VAR Example

D2K Course Staff

March 26 2019

Introduction

In the traditional multiple linear regression model with $i \in 1, ..., n$ observations, a response variable $\mathbf{Y}$ is modeled as a linear combination of predictor variables $\mathbf{X}_j$, $j \in 1, ..., p$:

$$\mathbf{Y}_i = \beta_0 + \beta_1 \mathbf{X}_{i1} + ... + \beta_p \mathbf{X}_{ip} + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(\mathbf{0}_p, \sigma^2 \mathbf{I}_p).$$

However, this is subject to the assumption that our observations are independently and identically distributed. One particular case where this assumption is almost surely violated is in the case of time series data. In this setting, we often see that our variables are correlated with themselves at previous time points; this is called *autocorrelation*. Below, we will show an example from the `uschange` data set from the `fpp2` package in R. The data set contains values for 5 quarterly US economic indicators from 1960 to 2016. The model will be fit using the `vars` package.

```
        Consumption     Income Production   Savings Unemployment
1970 Q1   0.6159862  0.9722610 -2.4527003 4.8103115          0.9
1970 Q2   0.4603757  1.1690847 -0.5515251 7.2879923          0.5
1970 Q3   0.8767914  1.5532705 -0.3587079 7.2890131          0.5
1970 Q4  -0.2742451 -0.2552724 -2.1854549 0.9852296          0.7
1971 Q1   1.8973708  1.9871536  1.9097341 3.6577706         -0.1
1971 Q2   0.9119929  1.4473342  0.9015358 6.0513418         -0.1
```

Autoregressive (AR) Process

For the case of one single variable, we write an autoregressive process for a variable Y as:

$$Y_t = \mu + \phi_1 Y_{t-1} + ... + \phi_k Y_{t-k} + \epsilon_t, \epsilon \overset{iid}{\sim} N(0, \sigma^2)$$

where the subscript denotes the time point of the observation. Essentially, this means that the value of Y at time t is a function of the value of Y at times $t-1, t-2, ..., t-k$, along with some random Gaussian noise. The difference between the times of the observations is called the *lag*, e.g. Y_{t-2} is considered to be lag 2 for Y_t and to be lag 1 for Y_{t-1}. The parameter k denotes the largest lag at which there is a statistically significant autocorrelation with the current time point, and the model is denoted as an AR(k) process. The value of the autocorrelation between Y_t and Y_{t-i} is simply the stanardized version of the parameter ϕ_i. The key assumptions for this model are stationarity, meaning that the mean and variance of the process does not change, and independence of the residuals ϵ_t.

Vector Autoregressive (VAR) Process

The vector autoregressive process extends the autoregressive process shown above to multiple variables. For example, if we have 3 variables X, Y, and Z, we write the model as:

$$X_t = \mu_X + \gamma_{X,1} Y_{t-1} + ... + \gamma_{X,k} Y_{t-k} + \alpha_{X,1} X_{t-1} + ... + \alpha_{X,1} X_{t-k} + \beta_{X,1} Z_{t-1} + ... + \beta_{X,k} Z_{t-k} + \epsilon_{X,t}, \epsilon_X \overset{iid}{\sim} N(0, \sigma_X^2)$$

$$Y_t = \mu_Y + \gamma_{Y,1} Y_{t-1} + ... + \gamma_{Y,k} Y_{t-k} + \alpha_{Y,1} X_{t-1} + ... + \alpha_{Y,k} X_{t-k} + \beta_{Y,1} Z_{t-1} + ... + \beta_{Y,k} Z_{t-k} + \epsilon_{Y,t}, \epsilon_Y \overset{iid}{\sim} N(0, \sigma_Y^2)$$

$Z_t = \mu_Z + \gamma_{Z,1} Y_{t-1} + ... + \gamma_{Z,k} Y_{t-k} + \alpha_{Z,1} X_{t-1} + ... + \alpha_{Z,1} X_{t-k} + \beta_{Z,1} Z_{t-1} + ... + \beta_{Z,k} Z_{t-k} + \epsilon_{Z,t}, \epsilon_Z \overset{iid}{\sim} N(0, \sigma_Z^2)$.

As above, the parameter k denotes the maximum time lags with significant parameters for X, Y. and Z (we have the variables have the same maximum lag). This is called a VAR(k) model. The assumption for stationarity is required here as well for all variables. The residuals ϵ_X, ϵ_Y and ϵ_Z need to be independently and identically distributed with respect to time, but they do not necessarily need to independent of one another contemporaneously. In other words, the model is valid if $\epsilon_{X,t}$ is correlated with $\epsilon_{Y,t}$ but not if $\epsilon_{X,t}$ is correlated with $\epsilon_{X,t-1}$. In fact, we can consider contemporaneous correlations between the residuals for each of the variables as a measure of the relative strength of the relationship between the variables at the same current time point.

Example

Model Fit

We fit a VAR(1) on the 5 variabls in the `uschange` data. The summary of the fit is shown below. In total, we will have 30 different estimated coefficients - the 5 variables at lag 1 and an intercept predicting each of the variables in the data set. At the bottom, we also see the covariance and correlation matrix of the residuals.

```
VAR Estimation Results:
========================
Endogenous variables: Consumption, Income, Production, Savings, Unemployment
Deterministic variables: const
Sample size: 186
Log Likelihood: -1210.858
Roots of the characteristic polynomial:
0.4274 0.4274 0.3343 0.1767 0.1486
Call:
VAR(y = uschange)


Estimation results for equation Consumption:
============================================
Consumption = Consumption.l1 + Income.l1 + Production.l1 + Savings.l1 + Unemployment.l1 + const

                Estimate Std. Error t value Pr(>|t|)
Consumption.l1   0.268008   0.139468   1.922   0.0562 .
Income.l1        0.129989   0.127401   1.020   0.3090
Production.l1   -0.016624   0.049104  -0.339   0.7354
Savings.l1      -0.002415   0.008199  -0.295   0.7687
Unemployment.l1 -0.056899   0.200536  -0.284   0.7769
const            0.465606   0.079398   5.864 2.12e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.6182 on 180 degrees of freedom
Multiple R-Squared: 0.1386, Adjusted R-squared: 0.1147
F-statistic: 5.792 on 5 and 180 DF,  p-value: 5.524e-05


Estimation results for equation Income:
=======================================
Income = Consumption.l1 + Income.l1 + Production.l1 + Savings.l1 + Unemployment.l1 + const

                 Estimate Std. Error t value Pr(>|t|)
Consumption.l1   0.5473627  0.1973053   2.774  0.00612 **
Income.l1       -0.2328240  0.1802352  -1.292  0.19809
Production.l1   -0.1538254  0.0694678  -2.214  0.02806 *
Savings.l1      -0.0005203  0.0115985  -0.045  0.96427
Unemployment.l1 -0.4883886  0.2836987  -1.722  0.08688 .
const            0.5572868  0.1123250   4.961 1.61e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.8746 on 180 degrees of freedom
Multiple R-Squared: 0.1462, Adjusted R-squared: 0.1225
F-statistic: 6.166 on 5 and 180 DF,  p-value: 2.667e-05


Estimation results for equation Production:
===========================================
Production = Consumption.l1 + Income.l1 + Production.l1 + Savings.l1 + Unemployment.l1 + const

                Estimate Std. Error t value Pr(>|t|)
Consumption.l1   0.802484   0.261965   3.063  0.00253 **
Income.l1       -0.098022   0.239301  -0.410  0.68257
Production.l1    0.269798   0.092233   2.925  0.00389 **
Savings.l1       0.004885   0.015400   0.317  0.75143
Unemployment.l1 -0.768606   0.376671  -2.041  0.04276 *
const           -0.142099   0.149135  -0.953  0.34196
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 1.161 on 180 degrees of freedom
Multiple R-Squared: 0.4421, Adjusted R-squared: 0.4266
F-statistic: 28.53 on 5 and 180 DF,  p-value: < 2.2e-16


Estimation results for equation Savings:
========================================
Savings = Consumption.l1 + Income.l1 + Production.l1 + Savings.l1 + Unemployment.l1 + const

               Estimate Std. Error t value Pr(>|t|)
Consumption.l1    2.1480     2.9318   0.733   0.4647
Income.l1        -3.7718     2.6781  -1.408   0.1607
Production.l1    -1.6412     1.0322  -1.590   0.1136
Savings.l1       -0.0984     0.1723  -0.571   0.5687
Unemployment.l1  -5.5491     4.2155  -1.316   0.1897
const             3.3048     1.6690   1.980   0.0492 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 13 on 180 degrees of freedom
Multiple R-Squared: 0.1247, Adjusted R-squared: 0.1004
F-statistic: 5.128 on 5 and 180 DF,  p-value: 0.0002016


Estimation results for equation Unemployment:
=============================================
Unemployment = Consumption.l1 + Income.l1 + Production.l1 + Savings.l1 + Unemployment.l1 + const

                Estimate Std. Error t value Pr(>|t|)
Consumption.l1  -0.110761   0.068311  -1.621  0.10668
Income.l1       -0.043782   0.062401  -0.702  0.48381
```

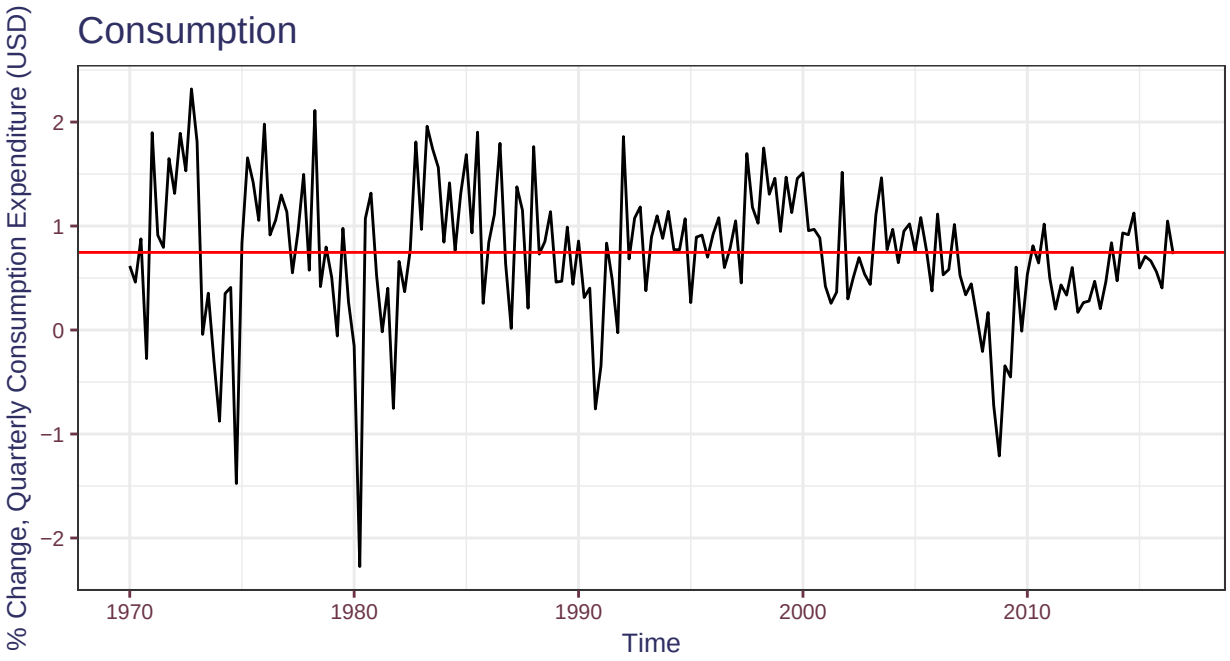
```
Production.l1   -0.021564   0.024051  -0.897  0.37113
Savings.l1       0.002682   0.004016   0.668  0.50512
Unemployment.l1  0.319604   0.098222   3.254  0.00136 **
const            0.122065   0.038889   3.139  0.00198 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.3028 on 180 degrees of freedom
Multiple R-Squared: 0.3499, Adjusted R-squared: 0.3319
F-statistic: 19.38 on 5 and 180 DF,  p-value: 2.057e-15



Covariance matrix of residuals:
             Consumption  Income Production  Savings Unemployment
Consumption      0.38221  0.2099     0.3696  -1.8884     -0.09289
Income           0.20989  0.7650     0.2323   8.1146     -0.04210
Production       0.36957  0.2323     1.3485  -1.1934     -0.23137
Savings         -1.88839  8.1146    -1.1934 168.8951      0.54695
Unemployment    -0.09289 -0.0421    -0.2314   0.5469      0.09169

Correlation matrix of residuals:
             Consumption  Income Production  Savings Unemployment
Consumption       1.0000  0.3882    0.51478 -0.23503      -0.4962
Income            0.3882  1.0000    0.22869  0.71391      -0.1590
Production        0.5148  0.2287    1.00000 -0.07908      -0.6580
Savings          -0.2350  0.7139   -0.07908  1.00000       0.1390
Unemployment     -0.4962 -0.1590   -0.65799  0.13898       1.0000
```

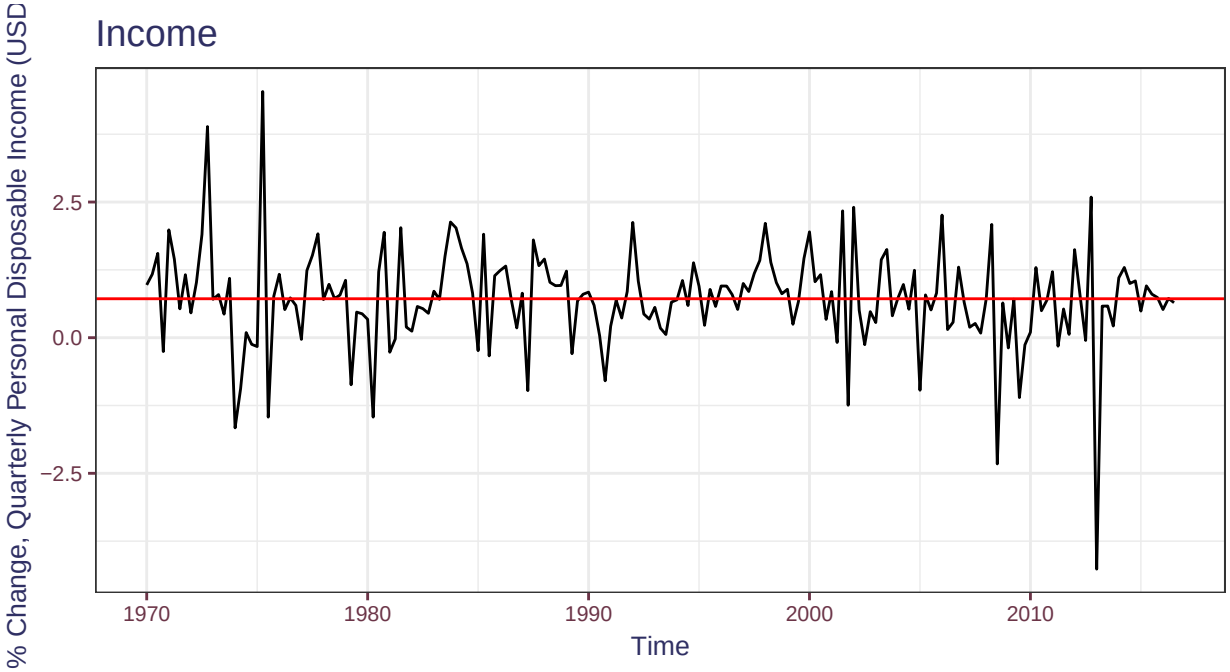
Diagnostics

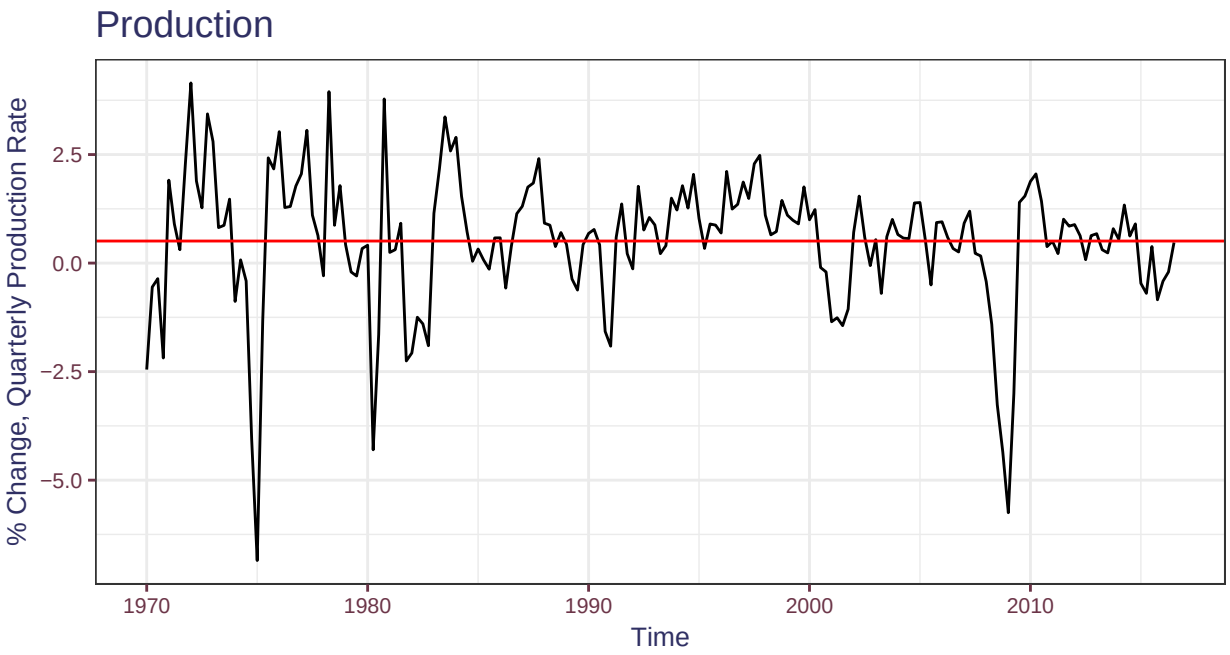
As when we fit any model with assumptions, we need to check the validity of our estimates.

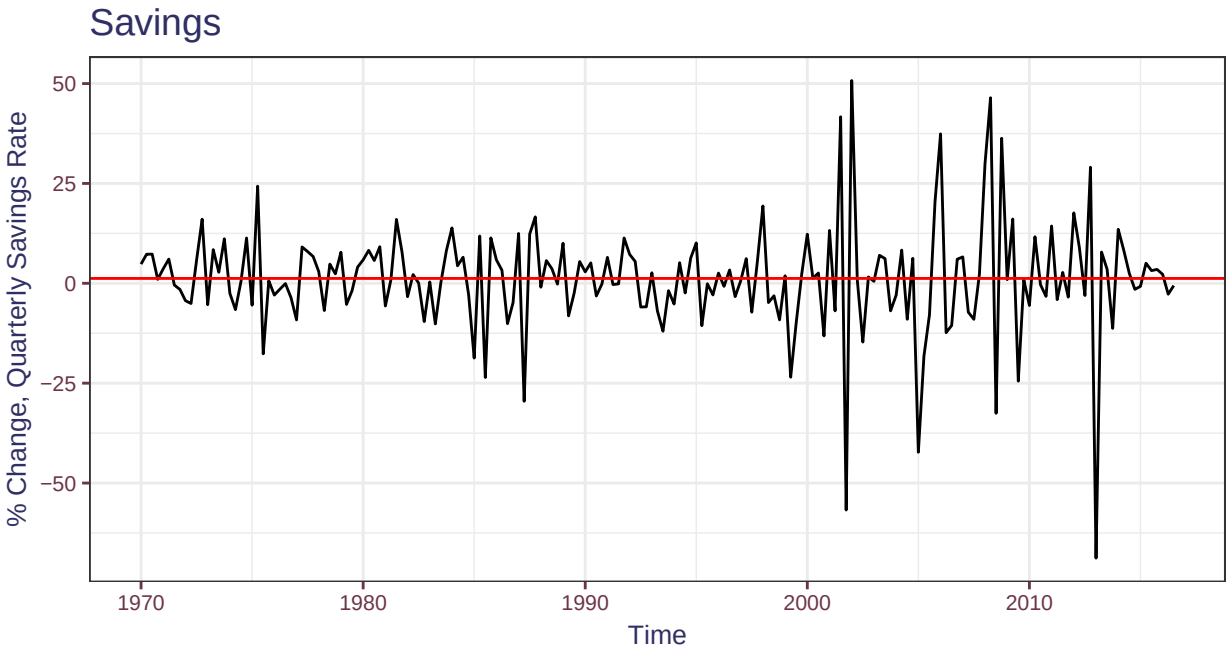
Stationarity

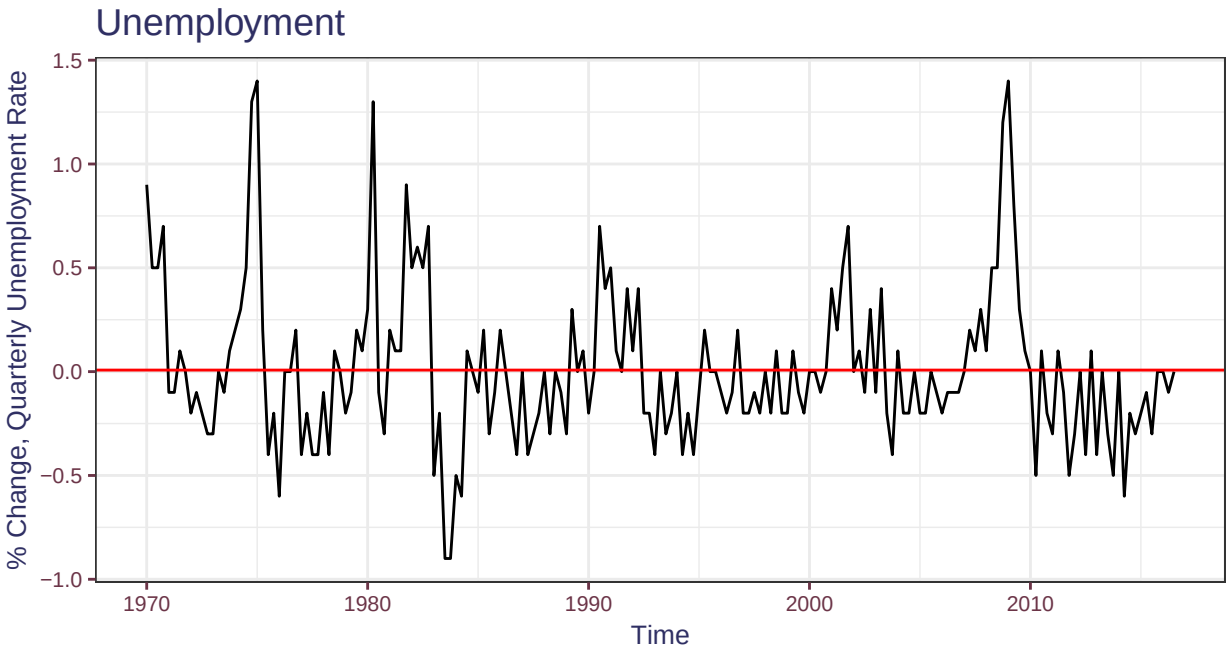
If the data are truly stationary, then the time series should not "drift" away from the global mean and the variance of the process should not change over time. The graphs below seem to indicate, at least visually, that the stationarity assumption is not violated. Mean stationarity can be tested formally with the Augmented Dickey-Fuller (ADF) test (not shown in the example here); a low p-value indicates stationarity.


Consumption
% Change, Quarterly Consumption Expenditure (USD)
2
1
0
−1
−2
1970
1980
1990
2000
2010
Time


Income
% Change, Quarterly Personal Disposable Income (USD
2.5
0.0
−2.5
1970
1980
1990
2000
2010
Time

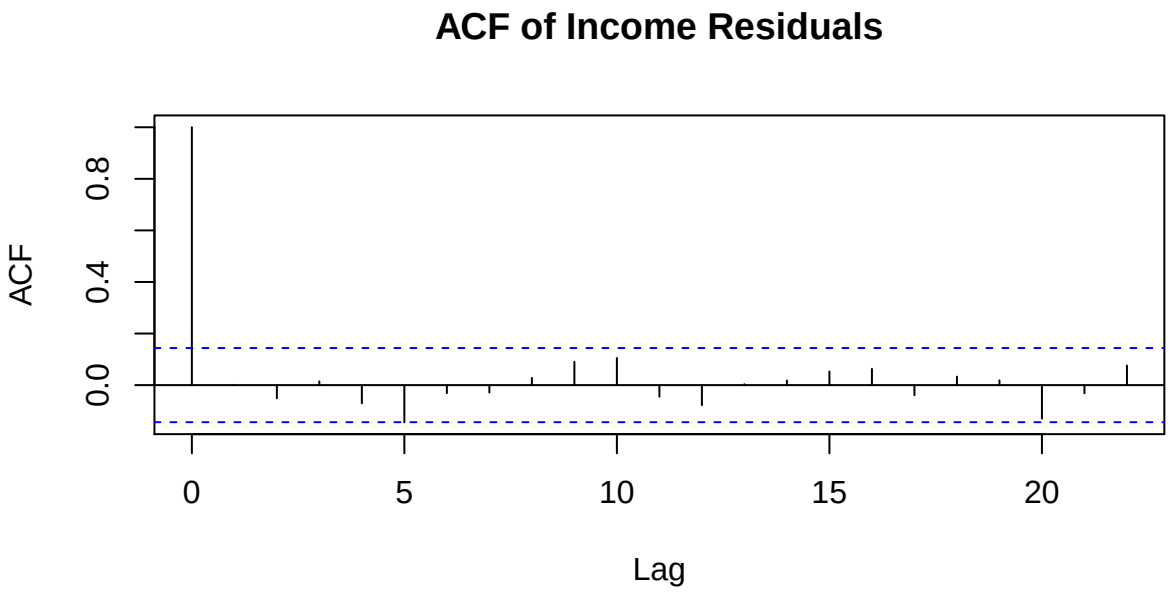

Production
% Change, Quarterly Production Rate
2.5
0.0
−2.5
−5.0
1970
1980
1990
2000
2010
Time

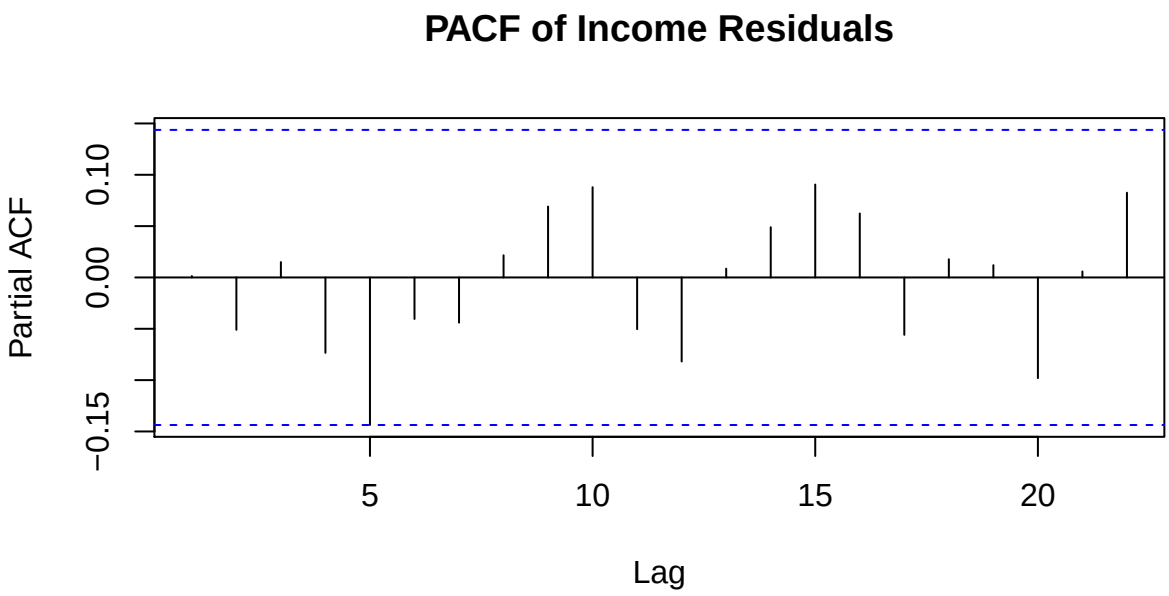

Savings
% Change, Quarterly Savings Rate
50
25
0
−25
−50
1970
1980
1990
2000
2010
Time



Autocorrelation

We also need to check that the residuals do not show any autocorrelation structure. To do this, we plot what are called the *autocorrelation function* and *partial autocorrelation function*. Autocorrelation is defined above in the introduction; partial autocorrelation is simply autocorrelation conditional on all previous time lags. The graphs simply show the autocorreation and partial autocorrelations for the residuals at different lags. If there are any large positive or negative values (besides at lag = 0) in the graph, there could still be some unexplained time dependent structure that has not been accounted for in the model as currently constructed.





Inference

Since the residuals are assumed to be distributed from a multivariate Gaussian distribution, we can use the typical linear regression hypothesis tests and p-values to evaluate the significance of the estimated parameters. As such, we can reliably use the p-values in the output above for testing individual parameters, and we an also use F-tests to test a whole set of parameters, e.g. one whole time lag or one entire variable in the model. In the model above, we can see that, for example, the percent change in personal disposable income in the current quarter is significantly related to the percent change in production rate and personal consumption expenditure in the previous quarter. On the other hand, the percent change in quarterly personal consumption expenditure for this quarter only seems to related to the change in quarterly personal consumption expenditure for the

previous quarter.

Forecasting

Predictions, along with confidence intervals, can be extracted as well:

```
$Consumption
          fcst      lower    upper       CI
[1,] 0.7384361 -0.4732810 1.950153 1.211717
[2,] 0.7462742 -0.5464069 2.038955 1.292681
[3,] 0.7465722 -0.5576102 2.050755 1.304182
[4,] 0.7471535 -0.5582624 2.052569 1.305416

$Income
          fcst      lower    upper       CI
[1,] 0.7337832 -0.9804359 2.448002 1.714219
[2,] 0.7113696 -1.1315370 2.554276 1.842907
[3,] 0.7181098 -1.1355611 2.571781 1.853671
[4,] 0.7159528 -1.1389689 2.570875 1.854922

$Production
          fcst     lower    upper       CI
[1,] 0.5055301 -1.770461 2.781522 2.275992
[2,] 0.5223765 -2.282101 3.326854 2.804477
[3,] 0.5317100 -2.440036 3.503456 2.971746
[4,] 0.5356099 -2.485071 3.556290 3.020680

$Savings
         fcst     lower    upper       CI
[1,] 1.717185 -23.75444 27.18881 25.47162
[2,] 1.117644 -25.88243 28.11772 27.00008
[3,] 1.244812 -25.95432 28.44394 27.19913
[4,] 1.200732 -26.01935 28.42081 27.22008

$Unemployment
              fcst      lower     upper        CI
[1,] 0.0012501129 -0.5922449 0.5947451 0.5934950
[2,] 0.0022514324 -0.6894025 0.6939054 0.6916540
[3,] 0.0007135897 -0.7218635 0.7232907 0.7225771
[4,] 0.0000337317 -0.7311414 0.7312089 0.7311751
```

This shows the forecast for the economic indicators for the next 4 quarters after the data set ends. Notice that the width of the confidence intervals gets wider the further out we go - this is typical, since any error in predition will propagate through time.

Assumption Violation

So what do we do if the assumptions of the VAR model are violated in one or more of the raw variables? In the case of heteroskedacity, we can attempt to apply the same tricks that we do in linear regression - some kind of transformation on the individual variables so that the variance remains constant throughout. In the case of a changing mean, we have several options:

- If the mean is changing linearly, we can use linear regression with respect to time and subtract the predicted mean out of the data.
- If the mean is changing nonlinearly, we can use some kind of nonparameteric regression method, such as a spline or a loess regression, with respect to time to predict the changing mean. Again, we then subtract the mean out of the data.
- We can estimate the changing mean using some kind of smoothing algorithm, such as kernel smoothing, exponential smoothing, or Holt's method, then subtract that mean out of the data.
- We can apply differencing, which simply means taking the difference between consecutive observations. This will create a new variables $\Delta Y_{t-1} = Y_t - Y_{t-1}, t \in 2, ..., n$ that we can then use.