MODEL SELECTION, DATA SPLITTING FOR ARMA TIME SERIES AND VISUALIZING SOME BOOTSTRAP CONFIDENCE REGIONS

by

Welagedara Arachchilage Dhanushka Madumali Welagedara

B.Sc., University of Ruhuna, 2015

M.S., University of Peradeniya, 2018

A Dissertation
Submitted in Partial Fulfillment of the Requirements for the
Doctor of Philosophy Degree

School of Mathematical and Statistical Sciences
in the Graduate School
Southern Illinois University Carbondale
August 2023

DISSERTATION APPROVAL

MODEL SELECTION, DATA SPLITTING FOR ARMA TIME SERIES AND VISUALIZING SOME BOOTSTRAP CONFIDENCE REGIONS

by

Welagedara Arachchilage Dhanushka Madumali Welagedara

A Dissertation Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in the field of Mathematics

Approved by:

Dr. David Olive, Chair

Dr. Seyed Yaser Samadi

Dr. Bhaskar Bhattacharya

Dr. Kwangho Choiy

Dr. Thushari Jayasekera

Graduate School
Southern Illinois University Carbondale
June 20, 2023

AN ABSTRACT OF THE DISSERTATION OF

Welagedara Arachchilage Dhanushka Madumali Welagedara, for the Doctor of Philosophy degree in Mathematics, presented on June 20, 2023, at Southern Illinois University Carbondale.

TITLE: MODEL SELECTION, DATA SPLITTING FOR ARMA TIME SERIES AND VISUALIZING SOME BOOTSTRAP CONFIDENCE REGIONS

MAJOR PROFESSOR: Dr. David Olive

ARMA model selection with criterion such as AIC and BIC tends not to select a consistent ARMA model with high probability. Hence data splitting is not reliable. One technique was fairly reliable with large sample sizes, and a modification also worked.

The DD plot for visualizing prediction regions can also be used to visualize three bootstrap confidence regions.

ACKNOWLEDGMENTS

First of all, I would like to thank my mentor Dr. David Olive for his dedicated support and guidance he has given me throughout my study for past four years. Without his help, advise, encouragement, this research and dissertation would not have happened. I feel blessed to have him as my Ph.D. advisor.

I would also like to thank all the professors and staff of School of Mathematical and Statistical Sciences at Southern Illinois University for their instructions and care throughout my time as a graduate student.

Last but not least, I want to say thank you to my parents and my husband, for all of your support and encouragement. You have pushed me to succeed and I could not have done it without you.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

A *time series* $Y_1, ..., Y_n$ consists of observations Y_t collected sequentially at times $1, ..., n$. We will use the *R* software notation and write a moving average parameter θ with a positive sign. Many references and software will write the model with a negative sign for the moving average parameters. For the time series models described below, we will assume that the errors e_t are independent and identically distributed (iid) with zero mean and variance σ^2. The backshift operator or lag operator B satisfies $BW_t = W_{t-1}$ and $B^j W_t = W_{t-j}$.

A *moving average* MA(q) times series is

$$Y_t = \tau + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t = \tau + (1 + \theta_1 B + \cdots + \theta_q B^q) e_t = \tau + \theta(B) e_t$$

where $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ and $\theta_q \neq 0$. Note that $E(Y_t) = \mu = \tau = \theta_0$ for $t \geq 1$. Since the e_t are iid, the Y_t are identically distributed, and $Y_j, Y_{j+q+1}, Y_{j+2(q+1)}$, ... are iid.

An *autoregressive* AR(p) times series is

$$Y_t = \tau + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \text{ or } (1 - \phi_1 \mathrm{B} - \cdots - \phi_\mathrm{p} \mathrm{B}^\mathrm{p}) \mathrm{Y_t} = \tau + \mathrm{e_t},$$

or $\phi(B) Y_t = \tau + e_t$ where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ and $\phi_p \neq 0$. If $E(Y_t) = \mu$ for $t \geq 1$, write $Y_t - \mu = \sum_{j=1}^{p} \phi_j (Y_{t-j} - \mu) + e_t$ to get $\tau = \phi_0 = \mu(1 - \sum_{j=1}^{p} \phi_j)$.

An *autoregressive moving average* ARMA(p, q) times series is

$$Y_t = \tau + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t,$$

or $\phi(B) Y_t = \tau + \theta(B) e_t$ where $\theta_q \neq 0$ and $\phi_p \neq 0$. The ARMA(0,q) model is the MA(q) model, and the ARMA(p,0) model is the AR(p) model. Again $\tau = \mu(1 - \sum_{j=1}^{p} \phi_j)$ if $p \geq 1$, and $\tau = \mu$ if $p = 0$. The ARMA(0,0) model is $Y_t = \mu + e_t$, often called the location model.

The results in this dessertation also apply to a time series X_t that follows an ARIMA(p, d, q) model with known d if the differenced time series model Y_t follows an ARMA(p, q) model. To describe ARIMA models, let the difference operator $\nabla = (1 - B)$. Let $Y_t = \nabla^d X_t = (1 - B)^d X_t$ be the differenced time series. The first difference is $Y_t = \nabla X_t = (1 - B)X_t = X_t - X_{t-1}$. The second difference is $Y_t = \nabla^2 X_t = \nabla(\nabla X_t) = X_t - 2X_{t-1} + X_{t-2}$. If X_t follows an ARIMA(p, d, q) model, want Y_t to follow a weakly stationary, causal, and invertible ARMA(p, q) = ARIMA$(p, 0, q)$ model. Typically $d = 0$ or 1, but occasionally $d = 2$. Usually $\tau = 0$ if $d > 1$. The ARIMA$(p, d = 1, q)$ model is $X_t = \tau + (1+\phi_1)X_{t-1} + (\phi_2 - \phi_1)X_{t-2} + \cdots + (\phi_p - \phi_{p-1})X_{t-p} - \phi_p X_{t-p-1} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t$. The ARIMA$(p, d, q)$ model can be written compactly as $\phi(B)\ \nabla^d\ X_t = \tau + \theta(B)e_t$. See Box and Jenkins (1976) for more on these models.

A *stochastic process* $\{Y_t, t \in \mathbb{T}\}$ is a collection of random variables where often $\mathbb{T} = \mathbb{Z}$, the set of integers. The observed time series is $\{Y_t\} = Y_1, ..., Y_n$. The *mean function* $\mu_t = E(Y_t)$ for $t \in \mathbb{Z}$. The *autocovariance function* $\gamma_{t,s} = Cov(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E(Y_t Y_s) - \mu_t \mu_s$ for $t, s \in \mathbb{Z}$. The *autocorrelation function* $\rho_{t,s} = Corr(Y_t, Y_s) = \dfrac{Cov(Y_t, Y_s)}{\sqrt{Var(T_t)Var(Y_s)}} = \dfrac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}}$ for $t, s \in \mathbb{Z}$.

A process $\{Y_t\}$ is weakly stationary if a) $E(Y_t) = \mu_t \equiv \mu$ is constant over time, and b) $\gamma_{t,t-k} = \gamma_{0,k}$ for all times t and lags k. Hence the covariance function $\gamma_{t,s}$ depends only on the absolute difference $|t - s|$. For a weakly stationary process $\{Y_t\}$, write the *autocovariance function* as $\gamma_k = Cov(Y_t, Y_{t-k})$ and the *autocorrelation function* as $\rho_k = corr(Y_t, Y_{t-k}) = \gamma_k/\gamma_0$. Note that the mean function $E(Y_t) = \mu$ and the variance function $V(Y_t) = Var(Y_t) = \gamma_0$ are constant and do not depend on t. The autocovariance and autocorrelation functions γ_k and ρ_k depend on the lag k but not on the time t.

We usually want the ARMA(p, q) model to be weakly stationary, causal, and invertible. Let $Z_t = Y_t - \mu$ where $\mu = E(Y_t)$ if $\{Y_t\}$ is weakly stationary and μ is some origin otherwise. Then the causal property implies that $Z_t = \sum_{j=1}^{\infty} \psi_j e_{t-j} + e_t$, which is an MA$(\infty)$ representation, where the $\psi_j \to 0$ rapidly as $j \to \infty$. Invertibility implies that $Z_t = \sum_{j=1}^{\infty} \chi_j Z_{t-j} + e_t$, which is an AR$(\infty)$ representation, where the $\chi_j \to 0$ rapidly as $j \to \infty$. We will make the usual assumption that the AR(∞) and MA(∞) parameters are square summable. Thus if the ARMA(p, q) model is weakly

stationary, causal, and invertible, then Y_t depends almost entirely on nearby lags of Y_t and e_t, not on the distant past. Also, the time series model $\approx$ AR(p_y) $\approx$ MA(q_y) for some positive integers p_y and q_y that do not depend on the sample size n.

Consider $\theta(B)$ and $\phi(B)$ as polynomials in B. An ARMA(p, q) model is invertible if all of the roots of the polynomial $\theta(B) = 0$ have modulus > 1, and weakly stationary if all of the roots of the polynomial $\phi(B) = 0$ have modulus > 1. (Let the complex number $W = W_1 + W_2\ i$ have modulus $|W| = W_1^2 + W_2^2$.) Hence the roots of both polynomials lie outside the unit circle. An AR(p) model is always invertible and an MA(q) model is always causal. For the AR(1) model, need $|\phi_1| < 1$. For the MA(1) model, need $|\theta_1| < 1$. For the ARMA(1,1) model, need $|\phi_1| < 1$ and $|\theta_1| < 1$.

Let τ_i stand for θ_i or ϕ_i. Let k stand for q or p, and let $\psi(B) = 1 - \tau_1 B - \tau_2 B^2 - \cdots - \tau_k B^k$ stand for $\phi(B)$ or $\theta(B)$. A necessary but not sufficient condition for the roots of $\psi(B) = 0$ to all be greater than 1 in modulus is $\tau_1 + \cdots + \tau_k < 1$ and $|\tau_k| < 1$.

CHAPTER 2

MODEL SELECTION

Let I be a time series model. The $AIC(I)$ statistic is used to pick a model from several ARIMA models. The model I_{min} with the smallest AIC is always of interest but often overfits: has too many unnecessary parameters. Imagine fitting an ARIMA(p, d, q) model where $d = 0, 1$ or 2 is fixed and p and q run from 0 to j for small j. The number of parameters in the model for fixed d is $p + q + 2$ where $\sigma = \sqrt{V(e_t)}$, τ, $\phi_1, ..., \phi_p$, $\theta_1, ..., \theta_q$ are the parameters. $AIC(I)$ tends to be large when the model does not have enough terms, to drop as needed terms are added, and then to rise as unnecessary terms are added. If AIC is scaled correctly (the penalty is $2(p + q)$ rather than $2(p + q)/n$) and $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, and models with $4 \leq \Delta(I) \leq 7$ are borderline. See Brockwell and Davis (1987, p. 269), Duong (1984), and Burnham and Anderson (2004).

Haile and Olive (2023a) extend regression variable selection notation to ARMA time series model selection as in the next few paragraphs. Consider regression models where the response variable Y is independent of the $p \times 1$ vector of predictors $\boldsymbol{x}$ given $\boldsymbol{x}^T\boldsymbol{\beta}$, written $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$. Many important regression models satisfy this condition, including multiple linear regression and generalized linear models (GLMs).

Following Olive and Hawkins (2005), a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{2.1}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Suppose that S is a subset

of I and that model (2.1) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{I/S} + \boldsymbol{x}_O^T\mathbf{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \mathbf{0}$ if $S \subseteq I$. The model using $\boldsymbol{x}^T\boldsymbol{\beta}$ is the full model.

To clarify notation, suppose $p = 4$, a constant $x_1 = 1$ corresponding to β_1 is always in the model, and $\boldsymbol{\beta} = (\beta_1, \beta_2, 0, 0)^T$. Then the $J = 2^{p-1} = 8$ possible subsets of $\{1, 2, ..., p\}$ that always contain 1 are $I_1 = \{1\}$, $S = I_2 = \{1, 2\}$, $I_3 = \{1, 3\}$, $I_4 = \{1, 4\}$, $I_5 = \{1, 2, 3\}$, $I_6 = \{1, 2, 4\}$, $I_7 = \{1, 3, 4\}$, and $I_8 = \{1, 2, 3, 4\}$. There are $2^{p-a_S} = 4$ subsets I_2, I_5, I_6, and I_8 such that $S \subseteq I_j$. Also, $\hat{\boldsymbol{\beta}}_{I_7} = (\hat{\beta}_1, \hat{\beta}_3, \hat{\beta}_4)^T$ is obtained by regressing Y on $\boldsymbol{x}_{I_7} = (x_1, x_3, x_4)^T$.

Let I_{min} correspond to the set of predictors selected by a variable selection method such as forward selection or backward elimination. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. Also use zero padding for the model I_{min}. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then the observed variable selection estimator $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. As a statistic, $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are J subsets. For example, if each subset contains at least one variable, then there are $J = 2^p - 1$ subsets.

For ARMA model selection, let the full model be an ARMA(p_{max}, q_{max}) model. For AR model selection $q_{max} = 0$, while for MA model selection $p_{max} = 0$. If model selection is restricted to AR models, Granger and Newbold (1977, p. 178) suggest using $p_{max} = 13$ for nonseasonal time series, quarterly seasonal time series, and short monthly seasonal time series. They recommend $p_{max} = 25$ for longer monthly seasonal time series. We may use $p_{max} = q_{max} = 5$ for ARMA model selection, and $q_{max} = 13$ for MA model selection. For ARMA model selection, there are $J = (p_{max} + 1)(q_{max} + 1)$ ARMA(p, q) submodels where p ranges from 0 to p_{max} and q ranges from 0 to q_{max}. For AR and MA model selection there are $J = p_{max} + 1$ and $J = q_{max} + 1$ submodels, respectively. See Example 1 where there are 36 submodels.

Example 2.1. Shown below is the aicmatrix of $\Delta(I) = AIC(I) - AIC(I_{\min})$ for the *R* `WWW usage` time series, which gives the number of users connected to the Internet through a server every minute where $n = 100$. First differences were used so $d = 1$. From this output, I_{min} is the ARIMA(5,1,4) model. Some interesting models are the ARIMA(3,1,0) model and the ARIMA(1,1,1) model.

```
aicmat(WWWusage,dd=1,pmax=5)
$aics             q
  p      0     1    2     3    4    5
  0 119.86 38.67 8.74  9.13 8.24 7.72
  1  18.10  3.16 5.11  3.44 3.96 5.14
  2  11.04  5.15 6.22  4.63 2.10 6.95
  3   0.85  2.80 4.48  3.27 3.62 5.29
  4   2.79  1.74 5.04  7.94 4.26 6.99
  5   4.72  6.50 2.40 10.50 0.00 1.63
```

Assume the true (optimal) model is an ARMA(p_S, q_S) model with $p_S \leq p_{max}$ and $q_S \leq q_{max}$. Let the selected model I be an ARMA(p_I, q_I) model. Then the model underfits unless $p_I \geq p_S$ and $q_I \geq q_S$. For AR model selection, the probability of underfitting goes to 0 if the Akaike (1973) AIC, Schwartz (1978) BIC, or Hurvich and Tsai (1989) AIC_C criterion are used, at least if the e_t are iid $N(0, \sigma^2)$. Also see Claeskens and Hjort (2008, pp. 39, 40, 45, 46), Hannan and Quinn (1979), and Shibata (1976).

More notation is needed for model selection. Let the full model be the AR(p_{max}), MA(q_{max}), or ARMA(p_{max}, q_{max}) model. Let $\boldsymbol{\beta}$ be a $b \times 1$ vector. For ARMA model selection, let $\boldsymbol{\beta} = (\boldsymbol{\phi}^T, \boldsymbol{\theta}^T)^T = (\phi_1, ..., \phi_{p_{max}}, \theta_1, ..., \theta_{q_{max}})^T$ with $b = p_{max} + q_{max}$. For AR model selection, let $\boldsymbol{\beta} = (\phi_1, ..., \phi_{p_{max}})^T$ with $b = p_{max}$, and for MA model selection, let $\boldsymbol{\beta} = (\theta_1, ..., \theta_{q_{max}})^T$ with $b = q_{max}$. Hence $\boldsymbol{\beta} = (\beta_1, ..., \beta_{p_{max}}, \beta_{p_{max}+1}, ..., \beta_{p_{max}+q_{max}})^T$. Let $S = \{1, ..., p_S, p_{max} + 1, ..., p_{max} + q_S\}$ index the true ARMA(p_S, q_S) model. If $S = \emptyset$ is the empty set, then the time series random variables $Y_1, ..., Y_n$ are iid. Let $I = \{1, ..., p_I, p_{max}+1, ..., p_{max}+q_I\}$ index the ARMA(p_I, q_I) model. Let $\hat{\boldsymbol{\beta}}_{I,0}$ be a $b \times 1$ estimator of $\boldsymbol{\beta}$ which is a obtained by padding $\hat{\boldsymbol{\beta}}_I$ with zeroes. If $\boldsymbol{\beta}_I = (\phi_1, ..., \phi_{p_I}, \theta_1, ..., \theta_{q_I})^T$,

then $\hat{\boldsymbol{\beta}}_{I,0} = (\hat{\phi}_1, ..., \hat{\phi}_{p_I}, 0, .., 0, \hat{\theta}_1, ..., \hat{\theta}_{q_I}, 0, ..., 0)^T$. If $q_I = 0$, then $\hat{\boldsymbol{\beta}}_{I,0} = (\hat{\phi}_1, ..., \hat{\phi}_{p_I}, 0, .., 0)^T$. If $p_I = 0$ then $\hat{\boldsymbol{\beta}}_{I,0} = (0, ..., .., 0, \hat{\theta}_1, ..., \hat{\theta}_{q_I}, 0, ..., 0)^T$. If $I = \emptyset$ with $p_I = q_I = 0$, then define $\hat{\boldsymbol{\beta}}_{I,0} = \mathbf{0}$, the $b \times 1$ vector of zeroes. The submodel I underfits unless $S \subseteq I$. Note that the full model, e.g. the ARMA(p_{max}, q_{max}) model, is a submodel.

For example, if $p_{max} = q_{max} = 5$, then $S = \{1, 6, 7\}$ corresponds to the ARMA(1,2) model, and $I = \{1, 6, 7, 8\}$ corresponds to the ARMA(1,3) model. Then $\hat{\boldsymbol{\beta}}_S = (\hat{\phi}_1, \hat{\theta}_1, \hat{\theta}_2)^T$, $\hat{\boldsymbol{\beta}}_{S,0} = (\hat{\phi}_1, 0, 0, 0, 0, \hat{\theta}_1, \hat{\theta}_2, 0, 0, 0)^T$, and $\hat{\boldsymbol{\beta}}_{I,0} = (\hat{\phi}_1, 0, 0, 0, 0, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, 0, 0)^T$.

The model I_{min} corresponds to the model that minimizes the AIC, AIC_C, or BIC criterion. Then the model selection estimator $\hat{\boldsymbol{\beta}}_{MS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$. Assume $\hat{\boldsymbol{\beta}}_{MS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$. Haile and Olive (2023a) gave the large sample theory for $\hat{\boldsymbol{\beta}}_{MS}$, and used bootstrap confidence regions for hypothesis testing.

2.1 UNDERFITTING

The following Olive and Hawkins (2005) theorem will be useful. A plot can be very useful if the OLS line can be compared to a reference line and if the OLS slope is related to some quantity of interest. Suppose that a plot of w versus z places w on the horizontal axis and z on the vertical axis. Then denote the OLS line by $\hat{z} = a + bw$. The following theorem shows that the plotted points in the FF, RR, and response plots will cluster about the identity line. Notice that the theorem is a property of OLS and holds even if the data does not follow an multiple linear regression model. Let $corr(x, y)$ denote the correlation between x and y. Let $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ be the hat matrix.

Theorem 2.2. Suppose that every submodel contains a constant and that $\boldsymbol{X}$ is a full rank matrix.

Response Plot: i) If $w = \hat{Y}_I$ and $z = Y$ then the OLS line is the identity line.

ii) If $w = Y$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [corr(Y, \hat{Y}_I)]^2 = R^2(I)$ and intercept $a = \overline{Y}(1 - R^2(I))$ where $\overline{Y} = \sum_{i=1}^n Y_i/n$ and $R^2(I)$ is the coefficient of multiple determination from the candidate model.

FF or EE Plot: iii) If $w = \hat{Y}_I$ and $z = \hat{Y}$ then the OLS line is the identity line. Note that $ESP(I) = \hat{Y}_I$

and $ESP = \hat{Y}$.

iv) If $w = \hat{Y}$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [corr(\hat{Y}, \hat{Y}_I)]^2 = SSR(I)/SSR$ and intercept $a = \overline{Y}[1 - (SSR(I)/SSR)]$ where SSR is the regression sum of squares.

RR Plot: v) If $w = r$ and $z = r_I$ then the OLS line is the identity line.

vi) If $w = r_I$ and $z = r$ then $a = 0$ and the OLS slope $b = [corr(r, r_I)]^2$ and

$$corr(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}} = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}.$$

Proof: Recall that $\boldsymbol{H}$ and $\boldsymbol{H}_I$ are symmetric idempotent matrices and that $\boldsymbol{H}\boldsymbol{H}_I = \boldsymbol{H}_I$. The mean of OLS fitted values is equal to $\overline{Y}$ and the mean of OLS residuals is equal to 0. If the OLS line from regressing z on w is $\hat{z} = a + bw$, then $a = \overline{z} - b\overline{w}$ and

$$b = \frac{\sum(w_i - \overline{w})(z_i - \overline{z})}{\sum(w_i - \overline{w})^2} = \frac{SD(z)}{SD(w)} corr(z, w).$$

Also recall that the OLS line passes through the means of the two variables $(\overline{w}, \overline{z})$.

(*) Notice that the OLS slope from regressing z on w is equal to one if and only if the OLS slope from regressing w on z is equal to $[corr(z, w)]^2$.

i) The slope $b = 1$ if $\sum \hat{Y}_{I,i} Y_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}_I^T \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

ii) By (*), the slope

$$b = [corr(Y, \hat{Y}_I)]^2 = R^2(I) = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = SSR(I)/SSTO.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

iii) The slope $b = 1$ if $\sum \hat{Y}_{I,i} \hat{Y}_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}^T \hat{\boldsymbol{Y}}_I = \boldsymbol{Y}^T \boldsymbol{H}\boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

iv) From iii),

$$1 = \frac{SD(\hat{Y})}{SD(\hat{Y}_I)}[corr(\hat{Y}, \hat{Y}_I)].$$

Hence

$$corr(\hat{Y}, \hat{Y}_I) = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}$$

and the slope

$$b = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})} corr(\hat{Y}, \hat{Y}_I) = [corr(\hat{Y}, \hat{Y}_I)]^2.$$

Also the slope

$$b = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(\hat{Y}_i - \overline{Y})^2} = SSR(I)/SSR.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

v) The OLS line passes through the origin. Hence $a = 0$. The slope $b = \boldsymbol{r}^T\boldsymbol{r}_I / \boldsymbol{r}^T\boldsymbol{r}$. Since $\boldsymbol{r}^T\boldsymbol{r}_I = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$ and $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I) = \boldsymbol{I} - \boldsymbol{H}$, the numerator $\boldsymbol{r}^T\boldsymbol{r}_I = \boldsymbol{r}^T\boldsymbol{r}$ and $b = 1$.

vi) Again $a = 0$ since the OLS line passes through the origin. From v),

$$1 = \sqrt{\frac{SSE(I)}{SSE}}[corr(r, r_I)].$$

Hence

$$corr(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}}$$

and the slope

$$b = \sqrt{\frac{SSE}{SSE(I)}}[corr(r, r_I)] = [corr(r, r_I)]^2.$$

Algebra shows that

$$corr(r, r_I) = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}. \quad \square$$

Next we give an argument, due to Rathnayake and Olive (2023), for the Mallows (1973) C_p criterion when each submodel contains a constant. Let submodel I have $k \leq p$ predictors including

a constant. Then

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n$$

where MSE is for the full model,and $C_p(I) \geq -p$. Assume the full model is one of the submodels considered with $C_p(full) = p$, e.g. forward selection, backward elimination, stepwise selection, and all subsets selection. Then $-p \leq C_p(I_{min}) \leq p$. Let $\boldsymbol{r}$ be the residual vector for the full model and $\boldsymbol{r}_I$ that for the submodel. Then the correlation

$$corr(r, r_I) = \sqrt{\frac{n-p}{C_p(I) + n - 2k}}$$

by Theorem 2. Thus $corr(r, r_{I_{min}}) \to 1$ as $n \to \infty$. Suppose S is not a subset of I. Under the model $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$, $corr(r, r_I)$ will not converge to 1 as $n \to \infty$, and for large enough n, $[corr(r, r_I)]^2 \leq \gamma < 1$. Thus $C_p(I) \to \infty$ as $n \to \infty$. Hence $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ if the zero mean iid errors have constant variance σ^2.

Write the AR(p) equations $Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t$ in matrix form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ or

$$\begin{bmatrix} Y_{p+1} \\ Y_{p+2} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & Y_p & Y_{p-1} & \dots & Y_1 \\ 1 & Y_{p+1} & Y_p & \dots & Y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Y_{n-1} & Y_{n-2} & \dots & Y_{n-p} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix} + \begin{bmatrix} e_{p+1} \\ e_{p+2} \\ \vdots \\ e_n \end{bmatrix} \tag{2.2}$$

where $\boldsymbol{X}$ is of full rank with more rows than columns $p + 1$ and $\boldsymbol{\beta} = (\phi_0, \boldsymbol{\phi}^T)^T = (\phi_0, \phi_1, ..., \phi_p)^T$.

If the C_p criterion is applied to the AR(1) model, then the AR(2) model, ..., then the AR(p) model, and if the full AR(p) model is good, then the probability of the C_p criterion underfitting goes to 0 as $n \to \infty$.

Heuristically, the underfitting arqument for the MA(q) model is similar. Suppose the MA(q) model is fitted and the residuals are obtained. Substitute the residuals for the errors to get a working

model $Y_t \approx \theta_0 + \theta_1 r_{t-1} + \cdots + \theta_q r_{t-q} + e_t$ in matrix form $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$

$$\begin{bmatrix} Y_{q+1} \\ Y_{q+2} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & r_q & r_{q-1} & \dots & r_1 \\ 1 & r_{q+1} & r_q & \dots & r_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & r_{n-1} & r_{n-2} & \dots & r_{n-q} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_q \end{bmatrix} + \begin{bmatrix} e_{q+1} \\ e_{q+2} \\ \vdots \\ e_n \end{bmatrix} \tag{2.3}$$

where $\boldsymbol{X}$ is of full rank with more rows than columns $q + 1$ and $\boldsymbol{\beta} = (\theta_0, \theta_1, ..., \theta_q)^T$. If the true dimension is $q_S \leq q$, then for large n, the residuals converge to the errors for MA(k) models with $q_S \leq k \leq q$. If $k < q_S$, then the model does not fit well so the residuals tend to larger in magnitude than the errors for large n. Hence the C_p criterion should be too large for $k < q_S$ due to both underfitting and the bad approximation of the residuals for the errors. Hence we expect the probability of underfitting using the working AR(k) models to go to zero.

Heuristically, the AIC and BIC criterion, given by Equation (2.4) below, are a lot like the C_p criterion for AR(p) and MA(q) models, so we expect the probability of underfitting to go to zero as $n \to \infty$. For ARMA(p, q) models, let $\log(\hat{L})$ be the log likelihood for the GMLE. Then the AIC and BIC criteria have the form $-2\log(\hat{L}) + (p + q)c(n)$ where $c(n) = 2$ for AIC and $c(n) = \log(n)$ for BIC. From McElroy and Politis (2020, p. 360), $-2\log(\hat{L}) \approx n\log(\hat{\sigma}_I^2) + a_n$ where $\hat{\sigma}_I^2$ is the GMLE of the error variance of model I and a_n is a constant that depends on n. Hence if I is an ARMA(p, q) model, take

$$AIC(I) = n\log(\hat{\sigma}_I^2) + 2(p + q) \text{ and } \mathrm{BIC(I)} = \mathrm{n}\log(\hat{\sigma}_\mathrm{I}^2) + (\mathrm{p} + \mathrm{q})\log(\mathrm{n}). \tag{2.4}$$

For AIC given by (2.4), let $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, and models with $4 \leq \Delta(I) \leq 7$ are borderline. See Brockwell and Davis (1987, p. 269), Duong (1984), and Burnham and Anderson (2004). Claeskens and Hjort (2008, pp. 39, 111) use slightly different formulas for AR(p) models Pötscher and Srinivasan (1994) multiply the Equation (2.4) formulas by $1/n$. In the literature and software, the criterion can take many forms since the criterion can

be multiplied by a positive constant, such as $1/n$, and a constant d_n can be added to the criterion without changing the model that minimizes the criterion. Parameters that are in every model, such as σ^2 and possibly a constant, can be absorbed in a constant d_n.

Two *tspack* functions are useful for illustrating least squares (OLS) applied to AR and MA time series. The function `arp` fits an AR(p) model to time series Y using OLS, and makes a response and residual plot. The function assumes $p < n - p$, and $(n - p) > 10p$ would be useful. The function `maq` fits an MA(q) model to time series Y using OLS and the residuals from the MA GMLE, and makes a response and residual plot. The function assumes $q < n - q$, and $(n - q) > 10q$ would be useful. The output below shows $n = 100$ sometimes gives useful estimates for MA(2) models, but $n = 1000$ works better. The term thetahat give the OLS estimates using Equation (2.3) while macoef gives the GMLE estimates.

```
#maq function fits OLS model to MA(q) model using MA(q) residuals
N=100
thet1 = -0.5
thet2 = 0.2
y <- arima.sim(n=N, list(ma=c(thet1,thet2)), innov = rnorm(N))
maq(Y=y,q=2)
$macoef
        ma1         ma2   intercept
-0.22283269  0.09356583 -0.01296065
$thetahat
 (Intercept)           V2           V3
-0.007946074 -0.220555232  0.086028835

y <- arima.sim(n=N, list(ma=c(thet1,thet2)), innov = rt(N,5))
maq(Y=y,q=2)
$macoef
```

```
       ma1        ma2  intercept
-0.5172023  0.1352083  0.0669141
$thetahat
(Intercept)          V2          V3
 0.05434624 -0.53197103  0.20387859

y <- arima.sim(n=N, list(ma=c(thet1,thet2)), innov = runif(N,min=-1,max=1))
maq(Y=y,q=2)
$macoef
        ma1         ma2   intercept
-0.46512124  0.26736265 -0.00300503
$thetahat
 (Intercept)           V2           V3
 0.004146452 -0.440160622  0.217768087

y <- arima.sim(n=N, list(ma=c(thet1,thet2)), innov = (rexp(N)-1))
maq(Y=y,q=2)
$macoef
       ma1        ma2  intercept
-0.5120401  0.0233581 -0.1096992
$thetahat
(Intercept)          V2          V3
-0.13061379 -0.49408186  0.04325158

N=1000
thet1 = -0.5
thet2 = 0.2
```

```
y <- arima.sim(n=N, list(ma=c(thet1,thet2)), innov = rnorm(N))
maq(Y=y,q=2)
$macoef
        ma1         ma2   intercept
-0.48106432  0.19112093  0.02784754
$thetahat
(Intercept)          V2          V3
 0.02789364 -0.47540835  0.17613383

y <- arima.sim(n=N, list(ma=c(thet1,thet2)), innov = rt(N,5))
 maq(Y=y,q=2)
$macoef
        ma1         ma2   intercept
-0.50790935  0.20230121 -0.01374076
$thetahat
(Intercept)          V2          V3
-0.01282107 -0.49810238  0.18459068

y <- arima.sim(n=N, list(ma=c(thet1,thet2)), innov = runif(N,min=-1,max=1))
maq(Y=y,q=2)
$macoef
         ma1          ma2    intercept
-0.485228859  0.174623138  0.007440461
$thetahat
(Intercept)          V2          V3
 0.00732724 -0.47524982  0.15228861
```

```
y <- arima.sim(n=N, list(ma=c(thet1,thet2)), innov = (rexp(N)-1))
maq(Y=y,q=2)
$macoef
        ma1         ma2   intercept
-0.50635001  0.24779131  0.01544221
$thetahat
(Intercept)          V2          V3
 0.01567763 -0.49909346  0.23753909

N=10000
thet1 = -0.5
thet2 = 0.2

y <- arima.sim(n=N, list(ma=c(thet1,thet2)), innov = rnorm(N))
maq(Y=y,q=2)
$macoef
         ma1          ma2    intercept
-0.502345461  0.188665700 -0.002417998
$thetahat
 (Intercept)           V2           V3
-0.002494226 -0.502138785  0.186856927

y <- arima.sim(n=N, list(ma=c(thet1,thet2)), innov = rt(N,5))
maq(Y=y,q=2)
$macoef
         ma1          ma2    intercept
```

```
-0.500973647  0.183052482  0.007249818
$thetahat
 (Intercept)           V2           V3
 0.007160961 -0.502497557  0.187213111

y <- arima.sim(n=N, list(ma=c(thet1,thet2)), innov = runif(N,min=-1,max=1))
maq(Y=y,q=2)
$macoef
        ma1         ma2   intercept
-0.48664358  0.19871333  0.00688937
$thetahat
 (Intercept)           V2           V3
 0.007045233 -0.485846742  0.196911171

y <- arima.sim(n=N, list(ma=c(thet1,thet2)), innov = (rexp(N)-1))
maq(Y=y,q=2)
$macoef
          ma1           ma2     intercept
-0.5210335442  0.2285755363  0.0002545913
$thetahat
  (Intercept)            V2            V3
 0.0002103866 -0.5167212868  0.2197263226
```

CHAPTER 3

LARGE SAMPLE THEORY FOR SOME MODEL SELECTION ESTIMATORS

Some notation is needed for the large sample theory. The Gaussian maximum likelihood estimator (GMLE) will be used. The Yule Walker and least squares estimators will also be used for AR(p) models. Let the r_i be the m (one step ahead) residuals where often $m = n$ or $m = n - p$. Under regularity conditions,

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{m} r_i^2}{m - p - q - c} \tag{3.1}$$

is a consistent estimator of σ^2 where often $c = 0$ or $c = 1$. See Granger and Newbold (1977, p. 85) and Pankratz (1983, p. 206). Let $\hat{\sigma}^2$ be the estimator of σ^2 produced by the time series model. Let

$$\mathbf{\Gamma}_n = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{bmatrix}.$$

The following large sample theorem for the AR(p) model is due to Mann and Wald (1943). Also see McElroy and Politis (2020, p. 333) and Anderson (1971, pp. 210-217). For large sample theory for MA and ARMA models, see Hannan (1973), Kreiss (1985), and Yao and Brockwell (2006). There is a strong regularity condition for the GMLE for the ARMA model. Assume the ARMA(p_S, q_S) model is the true model. If both $p > p_S$ and $q > q_S$, then the GMLE is not a consistent estimator. See Chan, Ling, and Yau (2020) and Hannan (1980). Pötscher (1990) shows how to estimate $\max(p_S, q_S)$ consistently.

Theorem 3.1. Let the iid zero mean e_i have variance σ^2, and let the time series have mean $E(Y_t) = \mu$.

a) Let $Y_1, ..., Y_n$ be a weakly stationary and invertible AR(p) time series, and let $\boldsymbol{\beta} =$

$(\phi_1, ..., \phi_p)$. Let $\hat{\boldsymbol{\beta}}$ be the Yule Walker estimator of $\boldsymbol{\beta}$. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}) \tag{3.2}$$

where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta}) = \sigma^2 \boldsymbol{\Gamma}_p^{-1}$. Equation (3.2) also holds under mild regularity conditions for the least squares estimator, and the GMLE of $\boldsymbol{\beta}$.

b) Let $Y_1, ..., Y_n$ be a weakly stationary, causal, and invertible MA(q) time series, and let $\boldsymbol{\beta} = (\theta_1, ..., \theta_q)$. Let $\hat{\boldsymbol{\beta}}$ be the GMLE. Under regularity conditions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_q(\mathbf{0}, \boldsymbol{V}). \tag{3.3}$$

where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta}) = \sigma^2 \boldsymbol{\Gamma}_q^{-1}$.

c) Let $Y_1, ..., Y_n$ be a weakly stationary, causal, and invertible ARMA(p, q) time series, and let $\boldsymbol{\beta} = (\phi_1, ..., \phi_p, \theta_1, ..., \theta_q)$ with $g = p + q$. Let $\hat{\boldsymbol{\beta}}$ be the GMLE. Under regularity conditions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{V}) \tag{3.4}$$

where $\boldsymbol{V}$ depends only on the autocorrelation function.

The main point of Theorem 3.1 is that the theory can hold even if the e_t are not iid $N(0, \sigma^2)$. The basic idea for the GMLE is that $\{Y_t\}$ satisfies an AR(∞) model which is approximately an AR(p_y) model, and the large sample theory for the AR(p_y) model depends on the zero mean error distribution through σ^2 by Theorem 3.1a). See Anderson (1971: ch. 5, 1977), Durbin (1959), Hamilton (1994, pp. 117, 429), Hannan and Rissanen (1982, p. 85), and Whittle (1953). When the e_t are iid $N(0, \sigma_e^2)$, $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{I}_1^{-1}(\boldsymbol{\beta})$, the inverse information matrix. Then for the AR(p) model, $\boldsymbol{V}(\boldsymbol{\phi}) = \sigma^2 \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi}) = \boldsymbol{I}_1^{-1}(\boldsymbol{\phi})$, while for the MA($q$) model, $\boldsymbol{V}(\boldsymbol{\theta}) = \sigma^2 \boldsymbol{\Gamma}_q^{-1}(\boldsymbol{\theta}) = \boldsymbol{I}_1^{-1}(\boldsymbol{\theta})$. See Box and Jenkins (1976, p. 241) and McElroy and Politis (2020, pp. 340-344).

CHAPTER 4

DATA SPLITTING

Data splitting is used to get valid inference. If the model was selected without using the time series, then the model has an asymptotic normal distribution that can be used for inference. If the entire time series is used to build or select the model, then the resulting model tends not to have an asymptotic normal distribution due to selection bias. If the first half of the time series is used to build or select the model, and that model is fit on the second half of the time series, then inference is valid (the model for the second half of the time series was selected without using the second half). See, for example, Hurvich and Tsai (1989). Time series models are often built or selected using the entire data set with transformations such as the log transformation, model selection with AIC, BIC, or AIC_C. Plots such as the ACF and PACF are also used to select the model.

An application of data splitting is to use a model selection method on H to get a model I. On the validation set V, fit time series model I. Then use the standard time series inference. For AR model selection and MA model selection, data splitting works if the selected model does not underfit. For the GMLE and an ARMA model, assume the ARMA(p_S, q_S) model is the true model. Then the selected model I is an ARMA(p_I, q_I) models. The model I should not underfit ($p_I \geq p_S$ and $q_I \geq q_S$), and needs $p_I = p_S$ or $q_I = q_S$ for valid inference.

4.1 SIMULATIONS

All of the simulations used four error types: etype = 1 for N(0,1) errors, etype = 2 for t distribution with tdf=degrees of freedom, etype = 3 for U(-1,1) errors, and etype=4 for EXP(1)−1 errors. The data splitting functions used $n_h = floor(n/2)$. Let pmax be the maximum AR order and qmax the maximum MA order. 1−undfit is the proportion of times a consistent model was selected.

For data splitting, the amount of underfitting often depended on the parameters in the model and the model selection method. For AR(p) data splitting, the *tspack* function `dsarsim` used

the built in "AIC" model selection from the *R* function `ar` and tended to underfit for $n < 20$ pmax. The *tspack* function `dsarsim2` used Equation (2.4) for AR(p) model selection with AIC for $n < 14(pmax)$ and BIC otherwise. AIC tended to overfit while BIC tended to underfit until n got larger. An AR(1) model with $\phi = 0.5$ and an AR(2) model with $\boldsymbol{\phi} = (0.5, 0.33)$ were used.

For MA(q) data splitting, the *tspack* function `dsmasim` used the Hyndman and Khandakar (2008) *forecast package* was used for data splitting programs. Also see Hyndman and Athanasopoulos (2018). Model selection was done with the auto.arima function using "AIC" model selection. This model selection tended to underfit for $n < 20$ pmax. The *tspack* function `dsmasim2` used Equation (2.4). This function can use a penalty similar to the second ARMA function described below. Neither AIC nor BIC tended to underfit, so the default program used BIC. An MA(1) model with $\theta = -0.5$ and an MA(2) model with $\boldsymbol{\theta} = (-0.5, 0.5)$ were used.

Table 4.1. AR Simulation Proportion of Underfitting, $n_H = n/2$, p=13, nruns=1000

n	dist	ϕ	R AIC	Eq. (2.4)
100	N	(0.5)	0.037	0.001
150	N	(0.5)	0.006	0.010
200	N	(0.5)	0.000	0.000
100	t	(0.5)	0.034	0.000
150	t	(0.5)	0.006	0.000
200	t	(0.5)	0.000	0.000
100	unif	(0.5)	0.029	0.001
150	unif	(0.5)	0.003	0.000
200	unif	(0.5)	0.001	0.000
100	exp	(0.5)	0.018	0.002
150	exp	(0.5)	0.004	0.001
200	exp	(0.5)	0.000	0.000
100	N	(0.5,0.33)	0.350	0.031
150	N	(0.5,0.33)	0.148	0.005
200	N	(0.5,0.33)	0.072	0.029
100	t	(0.5,0.33)	0.342	0.011
150	t	(0.5,0.33)	0.149	0.021
200	t	(0.5,0.33)	0.062	0.021
100	unif	(0.5,0.33)	0.358	0.019
150	unif	(0.5,0.33)	0.174	0.001
200	unif	(0.5,0.33)	0.075	0.041
100	exp	(0.5,0.33)	0.350	0.018
150	exp	(0.5,0.33)	0.146	0.001
200	exp	(0.5,0.33)	0.055	0.041

Table 4.2. MA Simulation Proportion of Underfitting, $n_H = n/2$, q=13, nruns=1000

n	dist	θ	R AIC	Eq. (2.4)
100	N	(−0.5)	0.104	0.002
150	N	(−0.5)	0.017	0.001
200	N	(−0.5)	0.002	0.000
100	t	(−0.5)	0.081	0.007
150	t	(−0.5)	0.014	0.001
200	t	(−0.5)	0.001	0.000
100	unif	(−0.5)	0.108	0.005
150	unif	(−0.5)	0.019	0.002
200	unif	(−0.5)	0.001	0.000
100	exp	(−0.5)	0.111	0.007
150	exp	(−0.5)	0.023	0.001
200	exp	(−0.5)	0.002	0.000
100	N	(−0.5,0.5)	0.083	0.011
150	N	(−0.5,0.5)	0.012	0.002
200	N	(−0.5,0.5)	0.002	0.000
100	t	(−0.5,0.5)	0.073	0.014
150	t	(−0.5,0.5)	0.009	0.000
200	t	(−0.5,0.5)	0.000	0.000
100	unif	(−0.5,0.5)	0.092	0.006
150	unif	(−0.5,0.5)	0.011	0.009
200	unif	(−0.5,0.5)	0.002	0.000
100	exp	(−0.5,0.5)	0.058	0.006
150	exp	(−0.5,0.5)	0.004	0.002
200	exp	(−0.5,0.5)	0.001	0.000

For ARMA(p_{max}, q_{max}) model selection, the underfitting using the AIC criterion was often severe. Using the Pötscher (1990) method to estimate $r = \max(p_S, q_s)$ often worked well. Also see Chan, Ling, and Yau (2020) and Pötscher and Srinivasan (1994). Let k_{max} be a positive integer such as $p_{max} = q_{max} = k_{max} = 5$. Fit the ARMA($k, k$) model for $k = 0, 1, ..., k_{max}$. For each of these $k_{max} + 1$ models, compute the BIC–type criterion $z(k) = \log(\hat{\sigma}_k^2) + 2k\log(n)/n$ where $\hat{\sigma}_k^2$ is the GMLE estimator of the error (or innovation) variance σ^2. This criterion is Equation (2.4) divided by n. The estimator $\hat{r}$ of r is the first local minimum of the series $z(0), z(1), ..., z(k_{max})$. Hence $\hat{r} = 0$ if $z(0) \leq z(1)$; $\hat{r} = 1$ if $z(0) > z(1)$ and $z(1) \leq z(2)$; $\hat{r} = 2$ if $z(0) > z(1)$, $z(1) > z(2)$, and $z(2) \leq z(3)$; $\hat{r} = k$ if $z(r) > z(r+1)$ for $0 \leq r < k$ and $z(k) \leq z(k+1)$ for $k = 0, ..., k_{max} - 1$; and $\hat{r} = k_{max}$ if

$z(k)$ is not a local minimum for any $k = 0, 1, ..., k-1$. Note that $r \leq k_{max}$ is necessary for $\hat{r}$ to be a consistent estimator of r.

The simplest ARMA model selection procedure uses the ARMA($\hat{r}, \hat{r}$) model as the final model. This model selection method had roughly 15% underfitting for $n = 1000$, 5% underfitting for $n = 1500$, 1% underfitting for $n = 2000$, while overfitting with $\hat{r} > r = max(p_S, q_S)$ was rare.

The *R* function `armamsel1` does this model selection while the function `armasim1` does the simulation. The 6 time series types are tstype=1 for an AR(1) model with $\phi = 0.5$, tstype=2 for an AR(2) model with $\boldsymbol{\phi} = (0.5, 0.33)$, tstype=3 for an MA(1) model with $\theta = -0.5$, tstype=4 for an MA(2) model with $\boldsymbol{\theta} = (-0.5, 0.5)$, tstype=5 for an ARMA(3,1) model with $\boldsymbol{\phi} = (0.7, 0.1, -0.4)$ and $\theta = 0.1$. Finally, tstype=6 allows the user to specify $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ for and ARMA(p, q) model with $p \geq 1$, $q \geq 1$, and $p, q \leq kmax$ where kmax is the largest value of r for the fitted ARMA(r, r) models, $r = 0, 1, ...,$kmax. The ARMA models were sensitive to the values of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$.

For ARMA(1,1) models, $(\phi, \theta) = (0.5, 0.2), (0.2, 0.1), (0.4, -0.1), (0.6, -0.3), (-0.2, 0.6)$, $(-0.4, 0.8), (-0.6, 1.0), (0.2, -0.5), (0.4, -0.7), (-0.2, -0.1), (-0.4, 0.1), (-0.6, 0.4), (-0.8, 0.6)$ worked well with $n = 1000$, but $(\phi, \theta) = (0.5, -0.5)$ did not. From Tables 4.3-4.8, model selection could work fairly well for n as low as 80, but often much larger sample sizes were needed.

A new ARMA model selection procedure first finds $\hat{r}$ as above. AIC and BIC type criteria can be multiplied by a positive constant, and constants can be added to a criterion without changing the model that minimizes the criterion. If a parameter is in all of the models, e.g. a constant or $\hat{\sigma}^2$, then with respect to the penalty, that parameter acts like adding a constant to the criterion. We used the Equation (2.4) AIC type criterion $AIC(I) = n \log(\hat{\sigma}_I^2) + 2(p + q)$ where I is an ARMA(p, q) model and $\hat{\sigma}_I^2$ is the GMLE estimator of the error (or innovation) variance σ_I^2. With this scaling, a decrease of AIC > 2 when one parameter is omitted suggests that the parameter was not needed. Let pen be a penalty such as pen=2 or pen=0. The algorithm computes the $crit = AIC(I) - pen$ for the ARMA($\hat{r}, \hat{r}$) model, and fits the ARMA($\hat{r} - i, \hat{r}$) and ARMA($\hat{r}, \hat{r} - i$) models for $i = 0, ..., \hat{r} - 1$. If one of the models has AIC(I) < crit, then the set crit = AIC(I) − pen. This process is repeated

at each step. The value of crit is updated only if a decrease of more than pen from the current value of crit is observed. The final model I is the model selected by this algorithm. This additional penalty decreased the amount of underfitting. Note that $2\hat{r}$ models are fitted after finding $\hat{r}$, which fits $k_{max} + 1$ models. This method is faster than computing the AIC for $(k_{max} + 1)^2$ models.

The *R* function `armasim2` has a default value of pen=2, but using another value, such as $2 + 10/n$ can be used. Take the ARMA$(p, \hat{r})$ or ARMA$(\hat{r}, q)$ model that has the smallest value of crit. Then at least one of p and q will equal $\hat{r}$. The function `armamsel2` does this model selection while the function `armasim2` does the simulation. Again *rtrue* gives the proportion of runs where $\hat{r} = r$, while *cfit* gives the proportion of runs where a consistent ARMA(p,q) model is selected. The selected model is consistent if i) $p = p_S$ and $q = q_S$, or if ii) $p = p_S$ and $q > q_S$, or if iii) $p > p_S$ and $q = q_S$. See discussion near Theorem 3.1. Underfitting occurs if $p < p_S$ or $q < q_S$ while overfitting that causes inconsistency occurs if $p > p_S$ and $q > q_S$.

Table 4.3. ARMA, Proportion Consistent Model is Selected, nruns=1000, tstype=1

n	dist	ϕ	R AIC	$\hat{r}$	I
50	N	(0.5)	0.653	0.698	0.698
80	N	(0.5)	0.766	0.868	0.859
100	N	(0.5)	0.752	0.931	0.931
200	N	(0.5)	0.845	0.991	0.999
350	N	(0.5)	0.888	0.999	0.999
500	N	(0.5)	0.852	0.999	0.999
800	N	(0.5)	0.863	1.000	1.000
1000	N	(0.5)	0.939	1.000	1.000
1500	N	(0.5)	0.924	1.000	1.000
2000	N	(0.5)	0.906	1.000	1.000
50	t	(0.5)	0.653	0.705	0.699
80	t	(0.5)	0.758	0.899	0.899
100	t	(0.5)	0.772	0.958	0.949
200	t	(0.5)	0.870	0.987	0.987
350	t	(0.5)	0.894	0.991	0.991
500	t	(0.5)	0.894	1.000	1.000
800	t	(0.5)	0.924	0.999	0.999
1000	t	(0.5)	0.875	1.000	1.000
1500	t	(0.5)	0.883	1.000	1.000
2000	t	(0.5)	0.892	1.000	1.000
50	unif	(0.5)	0.581	0.754	0.754
80	unif	(0.5)	0.752	0.875	0.875
100	unif	(0.5)	0.759	0.897	0.897
200	unif	(0.5)	0.879	0.999	0.999
350	unif	(0.5)	0.859	1.000	1.000
500	unif	(0.5)	0.818	1.000	1.000
800	unif	(0.5)	0.869	1.000	1.000
1000	unif	(0.5)	0.855	1.000	1.000
1500	unif	(0.5)	0.866	1.000	1.000
2000	unif	(0.5)	0.947	1.000	1.000
50	exp	(0.5)	0.739	0.697	0.681
80	exp	(0.5)	0.708	0.855	0.855
100	exp	(0.5)	0.741	0.956	0.956
200	exp	(0.5)	0.891	0.997	0.997
350	exp	(0.5)	0.889	0.998	0.998
500	exp	(0.5)	0.849	0.999	0.999
800	exp	(0.5)	0.887	1.000	1.000
1000	exp	(0.5)	0.871	1.000	1.000
1500	exp	(0.5)	0.981	1.000	1.000
2000	exp	(0.5)	0.939	1.000	1.000

Table 4.4. ARMA, Proportion Consistent Model is Selected, nruns=1000, tstype=2

n	dist	ϕ	R AIC	$\hat{r}$	I
50	N	(0.5, 0.33)	0.410	0.156	0.148
80	N	(0.5, 0.33)	0.469	0.065	0.059
100	N	(0.5, 0.33)	0.452	0.107	0.107
200	N	(0.5, 0.33)	0.507	0.081	0.081
350	N	(0.5, 0.33)	0.551	0.236	0.229
500	N	(0.5, 0.33)	0.632	0.415	0.415
800	N	(0.5, 0.33)	0.673	0.567	0.559
1000	N	(0.5, 0.33)	0.677	0.735	0.735
1500	N	(0.5, 0.33)	0.766	0.935	0.935
2000	N	(0.5, 0.33)	0.824	0.983	0.983
50	t	(0.5, 0.33)	0.371	0.141	0.139
80	t	(0.5, 0.33)	0.519	0.106	0.107
100	t	(0.5, 0.33)	0.624	0.045	0.045
200	t	(0.5, 0.33)	0.532	0.166	0.159
350	t	(0.5, 0.33)	0.521	0.186	0.186
500	t	(0.5, 0.33)	0.574	0.348	0.348
800	t	(0.5, 0.33)	0.729	0.648	0.648
1000	t	(0.5, 0.33)	0.781	0.714	0.699
1500	t	(0.5, 0.33)	0.769	0.857	0.857
2000	t	(0.5, 0.33)	0.809	0.989	0.989
50	unif	(0.5, 0.33)	0.379	0.078	0.069
80	unif	(0.5, 0.33)	0.389	0.145	0.145
100	unif	(0.5, 0.33)	0.459	0.126	0.126
200	unif	(0.5, 0.33)	0.590	0.157	0.157
350	unif	(0.5, 0.33)	0.619	0.191	0.191
500	unif	(0.5, 0.33)	0.588	0.385	0.385
800	unif	(0.5, 0.33)	0.587	0.567	0.567
1000	unif	(0.5, 0.33)	0.698	0.627	0.619
1500	unif	(0.5, 0.33)	0.686	0.896	0.889
2000	unif	(0.5, 0.33)	0.857	0.999	0.999
50	exp	(0.5, 0.33)	0.289	0.179	0.179
80	exp	(0.5, 0.33)	0.509	0.139	0.119
100	exp	(0.5, 0.33)	0.491	0.129	0.129
200	exp	(0.5, 0.33)	0.423	0.198	0.179
350	exp	(0.5, 0.33)	0.639	0.199	0.199
500	exp	(0.5, 0.33)	0.593	0.366	0.359
800	exp	(0.5, 0.33)	0.609	0.627	0.627
1000	exp	(0.5, 0.33)	0.589	0.657	0.657
1500	exp	(0.5, 0.33)	0.716	0.936	0.929
2000	exp	(0.5, 0.33)	0.768	0.979	0.979

Table 4.5. ARMA, Proportion Consistent Model is Selected, nruns=1000, tstype=3

n	dist	θ	R AIC	$\hat{r}$	I
50	N	(-0.5)	0.788	0.845	0.845
80	N	(-0.5)	0.845	0.876	0.876
100	N	(-0.5)	0.805	0.925	0.931
200	N	(-0.5)	0.816	0.997	0.997
350	N	(-0.5)	0.909	0.998	0.998
500	N	(-0.5)	0.859	0.999	0.999
800	N	(-0.5)	0.831	0.989	0.989
1000	N	(-0.5)	0.889	1.000	1.000
1500	N	(-0.5)	0.965	1.000	1.000
2000	N	(-0.5)	0.915	1.000	1.000
50	t	(-0.5)	0.708	0.815	0.815
80	t	(-0.5)	0.874	0.891	0.891
100	t	(-0.5)	0.851	0.952	0.952
200	t	(-0.5)	0.815	0.973	0.973
350	t	(-0.5)	0.836	0.999	0.999
500	t	(-0.5)	0.893	0.999	0.999
800	t	(-0.5)	0.865	0.998	0.998
1000	t	(-0.5)	0.916	1.000	1.000
1500	t	(-0.5)	0.929	1.000	1.000
2000	t	(-0.5)	0.928	1.000	1.000
50	unif	(-0.5)	0.788	0.832	0.832
80	unif	(-0.5)	0.787	0.925	0.925
100	unif	(-0.5)	0.815	0.956	0.956
200	unif	(-0.5)	0.863	0.976	0.976
350	unif	(-0.5)	0.894	0.986	0.986
500	unif	(-0.5)	0.926	1.000	1.000
800	unif	(-0.5)	0.881	1.000	1.000
1000	unif	(-0.5)	0.892	1.000	1.000
1500	unif	(-0.5)	0.919	1.000	1.000
2000	unif	(-0.5)	0.889	1.000	1.000
50	exp	(-0.5)	0.753	0.804	0.804
80	exp	(-0.5)	0.735	0.945	0.945
100	exp	(-0.5)	0.916	0.951	0.951
200	exp	(-0.5)	0.862	0.976	0.976
350	exp	(-0.5)	0.946	0.985	0.985
500	exp	(-0.5)	0.926	0.983	0.983
800	exp	(-0.5)	0.887	0.987	0.987
1000	exp	(-0.5)	0.865	1.000	1.000
1500	exp	(-0.5)	0.856	1.000	1.000
2000	exp	(-0.5)	0.928	1.000	1.000

Table 4.6. ARMA, Proportion Consistent Model is Selected, nruns=1000, tstype=4

n	dist	θ	R AIC	$\hat{r}$	I
50	N	(-0.5, 0.5)	0.435	0.415	0.489
80	N	(-0.5, 0.5)	0.679	0.456	0.447
100	N	(-0.5, 0.5)	0.745	0.667	0.675
200	N	(-0.5, 0.5)	0.915	0.956	0.956
350	N	(-0.5, 0.5)	0.901	0.999	0.999
500	N	(-0.5, 0.5)	0.886	0.999	0.999
800	N	(-0.5, 0.5)	0.855	1.000	1.000
1000	N	(-0.5, 0.5)	0.905	1.000	1.000
1500	N	(-0.5, 0.5)	0.886	1.000	1.000
2000	N	(-0.5, 0.5)	0.896	1.000	1.000
50	t	(-0.5, 0.5)	0.407	0.351	0.352
80	t	(-0.5, 0.5)	0.705	0.485	0.485
100	t	(-0.5, 0.5)	0.806	0.633	0.633
200	t	(-0.5, 0.5)	0.915	0.915	0.916
350	t	(-0.5, 0.5)	0.936	0.996	0.996
500	t	(-0.5, 0.5)	0.946	0.997	0.997
800	t	(-0.5, 0.5)	0.905	0.999	0.999
1000	t	(-0.5, 0.5)	0.908	1.000	1.000
1500	t	(-0.5, 0.5)	0.937	1.000	1.000
2000	t	(-0.5, 0.5)	0.966	1.000	1.000
50	unif	(-0.5, 0.5)	0.456	0.409	0.408
80	unif	(-0.5, 0.5)	0.737	0.617	0.618
100	unif	(-0.5, 0.5)	0.787	0.635	0.636
200	unif	(-0.5, 0.5)	0.918	0.908	0.905
350	unif	(-0.5, 0.5)	0.908	0.973	0.974
500	unif	(-0.5, 0.5)	0.947	0.998	0.998
800	unif	(-0.5, 0.5)	0.915	1.000	1.000
1000	unif	(-0.5, 0.5)	0.916	1.000	1.000
1500	unif	(-0.5, 0.5)	0.916	1.000	1.000
2000	unif	(-0.5, 0.5)	0.927	1.000	1.000
50	exp	(-0.5, 0.5)	0.322	0.339	0.341
80	exp	(-0.5, 0.5)	0.689	0.472	0.472
100	exp	(-0.5, 0.5)	0.769	0.661	0.661
200	exp	(-0.5, 0.5)	0.911	0.925	0.925
350	exp	(-0.5, 0.5)	0.927	0.989	0.989
500	exp	(-0.5, 0.5)	0.909	1.000	1.000
800	exp	(-0.5, 0.5)	0.965	1.000	1.000
1000	exp	(-0.5, 0.5)	0.901	1.000	1.000
1500	exp	(-0.5, 0.5)	0.948	1.000	1.000
2000	exp	(-0.5, 0.5)	0.889	1.000	1.000

Table 4.7. ARMA, Proportion Consistent Model is Selected, nruns=1000,tstype=5

n	dist	ϕ	θ	R AIC	$\hat{r}$	I
500	N	(0.7, 0.1, −0.4)	(0.1)	0.246	0.381	0.364
800	N	(0.7, 0.1, −0.4)	(0.1)	0.294	0.692	0.609
1000	N	(0.7, 0.1, −0.4)	(0.1)	0.364	0.648	0.619
1500	N	(0.7, 0.1, −0.4)	(0.1)	0.382	0.943	0.832
2000	N	(0.7, 0.1, −0.4)	(0.1)	0.378	1.000	0.954
500	t	(0.7, 0.1, −0.4)	(0.1)	0.213	0.354	0.324
800	t	(0.7, 0.1, −0.4)	(0.1)	0.418	0.654	0.584
1000	t	(0.7, 0.1, −0.4)	(0.1)	0.310	0.793	0.739
1500	t	(0.7, 0.1, −0.4)	(0.1)	0.380	0.939	0.793
2000	t	(0.7, 0.1, −0.4)	(0.1)	0.415	0.969	0.904
500	unif	(0.7, 0.1, −0.4)	(0.1)	0.219	0.409	0.339
800	unif	(0.7, 0.1, −0.4)	(0.1)	0.309	0.713	0.683
1000	unif	(0.7, 0.1, −0.4)	(0.1)	0.218	0.839	0.761
1500	unif	(0.7, 0.1, −0.4)	(0.1)	0.351	0.964	0.879
2000	unif	(0.7, 0.1, −0.4)	(0.1)	0.472	0.983	0.963
500	exp	(0.7, 0.1, −0.4)	(0.1)	0.279	0.362	0.301
800	exp	(0.7, 0.1, −0.4)	(0.1)	0.271	0.619	0.519
1000	exp	(0.7, 0.1, −0.4)	(0.1)	0.301	0.774	0.663
1500	exp	(0.7, 0.1, −0.4)	(0.1)	0.451	0.939	0.889
2000	exp	(0.7, 0.1, −0.4)	(0.1)	0.381	0.994	0.909

Table 4.8. ARMA, Proportion Consistent Model is Selected, nruns=1000, tstype=6

n	dist	ϕ	θ	R AIC	$\hat{r}$	I
50	N	(0.4)	(−0.7)	0.084	0.561	0.561
80	N	(0.4)	(−0.7)	0.222	0.732	0.721
100	N	(0.4)	(−0.7)	0.324	0.779	0.779
200	N	(0.4)	(−0.7)	0.519	0.979	0.979
350	N	(0.4)	(−0.7)	0.612	0.999	0.999
500	N	(0.4)	(−0.7)	0.722	1.000	1.000
800	N	(0.4)	(−0.7)	0.767	1.000	1.000
1000	N	(0.4)	(−0.7)	0.768	1.000	1.000
1500	N	(0.4)	(−0.7)	0.783	1.000	1.000
2000	N	(0.4)	(−0.7)	0.839	1.000	1.000
50	t	(0.4)	(−0.7)	0.073	0.609	0.609
80	t	(0.4)	(−0.7)	0.268	0.709	0.709
100	t	(0.4)	(−0.7)	0.273	0.771	0.771
200	t	(0.4)	(−0.7)	0.488	0.944	0.944
350	t	(0.4)	(−0.7)	0.619	0.999	0.999
500	t	(0.4)	(−0.7)	0.630	1.000	1.000
800	t	(0.4)	(−0.7)	0.810	0.991	0.991
1000	t	(0.4)	(−0.7)	0.811	1.000	1.000
1500	t	(0.4)	(−0.7)	0.852	1.000	1.000
2000	t	(0.4)	(−0.7)	0.929	1.000	1.000
50	unif	(0.4)	(−0.7)	0.069	0.559	0.559
80	unif	(0.4)	(−0.7)	0.258	0.771	0.782
100	unif	(0.4)	(−0.7)	0.289	0.709	0.709
200	unif	(0.4)	(−0.7)	0.399	0.953	0.953
350	unif	(0.4)	(−0.7)	0.621	0.971	0.971
500	unif	(0.4)	(−0.7)	0.658	0.982	0.982
800	unif	(0.4)	(−0.7)	0.841	0.984	0.984
1000	unif	(0.4)	(−0.7)	0.879	1.000	1.000
1500	unif	(0.4)	(−0.7)	0.875	1.000	1.000
2000	unif	(0.4)	(−0.7)	0.901	1.000	1.000
50	exp	(0.4)	(−0.7)	0.058	0.569	0.569
80	exp	(0.4)	(−0.7)	0.241	0.641	0.641
100	exp	(0.4)	(−0.7)	0.301	0.812	0.812
200	exp	(0.4)	(−0.7)	0.612	0.959	0.959
350	exp	(0.4)	(−0.7)	0.675	0.998	0.998
500	exp	(0.4)	(−0.7)	0.701	0.995	0.995
800	exp	(0.4)	(−0.7)	0.809	1.000	1.000
1000	exp	(0.4)	(−0.7)	0.879	1.000	1.000
1500	exp	(0.4)	(−0.7)	0.888	1.000	1.000
2000	exp	(0.4)	(−0.7)	0.915	1.000	1.000

CHAPTER 5

OUTLIER DETECTION

Outliers are cases that lie far away from the pattern set by the bulk of the data, and can be often be detected from the plot of t versus Y_t and from the response plot of $\hat{Y}_t$ versus Y_t with the identity line that has zero intercept and unit slope added as a visual aid. In both plots Y_t is on the vertical axis, and the vertical deviations of Y_t from the identity line are the residuals $\hat{e}_t = Y_t - \hat{Y}_t$. The residual plot of $\hat{Y}_t$ versus $\hat{e}_t$ is also useful.

The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \text{ if n is odd,} \tag{5.1}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \text{ if n is even.}$$

The notation $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ will also be used. The *sample median absolute deviation* is

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n). \tag{5.2}$$

Assume the time series Y_t is weakly stationary with an MA(∞) representation. Let $up = \text{MED}(n) + k\text{MAD}(n)$ and $low = \text{MED}(n) - k\text{MAD}(n)$ where $k = 6$ is the default. Make a new time series W_t where if $low \leq Y_t \leq up$, then $W_t = Y_t$. Otherwise, make W_t a missing value: let $W_t = NA$ if $Y_t < low$ or $Y_t > up$. This method is useful since software methods for handling missing values are widely available. See, for example, Jones (1980). The method may also be useful for handling heavy tailed time series, where the first or second moment of the Y_t does not exist. Since the W_t do not depend on the ARMA model, plug in $W_1, ..., W_n$ into the time series software in place of the $Y_1, ..., Y_n$ to to get robust estimators of other quantities, such as the ACF and PACF.

One alternative is to get a robustly Winzorize the time series W_t: if $Y_t > up$, then $W_t = max(Y_k \leq up)$. If $Y_t < low$, then $W_t = min(Y_k \geq low)$. If $low \leq Y_t \leq up$, then $W_t = Y_t$. Then fit the time series to W_t. A second alternative would set $W_t = \text{MED}(n)$ instead of NA. Variants would use the fitted time series to predict the W_t that were changed, fit the time series again, and perhaps

repeat this step. These methods impute the potential outliers, but the existing methods for handling missing values likely impute better.

For an MA(q) model, the $Y_j, Y_{j+q+1}, Y_{j+2(q+1)}$, ... are iid. Hence there are $q+1$ iid sequences starting at $j = 1, ..., (q+1)$. Since the sample percentiles of the iid sequences converge in probability to the population percentiles for fixed h, so do the sample percentiles of all of the data. Hence the sample median and sample median absolute deviation converge to the corresponding population quantities, and similar results hold for MA(∞) models. Haile and Olive (2023b) used similar results to justify a time series prediction interval. Lee and Scholtes (2014) also examine when percentiles of forecast errors of ARMA models are consistent.

To see why $k = 6$ is recommended, examine the approximate proportion of cases not changed to NA for several distributions when no outliers are present. See Table 5.1. Let MED(X) and MAD(X) be the population median and median absolute deviation. Notation for the random variables is as in Olive (2008, ch. 10; 2014, ch. 10). For uniform and discrete uniform data, $k = 2$ asymptotically covers 100% of the data, so $k = 6$ can have trouble detecting moderate outliers. For the data used in Figure 5.1, the value $k = 6$ only caused one outlier to be changed to NA. The main differences between this procedure and other time series methods in the literature, are a) potential outliers are replaced by missing values rather than the median of time series values close to the outlier, and b) some theory is given for why the robust estimators estimate percentiles of the time series.

Table 5.1. Probability X $\in$ [MED(X) $-$ 6MAD(X), MED(X) + 6MAD(X)]

distribution of X	MED(X)	MAD(X)	prob
Cauchy(μ, σ)	μ	σ	0.8949
double exponential(θ, λ)	θ	$\log(2)\lambda$	0.9844
exponential(θ, λ)	$\theta + \log(2)\lambda$	$\lambda/2.0781$	0.9721
logistic(μ, σ)	μ	$\log(3)\sigma$	0.9973
N(μ, σ^2)	μ	σ	0.9999
uniform(θ_1, θ_2)	$(\theta_1 + \theta_2)/2$	$(\theta_2 - \theta_1)/4$	1

The time series outlier literature is large. See, for example, Agnieszka and Magdalena (2018),

Basu and Meckesheimer (2007), Bhatia, et al. (2016), Chakhchoukh (2010), Chang, Tiao, and Chen (1988), Chen and Liu (1993), Choy (2001), de Luna and Genton (2001), Deutsch, Richards, and Swain (1990), Fox (1972), Justel, Peña, and Tsay (2001), Lawrence (2014), Ledolter (1989), Liu, Kumar, and Palomar (2019), Lucas, Franses, and Van Dijk (2009), Stockinger and Dutter (1987), Tsay (1986, 1988). Blázquez-Garcïa, et al. (2020) give a review, but omit "robust statistics" contributions like Allende and Heiler (1992), Bustos and Yohai (1986), Denby and Martin (1979), Kreiss (1985), Ma and Genton (2000), and Muler, Peña, and Yohai (2009). The robust statistics contributions attempt to robustify the time series estimating equations. For example, AR(p) models can be fit with OLS, and the OLS estimator can be replaced by a robust regression estimator.

Most of the literature has been for additive outliers, but there are other types of outliers in the literature. See, for example, Chan (1995). $Z_t = Y_t + W_t$ is an *additive outlier* if Y_t follows the time series model and W_t does not affect future values of the time series $Y - t + k$. $Z_t = Y_t$ is an *innovative outlier* (or innovation outlier or innovational outlier) if $Z_t = Y_t$ is far from the bulk of the data, and future values depend on Z_t according to the true time series model. The effects of an innovative outlier can take a long time to diminish. Thus an additive outlier only affects the value of the given observation while an innovational or innovative outlier affects all observations beyond the given time through the memory of the underlying ARMA process. Z_t is a *temporary change outlier* if its effects on future data diminish very quickly, perhaps at an $\exp(-\delta t)$ rate. The short sequence $Z_t, Z_{t+1}, ..., Z_{t+k-1}$ has a *distribution shift* if $Z_{t+j} = W_{t+j} + M$ for some constant M and $j = 0, 1, ..., k-1$. A *level shift* is an event that affects a time series at a particular time point whose effect becomes permanent. The following *R* code, modified from some code of Iturria, et al. (2019), is used to demonstrate two types of outliers shown in Figure 5.1. The bulk of the data are iid from a discrete uniform (1, ..., 100) distribution.

```
set.seed(100)
n <- 500
x <- sample(1:100,n,replace=TRUE)
```

```
x[70:90] <- sample(110:115,21,replace=TRUE) #distribution shift
x[25] <- 200 #abrupt transient anomaly, additve outlier
x[320] <- 170  #abrupt transient anomaly, additve outlier
plot(x,type="l",xlab="Time")
```

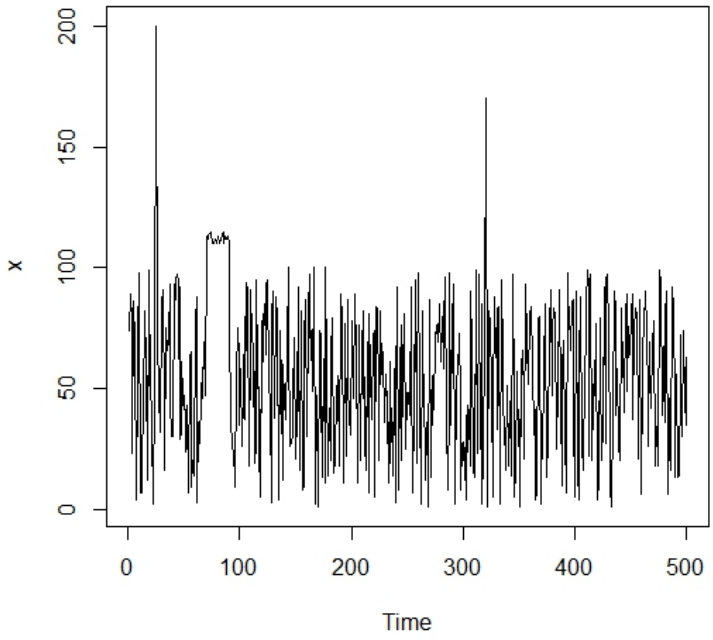


Figure 5.1. Artificial Time Series Has Outliers at t=25, 70-90, 320

The AR(p) model is useful for illustrating problems outliers cause. Repeat Equation (2.2) to write the AR(p) equations $Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t$ in matrix form $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ or

$$\begin{bmatrix} Y_{p+1} \\ Y_{p+2} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & Y_p & Y_{p-1} & \dots & Y_1 \\ 1 & Y_{p+1} & Y_p & \dots & Y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Y_{n-1} & Y_{n-2} & \dots & Y_{n-p} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix} + \begin{bmatrix} e_{p+1} \\ e_{p+2} \\ \vdots \\ e_n \end{bmatrix}$$

where $\boldsymbol{X}$ is of full rank with more rows than columns $p + 1$ and $\boldsymbol{\beta} = (\phi_0, \boldsymbol{\phi}^T)^T = (\phi_0, \phi_1, ..., \phi_p)^T$. Note that if Y_{p+1} is an outlier, then Y_{p+1} is an outlier in the kth row and kth column of $\boldsymbol{X}$ for

$k = 2, ..., p + 1$. Differencing can cause even more outliers in the data.

"Robust" multiple linear regression estimators can be applied to ARIMA$(p, d, 0)$ data or data from the dynamic linear model to create a "robust estimator." These estimators tend to work poorly for several reasons. First, the "high breakdown robust" multiple linear regression estimators that are practical to compute tend to be inconsistent with poor outlier resistance. See Hawkins and Olive (2002), Huber and Ronchetti (2009), and Olive (2017b, 2023b). M and GM estimators do not work well for multiple linear regression, and perform even worse for time series.

The Olive (2017b) `rmreg2` estimator will be used as the robust multiple linear regression estimator. The (ordinary) least squares estimator $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, $\hat{\boldsymbol{\phi}}_{0,OLS} = \overline{Y} - \hat{\boldsymbol{\phi}}_{OLS}^T\overline{\boldsymbol{x}}$, and $\hat{\boldsymbol{\phi}}_{OLS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$ where $\underline{\boldsymbol{\beta}} = (\boldsymbol{\phi}_O, \underline{\boldsymbol{\phi}}_{OLS}^T)^T$

Here $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$ are the usual estimated covariance matrices used when $\boldsymbol{w}_i = (\boldsymbol{x}_i, Y_i)^T$ are iid from some population. The `rmreg2` estimator plugs in robust covariance estimators in place of the classical estimators. More details follow. Let

$$\boldsymbol{w} = \begin{pmatrix} \boldsymbol{x} \\ Y \end{pmatrix}, \quad E(\boldsymbol{w}) = \boldsymbol{\mu}_{\boldsymbol{w}} = \begin{pmatrix} E(\boldsymbol{x}) \\ E(Y) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{x}} \\ \mu_Y \end{pmatrix}, \quad \text{and} \quad \text{Cov}(\boldsymbol{w}) = \boldsymbol{\Sigma}_{\boldsymbol{w}} =$$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{x}} & \boldsymbol{\Sigma}_{\boldsymbol{x},Y} \\ \boldsymbol{\Sigma}_{Y,\boldsymbol{x}} & \boldsymbol{\Sigma}_{Y,Y} \end{pmatrix}.$$

Let $(T, \boldsymbol{C}) = (\tilde{\boldsymbol{\mu}}_{\boldsymbol{w}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}})$ be a robust estimator of multivariate location and dispersion. Then the robust plug in estimator $\tilde{\phi}_0 = \tilde{\mu}_Y - \tilde{\boldsymbol{\phi}}^T\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}$ and $\tilde{\boldsymbol{\phi}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$. The robust estimator $(T, \boldsymbol{C})$ used will be the RMVN estimator of Olive (2017b), Olive and Hawkins (2010), and Zhang, Olive, and Ye (2012) that has been used to make robust estimators of multiple linear regression and multivariate linear regression. See Olive (2017b). The robust estimator has not yet been shown to be consistent for AR(p) data, but the robust estimator can be used as an outlier diagnostic.

Example 5.1. Here we examine outliers for the AR(p) model and use the Cryer and Chan (2008) *R* package `TSA` data set `deere1` which gives 82 consecutive values for the amount of de-

viation from a specified target value in an industrial machining process at Deere & Co. If there is an outlier at Y_k where k is not too close to 1 or n, then fitted values will use the outlier for $t = k + 1, ..., k + p$. So the outlier appears $p + 1$ times in the equations for the AR(p) model.

An AR(2) model will be used for the Deere time series, and the plot of the time series in Figure 5.2 shows that there is one large outlier, corresponding to case 27. Figure 5.3 shows the response and residual plots for the AR(2) model. Only one outlier, instead of two, appears in the fitted values since $\hat{\phi}_1 = 0.027$ is quite small. The plots for the robust fit are similar and are not shown.

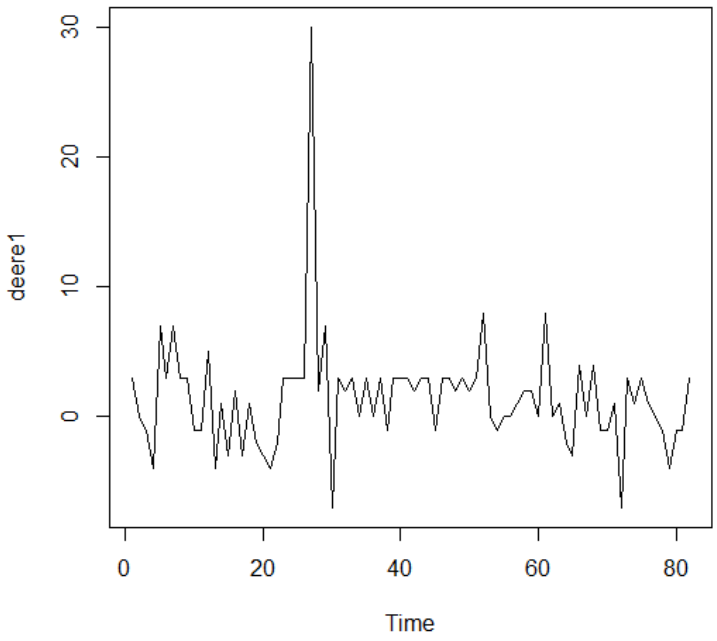


Figure 5.2. The Deere Time Series has One Outlier

The outlier Y_{27} is changed from 30 to a more reasonable value 8 to create "cleaned data." The clean fit $\hat{\boldsymbol{Z}} = \boldsymbol{Y} - \boldsymbol{r}$ where $\boldsymbol{r}$ are the residuals corresponding to an AR(2) model using the cleaned data. The clean fit fits all of the data, including the outlier, fairly well. Figure 5.3 shows the robust fit, using rmreg2, versus the clean fit. The identity line is tilted slightly away from the bulk of the data. Figure 5.4 shows the robust fit, using the AR(2) model with potential outliers replaced by

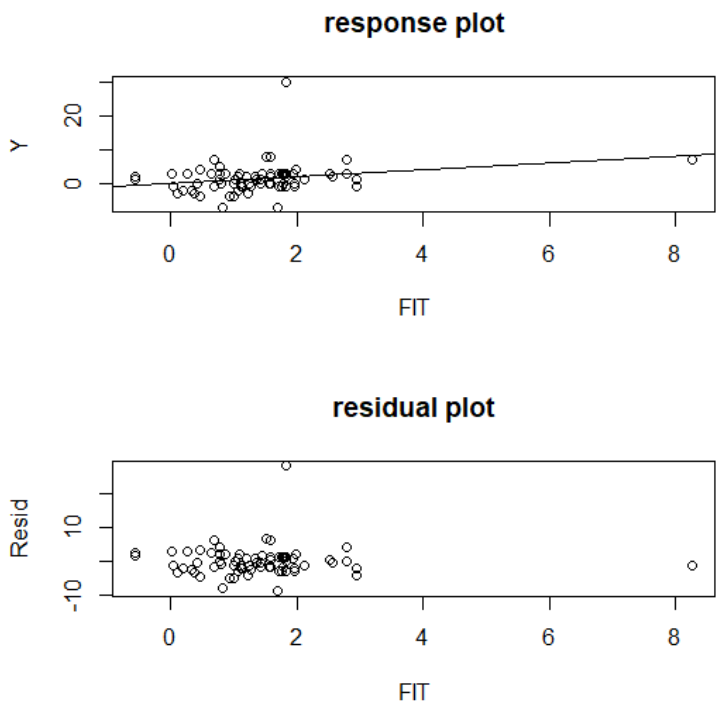


Figure 5.3. Response and Residual Plots for the AR(2) Model

missing values, versus clean fit. The missing values were replaced by median(Y) to get the fitted values. Now the identity line is not tilted away from the bulk of the data.

Next we added 2 more outliers to the data set: in the original data, cases Y_7 and Y_{76} were changed to 25 and 26. Figure 5.6 shows the robust fit, using rmreg2, versus the clean fit. The identity line fits the bulk of the data, but several large fitted values occur because of the outliers. Figure 5.7 shows the robust fit, using the AR(2) model with potential outliers replaced by missing values, versus clean fit. The missing values were replaced by median(Y) to get the fitted values. Now the identity line is not tilted away from the bulk of the data.

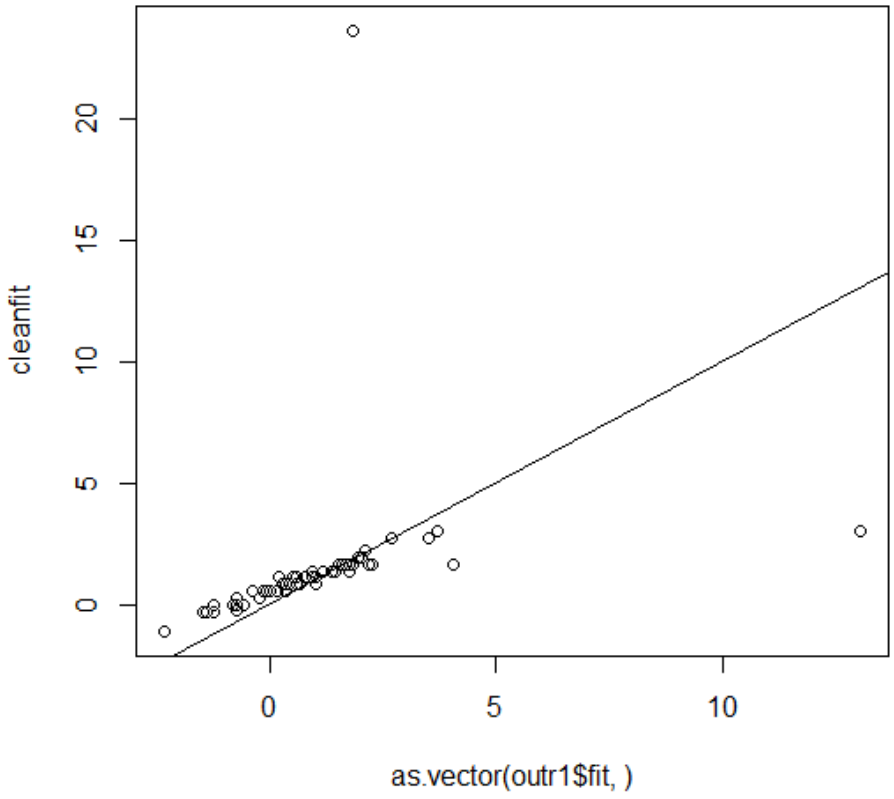


Figure 5.4. Robust Fitted Values from the Data with 1 Outlier Versus Fitted Values from the Cleaned Data

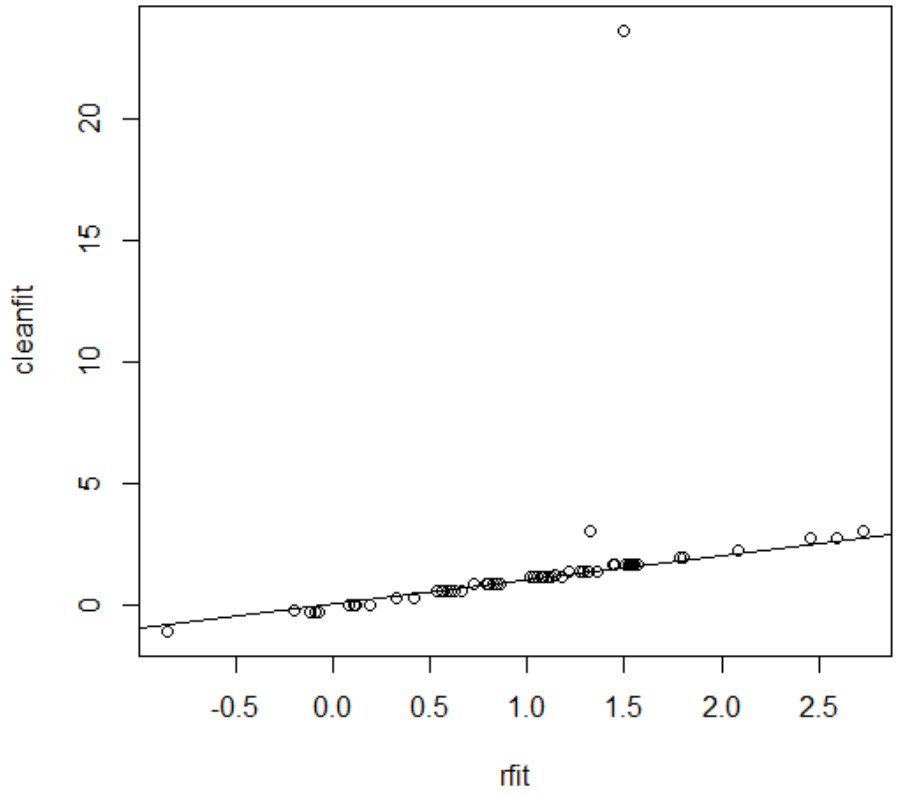


Figure 5.5. Fitted Values Using NA from the Data with 1 Outlier Versus Fitted Values from the Cleaned Data

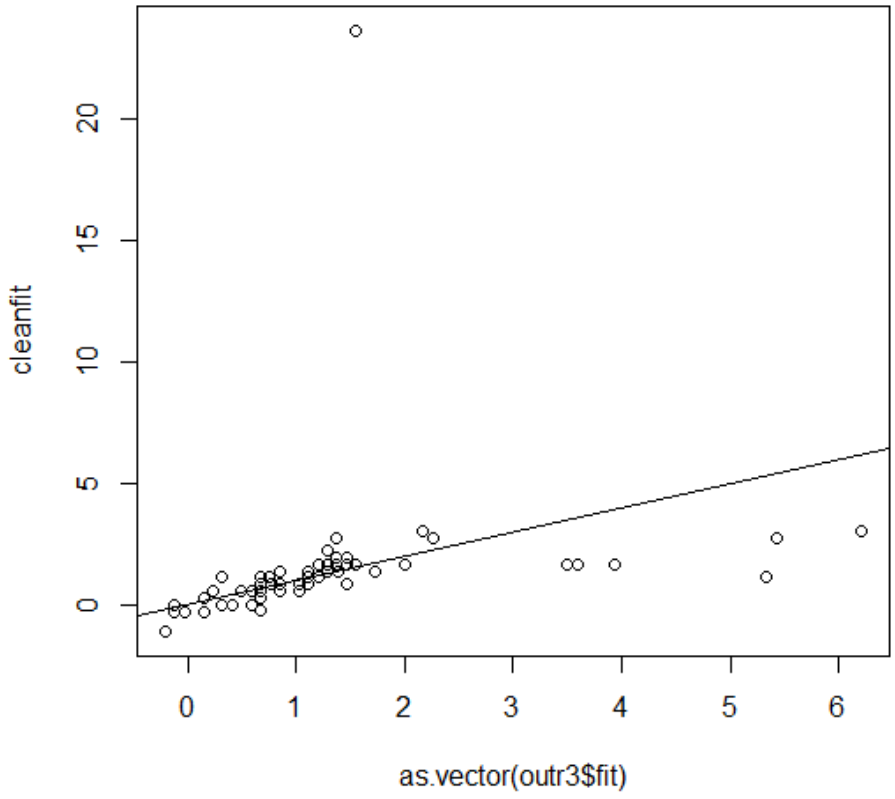


Figure 5.6. Robust Fitted Values from the Data with 3 Outliers Versus Fitted Values from the Cleaned Data

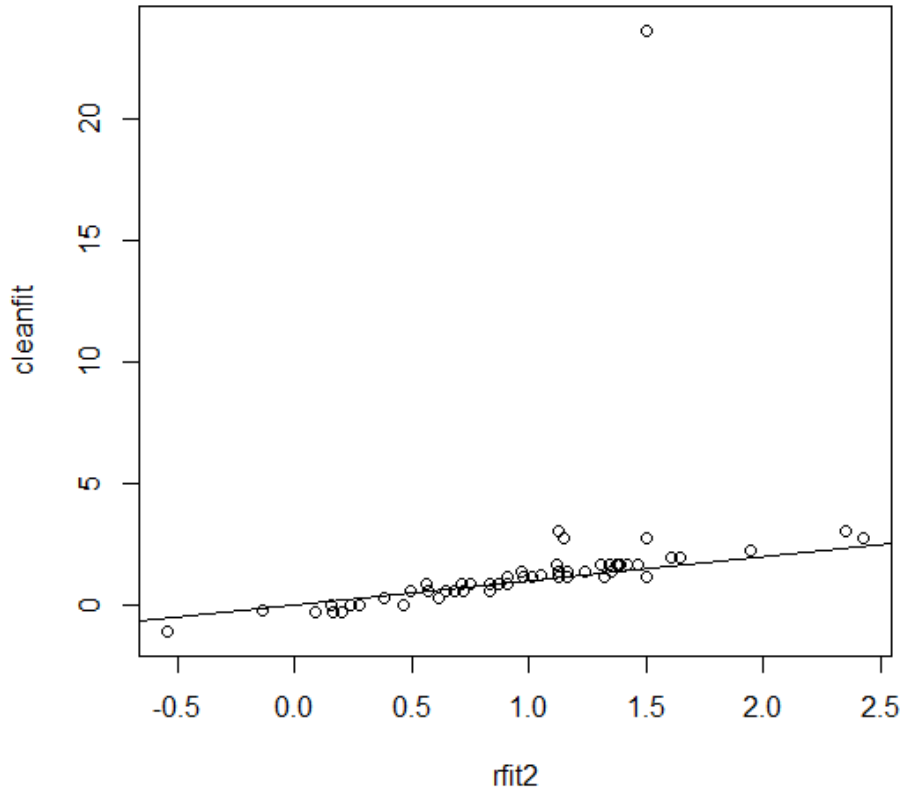


Figure 5.7. Fitted Values Using NA from the Data with 3 Outliers Versus Fitted Values from the Cleaned Data

Next cases Y_7 and Y_{76} were changed to 250 and 260. Figure 5.8 shows the robust fit, using rmreg2, versus the clean fit. The identity line fits the bulk of the data, but several large fitted values occur because of the outliers. Figure 5.9 shows the robust fit, using the AR(2) model with potential outliers replaced by missing values, versus clean fit. The missing values were replaced by median(Y) to get the fitted values. Now the identity line is not tilted away from the bulk of the data.

For this example, the fitted values from the AR(2) model, produced by replacing times series values outside of $[MED(Y) - 6MAD(Y), MED(Y) + 6MAD(Y)]$ by missing values NA, was effective. The cleaned fitted values, produced by fitting the AR(2) model to the data where the outlier was replaced by a more reasonable value, were made such that the single outlier was also

fit well. Ignoring the outlier, the fitted values using NA estimated the cleaned fitted values much better than the fitted values obtained using the rmreg2 estimator. Using missing values also gives robust estimators of other quantities produced by the software, such as the ACF, PACF, and some tests.

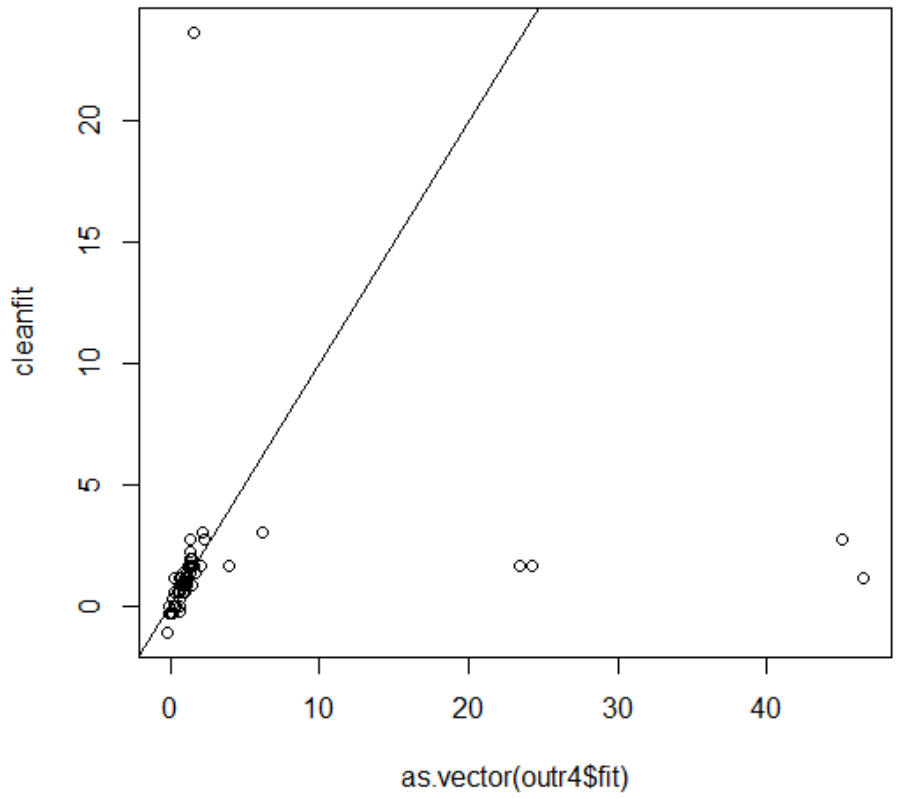


Figure 5.8. Robust Fitted Values from the Data with 3 Large Outliers Versus Fitted Values from the Cleaned Data

The *R* code below corresponds to the following.

a) Gives the plot of the time series. See Figure 5.2.

b) Gives the output table for the AR(2) model as well as the response and residual plots. See Figure 5.3.

c) The commands for this part fit a robust AR(2) model and gives the coefficient values and the response and residual plots.

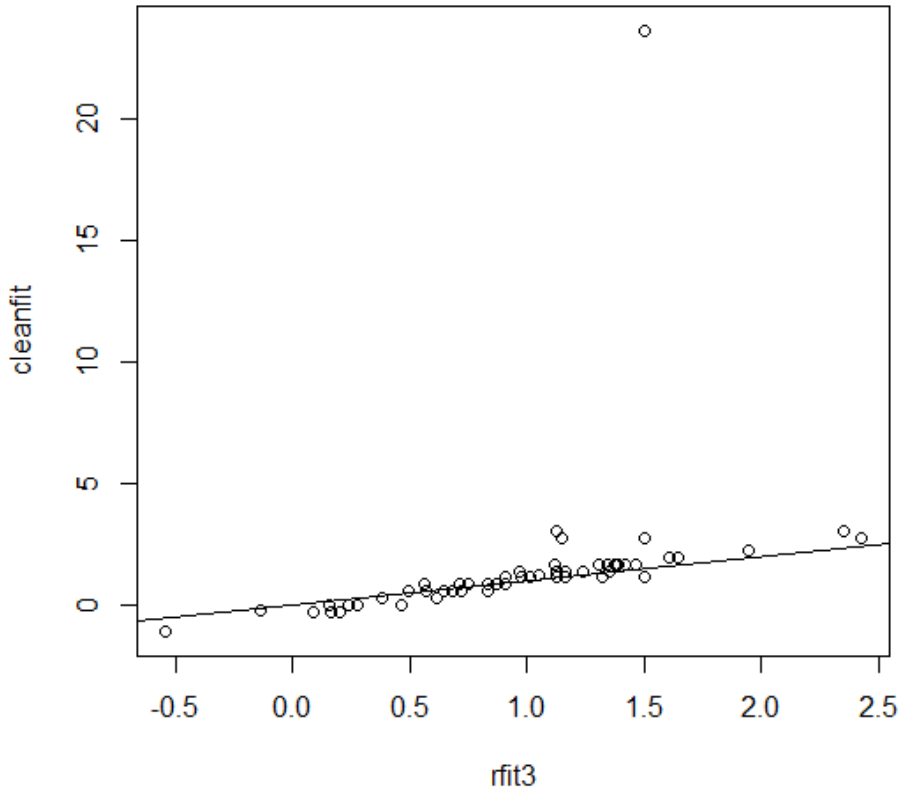


Figure 5.9. Fitted Values Using NA from the Data with 3 Large Outliers Versus Fitted Values from the Cleaned Data

d) The command for this part change the outlier from 30 to a more reasonable value 8 and refits the AR(2) model producing the output table for the AR(2) model, and the response and residual plots.

e) The commands for this part fit a robust AR(2) model on the cleaned data, giving the coefficient values for the AR(2) model, and the response and residual plots.

f) The commands for this part plot the fitted values from the robust AR(2) model fit to the data with the outlier versus the fitted values from the classical AR(2) model fit to the clean data. The fitted values are similar except for the outlier in Y_t and one of the outliers in $\hat{Y}_t$. See Figure 5.4.

g) The commands for this part change the values of two cases (7 and 76) to 25 and 26, and fits the robust estimator. Then the commands plot the fitted values versus the fitted values of the robust estimator to the cleaned data. The fitted values are tilted some.

h) Now the values of the two cases (7 and 76) are changed to 250 and 260. Then the commands plot the fitted values versus the fitted values of the robust estimator to the cleaned data. The fitted values for the bulk of the data are similar since big outliers are easier to detect. See Figure 5.8.

i) These commands change the potential outliers to NA. For the deere1 data set, case 27 is changed to NA. The output table and response and residual plots are given.

j) These commands take the data set from g) and change the potential outliers to NA. The three outliers got NA. The output table and response and residual plots are given.

k) These commands take the data set from h) and change the potential outliers to NA. The three outliers got NA. The output table and response and residual plots are given.

```
source("http://parker.ad.siu.edu/Olive/tspack.txt")
#library("TSA")
#data(deere1)
#plot(deere1)
```

```
deere1 <- c(3,0,-1,-4,7,3,7,3,3,-1,-1,5,-4,1,-3,2,-3,1,-2,-3,
-4,-2,3,3,3,3,30,2,7,-7,3,2,3,0,3,0,3,-1,3,3,3,2,3,3,-1,3,3,
2,3,2,3,8,0,-1,0,0,1,2,2,0,8,0,1,-2,-3,4,0,4,-1,-1,1,-7,3,1,
3,1,0,-1,-4,-1,-1,3)

#a)
plot(deere1,type="l",xlab="Time")  #Figure 5.2

#b)
out2 <- arima(deere1,c(2,0,0))
resplots(deere1,out2)  #Figure 5.3

#c)
outr1 <- robar(deere1,2)
outr1$phihat
#right click Stop on the plot twice
#For each Y outlier in an AR(p) model,
#there will be p outliers in the X matrix
#which could cause up to p outliers in the fitted values.
#So p+1 outliers in the XY matrix used to compute the robust estimator.

#d)
deerem2 <- deere1
deerem2[27] <- 8
out2m <- arima(deerem2,c(2,0,0))
resplots(deerem2,out2m)
```

```
#e)
outr2 <- robar(deerem2,2)
outr2$phihat    #robust fit to cleaned data

#f)
cleanfit <- as.vector(deere1) - as.vector(out2m$resid)
plot(as.vector(outr1$fit,),cleanfit)  #figure 5.4
abline(0,1)  #cleanfit fits all cases, including the outlier, fairly well

#plot(as.vector(outr2$fit,),cleanfit)
#abline(0,1)

#g)
deerem3 <- deere1
deerem3[c(7,76)] <- c(25,26)
outr3 <- robar(deerem3,2)
plot(as.vector(outr3$fit),outr2$fit) #right click Stop 2 times, hit Enter
abline(0,1)
plot(as.vector(outr3$fit),cleanfit)
abline(0,1)  #Figure 6
#identify(as.vector(outr3$fit),outr2$fit)
#NAs mean the identified points are off by 2
#74, 25, 5 instead of 76,27,7

#h)
deerem4 <- deere1
deerem4[c(7,76)] <- c(250,260)
```

```
outr4 <- robar(deerem4,2)
plot(as.vector(outr4$fit),outr2$fit) #right click Stop 2 times, hit Enter
abline(0,1)   #FFplot.eps
#identify(as.vector(outr4$fit),outr2$fit
#works better with massive outliers
#outliers are easy to spot with response plot
#since there Y values are outlying
plot(as.vector(outr4$fit),cleanfit)
abline(0,1) #Figure 5.8

##plot(as.vector(outr4$fit),cleanfit) #right click Stop 2 times, hit Enter
##abline(0,1)

#i)
YNA <- tsNA(deere1)$W
out5 <- arima(YNA,c(2,0,0))
resplots(YNA,out5)

rfit <- YNA-out5$res
rfit[is.na(rfit)]<-median(deere1)
rfit<-as.vector(rfit)
#plot(rfit,deere1)
plot(rfit,cleanfit)
abline(0,1) #Figure 5.5

#j)
W2 <- tsNA(deerem3)$W
out6 <- arima(W2,c(2,0,0))
```

```
resplots(W2,out6)

YNA2 <- W2
rfit2 <- YNA2-out6$res
rfit2[is.na(rfit2)]<-median(deere1)
rfit2<-as.vector(rfit2)
plot(rfit2,cleanfit)
abline(0,1) #Figure 5.7

#k)
W3 <- tsNA(deerem4)$W
out7 <- arima(W3,c(2,0,0))
resplots(W3,out7)

YNA3 <- W3
rfit3 <- YNA3-out7$res
rfit3[is.na(rfit3)]<-median(deere1)
rfit3<-as.vector(rfit3)
plot(rfit3,cleanfit)
abline(0,1) #Figure 5.9
```

CHAPTER 6

REAL DATA EXAMPLES

The main factors effecting the gasoline prices are global crude oil cost, refining costs, distribution and marketing costs, federal and state taxes, which are generally reflected in the wholesale costs that gasoline retailers pay to distributors. In addition to these factors, retail stations have to consider the local factors that can impact retail fuel prices such as store types (branded or unbranded), store location and their local competition, fuel delivery method, length of existing contracts with suppliers, volumes purchased, and specific store considerations and this becomes a main role in our life. Therefore for the implementation, two data sets will be used obtained;

- monthly Brent crude oil spot price:
  $https://github.com/rishabh89007/Time\_Series\_Datasets$
- weekly petrol prices:
  $https://data.world/makeovermonday/2020w17 - weekly - road - fuel - prices.$

Consider the monthly Brent crude oil spot price (dollars per barrel) with 396 observations collected over the period of 01/1990 - 12/2022. Brent is the leading global price benchmark for Atlantic basin crude oils. It is used to set the price of two-thirds of the world's internationally traded crude oil supplies. As well as weekly petrol prices (pence per litre) with 881 observations collected over the period of 06/09/2003 - 04/20/2020. Gasoline prices are determined largely by the laws of supply and demand. The two data sets were downloaded from each platforms as mentioned above and saved as "BSP.csv" and "fuel.csv" respectively.

For Brent crude oil spot price dataset;

```
> source("http://parker.ad.siu.edu/Olive/tspack.txt")
> library(forecast)
> d=read.csv("BSP.csv")
```

```
> price1ts=ts(d$price,frequency=12, start=c(1990,1))
> plot.ts(pricets,main="Brent Spot Oil Prices",xlab="Time",ylab="Price")
> d1=diff(log(d$price))
> pricets=ts(d1,frequency=12, start=c(1990,1))
> plot.ts(pricets, main="The Difference Series of the Logs of the Oil Price"
, xlab="Year", ylab="Price",type="o")
> oil<-d1
> acf(oil)
> pacf(oil)
> auto.arima(d1)
Series: d1
ARIMA(1,0,2) with zero mean
Coefficients:
         ar1      ma1      ma2
      0.8630  -0.5674  -0.3372
s.e.  0.0699   0.0787   0.0478
sigma^2 = 0.008995:  log likelihood = 371.34
AIC=-734.69   AICc=-734.58   BIC=-718.77
```

For petrol prices data set;

```
> source("http://parker.ad.siu.edu/Olive/tspack.txt")
> library(forecast)
> d=read.csv("fuel.csv")
> price1ts=ts(d$price,frequency=4, start=c(2003,6,9))
> plot.ts(pricets,main="Petrol Prices",xlab="Time",ylab="Price")
> d1=diff(d$price)
> pricets=ts(d1,frequency=4, start=c(2003,6,9))
> plot(d1,main="Differenced Petrol Prices",xlab="Time",ylab="Price",type="o")
```

For the analysing, the same methodology was used for both the data sets. Firstly consider the data set with monthly Brent crude oil spot price. The time series plot for this data is showing that the raw data is not stationary. To make them stationary, first difference is taken as in Figure 6.2.

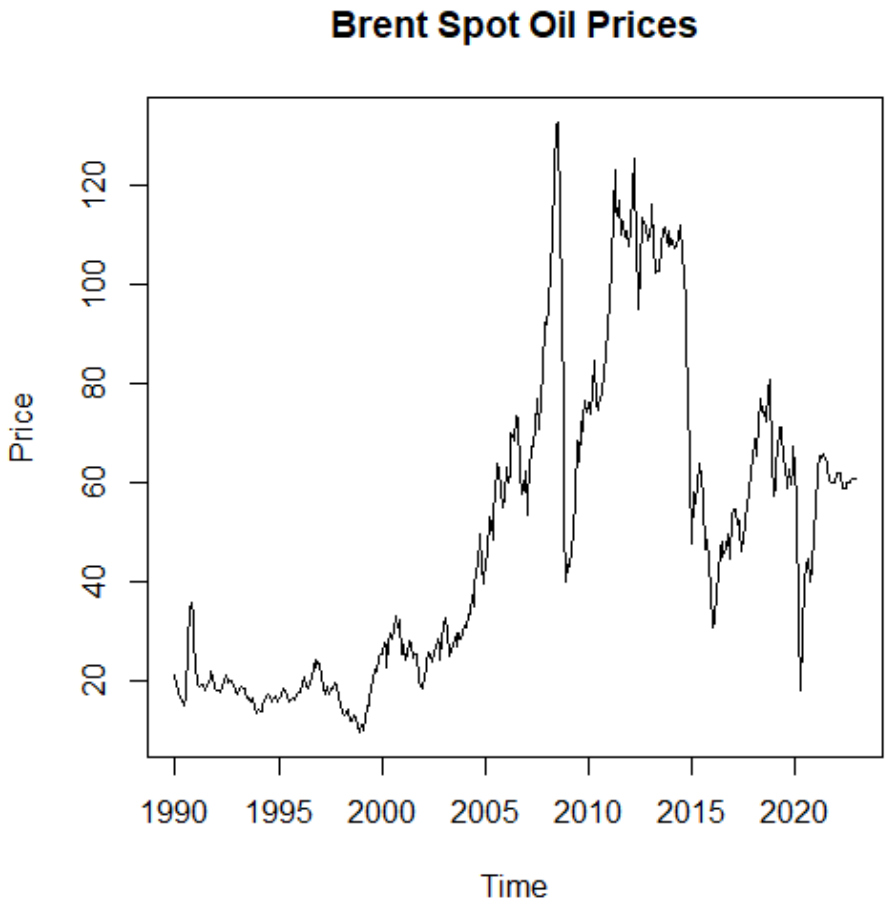


Figure 6.1. Time series plot of Brent Crude Oil Spot Price

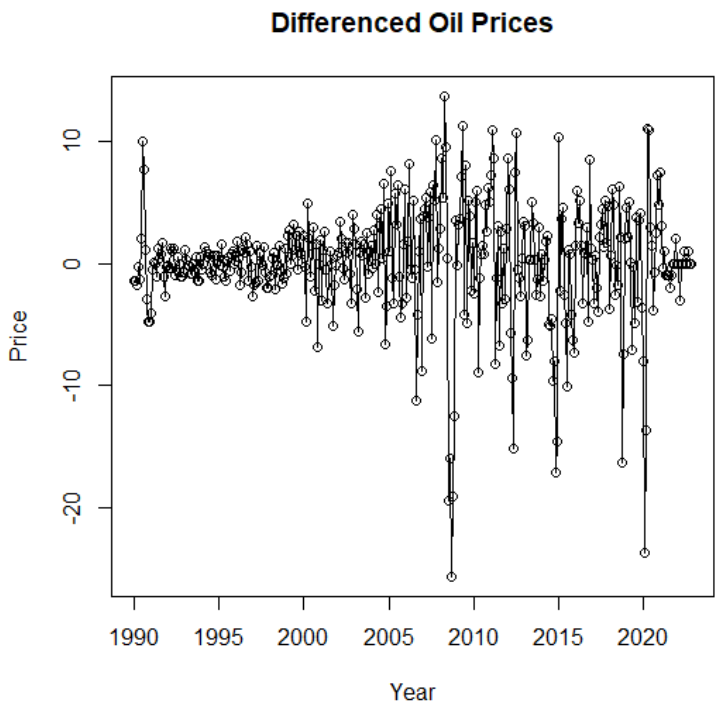


Figure 6.2. First Differenced Oil Price

Thought the first difference was taken, still the series not stationary as the variance increase with the time. Then the difference of logs of the price were taken and plotted the series as in Figure 6.3. Then the ACF and PACF plots are obtained for this new series and it seems AR(2) and MA(2) models would be a good fitting. But the result using `auto.arima`, generated the optimal parameter values in the model as ARMA(1,2) with AIC= -734.69.

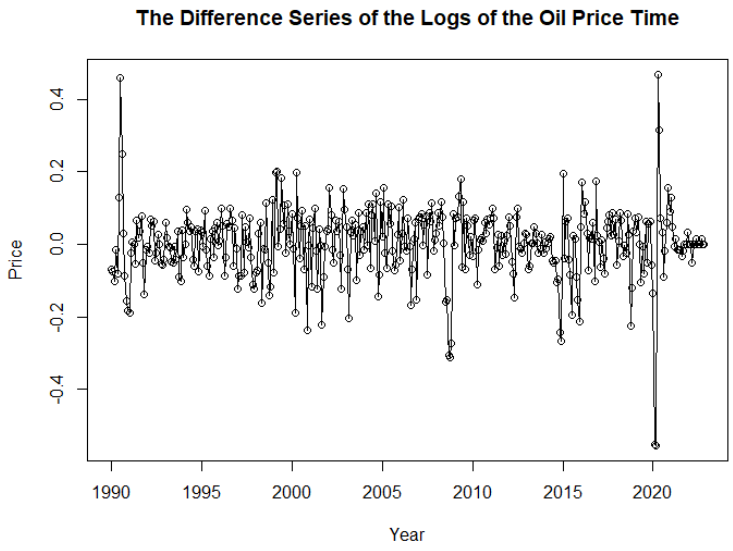


Figure 6.3. Difference Series of Logs of Oil Price

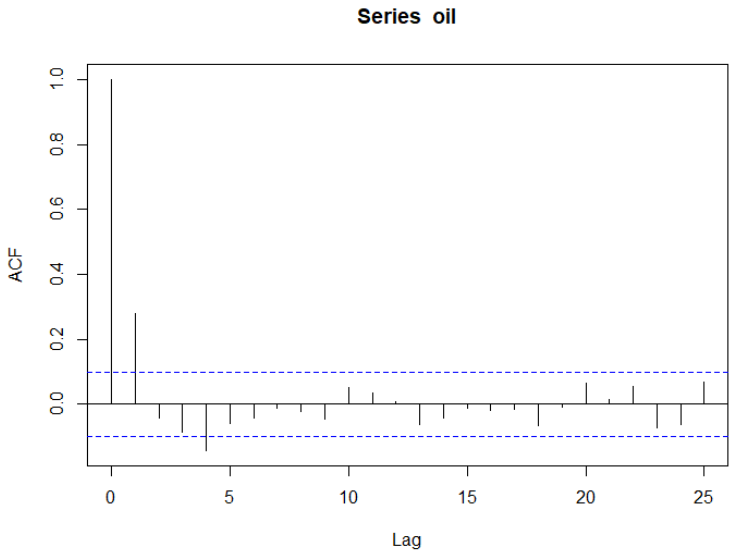


Figure 6.4. ACF plot for Difference Series of Logs of Oil Price

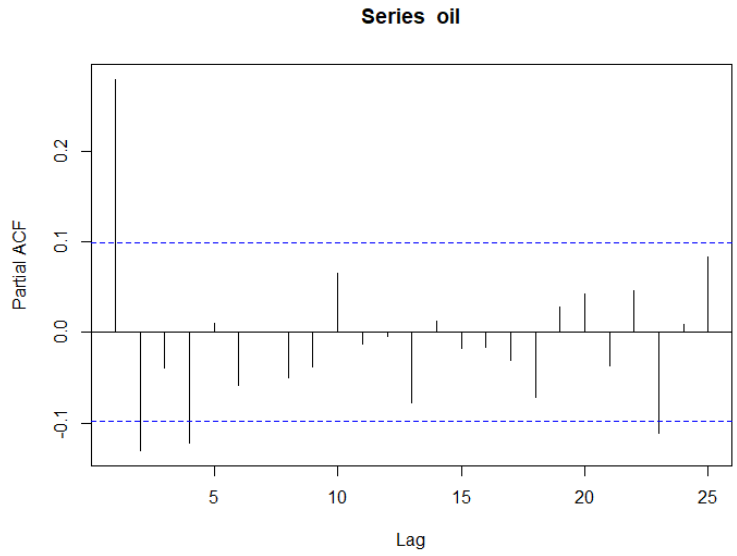


Figure 6.5. PACF plot for Difference Series of Logs of Oil Price

```
> auto.arima(d1)
Series: d1
ARIMA(1,0,2) with zero mean
Coefficients:
         ar1      ma1      ma2
      0.8630  -0.5674  -0.3372
s.e.  0.0699   0.0787   0.0478
sigma^2 = 0.008995:  log likelihood = 371.34
AIC=-734.69   AICc=-734.58   BIC=-718.77
```

According to the difference of logs of the price plot, it shows 6 visible outliers and then those potential outliers were replaced by missing values and the new series plot, ACF and PACF plots were obtained and followed the same analysing as above.

```
> identify(d1)
[1]   7 8 362 363 364 365
> y1<-d1
> y1[c(7,8,362,363,364,365)]<-NA
```

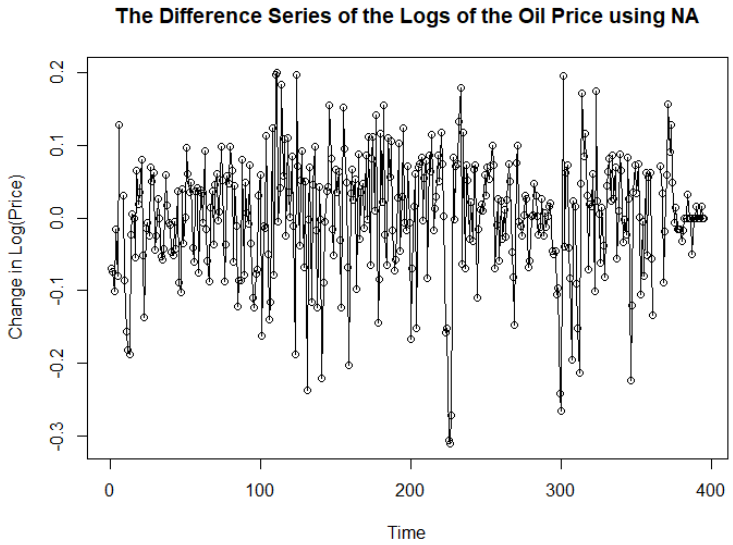


Figure 6.6. Difference Series of Logs of Oil Price using NA

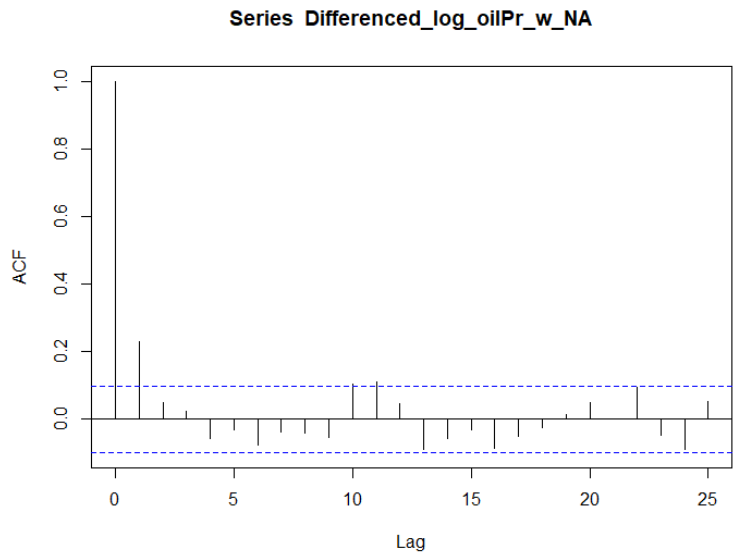


Figure 6.7. ACF plot for Difference Series of Logs of Oil Price using NA

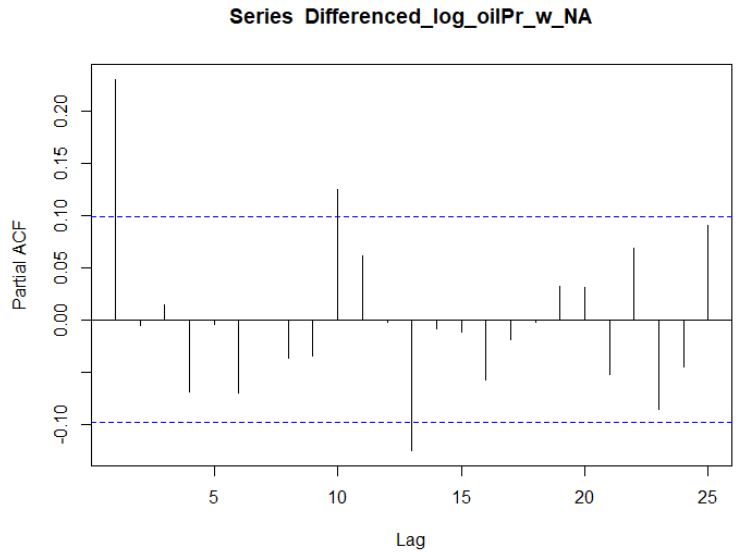


Figure 6.8. PACF plot for Difference Series of Logs of Oil Price using NA

```
> auto.arima(y2)
Series: y2
ARIMA(1,0,0) with zero mean
Coefficients:
         ar1
      0.2321
s.e.  0.0495
sigma^2 = 0.00668:  log likelihood = 422.62
AIC=-841.24   AICc=-841.21   BIC=-833.28
```

By the result using `auto.arima`, we obtained the model as AR(1) with AIC= -841.24 which is better than the ARMA(1,2) model due to the AIC, AICc and BIC criteria. This analysing implies that there is an effect from outliers to the model selection and making them missing values, we have the opportunity to get a consistent model for the data.

Then we will apply the algorithm `armamsel2` for doing the model selection and then will compare the results obtained from above.

```
> armamsel2(y1)
$rhat
[1] 1
$pI
[1] 1
$qI
[1] 1
```

```
> arima(y1,c(1,0,1))

Call:
arima(x = y2, order = c(1, 0, 1))

Coefficients:
         ar1     ma1  intercept
      0.2052  0.0282     0.0017
s.e.  0.2298  0.2359     0.0053

sigma^2 estimated as 0.006661:  log likelihood = 422.68,  aic = -837.36
```

This gives that the model as ARMA(1,1) since p_I=q_I=1 for this data set while the classical method with NA gives AR(1) as the model. Since q=q_I, this selected model from the algorithm is consistent. Since $\hat{r}$=1, only 8 models were fitted.

Similarly, the weekly Petrol price data showing the non-stationary behavior and then the first difference was taken as in Figure 6.8.

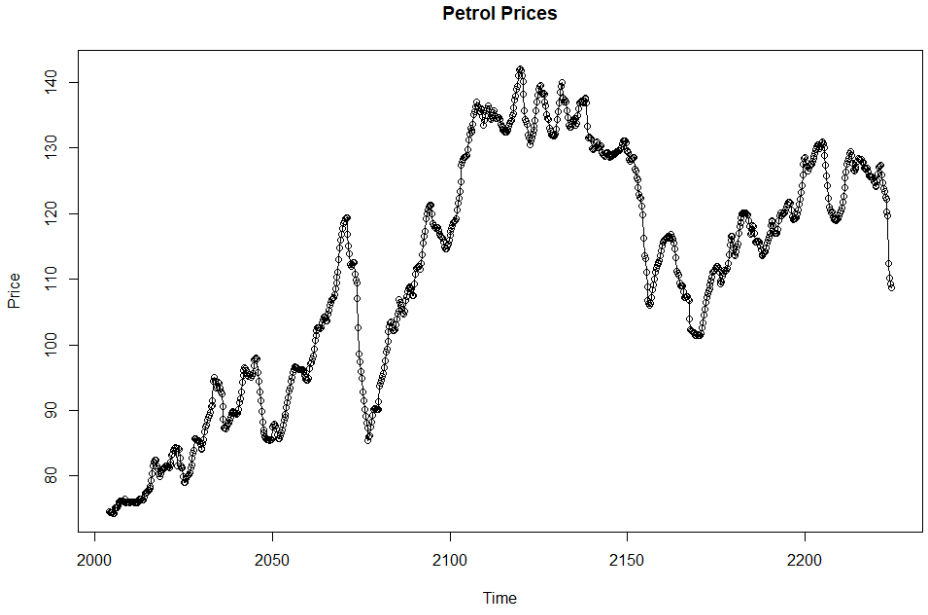


Figure 6.9. Time series plot of Weekly Petrol Price

Then the ACF and PACF plots are obtained for this differenced data and it seems a AR(1)

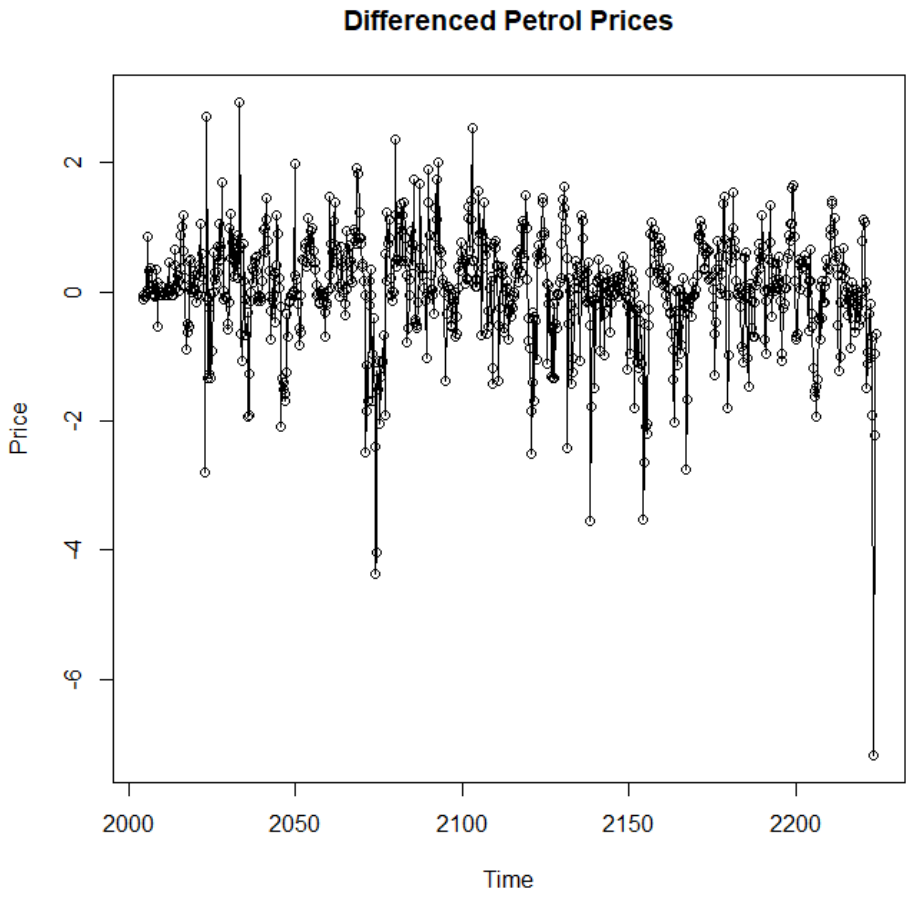


Figure 6.10. First Differenced Petrol Price

model would be a good fitting. But the result using `auto.arima`, generated the optimal parameter values in the model as ARMA(1,1) with AIC=1765.98.

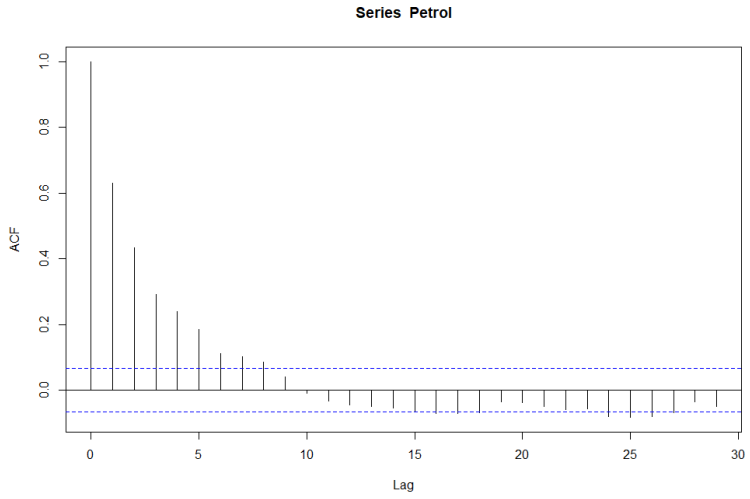


Figure 6.11. ACF plot for first differenced data

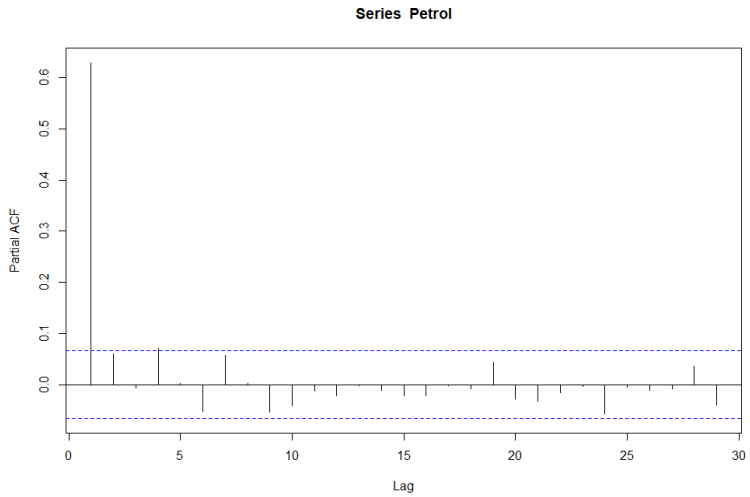


Figure 6.12. PACF plot for first differenced data

```
> auto.arima(d1f)
Series: d1
ARIMA(1,0,1) with zero mean
Coefficients:
         ar1      ma1
      0.6892  -0.0980
s.e.  0.0390   0.0541
sigma^2 = 0.4334:  log likelihood = -879.99
AIC=1765.98   AICc=1766.01   BIC=1780.32
```

According to the first differenced data plot, it shows 3 visible outliers and then those potential outliers were replaced by missing values and the new differenced plot, ACF and PACF plots were obtained and followed the same analysing as above.

```
> identify(d1f)
[1]   280  281  877
> y1f<-d1f
> y1f[c(280,281,877)]<-NA
```

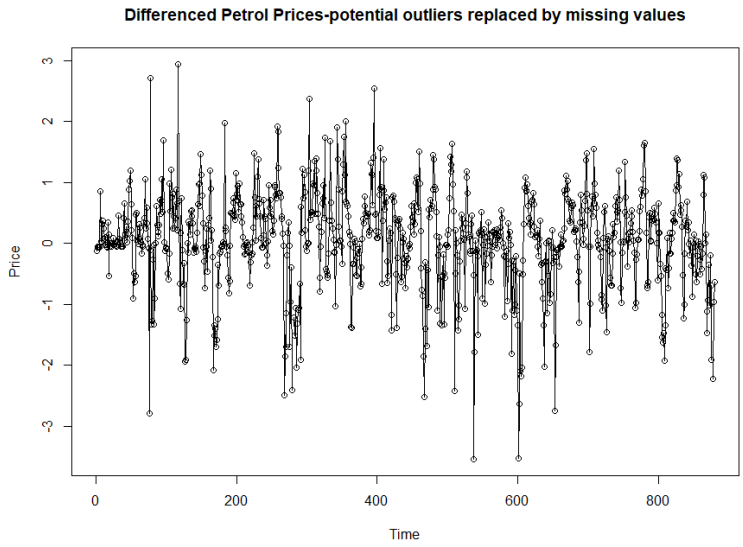


Figure 6.13. First Differenced Oil Price using NA

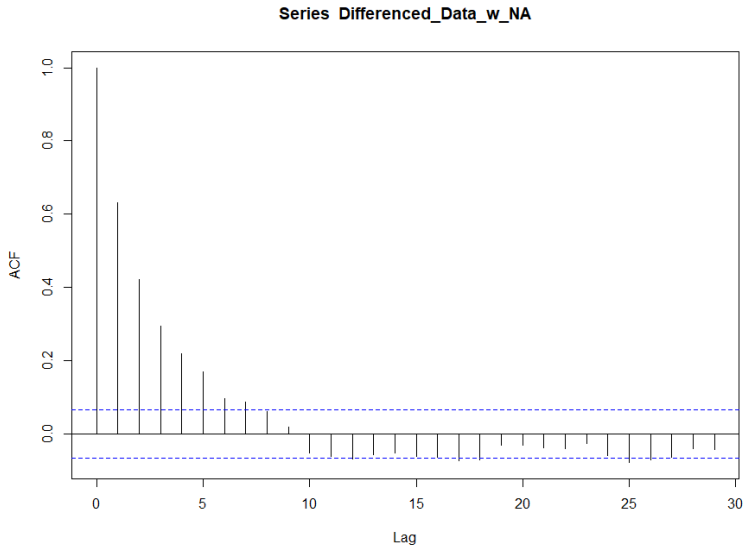


Figure 6.14. ACF plot for first differenced data using NA

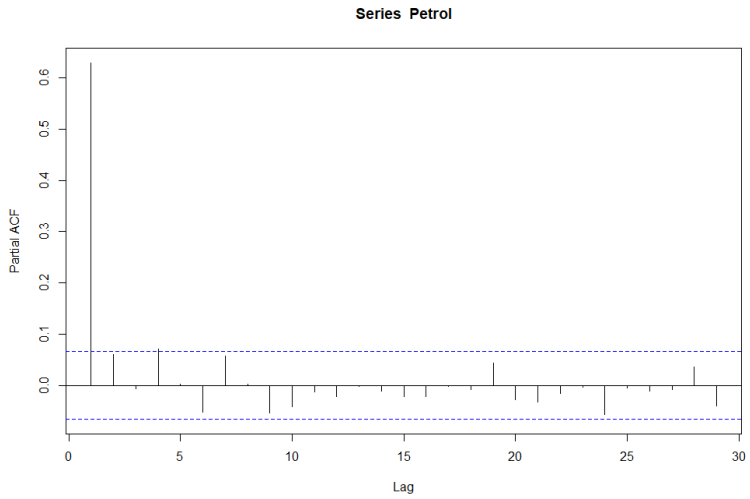


Figure 6.15. PACF plot for first differenced data using NA

```
> auto.arima(y1f)
Series: y1
ARIMA(1,0,0) with zero mean
Coefficients:
         ar1
      0.6454
s.e.  0.0258
sigma^2 = 0.3678:  log likelihood = -805.96
AIC=1615.91   AICc=1615.93   BIC=1625.47
```

By the result using `auto.arima`, we obtained the model as AR(1) with AIC=1615.91 which is better than the ARMA(1,1) model due to the AIC, AICc and BIC criteria. This analysing implies that there is an effect from outliers to the model selection and making them missing values, we have the opportunity to get an optimal model for the data.

Then we will apply the algorithm `armamsel2` for doing the model selection and then will compare the results obtained from above.

```
> armamsel2(y1f)
$rhat
[1] 1
$pI
[1] 1
$qI
[1] 1

> arima(y1,c(1,0,1))
Call:
arima(x = y1, order = c(1, 0, 1))
Coefficients:
         ar1      ma1  intercept
      0.6812  -0.0642     0.0494
s.e.  0.0391   0.0541     0.0598
sigma^2 estimated as 0.3665:  log likelihood = -804.88,  aic = 1617.76
```

This gives that the model as ARMA(1,1) since p_I=q_I=1 for this data set while the classical method with NA gives AR(1) as the model. Since p=p_I, this selected model from the algorithm is consistent. Since $\hat{r}$=1, only 8 models were fitted instead of 36 models.

CHAPTER 7

VISUALIZING SOME BOOTSTRAP CONFIDENCE REGIONS

A confidence interval is the likely range for the true score of your entire population and a confidence region in a single dimension is also called a confidence interval. The DD plot is a plot of the classical Mahalanobis distances MD_i versus robust Mahalanobis distances RD_i which uses to visualize prediction regions. Several bootstrap confidence intervals and regions are obtained by applying prediction intervals and regions to the bootstrap sample.

Notation: $P(A_n)$ is "eventually bounded below" by $1-\delta$ if $P(A_n)$ gets arbitrarily close to or higher than $1-\delta$ as $n \to \infty$. Hence $P(A_n) > 1-\delta-\epsilon$ for any $\epsilon > 0$ if n is large enough. If $P(A_n) \to 1-\delta$ as $n \to \infty$, then $P(A_n)$ is eventually bounded below by $1-\delta$. The actual coverage is $1-\gamma_n = P(Y_f \in [L_n, U_n])$, the nominal coverage is $1-\delta$ where $0 < \delta < 1$. The 90% and 95% large sample prediction intervals and prediction regions are common.

7.1 PREDICTION INTERVALS AND REGIONS

Consider predicting a future test value Y_f given training data $Y_1, ..., Y_n$. A large sample $100(1-\delta)\%$ *prediction interval* (PI) for Y_f has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n)$ is eventually bounded below by $1-\delta$ as the sample size $n \to \infty$. A large sample $100(1-\delta)\%$ PI is *asymptotically optimal* if it has the shortest asymptotic length: the length of $[\hat{L}_n, \hat{U}_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the *population shorth*: the shortest interval covering at least $100(1-\delta)\%$ of the mass.

Let the data $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ have joint pdf or pmf $f(\boldsymbol{y}|\theta)$ with parameter space Θ and support $\mathcal{Y}$. Let $L_n(\boldsymbol{Y})$ and $U_n(\boldsymbol{Y})$ be statistics such that $L_n(\boldsymbol{y}) \leq U_n(\boldsymbol{y})$, $\forall \boldsymbol{y} \in \mathcal{Y}$. Then $[L_n(\boldsymbol{y}), U_n(\boldsymbol{y})]$ is a 100 $(1-\delta)$ % *confidence interval* (CI) for θ if

$$P_\theta(L_n(\boldsymbol{Y}) \leq \theta \leq U_n(\boldsymbol{Y})) = 1-\delta$$

for all $\theta \in \Theta$. The interval $[L_n(\boldsymbol{y}), U_n(\boldsymbol{y})]$ is a large sample $100(1-\delta)$ % CI for θ if

$$P_\theta(L_n(\boldsymbol{Y}) \leq \theta \leq U_n(\boldsymbol{Y}))$$

is eventually bounded below by $1-\delta$ for all $\theta \in \Theta$ as the sample size $n \to \infty$.

A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$. A prediction region is *asymptotically optimal* if its volume converges in probability to the volume of the minimum volume covering region or the highest density region of the distribution of $\boldsymbol{x}_f$.

A *large sample* $100(1-\delta)\%$ *confidence region* for a vector of parameters $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

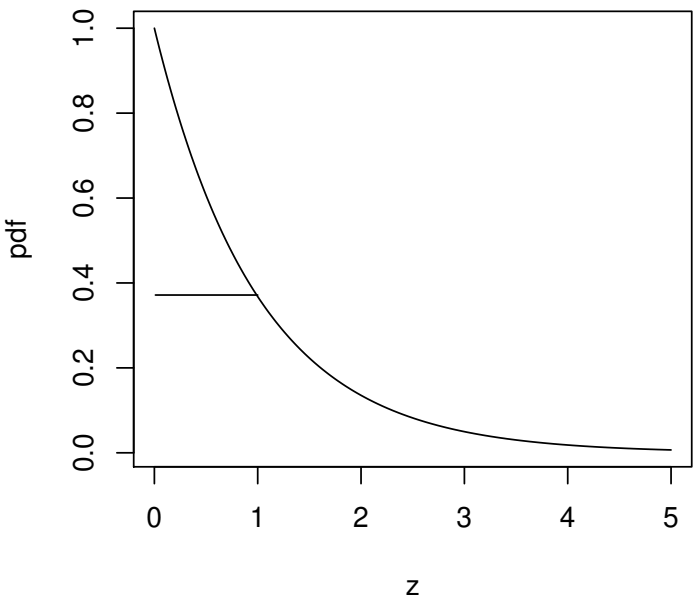


Figure 7.1. The 36.8% Highest Density Region is [0,1]

For a random variable Y, the $100(1-\delta)\%$ highest density region is a union of $k \geq 1$ disjoint intervals such that the mass within the intervals $\geq 1-\delta$ and the sum of the k interval lengths is as small as possible. Suppose that $f(z)$ is a unimodal pdf that has interval support, and that the pdf

$f(z)$ of Y decreases rapidly as z moves away from the mode. Let $[a, b]$ be the shortest interval such that $F_Y(b) - F_Y(a) = 1 - \delta$ where the cdf $F_Y(z) = P(Y \leq z)$. Then the interval $[a, b]$ is the $100(1-\delta)$ highest density region. To find the $100(1-\delta)\%$ highest density region of a pdf, move a horizontal line down from the top of the pdf. The line will intersect the pdf or the boundaries of the support of the pdf at $[a_1, b_1], ..., [a_k, b_k]$ for some $k \geq 1$. Stop moving the line when the areas under the pdf corresponding to the intervals is equal to $1 - \delta$. As an example, let $f(z) = e^{-z}$ for $z > 0$. See Figure 7.1 where the area under the pdf from 0 to 1 is 0.368. Hence [0,1] is the 36.8% highest density region. Often the highest density region is an interval $[a, b]$ where $f(a) = f(b)$, especially if the support where $f(z) > 0$ is $(-\infty, \infty)$.

The interpretation of a 100 $(1 - \delta)\%$ PI for a random variable Y_f is similar to that of a confidence interval (CI). Collect data, then form the PI, and repeat for a total of k times where the k trials are independent from the same population. If Y_{fi} is the ith random variable and PI_i is the ith PI, then the probability that $Y_{fi} \in PI_i$ for j of the PIs approximately follows a binomial$(k, \rho = 1-\delta)$ distribution. Hence if 100 95% PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times. If Y_f has a pdf, we often want $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1 - \delta$ as $n \to \infty$.

There are two big differences between CIs and PIs. First, the length of the CI goes to 0 as the sample size n goes to ∞ while the length of the PI converges to some nonzero number J, say. Secondly, many confidence intervals work well for large classes of distributions while many prediction intervals assume that the distribution of the data is known up to some unknown parameters. Usually the $N(\mu, \sigma^2)$ distribution is assumed, and the parametric PI may not perform well if the normality assumption is violated.

In the following theorem, if the open interval $(Y_{(k_1)}, Y_{(k_2)})$ was used, we would need to add the regularity condition that $Y_{\delta/2}$ and $Y_{1-\delta/2}$ are continuity points of $F_Y(y)$.

Definition 7.1. Let $Y_1, ..., Y_n, Y_f$ be iid. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of the training data. Let $k_1 = \lceil n\delta/2 \rceil$ and $k_2 = \lceil n(1 - \delta/2) \rceil$ where $0 < \delta < 1$. The large sample

$100(1-\delta)\%$ *percentile prediction interval* for Y_f is

$$[Y_{(k_1)}, Y_{(k_2)}]. \tag{7.1}$$

The bootstrap percentile confidence interval given by Equation (7.2) is obtained by applying the percentile prediction interval to the bootstrap sample $T_1, ..., T_B$.

Definition 7.2. The large sample $100(1-\delta)\%$ *bootstrap percentile confidence interval* for θ is an interval $[T^*_{(k_L)}, T^*_{(K_U)}]$ containing $\approx \lceil B(1-\delta) \rceil$ of the T^*_i. Let $k_1 = \lceil B\delta/2 \rceil$ and $k_2 = \lceil B(1-\delta/2) \rceil$. A common choice is

$$[T^*_{(k_1)}, T^*_{(k_2)}]. \tag{7.2}$$

Consider predicting a $p \times 1$ future test value $\boldsymbol{x}_f$, given past training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid. Much as confidence regions and intervals give a measure of precision for the point estimator $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$, prediction regions and intervals give a measure of precision of the point estimator $T = \hat{\boldsymbol{x}}_f$ of the future random vector $\boldsymbol{x}_f$.

For multivariate data, sample Mahalanobis distances play a role similar to that of residuals in multiple linear regression. Let the observed training data be collected in an $n \times p$ matrix $\boldsymbol{W}$. Let the $p \times 1$ column vector $T_n = T_n(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}_n = \boldsymbol{C}_n(\boldsymbol{W})$ be a dispersion estimator.

Definition 7.3. Let $x_{1j}, ..., x_{nj}$ be measurements on the jth random variable X_j corresponding to the jth column of the data matrix $\boldsymbol{W}$. The jth *sample mean* is $\overline{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj}$. The *sample covariance* S_{ij} estimates $\mathrm{Cov}(X_i, X_j) = \sigma_{ij} = E[(X_i - E(X_i))(X_j - E(X_j))]$, and

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j).$$

$S_{ii} = S_i^2$ is the *sample variance* that estimates the population variance $\sigma_{ii} = \sigma_i^2$.

Definition 7.4. Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ be the data where $\boldsymbol{x}_i$ is a $p \times 1$ vector. The sample mean or *sample*

mean vector

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = (\overline{x}_1, ..., \overline{x}_p)^T.$$

The sample covariance matrix

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

That is, the ij entry of $\boldsymbol{S}$ is the sample covariance S_{ij}. The *classical estimator of multivariate location and dispersion* is $(T_n, \boldsymbol{C}_n) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$.

Definition 7.5. The ith *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T_n(\boldsymbol{W}), \boldsymbol{C}_n(\boldsymbol{W})) = (\boldsymbol{x}_i - T_n(\boldsymbol{W}))^T \boldsymbol{C}_n^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T_n(\boldsymbol{W})) \tag{7.3}$$

for each point $\boldsymbol{x}_i$. Notice that D_i^2 is a random variable (scalar valued). Let $(T_n, \boldsymbol{C}_n) = (T_n(\boldsymbol{W}), \boldsymbol{C}_n(\boldsymbol{W}))$. Then

$$D_{\boldsymbol{x}}^2(T_n, \boldsymbol{C}_n) = (\boldsymbol{x} - T_n)^T \boldsymbol{C}_n^{-1}(\boldsymbol{x} - T_n).$$

Hence D_i^2 uses $\boldsymbol{x} = \boldsymbol{x}_i$.

Let the $p \times 1$ location vector be $\boldsymbol{\mu}$, often the population mean, and let the $p \times p$ dispersion matrix be $\boldsymbol{\Sigma}$, often the population covariance matrix. If $\boldsymbol{x}$ is a random vector, then the population squared Mahalanobis distance is

$$D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \tag{7.4}$$

and that the term $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$ is the $p-$dimensional analog to the z-score used to transform a univariate $N(\mu, \sigma^2)$ random variable into a $N(0, 1)$ random variable. Hence the sample Mahalanobis distance $D_i = \sqrt{D_i^2}$ is an analog of the absolute value $|Z_i|$ of the sample Z-score $Z_i = (X_i - \overline{X})/\hat{\sigma}$. Also notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

Next, we derive a prediction region for $\boldsymbol{x}_f$ if $(T_n, \boldsymbol{C}_n) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$, $\boldsymbol{\mu} = E(\boldsymbol{x})$, and $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{x})$ is nonsingular. Let $D = D(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Then $D_i \xrightarrow{D} D$ and $D_i^2 \xrightarrow{D} D^2$. Hence the sample percentiles of the D_i are consistent estimators of the population percentiles of D at continuity points of the cdf of D, and the sample percentiles of the D_i^2 are consistent estimators of the population percentiles of D^2 at continuity points of the cdf of D^2. Let $c = k_n = \lceil n(1-\delta) \rceil$. Then Olive (2013) showed that the hyperellipsoid

$$\mathcal{A}_n = \{\boldsymbol{x} : D_{\boldsymbol{x}}^2(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(c)}^2\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(c)}\} \tag{7.5}$$

is a large sample $100(1-\delta)\%$ prediction region under mild conditions, although regions with smaller volumes may exist.

To improve performance, we will use a correction factor $c = U_n$ where U_n decreases to k_n. U_n is defined under Equation (7.7). A problem with the prediction regions that cover $\approx 100(1-\delta)\%$ of the training data cases $\boldsymbol{x}_i$ (such as (7.5) for $c = k_n$), is that they have coverage lower than the nominal coverage of $1-\delta$ for moderate n. This result is not surprising since empirically statistical methods perform worse on test data than on training data. Empirically for many distributions, for $n = 20p$, the prediction region (7.5) applied to iid data using $c = k_n = \lceil n(1-\delta) \rceil$ tended to have undercoverage as high as $\min(0.05, \delta/2)$. The undercoverage decreases rapidly as n increases. (Referring to the next paragraph, taking $q_n \equiv 1-\delta$ does not take into account the unknown variability of $(\overline{\boldsymbol{x}}, \boldsymbol{S})$, which is another reason for undercoverage and the need for a correction factor.)

Let $q_n = \min(1-\delta+0.05, 1-\delta+p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1-\delta/2, 1-\delta+10\delta p/n), \quad \text{otherwise.} \tag{7.6}$$

If $1-\delta < 0.999$ and $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. Using

$$c = \lceil n q_n \rceil \tag{7.7}$$

in (7.8) decreased the undercoverage. Let $D_{(U_n)}$ be the $100q_n$th sample quantile of the D_i.

The nonparametric prediction region is due to Olive (2013). For the classical prediction region, see Chew (1966) and Johnson and Wichern (1988, pp. 134, 151). A future observation (random vector) $\boldsymbol{x}_f$ is in the region (7.8) if $D_{\boldsymbol{x}_f} \leq D^2_{U_n}$). If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ and $\boldsymbol{x}_f$ are iid, the nonparametric prediction region (7.8) is asymptotically optimal for a large class of elliptically contoured distributions since the volume of (8) converges in probability to the volume of the highest density region. (These distributions have a highest density region which is a hyperellipsoid determined by a population Mahalanobis distance.) Refer to the above paragraph for $D_{(U_n)}$. Let $P(D^2 \leq D^2_{1-\delta}) = 1 - \delta$ if $D^2_{1-\delta}$ is a continuity point of the cdf $F_{D^2}(y)$ and $D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \xrightarrow{D} D^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$.

Definition 7.6. Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from a distribution with mean $E(\boldsymbol{x}) = \boldsymbol{\mu}$ and nonsingular covariance matrix $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$. The large sample
$100(1 - \delta)\%$ *nonparametric prediction region* for a future value $\boldsymbol{x}_f$ is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(U_n)}\} \tag{7.8}$$

if $D^2_{1-\delta}$ is a continuity point of the cdf $F_{D^2}(y)$.

Highest density regions are usually hard to estimate for p not much larger than four, but many elliptically contoured distributions with a nonsingular population covariance matrix, including the multivariate normal distribution, have highest density regions that can be estimated by the nonparametric prediction region (7.8). For more about highest density regions, see Olive (2017b, pp. 148-155). If $\boldsymbol{x}_f$ has a pdf, we often want $P(\boldsymbol{x}_f \in \mathcal{A}_n) \rightarrow 1 - \delta$ as $n \rightarrow \infty$. A PI is a prediction region where $p = 1$.

Definition 7.7. Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Then the large sample $100(1-\delta)\%$ *classical prediction region* for multivariate normal data is

$$\{z : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi^2_{p,1-\delta}\}. \tag{7.9}$$

The nonparametric prediction region (7.8) is useful if $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from a distribution with a nonsingular covariance matrix, and the sample size n is large enough. The distribution could be continuous, discrete, or a mixture. The asymptotic coverage is $1 - \delta$ if D has a pdf, although prediction regions with smaller volume may exist. The nonparametric prediction region (7.8) contains U_n of the training data cases $\boldsymbol{x}_i$ provided that $\boldsymbol{S}$ is nonsingular, even if the model is wrong. For many distributions, the coverage started to be close to $1 - \delta$ for $n \geq 10p$ where the coverage is the simulated percentage of times that the prediction region contained $\boldsymbol{x}_f$. Olive (2013) suggests $n \geq 50p$ may be needed for the prediction region to have a good volume. Of course for any n there are distributions that will have severe undercoverage.

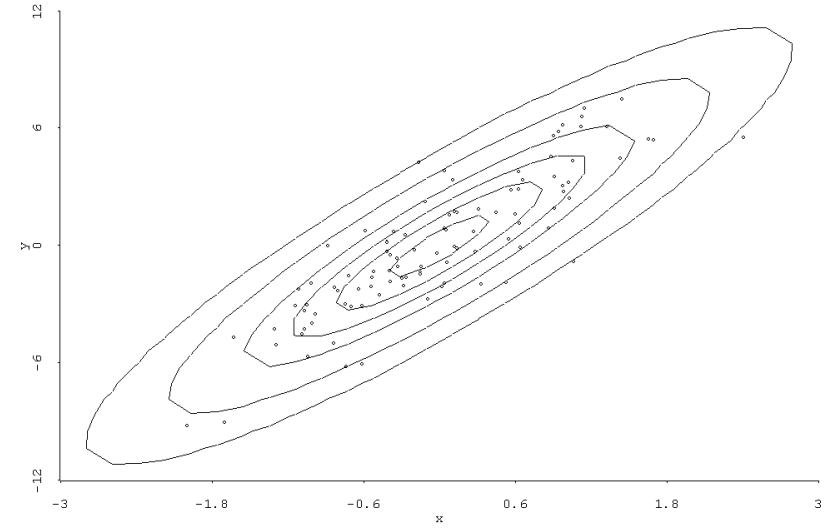
If $\boldsymbol{X}$ and $\boldsymbol{Z}$ have dispersion matrices $\boldsymbol{\Sigma}$ and $c\boldsymbol{\Sigma}$ where $c > 0$, then the dispersion matrices have the same shape. The dispersion matrices determine the shape of the hyperellipsoid $\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leq h^2\}$. Figure 7.2 was made with the *Arc* software of Cook and Weisberg (1999). The 10%, 30%, 50%, 70%, 90%, and 98% highest density regions are shown for two multivariate normal (MVN) distributions. Both distributions have $\boldsymbol{\mu} = \boldsymbol{0}$. In Figure 7.2a),

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 4 \end{pmatrix}.$$

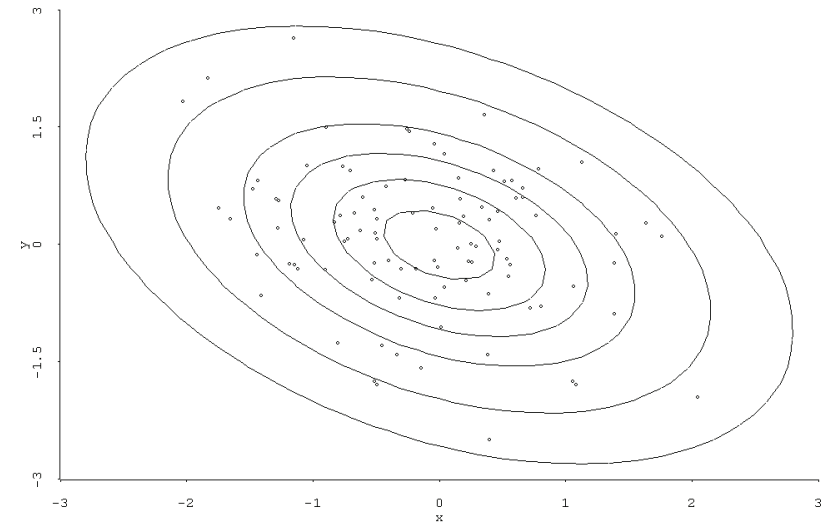
Note that the ellipsoids are narrow with high positive correlation. In Figure 7.2b),

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix}.$$

Note that the ellipsoids are wide with negative correlation. The highest density ellipsoids are superimposed on a scatterplot of a sample of size 100 from each distribution.



a)



b)

Figure 7.2. Highest Density Regions for 2 MVN Distributions

7.2 BOOTSTRAP CONFIDENCE REGIONS

For bootstrap confidence regions, if $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T_n^* - T_n) \xrightarrow{D} \boldsymbol{u}$, then the percentiles of $n(T_n - \boldsymbol{\theta}_0)^T \boldsymbol{C}_n^{-1}(T_n - \boldsymbol{\theta}_0)$ can be estimated with the sample percentiles of $n(T_n^* - T_n)^T \boldsymbol{C}_n^{-1}(T_n^* - T_n)$. Let $\boldsymbol{\theta}$ be a $g \times 1$ vector. For the correction factor below, and a nominal 95% confidence region, instead of using $D^2_{(\lceil 0.95B \rceil)}$ as the cutoff where $D^2_{(c)}$ is the cth order statistic of the D_i^2, the $100q_B$th sample quantile of the D_i^2, denoted by $D^2_{(U_B)}$, is used where $0.95B \leq U_B \leq 0.975B$ and $U_B \to 0.95B$ as B increases. Let $q_B = \min(1 - \delta + 0.05, 1 - \delta + g/B)$ for $\delta > 0.1$ and

$$q_B = \min(1 - \delta/2, 1 - \delta + 10\delta g/B), \quad \text{otherwise.} \tag{7.10}$$

If $1 - \delta < 0.999$ and $q_B < 1 - \delta + 0.001$, set $q_B = 1 - \delta$. This correction factor helps reduce undercoverage when $B \geq 50b$.

The following three confidence regions can be used for inference. The Olive (2017ab, 2018) prediction region method confidence region applies prediction region (7.8) to the bootstrap sample. Let the bootstrap sample be $T_1^*, ..., T_B^*$. Let $\overline{T}^*$ and $\boldsymbol{S}_T^*$ be the sample mean and sample covariance matrix of the bootstrap sample.

Definition 7.8. The large sample $100(1-\delta)\%$ prediction region method confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\} \tag{7.11}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

Olive (2017ab, 2018) also gave the modified Bickel and Ren (2001) confidence region that uses $\hat{\boldsymbol{\Sigma}}_A = n\boldsymbol{S}_T^*$.

Definition 7.9. The large sample $100(1 - \delta)\%$ modified Bickel and Ren confidence region is

$\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_{BT})}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_{BT})}\} \tag{7.12}$$

where the cutoff $D^2_{(U_{BT})}$ is the $100q_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - T_n)$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_B T)}$.

The hybrid confidence region is due to Pelawa Watagoda and Olive (2021).

Definition 7.10. Shift region (7.11) to have center T_n, or equivalently, change the cutoff of region (7.12) to $D^2_{(U_B)}$ to get the large sample $100(1 - \delta)\%$ hybrid confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\}. \tag{7.13}$$

Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if
$(T_n - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

Rajapaksha and Olive (2022) gave the following two confidence regions. The names of these confidence regions were chosen since they are similar to the Bickel and Ren and prediction region method confidence regions.

Definition 7.11. The large sample $100(1 - \delta)\%$ BR confidence region is

$$\{\boldsymbol{w} : n(\boldsymbol{w} - T_n)^T \boldsymbol{C}_n^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_{BT})}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{C}_n/n) \leq D^2_{(U_{BT})}\} \tag{7.14}$$

where the cutoff $D^2_{(U_{BT})}$ is the $100q_B$th sample quantile of the $D_i^2 = n(T_i^* - T_n)^T \boldsymbol{C}_n^{-1} (T_i^* - T_n)$ where q_B is found from (3) with $\boldsymbol{z}_i = T_i^*$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $n(T_n - \boldsymbol{\theta}_0)^T \boldsymbol{C}_n^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_{BT})}$.

Definition 7.12. The large sample $100(1 - \delta)\%$ PR confidence region for $\boldsymbol{\theta}$ is

$$\{\boldsymbol{w} : n(\boldsymbol{w} - \overline{T}^*)^T \boldsymbol{C}_n^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{C}_n/n) \leq D^2_{(U_B)}\} \tag{7.15}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = n(T_i^* - \overline{T}^*)^T \boldsymbol{C}_n^{-1}(T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $n(\overline{T}^* - \boldsymbol{\theta}_0)^T \boldsymbol{C}_n^{-1}(\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

The standard bootstrap confidence region is similar to what would be obtained if the classical prediction region (7.9) for multivariate normal data was applied to the bootstrap sample.

Definition 7.13. The large sample $100(1-\delta)\%$ standard bootstrap confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D^2_{1-\delta}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{1-\delta}\} \tag{7.16}$$

where $D^2_{1-\delta} = \chi^2_{g,1-\delta}$ or $D^2_{1-\delta} = d_n F_{g,d_n,1-\delta}$ where $d_n \to \infty$ as $n \to \infty$.

Much of the theory for the above confidence and prediction region appears in Olive (2023d, ch. 4, 5). If $n\boldsymbol{C}_n^{-1} = [\boldsymbol{S}_T^*]^{-1}$, then (7.14) and (7.15) are the modified Bickel and Ren (2001) and Olive (2017ab, 2018) prediction region method large sample $100(1-\delta)\%$ confidence regions for $\boldsymbol{\theta}$. Under regularity conditions, Bickel and Ren (2001) and Olive (2017b, 2018) proved that (7.11) and (7.12) are large sample confidence regions. Pelawa Watagoda and Olive (2021) gave simpler proofs. Pelawa Watagoda and Olive (2021) showed that under reasonable regularity conditions, i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, and iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$. Usually i) and ii) are proven using large sample theory. If $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma_u})$ with $\boldsymbol{\Sigma_u}$ nonsingular, then Pelawa Watagoda and Olive (2021) showed $\sqrt{n}(T_n - \overline{T}^*) \xrightarrow{P} \boldsymbol{0}$. Thus iii) and iv) hold if i) and ii) hold. If T_n is the sample mean or sample coordinatewise median, then see Bickel and Freedman (1981) and Rupasinghe Arachchige Don and Olive (2019). Then

$$D_1^2 = D^2_{T_i^*}(\overline{T}^*, \boldsymbol{C}_n/n) = \sqrt{n}(T_i^* - \overline{T}^*)^T \boldsymbol{C}_n^{-1} \sqrt{n}(T_i^* - \overline{T}^*),$$

$$D_2^2 = D^2_{\boldsymbol{\theta}}(T_n, \boldsymbol{C}_n/n) = \sqrt{n}(T_n - \boldsymbol{\theta})^T \boldsymbol{C}_n^{-1} \sqrt{n}(T_n - \boldsymbol{\theta}),$$

$$D_3^2 = D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{C}_n/n) = \sqrt{n}(\overline{T}^* - \boldsymbol{\theta})^T \boldsymbol{C}_n^{-1} \sqrt{n}(\overline{T}^* - \boldsymbol{\theta}), \quad \text{and}$$

$$D_4^2 = D^2_{T_i^*}(T_n, \boldsymbol{C}_n/n) = \sqrt{n}(T_i^* - T_n)^T \boldsymbol{C}_n^{-1} \sqrt{n}(T_i^* - T_n),$$

are well behaved. If $\boldsymbol{C}_n^{-1} \xrightarrow{P} \boldsymbol{C}^{-1}$, then $D_j^2 \xrightarrow{D} D^2 = \boldsymbol{u}^T \boldsymbol{C}^{-1} \boldsymbol{u}$, and (7.14) and (7.15) are large sample confidence regions. If $\boldsymbol{C}_n^{-1}$ is "not too ill conditioned," then $D_j^2 \approx \boldsymbol{u}^T \boldsymbol{C}_n^{-1} \boldsymbol{u}$ for large n, and the confidence regions (7.14) and (7.15) will have coverage near $1 - \delta$.

7.3 VISUALIZING THE NONPARAMETRIC PREDICTION REGION

Olive (2013) showed how to visualize the nonparametric prediction region (7.8) with the Rousseeuw and Van Driessen (1999) DD plot of classical distances versus robust distances on the vertical axis. See Section 7.5 where the exact same method will be used to visualize the bootstrap confidence region (7.11).

7.4 THE BOOTSTRAP

This section illustrates the nonparametric bootstrap with some examples. Suppose a statistic T_n is computed from a data set of n cases. The nonparametric bootstrap draws n cases with replacement from that data set. Then T_1^* is the statistic T_n computed from the sample. This process is repeated B times to produce the bootstrap sample $T_1^*, ..., T_B^*$. Sampling cases with replacement uses the empirical distribution.

Definition 7.14. Suppose that data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ has been collected and observed. Often the data is a random sample (iid) from a distribution with cdf F. The *empirical distribution* is a discrete distribution where the $\boldsymbol{x}_i$ are the possible values, and each value is equally likely. If $\boldsymbol{w}$ is a random variable having the empirical distribution, then $p_i = P(\boldsymbol{w} = \boldsymbol{x}_i) = 1/n$ for $i = 1, ..., n$. The *cdf of the empirical distribution* is denoted by F_n.

Example 7.1. Let $\boldsymbol{w}$ be a random variable having the empirical distribution given by Definition 14. Show that $E(\boldsymbol{w}) = \overline{\boldsymbol{x}} \equiv \overline{\boldsymbol{x}}_n$ and $\text{Cov}(\boldsymbol{w}) = \dfrac{n-1}{n}\boldsymbol{S} \equiv \dfrac{n-1}{n}\boldsymbol{S}_n$.

Solution: Recall that for a discrete random vector, the population expected value $E(\boldsymbol{w}) = \sum \boldsymbol{x}_i p_i$ where $\boldsymbol{x}_i$ are the values that $\boldsymbol{w}$ takes with positive probability p_i. Similarly, the population

covariance matrix

$$\text{Cov}(\boldsymbol{w}) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(\boldsymbol{w} - E(\boldsymbol{w}))^T] = \sum (\boldsymbol{x}_i - E(\boldsymbol{w}))(\boldsymbol{x}_i - E(\boldsymbol{w}))^T p_i.$$

Hence

$$E(\boldsymbol{w}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{1}{n} = \overline{\boldsymbol{x}},$$

and

$$\text{Cov}(\boldsymbol{w}) = \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \frac{1}{n} = \frac{n-1}{n} \boldsymbol{S}. \ \square$$

Example 7.2. If $W_1, ..., W_n$ are iid from a distribution with cdf F_W, then the empirical cdf F_n corresponding to F_W is given by

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(W_i \leq y)$$

where the indicator $I(W_i \leq y) = 1$ if $W_i \leq y$ and $I(W_i \leq y) = 0$ if $W_i > y$. Fix n and y. Then $nF_n(y) \sim$ binomial $(n, F_W(y))$. Thus $E[F_n(y)] = F_W(y)$ and $V[F_n(y)] = F_W(y)[1 - F_W(y)]/n$. By the central limit theorem,

$$\sqrt{n}(F_n(y) - F_W(y)) \xrightarrow{D} N(0, F_W(y)[1 - F_W(y)]).$$

Thus $F_n(y) - F_W(y) = O_P(n^{-1/2})$, and F_n is a reasonable estimator of F_W if the sample size n is large.

Suppose there is data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ collected into an $n \times p$ matrix $\boldsymbol{W}$. Let the statistic $T_n = t(\boldsymbol{W}) = T(F_n)$ be computed from the data. Suppose the statistic estimates $\boldsymbol{\mu} = T(F)$, and let $t(\boldsymbol{W}^*) = t(F_n^*) = T_n^*$ indicate that t was computed from an iid sample from the empirical distribution F_n: a sample $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ of size n was drawn with replacement from the observed sample $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. This notation is used for von Mises differentiable statistical functions in large sample theory. See Serfling (1980, ch. 6). The empirical distribution is also important for the influence function

(widely used in robust statistics). The *nonparametric bootstrap* draws B samples of size n from the rows of $\boldsymbol{W}$, e.g. from the empirical distribution of $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. Then T^*_{jn} is computed from the jth bootstrap sample for $j = 1, ..., B$.

Example 7.3. Suppose the data is 1, 2, 3, 4, 5, 6, 7. Then $n = 7$ and the sample median T_n is 4. Using R, we drew $B = 2$ bootstrap samples (samples of size n drawn with replacement from the original data) and computed the sample median $T^*_{1,n} = 3$ and $T^*_{2,n} = 4$.

```
b1 <- sample(1:7,replace=T)
b1
[1] 3 2 3 2 5 2 6
median(b1)
[1] 3
b2 <- sample(1:7,replace=T)
b2
[1] 3 5 3 4 3 5 7
median(b2)
[1] 4
```

7.5 VISUALIZING SOME BOOTSTRAP CONFIDENCE REGIONS

As mentioned in previous section, the DD plot will be used to visualize some bootstrap confidence regions. If a good robust estimator is used, Olive (2002) showed that the plotted points in a DD plot cluster about the identity line with zero intercept and unit slope if the $\boldsymbol{x}_i$ are iid from a multivariate normal distribution with nonsingular covariance matrix, while the plotted points cluster about some other line through the origin if the $\boldsymbol{x}_i$ are iid from a large family of nonnormal elliptically contoured distributions. For the robust estimator of multivariate location and dispersion, we recommend the RFCH or RMVE estimator. These two estimators $(T_n, \boldsymbol{C}_n)$ are such that $\boldsymbol{C}_n$ is a $\sqrt{n}$ consistent estimator of $a\text{Cov}(\boldsymbol{x})$ for a large class of elliptically contoured distributions where the constant $a > 0$ depends on the elliptically contoured distribution and the estimator RFCH or

RMVN, and $a = 1$ for the multivariate normal distribution with nonsingular covariance matrix. We will use the RMVN estimator in the software.

Example 7.4. We generated $\boldsymbol{x}_i \sim N_4(\mathbf{0}, \boldsymbol{I})$ for $i = 1, ..., 250$. The coordinatewise median was the statistic T_n. The nonparametric bootstrap was used with $B = 1000$. The DD plot of the bootstrap sample is shown in Figure 7.3. The plotted points cluster about the identity line. The vertical line MD = 2.9098 is the cutoff for the prediction region method confidence region (7.11). The long horizonal line RD = 3.0995 is the cutoff using the robust estimator. When $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}_T)$, then under mild regularity conditions, $\sqrt{n}(T_n - \overline{T}_n^*) \xrightarrow{P} \mathbf{0}$. The short horizontal line is RD = 2.8074 and MD = 2.8074 is approximately the cutoff that would be used by the standard bootstrap confidence region (mentally drop a vertical line from where the short horizontal line ends at the identity line). Variability in DD plots increases as RD increases. The *R* commands for making the plot are shown below.

```
source("http://parker.ad.siu.edu/Olive/mpack.txt")
x <-matrix(rnorm(1000),nrow=250,ncol=4)
out<-rhotboot(x)
ddplot4(out$mus)
$cuplim
   90.4%
2.809824
$ruplim
   90.4%
3.095542
$mvnlim
[1] 2.807479
```

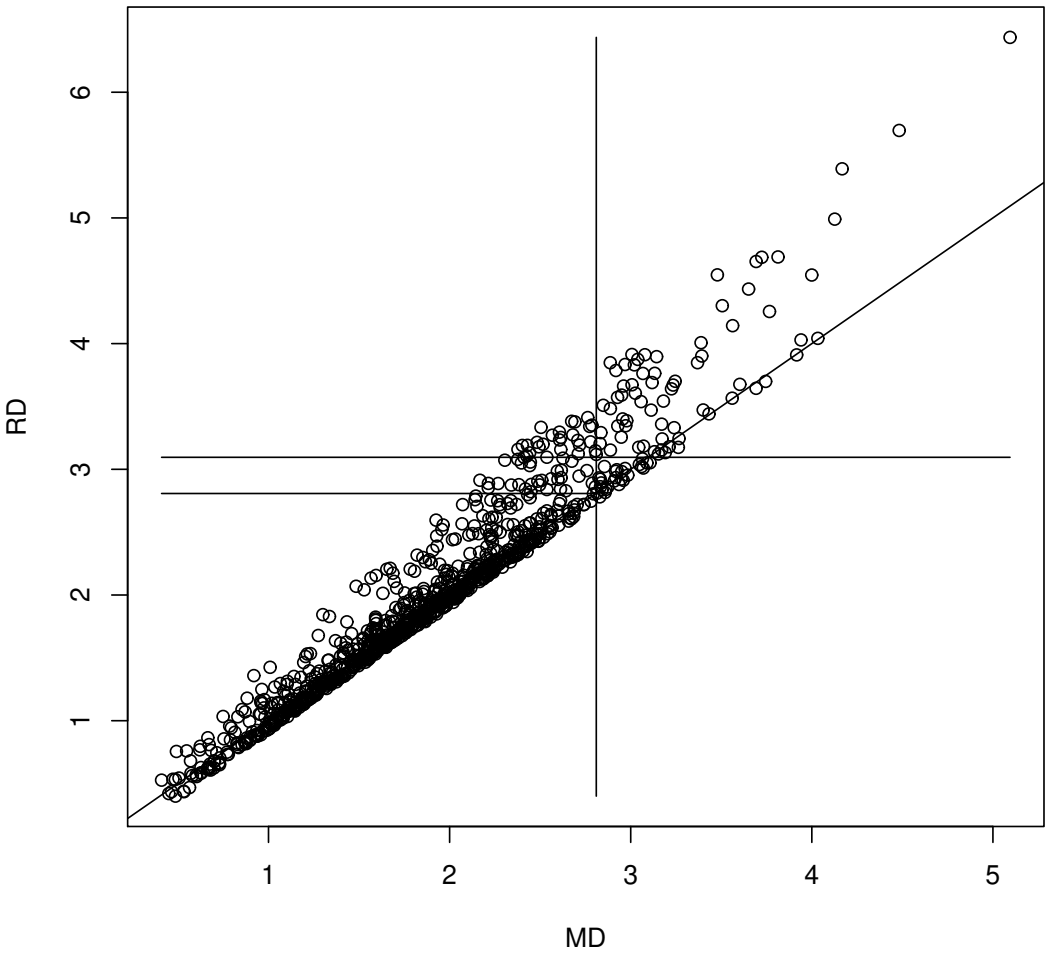


Figure 7.3. Visualizing the confidence region with a DD plot

Some *R* functions for bootstrapping several statistics are shown below.

```
source("http://parker.ad.siu.edu/Olive/slpack.txt")
args(bicboot)  #bootstrap min BIC model forward selection regression
#function (x, y, B = 1000)
args(FDAboot) #Bootstraps FDA betahat = first eigenvector.
#function (x, group, B = 1000)
args(fselboot2) #bootstrap min Cp model forward selection regression
#function (x, y, B = 1000, c = 0.01, aug = F)
args(lassoboot2) #bootstrap lasso or ridge regression for MLR
#function (x, y, B = 1000, regtype = 1, c = 0.01, aug = F)
args(LPHboot)  #bootstraps the Cox regression lasso, takes a few minutes
#function (x, time, status, B = 1000)
args(LRboot)  #bootstrap logistic regression  full model
#function (x, y, mv = c(1, 1), B = 1000, bin = T)
args(pcaboot) #Bootstraps PCA. Likely only accurate for positive eigenvalues
#function (x, corr = T, rob = F, B = 1000)
args(PHboot)  #bootstraps the Cox PH regression full model
#with the nonparametric bootstrap
#function (x, time, status, B = 1000)
args(PRboot)  #bootstraps the Poisson regression full model
#function (x, y, B = 1000)
args(regboot) #residual bootstrap for MLR
function (x, y, B = 1000)
args(rowboot) #nonparametric bootstrap for MLR
#function (x, y, B = 1000)

source("http://parker.ad.siu.edu/Olive/mpack.txt")
args(corboot)  #rowwise nonparametric bootstrap of the correlation matrix
```

```
#function (x, B = 1000) #stacks entries above the diagonal into a vector beta
args(rhotboot) #Bootstraps  RMVN center (med=F) or coordinatewise median.
#function (x, B = 1000, med = T)

source("http://parker.ad.siu.edu/Olive/tspack.txt")
args(arboot) Bootstraps AR(p) model selection using the parametric bootstrap
#function (Y, B = 100, pmax = 10, c = 0.01)
args(arboot2) #Bootstraps AR(p) model selection using the residual bootstrap.
#function (Y, B = 100, pmax = 10, c = 0.01)
args(maboot) #Bootstraps MA(q) model selection using the parametric bootstrap.
#function(Y,B=100,qmax=10,c=0.05)
args(maboot2) #Bootstraps MA(q) model selection using the residual bootstrap.
#function(Y,B=100,qmax=10,c=0.05)
```

CHAPTER 8

DISCUSSION

Plots and simulations were done in *R*. See R Core Team (2020). Programs are in the collection of functions *tspack.txt*. See (http://parker.ad.siu.edu/Olive/tspack.txt).

Two *tspack* functions are useful for illustrating least squares (OLS) applied to AR and MA time series. The function `arp` fits an AR(p) model to time series Y using OLS. The function assumes $p < n - p$, and $(n - p) > 10p$ would be useful. The function `maq` fits an MA(q) model to time series Y using OLS and the residuals from the MA GMLE. The function assumes $q < n - q$, and $(n - q) > 10q$ would be useful. Often $n \geq 1000$ was needed for the MA GMLE and OLS estimators to be close.

For AR(p) data splitting, the *tspack* function `dsarsim` used the built in "AIC" model selection from the *R* function `ar` and tended to underfit for $n < 20$ pmax. The *tspack* function `dsarsim2` used Equation (2.4) for AR(p) model selection with AIC for $n < 14(pmax)$ and BIC otherwise.

For MA(q) data splitting, the function `dsmasim` used `auto.arima` while `dsmasim2` used Equatin (2.4).

The function `armamsel1` performs Pötscher, B.M. (1990) ARMA model selection method, while the function `armasim1` does the simulation. The function `armamsel2` performs the new ARMA model selection method described in Section 2, while the function `armasim2` does the simulation.

The bootstrap is due to Efron (1979). Also see Efron (1982) and Olive (2017ab, 2023abcd). Rathnayake and Olive (2021) show how to bootstrap many variable selection estimators. Haile and Olive (2023a) show how to bootstrap AR, MA, and ARMA time series model selection estimators. See Haile, Zhang and Olive (2023) for prediction regions if n/p is small.

The `rpack` function ddplot4 applied to the bootstrap sample can be used to visualize the bootstrap prediction region method confidence region.

REFERENCES

[1] Agnieszka, D., and Magdalena, L. (2018), "Detection of Outliers in the Financial Time Series Using ARIMA Models," *Applications of Electromagnetics in Modern Techniques and Medicine (PTZE)*, 2018, 49-52.

[2] Allende, H., and Heiler, S. (1992), "Recursive Generalized M Estimates for Autoregressive Moving-Average Models," *Journal of Time Series Analysis,* 13, 1-18.

[3] Akaike, H. (1973), "Information Theory as an Extension of the Maximum Likelihood Principle," in *Proceedings, 2nd International Symposium on Information Theory*, eds. Petrov, B.N., and Csakim, F., Akademiai Kiado, Budapest, 267-281.

[4] Anderson, T.W. (1971), *The Statistical Analysis of Time Series*, Wiley, Hoboken, NJ.

[5] Basu, S., and Meckesheimer, M. (2007), "Automatic Outlier Detection for Time Series: an Application to Sensor Data," *Knowledge and Information Systems,* 11, 137-154.

[6] Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P. (2016), "Efficient and Consistent Robust Time Series Analysis," arXiv preprint arXiv:1607.00146, arxiv.org.

[7] Bickel, P.J., and Freedman, D.A. (1981), "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics*, 9, 1196-1217.

[8] Bickel, P.J., and Ren, J.–J. (2001), "The Bootstrap in Hypothesis Testing," in *State of the Art in Probability and Statistics: Festschrift for William R. van Zwet*, eds. de Gunst, M., Klaassen, C., and van der Vaart, A., The Institute of Mathematical Statistics, Hayward, CA, 91-112.

[9] Blázquez-Garcïa, A., Conde, A., Mori, U., and Lozano, J.A. (2020), "A Review on Outlier/Anomaly Detection in Time Series Data," (https://arxiv.org/pdf/2002.04236.pdf).

[10] Box, G.E.P, and Jenkins, G.M. (1976), *Time Series Analysis: Forecasting and Control*, revised ed., Holden-Day, Oakland, CA.

[11] Brockwell, P.J., and Davis, R.A. (1987), *Time Series: Theory and Methods*, Springer, New York, NY.

[12] Burnham, K.P., and Anderson, D.R. (2004), "Multimodel Inference Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, 33, 261-304.

[13] Bustos, O.H., and Yohai, V.J. (1986), "Robust Estimates for ARMA Models," *Journal of the American Statistician*, 81, 155-168.

[14] Chakhchoukh, Y. (2010), "A New Robust Estimation Method for ARMA Models," *IEEE Transactions on Signal Processing,* 58, 3512-3522.

[15] Chan, N.H., Ling, S., and Yau, C.Y. (2020), "Lasso-Based Variable Selection of ARMA Models," *Statistica Sinica*, 30, 1925-1948.

[16] Chan, W.-S. (1995), "Understanding the Effect of Time Series Outliers on Sample Autocorrelations, *Test*, 4, 179-186.

[17] Chang, I., Tiao, G.C., and Chen, C. (1988), "Estimation of Time Series Parameters in the Presence of Outliers," *Technometrics*, 30, 193-204.

[18] Chen, C. and Liu, L. (1993), "Joint Estimation of Model Parameters and Outlier Effects in Time Series," *Journal of the American Statistical Association*, 88, 284-297.

[19] Chew, V. (1966), "Confidence, Prediction and Tolerance Regions for the Multivariate Normal Distribution," *Journal of the American Statistical Association,* 61, 605-617.

[20] Choy, K. (2001), "Outlier Detection for Stationary Time Series," *Journal of Statistical Planning and Inference,* 99, 111-127.

[21] Claeskens, G., and Hjort, N.L. (2008), *Model Selection and Model Averaging*, Cambridge University Press, New York, NY.

[22] Cook, R.D., and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics,* Wiley, New York, NY.

[23] Cryer, J.D., and Chan, K.-S. (2008), *Time Series Analysis: with Applications in R*, 2nd ed., Springer, New York, NY.

[24] de Luna, X., and Genton, M.G. (2001), "Robust Simulation-Based Estimation of ARMA Models," *Journal of Computational and Graphical Statistics*, 10, 370-387.

[25] Denby, L., and Martin, R.D. (1979), "Robust Estimation of the First-Order Autoregressive Parameter," *Journal of the American Statistical Association*, 74, 365, 140-146.

[26] Deutsch, S.J., Richards, J.E., and Swain, J.J. (1990), "Effects of a Single Outlier on ARMA

Identification," *Communications in Statistics: Theory and Methods*, 19, 2207-2227.

[27] Duong, Q.P. (1984), "On the Choice of the Order of Autoregressive Models: a Ranking and Selection Approach," *Journal of Time Series Analysis,* 5, 145-157.

[28] Durbin, J. (1959), "Efficient Estimation of Parameters in Moving-Average Models," *Biometrika*, 46, 306-316.

[29] Efron, B. (1979), "Bootstrap Methods, Another Look at the Jackknife," *The Annals of Statistics*, 7, 1-26.

[30] Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans,* SIAM, Philadelphia, PA.

[31] Fox, A.J. (1972), "Outliers in Time Series," *Journal of the Royal Statistical Society: B*, 34, 350-363.

[32] Granger, C.W.J., and Newbold, P. (1977), *Forecasting Economic Time Series*, Academic Press, New York, NY.

[33] Haile, M.G., and Olive, D.J. (2023a), "Bootstrapping ARMA Time Series Models after Model Selection," preprint at (http://parker.ad.siu.edu/Olive/pptsboot.pdf).

[34] Haile, M.G., and Olive, D.J. (2023b), "Prediction Intervals for Some ARIMA Time Series," is at (http://parker.ad.siu.edu/Olive/pptspi.pdf).

[35] Haile, M.G., Zhang, L., and Olive, D.J. (2023), "Prediction Intervals and Regions for Random Walks and Renewal Processes" is at (http://parker.ad.siu.edu/Olive/pprwalkpi.pdf).

[36] Hamilton, J.D. (1994), *Time Series Analysis*, Princeton University Press, Princeton, NJ.

[37] Hannan, E.J. (1973), "The Asymptotic Theory of Linear Time-Series Models," *Journal of Applied Probability*, 10, 130-145.

[38] Hannan, E.J. (1980), "The Estimation of the Order of an ARMA Process," *The Annals of Statistics*, 8, 1071-1081.

[39] Hannan, E.J., and Quinn, B.G. (1979), "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society, B*, 41, 190-195.

[40] Hannan, E.J., and Rissanen, J. (1982), "Recursive Estimation of Mixed Autoregressive-

Moving Average Order," *Biometrika*, 69, 81-94.

[41] Hawkins, D.M., and Olive, D.J. (2002), "Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm," *Journal of the American Statistical Association,* (with discussion), 97, 136-148.

[42] Huber, P.J., and Ronchetti, E.M. (2009), *Robust Statistics,* 2nd ed., Wiley, Hoboken, NJ.

[43] Hurvich, C., and Tsai, C.L. (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297-307.

[44] Hyndman, R.J., and Athanasopoulos, G. (2018), *Forecasting: Principles and Practice*, 2nd edition, OTexts: Melbourne, Australia. https://OTexts.org/fpp2/

[45] Hyndman, R.J., and Khandakar, Y. (2008), "Automatic Time Series Forecasting: the Forecast Package for R." *Journal of Statistical Software*, 26, 1-22.

[46] Iturria, A., Carraso, J., Herrera, F., Charramendieta, S., and Intxausti, K. (2019), *otsad: Online Time Series Anomaly Detectors, R* Package version 0.2.0, (http://cran.r-project.org/package=otsad).

[47] Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

[48] Jones, R.H. (1980), "Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations," *Technometrics*, 22, 389-395.

[49] Justel, A., Peña, D., and Tsay, R.S. (2001), "Detection of Outlier Patches in Autoregressive Time Series," *Statistica Sinica,* 11, 651-673.

[50] Kreiss, J.P. (1985), "A Note on M-Estimation in Stationary ARMA Processes," *Statistics & Decisions*, 3, 317-336.

[51] Lawrence, C.J. (2014), "Robust Methods in Time Series Analysis," *Wiley StatsRef: Statistics Reference Online*.

[52] Ledolter, J. (1989), "The Effect of Additive Outliers on the Forecasts from ARIMA Models," *International Journal of Forecasting*, 5, 231-240.

[53] Lee, Y.S., and Scholtes, S. (2014), "Empirical Prediction Intervals Revisited," *International*

Journal of Forecasting, 30, 217-234.

[54] Liu, J., Kumar, S., and Palomar, D.P. (2019), "Parameter Estimation of Heavy-Tailed AR Model with Missing Data Via Stochastic EM," *IEEE Transactions on Signal Processing*, 67, 2159-2172.

[55] Lucas, A., Franses, P.H., and Van Dijk, D. (2009), *Outlier Robust Analysis of Economic Time Series*, Oxford University Press, Oxford, UK.

[56] Ma, Y., and Genton, M.G. (2000), "Highly Robust Estimation of the Autocovariance Function," *Journal of Time Series Analysis,* 21, 663-684.

[57] Mallows, C. (1973), "Some Comments on C_p," *Technometrics,* 15, 661-676.

[58] Mann, H.B., and Wald, A. (1943), "On the Statistical Treatment of Linear Stochastic Difference Equations," *Econometrica*, 11, 173-220.

[59] McElroy, T.S., and Politis, D.N. (2020), *Time Series: a First Course With Bootstrap Starter*, CRC Press Taylor & Francis, Boca Raton, FL.

[60] Muler, N., Peña, D., and Yohai, V. (2009), "Robust Estimation for ARMA Models," *The Annals of Statistics*, 37, 816-840.

[61] Olive, D.J. (2002), "Applications of Robust Distances for Regression," *Technometrics,* 44, 64-71.

[62] Olive, D.J. (2008), A Course in Statistical Theory, online course notes at (http://parker.ad.siu.edu/Olive/infbook.htm)

[63] Olive, D.J. (2013), "Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data," *International Journal of Statistics and Probability*, 2, 90-100.

[64] Olive, D.J. (2014), *Statistical Theory and Inference*, Springer, New York, NY.

[65] Olive, D.J. (2017a), *Linear Regression*, Springer, New York, NY.

[66] Olive, D.J. (2017b), *Robust Multivariate Analysis*, Springer, New York, NY.

[67] Olive, D.J. (2018), "Applications of Hyperellipsoidal Prediction Regions," *Statistical Papers*, 59, 913-931.

[68] Olive, D.J. (2023a), *Prediction and Statistical Learning*, online course notes, see

(http://parker.ad.siu.edu/Olive/slearnbk.htm).

[69] Olive, D.J. (2023b), *Robust Statistics*, online course notes at (http://parker.ad.siu.edu/Olive/robbook.html).

[70] Olive, D.J. (2023c), *Theory for Linear Models*, online course notes, (http://parker.ad.siu.edu/Olive/linmodbk.htm).

[71] Olive (2023d) *Large Sample Theory*: online course notes, (http://parker.ad.siu.edu/Olive/lsampbk.pdf).

[72] Olive, D. J., and Hawkins, D. M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

[73] Olive, D.J., and Hawkins, D.M. (2010), "Robust Multivariate Location and Dispersion," Preprint, see (http://parker.ad.siu.edu/Olive/pphbmld.pdf).

[74] Pankratz, A. (1983), *Forecasting with Univariate Box-Jenkins Models*, Wiely, New York, NY.

[75] Pelawa Watagoda, L.C.R., and Olive, D.J. (2021), "Bootstrapping Multiple Linear Regression after Variable Selection," *Statistical Papers,* 62, 681-700. See (http://parker.ad.siu.edu/Olive/ppboottest.pdf).

[76] Pötscher, B.M. (1990), "Estimation of Autoregressive Moving-Average Order Given an Infinite Number of Models and Approximation of Spectral Sensities," *Journal of Time Series Analysis*, 11, 165-179.

[77] Pötscher, B.M., and Srinivasan, S. (1994), "A Comparison of Order Estimation procedures for ARMA Models," *Statistica Sinica*, 4, 29-50.

[78] Rajapaksha, K.W.G.D.H., and Olive, D.J. (2022), "Wald Type Tests with the Wrong Dispersion Matrix," *Communications in Statistics: Theory and Methods*, to appear.

[79] Rathnayake, R.C., and Olive, D.J. (2023), "Bootstrapping Some GLM and Survival Regression Variable Selection Estimators," *Communications in Statistics: Theory and Methods*, 52, 2625-2645.

[80] Rousseeuw, P.J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

[81] Rupasinghe Arachchige Don, H.S., and Olive, D.J. (2019), "Bootstrapping Analogs of the One Way MANOVA Test," *Communications in Statistics: Theory and Methods*, 48, 5546-5558.

[82] R Core Team (2018), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

[83] Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics,* 6, 461-464.

[84] Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley.

[85] Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, 88, 486-494.

[86] Shibata, R. (1976), "Selection of the Order of an Autoregressive Model by Akaike's Information Criterion," *Biometrika*, 63, 117-126.

[87] Stockinger, N., and Dutter, R. (1987), "Robust Times Series Analysis: a Survey," *Kybernetika*, 23, 3-88.

[88] Tsay, R.S. (1986), "Time Series Model Specification in the Presence of Outliers," *Journal of the American Statistical Association*, 81, 132-141.

[89] Tsay, R.S. (1988), "Outliers, Level Shifts, and Variance Changes in Time Series," *Journal of Forecasting*, 7, 1-20.

[90] Whittle, P. (1953), "Estimation and Information in Stationary Time Series," *Arkiv för Matematik*, 2, 423-34.

[91] Yao, Q. and Brockwell, P.J. (2006), "Gaussian Maximum Likelihood Estimation for ARMA Models I: Time Series," *Journal of Time Series Analysis*, 27, 857-875.

[92] Zhang, J., Olive, D.J., and Ye, P. (2012), "Robust Covariance Matrix Estimation With Canonical Correlation Analysis," *International Journal of Statistics and Probability*, 1, 119-136.

VITA

Graduate School
Southern Illinois University Carbondale

Welagedara Arachchilage Dhanushka Madumali Welagedara

dhanushkawel@gmail.com

University of Ruhuna, Sri Lanka
Bachelor of Science, December 2015

University of Peradeniya, Sri Lanka
Master of Science, Mathematics, December 2018

Special Honors and Awards:
Dissertation Research Assistantship Award (Summer 2023)

Dissertation Paper Title:
Model Selection, Data Splitting for ARMA Time Series and Visualizing some Bootstrap Confidence Regions.

Major Professor: Dr. David Olive

Publications:
W.A.D.M. Welagedara, Lakshika S. Nawarathna Ruwan D. Nawarathna, "Forecasting the Sri Lankan Population with the Gompertz and Verhulst Logistic Growth Models", *Sri Lanka Journal of Economic Research*, Vol. 07(1): pp 1-12, 2019