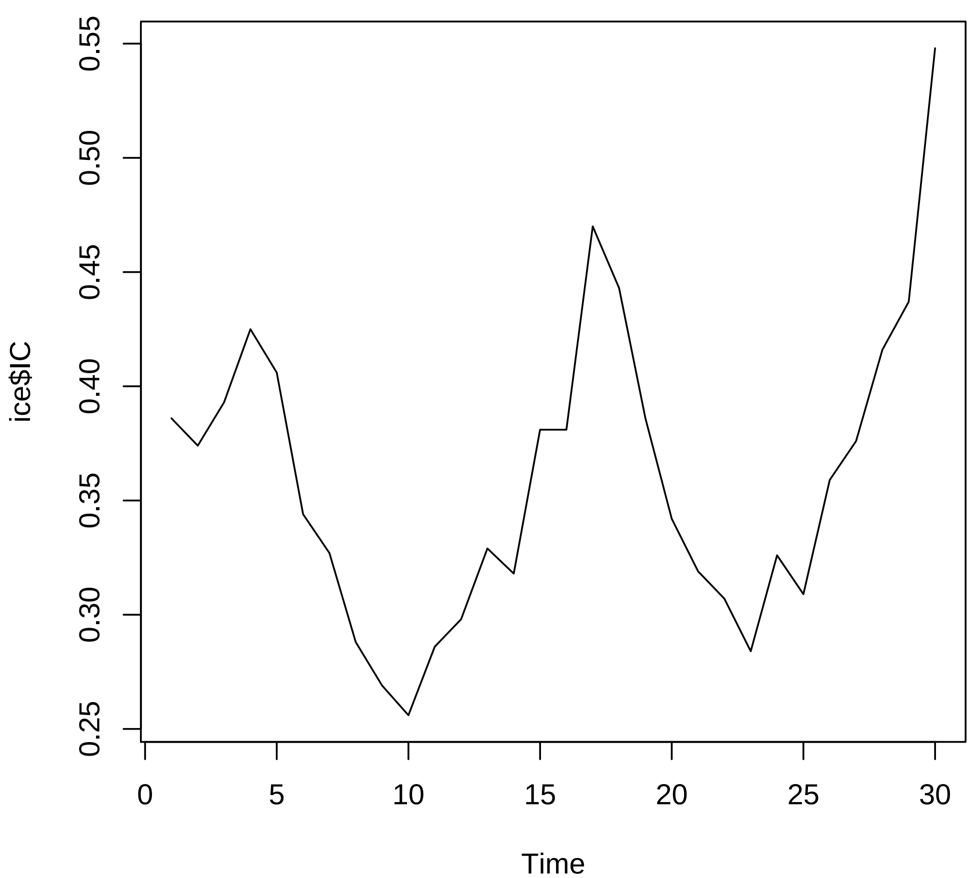
Homework 7 - STAT 511
Amal Agarwal

Answer 1

(a) • The given data was extracted and exploratory data analysis was conducted by plotting pairwise scatter plots. In particular the plot of consumption vs. date indicates positive autocorrelation as can be seen by the following plot:

```
> ice=read.csv("icecream.csv",sep=",")
> plot.ts(ice$IC)
> ice<-ice[-length(ice$date),]
```



Fitting a simple linear model

$$IC_i = \beta_0 + \beta_1 date_i + \beta_2 price_i + \beta_3 income_i + \beta_4 temp_i + \epsilon_i$$

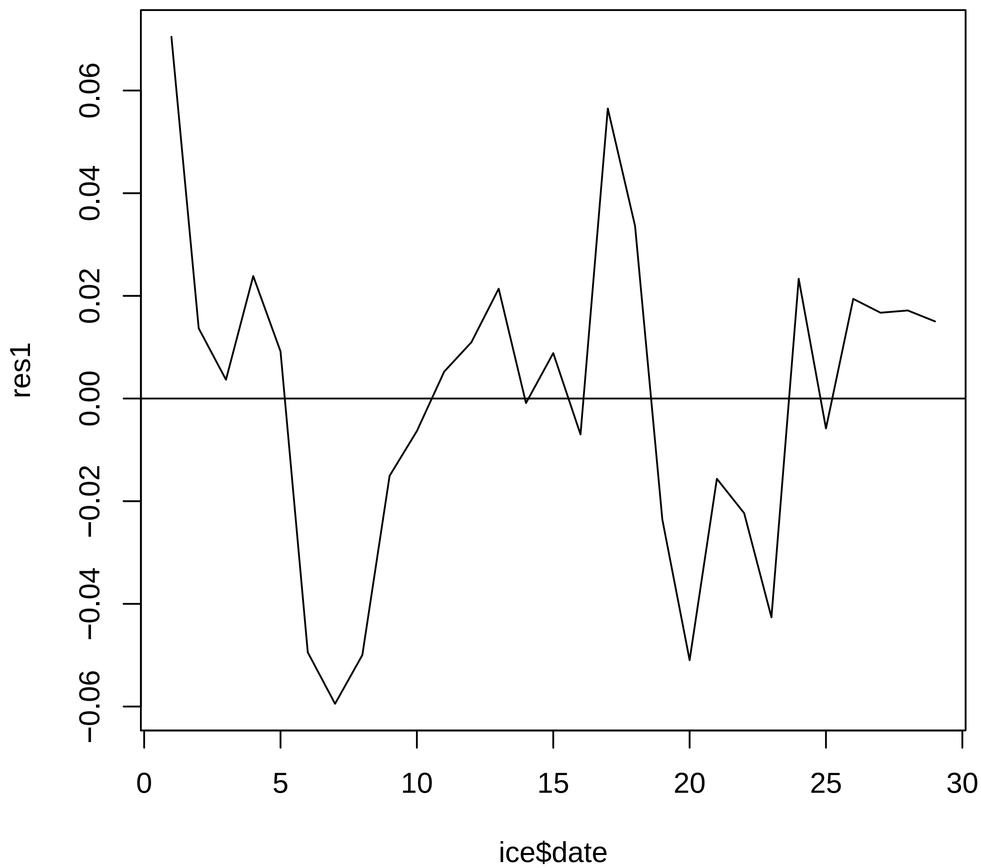
where $\epsilon_i \sim N(0, \sigma^2)$

and observing the summary as

```
> fit1=lm(IC~.,data=ice)
> summary(fit1)

Call:
lm(formula = IC ~ ., data = ice)

Residuals:
      Min        1Q    Median        3Q       Max
-0.059458 -0.015630  0.005229  0.017152  0.070472

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.666e-02  3.082e-01   0.281    0.781
date        -8.898e-06  1.478e-03  -0.006    0.995
price       -3.854e-01  8.141e-01  -0.473    0.640
income       2.629e-03  2.133e-03   1.233    0.230
temp         3.120e-03  4.419e-04   7.060 2.68e-07 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

Residual standard error: 0.03359 on 24 degrees of freedom
Multiple R-squared: 0.6948,	Adjusted R-squared: 0.6439
F-statistic: 13.66 on 4 and 24 DF,  p-value: 6.102e-06
```

we can infer the following things:

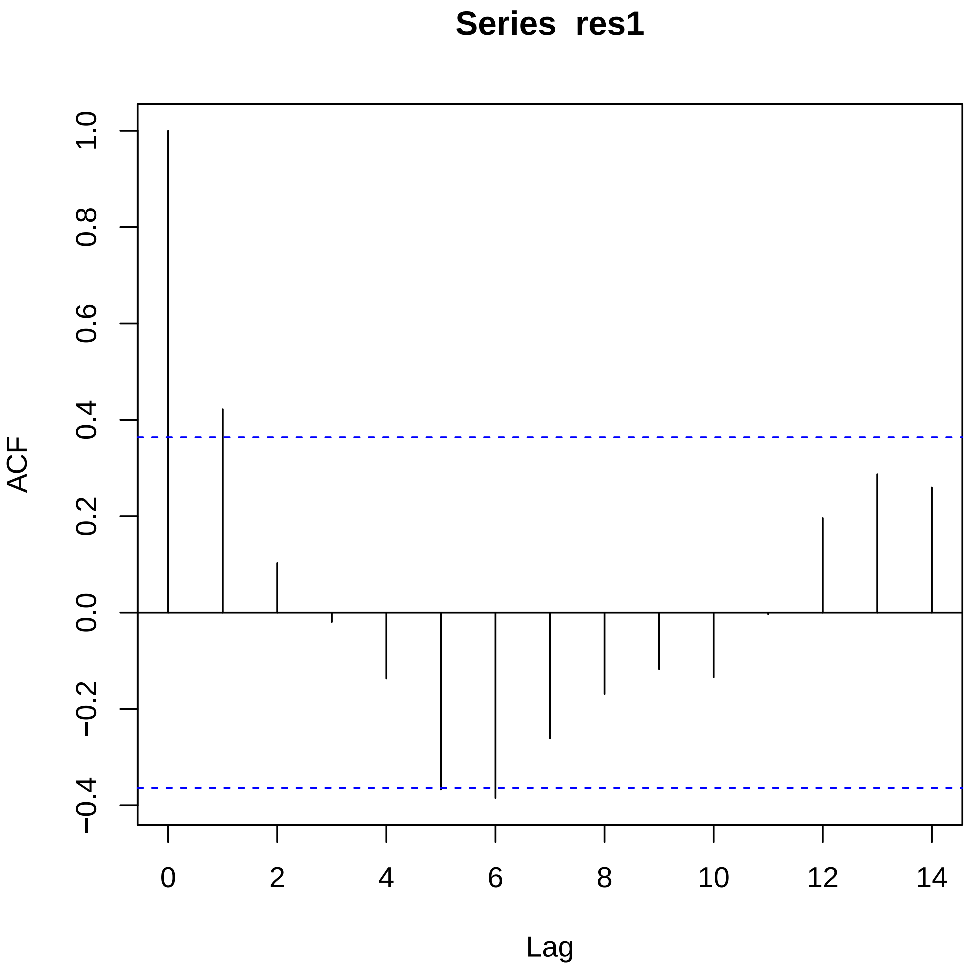
– Plotting the reiduals vs. date we get,



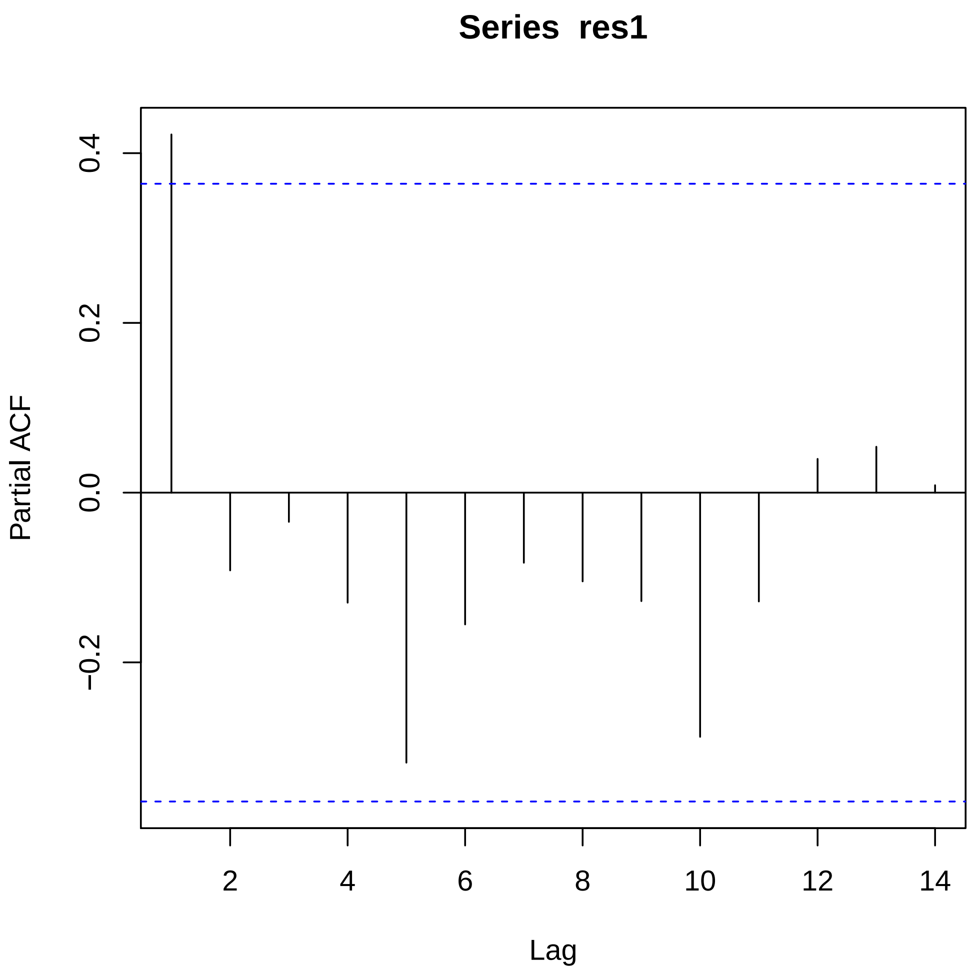
The above residual plot clearly confirms our hypothesis of positive autocorrelation under the fitted model. However, it also indicates a global periodic trend and suggests that if we include a sine/cosine term, we might be able to use our normal uncorrelated errors model.

– Testing for the autocorrelation:

```
> acf(res1)
```



```
> pacf(res1)
```



Tis confirms that only lag-1 autocorrelation is significant and so we can use an AR(1) time series model.

- Now fitting the following AR(1) correlated errors linear model:

$$IC = \beta_0 + \beta_1 date + \beta_2 price + \beta_3 income + \beta_4 temp + \epsilon$$

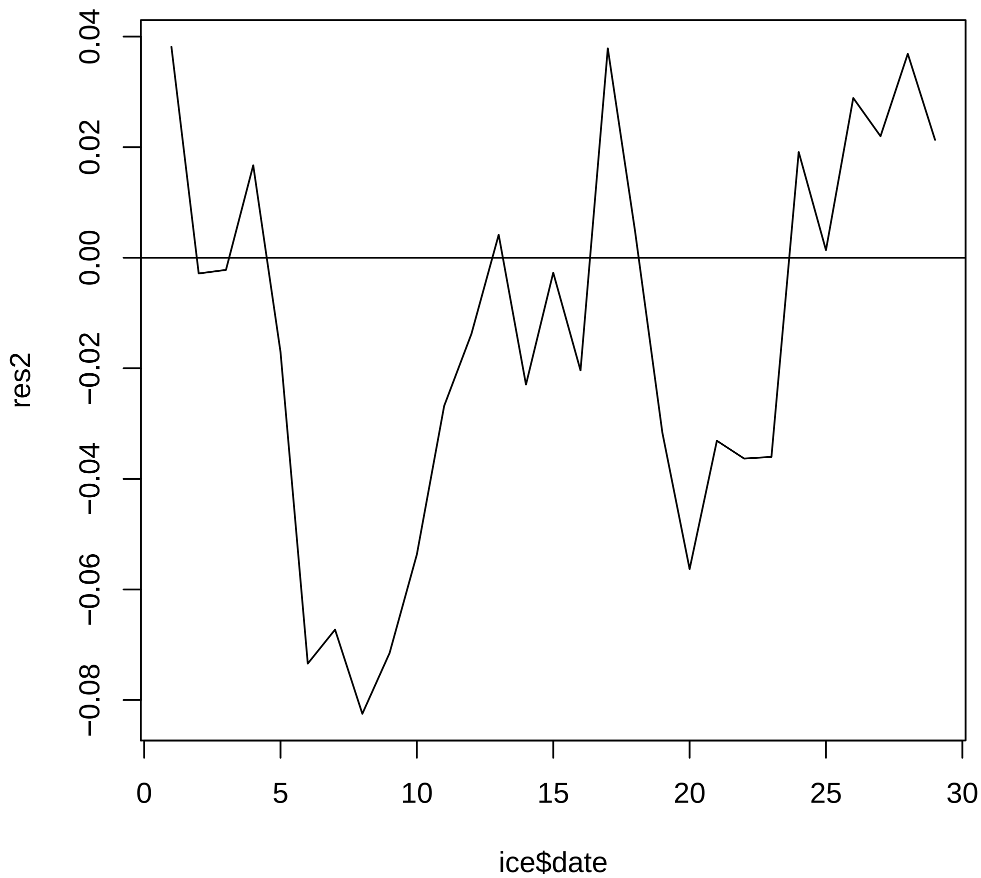
where $\epsilon \sim N(0, \Sigma), \Sigma_{ij} = \dfrac{\sigma_u^2}{1-\rho^2}\rho^{|i-j|}$

```
> library(nlme)
> fit2=gls(IC~.,data=ice,correlation=corAR1(),method="REML")
> summary(fit2)
```

```
Generalized least squares fit by REML
  Model: IC ~ .
```

```
  Data: ice
        AIC       BIC   logLik
  -76.34285 -68.09647 45.17142

Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
0.8757438

Coefficients:
                 Value Std.Error   t-value p-value
(Intercept)  0.5358471 0.2630945  2.036709  0.0529
date         0.0010757 0.0030762  0.349682  0.7296
price       -0.6250332 0.6652191 -0.939590  0.3568
income      -0.0015569 0.0019697 -0.790421  0.4370
temp         0.0024662 0.0006535  3.773659  0.0009

 Correlation:
       (Intr) date   price  income
date    0.089
price  -0.774 -0.040
income -0.687 -0.345  0.136
temp   -0.224 -0.150  0.028  0.170

Standardized residuals:
       Min         Q1        Med         Q3        Max
-1.3961197 -0.6097035 -0.2332330  0.2828391  0.6459124

Residual standard error: 0.05908097
Degrees of freedom: 29 total; 24 residual
```

Checking the residual plot again:



The above plot looks much more better. Based on the summary of the fit, it is clear that the p-values for the date, price and income are very large (greater than 0.05 significance criteria) which suggests that these predictor variables are not statistically significant in explaining the variation in consumption. Thus, these variables can be dropped from the model.

- Fitting the following correlated errors linear model:

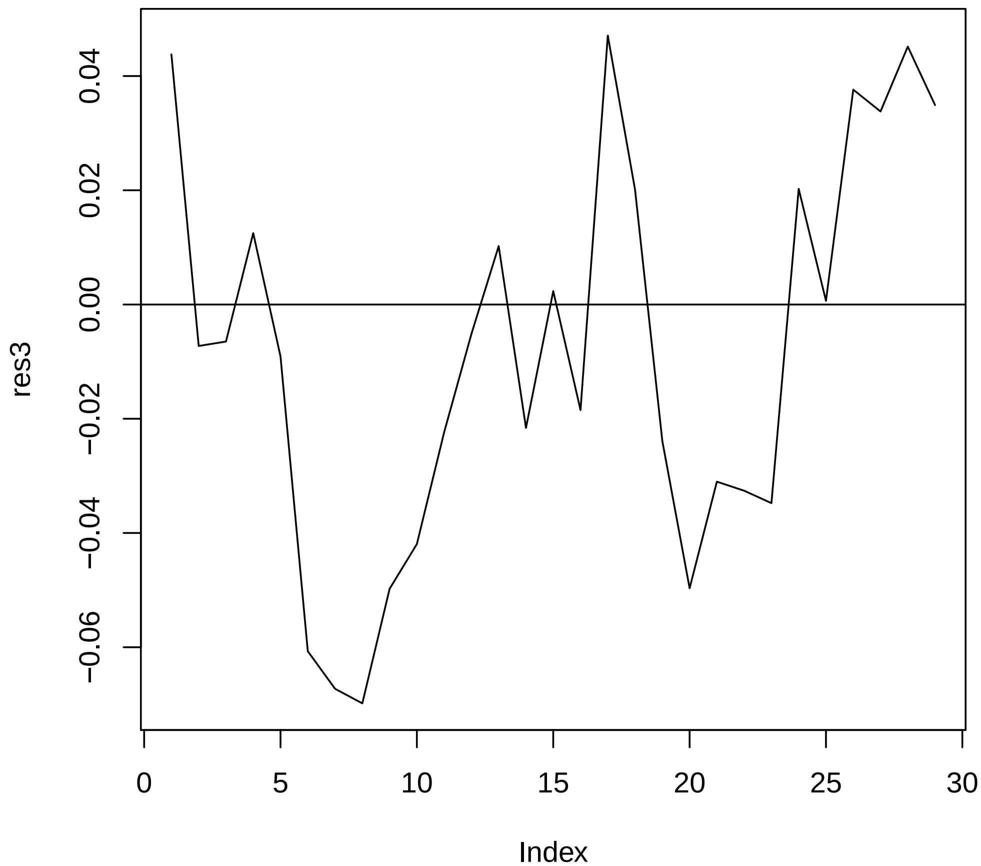
$$IC = \beta_1 temp + \epsilon$$

where $\epsilon \sim N(0, \Sigma), \Sigma_{ij} = \dfrac{\sigma_u^2}{1-\rho^2}\rho^{|i-j|}$

```
> fit3=gls(IC~temp,data=ice,correlation=corAR1(),method="REML")
> summary(fit3)
Generalized least squares fit by REML
  Model: IC ~ temp
  Data: ice
        AIC       BIC  logLik
  -101.0608 -95.87745 54.5304

Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
0.7557313

Coefficients:
                 Value  Std.Error  t-value p-value
(Intercept) 0.23545840 0.03462528 6.800188   0e+00
temp        0.00260358 0.00059407 4.382617   2e-04

 Correlation:
     (Intr)
temp -0.842

Standardized residuals:
       Min         Q1        Med         Q3        Max
-1.6894484 -0.7887911 -0.1756249  0.4859311  1.1391942

Residual standard error: 0.04133097
Degrees of freedom: 29 total; 27 residual
```

Checking the residual plot again:

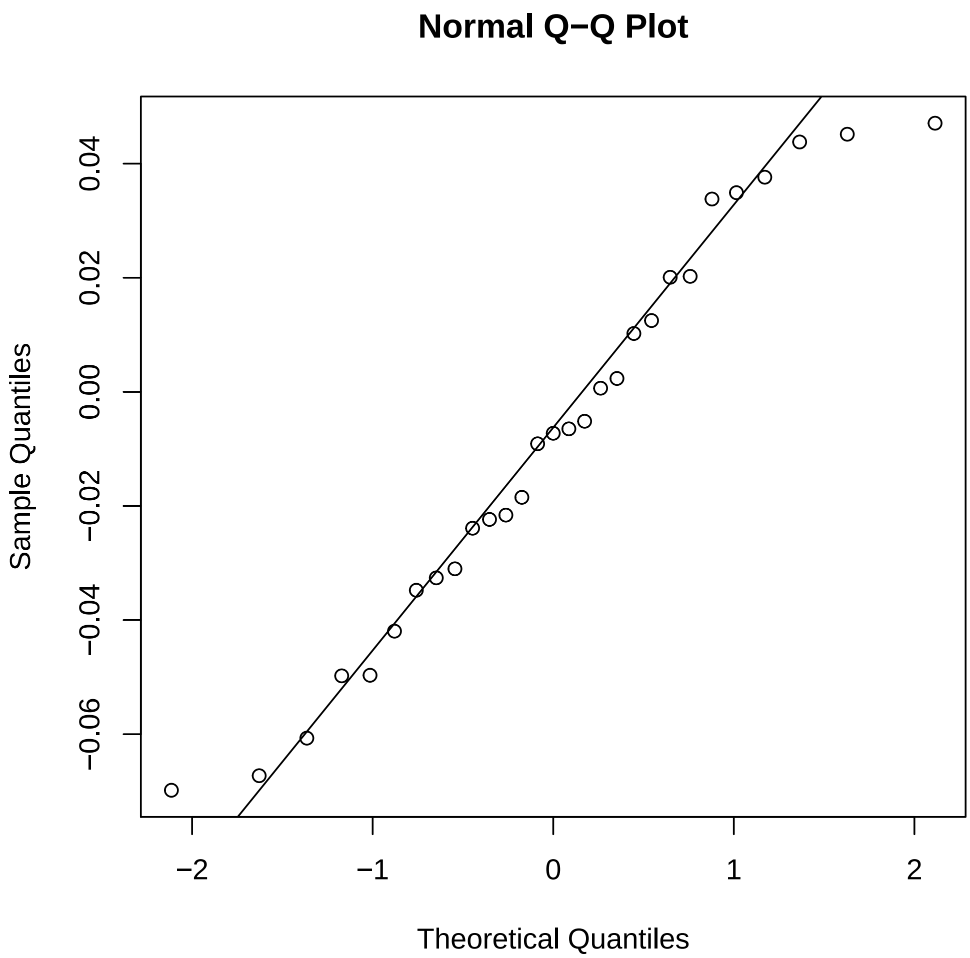
```
> res3=resid(fit3)
> plot(res3,type='l')
> abline(h=0)
```



The residual plot looks good now since the positive autocorrelation has been taken into account. There is no visible non-linear trend.

Clearly the estimate of ρ in our model is 0.7557313 which indicates a high postive lag 1 autocorrelation.

The qq plot of residuals is given as:



The above plot shows some short tails which are not significant. Hence our normality assumption is satisfied.

The estimated coeffiecients are given as:

```
> fit3$coeff

(Intercept)        temp
0.235458396 0.002603578
```

The positive slope of 0.002 shows that the mean consumption increases by 0.002 units with a unit increase in temperature. The postive value of intercept indiactes that the consumption at zero temperature. Note that the temperature is in Fahrenheit and thus it makes sense that even at 0^0 F, the consumtion is 0.235 units. Further the estimated value of the nuisance parameter σ_u^2 is

```
> X=cbind(1,ice$temp)
> n=29
> p=2
> rho.hat=0.7557313
> C.ar1=corAR1(rho.hat)
> C.ar1=Initialize(C.ar1,data=ice)
> R=corMatrix(C.ar1)
> W=solve(R[1:29,1:29]/(1-(rho.hat)^2))
> sigma2<-(t(res3)%*%W%*%res3)/(n-p)
> sigma2
             [,1]
[1,] 0.0007326174
```

(b) Using the procedure mentioned above, the p-value under t distribution for $H_0 : \beta_1 = 0$ can be calculated as

```
> Y<-solve(t(X)%*%W%*%X)
> F<-(as.numeric(fit3$coeff[2])^2)/(Y[2,2]*sigma2)
> p_val<-2*(1-pt((sqrt(F)),(n-p)))
> p_val

             [,1]
[1,] 0.0001598054
```

Note that this caluclated p value is same as the p value obtained from the summary of fitting the final model i.e. summary of fit 3 shown earlier.

(c) The predicted value at date=30 calculated using predict function is given as:

```
> ice=read.csv("icecream.csv",sep=",")
> newdata=ice[30,]
> Y30_cap=predict(fit3, newdata)
> Y30_cap
```

```
[1] 0.4203124
attr(,"label")
[1] "Predicted values"
```

which is slightly different from the true response

```
> ice$IC[30]
```

```
[1] 0.548
```

The prediction interval an be calculated as:

```
> H=X%*%Y%*%t(X)%*%W
> I=diag(1,29)
> V=((I-H)%*%solve(W)%*%t(I-H))
> h29<-V[29,29]
> h29
```

```
[1] 1.748117
```

```
> X30=matrix(c(1,ice$temp[30]),nrow=1,ncol=2)
> cf=(1/(1-rho.hat^2))+((rho.hat^2)*h29)+(X30%*%Y%*%t(X30))
> s=sqrt(fit3$sigma^2*cf)
> U=Y30_cap+s*qt(0.975,(n-p))
> L=Y30_cap-s*qt(0.975,(n-p))
> L
```

```
          [,1]
[1,] 0.2499013
attr(,"label")
[1] "Predicted values"
```

```
> U
```

```
          [,1]
[1,] 0.5907236
attr(,"label")
[1] "Predicted values"
```

Clearly the lower bound L=0.2499013 and upper bound U=0.5907236 contains the true response=0.548 at date=30.