Estimação IGP-M: Modelo ARIMA

Vinícius Barbosa Godinho

August 29, 2018

Introdução

A partir de uma série temporal, o intuito desse estudo é o de mostrar o tratamento dos dados, a partir do software R-Studio, bem como, escolher o melhor modelo de previsão ARIMA para a taxa mensal do IGP-M, a partir de 2000 e, consequentemente, estimar fora da amostra os dados um passo à frente.

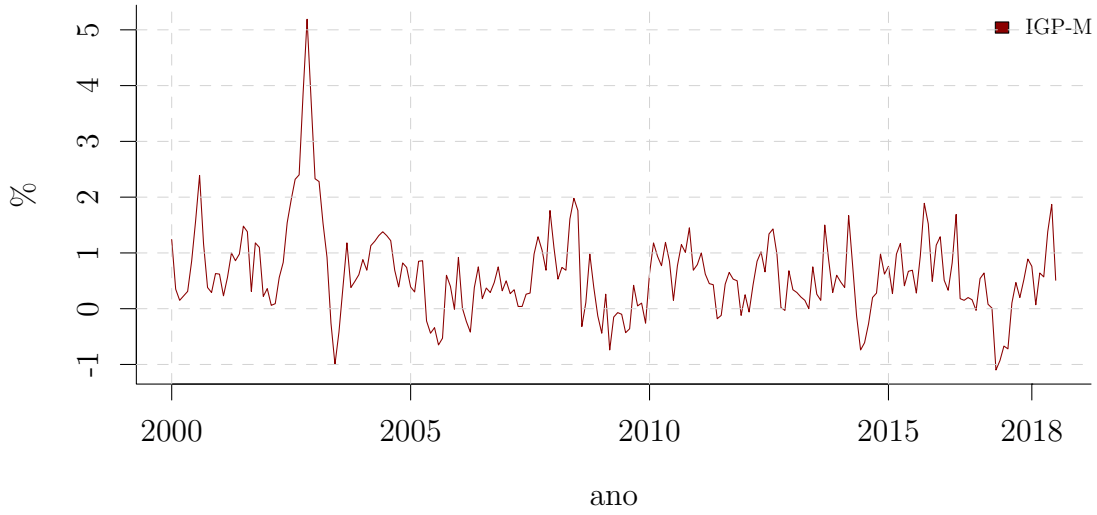
Modelos de regressão, normalmente, utilizam um método de estimação que apresenta uma relação da variável dependente com duas ou mais variáveis, exigindo equações múltiplas. Entretanto, os modelos ARIMA, são autoregressivos, ou seja, utilizam apenas a série temporal para realizar a estimativa, está baseado no comportamento da própria variável, assim não se relaciona com nenhuma outra variável a não ser o tempo.

Assim, a análise de séries temporais leva em consideração a forte relação das observações e dos erros ao longo do tempo, ou seja, é caracterizada pela violação do pressuposto de não autocorrelação. As variáveis e os resíduos não são independentes ao longo do tempo.

Análise Descritiva

Nessa seção é feita uma análise descritiva acerca da série analisada, a princípio traçamos a trajetória dos dados brutos, gráfico 1. Nessa análise é possível observar o comportamento da taxa mensal do IGP-M a partir do ano 2000, visivelmente é detectado um outlier no ano de 2004, a partir de 2005 percebe-se certa estabilidade. Para fazer uma análise estatística em séries temporais é necessário verificar, através de testes, se as variáveis possuem certas características, tais como: distribuição normal; possuem ou não componente sazonal; tendência; e raiz unitária.

Figure 1: Taxa Mensal IGP-M (2000-2018)



Fonte: Elaboração própria a partir de dados da FGV

É possível observar a partir da tabela , que a priori, uma análise das estatísticas descritivas demonstra que a série não deve possuir distribuição normal, pois sua mediana não coincide com a média e a curtose é maior que três. A série apresenta uma assimetria positiva, ou a direita, visto que a mediana é menor que

a média. Os dados abaixo apresentam a estatística descritiva dos dados em nível, analisando o coeficiente de variação mostra-se que a série não parece ser estável, já que este é de 126,27

Table 1: Estatística Descritiva

	IGPM-M
Média	0,616
Mediana	0,520
Máximo	5,19
Mínimo	-1,100
Desvio-Padrão	0,778
Coef. Variação	126,27
Assimetria	1,581
Curtose	9,546

Fonte: Elaboração própria a partir de dados FGV.

Normalidade

A análise de assimetria realizada pela estatística descritiva é confirmada pelo teste de Jarque-Bera de normalidade. Neste a hipótese nula é de que a série analisada segue uma distribuição normal. Estes valores devem ser analisados pela distribuição Qui-quadrado para 2 graus de liberdade, visto que o teste foi elaborado a partir dos coeficientes de curtose e de assimetria. Na série é observado um p-valor muito pequeno, portanto rejeita-se a hipótese nula de normalidade.

Tratamento dos Dados

Para determinar a melhor previsão para o IGP-M em um modelo ARIMA é detectado características da série, as principais são a volatilidade, correlação contemporânea. O motivo de separar a tendência de nossa variável, neste contexto, é de separar tendências de crescimento de longo prazo e variação sazonais de fenômenos exclusivamente cíclicos, ou aleatórios. Observa-se a seguinte equação:

$$Y_t = sazonalidade_t + tendencia_t + ciclo_t + \epsilon_t$$

No gráfico 2 é demonstrado a decomposição clássica da série.

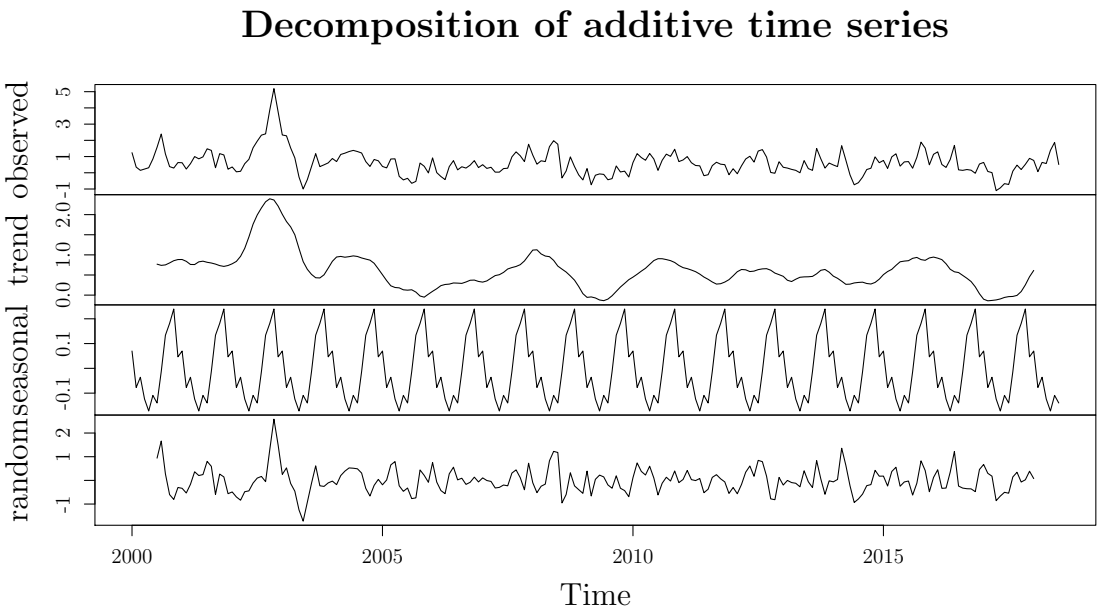
Sazonalidade

O componente sazonal, ou seja, a presença de um ciclo durante certo período pode causar instabilidade para a série, Enders (2014), destaca que utilizar uma série temporal ignorando o componente sazonal pode gerar uma variância alta na regressão. A priori, uma análise gráfica não é possível verificar a presença de sazonalidade na série, porém vamos utilizar outras ferramentas para constatar a ausência de sazonalidade. Segundo Gujarati (2003), um método de verificação do componente sazonal determinístico é inserir uma matriz de variáveis dummies mensais auxiliar de cada período de tempo regredido no modelo. Como a série analisada é mensal, o modelo testado pode ser descrito pela seguinte regressão:

$$Y_t = \alpha + \beta_2 M_2 + \beta_3 M_3 + ... * \beta_1 1 M_1 1 + \beta_1 2 M_1 2 + \epsilon_t$$

Nota-se que é incluído dummies para cada mês exceto o primeiro, para não ocorrer o problema de linearidade perfeita. Feita a regressão é efetuado um teste F, onde, $H_o : \beta_2 = \beta_3 = ... = \beta_1 1 = \beta_1 2 = 0$,

Figure 2: Decomposição Clássica IGP-M

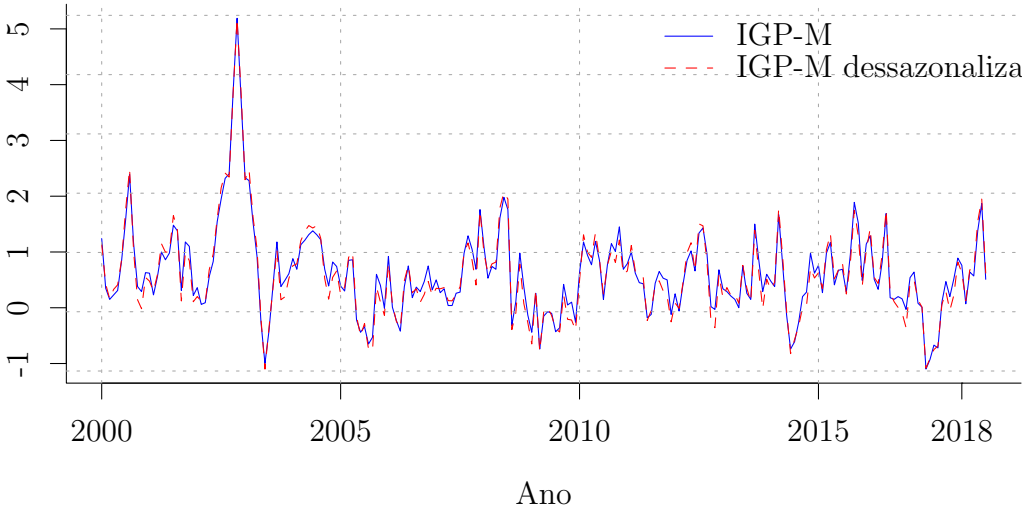


Fonte: Elaboração própria a partir de dados da FGV

ou seja é testado se todos os coeficientes mensais são, se forma conjunta, estatisticamente iguais a zero. O teste através da regressão de dummies testa a presença de uma sazonalidade mensal determinística nos dados.

É preciso verificar os p-valores de cada mês, se estes rejeitam a hipótese nula, é verificado que possuem componente sazonal. Todas as dummies aceitam a hipótese nula, portanto não é necessário dessazonalisar a série. Como forma de exercício apresentamos o gráfico 3 da série em dados brutos versus a dessazonalizada, nota-se que a trajetória é praticamente a mesma.

Figure 3: Comparação com IGP-M dessazonalizado

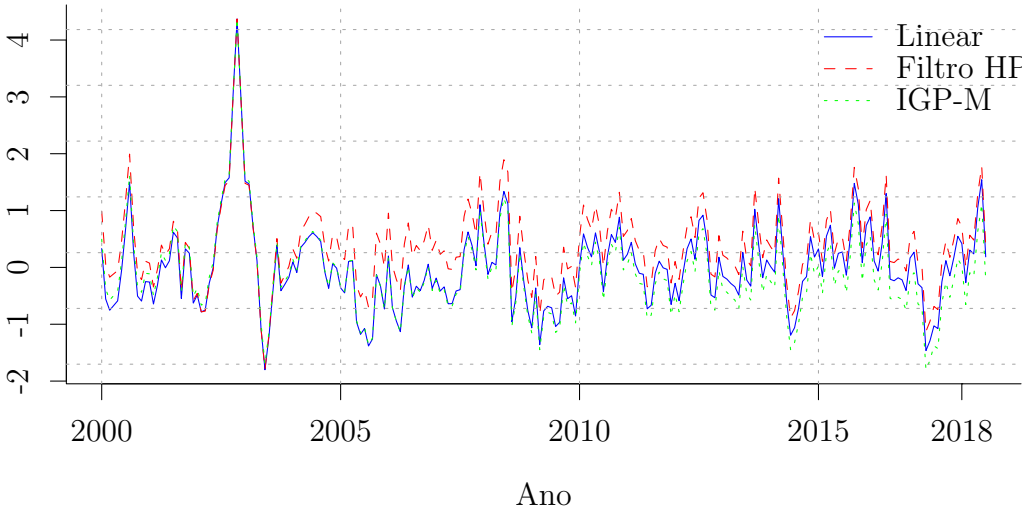


Fonte: Elaboração própria a partir de dados da FGV

Tendência

Existem várias maneiras de remover tendência de uma série, uma bastante usual é através da primeira diferença, mas como a série utilizada está em taxa, ou seja, já está diferenciada não vai ser utilizado esse método (apenas se a série for não estacionária, que será realizado testes mais para frente). Se a série apresentar uma tendência determinística de longo prazo, uma forma simples de tratamento é regredir a série contra o tempo e recolhendo os resíduos. Segundo Enders (2014) outro método de decompor a tendência é através do filtro de Hodrick and Prescott (1997). Este não assume uma tendência perfeitamente linear, e pode ser utilizado para remover tendências não determinísticas. Como podemos observar no gráfico 4 a série praticamente não possui tendência, portanto será utilizada a série bruta.

Figure 4: Métodos de remover tendência



Fonte: Elaboração própria a partir de dados da FGV

Outliers

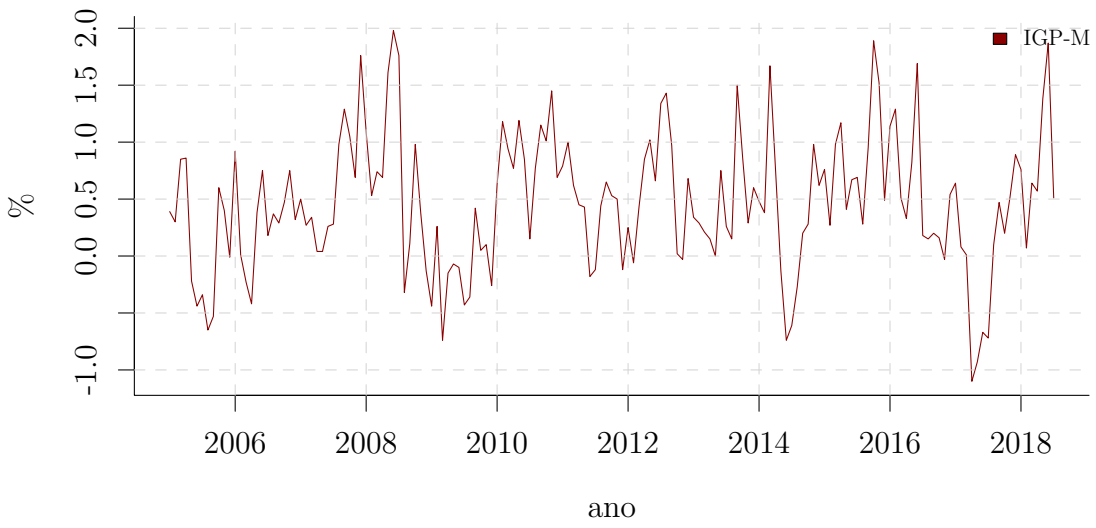
Analisando o gráfico da série, é possível verificar a presença de fortes outliers, para cima e para baixo, no ano de 2004. Portanto a série escolhida para a previsão é a partir de 2005, como apresentado no gráfico 5.

Estacionaridade

Para o caso de séries não estacionárias, é preciso determinar a ordem de integração, d , das mesmas, para se estimar um modelo ARIMA. Em suma uma série é estacionária quando a média, variância e covariância não varia no tempo, quando esta é estacionária seu valor converge para a média ao longo do tempo. Com uma série não estacionária não se pode fazer previsões, pois os valores não conhecidos podem ter caráter explosivo. Portanto, uma série estacionaria não apresenta mudança sistemática na média e na variância. As condições para que isso ocorra são:

i $E(y_t) = \mu paratodot$;

Figure 5: Série utilizada para a previsão (2005-2018)



Fonte: Elaboração própria a partir de dados da FGV

ii $Var(y_t) = E[(y_t - \mu)^2] paratodot$;

iii $Cov(y_t, y_{t-s} = Cov(y_t, y_{t+s}), s = 1, 2, 3, ..., n, paratodot$

Existem vários testes de estacionaridade, a seguir utilizaremos três dos mais usuais, o *D*ickey-Fuller, *D*ickey-Fuller Aumentado e *P*hillips-Perron, apresentando a metodologia *B*ox-Jekins.

O teste de *D*ickey-Fuller

Suponha o seguinte processo para uma série:

$$Y_t = \Theta Y_{t-1} + \varepsilon_t$$

Sendo, $\varepsilon_t \sim N(0)$, iid.

Se $|\Theta| < 1$, então o processo é estacionário, $\Theta = 1$ o processo apresenta raíz unitária, assim não estacionário. O diferencial do teste *D*ickey-Fuller, é que a determinação da ordem de integração é feita testando, na prática como as equações a seguir.

$$Y_t - Y_{t-1} = \Theta Y_{t-1} - Y_{t-1} + \varepsilon_t$$
$$\Delta Y_t = (\Theta - 1) Y_{t-1} + \varepsilon_t$$
$$\Delta Y_t = \pi Y_{t-1} + \varepsilon_t$$

A hipótese nula do teste é que $\pi = 0$. Desse modo, se $\pi = 0$, então $\Theta = 1$, consequentemente possui raíz unitária. Portanto, para testar esta hipótese é rodado a regressão da primeira diferença da série contra sua defasagem.

Listing 1: DF

```
Call:
lm(formula = diff(ig) ~ lag(ig, -1)[-length(ig)] - 1)

Residuals:
Min       1Q   Median       3Q      Max
-1.5989 -0.2178  0.0731  0.4607  1.3939

Coefficients:
Estimate Std. Error t value Pr(>|t|)
lag(ig, -1)[-length(ig)] -0.27334    0.05421  -5.042 1.23e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.519 on 161 degrees of freedom
Multiple R-squared:  0.1364,    Adjusted R-squared:  0.131
F-statistic: 25.42 on 1 and 161 DF,  p-value: 1.228e-06
```

Analisando o p-valor rejeita-se a hipótese nula, portanto a série é estacionária em nível, o que corrobora com a análise gráfica e que a série é uma taxa, portanto é como se estivesse em primeira diferença.

O teste de *Dickey-Fuller Aumentado (ADF)*

O teste ADF assume que os erros são não correlacionados, portanto no caso de processos autoregressivos esta correlação terá como consequência estimadores ineficientes e viesados. Como afirma González-Rivera (2013), a inferência só pode ser realizada se os resíduos não apresentarem autocorrelação, assim após realizar o teste ADF é necessário verificar a correlação dos resíduos através dos correlogramas. A desvantagem desse teste é determinar o número de defasagens utilizado, pois queremos utilizar o menor número possível de defasagens que garanta a não autocorrelação dos resíduos, nesse teste o número de defasagens foi escolhido pelo critério AIC, abaixo realizamos o teste com e sem constante.

Listing 2: ADF

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
Min       1Q   Median       3Q      Max
-1.61471 -0.20935  0.07579  0.46754  1.38930

Coefficients:
Estimate Std. Error t value Pr(>|t|)
z.lag.1     -0.26714    0.05924  -4.510 1.26e-05 ***
z.diff.lag  -0.02217    0.08148  -0.272    0.786
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5221 on 159 degrees of freedom
Multiple R-squared:  0.1366,    Adjusted R-squared:  0.1258
F-statistic: 12.58 on 2 and 159 DF,  p-value: 8.455e-06


Value of test-statistic is: -4.5095

Critical values for test statistics:
1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Portanto, em conformidade com o teste de *Dickey Fuller*, rejeita-se a hipótese nula, ou seja, a série é estacionária em nível. Como observado no correlograma abaixo, os resíduos não possuem autocorrelação, assim pode-se usar a inferência do teste.

Identificação dos modelos ARIMA

Para a escolha do modelo, é definido o mais parcimonioso possível, através da metodologia Box-Jenkins, o intuito é determinar os valores (p,d,q) do modelo ARIMA, segundo Moretin (1985), o procedimento de identificação consiste em duas partes:

- diferenciar a série Yt, tantas vezes quanto necessário, para se obter uma série estacionária, de modo que o processo seja reduzido a um ARMA(p,q). O número de diferenças necessárias (d) para que o processo se torne estacionário, é alcançado quando a FAC amostral decresce rapidamente para zero;

- identificar o processo ARMA(p,q) resultante, através da análise das autocorrelações e autocorrelações parciais. Pode-se verificar os gráficos das séries e das diferenças, suas médias e variâncias, suas autocorrelações e seus correlogramas com os respectivos intervalos de confiança. Os gráficos da FAC e FACP estão apresentados abaixo.

Os modelos selecionados são:
Modelo 1: AR(1) $Y_t = c + \phi_1 y_{t-1} + \varepsilon_t$
Modelo 2: MA(4) $Y_t = \mu + \Theta_1 \varepsilon_{t-1} + \Theta_2 \varepsilon_{t-2} + \Theta_3 \varepsilon_{t-3} + \Theta_4 \varepsilon_{t-4} + \varepsilon_t$
Modelo 3: ARMA(1,1) $Y_t = c + \phi_1 y_{t-1} + \Theta_1 \varepsilon_{t-1} + \varepsilon_t$
Modelo 4: ARMA(1,4) $Y_t = c + \phi_1 y_{t-1} + \Theta_1 \varepsilon_{t-1} + \Theta_2 \varepsilon_{t-2} + \Theta_3 \varepsilon_{t-3} + \Theta_4 \varepsilon_{t-4} + \varepsilon_t$

Para verificação dos modelos é necessário verificar se os resíduos da regressão são não autocorrelacionados. É usado o teste *Ljung – Box*, este tem como hipótese nula a independência de uma dada série de tempo, ou seja, os resíduos da regressão são conjuntamente não correlacionados ao longo do tempo.

Listing 3: ADF

```
> reg1.i <- arima(ig, order = c(1,0,0))
> Box.test(resid(reg1.i), lag=6, type='Ljung-Box', fitdf=1)

Box-Ljung test

data:  resid(reg1.i)
X-squared = 6.936, df = 5, p-value = 0.2254
```

```
> reg2.i <- arima(ig, order = c(0,0,4))
> Box.test(resid(reg2.i), lag=6, type='Ljung-Box', fitdf=1)

Box-Ljung test

data:  resid(reg2.i)
X-squared = 0.64082, df = 5, p-value = 0.9861

> reg3.i <- arima(ig, order = c(1,0,1))
> Box.test(resid(reg3.i), lag=6, type='Ljung-Box', fitdf=1)

Box-Ljung test

data:  resid(reg3.i)
X-squared = 4.2473, df = 5, p-value = 0.5144

> reg4.i <- arima(ig, order = c(1,0,4))
> Box.test(resid(reg4.i), lag=6, type='Ljung-Box', fitdf=1)

Box-Ljung test

data:  resid(reg4.i)
X-squared = 0.35395, df = 5, p-value = 0.9965
```

Analisando o teste *Box-Ljung*, em todos os modelos se aceita a hipótese nula, ou seja, os resíduos não são autocorrelacionados, portanto não descartamos nenhum modelo.

Para testar se os resíduos são normalmente distribuídos é utilizado o teste de *Shapiro-Wilk*, neste a hipótese nula é que as séries analisadas são normalmente distribuídas, abaixo o resultado do teste, onde aceita a normalidade em todos os modelos.

Listing 4: ADF

```
> shapiro.test(resid(reg1.i))

Shapiro-Wilk normality test

data:  resid(reg1.i)
W = 0.99166, p-value = 0.4623

> shapiro.test(resid(reg2.i))

Shapiro-Wilk normality test

data:  resid(reg2.i)
W = 0.99072, p-value = 0.3679

> shapiro.test(resid(reg3.i))

Shapiro-Wilk normality test
```

```
data:  resid(reg3.i)
W = 0.99081, p-value = 0.3766

> shapiro.test(resid(reg4.i))

Shapiro-Wilk normality test

data:  resid(reg4.i)
W = 0.99118, p-value = 0.4126
```

Como não descartamos nenhum modelo, é usado dois critérios de comparação, o *A*kaike (AIC) e o Bayesiano de *S*chwarz(BIC), quanto menor o critério maior é o poder de informação do modelo. Como é observado os critérios no AIC são bem próximos, com o melhor modelo o ARIMA(1,0,1), já pelo BIC o melhor modelo foi ARIMA(1,0,0).

Listing 5: ADF

```
> AIC(reg1.i)
[1] 236.8574
> AIC(reg2.i)
[1] 237.7422
> AIC(reg3.i)
[1] 236.4344
> AIC(reg4.i)
[1] 238.7
> BIC(reg1.i)
[1] 246.1387
> BIC(reg2.i)
[1] 256.3047
> BIC(reg3.i)
[1] 248.8094
> BIC(reg4.i)
[1] 260.3562
```

De acordo com González-Rivera(2013), uma forma de escolher o melhor modelo de previsão é através da soma do erro ao quadrado médio (SQE). Avaliando os erros de previsão para um horizonte de 11 meses, ou seja, estimação da série de janeiro de 2005 até dezembro de 2015, portanto será previsto os meses de janeiro a novembro de 2016.

Listing 6: ADF

```
> sqe1
[1] 0.2544828
> sqe2
[1] 0.2630116
> sqe3
[1] 0.2402772
> sqe4
[1] 0.2631178
>
```

Como podemos observar o menor SQE foi do modelo 3, ARIMA(1,0,1), utilizaremos ele como o modelo de previsão um passo a frente.