Linear Regression Models

GR5205/GU4205

Geraldine Hu (yh3499)
Kaimi Huang (kh2908)
Yang Han (yh3503)

Outline

- Life Expectancy Dataset
  - Introduction of the data and the research question
  - Simple Linear Regression using OLS Method
  - Multiple Linear Regression:
    - OLS vs. WLS
    - Diagnostic Tests
    - Back test

- Bitcoin Dataset
  - Background, the data, and the research question
  - Simple Linear Regression using OLS Method
  - Multiple Linear Regression:
    - OLS vs. GLS
    - Diagnostic Tests
    - Back test

- Summary

Life Expectancy

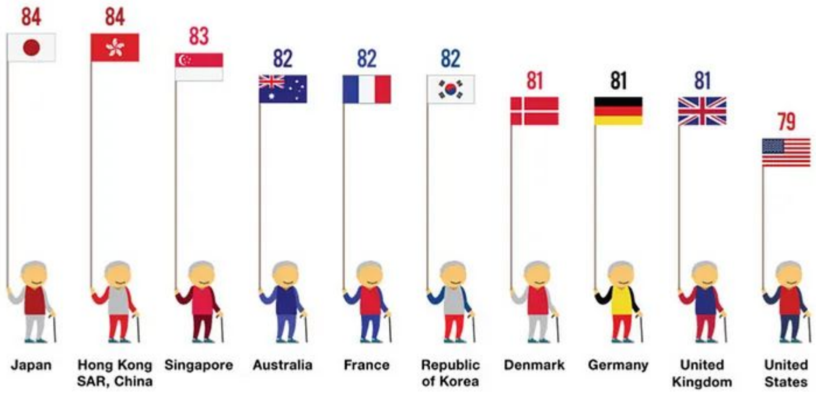
About our data:

The dataset contains life expectancy, health, immunization, economic and demographic information, a total of **21 variables**, about **179 countries** in the year of **2015**.

Data about Population, GDP, and Life Expectancy was collected from **World Bank Data**. Information about vaccinations for Measles, Hepatitis B, Polio, and Diphtheria, alcohol consumption, BMI, HIV incidents, mortality rates, and thinness was collected from **World Health Organization** public datasets. Information about schooling was collected from the **Our World in Data** which is a project of the University of Oxford.

Life Expectancy - Research Question

Among all the health, economic and demographic variables, which ones of them help predict a country's life expectancy in the year of 2015?



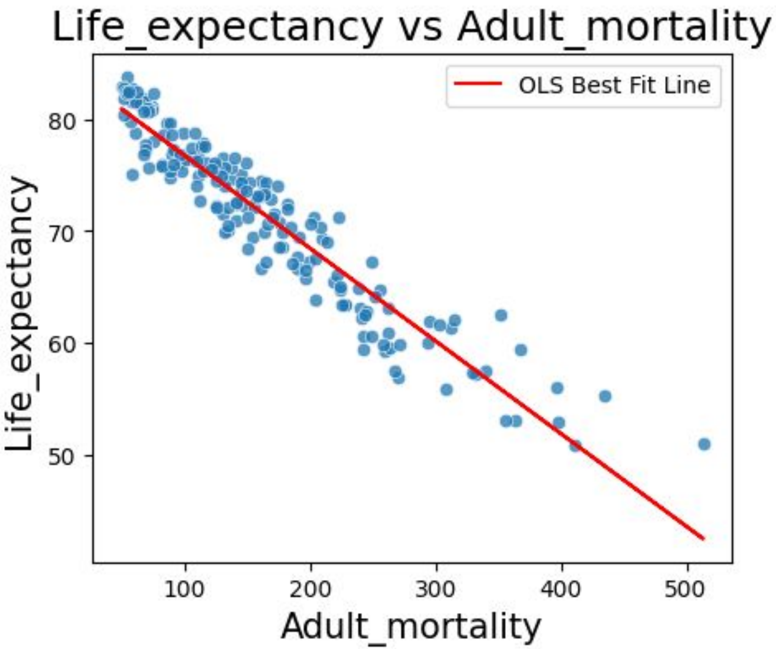
Source: World Bank Life Expectancy at Birth 2014

Life Expectancy - Simple Linear Regression

Ordinary Least Squares Regression:

Response: **Life_expectancy** (Average life expectancy of both genders)

Predictor: **Adult_mortality** (deaths of adults per 1000 population)



Life Expectancy - OLS Regression Results

OLS Regression Results

Dep. Variable:	Life_expectancy	R-squared:	0.901
Model:	OLS	Adj. R-squared:	0.900
Method:	Least Squares	F-statistic:	1609.
Date:	Thu, 27 Apr 2023	Prob (F-statistic):	8.79e-91
Time:	01:05:14	Log-Likelihood:	-415.01
No. Observations:	179	AIC:	834.0
Df Residuals:	177	BIC:	840.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.9897	0.384	221.054	0.000	84.231	85.748
Adult_mortality	-0.0826	0.002	-40.118	0.000	-0.087	-0.079

Omnibus:	0.324	Durbin-Watson:	2.218
Prob(Omnibus):	0.850	Jarque-Bera (JB):	0.110
Skew:	0.028	Prob(JB):	0.947
Kurtosis:	3.108	Cond. No.	388.

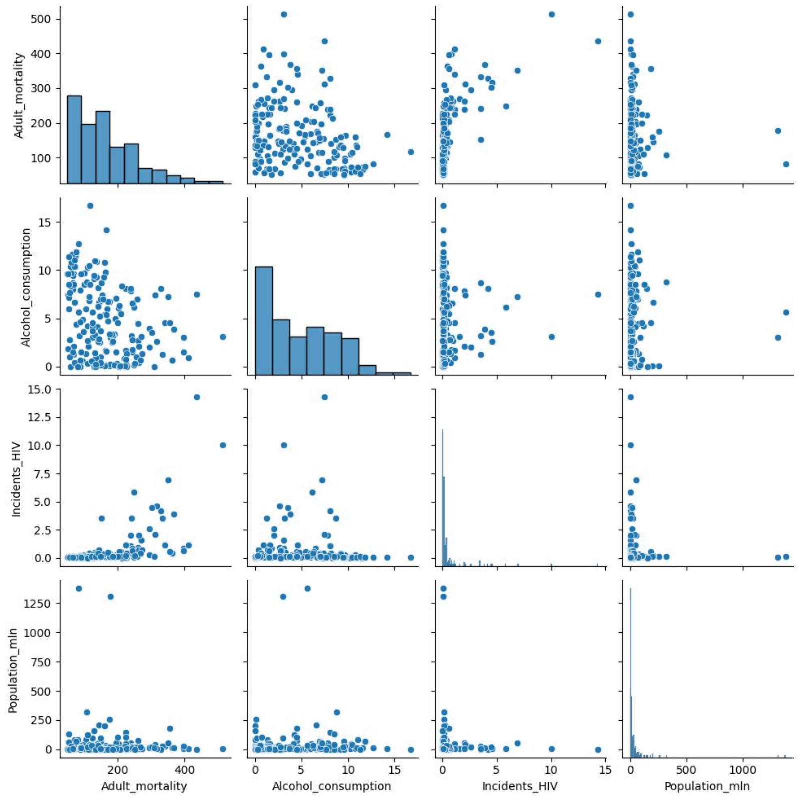
Life Expectancy - Multiple Linear Regression

Response: **Life_expectancy** (Average life expectancy of both genders)

Predictors:

Adult_mortality (deaths of adults per 1000 population)
Alcohol_consumption (Represents alcohol consumption that is recorded in liters of pure alcohol per capita with 15+ years old)
Incidents_HIV (Incidents of HIV per 1000 population aged 15-49)
Population_mln (Total population of a country in millions)

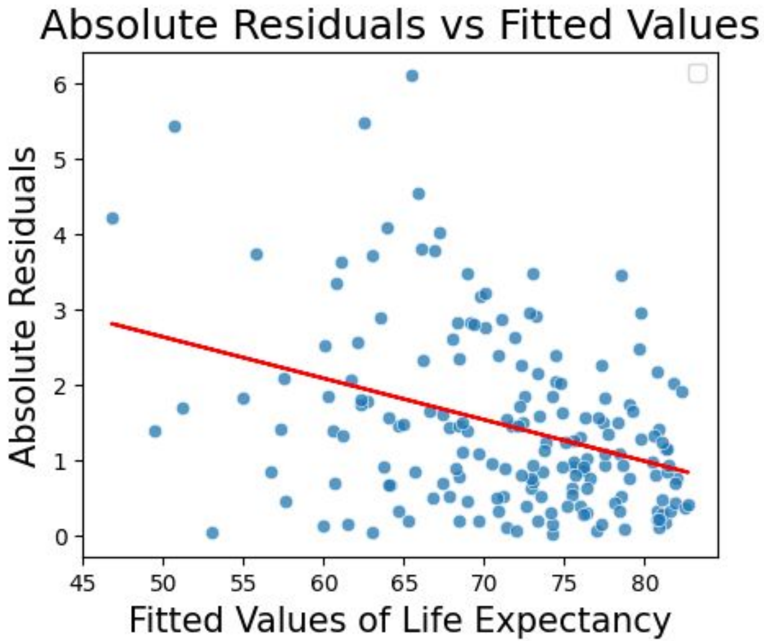
Life Expectancy - Multiple Linear Regression



Life Expectancy - WLS Multiple Linear Regression

Estimating the Weights for WLS:

1. Estimate the standard deviations (red line in the figure) by regressing the absolute residuals and the OLS fitted values of the response variable.
2. Estimate the variances by squaring the fitted values of the regression model
3. Calculate the weights by taking the reciprocal of the estimated variances

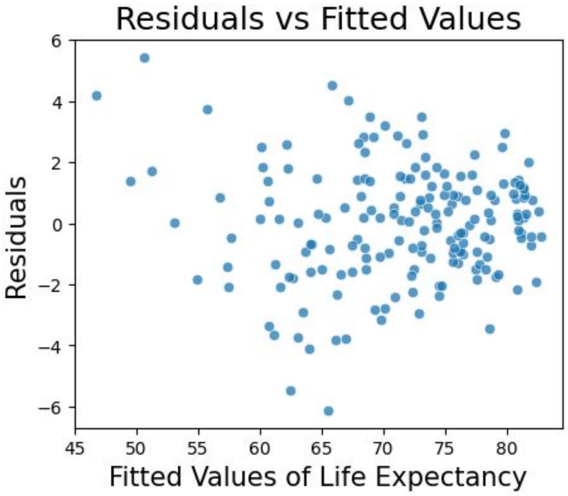


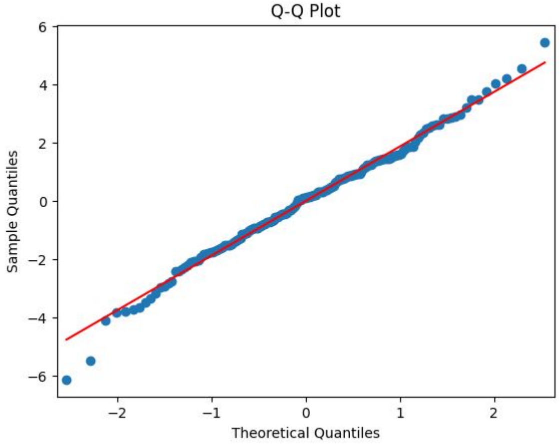
Life Expectancy - OLS vs. WLS Results

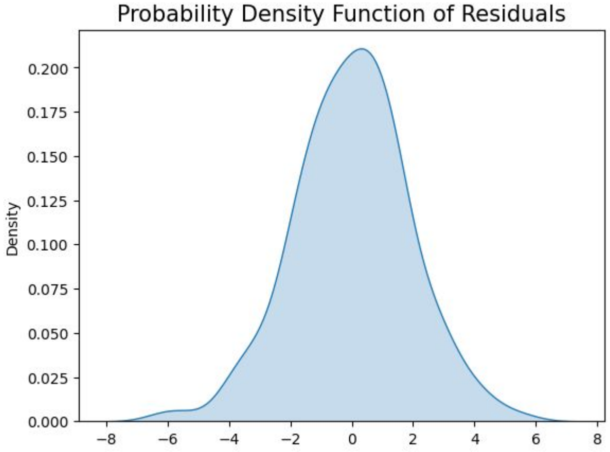
```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life_expectancy   R-squared:                       0.943
Model:                            OLS   Adj. R-squared:                  0.941
Method:                 Least Squares   F-statistic:                     715.5
Date:                Thu, 27 Apr 2023   Prob (F-statistic):          7.70e-107
Time:                        02:17:50   Log-Likelihood:                -366.01
No. Observations:                 179   AIC:                             742.0
Df Residuals:                     174   BIC:                             758.0
Df Model:                           4
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                  83.4165      0.446    187.167      0.000      82.537      84.296
Adult_mortality        -0.0849      0.002    -40.167      0.000      -0.089      -0.081
Alcohol_consumption     0.3395      0.041      8.286      0.000       0.259       0.420
Incidents_HIV           0.5900      0.112      5.259      0.000       0.369       0.811
Population_mln         -0.0007      0.001     -0.756      0.451      -0.003       0.001
==============================================================================
Omnibus:                        2.694   Durbin-Watson:                   2.204
Prob(Omnibus):                  0.260   Jarque-Bera (JB):                2.489
Skew:                          -0.136   Prob(JB):                        0.288
Kurtosis:                       3.509   Cond. No.                         613.
==============================================================================
```

```
                            WLS Regression Results
==============================================================================
Dep. Variable:        Life_expectancy   R-squared:                       0.941
Model:                            WLS   Adj. R-squared:                  0.940
Method:                 Least Squares   F-statistic:                     693.8
Date:                Sun, 30 Apr 2023   Prob (F-statistic):          9.57e-106
Time:                        17:28:23   Log-Likelihood:                -350.65
No. Observations:                 179   AIC:                             711.3
Df Residuals:                     174   BIC:                             727.2
Df Model:                           4
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                  83.7669      0.383    218.489      0.000      83.010      84.524
Adult_mortality        -0.0866      0.002    -39.427      0.000      -0.091      -0.082
Alcohol_consumption     0.3285      0.033     10.041      0.000       0.264       0.393
Incidents_HIV           0.5622      0.145      3.889      0.000       0.277       0.848
Population_mln         -0.0009      0.001     -1.064      0.289      -0.003       0.001
==============================================================================
Omnibus:                        1.002   Durbin-Watson:                   2.128
Prob(Omnibus):                  0.606   Jarque-Bera (JB):                1.107
Skew:                          -0.134   Prob(JB):                        0.575
Kurtosis:                       2.724   Cond. No.                         522.
==============================================================================
```

Life Expectancy - Diagnostic Tests







✓ **Linearity**

Life Expectancy - Diagnostic Tests

✓ **Normality**

Jarque-Bera Test Statistic: 2.49
P-value: 0.288
Null Hypothesis: The distribution of residuals is Normal.
Conclusion: Failed to reject the null hypothesis. There is no evidence to indicate the distribution is not Normal.

Skewness: -0.136
Kurtosis: 3.5

✓ **No Autocorrelation**

Durbin-Watson Test Statistic: 2.20
Null Hypothesis: There is no first-order autocorrelation in the residuals.
Conclusion: Failed to reject the null hypothesis. There is no evidence to indicate the presence of autocorrelation.

X **Heteroscedasticity**

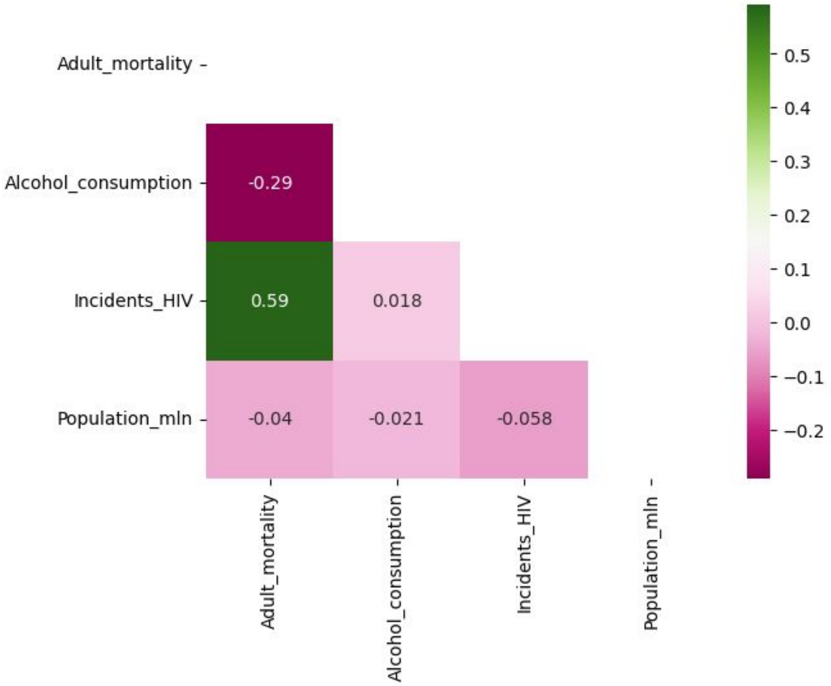
Breusch-Pagan Test Statistic: 68.10
P-value: 1.088e-14
Null Hypothesis: The variance of the residuals is constant (homoscedastic).
Conclusion: Reject the null hypothesis. There is strong evidence to indicate the presence of heteroscedasticity.

Life Expectancy - Diagnostic Tests

Variance Inflation Factors

Adult_mortality: 2.22
Alcohol_consumption: 1.59
Incidents_HIV: 1.52
Population_mln: 1.07

✓ **No Multicollinearity**



Life Expectancy - Backtesting

We back-tested both of the OLS and WLS models using 2014 life expectancy data:

OLS:

MSE: 3.58

R-squared: 0.944

WLS:

MSE: 3.57

R-squared: 0.945

The low MSE and high R-squared values indicate that both of the models fit 2014 data well.

Summary

When modeling the life expectancy of countries, the dataset exhibited heteroscedastic pattern in the error terms. In order to address the issue, we used the WLS method. The OLS and WLS methods turned out to generate very similar models with small MSE.

Given the fact that the OLS estimator was unbiased but inefficient, and the WLS estimator was biased but close to the OLS estimator and more efficient, we conclude that the WLS estimator was better.

In practice, when WLS is used to address heteroscedasticity, in order to obtain the best linear unbiased estimator (BLUE), it is important to use expert knowledge to decide the weights of the independent variables. If that knowledge is not available, then one can use estimated weights to obtain a robust WLS estimator. If the estimated coefficients differ substantially from the coefficients obtained by OLS, then iterate the WLS process by using residuals from the WLS fit to re-estimate the variance and then obtain revised weights. This process is called iteratively reweighted least squares.

Bitcoin

Bitcoin (BTC) is a cryptocurrency, a virtual currency designed to act as money and a form of payment outside the control of any one person, group, or entity, thus removing the need for third-party involvement in financial transactions. It is rewarded to blockchain miners for the work done to verify transactions and can be purchased on several exchanges.

Bitcoin was introduced to the public in 2009 by an anonymous developer or group of developers using the name Satoshi Nakamoto. It has since become the most well-known cryptocurrency in the world. Its popularity has inspired the development of many other cryptocurrencies. These competitors either attempt to replace it as a payment system or are used as utility or security tokens in other blockchains and emerging financial technologies.

Bitcoin

Research Question:

What time series variables can help predict the price of Bitcoin?

About our data:

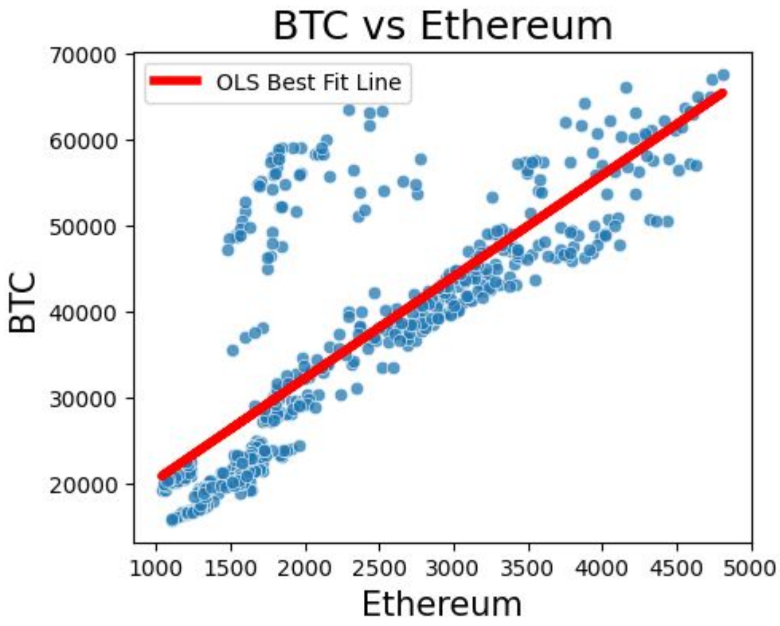
The dataset contains **549** historical data points from **2/02/2021 to 4/10/2023** that represent the daily dollar amounts in **6** variables: Bitcoin, Bitcoin Trade Volume, number of My Wallet transactions using Bitcoin per day, Nvidia Stock Price, Ethereum Stock Price, and Dow Jones Index.

Bitcoin - Simple Linear Regression

Ordinary Least Squares Regression:

Response Variable: **BTC** (Bitcoin Price)

Predictor: **Ethereum** (Ethereum Price)



Bitcoin - OLS Regression Results

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BTC   R-squared:                       0.636
Model:                            OLS   Adj. R-squared:                  0.635
Method:                 Least Squares   F-statistic:                     954.4
Date:                Thu, 27 Apr 2023   Prob (F-statistic):          5.03e-122
Time:                        01:23:58   Log-Likelihood:                -5747.8
No. Observations:                 549   AIC:                         1.150e+04
Df Residuals:                     547   BIC:                         1.151e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       8614.1776    964.796      8.929      0.000    6719.020    1.05e+04
Ethereum      11.7928      0.382     30.893      0.000      11.043      12.543
==============================================================================
Omnibus:                      224.787   Durbin-Watson:                   0.018
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              661.249
Skew:                           2.061   Prob(JB):                    2.58e-144
Kurtosis:                       6.451   Cond. No.                     6.69e+03
==============================================================================
```

Bitcoin - GLS Multiple Linear Regression

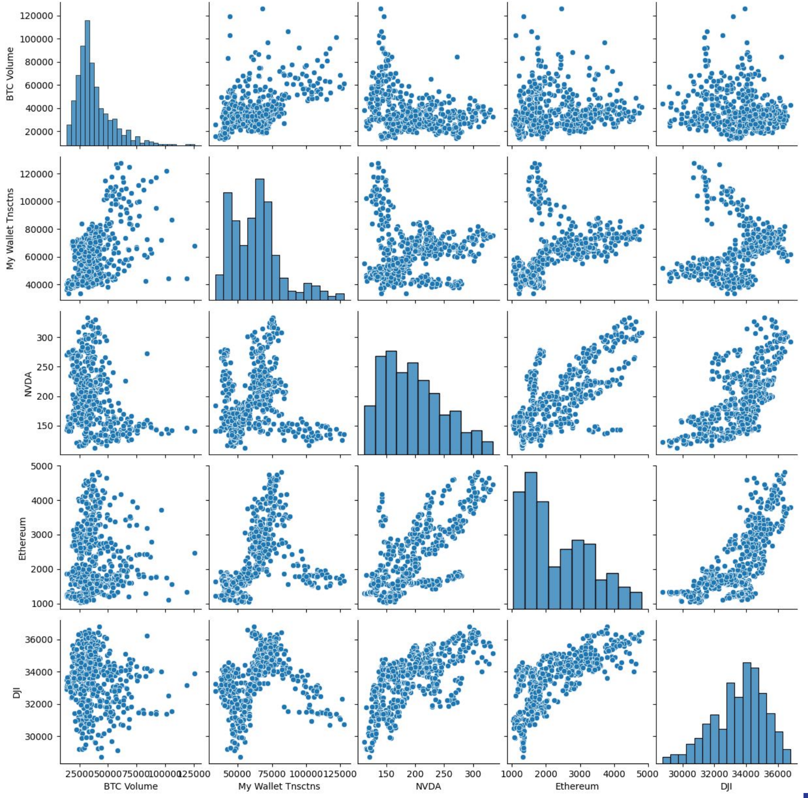
Using **Generalized Least Squares Method** for Multiple Linear Regression:

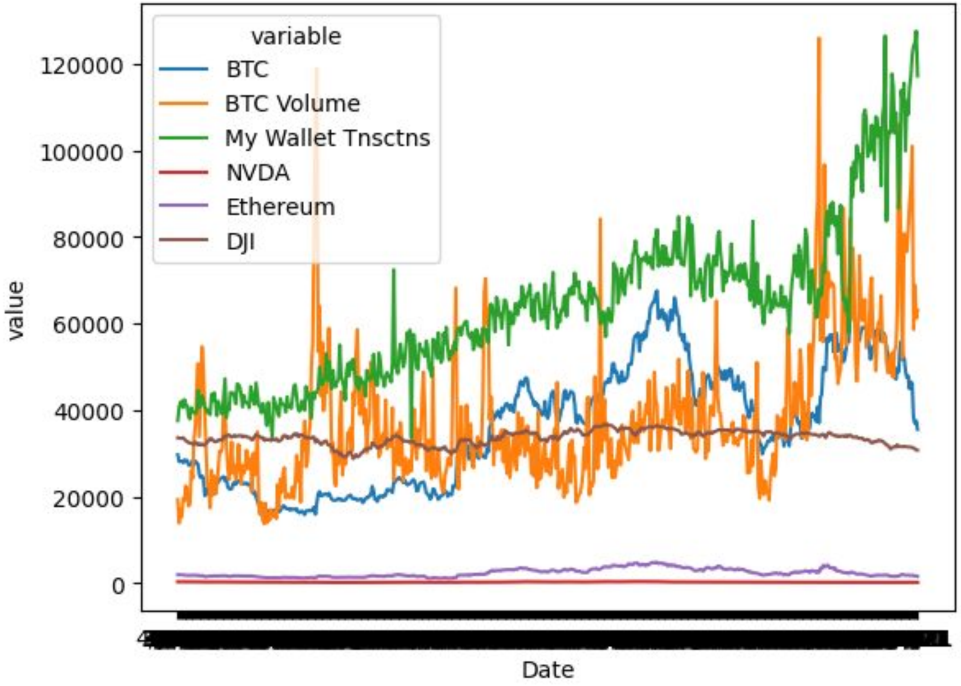
Response Variable: **BTC** (Bitcoin Price)

Predictors:

BTC Volume (Bitcoin Volume - mln)
Number of My Wallet transactions using Bitcoin per day
Nvidia Stock price
Ethereum Stock Price
DJI (Dow Jones Index)

Bitcoin - GLS Multiple Linear Regression





Bitcoin - GLS Multiple Linear Regression

Estimating the Covariance Matrix for GLS:

We assumed that the process generating the regression errors is stationary: That is, all of the errors have the same expectation (assumed to be 0) and the same variance (σ^2), and the covariance of two errors depends only upon their separation s in time:

$$C(\varepsilon_t, \varepsilon_{t+s}) = C(\varepsilon_t, \varepsilon_{t-s}) = \sigma^2 \rho_s$$

where ρs is the error autocorrelation at lag s. And our error covariance matrix would have this structure:

$$\mathbf{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix} = \sigma^2 \mathbf{P}$$

Bitcoin - GLS Multiple Linear Regression

Estimating the Covariance Matrix for GLS:

We chose the first-order auto-regressive process, AR(1), as our stationary time-series model:

$$\varepsilon_t = \phi\varepsilon_{t-1} + \nu_t$$

where the random shocks vt are assumed to be Gaussian white noise, NID(0,σ(v)^2). Under this model, ρ1 = φ, ρs = φ^s, and σ^2 = σ(v)^2/(1 − φ^2).

To estimate φ, we used OLS residuals and the AR(1) model.

Then, we estimated ρi using ρ1 = φ, ρs = φ^s.

Lastly, we estimated σ^2 by dividing the variance of the residuals of the AR(1) model by (1 − φ^2).

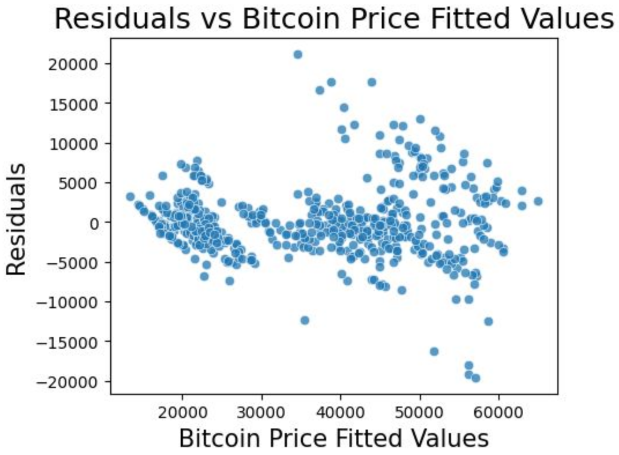
$$\mathbf{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix} = \sigma^2\mathbf{P}$$

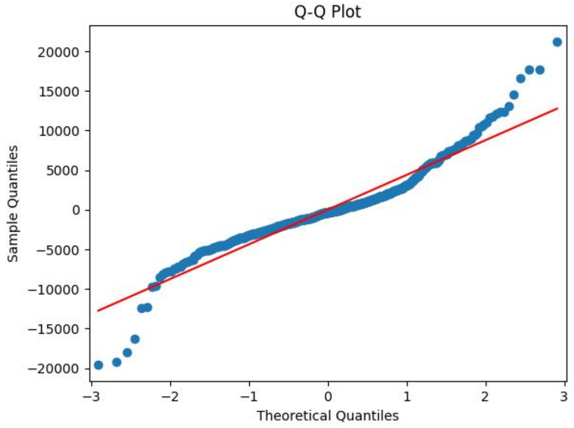
Bitcoin - OLS vs. GLS Results

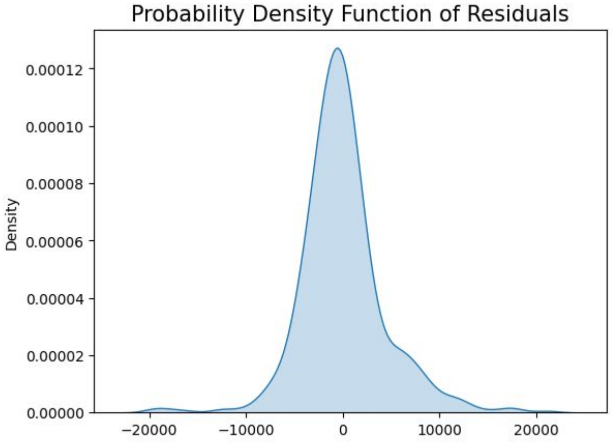
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BTC   R-squared:                       0.904
Model:                            OLS   Adj. R-squared:                  0.903
Method:                 Least Squares   F-statistic:                     1019.
Date:                Sun, 30 Apr 2023   Prob (F-statistic):          3.35e-273
Time:                        17:55:02   Log-Likelihood:                -5382.6
No. Observations:                 549   AIC:                         1.078e+04
Df Residuals:                     543   BIC:                         1.080e+04
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const             -3.104e+04   5817.850     -5.336      0.000   -4.25e+04   -1.96e+04
BTC Volume            0.0716      0.015      4.891      0.000       0.043       0.100
My Wallet Tnsctns     0.4106      0.014     29.504      0.000       0.383       0.438
NVDA                  9.4982      6.328      1.501      0.134      -2.932      21.928
Ethereum              7.3551      0.400     18.408      0.000       6.570       8.140
DJI                   0.5794      0.189      3.071      0.002       0.209       0.950
==============================================================================
Omnibus:                       75.498   Durbin-Watson:                   0.274
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              407.826
Skew:                           0.447   Prob(JB):                     2.77e-89
Kurtosis:                       7.127   Cond. No.                     2.58e+06
==============================================================================
```

```
                            GLS Regression Results
==============================================================================
Dep. Variable:                    BTC   R-squared:                       0.670
Model:                            GLS   Adj. R-squared:                  0.667
Method:                 Least Squares   F-statistic:                     220.2
Date:                Sun, 30 Apr 2023   Prob (F-statistic):          4.34e-128
Time:                        18:13:06   Log-Likelihood:                -4745.9
No. Observations:                 549   AIC:                             9504.
Df Residuals:                     543   BIC:                             9530.
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const             -3709.4961   5914.201     -0.627      0.531   -1.53e+04    7908.019
BTC Volume            0.0301      0.006      5.034      0.000       0.018       0.042
My Wallet Tnsctns     0.0970      0.011      8.813      0.000       0.075       0.119
NVDA                -11.8976      8.278     -1.437      0.151     -28.158       4.363
Ethereum              9.7779      0.415     23.552      0.000       8.962      10.593
DJI                   0.3614      0.201      1.795      0.073      -0.034       0.757
==============================================================================
Omnibus:                      145.932   Durbin-Watson:                   1.279
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              530.373
Skew:                           1.190   Prob(JB):                    6.78e-116
Kurtosis:                       7.186   Cond. No.                     1.29e+06
==============================================================================
```

Bitcoin - Diagnostic Tests







X No Linearity

Bitcoin - Diagnostic Tests

Jarque-Bera Test Statistic: 407.83
P-value: 2.77e-89
Null Hypothesis: The distribution of residuals is Normal.
Conclusion: Reject the null hypothesis. There is strong evidence to indicate the distribution is not Normal.

X No Normality

Skewness: 0.45
Kurtosis: 7.13 (leptokurtic)

X Autocorrelation

Durbin-Watson Test Statistic: 0.27
Null Hypothesis: There is no first-order autocorrelation in the residuals.
Conclusion: Reject the null hypothesis. There is strong evidence to indicate the presence of autocorrelation.

Breusch-Pagan Test Statistic: 180.39
P-value: 6.15e-38
Null Hypothesis: The variance of the residuals is constant (homoscedastic).
Conclusion: Reject the null hypothesis. There is strong evidence to indicate the presence of heteroscedasticity.

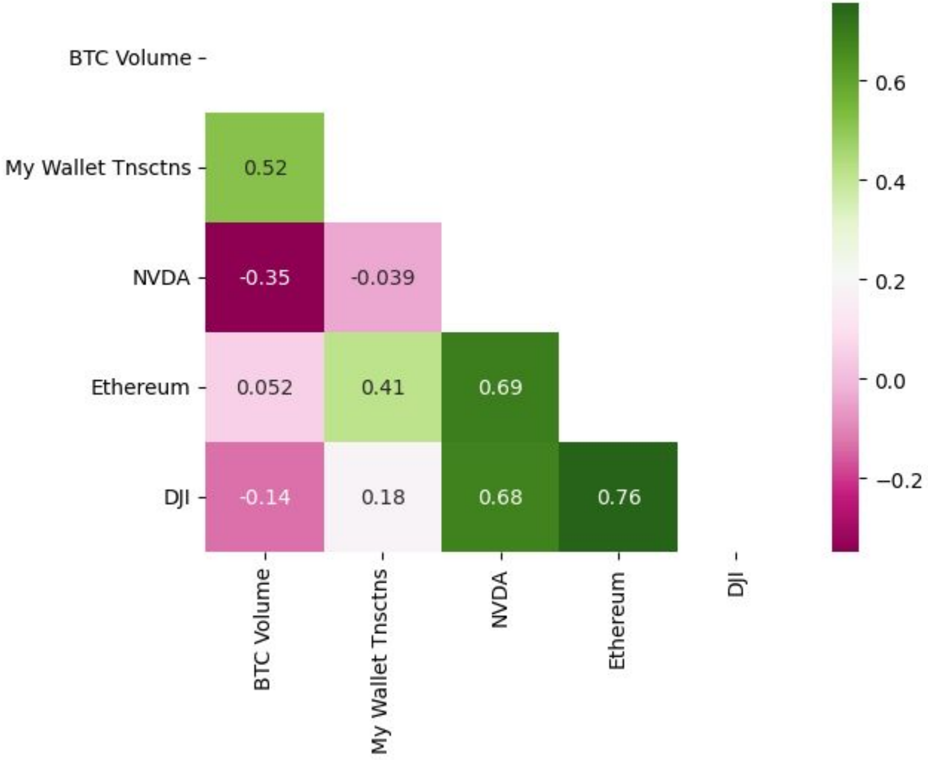
X Heteroscedasticity

Bitcoin - Diagnostic Tests

Variance Inflation Factors

BTC Volume: 10
My Wallet Bitcoin Transaction: 23.47
NVDA: 46.645
Ethereum: 19.30
DJI: 53.6

X **Multicollinearity**



Bitcoin - Backtesting

We back-tested both of the OLS and GLS models using daily data from 1/31/2020 to 2/01/2021:

OLS:

MSE: 534,946,834.87
R-squared: -8.35

GLS:

MSE: 64,468,795.13
R-squared: -0.127

The large MSE and negative R-squared values indicate that both of the OLS and the GLS models are predicting worse than the mean of the bitcoin price.

Summary

As you saw earlier, the Bitcoin dataset failed all of the diagnostic tests. We used GLS to model the Bitcoin price, hoping at least the issue of autocorrelation would be addressed. By comparing the results of OLS and GLS, the two methods indeed generated drastically different models.

So, which model was better? The OLS model was not appropriate as the data failed all of the assumptions required. As for the GLS model, the assumptions for linearity, normality, homoscedasticity, and multicollinearity were still not met. And the assumption of stationary regression errors might not be true in our data. Our covariance matrix might be erroneous. That means, the GLS estimator was not the BLUE, either. Furthermore, the backtest results confirmed that neither one was a good model for our time-series dataset.

In conclusion, neither OLS or GLS is suitable for time-series data. We should follow a procedure of tests, such as the Augmented Dickey-Fuller Test and the Johansen Test, to check stationarity and cointegration. If both of these tests fail, then we should consider Vector Autoregression and Vector Error Correction Model for time-series data.

Data Sources

Kaggle: https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated

Investopedia:https://www.investopedia.com/terms/b/bitcoin.asp#:~:text=Bitcoin%20(BTC)%20is%20a%20cryptocurrency,party%20involvement%20in%20financial%20transactions

Average life expectancy of both genders in different years from 2010 to 2015:
https://www.who.int/data/gho/data/indicators/indicator-details/GHO/life-expectancy-at-birth-(years)

Mortality-related attributes (infant deaths, under-five-deaths, adult mortality):
https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates

Alcohol consumption that is recorded in liters of pure alcohol per capita with 15+ years old:
https://www.who.int/data/gho/data/indicators/indicator-details/GHO/alcohol-recorded-per-capita-(15-)-consumption-(in-litres-of-pure-alcohol)

% of coverage of Hepatitis B (HepB3) immunization among 1-year-olds:
https://www.who.int/data/gho/data/indicators/indicator-details/GHO/hepatitis-b-(hepb3)-immunization-coverage-among-1-year-olds-(-)

% of coverage of Measles containing vaccine first dose (MCV1) immunization among 1-year-olds:
https://www.who.int/data/gho/data/indicators/indicator-details/GHO/measles-containing-vaccine-first-dose-(mcv1)-immunization-coverage-among-1-year-olds-(-)

% of coverage of Polio (Pol3) immunization among 1-year-olds:
https://www.who.int/data/gho/data/indicators/indicator-details/GHO/polio-(pol3)-immunization-coverage-among-1-year-olds-(-)

Data Sources

% of coverage of Diphtheria tetanus toxoid and pertussis (DTP3) immunization among 1-year-olds:
https://www.who.int/data/gho/data/indicators/indicator-details/GHO/diphtheria-tetanus-toxoid-and-pertussis-(dtp3)-immunization-coverage-among-1-year-olds-(-)

BMI: https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations

Incidents of HIV per 1000 population aged 15-49: https://data.worldbank.org/indicator/SH.HIV.INCD.ZS

Prevalence of thinness among adolescents aged 10-19 years. BMI < -2 standard deviations below the median:
https://www.who.int/data/gho/indicator-metadata-registry/imr-details/4805

GDP per capita in current USD: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?most_recent_year_desc=true

Total population in millions: https://data.worldbank.org/indicator/SP.POP.TOTL?most_recent_year_desc=true

Average years that people aged 25+ spent in formal education: https://ourworldindata.org/grapher/mean-years-of-schooling-long-run

Yahoo Finance:
https://finance.yahoo.com/quote/GOOG/history?period1=1644278400&period2=1675814400&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true

Bitcoin My Wallet Number of Transaction Per Day: https://data.nasdaq.com/data/BCHAIN/MWNTD-bitcoin-my-wallet-number-of-transaction-per-day

Time-Series Regression and Generalized Least Squares in R*:
https://socialsciences.mcmaster.ca/jfox/Books/Companion/appendices/Appendix-Timeseries-Regression.pdf

Questions?