The p-value for the Box–Pierce test is $p = 0.995$. The p-value for the Ljung–Box test is $p = 0.935$. We fail to reject H_0 for both tests. There is not enough statistical evidence to reject independence of the residuals. Prediction intervals for forecasted values of a time series require independence of the residuals, as well as normality. The p-value for the Shapiro–Wilk test for normality is calculated with

```
> shapiro.test(zhat)$p.value
[1] 0.9397434
```

which indicates that we can proceed as if the residuals are normally distributed. This is consistent with the shape of the histogram of the residuals. We can therefore proceed with forecasting for the AR(2) model based on the approximate independence and normality of the residuals.

Forecasting the level of Lake Huron for the next ten years. For the time series of Lake Huron levels $x_1, x_2, \ldots, x_{98}$, we would like to forecast lake levels into the future and give approximate 95% prediction intervals for the forecasted values assuming that the time series arises from a shifted AR(2) model with parameters estimated by maximum likelihood.

The first command in the R code below uses the `ar` function to estimate the parameters in the shifted AR(2) model via maximum likelihood estimation. The `predict` function then calculates the forecasted values and associated standard errors for the next five years. These standard errors can be used to calculate approximate 95% prediction interval limits.

```
> model = ar(LakeHuron, order.max = 2, aic = FALSE, method = "mle")
> predict(model, n.ahead = 5)        # predicted values and standard errors
$pred
Time Series:
Start = 1973
End = 1977
Frequency = 1
[1] 579.7889 579.5929 579.4310 579.3110 579.2261

$se
Time Series:
Start = 1973
End = 1977
Frequency = 1
[1] 0.6919689 1.0001831 1.1566954 1.2326978 1.2686176
```

The results are summarized in the table below. Notice that the forecasts trend monotonically toward $\bar{x} = 579$ feet and the standard errors increase as the time horizon increases. The increasing standard error is consistent with having less precision in the forecast as the time horizon increases.

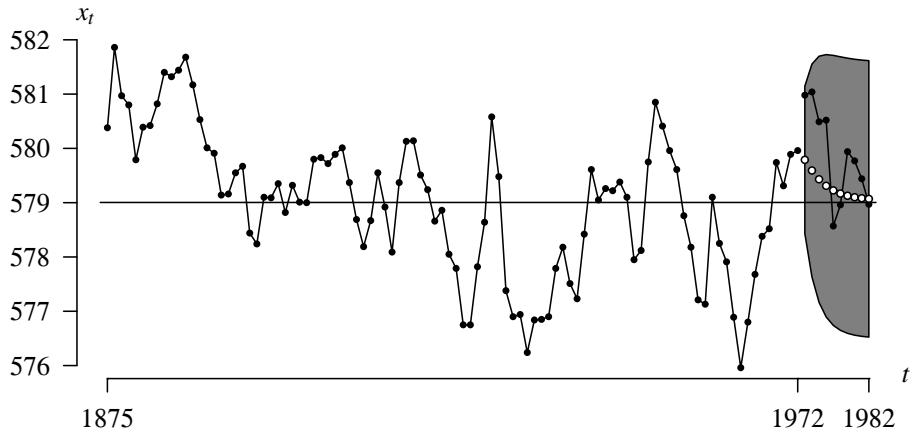
Time	$t = 99$	$t = 100$	$t = 101$	$t = 102$	$t = 103$
Year	1973	1974	1975	1976	1977
Forecast	579.79	579.59	579.43	579.31	579.23
Standard error	0.692	1.000	1.157	1.233	1.269
Lower prediction bound	578.43	577.63	577.16	576.89	576.74
Upper prediction bound	581.15	581.55	581.70	581.73	581.71

By increasing the `n.ahead` argument in the call to the `predict` function to 10, we can forecast the Lake Huron levels for the subsequent ten years. The figure that follows shows

- the original time series $x_1, x_2, \ldots, x_{98}$ as points (•) connected by lines,
- the first 10 forecasted lake levels $\hat{X}_{99}, \hat{X}_{100}, \ldots, \hat{X}_{108}$ as open circles (○),
- the 95% prediction intervals as a shaded region, and
- the next 10 actual average lake level values in July for the years 1973–1982 taken from the NOAA Great Lakes Experimental Research Laboratory website (provided by Lauren M. Fry),

  580.98, 581.04, 580.49, 580.52, 578.57, 578.96, 579.94, 579.77, 579.44, 578.97,

  as points (•) connected by lines.



There are four key observations concerning the figure.

- Even though the final three observations in the Lake Huron water level time series, which are $x_{96} = 579.31$, $x_{97} = 579.89$, and $x_{98} = 579.96$, show an increasing trend, the forecasts monotonically approach $\hat{\mu} = \bar{x} = 579$. The reason that the forecasts approach $\hat{\mu} = \bar{x} = 579$ in a damped exponential fashion is that we have assumed a stationary model which has long-term expected value μ.
- The widths of the prediction intervals increase as the forecasting time horizon h increases. These widths do not increase indefinitely, but rather approach a limit as $h \to \infty$.
- The random sampling variability in the observed time series values $x_1, x_2, \ldots, x_{98}$ is not apparent in the forecasted values $\hat{X}_{99}, \hat{X}_{100}, \ldots, \hat{X}_{108}$. Observed time series values tend to exhibit the typical random sampling variability; forecasted values for a stationary shifted AR(2) time series model tend to be smooth.
- The first actual value in the forecast region, $x_{99} = 580.98$ for the year 1973, nearly falls outside the associated 95% prediction interval. Even if the AR(2) model is a good fit for this time series, there is still a probability of approximately 0.05 that a future observation will fall outside the associated 95% prediction interval.

This concludes the brief introduction to the use of R in statistics. The `help` function, along with a statistics textbook are the best guides to learning all of the computational power that R has to offer in this area. The table that follows summarizes the functions introduced in this chapter.

function	description
ks.test	perform a Kolmogorov–Smirnov test
shapiro.test	perform a Shapiro–Wilk test
t.test	perform a t-test
aggregate	compute summary statistics for groups
var.test	perform an F test for comparing variances
aov	perform an analysis of variance
TukeyHSD	perform Tukey's honest significant difference test
lm	fit a linear model
coef	extract coefficients
predict	predict a value from a fitted statistical model
vcov	calculate a sample variance–covariance matrix
cor.test	perform a correlation test
I	inhibit interpretation
curve	draw a curve associated with a function
nls	fit a nonlinear least squares model
glm	fit a generalized linear model
layout	create a customized layout for multiple frame plots
plot.ts	plot a time series
acf	compute and plot a sample autocorrelation function
pacf	compute and plot a sample partial autocorrelation function
ar	fit an autoregressive model
arima	fit an ARIMA model
Box.test	perform Box–Pierce or Ljung–Box test for independence

The next chapter re-introduces matrices. Rather than considering a matrix as simply an elementary data structure in R, this time mathematical operations on a matrix are emphasized.

Exercises

26.1 The R code below conducts a Monte Carlo simulation that plots the estimated probability density function (histogram) of the p-value associated with the hypothesis test concerning the equality of the means of two standard normal populations based on $n = 3$ observations from each population.

```
> p = replicate(100000, t.test(rnorm(3), rnorm(3))$p.value)
> hist(p)
```

(a) Execute the code and comment on the results.

(b) Type `help(t.test)` to see if these commands can be modified to produce a $U(0, 1)$ distribution for the p-values. Implement any proposed alterations.

(c) Change one of the population distributions to a $N(1, 1)$ distribution. Comment on the results.

(d) Change the population distributions from standard normal to unit exponential (an exponential distribution with rate 1). Comment on the results.

26.2 Fit a simple linear regression model for the `cars` data set that forces the regression line to pass through the origin. Plot the data pairs and the associated fitted regression line.