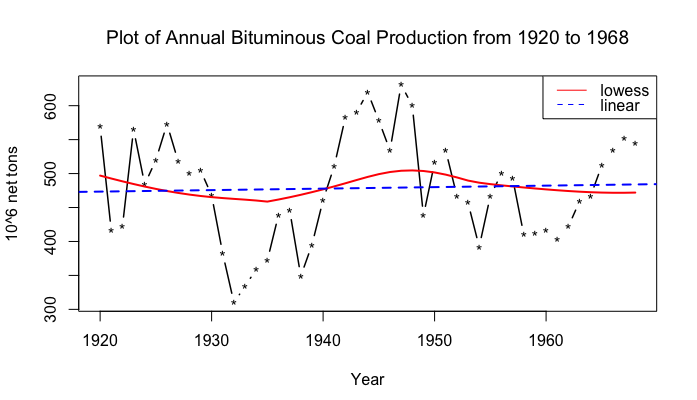
STAT7026 Assignment 2

## 1. Data

The time series bicoal.tons provided by U.S. Bureau of Mines stores the annual production of bituminous coal between 1920 and 1968 (in millions of net tons). Our goal is to explore the data and dig up any pattern hidden inside.

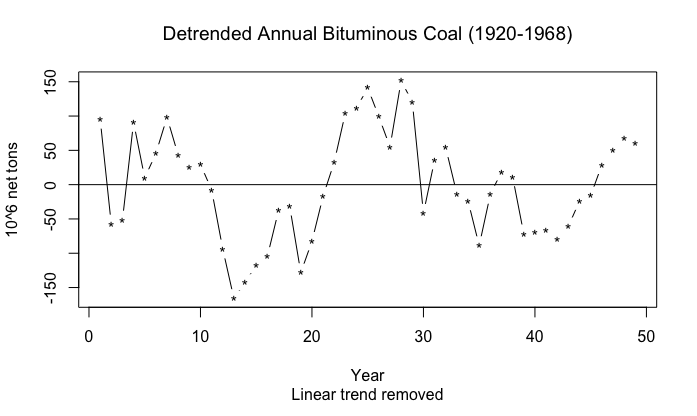
## 2. Plotting and Decomposing



We tried a log-scale plot and a square-root-scale plot right after this. However, they did not provide a better view of the data and looked almost identical, so there is no need to do such transformations.

After plotting the time series directly, we would like to decompose it into three main components.

* **Trend**: The solid curve is a LOWESS curve, while the dotted line is the linear trend candidate. Fitting a quadratic or even a cubic trend is possible, but nevertheless they seem to be overfitting in this scenario. In fact, a cubic trend might be accurate partially, but it really gets irresponsive starting from 1940. On the other hand, though not perfectly accurate, we believe a linear line is an adequate trend that basically represents the general tendency of coal production during this time period. To some extent, the coal production from 1920 to 1968 fluctuates, but it never deviates too much from this slightly increasing linear trend.
* **Seasonal component**: In this case we detect no seasonal component here. Even though the data shows sign of a drop in the 1930s (possibly due to the Great Depression which lasted from 1929 to 1939), then hits a peak in the 1940s (the World War II), then dives again in the 1950s. We cannot expect those rare events to happen every 10-20 years. In addition, the suspicious pattern hardly makes two whole cycles, thus we cannot confirm the existence of a seasonal component confidently.

****

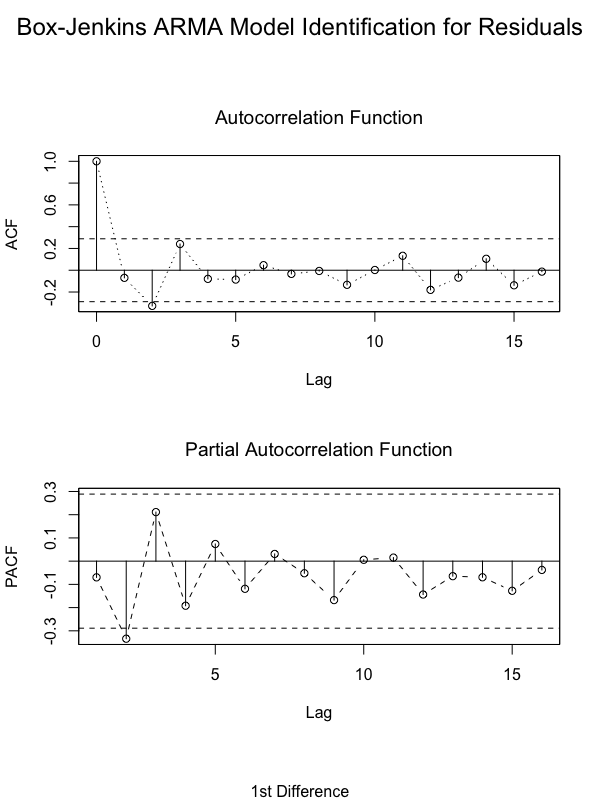
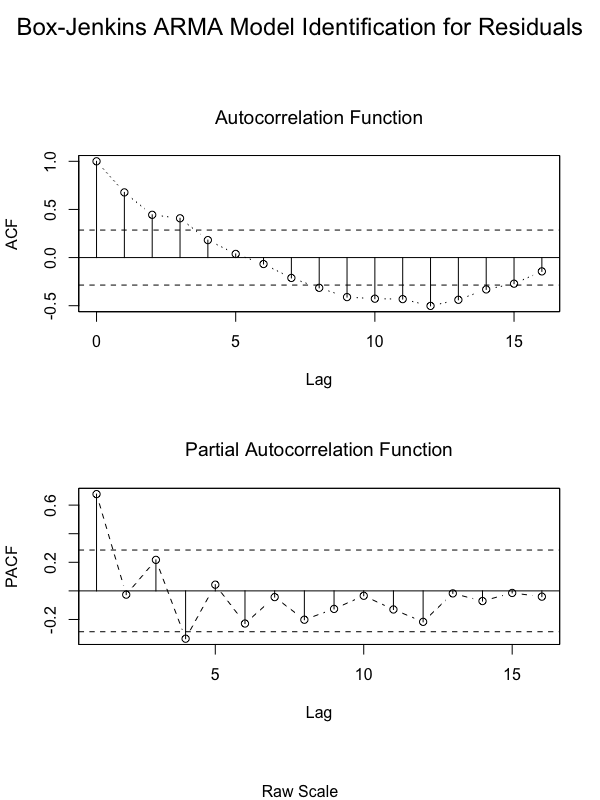
* **Irregular component**: After plotting the “detrended time series”, what remains should be some random noise. But the figure above still looks like the original data. This deja vu indicates that the residuals are dependent. We will try to fit the residuals with an AR model.

## 3. Fitting an AR Model

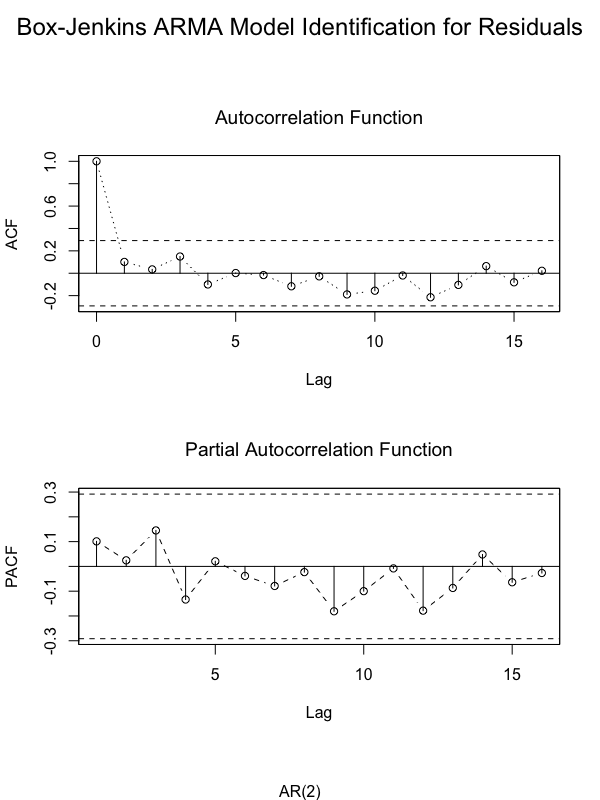
Firstly, we use Box-Jenkins method to test if the residual series is stationary. The Autocorrelation Function (ACF) is problematic, since it does not damp down to zero, as we can see the values go beyond the critical value from lag 8 to lag 14. This suggests that we might need to take the first differences, which agrees with our previous decision that a linear trend is selected. After doing so, we check the plot and corresponding ACF, PACF again.

Here we also checked the residual plots directly, it looks more like random noise than the previous non-differenced data. The plot is attached with other diagnostic plots later (See page 4).

This time, both ACF and PACF tail off. Specifically, the ACF damps down after lag 2, while the PACF has a spike at lag 2 then stays between two dotted lines after that. The number of lag 2 in PACF also determines the order of our candidate AR model. Hence, based on these two plots, we can deduce now the series is stationary, and an AR(2) model could be an appropriate fit for the irregular components.



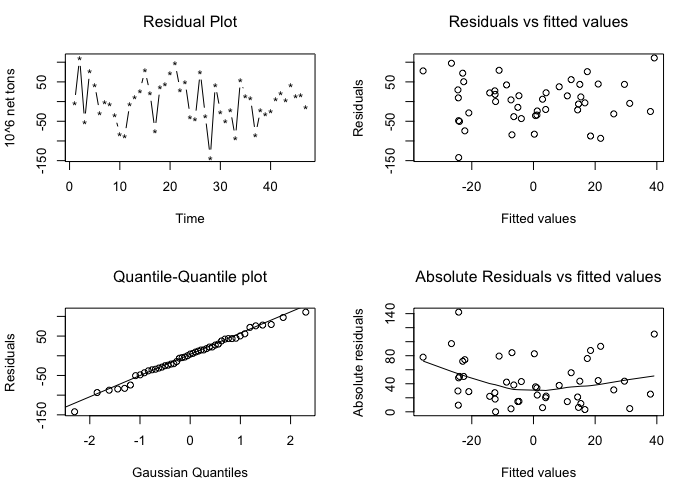
So after fitting an AR(2) model, we check the ACF and PACF for the third time. The plot is straightforward and confirms the legitimacy of AR(2) model.

In ACF, only lag 0 goes beyond the dotted lines. As in PACF, all spikes stay within the dotted lines.

Last but not least, we need to check the diagnostic plots of residuals:

* The residuals seem to have constant mean and variance.
* The residuals are normally distributed.
* The absolute residuals look like random noise.

So far, we believe this AR(2) model is adequate.



## 4. Improving the Model

Here we plan to discuss the possibility of improving our AR(2) model. Recall to the ACF and PACF plots after taking first difference. If we take an ARIMA model into account, the plots not only suggest an order of autoregressive terms but also the order of moving average terms , thus we have an ARIMA model candidate as ARIMA(2,1,2). But it is possible that an AR term and an MA term can cancel each other’s effects, so we would also try ARIMA(2,1,1) and ARIMA(1,1,2).

Due to page limit, we are not going to show the analytic comparison between those three models with our ARIMA(2,1,0) i.e. AR(2). But the idea is to produce the AIC and BIC values and find the model with the lowest values. If the values are really close, we pick the simpler model by the principle of parsimony.

## 5. Conclusion

Basically, we fit a linear trend and no seasonal component to the model and the residual of first difference follows an AR(2) model. The calculated coefficients of terms are presented above. Generally, this model gives us a relatively constant prediction in coal. However, man proposes, God disposes, further study shows that the reality was a skyrocketing increment in the 1970s. Hence, our model is not omnipotent based on a rather simple structure.