Price relations wheat in major domestic and international markets

Samita Paudel

6/28/2019

1 Notes

- Formulate simple linear regression
- Multicollinearlity check…refer to Variance Inflation Factor
- Plot time series
- Autocorrelation check
- Drop variables with high multicollinarity
- Use
  - Darbin-Watson test for autocorrelation test (D-statistic), or
  - Brueuch Godfrey test for autocorrelation
- Use
  - Bursch-pagan test for heteroscadesticity
  - Or plot histogram
- In time series analysis,
  - Use regression with OLS
  - Plot data of each series (Fuel, prices, temperature)
  - Test dickey fueller test (Use no trend, no drift), use all possibility
  - Use augmented dickey fueller test if dickey fueller test does not capture the essense
  - Perform detrending with first order difference

2 Why domestic price is studied at district level

- Until recently, before federal structure of governance was into force, planning, budgeting, service delivery and policy interventions in Agriculture were all excised through district level bodies. Each district was itself accountable for market information accrual and reporting. Hence, most reliable form of price series data would be district level itself.

- Districts present isolated markets and well organized customer segments surrounding that. For e.g., food grain retail market of kathmandu is drastically different from that of kailali, because while in the former consumer segment has a larger role to play in determining of market demand, the market of farwestern terai region has significant share of producer segments in determining what and when to produce.
- District present different socio-economic narrative for food commodity trade which is heavily affected by the geographical context of the district itself. For e.g., Rupandehi market is more closely tied to bordering Indian market, because of minimal to none customs intervention in cross-country trade of food grain. The price effects of indian districts are more easily reflected in border region market prices, than in distant market such as Jumla and Surkhet.
- District markets information systems are more organized than local level markets, mostly because there are factors that buffer price volatility in district level. For e.g., government intervene through subsidized input supply, facilitated by district level agriculture offices when prices of agricultural inputs are hightened. Also service delivery system, for example that providing subsidized farmers loan, is carried out at district level.

3 Why only few domestic markets were selected for study

- These markets either have a large production volumes or strong consumer segment. Because districts of terai, and mostly those of Central to Farwestern region, show consistently high annual production volume (Terai is also dubbed grain basket of Nepal) major producer districts in terai – Kailali, Rupandehi, Parsa are included in the study. At the same time Chitwan and Kathmandu districts have prominance of consumers.
- Some of the features that justify suitability of inclusion of abovementioned districts are presented:
  - Kailali district lies in farwestern region. It borders with India through Uttar Pradesh state. The district has huge chunks of land annually allocated to Wheat production (How much in total ???, what percentage of total wheat cultivated area ???). The farmer segment comprises mainly of Tharu community.
  - Rupandehi district is located in Western terai region. The district adjoining with Indian market of Uttar Pradesh state, likewise, has large volume of grain production arising from Wheat cultivation.
  - Parsa district lies in Central terai part of Nepal. It border with India through Bihar state and the district cultivates Wheat in large volume (areawise how much ???).
- Kathmandu and Chitwan markets are mostly dominanted by consumer segment, hence the retail price series of these districts are expected to differ from that of producer market districts.
- ???Some of the food market features of Kathmandu and Surkhet districts???

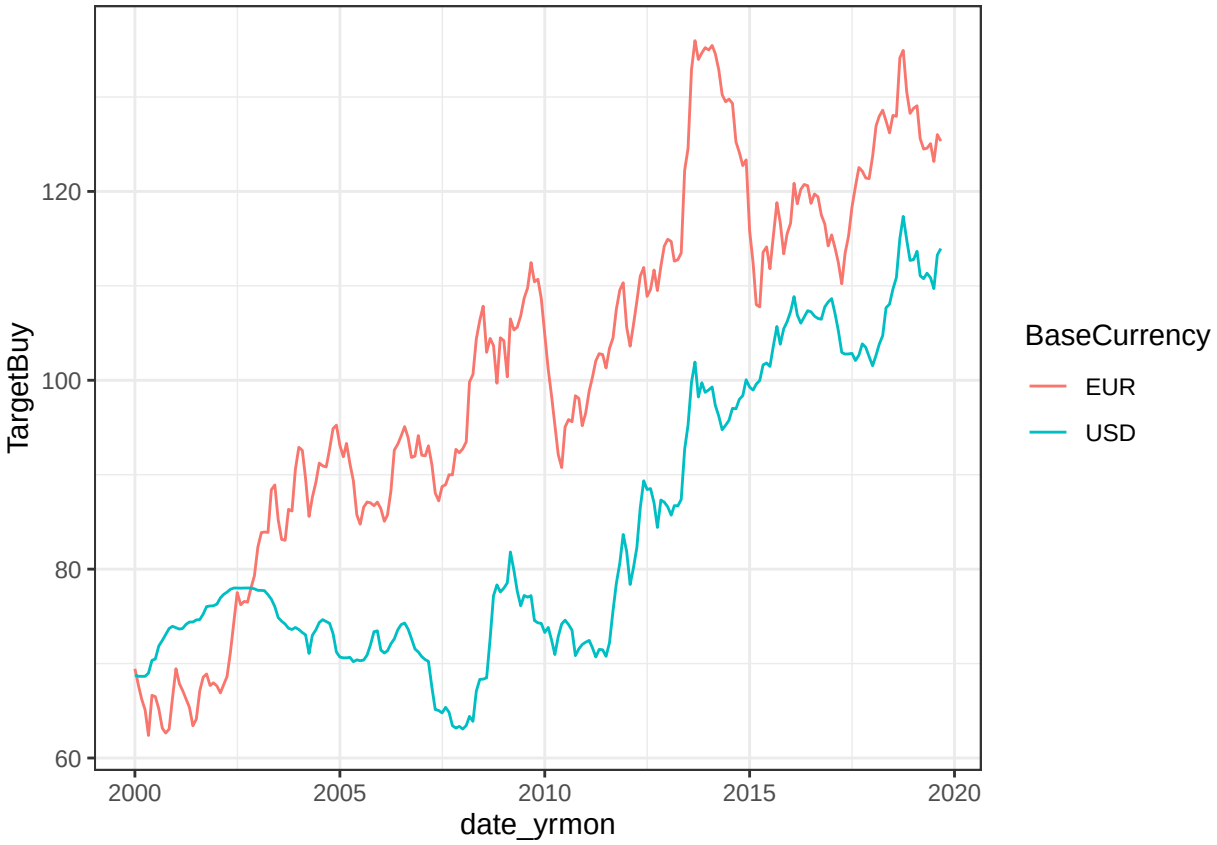
- Treatment of isolated markets into arbitrary category of consumer and producer districts will provide a more complete picture of the situation of region they represent – Farwestern terai, Western-central terai and Eastern-central terai regions, respectively. This form of classification is expected to improve interpretability and overall increase predictive accuray of model in the face of price shocks.

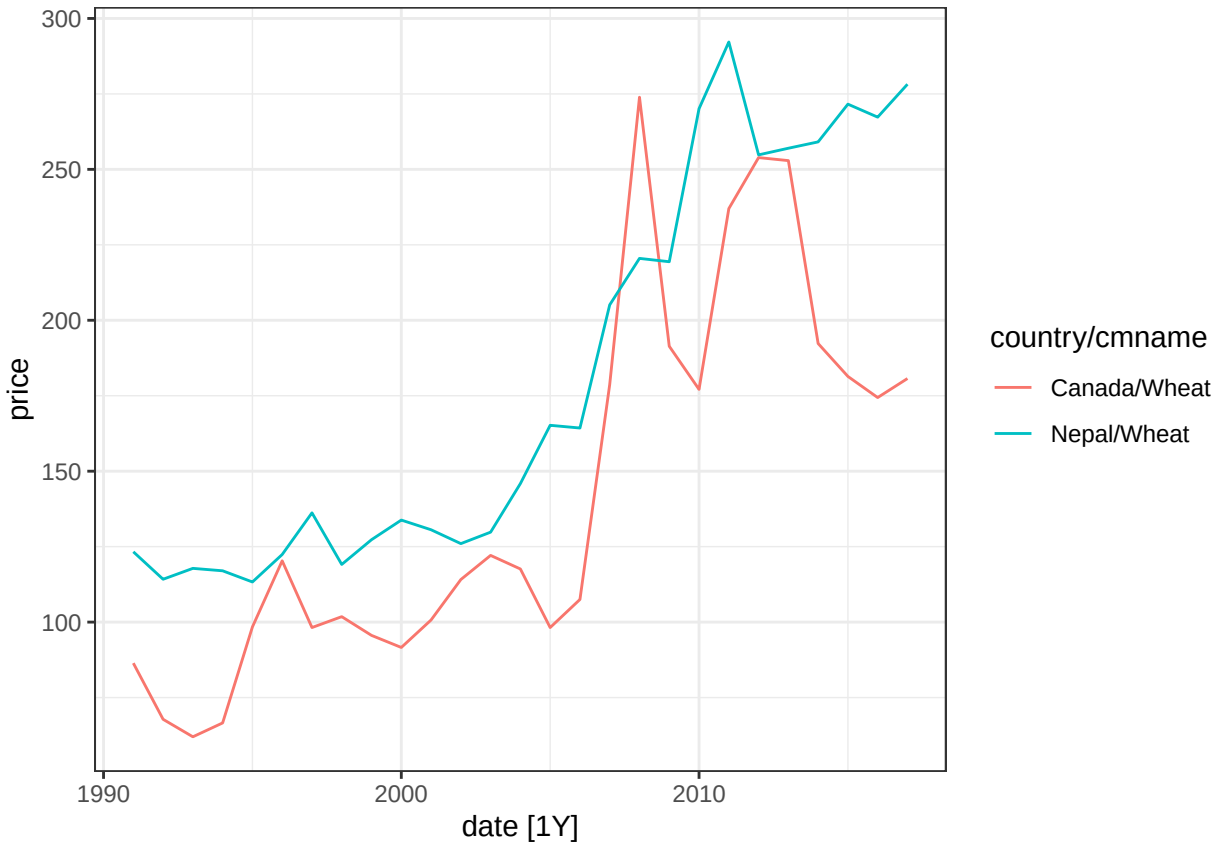
4 Dependent variables

4.1 Retail price of wheat in major domestic markets

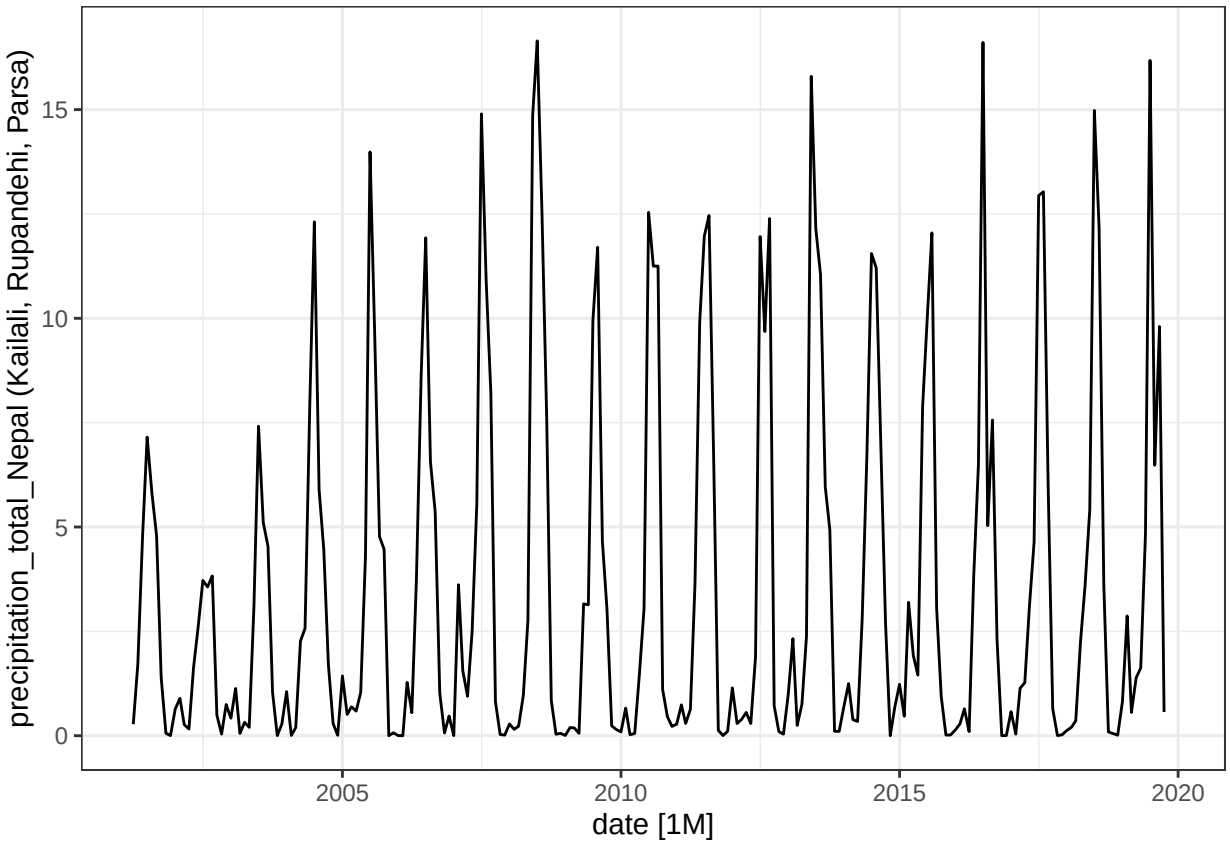
Price series of domestic markets were selected for study. The data represent imbalanced series of following 5 districts:

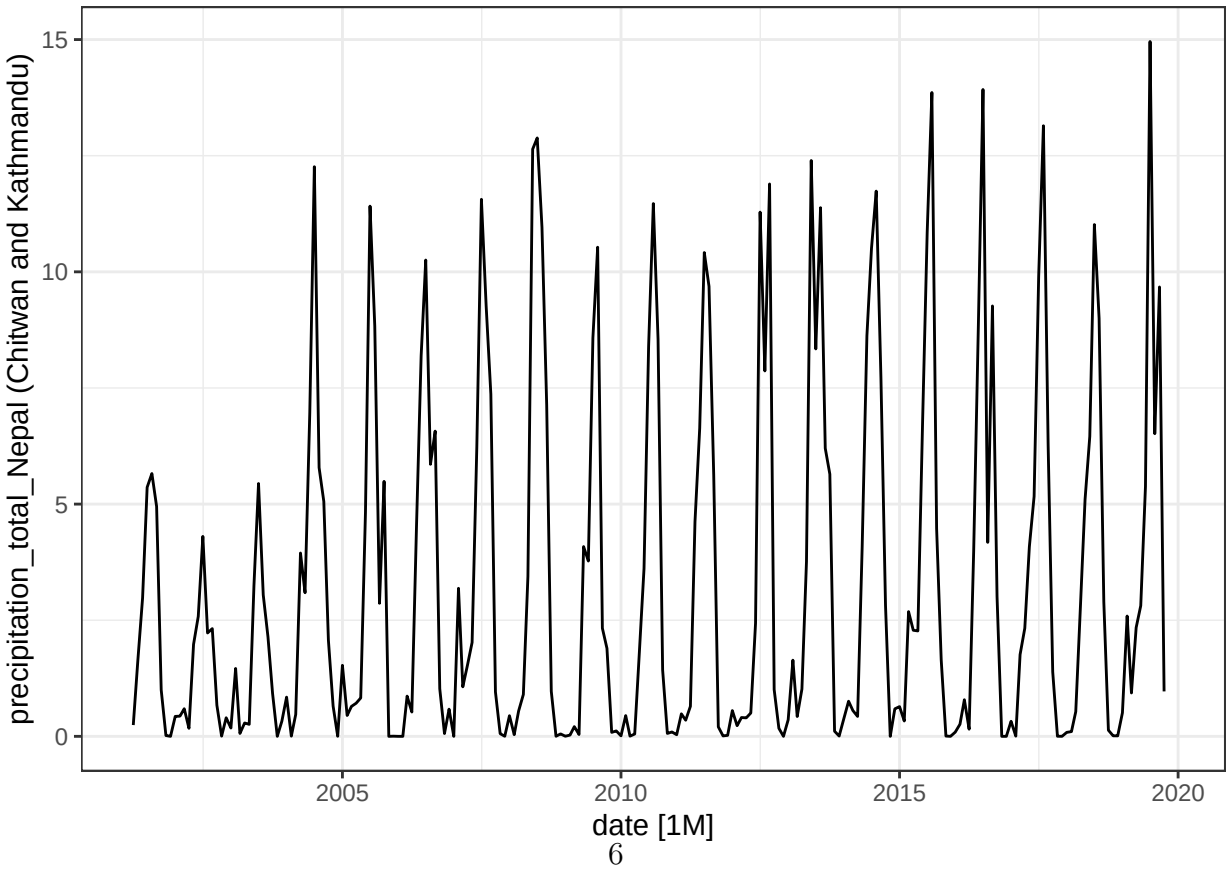
Kailali, Rupandehi, Parsa, Kathmandu, Chitwan




300
250
200
150
100
price
1990
2000
2010
date [1Y]
country/cmname
Canada/Wheat
Nepal/Wheat

4.2 Combined series





5 Geographical context of study districts

5.1 Study districts and market centres

A map of study districts.

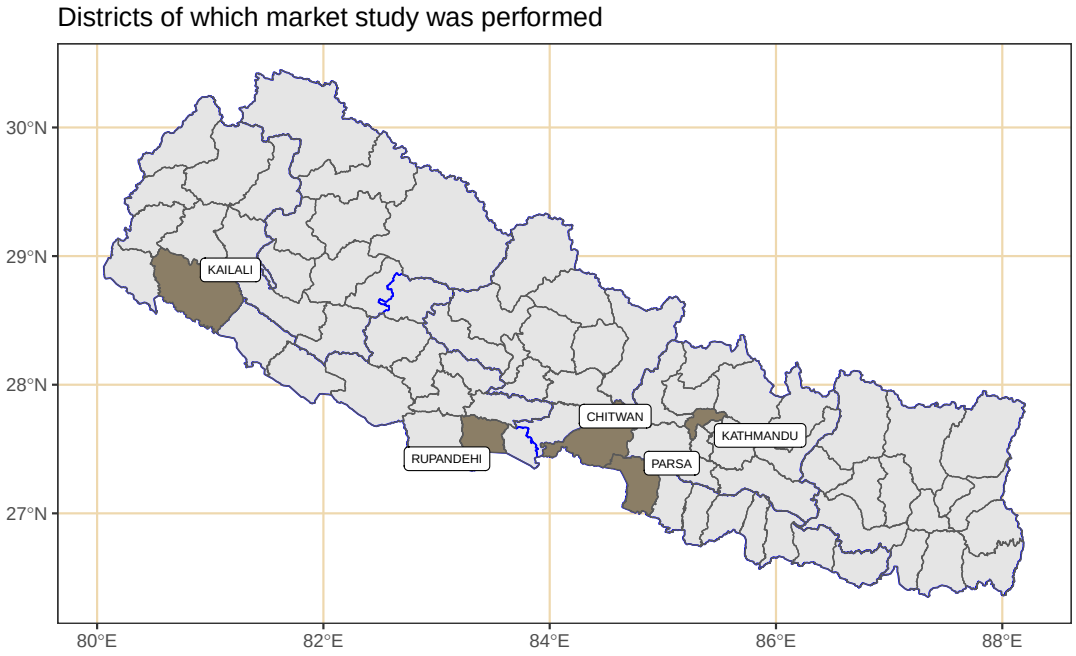


Figure 1: Geographical context of selected district markets

5.2 Aggregate series summary

Joint time series plot of rice and wheat retail prices aggregated over selected districtwise markets and comparison to national average price.

Time series plot of wheat retail price series is presented in above figures with some diagnostic plot alongside. The lineplot shows that there exist some time gaps at random periods. Autocorrelation of consecutive first order differenced lags is similarly shown.

6 Linear regression model formulation for price series

Table 1: ANOVA of regression between price of producer districts (as dependent variable) and 6 regressor variables (price of wheat in international market, price of wheat in indian bordering states, price of wheat in consumer districts, precipitation of consumer districts, precipitation of producer districts)

term	df	sumsq	meansq	statistic	p.value
price canada	1	736.012	736.012	196.169	0.000
price india	1	7604.395	7604.395	2026.798	0.000
price nepal	1	607.508	607.508	161.919	0.000
chitwan and kathmandu precipitation	1	0.070	0.070	0.019	0.891
total nepal chitwan and kathmandu precipitation	1	0.001	0.001	0.000	0.990
total nepal kailali rupandehi parsa					
fuel price	1	139.376	139.376	37.148	0.000
Residuals	150	562.789	3.752	NA	NA

The regression above (Table 1) is obtained on fitting the model Equation 1.

$$\begin{aligned} price_{\text{producer districts}} = {} & price_{\text{canada}} + price_{\text{indian bordering states}} \\ & + price_{\text{consumer districts}} + precipitation_{\text{producer districts}} \\ & + precipitation_{\text{consumer districts}} + fuel\ price_{\text{national average}} \end{aligned} \tag{1}$$

This presents a typical case of spurious relationship among variables, where variables having times series attributes show exceptionally high association among them. For example, all three aggregated wheat price series we consider in the regression (price in Canada, price in India, and price in Nepalese consumer markets) which show highly significant association. This is problematic and misleading, because without accounting for linear time trend they show very unstable variance, which increases at extremes (more recent time period). The phenomena of possible presence of heteroskedasticity is shown in the residual-vs-fit plot in Figure 2.

Simply after incorporating linear trend of time in the regression model (Equation 1), we get a different statistic.

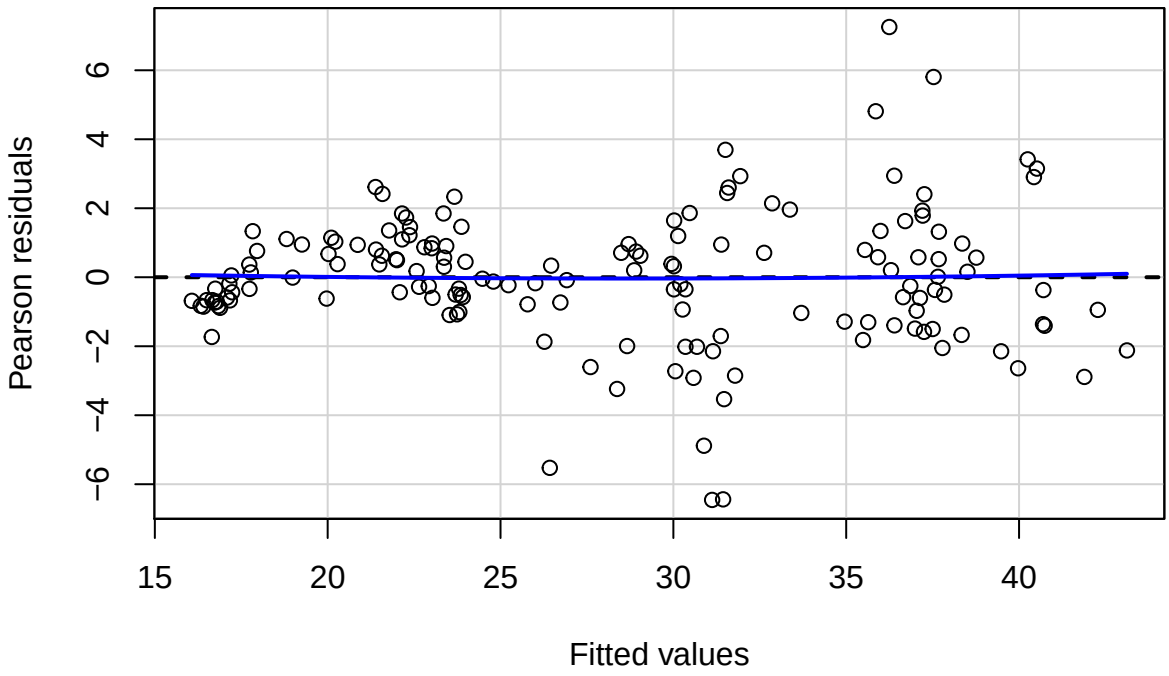


Figure 2: Residual (pearsons') versus fit plot of the linear regression without accounting for time attributes of the series

$$
\begin{aligned}
price_{\text{producer districts}} = {} & date + price_{\text{canada}} + price_{\text{indian bordering states}} \\
& + price_{\text{consumer districts}} + precipitation_{\text{producer districts}} \\
& + precipitation_{\text{consumer districts}} + fuel\ price_{\text{national average}}
\end{aligned} \tag{2}
$$

However, even then there is likely presence of biased variance, as shown in Figure 3 which uses Equation 2 for model fitting.

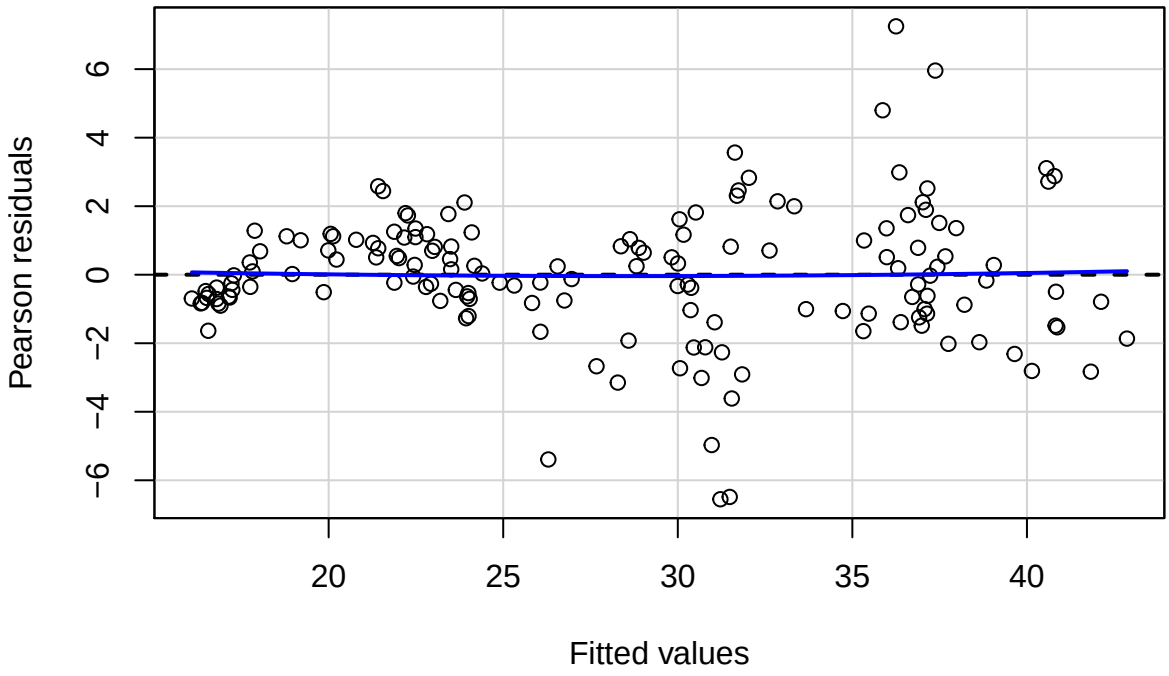


Figure 3: Residual (pearsons') versus fit plot of the linear regression after incorporating linear time trend of the price series

6.1 Testing for heteroskedasticity

Below we test the linear model (Equation 2 and its counterpart with consumer districts as dependent variable and same independent variables with linear time trend) for presence of heteroskedasticity using Breusch-Pagan test. The test fits a linear regression model to the residuals of a linear regression model (by default the same explanatory variables are taken as in the main regression model) and rejects if too much of the variance is explained by the additional explanatory variables.

Table 2: Breusch-Pagan test for heteroskedasticity of price series, modeled by the regression Equation ef(eqn:lm2) for two domestic price series (producers districts and consumer districts)

statistic	p.value	parameter	method
19.65260	0.0063710	7	studentized Breusch-Pagan test
26.73769	0.0003715	7	studentized Breusch-Pagan test

Breusch pagan test uses the null hypothesis of Homoscadesticity while testing studentized residuals. Hence, if the null hypothesis is rejected ($p < 0.05$), there is possible presence of heteroskedasticity.

A possible measure to removing non-stationary trend in the series is by differencing (with `diff`). However, before progressing we confirm that justifiable lag operations can infact render the series free of trends. For this, two popular unit test routines are performed – Augmented Dickey-Fueller test and KPSS test.

7 Unit root testing

7.1 Unit root (ADF and KPSS) test of retail price

The ADF, available in the function `adf.test()` (in the package `tseries`) implements the t-test of $H_0 : \gamma = 0$ in the regression, below.

$$\Delta Y_t = \beta_1 + \beta_2 t + \gamma Y_{t-1} + \sum_{i=1}^{m} \delta_i \Delta Y_{t-i} + \varepsilon_t \tag{3}$$

The null is therefore that x has a unit root. If only x has a non-unit root, then the x is stationary (rejection of null hypothesis).

The ADF test was parametrized with the alternative hypothesis of stationarity. This extends to following assumption in the model parameters;

$$-2 \leq \gamma \leq 0 \text{ or } (-1 < 1 + \phi < 1)$$

`k` in the function refers to the number of δ lags, i.e., $1, 2, 3, ..., m$ in the model equation.

The number of lags `k` defaults to `trunc((length(x)-1)^(1/3))`, where `x` is the series being tested. The default value of `k` corresponds to the suggested upper bound on the rate at which the number of lags, `k`, should be made to grow with the sample size for the general ARMA(p,q) setup `citation(package = "tseries")`.

For a Dickey-Fueller test, so only up to AR(1) time dependency in our stationary process, we set `k = 0`. Hence we have no δs (lags) in our test.

Table 3: Unit root test of log(price) series variables (both dependent and independent)

series	test	lprice pvalue	lprice tstatistic	lprice null accepted
price canada	adf	0.42	-2.37	TRUE
price india	adf	0.16	-3.00	TRUE
price nepal chitwan and kathmandu	adf	0.44	-2.32	TRUE
price nepal kailali rupandehi parsa	adf	0.26	-2.75	TRUE
fuel price	adf	0.09	-3.16	TRUE
price canada	kpss	0.01	1.60	FALSE
price india	kpss	0.01	4.29	FALSE
price nepal chitwan and kathmandu	kpss	0.01	3.47	FALSE
price nepal kailali rupandehi parsa	kpss	0.01	4.03	FALSE
fuel price	kpss	0.01	2.01	FALSE

The DF model can be written as:

$$Y_t = \beta_1 + \beta_2 t + \phi Y_{t-1} + \varepsilon_t$$

It can be re-written so we can do a linear regression of ΔY_t against t and Y_{t-1} and test if ϕ is different from 0. If only, ϕ is not zero and assumption above $(-1 < 1 + \phi < 1)$ holds, the process is stationary. If ϕ is straight up 0, then we have a random walk process – all white noise.

$$\Delta Y_t = \beta_1 + \beta_2 t + \gamma Y_{t-1} + \varepsilon_t$$

Alternative to above discussed tests, the Phillips-Perron test with its nonparametric correction for autocorrelation (essentially employing a HAC estimate of the long-run variance in a Dickey-Fuller-type test instead of parametric decorrelation) may be used. It is available in the function `pp.test()`.

7.2 Unit root test based lag order differencing determination

An alternative to decomposition for removing trends is differencing (Woodward, Gray, and Elliott 2017). We define the difference operator as,

Table 4: ANOVA of linear model fit with trend component of precipitation series with price(producer market) as dependent variable.

term	df	sumsq	meansq	statistic	p.value
date	1	8810.50	8810.50	2463.40	0.00
price_canada	1	16.17	16.17	4.52	0.04
price_india	1	12.83	12.83	3.59	0.06
price_nepal_chitwan_and_kathmandu	1	145.70	145.70	40.74	0.00
precipitation_total_ck_trend	1	10.54	10.54	2.95	0.09
precipitation_total_krp_trend	1	33.17	33.17	9.27	0.00
fuel_price	1	88.33	88.33	24.70	0.00
Residuals	149	532.91	3.58	NA	NA

$$\nabla x_t = x_t - x_{t-1}, \tag{4}$$

and, more generally, for order d

$$\nabla^d x_t = (1 - \mathbf{B})^d x_t, \tag{5}$$

Where $\mathbf{B}$ is the backshift operator (i.e., $\mathbf{B}^k x_t = x_{t-k}$ for $k \geq 1$).

Applying the difference to a random walk, the most simple and widely used time series model, will yield a time series of Gaussian white noise errors $\{w_t\}$:

$$\begin{aligned} \nabla(x_t &= x_{t-1} + w_t) \\ x_t - x_{t-1} &= x_{t-1} - x_{t-1} + w_t \\ x_t - x_{t-1} &= w_t \end{aligned} \tag{6}$$

We use an implementation of time series differencing based on optimal lag length in order to render series stationary. The first lag order differenced log(price) series (derived based on "kpss" test statistic) is shown in Figure 4. The kpss test, however, determined the precipitation series to be integrated of order null. This is due to insensitivity of test (model) to seasonal lag component, which is infact nicely captured by addtive or multiplicative trend decomposition (discussed ahead).

The first order differencing of price renders series stationary. However, precipitation series shows distinct components. Here, a linear model is fitted to the trend component, decomposed (into seasonal and trend components) from two precipitation series (producer and consumer districts), along with other terms of Equation 2 and the ANOVA output is presented.

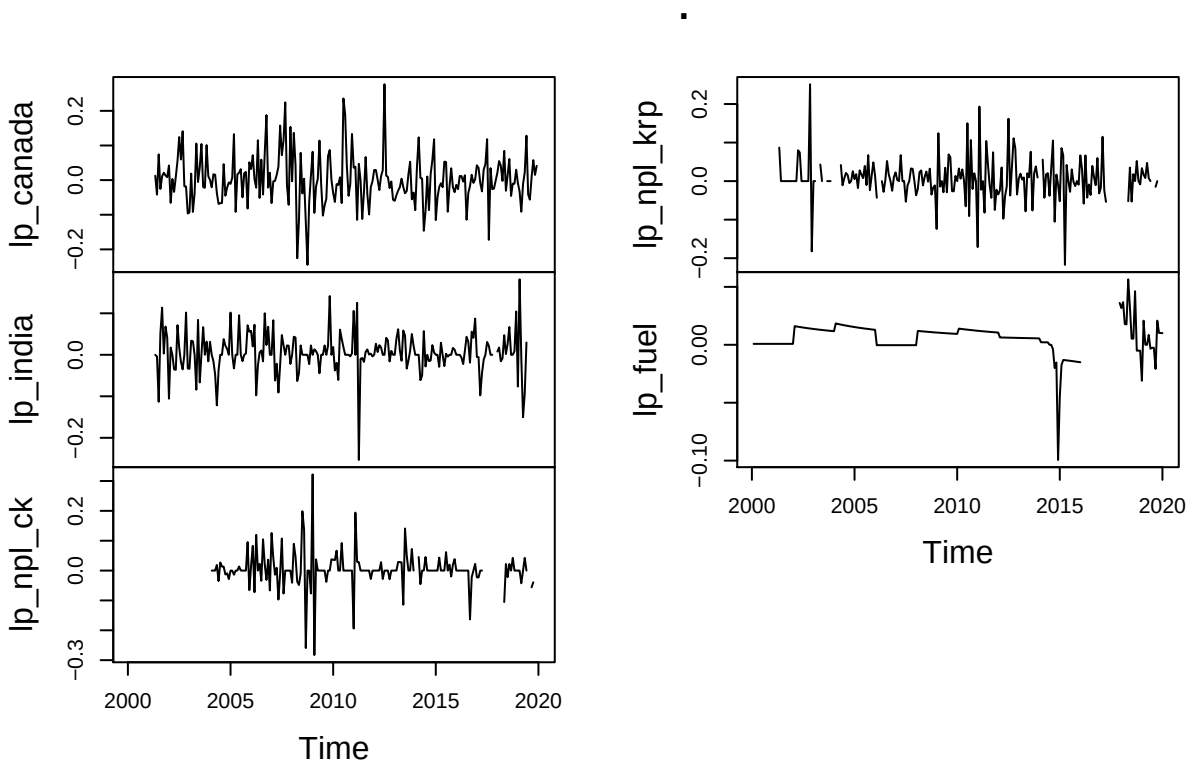


Figure 4: Plot of differenced time series for various lag length determined by kpss statistic.

Table 5: ANOVA of linear model fit with trend component of precipitation series with price(consumer market) as dependent variable.

term	df	sumsq	meansq	statistic	p.value
date	1	14695.17	14695.17	2146.39	0.00
price_canada	1	0.41	0.41	0.06	0.81
price_india	1	50.51	50.51	7.38	0.01
price_nepal_kailali_rupandehi_parsa	1	243.06	243.06	35.50	0.00
precipitation_total_ck_trend	1	0.45	0.45	0.07	0.80
precipitation_total_krp_trend	1	72.78	72.78	10.63	0.00
fuel_price	1	15.89	15.89	2.32	0.13
Residuals	149	1020.12	6.85	NA	NA

8 VAR model

8.1 VAR

VAR is a system regression model, i.e., there are more than one dependent variable. The regression is defined by a set of linear dynamic equations where each variable is specified as a function of an equal number of lags of itself and all other variables in the system. Any additional variable, adds to the modeling complexity by increasing an extra equation to be estimated.

The vector autoregression (VAR) model extends the idea of univariate autoregression to k time series regressions, where the lagged values of *all* k series appear as regressors. Put differently, in a VAR model we regress a *vector* of time series variables on lagged vectors of these variables. As for $AR(p)$ models, the lag order is denoted by p so the $VAR(p)$ model of two variables X_t and Y_t ($k = 2$) is given by a vector of equations (Equation 7).

$$\begin{aligned} Y_t &= \beta_{10} + \beta_{11}Y_{t-1} + \cdots + \beta_{1p}Y_{t-p} + \gamma_{11}X_{t-1} + \cdots + \gamma_{1p}X_{t-p} + u_{1t}, \\ X_t &= \beta_{20} + \beta_{21}Y_{t-1} + \cdots + \beta_{2p}Y_{t-p} + \gamma_{21}X_{t-1} + \cdots + \gamma_{2p}X_{t-p} + u_{2t}. \end{aligned} \tag{7}$$

The βs and γs can be estimated using OLS on each equation.

Simplifying this to a bivariate $VAR(1)$, we can write the model in matrix form as:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \mu_t \tag{8}$$

Where,

- Y_t, Y_{t-1} and μ_t are (2 x 1) column vectors
- β_0 is a (2 x 1) column vector
- β_1 is a (2 x 2) matrix

also,

$$Y_t = \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix}, \; Y_{t-1} = \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix}$$

$$\mu_t = \begin{pmatrix} \mu_{1t} \\ \mu_{2t} \end{pmatrix}, \beta_0 = \begin{pmatrix} \beta_{10} \\ \beta_{20} \end{pmatrix}, \beta_1 = \begin{pmatrix} \beta_{11} & \alpha_{11} \\ \alpha_{21} & \beta_{21} \end{pmatrix}$$

It is straightforward to estimate VAR models in `R`. A feasible approach is to simply use `lm()` for estimation of the individual equations. Furthermore, the `vars` package provides standard tools for estimation, diagnostic testing and prediction using this type of models.

Only when the assumptions presented below hold, the OLS estimators of the VAR coefficients are consistent and jointly normal in large samples so that the usual inferential methods such as confidence intervals and t-statistics can be used (Metcalfe and Cowpertwait 2009).

Two series $w_{x,t}$ and $w_{y,t}$ are bivariate white noise if they are stationary and their cross-covariances $\gamma_{xy}(k) = Cov(w_{x,t}, w_{y,t+k})$ satisfies

$$\gamma_{xx}(k) = \gamma_{yy}(k) = \gamma_{xy}(k) = 0 \text{ for all } k \neq 0$$

The parameters of a var(p) model can be estimated using the `ar` function in `R`, which selects a best-fitting order p based on the smallest information criterion values.

The structure of VARs also allows to jointly test restrictions across multiple equations. For instance, it may be of interest to test whether the coefficients on all regressors of the lag p are zero. This corresponds to testing the null that the lag order $p-1$ is correct. Large sample joint normality of the coefficient estimates is convenient because it implies that we may simply use an F-test for this testing problem. The explicit formula for such a test statistic is rather complicated but fortunately such computations are easily done using the `ttcode("R")` functions we work with in this chapter. Just as in the case of a single equation, for a multiple equation model we choose the specification which has the smallest $BIC(p)$, where

$$BIC(p) = \log\left[\det(\widehat{\Sigma}_u)\right] + k(kp+1)\frac{\log(T)}{T}.$$

with $\widehat{\Sigma}_u$ denoting the estimate of the $k \times k$ covariance matrix of the VAR errors and $\det(\cdot)$ denotes the determinant.

As for univariate distributed lag models, one should think carefully about variables to include in a VAR, as adding unrelated variables reduces the forecast accuracy by increasing the estimation error. This is particularly important because the number of parameters to be estimated grows qudratically to the number of variables modeled by the VAR.

Table 6: Model coefficients of VAR(AR(1)) model for for wheat log(price) series with two domestic (producer and consumer) markets as endogeneous variables and other price series, precipitation and fuel series as exogeneous regressors.

term	.response	estimate	std.error	statistic	p.value
lag(lp_npl_ck,1)	lp_npl_ck	0.848	0.032	26.572	0.000
lag(lp_npl_krp,1)	lp_npl_ck	0.055	0.041	1.356	0.176
constant	lp_npl_ck	0.110	0.044	2.502	0.013
lp_canada	lp_npl_ck	-0.003	0.015	-0.231	0.818
lp_india	lp_npl_ck	0.096	0.037	2.578	0.011
precip_npl_ck	lp_npl_ck	-0.002	0.004	-0.520	0.604

precip_npl_krp	lp_npl_ck	0.003	0.004	0.761	0.447
lp_fuel	lp_npl_ck	-0.011	0.027	-0.415	0.679
lag(lp_npl_ck,1)	lp_npl_krp	0.033	0.030	1.100	0.273
lag(lp_npl_krp,1)	lp_npl_krp	0.792	0.039	20.530	0.000
constant	lp_npl_krp	-0.060	0.042	-1.438	0.152
lp_canada	lp_npl_krp	-0.003	0.014	-0.205	0.838
lp_india	lp_npl_krp	0.102	0.035	2.914	0.004
precip_npl_ck	lp_npl_krp	-0.002	0.004	-0.457	0.648
precip_npl_krp	lp_npl_krp	0.002	0.003	0.700	0.485
lp_fuel	lp_npl_krp	0.079	0.025	3.175	0.002

–>

–>

–> –> –> –>

–>

Bibliography

Metcalfe, Andrew V, and Paul SP Cowpertwait. 2009. *Introductory Time Series with R*. Springer.

Woodward, Wayne A, Henry L Gray, and Alan C Elliott. 2017. *Applied Time Series Analysis with R*. CRC press.