Week 4 Lecture Note

Hyoeun Lee

Module 1 - Week 4

Time Series Regression

Time Series X_t, $t = 1, \dots, n$ is possibly influenced by $Z_{t1}, Z_{t2}, \dots, Z_{tq}$.

We express the genreal relation through the *linear regression model*

$$X_t = \beta_0 + \beta_1 Z_{t1} + \beta_2 Z_{t2} + \cdots + \beta_q Z_{tq} + W_t,$$

- $\beta_0, \dots, \beta_q$: unknown fixed regression coefficients
- $\{W_t\}$ is white noise (normally distributed) with variance σ_W^2.

Ordinary Least Sqruares (OLS) Method

In ordinary least squares (OLS), we minimize the *error sum of squares*

$$S = \sum_{t=1}^{n} W_t^2 = \sum_{t=1}^{n} (X_t - [\beta_0 + \beta_1 Z_{t1} + \beta_2 Z_{t2} + \cdots + \beta_q Z_{tq}])^2$$

with respect to β_i for $i = 0, 1, \dots, q$.

We denote the parameters obtained using OLS on the linear regression model $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_q$.

How do we estimate σ_W^2?

For $j = 1, \dots, n$,

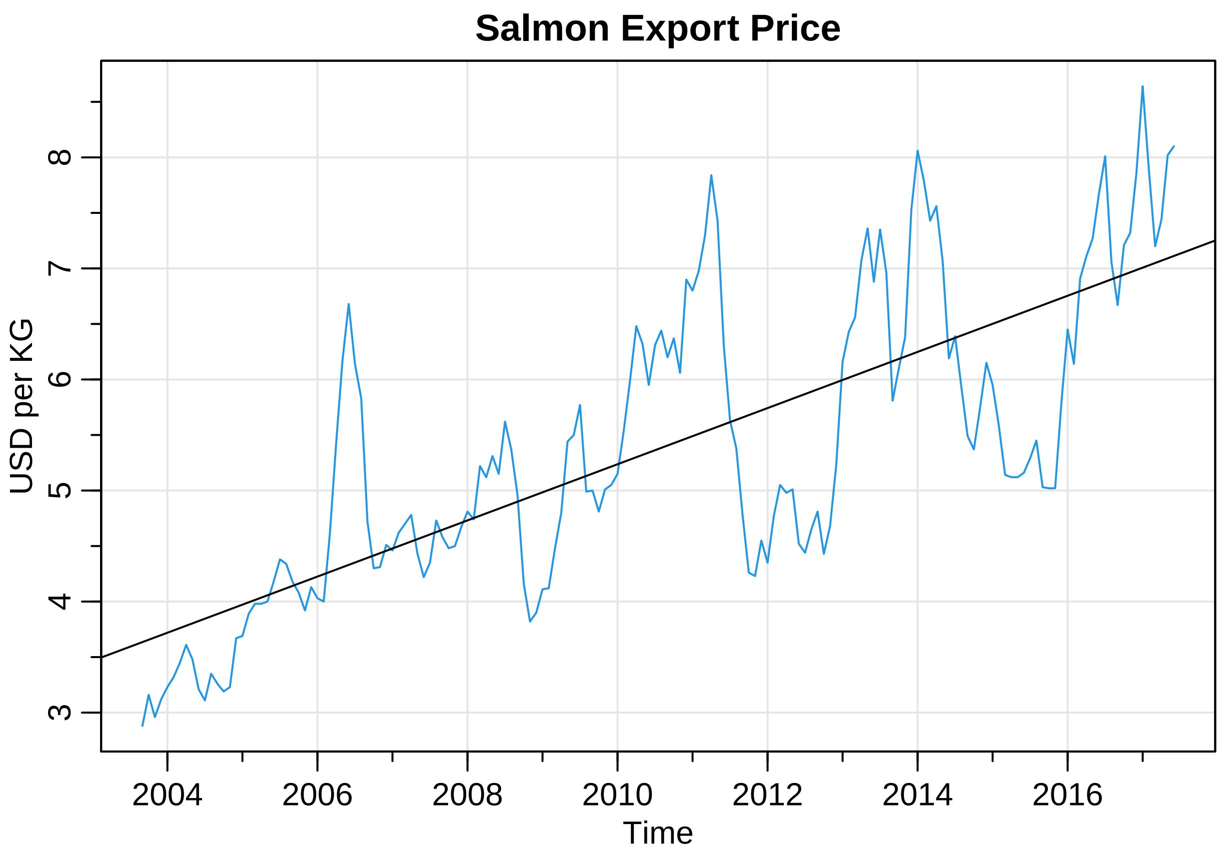
$$\hat{W}_j = X_j - (\hat{\beta}_0 + \hat{\beta}_1 Z_{t1} + \hat{\beta}_2 Z_{t2} + \cdots + \hat{\beta}_q Z_{tq})$$

Hence, we use

$$\hat{\sigma}_W^2 = \frac{\sum_{j=1}^{n} \hat{W}_j^2}{n - (q+1)}$$

Example: Estimating a linear trend of a commodity

```r
tsplot(salmon, col=4, ylab="USD per KG", main="Salmon Export Price")
fit <- lm(salmon~time(salmon), na.action=NULL)
abline(fit)
```



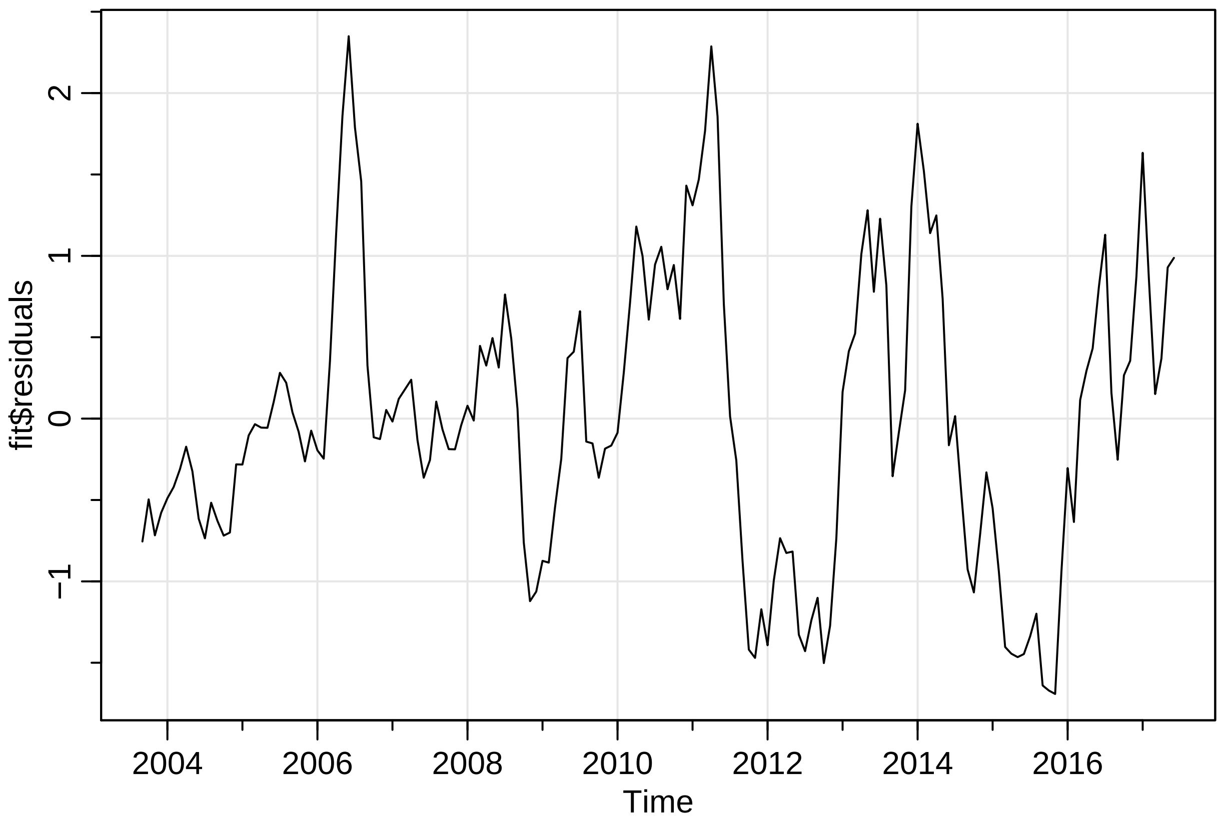

```r
summary(fit)
```

```
##
## Call:
## lm(formula = salmon ~ time(salmon), na.action = NULL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.69187 -0.62453 -0.07024  0.51561  2.34959
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -503.08947   34.44164  -14.61   <2e-16 ***
## time(salmon)    0.25290    0.01713   14.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8814 on 164 degrees of freedom
## Multiple R-squared:  0.5706, Adjusted R-squared:  0.568
## F-statistic: 217.9 on 1 and 164 DF,  p-value: < 2.2e-16
```

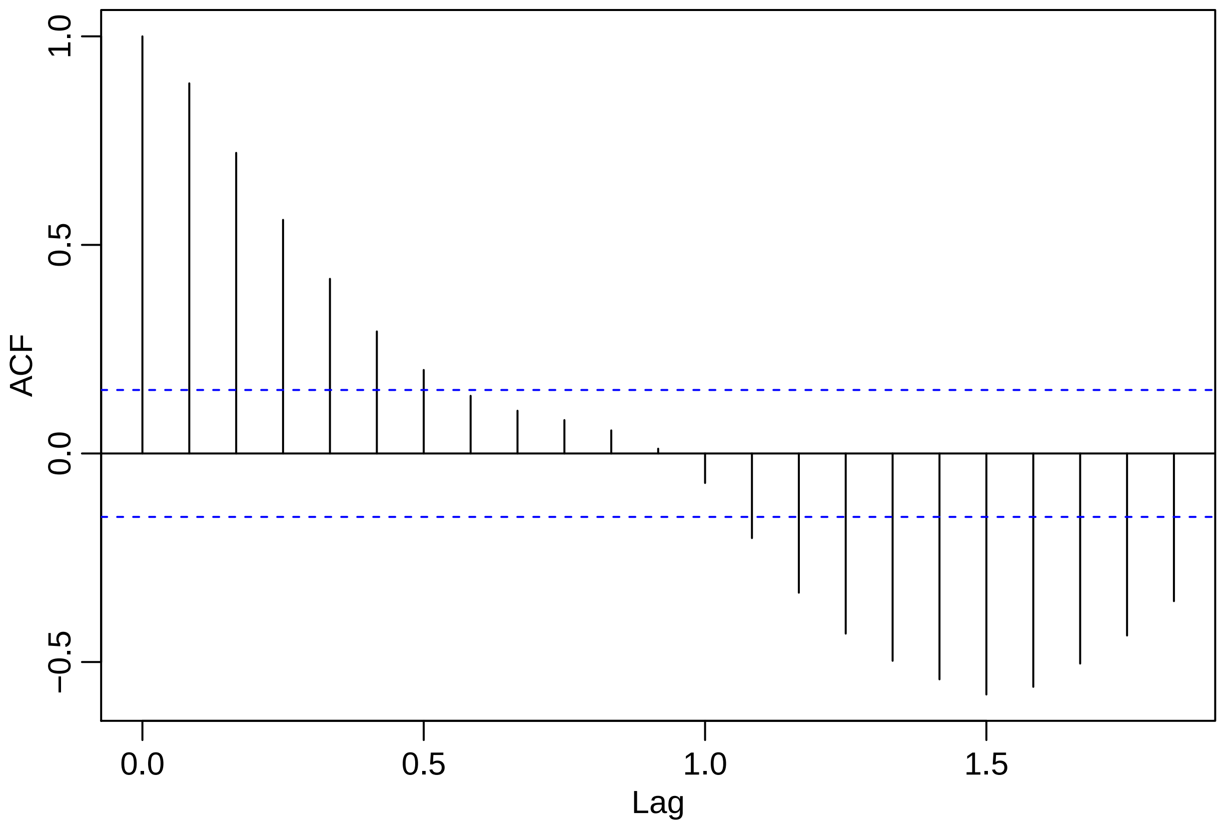
```r
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: salmon
##                Df Sum Sq Mean Sq F value    Pr(>F)
## time(salmon)    1 169.30 169.300  217.95 < 2.2e-16 ***
## Residuals     164 127.39   0.777
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
tsplot(fit$residuals)
```




```r
acf(fit$residuals)
```



ACF
1.0
0.5
0.0
−0.5
0.0
0.5
1.0
1.5
Lag

Model Selection Procedures

- Use the F test to compare one model against another
- Use the Test in a stepwise manner, by adding and/or deleting variables (stepwise regression)
- Evaluate each model on its on merit using Maximum likelihood Estimator for the variance,

$$\hat{\sigma}_k^2 = \frac{SSE_k}{n}$$

(Residual Sum of Squares with k regression coefficients)

Akaike (1974) suggested balancing the accuracy of the fit against the number of parameters in the model.

Akaike's Information Criterion (AIC)

$$AIC = \log \hat{\sigma}_k^2 + \frac{n + 2k}{n}$$

The model yielding the minimum AIC is the best model because: * small error $\hat{\sigma}_k^2$ * not overly complex

AIC, Bias corrected (AICc)

$$AICc = \log \hat{\sigma}_k^2 + \frac{n + k}{n - k - 2}$$

- A corrected form, suggested by Sugiura (1978), and expanded by Hurvich and Tsai (1989)
- based on small-sample distributional results for the linear regression model
- If sample size is relatively low [Burnham & Anderson (2002, ch. 7) suggests criteria $n/k < 40$] , AICc may be preferred.

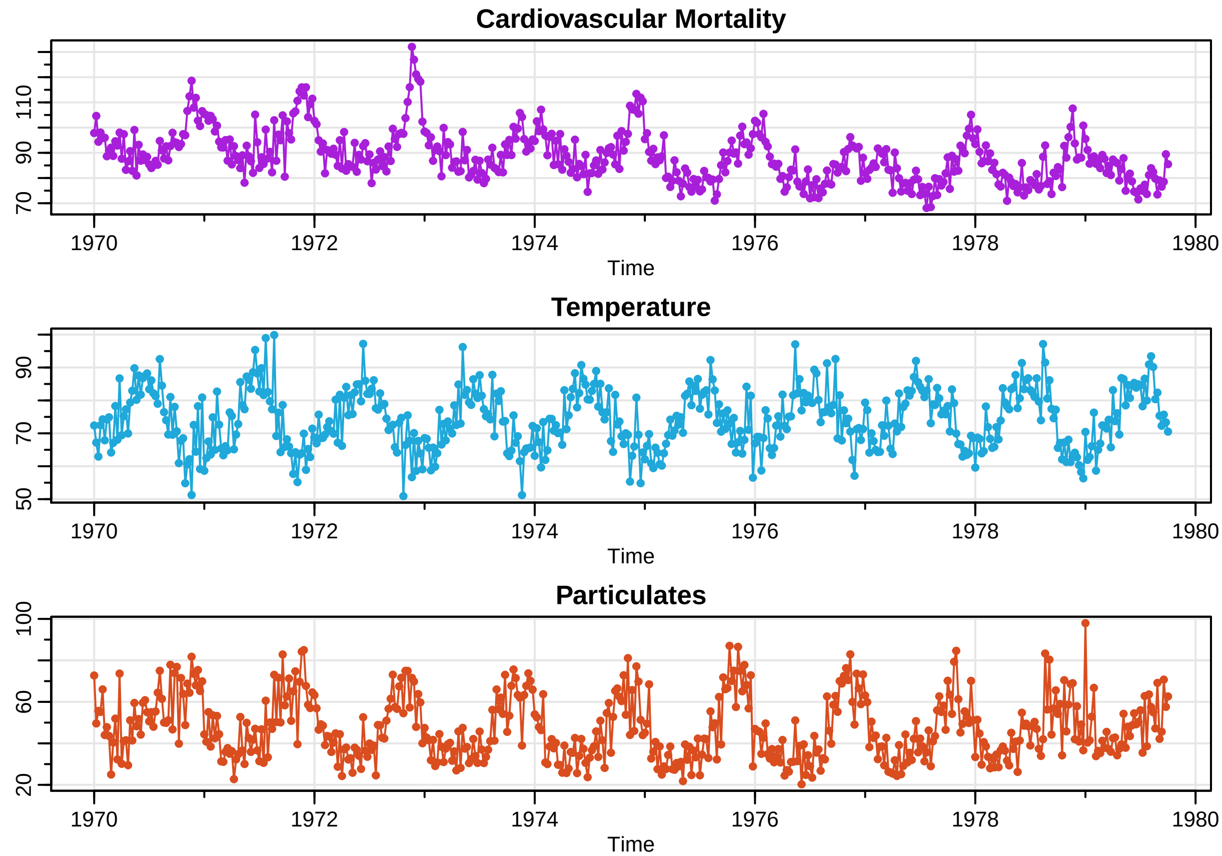
Bayesian Information Criterion (BIC)

$$BIC = \log \hat{\sigma}_k^2 + \frac{k \log n}{n}$$

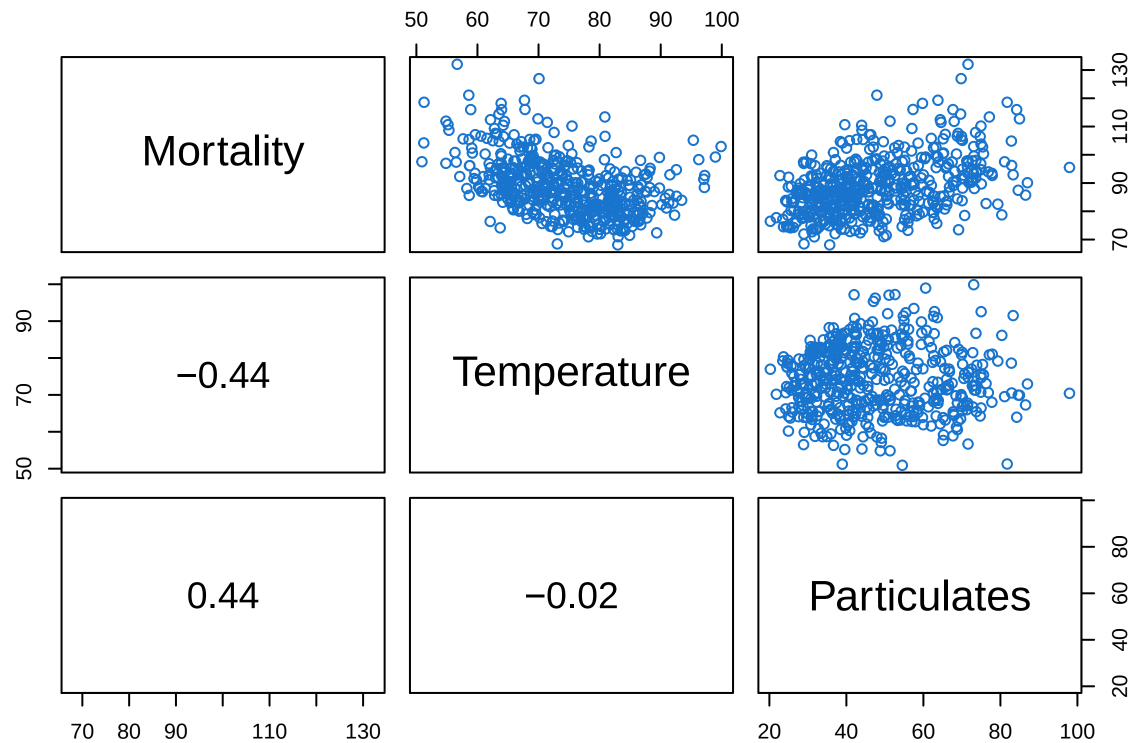
- penalty term based on Bayesian arguments, as in Schwarz (1978)
- penalty for additional parameters is more in BIC than AIC.

Example: Pollution, Temperature and Morality

```
culer = c(rgb(.66,.12,.85), rgb(.12,.66,.85), rgb(.85,.30,.12))
par(mfrow=c(3,1))
tsplot(cmort, main="Cardiovascular Mortality", col=culer[1], type="o", pch=19, ylab="")
tsplot(tempr, main="Temperature", col=culer[2], type="o", pch=19, ylab="")
tsplot(part, main="Particulates", col=culer[3], type="o", pch=19, ylab="")
```




```
##
panel.cor <- function(x, y, ...){
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- round(cor(x, y), 2)
  text(0.5, 0.5, r, cex = 1.75)
}
pairs(cbind(Mortality=cmort, Temperature=tempr, Particulates=part), col="dodgerblue3", lower.panel=panel
```




```
##
```

```r
temp = tempr - mean(tempr) # center temperature

cor(tempr, tempr^2)
```

```
## [1] 0.9972099
```

```r
cor(temp, temp^2)
```

```
## [1] 0.07617904
```

```
##
```

```r
temp = tempr - mean(tempr) # center temperature
temp2 = temp^2
trend = time(cmort) # time

fit1 = lm(cmort~ trend, na.action=NULL)
fit2 = lm(cmort~ trend+temp, na.action=NULL)
fit3 = lm(cmort~ trend + temp + temp2, na.action=NULL)
fit4 = lm(cmort~ trend + temp + temp2 + part, na.action=NULL)



fit=fit1
summary(fit) # regression results
```

```
##
```

```
## Call:
## lm(formula = cmort ~ trend, na.action = NULL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.445  -6.670  -1.366   5.505  40.107
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3297.6062   276.3132   11.93   <2e-16 ***
## trend         -1.6249     0.1399  -11.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.893 on 506 degrees of freedom
## Multiple R-squared:  0.2104, Adjusted R-squared:  0.2089
## F-statistic: 134.9 on 1 and 506 DF,  p-value: < 2.2e-16
```

```r
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: cmort
##            Df Sum Sq Mean Sq F value    Pr(>F)
## trend       1  10667 10666.9  134.87 < 2.2e-16 ***
## Residuals 506  40020    79.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

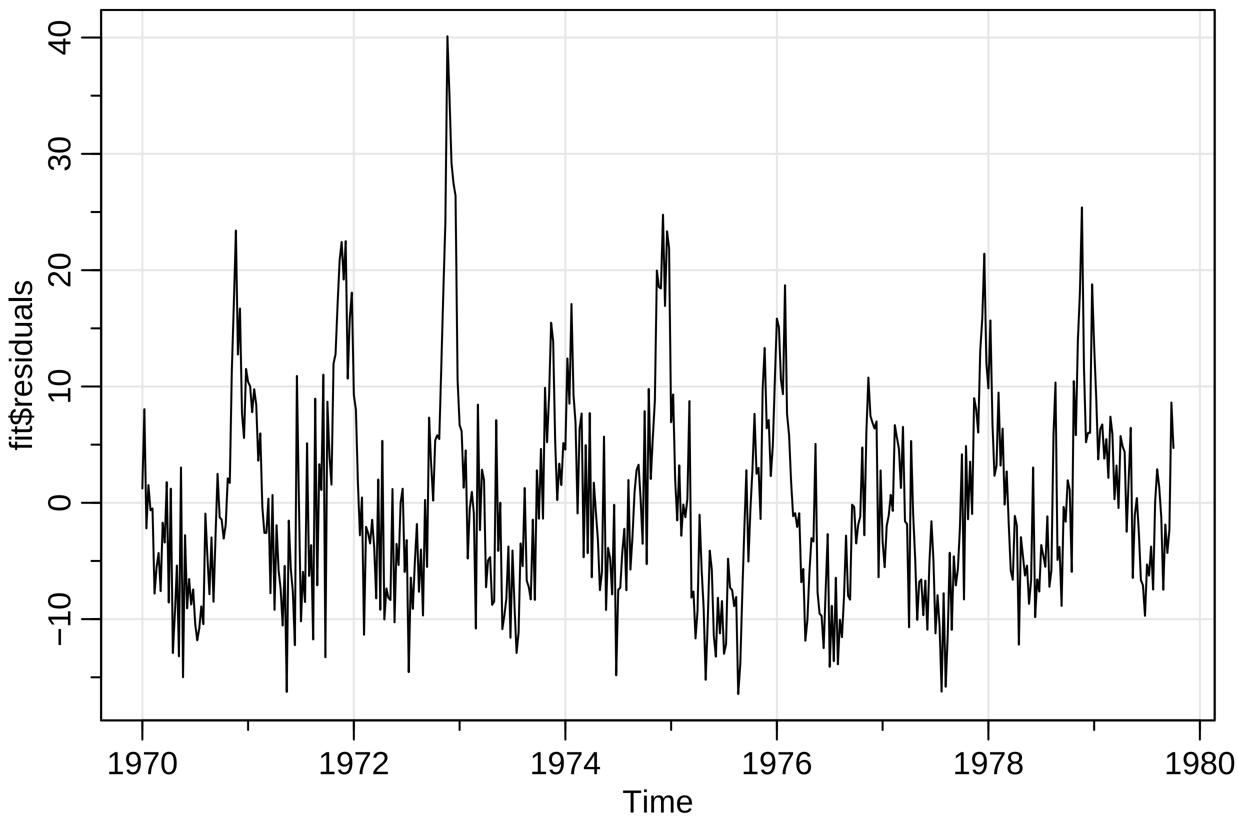
```r
num = length(cmort) # sample size
AIC(fit)/num - log(2*pi) # AIC
```

```
## [1] 5.37846
```

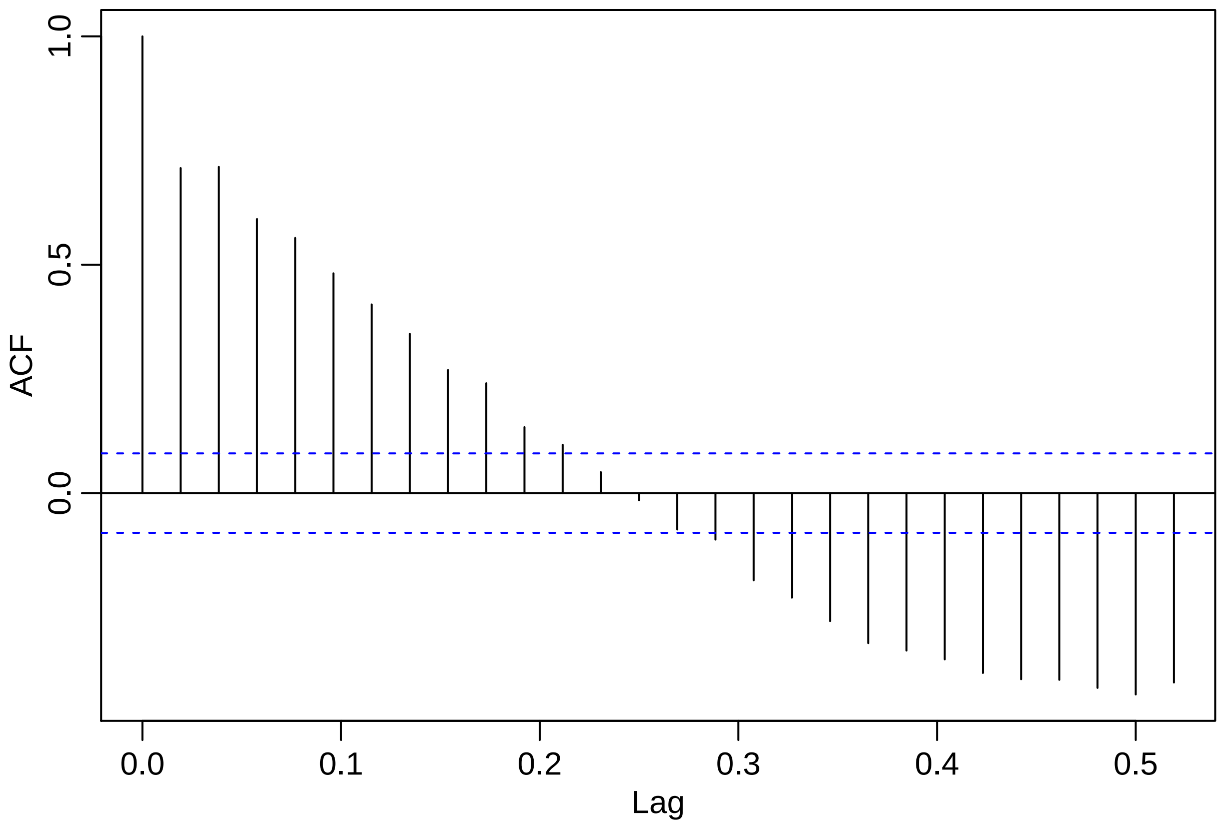
```r
BIC(fit)/num - log(2*pi) # BIC
```

```
## [1] 5.403443
```

```r
tsplot(fit$residuals)
```




```r
acf(fit$residuals)
```




```r
##
fit=fit2
summary(fit) # regression results
```

```
##
## Call:
## lm(formula = cmort ~ trend + temp, na.action = NULL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.846  -5.330  -1.207   4.701  33.306
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3125.75988  245.48233   12.73   <2e-16 ***
## trend         -1.53785    0.12430  -12.37   <2e-16 ***
## temp          -0.45792    0.03893  -11.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.887 on 505 degrees of freedom
## Multiple R-squared:  0.3802, Adjusted R-squared:  0.3778
## F-statistic: 154.9 on 2 and 505 DF,  p-value: < 2.2e-16
```

```r
anova(fit)
```

```
## Analysis of Variance Table
```

```
##
## Response: cmort
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## trend       1 10666.9 10666.9  171.48 < 2.2e-16 ***
## temp        1  8606.6  8606.6  138.36 < 2.2e-16 ***
## Residuals 505 31413.2    62.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

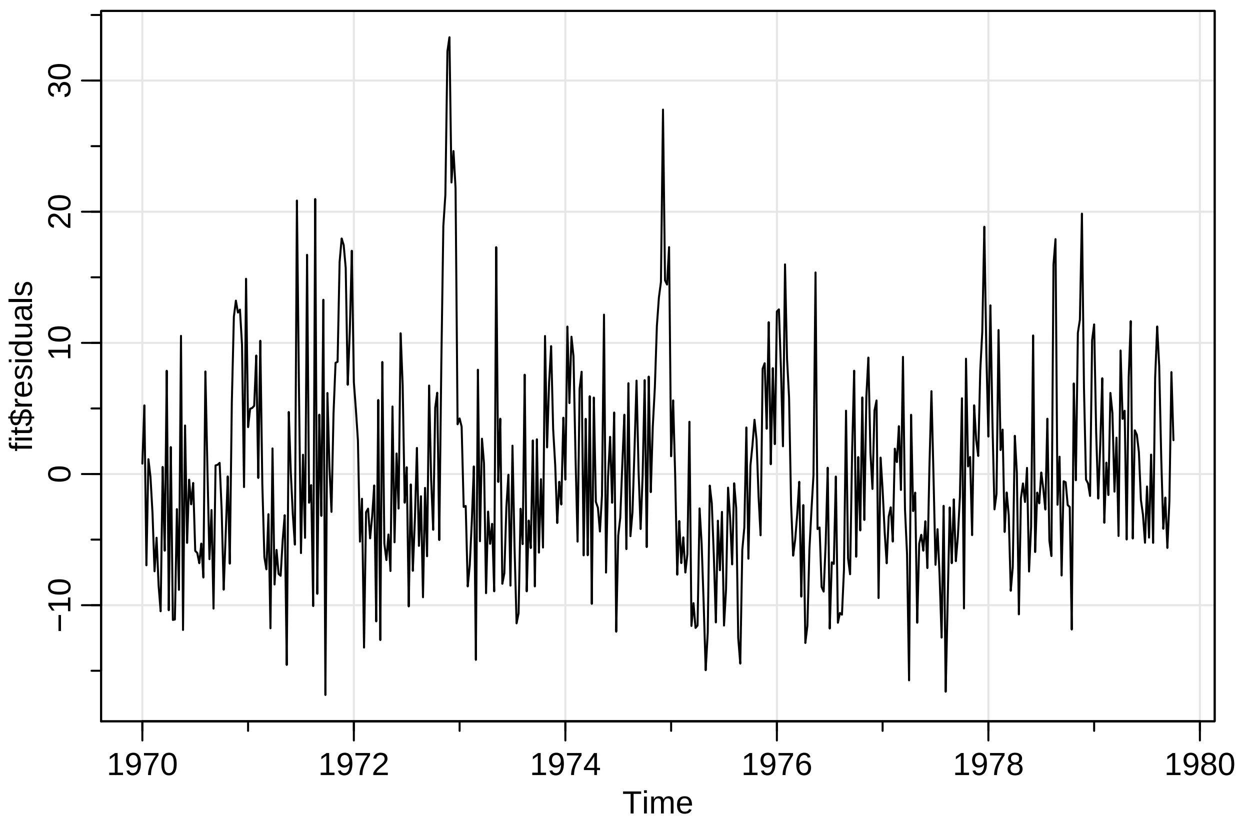
```r
num = length(cmort) # sample size
AIC(fit)/num - log(2*pi) # AIC
```

```
## [1] 5.14025
```

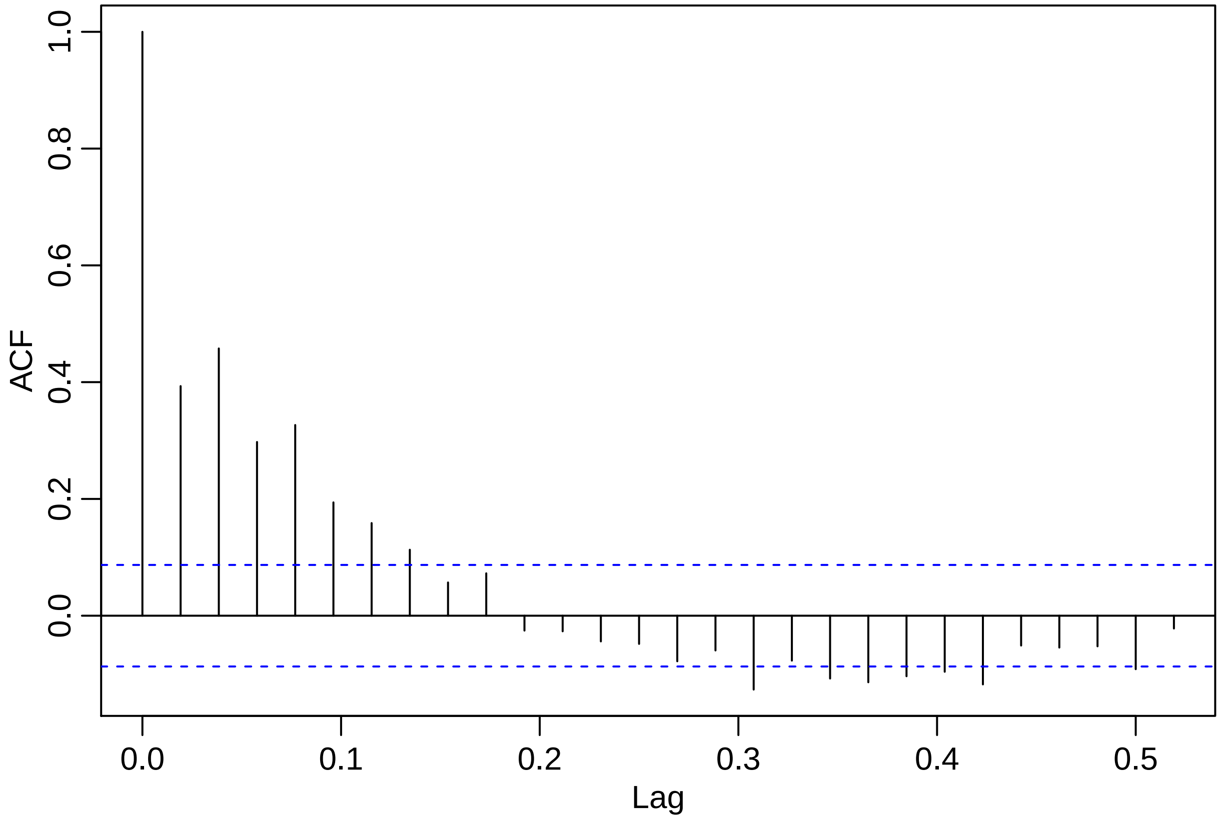
```r
BIC(fit)/num - log(2*pi) # BIC
```

```
## [1] 5.173561
```

```r
tsplot(fit$residuals)
```




```r
acf(fit$residuals)
```




```r
##
fit=fit3
summary(fit) # regression results
```

```
##
## Call:
## lm(formula = cmort ~ trend + temp + temp2, na.action = NULL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.464  -4.858  -0.945   4.511  34.939
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.038e+03  2.322e+02  13.083  < 2e-16 ***
## trend       -1.494e+00  1.176e-01 -12.710  < 2e-16 ***
## temp        -4.808e-01  3.689e-02 -13.031  < 2e-16 ***
## temp2        2.583e-02  3.287e-03   7.858 2.38e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.452 on 504 degrees of freedom
## Multiple R-squared:  0.4479, Adjusted R-squared:  0.4446
## F-statistic: 136.3 on 3 and 504 DF,  p-value: < 2.2e-16
```

```r
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: cmort
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## trend       1 10666.9 10666.9  192.11 < 2.2e-16 ***
## temp        1  8606.6  8606.6  155.00 < 2.2e-16 ***
## temp2       1  3428.7  3428.7   61.75 2.376e-14 ***
## Residuals 504 27984.5    55.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

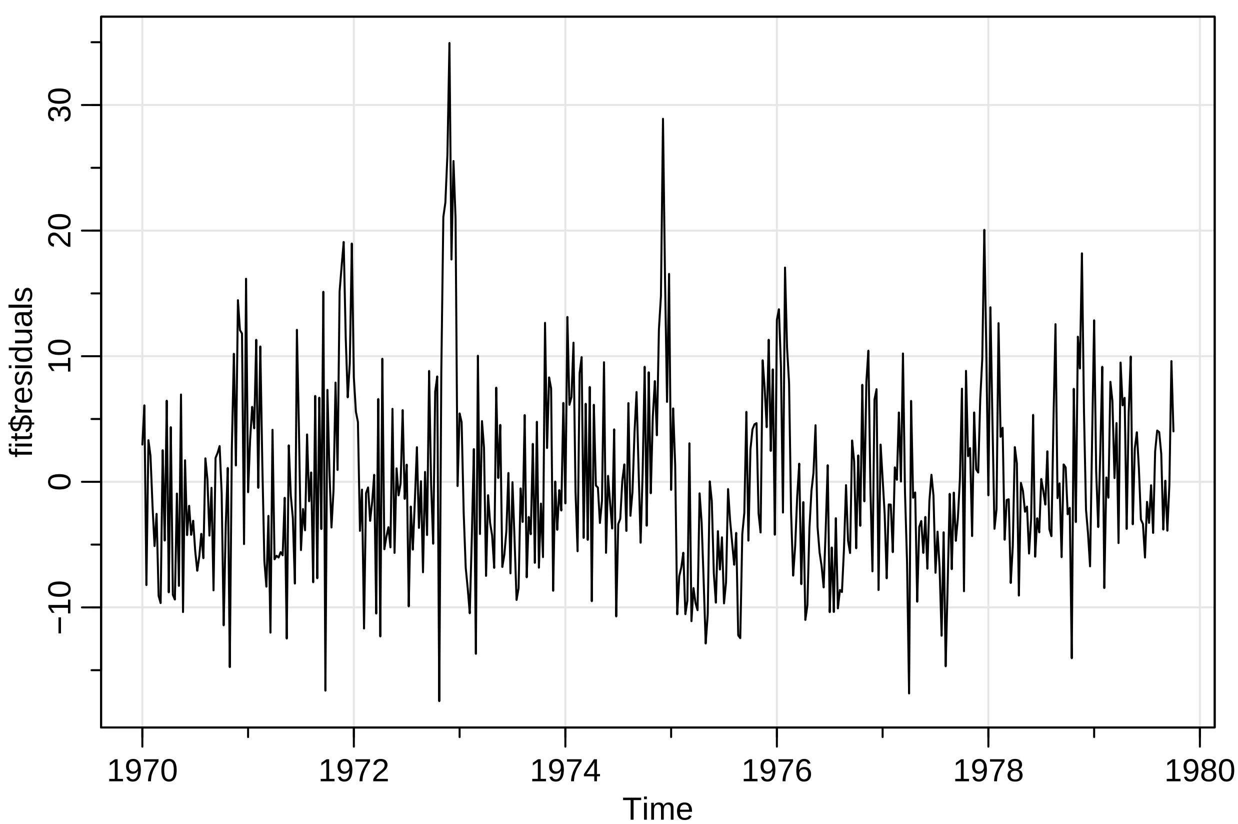
```r
num = length(cmort) # sample size
AIC(fit)/num - log(2*pi) # AIC
```

```
## [1] 5.028611
```

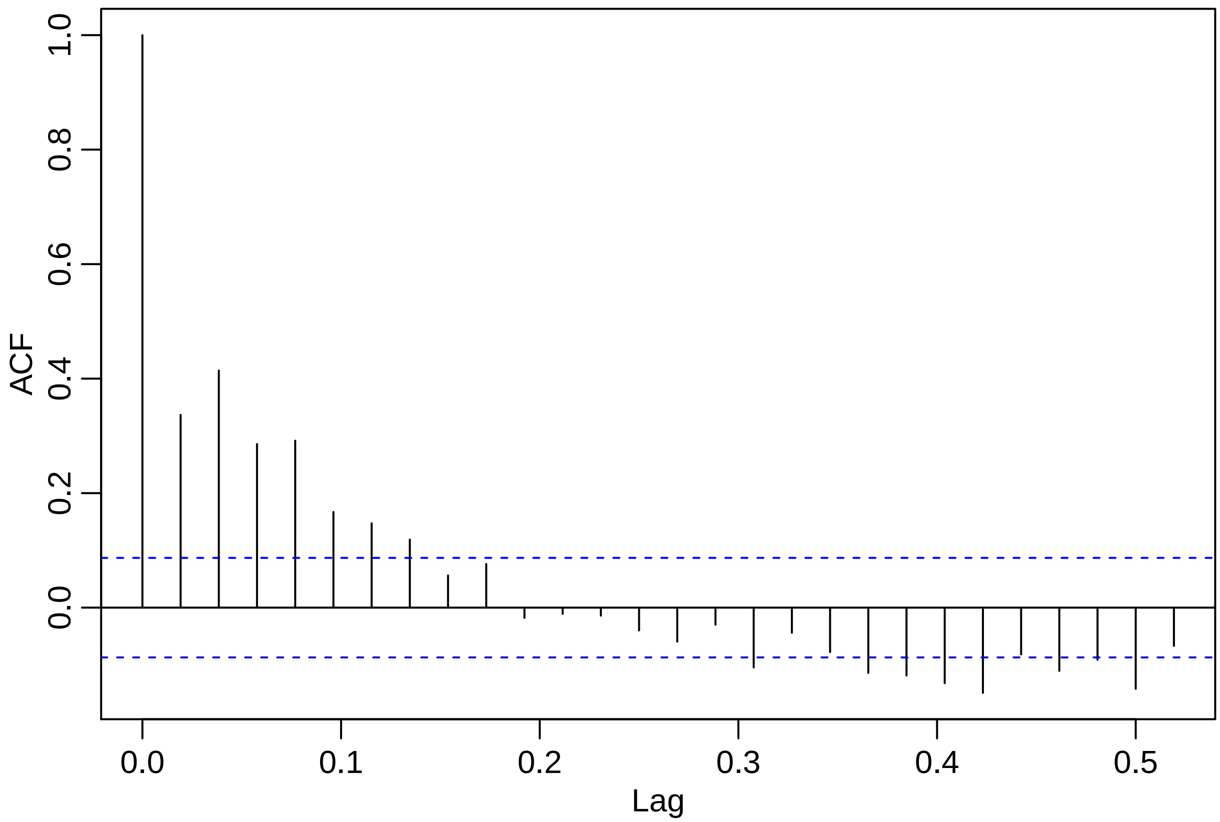
```r
BIC(fit)/num - log(2*pi) # BIC
```

```
## [1] 5.070249
```

```r
tsplot(fit$residuals)
```



```r
acf(fit$residuals)
```




```r
##
fit=fit4
summary(fit) # regression results
```

```
##
## Call:
## lm(formula = cmort ~ trend + temp + temp2 + part, na.action = NULL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.0760  -4.2153  -0.4878   3.7435  29.2448
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.831e+03  1.996e+02   14.19  < 2e-16 ***
## trend       -1.396e+00  1.010e-01  -13.82  < 2e-16 ***
## temp        -4.725e-01  3.162e-02  -14.94  < 2e-16 ***
## temp2        2.259e-02  2.827e-03    7.99 9.26e-15 ***
## part         2.554e-01  1.886e-02   13.54  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.385 on 503 degrees of freedom
## Multiple R-squared:  0.5954, Adjusted R-squared:  0.5922
```

```
## F-statistic:   185 on 4 and 503 DF,  p-value: < 2.2e-16
```

```r
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: cmort
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## trend       1 10666.9 10666.9 261.621 < 2.2e-16 ***
## temp        1  8606.6  8606.6 211.090 < 2.2e-16 ***
## temp2       1  3428.7  3428.7  84.094 < 2.2e-16 ***
## part        1  7476.1  7476.1 183.362 < 2.2e-16 ***
## Residuals 503 20508.4    40.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(lm(cmort~cbind(trend, temp, temp2, part)))) # Table 3.1
```

```
##                                  Df Sum Sq Mean Sq F value Pr(>F)
## cbind(trend, temp, temp2, part)   4  30178    7545     185 <2e-16 ***
## Residuals                       503  20508      41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

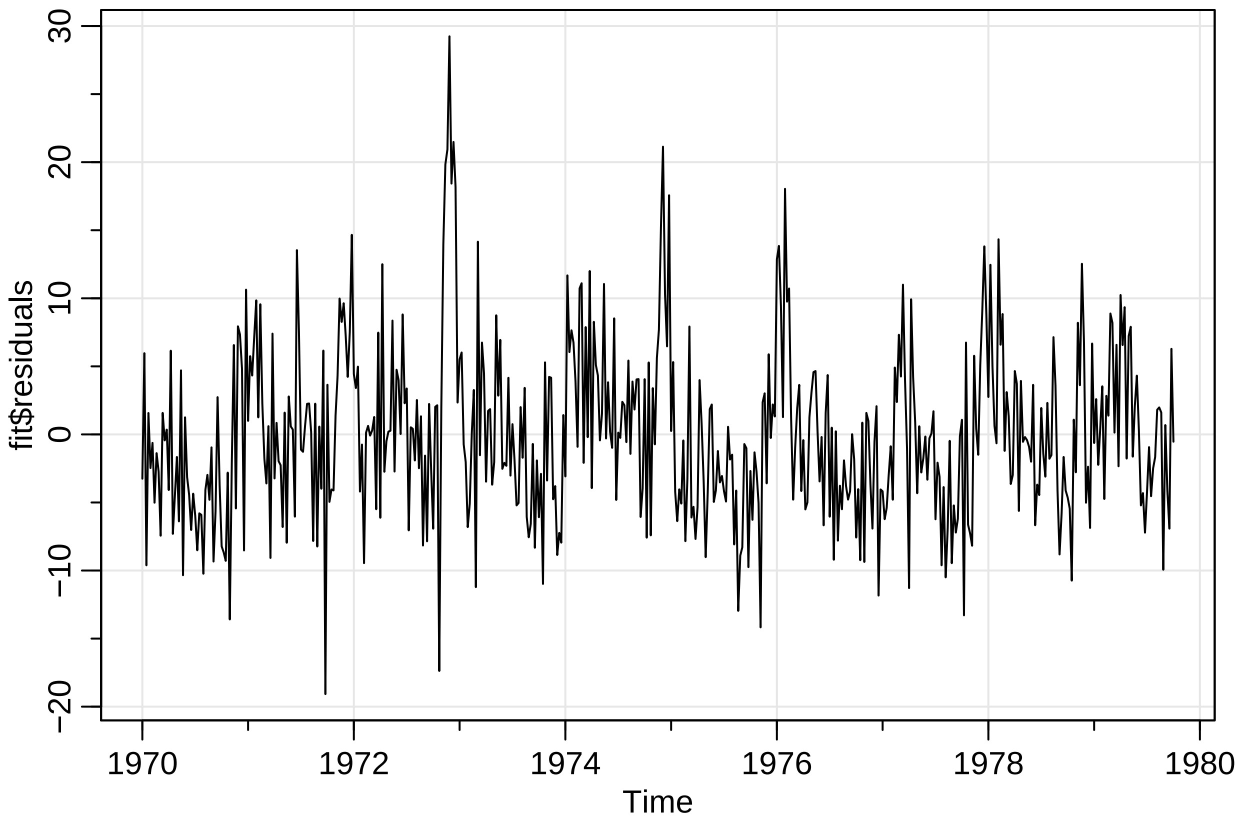
```r
num = length(cmort) # sample size
AIC(fit)/num - log(2*pi) # AIC
```

```
## [1] 4.721732
```

```r
BIC(fit)/num - log(2*pi) # BIC
```

```
## [1] 4.771699
```

```r
tsplot(fit$residuals)
```




```
acf(fit$residuals)
```

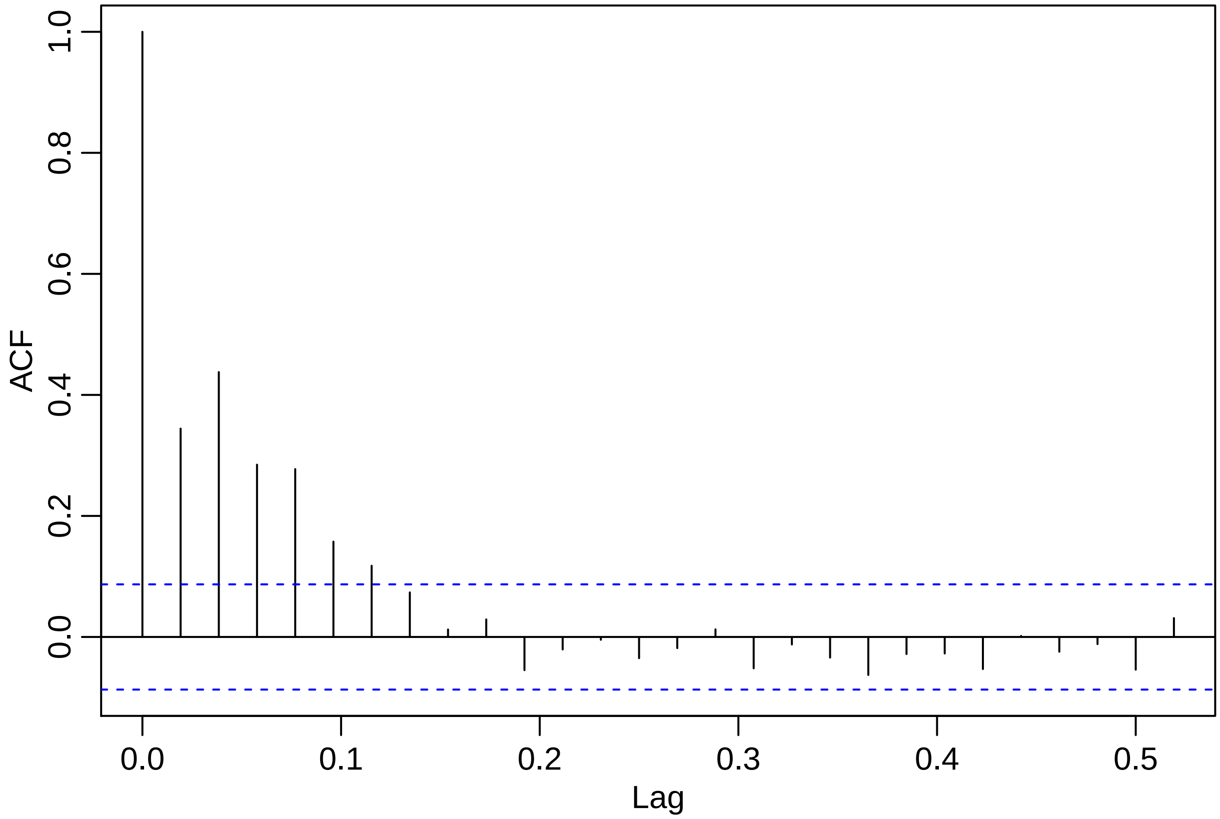


Table 3.2 *Summary Statistics for Mortality Models*

Model	k	SSE	df	MSE	R^2	AIC	BIC
(3.14)	2	40,020	506	79.0	.21	5.38	5.40
(3.15)	3	31,413	505	62.2	.38	5.14	5.17
(3.16)	4	27,985	504	55.5	.45	5.03	5.07
(3.17)	5	20,508	503	40.8	.60	4.72	4.77

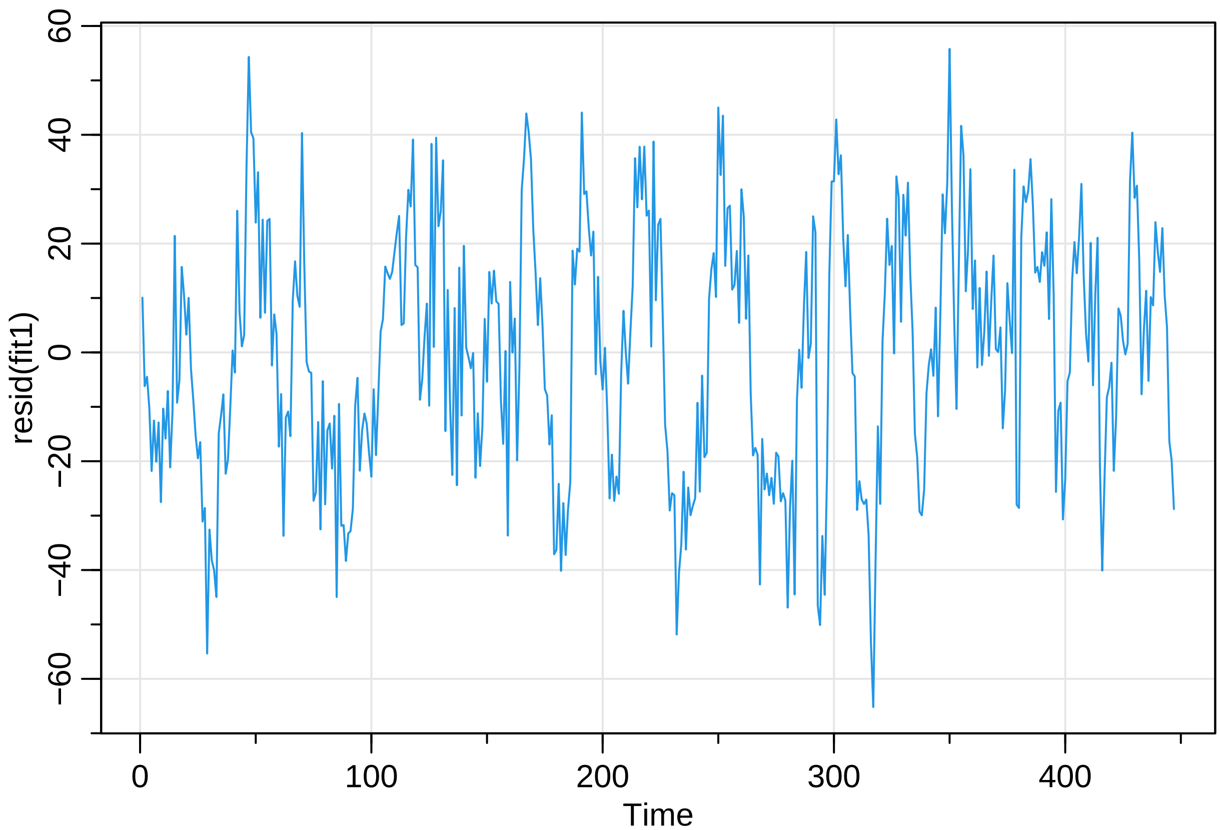
Figure 1: Table 3.2

```
fish = ts.intersect( rec, soiL6=lag(soi,-6) )
summary(fit1 <- lm(rec~ soiL6, data=fish, na.action=NULL))
```

```
##
## Call:
## lm(formula = rec ~ soiL6, data = fish, na.action = NULL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.187 -18.234   0.354  16.580  55.790
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   65.790      1.088   60.47   <2e-16 ***
## soiL6        -44.283      2.781  -15.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.5 on 445 degrees of freedom
## Multiple R-squared:  0.3629, Adjusted R-squared:  0.3615
## F-statistic: 253.5 on 1 and 445 DF,  p-value: < 2.2e-16
```

```
tsplot(resid(fit1), col=4) # residuals
```




```
##
#install.packages('dynlm')
library(dynlm)
summary(fit2 <- dynlm(rec~ L(soi,6)))
```

```
##
## Time series regression with "ts" data:
## Start = 1950(7), End = 1987(9)
##
## Call:
## dynlm(formula = rec ~ L(soi, 6))
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -65.187 -18.234   0.354  16.580  55.790
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   65.790      1.088   60.47   <2e-16 ***
## L(soi, 6)    -44.283      2.781  -15.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.5 on 445 degrees of freedom
## Multiple R-squared:  0.3629, Adjusted R-squared:  0.3615
## F-statistic: 253.5 on 1 and 445 DF,  p-value: < 2.2e-16
```

Exploratory Data Analysis

- To estimate autocorrelations with precision, we need to satisfy conditions of stationarity.
- Methods for coercing (transforming) nonstationary data to stationarity:

Detrend

detrend: remove the trend. First step in an exploratory analysis.

$$X_t = \mu_t + Y_t$$

,

- μ_t: trend
- Y_t: stationary process

Get a reasonable estimate of the trend component, $\hat{\mu}_t$, then work with the residuals

$$\hat{Y}_t = X_t - \hat{\mu}_t$$

If trend is random, differencing could be helpful.

Differencing

$$\nabla X_t = X_t - X_{t-1}$$

- Backshift operator

$$BX_t = X_{t-1}$$

log-transformation

Useful to equalize variance

$$Y_t = logX_t$$

$$\nabla X_t = (1 - B)X_t$$