Crime in New York City

Part II. Data Stationary

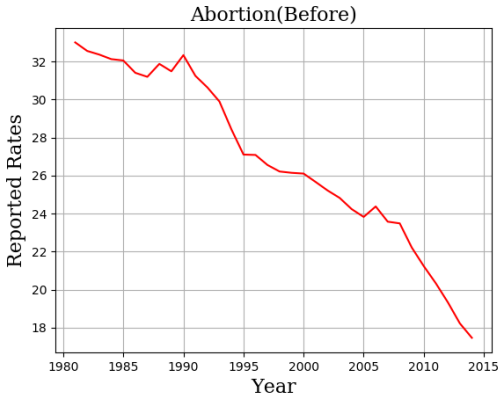
I. Stationary Examination

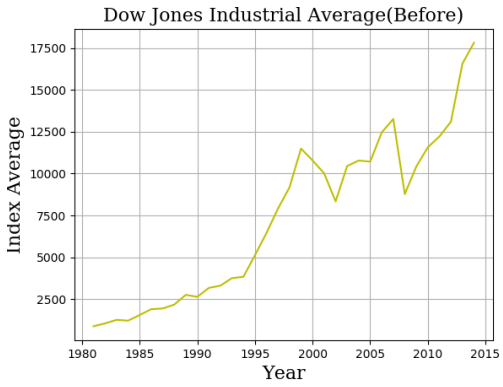
Stationary data has the property that its mean, variance, covariance and autocorrelation structure is stable with time and do not have trend or periodic effects. When using time series data, before establishing a model to fit the data we have, a stationary examination and processing is necessary if we expect a well-performed result.

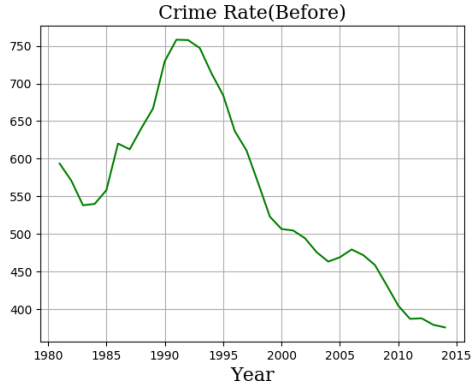
Also, in causality analysis, … (slides/paper)

II. First data analysis

The easiest step is to plot the current data we have and see if they are stable or not:







From the graph we can easily see all four features we used in this research show some upward or downward trends as year goes by. For a stationary series, the trend shouldn't present in figure, thus, by simply observing the graphs we know a stationary processing is needed here in our research.

III. Statistical Tests

We also used some statistic tests to help analyzing the data we have. The common two tests, as we listed below, are Augmented Dickey Fuller Test (ADF) and Kwiatkowski-Phillips-Schmidt-Shin Test (KPSS). They both have their advantages in

stationary examination.

ADF (Augmented Dickey Fuller) Test

ADF is more popular and commonly used, a unit root test for stationary,
The problem is, the type I error rate often occurs in ADF Test though it can handle complex models and widely used.

$$\Delta y_t = \alpha + \delta y_{t-1} + \Sigma_{i=1}^{n} \beta_i \Delta y_{t-i} + \varepsilon_t$$

- Null Hypothesis (H0): data series is non-stationary (has unit root, $\delta = 0$)
- Alternative Hypothesis (Ha): data series is stationary or trend-stationary. ($\delta < 0$)

Table. ADF result before/after stationary processing

	Abortion		Dow jones		Incarceration		Crime Rate	
Test Statistic	1.6780	-4.6622	-0.1516	-2.5718	-2.1849	-2.2808	-0.9739	-3.4022
p-value	0.9981	0.0001	0.9440	0.0990	0.2117	0.1782	0.7834	0.0109
Critical Value (1%)	-3.6461	-3.6535	-3.7377	-3.7377	-3.6535	-3.6535	-3.6535	-3.7239
Critical Value (5%)	-2.9541	-2.9572	-2.9922	-2.9922	-2.9572	-2.9572	-2.9572	-2.9865
Critical Value (10%)	-2.6160	-2.6176	-2.6357	-2.6357	-2.6176	-2.6176	-2.6176	-2.6328

* Left column is before while right is data after stationary processing.

Analysis

In table we can see the test statistics for abortion rate, Down Jones Index Average, incarceration rate, crime rate in ADF test is 1.6780, -0.1516, -2.1849, -0.9739 which is larger than critical value at 90% (10%), 95% (5%) and 99% (1%) confidence intervals, and this means the H0 should be accepted.
P-value is 0.9981, 0.9440, 0.2117, 0.7834 which is larger than significant level $(\alpha = 10\%, 5\%, 1\%)$, and this also tells us the null hypothesis (H0) holds.
Therefore, all features needs to be stationary.

KPSS (Kwiatkowski-Phillips-Schmidt-Shin) Test

This test tells us whether a time series is stationary with a linear trend or nonstationary (trend stationary). KPSS applies to those timeseries that have a trend in their plots.

$$\mathrm{KPSS} = \mathrm{n}^{-2} \Sigma_{i=1}^{n} \frac{S_t}{\bar{\sigma}^2}$$

Where S is the sum of residuals.

$$\mathrm{S_t} = \Sigma_{i=1}^{n} e_i$$

And $\bar{\sigma}^2$ is the estimate of variance of residuals.

Hypothesis

Null Hypothesis (H0): The data's trend is stationary
Alternative Hypothesis (Ha): The data series has a unit root(nonstationary)

Table. KPSS result before/after stationary processing

	Abortion		Dow jones		Incarceration		Crime Rate	
Test Statistic	0.4496	0.3250	0.4407	0.2250	0.3973	0.3597	0.3212	0.1994
p-value	0.0558	0.1000	0.0596	0.1000	0.0783	0.0945	0.1000	0.1000
Critical Value (1%)	0.3470	0.3470	0.3470	0.3470	0.3470	0.3470	0.3470	0.3470
Critical Value (5%)	0.4630	0.4630	0.4630	0.4630	0.4630	0.4630	0.4630	0.4630
Critical Value (10%)	0.7390	0.7390	0.7390	0.7390	0.7390	0.7390	0.7390	0.7390

* Left column is before while right is data after stationary processing.

Analysis:

In table we can see the test statistics for abortion, Dow jones, Incarceration in KPSS test is 0.4496, 0.4407, 0.3973 which is larger than critical value at 99% (1%) confidence intervals, and this means the H0 should be rejected. For critical value at 95%, 90%, the H0 should be accepted.

P-value is 0.0558, 0.0596, 0.0783, 0.1000 which is larger or equal than significant level $(\alpha = 10\%, 5\%, 1\%.)$, and this also tells us the null hypothesis holds.

Therefore, all features with trend are stationary.

Cases discussion:

Case 1:

Both tests are non-stationary

Case 2:

Both tests are stationary

Case 3:

ADF is non-stationary and KPSS is stationary

This is trend stationary. The data series has no unit root but its trend is stationary. Once the trend is removed the result series will be strict stationary, which means the mean and variance and covariance in this data series are not a function of time.

Case 4:

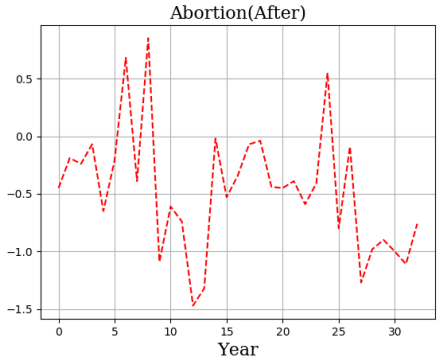
ADF is stationary and KPSS is non-stationary

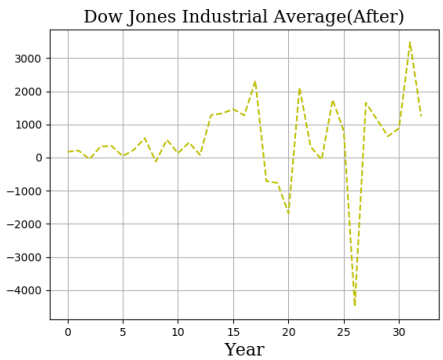
This is difference stationary. We should use differencing to make series stationary.

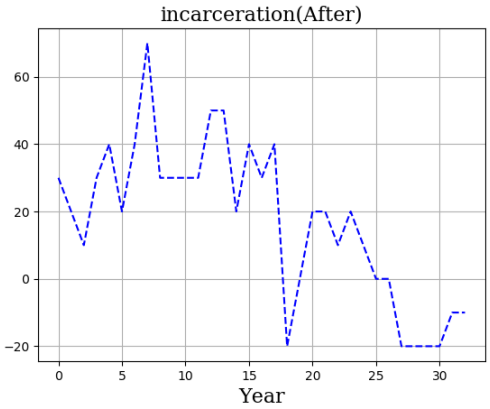
Therefore, we are facing the case 3, where all features are trend stationary.

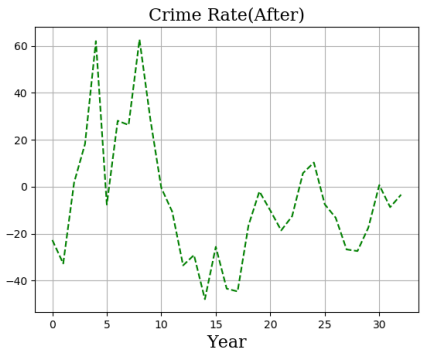
Stationary Process

In this step, our aim is to remove the trend in data we have. We will use differencing method, we calculate the differences to detrending data.









Before using these data, we still need to do normalization.

Part III. Linear Regression Model

I. First Model: linear

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

* Formula uses Wilkinson notation

Table. Result for Linear Regression Model

	Coefficient Estimates	Standard Error of Coefficients	t-statistic for H0	P value
Constant (β_0)	8.9658e-18	0.030266	2.9623e-16	1
Abortion (β_1)	0.036012	0.11155	0.32284	0.74922
Down Jones (β_2)	-0.12443	0.14244	-0.87353	0.3898
Incarceration (β_3)	-0.078681	0.13824	-0.56914	0.5738

Analysis

The p-values in first LR model examine the null hypothesis H0 that there is no effect on the corresponding row feature and in other words, the coefficient should be zero. Set the common statistic significant value to be 0.05, and compare with p value, we found that none of these features' coefficient should remain in this model.

For t value we have, the closer it to zero, the more likely there is no significant difference between predictors and response variable, which means H0 is true.

II. Second Regression Model: quadratic

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \epsilon$$

Table. Result for Quadratic

	Coefficient Estimates	Standard Error of Coefficients	t-statistic for H0	P value

β_0	-0.057056	0.065566	-0.87021	0.39358
β_1	0.033332	0.12092	0.27564	0.7854
β_2	-0.19545	0.15912	-1.2283	0.23232
β_3	-0.053268	0.2387	-0.22316	0.82547
β_{12}	-1.1568	0.76759	-1.5071	0.14601
β_{13}	0.44467	0.92644	0.47998	0.63598
β_{23}	-1.2112	0.86576	-1.399	0.17576
β_{11}	-0.2281	0.56288	-0.40523	0.68922
β_{22}	0.98563	0.76322	1.2914	0.20997
β_{33}	0.49607	0.53797	0.92211	0.36648

Analysis

The p-values in second is still too large for common significant level 0.005 but some rows are better than model 1.

Bibliography

Johnston, R. (2018, November). Historical abortion statistics, New York (USA). Retrieved November 2018, from http://www.johnstonsarchive.net/policy/abortion/usa/ab-usa-NY.html

Reproductive Health. (2017, November 16). Retrieved November 10, 2018, from https://www.cdc.gov/reproductivehealth/data_stats/abortion.htm

Brownlee, J. (2018, October 18). How to Check if Time Series Data is Stationary with Python. Retrieved November 21, 2018, from https://machinelearningmastery.com/time-series-data-stationary-python/

Charpentier, A. (2018). Unit Root Tests. [online] Freakonometrics. Available at: https://freakonometrics.hypotheses.org/12729 [Accessed 26 Nov. 2018].

Dss.princeton.edu. (2018). DSS - Interpreting Regression Output. [online] Available at: https://dss.princeton.edu/online_help/analysis/interpreting_regression.htm [Accessed 26 Nov. 2018].

Minitab. (2018). How to Interpret Regression Analysis Results: P-values and Coefficients. [online] Blog.minitab.com. Available at: http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients [Accessed 26 Nov. 2018].

Iordanova, T. (2018). Introduction To Stationary And Non-Stationary Processes. [online] Investopedia. Available at: https://www.investopedia.com/articles/trading/07/stationary.asp [Accessed 26 Nov.

2018].

Stat.yale.edu. (2018). Linear Regression. [online] Available at: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm [Accessed 26 Nov. 2018].

Statistics How To. (2018). Stationarity: Definition, Examples, Types - Statistics How To. [online] Available at: https://www.statisticshowto.datasciencecentral.com/stationarity/ [Accessed 26 Nov. 2018].

People.maths.bris.ac.uk. (2018). Time Series Introduction. [online] Available at: https://people.maths.bris.ac.uk/~magpn/Research/LSTS/STSIntro.html [Accessed 26 Nov. 2018].