Modelos SARIMA e Previsões

Lucas Domingues e Lucas Moschen

November 22, 2020

Questões Teóricas

Questão 6 (Capítulo 9)

Se $Y \sim N\left(\mu, \sigma^2\right)$, então X tal que $\log X = Y$ terá uma distribuição log-normal com $E[X] = e^{\mu+\sigma^2/2}$ e $Var[X] = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$. Baseado nessa definição, se $Y_t = \log Z_t$ e Y_t é Gaussiano, então

$$\hat{Z}_t(h) = \exp\left\{\hat{Y}_t(h) + \frac{1}{2}V_y(h)\right\}$$

$$V_Z(h) = \exp\left\{2\hat{Y}_t(h) + V_y(h)\right\}\left[\exp\{V_y(h)\} - 1\right]$$

Resposta: Sabemos que a distribuição de Y_{t+h} em um processo Gaussiano é $N\left(\hat{Y}_t(h), V_Y(h)\right)$, onde

$$\hat{Y}_t(h) = E[Y_{t+h}|Y_t, Y_{t-1}, ...]$$

Como $Z_{t+h} = \log Y_{t+h}$, pelo enunciado, a distribuição de $Z_{t+h}|Z_t, Z_{t-1}, ..$ é uma log-normal, tal que:

$$\hat{Z}_t(h) = E[Z_{t+h}|Z_t, Z_{t-1}, ...] = e^{\hat{Y}_t(h)+\frac{1}{2}V_Y(h)}$$

$$V_Z(h) = e^{2\hat{Y}_t(h)+V_Y(h)}\left(e^{V_Y(h)} - 1\right)$$

utilizando os valores do valor esperado e variância descritos. Isso demonstra a relação.

Questão 8 (Capítulo 9)

Considere o problema de encontrar a previsão linear ótima (erro quadrático médio mínimo) de um processo estacionário de média zero, $\{Z_t\}$, baseado em um número finito de observações, $Z_t, ..., Z_{t-r}$. Em resumo, queremos encontrar os coeficientes a_i na fórmula de previsão

$$\hat{Z}_t(h) = a_0 Z_t + a_1 Z_{t-1} + ... + a_r Z_{t-r}$$

que fornecem erro quadrático médio mínimo.

(a) Mostre que

$$\begin{aligned} E\left[(Z_{t+h} - \hat{Z}_t(h))^2\right] &= \gamma_0 - 2\sum_{i=0}^{r} a_i\gamma_{i+h} + \sum_{i=0}^{r}\sum_{j=0}^{r} a_i a_j \gamma_{i-j} \\ &= \gamma_0 - 2\boldsymbol{a}'\gamma_r(h) + \boldsymbol{a}'\Gamma_{r+1}\boldsymbol{a}, \end{aligned}$$

onde

$$\Gamma_{r+1} = \begin{bmatrix} \gamma_0 & \gamma_1 & ... & \gamma_r \\ \gamma_1 & \gamma_0 & ... & \gamma_{r-1} \\ ... & ... & ... & ... \\ \gamma_r & \gamma_{r-1} & ... & \gamma_0 \end{bmatrix}, \gamma_r(h) = \begin{bmatrix} \gamma_h \\ \gamma_{h+1} \\ ... \\ \gamma_{h+r} \end{bmatrix}, \boldsymbol{a} = \begin{bmatrix} a_0 \\ a_1 \\ ... \\ a_r \end{bmatrix}$$

Resposta: Temos a seguinte expressão:

$$\begin{aligned}
E\left[(Z_{t+h} - \hat{Z}_t(h))^2\right] &= E\left[\left(Z_{t+h} - \sum_{i=0}^{r} a_i Z_{t-i}\right)^2\right] \\
&= E\left[Z_{t+h}^2 - 2Z_{t+h}\sum_{i=0}^{r} a_i Z_{t-i} + \left(\sum_{i=0}^{r} a_i Z_{t-i}\right)^2\right] \\
&= E[Z_{t+h}^2] - 2\sum_{i=0}^{r} a_i E[Z_{t+h}Z_{t-i}] + \sum_{i=0}^{r}\sum_{j=0}^{r} a_i a_j E[Z_{t-i}Z_{t-j}] \\
&= \gamma_0 - 2\sum_{i=0}^{r} a_i \gamma_{h+i} + \sum_{i=0}\sum_{j=0} a_i a_j \gamma_{i-j} \\
&= \gamma_0 - 2\boldsymbol{a}'\gamma_r(h) + \boldsymbol{a}'\Gamma_{r+1}\boldsymbol{a}
\end{aligned}$$

onde a última igualdade vem da definição dada pelo enunciado e usando a notação de produto matricial.

(b) Encontre os a_i que minimizam o EQM e mostre que as equações resultantes são $\Gamma_{r+1}\boldsymbol{a} = \gamma_r(h)$.

Resposta: Queremos minimizar a expressão

$$\gamma_0 - 2\boldsymbol{a}^T\gamma_r(h) + \boldsymbol{a}^T\Gamma_{r+1}\boldsymbol{a}$$

Equivalente a minimizar

$$f(\boldsymbol{a}) = -2\boldsymbol{a}^T\gamma_r(h) + \boldsymbol{a}^T\Gamma_{r+1}\boldsymbol{a}$$

Para isso, fazemos

$$\nabla f(\boldsymbol{a}) = -2\gamma_r(h) + 2\Gamma_{r+1}\boldsymbol{a} = 0 \implies \Gamma_{r+1}\boldsymbol{a} = \gamma_r(h)$$

Vamos verificar a segunda derivada

$$\text{Hess}(f(a)) = 2\Gamma_{r+1}$$

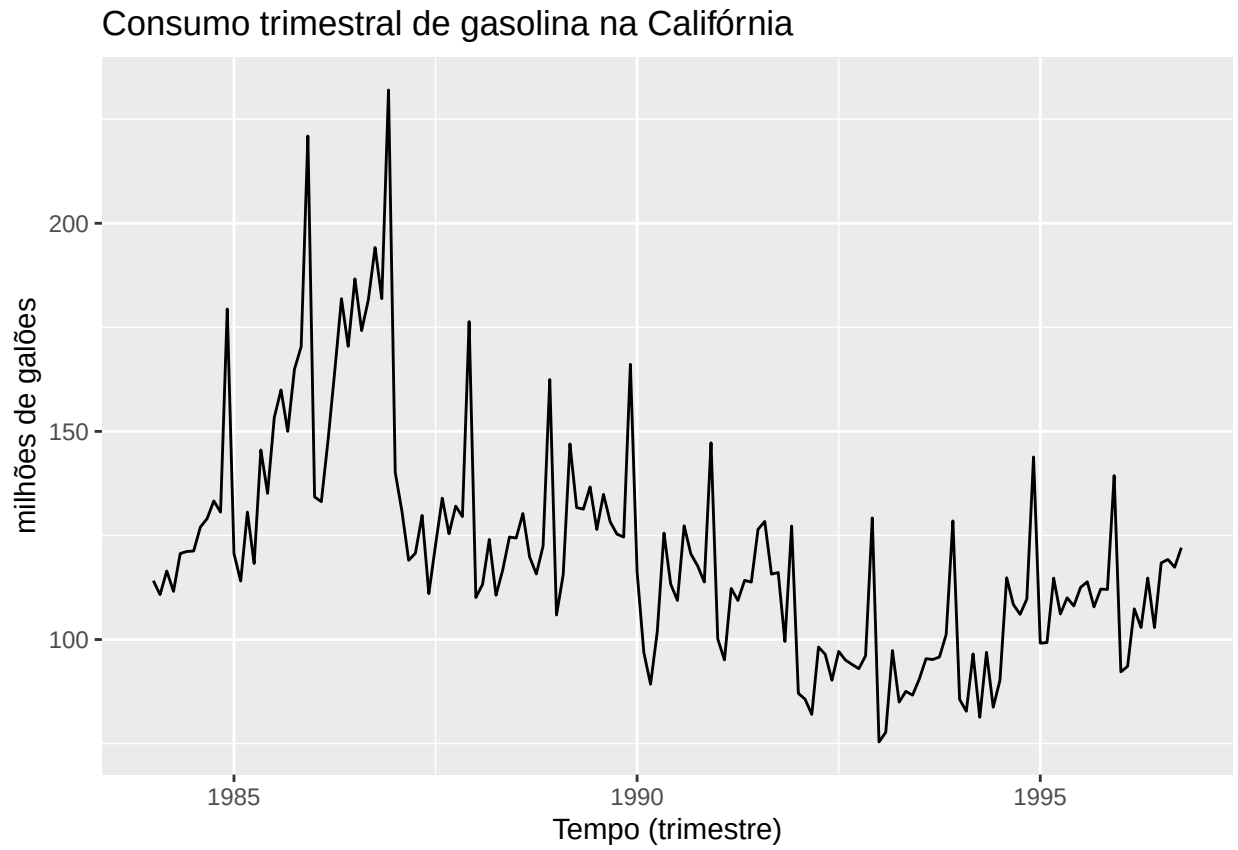
Como a matriz de covariância é positiva semi-definida e, nesse caso, definida, temos um mínimo local quando

$$\Gamma_{r+1}\boldsymbol{a} = \gamma_r(h),$$

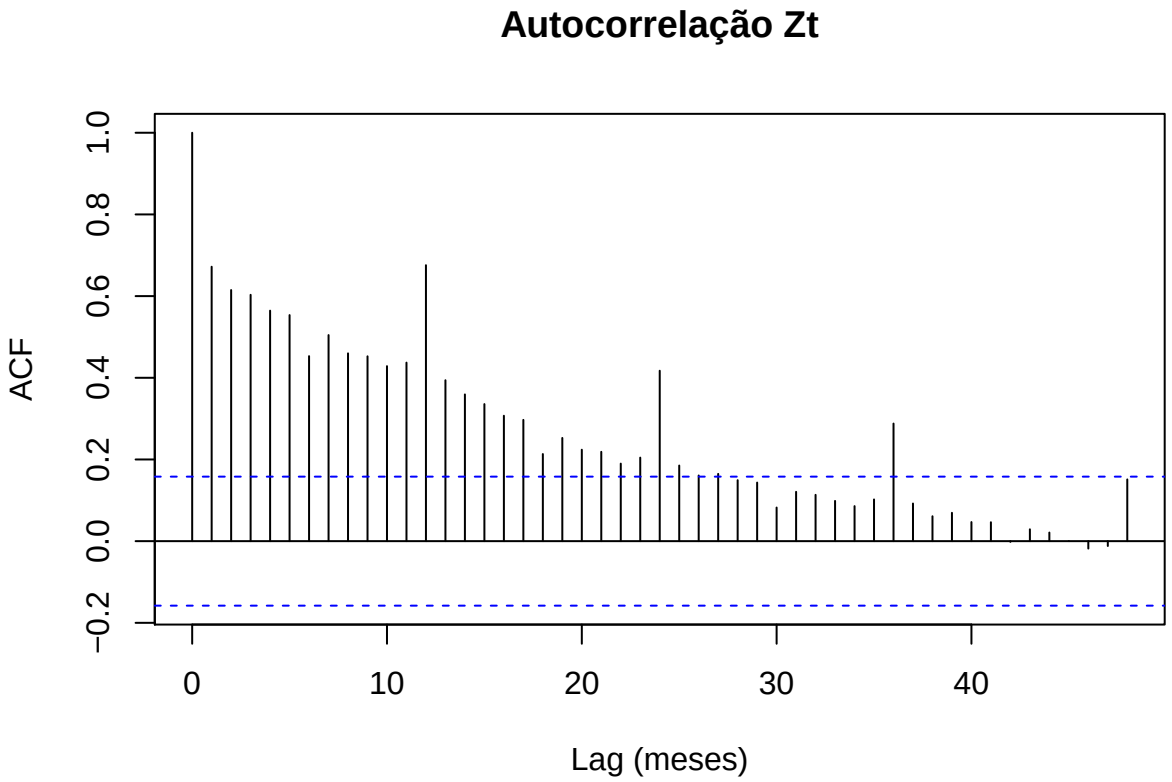
como queríamos provar.

Série Consumo - Questão 3 (Capítulo 10)

Usando um programa de computador apropriado, obtenha as autocorrelações estimadas para $Z_t, \Delta Z_t, \Delta_4 Z_t, \Delta\Delta_4 Z_t$, sendo Z_t a série de consumo de gasolina.

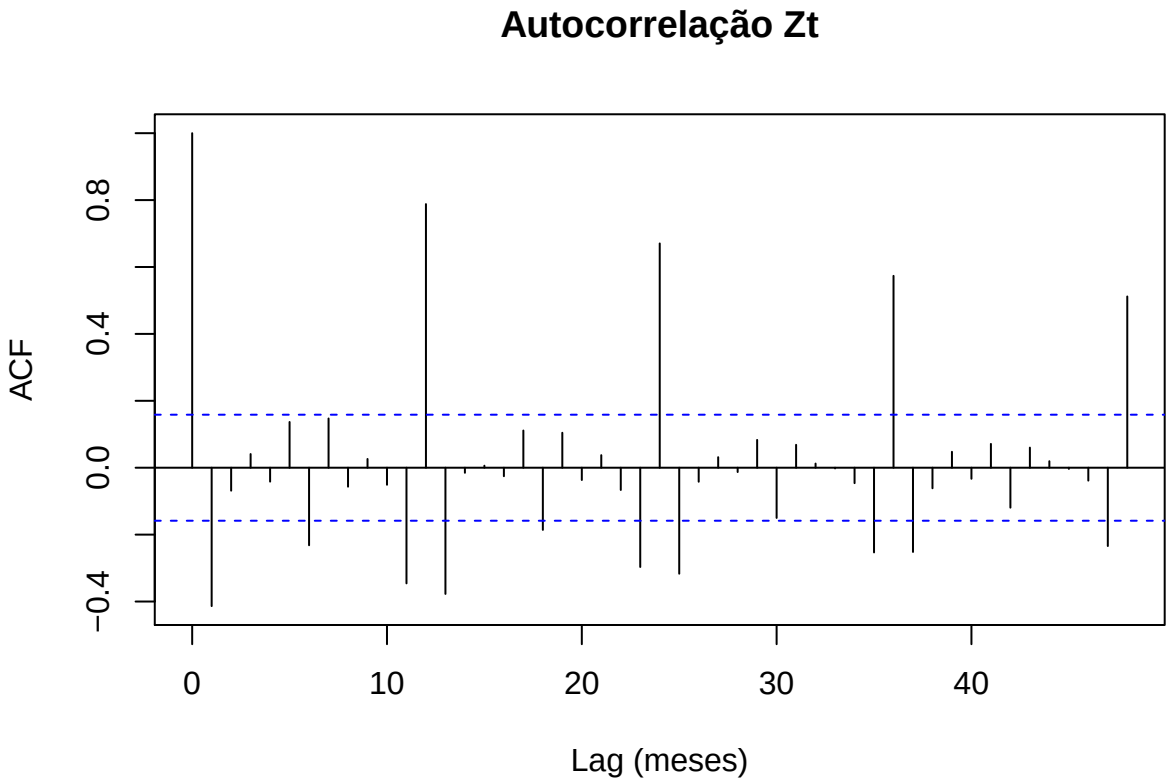


(a) O que você pode observar nas autocorrelações de Z_t?



Podemos observar um decaimento exponencial no gráfico acima que é não significante apenas após o lag 25. Também podemos observar que os lags 12, 23 e 36 tem picos fora do comum que indicam sazonalidade.

(b) A mesma pergunta para ΔZ_t.



Resposta: Podemos observar que grande parte da ACF foi anulada, apenas os lags que indicam uma diferença dos picos de sazonalidade, e os próprios picos.

(c) Qual das séries você consideraria estacionária?

Resposta: Não consideramos nenhuma das séries estacionária. Quando fizemos a primeira diferença, vários lags desapareceram, o que indica uma uma tendência na série original. Porém a segunda série também apresenta lags nos períodos 12, 24, 36 e 48, o que indica a sazonalidade que já havíamos notado.

(d) Utilizando um programa de identificação, sugira um ou mais modelos adequados para a série; obtenha as estimativas preliminares para os parâmetros.

Resposta: Vamos utilizar a função `auto.arima` para identificar o modelo.

```
## Series: consumo
## ARIMA(1,0,1)(0,1,1)[12]
##
## Coefficients:
##          ar1      ma1     sma1
##       0.9407  -0.2294  -0.7340
## s.e.  0.0324   0.0928   0.0739
##
## sigma^2 estimated as 93.81:  log likelihood=-527.47
## AIC=1062.94   AICc=1063.24   BIC=1074.77
##
## Training set error measures:
##                      ME     RMSE      MAE        MPE     MAPE      MASE
```

```
## Training set -0.3770845 9.201688 7.100073 -0.5534727 6.053567 0.4431577
##                     ACF1
## Training set 0.01366012
```

Vemos que o programa identifica um ARIMA(1,0,1)(0,1,1)[12]. Além disso, ele mostra as estimativas preliminares dos coeficientes.

Também podemos fazer uma identificação de um modelo alternativo feito artesanalmente. Primeiro fazemos uma transformação Box-Cox com o parâmetro λ ótimo. Depois podemos fazer um teste de estacionariedade. Porém o teste ADF acusa que não rejeitamos a hipótese nula (de não estacionariedade) a nível 0,05.

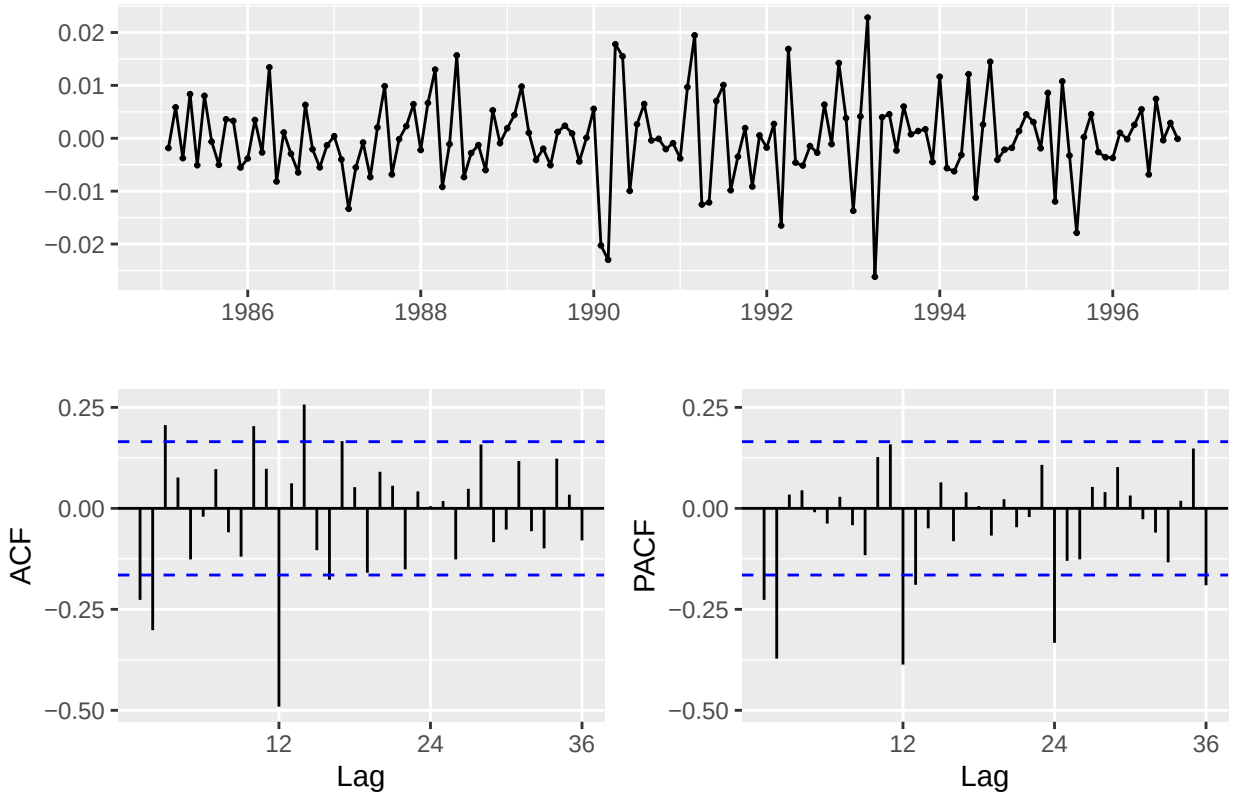
```
##
##  Augmented Dickey-Fuller Test
##
## data:  consumo.bc
## Dickey-Fuller = -3.2868, Lag order = 5, p-value = 0.07602
## alternative hypothesis: stationary
```

Portanto, fazemos uma diferença e refazemos o teste.

```
## Warning in adf.test(consumo.diff): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  consumo.diff
## Dickey-Fuller = -6.9311, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

Nesse caso possuimos a estacionaridade desejada. Como ainda temos a sazonalidade já observada, também fazemos a diferença sazonal e, assim:



Observe que a ACF (morte após lag 12) + PACF (decaimento exponencial) nos lags da sazonalidade indicam um componente MA sazonal. Vamos utilizar os critérios de informação para definir o modelo final. Os valores máximos de p e q serão 3 porque tanto a ACF quando a PACF parecem morrer cedo.

```
## [1] "Modelo com menor AIC"
```

```
##   p q       AIC       BIC      AICc
## 9 2 0 -1061.232 -1049.437 -1060.938
```

```
## [1] "Modelo com menor AICc"
```

```
##   p q       AIC       BIC      AICc
## 9 2 0 -1061.232 -1049.437 -1060.938
```

```
## [1] "Modelo com menor BIC"
```

```
##   p q       AIC       BIC      AICc
## 9 2 0 -1061.232 -1049.437 -1060.938
```

Em particular, as três informações escolheram o modelo ARIMA(2,1,0)(0,1,1)[12]. Por isso, esse é nosso modelo alternativo.

(e) Obtenha as estimativas finais para os parâmetros do(s) modelo(s) através de um programa de estimação; verifique se o(s) modelo(s) é (são) adequado(s).

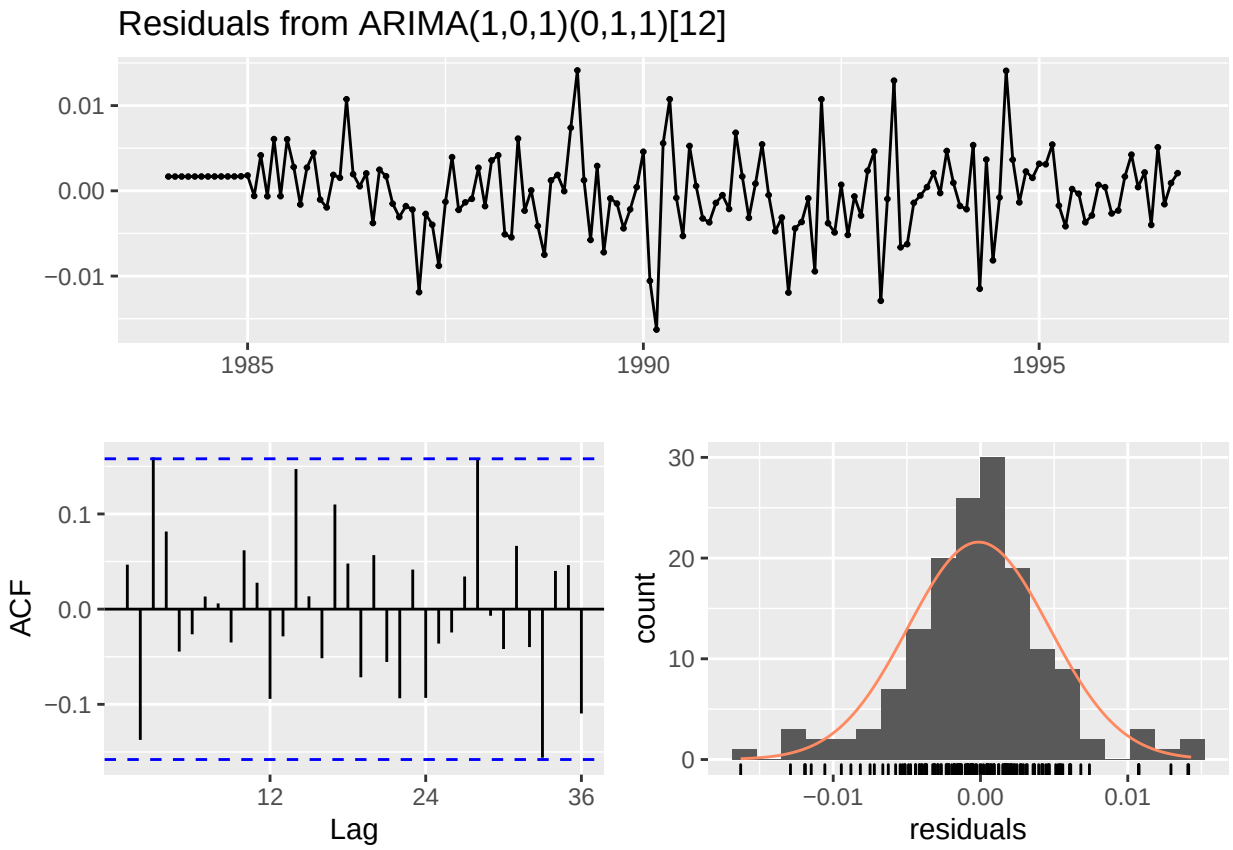
Resposta: Será utilizada a função `Arima`, que estima os parâmetros do modelo de dada ordem. Além disso, ela permite incluir o parâmetro `lambda` para a transformação de Box-Cox.

Vejamos o resultado da estimação para um modelo ARIMA(1,0,1)(0,1,1)[12] com parâmetro `lambda` escolhido automaticamente:

```
## Series: consumo
## ARIMA(1,0,1)(0,1,1)[12]
## Box Cox transformation: lambda= -0.5531393
##
## Coefficients:
##          ar1      ma1     sma1
##       0.9757  -0.3881  -1.0000
## s.e.  0.0227   0.0828   0.1081
##
## sigma^2 estimated as 2.547e-05:  log likelihood=536.49
## AIC=-1064.99   AICc=-1064.7   BIC=-1053.17
##
## Training set error measures:
##                       ME     RMSE      MAE        MPE     MAPE      MASE
## Training set 0.07607136 7.705328 5.810603 -0.2945207 4.892397 0.3626742
##                    ACF1
## Training set 0.1217969
```

Observamos que as estimativas atuais dos parâmetros são próximas daquelas preliminares (do exercício anterior). O parâmetro `lambda` também é estimado.

Vamos analisar os resíduos, através de seu gráfico, sua ACF e seu histograma:

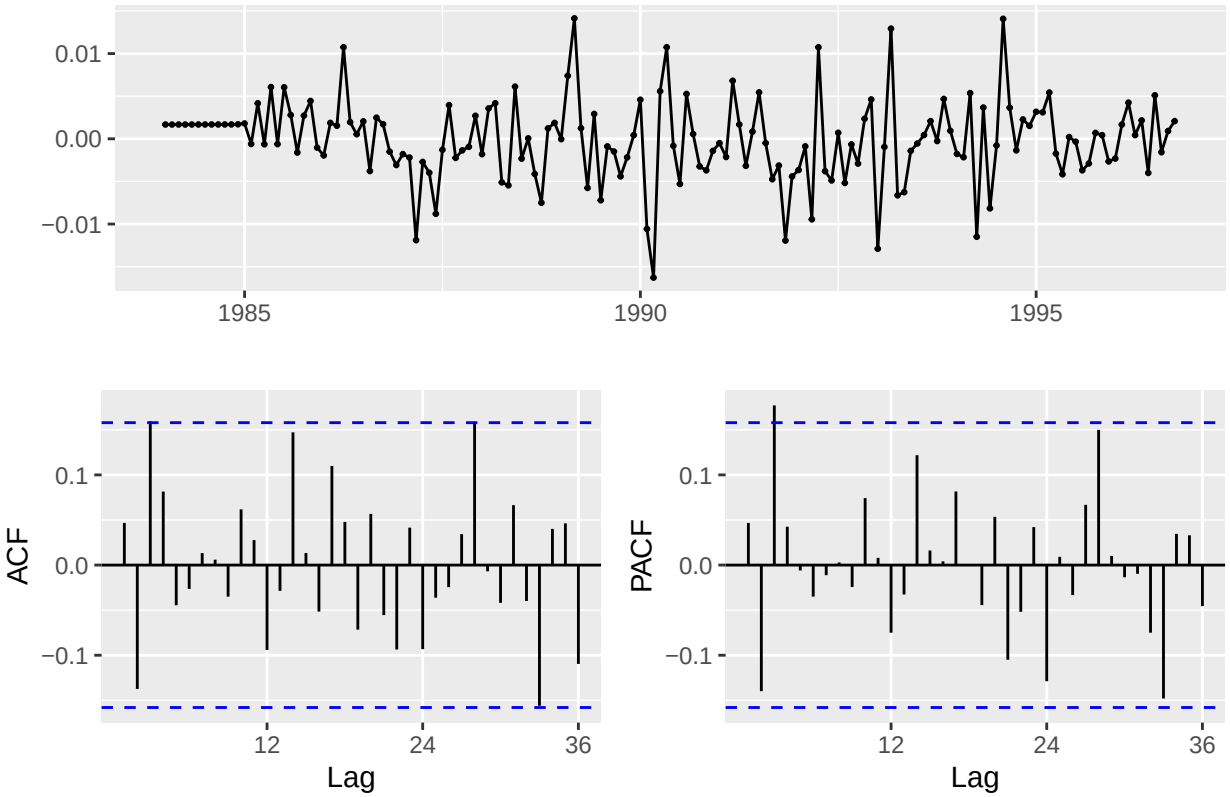



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,1)(0,1,1)[12]
## Q* = 23.808, df = 21, p-value = 0.3025
##
## Model df: 3.   Total lags used: 24
```

O teste de Ljung-Box nos resíduos nos dá boas notícias: não rejeitamos a hipótese de dados não correlacionados. Observamos que a ACF nos indica que os resíduos se comportam como esperaríamos de um ruído gaussiano. Todavia, o teste jarque.bera rejeita a hipótese nula de que a assimetria e o excesso de curtose são nulos. Isso nos dá evidência para a não normalidade.

```
##
##  Jarque Bera Test
##
## data:  est$residuals
## X-squared = 15.555, df = 2, p-value = 0.0004192
```

Vejamos o resultado de `ggtsdisplay` para os resíduos:

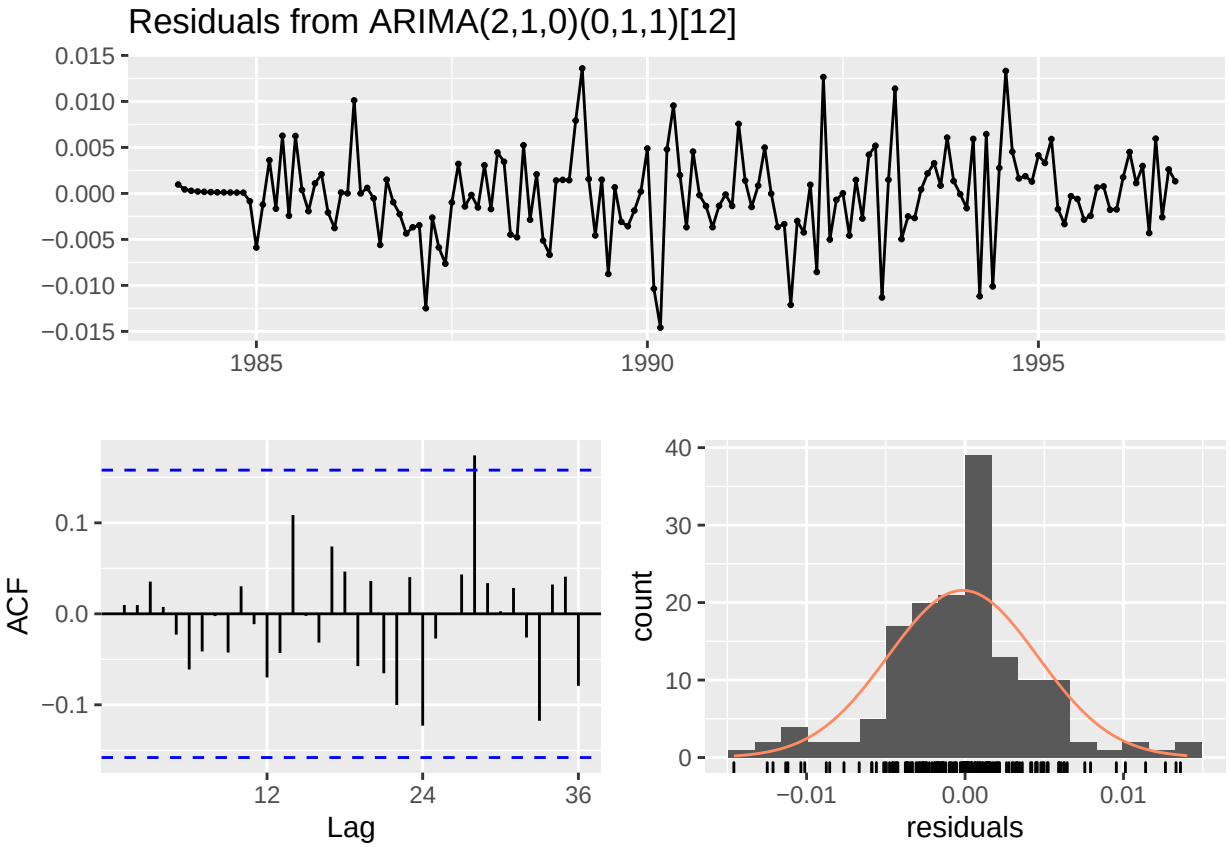


A PACF também se comporta dentro do intervalo de confiança para que consideremos o resíduo como um ruído branco.

Para o modelo alternativo ARIMA(2,1,0)(0,1,1)[12], façamos o mesmo processo.

```
## Series: consumo
## ARIMA(2,1,0)(0,1,1)[12]
## Box Cox transformation: lambda= -0.5531393
##
## Coefficients:
##         ar1      ar2     sma1
##       -0.34  -0.3075  -0.9999
## s.e.   0.08   0.0795   0.1078
##
## sigma^2 estimated as 2.474e-05:  log likelihood=534.62
## AIC=-1061.23   AICc=-1060.94   BIC=-1049.44
##
## Training set error measures:
##                       ME     RMSE      MAE        MPE     MAPE     MASE
## Training set -0.3753811 7.708961 5.572504 -0.4295857 4.704088 0.347813
##                    ACF1
## Training set 0.1034173
```

Observamos que alguns indicadores de erros nos mostram que o modelo é pior: tem maior erro média, maior raiz do erro quadrático e erro percentual, mas tem menor MAE, MAPE, MASE e ACF1.




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,1,0)(0,1,1)[12]
## Q* = 12.873, df = 21, p-value = 0.913
##
## Model df: 3.   Total lags used: 24
```

A ACF dos resíduis está particularmente boa! Além disso o teste de Ljung-Box não rejeita a hipótese nula de que não existe correlação, o que é uma boa notícia. Vamos ver o Jarque-Bera.

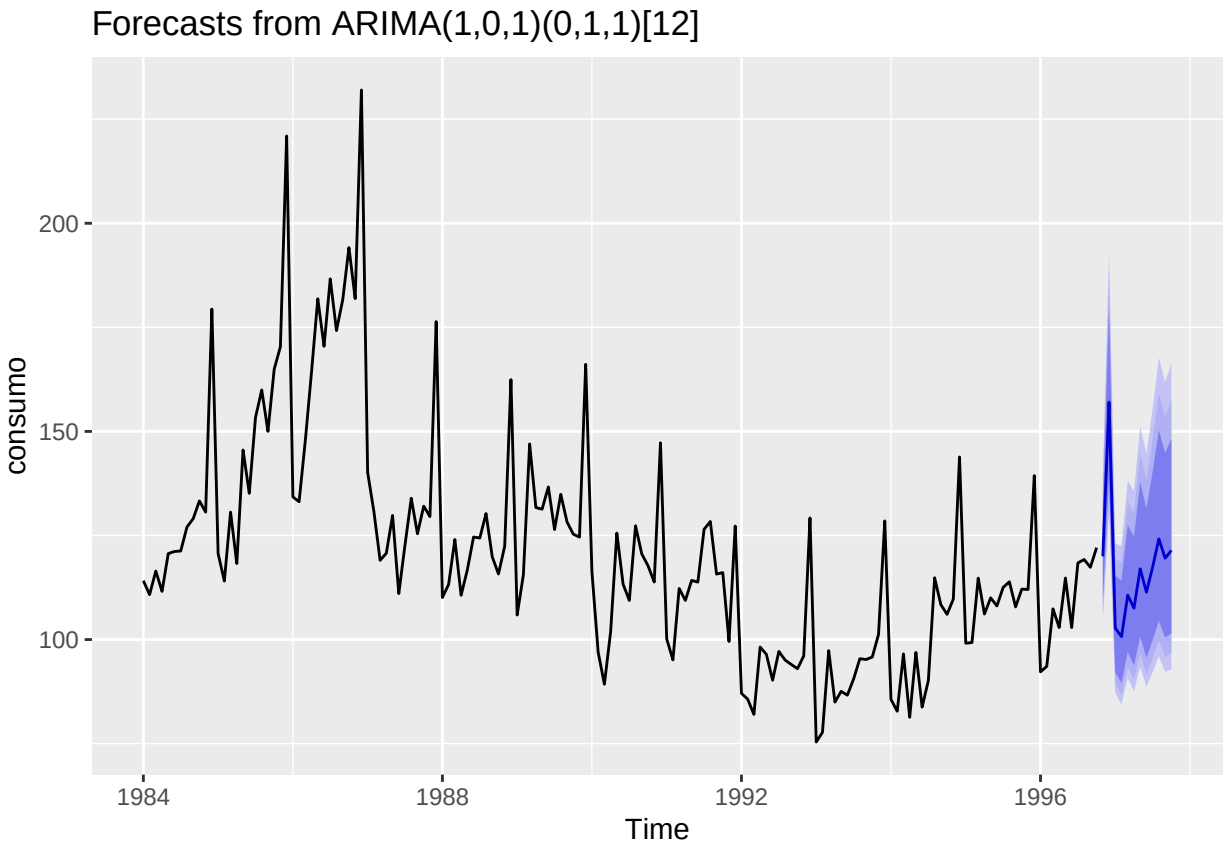
```
##
##  Jarque Bera Test
##
## data:  est2$residuals
## X-squared = 10.213, df = 2, p-value = 0.006057
```

O teste de Jarque Bera também rejeita a hipótese nula, o que não nos dá muito *insight* sobre o modelo ideial.

Vamos utilizar o primeiro modelo para previsão, dado que obteve menor RMSE.

(f) Obtenha previsões para 1974 utilizando o(s) modelo(s) estimado(s).

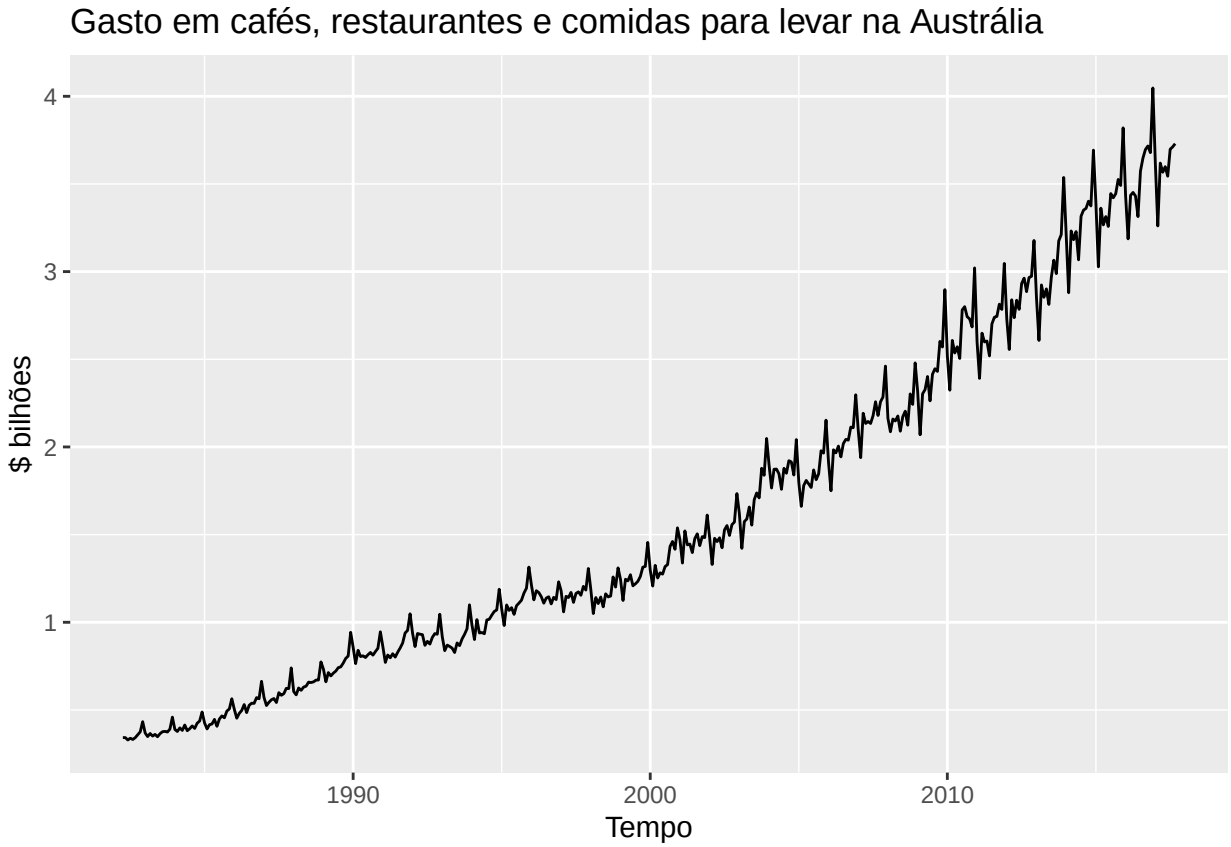
Resposta: O ano 1974 é no passado. A última observação é de outubro de 1996. Vamos prever 12 meses à frente utilizando nosso modelo e também mostrando os intervalos de confiança:



Visualmente, isso parece um tanto razoável.

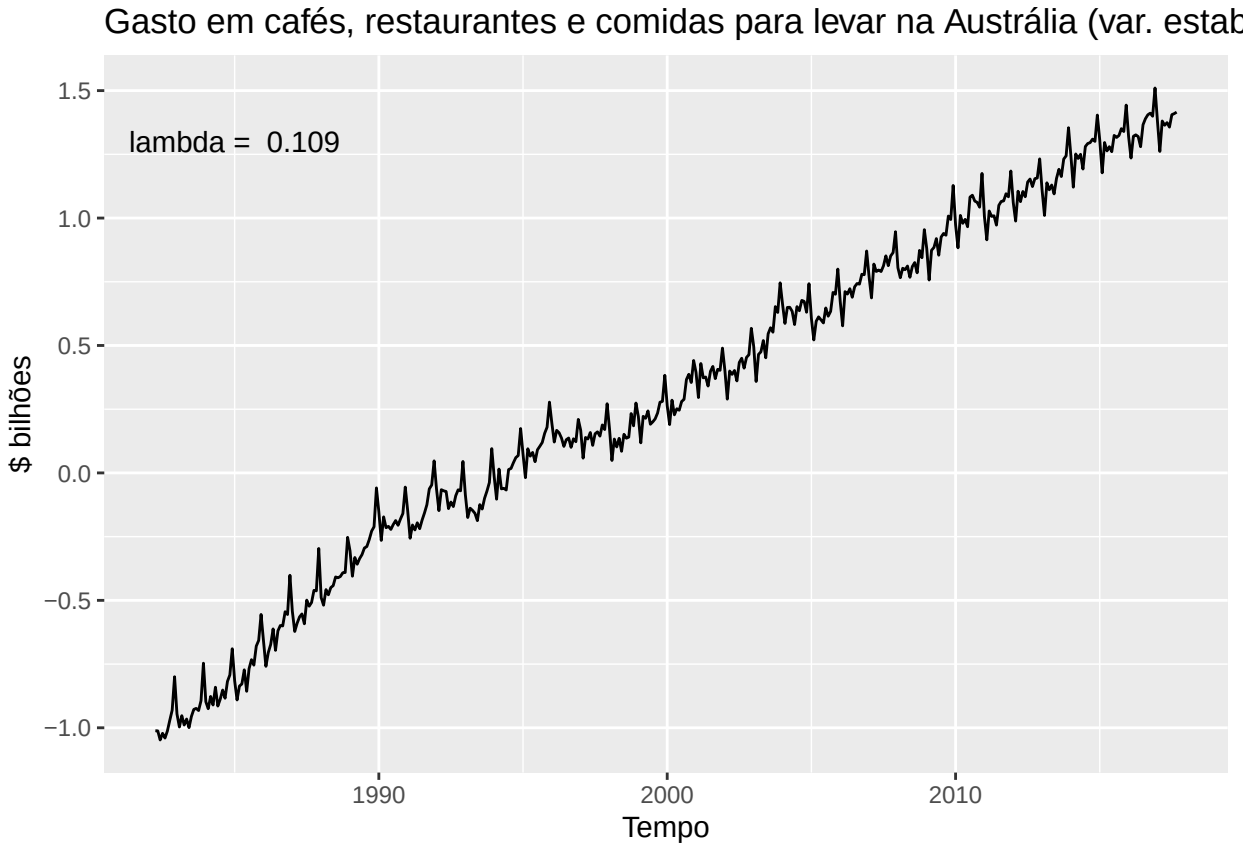
Série de serviços da Austrália

Essa série representa o gasto mensal total em serviços de cafés, restaurantes e comida para levar na Austrália em bilhões de dólares. Vamos visualizar a série:



Estabilização da variância

Observamos que a variância está aumentando, então é necessário estabilizá-la. Vamos utilizar a função `BoxCox`, que calcula automaticamente o parâmetro ótimo `lambda`. Vejamos o resultado:



Estacionariedade e diferenciação

Verifiquemos agora estacionariedade e necessidade de diferenciação. Inicialmente, utilizamos o teste ADF para a verificação da estacionariedade:

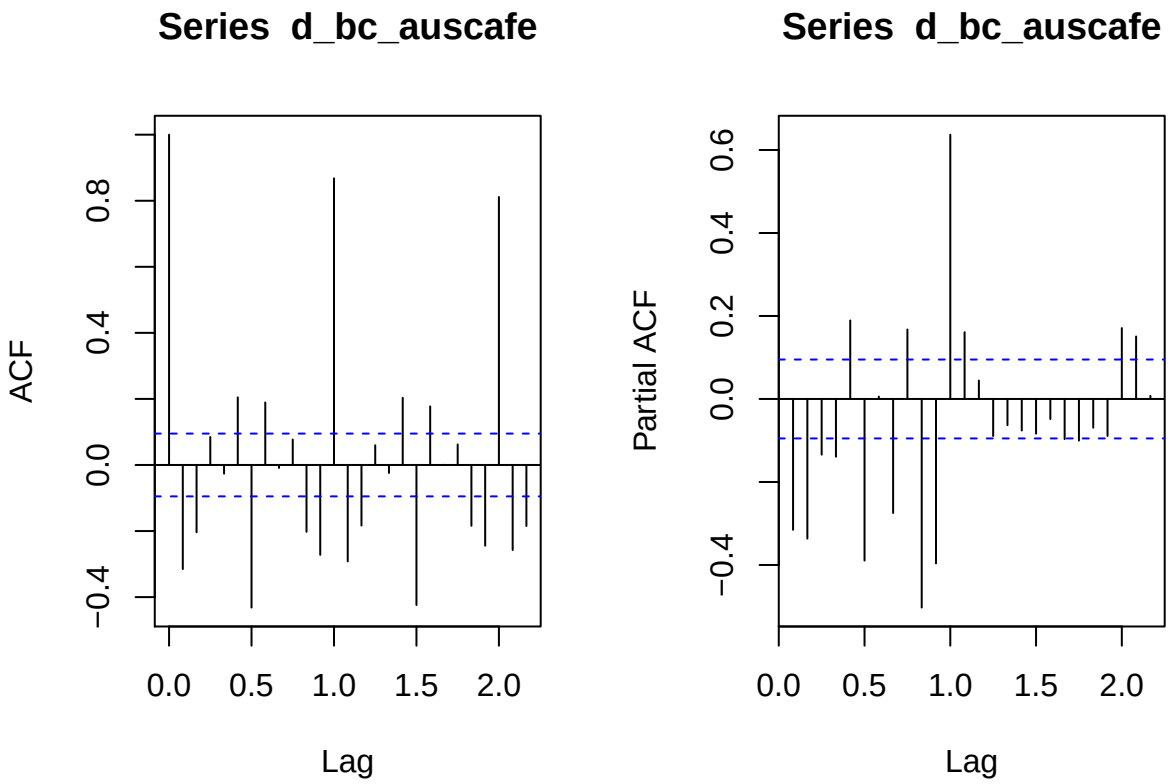
```
##
##  Augmented Dickey-Fuller Test
##
## data:  bc_auscafe
## Dickey-Fuller = -3.3838, Lag order = 7, p-value = 0.05669
## alternative hypothesis: stationary
```

Não rejeitamos a hipótese de não estacionariedade, logo decidimos por uma diferenciação. Diferenciando e testando:

```
## Warning in adf.test(d_bc_auscafe): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  d_bc_auscafe
## Dickey-Fuller = -13.205, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

Agora, sim, temos um modelo estacionário. Como estamos trabalhando com modelos SARIMA, verificamos também a necessidade de uma diferenciação sazonal. Verificamos isso com os gráficos da ACF e da PACF:

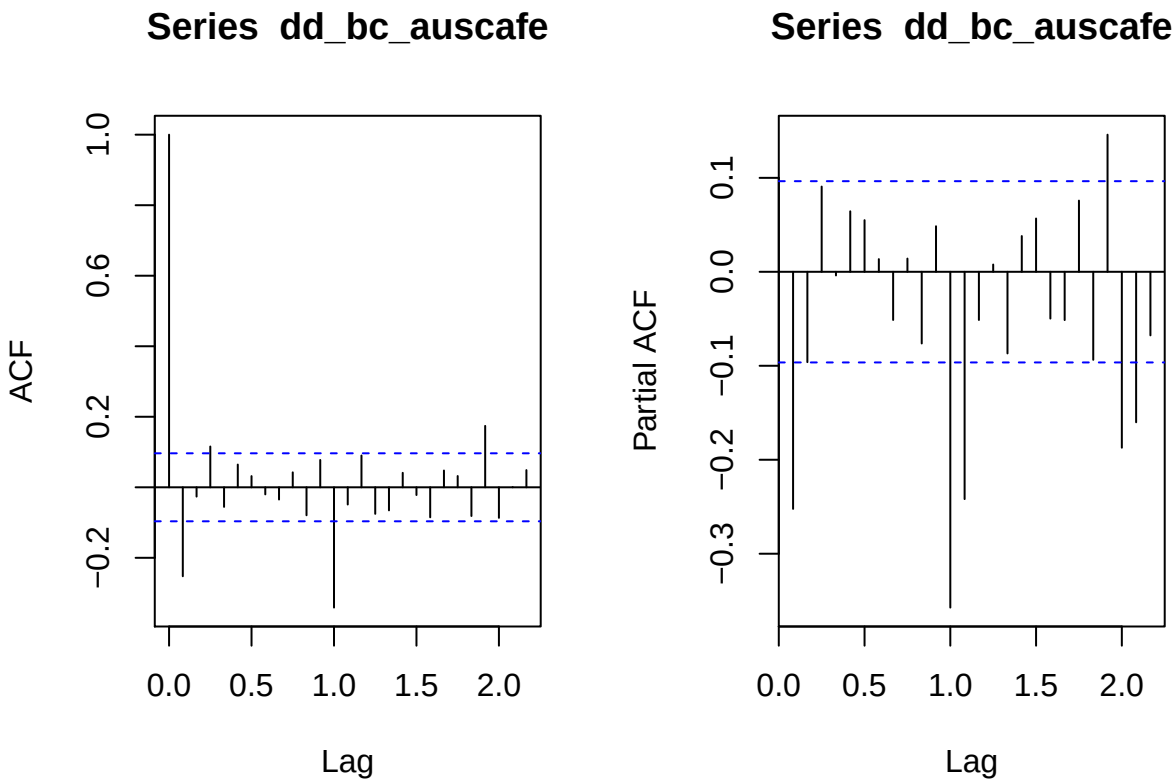


A presença de sazonalidade no lag 12 (e seus múltiplos e divisores) é clara, tanto nos gráficos da ACF e PACF, quanto no próprio gráfico da série. Decidimos portanto por realizar uma diferenciação sazonal de lag 12.

```
## Warning in adf.test(dd_bc_auscafe): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  dd_bc_auscafe
## Dickey-Fuller = -6.7817, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

Ainda mantemos a estacionariedade!



Os picos diminuíram de forma considerável. Os picos na ACF (1, 12 e 23) indicam um termo MA(1) (corte depois de $q > 1$) e um termo sazonal MA(1) (*spike* em lag 12). Os termos autoregressivos estão mais difíceis de interpretar se baseando apenas nos gráficos da ACF e PACF. Vamos utilizar uma abordagem de "gridsearch" para tentar encontrar o modelo mais simples que minimiza os critérios de informação.

Critério de informação

Vamos realizar um "gridsearch" entre 0 e 3 para os parâmetros p e P (com exceção de alguns modelos específicos que tiveram problemas na otimização durante a estimação).

p	P	AIC	BIC	AICc
0	0	-1882.555	-1866.461	-1882.457
0	1	-1884.761	-1864.644	-1884.614
0	2	-1883.760	-1859.619	-1883.553
0	3	-1891.388	-1863.224	-1891.111
1	0	-1881.042	-1860.925	-1880.895
1	1	-1883.137	-1858.997	-1882.931
1	2	-1882.114	-1853.950	-1881.837
2	0	-1887.723	-1863.583	-1887.517
2	1	-1889.001	-1860.837	-1888.725
2	2	-1887.733	-1855.545	-1887.376
2	3	-1893.137	-1856.926	-1892.690
3	0	-1885.751	-1857.587	-1885.475
3	2	-1885.708	-1849.497	-1885.262
3	3	-1891.188	-1850.953	-1890.640

Verificamos que minimizamos AIC e AICc com $p = 2$, $P = 3$, enquanto minimizamos o BIC com $p = P = 0$. Vamos trabalhar ambos os modelos.

Estimação

Vamos utilizar a função `Arima` para ajustar os modelos SARIMA(2, 1, 1)(3, 1, 1)[12] e SARIMA(0, 1, 1)(0, 1, 1)[12], com transformação de Box-Cox (automática).

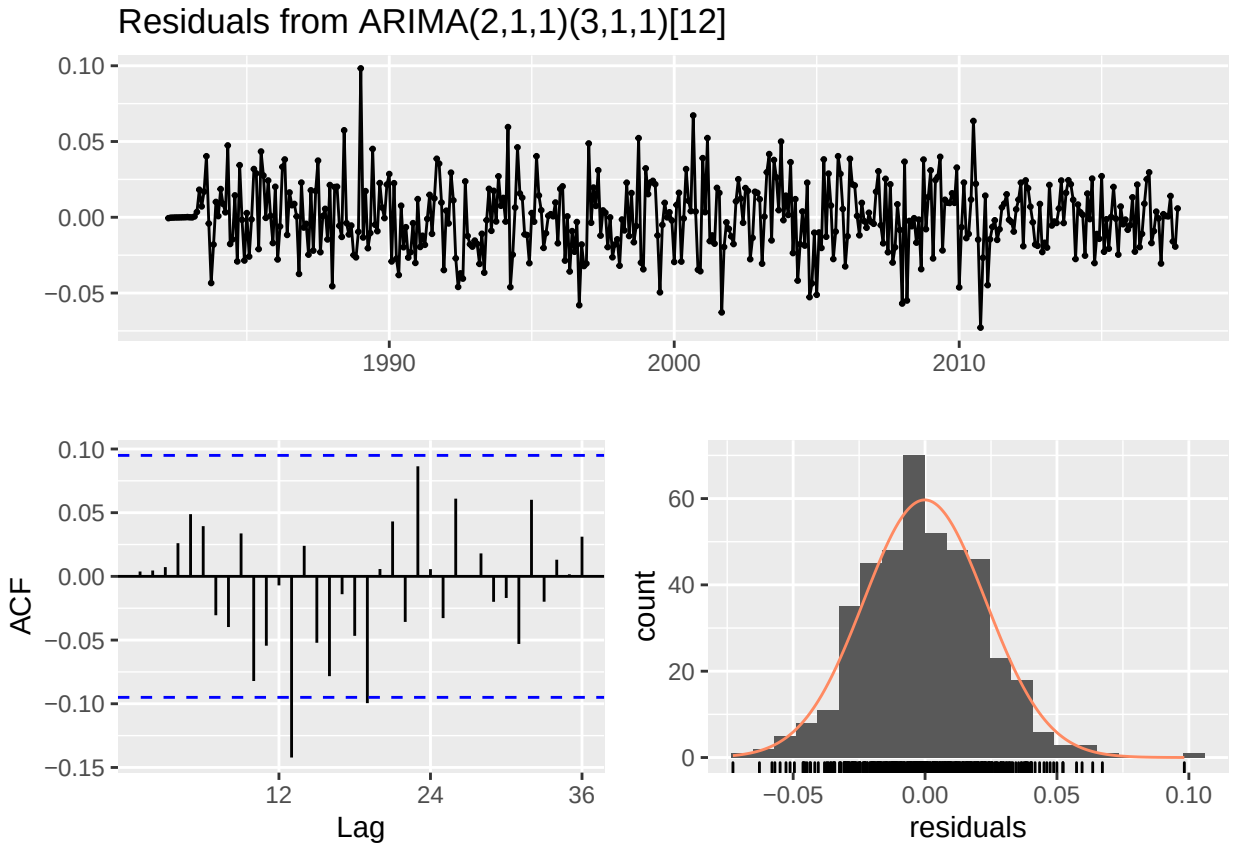
```
## Series: auscafe
## ARIMA(2,1,1)(3,1,1)[12]
## Box Cox transformation: lambda= 0.109056
##
## Coefficients:
##           ar1      ar2     ma1    sar1     sar2     sar3     sma1
##       -0.8834  -0.2952  0.5479  0.0017  -0.1058  -0.1699  -0.7419
## s.e.   0.2250   0.0682  0.2321  0.0740   0.0610   0.0599   0.0619
##
## sigma^2 estimated as 0.0005634:  log likelihood=955.47
## AIC=-1894.93   AICc=-1894.57   BIC=-1862.74
##
## Training set error measures:
##                         ME       RMSE        MAE         MPE     MAPE      MASE
## Training set -0.0006347659 0.03676802 0.02647835 -0.03142307 1.756252 0.2534141
##                       ACF1
## Training set -0.002408055
```

```
## Series: auscafe
## ARIMA(0,1,1)(0,1,1)[12]
## Box Cox transformation: lambda= 0.109056
##
## Coefficients:
##           ma1     sma1
##       -0.3649  -0.8204
## s.e.   0.0431   0.0325
##
## sigma^2 estimated as 0.0005856:  log likelihood=945.15
## AIC=-1884.3   AICc=-1884.24   BIC=-1872.23
##
## Training set error measures:
##                         ME       RMSE        MAE         MPE     MAPE      MASE
## Training set -0.0007623818 0.03690703 0.02723568 -0.03770534 1.833843 0.2606623
##                       ACF1
## Training set -0.006901538
```

Vale notar que, por mais que estamos deixando a função escolher o parâmetro `lambda`, ele escolhe o mesmo que encontramos anteriormente. (Existe uma vantagem em deixar a função fazer isso: o modelo todo é estimado com uma função só, então fica trivial de utilizar funções de previsão que automaticamente calculam intervalos preditivos, por exemplo.)

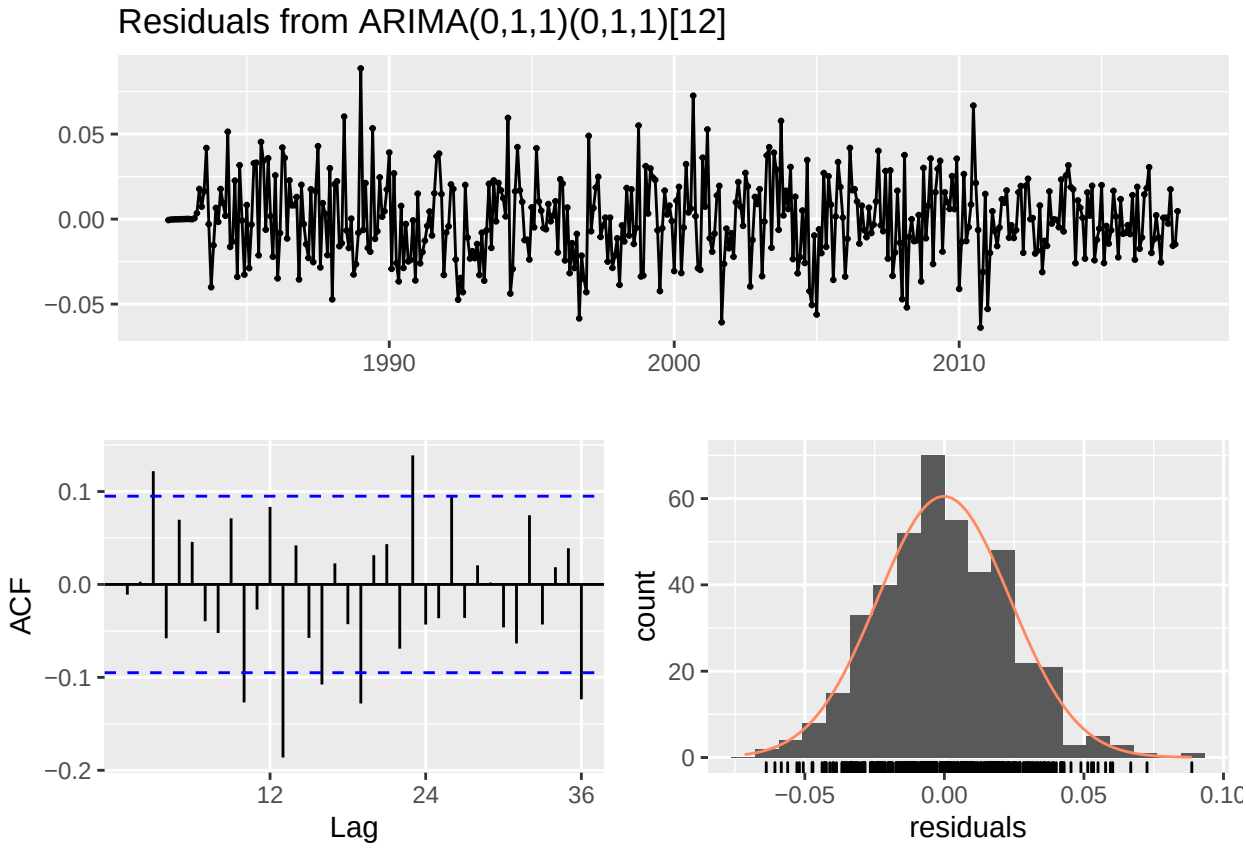
Diagnóstico

Precisamos verificar os resíduos das nossas estimações. Fazemos isso com a função `checkresiduals`, e também aplicamos um teste de Jarque-Bera (este tem hipótese nula de assimetria 0 e excesso de curtose 3).




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,1,1)(3,1,1)[12]
## Q* = 31.339, df = 17, p-value = 0.01816
##
## Model df: 7.   Total lags used: 24
```

```
##
##  Jarque Bera Test
##
## data:  mod1$residuals
## X-squared = 8.4329, df = 2, p-value = 0.01475
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,1)(0,1,1)[12]
## Q* = 69.58, df = 22, p-value = 7.7e-07
##
## Model df: 2.   Total lags used: 24
```

```
##
##  Jarque Bera Test
##
## data:  mod2$residuals
## X-squared = 4.0148, df = 2, p-value = 0.1343
```

Vamos discutir os resultados obtidos. O modelo SARIMA(2, 1, 1)(3, 1, 1)[12] rejeita as observações serem i.i.d., porém rejeita também a normalidade. Observamos valores extremos nos resíduos, assim como spike em lag 13. O modelo SARIMA(0, 1, 1)(0, 1, 1)[12] rejeita as observações serem i.i.d., mas não rejeita a normalidade! Vemos que várias autocorrelações estão fora do intervalo de confiança.

Com a posse de todas essas informações, acabamos por escolher o modelo SARIMA(0, 1, 1)(0, 1, 1)[12], que, mesmo que não seja perfeito, é um dos melhores no ajuste aos dados e também é o mais simples entre os dois. Observe ainda que este modelo é também escolhido pela função `auto.arima`:

```
## Series: auscafe
## ARIMA(0,1,1)(0,1,1)[12]
##
## Coefficients:
##           ma1     sma1
##       -0.3673  -0.5991
```

```
## s.e.   0.0443   0.0348
##
## sigma^2 estimated as 0.001673:  log likelihood=732.46
## AIC=-1458.92   AICc=-1458.87   BIC=-1446.85
```