Time Series: Assignment 1

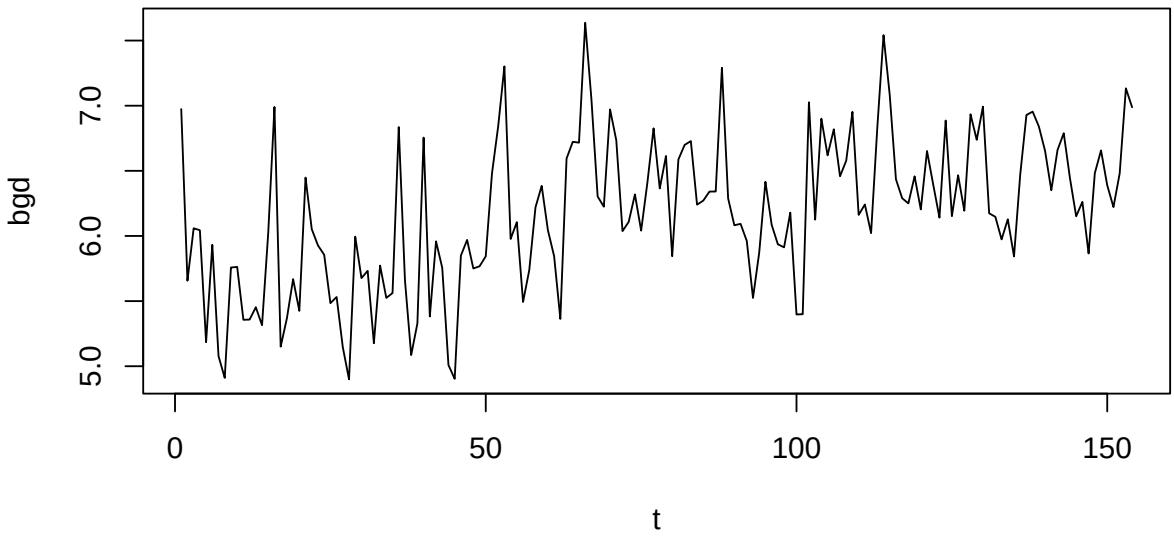
Xiao Heng (s2032451)

1. Firstly we read the two .txt files, and transfer the data as `ts()` form.

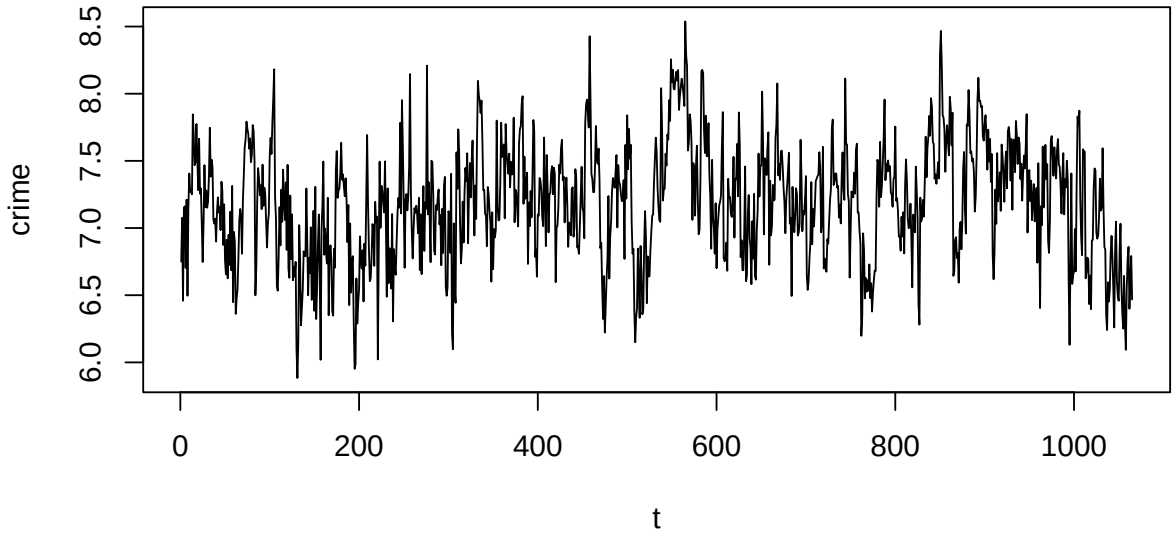
```
# input, header=FALSE, sep=""
bgd <-read.table("bgd.txt")
crime <- read.table("crime.txt")
# transfer into time series
bgd <- ts(bgd[,2])
crime <- ts(crime[,2])
```

Then plot the each time series.

```
# plot
plot(bgd, xlab="t")
```




```
plot(crime, xlab="t")
```



Notice that the mean of data `crime` is larger than the mean of data `bgd`. While for the data pattern, we cannot easily say what model they are following, since it seems noisy, so that we need further analysis in the following parts.

2. Following the Ljung-Box statistic formula:

$$Q_m = n(n+2)\sum_{k=1}^{m}[\frac{r_k^2}{n-k}],$$

and the corresponding hypotheses are:

H_0 : The data are independently distributed (no serial correlation);

H_1 : The data are not independently (serial correlation exists). distributed.

In this case, we could implement the Ljung-Box test on each time series respectively.

```
# Ljung-Box test
Box.test(bgd, type="Ljung-Box", lag=5)
```

```
##
##  Box-Ljung test
##
## data:  bgd
## X-squared = 99.702, df = 5, p-value < 2.2e-16
```

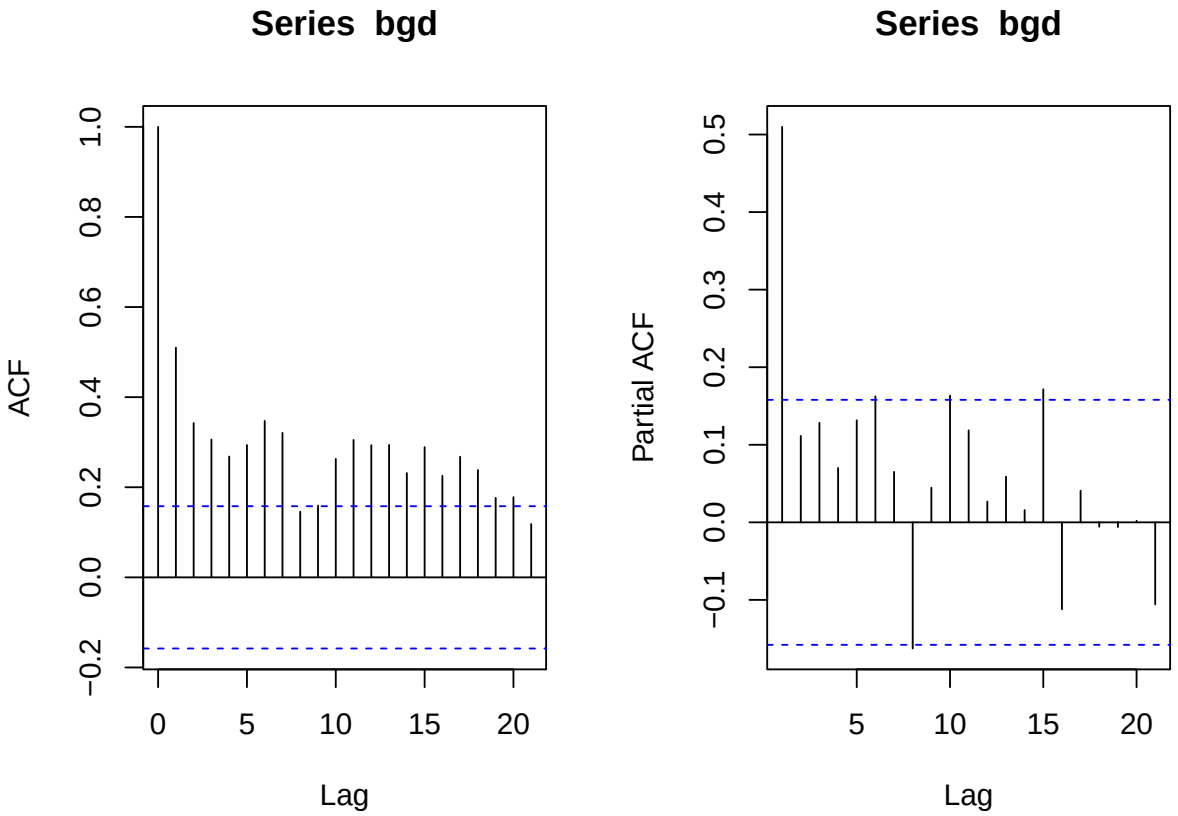
```
Box.test(crime, type="Ljung-Box", lag=5)
```

```
##
##  Box-Ljung test
##
## data:  crime
## X-squared = 1431.6, df = 5, p-value < 2.2e-16
```

As both tests' `p-value<0.05`, null hypothesis is rejected. We could say that both the data `bgd` and `crime` exhibit serial correlation, and there is potential relationship among the series, so that further modelling is necessary.

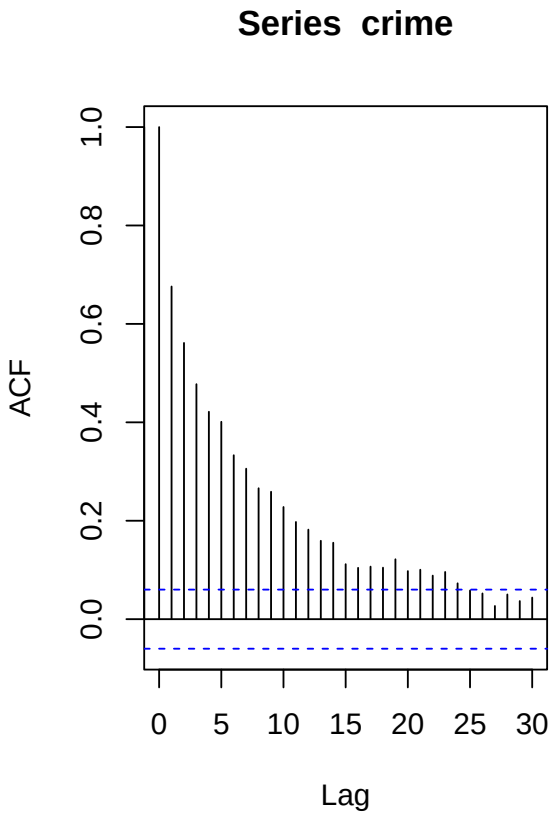
3. To choose the models and corresponding orders (and coefficients), we draw the correlograms for each dataset.

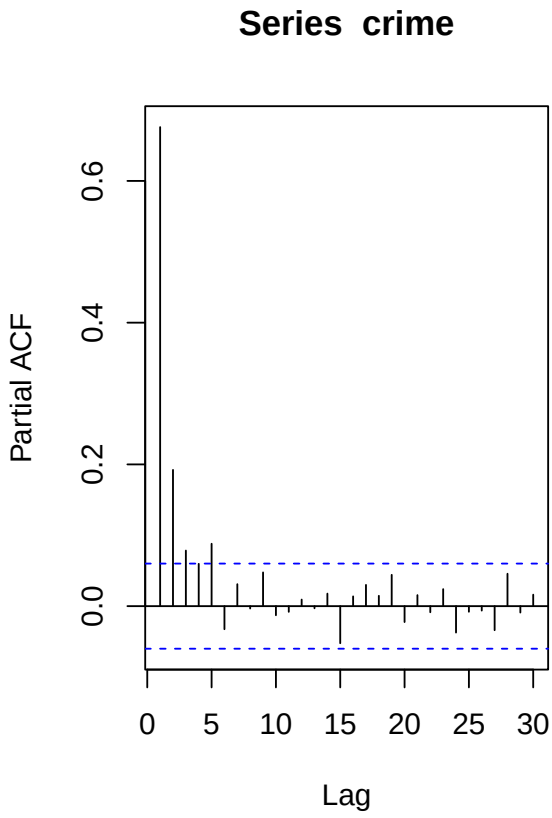
```
# plot acf and pacf for dataset bgd
par(mfrow=c(1,2))
acf(bgd)
pacf(bgd)
```



For `bgd`, its ACF trend gradually decreases while PACF shows a sharp drop after lag 1, with a value around 0.5. So we assume it follows an **AR(1)** process with $\alpha_1 \approx 0.5$.

```
# plot acf and pacf for dataset crime
par(mfrow=c(1,2))
acf(crime)
pacf(crime)
```





For `crime`, its ACF trend also gradually decreases while PACF shows a sharp drop after lag 2, with a value around 0.7 for lag 1 and 0.2 for lag 2. So we assume it follows an **AR(2)** process with $\alpha_1 \approx 0.7$ and $\alpha_2 \approx 0.2$.

4. To fit the two datasets with AR(2):

$$X_t - \mu = \alpha_1(X_{t-1} - \mu) + \alpha_2(X_{t-2} - \mu) + \omega_t,$$

where $\{\omega_t\} \sim (0, \sigma^2)$.

In R, we apply `arima()` to implement.

```
bgd_ar2 <- arima(x=bgd, order=c(2, 0, 0))
bgd_ar2
```

```
##
## Call:
## arima(x = bgd, order = c(2, 0, 0))
##
## Coefficients:
##          ar1     ar2  intercept
##       0.4611  0.1136     6.1873
## s.e.  0.0810  0.0819     0.0918
##
## sigma^2 estimated as 0.2392:  log likelihood = -108.55,  aic = 225.09
```

```
crime_ar2 <- arima(x=crime, order=c(2, 0, 0))
crime_ar2
```

```
##
## Call:
## arima(x = crime, order = c(2, 0, 0))
##
## Coefficients:
##          ar1     ar2  intercept
##       0.5475  0.1923     7.1822
## s.e.  0.0301  0.0301     0.0368
##
## sigma^2 estimated as 0.09852:  log likelihood = -277.44,  aic = 562.88
```

```
bgd_coef <- bgd_ar2$coef[1:3]
names(bgd_coef) <- c("a1","a2","mu")
bgd_coef
```

```
##        a1        a2        mu
## 0.4610688 0.1135770 6.1873383
```

```
crime_coef <- crime_ar2$coef[1:3]
names(crime_coef) <- c("a1","a2","mu")
crime_coef
```

```
##        a1        a2        mu
## 0.5474734 0.1923070 7.1821675
```

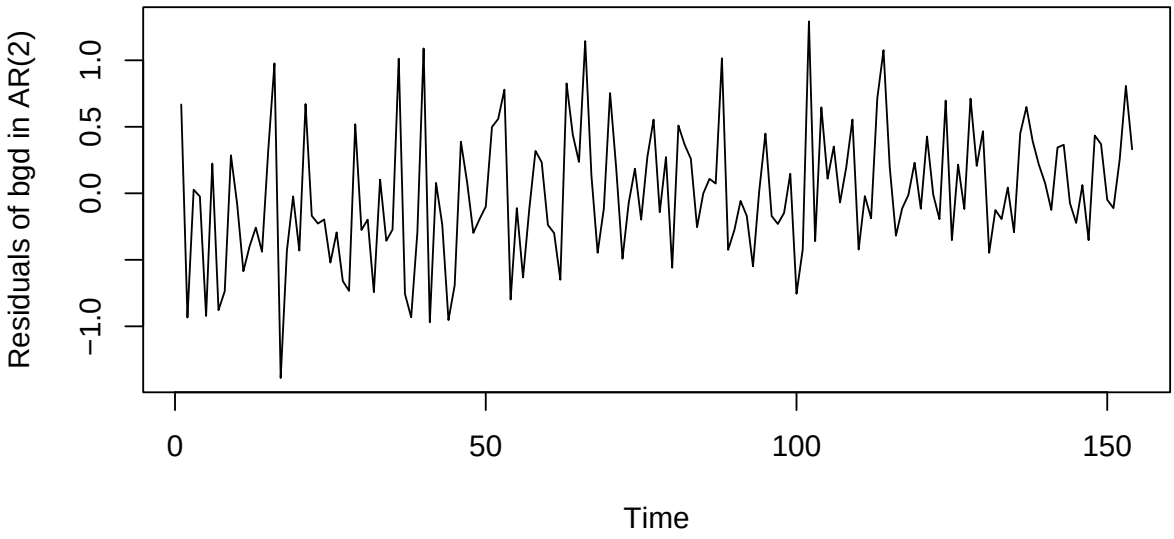
5. From the result of last part, we could concludes some features:

Differences: 1. the means of two datasets are significantly different, as the μ of crime dataset is apparently higher than the μ of dataset `bgd`; 2. both the coefficients of α_1 and α_2 of dataset `crime` is larger, indicating a higher serial correlation within dataset `crime`.

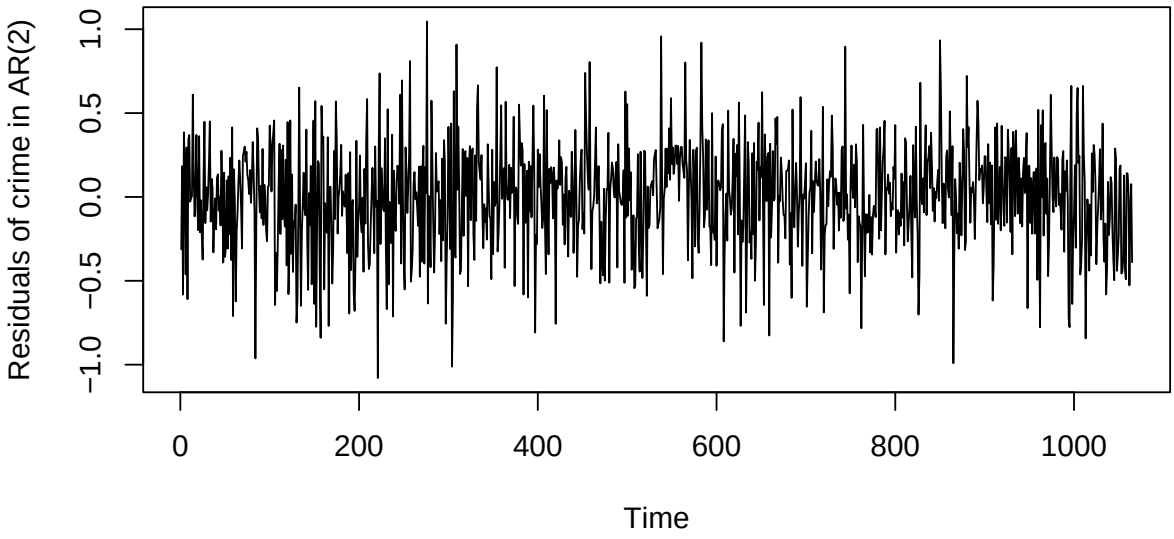
Similarities: Both of the dataset exhibit serial correlation (although different degree), which can be approximated by model like AR(2), implying a high relationship for one banknote to the one next to it (neighbourhood), and a slightly lower relationship to the lag 2 neighbourhood (seperated by the nearest banknote). As for further remote banknotes, the cocaine contaminations are almost unrelated.

6. As introduced in the lecture, for a model which fits well, the residuals $\{w_t\}$ will be approximately white noise, with constant variance. So we can assess the fit of the model by checking the residuals with 5 methods here (first 3 methods are from lecture slides, 4th is from R supplementary material, 5th is recommended by the assignment)

```
# 1. Plotting the residuals
bgd_res <- bgd_ar2$residuals
crime_res <- crime_ar2$residuals
plot(bgd_res, ylab="Residuals of bgd in AR(2)")
```



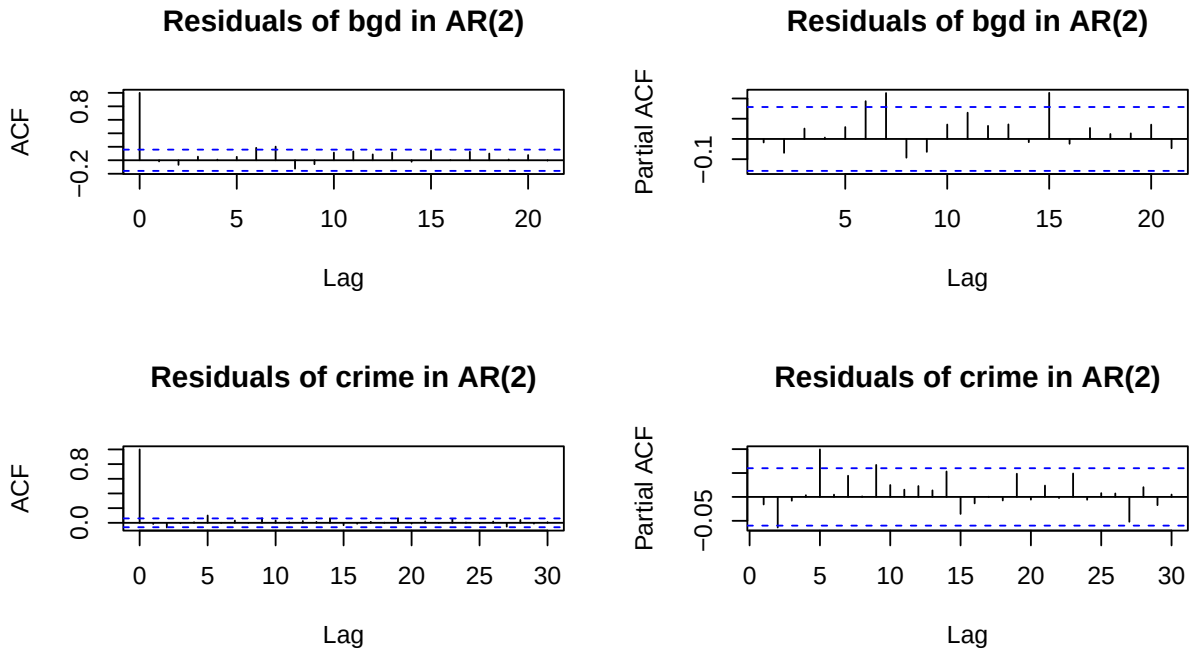

```
plot(crime_res, ylab="Residuals of crime in AR(2)")
```



In the first method, we directly plot the residuals. Visually, it could be viewed as white noise, with stationary zero mean and constant variance. But more precise methods are still needed.

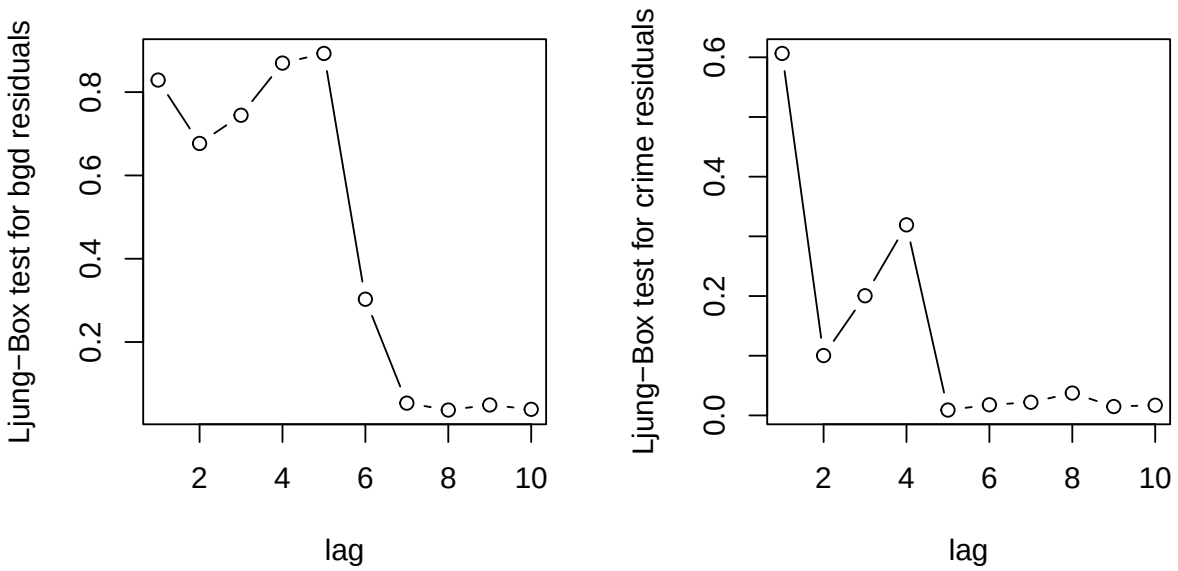
```
# 2. Obtaining a correlogram of the residuals
par(mfrow=c(2,2))
acf(bgd_res, main="Residuals of bgd in AR(2)")
pacf(bgd_res, main="Residuals of bgd in AR(2)")
```

```
acf(crime_res, main="Residuals of crime in AR(2)")
pacf(crime_res, main="Residuals of crime in AR(2)")
```



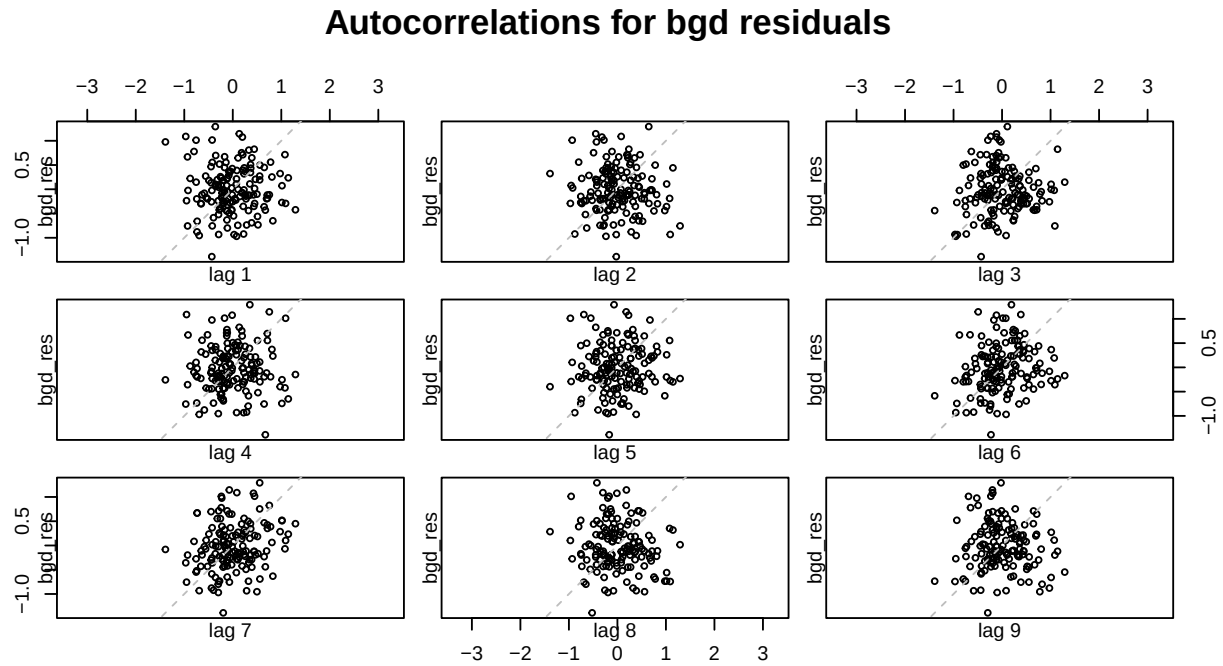
In the 2nd method, from the correlograms, although roughly most values are within the blue dotted lines (indicating not significantly different from 0), in some specific lags (like lag 6 and lag 7 for residuals of `bgd` and lag 5 for residuals of `crime`), the correlation is actually significant but in weak degree. For lag 0, both acf values are 1, because the data is completely related with itself.

```
# 3. Examining the Ljung-Box statistic of the residuals
bgd_p <- sapply(1:10, function(lag) Box.test(bgd_res, type="Ljung-Box", lag)$p.value)
crime_p <- sapply(1:10, function(lag) Box.test(crime_res, type="Ljung-Box", lag)$p.value)
par(mfrow=c(1,2))
plot(bgd_p,type="b", xlab="lag", ylab="Ljung-Box test for bgd residuals")
plot(crime_p,type="b", xlab="lag", ylab="Ljung-Box test for crime residuals")
```

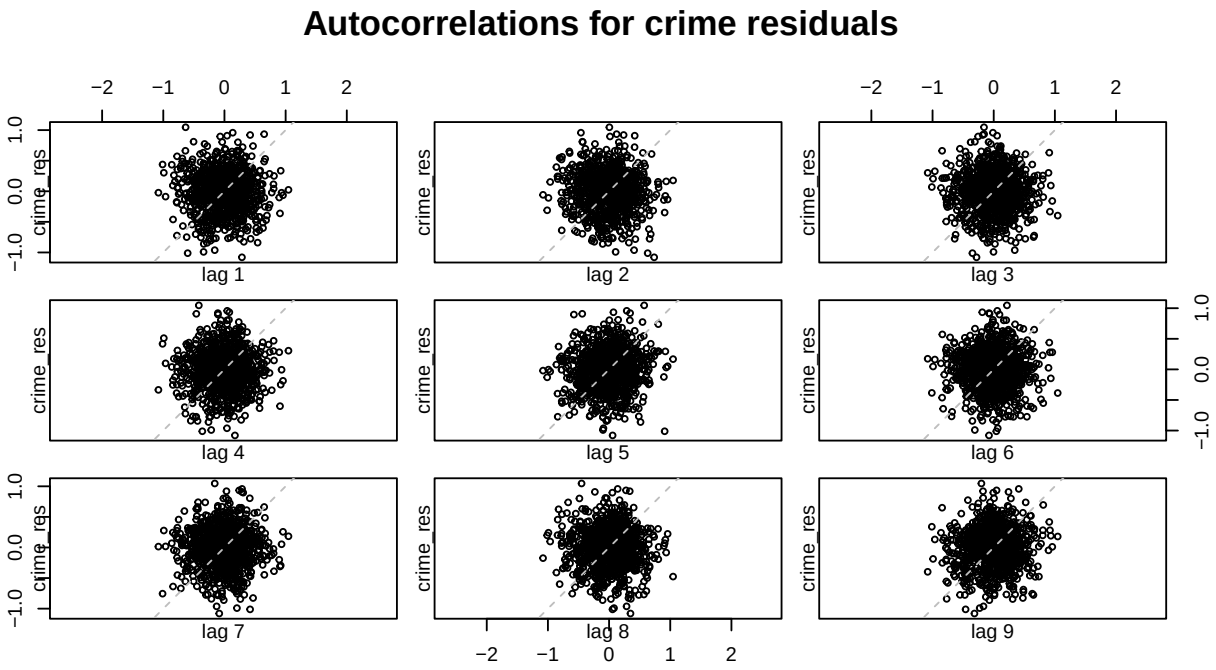


In 3rd method, Ljung-Box test tells us that the residuals of `bgd` under AR(2) might not be white noise within the first several lags, so fitting model for `bgd` may need further modification (As we concludes from the ACF and PACF of `bgd`, it tends to be more likely a AR(1) pattern). While for the result of residuals of `crime`, the AR(2) fits almost good enough, alhough the statistic is a little significant at lag 1 at a weak degree.

```
# 4. lag plot
lag.plot(bgd_res, type="p", lag=9, do.lines=FALSE, main="Autocorrelations for bgd residuals")
```

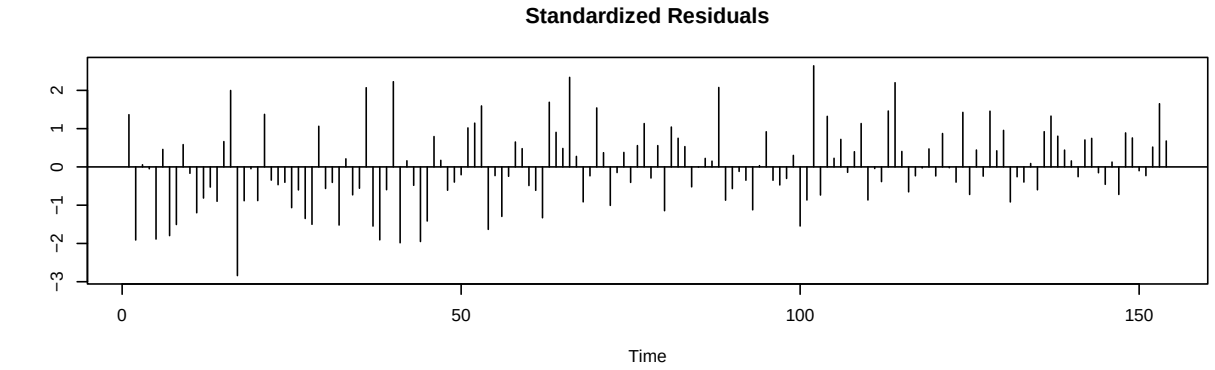


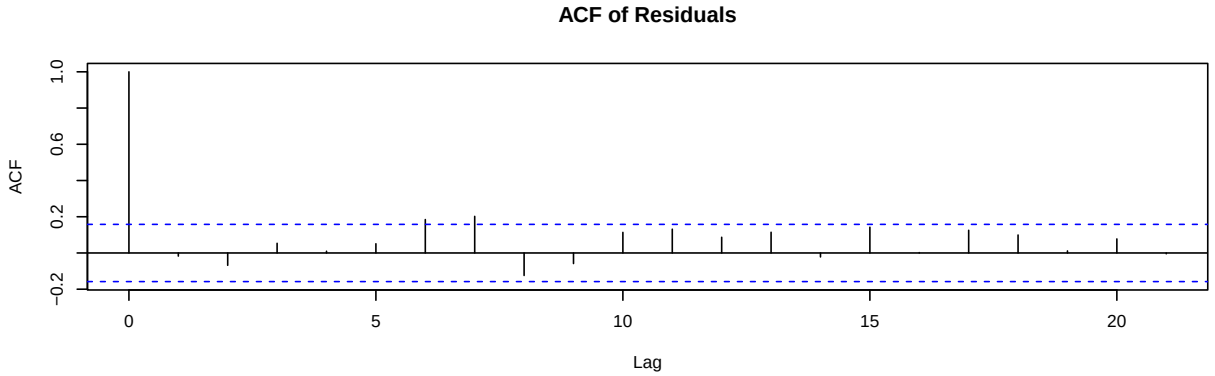
```
lag.plot(crime_res, type="p", lag=9, do.lines=FALSE, main="Autocorrelations for crime residuals")
```

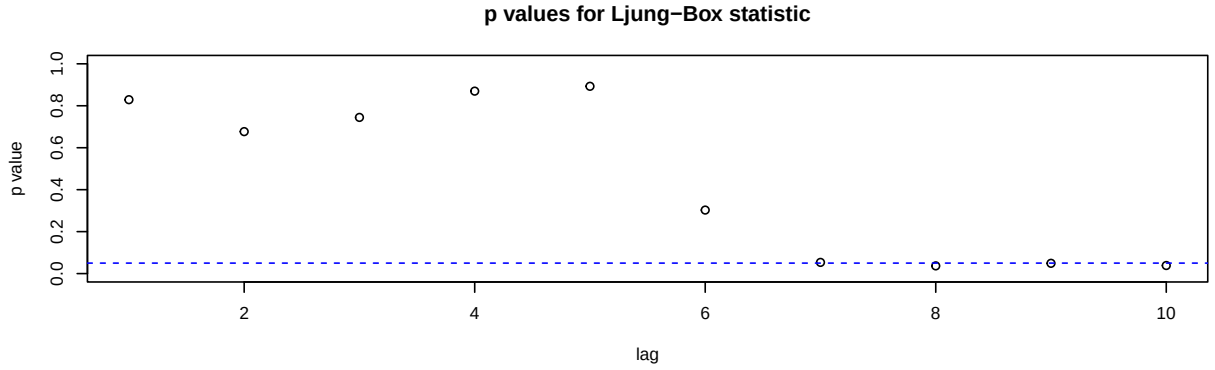


For the 4th method, almost all lags of residuals are showing a white noise pattern (uniformly distributed circle). More precise result has been obtained in previous methods.

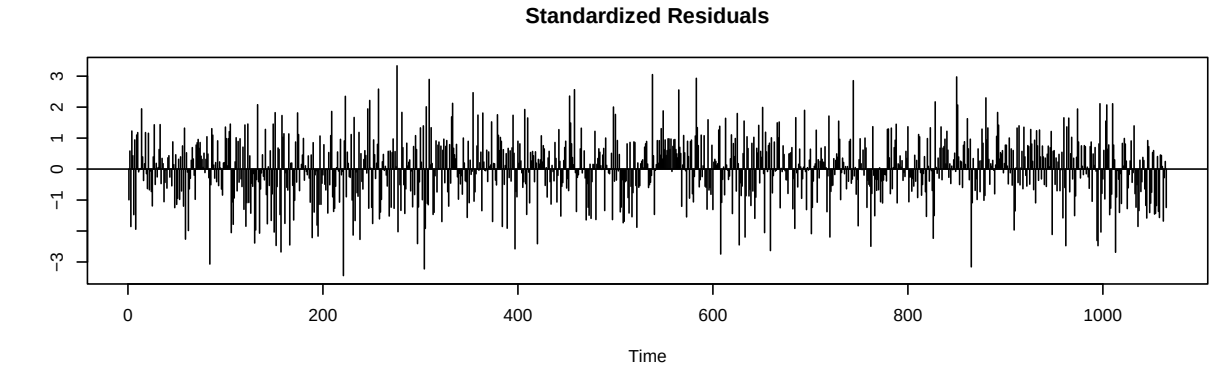
```
# 5. tsdiag
tsdiag(bgd_ar2)
```

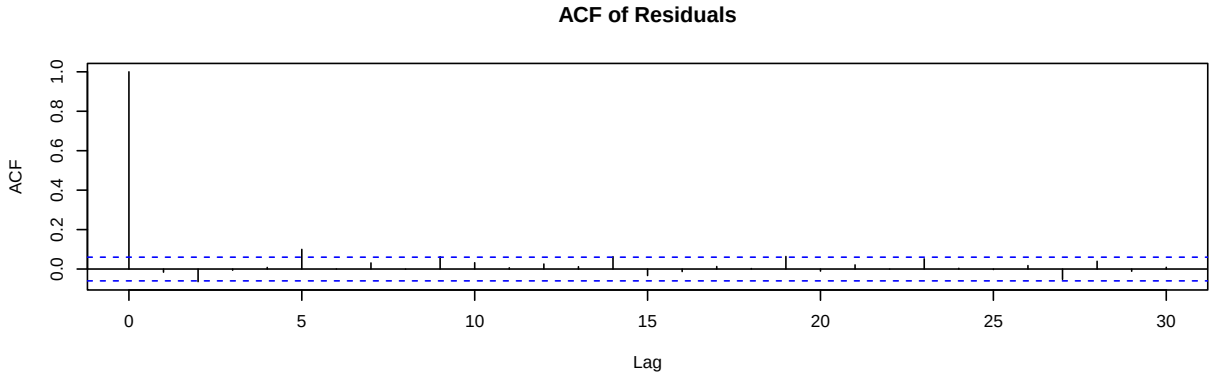


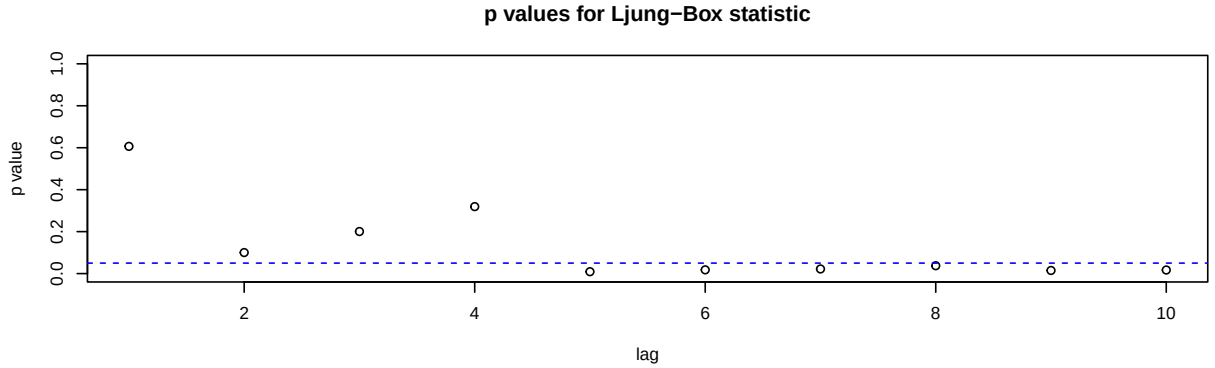





```
tsdiag(crime_ar2)
```







The final method directly adopts the fitted model and shows us several deagnostic graphes, which is really convenient. Apart from the upper plot of standardized residuals (which is also a visually result as before), the ACF and Ljung-Box test we've also discussed in previous methods.

Overall, the AR(2) fits well under `crime`, and the residuals can almost regarded as white noise; while the `bgd` may need further check because the fitting residuals are not random enough, so that AR(2) shows a weaker performance on this dataset.