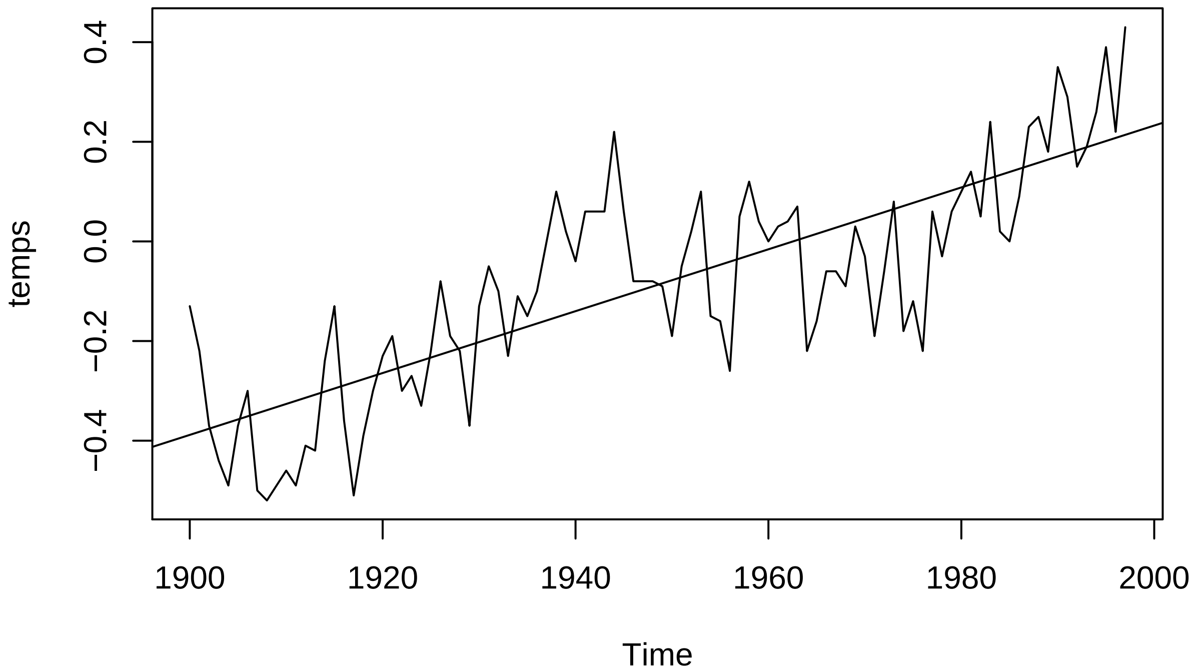
Stat 478: Final Exam, Part 2 Problem 1

Alex Towell (atowell@siue.edu)

2021-05-01

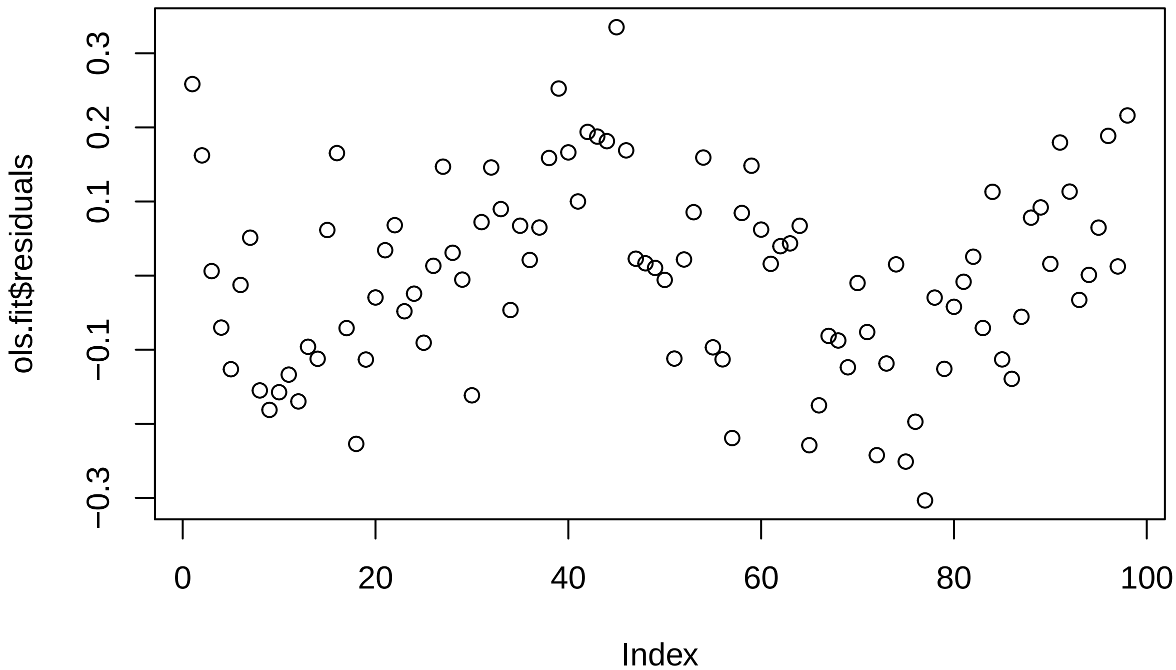
```r
#setwd("~/final_exam_478")

############################################################
# Consider the yearly global temperature data in period 1900-1997. The data set
# is given on blackboard. You may use the following to read in the data and make
# it a time series:
#  dt=read.table("your directory/globaltemps.txt", header=T)
#  temps=ts(dt$Temps, start=1900)
#############################################################
globaltemps.dt=read.table("globaltemps.txt", header=T)
temps=ts(globaltemps.dt$Temps, start=1900)

# preliminary analysis on the source of the data.
# we have a priori knowledge from scientific findings that the global
# average temperature is increasing, so we expect the data to show some
# positive trend. if the data ended up being ambiguous between, say, a random
# walk (without drift) and a deterministic trend, we would have a bias for the
# deterministic trend, or maybe an arima model with drift.

#############################################################
# part (a)
# Fit a simple linear regression model to the data, where y t is the yearly
# global temperature x t is time. Report the ANOVA table and summary for the
# model coefficients. Plot of the data with the least squares regression line
# overlaid.
#############################################################
ols.fit=lm(temps~time(temps),data=temps)
plot(temps)
abline(lm(temps ~ time(temps)))
```



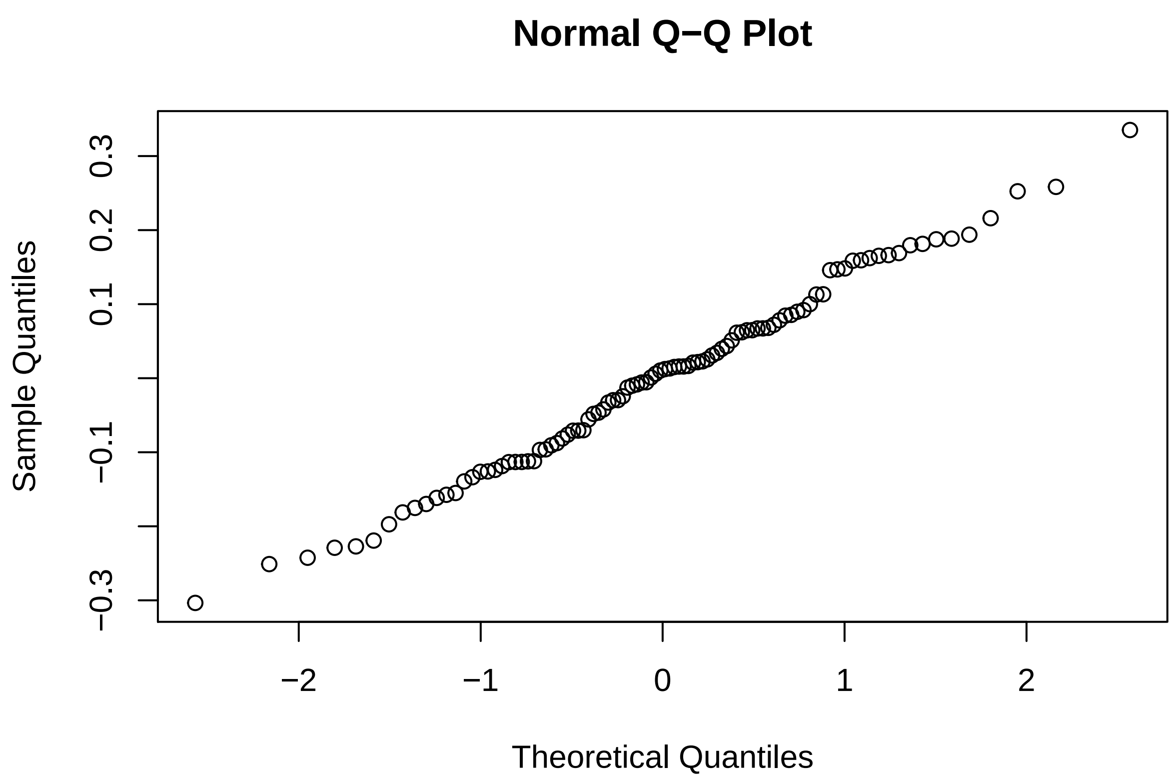

```
summary(ols.fit)
```

```
##
## Call:
## lm(formula = temps ~ time(temps), data = temps)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30352 -0.09671  0.01132  0.08289  0.33519
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.219e+01  9.032e-01  -13.49   <2e-16 ***
## time(temps)  6.209e-03  4.635e-04   13.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1298 on 96 degrees of freedom
## Multiple R-squared:  0.6515, Adjusted R-squared:  0.6479
## F-statistic: 179.5 on 1 and 96 DF,  p-value: < 2.2e-16
```

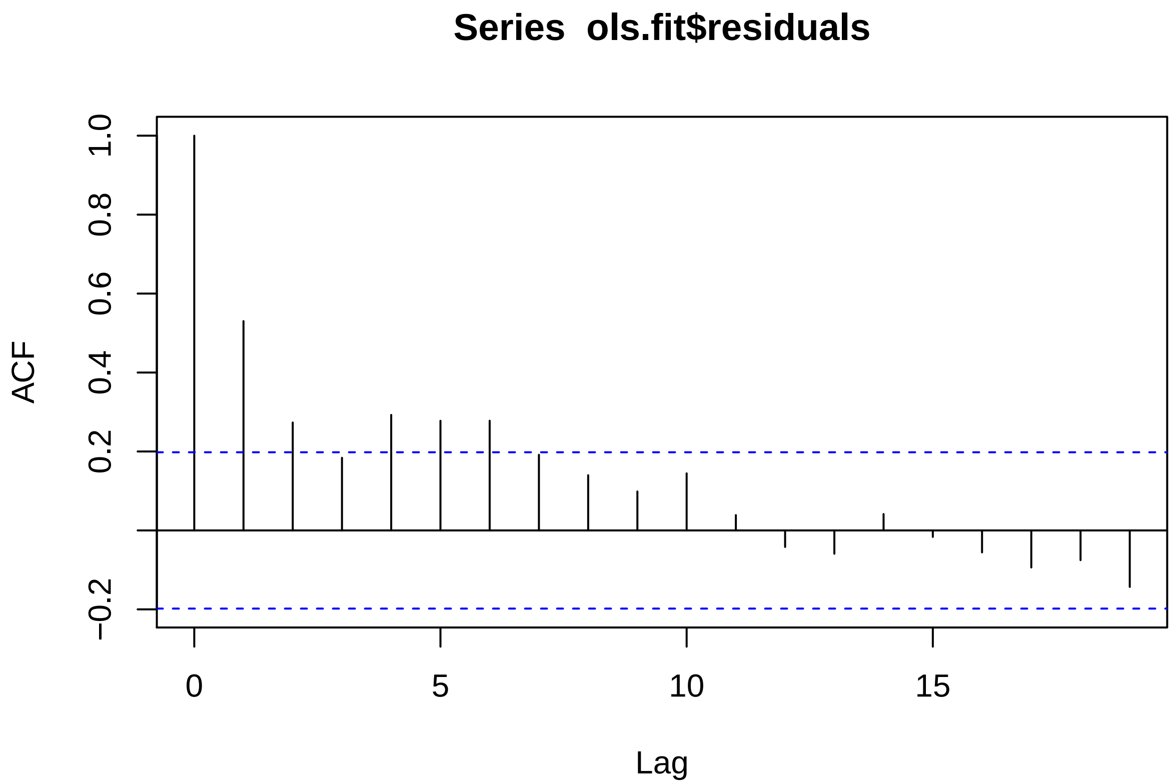
```
#############################################################
# part (b)
# Examine the residuals from your fitted model for normality and independence.
# Display the sample ACF. Do the residuals look to resemble a normal, zero mean
# white noise process?
#############################################################
plot(ols.fit$residuals)
```




```
qqnorm(ols.fit$residuals)
```




```
acf(ols.fit$residuals)
```

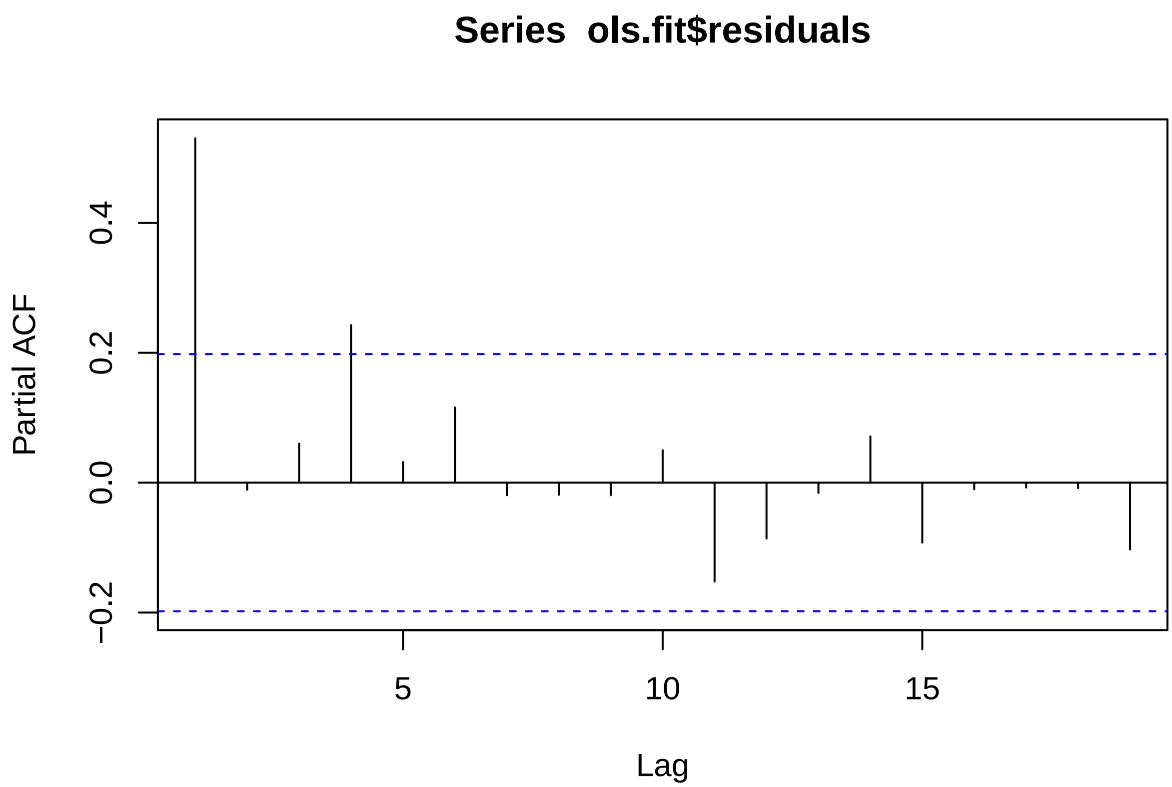



```
pacf(ols.fit$residuals)

# comments:
# the qq-plot supports a gaussian process. however, the residual plot
# seems to have a wave-like pattern and the ACF/PACF of the residuals
# indicate autocorrelation.
#
# maybe there is some unaccounted for seasonality in the data, e.g., global
# average temperatures are a very dynamic, complicated process affected by
# many potential covariates. so, the source of the autocorrelation may be a
# missing important covariate in the model.
#
# alternatively, the correlations in the residuals may be a result
# of a functional misspecification. i believe the trend is positive, but it
# is probably not a simple linear trend. perhaps a linear trend with an added
# seasonality component, or maybe the overall trend is just non-linear.
#
# it could also be the case that the assumption of uncorrelated error terms is
# violated, or some combination of all of the above.

##############################################################
# part (c)
# Conduct a Durbin-Watson test on the residuals. Comment on your conclusion
##############################################################
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```




```r
dwtest(ols.fit)
```

```
##
##  Durbin-Watson test
##
## data:  ols.fit
## DW = 0.86922, p-value = 1.783e-10
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
#dwtest(temps~time(temps))

# comments:
# H0: residuals have no autocorrelation
# p-value ~ .000. reject H0 at that p-value.
# we reject the null hypothesis that there is no autocorrelation.

##############################################################
# part (d)
# Use one iteration of the Cochrane-Orcutt procedure to estimated the regression
# coefficients. Also calculate the standard errors of the coefficients. Are the
# standard errors (from the Cochrane-Orcutt procedure) larger than the ones from
# simple linear regression?
##############################################################

# calculte phi fot the Cochrane Method
N=length(temps)
```

```r
phi.hat=lm(ols.fit$residual[2:N]~0+ols.fit$residual[1:(N-1)])$coeff
# transform y and x according to the Cochrane Method
y.trans=temps[2:N]-phi.hat*temps[1:(N-1)]
x.trans=time(temps)[2:N]-phi.hat*time(temps)[1:(N-1)]
# fit OLS regression with transformed data
coch.or=lm(y.trans~x.trans)
summary(coch.or)
```

```
##
## Call:
## lm(formula = y.trans ~ x.trans)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.267893 -0.069781  0.009598  0.069928  0.238791
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.0061132  0.7532273  -7.974 3.42e-12 ***
## x.trans      0.0067438  0.0008508   7.927 4.30e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1065 on 95 degrees of freedom
## Multiple R-squared:  0.3981, Adjusted R-squared:  0.3918
## F-statistic: 62.83 on 1 and 95 DF,  p-value: 4.298e-12
```

```r
# comments:
# larger.  The standard error of the estimate of beta1 from Cochrane procedure
# is .0008508 while the one from OLS is 4.635E-04.

##############################################################
# part (e)
# Instead of using a deterministic trend model, consider a model from the
# ARIMA(p, d, q) family.
# Choose a potential model and explain/defend your selection. Fit the model of
# your choice to the data
# and write out the the full model with estimated parameters.
##############################################################

library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

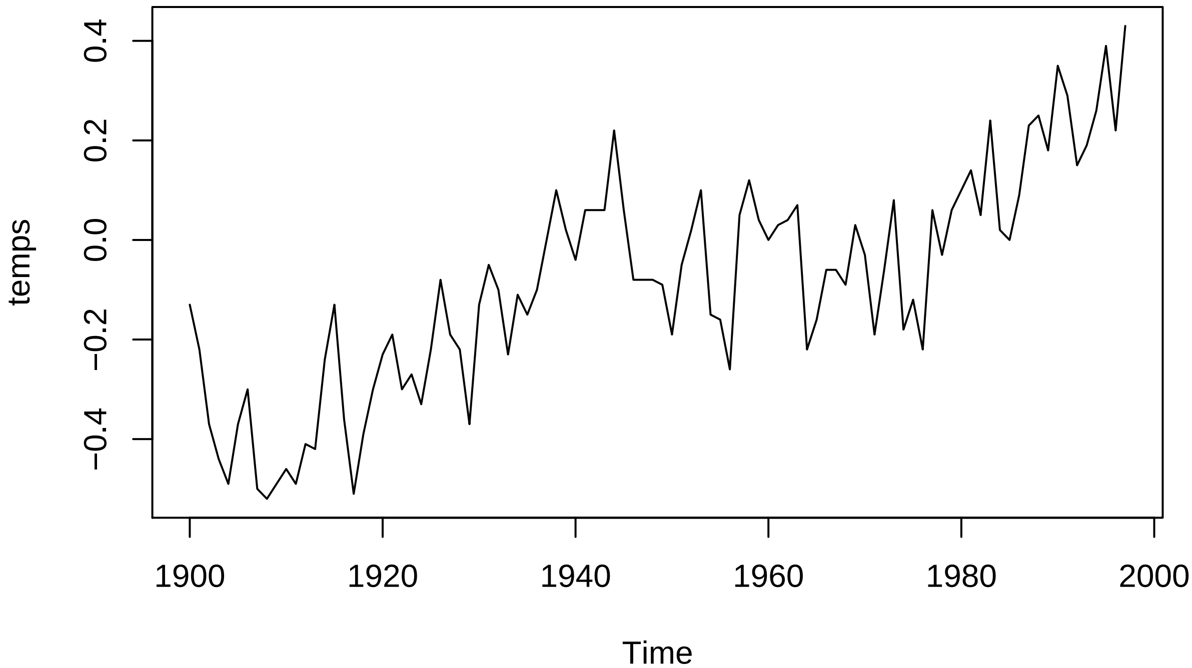
```r
library(TSA)
```

```
## Registered S3 methods overwritten by 'TSA':
##   method       from
##   fitted.Arima forecast
##   plot.Arima   forecast

##
## Attaching package: 'TSA'
```

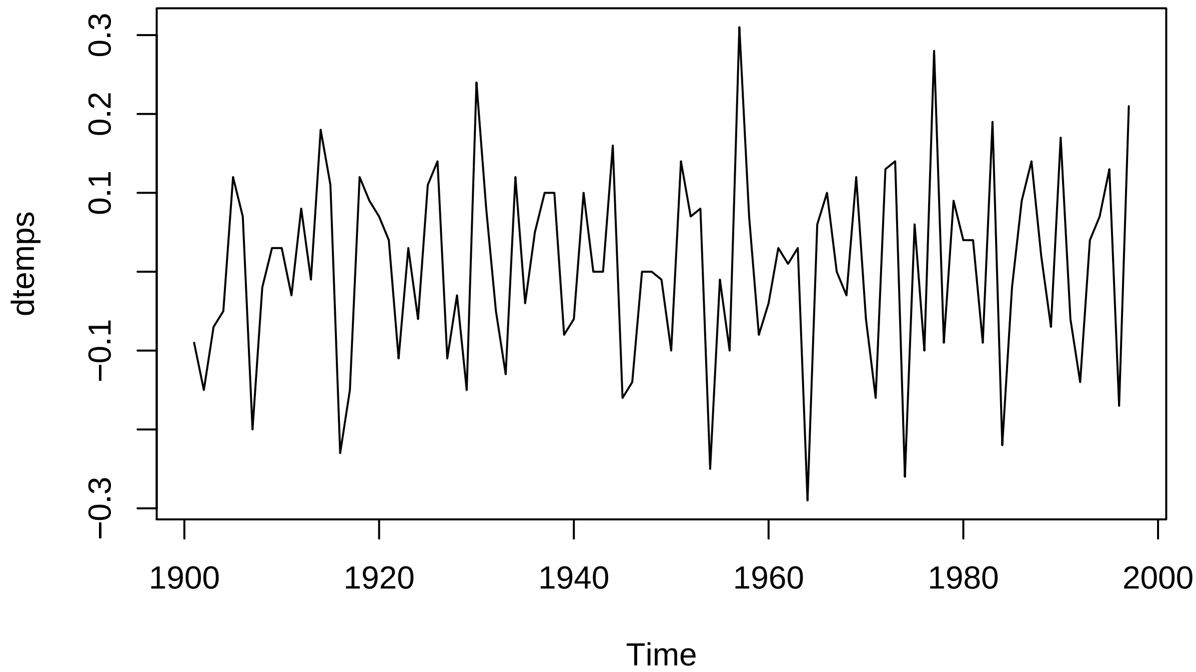
```
## The following objects are masked from 'package:stats':
##
##     acf, arima
```

```
## The following object is masked from 'package:utils':
##
##     tar
```

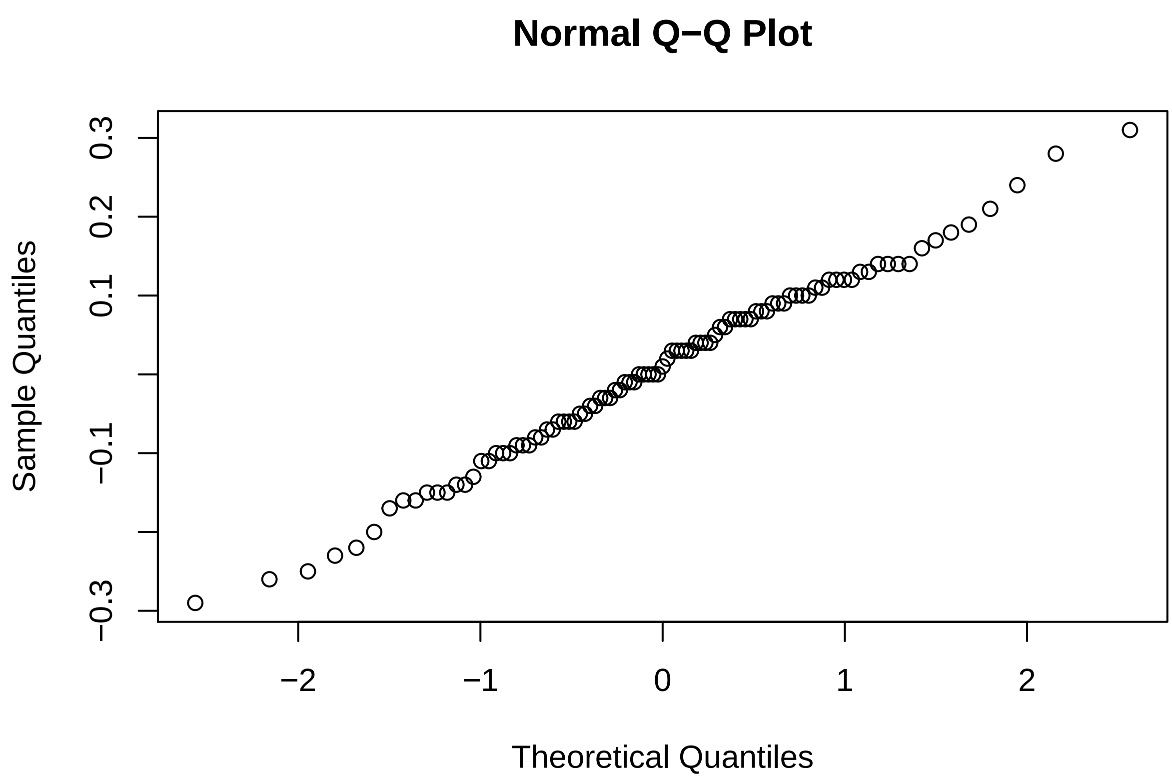
```
plot(temps)
```



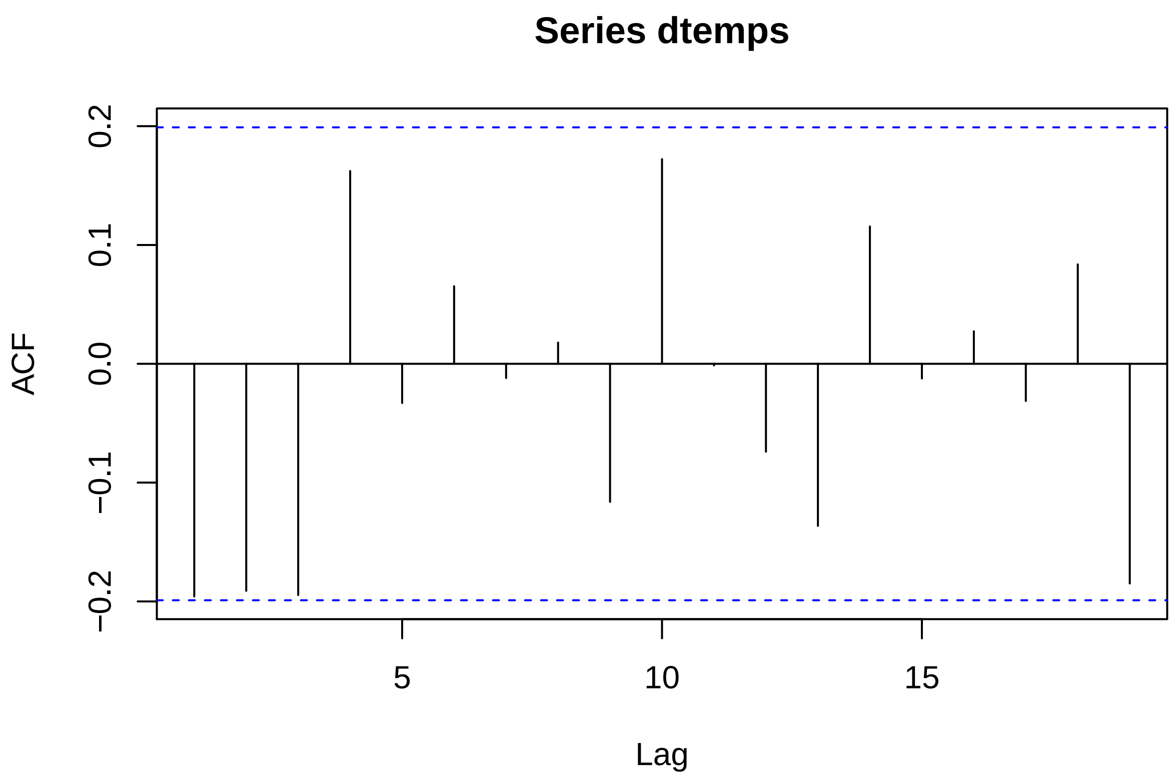

```
# not stationary, taking difference.
dtemps=diff(temps)

plot(dtemps)
```



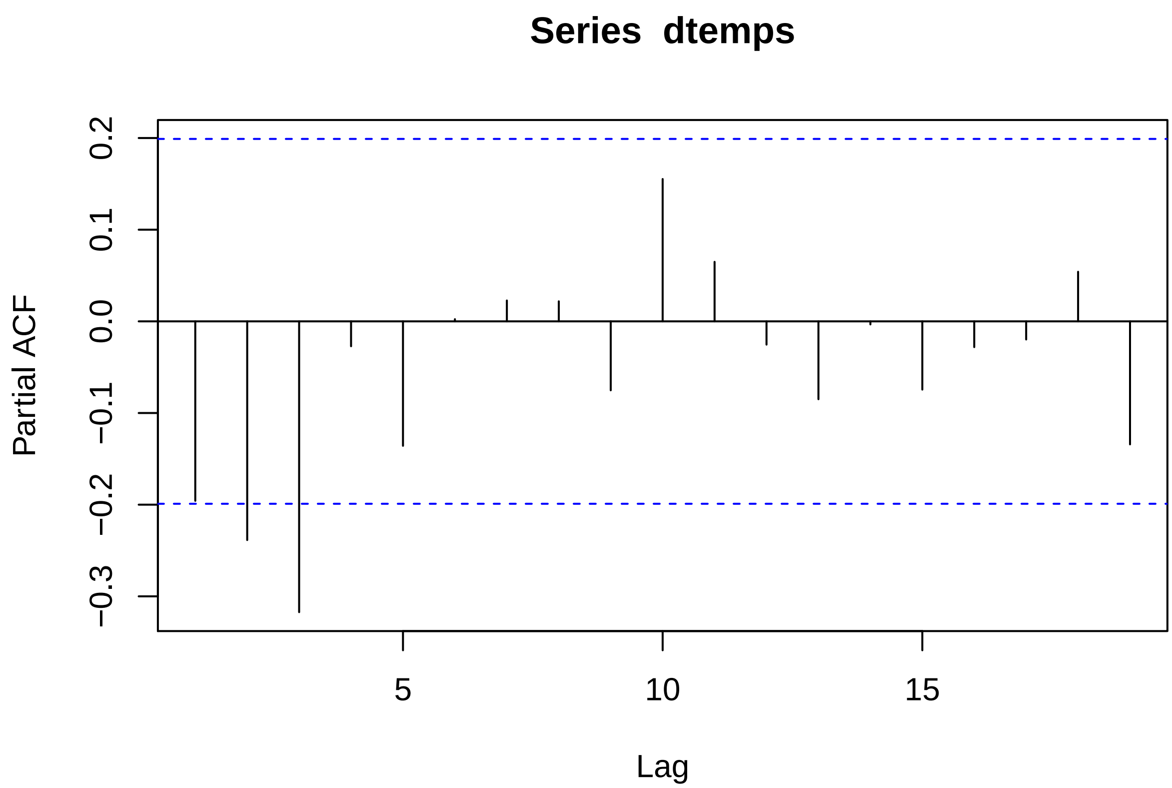

```
qqnorm(dtemps)
```




```
# the plot of the differenced process looks reasonably stationary,
# although i couldn't say if it was white noise.

acf(dtemps)
```

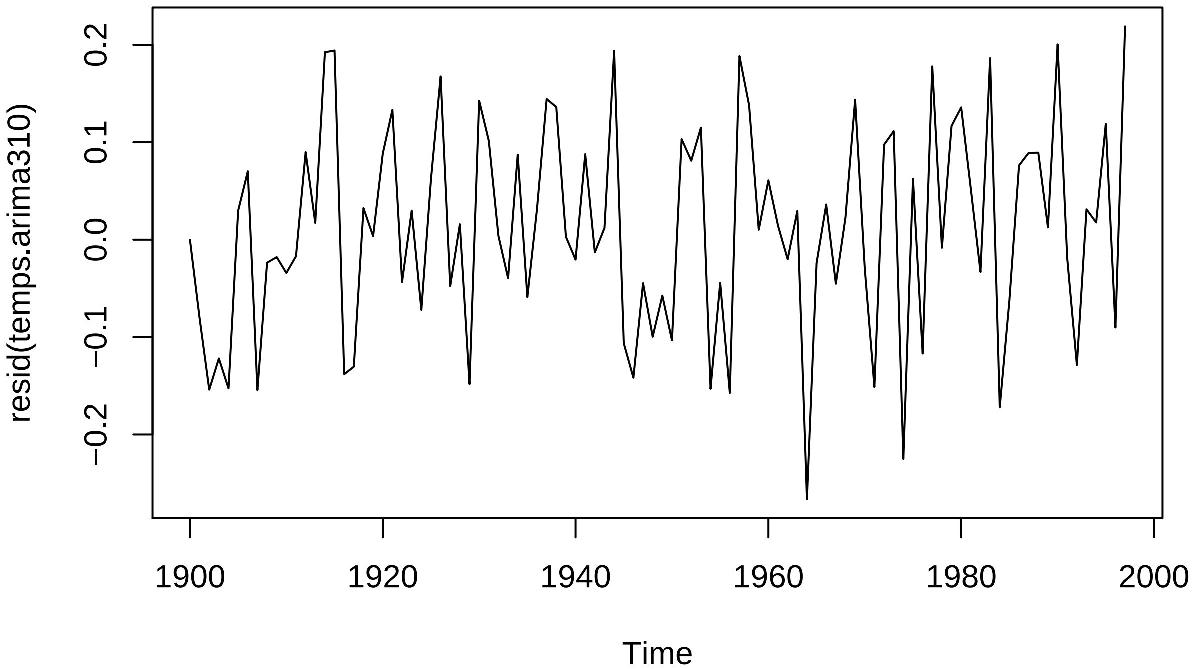



```
pacf(dtemps)
```

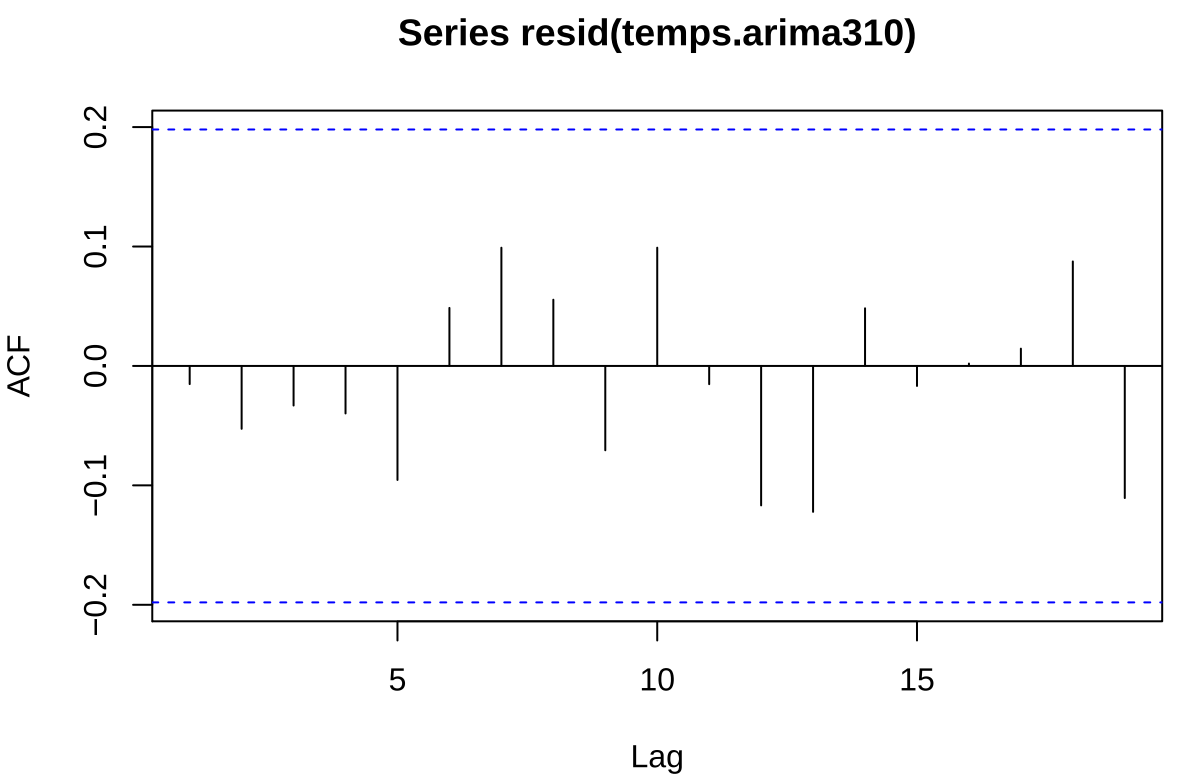



```
# it looks like there are autocorrelations in the data. the acf seems
# to support no MA components and the PACF seems to support ARIMA(0,1,3).
# so, preliminary conclusion: arima(p=0,d=1,q=3).
```

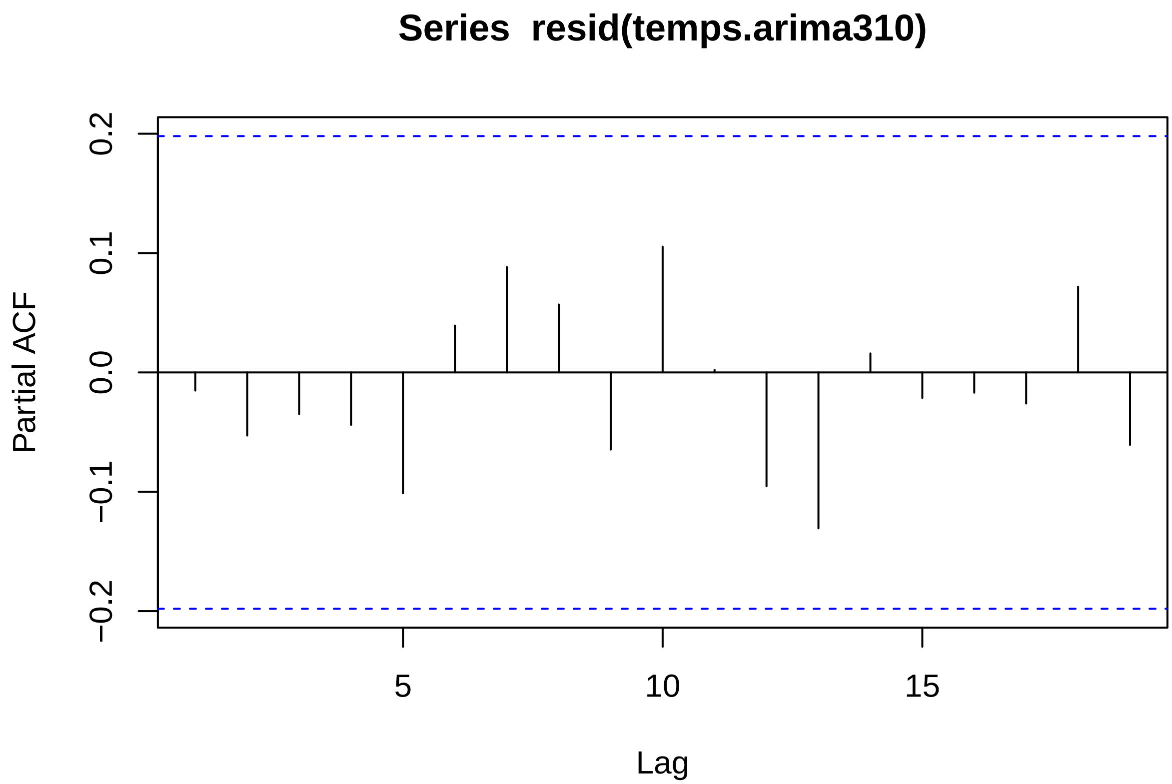
```
# let's fit that model to the data and do some tests
temps.arima310=Arima(temps,order=c(3,1,0))

# let's see if the evidence that the residuals model white noise
plot(resid(temps.arima310))
```



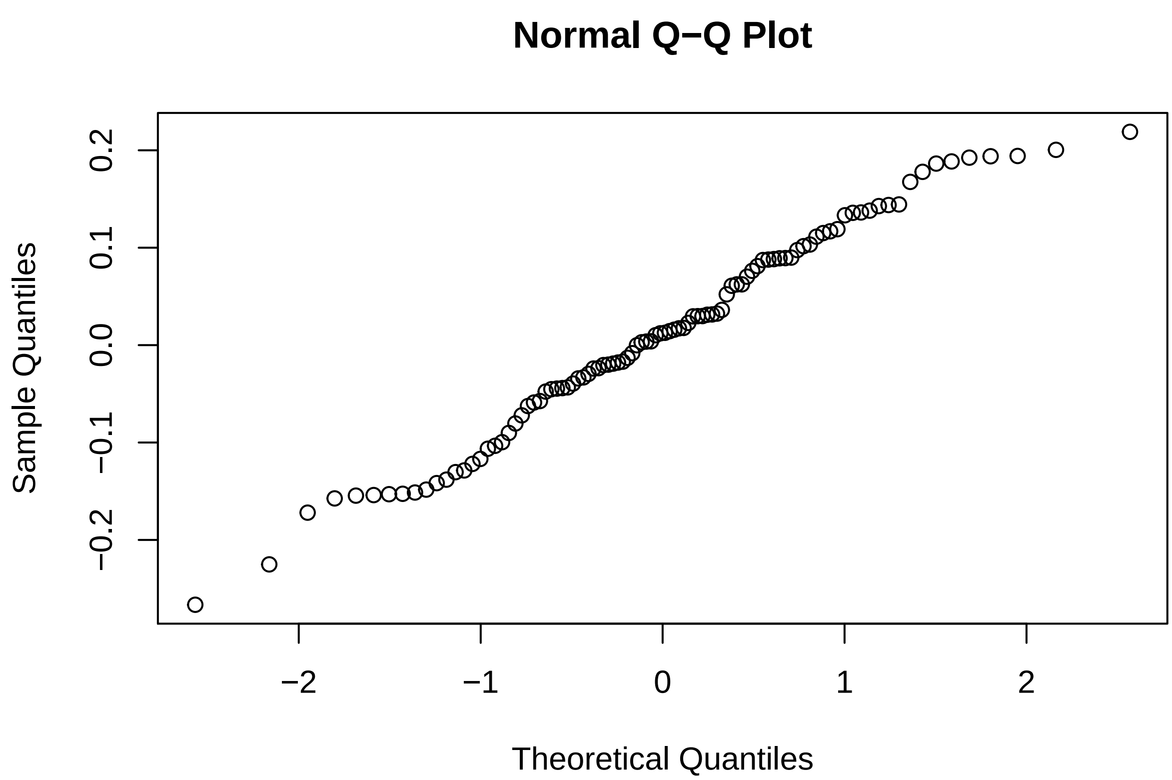

```
acf(resid(temps.arima310))
```



```r
pacf(resid(temps.arima310))
```




```r
# these plots look promising. no clear correlations in any of the above
# plots, suggesting that the residuals are uncorrelated.

qqnorm(resid(temps.arima310))
```




```
# the qq-plot looks somewhat gaussian, although that is not strictly
# necessary. we just want it to be a zero mean white noise process.

# let's look at a more objective test statistic, the Ljung-Box test.
Box.test(resid(temps.arima310,fitdf=3),lag=10,type="Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  resid(temps.arima310, fitdf = 3)
## X-squared = 4.8295, df = 10, p-value = 0.9023
```

```
# H0: the residuals of the model are independently distributed (0 correlation)
#
# this test produces a very p-value (.9), which we take to be very strong
# evidence in support of H0. we say the test statistic is compatible with the
# hypothesis of the residuals being independently distributed.

# out of curiosity, let's use auto.arima to have it select an ARIMA model
# based on the minimum AIC measure.
auto.arima(temps)
```

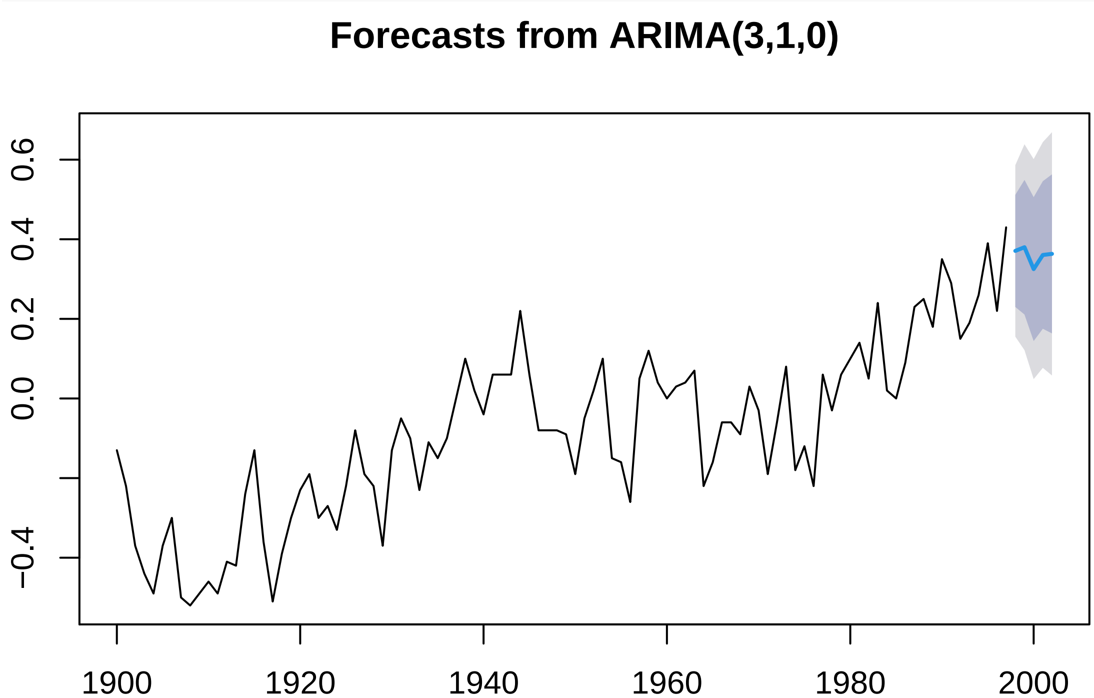
```
## Series: temps
## ARIMA(3,1,0)
##
## Coefficients:
##           ar1      ar2      ar3
##       -0.3351  -0.3245  -0.3388
## s.e.   0.0978   0.0979   0.0988
##
```

```
## sigma^2 estimated as 0.01207:  log likelihood=77.85
## AIC=-147.69   AICc=-147.26   BIC=-137.39
```

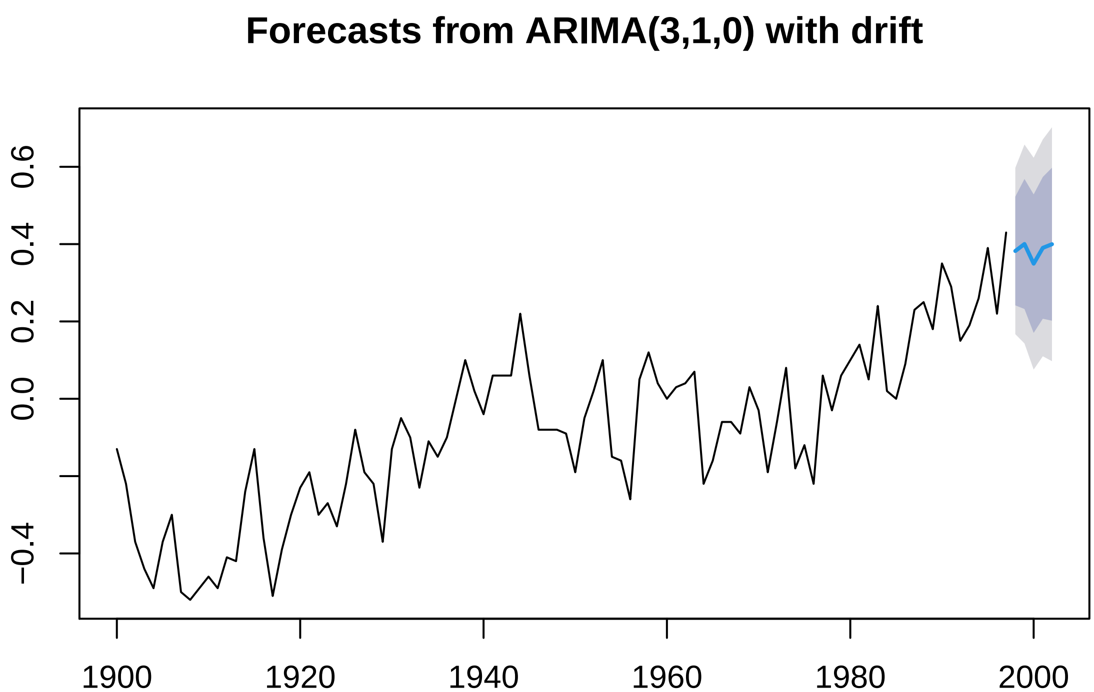
```
# it chose the same model, ar(3). we conclude that the ARIMA(3,1,0) model
# is a reasonable way to model the data in temps.
# the estimated paramters of the model are given by
coef(temps.arima310)
```

```
##        ar1        ar2        ar3
## -0.3351243 -0.3245378 -0.3387609
```

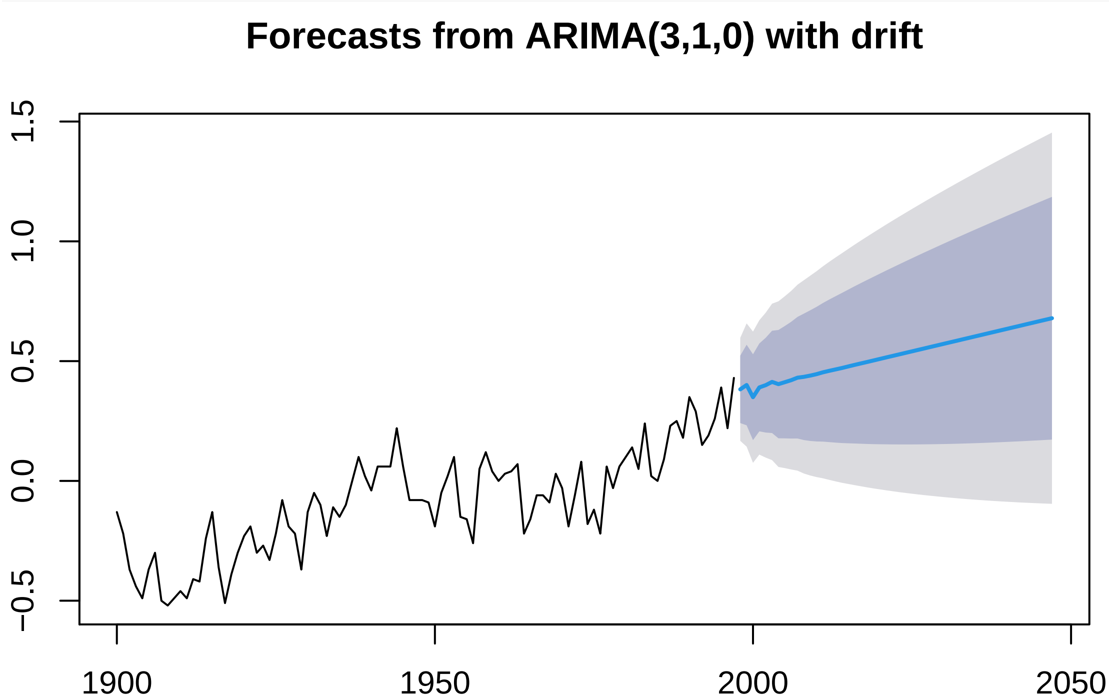
```
# theta=(-0.3351243,-0.3245378,-0.3387609)
# we can simulate drawing data the process with:
#     arima.sim(n=200,model=list(ar=c(-0.3351243,-0.3245378,-0.3387609)))

############################################################
# part (f)
# Use your model to forecast the global temperature for 1998-2003. Plot your
# forecast along with the prediction intervals.
############################################################
plot(forecast(temps,model=temps.arima310,h=5))
```



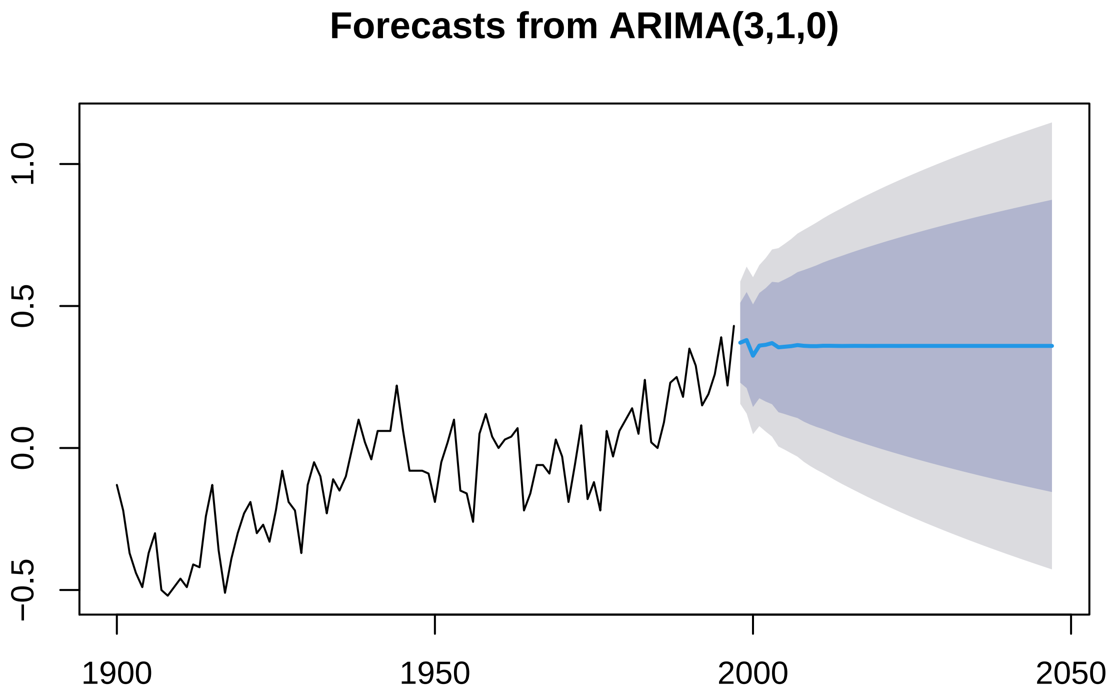

```
############################################################
# final comments
############################################################
# i have reservations about my chosen arima model. it's probably good for
# short-term forecasting, but it fails to model what i believe is an increasing
# average.

# let me try fitting the model to an arima model with drift.
temps.autoarima.d=auto.arima(temps,include.mean=T)
plot(forecast(temps.autoarima.d,h=5))
```




```
# this may be more realistic. out of curiosity, i forecast further ahead.

plot(forecast(temps.autoarima.d,h=50))
```




```
plot(forecast(temps.arima310,h=50))
```




```
# the drifting may be excessive here, overfitting to the data, although i would
# be curious to see if test data, if it were held out, would be supported by
# this forecast more than the alternative without drifting.
#
# anyway, all of this seems to call for a more indepth analysis with more
# attention paid to potential covariates and background knowledge about
# scientific findings.
```