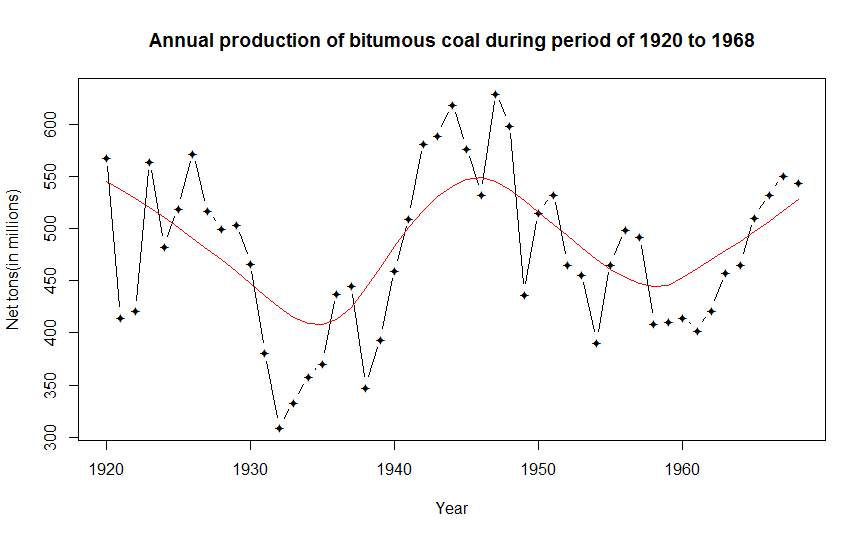
# Introduction

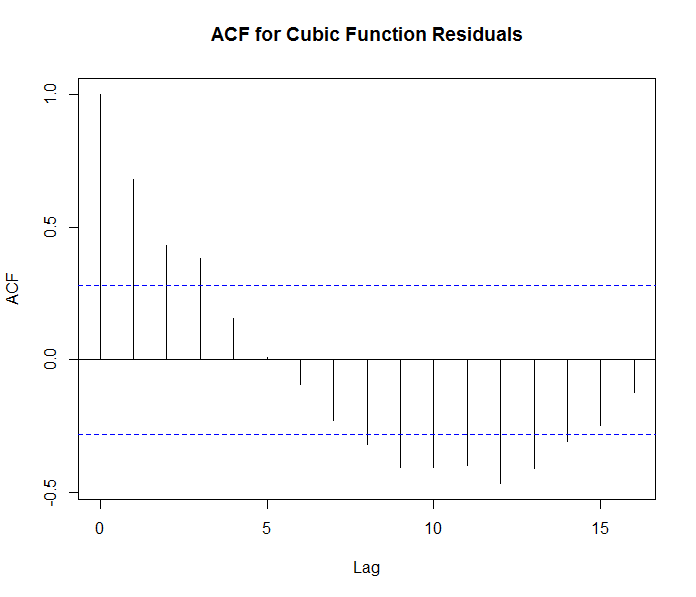
Bituminous coal is the most common type of coal in United States, accounting for over 50% of demonstrated reserve base. And it is the main source of providing electricity and heat, and also could extracted to produce steels. This report is trying to analyse the fluctuation and relationship of annual production of bituminous coal based on time series analysis, from period of 1920 to 1968. Analysis would present different dimensions to support and validate my final model and prediction, following with conclusion, which contains the final model.

# Time series Analysis

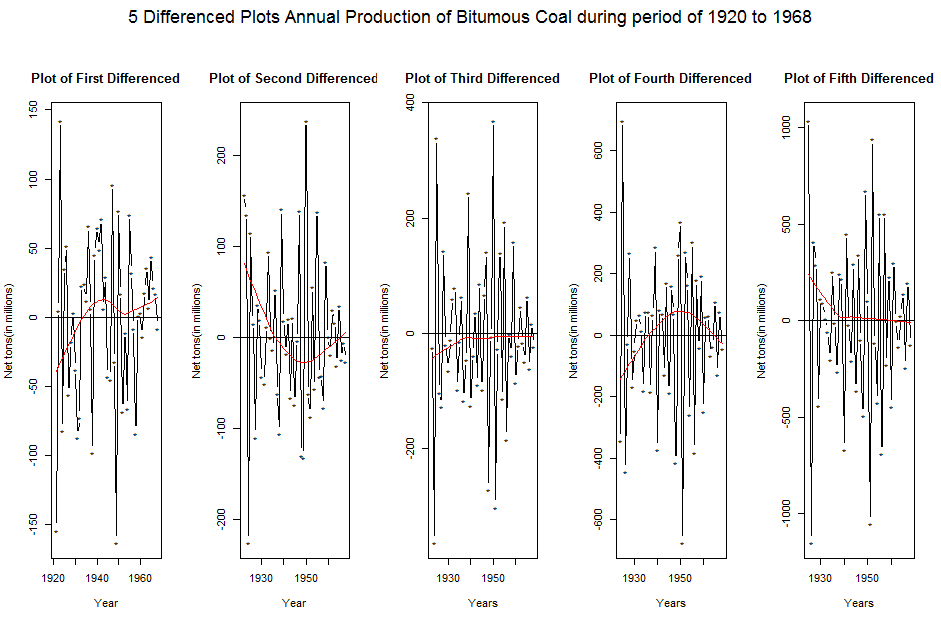
Trend Model

Time series based data and prediction is normally decomposed into trend, seasonal effect and its irregularity. At the first, it is reasonable to observe the overall trend and make a brief estimation of basic model.



The above graph draws the plot of this time series data and also fits the adjusted lowess curve to offer researchers basic trend about this graph. At first, we look into spread of time series as it progresses through time, the spread is roughly the same indicates the stabilised variance and transformation might not by required. Also, the applied log transformation doesn’t change a lot. As a result, non-transformation is applied in my model. Then, smooth lowess curve captures the overall trend of this time series. By detrending data, in other words, removing specific trend from time series dataset, there are two ways to detrend time series data. They are fitting lower order polynomial and differencing model respectively. The first option is fitting lower order polynomial, it is obvious that the trend goes down, goes up and goes down again, which is quintic-like or quartic-like polynomial, but fitting high orders into this small dataset (49 data) is problematic as it run out of degrees of freedom. As a result, fitting a cubic function could an ideal option.

However, the ACF (left hand side graph) suggests the issue of non-stationarity, as we can see from graph, the lag is not decreasing as the number increases. More specifically, the spike of lag 8 ,9 and 10 suggest it might be need differencing data or to employ different models.

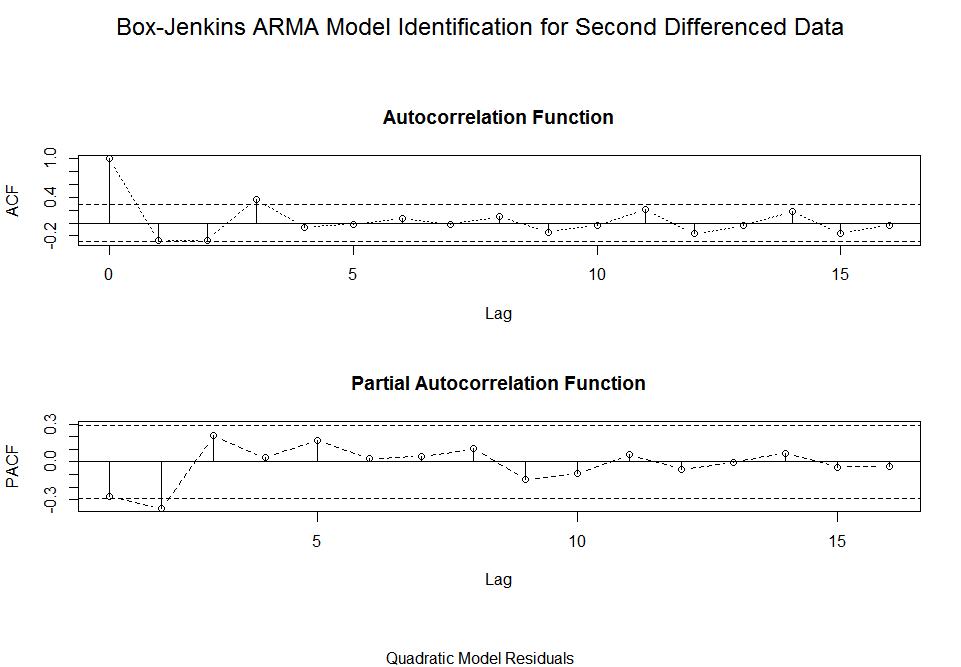


Since high orders polynomial function decreases degree of freedom and cubic function exist non-stationarity and might not fully explained the data. In this case, it is reasonable to detrend data by differencing data and try to capture more proper trend of dataset. Since differencing would bring down the highest order and reduce the loss of degree of freedom. From the graph we could see the first differencing shows the cubic function model and second differencing shows the quadratic function. I am strongly convinced by the original trend is a quartic function in this case and after second differenced (decrease power by 2), the second plot also shows the quadratic function which support my observation of “real trend” is a quartic function. Although third differenced and fifth differenced plot demonstrate constant trend, according to Anderson (1975), each successive differencing will decrease the variance of the series, but at some point, higher-order differencing will lead to an increase in variance. In this point, increasing differencing might not always a good idea and I am not convinced by the production of coal should be constant due to the effect of over-production or under-production of previous years, the production of coals would adjust to produce less or more respectively in the following year. As a result, based on my judgment, second differencing already offer a decent variance stabilised outcome in regardless of two abnormal (extreme value). While the first differencing’s residual plot doesn’t provide good diagnostic results compared to second differencing, which would be discussed later. Also, since it is a relatively small sample size data, it is reasonable to choose lower order polynomial while I also want to avoid over-differencing. As a result, the quadratic model with second differencing would suggest to explain the trend well.

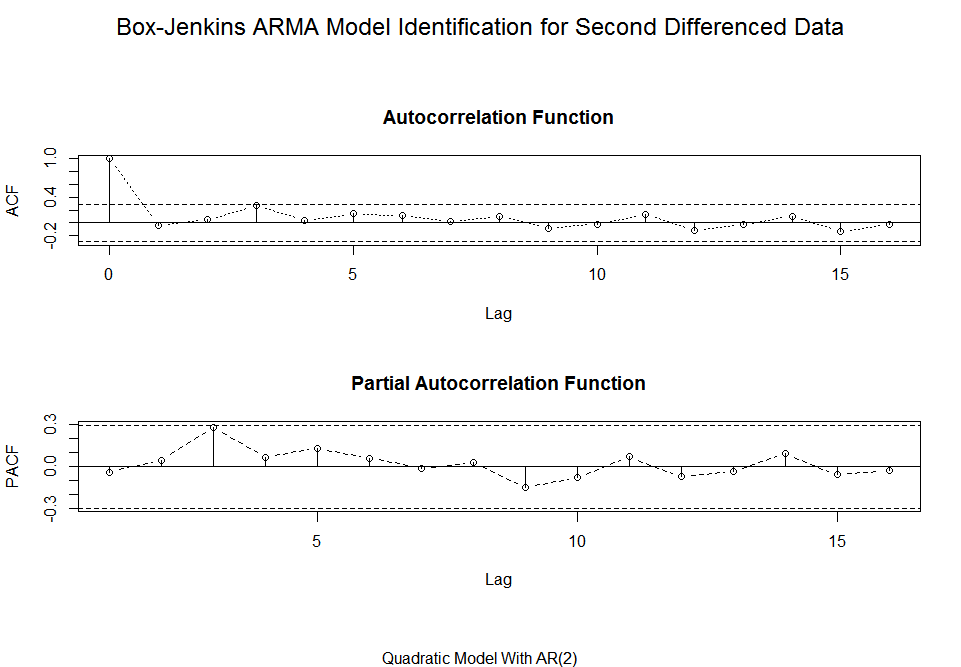
Seasonal Effect

When it comes to the seasonal effect, there is no seasonal effect observed and frequency equals to one might limit the exploration of seasonal effect. The production of coal fluctuates significantly and was mainly effected by unpredictable events, the production of coals from 1939 to 1945 soared up due to the Second World War and its energy and weapon requirements. Also, unpredictable frigid weather in early of 1947 also increased demand and production of coals. As a result, no seasonal effect would be fit into my final model.

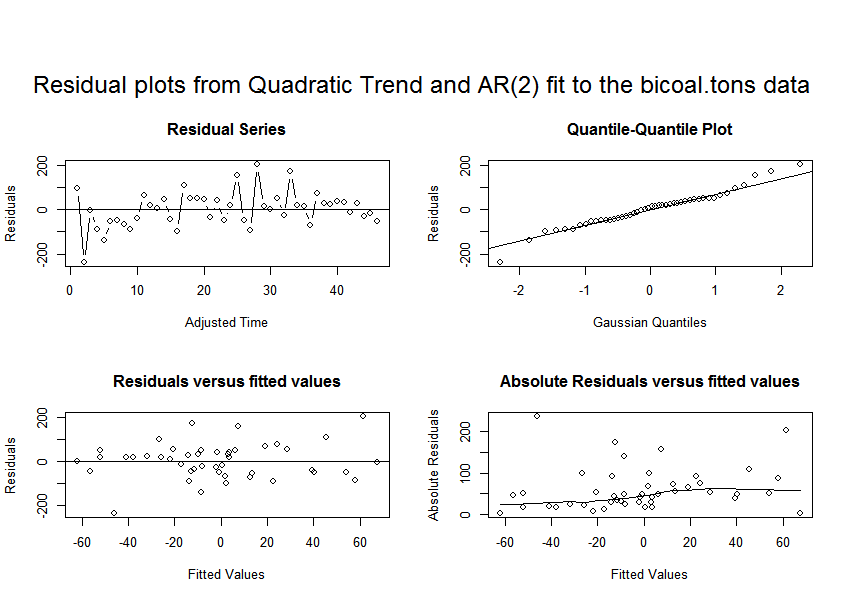
Irregularity



After fitting the trend, it is suggested to examine the residual plot that fit for dependence and fit AR model when it is necessary. As shown the above graph, it is suggested that ACF is not decay fast enough to zero after lag 10 although it shows stationary in my perspective. And PACF has spike at lag 2, which is also the last significant spike in the PACF, indicating the AR (2) would be able to explain irregular term. As a result, it is reasonable to fit AR (2) model to further examine its residual plot.



After fitting AR (2) model, ACF shows relatively better and smooth decay after lag 10 compare to one without fitting, and also there is no spike in ACF and PACF. As a result, AR (2) model is believe to fit in to explain the irregularity.



The residual from fit to irregular series is assessed and clean out the dependence, also diagnostic plots would be investigated further support and assumption and validate final model. The residual randomly plot in around the zero in residuals series.

In the qq plot, almost all the points consistent with qq line and no obvious curve observed, this linearity suggest the normal distribution except some tail points is affected by extreme values

As shown in residuals versus fitted values, the residuals bounce randomly around the zero line, expect the extreme value caused by unpredictable events. This suggests that the assumption that the relationship is linear is reasonable. Also, the variance is relatively stable and suggest no unequal error variances.

The Absolute Residuals versus fitted values also shows the relatively flat pattern, if I fit cubic function with first differenced, it shows bit trend of unequal variance. In this case, it further supports my modelling and valid assumption.

Overall speaking, the diagnostic is relatively good and support all the assumptions.

Model

Bituminous coals Production=Z(t)

Z(t)=71.90151501-6.9727354\*(t-1922) + 0.1251789\*(t-1922)^2 + X(t)

X(t)=-0.3081955 \* X(t-1)-0.2757575 \* X(t-2)

Conclusion

The quadratic function with second differenced data seems fit and explain dataset well and residual plots support its assumption. Since there are a few unpredictable events affect the fluctuation of coals production, the further researches should focus on analysing black swan effect and factors to accurately estimate trends and sort out better modelling.

The zero line and lowess curve (red line) is used to help researchers to compare visually effectiveness of differencing. Since quintic and cubic trend are regards as “real trend” in my perspective, the third and

Detrending means remove the specific trend off from time series dataset

Time series is normally decomposed into trend, seasonal and irregularity.

Difference try to fit different order differenced model to fit time series and remove the trend.

Use Each difference model to presents ACF and PACF and examine the stationarity

What is Robust PACF

Bituminous contains the widest range of carbon content and it is mainly used as a fuel to generate electricity, also used as a fuel to generate electricity.