W271 Summer 2022 Lecture Video Question Solutions Week 7

Contents

Week 7 Time Series Analysis Part 2

7.2 Classical Linear Regression: Revisit

Q: Write down the (statistical) consequences of the violation of each of the underlying assumptions of the classical linear regression model.

Solution: The assumptions of linear regression are:

(1) Linearity: The relationship between X and the mean of Y is linear

(2) Homoscedasticity: The variance of th residual is the same for any value of X

(3) Independence: Observationsare independent of each other

(4) Normality: For any fixed value of X Y is normally distrivuted

Violating assumptions (1) and (4) are not super serious.

Linearity is almost never true, and linear regression is a simplifying assumption. Furthermore, you can make the relationship nonlinear in the feature space by including various nonlinear transformations of the features as regressors.

Normality is similar in that it is almost never true but by the central limit theorem with large samples, it will hold for the mean response of Y given X, which is what linear regression is modeling.

Violating assumptions (2) and (3) are more serious. They lead to patterns in the residuals, which prevent the estimated coefficients in regression from being valid estimates and in turn making linear regression no longer the best unbiased linear predictor (BLUE), following the Gauss Markov Theorem.

However, while the classic linear regression model does not hold, there are transformations to the standard errors one can do to recover the optimal properties of linear regression, including unbiasedness and consistency. In particular, we can calculate heteroskedastic robust standard errors (or cluster robust standard errors in the case of grouping patterns) to adjust for violations of (3) homoskedasticity. For violations of (4) there are also standard errors that adjust for autocorrelation in the residuals such as Newey West standard

errors. Also, as we will see, we can sometimes do transformations of the input data to regression such as detrending or differencing to remove the autocorrleation and recover the independence assumption.

In short while violations of assumptions (2) and (3) make classical linear regression an invalid and suboptimal model, they can be adjusted for through alterations to the base regression framework.

7.3 Linear Time-Trend Regression

Q: When we shift the time index, (1) why is the interpretation of the slope coefficient not affected, and (2) why is the intercept affected?

Solution: Let's focus on a simple regression model with a just a time trend to keep things concrete, but the reasoning holds for the more general regression model.

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

With this $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{t}$ and $\hat{\beta}_1 = \frac{Cov(y_t,t)}{Var(t)}$.

Let's now use $t^* = t + \Delta$, which is a shifted time index.

$y_t = \beta_0 + \beta_1 t^* + \epsilon_t$

Hence:

$\hat{\beta}_1^* = \frac{Cov(y_t,t^*)}{Var(t^*)} = \frac{Cov(y_t,t+\Delta)}{Var(t+\Delta)} = \frac{Cov(y_t,t)+Cov(y_t,\Delta)}{Var(t)+Var(\Delta)} = \frac{Cov(y_t,t)}{Var(t)} = \hat{\beta}_1$

Intuitively, shifting the regression coefficient doesn't change things because the regression coefficient measures the marginal impact to the outcome variable for a unit change in the input. Shifting an input regressor cancels out when focusing on unit changes, meaning the slope coefficient is unchanged.

$\hat{\beta}_0^* = \bar{y} - \hat{\beta}_1 \bar{t^*} = \bar{y} - \hat{\beta}_1 (\bar{t + \Delta}) = \bar{y} - \hat{\beta}_1 \bar{t} - \hat{\beta}_1 \Delta = \hat{\beta}_0 - \hat{\beta}_1 \Delta$

Intuitively, shifting an input variable, changes the value when the input feature is zero. Since that is the point at which y equals the intercept, shifting input features will change the intercept.

Q: Take this estimated regression model and use regression diagnostics to examine if the underlying assumptions are satisfied.

Solution: The easiest regression diagnostic is to examine a residual vs. fitted plot. If the assumptions are satisfied, the residual vs. fitted plot should look like a random cloud that is consistent across the fitted values with no discernable trend. This would imply that the residuals are homeskedastic and there is a linear relationship with the outcome.

They can also show a lack of autocorrelation though this is more easily observed through an ACF plot of the residuals, which should exhibit no correlation across lags since the residauls are assumed to be independent across time.

```
#install.packages("astsa")
library(astsa)
```

```
## Warning: package 'astsa' was built under R version 4.1.1
```

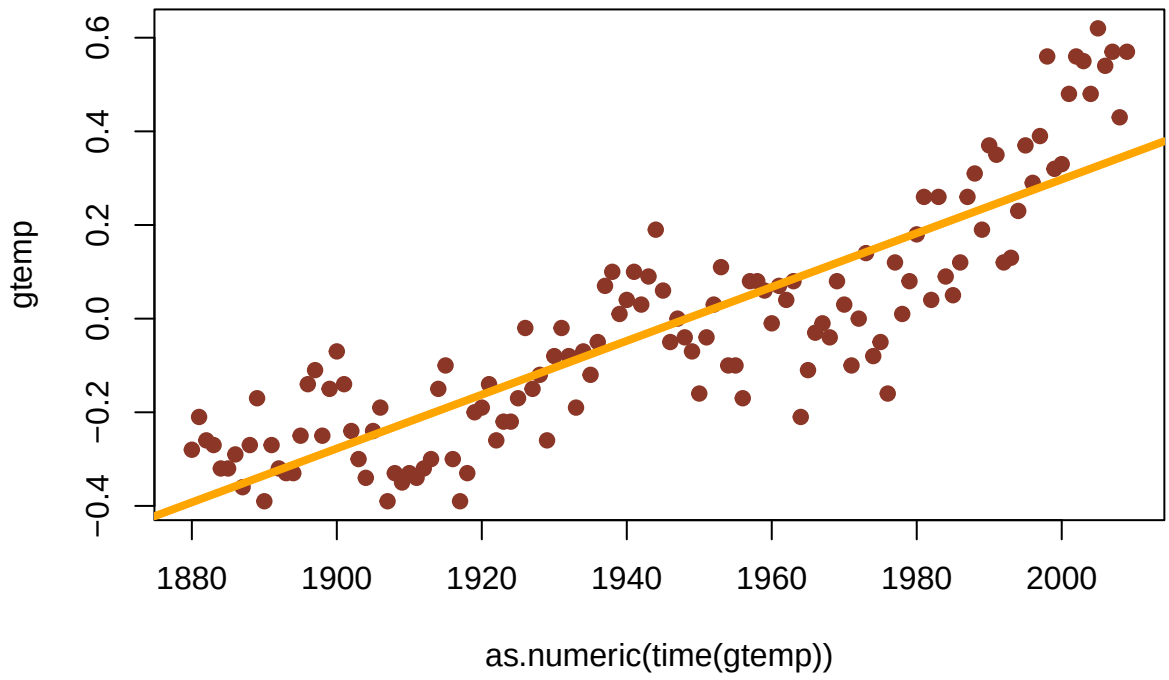
```r
data(gtemp)
head(gtemp)
```

```
## [1] -0.28 -0.21 -0.26 -0.27 -0.32 -0.32
```

```r
trend.model <- lm(gtemp ~ time(gtemp))
summary(trend.model)
```

```
##
## Call:
## lm(formula = gtemp ~ time(gtemp))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31946 -0.09722  0.00084  0.08245  0.29383
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.120e+01  5.689e-01  -19.69   <2e-16 ***
## time(gtemp)  5.749e-03  2.925e-04   19.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1251 on 128 degrees of freedom
## Multiple R-squared:  0.7511, Adjusted R-squared:  0.7492
## F-statistic: 386.3 on 1 and 128 DF,  p-value: < 2.2e-16
```

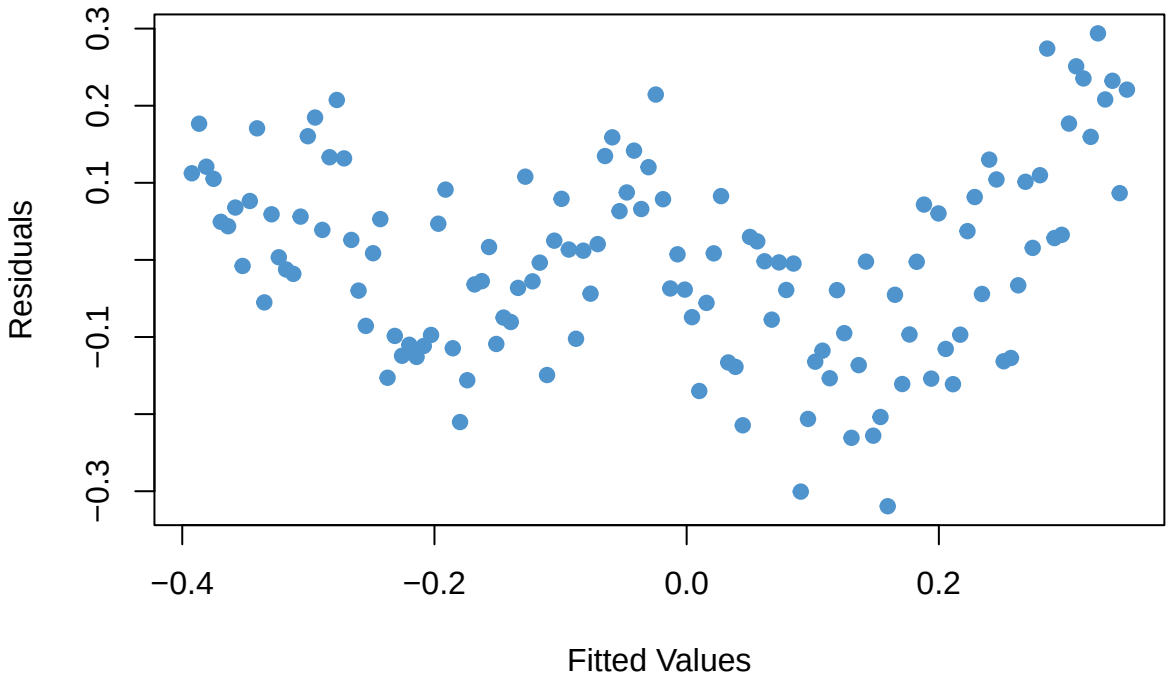
```r
plot(as.numeric(time(gtemp)), gtemp, type = "p", col = "tomato4", pch = 19)
abline(a  = trend.model$coefficients[1], b = trend.model$coefficients[2], lwd = 4, col = "orange1")
```



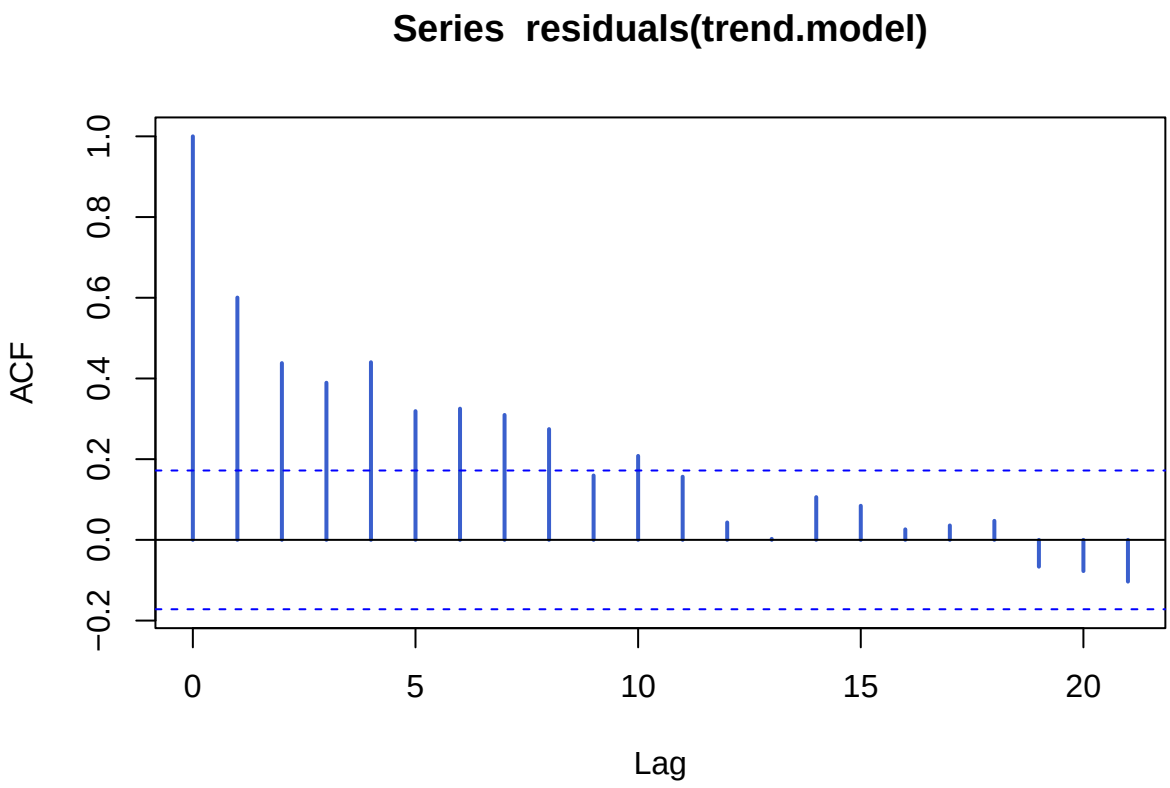
The residual vs fitted plot shows a clear nonlinear trend and also a trend with various fitted values, suggesting violations of linearity and homoskedasticity (though incorporating nonlinear terms may alleviate the heteroskedasticity).

The ACF plot shows some significant autocorrelation as well, suggesting that the residuals as is are not independent, violating another assumption of classical linear regression.

```
plot(predict(trend.model), residuals(trend.model), type = "p", pch = 19, col = "steelblue3",
     xlab = "Fitted Values", ylab = "Residuals")
```




```
acf(residuals(trend.model), lwd = 2, col = "royalblue3")
```



7.5 Time Series Regression, Example 2

Q: Estimate a model with lag SOI and use regression diagnostic to examine if the underlying assumptions are satisfied.

```
#you can find the data here: https://online.stat.psu.edu/stat510/book/export/html/690
soi <- read.table("~/Documents/Berkeley W271/Week 7 Time Series Analysis Part 2/soi.dat", sep = "\t")
recruit <- read.table("~/Documents/Berkeley W271/Week 7 Time Series Analysis Part 2/recruit.dat", sep =

soi <- ts(soi$V1)
recruit <- ts(recruit$V1)
fish <- ts.intersect(recruit, soiL6 = lag(soi, -6))
ts.model <- lm(recruit ~ soiL6, data = fish, na.action = NULL)
summary(ts.model)
```

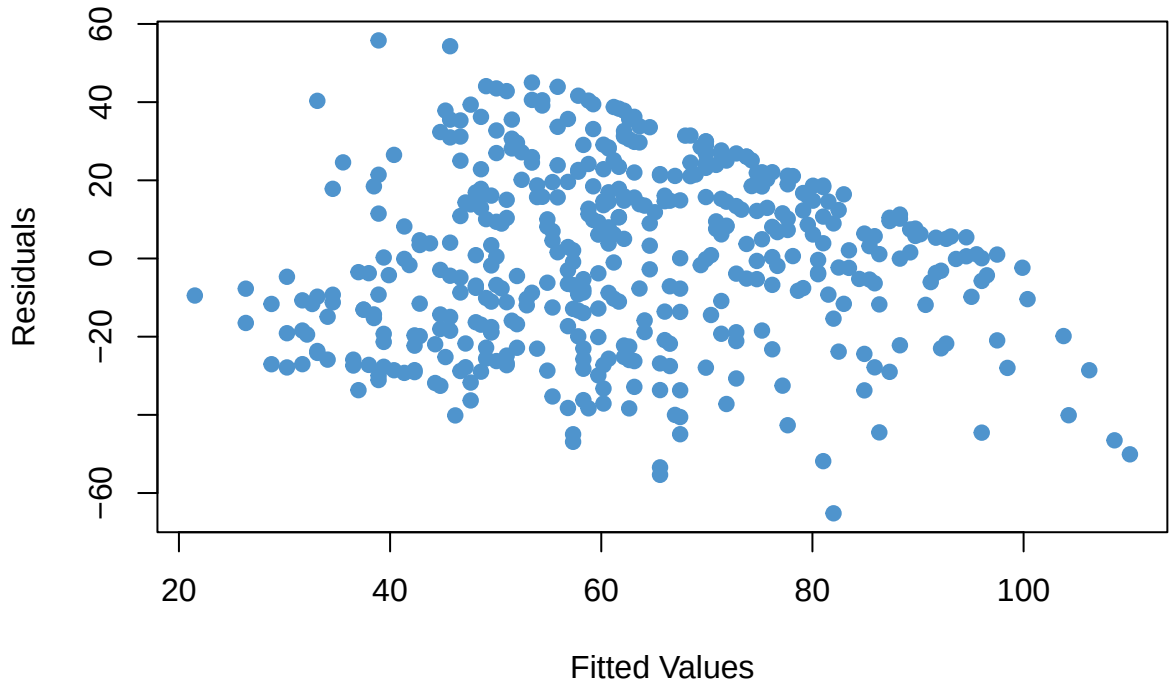
Solution:

```
##
## Call:
## lm(formula = recruit ~ soiL6, data = fish, na.action = NULL)
##
## Residuals:
```

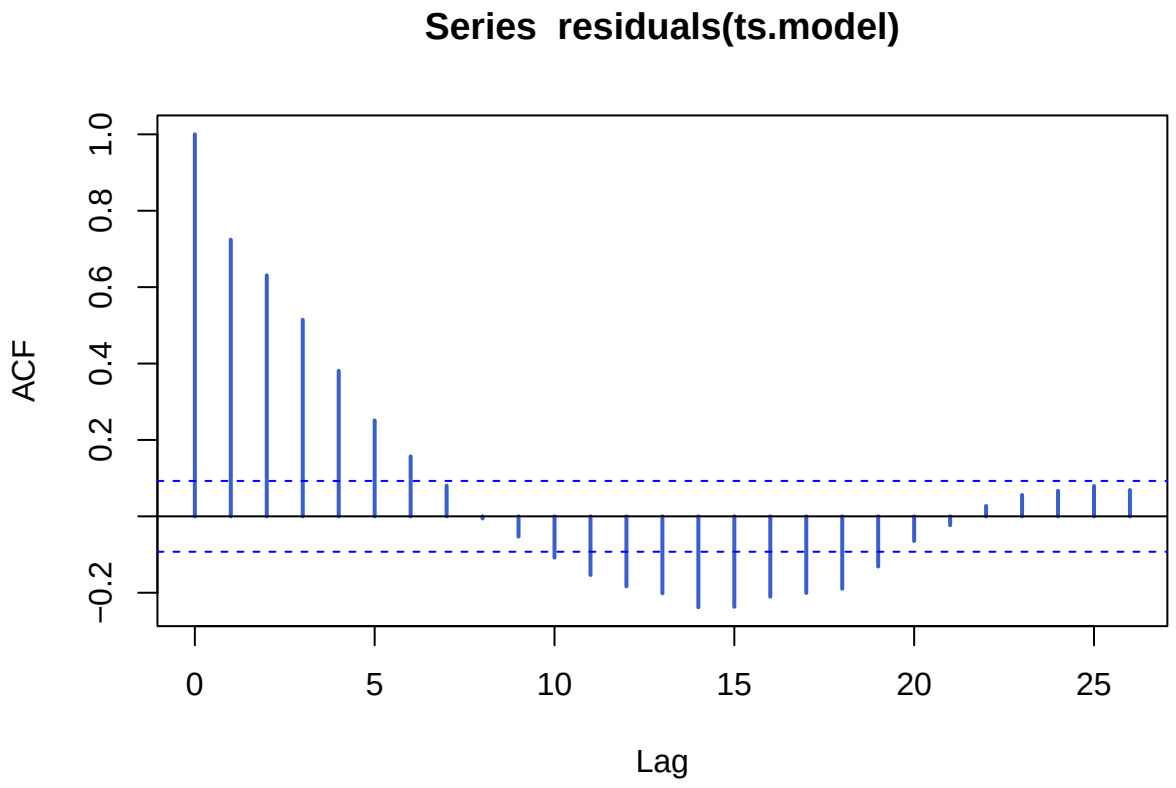
```
##     Min      1Q  Median      3Q     Max
## -65.187 -18.234   0.354  16.580  55.790
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   65.790      1.088   60.47   <2e-16 ***
## soiL6        -44.283      2.781  -15.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 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
```

The residual vs. fitted plot shows a clear trend (though it does not appear to be nonlinear), and the ACF plot shows both a general dependence across lags in the residuals as well as a somewhat season pattern of peaks and troughs. This violates homoskedasticity and independence of observations.

```
plot(predict(ts.model), residuals(ts.model), type = "p", pch = 19, col = "steelblue3",
     xlab = "Fitted Values", ylab = "Residuals")
```




```
acf(residuals(ts.model), lwd = 2, col = "royalblue3")
```



7.9 Autoregressive Models, Part 1A: Mathematical Formulation and Properties

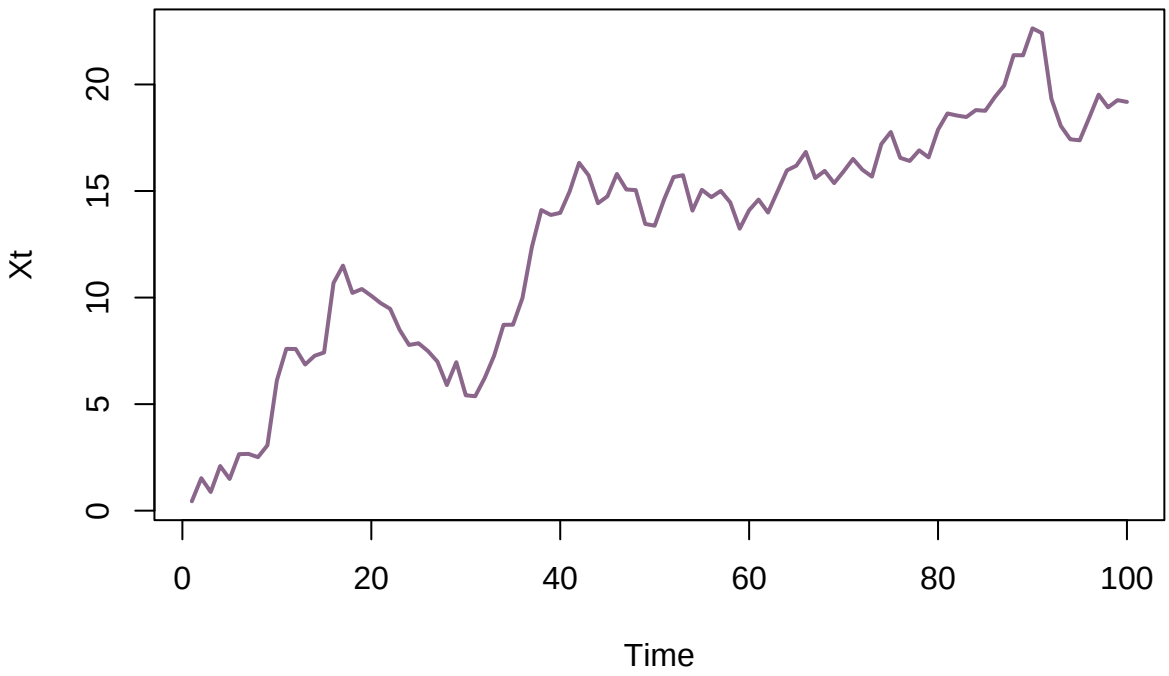
Q: Write an R program to simulate the series X_t under the condition that $\phi = 1$ and plot your simulated series.

Solution: If $\phi = 1$ then $x_t = \phi x_{t-1} + \omega_t = x_{t-1} + \omega_t$

This is a basic random walk, and we have what this looks like already. It will diverge over time as the variance explodes, resulting in a nonstationary time series.

```
phi <- 1
n <- 100
omega.var <- 1
drift <- 0
xt <- rnorm(n, mean = drift, sd = sqrt(omega.var))
xt <- cumsum(xt)

plot(1:n, xt, xlab = "Time", ylab = "Xt", col = "plum4", type = "l", lwd = 2)
```


Xt
0
5
10
15
20
0
20
40
60
80
100
Time