Statistics 435/535/711 16 March 2022
P. Shaman

On R^2, Adjusted R^2, Mean Squared Error, Penalty Factors, and Model Selection

Consider the regression model

(1) $$y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t .$$

There are k explanatory variables, and we require the disturbance term ε to satisfy assumptions (i)–(iii) cited on page 3 of the 12 January notes. Assume we have data y_t, $t = 1,\ldots n$. Let $b_0, b_1, \ldots, b_k$ be the least squares estimates of $\beta_0, \beta_1, \ldots, \beta_k$, respectively. Then the *fitted values* are

$$\hat{y}_t = b_0 + b_1 x_{1t} + \cdots + b_k x_{kt}, \quad t = 1, \ldots, n,$$

and the *residuals* are

$$e_t = y_t - \hat{y}_t, \qquad t = 1,..n.$$

We can think of e_t as an estimate of the value of ε_t realized when the data were generated.

A major issue we often address is model selection. That is, which explanatory variables should be included in a regression model? We may view (1) as one specific candidate model. That is, the k listed explanatory variables form one possible choice. The residual values measure the variation in the data that the model fit is unable to capture. For a good choice of model we want each of the explanatory variables to be statistically significant, and the residuals to be small relative to the variation in the y_t values, the original data.

The residual sum of squares (also called the error sum of squares) is

$$SSE = \sum_{t=1}^{n} e_t^2$$

and

$$s^2 = \frac{1}{n-k-1} SSE = \frac{1}{n-k-1} \sum_{t=1}^{n} e_t^2 .$$

That is, divide the residual sum of squares by its degrees of freedom (we lose one degree of freedom for each parameter estimated in the model). The square root of this

expression, *s*, is the so-called *standard error of the regression*, also called the *root mean squared error*. It is an estimate of the standard deviation of the disturbance term ε.

We measure the total variation in the y_t values, the original data, by

$$\sum_{t=1}^{n}(y_t-\bar{y})^2.$$

This is the sample variance of the y_t values without division by $n-1$. R^2, the percentage of the variation of the y_t values explained by the fitted regression model, is

$$R^2=1-\frac{SSE}{\sum_{t=1}^{n}(y_t-\bar{y})^2}=1-\frac{\sum_{t=1}^{n}e_t^2}{\sum_{t=1}^{n}(y_t-\bar{y})^2}.$$

R^2 necessarily increases as more explanatory variables are added to the model.

The denominator in the definition of R^2 doesn't depend on the model fit to the data, and thus R^2 is maximized if the residual sum of squares is minimized. The residual sum of squares can be made smaller and smaller by adding more and more explanatory variables to the model we fit (and R^2 can be made to approach 1). However, it doesn't make sense to simply maximize R^2: variables should be included in a model if they are statistically significant and enhance interpretation, not just to increase R^2. We can address this objection by modifying R^2 via the attachment of a penalty factor. First note that

$$s^2=\frac{1}{n-k-1}SSE=\frac{1}{1-\frac{k+1}{n}}\frac{1}{n}SSE.$$

As we add variables to a regression model $(1/n)$SSE decreases, but

$$\frac{1}{1-\frac{k+1}{n}}$$

increases—it acts as a penalty factor, a price to pay. Adjusted R^2 incorporates this penalty factor. It is defined as

$$R_a^2 = 1 - \frac{\frac{1}{n-k-1} SSE}{\frac{1}{n-1}\sum_{t=1}^{n}(y_t - \bar{y})^2} = 1 - \frac{s^2}{\frac{1}{n-1}\sum_{t=1}^{n}(y_t - \bar{y})^2}.$$

That is, adjusted R^2 is formed by dividing the two sums of squares in the definition of R^2 by their corresponding degrees of freedom. Then maximization of adjusted R^2 is equivalent to minimization of s^2.

This discussion indicates that adjusted R^2 can serve as a model selection criterion. We choose the model for which adjusted R^2 is maximum. The factor

$$\frac{1}{1 - \frac{k+1}{n}}$$

serves as a penalty factor, as we show below. Many model selection criteria have been proposed and studied in the literature. Two commonly used ones are Akaike's AIC and Schwarz's Bayesian criterion, SBC (also called BIC). These are

$$AIC = \log \frac{SSE}{n} + \frac{2(k+1)}{n}$$

and

$$SBC = \log \frac{SSE}{n} + \frac{(k+1)\log n}{n}.$$

The logarithm of s^2 is

$$\log s^2 = \log \frac{SSE}{n} - \log\left(1 - \frac{k+1}{n}\right) \approx \log \frac{SSE}{n} + \frac{k+1}{n}.$$

For each of these one chooses the model for which the criterion is smallest.

The representations of the three criteria allow a simple comparison, if we employ $\log s^2$. Each criterion is of the form

$$\log \frac{SSE}{n} + \frac{c(k+1)}{n}.$$

The value of c is 1 for s^2 (equivalently for adjusted R^2), 2 for AIC, and $\log n$ for SBC. That is, among these three criteria the penalty is least severe for s^2 (equivalently

for adjusted R^2), and most severe for SBC. Therefore, s^2 will tend to select models with more variables than AIC, and AIC will tend to select models with more variables than SBC.

We close this discussion by establishing a property of adjusted R^2. Suppose one wants to test the null hypothesis

$$H_0 : \beta_k = 0$$

vs. the alternative

$$H_0 : \beta_k \neq 0.$$

H_0 states that variable x_{kt} does not enter model (1). The hypothesis is tested via the *t* statistic, which can be calculated as follows in two steps:

1. Fit two models to the data, the *full* model given by (1), and the *reduced* model,

(2) $$y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_{k-1} x_{k-1,t} + \varepsilon_t .$$

The reduced model omits the variable being tested.

2. The square of the *t* statistic can be expressed as

$$t^2 = \frac{R^2_{full} - R^2_{reduced}}{(1-R^2_{full})/(n-k-1)} = \frac{SSE_{reduced} - SSE_{full}}{SSE_{full}/(n-k-1)}.$$

Consider the change in adjusted R^2 when we move from model (2) to model (1):

$$R_a^2(\,full) - R_a^2(reduced)$$

$$= 1 - \frac{SSR_{full}/(n-k-1)}{\sum_1^n (y_t - \bar{y})^2/(n-1)} - \left(1 - \frac{SSR_{reduced}/(n-k)}{\sum_1^n (y_t - \bar{y})^2/(n-1)} \right)$$

$$= \frac{SSR_{reduced}/(n-k) - SSR_{full}/(n-k-1)}{\sum_1^n (y_t - \bar{y})^2/(n-1)}.$$

Comparison of the last two displays and simple algebra show that

$$R_a^2(\,full) - R_a^2(reduced) > 0$$

if and only if $t^2 > 1$. Thus, adjusted R^2 increases when a variable is added if and only if the square of the *t* statistic for the new variable is greater than 1. If we test at the five per cent level, then (for *n* larger than 30, say) we require that the square of the *t* statistic exceed approximately 4 to justify adding the variable. Therefore, increase in adjusted R^2 is not a sufficiently stringent criterion for adding a new variable to a regression model. Stated another way, the penalty imposed is not large enough.

Summary and additional remarks

1. Selection criteria are widely used as an aid in choosing from among competing models which have been fit to the same data set. Two commonly used such criteria are Akaike's AIC and Schwarz's SBC. Later in the course, when we consider ARIMA modeling, we will be able to employ these criteria for model selection. I will encourage the use of AIC, rather than SBC.

2. The discussion presented also treats the adjusted *R*-square statistic employed in fitting regressions. A common feature among adjusted *R*-square, AIC, and SBC is described. While adjusted *R*-square can be employed as a model selection criterion, its use as such is discouraged.

ARIMA Models—Assorted Matters

A. Use of the backward shift operator notation with ARMA models

An ARMA(p,q) model with zero mean is specified by

$$y_t - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}. \tag{1}$$

Recall that the backward shift operator is defined by

$$By_t = y_{t-1},\ B^2 y_t = B(By_t) = By_{t-1} = y_{t-2}, \ldots, B^j y_t = y_{t-j}, \ldots.$$

Thus the ARMA(p,q) model (1) may be written as

$$\left(1 - \phi_1 B - \cdots - \phi_p B^p\right) y_t = \left(1 + \theta_1 B + \cdots + \theta_q B^q\right) \varepsilon_t,$$

or

$$\phi(B) y_t = \theta(B)\, \varepsilon_t,$$

where

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p, \quad \theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$$

are the autoregressive and moving average operators, respectively. The zeros of the polynomials

$$\phi(z) = z^p - \phi_1 z^{p-1} - \cdots - \phi_p, \quad \theta(z) = z^q + \theta_1 z^{q-1} + \cdots + \theta_q$$

play a central role in specification of the structure and the forecasting properties of the ARMA(p,q) process. These are the polynomials associated with the autoregressive and moving average operators. Note we have written the polynomials in descending rather than ascending order. In fact, either specification may be employed.

B. Parameterization of the mean for an ARMA model

The ARMA(p,q) model (1) above has zero mean. To incorporate nonzero mean into the model there are two choices.

(i) Introduce a constant ϕ_0 in (1). That is, write

$$y_t - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p} = \phi_0 + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}.$$

(ii) Adjust y_t by subtracting its mean μ. Write

$$(y_t - \mu) - \phi_1 (y_{t-1} - \mu) - \cdots - \phi_p (y_{t-p} - \mu) = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}.$$

The relation between the two options is given by

$$\mu = \frac{\phi_0}{1 - \phi_1 - \cdots - \phi_p}.$$

C. Estimation of ARMA parameters

One of the initial steps in fitting an ARMA model to data is the choice of the autoregressive order p and the moving average order q. To accomplish this we use both the sample autocorrelation function and the sample partial autocorrelation function. These are estimates of the population autocorrelation function and the population partial autocorrelation function. One is guided in the choice of p and q by the following properties of pure AR models, pure MA models, and mixed ARMA models.

	Model		
	AR(p)	MA(q)	ARMA(p,q)
acf	tails off	cuts off at lag q	tails off
pacf	cuts off at lag p	tails off	tails off

Estimation for the AR(p) model $y_t - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p} = \varepsilon_t$ is straightforward and one can apply least squares. One chooses values of the coefficients $\phi_1, \ldots, \phi_p$ for which

$$\sum_{t=p+1}^{n} \left(\dot{y}_t - \phi_1 \dot{y}_{t-1} - \cdots - \phi_p \dot{y}_{t-p} \right)^2, \quad \dot{y} = y_t - \bar{y}$$

is minimum. Thus, computer programs which handle multiple regression can be used. An alternative which is commonly employed is choice of the values of $\phi_1, \ldots, \phi_p$ which give the solution of the sample Yule–Walker equations,

(2) $$r_k = \phi_1 r_{k-1} + \phi_2 r_{k-2} + \cdots + \phi_p r_{k-p}, \quad k = 1, \ldots, p,$$

with

$$r_k = r_{-k} = \frac{\sum_{t=1}^{n-k} (y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2}.$$

Equation (2) with $k = 0$ is used to estimate $\sigma^2 = \text{var}(\varepsilon_t)$.

Many statistical packages use the Yule–Walker procedure to estimate autoregressive coefficients. The Yule–Walker method produces estimates with greater bias than the estimates created by the least squares method, and the Yule–Walker bias can be substantial. JMP uses least squares to estimate the autoregressive coefficients and the partial correlations. If the Yule–Walker method is employed to estimate the partial autocorrelations, large bias can result in the estimation of the pacf.

In the ARMA(p,q) case with $q > 0$, the estimation procedure is more complicated. Typically, approximate maximum likelihood estimation is employed.

D. ARIMA models

A time series which does not have constant mean $\text{E}y_t = \mu$ is not stationary. For example, if there is a trend component the series is not stationary. For many series differencing provides a means of achieving stationarity. If the d th-order difference

$$w_t = (1 - B)^d y_t$$

is a stationary ARMA(p,q) process, we say y_t is an ARIMA(p, d, q) process, and we write (the specification includes a zero mean for $w_t = (1 - B)^d y_t$)

$$\left(1 - \phi_1 B - \cdots - \phi_p B^p\right)(1 - B)^d y_t = \left(1 + \theta_1 B + \cdots + \theta_q B^q\right)\varepsilon_t,$$

or

$$\left(1 - \phi_1 B - \cdots - \phi_p B^p\right) w_t = \left(1 + \theta_1 B + \cdots + \theta_q B^q\right)\varepsilon_t,$$

or

$$\phi(B)w_t = \theta(B)\varepsilon_t.$$

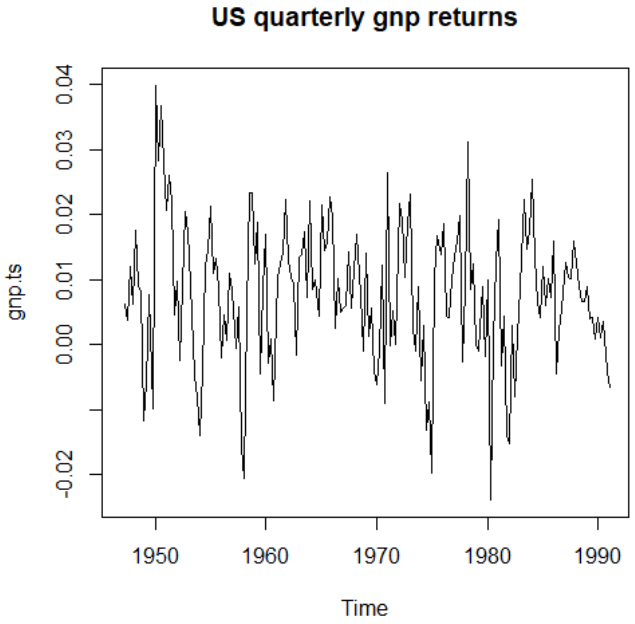
In practice one chooses d by examining visually the plot of y_t vs. *t*, the sample acf, and the sample pacf. One selects d large enough to eliminate a trend from the plot of y_t vs. *t*, and so that the sample acf and sample pacf tail off rapidly as the lag increases. It is unusual to select d to be greater than 2.

ARMA Model Fitting—U.S. Quarterly GNP

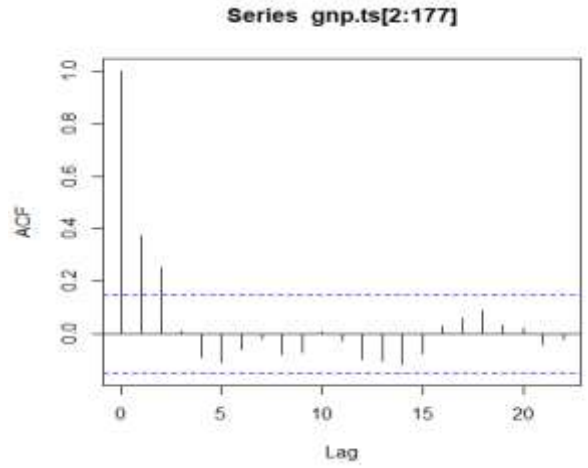
The data set qgnp.txt contains 176 observations of U.S. quarterly GNP returns, for 1947.2 to 1991.1, calculated as log returns. The series is seasonally adjusted. These data are analyzed in Example 2.1 and in Example 4.5 of Tsay.

We begin by plotting the data and giving the acf and pacf.

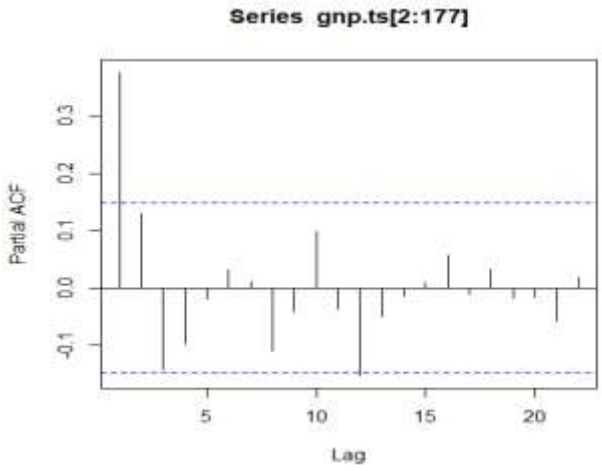
```
> qgnp<-read.csv("F:/Stat71122Spring/qgnp.txt")
> attach(qgnp)
> head(qgnp)
   return year quarter
1      NA 1947       1
2 0.00632 1947       2
3 0.00366 1947       3
4 0.01202 1947       4
5 0.00627 1948       1
6 0.01761 1948       2

> tail(qgnp)
      return year quarter
172  0.00085 1989       4
173  0.00420 1990       1
174  0.00108 1990       2
175  0.00358 1990       3
176 -0.00399 1990       4
177 -0.00650 1991       1

> gnp.ts<-ts(return,start=c(1947,1),freq=4)
> plot(gnp.ts,main="US quarterly gnp returns")
```



```
> acf(gnp.ts[2:177])
```




```
> pacf(gnp.ts[2:177])
```



The pacf plot suggests a low order AR fit, perhaps AR(1) or AR(3). Note that the lag 12 pacf value is barely significant, suggesting the deseasonalization was not thorough. I'm going to fit AR(1), AR(2), AR(3) and AR(4) models and compare them. It is common to experiment with some overfitting in choosing a model to fit the data. In judging the appropriateness of an AR model fit, we focus on the significance of the last coefficient of

the fitted model. If it is not significant, often we will refit with the AR order reduced by 1. But not always, as the following discussion shows.

To fit an AR(1) model, use the `arima` command as follows:

```
> gnp.ar1<-arima(gnp.ts,order=c(1,0,0))
```

The three arguments in the `order` specification are the values of *p*, *d*, and *q* for fitting an ARIMA(*p*, *d*, *q*) model. Brief output follows.

```
> gnp.ar1

Call:
arima(x = gnp.ts, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.3786     0.0077
s.e.  0.0698     0.0012

sigma^2 estimated as 9.801e-05:  log likelihood = 562.47,  aic = -
1118.94
```

For more detailed output, install the `lmtest` package and then load it from the library.

```
> library("lmtest")
Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric

Warning messages:
1: package 'lmtest' was built under R version 3.5.1
2: package 'zoo' was built under R version 3.5.1


> coeftest(gnp.ar1)

z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)
ar1       0.3786310  0.0698281  5.4223 5.883e-08 ***
intercept 0.0076889  0.0011993  6.4113 1.443e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let's digress briefly and show how this table is calculated.

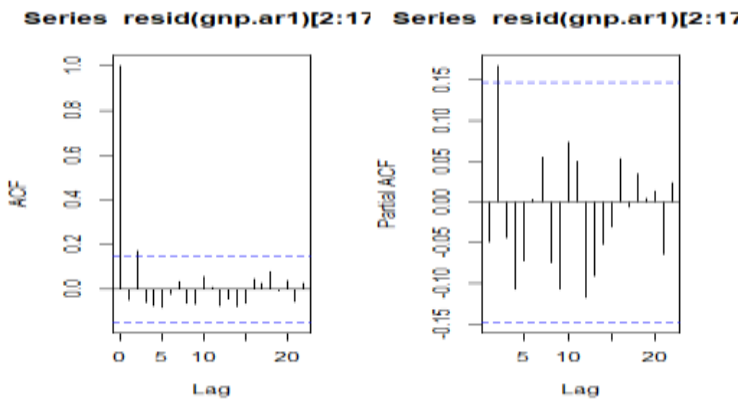
```
> coef(gnp.ar1)
        ar1   intercept
0.378631039 0.007688878
```

```
> #calculate standard errors of coefficients
> sqrt(diag(vcov(gnp.ar1)))
        ar1   intercept
0.069828084 0.001199277

> #calculate t ratios
> tratios<-coef(gnp.ar1)/sqrt(diag(vcov(gnp.ar1)))
> tratios
      ar1 intercept
 5.422332  6.411259

> #calculate p values
> 2*(1-pnorm(abs(tratios)))
         ar1    intercept
5.882657e-08 1.443230e-10
```

Note that the *p*-values are obtained from the normal distribution, rather than a *t* distribution.

Next, let's consider the residuals from this AR(1) fit.

```
> par(mfrow=c(1,2))
> acf(resid(gnp.ar1)[2:177])
> pacf(resid(gnp.ar1)[2:177])
```

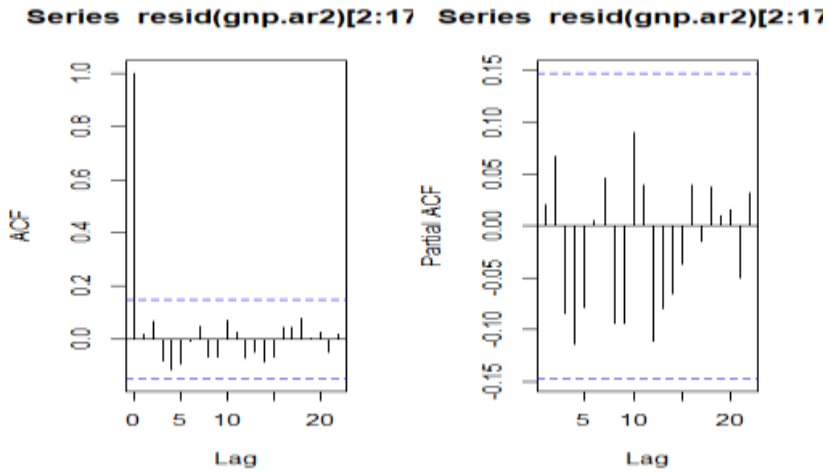


Both the lag 2 residual autocorrelation and the lag 2 residual partial autocorrelation are significant, and we conclude that the AR(1) is not adequate. An AR(2) model fit follows.

```
> gnp.ar2<-arima(gnp.ts,order=c(2,0,0))
> coeftest(gnp.ar2)

z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)
ar1       0.3288397  0.0745833  4.4090 1.038e-05 ***
ar2       0.1330454  0.0747537  1.7798   0.07511 .
intercept 0.0076455  0.0013683  5.5876 2.303e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The lag 2 coefficient is marginally significant.

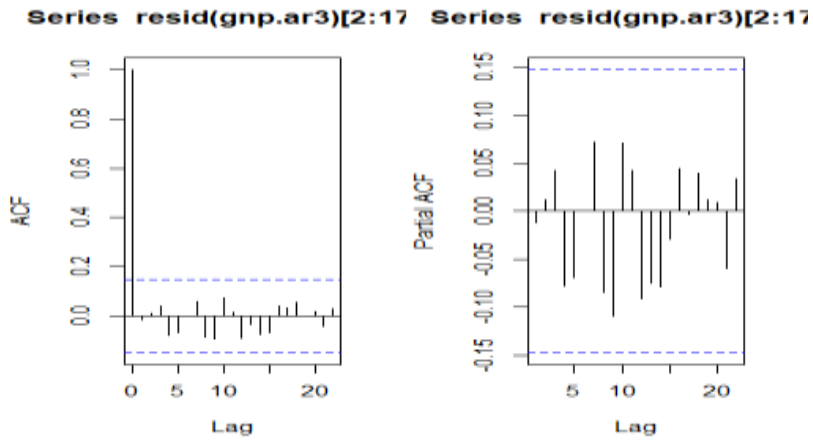
```
par(mfrow=c(1,2))
acf(resid(gnp.ar2)[2:177])
pacf(resid(gnp.ar2)[2:177])
```



The residual acf plot and residual pacf plot both suggest reduction to white noise has been achieved by the AR(2) fit. Nonetheless, let's proceed further and fit an AR(3) model.

```
> gnp.ar3<-arima(gnp.ts,order=c(3,0,0))
> coeftest(gnp.ar3)

z test of coefficients:

            Estimate Std. Error z value  Pr(>|z|)
ar1        0.3479915  0.0744573  4.6737 2.958e-06 ***
ar2        0.1793286  0.0778097  2.3047   0.02118 *
ar3       -0.1422596  0.0745231 -1.9089   0.05627 .
intercept  0.0076803  0.0011899  6.4545 1.086e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The lag 3 coefficient has *p*-value 0.056 and does merit attention.



These residual plots certainly suggest reduction to white noise has occurred. The AR(4) model fit follows.

```
> gnp.ar4<-arima(gnp.ts,order=c(4,0,0))
> coeftest(gnp.ar4)

z test of coefficients:

             Estimate Std. Error z value  Pr(>|z|)
ar1         0.3337150  0.0749047  4.4552 8.382e-06 ***
ar2         0.1966999  0.0786650  2.5005    0.0124 *
ar3        -0.1087163  0.0785343 -1.3843    0.1663
ar4        -0.0971933  0.0748107 -1.2992    0.1939
intercept   0.0077071  0.0010812  7.1285 1.015e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

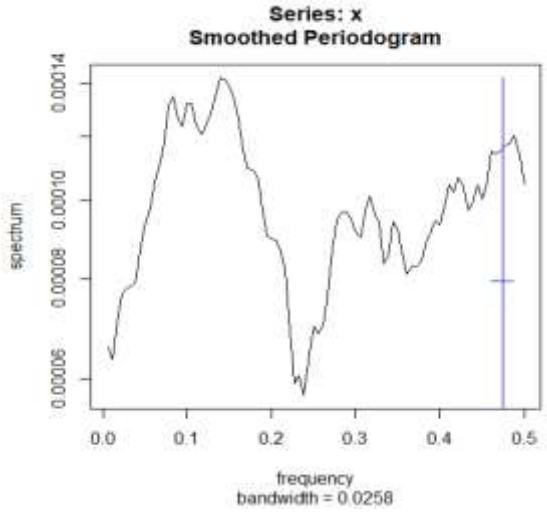
The last coefficient is certainly not significant. Let's examine the AIC values of the four fitted models.

```
> gnpaic<-c(gnp.ar1$aic,gnp.ar2$aic,gnp.ar3$aic,gnp.ar4$aic)
> gnpaic
[1] -1118.943 -1120.080 -1121.685 -1121.363
```

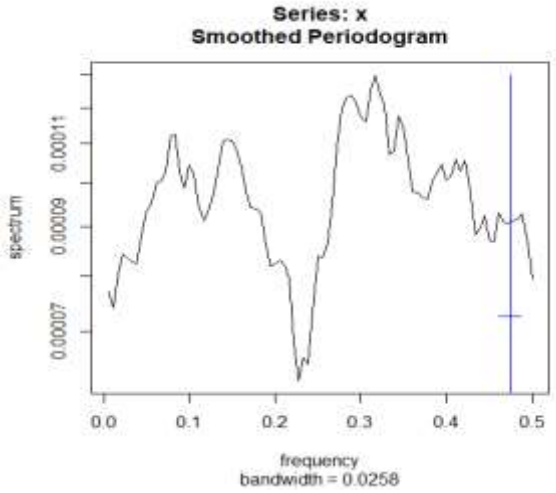
The four fits have similar features, as one can see from inspection of the regression coefficients. The residual acf is acceptable for all of the fits except perhaps the AR(1) model. The *t* ratios for the coefficients suggest that AR(3) should be chosen. The AIC criterion gives a slight edge to the AR(3) fit, with the AR(4) fit a close second best. I will choose the AR(4) model because of the interpretation it provides, as the discussion below makes clear.

Let's look at the residual spectral densities for the fits and compare these residual spectral densities to each other and to the density of the quarterly gnp return time series itself. If we have reduced to white noise with a model fit, that model's residual spectral density will be flat (constant)—the spectral density of white noise is flat.

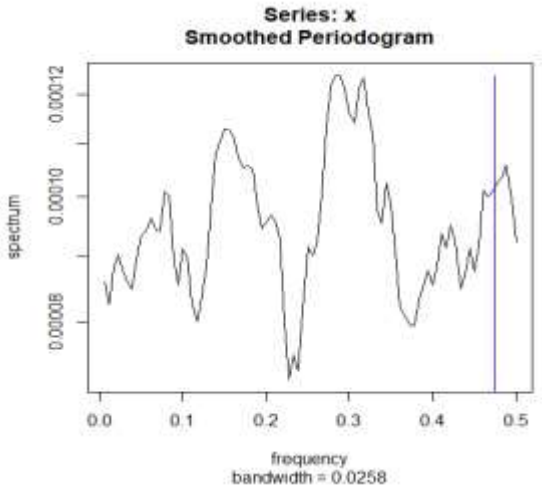
```
> spectrum(resid(gnp.ar2)[2:177],span=16)
```




```
> spectrum(resid(gnp.ar3)[2:177],span=16)
```




```
> spectrum(resid(gnp.ar4)[2:177],span=16)
```

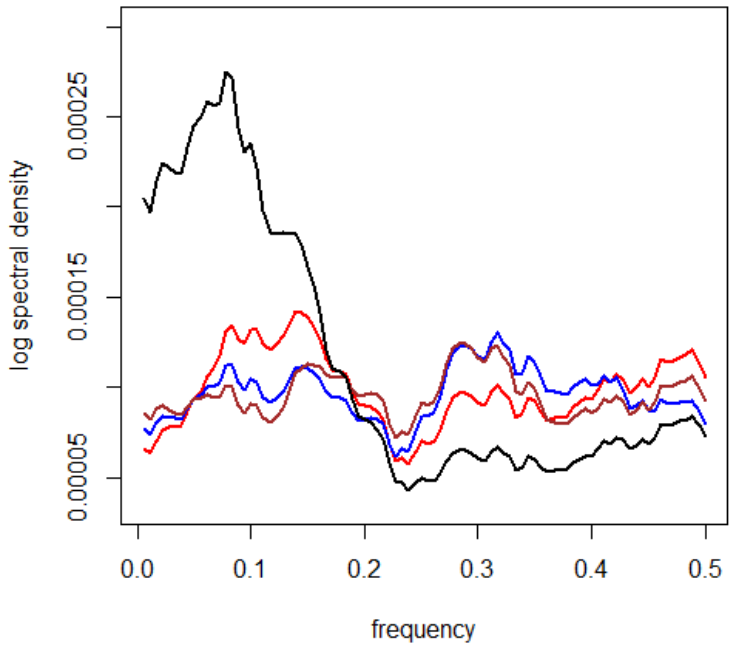


The residual spectral plot for the AR(2) fit indicates failure to reduce to white noise. The residual spectra for the AR(3) and AR(4) models are more flat, but perhaps appear to favor the AR(4) fit (we don't see the full extent of the blue line segment, but it is clearly quite long in the picture).

The next plot shows four spectral density estimates, one for the return series itself and the other three for the residuals from the AR(2), AR(3), and AR(4) fits. By comparing these spectra, we can see how the spectrum of the return series is gradually whitened as the AR fits progress from order 2 to order 4.

Some lines of R code are needed to prepare the plot.

```
> spar2<-spectrum(resid(gnp.ar2)[2:177],span=16,plot=F)
> spar3<-spectrum(resid(gnp.ar3)[2:177],span=16,plot=F)
> spar4<-spectrum(resid(gnp.ar4)[2:177],span=16,plot=F)
> spreturn<-spectrum(return[2:177],span=16,plot=F)

>
plot(spar2$freq,spar2$spec,ylim=c(0.000035,0.0003),type="l",lty=1,lwd=2
,col="red",xlab="frequency",ylab="log spectral density")
> lines(spar2$freq,spar3$spec,type="l",lty=1,lwd=2,col="blue")
> lines(spar2$freq,spar4$spec,type="l",lty=1,lwd=2,col="brown")
> lines(spar2$freq,spreturn$spec,type="l",lty=1,lwd=2,col="black")
```



The return series spectrum is in black, the residual AR(2) spectrum is in red, the residual AR(3) spectrum is in blue, and the residual AR(4) spectrum is in brown. Each of the

three AR fits described has produced a residual spectrum for which low frequency activity of the return series is attenuated and high frequency activity is enhanced, greatly flattening the return series spectrum. The picture shows that the AR(3) and AR(4) fits are better in reducing to white noise than the AR(2) fit, but it doesn't provide clear distinction between the AR(3) and AR(4) fits in this regard.

Next let's interpret three of the fitted models.

The AR(2) fit is

$$(1 - 0.32884B - 0.13305B^2)(y_t - 0.00765) = \varepsilon_t.$$

The autoregressive polynomial written in descending order is

$$z^2 - 0.32884\, z - 0.13305.$$

Let's use R code to find the zeros.

```
> coef(gnp.ar2)
        ar1         ar2   intercept
0.328839734 0.133045372 0.007645489

> zeros<-1/polyroot(c(1,-coef(gnp.ar2)[1:2]))
> zeros
[1]  0.5645189+0i -0.2356792+0i
```

Note that for this calculation only the first two estimated coefficients are wanted, because the third estimated coefficient is the intercept. The polynomial zeros are real-valued.

The AR(3) fit is

$$(1 - 0.34799B - 0.17933B^2 + 0.14226B^3)(y_t - 0.00768) = \varepsilon_t.$$

The first two lag coefficients are very close to those in the AR(2) fit. The autoregressive polynomial for the AR(3) fit is

$$z^3 - 0.34799\, z^2 - 0.17933\, z + 0.14226.$$

```
> #zeros of AR(3) polynomial
> zeros<-1/polyroot(c(1,-coef(gnp.ar3)[1:3]))
> zeros
[1]  0.4344004-0.2905982i -0.5208094-0.0000000i  0.4344004+0.2905982i
```

Thus, the zeros are -0.52081 and the complex conjugate pair $0.43440 \pm i\, 0.29060$. Let's find the amplitude and period for the complex conjugate pair. The calculation uses only the complex zero with positive imaginary part. The phase is calculated, followed by the period.

```
> #amplitude
Mod(zeros[3])
[1] 0.5226385
> #period
> 2*pi/Arg(zeros)[3]
[1] 10.65685
```

The result suggests the presence of a (rather weak) stochastic business cycle of length about 10.7 quarters. The estimate is that of a weak cycle because the amplitude of the complex zeros, 0.5226, is modest. The amplitudes of the complex zeros range between 0 and 1, and the strength of the cycle increases as the amplitude increases toward 1.

Finally, the AR(4) fit is

$$(1 - 0.33372B - 0.19670B^2 + 0.10872B^3 + 0.09719B^4)(y_t - 0.00771) = \varepsilon_t.$$

The autoregressive polynomial is

$$z^4 - 0.33372z^3 - 0.19670z^2 + 0.10872z + 0.09719.$$

This polynomial is written in descending order. However, the R command `polyroot` calculates the polynomial zeros for ascending order. We want to use descending order so that the amplitude calculation for the complex zeros will give values between 0 and 1. This is why the R command below includes a reciprocal calculation.

```
> #zeros of AR(4) polynomial
> zeros<-1/polyroot(c(1,-coef(gnp.ar4)[1:4]))
> zeros
[1]  0.5551679-0.3755875i -0.3883104-0.2560227i -0.3883104+0.2560227i
[4]  0.5551679+0.3755875i

> #amplitudes
> Mod(zeros)[3:4]
[1] 0.4651156 0.6702815
> #periods
> 2*pi/Arg(zeros)[3:4]
[1]  2.455626 10.563531
```

The four zeros occur in two complex conjugate pairs. One pair suggests a stochastic cyclical component with period about 10.6 quarters, and the second pair points to a weaker stochastic cyclical component with period about 2.5 quarters.

Tsay (pages 188–189) employs a Markov switching autoregressive model to fit the data. He identifies an expansion period averaging about 11.3 quarters and a contraction period averaging about 3.7 quarters.

The succession of the last three AR models is interesting. The AR(2) fit fails to display the cyclical component explicitly. The AR(3) model fit identifies the longer cyclical

component, and by moving to an AR(4) fit, we are able to also see the shorter (and weaker) cyclical component. We conclude that the AR(4) model reveals features of the time series that are not evidenced by the AR(3) model.

As noted above, the AR(4) model fit estimates the presence of stochastic cyclical components with periods 10.564 quarters and 2.456 quarters. These correspond to peaks in the estimated spectral density at approximately 1/10.565 = 0.095 cycles per quarter and 1/2.456 = 0.407 cycles per quarter. The latter peak is not prominent. That is, the component for the expansion period is stronger than the component for the contraction period.

Before ending this discussion of the GNP data set, let's look at one more candidate model. The estimated acf plot on page 11 suggests trying an MA(2).

```
> gnp.ma2<-arima(gnp.ts,order=c(0,0,2))
> coeftest(gnp.ma2)

z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)
ma1       0.3120841  0.0735627  4.2424 2.211e-05 ***
ma2       0.2713679  0.0678506  3.9995 6.348e-05 ***
intercept 0.0076812  0.0011625  6.6072 3.917e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> gnp.ma2$aic
[1] -1122.288
```

This MA(2) model fit has an AIC value lower than those of the AR(3) and AR(4) fits. The fitted MA(2) model is

$$y_t = 0.00768 + (1 + 0.31208B + 0.27137B^2)\,\varepsilon_t .$$

Let's rewrite this in autoregressive form. We have

$$(1+0.31208B + 0.27137B^2)^{-1}(\,y_t - 0.00768) = \varepsilon_t\,.$$

```
> delta<-c(rep(0,times=10))
> delta[1]<-1;delta[2]<--0.3120841
> for(j in 3:10){
+ j1<-j-1;j2<-j-2
+ delta[j]<--0.3120841*delta[j1]-0.2713679*delta[j2]
+ }
> delta
 [1]  1.0000000000 -0.3120841000 -0.1739714145  0.1389833192  0.0038357733
 [6] -0.0389126953  0.0111031277  0.0070945468 -0.0052271277 -0.0002939288
```

Thus, the autoregressive form of the MA(2) model fit is

$$(1 - 0.31208B - 0.17397B^2 + 0.13898B^3 + 0.00384B^4 - 0.03891B^5 + \cdots) \times (y_t - 0.00768) = \varepsilon_t .$$

We see that this MA(2) fit is approximately an AR(5) model that is similar to the AR(3) fit shown above. The MA(2) model has two coefficients, and thus we cannot expect to estimate two pseudocycles from it—that would require estimation of at least four coefficients.

My inclination is to stick with the estimated AR(4) model. The estimated periods obtained from it lead to a conclusion of two pseudocycles, consistent with the result of the Markov switching autoregressive model fit described by Tsay.

Summary and additional remarks

1. Autoregressions are models which provide great flexibility in describing time series data. They offer clear and convenient interpretation of the structure of the data. Specifically, with an autoregressive model, the present value of the series is given as a linear combination of a finite number of immediate past values, plus a new random noise value.

2. When one fits an AR(p) model, an autoregression of order p, one needs to select the order and judge whether suitable reduction to white noise has been achieved. It is common to fit several models with different values of p and use diagnostics to choose one of the fitted models. One should use all of the diagnostics to judge which model to select.

(i) The t ratios and p-values of the estimated coefficients are examined to judge significance. In doing so, one focuses primarily on the p-value of the highest lag coefficient.

(ii) One also examines the AIC values of the competing models. The lower the AIC value is, the better is the model.

(iii) If there is reduction to white noise for a model, the residual autocorrelations and residual partial autocorrelations will be insignificant.

(iv) Examine the residual spectral density plot on a log scale. If there is reduction to white noise, the density will be flat. One can use Bartlett's Kolmogorov–Smirnov statistic to test whether the density is sufficiently flat to be consistent with white noise (to be discussed in later notes). R output provides a line segment on the spectral plot which one can use to judge flatness visually.

In choosing a model, one should use all of these diagnostics, rather than rely on just one or several of them.

3. Complex-valued zeros of the autoregressive polynomial for an AR fit are often very useful in describing time series structure. This is the case for the U.S. GNP series considered in these notes. Autoregressions with complex-valued polynomial zeros exhibit cyclical behavior. The AR(4) model fit to the U.S. GNP time series suggests there are two stochastic cyclical components in the time series, one corresponding to an expansion period of length approximately 10.6 quarters, and a second corresponding to a contraction period of about 2.5 quarters.

4. In these notes, an AR(4) model is chosen to describe the U.S. GNP time series, even though the lag 4 estimated coefficient is not significant and the AR(3) fit has a slightly lower AIC value. The AR(4) model was chosen because of the interpretation it offers for the data.

Appendix: National Bureau of Economic Research

The National Bureau of Economic Research (NBER) is an organization made up of economists from academia, business and government. The NBER web site is http://www.nber.org/info.html. It begins with the following statement:

> Founded in 1920, the National Bureau of Economic Research is a private, nonprofit, nonpartisan research organization dedicated to promoting a greater understanding of how the economy works. The NBER is committed to undertaking and disseminating unbiased economic research among public policymakers, business professionals, and the academic community.

NBER's Business Cycle Dating Committee is generally accepted as the arbiter defining periods of recession and expansion. The following statement by the Committee is at http://www.nber.org/cycles/recessions.html:

> The NBER's Business Cycle Dating Committee maintains a chronology of the U.S. business cycle. The chronology comprises alternating dates of peaks and troughs in economic activity. A recession is a period between a peak and a trough, and an expansion is a period between a trough and a peak. During a recession, a significant decline in economic activity spreads across the economy and can last from a few months to more than a year. Similarly, during an expansion, economic activity rises substantially, spreads across the economy, and usually lasts for several years.
>
> In both recessions and expansions, brief reversals in economic activity may occur-a recession may include a short period of expansion followed by further decline; an expansion may include a short period of contraction followed by further growth. The Committee applies its judgment based on the above definitions of recessions and expansions and has no fixed rule to determine whether a contraction is only a short interruption of an expansion, or an expansion is only a short interruption of a contraction. The most recent example of such a judgment that was less than obvious was in 1980-1982, when the Committee determined that the contraction that began in 1981 was not a continuation of the one that began in 1980, but rather a separate full recession.
>
> The Committee does not have a fixed definition of economic activity. It examines and compares the behavior of various measures of broad activity: real GDP measured on the product and income sides, economy-wide employment, and real income. The Committee also may consider indicators that do not cover the entire economy, such as real sales and the Federal Reserve's index of industrial production (IP). The Committee's use of these indicators in conjunction with the broad measures recognizes the issue of double-counting of sectors included in both those indicators and

the broad measures. Still, a well-defined peak or trough in real sales or IP might help to determine the overall peak or trough dates, particularly if the economy-wide indicators are in conflict or do not have well-defined peaks or troughs.

The following table lists the recessions and expansions, as determined by NBER, during the period 1948–1991. The two contractions since 1991 are added at the bottom, to provide contemporary information. The counts in the table specify the number of months.

Peak	Trough	Contraction Peak to Trough	Expansion Trough to Peak
Feb 45	Oct 45	8	80
Nov 48	Oct 49	11	37
Jul 53	May 54	10	45
Aug 57	Apr 58	8	39
Apr 60	Feb 61	10	24
Dec 69	Nov 70	11	106
Nov 73	Mar 75	16	36
Jan 80	Jul 80	6	58
Jul 81	Nov 82	16	12
Jul 90	Mar 91	8	92
Mar 01	Nov 01	8	120
Dec 07	Jun 09	18	73
Mar 20	Apr 20	2	128

For 1948 to 1991 the average length of the nine contractions is 3.56 quarters, and the average length of the nine expansions is 16.63 quarters.