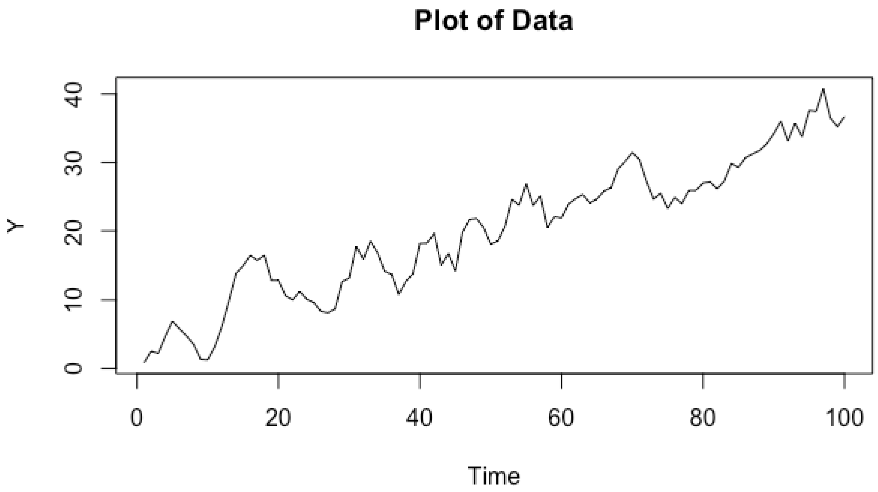
Question 2



After plotting the data, we notice the upward trend over time, which confirms that including time as an explanatory variable is reasonable.
So, the model is:

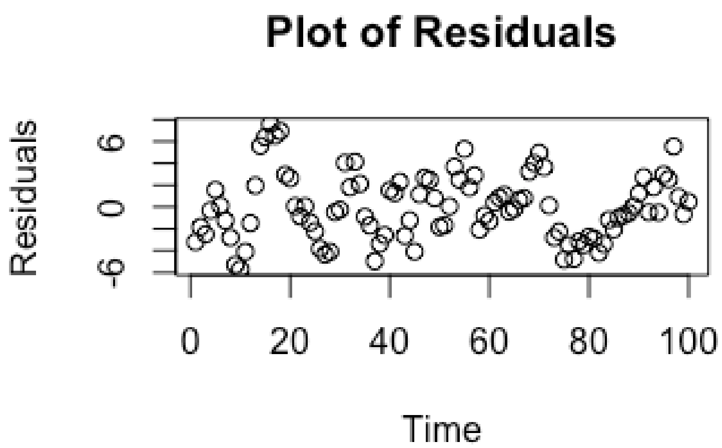
$$Y_t = \beta_0 + \beta_1 t + \epsilon_t$$

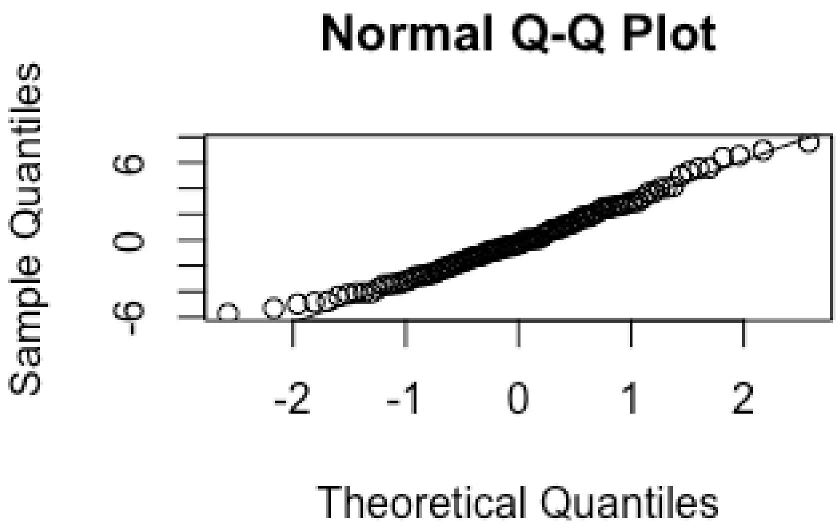
where $t = 1, \ldots, 100$

```
Residual Standard Error=3.0228
R-Square=0.9077
F-statistic (df=1, 98)=963.8667
p-value=0

            Estimate Std.Err t-value Pr(>|t|)
Intercept     3.7168  0.6091  6.1018        0
X             0.3251  0.0105 31.0462        0
```

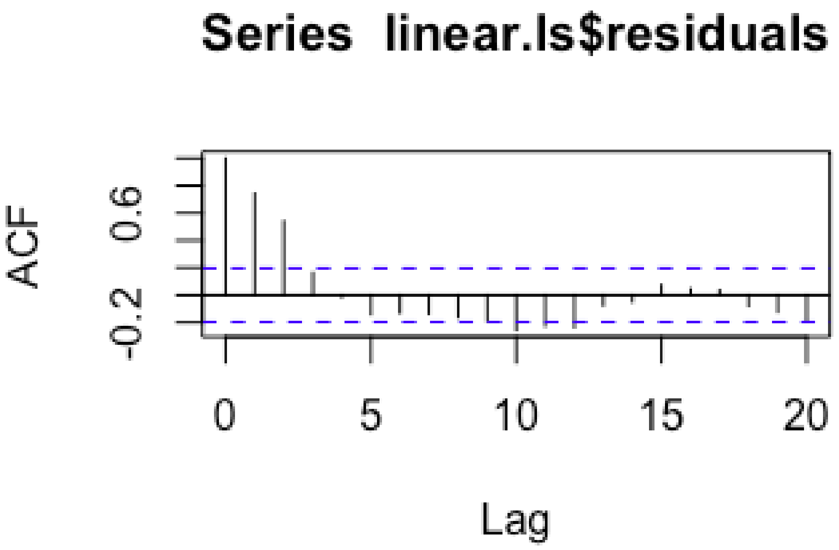
After fitting the linear model, we obtain $\widehat{\beta_0}$ = 3.7168 and $\widehat{\beta_1}$= 0.3251

Residual Analysis:

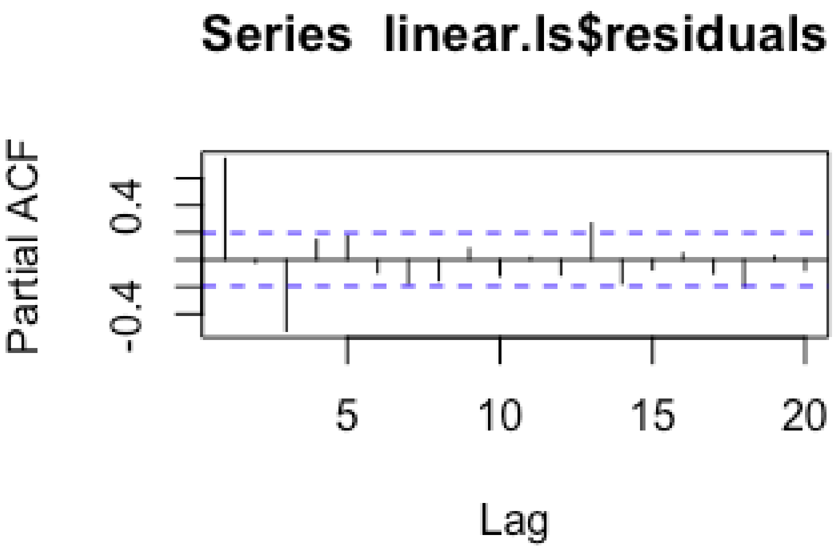




Normal Q-Q Plot
Sample Quantiles
-6
0
6
-2
-1
0
1
2
Theoretical Quantiles



Series linear.ls$residuals
ACF
-0.2
0.6
0
5
10
15
20
Lag



Series linear.ls$residuals
Partial ACF
-0.4
0.4
5
10
15
20
Lag

- The plot of residuals shows no obvious deviation from stationarity i.e. no changing mean or variance.
- From the Q-Q plot we can assume that residuals fit pretty well although there are deviations on the tails, which is normal.
- The SACF shows pattern consistent with AR (2) or AR (3) model (exponential decay).
- The SPACF shows a clear lag 1, 3 and 13 correlations, which is acceptable since there is 5% rejection region, while the partial correlations at the remaining lags are insignificant.

Our analysis suggests that $Y_t = \beta_0 + \beta_1 t + \epsilon_t$ where $\epsilon_t \sim AR(2 \; or \; 3)$

Or

$Y_t = \beta_0 + \beta_1 t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + Z_t$ where $Z_t \sim WN$ for AR (2)

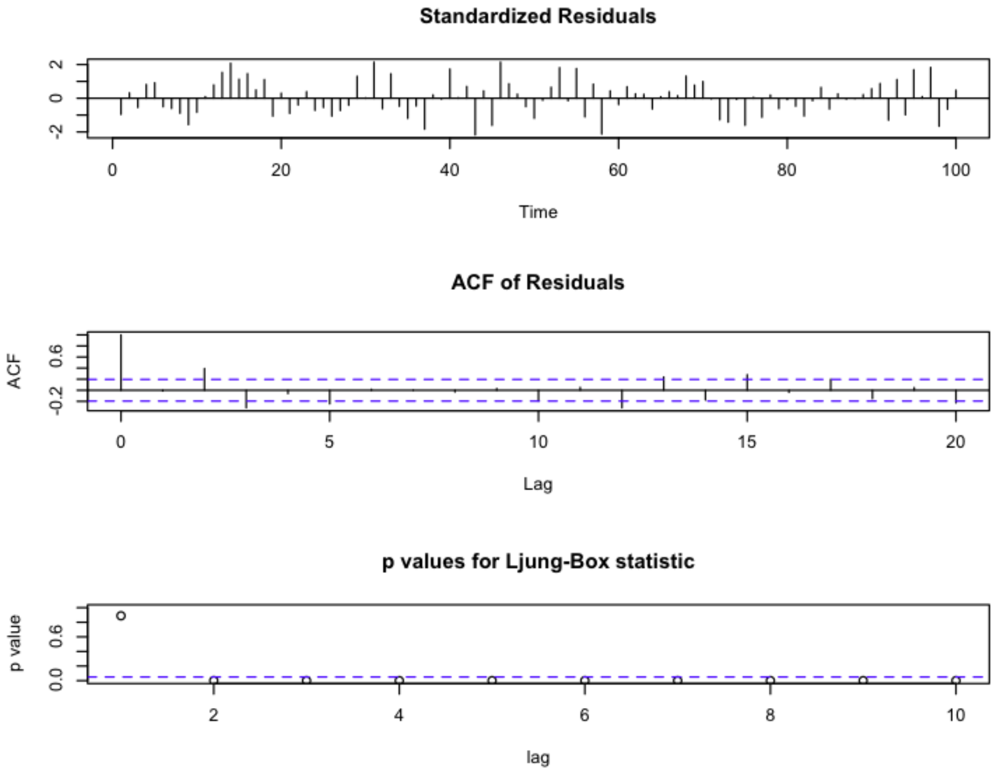
```
Call:
arima(x = linear.y, order = c(2, 0, 0), xreg = t, method = "ML")

Coefficients:
         ar1      ar2  intercept       t
      0.7630  -0.0252     3.3631  0.3307
s.e.  0.0997   0.0997     1.4371  0.0244

sigma^2 estimated as 3.956:  log likelihood = -211.06,  aic = 432.13
```

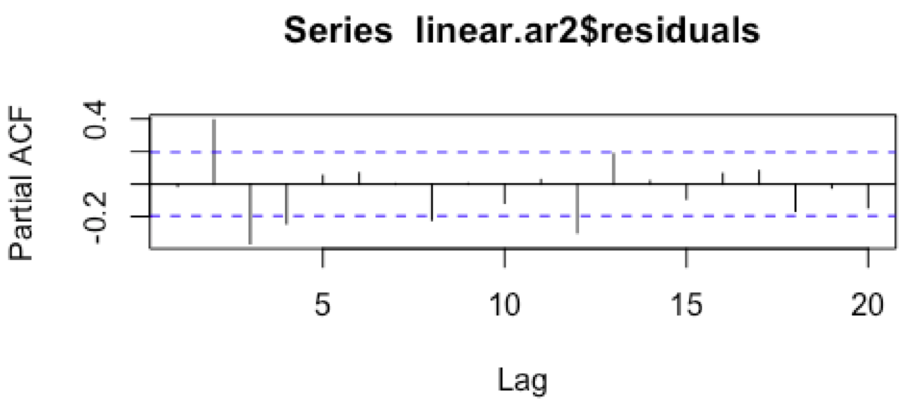
$Y_t = \beta_0 + \beta_1 t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \phi_3 \epsilon_{t-3} + Z_t$ where $Z_t \sim WN$ for AR (3)

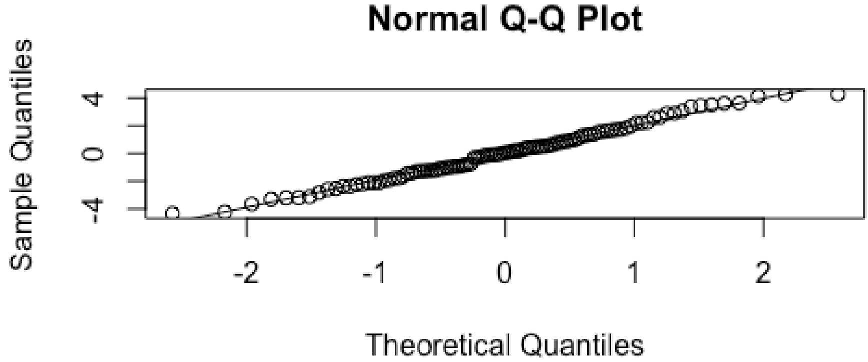
```
Call:
arima(x = linear.y, order = c(3, 0, 0), xreg = t, method = "ML")

Coefficients:
         ar1     ar2     ar3  intercept       t
      0.7450  0.3768  -0.518     3.6865  0.3255
s.e.  0.0839  0.1065   0.084     0.8646  0.0149

sigma^2 estimated as 2.86:  log likelihood = -195.3,  aic = 402.59
```

- The estimated variance for AR (2) model of WN process is 3.956 which is higher than the variance of the original residuals 3.0228.
- The estimated variance for AR (3) model of WN process is 2.86 which is lower than the variance of the original residuals 3.0228.
- Also, AIC for AR (2) is 432.12 and is much higher than AIC for AR (3) 402.59.
- When we compare AR (3) and AR (4) models we see that AIC is almost the same.
- So, we can conclude that AR (3) is a better model based on AIC statistic and rule of parsimony.

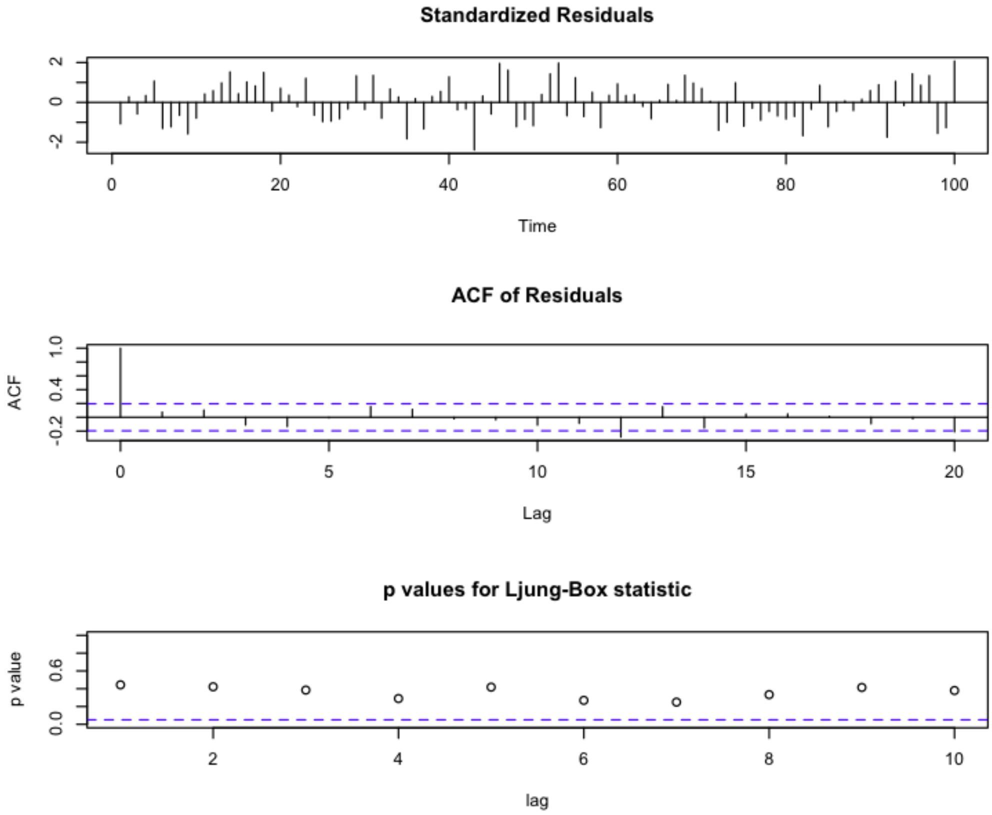
Now we conduct WN analysis for AR (2):

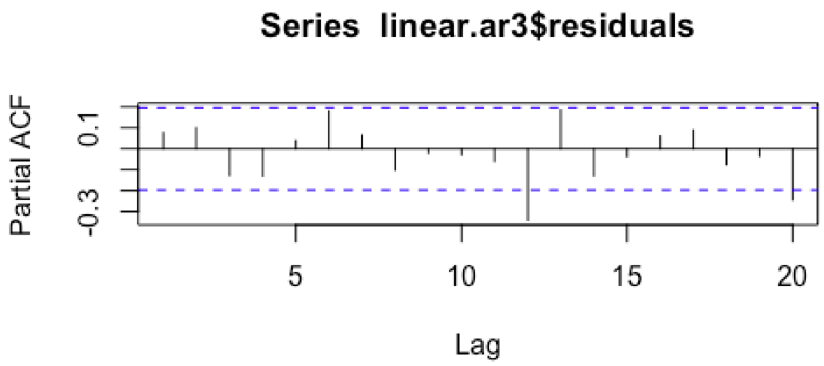


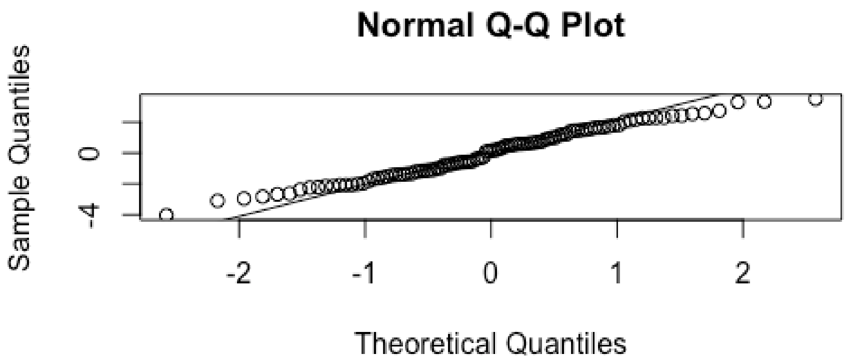




And we conduct WN analysis for AR (3):







- *For AR (2):*
- We see no obvious patterns of violation of stationarity and symmetry from the first plot.
- From ACF plot there are spikes at lags 2, 3, 5, 12, 13, 15 which suggests that there is a source of correlation that our model couldn't capture. So, we have evidence against WN.
- From third plot, we see that from lag 2 onwards all dots are below blue dashed line, which suggests that we should reject H_0: all correlations are zero.
- We notice spikes at multiple lags in the PACF plot too. From plot 5 we see normal plot fits well for AR (2) Model. However, so many spikes in ACF and PACF plots and rejection of H_0 in Ljung-Box test suggests that the process Z_t is not WN and AR (2) is not a good model.
- *For AR (3):*
- We see no obvious patterns of violation of stationarity and symmetry from the first plot.
- From ACF plot there is a spike at lag 12 which suggests that there may be a source of correlation that our model couldn't capture.
- From third plot, we see that all dots are above blue dashed line, which suggests that there is no evidence to reject H_0: all correlations are zero. So, we have no evidence against WN.
- We notice spike at lag 12 in the PACF plot too, however, behavior of Ljung-Box statistic suggests that spike at lag 12 is not significant.
- From plot 5 we see normal plot fits well for AR (3) Model.
- Hence, there are no evidences against normality or WN of Z_t. Moreover, we can use AR (3) model to test regression parameters and to construct CI and PI.

  Forecasting and PI:

  Forecasted values for Y_{101}, …, Y_{105} correspondingly:
  36.17514
  37.12937
  37.00516
  37.67401
  37.76018
  PI's for Y_{101}, …, Y_{105} correspondingly:
  (32.79310, 39.55717)
  (32.91197, 41.34677)
  (31.74043, 42.26990)
  (32.18717, 43.16084)
  (32.17685, 43.34351)

R code

```
## Question 2
mypath = "/Users/safurasuleymanovs/Desktop/4B/Stat/A4/"
data.set ="linear_y.txt"
linear.y = scan(paste(mypath,data.set,sep=""))
linear.ts = ts(scan(paste(mypath,data.set,sep="")))
plot(linear.ts, type="l",xlab="Time",ylab="Y", main="Plot of Data")
## After plotting the data we notice the upward trend which confirms
## that including time as an explanatory variable is reasonable
t <- c(1: length(linear.ts))
linear.ls <- lsfit(t,linear.ts)
ls.print(linear.ls)
## Residual Analysis
plot(t,linear.ls$residuals,main="Plot of Residuals", xlab="Time", ylab="Residuals")
qqnorm(linear.ls$residuals)
qqline(linear.ls$residuals)
acf(linear.ls$residuals)
pacf(linear.ls$residuals)
## Arima
linear.ar2 <- arima(linear.y,order=c(2,0,0),xreg=t,method="ML")
linear.ar3 <- arima(linear.y,order=c(3,0,0),xreg=t,method="ML")

tsdiag(linear.ar2)
tsdiag(linear.ar3)

pacf(linear.ar2$residuals)
qqnorm(linear.ar2$residuals)
qqline(linear.ar2$residuals)

pacf(linear.ar3$residuals)
qqnorm(linear.ar3$residuals)
qqline(linear.ar3$residuals)

## Forecating and PI
new.t <- 101:105
linear.pr <- predict(linear.ar3,n.ahead=5,newxreg=new.t,se.fit=TRUE)
## Predicted value
y.pred <- linear.pr$pred
y.sd <- linear.pr$se
## Upper bounds of PIs
U <- y.pred + 2*y.sd
## Lower bounds of PIs
L <- y.pred - 2*y.sd
```