Time Series Regression

Ian McLeod

November 5, 2017

Time Series Data and Regression

Millions of socio-economic time series are available from the Statistics Canada CANSIM database and tens of thousands of financial time series are available in R using the **quantmod** package (CRAN). Millions of climate, meteorological, hydrological, environmental and other types of scientific time series are also available.

In an MMSc. thesis my student L. King developed a comprehensive simulation method based on 30 years of daily meteorological observations on 7 variables (including precipation, daily max temperature, etc.) at 27 different locations in the Thames Valley watershed.

The multiple linear regression with some adjustments may be used to describe the relationship between time series. We start with the univariate model where we have one output variable $y_t, t = 1, \ldots, n$ observed at n successive times. In many cases these observation time correspond to actual time units such as seconds, days, quarters, years, etc. We assume there are p dependent or input variabes, $x_{t,j}, t = 1, \ldots, n \;\; j = 1, \ldots, p$. Sometimes lagged values of the input variables may also be included. The model may be written,

$$y_t = \beta_0 + \beta_1 x_{t,1} + \ldots + \beta_p x_{t,p} + e_t,$$

where $\beta_j, j = 1, \ldots, p$ are parameters and $e_t, t = 1, \ldots, n$ is the error term. The error is assumed stochastic with mean zero and constant variance σ^2. Variations of this model can be used in a wide variety of applications. Often more specific assumptions are needed to describe the error of noise component e_t.

Models for Time Series Regression

- OLS
- Regression with autocorrelated error
- Dynamic regression with autocorrelated error
- Regression with ARIMA-GARCH errors

Ordinary Least Squares (OLS)

OLS models assume that $e_t \sim IID(0, \sigma^2)$ or perhaps $e_t \sim NID(0, \sigma^2)$ where IID/NID are respectively independent/normal identically distributed with mean 0 and variance σ^2.

Advertising Example

A company seeks to determine the optimal mix of expenditures on advertising to maximize sales. This dataset was obtained from the book homepage for *Introduction to Statistical Learning and is discussed in their textbook. A PDF copy of this textbook is also available on the book homepage.

Three advertising expenditures on **Newspaper**, **Radio** and **TV** are varied over 200 weekly periods. Time series plots shown in Figure 1 suggest that there is little or no trends so perhaps all the series are uncorrelated. Actually, I surmise that this data is completely artifical since real data of this nature is of crucial importance to a business and is not usually made publically available. Figure 1 supports my conjecture since the complete lack of time series structure either in trends or seasonality is unusual although not impossible.

```
#ads <- read.csv("D:/Dropbox/R/2017/3859/data/advertising.csv")
ads <- cbind(week=1:nrow(ads), ads)
adsL <- gather(ads, key=media, value=expenditure, -week)
ggplot(adsL, aes(x=week, y=expenditure)) +
 #geom_point(color="blue") +
 geom_line() +
 facet_wrap(~media, scales="free_y")
```

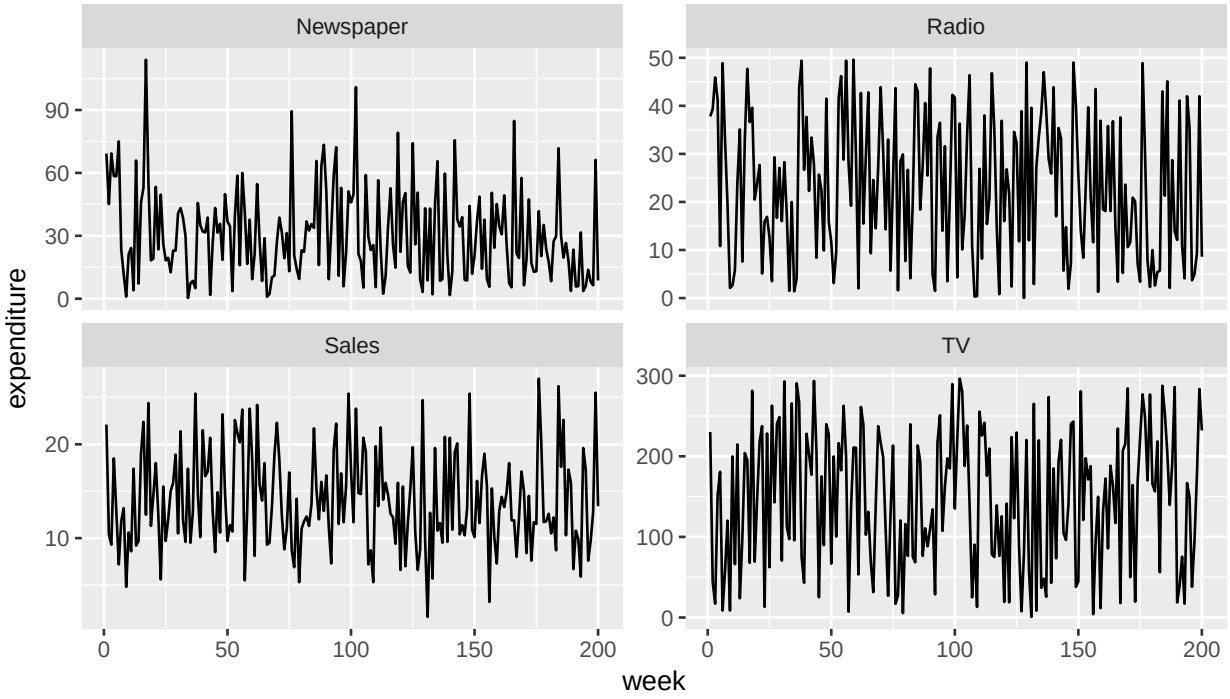


Figure 1: Time series plots of the Advertising dataset

Dependence in an observed time series is often assessed by its sample autocorrelation function (SACF) defined by,

$$r_k = \frac{\sum(z_t - \bar{z})(z_{t-k} - \bar{z})}{\sum(z_t - \bar{z})^2}$$

where $k = 1, ..., M$ where M is the maximum lag of interest. For most time series lag one is the most important lag since usually observations closer together in time are more highly correlated. In selecting a time series ARIMA model we usually take M between 15 and 50 depending on the series length n.

```
ggacf <- function(z) {
 bacf <<- acf(z, plot = FALSE)
 bacfdf <<- with(bacf, data.frame(lag=c(lag)[-1], acf=c(acf)[-1]))
 ggplot(data = bacfdf, mapping = aes(x = lag, y = acf)) +
  geom_hline(aes(yintercept = 0)) +
  geom_segment(mapping = aes(xend = lag, yend = 0), size=2) +
  geom_hline(aes(yintercept = 1.96/sqrt(length(z))), col="red", size=2) +
  geom_hline(aes(yintercept = -1.96/sqrt(length(z))), col="red", size=2)
}
ggacf(ads$Sales)
```

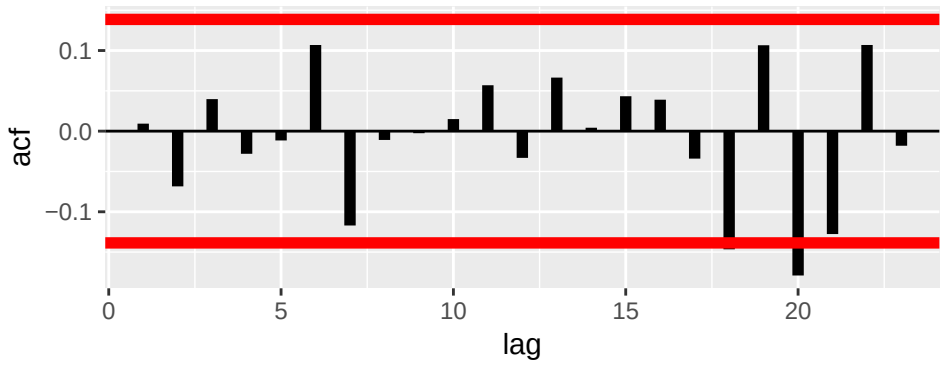


Figure 2: SACF for Sales

Asymptotically, if the IID assumption holds $\sqrt{n}r_k \rightarrow \text{NID}(0,1)$ for $k = 1, \ldots, M$. The benchmark 95% confidence limits are shown in red. This is only an informal check since due to randomness we expect 1 in 20 to exceed the limits.

Next we fit the OLS model using R's **lm()** function.

```
ans <- lm(Sales ~ TV+Radio+Newspaper, data=ads)
RSq <- 1-(sum(resid(ans)^2))/with(ads, sum((Sales - mean(Sales))^2))
out <- xtable(ans, caption="OLS of Sales on Advertising Variables")
print(out, type="latex")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9389	0.3119	9.42	0.0000
TV	0.0458	0.0014	32.81	0.0000
Radio	0.1885	0.0086	21.89	0.0000
Newspaper	-0.0010	0.0059	-0.18	0.8599

Table 1: OLS of Sales on Advertising Variables

From Table 1, **Newspaper** is not significant at 10% so it can be dropped from the model. Refitting the model does not change the other estimates very much as can be seen from Table 2.

```
ans <- lm(Sales ~ TV+Radio, data=ads)
RSq <- 1-(sum(resid(ans)^2))/with(ads, sum((Sales - mean(Sales))^2))
out <- xtable(ans, caption="OLS of Sales on TV and Radio")
print(out, type="latex")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9211	0.2945	9.92	0.0000
TV	0.0458	0.0014	32.91	0.0000
Radio	0.1880	0.0080	23.38	0.0000

Table 2: OLS of Sales on TV and Radio

In this model $R^2 = 90\%$ so the model may be useful provided it passes all diagnostic checks. Since we are dealing time series data, I recommend using R's function **tsdiag** to check the residuals for lack of independence.

```
tsdiag(arima(resid(ans)))
```

The basic time series diagnostic plot in Figure 3 is comprised of three panels:

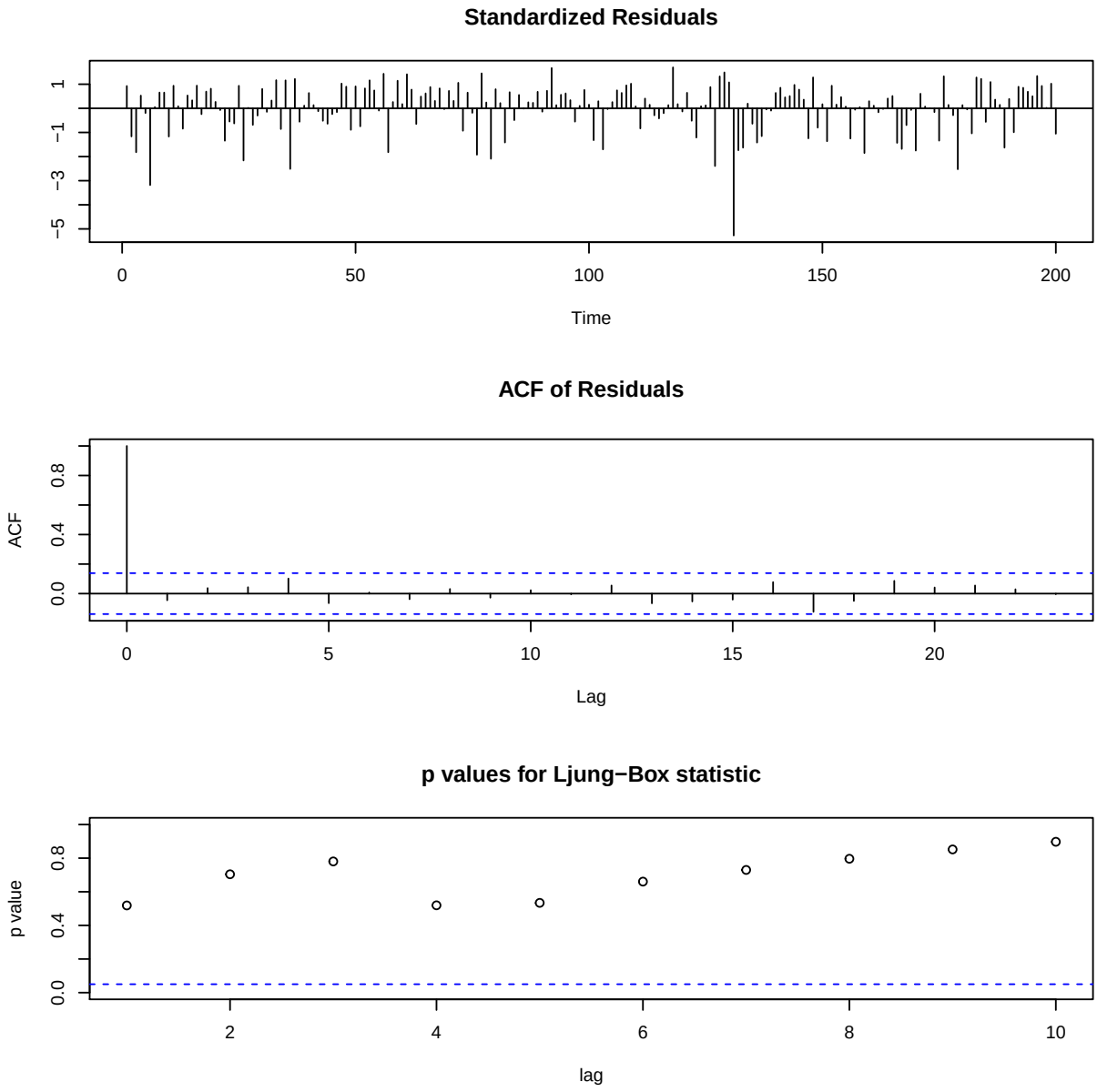


Figure 3: Time Series Diagnostic Checks for Residuals in OLS Model

1. time series plot of residuals
2. SACF plot of the residuals
3. P-value plot of the portmanteau test statistic $Q_m, m = 1, \ldots, M$

In the time series plot we look for any systematic departures from randomness such as trend, seasonality, clustering and outliers. *Volatility clustering* is an especially important departure that often arises when long financial time series of daily returns are used. In the present case, the time series panel merely suggests that there is one outlier relative to the normal distribution assumption since we see that at \$t=131 the value of the standard residual is -5.3. This is of little account since the sample size $n = 200$ is quite large and the IID assumption is not violated by this outlier.

The SACF plot is used to detect if there is strong autocorrelation present. Positive autocorrelation in the residuals occurs frequently when OLS models are fitted to time series. It is a common cause of *spurious* or *nonsense* correlation in time series regression. The value of the lag-one autocorrelation is of special interest because we normally expect that if there is correlation it will be largest at lag one. The seond panel does not indicate any departure from the IID assumption.

The third panel presents another informal diagnostic check for autocorrelation. Under the IID assumption the portmanteau test statistic Q_m,

$$Q_m = n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{n-k}$$

is approximately χ^2 distributed on m degrees of freedom. The plot shows the p-value of this test for $m = 1, \ldots, M$, where M may be chosen by the software or specified as an argument to **tsdiag()**.

A further type of major violation of the IID assumption is caused by clustering of volatility. The presence of such heteroscedasticity may also cause spurious inferences to be made. Most frequently this departure from IID occurs with long daily financial returns but some researchers have claimed to find this with economic and environmental data – I am skeptical since some of these claims are simply due model mis-specification and/or p-value hacking.

It turns out that an efficient all round test for the presence of *conditional volatility* can be obtained from the SACF of the squared residuals, $\hat{e}_t^{\,2}$ – see W. K. Li's book.

```
tsdiag(arima(resid(ans)^2))
```

There is no evidence of conditional heteroscedaticity in this dataset. Not surprising since it is artificial data!

The basic regression diagnostic checks and further modeling of this dataset will be discussed in a separate lecture note.

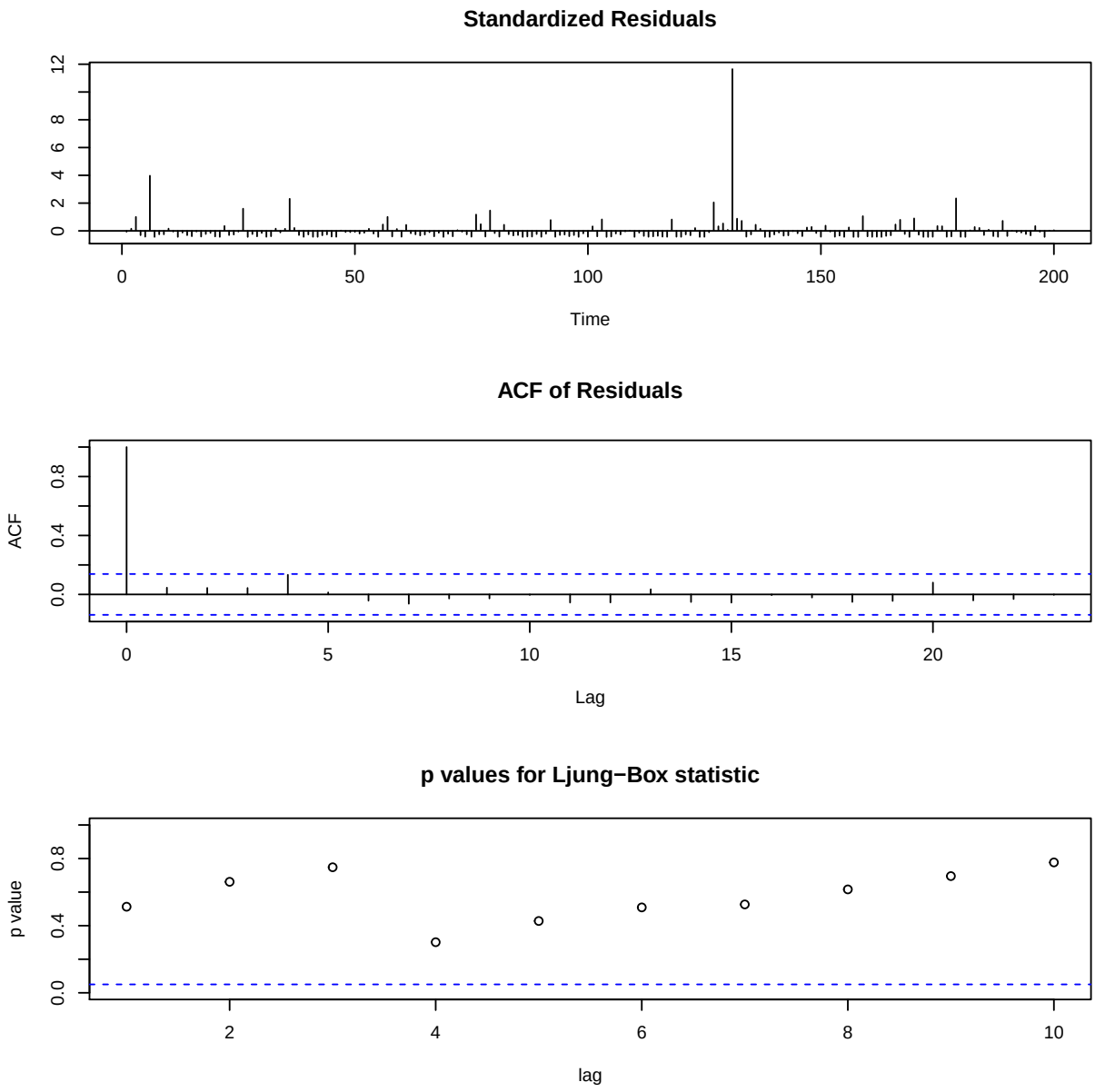


Figure 4: Squared Residuals Diagnostic Check

Regression with Autocorrelated Error

Often the error term is autocorrelated perhaps even nonstationary so the IID assumption is violated.

Famous Spurious Regression Example

In a paper published in a leading journal the authors claimed that they could predict level of the quarterly level of the UK stock market based on the UK car production six quarters prior and the index for the commodities market seven quarters previous. In other words, car production and the commodities market are leading indicators for stock market with a lags of six and seven quarters respectively. If this relationship were to continue to hold outside the of training data then one could make a lot of money!

In fact the authors used stepwise regression with a large number of possible variables resulting in an overfit model. Even more seriously the residuals in their fitted model were positively correlated so the statistical inference was completely inaccurate for this reason alone.

```
CGK <- matrix(c(z$UKCars, z$FTICom, z$FTI), ncol=3)
dimnames(CGK)[2] <- list(c("Cars", "FTI Commod.", "FTI"))
CGK <- ts(CGK, start=c(1952, 3), frequency=4)
xyplot(CGK, lwd=2, pch=16, type="o", cex=1, xlab="year")
```

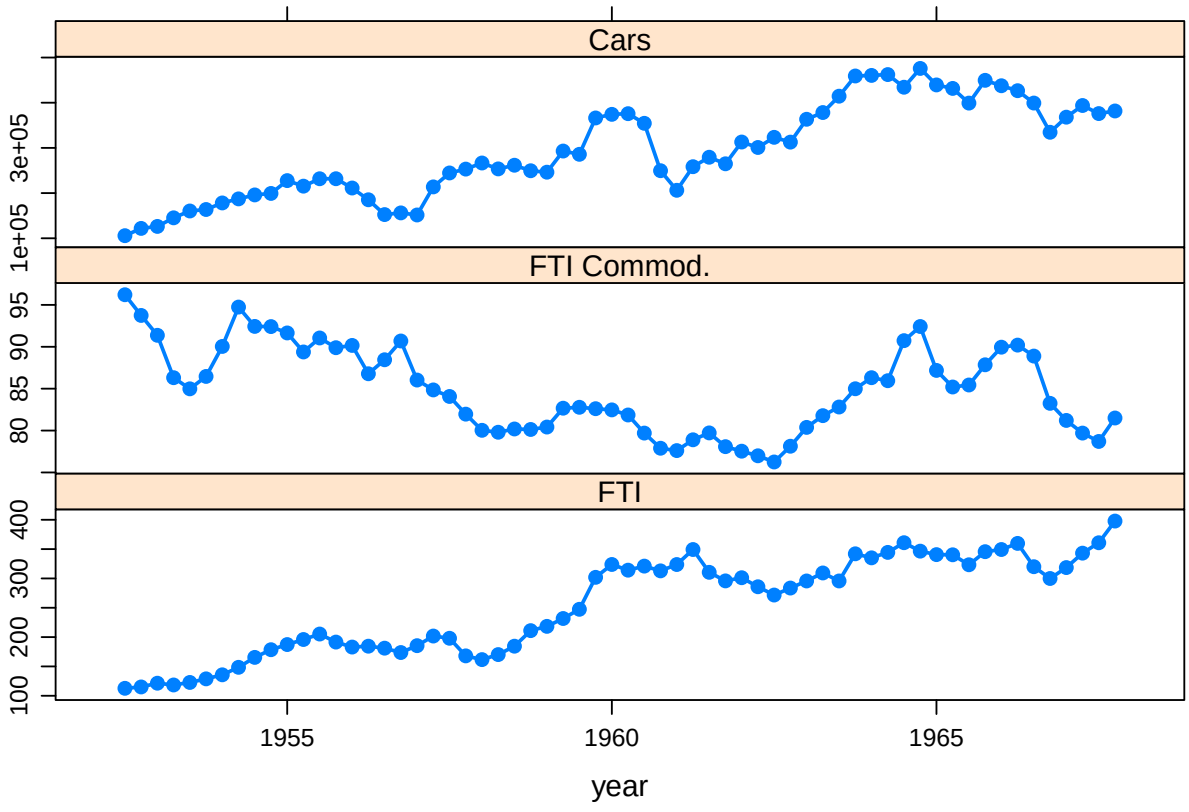


Figure 5: UK Stock Market Dataset

```
cars <- dplyr::lag(z$UKCars, 6)
FTICommod <- dplyr::lag(z$FTICom, 7)
```

```
FTI <-  z$FTI
dfCGK <- na.omit(data.frame(cars=cars, FTICommod=FTICommod, FTI=FTI))
ans <- lm(FTI ~ cars + FTICommod, data=dfCGK)
out <- xtable(ans, caption="OLS UK FTI On Lagged (6) Car Production and Lagged (7) FTI Commodities")
print(out, type="latex")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	594.5106	60.6512	9.80	0.0000
cars	0.0005	0.0000	15.10	0.0000
FTICommod	-5.5439	0.6727	-8.24	0.0000

Table 3: OLS UK FTI On Lagged (6) Car Production and Lagged (7) FTI Commodities

Examining the time series diagnostics, all three panels in Figure 6 indicates the residuals are not IID. The time series plot is too smooth. The SACF of the residuals shows a large value at lag one. And the portmaneau test shows all p-values are less than 5%.

```
lagOneACF <- c(acf(resid(ans), plot=FALSE, lag.max=1)$acf)[2]
tsdiag(arima(resid(ans)))
```

The value of the lag one residual autocorrelation is $r_1 = 0.453$. Since $n = 62$, the approximate sd under the null hypothesis of IID is about 0.13 so the result is significant at less than 0.1%.

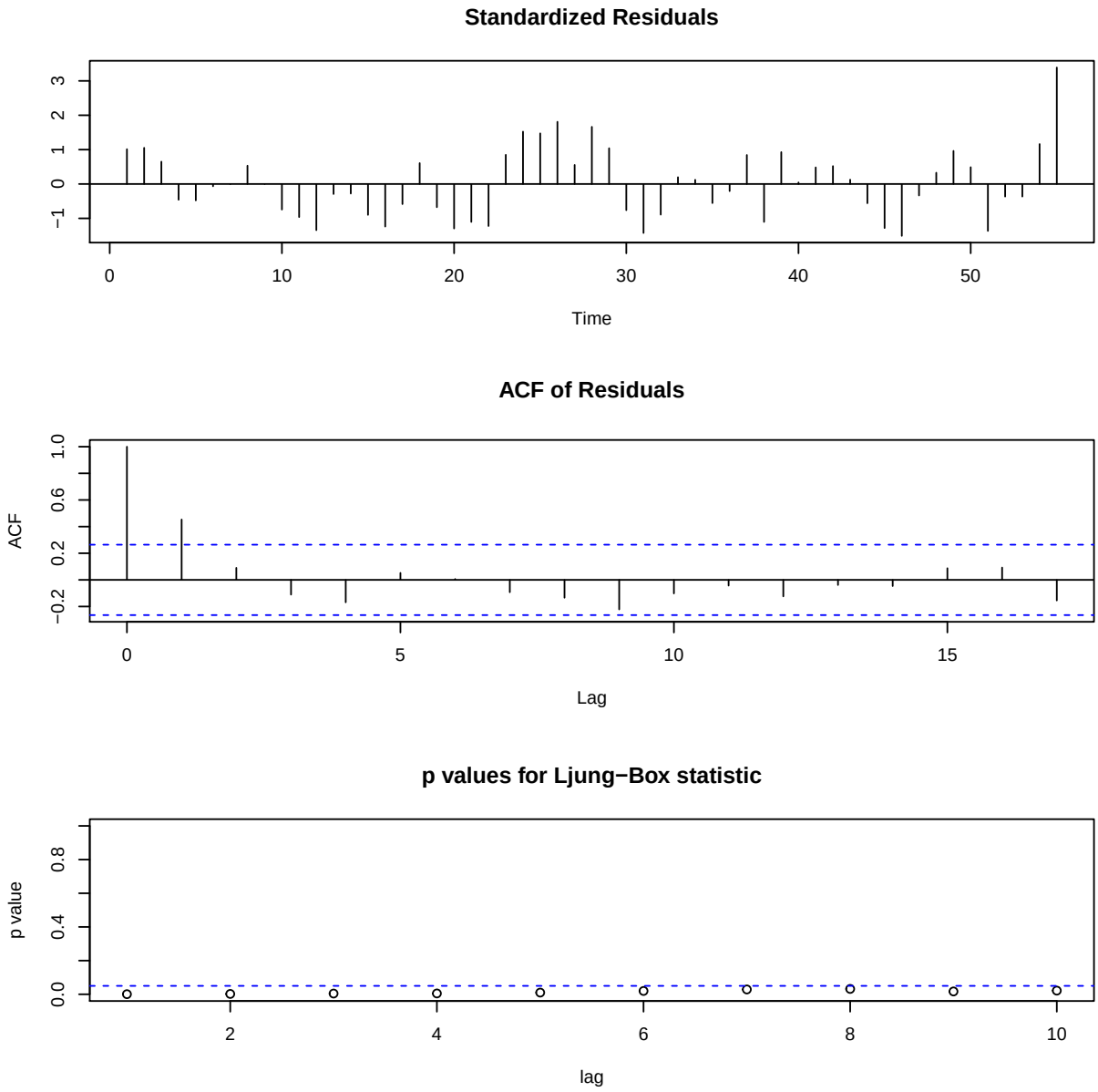


Figure 6: Time Series Diagnostic Checks for Residuals in OLS Model For UK FTI

When strong positive autocorrelation exists in the residuals, the simplest approach is to consider a model obtained by differencing all the variables in the regression.

In general, this new family of models may be written,

$$y_t = \beta_0 + \beta_1 \nabla x_{t,1} + \ldots + \beta_p \nabla x_{t,p} + \nabla e_t,$$

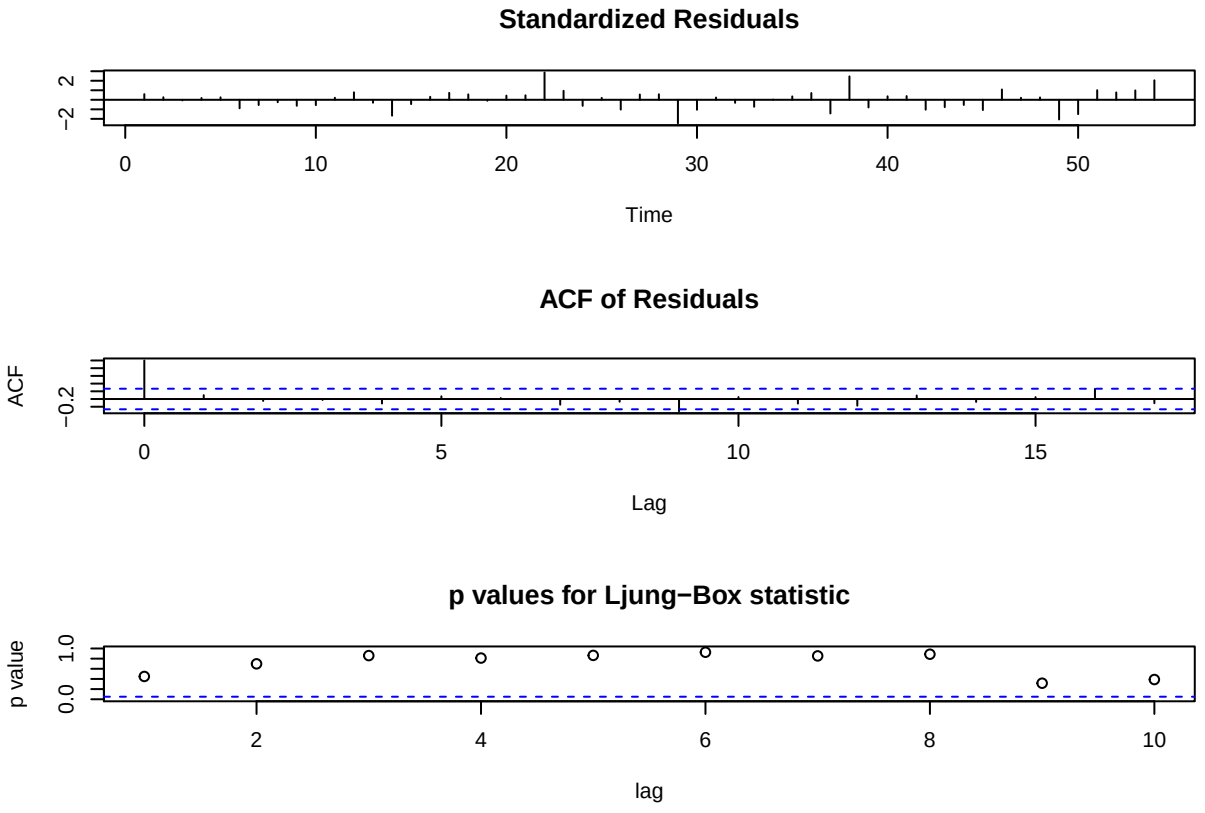
where ∇e_t is assumed IID ∇ is the first backward differencing operator so $\nabla x_{t,k} = x_{t,k} - x_{t-1,k}, k = 1, \ldots, p$. This model may be fit using **lm()**. In the fitted model none of the variables are significant at 5%.

```
dfCGK1 <- as.data.frame.matrix(diff(as.matrix.data.frame(dfCGK)))
ans <- lm(FTI ~ cars + FTICommod, data=dfCGK1)
out <- xtable(ans, caption="OLS with First Differences.")
print(out, type="latex")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.7122	2.5469	1.46	0.1511
cars	0.0001	0.0001	1.75	0.0854
FTICommod	-0.7857	1.1754	-0.67	0.5069

Table 4: OLS with First Differences.

The time series diagnostic checks do not suggest any model inadequacy.



More generally we may extend the regression by assuming that $e_t \sim \text{ARIMA}(p, d, q)$ where ARIMA denotes the family of ARIMA time series models. The model in the above example corresponds to an {ARIMA}(0,1,0). Some authors prefer to drop the intercept term when there is differencing. This more general regression with ARIMA errors may fit using **arima()**.

Intervention Analysis: Annual Nile Riverflow

```
stepIntervention <- ifelse(1870:1944 <= 1903, 0, 1)
ans <- lm(nile ~ stepIntervention)
plot(nile, xlab="year", ylab="average flow (cms)", col="blue", lwd=2)
lines(as.vector(time(nile)), fitted(ans), col=rgb(0.5,0.5,0.5, 0.5), lwd=4)
abline(v=1903, col=rgb(1,0,0,0.6), lty=2, lwd=3)
```

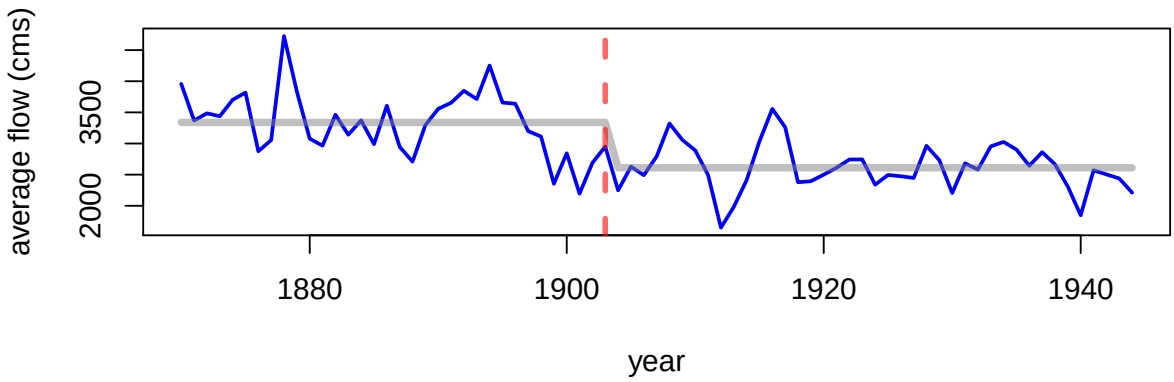


Figure 7: Average Annual Flow of the Nile River at Aswan, May 1870 to May 1945.

```
out <- xtable(ans, caption="OLS Fit. Corresponds to t-test.")
print(out, type="latex")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3337.7688	77.0351	43.33	0.0000
stepIntervention	-727.1008	104.1903	-6.98	0.0000

Table 5: OLS Fit. Corresponds to t-test.

```
tsdiag(arima(resid(ans)))
```

```
ans <- arima(nile, order=c(1,0,0), xreg=stepIntervention)
stargazer(ans, title="ARIMA Fit", header=FALSE)
```

```
tsdiag(ans)
```

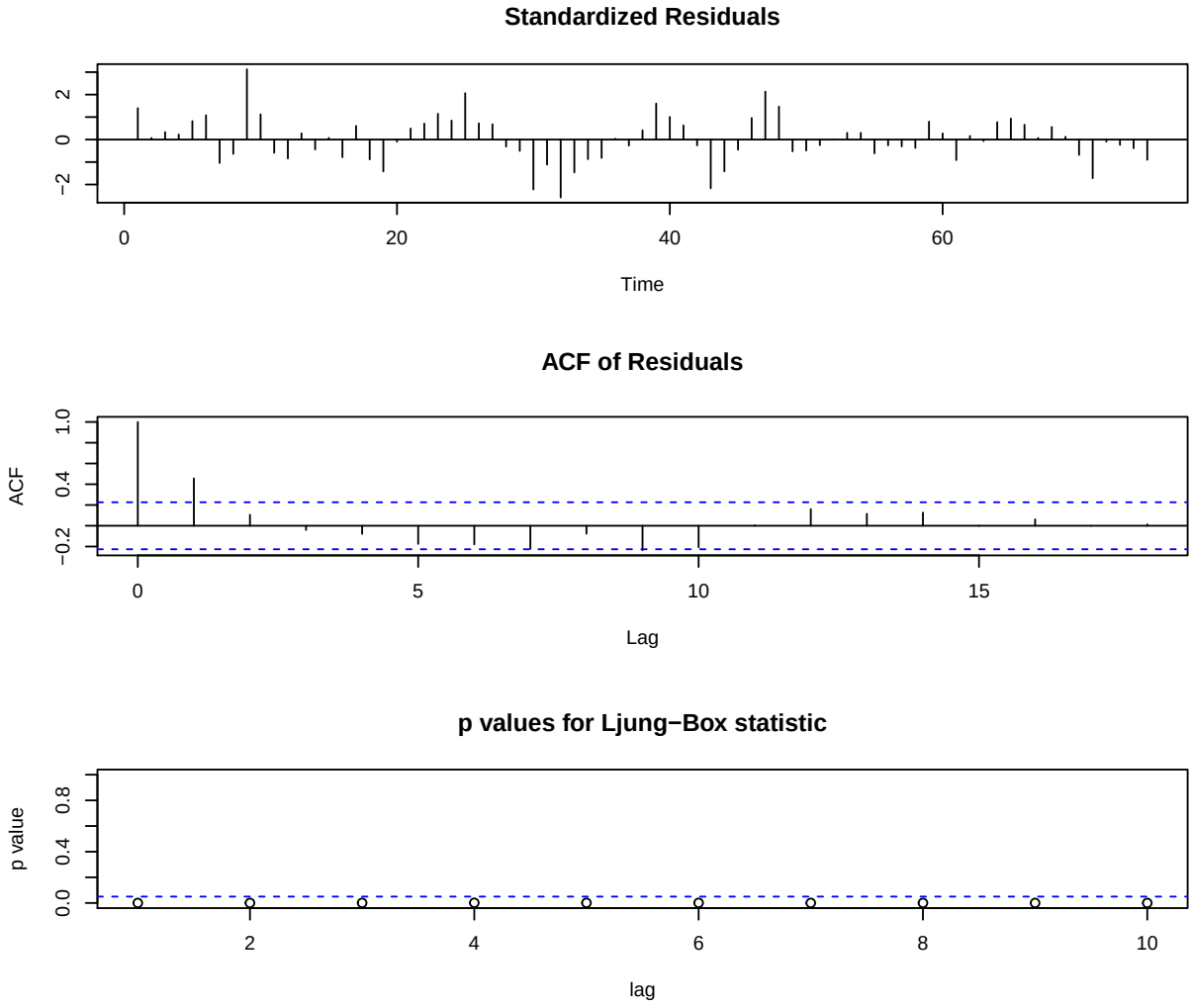
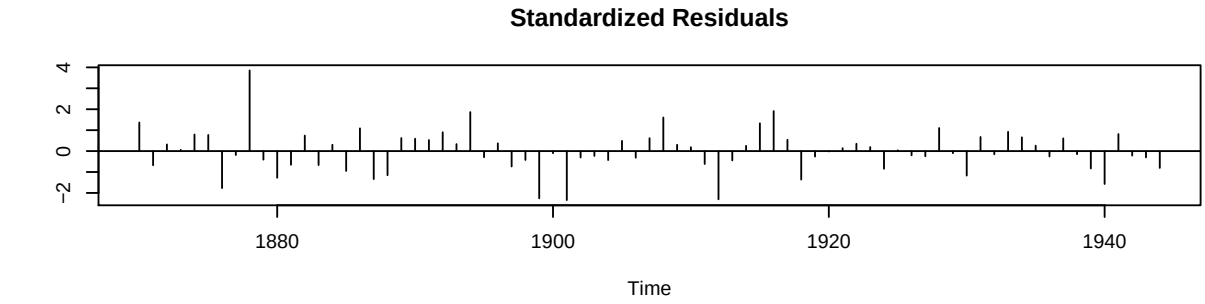
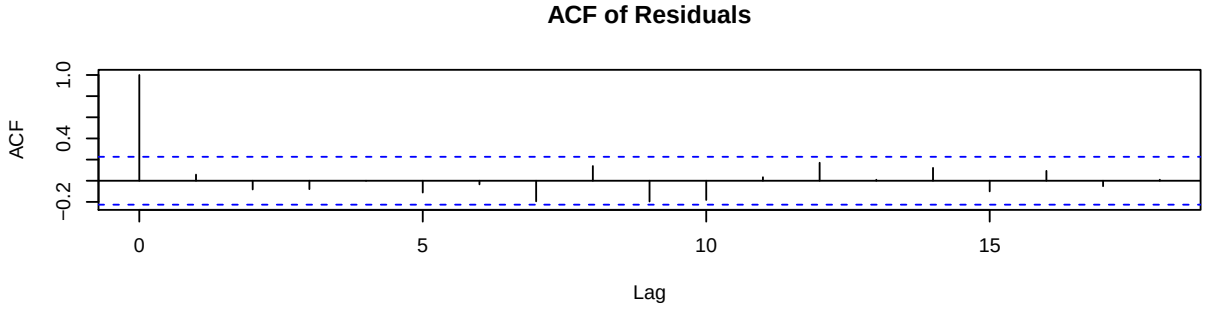


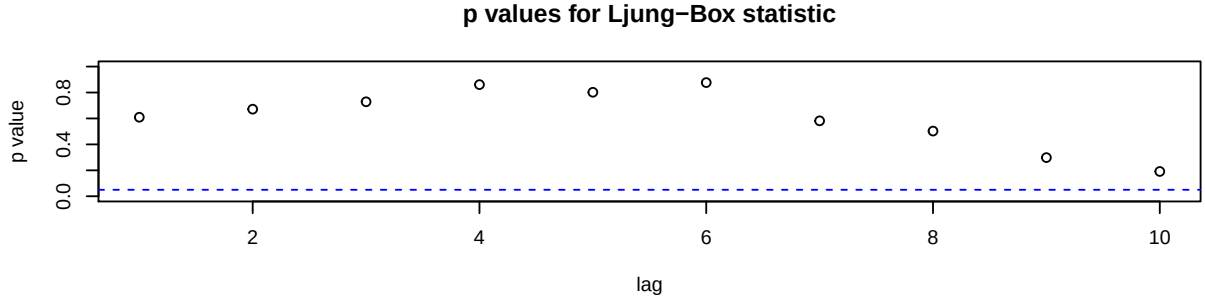
Figure 8: Diagnostic tests for OLS Fit

Table 6: ARIMA Fit

	Dependent variable:
	nile
ar1	0.468***
	(0.103)
intercept	3,351.329***
	(122.198)
stepIntervention	−747.302***
	(162.281)
Observations	75
Log Likelihood	−554.422
σ^2	153,799.600
Akaike Inf. Crit.	1,116.844
Note:	*p<0.1; **p<0.05; ***p<0.01







Dynamic Regression with Autocorrelated Errors

Transfer function and filtering

Consider an output time series $\{y_t\}$ and an input time series $\{x_t\}$ related by a linear filter,

$$y_t = x_t + \sum_{k=1}^{\infty} \nu_k x_{t-k}$$

.

Such linear filters are widely used in electrical engineering and in dynamic regression in econometric models. See Wikipedia article.

It is assumed that a bounded input change, replacing $\{x_t\}$ with $\{x_t + \Delta\}$ produces a bounded change in the output signal $\{y_t + \Delta_y\}$. This implies that $\nu_1 + \nu_2 + \ldots < \infty$ and the linear filter $\{\nu_t\}$ is said to be *stable*. The filter gain is given by

$$g = 1 + \nu_1 + \nu_2 + \ldots.$$

The gain g shows the long-run change in output signal given a unit change in input. So a Δ change in input produces a change of $g\Delta$ in the output.

Intervention Analysis

Regression with ARIMA-GARCH Errors

Appendix A. Durbin-Watson and Related Tests

Appendix B. Quantmod Package. Quantitative Financial Modelling Framework.