

# Interventions and Outliers

- ▶ We now discuss the situation of particular regions of a time series (or particular points along a time series) at which the behavior of the series changes or is unusual compared to the bulk of the series.
- ▶ Sometimes an outside agency causes the underlying process to change abruptly.
- ▶ In other cases, we simply observe an unusual value at a particular time (or at several times).

# Intervention Analysis

- ▶ When some outside force causes the behavior of a time series process to change abruptly, this is called an *intervention*.
- ▶ See the plot of the `airmiles` data in R, which shows the monthly airline passenger-miles in the U.S. from January 1996 to May 2005.
- ▶ The seasonality is clearly apparent, with higher travel in the summer and in December.
- ▶ There is an overall increasing linear trend, but there is a substantial drop in September 2001, clearly since the terrorist attacks that month caused a decrease in air travel.
- ▶ After the sudden drop, the air travel gradually continued to rise.
- ▶ In this case, the mean function changes at a specific time point.

# Causes of Interventions

- ▶ Sometimes an intervention is caused by a natural event, such as a weather disaster causing an abrupt change in population level for human or animal inhabitants of a certain location.
- ▶ Or sometimes the change can be manmade, such as a new seatbelt law or new speed limit causing changes in traffic death patterns.
- ▶ In the airmiles example, the intervention caused a change in the mean function.
- ▶ In other situations, the intervention could cause a change in the autocovariance function (which includes the variance function, i.e., the lag-0 autocovariance), but we will not discuss intervention analysis other than modeling the effects on the mean function.

# A Simple Single-Intervention Model

- ▶ Suppose we have the observed series  $Y_t = m_t + N_t$ , where  $N_t$  is some specific ARIMA process.
- ▶ The process  $\{N_t\}$  is what the process would look like if there were no intervention.
- ▶ The  $m_t$  is the change in the mean function due to the intervention, which occurs at time  $T$ .
- ▶ Before time  $T$ , we see that  $m_t$  is constant at zero.
- ▶ The pre-intervention time series data,  $\{Y_t\}$ ,  $t < T$ , can be used to specify the model for  $N_t$ .
- ▶ Define the *step function*  $S_t^{(T)}$  to be 1 if  $t \geq T$  and 0 otherwise; this is an indicator of being in the post-intervention time period.
- ▶ Then the *pulse function*  $P_t^{(T)} = S_t^{(T)} - S_{t-1}^{(T)}$  is 1 if  $t = T$  and 0 otherwise.

# More on the Single-Intervention Model

- ▶ If the intervention results in an immediate and permanent shift in the mean function, we can model  $m_t$  as:

$$m_t = \omega S_t^{(T)}$$

where  $\omega$  is the permanent change in the mean.

- ▶ If  $\omega > 0$ , the mean function shifts up at time  $T$ .
- ▶ If  $\omega < 0$ , the mean function shifts down at time  $T$ .
- ▶ We can formally test whether  $\omega = 0$  to determine whether the intervention has any significant effect at all.
- ▶ But note that the pre-intervention and post-intervention data are not two independent samples; they are autocorrelated.

# Delays and Gradual Onsets of Interventions

- ▶ If there is a known *delay* of  $d$  time units between the time of the intervention and when it takes effect, then we can model  $m_t$  as:

$$m_t = \omega S_{t-d}^{(T)}$$

- ▶ In other cases, the intervention may take effect on the mean function only gradually, eventually reaching its full effect.
- ▶ In this case, we can specify  $m_t$  with something similar to an  $AR(1)$  model:

$$m_t = \delta m_{t-1} + \omega S_{t-1}^{(T)}$$

with initial condition  $m_0 = 0$ .

- ▶ This is equivalent to

$$m_t = \omega \frac{1 - \delta^{t-T}}{1 - \delta}, \text{ for } t > T$$

and  $m_t = 0$  if  $t \leq T$ , where  $0 < \delta < 1$ .

## More on Gradual Onsets of Interventions

- ▶ As  $t$  grows large,  $m_t$  approaches  $\omega/(1 - \delta)$ , which is the long-run effect of the intervention on the mean function, the ultimate change in the mean function (see R example graph 11.3(b)).
- ▶ When  $t = T + \log(0.5)/\log(\delta)$ , the intervention effect reaches half of its ultimate value, so this time duration  $\log(0.5)/\log(\delta)$  is called the *half-life* of the intervention effect.
- ▶ If  $\delta$  is near 0, the half-life is small, and the full effect of the intervention is quickly felt.
- ▶ If  $\delta$  is near 1, the half-life is large, and the full effect of the intervention takes a long time to be felt.
- ▶ In the extreme case when  $\delta = 1$ , then  $m_t = \omega(t - T)$  for  $t \geq T$  and 0 otherwise.
- ▶ This corresponds to a linear intervention effect with slope  $\omega$  that continues to increase (if  $\omega > 0$ ) linearly over time (see R example graph 11.3(c)).

## Some Models for Types of Interventions

- ▶ If the intervention only affects the mean function at time  $T$ , we can model it as:

$$m_t = \omega P_t^{(T)}$$

where  $P_t^{(T)}$  is the pulse variable that is 1 if  $t = T$  and 0 otherwise.

- ▶ If the intervention effect dies out gradually, we can model it as:

$$m_t = \delta m_{t-1} + \omega P_t^{(T)}$$

or equivalently,  $m_t = \omega \delta^{t-T}$  for  $t \geq T$ .

- ▶ This would imply that the intervention immediately changes the mean function by  $\omega$ , but then its effect dies off over time geometrically by a factor of  $\delta$ .

# A Model for a Delayed Intervention

- ▶ We could also model a delay before the intervention takes effect; for example, for a one-time-unit delay:

$$m_t = \delta m_{t-1} + \omega P_{t-1}^{(T)}$$

with  $m_0 = 0$  (see R example graph 11.4(a)).

# The Backshift Operator

- ▶ Recall the backshift operator  $B$ , defined so that  $Bm_t = m_{t-1}$  and  $BP_t^{(T)} = P_{t-1}^{(T)}$ .
- ▶ Using the previous model:

$$\begin{aligned}m_t &= \delta m_{t-1} + \omega P_{t-1}^{(T)} \\ \Rightarrow m_t - \delta Bm_t &= \omega BP_t^{(T)} \\ \Rightarrow (1 - \delta B)m_t &= \omega BP_t^{(T)}\end{aligned}$$

so that

$$m_t = \frac{\omega B}{(1 - \delta B)} P_t^{(T)}.$$

- ▶ Furthermore, note that  $(1 - B)S_t^{(T)} = P_t^{(T)}$ , or equivalently  $S_t^{(T)} = [1/(1 - B)]P_t^{(T)}$ , so these models for  $m_t$  can be specified in terms of either the pulse function or the step function.

## More Complex Intervention Models

- ▶ The model

$$m_t = \frac{\omega_1 B}{(1 - \delta B)} P_t^{(T)} + \frac{\omega_2 B}{(1 - B)} P_t^{(T)}$$

describes an intervention effect that (after a one-time-unit delay) achieves a value of  $\omega_1 + \omega_2$  and then gradually fades to its limiting value of  $\omega_2$  (see R example graph 11.4(b) for the case when  $\omega_1 > 0, \omega_2 > 0$ ).

- ▶ The model

$$m_t = \omega_0 P_t^{(T)} + \frac{\omega_1 B}{(1 - \delta B)} P_t^{(T)} + \frac{\omega_2 B}{(1 - B)} P_t^{(T)}$$

describes an intervention effect that immediately achieves a value of  $\omega_0$ , then (after a one-time-unit delay) changes abruptly to a value of  $\omega_1 + \omega_2$  and then gradually fades to its limiting value of  $\omega_2$  (see R example graph 11.4(c) for the case when  $\omega_0 > 0, \omega_1 < 0, \omega_2 < 0$ ).

# Estimating the Parameters of the Intervention Model

- ▶ The ARIMA model for the  $N_t$  process is specified using the pre-intervention data.
- ▶ Then maximum likelihood can be used to estimate the model parameters, including any  $\omega$ 's and  $\delta$ , as well as the parameters for the ARIMA model specified for  $N_t$ .

# Modeling the Air Miles Data

- ▶ Based on the full time series plot of the airmiles data, there seemed to be an immediate change in the mean function at the intervention time  $T$  (September 2001).
- ▶ The intervention effect gradually gets smaller as time goes on.
- ▶ The following model was used for the intervention effect:

$$m_t = \omega_0 P_t^{(T)} + \frac{\omega_1}{1 - \omega_2 B} P_t^{(T)}$$

- ▶ Under this model, the immediate change to the mean response function is  $\omega_0 + \omega_1$ , while the effect  $k$  time units after  $T$  is  $\omega_1(\omega_2)^k$ .

## Specifying the Model for the Air Miles Data

- ▶ Based on the pre-intervention airmiles data (through August 2001), there appears to be nonstationarity and seasonality (with seasonal period  $s = 12$ ).
- ▶ We thus take both the first differences and the seasonal ( $s = 12$ ) differences.
- ▶ After doing that, based on the ACF and PACF, we tentatively specify an MA(1) model for the differenced and seasonally differenced pre-intervention data.
- ▶ For our final model, we will use a seasonal  $ARIMA(0, 1, 1) \times (0, 1, 0)_{12}$  model for  $N_t$  and also incorporate our intervention model for  $m_t$ .
- ▶ But from the standardized residual plot, we note a serious outlier, so we will wait to fit the final model until we discuss handling outliers.

# Outliers in Time Series

- ▶ Outliers are individual observations that are highly unusual relative to the pattern of the overall time series.
- ▶ They may be due to measurement error or data recording error, but also may occur because the underlying process briefly changes for some reason.
- ▶ We will define two types of outliers: the *additive outlier* (AO) and the *innovative outlier* (IO).
- ▶ An additive outlier at time  $T$  is essentially an intervention with a pulse response at time  $T$ .
- ▶ With an AO at time  $T$ , the time series is only affected at time  $T$ .
- ▶ With an IO at time  $T$ , the time series is affected at and after time  $T$ , but the effect of the IO grows less as time gets farther away from  $T$ .

# Modeling an Additive Outlier

- ▶ If a time series has an additive outlier (AO) at time  $T$ , it is perturbed additively such that:

$$Y'_t = Y_t + \omega_A P_t^{(T)}$$

where  $\{Y_t\}$  represents the *unperturbed* process.

- ▶ So  $Y'$  is the process that may be affected by outliers, and  $Y$  is what the process would be if there were no outliers.
- ▶ To summarize:  $Y'_T = Y_T + \omega_A$ , but  $Y'_t = Y_t$  for  $t \neq T$ , so the AO only affects the series at time  $T$ .

# Modeling an Innovative Outlier

- ▶ If a time series has an innovative outlier (IO) at time  $T$ , the error at time  $T$  is perturbed such that:  $e'_t = e_t + \omega_I$  at time  $T$  but  $e'_t = e_t$  when  $t \neq T$ .
- ▶ Recall that a stationary model can be written in general linear process form as a linear combination of earlier error terms:

$$Y_t = e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \dots$$

- ▶ So an IO will affect the process not only at time  $T$ , but at later times:

$$Y'_t = Y_t + \psi_{t-T} \omega_I$$

where this weight  $\psi_{t-T}$  is 0 when  $t < T$ , is 1 when  $t = T$ , and grows smaller as  $t$  increases past  $T$ .

- ▶ Thus the effect of the IO continues on throughout time, but it is most pronounced at time  $T$  and soon after that.

# Detecting Innovative Outliers

- ▶ The detection of an IO is based on the residuals  $a_t$ ,  $t = 1, 2, \dots, n$ .
- ▶ If the process has exactly one IO at time  $T$ , then  $a_T = \omega_I + e_T$  at that time, but for all other times,  $a_t = e_t$ .
- ▶ Since the unknown  $e_T$  is assumed to have mean zero, we can estimate  $\omega_I$  by  $\tilde{\omega}_I = a_T$ .
- ▶ Since  $\lambda_{1,T} = a_T/\sigma$  has a standard normal distribution under the null hypothesis of no outliers, we would reject this null at the 0.05 significance level and conclude there is an IO at time  $T$  if  $|a_T/\sigma| > 1.96$ .
- ▶ In practice, we must estimate  $\sigma$ , but for large samples, this does not affect the properties of the test very much.

# Checking All Observations for Innovative Outliers

- ▶ That previous decision rule assumes that we are checking whether one single observation, at some known time  $T$ , is an IO.
- ▶ In practice, we probably will not know beforehand where the IO could occur, so we will check *all* observations to see whether any are IOs.
- ▶ Since we are doing multiple simultaneous tests, we should do a Bonferroni correction to prevent false detection of IOs.
- ▶ We define the maximum absolute standardized residual to be  $\lambda_1 = \max_{1 \leq t \leq n} |a_t/\sigma|$  and we declare the observation with the largest absolute standardized residual to be an IO if  $\lambda_1$  exceeds the quantile of the standard normal distribution that cuts off area  $0.025/n$  in the upper tail.

# Properties of the Test for Innovative Outliers

- ▶ This procedure guarantees that the probability of false outlier detection of an outlier is no more than 0.05.
- ▶ But if ML is used to estimate  $\sigma$ , the power to detect an IO may be weakened, since the outlier will hinder the estimation of  $\sigma$ .
- ▶ To improve power, a robust estimator of  $\sigma$  can be used, such as  $\sqrt{2/\pi}$  times the mean absolute residual.
- ▶ If an outlier is found, it can be incorporated into the model (more on that later) and the process can be repeated to check for outliers in the revised model.

# Detecting Additive Outliers

- ▶ The detection of an AO at time  $T$  is more complicated since the residual  $a_t$  is affected by the AO at all times at or past  $T$ .
- ▶ The test statistic  $\lambda_{2,T}$  for checking for an AO at time  $T$  is a weighted function of the standardized residuals (given on page 258).
- ▶ In practice, we probably will not know beforehand where the AO could occur, so we will check *all* observations to see whether any are AOs, and we again do a Bonferroni correction to account for the simultaneous tests.

# Determining Whether an Outlier is Additive or Innovative

- ▶ We may not know beforehand whether an outlier is an IO or an AO.
- ▶ One possible rule, if we detect an outlier at time  $T$ , is to declare it an IO if  $|\lambda_{1,T}| > |\lambda_{2,T}|$  and declare it an AO otherwise.
- ▶ Once an outlier is found, we can incorporate it into the model using the `arima` or `arimax` functions.
- ▶ After the model is refit incorporating the outlier, we could repeat the check for outliers until no more outliers are detected.

## A Simulated Time Series Example with an Outlier

- ▶ In an R example, we simulate a time series following an  $ARIMA(1, 0, 1)$  model, but we include an additive outlier at time  $t = 10$  (see plot of data).
- ▶ The ACF of this data set shows a damped sine wave pattern, and the PACF cuts off after lag 1, so we tentatively specify an  $AR(1)$  model.
- ▶ The `detectAO` and `detectIO` functions were used on the fitted  $AR(1)$  object, and observations 9, 10, 11 were marked as potential AOs, and observations 10 and 11 were marked as potential IOs.
- ▶ Of all these, the test statistic identifying observation 10 as an AO had the largest absolute value, so we tentatively mark observation 10 as an AO and incorporate this into the model via a dummy variable in the `xreg` argument.

## More on the Simulated Example with an Outlier

- ▶ Once this AO at observation 10 is incorporated into the model, the estimated coefficients do change a good bit, and no further outliers are detected.
- ▶ But the diagnostics on the revised model shows a large lag-1 autocorrelation in the ACF plot of the residuals.
- ▶ This leads us to include an  $MA(1)$  term in the model, making the model an  $ARIMA(1, 0, 1) + AO$  at  $T = 10$ .
- ▶ This final model shows no further outliers, and the residuals resemble white noise without excessively large autocorrelations.

## A Real Time Series Example with an Outlier

- ▶ We now revisit the seasonal  $ARIMA(0, 1, 1) \times (0, 1, 1)_{12}$  model that we used for the `co2` data set in the `TSA` package in Chapter 10.
- ▶ The standardized residuals plot shows one quite large residual (for September 1998).
- ▶ The `detectAO` and `detectIO` functions mark observation 57 (which is September 1998) as a potential IO.
- ▶ We incorporate this IO into the model using the `io` argument in the `arimax` function.
- ▶ In the revised model, the MA and seasonal MA coefficients do not change much, but the IO effect is highly significant and the AIC is improved.
- ▶ The diagnostics show no problem, so we may use this as our final model.

## Returning to the Air Miles Data

- ▶ Recall that for the airmiles model, we chose a seasonal  $ARIMA(0, 1, 1) \times (0, 1, 0)_{12}$  model for  $N_t$  and also incorporate our intervention model for  $m_t$ , to account for the intervention due to the September 2001 event.
- ▶ But from the standardized residual plot, we noted a serious outlier, and we now use our outlier techniques to handle it.
- ▶ The `detectAO` function marks observation 25 as an AO, but due to the seasonal differencing and first differencing, we see that the observations 12 and 13 are really the ones that break from the pattern we would expect from their month and year (see plot of logged airmiles data).

- ▶ The December 2002 value also breaks from the pattern to some extent.
- ▶ These three points can be marked as AOs with the `xreg` arguments, and the nature of the intervention can be specified with the `xtransf` and `transfer` arguments.
- ▶ The fit of the model seems good (see plot of observed data with fitted values).

# The Estimated Intervention Effect for the Air Miles Data

- ▶ Recall the model that was used for the intervention effect:

$$m_t = \omega_0 P_t^{(T)} + \frac{\omega_1}{1 - \omega_2 B} P_t^{(T)}$$

- ▶ Under this model, the immediate change to the mean response function is  $\omega_0 + \omega_1$ , while the effect  $k$  time units after  $T$  is  $\omega_1(\omega_2)^k$ .
- ▶ Using our estimates of  $\omega_0, \omega_1, \omega_2$ , we see that the mean log-airmiles immediately changed by  $-0.0949 - 0.2715$ .
- ▶ The immediate percent reduction, in terms of airmiles, is  $[1 - \exp(-0.0949 - 0.2715)] \times 100\% = 31\%$ .
- ▶ The change in log-airmiles  $k$  months after September 2001 is  $-0.2715(0.8139)^k$ .
- ▶ See the example R code for a plot of this intervention effect over time.

# A Time Series Regression Example with an Outlier

- ▶ We now examine a public transportation time series regression example.
- ▶ We wish to use the monthly gasoline prices in Denver (from August 2000 to March 2006) to predict the monthly number of boardings on public transportation (both variables are log-transformed to account for right-skewness).
- ▶ Both series show increasing trend, especially the gas-price series, and the boardings series appears seasonal.
- ▶ An  $ARIMA(2, 1, 0)$  model was specified for the logged price data, based on its sample ACF and PACF.
- ▶ The sample CCF on prewhitened data based on this model shows significant positive contemporaneous cross-correlation (makes sense), and a significant CCF value at lag 15 (strange; do boardings lead price by 15 months? Probably not).

## More on Regression Example with an Outlier

- ▶ An OLS fit was done, and based on the residuals, a seasonal  $ARIMA(2, 0, 0) \times (1, 0, 0)_{12}$  model for the noise process was specified.
- ▶ Upon fitting this regression model, the AR(2) coefficient was not significant, so we removed it, and this improved the model.
- ▶ A plot of the standardized residuals shows a large positive residual for March 2004, and somewhat less notably, a couple of sizable negative residuals.
- ▶ The `detectAO` and `detectIO` functions find an AO at time 32 (March 2003) and an IO at time 44 (March 2004).
- ▶ Since the March 2003 outlier has a larger absolute test statistic, we incorporate this AO into the model.

## Completing the Boarding Regression Example

- ▶ The ACF of the residuals showed a significant lag-3 autocorrelation so the model for the noise process was altered to a  $ARIMA(1, 0, 3) \times (1, 0, 0)_{12}$  model.
- ▶ The  $MA(1)$  and  $MA(2)$  terms' coefficients were not significant, so they were fixed to be zero (not appearing in the model).
- ▶ After those alterations, no more outliers were detected, and the model diagnostics showed no problems.
- ▶ Note: In the model with the outlier unaccounted for, logged price was NOT a significant predictor of logged boardings.
- ▶ But when we account for the outlier, we see that logged gas price DOES have a significant (positive) effect on logged boardings.
- ▶ It turns out that in March 2003, there was a major snowstorm that shut down Denver and altered the apparent relationship between gas price and boardings.