console Terminal Jobs

R 4.1.0 · ~/

```
> summary(AMZN)
     Index              AMZN.Open          AMZN.High          AMZN.Low      
 Min.   :2007-01-03   Min.   :  35.29   Min.   :  37.07   Min.   :  34.68  
 1st Qu.:2010-09-02   1st Qu.: 144.07   1st Qu.: 145.91   1st Qu.: 142.10  
 Median :2014-05-07   Median : 325.44   Median : 329.30   Median : 321.52  
 Mean   :2014-05-06   Mean   : 813.05   Mean   : 821.62   Mean   : 803.50  
 3rd Qu.:2018-01-07   3rd Qu.:1217.51   3rd Qu.:1229.14   3rd Qu.:1210.00  
 Max.   :2021-09-09   Max.   :3744.00   Max.   :3773.08   Max.   :3696.79  
                      NA's   :1         NA's   :1         NA's   :1        
   AMZN.Close        AMZN.Volume          AMZN.Adjusted    
 Min.   :  35.03   Min.   :   881300   Min.   :  35.03  
 1st Qu.: 144.20   1st Qu.:  3051800   1st Qu.: 144.20  
 Median : 325.00   Median :  4385100   Median : 325.00  
 Mean   : 812.88   Mean   :  5506520   Mean   : 812.88  
 3rd Qu.:1229.14   3rd Qu.:  6587700   3rd Qu.:1229.14  
 Max.   :3731.41   Max.   :104329200   Max.   :3731.41  
 NA's   :1         NA's   :1           NA's   :1        
> 
```

chartSeries (name, subset , type)

Bollinger Bands



Augmented Dicky-fuller Test

Console Terminal Jobs

R 4.1.0 · ~/

```
> adf.test(na.omit(AMZN$AMZN.Adjusted))

	Augmented Dickey-Fuller Test

data:  na.omit(AMZN$AMZN.Adjusted)
Dickey-Fuller = -0.19847, Lag order = 15, p-value = 0.99
alternative hypothesis: stationary
```

ARIMA

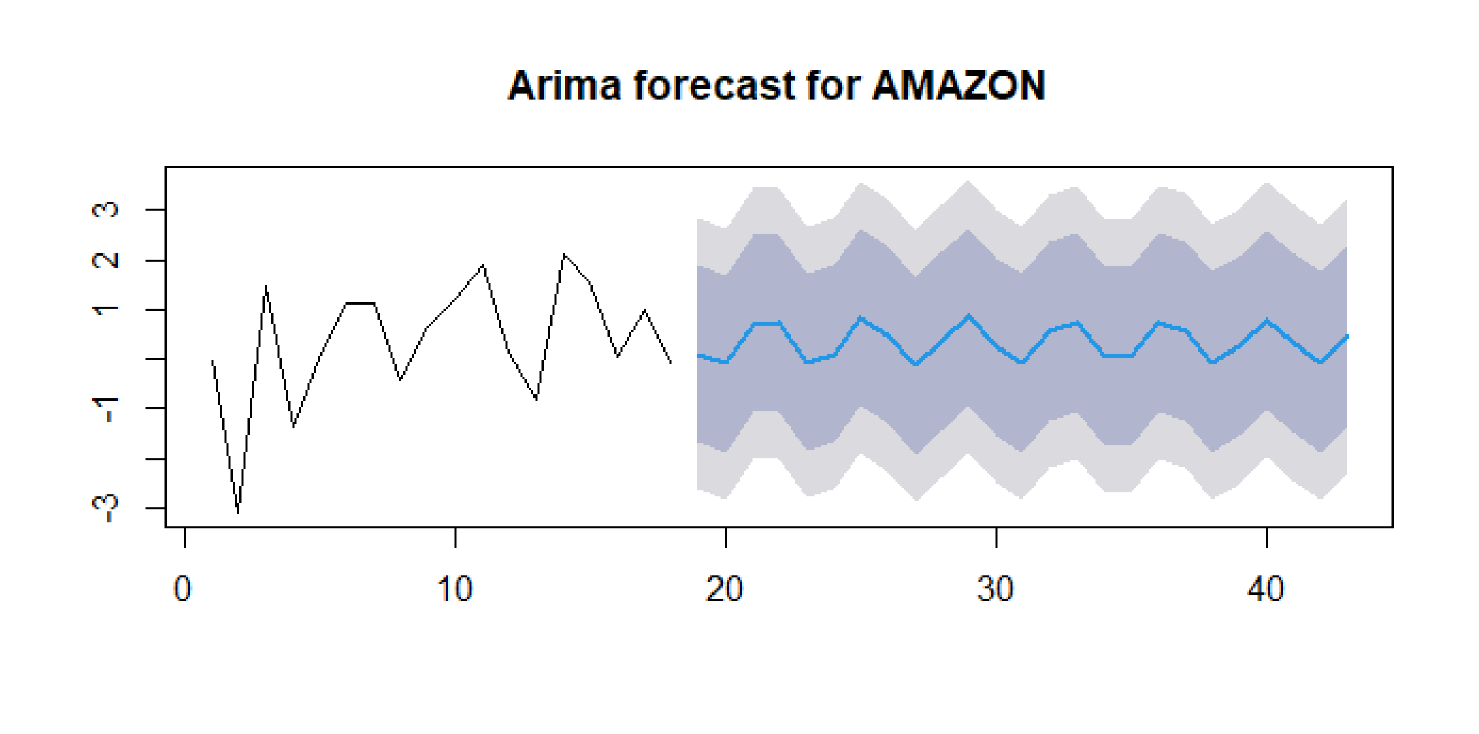
```r
AMZN_ret_train <- ret_AMZN[1:(0.9*length(ret_AMZN))]

AMZN_ret_test <- ret_AMZN[(0.9*length(ret_AMZN)+1):length(ret_AMZN)]

fit <- Arima(AMZN_ret_train, order = c(2,0,2))
```

```
> preds <- predict(fit,n.ahead=(length(ret_AMZN)-(0.9*length(ret_AMZN))))$pred
> preds
Time Series:
Start = 19
End = 36
Frequency = 1
 [1]  0.11248074 -0.09904106  0.71110815  0.72255378 -0.07189296  0.10839391
 [7]  0.84128606  0.48847115 -0.14268761  0.35420487  0.85122396  0.24586637
[13] -0.09392824  0.57952976  0.74926953  0.05024830  0.05284936  0.73531498
> 
```

FORECAST

```
> test_forecast <- forecast(fit,h=25)
> test_forecast
   Point Forecast      Lo 80    Hi 80     Lo 95    Hi 95
19     0.11248074 -1.6583849 1.883346 -2.595825 2.820786
20    -0.09904106 -1.8743016 1.676220 -2.814068 2.615986
21     0.71110815 -1.0635981 2.485814 -2.003071 3.425287
22     0.72255378 -1.0581141 2.503222 -2.000743 3.445850
23    -0.07189296 -1.8506769 1.706891 -2.792308 2.648522
24     0.10839391 -1.6769455 1.893733 -2.622047 2.838835
25     0.84128606 -0.9418906 2.624463 -1.885847 3.568419
26     0.48847115 -1.3008138 2.277756 -2.248004 3.224946
27    -0.14268761 -1.9305228 1.645148 -2.876946 2.591570
28     0.35420487 -1.4384242 2.146834 -2.387385 3.095794
29     0.85122396 -0.9413856 2.643833 -1.890336 3.592784
30     0.24586637 -1.5497089 2.041442 -2.500229 2.991962
31    -0.09392824 -1.8912217 1.703365 -2.842651 2.654795
32     0.57952976 -1.2188257 2.377885 -2.170818 3.329877
33     0.74926953 -1.0524097 2.550949 -2.006161 3.504700
34     0.05024830 -1.7509293 1.851426 -2.704415 2.804912
35     0.05284936 -1.7527565 1.858455 -2.708587 2.814285
36     0.73531498 -1.0688677 2.539498 -2.023944 3.494574
37     0.56829037 -1.2407028 2.377284 -2.198326 3.334907
38    -0.05836876 -1.8657900 1.749052 -2.822581 2.705843
39     0.25588260 -1.5559711 2.067736 -2.515108 3.026874
40     0.79252340 -1.0183279 2.603375 -1.976935 3.561982
41     0.35609573 -1.4581876 2.170379 -2.418611 3.130803
42    -0.06313197 -1.8774907 1.751227 -2.837954 2.711690
43     0.46438992 -1.3520440 2.280824 -2.313606 3.242386
```

Series



Result

- Above is the result that we obtain from a simple ARIMA(2,0,2) model.

- The deeply shaded region provides us the 80% confidence interval and the lightly shaded region provides the 95% confidence interval.

- The basic interpretation of a 95% confidence interval of the model tells us that the forecasted values will have a maximum deviation of +/- 2 as shown in the plot above, thus giving us a fair estimation of the values of the future stock indices.

02.

Regression Analysis For Prediction Of Real Estate

Real Estate Investment And Economy

- Real estate is one of the pillar industry of the national economy to enhance the development of investment, consumption and employment.

- In this paper we aim to analyze the relationship between real estate investment and economic growth by using the principle of liner regression, with the help of related data about the investment of the real estate and GDP of the city of Nanjing from 1998 to 2007.

- The statistics of the Nanjing city reflected that the outcome have largely increased from 28 billion Yuan in 1980 to 5 trillion Yuan in 2007, which increased value accounted for almost 5.5% of GDP.

- In this table the related data of the total output value of Nanjing and the total output value of the real estate in Nanjing is given.

Output Value

- In this table the related data of the total output value of Nanjing and the total output value of the real estate in Nanjing is given.

year	GDP of Nanjing[billion Yuan]	growth rate [%]	the investment of the real estate[billion Yuan]	the proportion [%]	growth rate[%]
1998	825.13	11.8	19.58	2.4	18.2
1999	899.42	10.6	28.23	3.1	15.0
2000	1021.30	12.3	41.8	4.1	28.5
2001	1150.30	11.1	47.80	4.2	14.3
2002	1297.57	12.8	56.92	4.4	13.7
2003	1576.33	15.0	65.36	4.1	11.8
2004	2067.18	17.3	78.22	3.8	25.8
2005	2411.11	15.1	88.30	3.7	10.9
2006	2773.78	15.1	123.85	4.5	28.7
2007	3283.78	15.7	156.00	4.8	19.8

Input Value

```r
Year=c(1998,1999,2000,2001,2002,2003,2004,2005,2006,2007)
GDP_of_Nanjing=c(825.13,899.42,1021.3,1150.3,1297.57,1576.33,2067.18,2411.11,
                 2773.78,3283.78)
data.frame(Year,GDP_of_Nanjing)
plot(Year,GDP_of_Nanjing)
m=lm(Year~GDP_of_Nanjing)
summary(m)
cor(Year,GDP_of_Nanjing)
barplot(GDP_of_Nanjing)
barplot(GDP_of_Nanjing,names.arg = Year,xlab ="Year",ylab="GDP of Nanjing",
        col="dark blue", border = "black")
```

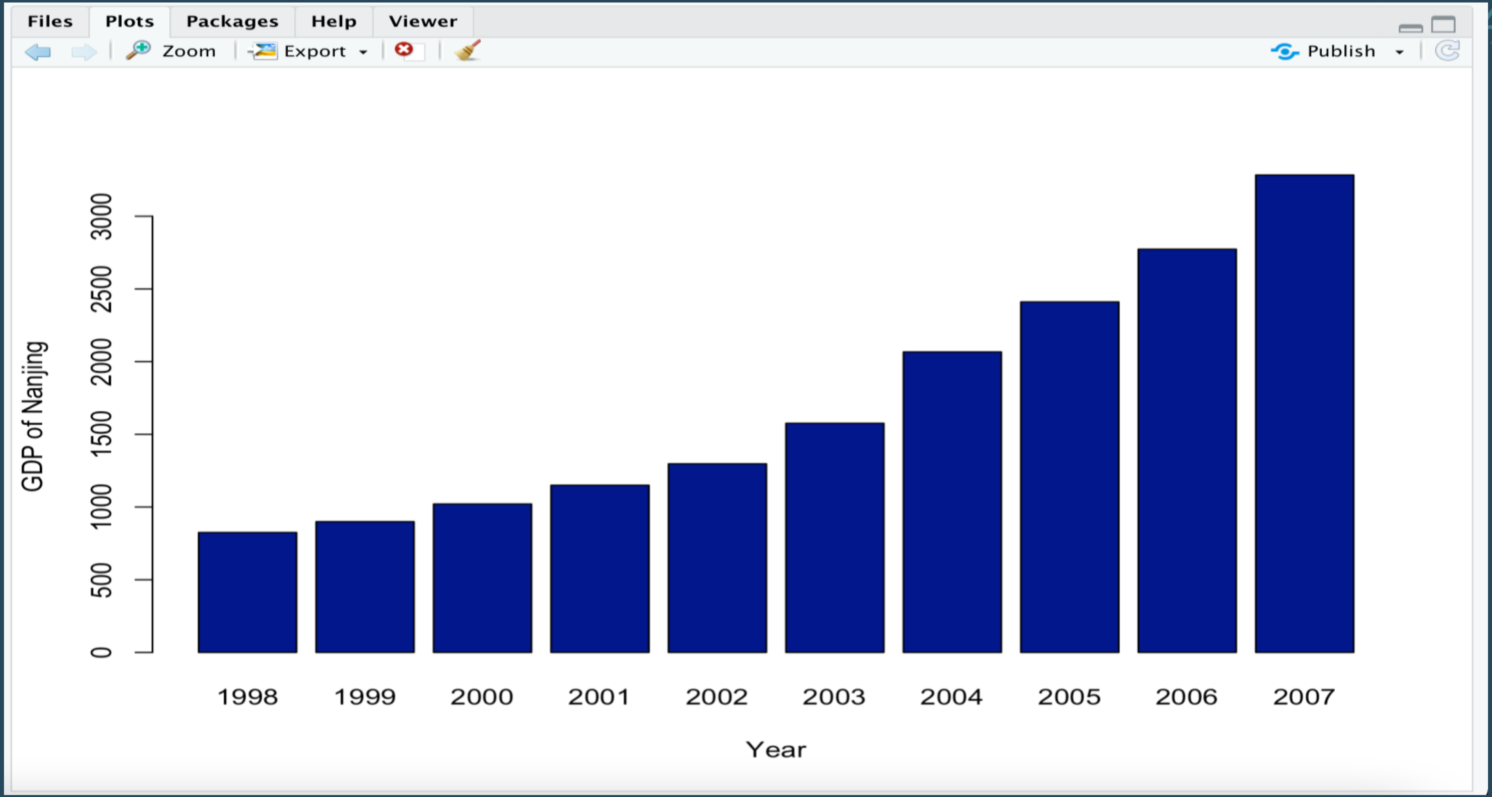
Output Value

```
Source

R 4.1.0 · ~/
> Year=c(1998,1999,2000,2001,2002,2003,2004,2005,2006,2007)
> GDP_of_Nanjing=c(825.13,899.42,1021.3,1150.3,1297.57,1576.33,2067.18,2411.11,
+                  2773.78,3283.78)
> data.frame(Year,GDP_of_Nanjing)
   Year GDP_of_Nanjing
1  1998         825.13
2  1999         899.42
3  2000        1021.30
4  2001        1150.30
5  2002        1297.57
6  2003        1576.33
7  2004        2067.18
8  2005        2411.11
9  2006        2773.78
10 2007        3283.78
> plot(Year,GDP_of_Nanjing)
> m=lm(Year~GDP_of_Nanjing)
> summary(m)

Call:
lm(formula = Year ~ GDP_of_Nanjing)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.41581 -0.52265  0.06434  0.44582  1.02544 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1.997e+03  6.141e-01 3251.34  < 2e-16 ***
GDP_of_Nanjing 3.406e-03  3.210e-04   10.61 5.45e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8272 on 8 degrees of freedom
Multiple R-squared:  0.9336,	Adjusted R-squared:  0.9254
```

Output Value

```
R 4.1.0 · ~/
4   2001          1150.30
5   2002          1297.57
6   2003          1576.33
7   2004          2067.18
8   2005          2411.11
9   2006          2773.78
10  2007          3283.78
> plot(Year,GDP_of_Nanjing)
> m=lm(Year~GDP_of_Nanjing)
> summary(m)

Call:
lm(formula = Year ~ GDP_of_Nanjing)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.41581 -0.52265  0.06434  0.44582  1.02544 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1.997e+03  6.141e-01 3251.34  < 2e-16 ***
GDP_of_Nanjing 3.406e-03  3.210e-04   10.61 5.45e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8272 on 8 degrees of freedom
Multiple R-squared:  0.9336,	Adjusted R-squared:  0.9254 
F-statistic: 112.6 on 1 and 8 DF,  p-value: 5.447e-06

> cor(Year,GDP_of_Nanjing)
[1] 0.966255
> barplot(GDP_of_Nanjing)
> barplot(GDP_of_Nanjing,names.arg = Year,xlab ="Year",ylab="GDP of Nanjing",
+         col="dark blue", border = "black")
> 
```

Bar graph


Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
GDP of Nanjing
0
500
1000
1500
2000
2500
3000
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
Year

Scatter Diagram

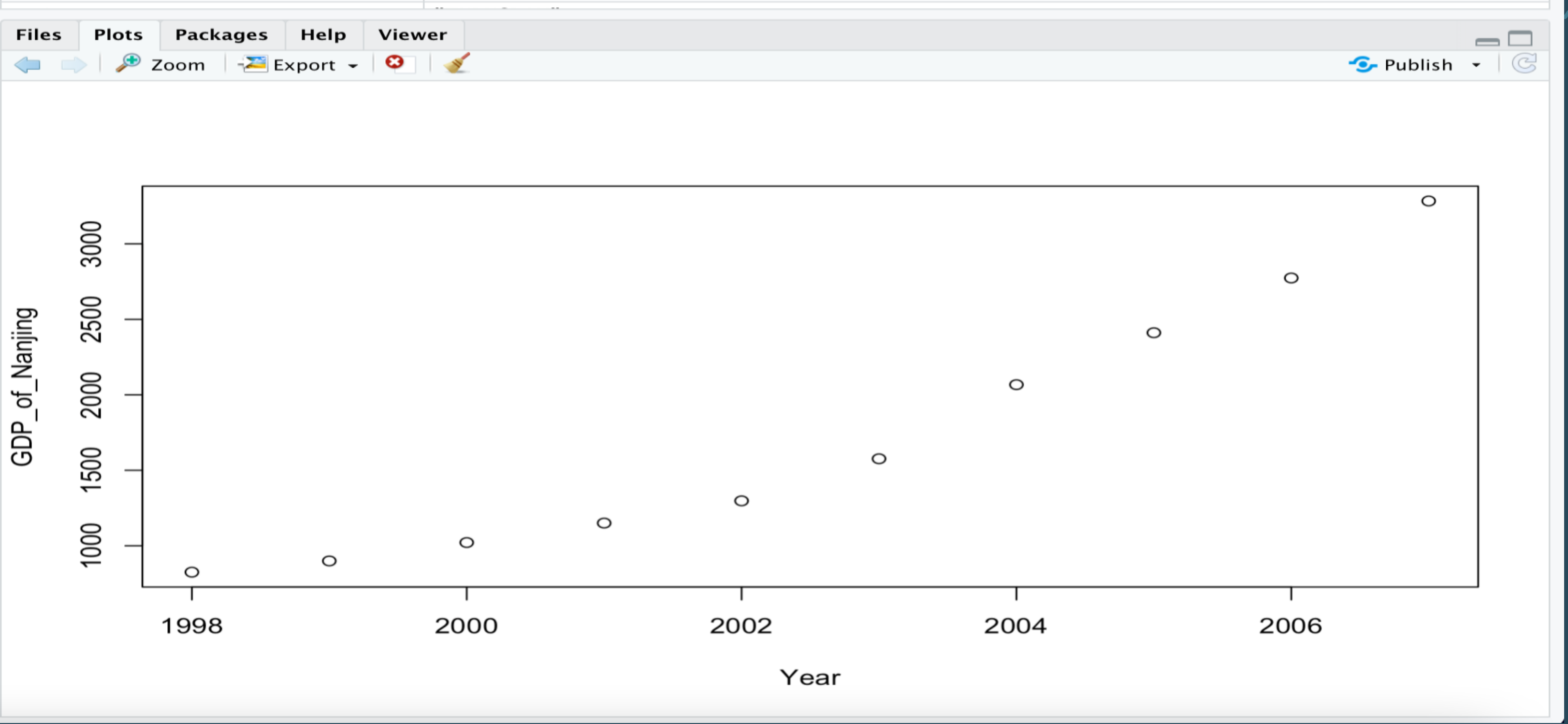


Table Analysis

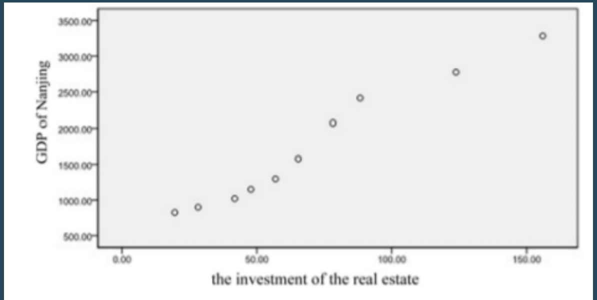
- Since 1998, nanjing’s real estate production value proportion of GDP.Increased year by year, even though it showed certain fluctuations some years, the overall trend was increasing.

- Secondly, gdp growth rate in nanjing increased year by year, although some fluctuations showed up in certain years. The overall trend was also enhancing. More specifically, it reflected that the speed of gdp’s increase in nanjing becoming increasingly faster.

- From the analysis above, we could reach the conclusion that the more proportion of real estate production which accounted for in nanjing’s gdp, the more of the increase rate in nanjing’s gdp

Scatter Diagram

- Using the spss software draws the scatter diagram of the investment of the real estate and the total output value of Nanjing
- The scatter diagram shows the linear relationship between the total output value of whole city and the investment of the real estate
- shows, there is a positive correlation between the investment of the real estate and total output value of nanjing.
- In this scatter diagram the distribution of all the points are similar to a straight line. This is consistent with the precondition of monadic linear



Scatter Diagram

- Let the total output value of the city be dependent variable Y and the value of the real estate investment be independent variable X, the model be:

$$Y=\beta 0+\beta 1X+\varepsilon. \quad (1)$$

where, β0 is regression constant, β1 is regression parameter, ε is random error.

Parameter Estimation

- There are usually two ways to do the parameter estimation for the model of simple linear regression.

- Two methods: least squares and method of Maximum Likelihood .

- SPSS software is used to calculate the parameters in this model

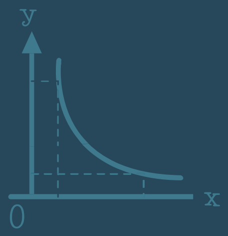
Table 2: Coefficienta

model		non-standardized coefficient		standard coefficient	t	Sig.
		B	standard error	trial version		
1	(constant)	338.364	109.524		3.089	.015
	the value of the real estate investment	19.718	1.345	.982	14.661	.000

a. independent variable: the city's GDP

Parameter Estimation

- The constants term of the regression model is 338.364
- The independent variable the regression coefficient of the real estate investment is 19.718.
- So the regression equation is: Y=19.718X+338.364.



Economic Significance

- The regression constant is positive. It means the total output value of the city is still increased without the investment of the real estate.
- The regression parameter is 19.718 it means the total output value will be increased 19.718 Yuan when the real estate investment was increased 1 Yuan.
- It is reasonable by economic viewpoint.
- Thus, the model has been checked by economic significance.

Statistical Significance

- The data given in the table show, R=0.982. It means a strong correlation between the value of the real estate investment and the city total output.
- In addition, the goodness of fit R2=0.964 ~ 1. It means the changing of city's total output value caused by the changing of the real estate investment value accounts for 96.4%of the changing of city's total output value.
- It shows a high fitting degree in variables.

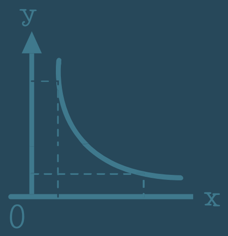


Table 3: Model summary

model	R	R^2	adjust R^2	standard error of estimate
1	.982[a]	.964	.960	172.56369

a. predictive variable:(constant), the value of the real estate investment

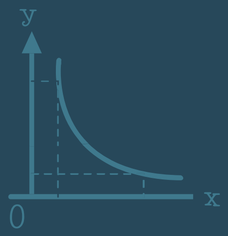
Statistical Significance

- In the following Anova table, under the loss 5%, the observed value of F test statistic is 214.944
- The significance probability of F Distribution is 0.000. It is equivalent to the probability of the establishment of test hypothesis- H0: regression coefficient B=0 is 0.00, so we reject the null hypothesis. It means the linear relationship is significance between dependent variable and independent variable.
- Therefore the goodness of fit of the regression equation is very good. In addition, from table , the regression coefficient is 14.661 after t test; the significance level of regression coefficient a= 0.00, therefore we reject null hypothesis of t test too. It also shows the same result.

model		square sum	df	mean square	F	Sig.
1	regression	6400651.666	1	6400651.666	214.944	.000[a]
	residual error	238225.822	8	29778.228		
	total	6638877.487	9			

Conclusion

- We reject the null hypothesis of f test. It means the linear the linear relationship is significance between dependent variable and independent variable.
- We reject the null hypothesis of t test also.
- From the analysis of the investment of the real estate we could see that real estate investment have a great effect on the advancement of economy, which reflected the mutual effect on real estate and national economy.



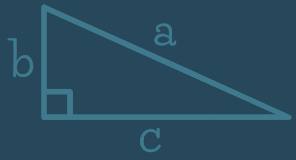
03.
Regression Analysis For Prediction Of Oil Price

Crude Oil

- With the continuous development of social economy, various countries in the world have a growing demand for crude oil.
- Crude oil has become the most important basic energy, and its fluctuated price brings great risks to the world economic development.
- It is beneficial to have accurate oil market forecasts in many sectors of economy.
- For example, both central banks and private sectors are using these predictions in many cases to generate macroeconomics plans and also, to measure risks.
- Sectors such as most of transport manufacturers, utility companies etc. are directly dependent on these forecasts.
- Even homeowners relay many of their decisions based on the oil prices.

Historical Price Of Crude Oil

- Modeling the price of oil is difficult, because of many fluctuations over time.
- The oil price can dramatically change over short time, which makes it very difficult to predict.
- In the past decade oil prices have fluctuated a lot, from $40 in 2003 to $140 in 2008.
- Oil demand and supply are quite inelastic in short term, that makes the price skyrocketed when the demand for oil exceeds supply.



Modeling

- The prediction is made using structural models.
- The price is predicted using linear regression models and will predict with mean square error or mean absolute error at the end .
- It uses multivariable linear regression model, to find the best fit.
- It assumes that spot price of oil is linearly correlated with each of the historical value of oil and the explanatory variables.
- For this linear regression, different time spans can be regressed.
- Hence depending on available data, different time spans for linear regression is assumed.

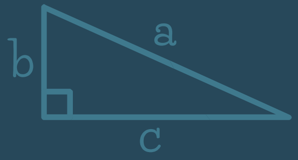
The Math

- In order to write regression formalism, we assume the price of oil at each given time period n is given by yn.

- It is presented by n×1 matrix Y.

- At each given time xpn refers to value of explanatory variable p at period n. And presented by p × n matrix X. Now at a given time n, the price of oil can be linearly regressed using the the explanatory variables as :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \in$$



The Math

- To understand in simpler form , we can write this as **y=Xβ +ε.**
- The formalism listed regresses the price of oil to the existing known events.
- In order to find a particular prediction the time span of n and n! Is shifted by that amount in the regression equation.
- Once we find the solution matrix β, regressing price of today and future can be easily calculated.
- The fitting can be optimized either using mean square error or mean absolute error (ε) methods.
- The expected value of solution of equation of matrix is

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

The Math

- To model the price, this project uses seven different variables.
- They are intended to correlate the oil price to the economy.

	Utilization	Production	Stock	Import	Consumption	SP500	Price	DGS10
Utilization	1							
Production	0.28	1						
Stock	0.182	0.508	1					
Import	-0.096	-0.636	-0.380	1				
Consumption	0.253	-0.065	-0.093	0.328	1			
SP500	0.375	0.529	0.274	-0.396	0.128	1		
Price	0.102	-0.056	-0.324	-0.003	-0.201	-0.098	1	
DGS10	0.104	-0.562	-0.437	0.513	0.282	-0.283	0.059	1

- Using these variables, price is predicted by shifting the oil price matrix (Y) for the amount of time in future that we would like to predict, and the matrix X gets lagged from matrix Y.
- Then solving equation for present values of X results in Y matrix that predict the oil price in future.

Conclusion

- Therefore ,these variable are used to regress various length of data to predict upto 12 month in advance.
- Based on this oil price listed above, forecast for 9 months were performed.
- It indeed predicted drop of the oil price in the first half of the year and increase of the price towards the end of the year.
- Ultimately, crude oil is a strategic material, and non-renewable resource.
- This analysis shows that most importantly, factors like weather, utilization, national policies, consumption and production have the most effect and will continue to affect the future price of crude oil.

THANK YOU!