Homework 4 Solution

Ian Cranfield SID [redacted]

Due 5/21/2020 5pm

```
library(dynlm)
library(pastecs)
library(openxlsx)
library(tidyverse)
library(data.table)
library(scales)
library(TSstudio)
library("readxl")
library(ggplot2)
require(cowplot)
library(psych)
library(corrplot)
library(gtable)
library(timeDate)
library(PerformanceAnalytics)
library(fpp2)

library(seasonal)
library(lubridate)
library(dplyr)
library(forecast)
library(timeSeries)
library(tseries)
library(strucchange)
library(tis)
require("datasets")
require(graphics)
library(vars)
library(lmtest)
```

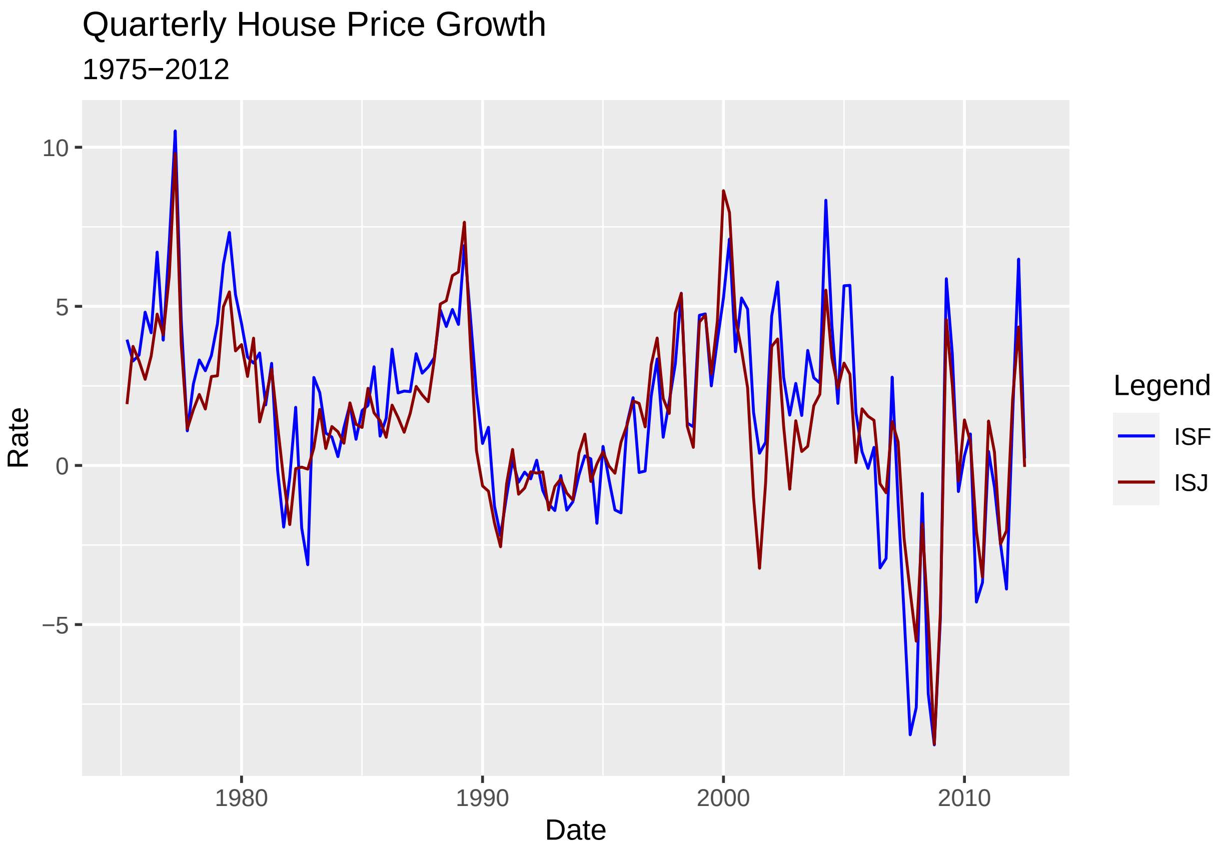
Problem 11.1

```
mac <- read_excel("Chapter11_exercises_data.xls", sheet = 1)[,1:7] %>%
  mutate(Date = as.Date(as.yearqtr(Date, format = "%YQ%q")),
         GSF = as.numeric(GSF),
         GSJ = as.numeric(GSJ),
         GAL = as.numeric(GAL)
         ) %>% na.omit()
```

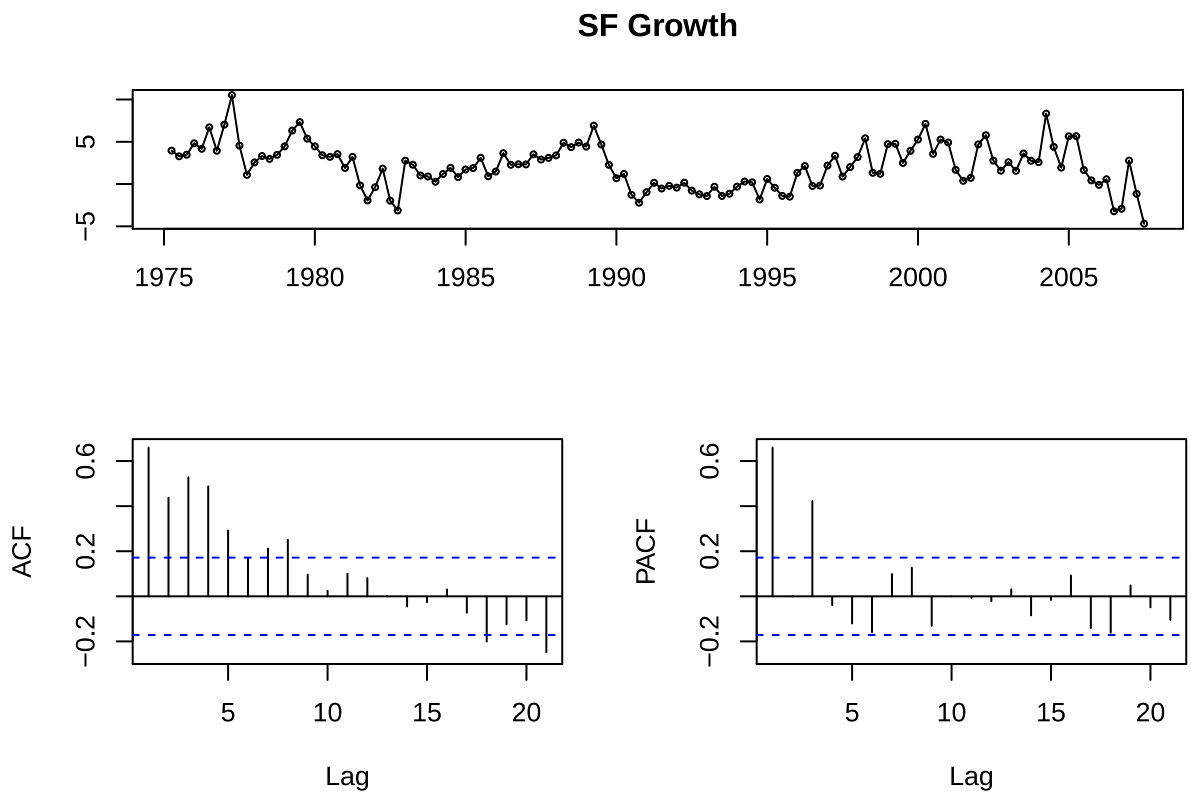
```
## New names:
```

```
## * `` -> ...8
## * `` -> ...9
```

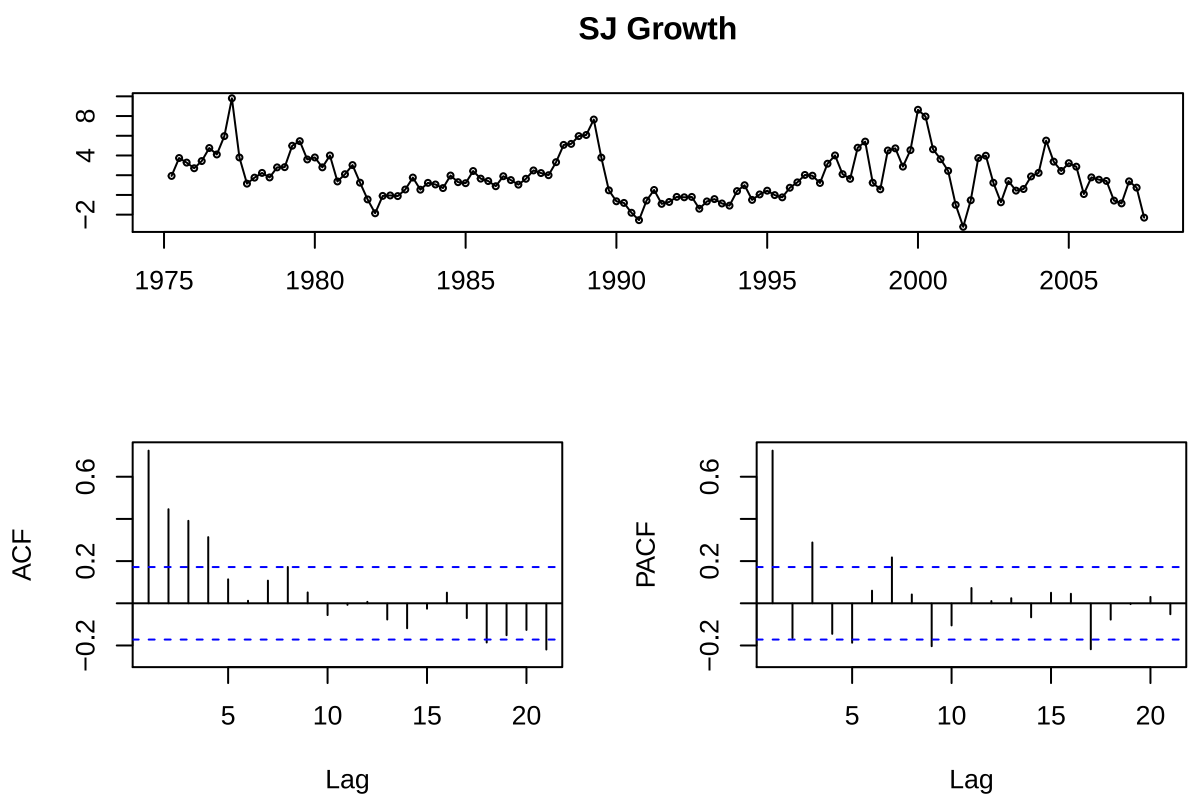
```
SF_growth <- ts(mac$GSF, start = 1975.25, end = 2007.50 , frequency = 4)

SJ_growth <- ts(mac$GSJ, start = 1975.25, end = 2007.50, frequency = 4)

AL_growth <- ts(mac$GAL, start = 1975.25, frequency = 4)


cols2 <- c("ISF"="blue", "ISJ"="dark red")
ggplot(mac, aes(x=Date)) +
  geom_line(aes(y=GSF, color="ISF")) +
  geom_line(aes(y=GSJ, color="ISJ")) +
  ggtitle("Quarterly House Price Growth", "1975-2012") +
  xlab("Date") +
  ylab("Rate") +
  scale_color_manual("Legend", values=cols2)
```



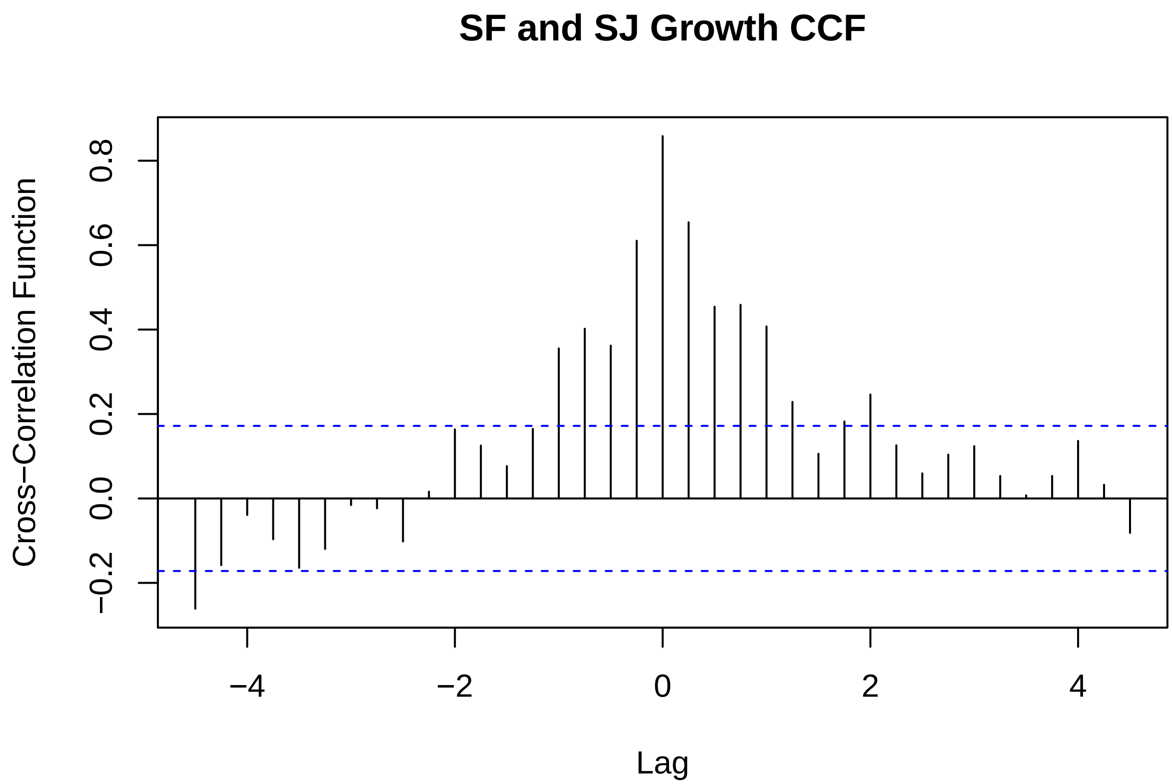

```
tsdisplay(SF_growth , main = "SF Growth")
```




```r
tsdisplay(SJ_growth , main = "SJ Growth")
```



```r
ccfvalues <- ccf(SF_growth, SJ_growth, ylab="Cross-Correlation Function", main = "SF and SJ Growth CCF")
```



The time series shows us that the Price Growth in San Francisco and in San Jose are highly associated with each other. It also looks like that the Housing Price Growth in SF comes before SJ suggesting that SF prices have some causation on SJ.

The ACF of GSF decays slowly to 0, while displaying some seasonality. The PACF shows significant lags at Q2 and Q4 (Lag 1 and 3).

The ACF of GSJ decays slowly to 0, and the PACF shows significant lags at Q2 and Q4 (Lag 1 and 3).

The Cross-Correlation Function of Growth Prices in SF and SJ show a high positive correlation at 0, and slightly higher significance with positive lags increasing indicating that San Francisco Price Growth most likely leads to San Jose housing price growth. Indicates one leads the other in high price growth. Maximally correlated when SF_growth at lag 0 .

```r
sf.sj <- data.frame((cbind(SF_growth[1:130], SJ_growth[1:130]))) %>% set_names("GSF", "GSJ")

VARselect(sf.sj)
```

```
## $selection
## AIC(n)  HQ(n)  SC(n) FPE(n)
##      3      3      3      3
##
## $criteria
##                1        2        3        4        5        6        7        8
## AIC(n) 1.260763 1.259156 1.003914 1.028296 1.034471 1.071311 1.014798 1.077252
## HQ(n)  1.317364 1.353491 1.135982 1.198098 1.242007 1.316580 1.297802 1.397989
## SC(n)  1.400138 1.491447 1.329121 1.446420 1.545511 1.675267 1.711671 1.867041
## FPE(n) 3.528188 3.522788 2.729665 2.797876 2.816523 2.924191 2.766069 2.947888
##                9       10
```

```
## AIC(n) 1.071033 1.080588
## HQ(n)  1.429504 1.476793
## SC(n)  1.953739 2.056210
## FPE(n) 2.934116 2.967940
```

```r
house.model <- VAR(sf.sj, p = 3)

summary(house.model)
```

```
##
## VAR Estimation Results:
## =========================
## Endogenous variables: GSF, GSJ
## Deterministic variables: const
## Sample size: 127
## Log Likelihood: -418.202
## Roots of the characteristic polynomial:
## 0.9124 0.6813 0.6813 0.5674 0.319 0.319
## Call:
## VAR(y = sf.sj, p = 3)
##
##
## Estimation results for equation GSF:
## ====================================
## GSF = GSF.l1 + GSJ.l1 + GSF.l2 + GSJ.l2 + GSF.l3 + GSJ.l3 + const
##
##        Estimate Std. Error t value Pr(>|t|)
## GSF.l1  0.27804    0.12489   2.226   0.0279 *
## GSJ.l1  0.65073    0.14912   4.364 2.72e-05 ***
## GSF.l2 -0.16813    0.13269  -1.267   0.2076
## GSJ.l2 -0.33706    0.17370  -1.940   0.0547 .
## GSF.l3  0.57698    0.12499   4.616 9.88e-06 ***
## GSJ.l3 -0.08235    0.15007  -0.549   0.5842
## const   0.16928    0.21193   0.799   0.4260
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 1.653 on 120 degrees of freedom
## Multiple R-Squared: 0.6244,  Adjusted R-squared: 0.6056
## F-statistic: 33.25 on 6 and 120 DF,  p-value: < 2.2e-16
##
##
## Estimation results for equation GSJ:
## ====================================
## GSJ = GSF.l1 + GSJ.l1 + GSF.l2 + GSJ.l2 + GSF.l3 + GSJ.l3 + const
##
##        Estimate Std. Error t value Pr(>|t|)
## GSF.l1 -0.06316    0.10844  -0.583 0.561319
## GSJ.l1  1.00212    0.12947   7.740 3.48e-12 ***
## GSF.l2 -0.19186    0.11521  -1.665 0.098453 .
## GSJ.l2 -0.29834    0.15081  -1.978 0.050200 .
## GSF.l3  0.37086    0.10852   3.417 0.000864 ***
## GSJ.l3 -0.01054    0.13030  -0.081 0.935652
## const   0.28633    0.18401   1.556 0.122327
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 1.435 on 120 degrees of freedom
## Multiple R-Squared: 0.637,   Adjusted R-squared: 0.6189
## F-statistic:  35.1 on 6 and 120 DF,  p-value: < 2.2e-16
##
##
##
## Covariance matrix of residuals:
##       GSF   GSJ
## GSF 2.733 1.688
## GSJ 1.688 2.060
##
## Correlation matrix of residuals:
##        GSF    GSJ
## GSF 1.0000 0.7112
## GSJ 0.7112 1.0000
```

```r
#SF explaining sj is more significant than sj explaining sf
```

`VARselect()` suggests to choose a model of up to 3 lags, which is consistent with the ACF and PACF plots. I chose the VAR model up to 3 lags because it had the smallest AIC and SC scores. I only included 130 observations for the data fram to reserve the last 20 observations for prediction exercises.

Problem 11.2

```r
grangertest(sf.sj$GSF ~ sf.sj$GSJ, order = 3)
```

```
## Granger causality test
##
## Model 1: sf.sj$GSF ~ Lags(sf.sj$GSF, 1:3) + Lags(sf.sj$GSJ, 1:3)
## Model 2: sf.sj$GSF ~ Lags(sf.sj$GSF, 1:3)
##   Res.Df Df      F    Pr(>F)
## 1    120
## 2    123 -3 6.5504 0.0003863 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
grangertest(sf.sj$GSJ ~ sf.sj$GSF, order = 3)
```

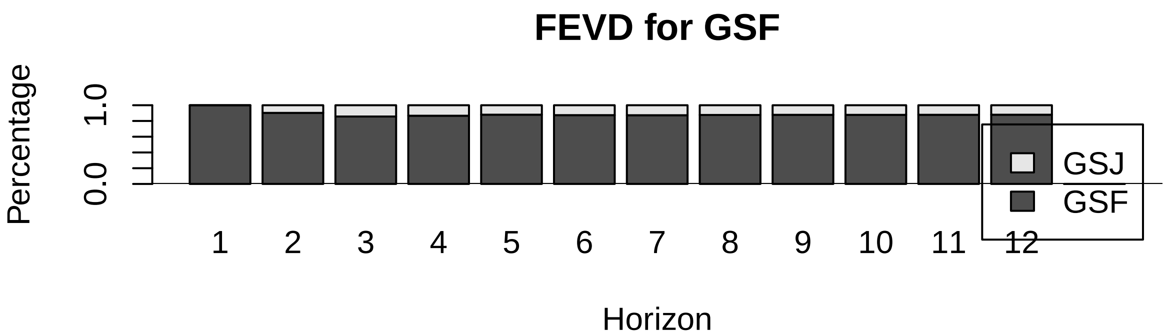
```
## Granger causality test
##
## Model 1: sf.sj$GSJ ~ Lags(sf.sj$GSJ, 1:3) + Lags(sf.sj$GSF, 1:3)
## Model 2: sf.sj$GSJ ~ Lags(sf.sj$GSJ, 1:3)
##   Res.Df Df      F  Pr(>F)
## 1    120
## 2    123 -3 4.0141 0.00921 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

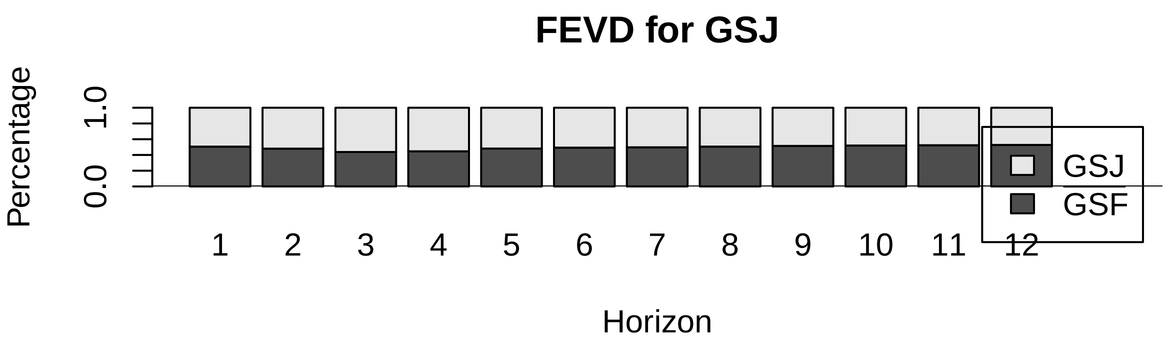
The first equation (`GSF ~ GSJ`) tests the effect of the San Jose Market on San Francisco's market. It has a large F-value and a really **low p-value of 0.00038** therefore we **reject the null hypothesis** and conclude that the San Jose market Granger-causes the San Francisco market, which means that San Jose market should have predictive ability for the San Francisco market.

The second equation (`GSJ ~ GSF`) tests the effect of the San Francisco Market on San Jose's Market. It has a large F-value and **low p-value of 0.009** therefore we **reject the null hypothesis** and conclude that San Francisco's Market Granger-causes San Jose's Market, which means that San Francisco also has predictive ability for the San Jose market.

This shows us that both markets are interdependent and a positive or negative shocks to either market will be transmitted to the other.

```
plot(fevd(house.model, n.ahead=12))
```

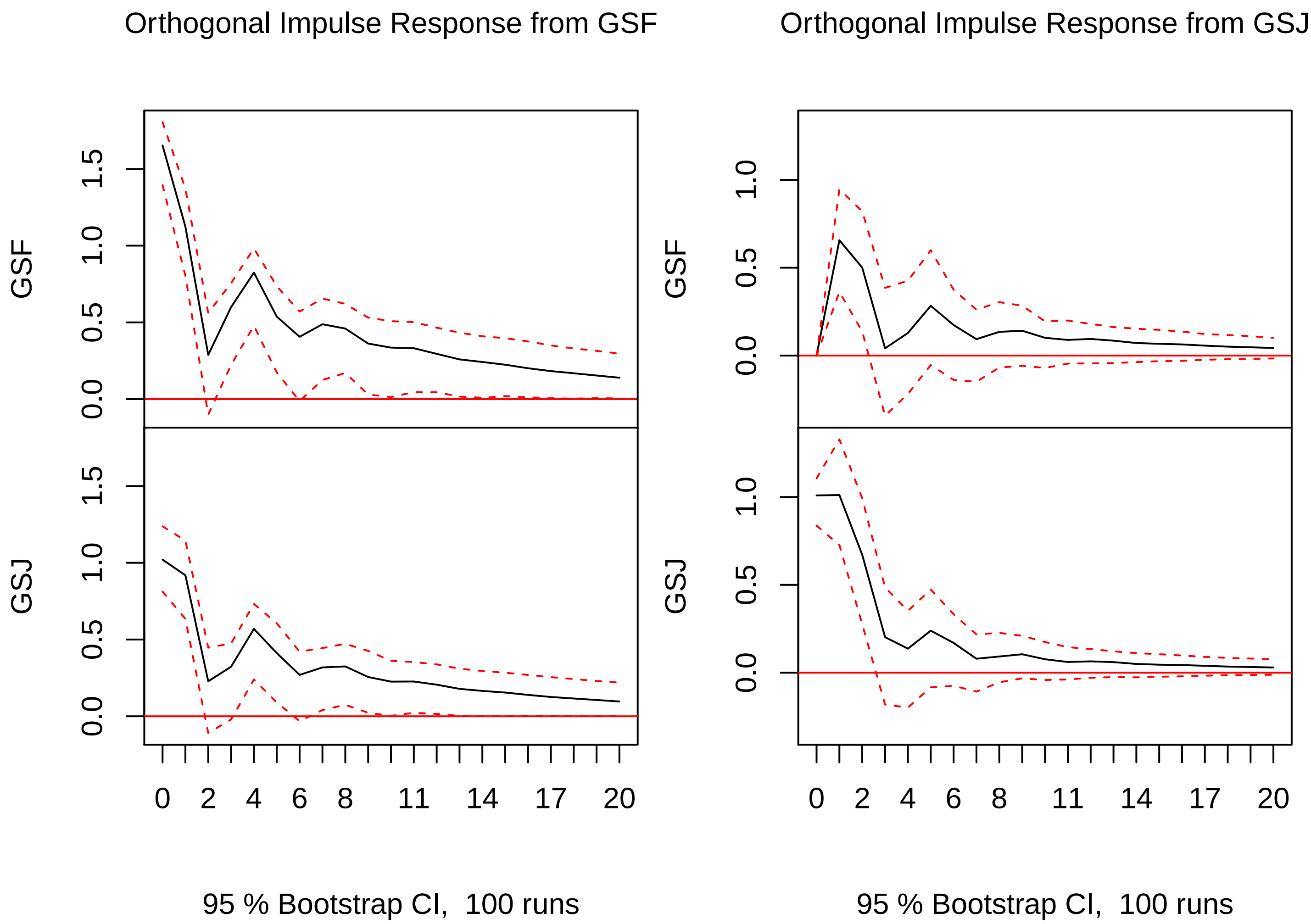




Variance Decomposition (Forecast Error Variance Decomposition) tells me more about the causality.

Problem 11.3

```
plot(irf(house.model, n.ahead = 20))
```



Looking at the IRF for GSF, a San Francisco shock affects the San Jose market for about 2 quarters and then goes to 0.

Looking at the IRF for GSJ, a San Jose shock affects the San Francisco market for longer periods, at about 8 or 9 quarters. For GSJ on GSJ there is a large effect from the shock which then sharply goes down and then goes to 0.

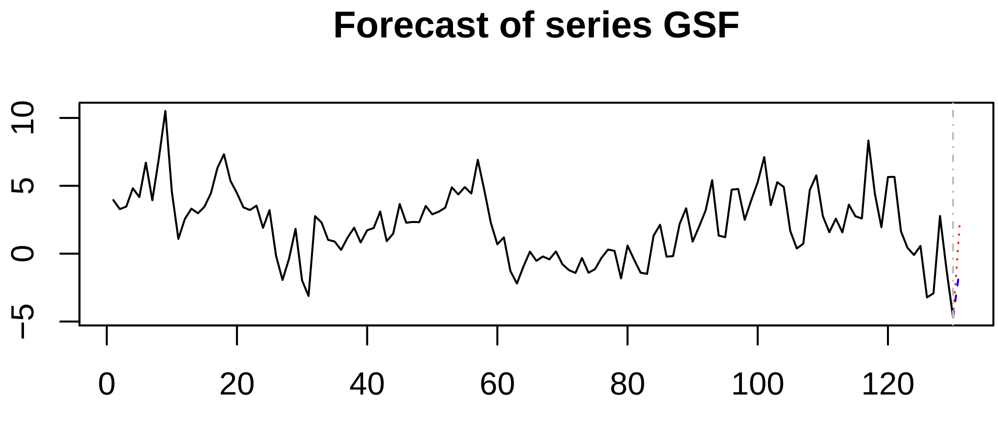
This information agrees with our findings regarding the estimation results of the VAR and the tests of Granger-causality that each market are affected by each other.

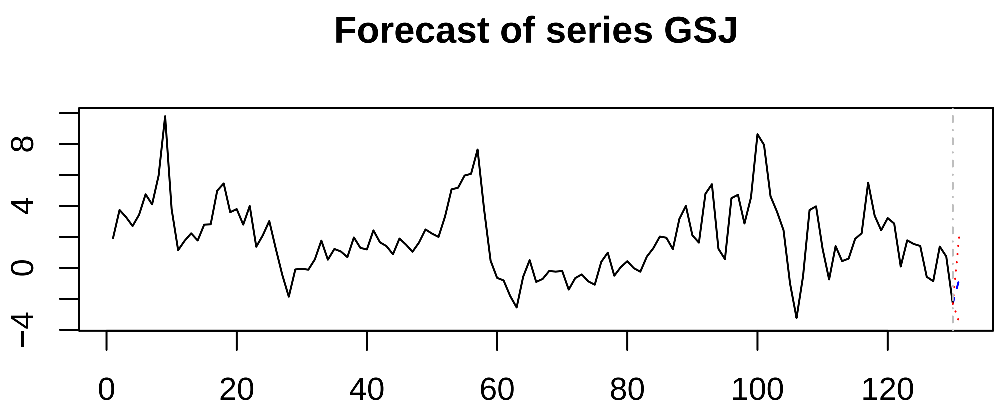
As higher prices and price growth in San Francisco's housing market increase, people tend to look at housing opportunities in San Jose because of more affordable housing. So from this, the price growth in San Jose is affected by the influx of people moving to that area so thats why we see an effect to the San Jose market at shorter lags when there is a shock in San Francisco's market.

San Jose attracts a lot of economic activity as well so there is some of the same migration going towards San Jose and San Francisco; however, it is much less so a shock in the San Jose market will effect the San Francisco market at later lags.

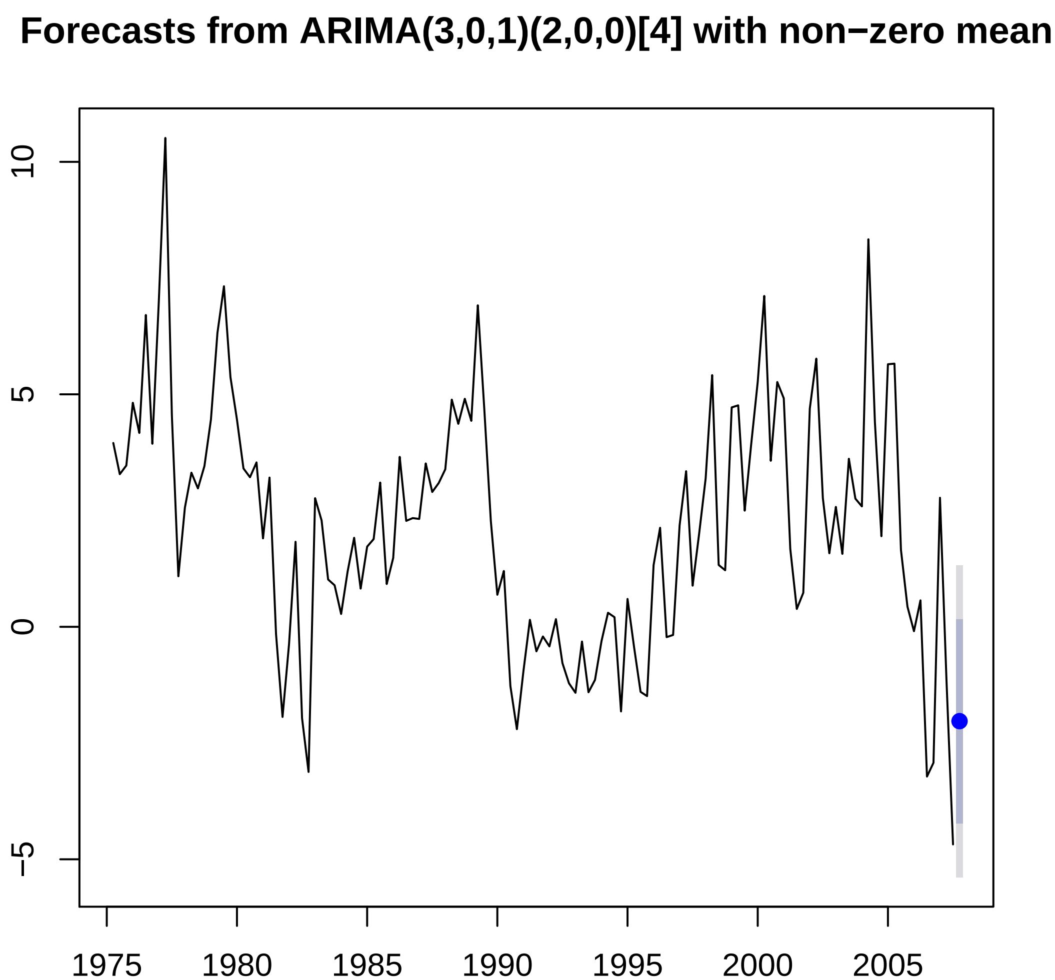
Problem 11.4

```r
predict <- predict(object = house.model, n.ahead = 1)

plot(predict)
```

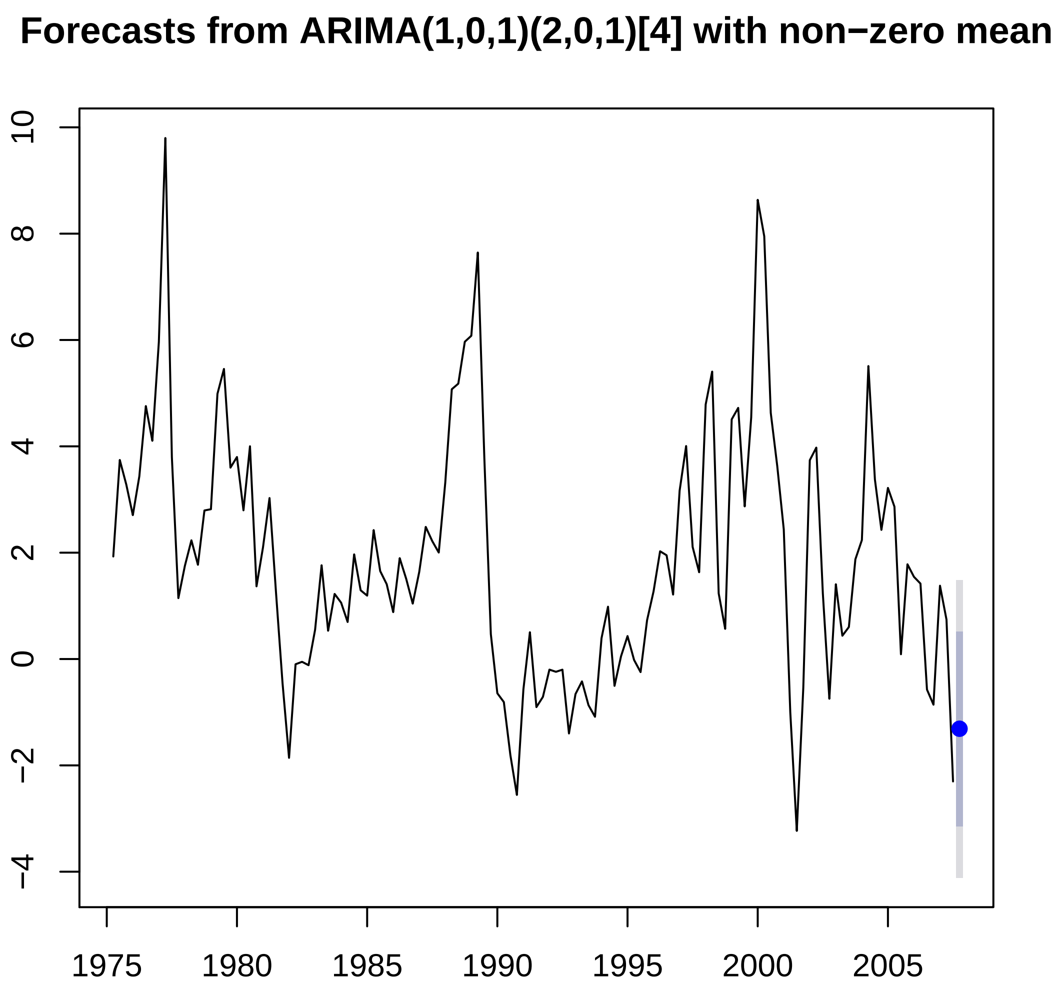





```r
fit1=auto.arima(SF_growth)

plot(forecast(fit1, h = 1))
```




```r
fit2 <- auto.arima(SJ_growth)

plot(forecast(fit2, h = 1))
```




```r
accuracy(fit1)
```

```
##                       ME     RMSE     MAE      MPE     MAPE      MASE
## Training set -0.02211634 1.664892 1.29464 46.56686 114.8161 0.6656678
##                     ACF1
## Training set 0.002764769
```

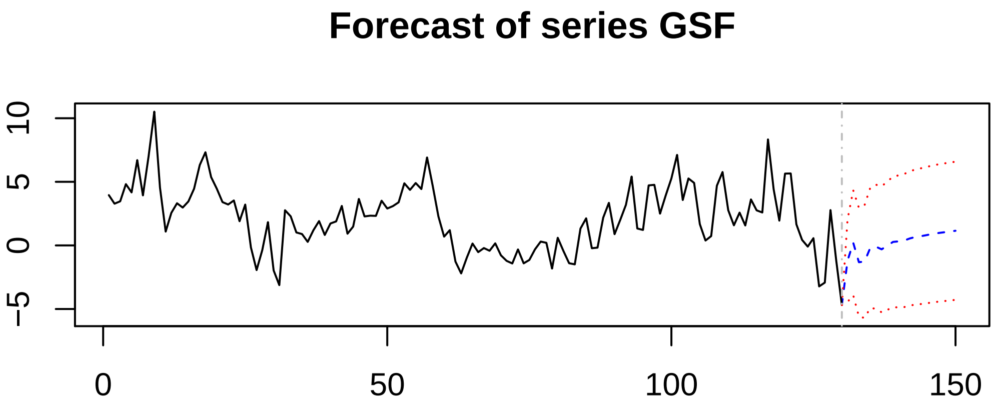
```r
accuracy(fit2)
```

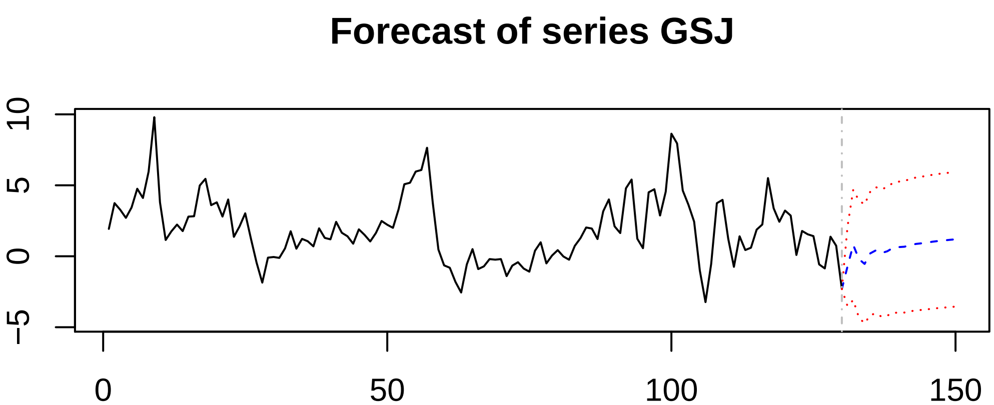
```
##                       ME     RMSE      MAE     MPE     MAPE      MASE
## Training set 0.005919342 1.394819 1.073992 56.2793 168.3939 0.5529809
##                     ACF1
## Training set -0.01843408
```

Forecasts of the VAR worked the best because the data is of stationary and ARIMA works well on non-stationary data. The MAPE scores of the ARIMA functions are high so the forecasts aren't strong.

Problem 11.5

```r
predict.multi <- predict(object = house.model, n.ahead = 20)

plot(predict.multi)
```



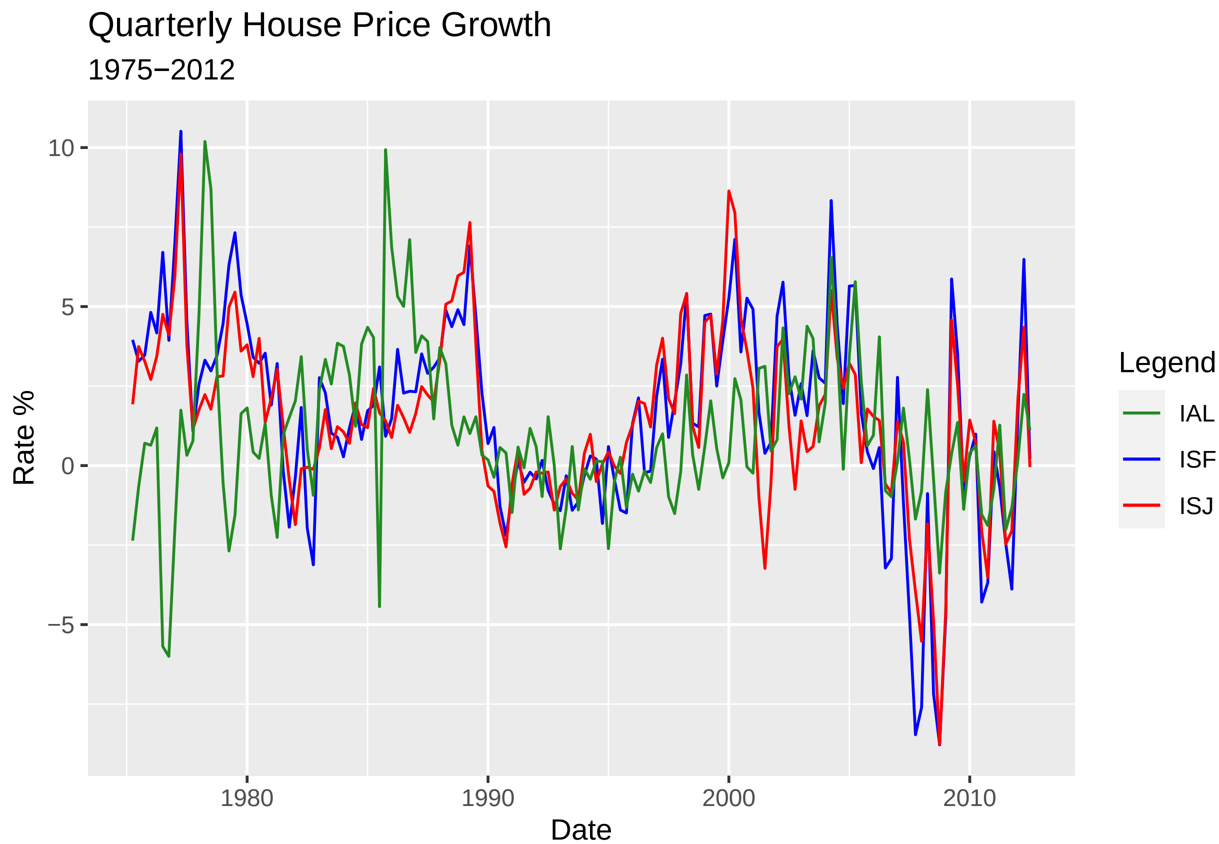


The multi-step forecasts do a good job of capturing the mean reversion; however, it did not track the affects of the great recession in 2008. The 95% confidence intervals are wide because of the uncertainty in both markets.

*Finding the growth rates and splitting into estimation and prediction for MSA3 is done in 11.6**

Problem 11.6

```
cols3 <- c("ISF"="blue", "ISJ"="red", "IAL" = "forest green")
ggplot(mac, aes(x=Date, color = factor(cols3))) +
  geom_line(aes(y=GSF, color="ISF")) +
  geom_line(aes(y=GSJ, color="ISJ")) +
  geom_line(aes(y=GAL, color = "IAL")) +
  ggtitle("Quarterly House Price Growth", "1975-2012") +
  xlab("Date") +
  ylab("Rate %") +
  scale_color_manual("Legend", values = cols3)
```



Choosing MSA3 as Albany-Schenectady-Troy in New York which is on the east coast and MSA1 and MSA2 which are on the West Coast and only about 50 miles apart.

Looking at the plot of the 3 MSAs, we can see that MSA3 (Albany) does not have strong co-movements with the other two MSAs in California. But, we can see that the great reccession in 2008 affected each market similarly.

```
AL.SF <- data.frame(cbind(mac$GSF[1:130], mac$GAL[1:130])) %>% set_names("GSF", "GAL" )

VARselect(AL.SF)
```

```
## $selection
## AIC(n)  HQ(n)  SC(n) FPE(n)
##      3      3      3      3
##
## $criteria
##                 1         2         3         4         5         6         7
## AIC(n)  2.808807  2.823468  2.575464  2.601271  2.619242  2.642840  2.662492
## HQ(n)   2.865407  2.917802  2.707532  2.771073  2.826778  2.888110  2.945496
## SC(n)   2.948181  3.055759  2.900671  3.019395  3.130282  3.246797  3.359365
## FPE(n) 16.590453 16.836758 13.140897 13.488472 13.739489 14.077073 14.369689
##                 8         9        10
## AIC(n)  2.707149  2.718237  2.741109
## HQ(n)   3.027886  3.076708  3.137313
## SC(n)   3.496938  3.600943  3.716731
## FPE(n) 15.044103 15.235233 15.617448
```

```
house.model1 <- VAR(AL.SF, p = 3)
```

```r
summary(house.model1)
```

```
##
## VAR Estimation Results:
## =========================
## Endogenous variables: GSF, GAL
## Deterministic variables: const
## Sample size: 127
## Log Likelihood: -521.412
## Roots of the characteristic polynomial:
## 0.9148 0.772 0.772 0.6539 0.2809 0.2809
## Call:
## VAR(y = AL.SF, p = 3)
##
##
## Estimation results for equation GSF:
## ====================================
## GSF = GSF.l1 + GAL.l1 + GSF.l2 + GAL.l2 + GSF.l3 + GAL.l3 + const
##
##        Estimate Std. Error t value Pr(>|t|)
## GSF.l1  0.69250    0.08284   8.360 1.30e-13 ***
## GAL.l1 -0.04570    0.07034  -0.650  0.51715
## GSF.l2 -0.26354    0.10115  -2.606  0.01033 *
## GAL.l2 -0.08384    0.07942  -1.056  0.29323
## GSF.l3  0.40368    0.08481   4.760 5.46e-06 ***
## GAL.l3  0.18768    0.06956   2.698  0.00798 **
## const   0.22131    0.23401   0.946  0.34620
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 1.728 on 120 degrees of freedom
## Multiple R-Squared: 0.5896,  Adjusted R-squared: 0.5691
## F-statistic: 28.74 on 6 and 120 DF,  p-value: < 2.2e-16
##
##
## Estimation results for equation GAL:
## ====================================
## GAL = GSF.l1 + GAL.l1 + GSF.l2 + GAL.l2 + GSF.l3 + GAL.l3 + const
##
##        Estimate Std. Error t value Pr(>|t|)
## GSF.l1  0.16196    0.10433   1.552   0.1232
## GAL.l1  0.55006    0.08860   6.209 7.93e-09 ***
## GSF.l2 -0.30778    0.12739  -2.416   0.0172 *
## GAL.l2 -0.13363    0.10003  -1.336   0.1841
## GSF.l3  0.22737    0.10681   2.129   0.0353 *
## GAL.l3  0.17537    0.08762   2.002   0.0476 *
## const   0.36701    0.29474   1.245   0.2155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 2.176 on 120 degrees of freedom
## Multiple R-Squared: 0.3355,  Adjusted R-squared: 0.3023
```

```
## F-statistic:  10.1 on 6 and 120 DF,  p-value: 5.061e-09
##
##
##
## Covariance matrix of residuals:
##         GSF     GAL
## GSF 2.98596 0.07776
## GAL 0.07776 4.73673
##
## Correlation matrix of residuals:
##         GSF     GAL
## GSF 1.00000 0.02068
## GAL 0.02068 1.00000
```

The San Francisco Growth regression has a stronger fit with a R-Squared of 59% where the Albany regression has a much weaker fit with a R-squared of 33%. The most signigicant term in each of the regression equations are its own lagged regressors and the cross-lagged terms are less significant showing the relationship of these two markets are weaker than the relationship of San Francisco and San Jose Markets.

```r
grangertest(AL.SF$GSF ~ AL.SF$GAL, order = 3)
```

```
## Granger causality test
##
## Model 1: AL.SF$GSF ~ Lags(AL.SF$GSF, 1:3) + Lags(AL.SF$GAL, 1:3)
## Model 2: AL.SF$GSF ~ Lags(AL.SF$GSF, 1:3)
##   Res.Df Df     F  Pr(>F)
## 1    120
## 2    123 -3 2.607 0.05485 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
grangertest(AL.SF$GAL ~ AL.SF$GSF, order = 3)
```

```
## Granger causality test
##
## Model 1: AL.SF$GAL ~ Lags(AL.SF$GAL, 1:3) + Lags(AL.SF$GSF, 1:3)
## Model 2: AL.SF$GAL ~ Lags(AL.SF$GAL, 1:3)
##   Res.Df Df      F  Pr(>F)
## 1    120
## 2    123 -3 2.4765 0.06468 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The **GSF** equation has a p-value of **5.5%** so we **fail to reject** the null hypothesis that the Albany Market does not Granger-cause the San Francisco Market.

The **GAL** equation, has a p-value of **6.4%** so, at the at the 5% significance level, we **fail to reject** the null hypothesis that the San Francisco Market does not Granger-cause the Albany Market.

I conclude that both markets are not strongly associated and therefore there is not much predictive power from San Francisco Market to Albany Market.