Syllabus.

- Chapter 1. Concepts in time series.
- Chapter 2. Univariate ARIMA models.
- Chapter 3. Model fitting and checking.
- Chapter 4. Prediction and model selection.
- Chapter 5. Outliers and influential observations.
- Chapter 6. Heterocedastic models.
- Chapter 7. Multivariate time series.

Chapter 5. Outliers and influential observations.

- 5.1.Types of outliers in time series.
- 5.2.Procedures for outlier identification and estimation.
- 5.3. Multiple outliers.
- 5.4.Influential observations.
- 5.5. Missing-value estimation.
- 5.6 Forecasting with outliers.

General assumptions.

- The outliers happen on a time series which can be modelled with the ARIMA (p,d,q):

$$\Phi(B)\nabla^d X_t = \Theta(B)a_t$$

- The model could be written in the AR or MA forms as follows:

$$\pi(B)X_t = \frac{\Phi(B)\nabla^d}{\Theta(B)}X_t = a_t \quad or \quad X_t = \frac{\Theta(B)}{\Phi(B)\nabla^d}a_t = \psi(B)a_t$$

Where a_t is Gaussian white noise, $\Phi(B)$ and $\Theta(B)$ have all the roots outside the unit circle and have no roots in common.

Additive outlier (AO). Definition.

Additive Outlier (AO). Is an isolated spike that corresponds to an external error or exogenous change of the time series at a particular time point.

$$Y_t = \left\{ \begin{array}{ll} X_t & t \neq \tau \\ X_t + \omega_A & t = \tau \end{array} \right\}$$

Where Y_t is the contaminated time series, τ is the time at which the outlier occurs and ω_A is the magnitude of the outlier.

An AO can have serious effects on the properties of the observed time series:

- It will affect the estimated residuals.
- It will affect the estimates of the parameter values.

Additive outlier (AO). Effect on the time series.

An alternative using the AR representation is:

$$Y_t = \omega_A I_t^{\tau} + \psi(B) a_t$$

where $I_t^{\tau}$is an indicator variable which is zero for all t except at time $t = \tau$.The expression above implies that an AO affects only one period of the contaminated series, does not matter the ARIMA process behind.

Additive outlier (AO).Effect on the estimated residuals.

Lets define the estimated residuals of the contaminated series as:

$$e_t = \pi(B)(X_t + \omega_A I_t^{\tau})$$

whereas, in the case of a non contaminated series, we have

$$a_t = \pi(B)X_t$$

The relation between them is, therefore,

$$e_t = a_t + \pi(B)\omega_A I_t^{\tau}$$

So the effect of the AO on the residuals depends on the π weights. In a pure AR model, p residuals will be affected.

Additive outlier (AO). Effect on the estimated parameters.

Consider a simple AR(1). The OLS estimate is :

$$\hat{\phi}_1 = \frac{\sum X_t X_{t-1}}{\sum X_t^2}$$

In a case of a time series contaminated with an AO we have:

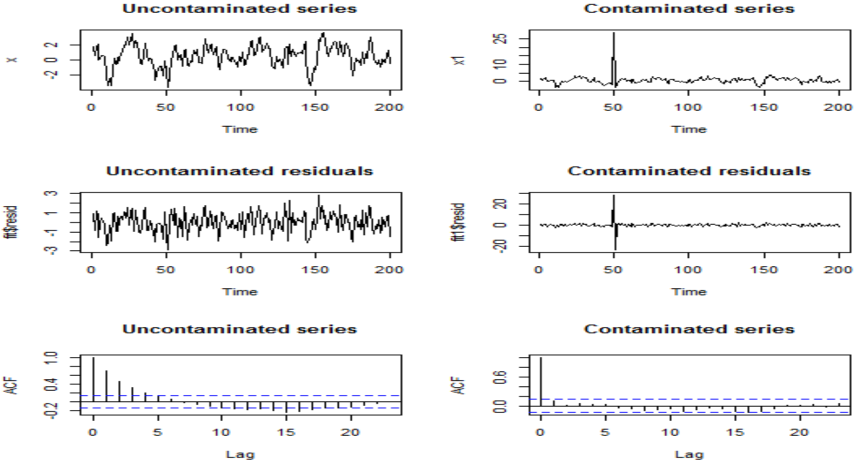
$$\hat{\phi}_1 = \frac{\sum Y_t Y_{t-1}}{\sum Y_t^2}$$

It can be shown that,

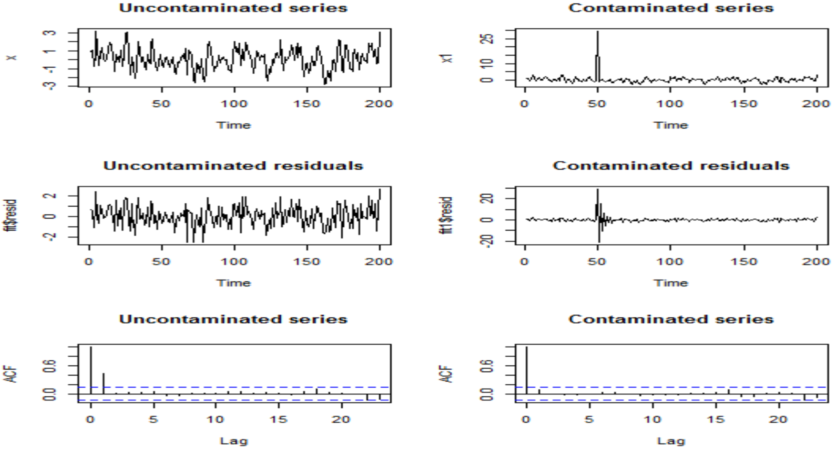
$$\omega_A \to \infty \Rightarrow \hat{\phi}_1 \to 0$$

In general, a large AO will push all the autocorrelation coefficients towards zero.

AO in a AR(1) with $\phi = 0{,}7$ and $\omega_A = 30$.



AO in a MA(1) with $\theta = 0{,}7$ and $\omega_A = 30$.



Innnovative (IO). Definition.

Innovative Outlier (IO). Isolated spike on the innovations series that correspond to an internal error or endogenous change of the time series at a particular time point.

$$Y_t = \left\{ \begin{array}{lr} X_t & t < \tau \\ X_t + \omega_I \psi_j & t = \tau + j \end{array} \right\}$$

Where Y_t is the contaminated time series, τ is the time at which the outlier occurs, ω_I is the magnitude of the outlier and ψ_j are the weights of $\psi(B) = \frac{\Theta(B)}{\Phi(B)\nabla^d}$.

Innovative outlier (IO). Effect on the time series.

An alternative using the MA representation is:

$$Y_t = \psi(B)(a_t + \omega_I I_t^\tau) = X_t + \psi(B)\omega_I I_t^\tau$$

where I_t^τ is an indicator variable which is zero for all t except at time $t = \tau$.The expression above implies that an IO affects several periods of the contaminated series with weigths depending on $\psi(B)$.

Innovative outlier (IO).Effect on the estimated residuals.

Multiplying by $\pi(B)$, we obtain the estimated residuals of the contaminated series as:

$$e_t = a_t + \omega_A I_t^{\tau}$$

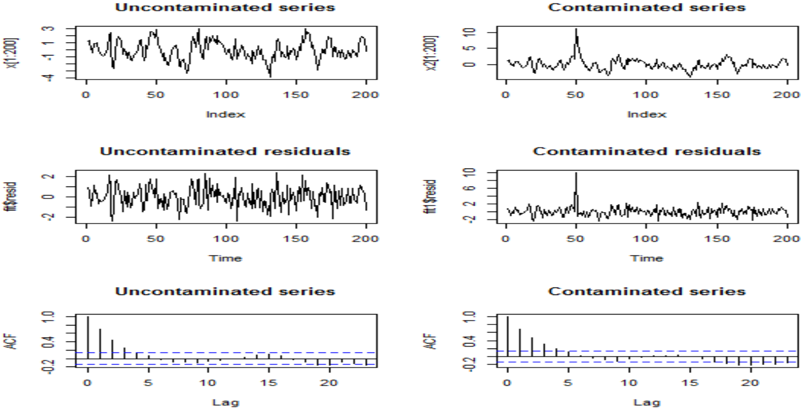
So the effect of an IO on the residuals only last one period and it is independent of the ARIMA model.

Effect on the estimated parameters: In a simple AR(1),

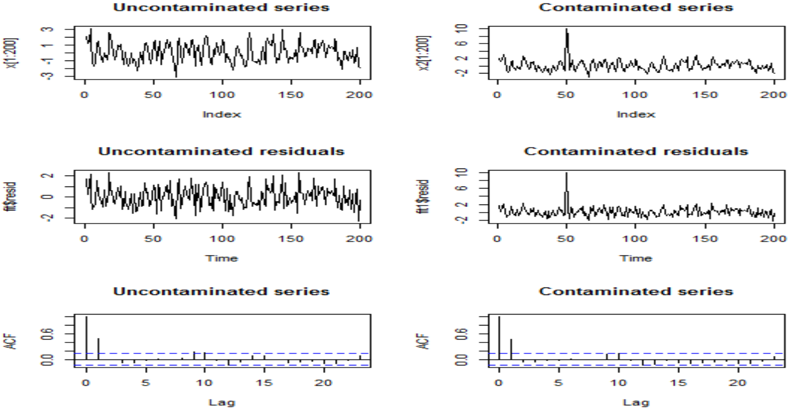
$$\omega_I \to \infty \Rightarrow \hat{\phi}_1 \to \phi$$

In general,for large samples, its effect can be neglected.

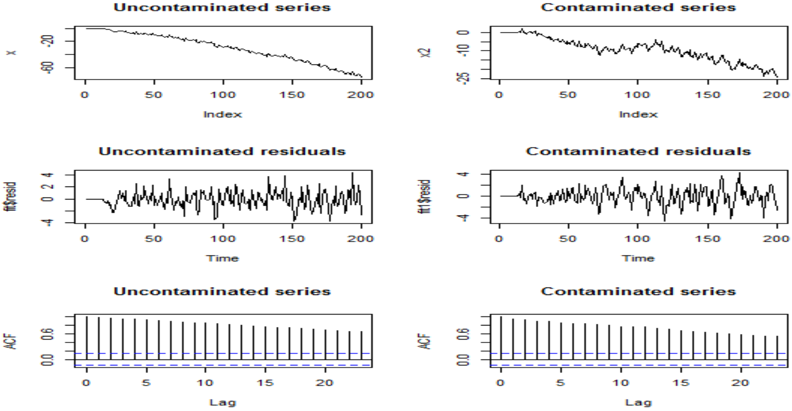
IO in a AR(1) with $\phi = 0{,}7$ and $\omega_I = 10$.



IO in a MA(1) with $\theta = 0{,}7$ and $\omega_I = 10$.



IO in a ARIMA(0,1,1)(0,1,1) with and $\omega_I = 3$.



Level shift outlier (LS). Definition.

Levl shift Outlier (LS). Corresponds to a modification of the global mean or level of the process starting from a specific point and continuing until the end of the observed sample.

$$Y_t = \left\{ \begin{array}{ll} X_t & t \neq \tau \\ X_t + \omega_L & t \geq \tau \end{array} \right\}$$

Where Y_t is the contaminated time series, τ is the time at which the outlier occurs and ω_L is the magnitude of the outlier.

A level shift can be seen as a sequence of AO's with the same magnitude.

Level shift (LS). Effect on the time series.

An alternative expression using the AR representation is:

$$Y_t = \omega_L S_t^\tau + \psi(B) a_t = \omega_L \left(\frac{1}{1-B} \right) I_t^\tau + \psi(B) a_t$$

where S_t^τ is an indicator variable which is zero for all $t < \tau$ and 1 afterwards. The expression above implies that an LS affects all the observations after the contamination.

Levl shift (LS).Effect on the estimated residuals.

Lets define the estimated residuals of the contaminated series as:

$$e_t = \pi(B)(X_t + \omega_A S_t^\tau)$$

The relation with the uncontaminated residuals is,assuming known parameters:

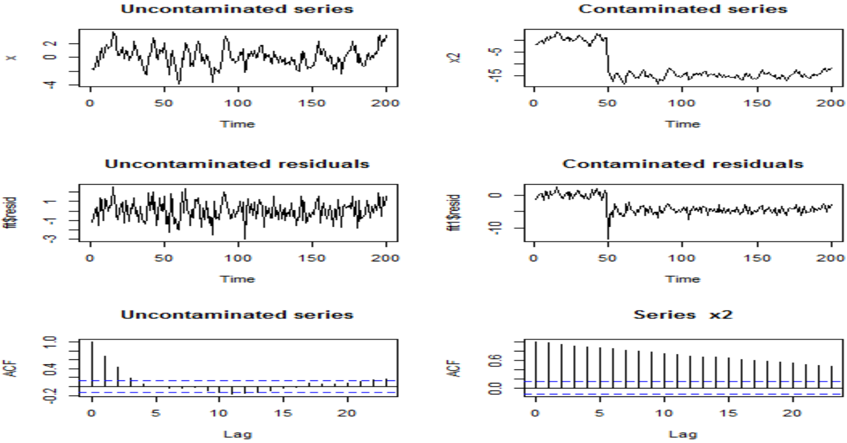
$$e_t = a_t + \pi(B)\omega_A S_t^\tau$$

All residuals after the LS can be affected and, therefore, the effect of the LS on the residuals and on the estimated parameters can be strong. The effect of a LS depends on (1) the ARIMA model and (2) the distance between the time of occurrence and the end of the observed sample. For $n-\tau$ not too small,

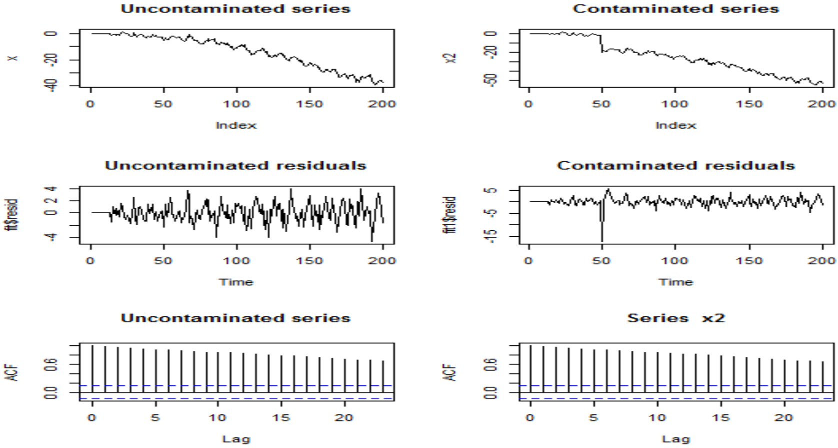
$$\omega_L \to \infty \Rightarrow \hat{\rho}(1) \to 1$$

A large LS will push the lag 1 autocorrelation coefficients towards one, indicating non-stationarity.

LS in an AR(1) with $\phi = 0{,}7$ and $\omega_L = 15$.



LS in a SARIMA(0,1,1)(0,1,1).



Outliers and intervention analysis.

The outliers we have seen can be considered as particular cases of interventions in a time series:

$$Y_t = \omega V(B) I_t^{\tau} + \psi(B) a_t$$

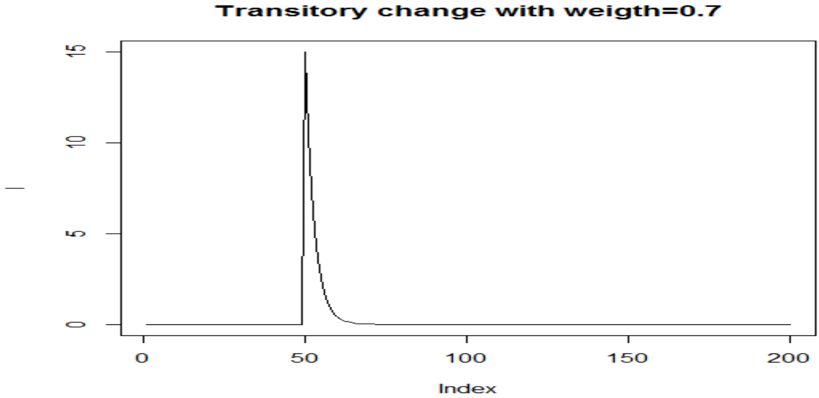
where $V(B)$ is the transfer function of the intervention.

Among others:

- AO: $V(B) = 1$.
- IO: $V(B) = 1/\pi(B)$.
- LS: $V(B) = 1/(1 - B)$.
- but also...

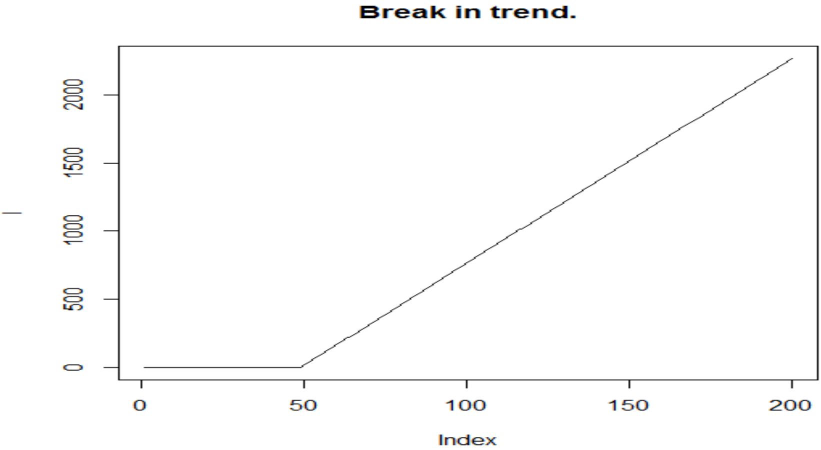
Outliers and intervention analysis.

Transitory change (TC): $V(B) = 1/(1-\delta B)$



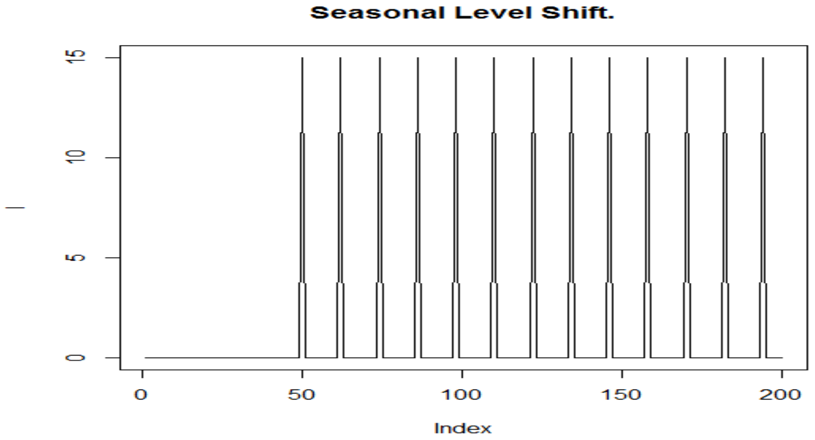
Outliers and intervention analysis.

Break in trend: $V(B) = 1/(1-B)^2$



Outliers and intervention analysis.

Seasonal Level Shift: $V(B) = 1/(1 - B_{12})$



Procedures for outlier identification and estimation.

In order to eliminate the effect of an outlier in a given time series it is necessary to:

- Detect the time at which the outlier happens.
- Identify the type.
- Remove its effect by estimating a model in which the outlier is incorporated.

Overview of an automatic procedure.

- Step 1: At each point, we analyze what will be the most likely type of outlier (likelihood ratio for the considered types).
- Step 2: We choose as the candidate outlier timepoint the one that has the smallest p value and as outlier type the corresponding outlier effect.
- Step 3: Fit the appropiate intervention model to remove the outlier.

One outlier with known parameters

- Consider the intervention model:

  $$Y_t = V(B)\omega I_t^{\tau} + X_t$$

  This model is a multiple regression model with autocorrelated residuals that can be estimated by Generalized Least Squares (GLS).
- To test the null hypothesis that the observation at time $t = \tau$ is not an outlier we use the likehood ratio:

  $$\lambda = \frac{\hat{\omega}}{\sqrt{var(\hat{\omega})}}$$

  which, assuming known parameters, follows a $N(0, 1)$ distribution.

One outlier with known parameters

- If the outlier type is unknown we should use the statistic:

$$\eta(\tau, I) = max_i|\lambda_i| \quad i = AO, IO, LS, \ldots$$

- And, if the period is also unknown, the statistic:

$$\eta = max_t(\tau_t, I_t) \quad t = 1, \ldots, n$$

- Criterion: An AO at time $t = \tau$ will be detected if

$$\eta = |\lambda_{A,\tau}| > C$$

where C is a predetermined constant (3-4).

Multiple outliers, unknown parameters.

- Masking effects:The method presented before works very well when the series has a single outlier or a few isolated outliers. However, sometimes, the series is subject to patches of outliers that may produce masking.
- The generalization of the model for multiple outliers is,

$$Y_t = \sum_{j=1}^{k} \omega_j V_{\tau,j}(B) I_t^{\tau,j} + X_t$$

- Using the model above, an iterative procedure can solve the problem of masking (Chen and Liu, 1993). Implemented in R with the library *tsoutlier.*

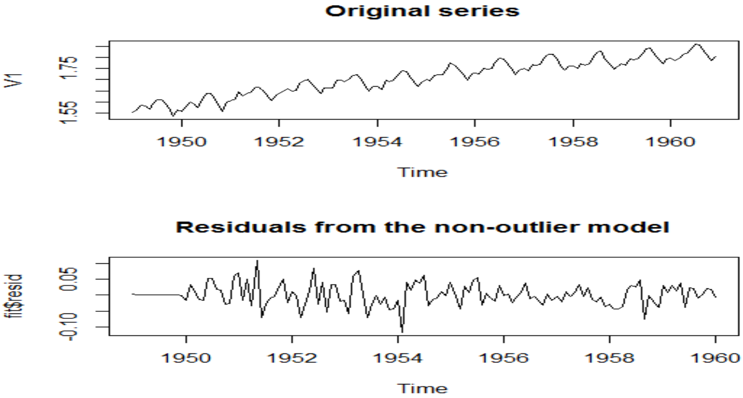
Multiple outliers, unknown parameters.

- Step 1: The model is estimated assuming no outliers.
- Step 2: Compute the likelihood ratio for any point and any outlier type.
- Step 3: If an outlier is detected, correct its effect on the residuals.
- Step 4: Compute again the likelihood ratio for the new residuals. Repeat until no new outliers can be identified.
- Step 5: Jointly estimate the sizes of the identified outleirs and te parameters using the intervention model,

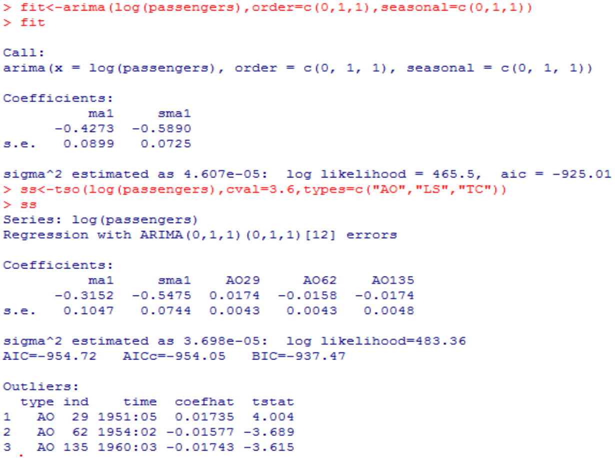
$$Y_t = \sum_{j=1}^{k} \omega_j V_{\tau,j}(B) I_t^{\tau,j} + X_t$$

- Step 6: Eliminate the non-significative outliers and re-start the procedure.

Performance of the procedure for passengers series.



Performance of the procedure for passengers series.



```
> fit<-arima(log(passengers),order=c(0,1,1),seasonal=c(0,1,1))
> fit

Call:
arima(x = log(passengers), order = c(0, 1, 1), seasonal = c(0, 1, 1))

Coefficients:
          ma1     sma1
      -0.4273  -0.5890
s.e.   0.0899   0.0725

sigma^2 estimated as 4.607e-05:  log likelihood = 465.5,  aic = -925.01
> ss<-tso(log(passengers),cval=3.6,types=c("AO","LS","TC"))
> ss
Series: log(passengers)
Regression with ARIMA(0,1,1)(0,1,1)[12] errors

Coefficients:
          ma1     sma1    AO29     AO62    AO135
      -0.3152  -0.5475  0.0174  -0.0158  -0.0174
s.e.   0.1047   0.0744  0.0043   0.0043   0.0048

sigma^2 estimated as 3.698e-05:  log likelihood=483.36
AIC=-954.72   AICc=-954.05   BIC=-937.47

Outliers:
  type ind    time  coefhat  tstat
1   AO  29 1951:05  0.01735  4.004
2   AO  62 1954:02 -0.01577 -3.689
3   AO 135 1960:03 -0.01743 -3.615
```

Influential observations and outliers.

A simple way to measure the influence of an observation is to assume that it is an AO. The Mahalanobis distance between the vector of MLE forecasts assuming no outliers in the series and the same vector obtained assuming an AO at this point is a measure of influence:

$$D_{\hat{X}}(\tau) = \frac{(\hat{X} - \hat{X}_{\tau}^{AO})'(\hat{X} - \hat{X}_{\tau}^{AO})}{h\hat{\sigma}_a^2}$$

where h is the number of parameters in the ARIMA model, $\hat{X}$ is the vector of h forecasts assuming no outliers, and $\hat{X}_{\tau}^{AO}$ is the vector of forcats assuming an AO.

Missing observations and outliers.

- The study of AO's and influential observations in time series is closely related to missing-value analysis because an outlier implies that the true value at this point is not observed.
- Therefore, a missing value (or a patch of missing values) could be interpolated assuming an AO(or a patch of AO's).

Forecasting with outliers.

- It has often been stressed that the prediction intervals computed from ARIMA models are too short, that is, forecast are out of the bounds more often than it will be expected.
- Sources of uncertainty:
  - noise of the model,
  - unknown parameters,
  - model uncertainty,
  - ...
  - and possible outliers in the forecast horizon.

Forecasting with outliers.

- A possibility to take account of those could be to consider the model:

$$Y_{n+k} = \sum_{i=1}^{4} \omega_{i,n+k} V_i(B) u_{i,n+k} + X_{n+k}$$

  where the terms $V_i(B)$ correspond to the four outliers types and ω_i are random variables with some specified distribution.

- The variables $u_{i,n+k}$ are Bernuilli variables indicating the probability of each outlier type, so that,

$$P(u_{i,n+k} = 1) = \alpha_i; \quad P(u_{i,n+k} = 0) = 1 - \alpha_i$$

  and $\sum \alpha_i = \alpha$ gives the probability of any outlier type affecting the forecast.