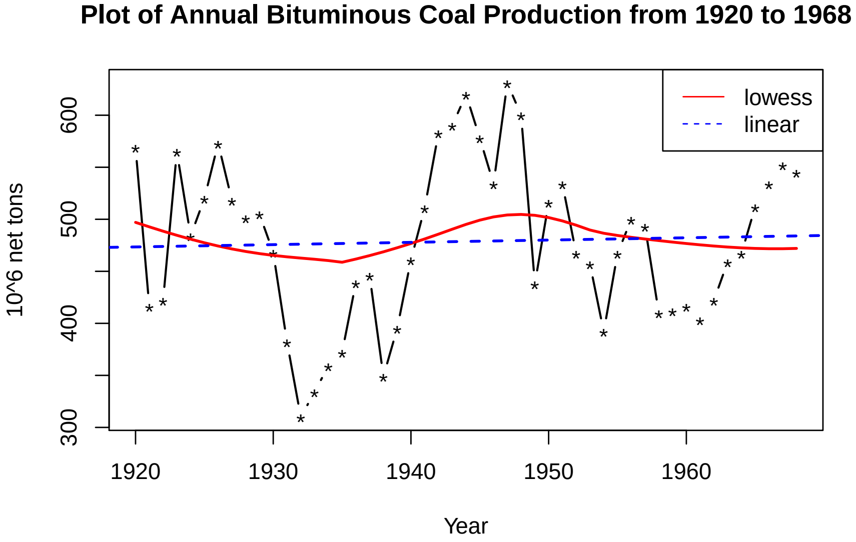
STAT7026 Assignment 2

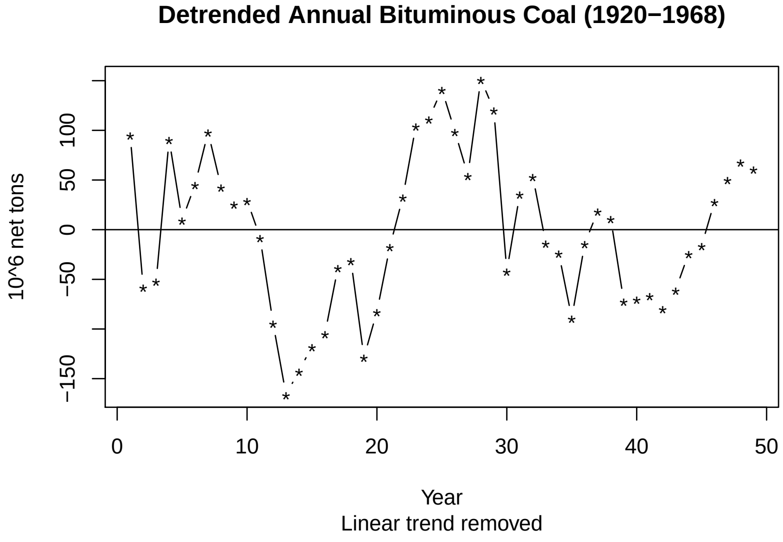
1. Decomposing

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). We tried a log-scale plot beforehand. However, it did not provide a better view of the data, so there is no need to do the log-transformation.



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

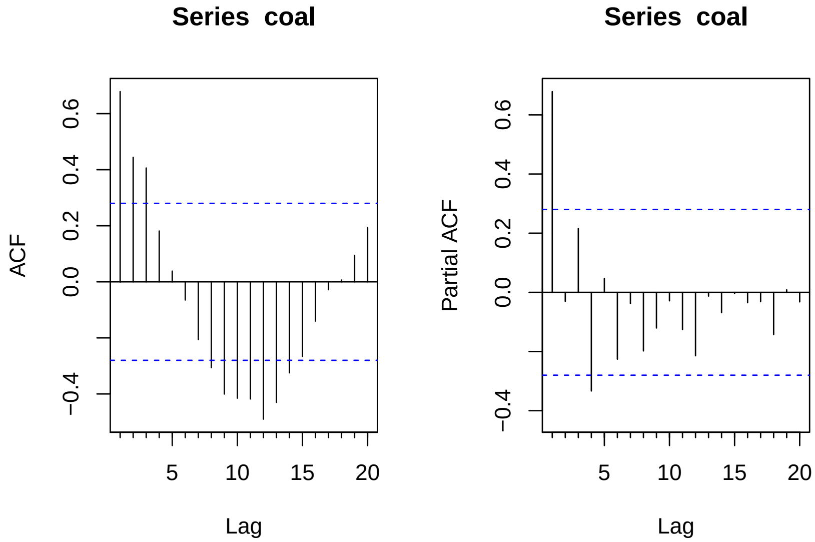
- **Trend**: The solid curve is a LOWESS curve, while the dotted line is the linear trend candidate. Though not perfectly accurate, we believe a linear line is an adequate trend that basically represents the of coal production during this time period. To some extent, the coal production from 1920 to 1968 fluctuates, but it never deviates too much from this linear trend. On the other hand, a quadratic or even a cubic trend is possible, but nevertheless they seem to be overfitting in this scenario.
- **Seasonal component**: In this case we detect no seasonal component here. Even though the data seems to have a drop-down 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 drops again in the 1950s. Notwithstanding that this back-and-forth changes look like a pattern, those data hardly make two whole cycles, thus we cannot confirm the existence of a seasonality confidently.



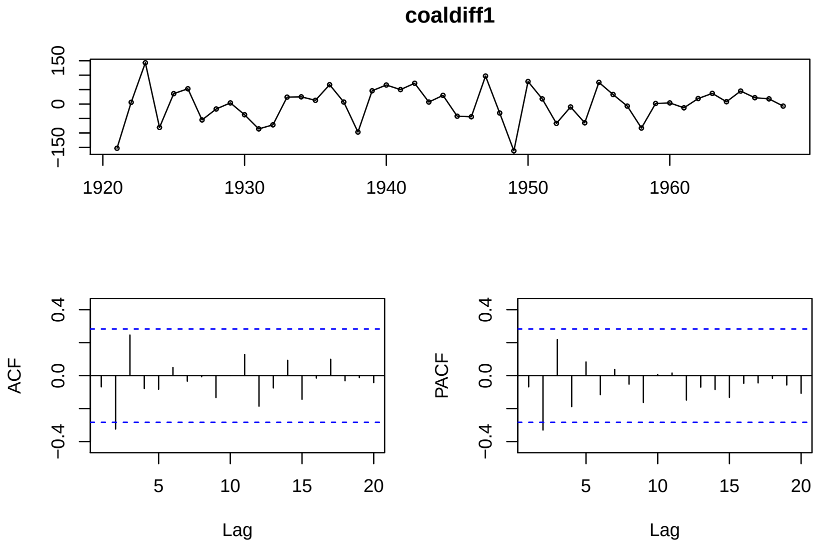
- **Irregular component**: After plotting the "detrended time series", what remains should be some random noise. But the figure above still looks like the original data. This deja vu indicates that the residuals are dependent.

2. Fitting an ARIMA Model

So far, we are a little bit worried that the coal data is not suitable to fit an ARMA model directly, as it's not stationary. To test it is really not stationary, we use the plot the Autocorrelation Function (ACF) and the partial Autocorrelation Function (PACF) for it.



The ACF does not cut off to the dotted region, 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.



This time both ACF and PACF tail off exponentially, the only two lags out of the region are lag 2 in ACf and lag 2 in PACF. Therefore, the order of autoregression terms $q = 2$ and the order of moving average terms $p = 2$, we have an ARIMA model candidate as ARIMA(2,1,2).

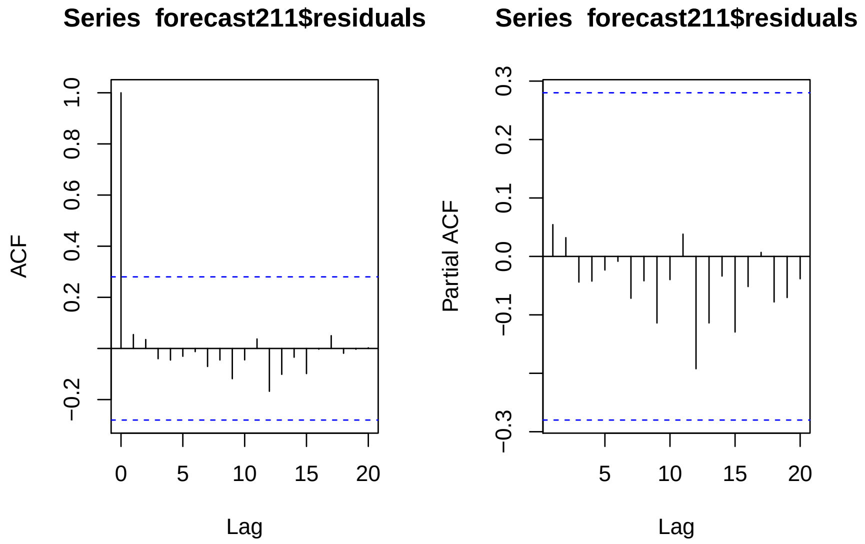
But it is possible that an AR term and an MA term can cancel each other's effects, so we would also try ARIMA(2,1,1) and ARIMA(1,1,2).

The AIC and BIC values returned by ARIMA table for all three candidates indicates that the ARIMA(2,1,1) model has the lowest AIC and BIC values, although the difference is not tremendous. On the other hand, it is a simpler model than our first candidate ARIMA(2,1,2). By the principle of parsimony, we pick ARIMA(2,1,1) as the model to fit our time series.

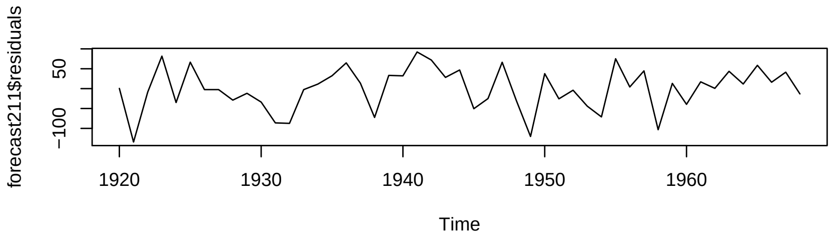
```
## Series: coal
## ARIMA(2,1,1)
##
## Coefficients:
##           ar1      ar2     ma1
##       -0.6648  -0.4336  0.6714
## s.e.   0.1947   0.1439  0.1730
##
## sigma^2 estimated as 3051:  log likelihood=-259.46
## AIC=526.91   AICc=527.84   BIC=534.4
```

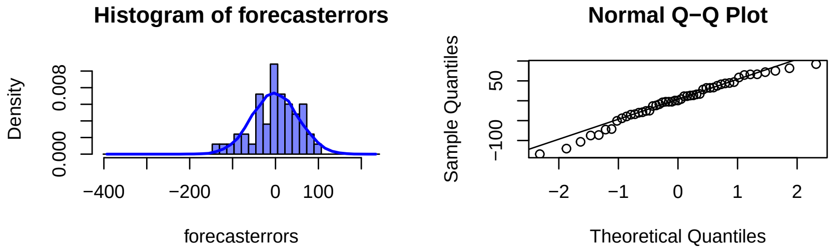
$$\begin{aligned}
\Delta Y &= Y_t - Y_{t-1} \\
\Delta Y &= \phi_2 Y_{t-2} + \phi_1 Y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t \\
&= -0.6648 Y_{t-2} - 0.4336 Y_{t-1} + 0.6714 \epsilon_{t-1} + \epsilon_t \text{ where } \epsilon_t, \epsilon_{t-1} \text{ are white noise error terms.}
\end{aligned}$$

3. Forecast



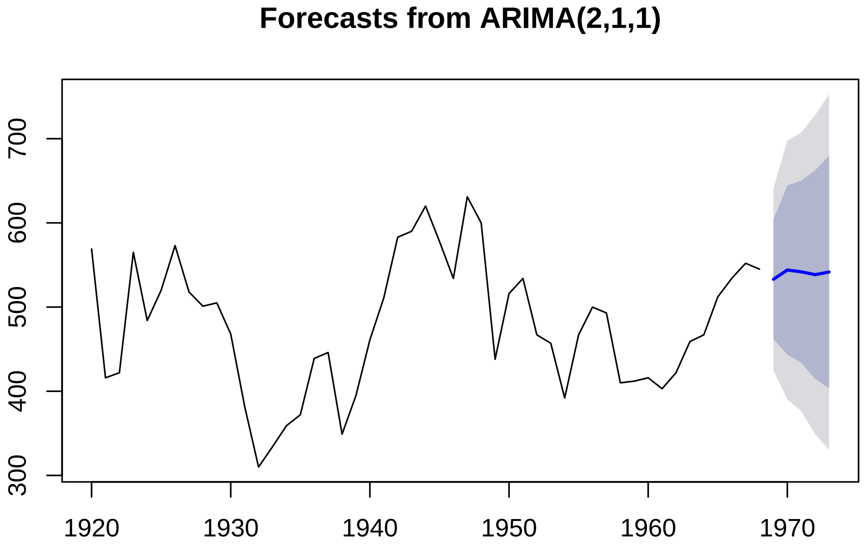
Now we are interested in if the forecast made by this model is plausible. First we can conduct a Ljung-Box test to see if there is little evidence for non-zero autocorrelations in the forecast errors. The returned p-value is $0.9993 > 0.05$, so the forecast errors are not correlated. Meanwhile, the ACF and PACF of forecast residuals both drop into the region as expected.





Furthermore, we investigate whether the forecast errors are normally distributed with mean zero and constant variance. The line plot shows that the variance of the forecast errors seems to be roughly constant over time. The histogram, together with normal Q-Q plot, shows that the forecast errors are roughly normally distributed and the mean seems to be close to zero. Therefore, it is plausible that the forecast errors are normally distributed with mean zero and constant variance.

Since successive forecast errors do not seem to be correlated, and the forecast errors seem to be normally distributed with mean zero and constant variance, the ARIMA(2,1,1) does seem to provide an adequate predictive model for the annual coal production.



Basically, we predict that the production would stay constant in the next 5 year.

4. References

- Using R for Time Series Analysis, http://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html
- Occam's razor, https://en.wikipedia.org/wiki/Occam%27s_razor
- Forecasting: principles and practice, https://www.otexts.org/fpp/
- Using AIC to Test ARIMA Models, https://coolstatsblog.com/2013/08/14/using-aic-to-test-arima-models-2/
- Terms "cut off" and "tail off" about ACF, PACF functions,https://stats.stackexchange.com/questions/241914/terms-cut-off-and-tail-off-about-acf-pacf-functions
- Summary of rules for identifying ARIMA models, https://people.duke.edu/~rnau/arimrule.htm